Analysis of Microdata

Rainer Winkelmann
Stefan Boes

Analysis of Microdata

With 38 Figures
and 41 Tables

Springer

Professor Dr. Rainer Winkelmann
Dipl. Vw. Stefan Boes

University of Zurich
Socioeconomic Institute
Zürichbergstrasse 14
8032 Zurich
Switzerland
E-mail: winkelmann@sts.unizh.ch
E-mail: boes@sts.unizh.ch

Cataloging-in-Publication Data

Library of Congress Control Number: 2005935030

ISBN-10 3-540-29605-0 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-29605-8 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media
springeronline.com

Printed in Germany

Cover design: Erich Kirchner
Production: Helmut Petri
Printing: Strauss Offsetdruck

SPIN 11573999 Printed on acid-free paper – 42/3153 – 5 4 3 2 1 0

Preface

The availability of microdata has increased rapidly over the last decades, and standard statistical and econometric software packages for data analysis include ever more sophisticated modeling options. The goal of this book is to familiarize readers with a wide range of commonly used models, and thereby to enable them to become critical consumers of current empirical research, and to conduct their own empirical analyses.

The focus of the book is on regression-type models in the context of large cross-section samples. In microdata applications, dependent variables often are qualitative and discrete, while in other cases, the sample is not randomly drawn from the population of interest and the dependent variable is censored or truncated. Hence, models and methods are required that go beyond the standard linear regression model and ordinary least squares. Maximum likelihood estimation of conditional probability models and marginal probability effects are introduced here as the unifying principle for modeling, estimating and interpreting microdata relationships. We consider the limitation to maximum likelihood sensible, from a pedagogical point of view if the book is to be used in a semester-long advanced undergraduate or graduate course, and from a practical point of view because maximum likelihood estimation is used in the overwhelming majority of current microdata research.

In order to introduce and explain the models and methods, we refer to a number of illustrative applications. The main examples include the determinants of individual fertility, the intergenerational transmission of secondary school choices, and the wage elasticity of female labor supply. The models presented, while chosen with economic applications in mind, should be equally relevant for other social sciences, for example, quantitative political science and sociology, and for empirical disciplines outside of the social sciences.

The book can be used as a textbook for an advanced undergraduate, a Master's or a first-year Ph.D. course on the topic of microdata analysis. In economics and related disciplines, such a course is typically offered after a first course on linear regression analysis. Alternatively, the book can also serve as a supplementary text to an applied microeconomics field course, such as

those offered in the areas of labor economics, health economics, and the like. Finally, it is intended as a reference for graduate students, researchers as well as practitioners who encounter microdata in their work. The mathematical prerequisites are not very high. In particular, the use of linear algebra is minimal. On the other hand, some background in mathematical statistics is useful although not absolutely necessary.

The book includes numerous exercises. Most of the exercises do not require the use of a computer. Rather, they typically present specific empirical results, and the task is to assess the validity of the procedure in that particular context and to provide a correct interpretation of the estimated parameters. In addition, we encourage the reader to develop practical skills in applied data analysis by re-estimating the examples we discuss, using a software of choice. For this purpose, we have made the datasets employed available at our homepage www.unizh.ch/sts/, both in ASCII format and in Stata 7 format.

An earlier version of the manuscript was used in a course of the same name taught by us for several years at the economics department of the University of Zurich. We thank the participants for numerous suggestions for improvement. We are heavily indebted to Markus Lipp and Adrian Bruhin for careful proofreading, to Markus in addition for creating all the figures, and to Deborah Bowen for improving our English.

Zurich, September 2005 *Rainer Winkelmann*
Stefan Boes

Contents

1 Introduction 1
1.1 What Are Microdata? 1
1.2 Types of Microdata 4
1.2.1 Qualitative Data 4
1.2.2 Quantitative Data 6
1.3 Why Not Linear Regression? 8
1.4 Common Elements of Microdata Models 10
1.5 Examples 11
1.5.1 Determinants of Fertility 11
1.5.2 Secondary School Choice 16
1.5.3 Female Hours of Work and Wages 17
1.6 Overview of the Book 19

2 From Regression to Probability Models 21
2.1 Introduction 21
2.2 Conditional Probability Functions 23
2.2.1 Definition 23
2.2.2 Estimation 24
2.2.3 Interpretation 25
2.3 Probability and Probability Distributions 29
2.3.1 Axioms of Probability 29
2.3.2 Univariate Random Variables 30
2.3.3 Multivariate Random Variables 31
2.3.4 Conditional Probability Models 34
2.4 Further Exercises 39

3 Maximum Likelihood Estimation 45
3.1 Introduction 45
3.2 Likelihood Function 46
3.2.1 Score Function and Hessian Matrix 48
3.2.2 Conditional Models 50

3.2.3 Maximization 50
3.3 Properties of the Maximum Likelihood Estimator 53
3.3.1 Expected Score 54
3.3.2 Consistency 55
3.3.3 Information Matrix Equality 56
3.3.4 Asymptotic Distribution 59
3.3.5 Covariance Matrix 60
3.4 Normal Linear Model 63
3.5 Further Aspects of Maximum Likelihood Estimation 67
3.5.1 Invariance and Delta Method 67
3.5.2 Numerical Optimization 69
3.5.3 Identification 74
3.5.4 Quasi Maximum Likelihood 76
3.6 Testing 76
3.6.1 Introduction 76
3.6.2 Restricted Maximum Likelihood 79
3.6.3 Wald Test 81
3.6.4 Likelihood Ratio Test 83
3.6.5 Score Test 86
3.6.6 Model Selection 88
3.6.7 Goodness-of-Fit 89
3.7 Pros and Cons of Maximum Likelihood 89
3.8 Further Exercises 90

4 Binary Response Models 95
4.1 Introduction 95
4.2 Models for Binary Response Variables 97
4.2.1 General Framework 97
4.2.2 Linear Probability Model 98
4.2.3 Probit Model 100
4.2.4 Logit Model 102
4.2.5 Interpretation of Parameters 104
4.3 Discrete Choice Models 107
4.4 Estimation 110
4.4.1 Maximum Likelihood 110
4.4.2 Perfect Prediction 113
4.4.3 Properties of the Estimator 114
4.4.4 Endogenous Regressors in Binary Response Models 116
4.4.5 Estimation of Marginal Effects 118
4.5 Goodness-of-Fit 122
4.6 Non-Standard Sampling Schemes 127
4.6.1 Stratified Sampling 127
4.6.2 Exogenous Stratification 127
4.6.3 Endogenous Stratification 128
4.7 Further Exercises 130

5 Multinomial Response Models 137
5.1 Introduction 137
5.2 Multinomial Logit Model 139
5.2.1 Basic Model 139
5.2.2 Estimation 140
5.2.3 Interpretation of Parameters 144
5.3 Conditional Logit Model 150
5.3.1 Introduction 150
5.3.2 General Model of Choice 151
5.3.3 Modeling Conditional Logits 152
5.3.4 Interpretation of Parameters 155
5.3.5 Independence of Irrelevant Alternatives 159
5.4 Generalized Multinomial Response Models 160
5.4.1 Multinomial Probit Model 161
5.4.2 Mixed Logit Models 163
5.4.3 Nested Logit Models 164
5.5 Further Exercises 166

6 Ordered Response Models 171
6.1 Introduction 171
6.2 Standard Ordered Response Models 174
6.2.1 General Framework 174
6.2.2 Ordered Probit Model 176
6.2.3 Ordered Logit Model 177
6.2.4 Estimation 179
6.2.5 Interpretation of Parameters 179
6.2.6 Single Indices and Parallel Regression 186
6.3 Generalized Threshold Models 188
6.3.1 Generalized Ordered Logit and Probit Models 188
6.3.2 Interpretation of Parameters 189
6.4 Sequential Models 194
6.4.1 Modeling Conditional Transitions 194
6.4.2 Generalized Conditional Transition Probabilities 197
6.4.3 Marginal Effects 197
6.4.4 Estimation 198
6.5 Interval Data 200
6.6 Further Exercises 202

7 Limited Dependent Variables 207
7.1 Introduction 207
7.1.1 Corner Solution Outcomes 208
7.1.2 Sample Selection Models 209
7.1.3 Treatment Effect Models 210
7.2 Tobin's Corner Solution Model 211
7.2.1 Introduction 211

7.2.2 Tobit Model . . . 212
7.2.3 Truncated Normal Distribution . . . 214
7.2.4 Inverse Mills Ratio and its Properties . . . 215
7.2.5 Interpretation of the Tobit Model . . . 218
7.2.6 Comparing Tobit and OLS . . . 221
7.2.7 Further Specification Issues . . . 223
7.3 Sample Selection Models . . . 224
7.3.1 Introduction . . . 224
7.3.2 Censored Regression Model . . . 226
7.3.3 Estimation of the Censored Regression Model . . . 228
7.3.4 Truncated Regression Model . . . 230
7.3.5 Incidental Censoring . . . 231
7.3.6 Example: Estimating a Labor Supply Model . . . 237
7.4 Treatment Effect Models . . . 239
7.4.1 Introduction . . . 239
7.4.2 Endogenous Binary Variable . . . 242
7.4.3 Switching Regression Model . . . 243
7.5 Appendix: Bivariate Normal Distribution . . . 246
7.6 Further Exercises . . . 247

8 Event History Models . . . 251
8.1 Introduction . . . 251
8.2 Duration Models . . . 254
8.2.1 Introduction . . . 254
8.2.2 Basic Concepts . . . 254
8.2.3 Discrete Time Duration Models . . . 259
8.2.4 Continuous Time Duration Models . . . 262
8.2.5 Key Element: Hazard Function . . . 265
8.2.6 Duration Dependence . . . 267
8.2.7 Unobserved Heterogeneity . . . 271
8.3 Count Data Models . . . 279
8.3.1 The Poisson Regression Model . . . 279
8.3.2 Unobserved Heterogeneity . . . 284
8.3.3 Efficient versus Robust Estimation . . . 289
8.3.4 Censoring and Truncation . . . 289
8.3.5 Hurdle and Zero-Inflated Count Data Models . . . 291
8.4 Further Exercises . . . 294

List of Figures . . . 297

List of Tables . . . 299

References . . . 301

Index . . . 309

1

Introduction

1.1 What Are Microdata?

This book is about the theory and practice of modeling microdata using statistical and econometric methods, in particular regression-type models, in which one variable is explained by a number of other variables. The defining feature of microdata – as we understand the term – is that their main dimension is cross-sectional, meaning that the basic sampling model is characterized by independence between observations. This excludes pure time series applications. Hybrid cases, such as panel data, can in principle be counted among microdata, in particular when the time dimension is short relative to the cross-sectional one, but we decided not to include such models in this book in order to keep the material covered manageable for a semester-long course. We recommend the textbooks by Baltagi (2005) and Hsiao (2003) for introductions to panel data methods.

Microeconometrics

All applications included in this book, and most of the literature we draw from, stem from the discipline of economics, reflecting our own background and preferences. Within economics, the subject matter of this book is also known as microeconometrics – the ensemble of econometric methods that have been developed to study microeconomic phenomena. In microeconometric studies, the empirical analysis is motivated by an economic question, and often such analyses start with a formal economic model or theory which is used to determine the quantities of interest and to derive testable hypotheses. The underlying model – in our case typically a microeconomic model where individual decisions and behavior are a function of exogenous parameters – offers guidance in the selection of the dependent and independent variables.

Economic Examples

Historically, many microeconometric methods have been developed with labor economic applications in mind. The three following examples are a reflection of this tradition. The human capital theory, for instance, predicts a positive relationship between wages, the dependent variable, and the level of education as a measure of human capital, the independent variable. Similarly, the simple static labor supply model posits that an exogenous wage rate defines the trade-off between consumption and leisure. Under utility maximization, the wage elasticity of labor supply, which can for example be measured by an individual's desired hours of work, depends on the individual preference structure and in particular on the relative magnitude of income and substitution effects, and thus, in principle, is indeterminate. Finally, anticipating a further example that will be used later on in this chapter, the number of children borne by a women is (or may be), among other things, a function of her labor market opportunities and thus her education.

Do We Need a Theory?

According to one school of thought, the more closely the empirical specification fits the underlying theoretical model, the more convincing the empirical analysis. Only with a fully **theory-based** analysis, as the argument goes, do the estimated parameters point to a well-defined economic interpretation and only then can the results be used for policy analysis.

While we have some sympathy for this point of view, it would be a mistake to require that all empirical analyses start with a fully fledged theoretical model. In some cases, a formal theory does not yet exist, and in others, the existing theories require modification. In these cases, empirical analysis has a **theory-building** function. Examples of intensive empirical activity without a well-established underlying theory are found in the current literature on the economic determinants of individual life-satisfaction (Frey and Stutzer, 2002, Layard, 2005), the literature on evaluating the effects of active labor market programs (Heckman, Lalonde and Smith, 1999), and the literature on the intergenerational transmission of education and income (Solon, 1999).

Importantly, the principles and empirical methods of analyzing microdata are largely independent of the underlying theory, if any, although the substantive – rather than the statistical – interpretation of the results may critically depend on it. Therefore, we feel justified in adhering to the principle of division of labor, i.e., focusing on the empirical models and mostly skipping the discussion of underlying theoretical models. This conceptual separation also underlines that the empirical methods covered in this book are not restricted to economic applications. The methods presented should be equally relevant for related social sciences, such as quantitative political science and sociology, as well as other disciplines, including biology and life-sciences. This, incidentally, is the reason for choosing the more general title of the book.

On the other hand, it would be wrong to introduce a further division of labor, one between econometric theory and data analysis. A main feature of microdata analysis is the almost symbiotic relationship between the empirical model and the data it is used for. Models are only defined and relevant in relation to certain types of data. Therefore, any student or researcher working with microdata needs to develop a good grasp of the underlying data structures as well as the associated empirical methods.

Defining Microdata

As the above remarks foreshadow, the notion of microdata that is used here encompasses a great variety of data types and applications. The most common situation is probably the one where microdata provide subjective or objective information on individual units such as persons, households or firms. This information may have been purposefully collected from surveys, or it may be the by-product of other activities (such as keeping and administering official tax or health records). In other instances, the observations can be a sample of transactions, such as supermarket-scanner and auction data, or a cross-section of countries.

The three most important features of microdata – as defined here – are that they are **cross-sectional**, that they are **observational**, and that they often have a **non-continuous measurement scale**. The term "observational" contrasts the collection of data from surveys and administrative records with those from a (randomized) experiment. While such "experimental" data are increasingly available in the social sciences, their use is restricted to very specific questions and applications, and the bulk of empirical work continues to rely on non-experimental data. Observational data may be subject to systematic sample selection, a problem that is discussed in detail in this book.

The different possibilities of scaling a variable are discussed in any introductory statistics course. These include the distinction between continuous and discrete variables, as well as the distinction between quantitative and qualitative (or categorical) variables. But when it comes to regression analysis with microdata, these distinctions are often forgotten and the linear regression model is inappropriately applied even when the dependent variable is measured on a non-continuous scale.

Micro versus Macrodata

Finally, note that microdata and microeconometrics can be usefully contrasted with macrodata and **macroeconometrics**, respectively. Macroeconometrics denotes the methods for the empirical study of macroeconomic phenomena based mostly on time series macrodata from national accounts. While the micro/macro distinction is inconsequential for the classical linear regression model – where it is largely a matter of taste and emphasis whether the model

is written with an i or with a t subscript – the distinction becomes important as soon as the standard assumptions of the linear regression model are violated. The typical departures from the standard assumptions are very different, depending on whether one deals with micro- or with macro data. An overview of the potential limitations of linear regression analysis when applied to microdata is given in Section 1.3.

1.2 Types of Microdata

The most basic distinction among types of microdata is certainly the one between **quantitative** and **qualitative** data. The latter are also referred to as **categorical**. Qualitative data are always discrete. The three types of qualitative data are binary, multinomial, and ordered. Quantitative data may be discrete or continuous. The separation between discrete and continuous quantitative data is a gradual one. While all measurements have finite precision and are therefore discrete in a strict sense, this may be ignored in most cases – we then also speak of quasi-continuous data. An exception are counts, where the discrete support should be taken into account.

Among quantitative data, one can further distinguish between data with restricted and unrestricted range. Variables may be non-negative: for example, many financial variables (like income), durations and counts. Alternatively, quantitative variables may be censored, truncated, or grouped. Although both discrete and continuous quantitative variables can be subject to censoring and truncation in principle, we only cover the continuous case in this book. Such variables – if used as dependent variable – are commonly referred to as **limited dependent variables**. Figure 1.1 illustrates the various types of microdata we consider in this book.

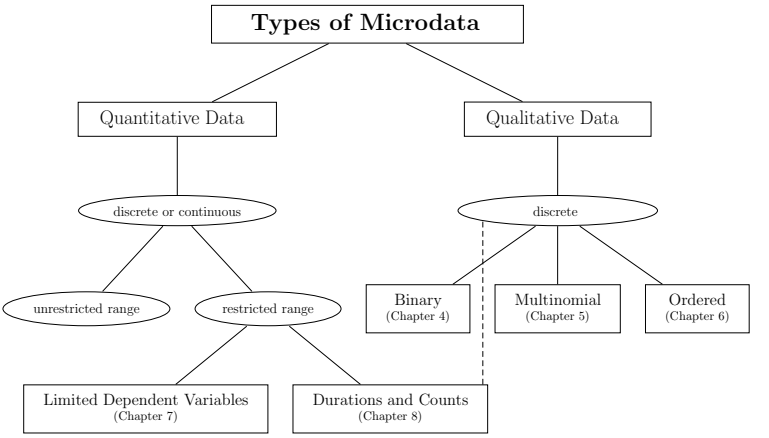
1.2.1 Qualitative Data

In practice, all these measurement types are frequently encountered in applied empirical work. First, consider the following examples of qualitative data.

Binary Variables

A binary variable has two possible outcomes and indicates the presence or absence of a certain property. It answers questions such as: Is a person gainfully employed at the day of the survey (yes/no)? Has a credit application been approved (yes/no)? Has an apprentice been retained in the training firm after completion of apprenticeship (yes/no)? Is a person's willingness-to-pay greater than the asking price (yes/no)?

Fig. 1.1. *Types of Microdata*



Multinomial Variables

A multinomial variable has three or more possible outcomes and indicates the quality of an object using a set of mutually exclusive and exhaustive *non-ordered* categories. Such variables can be used to describe the employment status of a person (full-time / part-time / unemployed / not in labor force), the field of study (humanities / social sciences / engineering) or the portfolio structure of households (stocks only / stocks and bonds / bonds only / none). If there are only two categories, multinomial variables reduce to binary variables.

Ordered Variables

An ordered variable has three or more possible outcomes and indicates the quality of an object using a set of mutually exclusive and exhaustive *ordered* categories, but differences between categories are not defined. Applications include questions like: How satisfied are you with your life (completely satisfied / somewhat satisfied / neutral / somewhat dissatisfied / completely dissatisfied)? How does a credit agency evaluate a lender (AAA / AA+ / ...)? Do you agree with the political program of the ruling party (strongly agree / agree / neutral / disagree / strongly disagree)?

1.2.2 Quantitative Data

The default assumptions for a quantitative dependent variable are that its support is the real line, and that observations form a random sample of the population. The first assumption is compatible with assuming in the linear regression model that the dependent variable is normally distributed, conditional on the regressors, since the normal distribution has support $\mathbb{R}$. The second assumption takes away the possibility of a systematic discrepancy between the population model and what one observes once the sample has been selected. As we will see in this book, both assumptions are frequently violated in microdata applications, and we provide some suggestive examples here.

Non-negative Variables

Wages of workers and prices of houses are non-negative and therefore cannot be normally distributed in a strict sense (although the normal distribution might be a satisfactory approximation). The same holds true for durations between events (such as the duration of unemployment, or time elapsed before an ex-convict is arrested again for a new crime). An additional feature of duration data is their implicit relationship to an underlying stochastic process, which explains why quite specialized methods have been developed for such data. Another example of continuous data with restricted support – not covered in this book – are proportions or share data, where the values necessarily lie between zero and one.

Non-negative Variables with Frequent Zeros

A common data situation is one where a continuous positive variable coexists with a discrete cluster of observations at zero. The prime example, studied by Tobin (1958), are the expenditures for a certain consumer good, measured per household and per period of time (for instance day, month, or year). Such data provide two kinds of information. First, they tell us whether a good was purchased or not, and second, they give us the purchased quantity, provided a positive amount of the item was purchased. From an economic point of view, this distinction corresponds to the difference between a corner and an interior solution to the household utility maximization problem. Thus, Wooldridge (2002) suggests that models for this type of data be referred to as "corner solution models".

Truncated Variables

A variable is truncated if all observations with realizations above or below a certain threshold are excluded from the sample. For example, if colleges only admit students with a certain minimum SAT (Standardized Aptitude Test) score, then the distribution of scores among admitted students is truncated

from below at the threshold level. The consequences of truncation are that the observed data (such as SAT scores among admitted students) are no longer representative for the population at large (the SAT scores among all high school graduates or college applicants), even if the sampling is otherwise random (every student with a passing SAT score has the same chance of being admitted). As we will see, it may nevertheless be possible to infer population parameters from such a sample, as long as we know both the truncation point and the distribution function of test scores in the population, up to some unknown parameters.

Censored Variables

A variable is said to be censored if for parts of the support of the variable, for instance the real line, only the interval rather than the actual value is observed in the data. An example is top-coding of income or wealth. In Germany, for example, social security contributions (for unemployment and health insurance as well as statutory pensions) are proportional to earnings up to a ceiling, beyond which they remain constant. If such social security earnings data report the top income, it means that the person earned at least that income – and possibly much more. A special case of censored data with known censoring points arises if earnings data are grouped, or categorical (such as income from 0 to 500, from 501 to 1000, etc.).

Another example of censoring occurs in duration analysis. Suppose we follow a sample of 15-year-old women and measure the time until first birth. If the study terminates ten years later, then we either have seen a first birth, in which case the duration is known, or we have not, in which case we only know that the time until first birth is greater than ten years. This is a censored observation. In contrast to truncation, censoring does not exclude those observations from the sample. Rather, they are retained, and their proportion is known. The problem of censoring is that the exact value – here for the duration until first birth – is not observed.

A more complex form of censoring arises if the censoring threshold itself is random. For example, wages (and hours of work) are only observed for workers. If workers differ systematically from non-workers, this may be a problem if the objective is to use observed wages to predict potential wages of a randomly selected person or non-worker. The traditional solution to this problem – typically referring to the labor supply of married women – has been to analyze the decision to work in a simple economic model without unemployment, where a woman works only if the wage offer exceeds a certain aspiration (or "reservation") wage (Gronau, 1974). In this case, we observe the wage which equals the wage offer. On the other hand, if a woman is observed not to work, we only know that the wage offers fall short of her reservation wage. Since the reservation wage can vary from person to person, partially depending on factors that are unobserved by the analyst, the threshold is now random.

Count Variables

A count variable answers the question of how often an event occurred, and the possible responses take the form of non-negative integers $\{0, 1, 2, \ldots\}$ (or $\{0, 1, 2, \ldots, n\}$ if there is an explicit upper bound). Examples include the number of patents annually awarded to a firm, the number of casualties from air traffic accidents per year, or the number of shares traded on a given day. An example of a count with an explicit upper bound is the number of days a worker does not report to work during a given week. Count data fill an intermediate position between qualitative and quantitative data. If the number of counts is relatively low, the responses should be treated as categories. As the number of counts increases, the difference between treating the counts as discrete or as continuous becomes increasingly negligible.

These examples cover most of the topics that we will encounter throughout this book. In applications such as these, the linear regression model tends to be inappropriate, and we will need to consider alternative models. Some general remarks about the shortcomings of the linear model are discussed next.

1.3 Why Not Linear Regression?

The workhorse for all applied empirical analyses of relationships between quantitative variables is the linear regression model.

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + u_i \tag{1.1}$$

It is easy to estimate and to interpret, and it provides optimal inference if the standard regularity assumptions are fulfilled, namely linearity in the parameters, uncorrelated errors, mean independence of the error term u_i and the regressors x_{il}, $l = 1, \ldots, k$, non-singular regressors, and homoscedasticity. Under these **Gauss-Markov assumptions**, the ordinary least squares (OLS) estimator is best linear unbiased. The additional assumption of normally distributed error terms has two further implications. First, the OLS estimator is asymptotically efficient among all possible estimators. Second, the small sample distribution of the OLS estimator is known, and exact inference can therefore be based on t- or F-statistics.

For the following arguments, it is useful to rewrite the linear regression model in terms of the **conditional expectation function**, since under the assumption of mean independence, we obtain

$$\mathrm{E}(y_i|x_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} \tag{1.2}$$

Here, $\mathrm{E}(y_i|x_i)$ is shorthand notation for $\mathrm{E}(y_i|x_{i1}, \ldots, x_{ik})$. Henceforth, let $x_i = (1, x_{i1}, \ldots, x_{ik})'$ denote the $(k+1) \times 1$-dimensional column vector of regressors (including a constant), where a' is the transpose of a. Furthermore, if we define

a conformable parameter vector $\beta = (\beta_0, \beta_1, \ldots, \beta_k)'$, again a $(k+1) \times 1$-dimensional column vector, we can express the linear combination on the right hand side of (1.2) conveniently as a scalar product, namely

$$\mathrm{E}(y_i|x_i) = x_i'\beta \tag{1.3}$$

In which sense does the linear model fail if the dependent variable is of any one of the types described in the previous section? We will follow the above order and start with qualitative dependent variables. If the dependent variable is binary, coded as either 0 or 1, the linear regression can be interpreted as a probability model, since $\mathrm{E}(y_i|x_i) = 0 \times P(y_i = 0|x_i) + 1 \times P(y_i = 1|x_i) = P(y_i = 1|x_i)$ and therefore, from (1.2), we get

$$P(y_i = 1|x_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} = x_i'\beta \tag{1.4}$$

One problem in this model are the predictions: clearly, it should be the case that $0 \leq \widehat{P(y = 1|x_0)} \leq 1$. However, the linearity means that this restriction must be violated for certain values x_0 of the regressors. Predictions outside of the admissible range are thus possible. Moreover, the model is heteroscedastic, because the variance of a binary variable conditional on the regressors is $\mathrm{Var}(y_i|x_i) = P(y_i = 1|x_i)[1 - P(y_i = 1|x_i)]$, which is a function of x_i.

If the dependent variable is multinomial, the linear model does not make sense at all since it is meaningless to model (or even compute) the expected value of a multinomial variable. Regression models for multinomial variables should rather directly model the probability distribution function. The same considerations apply to ordered variables. Again, the numerical coding of the outcomes is arbitrary. Any rank preserving recoding should leave the analysis unaffected. Hence, expectations are undefined and cannot be modeled.

In contrast, count data are quantitative and therefore have well-defined expectations. Nevertheless, the linear regression model is inappropriate as well. The problem is threefold. First, the expectation of a count must be non-negative. Again, this is not assured by the functional form (1.2). Second, non-negative variables often have a non-constant variance, so that the homoscedasticity assumption is violated. Admittedly, both of these points could casewise be addressed with standard methods. For example, in the absence of zero counts, one could take logarithms of the dependent variable to enforce a non-negative conditional expectation. Otherwise, non-linear least squares would be an option.

However, these quick fixes fail to address the third problem with counts, as with all other discrete dependent variables, that each outcome has a positive probability and it may be desirable to draw inferences about these distinct probabilities rather than on expectations only. Therefore, the general modeling strategy for discrete data is a shift away from conditional expectation models, such as (1.2), towards the class of conditional probability models.

As far as using the linear regression model for continuous microdata is concerned, one has to distinguish between applications that use limited dependent

variables and those that do not. For example, if the dependent variable is continuous with support over the real line, there is no a priori argument for not using the linear regression model. Indeed, this is the situation for which the linear regression model is best suited. If, however, the support of the dependent variable is limited to the positive real numbers, then the model should take this into account. Otherwise, if inference is based on the conditional expectation (1.2), predictions outside of the admissible range may result. Another related consequence, to be explored in detail later, is that marginal effects in such models should not be constant. This is very much like in the count data case. For example, one can take logarithms and estimate a log-linear model. However, if zeros are important, in particular in corner solution models, other models are required. Again, there are two desirable features. First, predictions should be restricted to the support of the data, and second, probability inferences should be possible regarding the positive mass at zero.

The argument against applying linear regression models in limited dependent variable situations is a different one. Here, the basic idea is that a relationship such as (1.2) holds in the population, and we would like to estimate the population parameters β. However, because of censoring or truncation, it is not advisable to take the observed sample as representative for the population and to estimate the linear regression model directly. Such an estimator will be biased. The reason for the failure of the estimator is that the crucial assumption of mean independence between the error terms and the regressors must fail under sample selection. As an example, consider wages that are truncated from below because low-income individuals are not required to file a tax return. Intuitively, if a regressor x_{il}, such as education, has a positive effect on wages, a low value of this regressor means that the unobserved component of the model must be positive and relatively large in order for the dependent variable to exceed the truncation threshold. On the other hand, a large value of such a regressor means that observations with smaller, or even negative, unobserved components are retained as well. Hence, there is a negative correlation between u_i and x_{il} in the selected sample at hand, and the OLS estimates systematically underestimate the population parameters. Similar considerations apply when the dependent variable is censored.

1.4 Common Elements of Microdata Models

We now have presented more than a handful of departures from the linear regression framework, as they are likely to be encountered by the practitioner dealing with microdata applications. At first sight, these departures do not seem to have much in common. But this appearance is deceiving. In fact, the methods for modeling such data are closely interrelated and based on a common principle, namely **maximum likelihood estimation**. The maximum likelihood principle is quite different from the least squares principle used to fit a regression line to data. Here, the starting point is a parametric distribu-

tion of the endogenous variable (or of the error term). Next, the parameters of the distribution are specified as a function of the exogenous variables, and finally, assuming an independent (cross-sectional) sample, the parameters of the model are estimated by the method of maximum likelihood.

In discrete data applications, the benefit of modeling the probability distribution function directly in terms of regressors and parameters is immense. With the emphasis shifted away from the conditional expectation function towards the **conditional probability function**, a much richer set of inferences becomes available. Essentially, we can analyze the *ceteris paribus* effect of a change in one regressor on the entire distribution of the dependent variable. In limited dependent variable applications, the essential role of the distributional assumption is to tie the population model and the sample model together and to allow inferences on population parameters to be made even if the sample is selective (i.e., non-random).

To summarize, in microdata applications, the data are often qualitative and discrete, while in other cases, the sample is not randomly drawn from the population of interest. Hence, models and methods are needed that go beyond the standard linear regression model and ordinary least squares. As we will see, maximum likelihood is the unifying principle for modeling and estimating microdata relationships. The purpose of this book is to motivate and introduce these models and methods, and to illustrate them in a number of applications. All the models discussed in this book are **parametric**. Nonparametric and semiparametric models induce additional complexity both in terms of estimation and in terms of interpretation. We refer to Pagan and Ullah (1999) and Horowitz (1998) for examples of these methods.

1.5 Examples

The book features three examples, each of which consists of a substantive research question and a dataset for analyzing this question. The examples are referred to repeatedly throughout the different sections of the book. Here, we start with a short introduction and provide some descriptive information on the three datasets. The examples have been chosen such that each highlights a specific methodological issue we consider typical for the analysis of microdata, while they jointly cover much of the spectrum of modeling requirements that can arise in applied empirical work. The examples are: the determinants of fertility, secondary school choice, and female hours of work and wages.

1.5.1 Determinants of Fertility

While individual fertility decisions – the number of children borne by a woman, or the number of children a women would like to have – depend on many factors, including social norms and values, marital status, health and the like,

there has been one factor, namely the women's education, that has been singled out for intensive empirical investigation in the past (Willis, 1974, Sander, 1992). The interest in education is easily understood. If higher education of women leads to fewer children per woman, then we have both an explanation for the fertility decline observed in the developed world during the second half of the last century, and a recipe for reducing high population growth rates in some parts of the developing world.

The empirical analysis of the determinants of fertility in this example is based on data from the US General Social Survey (GSS), an annual or biannual cross-section survey started in 1972. For the purpose of our analysis, we select every fourth year, starting in 1974 and ending in 2002. The survey contains, among other things, information on the number of children ever borne by a woman. If we use the information as it is given, we have a count variable. Alternatively, we can investigate the proportion of childless women, a binary variable. Before we look at some descriptive statistics, we have to think about how to account for the influence of age on the number of children. Clearly, age plays a major role, since young women tend to have fewer children than older ones, even if the eventual number of children – the so-called completed fertility – might be the same. One way to avoid the interfering effect of age is to restrict the analysis to older women: those beyond child-bearing age. A common cut-off age is 40 years. Another possibility is to treat fertility observations for younger women as censored, but this would require more elaborate methods and complicate the descriptive analysis.

Table 1.1 shows the distribution of the fertility variable, where all observations have been pooled over the different years. All in all, the sample includes 5,150 women aged 40 or above, 14.5 percent of whom are childless, and whose average number of children is almost 2.6.

Table 1.1. *Fertility Distribution*

number of children ever	Frequencies	
borne to women (age 40+)	Absolute	Relative
0	744	14.45
1	706	13.71
2	1,368	26.56
3	1,002	19.46
4	593	11.51
5	309	6.00
6	190	3.69
7	89	1.73
8 or more	149	2.89
Total	5,150	100.00

Source: GSS, waves 1974 to 2002 (four-year intervals)

Assume that we want to use these data to answer the following two questions:

1. Is there a downward trend in fertility? In other words, do earlier birth cohorts have a higher fertility than later ones?
2. If there is such a trend, to what extent can it be attributed to (or *explained* by) the rising education levels of women?

Notice here that we are looking for a statistical explanation (a compositional effect): more educated women have fewer children; the proportion of more educated women increases over time; hence, *average* fertility declines. We do not analyze the question *why* more educated women have fewer children (whether it is *because* of their education or for some other reason). However, many studies have investigated this issue and there are indeed good reasons to assume that education has a causal effect on fertility. Economists point out that higher education improves the earnings position of a woman on the labor market, and thus increases the opportunity costs of not working on the market, i.e., of having children and working at home.

With this background, we can now return to the data and ask what type of information should be extracted in order to shed light on the two research questions above. The first sensible step is to investigate whether *average* levels of fertility went down over time, and whether *average* levels of education increased. Given access to the raw data, these quantities should be simple to compute. There is a problem, however. From Table 1.1, we see that the last category is coded as an open-ended "eight or more". This is an instance of "censoring" that will concern us in greater detail later on. For the moment, we ignore the censoring and treat all women in this category as if they had exactly eight children.

Under this assumption, we can conduct the necessary comparisons as in Table 1.2 with year-by-year statistics. The first column gives the number of women above 40 in each of the GSS surveys. The second column gives the average number of children, whereas the third column shows the proportion of childless women. The final column shows the average education level, here measured by the average number of years a woman went to school.

When interpreting such data, we have to keep in mind that they are not the true population values but that they are calculated from a random sample of the population. Therefore, they are subject to sampling error. However, because the observation numbers per year are quite high – they range from a minimum of 410 observations in 1974 to a maximum of 989 observations in 1994 – the confidence intervals for the population parameters are small, as we see from the standard errors in parentheses. Thus, there seems to be clear evidence of a downward trend in fertility. Also, it might be possible that this downward trend can at least partially be explained by the increased levels of formal education among women.

Table 1.2. *Fertility and Average Education Level by Years*

Year	No. of observations	No. of children	Proportion of childless	Years of schooling
1974	410	3.17	0.09	11.07
		(0.10)	(0.01)	(0.16)
1978	445	2.73	0.14	11.00
		(0.09)	(0.02)	(0.15)
1982	577	2.96	0.14	11.05
		(0.09)	(0.01)	(0.14)
1986	470	2.70	0.16	11.34
		(0.09)	(0.02)	(0.14)
1990	431	2.50	0.15	12.41
		(0.08)	(0.02)	(0.15)
1994	989	2.40	0.15	12.78
		(0.06)	(0.01)	(0.10)
1998	911	2.42	0.15	12.94
		(0.06)	(0.01)	(0.11)
2002	917	2.36	0.16	13.25
		(0.06)	(0.01)	(0.10)

Source: GSS, waves 1974 to 2002 (four-year intervals), standard errors in parentheses

Exercise 1.1.

- Can the mean of a discrete variable, such as the number of children, be normally distributed? What does this imply for inference?
- Conduct a formal test of the hypothesis that the average number of children is the same in 1974 and in 2002.
- Is the difference in education levels between 1974 and 2002 statistically significant?

There is a saying that "If the only tool you've got is a hammer, every problem will look as a nail." The only tool we are familiar with at this stage is the linear regression model, so we may as well ask how a regression-based analysis might be used to answer the two research questions. Table 1.3 shows results for three different models. In each case, the dependent variable is the number of children ever borne by a woman. In the first model, the number of children is regressed on year dummies. Since a constant is included, one year has to be chosen as reference, here, the year 1974. The second model includes a linear time trend instead. Here, $t = 0$ for the year 1974, $t = 4$ for the year 1978, and so forth. Finally, the third model includes the linear trend and adds the years of schooling as a further control variable.

Table 1.3. *Linear Regression Analysis of Fertility*

Dependent variable: *Number of children ever borne by a woman*			
	Model 1	Model 2	Model 3
linear time trend		-0.026	-0.014
		(0.003)	(0.003)
years of schooling			-0.128
			(0.008)
year = 1978	-0.436		
	(0.129)		
year = 1982	-0.211		
	(0.122)		
year = 1986	-0.469		
	(0.128)		
year = 1990	-0.674		
	(0.130)		
year = 1994	-0.770		
	(0.111)		
year = 1998	-0.748		
	(0.112)		
year = 2002	-0.807		
	(0.112)		
constant	3.171	3.026	4.392
	(0.093)	(0.056)	(0.103)
R-squared	0.018	0.015	0.060
Observations	5,150		

Notes: Standard errors in parentheses

Exercise 1.2.

- Discuss the regression results. Which one is the preferred model?
- What is the predicted number of children in 1982 according to Models 1 and 2, respectively?
- How can you predict the number of children in 2000?
- Is education related to fertility? Can the trends in education level explain the observed trends in fertility?
- If you were asked to discuss the potential shortfalls of linear regression models in such an application, what would you say?

1.5.2 Secondary School Choice

Our second example relates to the schooling achievement of adolescents in Germany. One peculiar feature of the German schooling system is that students are separated relatively early into different school types, depending on performance and perceived ability. The comprehensive primary school lasts for four years only. After that, around the age of ten, students are placed into one of three types of secondary school, either *Hauptschule* (lower secondary school), *Realschule* (middle secondary school) or *Gymnasium* (upper secondary school). This placement seriously affects a student's future education and labor market prospects, as only Gymnasium provides direct access to the country's universities.

A frequent criticism of this system is that the tracking takes place too early, and that it cements inequalities in education across generations. As the argument goes, the early tracking decision – although formally based on the recommendation of the homeroom teacher, who assesses the child's academic performance – is heavily influenced by the parents. First, more educated parents will better prepare their children for primary school so that after four years of formal schooling, these children may still have an advantage. Second, they may intervene directly and influence the teacher's recommendation, and the teacher has little incentive to oppose such interventions.

The extent to which the mobility (or immobility) in educational attainment between parents and children is high or low can only be decided based on empirical evidence. Our example provides such evidence. The data are based on the German Socio-Economic Panel (GSOEP), a large annual household survey that was first collected in 1984. Specifically, we extracted a sample of 675 14-year old children born between 1980 and 1988. Of them, 29.5 percent attended *Hauptschule*, 29.5 percent *Realschule* and 41.0 percent *Gymnasium*. The following Table 1.4 shows a cross-tabulation of the school the child attended and the education of the parent.

Table 1.4. *Mother's Education and School Track of Child*

Educational	*School track at age 14*			
level of mother	*Hauptschule*	*Realschule*	*Gymnasium*	
7-10 years	55.12	25.20	19.69	100.00
10.5-12 years	28.09	34.16	37.75	100.00
12.5-18 years	3.88	14.56	81.55	100.00

Source: GSOEP, waves 1994 to 2002

Exercise 1.3.

- Describe the nature of the variable "school track".
- Based on the evidence in Table 1.4, is there any evidence for a positive relationship between the educational attainment of mother and child? How would you formally test for the presence of such a relationship?
- What other socio-economic factors might explain the placement of children in the different school tracks?
- If you want to estimate the *ceteris-paribus* effect of the mother's education on the child's school track, can you use a linear regression model? Why, or why not?

1.5.3 Female Hours of Work and Wages

The first two examples on fertility and schooling involved discrete and qualitative dependent variables. In our third and final example, we encounter two types of limited dependent variables, namely a corner solution application and a censored variable with random censoring threshold. We do not claim special credit for this example – in fact, the labor supply of women must be, together with the returns to schooling, one of the most intensively studied topics in microeconometrics. One reason for the popularity of the topic is certainly that the data required for such an analysis can be obtained from any standard labor force survey, which have been available for many years and for most countries. Another reason is that there is a wide variation in the labor force participation of women over time and across countries. Understanding the causes of this variation, and in particular the contribution of tax-, family-, and labor market policies, is of substantive interest.

We base the analysis on the publicly available dataset by Mroz (1987). Previous analyses of these data can also be found in the textbooks by Berndt (1990) and Wooldridge (2002). The dataset comprises a sample of 753 married women, 428 of whom had worked in the year prior to the interview (in 1975) and the remaining 325 of whom had not. Among the women who had worked, the total number of hours ranged from 12 to 4,950, with an average of 1,303 hours (or 27 hours per week, assuming a year has 48 working weeks). For working women, the data also contain information on the hourly wage, which is obtained by dividing annual earnings by annual hours of work. The average hourly wage amounts to USD 4.20. The data include further information on a number of variables that can be expected to affect hours and wages. Among these are the age and education level of the woman, her previous labor market

experience (measured in years of participation), her husband's income, and the presence of young and adolescent children in the household.

Suppose that we want to use these data in order to answer the following research question: What is the wage elasticity of female labor supply – by what percent will the hours of work change if the wage is increased by one percent? A simple linear regression of hours of work on wages and some other factors produces the following result for the Mroz data.

$$\begin{array}{lllll} \widehat{hours} = 1,665.6 - 22.7\,wage - 4.9\,nwifeinc - 300.6\,kidslt6 - 99.0\,kidsge6 \\ \quad\quad (92.2) \quad\quad (11.2) \quad\quad (3.5) \quad\quad (93.4) \quad\quad (27.9) \end{array}$$

$$n = 428, \;\; R^2 = 0.07$$

On the face of it, the estimated labor supply curve has a negative slope, and the elasticity, evaluated at the mean wage and mean hours, is $-22.7 \times 0.042/1303 = -0.07$ percent and thus very small.

But such an analysis has a number of problems. Most importantly, we do not know the wages of women who do not work. Hence, we can only estimate the above model with the subsample of 428 employed women. By doing so, we ignore that a wage increase may also have an effect at the extensive margin of labor supply: some women who did not work previously might be drawn into the labor market as their wage offer (or potential wage) increases. If we want to estimate the model using all observations, we need to predict the wage for women who do not work. What should this prediction be based on? We can model the wages as a function of other factors, such as education and experience. However, estimating the parameters of this regression using working women only without further adjustment is generally not a good idea, because women have self-selected into employment – partially based on their wages – and their wages therefore are not necessarily representative of all women. Once we have predicted wages for non-working women, based on a method that corrects for such self-selection, we can estimate a structural hours of work model ("structural" here means that the wages are included as a regressor – as opposed to a reduced-form model where wages are excluded but the wage determinants, such as education and experience, are included instead). But again, linear regression is inappropriate since we are now dealing with a corner solution outcome: many women report zero hours of work, and the estimation method should account for this discrete cluster of observations at zero.

Exercise 1.4.

- The minimum reported hourly wage is 12 cents. Is this a reasonable number? What should one do about it?
- Draw a simple labor supply diagram, with consumption on the y-axis and hours of leisure on the x-axis. What does the budget constraint look like? How can the effect of the husband's income and of children at home be represented in this diagram?
- Assume you want to model participation only. What type of dependent variable is this?
- What is the labor participation rate of women in your country? How can you find out?

1.6 Overview of the Book

The book is composed of seven chapters in addition to this introduction. In the next chapter, we will further motivate the probability-based approach that underlies all models for qualitative dependent variables. Accordingly, the concept of a **conditional expectation function** central to all regression analysis is replaced by the concept of a **conditional probability function**. The interpretation of such models, then, naturally can be based on what we refer to as **marginal probability effects**. The chapter provides some illustrations of these concepts, and it also reviews some basic results from mathematical statistics and probability theory that are required in the further analysis.

Chapter 3 introduces the theory of maximum likelihood estimation. We believe that a correct application and interpretation of likelihood-based models requires a good grasp of the underlying method, although not necessarily the ability to prove all the results. The chapter therefore tries to follow an intermediate approach, covering the essential aspects of estimation and inference. In Chapter 4, the binary response model is introduced. We present the basic probit and logit models, and discuss the estimation and interpretation of the parameters. We also consider non-standard sampling schemes. Binary response models only work for two response categories, so Chapter 5 introduces the multinomial extensions to more than two unordered categories. If more than two categories are ordered, this information should be taken into account, and the ordered response models discussed in Chapter 6 show how to do so.

Chapter 7 deals with models for limited dependent variables. After reviewing general results for the truncated normal distribution, we start with corner-solution models for mixed discrete-continuous data. The focus then shifts to censored regression models, first with known thresholds and then with random thresholds. Finally, Chapter 8 combines the discussion of duration models and count data models under the theme of "event history analysis", emphasizing the common aspects of the two types of data: whereas count data measure the number of events during a given period of time, duration data measure the time between them.

2

From Regression to Probability Models

2.1 Introduction

In this chapter, we introduce the general principles underlying the modeling of microdata, in particular qualitative response variables. Relative to the linear regression framework, the key element is a change in paradigm from modeling the **conditional expectation function** towards modeling the **conditional probability function**. There are two main reasons for this shift in focus. First, in many cases the expected value of a qualitative variable is simply not defined (for ordered and multinomial responses). And second, even where the choice exists (such as for count data that may be treated as qualitative or quantitative), the probability-based approach provides additional information: once the probabilities are known, the expected value is fully determined. The opposite does not hold. We begin with an example.

Example 2.1. Fertility

Consider the data from the U.S. General Social Survey on the number of children among women aged 40 or above. In Table 1.2 of the previous chapter, we displayed the average number of children by survey year. Each mean can be interpreted as an estimator for the true average in that year, and thus for the expectation conditional on the survey year, denoted by $\mathrm{E}(y_i|year_i)$. For example, the average number of children declined from 2.70 to 2.36 between 1986 and 2002, and this decline is statistically significant. We cannot tell from this mean comparison, however, what changes in the fertility distribution were responsible for the average decline. For example, the decline could result either from an increase in the number of childless women, or from a decline in the proportion of very large numbers of children. Depending on the practical question one wants to answer, this might make a difference.

In order to answer such questions, we can look at conditional relative frequency distributions instead. Table 2.1 shows these distributions for the years 1986 and 2002. They can be interpreted as estimators of the conditional probability distributions $f(y_i|year_i = 1986)$ and $f(y_i|year_i = 2002)$. The additional information allows for a more detailed analysis. We observe that the proportion of childless women has not changed much – if anything, it has declined – whereas the proportion of women with five or more children has decreased from 18.5 to 10 percent.

Table 2.1. *Conditional Relative Frequency Distributions*

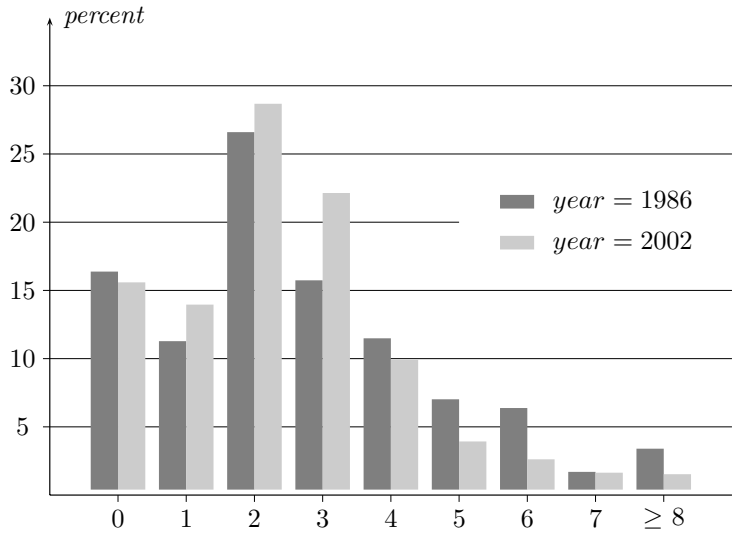
number of children ever borne to women (age 40+)	1986	2002	Total
0	16.38	15.59	15.86
1	11.28	13.96	13.05
2	26.60	28.68	27.97
3	15.74	22.14	19.97
4	11.49	9.92	10.45
5	7.02	3.93	4.97
6	6.38	2.62	3.89
7	1.70	1.64	1.66
8 or more	3.40	1.53	2.16
Total	100.00	100.00	100.00

Source: General Social Survey.

The change can also be illustrated graphically, as in Figure 2.1. A Pearson chi-squared test can be used to formally test the hypothesis of independence between the distribution of the number of children and the survey year.

Of course, if we want to include further dimensions in the problem – for example, account for the influence of other factors like the level of education or the family background – a simple analysis by cross-tabulation or graph is no longer possible. This holds true whether we consider the outcome "number of children", which is a count variable, or the outcome "childless" (yes/no), which is a binary variable. In this book, we introduce general methods for modeling conditional probabilities in such cases.

Fig. 2.1. *Relative Frequencies of Number of Children Ever Born*



2.2 Conditional Probability Functions

2.2.1 Definition

The goal of most microdata applications is to investigate the relationship between a dependent variable y_i and a vector of explanatory variables x_i. The concept of a **conditional expectation function** (CEF) is the key element of linear regression analysis as presented in any introductory econometrics course. The basic idea is to model and estimate the expectation of y_i conditional on x_i. Let

$$\mathrm{E}(y_i|x_i) = \mu(x_i;\beta) \tag{2.1}$$

denote the CEF of y_i given x_i. The conditional expectation depends on a set of unknown parameters β to be estimated from the data. The linear regression model is a special case obtained for

$$\mu(x_i;\beta) = x_i'\beta$$

The **conditional probability function** (CPF) is defined analogously as

$$P(y_i|x_i) = f(y_i|x_i;\theta) \tag{2.2}$$

In many applications, the parameter vector θ can be partitioned into one set of parameters β, which is part of the regression component – often entering via

a **linear index function** $x_i'\beta$ – and another set of **auxiliary parameters** that influence other aspects of the conditional probability model and do not interact directly with x_i.

There is one important special case in which CEF and CPF coincide. If the dependent variable is binary, $y_i \in \{0, 1\}$, then the conditional expectation is $\mathrm{E}(y_i|x_i) = 0 \times P(y_i = 0|x_i) + 1 \times P(y_i = 1|x_i) = P(y_i = 1|x_i)$.

2.2.2 Estimation

The shift in focus from the CEF to the CPF requires new techniques for estimating the parameters of the model. In CEF applications, the least squares criterion states that the (weighted or unweighted) sum of squared residuals should be minimized. A residual is defined as the difference between the actual (observed) outcome y_i and its predicted value $\hat{y}_i$, $\hat{u}_i = y_i - \hat{y}_i$, where $\hat{y}_i$ is an estimator of $\mathrm{E}(y_i|x_i)$. The estimator obtained by the least squares criterion is the value $\hat{\beta}_{LS}$ that minimizes $\sum_{i=1}^n \hat{u}_i^2$. If the CEF is linear, $\mathrm{E}(y_i|x_i) = x_i'\beta$, then we can write $\hat{y}_i = x_i'\hat{\beta}$ and $\hat{u}_i = y_i - x_i'\hat{\beta}$.

In CPF applications, the conditional expectation is often not defined. Thus, there is no regression error and residual-based methods do not work. Therefore, in such applications, the least squares criterion must be replaced by a different one. As we will see, a very general method is maximum likelihood estimation. It consists of deriving a **likelihood function** for the problem at hand, and maximizing the likelihood function with respect to the unknown parameters, which provides the **maximum likelihood estimator**.

What does a likelihood function look like? Our starting point is the CPF $f(y_i|x_i;\theta)$ for observation i. Assume that a random sample of n pairs of observations (y_i, x_i), $i = 1, \ldots, n$, is available. Define $y = (y_1, \ldots, y_n)'$ and $x = (x_1, \ldots, x_n)'$. Under **random sampling**, the joint probability function of the observed sample can be written simply as the product over all individual probabilities

$$f(y|x;\theta) = \prod_{i=1}^{n} f(y_i|x_i;\theta) \tag{2.3}$$

Seen as a function of the unknown parameter vector θ, this is a likelihood function. More generally, any function proportional to (2.3) is an equally well defined likelihood function, and therefore we can write

$$L(\theta; y, x) = c \prod_{i=1}^{n} f(y_i|x_i;\theta) \tag{2.4}$$

where $c > 0$ is a proportionality constant. The maximum likelihood estimator $\hat{\theta}_{ML}$ is defined as the value of θ that maximizes (2.4). Moreover, under some regularity assumptions, the maximum likelihood estimator exists, is unique, consistent and asymptotically efficient. In Chapter 3, we discuss in detail the concept of maximum likelihood estimation.

2.2.3 Interpretation

In a CEF model with $\mathrm{E}(y_i|x_i) = \mu(x_i;\beta)$, we are interested in how the conditional expectation changes, as a function of the parameters, if an explanatory variable increases. Similarly, in a CPF model with $P(y_i|x_i) = f(y_i|x_i;\theta)$ we want to know how the conditional probability changes, as a function of the parameters, if an explanatory variable increases.

Formally, in CEF models, the **marginal mean effect** (MME) of the l-th regressor is defined as $\partial \mathrm{E}(y_i|x_i)/\partial x_{il}$. In the linear regression model, the conditional expectation function is given by $x_i'\beta = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$, and the MME simplifies to

$$\frac{\partial \mathrm{E}(y_i|x_i)}{\partial x_{il}} = \beta_l \tag{2.5}$$

If $\mathrm{E}(y_i|x_i)$ is non-linear in β, the MME is not constant but rather depends on the values of x_i. In this case, it is usual in applied research to report the MME for a "typical" individual, for example by evaluating the MME at the sample mean of the regressors, or at the mode for a binary regressor. Alternatively, one can define the average marginal mean effect (AMME) as $\mathrm{E}_x[\partial \mathrm{E}(y_i|x_i)/\partial x_{il}]$, which can be estimated consistently by averaging over all MME's in the sample. Some simplifications arise if the model is of the generalized linear form $\mathrm{E}(y_i|x_i) = \mu(x_i'\beta)$ where $\mu(\cdot)$ is any continuous and differentiable function of the linear index $x_i'\beta$. In this case, the (absolute) marginal mean effect can be written as

$$\frac{\partial \mathrm{E}(y_i|x_i)}{\partial x_{il}} = \mu'(x_i'\beta)\beta_l \tag{2.6}$$

where $\mu'(\cdot)$ is shorthand notation for the first derivative of $\mu(\cdot)$ with respect to its argument, $d\mu(z)/dz$. The relative MME is defined as the ratio $[\partial \mathrm{E}(y_i|x_i)/\partial x_{il}]/\mathrm{E}(y_i|x_i)$.

In a generalized linear index model, the **discrete mean effect** of a unit change in one regressor (like the switch of a dummy variable from zero to one) can be computed as

$$\Delta \mathrm{E}(y_i|x_i) = \mu(x_i'\beta + \beta_l) - \mu(x_i'\beta) \qquad \text{(absolute change)} \tag{2.7}$$

$$\frac{\Delta \mathrm{E}(y_i|x_i)}{\mathrm{E}(y_i|x_i)} = \frac{\mu(x_i'\beta + \beta_l)}{\mu(x_i'\beta)} - 1 \qquad \text{(relative change)} \tag{2.8}$$

Exercise 2.1.

Determine (i) the range of $\mathrm{E}(y_i|x_i)$, (ii) the absolute and relative marginal mean effects for x_{il}, and (iii) the absolute and relative change in $\mathrm{E}(y_i|x_i)$ for a unit change in x_{il} for the following CEF's

- $\mathrm{E}(y_i|x_i) = \exp(x_i'\beta)$ for all $x_i \in I\!R$
- $\mathrm{E}(y_i|x_i) = \exp(x_i'\beta)/[1 + \exp(x_i'\beta)]$ for all $x_i \in I\!R$

In many cases, the linear index form applies to transformations of the original variables. The relevant marginal and discrete mean effects have to be calculated then with respect to the untransformed variables, which leads to modified expressions. For example, consider the conditional expectation function $\mathrm{E}(y_i|x_i) = \mu(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2)$. The MME of x_{i1} is then

$$\frac{\partial \mathrm{E}(y_i|x_i)}{\partial x_{i1}} = \mu'(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2) \times (\beta_1 + 2\beta_2 x_{i1})$$

By the same reasoning, if $\mathrm{E}(y_i|x_i) = \mu(\beta_0 + \beta_1 \log(x_{i1})\,)$, then

$$\frac{\partial \mathrm{E}(y_i|x_i)}{\partial x_{i1}} = \mu'(\beta_0 + \beta_1 \log(x_{i1})\,) \times \frac{\beta_1}{x_{i1}}$$

where $\log(z)$ denotes the natural logarithm of z, with first derivative given by $d\log(z)/dz = 1/z$.

An important observation is that the non-linear CEF $\mu(x_i'\beta)$ implies interactive effects, even in the absence of an explicit interaction term $x_{il}x_{im}$, since

$$\frac{\partial^2 \mathrm{E}(y_i|x_i)}{\partial x_{il}\partial x_{im}} = \mu''(x_i'\beta)\beta_l\beta_m \tag{2.9}$$

where $\mu''(\cdot)$ is shorthand notation for the second derivative of $\mu(\cdot)$ with respect to its argument, $d^2\mu(z)/(dz)^2$. This is unlike in the linear regression model with expectation function $\mathrm{E}(y_i|x_i) = x_i'\beta$, where cross-derivatives are zero unless the linear index contains an interaction term.

Exercise 2.2.

Suppose you have $\mathrm{E}(y_i|x_i) = \mu(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1}x_{i2})$.

- Derive the absolute and relative MME for x_{i1}.
- Derive the cross-derivative of the CEF with respect to x_{i1} and x_{i2}.

In a conditional probability model, the main quantities of interest are the **marginal probability effects** (MPE's). The MPE of the l-th exogenous variable is defined as partial derivative $\partial P(y_i|x_i)/\partial x_{il}$. For any given probability model, there are always as many MPE's as there are outcomes of the dependent variable. For example, if we consider a binary response variable, then there are two such MPE's, namely the marginal change of the probability of a zero, and the marginal change of the probability of a one, as one of the regressors changes and the others are kept constant. In general, for a qualitative response variable with J response categories, J distinct MPE's can be computed. Due to the adding-up constraints of a proper probability model, the J MPE's must add up to zero, so that only $J-1$ MPE's are linearly independent.

Again, a relatively simple expression for the MPE's is obtained if the conditional probability model is of the linear index form (as in most of the models in this book). In this case, we can write $P(y_i|x_i) = f(y_i|x_i'\beta)$, and therefore

$$\frac{\partial P(y_i|x_i)}{\partial x_{il}} = f'(y_i|x_i'\beta)\beta_l \tag{2.10}$$

where $f'(\cdot)$ is shorthand notation for the first derivative of $f(\cdot)$ with respect to the linear index, $\partial f(y_i|x_i'\beta)/\partial(x_i'\beta)$. Note that $f(y_i|x_i'\beta)$ may or may not depend on further auxiliary parameters. As for the MME's, we may evaluate the MPE's at the average of the regressors, or average over the MPE's to obtain effects that are unconditional on x_i.

The **discrete effect** on the probabilities associated with a unit increase in x_{il}, again under the linear index assumption, can be expressed as

$$\Delta P(y_i|x_i) = f(y_i|x_i'\beta + \beta_l) - f(y_i|x_i'\beta) \tag{2.11}$$

In order to obtain relative effects, the two last formulas are simply divided by $f(y_i|x_i'\beta)$. We will study these marginal and discrete probability effects in detail later on, when we introduce the specific models. However, a general comment can be made already based on (2.10) and (2.11): there is no reason to assume that $f(y_i|x_i'\beta)$ must be an increasing function in the linear index. Therefore, it is possible that both marginal and discrete probability effects have the opposite sign to β_l.

Exercise 2.3.

Suppose you have a binary dependent variable, $y_i \in \{0, 1\}$, with conditional probability function given by

$$f(y_i|x_i'\beta) = \left(\frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)}\right)^{y_i} \left(\frac{1}{1+\exp(x_i'\beta)}\right)^{1-y_i}$$

- Derive $P(y_i = 0|x_i)$ and $P(y_i = 1|x_i)$.
- Determine the absolute (relative) marginal probability effects for x_{il}.
- How does your answer change if you replace x_{il} by $\log x_{il}$.

Marginal probability effects can be used to approximate the discrete change in probabilities using the concept of differentials

$$\Delta P(y_i|x_i) \approx \frac{\partial P(y_i|x_i)}{\partial x_{il}} \Delta x_{il}$$

and the smaller the absolute change in x_{il}, the better the approximation. This concept can easily be extended to changes in two or more explanatory variables. For example, in the case of two explanatory variables x_{il} and x_{im}, we obtain by *totally differentiating*

$$\Delta P(y_i|x_i) \approx \frac{\partial P(y_i|x_i)}{\partial x_{il}} \Delta x_{il} + \frac{\partial P(y_i|x_i)}{\partial x_{im}} \Delta x_{im} \tag{2.12}$$

and we could ask how much both variables have to change such that the probabilities remain the same, $\Delta P(y_i|x_i) = 0$. Rearranging terms yields

$$\frac{\Delta x_{il}}{\Delta x_{im}} = -\frac{\partial P(y_i|x_i)/\partial x_{im}}{\partial P(y_i|x_i)/\partial x_{il}} \tag{2.13}$$

This ratio can be interpreted as **"iso-probability"** curve since we ask how much variation is needed in one regressor, Δx_{il}, to compensate for a given variation in another regressor, say $\Delta x_{im} = 1$, such that $\Delta P(y_i|x_i) = 0$.

2.3 Probability and Probability Distributions

Because all of the models discussed in this book require a good knowledge of the properties of random variables and probability distribution functions, we provide a brief review of some central results. For more comprehensive presentations we recommend introductions to mathematical statistics by DeGroot (1986) and by Hogg and Craig (1989). Following the conventions in the literature about probability and probability distributions, we use uppercase letters in this section to denote the names of random variables (e.g., Y or X), and lowercase letters to denote particular outcomes of the random variable (e.g., y or x). Moreover, for notational simplicity, we drop the subscript i for the rest of this chapter.

2.3.1 Axioms of Probability

To begin with, recall some elementary concepts from probability theory. Let Ω denote the sample space, i.e., the set of all possible elementary outcomes ω_i of an experiment, and call any subset of Ω an event A. In microdata applications, the fundamental experiment is a draw from the underlying population in order to measure the outcome of a variable. For example, the random draw could be from the set of all young men, in order to determine whether the selected male attends university or not. In this case, the sample space given by $\Omega = \{$not attending university, attending university$\}$. In this simple example, there are only four events (subsets of Ω): the empty set, Ω itself, "the person attends university", and "the person does not attend university". A priori, before the data are drawn, we do not know the outcome of the experiment, and hence, there is uncertainty about which outcome and which event will be observed. Uncertainty is modeled using the concept of probability. Formally, we call any real valued set function $P(A)$ on the sample space Ω a **probability measure**, if it fulfills the following **axioms of probability**

1. $0 \leq P(A) \leq 1$ for every event A.
2. $P(\Omega) = 1$.
3. If $A_1, A_2, \ldots, A_J$ is a finite or infinite sequence of *disjoint* events, then

$$P\left(\bigcup_{j=1}^{J} A_j\right) = \sum_{j=1}^{J} P(A_j)$$

In practice, we always work with **random variables**. A random variable Y is a function whose domain of definition is the set of elementary events ω_i and whose range is the set of real numbers. In the above example, we can let $Y = 1$ if $\omega =$"the person attends university". A **probability model** consists then of a sample space, a probability measure, and a random variable. In practice, we start immediately from a random variable, i.e., a set of numerical

outcomes together with their probabilities. We usually assume that a probability function depends on a set of parameters, denoted by θ. For example, the probability of attending university (the data), given that the university selects its students randomly on a 80/20 basis (the parameters). So far, the model does not include any explanatory variables. In Section 2.3.4 we include regressors in the probability model by extending it to a conditional probability model, as the latter plays a crucial role in implementing microdata models in the framework of maximum likelihood (see also the definition of the likelihood function in equation (2.4)). We first review some properties of univariate and multivariate random variables.

2.3.2 Univariate Random Variables

Consider univariate random variables first. A random variable Y is said to be **discrete** if it takes a finite number of values or is countably infinite. It is convenient to index the outcomes in increasing order and write the range of Y as $\{y_1, \ldots, y_J\}$ where $y_j \leq y_{j+1}$. Examples are the set $\{0, 1\}$ in the case of a binary variable with $J = 2$ distinct outcomes, or the set $\{0, 1, 2, 3, \ldots\}$ in the case of a count variable with a countably infinite number of values, $J = \infty$. The **probability function**

$$P(Y = y) = f(y) \qquad y \in \{y_1, \ldots, y_J\} \tag{2.14}$$

has the following properties

1. $0 \leq f(y) \leq 1$.
2. $f(y_1) + \ldots + f(y_J) = 1$

Consistent with the probability axioms, the first property ensures that probabilities are within the unit interval, the second property states that the probabilities of the J mutually exclusive outcomes add up to unity.

Exercise 2.4.

- The Poisson probability function is given by

$$f(y; \lambda) = \frac{\exp(-\lambda)\lambda^y}{y!} \qquad y = 0, 1, 2, 3, \ldots$$

Verify that $0 \leq f(y; \lambda) \leq 1$ and that $\sum_{y=0}^{\infty} f(y; \lambda) = 1$.

Moreover, the **cumulative distribution function** is defined – whenever appropriate – as

$$P(Y \leq y) = \sum_{y_j \leq y} f(y_j) = F(y)$$

from which it follows that $f(y_j) = F(y_j) - F(y_{j-1})$. For multinomial outcomes, the cumulative distribution function has no meaningful interpretation.

A **continuous** random variable is characterized by a set of infinitely divisible outcomes. Hence, the probability associated with any particular outcome must be zero, but we can calculate the probability that a particular outcome falls in a certain interval. This probability is defined as the corresponding area under the **density function** $f(y) \geq 0$

$$P(a \leq Y \leq b) = \int_a^b f(y)dy \tag{2.15}$$

From the probability axioms it must hold that the area under the density function over the complete support must equal one, $\int_{-\infty}^{\infty} f(y)dy = 1$. Moreover, the **cumulative density function** is defined as $F(y) = \int_{-\infty}^{y} f(t)dt = P(Y \leq y)$ from which we obtain an alternative expression of the interval probabilities, namely $P(a \leq Y \leq b) = F(b) - F(a)$. Another useful property of the cumulative distribution function is that $P(Y > y) = 1 - F(y)$.

Exercise 2.5.

- The density function of the exponential distribution is given by

$$f(y;\lambda) = \begin{cases} \lambda e^{-\lambda y} & \text{for } y \geq 0 \\ 0 & \text{else} \end{cases}$$

  where $\lambda > 0$. Verify that $\int_{-\infty}^{\infty} f(y;\lambda)\, dy = 1$.

2.3.3 Multivariate Random Variables

The concept of, and results for, multivariate random variables are essential in the analysis of microdata because on the one hand, we deal here with samples of observations, and on the other hand, we want to model probabilistic relationships between variables. The former forces one to think about how the sample is drawn from the population; the latter requires assumptions about how the independent or explanatory variables affect the dependent variable.

As a starting point, take the joint distribution of two **discrete** random variables denoted by X and Y with possible outcomes $x \in \{x_1, \ldots, x_K\}$ and

$y \in \{y_1, \ldots, y_J\}$, respectively. We define the **joint** (or in this case "bivariate") **probability function** as

$$P(X = x, Y = y) = f(x, y) \qquad y \in \{y_1, \ldots, y_J\},\ x \in \{x_1, \ldots, x_K\} \quad (2.16)$$

which is the joint probability that X takes the value x, and that Y takes the value y. As before, the probability axioms require that $0 \leq f(x, y) \leq 1$ and $\sum_x \sum_y f(x, y) = 1$. We obtain the cumulative distribution function by taking sums over all probabilities with values $x_k \leq x$ and $y_j \leq y$

$$P(X \leq x, Y \leq y) = \sum_{x_k \leq x} \sum_{y_j \leq y} f(x_k, y_j) = F(x, y)$$

and the requirement of at least ordinal scale extends to both random variables, X and Y. Of course, all these considerations can be generalized to more than two random variables. In this case it is more convenient to use matrix notation. However, since all ideas can be illustrated by using the simple bivariate case we concentrate here on the latter.

Based on any joint probability function for discrete random variables, one can obtain the univariate marginal and conditional probability functions. The marginal probability functions of X and Y are given by

$$f(x) = \sum_y f(x, y) \qquad \text{and} \qquad f(y) = \sum_x f(x, y) \quad (2.17)$$

whereas the conditional distributions are given by

$$f(x|y) = \frac{f(x, y)}{f(y)} \qquad \text{and} \qquad f(y|x) = \frac{f(x, y)}{f(x)} \quad (2.18)$$

With the definitions of conditional and marginal distributions, we can define **statistical independence** as follows. Two random variables X and Y are independent if, and only if $f(x, y) = f(x)f(y)$ or, equivalently, X and Y are independent if $f(x|y) = f(x)$ or $f(y|x) = f(y)$. If X and Y are not statistically independent, then they are called **statistically dependent**. Statistical independence has two important implications. First, if we know the outcome of X, this additional information does not change the probability of observing a particular value of Y. Second, if X and Y are independent, then any two functions $h(X)$ and $g(Y)$ are independent as well. As mentioned already, the basic sampling model in microdata applications is the random sampling approach, which implies that observations are independent.

Exercise 2.6.

Consider the following joint probability function

	$Y=0$	$Y=1$
$X=0$	2/6	1/4
$X=1$	1/4	1/6

- Derive the conditional expectation function $\mathrm{E}(y|x) = \sum_y yf(y|x)$.
- Are X and Y independent?

The definitions of joint, conditional and marginal density functions in the **continuous** case are largely analogous. Let $f(x, y)$ be the bivariate density function of two continuous random variables X and Y. Then

$$P(a_1 \leq X \leq a_2, b_1 \leq Y \leq b_2) = \int_{a_1}^{a_2} \int_{b_1}^{b_2} f(x, y)dxdy \tag{2.19}$$

is the joint probability that X falls into the interval (a_1, a_2) and Y falls into the interval (b_1, b_2). The cumulative density $F(x, y)$ can be obtained similarly to the univariate case by taking integrals, now in the X *and* Y dimension. Moreover, the marginal distributions are given by

$$f(x) = \int_{-\infty}^{\infty} f(x, y)dy \qquad \text{and} \qquad f(y) = \int_{-\infty}^{\infty} f(x, y)dx \tag{2.20}$$

and the conditional distributions are defined in the same way as in (2.18).

Exercise 2.7.

- Suppose that the joint density function is of the form

  $$f(x, y) = x + y \qquad 0 \leq x \leq 1 \quad 0 \leq y \leq 1$$

  This distribution is sometimes referred to as a "roof" distribution. Derive the conditional density $f(y|x)$ and the conditional expectation function

  $$\mathrm{E}(y|x) = \int_{-\infty}^{\infty} yf(y|x)$$

Exercises 2.6 and 2.7 are somewhat unrealistic for two reasons. First, they do not include any parameters, although parameters are essential for econometric modeling. Second, they suggest that microeconometric model-building starts with a joint probability or density function and derives the corresponding conditional function. But this is not the usual procedure. The reason is that it is both difficult and unnecessary to specify a full joint model. It is difficult, because in more realistic set-ups there are many explanatory variables and parameters, and except for special cases (such as if X and Y are multivariate normally distributed) the derivations may become cumbersome and the integrals required may not even have a closed-form solution. It is unnecessary because one can start in many cases directly with a **conditional model** and leave the joint distribution completely unspecified. For example, consider the following factorization of a joint probability function into a conditional and a marginal probability function

$$f(x, y; \theta, \gamma) = f(y|x; \theta) f(x; \gamma)$$

where θ and γ are two types of parameters. Importantly, θ appears only in the conditional probability function of $y|x$ and γ appears only in the marginal distribution of x. This means that all the information about θ is contained in $f(y|x; \theta)$ and $f(x; \beta)$ is a constant as far as θ is concerned. Therefore, it can be ignored in the modeling process of the relationship between Y and X. This simplification does not always work, for example if X is endogenous, but we start with the assumption of exogenous regressors X.

2.3.4 Conditional Probability Models

After these preliminaries, we are now in a position to formally define what we mean by conditional probability modeling of qualitative microdata. The essential approach is to leave the joint distribution of Y and X unspecified. Instead, we obtain a conditional model directly, as follows.

1. Select a simple univariate probability function $P(Y = y) = f(y; \theta)$. The main requirement for the choice of $f(y; \theta)$ is that it should have the same support as the observed variable.
2. Express the parameter θ as a function of X to obtain a conditional probability model $f(y|x; \theta)$.

This can be interpreted as follows. We want to model the data-generating process that has created the observations of the dependent variable. This requires us to identify and select an appropriate probability function. Finally, we make the parameters of the selected probability function dependent on the explanatory variables. By doing so, we obtain a conditional probability model. This approach is best illustrated with a few examples.

Example 2.2. Binary Responses

Suppose that Y is a binary variable. Obviously, the underlying distribution function is a Bernoulli distribution, which is fully determined by the success probability π. In this case, the counter-probability is $1-\pi$, and the probability function can be written compactly as

$$f(y;\pi) = \pi^y(1-\pi)^{1-y} \qquad y = 0, 1$$

In a conditional probability model, π is specified as a function of X. For example, with a single explanatory variable, we could let

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \tag{2.21}$$

in order to make sure that $0 \leq \pi(x) \leq 1$ for arbitrary values of $x \in \mathbb{R}$, and without any restrictions on the parameter space of β_0 and β_1. In the general case of k independent variables, we could instead write $\pi(x) = \exp(x'\beta)/[1 + \exp(x'\beta)]$. In Chapter 4, we provide alternative specifications of π in terms of the explanatory variables. But this simple example already illustrates the basic idea. Select the Bernoulli distribution as describing appropriately the nature of the dependent variable, and specify the parameter π as a function of X, adequately taking into account its range.

Exercise 2.8.

Let $Y|X$ be Bernoulli distributed with success probability $\pi(x)$ specified as in (2.21) and let $\beta_0 = 0.5$ and $\beta_1 = 0.8$.

- Draw a graph of the function $\pi(x)$.
- What happens to π as x goes to plus/minus infinity?
- Determine $P(Y = 1|X = 0)$, $P(Y = 0|X = 0)$ and $P(Y = 1|X = 1)$.

Example 2.3. Multinomial Responses

If Y is multinomial distributed with J unordered mutually exclusive outcomes, we could select the probability function of the multinomial distribution. This probability function has J parameters: the outcome probabilities π_j for each outcome. The probability function can be written compactly as

$$f(y; \pi_1, \ldots, \pi_J) = \pi_1^{d_1} \pi_2^{d_2} \cdots \pi_J^{d_J}$$

where $d_j = 1$ if $y = j$ and $d_j = 0$ otherwise. Note that the axioms of probability require $\sum_j p_j = 1$. Thus, one parameter is defined by the others and we can write

$$\pi_1 = 1 - \sum_{j=2}^{J} \pi_j$$

In a conditional probability model, the π's are specified as functions of X in compliance with the adding-up restriction. In Chapter 5, we will discuss several specifications of the parameters π_j, and we will learn which specifications have what implications for the properties of the conditional probability model.

Exercise 2.9.

Let $Y|X$ be multinomial distributed with $J = 3$ unordered and mutually exclusive outcomes.

- Suggest a specification, i.e., a functional form, of $\pi_j(x)$ such that $0 \leq \pi_j(x) \leq 1$ for all $j = 1, \ldots, 3$ and for all $x \in \mathbb{R}$, and such that $\sum_{j=1}^{3} \pi_j = 1$.

Example 2.4. Count Responses

Suppose that Y is a count variable distributed according to the Poisson probability function. Then we can write

$$f(y;\lambda) = \frac{\exp(-\lambda)\lambda^y}{y!} \qquad y = 0, 1, 2, 3, \ldots$$

An important property of the Poisson distribution is that expectation equals variance and both are determined by one parameter, λ, thus $\mathrm{E}(y) = \mathrm{Var}(y) = \lambda$. In order to obtain a conditional probability function, we need to observe that λ must be greater than zero for arbitrary values of X and the regression parameters β. If we let $\lambda(x) = \exp(\beta_0 + \beta_1 x)$, or more generally $\lambda(x) = \exp(x'\beta)$, the functional form guarantees that this condition is fulfilled. In Chapter 8, we present in greater detail various probability functions that can be used for count dependent variables.

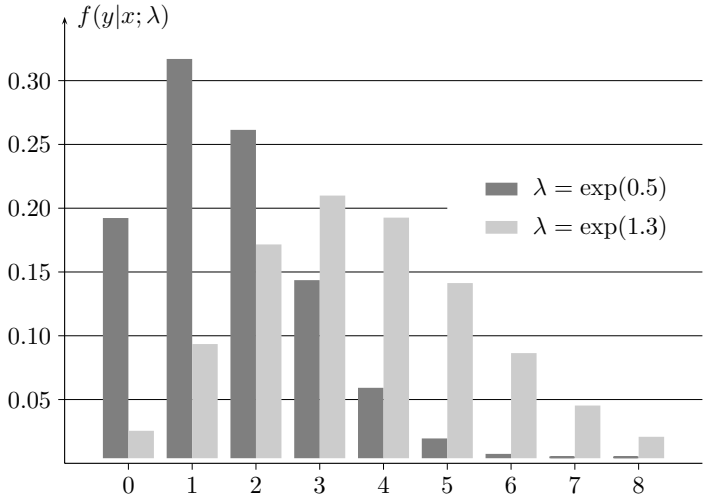
Exercise 2.10.

Suppose that $Y|X$ is Poisson distributed with $\lambda(x) = \exp(0.5 + 0.8x)$.

- Determine $P(Y = 0|X = 0)$ and $P(Y = 0|X = 1)$.
- Determine $P(Y = 2|X = 0)$ and $P(Y = 2|X = 1)$.

The essential idea behind the modeling of conditional probabilities can be illustrated graphically as well. Figure 2.2 refers to the set-up in Exercise 2.10. If $X = 1$, then Y is Poisson distributed with parameter $\lambda = 1.3$. If $X = 0$, however, then Y is Poisson distributed with parameter $\lambda = 0.5$. Thus, we can draw bar diagrams for the two Poisson probability functions, as in Figure 2.2. The dark gray bars plot the probability function with the low parameter value, the light gray bars plot the probability function with the higher parameter value. The changes in the probabilities show the effect of an increase of the regressor X from 0 to 1. For example, if X is a binary explanatory variable, this shows the difference in the probability function with or without a certain attribute. This is what we usually refer to as discrete probability effect, compare to equation (2.11). In the present case we can see, as expected, that the probability of small outcomes decreases, and the probability of large outcomes increases, as X changes from zero to one.

Fig. 2.2. *Predicted Poisson Probabilities*



In limited (but continuous) dependent variable models, distributional assumptions are also important, albeit for different reasons. Recall from Chapter 1 that limited dependent variables are subject to censoring or truncation, or are "naturally" limited in range. Even if the interest remains firmly focused on the conditional expectation function, the distributional model makes it possible to model the underlying population model based on a sample that is non-randomly selected. In principle, continuous conditional probability (or better, conditional density) models are constructed just as their discrete counterparts. We give here just two indicative examples. First, let Y be exponentially distributed with density function

$$f(y;\lambda) = \lambda \exp(-\lambda y) \qquad \lambda > 0 \tag{2.22}$$

In this model, we could specify λ as a function of the explanatory variables by assuming $\lambda = \exp(x'\beta)$. Second, the normal linear regression model can in fact be interpreted as a conditional density model. The density function of the $Normal(\mu, \sigma^2)$-distribution can be written as

$$f(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right] = \frac{1}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right) \tag{2.23}$$

where ϕ is used as a symbol for the density function of the standard normal distribution. In the standard linear regression model, we usually specify the conditional expectation function $\mathrm{E}(y|x)$ as a linear index function. Thus, we could specify the mean parameter μ of the normal distribution here as a

function of the explanatory variables in exactly the same way, $\mu(x) = x'\beta$, from which we obtain the normal linear model where

$$f(y|x;\theta) = \frac{1}{\sigma}\phi\left(\frac{y - x'\beta}{\sigma}\right) \tag{2.24}$$

where $\theta = (\beta' \ \sigma^2)'$. Moreover, if we would like to model a specific form of heteroscedasticity we could write the variance parameter σ^2 as a function of explanatory variables. Of course, in the latter case we should take into account the non-negativity constraint of a variance. But as before with qualitative dependent variables, we have an idea of how the observed dependent variable has been generated, for example by drawing from a normal distribution. Then we specify the parameters of the density in terms of the explanatory variables, and this yields our specification of a conditional probability model. In Chapters 7 and 8, we will present a number of models in which the dependent variable is continuous but limited in a certain way, and where we use the distributional assumption to infer the parameters of the population from the sample that has been drawn.

2.4 Further Exercises

Exercise 2.11 Consider the following conditional distribution of living-space per household in Switzerland in the years 1980 and 1990:

number of rooms	% of households in 1980	% of households in 1990
1	10.7	8.3
2	14.2	14.9
3	31.4	31.1
4	24.8	25.9
5 or more	18.9	19.8

Which of the following statements are supported by the data, which are not? What are the assumptions needed?

a) The average number of rooms per household increased between 1980 and 1990.

b) The median number of rooms per household increased between 1980 and 1990.

c) The number of households living in a flat with one room decreased between 1980 and 1990.

d) The proportion of households living in a flat with one room decreased between 1980 and 1990.

e) The number of rooms is exponentially distributed.

f) The number of rooms is Poisson distributed.

Exercise 2.12 In a study of smoking behavior, a random sample of 16-year-old males in the Canton of Zurich is asked to state the number of cigarettes smoked per day. There are 20 responses, and the average number of cigarettes is 3.6, with standard deviation 0.7.

Which of the following statements are supported by the data, which are not? What are the assumptions needed?

a) The average number of cigarettes consumed per day by 16-year-old males in Canton Zurich is 3.6.

b) The average number of cigarettes consumed per day by 16-year-old youth in Switzerland is 3.6.

c) The average number of cigarettes consumed per day by 16-year-old males in Canton Zurich is 7.7.

d) The average number of cigarettes consumed per day by 16-year-old male smokers is 3.6.

Exercise 2.13 Suppose that two random variables y and x are related in the population as in one of the following models

$$\text{Model 1:} \quad y = \alpha_0 + \alpha_1 x + u \qquad \text{Model 2:} \quad \log y = \beta_0 + \beta_1 x + u$$

where $\log(\cdot)$ denotes the natural logarithm, and u denotes a random term with $\mathrm{E}(u|x) = 0$.

a) For each of the models, determine the absolute and relative marginal mean effect $\partial \mathrm{E}(y|x)/\partial x$. Answer precisely.

b) For each of the models, determine the absolute and relative change in $\mathrm{E}(y|x)$ for a unit increase in x. Answer precisely.

c) How do your answers in a) and b) change, if you replace x by $\log x$?

Exercise 2.14 Suppose that y is Poisson distributed with probability function $f(y;\lambda) = \exp(-\lambda)\lambda^y/(y!)$, and specify a conditional probability model by assuming that $\lambda = \exp(\beta_0 + \beta_1 x)$.

a) Determine the absolute (relative) marginal probability effects (MPE's) $\partial P(y|x)/\partial x$ for $y = 0, 1, 2, 3$. Answer precisely.

b) How does your answer in a) change if you replace x by $\log x$?

Exercise 2.15 Let y denote a discrete random variable with probability function given by

y	1	2	3	4
$f(y)$	4/10	1/10	3/10	2/10

a) Find the probabilities $P(y = 2)$, $P(y \leq 2)$, and $P(y > 2)$.

b) Find $\mathrm{E}(y)$ and $\mathrm{Var}(y)$.

c) Draw $F(y)$ and find the median of y. Is the median a sensible measure for a discrete random variable? Why (not)? Which alternatives may be used instead?

