CHAPTER 17

Sample Sizes for Observational Studies

17.1 INTRODUCTION

In this chapter we deal with the problem of calculating sample sizes in various observational settings. There is a very diverse literature on sample size calculations, dealing with many interesting areas. We can only give you a feeling for some approaches and some pointers for further study.

We start the chapter by considering the topic of screening in the context of adverse effects attributable to drug usage, trying to accommodate both the "rare disease" assumption and the multiple comparison problem. Section 17.3 discusses sample-size considerations when costs of observations are not equal, or the variability is unequal; some very simple but elegant relationships are derived. Section 17.4 considers sample size consideration in the context of discriminant analysis. Three questions are considered: (1) how to select variables to be used in discriminating between two populations in the face of multiple comparisons; (2) given that m variables have been selected, what sample size is needed to discriminate between two populations with satisfactory power; and (3) how large a sample size is needed to estimate the probability of correct classification with adequate precision and power. Notes, problems, and references complete the chapter.

17.2 SCREENING STUDIES

A screening study is a scientific fishing expedition: for example, attempting to relate exposure to one of several drugs to the presence or absence of one or more side effects (disease). In such screening studies the number of drug categories is usually very large—500 is not uncommon—and the number of diseases is very large—50 or more is not unusual. Thus, the number of combinations of disease and drug exposure can be very large—25,000 in the example above. In this section we want to consider the determination of sample size in screening studies in terms of the following considerations: many variables are tested and side effects are rare. A cohort of exposed and unexposed subjects is either followed or observed. We have looked at many diseases or exposures, want to "protect" ourselves against a large Type I error, and want to know how many observations are to be taken. We proceed in two steps: First, we derive the formula for the sample size without consideration of the multiple testing aspect, then we incorporate the multiple testing aspect. Let

$$X_1 = \text{number of occurrences of a disease of interest (per 100,000 person-years, say) in the unexposed population}$$

Biostatistics: A Methodology for the Health Sciences, Second Edition, by Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley
ISBN 0-471-03185-2

X_2 = number of occurrences (per 100,000 person-years) in the exposed population

If X_1 and X_2 are rare events, $X_1 \sim \text{Poisson}(\theta_1)$ and $X_2 \sim \text{Poisson}(\theta_2)$. Let $\theta_2 = R\theta_1$; that is, the risk in the exposed population is R times that in the unexposed population ($0 < R < \infty$). We can approximate the distributions by using the variance stabilizing transformation (discussed in Chapter 10):

$$Y_1 = \sqrt{X_1} \sim N(\sqrt{\theta_1}, \sigma^2 = 0.25)$$
$$Y_2 = \sqrt{X_2} \sim N(\sqrt{\theta_2}, \sigma^2 = 0.25)$$

Assuming independence,

$$Y_2 - Y_1 \sim N\left(\sqrt{\theta_1}(\sqrt{R} - 1), \sigma^2 = 0.5\right) \tag{1}$$

For specified Type I and Type II errors α and β, the *number of events* n_1 and n_2 in the unexposed and exposed groups required to detect a relative risk of R with power $1 - \beta$ are given by the equation

$$n_1 = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2}{2(\sqrt{R} - 1)^2}, \qquad n_2 = Rn_1 \tag{2}$$

Equation (2) assumes a two-sided, two-sample test with an equal number of subjects observed in each group. It is an approximation, based on the normality of the square root of a Poisson random variable. If the prevalence, π_1, in the unexposed population is known, the number of subjects per group, N, can be calculated by using the relationship

$$N\pi_1 = n_1 \quad \text{or} \quad N = n_1/\pi_1 \tag{3}$$

Example 17.1. In Section 15.4, mortality was compared in active participants in an exercise program and in dropouts. Among the active participants, there were 16 deaths in 593 person-years of active participation; in dropouts there were 34 deaths in 723 person-years. Using an α of 0.05, the results were not significantly different. The relative risk, R, for dropouts is estimated by

$$R = \frac{34/723}{16/593} = 1.74$$

Assuming equal exposure time in the active participants and dropouts, how large should the sample sizes n_1 and n_2 be to declare the relative risk, $R = 1.74$, significant at the 0.05 level with probability 0.95? In this case we use a two-tailed test and $Z_{1-\alpha/2} = 1.960$ and $Z_{1-\beta} = 1.645$, so that

$$n_1 = \frac{(1.960 + 1.645)^2}{2(\sqrt{1.74} - 1)^2} = 63.4 \doteq 64 \quad \text{and} \quad n_2 = (1.74)n_1 = 111$$

for a total number of observed events $= n_1 + n_1 = 64 + 111 = 175$ deaths. We would need approximately $(111/34) \times 723 = 2360$ person-years exposure in the dropouts and the same number of years of exposure among the controls. The exposure years in the observed data are not split equally between the two groups. We discuss this aspect further in Note 17.1.

If there is only one observational group, the group's experience perhaps being compared with that of a known population, the sample size required is $n_1/2$, again illustrating the fact that comparing two groups requires four times more exposure time than comparing one group with a known population.

Table 17.1 Relationship between Overall Significance Level α, Significance Level per Test, Number of Tests, and Associated Z-Values, Using the Bonferroni Inequality

Number of Tests (K)	Overall α Level	Required Level per Test (α)	Z-Values	
			One-Tailed	Two-Tailed
1	0.05	0.05	1.645	1.960
2	0.05	0.025	1.960	2.241
3	0.05	0.01667	2.128	2.394
4	0.05	0.0125	2.241	2.498
5	0.05	0.01	2.326	2.576
10	0.05	0.005	2.576	2.807
100	0.05	0.0005	3.291	3.481
1000	0.05	0.00005	3.891	4.056
10000	0.05	0.000005	4.417	4.565

We now turn to the second aspect of our question. Suppose that the comparison above is one of a multitude of comparisons? To maintain a per experiment significance level of α, we use the Bonferroni inequality to calculate the per comparison error rate. Table 17.1 relates the per comparison critical values to the number of tests performed and the per experiment error rate. It is remarkable that the critical values do not increase too rapidly with the number of tests.

Example 17.2. Suppose that the FDA is screening a large number of drugs, relating 10 kinds of congenital malformations to 100 drugs that could be taken during pregnancy. A particular drug and a particular malformation is now being examined. Equal numbers of exposed and unexposed women are to be selected and a relative risk of $R = 2$ is to be detected with power 0.80 and per experiment one-sided error rate of $\alpha = 0.05$. In this situation $\alpha^* = \alpha/1000$ and $Z_{1-\alpha^*} = Z_{1-\alpha/1000} = Z_{0.99995} = 3.891$. The required number of events in the unexposed group is

$$n_1 = \frac{(3.891 + 0.842)^2}{2(\sqrt{2} - 1)^2} = \frac{22.4013}{0.343146} = 65.3 \doteq 66$$

$$n_2 = 2n_1 = 132$$

In total, $66 + 132 = 198$ malformations must be observed. For a particular malformation, if the congenital malformation rate is on the order of 3/1000 live births, approximately 22,000 unexposed women and 22,000 women exposed to the drug must be examined. This large sample size is not only a result of the multiple testing but also the rarity of the disease. [The comparable number testing only once, $\alpha^* = \alpha = 0.05$, is $n_1 = \frac{1}{2}(1.645 + 0.842)^2/(\sqrt{2} - 1)^2 = 18$, or 3000 women per group.]

17.3 SAMPLE SIZE AS A FUNCTION OF COST AND AVAILABILITY

17.3.1 Equal-Variance Case

Consider the comparison of means from two independent groups with the same variance σ; the standard error of the difference is

$$\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{4}$$

where n_1 and n_2 are the sample sizes in the two groups. As is well known, for fixed N the standard error of the difference is minimized (maximum precision) when

$$n_1 = n_2 = N$$

That is, the sample sizes are equal. Suppose now that there is a differential cost in obtaining the observations in the two groups; then it may pay to choose n_1 and n_2 unequal, subject to the constraint that the standard error of the difference remains the same. For example,

$$\frac{1}{10} + \frac{1}{10} = \frac{1}{6} + \frac{1}{30}$$

Two groups of equal sample size, $n_1 = n_2 = 10$, give the same precision as two groups with $n_1 = 6$ and $n_2 = 30$. Of course, the total number of observations N is larger, 20 vs. 36.

In many instances, sample size calculations are based on additional considerations, such as:

1. Relative cost of the observations in the two groups
2. Unequal hazard or potential hazard of treatment in the two groups
3. The limited number of observations available for one group

In the last category are case–control studies where the number of cases is limited. For example, in studying sudden infant death syndrome (SIDS) by means of a case–control study, the number of cases in a defined population is fairly well fixed, whereas an arbitrary number of (matching) controls can be obtained.

We now formalize the argument. Suppose that there are two groups, G_1 and G_2, with costs per observations c_1 and c_2, respectively. The total cost, C, of the experiment is

$$C = c_1 n_1 + c_2 n_2 \tag{5}$$

where n_1 and n_2 are the number of observations in G_1 and G_2, respectively. The values of n_1 and n_2 are to be chosen to minimize (maximum precision),

$$\frac{1}{n_1} + \frac{1}{n_2}$$

subject to the constraint that the total cost is to be C. It can be shown that under these conditions the required sample sizes are

$$n_1 = \frac{C}{c_1 + \sqrt{c_1 c_2}} \tag{6}$$

and

$$n_2 = \frac{C}{c_2 + \sqrt{c_1 c_2}} \tag{7}$$

The ratio of the two sample sizes is

$$\frac{n_2}{n_1} = \sqrt{\frac{c_1}{c_2}} = h, \qquad \text{say} \tag{8}$$

That is, if costs per observation in groups G_1 and G_2, are c_1 and c_2, respectively, then choose n_1 and n_2 on the basis of the ratio of the square root of the costs. This rule has been termed the *square root rule* by Gail et al. [1976]; the derivation can also be found in Nam [1973] and Cochran [1977].

If the costs are equal, $n_1 = n_2$, as before. Application of this rule can decrease the cost of an experiment, although it will increase the total number of observations. Note that the population means and standard deviation need not be known to determine the ratio of the sample sizes, only the costs. If the desired precision is specified—perhaps on the basis of sample size calculations assuming equal costs—the values of n_1 and n_2 can be determined. Compared with an experiment with equal sample sizes, the ratio ρ of the costs of the two experiments can be shown to be

$$\rho = \frac{1}{2} + \frac{h}{1+h^2} \tag{9}$$

If $h = 1$, then $\rho = 1$, as expected; if h is very close to zero or very large, $\rho = \frac{1}{2}$; thus, no matter what the relative costs of the observations, the savings can be no larger than 50%.

Example 17.3. (After Gail et al. [1976]) A new therapy, G_1, for hypertension is introduced and costs \$400 per subject. The standard therapy, G_2, costs \$16 per subject. On the basis of power calculations, the precision of the experiment is to be equivalent to an experiment using 22 subjects per treatment, so that

$$\frac{1}{22} + \frac{1}{22} = 0.09091$$

The square root rule specifies the ratio of the number of subjects in G_1 and G_2 by

$$\begin{aligned} n_2 &= \sqrt{\frac{400}{16}} n_1 \\ &= 5n_1 \end{aligned}$$

To obtain the same precision, we need to solve

$$\frac{1}{n_1} + \frac{1}{5n_1} = 0.09091$$

or

$$n_1 = 13.2 \quad \text{and} \quad n_2 = 66.0$$

(i.e., $1/13.2 + 1/66.0 = 0.09091$, the same precision). Rounding up, we require 14 observations in G_1 and 66 observations in G_2. The costs can also be compared as in Table 17.2.

A savings of \$3896 has been obtained, yet the precision is the same. The total number of observations is now 80, compared to 44 in the equal-sample-size experiment. The ratio of the savings is

$$\rho = \frac{6656}{9152} = 0.73$$

Table 17.2 Costs Comparisons for Example 17.3

	Equal Sample Size		Sample Size Determined by Cost	
	n	Cost	n	Cost
G_1	22	8800	14	5600
G_2	22	352	66	1056
Total	44	9152	80	6656

The value for ρ calculated from equation (9) is

$$\rho = \frac{1}{2} + \frac{5}{26} = 0.69$$

The reason for the discrepancy is the rounding of sample sizes to integers.

17.3.2 Unequal-Variance Case

Suppose that we want to compare the means from groups with unequal variance. Again, suppose that there are n_1 and n_2 observations in the two groups. Then the standard error of the difference between the two means is

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Let the ratio of the variances be $\eta^2 = \sigma_2^2/\sigma_1^2$. Gail et al. [1976] show that the sample size should now be allocated in the ratio

$$\frac{n_2}{n_1} = \sqrt{\frac{\sigma_2^2}{\sigma_1^2}\frac{c_1}{c_2}} = \eta h$$

The calculations can then be carried out as before. In this case, the cost relative to the experiment with equal sample size is

$$\rho^* = \frac{(h+\eta)^2}{(1+h^2)(1+\eta^2)} \tag{10}$$

These calculations also apply when the costs are equal but the variances unequal, as is the case in binomial sampling.

17.3.3 Rule of Diminishing Precision Gain

One of the reasons advanced at the beginning of Section 17.3 for distinguishing between the sample sizes of two groups is that a limited number of observations may be available for one group and a virtually unlimited number in the second group. Case–control studies were cited where the number of cases per population is relatively fixed. Analogous to Gail et al. [1976], we define a rule of diminishing precision gain. Suppose that there are n cases and that an unlimited number of controls are available. Assume that costs and variances are equal. The precision of the difference is then proportional to

$$\sigma\sqrt{\frac{1}{n} + \frac{1}{hn}}$$

where hn is the number of controls selected for the n cases.

We calculate the ratio P_h:

$$P_h = \frac{\sqrt{1/n + 1/hn}}{\sqrt{1/n + 1/n}}$$

$$= \sqrt{\frac{1}{2}\left(1 + \frac{1}{h}\right)}$$

This ratio P_h is a measure of the precision of a case–control study with n and hn cases and controls, respectively, relative to the precision of a study with an equal number, n, of cases and controls. Table 17.3 presents the values of P_h and $100(P_h - P_\infty)/P_\infty$ as a function of h.

Table 17.3 Comparison of Precision of Case Control Study with n and hn Cases and Controls, Respectively

h	P_h	$100[(P_h - P_\infty)/P_\infty]\%$
1	1.00	41
2	0.87	22
3	0.82	15
4	0.79	12
5	0.77	10
10	0.74	5
∞	0.71	0

This table indicates that in the context above, the gain in precision with, say, more than four controls per case is minimal. At $h = 4$, one obtains all but 12% of the precision associated with a study using an infinite number of controls. Hence, in the situation above, there is little merit in obtaining more than four or five times as many controls as cases. Lubin [1980] approaches this from the point of view of the logarithm of the odds ratio and comes to a similar conclusion.

17.4 SAMPLE-SIZE CALCULATIONS IN SELECTING CONTINUOUS VARIABLES TO DISCRIMINATE BETWEEN POPULATIONS

In certain situations, there is interest in examining a large number of continuous variables to explain the difference between two populations. For example, an investigator might be "fishing" for clues explaining the presence (one population) or absence (the other population) of a disease of unknown etiology. Or in a disease where a variety of factors are known to affect prognosis, the investigator may desire to find a good set of variables for predicting which subjects will survive for a fixed number of years. In this section, the determination of sample size for such studies is discussed.

There are a variety of approaches to the data analysis in this situation. With a large, say 50 or more, number of variables, we would hesitate to run stepwise discriminant analysis to select a few important variables, since (1) in typical data sets there are often many dependencies that make the method numerically unstable (i.e., the results coming forth from some computers cannot be relied on); (2) the more complex the mathematical model used, the less faith we have that it is useful in other situations (i.e., the more parameters that are used and estimated, the less confidence we can have that the result is transportable to another population in time or space; here we might be envisioning a discriminant function with a large number of variables); and (3) the multiple-comparison problems inherent in considering the large number of variables at each step in the stepwise procedure make the result of doubtful value.

One approach to the analysis is first to perform a *univariate screen*. This means that variables (used singly, that is, univariately) with the most power to discriminate between the two populations are selected. Second, use these univariate discriminating variables in the discriminant analysis. The sample-size calculations below are based on this method of analysis. There is some danger in this approach, as variables that univariately are not important in discrimination could be important when used in conjunction with other variables. In many practical situations, this is not usually the case. Before discussing the sample-size considerations, we will consider a second approach to the analysis of such data as envisioned here.

Often, the discriminating variables fall naturally in smaller subsets. For example, the subsets for patients may involve data from (1) the history, (2) a physical exam, and (3) some routine tests. In many situations the predictive information of the variables within each subset is roughly

the same. This being the case, a two-step method of selecting the predictive variables is to (1) use stepwise selection within subsets to select a few variables from each subset, and (2) combine the selected variables into a group to be used for another stepwise selection procedure to find the final subset of predictive variables.

After selecting a smaller subset of variables to use in the prediction process, one of two steps is usually taken. (1) The predictive equation is validated (tested) on a new sample to show that it has predictive power. That is, an F-test for the discriminant function is performed. Or, (2) a larger independent sample is used to provide an indication of the accuracy of the prediction. The second approach requires a larger sample size than merely establishing that there is some predictive ability, as in the first approach. In the next three sections we make this general discussion precise.

17.4.1 Univariate Screening of Continuous Variables

To obtain an approximate idea of the sample size needed to screen among k variables, the following is assumed: The variables are normally distributed with the same variance in each population and possibly different means. The power to classify into the two populations depends on δ, the number of standard deviations distance between the two populations means:

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}$$

Some idea of the relationship of classificatory power to δ is given in Figure 17.1.

Suppose that we are going to screen k variables and want to be sure, with probability at least $1 - \alpha$, to include all variables with $\delta \geq D$. In this case we must be willing to accept some variables with values close to but less than D. Suppose that at the same time we want probability at least $1 - \alpha$ of not including any variables with $\delta \leq fD$, where $0 < f < 1$. One approach is to look at confidence intervals for the difference in the population means. If the absolute value of the difference is greater than $fD + (1 - f)D/2$, the variable is included. If the

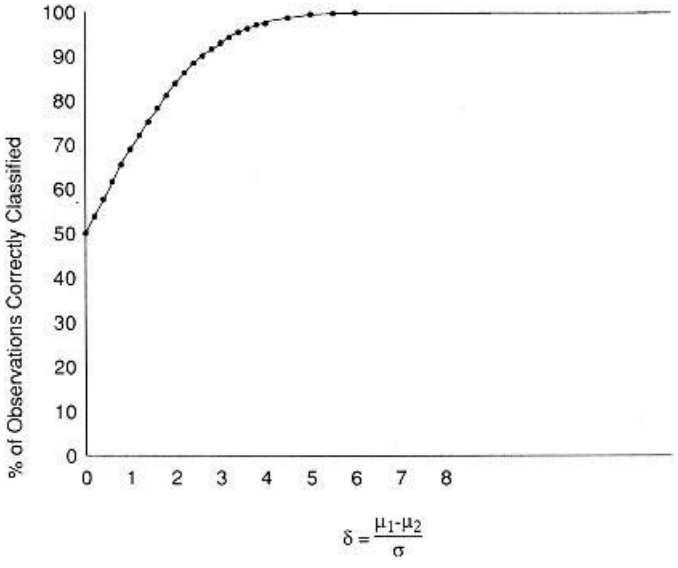

Figure 17.1 Probability of correct classification between $N(0, \sigma^2)$ and $N(\delta\sigma, \sigma^2)$ populations, assuming equal priors and $\delta\sigma/2$ as the cutoff values for classifying into the two populations.

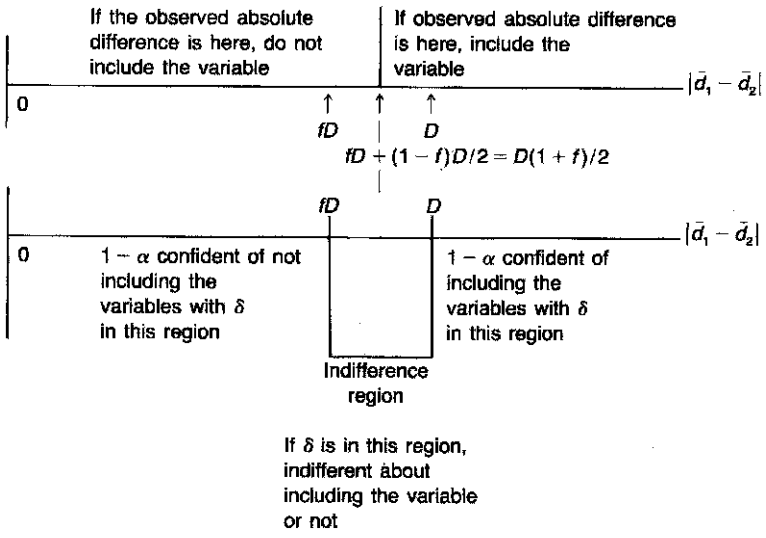

Figure 17.2 Inclusion and exclusion scheme for differences in sample means $|d_1 - d_2|$ from populations G_1 and G_2.

absolute value of the difference is less than this value, the variable is not included. Figure 17.2 presents the situation. To recap, with probability at least $1 - \alpha$, we include for use in prediction all variables with $\delta \geq D$ and do not include those with $\delta \leq fD$. In between, we are willing for either action to take place. The dividing line is placed in the middle.

Let us suppose that the number of observations, n, is large enough so that a normal approximation for confidence intervals will hold. Further, suppose that a fraction p of the data is from the first population and that $1 - p$ is from the second population. If we choose $1 - \alpha^*$ confidence intervals so that the probability is about $1 - \alpha$ that all intervals have half-width $\sigma(1 - f)D/2$, the result will hold.

