Chapter 11
Missing Data

Missing data are a fact of life in medical research. Subjects refuse to answer sensitive questions (e.g., questions about income or drug use), are unable to complete an MRI exam because of metallic implants, or drop out of studies and do not contribute further data. In each of these cases, data are "missing" or not complete. How should this be accommodated in a data analysis? Statistical computing packages will typically drop from the analysis all observations that are missing any of the variables (outcomes or predictors). So, for example, a linear regression predicting a patient's number of emergency room visits from their age, gender, race, income, and current drug use will drop any observation missing even one of those variables. Analysis of data using this strategy is called *complete case analysis* because it requires that the data be complete for *all* variables before that observation can be used in the analysis.

Complete case analysis is simple and the default for statistical analysis packages. But it can be inefficient and lead to biased estimates. Imagine a situation in which the first 20% of the sample is missing age information, the second 20% is missing gender information and so on, with the last 20% missing drug use information. Even though, in a sense, 80% of the predictor data is present, there will be no observations left for a complete case analysis.

Further, data are often missing for a reason related to the outcome under study. As examples, sicker patients may not show up for follow-up visits, leading to overly optimistic estimates based on the data present. Or those patients staying in the hospital longer may be the sicker ones (with the better-off patients having been discharged). This might lead us to the erroneous conclusion that longer stays in the hospital produce poorer outcomes, so why check-in in the first place? A basic message is that we need to think carefully about why the data are missing. This may influence how we will handle it and guide us to ways we can avoid biased estimates.

How can these drawbacks be overcome? If we could intelligently fill in the missing data to obtain a complete dataset then we could use standard methods without concern. Of course, we would need to account for the fact that the missing data are estimated and not actual measured quantities in our sample. This is the

E. Vittinghoff et al., *Regression Methods in Biostatistics*, Statistics for Biology and Health, DOI 10.1007/978-1-4614-1353-0_11,

basic idea behind *multiple imputation*, which we discuss in Sect. 11.5. Or perhaps, we could use members in the sample with complete data to represent those with missing data. For example, suppose heavy drug users tended to drop out of a study at twice the rate of other participants. Then we could "double-count" the heavy drug users who did not drop out of the study by weighting their contributions to the analysis more heavily. This is the basic idea behind *inverse probability weighting* (IPW) which we cover in Sect. 11.9.3. In either case, the key is to use the data on hand, along with anything we might know about why the data are missing in order to infer the missing data. Not surprisingly, this strategy will only work if the values of the missing data are, to some extent, predictable from the observed data.

We begin this chapter with some simple illustrations of what can go wrong when there is missing data. This naturally leads to consideration of why the data are missing and some more formal classifications of the missing data process in Sect. 11.2. We discuss some simple strategies that have been used in the past to accommodate missing data. We then consider common missing data scenarios: missing predictor values (with at least some of the associated outcomes being measured) and complete (or nearly complete) predictor values, but missing outcomes. For this latter situation, we consider three different ways in which the data came to be missing. The two strategies mentioned above—multiple imputation and inverse probability weighting—are then considered in more detail as principled approaches to missing data. In Sect. 11.9.1, we also describe situations with missing outcome data in longitudinal studies that can be addressed by using maximum-likelihood methods like mixed models. These "automatically" infer the missing data with the advantage of not requiring explicit modeling. Our focus throughout this chapter is on the effect that missing data has on estimation of regression coefficients, but missing data can also cause predictions to be biased.

11.1 Why Missing Data Can Be a Problem

To more clearly demonstrate why missing data can be a problem, we consider two examples using the HERS study (see Sect. 3.1). In the first, we consider linear regression of SBP on glucose level, BMI, and whether the person was Caucasian or not using only the data from the fourth visit. For that visit 443 of the 1,871 observations had missing data for glucose. The second considers a longitudinal data setting in which SBP is measured over two visits with the second one potentially missing, as would happen with participants dropping out of a study.

11.1.1 Missing Predictor in Linear Regression

Standard regression of SBP on blood glucose level (`glucose`), whether a person is Caucasian or not (`white`), and their BMI (`bmi`) using the 1,871 participants with

Table 11.1 Regression of SBP using a complete case analysis

```
. regress sbp glucose white bmi

      Source |       SS       df       MS              Number of obs =    1385
-------------+------------------------------           F(  3,  1381) =    2.69
       Model |  2855.36663     3  951.788878            Prob > F      =  0.0450
    Residual |  488496.255  1381   353.72647            R-squared     =  0.0058
-------------+------------------------------           Adj R-squared =  0.0037
       Total |  491351.622  1384  355.022848            Root MSE      =  18.808

------------------------------------------------------------------------------
         sbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     glucose |   .0294818   .0126344     2.33   0.020     .0046972    .0542665
       white |  -1.537906   1.689423    -0.91   0.363    -4.852019    1.776207
         bmi |   .0644021   .0934208     0.69   0.491      -.11886    .2476641
       _cons |    132.716    3.29506    40.28   0.000     126.2521    139.1799
------------------------------------------------------------------------------
```

Table 11.2 Regression of systolic blood pressure using imputed glucose values

```
. regress sbp imp_glucose white bmi

      Source |       SS       df       MS              Number of obs =    1750
-------------+------------------------------           F(  3,  1746) =    5.34
       Model |  5766.65623     3  1922.21874            Prob > F      =  0.0012
    Residual |  628318.844  1746  359.861881            R-squared     =  0.0091
-------------+------------------------------           Adj R-squared =  0.0074
       Total |   634085.5  1749  362.541738            Root MSE      =   18.97

------------------------------------------------------------------------------
         sbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 imp_glucose |   .0338782   .0122595     2.76   0.006     .0098333     .057923
       white |  -2.209204    1.49944    -1.47   0.141    -5.150092    .7316834
         bmi |   .1364681    .083551     1.63   0.103    -.0274025    .3003388
       _cons |    130.385   2.937854    44.38   0.000     124.6229    136.1471
------------------------------------------------------------------------------
```

data for visit 4 in HERS gives the output in Table 11.1. We can see that only 1,385 subjects are used in the complete case analysis. This is because, in addition to the 443 participants missing data on glucose, there are 85 missing values for SBP, 110 missing values for BMI, and 3 missing values for white (and some overlap in the missing data). We will concentrate on the missing glucose values to introduce the main ideas.

Glucose values are fairly strongly related to the other predictors, so there is some hope in filling in the missing values relatively accurately; a regression of glucose on SBP, BMI, white, current smoking status, and whether or not a woman develops diabetes has an R^2 of 0.44. We could use this regression to generate predicted values for 372 of the 443 of the missing glucose values—we cannot fill them all in because there is missing data for BMI, white, and diabetes. Using the predicted values in place of the missing glucose values, we can now use more of the data. Table 11.2 gives the regression results, where `imp_glucose` is equal to the actual value of glucose when it is available and the predicted (imputed) value of glucose when it is missing. Some of the regression coefficients are noticeably different,

Table 11.3 Regression of systolic blood pressure using multiply imputed glucose values

```
. mi estimate: regress sbp glucose white bmi

Multiple-imputation estimates                   Imputations     =          5
Linear regression                               Number of obs   =       1750
                                                Average RVI     =     0.0106
                                                Complete DF     =       1746
DF adjustment:   Small sample                   DF:     min     =    1046.77
                                                        avg     =    1557.86
                                                        max     =    1743.23
Model F test:       Equal FMI                   F(   3, 1644.5) =       4.57
Within VCE type:          OLS                   Prob > F        =     0.0034

------------------------------------------------------------------------------
         sbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     glucose |   .0269531   .0116979     2.30   0.021     .0039991     .049907
       white |  -2.199165   1.500637    -1.47   0.143    -5.142402    .7440727
         bmi |   .1467563   .0836029     1.76   0.079    -.0172179    .3107305
       _cons |   130.8553   2.930445    44.65   0.000     125.1077    136.6029
------------------------------------------------------------------------------
```

for example, the BMI coefficient has approximately doubled in size, has a smaller standard error, and has a smaller *p*-value. All the standard errors are smaller. This is an illustration of what is called *single imputation*, because we have filled in or imputed the missing data a single time.

But this is not quite legitimate. In this analysis, the software does not distinguish between the imputed glucose values and the actual measured values. So the information content of the dataset is overestimated and standard errors may be falsely small. A solution to this is to impute the glucose values but properly account for the actual amount of information available. One way to do this is to use *multiple imputation* which we describe in more detail in Sect. 11.5. Table 11.3 gives the results of such an analysis.

The results are very similar to the singly imputed analysis. Because we have not imputed a large portion of the data, the standard errors are only slightly increased in the multiply imputed approach compared to the singly imputed. Notably, the standard errors remain smaller than those from the complete case analysis.

Using imputation to handle the missing data for this example has had two benefits: it may have slightly reduced a bias in the original coefficients and we have been able to successfully utilize more of the data, thereby reducing the standard errors. Multiple imputation is a flexible methodology and can be used to impute not only the predictor, but also the outcomes.

11.1.2 Missing Outcome in Longitudinal Data

To illustrate the potential problems with drop out in longitudinal data, we used the HERS study, for which there is actually very little drop out. We consider the outcome of SBP using data only from baseline and year 1. In the complete dataset,

Table 11.4 Analysis of HERS data using complete data and generalized estimating equations

```
. xtgee sbp visit bmi baseline_dm, i(pptid) corr(exch) robust

GEE population-averaged model                 Number of obs      =       5368
Group variable:                      pptid    Number of groups   =       2761
Link:                              identity   Obs per group: min =          1
Family:                            Gaussian                 avg =        1.9
Correlation:                   exchangeable                 max =          2
                                              Wald chi2(3)       =      67.85
Scale parameter:                   357.8178   Prob > chi2        =     0.0000

                                (Std. Err. adjusted for clustering on pptid)
------------------------------------------------------------------------------
             |             Semirobust
         sbp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       visit |   .2836451   .3388382    0.84   0.403    -.3804657    .9477558
         bmi |   .1385708   .0598246    2.32   0.021     .0213167     .255825
 baseline_dm |   5.511153   .7814551    7.05   0.000      3.97953    7.042777
       _cons |     129.66   1.723401   75.23   0.000     126.2822    133.0379
------------------------------------------------------------------------------
```

the average SBP at baseline was 135.1 and at year 1 was 135.2, so very little change from baseline to year 1.

