Chapter 3

Naive Cross-Sectional Analysis of Longitudinal Data

In this chapter, we review some simple cross-sectional regression approaches and consider the implications of using these directly on longitudinal observations by ignoring the fact that observations are repeated on the same individual. That is, we naively treat all N observations as if they were independently sampled. We recognize that this independence assumption is extremely unlikely to hold because repeated observations of the outcome Y on the same individual are often more likely to be somewhat similar, or correlated in general, as compared to observations on different individuals. In addition, we also discuss that by ignoring the longitudinal structure of our data, we can lose the opportunity to disentangle potentially causal effects of interest (so-called *longitudinal effects*) from potentially spurious ones (*cross-sectional effects*) based on how the data was sampled. The point of this chapter is twofold: first, to demonstrate how to take advantage of longitudinal data in estimating potentially causal associations, and second, to highlight that ignoring correlation of repeated outcomes on the same unit can lead to reasonable estimates of regression coefficients, but leads to biased inference from biased standard errors. We discuss these themes mainly in the context of linear regression models.

3.1 Standard Estimators That Ignore Longitudinal Structure: Pitfalls

For now, we will assume the parameters of interest are the coefficients in a simple linear model. A simplistic approach to longitudinal data in this context is to merely ignore the longitudinal structure and apply standard regression estimators to the full data, ignoring entirely the fact that these observations reflect grouping when they are repeated observations on the same individual. For example, if the outcome is continuous, we might apply ordinary least squares to obtain regression estimators using the raw data $\{(Y_{ij}, X_{ij}) : i = 1, \ldots, m; j = 1, \ldots, n_i\}$. This is predicated on extending the basic linear regression model for cross-sectional data

$$E(Y_i|X_i = x_i) = b_0 + b_1 x_i$$

naively to the longitudinal version

$$E(Y_{ij} \mid X_{ij} = x_{ij}) = b_0 + b_1 x_{ij}. \tag{3.1}$$

What are the properties of naive (cross-sectional) estimators, ignoring the repeated measures (longitudinal) structure? We consider these questions in the simplest setting, the use of ordinary least squares to estimate the regression coefficients b_0 and b_1 in (3.1). For independent observations where the goal is regression of an outcome on covariates, ordinary least squares (OLS) is the typical choice. OLS focuses on the deviations from observed points and their expected values, given the associated risk factors, based on a model such as (3.1). The total amount of deviation is summarized by

$$RSS = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left((y_{ij} - E(Y_{ij}|X_{ij} = x_{ij})\right)^2 = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(y_{ij} - (b_0 + b_1 x_{ij})\right)^2. \tag{3.2}$$

As in standard linear regression, this quantity is known as the residual sum of squares (RSS). The least squares technique chooses the value of the parameters that minimize this residual sum of squares; that is, OLS provides the least squares estimates $\hat{b}_0$ and $\hat{b}_1$ to minimize the quantity (3.2). The solutions to this minimization problem can be found either by algebra or by using differential calculus and are given, in matrix form, by

$$\hat{\mathbf{b}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}. \tag{3.3}$$

Note that the residual sums of squares, (3.2), treats each observation "equally", making no distinction between longitudinal observations on the same individual from those on

different individuals. Does the fact that repeated observations on the same individual are likely to be correlated make $\hat{\mathbf{b}}$, as given by (3.3), an erroneous estimator in some way? Is it biased systematically so as to underestimate or overestimate the true population regression parameters of (3.1)? To investigate this issue, we take a closer look at the expectation of the random variable $\hat{\mathbf{b}}$ to see how it is affected by the correlation or covariance structure of the Y_{ij}'s. Note that

$$\begin{aligned} E(\hat{\mathbf{b}}|\mathbf{X}) &= E\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}|\mathbf{X}\right] \\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T E(\mathbf{Y}) \\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\mathbf{b} \\ &= I\mathbf{b} \\ &= \mathbf{b}, \end{aligned}$$

