Chapter 1

Introduction

1.1 Advantages of Longitudinal Studies

There are numerous advantages of longitudinal studies over cross-sectional studies. First, to the extent that repeated measurements from the same subject are not perfectly correlated, longitudinal studies are more powerful than cross-sectional studies for a fixed number of subjects. Stated in another way, to achieve a similar level of statistical power, fewer subjects are required in a longitudinal study. The reason for this is that repeated observations from the same subject, while correlated, are rarely if ever perfectly correlated. The net result is that the repeated measurements from a single subject provide more

independent information than a single measurement obtained from a single subject.

Second, in a longitudinal study, each subject can serve as his/her own control. For example, in a cross-over study, each subject can receive both experimental and control conditions. In general, intra-subject variability is substantially less than inter-subject variability, so a more sensitive or statistically powerful test is the result. As previously mentioned, in naturalistic or observational studies, the primary intervention of interest may also be time varying, so that naturalistic intra-subject changes in the intervention can be related to changes in the outcome of interest within individuals. Again, the net result is an exclusion of between subject variability from measurement error which results in more efficient estimators of treatment related effects than corresponding cross-sectional designs with the same number and pattern of observations.

Third, longitudinal studies allow an investigator to separate aging effects (*i.e.*, changes over time within individuals), from cohort effects (*i.e.*, differences between subjects at baseline). Such cohort effects are often mistaken for changes occurring within individuals. Without longitudinal data, one cannot differentiate these two competing alternatives.

Finally, longitudinal data can provide information about individual change, whereas cross-sectional data cannot. Statistical estimates of individual trends can be used to better understand heterogeneity in the population and the determinants of growth and change at the level of the individual.

1.2 Challenges of Longitudinal Data Analysis

Despite their advantages, longitudinal data are not without their challenges. Observations are not, by definition, independent and we must account for the dependency in data using more sophisticated statistical methods. The appropriate analytical methods are not as well developed, especially for more sophisticated models that permit more general forms of correlation among the repeated measurements. Often, there is a lack of available computer software for application of these more complex statistical models, or the level of statistical sophistication required of the user is beyond the typical level of the practitioner. In certain cases, for example nonlinear models for binary, ordinal, or nominal endpoints, parameter estimation can be computationally intensive due to the need for numerical or Monte Carlo simulation methods to evaluate the likelihood of nonlinear mixed-effects regression models.

An added complication that arises in the context of analysis of longitudinal data is the invariable presence of missing data. In some cases, a subject may be missing one of several measurement occasions; however, it is more likely that there are missing data due to attrition. Attrition, sometimes referred to as "drop-out," refers to a subject removing himself or herself from the study, prior to the end of the study. The data record for this subject therefore prematurely terminates. Several simple approaches to this problem have been proposed, none of which are statistically satisfactory. The simplest approach, termed a "completer analysis," limits the analysis to only those subjects that completed the study. Unfortunately, the available sample at the end of the study may have little resemblance to the sample initially randomized. Reasons for not completing the study may be confounded with the effects that the study was designed to investigate. For example, in a randomized clinical trial of a new drug versus placebo, only those subjects that did well on the drug may complete the study, giving the potentially false appearance of superiority of drug over placebo. The second simple approach is termed "Last Observation Carried Forward" (LOCF) and involves imputing the last available measurement to all subsequent measurement occasions. While things are somewhat better in the case of LOCF versus completer analyses, in an LOCF analysis, subjects

treated in the analysis as if they have had identical exposure to the drug may have quite different exposures in reality or their experience on the drug may be complicated by other factors that led to their withdrawal from the study that are ignored in the analysis. More rigorous statistical alternatives based on mixed-effects regression models with ignorable and nonignorable nonresponse are an important focus of this book. Nevertheless, the presence of missing data, and its treatment in the statistical analysis, is a complicating feature of longitudinal data, making analysis potentially far more complex than analysis of cross-sectional data. The advantage, however, is that all available data from each subject can be used in the analysis, leading to increased statistical power, the ability to estimate subject-specific effects, and decreased bias due to arbitrary exclusion of subjects with incomplete response or the simple imputation of values for the missing responses.