To quantify the change from baseline to visit 1, we used regression analyses of SBP on visit (baseline, coded as 0, or the year 1 visit, coded as 1), BMI and whether the participant had diabetes at baseline (yes/no). Since we have repeated measures, we could use either GEEs (via `xtgee`) or mixed models (via `xtmixed`) to analyze the data. Tables 11.4 and 11.5 give the results using the complete data.

The two analyses give virtually the same results. Focussing on the visit term, there is a small and nonstatistically significant increase from baseline to year 1 (estimated to be about 0.28), consistent with the raw data.

We next simulated drop out at year 1 on the basis of either the baseline SBP or the year 1 SBP, but keeping all the data for the baseline visit. In either case, those with higher SBP were dropped at higher rates than those with lower SBP. In the situation where drop out depended on baseline SBP, we "dropped" 1,461 participants at year 1 and "retained" 1,302. Those retained had average SBP at year 1 of 127.5 (range 85–196) and those dropped had average SBP 143.9 (range 93–220). So there is a distinct difference between those dropped and retained, but there is also considerable overlap. Importantly, in the incomplete data, the average SBP drops from 135.1 at baseline to 127.5 at year 1, quite different from the complete data.

We, therefore, anticipate trouble with the analysis using the incomplete data since the average SBP drops between baseline and the year 1 visit. Ideally, a technique that handles missing data well will give results similar to the analysis of the complete data (e.g., Table 11.4). Table 11.6 gives the regression coefficient tables for the situation where drop out depends on SBP at baseline.

Now we see a completely different story. The generalized estimating equations (GEEs) approach incorrectly estimates a highly statistically significant drop in SBP

Table 11.5 Analysis of HERS data using complete data and maximum likelihood

```
. xtmixed sbp visit bmi baseline_dm || pptid:

Mixed-effects REML regression                   Number of obs      =      5368
Group variable: pptid                           Number of groups   =      2761

                                                Obs per group: min =         1
                                                               avg =       1.9
                                                               max =         2

                                                Wald chi2(3)       =     73.13
Log restricted-likelihood = -22872.471          Prob > chi2        =    0.0000

------------------------------------------------------------------------------
         sbp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       visit |   .2843892    .338578     0.84   0.401    -.3792114     .9479898
         bmi |   .1392584   .0587622     2.37   0.018     .0240865     .2544302
 baseline_dm |   5.507891   .7583126     7.26   0.000     4.021625     6.994156
       _cons |   129.6413   1.677004    77.31   0.000     126.3544     132.9282
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
pptid: Identity              |
                  sd(_cons)  |   14.40895   .2784187     13.87346     14.9651
-----------------------------+------------------------------------------------
               sd(Residual)  |   12.28843   .1702939     11.95916     12.62678
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01)= 1055.76 Prob >= chibar2 = 0.0000
```

of 1.32 from baseline to year 1. Interestingly, the mixed model approach (which uses maximum likelihood, or ML, to fit the model) gives estimates similar to the complete data analysis with a small estimated increase which is not statistically significant. For the other coefficients, the two analyses give similar results, both to one another and to the complete data analyses.

Finally, we also simulated a dataset where drop out at year 1 depended on year 1 SBP in a fashion similar to that described above. This differs from the previous case in that whether or not a participant was included in the dataset depended on *unobserved* quantities. Table 11.7 gives the results with drop out that depends on SBP at year 1. Now both the analyses give very severely biased estimates of the visit effect, though other coefficients are little affected.

There are several important messages from this example. When drop out is dependent on previous, observed values, some analysis methods such as GEEs can give badly biased estimates whereas others such as mixed model methods, based on maximum likelihood, are less affected. The situation when drop out depends on unobserved values is much more serious and leads to severe bias using either method.

Table 11.6 Analysis of HERS data with drop out depending on baseline outcome using GEEs and ML

```
. xtgee sbp visit bmi baseline_dm if miss_mar==0, i(pptid) corr(exch) robust

------------------------------------------------------------------------------
             |             Semirobust
         sbp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       visit |  -1.320828    .431427    -3.06   0.002    -2.166409   -.4752463
         bmi |   .1041894   .0622733     1.67   0.094    -.0178641    .2262428
 baseline_dm |   5.787856    .813257     7.12   0.000     4.193901     7.38181
       _cons |   130.5635   1.790897    72.90   0.000     127.0534    134.0736
------------------------------------------------------------------------------

. xtmixed sbp nvisit bmi baseline_dm if miss_mar==0 || pptid:

------------------------------------------------------------------------------
         sbp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       visit |   .5912612   .4179003     1.41   0.157    -.2278083    1.410331
         bmi |   .1084238   .0625694     1.73   0.083    -.0142101    .2310576
 baseline_dm |   5.894762    .801877     7.35   0.000     4.323111    7.466412
       _cons |   130.4142   1.779439    73.29   0.000     126.9266    133.9019
------------------------------------------------------------------------------
```

Table 11.7 Analysis of HERS data with drop out depending on unobserved outcome using GEEs and ML

```
. xtgee sbp visit bmi baseline_dm if miss_nmar==0, i(pptid)corr(exch) robust

------------------------------------------------------------------------------
             |             Semirobust
         sbp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       visit |  -9.889191   .3840043   -25.75   0.000    -10.64183   -9.136557
         bmi |   .0962627   .0574965     1.67   0.094    -.0164284    .2089539
 baseline_dm |   4.985786   .7507309     6.64   0.000     3.514381    6.457192
       _cons |   131.0006   1.656733    79.07   0.000     127.7534    134.2477
------------------------------------------------------------------------------

. xtmixed sbp visit bmi baseline_dm if miss_nmar==0 || pptid:

------------------------------------------------------------------------------
         sbp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       visit |   -8.35655   .4240134   -19.71   0.000    -9.187601   -7.525499
         bmi |   .1043524   .0573602     1.82   0.069    -.0080715    .2167762
 baseline_dm |   5.027966     .73204     6.87   0.000     3.593194    6.462738
       _cons |   130.7572   1.630634    80.19   0.000     127.5613    133.9532
------------------------------------------------------------------------------
```

11.2 Classifications of Missing Data

The previous example has shown that the mechanism that causes the data to be missing can be very important. It is, therefore, useful to develop categorizations of missing data mechanisms that either are or are not likely to cause misleading results.

To motivate some of the considerations, we use the Steroids for Corneal Ulcer Trial (SCUT). SCUT was a randomized clinical trial to gauge the effectiveness of a steroid treatment (steroid eye drop versus a placebo) on visual acuity (VA) in people with bacterial corneal ulcers. The primary outcome of VA for SCUT was measured on a scale called logmar, which is short for logarithm (base 10) of the minimum angle of resolution. A logmar of 0 corresponds to 20/20 vision, a logmar of 1 to 20/200 vision, and, in general, a logmar of x corresponds to a vision of $20/(20\times 10^{x})$ on an eye chart. Follow-up measures were taken at 3 weeks, 3 months, and 12 months. The predictors, all measured at the enrollment visit, are baseline VA, ulcer location (whether it covered the center of the eye), ulcer size (in square mm), and the type of infecting organism (gram positive versus gram negative). It is easy to envision what the full or "complete" data would consist of for this example: all participants have all their predictors measured at baseline and outcome information at baseline and each of the three follow-up times.

For regression analyses, a key distinction with regard to missing information is whether or not we have a considerable percentage of observations for which the predictors are missing but we have a measured outcome. This is important because, in regression analyses we typically model the distribution of the outcome variable and treat the predictor variables as fixed (see Sect. 3.3.3). If the predictor variables are missing for some observations (e.g., glucose values in the HERS example) then we need a method for inferring those missing values and assumptions will have to be made with respect to their distribution.

In the HERS example above, we used multiple imputation to build a model, temporarily treating glucose as an *outcome* variable in a linear regression model. That model assumes that glucose follows a normal distribution (for fixed values of the predictors of that model). That is, we have to make a distributional assumption about a variable that was a predictor in the original model (a regression of blood pressure on glucose), something we did not have to do before.

In more complicated examples, with multiple missing predictors, we would have to account for not only the distribution of each missing predictor by itself but also the joint distribution, including aspects such as correlations between predictors. In the not uncommon situation where the predictors with missing values consist of nominal, ordinal, and skewed variable types, specifying a distribution for how they are all jointly associated is a daunting task.

A simpler situation to handle is when there is little or no missing information on the predictors and missing data are mainly in the outcome, or both outcome and predictors are missing (as when a participant drops out of a study). In such cases, we can focus on the outcome variable, for which we are already hypothesizing a distribution, and categorize the missing data mechanisms relatively succinctly.

11.2.1 Mechanisms for Missing Data

Because we will want to describe the way in which the data came to be missing, it is worthwhile to consider a formal statistical model and develop some notation.

In that spirit, we envision a "complete" dataset, where all the data are present. We will think of this in the context of a longitudinal cohort study with regularly scheduled observation times, but the ideas apply more generally. Our complete dataset would be one with all outcome and all predictors measured on each person for each visit. Next, consider one of the variables that actually has missing data.

Let R_{it} be 1 if the ith participant has a measured value of the variable with missing data at visit time t and zero if it has "become" missing. So R is a binary indicator of whether a data value is present or not. For each variable that has missing data, we can now classify various missing data mechanisms by how they relate to the probability that $R_{it} = 1$. If factors are unrelated to this probability, then they have no bearing on the missing data process.

A common practice with missing data in a longitudinal study is to look at baseline characteristics of participants who had missing data later in the study. If variables differ significantly between those with and those without missing data (e.g., their age, gender, or baseline value of the outcome) then we can begin to understand what is related to $R_{it} = 1$. For example, Splieth et al. (2005) obtained a baseline oral health assessment of all first- and second-grade schoolchildren in a city in Germany. They compared the oral health of children whose parents did and did not allow them to participate in a cavity prevention program and longitudinal follow-up. They found that the children not participating were older and had poorer dental health compared to the participants. Failure to recognize this selective participation would result in biased estimates of average values. The formal classification scheme we consider next takes the idea of relating missing data to baseline covariates a step further.

11.2.1.1 Missing Completely at Random (MCAR)

There are three common classifications of the missing data process. Data are said to be *missing completely at random* (MCAR) if $\mathrm{P}(R_{it} = 1)$ does not depend on any of the variables. For SCUT this would mean, for example, that the probability a logmar value at 3 months was missing was unrelated to the previous, current or future logmar values and also unrelated to visual acuity, ulcer location, ulcer size, or type of infecting organism. If we observed, for example, that participants with very poor logmar values at baseline were less likely to return then we would know that the MCAR scenario would not apply.

