ST 790

Statistical Methods for Analysis With Missing Data

Lecture Notes

M. Davidian and A. A. Tsiatis

Department of Statistics

North Carolina State University

1 Introduction and Motivation

1.1 Objectives and scope

OBJECTIVE OF THIS COURSE: The goal of this course is to provide a comprehensive overview of modern methods for statistical analysis in the presence of ***missing data***.

BIG PICTURE: In many practical settings, such as clinical trials, sample surveys, or agricultural experiments, data are to be collected according to a predetermined plan or experimental design. Given these data, the objective is to make inference about some aspect of an underlying population of interest.

However, things may not turn out as intended – data that were supposed to be collected are not collected or are not available.

- ***Non- or partial response in sample surveys.*** A survey may be conducted to estimate the proportion of the population of likely voters favoring a certain candidate or, as in the example we discuss shortly, estimate features of the distribution of income in a population of households. A sample of (randomly chosen) individuals from the population is to be contacted or sent a questionnaire.

  However, some members of the sample may not answer the phone or refuse to respond, or some may fail to return the questionnaire or provide only a partial response. The response of interest (candidate preference, household income) will be ***missing*** for such individuals.

- ***Dropout and noncompliance in clinical trials.*** A clinical trial may be conducted to compare two or more treatments in a certain patient population. Subjects are recruited and enrolled in the study; are randomly assigned to one of the treatments; and are supposed to take the treatment as directed and return to the clinic weekly, at which times an outcome measure is recorded.

  However, some subjects may fail to show up for any clinic visit beyond a certain point, "dropping out" of the study. Others may quit taking their assigned treatments as directed or quit taking them altogether. Still others may miss clinic visits sporadically. Here, part of the intended full set of longitudinal outcomes arising from taking assigned treatment and visiting the clinic as directed will be ***missing*** for these subjects.

- ***Surrogate measurements and missingness by design.*** In a nutrition study, the daily average percent fat intake in a certain population may be of interest, and a (random) sample of subjects from the population is recruited to participate. Accurate measurement of long-term fat intake requires subjects to keep a detailed "food diary" over a long period of time, which is both time-consuming for subjects to maintain and expensive for investigators to orchestrate. A simpler and cheaper measure of fat intake is to have subjects recall all the food they ate in the last 24 hours. This 24-hour recall may be correlated with long-term fat intake, but will obviously be an error-prone measurement of it. Such a measure is often referred to as a ***surrogate*** for the more detailed measure.

  To reduce costs and subject burden, the study may be designed so that all subjects provide a 24-hour recall (surrogate) measurement, but only a subsample of participants provide the expensive, time-consuming diary measurement; such a subsample is referred to as a ***validation sample***. Here, the expensive measure will be ***missing*** for all subjects who are not part of the validation sample.

  Note that in this example the fact that some subjects are missing the expensive measure is ***deliberate***; that is, the missingness is ***by design***. This is in contrast to the previous two examples, in which missingness is ***by happenstance*** and outside the control of the investigators.

It is a fact of life in almost all studies involving human subjects that some information that was intended to be collected is missing for some subjects. This may be because of oversight or mistakes by the personnel conducting the study or because some subjects refuse or are unable to provide information. As a result, ***missing data*** are a routine challenge to the data analyst in human studies.

PROBLEM: As mentioned at the outset, usually, interest in these settings focuses on making inference about some aspect of the distribution of the ***full data*** that were intended to be collected and would be observed if no data were missing.

When some of the intended full data are missing, depending on why and how they are missing, the validity of the desired inferences may be ***compromised***. In the next section, we present examples illustrating this principle.

APPROACHES: One possible approach to analysis in the presence of missing data is to just ***ignore*** the problem and analyze the ***observed data*** that were collected as if they were the intended, full data. As the examples demonstrate, this can lead to misleading conclusions in many situations.

Clearly, then, statistical methods are required that acknowledge the fact that some intended data are missing and that attempt to "correct" for this somehow. Such methods can be based only on the available, ***observed data***. The challenge is that the desired inference is on some aspect of the distribution of the ***full data***, but the full data are not entirely available.

This challenge has led to a vast literature on statistical models and methods for analysis in the presence of missing data, and this area continues to be a topic of vigorous current research.

Interestingly, although missing data have always been an issue in many areas of application, especially when human subjects are involved, it was not until the 1970s that formal approaches to describing and addressing them began to be developed in earnest. Prior to that time, missing data were viewed mainly as a computational challenge. For example, in designed experiments, missing data induces a lack of balance that complicates the required computations. Although today these computations are trivial, in the era pre-dating modern computing, this was a significant obstacle.

The focus on computation meant that the inferential challenges went largely unexplored. A fundamental paper published in the mid-1970s (Rubin, 1976), which laid out a framework for thinking about missing data, characterizing formally how they can arise and elucidating their possible implications for inference, catalyzed research on methods for taking missing data into appropriate account in inference.

OBJECTIVE, RESTATED: In this course, we will review major developments in the literature on statistical models and methods for analysis in the presence of missing data, including

- A formal framework for thinking about missing data that clarifies the assumptions that must be made/justified, and associated terminology
- "Naïve" methods and their drawbacks
- Three main classes of methods for drawing valid inferences in the presences of missing data under certain assumptions on how the missing data arise
- Methods for assessing the sensitivity of inferences to departures from assumptions on how missing data arise.

To set the stage, we now discuss examples that demonstrate the challenges involved.

1.2 Examples

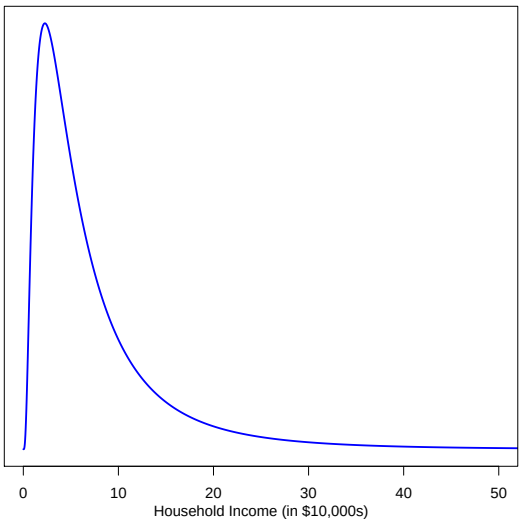
EXAMPLE 1: Nonresponse in sample surveys. Suppose a survey is conducted to learn about the financial status of households in the US population. A survey questionnaire is sent to members of a (random) sample of 1,000 individuals selected from this population. As is often the case with such surveys, many participants fail to answer some (or all) of the survey questions.

In particular, a large proportion of the survey participants fail to respond to the question asking them to report total annual household income.

Data from the US Census Bureau for 2012 indicate that median annual household income in the US is approximately $51,000. The so-called ***Gini coefficient*** of US household incomes is about 0.48; the Gini coefficient (or index) is a measure of statistical dispersion that is often used to reflect the extent of income inequality, where a Gini coefficient of 0 expresses perfect equality, with all members of the population having the same income; and a coefficient of 1 expresses maximal inequality, with one household having all the income.

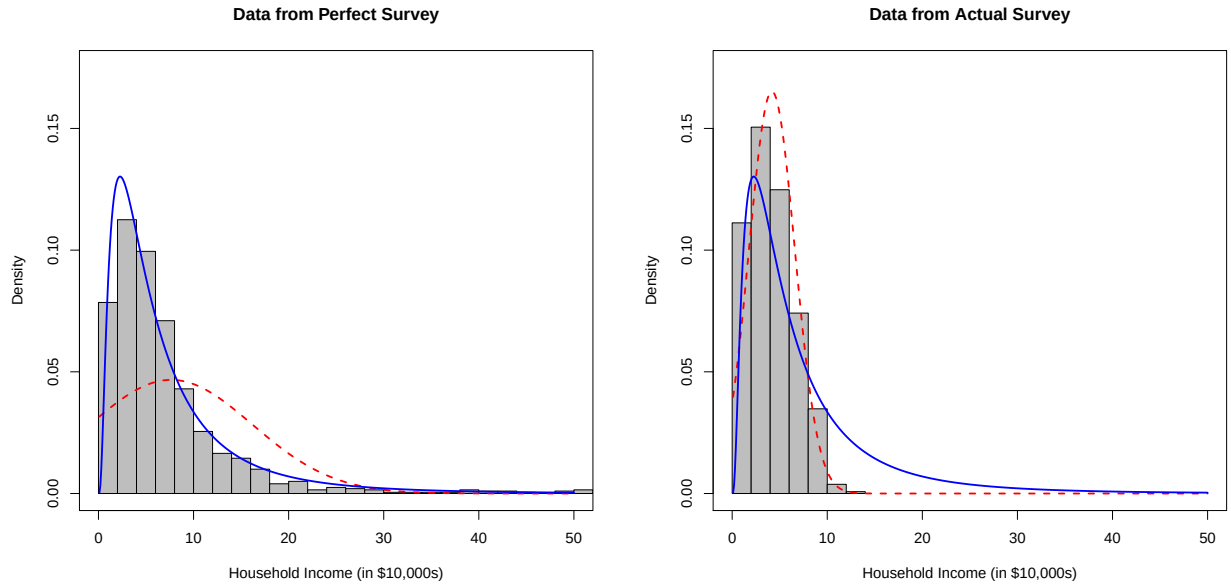
Figure 1.1 depicts a probability density with these features that is a likely representation of the true distribution of household incomes in the surveyed population. As might be expected, the true distribution of household incomes is extremely right skewed. Of course, in conducting the actual survey, we have no knowledge of what the true distribution of incomes looks like.

Figure 1.1: *Probability density of US household incomes with median $51,000 and Gini coefficient 0.48.*



If the density in Figure 1.1 were indeed that corresponding to the true distribution of household incomes, ***and*** if all 1,000 randomly chosen individuals were to respond to the survey question on household income (a "perfect survey"), then we would expect a histogram of the 1,000 survey responses to appear similar to the density in Figure 1.1. The left panel of Figure 1.2 shows what this might look like. Here, the sample median and mean are $50,556 and $74,060, respectively, with sample standard deviation $89,000. This figure also shows (red dashed line) a normal probability density with this mean and variance for comparison. Given the extreme skewness of the true distribution, the sample mean is predictably considerably larger than the sample median.

Figure 1.2: *Left panel: Histogram of ideal data that would have been collected from all 1000 survey participants. Right panel: Histogram of actual observed data collected from the subset of survey participants who responded. In each panel, the true income distribution is superimposed (blue solid) along with a normal distribution with the same mean and variance as that in the data (red dashed).*



In truth, only 66% (661) of the 1,000 survey participants responded to the question on household income. The right panel of Figure 1.2 shows a histogram of the reported household incomes for these respondents to the survey (the "actual survey"). The sample median and mean are $38,625 and $41,796, respectively, with standard deviation $24,152. Here, the mean and median are similar, and a normal distribution with this mean and variance (red dashed line) appears to be a reasonable description of the distribution of incomes among the respondents.

Armed with only the ***actual*** survey results, what can we say about the true distribution of household incomes in the population?