Yet another complicating feature of longitudinal data is that not only does the outcome measure change over time, but the values of the predictors or independent variables can also change over time. For example, in the measurement of the relationship between plasma level of a drug and health status, both plasma level (the predictor) and health status (the outcome) change over time. The goal here is to estimate the dynamic relationship between these

two variables over time. Note that this is a relationship that occurs within individuals, and may vary from individual to individual as well. While our overall objective may be to determine if a relationship exists between drug plasma level and health status in the population, we must be able to model this dynamic relationship within individuals, and reach an overall conclusion regarding whether such a relationship exists in the population. The treatment of time-varying covariates in analysis of longitudinal data permits much stronger statistical inferences about dynamical relationships than can be obtained using cross-sectional data. The price, however, is considerable added complexity to the statistical model.

Finally, in some cases, the repeated measurements involve different conditions that the same subjects are exposed to. A classic example is a cross-over design in which two or more treatments are given to the same subject in different orders. In these cases, the statistical inferences may be compromised by order or carry-over effects in which response to a one treatment may be conditional on exposure to a previous treatment. Dealing with carry-over or sequence effects is far from trivial, and is the statistical price paid for the stronger statistical inferences permitted by within-subject experimentation.

1.3 Some General Notation

To set the stage for the statistical discussion to follow, it is helpful to present a unified notation for the various aspects of the longitudinal design. We index the N subjects in the longitudinal study as

$$i = 1, \ldots, N \text{ subjects .}$$

For a balanced design in which all subjects have complete data, and are measured on the same occasions, we index the measurement occasions as

$$j = 1, \ldots, n \text{ observations .}$$

or in the unbalanced case of unequal numbers of measurements or different time-points for different subjects

$$j = 1, \ldots, n_i \text{ observations for subject i .}$$

The total number of observations are given by

$$\sum_{i}^{N} n_i \ .$$

The repeated responses, or outcomes, or dependent measures for subject i are denoted as the vector

$$\boldsymbol{y}_i \ = \ n_i \times 1 \ .$$

The values of the p predictors, or covariates, or independent variables for subject i on occasion j are denoted as (including an intercept term):

$$\boldsymbol{x}_{ij} \;=\; p \times 1 \; .$$

For time-invariant predictors (between subject, *e.g.*, sex), the values of $\boldsymbol{x}_{ij}$ are constant for $j = 1, \ldots, n_i$. For time-varying predictors (within-subject, *e.g.*, age), the $\boldsymbol{x}_{ij}$ can take on subject and time-point specific values. To describe the entire matrix of predictors for subject i, we use the notation

$$\boldsymbol{X}_i = n_i \times p \; .$$

It should be noted that not all of the statistical literature on analysis of longitudinal data uses this notation. Sometimes, in other sources, the notation

- $i = 1, \ldots, n$ subjects
- $j = 1, \ldots, t_i$ observations
- total number of observations $= \sum_i^n t_i$

is used.

1.4 Data Layout

In fixing ideas for the statistical development to follow, it is also useful to apply this previously described notation to describe a longitudinal dataset as follows.

subject	observation	response	covariates		
1	1	y_{11}	x_{111}	$\ldots$	x_{11p}
1	2	y_{12}	x_{121}	$\ldots$	x_{12p}
.	.	.	.	. .	
1	n_1	y_{1n_1}	x_{1n_11}	$\ldots$	x_{1n_1p}
.	.	.	.	. .	
.	.	.	.	. .	
.	.	.	.	. .	
.	.	.	.	. .	
N	1	y_{N1}	x_{N11}	$\ldots$	x_{N1p}
N	2	y_{N2}	x_{N21}	$\ldots$	x_{N2p}
.	.	.	.	. .	
N	n_N	y_{Nn_N}	x_{Nn_N1}	$\ldots$	x_{Nn_Np}

In this univariate design, n_i varies by subject, and if x_r is time-invariant

(between-subjects) then, for a given subject i, the covariate values are the same across time, namely, $x_{i1r} = x_{i2r} = x_{i3r} = \ldots = x_{in_ir}$.

