Longitudinal and Panel Data

Analysis and Applications in the Social Sciences

EDWARD W. FREES

University of Wisconsin–Madison

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011-4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

First published 2004

Printed in the United States of America

Typeface Times Roman 10/13 pt. *System* $\LaTeX\ 2_{\varepsilon}$ [TB]

A catalog record for this book is available from the British Library.

Library of Congress Cataloging in Publication Data

Frees, Edward W.

Longitudinal and panel data : analysis and applications in the social
sciences / Edward W. Frees.

p. cm.

Includes bibliographical references and index.

ISBN 0-521-82828-7 – ISBN 0-521-53538-7 (pb.)

1. Social sciences – Research – Statistical methods.
2. Longitudinal method. 3. Panel analysis.
4. Social sciences – Mathematical models.
5. Social sciences – Statistical methods. I. Title.

HA29.F6816 2004
$300'.72'7$ – dc22 2004043583

ISBN 0 521 82828 7 hardback
ISBN 0 521 53538 7 paperback

Contents

1
Introduction

Abstract. This chapter introduces the many key features of the data and models used in the analysis of longitudinal and panel data. Here, longitudinal and panel data are defined and an indication of their widespread usage is given. The chapter discusses the benefits of these data; these include opportunities to study dynamic relationships while understanding, or at least accounting for, cross-sectional heterogeneity. Designing a longitudinal study does not come without a price; in particular, longitudinal data studies are sensitive to the problem of attrition, that is, unplanned exit from a study. This book focuses on models appropriate for the analysis of longitudinal and panel data; this introductory chapter outlines the set of models that will be considered in subsequent chapters.

1.1 What Are Longitudinal and Panel Data?

Statistical Modeling

Statistics is about data. It is the discipline concerned with the collection, summarization, and analysis of data to make statements about our world. When analysts collect data, they are really collecting information that is quantified, that is, transformed to a numerical scale. There are many well-understood rules for reducing data, using either numerical or graphical summary measures. These summary measures can then be linked to a theoretical representation, or model, of the data. With a model that is calibrated by data, statements about the world can be made.

As users, we identify a basic entity that we measure by collecting information on a numerical scale. This basic entity is our *unit of analysis*, also known as the *research unit* or *observational unit*. In the social sciences, the unit of analysis is typically a person, firm, or governmental unit, although other applications can and do arise. Other terms used for the observational unit include *individual*, from the econometrics literature, as well as *subject*, from the biostatistics literature.

Regression analysis and *time-series analysis* are two important applied statistical methods used to analyze data. Regression analysis is a special type of multivariate analysis in which several measurements are taken from each subject. We identify one measurement as a *response*, or *dependent variable*; our interest is in making statements about this measurement, controlling for the other variables.

With regression analysis, it is customary to analyze data from a cross section of subjects. In contrast, with *time-series analysis*, we identify one or more subjects and observe them over time. This allows us to study relationships over time, the *dynamic* aspect of a problem. To employ time-series methods, we generally restrict ourselves to a limited number of subjects that have many observations over time.

Defining Longitudinal and Panel Data

Longitudinal data analysis represents a marriage of regression and time-series analysis. As with many regression data sets, *longitudinal data* are composed of a cross section of subjects. Unlike regression data, with longitudinal data we observe subjects over time. Unlike time-series data, with longitudinal data we observe many subjects. Observing a broad cross section of subjects over time allows us to study dynamic, as well as cross-sectional, aspects of a problem.

The descriptor *panel data* comes from surveys of individuals. In this context, a "panel" is a group of individuals surveyed repeatedly over time. Historically, panel data methodology within economics had been largely developed through labor economics applications. Now, economic applications of panel data methods are not confined to survey or labor economics problems and the interpretation of the descriptor "panel analysis" is much broader. Hence, we will use the terms "longitudinal data" and "panel data" interchangeably although, for simplicity, we often use only the former term.