Exercise 2.16 In the winter semester 2004/2005, a total of 23421 students enrolled in the University of Zurich. The relative frequencies conditional on gender are as follows:

field of study	men	women
theology	0.004	0.006
law	0.073	0.075
economics	0.103	0.040
medicine	0.046	0.052
vetsuisse	0.006	0.022
arts	0.174	0.289
sciences	0.058	0.050
Total	0.464	0.534

a) Describe the nature of the variable *field of study*.

b) Find the probability $f(economics)$ and the conditional probability $f(economics|woman)$.

c) Find the conditional probability $f(woman|economics)$.

Exercise 2.17 Let y denote a continuous random variable with density function $f(y) = 2y$ with $0 \leq y \leq 1$. Find the following probabilities

a) $P(y > 0.3)$

b) $P(y \geq 0.4)$

c) $P(0.25 < y < 0.75)$

Exercise 2.18 Let y denote a continuous random variable following a normal distribution with $y \sim Normal(3, 2)$.

a) Find $P(y \leq 2.5)$, $P(y > 4)$, and $P(|y - 3| > 1)$.

b) What distribution does $z = 2y + 3$ have?

c) What distribution does $z = 0.5(y - 3)^2$ have?

Exercise 2.19 Suppose that you have a random sample $y_1, \ldots, y_n$ of size $n = 100$ which has been drawn from a normally distributed population with $y_i \sim Normal(35, 100)$. Find $P(\bar{y} < 34)$.

Exercise 2.20 Suppose that you have two continuous random variables x and y with joint density function $f(x, y) = x + y$, $0 \leq x \leq 1$, $0 \leq y \leq 1$. Calculate $\text{Cov}(x, y)$.

Exercise 2.21 True or False? Evaluate the following statements critically.

a) The variance of a random variable Y with mean $\text{E}(y) = \mu$ can be expressed as $\text{Var}(y) = \text{E}(y^2) - \mu$.

b) $\text{E}(xy) = \text{E}(x)\text{E}(y)$ is a sufficient condition for independence of two random variables X and Y.

c) Let $Y_1 \sim Normal(\mu_1, \sigma_1^2)$ and $Y_2 \sim Normal(\mu_2, \sigma_2^2)$ denote two independent and normally distributed random variables. Then $Y_1 + 3Y_2 \sim Normal(\mu_1 + 3\mu_2, \sigma_1^2 + 3\sigma_2^2)$.

d) Let Y denote a Poisson distributed random variable with probability function $f(y) = \exp(-2)2^y/(y!)$, $y = 0, 1, 2, \ldots$. The probability $P(Y \geq 2)$ is 0.594.

e) The probability of observing at most one tail when flipping a fair coin five times is 0.233.

Exercise 2.22 Assume that you have access to a survey on video games. The data comprises 91 individuals aged 18 to 31 and contains the following information:

- Time spent playing video games in the week prior to the survey — in hours, no answer (-1)
- Do you like to play video games? — never played (1), very much (2), somewhat (3), not really (4), not at all (5)
- How often do you play video games? — daily (1), weekly (2), monthly (3), every semester (4)
- Gender — male (1), female (2)
- Age — in years
- Expected end-of-school grades — from 0 to 100 points

a) Define a research question that you could study with the data.

b) Think about an appropriate probability distribution function that has the same support as the dependent variable you chose in (a).

c) What parameters does the distribution function have? How would you specify the parameters in terms of the explanatory variables?

d) How would you proceed with the variables?

Exercise 2.23 Assume that you have access to data from the Child Health and Development Studies (CHDS), a comprehensive investigation of all pregnancies that occurred between 1960 and 1967 among women in the Kaiser Foundation Health Plan in the San Francisco East Bay area. The variables are:

- Birth weight — in ounces, unknown (999)
- Length of pregnancy — in days, unknown (999)
- First born — yes (1), no (2), unknown (9)
- Mother's age — in years
- Mother's height — in inches
- Mother's prepregnancy weight — in pounds.
- Smoking status of mother — not now (1), yes now (2)

a) Define a research question that you could study with the data.

b) Think about an appropriate probability distribution function that has the same support as the dependent variable you chose in (a).

c) What parameters does the distribution function have? How would you specify the parameters in terms of the explanatory variables?

d) How would you proceed with the variables?

3

Maximum Likelihood Estimation

3.1 Introduction

Consider the problem of estimating the unknown parameter of a population when the population distribution is known (up to the unknown parameter or unknown vector of parameters), and when a random sample of n observations has been drawn from that population. Common examples are sampling from a **Bernoulli distribution** with unknown parameter π, sampling from a **Poisson distribution** with unknown parameter λ, or sampling from a **normal distribution** with unknown parameters μ and σ^2. Generically, we can write the probability or density function of y_i, $i = 1, \ldots, n$ as $f(y_i; \theta)$, where y_i is the i-th draw from the population and θ is the unknown parameter. Throughout the book we assume independent sampling, i.e., the i-th draw from the population is independent from all other draws $i' \neq i$. Then, the joint probability function of the sample is simply the product of the individual probability functions:

$$f(y_1, \ldots, y_n; \theta) = \prod_{i=1}^{n} f(y_i; \theta) \tag{3.1}$$

The easiest way to introduce maximum likelihood estimation is in the context of such unconditional models, where there is only one variable y and no explanatory variable x. The extension of the maximum likelihood method to conditional models $f(y_i|x_i; \theta)$, such as those described in the previous chapter, does not change any of the fundamental principles, although some of the implementation issues become more complex.

3.2 Likelihood Function

In the case of a discrete random variable, equation (3.1) is the joint probability of the sample given the parameters. In the case of a continuous random variable, it is the joint density. Alternatively, we can interpret (3.1) not as a function of the random sample $y_1, \ldots, y_n$ given the parameter θ, but rather as a function of θ for a given random sample $y_1, \ldots, y_n$. When we do this, we call (3.1) the **likelihood function**, and we denote it by capital L:

$$L(\theta; y) = \prod_{i=1}^{n} L(\theta; y_i) = \prod_{i=1}^{n} f(y_i; \theta) \tag{3.2}$$

where $y = (y_1, \ldots, y_n)'$. $L(\theta; y_i)$ is the likelihood contribution of the i-th observation, and $L(\theta; y) = L(\theta; y_1, \ldots, y_n)$ is the likelihood function of the whole sample. Equation (3.2) says that, for any given sample y, the likelihood of having obtained the actual sample that we are using depends on the parameter θ. As the name suggests, the basic idea of maximum likelihood estimation is to find a set of parameter estimates, say $\hat{\theta}$, such that this likelihood is maximized. This principle is widely applicable. Whenever we can write down the joint probability function of the sample we can in principle use maximum likelihood estimation.

It now becomes important to distinguish between an **estimate** and an **estimator**. While the maximum likelihood estimator $\hat{\theta}$ is a random variable that assigns the corresponding maximizing value to each possible random sample $y_1, \ldots, y_n$, and is thus a function of the data, the estimate is the value taken by that function for a specific data set. The same distinction can be made for the likelihood function itself, or for any function of the likelihood function. For instance, for each point θ_p, $L(\theta_p; y)$ is a random variable, as are $\log L(\theta_p; y)$ or $\partial \log L(\theta_p; y)/\partial\theta$, since all these functions depend on the random sample that has been drawn. Of course, in practice, a single sample is the only information we have. However, the derivation of general properties of the maximum likelihood estimator, such as **consistency** or **asymptotic normality**, require the analysis of the behavior of the estimator in repeated samples, which can be conducted based on the assumption that we know the true data generating process $f(y; \theta_0)$.

We conclude this section with two additional remarks. First, the definition of the likelihood function in (3.2) is somewhat more restrictive than necessary. Any function that is proportional to (3.1) can serve as a likelihood function, that is, we require that $L(\theta; y_1, \ldots, y_n) = c \cdot \prod_{i=1}^{n} f(y_i; \theta)$ where c is any positive proportionality constant. This is the definition given by Fisher (1922) in his description of the method in its original form:

> *"The likelihood that any parameter (or set of parameters) should have any assigned value (or set of values) is proportional to the probability that if this were so, the totality of observations should be that observed."* (Fisher, 1922, p. 310)

Second, it is not the likelihood function itself, but rather the **log-likelihood function**, that is most often used in practice. By taking logarithms of equation (3.2) we obtain

$$\log L(\theta; y) = \sum_{i=1}^{n} \log f(y_i; \theta) \tag{3.3}$$

Since the log-likelihood function is a monotonically increasing function of $L(\theta; y)$, any maximizing value $\hat{\theta}$ of $\log L(\theta; y)$ must also maximize $L(\theta; y)$. Taking logarithms has (at least) two advantages. First, the likelihood function may become extremely small or large. For qualitative data, where $f(y_i; \theta)$ is a probability, and therefore $0 \leq f(y_i; \theta) \leq 1$, the product over a large sample can easily be so small that it moves below the range of floating point numbers that can be represented by a computer. Continuous data density functions $f(y_i; \theta)$ can be greater than 1 so that extremely large values are possible. All this is avoided by taking logarithms which converts products into sums. Second, the mathematical manipulations required to obtain analytical results for the value and the distribution of the maximum likelihood estimator are much simpler when they are based on sums, as in (3.3), rather than on products. Finally, using the log-likelihood function makes clear that neglecting the proportionality constant c is without loss of generality. This would be an additive term $\log c$ which does not depend on θ, and therefore is irrelevant in the maximization.

Example 3.1. Sampling from a Bernoulli Distribution (Part I)

Assume that a random sample of size n has been drawn from a Bernoulli distribution with parameter π. Then the likelihood function and the log-likelihood function have the form

$$L(\pi; y) = \prod_{i=1}^{n} (1-\pi)^{1-y_i} \pi^{y_i}$$

$$\log L(\pi; y) = \sum_{i=1}^{n} (1-y_i) \log(1-\pi) + y_i \log \pi$$

In order to illustrate that $L(\pi; y_1, \ldots, y_n)$ is a random variable, Figure 3.1 plots the likelihood function for two different samples of size $n = 5$. The sample $(0, 0, 0, 1, 1)$ has the likelihood function $L_1(\pi) = (1-\pi)^3\pi^2$. The sample $(0, 0, 1, 1, 1)$ has the likelihood function $L_2(\pi) = (1-\pi)^2\pi^3$. We see that both likelihood functions are bell-shaped with a unique maximum. In the first sample, where the proportions of ones is 40 percent, the maximum is reached at $\pi = 0.4$. In the second sample, where the proportions of ones is 60 percent, the maximum is reached at $\pi = 0.6$. This is no coincidence, as we will see later in this chapter.

Fig. 3.1. *Likelihood Function for the Bernoulli Example*

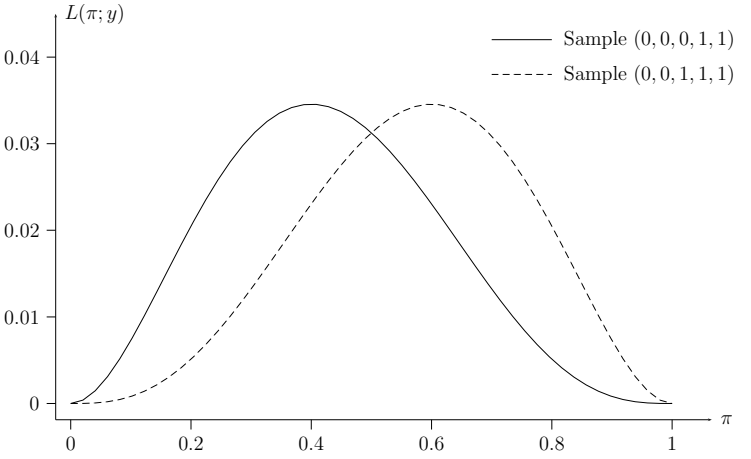


Figure 3.1 shows the likelihood functions for specific samples. Actually, different samples can have the same likelihood function. For example, the samples $(0, 0, 0, 1, 1)$ and $(1, 0, 0, 0, 1)$ both have the likelihood function $L_1(\pi)$. Since we can assign a probability to each sample, depending on the true population parameter π_0, we can assign a probability to each likelihood function as well. The likelihood function $L_1(\pi)$ is obtained whenever there are two ones and three zeros. The probability of such a sample is

$$\binom{5}{2}\pi_0^2(1-\pi_0)^3$$

For example, if $\pi_0 = 0.5$, then $L_1(\pi)$ has probability $10 \times 0.5^5 = 0.31$.

1

3.2.1 Score Function and Hessian Matrix

As indicated earlier, the main object of interest is not the likelihood function $L(\theta; y)$ but rather the log-likelihood function $\log L(\theta; y)$, and its first and second-order derivatives, $\partial \log L(\theta; y)/\partial\theta$ and $\partial^2 \log L(\theta; y)/\partial\theta\partial\theta'$. The first derivative of the log-likelihood function, $\partial \log L(\theta; y)/\partial\theta$, is called the **score function**, or simply **score**. We write $s(\theta; y)$. The second derivative of the log-likelihood function, $\partial^2 \log L(\theta; y)/\partial\theta\partial\theta'$ is commonly referred to as the **Hessian matrix**, or simply **Hessian**. We denote it as $H(\theta; y)$.

In almost all applications, θ is multidimensional – recall the basic regression framework, where there are k explanatory variables and $k + 1$ regression

parameters. Hence, as in least squares estimation, we will need to use some definitions from multivariate calculus in order to properly define $s(\theta; y)$ and $H(\theta; y)$. Assume that θ includes a total of p parameters $\theta = (\theta_1, \ldots, \theta_m)'$. Then we define $s(\theta; y)$ as $(p \times 1)$ column vector with the following property:

$$s(\theta; y) = \frac{\partial \log L(\theta; y)}{\partial \theta} = \begin{pmatrix} \dfrac{\partial \log L(\theta; y)}{\partial \theta_1} \\ \dfrac{\partial \log L(\theta; y)}{\partial \theta_2} \\ \vdots \\ \dfrac{\partial \log L(\theta; y)}{\partial \theta_p} \end{pmatrix}$$

Similarly, we define $H(\theta; y)$ as the $(p \times p)$ matrix with the following property:

$$H(\theta; y) = \frac{\partial^2 \log L(\theta; y)}{\partial \theta \partial \theta'} = \begin{pmatrix} \dfrac{\partial^2 \log L(\theta; y)}{(\partial \theta_1)^2} & \dfrac{\partial^2 \log L(\theta; y)}{\partial \theta_1 \partial \theta_2} & \cdots & \dfrac{\partial^2 \log L(\theta; y)}{\partial \theta_1 \partial \theta_p} \\ \dfrac{\partial^2 \log L(\theta; y)}{\partial \theta_2 \partial \theta_1} & \dfrac{\partial^2 \log L(\theta; y)}{(\partial \theta_2)^2} & \cdots & \dfrac{\partial^2 \log L(\theta; y)}{\partial \theta_2 \partial \theta_p} \\ \vdots & & \ddots & \vdots \\ \dfrac{\partial^2 \log L(\theta; y)}{\partial \theta_p \partial \theta_1} & \dfrac{\partial^2 \log L(\theta; y)}{\partial \theta_p \partial \theta_2} & \cdots & \dfrac{\partial^2 \log L(\theta; y)}{(\partial \theta_p)^2} \end{pmatrix}$$

Thus, the Hessian matrix collects all second-order partial derivatives on its main diagonal, and all cross partial derivatives on the off-diagonal elements. The Hessian matrix is by necessity **symmetric**, since the cross-derivatives are invariant to the order of differentiation. Moreover, if the log-likelihood function is concave in θ, $H(\theta; y)$ is said to be **negative definite**. In the scalar case, for $p = 1$, this simply means that the second derivative of the log-likelihood function is negative (for a more formal definition, see for example Greene, 2003).

Because of the additivity of terms in the log-likelihood function (3.3), the first and second derivatives are additive functions as well. That is, we can write

$$s(\theta; y) = \sum_{i=1}^{n} s(\theta; y_i) \qquad \text{where} \qquad s(\theta; y_i) = \frac{\partial \log L(\theta; y_i)}{\partial \theta}$$

and

$$H(\theta; y) = \sum_{i=1}^{n} H(\theta; y_i) \qquad \text{where} \qquad H(\theta; y_i) = \frac{\partial^2 \log L(\theta; y_i)}{\partial \theta \partial \theta'}$$

It is important to keep in mind that both score and Hessian depend on the sample $y = (y_1, \ldots, y_n)$ and are therefore random variables (they differ in repeated samples). Later on, we determine the expectation of the score vector and of the Hessian matrix, two expressions that are important in the further analysis of maximum likelihood estimation.

3.2.2 Conditional Models

All results mentioned so far require only minor modifications, if conditional rather than marginal probability models are considered. Recall the examples of conditional probability models from Chapter 2.

- $y_i|x_i$ is Bernoulli distributed with parameter $\pi_i = \exp(x_i'\beta)/[1+\exp(x_i'\beta)]$.
- $y_i|x_i$ is Poisson distributed with parameter $\lambda_i = \exp(x_i'\beta)$
- $y_i|x_i$ is normally distributed with parameters $\mu_i = x_i'\beta$ and σ^2.

In order to accommodate such models within the previous framework, we have to extend the assumption of random sampling to pairs of observations (y_i, x_i), requiring that the i-th draw is independent from all other draws $i' \neq i$. All we need to do then is to replace the marginal probability or density function $f(y_i;\theta)$ with the conditional probability or density function $f(y_i|x_i;\theta)$ implied by the model. Now, θ comprises the β's plus any other parameters of the model, for example σ^2 in the case of the normal linear model. Thus, if β is a $(k+1) \times 1$ vector, then the score functions in the Bernoulli and Poisson examples has dimension $(k+1) \times 1$ as well, and the Hessian matrix has dimension $(k+1) \times (k+1)$. In the case of a normal distribution, we have the dimensions $(k+2) \times 1$, and $(k+2) \times (k+2)$, respectively.

3.2.3 Maximization

The value of θ that maximizes $L(\theta;y)$ is called the **maximum likelihood estimator** (or **ML estimator**). We use the symbol $\hat{\theta}$, or $\hat{\theta}_{ML}$, if the type of estimation procedure used is not clear from the context. Since the logarithm is a monotonic transformation, any value $\hat{\theta}$ that maximizes $L(\theta;y)$ also maximizes $\log L(\theta;y)$. As a starting point, consider the problem of determining $\hat{\theta}$ such that

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \log L(\theta;y) \tag{3.4}$$

Here, Θ denotes the parameter space in which the parameter vector θ lies. Usually, we will assume that $\Theta = I\!R^p$ for some $p \geq 1$. This implies that Θ is unbounded, and we must live with the possibility that no ML estimator exists, even if the log-likelihood function is continuous with respect to θ.

However, if there is an interior solution to the maximization problem (as is almost always the case in the standard applications considered in this book), then we will find it by solving the necessary **first-order conditions** for a maximum, namely that the first derivative of the log-likelihood function, i.e., the score vector, is equal to zero

$$\left[\frac{\partial \log L(\theta;y)}{\partial \theta}\right]_{\theta=\hat{\theta}} = s(\hat{\theta};y) = 0 \tag{3.5}$$

For a necessary *and* sufficient condition, we require in addition that the Hessian matrix

$$\left[\frac{\partial^2 \log L(\theta; y)}{\partial\theta\,\partial\theta'}\right]_{\theta=\hat{\theta}} = H(\hat{\theta})$$

is negative definite, provided there is a solution at an inner point of the parameter space Θ. This maximum could be local or global. In "well-behaved" cases – and most problems considered in this book are well-behaved in this sense – the log-likelihood function is **globally concave**, from which it follows that the solution to the first-order condition gives the unique and global maximum of the log-likelihood function.

Example 3.2. Sampling from a Bernoulli Distribution (Part II)

Assume that a random sample of size n has been drawn from a Bernoulli distribution, as before. The score is given by

$$s(\pi; y) = \sum_{i=1}^{n} \frac{y_i - \pi}{\pi(1-\pi)}$$

Solving the first-order condition

$$\sum_{i=1}^{n} \frac{y_i - \hat{\pi}}{\hat{\pi}(1-\hat{\pi})} = 0$$

we find that

$$\hat{\pi} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$$

In the Bernoulli case, the ML estimator for the probability of a one is thus equal to the sample mean, i.e., the proportion of ones in the sample. The second derivative of the log-likelihood function is equal to

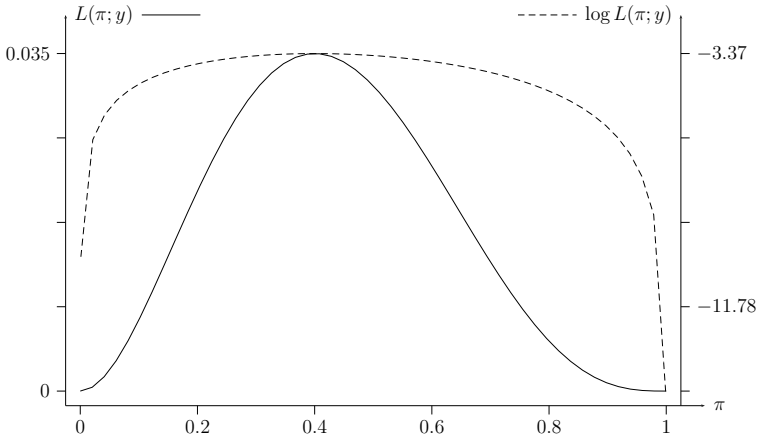
$$H(\pi; y) = \sum_{i=1}^{n} -\frac{y_i}{\pi^2} - \frac{1-y_i}{(1-\pi)^2}$$

This term is negative for all possible samples $(y_1, \ldots, y_n)$ as long as there is variation in y. And only in this case, an inner solution with well-defined score function and Hessian will exist. Otherwise, if all observed values are either one or zero, the likelihood function becomes either $L(\pi; y) = (1-\pi)^n$ or $L(\pi; y) = \pi^n$. The maximizing values of π are then zero or one, respectively, which is at the boundary of the parameter space. We say that the model predicts the outcome perfectly. In the conditional Bernoulli model, perfect prediction

causes a breakdown of the maximum likelihood method. For instance, if $\pi_i = \exp(x_i'\beta)/[1+\exp(x_i'\beta)]$, the parameter π_i cannot become zero or one for any finite value of β.

Now assume that a particular sample has been realized, for example $(0,0,0,1,1)$. Accordingly, the realization of the ML estimator $\hat{\pi}$ will be the estimate $\hat{\pi} = 0.4$. This situation is depicted in Figure 3.2, where the likelihood and log-likelihood functions for this particular sample are plotted. The maximum is reached at $\hat{\pi} = 0.4$. The corresponding value of the likelihood function is 0.035, and -3.37 for the log-likelihood function.

Fig. 3.2. *Likelihood and Log-Likelihood in the Bernoulli Example*



Exercise 3.1.

- Consider a random sample of size n from a population with exponential density function:

  $$f(y_i;\lambda) = \lambda \exp(-\lambda y_i) \qquad \lambda > 0,\ y_i > 0$$

  Find the ML estimator of λ. Check the first and second-order conditions for a maximum.

Exercise 3.2.

- Consider a random sample of size n from a population with Poisson probability function:

$$f(y_i;\lambda) = \frac{\exp(-\lambda)\lambda^{y_i}}{y_i!} \qquad \lambda > 0,\ y_i = 0, 1, 2, \ldots$$

  Find the ML estimator of λ. Check the first and second-order conditions for a maximum.

3.3 Properties of the Maximum Likelihood Estimator

In probability models, the rule of choosing parameters such that the likelihood of observing the actual data is maximized appears eminently sensible. However, we need to study more formally, whether, and under what conditions, such a method makes sense from a statistical point of view as well. Is the ML estimator a good estimator to use? Is it unbiased? Is it consistent? Does it use the information provided in the data efficiently? What is its distribution?

One important aspect of ML estimation is that the method indeed has desirable properties and a known sampling distribution, regardless of the particulars of the probability model, as long as the model is correctly specified. For then, the maximum likelihood estimator is

1. consistent
2. asymptotically normal
3. efficient

Consistency means that, as the sample size increases, the ML estimator tends in probability toward the true parameter value. Moreover, for large sample sizes, the ML estimator will have an **approximate normal distribution** centered around the true parameter value. Finally, **efficiency** means that the ML estimator will have a smaller (asymptotic) variance than other consistent estimators (technically, other consistent uniformly asymptotically normal estimators). We discuss what happens if the ML estimator is based on a misspecified model in Section 3.5.4.

There is no hard and fast rule for how large the sample size should be for these properties to be good approximations. With large cross-sectional datasets, however, the use of asymptotic approximations is usually not an issue. Little of general validity can be said about the **small sample properties** of the ML estimator. In most applications, the ML estimator $\hat{\theta}$ is a non-linear

function of the dependent variable and it will be **biased** in small samples. Its distribution can be far from normal – think about the distribution of the ML estimator $\hat{\pi}$ in a sample of size $n = 5$ obtained from a Bernoulli population; the possible values taken by this estimator are $(0, 0.2, 0.4, \ldots, 1)$. A common way to investigate the small sample properties of ML estimators is by means of **Monte Carlo simulations**. However, such simulations provide results for specific parameter values only, and one cannot prove general results in this way. For information about this issue see Gouriéroux and Monfort (1996).

3.3.1 Expected Score

A crucial property of the ML method is that $\mathrm{E}[s(\theta; y)]$, the expected score, if evaluated at the true parameter θ_0, is equal to zero. If $\mathrm{E}[s(\theta; y)]$ is a vector, this means that each element of the vector is equal to zero. As we will see, this zero expected score property implies consistency of the ML estimator. The result will be formally established for the case of a continuous random variable y because this simplifies the notation considerably. However, the result is perfectly general, subject to a regularity condition stated below.

There are two preliminary remarks. First, we have to be clear whether we are speaking about the score of a single observation $s(\theta; y_i)$ or the score of the sample $s(\theta; y)$. Since under random sampling, $s(\theta; y) = \sum_{i=1}^{n} s(\theta; y_i)$, it is sufficient to establish that $\mathrm{E}[s(\theta; y_i)] = 0$, and the result will follow. Second, a density function is called **regular**, if

$$\frac{\partial}{\partial \theta} \int f(y_i; \theta) dy_i = \int \frac{\partial}{\partial \theta} f(y_i; \theta) dy_i$$

i.e., the order of integration and differentiation can be exchanged. Regularity requires that the domain of integration is independent of θ. An example in which this condition is not satisfied arises if $f(y; a)$ is a uniform distribution $Uniform(0, a)$ and we treat the upper boundary as a parameter to be estimated. In all standard microeconometric models, however, regularity is always fulfilled. Since the integral over any density function is one, it follows from regularity that

$$\int \frac{\partial}{\partial \theta} f(y_i; \theta) dy_i = 0 \tag{3.6}$$

Next, we can express the expected score function using the standard rules regarding the expectation of a function $g(y)$ as

$$\begin{aligned}
\mathrm{E}[s(\theta; y_i)] &= \int s(\theta; y_i) f(y_i; \theta_0) dy_i \\
&= \int \frac{\partial \log f(y_i; \theta)}{\partial \theta} f(y_i; \theta_0) dy_i \\
&= \int \frac{1}{f(y_i; \theta)} \frac{\partial f(y_i; \theta)}{\partial \theta} f(y_i; \theta_0) dy_i
\end{aligned}$$

where the last equality applies the chain rule for taking derivatives of logarithmic functions. In this expression, θ_0 is understood to be the true, albeit unknown, parameter of the model. We see that if the expected score is evaluated at the true parameter θ_0, the $f(y_i;\theta_0)$-expressions cancel and we obtain

$$\mathrm{E}[s(\theta_0; y_i)] = \int \frac{\partial f(y_i;\theta)}{\partial \theta} dy_i \bigg|_{\theta=\theta_0} = 0$$

The expected score for each single observation, if evaluated at the true parameter, is zero, which was to be shown. As a consequence, the expected score vector of the full sample is zero as well.

Example 3.3. Sampling from a Bernoulli Distribution (Part III)

Assume that a random sample of size n has been drawn from a Bernoulli distribution with true parameter π_0, as before. The score function has been derived as

$$s(\pi; y) = \sum_{i=1}^{n} \frac{y_i - \pi}{\pi(1-\pi)}$$

Taking expectations, we obtain

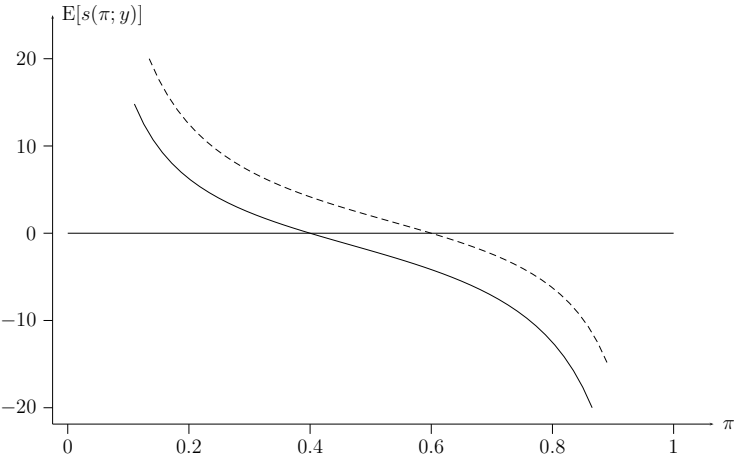
$$\mathrm{E}\left[s(\pi; y)\right] = \sum_{i=1}^{n} \frac{\mathrm{E}(y_i) - \pi}{\pi(1-\pi)} = \frac{n(\pi_0 - \pi)}{\pi(1-\pi)}$$

where $E(y_i) = \pi_0$. Evaluating this function of π at the true parameter, i.e., at the point $\pi = \pi_0$, we see that $\mathrm{E}[s(\pi; y)]_{\pi=\pi_0} = 0$ as required. Figure 3.3 plots two expected score functions for $\pi_0 = 0.4$ and $\pi_0 = 0.6$, respectively. In the case of $\pi_0 = 0.4$ and $n = 5$ we obtain $(2 - 5\pi)/[\pi(1-\pi)]$ (solid line), and we have $(3 - 5\pi)/[\pi(1-\pi)]$ in the case of $\pi_0 = 0.6$ (dashed line). Observe that the expected score is equal to zero at the true parameter value.

3.3.2 Consistency

It should now be intuitively clear why ML estimation is consistent. Under random sampling, the score function is a sum of independent components. By the law of large numbers, the sample score converges in probability to its expected value as the sample size increases beyond bounds. Now, the ML rule says to pick the ML estimator such that the sample score – and hence in the limit the expected score – is equal to zero. Since the zero expected score condition is only satisfied at the true parameter value, it must be the case that in the limit $\hat{\theta} = \theta_0$.

Fig. 3.3. *Expected Score Functions in the Bernoulli Example*



Exercise 3.3.

Consider a random sample of size n from a population with Poisson probability function:

$$f(y_i;\lambda) = \frac{\exp(-\lambda)\lambda^{y_i}}{y_i!}$$

where $y_i = 0, 1, 2, \ldots$ and $\mathrm{E}(y_i) \overset{1}{=} \mathrm{Var}(y_i) = \lambda$, and λ is the unknown parameter that is to be estimated.

- Derive the likelihood function for λ. Are there any terms that can be dropped due to the proportionality property?
- Derive the log-likelihood function and the score for λ. Confirm that the expected score at the true parameter λ_0 is equal to zero.

3.3.3 Information Matrix Equality

In order to go beyond consistency and analyze the variance and the limiting distribution of the ML estimator, we require some results on the **Fisher information matrix** and its relationship to the second derivative of the log-likelihood function – a matrix if θ is a vector:

$$H(\theta; y) = \frac{\partial^2 \log L(\theta; y)}{\partial\theta\partial\theta'} = \sum_{i=1}^{n} \frac{\partial^2 \log f(y_i; \theta)}{\partial\theta\partial\theta'}$$

The information matrix of a sample is simply defined as the negative expectation of the Hessian matrix, algebraically

$$I(\theta) = -\mathrm{E}[H(\theta; y)] \tag{3.7}$$

The information matrix is important in a number of ways in the development of maximum likelihood methodology. First, the information matrix can be used to assess whether the likelihood function is "well behaved". This relates to the issue of identification that will be discussed in Section 3.5.3. A lack of identification means that no matter how large the sample size, the information provided by it is insufficient to estimate the parameters of interest.

Second, the information matrix is important, because it is the inverse of the variance of the maximum likelihood estimator, a result we will derive in Section 3.3.4. And third, it links results for maximum likelihood estimation to an important result on the precision of estimators from general estimation theory, the so-called **Cramér Rao lower bound**. This result states that, under certain regularity conditions, the variance of a consistent estimator of a parameter θ will always be at least as large as $I(\theta)^{-1}$. As a consequence, since the maximum likelihood estimator reaches the Cramér Rao lower bound, it is **asymptotically efficient**.

A final result pertaining to $I(\theta)$ is the so-called **information matrix equality**. This equality establishes that the information matrix can be derived in two ways, either as minus the expected Hessian, as in equation (3.7), or alternatively as the variance of the score function, both evaluated at the true parameter θ_0. In other words,

$$\mathrm{Var}[s(\theta_0; y)] = -\mathrm{E}[H(\theta_0; y)] \tag{3.8}$$

The derivation of the variance of the score function is based on the following considerations. As before, $f(y_i; \theta)$ denotes the probability model for the i-th observation. The Hessian matrix for this observation can be written as

$$\begin{aligned} H(\theta; y_i) &= \frac{\partial s(\theta; y_i)}{\partial\theta'} \\ &= \frac{\partial}{\partial\theta'} \frac{\partial f(y_i; \theta)/\partial\theta}{f(y_i; \theta)} \\ &= \frac{\partial^2 f(y_i; \theta)/\partial\theta\partial\theta'}{f(y_i; \theta)} - \frac{\partial f(y_i; \theta)/\partial\theta}{f(y_i; \theta)^2} \frac{\partial f(y_i; \theta)}{\partial\theta'} \\ &= \frac{\partial^2 f(y_i; \theta)/\partial\theta\partial\theta'}{f(y_i; \theta)} - s(\theta; y_i) s(\theta; y_i)' \end{aligned}$$

Upon taking expectations, the first term on the right disappears since

$$\mathrm{E}\left[\frac{\partial^2 f(y_i; \theta)/\partial\theta\partial\theta'}{f(y_i; \theta)} \bigg|_{\theta=\theta_0} \right] = \int \frac{\partial^2 f(y_i; \theta)}{\partial\theta\partial\theta'} dy_i \bigg|_{\theta=\theta_0} = 0$$

and therefore

$$\mathrm{E}[H(\theta_0; y_i)] = -\mathrm{E}[s(\theta_0; y_i)s(\theta_0; y_i)'] = -\mathrm{Var}[s(\theta_0; y_i)]$$

as stated. The extension to the full sample score and Hessian functions follows from the additivity of the log-likelihood function. This is the information matrix equality.

Example 3.4. Sampling from a Bernoulli Distribution (Part IV)

Assume that a random sample of size n has been drawn from a Bernoulli distribution with parameter π, as before. The score function is given by

$$s(\pi; y) = \sum_{i=1}^{n} \frac{y_i - \pi}{\pi(1-\pi)}$$

The variance of the score is then

$$\mathrm{Var}[s(\pi; y)] = \sum_{i=1}^{n} \frac{\mathrm{Var}(y_i - \pi)}{\pi^2(1-\pi)^2} = \frac{n\pi_0(1-\pi_0)}{\pi^2(1-\pi)^2}$$

Evaluated at the true parameter value ($\pi = \pi_0$), this expression simplifies to

$$\mathrm{Var}[s(\pi; y)] = \frac{n}{\pi_0(1-\pi_0)}$$

The Hessian matrix (here a scalar) is

$$H(\pi; y) = \sum_{i=1}^{n} -\frac{1-y_i}{(1-\pi)^2} - \frac{y_i}{\pi^2}$$

with expectation

$$\mathrm{E}[H(\pi; y)] = \sum_{i=1}^{n} -\frac{1-\pi_0}{(1-\pi)^2} - \frac{\pi_0}{\pi^2}$$

Evaluating the expected Hessian matrix at the true parameter value, we obtain

$$\mathrm{E}[H(\pi_0; y_1, \ldots, y_n)] = -\frac{n}{\pi_0(1-\pi_0)}$$

which corresponds to the information matrix equality.

3.3.4 Asymptotic Distribution

All three properties of the ML estimator, consistency, efficiency, and asymptotic normality, can be summarized in a single result about convergence in distribution. Let $\hat{\theta}$ denote the ML estimator, θ the true parameter value (we drop the zero subscript from now on), and $I(\theta)$ the information matrix of the sample. Then

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} Normal(0, nI(\theta)^{-1})$$

where $\xrightarrow{d}$ stands for "converges in distribution" (see Amemiya, 1994, for an introductory treatment of convergence concepts for random variables). If one wants to emphasize the speed of this type of convergence, one also speaks of "root n convergence", since $\sqrt{n}$ is the value one needs to multiply $\hat{\theta}$ such that the limiting variance of $\sqrt{n}\,\hat{\theta}$ is a non-zero constant matrix.

For large but finite samples, we can therefore write the approximate distribution of $\hat{\theta}$ as

$$\hat{\theta} \overset{\text{app}}{\sim} Normal(\theta, -\mathrm{E}[H(\theta; y)]^{-1}) \tag{3.9}$$

We conclude that the ML estimator is **normally distributed**. Since the expected value of the limiting distribution is the true parameter value θ, the ML estimator is **consistent**. And since its asymptotic variance is the inverse of the information matrix, it is asymptotically **efficient**.

Proof

The main steps of a proof are as follows. Consider a first-order Taylor series approximation of $s(\hat{\theta})$ around the true parameter vector θ:

$$s(\hat{\theta}; y) \approx s(\theta; y) + H(\theta; y)(\hat{\theta} - \theta)$$

Since $s(\hat{\theta}; y)$ is equal to zero by virtue of the first-order condition of a maximum likelihood estimator, we have

$$\hat{\theta} - \theta \approx -H(\theta; y)^{-1} s(\theta; y)$$

or

$$\sqrt{n}(\hat{\theta} - \theta) \approx \left(-\frac{1}{n} H(\theta; y)\right)^{-1} \frac{1}{\sqrt{n}} s(\theta; y)$$

Score function and Hessian are sums of independent components, and law of large numbers and central limit theorems can be invoked. For increasing n, both $s(\theta; y)$ and $H(\theta; y)$ converge to their first moments $\mathrm{E}[s(\theta; y)] = 0$ and $\mathrm{E}[H(\theta; y)] = -I(\theta)$. The first convergence, that of $s(\theta; y)$ to $\mathrm{E}[s(\theta; y)]$, ensures consistency of the maximum likelihood estimator. In fact, this convergence

allows for a re-interpretation of the ML estimator as a **method of moments** estimator: the estimator is the value that solves the sample equivalent to the population moment restriction $\mathrm{E}[s(\theta; y)] = 0$. The asymptotic distribution follows from a central limit theorem, whereby

$$\frac{1}{\sqrt{n}} s(\theta; y) \xrightarrow{d} Normal(0, n^{-1} I(\theta)) \tag{3.10}$$

and therefore

$$\sqrt{n}(\hat{\theta} - \theta) \approx \left(-\frac{1}{n} H(\theta; y) \right)^{-1} \frac{1}{\sqrt{n}} s(\theta; y) \xrightarrow{d} Normal(0, nI(\theta)^{-1})$$

In practice, θ and thus $I(\theta)$ are not known. For the purpose of inference, the true variance covariance matrix of $\hat{\theta}$ can be replaced by a consistent estimator.

Exercise 3.4.

- Consider the ML estimator $\hat{\lambda} = \bar{y}$ of the Poisson parameter λ. What is the asymptotic distribution of $\hat{\lambda}$?

To summarize, the ML estimator has the following properties. The asymptotic distribution is centered at the true parameter θ and its variance goes to zero. Hence, we have mean squared convergence to zero and thus consistency. As mentioned earlier, $I(\theta)^{-1}$ is the smallest variance for any consistent estimator (linear or not). It is the so-called *Cramér-Rao lower bound*. Hence, the ML estimator is asymptotically efficient. Also, the asymptotic distribution is normal, which generates simple (asymptotic) procedures for inference. Of course, these properties have been derived under the assumption that the model is *correctly specified*, i.e., the data generating process, which comprises the population distribution model and the assumption of random sampling, must be valid. The researcher will usually feel more comfortable with the second assumption than with the first. However, it provides some consolation that, in important special cases, some or all of the desirable properties hold, even if the model is misspecified. This is an instance of **quasi-likelihood estimation**, which will be discussed in Section 3.5.4.

3.3.5 Covariance Matrix

Recall from the approximate distribution in (3.9) that the ML estimator $\hat{\theta}$ has variance $\mathrm{Var}(\hat{\theta}) = -\mathrm{E}[H(\theta; y)]^{-1}$. First, we should clarify the nature of the variance. If θ is a vector, say $\theta = (\theta_1, \ldots, \theta_p)'$, then the variance is defined as a matrix that collects on the main diagonal all p variances $\mathrm{Var}(\hat{\theta}_1), \ldots, \mathrm{Var}(\hat{\theta}_p)$,

and on the off-diagonal the $p \times (p-1)$ covariances $\text{Cov}(\hat{\theta}_l, \hat{\theta}_m)$ for $l \neq m$. Thus, we can write

$$\text{Var}(\hat{\theta}) = \begin{pmatrix} \text{Var}(\hat{\theta}_1) & \text{Cov}(\hat{\theta}_1, \hat{\theta}_2) & \cdots & \text{Cov}(\hat{\theta}_1, \hat{\theta}_p) \\ \text{Cov}(\hat{\theta}_1, \hat{\theta}_2) & \text{Var}(\hat{\theta}_2) & \cdots & \text{Cov}(\hat{\theta}_2, \hat{\theta}_p) \\ \vdots & & \ddots & \vdots \\ \text{Cov}(\hat{\theta}_p, \hat{\theta}_1) & \text{Cov}(\hat{\theta}_p, \hat{\theta}_2) & \cdots & \text{Var}(\hat{\theta}_p) \end{pmatrix}$$

This matrix is referred to as a **variance matrix**, **covariance matrix** or, maybe more accurately but also somewhat cumbersomely, as a **variance-covariance matrix**. It is symmetric and positive definite. The covariance matrix is important in every econometric analysis, because it is required for interval estimation and for hypothesis testing.

Unfortunately, $\text{Var}(\hat{\theta}) = -\text{E}[H(\theta; y)]^{-1}$ cannot be known in practice, since it depends on the unknown true parameter θ. Nevertheless, inference can be conducted when the true covariance matrix of $\hat{\theta}$ is replaced by a consistent estimator. It turns out that in the case of ML estimation, three asymptotically equivalent estimators exist. A first candidate is minus the expected Hessian, evaluated at the ML estimate of the parameter (instead of the true value θ):

$$\widehat{\text{Var}}(\hat{\theta})_1 = -\text{E}[H(\hat{\theta}; y)]^{-1}$$

It is frequently the case that the Hessian matrix is a highly nonlinear function of y, making it impossible to obtain an exact expression for the expected value. A second alternative estimator for the covariance matrix is the inverse of minus the *actual* Hessian matrix, evaluated at the maximum likelihood estimator

$$\widehat{\text{Var}}(\hat{\theta})_2 = [-H(\hat{\theta}; y)]^{-1}$$

This is the standard procedure incorporated into most of the software packages used for microeconometric models. Sometimes, even the computation of the Hessian is complicated. However, it follows from the information matrix equality (see Section 3.3.3) that another estimator of the covariance matrix is the variance of the score, which can be estimated by the **outer product** of the score (sometimes also referred to as **outer product of the gradient**)

$$\widehat{\text{Var}}(\hat{\theta})_3 = \left[\sum_{i=1}^{n} s(\hat{\theta}; y_i) s(\hat{\theta}; y_i)' \right]^{-1}$$

The practical relevance of these results is that the three estimators are asymptotically equivalent, and hence one can use whatever is most convenient.

It is also possible to combine two or three of these estimators. Indeed, it can be shown that the estimator

$$\widehat{\text{Var}}(\hat{\theta})_4 = \widehat{\text{Var}}(\hat{\theta})_2 \widehat{\text{Var}}(\hat{\theta})_3^{-1} \widehat{\text{Var}}(\hat{\theta})_2$$

has desirable properties. In particular, it is a consistent estimator of the covariance matrix of $\hat{\theta}$, even if the model is misspecified (the distributional assumption is violated). In this case, $\hat{\theta}$ may or may not be a consistent estimator of θ. If it remains consistent, even if some aspects of the model are misspecified, one also talks about **quasi-maximum likelihood estimation** (Gouriéroux, Monfort and Trognon, 1984).

In the context of hypothesis testing, one may also consider evaluating the above four covariance matrix estimators at the parameter under H_0, i.e., at $\theta = \theta_0$, rather than at $\hat{\theta}$. However, this does not make any difference, since $\hat{\theta}$ is a consistent estimator of θ. If H_0 is true, $\hat{\theta}$ converges in probability to θ_0, and the asymptotic covariance matrix is the same, regardless of whether $\hat{\theta}$ or θ_0 is used to evaluate the covariance matrix.

Example 3.5. Sampling from a Bernoulli Distribution (Part V)

Assume that a random sample of size n has been drawn from a Bernoulli distribution, as before. Recall that score and Hessian are given by

$$s(\pi; y) = \sum_{i=1}^{n} \frac{y_i - \pi}{\pi(1-\pi)}$$

and

$$H(\pi; y) = \sum_{i=1}^{n} -\frac{y_i}{\pi^2} - \frac{1-y_i}{(1-\pi)^2}$$

respectively. Therefore

$$\widehat{\mathrm{Var}}(\hat{\pi})_1 = \frac{\hat{\pi}(1-\hat{\pi})}{n}$$

$$\widehat{\mathrm{Var}}(\hat{\pi})_2 = \left(\sum_{i=1}^{n} \frac{y_i}{\hat{\pi}^2} + \frac{1-y_i}{(1-\hat{\pi})^2} \right)^{-1} = \left(\frac{n\bar{y}}{\hat{\pi}^2} + \frac{(n-n\bar{y})}{(1-\hat{\pi})^2} \right)^{-1} = \frac{\hat{\pi}(1-\hat{\pi})}{n}$$

$$\widehat{\mathrm{Var}}(\hat{\pi})_3 = \left(\sum_{i=1}^{n} \frac{(y_i - \hat{\pi})^2}{\hat{\pi}^2(1-\hat{\pi})^2} \right)^{-1} = \left(\frac{n\bar{y} - 2n\hat{\pi}\bar{y} + n\hat{\pi}^2}{\hat{\pi}^2(1-\hat{\pi})^2} \right)^{-1} = \frac{\hat{\pi}(1-\hat{\pi})}{n}$$

since $\hat{\pi} = \bar{y}$. In this simple case, all three estimators are identical.

Exercise 3.5.

- Consider the ML estimator $\hat{\lambda} = \bar{y}$ of the Poisson parameter λ. Compute the asymptotic variance of $\hat{\lambda}$ using the three alternative expressions.

3.4 Normal Linear Model

A leading example of ML estimation of a continuous response model is the normal linear model. Of course, we already know very well that the regression parameters can be estimated by ordinary least squares, without making any distributional assumption at all. Indeed, the Gauss-Markov results do not require the errors to be normally distributed. Still, it is instructive to obtain the ML estimator as well, and to compare it to the OLS result. The classical linear regression model is usually written as

$$y_i = x_i'\beta + u_i \qquad i = 1, \ldots, n$$

where $x_i = (1, x_{i1}, \ldots, x_{ik})'$ is a $(k+1) \times 1$-vector of explanatory variables and β is a conformable vector of parameters. Under the assumptions of mean independence and homoscedasticity, the regression parameters of the model can be estimated by **ordinary least squares** and the resulting estimator is the **best linear unbiased estimator**. Now, we assume in addition that u_i is normally distributed with mean 0 and variance σ^2. The resulting **normal linear model** is in the format of a conditional probability model because it follows from the above assumptions that

$$y_i | x_i \sim Normal(x_i'\beta, \sigma^2)$$

We will show now how the parameters of this model, β and σ^2, can be estimated by the maximum likelihood method. The density function for each observation can be written explicitly as

$$f(y_i|x_i; \beta, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{y_i - x_i'\beta}{\sigma}\right)^2\right]$$

Assuming a random sample of n pairs of observations (y_i, x_i), the log-likelihood function is

$$\begin{aligned}\log L(\beta, \sigma^2; y, x) &= \sum_{i=1}^{n} \log f(y_i|x_i; \beta, \sigma^2) \\ &= \sum_{i=1}^{n} \left[-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\sigma^2 - \frac{1}{2}\left(\frac{y_i - x_i'\beta}{\sigma}\right)^2 \right] \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i'\beta)^2 \qquad (3.11)\end{aligned}$$

The first term on the right-hand side can be dropped, because it does not depend on σ^2 or β. Thus, it can be absorbed into the proportionality constant. The first-order conditions for maximizing this log-likelihood are:

$$\frac{\partial \log L}{\partial \beta} = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i(y_i - x_i'\beta) \stackrel{!}{=} 0 \qquad (3.12)$$

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - x_i'\beta)^2 \stackrel{!}{=} 0 \qquad (3.13)$$

Note that the partial derivative $\partial \log L/\partial\beta$ is a vector of dimension $(k+1)\times 1$. Its first element is $\partial \log L/\partial\beta_0 = \sigma^{-2}\sum_{i=1}^{n}(y_i - x_i'\beta)$, its second $\partial \log L/\partial\beta_1 = \sigma^{-2}\sum_{i=1}^{n} x_{i1}(y_i - x_i'\beta)$, and so forth. The ML estimator $\hat{\beta}$ can be determined from the first equation, which can be rewritten as:

$$\sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} x_i x_i'\hat{\beta}$$

such that

$$\hat{\beta}_{ML} = \left(\sum_{i=1}^{n} x_i x_i'\right)^{-1}\left(\sum_{i=1}^{n} x_i y_i\right) \qquad (3.14)$$

Here, $\sum_{i=1}^{n} x_i x_i'$ is a $(k+1)\times(k+1)$ dimensional matrix with diagonal elements $\sum_{i=1}^{n} x_{il}^2$ and off-diagonal elements $\sum_{i=1}^{n} x_{il}x_{im}$ for $l \neq m$. To solve the system of $k+1$ linear equations, we make use of the concept of an **inverse matrix**: A^{-1} is the inverse of A if $A^{-1}A = I$, where I is a diagonal matrix with ones on the main diagonal and zeros elsewhere (see for example Greene, 2003, for further details). The inverse $\left(\sum_{i=1}^{n} x_i x_i'\right)^{-1}$ exists, as long as the explanatory variables $1, x_1, \ldots, x_k$ are linearly independent.

An inspection of the ML estimator $\hat{\beta}_{ML}$ in (3.14) shows that it is the same as the ordinary least squares estimator. Hence, under the assumptions of the normal linear model, and as far as the slope vector β is concerned, there is no difference between maximum likelihood estimation and least squares. The reason is that the score vector is proportional to the normal equations of the least squares problem.

The second parameter of the model, σ^2, can be estimated by maximum likelihood as well. In order to do so, replace β in (3.13) by its ML estimator $\hat{\beta}$, and define the residual $\hat{u}_i = y_i - x_i'\hat{\beta}$. Then

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2$$

This expression differs from the familiar variance estimator, since the divisor is n and not $n-k-1$. In other words, the ML estimator makes no adjustment for the degrees of freedom. Hence, it is biased in small samples. However, since maximum likelihood estimation relies on asymptotic properties anyway, and since the bias disappears quickly for large n, this is not something to worry about.

So far, we have taken for granted that the ML estimators obtained from the first-order conditions do indeed maximize the log-likelihood. Of course, we should also check the second-order conditions to see whether this is the case in the normal linear model. We need to evaluate the three second derivatives

$$\frac{\partial^2 \log L}{\partial\beta\partial\beta'} = -\frac{1}{\sigma^2}\sum_{i=1}^{n} x_i x_i'$$

$$\frac{\partial^2 \log L}{\partial(\sigma^2)^2} = \frac{n}{2\sigma^4} + \frac{1}{\sigma^6}\sum_{i=1}^{n}(y_i - x_i'\beta)^2$$

$$\frac{\partial^2 \log L}{\partial\beta\partial\sigma^2} = -\frac{1}{\sigma^4}\sum_{i=1}^{n} x_i(y_i - x_i'\beta)$$

Because of symmetry, $\partial^2 \log L/(\partial\beta\partial\sigma^2) = \partial^2 \log L/(\partial\sigma^2\partial\beta)$. Collecting terms, we get the Hessian matrix

$$H(\beta,\sigma^2;y,x) = \begin{bmatrix} -\frac{1}{\sigma^2}\sum_{i=1}^{n} x_i x_i' & -\frac{1}{\sigma^4}\sum_{i=1}^{n} x_i(y_i - x_i'\beta) \\ -\frac{1}{\sigma^4}\sum_{i=1}^{n} x_i(y_i - x_i'\beta) & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(y_i - x_i'\beta)^2 \end{bmatrix}$$

The dimension of this matrix is $(k+2)\times(k+2)$. It can be shown that it is a negative definite matrix, provided the x_i's are well-behaved and not collinear. In this case, the log-likelihood function of the normal linear model is globally concave, and $\hat{\beta}$ and $\hat{\sigma}^2$ are indeed the maximizing values.

The information matrix contains minus the expected values of the Hessian matrix:

$$I(\beta,\sigma^2) = \begin{bmatrix} \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i x_i' & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}$$

Its inverse provides the asymptotic covariance matrix of the maximum likelihood estimator in the normal linear regression model.

$$I(\beta,\sigma^2)^{-1} = \begin{bmatrix} \sigma^2(\sum_{i=1}^{n} x_i x_i')^{-1} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}$$

Finally, it follows that

$$\begin{pmatrix} \hat{\beta} \\ \hat{\sigma}^2 \end{pmatrix} \overset{\text{app}}{\sim} Normal\left[\begin{pmatrix} \beta \\ \sigma^2 \end{pmatrix}, \begin{pmatrix} \sigma^2(\sum_{i=1}^{n} x_i x_i')^{-1} & 0 \\ 0 & 2\sigma^4/n \end{pmatrix}\right]$$

An estimator of the covariance matrix can be obtained if we evaluate at the ML estimator $\hat{\sigma}^2$. To summarize, under the standard assumptions of the normal linear model, OLS and ML give the same estimator for β, whereas the estimators for σ^2 differ. The ML estimator $\hat{\sigma}^2_{ML}$ is biased, but consistent and asymptotically efficient.

Exercise 3.6.

- Show that the log-likelihood function, when evaluated at the ML estimates for $\hat{\sigma}^2_{ML}$ and $\hat{\beta}_{ML}$, is equal to $-0.5n(\log(2\pi)+\log\hat{\sigma}^2_{ML}+1)$.
- Suppose a normal linear model has been fitted by ordinary least squares using a sample of 500 observations. The sum of squared errors is 420. Had the model been estimated by maximum likelihood instead, what would the value of the log-likelihood function be?

We conclude with three further interesting observations about the normal linear model. First, we see that we have an instance where the distributional assumption is not critical for the good properties of the ML estimator. Since ML = OLS, and since OLS does not require any distributional assumption to establish unbiasedness and minimum variance in the class of linear unbiased estimators, we can immediately conclude that the ML estimator of β is, under the assumptions of the conditional normal model, unbiased and has minimum variance as well, even if the true distribution is not normal. Second, we gain a bit by making the additional normality assumption, because in the ML framework, we know that the ML estimator now has minimum variance in the class of *all* consistent estimators, not only the linear ones. Third, we should mention that the normal linear model is somewhat atypical (as were the examples given previously for the Bernoulli model without covariates), because it is possible to obtain the ML estimator as a closed-form solution to the first-order condition. In most microdata models, however, such a closed-form solution is not available and numerical solution algorithms will be needed. This, together with some other extensions, is discussed in the next section.

3.5 Further Aspects of Maximum Likelihood Estimation

In this section, we pick up some further issues of maximum likelihood estimation. We start with a discussion of the invariance property. After that, we discuss the estimation of models with non-linear score functions using numerical optimization routines. There are a number of such algorithms available, but we focus on the Newton-Raphson method. Finally, there are situations where the ML estimator does not exist. This may arise due to data deficiencies or modeling deficiencies, in particular a lack of identification.

3.5.1 Invariance and Delta Method

The study of invariance refers to the behavior of ML estimators when the model is reparameterized. Let $\hat{\theta}$ denote the ML estimator of θ. Assume that instead of θ, we are interested in $h(\theta)$, where h is an arbitrary, possibly non-linear function. The invariance property says that the ML estimator of $h(\theta)$ is simply $h(\hat{\theta})$. The intuition for this result is that ML estimation yields, under the assumptions of the model, a consistent estimator and that, for large enough samples, we can treat the estimator as if it were a constant.

Example 3.6. Parameter Constraints

Let $\vartheta = \ln \theta$ with inverse function $\theta = \exp(\vartheta)$. Assume one wants to impose the constraint that θ is non-negative. This is achieved by parameterizing the model in terms of ϑ (since $\exp(\vartheta)$ is nonnegative for all $\vartheta \in \mathbb{R}$) and obtaining the ML estimator for ϑ. The ML estimator for θ then is $\hat{\theta} = \exp(\hat{\vartheta})$. Applications for the log transformation arise, for example, when estimating the parameter σ^2 in a normal linear model, or when estimating the parameter λ in a Poisson model without covariates.

Example 3.7. Expectation of a Log-Normal Distribution

Let $\hat{\mu} = \bar{y}$ be the ML estimator for the mean μ of a homoscedastic normal population and a sample of size n. We know that $\mathrm{E}(\hat{\mu}) = \mu$ and $\mathrm{E}[\exp(\hat{\mu})] = \exp(\mu + 0.5\sigma^2/n)$. Hence, $\exp(\hat{\mu})$ is a biased estimator of $\exp(\mu)$. This is an exact result that holds for any sample size. However, from the maximum likelihood point of view, we are interested in the large sample behavior. Clearly, $\exp(\hat{\mu})$ is a consistent estimator of $\exp(\mu)$ since $\sigma^2/n \to 0$ as n increases. Therefore, invariance holds as required.

Example 3.8. Ratio of Parameters

For another application of the invariance principle, consider the normal linear model $y = \beta_0 + \beta_1 x + u$. The ML estimator of the ratio β_0/β_1 is equal to the ratio of the ML estimators $\hat{\beta}_0/\hat{\beta}_1$.

Exercise 3.7.

- Assume that y is Poisson distributed with parameter $\lambda = \exp(\vartheta)$. Use the likelihood function

$$L(\vartheta) = \prod_{i=1}^{n} \exp(-\exp(\vartheta)) \exp(\vartheta)^{y_i}$$

  to derive the ML estimator for ϑ. How does it compare to the usual ML estimator for λ?
- Assume that $\hat{\mu}$ and $\hat{\sigma}^2$ are the ML estimators of the parameter of the normal linear model. Find the ML estimator for $P(y > 0)$.

Delta Method

In order to obtain asymptotic standard errors for transformed estimators, we can use the **Delta method**. For a scalar parameter θ,

$$\widehat{\mathrm{Var}}[h(\hat{\theta})] = \left[\frac{\partial h(\hat{\theta})}{\partial \theta}\right]^2 \widehat{\mathrm{Var}}[\hat{\theta}]$$

For example, assume that $h(\hat{\theta}) = \exp(\hat{\theta})$. In this case $[\partial h(\hat{\theta})/\partial\theta]^2 = \exp(2\hat{\theta})$ and it follows that $\widehat{\mathrm{Var}}[h(\hat{\theta})] = \exp(2\hat{\theta})\widehat{\mathrm{Var}}[\hat{\theta}]$. If θ and h are vectors, the delta method requires computation of a quadratic form involving the vector of derivatives of the h-function with respect to the elements of θ and the covariance matrix of θ. Formally, we have

$$\widehat{\mathrm{Var}}[h(\hat{\theta})] = \left[\frac{\partial h(\hat{\theta})}{\partial \theta'}\right] \widehat{\mathrm{Var}}[\hat{\theta}] \left[\frac{\partial h(\hat{\theta})}{\partial \theta'}\right]' \tag{3.15}$$

Exercise 3.8.

Suppose a random sample of size n = 100 has been drawn from a Poisson distribution, with sample mean equal to $\bar{y} = 2.0$.

- Calculate the 95% confidence interval for λ.
- Determine the predicted probability $\widehat{P(Y = 1)}$.
- Determine the standard error of $\widehat{P(Y = 1)}$.
- What is the 95% confidence interval for $P(Y = 1)$?

3.5.2 Numerical Optimization

The maximum likelihood principle requires that we find the parameter values that yield the maximum of the (log-) likelihood function. We can proceed with the optimization using any of the following techniques:

- Trial and Error
- Graphical Methods
- Analytical Optimization
- Numerical Methods

The **trial and error** method is a very simple way to find the optimum of a function. We plug in trial values of the parameters, and compare the values of the likelihood function until we find its maximum. This method is generally very inefficient. Nevertheless, a systematical variant called **grid search** is sometimes used to get an idea about the range in which the maximum should be. **Graphical methods** are of practical relevance only if the function to be optimized is linear in the parameters (which is rarely the case), otherwise the results are fairly inaccurate. **Analytical optimization** requires the use of calculus, i.e., we set the score function equal to zero, solve for the maximizing parameters, and evaluate the Hessian matrix. This method is very efficient. However, in many cases a closed-form solution is not available due to the non-linearity of the log-likelihood function.

In this section, we introduce some **numerical methods** for maximizing the log-likelihood function when the first-order condition has no closed-form solution. This is only a very basic introduction to the topic. For more details, refer to textbooks such as Davidson and MacKinnon (1993), Gouriéroux and Monfort (1996), and Greene (2003).

The common element of all numerical maximization algorithms is that they are iterative, and that, departing from some more or less arbitrary starting values, the parameter updates account for the slope and the curvature of the

log-likelihood function at the current parameter value. If the slope is positive for the l-th element of the score vector, then we know that choosing an update $\hat{\theta}_l^{t+1} > \hat{\theta}_l^t$ will tend to improve the function that is to be maximized. However, the update step should not be too large, for then one can "overshoot" the target and move beyond the maximum of the likelihood function. Therefore, the curvature of the log-likelihood function can be used to determine the length of the step. If the curvature of the target function is large, the steps should be smaller than if the curvature is small.

Numerical optimization is an art rather than a science. There is no guarantee that the algorithm will find the correct answer. First, the log-likelihood function may be ill-conditioned, for example because of a lack of identification which means that the log-likelihood is flat in some parts of the parameter space. Second, the maximum that has been found may be a local maximum. In this case, different starting values may lead to different "ML estimators". In many of the standard microdata models considered in this book, it can be shown that the likelihood function is globally concave, and in such a case, algorithms will always perform well and find (to an arbitrary degree of precision) the unique global maximum.

The various available algorithms differ in the way that they compute the gradient and the curvature of the log-likelihood function. In some cases, these functions are provided in analytical form, whereas in others, they are based on numerical approximations. This distinction may be a matter of convenience, when one could in principle derive the analytical forms but does not do so, or of necessity, when the gradient or even the log-likelihood has no closed-form. The latter occurs, for example, in many models of unobserved heterogeneity or in mixture models, where the log-likelihood function includes high-dimensional integrals. In such cases, the optimization algorithms need to be combined with methods of **numerical integration**, or, if this is not possible, with simulation estimators (such as **maximum simulated likelihood**). Clearly, these methods have been helped greatly by the increased computing power, and it is fair to say that one can nowadays write down even the most complicated likelihood function and – provided that a ML estimator exists – be rather sure that a method of maximizing it within a reasonable amount of time can be implemented.

As we said before, a comprehensive treatment of existing numerical optimization routines would fill an entire book, and we cover only some elementary aspects of the issues typically encountered. We start with an example, in which the score equation has no closed-form solution, and then discuss how such models can be estimated with the **Newton Raphson algorithm**.

Example 3.9. The Log-Linear Normal Model

The log-linear normal model is a non-linear regression model, where the conditional model of y_i given x_i is a normal distribution with conditional expectation $E(y_i|x_i) = \exp(x_i'\beta)$ and constant variance σ^2. For simplification, assume that the true σ^2 is known, such that we do not have to find its maximizing value. The conditional density function can therefore be written as

$$f(y_i|x_i;\beta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{y_i - \exp(x_i'\beta)}{\sigma}\right)^2\right]$$

Given a random sample of size n, the log-likelihood of this model is

$$\log L(\beta; y, x) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \exp(x_i'\beta))^2$$

and the score vector wit respect to β is

$$s(\beta; y, x) = \frac{\partial \log L}{\partial \beta} = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i[y_i - \exp(x_i'\beta)]\exp(x_i'\beta)$$

The first-order condition for a maximum of the log-likelihood function with respect to β, $s(\hat{\beta}; y, x) = 0$, has no analytical solution.

Newton-Raphson Algorithm

A frequently used method of quadratic approximation is the **Newton-Raphson method**, or algorithm. It can be motivated as follows. Given any initial parameter estimate, say $\hat{\theta}^0$, we can obtain a second-order approximation of $\log L(\theta)$ around $\hat{\theta}^0$:

$$\log L^*(\theta) = \log L(\hat{\theta}^0) + s(\hat{\theta}^0)'(\theta - \hat{\theta}^0) + \frac{1}{2}(\theta - \hat{\theta}^0)'H(\hat{\theta}^0)(\theta - \hat{\theta}^0) \approx \log L(\theta)$$

where $s(\cdot)$ denotes the score and $H(\cdot)$ the Hessian of the log-likelihood function (see Greene, 2003, for further details on vector and matrix differentiation). Now, we can maximize $\log L^*(\theta)$ (rather than $\log L(\theta)$) with respect to θ, yielding a new parameter value which we call $\hat{\theta}^1$. The first-order condition of this simpler problem is

$$s(\hat{\theta}^0) + H(\hat{\theta}^0)(\hat{\theta}^1 - \hat{\theta}^0) = 0$$

or

$$\hat{\theta}^1 = \hat{\theta}^0 - [H(\hat{\theta}^0)]^{-1} s(\hat{\theta}^0)$$

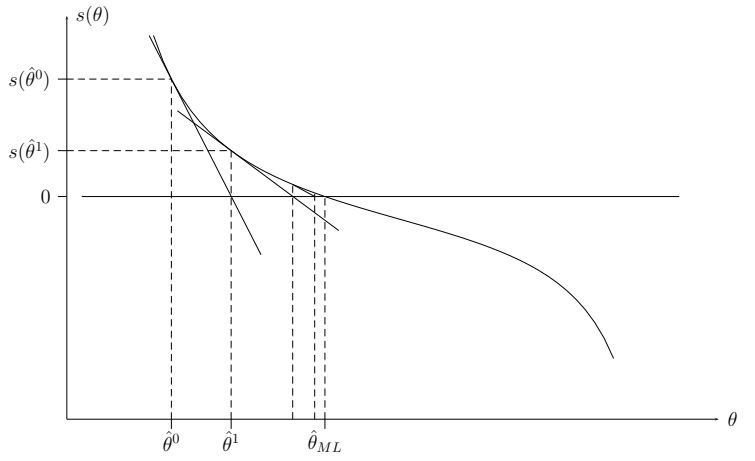
Thus, for an arbitrary starting value $\hat{\theta}^0$, the Newton-Raphson updating rule is given by

$$\hat{\theta}^{t+1} = \hat{\theta}^t - [H(\hat{\theta}^t)]^{-1} s(\hat{\theta}^t) \qquad t = 0, 1, \ldots \tag{3.16}$$

The iterative procedure ends when a predefined **convergence criterion** is satisfied. Possible criteria are the change in the value of the estimate $\hat{\theta}^{t+1} - \hat{\theta}^t$, the change in the log-likelihood values $\log L(\hat{\theta}^{t+1}) - \log L(\hat{\theta}^t)$, or the value of the gradient at the estimate $s(\hat{\theta}^t)$. Convergence occurs when any of these values, or a combination of them, are close to zero (say, smaller than 10^{-5} in absolute value).

Figure 3.4 illustrates graphically the steps involved in the Newton-Raphson algorithm. With an initial parameter value $\hat{\theta}_0$, we obtain the value of the score function $s(\hat{\theta}_0)$. The Hessian matrix $H(\hat{\theta}_0)$, in this case simply a scalar, is the slope of the score function evaluated at $\hat{\theta}_0$. The updating rule in (3.16) requires that we divide the value of the score function at the tangency point by the slope of the tangent, and this equals the distance $\hat{\theta}_0 - \hat{\theta}_1$. But this can be interpreted in terms of simple trigonometric relations. Specifically, the ratio of the distance on the vertical axis, $s(\hat{\theta}_0) - 0$, and the distance on the horizontal axis, $\hat{\theta}_0 - \hat{\theta}_1$, yields the slope of the secant between the two points under consideration. Thus, the updated value $\hat{\theta}_1$ can also be obtained by finding the intersection point of the tangent line with the zero line.

Fig. 3.4. *The Newton-Raphson Algorithm*



Example 3.10. Maximizing a Third-Order Polynomial

Suppose we want to find a maximum of the function $f(x) = x^3 - 3x + 1$. The first and second derivatives are $f'(x) = 3x^2 - 3$ and $f''(x) = 6x$, respectively. We can solve the first-order condition directly to obtain the candidate values -1 and 1. Because $f''(-1) < 0$, the first of the two candidate values is a local maximum of the function. Now, suppose that we use the Newton-Raphson algorithm instead. The update rule is

$$x_{t+1} = x_t - \frac{f'(x_t)}{f''(x_t)} = x_t - \frac{3x_t^2 - 3}{6x_t}$$

Starting, for example, with $x_0 = -2$, we obtain the update sequence -2, -1.25, -1.025, ... which converges rapidly to the local maximum. If we start the iterations with $x_0 = 2$, however, the algorithm does not find the local maximum but rather ends up at the local minimum.

Example 3.11. Newton-Raphson in the Log-Linear Normal Model

In order to implement the Newton-Raphson algorithm, we need to compute the score function and the Hessian matrix and evaluate them iteratively at the current parameter values:

$$s(\beta^t) = \left.\frac{\partial \log L}{\partial \beta}\right|_{\beta=\beta^t} = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i(y_i - \exp(x_i'\beta^t))\exp(x_i'\beta^t)$$

$$H(\beta^t) = \left.\frac{\partial^2 \log L}{\partial \beta\, \partial \beta'}\right|_{\beta=\beta^t} = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i x_i'(y_i - 2\exp(x_i'\beta^t))\exp(x_i'\beta^t)$$

A numerical approximation to the solution can be found by selecting starting values β_0, for example from a linear regression of $\log y$ on x, and then updating according to the Newton-Raphson formula, using the update $\hat{\beta}^{t+1} = \hat{\beta}^t - H(\beta^t)^{-1}s(\beta^t)$, until convergence is reached.

Other Optimization Algorithms

In the Newton-Raphson algorithm, the curvature – and therefore the step length – is determined by the actual Hessian of the sample, evaluated at

the current parameter value. If the expected Hessian is used instead, the resulting algorithm is referred to as the **method of scoring**. While this is a complication in comparison to the base method, since the expectation may not be easy to compute, a clear computational simplification is obtained if the Hessian is replaced by the outer product of the score, as proposed by Berndt et al. (1974), which is referred to as the **BHHH algorithm**. Still other approaches include the so-called **steepest ascent method** (where the score vector is multiplied by the scalar $s(\beta^t)'s(\beta^t)/s(\beta^t)'H(\beta^t)s(\beta^t)$), and the **quadratic hill-climbing method** that corrects for the possibility that the Hessian matrix may not be negative definite at points distant from the optimum. Of course, if the log-likelihood function is globally concave, such a provision is not necessary.

Depending on the application, it may also be preferable to use derivate free algorithms, such as grid search, simplex methods, and simulated annealing, or the **EM algorithm**, which is advantageous in missing data problems (see Greene, 2003, and the references cited therein, for further details).

From a practical point of view, one should note that results might be sensitive to the choice of starting values, and one should check the robustness of the solution by providing different starting values. Dependent on the problem at hand, some algorithms work better than others, some may fail to find the maximum, others will find the solution very fast. Thus, one should try different algorithms and compare the results.

3.5.3 Identification

There are two distinct types of problem that may cause maximum likelihood estimation to fail. The first one is a deficiency of the sample at hand, such as no variation in the dependent variable or no variation in an explanatory variable, or a combination of the two. A common manifestation of such a combination arises when binary dependent variables and binary explanatory variables are mixed and there is no variation in the dependent variable for a given value of the binary explanatory variable. This leads to a perfect prediction problem, and an associated unboundedness of the log-likelihood function. See also Section 3.2.3. To remedy such a problem, one can collect more data and hope that some useful variation arises in the additional observations.

A fundamentally different problem is that of identification, or the lack thereof. Identification is the study of what conclusions can and cannot be drawn in infinite samples, given specified combinations of assumptions and types of data.

> *"Identification problems cannot be solved by gathering more of the same kind of data. These inferential difficulties can be alleviated only by invoking stronger assumptions or by initiating new sampling processes that yield different kinds of data."* (Manski, 1995, p. 3/4)

Identification is a general issue in any econometric model and for any estimation procedure. It is not restricted to maximum likelihood estimation. The following examples are illustrative of the types of identification failure that may occur, and of the solution strategies for overcoming the lack of identification.

Identification by Functional Form

Suppose that a random sample of y_i is taken conditional on $x_i = 1, 2$. This can identify $\mathrm{E}(y_i|x_i = 1)$ and $\mathrm{E}(y_i|x_i = 2)$ but not, without further assumption, $\mathrm{E}(y_i|x_i = 1.5)$. Identification could be achieved in two ways:

- Change the sampling scheme and sample conditional on $x_i = 1.5$ as well.
- Add the assumption that the conditional expectation function has a particular functional form, for example

$$\mathrm{E}(y_i|x_i) = \beta_0 + \beta_1 x_i$$

Identification by Exclusion Restriction

In other cases, identification is achieved by imposing an exclusion restriction. Here is an example:

$$\mathrm{E}(y_i|male_i, female_i) = \beta_0 + \beta_1\, male_i + \beta_3\, female_i$$

$Male_i$ and $female_i$ are dummy variables that take the value 1 if person i is male or female, respectively, and 0 else. In this conditional expectation function, not all parameters β_0, β_1, and β_2 are identified, since $male_i = 1 - female_i$ and therefore

$$\mathrm{E}(y_i|male_i, female_i) = (\beta_0 + \beta_1) + (\beta_2 - \beta_1) female_i$$

In order to handle the identification problem, one can exclude either the constant, or one of the dummy variables. Exclusion restrictions are also a key when estimating the structural parameters from a reduced-form model in the context of simultaneous equations, but this is not discussed further here.

Identification in Probability Models

In a probability framework, two parameters θ_1 and θ_2 are said to be **observationally equivalent** if $f(y, \theta_1) = f(y, \theta_2)$ for all y. A parameter point θ_0 is then identifiable if there is no other $\theta \in \Theta$ which is observationally equivalent (Rothenberg, 1971). An example of a non-identified model, a mixture of two normal distributions, can be found in Maddala (1983, page 300). A sufficient condition for identification is that the information matrix $-\mathrm{E}[H(\theta; y)]$ is a nonsingular matrix.

3.5.4 Quasi Maximum Likelihood

The single most important drawback of maximum likelihood estimation is that it requires the specification of a true probability model. If the specified model is incorrect, the resulting ML estimator is inconsistent in general. This situation reflects a classic trade-off in empirical modeling: to be able to obtain strong results, which entails consistency, asymptotic normal distribution, and the ability to make inferences on conditional probability effects, one must also be prepared to make strong assumptions. The weaker the assumptions of a model, the more robust the estimator will be with regard to model misspecification, but also the weaker the inferences one can draw. An often employed practical approach is not to rely on a single estimate based on a single specification, but rather to employ various models. One can then perform formal specification tests or ask informally whether the key inferences across the different specifications are qualitatively similar.

Whether or not the ML estimator retains some of its desirable properties, even when the underlying probability model is misspecified, needs to be investigated on a case-by-case basis. The study of the behavior of maximum likelihood estimators under misspecification is the topic of so-called **quasi-likelihood estimation** (see, for instance, White, 1982, and Gouriéroux, Monfort and Trognon, 1984). Many frequently used ML estimators remain consistent even if the underlying distributional assumption is violated. The leading example is the estimation of β by maximum likelihood in the normal linear model (Section 3.4). Since the maximum likelihood estimator is identical to the ordinary least squares estimator, it inherits all the desirable properties of the latter, i.e., best linear unbiasedness, which do *not* depend on the distribution of $y_i|x_i$. Of course, we still must require that the conditional expectation function is correct. In fact, all of the univariate examples given in this chapter, as well as the logit and Poisson regression models to be introduced in detail later, are robust to distributional misspecification in the sense that the (conditional) expectation can be estimated consistently, even if the chosen distribution is the wrong one.

3.6 Testing

3.6.1 Introduction

After a model has been specified and the parameters estimated, the next step of the econometric analysis is to conduct **inference**, i.e., to generalize from the estimates obtained for the sample to the population parameters. In all the conditional probability models considered in this book, the focus of attention is on the index parameters that capture the effect of explanatory variables on the conditional probabilities, although the models frequently include other parameters as well (such as σ^2 in the normal linear model).

The ultimate goal of any empirical analysis is to learn something about the values of these parameters in the population, or functions thereof. The ML estimator gives us point estimates. In order to obtain interval estimates and to draw inferences on the population parameters, we need to account for the sampling variability. Recall that the inference step requires formulating a hypothesis and assessing whether, from a statistical point of view, the evidence found in the data speaks strongly against this hypothesis, in which case it is rejected. Testable hypotheses are formally equivalent to **restrictions** imposed on the parameter space.

Assume that we consider q such restrictions. In all generality, we can write the null hypothesis as

$$H_0 : c(\theta^0) = 0$$

with alternative hypothesis

$$H_0 : c(\theta^0) \neq 0$$

where $c(\theta^0)$ is a $(q \times 1)$-vector and $c(\cdot)$ is any linear or nonlinear function. The most common types of restriction in microdata applications are linear. In the simplest case, $q = 1$, and $c(\cdot)$ is a selection function that extracts the l-th element from the parameter vector θ, and we can simply write

$$H_0 : \ \theta_l^0 = 0$$

Another linear restriction involves linear combinations of two or more parameters such that

$$H_0 : \ a + b\,\theta_l^0 + c\,\theta_m^0 = 0$$

For $q > 1$, the framework allows to test joint restrictions such as

$$H_0 : \ \theta_l^0 = 0 \text{ and } \theta_m^0 = 0$$

Finally, an example of a genuine non-linear restriction is

$$H_0 : \ \theta_l^0 / \theta_m^0 = a$$

Whether a particular restriction is interesting or not, cannot be answered in a general way. It depends on the economic context. Also, remember that statistical significance is not to be confused with "economic significance" (see McCloskey, 1985).

Example 3.12. Wage Function

Consider the following example of a log-linear wage function

$$\log(wage) = \beta_0 + \beta_1\, exper + \beta_2\, exper^2 + \beta_3\, educ + u$$

In this model, each of the parameters can take a priori any value in $\mathbb{R}$. Alternatively, we can consider a model where $\beta_3 = 0$ such that education does not affect wages. This is a restriction, and we can test whether it holds. If we reject the restriction, we say that education has a "significant effect" on wages, or simply that "education is statistically significant". Two further examples of restrictions are $\beta_3 = 0.06$, in which case the return to schooling is approximately six percent, and $\beta_1 + 20 * \beta_2 = \beta_3$, in which case the return for one additional year of schooling is identical to the return of the 11th year of labor market experience.

Example 3.13. Weibull Distribution

Consider the following two-parameter density function for a positive continuous random variable y:

$$f(y; \lambda, \alpha) = \lambda\alpha y^{\alpha-1} \exp\left(-\lambda y^{\alpha}\right) \tag{3.17}$$

This is the density function of the **Weibull distribution**. The Weibull distribution has some importance in the analysis of duration data. Depending on the value of α, it can have an increasing ($\alpha > 1$) or decreasing ($\alpha < 1$) hazard rate. Under the restriction that $\alpha = 1$, the Weibull density function reduces to the exponential density with constant hazard rate.

When interpreting test results, it is important to understand that tests of restrictions are asymmetric by construction. If a restriction is rejected, then we know that the data provide strong evidence against it. Only in 5 (or 1, or 10) percent of all cases a restriction is rejected although it is correct. If a restriction is not rejected, then we *cannot* say that the data provide strong evidence for the validity of the restriction. The **power** of the test, i.e., the probability of rejecting the restriction when it is false, can be quite low. But with low power, the probability of not rejecting the restriction when it is false (the "Type-II" error) is high.

Exercise 3.9.

Spencer (1985) modeled the relationship between money demand and the price level in the following way

$$m_t = \theta_0 + \theta_1 Y_t + \theta_2 R_t + \delta P_t + u_t$$

where m_t are real money balances and P_t is the log of the price level. According to some theories, the price-level elasticity of the demand for real money balances should be zero. Spencer (1985) tests the hypothesis $\delta = 0$ using aggregate U.S. data for the years 1952 to 1982. He finds that "...the hypothesis of zero price level elasticity is (...) accepted for each of the two subperiods" and concludes that zero price elasticity "... receives strong support" (p. 490).

- Do you agree with these statements?

ML estimation offers three general methods that can be used to test linear or nonlinear restrictions, namely the **likelihood ratio test**, the **score test**, and the **Wald test**. The basic approach for all three tests is to construct a test statistic, derive its distribution under H_0 (i.e., under the assumption that the restriction is correct), and then compare the observed value of the test statistic for the sample to the distribution under H_0: if what we see is unlikely to occur given this distribution, then we reject the H_0. Here "unlikely" could, for example, mean that values such as the observed test value or greater occur in less than 5 percent of all cases in repeated samples.

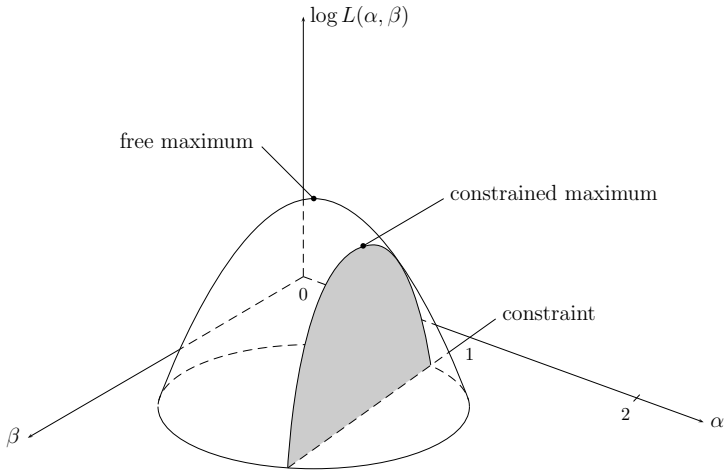
These tests are asymptotic tests. Therefore, the relevant test distribution is either the chi-squared or the standard normal distribution. The validity of the inference depends on the availability of a sample of sufficient size. The tests are asymptotically equivalent, but they differ in their small sample properties; although the analytical forms of the small sample distributions of the test statistics are in general unknown, they can be, and have been on occasion, evaluated through Monte Carlo experimentation. Before we present these tests and their properties, we need to introduce the concept of restricted maximum likelihood estimation.

3.6.2 Restricted Maximum Likelihood

Restrictions can be imposed in the model. We call the ML estimator that fulfills the restriction in the sample the **restricted ML estimator** $\hat{\theta}_r$. In general, $\hat{\theta}_r$ can easily be obtained by directly imposing the restriction and then maximizing the resulting log-likelihood function over the remaining parameters. This is, for example, straightforward whenever the restrictions set

certain parameters to some hypothesized values, such as zero. If a model is estimated by maximum likelihood subject to a restriction, we refer to this estimation procedure as **restricted maximum likelihood**.

Fig. 3.5. *Unrestricted and Restricted Maximization*



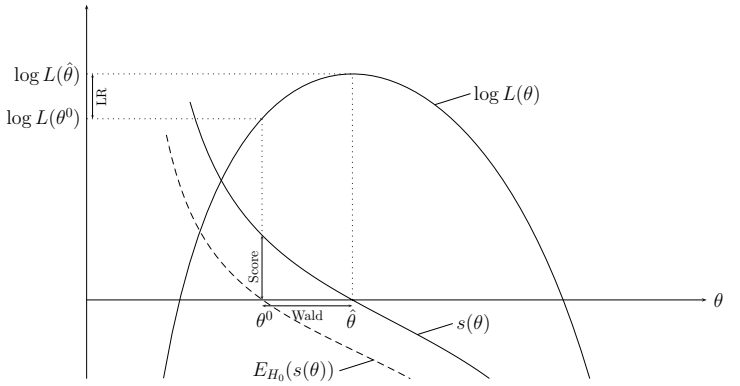
A schematic depiction of restricted maximum likelihood estimation is given in Figure 3.5. Assume that the log-likelihood function has two parameters, $\theta = (\alpha, \beta)'$, and that we consider the restriction $\alpha = 1$. The figure shows both the unrestricted and restricted maxima, and let the corresponding parameter values denote $\hat{\theta}_u = (\hat{\alpha}_u, \hat{\beta}_u)$ and $\hat{\theta}_r = (1, \hat{\beta}_r)$.

