Handling Missing Data in Longitudinal Studies

Missing data is a pervasive issue, particularly in longitudinal studies. It can be quite challenging maintain individuals in studies over time, some individuals may skip certain follow-up appointments, and so on. In the second half of this course we will study survival analysis, which handles a particular kind of missingness (dropout, which occurs when a subject is lost to follow-up for the remainder of the study), for now we will focus on more general mechanisms. We are considering missingness only in the outcome, which is to say that for some individuals we do not have a measure Y_{ij} for some time points.

Notation and Categorization of Missing Mechanisms

In addition to the standard notation that we have been using, we let R_{ij} represent the observation indicator for individual i at t_j. That is, if we observe Y_{ij} then $R_{ij} = 1$ and $R_{ij} = 0$ otherwise. Moreover, we can partition Y_i into Y_i^{O} for the observations which are observed, and Y_i^{M} for the observations which are missing. We categorize three distinct classes of missingness, based on the conditional distributions surrounding these quantities:

1. Data are said to be **Missing Completely at Random (MCAR)** if
$$f_{R_i}(r_i|Y_i^{\text{O}}, Y_i^{\text{M}}, X_i) = f_{R_i}(r_i|X_i).$$
2. Data are said to be **Missing at Random (MAR)** if
$$f_{R_i}(r_i|Y_i^{\text{O}}, Y_i^{\text{M}}, X_i) = f_{R_i}(r_i|Y_i^{\text{O}}, X_i).$$
3. Otherwise, data are said to be **Not Missing at Random (NMAR)**.

Put differently: data are MCAR when observations occur independently of the values of any outcomes. Data are MAR when observations occur independently of any **unobserved** values for the outcomes. Data are NMAR when there is a dependence between whether an observation is made and the unobserved values of that observation. While the definitions for these models occur based on the conditional distribution of R_i, it can be shown (e.g. consider Rubin and Little (2019) Chapter 6) that, if the data are MAR or MCAR (which is a specific type of MAR), we need only focus on $f(Y_i|X_i)$.

In addition to the mechanism that causes the missingness, we also discuss *patterns* of missingness in data. Data are said to follow a monotone missing pattern if they can be ordered in such a way such that, once one observation is missing for an individual, all future observations are missing. Dropout, where a patient leaves the study and never returns is one such way that data can exhibit monotone missingness. For examples see the illustration in Figure 1. Resolving issues with missingness, regardless of the underlying mechaism is more straightforward when the data follows a monotone missing pattern. Moreover, because of common causes for missingness, data often do follow a monotone pattern (or are very close to following a monotone pattern).

