ANALYSIS OF LONGITUDINAL STUDIES IN EPIDEMIOLOGY

NICHOLAS P. JEWELL, ALAN HUBBARD, & BRIANNA C. HEGGESETH[1]

August 18, 2015

[1]Send correspondence to Nicholas P. Jewell, Division of Biostatistics, School of Public Health, 140 Warren Hall #7360, University of California, Berkeley, CA 94720, USA. Tel: 510-642-4627, Fax: 510-643-5163, e-mail: jewell@stat.berkeley.edu

Chapter 1

Introduction

Epidemiology is concerned with the study of the distribution of (human) diseases across populations and sub-populations, a primary goal being the identification of factors that explain observed patterns of disease distribution. Many such investigations involve cross-sectional data that provide a snapshot of disease and risk factor distributions for a population over a fixed period of time. There is a considerable literature on statistical methods underlying the design and analysis of these studies—see Jewell (2003), Rothman, Greenland & Lash (2012), or Woodward (2014). Recent trends in epidemiological studies have been to employ *longitudinal studies* which require, explicitly or implicitly, repeated observation of individuals over possibly varying time periods. While basic cross-sectional statistical approaches carry over to the longitudinal setting, new methods are desirable and necessary for a variety of reasons that we outline in this section and discuss in detail in subsequent chapters.

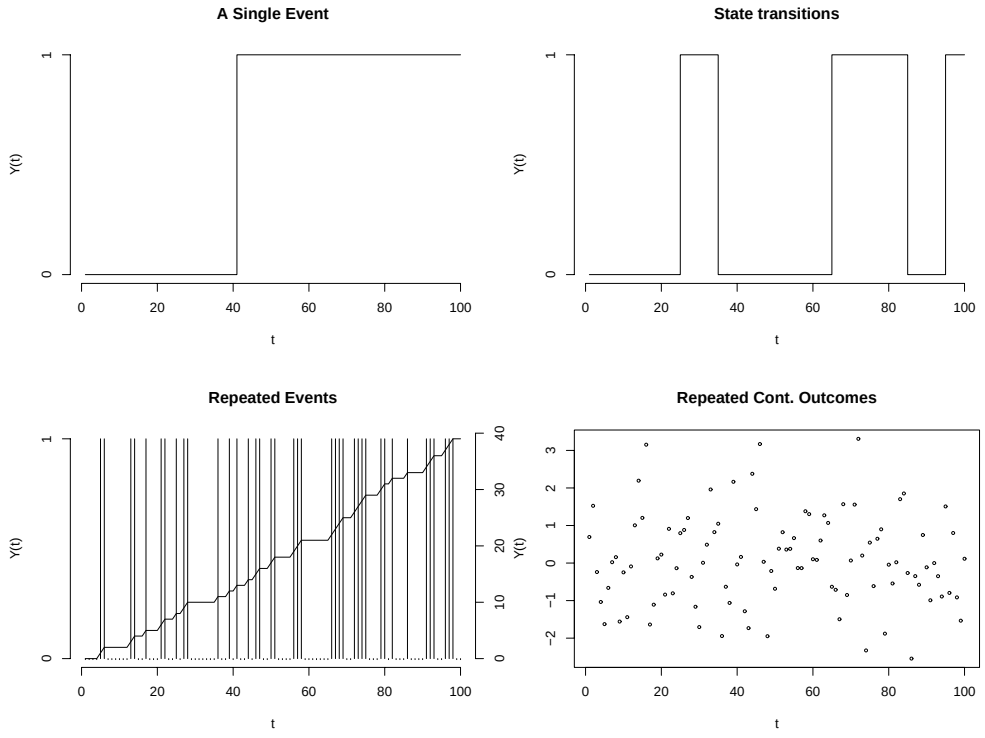
Fully exploiting the advantages of longitudinal data is a prime motivation for approaches that extend cross-sectional thinking. If a cross-sectional study is analogous to a single

photograph in time, then a longitudinal investigation reflects a *movie* of the same events, albeit one that is more like stop-motion than what we are used to seeing on a DVD. Nevertheless, in understanding what happened during an accident, say, it would be foolish to depend solely on a few still photographs when a movie of the entire incident was available. At the very least, cause and effect issues are usually easier to describe and understand with a movie as an aid.

1.1 Longitudinal Studies

As suggested by their name, longitudinal studies involve following individuals over time, thereby measuring a random outcome variable, Y, and p risk factors, $X_1, \ldots, X_p$, at least at two different points in time, and often more. For the moment, as is true for several examples explored in detail, the word 'time' here refers to *chronological time*, although we discuss other important choices. For example, in some applications 'time' might be geographical distance from a point source of an exposure. Notationally, the expression $Y(t)$ stands for the variable Y measured at time t, with similar convention for risk factors or explanatory variables. In general, the outcome Y can take various forms, reflecting binary, count, or continuous measurements. In the binary case, it is possible that once $Y(t) = 1$, it automatically remains that way; that is, $Y(s) = 1$ if $s \geq t$. This arises in mortality studies where Y measures whether an individual is alive ($Y(t) = 0$) or dead ($Y(t) = 1$) at time t. Sample paths for $Y(t)$, as t changes, are particularly simple in this case as the path starts at $Y(0) = 0$ and stays there until it jumps to $Y = 1$ where it then remains (as seen in Figure 1.1(a)). The properties of the outcome variable over time are therefore completely determined by those of the random variable, T, say, which defines

Figure 1.1: Schematics Showing Longitudinal Observations of Outcome Variables $Y(t)$ of Various Forms Over Time: (a) Binary Outcome Associated with Failure Event, (b) General Binary Outcome, (c) Count Outcome, (d) Continuous Outcome.



the time at which Y changes from 0 to 1. In other binary cases, say when denoting the presence and absence of a particular symptom (e.g. wheezing), Y can take the values of 0 and 1 in any particular order at various times. Figure 1.1 shows four possible schematics of a random variable Y observed longitudinally.

In Figure 1.1(c), Y represents a count which may arise from combining binary counts across groups of individuals (e.g. the number of AIDS diagnoses in a state in year t), or, alternatively, from individual measurements such as the total number of sexual partners by

age t. A similar remark applies to continuous outcome variables (Figure 1.1(d)) although we usually consider examples here where $Y(t)$ denotes a continuous random variable for an individual at time t e.g. the CD4 cell count, at t, for an individual suffering from HIV disease.

The primary focus of a longitudinal study is often to elucidate if and by how much the explanatory variables, $X_1, \ldots, X_p$, (or changes in these variables) cause changes in the outcome variable Y. Does taking anti-oxidant supplements make an individual less likely to get cancer before age 70? Does a mother stopping smoking alter her child's frequency of asthma symptoms? Additional kinds of questions may also be of interest. What types of HIV infected individuals share similar patterns of CD4 cell counts over time?

Another kind of longitudinal study involves the collection of repeated measurements of Y and $X_1, \ldots, X_p$ for a *single* entity over time, commonly referred to as a *time series*. For example, we may collect monthly counts of new incident cases of HIV infections in a community of drug-users over several years in order to determine the effect of a needle sharing program. Although time series methods are also worthy of study, they are outside of the scope of this text.

As in all scientific investigations, we are most interested in causal effects of explanatory factors since mere associations may be quite misleading with regard to future events. As in any epidemiological study, randomized experiments—where the primary factor of interest is randomly allocated to study participants—require fewer assumptions to infer causality than observational studies. Nevertheless, even in the absence of randomization, we wish to estimate the causal effects of specific risk factors as well as we can, and understand any assumptions that are necessary for our estimates to have a causal interpretation.