The discrepancy between the unrestricted and the restricted maximum can be expressed in three alternative ways, each of which forms the basis for one of the three tests. First, the unrestricted ML estimator $\hat{\alpha}_u$ differs from 1. The discrepancy between the two is the basis for the Wald test. Second, the restricted maximum $\log L(1, \hat{\beta}_r)$ of the log-likelihood function differs from the unrestricted maximum $\log L(\hat{\alpha}_u, \hat{\beta}_u)$. This discrepancy is the basis for the likelihood ratio test. Third and finally, the overall (unrestricted) score function evaluated at parameters of the restricted maximum is not equal to zero. The discrepancy between the score and zero is the basis for the score test.

Figure 3.6 summarizes the essence of the three tests, using the simplest possible example, where the parameter vector θ consists of one element only. Consider testing the hypothesis $\theta = \theta^0$ against the alternative $\theta \neq \theta^0$. This situation is particularly simple, since the restricted model has no additional (unrestricted) parameters that need to be estimated, and we can directly write $\hat{\theta}_r = \theta_0$. As the Figure shows, the Wald test is based on the difference $\hat{\theta} - \theta^0$,

which should be "small" if H_0 is true. If the hypothesis is true, then we know that the expected score at θ^0 is equal to zero. Of course, in any particular sample, the score $s(\theta^0)$ will be above or below its expected value. Then, the question is whether the observed discrepancy $s(\theta^0) - 0$ is "too large" to be explained through random sampling. Finally, the likelihood ratio test is simply twice the distance between $\log L(\hat{\theta})$ and $\log L(\theta^0)$.

Fig. 3.6. *Wald, Likelihood Ratio and Score Test*



All three tests ultimately depend on the distance $\hat{\theta} - \theta^0$ as well as on the curvature of the log likelihood function, $d^2 \log L / d\theta^2$. For a given distance, a greater curvature is associated with a larger change in the log likelihood. In a Wald test, the curvature is inversely related to the variance of the estimator $\hat{\theta}$. Hence, a larger curvature (in absolute value) leads to a smaller variance, meaning that large differences $\hat{\theta} - \theta^0$ are less likely under H_0. In a score test, the curvature is – via the information matrix equality – directly related to the variance of the score $s(\theta^0)$. Hence, a larger curvature leads to a larger variance, meaning that large deviations of $s(\theta)$ from zero become more likely, even if H_0 is true.

3.6.3 Wald Test

[1]

The basic variant of the Wald test (Wald, 1943), the so-called t- or z-test is the most important test in empirical practice for simple restrictions. Starting point of this test is the approximate distribution of the unrestricted ML estimator:

$$\hat{\theta} \overset{\text{app}}{\sim} Normal(\theta, \text{Var}(\hat{\theta})) \tag{3.18}$$

where $\text{Var}(\hat{\theta}) = -\text{E}[H(\theta; y)]^{-1}$. From this distribution, we neither know θ nor $\text{Var}(\hat{\theta})$. The Wald test replaces the unknown $\text{Var}(\hat{\theta})$ by a consistent estimator $\widehat{\text{Var}}(\hat{\theta})$ (three such estimators were proposed before), and the unknown mean θ by a hypothetical value θ^0. We thus obtain "the distribution of $\hat{\theta}$ under the null hypothesis".

In general, $\hat{\theta}$ is a vector of parameters. Denote the l-th element of $\hat{\theta}$ by $\hat{\theta}_l$. It follows from (3.18) that $\hat{\theta}_l$ is approximately univariate normally distributed

$$\hat{\theta}_l \overset{\text{app}}{\sim} Normal(\theta_l, \hat{\sigma}_{ll})$$

where the estimated variance $\hat{\sigma}_{ll}$ is the l-th diagonal element of the covariance matrix. Under the restriction $\theta_l = \theta_l^0$, the approximate sampling distribution of $\hat{\theta}_l$ is fully known. We can compare the evidence in the data – i.e., the observed estimate – with this known sampling distribution to assess whether the restriction appears valid or invalid.

For example, assume that $\theta_l^0 = 0$. It follows that under the null hypothesis $\hat{\theta}_l \sim Normal(0, \hat{\sigma}_{ll})$, and the z-statistic is given by

$$z = \frac{\hat{\theta}_l}{\sqrt{\hat{\sigma}_{ll}}} = \frac{\hat{\theta}_l}{\text{s.e.}} \overset{\text{app}}{\sim} Normal(0, 1)$$

where s.e. is the estimated standard error of $\hat{\theta}_l$. On occasion, this test statistic is referred to as a t-statistic, although it should not be taken literally, since the t-distribution is a small sample distribution and all inference under maximum likelihood rests on asymptotic arguments.

There are two ways to present the results of a hypothesis test, either through p-values or through z-values. The p-value for a two-sided test is defined as $2 \times [1 - \Phi(|z|)]$, where $|z|$ is the absolute value of the test statistic and Φ denotes the cumulative density function of the standard normal distribution. It gives the probability that we observe a value greater than $|z|$ in repeated samples. If the p-value is smaller than a pre-determined level of significance (for example 5%), the restriction (null hypothesis) is rejected. Alternatively, we can compare the z-value to the critical value of the standard normal distribution. For example, the critical value for a two-sided test at the 5% significance level is ± 1.96.

Such z-scores are routinely reported in published research papers and also produced by statistical and econometric software packages. It is important to remember that these packages assume a particular type of restriction, namely that the true parameter is zero. This restriction may not be relevant in some situations. For different restrictions, different z-values need to be computed. In applied econometrics, it is therefore customary to report standard errors rather than z-values.

The methodology easily extends to the case of a single restriction involving more than one parameter. Consider the linear restriction $a\,\theta_1 + b\,\theta_2 = 0$ (which is the same as $\theta_1 = -(b/a)\,\theta_2$). If the restriction is true, it follows that

$$a\,\hat{\theta}_1 + b\,\hat{\theta}_2 \sim Normal(0, a^2\sigma_{11} + 2ab\,\sigma_{12} + b^2\sigma_{22})$$

The z-value to test this restriction is therefore

$$z = \frac{a\hat{\theta}_1 + b\hat{\theta}_2}{\sqrt{a^2\hat{\sigma}_{11} + 2ab\hat{\sigma}_{12} + b^2\hat{\sigma}_{22}}}$$

where $\hat{\sigma}_{12}$ is the estimated covariance of the two parameters.

This is conceptionally straightforward. In practice, computation of the z-value requires knowledge of the estimated covariance matrix of the ML estimates (or at least of its relevant elements). Not all econometric software programs allow easy access to this matrix.

The Wald test is usually introduced in a more general form that allows for multiple restrictions (with a single restriction as a special case). The test statistic is a quadratic form of a restriction vector and a covariance matrix that has a chi-squared distribution when the restriction is valid. We do not present this test here, because it is not used very frequently in practice, since alternative tests are easier to compute. For single restrictions, the general Wald test is equivalent to the simple z-tests and offers no extra benefit. For multiple restrictions, it is much easier to use the likelihood ratio test. This test is discussed next.

3.6.4 Likelihood Ratio Test

As we have just seen, it is possible to impose the restriction and estimate the remaining parameters of the model by restricted maximum likelihood. Denote the value of the log-likelihood function at the restricted maximum by $\log L(\hat{\theta}_r)$, where the subscript "r" reminds us that the restriction has been imposed. Denote the value of the log-likelihood function at the unrestricted maximum by $\log L(\hat{\theta}_u)$, where the subscript "u" stands for "unrestricted". We know that $\log L(\hat{\theta}_r) \leq \log L(\hat{\theta}_u)$. Clearly, the restricted maximum is smaller than the unrestricted one (recall Figure 3.5), and this must be so for almost all restrictions. Only if the restriction is true in the sample (i.e., if the unrestricted estimator happens to satisfy the restriction and $\hat{\theta}_u = \hat{\theta}_r$), then will restricted and unrestricted maxima be the same.

The drop in the log-likelihood associated with imposing restrictions is an indicator of the disagreement between the restrictions and the data. If the drop is sufficiently large, the restrictions are rejected. Again, let q denote the number of restrictions. The critical value for rejection comes from a χ^2_q distribution, since it can be shown that if the restrictions are true

$$LRT = 2(\log L(\hat{\theta}_u) - \log L(\hat{\theta}_r)) \sim \chi^2_q$$

with q degrees of freedom. The test is referred to as a **likelihood ratio** test, although in fact, it involves the computation of a *difference* of two *log*-likelihood values. The main advantage of this test is that it is easy to perform.

It always works when a restricted and an unrestricted version of a model can be compared. The disadvantage is that two models need to be computed separately. This might have been an obstacle at times when computing was expensive, but it should no longer be nowadays.

Example 3.14. The Probability of Remaining Childless

One of our example datasets concerns the fertility choices of women in the U.S. (see Chapter 1.5.1). The dataset consists of 5,150 women aged 40 or older in the years 1974 to 2002. Suppose we are interested in factors that influence the probability of remaining childless. For example, we may ask a question such as "Are more educated women (those with a higher number of formal years of schooling) more likely to remain childless than less educated ones?" It is easy to find economic models that predict such a connection. In addition, social norms and individual values certainly influence the fertility decision. Hence, race might have an effect, and also the number of siblings of the woman herself. A time trend captures all other influences that have changed over time but are not explicitly controlled for in the model.

Table 3.1. *Probit Estimates of Fertility Decision*

	Model 1	Model 2	Model 3
constant	-1.060	-1.166	-1.503
	(0.022)	(0.047)	(0.116)
linear time trend		0.006	0.003
		(0.002)	(0.003)
years of education			0.031
			(0.007)
white			0.063
			(0.063)
number of siblings			-0.012
			(0.007)
Log-likelihood value	-2,126.9	-2,123.6	-2,107.1
Observations	5,150	5,150	5,150

Notes: Standard errors in parentheses.

In Table 3.1, the ML estimates from three different probit models are shown. For now, take the probit model that can be chosen if the outcome variable is binary (childless, yes/no) – see Chapter 4, where we provide more details on this model and its interpretation. The most basic one, Model 1, includes a constant only. In Model 2, a linear time trend is included. In Model 3, three additional explanatory variables were added, namely the years of education, the race of the woman (white yes/no), and the number of siblings.

In the terminology of hypothesis testing, the three models are **nested** since Model 1 is a restricted version of Model 2 (if we set the trend coefficient

to zero), and Model 2 is a restricted version of Model 3 (if we set the three coefficients of the additional variables to zero). As it must be the case, the log-likelihood value increases as we move towards the least restrictive model. The likelihood ratio test rejects the restriction if the increase of the test statistic is unlikely to occur under H_0: the restriction is true. We have:

- Model 1 against Model 2: LRT = 6.6. The 5% critical value from a chi-squared distribution with one degree of freedom is 3.8; therefore, the restriction is rejected.
- Model 2 against Model 3: LRT = 33.0. The 5% critical value from a chi-squared distribution with three degrees of freedom is 7.8; therefore, the restriction is rejected.

If we want to test the statistical significance of single parameters, we can use the z-test. For example, in Model 2, the z-value of the trend coefficient is $z = 0.006/0.002 = 3$, which is greater than the 5% critical value of 1.96 from a standard normal distribution for a two-sided test. This test is, at the same time, a test of Model 1, and the Wald and LR tests thus come to the same conclusion. If we consider Model 3, and ask whether the number of siblings has a significant effect on a woman's probability of remaining childless, we obtain a z-value of -1.57. Hence, we cannot reject the null hypothesis of no effect, *ceteris paribus*.

Example 3.15. The Normal Linear Model

We have seen in Section 3.4 that the ML estimator for β in the normal linear model coincides with the least squares estimator. In standard regression outputs, it is rarely the case that log-likelihood values or likelihood ratio test statistics are reported alongside the usual Wald test statistics, such as an F-test for overall significance or $z-$test statistics for single parameters. However, likelihood ratio test statistics can be derived directly from equation (3.11), the log-likelihood function of the normal linear model, which we restate here for convenience:

$$\log L(\beta, \sigma^2; y, x) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i'\beta)^2$$

In order to evaluate this expression at the ML estimators $\hat{\beta}$ and $\hat{\sigma}^2$, we notice that

$$\sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2 = n\hat{\sigma}^2$$

Therefore

$$\log L(\hat{\beta}, \hat{\sigma}^2; y, x) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\hat{\sigma}^2 - \frac{n}{2}$$

Moreover, let $\hat{\sigma}_u^2$ denote the estimated residual variance of the unrestricted model, and $\hat{\sigma}_r^2$ denote the estimated residual variance of the restricted model, and SSR the sum of squared residuals. The likelihood ratio test of the restricted model against the unrestricted model then has a relatively simple form:

$$\begin{aligned} LRT &= 2\left(-\frac{n}{2}\log\hat{\sigma}_u^2 - \left(-\frac{n}{2}\log\hat{\sigma}_r^2\right)\right) \\ &= n(\log\hat{\sigma}_r^2 - \log\hat{\sigma}_u^2) \\ &= n(\log SSR_r - \log SSR_u) \end{aligned}$$

In the special case where the restricted model is the one with all regressors excluded, we know that $SSR_r = SST$, the total sum of squares or total variation in the dependent variable, and therefore the likelihood ratio test statistic can be written as a function of the coefficient of determination R^2 as

$$LRT = -n\log(1 - R^2)$$

3.6.5 Score Test

A third, asymptotically equivalent test for a single restriction or a number of restrictions is the **score test** (Rao, 1948). Recall from Chapter 3.2.1 that the expectation of the score, evaluated at the true parameter, equals zero:

$$\mathrm{E}[s(\theta; y)] = 0$$

The score test is based on the limiting distribution of

$$\frac{1}{\sqrt{n}}s(\hat{\theta}_r; y) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} s(\hat{\theta}_r; y_i) \tag{3.19}$$

under H_0. This is the score with respect to the entire vector θ, but we evaluate it at the restricted estimates $\hat{\theta}_r$. If we would evaluate it at $\hat{\theta}_u$ instead, then (3.19) would be identical to zero, because this is just the first-order condition of the ML estimator. If the restrictions imposed by the null hypothesis are true, then (3.19) should not be statistically different from zero.

Evaluated at the true parameter vector, we know that the limiting distribution of the score is a normal distribution (see also equation (3.10)) with expectation zero and variance $\mathrm{Var}[s(\theta; y)]/n = -E[H(\theta; y)]/n$. Thus, under the null hypothesis H_0

$$ST = \left(\frac{s(\hat{\theta}_r; y)}{\sqrt{n}}\right)' \left(\frac{\text{Var}[s(\hat{\theta}_r; y)]}{n}\right)^{-1} \left(\frac{s(\hat{\theta}_r; y)}{\sqrt{n}}\right)$$

has a chi-squared distribution with q degrees of freedom. The variance $\text{Var}[s(\hat{\theta}_r; y)]$ can be estimated consistently by any of the three methods discussed in Chapter 3.3.5, i.e., the outer product of the gradient, minus the actual Hessian matrix, or minus the expected Hessian matrix, all evaluated at the restricted ML estimator $\hat{\theta}_r$. The small sample properties of the estimator may depend on which of the three matrices is used. In contrast to the other two test methods, the score test requires only the estimation of the restricted model, in order to obtain the restricted parameter estimates $\hat{\theta}_r$. This may be an advantage. On the downside, the score test requires a considerable amount of algebra that is specific to the unrestricted model under consideration.

Example 3.16. Score Test in the Poisson Model

As an example of a score test, consider sampling from a Poisson distribution with parameter λ, and assume a null hypothesis $\lambda_r = 1$. The score of the unrestricted model was derived earlier, and is given by

$$s(\lambda; y) = \sum_{i=1}^{n} \left(-1 + \frac{y_i}{\lambda}\right)$$

Moreover, the variance can be estimated by the sum of the squared score contributions:

$$\widehat{\text{Var}}[s(\lambda; y)] = \sum_{i=1}^{n} s(\lambda; y_i)^2$$

Therefore, the score test statistic for the hypothesis $\lambda_r = 1$ is given by

$$ST = \left.\frac{\left[\sum_{i=1}^{n}\left(-1 + \frac{y_i}{\lambda}\right)\right]^2}{\sum_{i=1}^{n}\left(-1 + \frac{y_i}{\lambda}\right)^2}\right|_{\lambda=1} = \frac{n(\bar{y} - 1)^2}{n^{-1}\sum_{i=1}^{n}(y_i - 1)^2}$$

Exercise 3.10.

- Show that for large samples the score statistic converges to the square of the Wald z-statistic.

3.6.6 Model Selection

The Wald, likelihood-ratio and score tests compare two models, of which one is a restricted version of the other. We say that the two models are **nested**. The restrictions we mean here are those on the possible range of values that parameters might take. Consider a comparison of the two models

$$\text{Model 1} : y_i|x_{i1} \sim Normal(\beta_0 + \beta_1 x_{i1}, \sigma^2)$$

$$\text{Model 2} : y_i|x_{i2} \sim Normal(\beta_0 + \beta_2 x_{i2}, \sigma^2)$$

We call such two models **nonnested**, because none of them can be obtained as a restricted version of the other. If we are forced nevertheless to decide which of the two models, Model 1 or Model 2, is the preferable one – an instance of so-called **model selection** – we cannot use any of the three tests discussed previously. We rather have to look for a different approach. A similar situation would arise for the following two models:

$$\text{Model 1'} : y_i|x_{i1} \sim Normal(\beta_0 + \beta_1 x_{i1}, \sigma^2)$$

$$\text{Model 2'} : y_i|x_{i1} \sim Normal(\exp(\beta_0 + \beta_1 x_{i1}), \sigma^2)$$

Again, the two models are nonnested. There is no parametric restriction that would transform (reduce) one model to the other.

A simple rule to select between two nonnested model is to estimate both models by maximum likelihood and choose the one with the higher value of the log-likelihood. This works well in the above examples, where the number of parameters is the same in the two models. If the number of parameters differs, this should be taken into account. Specifically, one can increase the log-likelihood of any model arbitrarily by adding more and more parameters. When comparing two models, the model with the larger number of parameters should accordingly be penalized for this relative abundance.

There are two common penalty functions. In the so-called **Akaike information criterion**, we choose the model that minimizes

$$AIC = -2 \log L(\hat{\theta}) + 2p$$

whereas in the **Schwarz information criterion**, we choose the model that minimizes

$$SIC = -2 \log L(\hat{\theta}) + p \log n$$

Here, p denotes the number of parameters and n is the number of observations. Clearly, if the number of parameters is the same in the two models, as in our two examples above, the penalty function does not matter and the choice between the two models indeed simply comes down to choosing the model with the higher log-likelihood function.

3.6.7 Goodness-of-Fit

Finally, we briefly consider the issue of goodness-of-fit in maximum likelihood estimation. In the standard linear models, the goodness of fit is usually assessed by the R^2, the proportion of the total variation in the dependent variable that is explained by the model.

In many non-linear microdata models, the underlying variance decomposition does not work and a standard R^2 measure is not available. One possible substitute is a log-likelihood comparison of the full model (with all regressors) and a constant-only model. Define

$$R^2_{pseudo} = 1 - \frac{\log L(\hat{\theta}_u)}{\log L(\hat{\theta}_r)}$$

Clearly, $\log L(\hat{\theta}_u) \geq \log L(\hat{\theta}_r)$. In the case of discrete data and regular restrictions, we have $\log L(\hat{\theta}_r) < 0$ and $\log L(\hat{\theta}_u) \leq 0$ such that $0 \leq R^2_{pseudo} \leq 1$. In fact, the likelihood-based pseudo R^2 was proposed by McFadden (1974a) in the context of conditional logit models for multinomial variables (see Chapter 5 for further details on this kind of models). It is important to note that the pseudo R^2 measure does not have an interpretation as the R^2 in the standard linear model. For the pseudo R^2, a value near one just indicates a better model fit than a value near zero.

For continuous data, the pseudo R^2 measure may not lie within the unit interval, since the value of the log-likelihood function can be positive or negative. Moreover, a larger value of the pseudo R^2 measure does not necessarily indicate a better fit.

From a practical point of view, apart from point estimates and standard errors, one should always report the value of the log-likelihood function and a test statistic, for example from a LR test, of the full model against the constant-only model. These statistics can be used to evaluate formally the statistical significance of regressors, which is in general more important than assessing the goodness-of-fit.

3.7 Pros and Cons of Maximum Likelihood

Maximum likelihood is not the only method for estimating microdata models. Alternative methods include

- ordinary least squares
- non-linear least squares
- (generalized) method of moments

The first method is taught in all introductory econometrics courses, the second and third methods, while also important, are not yet standard. One justification for concentration on maximum likelihood is its wide applicability and –

with the exception of OLS – it is the most frequently used method for micro-data applications in practice. Still, one needs to be aware that ML estimation has its limitations.

Disadvantages

- Specific assumptions on a parametric probability distribution are needed.
- Often, theory provides little guidance on the functional form and probability model, and one has to make somewhat arbitrary assumptions.
- In general, the ML estimator is not robust with respect to misspecification.
- The ML estimator does not always exist, since the likelihood function may become unbounded.

Advantages

- Convenience: all that one needs to do is to write down the likelihood function. First and second derivatives can be obtained numerically.
- The computer produces a numerical answer (as long as the problem is well defined, parameters are identified, etc.)
- There is a well established large sample theory (asymptotic normality, consistency, and efficiency)
- The invariance property generates flexibility in reformulating the model.
- The inference is simple.
- The formulation of the likelihood function forces one to carefully think about the problem, the way the sample was generated, etc.

3.8 Further Exercises

Exercise 3.11 Assume that you have a sample of n independent observations from a Poisson distribution with density function

$$f(y_i;\lambda) = \frac{\exp(-\lambda)\,\theta^{y_i}}{y_i!} \qquad y_i = 0, 1, 2, \ldots$$

where $\mathrm{E}(y_i) = \mathrm{Var}(y_i) = \lambda$.

a) Write down the log-likelihood function and derive the score.

b) Find the ML estimator of λ.

c) Show that the expectation of the score is zero.

d) Derive the Hessian and the information matrix.

e) Does the information equality hold?

f) Derive the variance of the ML estimator. Does the ML estimator reach the Cramér-Rao lower bound? Does this result hold in general?

Exercise 3.12 Suppose that $y_i|x_i$ is Poisson distributed with parameter λ_i in a sample of n independent observations. Furthermore, assume that λ_i is specified as $\lambda_i = \alpha + \beta x_i$.

a) Does this specification make sense?

b) Show that the first-order conditions of the ML estimators of α and β can be written as

$$\sum_{i=1}^{n} \frac{y_i - \mathrm{E}(y_i|x_i)}{\mathrm{Var}(y_i|x_i)} = 0$$

$$\sum_{i=1}^{n} \frac{[y_i - \mathrm{E}(y_i|x_i)]x_i}{\mathrm{Var}(y_i|x_i)} = 0$$

Exercise 3.13 Suppose you have sample of n independent observations from an exponential distribution with density function

$$f(y_i;\lambda) = \lambda e^{-\lambda y_i} \qquad y_i \geq 0,\ \lambda > 0$$

where $\mathrm{E}(y_i) = 1/\lambda$ and $\mathrm{Var}(y_i) = 1/\lambda^2$.

a) Find the ML estimator of λ. Is this estimator consistent? What is the asymptotic distribution of $\hat{\lambda}$?

b) Now, assume that you observe explanatory variables as well, and that the conditional density of $y_i|x_i$ is of exponential form with parameter $\lambda_i = 1/(\alpha + \beta x_i)$. Does this specification make sense?

c) Show that the first-order conditions of the ML estimators of α and β can be written as

$$\sum_{i=1}^{n} \frac{y_i - \mathrm{E}(y_i|x_i)}{\mathrm{Var}(y_i|x_i)} = 0$$

$$\sum_{i=1}^{n} \frac{[y_i - \mathrm{E}(y_i|x_i)]x_i}{\mathrm{Var}(y_i|x_i)} = 0$$

Exercise 3.14 Consider ML estimation of the parameter α of the **Pareto distribution** with density function

$$f(y_i;\alpha) = \alpha/y_i^{\alpha+1} \qquad \alpha > 0,\ y > 0$$

a) Find the ML estimator of α.

b) Find the LR, Wald and Score test statistics for $H_0 : \alpha = \alpha^0$ against $H_1 : \alpha \neq \alpha^0$.

Exercise 3.15 Assume that you have a sample of n independent observations from a **geometric distribution** with density function

$$f(y_i;\theta) = \theta(1-\theta)^{y_i} \qquad 0 < \theta < 1,\ y_i = 0, 1, 2, \ldots$$

a) Write down the log-likelihood function.

b) Find the ML estimator of θ.

c) Calculate the Fisher information matrix of the sample.

d) What is the asymptotic variance of the ML estimator?

Exercise 3.16 Assume that you have a sample of n independent observations from a distribution with density function

$$f(y_i;\theta) = \theta y_i^{\theta-1} \qquad \theta > 0,\ 0 < y_i < 1$$

a) Find the ML estimator of θ.

b) Find the asymptotic distribution of the ML estimator.

Exercise 3.17 Consider a simple linear regression model without constant and heteroscedastic error terms. Furthermore, suppose you have normally distributed errors and a sample of n independent pairs of observations (y_i, x_i) with conditional density function for each observation given by

$$f(y_i|x_i;\beta,\sigma^2) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2 x_i^2}} \exp\left[-\frac{1}{2}\left(\frac{y_i - \beta x_i}{\sqrt{\sigma^2 x_i^2}}\right)^2\right]$$

a) Derive the log-likelihood function and the score.

b) Find the ML estimators of β and σ^2.

c) What is the asymptotic distribution of $(\hat{\beta}'\ \hat{\sigma}^2)'$?

d) Compare your results with the GLS estimator. What do you conclude?

Exercise 3.18 Suppose a random sample of size $n = 100$ has been drawn from a Poisson distribution, with sample mean equal to $\bar{y} = 2.002$.

a) What is the distribution of the ML estimator $\hat{\lambda} = \bar{y}$?

b) Calculate a 95% confidence interval for λ.

c) Determine the ML estimator for $P(Y = 1)$.

d) Test $H_0:\ P(Y = 1) = 0.30$ against $H_1:\ P(Y = 1) \neq 0.30$.

Exercise 3.19 In the literature on the economics of happiness, one often finds statements such as "happiness is u-shaped in age". What does this statement mean? What type of regression specification is needed to arrive at such a conclusion?

Exercise 3.20 Depending on your answer to the previous exercise, can you give a different regression specification that would allow for a u-shaped age effect as well?

Exercise 3.21 Consider the model

$$\log y_i | x_i \sim Normal(x_i'\beta, \sigma^2)$$

where $\log y_i$ denotes log-earnings, and the linear index is given by $x_i'\beta = \beta_0 + \beta_1\, schooling_i + \beta_2\, age_i + \beta_3\, age_i^2$. How can you test whether age has no influence on log-earnings?

Exercise 3.22 The following earnings equation has been estimated by ML assuming normally distributed error terms:
(standard errors in parentheses)

$$\widehat{\log y_i} = \underset{}{8.5633} + \underset{(0.0409)}{0.3733}\, voc_i + \underset{(0.0510)}{0.6658}\, uni_i + \underset{(0.0072)}{0.1023}\, age_i - \underset{(0.0001)}{0.0010}\, age_i^2$$
$$\qquad\qquad (0.1421)$$

a) Determine the age at which earnings are maximized.

b) What (if anything) can you say about the standard error of this age?

4

Binary Response Models

4.1 Introduction

There can be little doubt that the logit and probit models discussed in this chapter are used more frequently in empirical practice than any other model for discrete or limited dependent variables. The simple reason is that binary dependent variables – outcomes that can be characterized as being either true or false, either one or zero – occur very frequently in real-life measurements.

Consider the following events: a person is unemployed, smokes, is childless, has visited a doctor during the last quarter, has had an extramarital affair, has been granted a bank loan, is willing to pay for a public project, and has been part of a treatment group in an experiment. The list could be continued almost indefinitely. Each of these events is, by its very definition, binary, i.e., it either applies or it does not apply. While microeconomic theory maintains mostly the fiction of continuous choices within convex budget sets, reality suggests otherwise. Most choices people make are necessarily discrete. One either buys a car, or one does not. It is not possible to buy a fraction of a car.

Binary response models can also be useful when the original dependent variable is multinomial, ordered, or even continuous. For example, assume that the variable *employment status* is coded with four possibilities: full-time employed (1), part-time employed (2), unemployed (3), or not in the labor force (4). Depending on the goal of the analysis, one may want to analyze, for example, the distinction between being employed (1+2) and not being employed (3+4), or the choice between full-time and part-time employment, conditional on employment.

As an example of an ordered variable, consider a survey question on individual happiness where responses are integer values between zero ("completely unhappy") and ten ("completely happy"). From this information, one can obtain a binary variable *happy/unhappy* by splitting the 0-10 scale into two parts, for example at the value of five. Finally, continuous variables can be transformed into, and therefore modeled as, binary events as well. An example

arises in business cycle research, where events such as "downturn" or "recession" are based on an underlying continuous variable, such as gross domestic product, or its rate of change. In such applications, where some dimension of variation in the dependent variable is ignored, one cannot expect binary models to be efficient. However, they may provide good first answers, the estimators may be robust, and they can be obtained relatively easily, as we will show in the following.

The remainder of this chapter is organized as follows. In Section 4.2, we start with a presentation of the models, discussing the pros and cons of the various specifications of the conditional probability functions. From the practitioner's point of view, a good understanding of the model assumptions is necessary, particularly with regard to a correct interpretation of parameter estimates. In general, the interpretation of parameters is not as straightforward as in the linear model, and we devote a separate part of this section to this issue. In Section 4.3, we show how the probit and logit models arise in models of discrete choice, where consumers are characterized by random utility functions.

In Section 4.4, we discuss estimation of the models by maximum likelihood, building on the general principles derived in the previous chapter. Of course, we do not expect anyone to actually write a program to estimate these models, as such programs are readily available in standard software packages. Yet some basic understanding of the estimation techniques is useful and can give insight into why estimation sometimes fails ("perfect prediction"). Finally, we consider various goodness-of-fit measures (Section 4.5) and estimation of binary response models when the sample is not random (Section 4.6).

4.2 Models for Binary Response Variables

4.2.1 General Framework

Binary response variables have a **Bernoulli probability function**

$$f(y_i|x_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i} \qquad y_i = 0, 1 \tag{4.1}$$

where $\pi_i = \pi(x_i)$ is shorthand notation for $P(y_i = 1|x_i)$, the conditional probability of observing outcome one, given the regressors. Note that $\mathrm{E}(y_i|x_i) = 0 \times (1-\pi_i) + 1 \times \pi_i = \pi_i$, so that the conditional expectation is equal to the conditional probability of observing outcome one. Binary response models can be distinguished by the way they parameterize π_i in terms of x_i. The standard models used in practice have in common that π_i is derived as a monotonic transformation of a linear index function

$$\pi_i = G(x_i'\beta) \tag{4.2}$$

where $x_i'\beta = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$ is defined as before. Since π_i denotes a probability, we usually require that $0 \leq G(x_i'\beta) \leq 1$. The fact that $G^{-1}(\pi_i)$ is linear is less restrictive than it might appear at first sight. Since the linearity assumption relates to the parameters and not to the explanatory variables x_i, the latter may include logarithms, polynomials, and interactions. In fact, any continuous function can be approximated to any desired degree of accuracy by such a linear-in-parameters specification. One consequence of the linear index assumption is that the partial derivative of π_i with respect to the l-th element in x_i, denoted by x_{il}, has the form

$$\frac{\partial \pi_i}{\partial x_{il}} = g(x_i'\beta)\beta_l \tag{4.3}$$

where $g(x_i'\beta) = dG(x_i'\beta)/d(x_i'\beta)$. From (4.3) it follows that partial derivatives are not constant in general, but rather depend on the specific values of x_i. We have already mentioned that $G(\cdot)$ needs to map a real number onto the unit interval. This is a matter of logical consistency when a probability is to be modeled. Two such transformations are the cumulative distribution function of the standard normal distribution (the **probit model**), and the cumulative distribution function of the logistic distribution (the **logit model**). If $G(\cdot)$ is the identity function, we obtain the **linear probability model**. This model violates the requirement for logical consistency. Yet we include it in our discussion, because it is commonly used in practice.

4.2.2 Linear Probability Model

In the **linear probability model (LPM)** we use the linear regression model to explain a binary outcome, which corresponds to the assumption that the response probabilities are entirely determined by the linear index $x_i'\beta$ and the transformation function is simply given by $G(x_i'\beta) = x_i'\beta$. This specifies

$$\pi_i = x_i'\beta = \mathrm{E}(y_i|x_i) \tag{4.4}$$

which is (almost) a standard linear regression model. It is special since the dependent variable y_i can take only two values, zero and one. Hence, the implicit regression error $u_i = y_i - x_i'\beta$ can take only two values, $0 - x_i'\beta$ and $1 - x_i'\beta$. It follows that

$$\begin{aligned}\mathrm{Var}(u_i|x_i) &= (x_i'\beta)^2 P(y_i = 0|x_i) + (1 - x_i'\beta)^2 P(y_i = 1|x_i) \\ &= (x_i'\beta)^2(1 - x_i'\beta) + (1 - x_i'\beta)^2 x_i'\beta \\ &= x_i'\beta(1 - x_i'\beta)\end{aligned}$$

Thus, the error term is heteroscedastic, which would need to be accounted for in estimation (by least squares) in order to obtain an efficient estimator and valid inference, either by using **generalized least squares (GLS)**, or by computing a heteroscedasticity consistent covariance matrix.

Exercise 4.1.

- How would you transform the model $y_i = x_i'\beta + u_i$ with $y_i \in \{0, 1\}$ in order to perform GLS estimation?
- What does the heteroscedasticity robust covariance matrix look like?

The LPM has a number of advantages. We analyze the binary outcomes within the familiar linear regression framework with heteroscedastic error terms, and therefore the model is easy to estimate. Parameters can be interpreted directly as marginal effects, and the approximation is good as long as we do not move too far away from the means of the explanatory variables. The estimator does not suffer from the non-existence problem possibly encountered in the probit and logit models (see Section 4.4.2). On the downside, the model does not properly restrict the range of $G(x_i'\beta) = x_i'\beta$ to the unit interval, as it should, because it is supposed to be a probability. Therefore, nonsensical predictions outside the (0,1) interval are possible if extreme values of x_i are considered.

Example 4.1. A Linear Probability Model of Being Childless

Consider again the question of whether more educated women are more likely to remain childless than less educated women (see Example 3.14). We analyze this question using our data from the GSS waves 1974 to 2002 (four-year intervals). The dependent variable *childless* is a binary outcome equal to one if the woman has no children, and equal to zero otherwise. The variable of main interest is the highest year of schooling (*educ*). Additionally, we include a linear *time* trend, and control for race (a dummy variable *white*) and the number of siblings (*sibs*). Least squares estimation yields the following output

$$\begin{array}{cccccc} \widehat{childless} = & 0.0397 & + 0.0006\, time & + 0.0076\, educ & + 0.0142\, white & - 0.0024\, sibs \\ & (0.0290) & (0.0005) & (0.0019) & (0.0132) & (0.0015) \end{array}$$

$$n = 5,150,\ R^2 = 0.0079$$

with heteroscedasticity robust standard errors in parentheses. From the estimates of the LPM we can see the effect of education on the probability of being childless directly: one more year of schooling increases, *ceteris paribus*, the probability of being childless by 0.76 **percentage points**. Now assume we want to predict the probability of being childless. Using the above estimates, we predict a probability of about 21.0 percent for a white woman surveyed in 1994 with twenty years of schooling and three siblings. However, if we use extreme values of the regressors, for example $t = 0$, $educ = 0$, $white = 0$, and $sibs = 23$, then we predict a probability of -1.4 percent, which is clearly nonsensical.

Exercise 4.2.

- How would you interpret the coefficient on *white*? Answer precisely.
- What is the predicted probability that a white woman with 12 years of schooling and one sister (recorded in 1978) will be childless?
- Compare to a woman with the same characteristics as before but with three sisters.
- Estimate $\mathrm{Var}(u_i|x_i)$ for both women. What do you conclude?

4.2.3 Probit Model

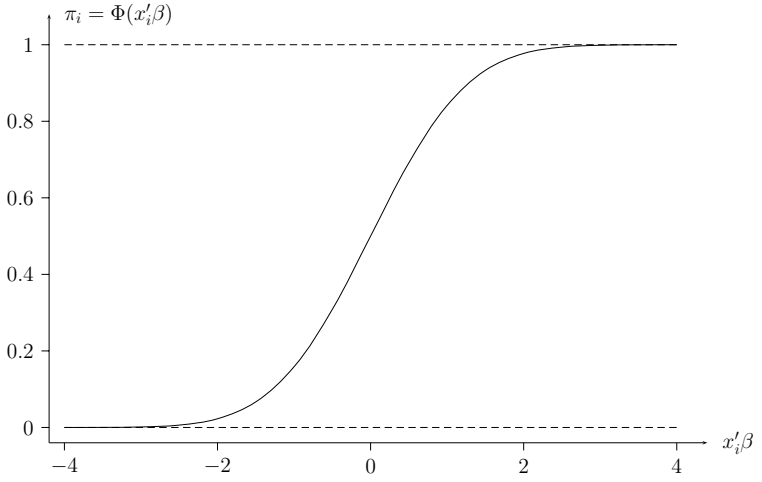
The **probit model** has a long history in the modeling of binary outcomes; early references can already be found in the 19th century. An overview is given, for example, in Finney (1971). In the probit model we assume that

$$\pi_i = G(x_i'\beta) = \Phi(x_i'\beta) = \int_{-\infty}^{x_i'\beta} \frac{1}{\sqrt{2\pi}} \exp[-(t^2/2)]\mathrm{d}t \tag{4.5}$$

where Φ is the cumulative density function of the *standard normal distribution*. Φ takes the real numbers as an argument, i.e., imposing no restriction on the linear index $x_i'\beta$, and maps them onto the unit interval (0,1), as required. A graphical illustration of the probability function in the probit model is given in Figure 4.1. In order to write the probit model in the form of a (conditional) probability model, we simply plug (4.5) into (4.1) and obtain

$$f(y_i|x_i) = [\Phi(x_i'\beta)]^{y_i}[1 - \Phi(x_i'\beta)]^{1-y_i} \qquad y_i = 0, 1$$

Fig. 4.1. *Probability Function in the Probit Model*



Exercise 4.3.

- Which part of $\Phi(x_i'\beta)$ is convex and which part is concave?

An interesting motivation of the probit model, alternatively to the mechanical approach above, starts with a linear model for a **latent** (unobserved) continuous variable y_i^* which is related to the linear index function and an additive error term u_i, formally

$$y_i^* = x_i'\beta + u_i \tag{4.6}$$

Latent variables provide a powerful tool in microeconometrics. This is our first encounter with this method; more applications will follow later on. If we were able to observe y_i^*, this would be a normal linear regression model. Provided that the Gauss Markov assumptions apply, the OLS estimator for β would be best, linear and unbiased. However, y_i^* is not observed, so OLS estimation is not an option. The next step is to specify how the latent model relates to the observed outcome. In the case of the probit model, we assume that

$$y_i = \begin{cases} 1 \text{ if } y_i^* \geq 0 \\ 0 \text{ if } y_i^* < 0 \end{cases}$$

or, more compactly,

$$y_i = I(y_i^* \geq 0) \tag{4.7}$$

where I is an indicator function that returns 1 if the argument is true and 0 if the argument is false. It follows that

$$\pi_i = P(y_i^* \geq 0|x_i) = P(x_i'\beta + u_i \geq 0|x_i) = P(u_i \geq -x_i'\beta|x_i)$$

At this stage, we need to make an assumption with respect to the distribution of u_i, so that we can derive the corresponding conditional probability function for y_i. If the u_i's are independently and normally distributed with mean 0 and variance σ^2, then we can proceed with

$$P\left(\frac{u_i}{\sigma} \geq -\frac{x_i'\beta}{\sigma}\Big|x_i\right) = 1 - \Phi\left(-\frac{x_i'\beta}{\sigma}\right) = \Phi\left(\frac{x_i'\beta}{\sigma}\right)$$

The last equality follows from symmetry of the standard normal distribution. We see that the probability depends on two parameters, β and σ. Unfortunately, this is one too many. Only the ratio β/σ is identified, but not the single parameters β and σ. For instance, if β and σ each are multiplied by a constant c, then the probability π_i remains unchanged. Hence, there are infinitely many parameter combinations that give the same probability. They are *observationally equivalent* and no amount of data can distinguish them from each other. Therefore, we need a normalization. Typically, we let $\sigma = 1$, such that $\pi_i = \Phi(x_i'\beta)$, which corresponds to the probability we stated earlier in (4.5).

At this point, it is important to note that the latent model is *just a tool* to derive the conditional probability model. Since y_i^* is not observed, its interpretation is not of interest, in general. Instead, interpretation should focus on what we originally intended to model, namely the conditional probabilities of observing the binary outcomes.

4.2.4 Logit Model

An early treatment of the **logit model** can be found in Berkson (1944) who considered this model in the context of estimating the effect of a continuous treatment (injection of varying amounts of a poisonous substance) on a binary outcome (death or survival) by the subject. The conditional probability function of the logit model is given by

$$\pi_i = G(x_i'\beta) = \frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)} = \Lambda(x_i'\beta) \tag{4.8}$$

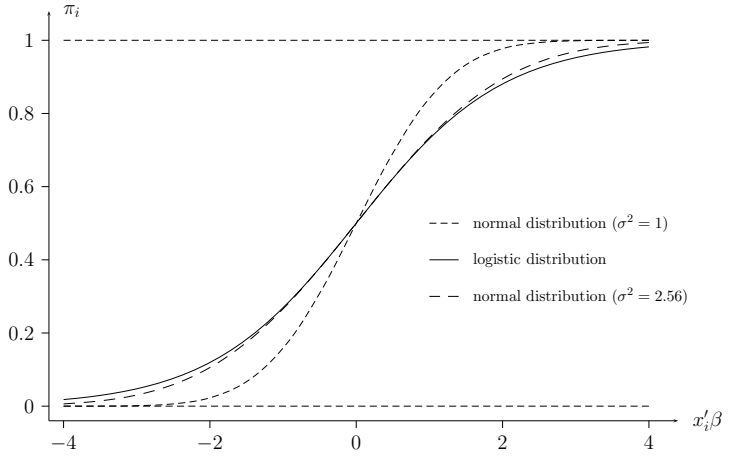
We will use the symbol Λ to denote the cumulative density function of the (standard) *logistic distribution*. In contrast to the probit model, choice probabilities are available in closed-form, i.e., we do not have to integrate in order to obtain the probability.

A repeatedly needed expression is the first derivative of Λ, i.e., the density function of the logistic distribution. It is given by

$$\frac{d\Lambda(z)}{dz} = \frac{\exp(z)}{1+\exp(z)} - \frac{\exp(z)}{[1+\exp(z)]^2}\exp(z) = \Lambda(z)[1-\Lambda(z)]$$

Despite the substantial difference in functional form, logit and probit turn out to be very similar, and the choice between the two makes little difference in practice. An illustration of the similarity is provided by Figure 4.2.

Fig. 4.2. *Comparison of Probit and Logit Model*



We can see that the logistic distribution function is flatter than the distribution function of the standard normal. For example, the slope in the center of the distribution (at 0) is $\Lambda(0)(1-\Lambda(0)) = 0.25$, compared to $\phi(0) \approx 0.4$. Moreover, the logistic distribution has a larger variance ($\pi^2/3$) than the standard normal distribution.

To make the two distributions comparable, some rescaling is required. In particular, we can consider normal distributions with non-unit variances σ^2 and ask what values of σ^2 would make the resulting probit model most similar to the logit model. A first thought would be to let $\sigma^2 = \pi^2/3$, since this equalizes their variances. Alternatively, we can choose σ^2 such that the slopes of the two distribution functions in the middle (at the value $x_i'\beta = 0$) are equalized. The equal-slopes condition can be written as

$$\Lambda(0)[1 - \Lambda(0)] = \frac{1}{\sigma}\phi(0)$$

from which it follows directly that $\sigma = 0.4/0.25 = 1.6$, and hence $\sigma^2 = 2.56$. Figure 4.2 additionally plots the cdf of a normal distribution with variance 2.56, which looks very much like the logistic distribution function. As we will see later, the factor 1.6 can be used to *approximate* the parameter estimates in the logit model, given that we have estimated a probit, by simply multiplying the probit coefficients by 1.6.

Exercise 4.4.

- Why do we have to multiply the coefficients by 1.6 rather than divide them by this factor?

As for the probit model, we can offer a derivation of the logit model based on the latent variable approach. Again, let y_i^* denote the latent (unobserved) continuous variable as specified in (4.6) with observation rule (4.7). Now assume that u_i has a standard logistic distribution, such that

$$P(u_i < z) = \frac{\exp(z)}{1 + \exp(z)}$$

It follows that

$$\pi_i = 1 - P(u_i < -x_i'\beta) = 1 - \frac{\exp(-x_i'\beta)}{1 + \exp(-x_i'\beta)} = \frac{\exp(x_i'\beta)}{1 + \exp(x_i'\beta)}$$

where the last equality follows from the symmetry of the logistic distribution. By comparing this expression with (4.8), we recognize the same conditional probability function.

4.2.5 Interpretation of Parameters

The probit and logit models are non-linear models since $G(\cdot)$ is a non-linear function. Therefore, the parameter β_l associated with the l-th element in x_i does not directly measure the marginal effect $\partial \mathrm{E}(y_i|x_i)/\partial x_{il} = \partial \pi_i/\partial x_{il}$. Rather, applying the chain rule of differentiation, we obtain the **marginal probability effect** (MPE)

$$MPE_{il} = \frac{\partial \pi_i}{\partial x_{il}} = \frac{dG(x_i'\beta)}{d(x_i'\beta)} \frac{\partial x_i'\beta}{\partial x_{il}} = g(x_i'\beta)\beta_l \tag{4.9}$$

where $g(x_i'\beta)$ denotes the first derivative of $G(x_i'\beta)$. Strictly speaking, there are two marginal probability effects $\partial P(y_i = 1)/\partial x_{il}$ and $\partial P(y_i = 0)/\partial x_{il}$. However, since $\partial P(y_i = 0)/\partial x_{il} = -\partial P(y_i = 1)/\partial x_{il}$ in the binary response model, it is common to refer to $\partial P(y_i = 1)/\partial x_{il} = \partial \pi_i/\partial x_{il}$ as "the" MPE. It follows that the change in the success probability due to a change in the l-th regressor by the amount Δx_{il} can be expressed by $\Delta \pi_{il} \approx [g(x_i'\beta)\beta_l]\Delta x_{il}$. The smaller Δx_{il}, the better this linear approximation. The probit marginal effects are obtained for $G(x_i'\beta) = \Phi(x_i'\beta)$, hence

$$MPE_{il}\big|_{\text{probit}} = \phi(x_i'\beta)\beta_l$$

The logit marginal effects are obtained for $G(x_i'\beta) = \Lambda(x_i'\beta)$, hence

$$MPE_{il}\big|_{\text{logit}} = \Lambda(x_i'\beta)[1 - \Lambda(x_i'\beta)]\beta_l$$

Three important aspects of the marginal probability effects in the probit and logit model are

- The sign of the marginal effect is equal to the sign of β_l
- The effect is largest for $x_i'\beta = 0$
- The effect varies among individuals

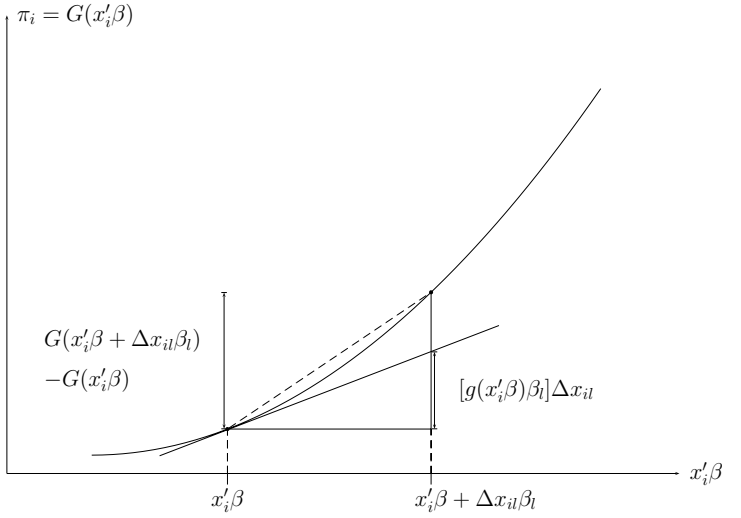
The second and third item follow since in both models, the densities peak at zero and the MPE's depend on x_i, respectively. In practice, one often wants to make statements about the expected effect, or the effect for a "typical" person. Two types of expectations are possible. The first is the expected marginal effect $\mathrm{E}_x[g(x_i'\beta)]\beta_l$, which can be estimated by

$$\widehat{AMPE}_l = \frac{1}{n}\sum_{i=1}^{n} g(x_i'\beta)\beta_l \tag{4.10}$$

where AMPE stands for the **average marginal probability effect**. Equation (4.10) includes calculating the marginal probability effect for each individual and then averaging over the individual effects. The second is the marginal effect for the expected explanatory variables $g(\mathrm{E}(x)'\beta)\beta_l$, which can be estimated by $g(\bar{x}'\beta)\beta_l$. Because of the non-linearity of $g(\cdot)$, the two expressions are not identical. However, they may be close in practice.

The discussion so far has implicitly assumed that the explanatory variables are continuous. In many applications, this is the exception rather than the rule. If explanatory variables are discrete (such as years of schooling) or even binary (such as marital status or a treatment indicator), computing the effect of an infinitesimal change of x_{il} can be highly inaccurate. Likewise, if x_{il} is continuous, we might be interested in a change that is much larger than required for the MPE's to be accurate.

Fig. 4.3. *Discrete versus Marginal Change in Nonlinear Models*



Instead, we define

$$\Delta\pi_{il} = G(x_i'\beta + \Delta x_{il}\beta_l) - G(x_i'\beta) \tag{4.11}$$

as the **discrete change** in the probabilities associated with a discrete change in the l-th regressor by the amount Δx_{il}. An example is a change of x_{il} from zero to one (such as if x_{il} was a dummy variable), or a change by one standard deviation. Again, one can compute the average effect, or the effect for average values. If the second method is chosen, one can also use the median value, or the modal value (rather than the arithmetic mean) for categorical variables, in order to define a "typical" individual. To obtain the per-unit change in the probability, we can divide (4.11) by Δx_{il}. In Figure 4.3, we illustrate the difference between marginal and discrete changes. The discrete change can be obtained by taking the difference of $G(\cdot)$ at two distinct values, $x_i'\beta$ and $x_i'\beta + \Delta x_{il}\beta_l$. By using marginal effects, we approximate this change in

$G(\cdot)$ linearly. The more convex or concave the function, the more inaccurate becomes the approximation.

Exercise 4.5.

- How can you use your answer in Exercise 4.3 to determine the range of $x_i'\beta$ in which the true (discrete) change is (under-) overestimated by the approximation $[g(x_i'\beta)\beta_l]\Delta x_{il}$?

The specific structure of the logit model offers an alternative way to describe the effect of explanatory variables. It is based on the so-called **odds**, defined as $P(y_i = 1)/P(y_i = 0)$. In the logit case, the odds are simply

$$\frac{P(y_i = 1|x_i)}{P(y_i = 0|x_i)} = \frac{\pi_i}{1 - \pi_i} = \exp(x_i'\beta) \tag{4.12}$$

Now consider the effect of a change in the l-th explanatory variable on the odds, which can be examined in two, albeit related, ways. First, we can look at the **factor change** in the odds associated with an increase in x_{il} by Δx_{il}. We have

$$\frac{\exp(x_i'\beta + \Delta x_{il}\beta_l)}{\exp(x_i'\beta)} = \exp(\Delta x_{il}\beta_l) \tag{4.13}$$

which simplifies to $\exp(\beta_l)$ if we look at a unit increase in x_{il}, which implies $\Delta x_{il} = 1$. The term $\exp(\beta_l)$ is also called the **odds ratio**. If an explanatory variable has no effect, the associated odds ratio is unity: increasing x_l by one unit leaves the odds unchanged. Similarly, a positive (negative) coefficient leads to an odds ratio greater (smaller) than one. Second, we can compute the **relative change** in the odds if the l-th regressor increases by Δx_{il}. We have

$$\frac{\exp(x_i'\beta + \Delta x_{il}\beta_l) - \exp(x_i'\beta)}{\exp(x_i'\beta)} = \exp(\Delta x_{il}\beta_l) - 1 \tag{4.14}$$

which can be interpreted as percentage change in the odds. Specifically, for an unit increase in x_{il}, the odds change by $100 \times [\exp(\beta_l) - 1]$ percent. From a practitioner's point of view, both measures – the odds ratio and the percentage change – incorporate the same information and are used equally often, and the choice between them is simply a matter of taste.

Sometimes the relation (4.12) is used as the defining characteristic of the logit model. In particular, taking logs yields $\log[\pi_i/(1-\pi_i)] = x_i'\beta$. Hence, the logit model is one in which the logarithmic odds (or "log-odds") are a linear function in the parameters β.

In practice, all elements of β are unknown. They are estimated from the data. Hence, marginal probability effects and discrete changes, just as the odds

ratio and the percentage change, are estimates as well. They have (a priori, before the sample is drawn) a sampling distribution with non-zero variance. Estimation of marginal effects as well as their standard errors is discussed in Section 4.4.5.

4.3 Discrete Choice Models

The probit and logit models, as discussed before, are used to model a binary dependent variable. In many cases, the realization of this variable can be interpreted as the outcome of an individual's choice between two alternatives. This raises the question of whether and how statistical models for binary dependent variables can be related to microeconomic models of choice, based on **utility maximization** subject to constraints. It turns out that such a link can in fact be made, and the class of **discrete choice models** has been developed to derive discrete probability models based on utility maximization (McFadden, 1974a, 1981). With the help of these models it is possible to estimate parameters of a utility function by observing choices made by different individuals.

Our starting point is the utility function $\mathcal{U}(z_{ij}, x_i)$. Here, $j = 0, 1$ is an index for the alternative under consideration. The utility function can be thought of as an indirect utility function, the parameters of which are attributes of the alternative, denoted as z_{ij}, and individual-specific characteristics x_i. An example of z_{ij} is the price; the vector x_i might include, for example, individual income. In the simplest version, we can specify a linear utility function with additive random error

$$\mathcal{U}(z_{ij}, x_i) = z'_{ij}\gamma + x'_i\beta_j + u_{ij} \qquad j = 0, 1 \tag{4.15}$$

Utilities have both a deterministic and a random component. The deterministic part is modeled by a linear index $z'_{ij}\gamma + x'_i\beta_j$ that varies across individuals *and* choices. The random error could stand for partial ignorance of the econometrician, but it could also capture intrinsic randomness in people's behavior. To reflect the nature of utility in (4.15), it is common to use the name **random utility** or **random utility maximization** when referring to the optimization rule derived from these utilities.

¿From equation (4.15), alternative 1 is chosen if it has a higher utility than alternative 0, algebraically

$$z'_{i1}\gamma + x'_i\beta_1 + u_{i1} > z'_{i0}\gamma + x'_i\beta_0 + u_{i0}$$

or, alternatively, if

$$u_{i1} - u_{i0} > -(z_{i1} - z_{i0})'\gamma - x'_i(\beta_1 - \beta_0) \tag{4.16}$$

Now assume that u_{1i} and u_{0i} are jointly normally distributed with mean 0, variances σ_1^2 and σ_0^2, and covariance σ_{10}. Applying the rules for the linear

combination of correlated normally distributed variables, we obtain that $u_2 - u_1 \sim N(0, \sigma^2)$ where $\sigma^2 = \sigma_1^2 - 2\sigma_{10} + \sigma_0^2$.

Hence, alternative 1 is chosen with probability

$$P(y_i = 1|z_{i1} - z_{i0}, x_i) = \Phi\left(\frac{(z_{i1} - z_{i0})'\gamma + x_i'(\beta_1 - \beta_0)}{\sqrt{\sigma_1^2 - 2\sigma_{10} + \sigma_0^2}}\right)$$

Again, we need to impose some normalizations to ensure identification. First, we see that only the difference $\beta = \beta_1 - \beta_0$ can be estimated. This is intuitively plausible. In our linear framework, changes in x_i affect the choice probability only inasmuch as they change the utility of one alternative more than the utility of the other. The absolute utility levels are inconsequential for the choice itself. For example, income might have an effect on the choice probability by increasing the utility of driving a car (private transport) and leaving the utility of taking public transport unchanged.

Second, we can multiply both sides of equation (4.16) by any constant without changing the choice situation and the choice probability. Thus the standard deviations of u_{i1}, u_{i0} and γ and β are only identified up to scale. Setting $\sigma_1^2 - 2\sigma_{10} + \sigma_0^2 = 1$, we obtain the estimable probit model

$$P(y_i = 1|z_{i1} - z_{i0}, x_i) = \Phi\Big((z_{i1} - z_{i0})'\gamma + x_i'\beta\Big) \tag{4.17}$$

Therefore, in practice, all one needs to do is to estimate a probit model including two types of explanatory variables, the first being the difference in the values of the attributes between the two alternatives for individual i, the second being the individual-specific characteristics. The constant of the model can be interpreted as the relative contribution of unobserved attributes to the utility of alternative 1. For example, if there are no individual characteristics, a positive constant means that the average individual favors alternative 1 over alternative 0, *ceteris paribus*, i.e., if the attributes are the same.

¿From (4.17), we see that γ can only be estimated if the choice-specific attributes $(z_{i1} - z_{i0})$ vary across individuals. If they do not vary, we have

$$P(y_i = 1|z_1 - z_0, x_i) = \Phi(x_i'\beta)$$

where $x_i'\beta = \beta_0^* + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$ and $\beta_0^* = \beta_0 + (z_1 - z_0)'\gamma$ is the new constant of the model. Hence, we cannot identify β_0 and γ separately, we can only identify β_0^*. The classic example of data with individual choice-specific attributes is in transportation research, where $j = 0, 1$ denotes the choice between private and public transport, and z_{ij} are things like travel time and travel cost.

If the random component of the utility function is independently **type-I extreme value**, or **Gumbel**, distributed with distribution function

$$F(a) = \exp(-\exp(-a))$$

then the *difference* between two random errors $u_{i1} - u_{i0}$ has a standard logistic distribution, and the discrete choice model takes the form of the binary logit model with

$$P(y_i = 1 | z_{i1} - z_{i0}, x_i) = \frac{\exp[(z_{i1} - z_{i0})'\gamma + x_i'\beta]}{1 + \exp[(z_{i1} - z_{i0})'\gamma + x_i'\beta]} \tag{4.18}$$

If we let choice-specific attributes be constant across individuals, we obtain the familiar expression of the logit model, $\Lambda(x_i'\beta)$. In this case, as before, we cannot identify the constant separately from γ.

Example 4.2. Contingent Valuation

One important area of application for discrete choice models is the estimation of willingness-to-pay for public goods (Hanemann, 1984). Such studies are based on surveys in which respondents have to choose between hypothetical alternatives, or scenarios. This type of discrete choice experiment is also known under the name of "contingent valuation". In *closed-end* contingent valuation surveys, people are asked whether they would be willing to contribute a certain amount t_i to a project or not. In *open-ended* contingent valuation surveys they are asked how much they are willing to contribute. We consider the closed-end method here. In this case, respondents can answer with either "yes" or with "no", and the amount is varied over respondents.

Let $\mathcal{U}_{i1} = \beta_1 + \gamma t_i + u_{i1}$ and $\mathcal{U}_{i0} = \beta_0 + u_{i0}$ denote a person's utility with and without the public good, respectively. Presumably, if it is a good, we should have that $\beta_1 > \beta_0$. Moreover, people dislike paying for it, so that $\gamma < 0$. A person is indifferent between the choices whenever

$$\gamma t_i = \beta_0 - \beta_1 + u_{i1} - u_{i0}$$

and the expected willingness-to-pay is

$$\mathrm{E}(t_i) = -\frac{\beta_1 - \beta_0}{\gamma}$$

If we assume that the errors of the model are normally distributed, the parameters $\beta = \beta_1 - \beta_0$ and γ can be estimated using a standard probit model, since

$$P(y_i = 1 | t_i) = \Phi(\beta + \gamma t_i)$$

Of course, the model can be extended by allowing for additional variables x_i. In this case, the willingness-to-pay is no longer constant but depends on these other factors as well. To obtain an overall willingness-to-pay, one can then average over the x's.

To summarize, the probit and logit models may be derived from economic behavior affecting the binary choice. In this case, it is possible to estimate parameters of an indirect utility function and to predict economic behavior if explanatory variables change. However, we do not need to have an underlying (micro-) economic model with behavioral interpretation (like utility maximization under constraints). Rather, we can treat the probit and logit as statistical models, more precisely conditional probability models, that describe the true data generating process of binary outcomes appropriately. This explains the more general chapter heading "Binary Response Models" instead of "Binary Choice Models", the latter of which is often found in the econometric literature.

4.4 Estimation

4.4.1 Maximum Likelihood

The parameters of binary response models can be estimated by the method of maximum likelihood. Here – as elsewhere in the book, unless mentioned otherwise – we assume random sampling. Under this assumption, the log-likelihood function can be written as

$$\log L(\beta; y, x) = \sum_{i=1}^{n} y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i) \tag{4.19}$$

As before, we substitute $\pi_i = G(x_i'\beta) = \Phi(x_i'\beta)$ for the probit model and $\pi_i = G(x_i'\beta) = \Lambda(x_i'\beta)$ for the logit model. The log-likelihood function (4.19) is essentially the same as the one in the Bernoulli model discussed in Chapter 3. The only difference is that π_i is now a function of x_i and β rather than a constant. As a consequence, the score of the log-likelihood function has an additional element, $\partial \pi_i / \partial \beta$. In particular,

$$s(\beta; y, x) = \sum_{i=1}^{n} \left(\frac{y_i - \pi_i}{\pi_i(1 - \pi_i)} \right) \frac{\partial \pi_i}{\partial \beta}$$

Therefore, the score vector of the binary response model can be written as

$$s(\beta; y, x) = \sum_{i=1}^{n} \left(\frac{y_i - G(x_i'\beta)}{G(x_i'\beta)[1 - G(x_i'\beta)]} \right) g(x_i'\beta) x_i \tag{4.20}$$

For the probit model, just replace $G(x_i'\beta) = \Phi(x_i'\beta)$ and $g(x_i'\beta) = \phi(x_i'\beta)$. For the logit model, one can further simplify the score (4.20) to

$$s(\beta; y, x) = \sum_{i=1}^{n} \big[y_i - \Lambda(x_i'\beta) \big] x_i \tag{4.21}$$

This follows, since $\Lambda(x_i'\beta) = \exp(x_i'\beta)/[1 + \exp(x_i'\beta)]$, and the first-order derivative with respect to β can be written as

$$\frac{\partial \Lambda(x_i'\beta)}{\partial \beta} = \Lambda(x_i'\beta)\big[1 - \Lambda(x_i'\beta)\big]x_i$$

The maximum likelihood estimator $\hat{\beta}$ solves the first-order conditions for a maximum, $s(\hat{\beta}; y, x) = 0$. Setting (4.20) equal to zero defines a system of $k + 1$ non-linear equations, and in general a closed-form solution for $\hat{\beta}$ is not available. A numerical solution can be obtained by using iterative optimization algorithms as discussed in Chapter 3.5.2.

Example 4.3. Probit and Logit Model of Being Childless

In Example 4.1, we obtained negative predictions from the linear probability model. To avoid this kind of problem, we now estimate the probit and the logit models, which explicitly take into account the requirement $0 \leq G(x_i'\beta) \leq 1$. The results are reported in Table 4.1.

Table 4.1. *Probit and Logit Estimates of Fertility Decision*

Dependent Variable: *childless*		
	Probit Model	Logit Model
linear time trend	0.0027	0.0050
	(0.0025)	(0.0047)
years of education	0.0314	0.0630
	(0.0071)	(0.0136)
white	0.0626	0.1287
	(0.0626)	(0.1182)
number of siblings	-0.0117	-0.0211
	(0.0071)	(0.0135)
constant	-1.5030	-2.6764
	(0.1158)	(0.2251)
Log-likelihood value	-2,107.11	-2,105.96
LR (χ_4^2)	39.56	41.86
Observations	5,150	5,150

Notes: Standard errors in parentheses.

Both models reject the constant-only model significantly. For example, the 1% critical value of the χ_4^2 distribution is 13.28, which is much less than the LR test statistics of 39.56 (probit) and 41.86 (logit). We recognize the same signs of the coefficients in both models, which implies that marginal effects do not differ in their signs between the two models. Note that the coefficients cannot be interpreted directly as measuring marginal effects because of the

nonlinearity in π. In Example 4.4, we will discuss the estimation and interpretation of marginal effects in greater detail. Finally, we can multiply the probit estimates by 1.6, for example $-1.50 \times 1.6 = -2.40$, and roughly obtain the logit estimates (here -2.68). Of course, the estimates do not exactly coincide since we used this factor just as an approximation.

Exercise 4.6.

- How can you test whether education is statistically significant in explaining the probability of being childless?
- Assume you estimated the logit model without controlling for race (*white*) and the *number of siblings*. The log-likelihood value using this specification is -2108.21. How can you test whether both variables have no effect on the probability of being childless?

The simple structure of the score function in the logit model, (4.21), gives an interesting interpretation. Define $\hat{u}_i = y_i - \hat{\Lambda}_i$, where $\hat{\Lambda}_i = \Lambda(x_i'\hat{\beta})$. Then, the first-order conditions can be re-interpreted as a simple moment restriction

$$\sum_{i=1}^{n} \hat{u}_i x_i = 0$$

Just as for the OLS estimator, the parameter values are chosen such that the empirical correlation between the residuals and the regressors is zero. Also, if x_i contains a constant, the solution must be such that

$$\sum_{i=1}^{n} (y_i - \hat{\Lambda}_i) = 0$$

or $\bar{y} = \bar{\hat{\Lambda}}$. Hence, the average of the predicted values is equal to the proportion of ones in the sample. In the probit model, though, such a simple relationship does not exist.

So far, we have only studied the necessary condition for a maximum of the log-likelihood function. To check whether the solution to the score equation is indeed a unique and global maximum, we also need to inspect the Hessian matrix. Taking derivatives of (4.20) and expectations thereof, we find that

$$\mathrm{E}\left[H(\beta; y, x)\right] = \sum_{i=1}^{n} -\frac{g(x_i'\beta)^2 x_i x_i'}{G(x_i'\beta)[1 - G(x_i'\beta)]} \tag{4.22}$$

This matrix is negative semidefinite for all possible values of β. Minus its inverse is the asymptotic covariance matrix of the ML estimator.

This result is not a direct proof that the log-likelihood function is globally concave in all samples, since it refers to the expectation of the Hessian. In the logit case, the situation is simple, since the actual Hessian

$$H(\beta; y, x) = \sum_{i=1}^{n} -\Lambda(x_i'\beta)\big[1 - \Lambda(x_i'\beta)\big]x_i x_i'$$

does not depend on y, so that it coincides with (4.22). For the probit model, the situation is a bit more complicated. But again, it can be shown that the log-likelihood function is globally concave (for example Maddala, 1983: 63). This is reassuring, because it means that there is a unique (global) maximum. Hence, the solution of the first-order condition, if it exists, gives us the maximum likelihood estimator.

Alas, there is not always a solution. Apart from the standard requirement that the regressors are not linearly dependent, a more subtle issue can arise in such models: one commonly referred to as "perfect prediction". This is discussed next.

4.4.2 Perfect Prediction

Perfect prediction can occur in connection with binary explanatory variables. Suppose there is a regressor (call it d_i) such that whenever $d_i = 1$, the dependent variable is equal to one. When $d_i = 0$, the dependent variable is either 1 or 0. Hence, the correlation between d_i and y_i is not perfect; it can even be quite low. In the context of a linear regression model, the situation would not pose any problem.

In a logit or probit model, however, such a situation makes estimation of the effect of d_i on $P(y_i = 1|d_i, x_i)$ impossible. In order to understand why this is the case, consider the log-likelihood function with d_i as additional regressor

$$\begin{aligned}\log L(\beta, \delta) &= \sum_{i=1}^{n} y_i \log G(x_i'\beta + \delta d_i) + (1 - y_i)\log[1 - G(x_i'\beta + \delta d_i)] \\ &= \sum_{d_i=1} \log G(x_i'\beta + \delta) \\ &\qquad + \sum_{d_i=0} y_i \log G(x_i'\beta) + (1 - y_i)\log[1 - G(x_i'\beta)]\end{aligned}$$

where $G(x_i'\beta + \delta d_i) = \Phi(x_i'\beta + \delta d_i)$ for the probit model and $G(x_i'\beta + \delta d_i) = \Lambda(x_i'\beta + \delta d_i)$ for the logit model. The simplification for the first summand on the right arises since, by assumption, $y_i = 1$ whenever $d_i = 1$. But considered as a function of δ, only the first term is relevant. Thus, the maximization of the log-likelihood requires that the first term on the right-hand side becomes as large as possible, in this case zero, which means that $\delta \to \infty$. Hence, a well defined maximum-likelihood estimator for δ in the interior of the parameter space does not exist.

Perfect prediction can arise in four different ways. First, $y_i = 1$ whenever $d_i = 1$, as assumed so far. Second, $y_i = 0$ whenever $d_i = 1$, third $y_i = 1$ whenever $d_i = 0$, and fourth $y_i = 0$ whenever $d_i = 0$. In the last case, for example, the offending term of the likelihood is

$$\sum_{d_i=0} \log \left[1 - G(x_i'\beta)\right]$$

which is maximized by letting the constant go to $-\infty$.

The following remarks apply. Perfect prediction is in general not an identification problem but a problem that arises in specific samples and that usually would disappear if one were to collect more data or another sample. Naturally, the more likely the problem is to occur, the smaller the number of observations for which $d_i = 1$. In the extreme, if there is only a single i for which $d_i = 1$, perfect prediction must occur. Dummy variables with few ones typically arise in problems where dummies are used to assign group membership and there are many groups. For example, in a model of job mobility, where the dependent variable is 1 if a person changed the job in a given year, and 0 otherwise, we want to control for occupational status. The finer the selected classification (one distinguishes so-called one-digit, two-digit, and three-digit occupations), for a given data size, the fewer the number of observations in a given occupation and the greater the probability that only non-movers are observed within the occupation (since changing jobs is a relatively rare event). To remedy the problem, the offending binary explanatory variable needs to be dropped. Most econometric software packages will do this automatically and give a warning message.

4.4.3 Properties of the Estimator

As long as the model is correctly specified, the standard results for ML estimation apply. To recapitulate, the ML estimator is consistent, asymptotically efficient, and asymptotically normally distributed. In the context of binary response models we get the asymptotic distribution

$$\hat{\beta} \sim Normal\left(\beta, -\mathrm{E}\left[H(\beta; y, x)\right]^{-1}\right)$$

The formula for the expected Hessian was given in (4.22). Its estimated counterpart is

$$\hat{H}(\hat{\beta}; y, x) = \sum_{i=1}^{n} -\frac{g(x_i'\hat{\beta})^2 x_i x_i'}{G(x_i'\hat{\beta})[1 - G(x_i'\hat{\beta})]}$$

A major concern is the robustness of these desirable properties as some of the model assumptions are violated. For the linear model, the consequences of violating standard assumptions, such as endogeneity of a regressor or a

non-standard covariance matrix of the regression error, are relatively easy to derive, and are well understood. In particular, it is the case that the ordinary least squares estimator in the standard linear model remains unbiased even if the errors are autocorrelated or heteroscedastic. Moreover, the solution required to obtain a consistent estimator of the covariance matrix of the estimator is relatively simple.

Unfortunately, heteroscedasticity is much more of a problem in the binary response model. The effect of heteroscedasticity in the probit model was explored in detail by Yatchew and Griliches (1985). They show that if heteroscedasticity is unrelated to the explanatory variables then no bias results. In this case just the scaling is affected, but the scaling is arbitrary anyway. A serious problem arises, however, if the heteroscedasticity is related to the explanatory variables. For instance, let

$$\sigma_i^2 = \exp(\gamma_1 + \gamma_2 x_{i1})$$

In this case, we can write

$$\frac{x_i'\beta}{\sigma_i} = \frac{x_i'\beta}{\sqrt{\exp(\gamma_1 + \gamma_2 x_{i1})}}$$

and there is no hope that the standard probit or logit model can estimate β consistently. However, as long as the form of heteroscedasticity is known, we can account for it in the log-likelihood function and, with some adequate normalization and parameterization, get consistent estimates.

Exercise 4.7.

Assume a probit model with heteroscedasticity of known form with $\sigma_i^2 = \exp(\gamma_1 + \gamma_2 x_{i1})$.

- Does this specification of the variance make sense? Why (not)?
- Derive the outcome probabilities.
- Which normalization is required for identification?

Of course, endogeneity is a concern as well. If ignored, the maximum likelihood estimator will be inconsistent, as would be the ordinary least squares estimator in the linear model. How to handle binary response models with continuous and binary endogenous regressors is discussed next.

4.4.4 Endogenous Regressors in Binary Response Models

In probit and logit applications, the assumption of exogenous regressors is often questionable. A good example is provided by Evans and Schwab (1995), who analyze the effect of attending a catholic school on the probability of graduating from high school. They rightly point out that the catholic school dummy might be endogenous:

> *"Consider a child whose parents care a great deal about his welfare. We would expect this child to do well in school for two reasons. First, his parents will see that he attends a better than expected school and will be more willing to pay the cost for sending him to a private school. Second, he will succeed in part because of factors that cannot be observed but are under his parents' control. They will spend more time reading to him, they will stress the importance of good grades, and they will see that he does his homework. A single-equation model would mistakenly attribute all of this child's success to his private school."*
> (Evans and Schwab, 1995, p. 961)

In this example, the **endogenous regressor** is binary itself. In other applications the endogenous regressor is a continuous variable. For example, Costa (1995) estimates a model where the retirement decision is modeled as a function of the pension benefit. The pension benefit is potentially endogenous as it depends on past retirement decisions, and thus on the unobserved "disutility of work". Finally, it is also possible that the *dependent variable* is continuous and the endogenous regressor is binary. This possibility is discussed later in Chapter 7 on estimating treatment effects.

We now discuss possible approaches to estimating probit models with endogenous regressors. For simplicity, we limit our discussion to models with a single endogenous regressor, and for notational convenience, we drop the subscript i in this section. Formally, endogeneity in binary response models can be represented in a two equation system:

Model 1: Continuous endogenous regressor

$$y_1^* = \alpha_1 y_2 + \beta_1 x_1 + u_1$$

$$y_2 = \delta_1 x_1 + \delta_2 x_2 + v_2$$

The observed binary outcome is $y_1 = I(y_1^* > 0)$. The conditions for identification are the same as those in the usual simultaneous equations model. Either u_1 and v_2 are independent, or else $\delta_2 \neq 0$, i.e., x_2 is an instrument.

Model 2: Binary endogenous regressor

$$y_1^* = \alpha_1 y_2 + \beta_1 x_1 + u_1$$
$$y_2^* = \delta_1 x_1 + \delta_2 x_2 + v_2$$

The observed binary outcomes are $y_1 = I(y_1^* > 0)$ and $y_2 = I(y_2^* > 0)$. Because of the non-linearity of the indicator function, this model is identified even if u_1 and v_2 are correlated and $\delta_2 = 0$.

Endogeneity in these recursive systems arises from correlation between the two equation errors u_1 and v_2, thus capturing for example the effect of common **omitted variables**. One reason why we restrict our attention to recursive systems is that it can be shown, for Model 2 with binary endogenous regressor, that a full simultaneous model is logically inconsistent (Maddala, 1983, p. 119). We point out that if the first equation in Model 2 depends on the latent index y_2^* rather than the binary realization, then we are effectively back to the structure of Model 1. In both models, we are interested in obtaining consistent estimates of α_1. x_2 denotes a set of potential instruments.

Probit Model with Continuous Endogenous Regressor

A method for estimating Model 1 has been proposed by Rivers and Vuong (1988). See also Wooldridge (2002) for a simple discussion of the procedure. Assume that u_1 and v_2 are bivariate normally distributed with zero mean, correlation ρ, and variances 1 and σ_v^2, respectively. Thus, $u_1 = \theta_1 v_2 + \varepsilon_1$, where $\varepsilon_1 \sim Normal(0, 1 - \rho^2)$ (see Appendix 7.5), and we can write the first equation *conditional* on v_2 as

$$y_1^* = \alpha_1 y_2 + \beta_1 x_1 + \theta_1 v_2 + \varepsilon_1 \tag{4.23}$$

Of course, we do not know v_2, but we can replace it by an estimate. The Rivers and Vuong (1988) two-step approach is then to estimate in *Step 1* the second equation by OLS to get residuals $\hat{v}_2$. In *Step 2*, run a probit of y_1 on y_2, x_1 and $\hat{v}_2$, to obtain consistent estimators of the probit equation. The probit parameters are estimated only up to scale, with factor $(1 - \rho^2)^{-1/2}$. An estimate for ρ is $\hat{\rho} = \hat{\theta}_1 \hat{\sigma}_{v_2}$, where $\hat{\sigma}_{v_2}$ is the square root of the usual error variance estimator from the first-stage regression.

A nice feature of the Rivers and Vuong approach is that it leads to a simple test for exogeneity. A z-test of the null hypothesis $H_0 : \theta_1 = 0$ tests whether y_2 is exogenous. If there is evidence of endogeneity and we apply a two-step procedure to find consistent estimators, the usual probit standard errors are not valid. The asymptotic variance of the estimated probit parameters needs to be adjusted to account for the first stage estimation (see Rivers and Vuong, 1988). A closely related procedure for a Tobit model (see Chapter 7 for further details on this model) with an endogenous continuous regressor has been developed by Smith and Blundell (1986).

As an alternative, one might be tempted to estimate the model in a two-stage least squares fashion, i.e., estimate the model

$$y_1^* = \alpha_1 \hat{y}_2 + \beta_1 x_1 + \epsilon_1$$

where $\hat{y}_2$ is the prediction from a first-stage OLS regression. If one assumes a linear probability model, this approach works well. In the probit model, however, the regression parameters are only estimated up to scale, and the scaling factor, the standard deviation of $\epsilon_1 = \alpha_1(y_2 - \hat{y}_2) + u_1$ is unknown and cannot be estimated in this model. This method is therefore not recommended, although one could use it to test hypotheses such as $\alpha_1 = 0$, or obtain valid ratios of coefficients, since these do not depend on scaling.

Probit Model with Binary Endogenous Regressor

This is the model considered by Evans and Schwab (1995). We can write the two-equation model compactly as

$$y_1 = I(\alpha_1 y_2 + \beta_1 x_1 + u_1 > 0) \tag{4.24}$$

and

$$y_2 = I(\delta_1 x_1 + \delta_2 x_2 + v_2 > 0), \tag{4.25}$$

where u_1 and v_2 are independent of x_1 and x_2 and bivariate standard normally distributed. If $\rho \neq 0$, then probit estimation of (4.24) is inconsistent for α_1 and β_1. Note that there is no simple reduced-form as before. Two-stage estimation is therefore not an option. Rather, one can proceed by **full-information maximum likelihood estimation**. This is described in detail in Wooldridge (2002, p. 477). The idea is to find the expressions for the four joint probabilities $P(y_1 = 1, y_2 = 1)$, $P(y_1 = 0, y_2 = 1)$, $P(y_1 = 1, y_2 = 0)$, and $P(y_1 = 0, y_2 = 0)$. Due to the recursive structure of the underlying model, this is not the same as the standard **bivariate probit model**, where $y_1 = I(\beta_1 x_1 + u_1 > 0)$ and $y_2 = I(\delta_2 x_2 + v_2 > 0)$ and α is not estimated.

4.4.5 Estimation of Marginal Effects

In Section 4.2.5, we derived general statements of the marginal effects in binary response models. In practice, computation of marginal effects requires that the unknown true parameter is replaced by an estimator, here the ML estimator. Hence, the marginal effects are random variables as well, and we are not only interested in their point estimation, but also in sampling variability.

Let $\hat{\beta}$ be the ML estimator of β. The following properties of ML estimation can be exploited:

1. *Invariance* says that the ML estimator of any function $h(\beta)$ is simply $h(\hat{\beta})$. For example, consider estimating the odds ratio $or = \exp(\beta)$ and let $\hat{\beta}$ be the ML estimator in the logit model. Then $\widehat{or} = \exp(\hat{\beta})$.

2. The *Delta method* says that

$$\widehat{\text{Var}}[h(\hat{\beta})] = \left[\frac{\partial h(\beta)}{\partial \beta'}\right]_{\hat{\beta}} \widehat{\text{Var}}(\hat{\beta}) \left[\frac{\partial h(\beta)}{\partial \beta'}\right]'_{\hat{\beta}}$$

This approximation improves with larger sample size. For the odds ratio in the above example, we obtain

$$\widehat{\text{Var}}(\widehat{or}) = \exp(2\hat{\beta})\widehat{\text{Var}}(\hat{\beta})$$

3. The function $h(\hat{\theta})$ is approximately normally distributed with

$$h(\hat{\theta}) \overset{app}{\sim} Normal\left(h(\theta), \widehat{\text{Var}}[h(\hat{\theta})]\right)$$

or in our example $\widehat{or} \overset{app}{\sim} Normal\left(or, \widehat{\text{Var}}(\widehat{or})\right)$

We can apply these considerations to the estimation of probabilities and marginal probability effects in the probit and logit model. Consider probabilities first. To estimate probabilities, we simply replace β by $\hat{\beta}$ in the conditional probability expressions to obtain **predicted probabilities**, formally

$$\hat{\pi}_i = G(x_i'\hat{\beta})$$

Since the idea is the same for the probit and logit model, we can use the generic function $G(x_i'\hat{\beta})$. Furthermore, the asymptotic variance of the prediction using the delta rule can be derived as

$$\widehat{\text{Var}}\,(\hat{\pi}_i) = [g(x_i'\hat{\beta})x_i']\widehat{\text{Var}}(\hat{\beta})[g(x_i'\hat{\beta})x_i']' = g(x_i'\hat{\beta})^2\left[x_i'\widehat{\text{Var}}(\hat{\beta})x_i\right]$$

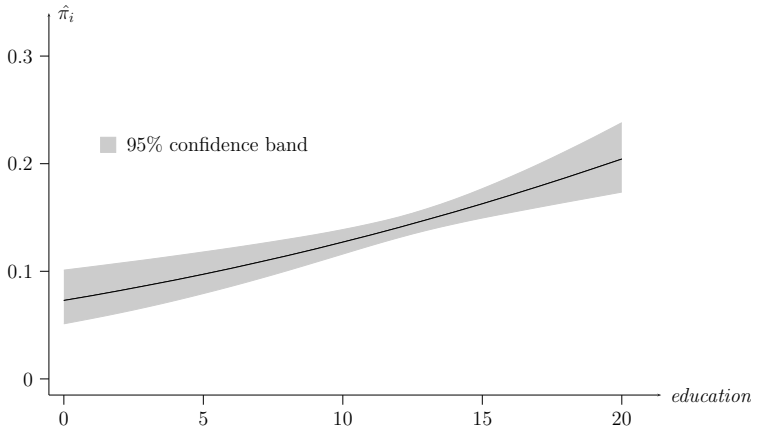
As the point estimate for the prediction itself, this expression depends on the x_i vector used. For the marginal probability effects, the same principles apply. For example, the point estimation of the marginal probability effects yields $\widehat{MPE}_{il} = g(x_i'\hat{\beta})\hat{\beta}_l$. The computation of the standard errors via the delta rule is again feasible, although more complex. Greene (2003) provides the details.

Example 4.4. Are More Educated Women More Likely Childless?

Now, we want to interpret the results reported in Table 4.1, particularly the estimated parameters, in order to shed some light on the research question posed: Is there a positive relationship between being childless and the years of education? And if so, how can we quantify this relationship?