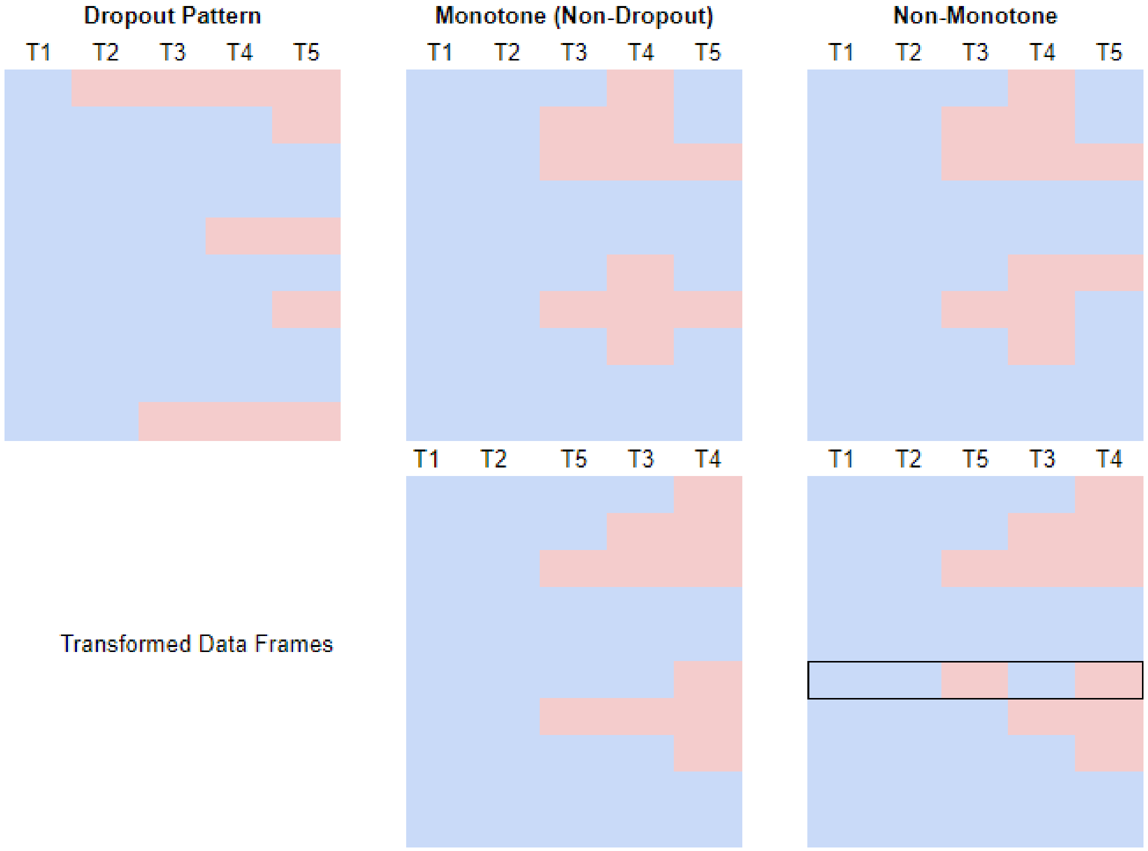

Figure 1: Missing patterns where observed values for 5 times points are shown in blue and missing values are shown in red. This compares (left) a drop out missing pattern, to a (center) monotone non-drop out pattern, to a (right) non-monotone pattern. The second row shows the transformed dataframes (where applicable) demonstrating the monotonicity (or lack thereof) for each of the provided datasets. The outlined row in the transformed data frames illustrate the issues with the non-monotone pattern.

Impacts of Missingness

When missingness occurs completely at random, it can actually be ignored in our analysis. That is because $f(Y_i|X_i) = f(Y_i^{\text{O}}|X_i)$, and so modelling just the observed values provides the correct distribution of interest. This can either be accomplished by running a *complete case* analysis, which occurs simply by dropping any responses which do not have all values recorded or (for some methods) an *available data* method. Available data methods use all available data, which may include only partial realizations for some individuals.

Even when the mechanism is MCAR, there are still issues that arise. First, conducting a complete case analysis (potentially) removes plenty of data that is avaialble, which renders the estimators far less efficient than they otherwise would be. Second, if available data are kept, then missing data forces every dataset to be *unbalanced.* As a result, any methods which rely on balanced data cannot be estimated using just the available data, even when they are MCAR. Third, there is no way of testing whether or not data are MCAR. As a result, it is a strong assumption, and one which likely is not permissible in practice.

If the data are MAR rather than MCAR, it is no longer the case that a complete case analysis is valid. This is because responders and non-responders may differ from one another, they just do so in a way that is predictable from the observed data. However, exploiting the aforementioned result, it can be shown that any techniques which rely on a **correctly specified** model of $f(Y_i|X_i)$ will produce valid estimators under MAR missingness. In particular, likelihood-based techniques will provide valid estimates if the missing data are MAR, while techniques like GEE will not. Techniques which do not specify this complete distribution, or applying techniques when these are misspecified, will generally result in biased estimates and invalid inference.

Finally, when data are NMAR, even techniques which correctly model $f(Y_i|X_i)$ are invalid. In this case we need to specifically model the missing mechanism, $f(R_i|Y_i^{\text{O}}, Y_i^{\text{M}}, X_i)$ and incorporate such a model into our analysis in order for the analysis to be valid.

Techniques for Addressing Missingness

There are several techniques for correcting for the impacts of missingness in observed data. In fact, it is a topic of research that is not only rapidly expanding, but which would take a full length course to adequately introduce. Still, becoming familiar enough with some techniques will greatly improve your capacity to accurately analyze real-world data.

