MCAR	Missing Completely at Random
MCMC	Markov Chain Monte Carlo
ML	Maximum Likelihood
MLE	Maximum Likelihood Estimate
MNAR	Missing Not at Random
MRM	Mixed-effects Regression Model
NAS	National Academy of Sciences
OPO	Organ Procurement Organization
OPTN	Organ Procurement Transplantation Network
REML	Restricted Maximum Likelihood
SSCP	Sum of Squares and Cross Products
SSRI	Selective Serotonin Reuptake Inhibitor
TCA	Tricyclic Antidepressant
UN	Unstructured
ZAP	Zero-altered Poisson
ZIP	Zero-inflated Poisson

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 crossover 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 when compared to 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

Longitudinal Data Analysis. By Donald Hedeker and Robert D. Gibbons 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 studymissing data,completer analysis. 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 it 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 must 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 crossover 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

$$y_i = n_i \times 1.$$

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

$$x_{ij} = p \times 1.$$

For time-invariant predictors (between subject, *e.g.*, sex), the values of x_{ij} are constant for $j = 1, \ldots, n_i$. For time-varying predictors (within-subject, *e.g.*, age), the x_{ij} can take on subject- and timepoint-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 literature on longitudinal data analysis uses this notation. In other sources, the following notation is sometimes used:

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

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}	$\cdots$	x_{11p}
1	2	y_{12}	x_{121}	$\cdots$	x_{12p}
.	.	.	.	..	
1	n_1	y_{1n_1}	x_{1n_11}	$\cdots$	x_{1n_1p}
.	.	.	.	..	
.	.	.	.	..	
.	.	.	.	..	
.	.	.	.	..	
N	1	y_{N1}	x_{N11}	$\cdots$	x_{N1p}
N	2	y_{N2}	x_{N21}	$\cdots$	x_{N2p}
.	.	.	.	..	
N	n_N	y_{Nn_N}	x_{Nn_N1}	$\cdots$	x_{Nn_Np}

In this univariate design, n_i varies by subject and so the number of data lines per subject can vary. In terms of the covariates, if x_r is time-invariant (*i.e.*, a between-subjects variable) then, for a given subject i, the covariate values are the same across time, namely, $x_{i1r} = x_{i2r} = x_{i3r} = \ldots = x_{in_ir}$.

The above layout depicts what is called a 2-level design in the multilevel [Goldstein, 1995] and hierarchical linear modeling [Raudenbush and Bryk, 2002] literatures. Namely, repeated observations at level 1 are nested within subjects at level 2. In some cases, subjects themselves are nested within sites, hospitals, clinics, workplaces, etc. In this case, the design has three levels with level-2 subjects nested within level-3 sites. This book primarily focuses on 2-level designs and models, with Chapter 13 covering 3-level extensions.

1.5 ANALYSIS CONSIDERATIONS

There are several different features of longitudinal studies that must be considered when selecting an appropriate longitudinal analysis. First, there 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 nonlinear 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), or nominal (republican, democrat, independent), more complex nonlinear 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 (a) homogeneous or heterogeneous variances and/or (b) 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 violating 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 observations 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 principal 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 and 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 generalized 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 a restrictive assumption in many situations and therefore, GEE models have somewhat limited applicability to incomplete longitudinal data.

1.7 THE SIMPLEST LONGITUDINAL ANALYSIS

The simplest possible longitudinal design consists of a single group and two measurement occassions. A paired t-test can be used to determine if there is significant average change between two timepoints. For this, note that there are N subjects, and the pre-test and post-test measurements for subject i are denoted y_{i1} and y_{i2} respectively. The difference or change score for subject i is denoted $d_i = y_{i2} - y_{i1}$. To test the difference between pre-test and post-test measurements, the null hypothesis can be written as: $H_0 : \mu_{y_1} = \mu_{y_2}$ which is the same as same as writing $H_0 : (\mu_{y_2} - \mu_{y_1}) = 0$. The test statistic is computed as

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

and is distributed as Student's t on $N-1$ degrees of freedom. Notice that we can perform the same test using a regression model, where the difference between pre- and post-test measurements is

$$d_i = \beta_0 + e_i.$$

Assumming normality, we can test $H_0 : \beta_0 = 0$, by computing the ratio of $\hat{\beta}_0$ to its standard error, which also has a t-distribution on $N-1$ degrees of freedom.

1.7.1 Change Score Analysis

Now consider a slightly more complex situation in which we have randomized subjects into two groups, a treatment and a control group. The groups are designated by $x_i = 0$ for controls and $x_i = 1$ for the treatment group. A regression model for the change score is given by

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

Note that in this model, β_0 reflects the average change for the control group, and β_1 reflects the difference in average change between the two groups. 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.

$$\begin{aligned} d_i &= \beta_0 + \beta_1 x_i + e_i, \\ y_{i2} - y_{i1} &= \beta_0 + \beta_1 x_i + e_i, \\ y_{i2} &= y_{i1} + \beta_0 + \beta_1 x_i + e_i. \end{aligned}$$

In many ways, this is an overly restrictive assumption, since there is typically no *a priori* reason why a unit change in the pre-test score should translate to a unit change in the post-test measurement.

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 using ANCOVA in nonrandomized settings, where groups are not necessarily similar in terms of pre-test scores (Lord's paradox, see Allison [1990]; Bock [1975]; Maris [1998]; Wright [2005]).

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.*, $H_0 : \beta_1 = 0$).

$$\begin{aligned} d_i &= \beta_0 + \beta_1 x_i + \beta_2 y_{i1} + e_i, \\ y_{i2} - y_{i1} &= \beta_0 + \beta_1 x_i + \beta_2 y_{i1} + e_i, \\ y_{i2} &= \beta_0 + \beta_1 x_i + (1 + \beta_2) y_{i1} + e_i. \end{aligned}$$

As can be seen from the equations above, there is no difference whatsoever between the two alternative models in terms of the treatment effect.

1.7.4 Example

To illustrate application of these simple models for the analysis of pre-test versus post-test change, we applied them to the Television School and Family Smoking Prevention and Cessation Project [Flay et al., 1988]. This study 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

Here, we will ignore the clustering of students within classrooms and schools (see Chapter 13 for a description of methods to deal with this) and will concentrate on the potential change across the two study timepoints.

The first hypothesis to be tested is whether there was any overall change across time. The mean pre-intervention score is 2.069 and the mean post-intervention score is 2.662. Thus, there was an overall increase in knowledge scores of 0.59 units. A simple change score analysis (*i.e.*, paired *t*-test) reveals that this difference is significant ($t_{1599} = 15.01, p <$

.0001). Alternatively, a regression model treating the change score as the dependent variable, with an intercept and no regressors, yields identical results: $\hat{\beta}_0 = .5925$, standard error $= .0395$, $t_{1599} = 15.01, p < .0001$.