To begin with, we note that the estimated coefficient of education is positive, confirming our hypothesis that more educated women are more likely to be childless. We can quantify the effect by calculating predicted probabilities evaluated at *each* level of education, setting all other variables to the mean. In Figure 4.4, we plot the results together with a 95% confidence band obtained from the asymptotic variance derived earlier. We only show the graph for the probit model since the logit model gives about the same shape. We see that the probability of being childless increases from 7.3 percent for women without schooling to 20.4 percent for women with 20 years of schooling. This increase is statistically significant, and even if we compare the last prediction to the prediction at the average schooling level of 12.24 years ($\hat{\pi} = 0.142$), the increase remains significant. As expected, the confidence band around the mean values is the smallest, with higher variation for the minimum and maximum values of *educ*.

Fig. 4.4. *Predicted Probabilities by Years of Education*



We can further investigate the changes in probabilities by calculating marginal effects. In Table 4.2, we report the predicted linear index, evaluated at the means, which is an estimator of the expected linear index. Then, we calculate the value of the density function, $g(\bar{x}'\hat{\beta})$ and multiply it by the coefficient to obtain the MPE. In the probit model we get a value of 0.0071, in the logit model we get a value of 0.0077. This implies that an increase in education by one unit increases the probability of being childless by about 0.71 **percentage points** in the probit model (0.77 in the logit).

Table 4.2. *The Effect of Education on the Probability of Being Childless*

	Probit Model	Logit Model
Marginal Probability Effects		
years of schooling ($\hat{\beta}_{educ}$)	0.0314	0.0630
$\bar{x}'\hat{\beta}$	-1.0699	-1.8030
$g(\bar{x}'\hat{\beta})$	0.2251	0.1415
$g(\bar{x}'\hat{\beta})\hat{\beta}_{educ} = \widehat{MPE}$	0.0071	0.0077
Discrete Changes in the Predicted Probabilities		
min → max	0.1315	0.1410
12 years → 16 years	0.0300	0.0331
∓ (standard deviation)/2 from the mean	0.0231	0.0250
Odds (Changes with $\Delta educ = 1$)		
Factor Change (Odds Ratio)		1.0651
		(0.0145)
Relative Change		6.51%
		(1.45%)

Notes: Mean values $\bar{x} = (16.748\ 12.245\ 0.841\ 4.277)'$ for *education, time trend, white* and *number of siblings*, respectively. For illustration purposes, we report the standard errors (in parentheses) only for the odds ratio and the relative change.

¿From a practical point of view, it might be of more interest to consider a discrete change in the probabilities associated with an increase in education by four years, 20 years, or one standard deviation. Table 4.2 reports these measures. For example, the change in education from the minimum (0) to the maximum (20) years of schooling increases the probability of being childless by about 13 or 14 **percentage points** (see also Figure 4.4 for the probit case). Moreover, the change in probabilities associated with a change in schooling from 12 to 16 years can be interpreted as the effect of attending college. Here, this *decreases* the probability of *having children* by around three percentage points. Finally, we can look at the odds. The odds ratio indicates that one more year of schooling increases the odds by the factor 1.0651. Or alternatively, the relative change in the odds due to an additional year of schooling is 6.51 **percent**.

4.5 Goodness-of-Fit

A number of suggestions have been made for how to evaluate the overall quality of a binary response model. These suggestions fall into one of two categories. The first approach attempts to mimic the R^2 measure used in the linear model, i.e., to construct a measure between zero and one, where low values indicate a poor fit and high values indicate a good fit. The second approach assesses the predictive performance of the model.

R^2 measures are not directly applicable in non-linear models such as binary response models, since we do not have a proper variance decomposition result. However, a number of so-called **"pseudo R^2"** measures have been suggested. A first one, already presented in Chapter 3.6.7, was put forward by McFadden (1974a). He observes that for discrete response models, the value of the log-likelihood function is always negative, so that $\log L(\hat{\beta}_u) \geq \log L(\hat{\beta}_r)$ implies $|\log L(\hat{\beta}_u)| \leq |\log L(\hat{\beta}_r)|$. In this expression, $\log L(\hat{\beta}_r)$ is the value of the (maximized) log-likelihood function in the constant-only model and $\log L(\hat{\beta}_u)$ is the (maximized) log-likelihood value in the full model. It follows that

$$0 \leq 1 - \frac{\log L(\hat{\beta}_u)}{\log L(\hat{\beta}_r)} = R^2_{\text{McFadden}} \leq 1$$

The McFadden R^2 will be zero if the full model has no explanatory power. In this case, all slope parameters are zero and restricted and unrestricted models are the same. The McFadden R^2 will be one if the model is a perfect predictor (then $\hat{\pi}_i = 1$ whenever $y_i = 1$ and $\hat{\pi}_i = 0$ whenever $y_i = 0$), although this upper bound cannot be reached for finite parameter values. We explicitly indicate the measure as "McFadden R^2" since other pseudo R^2 measures have been proposed in the context of binary response models.

One of these is the pseudo R^2 measure proposed by McKelvey and Zavoina (1975). It is based on a for the *latent* linear model $y_i^* = x_i'\beta + u_i$. In particular, if we let $\hat{y}_i^* = x_i'\hat{\beta}$, then we can write

$$R^2_{\text{MZ}} = \frac{SSE^*}{SSR^* + SSE^*} = \frac{\sum_{i=1}^n (\hat{y}_i^* - \bar{\hat{y}}^*)^2}{n\sigma^2 + \sum_{i=1}^n (\hat{y}_i^* - \bar{\hat{y}}^*)^2}$$

where SSE^* denotes the explained sum of squares, and SSR^* denotes the "residual" sum of squares of the *latent* model. For the pseudo R^2 in the probit model σ^2 equals one, in the logit model σ^2 equals $\pi^2/3$.

An alternative approach is to look at the **proportion of correct predictions**. Based on the parameter estimates and the values of the explanatory variables, we can predict whether an observation will be a "success" or a "failure". For example, a bank needs to decide whether a credit applicant should be classified as a "defaulter" (and therefore be denied credit), or not, based on what the bank knows about the person, for example her credit history and other socio-economic characteristics. Thus, it becomes important to know how

applicants can be classified into "good" and "bad". Obviously, the bank has an interest in determining how well it can actually discriminate between the two groups of customers.

In our setup, since the prediction is binary, as is the actual outcome, the results can be summarized in a two-way contingency table. Note that this can be done **in-sample** with observations taken from the actual sample, or **out-of-sample** using an independent validation sample.

Table 4.3. *Contingency Table for Binary Predictions*

		Actual	
		Success	Failure
Model		True Positives	False Positives
	Success	$\pi(1-\alpha)$	$(1-\pi)\beta$
		False Negatives	True Negatives
	Failure	$\pi\alpha$	$(1-\pi)(1-\beta)$

In Table 4.3, we observe that there are two types of errors. α is the probability of predicting failure for an actual success. β is the probability of predicting success when the actual outcome is failure. p is the probability of a success. So far, we have not mentioned how the classification should be performed once the binary response model has been estimated. The basic idea is that we use the model to assign to each observation a "rating" r_i. This can be either the predicted probability of a success $\hat{p}_i = f(x_i'\hat{\beta})$, or simply the linear predictor $x_i'\hat{\beta}$. As a matter of fact, since the one is a monotonous function of the other and the classification rule requires only ordinality of the rating, the predictions will be identical. Secondly, we define an arbitrary cut-off value t, such that $\hat{y}_i = 1$ if $r_i \geq t$ and $\hat{y}_i = 0$ else. Interestingly, we can now compute the classification error probabilities $\alpha(t)$ and $\beta(t)$ for all possible threshold levels t by

$$\alpha(t) = 1 - P(r \geq t|y = 1)$$
$$\beta(t) = P(r \geq t|y = 0)$$

Note that these are conditional probabilities, conditioned on observing either the outcome one or the outcome zero. The unconditional probability of a misclassification is then

$$P(wrong\ prediction; t) = \pi\alpha(t) + (1-\pi)\beta(t)$$

We see that minimizing the unconditional probability of a misclassification is not that meaningful. For example, it can be the case that the outcome "success" is very rare, i.e., $P(y = 1)$ is close to zero. In this case, one would tend to choose t such that $\beta(t) = P(r \geq t|y = 0)$ is small. Since $\beta(t)$ is the error of predicting a success when we actually have a failure, this amounts to predicting a failure for everyone (or choosing a very large t).

The alternative is to focus on $\alpha(t)$ and $\beta(t)$ themselves. There exists a trade-off between these two errors: the higher $\alpha(t)$, the lower $\beta(t)$, and vice versa. For a given sample, and a given set of ratings and outcomes, all $(\alpha(t), \beta(t))$ combinations can be summarized in a so-called **Receiver Operating Characteristic** (ROC) curve. These curves can be estimated for a given sample, consisting of binary outcomes and ratings. The aforementioned probabilities are then replaced by relative frequencies.

Table 4.4. *Predicting Binary Outcomes*

Observation	Actual Outcome	Rating (r)
1	0	0.10
2	0	0.15
3	1	0.16
4	1	0.17
5	0	0.17
6	1	0.19
7	0	0.20
8	1	0.27

Consider Table 4.4 with eight observations, sorted by their rating r. The ROC curve is constructed by considering all distinct rating values observed in the data as cut-offs t, and then computing the relative frequencies $f(r \geq t|y = 0)$ as an estimate of $\beta(t)$, and $1 - f(r \geq t|y = 1)$ as an estimate of $\alpha(t)$, respectively. The estimates are listed in Table 4.5.

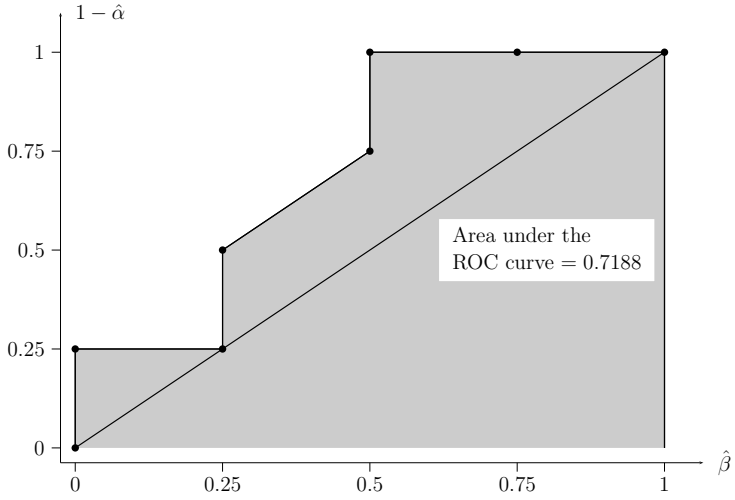
Table 4.5. *Sensitivity and Specificity*

Cut-Off Values (t)	$f(r \geq t\|y = 0) = \hat{\beta}$	$f(r \geq t\|y = 1) = 1 - \hat{\alpha}$	$\hat{\alpha}$
0.10	4/4	4/4	0/4
0.15	3/4	4/4	0/4
0.16	2/4	4/4	0/4
0.17	2/4	3/4	1/4
0.19	1/4	2/4	2/4
0.20	1/4	1/4	3/4
0.27	0/4	1/4	3/4

For example, if the cut-off value is less than or equal to 0.1, all four actual failures in Table 4.4 are classified as successes. Therefore, $\hat{\beta}(0.1) = 4/4 = 1$. On the other hand, all four actual successes are classified as successes as well, such that $\hat{\alpha}(0.1) = 1-4/4 = 0$. Now, if the cut-off is increased to 0.17, say, the situation is different. Half of all actual failures have ratings below 0.17, and thus are classified as failure. Hence, $\hat{\beta}$ has decreased to 0.5. Using the same cut-off value, three of the four actual successes are classified as success, so that the

relative frequency of a false negative has increased to 0.25. Sometimes, $1-\hat{\alpha}$ is called the **sensitivity** of a classification, and $1-\hat{\beta}$ is called the **specificity**. The ROC curve is obtained by plotting $1-\hat{\alpha}$ against $\hat{\beta}$, as shown in Figure 4.5 for the above example.

Fig. 4.5. *Receiver Operating Characteristic Curve for Artificial Data*



ROC curves can be used for two different, albeit related, purposes. First, one can determine the optimal cut-off point to be used for classification, defined as the point that minimizes the sum of the two conditional classification error frequencies $\hat{\alpha}+\hat{\beta}$. Graphically, it is the point that is obtained when the 45^o line is shifted in parallel to the northwest until it is just tangent to the ROC curve. In Figure 4.5, it is the point where $t = 0.16$, $\hat{\alpha} = 0$, $\hat{\beta} = 0.5$, and $\hat{\alpha}+\hat{\beta} = 0.5$.

Second, one can assess the goodness-of-fit of a given model. The benchmarks are a model with perfect prediction and a random model. A model with perfect prediction is one in which the ratings of the success group and the failure group do not overlap, i.e., the ratings of the successes are to the right of the ratings of the failures. In this case, perfect discrimination is possible. More specifically, there exists a value t such that $f(r \geq t|y = 1) = 1$ and $f(r \geq t|y = 0) = 0$. The associated ROC curve is determined by the points (0/0), (0/1), and (1/1) in Figure 4.5.

In the case of complete randomness, the rating distributions of the two groups are identical: successes do not tend to have higher rating values than failures. Clearly, in such a case the rating is uninformative for the classifica-

tion, and $f(r \geq t|y = 1) = f(r \geq t|y = 0)$. Therefore, $\hat{\alpha} + \hat{\beta} = 1 - f(r \geq t|y = 1) + f(r \geq t|y = 0) = 1$. In the ROC curve, the complete randomness is represented by the 45^o line. In practice, we observe intermediate cases in which the ROC curve is located somewhere above the diagonal in the graph. The more it moves to the northwest corner, the higher the discriminatory power.

A simple measure of fit is then the area under the ROC curve. It has a minimum of 0.5 and a maximum of 1. The area in our example, shown in Figure 4.5, is 0.7188. The area under the ROC curve is also a possible measure for selecting between two alternative models producing different ratings. There are two possibilities. First, one ROC curve dominates the other, i.e., lies to the northeast of it. Then the decision is unambiguous and we choose the model with the dominating ROC curve. Second, the two ROC curves intersect. Then one can select the model with the larger area. This approach is closely related to the problem of comparing two Lorenz curves and the Gini coefficients, respectively.

Example 4.5. Goodness-of-Fit in the Analysis of Being Childless

In our probit and logit analysis of the fertility decision (Table 4.1), we reported the likelihood ratio statistic of the null hypothesis that all coefficients are zero. Thus, we compared the full model with a constant-only model and we found significant grounds to reject the null hypothesis. This test on overall significance can be supplemented by the goodness-of-fit measures reported in Table 4.6.

Table 4.6. *Goodness-of-Fit Measures*

	Probit Model	Logit Model
McFadden's R^2	0.009	0.010
McKelvey and Zavoina's R^2	0.018	0.021
Area under the ROC Curve	0.5828	0.5829

These measures provide positive support for the logit model. However, it is important to note that measures of fit provide just a *rough* indication of the model that should be supported. They *do not give any evidence* that this model is optimal or best.

4.6 Non-Standard Sampling Schemes

4.6.1 Stratified Sampling

In some applications, the outcome of interest might be a rare event, hence, the proportion of ones in the sample is very low. Consider the following example.

Example 4.6. Dutch Migration to New Zealand

Hartog and Winkelmann (2003) analyze post-war Dutch migration to New Zealand. The Netherlands have a population of around 15 million, about two thirds of which are of working age. In 1986, there were approximately 20,000 Dutch working-age migrants, i.e., the ratio of working-age migrants to non-migrants is 1 to 500 and the unconditional probability of being a migrant is 0.2 percent. Thus, if a random sample of 10,000 is drawn, one would expect only 20 migrants to be included.

There is nothing wrong with this *per se*, and the ML estimator in the binary response model would be consistent. However, it turns out that one can obtain, for a given sample size (and thus given data collection cost), more precise parameter estimates if one oversamples the rare outcome and undersamples the frequent outcome. In the above example, one could base the analysis, for example, on 500 observations of migrants and 500 observations on non-migrants.

It is costly to collect data. Therefore, it may be preferable to oversample the group with the rare outcome – which we call stratification – rather than randomly sampling the two groups in the same proportion as in the population. Two cases of stratified sampling need to be distinguished:

- Exogenous stratification, based on the x's.
- Endogenous stratification, based on the y's.

4.6.2 Exogenous Stratification

To understand the consequences of **exogenous stratification** on the properties of the maximum likelihood estimator, consider the joint probability function for $f(y_i, x_i)$, rather than the conditional model $f(y_i|x_i)$ that we usually start with. We can always write the joint probability function as the product of a conditional and a marginal distribution, formally

$$f(y_i, x_i; \beta, \gamma) = f(y_i|x_i; \beta)h(x_i; \gamma)$$

for some marginal distribution $h(x_i; \gamma)$. Here, β and γ are two vectors of parameters characterizing the joint probability function. Now assume we can

separate the parameter vector of interest, β, which is only included in the conditional part, and the parameter vector γ as characterizing only the marginal distribution. For example, in the binary response model, we could have

$$f(y_i, x_i; \beta, \gamma) = \pi_i^{y_i}(1-\pi_i)^{1-y_i} h(x_i; \gamma)$$

where $\pi_i = G(x_i'\beta)$, as before. Assuming independent observations, the likelihood function can be written as the product over these joint probabilities and, after taking logs, we obtain the log-likelihood function

$$\log L(\beta, \gamma; y, x) = \sum_{i=1}^{n} \left[y_i \log \pi_i + (1-y_i) \log(1-\pi_i) + \log h(x_i; \gamma) \right]$$

Importantly, as long as the density of x_i does not depend on β, it does not affect the estimator at all. Hence, in the maximization of the log-likelihood function over β, the parameter vector of interest, we can neglect the term $\log h(x_i; \gamma)$ by the proportionality assumption. Moreover, we can draw data from a different distribution $h^*(x)$, and the estimator will still be the same. Hence, we see that *if* we can find explanatory variables that have a high predictive power for the probability of observing the rare event, then we can draw overproportionally from that segment of the population without invalidating the inference. The only problem is that such good predictors x may not always exist. In addition, it may not be possible to tell in advance what these regressors should be, before one actually collected the data and conducted the analysis.

4.6.3 Endogenous Stratification

Usually, sampling based on the outcome of the endogenous variable causes the maximum likelihood estimator to be seriously biased. However, in the logit model, the consequences of estimation based on an **endogenously stratified** sample are actually less severe. Consider the following situation: The outcome zero is rarely, the outcome one is frequently. To save data collection cost, observations with $y_i = 1$ are only included in the sample with probability p.

The joint probability for the event $y_i = 1$ and *observation included in the sample* is then

$$P(y_i = 1, in\ sample | x_i) = \frac{p \exp(x_i'\beta)}{1 + \exp(x_i'\beta)} = p\Lambda_i$$

The probability of a zero is not modified, so that

$$P(y_i = 0 | x_i) = \frac{1}{1 + \exp(x_i'\beta)} = 1 - \Lambda_i$$

The likelihood contribution is then

$$P(y_i = 1|in\ sample, x_i) = \frac{p\Lambda_i}{(1-\Lambda_i) + p\Lambda_i} = \frac{p\exp(x'\beta)}{1 + p\exp(x'\beta)}$$

which can be rewritten as

$$P(y_i = 1|in\ sample, x_i) = \frac{\exp(\ln p + x'\beta)}{1 + \exp(\ln p + x'\beta)} \tag{4.26}$$

where $\ln p < 0$. Similarly, we can obtain $P(y_i = 0|in\ sample, x_i)$. From (4.26) we can see that only the intercept term β_0 is modified. If p is known (as would be the case in the above example) all parameters of the original model (and therefore also all marginal effects) can be recovered from the selected sample. In particular, the following two steps are required:

1. Estimate a standard logit model using the reduced (endogenously stratified) sample.
2. Keep the slope coefficients and subtract $\ln p$ from the estimated constant $\hat{\beta}_0$ to obtain estimates for the full model.

What this result says is that if we have a sample with n_1 ones and $n - n_1$ zeros, where n_1 is large, we might as well draw a random subsample of the n_1 ones, with sampling probability p, and estimate the parameters of interest using the reduced sample. The procedure will be consistent. However, it will not be efficient to discard part of the observations with ones *if they are already collected*. The real relevance of this result refers to a situation as mentioned in the introduction: It is costless to observe y (and thus the fraction of ones and zeros in the underlying sample), but it is costly to collect the x variables. Then, it is preferable, on efficiency considerations, to stratify on y and collect data on x roughly on a 50/50 basis.

Two concluding remarks. First, this simple procedure only works for the logit model but not for the probit model or the linear probability model. In these other cases, the endogenous sample selection leads to a modified likelihood function that can no longer be estimated by the base model. Second, an intuition for the efficiency gain of a sampling scheme based on a 50/50 collection can be obtained by looking at the Hessian matrix of the logit log-likelihood. It involves the term $\hat{\pi}(1 - \hat{\pi})$ which becomes larger (and hence smaller when inverted) as the "average" $\hat{\pi}$ (the mean of the dependent variable in the logit case) approaches 0.5.

4.7 Further Exercises

Exercise 4.8 Consider ML estimation of the linear probability model

$$f(y_i|x_i) = (x_i'\beta)^{y_i}(1 - x_i'\beta)^{1-y_i} \qquad y_i = 0, 1$$

Derive the score vector of the log-likelihood function and show that the maximum likelihood estimator is equal to the GLS estimator. Which problems might arise?

Exercise 4.9 Suppose a binary response model that has been specified using a Weibull distribution with probabilities

$$\pi_i = 1 - \exp[-\exp(x_i'\beta)]$$

This specification is also known as a complementary log-log transformation, since we can write $\log[-\log(1 - \pi_i)] = x_i'\beta$.

a) Does this specification make sense? Compare to the probit and logit transformations.

b) Derive the marginal probability effects for x_{il}.

c) Write down the likelihood and the log-likelihood function for a sample of n independent observations.

d) Derive the score. How can you obtain an estimate of β?

Exercise 4.10 Consider a constant-only logit model, i.e., the probability of observing the outcome one is specified as $\pi = \Lambda(\beta) = [1 + \exp(-\beta)]^{-1}$.

a) Write down the likelihood and the log-likelihood function for a sample of n independent observations.

b) Derive the score function and find the ML estimator of β.

c) Derive the Hessian matrix and verify that $-E[H(\beta)] = n\pi(1 - \pi)$. Find the asymptotic distribution of $\hat{\beta}$?

d) Let $\bar{y} = 0.4$ and $n = 100$. Test the hypothesis that an individual chooses both alternatives with the same probability.
   (i) Formulate the null hypothesis in terms of the parameter β.
   (ii) Calculate the LR test statistic.
   (iii) Calculate the Score test statistic.
   (iv) Calculate the Wald test statistic. (Hint: Use the more general form of the Wald test, $Wald = (\hat{\beta} - \beta_0)\,\widehat{\mathrm{Var}}[\hat{\beta}]^{-1}\,(\hat{\beta} - \beta_0)$, where β_0 is the value of the parameter stated in the null hypothesis. This test statistic is asymptotically χ^2-distributed with degrees of freedom equal to the number of restrictions.)
   (v) Can you reject the null hypothesis at the 5% level of significance? What do you conclude?

Exercise 4.11 Suppose you are interested in estimating the effect of a dummy variable d on a binary response variable y. The joint probability function of y and d in a random sample of size $n = 100$ is as follows:

	$y = 0$	$y = 1$
$d = 0$	0.20	0.23
$d = 1$	0.38	0.19

a) Are y and d independent?

b) Obtain the maximum likelihood estimators for $\pi^1 = P(y_i = 1|d_i = 1)$ and $\pi^0 = P(y_i = 1|d_i = 0)$.

c) Now assume that the model has been specified as a logit model, where

$$\pi_i = P(y_i = 1|d_i) = \frac{\exp(\beta_0 + \beta_1 d_i)}{1 + \exp(\beta_0 + \beta_1 d_i)}$$

Exploit the ML invariance property to find the maximum likelihood estimators for β_0 and β_1.

Exercise 4.12 Consider again the table in Exercise 4.11 and assume that the relationship between y and d has been specified as a logit model as

$$\pi_i = P(y_i = 1|d_i) = \frac{\exp(\beta_0 + \beta_1 d_i)}{1 + \exp(\beta_0 + \beta_1 d_i)}$$

a) Find the likelihood and the log-likelihood function of the sample.

b) Find the score function and the Hessian matrix?

c) Obtain the ML estimates of β_0 and β_1 by solving the first-order conditions and using the information about the sample above.

d) Show that the ML estimates in c) indeed *maximize* the log-likelihood function in a).

e) Calculate the value of the log-likelihood function evaluated at the ML estimates.

f) Verify that the ML estimator of β_0 under the assumption $\beta_1 = 0$, the restricted ML estimator, is given by $\hat{\beta}_0^r = \log[\bar{y}/(1 - \bar{y})]$ where $\bar{y} = (\sum_{i=1}^n y_i)/n$. Calculate the ML estimate given the information about the sample above.

g) Calculate the value of the log-likelihood function evaluated at the restricted ML estimates.

h) Test the hypothesis $H_0 : \beta_1 = 0$ using a likelihood ratio test. Can you reject the null hypothesis at the 5% (1%) level of significance? What do you conclude?

Exercise 4.13 Evans and Schwab (1995) analyze the effect of attending a catholic school on graduation from high school. They obtain results of the form

$$P(\textit{high school graduate}) = \Phi[\, 1.2 + \underset{(0.056)}{0.777}\ \textit{catholic school} + \underset{(0.029)}{0.041}\ \textit{female} + \underset{(0.045)}{0.132}\ \textit{black}\,]$$

where asymptotic standard errors are reported in parentheses.

a) Write down the latent model that the estimation is based upon, including the assumptions for the error term.

b) What do the authors want to show by estimating such a model?

c) Test each of the three variables displayed for statistical significance. What do you conclude?

d) What is the predicted probability that a black female student attending a public high school will graduate?

e) How would the predicted probability of graduation change if the black female student went to a catholic high-school?

f) What is the interpretation of the difference between your answers in d) and e)?

g) Without doing the computations, would you expect that the "marginal effect" of catholic school is larger for black or for non-black students? Why?

h) Formulate an alternative model that would allow you to test whether the catholic school coefficient is the same for black and non-black students. Explain.

Exercise 4.14 Bantle and Haisken-DeNew (2002) report the following odds ratios from a logit regression. The dependent variable is 1 if the adolescent smokes, and 0 otherwise. The authors use data from the 1999 wave of the German Socio-Economic Panel. The estimates are based on a subset of $n = 830$ observations for 16- to 19-year-old youth who live with their parents.

	Odds Ratio	Standard Error
east german	1.4220	(0.2526)
gender	1.0762	(0.2116)
large city	0.8188	(0.3235)
only father smokes	3.5889	(0.6997)
only mother smokes	2.5097	(0.5851)
both mother and father smoke	4.4105	(0.7120)
further variables	⋮	

a) What do you think the authors want to prove or disprove?

b) How do you interpret the value of 4.4105 for the variable *both mother and father smoke*?

c) What is the odds ratio that results if a variable does not effect the probability of smoking?

d) Test whether *both mother and father smoke* has no effect on youth smoking.

e) Derive the logit coefficients of the model.

f) Can you compute from the information provided the probability that a young woman living in a small West German city with non-smoking parents smokes herself ? If so, how? If not, why not?

Exercise 4.15 Suppose you are interested in the survival probability of passengers aboard Titanic. The dataset *titanic.asc* (available on the author's homepage) contains information about 2201 passengers, including the gender, whether child or adult, the class traveled, and a binary variable indicating survival. A logit model gives the following results:

Dependent variable: *alive*		
	Estimate	Standard Error
male	-2.420	(0.140)
adult	-1.062	(0.244)
class 2 (1=yes)	-1.018	(0.195)
class 3 (1=yes)	-1.778	(0.172)
class 4 (1=yes)	-0.858	(0.157)
constant	3.105	(0.298)
Log-likelihood value		-1,105.03
Log-likelihood value (constant only model)		-1,384.72

The mean values are given by 0.95 (*male*), 0.79 (*adult*), 0.13 (*class 2*), 0.32 (*class 3*), 0.40 (*class 4*).

a) Which research question could you study with such a model?

b) Write down the formal model the estimation is based upon. What assumptions are made?

c) Use the invariance property and the Delta method to calculate estimates and to obtain standard errors of the odds ratios.

d) Interpret the results in the light of your research question using
   (i) discrete probability effects and predicted probabilities
   (ii) odds ratios

e) Conduct a LR test to test the full model against a constant-only model. What do you conclude?

Exercise 4.16 Use the data in *mroz.dta* to estimate a labor force participation model using the LPM, the probit and logit models. Include nonwife income, the years of schooling, a quadratic form in experience, age, and the number of children (younger than six years, and between six and 18 years) as explanatory variables.

a) Estimate the three models and report your results along with the value of the log-likelihood function (probit and logit), and an appropriate test statistic for the null hypothesis of a constant-only model.

b) How would you interpret the coefficient of years of schooling in the three models using
   (i) marginal probability effects
   (ii) discrete probability effects
   (iii) predicted probabilities

c) Does the presence of children affect the labor force participation of women? How can you quantify the effects (if any)?

d) Do the models have explanatory power compared to a constant-only model?

e) Why might education be an endogenous variable in the labor force participation?

f) Which variables might serve as instruments for the years of schooling? Test the null hypothesis that education is exogenous. What do you conclude?

Exercise 4.17 Consider a dataset with $n = 1,000$ pairs of observations (y_i, d_i), where y_i and d_i are binary variables. Assume that you have 326 observations with $y_i = 0$ and $d_i = 0$, 478 observations with $y_i = 1$ and $d_i = 0$, and 196 observations with $y_i = 1$ and $d_i = 1$. Which problem arises if you want to estimate a probit (logit) model of y on d in this dataset?

Exercise 4.18 True or False? Evaluate the following statements critically.

a) Let $\pi_i = \Phi(\beta_0+\beta_1 x_i+\beta_2 x_i^2)$ with continuous regressor x_i. The parameter β_1 measures the marginal probability effect of x_i on the probability of observing the outcome 1.

b) Let $\pi_i = \Phi(\beta_0 + \beta_1 x_i)$ with continuous regressor x_i and $\beta_1 > 0$. The change required in x_i to increase π_i from 0.4 to 0.5 is larger than the change required in x_i to increase π_i from 0.8 to 0.9.

c) Let $\pi_i = \Lambda(\beta_0 + \beta_1 d_i)$ with binary regressor d_i. The parameter β_1 measures the discrete probability effect of d_i on the probability of observing the outcome 1.

d) The linear probability model is more robust than the probit model, since we do not require normally distributed error terms.

Exercise 4.19 Suppose you want to analyze the determinants of secondary school choice in Germany. Use the data in *school.dta* and generate a binary variable, say *gym*, which equals one if the pupil attends *Gymnasium*, and zero if the pupil attends *Real-/Hauptschule*.

a) Define a research question that you could study with the data. Which explanatory variables would include in your analysis?

b) Estimate a probit or logit model in order to answer your research question, and report your results along with the value of the log-likelihood function, and an appropriate test statistic for the null hypothesis of a constant-only model.

c) Does the model have explanatory power compared to a constant-only model?

d) Interpret your results using marginal or discrete probability effects, and predicted probabilities.

e) Critically evaluate your results. Are they plausible? Which problems might arise? Do you have all the relevant information?

Exercise 4.20 Suppose you have a logit model with $\pi = P(y_i = 1|x_i) = \Lambda(x_i'\beta)$ in the population of interest. Now suppose that you do not have a random sample from the population, but the sample is drawn in the following way:

- With probability φ you draw an observation from the conditional distribution of (y_i, x_i) given $y_i = 1$.
- With probability $1 - \varphi$ you draw an observation from the conditional distribution of (y_i, x_i) given $y_i = 0$.

Furthermore, assume that x_i is a discrete random variable. Find $\tilde{\pi}_i = \tilde{P}(y_i = 1|x_i)$ in the sample.

5

Multinomial Response Models

5.1 Introduction

In this chapter, we turn our attention to probability models for the analysis of multinomial data. Recall from Chapter 1.2.1 that multinomial variables are characterized by a set of mutually exclusive and exhaustive non-ordered categories. Multinomial dependent variables appear in a number of microdata applications and we will start with some examples.

McFadden (1974b) considers three alternative travel modes (*car*, *bus*, and *train*) and investigates the demand behavior of urban commuters. Schmidt and Strauss (1975) analyze the effects of gender, race, education and labor market experience on occupational attainment. They divide individuals into the categories *menial*, *blue collar*, *craft*, *white collar*, and *professional occupations*. Terza (2002) shows that alcohol abuse significantly reduces the probability of being employed. In his study, the outcomes *employed*, *unemployed*, and *out of labor force* specify the multinomial character of an individual's employment status. Berger (1988) examines the relationship between predicted future earnings and the major field of study, distinguishing between *business*, *liberal arts*, *engineering*, *science*, and *education*. Alvarez et al. (2000) study the relevance of policy issues and the state of the economy in multiparty elections using the 1987 British general election, where British voters chose among the *Conservatives*, the *Social Democrat Party*, the *Liberals*, and the *Labour Party*. Other examples include the reasons for early retirement, industry affiliation, choice of insurance, choice of a certain brand of beer, or the portfolio structure of households (see Example 5.1 for possible outcomes).

In the previous chapter, we introduced the probit and logit models for binary outcomes. In the case of a multinomial response we can still make use of these models if we dichotomize the various outcomes. Consider, for example, the travel mode choice of urban commuters to go to work by car, bus, or train. In terms of binary outcomes, we could proceed in two ways. First, we could generate a binary variable indicating, say, public transport mode (bus or train). As a result, we would restrict our analysis to the choice

between private and public transport. Second, we could draw two subsamples in which we keep only observations choosing one of the alternatives car/bus (subsample one), or one of the alternatives car/train (subsample two), and recode the choice variable to 0/1. Begg and Gray (1986) show that estimating these two binary logits provides consistent estimates of the parameters of the corresponding multinomial model.

However, by dichotomizing a multinomial variable we "throw away" information in the data, and such an estimator cannot be efficient. Therefore, we want to explicitly model the choice across the whole range of outcomes.

Example 5.1. Multinomial Outcomes

- Reasons for early retirement
  (*company restructuring*, *health problems*, *personal reasons*)
- Insurance choice
  (*low deductible*, *high deductible*, *extra insurance*)
- Brand choice
  (*Becks*, *Krombacher*, *Heineken*, *Budweiser*, *Cardinal*)
- Industry affiliation
  (*agriculture*, *manufacturing*, *construction*, *trade*, *services*)
- Portfolio structure of households
  (*stocks only*, *stocks and bonds*, *bonds only*, *none*)

Chapter 5 is organized as follows. In Section 5.2, we introduce the multinomial logit model and give special attention to the interpretation of parameters. In Section 5.3, we generalize the multinomial logit model to choice-specific attributes. Within the random utility framework, we present the conditional logit model and discuss its advantages and disadvantages in economic applications. Finally, in Section 5.4, we consider several alternatives, namely the mixed logit, the multinomial probit, and the nested logit, which circumvent the problems of the basic models.

5.2 Multinomial Logit Model

5.2.1 Basic Model

The simplest multinomial response model is the **multinomial logit** (MNL), which is an extension of the binary logit to more than two response categories. The MNL model can be justified either in a mechanical way or within a discrete choice model. We restrict our attention to the former here, as the latter will be the subject of Section 5.16.

To begin with, let us suppose that there are J unordered outcomes of the dependent variable y_i. These outcomes are coded, without loss of generality, as $1, 2, \ldots, J$. For example, a commuter might choose among $J = 3$ alternative modes of transportation, including the options car (coded by 1), bus (2), and train (3). With multinomial data, the numerical coding is entirely arbitrary. We could also have chosen train (1), car (2), and bus (3), which reflects the unordered nature of multinomial data, or car (10), bus (27), and train (389). Assigning the values $1, \ldots, J$ to the outcomes simply contributes to notational simplicity. Let

$$P(y_i = j|x_i) = \pi_{ij} \tag{5.1}$$

denote the probability that individual i chooses alternative j, given her characteristics x_i, where x_i has dimension $(k+1)\times 1$, as before. For each individual, there are J such probabilities. For example, π_{i1} might be individual i's probability to commute by car, given her income, and π_{i2} and π_{i3} her probabilities to commute by bus and train, respectively.

Now we would like to specify the relationship between the probabilities and the vector of individual characteristics x_i, more specifically, a linear index $x_i'\beta_j$ with outcome-specific parameter vectors β_j. We use outcome-specific parameters to allow the effect of a change in one regressor to be different for each outcome probability. A sensible specification must observe that all probabilities lie between zero and one, and that they add up to unity. Suppose we let

$$\pi_{ij} = \frac{\exp(x_i'\beta_j)}{\sum_{r=1}^{J} \exp(x_i'\beta_r)} \qquad j = 1, \ldots, J \tag{5.2}$$

By definition of the exponential function, all probabilities are greater than zero, and, by construction, they fulfill the requirements of being less than one and adding up to unity over all categories. So far, the model has J parameter vectors $\beta_1, \ldots, \beta_J$, each of them with $k + 1$ elements. This leads to a total number of $(k + 1) \times J$ different parameters.

In fact, it turns out that there are too many parameters in the model and not all of them are identified. This can best be seen with some mathematical rearrangements. If we divide numerator *and* denominator in (5.2) by $\exp(x_i'\beta_1)$, and therefore do not change the probabilities, then we obtain

$$\pi_{i1} = \frac{1}{1 + \sum_{r=2}^{J} \exp(x_i'(\beta_r - \beta_1))}$$

$$\pi_{ij} = \frac{\exp(x_i'(\beta_j - \beta_1))}{1 + \sum_{r=2}^{J} \exp(x_i'(\beta_r - \beta_1))} \qquad j = 2, \ldots, J \tag{5.3}$$

Thus, the probabilities only depend on the *differences* between the parameter vectors. The levels themselves are unidentified since any constant can be added to each of the β's, such as $\beta_j + d$ instead of $\beta_j \; \forall j$, and this constant will cancel out in the differencing. Hence, we might as well let $\beta_1 \equiv 0$, i.e., the parameters of the first alternative are set to zero. This normalization yields the standard **multinomial logit model** as

$$\pi_{i1} = \frac{1}{1 + \sum_{r=2}^{J} \exp(x_i'\beta_r)}$$

$$\pi_{ij} = \frac{\exp(x_i'\beta_j)}{1 + \sum_{r=2}^{J} \exp(x_i'\beta_r)} \qquad j = 2, \ldots, J \tag{5.4}$$

Of course, the identification restriction $\beta_1 \equiv 0$ is chosen arbitrarily. As an alternative, we could have set any other β_j equal to zero. The category j with normalization $\beta_j \equiv 0$ is called the *base category* or *baseline*, which provides the "reference point" for all other alternatives. The choice of base category needs to be kept in mind when we interpret the model parameters.

5.2.2 Estimation

In order to proceed with ML estimation, we need to rewrite (5.4) as a conditional probability function. Remember that *each* individual makes exactly *one* choice since alternatives are mutually exclusive. In this case the multinomial probability function can be written as

$$f(y_i|x_i; \beta_2, \ldots, \beta_J) = (\pi_{i1})^{d_{i1}} (\pi_{i2})^{d_{i2}} \cdots (\pi_{iJ})^{d_{iJ}} = \prod_{j=1}^{J} (\pi_{ij})^{d_{ij}} \tag{5.5}$$

where the probabilities π_{ij} are given by (5.4), and d_{ij} is defined as a binary indicator with

$$d_{ij} = \begin{cases} 1 & \text{if individual } i \text{ chooses alternative } j \; (y_i = j) \\ 0 & \text{otherwise} \end{cases}$$

Hence, the probability function in (5.5) can be interpreted as individual i's probability of observing his or her actual response.

Assuming a sample of n independent pairs of observations (y_i, x_i), we can write the log-likelihood function of the sample as

$$\log L(\beta_2, \ldots, \beta_J; y, x) = \sum_{i=1}^{n} \sum_{j=1}^{J} d_{ij} \log \pi_{ij} \tag{5.6}$$

The necessary condition for a maximum of (5.6) requires us to set the score equal to zero. However, since each of the probabilities π_{ij} depends on the $J-1$ parameter vectors, the derivation of the score in the MNL model is more complicated than in the binary logit (see Chapter 4.4). We will go through its derivation step by step. The first derivative of (5.6) with respect to any of the parameter vectors β_s yields

$$\frac{\partial \log L(\beta_2, \ldots, \beta_J; y, x)}{\partial \beta_s} = \sum_{i=1}^{n} \sum_{j=1}^{J} d_{ij} \frac{1}{\pi_{ij}} \frac{\partial \pi_{ij}}{\partial \beta_s} \qquad s = 2, \ldots, J \tag{5.7}$$

(remember that $\beta_1 \equiv 0$). Now we need to know the partial derivatives of π_{ij} with respect to β_s. They are given by

$$\frac{\partial \pi_{ij}}{\partial \beta_s} = \begin{cases} -\pi_{ij}\pi_{is} x_i & s \neq j \\ \pi_{is}(1 - \pi_{is}) x_i & s = j \end{cases}$$

since we have to distinguish whether the vector β_s only appears in the denominator, or in both numerator and denominator of π_{ij}. We can plug these partial derivatives into (5.7) and obtain

$$\begin{aligned} \frac{\partial \log L(\beta_2, \ldots, \beta_J; y, x)}{\partial \beta_s} &= \sum_{i=1}^{n} \left(d_{is}(1 - \pi_{is}) x_i + \sum_{j \neq s} -d_{ij}\pi_{is} x_i \right) \\ &= \sum_{i=1}^{n} (d_{is} - \pi_{is}) x_i \qquad s = 2, \ldots, J \end{aligned} \tag{5.8}$$

The last equality follows since the d_{ij}'s add-up to unity, and therefore $d_{is}\pi_{is} + \sum_{j \neq s} d_{ij}\pi_{is} = \pi_{is}$. The score function as a vector of all first-order derivatives does not have a closed-form solution for the β's, and hence numerical methods are required to get the maximum, as presented in Chapter 3.5.2.

The sufficient condition for a maximum needs to be investigated by the second-order and cross-derivatives of the log-likelihood function, which yield the Hessian matrix

$$\begin{aligned} \frac{\partial^2 \log L(\beta_2, \ldots, \beta_J; y, x)}{\partial \beta_s \, \partial \beta_s'} &= -\sum_{i=1}^{n} \pi_{is}(1 - \pi_{is}) x_i x_i' \qquad s = 2, \ldots, J \\ \frac{\partial^2 \log L(\beta_2, \ldots, \beta_J; y, x)}{\partial \beta_s \partial \beta_t'} &= \sum_{i=1}^{n} \pi_{is}\pi_{it} x_i x_i' \qquad \forall s \neq t \end{aligned} \tag{5.9}$$

One can show that the Hessian is negative definite, which implies that the likelihood function is globally concave, ensuring the uniqueness of the maximum. If the model is correctly specified the ML estimator is consistent, efficient, and asymptotically normal. The covariance matrix of the ML estimator can be derived from the Hessian, which does not depend on y_i (or d_{ij}), and hence, its expected value is equal to the actual value. A consistent estimator of the covariance matrix can be obtained by inverting the Hessian evaluated at the ML estimates. Note that the Hessian is not block-diagonal. The cross-derivatives are non-zero and therefore, estimation of the joint model is not equivalent to estimating "dichotomized" logit models for one outcome at a time.

Example 5.2. Secondary School Choice

The secondary school choice of students in Germany is characterized by three mutually exclusive schooling tracks: *Hauptschule*, *Realschule*, and *Gymnasium*. The choice of schooling track is an important determinant of the child's future development, and it is argued that this choice is mainly driven by the parents and their educational level. In order to shed some empirical light on the mobility of educational attainment, we can use a MNL model with schooling track (*school*) as a dependent variable, and analyze how the mother's education affects secondary school choice. We can think of various other explanatory variables that should be controlled for, examples being mother's employment level, household income and household size, and parity. Using the data in *school.dta* we get the results in Table 5.1.

We choose *Hauptschule* as our base category. Hence, we obtain two vectors of parameters, one associated with *Realschule* and one associated with *Gymnasium*. Table 5.1 lists the ML estimates together with asymptotic standard errors calculated by the inverse of the Hessian matrix (evaluated at the ML estimates). We can use the information provided to test whether a coefficient is significantly different from zero. For example, we can construct a Wald test (or z-test) for the coefficient *mother's educational level* associated with the outcome *Realschule* and we obtain $z = 0.303/0.079 = 3.835$. The value of the test statistic is larger than the 1% critical value of the standard normal distribution for a two-sided test (which is 2.576), and therefore, the parameter is significantly different from zero.

Of course, we can extend the testing procedure to test whether more parameters are *jointly* different from zero by using a likelihood ratio test. For example, we can test the statistical significance of *all* parameters (except the constants). The appropriate LR test statistic is reported in Table 5.1 and has a value of 225.98. Since we compare the overall model with a constant-only model, we have to restrict 26 parameters to zero, and the test statistic is asymptotically χ^2-distributed with 26 degrees of freedom. The 1% critical value of this distribution is 45.64, which is much less than the value of the test statistic, and therefore we reject the constant-only model. Note that

Table 5.1. *Multinomial Logit Estimates of Secondary School Choice*

Dependent variable: *school*		
Base category: *Hauptschule* (1)	(2)	(3)
	Realschule	*Gymnasium*
mother's educational level in years	0.303	0.664
	(0.079)	(0.081)
mother's employment level (0/1)	0.363	0.454
	(0.160)	(0.168)
logarithmic household income	0.417	1.731
	(0.223)	(0.282)
logarithmic household size	-1.226	-1.633
	(0.446)	(0.484)
birth order	-0.118	-0.261
	(0.126)	(0.136)
constant	-6.913	-24.454
	(2.418)	(3.044)
Observations	675	
Log-likelihood value	-619.85	
LR (χ^2_{26})	225.98	

Notes: Standard errors in parentheses. Further controls: year dummies 1995-2002

constant-only in the context of the MNL model means that separate constants for *Realschule* and *Gymnasium* are allowed.

So far, we have neglected the interpretation of parameters. What does it mean if one parameter is positive or negative? What does it mean if the parameter takes a value of 0.303, as the parameter of mother's education in the *Realschule* equation, or 1.731, as the parameter of logarithmic household income in the *Gymnasium* equation? Moreover, does a positive sign of one coefficient imply that the explanatory variable positively affects the probability of choosing the associated alternative? This issue will be discussed next.

Exercise 5.1.

- Can you reject (at the 5% significance level) the hypothesis that the coefficient of birth order associated with *Gymnasium* is zero?
- Can you reject (at the 5% significance level) the hypothesis that the coefficients of birth order are *jointly* zero?
  (Hint: The log-likelihood value under this hypothesis is -621.72.)
- Calculate the pseudo R^2 measure proposed by McFadden (1974a).

5.2.3 Interpretation of Parameters

Similar to the probit and logit case, the parameters in the MNL model do not measure the effect of an explanatory variable on the outcome probabilities directly due to the nonlinear form. Moreover, we have $(J-1)\times(k+1)$ parameters in the model, which does not make interpretation especially easy. We begin with two important observations. First of all, it is essential to keep in mind which category is the base category. All coefficients need to be interpreted *relative* to the base category. And second, as with all discrete response models, it is generally easier to argue in terms of probabilities, either by calculating predictions or by investigating their changes (discrete or partial). Furthermore, like in the binary logit model, the odds (or odds ratios) provide an alternative means of interpretation.

The **odds** are in fact the simplest way to interpret the parameters in the MNL model. The odds of alternative j versus the base category (alternative 1) can be written as

$$\frac{\pi_{ij}}{\pi_{i1}} = \exp(x_i'\beta_j) \qquad j = 2,\ldots,J \tag{5.10}$$

since the denominator of the probabilities in (5.4) cancels out. With this result, the effect of an increase in the l-th explanatory variable by Δx_{il} on the odds can be expressed in terms of the odds before and after the change. Specifically, the **factor change** in the odds is

$$\frac{\exp(x_i'\beta_j + \Delta x_{il}\beta_{jl})}{\exp(x_i'\beta_j)} = \exp(\Delta x_{il}\beta_{jl})$$

where β_{jl} is the l-th element in the parameter vector β_j. It follows that for a unit change in x_{il}, $\Delta x_{il} = 1$, the odds of alternative j *relative* to the base category change by the factor $\exp(\beta_{jl})$, holding all other variables constant. As in the binary logit case, we call $\exp(\beta_{jl})$ the **odds ratio**. For example, a positive coefficient β_{jl} implies that the odds ratio is larger than one and that increasing x_{il} raises the probability of category j *relative* to the probability of the base category. Thus, information on the sign of the coefficient alone is uninformative; we also need to know the base category.

We can generalize the discussion to the odds of any two outcomes j versus m and obtain

$$\frac{\pi_{ij}}{\pi_{im}} = \exp[x_i'(\beta_j - \beta_m)] \qquad \forall j \neq m \tag{5.11}$$

with factor change $\exp[\Delta x_{il}(\beta_{jl}-\beta_{ml})]$ for a change Δx_{il} in the l-th element of x_i, and odds ratio $\exp(\beta_{jl}-\beta_{ml})$. Hence, increasing x_{il} raises the probability of alternative j *relative* to the probability of alternative m if, and only if, the parameter associated with alternative j, β_{jl}, is larger than the parameter associated with alternative m, β_{ml}. Therefore, in the MNL model we always obtain *relative* statements.

If we consider the **relative change** in the odds of alternative j versus the base category, then we obtain

$$\frac{\exp(x_i'\beta_j + \Delta x_{il}\beta_{jl}) - \exp(x_i'\beta_j)}{\exp(x_i'\beta_j)} = \exp(\Delta x_{il}\beta_{jl}) - 1$$

which simplifies to $\exp(\beta_{jl}) - 1$ with a unit increase in x_{il}, $\Delta x_{il} = 1$. Moreover, for small values of β_{jl}, we have $\exp(\beta_{jl}) - 1 \approx \beta_{jl}$. This is an attractive feature of the odds since we can interpret parameters directly.

A final observation about the odds in the MNL model is that taking logs in equation (5.11) shows linearity in the logarithmic odds. Historically, the MNL model was developed in the biometrics and statistics literature by modeling these logarithmic odds (see for example Gurland, Lee and Dahm, 1960, for a special case of the MNL model); early references can also be found in the social sciences (Theil, 1969, 1970).

Exercise 5.2.

- Provide a formula that generalizes the *relative change* in the odds to the comparison of any two alternatives j and m, $\forall j \neq m$.

While the interpretation of parameters in terms of odds is relatively simple, we lose important information which is, in principle, directly available and which may be of interest in empirical applications. More specifically, the odds only provide information about the *ratio* of probabilities (or the changes therein), but not about their *levels*. Consider the following Example 5.3.

Example 5.3. Odds versus Probabilities

Suppose we have three alternatives – car (1), bus (2) and train (3) – with the odds "car versus train" given by $p_2/p_1 = 3/4$, and "car versus bus" given by $p_3/p_1 = 3/4$. Since the probabilities must add up to unity, we can solve for $p_1 = 0.4$ and $p_2 = p_3 = 0.3$. Now assume that a unit increase in one regressor changes the odds to $p_2/p_1 = 3/5$ and $p_3/p_1 = 2/5$. Again under the adding-up restriction, we can solve for the probabilities and obtain $p_1 = 0.5$, $p_2 = 0.3$, and $p_3 = 0.2$. This implies that the probability of choosing to go by car increases by 0.1, whereas the probability of choosing the train decreases by 0.1, and p_2 remains the same.

Exercise 5.3.

Consider Example 5.3 again. Let alternative car be the base category, and denote the unit increase in one regressor, say the l-th element in x_i, by $\Delta x_{il} = 1$.

- Calculate the odds ratios $\exp(\beta_{2l})$ and $\exp(\beta_{3l})$ by comparing the odds before and after the change.
- Calculate the odds of bus versus train.

In the simple case of Example 5.3 with $J = 3$ alternatives, the calculation of probability levels (and their changes) from the odds is tractable, but with increasing J it becomes cumbersome. Hence, it is desirable to interpret the parameters directly in terms of the probabilities and their changes (discrete and partials). Unfortunately, the relationship between the parameters and the marginal or discrete changes of probabilities in the MNL model is not as straightforward as in the binary logit model.

First, consider the **discrete change** in the probabilities due to the change in the l-th element in x_i by Δx_{il}. Formally, this can be expressed by

$$\begin{aligned}\Delta\pi_{ij} &= P(y_i = j|x_i + \Delta x_{il}) - P(y_i = j|x_i) \qquad j = 2, \ldots, J \\ &= \frac{\exp(x_i'\beta_j + \Delta x_{il}\beta_{jl})}{1 + \sum_{r=2}^{J} \exp(x_i'\beta_r + \Delta x_{il}\beta_{rl})} - \frac{\exp(x_i'\beta_j)}{1 + \sum_{r=2}^{J} \exp(x_i'\beta_r)}\end{aligned} \tag{5.12}$$

The interpretation of (5.12) is as follows: Given a *ceteris paribus* change in x_{il} by Δx_{il}, the probability of observing outcome j changes by $\Delta\pi_{ij}$. However, the sign of the probability effect does not need to coincide with the sign of the parameter (as in the binary logit model). For example, if the l-th element in β_j is positive, the numerator increases with increasing x_{il}. But the denominator may increase even more – depending on the magnitude of $\beta_{rl}, \forall r \neq j$. Hence, the direction of the overall effect is ambiguous.

Since the sum over all probabilities must equal one, the change in one probability is determined by the $J-1$ other changes. Hence, we can write the change in π_{i1} simply as

$$\Delta\pi_{i1} = -\sum_{j=2}^{J} \Delta\pi_{ij} \tag{5.13}$$

In order to obtain per-unit changes, we can divide the overall changes by Δx_{il}. Note that the exact amount of change depends on the specific values to which we fix the regressors. We already discussed this issue in Chapter 4 and do not consider it further here.

In the case of a continuous regressor, it might be more convenient to calculate the marginal change in the probabilities. Algebraically, we compute the **marginal probability effect** (MPE) by taking the first derivative with respect to the variable of interest, and obtain

$$MPE_{i1l} = \frac{\partial \pi_{i1}}{\partial x_{il}} = -\pi_{i1} \sum_{r=2}^{J} \pi_{ir} \beta_{rl}$$

$$MPE_{ijl} = \frac{\partial \pi_{ij}}{\partial x_{il}} = \pi_{ij} \left[\beta_{jl} - \sum_{r=2}^{J} \pi_{ir} \beta_{rl} \right] \qquad j = 2, \ldots, J \tag{5.14}$$

where MPE_{ijl} denotes the marginal probability effect (MPE) for individual i of choosing alternative j if the l-th element in x_i is increased by a small unit. We recognize from (5.14) that the MPE of one outcome generally depends on the probabilities (and parameters) of all other outcomes. For example, if β_{jl} is positive, then the MPE does not need to be positive either, and hence, its sign is again ambiguous.

Exercise 5.4.

- Show that $\sum_{j=1}^{J} MPE_{ijl} = 0$.

In order to obtain predicted probabilities, we can make use of the invariance property of ML estimation. Since the probabilities in (5.4) are a function of the unknown parameters, and are therefore unknown themselves, we have to estimate them. With help of the invariance property, this reduces to simply replacing the true parameters by their ML estimates. Furthermore, we can calculate standard errors of the prediction using the Delta method, as discussed in detail in Chapters 3 and 4. If we consider individual predictions, we can summarize the information by computing the mean and the standard deviation, or the minimum and maximum predictions. We can also plot the predicted probabilities at various levels of one explanatory variable, fixing the others to their means (or other interesting values), and so we can examine the effect of this variable on the probabilities.

In the same manner, i.e., using the invariance property and the Delta method, we can estimate the odds ratios, as well as discrete and marginal probability effects. Since probability effects depend on x_i, we should summarize the information either by calculating the expected effects, or by calculating the effect evaluated at the expected characteristics. The former can be estimated by the **average marginal probability effects**

$$\widehat{AMPE}_{jl} = \frac{1}{n} \sum_{i=1}^{n} \widehat{MPE}_{ijl} \qquad j = 1, \ldots, J \tag{5.15}$$

where $\widehat{MPE}_{ijl}$ are the MPE's defined in (5.14) evaluated at the ML estimates $\hat{\beta}_j$, $\forall j$. The latter are calculated by using (5.14) and replacing the vector of characteristics x_i by its mean values $\mathrm{E}(x_i)$, which can be estimated by the sample average over the x_i's.

Example 5.4. Secondary School Choice

We now want to analyze the estimation results of Example 5.2, in particular the parameter estimates in Table 5.1. We focus on the intergenerational transmission of educational attainment, i.e., we want to know how an increase in mother's educational level changes, *ceteris paribus*, the probabilities of attending *Hauptschule*, *Realschule*, and *Gymnasium*. From the discussion above, we can investigate this relationship by means of predicted probabilities (or the changes therein) as well as the odds ratios. The following tables, 5.2 and 5.3, report the measures.

Table 5.2. *Predicted Probabilities and Mother's Educational Level*

mother's educational level in years	(1) *Hauptschule*	(2) *Realschule*	(3) *Gymnasium*
7	0.6959	0.2448	0.0594
9	0.5085	0.3279	0.1636
10	0.4003	0.3495	0.2502
10.5	0.3464	0.3519	0.3017
11	0.2945	0.3481	0.3574
11.5	0.2459	0.3383	0.4159
12	0.2017	0.3229	0.4754
13	0.1291	0.2799	0.5910
13.5	0.1010	0.2548	0.6443
14	0.0780	0.2289	0.6932
14.5	0.0595	0.2033	0.7372
15	0.0450	0.1787	0.7763
16	0.0251	0.1348	0.8401
18	0.0073	0.0719	0.9209

Note: All other explanatory variables are fixed at their means.

In Table 5.2 we list the predicted probabilities for each level of mother's educational level. We can see that a child has a 69.6 percent probability of attending *Hauptschule* if the mother has seven years of schooling, and if all other explanatory variables are fixed at their means. With increasing levels of education, the probability decreases monotonically from 69.6 to 0.7 percent, *ceteris paribus*. The opposite holds for the child's probability of attending *Gymnasium*, which is at its highest 92.1 percent, if the child's mother has 18 years of schooling (which corresponds to a doctorate). The probability of

attending *Realschule* remains at around 30 percent with mother's educational level between seven and 14 years, but then decreases to 7.2 percent if the child's mother has a tertiary education. Hence, there seems to be empirical evidence for intergenerational transmission of education. Although Table 5.2 provides an in-depth analysis of this relationship, the tabulation of predicted probabilities by various levels of an explanatory variable becomes intractable with large J and many distinct values of the regressor. We summarize the information in the first part of Table 5.3.

Table 5.3. *The Effect of Mother's Education on Secondary School Choice*

	(1) *Hauptschule*	(2) *Realschule*	(3) *Gymnasium*
Discrete Changes in the Predicted Probabilities			
min → max	-0.6886	-0.1729	0.8615
9 years → 10 years	-0.1082	0.0216	0.0865
10 years → 13 years	-0.2712	-0.0697	0.3408
13 years → 16 years	-0.1041	-0.1450	0.2491
Marginal Changes in the Predicted Probabilities			
$\widehat{MPE}$	-0.0941	-0.0243	0.1183
$\widehat{AMPE}$	-0.0800	-0.0091	0.0891
Odds Ratios			
		1.3540	1.9420
		(0.1072)	(0.1582)

Notes: The AMPE is calculated as in (5.15). All other effects are calculated by fixing the explanatory variables at their means. For illustration purposes, we report the standard errors (in parentheses) only for the odds ratio.

In particular, the change in mother's educational level from 13 to 16 years, which can be interpreted as the effect of an university degree, yields a decrease in the probability of attending *Hauptschule* or *Realschule* by 24.9 **percentage points**. The marginal changes are calculated in two ways, as the marginal probability effect evaluated at the means, and as the average MPE. We can see that the two measures differ slightly because of the nonlinearity in (5.14). A value of 0.0891, for example, is interpreted as follows: If we increase the years of schooling by one unit, then the probability of attending *Gymnasium* increases on average by about 8.91 percentage points.

The odds ratios for comparing *Realschule* with *Hauptschule*, as well as *Gymnasium* with *Hauptschule*, are significantly larger than one. Both can be tested using a Wald test. We have, for example, $z = (1.3540 - 1)/0.1072 = 3.3022$, which is larger than 1.96, the 5% critical value of the standard normal for a two-sided test. From the point estimates, we conclude that with one more year of schooling by the child's mother, the probability of attend-

ing middle or upper secondary school increases *relative* to the probability of attending lower secondary school. Moreover, the factor change is larger when comparing *Gymnasium* with *Hauptschule* than when comparing *Realschule* with *Hauptschule*.

Exercise 5.5.

- What exactly is the difference in the interpretation of $\widehat{MPE}$ and $\widehat{AMPE}$ in Table 5.3?
- Calculate the odds ratio for the comparison of *Realschule* versus *Gymnasium*.
- Calculate the odds ratios for a change in mother's employment level from 0 to 1 for all possible comparisons.

5.3 Conditional Logit Model

5.3.1 Introduction

The discussion so far has neglected an important aspect of multinomial response modeling. In many applications, one is interested in how attributes of the *choice alternative* affect the individual decision. A typical example appears in studies of travel demand behavior. Let the alternatives again be car, bus, and train, and consider the **choice-specific attributes**

- travel time and
- cost of transport

These attributes depend on the alternative as well as on the individual him-/herself. For example, the travel time for individual i when commuting to work by car might be 20 minutes, whereas by train, the travel time might be 35 minutes. Moreover, the travel time varies across individuals, for instance due to the distance between home and workplace, and therefore another individual might have travel times of 45 minutes by car and 40 minutes by train, respectively. We denote choice-specific attributes by z_{ij} with subscript ij to stress the two-dimensional variation. The standard multinomial logit model, as presented in the previous section, only allows for individual-specific characteristics x_i (such as income and gender), which do not vary across alternatives, and therefore only have subscript i.

In order to generalize the MNL model, we now demonstrate how we may formulate a general model of choice including both individual characteristics

and choice-specific attributes. The analysis of multinomial responses indeed has its origin within the framework of **discrete choice models**, in particular in the **random utility model** introduced by Thurstone (1927) and studied by Marschak (1960) and Block and Marschak (1960), and in **Luce's theory of individual choice behavior** (Luce, 1959). McFadden (1968, 1974a) shows how Luce's model can be parameterized to obtain a statistical model that can be used to analyze individual choice data, and calls this the **conditional logit** (CL). This initiated an enormous literature on discrete choice modeling, predominantly with applications in studies of travel demand behavior (see McFadden, 1974b, or Ben-Akiva and Lerman, 1985).

5.3.2 General Model of Choice

In Chapter 4.3, we motivated binary responses within the framework of discrete choice models. The basic idea was to assume the existence of a latent variable $\mathcal{U}(z_{ij}, x_i)$, the indirect **utility function** of individual i when choosing alternative j, which is a function of individual characteristics x_i and choice-specific attributes z_{ij}. We restricted our attention to $J = 2$ alternatives, however, the model can be readily extended to an arbitrary number of alternatives $J \geq 2$. Assume that we specify a linear utility function as in equation (4.15)

$$\mathcal{U}_{ij} = \mathcal{U}(z_{ij}, x_i) = z'_{ij}\gamma + x'_i\beta_j + u_{ij} \qquad j = 1, \ldots, J \tag{5.16}$$

with systematic part $\mu_{ij} = z'_{ij}\gamma + x'_i\beta_j$ and additive error term u_{ij}. As before, the random error might capture partial ignorance of the econometrician as well as intrinsic randomness in an individual's behavior. Under **(random) utility maximization**, an individual i chooses alternative j if, and only if, the utility of alternative j is the largest of all utilities, formally $\mathcal{U}_{ij} = \max(\mathcal{U}_{i1}, \ldots, \mathcal{U}_{iJ})$. Making a parametric assumption on the error terms u_{ij} produces a parametric probability law of the multinomial outcomes. However, since

$$\begin{aligned} \pi_{ij} &= P(y_i = j | x_i, z_{i1}, \ldots, z_{iJ}) \qquad j = 1, \ldots, J \\ &= P(\mathcal{U}_{ij} > \mathcal{U}_{im}, \forall m \neq j | x_i, z_{i1}, \ldots, z_{iJ}) \end{aligned} \tag{5.17}$$

it is difficult to find a parametric distribution that leads to simple algebraic forms and does not require multiple integration. Fortunately, it can be shown (see for example Maddala, 1983: pp. 60/61) that if the u_{ij}'s ($j = 1, \ldots, J$) are *independently* and identically **type-I extreme value** distributed with density function $f(u) = \exp(-u - \exp(-u))$, then

$$\pi_{ij} = \frac{\exp(\mu_{ij})}{\sum_{r=1}^{J} \exp(\mu_{ir})} \qquad j = 1, \ldots, J \tag{5.18}$$

In the special case of $J = 2$ alternatives, individual i chooses alternative 1 if, and only if, $\mathcal{U}_{i1} > \mathcal{U}_{i2}$. Under the assumption of independent type-I extreme value distributed error terms, the resulting model is the binary logit model, as stated already in equation 4.18.

5.3.3 Modeling Conditional Logits

From the general model in (5.18) we can further investigate the relationship between outcome probabilities and explanatory variables. Recall that μ_{ij} is the systematic part of the utility function, which comprises a linear index of individual characteristics $x_i'\beta_j$ *and* a linear index of choice-specific attributes $z_{ij}'\gamma$. First, consider the case in which **only choice-specific attributes** are available ($\beta_j = 0\ \forall j$), such that $\mu_{ij} = z_{ij}'\gamma$. For example, we want to explain the demand for various travel modes (car, bus, train) by only including travel time and travel cost as regressors. Then, the relevant probability model is given by

$$\pi_{ij} = P(y_i = j|z_{i1}, \ldots, z_{iJ}) = \frac{\exp(z_{ij}'\gamma)}{\sum_{r=1}^{J} \exp(z_{ir}'\gamma)} \qquad j = 1, \ldots, J \tag{5.19}$$

This model is called the **conditional logit model** as proposed by McFadden (1968, 1974a). The model is named conditional logit (CL) since it reduces to a logit model in the case of two alternatives, and its specific form is reminiscent of the form of conditional probabilities. In principle, the same argument holds for (5.18), but we will follow the usual terminology and call only (5.19) the CL model.

It is useful for understanding the CL model to rewrite the probabilities in (5.19) by dividing both numerator and denominator by $\exp(z_{i1}'\gamma)$. This transformation does not change the probabilities and we obtain

$$\pi_{ij} = \frac{\exp[(z_{ij} - z_{i1})'\gamma]}{\sum_{r=1}^{J} \exp[(z_{ir} - z_{i1})'\gamma]} \qquad j = 1, \ldots, J \tag{5.20}$$

where $z_{ij} - z_{i1}$ is a comparison term of the choice attribute z_{ij} relative to z_{i1}. Equation (5.20) has an interesting interpretation. If each choice has the same attribute, i.e., $z_{i1} = \ldots = z_{iJ}$, then alternative j is chosen with probability $1/J$. This simply reflects the idea that we have no variation to explain preference for one of the alternatives, and hence, alternatives are chosen randomly with equal probabilities. Moreover, we can introduce a set of choice specific intercepts α_j with $\alpha_1 = 0$. These intercepts measure the *relative* preference for the choice of alternative j compared to alternative 1, given that all attributes were the same. We will show later how we can include these choice-specific intercepts in the vector z_{ij} by defining a set of dummy variables indicating alternatives (see Table 5.5) and how exactly we interpret the parameters γ.

In the CL model, the parameters are no longer choice-specific (γ carries no subscript j). No normalization of γ is required. This is a natural approach, since choices are based on the attributes of the alternatives, and only these attributes matter. For instance, we want to learn something about by how much a *ceteris paribus* reduction in the travel time by bus, e.g., through the

introduction of an express bus service, would affect the probability that this travel mode is selected by an individual. Furthermore, we should be perfectly aware of the implication of the *ceteris paribus* condition which means *given all other travel times are unchanged*.

The following example, 5.5, illustrates how the data have to be arranged in the CL model. In addition, we make clear how individual-specific characteristics have to be distinguished from choice-specific attributes.

Example 5.5. Data Organization in the Conditional Logit Model

Consider a subsample of the data used in Paap and Franses (2000) to analyze the impact of marketing-mix variables on the choice of cracker brands. Table 5.4 lists the choices made at each of the purchase occasions, i.e., instances where a customer bought one of 4 cracker types available in the local supermarket. The choice includes the three major cracker brands *Sunshine*, *Keebler*, and *Nabisco* together with a category *Private Label*, which comprises various local brands. The chosen alternative is indicated by a "1". We report the different prices as well as a dummy variable equaling one if the cracker brand was on display. These variables vary across purchases *and* alternatives. In contrast to this, an individual-specific characteristic like gender does not vary across alternatives.

Table 5.4. *Cracker Brand Choice*

i	Cracker Brand	*brand choice* (0/1)	*price*	*on display* (0/1)	*male* (0/1)
1	*Sunshine*	0	1.03	0	1
1	*Keebler*	0	1.09	0	1
1	*Nabisco*	1	0.89	0	1
1	*Private Label*	0	0.78	0	1
2	*Sunshine*	1	0.69	1	0
2	*Keebler*	0	1.05	0	0
2	*Nabisco*	0	0.89	1	0
2	*Private Label*	0	0.65	0	0
3	*Sunshine*	0	1.05	0	0
3	*Keebler*	1	0.99	1	0
3	*Nabisco*	0	1.29	0	0
3	*Private Label*	0	0.59	0	0

Notes: The original dataset contains information on 3292 purchases of crackers. For illustration purposes we display only three purchases and a hypothetical variable *male*.

Now assume that **only individual-specific characteristics** are available ($\gamma = 0$) such that $\mu_{ij} = x_i'\beta_j$. For example, we want to analyze travel demand behavior just by using the explanatory variables income, gender and age. In this case the general model reduces to

$$\pi_{ij} = P(y_i = j|x_i) = \frac{\exp(x_i'\beta_j)}{\sum_{r=1}^{J} \exp(x_i'\beta_r)} \qquad j = 1, \ldots, J$$

which suffers from the same identification problem as the probabilities stated in (5.2). Hence, we need a normalization on the parameters, and we let $\beta_1 \equiv 0$. This yields the MNL model that we have already analyzed in Section 5.2.

In many applications, **both kind of regressors**, choice-specific attributes and individual characteristics, are available and of interest. Therefore, it is useful to formulate a **hybrid model** of MNL and CL in which we have the systematic part $\mu_{ij} = z_{ij}'\gamma + x_i'\beta_j$. From (5.18), the probabilities are given by

$$\pi_{ij} = P(y_i = j|x_i, z_{i1}, \ldots, z_{iJ}) = \frac{\exp(z_{ij}'\gamma + x_i'\beta_j)}{\sum_{r=1}^{J} \exp(z_{ir}'\gamma + x_i'\beta_r)} \tag{5.21}$$

for $j = 1, \ldots, J$. However, in this form the probabilities are unidentified since we can add any constant d to each of the β_j's and this term would cancel out without changing the probabilities. Hence, we need again, as in the MNL model, a normalization to identify all parameters. One possibility is to write the probabilities as follows

$$\begin{aligned}
\pi_{i1} &= \frac{1}{1 + \sum_{r=2}^{J} \exp[(z_{ir} - z_{i1})'\gamma + x_i'\beta_r]} \\
\pi_{ij} &= \frac{\exp[(z_{ij} - z_{i1})'\gamma + x_i'\beta_j]}{1 + \sum_{r=2}^{J} \exp[(z_{ir} - z_{i1})'\gamma + x_i'\beta_r]} \qquad j = 2, \ldots, J
\end{aligned} \tag{5.22}$$

The probabilities in (5.22) are derived by dividing (5.21) by $\exp(z_{i1}'\gamma + x_i'\beta_1)$ and setting $\beta_1 \equiv 0$. With this normalization, alternative 1 is called the base category, baseline, or reference category.

A second possibility of combining the two types of regressors, x_i and z_{ij}, is to create a set of dummy variables, one for each alternative, and then multiply them by the individual-specific characteristics. These interaction terms are then choice-specific as well, and we can reduce the model to the form of the CL model, where the z_{ij}'s contain (true) choice-specific attributes as well as interaction terms. For identification, we have to exclude the interaction terms of one category, the base category. Table 5.5 shows how the data have to be manipulated in the case of $J = 5$ alternatives and one individual-specific characteristic, denoted by x_{il}.

Table 5.5. *Data Transformation in the Conditional Logit Model*

i	Alternatives (j)	d_1	d_2	d_3	d_4	d_5	x_{il}	$x_{il}d_2$	$x_{il}d_3$	$x_{il}d_4$	$x_{il}d_5$
1	1	1	0	0	0	0	20	0	0	0	0
1	2	0	1	0	0	0	20	20	0	0	0
1	3	0	0	1	0	0	20	0	20	0	0
1	4	0	0	0	1	0	20	0	0	20	0
1	5	0	0	0	0	1	20	0	0	0	20

The dummy variables d_m, $m = 1, \ldots, J$, indicate whether $m = j$. The individual-specific regressor x_i has the hypothetical value 20 for each alternative. By multiplying x_{il} by d_m, $m = 2, \ldots, 5$, we obtain $J - 1 = 4$ choice-specific attributes. Now it becomes clear why we have to exclude one interaction term, here the interaction with d_1. If we included this term in the regression, then the term $\sum_m x_i d_m = 20$ would be constant across alternatives, and therefore could not be distinguished from a constant in the model. The category for which we drop the interaction term is called the base category.

In principle, we can generate these interaction terms with a whole vector of explanatory variables x_i. Moreover, if x_i includes a constant term, which we usually assume, then the interactions will generate the choice-specific intercepts that we introduced earlier when we discussed the CL probabilities in (5.20). Hence, in order to estimate a MNL within the framework of the CL model we have to extend the vector of (true) choice-specific attributes, z_{ij}, by all interactions $x_i d_m$ for $m = 2, \ldots, J$. In this way, the MNL model can be seen as a special case of the CL model.

ML estimation of the parameters in the CL model is basically the same as in the MNL model. The only thing we have to change, compared to the derivation of the log-likelihood function in Section 5.2.2, is to replace the MNL probabilities by the probabilities of the CL model. Assuming an independent sample of size n with observations $(y_i, z_{i1}, \ldots, z_{iJ})$, we obtain the log-likelihood function $\log L(\gamma; y, z_1, \ldots, z_J)$, the score function $s(\gamma; y, z_1, \ldots, z_J)$, and the Hessian matrix $H(\gamma; y, z_1, \ldots, z_J)$ in a straightforward manner. Inference can be based on the well-known ML properties (asymptotic normal distribution, invariance).

5.3.4 Interpretation of Parameters

Although the response probabilities in the CL and in the MNL have a similar form, the interpretation of parameters in both models is very different. Consider, for example, the **odds**. In the CL model, the general odds of comparing two alternatives j and m are given by

$$\frac{\pi_{ij}}{\pi_{im}} = \exp[(z_{ij} - z_{im})'\gamma] \qquad m \neq j \tag{5.23}$$

In (5.23), we can see that in the CL model the values of the regressors differ but the parameter γ is the same for each possible comparison. In other words,

the *difference in attributes* $(z_{ij} - z_{im})$ is what matters. This has to be distinguished from the general odds in the MNL model, equation (5.11), in which the individual-specific characteristics do not vary across outcomes (x_i has no subscript j), but the parameters might differ for each category. Hence, in the MNL model the *difference in parameters* $\beta_j - \beta_m$ is what matters.

Exercise 5.6.

- Derive the general odds in the model with both individual-specific characteristics and choice-specific attributes, equation (5.22). How does your result compare to the odds in the MNL and CL model?

Likewise, we can compare the marginal probability effects in the MNL and in the CL model. In the latter, the **marginal probability effect** of the l-th element in z_{ij} on π_{ij} is given by

$$MPE_{ijjl} = \frac{\partial \pi_{ij}}{\partial z_{ijl}} = \pi_{ij}(1 - \pi_{ij})\gamma_l \tag{5.24}$$

The MPE_{ijjl} describes the marginal change in the probability of outcome j accorded to a marginal increase in the l-th attribute of the *same* alternative j. It may also be of interest how the probability π_{ij} marginally changes if the l-th element of any other alternative m is increased by a small unit. This can be calculated as

$$MPE_{ijml} = \frac{\partial \pi_{ij}}{\partial z_{iml}} = -\pi_{ij}\pi_{im}\gamma_l \qquad m \neq j \tag{5.25}$$

Unlike in the MNL model, the sign of the parameters already gives information about the sign of the marginal probability effects. In particular, the sign of γ_l is the same as the sign of MPE_{ijjl}, e.g., if we estimate $\hat{\gamma}_l > 0$ then MPE_{ijjl} is positive as well. In contrast, γ_l and MPE_{ijml} are opposite in sign to each other.

Predicted probabilities in the CL model are obtained, as before, by evaluating (5.19) at the ML estimates and specific values of z. The same applies if we want to estimate the odds or the MPE's and we can use the delta method to obtain standard errors of these measures. Based on the CL model we can predict the choice probability for a *category* h that is not considered in the estimation procedure, but for which we know a vector of attributes z_{ih}. This has to be distinguished from the MNL model, which tells us something about the choice probability of a new *individual* with characteristics x_i.

Example 5.6. Cracker Brand Choice: Marketing and Preferences

Paap and Franses (2000) analyze the impact of marketing-mix variables on the choice of cracker brands. In a sample of 3,292 purchases of crackers, the choice is among three major cracker brands – *Sunshine*, *Keebler*, and *Nabisco* – and various local brands that are collected under *Private Label*. The dataset contains information about the chosen brand in each purchase (*brand choice*) as well as the *price* of all crackers and two binary variables indicating whether the cracker brand was on special *display* in the shop or *featured* in an advertisement in the newspaper. This information is available for each purchase, and hence we have three choice specific attributes.

Suppose we want to know whether marketing-mix variables have an important impact on brand choice, or whether purchases are mainly driven by preferences for a certain brand. Table 5.6 reports results of a CL analysis that can be used to answer this research question.

Table 5.6. *Cracker Brand Choice: A Conditional Logit Analysis*

Dependent Variable: *brand choice*			
	Model 1	Model 2	Model 3
price	-0.969	-3.448	-3.332
	(0.084)	(0.202)	(0.205)
sunshine		-0.559	-0.608
		(0.088)	(0.090)
keebler		-0.008	-0.066
		(0.115)	(0.116)
nabisco		1.965	1.886
		(0.094)	(0.099)
featured			0.259
			(0.101)
on display			0.144
			(0.061)
Log-likelihood value	-4,498.52	-3,364.90	-3,358.05
LR (H_0 : constant-only model)	130.32	2,397.56	2,411.27
Observations	3,292	3,292	3,292

Model 1 starts with a simple specification with price as single explanatory variable. The negative coefficient indicates, on the one hand, that a *ceteris paribus* increase in the price of cracker brand j decreases the probability of choosing this brand. On the other hand, the increase in the price of any other cracker brand m increases the probability of choosing alternative j.

However, our simple specification neglects an important aspect of cracker brand choice. Individuals might have a preference for a certain cracker brand, and therefore might choose this brand although it is relatively expensive. This would imply that we *underestimate* the absolute price effect in Model 1, and we should extend our analysis to allow for alternative specific intercepts. Recall that alternative specific intercepts describe the initial preferences of individuals. In order to include these intercepts, we choose *Private Label* as base category. Model 2 displays the estimation results for this specification, and we can see that the price effect is much more negative than before. We may formally test the two models against each other, since Model 1 is a restricted version of Model 2. A LR test shows a test statistic of $LRT = 2 \times [-3,364.90 - (-4,498.52)] = 2,267.24$ which is much larger than conventional critical values of a χ^2_3 distribution, and hence, we can reject Model 1.

Model 3 controls for further marketing-mix variables. Again, we can reject the restricted Model 2 against Model 3 with a LR test. To begin with, we can examine initial preferences of individuals. With help of the constants of the *Sunshine*, *Keebler*, and *Nabisco* equations, we can predict the choice probabilities, given all other characteristics are equal. This yields $\hat{p}_1 = 0.06$, $\hat{p}_2 = 0.10$, $\hat{p}_3 = 0.73$, and the probability of choosing a private label $\hat{p}_4 = 0.11$, which implies a strong preference for *Nabisco* crackers.

Now it could be of interest to determine how probabilities change by the usage of marketing instruments. How can we investigate the effect of a marginal increase in prices on the probabilities? Particularly, how can we interpret the coefficient of *price*? Here, we have to distinguish between the change in the probability of choosing cracker brand j if the price of j increases (own price effects), or if the price of any other alternative m increases (cross-price effects). Table 5.7 reports all possible effects.

Table 5.7. *Marginal Probability Effects*

$\widehat{MPE}_{jm,price} \times 0.1$

	j	(1) *Sunshine*	(2) *Keebler*	(3) *Nabisco*	(4) *Private Label*
m (1) *Sunshine*		-0.0209			
(2) *Keebler*		0.0014	-0.0206		
(3) *Nabisco*		0.0125	0.0123	-0.0822	
(4) *Private Label*		0.0069	0.0068	0.0574	-0.0711

Notes: All explanatory variables are fixed at their means. The predicted probabilities evaluated at the ML estimates are $\hat{p}_1 = 0.0671$, $\hat{p}_2 = 0.0663$, $\hat{p}_3 = 0.5579$, $\hat{p}_4 = 0.3087$

As expected, the probabilities of choosing brand j increases, if the own price decreases, or if the price of competing brands increases. Note that MPE's in the CL model are symmetric. This means that an increase in the price of alternative m on the probability of alternative j is the same as an increase

in the price of j on the probability of m. How do we obtain the numbers in Table 5.7? With equations (5.24) and (5.24) we have, for example, $\widehat{MPE}_{13} = -\hat{p}_1\hat{p}_3(-3.332) = 0.125$ which is multiplied by 0.1 to obtain the change in probabilities associated with a change in price by 0.1, or 10 cents, as prices are measured in US dollars. Hence, the effect 0.0125 means that if the price of *Nabisco* crackers increases by 10 cents, and if the prices of all competing brands remain unchanged, then the probability of choosing *Sunshine* crackers increases by 1.25 percentage points. Of course, MPE's can also be used in evaluating the quantitative effects of other marketing-mix variables.

Exercise 5.7.

- Calculate the MPE's of the variable *on display*.
- Critically evaluate the MPE's of *on display*.

5.3.5 Independence of Irrelevant Alternatives

The CL and MNL models have been criticized for making an implicit restrictive assumption, namely that the odds of comparing alternatives j and m (here for the CL model)

$$\frac{\pi_{ij}}{\pi_{im}} = \exp((z_{ij} - z_{im})'\gamma)$$

only depend on the attributes (or parameters) of the two alternatives j and m and not on the attributes, or even presence, of other alternatives. This property is known as **Independence of Irrelevant Alternatives** (IIA). To understand why this property may be undesirable, consider a model for choice of transport mode (the classic "red bus, blue bus" example). Assume that initially, the choice is between a blue bus and a car, and that the odds are 1/1, that is, the probability of either choice is 0.5. If one were to add a third choice to the analysis, a red bus, one would expect that the odds of taking a blue bus against taking a car drop to say, 1/2, since customers can be expected to perceive buses as rather similar and choose red and blue buses on an equal basis. But in the CL model this will not happen, the odds will remain 1/1. As a consequence, the model predicts too high a joint probability for very similar alternatives.

The IIA property results from the specific structure of the outcome probabilities. Moreover, in the general model of choice, the independence assumption of error terms leads to the IIA property, and therefore, IIA is also a

result of the decision-maker's behavior. But this can yield serious problems. From Chapter 3 we know that ML estimation loses all its good properties (consistency, asymptotic efficiency, and asymptotic normality) if the model is misspecified. Hence, if individual choices are in fact not conformable to IIA then our model *is* misspecified.

A test for the validity of the IIA assumption has been developed by Hausman and McFadden (1984). Their basic idea is that if an alternative (or set of alternatives) is irrelevant, exclusion of it should not change parameter estimates systematically. In other words, inclusion of irrelevant alternatives leads to consistent but inefficient estimates, whereas omitting relevant alternatives – where "relevant" means that the remaining alternatives are not independent of them – will cause inconsistency. This is the setup needed for a **Hausman specification test**. Let $\hat{\gamma}_r$ denote the parameter estimates from the reduced choice set, and $\widehat{\text{Var}}[\hat{\gamma}_r]$ the corresponding estimated covariance matrix. Let $\hat{\gamma}_f$ denote the estimates of the same parameters from the full choice set, and $\widehat{\text{Var}}[\hat{\gamma}_f]$ denote their estimated covariance matrix. It is important to know that some parameters estimated in the full set of choices may not be identified in the restricted choice set, in which case $\hat{\gamma}_f$ refers to the estimates of a subvector that is identified in both choice sets. The test statistic is given by

$$H = (\hat{\gamma}_f - \hat{\gamma}_r)' \left[\widehat{\text{Var}}(\hat{\gamma}_r) - \widehat{\text{Var}}(\hat{\gamma}_f)\right]^{-1} (\hat{\gamma}_f - \hat{\gamma}_r)$$

Under the null hypothesis that IIA holds, H is asymptotically χ^2-distributed with degrees of freedom equal to the number of elements in $\hat{\gamma}_f$. Note that we have to conduct separate tests for each possible reduced set of alternatives.