using 1) that we condition on the observed values of the explanatory variables, x_{ij}—that is, hold them constant—2) and properties of expectations. The set of equations above establishes a key point, that $E(\hat{\mathbf{b}}|\mathbf{X})$ does not depend at all on the covariance or interrelationship of the Y_{ij}'s, with the result that the ordinary least squares estimate $\hat{\mathbf{b}}$ is just as valid an estimator of $\mathbf{b}$, in the model (3.1), with longitudinal data as it is for cross-sectional observations, at least in terms of bias. A similar result holds, at least in large samples, for maximum likelihood estimators of regression parameters in other forms of regression models such as Poisson or logistic regression. So far so good for the naive approach. But remember that this logic is predicated on the validity of the model (3.1) and pays no attention to the variability—and estimates thereof—of the estimator $\hat{\mathbf{b}}$.

3.2 Modeling Changes in Outcomes vs. Changes in Explanatory Variables

Note that the model (3.1) makes no distinction between changes in the explanatory variable X that occur over longitudinal observations within the same individual and changes of X across different individuals. For instance, consider the example of data collected longitudinally from initiation of HAART among HIV+ patients (Section 1.3.1), and examining CD4 count versus time since beginning of the study. In this data set, the time of measurements of subjects are not chosen by the researcher, but are passively collected as the patient comes into clinic. Something that is certainly true here and perhaps in many other contexts is that times chosen by the patient to come in (or scheduled by the physician) can be related to patient characteristics that are also related to the outcomes

of interest, in our case CD4 count. For instance, the time of the first measurement after initiation of HAART can be related to the progression of HIV disease. Thus, we have the possibility of confounding in the relationship (e.g., linear trend) of CD4 and time if, for instance, only the first measurement is used (or equivalently, as we demonstrate below, if a naive cross-sectional model is fit). As a concrete example, assume that the rate of CD4 increase post-HAART is the same for every individual, but the CD4 at baseline (beginning of HAART) is different among individuals. Also, assume those with higher baseline CD4 counts tend to show up at clinic at earlier dates - see Figure 3.1A. The figure shows an extreme example where if one fits a naive cross-sectional model, for instance a model such as 3.1 to data as represented in Figure 3.1B, then the resulting estimate of the trend will underestimate the true, longitudinal trend within individuals. However, this confounding by the association of visit time with baseline characteristics of the individual is simply removed, in this idealized case, by looking at the change in CD4 count versus change in time (Figure 3.1C), that is by recognizing the longitudinal structure of the data: changes of CD4 count versus change in time. This notion of looking at changes in explanatory variables versus changes in outcomes can provide longitudinal studies an edge over cross-sectional studies in circumstances where measurement time is related to prognostic factors. However, to take advantage of the longitudinal structure of the data one must properly parameterize the regression models to separate out the "longitudinal" effects from those due to a type of selection bias or confounding.

Of course, with cross-sectional observations we never see changes in an individual's risk factor since all explanatory variables are observed but once per individual. However, one of the great advantages of a longitudinal study is that we often can compare response to changes in an individual's risk level in addition to comparing individuals who differ in their levels of these same variables. Of course, this is only possible for time-dependent covariates. For such variables, X, which always includes the "time" of the longitudinal observation, we can model distinct effects for (i) longitudinal changes in X, and (ii) changes in X across individuals. To achieve this goal, we can use a simple extension of the model (3.1) given by