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

Tobacco and Health Knowledge Scale (THKS)
Subgroup Descriptive Statistics
Pre-Intervention, Post-Intervention, and Difference

	CC = no		CC = yes	
	TV = no	TV = yes	TV = no	TV = yes
N	421	416	380	383
Pre- Int 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 post-intervention THKS means for the four groups in Figure 1.1.

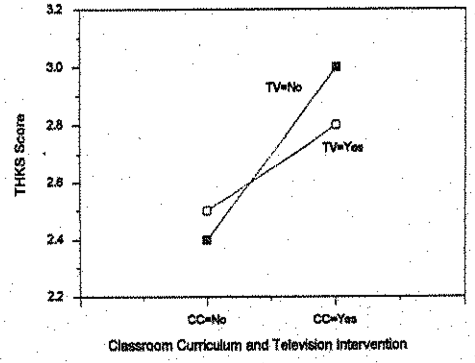

Figure 1.1. Post-intervention THKS means.

Next, we compute a regression analysis for post-intervention knowledge scores. Here, we code the treatment variables CC and TV as simple dummy-codes with 0 indicating no treatment and 1 indicating treatment exposure. The model with main effects and interaction of CC and TV effects yields a significant main effect of CC and a CC by TV interaction. The TV effect approaches significance ($p < .06$).

Variable	Parameter Estimate	Standard Error	*t* Value	Pr > | *t* |
Intercept	2.3611	0.0665	35.52	<.0001
CC	0.6074	0.0965	6.29	<.0001
TV	0.1774	0.0943	1.88	0.0600
CC by TV	−0.3234	0.1365	−2.37	0.0180

From this model, we can obtain the estimated post-intervention means of the four groups, based on our coding of the variables CC and TV:

- CC no TV no $= \hat{\beta}_0 = 2.3611$
- CC yes TV no $= \hat{\beta}_0 + \hat{\beta}_1 = 2.3611 + 0.6074 = 2.9685$
- CC no TV yes $= \hat{\beta}_0 + \hat{\beta}_2 = 2.3611 + 0.1774 = 2.5385$
- CC yes TV yes $= \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 = 2.3611 + 0.6074 + 0.1774 - 0.3234 = 2.8225$

These agree with the observed means, within rounding error, since the model is a saturated one in terms of the four groups. The ANCOVA model, adjusting for pre-intervention knowledge scores, yields significant main effects and interactions.

Variable	Parameter Estimate	Standard Error	*t* Value	Pr > | *t* |
Intercept	1.6613	0.0844	19.69	<.0001
PRETHKS	0.3252	0.0259	12.58	<.0001
CC	0.6406	0.0921	6.95	<.0001
TV	0.1987	0.0900	2.21	0.0273
CC by TV	−0.3216	0.1303	−2.47	0.0136

Here, using the sample average on the pre-intervention scores (= 2.069), we can calculate the adjusted means for the four groups as

- CC no TV no $= \hat{\beta}_0 + 2.069 \times \hat{\beta}_1 = 1.6613 + 2.069 \times 0.3252 = 2.3341$
- CC yes TV no $= \hat{\beta}_0 + 2.069 \times \hat{\beta}_1 + \hat{\beta}_2 = 1.6613 + 2.069 \times 0.3252 + 0.6406 = 2.9747$
- CC no TV yes $= \hat{\beta}_0 + 2.069 \times \hat{\beta}_1 + \hat{\beta}_3 = 1.6613 + 2.069 \times 0.3252 + 0.1987 = 2.5328$
- CC yes TV yes $= \hat{\beta}_0 + 2.069 \times \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 + \hat{\beta}_4 = 1.6613 + 2.069 \times 0.3252 + 0.6406 + 0.1987 - 0.3216 = 2.8518$

These are not very different from the estimated marginal means, which is not surprising since these four groups did not differ much in terms of their pre-intervention scores.

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

Variable	Parameter Estimate	Standard Error	t Value	Pr > | t |
Intercept	0.2090	0.0757	2.76	0.0058
CC	0.7094	0.1100	6.45	<.0001
TV	0.2429	0.1074	2.26	0.0239
CC by TV	−0.3180	0.1556	−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	1.6613	0.0844	19.69	<.0001
PRETHKS	−0.6748	0.0259	−26.10	<.0001
CC	0.6406	0.0921	6.95	<.0001
TV	0.1987	0.0900	2.21	0.0273
CC by TV	−0.3216	0.1303	−2.47	0.0136

These analyses reveal that both CC and TV interventions impact knowledge scores, however, their effects are nonadditive. CC increases knowledge gains overall, however the TV intervention decreases the gain in knowledge when CC is present and increases the gain in knowledge when CC is absent.

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.

CHAPTER 2

ANOVA APPROACHES TO LONGITUDINAL DATA

There are two classical approaches to the analysis of longitudinal data. The first is variably called univariate mixed-model, split-plots, or repeated measures ANOVA, and the second is based on multivariate ANOVA (MANOVA). Both models assume interval measurement and normally distributed errors that are homogeneous across groups. In some cases, normality and homogeneity of variance can be brought about through transformation (*e.g.*, natural log transformation). For both models, the primary focus is on comparison of group means, and neither model is informative about individual growth curves (*i.e.*, subject-specific trends). Furthermore, the timepoints are assumed to be fixed across subjects (either evenly or unevenly spaced) and are treated as a classification variable in the ANOVA or MANOVA model. This precludes analysis of unbalanced designs in which different subjects are measured on different occasions. Both models are based on least-squares estimation and are therefore adversely affected by outliers and missing data. While the ANOVA model can handle some missing data (*i.e.*, there are methods for unbalanced ANOVA), the MANOVA model cannot handle any missing data. In terms of the variance–covariance structure for the responses ($\boldsymbol{y}_i$), the ANOVA model assumes compound symmetry (*i.e.*, equal variances and covariances over time), whereas the MANOVA model makes no assumption regarding the specific form of the variance–covariance structure. While this is an important advantage of MANOVA over ANOVA, it is tempered by the larger limitation of requiring complete data for all subjects in the MANOVA model. As such, application of MANOVA must follow deletion of all subjects without complete data, which is essentially a completer analysis, and is prone to substantial bias in that the composition of subjects that complete the study can be quite different from the composition of the subjects at the time of randomization.

Longitudinal Data Analysis. By Donald Hedeker and Robert D. Gibbons

In this chapter, we describe the univariate ANOVA model for within-subject designs in detail. In the social and behavioral sciences literature, this model and extensions of it are often referred to as a "Repeated Measures ANOVA." In Chapter 3, we present the alternative multivariate approach. While these approaches are no longer recommended for routine application (if at all), they are important in that they fix ideas for the development of the more modern and advanced methods that are the primary focus of this book.