5.4 Generalized Multinomial Response Models

In this section, we briefly present three ways to overcome the IIA problem in the CL and MNL models. These include

- Multinomial probit models
- Mixed logit models
- Nested logit models

All these models relax the independence assumption of error terms u_{ij}, albeit in very different ways. Multinomial probit models are based on a **probit specification** of the choice probabilities. From Chapter 4 we know that with binary responses, the choice between probit and logit is largely a matter of taste. However, unlike with binary data, with multinomial response variables the choice *does* make a difference since flexible correlation structures are possible. Mixed logit models extend the basic models by assuming the existence of **additional error terms**. A distributional assumption with flexible correlation structure then relaxes the IIA property. Finally, nested logit models explicitly model a **tree structure** of choices, thereby allowing for dependencies among the choice set.

5.4.1 Multinomial Probit Model

The **multinomial probit** (MNP) model is frequently motivated in the general model of choice that we presented in Section 5.3.2. Recall that under random utility maximization, alternative j is chosen with probability

$$\begin{aligned}\pi_{ij} &= P(y_i = j|x_i, z_{i1}, \ldots, z_{iJ}) \qquad j = 1, \ldots, J \\ &= P(\mathcal{U}_{ij} > \mathcal{U}_{im}, \forall m \neq j|x_i, z_{i1}, \ldots, z_{iJ}) \\ &= P(u_{ij} - u_{im} > \mu_{im} - \mu_{ij}, \forall m \neq j|x_i, z_{i1}, \ldots, z_{iJ}) \\ &= P(u_{im} - u_{ij} < \mu_{ij} - \mu_{im}, \forall m \neq j|x_i, z_{i1}, \ldots, z_{iJ}) \end{aligned} \tag{5.26}$$

where $\mathcal{U}_{ij} = \mu_{ij} + u_{ij} = z'_{ij}\gamma + x'_i\beta_j + u_{ij}$ denotes individual i's indirect utility function of alternative j. The basic idea in the MNP model is to assume that the error terms u_{ij} are **jointly normally distributed** with mean 0 and flexible covariance matrix Σ. With a total of J alternatives, the most general form of covariance matrix is

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \cdots & \sigma_{1J} \\ \vdots & \ddots & \vdots \\ \sigma_{J1} & \cdots & \sigma_J^2 \end{pmatrix}_{J \times J}$$

where σ_j^2 denotes the variance of u_{ij}, $j = 1, \ldots, J$, and σ_{jm} denotes the covariance of u_{ij} and u_{im}, $\forall m \neq j$. Note that because of its symmetry, the covariance matrix Σ has $J(J+1)/2$ unique elements – the J variances σ_j^2 and the $J(J-1)/2$ covariances σ_{jm}. However, not all elements in Σ are identified, nor are all parameters of the utility function identified. The identification problem in the parameters of the utility function follows from the unidentified location of $\mathcal{U}_{ij}$. We can add any term $x'_i d$ to $\mathcal{U}_{ij}$ and this term will drop out in the utility comparisons of (5.26). This implies that the β_j's are not identified and we have to choose one category, say alternative 1, as the **base category**, and set its parameter vector β_1 to zero. This is analogous to the MNL model.

The identification problem in Σ is not that straightforward. Consider the simple case of three alternatives where the covariance matrix looks like

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{pmatrix}$$

with $J(J+1)/2 = 6$ unique elements. We begin with the derivation of the three outcome probabilities, and then discuss the identification issue. The probability of choosing alternative 1 can be derived from (5.26) as

$$\pi_{i1} = P(\nu_{i21} < \mu_{i1} - \mu_{i2}, \nu_{i31} < \mu_{i1} - \mu_{i3}|x_i, z_{i1}, \ldots, z_{iJ})$$

where $\nu_{i21} = u_{i2} - u_{i1}$ and $\nu_{i31} = u_{i3} - u_{i1}$. Since the error terms are assumed to be normally distributed with covariance matrix Σ, the differenced error terms ν_{i21} and ν_{i31} are normally distributed as well with covariance matrix

$$\Sigma^{(1)} = \begin{pmatrix} \sigma_1^2 + \sigma_2^2 - 2\sigma_{12} & \sigma_1^2 - \sigma_{12} - \sigma_{13} + \sigma_{23} \\ \sigma_1^2 - \sigma_{12} - \sigma_{13} + \sigma_{23} & \sigma_1^2 + \sigma_3^2 - 2\sigma_{13} \end{pmatrix} = \begin{pmatrix} \omega_{11}^{(1)} & \omega_{12}^{(1)} \\ \omega_{12}^{(1)} & \omega_{22}^{(1)} \end{pmatrix}$$

The normality assumption yields to a two-dimensional integral for π_{i1} of the form

$$\pi_{i1} = \int_{-\infty}^{\mu_{i1}-\mu_{i2}} \int_{-\infty}^{\mu_{i1}-\mu_{i3}} \phi_2^{(1)}(\nu_{i21}, \nu_{i31}) d\nu_{i21} d\nu_{i31} \tag{5.27}$$

where $\phi_2^{(1)}$ stands for the **bivariate normal density function** of the differenced error terms with zero mean and covariance matrix $\Sigma^{(1)}$. The probability of choosing alternative 2 can be derived similarly as

$$\pi_{i2} = P(\nu_{i12} < \mu_{i2} - \mu_{i1}, \nu_{i32} < \mu_{i2} - \mu_{i3} | x_i, z_{i1}, \ldots, z_{iJ})$$

where $\nu_{i12} = u_{i1} - u_{i2}$ and $\nu_{i32} = u_{i3} - u_{i2}$. As before, the differenced error terms ν_{i12} and ν_{i32} are normally distributed but now with covariance matrix

$$\Sigma^{(2)} = \begin{pmatrix} \sigma_1^2 + \sigma_2^2 - 2\sigma_{12} & \sigma_2^2 - \sigma_{12} - \sigma_{23} + \sigma_{13} \\ \sigma_2^2 - \sigma_{12} - \sigma_{23} + \sigma_{13} & \sigma_2^2 + \sigma_3^2 - 2\sigma_{23} \end{pmatrix} = \begin{pmatrix} \omega_{11}^{(2)} & \omega_{12}^{(2)} \\ \omega_{12}^{(2)} & \omega_{22}^{(2)} \end{pmatrix}$$

From this, the probability of observing alternative 2 can be calculated as

$$\pi_{i2} = \int_{-\infty}^{\mu_{i2}-\mu_{i1}} \int_{-\infty}^{\mu_{i2}-\mu_{i3}} \phi_2^{(2)}(\nu_{i12}, \nu_{i32}) d\nu_{i12} d\nu_{i32} \tag{5.28}$$

where $\phi_2^{(2)}$ stands for the bivariate normal density function with zero mean and covariance matrix $\Sigma^{(2)}$. The probability of choosing the third alternative is simply $\pi_{i3} = 1 - \pi_{i1} - \pi_{i2}$. We can see that estimation of the MNP model requires evaluating two two-dimensional integrals over a bivariate normal density function, or in general $J-1$ integrals of dimension $(J-1)$ over a multivariate normal density with dimension $J-1$. The reduction in dimension from J to $J-1$ follows from the $J-1$ possible *comparisons* of utilities.

However, not all elements in $\Sigma^{(1)}$ and $\Sigma^{(2)}$ are identified. More specifically, we only have three unique elements in both variances since

$$\begin{aligned} \omega_{11}^{(1)} &= \omega_{11}^{(2)} \\ \omega_{11}^{(1)} - \omega_{12}^{(1)} &= \omega_{12}^{(2)} \\ \omega_{22}^{(1)} - \omega_{12}^{(1)} &= \omega_{22}^{(2)} - \omega_{12}^{(2)} \end{aligned}$$

Hence, $\Sigma^{(2)}$ is fully determined by $\Sigma^{(1)}$. But the three unique elements in $\Sigma^{(1)}$ are still not identified since the scale of utilities is not identified: we can multiply all the $\mathcal{U}_{ij}$'s by a positive constant c and this does not change the maximizing $\mathcal{U}_{ij}$. For example, the inequality $\nu_{i12} < \mu_{i2} - \mu_{i1}$ is still fulfilled if we multiply both sides by $c > 0$. We solve this problem by fixing the upper-left element in $\Sigma^{(1)}$ to one, i.e., we set $\omega_{11}^{(1)} = 1$, and obtain

$$\tilde{\Sigma}^{(1)} = \begin{pmatrix} 1 & \tilde{\omega}_{12}^{(1)} \\ \tilde{\omega}_{12}^{(1)} & \tilde{\omega}_{22}^{(1)} \end{pmatrix} \tag{5.29}$$

with two unique elements. Now the question arises, how we can achieve the normalization in (5.29) from a normalization of the original covariance matrix Σ. It turns out that one possibility is to fix the error term covariances of the base category to 1/2, its variance to 1, and the variance of category 2 to one. Formally, we let $\sigma_1^2 = \sigma_2^2 = 1$ and $\sigma_{12} = \sigma_{13} = 1/2$. In this case, we would obtain $\sigma_{23} = \tilde{\omega}_{12}^{(1)}$ and $\sigma_3^2 = \tilde{\omega}_{22}^{(1)}$ which implies that we can recover all non-normalized parameters in Σ from the elements in $\tilde{\Sigma}^{(1)}$.

In general, we can identify a maximum of $J(J-1)/2 - 1$ parameters in Σ, which is exactly the number of unique elements in the normalized covariance matrix $\tilde{\Sigma}^{(1)}$ with general dimension $(J-1) \times (J-1)$. Hence, we have to normalize $J+1$ parameters for identification. In order to impose this normalization, we fix the covariances of the base category to 1/2, its variance to 1, and choose the second category as **scale category** by fixing its error term variance to 1. In principle, we could impose any other structure on the covariance matrix Σ. Here, we have to distinguish between *normalization* and *restriction*. A restriction is any assumption about the values of parameters or relationships between them, whereas a normalization is a necessary condition imposed on the parameter space to achieve identification, such that we are able to estimate the model. We can check whether the restrictions that we imposed are sufficient to normalize the model. This can be done by transforming the stated covariance matrix Σ to $\tilde{\Sigma}^{(1)}$. If we can recover all the original parameters in Σ from the elements in the transformed matrix $\tilde{\Sigma}^{(1)}$, then the model is sufficiently normalized.

Having expressions for the probabilities and identified parameters, one can set the log-likelihood function in the same manner as in Section 5.2.2. In general, evaluation of a $J-1$ dimensional integral over a joint normal density is not a trivial task. Traditionally, it has been considered impossible if the number of categories exceeds three. However, recent progress in computational methods has alleviated this problem (see for example Genz, 1992, Bolduc, 1999, Train, 2003, Cappellari and Jenkins, 2003).

5.4.2 Mixed Logit Models

Mixed logit (MXL) models are the second alternative to relax the IIA assumption that we consider in this book. MXL models assume that the utility function $\mathcal{U}_{ij}$ depends on a deterministic component μ_{ij}, an independently and identically distributed random component u_{ij}, and, deviating from the standard model, additional random terms, grouped together in ε_{ij}. These terms can be a function of the data, potentially modeling the presence of correlation and heteroscedasticity. For notational simplicity, let the deterministic component be determined by the CL specification $\mu_{ij} = z_{ij}'\gamma$ such that we can write the random utility function as

$$\mathcal{U}_{ij} = z'_{ij}\gamma + \varepsilon_{ij} + u_{ij}$$

Now let $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iJ})'$ denote the vector of additional random terms for all choices $j = 1, \ldots, J$. Furthermore, assume that ε_i is distributed as $\varepsilon_i \sim f(\varepsilon_i; \alpha)$, where f denotes a general density function with fixed parameters α. As in the MNL/CL case, assume that the u_{ij}'s, $j = 1, \ldots, J$, are independently and identically type-I extreme value distributed. Therefore, individual i chooses alternative j *conditional* on ε_i with probability

$$P(y_i = j|\varepsilon_i, z_{i1}, \ldots, z_{iJ}) = \frac{\exp(z'_{ij}\gamma + \varepsilon_{ij})}{\sum_{r=1}^{J} \exp(z'_{ir}\gamma + \varepsilon_{ir})} \tag{5.30}$$

This corresponds to the probabilities specified in the standard CL model, with additional conditioning terms ε_i. However, since ε_i is not observable, we cannot calculate the probabilities in (5.30) for a given set of parameters. In order to obtain probabilities that depend only on the parameters we have to **integrate out** the ε_i's. This means that we have to calculate the integral of the conditional probability (5.30) over all possible values of ε_i, formally

$$P(y_i = j|z_{i1}, \ldots, z_{iJ}) = \int_{\Upsilon(\varepsilon)} \frac{\exp(z'_{ij}\gamma + \varepsilon_{ij})}{\sum_{r=1}^{J} \exp(z'_{ir}\gamma + \varepsilon_{ir})} f(\varepsilon_i; \alpha) d\varepsilon_i \tag{5.31}$$

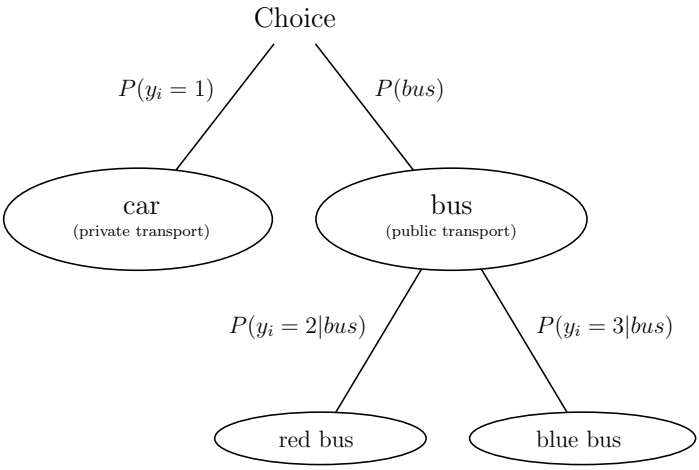
where $\Upsilon(\varepsilon_i)$ is the support of ε_i. In general, this choice probability does not have a mathematical closed-form expression that can be solved analytically. Nevertheless, the closed-form of the *conditional* probability in (5.30) allows for evaluation at different values of ε_i drawn from its density function $f(\varepsilon_i; \alpha)$. Taking an average probability over the different draws and varying α to maximize the log-likelihood function of the sample yields consistent and asymptotically normal estimates. In the literature, this method is called **simulated maximum likelihood**. We do not consider these methods in greater depth here; the interested reader is referred to Train (2003). MXL models have been proven computationally tractable and they provide great flexibility in specification of $f(\varepsilon_i; \alpha)$, and therefore are very attractive in empirical applications.

5.4.3 Nested Logit Models

As a third alternative, we briefly discuss **nested logit models**, also called **hierarchical response models**. The basic idea of this class of models is to group similar alternatives into **nests** within which the IIA assumption holds, while allowing the variances to differ across these nests. To illustrate the basic principles and refer back to the simple example in Section 5.3.5, we consider a tree structure describing the choice between a car, a blue bus, and a red bus. Figure 5.1 shows a possible decision tree.

In the above example, we assume that an individual first chooses between a bus and car (or call it public and private transport), and – conditional on

Fig. 5.1. *Transport Mode Decision Tree*



bus preference – the decision is between the red and the blue one. Further, assume that utility is again given by

$$\mathcal{U}_{ij} = z'_{ij}\gamma + x'_i\beta_j + u_{ij} = \mu_{ij} + u_{ij}$$

where $j = 1, 2, 3$ with car (1), red bus (2), blue bus (3), and the error terms are trivariate extreme value distributed. Then, the probabilities of all three alternatives are determined by

$$P(y_i = 1|x_i, z_{i1}, z_{i2}, z_{i3}) = \frac{\exp(\mu_{i1})}{\exp(\mu_{i1}) + [\exp(\varrho^{-1}\mu_{i2}) + \exp(\varrho^{-1}\mu_{i3})]^{\varrho}}$$

$$P(y_i = 2|bus, x_i, z_{i2}, z_{i3}) = \frac{\exp(\varrho^{-1}\mu_{i2})}{\exp(\varrho^{-1}\mu_{i2}) + \exp(\varrho^{-1}\mu_{i3})} \tag{5.32}$$

with "similarity" parameter ϱ. If $\varrho = 1$ the nested logit model reduces to the standard CL model. Note that with these two probabilities, we can calculate all other probabilities of interest

$$P(bus|x_i, z_{i1}, z_{i2}, z_{i3}) = 1 - P(y_i = 1|x_i, z_{i1}, z_{i2}, z_{i3})$$

$$P(y_i = 3|y \neq 1, x_i, z_{i2}, z_{i3}) = 1 - P(y_i = 2|y \neq 1, x_i, z_{i2}, z_{i3})$$

$$\begin{aligned} &P(y_i = 2|x_i, z_{i1}, z_{i2}, z_{i3}) = \\ &\qquad P(y_i = 2|y \neq 1, x_i, z_{i2}, z_{i3}) \cdot P(bus|x_i, z_{i1}, z_{i2}, z_{i3}) \end{aligned} \tag{5.33}$$

For more general tree structures, joint cumulative density functions of the **generalized extreme value** form have been developed. McFadden (1984) gives a detailed treatment of hierarchical response models. In order to specify the nested logit model, one has to partition choices and impose an appropriate tree structure. In some applications, the partitioning is naturally given or can be based on an underlying economic theory. But in general, the hierarchical structure of choices is assumed by the researcher and results might be sensitive to the particular specification. This is a possible disadvantage of nested logit models.

5.5 Further Exercises

Exercise 5.8 Which alternative normalization could be imposed in the MNL model with probabilities specified as in equation (5.2) in order to identify the parameters? What consequence does this have for the interpretation?

Exercise 5.9 Suppose you want to estimate a multinomial response model in a dataset with the following response pattern:

y	(1) *car*	(2) *bus*	(3) *train*
n	26	53	31

Let π_j denote the probability of choosing transport mode j.

a) Derive the likelihood function $L(\pi_1, \pi_2, \pi_3; y)$. What restriction has to be imposed on the parameters?

b) Obtain the ML estimates of the probabilities π_1, π_2, and π_3.

c) Now suppose the model has been specified as a MNL model without covariates. How many parameters does the model have? Obtain the ML estimates of the parameters.

d) How can you obtain standard errors of the estimates in b) and c).

Exercise 5.10 Suppose you want to analyze the choices that tourists make when selecting between different holiday destinations. To simplify the discussion assume that the choice is among the alternatives (1) *Switzerland*, (2) *Canary Islands*, (3) *Balearic Islands*, and (4) *North America*.

a) Which variables could be important determinants of a tourist's choice? List at least 3 choice-specific attributes, and 3 individual-specific attributes.

b) How would you estimate the effect of the regressors listed in a) on the outcome probabilities? Which model do you use? Why?

Exercise 5.11 Suppose you estimate a MNL model using three labor force states: (1) *employment*, (2) *unemployment*, and (3) *out of labor force*. *Out of labor force* is the omitted reference state. You obtain the following parameter estimates:

	employment	*unemployment*
male	3.50	2.90
age	0.10	0.008
$(age)^2$	-0.002	-0.0006
male$*$*age*	-0.03	-0.02
constant	-1.90	-1.40

a) True or false?
The odds of being *employed* relative to being *out of the labor force* are $(\exp(3.5) - 1) \times 100$ percent higher for men than for women.

b) What is the predicted probability that a 20-year-old woman is *employed*, *unemployed*, and *out of the labor force*, respectively?

c) What is the predicted probability that a 20-year-old man is *employed*, *unemployed*, and *out of the labor force*, respectively?

d) How do you interpret the difference in b) and c)?

Exercise 5.12 Schmidt and Strauss (1975) apply a MNL model to predict occupations. The occupations are classified in five groups: (1) *menial workers*, (2) *blue collar workers*, (3) *craft workers*, (4) *white collar workers*, and (5) *professional workers*. They obtain the following results:

	blue collar	*craft*	*white collar*	*professional*
education	-0.1238	0.0490	0.2163	0.4128
	(-2.71)	(0.92)	(4.17)	(7.59)
experience	-0.0243	-0.0096	-0.0168	-0.0013
	(-2.74)	(-0.94)	(-1.70)	(-0.12)
white	1.244	2.747	2.8517	1.879
	(4.46)	(5.02)	(5.11)	(3.83)
male	0.7988	2.138	-0.8087	0.2263
	(3.23)	(5.34)	(-3.14)	(0.80)
constant	1.293	-4.086	-3.358	-6.025
	(2.18)	(-4.56)	(-4.05)	(-7.11)

Notes: Coefficients and z-ratios. *Education* and *experience* are measured in years.

a) Which category has been chosen as the base category?

b) How does the probability of being a *blue collar worker* change relative to the probability of being a *professional worker*, given an additional year of schooling?

c) How does the probability of being a *blue collar worker* change relative to the probability of being a *menial worker*, given an additional year of schooling?

d) Calculate the predicted probability of each occupation for a white man with 12 years of schooling and 25 years of experience.

e) Calculate the predicted probability of each occupation for a black man with 12 years of schooling and 25 years of experience. Interpret the differences compared to the predictions in d).

Exercise 5.13 Suppose you are interested in the location choice of firms and its relationship to environmental regulations. Stafford (2000) uses a dataset including 1,548 facilities in 48 states in the continental U.S. that managed hazardous waste. A CL analysis yields the following results:

	Coefficient	St. Err.
demand in log(tons)	-0.0065	(0.0328)
neighboring demand in log(tons)	-0.0013	(0.0387)
mean corporate income tax rate	-3.8463	(1.5798)
average annual wage in log(dollar)	0.4622	(0.6377)
number of production hours in log(million)	0.6784	(0.0733)
average cost of energy in log(dollar/BTU)	-2.6795	(0.4289)
— *if recovery system* in log(dollar/BTU)	1.4003	(0.2861)
average construction cost in log(dollar)	3.6169	(0.9378)
total area in 100 million acres	0.1630	(0.1460)
population in 100 thousand people	-1.1100	(0.4310)
spending on environmental programs index	-0.0052	(0.0025)
stringency of environmental policies index	0.0136	(0.0065)

a) Describe the nature of the dependent variable *location choice*. Why are environmental regulations particularly interesting when considering hazardous waste management firms?

b) Interpret the coefficient on the tax rate. Is the sign plausible? What does the sign imply for the marginal probability effects?

c) How does an increase in the average energy costs in location j affect the probability of choosing location j? What happens if the costs in location $m \neq j$ increase?

d) How do the environmental policy variables affect the location decision of hazardous waste management firms?

Exercise 5.14 Hoffmann and Duncan (1988) study the choice of marital and welfare status of divorced or separated women, where the choice is between (1) *remarriage*, (2) remaining *single with welfare receipt*, and (3) remaining *single without welfare receipt*. In particular, they focus on the role of income in determining the woman's choice, and distinguish between exogenous income (the income available at zero hours of work) and the after-tax wage rate. They report results of the form:

	(1)	(2)	(3)
wage rate	1.102	1.102	1.102
	(0.107)	(0.107)	(0.107)
husbands income	-0.018	-	-
	(0.017)	-	-
AFDC income	-	0.192	-
	-	(0.051)	-
nonlabor income	-0.011	-	0.215
	(0.090)	-	(0.076)
constant	-2.408	-	-2.587
	(0.542)	-	(0.499)

Notes: Standard errors in parentheses. AFDC = Aid to Females with Dependent Children.

a) Critically evaluate the use of a CL model. Does exogenous income vary across alternatives? What are the components of exogenous income? Does the after-tax wage rate vary across alternatives?

b) Why do the authors report only one coefficient for husbands income and AFDC income? Why do they report two different coefficients for nonlabor income and the constant?

c) Interpret the coefficients on wage rate and AFDC income.

Exercise 5.15 Use the dataset from Greene (2003: Table F21.2) to study the travel mode choice of individuals traveling between Sydney and Melbourne. The dataset contains 210 individuals choosing among four alternatives: (1) *air*, (2) *train*, (3) *bus*, and (4) *car*.

a) Describe the nature of the variable travel mode choice? What explanatory variables are available? How are the data organized?

b) Estimate a CL model and report your results along with the value of the maximized log-likelihood function, and a LR test statistic under the null hypothesis of choice-specific intercepts only.

c) Interpret the results using discrete or marginal probability effects, predicted probabilities, and odds ratios.

d) What problem might arise in the model above? What solution do you propose?

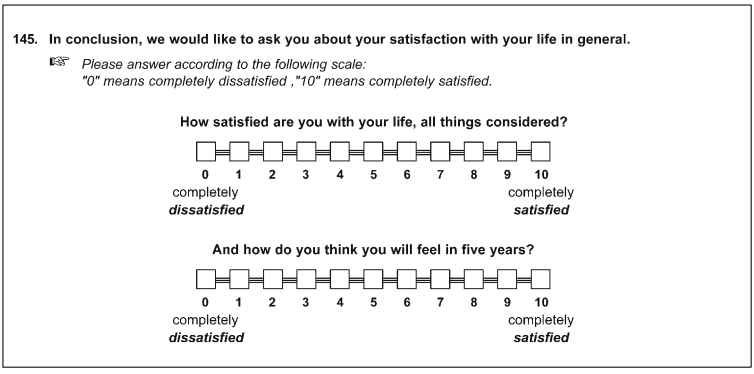
6

Ordered Response Models

6.1 Introduction

In Chapter 5, we discussed models for multinomial data, i.e., models in which the response variable is characterized by more than two categorical outcomes, and whereby the ranking of responses is purely arbitrary. In some empirical applications, however, a multinomial choice can have an inherent ordering. For example, in the individual questionnaire of the German Socioeconomic Panel (GSOEP) we can find questions like the one displayed in Figure 6.1.

Fig. 6.1. *Life-Satisfaction Question in the GSOEP*



The given answer categories range from zero to ten and have an ordinal scale since a response "4" instead of "8" contains information – we know that people answering a satisfaction level of "8" are happier than those answering a level of "4" – and this information should be used for estimation.

Ordered responses appear in various fields of research, and we illustrate this with some examples.

Example 6.1. Ordered Responses

- Aitchison and Silvey (1957) propose an ordered response model to analyze experiments where insects were subjected to various doses of poison and respond, for example, with *unaffected*, *slightly affected*, *moribund*, or *dead*.
- McKelvey and Zavoina (1975) illustrate the ordered probit model with an analysis of voting behavior which is characterized by voting *against*, *weakly for*, or *strongly for* a Medicare bill.
- Winship and Mare (1984) estimate a simple model of educational transmission where schooling is measured as *less than 8 years*, *between 8 and 11 years*, *12 years*, or *more than 13 years*.
- Ederington (1985) investigates the performance of the ordered probit model in predicting bond ratings, where industrial bond issues are rated as *Aaa*, *Aa*, *A*, *Baa*, *Ba*, or *B*.
- Cutler and Richardson (1998) examine the relationship between different forms of disease and self-reported health status, the latter being classified as *excellent*, *very good*, *good*, *fair*, or *poor*.
- Kaiser and Spitz (2002) propose the use of ordered probit models to study qualitative information in business surveys when respondents are asked, for example, whether their total sales *increased*, *remained the same*, or *decreased* in the preceding quarter.

In all these examples, the ordered responses are mutually exclusive and exhaustive, such as with purely multinomial data. The additional and special feature of an ordered response is that its answer categories can be ranked from low to high (or vice versa), but the particular values assigned to the outcomes remain arbitrary, as long as they preserve the order. Thus the sequence 1, 2, 3 embodies the same information as the sequence 13, 22, 27, or the sequence -5, 11, 100. This in turn implies that ordered responses do not have origins, or units of measurement, and that expectations, variances, and covariances have no meaning. For notational simplicity, we adopt the convention that the dependent variable is coded as

$$y_i = 1, 2, \ldots, J \tag{6.1}$$

where "1" < "2" < ... < "J". From a statistical point of view, we could analyze an ordered dependent variable with multinomial response models like the ones presented in Chapter 5 since we have a multinomial dependent variable. However, conditional probability models that ignore the ranking of responses

are likely to give inefficient estimation results compared to models that appropriately account for the additional information.

Yet we should warn the reader of a mistake that sometimes occurs in empirical practice and which might be the result of the (conventional) coding given in equation (6.1). In many applications, the ordinal outcomes with values $1, \ldots, J$ are analyzed by using the linear regression model and ordinary least squares. This would presume that the difference between a "1" and a "2" is the same as the difference between a "8" and a "9", whereas in fact, we only have the ranking. Moreover, for example in the satisfaction response, we cannot conclude from ordinal information that people answering a "2" are twice as happy as people answering a "1".

To summarize, one should think carefully about the dependent variable under consideration. Is it really measured on an ordinal scale? If so, models for multinomial data are feasible but inefficient since they ignore the ordering information. The linear regression model cannot be appropriate either due to the implicit assumption of an interval scale. The purpose of this chapter is to introduce econometric models that take into account the special feature of ordered responses.

The chapter proceeds as follows. In Section 6.2, we present the standard ordered response models, the ordered probit and ordered logit models. As we will see, the basic idea in modeling ordinal outcomes is to use a latent variable and a threshold mechanism in the same manner as presented in Chapter 4 for binary responses, but now with more than one constant threshold parameter. In Section 6.3, we generalize the basic threshold mechanism by making the thresholds themselves functions of the explanatory variables. This relaxes the single index assumption imposed by the standard models and provides much more flexibility for interpretation in terms of marginal probability effects. Section 6.4 discusses an alternative, sequential mechanism, where the categories are reached successively. Finally, we consider a special kind of ordinal data in Section 6.5 which is referred to as interval data. In this case, an ordered dependent variable is the result of a coarsening of data, for example due to limitations in data availability.

6.2 Standard Ordered Response Models

6.2.1 General Framework

Models for ordered dependent variables are usually motivated by an underlying continuous but latent process y_i^* given by

$$y_i^* = x_i'\beta + u_i \qquad i = 1, \ldots, n \tag{6.2}$$

with deterministic component $x_i'\beta$ (the linear index of regressors), and random terms u_i, which are assumed to be independently and identically distributed with distribution function $F(u)$ with mean zero and constant variance (see, for example, McKelvey and Zavoina, 1975). Since we cannot observe the latent continuous variable y_i^*, but instead observe y_i with discrete values $1, \ldots, J$, we need to find a mechanism that relates y_i^* and y_i. A sensible mechanism accounts for the ordering information in y_i and we assume that

$$y_i = j \quad \text{if and only if} \quad \kappa_{j-1} < y_i^* \leq \kappa_j \qquad j = 1, \ldots, J \tag{6.3}$$

This mechanism is called **threshold mechanism** since the J outcomes are obtained by dividing the real line, represented by y_i^*, into J intervals, using $J + 1$ constant but unknown threshold parameters $\kappa_0, \ldots, \kappa_J$. In order to ensure well-defined intervals, we need to assume ascending thresholds such that $\kappa_0 < \ldots < \kappa_J$. We code the intervals from 1 to J and account for the ordering information since higher values of y_i^* yield higher outcomes of y_i. The full set of outcomes, or intervals, is given by

$$\begin{aligned}
y_i = 1 &\text{ if and only if } \quad \kappa_0 < y_i^* \leq \kappa_1 \quad \Leftrightarrow \quad \kappa_0 - x_i'\beta < u_i \leq \kappa_1 - x_i'\beta \\
y_i = 2 &\text{ if and only if } \quad \kappa_1 < y_i^* \leq \kappa_2 \quad \Leftrightarrow \quad \kappa_1 - x_i'\beta < u_i \leq \kappa_2 - x_i'\beta \\
&\vdots \qquad\qquad\qquad\qquad \vdots \qquad\qquad\qquad\qquad \vdots \\
y_i = J &\text{ if and only if } \quad \kappa_{J-1} < y_i^* \leq \kappa_J \Leftrightarrow \kappa_{J-1} - x_i'\beta < u_i \leq \kappa_J - x_i'\beta
\end{aligned}$$

where it is understood that $\kappa_0 = -\infty$ and $\kappa_J = \infty$ to cover the entire real line. Hence, the number of unknown threshold parameters reduces to $J - 1$. The distributional assumption on the error terms yields the conditional probability function of the latent variable, $f(y_i^*|x_i)$ which allows for an illustrative representation of the threshold mechanism. Figure 6.2 plots a symmetric density function $f(y_i^*|x_i)$ with mean $x_i'\beta$. On the horizontal axis, we display the latent variable y_i^*, which is, according to the threshold mechanism in (6.3), divided into $J = 5$ intervals. The utmost left interval corresponds to $y_i = 1$, the utmost right interval corresponds to $y_i = 5$, respectively. In Figure 6.3, we have basically the same representation of the threshold mechanism, but now in terms of the error terms u_i. Therefore, the density function has zero mean and the thresholds are determined by $\kappa_j - x_i'\beta$.

Fig. 6.2. *Threshold Mechanism in Terms of* y_i^*

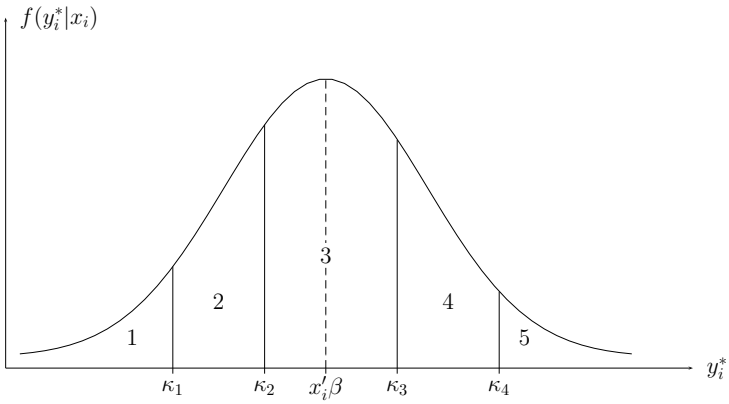
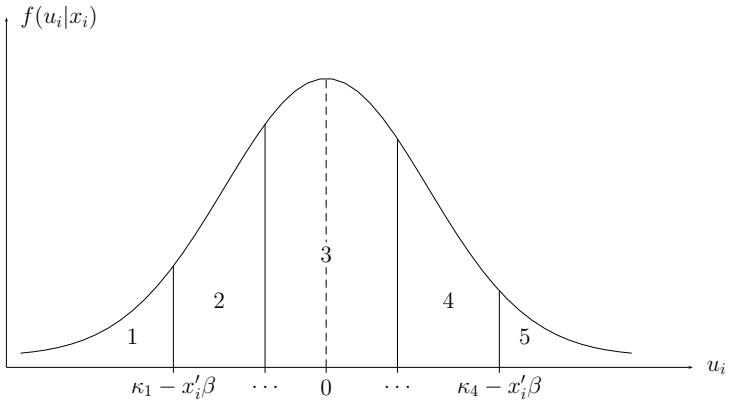


Fig. 6.3. *Threshold Mechanism in Terms of* u_i



Once the distribution function $F_1(u)$ has been specified, the probability of a particular outcome is determined by the area under the density function between the relevant thresholds. For example, the probability of observing $y_i = 3$ is the area under $f(y_i^*|x_i)$ between κ_2 and κ_3. For $y_i = 1$ it would be the area in the left tail, for $y_i = J$ the area in the right tail of the density function. In general, we can write the probabilities as

$$\pi_{ij} = P(y_i = j|x_i) = F(\kappa_j - x_i'\beta) - F(\kappa_{j-1} - x_i'\beta) \tag{6.4}$$

for $j = 1, \ldots, J$. Note that we assumed $\kappa_0 = -\infty$ and $\kappa_J = \infty$ such that $F(-\infty) = 0$ and $F(\infty) = 1$. As long as the explanatory variables x contain

an intercept, $F(\kappa_j - x_i'\beta) = F[(\kappa_j + c) - (x_i'\beta + c)]$, where c is an arbitrary constant. From this simple rearrangement we see that the intercept in $x_i'\beta$ cannot be distinguished from one of the threshold parameters. In order to solve this identification problem, a normalization is required. Two possibilities come in mind. First, we can fix one of the threshold parameters, for example $\kappa_1 = 0$. Second, we can drop the constant term from the set of regressors. In the remainder of this chapter, if not mentioned otherwise, we will assume that x does not contain an intercept and has the dimension $k \times 1$. As will become clear below, this simplifies interpretation, and it is the default option in many software packages.

The most commonly known models for ordered responses, the ordered probit model and the ordered logit model, can be distinguished in the way they specify the distribution function of the error terms. In the former, it is assumed that $F(u)$ is the standard normal distribution, in the latter, $F(u)$ is assumed to be the (standard) logistic distribution. These two models will be the subject of the following two sections, 6.2.2 and 6.2.3.

6.2.2 Ordered Probit Model

In the **ordered probit model**, we assume that the error terms follow a **standard normal distribution**, $F(u) = \Phi(u)$. In this case, the probabilities can be written as

$$\pi_{ij} = \Phi(\kappa_j - x_i'\beta) - \Phi(\kappa_{j-1} - x_i'\beta) \qquad j = 1, \ldots, J \tag{6.5}$$

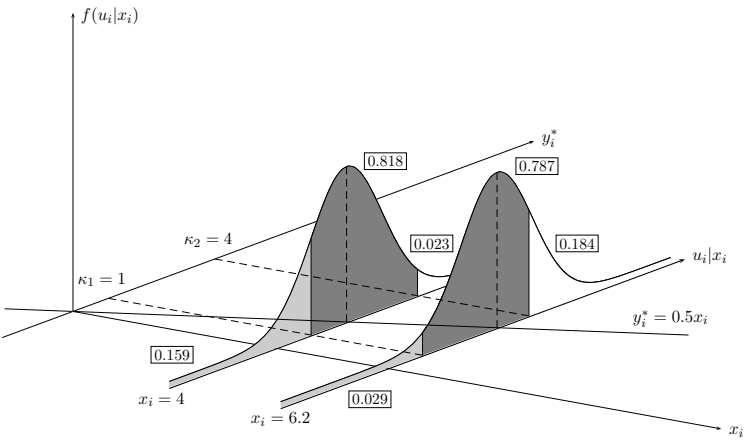
The assumption of a *standard* normal distribution is, like in the binary probit model, necessary to achieve identification of parameters. With general variance σ^2, the probabilities would be given by

$$\pi_{ij} = \Phi\left(\frac{\kappa_j - x_i'\beta}{\sigma}\right) - \Phi\left(\frac{\kappa_{j-1} - x_i'\beta}{\sigma}\right) \qquad j = 1, \ldots, J$$

and we see that only the ratios κ_j/σ and β/σ are identified. If we multiply each of the parameters (κ, β, σ) by a constant c, this constant cancels out in the ratios and all probabilities remain unchanged. Hence, we need a normalization to identify the parameters and we solve this issue by setting $\sigma = 1$.

Figure 6.4 illustrates, similar to Kennedy and Becker (1982), the ordered probit model by means of a simple example with $J = 3$ response categories and one explanatory variable x. The three axes describe the latent variable y_i^*, the explanatory variable x, and the density of the error term conditional on x, respectively. For the moment, assume that we know the true parameter values in the model (we will deal with the subject of estimation in Section 6.2.4). In this example, the only slope parameter is given by 0.5, the thresholds are given by 1 and 4. The conditional density of the error term is plotted for two distinct values of x, $x = 4$ and $x = 6.2$, and the probabilities of observing the outcomes $y = 1, 2, 3$, *conditional* on x, are characterized by the light gray, the

Fig. 6.4. *Graphical Illustration of the Ordered Probit Model*



dark gray, and the white shaded areas under the conditional density. Figure 6.4 visualizes that probabilities generally depend on the specific value(s) of x. As mentioned before, they are determined by the particular areas under the density function with fixed integration limits $-\infty$, κ_1, κ_2, and ∞. But the location of the density varies with x, and therefore the probabilities for $x = 4$ are different from those for $x = 6.2$. For example, the probability of $y = 1$ given $x = 4$ is equal to 0.159, and decreases to 0.029 for $x = 6.2$.

6.2.3 Ordered Logit Model

The **ordered logit model** can be derived in the same way as its probit counterpart. More precisely, we specify a latent variable y_i^* and assume that the error terms are independently and identically **logistically distributed**. The model may be illustrated graphically as in Figure 6.4 with a logistic density function, and the probabilities are given by

$$\pi_{ij} = \Lambda(\kappa_j - x_i'\beta) - \Lambda(\kappa_{j-1} - x_i'\beta) \qquad j = 1, \ldots, J \tag{6.6}$$

where $F(\cdot) = \Lambda(\cdot)$ is shorthand notation for the cumulative density function of the logistic distribution.

Another way to motivate the ordered logit model is based on the odds, and we will cover this approach in greater[1]detail here to explain why this model is also known as a **proportional odds model**. Consider two events $y_i \leq j$ and $y_i > j$ with probabilities $P(y_i \leq j|x_i)$ and $P(y_i > j|x_i)$. In a model with J ordered outcomes there are $J - 1$ such comparisons. Then form the ratio of the two probabilities

$$\frac{P(y_i \leq j|x_i)}{P(y_i > j|x_i)} \qquad j = 1, \ldots, J-1 \tag{6.7}$$

adopting the term **odds** of the events $y_i \leq j$ versus $y_i > j$ from the previous chapters. Now assume that the logarithmic odds are linear in the parameters, formally

$$\log \frac{P(y_i \leq j|x_i)}{P(y_i > j|x_i)} = \kappa_j - x_i'\beta \qquad j = 1, \ldots, J-1 \tag{6.8}$$

which is the same as saying that

$$\frac{P(y_i \leq j|x_i)}{P(y_i > j|x_i)} = \exp(\kappa_j - x_i'\beta) = \exp(\kappa_j)\exp(-x_i'\beta) \tag{6.9}$$

As before, the vector of explanatory variables x does not contain an intercept for identification of κ_j. For each of the possible odds, we assume a specific intercept κ_j and the ordering information in y_i is accounted for by assuming that $\kappa_1 < \kappa_2 < \ldots < \kappa_{J-1}$. We choose a negative sign of $x_i'\beta$ which is arbitrary but relates the results below to those of the latent variable approach. Note that from (6.9) the relative odds of comparison j and comparison m do not depend on x but only on the threshold parameters,

$$\frac{P(y_i \leq j|x_i)/P(y_i > j|x_i)}{P(y_i \leq m|x_i)/P(y_i > m|x_i)} = \frac{\exp(\kappa_j)}{\exp(\kappa_m)} \tag{6.10}$$

This property gives rise to the name **proportional odds model**.

For the purpose of formulating a conditional probability model and estimation by ML we need probability expressions rather than odds. From

$$\frac{P(y_i \leq j|x_i)}{P(y_i > j|x_i)} = \frac{P(y_i \leq j|x_i)}{1 - P(y_i \leq j|x_i)}$$

we obtain

$$P(y_i \leq j|x_i) = \frac{\exp(\kappa_j - x_i'\beta)}{1 + \exp(\kappa_j - x_i'\beta)} = \Lambda(\kappa_j - x_i'\beta)$$

and therefore

$$\begin{aligned} \pi_{ij} = P(y_i = j|x_i) &= P(y_i \leq j|x_i) - P(y_i \leq j-1|x_i) \\ &= \Lambda(\kappa_j - x_i'\beta) - \Lambda(\kappa_{j-1} - x_i'\beta) \end{aligned} \tag{6.11}$$

where it is understood that $\kappa_0 = -\infty$ and $\kappa_J = \infty$. If we compare (6.11) with (6.6), we notice that both probability expressions are the same. Hence, both approaches, the latent variable and the proportional odds approach, yield the same probability function and we call them **observationally equivalent**.

6.2.4 Estimation

In order to proceed with ML estimation of the parameter vector β and the $J-1$ threshold parameters $\kappa_1, \ldots, \kappa_{J-1}$, we need to rewrite the general probabilities in (6.4), or the specific ones in (6.5) and (6.6), into a conditional probability function. We have

$$f(y_i|x_i;\beta,\kappa_1,\ldots,\kappa_{J-1}) = (\pi_{i1})^{d_{i1}} \cdots (\pi_{iJ})^{d_{iJ}} = \prod_{j=1}^{J} (\pi_{ij})^{d_{ij}} \tag{6.12}$$

where d_{ij} is defined as a binary indicator equal to one if $y_i = j$ and equal to zero otherwise (see also Chapter 5.2.2). For a sample of n independent pairs of observations (y_i, x_i), the likelihood function is given by

$$L(\beta,\kappa_1,\ldots,\kappa_{J-1};y,x) = \prod_{i=1}^{n}\prod_{j=1}^{J}(\pi_{ij})^{d_{ij}}$$

Taking logarithms converts products into sums and we can write the log-likelihood function as

$$\log L(\beta,\kappa_1,\ldots,\kappa_{J-1};y,x) = \sum_{i=1}^{n}\sum_{j=1}^{J} d_{ij}\log\pi_{ij} \tag{6.13}$$

For the ordered probit model we plug in the probabilities in (6.5); for the ordered logit we use the probabilities in (6.6). ML estimation of the parameters yields consistent, asymptotically efficient, and asymptotically normally distributed estimators. We can use LR, Wald, and Score tests to test for general restrictions, and the invariance property together with the Delta method for estimation and inference of predicted probabilities, odds ratios, or marginal probability effects, which will be discussed in the next section, 6.2.5, as different ways of interpreting the ordered response model.

6.2.5 Interpretation of Parameters

As with all the models discussed in this book, it is of substantial interest to interpret the parameters of the ordered response model correctly. What does it mean for an element in β to be "large" or "small"? It is tempting to interpret coefficients in terms of the latent model, since this part of the model is known from any introductory course in econometrics as the linear regression model with dependent variable y_i^*. However, since the variance of the error term is normalized, β is only identified up to scale. Moreover, y_i^*, being an artificial construct, is not of interest in general. Potentially more interesting is a comparison based on **compensating variation**, the variation in two regressors such that the latent variable does not change. Let x_{il} denote the l-th element in x_i and β_l the corresponding parameter, and let m index the m-th

elements in both vectors x_i and β, respectively. Now consider a change in x_{il} and x_{im} at the same time, such that $\Delta y_i^* = 0$ (and therefore all probabilities are unchanged). This requires

$$0 = \beta_l \Delta x_{il} + \beta_m \Delta x_{im} \qquad \text{or} \qquad \frac{\Delta x_{il}}{\Delta x_{im}} = -\frac{\beta_m}{\beta_l} \tag{6.14}$$

If, for example, x_{il} is logarithmic income and x_{im} is a dummy variable indicating unemployment, then the above fraction gives a trade-off ratio: the relative increase in income required to compensate for the negative effect of unemployment (assuming that $\beta_l > 0$ and $\beta_m < 0$).

Exercise 6.1.

- In the above example, why does the fraction $-\beta_m/\beta_l$ give the *relative* increase in income to compensate for unemployment?

The interpretation is moved closer to the observed outcomes y_i if we take into account the threshold parameters κ_j. In some applications, it could be interesting to know how much an explanatory variable should change to reach the next higher response category. For this purpose, one could consider the ratio of the interval length to the parameter, $(\kappa_j - \kappa_{j-1})/\beta_l$. The smaller this ratio (in absolute terms), the smaller the maximum change in x_{il} required to move the response from $y_i = j$ to $y_i = j + 1$.

However, all these measures stop short of the most natural way of interpreting the parameters in conditional probability models such as ordered response models, namely in terms of discrete or marginal probability effects.

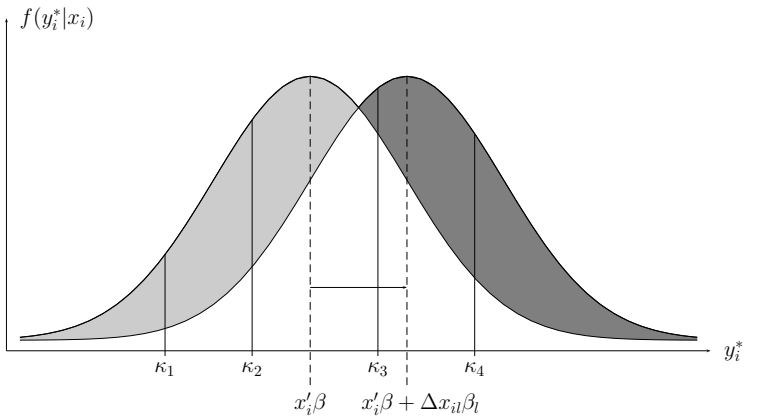
The *ceteris paribus* effect of a discrete change in one explanatory variable, say the l-th element in x_i, is simply the difference between the probabilities before and those after the change, given the values of the other variables. Mathematically,

$$\begin{aligned}\Delta\pi_{ij} &= P(y_i = j|x_i + \Delta x_{il}) - P(y_i = j|x_i) \\ &= \big[F(\kappa_j - x_i'\beta - \Delta x_{il}\beta_l) - F(\kappa_{j-1} - x_i'\beta - \Delta x_{il}\beta_l)\big] \\ &\quad - \big[F(\kappa_j - x_i'\beta) - F(\kappa_{j-1} - x_i'\beta)\big]\end{aligned} \tag{6.15}$$

This **discrete probability effect** can be illustrated graphically in two ways. First, consider Figure 6.5 which is based on Figure 6.2. The discrete change in x_{il} by $\Delta x_{il} > 0$ shifts the mean of the conditional density function of y_i^* to the right (given $\beta_l > 0$), or to the left (given $\beta_l < 0$). We plot a right shift in $f(y_i^*|x_i)$, thereby depicting the situation for a positive coefficient. The change in probabilities can be analyzed by investigating the light gray and dark gray shaded areas, where light gray indicates a decrease and dark gray an increase in the area under the density function due to the change in x_{il}.

The probabilities of observing $y_i = 1$ (the area left of κ_1) and $y_i = 2$ (the area between κ_1 and κ_2) clearly decrease since both areas decrease. The opposite holds for the probabilities of $y_i = 4$ and $y_i = 5$, both probabilities clearly increase since the relevant areas under the density function increase. In order to determine the sign of the probability effect for category $y_i = 3$, however, we have to weight a negative *and* a positive change in the area between κ_2 and κ_3. By visual inspection, the light gray area is larger than the dark gray area and we expect a negative probability effect for $y_i = 3$. As we will see below, it is a mathematical necessity that the sign of probability effects changes exactly once when moving from small to large outcomes of y_i, and that the sign is ambiguous for the category in which the mean $x_i'\beta$ lies.

Fig. 6.5. *Shift in Density Due to a Change $\Delta x_{il} > 0$ ($\beta_l > 0$)*

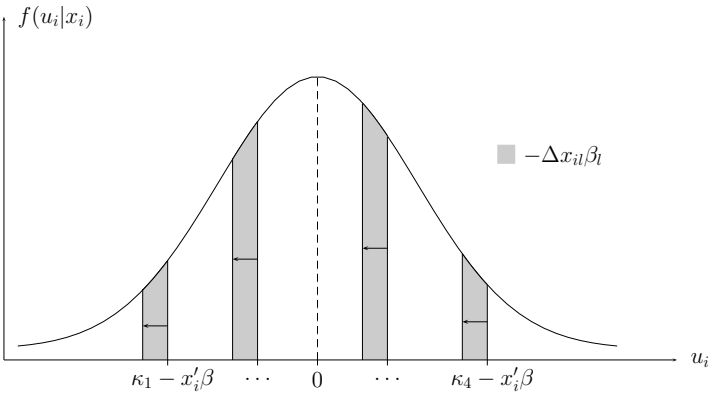


Exercise 6.2.

- Draw a diagram similar to Figure 6.5 with negative coefficient $\beta_l < 0$.

An alternative illustration of discrete probability effects is based on the conditional density of the error term, $f(u_i|x_i)$. In Figure 6.6 we plot the same density function as in Figure 6.3, and the change $\Delta x_{il} > 0$ now *shifts the thresholds* $\kappa_j - x_i'\beta$ to the left (again assuming a positive coefficient β_l). The change in the areas due to the change in one regressor is marked with the light gray shaded areas. Again, the sign of the probability effects is determined by these changes and we clearly have negative signs for $y_i = 1, 2$, clearly positive signs for $y_i = 4, 5$, and an ambiguous effect on the probability of $y_i = 3$.

Fig. 6.6. *Discrete Probability Effects in the Standard Model*



Exercise 6.3.

- Draw a diagram similar to Figure 6.6 with negative coefficient $\beta_l < 0$.

The **marginal probability effect** (MPE) of the l-th element in x_i can be obtained in general form from equation (6.4), or specifically from equations (6.5) and (6.6), by taking first derivatives,

$$MPE_{ijl} = \frac{\partial \pi_{ij}}{\partial x_{il}} = \big[f(\kappa_{j-1} - x_i'\beta) - f(\kappa_j - x_i'\beta)\big]\beta_l \tag{6.16}$$

with density function $f(z) = dF(z)/dz$. In general, the MPE's are functions of the covariates and therefore vary across individuals. In order to calculate **average marginal probability effects** (AMPE's) we have to take expectations of (6.16) with respect to x, which is estimated consistently by replacing β by its ML estimate $\hat{\beta}$ and averaging over the sample. Apart from calculating average effects, we can also report the effects evaluated at the averages or other interesting values, and thereby obtain the effect for a "typical" person. We discussed this issue already in Chapters 4 and 5 and therefore do not consider it further here. The MPE can be used to approximate the discrete change in a probability with $\Delta\pi_{ij} \approx MPE_{ijl}\Delta x_{il}$, and the smaller the absolute change in x_{il}, the better the approximation. Note that despite their intuitive appeal, discrete and marginal probability effects are rarely reported in empirical applications with ordinal data.

Exercise 6.4.

The specific structure of the ordered logit model allows for an alternative way of interpreting the parameters which is based on the odds given in equation (6.9).

- Derive the factor change in the odds for a general change Δx_{il} and the special case $\Delta x_{il} = 1$ (the odds ratio).
- Derive the relative change in the odds for a general change Δx_{il} and the special case $\Delta x_{il} = 1$.
- How would you interpret these measures?

Example 6.2. Secondary School Choice

In the Examples 5.2 and 5.4 we analyzed the secondary school choice of young people in Germany by means of a MNL model, and we identified the mother's educational level (in years of schooling) as an important factor. This was reflected by significantly higher probabilities of attending *Gymnasium*, the highest secondary school level, compared to *Realschule* and *Hauptschule* with increasing number of years the mother spent in formal schooling.

The dependent variable *school* with its three categorical outcomes incorporates ordering information. This results from the hierarchical structure of the German school system, which classifies school tracks into lower secondary school (*Hauptschule*), intermediate secondary school (*Realschule*), and upper secondary school (*Gymnasium*), with better long-term educational opportunities, the higher the schooling level. Hence, we could analyze the same research question as before with an ordered response model.

Table 6.1 displays the estimation results of the ordered probit and the ordered logit model, respectively. For each model, we get estimates of the regression vector (excluding a constant) and two threshold parameters $\hat{\kappa}_1$ and $\hat{\kappa}_2$. Since we estimated both models with ML, we can use a simple Wald test for the hypothesis of a zero parameter value. For example, the z-statistic of the ordered probit coefficient of mother's educational level is equal to 9.86, and therefore the parameter is significantly different from zero. The overall significance is tested using the reported LR test statistics and we conclude that both models significantly reject the model with only two thresholds.

In order to investigate the intergenerational transmission of education, we calculate the MPE of mother's educational level. Table 6.2 reports the MPE's in both models and compares them to the effects obtained in the MNL model. All models yield approximately the same MPE's, and on average, an increasing

Table 6.1. *Ordered Probit and Logit Estimates of Secondary School Choice*

Dependent variable: *school*		
	Ordered Probit	Ordered Logit
mother's educational level in years	0.276	0.475
	(0.028)	(0.052)
mother's employment level (0/1)	0.197	0.353
	(0.070)	(0.120)
logarithmic household income	0.662	1.165
	(0.104)	(0.191)
logarithmic household size	-0.685	-1.187
	(0.201)	(0.343)
birth order	-0.099	-0.160
	(0.058)	(0.098)
$\hat{\kappa}_1$	9.077	15.955
	(1.098)	(2.024)
$\hat{\kappa}_2$	10.012	17.521
	(1.106)	(2.045)
Log-likelihood value	-628.62	-627.42
LR (χ^2_{13})	208.45	210.84
Observations	675	675

Notes: Standard errors in parentheses. Further controls: year dummies 1995-2002

level of mother's schooling reduces the probability of attending *Hauptschule* and *Realschule*, and increases the probability that a child visits *Gymnasium*. More specifically, the AMPE of -0.0791 in the ordered probit model means that if the mother has one additional year of schooling, the probability of attending *Gymnasium* increases **on average by about 8.65 percentage points**.

Apart from point estimates, we also report the standard errors of the MPE's, which can be used, for example, to test whether an effect is statistically different from zero. Take the MPE of -0.0866 from the ordered probit model, with standard error 0.0104 in parentheses. The ratio of both forms a z-statistic of 8.33 and we significantly reject the null hypothesis of a zero MPE. Note that standard errors in the ordered probit and in the ordered logit model are roughly the same, whereas in the multinomial logit model, the standard errors are slightly higher. This is not a proof but it supports our hypothesis that if the dependent variable is indeed an ordered variable, then the multinomial logit model provides consistent results, but is inefficient compared to ordered response models.

Table 6.2. *Marginal Probability Effects of Mother's Educational Level*

	Hauptschule	*Realschule*	*Gymnasium*
Ordered Probit			
$\widehat{MPE}$	-0.0866	-0.0203	0.1069
	(0.0104)	(0.0062)	(0.0128)
$\widehat{AMPE}$	-0.0791	-0.0074	0.0865
	(0.0088)	(0.0035)	(0.0092)
Ordered Logit			
$\widehat{MPE}$	-0.0859	-0.0283	0.1142
	(0.0107)	(0.0085)	(0.0142)
$\widehat{AMPE}$	-0.0803	-0.0081	0.0884
	(0.0092)	(0.0041)	(0.0089)
Multinomial Logit			
$\widehat{MPE}$	-0.0941	-0.0242	0.1183
	(0.0145)	(0.0139)	(0.0163)
$\widehat{AMPE}$	-0.0800	-0.0091	0.0891
	(0.0127)	(0.0115)	(0.0098)

Notes: The MPE's are calculated by fixing the explanatory variables at their means, the AMPE's are calculated as average of the individual effects. Bootstrapped standard errors in parentheses.

Exercise 6.5.

Consider the estimation results in Table 6.1. In the ordered probit model, the value of the linear index, if evaluated at the mean of the explanatory variables, is 9.77.

- What is, for a thus defined average child, the predicted probability of attending *Hauptschule*, *Realschule* and *Gymnasium*, respectively?
- What is the *ceteris paribus* change in the outcome probabilities if the mother has one additional year of schooling?
- Calculate the discrete probability changes for a change in mother's schooling level from 9 to 10, from 10 to 13, and from 13 to 16 years. (Note that the mean of mother's schooling level is 11.44 years.)

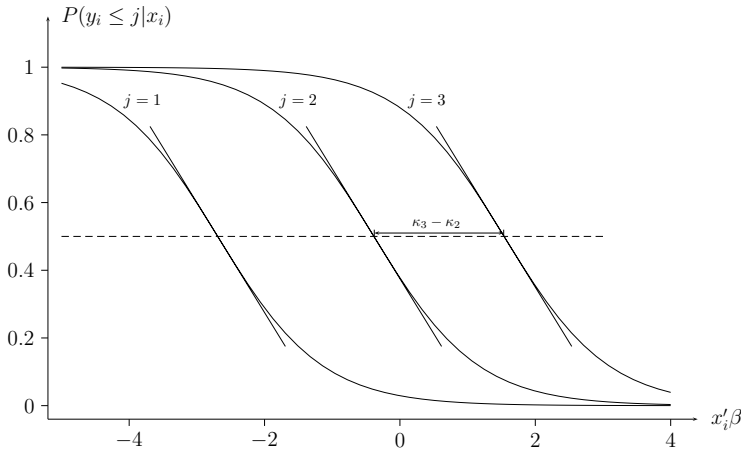
6.2.6 Single Indices and Parallel Regression

An often-raised criticism received by the ordered probit and ordered logit models is the implicitly imposed **parallel regression assumption**. This property is best understood by inspecting the cumulative probabilities of the events $y_i \leq j$, which are given by

$$P(y_i \leq j|x_i) = F(\kappa_j - x_i'\beta) \qquad j = 1, \ldots, J-1 \tag{6.17}$$

In Figure 6.7 we draw three such cumulative probabilities with different threshold values $\kappa_1 < \kappa_2 < \kappa_3$ and linear index $x_i'\beta$ on the horizontal axis. Now compare, for example, the two cumulative probability functions $P(y_i \leq 2|x_i)$ and $P(y_i \leq 3|x_i)$. These functions only differ in their constant threshold values with equal linear indices. This, in turn, implies that $P(y_i \leq 3|x_i)$ is obtained by a parallel shift of $P(y_i \leq 2|x_i)$ to the right by $\kappa_3 - \kappa_2$ units, but the function's shape is not changed at all. Hence, the slopes for a given value on the vertical axis (marked by the dashed line) remain the same for each cumulative probability, and this property is referred to as parallel regression.

Fig. 6.7. *Parallel Regression Assumption*



Mathematically, if we fix the cumulative probabilities to a particular value, say $F(\kappa_j - x_i'\beta) = 0.5$, then we *must* have $\kappa_j - x_i'\beta = 0$ for all $j = 1, \ldots, J-1$. But in this case, the slopes of the cumulative probabilities do not vary by j which can be seen from the first derivative with respect to $x_i'\beta$

$$\frac{\partial P(y_i \leq j|x_i)}{\partial(x_i'\beta)} = -f(\kappa_j - x_i'\beta) = -f(0) \qquad j = 1, \ldots, J-1$$

Likewise, the partial derivatives of (6.17) with respect to the l-th element in x_i are constant across j since

$$\frac{\partial P(y_i \leq j|x_i)}{\partial x_{il}} = -f(0)\beta_l \qquad j = 1, \ldots, J-1$$

The heart of the matter is the **single index assumption**, which means that the linear index of regressors $x_i'\beta$ appears as a common element in the argument of *each* cumulative probability. In terms of parallel regression, this allows us to plot all functions $P(y_i \leq j|x_i)$ in the same diagram with the horizontal distance determined by the difference in the constant threshold parameters but otherwise equal functional shape.

But what implications does the single index assumption have for the interpretation of parameters in terms of marginal probability effects? To pinpoint why single indices restrict the analysis of ordered responses, one should first observe that the relative magnitude of MPE's, i.e., the slope of the iso-probability curve (see Chapter 2.2.3), is not allowed to vary across outcomes and individuals:

$$\frac{MPE_{ijl}}{MPE_{ijm}} = \frac{\beta_l}{\beta_m} \tag{6.18}$$

This ratio does not depend on j, nor does it depend on x_i. In other words, the **relative marginal probability effects** are constant and the same in each part of the outcome distribution, that is for each j. It is not possible, for example, that x_{il} (e.g., income) is relatively more important than x_{im} (e.g., unemployment) for low responses than it is for high ones. This property holds for whatever choice of $F(u)$.

A second way to illustrate the limitations of the standard models is to examine the sign of MPE's. With $F(u)$ being either the standard normal or logistic distribution, the density function $f(u)$ is bell-shaped with maximum at 0. It follows from equation (6.4) and Figure 6.5 that

$$\begin{aligned} \operatorname{sgn}(MPE_{ijl}) &= -\operatorname{sgn}(\beta_l) \quad \text{if} \quad \kappa_{j-1} < x_i'\beta \ \text{ and } \ \kappa_j \leq x_i'\beta \\ \operatorname{sgn}(MPE_{ijl}) &= \operatorname{sgn}(\beta_l) \quad \text{if} \quad \kappa_{j-1} \geq x_i'\beta \ \text{ and } \ \kappa_j > x_i'\beta \end{aligned} \tag{6.19}$$

where $\operatorname{sgn}(z)$ is a function extracting the sign of z. If $\kappa_{j-1} < x_i'\beta$ and $\kappa_j > x_i'\beta$, the sign of the MPE is indeterminate. This is referred to as **single crossing property**: as one moves from the probability of the smallest outcome to the probability of the largest outcome, the MPE's are either first negative and then positive, or they are first positive and then negative. A reversion is precluded by the assumptions of the model.

Brant (1990) proposes a procedure to test the single index assumption based on binary regressions. However, we do not consider this test further here and discuss a more general ordered response model in the following section, 6.3, which can be used to test single indices with a familiar LR test.

6.3 Generalized Threshold Models

Had we selected the topics of this book based on past and current practice in empirical economic research with ordinal data only, we could have ended this chapter here. The standard ordered probit and logit models fully dominate the literature. The relative simplicity of these models, to which they certainly owe their popularity, is also a problem, however. In many applications, they may be too simple, precluding a meaningful interpretation of the parameters, and generalized models, developed in other fields, may increasingly find their way into mainstream econometric work as well.

6.3.1 Generalized Ordered Logit and Probit Models

The single index assumption in the ordered probit and logit models restricts the modeling of ordered responses, in particular when considering marginal probability effects. When searching for more flexible response patterns, one will need to look for a richer parametric model where index functions are allowed to vary across response categories. In this section, we present a very flexible ordered response model that relaxes this assumption.

The basic idea is to make the threshold parameters linear functions of the covariates (see Maddala, 1983, and Terza, 1985). Let

$$\kappa_{ij} = \tilde{\kappa}_j + x_i'\gamma_j \qquad j = 1, \ldots, J \tag{6.20}$$

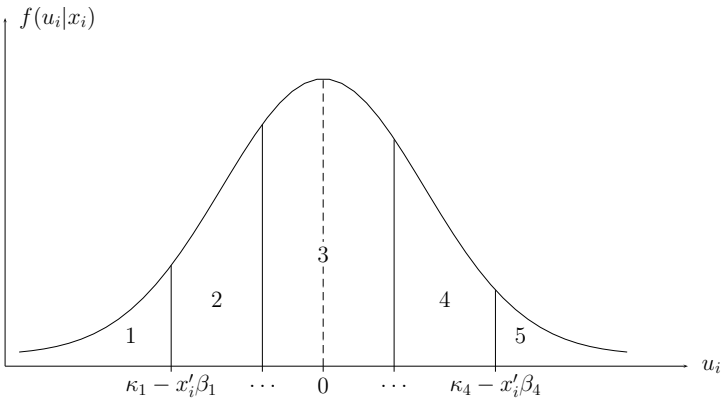
where x_i is a $(k \times 1)$-dimensional vector of explanatory variables excluding a constant, as before. Because of the assumption in (6.20), the model is referred to as a **generalized threshold model**. Substitution of κ_{ij} for κ_j in equation (6.4) yields

$$\begin{aligned}\pi_{ij} &= F(\tilde{\kappa}_j + x_i'\gamma_j - x_i'\beta) - F(\tilde{\kappa}_{j-1} + x_i'\gamma_{j-1} - x_i'\beta) \\ &= F(\tilde{\kappa}_j - x_i'\beta_j) - F(\tilde{\kappa}_{j-1} - x_i'\beta_{j-1})\end{aligned} \tag{6.21}$$

where $\beta_j = \beta - \gamma_j$ since we cannot identify β and γ_j separately, and it is understood that $\tilde{\kappa}_0 = -\infty$ and $\tilde{\kappa}_J = \infty$ such that $F(-\infty) = 0$ and $F(\infty) = 1$.

Figure 6.8 illustrates the greater flexibility in modeling ordered responses. The parameter vectors are now allowed to vary across outcomes and we have a specific vector β_j for each threshold. Once we have specified the distribution function $F(u)$, we obtain the probability of a particular outcome j by integrating over the density function $f(u)$ with integration limits $\kappa_{j-1} - x_i'\beta_{j-1}$ and $\kappa_j - x_i'\beta_j$. Like in the standard model, two distributional assumptions come to mind: the standard normal distribution $\Phi(u)$, which yields the **generalized ordered probit model** and the logistic distribution $\Lambda(u)$, which yields the **generalized ordered logit model**.

Fig. 6.8. *Generalized Threshold Mechanism*



From the probability function in (6.21) and Figure 6.8 we can see that the model now contains $J-1$ parameter vectors $\beta_1, \ldots, \beta_{J-1}$ plus $J-1$ thresholds $\tilde{\kappa}_1, \ldots, \tilde{\kappa}_{J-1}$ that can be estimated jointly by maximum likelihood. Under the hypothesis

$$H_0: \quad \beta_1 = \ldots = \beta_{J-1} \tag{6.22}$$

the generalized threshold model nests the standard model. This can be used to discriminate between the two models with a simple LR test, which implicitly tests for the single index assumption. Clearly, the proliferation of parameters, in particular when J is large, is a potential disadvantage of the generalized model. However, a LR test can be conducted, and one can economize by imposing partial restrictions, such as $\beta_2 = \beta_3$, while allowing parameters to differ in other parts of the outcome distribution.

6.3.2 Interpretation of Parameters

The generalized threshold model provides much more flexibility in analyzing the effects of explanatory variables in the outcome probabilities. This can be seen in two ways. First, consider the cumulative probability of the event $y_i \leq j$ which is given by

$$P(y_i \leq j|x_i) = F(\tilde{\kappa}_j - x_i'\beta_j) \tag{6.23}$$

In Section 6.2.6, we stated that the first derivatives of the cumulative probabilities with respect to any regressor x_{il} are equal for all j when fixing the probability to a certain value. But this does not need to hold in the generalized model since

$$\frac{\partial P(y_i \leq j|x_i)}{\partial x_{il}} = -f(\tilde{\kappa}_j - x_i'\beta_j)\beta_{jl} \tag{6.24}$$

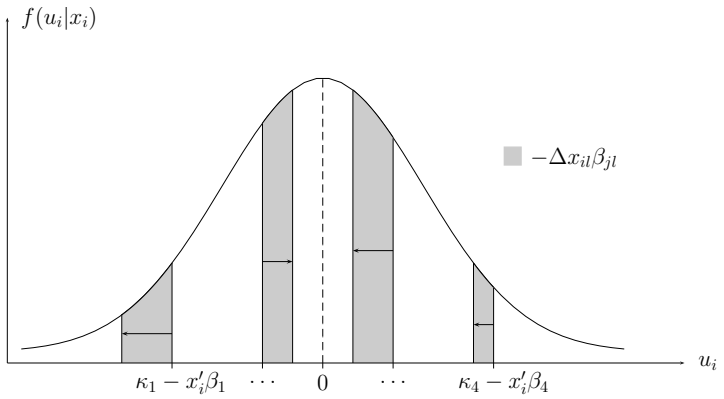
Even if $\tilde{\kappa}_j - x_i'\beta_j$ is fixed $\forall j$, β_{jl} possibly varies across j and therefore does not impose the parallel regression assumption. Second, consider the MPE of one regressor x_{il}, which is given by

$$MPE_{ijl} = f(\tilde{\kappa}_{j-1} - x_i'\beta_{j-1})\beta_{j-1,l} - f(\tilde{\kappa}_j - x_i'\beta_j)\beta_{jl} \tag{6.25}$$

From (6.25) it follows that the relative magnitude of marginal probability effects, MPE_{ijl}/MPE_{ijm}, no longer needs to be constant, and that MPE's do not need to switch the sign exactly once when moving from the lowest category to the highest. Rather, these effects can be determined empirically.

Figure 6.9 illustrates what probability effects might look like in the generalized model. It is important to note that the shift in thresholds does not need to be constant and in the same direction for each of the $J-1$ thresholds.

Fig. 6.9. *Discrete Probability Effects in the Generalized Model*



Exercise 6.6.

- Can you make a statement about the sign and relative magnitude of the parameters β_{1l}, β_{2l}, β_{3l}, and β_{4l} in Figure 6.9?

More specific results are obtained in the generalized ordered logit model in which $P(y_i \leq j|x_i) = \Lambda(\tilde{\kappa}_j - x_i'\beta_j)$, and therefore the odds of $y_i \leq j$ versus $y_i > j$ can be written as

$$\frac{P(y_i \leq j|x_i)}{P(y_i > j|x_i)} = \frac{\Lambda(\tilde{\kappa}_j - x_i'\beta_j)}{1 - \Lambda(\tilde{\kappa}_j - x_i'\beta_j)} = \exp(\tilde{\kappa}_j - x_i'\beta_j) \tag{6.26}$$

Equation (6.26) shows that the effects of covariates on the log-odds are now category-specific.

As mentioned before, the flexible modeling of ordered responses with generalized thresholds does not come without costs. First, the constraint of an ascending order of thresholds in the standard model to obtain a well-defined probability function now extends to

$$\kappa_j - x_i'\beta_j \geq \kappa_{j-1} - x_i'\beta_{j-1} \tag{6.27}$$

Obviously, not only are the constants restricted in a hierarchical manner; the slope parameters must also be taken into account. Second, fulfilling (6.27) together with a larger number of parameters in the model increases computation times considerably.

Example 6.3. Income and Happiness

In the introduction of this chapter, we gave an example of a survey question in which respondents were asked to state their satisfaction with life in general, with answer categories ranging from *completely dissatisfied* "0" to *completely satisfied* "10" on the eleven-point scale (see Figure 6.1). In the literature on reported subjective well-being this variable is usually referred to as *happiness*. In particular, the relationship between income and happiness in cross-sectional studies has gained much attention in the past with the common finding that increasing income raises a person's happiness (see, for example, Frey and Stutzer, 2002). Since the dependent variable is measured on an ordinal scale, standard ordered response models are applied. However, from Section 6.2.6 we know that these models impose single indices precluding a flexible analysis in terms of MPE's, and it might be interesting to compare the results of the standard models with those of the generalized alternatives.

Table 6.3. *Recoding of Happiness Responses and Relative Frequencies*

New scale	Original scale	Relative Frequencies
(1) or *very unhappy*	"0" to "5"	0.246
(2) or *unhappy*	"6" or "7"	0.341
(3) or *happy*	"8"	0.266
(4) or *very happy*	"9" or "10"	0.147

Using data from the German Socio-Economic Panel (GSOEP) waves 1984 to 1997, we analyze the effect of logarithmic real income on happiness. We focus on single men and control for a second-order polynomial in *age*, and two dummy variables indicating *good health* status and whether the person is *unemployed*. For simplification, we reduce the total number of categories to $J = 4$ by recoding the dependent variable as listed in Table 6.3.

Table 6.4 reports the results of the ordered logit and the generalized ordered logit models. Since the dependent variable has four categories, we get three threshold parameters in the standard model, and three separate parameter vectors in columns (1) – (3) for the generalized model. Looking at the estimation results, we recognize that the parameters of some variables (like age or unemployment) do not differ much between the two models, whereas the three income coefficients in the generalized model change considerably compared to the standard model. We can formally test for these differences by conducting a likelihood ratio test. The appropriate test statistic for the null hypothesis $\beta_1 = \beta_2 = \beta_3$ is twice the difference of the log-likelihood values, or 25.54. With a p-value of 0.004 we can clearly reject the standard model in favor of the generalized model, and thereby reject the hypothesis of single indices. Alternatively, we can test whether the parameters of income are the only ones that differ across j with equal parameters for all other variables.

Table 6.4. *(Generalized) Ordered Logit Estimates of Happiness Equation*

Dependent variable: *happiness*				
	Ordered	Generalized Ordered Logit		
	Logit	(1)	(2)	(3)
logarithmic real income	0.064	0.133	0.038	-0.026
	(0.058)	(0.059)	(0.059)	(0.081)
age	-0.074	-0.085	-0.064	-0.060
	(0.017)	(0.020	(0.018)	(0.024)
$age^2 \times 10^{-2}$	0.077	0.083	0.068	0.068
	(0.016)	(0.018	(0.016)	(0.022)
good health	1.748	1.763	1.503	1.812
	(0.146)	(0.146	(0.158)	(0.294)
unemployed	-1.223	-1.210	-1.287	-1.010
	(0.165)	(0.182	(0.192)	(0.320)
$\hat{\kappa}_1$	-0.559	-0.222		
	(0.606)	(0.691)		
$\hat{\kappa}_2$	1.052		0.757	
	(0.608)		(0.641)	
$\hat{\kappa}_3$	2.527			2.121
	(0.610)			(0.856)
Log-likelihood value	-5,680.08		-5,667.31	
LR	220.70		543.14	
Observations	4,413		4,413	

Notes: Standard errors in parentheses. Data: GSOEP 1984–1997.

But what implications do these statistical differences in parameters have for their interpretation? In Table 6.5, we report the MPE's of income on happiness. In the standard model, we obtain negative effects on the categories *very unhappy* and *unhappy*, and positive effects on the categories *happy* and *very happy*. Specifically, $\widehat{MPE} = -0.0114$ means that the probability of being *very unhappy* decreases by 0.0114 percentage points if income is increased by 1 percent, given that all regressors are fixed to their means, i.e., given the characteristics of an "average" individual. Similar, $\widehat{AMPE} = 0.0078$ means that on average the probability of being *very happy* increases by 0.0078 percentage points given a 1 percent increase in income.

Table 6.5. *Marginal Effects of Income on Happiness*

	4 Categories in original scale from 0–10			
	0–5	6/7	8	9/10
	very unhappy	*unhappy*	*happy*	*very happy*
Ordered Logit				
$\widehat{MPE}$	-0.0114	-0.0039	0.0081	0.0073
	(0.0069)	(0.0024)	(0.0049)	(0.0044)
$\widehat{AMPE}$	-0.0107	-0.0038	0.0067	0.0078
	(0.0065)	(0.0023)	(0.0041)	(0.0047)
Generalized Ordered Logit				
$\widehat{MPE}$	-0.0236	0.0144	0.0121	-0.0029
	(0.0077)	(0.0091)	(0.0100)	(0.0070)
$\widehat{AMPE}$	-0.0220	0.0133	0.0118	-0.0031
	(0.0072)	(0.0086)	(0.0097)	(0.0074)
Multinomial Logit				
$\widehat{MPE}$	-0.0242	0.0168	0.0115	-0.0041
	(0.0078)	(0.0099)	(0.0117)	(0.0078)
$\widehat{AMPE}$	-0.0224	0.0162	0.0109	-0.0048
	(0.0072)	(0.0095)	(0.0114)	(0.0082)

Notes: The MPE's are calculated by fixing the explanatory variables at their means, the AMPE's are calculated as the average of the individual effects. Bootstrapped standard errors in parentheses.

In contrast to that, the generalized model yields a negative MPE on the highest category. For example, increasing income by 1 percent reduces the probability of being *very happy* on average by -0.0031 percentage points. By assumption, such a negative MPE is ruled out in the standard model with a single index and positive income coefficient. Furthermore, the negative effect on the probability of being *very unhappy* is higher than in the standard model,

and the effect of income on the probability of being *unhappy* is positive rather than negative.

In addition to ordered response models we also calculate the MPE's obtained from the MNL model, which turn out to be very similar to those in the generalized threshold model with slightly higher standard errors. This confirms our hypothesis of a violated single index assumption. We know that the standard model is inconsistent if this assumption is not supported by the data, and the generalized threshold model offers a consistent alternative appropriately accounting for the ordered information in happiness responses. The MNL model is consistent whether or not the data fulfill the single index assumption, but inefficient when the dependent variable is indeed ordered.

Exercise 6.7.

- From the information given in Table 6.5, can you determine the change in probabilities associated with a doubling of income? If so, how? If not, why not?

6.4 Sequential Models

6.4.1 Modeling Conditional Transitions

In some applications, the ordinal information of the dependent variable arises from a sequential mechanism in which categories can only be reached successively. Consider, for example, a classification of unemployment with the categories (1) *short-term unemployment*, (2) *medium-term unemployment*, and (3) *long-term unemployment*, and suppose we are interested in the determinants of being in one of the three states. Obviously, if a person is currently classified in medium-term unemployment, she must have been classified in short-term unemployment before. With the same argument, Category 3 can only be reached if the person was classified in Categories 1 and 2 before. In this section, we will discuss models for such ordinal outcomes which explicitly account for the ordered *and* sequential information in the response variable.

The basic idea of a **sequential model**(Fienberg, 1980, Tutz, 1991) is that of an underlying response mechanism starting in the first category. The individual decides whether to stay in the first category, or to choose a higher one. This binary decision is coded by 0/1 and $y_i^{(1)} = 1$ indicates whether the process stops in the first category, which implies that we observe $y_i = 1$, or if the process continues, $y_i^{(1)} = 0$. Conditional on the continuation of the

process, the transition from the second to the third category is determined by $y_i = 2$ against $y_i > 2$, and the process stops if $y_i^{(2)} = 1$, and so on. In general, the transition from category j to $j+1$ for $j = 1, \ldots, J-1$ is indicated by

$$y_i^{(j)} = \begin{cases} 1 & \text{if process stops, no transition to category } j+1 \\ 0 & \text{if process continues, transition to category } j+1 \end{cases} \tag{6.28}$$

Note that category J is chosen with probability one given that category $J-1$ is rejected, and therefore we do not need to model the last transition. To summarize, we assume a **conditional model of transitions**, i.e., given that category j is reached, does the process stop, or continue?

Sequential models can also be motivated by concepts of discrete duration analysis. A common approach in the literature of grouped duration data is to set an indicator variable equal to one if an individual exits the initial state in a certain *time interval*, and zero if the individual remains in the initial state. We will discuss these kinds of models in Chapter 8.

In order to formulate a conditional probability model for the ordered responses we start with the conditional transition probabilities. The probability of event $y_i = j$ conditional on $y_i \geq j$, i.e., conditional on "the process having continued $j-1$ times", is assumed to be

$$P(y_i = j | y_i \geq j, x_i) = F(\alpha_j + x_i'\beta) \tag{6.29}$$

where α_j is a category-specific constant, $x_i'\beta$ is the linear index of explanatory variables (excluding a constant), and $F(\cdot)$ is a monotonic transformation mapping $\alpha_j + x_i'\beta$ onto the unit interval. Three special cases of this model are prevalent in the literature, depending on the assumption about $F(\alpha_j + x_i'\beta)$.

- **Sequential Logit Model**

  with $F(\alpha_j + x_i'\beta) = \Lambda(\alpha_j + x_i'\beta)$

- **Sequential Probit Model**

  with $F(\alpha_j + x_i'\beta) = \Phi(\alpha_j + x_i'\beta)$

- **Proportional Hazards Model**

  with $F(\alpha_j + x_i'\beta) = 1 - \exp(-\exp(\alpha_j + x_i'\beta))$

In general, sequential models differ from the threshold models as presented in the preceding sections, 6.2 and 6.3. However, Läärä and Matthews (1985) show the observational equivalence of the proportional hazards model and a threshold model with type-I extreme value distribution $F(u) = 1 - \exp(-\exp(u))$ and single index.

Exercise 6.8.

- Critically evaluate a **linear sequential model** in which it is assumed that $F(\alpha_j + x_i'\beta) = \alpha_j + x_i'\beta$.

Once we have determined the conditional transition probabilities, the unconditional probabilities $P(y_i = j|x_i)$, i.e., unconditional on $y_i \geq j$ but still conditional on x_i, are obtained from the recursive relationship

$$\pi_{ij} = P(y_i = j|x_i) = P(y_i = j|y_i \geq j, x_i)P(y_i \geq j|x_i) \tag{6.30}$$

In fact, it turns out that the conditional transition probabilities *fully* characterize the probability function of y_i. For example,

$$\begin{aligned}
P(y_i = 1) &= P(y_i = 1|y_i \geq 1)P(y_i \geq 1) \;=\; P(y_i = 1|y_i \geq 1) \\
P(y_i = 2) &= P(y_i = 2|y_i \geq 2)P(y_i \geq 2) \\
&= P(y_i = 2|y_i \geq 2)[1 - P(y_i = 1|y_i \geq 1)] \\
P(y_i = 3) &= P(y_i = 3|y_i \geq 3)P(y_i \geq 3) \\
&= P(y_i = 3|y_i \geq 3)[1 - P(y_i = 1) - P(y_i = 2)] \\
&= P(y_i = 3|y_i \geq 3)[1 - P(y_i = 1|y_i \geq 1)][1 - P(y_i = 2|y_i \geq 2)]
\end{aligned}$$

where we dropped x_i in the conditioning for notational simplicity. In general, we can write

$$\begin{aligned}
\pi_{ij} &= P(y_i = j|y_i \geq j, x_i) \prod_{r=0}^{j-1} [1 - P(y_i = r|y_i \geq r, x_i)] \\
&= F(\alpha_j + x_i'\beta) \prod_{r=0}^{j-1} [1 - F(\alpha_r + x_i'\beta)]
\end{aligned} \tag{6.31}$$

and it is understood that $\alpha_0 = -\infty$ such that $F(-\infty) = 0$. An important feature of the sequential model is that no restrictions on the parameter space are required to ensure the existence of a proper probability function. This simplifies estimation considerably. On the downside, due to the increasing number of terms in (6.31), the calculation of marginal probability effects becomes somewhat intractable, in particular for large J, since repeated application of product and chain rules is required. We will discuss both issues, estimation and interpretation of parameters, after the following section, 6.4.2, which presents a useful extension of the basic model.