1.5 Analysis Considerations

There are several different features of longitudinal studies that must be considered when selecting an appropriate longitudinal analysis. First, is the form of the outcome or response measure. If the outcome of interest is continuous and normally distributed, much simpler analyses are usually possible (*e.g.*, a mixed-effects linear regression model). By contrast, if the outcome is continuous, but does not have a normal distribution (*e.g.*, a count) then alternative non-linear models (*e.g.*, a mixed-effects Poisson regression model) can be considered. For qualitative outcomes, such as binary (yes or no), ordinal (*e.g.*, sad, neutral, happy), nominal (republican, democrat, independent), more complex non-linear models are also typically required.

Second, the number of subjects N is an important consideration for selecting a longitudinal analysis method. The more advanced models (*e.g.*, generalized mixed-effects regression models) that are appropriate for analysis of unbalanced longitudinal data, are based on large sample theory and may be

inappropriate for analysis of small N studies (*e.g.*, $N < 50$).

Third, the number of observations per subject n_i is also an important consideration when selecting an analytic method. For $n_i = 2$ for all subjects, a simple change score can be computed and the data can be analyzed using methods for cross-sectional data, such as ANCOVA. When $n_i = n$ for all subjects, the design is said to be balanced, and traditional ANOVA or MANOVA models for repeated measurements (*i.e.*, traditional mixed-effects models or multivariate growth curve models) can be used. In the most general case where n_i varies from subject to subject, more general methods are required (*e.g.*, generalized mixed-effects regression models), which are the primary focus of this book.

Fourth, the number and type of covariates is an important consideration for model selection for $E(\boldsymbol{y}_i)$. In the one sample case, we may only have interest in characterizing the rate of change in the population over time. Here, we can use a random-effects regression model, where the parameters of the growth curve are treated as random effects and allowed to vary from subject to subject. In the multiple sample case (*e.g.*, comparison of one or more treatment conditions to control), the model consists of one or more categorical covariates that contrast the various treatment conditions in the design. In

the regression case, we may have a mixture of continuous and/or categorical covariates, such as age, sex, and race. When the covariates take on time-specific values (*i.e.*, time-varying covariates), the statistical model must be capable of handling these as well.

Fifth, selection of a plausible variance-covariance structure for the $V(\boldsymbol{y}_i)$ is of critical importance. Different model specifications lead to homogeneous or heterogeneous variances, and/or homogeneous or heterogeneous covariances of the repeated measurements over time. Furthermore, residual autocorrelation among the responses may also play a role in modeling the variance-covariance structure of the data.

Each of these factors is important for selecting an appropriate analytical model for analysis of a particular set of longitudinal data. In the following chapters, greater detail on the specifics of these choices will be presented.

1.6 General Approaches

There are several different general approaches to the analysis of longitudinal data. To provide an overview, and to fix ideas for further discussion and more detailed presentation, we present the following outline.

The first approach, which we refer to as the "derived variable" approach, involves the reduction of the repeated measurements into a summary variable. In fact, once reduced, this approach is strictly not longitudinal, since there is only a single measurement per subject. Perhaps the earliest example of the analysis of longitudinal data was presented by Student (1908) in his illustration of the t-test. The objective of the study (Cushny and Peebles, 1905) was to determine changes in sleep as a function of treatment with the hypnotic drug scopolomine. Although hours of sleep were carefully measured by the investigators, day to day variability presented statistical challenges in detecting the drug effect using large sample methods available at the time. Student (Gossett) proposed the one sample t-test to test if the average difference between experimental and control conditions was zero.

Examples of derived variables include (a) average across time, (b) linear trend across time, (c) carrying the last observation forward, (d) computing a change score, and (e) computing the area under the curve. A critical problem with all of these approaches is that our uncertainty in the derived variable is proportional to the number of measurements for which it was computed. In the unbalanced case (*e.g.*, drop-outs), different subjects will have different numbers of measurements and hence different uncertainties, therefore violat-

ing the commonly made assumption of homoscedasticity. Furthermore, by reducing multiple repeated measurements to a single summary measurement, there is typically, a substantial loss of statistical power. Finally, use of time-varying covariates is not possible when the temporal aspect of the data has been removed.