With $\mathbf{X}$ representing all the predictors and $\mathbf{Y}$ representing all the outcomes, this can be formally stated as

$$\mathrm{P}(R_{it} = 1|\mathbf{Y}, \mathbf{X}) = \mathrm{P}(R_{it} = 1), \tag{11.1}$$

i.e., the probability of the data being missing is not associated with any part of the data. Another way to interpret this is that knowing the values of the outcomes and predictors would not change our estimate of the likelihood that a particular data

value is missing. While a useful conceptual "baseline" definition, MCAR is often not a reasonable assumption. For example, in longitudinal studies where there is missing data, there is almost invariably more missing data later in the study. So, at the very least, the predictor time or visit would be associated with the probability that an observation is missing.

11.2.1.2 Covariate-Dependent Missing Completely at Random (CD-MCAR)

A minor, but important, variation of this definition is *covariate-dependent missing completely at random* (CD-MCAR), which is mainly applicable to missing *outcome* data. In this situation, the probability of the outcome being missing can depend on the predictors which are part of the statistical model but does not depend on the other outcomes. With $\mathbf{X}^{\text{obs}}$ representing all the observed information for predictors which will be included in our model, we would formally write this as

$$\mathrm{P}\left(R_{it} = 1|\mathbf{Y}, \mathbf{X}\right) = \mathrm{P}\left(R_{it} = 1|\mathbf{X}^{\text{obs}}\right). \tag{11.2}$$

For SCUT this would mean, for example, that the probability a logmar value was missing was unrelated to the 3 weeks, 3 months, or 12 months logmar values but could be related to visit, VA, ulcer location, ulcer size, or type of infecting organism. If we observed, after accounting for differences due to the predictors, that participants with very poor logmar values at 3 weeks were more likely to return at 3 months then we would know that the covariate-dependent MCAR scenario would not apply.

11.2.1.3 Missing at Random (MAR)

A yet more flexible specification is that data are *missing at random* (MAR). This assumption handles a variety of more plausible scenarios. In MAR, the probability of missing data may depend not only on the covariates in the model but also on observed outcomes.

With $\mathbf{Y}^{\text{obs}}$ representing all the observed outcome information, formally this would be written as

$$\mathrm{P}\left(R_{it} = 1|\mathbf{Y}, \mathbf{X}\right) = \mathrm{P}\left(R_{it} = 1|\mathbf{Y}^{\text{obs}}, \mathbf{X}^{\text{obs}}\right). \tag{11.3}$$

In the SCUT example, the MAR scenario would allow for people with worse VA at 3 weeks or 3 months to be missing more frequently at 12 months and also to depend on visit, baseline logmar, VA, ulcer location, ulcer size, or type of infecting organism. In the HERS example, in Table 11.6, we artificially created data that was MAR.

11.2.1.4 Missing Not at Random (MNAR)

Finally, it may be that the probability a data value is missing depends on unobserved quantities, for example, the outcome we would have measured were it not missing. For instance, consider SCUT patients with identical baseline and 3 week visual acuities. Suppose the ones whose VA did not improve are more likely to make the 3 month visit (to get checked by the doctors). Then the fact that the data are missing would depend on the unobserved 3-month outcome. This scenario is called *missing not at random* or MNAR. In the HERS example, in Table 11.7, we artificially created data that was MNAR.

More formally, simplification of the model for $\mathrm{P}(R_{it} = 1|\mathbf{Y}, \mathbf{X})$ would not be possible as we did in, for example, (11.2). Unfortunately, but perhaps not surprising and because MNAR depends on *unobserved* quantities, we cannot verify or rule out a MNAR process from the observed data alone. Instead, if we suspect the data are MNAR the best we can do is conduct sensitivity analyses. One way to do so is via multiple imputation, described in Sect. 11.5.

Why are these characterizations important? Their utility is that we can now describe more carefully when standard types of analyses can be expected to give answers free of bias due to the missing data. We give more details and caveats beginning in Sect. 11.5 but in essence:

- When the data are MCAR, any method of analysis will give unbiased answers.
- When the outcome data are CD-MCAR, and those covariates are included in the statistical model, any method of analysis will give unbiased answers for regression coefficients and predicted values. Care still needs to be taken with calculations that average over values of the covariates (e.g., an average of the predicted values, or estimation of marginal effects) because those may not have the same distribution of covariate values as in the complete data.
- When the outcome data are MAR, correctly specified, likelihood-based analysis methods (e.g., mixed models) will give unbiased answers, but other methods (e.g., GEEs) may not.
- When the data are MNAR, any standard method of analysis may be biased.

Moving from MCAR to CD-MCAR accommodates the common situation in which missing data depend on measured covariates. Going from CD-MCAR to MAR allows even more elaborate dependence of the missing data—on measured covariates *and* outcomes—and will therefore include missing data mechanisms that have a higher chance of being applicable in practice. This makes likelihood-based methods especially attractive because they can continue to give unbiased answers even if the data are MAR. To reflect this fact, data which are MAR (or the more stringent requirements of MCAR or CD-MCAR) are sometimes called "ignorable". Notably, although we postulate the MAR condition in terms of (11.3), if we are using likelihood-based methods, we need not specify a explicit statistical model for it, no matter how complicated the dependence might be. Instead, we can focus on developing a model for the complete data. This avoids being distracted by modeling a missingness mechanism which is likely to be imperfectly understood.

11.3 Simple Approaches to Handling Missing Data

We begin our discussion of methods of addressing missing data with a number of simple (and sometimes simplistic) methods that have been used previously. We return to the context of the HERS and SCUT trials.

11.3.1 Include a Missing Data Category

A simple approach to completing a dataset with missing values in a categorical *predictor* is to define a separate category for the missing values. In Sect. 4.3, we note that women in the HERS cohort responded to a question about how physically active they considered themselves compared to other women of their age. The five-level response ranged from "much less active" to "much more active", and was coded in order from 1 to 5. A solution to missing data for this predictor is to define a category designated as "missing". Physical activity is then analyzed as a categorical variable with six categories with all observations in the sample having a defined value for this variable. This is appealing because it avoids imputing values to incomplete observations but allows all observations to be used in the analysis.

Unfortunately, this can create biased estimates for other regression coefficients in the model, even when the data are MCAR. The reason for this is that, for the subset coded as missing, we are not adjusting for the value of physical activity, whereas for the rest of the data we are. So regression coefficients (for predictors other than physical activity) that are estimated from the model using the six category version of physical activity are a blend of the coefficient before and after adjustment for physical activity. Bias is introduced when the unadjusted and adjusted coefficients differ and there is a sizeable percentage of observations in the missing data category. On the other hand, if the adjusted and unadjusted coefficients are similar and the percentage of observations in the missing data category is small, little bias will be introduced.

11.3.2 Last Observation or Baseline Carried Forward

In SCUT, vision tends to improve rapidly in the first month as the infection is treated and has usually stabilized by 3 months. As patients feel better, they are less likely to return to the clinic for follow-up appointments and nearly 30% of 12 month visual measurements are missing due to loss to follow-up.

One approach to handling a missing 12-month outcome value in the SCUT trial is to use (or "carry forward") a patient's 3 month VA measure. If the 3-month value is not available the 3-week (or, if that is missing, the baseline value) value would be used. This approach is called *last observation carried forward* (LOCF) because missing values are filled in with the last available value from the same person. This

approach can be used with either outcomes or predictors. The LOCF approach has the appeal of using the most proximate available VA measure to complete the data. It has been argued that this is a conservative method because it assumes no change in measured values for the missing data.

The method has substantial disadvantages. In SCUT, for instance, visual acuity improves substantially from 3 weeks to 3 months. Hence, LOCF would be implausible for such data and almost certainly underestimate VA if values are carried forward, potentially leading to biased estimates. Second, a single value is substituted for the missing value. As with single imputation, if a standard analysis is then applied to the completed data set, this uncertain, filled-in value is treated as if it were an actual measurement and leads to falsely precise analyses. This is a concern whether or not carrying forward values is approximately correct, on average.

Consider a study of people initiating an experimental treatment to reduce hypertension with repeated measures of their blood pressure, subject to missing values. If the missing values are due to study dropout and the participants must discontinue the experimental treatment, then we might reasonably expect that the blood pressure values would return to pretreatment levels. This would be captured by using the baseline value (rather than the last value prior to dropout) to fill in missing values. This approach is termed baseline value carried forward (BCF) and it is very similar in spirit and execution to LOCF except that a baseline value is used to replace the missing value. While imputing using the baseline value might be reasonable for the above example, the immediate return to baseline assumption may not be plausible in other contexts. BCF, like LOCF, under-accounts for the variation due to the single imputed value.

11.3.2.1 Other Single Imputation Approaches

Other approaches use information from the remainder of the data set to infer a single value. Suppose values of a variable like income are missing in a sample. A typical value, such as the mean or median of observed values, could be used. While this can generate a reasonable value for a continuous value, like income, mean values would produce an implausible value for a categorical value, like race. For categorical variables, the method could be adapted to impute the race as the most common answer (e.g., white) if the variable is categorical. The main advantage of all of these "single imputation" approaches is their simplicity in generating the imputation (substituting means, modes, or previously measured values). However, this simplicity may reflect a lack of critical thinking about the relationship of missing data to observed data. In the SCUT trial for example, a better imputation for a missing 3 month VA measure might be to use the 3-week value augmented by the expected change in VA from 3 weeks to 3 months.

With a variable such as income, it is highly possible that the value to be measured contributes to the chance that it will not be observed, which might lead to data that are MNAR. A better approach to imputation might use values of other covariates such as zip code, age, and/or education to predict the missing values of income. If those covariates were able to account for the dependence of missingness on income,

then the data would be MAR. Thus, superior imputations will need to be informed by a model for the data and for the mechanism which underlies the missing values. Methods such as LOCF or BCF skip this crucial step of model development.

Furthermore, any single imputation approach that applies standard analysis methods to the completed data set can seriously underestimate the variation in the data set, giving standard errors that are too small and CIs which are too narrow. These deficiencies can be corrected by applying the method of multiple imputation which we discuss in Sect. 11.5.

11.4 Methods for Handling Missing Data

We now return to more general approaches for handling missing data. The recommended methods depend on both the pattern of missing data (drop out from the study, missing predictors only, etc.) and the missing data mechanism. A key distinction is whether there is missing data for the predictor variables with at least some of those instances having observed values of the outcome. In such a case, we recommend using multiple imputation, described in more detail in Sect. 11.5.

