2 Naïve Methods

Before discussing methods for taking account of missingness when the missingness pattern can be assumed to be MAR in the next three chapters, we review some simple methods for handling missingness. We refer to these as ***naïve*** methods because they are mainly ad hoc in nature and not necessarily based on a principled framework for addressing missing data problems. Nonetheless, these methods have been widely used in practice.

2.1 Complete or available case analysis

In Chapter 1, we have already examined the consequences of proceeding with analysis based on only the ***complete cases***, that is, the data from individuals on whom the full data (all data intended to be collected) are observed; or the ***available cases***, that is, the observed data on all individuals.

COMPLETE CASE ANALYSIS: Historically, before the advent of modern computing, much importance was attached to having what has often been called in the missing data literature a ***rectangular data set***. That is, if data on M variables were to be collected on each of N individuals, the intended data could be represented as a $N \times M$ rectangular array, with a row of M values (data record) corresponding to each of the N individuals. Implementation of common analysis methods depended on this rectangular structure, as necessary computations were often simpler and less time-consuming than those when this structure did not hold. In fact, (finite-sample) properties of some methods that were straightforward under a rectangular structure could become more challenging to deduce.

Complete case analysis was thus viewed primarily as a way to preserve rectangular structure and thus simplify computations and understanding of properties. However, putting aside the issues associated with failure to appreciate the implications of the missing data mechanism, the ***loss of information*** could be considerable. For example, as noted by Molenberghs and Kenward (2007), with $M = 20$ variables, 10% missing data on each, and (somewhat unrealistically) missingness in each variable happening ***independently*** of that in all the others, the probability of observing a complete case is around 0.13. Even in the more realistic setting where the missingness in variables is correlated, it is clear that precision and power of desired inferences will be diminished.

More ominously, as we demonstrated in Chapter 1 in the simple cases of estimation of a single mean and univariate regression analysis, unless the missingness mechanism is MCAR, the potential for ***biased inferences*** is substantial. It is not hard to imagine that the implications for conducting a complete case analysis in more complicated settings are similar.

AVAILABLE CASE ANALYSIS: Methods that do not necessarily require the data to be representable in a rectangular array can also be impacted by simply carrying out the analysis based on the ***available or observed data*** without consideration of the consequences of missingness. In Chapter 1, we considered the setting of longitudinal data analysis under an assumed model for the ***population average*** of an outcome ascertained over time using GEEs, where interest focuses on inference on parameters in this assumed model.

By construction, GEEs can handle easily unequal numbers of longitudinal outcomes across individuals, so there is no computational challenge, and available software allows for this possibility. However, if the numbers of outcomes available per individual differ because, among some intended number T outcomes to be collected over time, some are missing, then care must be taken. We saw that, in the case of ***dropout***, if the resulting observed data are analyzed as if the differing numbers of observations were planned and not the result of some missing data mechanism, then inference on parameters characterizing the population average outcome can be compromised. Thus, even though the analysis is entirely feasible to conduct, it must be interpreted through the lens of missing data.

BOTTOM LINE: Although complete case and available case approaches to handling missing data have the appeal of being simple and straightforward to implement, they are not recommended because they do not take appropriate account of the missingness mechanism. Accordingly, we do not discuss them further.

2.2 Simple imputation methods

Given the historical context for desirability of having a rectangular data set, an alternative approach to achieving this based on “filling in” the missing values has been advocated.

IMPUTATION: In particular, the idea is to ***impute*** missing values in the data set based on the observed data and then to carry out the analysis that would be undertaken if the full data were all observed. Several ***simple imputation*** approaches have been proposed. Such approaches are appealing due to their simplicity but, as we demonstrate shortly, can be ***dangerous***.

Most simple imputation methods can be shown to require that the missingness mechanism be MCAR for the ensuing inferences to be valid; for example, for ***consistency*** of estimators for parameters of interest to hold. In fact, some methods do not necessarily yield consistent inferences even under MCAR, as we will see in the next section.

A further drawback is that measures of uncertainty will be ***distorted***. Specifically, the standard approach is to proceed as if the imputed observations are the actual, intended observations; that is, adopt standard error estimates, confidence intervals, and so on arising from the usual formulæ, with no acknowledgement that the imputed data are derived from the observed data. Intuitively, this fails to account for the uncertainty due to the imputation of missing values.