Suppose we were to assume that there is no difference in the distribution of income between individuals who respond to the survey and those who don't; that is, the probability that an individual responds to the survey does not depend on his/her income. In this case, we would view the 661 responses as a random sample from the population (albeit smaller than the one we set out to obtain) and would conclude that the distribution of household incomes in this population is approximately normally distributed, with median/mean around $40,000. Of course, with no knowledge of the true distribution of incomes and having data only from those who responded, we have no way of knowing if this is a reasonable assumption.

It is not far-fetched to think that individuals with higher incomes might be less likely to divulge their income amounts than those with lower incomes. Suppose we were to assume instead that the probability of nonresponse increases with income and in fact rises dramatically for incomes larger than $100,000. In this case, we would not view the 661 responses as a random (representative) sample from the population – under this assumption, individuals with lower incomes are clearly overrepresented in the sample.

The right hand panel of Figure 1.2 seems consistent with this assumption; however, again, we have no way of knowing if this is really the case. If we were to adopt this assumption, then clearly we would not believe that, for example, the raw sample median and mean from the actual survey results are believable estimates of the true median and mean in the population. To obtain more realistic estimates, we would need a way of incorporating this assumption with the actual survey data.

This example illustrates two main points:

- Different assumptions about the nature of nonresponse, more precisely, about the probability of nonresponse given income, will lead to different inferences about the true distribution of interest.
- Assumptions about the nature of nonresponse are unverifiable from the available data because we have no information in the data about the distribution of incomes among nonrespondents.

We can formalize the situation as follows. Let Y be a random variable denoting household income for an individual in this population. Let R be a random variable taking the value 1 if Y is observed for the individual and the value 0 if it is not.

Then it follows that

(i) If we were to assume that there is no difference in the distribution of income between individuals who respond to the survey and those who do not, we believe that

$$p_{Y|R}(y|R = 1) = p_{Y|R}(y|R = 0), \tag{1.1}$$

where $p_{Y|R}(\cdot|\cdot)$ is the density of income Y conditional on response status R. Because R takes on only the values 0 and 1, (1.1) states that Y and R are independent, and (1.1) can be expressed equivalently as

$$p_{R|Y}(r|y) \text{ does not depend on } y,$$

where $p_{R|Y}(r|y)$ is the density of R conditional on Y, so that $p_{R|Y}(r|y) = p_R(r)$, where $p_R(r)$ is the marginal density of R. That is, the probability that an individual responds to the survey does not depend on his/her income.

(ii) If we were to assume that the probability of nonresponse increases with income, then clearly we believe that Y and R are not independent and that

$$p_{R|Y}(r|y) \text{ depends on } y.$$

Under this assumption, we might try to specify the form of $p_{R|Y}(r|y)$ as a function of y reflecting our belief about nonresponse.

Of course, in either of (i) or (ii), the form of $p_{R|Y}(r|y)$ cannot be verified from the data.

As we will discuss shortly (i) corresponds to the case of data ***missing completely at random***, and (ii) corresponds to the case of data ***missing not at random***.

EXAMPLE 2: Dropout in a clinical trial. Consider a randomized clinical trial involving two treatments coded as 0 and 1 in which participants are randomized at ***baseline*** (time $t_1 = 0$) to receive one treatment or the other and then followed for a specified period of time. Subjects have the outcome of interest and other health status information measured at baseline, immediately prior to starting treatment, and then are to return to the clinic at additional specified times $t_2 < \cdots < t_T$ at which the outcome and other health information are to be ascertained.

Let Y_j be the outcome that is to be collected at time t_j, $j = 1, \ldots, T$, and let V_j be additional information collected at time t_j. Interest ordinarily focuses on the difference in mean outcome between treatments 0 and 1.

As noted previously, it is common in medical and pharmaceutical research for participants to ***drop out*** of such studies. By "drop out," we mean that a subject fails to show up for clinic visits after a certain point, so we are unable to obtain the required outcome measures after that point. (Dropout also occurs when subjects cease to comply with their assigned treatments, as mentioned earlier.)

More formally, we say that a subject is a ***dropout*** at time t_j if s/he appeared at the clinic at times $t_1, t_2, \ldots, t_{j-1}$ but then ceased to appear at time t_j onward. We assume in this example that all subjects are observed at baseline, so the first possible time at which dropout can take place is t_2. Thus, all information collected at baseline, (Y_1, V_1), is available for all subjects. If a subject is a dropout at time t_j, $j \geq 2$, $Y_1, \ldots, Y_{j-1}$ would be observed for the subject, but $Y_j, \ldots, Y_T$ would be ***missing***. Similarly, $V_1, \ldots, V_{j-1}$ would also be observed, but $V_j, \ldots, V_T$ would be missing.

It is often the case that interest focuses not only on the mean outcome at the final follow-up time t_T but also on the ***evolution*** of mean outcome over the entire follow-up period and how the pattern of mean outcome compares between the two treatments. This may be implemented by positing a regression model for mean outcome as a function of time, fitting the model to the data by some appropriate technique, and making inference on differences in the regression parameters. For example, it may be appropriate to assume a ***linear model*** in each treatment group $a = 0, 1$, such as

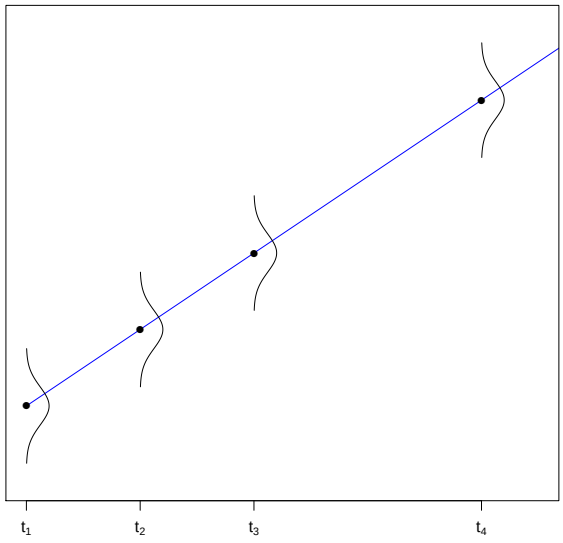
$$E(Y_j) = \beta_{0,a} + \beta_{1,a} t_j, \quad a = 0, 1, \tag{1.2}$$

in which case interest would focus on the difference in slopes $\beta_{1,1} - \beta_{1,0}$.

If the ***full data*** on outcome, $Y_1, \ldots, Y_T$, were collected on each study participant at the specified times $t_1, \ldots, t_T$, then inference on this comparison would be straightforward; one could use standard methods for fitting so-called population average models to longitudinal data for this purpose to obtain consistent estimators for $\beta_{1,1}$ and $\beta_{1,0}$ and their difference.

How is inference on the difference $\beta_{1,1} - \beta_{1,0}$ affected by dropout? To get a sense of this, consider treatment 1 and inference on $\beta_{1,1}$. Suppose ***in truth*** the relationship between mean outcome and time is a straight line as in (1.2) and Y_j is in fact normally distributed with the same variance at each time point. Figure 1.3 depicts this situation.

Figure 1.3: *Hypothetical outcome distributions over time for the clinical trial example in the case* $T = 4$.



Consider three cases:

- ***Case 1: Dropout is unrelated to outcome.*** Suppose that subjects who drop out of the study do so because they move to other states for work or family considerations that have nothing to do with health status. In this case, it is reasonable to believe that at each time t_j the distribution of outcomes among those who move and leave the study is the same as that for those who continue to participate.

  The size of the sample of subjects observed at each time point may be reduced from that intended, so that the precision of the estimator for $\beta_{1,1}$ may be lower than hoped, but a consistent estimator can still be obtained by applying the standard methods to the ***observed data***.

- ***Case 2: Dropout is related only to information that is observed.*** Suppose that subjects who drop out of the study at any time t_j, $j \geq 2$, do so based on the histories of their outcome $(Y_1, \ldots, Y_{j-1})$ and other information $(V_1, \ldots, V_{j-1})$ being collected in the study. A subject's physician may decide after review of his/her evolving outcome and other information to remove the subject from the study and place him/her on another treatment. For example, in a study comparing anti-hypertensive agents with systolic blood pressure as the outcome, this decision may be based not only on the subject's blood pressure readings but also on his/her heart rate, cholesterol levels, and other health status information.

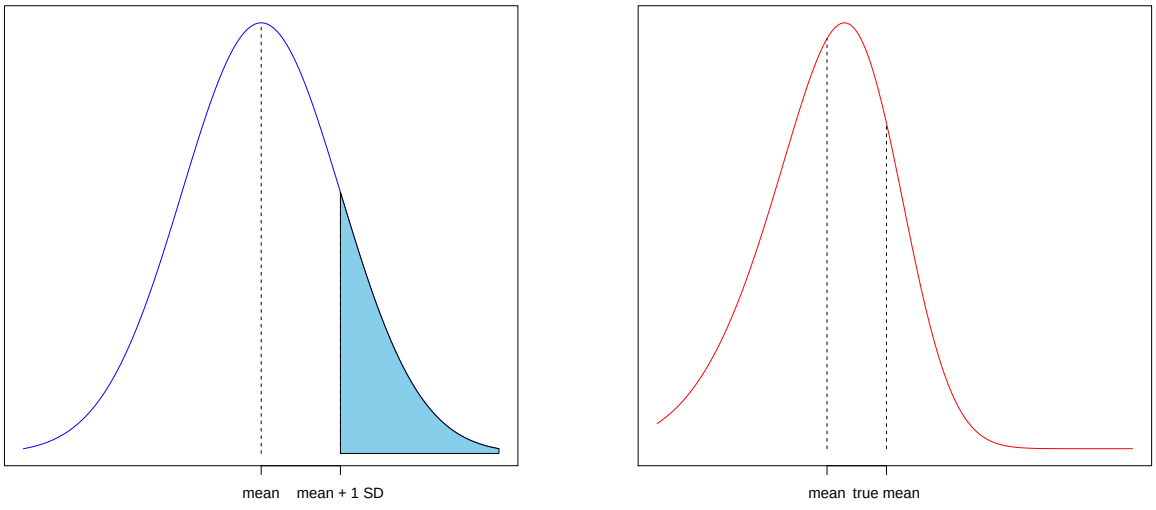
In this situation, for subjects with identical histories through time t_{j-1}, some drop out and some do not, so that, ***conditional*** on the history, whether or not a subject drops out is essentially at random. This implies that the distribution of outcomes that would be observed at time t_j and the distribution of outcomes that would be missing at t_j, ***conditional on*** the previously observed outcomes and other information, $\{(Y_1, V_1), \ldots, (Y_{j-1}, V_{j-1})\}$, are the ***same***.

Will the standard analysis applied to the observed data result in valid inferences on $\beta_{1,1}$ in this case?

For definiteness, consider the simple, albeit contrived, situation where the decision to drop out at t_j depends only on a subject's last observed outcome. In particular, suppose that, for a subject whose observed outcome at t_{j-1} is greater than the mean outcome at t_{j-1} + 1 standard deviation (SD), s/he will drop out of the study with probability 0.9 and continue with probability 0.1. A subject whose observed outcome at t_{j-1} is less than mean outcome at t_{j-1} + 1 SD does not drop out.