1. **Complete Case Analysis:** A complete case analysis occurs when all observations that have *any* missingness are simply excldued from the analysis. This leaves us with a subset of only those who were completely observed (to so called "complete cases") and then any technique can be applied on these data. As a general rule, complete case analyses are valid when the data are MCAR, however, as previously discussed there are concerns regarding the efficiency of such estimators. In general, a complete case analysis should be avoided except for in two situations: (1) if you are trying to quickly test out an analysis, and plan on doing a more suitable correction after, but want to understand performance or general trends, and (2) if the missingness is an **incredibly** small percentage of the full data and you **know** that it is MCAR.

2. **Available Data Analysis:** An available data analysis overcomes some of the issues with the complete case analysis by including all outcomes that were observed (and ignoring those that were not). This will render the data unbalanced, and as such will restrict available to only those which accommodate unbalanced data. In general, in order for an available data analysis to be valid, the specified models (for means and covariances, for instance) must be correct for both the missing and non-missing outcomes (conditional on X_i), which is only guaranteed to be the case for MCAR data. If you have data which you know to be MCAR, running an available data analysis is a good way to overcome the efficiency concerns with a complete case analysis, without needing to perform additional modelling. Note though that the assumption of MCAR data is strong and untestable!

3. **Weighting Techniques:** If the missingness depends on specific traits (both observable and unobservable) of each individual, then this means that some people will have a higher tendency to be included (or to have their observations made) as compared to others. The idea with weighting techniques (which represent a whole class of different methods) is to compute the probability that a particular individual (who was observed) was included in the sample. Then, we can construct a *pseudo-population* by weighting each observed individual to "count for" multiple people in the population who we could have observed, but did not. If one individuals inclusion probability was 0.5, then we can think of this as saying: "In the full population there are two people who would have responded as this individual did, but one of them was unobserved due to missingness. As a result, we will count the observed individual for $\frac{1}{0.5} = 2$ individuals in our pseudo-population, to account for them and the individual like them, but who is missing."

4. **Imputation Techniques:** Imputation is a family of techniques which is based on the idea of filling in all of the missing values in the dataset, so that missingness is not a concern. The key decision for imputation techniques is the method used to determine how to fill-in the values. The strength of imputation relies entirely on the strength of the method used to impute the missing values. There are some procedures which are "ad-hoc", and correspondingly do not perform particularly well in general. These would include, for instance mean imputation or last observation carried forward, and are generally advised against. Instead, imputation techniques should be based on specific (conditional) distributions that are posed, and fit to the data, and then sampled from; the techniques will be valid whenever the posited model is valid, and this acts as a method of making explicit assumptions about what drives the missingness. Note, likelihood methods which model $f(Y_i|X_i)$ can be viewed as a type of imputation, which underscores why these methods are valid with MAR data. An imputation technique is deemed either **single imputation** where one dataset is imputed and analyzed or **multiple imputation** where (as it sounds) multiple imputed datasets are analyzed. Multiple imputation is preferable as it better accounts for the sampling variability in the imputation procedure.

In the following sections we will explore in more detail both weighting techniques and imputation techniques. Becoming familiar and comfortable with these methods will assist you not only in this course, but in your future statistics courses, as (generally) these methods are applicable outside of the setting of longitudinal data.

Weighting Methods

Consider the special case of dropout to illustrate the idea of weighting methods. We can define a “dropout indicator”, which represents the observation time that the individual drops out of the study and does not return. We denote this $D_i = 1 + \sum_{j=1}^{K} R_{ij}$. A fully observed individual will thus has $D_i = K + 1$. Now, consider the probability that an individual who made it to at least stage j, is observed past stage j. That is, define

$$\pi_{ij} = P(D_i > j | D_i \geq j).$$

These probabilities could be estimated, for instance, through the use of sequential logistic regressions, and could take into account all observed values of $\{Y_i, X_i, R_i\}$ up to time j. Once we have these probabilities, there are two general techniques for incorporating this information into an analysis: we can either perform a weighted complete case analysis, or a weighted available data analysis.