For situations in which the predictors are mostly complete and the issue is data missing in the outcome variable, we divide our presentation and recommendations by the mechanism of the missing data: missing completely at random (MCAR—Sect. 11.7), covariate- dependent missing completely at random (CD-MCAR—Sect. 11.8) or missing at random (MAR—for hierarchical analyses only and in Sect. 11.9). When the data are MCAR or CD-MCAR, relatively simple approaches may suffice. For data that are MAR, several approaches are possible.

11.5 Missing Data in the Predictors and Multiple Imputation

The first distinction in recommended analysis strategies is whether there is missing data in the predictors (even if there is also missing data in the outcomes) and the missing data can be assumed to be MAR. With missing predictor data, we recommend the approach of multiple imputation, which we introduced briefly in Sect. 11.1.2. The basic idea is not only to fill in a reasonable value for the missing data but also to incorporate some random error. While it may seem counterproductive to add in random error, it is a convenient device for properly reflecting the degree of uncertainty due to the missing data. By doing it a number of times (hence the adjective multiple in multiple imputation), we can get valid estimates of standard errors (and hence CIs and p-values), and by averaging the results, not have them unduly affected by the random error. It turns out, perhaps surprisingly, that the process does not need to be repeated very many times. A typical number is five or ten.

Table 11.8 Regression model for imputing glucose

```
. regress glucose bmi csmker white sbp diabetes

      Source |       SS       df       MS              Number of obs =    1355
-------------+------------------------------           F(  5,  1349) =  213.11
       Model |  1019590.52     5  203918.103            Prob > F      =  0.0000
    Residual |  1290806.09  1349  956.861448            R-squared     =  0.4413
-------------+------------------------------           Adj R-squared =  0.4392
       Total |  2310396.61  1354  1706.34905            Root MSE      =  30.933

------------------------------------------------------------------------------
     glucose |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bmi |   .4921757   .1568229     3.14   0.002     .1845325    .7998189
      csmker |   1.183684   2.603247     0.45   0.649    -3.923168    6.290536
       white |   9.180278   2.863755     3.21   0.001     3.562382    14.79817
         sbp |   .0342977   .0447849     0.77   0.444    -.0535579    .1221532
    diabetes |   60.45312   1.977712    30.57   0.000     56.57339    64.33285
       _cons |   69.66885   8.108395     8.59   0.000     53.76242    85.57528
------------------------------------------------------------------------------
```

The steps in multiple imputation are essentially as follows:

(1) Specify a probabilistic model for how to fill in the missing data.
(2) Using the model, fill in (impute) the missing data with random error.
(3) Repeat the imputation a small number of times (e.g., five) so you end up with multiple versions of the data set, each with somewhat different values of the imputed variable(s).
(4) For each of the imputed data sets, calculate the quantities of interest (e.g., the regression coefficients).
(5) Average the quantity of interest across the imputed data sets.
(6) Calculate a standard error based on the average of the model-based variation plus the variation in the calculated quantities of interest across the imputed data sets.

The first step, of specifying the imputation model, is the most difficult and involves building a regression model for the variable with missing data; the subject of this entire book! The remaining steps are relatively automatic and are handled by statistical software.

For the HERS example of Sect. 11.1.2, the variable we imputed was glucose. Our probabilistic model was a linear regression model for glucose with predictors of SBP, BMI, being white, current smoking status, and development of diabetes. The standard assumption of a linear regression model is that the error terms are normally distributed. Table 11.8 gives the output from fitting that regression equation.

Given values of BMI, current smoking status, being white, SBP, and development of diabetes, we can use the regression equation to generate a predicted glucose value for those with missing values. But in multiple imputation we do more. The regression output from the table also gives the value of the root mean square error, 30.933, which quantifies the degree of uncertainty (i.e., the error term) in the regression equation. Under the assumptions of the linear regression model, those

Table 11.9 Stata code for imputing glucose

```
mi set wide
mi register imputed glucose
mi impute reg glucose bmi csmker white sbp diabetes, add(5) rseed(271828) ///
>     force
mi estimate: regress sbp glucose white bmi
```

errors are normally distributed, with means zero and standard deviation 30.933. So to impute the missing values of glucose, we calculate the predicted value of glucose and then add a normally distributed error term with standard deviation 30.933.

As an example, one of the HERS participants who had a missing glucose measurement had a BMI of 24.68, was not a current smoker, was white, had a SBP of 130, and was not diabetic. Using the coefficients from Table 11.8, her predicted glucose value is 95.45. To impute a value for her, we would add a normally distributed error term with mean zero and standard deviation 30.933. Using the `rnormal(0,30.9333)` command twice in Stata to generate random normal variables with the correct mean and standard deviation gave the values 42.98 and −13.34. So her imputed glucose value for the first imputed data set would be $95.45 + 42.98 = 138.43$ and would be $95.45 - 13.34 = 82.11$ for the second. This process is repeated for each of the missing glucose values and each imputed data set.

Next, for each imputed dataset, we perform the regression of SBP on glucose, BMI, and white. Suppose our interest lies in understanding the relationship between SBP and BMI. We would have five estimates of the regression coefficient, each slightly different due to the different imputed values. Averaging those five estimates gives us our multiple imputation estimate and the standard error is calculated both from the model-based standard errors and the variation in the coefficients from imputed data set to imputed data set, which measures the amount of uncertainty due to imputing the values of glucose.

Across the five imputations, the values of the coefficient for BMI were 0.145, 0.149, 0.138, 0.150, and 0.152 with an average of 0.147. The model-based standard errors were 0.083, 0.083, 0.084, 0.083, and 0.084 with an average of 0.083. So the estimated coefficient for BMI from the multiple imputation is 0.147 and the standard error is slightly higher than the average of the model-based standard errors (due to the imputation to imputation variability) and is equal to 0.084.

While no one step in the multiple imputation process is difficult, conducting the multiple imputations and analyses and assembling the results is tedious and could be error-prone if done manually. So programs like Stata automate the process. The results reported in Table 11.3 were generated with the Stata code given in Table 11.9.

11.5.1 Remarks About Using Multiple Imputation

In the HERS example, we used the outcome of the original analysis (SBP) to impute glucose. And then we turned around and used the imputed glucose value in the

regression of SBP on glucose and the other predictors. This may seem like cheating but is actually needed to obtain unbiased estimates (Little 1992). In fact, multiple imputation does not distinguish the roles of outcome and predictor, but instead regards all the variables on an equal footing. So, whenever variables are associated with one another, it becomes important to utilize that information in the imputation model. And it is usually the case that we include predictors (e.g., glucose) in our modeling precisely because we expect them to be associated with the outcome. So if those predictors are missing, it is important to use the original outcome variable in the imputation modeling.

Multiple imputation has a long history of use in sample surveys. Many surveys (like NHANES, the National Health and Nutrition Examination Survey) are extremely comprehensive and are released for public use. In this case, it is difficult to predict which analyses will be undertaken, which variables will be used in analyses, and which will be treated as outcome variables. Because multiple imputation does not distinguish outcomes from predictors, users need not worry about the distinction with regards how the data were imputed. Contrast this with a method like creating a missing data category, which only works for categorical predictors. Furthermore, the imputation algorithms may be quite complicated and difficult to implement, but the analysis of the data is straightforward using routines like those available in Stata. So once a multiple imputation is produced, it can be used in a variety of situations.

Because of the potential for using a multiply imputed data set for many purposes, when imputing missing data it is important to err on the side of flexibility rather than parsimony. If the model for imputation is overly simplistic, those assumptions will be built into the portion of the data that has been imputed. For example in HERS, if the relationship between glucose and BMI were non-linear, but the imputation model assumed it to be linear, then predictions might be biased. Or if we assumed there was no interaction between BMI and race in imputing glucose and if a later analysis searched for interactions, interaction effects would be attenuated.

11.5.2 Approaches to Multiple Imputation

In practice, the pattern of missing data across variables can be quite different. In the HERS example, BMI and current smoking status (yes/no) had missing data in addition to glucose, whereas race (white versus other) had very little missing data. With multiple missing variables, a number of complications arise. First, how should we handle variables with distributions other than normal such as current smoking status, which is binary? Second, if we want to impute glucose using SBP, BMI, race, and current smoking status, what do we do about the missing data for BMI and current smoking status? Third, once the data are filled in, how should we update the parameter estimates? There are three main approaches for dealing with multiple imputation across a number of variables: iterative chained imputation, multivariate normal (MVN) imputation, and Monte Carlo Markov chain. We describe in more detail the first two.

11.5.2.1 Iterative Chained Equations Imputation

Using iterative chained equations (ICEs) imputation, we build regression models for each of the variables with missing data, in turn treating them as outcome variables and using the rest of the variables as possible predictors. For the HERS example, in addition to the model we have already built for glucose, we would need models for BMI and current smoking status. Because current smoking status is a binary variable, a logical imputation model would be to use a logistic regression with predictors of SBP, BMI, race, and glucose. From the logistic regression model, we would get a predicted probability of being a current smoker. We would then generate a random binary outcome with the predicted probability (which could be done using the `rbinomial` command in Stata). Once the value of current smoking was imputed, this could be used as a predictor in a regression model to impute BMI. These regression equations are used to fill in each of the variables in turn. The whole process is repeated a number of times to reach a "steady state" so that the results do not depend on the order in which the variables are imputed.

An important advantage of this approach is the ability to tailor the model for imputing each variable, both with respect to its distribution (e.g., normal, binary, or multiple categories) as well as the inclusion of predictors, possibly with non-linear terms and/or interactions. Currently, Stata allows regression models of the following types via its `mi impute chained` command: linear regression (regular, truncated, and interval), logistic (binary, ordinal, and multinomial), Poisson, and negative binomial. This is also its most important disadvantage: a regression model has to be constructed for each of the variables for which there is a significant percentage of missing data. With, say, 20 variables with missing data, the regression modeling effort increases 20-fold, even though this may not be the scientific focus of the analysis. These regression models need to be built with care so as not to introduce out of range or implausible imputed values.

11.5.2.2 Multivariate Normal Imputation

A simpler to use method available in many statistical software packages is to impute the missing data assuming all the variables follow a joint, normal distribution. While this is invariably an incorrect assumption when there are a number of variables with missing data, it has often been found to perform well in practice. This is because, even though the distributional assumptions may be suspect, imputation assuming a MVN distribution still retains the proper average values and correlations among the variables. When the later analysis depends only on such quantities (as when the ultimate analysis is a linear regression) this method may suffice.