At baseline t_1, the distribution of outcomes in the population is normally distributed, as in the left panel of Figure 1.4. As shown, subjects for whom outcome at baseline is greater than mean + 1 SD, that is, for whom baseline outcome is in the shaded region, will drop out at t_2 with high probability (0.9).

Figure 1.4: *Outcome distributions at t_1 (baseline, left panel) and t_2 (right panel). Subjects with observed baseline outcome in the shaded region of the right panel drop out with probability 0.9.*



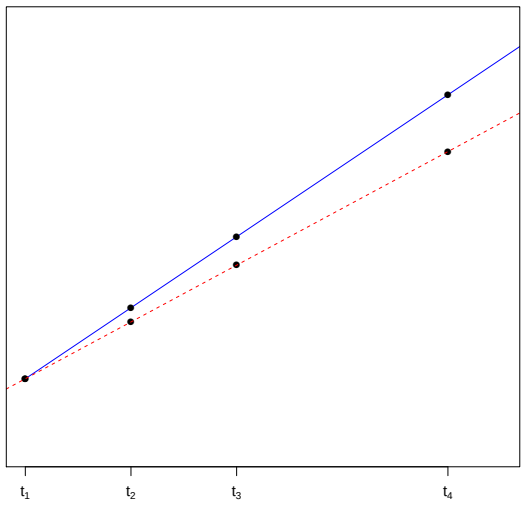
It is often the case that successive repeated outcomes on the same subject are highly correlated; if a subject's outcome is "high" at time $j-1$, it is "high" at time j. Thus, it is likely that the outcomes at t_2 that are not observed for the subjects who dropped out would have also been greater than mean + 1 SD at t_2.

The distribution of ***observed*** outcomes at t_2 would thus look like that depicted in the right panel of Figure 1.4. This is to be distinguished from the distribution of ***all*** outcomes at t_2, which as in Figure 1.3 is normally distributed. Relative to that distribution, the distribution of observed outcomes will be left-skewed as shown. Accordingly, an estimate of mean outcome at t_2 based on the observed data at t_2 will be smaller than the actual mean of the (normal) distribution of all outcomes at t_2, as in the figure.

If we extend this reasoning to dropout at $t_3, \ldots, t_T$, at each of these times the distribution of the ***observed*** outcomes will appear more and more skewed to the left as subjects whose observed outcomes at the previous time were greater than mean + 1 SD would be highly likely to have dropped out. Estimated mean outcome at each time based on the observed data will be smaller and smaller relative to the actual mean of the (normal) distribution of all outcomes at that time.

As a result, the apparent relationship of mean outcome to time based on the observed data will be ***attenuated*** relative to the true relationship, as shown in Figure 1.5. Accordingly, if we were to apply standard methods for fitting (1.2) to the observed data, the estimator for $\beta_{1,1}$ would be ***biased*** downward.

Figure 1.5: *True mean outcome over time (blue solid line) and apparent mean outcome over time based on the observed data under dropout (red dashed line).*



- ***Case 3: Dropout is related to information that is not observed.*** Now suppose that at each time t_j, $j \geq 2$, subjects who would have large outcomes Y_j at t_j are more likely to dropout. For example, in the anti-hypertension study above, a subject might take his blood pressure right before a scheduled clinic visit and, if it is high, decide to terminate his participation in the study, feeling the agent he is taking is not helping. More generally, suppose that at each t_j the probability of dropping out increases with the value of Y_j.

  In this scenario, dropout is related to information that is ***not observed*** in the sense that it will not be recorded in the database available to the analyst. Intuitively, if subjects with large values of the outcome are more likely to dropout, the apparent distribution of outcomes at each t_j, $j \geq 2$ will appear skewed to the left as in the previous setting, and estimation of the relationship of mean outcome to time based on the observed data will again be compromised.

  A litte thought suggests that this situation may be even more challenging than the previous one. There, the information implicated in dropout, the outcome observed at the last clinic visit, is available to the data analyst, raising the possibility that something could be done to "correct" the attenuated estimate of slope. Here, however, the information implicated in dropout is ***not available***, making this possibility seem fairly hopeless.

As in the previous example, we can represent these three situations formally in terms of suitable notation. As above, we assume that if a subject has not yet dropped out by t_j, we observe both Y_j and V_j at t_j, and if a subject does drop out at t_j, the pair (Y_k, V_k) is missing for $j \leq k \leq T$.

Let $R = (R_1, \ldots, R_T)$ be a vector whose elements correspond to each time point, such that

$$\begin{aligned} R_j &= 1 \quad \text{if } (Y_j, V_j) \text{ is observed} \\ &= 0 \quad \text{if } (Y_j, V_j) \text{ is missing,} \end{aligned}$$

$j = 1, \ldots, T$. Here, because all intended information is always observed at baseline, $R_1 = 1$ for all subjects. Moreover, because subjects who drop out never return, the vector R only can take on possible values $r^{(j)}$, $j = 1, \ldots, T$, of the form

$$r^{(1)} = (1, 0, \ldots, 0), \quad r^{(2)} = (1, 1, 0, \ldots, 0), \quad \ldots, \quad r^{(T)} = (1, 1, \ldots, 1). \tag{1.3}$$

That is, $r^{(j)}$ represents dropout at the $(j+1)th$ time point (so the subject is last seen at the jth time), where "dropout at time $(T+1)$" corresponds to never dropping out and completing all T clinic visits.

Write $Y = (Y_1, \ldots, Y_T)^T$ and $V = (V_1^T, \ldots, V_T^T)^T$. In principle, the probability of dropping out could depend on any aspect of Y and V, which are the data intended to be collected. With this notation, we can characterize each of the above cases in terms of this notation and the following conditions on the probability of observing dropout pattern $r^{(j)}$, $j = 1 \ldots, T$:

$$\begin{aligned}
&\text{Case 1:} \quad \text{pr}(R = r^{(j)} \mid Y, V) = \text{pr}(R = r^{(j)}) \\
&\text{Case 2:} \quad \text{pr}(R = r^{(j)} \mid Y, V) = \text{pr}(R = r^{(j)} \mid Y_1, \ldots, Y_j, V_1, \ldots, V_j) \\
&\text{Case 3:} \quad \text{pr}(R = r^{(j)} \mid Y, V) \text{ depends on } Y_{j+1}, \ldots, Y_T, V_{j+1}, \ldots, V_T.
\end{aligned}$$

As we will discuss shortly, each case may be characterized in terms of the following terminology:

Case 1 corresponds to data ***missing completely at random***
Case 2 corresponds to data ***missing at random***
Case 3 corresponds to data ***missing not at random***.

1.3 General framework and taxonomy of missing data mechanisms

The foregoing examples are special cases of the general framework for thinking about missing data problems that we now present. Throughout this course, we will often introduce models and methods in terms of the notation here and place special cases in this general context. We will consider problems where data are to be collected on each of N ***individuals***, which may be subjects in a clinical trial, households in a survey, plots in an agricultural experiment, and so on.

NOTATION FOR FULL DATA AND OBSERVATION INDICATOR: We let Z denote the ***full data*** intended to be collected on each individual, which can be partitioned into K components, i.e.,

$$Z = (Z_1, \ldots, Z_K), \tag{1.4}$$

each of which may be vector-valued. We assume that the elements of each component are either all missing or all observed. Define

$$R = (R_1, \ldots, R_K) \tag{1.5}$$

to be the K-dimensional vector of (scalar) indicators of whether or not the corresponding components of Z are ***observed***; that is, for $k = 1, \ldots, K$,

$$R_k = 1 \text{ if } Z_k \text{ is observed} \tag{1.6}$$

$$= 0 \text{ if } Z_k \text{ is missing.} \tag{1.7}$$

We allow the possibility that any of the components of Z in (1.4) may be observed or missing. Thus, it is clear that the vector R in (1.5) can take on 2^K possible values, where we denote a possible value of R as

$$r = (r_1, \ldots, r_K),$$

and the r_k are equal to 0 or 1. In some settings, as with dropout, the possible values of r may be restricted by the context.

For convenience, we also define

$$\overline{R} = (1 - R_1, \ldots, 1 - R_K), \tag{1.8}$$

the K-dimensional vector of indicators of whether or not the corresponding components of Z are ***missing***, with possible values $\bar{r}$.

The nature of Z and R depends on the particular setting:

- ***Regression.*** Suppose the full data to be collected are a scalar outcome Y and a $p-$dimensional vector of covariates X, and interest focuses on estimation of a parameter β in a parametric model $\mu(x;\beta)$ for $E(Y|X = x)$. Suppose that it is possible for Y to be missing, and that either the covariates are all observed or all are missing. Then $K = 2$, and

  $$Z = (Z_1, Z_2) = (Y, X). \tag{1.9}$$

  Alternatively, if it is possible for each covariate to be observed or missing regardless of the others, then $K = p + 1$, and

  $$Z = (Z_1, Z_2, \ldots, Z_{p+1}) = (Y, X_1, \ldots, X_p).$$

  In (1.9), there are $2^2 = 4$ possible values r: $(1,1)$, $(1,0)$, $(0,1)$, and $(0,0)$. If in fact it is the case that X is always observed, and only Y can be missing, then the possible values r are restricted to $(1,1)$ and $(0,1)$.

- ***Regression, continued.*** In the previous example, interest focuses on a model $\mu(x;\beta)$ for $E(Y|X = x)$. However, it may be that additional data V are collected. Although these data may not be relevant to the inference of interest, which involves the relationship between Y and X only, as we will discuss, they may be useful for rendering assumptions about the mechanism of missingness plausible. Suppose there is an additional q-dimensional vector of such variables and that it is possible for Y to be missing and that both covariates and additional variables are either all observed or all missing. Then $K = 3$, and

  $$Z = (Z_1, Z_2, Z_3) = (Y, X, V).$$

If each element of X and V could be missing or not, then Z would have $1 + p + q$ components analogous to the above.

- ***Longitudinal study.*** Suppose as in the dropout example there are T pre-specified times at which observations are to be taken on each participant in the study, where the first time corresponds to baseline. Suppose further that, at time j, a scalar outcome Y_j, a vector of covariates of interest X_j, and a vector of additional variables V_j are to be collected on each participant. If a participant shows up at time j, all of (Y_j, X_j, V_j) are observed; otherwise, all are missing. Here, $K = T$, $R = (R_1, \ldots, R_T)$, and

$$Z = (Z_1, \ldots, Z_T) = \{(Y_1, X_1, V_1), (Y_2, X_2, V_2), \ldots, (Y_T, X_T, V_T)\}.$$

  If the only pattern of missingness possible is that of dropout, as in the example above, and all subjects are observed at baseline, then the values r are restricted to the T possible values

$$(1, 0, \ldots, 0), \quad (1, 1, 0, \ldots, 0), \quad \ldots, \quad (1, 1, \ldots, 1) \tag{1.10}$$

  as in (1.3). Under this condition, missingness is referred to as ***monotone***, for obvious reasons.