Weighted Complete Case

If we consider just the full responders, then the probability that any one of them was included in the sample is given by

$$\pi_i = P(D_i > K) = \prod_{j=1}^{K} P(D_i > j | D_i \geq j) = \prod_{j=1}^{K} \pi_{ij}.$$

Individuals who have a large π_i are likely to have been observed and those with a small π_i are unlikely to have been observed. As a result, if we consider just the complete cases, those with a large π_i are likely to be well represented within the sample (compared to the overall population) whereas those with a small π_i are likely to be underrepresented (compared to the overall population). If we consider two individuals, $\pi_i = 0.95$ and $\pi_{i'} = 0.05$ then it is clear that for people like individual i we have a good sample of observed values in our complete cases. If there were 100 people in the full sample who were like individual i, we would expect to have 95 of them in the complete sample, only missing values for 5 of them. For individual i', however, this is not the case: if there were 100 individuals like this person to begin, we would only expect 5 in our complete case, missing 95 of them! If we want to re-balance our sample, then counting individual i for $w_i = \frac{1}{0.95} \approx 1.05$ individuals will mean that the 95 observed individuals (each counting at w_i) will represent 100 people in the weighted sample. Similarly, if we take $w_{i'} = \frac{1}{0.05} = 20$, then the 5 individuals each weighted at $w_{i'}$ will also represent 100 people. Based on this logic, we can then conduct a complete case analysis where each individual is weighted using

$$w_i = \frac{1}{\pi_i},$$

and so long as the data are MAR and our models are correctly specified, this analysis will be valid.

Weighted Available Data

While the complete case analysis presented above is valid under the specified assumptions, it does not make efficient use of the available data. Just like with a complete case analysis under the MCAR assumption, a complete case weighted analysis with MAR data will be less efficient than a similar analysis which uses all of the observed data. Instead of using a single weight for each individual, we can think of computing the weights for each individual at each time point, j. That is,

$$w_{ij} = \frac{1}{P(D_i > j)} = \left[\prod_{\ell=1}^{j} \pi_{i\ell}\right]^{-1}.$$

Then these weights can be incorporated into an analysis which exploits all available data, where an individuals observations at time j are given weight based on $R_{ij}w_{ij}$. Note that, depending on the specific modelling approach for estimating π_{ij} it may be the case that some weights estimated are incredibly large (when $\pi_{ij} \approx 0$). These weights may have undue influence on the estimation procedure, leading to a loss of precision. When using these weighting schemes it is worth considering the distribution of weights, in full, to ensure that there do not appear to be any which have too strong of an influence on the outcome. If so, other techniques to account for the missingness (or alternative *stabilized weights* can be employed).

This type of weighting is commonly employed for the GEE approach. Recall that, in general, the GEE estimators for β solve

$$\sum_{i=1}^{n} D_i' V_i^{-1}\left\{Y_i - \mu_i(\beta)\right\} = 0.$$

An available data analysis for GEE then would include

$$\sum_{i=1}^{n} D_i^{O'} V_i^{O-1}\left\{Y_i^O - \mu_i^O(\beta)\right\} = 0,$$

where the observed indicators on each of the corresponding components represents inclusion of only the values observed from each individual. A weighted analysis of GEE, often called IPW-GEE (for inverse probability weighted GEE) can proceed by defining a weight matrix for each individual, $W_i = \text{diag}(R_{ij}w_{ij} \mid j = 1, \dots, K)$. Then, if this weight Matrix is correctly specified, the IPW-GEE estimators can be obtained by solving

$$\sum_{i=1}^{n} D_i' V_i^{-1} W_i\left\{Y_i - \mu_i(\beta)\right\} = 0.$$

It is worth noting that, if V_i is not a diagonal matrix (that is, we do not assume working independence) then we **must** observed X_{ij} for **all** time points, even when we do not observe the corresponding Y_{ij}, in order for these estimating equations to be computable. Note that this is not as strong of an assumption as it may first seem. In an analysis which focuses predominantly on baseline factors, it may very well be the case that X_i is completely observed at all time points, even when the outcome is not. Second, recall that V_i need not be correctly specified for consistent estimation of β. The correct specification of the variance matrix simply improves efficiency. As a result, if there are concerns with the observed values for X_i, then we can simply employ a working independence assumption, sacrificing efficiency for valid estimation. While V_i need not be correctly specified, W_i must be.