If n is large, the pooled variance is approximately σ and the half-interval has width (in standard deviation units) of about

$$\sqrt{\frac{1}{Np} + \frac{1}{N(1-p)}} Z_{1-\alpha^*}$$

where $Z_{1-\alpha^*}$ is the $N(0, 1)$ critical value. To make this approximately $(1 - f)D/2$, we need

$$N = \frac{4z^2_{1-\alpha^*}}{p(1-p)D^2(1-f)^2} \tag{11}$$

In Chapter 12 it was shown that $\alpha^* = \alpha/2k$ was an appropriate choice by Bonferroni's inequality. In most practical situations, the observations tend to vary together, and the probability of all the confidence statements holding is greater than $1 - \alpha$. A slight compromise is to use $\alpha^* = [1 - (1 - \alpha)^{1/k}]/2$ as if the tests are independent. This α^* was used in computing Table 17.4.

From the table it is very clear that there is a large price to be paid if the smaller population is a very small fraction of the sample. There is often no way around this if the data need to be collected prospectively before subjects have the population membership determined (by having a heart attack or myocardial infarction, for example).

Table 17.4 Sample Sizes Needed for Univariate Screening When $f = \frac{2}{3}$[a]

	$p = 0.5$			$p = 0.6$			$p = 0.7$			$p = 0.8$			$p = 0.9$		
D	0.5	1	2	0.5	1	2	0.5	1	2	0.5	1	2	0.5	1	2
$k = 20$	2121	527	132	2210	553	136	2525	629	157	3315	829	204	5891	1471	366
	2478	616	153	2580	642	157	2950	735	183	3872	965	238	6881	1717	429
	3289	825	204	3434	859	213	3923	978	242	5151	1288	319	9159	2287	570
$k = 100$	2920	721	179	3043	761	187	3477	867	217	4565	1139	285	8118	2028	506
	3285	820	204	3421	854	213	3910	978	242	5134	1284	319	9129	2282	570
	4118	1029	255	4288	1071	268	4905	1224	306	6435	1607	400	11445	2860	714
$k = 300$	3477	867	217	3625	905	225	4140	1033	255	5436	1356	336	9665	2414	604
	3846	961	238	4008	999	247	4577	1143	285	6010	1500	374	10685	2669	667
	4684	1169	289	4879	1220	302	5576	1394	349	7323	1828	455	13018	3251	812

[a]For each entry the top, middle, and bottom numbers are for $\alpha = 0.10$, 0.05, and 0.01, respectively.

17.4.2 Sample Size to Determine That a Set of Variables Has Discriminating Power

In this section we find the answer to the following question. Assume that a discriminant analysis is being performed at significance level α with m variables. Assume that one population has a fraction p of the observations and that the other population has a fraction $1 - p$ of the observations. What sample size, n, is needed so that with probability $1 - \beta$, we reject the null hypothesis of no predictive power (i.e., Mahalanobis distance equal to zero) when in fact the Mahalanobis distance is $\Delta > 0$ (where Δ is fixed and known)? (See Chapter 13 for a definition of the Mahalanobis distance.)

The procedure is to use tables for the power functions of the analysis of variance tests as given in the CRC tables [Beyer, 1968 pp. 311–319]. To enter the charts, first find the chart for $v_1 = m$, the number of predictive variables.

The charts are for $\alpha = 0.05$ or 0.01. It is necessary to iterate to find the correct sample size n. The method is as follows:

1. Select an estimate of n.
2. Compute
$$\phi_n = \Delta\sqrt{\frac{p(1-p)}{m+1}} \times \sqrt{n} \tag{12}$$
This quantity indexes the power curves and is a measure of the difference between the two populations, adjusting for p and m.
3. Compute $v_2 = n - 2$.
4. On the horizontal axis, find ϕ and go vertically to the v_2 curve. Follow the intersection horizontally to find $1 - \widetilde{\beta}$.
5. **a.** If $1 - \widetilde{\beta}$ is greater than $1 - \beta$, decrease the estimate of n and go back to step 2.
   b. If $1 - \widetilde{\beta}$ is less than $1 - \beta$, increase the estimate of n and go back to step 2.
   c. If $1 - \widetilde{\beta}$ is approximately equal to $1 - \beta$, stop and use the given value of n as your estimate.

Example 17.4. Working at a significance level 0.05 with five predictive variables, find the total sample size needed to be 90% certain of establishing predictive power when $\cdot\Delta = 1$ and $p = 0.34$. Figure 17.3 is used in the calculation.

We use
$$\phi_n = 1 \times \sqrt{\frac{0.3 \times 0.7}{5+1}}\sqrt{n} = 0.187\sqrt{n}$$

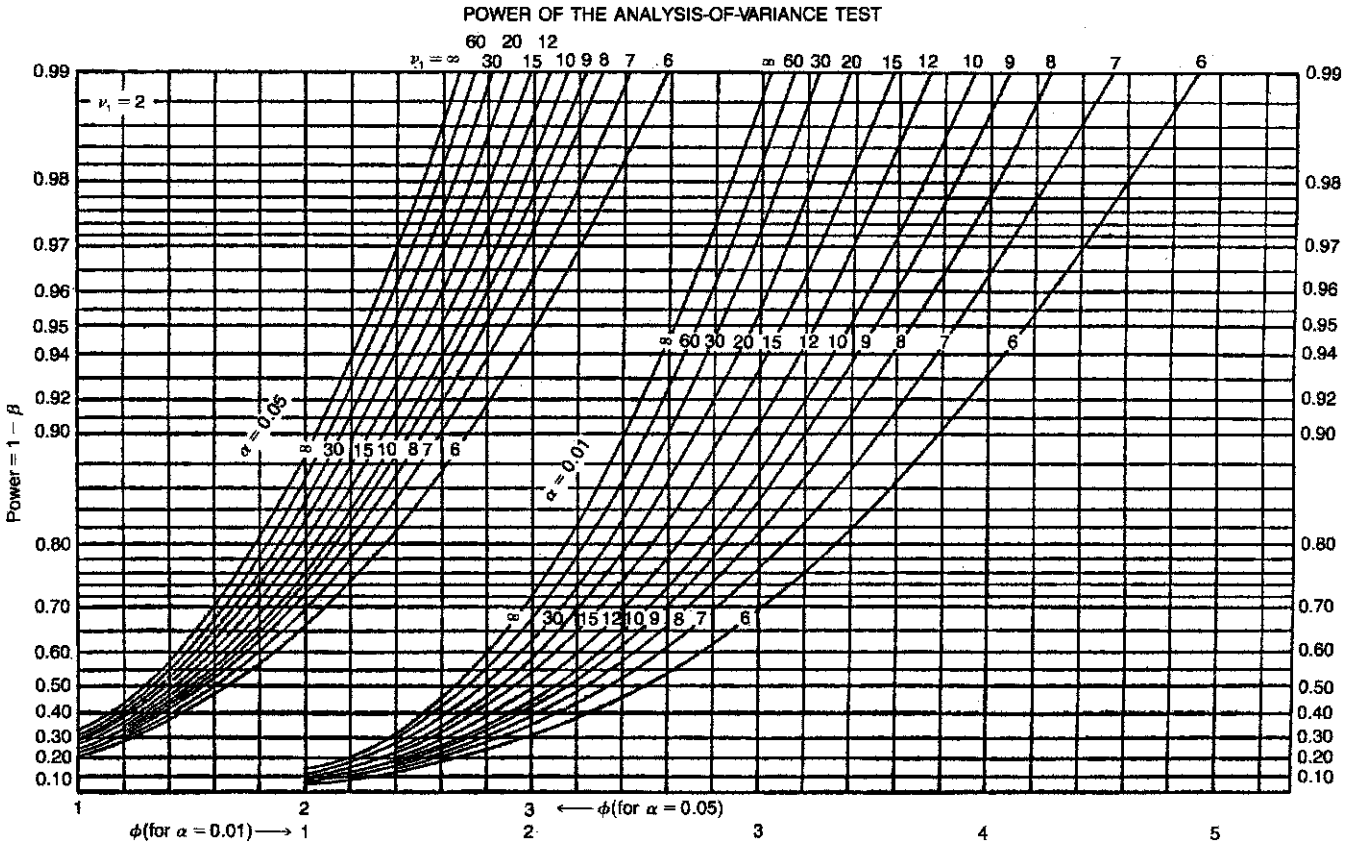

Figure 17.3 Power of the analysis of variance test. (From Beyer [1968].)

The method proceeds as follows:

1. Try $n = 30, \phi = 1.024, v_2 = 28, 1 - \beta \doteq 0.284$.
2. Try $n = 100, \phi = 1.870, v_2 = 98, 1 - \beta \doteq 0.958$.
3. Try $n = 80, \phi = 1.672, v_2 = 78, 1 - \beta \doteq 0.893$.
4. Try $n = 85, \phi = 1.724, v_2 = 83, 1 - \beta \doteq 0.92$.

Use $n = 83$. Note that the method is somewhat approximate, due to the amount of interpolation (rough visual interpretation) needed.

17.4.3 Quantifying the Precision of a Discrimination Method

After developing a method of classification, it is useful to validate the method on a new independent sample from the data used to find the classification algorithm. The approach of Section 17.4.2 is designed to show that there is some classification power. Of more interest is to be able to make a statement on the amount of correct and incorrect classification. Suppose that one is hoping to develop a classification method that classifies correctly $100\pi\%$ of the time.

To estimate with $100(1 - \alpha)\%$ confidence the correct classification percentage to within $100\varepsilon\%$, what number of additional observations are required? The confidence interval (we'll assume n large enough for the normal approximation) will be, letting c equal the number of n trials correctly classified,

$$\frac{c}{n} \pm \sqrt{\frac{1}{n}\frac{c}{n}\left(1 - \frac{c}{n}\right)} z_{1-\alpha/2}$$

where $z_{1-\alpha/2}$ is the $N(0, 1)$ critical value. We expect $c/n \doteq \pi$, so it is reasonable to choose n to satisfy $z_{1-\alpha/2} = \varepsilon\sqrt{\pi(1 - \pi)/n}$. This implies that

$$n = z^2_{1-\alpha/2}\pi(1 - \pi)/\varepsilon^2 \tag{13}$$

where $\varepsilon =$ (predicted - actual) probability of misclassification.

Example 17.5. If one plans for $\pi = 90\%$ correct classification and wishes to be 99% confident of estimating the correct classification to within 2%, how many new experimental units must be allowed? From Equation (13) and $z_{0.995} = 2.576$, the answer is

$$n = (2.576)^2 \times \frac{0.9(1 - 0.9)}{(0.02)^2} \doteq 1493$$

17.4.4 Total Sample Size for an Observational Study to Select Classification Variables

In planning an observational study to discriminate between two populations, if the predictive variables are few in number and known, the sample size will be selected in the manner of Section 17.4.2 or 17.4.3. The size depends on whether the desire is to show some predictive power or to have desired accuracy of estimation of the probability of correct classification. In addition, a different sample is needed to estimate the discriminant function. Usually, this is of approximately the same size.

If the predictive variables are to be culled from a large number of choices, an *additional* number of observations must be added for the selection of the predictive variables (e.g., in the manner of Section 17.4.1). Note that the method cannot be validated by application to the observations used to select the variables and to construct the discriminant function: This would lead to an exaggerated idea of the accuracy of the method. As the coefficients and variables were chosen specifically for these data, the method will work better (often considerably better) on these data than on an independent sample chosen as in Section 17.4.2 or 17.4.3.

NOTES

17.1 Sample Sizes for Cohort Studies

Five major journals are sources for papers dealing with sample sizes in cohort and case–control studies: *Statistics in Medicine*, *Biometrics, Controlled Clinical Trials, Journal of Clinical Epidemiology*, and the *American Journal of Epidemiology*. In addition, there are books by Fleiss [1981], Schlesselman [1982], and Schuster [1993].

A cohort study can be thought of as a cross-sectional study; there is no selection on case status or exposure status. The table generated is then the usual 2×2 table. Let the sample proportions be as follows:

	Exposure	**No Exposure**	
Case	p_{11}	p_{12}	$p_{1\cdot}$
Control	p_{21}	p_{22}	$p_{2\cdot}$
	$p_{\cdot 1}$	$p_{\cdot 2}$	1

If p_{11}, $p_{1\cdot}$, $p_{2\cdot}$, $p_{\cdot 1}$, and $p_{\cdot 2}$ estimate π_{11}, $\pi_{1\cdot}$, $\pi_{2\cdot}$, $\pi_{\cdot 1}$, and $\pi_{\cdot 2}$, respectively, then the required total sample size for significance level α, and power $1 - \beta$ is approximately

$$n = \frac{\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2 \pi_{11}\pi_{1\cdot}\pi_{2\cdot}\pi_{\cdot 1}\pi_{\cdot 2}}{(\pi_{11} - \pi_{1\cdot}\pi_{\cdot 1})^2} \tag{14}$$

Given values of $\pi_{1\cdot}$, $\pi_{\cdot 1}$, and $R = (\pi_{11}/\pi_{\cdot 1})/(\pi_{12}/\pi_{\cdot 2}) =$ the relative risk, the value of π_{11} is determined by

$$\pi_{11} = \frac{R\pi_{\cdot 1}\pi_{1\cdot}}{R\pi_{\cdot 1} + \pi_{\cdot 2}} \tag{15}$$

The formula for the required sample size then becomes

$$n = \left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2 \frac{\pi_{\cdot 1}}{1 - \pi_{\cdot 1}} \frac{1 - \pi_{\cdot 1}}{\pi_{\cdot 1}} \left[1 + \frac{1}{\pi_{\cdot 1}(R - 1)}\right]^2 \tag{16}$$

If the events are rare, the Poisson approximation derived in the text can be used. For a discussion of sample sizes in $r \times c$ contingency tables, see Lachin [1977] and Cohen [1988].

17.2 Sample-Size Formulas for Case–Control Studies

There are a variety of sample-size formulas for case–control studies. Let the data be arranged in a table as follows:

	Exposed	**Not Exposed**	
Case	X_{11}	X_{12}	n
Control	X_{21}	X_{22}	n

and

$$P[\text{exposure}|\text{case}] = \pi_1, \qquad P[\text{exposure}|\text{control}] = \pi_2$$

estimated by $P_1 = X_{11}/n$ and $P_2 = X_{21}/n$ (we assume that $n_1 = n_2 = n$). For a two-sample, two-tailed test with

$$P[\text{Type I error}] = \alpha \quad \text{and} \quad P[\text{Type II error}] = \beta$$

the approximate sample size per group is

$$n = \frac{[Z_{1-\alpha/2}\sqrt{2\overline{\pi}(1-\overline{\pi})} + Z_{1-\beta}\sqrt{\pi_1(1-\pi_1)+\pi_2(1-\pi_2)}]^2}{(\pi_1-\pi_2)^2} \tag{17}$$

where $\overline{\pi} = \frac{1}{2}(\pi_1 + \pi_2)$. The total number of subjects is $2n$, of which n are cases and n are controls. Another formula is

$$n = \frac{[\pi_1(1-\pi) + \pi_2(1-\pi_2)](Z_{1-\alpha/2} + Z_{1-\beta})^2}{(\pi_1-\pi_2)^2} \tag{18}$$

All of these formulas tend to give the same answers, and underestimate the sample sizes required. The choice of formula is primarily a matter of aesthetics.

The formulas for sample sizes for case–control studies are approximations, and several corrections are available to get closer to the exact value. Exact values for equal sample sizes have been tabulated in Haseman [1978]. Adjustment for the approximate sample size have been presented by Casagrande et al. [1978], who give a slightly more complicated and accurate formulation. See also Lachin [1981, 2000] and Ury and Fleiss [1980].

Two other considerations will be mentioned. The first is unequal sample size. Particularly in case–control studies, it may be difficult to recruit more cases. Suppose that we can select n observations from the first population and rn from the second ($0 < r < \infty$). Following Schlesselman [1982], a very good approximation for the exact sample size for the number of cases is

$$n_1 = n\left(\frac{r+1}{2r}\right) \tag{19}$$

and for the number of controls

$$n_2 = n\left(\frac{r+1}{2}\right) \tag{20}$$

where n is determined by equation (17) or (18). The total sample size is then $n((r+1)^2/2r)$. Note that the number of cases can never be reduced to more than $n/2$ no matter what the number of controls. This is closely related to the discussion in Section 17.3. Following Fleiss et al. [1980], a slightly improved estimate can be obtained by using

$$n_1^* = n_1 + \frac{r+1}{r\Delta} = \text{number of cases}$$

and

$$n_2^* = rn_1^* = \text{number of controls}$$

A second consideration is cost. In Section 17.3 we considered sample sizes as a function of cost and related the sample sizes to precision. Now consider a slight reformulation of the problem in the case–control context. Suppose that enrollment of a case costs c_1 and enrollment of a control costs c_2. Pike and Casagrande [1979] show that a reasonable sample size approximation is

$$n_1 = n\left(1 + \sqrt{\frac{c_1}{c_0}}\right)$$

$$n_2 = n\left(1 + \sqrt{\frac{c_0}{c_1}}\right)$$

where n is defined by equations (17) or (18).

Finally, frequently case–control study questions are put in terms of odds ratios (or relative risks). Let the odds ratio be $R = \pi_1(1-\pi_2)/\pi_2(1-\pi_1)$, where π_1 and π_2 are as defined at the beginning of this section. If the control group has known exposure rate π_2, that is, $P[\text{exposure}|\text{control}] = \pi_2$, then

$$\pi_1 = \frac{R\pi_2}{1+\pi_2(R-1)}$$

To calculate sample sizes, use equation (17) for specified values of π_2 and R.

Mantel [1983] gives some clever suggestions for making binomial sample-size tables more useful by making use of the fact that sample size is "inversely proportional to the square of the difference being sought, everything else being more or less fixed."

Newman [2001] is a good reference for sample-size questions involving survival data.

17.3 Power as a Function of Sample Size

Frequently, the question is not "How big should my sample size be" but rather, "I have 60 observations available; what kind of power do I have to detect a specified difference, relative risk, or odds ratio?" The charts by Feigl illustrated in Chapter 6 provided one answer. Basically, the question involves inversion of formulas such as given by equations (17) and (18), solving them for $Z_{1-\beta}$, and calculating the associated area under the normal curve. Besides Feigl, several authors have studied this problem or variations of it. Walter [1977] derived formulas for the smallest and largest relative risk, R, that can be detected as a function of sample size, Type I and Type II errors. Brittain and Schlesselman [1982] present estimates of power as a function of possibly unequal sample size and cost.

17.4 Sample Size as a Function of Coefficient of Variation

Sometimes, sample-size questions are asked in the context of percent variability and percent changes in means. With an appropriate, natural interpretation, valid answers can be provided. Specifically, assume that by *percent variability* is meant the coefficient of variation, call it V, and that the second mean differs from the first mean by a factor f.

Let two normal populations have means μ_1 and μ_2 and standard deviations σ_1 and σ_2. The usual sample-size formula for two independent samples needed to detect a difference $\mu_1 - \mu_2$ in means with Type I error α and power $1-\beta$ is given by

$$n = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2(\sigma_1^2 + \sigma_2^2)}{(\mu_1 - \mu_2)^2}$$

where $z_{1-\gamma}$ is the $100(1-\gamma)$th percentile of the standard normal distribution. This is the formula for a two-sided alternative; n is the number of observations per group. Now assume that $\mu_1 = f\mu_2$ and $\sigma_1/\mu_1 = \sigma_2/\mu_2 = V$. Then the formula transforms to

$$n = (z_{1-\alpha/2} + z_{1-\beta})^2 V^2 \left[1 + \frac{2f}{(f-1)^2}\right] \tag{21}$$

The quantity V is the usual coefficient of variation and f is the ratio of means. It does not matter whether the ratio of means is defined in terms of $1/f$ rather than f.

Sometimes the problem is formulated with the variability V as specified but a percentage change between means is given. If this is interpreted as the second mean, μ_2, being a percent change from the first mean, this percentage change is simply $100(f-1)\%$ and the formula again applies. However, sometimes, the relative status of the means cannot be specified, so an

interpretation of *percent change* is needed. If we know only that $\sigma_1 = V\mu_1$ and $\sigma_2 = V\mu_2$, the formula for sample size becomes

$$n = \frac{V^2(z_{1-\alpha/2} + z_{1-\beta})^2}{\left((\mu_1 - \mu_2)/\sqrt{\mu_1\mu_2}\right)^2}$$

The quantity $\left((\mu_1 - \mu_2)/\sqrt{\mu_1\mu_2}\right)$ is the proportional change from μ_1 to μ_2 as a function of their geometric mean. If the questioner, therefore, can only specify a percent change, this interpretation is quite reasonable. Solving equation (21) for $z_{1-\beta}$ allows us to calculate values for power curves:

$$z_{1-\beta} = -z_{1-\alpha/2} + \frac{\sqrt{n}|f-1|}{V\sqrt{f^2+1}} \tag{22}$$

A useful set of curves as a function of n and a common coefficient of variation $V = 1$ can be constructed by noting that for two coefficients of variation V_1 and V_2, the sample sizes $n(V_1)$ and $n(V_2)$, as functions of V_1 and V_2, are related by

$$\frac{n(V_1)}{n(V_2)} = \frac{\sigma_1^2}{\sigma_2^2}$$

for the same power and Type I error. See van Belle and Martin [1993] and van Belle [2001].

PROBLEMS

17.1 **(a)** Verify that the odds ratio and relative risk are virtually equivalent for

$$P[\text{exposure}] = 0.10, \qquad P[\text{disease}] = 0.01$$

in the following two situations:

$$\pi_{11} = P[\text{exposed and disease}] = 0.005$$

and $\pi_{11} = 0.0025$.

(b) Using equation (2), calculate the number of disease occurrences in the exposed and unexposed groups that would have to be observed to detect the relative risks calculated above with $\alpha = 0.05$ (one-tailed) and $\beta = 0.10$.

(c) How many exposed persons would have to be observed (and hence, unexposed persons as well)?

(d) Calculate the sample size needed if this test is one of K tests for $K = 10, 100,$ and 1000.

(e) In part (d), plot the logarithm of the sample size as a function of $\log K$. What kind of relationship is suggested? Can you state a general rule?

17.2 (After N. E. Breslow) Workers at all nuclear reactor facilities will be observed for a period of 10 years to determine whether they are at excess risk for leukemia. The rate in the general population is 7.5 cases per 100,000 person-years of observation. We want to be 80% sure that a doubled risk will be detected at the 0.05 level of significance.