2.1 SINGLE-SAMPLE REPEATED MEASURES ANOVA

In the single-sample case, the model is referred to as the randomized blocks ANOVA. In this case, we have no intervention or group effects, but are simply using the model to characterize rates of change over time. With $i = 1, \ldots, N$ subjects and $j = 1, \ldots, n$ measurement occasions, the randomized blocks ANOVA is given by the linear model,

$$y_{ij} = \mu + \pi_i + \tau_j + e_{ij} \ , \tag{2.1}$$

where μ = the grand mean, π_i = the individual difference component for subject i, which is assumed to be constant over time, τ_j = the effect of time, assumed to be the same for all subjects, and e_{ij} = the error for subject i on occasion j. We also assume that the random components are distributed as $\pi_i \sim N(0, \sigma_\pi^2)$, where σ_π^2 is the between-subjects variance, and $e_{ij} \sim N(0, \sigma_e^2)$, with σ_e^2 as the within-subjects variance. Notice that this is a mixed model because it includes both random (π_i) and fixed (τ_j) parameters.

2.1.1 Design

The data for the model above can be represented by the following two-factor design of subjects crossed with timepoints:

	Timepoint			
Subject	1	2	...	n
1	y_{11}	y_{12}	...	y_{1n}
2	y_{21}	y_{22}	...	y_{2n}
.	.	.	...	.
.	.	.	...	.
N	y_{N1}	y_{N2}	...	y_{Nn}

In this design there is one observation per cell. In other words, each subject is observed once at each timepoint. The design is similar to a randomized blocks design with subjects as blocks. In the simple case of $n = 2$, the design and subsequent analysis is identical to a paired t-test, in terms of testing of the time effect.

In terms of model assumptions, we assume that

$$\begin{aligned}
\sum_{j=1}^{n} \tau_j &= 0, \\
E(y_{ij}) &= \mu + \tau_j, \\
V(y_{ij}) &= V(\mu + \tau_j + \pi_i + e_{ij}) = \sigma_\pi^2 + \sigma_e^2, \\
C(y_{ij}, y_{i'j}) &= 0 \ for \ i \neq i', \\
C(y_{ij}, y_{ij'}) &= \sigma_\pi^2 \ for \ j \neq j'.
\end{aligned}$$

Here, $E(\cdot)$, $V(\cdot)$, and $C(\cdot)$ denote expectation, variance, and covariance respectively. Note that the first covariance statement indicates that subjects are independent of each other, whereas the second covariance statement indicates that the covariance is σ_π^2 for any two repeated measures within the same subject. This covariance can be expressed as a correlation, which reflects the magnitude of the within-subject association in a more interpretable metric:

$$Corr(y_{ij}, y_{ij'}) = \frac{\sigma_\pi^2}{\sigma_\pi^2 + \sigma_e^2} \, . \tag{2.2}$$

This correlation is termed the intraclass correlation. Note that it is the same for all longitudinal pairs of measures and therefore represents the average correlation of y from any two timepoints. In the above formulation, because variances cannot be negative, the intraclass correlation ranges from 0 to 1; it equals 0 if subjects explain none of the variance (*i.e.*, $\sigma_\pi^2 = 0$), and it equals 1 if subjects explain all of the variance (*i.e.*, $\sigma_e^2 = 0$). Thus, as described later, it can also be interpreted as the proportion of the total variance that is attributable to subjects.

Given the above assumptions, the variance covariance matrix of the repeated measures has the "compound symmetry" structure:

$$\Sigma \boldsymbol{y}_i = \begin{bmatrix} \sigma_e^2 + \sigma_\pi^2 & \sigma_\pi^2 & \sigma_\pi^2 & \cdots & \sigma_\pi^2 \\ \sigma_\pi^2 & \sigma_e^2 + \sigma_\pi^2 & \sigma_\pi^2 & \cdots & \sigma_\pi^2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sigma_\pi^2 & \cdots & \sigma_\pi^2 & \sigma_e^2 + \sigma_\pi^2 & \sigma_\pi^2 \\ \sigma_\pi^2 & \sigma_\pi^2 & \cdots & \sigma_\pi^2 & \sigma_e^2 + \sigma_\pi^2 \end{bmatrix} , \tag{2.3}$$

where the variance is homogeneous across time $(\sigma_e^2 + \sigma_\pi^2)$, the covariances are homogeneous across time (σ_π^2), and the correlation is given by $\sigma_\pi^2/(\sigma_e^2 + \sigma_\pi^2)$. Unfortunately, compound symmetry is not very realistic for longitudinal data. First, variances often change over time, where subjects are generally more similar at the start of the trial than at the end of the trial where some have responded to treatment and others have not. Second, covariances close in time are usually greater than covariances that are further separated in time.

In the balanced case (*i.e.*, no missing data and all subjects measured on the same occasions), the ANOVA table for a model with random subject effects and fixed time effects is of the following form:

Source	df	SS	MS	E(MS)
Subjects	$N-1$	$\text{SS}_S = n\sum_{i=1}^{N}(\bar{y}_{i.} - \bar{y}_{..})^2$	$\frac{\text{SS}_S}{N-1}$	$\sigma_e^2 + n\sigma_\pi^2$
Time	$n-1$	$\text{SS}_T = N\sum_{j=1}^{n}(\bar{y}_{.j} - \bar{y}_{..})^2$	$\frac{\text{SS}_T}{n-1}$	$\sigma_e^2 +$ $N\sum(\tau_j - \tau_.)^2$
Residual	$(N-1)$ $\times(n-1)$	$\text{SS}_R = \sum_{i=1}^{N}\sum_{j=1}^{n}(y_{ij}$ $-\bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2$	$\frac{\text{SS}_R}{(N-1)(n-1)}$	σ_e^2
Total	$Nn-1$	$\text{SS}_y = \sum_{i=1}^{N}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2$		

where

SS = sum of squares

MS = mean squares

E(MS) = expected MS

$\bar{y}_{..}$ = grand mean (averaged over time and subjects)

$\bar{y}_{i.}$ = subject mean ($i = 1, \ldots, N$)

$\bar{y}_{.j}$ = timepoint mean ($j = 1, \ldots, n$)

Tests of hypothesis are constructed as follows:

$$\begin{array}{lll} H_S: & \sigma_\pi^2 = 0 & F_S = \frac{\text{MS}_S}{\text{MS}_R} \overset{H_S}{\sim} F_{N-1,(N-1)(n-1)} \\ H_T: & \tau_1 = \tau_2 = \ldots = \tau_n = 0 & F_T = \frac{\text{MS}_T}{\text{MS}_R} \overset{H_T}{\sim} F_{n-1,(N-1)(n-1)} \end{array} \tag{2.4}$$