6.4.2 Generalized Conditional Transition Probabilities

A natural generalization of the basic model is to assume category-specific parameter vectors β_j, again to allow for different effects on the conditional transition probabilities. In this case, the probability of any $y_i = j$ conditional on the process having continued $j-1$ times can be specified as

$$P(y_i = j | y_i \geq j, x_i) = F(\alpha_j + x_i' \beta_j) \tag{6.32}$$

which yields the probability function

$$P(y_i = j | x_i) = F(\alpha_j + x_i' \beta_j) \prod_{r=0}^{j-1} [1 - F(\alpha_r + x_i' \beta_r)] \tag{6.33}$$

Again, it is understood that $\alpha_0 = -\infty$ such that $F(-\infty) = 0$. The equivalence of proportional hazards and proportional odds model does not hold anymore when parameters are different in each conditional transition.

6.4.3 Marginal Effects

The calculation of marginal probability effects in the standard or generalized sequential model is not especially easy due to the increasing number of terms in (6.31) and (6.33). In particular for large J, repeated product and chain rules are required, and terms become more and more complicated. However, it turns out that MPE's in sequential models should be calculated successively, starting with the first category,

$$MPE_{i1l} = f(\alpha_1 + x_i' \beta_1) \beta_{1l} \tag{6.34}$$

where $f(z) = dF(z)/dz$. Equation (6.34) gives the approximate change in the probability given a one-unit increase in the l-th element of x_i. We directly report the MPE of the generalized model with category-specific parameter vector β_1; the effect in the standard model can be obtained as a special case if we replace β_1 by β. For each of the $j = 2, \ldots, J$ remaining categories, the MPE's can be calculated as

$$\begin{aligned} MPE_{ijl} = {} & f(\alpha_j + x_i' \beta_j) \beta_{jl} \prod_{r=1}^{j-1} [1 - F(\alpha_r + x_i' \beta_r)] \\ & - F(\alpha_j + x_i' \beta_j) \sum_{r=1}^{j-1} MPE_{irl} \end{aligned} \tag{6.35}$$

where the MPE's in the basic model are obtained if we let $\beta_1 = \ldots = \beta_{J-1}$. It follows from (6.34) and (6.35) that in the generalized sequential model, like in

the generalized threshold model, the effects are local, as the parameter vectors possibly (but not necessarily) vary by category.

Exercise 6.9.

- Verify the validity of equation (6.35) and that it can be used for the MPE's in the basic model under the restriction $\beta_1 = \ldots = \beta_{J-1}$.

6.4.4 Estimation

Estimation of sequential models is more a matter of appropriately transforming the dataset than of handling multiple products in the conditional probability function. From Section 6.4.1 we know that the sequential mechanism reduces the overall model in a sequence of binary decisions, and the binary variable indicating whether the process stops or continues in category j can be used as dependent variable in each binary decision.

In the basic model, the parameter vector β is identical for all categories, only the intercepts α_j differ across j. In order to illustrate the data handling for the purpose of estimation, we consider a simple artificial dataset. Let the total number of categories be given by $J = 4$, and assume that two covariates x_1, x_2 are available. The information for a total of $n = 4$ observations is summarized in Table 6.6.

Table 6.6. *Artificial Dataset*

i	y_i	x_{i1}	x_{i2}
1	2	3.2	5.4
2	1	1.1	6.8
3	4	2.5	4.2
4	3	1.7	0.3

We now expand the data such that the maximum of transitions for each observation is related to the ordered response variable. Let α_j denote a dummy variable indicating the transitions $j = 1, \ldots, J-1$ and index the expanded data rows by *ind*. Then, the data have to be organized as in Table 6.7.

The first observation has been doubled since the process continued once (from Category 1 to Category 2), and then stopped. The second observation has not been expanded since $y_i = 1$. The fourth observation has been tripled since two transitions occurred before the process stopped. The third observation has been tripled as well although actual outcome was $y_3 = 4$. The reason is that reaching the last category is implied by not stopping in the

Table 6.7. *Data Organization in the Standard Sequential Model*

i (ind)	y_i	$y_i^{(j)}$	α_1	α_2	α_3	x_{i1}	x_{i2}
1 (1)	2	0	1	0	0	3.2	5.4
1 (2)	2	1	0	1	0	3.2	5.4
2 (1)	1	1	1	0	0	1.1	6.8
3 (1)	4	0	1	0	0	2.5	4.2
3 (2)	4	0	0	1	0	2.5	4.2
3 (3)	4	0	0	0	1	2.5	4.2
4 (1)	3	0	1	0	0	1.7	0.3
4 (2)	3	0	0	1	0	1.7	0.3
4 (3)	3	1	0	0	1	1.7	0.3

previous category, which has been indicated by setting the dummy $y_i^{(3)} = 0$. Now, the model can be estimated by a binary regression model, for example the binary logit. As dependent variable we choose the binary outcome $y_i^{(j)}$, as explanatory variables we include the dummy variables α_j (the category-specific constants) and x_1, x_2, and we need to cluster the observations by i in order to obtain valid standard errors.

In the generalized model, data handling is somewhat different, although not substantially more difficult. Consider the same example as before. Taking into account that parameters β_j now differ among the ordinal responses, we explicitly include the expanding index for each observation in the regression. Hence, we rewrite the data as in Table 6.8.

Table 6.8. *Data Organization in the Generalized Sequential Model*

i	ind	y_i	$y_i^{(j)}$	x_{1i}	x_{2i}
1	1	2	0	3.2	5.2
1	2	2	1	3.2	5.2
2	1	1	1	1.1	6.8
3	1	4	0	2.5	4.2
3	2	4	0	2.5	4.2
3	3	4	0	2.5	4.2
4	1	3	0	1.7	0.2
4	2	3	0	1.7	0.2
4	3	3	1	1.7	0.2

The estimation procedure is as follows: we split estimation into $J - 1$ steps, i.e., we estimate $J - 1$ binary regression models conditional on the particular values of ind. In the first step, we include all observations since all observations have at least $ind = 1$. In the second step, we drop observations with $y_i = 1$, since these observations were not expanded and include only those with $ind = 2$. In the third step, we also drop all observations with

$y_i = 2$, and so on. We call the included observations "at risk" since only these observations might continue in the process. In each step, we use $y_i^{(j)}$ as the dependent variable and include the regressors x_1 and x_2 as well as an intercept in the binary model, thereby getting category-specific parameters α_j and β_j. The structure of the generalized sequential model allows for simply adding up the log-likelihood values of each binary step to obtain the overall log-likelihood. This value can be used for model selection, e.g., to discriminate between the generalized sequential and the generalized threshold model.

6.5 Interval Data

We conclude this chapter by considering a special kind of ordered dependent variables. As already mentioned in the introduction, the coarsening of a variable, for example due to limited data availability, might produce an ordered and discrete outcome, although the variable could, in principle, have been measured continuously. A typical example is income data in household surveys. People may find it easier to state their income *class* rather than their exact income (and therefore express less reservations). Consider, for instance, the categorization scheme in Table 6.9.

Table 6.9. *Example of an Income Classification*

Code (y_i)	Income class
1	$y_i^* \leq 30000$
2	$30000 < y_i^* \leq 60000$
3	$60000 < y_i^* \leq 100000$
4	$100000 < y_i^*$

Such variables are also called **interval data**. Formally, this setting is closely related to the standard ordered response models as presented in Section 6.2. Again, a dependent variable with a relatively small number of ordered (otherwise arbitrarily coded) outcomes is observed, and it is useful to think of a latent model

$$y_i^* = x_i'\beta + u_i \qquad u_i \sim N(0, \sigma^2) \tag{6.36}$$

describing, for example, the well-known linear model of a Mincerian earnings function. Let $\kappa_0, \ldots, \kappa_J$ denote the threshold values and assume again that $\kappa_0 = -\infty$ and $\kappa_J = \infty$. The specific feature of interval data is that all remaining thresholds are known as well; in the above example they are given by $\kappa_1 = 30'$, $\kappa_2 = 60'$, and $\kappa_3 = 100'$, unlike in the standard model in which the κ_j's are unknown and have to be estimated from the data.

The latent model is converted into observed responses, using the same classification rule as in equation (6.3). We have

$$\begin{aligned}\pi_{ij} = P(y_i = j|x_i) &= P(\kappa_{j-1} < y_i^* \leq \kappa_j|x_i) \\ &= P(y_i^* \leq \kappa_j|x_i) - P(y_i^* \leq \kappa_{j-1}|x_i) \end{aligned} \tag{6.37}$$

However, since the threshold values are known, the situation changes somewhat. Consider the probability that $y_i^* \leq \kappa_j$,

$$\begin{aligned} P(y_i^* \leq \kappa_j|x_i) &= P(u_i \leq \kappa_j - x_i'\beta|x_i) \\ &= P\left(\frac{u_i}{\sigma} \leq \frac{\kappa_j - x_i'\beta}{\sigma}\Big|x_i\right) \\ &= \Phi\left(\frac{\kappa_j - x_i'\beta}{\sigma}\Big|x_i\right) \end{aligned} \tag{6.38}$$

where the last equality follows from the normality assumption of the error terms. In contrast to the ordered probit model, we are now able to identify all parameters of the latent model, i.e., we can estimate β as well as σ.

Table 6.10. *Log-Transformation of Income Classification*

Code (y_i)	Income class	Transformed income class
1	$y_i^* \leq 30000$	$\log y_i^* \leq 10.3$
2	$30000 < y_i^* \leq 60000$	$10.3 < \log y_i^* \leq 11.0$
3	$60000 < y_i^* \leq 100000$	$11.0 < \log y_i^* \leq 11.5$
4	$100000 < y_i^*$	$11.5 < \log y_i^*$

Of course, this method relies on the knowledge of the distribution of u_i (and thereby that of y_i^*). In the income example, we assumed that y_i^*, conditional on x_i, was normally distributed. This assumption is not realistic since the normal distribution is symmetric, whereas income distributions tend to be skewed to the left, with a steep increase on the left and a long right tail. Moreover, with y_i^* being normally distributed, negative income values are possible. These features can be accommodated if we let the latent model be given by

$$\log y_i^* = x_i'\beta + u_i \qquad u_i \sim N(0, \sigma^2) \tag{6.39}$$

If $\log y_i^*$ is normally distributed, then y_i^* has a **log-normal distribution** with the desired features (positive, skewed to the left). In the above example, the threshold values have to be modified and the classification scheme would look like the one in Table 6.10. With this transformation of the threshold values, we can apply the same principles as described in Section 6.2.

6.6 Further Exercises

Exercise 6.10 Discuss the use of ordered response models in the following empirical applications (see also Checkovich and Stern, 2002):

a) In what sense can the variable *number of hours devoted per day to a specific activity* (such as taking care of ones elderly parents) be interpreted as an ordinal variable?

b) Assume that a dataset contains a number of health impairment indicators such as: problem getting in or out of bed (yes/no), problem dressing (yes/no), problem eating (yes/no), etc. Such data are often analyzed with ordered probit models – Explain.

Exercise 6.11 Derive a general expression of the score function and the Hessian matrix in the ordered probit and the ordered logit model. How can you obtain the ML estimators of β and κ in both models?

Exercise 6.12 Suppose you want to estimate an ordered response model in a dataset with the following response pattern:

y	(1) *bad*	(2) *so-so*	(3) *good*
n	26	53	31

Suppose the model has been specified as an ordered logit model with two threshold parameters, κ_1 and κ_2, and without further covariates.

a) Derive the likelihood and log-likelihood function of the model.

b) Obtain the ML estimates of κ_1 and κ_2.

c) Obtain the ML estimates of the outcome probabilities π_1, π_2, and π_3.

d) How can you obtain standard errors of the estimates in b) and c).

Exercise 6.13 Suppose you have a threshold model as described in Section 6.2.1 in which the error terms follow a type-I extreme value distribution?

a) Derive the probabilities of the J ordered outcomes.

b) Show that this model and the proportional hazards model (see Section 6.4.1) are observationally equivalent if the thresholds are appropriately reparameterized.

Exercise 6.14 Suppose you are interested in the determinants of health status in Switzerland. In particular, you want to know how problems with own children (measured as binary variable, 1=yes) affect a person's health. For your analysis you have access to data from the Swiss Household Panel 2000, which includes a variable indicating self-reported health status coded as: (1) *not in good health*, (2) *good health*, and (3) *very good health*. You estimate two ordered logit models:

Dependent variable: *health status*				
	Model 1		Model 2	
problems with children	-0.352	(0.136)	-0.355	(0.136)
male	0.142	(0.086)	0.099	(0.088)
age	-0.019	(0.003)	-0.018	(0.003)
log(*income*)	0.399	(0.077)	0.339	(0.082)
highest education achieved (reference group *compulsory school*)				
apprenticeship (1=yes)			0.312	(0.115)
university (2=yes)			0.376	(0.153)
$\hat{\kappa}_1$	1.945	(0.930)	1.560	(0.975)
$\hat{\kappa}_2$	4.699	(0.936)	4.322	(0.980)
Log-likelihood value	-2,090.87		-2,086.79	
Observations	2,213		2,213	

Notes: Standard errors in parentheses.

The mean values are given by 0.11 (*problems with children*), 0.42 (*male*), 48.82 (*age*), 11.43 (log(*income*)), 0.63 (*apprenticeship*), 0.18 (*university*).

a) Explain the special feature of the variable *health*.

b) Interpret the coefficients on *problems with children*, *apprenticeship*, and *university* in Model 2 using discrete probability effects.

c) Calculate the marginal probability effects of log(*income*) in Model 2 and use your results to calculate the approximate change in the probabilities, given an one percent increase in income.

d) Test Model 1 against Model 2. What do you conclude?

Exercise 6.15 Human life expectancy has increased considerably over the last century. What is the value of an additional, or for that matter any, year of life? Recent literature in health economics recommends adjusting for health impediments by defining a weight function $d_{it} \in (0, 1)$. The value of d_{it} is equal to one if individual i is in perfect health at time t, and it is zero if the individual has died. Intermediate values apply to individuals with impaired health, with values closer to one or to zero, depending on the severity. The sum $QALY = \sum_{t=0}^{T_i} d_{it}$ gives the "quality-adjusted life years", where T_i is the remaining expected lifetime.

This approach is exemplified in Cutler and Richardson (1997, 1998) who use a large survey providing information on self-reported health, apart from the incidence of a number of diseases. Answer categories include (1) *excellent*, (2) *very good*, (3) *good*, (4) *fair*, and (5) *poor*. In order to obtain $QALY$ weights, they use the following estimation results of an ordered probit model:

Dependent Variable: *health status*			
	Coefficient	Standard Error	$QALY$ weight
arthritis	-0.578	(0.010)	0.79
diabetes	-0.927	(0.018)	0.66
stroke	-0.692	(0.033)	0.74
asthma	-0.708	(0.014)	0.74
bronchitis	-0.370	(0.019)	0.86
hearing impairment	-0.200	(0.010)	0.93
paralysis	-0.873	(0.034)	0.68
age	-0.011	(0.0004)	
$age^2 \times 10^{-4}$	-0.005	(0.0005)	
male	-0.028	(0.011)	
*male*age*	0.010	(0.0006)	
$male*age^2 \times 10^{-2}$	-0.014	(0.0007)	
$\hat{\kappa}_1$	-2.98	(0.010)	
$\hat{\kappa}_2$	-2.13	(0.008)	
$\hat{\kappa}_3$	-1.10	(0.008)	
$\hat{\kappa}_4$	-0.28	(0.008)	
Observations	246,625		

a) Which normalization has been chosen in this estimation?

b) All the cut points are negative. Does this mean that an average person has a high probability of reporting poor health?

c) The $QALY$ weights have been computed using the formula

$$QALY \text{ weight}_l = 1 - \frac{-\beta_l}{\kappa_4 - \kappa_1}$$

What is the implicit assumption on the value of the $QALY$ index for individuals below κ_1 (or above κ_4)? Critically evaluate this approach.

Exercise 6.16 The following questions relate to Kockelman and Kweon (2002), who analyze a large sample of car accidents. The goal of their analysis is to explain the severity of driver injuries, which are classified into four categories: (1) *no injury*, (2) *minor injury*, (3) *severe injury*, and (4) *fatal injury*. Consider the following results from an ordered probit estimation:

Dependent Variable: *severity of injury*		
	Coefficient	p-value
age of driver (in years) $\times 10^{-2}$	0.3171	0.0352
age squared $\times 10^{-4}$	-0.1411	0.3877
male	-0.2665	0.0000
police-reported alcohol involvement (1=yes)	0.2435	0.0000
driver has a past traffic violation on record (1=yes)	-0.0877	0.0000
vehicle age (1999 – model year) $\times 10^{-2}$	0.7045	0.0000
driver driving a pick-up truck (1=yes)	-0.1915	0.0000
driver driving a minivan (1=yes)	-0.1549	0.0000
driver driving a SUV (1=yes)	-0.2171	0.0000
driver driving a MDT (1=yes)	-0.7743	0.0000
driver driving any other type of vehicle (1=yes) (but not a passenger car, e.g., motorcycles, full-sized vans, or buses)	0.1103	0.0000
collision partner is a pickup (1=yes)	0.1458	0.0000
collision partner is a minivan (1=yes)	0.0598	0.0893
collision partner is a SUV (1=yes)	0.0935	0.0000
collision partner is a MDT (1=yes)	0.5444	0.0000
collision partner any other type of vehicle (1=Y) (but not a passenger car, e.g., motorcycles, full-sized vans, or buses)	0.1439	0.0000
head-on crash type (1=yes)	0.5870	0.0000
rear-end crash type (1=yes)	-0.1935	0.0000
rollover (1=yes)	1.0908	0.0000
left-side impact (1=yes)	-0.0451	0.0032
right-side impact (1=yes)	-0.1933	0.0000
$\hat{\kappa}_1$	0.0000	-
$\hat{\kappa}_2$	1.2290	0.0000
$\hat{\kappa}_3$	2.3272	0.0000
Observations	65,510	

Notes: SUV = sport utility vehicle; MDT = medium-duty truck.

a) Write down the formal model that the estimation is based upon.

b) Can you describe the type of accident that leads to the highest fatality rate?

c) Compute the predicted probability of a fatal injury for your chosen scenario in question b).

d) Is it, as many people claim, dangerous to drive a sports utility vehicle? Do you have all the information required to answer such a question? Why or why not?

7

Limited Dependent Variables

7.1 Introduction

In this chapter, we present methods for mixed discrete and continuous – or "limited" – dependent variables. The coexistence of discrete and continuous aspects of a dependent variable is encountered in a wide range of different – and at first glance unrelated – applications, see the following example, 7.1. Upon closer inspection, however, it turns out that the formal structure of these models is very similar, so that they can be usefully presented in a single chapter, using a unified notation and estimation methodology.

Example 7.1. Applications

- The total hours of work in an economy can be computed as the number of workers times the average number of hours worked per worker. How will a wage increase affect total labor supply?
- If you try to estimate the average income of the Swiss population using data on individual tax returns, do you think you are going to underestimate or overestimate true income?
- Are wages of workers representative of potential wages of all persons?
- The observed health status may be lower among people with mandatory health insurance than among those with additional private insurance. Can you conclude that private insurance makes people sick?

In the first example, we have to realize that a sizeable fraction of the working-age population does not participate in the labor market. These people have a labor supply of zero hours. In microeconomic terminology, their utility-maximizing choice between leisure and consumption is at the corner

of the budget set – hence we call models for this type of situation **corner solution models**. The second and third examples allude to the possibility of a discrepancy between what we see in the data and what holds in the underlying population. This is an instance of partial observability, or **sample selection**. Finally, in the fourth example we face the question of whether the assignment to health insurance is random. In that case, the effect of insurance on health can be interpreted as a causal "treatment" effect, and the follow-up question of what should be done – how the causal effect can be estimated – if the assignment is non-random. This estimation problem is addressed in a subsection on **treatment effect models**. To summarize, in this chapter we distinguish between three main types of limited dependent variable models.

- Corner solution models
- Sample selection models
  - Censoring
  - Truncation
  - Incidental censoring
- Treatment effect models
  - Endogenous dummy variable
  - Switching regression

7.1.1 Corner Solution Outcomes

In studies of consumer demand, one typically observes that many households abstain from buying a certain product, or group of products, in any given period of time. Other households buy various positive quantities. Hence, expenditures are zero for a non-trivial proportion of the population, and positive and continuous for the rest. Similarly, if we are interested in labor supply behavior, i.e., the number of hours of work per week, we will find that many people work zero hours per week, while others work one, two, or more hours.

In either case, the underlying data generating process can be conceptualized as arising from utility maximization under constraints, where both interior (i.e., continuous) solutions and corner (i.e., discrete) solutions are possible, depending on the person's preferences and constraints. Following Wooldridge (2002), we therefore refer to such dependent variables as **corner solution outcomes**. One objective of this chapter is to formulate and describe econometric models that jointly describe a binary outcome – whether we have a corner solution or an interior solution – *and* the continuous value of the interior solution if it applies.

7.1.2 Sample Selection Models

The common feature of **sample selection models** is that a standard continuous regression model is assumed for the population, and the purpose of the analysis is to draw inferences about the parameters of this underlying population model. However, the sampling or data collection process is such that we cannot observe the full range of the population model. Either the dependent variable is only **partially observed**, or certain observations are **excluded** from the sample altogether. The nature of the problem depends on the type of **selection rule**, whether censoring, truncation, or incidental censoring.

With **censoring**, the precise value of the underlying continuous variable is unobserved if it falls below or above a certain censoring (or threshold) value, i.e., we know that the continuous dependent variable falls within the censoring range, but not the exact value. The censoring value is known. In contrast to interval data discussed in the previous chapter, the dependent variable is fully observable in the non-censored range. An example of censoring is information on personal wealth that has been top-coded for confidentiality reasons. Top-coding means that the information is given in dollar amounts up to a maximum value, and as a class indicator for all wealth amounts beyond that maximum value. Another example of top-coding was given in Table 1.1, where the largest category for the number of children ever borne to a woman was listed as "8 or more".

With **truncation**, the whole observation is excluded from the sample if the underlying continuous variable falls below or above a certain truncation point. For example, most countries have a minimum amount of earnings below which no tax-return needs to be filed. Therefore, the distribution of earnings obtained from tax records is truncated below this point. Without further assumptions, we cannot use the tax distribution to estimate parameters of the earnings distribution in the whole population. Generalizations of censoring and truncation are obtained if the selection rule does not depend on the outcome of the underlying continuous variable, wealth or income in the above examples, but rather on the realization of a secondary process. This aspect will be discussed in the section on **incidental selection** models.

The distinction between corner solution and sample selection models is first and foremost a matter of interpretation. In the former type of model, the discrete outcome is an economic choice with a meaningful interpretation in the population. In the latter, the discrete outcome reflects a data deficiency. It tells us that some value in the censored range has been realized, without any way of knowing which one. As a consequence, the quantities of interest in the two models are quite different, as we will see in detail below.

Nevertheless, the previous literature has not always been sufficiently clear on this point. Potential confusion can arise because the log-likelihood function is partly continuous and partly discrete in both applications, and it can even be identical in a certain corner solution model and a certain censored regression model – the so-called **Tobit model**. To avoid this confusion, we use the term

"Tobit" only when applied to corner solution applications, as in the original paper by Tobin (1958).

7.1.3 Treatment Effect Models

There has recently been much interest in the econometrics community in developing and applying methods of program evaluation (see Heckman and Vytlacil, 2005, for a recent exposition). For example, a training program is offered to a group of unemployed persons, and one may be interested in finding out whether the persons are better off (for example, in terms of re-employment probability or wages) with the training than they would have been without it. Since this type of estimation problem has much in common with analyzing the effectiveness of a medical treatment, say, and since some of the methods were adopted from an earlier literature in biometrics, the term "treatment effect" has become quite common in econometrics as well, notwithstanding the fact that related models have been used much earlier under different names. An exemplary "treatment" considered in economic applications is, then, whether a person participated in such a training program ($d = 1$), or whether he/she did not participate ($d = 0$).

Interestingly from our point of view, treatment effect models thus defined share three main features with the selection models introduced in the previous section. First, a constituent feature of a "treatment", as defined in this literature, is that it is binary, whereas the outcome variable is often continuous. Therefore, treatment models and censored models both combine discrete and continuous elements. Second, there is a problem of partial observability: we cannot observe the hypothetical (or **counterfactual**) outcome for treated people without the treatment. Similarly, we cannot observe the hypothetical outcome for the control group with the treatment. And third, as censoring may be non-random and related to the outcome of interest, it may be that the assignment to treatment and control is not random but rather related to the outcome as well. It should not come as too great a surprise, then, that simple OLS often fails in treatment models, and that formal methods of estimating the causal effect of a treatment are closely related to the sample selection methods. This is further discussed in Section 7.4.

7.2 Tobin's Corner Solution Model

7.2.1 Introduction

Consider the amount of an individual's expenditure for health services in the previous month, or the hours of work spent in the labor market. Such data have two main characteristics:

1. they are non-negative
2. they have a cluster of observations at "zero".

Therefore, a coherent econometric model should be able to account for these two features. It should attribute a positive probability mass to the discrete outcome "zero", and it should exclude negative predictions. A simple linear regression model is inappropriate in this context for the following reasons: first, it ignores that the dependent variable, consumer expenditures, cannot be negative. Ignoring this restriction will be inefficient. Moreover, predictions outside of the admissible range of values are possible. Second, the linear regression model imposes constant marginal effects, which is unrealistic in a model where the dependent variable is bounded from below. Also, it is not possible to impose the non-negativity and non-constant marginal effects by log-transforming the dependent variable, since the logarithm of zero is not defined. Moreover, any purely continuous data model does not account for the cluster at zero.

The main limitation of the linear regression model therefore is that it cannot be used to analyze the quantities of interest in such corner solution models. While the regression is a model for the expectation $\mathrm{E}(y|x)$ (conditional on x but unconditional on y being positive), in the corner solution model, we also want to learn something about

- $P(y = 0|x) = 1 - P(y > 0|x)$, the probability of having zero expenditures, and
- $\mathrm{E}(y|y > 0, x)$, the expected expenditures conditional on having positive expenditures.

For notational simplicity, we will drop the individual subscript i for the rest of this chapter, if it is not necessary. By the law of iterated expectations,

$$\begin{aligned}\mathrm{E}(y|x) &= P(y = 0|x) \times 0 + P(y > 0|x) \times \mathrm{E}(y|y > 0, x) \\ &= P(y > 0|x) \times \mathrm{E}(y|y > 0, x) \end{aligned} \tag{7.1}$$

Therefore, a coherent model for such data should combine both a binary part, ensuring that $0 \leq P(y > 0|x) \leq 1$, and a positive regression part, ensuring that $\mathrm{E}(y|y > 0, x) > 0$. For example, one could model the binary part as a probit model and the positive part as a log-linear model. Indeed, this suggestion was made by advocates of so-called "two-part models" (see, for example, Duan et al., 1983).

7.2.2 Tobit Model

The classical solution put forward by Tobin (1958) was a different one. He proposed a model where both aspects, $P(y > 0|x)$ and $\mathrm{E}(y|y > 0, x)$, are manifestations of a common underlying latent variable model for y^*. Let

$$y^* = x'\beta + u \qquad u|x = Normal(0, \sigma^2) \tag{7.2}$$

In the Tobit model, this latent model has no interesting interpretation *per se*; it is a purely artificial device. The object of interest is the discrete-continuous y, which is defined as follows:

$$y = \max(0, y^*) \tag{7.3}$$

which means that $y = y^*$ if $y^* > 0$ and $y = 0$ if $y^* \leq 0$. The two equations (7.2) and (7.3) fully characterize the Tobit model, and a number of properties can be derived directly. The probability of a zero is given by

$$P(y = 0|x) = P(y^* \leq 0|x) = P\left(\frac{u}{\sigma} \leq -\frac{x'\beta}{\sigma} \,\Big|\, x\right) = 1 - \Phi\left(\frac{x'\beta}{\sigma}\right) \tag{7.4}$$

and for positive observations, the density is

$$f(y|x; \beta, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{y - x'\beta}{\sigma}\right)^2\right] = \frac{1}{\sigma}\phi\left(\frac{y - x'\beta}{\sigma}\right) \tag{7.5}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ denote the cumulative density function and density function of the standard normal distribution. Given an independent sample of observations, the two parameters σ and β can be estimated by maximum likelihood, and the likelihood function is given by

$$\begin{aligned} L(\beta, \sigma; y, x) &= \prod_{i=1}^{n} f(y_i = 0)^{I(y_i=0)} f(y_i, y_i > 0)^{I(y_i>0)} \\ &= \prod_{i=1}^{n} \left[1 - \Phi(x_i'\beta/\sigma)\right]^{I(y_i=0)} \left[\frac{1}{\sigma}\phi\left(\frac{y_i - x_i'\beta}{\sigma}\right)\right]^{I(y_i>0)} \end{aligned} \tag{7.6}$$

where $I(\cdot)$ is an indicator function that returns the value one if the statement in parentheses is true and zero otherwise. It can be shown that the Tobit log-likelihood function is globally concave (Olsen, 1978). The MLE is asymptotically normally distributed with the variance matrix given, for example, in Maddala (1983, p. 155) and Amemiya (1985, p. 373).

Example 7.2. Female Hours of Work

The following table shows the OLS and Tobit estimates for a dataset of 753 women, 428 of whom had positive hours of work in 1975, while the remaining 325 women did not work for pay in 1975. The data were extracted from the Michigan Panel Study of Income Dynamics (PSID). The subsample employed here was used by Mroz (1987), where further details can be found.

The dependent variable is the annual hours of work (standard errors in parentheses). Here, we report estimates from a **reduced-form** specification, where the explanatory variables include human capital proxies and a number of variables capturing the preferences towards leisure, and thereby the marginal ratio of substitution between leisure and consumption. A variable that is conspicuously missing is the wage rate. Including the wage would lead to a **structural labor supply equation**. Estimation of such a structural equation introduces additional econometric complexities – we do not observe the wages for women who do not work – that are not accounted for by the simple Tobit model. These will be discussed later in Section 7.3.6.

Table 7.1. *Tobit and OLS Estimates of Female Hours of Work*

Dependent Variable: *hours of work*			
	Tobit	OLS (all)	OLS (positive)
nonwife income	-8.81	-3.45	0.44
	(4.46)	(2.54)	(3.61)
years of schooling	80.65	28.76	-22.79
	(21.58)	(12.95)	(16.43)
years of experience	131.56	65.67	47.01
	(17.28)	(9.96)	(14.56)
$(\textit{years of experience})^2$	-1.86	-0.70	-0.51
	(0.54)	(0.32)	(0.44)
age	-54.41	-30.51	-19.66
	(7.42)	(4.36)	(5.89)
# kids less than 6 years	-894.02	-442.09	-305.72
	(111.88)	(58.85)	(96.45)
# kids between 6 and 18	-16.22	-32.78	-72.37
	(38.64)	(23.18)	(30.36)
constant	965.31	1,330.48	2,056.64
	(446.44)	(270.78)	(346.48)
$\hat{\sigma}$	1,122.0	750.2	725.7
R-squared		0.27	0.14
Log-likelihood value	-3,819		
Observations	753	753	428

Notes: Standard errors in parentheses.

Three models were estimated. The first column gives the Tobit estimates, the second the OLS estimates using all observations (the dependent variable "hours" is either zero or positive) while the third column gives the OLS estimates on the sample of 428 positive observations. Clearly, the point estimates differ widely among the three models. In order to understand why this is so, and how the parameters should be interpreted, we require additional results on truncated normal distributions and conditional expectations in the Tobit model which will be discussed next.

7.2.3 Truncated Normal Distribution

In order to provide a full analysis of the properties of the Tobit model, we need to make use of some elementary statistical results related to the truncated normal distribution. These results will be needed later on, in the discussion of diverse sample selection models and treatment effect models as well.

Consider a random variable y with density function $f(y)$. Let y be truncated from below at a. The truncated density function $f(y|y > a)$ is obtained by rescaling the original density function $f(y)$

$$f(y|y > a) = \frac{f(y)}{P(y > a)} = \frac{f(y)}{1 - F(a)} \tag{7.7}$$

where $F(a)$ is the cumulative density function of y at point a, so that the density integrates to one over the range above a. Therefore, if y is normally distributed with expected value μ and variance σ^2, the truncated density function can be written as

$$f(y|y > a) = \frac{1}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right) \Big/ \left[1 - \Phi\left(\frac{a-\mu}{\sigma}\right)\right] \tag{7.8}$$

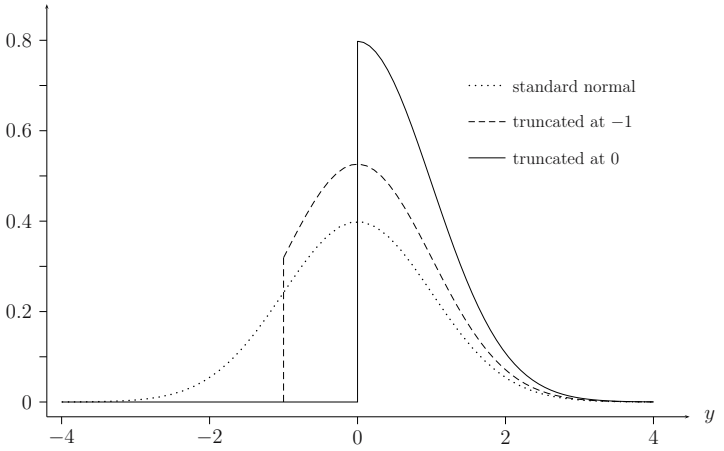
where ϕ and Φ are the density and cumulative density functions of the standard normal distribution, respectively.

Figure (7.1) shows three density functions: the untruncated standard normal distribution and two truncated-from-below standard normal distributions with truncation points $a = -1$ and $a = 0$, respectively.

Expectation of $Normal(0, 1)$ Variable Conditional on Truncation

In the following, we frequently require an expression for the expectation of a truncated normal distribution. First, consider the case of the **standard normal distribution**. Let $u \sim Normal(0, 1)$. Then for any c

Fig. 7.1. *Density Functions of Truncated Normal Distributions*



$$\begin{aligned} \mathrm{E}(u|u > c) &= \int_c^\infty u f(u|u > c) du \\ &= \frac{1}{1 - \Phi(c)} \int_c^\infty u\phi(u) du = \frac{\phi(c)}{1 - \Phi(c)} \end{aligned} \tag{7.9}$$

The last equality follows since

$$\phi'(u) = \frac{d}{du}\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}\right) = -u\phi(u)$$

and therefore

$$\int_c^\infty u\phi(u) du = -\phi(u)\Big|_c^\infty = -\phi(\infty) - (-\phi(c)) = \phi(c)$$

The result that the conditional expectation can be written as a ratio of density $\phi(c)$ and complementary cumulative density function $1 - \Phi(c)$ depends crucially on the distributional assumption. For distributions other than the normal distribution, the conditional expectation does not, in general, have such a simple expression.

7.2.4 Inverse Mills Ratio and its Properties

The quantity $\lambda(\delta) = \phi(\delta)/\Phi(\delta)$ for $\delta \in (-\infty, \infty)$ is also known as the **inverse Mills ratio**. It is defined as the ratio of density and cumulative density of the standard normal distribution. In the present case, since $\phi(c)/[1 - \Phi(c)] = \phi(-c)/\Phi(-c)$, we can rewrite the central result (7.9) as

$$\mathrm{E}(u|u > c) = \lambda(-c) \tag{7.10}$$

The inverse Mills ratio $\lambda(\delta)$ plays a very important role in the further development, and therefore we will list some of its main properties here. First, it is bounded from below at 0. This follows directly, since both the numerator $\phi(\delta)$ and the denominator $\Phi(\delta)$ are non-negative quantities. Second, the inverse Mills ratio is also bounded from below at $-\delta$. This bound follows, since $\lambda(\delta) = \mathrm{E}(u|u > -\delta)$, where u has a standard normal distribution, and it must be the case that an expectation is always greater or equal than its smallest value, in this case $-\delta$. Combining these two results for all possible values of δ (positive or negative), we obtain a lower bound for the inverse Mills ratio:

$$\lambda(\delta) \geq \max(0, -\delta) \tag{7.11}$$

This bound ensures that $\lambda(\delta) + \delta > 0$ for all values of δ. Moreover, we would like to know how $\lambda(\delta)$ changes as δ increases. Direct differentiation of $\lambda(\delta) = \phi(\delta)/\Phi(\delta)$ gives

$$\lambda'(\delta) = -\lambda(\delta)[\lambda(\delta) + \delta] \tag{7.12}$$

This expression is always negative, which proves the result that the inverse Mills ratio is monotonically decreasing in δ. In principle, the ratio of density and distribution function can be computed for any continuous univariate distribution $F(\delta)$. Since the inverse Mills ratio is the first derivative of the logarithmic distribution function, it is decreasing in δ as long as $\log[F(\delta)]$ is concave. This property is referred to as **log-concavity** (Heckman and Honoré, 1990). A sufficient condition for log-concavity of a cumulative density function is log-concavity of the density itself. It is easy to verify that this sufficient condition is fulfilled in the case of a standard normal distribution (the second derivative of $\ln \phi(z)$ is -1), which is another way of showing that the inverse Mills ratio $\lambda(\delta)$ is decreasing in δ.

In the present case, we have $\delta = -c$ so that we know that $\mathrm{E}(u|u > c)$ is monotonically *increasing* in c.

Exercise 7.1.

- Calculate the expected value of a standard normal density that is truncated below zero.
- Derive equation (7.12) explicitly.

Expectation of $Normal(\mu, \sigma^2)$ Variable Conditional on Truncation

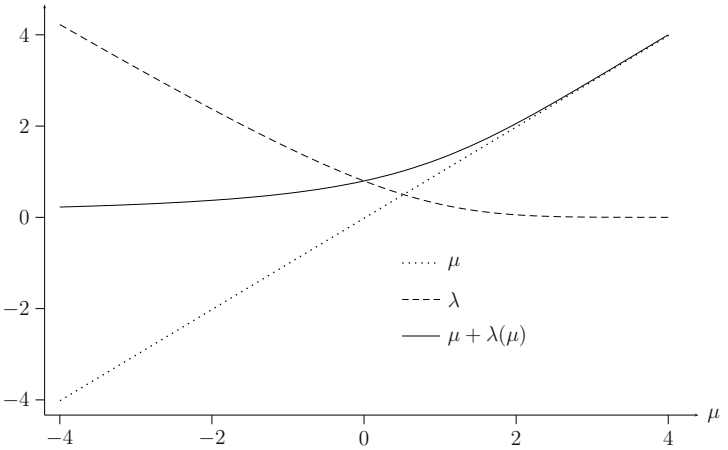
The above results generalize directly to $Normal(\mu, \sigma^2)$ distributions. Since for $u \sim Normal(0, 1)$, we know that $y = \mu + \sigma u$ has a $Normal(\mu, \sigma^2)$ distribution, we obtain the following conditional expectation for truncation below a:

$$\begin{aligned} \mathrm{E}(y|y > a) &= \mathrm{E}(\mu + \sigma u|\mu + \sigma u > a) \\ &= \mu + \sigma \mathrm{E}\left(u \left| u > \frac{a-\mu}{\sigma}\right.\right) \\ &= \mu + \sigma \frac{\phi(\alpha)}{1 - \Phi(\alpha)} \end{aligned} \tag{7.13}$$

where $\alpha = (a - \mu)/\sigma$. Often, we will require results for truncation-from-below **at zero**. In this case, the expected value of the truncated variable can be written as

$$\begin{aligned} \mathrm{E}(y|y > 0) &= \mu + \sigma \frac{\phi(-\mu/\sigma)}{1 - \Phi(-\mu/\sigma)} \\ &= \mu + \sigma \frac{\phi(\mu/\sigma)}{\Phi(\mu/\sigma)} \\ &= \mu + \sigma \lambda(\mu/\sigma) \end{aligned} \tag{7.14}$$

Fig. 7.2. *Expected Value and Inverse Mills Ratio for Truncated Normal Distribution*



The relationship between truncated and untruncated expectations for truncation from below at zero, as a function of μ, is depicted in Figure 7.2. Here, the point of departure is a normal distribution with unit variance $N(\mu, 1)$. Therefore, $\mathrm{E}(y|y > 0) = \mu + \lambda(\mu)$. We see, as it should be, that the conditional expectation never falls below zero, even if μ is negative and large in absolute value. Clearly, the more the distribution is truncated, i.e., the smaller the mean μ relative to the given truncation point (here 0), the larger the discrepancy between the truncated and untruncated expectations.

So far, all results are valid for truncation from below (from the left, a lower bound). Similar results exist for truncation from above (from the right, imposing an upper bound). A derivation similar to the one in (7.9) shows that

$$\mathrm{E}(y|y < a) = \mu - \sigma\frac{\phi(\alpha)}{\Phi(\alpha)} \tag{7.15}$$

Exercise 7.2.
Let $u \sim Normal(0, 1)$.

- Show that

$$\mathrm{E}(u|u \leq c) = -\frac{\phi(c)}{\Phi(c)}$$

  where $\phi(\cdot)$ and $\Phi(\cdot)$ are density and cumulative density of the standard normal distribution, respectively.
- Verify that $\mathrm{E}(u) = \mathrm{E}(u|u \leq c)P(u \leq c) + \mathrm{E}(u|u > c)P(u > c) = 0$.

Finally, truncation obviously affects the higher-order moments of the distribution as well. For example, if $y \sim Normal(\mu, \sigma^2)$, one can show that

$$\mathrm{Var}(y|y > a) = \sigma^2(1 - \lambda(\alpha)(\lambda(\alpha) - \alpha)) \tag{7.16}$$

where $\alpha = (a-\mu)/\sigma$ and $\lambda(\alpha) = \phi(\alpha)/\Phi(\alpha)$ as before. Therefore, a truncated normal distribution is **heteroscedastic** – it has a variance that depends on μ – even if the underlying population distribution is homoscedastic.

7.2.5 Interpretation of the Tobit Model

We are now in a position to establish the central properties of the Tobit model, in particular the conditional expectation function (CEF) and the marginal effects. We start with the latent model. $\mathrm{E}(y^*|x)$ is simple – since linear in β – but not interesting, because it lacks a meaningful substantive interpretation. The other aspects of the model are more complicated, since they are non-linear functions of the parameters. In particular, we can distinguish

- $P(y = 0|x)$
- $P(y > 0|x)$
- $\mathrm{E}(y|x)$
- $\mathrm{E}(y|x, y > 0)$

All of these entities can be of substantive interest, depending on the question one wants to address. For example, in the labor supply of women, one may

want to find out how, say, education affects the probability of participating in the labor force. Similarly, one can study how the expected hours among workers, or the expected hours among all women, vary as the level of education increases. These entities are not unrelated in the Tobit model. First, by the basic laws of probability, $P(y = 0|x) = 1 - P(y > 0|x)$ and $\mathrm{E}(y|x) = P(y > 0|x)\mathrm{E}(y|x, y > 0)$. Second, they all depend on the same basic two parameters β and σ^2. Based on the Tobit specification, we have

$$P(y > 0|x) = \Phi(x'\beta/\sigma) \tag{7.17}$$

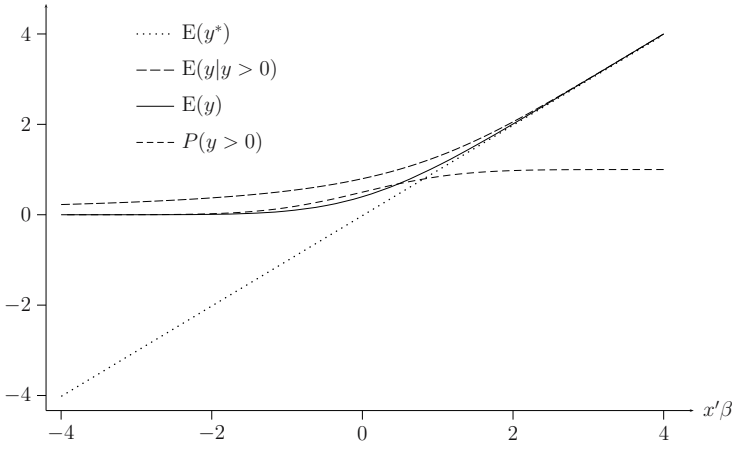
$$\mathrm{E}(y|y > 0, x) = x'\beta + \sigma\lambda(x'\beta/\sigma) \tag{7.18}$$

and

$$\mathrm{E}(y|x) = \Phi(x'\beta/\sigma)[x'\beta + \sigma\lambda(x'\beta/\sigma)] \tag{7.19}$$

The three functions (7.17) - (7.19) are plotted in Figure 7.3 against $\mu = x'\beta$. We see that $\mathrm{E}(y|y > 0, x) > \mathrm{E}(y|x) > \mathrm{E}(y^*|x)$. We can use the figure to make the following observation about the slopes, and thus implicitly about the marginal effects. For a given increase in μ, for instance $\Delta\mu = \beta_l \Delta x_l$, where $\beta_l > 0$ and $\Delta x_l > 0$, we see that $\Delta\mathrm{E}(y|y > 0, x) < \Delta\mathrm{E}(y|x) < \Delta\mathrm{E}(y^*|x) = \Delta\mu$. The inequalities need to be reversed if μ is decreased. A more formal look at marginal effects comes next.

Fig. 7.3. *Conditional Expectation Functions in the Tobit Model ($\sigma^2 = 1$)*



Marginal Effects

The marginal effect for the probability of a corner solution is simply

$$\frac{\partial P(y=0|x)}{\partial x_l} = -\phi(x'\beta/\sigma)\beta_l \tag{7.20}$$

To obtain the first derivative of the **conditional expectation** of the Tobit model, we make use of the result on the first derivative of the inverse Mills ratio (7.12). Therefore

$$\frac{\partial \mathrm{E}(y|y>0,x)}{\partial x_l} = \beta_l\{1-\lambda(x'\beta/\sigma)[x'\beta/\sigma+\lambda(x'\beta/\sigma)]\} \tag{7.21}$$

Since the unconditional expectation in the Tobit model is given by $\mathrm{E}(y|x) = P(y>0|x)\mathrm{E}(y|y>0,x)$, the overall marginal effect $\partial \mathrm{E}(y|x)/\partial x_l$ can be written in general as

$$\frac{\partial \mathrm{E}(y|x)}{\partial x_l} = \frac{\partial P(y>0|x)}{\partial x_l}\mathrm{E}(y|y>0,x) + P(y>0|x)\frac{\partial \mathrm{E}(y|y>0,x)}{\partial x_l} \tag{7.22}$$

This derivative has an interesting economic interpretation. The overall effect is the sum of an effect at the

- **extensive margin** (e.g., how much more likely a person is to join the labor force as education increases times the expected hours)

  and an effect at the

- **intensive margin** (e.g., how much the expected hours of work increase for workers as education increases times the probability of participation)

Under the assumptions of the Tobit model, the formula for the marginal effects of the **unconditional expectation** is remarkably simple. It can be verified by plugging in the appropriate expressions in (7.22) that

$$\frac{\partial \mathrm{E}(y|x)}{\partial x_l} = \beta_l \Phi(x'\beta/\sigma) \tag{7.23}$$

Hence, the overall marginal effects are just a scaled version of β_l. Moreover, the relative marginal effects

$$\frac{\partial \mathrm{E}(y|x)/\partial x_l}{\partial \mathrm{E}(y|x)/\partial x_m} = \frac{\beta_l}{\beta_m} \tag{7.24}$$

are constant. We also see that the larger $x'\beta$, i.e., the larger $\Phi(x'\beta/\sigma)$ and the smaller the probability of a zero, the more similar the models for y and y^*, and thus the conditional expectations and the marginal probability effects.

As fewer and fewer individuals are observed at the corner outcome, the Tobit model converges to the normal linear model.

Exercise 7.3.

- Verify the validity of equation (7.23).
- What is the marginal effect of a woman's education on the expected number of (total) hours in the Mroz example?

7.2.6 Comparing Tobit and OLS

In this section, we show that application of the linear regression model to corner solution problems leads to a biased estimator of β. The reason is that OLS estimates a linear CEF – indeed, OLS is the **best linear predictor** – whereas the true model has a non-linear CEF. However, one should keep in mind that β alone is not of great interest in the corner solution model. At the end of the day, we are interested in marginal effects. While OLS fails to capture the non-linearity of marginal effects, the full-sample OLS estimates may actually provide a good approximation of the true marginal effects at the average values of the explanatory variables.

In principle, there are two possibilities to estimate a linear regression model for corner solution data. First, one may discard part of the sample and include only non-limit observations. Alternatively, one may use the full sample including the observations for which $y_i = 0$. Both approaches lead to a bias towards zero. As we have seen previously,

$$\left| \frac{\partial \mathrm{E}(y|y > 0, x)}{\partial x_l} \right| < |\beta_l| \tag{7.25}$$

and

$$\left| \frac{\partial \mathrm{E}(y|x)}{\partial x_l} \right| < |\beta_l| \tag{7.26}$$

The bias arises since OLS estimates a linear approximation of the left-hand side marginal mean effects. An alternative way of looking at this problem is to interpret the bias as an **omitted variable** problem. For example, remember that the CEF for the positive part of the model was given by

$$\mathrm{E}(y|x, y > 0) = x'\beta + \sigma\lambda(x'\beta/\sigma) \tag{7.27}$$

The error term in a regression of y on x without including the second term is

$$v = \sigma\lambda(x'\beta/\sigma) + u$$

where u is a "well-behaved" error with mean independence. However, correlation between the regressor x and v exists, since x is in general correlated with the inverse Mills ratio. The direction of the bias depends on the sign of the correlation between the regressor x and λ. As we recall from Figure 7.2, λ decreases in its argument. For positive β_l, there is a negative correlation between x and v and the bias is downward. For negative β_l, there is a positive correlation and the bias is upward. Hence, in general we have a bias towards zero, or **attenuation bias**. This result holds unambiguously for a single regressor or for a set of orthogonal regressors. In other cases, cross-correlations may complicate the analysis.

To get an idea about the magnitude of the bias, recall that the Tobit marginal effects are given by $\partial \mathrm{E}(y|x)/\partial x_l = \beta_l \Phi(x'\beta/\sigma)$. The full-sample OLS coefficients, say $\hat{\gamma}_l$, are direct estimates of the marginal effects. Since the Tobit coefficients $\hat{\beta}_l$ are multiplied by a factor with a value between zero and one, it will tend to be the case that $|\hat{\beta}_l| > |\hat{\gamma}_l|$. Moreover, the Tobit coefficients will exceed the OLS estimates roughly by a factor of $\Phi(\bar{x}'\hat{\beta}/\hat{\sigma})$, where $\bar{x}$ are the mean values of the explanatory variables. This approximation is less appropriate for discrete explanatory variables, where the non-linearity may lead to larger discrepancies.

Example 7.3. Female Hours of Work, Continued

Recall the labor supply example using data on 753 married women, 325 of whom did not work for pay in 1975. The Tobit and OLS results were shown in Table 7.1. Not surprisingly, we find that the Tobit estimates are much larger (in absolute value) than the OLS estimates. This is an illustration of the bias in the underlying β estimates of OLS in the corner solution case. However, it must be emphasized again that the latent model is not of any interest *per se* in this context. Most importantly, one should not think that the greater coefficient estimates in the Tobit model must imply a greater response of the dependent variable to changes in the independent variable in the Tobit model. Recall the formula for the marginal effects in the Tobit model. To compare the first and second columns, we need to consider the unconditional marginal effect $\partial \mathrm{E}(y_i|x_i)/\partial x_{il}$. In the OLS model, this is simply the estimated coefficient. In the Tobit model, we have $\partial \mathrm{E}(y_i|x_i)/\partial x_{il} = \beta_l \Phi(x_i'\beta/\sigma)$. The average adjustment factor is approximately equal to the probability of a positive observation. This probability varies from person to person, with an average of 0.65. Hence, the marginal effect of *nonwifeinc* for a woman with average participation probability is $(0.65) \cdot (-8.81) = -5.7$ in the Tobit model, compared to -3.5 in the OLS model.

7.2.7 Further Specification Issues

As always in maximum likelihood estimation, we require in general that the model be correctly specified. In the Tobit model, this means in particular that the errors in the latent model are normally distributed and that they have a constant variance. Violations of these assumptions would mean that the likelihood function is misspecified, that the formulas for the unconditional expectation $\mathrm{E}(y|x)$ and the conditional expectation $\mathrm{E}(y|y > 0, x)$ are wrong, and that the estimator is no longer consistent. In this sense, OLS is much more robust since the desirable property of unbiasedness does not depend on the error distribution – provided mean independence holds.

One important restriction of the Tobit model is that the same process drives the probability of a corner solution and the conditional expectation of positive outcomes. For example, consider the labor supply application and the question of how age affects hours of work. In the standard Tobit model, it is precluded a priori, by functional form assumption, that a factor such as age could, for example, reduce the probability of participation (because of early retirement), but increase hours of work conditional on participation (i.e., older workers might have a lower part-time rate than younger ones).

One way to evaluate the appropriateness of the Tobit model is to estimate a simple probit model, where the binary outcome is either $y = 0$ or $y > 0$. In this model, we know that we estimate the scaled coefficients $\gamma = \beta/\sigma$. We can then compare $\hat{\gamma}$ from the probit with the ratio of the Tobit estimates $\hat{\beta}/\hat{\sigma}$. The two will never be identical because of sampling error. But they should be in the same order of magnitude if the Tobit model, as represented by equations (7.2) and (7.3), is correct. It would be a worrying sign if, for instance, one of the coefficients switched its sign.

In this case, one might instead consider the **two-part model** that was first proposed by Cragg (1971). In particular, he suggested to separately estimate a binary model for $P(y = 0|x)$ and a truncated-at-zero model for $\mathrm{E}(y|y > 0, x)$. An interesting aspect of this approach is that the standard Tobit model is nested, and the restriction implied by it can therefore be tested by a simple likelihood ratio test. Other two-part models that do not nest the standard Tobit model are discussed by Duan et al. (1983).

7.3 Sample Selection Models

7.3.1 Introduction

Sample selection models are models of **partial observability**. It is useful to distinguish between two cases. Under **censoring**, the precise value of the dependent variable is unobserved if it falls below or above a certain threshold value. The partial observability may be due to things like coding, sampling process or other factors. For the censored observations, we know that the continuous dependent variable falls in the censoring range but not the precise value. In the non-censored range the dependent variable is fully observable. Clearly, this is not a corner solution problem. Examples of censored dependent variables include "top-coding" of income, age at first birth, which is censored (unobserved) for childless women before the end of their reproductive period, the duration of marriage or unemployment, or wage offers that are only observed for workers.

With **truncation**, the whole observation is excluded from the sample if the underlying variable falls below or above a certain truncation point. Both censoring and truncation are always defined in relation to the **dependent** variable, either in a univariate setting or in a regression context. The following example illustrates the two concepts, using artificial data.

Example 7.4. Artificial Data with Censoring and Truncation

In Table 7.2, the first column shows ten pseudo-random numbers that have been independently drawn from a $Normal(0,1)$-distribution. They are arranged in increasing order.

Table 7.2. *Artificial Data with Right-Censoring and Truncation at One*

Observation	Full Sample	Censored	Truncated
1	-1.70	-1.70	-1.70
2	-0.80	-0.80	-0.80
3	-0.65	-0.65	-0.65
4	-0.62	-0.62	-0.62
5	-0.19	-0.19	-0.19
6	0.05	0.05	0.05
7	0.16	0.16	0.16
8	0.81	0.81	0.81
9	1.27	.	
10	1.62	.	
$\bar{x}$	-0.005	-0.367	-0.367

Suppose, instead, that observations are censored from above at one. This means that the exact values for the two observations 9 and 10 are missing.

It is only known that their value exceeds one. This is shown in the second column. Despite the censoring, the sample still consists of ten observations. This changes once we consider a sample that is truncated from above at one. In this case, the sample only includes the eight non-truncated observations, as seen in the third column of Table 7.2. Naturally, we find that the means of the censored (using non-censored observations only) and truncated samples, respectively, are smaller than the mean of the full sample.

The central insight of this chapter is that we can draw inferences on the expectation μ in the full sample, *even if* we can only see the censored or truncated data. The crucial **identifying assumption** is that the distribution of the population is known up to the parameters. In this chapter, we exclusively deal with the normal distribution, although the principles and methods directly apply to any other distribution function as well.

Example 7.5. Estimation of Population Parameters with Censoring

From the second column of Table 7.2, we know that 20 percent of the observations are non-censored, and that the censored mean is -0.367. These two pieces of information can be used to estimate $\hat{\mu}$ as follows. Because of normality, we know that

$$P(\text{censored}) = P(y > 1) = 1 - \Phi\left(\frac{1-\mu}{\sigma}\right)$$

Hence, we can use the analogy principle to obtain estimates by considering the corresponding sample relationship.

$$\frac{1-\hat{\mu}}{\hat{\sigma}} = \Phi^{-1}(0.8) = 0.842$$

and therefore $\hat{\mu} = 1 - 0.842\,\hat{\sigma}$. Moreover $\mathrm{E}(y|y<1) = \mu - \sigma\lambda((1-\mu)/\sigma)$, and thus

$$\hat{\mu} = -0.367 + \hat{\sigma}\frac{\phi(0.842)}{0.8} = -0.367 + 0.350\,\hat{\sigma}$$

We obtain two linear equations in two unknowns with solution $\hat{\mu} = 0.034$ and $\hat{\sigma} = 1.147$. These are reasonably close to the true values. Alternative identification strategies could be based on higher-order moment restrictions (e.g., equation (7.16)).

7.3.2 Censored Regression Model

Suppose that the population of interest follows a normal linear regression model such that

$$y = x'\beta + u \qquad u|x \sim Normal(0, \sigma^2) \tag{7.28}$$

where u and x are independent. The parameters of interest are the regression parameters β. If y were observed, we could obtain the best linear and unbiased estimator by simple OLS estimation. However, y is not fully observed. We observe only w such that

$$w = \max(y, c) \tag{7.29}$$

where c is the individual censoring point. Thus, data are censored **from below** at c. If $c = 0$, for example, we have a model with censoring from below at zero. In other words, negative values of y are not observed, but records with negative y are kept in the sample. This situation is formally the same as Tobin's corner solution model in Section 7.2. However, the interpretation is different (see Section 7.1.2), and in order to minimize confusion between these two models, we use different notation: in the Tobit model, the linear model is for y^* while we observe y. In the censored model, the linear model is for y while we observe w. In both cases, y is the main variable of interest.

In the censored regression model, we therefore have two possibilities:

- Either $w = y$, with probability

$$f(w, w > c|x; \beta, \sigma) = (1/\sigma)\,\phi\left(\frac{w - x'\beta}{\sigma}\right) \tag{7.30}$$

- or $w = c$ in which case we know that $y \leq c$ falls in the censoring area, and the probability of this event is

$$P(y \leq c|x) = \Phi\left(\frac{c - x'\beta}{\sigma}\right). \tag{7.31}$$

With censoring **from above**, the observation rule needs to be rewritten as

$$w = \min(y, c) \tag{7.32}$$

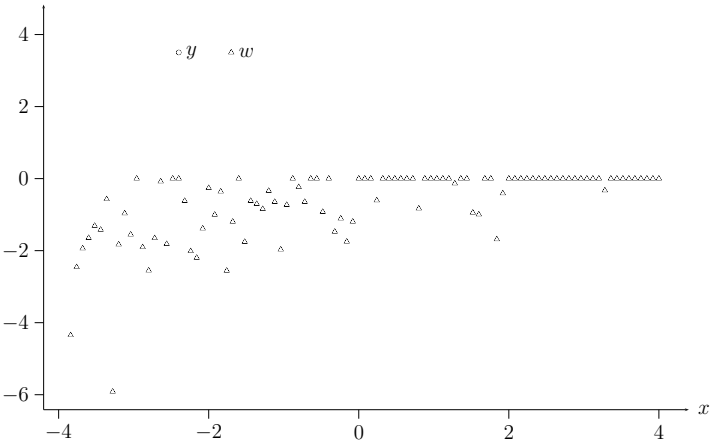
In this case, the censoring probability is

$$P(y \geq c|x) = 1 - \Phi\left(\frac{c - x'\beta}{\sigma}\right). \tag{7.33}$$

Example 7.6. Regression Analysis with Censoring

The following example illustrates the effect of censoring from above at zero. 100 data pairs (y_i, x_i) were generated according to the following rule. First, 100 values for the regressor x_i were chosen as equally spaced on the interval $(-4, +4)$. Second, the dependent variable y_i was generated as $y_i = 0 + 0.5 \times x_i + u_i$ where u_i is a sequence of pseudo-random numbers from a standard normal distribution (with seed 123 in Stata). Finally, the censored variable w_i was obtained using the "min"-rule: $w_i = \min(y_i, 0)$. The circles in Figure 7.4 represent uncensored data points, whereas the triangles show the data after censoring.

Fig. 7.4. *Simulated Data for Regression with and without Censoring*



If we attempt to estimate the slope parameter β by OLS using uncensored observations only, we obtain a slope parameter of $\hat{\beta}_{ols} = 0.242$ (with standard error 0.034). Clearly, this parameter is much smaller than the true parameter of 0.5. Moreover, it has no useful interpretation (such as best linear predictor, or as an approximation to a nonlinear CEF); it simply is a biased estimator, and the direction of the bias with any censoring is towards zero.

7.3.3 Estimation of the Censored Regression Model

The obvious estimation method for the model parameters is ML. For example, with censoring **from below** at c_i the log-likelihood function is given by

$$\log L(\beta, \sigma) = \sum_{y_i \leq c_i} \log \left[\Phi \left(\frac{c_i - x_i'\beta}{\sigma} \right) \right] + \sum_{y_i > c_i} -\frac{1}{2} \left[\log(2\pi) + \log \sigma^2 + \frac{(y_i - x_i'\beta)^2}{\sigma^2} \right] \tag{7.34}$$

For $c_i = 0$, this is exactly the likelihood function of Tobin's corner solution model (see equation (7.6)). The main difference is the information transmitted by the discrete part. In the corner solution model, the discrete part gives the probability of a zero, and this specific value has a well-defined quantitative meaning. In the censored regression model, however, the discrete part describes an interval of possibilities, namely that y_i takes some value below the censoring value c_i. This is clearly different.

Censoring from below at $c_i = 0$ is a special case for which it appears almost impossible to find a good economic application. It is much more common to encounter a varying censoring level, possibly in combination with censoring from above. A typical example of this combination arises in the analysis of durations of spells, such as unemployment or time to first birth. If the terminating event has not occurred at the time when the observation is made, the true (eventual) duration is unknown and the measured duration is censored from above. Another model with varying censoring level, the so-called **incidental censoring** model, is introduced in Section 7.3.5.

Example 7.7. Determinants of Time to First Birth

The duration, or "waiting time", until first birth is an important variable in the analysis of population dynamics. For example, if there is a general trend of women to postpone the first birth (e.g., have a first child on average t 25 rather than at 22), this has a negative effect on current fertility levels (number of births per thousand women, or the total fertility rate). However, this effect may only be temporary if women eventually catch up and have the same number of children as before.

In a cross-section of women of all ages, the duration is subject to censoring. Suppose, that the reproductive period starts at age 14. If a women is 35 and had her first birth at 23, we can say for sure that the duration until first birth was nine years. However, for a woman aged 23 without a first child, we cannot observe the duration until first birth. We only know that the duration is greater than nine years. This observation is censored from above.

In this application, we want to estimate how education affects the time to first birth (TFB). The general hypothesis is that longer education time, and

the associated increased opportunity cost of time away from the labor market, should increase the time to first birth. The data stem from the U.S. General Social Survey for the year 2002. We keep all women aged less than 40 or with known age at first birth, which gives a sample of 1,371 observations. There are two types of women:

Type A: the woman had a first child and thus TFB is known (uncensored; $n = 1{,}154$). The average TFB for uncensored observations is 8.7 years.

Type B: the woman is childless. In this case, either $TFB > age - 14$ or she remains childless (right censored; $n = 217$).

This is an example of a variable with varying but known censoring threshold. We observe $TFB|TFB < age - 14$. Hence, we have two modeling options. Either, we estimate by OLS using non-censored observations only; or we perform ML estimation of the censored model using all observations. In this set-up, OLS should be subject to attenuation bias. Table 7.3 displays the regression results. Since TFB is a non-negative variable, we have applied a log-transformation first. This is an example of a so-called "accelerated failure time model" that will be discussed in more detail in Chapter 8 on duration models. Here, coefficients estimate the relative marginal effects. A positive sign means that an increase in the associated regressor leads to a postponement of first birth. For example, an additional year of schooling is predicted to increase the time to first birth by approximately 7.7 percent. The estimated effect in the non-censored OLS model is about 15 percent smaller.

Table 7.3. *Censored Regression of Time to First Birth*

Dependent Variable: $\log(TFB)$		
	Censored Regression	OLS (non-censored)
years of education	0.077	0.066
	(0.005)	(0.005)
white	0.238	0.295
	(0.036)	(0.034)
number of siblings	-0.012	-0.005
	(0.005)	(0.005)
lived in city at age 16	0.007	-0.017
	(0.030)	(0.029)
immigrant	0.144	0.135
	(0.046)	(0.044)
Log-likelihood value	-1,115.4	
Observations	1,371	1,154

Notes: Standard errors in parentheses.

7.3.4 Truncated Regression Model

As for censoring, the starting point for the truncated regression model is the normal linear model

$$y = x'\beta + u, \qquad u|x \sim Normal(0, \sigma^2) \tag{7.35}$$

In a truncated sample, we no longer observe a random sample from the underlying population. Instead, the observed sample is selected systematically, based on the realization of the dependent variable. With **truncation from above**, we only observe (y, x) if $y \leq c$. The probability model for the observed data is then

$$g(y|y \leq c, x) = \frac{f(y|x)}{F(c|x)} \tag{7.36}$$

With **truncation from below**, we only observe (y, x) if $y \geq c$. The probability model for the observed data is then

$$g(y|y \geq c, x) = \frac{f(y|x)}{1 - F(c|x)} \tag{7.37}$$

These are general expressions that apply to any conditional distribution model $f(y|x)$. If the underlying model is normal, as assumed here, we can substitute the density and cumulative density function of the $Normal(x'\beta, \sigma)$-distribution for f and F, respectively.

The consequences for OLS are the same as in the censored model, i.e., the parameter estimates are biased towards zero. Instead, maximum likelihood estimation yields a consistent and asymptotically efficient estimator. For example, with truncation from below at zero, the log-likelihood function for an independent sample of n observations can be written as

$$\log L = \sum_{i=1}^{n} -\frac{1}{2}\left[\log(2\pi) + \log\sigma^2 + \frac{(y_i - x_i'\beta)^2}{\sigma^2}\right] - \log\Phi\left(\frac{x_i'\beta}{\sigma}\right) \tag{7.38}$$

If it were not for the second term on the right, this would be exactly the log-likelihood function of the normal linear regression model. $\Phi(x'\beta/\sigma)$ is the probability that an observation is not truncated. As this probability approaches one, the adjustment term goes to zero and the standard model is reached in the limit.

The truncated regression model is rarely used in practice. For example, it is applied in the Cragg (1971) model, where the positive part of the Tobit corner solution model is estimated, based on the likelihood function (7.38), using non-limit observations only. This approach can be used to test the assumption that the corner-solution part and the positive part can be described well by the same underlying latent process, with identical coefficients. Even if Tobin's restriction is valid, (7.38) could be used to estimate the model parameters.

However, this clearly would be inefficient in comparison with estimating the Tobit model using all data.

Exercise 7.4.

- In a poverty survey, only families with income of 20,000 USD or below are included. What is the relationship between observed incomes in that sample and the overall U.S. income distribution?
- The list of results for a road race only includes those participants who finished the race within a given maximum time allowance. What can we say about the average finishing time among all starters?

7.3.5 Incidental Censoring

A very useful and widely applicable generalization of the censoring model arises if the censoring threshold is modeled as a random variable. This is referred to as **incidental censoring**, or **self-selection** (Heckman, 1979). For example, with censoring from below, we observe $y|y > c$. So far, c was a known (although potentially variable) value. Now, we assume instead that c is a random variable as well.

Example 7.8. The Distribution of Wages

Wages are, by definition, only observed for workers. Still, it is possible to think of the wages that non-workers would receive if they decided to start working. These are potential wages, or wage offers. Therefore, we can say that the wage offers are equal to the observed wages of workers, whereas they cannot be observed for non-workers. In the standard labor economic paradigm, people choose to work if the marginal benefit of doing so (the wage times the marginal utility from consumption) exceeds the marginal cost (the lost utility due to time not spent on other activities, including leisure). The wage that just equalizes marginal benefits and costs is called the **reservation wage**. Therefore, the condition for observing a wage is that the wage offer w_o exceeds the reservation wage w_r. Consequently, the distribution of wages of workers is a conditional distribution $f(w_o|w_o > w_r)$. Since w_r is in general unknown, it is, from the econometrician's perspective, a random variable. This is an instance of incidental censoring, and appropriate methods are needed if one wants to use wages of workers in order to make inferences on the unconditional distribution of wage offers $f(w_o)$.

When y *and* c are stochastic, we can rewrite the conditional model of interest, $f(y|y > c) = f(y|y - c > 0)$, or, in general notation, $f(y_1|y_2 > 0)$. Clearly, the standard censoring model is obtained as a special case if $y_1 = y_2$. At the other extreme, when y_1 and y_2 are stochastically independent, $f(y_1|y_2 > 0) = f(y_1)$ and there is no selection problem. The censoring is random, and again, the question arises whether the relevant quantities of the underlying population can be estimated from the censored sample.

In order to model intermediate cases (where y_1 and y_2 are neither identical nor independent), the standard approach is to assume that y_1 and y_2 have a bivariate normal distribution with correlation ρ. This includes the polar cases of $\rho = 0$ (independence) and identity ($\rho = -1$ or $\rho = 1$). The bivariate normal distribution undoubtedly is a strong assumption. For generalizations, see, for example, a model with bivariate t-distribution as in Heckman, Tobias and Vytlacil (2003), or semi-parametric approaches as in Powell (1984).

The Regression Model

The full regression model with incidental censoring is often referred to as the **Heckman model** (Heckman, 1979), who himself used the term **self-selection model**. The prototypical model consists of two equations, an outcome equation and a selection equation, together with an observation rule:

$$\textbf{outcome equation:} \qquad y_1 = x'\beta + \sigma_1 u_1 \tag{7.39}$$

$$\textbf{selection equation:} \qquad y_2 = z'\gamma + u_2 \tag{7.40}$$

where the set of explanatory variables x and z may overlap. The **observation rule** is that y_1 is non-censored and observed for $y_2 > 0$. Moreover, u_1 and u_2 are assumed to have a standard bivariate normal distribution, independently of x and z, with means zero, unit variances and correlation ρ. Thus, the variance of y_1 is $\mathrm{Var}(y_1) = \sigma_1^2$, whereas the variance of y_2 is $\mathrm{Var}(y_2) = 1$. This normalization is required since we observe, as in the probit model, only two outcomes of the selection equation, namely $y_2 > 0$ (in which case y_1 is observed), or $y_2 \leq 0$ (in which case y_1 is not observed). The object of interest is β, the marginal effect in the outcome equation.

Why not OLS?

Direct estimation of the outcome equation by OLS, using the non-censored observations, does not lead to a consistent estimator of β, since under the assumptions of the model $\mathrm{E}(y_1|y_2 > 0, x) \neq x'\beta$. But what exactly is the expected value of y_1 conditional on selection in this set-up?

In order to derive the answer to this question, we need a result on the conditional expectation in a truncated bivariate normal distribution. First, for a standard bivariate normal distribution, we know from Appendix 7.5 that $\mathrm{E}(u_1|u_2) = \rho u_2$, where ρ is the coefficient of correlation. Therefore

$$\mathrm{E}(u_1|u_2 > c) = \mathrm{E}(\rho u_2|u_2 > c) = \rho \mathrm{E}(u_2|u_2 > c) = \rho \frac{\phi(c)}{1 - \Phi(c)}$$

With this result, we can now derive $\mathrm{E}(y_1|y_2 > 0)$:

$$\begin{aligned} \mathrm{E}(y_1|y_2 > 0, x) &= x'\beta + \sigma_1 \mathrm{E}(u_1|y_2 > 0, x) \\ &= x'\beta + \sigma_1 \mathrm{E}(u_1|u_2 > -z'\gamma, x) \\ &= x'\beta + \sigma_1 \mathrm{E}(\rho u_2|u_2 > -z'\gamma, x) \\ &= x'\beta + \sigma_1 \rho \lambda(z'\gamma) \end{aligned} \tag{7.41}$$

The second term on the right can be either positive or negative, depending on the sign of ρ. Thus, for example, $\mathrm{E}(y_1|y_2 > 0, x) > \mathrm{E}(y_1|x)$ if $\rho > 0$. In this situation, we say that the sample is **positively selected**, since the expected value in the sample exceeds the expected value in the population. On the other hand, if $\rho < 0$, we speak of **negative selection**.

As far as OLS with non-censored observations is concerned, there are two immediate problems. First, the estimate of the intercept β_0 is biased unless $\rho = 0$. Second, and in addition, the estimates of the slope parameters are biased as long as x and $\lambda(z'\gamma)$ are correlated. This is an "omitted variable bias"-type problem. The existence – and possible direction – of such a bias, depends on the particular application.

We illustrate this issue in the context of a classic example, the estimation of mean wage offers (or workers' productivity) from a sample of workers and their wages. Possibly the first to recognize the problem of sample selection bias in this application was Roy (1951). Heckman (1979) and Gronau (1974) explored its implications for estimation theory. We suggest that this approach be referred to as the "Gronau/Heckman/Roy wage model".

Example 7.9. The Gronau/Heckman/Roy Wage Model

Let w_o denote the wage offer and w_r denote the reservation wage (both in logarithms). Wage offers and reservation wages in the population are determined as follows:

$$w_o = x_o'\beta_o + u_o \tag{7.42}$$

$$w_r = x_r'\beta_r + u_r \tag{7.43}$$

where u_o, u_r are bivariate normally distributed, independently of x_o and x_r, with means zero, variances σ_o^2 and σ_r^2, and covariance σ_{or} (see Appendix 7.5). A person works (and the wage is observed, $w = w_o$) as long as $w_o > w_r$. Thus

$$\begin{aligned} P(w_o > w_r|x_o, x_r) &= P(x_o'\beta_o + u_o > x_r'\beta_r + u_r) \\ &= P(u_o - u_r > x_r'\beta_r - x_o'\beta_o) \\ &= \Phi\left(\frac{x_o'\beta_o - x_r'\beta_r}{\sigma}\right) \end{aligned} \tag{7.44}$$

where $\sigma = \mathrm{Var}(u_o - u_r)$. For workers

$$\begin{aligned} \mathrm{E}(w_o|w_o > w_r) &= \mathrm{E}(w_o|w_o - w_r > 0) \\ &= x_o'\beta_o + \mathrm{E}(u_o|u_o - u_r > x_r'\beta_r - x_o'\beta_o) \\ &= x_o'\beta_o + \frac{\mathrm{Cov}(u_o, u_o - u_r)}{\sigma}\lambda\left(\frac{x_o'\beta_o - x_r'\beta_r}{\sigma}\right) \end{aligned} \tag{7.45}$$

We observe that x_o must appear in the inverse Mills ratio, because x_o affects the wage offer, and the wage offer affects the decision to work. The sign of the inverse Mills ratio is positive if and only if $\sigma_o^2 - \sigma_{or} > 0$, or, equivalently, $\sigma_o/\sigma_r - \rho_{or} > 0$. Positive selection means that average wages of workers are higher than average wage offers of all persons. It is sufficient for positive selection that either $\rho_{or} < 0$ or $\sigma_o > \sigma_r$. To illustrate the latter effect, assume that reservation wages and wage offers are perfectly correlated – this is the single factor assumption used for example by Borjas (1999) – though not in Borjas (1987) – in the context of immigrant wages. Then, choosing work will tend to be favorable for above-average individuals only if their wage premium exceeds their increase in reservation wages, i.e., if $\sigma_o > \sigma_r$.

Estimation of the Model

The two main estimation methods of the model are either the **Heckman twp-step method** (Heckit), or full maximum likelihood. Two-step estimation implements a moment estimator of the conditional expectation function (7.42) in the censored sample, where the unknown $\lambda(z'\gamma)$ is replaced by a consistent estimator:

Step 1 Estimate the probit model $P(y_2 > 0|z) = \Phi(z'\gamma)$ and use the estimated coefficient $\hat{\gamma}$ to predict the inverse Mills ratio $\hat{\lambda}$ for each observation.

Step 2: Regress y_1 on x and $\hat{\lambda}$, using the sample of non-censored observations only.

Heckman (1979) shows that this two-step estimator is consistent, and provides a formula for the correct standard errors of the second-stage regression. A fully efficient estimator can be obtained by opting for the maximum likelihood estimator instead. The caveat of potentially excessive computing cost certainly no longer applies. In this case, the likelihood function has two elements. First, the likelihood contribution of an individual with observed wages is given by

$$\begin{aligned} f(y_1|y_2 > 0)f(y_2 > 0) &= f(y_1, y_2 > 0) \\ &= \phi_2(u_1, u_2 > -z'\gamma) \\ &= \int_{-z'\gamma}^{\infty} \phi_2\left(\frac{y_1 - x'\beta}{\sigma}, u_2\right) du_2 \end{aligned} \tag{7.46}$$

where $\phi_2(\cdot)$ is the bivariate standard normal distribution. To evaluate this expression, the trick is to change the order of the conditioning, i.e., to work with $f(y_2 > 0|y_1)f(y_1)$ rather than $f(y_1|y_2 > 0)f(y_2 > 0)$. The conditional distributions of the bivariate normal are derived in Appendix 7.5. They are normal distributions as well

$$u_2|u_1 \sim N(\rho u_1, 1 - \rho^2)$$

Hence, the density functions for non-limit observations can be written as

$$\begin{aligned} f(y_1, y_2 > 0) &= f(u_2 > -z'\gamma|y_1)f(y_1) \\ &= f(u_2 < z'\gamma|y_1)f(y_1) \\ &= \Phi\left(\frac{z'\gamma + \rho(y_1 - x'\beta)/\sigma}{\sqrt{1-\rho^2}}\right)\frac{1}{\sigma}\phi\left(\frac{y_1 - x'\beta}{\sigma}\right) \end{aligned} \tag{7.47}$$

and the log-likelihood for an observation with observed y_1 is

$$\log L_{nci} = -\log(\sigma) + \frac{1}{2}\left(\frac{y_{1i} - x_i'\beta}{\sigma}\right)^2 + \log\Phi\left(\frac{z_i'\gamma + (y_{1i} - x_i'\beta)\rho/\sigma}{\sqrt{1-\rho^2}}\right)$$

Second, if y_1 is unobserved, we know that $y_2 < 0$, or, equivalently, $u_2 < -z'\gamma$. The associated log-likelihood contribution is

$$\log L_{ci} = \log\Phi(-z_i'\gamma) = \log[1 - \Phi(z_i'\gamma)]$$

Thus, the log-likelihood function of the full sample is simply

$$\log L = \sum_{y_{2i}<0} \log L_{ci} + \sum_{y_{2i}\geq 0} \log L_{nci} \tag{7.48}$$

Identification

In the model we distinguish between the covariates x that affect the outcome, and the covariates z that affect the selection. Economic considerations often dictate that all variables that affect the outcome should also affect the selection. This is certainly the case if the selection is "self-selection", i.e., made by the individual who should usually take the outcome into account when making the decision.

In addition, there may be variables that affect the selection but not the outcome. Such "instruments" are highly desirable for precise estimation, although they may be difficult to find in a given dataset and application. In the Gronau-Heckman-Roy model, all variables affecting reservation wages but not wage offers are such instruments. Many people will find it reasonable to exclude the presence of small children from the wage offer equation, although it certainly should affect a woman's reservation wage.

In principle, the model is identified even if the variables in x and z are the same. However, such identification "by functional form" is not very convincing, because it depends on the normality assumption being exactly correct.

Example 7.10. Female Wages

This example is a continuation of Example 7.2, where female hours of work were analyzed in the context of Tobin's corner solution model. Now, we focus on wages instead, and directly estimate the Gronau-Heckman-Roy wage model. Again, we use the Mroz (1987) dataset on 753 women from the Michigan Panel Study of Income Dynamics. We only observe wages for the 428 women who work, whereas the remaining 325 wage observations are censored. This is an instance of incidental censoring, because the reservation wage will differ between people, depending on systematic factors including, for example, the presence of small children, but on unobservable factors as well.

Table 7.4 shows the OLS estimates for the non-censored sample, as well as the maximum likelihood estimates of the Heckman model. The dependent variable is the logarithmic wage. We assume that the variables *other income*, *age*, *number of young children* and *number of older children* affect the decision to work, but do not affect productivity and thus also not wage offers directly. In other words, these four factors are excluded from the wage equation.

Table 7.4. *Wage Offer Equation for Married Women*

Dependent Variable: ln(*wage*)		
	OLS (non-censored)	Maximum Likelihood
years of education	0.1075	0.1083
	(0.0141)	(0.0148)
years of experience	0.0416	0.0428
	(0.0132)	(0.0148)
$(\textit{years of experience})^2$	-0.00081	-0.0008
	(0.00039)	(0.0004)
constant	-0.522	-0.5526
	(0.1986)	(0.2603)
$\hat{\rho}$		0.0266
		(0.1471)
Observations	428	753

Notes: Standard errors in parentheses.

The results show no evidence of a sample selection problem in estimating the wage offer equation. The correlation coefficient has a very small t statistic (0.181), and so we cannot reject that it is zero and, by implication, that the selected sample is representative for the whole population, i.e., that $\mathrm{E}(y_1|y_2 > 0, x) = \mathrm{E}(y_1|x)$. This conclusion is confirmed informally by the fact that there are only minute differences in the estimated OLS and ML coefficients.

Exercise 7.5.

- Based on the structure of the Gronau-Heckman-Roy model, what does it mean in the wage model that we find no selection effect? Are the two errors u_o and u_r independent or correlated?

7.3.6 Example: Estimating a Labor Supply Model

The estimation of structural labor supply functions is a methodologically demanding problem since corner solution outcomes and censoring occur simultaneously (see Killingsworth, 1983, for a survey of empirical labor supply models). Consider the following individual-level labor supply function

$$h = x'\beta + \gamma w + \varepsilon \tag{7.49}$$

where h is hours of work, x is a vector of personal characteristics, w is the wage rate for the individual, and ε is the error term. The primary object of interest here is γ, since $\gamma \times w/h$ is the **wage elasticity** of labor supply.

There are at least two problems with estimating such a labor supply equation directly: first, h is nonnegative and has a cluster at zero; this is a corner solution issue. Second, w is unobservable for those with $h = 0$; this is an instance of incidental censoring. One might be tempted to just estimate the labor supply function using a sample of workers (i.e., those with $h > 0$). However, for the reasons discussed in this chapter, this solution does not work because the mean independence assumption of the classical linear regression model no longer holds. In particular, we have

$$\mathrm{E}(\varepsilon|h > 0) = \mathrm{E}(\varepsilon|\varepsilon > -x'\beta - \gamma w) \tag{7.50}$$

This residual has a non-zero mean, and in addition, it is correlated with x and w. Hence there is bias in both constant and slopes. To address this problem, we could implement a Tobit-type estimator for the sub-sample of workers, i.e., a truncated regression model. However, such an approach would not use the information efficiently. Moreover, it would rest on the assumption that mean independence between w and ε holds in the latent model (7.49). This would rule out omitted variables that increase both a person's productivity and the hours of work.

Therefore, it is more promising to approach the estimation problem in a general set-up, where wages are modeled simultaneously with hours in a recursive system of equations, and where the estimation problem takes into account that wage observations are censored, since wages are only observed for workers. Specifically, suppose that the labor supply equation is

$$h^* = x'\gamma + \gamma w + \varepsilon \tag{7.51}$$

with $h = h^*$ if $h^* > 0$ and $h = 0$ if $h^* \leq 0$. Furthermore, suppose that the market wage equation is

$$w = z'\alpha + u \tag{7.52}$$

and that the two error terms ε and u have a bivariate normal distribution with covariance $\sigma_{\varepsilon u}$.