Second, perhaps the simplest, but most restrictive model is the ANOVA for repeated measurements (Winer, 1971). The model assumes compound symmetry which implies constant variances and covariances over time. Clearly such an assumption has little, if any validity for longitudinal data. Typically, variances increase with time because some subjects respond and others do not, and covariances for proximal occasions are larger than covariances for distal occasions. The model allows each subject to have his or her own trend line, however, the trend lines can only differ in terms of their intercepts, which implies that subjects deviate at baseline, but are consistent thereafter. It is more likely, of course, that subjects will deviate systematically from the overall trend both at baseline and in terms of the rate that they change over time (*i.e.*, their slope).

Third, MANOVA models have also been proposed for analysis of longitudinal data (see Bock, 1975). In the multivariate case, the repeated obser-

vations are generally transformed to orthogonal polynomial coefficients, and these coefficients (*e.g.*, constant, linear, quadratic growth rates) are then used as multivariate responses in a MANOVA. The principle disadvantages of this approach is that it does not permit missing data, or different measurement occasions for different subjects.

Fourth, generalized mixed-effects regression models, which form the primary emphasis of this book, are now quite widely used for analysis of longitudinal data. These models can be applied to both normally distributed continuous outcomes as well as categorical outcomes and other nonnormally distributed outcomes such as counts that have a Poisson distribution. Mixed-effects regression models are quite robust to missing data, irregularly spaced measurement occasions, and can easily handle both time-invariant and time-varying covariates. As such, they are among the most general of the methods for analysis of longitudinal data. They are sometimes called "full-likelihood" methods, because they make full use of all available data from each subject. The advantage is that missing data are ignorable if the missing responses can be explained either by covariates in the model, or by the available responses from a given subject. The disadvantage is that full-likelihood methods are more computationally complex than quasi-likelihood methods, such as gener-

alized estimating equations (GEE).

Fifth, covariance pattern models (Jennrich and Schluchter, 1986) can also be used to analyze longitudinal data. Here, the variance-covariance matrix of the repeated outcomes is modeled directly, and there is no attempt at distinguishing within-subjects variance from between-subjects variance, as is the case with mixed-effects regression models. Typically, the variance-covariance matrix is modeled in terms of a relatively small number of parameters and full-likelihood estimation methods are used.

Sixth, GEE models are often used as a very general and computationally convenient alternative to mixed-effects regression models. They can be used to fit a wide variety of types of outcome measures, and do not require complex numerical evaluation of the likelihood for nonlinear models. The disadvantage, is that missing data are only ignorable if the missing data are explained by covariates in the model. This is an overly restrictive assumption and therefore, GEE models have only limited applicability to incomplete longitudinal data.

1.7 The Simplest Longitudinal Analysis

A paired t-test can be used to determine if there is significant average change between two timepoints. For this, note that:

- $i = 1, \ldots, N$ subjects
- y_{i1} = pre-test
- y_{i2} = post-test
- $d_i = y_{i2} - y_{i1}$ = post to pre change score

The null hypothesis can be written as:

$$H_0 : \mu_{y_1} = \mu_{y_2} \quad \text{same as} \quad H_0 : (\mu_{y_2} - \mu_{y_1}) = 0$$

The test statistic is computed as:

$$\begin{aligned} t &= \bar{d} \,/\, \left(s_d/\sqrt{N}\right) \\ &= \bar{d} \,/\, \left(\sqrt{\left[\sum_i d_i^2 - (\sum_i d_i)^2/N\right] /(N-1)} \,/\, \sqrt{N}\right) \\ &\overset{H_0}{\sim} t_{N-1} \end{aligned}$$

Notice, that we can do the same test using a regression model:

$$d_i = \beta_0 + e_i \ ,$$

and testing $H_0 : \beta_0 = 0$

1.7.1 Change Score analysis

Suppose there is a grouping variable

- $x_i = 0$ for controls
- $x_i = 1$ for treatment group

A regression model for the change score is given by:

$$d_i \;=\; \beta_0 \;+\; \beta_1 x_i \;+\; e_i \;.$$

Hypothesis testing is as follows:

- testing $H_0 : \beta_0 = 0$ tests whether the average change is equal to zero for the control group
- testing $H_0 : \beta_1 = 0$ tests whether the average change is equal for the two groups