Testing the significance of the time effect is typically the focus, because we generally usually assume that $\sigma_\pi^2 > 0$ (*i.e.*, subjects have significant influence on their longitudinal data). To quantify the degree of the subject effect, the intraclass correlation (ICC) for this design describes the relative magnitude of σ_π^2, namely

$$\text{ICC} = \frac{\hat{\sigma}_\pi^2}{\hat{\sigma}_\pi^2 + \hat{\sigma}_e^2}\,. \tag{2.5}$$

As subjects' data are highly correlated, the ICC becomes large. Here, the variance parameters are estimated as

$$\hat{\sigma}_\pi^2 = (\text{MS}_S - \text{MS}_R)\,/\,n \tag{2.6}$$

and

$$\hat{\sigma}_e^2 = \text{MS}_R. \tag{2.7}$$

When $\text{MS}_S \leq \text{MS}_R$, then $\hat{\sigma}_\pi^2 = 0$. Notice that the ICC is akin to a R^2 statistic; it represents the proportion of (unexplained) variation due to subjects. In the present model, the term "unexplained" refers to the variation that is not explained by the time effect. More generally, in models with more independent variables or covariates, "unexplained" refers to variation not explained by the set of independent variables. Thus, the value of the ICC can vary depending on model covariates. As more and more of the between-subject variability is explained by model covariates, the value of the ICC decreases.

2.1.2 Decomposing the Time Effect

As mentioned, the testing of the time effect is typically the focus in the current design. However, the overall test of the null hypothesis of no difference over time,

$$H_T: \tau_1 = \tau_2 = \ldots = \tau_n = 0 \tag{2.8}$$

is a very global test. It tests whether there is any difference in the population means across time, namely,

$$H_T: \mu_1 = \mu_2 = \ldots = \mu_n. \tag{2.9}$$

For more specific time-related comparisons, it is useful to construct a set of $n-1$ contrasts $L_{j'}$ of the timepoint means as follows:

$$L_{j'} = \sum_{j=1}^{n} c_{j'j}\,\bar{y}_{.j}\,, \qquad j' = 1, \ldots, n-1, \tag{2.10}$$

where $c_{j'j}$ represent contrast coefficients. Specific examples of sets of contrast coefficients will be given subsequently. Note, for a given contrast $L_{j'}$, there is a restriction that the sum of these contrast coefficients equals 0 across the n timepoints,

$$\sum_{j=1}^{n} c_{j'j} = 0. \tag{2.11}$$

Tests of these contrasts ($H_{j'}: L_{j'} = 0$) can be obtained using

$$\text{MS}_{j'} = \text{SS}_{j'} = \frac{NL_{j'}^2}{\sum_{j=1}^{n} c_{j'j}^2} \tag{2.12}$$

in terms of either an F or t statistic:

$$F_{j'} = \frac{\text{MS}_{j'}}{\text{MS}_R} \overset{H_{j'}}{\sim} F_{1,(N-1)(n-1)},$$

$$t_{j'} = \frac{L_{j'}}{\sqrt{\text{MS}_R\left[\sum_{j=1}^{n} \frac{c_{j'j}^2}{N}\right]}} \overset{H_{j'}}{\sim} t_{(N-1)(n-1)}.$$

Note that these two tests are the same, the F statistic is simply the square of the t statistic.

A set of $n-1$ contrasts partitions the variation attributable to time (*i.e.*, differences across time) in terms of specific timepoint comparisons. If the set of $n-1$ contrasts are orthogonal (*i.e.*, independent of each other), then

$$\text{SS}_T = \sum_{j'=1}^{n-1} \text{SS}_{j'}, \tag{2.13}$$

and we have an independent partitioning of the variation due to time. If the set of contrasts are not orthogonal, then adding the contrast sums of squares together does not equal SS $_T$.

Selecting a set of $n-1$ contrasts clearly depends on the set of scientific questions that are of interest in a particular study. Below, we present several sets that are commonly (and perhaps not so commonly) applied in longitudinal models, and we indicate some of the characteristics of each set. Before selecting a particular set for a given analysis, the analyst should think carefully and match the set with the scientific aims of the study. By choosing a set of contrasts, one is choosing how changes in the dependent variable over time are modeled in an analysis. It should be noted, however, that the overall test of the time effect (*i.e.*, the F-test of $H_T: \tau_1 = \tau_2 = \ldots = \tau_n = 0$) is unaffected by the choice of contrasts (as long as there are $n-1$ contrasts in a set). It is the decomposition of this overall test that changes as the set of contrasts varies.

2.1.2.1 Trend Analysis—Orthogonal Polynomial Contrasts One approach to testing specific time-related contrasts is to characterize the $n-1$ time effects as $n-1$ orthogonal polynomials (see Bock [1975], Draper and Smith [1981], or Fleiss [1986]). For example, the $(n-1) \times n$ contrast matrix for $n=4$ is

$$C = \begin{bmatrix} -3 & -1 & 1 & 3 \\ 1 & -1 & -1 & 1 \\ -1 & 3 & -3 & 1 \end{bmatrix} \begin{array}{ll} \div\sqrt{20} & \text{linear} \\ \div\sqrt{4} & \text{quadratic} \\ \div\sqrt{20} & \text{cubic} \end{array} \tag{2.14}$$

for linear, quadratic, and cubic trend components. Here, the rows of the matrix indicate the $n-1$ polynomial contrasts, and the columns indicate the n timepoints. Thus, the values in a given row are the contrast coefficient values $c_{j'j}$ for a particular contrast $L_{j'}$. The division sign to the right of the matrix indicates that the elements of each row are to be divided by the indicated square root quantity for that row. These quantities are simply the sum of squares of the row elements, and so dividing the elements of the matrix by these yields polynomial contrasts (*i.e.*, linear, quadratic, and cubic) that are on the same scale and thus can be more directly compared to each other. As the name implies, orthogonal polynomials are orthogonal (*i.e.*, independent of each other). Additionally, they can be useful for determining (a) the "degree" of change across time and (b) the relative contribution of each polynomial component of the trend. While the above matrix is appropriate if the timepoints are equally spaced, it is possible to generalize orthogonal polynomials for unequally spaced timepoints. We will describe this in more detail in Chapter 5. Also, for some problems it is common to specify fewer than $n-1$ contrasts, especially as n gets large, because higher-order polynomials beyond cubic, or so, can be difficult to interpret and hard to justify.