For example, in Sect. 11.5.3 we wish to impute the binary predictor variable race, which is coded as 1 for white and 0 otherwise. Recall that when a variable is coded as 0 and 1, its mean is equal to proportion of observations falling in the category coded as 1. When imputing such a variable, MVN imputation will generate a continuous variable in its place, but one which will have the proper mean (in the sense that the

mean will properly reflect the proportion falling in category 1). Of course, care must be taken when using such a variable in an analysis: since it is no longer categorical it cannot be treated as such in a prediction equation. Software packages such as SAS allow the user to round off to 0 or 1 to recover this aspect of the data.

Although MVN imputation often gives sensible answers, in some cases it may be important to retain more detailed aspects of the distribution (e.g., the proportion exceeding a threshold), and MVN imputation may lead to suspect conclusions. Another situation in which the multivariate normal assumption is not satisfactory is when one or more variable is a nominal categorical variable, e.g., marital status (single and never married, married, divorced).

If most of the variables to be imputed are approximately normally distributed and there are no nominal categorical variables, then it is probably safe to use MVN imputation, which is often easier to implement in practice. However, if there are nominal categorical variables, or the predictors are highly nonnormally distributed, then iterative chained imputation is the recommended approach.

11.5.3 Multiple Imputation for HERS

We demonstrate the use of ICEs imputation and MVN imputation using the HERS dataset and two regression analyses: regression of SBP on glucose, BMI, and race (white or not), which has missing data on two continuous predictors (glucose and BMI) and the regression of SBP on glucose, current smoking status (yes/no), and race, which has missing data on a continuous predictor (glucose) and a binary predictor (smoking status).

Using the ICE methodology, we built linear regression models for glucose and BMI and a logistic regression model for current smoking status. We considered two approaches to modeling: parsimonious and flexible. In the parsimonious approach, we included the other variables in the imputation model as is. So, for example, the parsimonious imputation model for BMI was a linear regression with predictors of SBP, glucose, race, and current smoking status. In the flexible approach, we included all two way interactions and quadratic versions of numerical predictors. So, for example, the flexible imputation model for current smoking status was a logistic regression with predictors of glucose, BMI, and SBP, the squared versions of each of those, race and all the two way interactions such as race by BMI, race by SBP, BMI times SBP, etc.

We compared this to the MVN approach, which assumes that SBP, glucose, BMI, and current smoking status are MVN and imputes the values under that assumption. Table 11.10 lists the sample sizes, regression coefficients, and p-values for a complete case analysis and the three approaches to multiple imputation. Similarly, Table 11.11 lists the values for a regression of SBP on glucose, race, and current smoking status. We might expect the MVN approach to do more poorly for this model since current smoking status is a binary variable.

Table 11.10 HERS model fit comparisons with different multiple imputation strategies: regression of SBP on glucose, race, and BMI

		Parameter estimates			*p*-values		
MI method	*N*	Glucose	Race	BMI	Glucose	Race	BMI
Complete case	1385	0.030	−1.54	0.06	0.02	0.36	0.49
ICE parsimony	1871	0.029	−2.52	0.13	0.02	0.09	0.11
ICE flexible	1871	0.028	−2.53	0.14	0.02	0.09	0.09
MVN	1871	0.030	−2.44	0.14	0.02	0.10	0.10

Table 11.11 HERS model fit comparisons with different multiple imputation strategies: regression of SBP on glucose, race, and current smoking status

		Parameter estimates			*p*-values		
MI method	*N*	Glucose	Race	Smoke	Glucose	Race	Smoke
Complete case	1370	0.028	−2.04	−1.55	0.02	0.23	0.32
ICE parsimony	1871	0.032	−2.75	−0.96	0.007	0.06	0.49
ICE flexible	1871	0.032	−2.77	−0.94	0.006	0.06	0.49
MVN	1871	0.033	−2.68	−1.01	0.005	0.07	0.46

The imputation analyses differ from the complete case analyses in several important aspects:

- The imputation methods are based on imputed versions of the complete data set with 1,871 observations.
- For a number of the coefficients, the imputations give materially different estimates of the coefficients compared to the complete case analysis, e.g., the coefficient for race.
- The imputations, which use all the observed data, often have smaller *p*-values than the complete case analysis.

Turning to comparisons among the various imputation methods, we observe that

- The flexible and parsimonious approaches to ICE gave virtually the same answers.
- The MVN approach gave somewhat different answers than the two ICE approaches, but all three imputation approaches gave answers similar to one another and somewhat different than the complete case analysis.
- The MVN approach seemed to do a creditable job even when imputing the binary variable, race.

The example serves to illustrate both the advantages and disadvantages of multiple imputation. It uses all the observed data while properly reflecting the fact some of the data are missing. It may have reduced the bias in some of the regression coefficients. It properly reflects the fact some of the data are missing but allows for reduced standard errors and generally smaller *p*-values. But it came at the cost of either having to construct a model for each of the original predictor variables (for ICE) or hypothesize a MVN model for all the predictor variables that had substantial missing data and led to a somewhat more complicated overall analysis.

11.6 Deciding Which Missing Data Mechanism May Be Applicable

The key to using multiple imputation is to build regression models to fill in predictors or outcomes that have missing data. When the predictors have missing data, the outcome variables will usually be part of the imputation models. In the next few sections, we consider situations where the main missing data are *outcome* data. Our recommended strategies depend on which missing data mechanism is to be assumed so we give some guidance here as to how to choose.

As noted above, e.g., (11.2), the different missing data mechanisms are distinguished by dependence of the probability of the data being missing on different quantities. In CD-MCAR dependence is on covariates and, in MAR, dependence is on the outcome and possibly also on covariates. Distinguishing between these cases can be done in a descriptive manner or using a more formal statistical model.

For example, in the SCUT trial and considering missing outcome (visual acuity) data at the 3-month visit, we would calculate descriptive statistics for those with and without missing data. If the average ulcer size, the proportion with the ulcer in the center of the eye, or the proportion of gram positive infections (all measured at baseline) differed between those with a missing VA measurement at 3 months and those with it present, then we would know the data could not be considered MCAR, but instead would be at least CD-MCAR. We could formally test the association by conducting a t-test for ulcer size or χ^2 tests for whether the ulcer is in the center of the eye or type of infection across the missingness groups.

Alternatively, we could define an indicator variable R_i, equal to 1 if the 3-month measurement was present and zero otherwise, and conduct a logistic regression to assess the association of missingness with the covariates. If we found that any of the covariates is associated with missingness, it would establish that the data could not be MCAR.

By further considering previously measured outcomes (e.g., the value of VA at 3 weeks), we can check to see if the CD-MCAR assumption is inadequate. If the VA at 3 weeks differed between the groups with 3 month VA data present and absent that would suggest the missing data mechanism to be at least MAR. More rigorously, if VA at 3 weeks was predictive of missing VA at 3 months in a logistic regression model that also contained the covariates that were related to missingness then we would know that the assumption of CD-MCAR was inadequate.

With substantial amounts of missing data, it is invariably good practice to conduct descriptive analyses to understand to what extent the missing data are associated with measured variables. As noted above this can help rule out simpler mechanism such as MCAR (which rarely holds in practice) and CD-MCAR. As noted earlier, because MNAR depends on *unobserved* quantities, we cannot verify or rule out a MNAR process from the observed data alone.

11.7 Missing Outcomes, Missing Completely at Random

We now consider datasets for which there is missing data in the outcomes but where any missing data in the predictor variables is negligible or occurs along with missing data in the outcome (as when a participant drops out of a study). The easiest case to deal with is when the data are MCAR, i.e., the missing data are totally unrelated to either the other outcomes or the predictors. In this case, ignoring the missing data does not cause bias and simply leaving the missing data out of the analysis properly reflects the amount of information available. In this case, complete case analysis using any of the usual statistical analysis strategies (e.g., linear regression or logistic regression) is the recommended strategy. It will automatically be adopted by any of the usual statistical packages, including Stata, if you conduct the usual analysis in the presence of missing data.

We again return to the HERS dataset and we fit a model to predict systolic blood pressure (SBP) from BMI, race (white or not), whether the participant was on medication to control their blood pressure (yes/no) and the interaction of BMI and blood pressure medication. From that model, the coefficient of BMI in the on-medication group was 0.24, with a standard error of 0.06. So with each increase in BMI of one unit, there is an associated increase in SBP of about 0.24. However, in the off-medication group, the coefficient is 0.52 with a standard error of 0.11. This is not surprising as we would expect those on medication to have their blood pressure better controlled and less associated with BMI.

We again artificially create missing data to illustrate the consequences. Using a random mechanism, we dropped 75% of the data and refit the above model, so the missing data mechanism is MCAR. The on-medication BMI coefficient was 0.19 with a standard error of 0.10 and the off-medication coefficient was 0.56 with a standard error of 0.21. So, even though we have dropped 75% of the data, the two coefficients are similar to those obtained from the full dataset, as expected. Using GEEs gave virtually the same results, with coefficients of 0.18 and 0.56, respectively.

11.8 Missing Outcomes, Covariate-Dependent Missing Completely at Random

The next level of missing data occurs with data where missingness may depend on a covariate that is in the analysis model as a predictor, but does not depend on other variables (either other outcomes or variables not in the model), that is, covariate-dependent missing completely at random (CD-MCAR). Under CD-MCAR, a complete case analysis yields unbiased estimates of regression coefficients and predictions for given values of the covariates using any of the regression methods we have described. However, quantities that require averaging over members of the sample may not be correct.

Using the HERS data, as in the previous section, we again randomly dropped 75% of the data, but this time all the dropped data was from the on-medication subgroup, which makes up about 80% of the full dataset. This missing data mechanism would be CD-MCAR because it depends on whether the participant is on hypertension medication or not, but not on other variables. We fit the same model as described in the previous section and obtained an on-medication coefficient for BMI of 0.23 with a standard error of 0.16 and an off-medication coefficient of 0.59 with a standard error of 0.12, with the coefficients again quite similar to the full dataset.

But suppose we were interested in the average increase in SBP associated with a one unit increase in BMI. Since the no-medication participants make up about 80% of the cohort, the average increase is a weighted average of the two coefficients: $0.30 = 0.8(0.24) + 0.2(0.52)$. Being more careful with the calculations, the exact value is actually 0.29. But in the CD-MCAR scenario, the proportion of on-medication participants is only about 24%. And so the average increase will be misestimated as 0.51, because the off-medication participants are weighted too heavily.