  If participants show up sporadically, ***intermittent*** or ***nonmonotone*** patterns of missingness are possible; for example, if $T = 5$, $r = (1, 1, 0, 0, 1)$. In principle, if it is possible for all information to be missing at baseline, then there are 2^T possible patterns.

NOTATION FOR OBSERVED DATA: Given a particular definition of Z,there are two types of notation for the ***observed data*** seen under possible missingness of some components of Z.

- For a specific pattern of missingness r, we write $Z_{(r)}$ to denote the subset of components of Z that is observed under the ***given missingness pattern*** r and $Z_{(\bar{r})}$ to denote the subset that is missing.

  For example, if $K = 3$, so that $Z = (Z_1, Z_2, Z_3)$, with $r = (1, 0, 1)$, $Z_{(r)} = (Z_1, Z_3)$, and $Z_{(\bar{r})} = Z_2$.

- With these definitions, it should be clear that the data that are available when some data are missing may be written as

$$(R, Z_{(R)}), \qquad \text{where} \qquad Z_{(R)} = \sum_r Z_{(r)}\, I(R = r). \tag{1.11}$$

  We refer to (1.11) as the ***observed data***. It is important to recognize that, because R is a random vector, $Z_{(R)}$ is not the same as $Z_{(r)}$ for a fixed, particular value r. Rather, $Z_{(R)}$ depends on R, and thus both R and $Z_{(R)}$ are necessary to characterize fully which components of Z are observed for a randomly chosen individual.

- In the missing data literature, it is standard to refer to components of Z that may be observed and missing by rearranging and partitioning Z as

$$Z = (Z^{obs}, Z^{mis}),$$

where Z^{obs} comprises the components of Z that are observed and Z^{mis} comprises those that are missing.

This notation is appealing for its simplicity, but it is not precise. It is important to recognize that Z^{obs} and Z^{mis} are ***not*** of fixed dimension. In the context of the above notation, for a randomly chosen individual with pattern R, one can think loosely of

$$Z^{obs} = Z_{(R)} \quad \text{and} \quad Z^{mis} = Z_{(\overline{R})}.$$

However, although it is tempting to refer to Z^{obs} alone as the observed data, this is not strictly correct; as noted above, both R and $Z_{(R)}$ are required to characterize fully the ***observed data*** on a randomly chosen individual, as in (1.11).

In this course, we will be precise and use the notation in (1.11) to refer to the ***observed data***.

DROPOUT: In the special case of a longitudinal study in which missingness arises because of ***dropout***, the following notation will sometimes be convenient. If full data $Z = (Z_1, \ldots, Z_T)$ are to be collected at each of a series of T prespecified times $t_1 < \cdots < t_T$, define

$$D = 1 + \sum_{j=1}^{T} R_j. \tag{1.12}$$

The random variable D defined in (1.12) thus represents the time at which dropout occurs. E.g., if $D = j$, then a subject is observed at times $t_1, \ldots, t_{j-1}$ ($R_1 = \cdots = R_{j-1} = 1$) and is a dropout at time t_j ($R_j = \cdots = R_T = 0$). In this case, for fixed j, we modify the definition above and write $Z_{(j)} = (Z_1, \ldots, Z_{j-1})$ to denote components of Z observed under dropout at time j. Note that $D = T + 1$ corresponds to observing the full data Z.

In this case, the ***observed data*** are $(D, Z_{(D)})$.

SAMPLE DATA: For a sample of N (randomly sampled) individuals from the population, we index individuals by $i = 1, \ldots, N$.

- Denote the full data that could potentially be collected on individual i by Z_i, with components $Z_{i1}, Z_{i2}, \ldots, Z_{iK}$. Thus, if available, the full data for the sample would be ***independent and identically distributed*** (iid) Z_i, $i = 1, \ldots, N$. We sometimes use the shorthand notation $\underset{\sim}{Z} = \{Z_i, i = 1, \ldots, N\}$ to denote the full sample data.
- Denote the observation indicator for individual i by R_i (analogously, D_i in the case of dropout).
- If some data are missing, from (1.11), the observed data for the sample would be iid $(R_i, Z_{(R_i)i})$, $i = 1, \ldots, N$ (analogously, $(D_i, Z_{(D_i)i})$). Write $(\underset{\sim}{R}, \underset{\sim}{Z}_{(\underset{\sim}{R})})$, for example, to denote the observed sample data.

INFERENCE: Armed with this notation, we may now state precisely the inferential problem.

The full data Z are the data ***intended*** to be collected, and the objective is to make inference on some aspect of the distribution of Z. Let $p_Z(z)$ denote the probability density of Z.

- Interest may focus on the ***mean*** or other characteristic of one or more components of Z. In a longitudinal study with $Z = (Y_1, \ldots, Y_T)$ for real-valued outcome Y, estimation of $E(Y_T)$, mean outcome at the final time point, or the covariance matrix of $(Y_1, \ldots, Y_T)^T$ may be important.
- In many settings, a ***parametric*** model may be adopted. Recall that a statistical model is a class of probability distributions that is assumed to have generated the data (and thus is assumed to contain the true density). Interest may focus on parameters that characterize this class of assumed probability distributions.

  If a model $p_Z(z;\theta)$, is posited to describe $p_Z(z)$, where θ is a finite-dimensional vector of parameters, then the goal is to make inference on θ or, partitioning θ as $(\beta^T, \eta^T)^T$, say, on a subset β of elements of θ.

  As a simple example, if the components of $Z = (Z_1, \ldots, Z_K)$ are each real-valued random variables, Z might be assumed to be K-variate normal, with mean vector β and covariance matrix with distinct elements η.

- Instead of a fully parametric model, a ***semiparametric*** model may be adopted for $p_Z(z)$. Here, the class of probability distributions assumed is described by both finite-dimensional parametric and infinite-dimensional components. A member of the assumed class of probability densities may be written $p_Z\{z; \beta, \eta(\cdot)\}$, where β is a finite-dimensional vector of parameters and $\eta(\cdot)$ is an infinite-dimensional component. Interest ordinarily focuses on inference on β.

  For example, if $Z = (Y, X)$, one may posit a parametric model $\mu(x; \beta)$ for $E(Y|X = x)$, leaving the rest of the joint density ***unspecified***.

If the full data Z_i, $i = 1, \ldots, N$, were available for all N individuals, inference would proceed using methods appropriate for the type of full data model adopted and the specific goal (e.g., inference on β). However, when some components of Z may be missing, the challenge is to how to achieve valid inference using the observed data $(R_i, Z_{(R_i)i})$, $i = 1, \ldots, N$.

MISSING DATA MECHANISMS: We now present general, formal definitions of ***missing data mechanisms***. The terminology and taxonomy of these mechanisms are given in a seminal *Biometrika* paper by Don Rubin (Rubin, 1976).

Recall as above that, in principle, the probability of a particular missingness pattern can depend on aspects of the full data intended to be collected, Z. The following definitions formalize this.

- ***Missing Completely at Random (MCAR)***. Data are ***missing completely at random*** if

$$\text{pr}(R = r|Z) \text{ does not depend on } Z. \tag{1.13}$$

  MCAR as in (1.13) states that R and Z are ***independent***, which we write as

$$R \perp\!\!\!\perp Z.$$

  Under MCAR as in (1.13), the probabilities of the possible missingness patterns are ***constants***; i.e.,

$$\text{pr}(R = r|Z) = \pi(r),$$

  where $\pi(r)$ is a constant for each possible value r.

  Situations in which MCAR (1.13) is plausible are those in which it is clear that missingness has nothing to do with the issues under study, as in the previous example in which subjects in a clinical trial drop out because they move away for work or family considerations. In the surrogate measurement example at the beginning of this chapter, if the validation sample is constructed by selecting members from the entire study sample with probability 0.1, say, then missingness of the expensive diary measurement is completely at random.

In settings like the regression example in (1.9), in which interest focuses on the relationship between outcome Y and covariates X through a regression model for $E(Y|X = x)$, the inference of interest is ***conditional*** on X. Here, when X is always observed and only Y may be missing, a convention that is sometimes adopted is to refer to the case where

$$\text{pr}(R = r|Z) = \text{pr}(R = r|Y, X) = \text{pr}(R = r|X) = \pi(r, X), \tag{1.14}$$

so that missingness depends only on the always-observed X, as (conditional) MCAR. In (1.14), R is independent of Z ***conditional*** on X, which we write as

$$R \perp\!\!\!\perp Z \mid X.$$

Because the desired inferences focus on aspects of the conditional distribution of Y given X, this conditional independence is analogous to the strict independence implied by (1.13).

In this course, we do not adopt this convention and restrict our notion of MCAR to refer to the case where the probabilities in (1.13) are constants. Accordingly, we consider the situation in (1.14) to be a special case of ***missing at random***, defined next.

- ***Missing at Random (MAR)***. Data are said to be ***missing at random*** if

$$\text{pr}(R = r|Z) = \text{pr}(R = r|Z_{(r)}) = \pi(r, Z_{(r)}), \tag{1.15}$$

say, so that the probability of missingness pattern r as a function of Z depends only on the components of Z that ***are observed*** under r.

In the particular case of ***dropout*** in a longitudinal study, as described above, MAR has a simple interpretation. Here, possible values of r are as defined in (1.3); accordingly, for any of these possible values $r^{(j)}$, there is a corresponding value of j, and we can rewrite (1.15) in terms of D defined in (1.12). For $r^{(j)}$ corresponding to j, we can write $Z_{(r^{(j)})} = (Z_1, \ldots, Z_j)$ and $Z_{(\bar{r}^{(j)})} = (Z_{j+1}, \ldots, Z_T)$.

With these definitions, it can be shown that we can write (1.15) with $r = r^{(j)}$ equivalently as

$$\text{pr}(D = j + 1|Z) = \text{pr}(D = j + 1|Z_{(r^{(j)})}, Z_{(\bar{r}^{(j)})}) = \text{pr}(D = j + 1|Z_{(r^{(j)})}). \tag{1.16}$$

This makes explicit that it is possible to think of MAR as implying that dropout is a ***sequentially random*** process; whether or not a subject still in the study drops out at time j is at random depending only on the history of observed data to that point.

In the surrogate measurement example, if the validation sample is constructed by selecting members from the entire study sample depending on whether or not their 24-hour recall measurements exceed some threshold, then the missingness of the expensive diary measurement depends on a variable that is always observed, and hence is MAR. Here, as noted earlier, MAR is ***by design***.

Later in this chapter, we consider how, when MAR is thought to hold, the observed data may be used to "correct" for missingness.

- ***Missing Not at Random (MNAR)***. Data are said to be ***missing not at random*** if

$$\text{pr}(R = r|Z) \text{ depends on components of } Z \text{ not observed when } R = r. \quad (1.17)$$

Under MNAR (1.17), missingness depends on data that are not observed. Intuitively, if this mechanism governs the missingness, it does not seem possible to use the observed data to "correct" for missingness. We will discuss this in detail later in the course.