In order to estimate this recursive system of equations, we proceed in three steps. In the first, we substitute $w = z'\alpha + u$ into the labor supply equation and estimate a **reduced-form** participation equation based on a probit model. Specifically

$$h^* = x'\beta + \gamma(z'\alpha + u) + e = r'\delta + v \tag{7.53}$$

where r is the union of the set of variables in x and z and $v = \gamma u + \varepsilon$. We can just estimate this Tobit equation using the full sample of workers and non-workers. There is no selection issue and the estimated vector of coefficients $\hat{\delta}$ will be a consistent estimate of the true δ. Alternatively, we could estimate δ using a probit model with the two outcomes $h^* > 0$ and $h^* \leq 0$. Note, however, that even if we can estimate δ consistently, we cannot fully recover the underlying coefficients of interest (β, γ, and α) because they are not separately identified in this reduced-form equation.

In the second step, we can run an OLS wage equation on the sample of workers, i.e., those persons who choose to participate in the labor market. However, we need to include the inverse Mills ratio at this stage to correct for the selectivity effect. In particular, the conditional expectation of wages for workers can be written under the bivariate normal assumption as

$$E(u|v > -r'\delta) = \rho_{uv}\lambda = \rho_{uv}\frac{\phi(-r'\delta)}{1 - \Phi(-r'\delta)} \tag{7.54}$$

where we assume that the variance of v is 1 (this is the standard normalization used in probit anyway). Since $v = \gamma u + \varepsilon$, the above conditional expectation is not equal to zero even if u and ε were uncorrelated. Thus, a consistent estimator of α can be obtained in the following manner: Use the $\hat{\delta}$ estimated from stage one to predict λ, using the formula above. Then, regress wages against the vector z and the variable $\hat{\lambda}$, using the sample of workers only.

Once we get an unbiased estimate $\hat{\alpha}$ from stage two, we can construct an unbiased predictor $\hat{w} = z'\hat{\alpha}$ for all persons, workers and non-workers alike. Then, in the third step, we may run, say, a Tobit equation of h on x and $\hat{w}$. The coefficient of $\hat{w}$ is a consistent estimator of γ in the labor supply equation.

7.4 Treatment Effect Models

7.4.1 Introduction

This chapter offers a short introduction to the treatment literature. For more detailed discussions, see Heckman and Rob (1985), Holland (1986), Moffitt (1991), Rubin (1974), and Wooldridge (2002, Chapter 18). At the beginning of each evaluation, one needs to specify the outcome variable of interest, y. Let y_1 be the outcome (e.g., employment, wages, crime) for an individual if he participates in a certain social program (e.g., retraining, education, community service), and let y_0 be the outcome if he does not participate. The causal effect of the treatment then is defined as $y_1 - y_0$. The fundamental problem of program evaluation is that we observe either y_1 or y_0, but never both. Identification of the treatment effect requires additional assumptions.

Let d denote an indicator for treatment. If $d = 1$, then the individual received the treatment, if $d = 0$, the individual is part of the **control group** and did not receive the treatment. We are not interested in individual outcomes $y_1 - y_0$, which differ from person to person, but rather in population averages. The **average treatment effect** is defined as

$$ATE = \mathrm{E}(y_1 - y_0) \tag{7.55}$$

whereas the **average treatment effect on the treated** is

$$ATT = \mathrm{E}(y_1 - y_0 | d = 1) \tag{7.56}$$

We can directly identify from the data $\mathrm{E}(y_1|d = 1)$ and $\mathrm{E}(y_0|d = 0)$ since $y = y_0 + d(y_1 - y_0)$, but this is not sufficient in itself to estimate the ATE or the ATT. Consider the ATT first. By definition,

$$\mathrm{E}(y_1 - y_0|d = 1) = \mathrm{E}(y_1|d = 1) - \mathrm{E}(y_0|d = 1) \tag{7.57}$$

The second term is unobserved. If we *assume* that the expected outcome without treatment is the same for treatment and control group – formally, this is **mean independence** between y_0 and d – then we can replace the unobserved term $\mathrm{E}(y_0|d = 1)$ in (7.57) by the observed term $\mathrm{E}(y_0|d = 0)$ and estimate the ATT in this way. The required assumption for identifying the ATE are somewhat stronger. Here, it is not sufficient that y_0 and d are independent, but we also require that y_1 and d are independent. This is so, because we can write

$$ATE = P(d = 1)\mathrm{E}(y_1 - y_0|d = 1) + P(d = 0)\mathrm{E}(y_1 - y_0|d = 0) \tag{7.58}$$

Hence, we also require an estimate of $\mathrm{E}(y_1|d = 0)$ in order to estimate the average treatment effect.

In a **randomized experiment**, treatment and outcome are by definition independent, $ATE = ATT = \mathrm{E}(y|d = 1) - \mathrm{E}(y|d = 0)$, and one can estimate the average treatment effect by a simple regression of y on d.

Example 7.11. Randomized Experiments in Economics

Randomized experiments are the exception rather than the rule in empirical economic and social research. One can distinguish between **laboratory experiments** and so-called **field experiments**. While randomization in laboratory experiments is easily achieved, the scope of such experiments is limited. People may behave differently in laboratory and "real world" settings, and many important policy questions cannot be addressed. How will a tax, social security or health reform affect labor supply or demand for health services? What is the benefit of a re-training program for the participants? To answer such questions, large-scale social field experiments have been conducted in the U.S. and elsewhere since the 1960's (see, for instance, Burtless, 1995, and Heckman, 1995).

Unfortunately, there is ample evidence that genuine randomization in social experiments is very hard to achieve. When dealing with human subjects - unlike land plots as in Fisher's original paradigm - human interference is likely to invalidate pure randomization. Consider the example of an active labor market policy that is aimed at improving the re-employment probability of the unemployed. First, some government agency has to perform the random selection of treatment and control group (those that are denied the treatment). Since the administrators involved often have an incentive to exaggerate the benefits of the program, they might find subtle ways for boycotting the randomization by allocating "better" individuals to the treatment. Secondly, unlike in some clinical trials where true placebos exist and the subject may be unaware of belonging to the control group, participants in social experiments who are denied the treatment generally know about it, and they may gain access to close substitutes for the experimental treatment, i.e., partake in another training program. Again, the comparability between treatment and control is compromised. And thirdly, when the program success is measured over time, attrition may become a problem. If attrition rates depend on the outcome of interest, bias may arise.

Such failure of genuine randomization is not equally likely in all field experiments, though. Consider the following three recent examples for field experiments where randomization may be successful.

- In order to find out whether employers' hiring practices in the U.S. discriminate against blacks, Bertrand and Mullainathan (2004) sent out thousands of resumes to different manual and service sector job openings (advertised in Boston and Chicago Newspapers). They randomized the names of the applicants: e.g., Lakisha versus Emily, Jamal versus Greg, and kept the

résumés otherwise identical. In this case, the coefficient β in the regression model

$$callback_i = \alpha + \beta\, black\, name_i + u_i$$

identifies the treatment effect. They indeed found evidence of discrimination. Individuals with white-sounding given names were about twice as likely to get callbacks.

- Fehr and Goette (2004) conducted a field experiment involving two bicycle messenger companies in Zurich. In one company, messengers were randomly selected into two groups. Each group received a temporary increase in pay – the share of generated revenue – during a different week. In the other company, nothing changed. Wage records make it possible to estimate the treatment effect on working hours and effort. The authors find no evidence of intertemporal substitution.
- The Progresa program in Mexico provides poor mothers in rural Mexico with education grants if their children regularly attend school. It was preceded by a genuine experiment, where communities were randomly selected to participate. School attendance of eligible children increased substantially in the communities treated (Schultz, 2004).

In **observational data**, however, the independence assumption is not plausible. It can be somewhat weakened by assuming that selection into treatment and outcome are independent, *conditional* on regressors x, such that $\mathrm{E}(y_0|d = 1, x) = \mathrm{E}(y_0|d = 0, x)$ and $\mathrm{E}(y_1|d = 1, x) = \mathrm{E}(y_1|d = 0, x)$. A possible regression model then is

$$y = x'\beta + \delta d + \varepsilon \tag{7.59}$$

where d and ε are independent, conditional on x, and $ATE = \delta$. Of course, other functional forms might be more appropriate. For example, if we interact d and x, we obtain a switching regression model.

However, estimating model (7.59) by OLS is not very promising either, since the crucial assumption of **conditionally independent** selection into treatment is likely to be too strong in most applications, and $\hat{\delta}_{ols}$ then is a biased estimator of the ATE. There are several possibilities to account for the possible correlation between d and ε. One is **instrumental variable** estimation, i.e., to find a variable that affects selection into treatment but not the outcome itself. Another approach, in the spirit of this chapter, is to specify a joint distribution for d and ε – typically a bivariate normal distribution – and estimate the parameters of the model by **maximum likelihood** or using a two-step approach. We illustrate this approach in two settings, first the endogenous binary variable model and second the switching regression model.

7.4.2 Endogenous Binary Variable

The outcome equation is

$$y = x'\beta + \delta d + \varepsilon \tag{7.60}$$

where ε and x are mean independent. Selection into treatment is based on the **latent model**

$$d = I(z'\gamma + u > 0) \tag{7.61}$$

If we further assume that the joint distribution of u and ε is a **bivariate normal** distribution with variance matrix

$$\Sigma = \begin{bmatrix} 1 & \rho\sigma \\ \rho\sigma & \sigma^2 \end{bmatrix}$$

then we can immediately derive the expectation of the dependent variable conditional on treatment, since

$$\begin{aligned} \mathrm{E}(y|x, d=1) &= x'\beta + \delta + \mathrm{E}(\varepsilon|d=1) \\ &= x'\beta + \delta + \mathrm{E}(\varepsilon|u > -z'\gamma) \\ &= x'\beta + \delta + \rho\sigma\frac{\phi(z'\gamma)}{\Phi(z'\gamma)} \end{aligned} \tag{7.62}$$

For non-participants, the counterpart is

$$\mathrm{E}(y|x, d=0) = x'\beta - \rho\sigma\frac{\phi(z'\gamma)}{1-\Phi(z'\gamma)} \tag{7.63}$$

The difference between participants and non-participants then is

$$\mathrm{E}(y|x, d=1) - \mathrm{E}(y|x, d=0) = \delta + \rho\sigma\frac{\phi(z'\gamma)}{\Phi(z'\gamma)[1-\Phi(z'\gamma)]} \tag{7.64}$$

Without correction for endogeneity, the OLS coefficient on the treatment dummy estimates this difference. If $\rho > 0$, it follows that $\mathrm{E}(y|d=1) - \mathrm{E}(y|d=0) > \delta$. The treatment effect is overestimated by OLS; there is an upward bias. A consistent estimator of the treatment effect can be obtained from a Heckman-type two-stage estimator, or by full maximum likelihood estimation. Both procedures rely on joint normality of ε and u.

Estimation Under Weaker Assumptions

The model with a binary endogenous variable can also be estimated under weaker assumptions, i.e., without assuming bivariate normality of u and ε. There are two approaches that differ in the restrictiveness of their assumptions.

1. **Plug-in solution:** Replace d in the outcome equation by an estimate of $\mathrm{E}(d)$. For example, with the above probit-type self-selection model, one would use $\Phi(z'\hat{\gamma})$. The standard errors need to be adjusted in order to account for the two-step estimation.
2. **IV-solution:** If an instrument is available, estimation can proceed by **two-stage least squares** (2SLS) as usual. First, d is regressed on all exogenous variables of the model, and second, $\hat{d}$ is substituted in the second-stage OLS.

To understand why the first method works, marginalize the outcome equation with respect to d:

$$\mathrm{E}(y|x,d) = x'\beta + \delta d + \mathrm{E}(\varepsilon|d) \tag{7.65}$$

Taking expectations over d using the law of the iterated expectation, we get

$$\mathrm{E}(y|x,z) = \mathrm{E}_d[\mathrm{E}(y|x,d)] = x'\beta + \delta\mathrm{E}(d|z) + \mathrm{E}_d[\mathrm{E}(\varepsilon|d)] \tag{7.66}$$

But the last term is zero, since the marginal distribution of ε has zero mean by assumption. Thus we can write

$$y = x'\beta + \delta\mathrm{E}(d|z) + v \tag{7.67}$$

where v and $\mathrm{E}(d|z)$ are independent. It is required for this method that $\mathrm{E}(d|z)$ is correctly specified, which means in the context of this model that u has a normal distribution. It is not required, however, that ε is normally distributed, or that the joint distribution of u and ε is known.

Exercise 7.6.
Show that in the above model with bivariate normality of u and ε, it holds that

$$\mathrm{E}(y|x,z) = P(d=0)\mathrm{E}(y|d=0) + P(d=1)\mathrm{E}(y|d=1) = x'\beta + \delta\Phi(z'\gamma)$$

Instrumental variable estimation is even more general, since no assumption about the distribution of u is required. On the other hand, we need an instrument which, strictly speaking, is not necessary under the other two approaches.

7.4.3 Switching Regression Model

An early formulation of a switching regression model is Lee's (1978) study of union and non-union wages. Another example is Willis and Rosen (1979),

who analyze the wages of people with and without college degree in such a self-selection framework. The switching regression model is defined by two separate equations of potential outcomes

$$y_0 = x'\beta_0 + u_0 \tag{7.68}$$

$$y_1 = x'\beta_1 + u_1 \tag{7.69}$$

As before, y_0 is the outcome without treatment and y_1 is the outcome with treatment. The model can thus alternatively be written as

$$y = x'\beta_0 + (x'\beta_1 - x'\beta_0)d + u_0 + (u_1 - u_0)d \tag{7.70}$$

We see that this is a generalization of the endogenous dummy variable model, since the latter is obtained if all slopes are the same, as are the errors, but only the intercepts are allowed to differ. The average treatment effect (ATE) in the switching regression model is defined as

$$\mathrm{E}(y_1 - y_0|x) = x'(\beta_1 - \beta_0) \tag{7.71}$$

and the treatment effect on the treated is

$$\mathrm{E}(y_1 - y_0|x, d = 1) = x'(\beta_1 - \beta_0) + \mathrm{E}(u_1 - u_0|d = 1) \tag{7.72}$$

All treatment effects depend on x. The unconditional treatment effect is obtained by taking expectations over x. Similarly, one can define a local average treatment effect (LATE) and a marginal treatment effect (MT), see Heckman, Tobias and Vytlacil (2003).

The switching regression model is completed by specifying the selection into treatment. We assume as in (7.61) that

$$d = I(z'\gamma + u_d > 0) \tag{7.73}$$

Selection bias can now arise in three different ways. Participation in the treatment may be correlated with u_1, i.e., the outcome in the treated state, with u_0, i.e., the outcome in the control state, or both. In the Roy model, for instance, we have $d = I(y_1 - y_0 > 0) = I(u_1 - u_0 > x'\beta_0 - x'\beta_1)$, i.e., selection based on gains or **relative advantage**.

One could attempt to estimate the two equations for treatment and control separately by OLS, using the respective samples (i.e., the y_0 equation is estimated using all observations that were not treated; the y_1 equation is estimated using all observations that received the treatment). However, this is unlikely to produce good results, as in each of the cases, we face a standard censored regression problem, only in this case, the problem is two-sided, because observations that are censored in one equation (we do not see the counterfactual outcome with treatment for those who were not treated) are non-censored in the other, and vice versa. Accordingly, we usually have that

$$\mathrm{E}(y_0|x, d = 0) = x'\beta_0 + \mathrm{E}(u_0|d = 0) \neq x'\beta_0 \tag{7.74}$$

$$\mathrm{E}(y_1|x, d = 1) = x'\beta_1 + \mathrm{E}(u_1|d = 1) \neq x'\beta_1 \tag{7.75}$$

Specific expressions can be obtained if we assume that u_0, u_1 and u_d have a trivariate normal distribution. For instance,

$$\mathrm{E}(y_1|x, d = 1) = x'\beta_1 + \mathrm{E}(u_1|u_d > -z'\gamma) = x'\beta_1 + \rho_1\sigma_1\frac{\phi(z'\gamma)}{\Phi(z'\gamma)}$$

where ρ_1 is the correlation between u_1 and u_d. Again, the parameters can be estimated consistently by maximum likelihood, or by appropriately defined two-step estimators.

Example 7.12. The Union/Non-Union Wage Differential

Lee (1978) assumed separate wage equations for union workers and for non-union workers. The model is

$$\ln w_{ui} = x'_{ui}\theta_{ui} + \varepsilon_{ui}$$

$$\ln w_{ni} = x'_{ni}\theta_{ni} + \varepsilon_{ni}$$

$$I_i^* = \delta_0 + \delta_1(\ln w_{ui} - \ln w_{ni}) + z_i'\delta_2 + v_i$$

where $\varepsilon_{ui} \sim Normal(0, \sigma_u^2)$, $\varepsilon_{ni} \sim Normal(0, \sigma_n^2)$, $v \sim Normal(0, \sigma_v^2)$, and a worker belongs to a union (and the wage is determined by the first equation) if $I_i^* \geq 0$. The model is indeed an extended version of the Roy model. Selection bias occurs, even though the errors in the three equations are independent. This is so, because the difference $\ln w_{ui} - \ln w_{ni}$, and therefore the errors in the first two equations, affect the decision of a worker to become a unionized or a non-unionized worker, just as in the Roy model.

Lee proposes two-stage estimation of the model. In a first step, a reduced-form selection equation is obtained after substituting the wage equations into the selection equation:

$$\begin{aligned} I_i^* &= \delta_0 + \delta_1(x'_{ui}\theta_{ui} + \varepsilon_{ui} - x'_{ni}\theta_{ni} - \varepsilon_{ni}) + z_i'\delta_2 + v_i \\ &= \gamma_0 + \gamma_1 w_i + \xi_i \end{aligned}$$

where w_i contains all exogenous variables in x and z. The parameters can be estimated by probit. Since the union wage equation, conditional on union status, is

$$\ln w_{ui} = x'_{ui}\theta_{ui} + \sigma_{u\xi}\frac{\phi(\psi_i)}{\Phi(\psi_i)} + \eta_{ui}$$

$$\ln w_{ni} = x'_{ni}\theta_{ni} - \sigma_{n\xi}\frac{\phi(\psi_i)}{1 - \Phi(\psi_i)} + \eta_{ni}$$

where $\psi_i = \gamma_0 + \gamma_1 w_i$, one can estimate the two equations by OLS, using observations on the subsamples, where ψ_i is replaced by $\hat{\psi}_i = \hat{\gamma}_0 + \hat{\gamma}_1 w_i$.

7.5 Appendix: Bivariate Normal Distribution

If the joint density function of two random variables (u_1, u_2) has the form

$$f(u_1, u_2) = \frac{1}{2\pi} \frac{1}{\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2} \left(\frac{u_1^2 + u_2^2 - 2\rho u_1 u_2}{1-\rho^2} \right) \right\} \tag{7.76}$$

u_1 and u_2 are said to have a **bivariate standard normal distribution**. The marginal distributions are standard normal: $u_1 \sim Normal(0,1)$ and $u_2 \sim Normal(0,1)$. The correlation between u_1 and u_2 is given by ρ. The basic rule for conditional densities is that $f(u_1|u_2) = f(u_1, u_2)/f(u_2)$. Since

$$f(u_2) = \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} u_2^2 \right\} \tag{7.77}$$

and using some simplifications, we find that the ratio $f(u_1, u_2)/f(u_2)$ can be written as

$$f(u_1|u_2) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2} \left(\frac{(u_1 - \rho u_2)^2}{1-\rho^2} \right) \right\} \tag{7.78}$$

In other words, $u_1|u_2$ has a normal distribution as well, and

$$u_1|u_2 \sim Normal(\rho u_2, 1-\rho^2) \tag{7.79}$$

The conditional expectation function of the standard bivariate normal distribution is given by

$$\mathrm{E}(u_1|u_2) = \rho u_2 \tag{7.80}$$

The results can be generalized to allow for marginal normal distributions with non-zero means and non-unit variances. Let

$$z_1 = \mu_1 + \sigma_1 u_1$$
$$z_2 = \mu_2 + \sigma_2 u_2$$

so that $z_1 \sim Normal(\mu_1, \sigma_1^2)$ and $z_2 \sim Normal(\mu_2, \sigma_2^2)$, and $(z_1, z_2) \sim Bivariate\ Normal(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ denotes the general **bivariate normal distribution**. For the conditional expectation, we obtain

$$\begin{aligned} \mathrm{E}(z_1|z_2) &= \mu_1 + \sigma_1 \mathrm{E}(u_1|z_2) \\ &= \mu_1 + \sigma_1 \rho \left(\frac{z_2 - \mu_2}{\sigma_2} \right) \\ &= \mu_1 + \rho \frac{\sigma_1}{\sigma_2} (z_2 - \mu_2) \\ &= \mu_1 + \frac{\sigma_{12}}{\sigma_2^2} (z_2 - \mu_2) \end{aligned} \tag{7.81}$$

This result shows that under bivariate normality, it is the case that $\mathrm{E}(z_1|z_2) = \alpha + \beta z_2$, where $\beta = \sigma_{12}/\sigma_2^2$ and $\alpha = \mu_1 - \beta\mu_2$. Thus, the conditional expectation function of the bivariate normal distribution is a linear function, or a "linear regression".

7.6 Further Exercises

Exercise 7.7 Let y denote a continuous random variable with density function $f(y) = 2y$ with $0 \leq y \leq 1$. Find

a) $\mathrm{E}(y)$

b) $f(y|y > 0.5)$

c) $\mathrm{E}(y|y > 0.5)$

Exercise 7.8 Let y denote a continuous random variable following a normal distribution with $y \sim Normal(3, 4)$. Find

a) $f(y|y > 3)$

b) $\mathrm{E}(y|y > 3)$

c) $\mathrm{Var}(y|y > 3)$

Exercise 7.9 Consider the latent model $y^* = x'\beta + u$. Assume that u is normally distributed with mean zero and variance σ^2. You observe $y = \min(0, y^*)$.

a) Find $\mathrm{E}(y|y < 0, x)$.

b) Write down the likelihood function of the model.

c) How can you obtain a two-step estimator of β? Describe in detail the first and the second step involved in the procedure.

Exercise 7.10 Consider the latent model $y^* = x'\beta + u$. Assume that u is normally distributed with mean zero and variance σ^2. You observe $y = \max(t, y^*)$ where t is a fixed and known number. Write down the likelihood function in this situation and compare with the special case $t = 0$.

Exercise 7.11 In which of the following situations is it of interest to identify the parameters of the latent model?

a) When estimating the determinants of hours of work.

b) When estimating the determinants of wage offers.

c) When estimating the counterfactual wages of immigrants at their origin (i.e., had they not migrated).

Exercise 7.12 Suppose you consider the Tobit model for estimating the determinants of individual fertility (i.e., the number of children ever born). What are alternative models, and how do you see their respective advantages and disadvantages?

Exercise 7.13 The following table reconsiders the labor supply example, based on the Mroz data using 753 observations from the Panel Study of Income Dynamics. The dependent variable is the annual hours of work by married women. The first column reprints the Tobit results from Table 7.1. The second and third columns show the results for the more flexible Cragg model. Column 2 is a probit for zero hours versus positive hours. Column 3 is a truncated at zero model for positive hours.

Dependent Variable: *hours of work*			
		Cragg Model	
	Tobit	Probit	Truncated
nonwife income	-8.8	-0.012	0.2
	(4.5)	(0.005)	(5.2)
years of education	80.6	0.131	-29.9
	(21.6)	(0.025)	(22.8)
years of experience	131.6	0.123	72.6
	(17.3)	(0.019)	(21.2)
(*years of experience*)^2	-1.9	-0.002	-0.9
	(0.5)	(0.001)	(0.6)
age	-54.4	-0.053	-27.4
	(7.4)	(0.008)	(8.3)
# kids less than 6 years	-894.0	-0.868	-484.7
	(111.9)	(0.119)	(153.8)
# kids between 6 and 18	-16.2	0.036	-102.7
	(38.6)	(0.043)	(43.5)
constant	965.3	0.270	2,123.5
	(446.4)	(0.509)	(483.3)
$\hat{\sigma}$	1,122.0		850.8
	(41.6)		(43.8)
Log-likelihood value	-3,819.0	-401.3	-3,390.6
Observations	753	753	428

Notes: Standard errors in parentheses.

a) Write down the formal models on which these estimates are based.

b) Do the parameter estimates appear "plausible" to you?

c) How can you test the Tobit model against the Cragg model?

d) What is the distribution of your test statistic, and what do you conclude?

Exercise 7.14 Suppose you want to estimate the effect of alcohol abuse on wages (e.g., Hamilton and Hamilton, 1997). Which econometric framework would you use? What are the potential problems that you could encounter?

Exercise 7.15 Use the data in *mroz.dta* to estimate the following linear regression for logarithmic wages of female workers:

$$lwage = \beta_0 + \beta_1\, educ + \beta_2\, exper + \beta_3\, expersq + u$$

Next, generate a new variable, say *clwage*, such that

$$clwage = \begin{cases} lwage & \text{if } lwage \geq 0.5 \\ \text{unobserved} & \text{else} \end{cases}$$

Now, perform OLS and censored regression, replacing *lwage* with *clwage*. How do your estimates of the two models compare? What happens as you increase the censoring point?

Exercise 7.16 Fertility data from the General Social Survey were introduced in Chapter 1.5.1. The data provide information on the number of children and education levels for a random sample of 5,150 women over the years 1974 to 2002. It was mentioned then that the number of children is censored from above at eight.

a) What does this imply for the linear least squares estimator of the model

$$children = \beta_0 + \beta_1 educ + \beta_2 time + u$$

b) Estimate the appropriate censored regression model and interpret your results.

Exercise 7.17 A simple migration model posits that workers migrate if the wage in the host country w_h exceeds the wage in the country of origin w_o. Let wages be determined as follows:

$$w_o = \mu_o + \eta_o \nu$$
$$w_h = \mu_h + \eta_h \nu$$

where ν is a measure of skills that are perfectly transferable and η_j gives the rate of returns to skills in country j. Assume that $\nu \sim Normal(0, 1)$.

a) Determine the variances, covariance and correlation of w_o and w_h.

b) In terms of the Roy model, what is the condition that migrants are positively selected, i.e., that $\mathrm{E}(w_h | w_h > w_o) > \mathrm{E}(w_h)$? Provide an economic intuition of your result. What does it imply, for instance, for the wages of Swedish migrants to the U.S.A.?

Exercise 7.18 Consider the following artificially generated data:

$$y_1 = 0.5 + 0.5x + u_1$$
$$y_2 = 0.7 + 0.3x + u_2$$

y_1 is observed whenever $y_1 > y_2 + z$. x and z are drawn from a standard uniform distribution, whereas u_1 and u_2 have standard normal distributions.

Assume that the first equation is estimated using the Heckman selection model with z as an instrument. What are the probability limits of the two parameters of the outcome equation, $y_1 = \beta_0 + \beta_1 x + u_1$, of the three parameters of the selection equation, $P(y_1 \text{ is observed}) = \gamma_0 + \gamma_1 x + \gamma_2 z$, and of ρ and σ?

Exercise 7.19 Consider the latent model

$$y_1^* = x_1'\beta_1 + u_1$$
$$y_2^* = x_2'\beta_2 + u_2$$

where the vector of error terms $(u_1, u_2)'$ follows a bivariate normal distribution. You observe (y_1, y_2, x_1, x_2), where $y_1 = \max(0, y_1*)$, and $y_2 = y_2^*$ if $y_1^* > 0$ and $y_2 = 0$ otherwise. Find the likelihood function for a sample of n independent observations.

Exercise 7.20 A popular estimator of treatment effects is the so-called **difference-in-differences** estimator. In this case, the outcome for the treatment group ($d = 1$) and for the control group ($d = 0$) is observed twice, once before the treatment ($t = 0$) and once after the treatment ($t = 1$). The treatment effect on the treated is then

$$\mathrm{E}(y_1 - y_0 | d = 1, t = 1) = \mathrm{E}(y_1 | d = 1, t = 1) - \mathrm{E}(y_0 | d = 1, t = 1) \quad (7.82)$$

where $\mathrm{E}(y_0 | d = 1, t = 1)$ is the counterfactual outcome. Assume that $\mathrm{E}(y_0 | d = 1, t = 1) - \mathrm{E}(y_0 | d = 1, t = 0) = \mathrm{E}(y_0 | d = 0, t = 1) - \mathrm{E}(y_0 | d = 0, t = 0)$ i.e., the *change* for the treatment group in the absence of the treatment would have been the same as the actual *change* observed for the control group.

a) How does this assumption allow you to identify the treatment effect in (7.82)?

b) In what sense is this assumption weaker than assuming $\mathrm{E}(y_0 | d = 1, t = 1) = \mathrm{E}(y_0 | d = 0, t = 1)$?

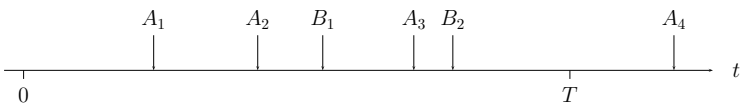
8

Event History Models

8.1 Introduction

The discussion so far has excluded the dimension *time* almost entirely from the modeling of microdata relationships. In this chapter, we will extend our view to models in which the dependent variable is directly related to the factor time, while maintaining the idea of cross-sectional and thus independent observations. More specifically, we will consider two types of models, **count data** and **duration models**. As we will see, both approaches are closely related, and we emphasize this connection by discussing them under the common notion **event history models**. An event history, as we understand the term, is a record of when, if at all, specific events occurred to an individual over time. Count data models then describe the number of occurrences of a particular event in a given time interval, whereas duration models describe the time that has elapsed between occurrences of two particular events. Figure 8.1 gives an example of an individual event history with events $A_{1,\ldots,4}$ and $B_{1,2}$.

Fig. 8.1. *Example of an Individual Event History*



Duration analysis has its origins in what is typically referred to as **survival analysis**, i.e., the study of survival times of a subject and its dependence on a treatment and individual characteristics. For example, an experiment may be performed in which laboratory animals are exposed to different doses of a toxic substance. The experimenter then observes how long the animal survives

under the different doses. Analyzing the *survival time* is a more complex issue than analyzing the event "survival until T", an example of a binary outcome. Whereas for the latter question, logit or probit analysis would work, different methods are generally needed for the former.

In economic applications, "survival" typically takes a different meaning. A subject is observed in a specific state (being unemployed, being without further arrest, etc.), and we call the time spent in this state an episode or **spell**. A spell ends when a terminal event occurs (find a job, be arrested again, etc.), and we want to model the duration of the spell as a function of explanatory variables. In Figure 8.1, this could be, for example, the time between the events B_1 and A_3.

Example 8.1. Applications of Duration Data

- the duration of unemployment (Lancaster, 1979)
- the duration of a strike (Kennan, 1985)
- the number of months since release from prison before an ex-convict is arrested again for a crime (Beck and Shipley, 1989)
- the time it takes to sell a house from the moment it is listed for sale (Genesove and Mayer, 1997)
- the age at first birth (Ermisch and Ogawa, 1994)

Individual event histories, as shown in Figure 8.1, allow for an alternative method of data analysis. Instead of focusing on the survival time we could *count* the occurrences of a particular event within a fixed time interval. For example, we may examine the time until a car breaks down for the first time (a duration variable), or instead, we may analyze the number of breakdowns of ten-year-old cars (a count variable). In terms of Figure 8.1, the number of events A within the fixed time interval $(0, T)$ may be considered as a count (in this case with an outcome of 3).

Count data models are characterized by a dependent variable that only takes non-negative integer values, the number of occurrences. Compared to multinomial and ordered dependent variables (see Chapters 5 and 6), the sequence $\{0, 1, 2, \ldots\}$ *has* a quantitative meaning for counts, and therefore cannot be replaced by any other (arbitrary) sequence of numbers. Thus, when modeling such data, we need to take into account the implied cardinality. Moreover, as with all discrete data models considered in this book, each possible outcome has a positive probability, and we want to continue being able to draw inferences about these outcomes. Therefore, we will direct our attention to the class of conditional probability models for counts.

The following list of examples shows how diverse empirical applications of count data can be.

Example 8.2. Applications of Count Data

- the number of patents (Hausman, Hall and Grilliches, 1984)
- the number of strikes per month in a certain industry, e.g., manufacturing (Kennan, 1985)
- the number of airline incidents per thousand scheduled departures within one year (Rose, 1990)
- the number of doctor visits (Winkelmann, 2004)
- the number of children born to women aged 40 to 65 (Winkelmann and Zimmermann, 1994)

Duration and count data have attracted much interest in recent years, and a number of very complex and highly specialized models are available for both kinds of data. In this chapter, we confine ourselves to the basic elements and models in duration and count data analysis. For more advanced treatments of duration data, we refer to Kiefer (1988), Lancaster (1990), and van den Berg (2001); for count data we recommend the textbooks by Winkelmann (2003) and Cameron and Trivedi (1998).

This chapter is organized as follows. Section 8.2 presents the basic tools to analyze duration data. We begin with two examples, an artificial and a real data example, to illustrate the main concepts involved in duration analysis, such as the hazard rate and the survivor function. We distinguish between discrete time and continuous time duration models, and discuss several problems arising in the estimation of such models. A significant part is devoted to the key element of duration analysis, the hazard function, and we introduce the concepts of duration dependence and unobserved heterogeneity.

The subject of Section 8.3 is the analysis of count data. We start with the basic count data model, the Poisson regression model. We show how the model parameters are interpreted in terms of marginal mean and probability effects, we estimate the model using ML methods, and point out the relationship between count data and duration analysis. The remainder of this section addresses some of the shortcomings of the Poisson regression model, such as unobserved heterogeneity, censoring or truncation, and frequent zeros.

8.2 Duration Models

8.2.1 Introduction

As always when contemplating a new class of models tailored to a specific data situation, we should ask ourselves: what is different? Why are the discrete and continuous data models studied so far, including the linear regression model and OLS, insufficient? At first glance, "time" appears as a quite normal dependent variable, not so different from income or the number of children. Time, like income and children, is certainly a non-negative variable, and this should be accounted for in modeling and estimation. But this could be done easily using a log-transformation or a non-linear expectation function.

What, then, is so special about time? There are two points to consider. First, censoring is ubiquitous in duration data. As an example, reconsider Figure 8.1, where you will find that at observation time T, the ultimate length of the third spell of type "A", the one lasting from A_3 to A_4, is unknown. Similarly, the second spell of type "B" is ongoing at time T. Second, and more importantly, the expected duration is often of less substantive interest than a related concept, the conditional exit rate, or hazard rate. This is specific to duration data, and we will provide a formal definition below. Here, we only give two suggestive examples why the hazard rate is of particular interest to economists. Both are related to unemployment.

First, to study the incentive effect of temporary limited unemployment benefits, we are interested in particular in the hazard rate around the expiry date of the benefit. We would expect a spike at that point in time. Such an effect will be much more clearly seen in the hazard rate than in the distribution of exit times. It cannot be seen though if looking at expectations only. Second, a prominent hypothesis about unemployment is the "hysteresis" hypothesis: unemployment breeds unemployment in a sense that the longer an unemployment spell lasts, the less likely it is that the person will find work again. This hypothesis is directly related to the hazard rate. It posits that the hazard rate should fall over time, an instance of "negative duration dependence".

8.2.2 Basic Concepts

Duration analysis, generally speaking, is about the *modeling of time spent within a particular state until moving on to another state.* The length of this time interval is usually referred to as **spell length** or **survival time**. We distinguish between **discrete** and **continuous** survival times. In the former, we assume that spell lengths can be measured in infinitely small units, whereas in the latter, we assume that time is either intrinsically discrete (e.g., business cycles) or marked off in discrete time intervals, although the underlying process could have been measured continuously (e.g., number of years). Durations may incorporate very complex data patterns since individuals may enter

the state of interest more than once, resulting in repeated spells, but we will focus on the case of **single spells** here.

Table 8.1 illustrates some of the main concepts encountered in duration analysis. Consider 100 individuals for whom a spell starts at $t = 1$. It is assumed that 20 percent of all spells end during the first period, such that at the beginning of the second period, 80 persons remain. The proportion of spells that end during period t is called the **exit rate** $\lambda(t)$, which is also referred to as the **hazard rate**. If the hazard rate stays constant at 20 percent, 16 spells will be terminated in the second period, 12.8 spells in the third, and so forth. We call the proportion of spells that have *not* been terminated before period t the **survivor function** $S(t)$. $f(t)$ is the (discrete) probability function for the duration of the spell. In the above example, 20 percent of all spells last for one period, 16 percent last for two periods, and so on. Since we assume a constant hazard rate, the probabilities of the spell duration decline according to a geometric distribution with $f(t) = p(1-p)^{t-1}$ where $\lambda(t) = p$ denotes the constant hazard rate.

Table 8.1. *An Artificial Data Example*

Period	Number at start of period $S(t) \times 100$	Exit rate $\lambda(t) \times 100$	Exits during period $S(t) \times \lambda(t) \times 100$
1	100	20	20
2	80	20	16
3	64	20	12.8
4	51.2	20	10.14
⋮	⋮	⋮	⋮

In empirical practice, hazard rates are rarely constant, and we want to infer the shape of the hazard function from the data. Moreover, it may be interesting to analyze the factors that determine the length of a spell. Consider the following example, 8.3, from the General Social Survey (GSS) dataset introduced in Chapter 1.5.1.

Example 8.3. Age at First Birth as a Duration

In 2002, the GSS sample contains information about a total of 1,517 women. For these women, we observe, among other things, the actual age and, if applicable, the age at first birth as **retrospective** information, which means that the information is obtained by asking each woman in 2002 about an event, here the first birth, that occurred in the past.

The variable *age at first birth* can be interpreted as a duration. For example, very few births are observed before the age of 15. Thus, if we take 15

years as a *de facto* starting point of the reproductive process for an average woman, then *age at first birth* minus 15 is the duration until first birth. For a 24-year-old first-time mother, for instance, this would be nine years. *Age at first birth* is an important variable in demography and population economics. For example, the first effect of increased access to higher education on female fertility decisions is commonly a postponement of the age at first birth. The effect on the total number of children a woman will eventually have is less clear-cut.

In the GSS data on age at first birth, right-censoring is a potential problem. Typically, such household survey data are collected as a random sample from the adult population. Hence, women of all ages are represented in proportion to the age distribution of the population. In particular, some women are beyond their physical child-bearing, or reproductive, period, say 40 or 45, while others are not. For the former group, we have all the information we require. Obviously, not all women have a child and therefore age at first birth is not defined. However, we can conduct an analysis of age at first birth conditional on having a child, for example by computing the expected age for different education groups, or running a regression. Unfortunately, though, we lose many observations if we look only at older women. Moreover, if there are important cohort effects, then the fertility behavior of younger women is more relevant for current and future fertility trends than the completed fertility pattern of previous cohorts. For both reasons, it may be desirable to include women under the age of 40 as well. This immediately gives rise to a censoring problem, as women under 40 who do not have a child may have one at a later age. For these women, the age at first birth is a censored variable, and excluding these observations from the analysis would lead to the same biases that were already discussed in Chapter 7.

In Table 8.1, we have seen how many people exit a state at each point in time – starting from a population of 100 – if the hazard rate is known. Usually, the procedure is the opposite: we observe exit times and want to draw inferences about the unknown underlying hazard rate. A complicating factor is that we typically do not observe the exit times for all spells, as some of them are right-censored. Continuing the fertility example, Table 8.2 shows a so-called **lifetable** for the variable *age at first birth* of women in the GSS 2002. Out of a total of 1,517 women in the dataset we keep 1,371 observations with age at first birth less than 40 years, of which 217 are censored and 1,154 are non-censored.

The hazard rate is now simply the ratio of the number of births divided by the number at risk for a given age group, or time interval. Formally,

$$\lambda(t) = \frac{n(t)}{r(t)} \tag{8.1}$$

Table 8.2. *Lifetable of Age at First Birth Among Women in the GSS*

Age at first birth	At risk	Exit/Birth	Censored	Hazard Rate $\lambda(t)$	Survivor Function $\hat{S}(t)$
11	1,371	1	0	0.0007	0.9993
13	1,370	3	0	0.0022	0.9971
14	1,367	2	0	0.0015	0.9956
15	1,365	14	0	0.0103	0.9854
16	1,351	40	0	0.0296	0.9562
17	1,311	66	0	0.0503	0.9081
18	1,245	97	0	0.0779	0.8373
19	1,148	106	12	0.0923	0.7600
20	1,030	122	10	0.1184	0.6700
21	898	123	8	0.1370	0.5782
22	767	97	16	0.1265	0.5051
23	654	72	19	0.1101	0.4495
24	563	61	18	0.1083	0.4008
25	484	60	17	0.1240	0.3511
26	407	45	11	0.1106	0.3123
27	351	45	12	0.1282	0.2723
28	294	37	11	0.1259	0.2380
29	246	32	5	0.1301	0.2070
30	209	38	8	0.1818	0.1694
31	163	18	10	0.1104	0.1507
32	135	19	8	0.1407	0.1295
33	108	17	8	0.1574	0.1091
34	83	7	11	0.0843	0.0999
35	65	13	7	0.2000	0.0799
36	45	7	9	0.1556	0.0675
37	29	7	5	0.2414	0.0512
38	17	3	2	0.1765	0.0422
39	12	2	10	0.1667	0.0351

where $n(t)$ denotes the number of exits, and $r(t)$ denotes the number at risk in t. The number at risk is affected by two types of events, one being a birth at age t, and the other being a censoring at age t. For example, the risk set at age 20 includes 1,030 women. 122 of them give first birth, while 10 have not given first birth but are not observed any further. Therefore, the risk set in the next period, at age 21, is $1,030 - 122 - 10 = 898$. The survivor function (in the case of censoring) may be estimated by the **Kaplan-Meier** (or product-limit) estimator. Formally,

$$\hat{S}(t) = \prod_{s<t} [1 - \lambda(s)] \tag{8.2}$$

The **empirical survivor function** $\hat{S}(t)$ is the estimated proportion of survivors at age t. For example, at age 11 it is simply one minus the exit rate at age 11, at age 13 the proportion $\hat{S}(13)$ is estimated as $\hat{S}(11)$ times one minus

the exit rate at age 13, and so forth. Both the hazard rate and the empirical survivor function can be illustrated graphically as in Figures 8.2 and 8.3.

Fig. 8.2. *Empirical Hazard Rates of Age at First Birth*

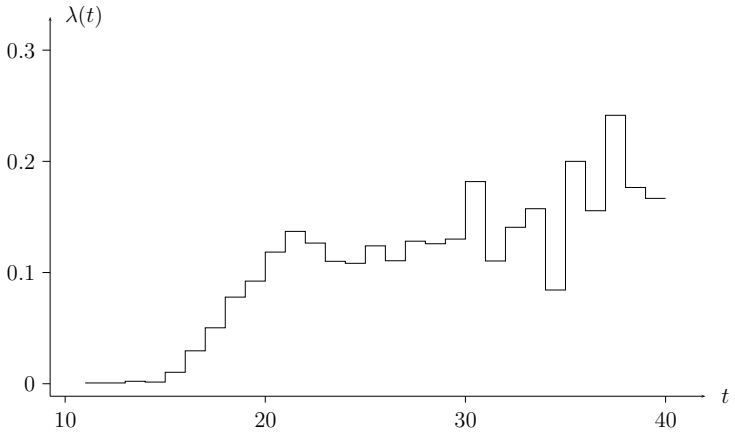
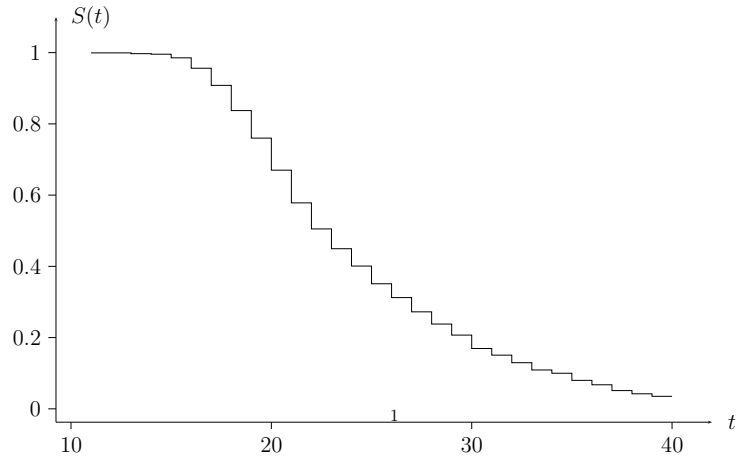


Fig. 8.3. *Kaplan-Meier Estimator of the Survivor Function*



The Kaplan-Meier estimator presumes that the number at risk at time t is simply the number at risk at the beginning of t, and that censored observations drop out of the sample at the end of t. This assumption is reasonable as long as

we consider intrinsically discrete data. However, with grouped duration data it may be more convenient to take into account that individuals will leave *during* the time interval, and not at the beginning or the end. The **lifetable method** adjusts the number at risk in each time interval by producing an estimate of the average number at risk at the midpoint of the interval. For example, if we assume that individuals leave evenly distributed over the time interval, we can formulate an adjusted number at risk for the time interval t as

$$\tilde{r}(t) = r(t) - \frac{m(t)}{2} \tag{8.3}$$

where $m(t)$ is the number of censored observations in t. The within interval exit rate is defined as $\tilde{\lambda}(t) = n(t)/\tilde{r}(t)$, and an estimator of the survivor function can be formulated similar to (8.2) using the adjusted exit rate. As long as censoring does not occur, the lifetable and Kaplan-Meier estimates are the same. However, the more observations are censored, the higher the adjusted exit rate and thus, the more the proportion of survivors is overestimated by the Kaplan-Meier estimator.

8.2.3 Discrete Time Duration Models

Discrete time duration models can be estimated by standard logit or probit routines. The key element is the **discrete time hazard**, i.e., the probability of exit during period t conditional on being at risk in t. Let $\mathcal{T}$ denote the discrete survival time with outcomes $t \in I\!N = \{1, 2, 3, \ldots\}$ which can be days, months, years, or simply an index of periods. We use the symbol $\mathcal{T}$ for the dependent variable, emphasizing that we are modeling time. The **hazard function** of $\mathcal{T}$ can be written as

$$\lambda_{it} = \lambda(t_i|x_{it}) = P(\mathcal{T}_i = t_i|\mathcal{T}_i \geq t_i, x_{it}) \tag{8.4}$$

where x_{it} denotes a vector of regressors, potentially varying over time. The question is how λ_{it} can be specified as a function of x_{it}, or more precisely, as a function of the linear index $x'_{it}\beta$. Since λ_{it} is a probability, we require a functional form that maps $x'_{it}\beta$ onto the unit interval. From Chapter 4, we know two such transformations: logit and probit. In the logit case, we let

$$\log\left(\frac{\lambda_{it}}{1-\lambda_{it}}\right) = \alpha_t + x'_{it}\beta$$

In the probit case, we would formulate the model as $\lambda_{it} = \Phi(\alpha_t + x'_{it}\beta)$. The term α_t is a time-varying constant that captures **duration dependence**, i.e., how the hazard rates vary over time for constant regressors. This could be a simple linear trend $\alpha_t = \delta_1 t$, a logarithmic form with $\alpha_t = \delta_1 \log(t)$, or alternatively a flexible dummy specification with $\alpha_t = \delta_2 d_{t=2} + \ldots + \delta_{t_{max}} d_{t=t_{max}}$, given that observed exit times are 1, 2 until t_{max}, and $d_{t=2}$ indicates whether

$t = 2$, and so on. For estimation, we need to create a separate record for each point of time at which an individual is known to be at risk. The dependent variable takes the value 1 if the event occurs (this is the last record because after the event, the person is no longer at risk) and 0 otherwise. Censored observations are coded as 0 until they leave the sample. To see how the dataset needs to be organized, consider the following example, 8.4.

Example 8.4. Early Retirement

Suppose we are interested in modeling the retirement decision. Under what conditions do men choose early retirement? Specifically, what is the influence of health on this decision? Assume that health can be measured with a simple indicator, where $health_t = 1$ means "good health" and $health_t = 0$ means "no good health". Table 8.3 shows a hypothetical sample of 10 men. Their retirement age varies between 62 and 65 years. Of the 10 men, three retire at the age of 62, one at the age of 63, one at the age of 64, three at the official retirement age of 65, and two observations are censored. These are duration data since we can consider the variable "time to retirement", where $t = 1$ means retirement at age 62, $t = 2$ means retirement at age 63 and so forth. In this situation, $health_{62}$, for example, measures the health status at age 62, which may differ from the health status at age 63, $health_{63}$, a case of a time-varying explanatory variable.

Table 8.3. *Retirement Age and Health: Fictitious Data*

i	*Age*	*Censored*	$health_{62}$	$health_{63}$	$health_{64}$
1	62	0	0	1	1
2	65	0	1	1	1
3	63	0	1	1	0
4	62	1	1	1	1
5	64	0	0	0	0
6	65	0	1	1	1
7	62	0	0	0	0
8	62	0	1	0	1
9	63	1	0	1	1
10	65	0	1	1	0

In order to estimate a discrete time hazard rate model, the data needs to be reorganized as indicated above. At age 62, 10 persons are at risk, three of whom exit (retire) at that age with one observation censored. Hence, only the remaining six observations are relevant at the next stage, when modeling whether to retire at age 63 or not. All in all, there are 20 relevant observations for estimation, as seen in Table 8.4. Assuming that all men have to retire at the age of 65, the hazard rate in the last period is one and cannot be

estimated. The dependent variable is *Exit*, the explanatory variables include the two dummy variables d_{63} and d_{64} to model the non-constant hazard, and the variable *health*.

Table 8.4. *Retirement Age and Health in Binary Data Format*

i	*Age*	*Exit*	d_{63}	d_{64}	*health*
1	62	1	0	0	0
2	62	0	0	0	1
2	63	0	1	0	1
2	64	0	0	1	1
3	62	0	0	0	1
3	63	1	1	0	1
4	62	0	0	0	1
5	62	0	0	0	0
5	63	0	1	0	0
5	64	1	0	1	0
6	62	0	0	0	1
6	63	0	1	0	1
6	64	0	0	1	1
7	62	1	0	0	0
8	62	1	0	0	1
9	62	0	0	0	0
9	63	0	1	0	1
10	62	0	0	0	1
10	63	0	1	0	1
10	64	0	0	1	0

If we use a logit model to estimate the parameters of the model, we obtain the following result:

$$\log\left(\frac{\hat{\lambda}_{it}}{1-\hat{\lambda}_{it}}\right) = -0.107 - 0.443\, d_{63} - 0.422\, d_{64} - 1.379\, health_i$$

Thus, the hazard rate declines over time. Moreover, healthy persons are less likely to retire early than unhealthy ones.

In principle, estimation of discrete time hazard models is straightforward (see Jenkins, 1995). However, there are a number of problems as well. First, the resulting pooled sample may have a large dimension. In the above example, the increase was from 10 person records to 20 person-year records. In the GSS example on age at first birth, this procedure would lead to a total of 32,168 person-years (starting from 1,371 observations). Thus, the approach becomes impracticable if the number of exit points is large. The finer the time

measurement, the more observations there will be. This is one of the reasons why continuous time models are more suitable in many situations. The other reason is that with discrete time hazard models, we can only predict hazard rates for times at which exits are observed. Consequently, it is not possible either to draw inferences on average durations directly. All these concerns are overcome if we use continuous time duration models.

8.2.4 Continuous Time Duration Models

For all conditional probability models discussed so far, we typically started with a parametric distribution function, parameterized the model in terms of explanatory variables, and then estimated the parameters by maximum likelihood. We could choose the same approach for duration data. Since durations are non-negative, we should start with a continuous distribution $f(t)$ with support over $\mathbb{R}^+$, such as the log-normal distribution or the exponential distribution. More specifically, the conditional model we are thinking of has the form $f(t_i|x_i)$. This imposes some constraints on the type of explanatory variables we can employ. In particular, x_i has no time subscript, i.e., it should be unrelated to time and length of the duration, and is therefore called **time-invariant**. This refers to a situation in which, for example, the regressors are measured at the beginning of a spell. Relaxing this restriction is relatively easy in discrete time duration models, but more tricky in continuous time models and requires so-called episode splitting (see Collett, 2003: Chapter 8).

Example 8.5. The Exponential Model

The conditional probability approach can be exemplified with the exponential distribution. We have $f(t;\lambda) = \lambda \exp(-\lambda t)$ and $\mathrm{E}(t) = \lambda^{-1}$. Therefore, we may parameterize $\mathrm{E}(t_i|x_i) = \lambda_i^{-1}$ with $\lambda_i = \lambda(t_i|x_i) = \exp(x_i'\beta)$, which implies the conditional probability model

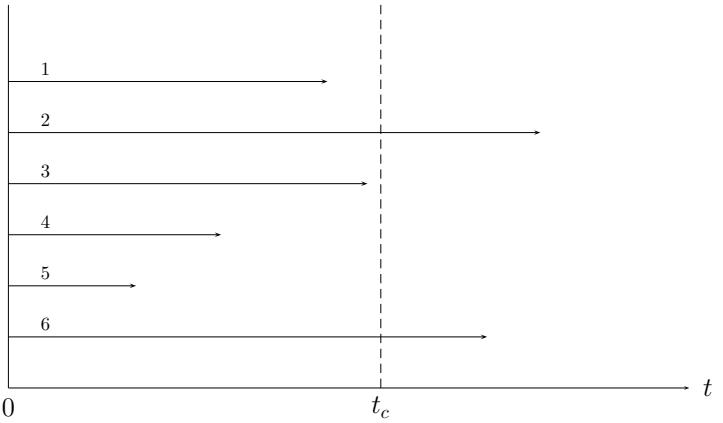
$$f(t_i|x_i;\beta) = \exp(x_i'\beta)\exp[-\exp(x_i'\beta)t_i] \tag{8.5}$$

Next, we show how to estimate β by ML when we have a sample of spells, some of which may be ongoing and hence be right-censored.

Maximum likelihood estimation of duration models can involve quite elaborate procedures since, depending on the way the data are collected, non-random sampling and partial observability are frequent problems. A major distinction is that between stock and flow sampling. **Stock sampling** refers to the analysis of spells that are ongoing at a certain point of time. Flow sampling refers to the analysis of spells that begin at a certain point in time (or during a certain interval of time). For example, consider the duration of

unemployment, i.e., the time until an unemployed person is employed again. In this case, stock sampling would mean drawing a sample of all unemployed individuals, say on the first of January of a given year, and then seeing how long it takes for these individuals to become reemployed. Two issues in such samples are left-censoring, where it may be unknown to the analyst when the unemployment spell actually started, and length-biased sampling, where long unemployment spells will be overrepresented in the sample. While techniques have been developed to deal with these situations, here we restrict our attention to the second type of sampling, **flow sampling**, as it occurs when data are taken from inflow records into the unemployment register during a given month. In the case of flow sampling, the only issue is right censoring, i.e., that some spells may not have ended before the observation period ended.

Fig. 8.4. *Flow Sampling With Censoring*



Flow sampling is depicted in Figure 8.4. Six spells start at time $t = 0$. This could be any real time date, such as the first of January, 2004. The observation window lasts until t_c. For example, t_c can be the present time. Clearly, for all spells that have not ended before the present time, it is impossible to know – at present – how long they will ultimately last. All that can be said about such spells is that they lasted for a duration of at least t_c. In Figure 8.4, the second and the sixth spell are censored, whereas all other spells are completed. Formally, we can write for the observed duration that

$$t_i = \min(t_i^*, t_c) \tag{8.6}$$

where t_i^* is the (potentially unobserved) true duration. If the censoring time varies across individuals, simply substitute t_{ci} for t_c in equation (8.6). The probability that t_i is censored can be written as

$$P(t_i^* > t_c|x_i) = 1 - F(t_c|x_i;\beta) \tag{8.7}$$

For noncensored observations, we know the density $f(t_i|x_i;\beta)$. Define a censoring indicator $d_i = 1$ if the observation is uncensored, and $d_i = 0$ if the observation is censored. Then, the conditional probability function for observation i can be written as

$$f(t_i|x_i;\beta)^{d_i}[1 - F(t_i|x_i;\beta)]^{1-d_i}$$

from which we obtain the log-likelihood function, assuming a sample of n independent pairs of observations (t_i, x_i)

$$\log L(\beta;t,x) = \sum_{i=1}^{n} d_i \log f(t_i|x_i;\beta) + (1-d_i)\log[1 - F(t_i|x_i;\beta)] \tag{8.8}$$

Example 8.6. ML Estimation of the Exponential Model

Suppose that t_i is exponentially distributed with $\lambda_i = \exp(x_i'\beta)$. In this case

$$\begin{aligned}\log L(\beta;t,x) &= \sum_{i=1}^{n} d_i[x_i'\beta - \exp(x_i'\beta)t_i] - (1-d_i)\exp(x_i'\beta)t_i \\ &= \sum_{i=1}^{n} d_i x_i'\beta - \exp(x_i'\beta)t_i \end{aligned} \tag{8.9}$$

The log-likelihood function is non-linear in the parameters, so that iterative methods are required to obtain the maximum likelihood estimator. In the special case where x_i contains a constant only, we obtain

$$\log L(\beta;t) = \sum_{i=1}^{n} \beta d_i - \exp(\beta)t_i$$

from which it follows that

$$s(\beta;t) = \sum_{i=1}^{n} d_i - \exp(\beta)\sum_{i=1}^{n} t_i$$

and therefore the maximum likelihood estimator can simply be written as

$$\hat{\beta} = \log\left(\frac{n_{nc}}{\sum_{i=1}^{n} t_i}\right) = \log\left(\frac{n_{nc}/n}{\bar{t}}\right)$$

where n_{nc} is the number of non-censored observations. In this case, $\exp(\hat{\beta})$ is the inverse of the ratio of the average exit time (of censored and uncensored observations) divided by the fraction of non-censored observations in the sample. Similarly, the maximum likelihood estimator for the expected duration is the average exit time divided by the fraction of non-censored observations. If all observations are non-censored, this is the sample mean, as it should be.

8.2.5 Key Element: Hazard Function

The exponential model is easy to formulate and to estimate. Nevertheless, it is rarely used in practice. In order to understand why the exponential model is overly restrictive and therefore of limited use in most economic applications, a good place to start looking is the hazard function. Hazard functions were already introduced in the context of discrete time duration models. Here, we provide the corresponding results for continuous time duration models.

It is important to understand that one can build a duration model directly around the hazard function, $\lambda_i = \lambda(t_i|x_i)$, an entity that obviously provides more information than the conditional expectation function, as it shows the conditional exit rates for all values of t. Indeed, the hazard function uniquely determines the density function, and therefore the hazard rate fully characterizes the conditional duration model. Let $\mathcal{T} \geq 0$ denote the continuous duration of a spell. The cumulative density function of $\mathcal{T}$ is defined as

$$F(t) = P(\mathcal{T} \leq t) \qquad t \geq 0 \tag{8.10}$$

The **survivor function** shows the probability of surviving past t, which is defined as the complement of the cumulative density

$$S(t) = 1 - F(t) = P(\mathcal{T} > t) \tag{8.11}$$

Due to the relationship between survivor function and cumulative density, the latter is sometimes referred to as **failure function** $F(t)$. The density of $\mathcal{T}$ is the first derivative of the cumulative density function with respect to t

$$f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt} \tag{8.12}$$

For any $h > 0$, the probability of exiting the spell during the interval $[t, t+h)$, given survival up to time t, can be written as $P(t \leq \mathcal{T} < t + h|\mathcal{T} \geq t)$. The **hazard function** for $\mathcal{T}$ is defined as

$$\lambda(t) = \lim_{h \to 0} \frac{P(t \leq \mathcal{T} < t + h|\mathcal{T} \geq t)}{h} \tag{8.13}$$

Hence, the hazard function is the limit of the probability that the spell is completed during the interval $[t, t + h)$, given that it has not been completed before time t, for the limit $h \to 0$, and thus, it is an "instantaneous exit rate" per unit of time (rather than a proper probability). We can express the hazard function in terms of density and cumulative density. First

$$P(t \leq \mathcal{T} < t + h|\mathcal{T} \geq t) = \frac{P(t \leq \mathcal{T} < t + h)}{P(\mathcal{T} \geq t)} = \frac{F(t + h) - F(t)}{1 - F(t)}$$

and therefore,

$$\lambda(t) = \lim_{h \to 0} \frac{F(t + h) - F(t)}{h} \frac{1}{1 - F(t)} = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)} \tag{8.14}$$

Because the derivative of $S(t)$ is $-f(t)$ it follows that

$$\lambda(t) = -\frac{d \log S(t)}{dt} \tag{8.15}$$

By integrating both sides, we get the **integrated hazard function**

$$\begin{aligned} H(t) &= \int_0^t \lambda(s)ds = \int_0^t -\frac{d \log S(s)}{ds} ds = -\log S(t) + \log S(0) \\ &= -\log S(t) \end{aligned} \tag{8.16}$$

and therefore,

$$F(t) = 1 - \exp\left[-\int_0^t \lambda(s)ds\right] \tag{8.17}$$

Differentiation of (8.17) gives the density of $\mathcal{T}$ as a function of the hazard rate, and we have

$$f(t) = \lambda(t) \exp\left[-\int_0^t \lambda(s)ds\right] \tag{8.18}$$

Therefore, all probabilities can be computed as a function of the hazard rate with integrating over $f(t)$ or directly using $F(t)$. Moreover, once we have specified a functional form for the hazard rate, we can also derive the survivor function $S(t)$ and the integrated hazard function $H(t)$.

Exercise 8.1.

- Suppose that you have $a_1 < a_2$. Derive $P(\mathcal{T} \geq a_2 | \mathcal{T} \geq a_1)$ and $P(a_1 \leq \mathcal{T} < a_2 | \mathcal{T} \geq a_1)$ in terms of $\lambda(t)$.

Example 8.7. Constant Hazard Function

The simplest duration model is one with a constant hazard function

$$\lambda(t) = \lambda$$

From (8.17) we obtain

$$F(t) = 1 - \exp\left[-\int_0^t \lambda ds\right] = 1 - \exp(-\lambda t)$$

which is the cumulative density of the exponential distribution. Conversely, if $\mathcal{T}$ has an exponential distribution, its hazard is constant. An exponentially

distributed random variable with hazard λ has expected value $1/\lambda$. The hazard function must be nonnegative - hence it is natural to model it as an exponential function, namely

$$\lambda(t_i|x_i) = \exp(x_i'\beta) \tag{8.19}$$

Alternatively, the hazard function can be written as $\log \lambda(t_i|x_i) = x_i'\beta$. With these considerations, we may now understand why the hazard function of the exponential model is too restrictive to be useful in practice. The two main drawbacks discussed in the literature are as follows. First, the hazard function is assumed to be constant over, and thus independent of, time. Second, the hazard is fully known, once the regressors are given and the parameters are estimated. There is no room for additional unobserved heterogeneity. The more general duration models used in practice address one of these concerns, or both, and we will discuss these issues next.

8.2.6 Duration Dependence

When the hazard is not constant over time, we say that the process exhibits **duration dependence**. With $d\lambda(t)/dt > 0$ the duration dependence is positive at time t. If $d\lambda(t)/dt > 0$ for all t, then we have monotonous positive duration dependence. The probability of leaving the spell increases the longer one has been in the spell. Of course, we may also have monotonous negative duration dependence, or a hazard function that is non-monotonous in t.

A first possibility to model monotonous duration dependence is obtained by formulating the logarithmic hazards as a function of regressors *and* logarithmic survival times, formally

$$\log \lambda(t_i|x_i) = x_i'\beta + \delta \log t_i \tag{8.20}$$

With this assumption, the hazard function can be written as $\lambda(t_i|x_i) = \exp(x_i'\beta)t_i^{\delta}$ and the cumulative density function is given by

$$\begin{aligned} F(t_i|x_i) &= 1 - \exp\left[-\int_0^{t_i} \exp(x_i'\beta)s^{\delta} ds\right] \\ &= 1 - \exp\left(-\frac{\exp(x_i'\beta)}{\delta+1} t_i^{\delta+1}\right) \end{aligned} \tag{8.21}$$

If we define $\alpha = \delta + 1$ and $\lambda_i = \exp(x_i'\beta)/\alpha$, then (8.22) simplifies to

$$F(t_i|x_i) = 1 - \exp\left(-\lambda_i t_i^{\alpha}\right) \tag{8.22}$$

and

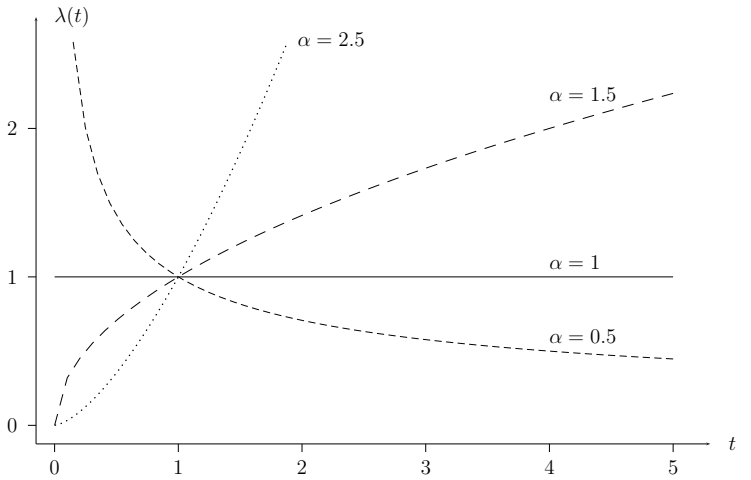
$$f(t_i|x_i) = \lambda_i \alpha t_i^{\alpha-1} \exp\left(-\lambda_i t_i^{\alpha}\right) \tag{8.23}$$

This is the density function of the **Weibull** distribution. In order to ensure a proper density function, we need to impose the restriction $\alpha = \delta + 1 > 0$. In the case of $\alpha = 1$, the Weibull distribution reduces to the exponential distribution with $\lambda_i = \exp(x_i'\beta)$. The hazard function of the Weibull distribution with parameter α can be written as

$$\lambda(t_i|x_i) = \frac{f(t_i|x_i)}{1 - F(t_i|x_i)} = \lambda_i \alpha t_i^{\alpha-1} = \exp(x_i'\beta)t^{\alpha-1} \tag{8.24}$$

which can also be obtained directly from (8.20) by simply exponentiating and replacing δ by $\alpha - 1$. The shape of these hazard functions for different values of α is illustrated in Figure 8.5. If $\alpha > 1$ (or $\delta > 0$), the hazard is monotonically increasing and we have positive duration dependence. For $\alpha < 1$ the hazard rate is monotonically decreasing (negative duration dependence). The Weibull distribution therefore offers a simple way to capture monotonous duration dependence.

Fig. 8.5. *Weibull Hazard Functions*



Other examples of parametric models with non-constant hazard functions include

- the **Gompertz distribution**, where

$$\lambda(t) = \lambda \exp(\gamma t) \qquad \lambda > 0 \tag{8.25}$$

- the **log-logistic distribution**, where

$$\lambda(t) = \frac{\psi^{\frac{1}{\gamma}} t^{\frac{1}{\gamma}-1}}{\gamma\left[1+(\psi t)^{\frac{1}{\gamma}}\right]} \qquad \psi > 0,\ \gamma > 0 \tag{8.26}$$

- the **log-normal distribution**, where

$$\lambda(t) = \frac{\frac{1}{t\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\log(t)-\mu}{\sigma}\right)^2\right]}{1-\Phi\left(\frac{\log(t)-\mu}{\sigma}\right)} \qquad \sigma > 0 \tag{8.27}$$

The Gompertz model differs from the Weibull model in that logarithmic hazards are assumed to be a *linear* function of survival times, and we may parameterize the model as $\lambda_i = \exp(x_i'\beta)$. With positive (negative) shape parameter γ, the hazard is monotonically increasing (decreasing) over time, and with $\gamma = 0$ the Gompertz model reduces to the exponential model with constant hazard. The log-logistic model is usually parameterized as $\psi_i = \exp(-x_i'\beta)$ (the reason for the negative sign will become clear below) and, depending on the shape parameter γ, the hazard rate is monotonically decreasing ($\gamma \geq 1$), or first increasing and then decreasing ($0 < \gamma < 1$). The shape of the hazard function in the log-normal model is very similar to the log-logistic model with $0 < \gamma < 1$, and one usually assumes that the mean parameter is specified as linear index of regressors, $\mu_i = x_i'\beta$.

Proportional Hazards Models

The exponential model, the Weibull model, and the Gompertz model have in common that they can be formulated as a **proportional hazards (PH) model**. In general, a PH model can be written as

$$\lambda(t_i|x_i) = \kappa(x_i)\lambda_0(t_i) \tag{8.28}$$

where $\kappa(\cdot)$ is a nonnegative function of x_i and $\lambda_0(t_i) > 0$ is the so-called **baseline hazard**. The baseline hazard, and therefore the nature of duration dependence, is common to all individuals. The factor $\kappa(x_i)$ then determines the position of the individual hazard rate by shifting the baseline hazard proportionally up or down. For example, in the Weibull model we have $\kappa(x_i) = \exp(x_i'\beta)$ and $\lambda_0(t_i) = t_i^{\delta}$. In particular, if we assume that $\kappa(x_i) = \exp(x_i'\beta)$, then we can interpret β in terms of **relative hazards** for a given change in the l-th element in x_i,

$$\frac{\lambda(t_i|x_i + \Delta x_{il})}{\lambda(t_i|x_i)} = \exp(\Delta x_{il}\beta_l) \tag{8.29}$$

For a unit change in x_{il}, $\Delta x_{il} = 1$, we obtain the factor change $\exp(\beta_l)$ in the relative hazards, which is also referred to as **hazard ratio**.

Accelerated Failure Time Models

The log-logistic and the log-normal hazard functions cannot be classified as PH models. Alternatively, both models may be summarized in the class of **accelerated failure time (AFT) models**, for which logarithmic survival times are expressed as a linear index of covariates and an additive error term,

$$\log t_i = x_i'\beta + u_i \tag{8.30}$$

A distributional assumption about u_i determines the survival time distribution, and thereby the functional form of the hazard function. For example, if error terms are normally distributed then we obtain the log-normal model, and if the error terms follow a logistic distribution then we get the log-logistic model. Moreover, it turns out that the Weibull model (and therefore also the exponential model) as the only PH models can be formulated as AFT models, if we assume extreme value distributed error terms (see, for example, Cox and Oakes, 1971: p. 71). In order to see how AFT models accelerate the failure time, rewrite (8.30) as

$$\log \psi_i t_i = u_i \tag{8.31}$$

where we define $\psi_i = \exp(-x_i'\beta)$, as in the log-logistic model. The term ψ_i now changes the time scaling in the sense that if $\psi_i > 1$, then survival times are shortened (or failure is accelerated), and if $\psi_i < 1$, then survival times are lengthened. One way to interpret an AFT model is to consider the model formulation in (8.30) directly. Here, a unit increase in the l-th regressor x_{il} changes the expected survival time by approximately $\beta_l \times 100$ percent, or by exactly $[\exp(\beta_l) - 1] \times 100$ percent. The term $\exp(\beta_l)$ is sometimes referred to as a **time ratio**.

While these parametric models are simple to estimate and to interpret, they may be overly restrictive and, for this reason, misspecified. Therefore, more flexible models have been proposed, the leading example being the **Cox proportional hazards model** (Cox, 1972, 1975). The regression parameters of this model can be estimated by **partial likelihood** methods without even specifying a functional form for the hazard function (see also Kalbfleisch and Prentice, 2002, or Klein and Moeschberger, 2003). Of course, this is not very interesting if learning something about the shape of the hazard function is a main concern of the analysis, as is often the case in econometric applications. In such cases, the exponential model with a **piecewise linear hazard function** has become increasingly popular (see Lancaster, 1990, for further details).

Duration dependence is an issue of substantive economic interest in many applications. This is one possible way to capture the dynamics of the underlying exit process. For example, in the analysis of unemployment duration, negative duration dependence implies that the probability of finding a job

decreases as the length of the unemployment spell increases. Reason for such a decline can be human capital depreciation or stigma effects. In any case, this will tend to generate a substantial fraction of long-term unemployment.

8.2.7 Unobserved Heterogeneity

Unfortunately, it is empirically difficult to separate the effects of duration dependence from the effects of **unobserved heterogeneity**, also referred to as **frailty**. One may find evidence for "spurious" duration dependence even in a true world with constant hazard, as long as there is unobserved heterogeneity. To understand the effect of unobserved heterogeneity in duration models, we consider the example of the exponential duration model. Notice that in (8.19) the hazard is fully specified once we know x_i and β. Alternatively, we could formulate the hazard function in log-form with an additive error term ε_i as

$$\log \lambda(t_i|x_i, \varepsilon_i) = x_i'\beta + \varepsilon_i \tag{8.32}$$

Now, there are two sources of heterogeneity, i.e., differences in hazard rates between observation units, one being observed and one being unobserved. In (8.32), the random error ε_i captures the unobserved heterogeneity. For example, such unobserved heterogeneity can arise if there are additional variables that affect the hazard but are unobserved by the econometrician. Because we can rephrase (8.32) as

$$\lambda(t_i|x_i, u_i) = \exp(x_i'\beta)u_i \tag{8.33}$$

where $u_i = \exp(\varepsilon_i)$, we refer to this model as one with **multiplicative heterogeneity**. Without loss of generality, we let $\mathrm{E}(u_i|x_i) = 1$ such that

$$\mathrm{E}[\lambda(t_i|x_i, u_i)|x_i] = \exp(x_i'\beta)\mathrm{E}(u_i|x_i) = \exp(x_i'\beta) \tag{8.34}$$

We know that conditional on u_i, the duration is exponentially distributed with

$$F(t_i|x_i, u_i) = 1 - \exp(-\exp(x_i'\beta)u_i t_i) \tag{8.35}$$

Since we do not observe u_i, we have to reformulate the model in terms of what we can observe, and the technique for doing so is to **marginalize** (8.35) with respect to u_i, or in other words, **to integrate out** the unobservables. Let $g(u_i|x_i)$ denote the density function of u_i given x_i, with support over the positive real line. Then we obtain the marginal distribution function by integrating the product of $F(t_i|x_i, u_i)$ and $g(u_i|x_i)$ over u_i:

$$F(t_i|x_i) = \int_0^\infty F(t_i|x_i, u_i)g(u_i|x_i)du_i \tag{8.36}$$

In parametric models of unobserved heterogeneity, we have to specify a suitable density function for $g(u_i|x_i)$. It turns out that the gamma distribution

is a suitable candidate, since it leads to relatively simple derivations and closed-form solutions. In semi-parametric models, we avoid such a parametric assumption, and instead approximate the density of u_i with discrete mass-points. Here, we focus on the parametric case. If u_i is **gamma distributed** with parameters $\theta > 0$ and $\gamma > 0$ the density function can be written as

$$g(u_i|x_i) = \frac{\gamma^\theta}{\Gamma(\theta)} u_i^{\theta-1} e^{-\gamma u_i} \tag{8.37}$$

where $\Gamma(\theta)$ is the **gamma function** defined as $\int_0^\infty z^{\theta-1}e^{-z}dz$. In order to impose the normalization $\mathrm{E}(u_i|x_i) = \theta/\gamma = 1$, we let $\theta = \gamma$. Therefore,

$$\begin{aligned} F(t_i|x_i) &= \int_0^\infty \left[1 - \exp(-\exp(x_i'\beta)u_i t_i)\right] \frac{\theta^\theta}{\Gamma(\theta)} u_i^{\theta-1} e^{-\theta u_i} du_i \\ &= 1 - \int_0^\infty \exp(-\exp(x_i'\beta)u_i t_i) \frac{\theta^\theta}{\Gamma(\theta)} u_i^{\theta-1} e^{-\theta u_i} du_i \\ &= 1 - \frac{\theta^\theta}{\Gamma(\theta)} \int_0^\infty u_i^{\theta-1} e^{-(\theta+\exp(x_i'\beta)t_i)u_i} du_i \\ &= 1 - \left(\frac{\theta}{\theta + \exp(x_i'\beta)t_i}\right)^\theta \end{aligned} \tag{8.38}$$

where the last equality follows because the integrand can be expanded such that it equals the density function of a $Gamma(\theta, \theta+\exp(x_i'\beta)t_i)$ distribution, and hence must integrate to unity. Therefore, we obtain

$$F(t_i|x_i) = 1 - \left[1 + \theta^{-1}\exp(x_i'\beta)t_i\right]^{-\theta} \tag{8.39}$$

and

$$f(t_i|x_i) = \exp(x_i'\beta)\left[1 + \theta^{-1}\exp(x_i'\beta)t_i\right]^{-\theta-1} \tag{8.40}$$

from which it follows that the hazard rate in the exponential model with gamma distributed unobserved heterogeneity can be written as

$$\lambda(t_i|x_i) = \frac{f(t_i|x_i)}{1 - F(t_i|x_i)} = \exp(x_i'\beta)\left[1 + \theta^{-1}\exp(x_i'\beta)t_i\right]^{-1} \tag{8.41}$$

For $\theta^{-1} \to 0$, we obtain the hazard function of the simple exponential model. Note that θ^{-1} is the variance of the gamma distributed heterogeneity term u_i – hence in this degenerate case, there is no heterogeneity. For $\theta^{-1} > 0$ the hazard rate becomes a decreasing function of t_i, which was to be shown. This is an example of "spurious" duration dependence.

Example 8.8. Unobserved Heterogeneity across Groups

The nature of unobserved heterogeneity and negative duration dependence (declining hazard function over time) can also be depicted in a more elementary way by returning to a discrete example. Assume that the population is composed of two groups. The first group has a hazard rate of 0.1, the second group has a hazard rate of 0.5. Therefore, the hazard rate is constant in each group, but there is heterogeneity. Now assume that the heterogeneity is unobserved, and ignored by the analyst. Also assume that the two groups possess equal weight in the population. For example, if we start out with 100 persons from group 1 and 100 persons from group 2, there are 140 survivors after one period. The initial hazard rate is 60/200=0.30, or 30 percent. This is shown in Table 8.5.

Table 8.5. *Unobserved Heterogeneity and Spurious Duration Dependence*

	Survivors			Hazard Rate
Period	Group 1	Group 2	Total	Total
1	100	100	200	0.30
2	90	50	140	0.24
3	81	25	106	0.19
4	72.9	12.5	85.4	
⋮	⋮	⋮	⋮	⋮

In the next period, the composition of the pool of survivors has changed. Among survivors, there are 90 persons, or 64 percent, of type "low hazard". The remaining 50 persons are of type "high hazard". The initial split was 50/50. Hence, what will happen is that in the next period, the *average* hazard rate is lower, since the low hazard types are now overrepresented. This is documented in the last column of Table 8.5. The average hazard drops from 30 percent to 24 percent, and then to 19 percent in the next period, when the high hazard types have been reduced to 15 percent of the surviving population. The important point is the following: if we think that we are confronted with a homogeneous population, the empirical evidence will look as if there is negative duration dependence. The hazard rate falls over time. But this is spurious. The falling hazard rate has no behavioral interpretation; rather, it is a direct consequence of heterogeneity that is unaccounted for, and the resulting compositional changes.

To return to our previous example of the duration of unemployment spells, what happens is this: individuals with high hazard rates exit unemployment early and then are no longer part of the risk set. As time goes on, this selection process yields risk sets that contain individuals with lower and lower (albeit individually constant) exit rates, and the observed hazard falls over time. It is extremely difficult to distinguish hazard rates that are truly declining over time from simple variation in hazard rates across individuals.

In a regression context, whether unobserved heterogeneity should be a major concern or not, depends on the questions one wants to answer. If the main goal of the analysis is consistent estimation of β, then unobserved heterogeneity does not necessarily invalidate the estimators, as long as ε_i and x_i are independent. If the goal is to identify duration dependence, it is imperative to properly account for unobserved heterogeneity.

Duration Dependence and Unobserved Heterogeneity

Parametric models that account for both duration dependence *and* unobserved heterogeneity can be formulated as well. Our starting point is a generalized form of the hazard function in (8.33)

$$\lambda(t_i|x_i, u_i) = \tilde{\lambda}(t_i|x_i)u_i \tag{8.42}$$

with the hazard $\tilde{\lambda}(t_i|x_i)$ depending on observable characteristics, possibly allowing for duration dependence, and the multiplicative error term u_i capturing unobserved heterogeneity. From the relationship between the hazard function, the integrated hazard function, and the cumulative distribution of the survival time, it follows that

$$F(t_i|x_i, u_i) = \exp\left[-\int_0^{t_i} \tilde{\lambda}(s_i|x_i)u_i ds_i\right] = 1 - \exp\left[-\tilde{H}(t_i|x_i)u_i\right] \tag{8.43}$$

As before, we assume that unobserved heterogeneity is distributed according to a gamma distribution with unit expectation, and we need to integrate out the error term u_i. It can be shown that

$$\begin{aligned} F(t_i|x_i) &= \int_0^{\infty} [1 - \exp\left[-\tilde{H}(t_i|x_i)u_i\right] \frac{\theta^{\theta}}{\Gamma(\theta)} u_i^{\theta-1} e^{-\theta u_i} du_i \\ &= 1 - \left[1 + \theta^{-1}\tilde{H}(t_i|x_i)\right]^{-\theta} \end{aligned} \tag{8.44}$$

which implies the hazard function

$$\lambda(t_i|x_i) = \tilde{\lambda}(t_i|x_i)\left[1 + \theta^{-1}\tilde{H}(t_i|x_i)\right]^{-1} \tag{8.45}$$

Now, we have two sources of duration dependence. A "true" duration dependence stemming from the hazard function $\tilde{\lambda}(t_i|x_i)$, which can be any of the

hazard functions presented in Section 8.2.6, and an additional "spurious" duration dependence arising from unobserved heterogeneity, which may amplify or diminish the "true" duration dependence. In the degenerate case $\theta^{-1} \to 0$ (zero variance of u_i) the frailty part drops out and we get a basic duration dependence model.

Example 8.9. Age at First Birth of Women in the GSS

In this example, we analyze the variable *age at first birth* of women using data drawn from the General Social Survey (GSS) in 2002. We investigated the age at first birth already in Section 8.2.2 by means of a lifetable and the Kaplan-Meier estimator. In Figures 8.2 and 8.3, we plotted the hazard rates and the empirical survivor function and the inspection of both figures is an initial step in analyzing duration data. At first, the hazard rate is flat, then increasing with a convex shape, and finally, slightly decreasing (with a lot of variation around the trend). Therefore, empirical evidence suggests first a positive and then a negative duration dependence, which is a common finding for the timing of first birth (see, for example, Morgan, 1996, Gustafsson, 2001).

Now, we want to examine the factors that affect the time until a woman gives first birth. In particular, higher educational attainment of the mother is known to postpone the age at first birth (Gustafsson et al., 2002), and we want to determine whether and how the number of years in formal schooling change the duration until first birth. Further controls include the number of siblings, two dummy variables, *white* and *immigrant*, and two variables indicating whether the woman was living in a family with low income (less than average income), and whether the woman was living in a city at age 16, respectively. We report estimates from four parametric duration models, the exponential model, the Weibull model, the log-normal model, and the log-normal model with gamma distributed frailty.

The results are displayed in Table 8.6. First of all, note that all models are estimated in the AFT structure for which a *ceteris paribus* unit increase in one regressor x_{il} changes the expected duration until first birth by approximately $\beta_l \times 100$ percent. Turning to the estimated coefficients, we find that women with a higher educational level have a higher expected duration until they give birth to a first child. For each model, we can clearly reject the null hypothesis of no effect. For example, in the exponential model, we have a point estimate of 0.047 with standard error 0.011 and z-statistic of about 4.3, which is higher than the 1% critical value of the standard normal distribution for a two-sided test. The specific effects of one additional year of schooling on the expected duration until first birth range between 2.6 percent (Weibull model) and 4.7 percent (exponential model).