(a) Calculate the number of leukemia cases that must be detected among the nuclear plant workers.

(b) How many workers must be observed? That is, assuming the null hypothesis holds, how many workers must be observed to accrue 9.1 leukemia cases?

(c) Consider this as a binomial sampling problem. Let $\pi_1 = 9.1/$answer in part (b), and let $\pi_2 = 2\pi_1$. Now use equation (17) to calculate $n/2$ as the required sample size. How close is your answer to part (b)?

17.3 (After N. E. Breslow) The rate of lung cancer for men of working age in a certain population is known to be on the order of 60 cases per 100,000 person-years of observation. A cohort study using equal numbers of exposed and unexposed persons is desired so that an increased risk of $R = 1.5$ can be detected with power $1 - \beta = 0.95$ and $\alpha = 0.01$.

(a) How many cases will have to be observed in the unexposed population? The exposed population?

(b) How many person-years of observation at the normal rates will be required for either of the two groups?

(c) How many workers will be needed assuming a 20-year follow-up?

17.4 (After N. E. Breslow) A case–control study is to be designed to detect an odds ratio of 3 for bladder cancer associated with a certain medication that is used by about one person out of 50 in the general population.

(a) For $\alpha = 0.05$, and $\beta = 0.05$, calculate the number of cases and number of controls needed to detect the increased odds ratio.

(b) Use the Poisson approximation procedure to calculate the sample sizes required.

(c) Four controls can be provided for each case. Use equations (19) and (20) to calculate the sample sizes. Compare this result with the total sample size in part (a).

17.5 The sudden infant death syndrome (SIDS) occurs at a rate of approximately three cases per 1000 live births. It is thought that smoking is a risk factor for SIDS, and a case–control study is initiated to check this assumption. Since the major effort was in the selection and recruitment of cases and controls, a questionnaire was developed that contained 99 additional questions.

(a) Calculate the sample size needed for a case–control study using $\alpha = 0.05$, in which we want to be 95% certain of picking up an increased relative risk of 2 associated with smoking. Assume that an equal number of cases and controls are selected.

(b) Considering smoking just one of the 100 risk factors considered, what sample sizes will be needed to maintain an $\alpha = 0.05$ per experiment error rate?

(c) Given the increased value of Z in part (b), suppose that the sample size is not changed. What is the effect on the power? What is the power now?

(d) Suppose in part (c) that the power also remains fixed at 0.95. What is the minimum relative risk that can be detected?

(e) Since smoking was the risk factor that precipitated the study, can an argument be made for not testing it at a reduced α level? Formulate your answer carefully.

***17.6** Derive the square root rule starting with equations (4) and (5).

***17.7** Derive formula (16) from equation (14).

17.8 It has been shown that coronary bypass surgery does not prolong life in selected patients with relatively mild angina (but may relieve the pain). A surgeon has invented a new

bypass procedure that, she claims, will prolong life substantially. A trial is planned with patients randomized to surgical treatment or standard medical therapy. Currently, the five-year survival probability of patients with relatively mild symptoms is 80%. The surgeon claims that the new technique will increase survival to 90%.

(a) Calculate the sample size needed to be 95% certain that this difference will be detected using an $\alpha = 0.05$ significance level.

(b) Suppose that the cost of a coronary bypass operation is approximately \$50,000; the cost of general medical care is about \$10,000. What is the most economical experiment under the conditions specified in part (a)? What are the total costs of the two studies?

(c) The picture is more complicated than described in part (b). Suppose that about 25% of the patients receiving the medical treatment will go on to have a coronary bypass operation in the next five years. Recalculate the sample sizes under the conditions specified in part (a).

***17.9** Derive the sample sizes in Table 17.4 for $D = 0.5$, $p = 0.8$, $\alpha = 0.5$, and $k = 20, 100, 300$.

***17.10** Consider the situation in Example 17.4.

(a) Calculate the sample size as a function of m, the number of variables, by considering $m = 10$ and $m = 20$.

(b) What is the relationship of sample size to variables?

17.11 Two groups of rats, one young and the other old, are to be compared with respect to levels of nerve growth factor (NGF) in the cerebrospinal fluid. It is estimated that the variability in NGF from animal to animal is on the order of 60%. We want to look at a twofold ratio in means between the two groups.

(a) Using the formula in Note 17.4, calculate the sample size per group using a two-sided alternative, $\alpha = 0.05$, and a power of 0.80.

(b) Suppose that the ratio of the means is really 1.6. What is the power of detecting this difference with the sample sizes calculated in part (a)?

REFERENCES

Beyer, W. H. (ed.) [1968]. *CRC Handbook of Tables for Probability and Statistics*, 2nd ed. CRC Press, Cleveland, OH.

Brittain, E., and Schlesselman, J. J. [1982]. Optimal allocation for the comparison of proportions. *Biometrics*, **38**: 1003–1009.

Casagrande, J. T., Pike, M. C., and Smith, P. C. [1978]. An improved approximate formula for calculating sample sizes for comparing two binomial distributions. *Biometrics*, **34**: 483–486.

Cochran, W. G. [1977]. *Sampling Techniques*, 3rd ed. Wiley, New York.

Cohen, J. [1988]. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Lawrence Erlbaum Associates, Hillsdale, NJ.

Fleiss, J. L., Levin, B., and Park, M. C. [2003]. *Statistical Methods for Rates and Proportions*, 3rd ed. Wiley, New York.

Fleiss, J. L., Tytun, A., and Ury, H. K. [1980]. A simple approximation for calculating sample sizes for comparing independent proportions. *Biometrics*, **36**: 343–346.

Gail, M., Williams, R., Byar, D. P., and Brown, C. [1976]. How many controls. *Journal of Chronic Diseases*, **29**: 723–731.

Haseman, J. K. [1978]. Exact sample sizes for the use with the Fisher–Irwin test for 2×2 tables. *Biometrics*, **34**: 106–109.

Lachin, J. M. [1977]. Sample size determinations for $r \times c$ comparative trials. *Biometrics*, **33**: 315–324.

Lachin, J. M. [1981]. Introduction to sample size determination and power analysis for clinical trials. *Controlled Clinical Trials*, **2**: 93–113.

Lachin, J. M. [2000]. *Biostatistical Methods*. Wiley, New York.

Lubin, J. H. [1980]. Some efficiency comments on group size in study design. *American Journal of Epidemiology*, **111**: 453–457.

Mantel, H. [1983]. Extended use of binomial sample-size tables. *Biometrics*, **39**: 777–779.

Nam, J. M. [1973]. Optimum sample sizes for the comparison of a control and treatment. *Biometrics*, **29**: 101–108.

Newman, S. C. [2001]. *Biostatistical Methods in Epidemiology*. Wiley, New York.

Pike, M. C., and Casagrande, J. T. [1979]. Cost considerations and sample size requirements in cohort and case-control studies. *American Journal of Epidemiology*, **110**: 100–102.

Schlesselman, J. J. [1982]. *Case–Control Studies: Design, Conduct, Analysis*. Oxford University Press, New York.

Schuster, J. J. [1993]. *Practical Handbook of Sample Size Guidelines for Clinical Trials*. CRC Press, Boca Raton, FL.

Ury, H. K., and Fleiss, J. R. [1980]. On approximate sample sizes for comparing two independent proportions with the use of Yates' correction. *Biometrics*, **36**: 347–351.

van Belle, G. [2001]. *Statistical Rules of Thumb*. Wiley, New York.

van Belle, G., and Martin, D. C. [1993]. Sample size as a function of coefficient of variation and ratio of means. *American Statistician*, **47**: 165–167.

Walter, S. D. [1977]. Determination of significant relative risks and optimal sampling procedures in prospective and retrospective comparative studies of various sizes. *American Journal of Epidemiology*, **105**: 387–397.

CHAPTER 18

Longitudinal Data Analysis

18.1 INTRODUCTION

One of the most common medical research designs is a "pre–post" study in which a single baseline health status measurement is obtained, an intervention is administered, and a single follow-up measurement is collected. In this experimental design, the *change* in the outcome measurement can be associated with the *change* in the exposure condition. For example, if some subjects are given placebo while others are given an active drug, the two groups can be compared to see if the change in the outcome is different for those subjects who are actively treated as compared to control subjects. This design can be viewed as the simplest form of a prospective longitudinal study.

Definition 18.1. A *longitudinal study* refers to an investigation where participant outcomes and possibly treatments or exposures are collected at multiple follow-up times.

A longitudinal study generally yields multiple or "repeated" measurements on each subject. For example, HIV patients may be followed over time and monthly measures such as CD4 counts or viral load are collected to characterize immune status and disease burden, respectively. Such repeated-measures data are correlated within subjects and thus require special statistical techniques for valid analysis and inference.

A second important outcome that is commonly measured in a longitudinal study is the time until a key clinical event such as disease recurrence or death. Analysis of event-time endpoints is the focus of *survival analysis*, which is covered in Chapter 16.

Longitudinal studies play a key role in epidemiology, clinical research, and therapeutic evaluation. Longitudinal studies are used to characterize normal growth and aging, to assess the effect of risk factors on human health, and to evaluate the effectiveness of treatments.

Longitudinal studies involve a great deal of effort but offer several benefits, which include:

1. *Incident events recorded.* A prospective longitudinal study measures the new occurrence of disease. The timing of disease onset can be correlated with recent changes in patient exposure and/or with chronic exposure.

2. *Prospective ascertainment of exposure*. In a prospective study, participants can have their exposure status recorded at multiple follow-up visits. This can alleviate recall bias where subjects who subsequently experience disease are more likely to recall their exposure (a form of measurement error). In addition, the temporal order of exposures and outcomes is observed.

Biostatistics: A Methodology for the Health Sciences, Second Edition, by Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley
ISBN 0-471-03185-2

3. *Measurement of individual change in outcomes*. A key strength of a longitudinal study is the ability to measure change in outcomes and/or exposure at the individual level. Longitudinal studies provide the opportunity to observe individual patterns of change.

4. *Separation of time effects: cohort, period, age*. When studying change over time, there are many time scales to consider. The *cohort scale* is the time of birth, such as 1945 or 1963; *period* is the current time, such as 2003; and *age* is (period − cohort), for example, $58 = 2003 - 1945$, and $40 = 2003 - 1963$. A longitudinal study with measurements at times $t_1, t_2, \ldots, t_n$ can simultaneously characterize multiple time scales such as age and cohort effects using covariates derived from the calendar time of visit and the participant's birth year: the age of subject i at time t_j is $\text{age}_{ij} = t_j - \text{birth}_i$; and their cohort is simply $\text{cohort}_{ij} = \text{birth}_i$. Lebowitz [1996] discusses age, period, and cohort effects in the analysis of pulmonary function data.

5. *Control for cohort effects*. In a cross-sectional study the comparison of subgroups of different ages combines the effects of aging and the effects of different cohorts. That is, comparison of outcomes measured in 2003 among 58-year-old subjects and among 40-year-old subjects reflects both the fact that the groups differ by 18 years (aging) and the fact that the subjects were born in different eras. For example, the public health interventions, such as vaccinations available for a child under 10 years of age, may differ in 1945–1955 compared to the preventive interventions experienced in 1963–1973. In a longitudinal study, the cohort under study is fixed, and thus changes in time are not confounded by cohort differences.

An overview of longitudinal data analysis opportunities in respiratory epidemiology is presented in Weiss and Ware [1996].

The benefits of a longitudinal design are not without cost. There are several challenges posed:

1. *Participant follow-up*. There is the risk of bias due to incomplete follow-up, or dropout of study participants. If subjects who are followed to the planned end of a study differ from subjects who discontinue follow-up, a naive analysis may provide summaries that are not representative of the original target population.

2. *Analysis of correlated data*. Statistical analysis of longitudinal data requires methods that can properly account for the intrasubject correlation of response measurements. If such correlation is ignored, inferences such as statistical tests or confidence intervals can be grossly invalid.

3. *Time-varying covariates*. Although longitudinal designs offer the opportunity to associate changes in exposure with changes in the outcome of interest, the direction of causality can be complicated by "feedback" between the outcome and the exposure. For example, in an observational study of the effects of a drug on specific indicators of health, a patient's current health status may influence the drug exposure or dosage received in the future. Although scientific interest lies in the effect of medication on health, this example has reciprocal influence between exposure and outcome and poses analytical difficulty when trying to separate the effect of medication on health from the effect of health on drug exposure.

18.1.1 Example studies

In this section we give some examples of longitudinal studies and focus on the primary scientific motivation in addition to key outcome and covariate measurements.

Child Asthma Management Program

In the Child Asthma Management Program (CAMP) study, children are randomized to different asthma management regimes. CAMP is a multicenter clinical trial whose primary aim is evaluation of the long-term effects of daily inhaled anti-inflammatory medication use on asthma status and lung growth in children with mild to moderate asthma [The Childhood Asthma Management

Program Research group, 2000]. Outcomes include continuous measures of pulmonary function and categorical indicators of asthma symptoms. Secondary analyses have investigated the association between daily measures of ambient pollution and the prevalence of symptoms. Analysis of an environmental exposure requires specification of a lag between the day of exposure and the resulting effect. In the air pollution literature, short lags of 0 to 2 days are commonly used [Samet et al., 2000; Yu et al., 2000]. For both the evaluation of treatment and exposure to environmental pollution, the scientific questions focus on the association between an exposure (treatment, pollution) and health measures. The within-subject correlation of outcomes is of secondary interest, but must be acknowledged to obtain valid statistical inference.

Cystic Fibrosis Foundation Registry

The Cystic Fibrosis Foundation maintains a registry of longitudinal data for subjects with cystic fibrosis. Pulmonary function measures, such as the 1-second forced expiratory volume (FEV1), and patient health indicators, such as infection with *Pseudomonas aeruginosa*, have been recorded annually since 1966. One scientific objective is to characterize the natural course of the disease and to estimate the average rate of decline in pulmonary function. Risk factor analysis seeks to determine whether measured patient characteristics such as gender and genotype correlate with disease progression or with an increased rate of decline in FEV1. The registry data represent a typical observational design where the longitudinal nature of the data are important for determining individual patterns of change in health outcomes such as lung function.

Multicenter AIDS Cohort Study

The Multicenter AIDS Cohort Study (MACS) enrolled more than 3000 men who were at risk for acquisition of HIV1 [Kaslow et al., 1987]. This prospective cohort study observed $N = 479$ incident HIV1 infections and has been used to characterize the biological changes associated with disease onset. In particular, this study has demonstrated the effect of HIV1 infection on indicators of immunologic function such as CD4 cell counts. One scientific question is whether baseline characteristics such as viral load measured immediately after seroconversion are associated with a poor patient prognosis as indicated by a greater rate of decline in CD4 cell counts. We use these data to illustrate analysis approaches for continuous longitudinal response data.

HIVNET Informed Consent Substudy

Numerous reports suggest that the process of obtaining informed consent in order to participate in research studies is often inadequate. Therefore, for preventive HIV vaccine trials a prototype informed consent process was evaluated among $N = 4892$ subjects participating in the Vaccine Preparedness Study (VPS). Approximately 20% of subjects were selected at random and asked to participate in a mock informed consent process [Coletti et al., 2003]. Participant knowledge of key vaccine trial concepts was evaluated at baseline prior to the informed consent visit, which occurred during a special three-month follow-up visit for the intervention subjects. Vaccine trial knowledge was then assessed for all participants at the scheduled six-, 12-, and 18-month visits. This study design is a basic longitudinal extension of a pre–post design. The primary outcomes include individual knowledge items and a total score that calculates the number of correct responses minus the number of incorrect responses. We use data on a subset of men and women VPS participants. We focus on subjects who were considered at high risk of HIV acquisition, due to injection drug use.

18.1.2 Notation

In this chapter we use Y_{ij} to denote the outcome measured on subject i at time t_{ij}. The index $i = 1, 2, \ldots, N$ is for subject, and the index $j = 1, 2, \ldots, n$ is for observations within a subject. In a designed longitudinal study the measurement times will follow a protocol with

a common set of follow-up times, $t_{ij} = t_j$. For example, in the HIVNET Informed Consent Study, subjects were measured at baseline, $t_1 = 0$, at six months after enrollment, $t_2 =$ six months, and at 12 and 18 months, $t_3 = 12$ months, $t_4 = 18$ months. We let X_{ij} denote covariates associated with observation Y_{ij}. Common covariates in a longitudinal study include the time, t_{ij}, and person-level characteristics such as treatment assignment or demographic characteristics.

Although scientific interest often focuses on the mean response as a function of covariates such as treatment and time, proper statistical inference must account for the within-person correlation of observations. Define $\rho_{jk} = \text{corr}(Y_{ij}, Y_{ik})$, the within-subject correlation between observations at times t_j and t_k. In the following section we discuss methods for exploring the structure of within-subject correlation, and in Section 18.5 we discuss estimation methods that model correlation patterns.

18.2 EXPLORATORY DATA ANALYSIS

Exploratory analysis of longitudinal data seeks to discover patterns of systematic variation across groups of patients, as well as aspects of random variation that distinguish individual patients.

18.2.1 Group Means over Time

When scientific interest is in the average response over time, summary statistics such as means and standard deviations can reveal whether different groups are changing in a similar or different fashion.

Example 18.1. Figure 18.1 shows the mean knowledge score for the informed consent subgroups in the HIVNET Informed Consent Substudy. At baseline the intervention and control groups have very similar mean scores. This is expected since the group assignment is determined by randomization that occurs after enrollment. At an interim three-month visit the intervention subjects are given a mock informed consent for participation in a hypothetical phase III vaccine efficacy trial. The impact of the intervention can be seen by the mean scores at the six-month visit. In the control group the mean at six months is 1.49(SE = 0.11), up slightly from the baseline mean of 1.16(SE = 0.11). In contrast, the intervention group has a six-month mean score of 3.43(SE = 0.24), a large increase from the baseline mean of 1.09(SE = 0.24). The intervention and control groups are significantly different at six months based on a two-sample t-test. At later follow-up times, further change is observed. The control group has a mean that increases to 1.98 at the 12-month visit and to 2.47 at the 18-month visit. The intervention group fluctuates slightly with means of 3.25(SE = 0.27) at month 12 and 3.76(SE = 0.25) at 18 months. These summaries suggest that the intervention has a significant effect on knowledge, and that a small improvement is seen over time in the control group.

Example 18.2. In the MACS study we compare different groups of subjects formed on the basis of their initial viral load measurement. Low viral load is defined by a baseline value less than 15×10^3, medium as 15×10^3 to 46×10^3, and high viral load is classified for subjects with a baseline measurement greater than 46×10^3. Table 18.1 gives the average CD4 count for each year of follow-up. The mean CD4 declines over time for each of the viral load groups. The subjects with the lowest baseline viral load have a mean of 744.8 for the first year after seroconversion and then decline to a mean count of 604.8 during the fourth year. The $744.8 - 604.8 = 140.0$-unit reduction is smaller than the decline observed for the medium-viral-load group, $638.9 - 470.0 = 168.9$, and the high-viral-load group, $600.3 - 353.9 = 246.4$. Therefore, these summaries suggest that higher baseline viral-load measurements are associated with greater subsequent reduction in mean CD4 counts.

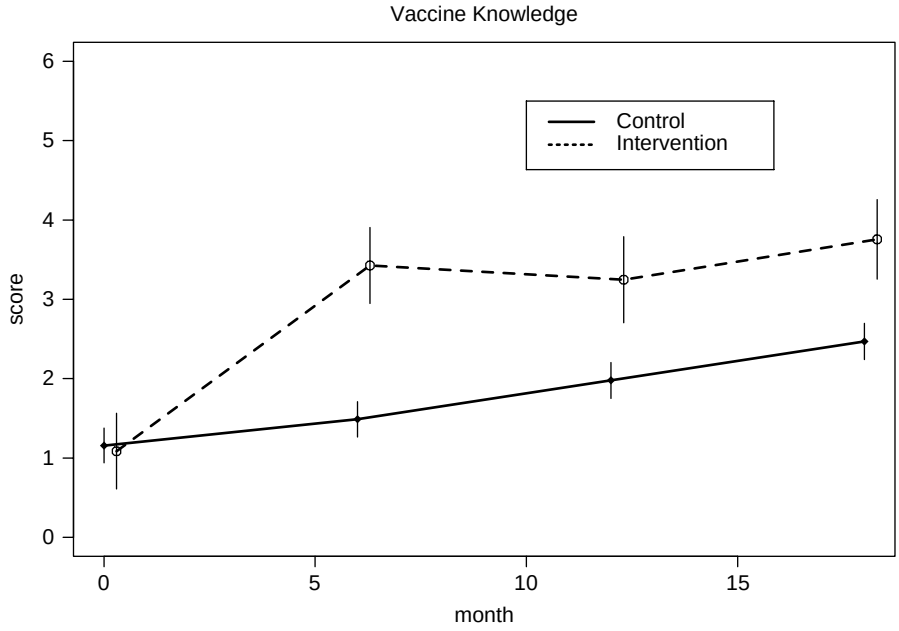

Figure 18.1 Mean knowledge scores over time by treatment group, HIVNET informed consent substudy.

Table 18.1 Mean CD4 Count and Standard Error over Time[a]

	Baseline Viral Load					
	Low		Medium		High	
Year	Mean	SE	Mean	SE	Mean	SE
0–1	744.8	35.8	638.9	27.3	600.3	30.4
1–2	721.2	36.4	588.1	25.7	511.8	22.5
2–3	645.5	37.7	512.8	28.5	474.6	34.2
3–4	604.8	46.8	470.0	28.7	353.9	28.1

[a]Separate summaries are given for groups defined by baseline viral load level.