FUNDAMENTAL CHALLENGE: This taxonomy of missing data mechanisms clarifies the extent to which missing data can complicate inference. Obviously, MCAR is the least problematic, while MNAR seems to pose a significant obstacle, and MAR is somewhere in between.

The central challenge is that it is ***not possible*** to test if a particular mechanism holds based on the observed data. As a consequence, the data analyst must adopt an ***assumption*** about the mechanism ***without being able to verify*** its plausibility from the data.

- Clearly, it is not possible to distinguish between MNAR and MAR from the observed data only. Intuitively, because MNAR mechanisms depend on data that are not observed, it is impossible to test if a mechanism that depends on data that may be missing is a more plausible explanation for why some observations are missing than one that depends only on observed data.
- Thus, if one adopts an assumption of MAR, it must be defensible on scientific, subject matter, and/or practical grounds, because it cannot be validated from the data.
- If one is willing to assume a MAR mechanism, it ***is*** possible to test if the mechanism is in fact MCAR; i.e., if one assumes the mechanism depends only on observed data, it is clearly possible to assess the strength of that dependence from the observed data. ***However***, this is predicated on the MAR assumption holding, which is itself ***not*** verifiable.

- If one is willing to adopt an assumption of MAR, as we will discuss in the next four chapters, several general approaches are available to facilitate desired inferences based on the observed data, which we will cover in detail. Of course, the validity of these inferences is predicated on the MAR assumption holding.
- If one is unwilling or unable to justify a MAR assumption, implying the belief that the mechanism is MNAR, as we will discuss, inference based on the observed data involves even more serious challenges.

When the individuals are ***humans***, unless missingness is by design as in the surrogate measurement example, MCAR is not a realistic assumption in general. The major issue is willingness to assume that the missingness mechanism is MAR. To justify a MAR assumption, sufficiently rich information that is ideally always observed must be collected and be available to the data analyst.

For example, if $Z = (Y, X, V)$ and interest focuses on β in a model $\mu(x;\beta)$ for $E(Y|X = x)$, and Y and/or X may be missing, collection of additional information V that is always observed and that may be associated with missingness can render MAR plausible.

For these reasons, recent guidelines for handling missing data in clinical trials published by the National Research Council (2010) emphasize taking steps at the design stage to prevent missing data and, recognizing that some missing data are unavoidable, collecting rich information that can be used to support MAR.

1.4 More examples

We are now in a position to explore more precisely the consequences of missing data for inference. We consider three common data-analytic situations.

EXAMPLE 1. Estimation of the mean. Suppose that the full data are simply $Z = Y$, where Y is a real-valued outcome random variable, and the objective is to estimate $\mu = E(Y)$ based on a random sample of N individuals. If full data were available on all individuals, Y_i, $i = 1, \ldots, N$, the obvious estimator is the sample mean

$$\widehat{\mu}^{full} = N^{-1} \sum_{i=1}^{N} Y_i.$$

Suppose that Y is missing for some individuals, where $R = 1$ if Y is observed and $R = 0$ if it is missing.

In this case, Y is available only if $R = 1$, and the observed data $(R, Z_{(R)})$ can be written as (R, RY). The observed data from the N individuals can then be written as

$$(R_i, R_i Y_i), \quad i = 1, \ldots, N.$$

A natural estimator for μ based on the observed data is the sample mean of the observed outcomes,

$$\widehat{\mu}^c = \frac{\sum_{i=1}^N R_i Y_i}{\sum_{i=1}^N R_i}, \tag{1.18}$$

where the superscript "c" emphasizes that (1.18) depends only on the so-called ***complete cases***; i.e., the individuals for whom the full data are entirely observed.

What are the properties of $\widehat{\mu}^c$ in (1.18) under different missingness mechanisms?

In this simple situation, R takes on only two values, 1 or 0. First consider MCAR. Under MCAR,

$$\text{pr}(R = 1|Y) = \text{pr}(R = 1) = \pi, \tag{1.19}$$

where $\pi > 0$ is a constant, and thus $\text{pr}(R = 0) = 1 - \pi$. Thus, the missingness probabilities do not depend on Y, and $R \perp\!\!\!\perp Y$.

It follows that, as $N \to \infty$, by the ***weak law of large numbers***,

$$\widehat{\mu}^c = \frac{N^{-1} \sum_{i=1}^N R_i Y_i}{N^{-1} \sum_{i=1}^N R_i} \xrightarrow{p} \frac{E(RY)}{E(R)} = \frac{E(R)E(Y)}{E(R)} = \mu$$

by the independence of R and Y. Thus, if the missingness mechanism is MCAR, the "complete case estimator" (1.18) is a ***consistent*** estimator for μ. This coincides with the intuition we discussed earlier: under MCAR, where missingness of Y has nothing to do with Y, the sample for whom Y is observed is a smaller but still representative sample from the population, so we expect $\widehat{\mu}^c$ to yield valid (albeit less precise than hoped) inference on μ.

In this simple setting, the only other possible missingness mechanism is MNAR, under which

$$\text{pr}(R = 1|Y) = \pi(Y), \tag{1.20}$$

say. Under (1.20), using $E(R|Y) = \text{pr}(R = 1|Y) = \pi(Y)$, as $N \to \infty$,

$$\widehat{\mu}^c = \frac{N^{-1} \sum_{i=1}^N R_i Y_i}{N^{-1} \sum_{i=1}^N R_i} \xrightarrow{p} \frac{E(RY)}{E(R)} = \frac{E\{\, E(RY|Y)\,\}}{E\{\, E(R|Y)\,\}} = \frac{E\{\, Y\, E(R|Y)\,\}}{E\{\, E(R|Y)\,\}} = \frac{E\{Y\,\pi(Y)\}}{E\{\pi(Y)\}} \neq E(Y) = \mu$$

in general.

In fact, if $\pi(y)$ is an ***increasing*** function of y, so that the probability of observing Y increases with y, then (why?)

$$\frac{E\{\, Y\, \pi(Y)\}}{E\{\pi(Y)\}} > \mu.$$

These calculations show that, under MNAR, the complete case estimator $\widehat{\mu}^c$ need not be a consistent estimator for μ. The profound difficulty is that, because Y is missing when $R = 0$, there is no way to model and estimate $\pi(y) = \text{pr}(R = 1|Y = y)$ from the observed data. In fact, based only on the observed data, we cannot identify if the missingness mechanism is MCAR or MNAR. Thus, it seems hopeless to find an alternative estimator that would be consistent.

Now suppose that, ***in addition*** to Y, a set of variables V is collected on all individuals, so that the full data are now $Z = (Z_1, Z_2) = (Y, V)$. Suppose further that V is ***always*** observed, even if Y is missing.

Here, $R = (R_1, R_2)$ can take on only the two values $(1, 1)$ and $(0, 1)$. Let $C = 1$ if $R = (1, 1)$ and $C = 0$ if $R = (0, 1)$. Here, C is an indicator of observing a ***complete case***. Then we may summarize the observed data $(R, Z_{(R)})$ equivalently as (C, CY, V), and the observed data on the N individuals are

$$(C_i, C_i Y_i, V_i), \quad i = 1, \ldots, N.$$

Suppose we are willing to assume that missingness of Y depends only on V and not on Y, i.e.,

$$\text{pr}(C = 1|Y, V) = \text{pr}(C = 1|V) = \pi(V), \tag{1.21}$$

where $\pi(v) > 0$ for all v, so that $\text{pr}(C = 0|V) = 1 - \pi(V)$, and thus $C \perp\!\!\!\perp Y|V$ and indeed $R \perp\!\!\!\perp Y|V$. This missingness mechanism is MAR, depending on data that are ***always observed***.

Note that V may be related to both Y and C, in which case, were V not available, it still might be that

$$\text{pr}(C = 1|Y) \quad \text{depends on } Y.$$

Thus, collection of the additional information V facilitates the assumption of MAR.

We now consider the behavior of $\widehat{\mu}^c$ if (1.21) holds. Under (1.21), as $N \to \infty$,

$$\widehat{\mu}^c = \frac{N^{-1}\sum_{i=1}^N C_i Y_i}{N^{-1}\sum_{i=1}^N C_i} \xrightarrow{p} \frac{E(CY)}{E(C)} = \frac{E\{\, E(CY|Y, V)\,\}}{E\{\, E(C|Y, V)\,\}} = \frac{E\{\, YE(C|Y, V)\,\}}{E\{\, E(C|Y, V)\,\}} = \frac{E\{Y\pi(V)\}}{E\{\pi(V)\}} \neq E(Y) = \mu$$

in general, because Y and V may be correlated. Thus, under MAR, $\widehat{\mu}^c$ is not necessarily a consistent estimator for μ.

Thus, if missingness is not MCAR, the sample mean based on the observed data only is ***not*** a consistent estimator for μ in general.

However, a fundamental difference between MAR and MNAR is that, because V is available, it is possible to construct alternative estimators for μ that ***are consistent*** under this form of MAR. To get a sense of this, consider the following intuitive argument.

For a randomly chosen individual having a particular value of V, by (1.21), the probability that Y is observed for the individual (equivalently, the probability of being a ***complete case***), is $\pi(V)$.

Thus, the individual can be thought of as representing $1/\pi(V)$ randomly chosen individuals, some of whom will have Y missing. Consider the estimator

$$\widehat{\mu}^{ipw} = N^{-1} \sum_{i=1}^{N} \frac{C_i Y_i}{\pi(V_i)}. \tag{1.22}$$

The superscript "*ipw*" stands for ***inverse probability weighting***, acknowledging that contributions in (1.22) from complete cases are ***weighted*** by the reciprocal (inverse) of the probability of being a complete case so that they represent themselves and others like them for whom Y may be missing.

The behavior of (1.22) as $N \to \infty$ may be deduced as follows:

$$\begin{aligned} \widehat{\mu}^{ipw} &= N^{-1} \sum_{i=1}^{N} \frac{C_i Y_i}{\pi(V_i)} \xrightarrow{p} E\left\{\frac{CY}{\pi(V)}\right\} = E\left[E\left\{\frac{CY}{\pi(V)}\,\middle|\, Y, V\right\}\right] = E\left\{\frac{Y}{\pi(V)} E(C \mid Y, V)\right\} \\ &= E\left\{\frac{Y}{\pi(V)}\pi(V)\right\} = \mu; \end{aligned}$$

here, $\pi(V)/\pi(V)$ is equal to 1 because $\pi(v) > 0$ for all v. Thus, $\widehat{\mu}^{ipw}$ is a consistent estimator for μ.

This argument demonstrates two main points:

- When the missingness mechanism is MAR, it is possible to construct estimators based on the observed data that "correct" for the missingness; i.e., that are ***consistent*** estimators for population quantities of interest.
- ***Inverse probability weighting of complete cases*** is one general approach to doing so. We will discuss this in detail in Chapter 5 of this course.

NOTATIONAL CONVENTIONS: The foregoing example is perhaps the simplest setting possible in which to discuss missing data. To facilitate our discussion of missing data in more complex statistical models in the rest of this course, we need appropriate notation, as demonstrated by the next two examples and in Section 1.5. Accordingly, before introducing this material, we describe the ***notational conventions*** we adopt henceforth.