The correctly blended average can be calculated using Stata's `margins` command as shown in Table 11.12. In that table, `bmi_ctr` is the centered value of BMI (i.e., it has mean zero) and `sbp_cdmcar` is SBP with values missing due to the CD-MCAR mechanism. The margins command estimates the value of SBP at the mean value of BMI and at one unit above the mean. Using just the estimation sample gives an associated increase in SBP of about $0.51 (= 135.1076 - 134.6008)$. However, using the `noesample` option generates an estimate for the entire sample, recovering the proper weighting of the on- and off-medication subgroups, and gives an estimate of about $0.30 (= 133.2468 - 132.9490)$, quite close to the full data estimate.

As in the MAR scenario, for CD-MCAR the particular analysis method makes little difference. Using GEEs gave virtually the same answers as the mixed-model approaches reported above.

11.9 Missing Outcomes for Longitudinal Studies, Missing at Random

Longitudinal studies with a planned observation schedule invariably have at least some missing data. Although attempts are usually made to have participants return for every scheduled visit (e.g., yearly), some drop out of the study, either voluntarily or involuntarily (e.g., death), or miss visits. A consequence is that all data that would have been collected at that visit (either outcomes or predictors) will be missing. So use of analysis strategies to minimize bias due to missing data are essential. For example, the Osteoarthritis Initiative, a well-conducted cohort study, enrolled 4,796 individuals, attempting to collect data yearly. After 1 year, 94% were still

Table 11.12 Using the margins command with CD-MCAR missing data

```
. xtmixed sbp_cdmcar c.bmi_ctr white htnmeds htnmeds#c.bmi_ctr || pptid:

Mixed-effects REML regression                   Number of obs      =      2291
Group variable: pptid                           Number of groups   =       972

                                                Obs per group: min =         1
                                                               avg =       2.4
                                                               max =         6

                                                Wald chi2(4)       =     30.08
Log restricted-likelihood = -9527.0924          Prob > chi2        =    0.0000

------------------------------------------------------------------------------
  sbp_cdmcar |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     bmi_ctr |   .5946715   .1249834     4.76   0.000     .3497085    .8396344
       white |   .6615583   2.262646     0.29   0.770    -3.773146    5.096263
     htnmeds |  -2.858138   .9904139    -2.89   0.004    -4.799314   -.9169625
             |
     htnmeds#|
   c.bmi_ctr |
           1 |  -.3666237   .1957226    -1.87   0.061    -.750233    .0169856
             |
       _cons |   134.6577   2.258123    59.63   0.000     130.2318    139.0835
------------------------------------------------------------------------------
------------------------------------------------------------------------------
  Random-effects parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
pptid: Identity              |
                   sd(_cons) |   14.58279   .4669968     13.69562    15.52742
-----------------------------+------------------------------------------------
                sd(Residual) |   11.41371   .2233941     10.98415    11.86006
------------------------------------------------------------------------------
LR test versus linear regression: chibar2(01)=736.07 Prob >= chibar2= 0.0000

. margins, at(bmi_ctr=0) at(bmi_ctr=1)

Predictive margins                                  Number of obs   =       2291

Expression   : Linear prediction, fixed portion, predict()

1._at        : bmi_ctr         =           0
2._at        : bmi_ctr         =           1

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         _at |
          1  |   134.6008   .5746835   234.22   0.000     133.4745    135.7272
          2  |   135.1076   .5946754   227.20   0.000     133.9421    136.2732
------------------------------------------------------------------------------

. margins, at(bmi_ctr=0) at(bmi_ctr=1) noesample

Predictive margins                                  Number of obs   =       9157

Expression   : Linear prediction, fixed portion, predict()
1._at        : bmi_ctr         =           0
2._at        : bmi_ctr         =           1

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         _at |
          1  |   132.9490   .6731565   197.50   0.000     131.6297    134.2684
          2  |   133.2468   .6880263   193.67   0.000     131.8983    134.5953
------------------------------------------------------------------------------
```

being followed, with 6% dead or lost to follow-up and, after 2 years, 90% were still being followed. If the data are MCAR or CD-MCAR then the analysis strategies suggested above will work for longitudinal data. But in a longitudinal study, it is quite possible that missingness is related to previously measured *outcomes*, making the data MAR. For example, in the OAI, a patient with an MRI (magnetic resonance image) showing advanced osteoarthritis at one visit may be less likely to come in for the next visit, since it would entail lengthy data collection and another MRI.

The situation of MAR represents a reasonable one for a wide variety of missing data problems. This is a middle ground between MCAR and MNAR for which the choice of analysis strategy can make a difference. Three general approaches have been suggested for dealing with MAR data in longitudinal studies: use maximum-likelihood based methods, use inverse weighting methods, or use multiple imputation.

11.9.1 ML and MAR

In Sect. 11.1.2, we contrasted the use of generalized estimating equations and linear mixed models in a particular example. Under a MAR situation we showed that the generalized estimating equations approach gave biased results whereas the linear mixed-model analysis did not. This result generalizes to the wider class of models fit by maximum likelihood (see Sect. 5.6 for the definition of maximum likelihood). Namely that a simple strategy for dealing with MAR data is to use approaches wherein the models are fit by the method of maximum likelihood, such as the random effects models described in Sect. 7.5. As long as the model is correct in both its fixed and random components, this fitting technique leads to methods that are not biased. A more detailed explanation as to why maximum likelihood avoids bias with MAR data is given below in Sect. 11.10.

Commonly used approaches which use maximum likelihood include linear mixed-model analyses (Stata `xtmixed` or `xtreg` [with the mle option]; SAS Proc MIXED; SPSS linear mixed model routines) and random effects logistic or Poisson regression models (Stata `xtlogit`, `xtmelogit`, `xtpoisson`, `xtmepoisson`, and others; SAS Proc NLMIXED). The primary method for longitudinal data which does *not* use maximum likelihood is GEEs (see Sect. 7.4), which is therefore subject to bias under MAR data.

When maximum-likelihood methods are a natural analysis strategy, we generally recommend them since they obviate the need to model the missingness mechanism. And for studies that are not on a regularly scheduled visit time, it is not clear what data should be imputed. When following a maximum-likelihood analysis strategy and for cases where there is a significant portion of missing outcome data, care should be taken on model diagnostics (e.g., checking for interactions and correct specification of the variance–covariance structure). This is because the ability of maximum likelihood to adjust for missingness depends on specifying a correct or nearly correct model.

11.9.2 Multiple Imputation

In a longitudinal study with MAR missing data, maximum-likelihood methods automatically correct for missing data without having to specify a model for the missingness. But multiple imputation is also a viable method, building a model to impute the missing outcomes based on the covariates and previously measured outcomes.

There are, however, circumstances in which multiple imputation is to be recommended over maximum likelihood. If the preferred analysis strategy is GEEs (or another, non-likelihood-based method) then multiple imputation is an attractive strategy to deal with missing data. This is because it can reduce the bias associated with the use of non-likelihood-based methods under MAR missing data.

So far we have assumed that, when missingness is dependent on the predictors, these are predictors that can be included in the analysis model. This will not always be the case, for example, if drop out in a longitudinal study depends on a mediator. In SCUT, for example, suppose that individuals whose ulcers have cleared by three weeks are less likely to return at 3 months since it is not as urgent for them to visit the clinic. To properly account for missingness in an ML analysis, we would need to include presence of an ulcer at 3 months in the model. But this will also adjust away some of the treatment effect, which we do not wish to do. This would be an example of a situation in which a variable (presence of an ulcer at 3 weeks) is needed to make the MAR assumption plausible but is not useful for the analysis model. This is another situation in which multiple imputation is an attractive approach: we can use the mediator in the imputation model, but leave it out of the analysis model.

11.9.3 Inverse Probability Weighting

Another family of methods which use the MAR assumption are those based on *inverse probability weighting*. The basic idea is to use complete observations to represent incomplete observations, just as we did for potential outcomes in Subsect. 9.1.8. For instance, in the SCUT example, suppose that we could make the assumption that the probability of missing visual acuity (VA) at the second (3 month) visit depended only on the distance the patient lives from the clinic and their VA at enrollment.

In that case, for patient i at visit 2

$$\mathrm{P}(R_{i2} = 1|\mathbf{Y}, \mathbf{X}) = \mathrm{P}(R_{i2} = 1|\mathbf{X}) \tag{11.4}$$

or more specifically it is equal to $\mathrm{P}(R_{i2} = 1|x_{1i}, x_{2i})$, where x_{1i} is the distance patient i lives from the clinic and x_{2i} is his or her VA at baseline. This simplification means that we postulate covariate-dependent MCAR for the missing outcomes.

Suppose, for specific values of clinic distance and VA, the probability of observing the visit 2 outcome (11.4) is equal to 1/2. Then only about half of the patients with these values of x_{1i} and x_{2i} will have VA data at the second visit and it would be reasonable to "double-count" their values to represent the missing values. Similarly, if the probability were 1/3, we would observe only about 1/3 of the outcomes and it would be reasonable to "triple-count" the participants for whom we observed the outcome. In general, we would up-weight observed outcomes by one divided by the probability of being measured, hence the name, inverse probability weighting. This is the spirit that underlies inverse weighting methods.

Many statistical packages allow the incorporation of weights, but care must be taken. Often, for example the `_weight` statement in SAS, a weight of 2 would represent 2 actual measured observations with the same value. This is distinct from our situation in which a weight of 2 would mean we are using a single measured value to represent itself and an unmeasured value. The weights that are needed for inverse weighting estimation are sometimes called *probability* or *sampling* weights and are implemented for many of the commands in Stata using the `pweight` option. Using the more standard weighting as in SAS gives the correct estimate, but incorrectly implies there is more actual measured data and hence will give standard errors and p-values that are too small and CIs that are too narrow. For some routines, this can be corrected by using robust standard errors.

11.9.3.1 Comments on Inverse Probability Weighting

Inverse probability estimates require that we specify or estimate the *probability of observing* an outcome at 3 months. We might do this by developing a regression model, like logistic regression, for the probability of a measured value in terms of observed data (much like the propensity score method discussed in Sect. 9.4.3). Other methods discussed in this chapter based on the MAR assumption rely on postulating a correct model for the outcomes. For example, in the SCUT trial, we might postulate a linear mixed-effects model for the VA measures. These approaches use MI- or ML- based estimation and are able to avoid specifying a model for the missing data mechanism but depend on the correctness of the outcome model to adjust for missing data. In contrast, inverse weighting adjusts for missing data through the weighting scheme and does not depend as strongly on the correctness of the outcome model. Inverse weighting has been suggested in situations using analyses such as generalized estimating equations.