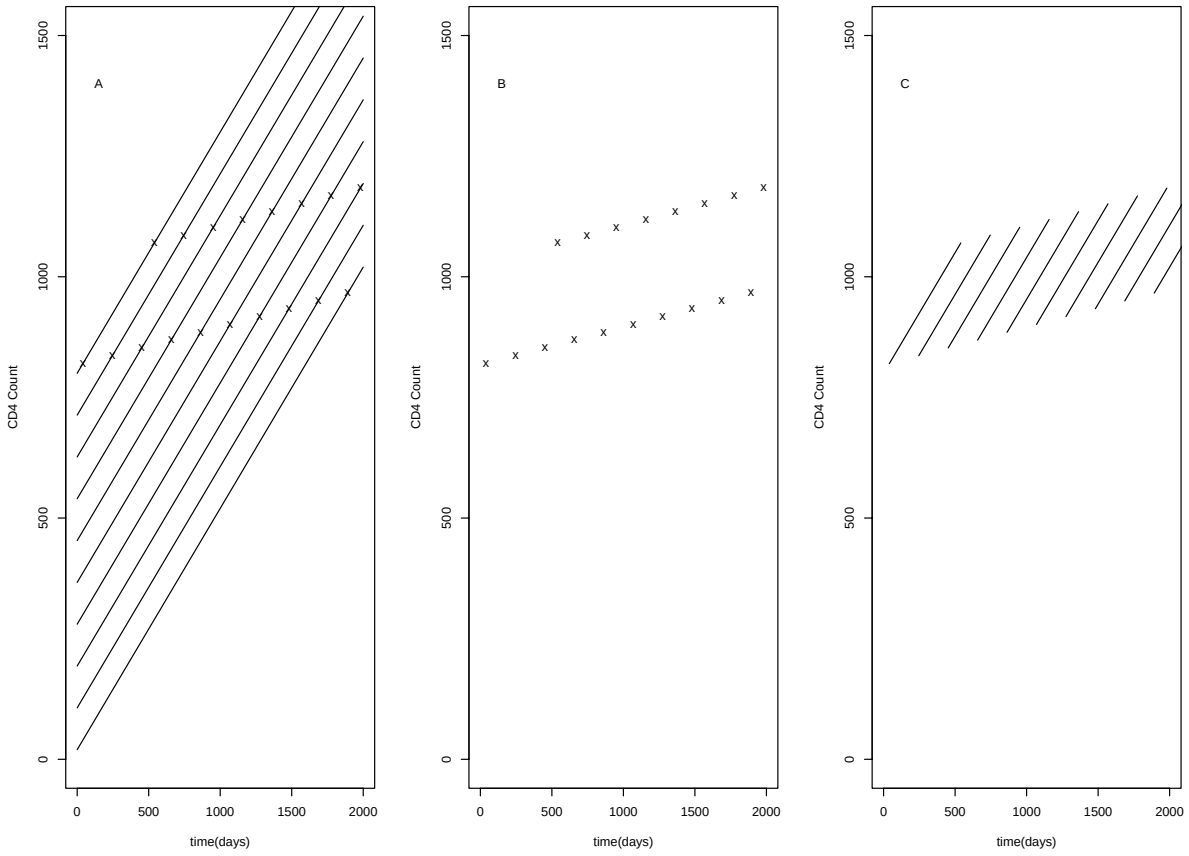
$$E(Y_{ij} \mid X_{i1} = x_{i1}, X_{ij} = x_{ij}) = b_0 + b_{CS}x_{i1} + b_L(x_{ij} - x_{i1}), \tag{3.4}$$

which can be equivalently written as

$$E(Y_{ij} \mid X_{i1} = x_{i1}, X_{ij} = x_{ij}) = b_0 + (b_{CS} - b_L)x_{i1} + b_L x_{ij}. \tag{3.5}$$

In the version of this model, (3.4), we see that the explanatory variable appears in two forms, the first as a time-independent covariate, X_{i1}, which measures its initial value, and then as a time-dependent covariate, $X_{ij} - X_{i1}$, which captures the *change* in X_i over the longitudinal measurements. There are now two regression coefficients of interest: b_{CS} which quantifies the cross-sectional effect of variation in X_{i1}, and b_L for the effect of longitudinal

Figure 3.1: CD4 count vs. time. A) true trends with sampling points for 10 hypothetical individuals, B) treating the data as cross-sectional data, C) looking at change in CD4 vs. change in time (longitudinal effect)