Example 1.1: Divorce Rates Figure 1.1 shows the 1965 divorce rates versus AFDC (Aid to Families with Dependent Children) payments for the fifty states. For this example, each state represents an observational unit, the divorce rate is the response of interest, and the level of AFDC payment represents a variable that may contribute information to our understanding of divorce rates.

The data are observational; thus, it is not appropriate to argue for a causal relationship between welfare payments (AFDC) and divorce rates without invoking additional economic or sociological theory. Nonetheless, their relation is important to labor economists and policymakers.

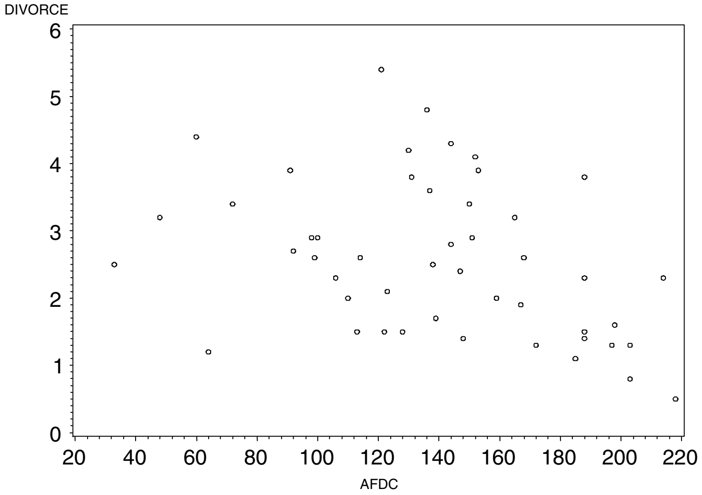

Figure 1.1. Plot of 1965 divorce rates versus AFDC payments. (*Source: Statistical Abstract of the United States.*)

Figure 1.1 shows a negative relation; the corresponding correlation coefficient is −.37. Some argue that this negative relation is counterintuitive in that one would expect a positive relation between welfare payments and divorce rates; states with desirable economic climates enjoy both a low divorce rate and low welfare payments. Others argue that this negative relationship is intuitively plausible; wealthy states can afford high welfare payments and produce a cultural and economic climate conducive to low divorce rates.

Another plot, not displayed here, shows a similar negative relation for 1975; the corresponding correlation is −.425. Further, a plot with both the 1965 and 1975 data displays a negative relation between divorce rates and AFDC payments.

Figure 1.2 shows both the 1965 and 1975 data; a line connects the two observations within each state. These lines represent a change over time (dynamic), not a cross-sectional relationship. Each line displays a positive relationship; that is, as welfare payments increase so do divorce rates for each state. Again, we do not infer directions of causality from this display. The point is that the dynamic relation between divorce and welfare payments within a state differs dramatically from the cross-sectional relationship between states.

Some Notation

Models of longitudinal data are sometimes differentiated from regression and time-series data through their double subscripts. With this notation, we may

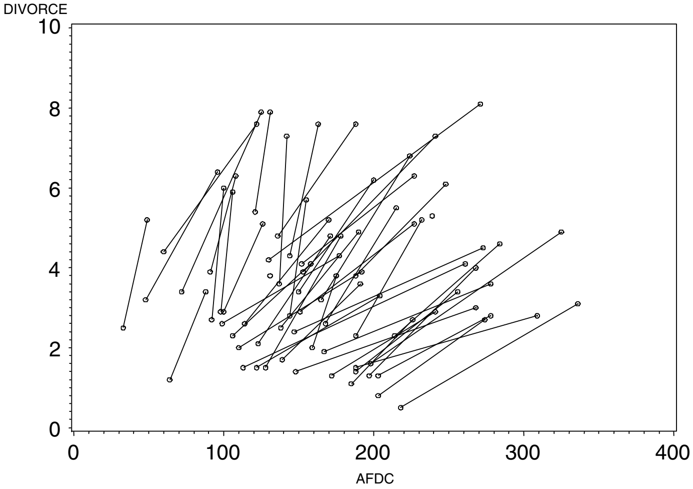

Figure 1.2. Plot of divorce rate versus AFDC payments from 1965 and 1975.