- Many problems we will consider involve a posited ***parametric model*** for the probability density of the ***full data*** Z in terms of a finite dimensional parameter θ, $p_Z(z;\theta)$. Suppose θ is $(p \times 1)$.

  For any real-valued function $f(z;\theta)$, say,

  $$\frac{\partial}{\partial\theta}\{f(z;\theta)\} \tag{1.23}$$

  denotes the $(p \times 1)$ vector of partial derivatives of $f(z;\theta)$ with respect to the components of θ. Technically, if we wish to discuss (1.23) evaluated at a particular value of θ, θ^*, say, we should write

  $$\frac{\partial}{\partial\theta}\{f(z;\theta)\}\bigg|_{\theta=\theta^*}. \tag{1.24}$$

  As shorthand, we will often write this as

  $$\frac{\partial}{\partial\theta}\{f(z;\theta^*)\} \tag{1.25}$$

  It should be understood from (1.25) that we mean that differentiation is with respect to the second argument of $f(z;\theta)$, and then θ^* is substituted into the result.

  Thus, when we define, for example,

  $$g(z;\theta) = \frac{\partial}{\partial\theta}\{f(z;\theta)\} \quad (p \times 1),$$

  this refers to the $(p\times 1)$ vector function of θ that is the consequence of differentiation with respect to the second argument as in (1.23), and $g(z;\theta^*)$ is then shorthand for (1.24).

  Similarly,

  $$G(z;\theta) = \frac{\partial^2}{\partial\theta\partial\theta^T}\{f(z;\theta)\}$$

  denotes the $(p\times p)$ matrix of second partial derivatives of $f(z;\theta)$ with respect to the components of θ, and

  $$\frac{\partial^2}{\partial\theta\partial\theta^T}\{f(z;\theta^*)\}$$

  and $G(z;\theta^*)$ are defined analogously.

- We will often need to take care in discussing expectations. We will use notation, shown here for the case Z is continuous,

  $$E_\theta\{f(Z;\theta^*)\} = \int f(z;\theta^*)\, p_Z(z;\theta)\, dz; \tag{1.26}$$

  (1.26) allows the possibility that expectation is with respect to the density of Z evaluated at a value θ, whereas $f(z;\theta^*)$ is evaluated at a possibly different value θ^*. $E_\theta\{f(z;\theta)\}$ refers to the case where this value is the same.

For a posited model $p_Z(z;\theta)$, we will often assume that it is ***correctly specified***, by which we mean that there is a value θ_0 such that $p_Z(z;\theta_0)$ is the true density of Z. We will often write, for a function $f(z)$, say,

$$E\{f(Z)\} = E_{\theta_0}\{f(Z)\} = \int f(z)\, p_Z(z;\theta_0)\, dz,$$

dropping the subscript θ_0, when the expectation is with respect to the true density of Z. If $f(z)$ is also indexed by θ, i.e., $f(z;\theta)$, as above,

$$E\{f(Z;\theta)\} = E_{\theta_0}\{f(Z;\theta)\} = \int f(z;\theta)\, p_Z(z;\theta_0)\, dz,$$

where θ need not be equal to θ_0.

- All of the above conventions apply more generally to similar quantities and will be applied without comment henceforth. Occasionally, when what is meant is clear from the context, we may use less formal notation to streamline the presentation.

We now consider two examples in which we make use of these conventions.

EXAMPLE 2. Missingness in regression analysis. Consider the regression situation in (1.9) with full data $Z = (Y, X)$, where Y is a scalar outcome and X is a vector of covariates.

Suppose that it is possible for Y to be missing, and all of X is either observed or missing, so that $R = (R_1, R_2)$, where $R_1 = 1$ if Y is observed and 0 if it is missing, and $R_2 = 1$ if X is observed and 0 if it is missing. There are thus four possible missingness patterns r:

$$r = (1,1),\ \text{both } Y \text{ and } X \text{ observed}; \quad r = (1,0),\ Y \text{ observed}, X \text{ missing};$$

$$r = (0,1),\ Y \text{ missing}, X \text{ observed}; \quad r = (0,0),\ \text{both } Y \text{ and } X \text{ missing}.$$

Write

$$\pi\{r, (Y,X)\} = \text{pr}(R = r \,|\, Y, X).$$

Suppose further that interest focuses on inference on β in the regression model

$$E(Y|X = x) = \mu(x;\beta), \quad \beta\ (p \times 1),$$

which we assume here is ***correctly specified***, so that, as above, that there exists β_0 for which $\mu(x;\beta_0)$ coincides with the true function $E(Y|X = x)$ of x. We make no other assumptions about the distribution of the full data Z; thus, this is a ***semiparametric*** model.

A ***complete case*** in this setting is an individual for whom both Y and X are observed; i.e., for whom $R = (1,1)$. By default, when an analysis data set contains missing observations, most statistical software, such as SAS, will ***disregard*** data records where one or more analysis variables are missing and carry out the analysis based only on the complete cases.

What is the implication of this default approach for inference on β?

Here, the observed data from a sample of N individuals are

$$(R_i, R_{1i}Y_i, R_{2i}X_i), \quad i = 1, \ldots, N.$$

Suppose the software solves the usual ***estimating equation*** for ordinary least squares (OLS). If based only on the complete cases, the $(p \times 1)$ estimating equation being solved is

$$\sum_{i=1}^{N} I\{R_i = (1,1)\} \frac{\partial}{\partial \beta}\{\mu(X_i;\beta)\}\{Y_i - \mu(X_i;\beta)\} = 0, \tag{1.27}$$

where $I(\cdot)$ is the indicator function such that $I(A) = 1$ if the event A is true and 0 otherwise.

As is well-known in the study of regression, if, under regularity conditions, an ***estimating equation*** like (1.27) is such that

$$E_\beta \left[I\{R = (1,1)\} \frac{\partial}{\partial \beta}\{\mu(X;\beta)\}\{Y - \mu(X;\beta)\} \right] = 0 \tag{1.28}$$

for all β, then the estimating equation is said to be ***unbiased***. If (1.28) holds at $\beta = \beta_0$, then the complete case estimator $\widehat{\beta}^c$ obtained from solving (1.27) will converge in probability as $N \to \infty$ to β_0; i.e., $\widehat{\beta}^c$ is ***consistent***.

We discuss this situation in generality in Section 1.5.

By a conditioning argument, conditioning on (Y, X) inside the expectation in (1.28) (try it), (1.28) can be rewritten as

$$E_\beta \left[\pi\{(1,1),(Y,X)\} \frac{\partial}{\partial \beta}\{\mu(X;\beta)\}\{Y - \mu(X;\beta)\} \right] = 0. \tag{1.29}$$

Thus, examining consistency of the solution $\widehat{\beta}^c$ to (1.27) under different assumptions on the missingness mechanism boils down to examining the root of (1.29).

Case 1. $\pi\{(1,1),(Y,X)\}$ ***depends only on*** X**.** Here, we can write $\pi\{(1,1),X)\}$, and the left hand side of (1.29) becomes

$$\begin{aligned} E_\beta &\left[\pi\{(1,1),X\} \frac{\partial}{\partial\beta}\{\mu(X;\beta)\}\{Y-\mu(X;\beta)\} \right] \\ &= E\left(E_\beta \left[\pi\{(1,1),X\} \frac{\partial}{\partial\beta}\{\mu(X;\beta)\}\{Y-\mu(X;\beta)\} \,|X\right] \right) \\ &= E\left(\pi\{(1,1),X\} \frac{\partial}{\partial\beta}\{\mu(X;\beta)\} E_\beta\left[\{Y-\mu(X;\beta)\}\,|X\right] \right) \end{aligned} \tag{1.30}$$

Because $\mu(x;\beta)$ is ***correctly specified***, setting $\beta = \beta_0$ gives

$$E\left[\{Y-\mu(X;\beta_0)\}\,|X\right] = 0,$$

which implies (1.29) holds with $\beta = \beta_0$. Thus, $\widehat{\beta}^c$ is a consistent estimator.

Note that in the foregoing development we did not specify the missingness mechanism. We now examine this issue, continuing to assume that $\pi\{(1,1),(Y,X)\}$ depends only on X.

- Suppose that X can be missing but Y is always observed. Then

  $$\pi\{(0,1),X\} = \pi\{(0,0),X\} = 0 \quad \text{because } Y \text{ is always observed.}$$

  Because it must be that $\sum_r \pi(r,X) = 1$, where this sum is over the four possible missingness patterns, it follows that

  $$\pi\{(1,0),X\} = 1 - \pi\{(1,1),X\} \quad \text{depends on } X. \tag{1.31}$$

  From (1.31), the probability of observing the missingness pattern where Y is observed but X is missing depends on the unobserved X; thus, the missingness mechanism is MNAR.

  Thus, the complete case estimator is consistent even though the missing data are MNAR.

- Suppose conversely that Y can be missing but X is always observed. Then

  $$\pi\{(1,0),X\} = \pi\{(0,0),X\} = 0 \quad \text{because } X \text{ is always observed,}$$

  $$\pi\{(0,1),X\} = 1 - \pi\{(1,1),X\} \quad \text{depends on } X.$$

  Here, the probability of observing the missingness pattern where X is observed but Y is missing depends only on X, which is itself always observed. Thus, the missingness mechanism is MAR.

  This shows that the complete case estimator is consistent under MAR.

In conclusion, ***regardless*** of whether or not the missingness mechanism is MAR or MNAR, the complete case OLS estimator will be consistent if $\pi\{(1,1),(Y,X)\}$ depends only on X.

One should ***not*** conclude from this that the complete case analysis always yields valid inferences, as the next case shows.

Case 2. $\pi\{(1,1),(Y,X)\}$ ***depends on*** Y***.*** Under this condition, the probability of observing a complete case depends on Y alone or on both Y and X. In general, we can write the left hand side of (1.29) as

$$\begin{aligned} & E_\beta\left[\pi\{(1,1),(Y,X)\}\,\frac{\partial}{\partial\beta}\{\mu(X;\beta)\}\{Y-\mu(X;\beta)\}\right] \\ & \qquad = E\left(E_\beta\left[\pi\{(1,1),(Y,X)\}\,\frac{\partial}{\partial\beta}\{\mu(X;\beta)\}\{Y-\mu(X;\beta)\}\,|X\right]\right) \\ & \qquad = E\left(\frac{\partial}{\partial\beta}\{\mu(X;\beta)\}\,E_\beta\left[\pi\{(1,1),(Y,X)\}\,\{Y-\mu(X;\beta)\}\,|X\right]\right). \end{aligned}$$

This depends critically on the quantity

$$E\left[\pi\{(1,1),(Y,X)\}\,\{Y-\mu(X;\beta)\}\,|X\right],$$

which is the conditional (given X) covariance between $\pi\{(1,1),(Y,X)\}$ and $\{Y-\mu(X;\beta)\}$, which typically would be expected to be ***nonzero*** even if $\beta=\beta_0$.

Accordingly, $\widehat{\beta}^c$ solving (1.27) is almost certainly an ***inconsistent*** estimator in general under these conditions.