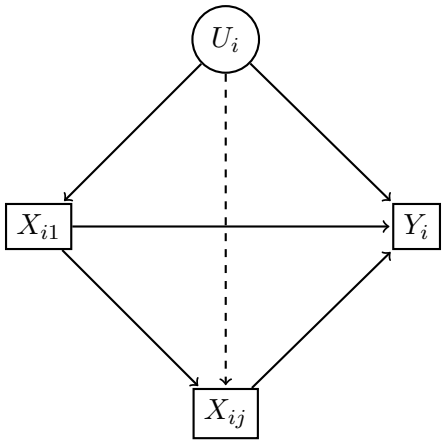


variation in X_i. The cross-sectional coefficient is interpreted as the change in the mean of the response variable Y associated with comparing two *different* individuals whose initial values of X differ by one unit on the scale of X. On the other hand, b_L captures the change in the mean of the response Y associated with a change of X *over time*, or equivalently, when comparing two different longitudinal observations on the *same* individual. With reference to Figure 3.1, the slope through the baseline points in B is b_{CS}, whereas the slopes in C correspond to b_L.

When is it that this simple approach, which one can think of as adjusting for the baseline value of our explanatory variable of interest, X_{i1}, helps to retrieve the true longitudinal association (how changes in X are associated with changes in the mean of Y within a person)? There are several related ways to think about this, but perhaps the easiest is graphical. Consider the association of X_{ij} with Y_{ij} (the explanatory variable of interest and outcome for person i, time j) in the presence of both the baseline measurement, X_{i1} and a possible time-fixed unmeasured confounder, U_i. Figure 3.2 has a graph showing the possible connections between these variables. Consider first that the dashed line (from U_i to X_{ij}) is not there, so the only association of this unmeasured confounder and the current value of the explanatory variable is through the baseline measurement. As a concrete example, in our HIV data set, there can be unmeasured confounders that are related both to baseline viral load (X_{i1}) and the CD4 count at time j (Y_{ij}), but have no independent relationship to viral load at time j (X_{ij}). For instance, if U_i is socio-economic status, it might be argued that once I know the baseline viral load, baseline factors such as socio-economic status provide no additional information about current viral load. In that case, given the rules for adjusting for confounding in causal graphs (see Pearl, 2000 and Jewell, 2003), adjusting for baseline viral load is sufficient to adjust for confounding in Figure 3.2, and thus using model (3.4) successfully adjusts for confounding. However, if this same arrow (from U_i to X_{ij}) is solid (exists), the change in X ($X_{ij} - X_{i1}$) is also confounded by U_i and so using model (3.4) does not remove the impacts of unmeasured baseline confounders (the U_i). Thus one can not retrieve the true association by simple adjustment by the baseline value of X. However, it still seems sensible that the baseline value of the covariate of interest will serve as a reasonable proxy for the unmeasured baseline confounders.

As discussed in the previous paragraph, there is no reason why these two regression effects of X, the cross-sectional (b_{CS}) and the longitudinal (b_L), should be the same. In fact, if b_L is the coefficient of interest, and one fits a naive model (3.1), then the estimate, $\hat{\mathbf{b}}$ will only be unbiased if $b_{CS} = b_L$. This can happen, for example in the CD4 example, if the time of clinic visits is unrelated in any way to the baseline CD4 count. This section has ignored the other part of statistical estimation—that is the estimates of uncertainty of the coefficient estimates (the standard errors). In the next section, we show that, ignoring the potential bias in the coefficients from estimating longitudinal effects from naive cross-

Figure 3.2: A graph showing the relationship of time varying covariates on longitudinal outcomes - dashed line means possible link.



sectional data, this method in general also leads to biased estimation of the variance of these coefficients. Thus, there are two sources of potential bias from naive approaches - estimation and inference. We discussed approaches for unbiased estimation in this section, however fitting a model such as (3.4) will not automatically provide proper standard errors in the presence of correlated longitudinal measurements. That is a separate issue discussed in the next section.