Example 18.1. (*continued*) In the HIVNET informed consent substudy we saw a substantial improvement in the knowledge score. It is also relevant to consider key individual items that comprise the total score, such as the "safety item" or "nurse item." Regarding safety, participants were asked whether it was true or false that "Once a large-scale HIV vaccine study begins, we can be sure the vaccine is completely safe." Table 18.2 shows the number of responding subjects at each visit and the percent of subjects who correctly answered that the safety statement is false. These data show that the control and intervention groups have a comparable understanding of the safety item at baseline with 40.9% answering correctly among controls, and 39.2% answering correctly among the intervention subjects. A mock informed consent was administered at a three-month visit for the intervention subjects only. The impact of the intervention appears modest, with only 50.3% of intervention subjects correctly responding at six months. This represents a 10.9% increase in the proportion answering correctly, but a two-sample comparison of intervention and control proportions at six months (e.g., 50.3% vs. 42.7%) is not significant

Table 18.2 Number of Subjects and Percent Answering Correctly for the Safety Item from the HIVNET Informed Consent Substudy

	Control Group		Intervention Group	
Visit	*N*	% Correct	*N*	% Correct
Baseline	946	40.9	176	39.2
six-month	838	42.7	171	50.3
12-month	809	41.5	163	43.6
18-month	782	43.5	153	43.1

Table 18.3 Number of Subjects and Percent Answering Correctly for the Nurse Item from the HIVNET Informed Consent Substudy

	Control Group		Intervention Group	
Visit	*n*	% Correct	*n*	% Correct
Baseline	945	54.1	176	50.3
six-month	838	44.7	171	72.1
12-month	808	46.3	163	60.1
18-month	782	48.2	153	66.0

statistically. Finally, the modest intervention impact does not appear to be retained, as the fraction correctly answering this item declines to 43.6% at 12 months and 43.1% at 18 months. Therefore, these data suggest a small but fleeting improvement in participant understanding that a vaccine studied in a phase III trial cannot be guaranteed to be safe.

Other items show different longitudinal trends. Subjects were also asked whether it was true or false that "The study nurse will decide who gets the real vaccine and who gets the placebo." Table 18.3 shows that the groups are again comparable at baseline, but for the nurse item we see a large increase in the fraction answering correctly among intervention subjects at six months with 72.1% answering correctly that the statement is false. A cross-sectional analysis indicates a statistically significant difference in the proportion answering correctly at six months with a confidence interval for the difference in proportions of $(0.199, 0.349)$. Although the magnitude of the separation between groups decreases from 27.4% at six months to 17.8% at 18 months, the confidence interval for the difference in proportions at 18 months is $(0.096, 0.260)$ and excludes the null comparison, $p_1 - p_0 = 0$. Therefore, these data suggest that the intervention has a substantial and lasting impact on understanding that research nurses do not determine allocation to real vaccine or placebo.

18.2.2 Variation among Subjects

With independent observations we can summarize the uncertainty or variability in a response measurement using a single variance parameter. One interpretation of the variance is given as one-half the expected squared distance between any two randomly selected measurements, $\sigma^2 = \frac{1}{2}E[(Y_i - Y_j)^2]$. However, with longitudinal data the "distance" between measurements on different subjects is usually expected to be greater than the distance between repeated measurements taken on the same subject. Thus, although the total variance may be obtained with outcomes from subjects i and i' observed at time t_j, $\sigma^2 = \frac{1}{2}E[(Y_{ij} - Y_{i'j})^2]$ [assuming that $E(Y_{ij}) = E(Y_{i'j}) = \mu$], the expected variation for two measurements taken on the same person

(subject i) but at times t_j and t_k may not equal the total variation σ^2 since the measurements are correlated: $\sigma^2(1-\rho_{jk}) = \frac{1}{2}E[(Y_{ij}-Y_{ik})^2]$ [assuming that $E(Y_{ij}) = E(Y_{ik}) = \mu$]. When $\rho_{jk} > 0$, this shows that *between-subject variation* is greater than *within-subject variation*. In the extreme, $\rho_{jk} = 1$ and $Y_{ij} = Y_{ik}$, implying no variation for repeated observations taken on the same subject.

Graphical methods can be used to explore the magnitude of person-to-person variability in outcomes over time. One approach is to create a panel of individual line plots for each study participant. These plots can then be inspected for both the amount of variation from subject to subject in the overall "level" of the response and the magnitude of variation in the "trend" over time in the response. Such exploratory data analysis can be useful for determining the types of correlated data regression models that would be appropriate. In Section 18.5 we discuss random effects regression models for longitudinal data. In addition to plotting individual series, it is also useful to plot multiple series on a single plot, stratifying on the value of key covariates. Such a plot allows determination of whether the type and magnitude of intersubject variation appears to differ across the covariate subgroups.

Example 18.2. (*continued*) In Figure 18.2 we plot an array of individual series from the MACS data. In each panel the observed CD4 count for a single subject is plotted against the times that measurements were obtained. Such plots allow inspection of the individual response patterns and whether there is strong heterogeneity in the trajectories. Figure 18.2 shows that there can be large variation in the "level" of CD4 for subjects. Subject ID = 1120 in the upper right corner has CD4 counts greater than 1000 for all times, while ID = 1235 in the lower left corner has all measurements below 500. In addition, individuals plots can be evaluated for the change over time. Figure 18.2 indicates that most subjects are either relatively stable in their measurements over time, or tend to be decreasing.

In the common situation where we are interested in correlating the outcome to measured factors such as treatment group or exposure, it will also be useful to plot individual series stratified by covariate group. Figure 18.3 takes a sample of the MACS data and plots lines for each subject stratified by the level of baseline viral load. This figure suggests that the highest viral load group has the lowest mean CD4 count and suggests that variation among measurements may also be lower for the high baseline viral-load group compared to the medium- and low-viral-load groups. Figure 18.3 can also be used to identify those who exhibit time trends that differ markedly from the profiles of others. In the high-viral-load group there is a person who appears to improve dramatically over time, and there is a single unusual measurement where the CD4 count exceeds 2000. Plotting individual series is a useful exploratory prelude to more careful confirmatory statistical analysis.

18.2.3 Characterizing Correlation and Covariance

With correlated outcomes it is useful to understand the strength of correlation and the pattern of correlations across time. Characterizing correlation is useful for understanding components of variation and for identifying a variance or correlation model for regression methods such as mixed-effects models or *generalized estimating equations* (GEEs), discussed in Section 18.5.2. One summary that is used is an estimate of the *covariance matrix*, which is defined as

$$\begin{bmatrix} E[(Y_{i1}-\mu_{i1})^2] & E[(Y_{i1}-\mu_{i1})(Y_{i2}-\mu_{i2})] & \cdots & E[(Y_{i1}-\mu_{i1})(Y_{in}-\mu_{in})] \\ E[(Y_{i2}-\mu_{i2})(Y_{i1}-\mu_{i1})] & E[(Y_{i2}-\mu_{i2})^2] & \cdots & E[(Y_{i2}-\mu_{i2})(Y_{in}-\mu_{in})] \\ \vdots & & \ddots & \cdots \\ E[(Y_{in}-\mu_{in})(Y_{i1}-\mu_{i1})] & E[(Y_{in}-\mu_{in})(Y_{i2}-\mu_{i2})] & \cdots & E[(Y_{in}-\mu_{in})^2] \end{bmatrix}$$

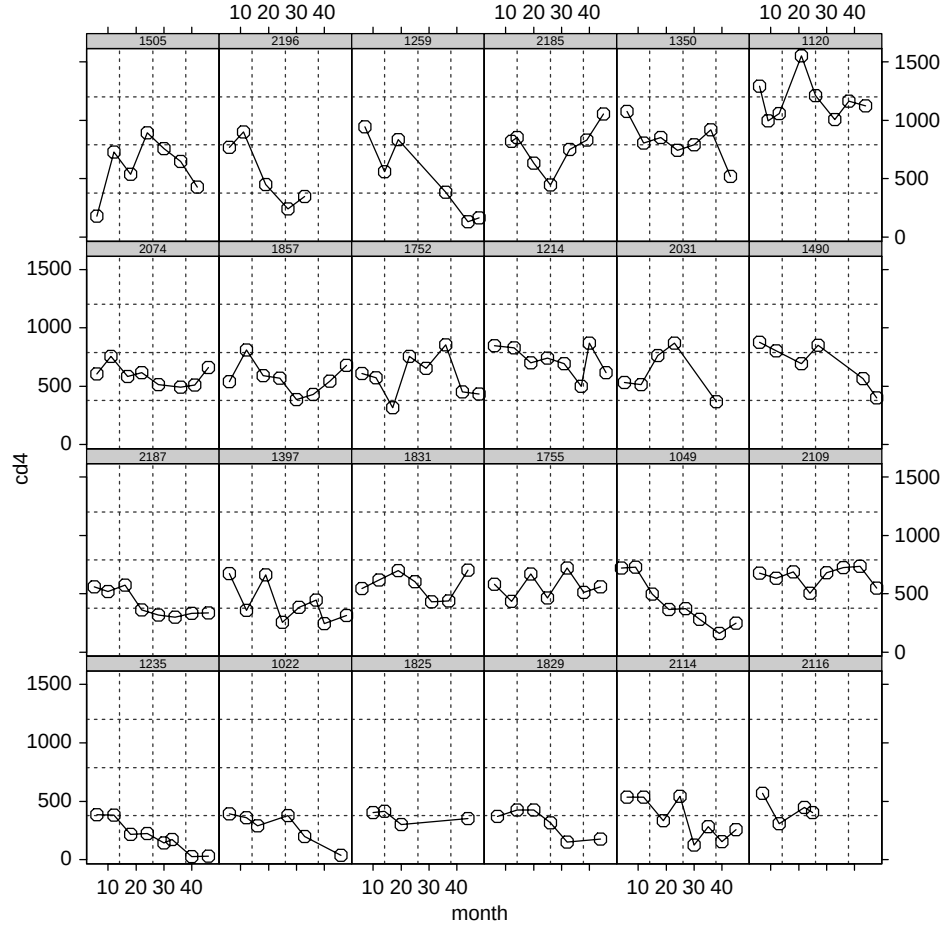

Figure 18.2 A sample of individual CD4 trajectories from the MACS data.

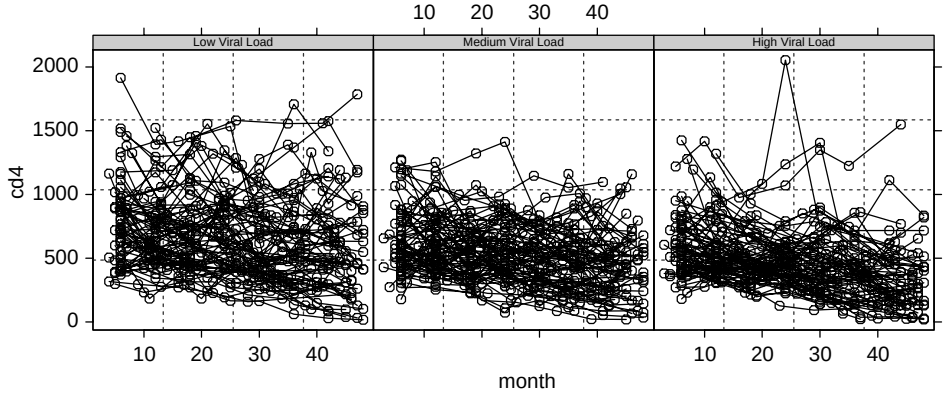

Figure 18.3 Individual CD4 trajectories from the MACS data by tertile of viral load.

The covariance can also be written in terms of the variances σ_j^2 and the correlations ρ_{jk}:

$$\text{cov}(Y_i) = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_{12} & \cdots & \sigma_1\sigma_n\rho_{1n} \\ \sigma_2\sigma_1\rho_{21} & \sigma_2^2 & \cdots & \sigma_2\sigma_n\rho_{2n} \\ \vdots & & \ddots & \vdots \\ \sigma_n\sigma_1\rho_{n1} & \sigma_n\sigma_2\rho_{n2} & \cdots & \sigma_n^2 \end{bmatrix}$$

Finally, the *correlation matrix* is given as

$$\text{corr}(Y_i) = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & \ldots & \rho_{2n} \\ \vdots & & \ddots & \vdots \\ \rho_{n1} & \rho_{2n} & \ldots & 1 \end{bmatrix}$$

which is useful for comparing the strength of association between pairs of outcomes, particularly when the variances σ_j^2 are not constant. Sample estimates of the correlations can be obtained using

$$\widehat{\rho}_{jk} = \frac{1}{N-1} \sum_i \frac{(Y_{ij} - \overline{Y}_{\cdot j})}{\widehat{\sigma}_j} \frac{(Y_{ik} - \overline{Y}_{\cdot k})}{\widehat{\sigma}_k}$$

where $\widehat{\sigma}_j^2$ and $\widehat{\sigma}_k^2$ are the sample variances of Y_{ij} and Y_{ik}, respectively (i.e., across subjects for times t_j and t_k).

Graphically, the correlation can be viewed using plots of Y_{ij} vs. Y_{ik} for all possible pairs of times t_j and t_k. These plots can be arranged in an array that corresponds to the covariance matrix and patterns of association across rows or columns can reveal changes in the correlation as a function of increasing time separation between measurements.

Example 18.1. (*continued*) For the HIVNET informed consent data, we focus on correlation analysis of outcomes from the control group. Parallel summaries would usefully characterize the similarity or difference in correlation structures for the control and intervention groups. The correlation matrix is estimated as follows:

	Month 0	**Month 6**	**Month 12**	**Month 18**
Month 0	1.00	0.471	0.394	0.313
Month 6	0.471	1.00	0.444	0.407
Month 12	0.394	0.444	1.00	0.508
Month 18	0.313	0.407	0.508	1.00

The matrix suggests that the correlation in outcomes from the same person is slightly decreasing as the time between the measurements increases. For example, the correlation between knowledge scores from baseline and month 6 is 0.471, while the correlation between baseline and month 12 decreases to 0.394, and decreases further to 0.313 for baseline and month 18. Correlation that decreases as a function of time separation is common among biomedical measurements and often reflects slowly varying underlying processes.

Example 18.2. (*continued*) For the MACS data the timing of measurement is only approximately regular. The following displays both the correlation matrix and the covariance matrix:

	Year 1	Year 2	Year 3	Year 4
Year 1	92,280.4	[0.734]	[0.585]	[0.574]
Year 2	63,589.4	81,370.0	[0.733]	[0.695]
Year 3	48,798.2	57,457.5	75,454.5	[0.806]
Year 4	55,501.2	63,149.9	70,510.1	101,418.2

The correlations are shown in brackets above. The variances are shown on a diagonal below the correlations. For example, the standard deviation among year 1 CD4 counts is $\sqrt{92{,}280.4} = 303.8$, while the standard deviations for years 2 through 4 are $\sqrt{81{,}370.0} = 285.3$, $\sqrt{75{,}454.5} = 274.7$, and $\sqrt{101{,}418.2} = 318.5$, respectively. Below the diagonal are the covariances, which together with the standard deviations determine the correlations. These data have a correlation for measurements that are one year apart of 0.734, 0.733, and 0.806. For measurements two years apart, the correlation decreases slightly to 0.585 and 0.695. Finally, measurements that are three years apart have a correlation of 0.574. Thus, the CD4 counts have a within-person correlation that is high for observations close together in time, but the correlation tends to decrease with increasing time separation between the measurement times.

An alternative method for exploring the correlation structure is through an array of scatter plots showing CD4 measured at year j versus CD4 measured at year k. Figure 18.4 displays these scatter plots. It appears that the correlation in the plot of year 1 vs. year 2 is stronger than for year 1 vs. year 3, or for year 1 vs. year 4. The sample correlations $\widehat{\rho}_{12} = 0.734$, $\widehat{\rho}_{13} = 0.585$, and $\widehat{\rho}_{14} = 0.574$ summarize the linear association presented in these plots.

18.3 DERIVED VARIABLE ANALYSIS

Formal statistical inference with longitudinal data requires either that a univariate summary be created for each subject or that methods for correlated data are used. In this section we review and critique common analytic approaches based on creation of summary measures.

A *derived variable analysis* is a method that takes a collection of measurements and collapses them into a single meaningful summary feature. In classical multivariate methods principal component analysis is one approach for creating a single major factor. With longitudinal data the most common summaries are the average response and the time slope. A second approach is a pre–post analysis which analyzes a single follow-up response in conjunction with a baseline measurement. In Section 18.3.1 we first review average or slope analyses, and then in Section 18.3.2 we discuss general approaches to pre–post analysis.

18.3.1 Average or Slope Analysis

In any longitudinal analysis the substantive aims determine which aspects of the response trajectory are most important. For some applications the repeated measures over time may be averaged, or if the timing of measurement is irregular, an area under the curve (AUC) summary can be the primary feature of interest. In these situations statistical analysis will focus on $\overline{Y}_i = 1/n \sum_{j=1}^{n} Y_{ij}$. A key motivation for computing an individual average and then focusing analysis on the derived averages is that standard methods can be used for inference such as a two-sample t-test. However, if there are any incomplete data, the advantage is lost since either subjects with partial data will need to be excluded, or alternative methods need to be invoked to handle the missingness. Attrition in longitudinal studies is unfortunately quite common, and thus derived variable methods are often more difficult to apply validly than they may first appear.

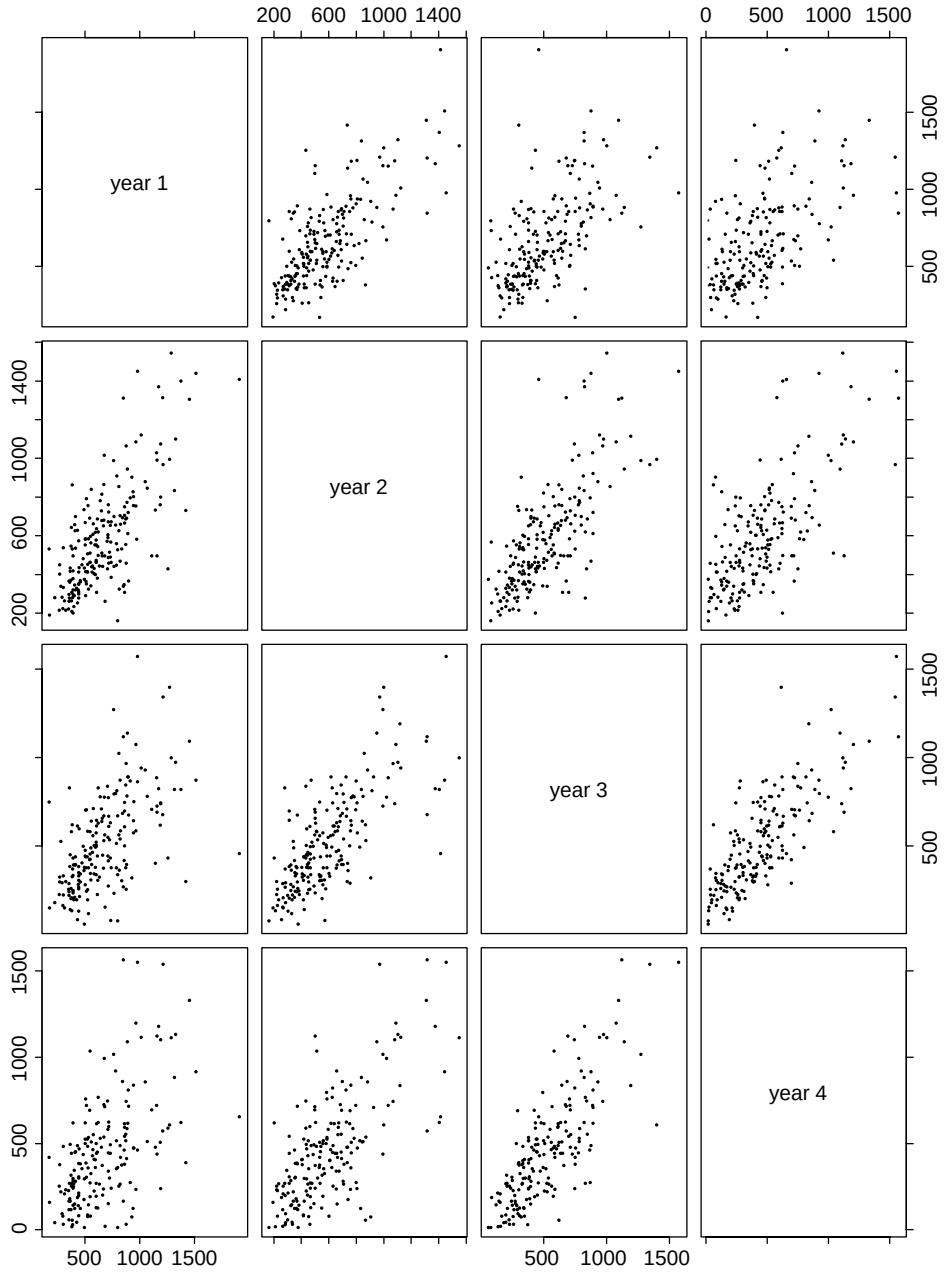

Figure 18.4 Scatter plots of CD4 measurements (counts/mL) taken at years 1 to 4 after seroconversion.

Example 18.1. (*continued*) In the HIVNET informed consent study, the goal is to improve participant knowledge. A derived variable analysis to evaluate evidence for an effect due to the mock informed consent process can be conducted using $\overline{Y}_i = (Y_{i1} + Y_{i2} + Y_{i3})/3$ for the post-baseline times $t_1 =$ six months, $t_2 = 12$ months, and $t_3 = 18$ months. The following table summarizes the data for subjects who have all three post-baseline measurements:

Group	Baseline *N*	Final *N*	Mean	SE	95% CI
Control	947	714	2.038	0.095	
Intervention	177	147	3.444	0.223	
Difference			1.406	0.243	[0.928, 1.885]

First, notice that only $714/947 = 75.4\%$ of control subjects, and $147/177 = 83.1\%$ of intervention subjects have complete data and are therefore included in the analysis. This highlights one major limitation to derived variable analysis: There may be selection bias due to exclusion of subjects with missing data. We discuss missing data issues in Section 18.6. Based on the data above, we would conclude that there is a statistically significant difference between the mean knowledge for the intervention and control groups with a two-sample t-test of $t = 5.796$, $p < 0.001$. Analysis of the single summary for each subject allows the repeated outcome variables to be analyzed using standard independent sample methods.