We now review a few simple imputation approaches and, for one of these, provide a detailed demonstration of the implications for inference.

UNCONDITIONAL MEAN IMPUTATION: A natural and simple method for imputing missing values on a particular continuous variable is to use the ***average*** of the observed values on that variable from the individuals on whom that variable is observed. This is referred to as ***unconditional mean imputation*** to reflect the fact that the imputed value does not use (so condition on) other information on an individual for whom the variable is missing, as in the method we discuss below.

When a variable is discrete (categorical), an analogous approach is to use the ***mode*** of the observed values on that variable from the individuals on whom that variable is observed.

Although these approaches are straightforward to implement, they seem likely to be problematic as far as inference is concerned.

REGRESSION IMPUTATION: A seemingly more sophisticated approach is to posit a ***regression model*** for missing values as a function of observed variables and use the ***predicted values*** from a fit of such a model as the imputed data. We examine the properties of this method in a simple example, which suffices to illustrate the potential drawbacks.

Suppose that the full data are $Z = (Y_1, Y_2)$, so that $K = 2$, corresponding to a baseline and follow-up measurement of some outcome, say. Suppose further that Y_1 is always observed, so that $R_1 = 1$ always, and Y_2 may be missing. Write $Y = (Y_1, Y_2)^T$ to be the full data vector of outcomes, and let $\widetilde{Y}_1 = (1, Y_1)^T$. The objective is to estimate

$$\mu_2 = E(Y_2). \tag{2.1}$$

To impute Y_2 for those individuals for whom it is missing, suppose we posit a linear regression model

$$E(Y_2|Y_1) = \beta_0 + \beta_1 Y_1 = \widetilde{Y}_1^T\beta, \quad \beta = (\beta_0, \beta_1)^T. \tag{2.2}$$

For a random sample of N individuals, suppose we fit this model by OLS using the data from the ***complete cases*** for whom $R_2 = 1$. This OLS estimator is given by

$$\widehat{\beta} = \left\{\sum_{i=1}^N R_{i2}\widetilde{Y}_{i1}\widetilde{Y}_{i1}^T\right\}^{-1}\left\{\sum_{i=1}^N R_{i2}\widetilde{Y}_{i1}Y_{i2}\right\}. \tag{2.3}$$

We then ***impute*** Y_2 for individuals i for whom it is missing ($R_{i2} = 0$) by the ***predicted values***

$$\widehat{Y}_{i2} = \widetilde{Y}_{i1}^T\widehat{\beta} = \widehat{\beta}_0 + \widehat{\beta}_1 Y_{i1}.$$

The ***regression imputation*** estimator for μ_2 in (2.1) substitutes $\widehat{Y}_{i2}$ for the missing Y_{i2} for all individuals i in the sample for whom it is missing. We can write this estimator as

$$\widehat{\mu}_2^{RIMP} = N^{-1}\sum_{i=1}^N \left\{R_{i2}Y_{i2} + (1 - R_{i2})\widehat{Y}_{i2}\right\}. \tag{2.4}$$

When is $\widehat{\mu}_2^{RIMP}$ in (2.4) a consistent estimator for μ_2?

Consider the properties of $\widehat{\mu}_2^{RIMP}$ under one or both of the following conditions:

(i) $E(Y_2|Y_1) = \beta_0^{(0)} + \beta_1^{(0)}Y_1 = \widetilde{Y}_1^T\beta^{(0)}$ for some $\beta^{(0)} = (\beta_0^{(0)}, \beta_1^{(0)})^T$; that is, the linear regression model (2.2) used for imputation is ***correctly specified***.

(ii) $\text{pr}(R_2 = 1|Y) = \text{pr}(R_2 = 1|Y_1) = \pi_2(Y_1)$, say; that is, the ***missingness mechanism is MAR***.