3.3 Incorrect Estimation of Variability with Cross-Sectional Analysis

We now return to consideration of the variability of the ordinary least squares estimator, $\hat{\mathbf{b}}$. Using the results of Chapter 1.8, one can show that

$$\begin{aligned} Var(\hat{\mathbf{b}}) &= Var\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}\right] \\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T Var(\mathbf{Y})\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} \\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}, \end{aligned} \tag{3.6}$$

where $\mathbf{V}$ is just the variance-covariance matrix of $\mathbf{Y}$. Our first reaction to (3.6) is that if the variance-covariance matrix of $\mathbf{Y}$ is known, or at least can be estimated, then we can immediately estimate the variance of $\hat{\mathbf{b}}$. Of more immediate relevance to our present

discussion is the fact that, unlike the mean $E(\hat{\mathbf{b}})$, the variance, $Var(\hat{\mathbf{b}})$ depends on $\mathbf{V}$ in a crucial way. This has an immediate effect on the property of the ordinary least squares estimate $\hat{\mathbf{b}}$, and particularly its estimated variance, when we use formulae that assume naively that the longitudinal observations in $\mathbf{Y}_i$ are independent, that is that $\mathbf{V}_i = \sigma^2\mathbf{I}$, where $\mathbf{I}$ is the $n_i \times n_i$ identity matrix. This is, of course, a supposition that underlies most cross-sectional regression models. If we take $\mathbf{V}_i = \sigma^2\mathbf{I}$, then (3.6) shows that

$$\begin{aligned} Var(\hat{\mathbf{b}}) &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\sigma^2\mathbf{I}\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} \\ &= \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}, \end{aligned} \tag{3.7}$$

so that estimation of $Var(\hat{\mathbf{b}})$ depends solely on estimation of the constant σ^2. However, if the assumption $\mathbf{V}_i = \sigma^2\mathbf{I}$ is incorrect then the difference between (3.7) and (3.6) introduces bias into estimation of the variability of $\hat{\mathbf{b}}$, translating into incorrect inference (e.g. statistical tests and confidence intervals).

As an aside, we note here that you still need to estimate $\mathbf{V}$ to use formulae like (3.6), or even the simple (3.7), to produce an estimate of $Var(\hat{\mathbf{b}})$. In standard cross-sectional regression methods, where (3.7) holds, an estimate is easily constructed by estimating σ^2. We do not observe the residuals but with our estimate of the regression coefficients in hand, say $\hat{\mathbf{b}}$, we can in turn estimate e_{ij} by r_{ij}:

$$\begin{aligned} r_{ij} &= y_{ij} - \left(\hat{b}_0 + \hat{b}_1 x_{ij1} + \hat{b}_2 x_{ij2} + \cdots + \hat{b}_p x_{ijp}\right) \\ &= y_{ij} - \mathbf{x}_{ij}^T\hat{\mathbf{b}}. \end{aligned}$$

To obtain an estimate of $\sigma^2 = Var(e_{ij})$, we then take the average of the squared residuals

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(r_{ij}\right)^2, \tag{3.8}$$

using the fact that the mean $E(e_{ij}) = 0$, reflected by $\frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i} r_{ij} = 0$ when ordinary least squares is used to give $\hat{\mathbf{b}}$ in (3.3). The estimate $\hat{\sigma}^2$ can then be plugged into the assumption $\mathbf{V}_i = \sigma^2 I$ to give $\hat{\mathbf{V}}_i = \hat{\sigma}^2\mathbf{I}$ for use in (3.7). That is,

$$\widehat{Var}(\hat{\mathbf{b}}) = \hat{\sigma}^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}. \tag{3.9}$$

The diagonal elements of the matrix (3.9) then provide the estimated variances of each of the estimated regression coefficients, $\hat{b}_0, \hat{b}_1, \ldots, \hat{b}_p$. The square roots of these variances yield the standard errors necessary for confidence interval construction for the regression coefficients.