1.2 Notation

Longitudinal studies collect repeated observations on a sample of individuals. In both creating a database and in analyzing data, we need to distinguish between observations on the same and on different individuals. For the outcome variable Y we do this notationally by using two indices: the first, i, to indicate an individual, and the second, j, to cover multiple observations for a single person. Thus, the random variable Y_{ij} represents the j^{th} observation on the i^{th} individual. The mean of Y_{ij} is denoted by $\mu_{ij} = E(Y_{ij})$. We use m to denote the number of individuals sampled, and n_i the number of longitudinal observations taken for the i^{th} person. The index i thus runs from 1 to m, and the index j from 1 to n_i for the i^{th} individual. It is possible that $n_i = 1$ for some individuals, although if this happens for all i, a longitudinal study collapses to being merely cross-sectional, that is, one observation per person. The total number of observations over all individuals is given by $N = \sum_{i=1}^{m} n_i$.

Similar notation can be used for any explanatory variable X. Thus X_{ij} again denotes the j^{th} observation of the variable X on the i^{th} individual. It is useful, conceptually and practically, to distinguish two types of risk variables: *time-dependent covariates* and *time-fixed covariates* for which $X_{ij} = X_i$ for all i and j. A time-fixed covariate, like gender, remains the same over all longitudinal observations of the same individual. On the other hand, a time-dependent covariate, such as blood pressure, may vary from observation to observation on a particular individual. Time-dependent covariates are particularly useful for causal inference in longitudinal settings since they allow us to examine whether changes in risk variables lead to a change in outcome *within the same person*—the so-called longitudinal effect of X, as opposed to comparison of *different* individuals at varying levels of

X, all that is available with cross-sectional information. Note that we can choose to treat a time-dependent covariate as time-fixed: this is common, for example, when X is the baseline value of a variable of interest such as blood pressure. In this case, we focus on the effect of an individual's baseline blood pressure on an outcome's pattern in future longitudinal observations by taking the baseline value of X (that is X_{i1}) as time fixed although we are well aware that blood pressure is likely to also vary over time.

Databases keep track of longitudinal observations in two formats: *long* and *wide*. Table 1.3 lists a small extract of data relating drug and alcohol use to a simple measure of teenage sexual activity, an example introduced in Section 1.3.3—the format here is *long* with one line provided for every unique observation on the same individual. This format requires a variable—*Id. Number* in Table 1.3—that unequivocally identifies the single individual associated with the observation.

The same data can also be recorded in *wide* format with one line per *individual*. Referring again to Table 1.3, this data file would look essentially the same for the first teenager (10122) since there is only one observation here, but compresses all 23 lines of information for individual 10123 onto a single line. Table 1.4 shows the data extract for individuals 10122 and 10124 from Table 1.3 in wide format. Note that data are missing for teenager (10122) on all dates where no information was collected for that individual. In wide format, it is necessary to define variables (or columns in the database) for all unique times of observation over all individuals; this can make the wide format particularly cumbersome when there is a wide variety of observation times over the study participants as we can begin to see from Table 1.3. It is usually necessary to keep track of specific times at which observations are made for each individual since this may be important in either predicting

the outcome, or understanding the correlation between two longitudinal observations on the same individual, or both. For example, we discuss in Chapter 6.5 the possibility that higher levels of sexual activity might be expected on weekends; also, we would anticipate that two measures of a teenager's sexual activity might be more correlated for two days close in time than if the observations were separated by a long period.

The description of relationships between an outcome variable Y and explanatory variables X is one of our primary goals. Such relationships are most easily described using mathematical formulae with appropriate notation. This is particularly true when we want to understand the joint effects of several explanatory variables on Y simultaneously. In writing formulae simply, it is convenient to represent the data for a single individual in *vector* notation. Specifically, the n_i measurements on the i^{th} individual can be represented by the column vector

$$\mathbf{Y}_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{pmatrix}.$$

Here $\mathbf{Y}_i$ is a $n_i \times 1$ vector representing a multivariate random variable with mean μ_i, with

$$\mu_i = \begin{pmatrix} \mu_{i1} \\ \mu_{i2} \\ \vdots \\ \mu_{in_i} \end{pmatrix},$$

and $\mu_{ij} = E(Y_{ij})$. Thus μ_{ij} is just the mean of the j^{th} observation on the i^{th} participant. Referring to Table 1.1 of Section 1.3.1, a mean μ_{23} designates the mean of a random variable Y_{23}, the CD4 count for the second individual (with *Id. Number* = 2) measured at their third visit time (on day 13 after the beginning of the study for this participant). The distribution of Y_{23} refers to potential repeated measurements of the second individual on

their third clinic visit, although only one realization is observed in this case.

The $n_i \times n_i$ variance-covariance matrix of the random variable $\mathbf{Y_i}$ is described by

$$\mathbf{V}_i = \begin{pmatrix} v_{i11} & v_{i12} & \cdots & v_{i1n_i} \\ v_{i21} & v_{i22} & \cdots & v_{i2n_i} \\ \vdots & \vdots & \ddots & \vdots \\ v_{in_i1} & v_{in_i2} & \cdots & v_{in_in_i} \end{pmatrix}. \tag{1.1}$$

The diagonal terms, v_{ijj} are just the variances of the single observations Y_{ij}, respectively. In some examples, we may wish to assume that these diagonal elements are all the same, that is, that the variance of Y_{ij} stays the same over j, the repeated observations. However, this may not represent what is observed; for example, in some cases the variance of the outcome may increase over time, particular if the values of Y_{ij} tend to increase longitudinally. Off the diagonal, the covariance term $v_{ijk} = Cov(Y_{ij}, Y_{ik})$ yields the covariance between the j^{th} and k^{th} longitudinal observation (for the i^{th} individual). Later we also discuss the related *correlation matrix* of the repeated observations in $\mathbf{Y_i}$, a scale-invariant version of the covariance matrix $\mathbf{V}_i$.

Similar vector notation is used to denote repeated observations on a risk factor (or covariate). In most examples, we want to keep track of several covariates, or explanatory variables, simultaneously. We achieve this by combining the vectors of p separate covariates, $X_1, X_2, \ldots, X_p$ into a large $n_i \times (p+1)$ *matrix* for the i^{th} individual as follows:

$$\mathbf{X}_i = \begin{pmatrix} 1 & X_{i11} & X_{i12} & \ldots & X_{i1p} \\ 1 & X_{i21} & X_{i22} & \ldots & X_{i2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{in_i1} & X_{in_i2} & \ldots & X_{in_ip} \end{pmatrix}.$$

For example, in the study of HIV patients described in Section 1.3.1 (see Table 1.1), if CD4 T-cell count is the outcome variable Y, then possible explanatory variables might

be log(viral load) (X_1), gender (X_2), age at initiation of therapy (X_3), and time since initiation of therapy (X_4). Further explanatory variables might be constructs of these basic covariates including quadratic and interaction terms. Including the first column of all 1's allows straightforward inclusion of an 'intercept' term in regression models for $\mathbf{Y}_i$ as we see below. Note that the second column of $\mathbf{X}_i$ gives the longitudinal observations of the first covariate X_1, the third column the longitudinal observations of X_2, and so on.

While the notation appears daunting, it is valuable to understand it carefully since the covariance or correlation structure (1.1) underlying $\mathbf{Y}_i$ represents how multiple observations on the same individual affect each other, at this point not accounting for the effects of other explanatory variables. First, we note that the matrix $\mathbf{V}_i$ is *symmetric*; that is $v_{ijk} = v_{ikj}$ since the covariance between Y_{ij} and Y_{ik} is the same as the covariance between Y_{ik} and Y_{ij}. If $v_{ijk} = 0$ for all j and k, then we can infer that repeated outcome observations on the i^{th} individual are uncorrelated.