distinguish among responses by subject and time. To this end, define y_{it} to be the response for the ith subject during the tth time period. A longitudinal data set consists of observations of the ith subject over $t = 1, \ldots, T_i$ time periods, for each of $i = 1, \ldots, n$ subjects. Thus, we observe

$$\begin{gathered}\text{first subject} - \{y_{11}, y_{12}, \ldots, y_{1T_1}\}\\ \text{second subject} - \{y_{21}, y_{22}, \ldots, y_{2T_2}\}\\ \vdots\\ n\text{th subject} - \{y_{n1}, y_{n2}, \ldots, y_{nT_n}\}.\end{gathered}$$

In Example 1.1, most states have $T_i = 2$ observations and are depicted graphically in Figure 1.2 by a line connecting the two observations. Some states have only $T_i = 1$ observation and are depicted graphically by an open-circle plotting symbol. For many data sets, it is useful to let the number of observations depend on the subject; T_i denotes the number of observations for the ith subject. This situation is known as the *unbalanced data* case. In other data sets, each subject has the same number of observations; this is known as the *balanced data* case. Traditionally, much of the econometrics literature has focused on the balanced data case. We will consider the more broadly applicable unbalanced data case.

Prevalence of Longitudinal and Panel Data Analysis

Longitudinal and panel databases and models have taken on important roles in the literature. They are widely used in the social science literature, where panel data are also known as *pooled cross-sectional time series*, and in the natural sciences, where panel data are referred to as *longitudinal data*. To illustrate their prevalence, consider that an index of business and economic journals, ABI/INFORM, lists 326 articles in 2002 and 2003 that use panel data methods. Another index of scientific journals, the ISI Web of Science, lists 879 articles in 2002 and 2003 that use longitudinal data methods. Note that these are only the applications that were considered innovative enough to be published in scholarly reviews.

Longitudinal data methods have also developed because important databases have become available to empirical researchers. Within economics, two important surveys that track individuals over repeated surveys include the Panel Survey of Income Dynamics (PSID) and the National Longitudinal Survey of Labor Market Experience (NLS). In contrast, the Consumer Price Survey (CPS) is another survey conducted repeatedly over time. However, the CPS is generally not regarded as a panel survey because individuals are not tracked over time. For studying firm-level behavior, databases such as Compustat and CRSP (University of Chicago's Center for Research on Security Prices) have been available for over thirty years. More recently, the National Association of Insurance Commissioners (NAIC) has made insurance company financial statements available electronically. With the rapid pace of software development within the database industry, it is easy to anticipate the development of many more databases that would benefit from longitudinal data analysis. To illustrate, within the marketing area, product codes are scanned in when customers check out of a store and are transferred to a central database. These *scanner data* represent yet another source of data information that may inform marketing researchers about purchasing decisions of buyers over time or the efficiency of a store's promotional efforts. Appendix F summarizes longitudinal and panel data sets used worldwide.

1.2 Benefits and Drawbacks of Longitudinal Data

There are several advantages of longitudinal data compared with either purely cross-sectional or purely time-series data. In this introductory chapter, we focus on two important advantages: the ability to study dynamic relationships and to model the differences, or *heterogeneity*, among subjects. Of course, longitudinal data are more complex than purely cross-sectional or times-series data and so there is a price to pay in working with them. The most important drawback is the difficulty in designing the sampling scheme to reduce the problem of subjects leaving the study prior to its completion, known as *attrition*.

Dynamic Relationships

Figure 1.1 shows the 1965 divorce rate versus welfare payments. Because these are data from a single point in time, they are said to represent a *static* relationship.

For example, we might summarize the data by fitting a line using the method of least squares. Interpreting the slope of this line, we estimate a *decrease* of 0.95% in divorce rates for each $100 increase in AFDC payments.