We have several concerns about the use of inverse probability methods and cannot recommend them in general. In many situations, the probability of a measured value can be small, leading to large inverse weights. The large weight given to a few observations means that these values significantly influence the results, leading to unstable estimates and loss of efficiency.

If IPW is used, weights should be carefully monitored. And even if weights are not large, inverse weighting can be notably less efficient than an analysis based on a carefully chosen model for the complete data. In many, if not most, situations,

a plausible model for missingness is poorly understood. It is, therefore, often more natural to build a model for the complete data and apply methods based on maximum likelihood.

11.10 Technical Details About Maximum Likelihood and Data Which are Missing at Random

We have stated earlier that methods of fitting models using maximum likelihood give valid estimates even when the data are MAR. In this section, we give some explanation as to why that is so and contrast maximum-likelihood with multiple imputation. The comparison rests on a particular way in which maximum likelihood estimates can be calculated, called the Expectation–Maximization Algorithm, or *EM algorithm* for short, an approach that has often been of utility in missing data problems. The EM algorithm operates by starting with a guess as to the values of the estimates and improves them using an expectation calculation and then a maximization calculation. The expectation and maximization calculations are repeated until the estimates stabilize. This gives the same answer as directly finding the maximum of the likelihood of the observed data.

11.10.1 An Example of the EM Algorithm

Suppose we wanted to estimate the average number of emergency room visits per person in a year for the population of people served by a particular emergency room. But suppose we only had emergency-room (ER) data and did not know the size of the population who might use that emergency room. If we had data for everyone, we would just calculate the average value. But we have a problem since we do not have a record of those who did not visit the emergency room that year, that is, those whose outcome is equal to 0. And clearly calculating the average among those who *were* seen in the emergency room will drastically overestimate the average.

If we had a preliminary estimate of the average and a probabilistic model for how often people visit the emergency room, we could predict how many we would expect to have a zero value. One such model is the Poisson distribution, for which the probability of an individual visiting the ER exactly x times during the year, $P(x)$, is given by the formula

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \tag{11.5}$$

where λ is the average number of visits per year and $x!$ is "x-factorial", e.g., $6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720$, and, by convention, $0! = 1$. Plugging a 0 in for x

in (11.5), the probability of a person not visiting the ER in a year is $\frac{\lambda^0 e^{-\lambda}}{0!} = e^{-\lambda}$. So, if the average is 0.1 visits per year, the probability of no visits is $e^{-0.1} = 0.905$, and we would predict about 90.5% of the people in the population will have no visits.

Suppose our data consist of 1,232 separate people who visited the ER. Of those people, 1,171 visited 1 time, 57 visited twice, and 4 visited 3 times. So there was a total of $1{,}171+2 \times 57 + 3 \times 4 = 1{,}297$ visits. We are certain that there many people who visited 0 times, but how many? The EM algorithm works by "filling in" the missing data (the number who visited 0 times) making the problem a simple one.

Suppose we start with an initial guess of 0.25 visits per person per year. Then the probability of zero visits would be $e^{-0.25} = 0.779$ and the probability of at least one visit would be $1 - 0.779 = 0.221$. That is, there should be $0.779/0.221$ or 3.52 as many people we did *not* see compared to how many we did see visit the ER. So we would expect that there are $3.52 \times 1{,}232 = 4{,}337$ people with zero visits. This is the expectation step of the EM algorithm.

Next we use our data to find the maximum-likelihood estimate of the average simply by calculating the arithmetic average using the filled in data. The total number of visits was 1,297 and we expect there were $4{,}337+1{,}232 = 5{,}569$ people, for an average of $1{,}297/5{,}569 = 0.233$. This is the maximization step. So we can see that our initial guess was too high and the average rate has tended lower.

With our new estimate of the average, we can calculate an updated probability of not visiting: $e^{-0.233} = 0.792$. And now we expect that there are $0.792/0.208$ or about 3.81 times as many zero visit people as those we actually saw in the ER for a expected number of $3.81 \times 1{,}232 = 4{,}698$. So we can further update our estimate of the average as $1{,}297/(4{,}698 + 1{,}232) = 0.219$. Repeating this process many times, the estimate converges to 0.104. This can easily be calculated using a spreadsheet program such as Excel.

The maximum-likelihood estimate can also be calculated directly. It corresponds to finding the value of λ that maximizes the quantity[1]

$$\log L = 1297 \log \lambda - 1232\lambda - 1232 \log(1 - e^{\lambda}). \tag{11.6}$$

Numerically calculating the maximum of (11.6) also gives the value 0.104. It is also possible to find the maximum-likelihood estimate graphically using the Stata commands given in Table 11.13. The resulting plot is shown in Fig. 11.1.

[1]Recall that the likelihood is the probability of observing the data. The probability of a specific count for a Poisson model, conditional on being 1 or greater is given by $P(x) = \frac{\lambda^x e^{-\lambda}}{x!(1-e^{-\lambda})}$. The product over the entire sample is given by $L = \frac{\lambda^{\Sigma x_i} e^{-n\lambda}}{\Pi x_i!(1-e^{-\lambda})^n}$, where n is the sample size, and x_i is the count for individual i. It is equivalent and easier to maximize the logarithm of L. We can also ignore the factorial term which does not depend on λ, giving $\log L = \Sigma x_i \log \lambda - n\lambda - n \log(1 - e^{\lambda})$.

Table 11.13 Stata commands for plotting the log likelihood

```
clear
set obs 100
gen lambda=_n/10000+.1
gen logL=1297*ln(lambda)-1232*lambda-1232*ln(1-exp(-lambda))
twoway line logL lambda, ytitle("log(likelihood)") xtitle({&lambda})
```

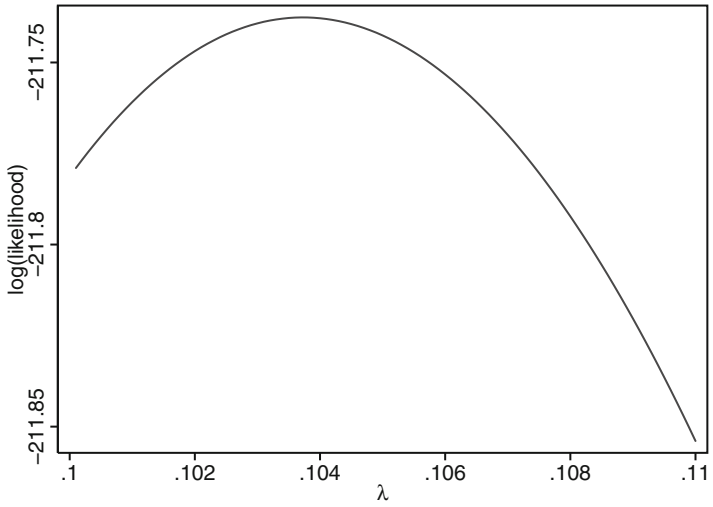

Fig. 11.1 Plot of log-likelihood versus the average rate, λ

11.10.2 The EM Algorithm Imputes the Missing Data

The example above illustrates a typical feature of the EM algorithm: using the observed data and the probability model, the EM algorithm fills in the missing data. The analysis of the data then proceeds using the "complete" dataset. The same algorithm can be applied to longitudinal data with missing outcomes. In that case, maximum likelihood is equivalent to filling in the data not observed due to, e.g., drop out or missed visits, using the longitudinal data mixed model. The parameter estimates are then calculated using the complete dataset. As long as the missing data can be reliably predicted from the observed data (which is the case if the longitudinal data model is correct and the MAR assumption holds), the analysis based on the complete dataset is free of bias due to missing data.

Using maximum-likelihood methods with modern computers does not appear to explicitly handle missing data (it just requires the push of a button on a computer). However, when viewed through the lens of the EM algorithm, it is implicitly filling in the missing data based on the assumed probability model being used to fit the data.

11.10.3 ML Versus MI with Missing Outcomes

Maximum likelihood via the EM algorithm may appear to be virtually the same as multiple imputation. Although there are similarities there are also important differences. Perhaps the primary one is that, under MAR, maximum likelihood implicitly selects the right model for filling in the missing data—no model specification is necessary as it is in multiple imputation.

However, because maximum likelihood implicitly assumes a model for "imputation" it cannot be varied. Multiple imputation gives the analyst more options. For example, nonignorable missing data models can be used to check sensitivity to the assumption of MAR. Or violations of the model assumptions can be checked, e.g., what if the assumption of a Poisson distribution was incorrect in the example above? MI also allows the use of techniques other than ML to obtain parameter estimates after the data are imputed.

However, if (a) the model being used for multiple imputation is the same as the one implicitly used by ML, (b) the imputation was performed so many times that the imputation error was negligible, and (c) once imputed, maximum likelihood was used to find parameter estimates, then MI and ML would give the same answers.

11.11 Methods for Data that are Missing Not at Random

We have mentioned previously that standard analysis methods can be biased when the data are MNAR. And, since it is impossible to figure out if the data are MNAR from the observed data, the main strategies to assess the potential impact of MNAR data are sensitivity analyses. Sensitivity analyses proceed by positing a spectrum of MNAR models with checks as to the seriousness of the violation of the MCAR or MAR assumptions required to qualitatively overturn the results of an analysis. If "small" departures from MCAR or MAR lead to different conclusions then the results are taken as tenuous. If "large" departures are required to change the results, then more confidence can be placed in the conclusions. To be convincing, the posited MNAR models and degree of departure from MCAR or MAR need to be defensible in context, which tends to be highly problem specific and so it is difficult to recommend generally applicable strategies. We describe briefly three approaches to MNAR data: pattern mixture models, multiple imputation, and selection models.

11.11.1 Pattern Mixture Models

Consider a study of cognitive decline in which the participants who dropped out had much higher rates of depression at baseline than those with complete data.

We would be concerned that the data was MAR or MNAR and, especially if the rates of decline were quite different, that we might be obtaining biased estimates. What about more detailed comparisons of those with different degrees of missing data?

Our approach to missing data to this point has been what is called a *selection model* approach. We have thought of the observed data as arising from a two-step process. In the first step, the complete data are generated. In the second step (via a process we have described as MCAR, CD-MCAR, MAR, or MNAR), certain of the data are selected for us to observe; the rest is missing.