In other applications, scientific interest centers on the rate of change over time and therefore an individual's slope may be considered as the primary outcome. Typically, each subject in a longitudinal study has only a small number of outcomes collected at the discrete times specified in the protocol. For example, in the MACS data, each subject was to complete a study visit every 6 months and with complete data would have nine measurements between baseline and 48 months. If each subject has complete data, an individual summary statistic can be computed as the regression of outcomes Y_{ij} on times t_j: $Y_{ij} = \beta_{i,0} + \beta_{i,1} t_j + \epsilon_{ij}$; and $\widehat{\beta}_i$ is the ordinary least squares estimate based on data from subject i only. In the case where all subjects have the same collection of measurement times and have complete data, the variation in the estimated slope, $\widehat{\beta}_{i,1}$, will be equal across subjects provided that the variance of ϵ_{ij} is also constant across subjects. Therefore, if

1. The measurement times are common to all subjects: $t_1, t_2, \ldots, t_n$,
2. Each subject has a complete collection of measurements: $Y_{i1}, Y_{i2}, \ldots, Y_{in}$,
3. The within-subject variation $\sigma_i^2 = \text{var}(\epsilon_{ij})$ is constant across subjects: $\sigma_i^2 \equiv \sigma^2$,

then the summaries $\widehat{\beta}_{i,1}$ will have equal variances attributable to using simple linear regression to estimate individual slopes. If any of points 1 to 3 above do not hold, the variance of individual summaries may vary across subjects. This will be the case when each subject has a variable number of outcomes, due to missing data.

When points 1 to 3 are satisfied, simple inference on the derived outcomes $\widehat{\beta}_{i,1}$ can be performed using standard two-sample methods or regression methods. This allows inference regarding factors that are associated with the rate of change over time. If any of points 1 to 3 do not hold, mixed model regression methods (Section 18.5) may be preferable to simple derived variable methods. See Frison and Pocock [1992, 1997] for further discussion of derived variable methods.

Example 18.2. (*continued*) For the MACS data, we are interested in determining whether the rate of decline in CD4 is correlated with the baseline viral load measurement. In Section 18.2 we looked at descriptive statistics comparing the mean CD4 count over time for categories of viral load. We now explore the association between the rate of decline and baseline viral load by obtaining a summary statistic, using the individual time slope $\widehat{\beta}_i$ obtained from a regression of the CD4 count Y_{ij} on measurement time t_{ij}. Figure 18.5 shows a scatter plot of the individual slope estimates plotted against the log of baseline viral load. First notice that plotting symbols of different sizes are used to reflect the fact that the number of measurements per subject, n_i,

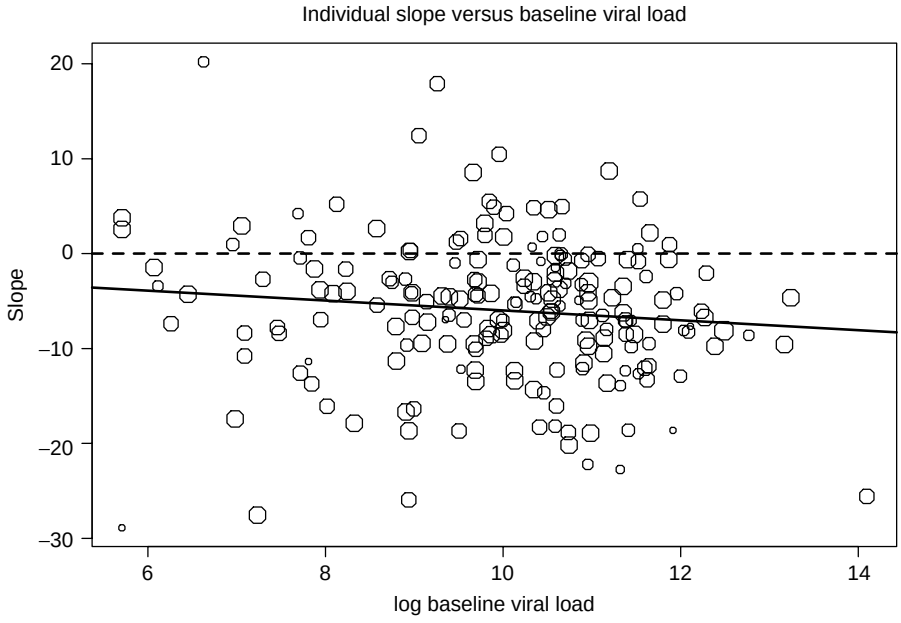

Figure 18.5 Individual CD4 slopes (count/month) vs. log of baseline viral load, MACS data.

is not constant. The plotting symbol size is proportional to n_i. For the MACS data we have the following distribution for the number of observations per subjects over the first four years:

	Number of Observations, n_i								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Number of subjects	5	13	8	10	25	44	82	117	3

For Figure 18.5 the $(5 + 13) = 18$ subjects with either one or two measurements were excluded as a summary slope is either unestimable ($n_i = 1$) or highly variable ($n_i = 2$). Figure 18.5 suggests that there is a pattern of decreasing slope with increasing log baseline viral load. However, there is also a great deal of subject-to-subject variation in the slopes, with some subjects having $\widehat{\beta}_{i,1} > 0$ count/month, indicating a stable or increasing trend, and some subjects having $\widehat{\beta}_{i,1} < 15$ count/month, suggesting a steep decline in their CD4. A linear regression using the individual slope as the response and log baseline viral load as the predictor yields a p-value of 0.124, implying a nonsignificant linear association between the summary statistic $\widehat{\beta}_{i,1}$ and log baseline viral load.

A categorical analysis using tertiles of baseline viral load parallels the descriptive statistics presented in Table 18.1. The average rate of decline in CD4 can be estimated as the mean of the individual slope estimates:

	***N* Subjects**	**Average Slope**	**SE**
Low viral load	66	−5.715	1.103
Medium viral load	69	−4.697	0.802
High viral load	65	−7.627	0.789

We find similar average rates of decline for the medium- and low-viral-load groups and find a greater rate of decline for the high-viral-load group. Using ANOVA, we obtain an F-statistic of 2.68 on 2 and 197 degrees of freedom, with a p-value of 0.071, indicating that we would not reject equality of average rates of decline using the nominal 5% significance level.

Note that neither simple linear regression nor ANOVA accounts for the fact that response variables $\widehat{\beta}_{i,1}$ may have unequal variance due to differing n_i. In addition, a small number of subjects were excluded from the analysis since a slope summary was unavailable. In Section 18.5 we discuss regression methods for correlated data that can efficiently use all of the available data to make inferences with longitudinal data.

18.3.2 Pre–Post Analysis

In this section we discuss analytic methods appropriate when a single baseline and a single follow-up measurement are available. We focus on the situation where interest is in the comparison of two groups: $X_i = 0$ denotes membership in a reference or control group; and $X_i = 1$ denotes membership in an exposure or intervention group. Assume for each subject i that we have a baseline measurement denoted as Y_{i0} and a follow-up measurement denoted as Y_{i1}. The following table summarizes three main analysis options using regression methods to characterize the two-group comparison:

Follow-up only:	$Y_{i1} = \beta_0 + \beta_1 X_i + \epsilon_i$
Change analysis:	$Y_{i1} - Y_{i0} = \beta_0^* + \beta_1^* X_i + \epsilon_i^*$
ANCOVA:	$Y_{i1} = \beta_0^{**} + \beta_1^{**} X_i + \beta_2^{**} Y_{i0} + \epsilon_i^{**}$

Since X_i is a binary response variable we can interpret the coefficients β_1, β_1^*, and β_1^{**} as differences in means comparing $X_i = 1$ to $X_i = 0$. Specifically, for the follow-up only analysis the coefficient β_1 represents the difference in the *mean response at follow-up* comparing $X_i = 1$ to $X_i = 0$. If the assignment to $X_i = 0/1$ was randomized, the simple follow-up comparison is a valid causal analysis of the effect of the treatment. For change analysis the coefficient β_1^* is interpreted as the difference between the *average change* for $X_i = 1$ as compared to the average change for $X_i = 0$. Finally, using ANCOVA estimates β_1^{**}, which represents the difference in the mean follow-up outcome comparing exposed ($X_i = 1$) to unexposed ($X_i = 0$) subjects who are *equal in their baseline response*. Equivalently, we interpret β_1^{**} as the comparison of treated versus control subjects after adjusting for baseline.

It is important to recognize that each of these regression models provides parameters with different interpretations. In situations where the selection of treatment or exposure is not randomized, the ANCOVA analysis can control for "confounding due to indication," or where the baseline value Y_{i0} is associated with a greater or lesser likelihood of receiving the treatment $X_i = 1$. When treatment is randomized, Frison and Pocock [1992] show that $\beta_1 = \beta_1^* = \beta_1^{**}$. This result implies that for a randomized exposure each approach can provide a valid estimate of the average causal effect of treatment. However, Frison and Pocock [1992] also show that the most *precise* estimate of β_1 is obtained using ANCOVA, and that final measurement analysis is more precise than the change analysis when the correlation between baseline and follow-up measurements is less than 0.50. This results from $\text{var}(Y_{i1} - Y_{i0}) = 2\sigma^2(1 - \rho)$, which is less than σ^2 only when $\rho > \frac{1}{2}$.

Example 18.1. (*continued*) To evaluate the effect of the HIVNET mock informed consent, we focus analysis on the baseline and six-month knowledge scores. The following tables give

inference for the follow-up, Y_{i1}:

Group	N	6-month Mean	SE	95% CI
Control	834	1.494	0.111	
Intervention	169	3.391	0.240	
Difference		1.900	0.264	[1.375, 2.418]

and for the change in knowledge score, $Y_{i1} - Y_{i0}$, for the 834/947 control subjects and 169/177 intervention subjects who have both baseline and six-month outcomes:

Group	N	Mean Change	SE	95% CI
Control	834	0.243	0.118	
Intervention	169	2.373	0.263	
Difference		2.130	0.288	[1.562, 2.697]

The correlation between baseline and month 6 knowledge score is 0.462 among controls and 0.411 among intervention subjects. Since $\rho < 0.5$, we expect an analysis of the change in knowledge score to lead to a larger standard error for the treatment effect than a simple cross-sectional analysis of scores at the six-month visit.

Alternatively, we can regress the follow-up on baseline and treatment:

Coefficients	Estimate	SE	Z-value
(Intercept)	0.946	0.105	9.05
Treatment	1.999	0.241	8.30
Baseline (Y_{i0})	0.438	0.027	16.10

In this analysis the estimate of the treatment effect is 1.999, with a standard error of 0.241. The estimate of β_1 is similar to that obtained from a cross-sectional analysis using six-month data only, and to the analysis of the change in knowledge score. However, as predicted, the standard error is smaller that the standard error for each alternative analysis approach. Finally, in Figure 18.6, the six-month knowledge score is plotted against the baseline knowledge score. Separate regression lines are fit and plotted for the intervention and control groups. We see that the fitted lines are nearly parallel, indicating that the ANCOVA assumption is satisfied for these data.

For discrete outcomes, different pre–post analysis options can be considered. For example, with a binary baseline, $Y_{i0} = 0/1$, and a binary follow-up, $Y_{i1} = 0/1$, the difference, $Y_{i1} - Y_{i0}$, takes the values $-1, 0, +1$. A value of -1 means that a subject has changed from $Y_{i0} = 1$ to $Y_{i1} = 0$, while $+1$ means that a subject has changed from $Y_{i0} = 0$ to $Y_{i1} = 1$. A difference of 0 means that a subject had the same response at baseline and follow-up and does not distinguish between $Y_{i0} = Y_{i1} = 0$ and $Y_{i0} = Y_{i1} = 1$. Rather than focus on the difference, it is useful to consider an analysis of change by subsetting on the baseline value. For example, in a comparative study we can subset on subjects with baseline value $Y_{i0} = 1$ and then assess the difference between intervention and control groups with respect to the percent that respond $Y_{i1} = 1$ at follow-up. This analysis allows inference regarding differential change from 0 to 1 comparing

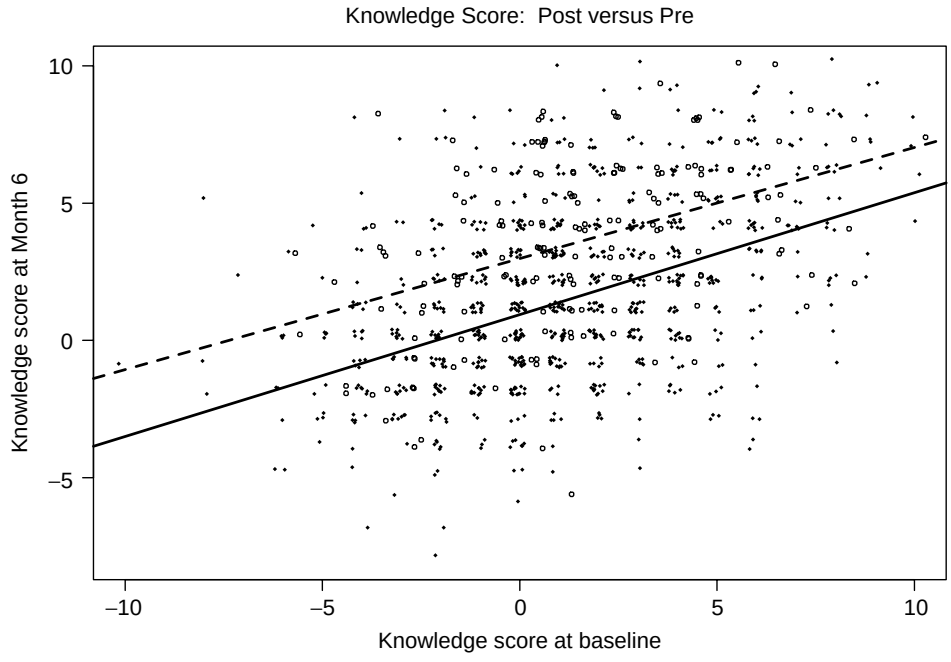

Figure 18.6 Month 6 knowledge score vs. baseline knowledge score (jittered), HIVNET informed consent substudy. Open points and dashed line represent intervention; solid points and line represent control.

the two groups. When a response value of 1 indicates a positive outcome, this analysis provides information about the "corrective" potential for intervention and control groups. An analysis that restricts to subjects with baseline $Y_{i0} = 1$ and then comparing treatment and control subjects at follow-up will focus on a second aspect of change. In this case we are summarizing the fraction of subjects that start with $Y_{i0} = 1$ and then remain with $Y_{i1} = 1$ and thus do not change their outcome but rather, maintain the outcome. When the outcome $Y_{ij} = 1$ indicates a favorable status, this analysis summarizes the relative ability of intervention and control groups to "maintain" the favorable status. Statistical inference can be based on standard two-sample methods for binary data (see Chapter 6). An analysis that summarizes current status at follow-up stratifying on the baseline, or previous outcome, is a special case of a transition model (see Diggle et al. [2002, Chap. 10]).

Example 18.1. (*continued*) The HIVNET informed consent substudy was designed to evaluate whether an informed consent procedure could correct misunderstanding regarding vaccine trial conduct and to reinforce understanding that may be tentative. In Section 18.2 we saw that for the safety item assessment at six months the intervention group had 50% of subjects answer correctly as compared to only 43% of control subjects. For the nurse item the fractions answering correctly at six months were 72% and 45% for intervention and control groups, respectively. By analyzing the six-month outcome separately for subjects that answered incorrectly at baseline, $Y_{i0} = 0$, and for subjects that answered correctly at baseline, $Y_{i0} = 1$, we can assess the mechanisms that lead to the group differences at six months: Does the intervention experience lead to greater rates of "correction" where answers go from $0 \rightarrow 1$ for baseline and six-month assessments; and does intervention appear to help "maintain" or reinforce correct knowledge by leading to increased rates of $1 \rightarrow 1$ for baseline and six-month responses?

The following table stratifies the month 6 safety knowledge item by the baseline response:

Safety Item

"Correction" : $Y_{i0} = 0$		Percent Correct	"Maintain" : $Y_{i0} = 1$		Percent Correct
	N	$Y_{i1} = 1$		N	$Y_{i1} = 1$
Control	488	160/488 = 33%	Control	349	198/349 = 57%
Intervention	105	43/105 = 41%	Intervention	65	42/65 = 65%

This table shows that of the 105 intervention subjects that answered the safety item at baseline incorrectly, a total of 43, or 41%, subsequently answered the item correctly at the 6-month follow-up visit. In the control group only 160/488 = 33% answered this item correctly at six months after they had answered incorrectly at baseline. A two-sample test of proportions yields a p-value of 0.118, indicating a nonsignificant difference between the intervention and control groups in their rates of correcting knowledge of this item. For subjects that answered this item correctly at baseline, 42/65 = 65% of intervention subjects and 198/349 = 57% of control subjects continued to respond correctly. A two-sample test of proportions yields a p-value of 0.230, indicating a nonsignificant difference between the intervention and control groups in their rates of maintaining correct knowledge of the safety item. Therefore, although the intervention group has slightly higher proportions of subjects that switch from incorrect to correct, and that stay correct, these differences are not statistically significant.

For the nurse item we saw that the informed consent led to a large fraction of subjects who answered the item correctly. At six months the intervention group had 72% of subjects answer correctly, while the control group had 45% answer correctly. Focusing on the mechanisms for this difference we find:

Nurse Item

"Correction" : $Y_{i0} = 0$		Percent Correct	"Maintain" : $Y_{i0} = 1$		Percent Correct
	N	$Y_{i1} = 1$		N	$Y_{i1} = 1$
Control	382	122/382 = 32%	Control	455	252/455 = 55%
Intervention	87	59/87 = 68%	Intervention	85	65/85 = 76%

Thus intervention led to a correction for 68% of subjects with an incorrect baseline response compared to 32% among controls. A two-sample test of proportions yields a p-value of <0.001, and a confidence interval for the difference in proportions of (0.250, 0.468). Therefore, the intervention has led to a significantly different rate of correction for the nurse item. Among subjects who correctly answered the nurse item at baseline, only 55% of control subjects answered correctly again at month 6, while 76% of intervention subjects maintained a correct answer at six months. Comparison of the proportion that maintain correct answers yields a p-value of <0.001 and a 95% confidence interval for the difference in probability of a repeat correct answer of (0.113, 0.339). Therefore, the informed consent intervention led to significantly different rates of both correction and maintenance for the safety item.

These categorical longitudinal data could also be considered as multiway contingency tables and analyzed by the methods discussed in Chapter 7.

18.4 IMPACT OF CORRELATION ON INFERENCE

For proper analysis of longitudinal data the within-subject correlation needs to be addressed. In Section 18.3.1 we discussed one method that avoids considering correlation among repeated measures by reducing the multiple measurements to a single summary statistic. In situations where there are variable numbers of observations per subject, alternative approaches are preferable. However, to analyze longitudinal outcomes, either a model for the correlation needs to be adopted or the standard error for statistical summaries needs to be adjusted. In this section we discuss some common correlation models and discuss the impact of the correlation on the standard errors and sample size.

18.4.1 Common Types of Within-Subject Correlation

The simplest correlation structure is the *exchangeable* or *compound symmetric* model, where

$$\text{corr}(Y_i) = \begin{bmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \rho & \rho & 1 & \cdots & \rho \\ \vdots & & & \ddots & \vdots \\ \rho & \rho & \rho & \cdots & 1 \end{bmatrix}$$

In this case the correlation between any two measurements on a given subject is assumed to be equal, $\text{corr}(Y_{ij}, Y_{ik}) = \rho_{jk} \equiv \rho$. The longitudinal outcomes form a simple "cluster" of responses, and the time ordering is not considered when characterizing correlation.

In other models the measurement time or measurement order is used to model correlation. For example, a *banded* correlation is

$$\text{corr}(Y_i) = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \rho_2 & \cdots & \rho_{n-2} \\ \rho_2 & \rho_1 & 1 & \rho_1 & \cdots & \rho_{n-3} \\ \rho_3 & \rho_2 & \rho_1 & 1 & \cdots & \rho_{n-4} \\ \vdots & & & \ddots & \vdots & \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \rho_{n-4} & \cdots & 1 \end{bmatrix}$$

and an *autoregressive* structure is

$$\text{corr}(Y_i) = \begin{bmatrix} 1 & \rho^{|t_1-t_2|} & \rho^{|t_1-t_3|} & \cdots & \rho^{|t_1-t_n|} \\ \rho^{|t_2-t_1|} & 1 & \rho^{|t_2-t_3|} & \cdots & \rho^{|t_2-t_n|} \\ \rho^{|t_3-t_1|} & \rho^{|t_3-t_2|} & 1 & \cdots & \rho^{|t_3-t_n|} \\ \vdots & & & \ddots & \vdots \\ \rho^{|t_n-t_1|} & \rho^{|t_n-t_2|} & \rho^{|t_n-t_3|} & \cdots & 1 \end{bmatrix}$$

Each of these models is a special case of a serial correlation model where the distance between observations determines the correlation. In a banded model correlation between observations is determined by their order. All observations that are adjacent in time are assumed to have an equal correlation: $\text{corr}(Y_{i1}, Y_{i2}) = \text{corr}(Y_{i2}, Y_{i3}) = \cdots = \text{corr}(Y_{in-1}, Y_{in}) = \rho_1$. Similarly, all observations that are two visits apart have correlation ρ_2, and in general all pairs of observations that are k visits apart have correlation ρ_k. A banded correlation matrix will have a total of $n-1$ correlation parameters. The autoregressive correlation model uses a single correlation parameter and assumes that the time separation between measurements determines their correlation through the model

$\text{corr}(Y_{ij}, Y_{ik}) = \rho^{|t_j - t_k|}$. Thus, if $\rho = 0.8$ and observations are 1 unit apart in time, their correlation will be $0.8^1 = 0.8$, while if they are 2 units apart, their correlation will be $0.8^2 = 0.64$. In an autoregressive model the correlation will decay as the distance between observations increases.

There are a large number of correlation models beyond the simple exchangeable and serial models given above. See Verbeke and Molenberghs [2000] and Diggle et al. [2002] for further examples.