2.1.2.2 Change Relative to Baseline—Reference Cell Contrasts In some cases, we are less concerned with the form of the growth curves and are more concerned about testing whether any change has occurred whatsoever. In this case, we can construct contrasts for each timepoint relative to the first timepoint, presumably baseline, as follows (again, for the case of 4 timepoints):

$$C = \begin{bmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix}. \tag{2.15}$$

Again, here the rows indicate the contrasts and the columns the timepoints, with the values indicating the contrast coefficient values. Denoting the four timepoints as T1, T2, T3, and T4, the above three contrasts represent the timepoint differences of T2 versus T1, T3 versus T1, and T4 versus T1, respectively. These contrasts are not orthogonal, and they are sometimes called simple contrasts. Although the above matrix has the first timepoint as the reference cell, of course any of the timepoints can be treated as the reference cell (*e.g.*, the last timepoint).

2.1.2.3 Consecutive Time Comparisons—Profile Contrasts In other cases, we may have interest in determining whether each consecutive timepoint is significantly different from the immediately previous timepoint. These contrasts are sometimes referred to as "profile contrasts," and are constructed as follows:

$$C = \begin{bmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix} \tag{2.16}$$

to test T2-T1, T3-T2, T4-T3, respectively. Profile contrasts are not orthogonal, but are useful for identifying when change begins (and ends). Use of profile contrasts in repeated measures ANOVA and MANOVA models is sometimes called "profile analysis," as described in detail by Morrison [1976].

2.1.2.4 Contrasting Each Timepoint to the Mean of Subsequent Timepoints—Helmert Contrasts When interest is in contrasting each timepoint to the mean of all subsequent timepoints, we can construct Helmert contrasts [Bock, 1975] as follows:

$$C = \begin{bmatrix} 1 & -1/3 & -1/3 & -1/3 \\ 0 & 1 & -1/2 & -1/2 \\ 0 & 0 & 1 & -1 \end{bmatrix} \tag{2.17}$$

for T1 versus the average of T2, T3, and T4; T2 versus the average of T3 and T4; and T3 versus T4. Helmert contrasts are orthogonal, and are useful for "ordered" tests. They can also be reversed to compare each timepoint to the mean of the previous timepoints:

$$C = \begin{bmatrix} -1 & 1 & 0 & 0 \\ -1/2 & -1/2 & 1 & 0 \\ -1/3 & -1/3 & -1/3 & 1 \end{bmatrix} \tag{2.18}$$

for T2 versus T1; T3 versus the average of T1 and T2; and T4 versus the average of T1, T2, and T3. In this form, these contrasts are somewhat similar to profile contrasts, except that the reference is all prior timepoints rather than simply the single previous timepoint.

2.1.2.5 Contrasting Each Timepoint to the Mean of Others—Deviation Contrasts When there is interest in contrasting each timepoint to the mean of all other timepoints, deviation contrasts can be computed as

$$C = \begin{bmatrix} 1 & -1/3 & -1/3 & -1/3 \\ -1/3 & 1 & -1/3 & -1/3 \\ -1/3 & -1/3 & 1 & -1/3 \end{bmatrix} \tag{2.19}$$

for T1 versus the average of T2, T3, and T4; T2 versus the average of T1, T3, and T4; and T3 versus the average of T1, T2, and T4. Here, all timepoints except the last one are compared to the average of the other timepoints. Which timepoint is excluded can, of course, be changed. Deviation contrasts are not orthogonal, but are useful for situations in which there is "vague prior knowledge" about the changes over time.

2.1.2.6 Multiple Comparisons Although $n-1$ multiple comparisons can often be specified *a priori*, we are nevertheless faced with making multiple comparisons that will potentially lead to elevated experiment-wise Type I error rates. The most conservative solution is to use the so-called Bonferroni corrected α level, which is given as $\alpha^* = \alpha/(n-1)$. Here, each test is evaluated at the modified α^* level to ensure that the experiment-wise Type I error rate is not inflated.

A far less conservative alternative is to apply the so-called Fisher protected test logic, in which each individual test is conducted at the α level, but the individual tests are only applied when the global test $H_T : \tau_1 = \tau_2 = ... = \tau_n = 0$ is rejected. For the special case of orthogonal polynomials, we can start with the highest-order polynomial and eliminate each degree polynomial in a backwards manner until we encounter the first significant one. In a simulation study, Hummel and Sligo [1971] support the use of Fisher protected tests if n is not too large. Also, from Rosner [1995] (page 319):

> "If a few linear contrasts, which have been specified in advance, are to be tested, then it may not be necessary to use a multiple-comparisons procedure, since if such procedures are used, there will be less power to detect differences for linear contrasts whose means are truly different from zero. Conversely, if many contrasts are to be tested, which have not been specified before looking at the data, then multiple-comparisons procedures may be useful in protecting against declaring too many significant differences."

Furthermore, some notable statisticians even argue against using any kind of adjustment for multiple comparisons [Cook and Farewell, 1996; Rothman, 1990; Saville, 1990]. As one can appreciate, it is not always clear whether to adjust or how to adjust for multiple comparisons. Additionally, there are many multiple comparison adjustment procedures to choose from (*e.g.*, see Westfall et al. [1999]). In this book we will generally use the aforementioned Fisher protected test logic, specifying the contrasts *a priori*. The reader should realize, though, that this is a "grey area" of statistics with many varying opinions.

2.2 MULTIPLE-SAMPLE REPEATED MEASURES ANOVA

In the multiple-sample case, the ANOVA model for repeated measurements is referred to as a "split-plots" ANOVA model. This is a common design in randomized clinical trials, where subjects are randomized to different treatment groups and followed across time. Here, with $h = 1, \ldots, s$ groups, $i = 1, \ldots, N_h$ subjects in group h (with $N = \sum_{h=1}^{s} N_h$), and $j = 1, \ldots, n$ timepoints, the ANOVA model is:

$$y_{hij} = \mu + \gamma_h + \tau_j + (\gamma\tau)_{hj} + \pi_{i(h)} + e_{hij}, \tag{2.20}$$

where

μ = grand mean,

γ_h = effect of group h $(\sum_h \gamma_h = 0)$,

τ_j = effect of time j $(\sum_j \tau_j = 0)$,

$(\gamma\tau)_{hj}$ = interaction effect of time j and group h $[\sum_h \sum_j (\gamma\tau)_{hj} = 0]$,

$\pi_{i(h)}$ = individual difference component for subject i nested in group h,

and e_{hij} = error for subject i in group h at time j.