Imputation Techniques

The process of imputation, whereby we fill in values that are otherwise missing from the dataset, is useful particularly when specifying a model for the missingness (based on the observed data as well as any available auxiliary information) is possible. In contrast to weighting techniques, imputation techniques directly model the conditional distributions for the data, and use this to fill-in information. In general there are two choices to make when employing imputation: (1) single versus multiple imputation, and (2) the model used to impute from. It is always preferable to use multiple imputation, whenever it is possible, and we will focus on it. The idea with multiple imputation is to simply repeat the imputation process many times, and combine the resulting estimates; failing to do so will have the effect of underestimating the variability in the standard errors (which will produce confidence intervals which are too small). For multiple imputation, we think of doing the imputation process m times, which results in m estimators for $\widehat{\beta}^{(k)}, k = 1, \dots, m$, We can then define

$$\widehat{\beta} = \frac{1}{m}\sum_{k=1}^{m}\widehat{\beta}^{(k)},$$

and we further take

$$\widehat{\text{cov}}\left(\widehat{\beta}\right) = \frac{1}{m}\sum_{k=1}^{m}\text{cov}\left(\widehat{\beta}^{(k)}\right) + \frac{m+1}{m(m-1)}\sum_{k=1}^{m}\left(\widehat{\beta}^{(k)} - \widehat{\beta}\right)\left(\widehat{\beta}^{(k)} - \widehat{\beta}\right)'.$$

Here, the covariance of each $\widehat{\beta}^{(k)}$ is estimated based on the technique used to estimate β. This averaging effect will produce valid estimators, and we can see that as m increases, the additional variance in our estimator tends to 0.

Knowing how to combine the multiple estimators based on imputed data, we have to make a decision regarding how to actually impute the values. There are several different methods, and what is possible will depend on the pattern of missingness (e.g., whether or not the missingness is monotone). Suppose that the missingness is monotone (as will often be the case), and suppose further that this is ordered in such a way that it resembles dropout. The idea with imputation with these data will be to first impute values at $t = 2$ using $\{Y_{i1}, X_{i1}\}$. Then, once everyone has an imputed value, we will use $\{\widehat{Y}_{i2}, X_{i2}, Y_{i1}, X_{i1}\}$ to impute values for Y_{i3}. We can continue this through to Y_{iK}, using the previously imputed values as predictors wherever necessary. Conceptually, each of these imputation models can be based on a suitable GLM for the outcome, say given by

$$g\left(E\left\{Y_{i\ell}\middle| Y_{i1}, \dots, Y_{i,\ell-1}, X_i\right\}\right) = Z_{i\ell}'\gamma_\ell.$$

Note that these models are imputation models not analysis models, and so these are distinct from the marginal models discussed earlier in the term. In theory these models could be fit and used to estimate the mean for each individual according to the iterative process above, however, this presents a problem. Prediction in GLMs is deterministic not random.

Our goal with multiple imputation was to account for added variability by sampling repeatedly from a set of possible values to give us a robust estimator. If we were to simply use the imputation model for prediction, this would not allow for different estimates of the imputed values. Instead, we wish to predict from the distribution, which would involve sampling from the residuals as well. In a linear regression, this amounts to taking the estimated mean $Z_{i\ell}'\widehat{\gamma}_\ell$ and adding onto it a draw from a normally distributed error term, with variance given by $\widehat{\sigma}_\ell^2$ from the OLS fit. In the event of a binomial outcome, we can take a draw from a binomial random variable where the probability of success is given by $\text{expit}(Z_{i\ell}'\widehat{\gamma}_\ell)$. Drawing from the residual distribution produces randomness in our estimates, and more closely accounts for the uncertainty in our imputation, however, it does not go far enough. Recall that we are using imputed values as predictors in our models. If the imputed values were exactly correct, that would lead to consistent estimates of the γ parameters; however, since these values are not actually observed, we are in fact **underestimating** the total variation needed to account for the imputation techniques.

To illustrate this point dramatically, imagine a trial with $K = 100$, and an imputation occurring for an individual with only the first observation recorded. At the 100th time step, the model we are using is based on $E[Y_{i,100}|Y_{i,1}, \dots, Y_{i,99}]$ where for this individual, $\{Y_{i,2}, Y_{i,3}, \dots, Y_{i,99}\}$ are all estimated values. The errors that propagate through using these predictions in place of truly observed values is going to grow very large, where we will be incredibly uncertain about the estimates at later stages of the analysis.