18.4.2 Variance Inflation Factor

The impact of correlated observations on summaries such as the mean of all observations taken over time and across all subjects will depend on the specific form of the within-subject correlation. For example,

$$\overline{Y} = \frac{1}{\sum_i n_i} \sum_{i=1}^{N} \sum_{j=1}^{n_i} Y_{ij}$$

$$\text{var}(\overline{Y}) = \frac{1}{(\sum_i n_i)^2} \sum_{i=1}^{N} \left[\sum_{j=1}^{n_i} \text{var}(Y_{ij}) \; + \; \sum_{j=1}^{n_i - 1} \sum_{k=(j+1)}^{n_i} 2 \times \text{cov}(Y_{ij}, Y_{ik}) \right]$$

If the variance is constant, $\text{var}(Y_{ij}) = \sigma^2$, we obtain

$$\text{var}(\overline{Y}) = \frac{\sigma^2}{(\sum_i n_i)^2} \sum_{i=1}^{N} \left[n_i \; + \; \sum_{j=1}^{n_i - 1} \sum_{k=(j+1)}^{n_i} 2 \times \text{corr}(Y_{ij}, Y_{ik}) \right]$$

Finally, if all subjects have the same number of observations, $n_i \equiv n$, and the correlation is exchangeable, $\rho_{jk} \equiv \rho$, the variance of the mean is

$$\text{var}(\overline{Y}) = \frac{\sigma^2}{Nn} [1 \; + \; (n-1)\rho]$$

The factor $[1 + (n-1) \cdot \rho]$ is referred to as the *variance inflation factor*, since this measures the increase (when $\rho > 0$) in the variance of the mean (calculated using $N \cdot n$ observations) that is due to the within-subject correlation of measurements.

To demonstrate the impact of correlation on the variance of the mean, we calculate the variance inflation factor, $1 + (n-1)\rho$, for various values of cluster size, n, and correlation, ρ, in Table 18.4. This shows that even very small within-cluster correlations can have an important impact on standard errors if clusters are large. For example, a variance inflation factor of 2.0 arises with ($\rho = 0.001, n = 1001$), ($\rho = 0.01, n = 101$), or ($\rho = 0.10, n = 11$). The variance

Table 18.4 Variance Inflation Factors

n Cluster Size	ρ				
	0.001	0.01	0.02	0.05	0.1
2	1.001	1.01	1.02	1.05	1.10
5	1.004	1.04	1.08	1.20	1.40
10	1.009	1.09	1.18	1.45	1.90
100	1.099	1.99	2.98	5.95	10.90
1000	1.999	10.99	20.98	50.95	100.90

inflation factor becomes important when planning a study. In particular, when treatment is given to groups of subjects (e.g., a cluster randomized study), the variance inflation factor needs to be estimated to power the study properly. See Koepsell et al. [1991] or Donner and Klar [1994, 1997] for a discussion of design and analysis issues in cluster randomized studies. For longitudinal data each subject is a "cluster," with individual measurements taken within each subject.

18.5 REGRESSION METHODS

Regression methods permit inference regarding the average response trajectory over time and how this evolution varies with patient characteristics such as treatment assignment or other demographic factors. However, standard regression methods assume that all observations are independent and if applied to longitudinal outcomes may produce invalid standard errors. There are two main approaches to obtaining valid inference: A complete model that includes specific assumptions regarding the correlation of observations within a subject can be adopted and used to estimate the standard error of regression parameter estimates; general regression methods can be used and the standard errors can be corrected to account for the correlated outcomes. In the following section we review a regression method for continuous outcomes that models longitudinal data by assuming random errors within a subject and random variation in the trajectory among subjects.

18.5.1 Mixed Models

Figure 18.7 presents hypothetical longitudinal data for two subjects. In the figure monthly observations are recorded for up to one year, but one person drops out prior to the eight-month visit, and thus the observations for months 8 through 12 are not recorded. Notice that each subject

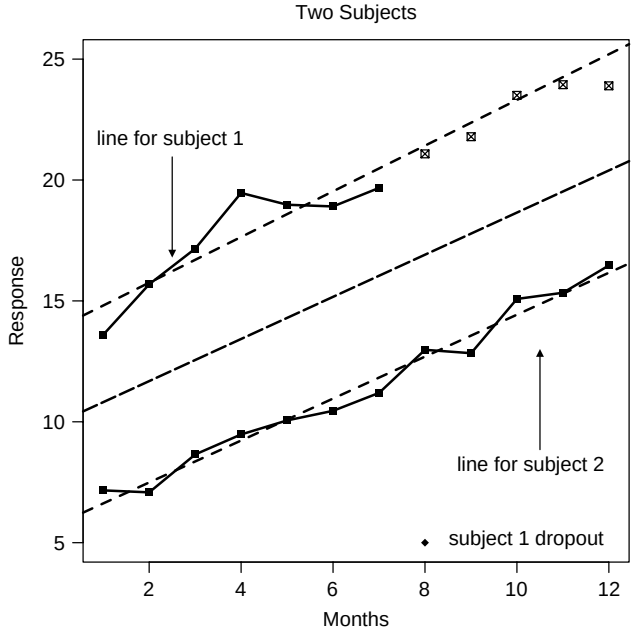

Figure 18.7 Hypothetical longitudinal data for two subjects. Each subject has an individual linear trajectory, and one subject has incomplete data due to dropout.

appears to be tracking his or her own linear trajectory but with small fluctuations about the line. The deviations from the individual observations to the individual's line are referred to as the within-subject variation in the outcomes. If we only had data for a single subject, these would be the typical error terms in a regression equation. In most situations the subjects in a study represent a random sample from a well-defined target population. In this case the specific individual line that a subject happens to follow is not of primary interest, but rather the *typical* linear trajectory and perhaps the magnitude of subject-to-subject variation in the longitudinal process. A dashed line in the center of Figure 18.7 shows the average of individual linear-time trajectories. This average curve characterizes the average for the population as a function of time. For example, the value of the dashed line at month 2 denotes the cross-sectional mean response if the two-month observation for all subjects was averaged. Similarly, the fitted value for the dashed line at 10 months represents the average in the population for the 10-month measurement. Therefore, the average line in Figure 18.7 represents both the typical trajectory and the population average as a function of time.

Linear mixed models make specific assumptions about the variation in observations attributable to variation within a subject and to variation among subjects. The within-subject variation is seen in Figure 18.7 as the deviation between individual observations, Y_{ij}, and the individual linear trajectory. Let $\beta_{i,0} + \beta_{i,1}X_{ij}$ denote the line that characterizes the observation path for subject i. In this example X_{ij} denotes the time of measurement j on subject i. Note that each subject has an individual-specific intercept and slope. Within-subject variation is seen in the magnitude of variation in the deviation between the observations and the individual trajectory, $Y_{ij} - (\beta_{i,0} + \beta_{i,1}X_{ij})$. The between-subject variation is represented by the variation among the intercepts, $\text{var}(\beta_{i,0})$, and the variation among subjects in the slopes, $\text{var}(\beta_{i,1})$.

If parametric assumptions are made regarding the within- and between-subject components of variation, maximum likelihood methods can be used to estimate the regression parameters which characterize the population average, and the variance components which characterize the magnitude of within- and between-subject heterogeneity. For continuous outcomes it is convenient to assume that within-subject errors are normally distributed and to assume that intercepts and slopes are normally distributed among subjects. Formally, these assumptions are written as:

$$\begin{aligned}
\text{within-subjects} \quad : \quad E(Y_{ij} \mid \beta_i) &= \beta_{i,0} + \beta_{i,1}X_{ij} \\
Y_{ij} &= \beta_{i,0} + \beta_{i,1}X_{ij} + \epsilon_{ij} \\
\epsilon_{ij} &\sim N(0, \sigma^2) \\
\text{between-subjects} \quad : \quad \begin{pmatrix} \beta_{i,0} \\ \beta_{i,1} \end{pmatrix} &\sim N\left[\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \begin{pmatrix} D_{00} & D_{01} \\ D_{10} & D_{11} \end{pmatrix} \right]
\end{aligned}$$

The model can be rewritten using $b_{i,0} = (\beta_{i,0} - \beta_0)$ and $b_{i,1} = (\beta_{i,1} - \beta_1)$:

$$Y_{ij} = \underbrace{\beta_0 + \beta_1 X_{ij}}_{\text{systematic}} + \underbrace{b_{i,0} + b_{i,1}X_{ij} + \epsilon_{ij}}_{\text{random}} \tag{1}$$

In this representation the terms $b_{i,0}$ and $b_{i,1}$ represent deviations from the population average intercept and slope, respectively. These "random effects" now have mean 0 by definition, but their variance and covariance is still given by the elements of the matrix D. For example, $\text{var}(b_{i,0}) = D_{00}$ and $\text{var}(b_{i,0}) = D_{11}$. In equation (1) the "systematic" variation in outcomes is given by the regression parameters β_0 and β_1. These parameters determine how the average for subpopulations differs across distinct values of the covariates, X_{ij}.

In equation (1) the random components are partitioned into the observation-level and subject-level fluctuations:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \underbrace{b_{i,0} + b_{i,1} X_{ij}}_{\text{between-subject}} + \underbrace{\epsilon_{ij}}_{\text{within-subject}}$$

A more general form is

$$Y_{ij} = \underbrace{\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}}_{\text{fixed effects}} + \underbrace{b_{i,0} + b_{i,1} X_{i1} + \cdots + b_{i,q} X_{iq}}_{\text{random effects}} + \epsilon_{ij}$$

$$Y_{ij} = X'_{ij}\beta + Z'_{ij} b_i + \epsilon_{ij}$$

where $X'_{ij} = [X_{ij,1}, X_{ij,2}, \ldots, X_{ij,p}]$ and $Z'_{ij} = [X_{ij,1}, X_{ij,2}, \ldots, X_{ij,q}]$. In general, we assume that the covariates in Z_{ij} are a subset of the variables in X_{ij} and thus $q < p$. In this model the coefficient of covariate k for subject i is given as $(\beta_k + b_{i,k})$ if $k \leq q$ and is simply β_k if $q < k \leq p$. Therefore, in a linear mixed model there may be some regression parameters that vary among subjects, while some regression parameters are common to all subjects. For example, in Figure 18.7 it is apparent that each subject has his or her own intercept, but the subjects may have a common slope. A *random intercept model* assumes parallel trajectories for any two subjects and is given as a special case of the general mixed model:

$$Y_{ij} = \beta_0 + \beta_1 X_{i1} + b_{i,0} + \epsilon_{ij}$$

In this model the intercept for subject i is given by $\beta_0 + b_{i,0}$, while the slope for subject i is simply β_1, since there is no additional random slope, $b_{i,1}$, in the random intercept model.

Laird and Ware [1982] discuss the linear mixed model and specific methods to obtain maximum likelihood estimates. Although linear mixed models can be computationally difficult to fit, modern software packages contain excellent numerical routines for estimating parameters and computing standard errors. For example, the SAS package contains the MIXED procedure and S-PLUS has the lme() function.

Example 18.2. (*continued*) In Section 18.3.1 we explored the change over time in CD4 counts for groups of subjects according to their baseline viral load value. Using linear mixed models we can estimate the average rate of decline for each baseline viral load category, and test for differences in the rate of decline.

To test for differences in the rate of decline, we use linear regression with

$$\begin{aligned} E(Y_{ij} \mid X_{ij}) = \beta_0 &+ \\ &\beta_1 \cdot \text{month} + \\ &\beta_2 \cdot I(\text{medium viral load}) + \\ &\beta_3 \cdot I(\text{high viral load}) + \\ &\beta_4 \cdot \text{month} \cdot I(\text{medium viral load}) + \\ &\beta_5 \cdot \text{month} \cdot I(\text{high viral load}) \quad . \end{aligned}$$

Here $X_{ij,3} = I(\text{medium viral load}) = 1$ if subject i has a medium value for baseline viral load and otherwise $= 0$, and $X_{ij,4} = I(\text{high viral load}) = 1$ if subject i has a high baseline viral load and otherwise $= 0$. Using this regression model, the average slope for the low baseline viral category is given by β_1, while the average slope for the other viral load categories are given by $(\beta_1 + \beta_4)$ and $(\beta_1 + \beta_5)$ for the medium- and high-viral-load categories, respectively. If the

estimate of β_4 is not significantly different from 0, we cannot reject equality of the average rates of decline for the low- and medium-viral-load subjects. Similarly, inference regarding β_5 determines whether there is evidence that the rate of decline for high-viral-load subjects is different than for low-viral-load subjects.

The linear mixed model is specified by the regression model for $E(Y_{ij} \mid X_{ij}) = \mu_{ij}$ and assumptions about random effects. We first assume random intercepts, $Y_{ij} = \mu_{ij} + b_{i,0} + \epsilon_{ij}$, and then allow random intercepts and slopes, $Y_{ij} = \mu_{ij} + b_{i,0} + b_{i,1} \cdot \text{month} + \epsilon_{ij}$. Maximum likelihood estimates are presented in Tables 18.5 and 18.6. In Table 18.5 the mixed model assumes that each subject has a random intercept, $b_{i,0}$, but assumes a common slope. In this model there are two estimated variance components: $162.5 = \widehat{\sigma} = \sqrt{\widehat{\text{var}}(\epsilon_{ij})}$ and $219.1 = \sqrt{\widehat{D_{00}}} = \sqrt{\widehat{\text{var}}(b_{i,0})}$. The total variation in CD4 is estimated as $162.5^2 + 219.1^2 = 272.8^2$, and the proportion of total variation that is attributed to within-person variability is $162.5^2/272.8^2 = 35\%$ with $219.1^2/272.8^2 = 65\%$ of total variation attributable to individual variation in their general level of CD4 (e.g., attributable to random intercepts).

Estimates from Table 18.5 are interpreted as follows:

- *(Intercept)* $\widehat{\beta}_0 = 803.4$. The intercept is an estimate of the mean CD4 count at seroconversion (i.e., month = 0) among the low-viral-load subjects.
- *month* $\widehat{\beta}_1 = -5.398$: Among subjects in the low-viral-load group, the mean CD4 declines -5.398 units per month.
- *I[Medium Viral Load]* $\widehat{\beta}_2 = -123.72$. At seroconversion the average CD4 among subjects with a medium value for baseline viral load is 123.72 units lower than the average CD4 among the low-viral-load subjects.
- *I[High Viral Load]* $\widehat{\beta}_3 = -146.40$. At seroconversion the average CD4 among subjects with a high value for baseline viral load is 146.40 units lower than the average CD4 among the low-viral-load subjects.
- *month * I[Medium Viral Load]* $\widehat{\beta}_4 = 0.169$. The rate of decline for subjects in the medium-viral-load category is estimated to be 0.169 count/month higher than the rate of decline among subjects with a low-baseline viral load. The rate of change in mean CD4 is estimated as $-5.398 + 0.169 = -5.229$ counts/month among subjects with medium-baseline viral load.

Table 18.5 Linear Mixed Model Results for the CD4 Data Assuming Random Intercepts[a]

Linear mixed-effects model fit by maximum likelihood

Data: MACS

AIC	BIC	logLik
19838.98	19881.38	−9911.491

Random effects:

Formula: ~ 1 | id

	(Intercept)	Residual
StdDev:	219.1106	162.5071

Fixed effects: cd4 ~ month * vcat

	Value	Std.Error	DF	t-value	p-value
(Intercept)	803.356	29.712	1250	27.04	<.0001
month	−5.398	0.578	1250	−9.34	<.0001
I[Medium Viral Load]	−123.724	42.169	223	−2.93	0.0037
I[High Viral Load]	−146.401	42.325	223	−3.46	0.0006
month * I[Medium Viral Load]	0.169	0.812	1250	0.21	0.8351
month * I[High Viral Load]	−1.968	0.817	1250	−2.41	0.0162

[a]Output from S-PLUS.

Table 18.6 Linear Mixed Model Results for the CD4 Data Assuming Random Intercepts and Slopes[a]

Linear mixed-effects model fit by maximum likelihood

Data: MACS

AIC	BIC	logLik
19719.85	19772.84	−9849.927

Random effects:

Formula: ~ 1 + month | id

Structure: General positive-definite

	StdDev	Corr
(Intercept)	244.05874	(Inter
month	5.68101	−0.441
Residual	142.22835	

Fixed effects: cd4 ~ month * vcat

	Value	Std.Error	DF	t-value	p-value
(Intercept)	803.509	31.373	1250	25.61	<.0001
month	−5.322	0.857	1250	−6.21	<.0001
I[Medium Viral Load]	−125.548	44.536	223	−2.82	0.0053
I[High Viral Load]	−142.177	44.714	223	−3.18	0.0017
month * I[Medium Viral Load]	0.159	1.205	1250	0.13	0.8954
month * I[High Viral Load]	−2.240	1.212	1250	−1.85	0.0648

[a]Output from S-PLUS.

- *month * I[High Viral Load]* $\widehat{\beta}_5 = -1.967$. The rate of decline for subjects in the high-viral-load category is estimated to be -1.967 counts/month lower than the rate of decline among subjects with a low-baseline viral load. The rate of change in mean CD4 is estimated as $-5.398 - 1.967 = -7.365$ counts/month among subjects with a high-baseline viral load.

Although the regression output also includes standard errors for each of the regression estimates, we defer making inference since a model with random intercepts and random slopes appears more appropriate and affects the resulting confidence intervals or tests for the regression estimates (see Table 18.6).

In Table 18.6 we present maximum likelihood estimates assuming random intercepts and random slopes. To assess whether the additional flexibility is warranted, we can evaluate the improvement in the fit to the data as measured by the maximized log likelihood. The maximized log likelihood for random intercepts is -9911.49 (see Table 18.5), while the maximized log likelihood is increased by 61.56 to -9849.93 when also allowing random intercepts. A formal likelihood ratio test is possible since the random intercepts and random intercepts plus slopes form nested models, but since the null hypothesis restriction involves $D_{11} = 0$, which is on the boundary of the allowable values for variance components (i.e., $D_{11} \geq 0$), the null reference distribution is of nonstandard form [Stram and Lee, 1994; Verbeke and Molenberghs, 2000]. However, the increase in maximized log likelihood of 61.56 is quite substantial and statistically significant with $p < 0.001$. Although the variance assumptions can be further relaxed to allow serial correlation in the measurement errors, ϵ_{ij}, the improvement in the maximized log likelihood is small and does not substantially affect the conclusions. We refer the reader to Diggle et al. [2002] and Verbeke and Molenberghs [2000] for further detail regarding linear mixed models that also include serial correlation in the errors.

Table 18.6 gives estimates of the variance components. For example, the standard deviation in intercepts is estimated as $\sqrt{\widehat{D}_{00}} = 244.1$ and the standard deviation of slopes is given

as $\sqrt{\widehat{D}_{11}} = 5.681$. Under the assumption of normally distributed random effects, these estimates imply that 95% of subjects with a low-baseline viral load would have a *mean* CD4 at seroconversion between $803.5 - 1.96 \times 244.1 = 325.1$ and $803.5 + 1.96 \times 244.1 = 1281.9$. We emphasize that this interval is for individual values of the mean CD4 at baseline rather than for individual measurements at baseline. The interval (325.1, 1281.9) does not include the measurement variation attributable to ϵ_{ij} so only describes the variation in the means, $\beta_0 + b_{i,0}$, and not the actual CD4 measurements, $Y_{ij} = \beta_0 + b_{i,0} + \epsilon_{ij}$. Similarly, 95% of low-viral-load subjects are expected to have a slope of $-5.322 \pm 1.96 \times 5.681 = (-16.456, 5.813)$ counts/month.

The estimated regression parameters can be used to make inference regarding the average rate of decline for each of the baseline viral load categories. For example, $\widehat{\beta}_4 = 0.159$ estimates the difference between the rate of decline among medium-viral-load subjects and low-viral-load subjects and is not significantly different from 0 using the standardized regression coefficient as test statistic: $0.159/1.205 = 0.13$ with $p = 0.8954$. Although the estimated rate of decline is lower for the high-viral-load group, $\widehat{\beta}_5 = -2.240$, this is also not significantly different from 0 with p-value 0.0648. It is important to point out that inference using linear mixed models can be quite sensitive to the specific random effects assumptions. If a random intercepts model were used, the comparison of high- versus low-viral-load group slopes over time becomes statistically significant, as seen in Table 18.5, where the p-value for testing $H_0 : \beta_5 = 0$ is $p = 0.0162$, which would naively lead to rejection of the null hypothesis. This inference is invalid, as it assumes that slopes do not vary among individuals, and the data clearly suggest between-subject variation in slopes.

Residual plots can be useful for checking the assumptions made by the linear mixed model. However, there are two types of residuals that can be used. First, the *population residuals* are defined as

$$\begin{aligned} R_{ij}^P &= Y_{ij} - (\widehat{\beta}_0 + \widehat{\beta}_1 X_{ij,1} + \cdots + \widehat{\beta}_p X_{ij,p}) \\ &= Y_{ij} - X_{ij}' \widehat{\beta} \end{aligned}$$

The population residuals measure the deviation from the individual measurement to the fitted population mean value. These residuals contain all components of variation, including between- and within-subject deviations since

$$Y_{ij} - X_{ij}' \beta = Z_{ij}' b_i + \epsilon_{ij}$$

The population residuals can be used to evaluate evidence for systematic departures from linear assumptions. Similar to standard multiple regression, plots of residuals versus predictors can be inspected for curvature.

Individual random effects b_i can also be estimated and used to form a second type of residual. Under the linear mixed model, these random effects are typically not estimated simply by using subject i data only to estimate b_i, but rather by using both the individual data $Y_{i1}, Y_{i2}, \ldots, Y_{i,n_i}$ and the assumption that random effects are realizations from a normal distribution among subjects. Empirical Bayes' estimates of b_i balance the assumption that b_i is intrinsic to generating the data Y_{ij} in addition to the assumption that the distribution of b_i is multivariate normal with mean 0. Thus, empirical Bayes' estimates are typically closer to 0 than estimates that would be obtained solely by using individual i data. See Carlin and Louis [1996] for more detail on empirical Bayes' estimation. Using the estimated random effects provides a second residual:

$$\begin{aligned} R_{ij}^W &= Y_{ij} - (\widehat{\beta}_0 + \widehat{\beta}_1 X_{ij,1} + \cdots + \widehat{\beta}_p X_{ij,p}) \\ &\quad - (\hat{b}_{i,0} + \hat{b}_{i,1} X_{ij,1} + \cdots + \hat{b}_{i,q} X_{ij,q}) \\ &= Y_{ij} - X_{ij}' \widehat{\beta} - Z_{ij}' \widehat{b}_i \end{aligned}$$

If the regression parameter β and the random effects b were known rather than estimated, the residual R_{ij}^W would equal the within-subject error ϵ_{ij}. The within-subject residuals R_{ij}^W can be used to assess the assumptions regarding the within-subject errors.