Suppose first that both (i) and (ii) hold. We first deduce the behavior of $\widehat{\beta}$ in (2.3). By the weak law of large numbers and under regularity conditions,

$$\widehat{\beta} \xrightarrow{p} \left\{E(R_2\widetilde{Y}_1\widetilde{Y}_1^T)\right\}^{-1} E(R_2\widetilde{Y}_1Y_2). \tag{2.5}$$

Consider the second component in the right most term in (2.5), which is $\{E(R_2Y_2), E(R_2Y_1Y_2)\}^T$:

$$\begin{aligned} E(R_2Y_1Y_2) &= E\{E(R_2|Y_1, Y_2)Y_1Y_2\} = E\{E(R_2|Y_1)Y_1Y_2\} = E\{\pi_2(Y_1)Y_1Y_2\} \quad \text{by (ii)} \\ &= E\{\pi_2(Y_1)Y_1E(Y_2|Y_1)\} = E\{\pi_2(Y_1)Y_1\widetilde{Y}_1^T\}\beta^{(0)}; \end{aligned}$$

a similar argument holds for the first component. Thus, under (i),

$$E(R_2\widetilde{Y}_1Y_2) = E\{\pi_2(Y_1)\widetilde{Y}_1\widetilde{Y}_1^T\}\beta^{(0)}.$$

By a similar argument, the other term in (2.5) satisfies

$$E(R_2 \widetilde{Y}_1 \widetilde{Y}_1^T) = E\{\pi_2(Y_1) \widetilde{Y}_1 \widetilde{Y}_1^T\}.$$

It follows that

$$\widehat{\beta} \xrightarrow{p} \Big[E\{\pi_2(Y_1) \widetilde{Y}_1 \widetilde{Y}_1^T\} \Big]^{-1} E\{\pi_2(Y_1) \widetilde{Y}_1 \widetilde{Y}_1^T\} \beta^{(0)} = \beta^{(0)}. \tag{2.6}$$

Thus, under (i) and (ii), using manipulations analogous to those above and (2.6),

$$\begin{aligned}
\widehat{\mu}_2^{RIMP} &= N^{-1} \sum_{i=1}^{N} \Big\{ R_{i2} Y_{i2} + (1 - R_{i2}) \widetilde{Y}_{i1}^T \widehat{\beta} \Big\} \\
&\xrightarrow{p} E(R_2 Y_2) + E\{(1 - R_2) \widetilde{Y}_1^T\} \beta^{(0)} = E\{\pi_2(Y_1) Y_2\} + E\Big[\{1 - \pi_2(Y_1)\} \widetilde{Y}_1^T \beta^{(0)} \Big] \\
&= E\{\pi_2(Y_1) Y_2\} + E[\{1 - \pi_2(Y_1)\} E(Y_2 | Y_1)] \\
&= E\{\pi_2(Y_1) E(Y_2 | Y_1)\} + E[\{1 - \pi_2(Y_1)\} E(Y_2 | Y_1)] \\
&= E\{E(Y_2 | Y_1)\} = E(Y_2) = \mu_2.
\end{aligned}$$

That is, under (i) and (ii), corresponding to a ***correctly specified imputation model*** and ***MAR***, $\widehat{\mu}_2^{RIMP}$ is a ***consistent estimator*** for μ_2.

What happens if (i) or (ii) does not hold?

Suppose that (ii) holds (MAR) but (i) (correctly specified imputation model) does not. Here,

$$\widehat{\beta} \xrightarrow{p} \beta^* = \Big[E\{\pi_2(Y_1) \widetilde{Y}_1 \widetilde{Y}_1^T\} \Big]^{-1} E\{\pi_2(Y_1) \widetilde{Y}_1 E(Y_2 | Y_1)\},$$

so that

$$\begin{aligned}
\widehat{\mu}_2^{RIMP} &= N^{-1} \sum_{i=1}^{N} \Big\{ R_{i2} Y_{i2} + (1 - R_{i2}) \widetilde{Y}_{i1}^T \widehat{\beta} \Big\} \xrightarrow{p} E\{\pi_2(Y_1) Y_2\} + E[\{1 - \pi_2(Y_1)\} \widetilde{Y}_1^T] \beta^* \\
&= E(Y_2) + E[\{1 - \pi_2(Y_1)\}(\widetilde{Y}_1^T \beta^* - Y_2)] = \mu_2 + E[\{1 - \pi_2(Y_1)\}(\widetilde{Y}_1^T \beta^* - Y_2)]
\end{aligned}$$

which evidently is not equal to μ_2 in general.

Likewise, if (i) holds (correctly specified imputation model) but (ii) (MAR) does not, we now have

$$\text{pr}(R_2 = 1 | Y) = \pi_2(Y),$$

say, depending on both Y_1 and Y_2, the latter of which may be unobserved. Even if the regression model (2.2) is correctly specified, the complete case OLS estimator for β (2.3) is not consistent in general (try it), and it can be shown (try it) that $\widehat{\mu}_2^{RIMP}$ does not converge in probability to μ_2 in general.