In contrast, Figure 1.2 shows changes in divorce rates for each state based on changes in welfare payments from 1965 to 1975. Using least squares, the overall slope represents an *increase* of 2.9% in divorce rates for each $100 increase in AFDC payments. From 1965 to 1975, welfare payments increased an average of $59 (in nominal terms) and divorce rates increased 2.5%. Now the slope represents a typical time change in divorce rates per $100 unit time change in welfare payments; hence, it represents a *dynamic* relationship.

Perhaps the example might be more economically meaningful if welfare payments were in real dollars, and perhaps not (for example, deflated by the Consumer Price Index). Nonetheless, the data strongly reinforce the notion that dynamic relations can provide a very different message than cross-sectional relations.

Dynamic relationships can only be studied with repeated observations, and we have to think carefully about how we define our "subject" when considering dynamics. Suppose we are looking at the event of divorce on individuals. By looking at a cross section of individuals, we can estimate divorce rates. By looking at cross sections repeated over time (without tracking individuals), we can estimate divorce rates over time and thus study this type of dynamic movement. However, only by tracking repeated observations on a sample of individuals can we study the duration of marriage, or time until divorce, another dynamic event of interest.

Historical Approach

Early panel data studies used the following strategy to analyze pooled cross-sectional data:

- Estimate cross-sectional parameters using regression.
- Use time-series methods to model the regression parameter estimators, treating estimators as known with certainty.

Although useful in some contexts, this approach is inadequate in others, such as Example 1.1. Here, the slope estimated from 1965 data is −0.95%. Similarly, the slope estimated from 1975 data turns out to be −1.0%. Extrapolating these negative estimators from different cross sections yields very different results from the dynamic estimate: a positive 2.9%. Theil and Goldberger (1961E) provide an early discussion of the advantages of estimating the cross-sectional and time-series aspects simultaneously.

Dynamic Relationships and Time-Series Analysis

When studying dynamic relationships, univariate time-series analysis is a well-developed methodology. However, this methodology does not account for relationships among different subjects. In contrast, multivariate time-series analysis does account for relationships among a limited number of different subjects. Whether univariate or multivariate, an important limitation of time-series analysis is that it requires several (generally, at least thirty) observations to make reliable inferences. For an annual economic series with thirty observations, using time-series analysis means that we are using the same model to represent an economic system over a period of thirty years. Many problems of interest lack this degree of stability; we would like alternative statistical methodologies that do not impose such strong assumptions.

Longitudinal Data as Repeated Time Series

With longitudinal data we use several (repeated) observations of many subjects. Repeated observations from the same subject tend to be correlated. One way to represent this correlation is through dynamic patterns. A model that we use is the following:

$$y_{it} = \mathrm{E}y_{it} + \varepsilon_{it}, \quad t = 1, \ldots, T_i, \;\; i = 1, \ldots, n, \tag{1.1}$$

where ε_{it} represents the deviation of the response from its mean; this deviation may include dynamic patterns. Further, the symbol E represents the expectation operator so that $\mathrm{E}y_{it}$ is the expected response. Intuitively, if there is a dynamic pattern that is common among subjects, then by observing this pattern over many subjects, we hope to estimate the pattern with fewer time-series observations than required of conventional time-series methods.

For many data sets of interest, subjects do not have identical means. As a first-order approximation, a linear combination of known, *explanatory* variables such as

$$\mathrm{E}y_{it} = \alpha + \mathbf{x}_{it}'\boldsymbol{\beta}$$

serves as a useful specification of the mean function. Here, $\mathbf{x}_{it}$ is a vector of explanatory, or *independent*, variables.

Longitudinal Data as Repeated Cross-Sectional Studies

Longitudinal data may be treated as a repeated cross section by ignoring the information about individuals that is tracked over time. As mentioned earlier, there are many important repeated surveys such as the CPS where subjects are not tracked over time. Such surveys are useful for understanding *aggregate* changes in a variable, such as the divorce rate, over time. However, if the interest is in studying the time-varying economic, demographic, or sociological characteristics *of an individual* on divorce, then tracking individuals over time is much more informative than using a repeated cross section.