Example 18.2. (*continued*) We use the random intercepts and random slopes model for the CD4 data to illustrate residual analysis for linear mixed models. The population residuals are plotted in Figure 18.8, and the within-subject residuals are plotted in Figure 18.9. First, no violation of the linearity assumption for month is apparent in either of these plots. Second, the population residuals are weakly suggestive of an increasing variance over time. However, it is important to note that under the assumption of random intercepts and random slopes, the total variance, $\text{var}(b_{i,0} + b_{i,1} \cdot \text{month} + \epsilon_{ij})$, may be an increasing or decreasing function of time. The population residuals suggest right skewness in the cross-sectional distribution of CD4. Since the within-subject residuals do not appear skewed, the population residuals suggest that the random effects may not be normally distributed. Figure 18.10 presents histograms of the estimated intercepts and slopes obtained using ordinary linear regression for subject i data rather than the empirical Bayes estimates. The histograms for the individual intercepts appear to be right skewed, while the individual slopes appear symmetrically distributed. Therefore, residual analysis coupled with exploratory analysis of individual regression estimates suggests that linearity assumptions appear satisfied, but normality of random effects may be violated. The linear mixed model is known to be moderately robust to distributional assumptions, so large-sample inference regarding the average rate of decline for baseline viral load groups can be achieved.

Mixed models can be adopted for use with categorical and count response data. For example, random effects can be included in logistic regression models for binary outcomes and can be included in log-linear models for count data. Maximum likelihood estimation for these models requires specialized software. Extensions of mixed models to alternate regression contexts is discussed in Chapters 7 and 9 of Diggle et al. [2002].

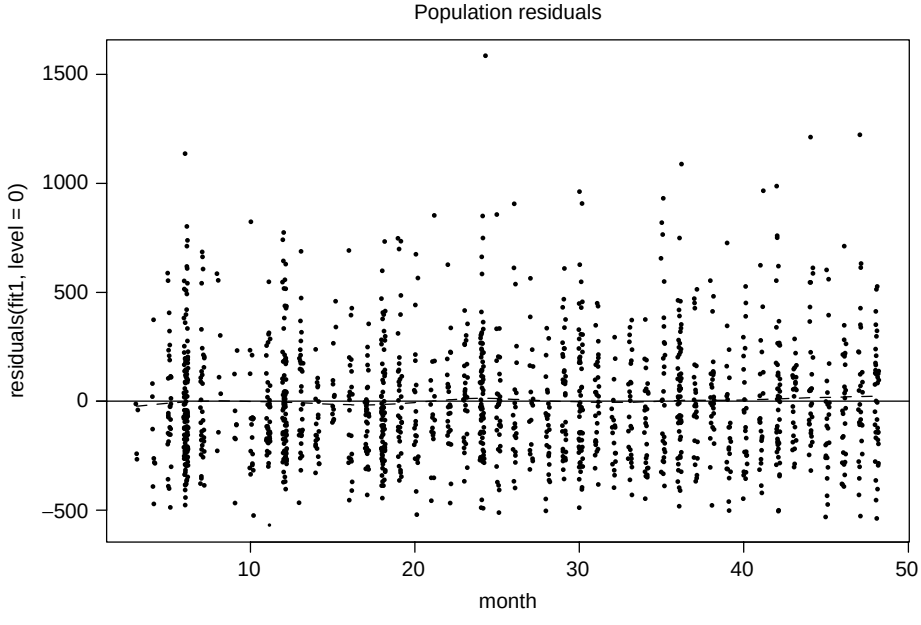

Figure 18.8 Population residuals, R_{ij}^P, vs. visit month for the MACS CD4 data. The dashed line is a smooth curve through the residuals.

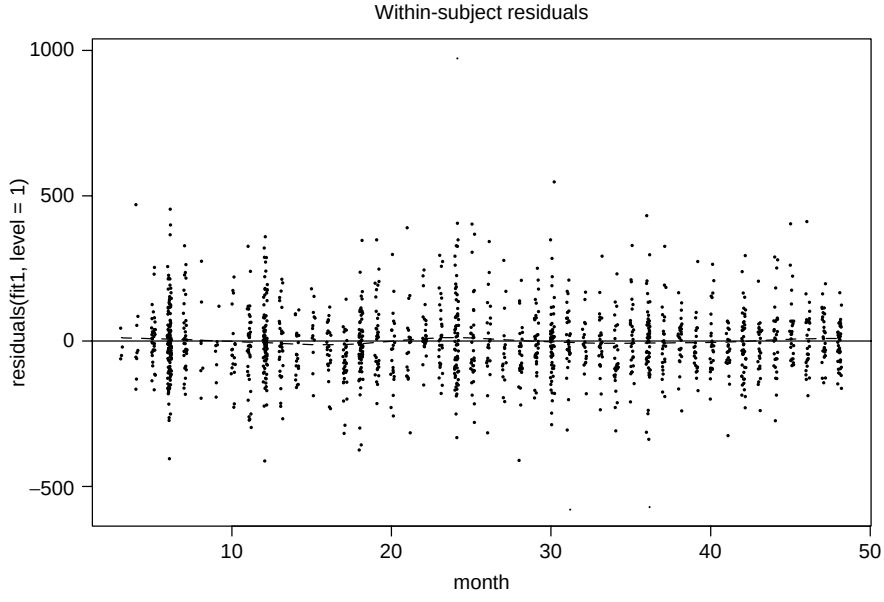

Figure 18.9 Within-subject residuals, R_{ij}^{W}, vs. visit month for the MACS CD4 data. The dashed line is a smooth curve through the residuals.

18.5.1.1 Summary

- Linear mixed models permit regression analysis with correlated data.
- Mixed models specify variance components that represent within-subject variance in outcomes and between-subject variation in trajectories.
- Linear mixed model parameters can be estimated using maximum likelihood.

18.5.2 Generalized Estimating Equations

A second regression approach for inference with longitudinal data is known as *generalized estimating equations* (GEE) [Liang and Zeger, 1986]. In this approach two models are specified. First, a regression model for the mean response is selected. The form of the regression model is completely flexible and can be a linear model, a logistic regression model, a log-linear model, or any generalized linear model [McCullagh and Nelder, 1989]. Second, a model for the within-subject correlation is specified. The correlation model serves two purposes: It is used to obtain weights (covariance inverse) that are applied to the vectors of observations from each subject to obtain regression coefficient estimates; and the correlation model is used to provide model-based standard errors for the estimated coefficients.

A regression model specifies a structure for the mean response, $\mu_{ij} = E(Y_{ij} \mid X_{ij})$, as a function of covariates. For longitudinal data the mean μ_{ij} has been called the *marginal mean* since it does not involve any additional variables, such as random effects, b_i, or past outcomes, Y_{ij-1}. Mixed models consider means conditional on random effects, and transition models include past outcomes as covariates. Adding additional variables leads to subtle changes in the interpretation of covariate coefficients, which becomes particularly important for nonlinear models such as logistic regression. See Diggle et al. [2002, Chaps. 7 and 11] for further discussion.

GEE has two important robustness properties. First, the estimated regression coefficients, $\widehat{\beta}$, obtained using GEE are broadly valid estimates that approach the correct value with increasing

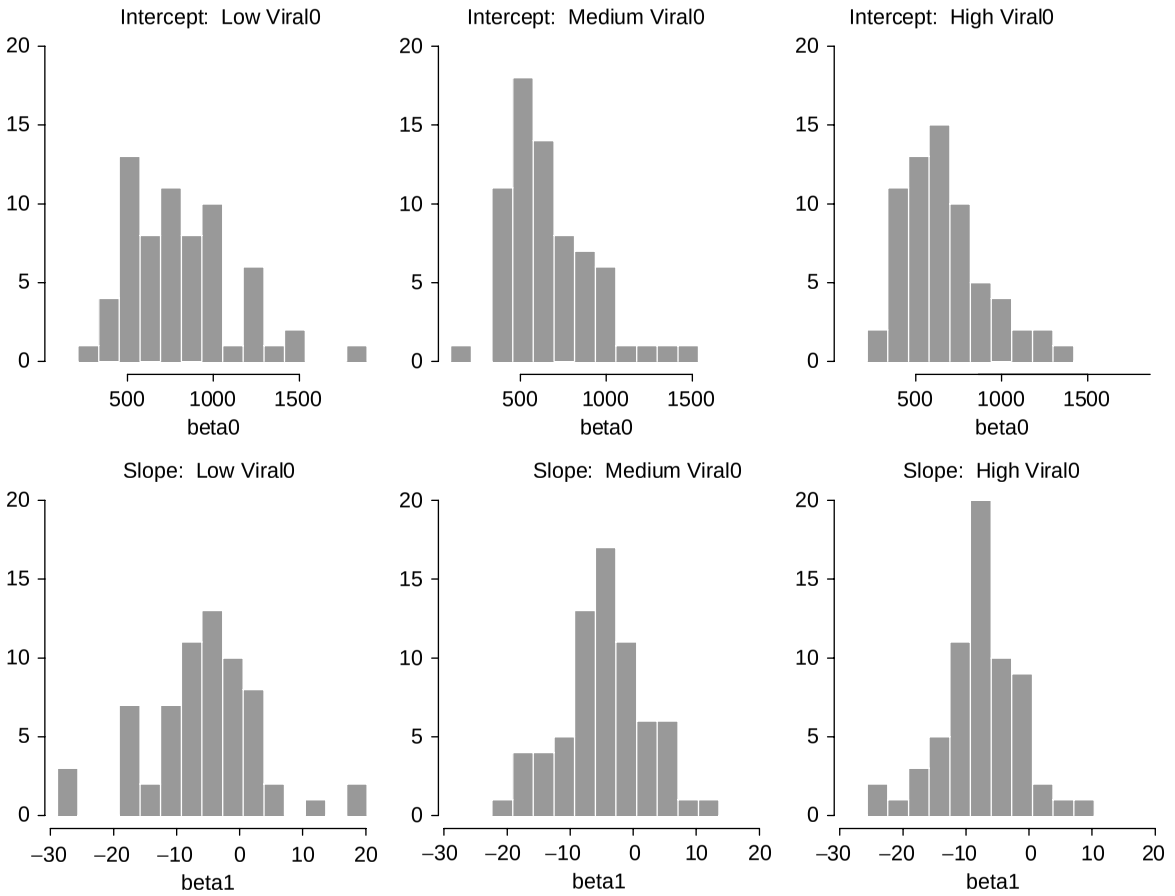

Figure 18.10 Estimates of individual intercepts and slopes by baseline viral load category for the MACS CD4 data.

sample size regardless of the choice of correlation model. In this respect the correlation model is used simply to weight observations, and a good correlation model choice can lead to more precise estimation of regression coefficients than can a poor choice. Based on optimal estimation theory (e.g., Gauss–Markov theory), the best correlation model choice for efficiency of estimation is the true correlation structure. Second, the correlation choice is used to obtain model-based standard errors, and these do require that the correlation model choice is correct in order to use the standard errors for inference. A standard feature of GEE is the additional reporting of *empirical standard errors*, which provide valid estimates of the uncertainty in $\widehat{\beta}$, even if the correlation model is not correct. Therefore, the correlation model can be any model, including one that assumes observations are independent, and proper large-sample standard errors obtained using the empirical estimator. Liang and Zeger [1993] provide an overview of regression methods for correlated data, and Hanley et al. [2003] give an introduction to GEE for an epidemiological audience.

Example 18.2. (*continued*) We return to the CD4 data and use GEE to investigate whether the rate of decline in CD4 over the first 48 months postseroconversion seems to depend on the baseline viral load category. Table 18.7 presents the estimates obtained using GEE and an independence correlation model. Standard errors using the independence correlation model are identical to those obtained from linear regression and are labeled as "model-based." In this application the key feature provided by GEE are the "empirical" standard errors, which are generally valid estimates of the uncertainty associated with the regression estimates. Notice that most of the empirical standard errors are larger than the naive model-based standard errors, which assume that the data are independent. However, corrected standard errors can be either larger or smaller than standard errors obtained under an independence assumption, and the nature of the covariate and the correlation structure interact to determine the proper standard error. It is an oversimplification to state that correction for correlation will lead to larger standard errors. Using GEE we obtain conclusions similar to that obtained using linear mixed models: The high-viral-load group has a steeper estimated rate of decline, but the difference between low and high groups is not statistically significant.

Example 18.1. (*continued*) GEE is particularly useful for binary data and count data. We now turn to analysis of the nurse item from the HIVNET informed consent study. We need to choose a regression model and a correlation model. For our first analysis we assume a common proportion answering correctly after randomization. For this analysis we create the covariate "post," which takes the value 1 if the visit occurs at month 6, 12, or 18, and takes the value 0 for the baseline visit. We use the variable "ICgroup" to denote the intervention and control group, where $\text{ICgroup}_{ij} = 1$ for all visits $j = 1, 2, 3, 4$ if the subject was randomized to the mock informed consent, and $\text{ICgroup}_{ij} = 0$ for all visits, $j = 1, 2, 3, 4$, if the subject was randomized to the control group. Since the response is binary, $Y_{ij} = 1$ if the item was correctly answered by subject i at visit j and 0 otherwise, we use logistic regression to characterize the

Table 18.7 GEE Estimates for the CD4 Data Using an Independence Working Correlation Model

		Standard Error		Z-statistic	
	Estimate	Model	Empirical	Model	Empirical
(Intercept)	792.897	26.847	36.651	29.534	21.633
Month	−4.753	0.950	1.101	−5.001	−4.318
I(Medium viral load)	−121.190	37.872	46.886	−3.200	−2.585
I(high viral load)	−150.705	37.996	45.389	−3.966	−3.320
Month · I(medium viral load)	−0.301	1.341	1.386	−0.224	−0.217
Month · I(high viral load)	−1.898	1.346	1.297	−1.410	−1.464

probability of a correct response as a function of time and treatment group:

$$\begin{aligned}\text{logit}P(Y_{ij} = 1 \mid X_i) = \beta_0 + \\ \beta_1 \cdot \text{post}_{ij} + \\ \beta_2 \cdot \text{ICgroup}_{ij} + \\ \beta_3 \cdot \text{ICgroup}_{ij} \cdot \text{post}_{ij}\end{aligned}$$

Since the visits are equally spaced and each subject is scheduled to have a total of four measurements, we choose to use an unstructured correlation matrix. This allows the correlations ρ_{jk} to be different for each pair of visit times (j, k).

In Table 18.8 we provide GEE estimates obtained using the SAS procedure GENMOD. The estimated working correlation is printed and indicates correlation that decreases as the time separation between visits increases. For example, the estimated correlation for Y_{i1} and Y_{i2} is $\widehat{\rho}_{12} = 0.204$, while for Y_{i1} and Y_{i3}, $\widehat{\rho}_{13} = 0.194$, and for Y_{i1} and Y_{i4}, $\widehat{\rho}_{14} = 0.163$. The correlation between sequential observations also appears to increase over time with $\widehat{\rho}_{23} = 0.302$ and $\rho_{34} = 0.351$.

Regression parameter estimates are reported along with the empirical standard error estimates. These parameters are interpreted as follows:

- *(Intercept)* $\widehat{\beta}_0 = 0.1676$. The intercept is an estimate of log odds of a correct response to the nurse item at baseline for the control group. This implies an estimate for the probability of

Table 18.8 GEE Analysis of the Nurse Item from the HIVNET Informed Consent Study[a]

GEE Model Information

Correlation Structure	Unstructured
Subject Effect	id (1123 levels)
Number of Clusters	1123
Correlation Matrix Dimension	4
Maximum Cluster Size	4
Minimum Cluster Size	1

Working Correlation Matrix

	Col1	Col2	Col3	Col4
Row1	1.0000	0.2044	0.1936	0.1625
Row2	0.2044	1.0000	0.3022	0.2755
Row3	0.1936	0.3022	1.0000	0.3511
Row4	0.1625	0.2755	0.3511	1.0000

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	0.1676	0.0652	0.0398	0.2954	2.57	0.0102
Post	−0.3238	0.0704	−0.4618	−0.1857	−4.60	<.0001
ICgroup	−0.1599	0.1643	−0.4819	0.1622	−0.97	0.3306
ICgroup*Post	1.0073	0.2012	0.6128	1.4017	5.01	<.0001

[a]Output from SAS procedure GENMOD.

a correct response at baseline among controls of $\exp(0.1676)/[1+\exp(0.1676)] = 0.5418$, which agrees closely with the observed proportion presented in Table 18.3.

- *Post* $\widehat{\beta}_1 = -0.3238$. The coefficient of Post is an estimate of the log of the odds ratio comparing the odds of a correct response among control subjects after randomization (either month 6, 12, or 18) relative to the odds of a correct response among the control group at baseline. Since the odds ratio estimate is $\exp(-0.3238) = 0.7234 < 1$, the odds of a correct response is lower after baseline. A test for equality of odds comparing postbaseline to baseline yields a p-value $p < 0.001$.
- *ICgroup* $\widehat{\beta}_2 = -0.1599$. The coefficient of ICgroup is an estimate of the log of the odds ratio comparing the odds of a correct response among intervention subjects at baseline relative to the odds of a correct response among the control subjects at baseline. Since the assignment to treatment and control was based on randomization, we expect this odds ratio to be 1.0, and the log odds ratio estimate is not significantly different from 0.0.
- *ICgroup* * *Post* $\widehat{\beta}_3 = 1.0073$. This interaction coefficient measures the difference between the comparison of treatment and control after randomization and the comparison of treatment and control at baseline. Specifically, $(\beta_3+\beta_2)$ represents the log odds ratio comparing the odds of a correct response among intervention subjects postbaseline to the odds of a correct response among control subjects postbaseline. Since β_2 represents the group comparison at baseline, $\beta_3 = (\beta_3 + \beta_2) - \beta_2$, or β_3 measures the difference between the comparison after baseline and the group comparison at baseline. Therefore, the parameter β_3 becomes the primary parameter of interest in this study, as it assesses the change in the treatment/control comparison that is attributable to the intervention. A test of $\beta_3 = 0$ is statistically significant with $p < 0.001$.

GEE is a convenient analysis tool for the informed consent data, as it allows inference regarding the differences between treatment and control groups over time. A standard logistic regression model is adopted and valid standard errors are calculated that account for the within-subject correlation of outcomes.

In Table 18.8 we used a single time variable that was an indicator for the postbaseline visits at six, 12, and 18 months. However, inspection of crude proportions responding correctly suggest that the treatment/control comparison may be decreasing over time. For example, in Table 18.3 we see (treatment, control) proportions of (72.1%, 44.7%) at month 6, (60.1%, 46.3%) and (66.0%, 48.2%) at months 12 and 18. To assess whether the treatment effect appears to be decreasing over time, we fit a second logistic regression model that uses indicator variables for months 6, 12, and 18. Table 18.9 presents GEE estimates using an exchangeable working correlation model. In this model the coefficient of month6*ICgroup contrasts the treatment/control log odds ratio at the six-month visit and at baseline. Similar to our earlier analysis, this difference in time-specific log odds ratios is the primary treatment effect observed at six months. Similarly, the coefficients of month12*ICgroup and month18*ICgroup represent treatment effects at 12 and 18 months. Each of the estimated differences in log odds ratios are significant as indicated by the individual p-values in Table 18.9. In addition, we contrast the observed treatment effect at six months with the treatment effect observed at 12 and 18 months. The difference between the estimated coefficient of month6*ICgroup and month12*ICgroup assesses the change in the treatment effect and is estimated as $1.3232 - 0.7362 = -0.5871$. A test of this contrast yields a p-value of 0.0035, indicating a different treatment effect at 12 months as compared to the treatment effect at 6 months. A similar analysis for the 18-month effect as compared to 6 months is barely statistically significant with $p = 0.041$. Therefore, there is evidence that the effect of the intervention may be changing over time. Once again GEE provides a general tool for evaluating the evolution of mean outcomes over time for different subgroups of subjects.

There are a number of extensions of the GEE approach introduced by Liang and Zeger [1986]. More flexible and tailored dependence models have been proposed for binary data [Lipsitz et al.,

Table 18.9 GEE Analysis of the Nurse Item from the HIVNET Informed Consent Study[a]

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	0.1644	0.0653	0.0364	0.2923	2.52	0.0118
month6	−0.3803	0.0839	−0.5448	−0.2158	−4.53	<.0001
month12	−0.3261	0.0854	−0.4934	−0.1587	−3.82	0.0001
month18	−0.2460	0.0886	−0.4197	−0.0723	−2.78	0.0055
ICgroup	−0.1536	0.1639	−0.4748	0.1676	−0.94	0.3487
month6*ICgroup	1.3232	0.2319	0.8687	1.7777	5.71	<.0001
month12*ICgroup	0.7362	0.2358	0.2739	1.1984	3.12	0.0018
month18*ICgroup	0.9101	0.2273	0.4647	1.3556	4.00	<.0001

Contrast Estimate Results

Label	Estimate	Standard Error	Alpha	Confidence Limits		Chi-Square
Effect at 12 versus 6	−0.5871	0.2014	0.05	−0.9817	−0.1924	8.50
Effect at 18 versus 6	−0.4131	0.2023	0.05	−0.8097	−0.0166	4.17

Contrast Estimate Results

Label	Pr > ChiSq
Effect at 12 versus 6	0.0035
Effect at 18 versus 6	0.0412

[a]Output from SAS procedure GENMOD.