MORAL: In this special case, the regression imputation method yields consistent inference if ***both*** the imputation model is correctly specified and the missingness mechanism if MAR. However, in more complex settings, it is not clear that this need hold.

Critically, even if (i) and (ii) do hold, the usual standard error for a sample mean is not a valid estimator of the precision of the estimator $\widehat{\mu}_2^{RIMP}$ in (2.4). The approach in practice is to use the usual formula as if the imputed values were actual observations on the outcome, which would clearly lead to incorrect assessment of uncertainty. (For fun, try deriving a valid standard error estimator.)

OTHER SIMPLE IMPUTATION METHODS: There are still other simple imputation approaches that have been widely used in practice. For example, so-called ***hot deck imputation*** is based on using observed values from “matching” individuals to “fill in” values for those individuals for whom they are missing based on some “matching” strategy.

In general, although in our simple regression imputation example a consistent estimator is possible under MAR and a correct imputation model, most simple imputation methods in fact require that the missingness mechanism be MCAR for consistent inference. Moreover, the usual formulæ for estimators of precision, although commonly used in practice, are incorrect because of failure to take into account the uncertainty due to imputation.

We do not discuss such simple imputation methods further in this course, as principled methods for handling missing data are available that do not have these drawbacks.

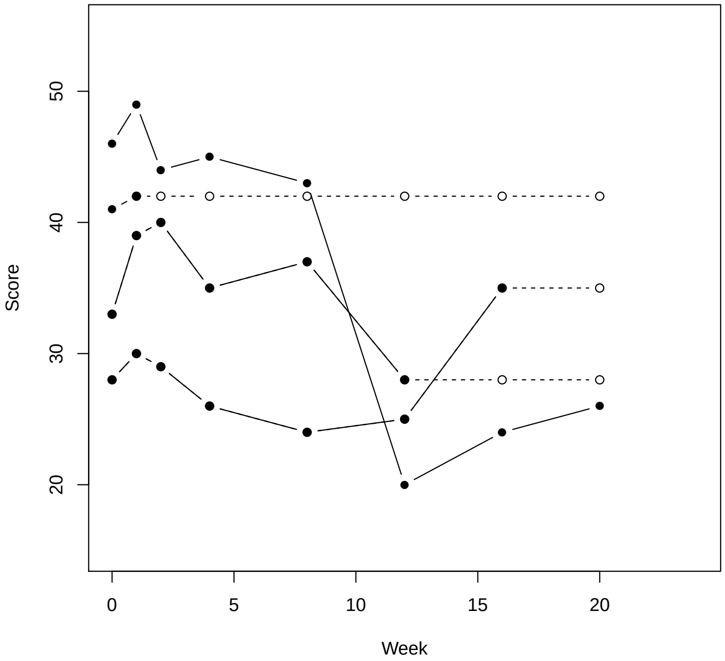
Although simple imputation methods are problematic, the idea of “filling in” missing values in some principled way has considerable practical appeal. In Chapter 4, we discuss ***multiple imputation***, which provides a framework in which imputation may be justified and incorporates estimators of precision that acknowledge that imputation has taken place.

2.3 Last Observation Carried Forward (LOCF)

A variation on the theme of simple imputation is the method called ***last observation carried forward***, commonly abbreviated LOCF. This approach is most common in settings where a variable is ascertained repeatedly over time and may be missing in a monotone or nonmonotone fashion, although it is particularly popular when missingness is due to ***dropout*** (monotone missingness). Accordingly, we discuss LOCF in this setting.

The name of this method describes its implementation: In a longitudinal study in which data are intended to be collected at T time points, for an individual who drops out at time j, his/her missing values at times $j, \ldots, T$ are replaced by his/her last observed value (i.e., the value ascertained at time $j-1$). From this standpoint, LOCF can be viewed as a form of imputation. Figure 2.1 depicts longitudinal profiles for four hypothetical individuals, where LOCF has been used to "fill in" the missing values for the three individuals who drop out prior to the final time point at $T = 20$ weeks.