The distributional assumptions for this model are the same as the previous randomized blocks ANOVA, namely,

$$\pi_{i(h)} \sim N(0, \sigma_\pi^2) \quad \text{and} \quad e_{hij} \sim N(0, \sigma_e^2),$$

which imply the same compound symmetry structure for $V(\boldsymbol{y}_i)$ as in (2.3). Also, as in the randomized blocks ANOVA, the model is a mixed model because subjects are considered random effects and group and time are considered fixed effects. The data are assumed to be balanced in terms of n (*i.e.*, timepoints), but not necessarily in terms N_h, (*i.e.*, group sample sizes). An example data layout is

		Timepoint			
Group	Subject	1	2	...	n
1	1	y_{111}	y_{112}	...	y_{11n}
1	2	y_{121}	y_{122}	...	y_{12n}
1	.	.	.	...	.
1	.	.	.	...	.
1	N_1	$y_{1N_1 1}$	$y_{1N_1 2}$	...	$y_{1N_1 n}$
.	.	.	.	...	.
.	.	.	.	...	.
.	.	.	.	...	.
.	.	.	.	...	.
s	1	y_{s11}	y_{s12}	...	y_{s1n}
s	2	y_{s21}	y_{s22}	...	y_{s2n}
s	.	.	.	...	.
s	.	.	.	...	.
s	N_s	y_{sN_s1}	y_{sN_s2}	...	y_{sN_sn}

Here, subjects are nested within groups and crossed with the time factor. Thus, each subject is observed only within a single group, and each subject is observed once at each timepoint. The ANOVA table is as follows:

Source	df	SS	MS	E(MS)
Group	$s-1$	$\mathrm{SS}_G = n\sum_{h=1}^{s} N_h\,(\bar{y}_{h..} - \bar{y}_{...})^2$	$\frac{\mathrm{SS}_G}{s-1}$	$\sigma_e^2\ + n\sigma_\pi^2 + D_G$
Time	$n-1$	$\mathrm{SS}_T = N\sum_{j=1}^{n}(\bar{y}_{..j} - \bar{y}_{...})^2$	$\frac{\mathrm{SS}_T}{n-1}$	$\sigma_e^2\ + D_T$
Group × Time	$(s-1)$ $\times(n-1)$	$\mathrm{SS}_{GT} = \sum_{h=1}^{s}\sum_{j=1}^{n} N_h\,(\bar{y}_{h.j}$ $-\bar{y}_{h..} - \bar{y}_{..j} + \bar{y}_{...})^2$	$\frac{\mathrm{SS}_{GT}}{(s-1)(n-1)}$	$\sigma_e^2\ + D_{GT}$
Subjects in Grps	$N-s$	$\mathrm{SS}_{S(G)} = n\sum_{h=1}^{s}\sum_{i=1}^{N_h}(\bar{y}_{hi.} - \bar{y}_{h..})^2$	$\frac{\mathrm{SS}_{S(G)}}{N-s}$	$\sigma_e^2 + n\sigma_\pi^2$
Residual	$(N-s)$ $\times(n-1)$	$\mathrm{SS}_R = \sum_{h=1}^{s}\sum_{i=1}^{N_h}\sum_{j=1}^{n}(y_{hij}$ $-\bar{y}_{h.j} - \bar{y}_{hi.} + \bar{y}_{h..})^2$	$\frac{\mathrm{SS}_R}{(N-s)(n-1)}$	σ_e^2
Total	$Nn-1$	$\mathrm{SS}_y = \sum_{h=1}^{s}\sum_{i=1}^{N_h}\sum_{j=1}^{n}(y_{hij} - \bar{y}_{...})^2$		

Here, D_G, D_T, and D_{GT} represent differences among groups, timepoints, and group by time interaction, respectively. Also, in terms of notation, the bar (*i.e.*, $\bar{y}$) indicates averaging and the dot subscript indicates the unit(s) that the averaging is over. Thus,

$\bar{y}_{...}$ = average across groups, timepoints, and subjects,
$\bar{y}_{h..}$ = average for group h across timepoints and subjects,
$\bar{y}_{..j}$ = average for timepoint j across groups and subjects,
$\bar{y}_{hi.}$ = average for subject i in group h across timepoints,
$\bar{y}_{h.j}$ = average for group h at timepoint j across subjects.

2.2.1 Testing for Group by Time Interaction

The group by time interaction, which is typically the test of primary interest, is constructed as

$$H_{GT} : D_{GT} = 0, \qquad F_{GT} = \frac{\mathrm{MS}_{GT}}{\mathrm{MS}_R} \overset{H_{GT}}{\sim} F_{(s-1)(n-1),(N-s)(n-1)}. \tag{2.21}$$

If the null hypothesis of no group by time interaction is rejected, we conclude that (a) the between-group differences are not the same across time, (b) the between-group curves across time are not parallel, and (c) group and time effects are confounded with the interaction and cannot be separately tested (or estimated). In this case, there is no "one" overall group effect, because it varies across time. Likewise, there is no single overall time effect, because it varies by groups.

If the null hypothesis of no group by time interaction cannot be rejected, then tests of the main effects of time and group are, respectively,

$$H_T : \tau_1 = \tau_2 = \ldots = \tau_n = 0, \qquad F_T = \frac{\mathrm{MS}_T}{\mathrm{MS}_R} \overset{H_T}{\sim} F_{n-1,(N-s)(n-1)}, \tag{2.22}$$

$$H_G : \gamma_1 = \gamma_2 = \ldots = \gamma_s = 0, \qquad F_G = \frac{\mathrm{MS}_G}{\mathrm{MS}_{S(G)}} \overset{H_G}{\sim} F_{s-1,N-s}. \tag{2.23}$$

In this case, the main effects of time (H_T) and group (H_G) are separately and independently testable. Note that the correct denominator for the F test of the group effect is not the usual error term MS_R, but instead the subjects within groups mean squares $\mathrm{MS}_{S(G)}$.

2.2.2 Testing for Subject Effect

To test for the significance of random subject effects, we construct the following statistic:

$$H_{S(G)} : \sigma_\pi^2 = 0, \qquad F_{S(G)} = \frac{\mathrm{MS}_{S(G)}}{\mathrm{MS}_R} \overset{H_{S(G)}}{\sim} F_{N-s,(N-s)(n-1)}. \tag{2.24}$$

As in the randomized blocks ANOVA, we generally assume that $\sigma_\pi^2 > 0$, and estimate the intraclass correlation as

$$ICC = \hat{\sigma}_\pi^2/(\hat{\sigma}_\pi^2 + \hat{\sigma}_e^2) \tag{2.25}$$

The intraclass correlation represents the proportion of (unexplained) variation in the dependent variable that is due to subjects. Here, "unexplained" refers to the variation not explained by the fixed effects of the model: group, time, and group by time. As the intraclass correlation approaches zero, we can conclude that there is little correlation among the repeated observations over time, and a traditional fixed-effects ANOVA model will suffice. However, this is rarely the case, it is much more common for the intraclass correlation to be moderately large for most longitudinal data.

2.2.3 Contrasts for Time Effects

As in the single-group case, it is advantageous to characterize time and group by time effects using time-related contrasts. These include the same sets of contrasts described earlier for single-group designs, namely, orthogonal polynomials, profile contrasts, Helmert contrasts etc. As before, it may be necessary to consider the multiple comparisons issue and utilize Fisher protected tests or Bonferroni correction for the $(n-1)(s-1)$ group by time contrasts (and for the $n-1$ time contrasts).