It is easy to imagine that missingness mechanisms for which $\pi\{(1,1),(Y,X)\}$ depends on Y are likely to be MNAR. However, it is possible for the missingness mechanism to be such that $\pi\{(1,1),(Y,X)\}$ depends on Y and the mechanism is MAR.

To see this, suppose Y is always observed but X can be missing and that the probability that X is missing depends only on Y. Then

$$\pi\{(0,1),(Y,X)\}=\pi\{(0,0),(Y,X)\}=0 \quad \text{because } Y \text{ is always observed,}$$

$$\pi\{(1,1),(Y,X)\} \quad \text{and} \quad \pi\{(1,0),(Y,X)\} \quad \text{depend only on } Y.$$

Then the probabilities of all possible missingness patterns depend at most only on Y, which is always observed, so that the mechanism is indeed MAR.

In conclusion, if the missing mechanism is MAR or MNAR, the complete case OLS estimator is inconsistent in general when $\pi\{(1,1),(Y,X)\}$ depends on Y.

Case 3. $\pi\{(1,1),(Y,X)\}$ ***does not depend on*** (Y,X)***.*** This implies that $\pi\{(1,1),(Y,X)\} = \pi\{(1,1)\}$ is a ***constant***, which is of course true if the missingness mechanism is MCAR. Here, $\pi\{(1,1)\}$ factors out of the left hand side of (1.29), and thus with a correctly specified model $\mu(x;\beta)$, (1.29) is trivially true. This confirms that, under MCAR, the complete case estimator $\widehat{\beta}^c$ is consistent.

MORAL: This example illustrates how complicated and sometimes nonintuitive the consequences of ignoring missing data can be.

EXAMPLE 3. Missingness in longitudinal regression analysis. Consider the simple longitudinal situation where a scalar outcome is collected at each of T times $t_1, \ldots, t_T$, with no additional covariates or other information, so that the full data are

$$Z = (Y_1, \ldots, Y_T).$$

Letting $Y = (Y_1, \ldots, Y_T)^T$, interest focuses on a model for $E(Y)$ of the form

$$\mu(\beta) = \{\mu_1(\beta), \ldots, \mu_T(\beta)\}^T,$$

where $\mu_j(\beta)$ depends only on time; e.g.,

$$\mu_j(\beta) = \beta_0 + \beta_1 t_j \quad \text{or} \quad \mu_j(\beta) = 1/\{1 + e^{-(\beta_0+\beta_1 t_j)}\},$$

and β is $(p \times 1)$. Let $R = (R_1, \ldots, R_T)$.

Assume missingness is ***monotone***, so that the possible values R can take on are

$$r^{(0)} = (0,0,\ldots,0), \quad r^{(1)} = (1,0,\ldots,0), \quad \ldots, \quad r^{(T)} = (1,1,\ldots,1),$$

where $r^{(0)}$ corresponds to outcome missing at all times and $r^{(T)}$ to availability of the full data as before.

If the full data were available on all individuals, a standard approach to estimation of β would be to solve a ***generalized estimating equation*** (GEE) of the form

$$\sum_{i=1}^{N} \mathcal{D}^T(\beta)\mathcal{V}^{-1}(\beta) \begin{pmatrix} Y_{i1} - \mu_1(\beta) \\ \vdots \\ Y_{iT} - \mu_T(\beta) \end{pmatrix} = 0, \tag{1.32}$$

where $\mathcal{D}^T(\beta)$ is a $(p \times T)$ matrix of partial derivatives of $\mu(\beta)$ multiplied by a $(T \times T)$ working covariance matrix $\mathcal{V}(\beta)$, both of which depend on the observation times.

If the model $\mu(\beta)$ is ***correctly specified*** in the same sense as in the previous example, then it is well known that (1.32) is an ***unbiased estimating equation*** and the resulting estimator for β is consistent.

A common default for statistical software implementing estimation of β by solving (1.32) when some elements of Y may be missing is to base this on all of the ***available (observed) data***.

What are the implications for inference on β in this case?

In the monotone missingness setting here, the available data for individual i are the observed data consisting of R_i and, when $R_i = r^{(j)}$, the vector of observed outcomes $(Y_{i1}, \ldots, Y_{ij})^T$.

The contribution to the GEE for an individual with missingness pattern $r^{(j)}$ thus depends only on the vector of observed outcomes $(Y_1, \ldots, Y_j)^T$ and the corresponding submatrices $\mathcal{D}_j^T(\beta)$ ($p \times j$) and $\mathcal{V}_j(\beta)$ ($j \times j$) of $\mathcal{D}^T(\beta)$ and $\mathcal{V}(\beta)$, respectively.

A little thought reveals that the GEE being solved may be written as

$$\sum_{i=1}^{N} \left\{ \sum_{j=1}^{T} I(R_i = r^{(j)})\, \mathcal{D}_j^T(\beta) \mathcal{V}_j^{-1}(\beta) \begin{pmatrix} Y_{i1} - \mu_1(\beta) \\ \vdots \\ Y_{ij} - \mu_j(\beta) \end{pmatrix} \right\} = 0. \tag{1.33}$$

(Note that there is no contribution for $j = 0$, as no outcomes are observed.)

As in the simpler univariate regression case, the estimator $\widehat{\beta}^{obs}$ found by solving (1.33) will converge in probability as $N \to \infty$ to β_0 if the expectation of the summand is equal to zero at the true value β_0. This will be true if, at $\beta = \beta_0$,

$$E_\beta \left\{ I(R_i = r^{(j)})\, \mathcal{D}_j^T(\beta) \mathcal{V}_j^{-1}(\beta) \begin{pmatrix} Y_{i1} - \mu_1(\beta) \\ \vdots \\ Y_{ij} - \mu_j(\beta) \end{pmatrix} \right\} = 0 \quad \text{for all } j = 1, \ldots, T. \tag{1.34}$$

Now recall that

$$E\{I(R = r)|Y\} = \text{pr}(R = r|Y) = \pi(r, Y),$$

say. Consider the left hand side of (1.34) under different conditions on $\pi(r, Y)$.

- $\pi(r, Y)$ ***does not depend on*** Y**.** In this case $\pi(r, Y) = \pi(r)$, a constant for all $r = r^{(j)}$, $j = 0, \ldots, T$, and the missingness mechanism is MCAR. Thus, the left hand side of (1.34) becomes

  $$E_\beta \left\{ \text{pr}(R_i = r^{(j)}|Y)\, \mathcal{D}_j^T(\beta) \mathcal{V}_j^{-1}(\beta) \begin{pmatrix} Y_{i1} - \mu_1(\beta) \\ \vdots \\ Y_{ij} - \mu_j(\beta) \end{pmatrix} \right\} = \pi(r^{(j)}) \mathcal{D}_j^T(\beta) \mathcal{V}_j^{-1}(\beta) E_\beta \left\{ \begin{pmatrix} Y_{i1} - \mu_1(\beta) \\ \vdots \\ Y_{ij} - \mu_j(\beta) \end{pmatrix} \right\}$$

  which clearly is equal to 0 at $\beta = \beta_0$ when $\mu(\beta)$ is correctly specified.

  Thus, under MCAR, $\widehat{\beta}^{obs}$ is a consistent estimator.

- $\pi(r^{(j)}, Y)$ ***depends only on*** $Y_1, \ldots, Y_j$**.** Here, missingness probabilities for each missingness pattern depend only on the data observed under the pattern, so this corresponds to MAR. Write $\pi(r^{(j)}, Y) = \pi(r^{(j)}, Y_1, \ldots, Y_j)$. Conditioning on $(Y_1, \ldots, Y_j)$, the left hand side of (1.34) can be written as

$$\mathcal{D}_j^T(\beta)\mathcal{V}_j^{-1}(\beta)\ E_\beta\left\{\pi(r^{(j)}, Y_1, \ldots, Y_j)\begin{pmatrix} Y_{i1} - \mu_1(\beta) \\ \vdots \\ Y_{ij} - \mu_j(\beta) \end{pmatrix}\right\}.$$

  The kth element in the expectation in this expression, $k = 1, \ldots, j$, is equal to

$$\text{cov}\{\pi(r^{(j)}, Y_1, \ldots, Y_j), Y_k\}.$$

  This covariance is almost certainly not equal to zero in general. Thus, under MAR, $\widehat{\beta}^{obs}$ is not a consistent estimator in general.

It should be clear that, in the case of MNAR, similar considerations apply.

In conclusion, the estimator for β obtained by solving a GEE based on the available, observed data under monotone missingness is only guaranteed to be consistent if the missingness mechanism is MCAR.

In Chapter 5, we will study how ***inverse probability weighting*** is one approach to "correcting" the analysis based on the available data here and the complete case analysis in univariate regression when the missingness is MAR.

MORAL: This example and the previous example demonstrate the general principle that great care must be taken with default analyses carried out by standard software for many popular statistical methods. Complete case or observed data analyses have the potential to lead to invalid inferences when data are missing, and the data analyst must be aware of these pitfalls and think carefully about his/her beliefs about the nature of the missingness. The potential for misleading inferences is high by users who assume that, because the default analysis is available in the software, it is sound. This is a particular danger when the users are not well-versed in statistics.

1.5 Review of estimating equations

Before we discuss approaches to addressing missing data in subsequent chapters, we review an important class of estimators that will arise later in the course. We have already discussed special cases in Section 1.4, where we considered two common situations in which interest focuses on parameters in a model for some aspect of the joint density of the full data.

- In ***EXAMPLE 2***, we considered the case where $Z = (Y, X)$, and interest focuses on the parameter β ($p \times 1$) in an assumed model

$$E(Y|X = x) = \mu(x; \beta). \tag{1.35}$$

  As discussed in Section 1.3, with no further assumptions on the distribution of Z, $p_Z(z; \theta)$, say, this is a semiparametric model, where β is a component of θ as described earlier. Alternatively, one could make a full distributional assumption on Z, which might involve taking the conditional distribution of Y given X to be normal with mean (1.35) and variance σ^2.

  In either case, with full data (Y_i, X_i), $i = 1, \ldots, N$ available, a natural approach to estimation of the parameter of interest β is to solve the usual least squares ***estimating equation***

$$\sum_{i=1}^{N} \frac{\partial}{\partial \beta}\{\mu(X; \beta)\}\{Y_i - \mu(X_i; \beta)\} = 0. \tag{1.36}$$

  In the case where Y given X is assumed normally distributed as above, solving (1.36) in fact corresponds to finding the maximum likelihood estimator for β.

- In ***EXAMPLE 3***, we considered the case where $Z = (Y_1, \ldots, Y_T)$; the Y_j, $j = 1, \ldots, T$, are scalar outcomes collected at times $t_1, \ldots, t_T$; $Y = (Y_1, \ldots, Y_T)^T$; and interest focuses on β ($p \times 1$) in a model

$$E(Y) = \mu(\beta) = \{\mu_1(\beta), \ldots, \mu_T(\beta)\}^T \tag{1.37}$$

  As we discussed, it is standard in this context to posit a semiparametric model (1.37), leaving the rest of the joint density of Z unspecified. The standard approach to estimation of β is to solve the ***GEEs***

$$\sum_{i=1}^{N} \mathcal{D}^T(\beta)\mathcal{V}^{-1}(\beta) \begin{pmatrix} Y_{i1} - \mu_1(\beta) \\ \vdots \\ Y_{iT} - \mu_T(\beta) \end{pmatrix} = 0, \tag{1.38}$$

  where $\mathcal{D}^T(\beta)$ is the ($p \times T$) partial derivative matrix of $\mu(\beta)$, and $\mathcal{V}(\beta)$ is a ($T \times T$) working covariance matrix.