It is also helpful to have notation that describes all the data simultaneously. This can be achieved, for example, by lumping all the response vectors $\mathbf{Y}_i$ into one large vector of

dimension $N \times 1$:

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_m \end{pmatrix} = \begin{pmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ Y_{22} \\ \vdots \\ Y_{2n_2} \\ \vdots \\ Y_{m1} \\ Y_{m2} \\ \vdots \\ Y_{mn_m} \end{pmatrix}. \tag{1.2}$$

For convenience and to simply express some arithmetic calculations, we often use the *transpose* of a matrix, $\mathbf{A}$, denoted by $\mathbf{A}^T$, defined as follows: if the matrix $\mathbf{A}$ has element a_{ij} in the ith row and jth column, then $\mathbf{A^T}$ has element a_{ij} in the jth row and ith column. Necessarily, $\mathbf{A^T}$ has dimension $l \times k$ if $\mathbf{A}$ has dimension $k \times l$. For example,

$$\mathbf{Y}^T = \begin{pmatrix} \mathbf{Y}_1^T & \mathbf{Y}_2^T & \dots & \mathbf{Y}_m^T \end{pmatrix}$$

has dimension $1 \times N$. The transpose of a matrix is therefore just a 'reflection' of the original where the first row is now the first column, and so on. Apart from its mathematical advantages, we often use the transpose to make it easier to write a long column vector on a 'horizontal page', as compared to the awkwardness of (1.2), for example.

Similarly, the matrices, $\mathbf{X}_i$ can be stacked 'on top of each other' to give an overall matrix $\mathbf{X}$, of dimension $N \times (p+1)$, where recall that p is the number of distinct explanatory variables. The vector $\mathbf{Y}$ is a very large dimensional vector representing all outcome measurements on all individuals. The mean, or expectation, of $\mathbf{Y}$ is

$$\boldsymbol{\mu} = E(\mathbf{Y}) = \begin{pmatrix} \boldsymbol{\mu}_1^T & \boldsymbol{\mu}_2^T & \dots & \boldsymbol{\mu}_m^T \end{pmatrix}^T,$$

(recalling that each μ_i is itself an $n_i \times 1$ vector), and its variance-covariance matrix, in "block" form, is just

$$\mathbf{V} = \text{variance}(\mathbf{Y}) = \begin{pmatrix} \mathbf{V}_1 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_2 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{V}_3 & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{V}_m \end{pmatrix}. \tag{1.3}$$

The dimension of this matrix is $N \times N$, with each non-zero block $\mathbf{V}_i$ given by (1.1). The blocks of zeros off the 'diagonal' of $\mathbf{V}$ reflect a key assumption we make throughout, namely that observations on different individuals are independent and thus uncorrelated.

We now make several observations about means and variance-covariance matrices that will be used later. First, the *identity matrix* of size $k \times k$ is a square matrix $\mathbf{I}$ that has diagonal elements all equal to 1, and all other elements—off the diagonal—equal to 0. That is,

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}.$$

The identity matrix has the property that, for *any* $k \times k$ matrix $\mathbf{A}$, $\mathbf{IA} = \mathbf{AI} = \mathbf{A}$, using matrix multiplication. (Look at Appendix 1 for an elementary introduction to matrices and their manipulation and application.) The *inverse* of a $k \times k$ square matrix, $\mathbf{A}$, is denoted by $\mathbf{A^{-1}}$, also of dimensions $k \times k$, and has the property that

$$\mathbf{A}\mathbf{A^{-1}} = \mathbf{A^{-1}}\mathbf{A} = \mathbf{I}. \tag{1.4}$$

Recall that, for any two random variables, X and Y, we have that $E(X+Y) = E(X) + E(Y)$, and $Var(X+Y) = Var(X) + 2Cov(X,Y) + Var(Y)$. Also, for any constant c,

$E(cX) = cE(X)$ and $Var(cX) = c^2 Var(X)$ (see Section 1.8 for a more complete list of the properties of expectation and variance). There are simple matrix analogies of these results for a multidimensional random variable $\mathbf{Y} = (Y_1 \dots Y_n)^T$, of dimension $n \times 1$—such as (1.2)—and a constant matrix $\mathbf{c}$, of dimension $k \times n$, say. Specifically,

$$\begin{aligned} E(\mathbf{cY}) &= E\begin{pmatrix} c_{11}Y_1 + c_{12}Y_2 + \cdots + c_{1n}Y_n \\ c_{21}Y_1 + c_{22}Y_2 + \cdots + c_{2n}Y_n \\ \vdots \\ c_{k1}Y_1 + c_{k2}Y_2 + \cdots + c_{kn}Y_n \end{pmatrix} \\ &= \begin{pmatrix} c_{11}E(Y_1) + c_{12}E(Y_2) + \cdots + c_{1n}E(Y_n) \\ c_{21}E(Y_1) + c_{22}E(Y_2) + \cdots + c_{2n}E(Y_n) \\ \vdots \\ c_{k1}E(Y_1) + c_{k2}E(Y_2) + \cdots + c_{kn}E(Y_n) \end{pmatrix} \\ &= \mathbf{c}E(\mathbf{Y}), \end{aligned}$$

and

$$\begin{aligned} Var(\mathbf{cY}) &= Var\begin{pmatrix} c_{11}Y_1 + c_{12}Y_2 + \cdots + c_{1n}Y_n \\ c_{21}Y_1 + c_{22}Y_2 + \cdots + c_{2n}Y_n \\ \vdots \\ c_{k1}Y_1 + c_{k2}Y_2 + \cdots + c_{kn}Y_n \end{pmatrix} \\ &= \mathbf{c}Var(\mathbf{Y})\mathbf{c^T} = \mathbf{cVc^T}, \end{aligned}$$

if $\mathbf{V} = Var(\mathbf{Y})$ is the variance-covariance matrix of $\mathbf{Y}$ as in (1.3).

1.3 Data Sets

The notation of Section 1.2 can be overwhelming outside of the context of specific examples. So we now turn to a brief introduction to several data sets that we use in later chapters to

illustrate both concepts and statistical techniques appropriate to the analysis of longitudinal studies. All of the data sets can be found on the web site associated with the book, in addition to others required for exercises.

1.3.1 HAART Therapy on HIV Patients

Deeks, et al. (1999) report the results from a longitudinal study of HIV-infected adults undergoing Highly Active Anti-Retroviral Therapy (HAART) at the University of California, San Francisco AIDS program at San Francisco General Hospital (SFGH). SFGH is an urban, university-based public hospital clinic that provides comprehensive primary care to HIV-infected adults. Patients were identified through an administrative database that records outpatient visits. The names of all patients seen at least three times by the same clinician between March 1996 and September 1997 were identified. Medical records were reviewed to identify those who were eligible for this study. Patients were included in this analysis if they received at least 16 weeks of continuous therapy with an anti-retroviral regimen containing indinavir, ritonavir or nelfinavir. To allow at least 48 weeks of subsequent follow-up, only patients who initiated therapy before November 1997 are included in this data set. The following data was obtained during the initial review: date of birth, gender, and length of previous exposure to each individual anti-retroviral agent. Once patients were identified, their medical records were reviewed every 3-4 months until November 1998. Plasma HIV RNA assays were performed using a branched DNA (bDNA) assay (Chiron Corp., Emeryville, CA, USA). Before 1 July 1996, HIV RNA tests were performed using an earlier version of the bDNA assay (Quantiplex HIV bDNA version 1.0; lower limit of quantification 10000 copies/ml). If these assays were not available, HIV RNA results from an experimental reverse transcriptase-polymerase chain reaction (RT-PCR) assay were used

(Immuno-Diagnostic Laboratories, San Leandro, California, USA). After 1 July 1996, all HIV RNA determinations were performed with the bDNA assay, version 2.0 (lower limit of quantification 500 copies/ml). After March 1998, all HIV RNA samples below the level of quantification with version 2.0 were re-analysed with version 3.0 (lower limit of quantification 50 copies/ml). Routine CD4 T-cell phenotyping was performed at the SFGH central laboratory.