Heterogeneity

By tracking subjects over time, we may model subject behavior. In many data sets of interest, subjects are unlike one another; that is, they are *heterogeneous*. In (repeated) cross-sectional regression analysis, we use models such as $y_{it} = \alpha + \mathbf{x}'_{it}\boldsymbol{\beta} + \varepsilon_{it}$ and ascribe the uniqueness of subjects to the disturbance term ε_{it}. In contrast, with longitudinal data we have an opportunity to model this uniqueness. A basic longitudinal data model that incorporates heterogeneity among subjects is based on

$$\mathrm{E}y_{it} = \alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta}, \quad t = 1, \ldots, T_i,\ i = 1, \ldots, n. \tag{1.2}$$

In cross-sectional studies where $T_i = 1$, the parameters of this model are unidentifiable. However, in longitudinal data, we have a sufficient number of observations to estimate $\boldsymbol{\beta}$ and $\alpha_1, \ldots, \alpha_n$. Allowing for *subject-specific parameters*, such as α_i, provides an important mechanism for controlling heterogeneity of individuals. Models that incorporate heterogeneity terms such as in Equation (1.2) will be called *heterogeneous models*. Models without such terms will be called *homogeneous models*.

We may also interpret heterogeneity to mean that observations from the same subject tend to be similar compared to observations from different subjects. Based on this interpretation, heterogeneity can be modeled by examining the sources of correlation among repeated observations from a subject. That is, for many data sets, we anticipate finding a positive correlation when examining $\{y_{i1}, y_{i2}, \ldots, y_{iT_i}\}$. As already noted, one possible explanation is the dynamic pattern among the observations. Another possible explanation is that the response shares a common, yet unobserved, subject-specific parameter that induces a positive correlation.

There are two distinct approaches for modeling the quantities that represent heterogeneity among subjects, $\{\alpha_i\}$. Chapter 2 explores one approach, where $\{\alpha_i\}$ are treated as fixed, yet unknown, parameters to be estimated. In this case, Equation (1.2) is known as a *fixed-effects* model. Chapter 3 introduces the second approach, where $\{\alpha_i\}$ are treated as draws from an unknown population and thus are random variables. In this case, Equation (1.2) may be expressed as

$$\mathrm{E}(y_{it} \mid \alpha_i) = \alpha_i + \mathbf{x}'_{it}\boldsymbol{\beta}.$$

This is known as a *random-effects* formulation.

Heterogeneity Bias

Failure to include heterogeneity quantities in the model may introduce serious bias into the model estimators. To illustrate, suppose that a data analyst mistakenly uses the function

$$\mathrm{E} y_{it} = \alpha + \mathbf{x}_{it}' \boldsymbol{\beta},$$

when Equation (1.2) is the true function. This is an example of heterogeneity bias, or a problem with data aggregation.

Similarly, one could have different (heterogeneous) slopes

$$\mathrm{E} y_{it} = \alpha + \mathbf{x}_{it}' \boldsymbol{\beta}_i$$

or different intercepts and slopes

$$\mathrm{E} y_{it} = \alpha_i + \mathbf{x}_{it}' \boldsymbol{\beta}_i.$$

Omitted Variables

Incorporating heterogeneity quantities into longitudinal data models is often motivated by the concern that important variables have been omitted from the model. To illustrate, consider the true model

$$y_{it} = \alpha_i + \mathbf{x}_{it}' \boldsymbol{\beta} + \mathbf{z}_i' \boldsymbol{\gamma} + \varepsilon_{it}.$$