Notice that the change score analysis is equivalent to regressing the post-treatment measurement on the treatment variable, using the pre-treatment measurement as a covariate with slope equal to one.

$$d_i = \beta_0 + \beta_1 x_i + e_i$$

$$y_{i2} - y_{i1} = \beta_0 + \beta_1 x_i + e_i$$

$$y_{i2} = y_{i1} + \beta_0 + \beta_1 x_i + e_i$$

1.7.2 Analysis of covariance of post-test scores

When the slope describing the relationship between the pre-test and post-test score is not one (*i.e.*, $\beta_2 \neq 1$), then we have an ANCOVA model for the post-test score, *i.e.*,

$$y_{i2} = \beta_0 + \beta_1 x_i + \beta_2 y_{i1} + e_i \ .$$

In terms of hypothesis testing we have

- testing $H_0 : \beta_0 = 0$ tests whether the average post-test is equal to zero for the control group subjects with zero pre-test
- testing $H_0 : \beta_1 = 0$ tests whether the post-test is equal for the two groups, given the same value on the pre-test (*i.e.*, conditional on pre-test)

- testing $H_0 : \beta_2 = 0$ tests whether the post-test is related to the pre-test, conditional on group

Note that change score analysis and ANCOVA answer different questions. Change score analysis tests if the average change is the same between the groups, whereas ANCOVA tests if the post-test average is the same between groups for sub-populations with the same pre-test values (*i.e.*, is the conditional average the same between the groups). The choice of which to use depends on the question of interest. The two models often yield similar conclusions for a test of the group effect. If subjects are randomized to group, then ANCOVA is more efficient (*i.e.*, more powerful), however, one must be careful in non-randomized settings, where groups are not necessarily similar in terms of pre-test scores (Lord's paradox, see Bock, 1975; Allison, 1990)

1.7.3 ANCOVA of change scores

Many practitioners have argued whether it is better to use the post-treatment ANCOVA (adjusting for pre-treatment) or to use ANCOVA on change scores adjusting for pre-treatment. With a bit of algebra, it is easy to show that the two approaches are identical for testing the null hypothesis of no treatment effect (*i.e.*, $H_0 : \beta_1 = 0$).

$$d_i = \beta_0 + \beta_1 x_i + \beta_2 y_{i1} + e_i$$

$$y_{i2} - y_{i1} = \beta_0 + \beta_1 x_i + \beta_2 y_{i1} + e_i$$

$$y_{i2} = \beta_0 + \beta_1 x_i + (1 + \beta_2) y_{i1} + e_i$$

1.7.4 Example

The Television School and Family Smoking Prevention and Cessation Project (Flay, *et al.*, 1988) was designed to increase knowledge of the effects of tobacco use in school-age children. Characteristics of the sample are as follows:

- *sample* - 1600 7th-graders - 135 classrooms - 28 LA schools
  - between 1 to 13 classrooms per school
  - between 2 to 28 students per classroom
- *outcome* - knowledge of the effects of tobacco use
- *timing* - students tested at pre and post-intervention

- *design* - schools randomized to
  - a social-resistance classroom curriculum (CC)
  - a media (television) intervention (TV)
  - CC combined with TV
  - a no-intervention control group

The first hypothesis to be tested was whether there was change across time. Summary statistics listed below reveal that there was an overall increase in knowledge scores of 0.59 units.

Variable	N	Mean	Std Dev	Minimum	Maximum
PRETHKS	1600	2.06938	1.26018	0	6.00000
POSTHKS	1600	2.66188	1.38293	0	7.00000
THKSdelt	1600	0.59250	1.57932	-5.00000	6.00000

A simple change score analysis revealed that this difference was significant ($t = 15.01, p < .0001$). Alternatively, a regression model with intercept and no regressors, yielded the identical result:

Variable	Parameter Estimate	Standard Error	t Value	Pr > | t |
Intercept	0.59250	0.03948	15.01	< .0001

Next, we examine the effect of the CC and TV interventions on change in knowledge. Summary statistics for the 2 x 2 design are given below.