Table 1.1 provides a brief extract from the data, showing longitudinal information (in long format) on the immunological variables CD4 T-cell count, the natural logarithm of viral load, gender (male = '1', female = '2'), and age at initiation of therapy. The origin of the time variable (days), at which measurements are taken, is also the initiation of therapy—and, as is obvious, the scale is in days. The covariates gender and age are time-fixed and need be measured only once. On the other hand, the viral load and CD4 counts are time-dependent with values measured at each outpatient visit at the noted number of days after therapy started. In analyses of this example, we focus on regression models that attempt to 'explain' variation in patients' CD4 cell counts by patterns of comtemperaneous and past measurements of their viral load, accounting for the possible effects of time since the initiation of HAART and basic demographic information including gender and age.

1.3.2 A Water Filter Intervention Trial

Previous trials of in-home drinking water interventions in general populations attributed up to 40% of gastrointestinal illness to the consumption of improperly treated municipal drinking water whereas others have found no increase in risk. A study conducted from April 2000 to May 2001 in the city of San Francisco (Colford et al., 2005) was a triple-blinded,

Table 1.1: Extract of Data from SFGH/HAART Study

Id. Number	days	CD4 count[a]	log(viral load)[a]	gender	age
1	39	45	2.70	1	32.0
1	137	119	5.22	1	32.0
1	147	113	.	1	32.0
1	179	74	5.20	1	32.0
1	187	95	.	1	32.0
1	298	137	3.87	1	32.0
1	335	.	5.07	1	32.0
1	354	167	5.14	1	32.0
1	411	.	4.66	1	32.0
1	1684	427	.	1	32.0
2	0	196	5.68	1	44.0
2	7	369	3.93	1	44.0
2	13	353	4.11	1	44.0
2	27	474	3.55	1	44.0
2	55	425	3.10	1	44.0
2	111	493	2.70	1	44.0
2	139	464	2.70	1	44.0
2	167	448	2.70	1	44.0
2	195	427	2.70	1	44.0
2	223	460	2.70	1	44.0
2	251	484	2.70	1	44.0
2	279	513	2.70	1	44.0
3	28	46	5.94	1	40.5
3	84	.	6.00	1	40.5
3	146	41	5.72	1	40.5
3	189	53	5.30	1	40.5
3	244	31	5.64	1	40.5
3	286	32	5.96	1	40.5
3	377	26	5.96	1	40.5
3	420	29	5.70	1	40.5
3	455	29	5.70	1	40.5

[a] Here . refers to a missing value

randomized, controlled trial of a drinking water intervention used to estimate the risk of gastrointestinal illness due to municipal drinking water among HIV-positive individuals. The participants were fifty HIV-positive patients who primarily consumed municipal tap water at home. These patients were randomized to use active ($n = 24$) or sham ($n = 26$) water treatment devices that were identical in external appearance. The active device had a 1-micron filter and an ultraviolet light used to rid the water of pathogenic microbes; the sham device consisted of an empty filter casing with no ultraviolet treatment. Triple-blinding here refers to the fact that the nature of an installed water-treatment device for a particular subject was hidden from the subject, installer and investigator.

The main outcome measure was called 'highly credible gastrointestinal illness' (HCGI), a previously published measure that includes symptoms of diarrhea, nausea, vomiting, and abdominal cramps. Events were determined by reading daily diary entries the participants were responsible for keeping. Because some subjects were followed longer than others and some data is simply missing (individuals failed to fill out their diary), outcomes for subjects were measured for varying number of days.

Table 1.2 displays an extract of the data, again in long format, with daily observations of the prevalence of gastrointestinal symptoms. In subsequent chapters, we seek to compare the risk of HCGI across the two randomized groups using the reported longitudinal observations on all subjects.

Table 1.2: Extract of Data from Study of Effect of Water Filters on Incidence of Gastrointestinal Symptoms

Id. Number	Date	HCGI[a]	group[b]
A7283	14780	.	6
A7283	14781	0	6
A7283	14782	0	6
A7283	14783	0	6
A7283	14784	0	6
A7283	14785	0	6
A7283	14786	0	6
A7283	14796	0	6
C1632	14738	.	7
C1632	14739	.	7
C1632	14740	0	7
C1632	14741	0	7
C1632	14742	0	7
C1632	14743	0	7
C1632	14744	1	7
C1632	14745	0	7
C1632	14746	0	7
C1632	14747	0	7
C1632	14748	0	7
C1632	14750	0	7
C1632	14751	1	7

[a] HCGI represents evidence of "highly credible gastrointestinal illness"; . refers to a missing value
[b] 6 refers to use of the active treatment device, 7 to use of the sham device

1.3.3 The Effect of Drug and Alcohol Use on Teenage Sexual Activity

Minnis & Padian (2001) conducted a longitudinal study of teenagers in San Rafael, California to investigate the association between drug and alcohol use on a specific day and sexual activity on the same day. Participants were asked to keep track of their activities over approximately one month and binary indicator variables were created to show whether drug/alcohol use and/or sexual activity were reported for each 24 hour period. The data was originally collected and stored in wide format, where all measurements of the same unit (individual) are on a single row (see an extract in Table 1.4), but converted to so-called long format for analysis (see Table 1.3). It is important to note that the data gives the date of report which refers to activities in the *previous* day. Data are available for 109 teenagers for whom information on 1 to 33 different days are available. The average number of longitudinal observations is 16, with the total number of data points (that is, teenager-days) equal to 1,708. In this example, we will use various regression models to link drug/alcohol use to the occurrence of sexual activity.

1.3.4 World Cup Soccer Data

The World Cup in soccer has been held every four years since 1930, except for 1942 and 1946 during and immediately following World War II. We collected and collated data on the number of goals scored by a single team in every World Cup game played in 17 competitions—note that each game thus provides two data values, one for each team. A summary for an extract of the data—for World Cup 2002—is shown in Table 1.5. The continent covariate gives the continent of a team with 0 referring to South America, 1 to Europe, and 2 to

Table 1.3: Extract of Data from Teenage Survey on Drug/Alcohol Use and Sexual Activity

Id. Number	Date	Drug/Alcohol Use	Sexual Activity
10122	03 Jun 98	yes	no
10123	04 Jun 98	no	no
10123	05 Jun 98	no	no
10123	06 Jun 98	yes	no
10123	07 Jun 98	no	no
10123	08 Jun 98	no	no
10123	09 Jun 98	no	no
10123	12 Jun 98	no	no
10123	14 Jun 98	yes	no
10123	16 Jun 98	no	no
10123	17 Jun 98	no	no
10123	18 Jun 98	no	yes
10123	19 Jun 98	no	no
10123	20 Jun 98	no	no
10123	21 Jun 98	no	no
10123	23 Jun 98	no	no
10123	25 Jun 98	no	yes
10123	28 Jun 98	no	no
10123	29 Jun 98	no	yes
10123	01 Jul 98	no	yes
10123	02 Jul 98	no	no
10123	03 Jul 98	no	no
10123	04 Jul 98	no	no
10123	05 Jul 98	no	no
10124	04 Jun 98	no	no
10124	07 Jun 98	no	no
10124	08 Jun 98	no	no

Table 1.4: Extract of Data from Teenage Survey on Drug/Alcohol Use and Sexual Activity in Wide Format