1991; Carey et al., 1993], and extension for multiple survival times has been developed [Wei et al., 1989; Lee et al., 1992].

Summary

- GEE permits regression analysis with correlated continuous, binary, or count data.
- GEE requires specification of a regression model and a working correlation model.
- Two standard error estimates are provided with GEE: a model-based standard error that is valid if the correlation model is specified correctly; and empirical standard errors that are valid even if the correlation model is not correct provided that the data contain a large number of independent clusters.
- Estimation with GEE does not involve a likelihood function; rather, it is based on the solution to regression equations that use models only for the mean and covariance.

18.6 MISSING DATA

One of the major issues associated with the analysis of longitudinal data is missing data, or more specifically, *monotone missing data*, which arise when subjects drop out of the study. It is assumed that once a participant drops out, he or she provides no further outcome information. Missing data can lead to biased estimates of means and/or regression parameters when the probability of missingness is associated with outcomes. In this section we first review a standard taxonomy of missing data mechanisms and then briefly discuss methods that can be used to alleviate bias due to attrition. We also discuss some simple exploratory methods that can help determine whether subjects who complete the longitudinal study appear to differ from those who drop out.

18.6.1 Classification of Missing Data Mechanisms

To discuss factors that are associated with missing data, it is useful to adopt the notation $R_{ij} = 1$ if observation Y_{ij} is observed, and $R_{ij} = 0$ if Y_{ij} is missing. Let $R_i = (R_{i1}, R_{i2}, \ldots, R_{in})$. Monotone missing data imply that if $R_{ij} = 0$, then $R_{ij+k} = 0$ for all $k > 0$. Let Y_i^O denote the subset of the outcomes $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in})$ that are observed, and let Y_i^M denote the missing outcomes. For longitudinal data a missing data classification is based on whether observed or unobserved outcomes are predictive of missing data [Laird, 1988]:

$$\text{Missing completely at random (MCAR)}: \quad P(R_i \mid Y_i^O, Y_i^M, X_i) = P(R_i \mid X_i)$$

$$\text{Missing at random (MAR)}: \quad P(R_i \mid Y_i^O, Y_i^M, X_i) = P(R_i \mid Y_i^O, X_i)$$

$$\text{Nonignorable (NI)}: \quad P(R_i \mid Y_i^O, Y_i^M, X_i) \text{ depends on } Y_i^M$$

In Figure 18.7 an example of monotone missing data is presented. For subject 1, all observations after the 7-month visit are missing. If the reason that these observations are missing is purely unrelated to outcomes (observed or not), the missing data are called *MCAR*. However, if the observed data are predictive of missingness, the missing data are called *MAR*, and the mechanism introduces a form of selection bias. MAR data could occur if an attending physician decides to disenroll any participant who appears to be failing treatment, particularly when the decision is based on the value of past measurements or factors associated with the past outcomes, Y_{ij}. Finally, the unobserved outcomes may be associated with missingness if, for example, subjects who are the most ill refuse to travel to attend their scheduled study visit.

The missing data taxonomy translates directly into implications for potential selection bias. If data are MCAR, both the missing and the observed outcomes are representative of the source population. Therefore, when data are MCAR, standard statistical summaries based on the observed data remain valid. However, if data are MAR or NI, summaries based on the available cases may be biased. Returning to Figure 18.7, if the dropout for patient 1 is indicative of a general process by which those subjects who have a high response value do not return for study, the observed mean for the measured outcomes will not be representative of what would be observed had the entire population been followed. In this example, the mean among available subjects would underestimate the population mean for later months.

Formally, we write $E(Y_{ij} \mid X_i, R_{ij} = 1)$ to denote the expected response conditional on responding, and we write $E(Y_{ij} \mid X_i)$ for the target of inference. If the data are MCAR, then $E(Y_{ij} \mid X_i, R_{ij} = 1) = E(Y_{ij} \mid X_i)$. However, if data are either MAR or NI, then $E(Y_{ij} \mid X_i, R_{ij} = 1) \neq E(Y_{ij} \mid X_i)$, implying that the available data, $R_{ij} = 1$, may not provide valid estimates of population parameters.

In any given application, serious thought needs to be given to the types of processes that lead to missing data. External information can help determine whether missingness mechanisms may be classified as MCAR, MAR, or NI. Unfortunately, since NI missingness implies that unobserved data, Y_i^M, predicts dropout, we cannot empirically test whether data are NI vs. MAR or MCAR. Essentially, one would need the unobserved data to check to see if they are associated with missingness, but these data are missing! The observed data can be used to assess whether the missingness appears to be MAR or MCAR. First, the dropout time can be considered a discrete-time "survival" outcome, and methods introduced in Chapter 16 can be used to assess whether past outcomes $Y_{ij-1}, Y_{ij-2}, \ldots$ are predictive of dropout, $R_{ij} = 0$. Second, each subject will have a dropout time, or equivalently, a "last measurement" time, with those completing the study having the final assessment time as their time of last measurement. The longitudinal data can be stratified according to the dropout time. For example, the mean at baseline can be calculated separately for those subjects that dropout at the first visit, second visit, through those that complete the study. Similarly, the mean response at the first follow-up visit can be computed for all subjects who have data for that visit. Such analyses can be used to determine whether the outcomes for the dropout subjects appear to be different from those

of the "completers." Naturally, subjects who are lost can only be compared to others at the visit times prior to their dropout. These exploratory analyses are complementary: The first approach assesses whether outcomes predict dropout, and the second approach evaluates whether the dropout time predicts the outcomes. An example of such modeling can be found in Zhou and Castelluccio [2004].

18.6.2 Approaches to Analysis with Missing Data

There are several statistical approaches that attempt to alleviate bias due to missing data. General methods include:

1. *Data imputation*. See Little and Rubin [1987], Schafer [1997], or Koepsell and Weiss [2003] for more information on imputation methods. Imputation refers to "filling in" missing data. Proper methods of imputation use multiple imputation to account for uncertainty in the missing data. Imputation methods require that a model be adopted that links the missing data to the observed data.

2. *Data modeling*. In this method the missing data process and the longitudinal data are both modeled using maximum likelihood for estimation. Use of a linear mixed model estimated with maximum likelihood is one example of this approach. However, to correct validly for MAR missingness, the mean and the covariance must be specified correctly. See Verbeke and Molenberghs [2000] for more details.

3. *Data weighting*. Nonresponse methods with available data are used to weight the observed data to account for the missing data. Use of inverse probability weighting or nonresponse weighting can be applied to general statistical summaries and has been proposed to allow for use of GEE in MAR situations. See Robins et al. [1995] for the statistical theory and Preisser et al. [2002] for a simulation study of the performance of weighted GEE methods.

However, it is important to note that these methods are designed to address data that are assumed to be MAR rather than the more serious nonignorable (NI) missing data. Nonignorable missing data can lead to bias, which cannot be corrected simply through modeling and estimation of the dropout model and/or the response model since unidentifiable parameters that link the probability of missingness to the unobserved data are needed. Therefore, reliance on statistical methods to correct for bias due to attrition either requires an untestable assumption that the data are MAR or requires some form of sensitivity analysis to characterize plausible estimates based on various missingness assumptions. See Diggle et al. [2002, Chap. 13] for discussion and illustration.

Example 18.1. (*continued*) In the HIVNET Informed Consent Study, there was substantial missing data due to attrition. In Tables 18.2 and 18.3 we see a decreasing number of subjects over time. In the control group there are 946 subjects with baseline data and only 782 with 18-month data. Is the knowledge score for subjects who complete the study different from the score for those who dropout? Figure 18.11 shows the mean response over time stratified by dropout time. For example, among subjects that dropout at the 12-month visit, their mean knowledge score at baseline and 6 months is plotted. This plot suggests that subjects who complete only the baseline interview have a lower mean baseline knowledge score than that of all other subjects. In addition, for subjects who complete the study, the average knowledge score at six and 12 months appears greater than the mean knowledge score among subjects who do not complete the 18-month visit. Thus, Figure 18.11 suggests that the completers and the dropout subjects differ with respect to their knowledge scores. Any analysis that does not account for differential dropout is susceptible to selection bias.

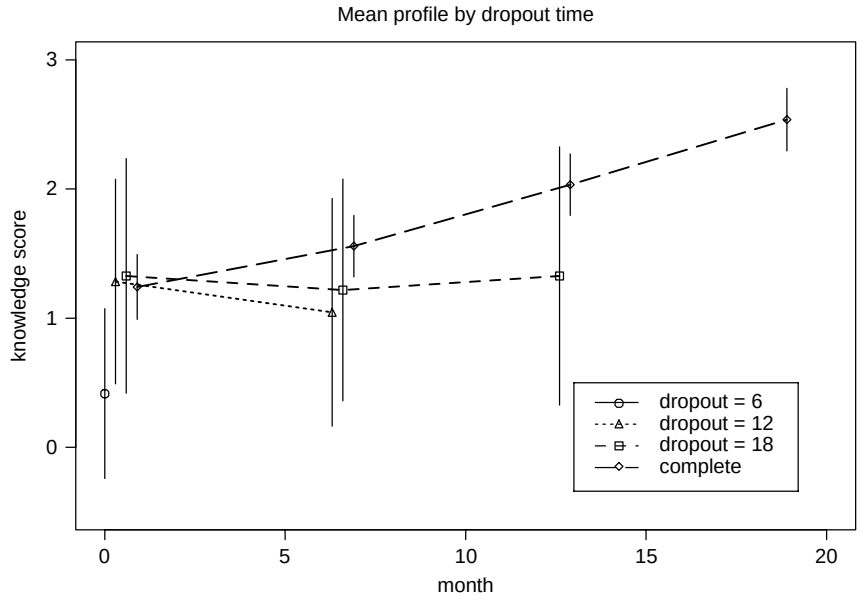

Figure 18.11 Patterns of mean knowledge score by dropout time for the control group. HIVNET informed consent substudy.

18.7 SUMMARY

Longitudinal data provide unique opportunities for inference regarding the effect of an intervention or an exposure. Changes in exposure conditions can be correlated with changes in outcome conditions. However, analysis of longitudinal data requires methods that account for the within-subject correlation of repeated measures. Texts by Diggle et al. [2002], Verbeke and Molenberghs [2000], Brown and Prescott [1999], and Crowder and Hand [1990] provide comprehensive discussions of statistical methods for the analysis of longitudinal data. There are a number of additional issues that warrant attention but are beyond the scope of this book.

NOTES

18.1 Nonlinear Mixed Models

We have introduced linear mixed models and GEE. However, mixed models have also been extended to logistic regression and other nonlinear model settings. See Diggle et al. [2002, Chap. 8 and 11] for illustrations.

18.2 Models for Survival and Repeated Measurements

In many longitudinal studies information on both repeated measurements and on ultimate time until death or key clinical endpoint is collected. Methods have been developed to analyze such data jointly. See Hogan and Laird [1997a, b] for an overview of approaches for the joint analysis of survival and repeated measures.

18.3 Models for Time-Dependent Covariates

In designed experiments the exposures X_{ij} may be controlled by the investigator. However, in many observational studies, exposures or treatments that are selected over time may be related to

past health outcomes. For example, subjects with low values of CD4 may be more likely to be exposed to a therapeutic agent. Analysis of such serial data to assess the effect of the intervention is complicated by the feedback between outcome and exposure. Robins [1986] and Robins et al. [1999] have identified proper causal targets of inference and methods for estimation in settings where time-varying covariates are both causes and effects. See Diggle et al. [2002, Chap. 12].

PROBLEMS

18.1 This exercise considers the interplay between the covariate distribution and the correlation. For each of the following scenarios, assume that there are a total of N pairs of observations, (Y_{i1}, Y_{i2}), with covariates (X_{i1}, X_{i2}). Assume that the covariate is binary: $X_{ij} = 0$ or $X_{ij} = 1$, denoting control and treatment exposures. Let $\overline{Y}_1$ denote the mean of all observations where $X_{ij} = 1$, and let $\overline{Y}_0$ denote the mean of all observations where $X_{ij} = 0$. Assume a constant variance $\sigma^2 = \text{var}(Y_{ij} \mid X_{ij})$ and a correlation $\rho = \text{corr}(Y_{i1}, Y_{i2})$.

- **(a)** Assume that half of the subjects are assigned to control for both visits, $(X_{i1}, X_{i2}) = (0, 0)$, and half of the subjects are assigned to intervention for both visits, $(X_{i1}, X_{i2}) = (1, 1)$. What is the variance of the estimated mean difference, $\widehat{\Delta} = (\overline{Y}_1 - \overline{Y}_0)$?
- **(b)** Assume that subjects change their treatment over time with half of the subjects are assigned to control and then treatment, $(X_{i1}, X_{i2}) = (0, 1)$, and half of the subjects assigned to treatment and then control, $(X_{i1}, X_{i2}) = (1, 0)$. This design is referred to as a *crossover study*. What is the variance of the estimated mean difference $\widehat{\Delta} = (\overline{Y}_1 - \overline{Y}_0)$?
- **(c)** Comment on the advantages and disadvantages of these two study designs.

18.2 Consider a study with a single prerandomization measurement, Y_{i0}, and a single postrandomization measurement, Y_{i1}. For any constant a we can define the average contrast, $\overline{D}(a) = \text{mean}[d_i(a)]$, where $d_i(a) = Y_{i1} - aY_{i0}$. Let $\overline{D}_0(a)$ denote the mean for the control group, and let $\overline{D}_1(a)$ denote the mean for the intervention group. Assume that $\sigma^2 = \text{var}(Y_{ij})$ for $j = 0, 1$, and let $\rho = \text{corr}(Y_{i0}, Y_{i1})$. We assume that the subjects are randomized to treatment and control after randomization at baseline. Therefore, the following table illustrates the mean response as a function of treatment and time:

	Control	**Intervention**
Baseline	μ_0	μ_0
Follow-up	μ_1	$\mu_1 + \Delta$

- **(a)** Show that the expected value of $\widehat{\Delta}(a) = \overline{D}_1(a) - \overline{D}_0(a)$ equals Δ for any choice of a.
- **(b)** When $a = 0$, we effectively do not use the baseline value, and $\widehat{\Delta}(0)$ is the difference of means at follow-up. What is the variance of $\widehat{\Delta}(0)$?
- **(c)** When $a = 1$, we effectively analyze the change in outcomes since $d_i(1) = Y_{i1} - Y_{i0}$. What is the variance of $\widehat{\Delta}(1)$?
- **(d)** What value of a leads to the smallest variance for $\widehat{\Delta}(a)$?

18.3 Use the data from the Web page to perform GEE analysis of the HIVNET Informed Consent Substudy "safety" item.

18.4 For the random intercepts and slopes model given in Table 18.6, the proportion of total variation that is attributable to within-subject variation is not constant over time. Compute estimates of the proportion of total variation at 0, 12, 24, and 36 months that is attributable to within-subject variation, ϵ_{ij}, as opposed to between subject variation, $b_{i,0} + b_{i,1} \cdot \text{month}$.

18.5 For the HIVNET Informed Consent Substudy data, create pairs of plots:

(a) Plot month 12 vs. month 6 knowledge score. Add a pair of lines that show the ordinary least squares estimate for the intervention and the control group.

(b) Plot month 18 vs. month 12 knowledge score. Add a pair of lines that shows the ordinary least squares estimate for the intervention and the control group.

(c) Do these plots suggest that there are additional differences between the intervention and control groups that is not captured by the difference that manifests at the six-month visit?

18.6 For the NURSE and SAFETY items from the HIVNET Informed Consent Substudy, evaluate the transition from incorrect to correct, and from correct to correct again, for the times (six-month $\rightarrow$ 12-month visit) and (12-month $\rightarrow$ 18-month visit). Is there evidence that the intervention and control groups differ in terms of the "correction" and "maintenance" of knowledge at the later times?

REFERENCES

Brown, H., and Prescott, R. [1999]. *Applied Mixed Models in Medicine*. Wiley, New York.

Carlin, B. P., and Louis, T. A. [1996]. *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall, London.

Coletti, A. S., Heagerty, P. J., Sheon, A. R., Gross, M., Koblin, B. A., Metzger, D. S., and Seage G. R. [2003]. Randomized, controlled evaluation of a prototype informed consent process for HIV vaccine efficacy trials. *Journal of Acquired Immune Deficiency Syndrome*, **32**: 161–169.

Carey, V., Zeger, S. L., and Diggle, P. [1993]. Modelling multivariate binary date with alternating logistic regressions. *Biometrika*, **80**: 517–526.

Crowder, M. J., and Hand, D. J. [1990]. *Analysis of Repeated Measures*. Chapman & Hall, New York.

Diggle, P. J., Heagerty, P. J., Liang, K.-Y., and Zeger, S. L. [2002]. *Analysis of Longitudinal Data*. Oxford University Press, Oxford.

Donner, A., and Klar, N. [1994]. Cluster randomization trials in epidemiology: theory and application. *Journal of Statistical Planning and Inference*, **42**: 37–56.

Donner, A., and Klar, N. [1997]. Statistical considerations in the design and analysis of community intervention trials. *Journal of Clinical Epidemiology*, **49**: 435–439.

Frison, L. J., and Pocock, S. J. [1992]. Repeated measures in clinical trials: analysis using summary statistics and its implication for design. *Statistics in Medicine*, **11**: 1685–1704.

Frison, L. J., and Pocock, S. J. [1997]. Linearly divergent treatment effects in clinical trials with repeated measures: efficient analysis using summary statistics. *Statistics in Medicine*, **16**: 2855–2872.

Hanley, J. A., Negassa, A., deB. Edwardes, M. D., and Forrester, J. E. [2003]. Statistical analysis of correlated data using generalized estimating equations: an orientation. *American Journal of Epidemiology*, **157**: 364–375.

Hogan, J. W., and Laird, N. M. [1997a]. Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine*, **16**: 239–257.

Hogan, J. W., and Laird, N. M. [1997b]. Model-based approaches to analysing incomplete longitudinal and failure time data. *Statistics in Medicine*, **16**: 259–272.

Kaslow, R. A., Ostrow, D. G., Detels, R., et al. [1987]. The Multicenter AIDS Cohort Study: rationale, organization and selected characteristics of the participants. *American Journal of Epidemiology*, **126**: 310–318.

Koepsell, T. D., and Weiss, N. S. [2003]. *Epidemiological Methods: Studying the Occurrence of Illness*. Oxford University Press, New York.

Koepsell, T. D., Martin, D. C., Diehr, P. H., Psaty, B. M., Wagner, E. H., Perrin, E. B., and Cheadle, A. [1991]. Data analysis and sample size issues in evaluations of community-based health promotion and disease prevention programs: a mixed model analysis of variance approach. *American Journal of Epidemiology*, **44**: 701–713.

Laird, N. M. [1988]. Missing data in longitudinal studies. *Statistics in Medicine*, **7**: 305–315.

Laird, N. M., and Ware, J. H. [1982]. Random-effects models for longitudinal data. *Biometrics*, **38**: 963–974.

Lebowitz, M. D. [1996]. Age, period, and cohort effects. *American Journal of Respiratory Critical Care Medicine*, **154**: S273–S277.

Lee, E. W., Wei, L. J., and Amato, D. A. [1992]. Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In *Survival Analysis: State of the Art*, J. P. Klein and P. K. Joel (eds.). Kluwer Academic Publishers, Dordrecht.

Liang, K.-Y., and Zeger, S. L. [1986]. Longitudinal data analysis using generalised linear models. *Biometrika*, **73**: 13–22.

Liang, K.-Y., and Zeger, S. L. [1993]. Regression analysis for correlated data. *Annual Review of Public Health*, **14**: 43–68.

Lipsitz, S., Laird, N., and Harrington, D. [1991]. Generalized estimating equations for correlated binary data: using odds ratios as a measure of association. *Biometrika*, **78**: 153–160.

Little, R. J. A., and Rubin, D. B. [2002]. *Statistical Analysis with Missing Data*, 2nd ed. Wiley, New York.

McCullagh, P., and Nelder, J. A. [1989]. *Generalized Linear Models, 2nd ed.* Chapman & Hall, New York.

Preisser, J. S., Lohman, K. K., and Rathouz, P. J. [2002]. Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random. *Statistics in Medicine*, **21**: 3035–3054.

Robins, J. M. [1986]. A new approach to causal inference in mortality studies with sustained exposure periods: application to control of the healthy worker survivor effect. *Mathematical Modelling*, **7**: 1393–1512.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. [1995]. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, **90**: 106–121.

Robins, J. M., Greenland, S., and Hu, F.-C. [1999]. Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome (with discussion). *Journal of the American Statistical Association*, **94**: 687–712.

Samet, J. M., Dominici, F., Curriero, F. C., Coursac, I., and Zeger, S. L. [2000]. Fine particulate air pollution and mortality in 20 US cities. *New England Journal of Medicine*, **343**(24): 1798–1799.

Schafer, J. L. [1997]. *Analysis of Incomplete Multivariate Data*. Chapman & Hall, New York.

Stram, D. O., and Lee, J. W. [1994]. Variance component testing in the longitudinal mixed model. *Biometrics*, **50**: 1171–1177.

The Childhood Asthma Management Program Research Group [2002]. Long-term effects of budesonide or nedocromil in children with asthma. *New England Journal of Medicine*, **343**(15): 1054–1063.

Verbeke, G., and Molenberghs, G. [2000]. *Linear Mixed Models for Longitudinal Data*. Springer-Verlag, New York.

Wei, L. J., Lin, D., and Weissfeld, L. [1989]. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, **84**: 1065–1073.

Weiss, S. T., and Ware, J. H. [1996]. Overview of issues in the longitudinal analysis of respiratory data. *American Journal of Respiratory Critical Care Medicine*, **154**: S208–S211.

Yu, O., Sheppard, L., Lumley, T., Koenig, J., and Shapiro, G. [2000]. Effects of ambient air pollution on symptoms of asthma in Seattle-area children enrolled in the CAMP study. *Environmental Health Perspectives*, **108**: 1209–1214.

Zhou, X.-H., and Castelluccio, P. [2004]. Adjusting for non-ignorable verification bias in clinical studies for Alzheimer's disease. *Statistics in Medicine*. **23**: 221–230.