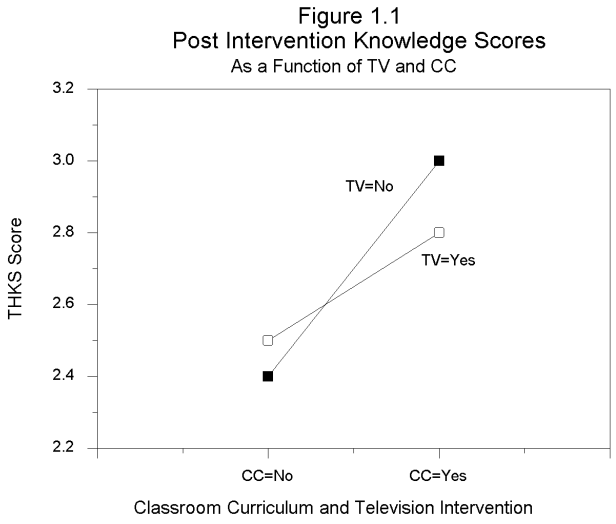
Tobacco and Health Knowledge Scale

Subgroup Descriptive Statistics

Pretest, Post-Intervention, and Difference

	CC = no		CC = yes	
	TV = no	TV = yes	TV = no	TV = yes
N	421	416	380	383
Pretest mean	2.152	2.087	2.050	1.979
sd	1.182	1.288	1.285	1.286
Post-Int mean	2.361	2.539	2.968	2.823
sd	1.296	1.437	1.405	1.312
Difference	0.209	0.452	0.918	0.844

Does change across time vary by CC, TV, or both? To test this hypothesis, we begin by graphically displaying the changes for each of the four groups in Figure 1.1.



Next, we compute a regression analysis for Post intervention knowledge scores. The model with main effects and interaction of CC and TV effects yielded a significant main effect of CC and a CC by TV interaction. The TV effect approached significance ($p<.06$).

Variable	Parameter Estimate	Standard Error	t Value	Pr > | t |
Intercept	2.36105	0.06646	35.52	<.0001
CC	0.60738	0.09649	6.29	<.0001
TV	0.17742	0.09427	1.88	0.0600
CCTV	-0.32338	0.13652	-2.37	0.0180

Adjusting for pre-intervention knowledge scores yielded significant main effects and interactions.

Variable	Parameter Estimate	Standard Error	t Value	Pr > | t |
Intercept	1.66126	0.08436	19.69	<.0001
PRETHKS	0.32518	0.02585	12.58	<.0001
CC	0.64055	0.09210	6.95	<.0001
TV	0.19871	0.08996	2.21	0.0273
CCTV	-0.32162	0.13025	-2.47	0.0136

Regression on difference scores yielded quite similar results to the baseline adjusted post-intervention regression model.

Variable	Parameter Estimate	Standard Error	t Value	Pr > | t |
Intercept	0.20903	0.07573	2.76	0.0058
CC	0.70939	0.10995	6.45	<.0001
TV	0.24290	0.10742	2.26	0.0239
CCTV	-0.31798	0.15556	-2.04	0.0411

Finally, adding pre-intervention knowledge scores as a covariate, produces identical intervention effects to the model for post-intervention scores.

Variable	Parameter Estimate	Standard Error	t Value	Pr > | t |
Intercept	1.66126	0.08436	19.69	<.0001
PRETHKS	-0.67482	0.02585	-26.10	<.0001
CC	0.64055	0.09210	6.95	<.0001
TV	0.19871	0.08996	2.21	0.0273
CCTV	-0.32162	0.13025	-2.47	0.0136

1.8 Summary

Longitudinal studies represent enormous advantages over cross-sectional studies in terms of providing foundations for causal inference. It is not surprising that the required level of statistical sophistication required for the analysis of longitudinal data is more advanced as well. Despite their advantages, longitudinal data are not without their challenges. The treatment of missing data plays a far greater role in longitudinal studies than it does in cross-sectional studies and analysis of naturalistic or observational longitudinal data is complicated by numerous sources of bias due to selection effects. In the following chapters, we provide a variety of approaches to the analysis of different types of longitudinal data, present their strengths and limitations, and illustrate their use with real examples.