Id.	Date1	Date2	Date3	Date4	Use1[a]	Use2[a]	Use3[a]	Use4[a]	SA1[a]	SA2[a]	SA3[a]	SA4[a]
10122	6/3/98	6/4/98	6/7/98	6/8/98	yes	.	.	.	no	.	.	.
10124	6/3/98	6/4/98	6/7/98	6/8/98	.	no	no	no	.	no	no	no

[a] Here . refers to a missing value

Table 1.5: Summary of Extract of Data from World Cup Soccer Results

Year	Continent	Goals	Teams
2002	0	0	4
2002	0	1	7
2002	0	2	5
2002	0	3	2
2002	0	4	1
2002	0	5	1
2002	1	0	16
2002	1	1	25
2002	1	2	10
2002	1	3	9
2002	1	4	1
2002	1	8	1
2002	2	0	17
2002	2	1	15
2002	2	2	12
2002	2	3	2

all other continents (including Africa, Asia, Australasia, and Central and North America). Thus, in 2002, a South American team scored no goals in a game on four occasions, and one goal in 7 games; a European team scored two goals on ten occasions, and so on. The data therefore provides crude information on the number of goals that a team scores in a single game, and allows comparison of goal scoring rates by continent of the team and year of the competition. Although this data set is far from epidemiological, we use it to illustrate some issues with Poisson regression models in Chapter 4, in part because violation of assumptions allows insight into extensions to simple Poisson regression methods. The full data set contains 1,286 observations for 643 games in the 17 competitions.

1.3.5 The Western Collaborative Group Study

Rosenman et al. (1975) introduce data arising from the Western Collaborative Group Study (WCGS), a long-term follow-up study of employed men, aged 39 to 59 years old, from 10 Californian companies. Investigators focused primarily on incidence of coronary heart disease (CHD) during follow-up, and measured a number of possible risk factors including lifestyle variables (e.g. cigarette smoking), physiological variables (e.g. serum cholesterol), and behavioral characteristics (e.g. Type A/B personality type). Here we focus on CHD incidence information of 3,154 men all of whom completed 9 years of follow-up over the calendar period ranging from about 1960 to 1970, and provided baseline information on many of these risk factors; an extract of the WCGS data is shown in Table 1.6.

Table 1.6: Extract of Data from WCGS Study

Id	Age	Height	Dbp[a]	Chol[b]	Ncigs[c]	Dibpat[d]	Chd69[e]	Time169[f]
2001	49	73	76	225	25	1	0	1664
2002	42	70	84	177	20	1	0	3071
2003	42	69	78	181	0	0	0	3071
2004	41	68	78	132	20	0	0	3064
2005	59	70	86	255	20	0	1	1885
2006	44	72	90	182	0	0	0	3102
2007	44	72	84	155	0	0	0	3074
2008	40	71	60	140	0	1	0	3071
2009	43	72	76	149	25	0	0	3064
2010	42	70	90	325	0	1	0	1032
2011	53	69	94	223	25	1	0	3091
2013	41	67	96	271	20	1	0	3081
2014	50	72	90	238	50	1	1	1528
2017	43	72	80	189	30	0	0	3072
2018	44	71	80	140	0	0	0	3102
2019	54	70	88	247	3	0	0	1360
2020	45	67	80	220	9	0	1	2426
2021	44	75	90	176	0	1	0	3071

[a]Dbp = diastolic blood pressure at baseline
[b]Chol = cholesterol at baseline
[c]Ncigs = number cigarettes per day at baseline
[d]Dibpat = behavoir type, 0 = type A, 1 = type B
[e]Chd69 = indicator of coronary heart disease by 1969
[f]Time169 = time of CHD in days since baseline

Table 1.7: Extract of Data from Leptosporosis Study

Year	Week	Day	Cases	Rain(mm)	TMAX[a]	TMIN[b]	TMED[c]
1996	1	2	0	0	31.5	25.4	27.9
1996	1	3	0	3	29.8	25.8	27.4
1996	1	4	0	0	31.2	24.6	27.7
996	1	5	0	0	32.3	25.5	28.4
1996	1	6	0	0	31.4	26	28.2
1996	1	7	0	4	32	25.5	27.9
1996	2	8	0	5	31.6	25.1	28.2
1996	2	9	0	0	31.2	26	27.9
1996	2	10	0	0	31.2	24.8	27.7
1996	2	11	0	0	30.8	25.6	27.9
1996	2	12	0	0	30.9	26.2	27.7
1996	2	13	0	1	32.4	25.6	28.4
1996	2	14	0	0	31.6	26.2	28.6
1996	3	15	0	0	31.6	25.8	28.1
1996	3	16	0	2	30.1	26	27.9
1996	3	17	0	0	32	25.8	28.1
1996	3	18	0	0	32.6	26.4	28.9
1996	3	19	0	10	29.2	23.5	26.8
1996	3	20	0	12	29.1	23	25.8
1996	3	21	1	0	31.2	23	27.7
1996	4	22	0	0	31.6	24.2	27.8
1996	4	23	2	0	32	25.6	28.4

[a]TMAX = maximum temperature (°C)
[b]TMIN = minimum temperature (°C)
[c]TMED = median temperature (°C)

1.3.6 Leptospirosis

Leptospirosis is a bacterial disease that affects humans and animals. In humans it causes a wide range of symptoms with around 5–10% of infected individuals suffering severe forms of the disease, and, on rare occasions, death. Symptoms of leptospirosis include high fever, severe headache, chills, muscle aches, and vomiting, and may include jaundice (yellow skin and eyes), red eyes, abdominal pain, diarrhea, or a rash. Outbreaks of leptospirosis are usually caused by exposure to water contaminated with the urine of infected animals, typically following heavy rainfall with subsequent sewer flooding. Urban outbreaks in large Latin American city slums are assumed to result from poor sanitation infrastructure and proliferation of rodent populations.

Investigators have demonstrated a strong correlation between rainfall patterns and leptospirosis incidence (see, for example, Kupek et al., 2000). The data used here arose from surveillance data in an infectious disease hospital in Salvador, Brazil, an institution that accounts for 95% of case notifications in the city (Flannery et al., 2001). A subset of the data is shown in Table 1.7, giving cases admitted per calendar day for a five year period from March 1996 to March 2001. In addition, meteorological information on daily rainfall, temperature (maximum, minimum, and median), and relative humidity for the same period were also collected. One goal for data analysis is estimation of the lag time between high rainfall days and days of high case counts, providing insight into the disease's incubation period in addition to suggesting appropriate time periods for possible intervention after periods of heavy rain. Furthermore, there is considerable interest in teasing out the separate influences of temperature, rainfall amount and frequency on leptospirosis incidence.

Note that the leptospirosis observations are not on single subjects but are ecological,

that is, summarize the experience of all individuals in Salvador on a single day. Chapter 10 gives a brief introduction to longitudinal analysis of such grouped data, where interest nevertheless focuses on interpretation at the individual level (extract of data in Table 1.7).