Figure 2.1: *Observed longitudinal profiles for four individuals (solid lines and symbols) with last observation carried forward (dashed lines and clear symbols) for the three individuals who drop out before the final time point (20 weeks).*



We discuss LOCF separately from other naïve imputation methods because of the controversy it has engendered. LOCF is especially common in the analysis of ***clinical trials*** in which the primary outcome is collected ***longitudinally***. Its use is a matter of considerable debate, particularly in pharmaceutical research and the regulatory context. The books by Molenberghs and Kenward (2007) and O'Kelly and Rattich (2014) offer discussions of the main issues and cite the substantial literature on this topic.

As noted by Molenberghs and Kenward (2007, Chapter 4), there have been attempts to justify LOCF on scientific grounds.

- In some settings, interest focuses on inference having to do with the ***last observed outcome measure***; that is, the outcome observed at the time point prior to when an individual might drop out of a study. Of course, whether or not this corresponds to a meaningful scientific question can only be assessed in the specific context. If it does, then an argument can be made for an analysis based on LOCF.

- A common contention is that LOCF represents a ***conservative*** analysis in the following sense. In a clinical trial comparing an experimental treatment to a control, when outcome for individuals assigned to the experimental treatment is expected to ***improve*** over time, replacing missing values from the time of dropout forward will make the longitudinal profile for an individual assigned to the experimental treatment look ***worse*** than it presumably would have if he/she had continued on the experimental treatment. This would have the effect of "handicapping" the experimental treatment in the comparison to the control treatment.

  Thus, if the experimental treatment nonetheless shows a statistically significant difference from the control on the basis of a measure like mean outcome at the final observation time T despite this "handicap," the argument is that this is evidence in support of the superiority of the experimental treatment.

  As exhibited by Molenberghs and Kenward (2007, Chapter 4), this argument ***does not*** have a rigorous basis, and scenarios in which such conservatism does not hold can be constructed.

EXAMPLE: To get a sense of the properties of the LOCF method, we consider the following example. Suppose that observations on some outcome of interest are planned to be taken at times $t_1, \ldots, t_T$. The full data are

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

Define $R = (R_1, \ldots, R_T)$ as usual. Suppose further that individuals in the study may drop out but that all individuals are observed at baseline, so $R_1 = 1$ for all. Interest focuses on estimation of mean outcome at the final (Tth) time point; i.e.,

$$\mu_T = E(Y_T).$$

The LOCF estimator for μ_T uses Y_T when it is observed and otherwise substitutes the last observed value of outcome in place of Y_T when it is not. It is convenient to express the estimator using the dropout notation defined in (1.12), i.e., $D = 1 + \sum_{j=1}^{T} R_j$. Using this notation, it is straightforward to observe that the LOCF estimator based on a sample of N individuals is then

$$\widehat{\mu}_T^{LOCF} = N^{-1} \sum_{i=1}^{N} \sum_{j=1}^{T} I(D_i = j+1) Y_{ij}. \tag{2.7}$$

Here, if dropout is at time $j+1$, $j < T$, the last observed outcome is that at time j, Y_j, so it is used in place of Y_T in the average in (2.7). Otherwise, if $j = T$, the observed Y_T is used.

What is the asymptotic behavior of $\widehat{\mu}_T^{LOCF}$? From (2.7),

$$\begin{aligned}
\widehat{\mu}_T^{LOCF} &= N^{-1} \sum_{i=1}^{N} \sum_{j=1}^{T} I(D_i = j+1) Y_{ij} \\
&= N^{-1} \sum_{i=1}^{N} \left\{ Y_{iT} + \sum_{j=1}^{T} I(D_i = j+1) Y_{ij} - \sum_{j=1}^{T} I(D_i = j+1) Y_{iT} \right\} \\
&= N^{-1} \sum_{i=1}^{N} \left\{ Y_{iT} - \sum_{j=1}^{T-1} I(D_i = j+1)(Y_{iT} - Y_{ij}) \right\} \\
&\xrightarrow{p} E(Y_T) - \sum_{j=1}^{T-1} E\{I(D = j+1)(Y_T - Y_j)\} \\
&= E(Y_T) - \sum_{j=1}^{T-1} E\{\text{pr}(D = j+1|Z)(Y_T - Y_j)\},
\end{aligned} \tag{2.8}$$

where the second equality follows because $\sum_{j=1}^{T} I(D = j+1) = 1$, and the last equality (2.8) follows by a conditioning argument (try it).