2.2.3.1 Orthogonal Polynomial Partition of SS An example orthogonal polynomial decomposition of the time effect for a four-timepoint design is given by

$$C = \begin{bmatrix} -3 & -1 & 1 & 3 \\ 1 & -1 & -1 & 1 \\ -1 & 3 & -3 & 1 \end{bmatrix} \begin{matrix} \div\sqrt{20} & c_1, \\ \div\sqrt{4} & c_2, \\ \div\sqrt{20} & c_3. \end{matrix} \tag{2.26}$$

The decomposition of the Time sum of squares is given by

Time	df	SS
Linear	1	$\mathrm{SS}_{T_1} = N\, c_1\, \bar{y}_{..}\, \bar{y}'_{..}\, c'_1$
Quadratic	1	$\mathrm{SS}_{T_2} = N\, c_2\, \bar{y}_{..}\, \bar{y}'_{..}\, c'_2$
⋮	⋮	⋮
$(n-1)$th	1	$\mathrm{SS}_{T_{n-1}} = N\, c_{n-1}\, \bar{y}_{..}\, \bar{y}'_{..}\, c'_{n-1}$
	$n-1$	SS_T

where $\bar{y}_{..}$ is the $n \times 1$ vector of timepoint means (*i.e.*, the averages at each timepoint, averaging over groups and subjects), and c_j is the $1 \times n$ vector of contrasts of order j. Corresponding F-statistics for each trend component are

$$\begin{aligned} F_{T_{n-1}} &= \mathrm{SS}_{T_{n-1}} \,/\, \mathrm{MS}_R, \\ &\;\;\vdots \\ F_{T_1} &= \mathrm{SS}_{T_1} \,/\, \mathrm{MS}_R. \end{aligned}$$

Typically, we determine the polynomial of least degree by working backwards. Similarly, the Group by Time sum of squares is decomposed as

G × T	df	SS
Linear	$s-1$	$\mathrm{SS}_{GT_1} = \sum_h N_h\, c_1\, \bar{y}_{h.}\, \bar{y}'_{h.}\, c'_1 - \mathrm{SS}_{T_1}$
Quadratic	$s-1$	$\mathrm{SS}_{GT_2} = \sum_h N_h\, c_2\, \bar{y}_{h.}\, \bar{y}'_{h.}\, c'_2 - \mathrm{SS}_{T_2}$
⋮	⋮	⋮
$(n-1)$th	$s-1$	$\mathrm{SS}_{GT_{n-1}} = \sum_h N_h\, c_{n-1}\, \bar{y}_{h.}\, \bar{y}'_{h.}\, c'_{n-1} - \mathrm{SS}_{T_{n-1}}$
	$(s-1)(n-1)$	SS_{GT}

where $\bar{y}_{h.}$ is the $n \times 1$ vector of timepoint means for group h, and c_j is the $1 \times n$ vector of contrasts of order j. F-statistics are given by

$$\begin{aligned} F_{GT_{n-1}} &= \frac{\mathrm{SS}_{GT_{n-1}} \,/\, (s-1)}{\mathrm{MS}_R}, \\ &\;\;\vdots \\ F_{GT_1} &= \frac{\mathrm{SS}_{GT_1} \,/\, (s-1)}{\mathrm{MS}_R}. \end{aligned}$$

Again, the order of the polynomial of least degree for the Group by Time interaction is determined working backwards from highest degree (most complex) to lowest degree (*i.e.*, linear).

2.2.4 Compound Symmetry and Sphericity

For both univariate randomized blocks and split-plot ANOVA models, the variance of the vector of responses is given by

$$V(\boldsymbol{y}_i) = \sigma_\pi^2 \mathbf{1}_n \mathbf{1}'_n + \sigma_e^2 I_n. \tag{2.27}$$

As mentioned, the compound symmetry (CS) structure is given by

$$\begin{aligned} V(y_{ij}) &= \sigma_\pi^2 + \sigma_e^2 && \forall\, j, \\ C(y_{ij}, y_{ij'}) &= \sigma_\pi^2 && \forall\, j \text{ and } j' \;\; (j \neq j'), \end{aligned} \tag{2.28}$$

where ∀ is the mathematical symbol denoting "for all." CS implies that variances are equal across time and that the covariances are all equal. Likewise, the correlation between responses across timepoints, given by the intraclass correlation,

$$Corr(y_{ij}, y_{ij'}) = \sigma_\pi^2 \,/\, (\sigma_\pi^2 + \sigma_e^2), \tag{2.29}$$

is the same across all pairs of timepoints. The CS assumption is highly restrictive, and often unrealistic (especially as n gets large). CS is a special case of the more general situation, sphericity, under which F-tests for time-related terms from ANOVA models are valid. If sphericity holds, then these F-tests are valid, otherwise, if sphericity doesn't hold, then these F-tests are generally too liberal.

2.2.4.1 Sphericity Sphericity, sometimes called circularity, can be expressed in different ways. The most general is that all variances of all pairwise differences between variables are equal

$$\begin{aligned} V(y_{ij} - y_{ij'}) &= V(y_{ij}) + V(y_{ij'}) - 2C(y_{ij}, y_{ij'}) \\ &= \text{constant} \; \forall \, j \text{ and } j'. \end{aligned}$$

The variance–covariance structure of compound symmetry satisfies this condition because the variances are all the same, as are the covariances. More generally, Crowder and Hand [1990] (page 50) note:

> "MS ratios derived by the univariate approach follow exact F-distributions if and only if the covariance matrix of the orthonormal contrasts has equal variances and zero covariances."

This statement is equivalent to

$$\underset{(n-1)\times n}{C} \quad \underset{n\times n}{V(\boldsymbol{y}_i)} \quad \underset{(n-1)\times n}{C'} \quad = \quad \text{constant} \quad \underset{(n-1)\times(n-1)}{I_{t-1}}, \tag{2.30}$$

where C is a matrix of orthonormal polynomials (*i.e.*, orthogonal polynomials with unit variance). Testing whether sphericity holds is therefore a test of whether the matrix product above results in the form on the right-hand side of the equality. A chi-square goodness-of-fit test of this was developed by Mauchly [1940] and is implemented in many standard statistical software packages. As Mauchly noted, this sphericity test has relatively low statistical power for small sample sizes. Alternatively, for large samples, the test is likely to be significant even though the effect on the F-test may be negligible. The sphericity test is sensitive to departures from normality and to the presence of outliers. As such, it should be used as a guide and not as a strict rule.

