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 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.

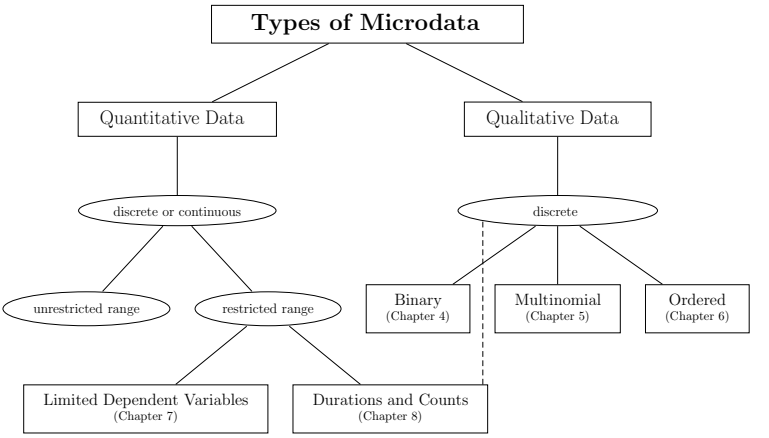
1.2.1 Qualitative Data

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.