The estimated coefficients are relatively stable when we compare the four models. Yet, there is clear statistical evidence that the Weibull model is preferred over the exponential model. The estimated parameter $\hat{\alpha} = 4.423$ indi-

Table 8.6. *Duration Analysis of Age at First Birth*

Dependent variable: *age at first birth*				
	Exponential	Weibull	Log-Normal	Log-Normal (Gamma Frailty)
years of education	0.047	0.026	0.031	0.032
	(0.011)	(0.002)	(0.002)	(0.002)
number of siblings	-0.015	-0.007	-0.005	-0.003
	(0.010)	(0.002)	(0.002)	(0.002)
white	0.046	0.063	0.083	0.091
	(0.072)	(0.016)	(0.015)	(0.013)
immigrant	0.094	0.034	0.056	0.060
	(0.093)	(0.021)	(0.019)	(0.017)
low income at age 16	-0.027	0.007	-0.021	-0.038
	(0.074)	(0.017)	(0.015)	(0.014)
lived in city at age 16	0.037	0.026	0.008	-0.004
	(0.061)	(0.014)	(0.012)	(0.011)
constant	2.707	2.890	2.696	2.615
	(0.170)	(0.037)	(0.036)	(0.034)
$\hat{\alpha}$		4.423		
		(0.095)		
$\hat{\sigma}$			0.219	0.150
			(0.005)	(0.006)
$\hat{\theta}^{-1}$				0.673
				(0.076)
Observations	1,371	1,371	1,371	1,371
Log-Likelihood value	-1,400.51	-258.35	-72.93	-8.70
LR	31.37	196.44	265.58	316.08
AIC	2,815.01	532.69	161.87	35.40

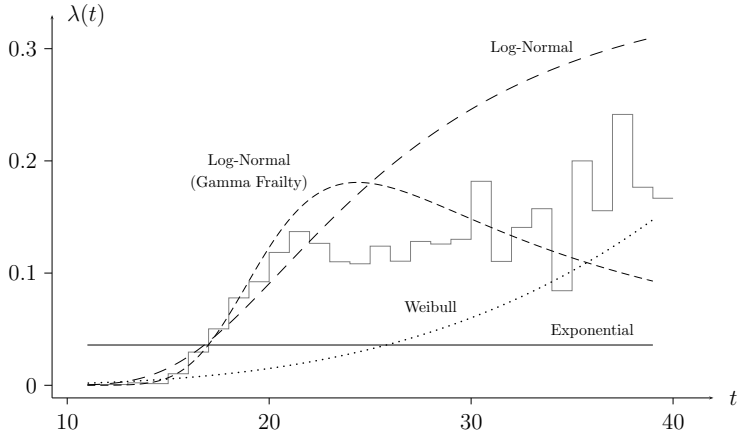
Notes: Standard errors in parentheses. Data: GSS 2002.

cates positive duration dependence and a likelihood ratio test between both models clearly rejects the null hypothesis of $\alpha = 1$. However, a coefficient of larger than two in the Weibull model suggests that the hazard rate is over-proportionally increasing over time. From Figure 8.2 we know that this is true only for short durations, and does not hold for long durations. Hence, we should consider a model that allows for more flexible shapes of the hazard function. Moreover, we control for very few factors, and therefore unobserved heterogeneity might be a serious concern. We address both issues by estimating a log-normal model with and without unobserved heterogeneity. And indeed, the log-normal model is preferred over the Weibull model, and, once we allow for unobserved heterogeneity in the log-normal model, we also reject the no-frailty model.

But what does this imply for the shape of the hazard function? Figure 8.6 shows four hazard functions for each of the four estimated models, evaluated at the sample means. We can see that the exponential and the Weibull model

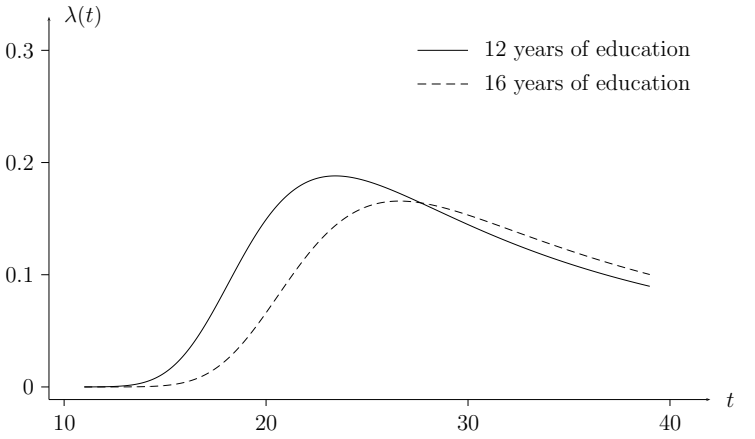
describe the empirical hazard rates very insufficiently. The log-normal model covers the hazard rates for low durations quite well, but deviates for high durations. Only the hazard function of the log-normal model with unobserved heterogeneity describes the hazard rates pretty well over the whole range of durations, which is also reflected by the high value of the log-likelihood function compared to the other models.

Fig. 8.6. *Parametric versus Empirical Hazard Functions*



In Figure 8.7, we plot the hazard function of the log-normal model with gamma frailty for two distinct values of the schooling variable, 12 and 16 years (fixing all other regressors to their means). The graphs show that a woman with less schooling, in this case 12 years, tends to give first birth earlier than a woman with higher schooling (16 years), all else being equal. However, we cannot observe a significant decline in average hazard rates, which confirms the hypothesis that higher educational attainment postpones the age at first birth but it does not significantly reduce the total number of children.

Fig. 8.7. *Hazard Functions by Educational Level*



Exercise 8.2.

- How would you interpret the coefficient of *white* in Table 8.6?
- Formally test the log-normal model against the Weibull model, and the log-normal model with gamma distributed unobserved heterogeneity against the log-normal model without heterogeneity using likelihood ratio tests.

8.3 Count Data Models

8.3.1 The Poisson Regression Model

The Poisson regression model is probably the most prominent member of the class of count data models, i.e., models that describe the number of occurrences of a particular event within a fixed time interval. The Poisson model for count data serves much the same function as the normal linear model for continuous data: it is used as a benchmark model against which more generalized (and maybe more suitable) models can be compared. Therefore, it is important to gain a good grasp of the underlying assumptions and properties of this standard model before turning to more general alternatives.

We start with a brief review of the Poisson distribution. Let y denote a random variable with support $\mathbb{N} \cup \{0\} = \{0, 1, 2, \ldots\}$, which is said to be **Poisson distributed** if and only if the probability function is given by

$$P(y = j) = \frac{e^{-\lambda}\lambda^j}{j!} \qquad \lambda > 0,\ j = 0, 1, 2, \ldots \tag{8.46}$$

where $j!$ denotes the factorial of j. In shorthand notation, we can write (8.46) as $y \sim Poisson(\lambda)$. The Poisson distribution is a one-parameter distribution and the parameter λ uniquely determines mean and variance. In particular, it can be shown that $\mathrm{E}(y) = \mathrm{Var}(y) = \lambda$. The equality of expectation and variance is also known as **equidispersion**, which is an important property of the Poisson distribution. Unfortunately, it is frequently violated in empirical applications, either because of **overdispersion** (the variance is larger than the mean) or **underdispersion** (the variance is smaller than the mean).

Exercise 8.3.

- Prove that expectation and variance of a Poisson distributed random variable are given by the parameter λ.

Two further properties invoked by the Poisson distribution should be mentioned. First, the ratio of probabilities of two successive outcomes j and $j+1$ is determined by

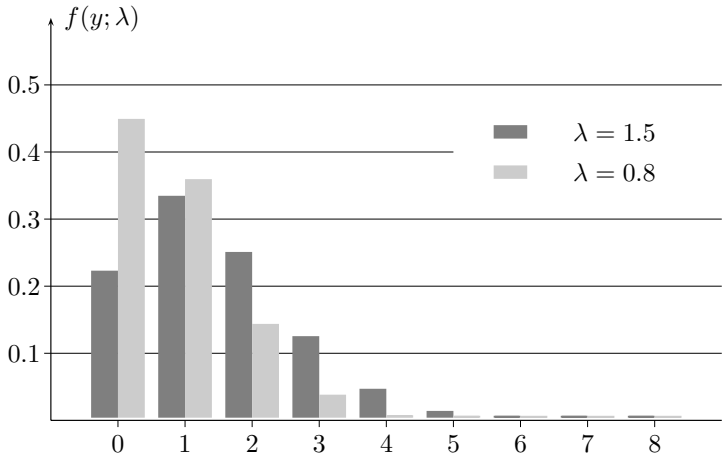
$$\frac{P(y = j)}{P(y = j + 1)} = \frac{j + 1}{\lambda} \qquad j = 0, 1, 2, \ldots \tag{8.47}$$

If the parameter λ is within the unit interval, $0 < \lambda < 1$, then this ratio is always $(\forall j)$ greater than one, such that the probabilities are strictly decreasing for increasing values of j. For $\lambda > 1$, the probabilities are first increasing and then decreasing. Both cases are illustrated in Figure 8.8. Second, the first derivative of (8.47) with respect to λ can be written as

$$\frac{\partial P(y=j)}{\partial \lambda} = \frac{e^{-\lambda}(-1)\lambda^j}{j!} + \frac{e^{-\lambda} j \lambda^{j-1}}{j!} = \frac{P(y=j)}{\lambda}(j-\lambda) \tag{8.48}$$

The sign of the first derivative is entirely determined by the sign of $j - \lambda$. If $j < (>)\lambda$, then the probabilities decrease (increase) with increasing λ. Both properties will become important when we consider discrete and marginal effects of regressors on the count outcomes.

Fig. 8.8. *The Shape of the Poisson Distribution*



The **Poisson regression model** can be derived from the distribution function stated in (8.46) by making the following two assumptions. First, we assume that the conditional distribution of y_i, given a set of regressors x_i, is Poisson distributed with parameter $\lambda_i = \lambda(x_i; \beta)$, formally

$$y_i | x_i \sim Poisson(\lambda_i) \tag{8.49}$$

And second, we parameterize λ_i in terms of x_i and β such that

$$\lambda_i = \exp(x_i'\beta) \tag{8.50}$$

where $x_i'\beta$ is the usual linear index of regressors including a constant. Both assumptions can be combined to obtain the conditional probability function of the Poisson regression model

$$f(y_i | x_i; \beta) = \frac{\exp(-\exp(x_i'\beta))\exp(y_i x_i'\beta)}{y_i!} \qquad y_i = 0, 1, 2, \ldots \tag{8.51}$$

with conditional expectation function $\mathrm{E}(y_i|x_i) = \exp(x_i'\beta)$ and conditional variance function $\mathrm{Var}(y_i|x_i) = \exp(x_i'\beta)$. This implies that, by assumption,

the Poisson model is heteroscedastic. This is a natural approach for non-negative random variables since the lower the linear index of regressors, and thus the expectation of y, the smaller the variance of y.

Interpretation of Parameters

The parameters of the Poisson model can be interpreted in two ways. First, since expectations of counts are well-defined, we may interpret the parameters in terms of marginal or discrete mean effects. And second, as with all discrete probability models, we may interpret the parameters in terms of marginal probability effects. The **marginal mean effect** is given by the first derivative of the conditional expectation function with respect to the l-th element in x_i, algebraically

$$\frac{\partial \mathrm{E}(y_i|x_i)}{\partial x_{il}} = \exp(x_i'\beta)\beta_l = \mathrm{E}(y_i|x_i)\beta_l \tag{8.52}$$

Equation (8.52) depends on the particular values of x_i, and it may therefore be more convenient to report the expected (or average) marginal effect, or alternatively, the effect evaluated for a "typical" individual (for example by fixing the regressors to their means). We already discussed this issue in the previous chapters and do not consider it further here.

Alternatively, we may report the relative change in the expectation function associated with a small change in one regressor. We obtain

$$\frac{\partial \mathrm{E}(y_i|x_i)/\mathrm{E}(y_i|x_i)}{\partial x_{il}} = \beta_l \tag{8.53}$$

This expression is constant for all i and has the interpretation of a **semi-elasticity**: if x_{il} increases by one unit, then $\beta_l \times 100$ gives the percentage change in $\mathrm{E}(y_i|x_i)$. Moreover, if x_{il} is in logarithmic form, then β_l measures the elasticity of the conditional expectation function with respect to x_{il}.

The relative change in $\mathrm{E}(y_i|x_i)$ associated with a **discrete change** in x_{il} by Δx_{il} can be written as

$$\begin{aligned}\frac{\mathrm{E}(y_i|x_i + \Delta x_{il}) - \mathrm{E}(y_i|x_i)}{\mathrm{E}(y_i|x_i)} &= \frac{\exp(x_i'\beta + \Delta x_{il}\beta_l) - \exp(x_i'\beta)}{\exp(x_i'\beta)} \\ &= \exp(\Delta x_{il}\beta_l) - 1 \end{aligned} \tag{8.54}$$

For example, if x_{il} is a dummy variable switching from 0 to 1, then $\exp(\beta_l) - 1$ gives the relative impact of the dummy variable on the expected value of y. For small values of β_l, we have $\exp(\beta_l) \approx \beta_l$.

An important observation can be drawn from the marginal mean effects. The Poisson model implies interactive effects even in the absence of an explicit interactive term $x_l x_m$, since

$$\frac{\partial^2 \mathrm{E}(y_i|x_i)}{\partial x_{il}\partial x_{im}} = \exp(x_i'\beta)\beta_l\beta_m = \mathrm{E}(y_i|x_i)\beta_l\beta_m \neq 0$$

This is unlike in the normal linear model with linear expectation function $\mathrm{E}(y_i|x_i) = x_i'\beta$ where interactive effects are zero unless the linear index contains an interactive term.

A second way to interpret the parameters in the Poisson model is to consider **marginal probability effects**. Clearly, the focus on the whole distribution of outcomes (and its response to small changes in regressors) provides much more information than focusing solely on mean effects. In other words, if we specify a conditional probability model for the dependent count variable, then we should also exploit the information provided by such a model. We obtain from (8.48)

$$\frac{\partial P(y_i = j|x_i)}{\partial x_{il}} = P(y_i = j|x_i)\big[j - \exp(x_i'\beta)\big]\beta_l \tag{8.55}$$

As mentioned previously, the marginal probability effects follow a specific pattern imposed by the model assumptions. Specifically, we have

$$\begin{aligned}\operatorname{sgn}(\partial P(y_i = j|x_i)/\partial x_{il}) &= -\operatorname{sgn}(\beta_l) \quad \text{if and only if} \quad j < \exp(x_i'\beta)\\ \operatorname{sgn}(\partial P(y_i = j|x_i)/\partial x_{il}) &= \operatorname{sgn}(\beta_l) \quad \text{if and only if} \quad j > \exp(x_i'\beta)\end{aligned}$$

Hence, if we increase the value of y, starting from the lowest outcome, then marginal probability effects are either first negative and turning to a positive value when y crosses its expected value $\exp(x_i'\beta)$, or vice versa. This is a **single crossing property**, similar to the one discussed in Chapter 6 in the context of ordered probit and logit models.

Estimation

Assuming a sample of n independent pairs of observations (y_i, x_i), the specification of a conditional probability model in (8.51) allows for straightforward application of ML methods. The log-likelihood function is

$$\log L(\beta; y, x) = \sum_{i=1}^{n} -\exp(x_i'\beta) + y_i x_i'\beta - \ln(y_i|) \tag{8.56}$$

The first-order condition for a maximum of (8.56) requires us to set the score function, which is given by

$$s(\beta; y, x) = \sum_{i=1}^{n} \big[y_i - \exp(x_i'\beta)\big]x_i \tag{8.57}$$

equal to zero. Since $s(\beta; y, x)$ is a non-linear function in β, we need to solve the first-order conditions $s(\hat{\beta}; y, x) = 0$ using an iterative algorithm like the Newton-Raphson method. The Hessian matrix in the Poisson regression model can be written as

$$H(\beta; y, x) = -\sum_{i=1}^{n} \exp(x_i'\beta) x_i x_i'$$

One can show that the Hessian is negative definite such that the log-likelihood function is globally concave, and the ML estimator $\hat{\beta}$ indeed maximizes (8.56). From ML theory we know that $\hat{\beta}$ is consistent, asymptotically efficient and asymptotically normal with

$$\hat{\beta} \overset{\text{app}}{\sim} Normal\left(\beta, \left[\sum_{i=1}^{n} \exp(x_i'\beta) x_i x_i'\right]^{-1}\right) \tag{8.58}$$

Since the Hessian matrix is independent of y, the expected and the actual Hessian coincide, resulting in the following estimator of the variance

$$\widehat{\text{Var}}_1(\hat{\beta}) = \left[\sum_{i=1}^{n} \exp(x_i'\hat{\beta}) x_i x_i'\right]^{-1}$$

An alternative variance estimator is the variance of the score, which can be estimated by summing over the outer product of the score

$$\widehat{\text{Var}}_2(\hat{\beta}) = \left[\sum_{i=1}^{n} (y_i - \exp(x_i'\hat{\beta}))^2 x_i x_i'\right]^{-1}$$

Asymptotically, the two estimators $\widehat{\text{Var}}_1$ and $\widehat{\text{Var}}_2$ are equivalent, as long as the model is correctly specified, because then $\text{Var}(y_i|x_i) = \text{E}[y_i - \exp(x_i'\beta)]^2 = \exp(x_i'\beta)$. However, in finite samples, the two formulas generally differ.

Duration Analysis and Count Data

In the introduction, we mentioned the close relationship between count data and duration analysis. The theoretical link between the two approaches is provided now. Let t_s denote the duration between two successive events $s-1$ and s, and let τ_j denote the **arrival time** of the j-th event, which is calculated as

$$\tau_j = \sum_{s=1}^{j} t_s \qquad j = 1, 2, \ldots \tag{8.59}$$

Furthermore, write $N(T)$ for the total number of events that have occurred within the fixed time interval $(0, T)$. By definition, it must hold that the probability of at most $j-1$ events having occurred before T equals the probability that the arrival time of the j-th event is greater than T. Formally,

$$P(N(T) < j) = P(\tau_j > T) = 1 - F_j(T) \tag{8.60}$$

where F_j is the cumulative density function of the arrival time τ_j. Moreover, the probability that exactly j events occurred in $(0, T)$ is

$$\begin{aligned} P(N(T) = j) &= P(N(T) < j + 1) - P(N(T) < j) \\ &= P(\tau_{j+1} > T) - P(\tau_j > T) \\ &= F_j(T) - F_{j+1}(T) \end{aligned} \tag{8.61}$$

Equation (8.61) characterizes the fundamental relationship between count data and duration data. If the distribution function F_j is known for all arrival times τ_j, then the distribution of $N(T)$ can be obtained by (8.61). In general, the specific form of F_j depends on a convolution of the underlying densities of t_s, and convolution is a rather complex operation. However, derivations simplify considerably if we assume that waiting times are identically and independently distributed with a common distribution. For example, one can show that exponentially distributed waiting times with constant hazard λ result in probabilities $P(N(T) = j)$ that follow a Poisson distribution with single parameter λ.

8.3.2 Unobserved Heterogeneity

In the Poisson regression model, we assume that $y_i|x_i$ is Poisson distributed with parameter $\lambda_i = \exp(x_i'\beta)$, and λ_i uniquely determines mean and variance of the conditional distribution. This presumes that the analyst is able to account for the full amount of individual heterogeneity. Once the explanatory variables x_i and β are known, the conditional probability model is fully specified. There is no room for additional, unobserved heterogeneity.

Unfortunately, given that the analyst does *not* observe all relevant information, the conditional distribution of $y_i|x_i$ cannot be Poisson, and thus the Poisson model must be misspecified. This can be seen as follows. Similar to Section 8.2.7, we may write the Poisson parameter as

$$\tilde{\lambda}_i = \exp(x_i'\beta + \varepsilon_i) \tag{8.62}$$

with error term ε_i capturing unobserved heterogeneity. We can rewrite the parameter $\tilde{\lambda}_i$ as

$$\tilde{\lambda}_i = \exp(x_i'\beta)\exp(\varepsilon_i) = \exp(x_i'\beta)u_i = \lambda_i u_i \tag{8.63}$$

where $u_i = \exp(\varepsilon_i)$. Now, assume that y_i conditional on x_i *and* u_i is Poisson distributed with parameter $\tilde{\lambda}_i$, and that u_i is distributed independently of x_i with $\mathrm{E}(u_i|x_i) = 1$ and $\mathrm{Var}(u_i|x_i) = \sigma_u^2$. Since we observe y_i and x_i, but not u_i, we are interested in the distribution of y_i conditional on x_i, but unconditional on u_i. By the law of iterated expectations and the decomposition of variance (see Hogg and Craig, 1978, p. 349), we obtain

$$\mathrm{E}(y_i|x_i) = \mathrm{E}_u(\tilde{\lambda}_i|x_i) = \exp(x_i'\beta)\,\mathrm{E}(u_i|x_i) = \lambda_i \tag{8.64}$$

$$\mathrm{Var}(y_i|x_i) = \mathrm{E}_u(\tilde{\lambda}_i|x_i) + \mathrm{Var}_u(\tilde{\lambda}_i|x_i) = \lambda_i + \sigma_u^2\lambda_i^2 \tag{8.65}$$

For $\sigma_u^2 > 0$, we have $\text{Var}(y_i|x_i) > \text{E}(y_i|x_i)$. Hence, unobserved heterogeneity causes **overdispersion**, and therefore violates the equality of mean and variance (equidispersion) in the Poisson model.

Example 8.10. Unobserved Heterogeneity across Groups

Assume two homogeneous groups of equal size in the population of interest. Each group is characterized by a Poisson distributed random variable, denoted by y_1 and y_2, with parameters $\lambda_1 = 0.5$ and $\lambda_2 = 1.5$, respectively. Due to the lack of data availability, the analyst cannot distinguish between both groups. The results above yield

$$\begin{aligned}
\text{E}(y) &= 0.5 * \text{E}(y_1) + 0.5 * \text{E}(y_2) \\
&= 0.5 * 0.5 + 0.5 * 1.5 = 1 \\
\text{Var}(y) &= 0.5 * \text{Var}(y_1) + 0.5 * \text{Var}(y_2) \\
&\quad +0.5 * (\text{E}(y_1) - \text{E}(y))^2 + 0.5 * (\text{E}(y_2) - \text{E}(y))^2 \\
&= 0.5 * 0.5 + 0.5 * 1.5 + 0.5 * (-0.5)^2 + 0.5 * 0.5^2 = 1.25
\end{aligned}$$

Thus, the unconditional variance of y in the population is greater than its unconditional mean, and therefore, the population cannot be Poisson distributed.

In order to study the effect of unobserved heterogeneity on the full conditional probability model (rather than the first two moments only), we need to integrate out the unobservables u_i. This requires that we know the distribution function $g(u_i|x_i)$. In this case, the probability function in terms of observables can be written as

$$f(y_i|x_i) = \int_0^\infty f(y_i|x_i, u_i) g(u_i|x_i) du_i \tag{8.66}$$

For instance, if $f(y_i|x_i, u_i)$ is of the Poisson form, we get

$$f(y_i|x_i) = \frac{\lambda_i^{y_i}}{y_i!} \int_0^\infty e^{-\lambda_i u_i} u_i^{y_i} g(u_i|x_i) du_i \tag{8.67}$$

where $\lambda_i = \exp(x_i'\beta)$. Whether or not a closed-form solution of (8.66) or (8.67) exists crucially depends on the assumption about $g(u_i|x_i)$. For example, if $u_i|x_i$ is gamma distributed with the density function given by (8.37), then we obtain from (8.67)

$$f(y_i|x_i) = \frac{\lambda_i^{y_i}}{y_i!} \int_0^\infty e^{-\lambda_i u_i} u_i^{y_i} \frac{\gamma^\theta}{\Gamma(\theta)} u_i^{\theta-1} e^{-\gamma u_i} du_i \tag{8.68}$$

In order to impose the normalization $\mathrm{E}(u_i|x_i) = 1$, we restrict the parameters such that $\gamma = \theta$. Therefore,

$$\begin{aligned} f(y_i|x_i) &= \frac{\lambda_i^{y_i}}{y_i!} \int_0^\infty e^{-\lambda_i u_i} u_i^{y_i} \frac{\theta^\theta}{\Gamma(\theta)} u_i^{\theta-1} e^{-\theta u_i} du_i \\ &= \frac{\lambda_i^{y_i}}{y_i!} \frac{\theta^\theta}{\Gamma(\theta)} \int_0^\infty e^{-(\lambda_i+\theta)u_i} u_i^{y_i+\theta-1} du_i \\ &= \frac{\lambda_i^{y_i}}{\Gamma(y_i+1)} \frac{\theta^\theta}{\Gamma(\theta)} \frac{\Gamma(y_i+\theta)}{(\lambda_i+\theta)^{y_i+\theta}} \\ &= \frac{\Gamma(y_i+\theta)}{\Gamma(y_i+1)\Gamma(\theta)} \left(\frac{\lambda_i}{\lambda_i+\theta}\right)^{y_i} \left(\frac{\theta}{\lambda_i+\theta}\right)^{\theta} \end{aligned} \tag{8.69}$$

where the third equality follows from an expansion of the integrand such that it represents the density function of a $Gamma(y_i + \theta, \lambda_i + \theta)$ distribution, and therefore integrates to one, and $\Gamma(y_i + 1) = y_i!$ since y_i is integer-valued. Equation (8.69) shows the probability function of the **negative binomial distribution** with $\mathrm{E}(y_i|x_i) = \lambda_i$ and $\mathrm{Var}(y_i|x_i) = \lambda_i + \theta^{-1}\lambda_i^2$.

The **Negbin regression models** can be derived from (8.69) for a suitable specification of the parameter θ, with λ_i given by $\exp(x_i'\beta)$. Two parameterizations are known in the literature. The **Negbin II model** assumes that $\theta^{-1} = \sigma^2$ is constant, such that the variance function can be written as

$$\mathrm{Var}(y_i|x_i) = \exp(x_i'\beta) + \sigma^2[\exp(x_i'\beta)]^2 \tag{8.70}$$

This corresponds to the variance function that we obtained in (8.65). In the **Negbin I model**, the parameter θ is allowed to vary across individuals with specification $\theta_i^{-1} = \sigma^2/\lambda_i$ and the variance function is given by

$$\mathrm{Var}(y_i|x_i) = (1 + \sigma^2)\exp(x_i'\beta) \tag{8.71}$$

For $\sigma^2 \to 0$, the Negbin regression models converge to the Poisson model, which can be used as a basis for a LR or Wald test. Since $\sigma^2 > 0$, the Poisson model is obtained as a degenerate case at the boundary of the parameter space, which needs to be taken into account when testing such restrictions (see Winkelmann, 2003, Chapter 3.4.1 for further details).

Example 8.11. Schooling and Fertility

In this example we investigate the effect of women's schooling on fertility using data from the GSS waves 1974 to 2002. The individual fertility decision is measured by the *number of children ever borne to a woman*, and thus is a count variable. We focus on women aged 40 years or older to account for the fact that younger women may not have reached the end of their fertility, or are still attending school. The marginal distribution of the number of children was displayed in Table 1.1. In particular, the effect of schooling on fertility has

attracted much interest in the past (see for example Willis, 1973), since raising educational levels might explain the observed downward trend in fertility.

We analyze this hypothesis with three different count data models, the standard Poisson model, and the Negbin I and II regression models to allow for unobserved heterogeneity. Further controls include a time trend, dummy variables indicating race (*white*), immigrant status, living in a city at age 16, and having less-than-average income at age 16. Table 8.7 reports the results.

Table 8.7. *Estimation Results of Fertility Decision*

Dependent variable: *number of children ever born to women*

	Poisson	Negbin II	Negbin I
years of education	-0.0442	-0.0452	-0.0423
	(0.0028)	(0.0033)	(0.0032)
linear time trend	-0.0055	-0.0055	-0.0051
	(0.0010)	(0.0012)	(0.0012)
white	-0.1366	-0.1376	-0.1267
	(0.0230)	(0.0268)	(0.0262)
immigrant	-0.0808	-0.0791	-0.0815
	(0.0276)	(0.0316)	(0.0314)
low income at age 16	0.0115	0.0116	0.0096
	(0.0212)	(0.0246)	(0.0241)
lived in city at age 16	-0.0555	-0.0549	-0.0527
	(0.0190)	(0.0218)	(0.0216)
constant	1.7080	1.7208	1.6722
	(0.0387)	(0.0459)	(0.0442)
$\hat{\sigma}$		0.1244	0.3019
		(0.0107)	(0.0283)
Observations	5,150	5,150	5,150
Log-likelihood value	-10,116.63	-10,014.05	-10,032.33
LR of H_0 : constant-only	484.28	355.04	318.49
LR of H_0 : $\sigma = 0$		205.17	168.62

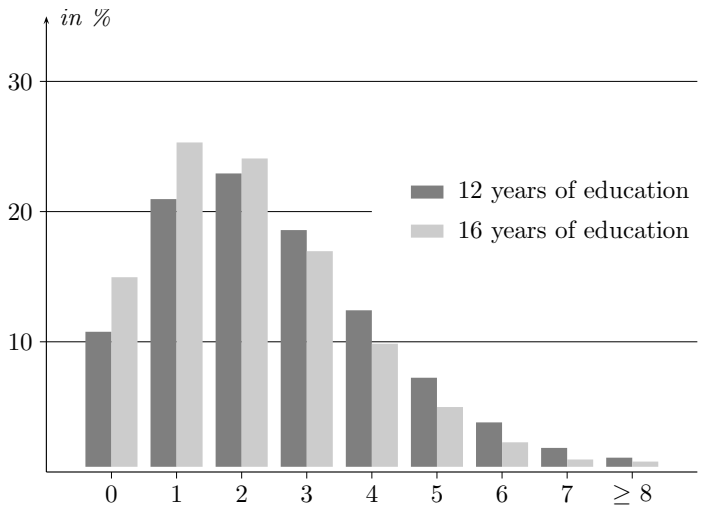
Notes: Standard errors in parentheses. Data: GSS 1974-2002 (4-year intervals).

All models indicate that schooling has a negative effect on the number of children. The value -0.0442 in the Poisson model can be interpreted in two ways. First, in terms of mean effects, the increase in schooling by one year reduces the expected number of children by about 4.42 percent. The exact change can be calculated as $[\exp(-0.0442)-1]\times 100 = -4.32$ percent. In terms of marginal probability effects, we need to evaluate (8.55) for each individual and outcome. For example, consider an individual with average characteristics. In this case, the conditional expectation is estimated as $\hat{\lambda} = 2.54$, and a marginal increase in schooling increases the probability of a zero outcome by $\exp(-2.54) * (-2.54) * (-0.04) = 0.0088$, or 0.88 percentage points.

The coefficients in the Negbin I and II models do not differ much from those in the Poisson model. However, if we look at the parameter estimate of σ, then we can see that there is evidence of unobserved heterogeneity. This can be formally tested, using the LR test statistics reported in the last row of Table 8.7, and we reject the absence of heterogeneity (and thereby the Poisson model) significantly. Moreover, from a statistical point of view, the Negbin II model is preferred over the Negbin I model.

The Negbin regression models can also be interpreted in terms of marginal mean or probability effects. In order to shed some more light on our discussion about the effect of schooling on fertility, we choose another means of interpretation. We calculate predicted probabilities for two levels of schooling, 12 and 16 years, showing the difference between a high school and a college degree. The predictions of the Negbin II model are depicted in Figure 8.9, given that all other regressors are fixed to their means. We can see that, *ceteris paribus*, the probability function differs conditional on the chosen levels, and that higher education reduces the probability of three or more children.

Fig. 8.9. *Predicted Probabilities by Educational Level*



8.3.3 Efficient versus Robust Estimation

The Negbin regression model is an example of an efficient estimation strategy. If the assumptions of the Poisson model are violated because unobserved heterogeneity is present, and the unobserved heterogeneity term is gamma distributed independently of the regressors, then the estimation of the Negbin model is the best way to proceed.

What would be the consequences of estimating the Poisson model instead? As we have shown, unobserved heterogeneity causes overdispersion, violating one of the restrictions of the Poisson distribution. However, the problem is more general. There are many possible reasons, apart from unobserved heterogeneity, why the conditional variance in the Poisson model would depart from the conditional mean. Interestingly, the consequences of such a departure are very similar to those of heteroscedasticity in the linear regression model: the parameter estimates remain consistent, but the usual variance matrix is inconsistent and the estimator is inefficient.

Thus, one possible strategy is not to pursue an efficient estimation at all, but rather use the Poisson model in combination with a corrected estimator of the standard errors. The formula for such "robust" standard errors is given, for instance, in Winkelmann (2003). The result that the parameters of the Poisson regression model are consistently estimated under this kind of misspecification has been derived by Hausman, Hall and Griliches (1984). The procedure of using the Poisson estimates together with robust standard errors is also referred to as "quasi-maximum likelihood estimation" (QML). The potential advantage of QML over efficient alternatives is that the latter are in general not even consistent if the assumptions underlying the efficient estimation are violated (as it is the case if the unobserved heterogeneity term is not gamma distributed).

8.3.4 Censoring and Truncation

While unobserved heterogeneity can be interpreted as a lack of observability in the *independent* variables, we now turn our attention to limitations in the observability of the *dependent* variable. More specifically, we cover two forms of limited data observability for counts, **truncation** and **censoring**, adopting methods from Chapter 7 for continuous dependent variables.

In the case of a **truncated** count variable, certain observations (y_i, x_i) are omitted entirely from the sample at hand. For example, if we ask unemployed workers about the number of unemployment spells, then we observe at least one unemployment spell. This is an example of (left) truncation at zero. Now suppose that we have a truncated-at-zero Poisson model. In this case, the observed data probability function can be written as

$$g(y_i|y_i > 0, x_i) = \frac{f(y_i, y_i > 0|x_i)}{f(y_i = 0|x_i)} \tag{8.72}$$

$$= \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i![1 - \exp(-\lambda_i)]} \qquad y_i = 1, 2, \ldots \tag{8.73}$$

with parameter λ_i specified like in the regular Poisson model as $\lambda_i = \exp(x_i'\beta)$. The general form of the probability function in (8.72) can easily be modified to account for other truncation values. Moreover, application to the Negbin models is straightforward if we replace f by the probability function of the negative binomial distribution.

Exercise 8.4.

Assume that y_i follows a truncated-at-zero Poisson distribution.

- Derive mean and variance of y_i.
- Does the model display under-, equi-, or overdispersion?

In the case of **censoring**, the dependent count variable is only observed over a limited range. A typical example arises in survey questions with top-coding of answer categories, such as "x or more" for the number of patents per year, which is an example of right-censoring. We can deal with a censored count variable if we introduce an observation mechanism of the form

$$c_i = \begin{cases} 1 & \text{for } y_i^* \in A = \{0, \ldots, a\} \\ 0 & \text{for } y_i^* \in A = \{a+1, a+2, \ldots\} \end{cases} \tag{8.74}$$

for some positive integer a. Here, y_i^* denotes the count variable in the population of interest with probability function $f(y_i^*)$, and c_i is referred to as selection variable. We observe $y_i = y_i^*$ if $c_i = 1$, but in the case of $c_i = 0$ the only information we have is that $y_i^* > a$. It follows that $P(c_i = 1) = F(a)$, where $F(a) = \sum_{j=0}^{a} f(j)$, and $P(c_i = 0) = 1 - F(a)$. Thus, the probability function of the observed count variable y_i can be written as

$$g(y_i|x_i, c_i) = f(y_i)^{c_i}[1 - F(a)]^{1-c_i} \tag{8.75}$$

The observation mechanism in (8.74) and the latent count are based on a single population model, i.e., the observation rule is entirely determined by the distribution of y_i^* and an exogenous censoring value a. This approach can be generalized to incidental censoring if we separate the processes for the count and the observation mechanism using a particular correlation structure (see Terza, 1998, or Winkelmann, 1998, for further details).

8.3.5 Hurdle and Zero-Inflated Count Data Models

The last two generalizations of the Poisson regression model that we discuss here address a problem that appears in many empirical applications, namely that zero outcomes are observed more frequently than is likely to be compatible with the standard Poisson or Negbin regression models. Here, we speak of "extra" or "excess zeros".

The first of the two generalizations is the **hurdle count data model**. The basic idea is to combine a binary model (indicating whether the outcome of the count is below or above the hurdle, where the hurdle typically is set at 0/1) with a truncated count data model for outcomes above the hurdle. These models are thus also referred to as **two-part models** (see also Chapter 7.2.7 for a related continuous variable model).

In the **hurdle-at-zero** Poisson model we assume that the statistical process generating the outcome $y_i = 0$ differs from the one generating positive integers. Let $f_1(0; \lambda_{1i})$ denote the probability of a zero outcome in a Poisson distribution with parameter λ_{1i}. If the hurdle is crossed, i.e., for $j = 1, 2, \ldots$, we assume that the dependent variable follows a truncated-at-zero Poisson distribution $f_2(j|j > 0; \lambda_{2i})$. This yields the two-part probability model

$$
\begin{aligned}
P(y_i = 0|x_i) &= f_1(0; \lambda_{1i}) = e^{-\lambda_{1i}} \\
P(y_i = j|x_i) &= \left[1 - f_1(0; \lambda_{1i})\right] \frac{f_2(j; \lambda_{2i})}{1 - f_2(0; \lambda_{2i})} \qquad (8.76) \\
&= \frac{1 - e^{-\lambda_{1i}}}{1 - e^{-\lambda_{2i}}} \frac{e^{-\lambda_{2i}} \lambda_{2i}^j}{j!} \qquad j = 1, 2, \ldots
\end{aligned}
$$

where we specify $\lambda_{1i} = \exp(x_i'\beta_1)$ and $\lambda_{2i} = \exp(x_i'\beta_2)$. The hurdle model nests the standard Poisson model under the restriction $\beta_1 = \beta_2$, and therefore a simple LR test can be conducted for model discrimination. A hurdle negative binomial model can be constructed in a similar way (see Pohlmeier and Ulrich, 1995).

Exercise 8.5.

- Derive the conditional expectation function $\mathrm{E}(y_i|x_i)$ in the hurdle-at-zero Poisson model.
- What are the marginal mean effects?

Example 8.12. The Determinants of Labor Mobility

The number of job changes by an individual during a given interval of time is a typical count variable. The possible values are $0, 1, \ldots$ with no obvious

upper bound. The German Socio-Economic Panel (GSOEP) included in the first wave a question on the total number of changes during the previous ten years, i.e., between 1974-84. There were 1,962 observations on men aged between 35 and 64, excluding self-employed persons and civil servants (see Winkelmann, 2003, for a more detailed exposition of this example).

The unconditional mean number of job changes in the data is 0.540. If the number of truly changes were Poisson distributed, this would imply a probability of zero changes given by $P(y = 0) = \exp(-0.540) = 0.583$, or 58.3 percent. In fact, a zero is recorded for 67.9 percent of all observations, indicating the possible presence of excess zeros. This is confirmed by a formal regression analysis. Table 8.8 shows the results from a Poisson and a hurdle Poisson model, using the variables *years of schooling*, *years labor market experience in 1974*, *union membership* and *German nationality* as explanatory variables. The likelihood ratio test statistic is, under H_0 : Poisson model is valid, chi-squared distributed with six degrees of freedom. The observed test statistic of 232.94 has a p-value of zero.

Table 8.8. *Estimation Results of Number of Job Changes*

		Hurdle Poisson	
	Poisson	Hurdle 0/1+	Hurdle 1+
education$*10^{-1}$	-0.138	0.133	-0.600
	(0.137)	(0.170)	(0.218)
experience$*10^{-1}$	-0.770	-0.758	-0.403
	(0.111)	(0.148)	(0.156)
$experience^2 * 10^{-2}$	0.119	0.107	0.085
	(0.037)	(0.048)	(0.050)
union	-0.292	-0.268	-0.167
	(0.065)	(0.084)	(0.097)
german	-0.368	-0.330	-0.206
	(0.076)	(0.101)	(0.108)
Log-likelihood value	-2,044	-1,928	
Observations	1,962	1,962	

Notes: Standard errors in parentheses. Data: GSOEP 1984.

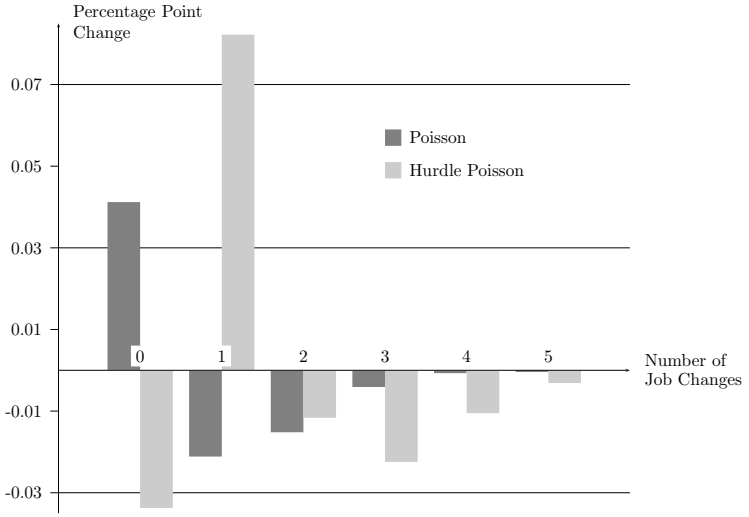
Importantly, the marginal probability effects in the hurdle model differ substantially from those in the Poisson model. The hurdle model is no longer bound by the single-crossing property, since it can be shown that

$$\frac{\partial P(y_i = j|x_i)}{\partial x_{il}} = P(y_i = j|x_i) \left[\frac{\lambda_{1i} e^{-\lambda_{1i}}}{1 - e^{-\lambda_{1i}}} \beta_{1l} - \frac{\lambda_{2i} e^{-\lambda_{2i}}}{1 - e^{-\lambda_{2i}}} \beta_{2l} + (j - \lambda_{2i}) \beta_{2l} \right]$$

Figure 8.10 displays the predicted probability changes associated with ten additional years of education.

Based on the hurdle model, such an increase reduces the probability of no job change by about 3 percentage points, whereas the Poisson model pre-

Fig. 8.10. *Marginal Probability Effects of Education: Poisson and Hurdle Poisson*



dicts an increase. Similarly, in the hurdle Poisson model, we find that more education increases the probability of one job change, whereas the simple Poisson model predicts a decrease. Using the hurdle model, we thus come to conclusions with regard to marginal probability effects that are diametrically opposed to those obtained from the Poisson model. This is a powerful illustration of the idea that a sufficiently flexible model is imperative for obtaining meaningful marginal probability effects.

Apart from a hurdle-at-zero approach, the problem of **excess zeros** can be handled with a **zero-inflated** count data model. The basic idea in a zero-inflated model is to introduce a binary variable c_i indicating whether the observation is zero ($c_i = 1$), or stems from a standard count data distribution with support over the non-negative integers ($c_i = 0$).

Let ω denote the probability of $c_i = 1$ and suppose a latent count variable y_i^* that is Poisson distributed with parameter $\lambda_i = \exp(x_i'\beta)$. The observed count y_i is given by

$$y_i = \begin{cases} 0 & \text{if } c_i = 1 \\ y_i^* & \text{if } c_i = 0 \end{cases} \tag{8.77}$$

and has probability function

$$f(y_i|x_i) = \omega_i d_i + (1 - \omega_i)\frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} \qquad y_i = 0, 1, 2, \ldots \tag{8.78}$$

where $d_i = I(y_i = 0)$. Note that the probability of a zero outcome now is $P(y_i = 0|x_i) = \omega + (1 - \omega)e^{\lambda_i}$, which is strictly greater than the probability of a zero outcome in the standard Poisson model. We conclude that in the zero-inflated model, the zero outcome arises from two types of regimes, either from regime 1 ($c_i = 1$) with probability ω, or from regime 2 ($c_i = 0$ and $y_i^* = 0$) with probability $1 - \omega$. In contrast, the structure of the hurdle model implies that the probability of a zero outcome is determined by a (single) process describing this outcome.

Exercise 8.6.

- Derive the conditional expectation function $\mathrm{E}(y_i|x_i)$ in the zero-inflated Poisson model.

8.4 Further Exercises

Exercise 8.7 Consider the case of discrete duration data with discrete hazard $P_{it} = P(t_i = t|t_i \geq t)$ under the assumption of inflow sampling and right-censoring.

a) What is the likelihood contribution for a completed spell?

b) What is the likelihood contribution for a censored spell?

c) Derive the likelihood and the log-likelihood function for a sample of n independent observations. What do you conclude?

Exercise 8.8 How would the likelihood function in Problem 8.7 change, if we allowed for

a) left-truncated spell data, also referred to as "delayed entries"?

b) right-truncated spell data, also referred to as "outflow sample"?

Exercise 8.9 Compare the model formulation and the estimation method of discrete time hazard models and sequential models as discussed in Chapter 6.4 for ordered dependent variables.

a) Where do you see similarities?

b) What are the differences?

Exercise 8.10 Derive the expectation and the variance of the gamma distribution with two parameters θ and γ as given in equation (8.37). How do your results simplify in the special case $\theta = \gamma$?

Exercise 8.11 Verify that integrating out the unobserved heterogeneity in equation (8.44) yields the stated expression.

Exercise 8.12 Research papers are usually subject to peer-review prior to publication (or rejection) by an academic journal. The length of the refereeing process is a source of some controversy.

Ten manuscripts were received by an editorial office during a given week and sent out to the referees. The following data shows the elapsed time (in days) between the moment the manuscript was sent out and the receipt of the referee report.

27, 51, 128, 63, 40, 47, 246, 106

Two further reports were still outstanding at the end of the observation period (after 250 days).

a) Assume that the refereeing time is exponentially distributed. Estimate the return rate of referee reports by maximum likelihood. Account for the fact that two reports have not arrived yet.

b) How do you interpret the return rate? What is the expected duration of obtaining a referee report?

Exercise 8.13 Derive the modal value(s) of the probabilities in the Poisson model conditional on the parameter λ.

Exercise 8.14 Derive the conditional variance function $\text{Var}(y_i|x_i)$ in the hurdle-at-zero Poisson model.

Exercise 8.15 Assume that a random variable y_i follows a truncated-at-one Poisson distribution.

a) Derive the probability function of y_i.

b) Derive mean and variance of y_i. Does the truncated-at-one model display under-, equi-, or overdispersion?

Exercise 8.16 One area where count data models can be usefully employed is in the study of worker absenteeism. The following table shows regression results for a sample of 604 workers employed in a UK factory (see also Barmby et al., 2001). The dependent variable is the annual number of non-condonable absences. The explanatory variables include the classification of the worker in the sickpay scheme (A: entitled to full benefits and bonus payments; B and C: entitled to successively reduced benefits, no bonus payments; based on experience rating). Workers may work four-day weeks or five-day weeks, which is captured by a dummy variable. Four-day workers may work the same number of weekly hours (and thus longer daily hours).

	Poisson	Negative Binomial
female	0.283	0.267
	(0.098)	(0.108)
grade B	0.547	0.564
	(0.093)	(0.094)
grade C	0.710	0.718
	(0.111)	(0.115)
hourly wage	-0.404	-0.406
	(0.291)	(0.270)
daily hours	0.001	-0.009
	(0.024)	(0.036)
five days	-0.217	-0.319
	(0.140)	(0.151)
constant	2.850	3.018
	(1.017)	(0.899)
Observations	604	604
Log-likelihood value	-2,955.9	-1,845.8

Notes: Standard errors in parentheses.

a) Test the Poisson model against the constant-only model and the negative binomial model.

b) Do you find large differences in the estimated coefficients? Explain.

c) If the probability of an absence was constant on each day, what would the expected difference in absences between five-day workers and four-day workers be? What do you find in the data?

Exercise 8.17 Suppose you want to examine if living in a city (a binary indicator d_i) affects the number of children ever borne to a woman. Let y_i denote the number of children and assume that $y_i|d_i$ is Poisson distributed with parameter $\lambda_i = \exp(\beta_0 + \beta_1 d_i)$. You have dataset with 100 women and 40 percent are living in a city. On average, women living in a city have 2.5 children, whereas women not living in city have 3 children.

a) Write down the log-likelihood function of the model and derive the score function and the Hessian matrix.

b) Calculate the ML estimates of β_0 and β_1.

c) Write down the restricted log-likelihood function under the null hypothesis that living in a city does not affect the number of children. Calculate the ML estimates under this hypothesis.

d) Test the hypothesis of no effect using a Score test and a LR test, respectively. What do you conclude?

e) How do your answers change if you increase the sample size to 1,000 and the proportions and sample means remain as given above?

List of Figures

1.1 *Types of Microdata* 5

2.1 *Relative Frequencies of Number of Children Ever Born* 23
2.2 *Predicted Poisson Probabilities* 38

3.1 *Likelihood Function for the Bernoulli Example* 48
3.2 *Likelihood and Log-Likelihood in the Bernoulli Example* 52
3.3 *Expected Score Functions in the Bernoulli Example* 56
3.4 *The Newton-Raphson Algorithm* 72
3.5 *Unrestricted and Restricted Maximization* 80
3.6 *Wald, Likelihood Ratio and Score Test* 81

4.1 *Probability Function in the Probit Model* 100
4.2 *Comparison of Probit and Logit Model* 102
4.3 *Discrete versus Marginal Change in Nonlinear Models* 105
4.4 *Predicted Probabilities by Years of Education* 120
4.5 *Receiver Operating Characteristic Curve for Artificial Data* 125

5.1 *Transport Mode Decision Tree* 165

6.1 *Life-Satisfaction Question in the GSOEP* 171
6.2 *Threshold Mechanism in Terms of* y_i^* 175
6.3 *Threshold Mechanism in Terms of* u_i 175
6.4 *Graphical Illustration of the Ordered Probit Model* 177
6.5 *Shift in Density Due to a Change* $\Delta x_{il} > 0$ *(*$\beta_l > 0$*)* 181
6.6 *Discrete Probability Effects in the Standard Model* 182
6.7 *Parallel Regression Assumption* 186
6.8 *Generalized Threshold Mechanism* 189
6.9 *Discrete Probability Effects in the Generalized Model* 190

7.1 *Density Functions of Truncated Normal Distributions* 215

7.2 *Expected Value and Inverse Mills Ratio for Truncated Normal Distribution* 217
7.3 *Conditional Expectation Functions in the Tobit Model* ($\sigma^2 = 1$) . 219
7.4 *Simulated Data for Regression with and without Censoring* 227

8.1 *Example of an Individual Event History* 251
8.2 *Empirical Hazard Rates of Age at First Birth* 258
8.3 *Kaplan-Meier Estimator of the Survivor Function* 258
8.4 *Flow Sampling With Censoring* 263
8.5 *Weibull Hazard Functions* 268
8.6 *Parametric versus Empirical Hazard Functions* 277
8.7 *Hazard Functions by Educational Level* 278
8.8 *The Shape of the Poisson Distribution* 280
8.9 *Predicted Probabilities by Educational Level* 288
8.10 *Marginal Probability Effects of Education: Poisson and Hurdle Poisson* 293

List of Tables

1.1 *Fertility Distribution* 12
1.2 *Fertility and Average Education Level by Years* 14
1.3 *Linear Regression Analysis of Fertility* 15
1.4 *Mother's Education and School Track of Child* 16

2.1 *Conditional Relative Frequency Distributions* 22

3.1 *Probit Estimates of Fertility Decision* 84

4.1 *Probit and Logit Estimates of Fertility Decision* 111
4.2 *The Effect of Education on the Probability of Being Childless* . . 121
4.3 *Contingency Table for Binary Predictions* 123
4.4 *Predicting Binary Outcomes* 124
4.5 *Sensitivity and Specificity* 124
4.6 *Goodness-of-Fit Measures* 126

5.1 *Multinomial Logit Estimates of Secondary School Choice* 143
5.2 *Predicted Probabilities and Mother's Educational Level* 148
5.3 *The Effect of Mother's Education on Secondary School Choice* . . 149
5.4 *Cracker Brand Choice* 153
5.5 *Data Transformation in the Conditional Logit Model* 155
5.6 *Cracker Brand Choice: A Conditional Logit Analysis* 157
5.7 *Marginal Probability Effects* 158

6.1 *Ordered Probit and Logit Estimates of Secondary School Choice* 184
6.2 *Marginal Probability Effects of Mother's Educational Level* 185
6.3 *Recoding of Happiness Responses and Relative Frequencies* 191
6.4 *(Generalized) Ordered Logit Estimates of Happiness Equation* . . 192
6.5 *Marginal Effects of Income on Happiness* 193
6.6 *Artificial Dataset* 198
6.7 *Data Organization in the Standard Sequential Model* 199

6.8 *Data Organization in the Generalized Sequential Model* 199
6.9 *Example of an Income Classification* 200
6.10 *Log-Transformation of Income Classification* 201

7.1 *Tobit and OLS Estimates of Female Hours of Work* 213
7.2 *Artificial Data with Right-Censoring and Truncation at One* . . . 224
7.3 *Censored Regression of Time to First Birth* 229
7.4 *Wage Offer Equation for Married Women* 236

8.1 *An Artificial Data Example* 255
8.2 *Lifetable of Age at First Birth Among Women in the GSS* 257
8.3 *Retirement Age and Health: Fictitious Data* 260
8.4 *Retirement Age and Health in Binary Data Format* 261
8.5 *Unobserved Heterogeneity and Spurious Duration Dependence* . . 273
8.6 *Duration Analysis of Age at First Birth* 276
8.7 *Estimation Results of Fertility Decision* 287
8.8 *Estimation Results of Number of Job Changes* 292

References

Aitchison, J., and S. Silvey (1957): "The Generalization of Probit Analysis to the Case of Multiple Responses," *Biometrika*, 44, pp. 131–140.

Amemiya, T. (1985): *Advanced Econometrics*, Harvard University Press.

Amemiya, T. (1994): *Introduction to Statistics and Econometrics*, Harvard University Press.

Baltagi, B.H. (2005): *Econometric Analysis of Panel Data* 3rd ed. Wiley.

Bantle, C., and J.P. Haisken-DeNew (2002): "Smoke Signals: The Intergenerational Transmission of Smoking Behavior," *DIW Discussion Papers No. 277.*

Barmby, T., M. Nolan, and R. Winkelmann (2001): "Contracted Workdays and Absence," *Manchester School*, 69, pp. 269–275.

Beck, A.J., and B.E. Shipley (1989): "Recidivism of Prisoners Released in 1983," Special report, Bureau of Justice Statistics.

Ben-Akiva, M., and S.R. Lerman (1985): *Discrete Choice Analysis: Theory and Application to Travel Demand*, MIT Press, Cambridge, MA.

Berger, M.C. (1988): "Predicted Future Earnings and Choice of College Major," *Industrial and Labor Relations Review*, 41, pp. 418–429.

Berkson, J. (1944): "Application of the Logistic Function to Bio-Assay," *Journal of the American Statistical Association*, 39, pp. 357–365.

Berndt, E.K., B.H. Hall, R.E. Hall, and J.A. Hausman (1974): "Estimation and Inference in Nonlinear Structural Models," *Annals of Economic and Social Measurement* 3/4, 653-665.

Berndt, E.R. (1990): *The Practice of Econometrics*, Addison-Wesley.

Bertrand, B., and S. Mullainathan (2004): "Are Emily and Brendan More Employable than Latoya and Tyrone? Evidence on Racial Discrimination in the Labor Market from a Large Randomized Experiment," *American Economic Review*, 94, pp. 991–1013.

Block, H., and J. Marschak (1960): "Random Orderings and Stochastic Theories of Response," in: I. Olkin (ed.), *Contributions To Probability And Statistics*, Stanford University Press, Stanford.

Bolduc, D. (1999): "A Practical Technique to Estimate Multinomial Probit Models in Transportation," *Transportation Research Part B*, 33, pp. 63–79.

Borjas, G.J. (1987): "Self-Selection and the Earnings of Immigrants," *American Economic Review* 77(4), pp. 531–53.

Borjas, G.J. (1999): "Immigration and Welfare Magnets," *Journal of Labor Economics*, 17, pp. 607-637.

Burtless, G. (1995): "The Case for Randomized Field Trials in Economic and Policy Research," *Journal of Economic Perspectives*, 9, pp. 63–84.

Cameron, A.C., and P.K. Trivedi (1998): *Regression Analysis of Count Data*, Cambridge University Press, New York.

Cappellari, L., and S.P. Jenkins (2003): "Multivariate Probit Regression Using Simulated Maximum Likelihood," *Stata Journal*, 3, pp. 278-294.

Checkovich, T., and S. Stern (2002): "Shared caregiving responsibilities of adult siblings with elderly parents", *Journal of Human Resources*, 37, pp. 441–478.

Collett, D. (2003): *Modelling Survival Data in Medical Research*, 2nd ed., Chapman & Hall, London.

Costa, D.L. (1995): "Pensions and Retirement: Evidence from Union Army Veterans," *Quarterly Journal of Economics*, 110, pp. 297–319.

Cox, D.R. (1972): "Regression Models and Life-Tables (with discussion)," *Journal of the Royal Statistical Society, Series B*, 34, pp. 187–220.

Cox, D.R. (1975): "Partial Likelihood," *Biometrika*, 62, pp. 269–276.

Cox, D.R., and D. Oakes (1984): *Analysis of Survival Data*, Chapman & Hall, London.

Cragg, J.G. (1971): "Some statistical models for limited dependent variables with application to the demand for durable goods," *Econometrica*, 39, pp. 829–844.

Cramer, J.S. (1986): *Econometric Applications of Maximum Likelihood Methods*, Cambridge University Press.

Cutler D., and E. Richardson (1997): "Measuring the Health of the U. S. Population", *Brookings Papers on Economic Activity, Microeconomics*, pp. 217–271.

Cutler D., and E. Richardson (1998): "The Value of Health: 1970-1990," *American Economic Review*, 88, pp. 97–100.

Davidson, R., and J.G. MacKinnon (1993): *Estimation and Inference in Econometrics*, Oxford University Press.

DeGroot, M.H. (1986): *Probability and Statistics*, 2nd ed. , Addison-Wesley.

Duan, N., W.G. Manning, C.N. Morris, and J.P. Newhouse (1983): "A comparison of alternative estimators for the demand for medical care," *Journal of Business and Economics Statistics* 1, pp. 115–126.

Ederington, L.H. (1985): "Classification Models and Bond Ratings," *Financial Review*, 20, pp. 237–262.

Ermisch, J.F., and N. Ogawa (1994): "Age at Motherhood in Japan," *Journal of Population Economics*, 7, pp. 393–420.

Evans, W., and R. Schwab (1995): "Finishing High School and Starting College: Do Catholic Schools Make a Difference?," *Quarterly Journal of Economics*, 110, pp. 941–974.

Fehr, E., and L. Goette (2004): "Do Workers Work More When Wages Are High?" *IZA Discussion Paper No. 1002.*

Fienberg, S.E. (1980): *The Analysis of Cross-Classified Categorical Data.* MIT Press, Cambridge, MA.

Finney, D.J. (1971): *Probit Analysis*, 3rd ed., Cambridge University Press.

Frey, B.S., and A. Stutzer (2002): *Happiness and Economics: How the Economy and Institutions Affect Human Well-Being*, Princeton University Press.

Genesove D., and C.J. Mayer (1997): "Equity and Time to Sale in the Real Estate Market," *American Economic Review*, 87, pp. 255–269.

Genz, A. (1992): "Numerical Computation of Multivariate Normal Probabilities," *Journal of Computational and Graphical Statistics*, 1, pp. 141–149.

Gouriéroux, C., A. Monfort, and A. Trognon (1984): Pseudo-Maximum Likelihood Methods: Theory, *Econometrica*, 52, pp. 681–700.

Gouriéroux, C., and A. Monfort (1996): *Simulation-based Econometric Methods*, Oxford University Press, Oxford.

Greene, W.H. (2003): *Econometric Analysis*, 5th ed., Prentice-Hall.

Gronau, R. (1974): "Wage comparisons - a Selectivity Bias," *Journal of Political Economy* 82, pp. 1119–1143.

Gurland, J., I. Lee, and P.A. Dahm (1960): "Polychotomous Quantal Response in Biological Assay," *Biometrics*, 16, pp. 382–398.

Gustafsson, S.S. (2001): "Optimal Age at Motherhood. Theoretical and Empirical Considerations on Postponement of Maternity in Europe," *Journal of Population Economics*, 14, pp. 225–247.

Gustafsson, S.S., E. Kenjoh, and C.M.M.P. Wetzels (2002): "The role of Education in Postponement of Maternity in Britain, Germany, The Netherlands and Sweden," in: Ruspini, E., and A. Dale (eds.), *The Gender Dimension of Social Change: The Contribution of Dynamic Research to the Study of Women's Life Courses*, The Policy Press, Bristol, pp. 55–79.

Hamilton, V., and B. Hamilton (1997): "Alcohol and earnings: Does drinking yield a wage premium?," *Canadian Journal of Economics*, 30, pp. 135–151.

Hanemann, W.M. (1984): "Welfare Evaluations in Contingent Valuation experiments with Discrete Responses," *American Journal of Agricultural Economics*, 66, pp. 332–341.

Hartog, J., and R. Winkelmann (2003): "Comparing migrants to non-migrants: The case of Dutch migration to New Zealand," *Journal of Population Economics*, 16, pp. 683-705.

Hausman, J.A., B.H. Hall, and Z. Griliches (1984): "Econometric Models for Count Data with an Application to the Patents-R&D Relationship," *Econometrica*, 52, pp. 909–938.

Hausman, J.A., and D.L. McFadden (1986): "A Specification Test for the Multinomial Logit Model," *Econometrica*, 52, pp. 1219–1240.

Heckman, J.J. (1974): "Shadow Prices, Market Wages, and Labor Supply," *Econometrica*, 42, pp. 679–694.

Heckman, J.J. (1979): "Sample Selection as a Specification Error, *Econometrica*, 47, pp. 153–161.

Heckman, J.J., J.L. Tobias, and E. Vytlacil (2003): "Simple Estimators for Treatment Parameters in a Latent Variable Framework," *Review of Economics and Statistics*, 85, pp. 748–755.

Heckman, J.J., R.J. Lalonde, and J.A. Smith (1999): "The Economics and Econometrics of Active Labor Market Programs," in: O.C. Ashenfelter, and D. Card (eds.) *Handbook of Labor Economics*, Volume 3A, Elsevier North-Holland, Chapter 31.

Heckman, J.J., and B. Honoré (1990): "The empirical content of the Roy model," *Econometrica*, 58, pp. 1121–1149.

Heckman, J.J., and E. Vytlacil (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 73, pp. 669–738.

Heckman, J.J., and J.A. Smith (1995): "Social Experiments Assessing the Case for Social Experiments," *Journal of Economic Perspectives*, 9, pp. 85–110.

Heckman, J.J., and R. Robb (1985): "Alternative Methods for Estimating The Impact of Interventions," in: J. Heckman, and B. Singer (eds.), *Longitudinal Analysis of Labor Market Data*, Cambridge (USA).

Hoffmann, S.D., and G.J. Duncan (1988): "Multinomial and Conditional Logit Discrete-Choice Models in Demography," *Demography*, 25, pp. 415–427.

Hogg, R.V., and A.T. Craig (1989): *Introduction to Mathematical Statistics*, 4th ed., Macmillan.

Holland, P.W. (1986): "Statistics and Causal Inference," *Journal of the American Statistical Association*, 81, pp. 945–960.

Horowitz, J.L. (1998): *Semiparametric Methods in Econometrics*, Lecture Notes in Statistics Vol.131, Springer.

Hsiao, C. (2003): *Analysis of Panel Data*, Econometric Society Monographs No.34, Cambridge University Press.

Jenkins, S.P. (1995): "Easy Ways to Estimate Discrete Time Duration Models," *Oxford Bulletin of Economics and Statistics*, 57, pp. 129–138.

Kaiser, U., and A. Spitz (2002): "Quantification of Qualitative Data Using Ordered Probit Models," in: G. Poser, and D. Bloesch (eds.), *Economic Survey and Data Analysis, CIRET Conference Proceedings* Paris, pp. 325–343.

Kalbfleisch, J.D., and R.L. Prentice (2002): *The Statistical Analysis of Failure Time Data*, 2nd ed., Wiley, New York.

Kennan, J. (1985): "The Duration of Contract Strikes in U.S. Manufacturing," *Journal of Econometrics*, 28, pp. 5–28.

Kiefer, N.M. (1988): "Economic Duration Data and Hazard Functions," *Journal of Economic Literature*, 26, pp. 646–679.

Killingsworth, M. (1983): *Labor Supply*, Cambridge University Press.

Klein, J.P., and M.L. Moeschberger (2003): *Survival Analysis: Techniques for Censored and Truncated Data*, 2nd ed., Springer, New York.

Kockelman, K.M., and Y.-J. Kweon (2002): "Driver injury severity: an application of ordered probit models," *Accident Analysis & Prevention*, 34, pp. 313–321.

Läärä, E., and J.N. Matthews (1985): "The Equivalence of two models for ordinal data," *Biometrika*, 72, pp. 206–207.

Lancaster, T. (1979): "Econometric Methods for the Duration of Unemployment," *Econometrica*, 47, pp. 939–956.

Lancaster, T. (1990): *The Econometric Analysis of Transition Data*, Cambridge University Press.

Layard, R. (2005): *Happiness: Lessons from a New Science*, Penguin Press.

Lee, L.-F. (1978): "Unionism and wage rates: a simultaneous equations model with qualitative and limited dependent variables," *International Economic Review*, 19, pp. 415–433.

Luce, R. D. (1959): *Individual Choice Behavior*, Wiley, New York.

Luce, R. D., and P. Supes (1965): "Preference, Utility, and Subjective Probability," in: R. Luce, R. Bush, and E. Galanter (eds.), *Handbook Of Mathematical Psychology*, Wiley, New York.

Maddala, G.S. (1983): *Limited Dependent and Qualitative Variables in Econometrics*, Cambridge University Press.

Manski, C.F. (1995): *Identification Problems in the Social Sciences*, Harvard University Press.

Marschak, J. (1960): "Binary Choice Constraints on Random Utility Indicators," in: K. Arrow (ed.), *Stanford Symposium On Mathematical Methods In The Social Sciences*, Stanford University Press, Stanford.

McCloskey, D.N. (1985): The Loss Function Has Been Mislaid: The Rhetoric of Significance Tests, *American Economic Review*, 75, 201-205.

McCullagh, P. (1980): "Regression Models for Ordinal Data," *Journal of the Royal Statistical Society. Series B (Methodological)*, 42, pp. 109–142.

McFadden, D.L. (1968): "The Revealed Preferences of a Public Bureaucracy," Department of Economics, University of California, Berkeley.

McFadden, D.L. (1974a): "Conditional Logit Analysis of Qualitative Choice Behavior," in: P. Zarembka (ed.), *Frontiers In Econometrics*, Academic Press, New York, pp. 105–142.

McFadden, D.L. (1974b): "The Measurement of Urban Travel Demand," *Journal of Public Economics*, 3, pp. 303–328.

McFadden, D.L. (1981): "Econometric Models of Probabilistic Choice," in: C.F. Manski, and D. McFadden (eds.), *Structural Analysis of Discrete Data with Econometric Applications*, MIT Press.

McFadden, D.L. (1984): "Econometric Analysis of Qualitative Response Models," in: Z. Griliches and M. Intrilligator (eds.), *Handbook of Econometrics*, Vol. II, Elsevier, Amsterdam, pp. 1396–1456.

McKelvey, R.D. and W. Zavoina (1975): "A statistical model for the analysis of ordinal level dependent variables," *Journal of Mathematical Sociology*, 4, pp. 103–120.

Moffitt, R. (1991): Program Evaluation with Nonexperimental Data, *Evaluation Review*, 15:291-314.

Morgan, S.P. (1996): "Characteristic Features of Modern American Fertility," *Population and Development Review*, 22, *Supplement: Fertility in the United States: New Patterns, New Theories*, pp. 19-63.

Mroz, T.A. (1987): "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions," *Econometrica*, 55, pp. 765–799

Olsen, R.J. (1978): "Note on the Uniqueness of the Maximum Likelihood Estimator for the Tobit Model," *Econometrica*, 46, pp. 1211–1215.

Paap, R., and P.H. Franses (2000): "A Dynamic Multinomial Probit Model for Brand Choice with Different Long-Run Effects of Marketing-Mix Variables," *Journal of Applied Econometrics*, 15, pp. 717–744.

Pagan, A., and A. Ullah (1999): *Nonparametric Econometrics*, Cambridge University Press.

Pohlmeier, W., and V. Ulrich (1995): "An Econometric Model of the Two-Part Decision Making Process in the Demand for Health Care," *Journal of Human Resources*, 30, pp. 339–361.

Powell, J.L. (1984): "Least squares absolute deviations estimation for the censored regression model," *Journal of Econometrics*, 25, pp. 303–325.

Rao, C.R. (1948) Large sample tests of hypotheses involving several parameters with applications to problems of estimation, *Proceedings of the Cambridge Philosophical Society* 44, 50-57.

Rose, N.L. (1990): "Profitability and Product Quality: Economic Determinants of Airline Safety Performance," *The Journal of Political Economy*, 98, pp. 944–964.

Rothenberg, T.J. (1971): Identification in parametric models, *Econometrica*, 39, 577-591.

Roy, A.D. (1951): "Some Thoughts on the Distribution of Earnings," *Oxford Economic Papers*, 3, pp. 135–146.

Rubin, D.B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Education Psychology*, 66, pp. 688–701.

Sander, W. (1992): "The Effect of Women's Schooling on Fertility," *Economics Letters*, 40, pp. 229-233.

Schmidt, P., and R.P. Strauss (1975): "The Prediction of Occupation Using Multiple Logit Models," *International Economic Review*, 16, pp. 471–486.

Schultz, T.P. (2004): "School Subsidies for the Poor: Evaluating the Mexican Progresa Poverty Program," *Journal of Development Economics*, 74, pp. 199–250.

Smith, R., and R. Blundell (1986): "An Exogeneity Test for a Simultaneous Equation Tobit Model with an Application to Labor Supply," *Econometrica*, 54, pp. 679–686.

Solon, G. (1999): "Intergenerational Mobility in the Labor Market," in: O.C. Ashenfelter, and D. Card (eds.) *Handbook of Labor Economics*, Volume 3A, Elsevier North-Holland, pp. 1761-1800.

Spencer, D.E. (1985): "Money demand and the price level," *Review of Economics and Statistics*, 67(3), 490-496.

Stafford, S.L. (2000): "The Impact of Environmental Regulations on the Location of Firms in the Hazardous Waste Management Industry," *Land Economics*, 76, pp. 569–589.

Terza, J.V. (1985): "Ordinal Probit: A Generalization," *Communications in Statistics – Theory and Methods*, 14, pp. 1–11.

Terza, J.V. (1998): "Estimating Count Data Models with Endogenous Switching: Sample Selection and Endogenous Treatment Effects," *Journal of Econometrics*, 84, pp. 129–154.

Terza, J.V. (2002): "Alcohol Abuse and Employment: A Second Look," *Journal of Applied Econometrics*, 17, pp. 393–404.

Theil, H. (1969): "A Multinomial Extension of the Linear Logit Model," *International Economic Review*, 10, pp. 251–259.

Theil, H. (1970): "On the Estimation of Relationships Involving Qualitative Variables," *American Journal of Sociology*, 76, pp. 103–154.

Thurstone, L. (1927): "A Law of Comparative Judgment," *Psychological Review*, 34, pp. 273–286.

Tobin, J. (1958): "Estimation of Relationships for Limited Dependent Variables," *Econometrica*, 26, pp. 24–36.

Train, K.E. (2003): *Discrete Choice Methods with Simulation*, New York, Cambridge University Press.

Tutz, G. (1991): "Sequential Models in Categorical Regression," *Computational Statistics and Data Analysis*, 11, pp. 275–295.

Van den Berg (2001): "Duration Models: Specification, Identification, and Multiple Durations," in: J.J. Heckman, and E. Leamer (eds.) *Handbook of Econometrics*, Volume V, North-Holland, Amsterdam.

Wald, A. (1943) Tests for statistical hypotheses concerning several parameters when the number of observations is large, *Transactions of the American Mathematical Society* 54, 426-482.

White, H. (1982): Maximum Likelihood Estimation of Misspecified Models, *Econometrica*, 50, 1-25.

Willis, R.J. (1973): "A New Approach to the Economic Theory of Fertility Behavior," *Journal of Political Economy*, 81, pp. S14–S64. Reprinted in: T.W. Schultz (ed.) (1974): *Economics of the Family: Marriage, Children and Human Capital*, University of Chicago Press, Chicago and London.

Willis, R.J., and S. Rosen (1979): "Education and Self-Selection," *Journal of Political Economy*, 87, pp. S7–36.

Winkelmann, R. (1998): "Count Data Models with Selectivity," *Econometric Reviews*, 17, pp. 339–359.

Winkelmann, R. (2003): *Econometric Analysis of Count Data*, 4th ed., Springer Verlag, Berlin.

Winkelmann, R. (2004): "Health Care Reform and the Number of Doctor Visits - An Econometric Analysis," *Journal of Applied Econometrics*, 19, pp. 455–472.

Winkelmann, R., and K.F. Zimmermann (1994): "Count Data Models for Demographic Data," *Mathematical Population Studies*, 4, pp. 205–221.

Winship, C., and R.D. Mare (1984): "Regression Models with Ordinal Variables," *American Sociological Review*, 49, pp. 512–525.

Wooldridge, J.M. (2002): *Econometric Analysis of Cross Section and Panel Data*, MIT Press.

Yatchew, A., and Z. Griliches (1985): "Specification error in probit models," *Review of Economics and Statistics*, 67, pp. 134–139.

Index

accelerated failure time (AFT) model, 229, 270, 275
Akaike information criterion (AIC), 88
attenuation bias, 222
auxiliary parameters, 24
average marginal mean effect (AMME), 25
average marginal probability effect (AMPE), 104, 147, 182
average treatment effect (ATE), 239, 242, 244
average treatment effect on the treated (ATT), 239, 244
axioms of probability, 29

base category, 140, 161
baseline hazard, 269
Bernoulli distribution, 35, 47, 51, 55, 58, 62, 97
best linear predictor, 221, 227
best linear unbiased estimator, 8, 63, 76, 226
binary response models, 95–135
bivariate normal distribution, 162, 232, 233, 235, 238, 242, 246

categorical data, 4
censored regression model, 226
censoring, 7, 209, 224, 260, 263, 289
choice-specific attributes, 107, 150, 153
compensating variation, 28, 179
complementary log-log transformation, 130
conditional distribution, 32
conditional expectation function (CEF)
  and limited dependent variables, 38
  basic concepts, 21–25
  Bernoulli distribution, 24, 97
  bivariate normal distribution, 246
  Heckman model, 234
  identification, 75
  linear regression model, 8
  Poisson regression model, 280
  quasi-likelihood estimation, 76
  Tobit model, 218
conditional logit (CL) model, 150–160
conditional probability, 32
conditional probability function (CPF)
  and limited dependent variables, 38
  basic concepts, 21–27
conditional probability model
  Bernoulli distribution, 24, 35, 97–102
  exponential distribution, 38, 262
  likelihood function, 50
  log-normal distribution, 262
  multinomial distribution, 36, 140, 179
  negative binomial distribution, 285
  normal distribution, 39
  normal linear model, 63–66
  Poisson distribution, 37, 280
  specification, 30, 34
conditional transition probability, 196
conditional transitions, 194
consistency, 55
contingent valuation, 109
convergence criterion, 72
convergence in distribution, 59

corner solution, 6
corner solution model, 208, 211–223
count data models, 252, 279–294
covariance matrix, 61
covariance matrix estimator
  actual Hessian, 61
  expected Hessian, 61
  outer product of the score, 61
Cox proportional hazards model, 270
cracker brand choice, 153, 157
Cragg model, 223, 230, 248
Cramér Rao lower bound, 57
cross-sectional data, 1
cumulative density function
  multivariate, 33
  univariate, 31
cumulative distribution function
  multivariate, 32
  univariate, 31

Delta method, 68, 119
density function
  multivariate, 33
  univariate, 31
dependence, 32
determinants of fertility, 12–15, 21, 84, 99, 111, 119, 126, 228, 255, 275, 286
difference-in-differences estimator, 250
discrete choice models, 107, 151
discrete mean effect, 25, 281
discrete probability effect, 27, 37, 105, 146, 180
duration analysis, 251, 254
duration analysis and count data, 283
duration dependence, 259, 267
duration models, 254–278

early retirement, 260
empirical survivor function, 257
endogeneity, 116–118, 242–243
equidispersion, 279
estimate, 46
estimator, 46
event history, 251
event history models, 251–297
exclusion restriction, 235, 236
exit rate, 255
expected score, 54
experimental data, 3, 240
exponential distribution, 31, 38, 52, 91, 262, 264, 284
exponential model, 262, 264, 266
extensive margin, 220
extreme value distribution, 108, 151, 164, 165, 195, 202, 270

female hours of work and wages, 17–19, 213, 222, 236
Fisher information matrix, 56
flow sampling, 263
frailty, 271

gamma distribution, 272, 294
gamma function, 272
Gauss-Markov assumptions, 8
generalized extreme value distribution, 166
generalized least squares (GLS), 98, 130
generalized linear model, 25
generalized method of moments, 89
generalized ordered logit model, 188
generalized ordered probit model, 188
generalized threshold models, 188–194
geometric distribution, 92
Gompertz distribution, 268
goodness-of-fit
  R-squared, 89
  proportion of correct predictions, 122
  pseudo R-squared, 89, 122
Gronau/Heckman/Roy wage model, 233
Gumbel distribution, 108

happiness and income, 191
Hausman specification test, 160
hazard function
  continuous time, 265, 267
  discrete time, 259
hazard rate, 255, 256
hazard ratio, 269
Heckman model, 232–237
Heckman two-step method, 234
Hessian matrix, 48–51, 57, 65, 69, 74, 112, 114, 129, 141, 155, 282
heterogeneity, 271, 284
heteroscedasticity, 9, 39, 98, 115, 163, 218, 281
hierarchical response model, 164
hurdle count data model, 291

hurdle-at-zero Poisson model, 291
hybrid model of MNL and CL, 154

identification, 57, 74, 101, 108, 139, 154, 161, 176, 201, 225, 235
  by exclusion restriction, 75
  by functional form, 75
  in probability models, 75
in-sample, 123
incidental censoring, 231
independence, 32
independence of irrelevant alternatives, 159
independent sampling, 45
indicator function, 101
individual-specific characteristics, 107, 150, 153, 154
information matrix, 57
information matrix equality, 57
instrumental variables, 117, 235, 241, 243
integrated hazard function, 266
intensive margin, 220
interactive effects, 26
interior solution, 6
interval data, 200
invariance property, 67, 118
inverse matrix, 64
inverse Mills ratio, 215
iso-probability curve, 28, 187

joint cumulative density function, 33
joint cumulative distribution function, 32
joint density function, 33
joint probability function, 24, 32, 45

Kaplan-Meier estimator, 257

latent variable, 101, 151, 174, 200, 212, 242
least squares, 8, 24, 63, 64, 89, 98
lifetable, 256
lifetable method, 259
likelihood function, 24, 46–53, 179, 212
likelihood ratio test, 83–86
limited dependent variables, 207–250
linear index, 9, 24, 25, 97, 107, 139, 152, 174, 187, 195
linear probability model (LPM), 9, 98
linear regression model, 3, 8–10, 23–26, 63, 98, 113, 173, 211, 221, 226, 230, 237, 246
linear sequential model, 196
local average treatment effect (LATE), 244
log-concavity, 216
log-likelihood function, 47, 63, 110, 141, 155, 179, 228, 230, 235, 282
log-linear normal model, 71
log-logistic distribution, 269
log-normal distribution, 67, 71, 201, 269
logistic distribution, 177
logit model, 102, 259
Luce's model, 151

macrodata, 3
macroeconometrics, 3
marginal distribution, 32, 33
marginal mean effect (MME), 25, 220, 281
marginal probability, 32
marginal probability effect (MPE), 19, 27, 104, 119, 147, 156, 182, 187, 190, 197, 220, 282
marginal treatment effect (MT), 244
maximum likelihood estimation, 45–93
  asymptotic efficiency, 57
  asymptotic normality, 59
  consistency, 55
  covariance matrix, 61
  definition, 24, 50
  inference, 76–89
  properties in large samples, 53
  properties in small samples, 53
maximum simulated likelihood, 70, 164
mean independence, 239
method of moments, 60, 89, 112
microdata, 1, 3, 4, 10
microeconometrics, 1, 3
mixed logit (MXL) model, 163–164
model selection, 88
multinomial distribution, 36, 140, 179
multinomial logit (MNL) model, 139–150
multinomial probit (MNP) model, 161–163
multinomial response models, 137–169

multivariate normal distribution, 161, 246

negative binomial distribution, 286
Negbin regression model, 286
nested alternatives, 164
nested logit model, 164–166
nested models, 88
Newton-Raphson algorithm, 71
non-linear least squares, 9, 89
nonnested models, 88
nonparametric models, 11
normal density function, 38
normal distribution, 100
  truncated density, 214
  truncated moments, 214, 216, 218
normal linear model, 38, 63–67, 76, 85, 101, 221, 226, 230
normalization, 101, 108, 115, 140, 152, 154, 163, 176, 238, 272, 286
numerical integration, 70
numerical optimization, 69–74
  BHHH algorithm, 74
  derivative free algorithms, 74
  method of scoring, 74
  Newton-Raphson algorithm, 71
  quadratic hill-climbing method, 74
  steepest ascent method, 74

observational data, 3, 241
observational equivalence, 75, 101, 178, 195, 202
observations at risk, 200, 259
odds, 106, 144, 155, 159, 177, 183, 191
odds ratio, 106, 118, 144, 183
omitted variables, 117, 221
optimization, 69
  analytical, 69
  graphical, 69
  numerical, 69–74
  trial and error, 69
ordered logit model, 177
ordered probit model, 176
ordered response models, 171–205
ordinary least squares (OLS), 8, 24, 63, 64, 76, 89, 90, 101, 112, 118, 173, 221, 223, 226, 227, 232, 234, 241
out-of-sample, 123
outcome equation, 232
outcome-specific parameters, 139, 244
overdispersion, 279, 285

panel data, 1
parallel regression assumption, 186–187
parameter constraints, 67
parameters, 23–28
parametric models, 11
Pareto distribution, 91
partial likelihood methods, 270
partial observability, 224
perfect prediction, 51, 113, 125
piecewise linear hazard function, 270
Poisson distribution, 37, 87, 279, 284
Poisson regression model, 279, 280
power of a test, 78
predicted probabilities, 119, 147, 158, 179, 288
probability function
  multivariate, 32
  univariate, 30
probability measure, 29
probability model, 29, 75, 76
probit model, 100, 259
product-limit estimator, 257
proportional hazards (PH) model, 195, 269
proportional odds model, 177
proportionality, 46

qualitative data, 4
  binary, 4, 95–135
  multinomial, 5, 137–169
  ordered, 5, 171–205
quantitative data, 4, 6
  censored, 7, 209, 262, 289
  counts, 8, 251, 279–294
  durations, 7, 251, 254–278
  non-negative, 6, 208, 211–223
  truncated, 6, 209, 289
quasi-likelihood estimation, 76, 289

random sample, 24, 32, 46
random utility maximization, 107, 151
random variable, 29
  multivariate, continuous, 33
  multivariate, discrete, 31
  univariate, continuous, 31
  univariate, discrete, 30
randomized experiment, 240

rare events, 127
receiver operating characteristic (ROC) curve, 124
reduced-form model, 18, 213, 238, 245
regression parameters, 23
regular density function, 54
relative marginal probability effects, 187
residual, 24
restricted maximum likelihood, 79–81
restrictions, 77
retrospective information, 255
robust estimation, 289

sample selection models, 209, 224–238
sample space, 29
scalar product, 9
scale category, 163
Schwarz information criterion (SIC), 88
score function, 48, 50, 64, 110, 141, 155, 282
score test, 86–87
secondary school choice, 16–17, 142, 148, 183
selection equation, 232
selection of dependent and independent variables, 1
self-selection, 231
self-selection model, 232–237
sensitivity, 125
sequential logit model, 195
sequential mechanism, 194
sequential models, 194–200
sequential probit model, 195
single crossing property, 187, 282
single index assumption, 186–188
specificity, 125
spells, 252
stock sampling, 262
stratified sampling, 127–129
structural labor supply model, 237
structural model, 18, 213
sum of squared residuals, 24
survival analysis, 251
survival time, 254
  continuous, 254, 262
  discrete, 254, 259
survivor function, 255, 265
  Kaplan-Meier estimator, 257
switching regression model, 243

theory-based analysis, 2
theory-building analysis, 2
threshold mechanism, 174, 188
time ratio, 270
time-invariant regressor, 262
tobit model, 212, 218, 223
top-coding, 7, 209, 224
trade-off ratio, 180
treatment effect models, 210, 239–245
truncated regression model, 230
truncation, 6, 209, 224, 289
two-part model, 223, 291
two-stage least squares (2SLS), 243
type-I extreme value distribution, 108, 151, 164, 195, 202

underdispersion, 279
union/non-union wage differentials, 245
unobserved heterogeneity, 70, 271–278, 284–288
utility maximization, 107

wage function, 78
Wald test, 81–83, 130
Weibull distribution, 78, 130, 268
willingness-to-pay, 109

zero-inflated count data model, 293