If the sphericity assumption is rejected, or deemed implausible, one can use a multivariate repeated measures analysis (MANOVA), which allows for general $V(\boldsymbol{y}_i)$ but does not allow for any missing data across time. In the following chapter we discuss the MANOVA approach to analysis of longitudinal data. Other classical statistical alternatives include adjusted univariate F-tests as described by Greenhouse and Geisser [1959] and by Huynh and Feldt [1976], both of which are generally overly conservative.

2.3 ILLUSTRATION

Bock [1975] presents data on vocabulary growth measured in a cohort of 64 students at the University of Chicago Laboratory school. The longitudinal data consist of repeated measurements of the vocabulary section of the cooperative reading test [Davis, 1950]. Alternate forms of the test were administered in eighth through eleventh grade. Since this age range marks the period of time that physical growth begins to decelerate, he hypothesized that a similar deceleration might be observed in the acquisition of new vocabulary as well. Figure 2.1 displays a plot of the average scores versus grade and visually suggests that the rate of change is indeed decelerating.

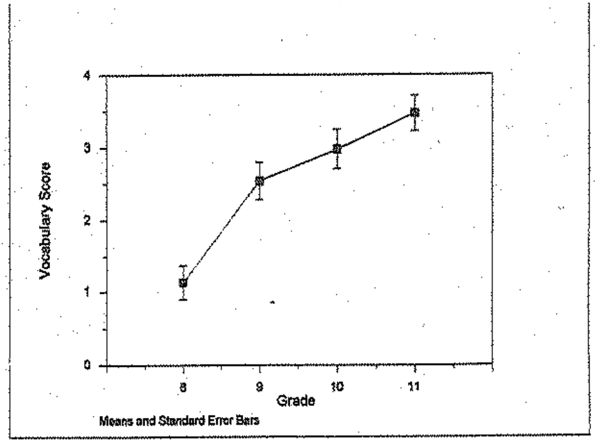

Figure 2.1. Average vocabulary scores of 64 students.

Summary statistics (means, standard deviations, and correlations) are presented in Table 2.1. Inspection of these summary statistics reveals that the variances and covariances are reasonably homogeneous over time, supporting the assumption of compound symmetry underlying the repeated measures ANOVA. Performing Mauchly's test yields a chi-square statistic of 6.32, on five degrees of freedom, which is not statistically significant. Thus, the assumption of sphericity, and therefore compound symmetry, is reasonable for these data.

Table 2.1. Means, Standard Deviations, and Correlations for the Vocabulary-Growth Data

Grade	Mean	SD	Correlations			
8	1.137	1.889	1.000			
9	2.542	2.085	.810	1.000		
10	2.988	2.169	.868	.785	1.000	
11	3.472	1.925	.785	.757	.811	1.000

This example can be used to nicely illustrate the simplest case of a one-sample repeated measures ANOVA. In the model, subject represents a random effect, and time represents a fixed effect. The ANOVA is illustrated in Table 2.2, which presents the ANOVA results.

Table 2.2. Repeated Measures ANOVA Results for the Vocabulary-Growth Data

Source	df	SS	MS	F	$p <$
Subjects	63	$\text{SS}_S = 873.60$	13.87	16.91	.0001
Grade (*i.e.*, "Time")	3	$\text{SS}_T = 194.34$	64.78	79.02	.0001
Residual	189	$\text{SS}_R = 154.94$	0.82		
Total	255	$\text{SS}_y = 1{,}222.88$			

The estimate of the error variance, which equals MS $_R$, is

$$\hat{\sigma}_e^2 = \frac{\text{SS}_R}{(N-1)(n-1)} = \frac{154.94}{189} = 0.82.$$

The estimate of the subject variance is gotten as

$$\hat{\sigma}_\pi^2 = \frac{\text{MS}_S - \text{MS}_R}{n} = \frac{13.87 - 0.82}{4} = 3.26,$$

and the intraclass correlation equals

$$ICC = \frac{\hat{\sigma}_\pi^2}{\hat{\sigma}_\pi^2 + \hat{\sigma}_e^2} = \frac{3.26}{3.26 + 0.82} = .80.$$

Thus, as one would expect, there is a tremendous effect of subjects on their vocabulary scores. 80% of the variation in vocabulary, that is not explained by grade, is attributable to subjects.

In terms of the grade effect, the ANOVA table reveals that we must reject the null hypothesis of no grade effect. This is clearly supported by Figure 2.1, which shows that vocabulary generally increases with grade. However, to obtain a more sensitive analysis, we can examine the significance of the individual polynomial terms based on the 4-timepoint orthogonal polynomial matrix:

$$C = \begin{bmatrix} -.67082 & -.22361 & .22361 & .67082 \\ .5 & -.5 & -.5 & .5 \\ -.22361 & .67082 & -.67082 & .22361 \end{bmatrix}.$$

Premultiplying the 4×1 vector of column means

$$\bar{y} = \begin{bmatrix} 1.14 \\ 2.54 \\ 2.99 \\ 3.47 \end{bmatrix},$$

by C, we obtain the following orthogonal estimates:

$$\text{Linear} = 1.67, \quad \text{Quadratic} = -0.46, \quad \text{Cubic} = 0.22.$$

Note that the squares of these estimates multiplied by 64 (*i.e.*, the number of subjects) are the numerators of the F-ratios for linear, quadratic, and cubic trends shown in Table 2.3. The denominator is $\hat{\sigma}_e^2 = 0.82$ and the denominator degrees of freedom equals 189 for these F-ratios. Clearly, the positive linear trend is highly significant. However, the significant quadratic term, coupled with its negative sign, reveals that the observed deceleration in the growth rate is statistically significant. Finally, the cubic term is only marginally significant suggesting that the deceleration reverses to some extent with increasing age (see Figure 2.1). Examining the trend estimates, we see that the cubic estimate also pales in comparison with the dominant linear and moderate quadratic trend components: There is clearly diminishing importance as the order of the polynomial is increased.

Table 2.3. Orthogonal Polynomial Decomposition of the Grade Effect

Source	df	SS	F	$p <$
Grade				
Linear	1	$\text{SS}_{\text{T}_1} = 177.58$	216.63	.0001
Quadratic	1	$\text{SS}_{\text{T}_2} = 13.58$	16.56	.0001
Cubic	1	$\text{SS}_{\text{T}_3} = 3.17$	3.86	.051

Taken together, these results indicate a decelerating positive trend across grade, supporting the notion that vocabulary acquisition is slowing down as students approach maturity.

2.4 SUMMARY

In summary, the ANOVA approach to analysis of longitudinal data represents a well-understood and well-developed statistical methodology. In addition, there is considerable available computer software for ANOVA computation. The results are based on relatively simple and noniterative calculations. Unfortunately, the ANOVA model for repeated measurements assumes sphericity, which is unrealistic for most applications where variances tend to increase with time and correlation decreases with increasing intervals in time. Other limitations include limited treatment of missing data, and the requirement that all subjects are measured on the same occasions.