1.4 Regression Models

With some notation in hand, and a few motivating data sets in mind, we now focus on our primary questions of interest, namely how changes in the levels of the covariates, $X_1, X_2, \ldots, X_p$, are associated with changes in the outcome Y. Many of the regression models considered in this book all take the following form when applied to data on the i^{th} individual

$$g\left[E(Y_{ij})|X_{ij1} = x_{ij1}, \ldots, X_{ijp} = x_{ijp}\right] = b_0 + b_1 x_{ij1} + \cdots + b_p x_{ijp}, \tag{1.5}$$

where the link function g depends, in part, on the nature of the outcome variable Y, and $b_0, b_1, \ldots, b_p$ are regression coefficients whose interpretation in turn depends on the choice of g. In short, (1.5) assumes that some function of the mean of the outcome depends linearly on the values of the set of covariates. For example, with continuous outcome data, a *linear* regression model, where $g(y) = y$, is often used:

$$E(Y_{ij} \mid X_{ij1} = x_{ij1}, \ldots, X_{ijp} = x_{ijp}) = b_0 + b_1 x_{ij1} + \cdots + b_p x_{ijp}. \tag{1.6}$$

In this case, (1.6) shows that the mean of Y_{ij} varies linearly with the covariates $X_1, X_2, \ldots, X_p$. The regression coefficient b_k, for any k with $1 \leq k \leq p$, is then interpreted as the change in the mean of Y_{ij} associated with a unit (on the relevant scale) increase in x_{ijk}, holding all other covariates in the model fixed. For example, with the HAART data of Section 1.3.1, we might use Y_{ij} for the j^{th} measurement of the CD4 cell count on the i^{th} patient,

with $X_{ij1}, X_{ij2}, X_{ij3}, X_{ij4}$ representing the measurement of log viral load, age (at initiation of HAART therapy), time since the beginning of therapy, and gender, respectively, at the same time on the same patient. For simplicity in interpreting the coefficients above, we assumed that the covariates are all separate risk factors so that (1.6) does not include interaction terms or other similar constructed covariates. In addition, note that (1.6) assumes that the regression coefficients are the same for all individuals. It is easy to relax both of these restrictions and we will soon have the opportunity to think about this more when we turn to several of our data analyses.

The model (1.6) only describes how the (conditional) mean of Y_{ij}, given a specified set of values for the covariates, changes as you look at different values for $x_{ij1}, x_{ij2}, \ldots, x_{ijp}$. An alternative way of approaching this model, which opens the door to additional assumptions about the *distribution* of Y_{ij} rather than just its mean, is as follows: given $X_{ij1} = x_{ij1}, X_{ij2} = x_{ij2}, \ldots, X_{ijp} = x_{ijp}$, for the j^{th} observation on the i^{th} individual, we can describe the outcome as follows:

$$Y_{ij} = b_0 + b_1 x_{ij1} + \cdots + b_p x_{ijp} + e_{ij}, \tag{1.7}$$

where e_{ij} is the 'error' or residual that describes how the random variable Y_{ij} varies around its mean value $b_0 + b_1 x_{ij1} + \cdots + b_p x_{ijp}$. Equation (1.7) implies (1.6) so long as we assume that

$$E(e_{ij} \mid X_{ij1} = x_{ij1}, \ldots, X_{ijp} = x_{ijp}) = 0. \tag{1.8}$$

As we are holding the covariates fixed in these models, (1.7) shows that the variability of Y_{ij} is completely determined by the variability of e_{ij}. Using the notation of Section 1.2, this is symbolically represented by

$$Var(\mathbf{Y_i}) = \mathbf{V_i} = \mathbf{Var}(\mathbf{e_i}), \tag{1.9}$$

where $\mathbf{e}_i = \begin{pmatrix} e_{i1} & \dots & e_{in_i} \end{pmatrix}^T$. In particular, this specifies that the covariance structure of the repeated outcomes Y_{ij} is exactly that of the repeated residuals e_{ij}.

If the outcome Y describes counts of certain events over time, such as the number of gastrointestinal illness during follow-up, then a Poisson regression model is often appropriate– see Chapter 4. In this case, the mean of Y represents the rate of occurrence of the event per unit time, and the link function is chosen to be $g(y) = \log(Y)$. In this case, the regression coefficient b_k, measures the log of the relative rate of the outcome associated with a unit increase in x_{ijk}, holding all other covariates in the model fixed.

Finally, the ubiquitous logistic regression is often used if the outcome Y is binary, where now $E(Y) = Pr(Y = 1)$. Here, an appropriate link function is $g(y) = \log\left[\frac{y}{1-y}\right]$, the log odds function. Now, the regression coefficient b_k in (1.6) can be interpreted as the log odds ratio associated with a unit (on the relevant scale) increase in x_{ijk}, holding all other covariates in the model fixed. For introductions to cross-sectional logistic regression models see Jewell (2003), Hosmer and Lemeshow (2000), and Woodward (2004).

Assuming (1.6) and using the matrix notation of Section 1.2, we can write the model for all the longitudinal data for the i^{th} person as

$$g\left[E(\mathbf{Y}_i)|\mathbf{X}_i = \mathbf{x}_i\right] = \mathbf{x}_i\boldsymbol{b},$$

where the regression coefficient vector $\boldsymbol{b} = (b_0, b_1, \dots, b_p)^T$. Similarly for the entire data set, we achieve the most succinct description of the model through

$$g\left[E(\mathbf{Y})|\mathbf{X} = \mathbf{x}\right] = \mathbf{x}\boldsymbol{b}. \tag{1.10}$$

As we alluded to above, this overall model, (1.10), makes a restrictive assumption that we wish to relax in some applications, namely that the regression coefficients, $\boldsymbol{b}$, do not

depend on i, that is, are the same for all individuals. In random and mixed effects models–see Chapters 5 and 7—we explicitly consider the possibility that either intercept or some slope coefficients in $\boldsymbol{b}$ may vary from individual to individual.

1.5 Overview of Chapters

This book introduces the reader to current statistical techniques used to collect and analyze data arising from various kinds of longitudinal epidemiological studies. We assume familiarity with epidemiologic methods for cross-sectional data including the ubiquitous linear and logistic regression models. Chapter 2 discusses introductory graphical tools to present longitudinal data that are helpful in beginning to think about the scales of both the outcome and explanatory variables and possible regression models that link these. Chapter 3 considers simple summary measures of longitudinal observations for each individual with a view to then applying cross-sectional techniques on these summary outcomes. For repeated longitudinal outcomes this approach naturally leads to regression models for count data and allows us to investigate Poisson regression. Chapter 4 introduces the additional complexity and opportunities presented by longitudinal data, introducing the major themes of subsequent chapters. We first develop intuition by thinking about what might go wrong with a naive application of cross-sectional regression strategies to longitudinal data.

Chapters 5 through 7 consider regression models for more complex longitudinal data, starting with continuously scaled outcome variables, and then extending the ideas to accommodate binary outcomes. A common feature of our approach is to first consider extending simple cross-sectional methods—linear and logistic regression—to the longitudinal setting, determining the limitations of such a strategy to provide a context for more complicated

techniques.

Chapter 8 considers a somewhat different type of research question in which the main interest in modeling the heterogeneity in longitudinal patterns through a finite set of groups. A rich set of tools have been developed to analyze the effect of baseline characteristics on the groups membership. We give a brief introduction to these ideas and finite mixture models in Chapter 8.

Chapter 9 introduces the ideas of causal inference to longitudinal data structures. We recommend that the reader review material on causal graphs, counterfactuals, causal measures of association, and confounding in the simpler cross-sectional setting (Jewell, 2003, Chapter 8), before tackling this material.

1.6 Comments and Further Reading

Several books are now available that directly address the analysis of longitudinal data in a variety of formats. The closest of these to our approach is the excellent monograph by Diggle et al (2002). This book is not solely concerned with epidemiological applications and does not cover in any detail the material of Chapters 2, 4, 8–9. Other books, including Verbeke and Molenberghs (2000), go into much more detail, focusing primarily on the material of Chapters 5–7. Both of these texts are at a higher mathematical level than used here where we depend more heavily on intuition derived from cross-sectional methods.

Other books include Fitzmaurice, Laird, and Ware (2004), and Singer and Willett (2003).