Instead, we will also add in additional variability by drawing the estimated $\widehat{\gamma}_\ell$ from a distribution as well. For those of you who have seen Bayesian analysis, we wish to sample $\widehat{\gamma}_\ell$ from its posterior distribution. For those of your who are unfamiliar with Bayesian analysis, we can simply think of treating our estimators as being random variables where the added variation is due (in part) to the fact that we are using imputed values to estimate them. Then, regression based imputation proceeds by fitting a suitable regression model at the first step, randomly sampling the estimated regression parameters from a suitable posterior distribution, and then drawing imputed values from the residual distribution of the model (based on the predicted mean). These values are then treated as observed and this sequence repeats. We then repeat the entire analysis several times, and combine the estimators through the multiple imputation procedure discussed above.

Predictive Mean Matching

As an alternative to the aforementioned procedure, we can use the sequential regression models in a slightly different way. If we fit the model, and randomly draw the regression coefficients exactly as above, then we could generate a mean prediction for every observation in the data (including those that were observed). Then, for each missing value, we consider the κ closest values to its mean prediction, among the individuals with an observed value. Here κ is some integer selected in advance. Among these κ individuals, we randomly select one of them, and use their observed value as the imputed value for the individual. We then proceed at the next step. Note that here the predictions being used for matching are based on the means, and random errors need not be sampled. Just as with regression techniques, these matches and imputed values are then repeated many times, and combined using the multiple imputation formula.

There are two primary benefits to predictive mean matching over regression-based imputation. First, it is generally more robust to misspecification in the regression models. Intuitively, this is because we only need the models to tell us which individuals are similar, rather than predicting their outcomes specifically. Second, because the imputed values are borrowed from actual observations, it is guaranteed that all values of the outcome are plausible (which may not be the case with regression values).

Likelihood as an Imputation Method

We have noted that, when data are MAR, likelihood based techniques (e.g., linear mixed effect models or transition models) can accommodate missingness. This assumes that the likelihood is correctly specified, but that assumption is already required for the validity of these methods. In such a situation we can see that $f(Y_i^{\text{O}}|X_i) = f(Y_i|X_i) = f(Y_i^{\text{M}}|X_i)$, permitting analysis to proceed. One way that we can view this is as a form of (single) imputation, where the values for the Y_i^{M} are imputed based on the model $E[Y_i^{\text{M}}|Y_i^{\text{O}}, X_i]$. This can often times be employed, alongside an Expectation-Maximization (EM) algorithm to get likelihood-based parameter estimates in missing data situations.

When ML estimators are available, they are often preferable to multiple imputation, since they are typically more efficient and less computationally demanding. However, multiple imputation does have several benefits which are worth considering over a likelihood based analysis.

1. Multiple imputation can leverage auxiliary information, which is not directly relevant to the analysis, but which may help predict missing values.

2. Multiple imputation can be easily generalized to predict missing values for the covariates in addition to the outcomes, which is not as easily accommodated by likelihood techniques.

3. Multiple imputation allows you to use techniques which are not likelihood based (e.g. GEE) under the assumption of MAR data.

4. Multiple imputation also allows a sensitivity analysis to be conducted, by simply altering the imputation models and seeing how much the parameter estimates change. This provides an in-built way of determining how sensitive the estimates are to different mechanisms, and removes uncertainty regarding incorrect models.

NMAR Data and Conclusions

The techniques presented allow us to overcome issues with data that are MCAR or MAR, but not when the data are NMAR. In this case the only recourse that you have as an analyst is to jointly model (Y_i, R_i), which will depend on the specific situation. Still, in many situations MI or weighting will provide flexible ways to accommodate missingness, both in the longitudinal setting and beyond.

Analyses which ignore missigness (without explicitly mentioning that they are performing a complete case or available data analysis on the assumption of MCAR missingness) should not be trusted. Analyses which use ad-hoc imputation techniques (such as last observation carried forward) are also susceptible to incredible bias, and are acceptable in only very limited contexts. At a minimum, models for missing data provide a way of making explicit the assumptions that you are making, and allow you to test the sensitivity of your estimates to these assumptions. Handling missing data in a way which is valid should be considered essential for any analysis in the “real world”.