M-estimators. The estimators defined by solving (1.35) and (1.38) are examples within a general class of estimators known as ***M-estimators***. M-estimators are covered in detail in ST 793. Because inferences of interest in so many common full data problems are addressed through M-estimators, we will be considering in future chapters how to handle these analyses when some data are missing. Accordingly, we provide a brief, generic review here of the essential elements of M-estimation in full data problems.

Assume that we have a sample of full data Z_i, $i = 1, \ldots, N$, which are iid with density $p_Z(z)$. Suppose interest focuses on a finite-dimensional parameter θ ($p \times 1$); here, θ may fully characterize the density, in which case we can write $p_Z(z;\theta)$, or θ may fully characterize aspects of the distribution of interest. Suppose that θ_0 is the true value of θ; i.e., the value of θ such that $p_Z(z;\theta_0)$ is the true density generating the data, or, as in the case of (1.35), for example, leads to the true function $E(Y|X = x)$.

A M-estimator for θ is defined as the solution $\widehat{\theta}$ (assuming a solution exists) to the ($p \times 1$) system of equations

$$\sum_{i=1}^{N} M(Z_i;\widehat{\theta}) = 0, \tag{1.39}$$

where the p-dimensional function $M(z;\theta)$ is such that

$$E_\theta\{M(Z;\theta)\} = 0.$$

The function $M(z;\theta)$ is also such that $E_\theta\{M(Z;\theta)^T M(Z;\theta)\} < \infty$, and $E_\theta\{M(Z;\theta)M(Z;\theta)^T\}$ ($p \times p$) is positive definite for all θ.

Estimating function. $M(z;\theta)$ satisfying these conditions is referred to as an ***estimating function***, as it defines an ***estimating equation***. For example, in the case of least squares regression in (1.36),

$$M(Z;\beta) = \frac{\partial}{\partial\beta}\{\mu(X;\beta)\}\{Y - \mu(X;\beta)\},$$

which satisfies $E_\theta\{M(Z;\theta);\theta\} = 0$ as long as the model $\mu(x;\beta)$ for $E(Y|X = x)$ is correctly specified.

Note that if we have a parametric model for the full data $p_Z(z;\theta)$, then taking

$$M(z;\theta) = \frac{\partial}{\partial\theta}\log\{p_Z(z;\theta)\},$$

i.e., the score, yields the maximum likelihood estimator for θ under this model, so that the maximum likelihood estimator is a M-estimator. Here, the estimating equation is the usual ***score equation***, and the estimating function is the score.

Approximate large sample distribution of M-estimator. Under suitable regularity conditions, which we do not state formally, it may be shown that M-estimators are consistent and asymptotically normal; these conditions are reviewed in ST 793. For consistency of the estimator $\widehat{\theta}$, for our purposes, it suffices to note that the M-estimating equation is ***unbiased*** in the sense we described earlier.

An approximate (normal) sampling distribution for $\widehat{\theta}$ can be found via a standard Taylor series argument. Multiply (1.39) by $N^{-1/2}$ and expand to linear terms about θ_0:

$$0 = N^{-1/2}\sum_{i=1}^{N} M(Z_i;\widehat{\theta}) = N^{-1/2}\sum_{i=1}^{N} M(Z_i;\theta_0) + \left\{N^{-1}\sum_{i=1}^{N}\frac{\partial}{\partial\theta^T}M(Z_i;\theta^*)\right\}N^{1/2}(\widehat{\theta}-\theta_0),$$

where θ^* is a value intermediate between $\widehat{\theta}$ and θ_0. By the consistency of $\widehat{\theta}$, it follows that

$$\left\{N^{-1}\sum_{i=1}^{N}\frac{\partial}{\partial\theta^T}M(Z_i;\theta^*)\right\} \xrightarrow{p} E\left\{\frac{\partial}{\partial\theta^T}M(Z;\theta_0)\right\}.$$

Assuming this matrix is nonsingular, rearranging, we have

$$N^{1/2}(\widehat{\theta}-\theta_0) \approx -\left[E\left\{\frac{\partial}{\partial\theta^T}M(Z;\theta_0)\right\}\right]^{-1} N^{-1/2}\sum_{i=1}^{N} M(Z_i;\theta_0). \tag{1.40}$$

Applying Slutsky's theorem and the central limit theorem to (1.40), it follows that

$$N^{1/2}(\widehat{\theta}-\theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \left[E\left\{\frac{\partial}{\partial\theta^T}M(Z;\theta_0)\right\}\right]^{-1}\text{var}\{M(Z;\theta_0)\}\left[E\left\{\frac{\partial}{\partial\theta^T}M(Z;\theta_0)\right\}\right]^{-1\,T}\right), \tag{1.41}$$

where

$$\text{var}\{M(Z;\theta_0)\} = E\{M(Z;\theta_0)M(Z;\theta_0)^T\}.$$

The notation in (1.41) is a shorthand representation of the fact that the expression on the left hand side converges in distribution (law) to a normal random vector with mean zero and covariance matrix as shown on the right hand side.

We can use the large sample result (1.41) to deduce an approximate sampling distribution for $\widehat{\theta}$. By the law of large numbers,

$$N^{-1}\sum_{i=1}^{N}\frac{\partial}{\partial\theta^T}M(Z_i;\theta_0) \xrightarrow{p} E\left\{\frac{\partial}{\partial\theta^T}M(Z;\theta_0)\right\}$$

and

$$N^{-1}\sum_{i=1}^{N} M(Z_i;\theta_0)M({}_iZ;\theta_0)^T \xrightarrow{p} E\{M(Z;\theta_0)M(Z;\theta_0)^T\}.$$

Substituting $\widehat{\theta}$ for θ_0 in these expressions, we arrive at the so-called ***sandwich estimator*** for the covariance matrix of the normal distribution in (1.41), given by

$$\left\{N^{-1}\sum_{i=1}^{N}\frac{\partial}{\partial\theta^T}M(Z_i;\widehat{\theta})\right\}^{-1}\left\{N^{-1}\sum_{i=1}^{N} M(Z_i;\widehat{\theta})M({}_iZ;\widehat{\theta})^T\right\}\left\{N^{-1}\sum_{i=1}^{N}\frac{\partial}{\partial\theta^T}M(Z_i;\widehat{\theta})\right\}^{-1\,T}. \tag{1.42}$$

Combining (1.41) and (1.42), we have

$$\widehat{\theta} \dot{\sim} \mathcal{N}\left(\theta_0, \left\{\sum_{i=1}^{N} \frac{\partial}{\partial \theta^T} M(Z_i;\widehat{\theta})\right\}^{-1} \left\{\sum_{i=1}^{N} M(Z_i;\widehat{\theta}) M({}_iZ;\widehat{\theta})^T\right\} \left\{\sum_{i=1}^{N} \frac{\partial}{\partial \theta^T} M(Z_i;\widehat{\theta})\right\}^{-1\,T}\right) \qquad (1.43)$$

where "$\dot{\sim}$" denotes "approximately distributed as." Note that in (1.43), because we have rescaled (1.41), the N^{-1} terms cancel.

The approximate result (1.43) may be used in practice to derive approximate (large sample) standard errors and confidence intervals for components of the M-estimator $\widehat{\theta}$ in the usual way.

1.6 Objectives and outline of this course

The previous sections provide ample demonstration of the serious consequences of ignoring the issue of missing data through the use of complete case or observed data analyses. This problem is exacerbated in practice by the fact that popular software carries out these analyses by default.

OBJECTIVE: Despite the pervasiveness of missing data in numerous application areas and the considerable research on methods to handle them over the past several decades, many degree programs in statistics or biostatistics do not offer a full-fledged course on methods for analysis in the presence of missing data. The first time many graduates encounter the challenge of doing an analysis when some data are missing is in applied or collaborative work after completing their programs.

This course is meant to address this gap in training. We will cover the key concepts and popular methodology in some detail. We will also cover the application of the methods.

CHALLENGE: There is a vast and still-evolving literature on models and methods for inference in the presence of missing data. It would be impossible in a single course to cover this extensive literature.

Accordingly, we will present key concepts and principles and discuss major classes of models and methods with the goal of developing the foundation and background necessary for further study of this literature and for understanding of how the methods can be implemented in practice.

IMPLEMENTATION AND SOFTWARE: An obstacle to the broad use of principled missing data methods is the dearth of available, general purpose software. Although there are software implementations for some methods in SAS and R, these have not yet gained widespread acceptance and use. The Comprehensive R Archive Network (CRAN) has several user-contributed packages, but many of these are relevant to specialized situations.

In this course, we will discuss, demonstrate, and use of a few of the software implementations that are generally applicable. We will also reinforce understanding of the formulation of the methods by programming them in specific settings. We will make use of simulation studies to explore the implications of missing data and the performance of methods.

SCOPE: As we noted earlier, missing data are a particular problem in studies of humans, and entire books (e.g., Molenberghs and Kenward, 2007; O'Kelly and Ratitch, 2014) are devoted to handling missing data in the analysis of clinical trials. Accordingly, many of the situations and examples we will discuss will involve clinical trials. However, the approaches and insights gained are broadly relevant to other application areas.

OUTLINE: A brief outline of the major topics to be covered is as follows

Chapter 1. Introduction and Motivation

- Challenges posed by missing data
- General statistical framework
- Missing data mechanisms
- Review of estimating equations

Chapter 2. Naïve Methods

- Complete case and available case analysis
- Single imputation methods
- Last Observation Carried Forward

Chapter 3. Likelihood-based Methods Under MAR

- Review of maximum likelihood inference for full data
- Factorization of the density of (R, Z)
- Observed data likelihood and ignorability
- Expectation-Maximization (EM) algorithm

- Missing information principle
- Bayesian inference

Chapter 4. Multiple Imputation Methods Under MAR

- Fundamentals of multiple imputation
- Proper versus improper imputation
- Rubin's variance formula
- Asymptotic results
- Imputation from a multivariate normal distributio
- Multiple Imputation by Chained Equations (MICE)

Chapter 5. Inverse Probability Weighting Methods Under MAR

- Illustrative examples of inverse probability weighting
- Weighted generalized estimating equations for longitudinal data with dropout
- Inverse probability weighting at the occasion level
- Inverse probability weighting at the individual level
- Doubly robust estimation

Chapter 6. Pattern Mixture Models

- Introduction and rationale
- Modeling strategies

Chapter 7. Sensitivity Analysis to Deviations from MAR

- Challenges under Missing Not At Random
- Example: Single outcome
- Example: Single outcome with auxiliary information
- Example: Longitudinal outcomes with dropout