1.7 Problems

Question 1.1 For each of the data set fragments, given in Tables 1.1–1.6, provide data matrices for suitably chosen outcome and explanatory variables (hint: first define $\mathbf{Y}$ and $\mathbf{X}$ for one individual, then extend the definition to all individuals in the dataset).

Question 1.2 Suppose a study followed a simple random sample of ten individuals for approximately six months. During clinic visits over the follow-up period, blood pressure measurements (diastolic, a continuous variable) were taken, and a simple questionnaire related to issues of stress was administered. The questions regarding stress were summarized by a simple binary score (High stress = 1; Low stress = 0). The number of visits to the clinic during follow-up varied among individuals: 3 came only once, 6 came twice, and 1 came four times. Using the notation introduced in this chapter:

(a) What is the value of m?

(b) What are the values of n_i for each i?

(c) Symbolically, write down the entire data set (order the individuals from those with fewest to most visits) treating the blood pressure measurement as the outcome variable of interest–be specific about the dimension of the vectors and matrices you use.

(d) Symbolically, write down a variance-covariance matrix of the blood pressure measurements for one of the subjects who had more than one visit, stating any assumptions you may use.

(e) Write down a linear model that relates properties of the blood pressure measurement at a particular time to the binary measure of stress at the same time on the same

individual. Repeat using vector notation.

(f) Write down the equivalent linear model for the entire data set using matrix notation.

(g) Write down a linear model that allows the effect of stress on (the mean) of diastolic blood pressure to vary with time since the beginning of the study.

Question 1.3 Consider a clustered, longitudinal study of diarrheal disease in children in Ecuador [Eisenberg et al., 2006]. Though not precisely how this study was done, assume the study population was collected by: i) a random sample of 10 *independent* villages, ii) within those villages, a random sample of 35 households that contained children under the age of 5, and iii) within the households, all children under the age of 5. At three annual visits, household data were collected on:

I) the (binary) outcome of diarrhea over the last week for each child

II) whether or not any water treatment was currently being used at the household level, and

III) the proportion of villagers that traveled to the closest main town over the past month (village-level "movement" variable).

Remaining as close as possible to the notation provided in this chapter,

(a) What is the value of m (number of independent units)?

(b) What is a good notation for the outcome and explanatory variables, including the appropriate indices, and the range of each index, e.g., $i = 1, \ldots, m$, $j = 1, \ldots,$ __?

(c) Provide a (symbolic) logistic (logit-linear) model that relates the probability of diarrhea for a particular child for at least one of the three visits given the two explanatory variables listed above - assume that all observations have the same relationship (coefficients) with the explanatory variables.

Question 1.4 Consider a simple experiment of measuring cholesterol twice on each of m (independent) individuals. In addition, we will assume a simple random effects model of the form:

$$Y_{ij} = \mu + \alpha_i + e_{ij}$$

for the jth measurement ($j = 1, 2$) on individual i ($i = 1, \ldots, m$), where, $E(\alpha_i) = 0$ $E(e_{ij}) = 0$, $cov(e_{i1}, e_{i2}) = 0$, $cov(\alpha_i, e_{ij}) = 0$, μ is a constant and thus the mean of Y_{ij}, the variance betwee individuals is $var(\alpha_i) = \sigma^2_\alpha$, variance within individuals (between measurements) is $var(e_{ij}) = \sigma^2_e$.

Using the rules and definitions provided in Section 1.8 (one does not necessarily need them all), demonstrate that the correlation of measurements made on the same subject, $Cor(Y_{i1}, Y_{i2}) = \rho$, is:

$$\rho = \frac{\sigma^2_\alpha}{\sigma^2_\alpha + \sigma^2_e}.$$

Show all steps, and reference which definitions and rules were used to arrive at that particular step. Note that $E(X)$ is expectation of X, $Cov(X, Y)$ is covariance of X and Y, $Var(Y)$ is variance of Y, $Cor(X, Y)$ is correlation of X and Y, $SD(X)$ is standard deviation of X. For a simple random effects model, this ρ is often called the *intraclass correlation coefficient*.

Question 1.5 (Data-generating distributions for repeated measures data.) Simulate data

in STATA from the model implied by the data-generating description in *Question 1.4*. Assume that random variables are normally distributed and that the following parameters are given:

- $\sigma_\alpha^2 = 0.5$
- $cor(Y_{ij}, Y_{ij'}) \equiv \rho = 0.3$
- $\mu = EY_{ij} = 10$

Simulate data at sample sizes of $m = 20, 100$, and $1,000$ always with $n_1 = \cdots = n_m = 2$. For each of these simulations, store estimates of ρ and μ. Repeat this procedure 100 times. Create the following:

(a) STATA code used to generate the simulated data and estimates of ρ and μ.

(b) A plot of these estimates versus the sample size m, separately for ρ and μ. Each point on the plots should be a boxplot, indicating the simulated empirical distribution of ρ and μ at sample sizes $20, 100$, and $1,000$ over the 100 simulations. For ease of interpretation, each plot should have a horizontal line indicating the true value of ρ and μ.

(c) A short explanation of what these plots show.

Question 1.6 (Review of linear regression.) Run the following STATA code to simulate correlated data with a continuous outcome:

```
clear
```

```
**n=100
set obs 100
gen X1 = 8*runiform()
gen X2 = 0.8*X1 + rnormal(-0.5,2)
scalar b0 = -3
scalar b1 = 1.0
scalar b2 = -1.25
gen Y = b0+b1*X1+b2*X2 + rnormal(0, 0.5)
```

(a) Given the preceding code, describe the data-generating distribution. How were X_1 and X_2 generated? Are they conditional distributions or univariate distributions? Be sure to include the model of regression (e.g. $E[Y|X = x] = a + bx$, $a = 3$, $b = 5$).

(b) Calculate the predicted value at $X_1 = 0$, $X_2 = 1$.

(c) What is the true change in the mean of Y when X_1 changes by 0.5, keeping X_2 fixed?

(d) Now run the simulation for three sample sizes, $n = 100, 500$ and $1,000$. Repeat (b) and (c) for the estimated model produced by the simulated data at all three sample sizes, and also provide a 95% confidence interval for each estimate.

(e) Interpret to the best of your ability every number in the output on the row of results starting with X_2 (for the sample size of $n = 100$).

(f) Calculate and describe what happens to the bias of the estimate of b_1 as the sample size increases.

Question 1.7 (Review of logistic regression.) Run the following STATA code to simulate correlated data with a binary outcome:

```
clear
**n=100
set obs 100
gen X1=5*runiform()
gen X2=0.5*X1+rnormal(0,2)
scalar b0 = -2
scalar b1 = 2.0
scalar b2 = -2.25
gen logitPY = b0+b1*X1+b2*X2
gen PY=1/(1+exp(-logitPY))
gen Y = rbinomial(1, PY)
```

(a) Based on the preceding code, describe the data-generating distribution including the model of regression (e.g. $logit[E[Y|X = x]] = a+bx$, $a = 3$, $b = 5$). More specifically, what is the distribution of Y? Is it conditional? How is its parameter defined?

(b) Calculate the predicted value at $X_1 = 0$, $X_2 = 1$. What is a predicted value in the context of logistic regression?

(c) What is the true odds ratio when X_1 changes by 0.5, keeping X_2 fixed?

(d) Now run the simulation for three sample sizes, $n = 100, 500$ and $1,000$. Repeat (b) and (c) for the estimated model produced by the simulated data at all three sample sizes, and also provide a 95% confidence interval for the odds ratios from (c).

(e) Interpret to the best of your ability every number in the output on the row of results starting with X_1 (for the sample size of $n = 500$).