In general, the primary point here is that the assumption $\mathbf{V}_i = \sigma^2 I$ is extremely unlikely to hold with longitudinal observations making the variance estimate (3.9) biased for use in inference about $\mathbf{b}$. This does not preclude the use of the ordinary least squares estimator, subject to careful choice of an appropriate longitudinal model as discussed above. But, obtaining an appropriate estimate of $Var(\hat{\mathbf{b}})$ from (3.6) will require an estimate of $\mathbf{V}_i$, using more realistic assumptions about the correlation structure of the Y_{ij}'s, that is the form of $\mathbf{V}_i$. The method using estimated residuals, briefly described above for $\hat{\sigma}^2$ in ordinary least squares, is the precursor to the general methods we describe in detail in Chapters 5 and 6.

3.3.1 Simple Example of Incorrect Variance Assessment if Correlation of Longitudinal Observations is Ignored

We illustrate the issues surrounding variance estimation by considering a very simple example arising from the HAART data of Chapter 1.3.1. In a preliminary analysis of this data we may wish to estimate the *mean viral load amongst patients in the early part of the study*, using the first two longitudinal observations available. Ignoring covariate information, including the timing of the measurement, consider the following model

$$Y_{ij} = b_0 + b_{0i} + e_{ij}, \tag{3.10}$$

with $i = 1, \ldots, m$ and $j = 1, 2$ where the terms, b_{0i}, are realizations of an unobserved random variable, varying from individual to individual. This random variable, b_{0i} is assumed to have mean 0 ($E(b_{0i}) = 0$) and to be independent of the error terms e_{ij}. The coefficients $b_0 + b_{0i}$ can be thought of as *random intercepts*; in this case, we have no slope parameters to worry about. The model (3.10) is a simple example of a *mixed effects model* which we consider more closely in Chapters 5 and 7. As always, we assume that $E(e_{ij}) = 0$, so that $E(Y_{ij}) = b_0$, interpreted as the overall population mean viral load, averaging over all measurements and the entire patient population. Conditional on examining only the i^{th} patient, that is given b_{0i}, $E(e_{ij})$ is still 0, so that for this patient $E(Y_{ij}|b_{0i}) = b_0 + b_{0i}$, the mean viral load for this individual. Conceptually, then, the term b_{0i} can be interpreted as the amount by which the i^{th} patient's longitudinal viral load varies from the overall population mean b_0.

The variance of the error terms e_{ij}, denoted by ${\sigma_e}^2$, is known as the *within* person variance since it measures the variability of viral load from measurement to measurement *for the same person*. On the other hand, the variance of the random person-means, or intercepts, b_{0i}, is denoted by ${\sigma_b}^2$, and represents the *between person* variance, the variation of the individual viral load means across individuals. The total variability of a person's

viral load measured at a specific time (Y_{ij}) is simply $Var(Y_{ij}) = {\sigma_b}^2 + {\sigma_e}^2$, the sum of these two components of variation in our measurements. Of course, if ${\sigma_b}^2 = 0$, each individual has the same mean viral load longitudinally, and the only source of variability arises from the error terms e_{ij}.

Non-zero variation in the individual terms b_{0i} induces correlation among the longitudinal observations for a fixed individual because they all share a common mean, $b_0 + b_{0i}$. In fact, from first principles (see Problem 3.1), we can see that the correlation between longitudinal measurements within an individual is given by

$$\rho = \frac{{\sigma_b}^2}{{\sigma_b}^2 + {\sigma_e}^2}. \tag{3.11}$$

If our interest is in the model (3.10), then our focus is on estimation of the population mean b_0 and not on the random individual effects b_{0i}. The ordinary least squares estimate of b_0 is just the sample average of all the observations

$$\bar{Y} = \frac{1}{2m} \sum_{i=1}^{m} \sum_{j=1}^{2} Y_{ij} \equiv \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{2} Y_{ij}. \tag{3.12}$$