Assume that we do not have available the variables represented by the vector $\mathbf{z}_i$; these *omitted variables* are also said to be *lurking*. If these omitted variables do not depend on time, then it is still possible to get reliable estimators of other model parameters, such as those included in the vector $\boldsymbol{\beta}$. One strategy is to consider the deviations of a response from its time-series average. This yields the derived model

$$\begin{aligned} y_{it}^* = y_{it} - \bar{y}_i &= (\alpha_i + \mathbf{x}_{it}' \boldsymbol{\beta} + \mathbf{z}_i' \boldsymbol{\gamma} + \varepsilon_{it}) - (\alpha_i + \bar{\mathbf{x}}_i' \boldsymbol{\beta} + \mathbf{z}_i' \boldsymbol{\gamma} + \bar{\varepsilon}_i) \\ &= (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)' \boldsymbol{\beta} + \varepsilon_{it} - \bar{\varepsilon}_i = \mathbf{x}_{it}^{*\prime} \boldsymbol{\beta} + \varepsilon_{it}^*, \end{aligned}$$

where we use the response time-series average $\bar{y}_i = T_i^{-1} \sum_{t=1}^{T_i} y_{it}$ and similar quantities for $\bar{\mathbf{x}}_i$ and $\bar{\varepsilon}_i$. Thus, using ordinary least-square estimators based on regressing the deviations in $\mathbf{x}$ on the deviations in y yields a desirable estimator of $\boldsymbol{\beta}$.

This strategy demonstrates how longitudinal data can mitigate the problem of *omitted-variable bias*. For strategies that rely on purely cross-sectional data, it is well known that correlations of lurking variables, $\mathbf{z}$, with the model explanatory variables, $\mathbf{x}$, induce bias when estimating $\boldsymbol{\beta}$. If the lurking variable is time-invariant, then it is perfectly collinear with the subject-specific variables α_i. Thus, estimation strategies that account for subject-specific parameters also account for time-invariant omitted variables. Further, because of the collinearity between subject-specific variables and time-invariant omitted variables, we may interpret the subject-specific quantities α_i as proxies for omitted variables. Chapter 7 describes strategies for dealing with omitted-variable bias.

Efficiency of Estimators

A longitudinal data design may yield more efficient estimators than estimators based on a comparable amount of data from alternative designs. To illustrate, suppose that the interest is in assessing the average change in a response over time, such as the divorce rate. Thus, let $\bar{y}_{\bullet 1} - \bar{y}_{\bullet 2}$ denote the difference between divorce rates between two time periods. In a repeated cross-sectional study such as the CPS, we would calculate the reliability of this statistic assuming independence among cross sections to get

$$\mathrm{Var}\,(\bar{y}_{\bullet 1} - \bar{y}_{\bullet 2}) = \mathrm{Var}\,\bar{y}_{\bullet 1} + \mathrm{Var}\,\bar{y}_{\bullet 2}.$$

However, in a panel survey that tracks individuals over time, we have

$$\mathrm{Var}\,(\bar{y}_{\bullet 1} - \bar{y}_{\bullet 2}) = \mathrm{Var}\,\bar{y}_{\bullet 1} + \mathrm{Var}\,\bar{y}_{\bullet 2} - 2\,\mathrm{Cov}\,(\bar{y}_{\bullet 1}, \bar{y}_{\bullet 2})\,.$$

The covariance term is generally positive because observations from the same subject tend to be positively correlated. Thus, other things being equal, a panel survey design yields more efficient estimators than a repeated cross-section design.

One method of accounting for this positive correlation among same-subject observations is through the heterogeneity terms, α_i. In many data sets, introducing subject-specific variables α_i also accounts for a large portion of the variability. Accounting for this variation reduces the mean-square error and standard errors associated with parameter estimators. Thus, we are more efficient in parameter estimation than for the case without subject-specific variables α_i.

It is also possible to incorporate subject-invariant parameters, often denoted by λ_t, to account for period (temporal) variation. For many data sets, this does not account for the same amount of variability as $\{\alpha_i\}$. With small numbers of time periods, it is straightforward to use time dummy (binary) variables to incorporate subject-invariant parameters.