(f) Calculate and describe what happens to the estimated standard deviation (that is, the standard error) of the estimate of b_1 as the sample size increases.

Question 1.8 (Proof by Matching) Consider a hierarchical random effects model with independent units indexed by i (say people), sub-units by j (say different days), and sub-sub units k (say repeated measures on the same day). Below we define a model, and provide a list of steps that prove the form of the correlation of two observations on the same (i, j) but different k's (say, same individual and day, but different repeated measures). For these steps, match the part of the model description below or property from **Section 1.8** that allows one to go from one step (one row) to the next step (row).

Steps 1 and 9 have been filled in, as they are the beginning of two parts of the proof. Note that it may be possible to include more than one of these per step. Note also that the choice of symbols and indices in the parameter definitions and properties is arbitrary and does not necessarily correspond to notation in steps (e.g. $Cov(Y_{ij}, Y_{ik})$ is the same as $Cov(X, Z)$).

Description of the Model

1. $Y_{ijk} = \mu + \alpha_i + \alpha_{ij} + e_{ijk}$

2. $E(\alpha_i) = E(\alpha_{ij}) = E(e_{ijk}) = 0$

3. α_i, α_{ij}, and e_{ijk} are independent of each other, both for different indices in the same random variable (e.g. α_i is independent of $\alpha_{i'}$ for $i \neq i'$) and between these random

variables (e.g. α_i is independent of α_{ij}).

4. $Var(\alpha_i) = \sigma_1^2$

5. $Var(\alpha_{ij}) = \sigma_2^2$

6. $Var(e_{ijk}) = \sigma_e^2$

Steps

1. For $k \neq k'$, $Cov(Y_{ijk}, Y_{ijk'}) =$ _____

2. $E\left[(Y_{ijk} - E[Y_{ijk}])(Y_{ijk'} - E[Y_{ijk'}])\right] =$ _____

3. $E\left[(\mu + \alpha_i + \alpha_{ij} + e_{ijk} - E[Y_{ijk}])(\mu + \alpha_i + \alpha_{ij} + e_{ijk'} - E[Y_{ijk'}])\right] =$ _____

4. $E\left[(\mu + \alpha_i + \alpha_{ij} + e_{ijk} - \mu)(\mu + \alpha_i + \alpha_{ij} + e_{ijk'} - \mu)\right] =$ _____

5. $E[\alpha_i^2] + 2E[\alpha_i\alpha_{ij}] + E[\alpha_i e_{ijk'}] + E[\alpha_{ij}^2] + E[\alpha_{ij}e_{ijk'}] + E[\alpha_i e_{ijk}] +$
   $E[\alpha_{ij}e_{ijk}] + E[e_{ijk}e_{ijk'}] =$ _____

6. $E[\alpha_i^2] + 2E[\alpha_i]E[\alpha_{ij}] + E[\alpha_i]E[e_{ijk'}] + E[\alpha_{ij}^2] + E[\alpha_{ij}]E[e_{ijk'}] +$
   $E[\alpha_i]E[e_{ijk}] + E[\alpha_{ij}]E[e_{ijk}] + E[e_{ijk}]E[e_{ijk'}] =$ _____

7. $E[\alpha_i^2] + E[\alpha_{ij}^2] =$ _____

8. $\sigma_1^2 + \sigma_2^2$ _____

9. $Var(Y_{ijk}) = Var(\mu + \alpha_i + \alpha_{ij} + e_{ijk}) =$ _____

10. $Var(\mu) + Var(\alpha_i) + Var(\alpha_{ij}) + Var(e_{ijk}) = Var(\alpha_i) +$
    $Var(\alpha_{ij}) + Var(e_{ijk}) =$ _____

11. $\sigma_1^2 + \sigma_2^2 + \sigma_3^2$ _____

12. $Cor(Y_{ijk}, Y_{ijk'}) = \dfrac{\sigma_1^2 + \sigma_2^2}{\sigma_1^2 + \sigma_2^2 + \sigma_3^2}$, $k \neq k'$ _____

1.8 Definitions and Rules

i) $E[aY] = aE[Y]$ if a is a constant

ii) Expectation of a sum is the sum of the expectations: $E\left[\sum_{i=1}^{n} Y_i\right] = \sum_{i=1}^{n} E[Y_i]$.

iii) $E[X * Y] = E[X]E[Y]$ if X and Y are independent random variables

iv) $Var(aY) = a^2 Var(Y)$ if a is a constant.

v) $Var(Y) = E[(Y - E[Y])^2] = E[Y^2] - E[Y]^2$

vi) $SD(Y) = \sqrt{Var(Y)}$

vii) $Var(\alpha_i + e_{ij}) = Var(\alpha_i) + Var(e_{ij})$ if α_i and e_{ij} are independent random variables with $Cov(\alpha_i, e_{ij}) = 0$ - that is the variance of a sum is the sum of the variances if the random variables are independent.

viii) $Cov(Y_{i1}, Y_{i2}) = E\left[[Y_{i1} - EY_{i1}][Y_{i2} - EY_{i2}]\right] = E[Y_{i1}Y_{i2}] - E[Y_{i1}]E[Y_{i2}]$.

ix) In general, $Var(Y_{i1} + Y_{i2}) = Var(Y_{i1}) + Var(Y_{i2}) + 2Cov(Y_{i1}, Y_{i2})$ or the variance of a sum of two random variables is the sum of the variances plus 2 times the covariance.

x) $Cor(Y_{ij}, Y_{ik}) = \dfrac{Cov(Y_{ij}, Y_{ik})}{SD(Y_{ij})SD(Y_{ik})}$.

1.9 References

Colford, J.M. Jr., Saha, S.R., Wade, T.J., Wright, C.C., Vu, M., Charles, S., Jensen, P., Hubbard, A., Levy, D.A., Eisenberg, J.N.S. (2005) A pilot randomized, controlled trial of an in-home drinking water intervention among HIV+ persons. *J. Water Health.* **03**, 173-184.

Deeks, S.G., Hecht, F.M., Swanson, M., et al. (1999) HIV RNA and CD4 cell count response to protease inhibitor therapy in an urban AIDS clinic: response to both initial and salvage therapy. *AIDS.* **13**, F35-F43.

Eisenberg, J.N.S., Cevallos, W., Ponce, K., Levy, K., Bates, S.J., Scott, J.C., Hubbard, A., Vieira, N., Endara, P., Espinel, M., et al. (2006). Environmental change and infectious disease: how new roads affect the transmission of diarrheal pathogens in rural Ecuador. *Proceedings of the National Academy of Sciences.* **103(51)**, 19460-19465.

Jewell, N.P. (2003) *Statistics for Epidemiology.* Boca Raton, FL: Chapman & Hall/CRC Press.

Kupek, E., de Sousa Santos Faversani, M.C., de Souza Philippi, J.M. (2000) The relationship between rainfall and human leptospirosis in Florianopolis, Brazil, 1991-1996. *Brazilian J. Infectious Disease.* **4(3)**, 131-134.

Minnis, A.M., Padian, N.S. (2001) Reliability of adolescents' self-reported sexual behavior: a comparison of two diary methodologies. *J. Adolescent Health.* **28(5)**, 394-403

ROSENMAN, R.H., ET AL. (1975) Final follow-up of $8^1/_2$ years in the Western Collaborative Group Study. *J. Amer. Medical Association.* **233**, 872-877.

ROTHMAN, K., GREELAND, S., & LASH, T.L.(2012) *Modern Epidemiology (3rd Edition).* Philadelphia, PA: Lippincott, Williams and Wilkins.

WOODWARD, M. (2014) *Epidemiology: Study Design and Data Analysis (3rd Edition).* Boca Raton, FL: Chapman & Hall/CRC Press.