If we naively assume that *all* observations are independent—that is there is no correlation between Y_{i1} and Y_{i2}, or ${\sigma_b}^2 = 0$—then estimation of the variance of this simple average relies on the usual formula

$$\widehat{Var}(\bar{Y}) = \frac{1}{2m} \frac{1}{2m-1} \sum_{i=1}^{m} \sum_{j=1}^{2} \left(Y_{ij} - \bar{Y}\right)^2, \tag{3.13}$$

remembering that there are $2m$ total observations. In large samples, that is when m is big, the variance estimate (3.13) approaches its expected value, easily seen to be

$$E\left(\widehat{Var}(\bar{Y})\right) = \frac{{\sigma_b}^2 + {\sigma_e}^2}{2m}. \tag{3.14}$$

However, because of the correlation between Y_{i1} and Y_{i2}, we have

$$\begin{aligned} Var(\bar{Y}) &= Var\left(\frac{1}{2m}\sum_{i=1}^{m}\sum_{j=1}^{2}Y_{ij} \equiv \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{2}Y_{ij}\right) \\ &= \frac{1}{4m^2}\sum_{i=1}^{m}Var\left(Y_{i1}+Y_{i2}\right), \quad \text{since individuals are independent} \\ &= \frac{1}{4m^2}\sum_{i=1}^{m}\left(Var(Y_{i1})+2Cov(Y_{i1},Y_{i2})+Var(Y_{i2})\right) \\ &= \frac{1}{4m^2}\sum_{i=1}^{m}\left(4{\sigma_b}^2+2{\sigma_e}^2\right), \text{from Problem 3.1} \\ &= \frac{2{\sigma_b}^2+{\sigma_e}^2}{2m}. \end{aligned} \tag{3.15}$$

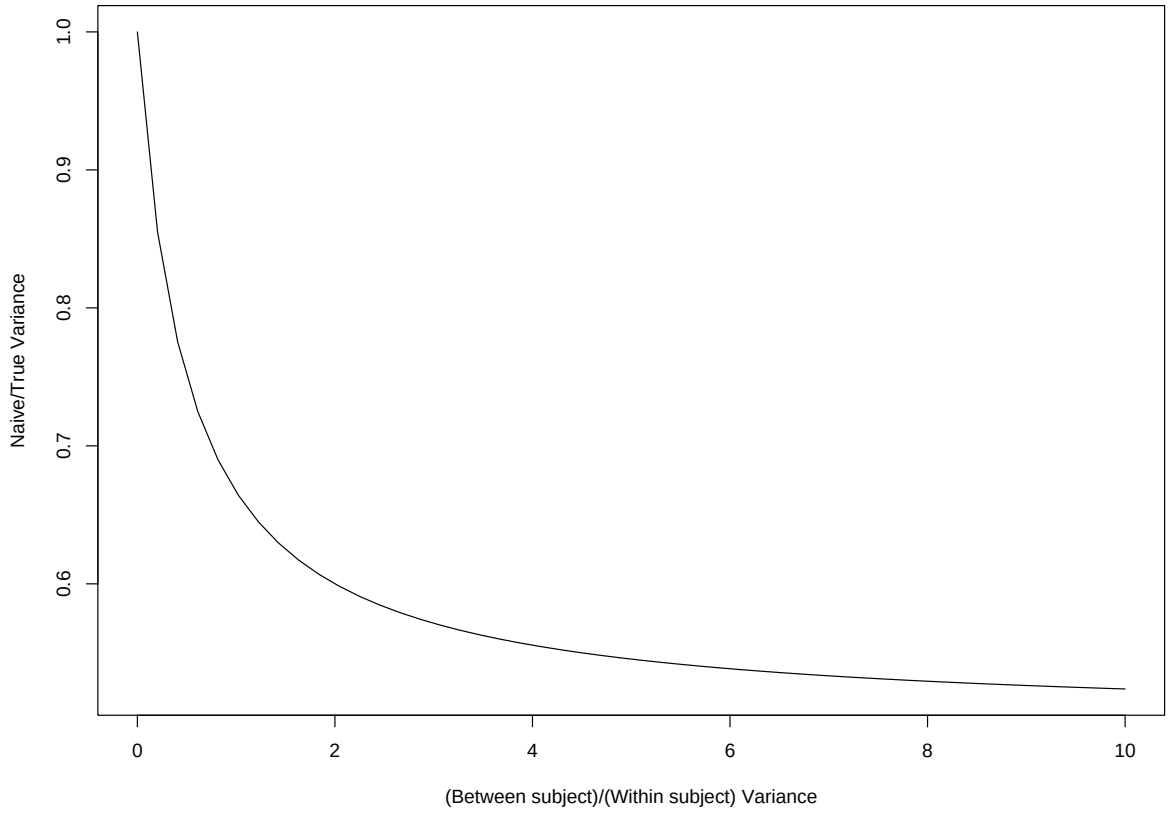
The key point is now clear: the true variance of the (ordinary least squares) estimator $\bar{Y}$, given by (3.15), is estimated incorrectly by the traditional variance formula for an average, (3.13), whose value in large samples approximates (3.14), clearly different from (3.15). In this case, the naive assumption that all observations are independent leads us to underestimate the variance, the size of the error depending on how large ${\sigma_b}^2$ is in regard to ${\sigma_e}^2$. Only when ${\sigma_b}^2 = 0$, in which case the longitudinal observations truly are independent, is the ordinary least squares variance estimator unbiased for the true variance. Further, in this simple example, the variance estimator (3.13) systematically *underestimates* the real variance (3.15) so that its use leads to incorrectly small p-values, and incorrectly too narrow confidence intervals. This reflects common sense since the naive estimator assumes that all $2m$ observations are uncorrelated, whereas in fact the within-person correlation ρ (3.11) indicates that each individual contributes somewhat less than 2 independent observations on viral load. How far off can the naive variance estimate tend to be? Figure 3.3 plots the ratio of what the naive variance formula is estimating over the true variance of $\bar{Y}$ against the ratio of between and within variances; that is, a plot of