Other things equal, standard errors become smaller and efficiency improves as the number of observations increases. For some situations, a researcher may obtain more information by sampling each subject repeatedly. Thus, some advocate that an advantage of longitudinal data is that we generally have more observations, owing to the repeated sampling, and greater efficiency of estimators compared to a purely cross-sectional regression design. The danger of this philosophy is that generally observations from the same subject are related.

Thus, although more information is obtained by repeated sampling, researchers need to be cautious in assessing the amount of additional information gained.

Correlation and Causation

For many statistical studies, analysts are happy to describe associations among variables. This is particularly true of forecasting studies where the goal is to predict the future. However, for other analyses, researchers are interested in assessing causal relationships among variables.

Longitudinal and panel data are sometimes touted as providing "evidence" of causal effects. Just as with any statistical methodology, longitudinal data models in and of themselves are insufficient to establish causal relationships among variables. However, longitudinal data can be more useful than purely cross-sectional data in establishing causality. To illustrate, consider the three ingredients necessary for establishing causality, taken from the sociology literature (see, for example, Toon, 2000EP):

- A statistically significant relationship is required.
- The association between two variables must not be due to another, omitted, variable.
- The "causal" variable must precede the other variable in time.

Longitudinal data are based on measurements taken over time and thus address the third requirement of a temporal ordering of events. Moreover, as previously described, longitudinal data models provide additional strategies for accommodating omitted variables that are not available in purely cross-sectional data.

Observational data do not come from carefully controlled experiments where random allocations are made among groups. Causal inference is not directly accomplished when using observational data and only statistical models. Rather, one thinks about the data and statistical models as providing relevant empirical evidence in a chain of reasoning about causal mechanisms. Although longitudinal data provide stronger evidence than purely cross-sectional data, most of the work in establishing causal statements should be based on the theory of the substantive field from which the data are derived. Chapter 6 discusses this issue in greater detail.

Drawbacks: Attrition

Longitudinal data sampling design offers many benefits compared to purely cross-sectional or purely time-series designs. However, because the sampling structure is more complex, it can also fail in subtle ways. The most common failure of longitudinal data sets to meet standard sampling design assumptions is through difficulties that result from *attrition*. In this context, attrition refers to a gradual erosion of responses by subjects. Because we follow the same subjects over time, nonresponse typically increases through time. To illustrate, consider the U.S. Panel Study of Income Dynamics (PSID). In the first year (1968), the nonresponse rate was 24%. However, by 1985, the nonresponse rate grew to about 50%.

Attrition can be a problem because it may result in a *selection bias*. Selection bias potentially occurs when a rule other than simple random (or stratified) sampling is used to select observational units. Examples of selection bias often concern endogenous decisions by agents to join a labor pool or participate in a social program. Suppose that we are studying a solvency measure of a sample of insurance firms. If the firm becomes bankrupt or evolves into another type of financial distress, then we may not be able to examine financial statistics associated with the firm. Nonetheless, this is exactly the situation in which we would anticipate observing low values of the solvency measure. The response of interest is related to our opportunity to observe the subject, a type of selection bias. Chapter 7 discusses the attrition problem in greater detail.

1.3 Longitudinal Data Models

When examining the benefits and drawbacks of longitudinal data modeling, it is also useful to consider the types of inference that are based on longitudinal data models, as well as the variety of modeling approaches. The type of application under consideration influences the choice of inference and modeling approaches.

Types of Inference

For many longitudinal data applications, the primary motivation for the analysis is to learn about the effect that an (exogenous) explanatory variable has on a response, controlling for other variables, including omitted variables. Users are interested in whether estimators of parameter coefficients, contained in the vector $\boldsymbol{\beta}$, differ in a statistically significant fashion from zero. This is also the primary motivation for most studies that involve regression analysis; this is not surprising given that many models of longitudinal data are special cases of regression models.

Because longitudinal data are collected over time, they also provide us with an ability to predict future values of a response for a specific subject. Chapter 4 considers this type of inference, known as *forecasting*.