Define

$$\lambda_j(Z) = \text{pr}(D = j | D \geq j, Z),\ j = 1, \dots, T; \quad \lambda_{T+1}(Z) = \text{pr}(D = T+1 | D \geq T+1, Z) = 1,$$

$$\overline{\pi}_j(Z) = \prod_{k=1}^{j} \{1 - \lambda_k(Z)\},\ j = 1, \dots, T.$$

Consider the second term in (2.8). It is straightforward to see that

$$\text{pr}(D = j+1 | Z) = \overline{\pi}_j(Z)\, \lambda_{j+1}(Z).$$

(Verify this.)

We can then rewrite (2.8) as

$$\widehat{\mu}_T^{LOCF} \xrightarrow{p} E(Y_T) - \sum_{j=1}^{T-1} E\{\overline{\pi}_j(Z)\lambda_{j+1}(Z)(Y_T - Y_j)\}. \tag{2.9}$$

What does (2.9) imply? Clearly, $\widehat{\mu}_T^{LOCF}$ is not a consistent estimator for $\mu_T = E(Y_T)$ in general.

If there were ***no dropout***, then $\text{pr}(D = j + 1|Z) = \overline{\pi}_j(Z)\lambda_{j+1}(Z) = 0$ for all $j = 1, \dots, T - 1$, and the estimator is ***consistent***, as is obvious.

Suppose that the missingness mechanism is ***MCAR***, so that $\text{pr}(D = j + 1|Z)$ does not depend on Z and hence $\lambda_j(Z)$ $\overline{\pi}_j(Z)$ are ***constants*** λ_j and $\overline{\pi}_j$, say, for each j. In this case, the right hand side of (2.9) becomes

$$E(Y_T) - \sum_{j=1}^{T-1} \overline{\pi}_j\lambda_{j+1}E(Y_T - Y_j). \tag{2.10}$$

The expression (2.10) implies the following.

- If in fact $E(Y_j)$ is the same for all j, then the second term in (2.10) is equal to zero, and $\widehat{\mu}_T^{LOCF}$ is a consistent estimator for μ_T.
- If outcomes ***increase over time***, so that $E(Y_{j+1}) \geq E(Y_j)$ for $j = 1, \dots, T - 1$, then note that the second term in (2.10) will be nonnegative. In this case, the estimator will converge in probability to a value ***less than or equal to*** μ_T and will be ***inconsistent***. Thus, in this special case, under MCAR, if μ_T is the expected outcome under an experimental treatment, the estimator would indeed be "conservative" in the sense described earlier.
- Overall, (2.10) shows that, ***even under MCAR***, the LOCF estimator $\widehat{\mu}_T^{LOCF}$ is ***not consistent*** in general.

Clearly, under MAR or MNAR, from (2.10), the estimator will be inconsistent in general.

MORAL: This example provides a simple illustration of the problems that arise with the LOCF approach. The literature on LOCF contains many further arguments regarding its supposed advantages, drawbacks, and interpretation. Our position is that LOCF is an ad hoc method that, despite its simplicity and supposed interpretations, is not based on a principled framework. Accordingly, we do not discuss it further.

2.4 Discussion

The takeaway message of this chapter is that ad hoc approaches to accounting for missing data that are not based on a formal, principled statistical framework are likely to lead to erroneous inferences.

As noted in the previous chapter, it is unlikely that the mechanism governing missingness is MCAR in many practical situations, particularly those involving human subjects. Thus, development of principled approaches to handling missing data is most relevant when the mechanism is MAR or MNAR.

We also remarked informally that progress should be possible when it is reasonable to assume that the mechanism is MAR – because, under MAR, missingness depends only on data that are observed, it should be possible to incorporate an "adjustment" for missingness based on these observed data. (Of course, recall that it is not possible to verify the assumption of MAR based on the observed data, so it must be justified based on subject matter/scientific grounds.) Methods for handling missingness seem much more problematic under MNAR, where missingness depends on data that are not observed.

Accordingly, in the next three chapters, we cover in detail three main, principled approaches to inference in the presence of missing data under a MAR mechanism. In the next chapter, we begin by considering ***likelihood-based*** methods; in doing so, we will examine ways to represent the joint distribution of (R, Z), the full data and missingness indicators, that provide the formal basis for methods under both MAR and MNAR.