$$\frac{{\sigma_b}^2+{\sigma_e}^2}{2{\sigma_b}^2+{\sigma_e}^2} \qquad \text{against} \qquad \frac{{\sigma_b}^2}{{\sigma_e}^2}.$$

Note that when ${\sigma_b}^2 = 0$, all observations *are* independent, and the expected value of the naive variance estimator is the true variance, that is their ratio is 1. As ${\sigma_b}^2$ grows in size as compared to ${\sigma_e}^2$, the naive variance estimator increasingly underestimates the true variance by as much as a half if ${\sigma_b}^2$ is very much bigger than ${\sigma_e}^2$.

Intuition now correctly tells us that as the number of longitudinal observations, n_i, increases, this underestimation of variability becomes increasingly serious, keeping the lon-

Figure 3.3: Effect of Ignoring Correlation in Variance Estimation: Naive/True Variance, given by $(\sigma_\alpha{}^2+\sigma_e{}^2)/(2\sigma_\alpha{}^2+\sigma_e{}^2)$; Between/Within Variance $= \sigma_\alpha{}^2/\sigma_e{}^2$



gitudinal correlation ρ the same. As we add one additional longitudinal observation per person, the naïve estimate assumes that this provides one additional independent observation whereas the presence of longitudinal correlation means that somewhat less than this information is provided by an additional longitudinal measurement. The case when $n_i = 3$ is considered in Question 3.2.

3.4 Problems

Question 3.1 Consider the model (3.10) where there are now three longitudinal observations per person (i.e. $n_i = 3$ for all i) for m different individuals. The estimator $\bar{y}$ of the population mean b_0 is defined, as in (3.12), to be the sample average. Derive the analogous formulae to (3.13) and (3.14) for the naive estimate of the variance of $\bar{y}$, that is $\widehat{Var}(\bar{y})$, and the expectation of this naive estimator, $E(\widehat{Var}(\bar{y}))$. Finally, calculate the true variance of $\bar{y}$ in terms of ${\sigma_\alpha}^2, {\sigma_e}^2$, and m. Compare the true and expected value of the naive variance estimator as in Figure 3.3.