The focus of Chapter 4 is on the "estimation" of random variables, known as *prediction*. Because future values of a response are, to the analyst, random variables, forecasting is a special case of prediction. Another special case involves situations where we would like to predict the expected value of a future response from a specific subject, conditional on *latent* (unobserved) characteristics associated with the subject. For example, this conditional expected value is known in insurance theory as a *credibility premium*, a quantity that is useful in pricing of insurance contracts.

Social Science Statistical Modeling

Statistical models are mathematical idealizations constructed to represent the behavior of data. When a statistical model is constructed (designed) to represent a data set with little regard to the underlying functional field from which the data emanate, we may think of the model as essentially data driven. For example, we might examine a data set of the form $(x_1, y_1), \ldots, (x_n, y_n)$ and posit a regression model to capture the association between x and y. We will call this type of model a *sampling-based model*, or, following the econometrics literature, we say that the model arises from the *data-generating process*.

In most cases, however, we will know something about the units of measurement of x and y and anticipate a type of relationship between x and y based on knowledge of the functional field from which these variables arise. To continue our example in a finance context, suppose that x represents a return from a market index and that y represents a stock return from an individual security. In this case, financial economics theory suggests a linear regression relationship of y on x. In the economics literature, Goldberger (1972E) defines a *structural model* to be a statistical model that represents causal relationships, as opposed to relationships that simply capture statistical associations. Chapter 6 further develops the idea of causal inference.

If a sampling-based model adequately represents statistical associations in our data, then why bother with an extra layer of theory when considering statistical models? In the context of binary dependent variables, Manski (1992E) offers three motivations: interpretation, precision, and extrapolation.

Interpretation is important because the primary purpose of many statistical analyses is to assess relationships generated by theory from a scientific field. A sampling-based model may not have sufficient structure to make this assessment, thus failing the primary motivation for the analysis.

Structural models utilize additional information from an underlying functional field. If this information is utilized correctly, then in some sense the structural model should provide a better representation than a model without this information. With a properly utilized structural model, we anticipate getting more precise estimates of model parameters and other characteristics. In practical terms, this improved precision can be measured in terms of smaller standard errors.

At least in the context of binary dependent variables, Manski (1992E) feels that extrapolation is the most compelling motivation for combining theory from a functional field with a sampling-based model. In a time-series context, extrapolation means forecasting; this is generally the main impetus for an analysis. In a regression context, extrapolation means inference about responses for sets of predictor variables "outside" of those realized in the sample. Particularly for public policy analysis, the goal of a statistical analysis is to infer the likely behavior of data outside of those realized.

Modeling Issues

This chapter has portrayed longitudinal data modeling as a special type of regression modeling. However, in the biometrics literature, longitudinal data models have their roots in multivariate analysis. Under this framework, we view the responses from an individual as a vector of responses; that is, $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iT})'$. Within the biometrics framework, the first applications are referred to as *growth curve* models. These classic examples use the height of children as the response to examine the changes in height and growth, over time (see Chapter 5). Within the econometrics literature, Chamberlain (1982E, 1984E) exploited the multivariate structure. The multivariate analysis approach is most effective with balanced data at points equally spaced in time. However, compared to the regression approach, there are several limitations of the multivariate approach. These include the following:

- It is harder to analyze missing data, attrition, and different accrual patterns.
- Because there is no explicit allowance for time, it is harder to forecast and predict at time points between those collected (interpolation).

Even within the regression approach for longitudinal data modeling, there are still a number of issues that need to be resolved in choosing a model. We have already introduced the issue of modeling heterogeneity. Recall that there are two important types of models of heterogeneity, fixed- and random-effects models (the subjects of Chapters 2 and 3).

Another important issue is the structure for modeling the dynamics; this is the subject of Chapter 8. We have described imposing a serial correlation on the disturbance terms. Another approach, described in Section 8.2, involves using lagged (endogenous) responses to account for temporal patterns. These models are important in econometrics because they are more suitable for structural modeling where a greater tie exists between economic theory and statistical modeling than models that are based exclusively on features of the data. When