A very different approach uses what is called a *pattern mixture model*. In this approach, the data are divided into categories according to the pattern of missing data, akin to dividing subjects into those with complete and incomplete data. For example, consider a cohort study where everyone has a baseline observation and there are four planned follow-up visits. Further, suppose the only missing data are because of dropout from the study. Then there are five possible data patterns with regard to presence or absence of data: complete data, missing only visit 5 (i.e., dropout after visit 4), missing visits 4 and 5 (dropout after visit 3), missing visits 3, 4, and 5, and missing visits 2, 3, 4, and 5.

We can now think about dividing up the data according to the missing data pattern and analyzing the data from each pattern separately. The advantage of this approach is that we do not need to think about the missing data mechanism (e.g., MCAR versus MAR). We immediately run in to a problem, however. Returning to the cognitive decline example, what are we to assume about the rate of decline for the participants for whom we only have baseline data? Because we only have a single time point, this group contains no information about the decline over time. If we wish to proceed, we have to make certain assumptions. For example, if we believe the rates of decline are linear over the course of the five year study, we might assume that the rate is the same as that for the group with data for visits 1 and 2 (for which we *can* estimate a linear decline). Or we might assume it is the same as the subgroup with complete data.

If it is reasonable to make simplifying assumptions then the pattern mixture approach is very attractive. Simply by including a categorical predictor for missing data pattern and allowing interactions of key components with that predictor allows the use of standard software packages to accommodate missing data. In the absence of interaction, the analysis gives estimates of the (assumed) common effect. In the presence of interaction, weighted estimates (weighted by the proportion in each missing data pattern) give an estimate of the overall effect.

Unfortunately, it is often the case that there is little guidance in the data as to what models are appropriate and strong assumptions must be made with little opportunity to check them. Further, there are often a multitude of different missing data patterns (it is rarely as simple as described above) which must be grouped subjectively into a manageable, smaller number of categories, each with reasonable sample sizes. These considerations limit the use of pattern mixture models as robust data analysis methods. However, they can still be useful as sensitivity analyses: by varying

the assumptions needed to fit such models, a variety of MNAR missing data mechanisms can be accommodated. See Little (1993, 1995), and Verbeke and Molenberghs (2000) for more in-depth discussion.

11.11.2 Multiple Imputation Under MNAR

Another possible approach to assess sensitivity of results to MNAR missingness is to use multiple imputation but hypothesize an imputation model that allows dependence between the probability that data are missing and the value that would be observed if $R = 1$. Subak et al. (2009) give an example of a trial to encourage weight loss in women with incontinence problems. Their primary analysis imputed end of study values by assuming that women who dropped out of the study, on average, lost no weight, a MNAR mechanism.

11.11.3 Joint Modeling of Outcomes and the Dropout Process

A third strategy is to directly hypothesize a joint model for the complete data and the missing data process and use the observed data to simultaneously estimate the parameters of both models (e.g., Diggle and Kenward 1994). Not surprisingly, it is difficult to estimate such a model from observed data and they are highly sensitive to the assumed form of the model, something which is not easily checked from the observed data.

11.12 Summary

Missing data are common and many of the simple methods of handling missing data, such as a complete case analysis (the default for most statistical analysis programs), can give misleading results. If it is the predictor variables that are missing in a dataset, we recommend the strategy of multiple imputation. When the main issue is dealing with missing outcomes in a longitudinal study, maximum-likelihood methods are often a good choice. When they are properly specified, they will give valid inference when the data are MAR, whereas generalized estimating equation methods may not. When the analyst needs to exclude important predictors of missingness, in particular mediators, from the outcome model, multiple imputation, and IPW can be useful strategies. Finally, when data are MNAR, pattern mixture models and sensitivity analyses using multiple imputation are recommended.

All techniques for handling missing data require assumptions about how the missing data relate to the observed data. Because the data are missing, these

assumptions cannot be empirically verified. The assumptions are clear in multiple imputation (where we model the missing data), IPW (where we model the probability of missingness), and pattern mixture modeling (where we must make assumptions about covariate effects across missing data patterns). When using maximum-likelihood-based techniques to handle missing-at-random data, the assumptions are inherent and revolve around correct specification of the model, including the variances and correlations in longitudinal data. Because assumptions cannot be verified from the data on hand, it is always a good idea to try a number of techniques of handling missing data to check sensitivity of the conclusions (Hogan et al. 2004).

11.13 Further Notes and References

An attraction of approaching missing data through inverse probability weighting is that it adjusts for missing data through the weighting scheme and does not depend as strongly on the correctness of the outcome model. However, we have noted that it can lead to unstable weights and inefficient analyses. This is an ongoing area of research, with investigations into ways to stabilize the weights, for example, using what is called "robit" regression instead of logistic regression to estimate the probabilities of missingness (Kang and Schafer 2007). Another promising avenue of research is to hedge bets between having to get the outcome or weighting models correct, by using what are known as doubly robust methods (Kang and Schafer 2007). These can correct for missing data when either the model for the inverse weights is correct or the regression model is correct.

The forms of multiple imputation we have illustrated are based on regression models, but there are other alternatives. Scheuren (2005) gives a historical survey of multiple imputation and describes other methods such as "hot deck imputation" (the name comes from a deck of paper "cards" on which data were stored in the early days of the Census Bureau).

Of course, missing data can also occur in situations requiring more complex analyses. For example, there could be missing predictor information in a setting with clustering by facility, physician, and patient. In such a case, just as described in Chap. 7, hierarchical, repeated measures or longitudinal data models must be used to properly impute missing values. Survival analysis is another situation for which imputation of missing predictor information might be required. For survival analysis, the "outcome" consists not only of follow-up time, but also whether censoring has occurred. Both sources of information should be used for imputation, but it is not always clear how to do so. For example, the suggestion to include both the log of the follow-up time and the censoring indicator as predictors in the imputation model can be too simplistic and lead to bias (White and Royston 2009).

11.14 Problems

Problem 11.1. Give an example of a data sampling regime in your research area that is likely to be MAR but not MCAR or CD-MCAR. Briefly explain why.

Problem 11.2. Perform a single imputation for the HERS visit 4 data and verify the results of Table 11.2. Regress glucose on SBP, BMI, ethnicity (white/not white), current smoking status, and diabetes status. Obtain the predicted values for glucose. Create an imputed glucose variable which is equal to the actual glucose value if it is not missing and equal to the predicted value if it is. Using this imputed glucose variable, reproduce the regression of SBP on glucose, white, and BMI given in Table 11.2.

Problem 11.3. How far off are the results when a poor imputation model is used? Singly impute the glucose values (as in Problem 11.2) but using a regression model that contains only current smoking status. How good is this imputation model? Next, compare the estimated effect of glucose on SBP and its statistical significance using this imputation model to the results in Tables 11.1 and 11.2.

Problem 11.4. With the HERS visit 4 data, use the code in Table 11.9 to impute the glucose values. Calculate the SD among the imputed values in glucose to verify that the SD is about 30.9. Hints: the Stata command `egen sd_glu_imp=rowsd(_?_glucose)` will calculate the standard deviation of the glucose values across the imputed datasets. Summarize those for which the original glucose measurement was missing.

Problem 11.5. What kind of imputation model would you use to impute missing physical activity data in the HERS study? Recall that that variable was a response to a question about how physically active the women considered themselves compared to other women of their age. The five-level response ranged from "much less active" to "much more active," and was coded in order from 1 to 5. Briefly explain why.

Problems 11.6–11.9 use the data sets `bpmisslong` and `bpmisswide`. The data are based on measurements of SBP in the HERS study. The data set allows us to compare methods of analysis with complete data and under simulated missing data. In the data sets are missing data indicators (`miss_mar` for `bpmisslong` and `miss_mar1` for `bpmisswide`) which have value 1 to flag SBP values which should be dropped to simulate data which displays MAR missingness. In particular, year 1 values from patients with higher baseline SBP are flagged more frequently and hence will be simulated as missing. You can consult the course website for the data sets and more complete documentation and details on Stata code.

Problem 11.6. Using `bpmisswide`,

(a) Calculate and compare the year 1 SBP (`year1_sbp`) for the complete data and for patients who in the simulated missingness setting would have an available year 1 SBP (i.e., `miss_year` equal to 0).

(b) Calculate and compare the change in SBP (`year1_sbp - base_sbp`). What is the mean change in the full sample? What is the mean change restricted among those with available year 1 values in the simulated missingness setting (`miss_year` equal to 0)?
(c) Based on (a) and (b) above, how has the simulated missing data mechanism affected estimates of mean of year 1 SBP values and change in SBP from baseline to year 1?

Problem 11.7. Using `bpmisslong`, fit a GEE model with SBP as the outcome and visit (`visit`) as the predictor. In Stata, the command would be `xtgee sbp visit, i(pptid) corr(exch)`. Compare a GEE model which uses the full data to one restricted to nonmissing data (`miss_year` equal to 0). What do you conclude about GEE with MAR missingness?

Problem 11.8. Using `bpmisslong`, fit a mixed linear regression model with SBP as the outcome and visit (`visit`) as the predictor. In Stata, the command would be `xtmixed sbp visit || pptid:`. Compare the mixed model which uses the full data to one restricted to nonmissing data (`miss_mar` equals 0). Compare the results with the GEE results in Problem 11.7. How do you explain the difference in results between the GEE and a linear mixed model with MAR missing data?

Problem 11.9. Using `bpmisswide`,

(a) Attempt to mimic the effects of multiple imputation by performing imputation to fill in SBP values flagged as missing in the simulated scenario. You may choose the imputation model but it should include baseline SBP, BMI at baseline and year 1 as well as diabetes. In Stata, it will be simplest to perform multivariate normal-based imputations.
(b) Fit a GEE model (as in Problem 11.7) with multiple imputation. How do the results compare to the results in Problem 11.7? *Note, to fit the GEE model you will need to convert the data from a wide to long format. In Stata, this can be done with the* `mi convert` *command*.
(c) Fit a mixed model (as in Problem 11.8) with multiple imputation. How do the results compare to the results in Problem 11.8?

Problem 11.10. The data set `multivisitsbp` extends the HERS SBP data to a series of up to six visits and borrows the set-up used in Problems 11.6–11.9 to simulate missing data through a missing data indicator `miss_mar`.

(a) To mimic an analysis on complete data, examine a series of models (ignoring the missing data indicators). Fit a GEE model with terms for time (`visit`) and BMI (`bmi`). Then, fit a series of mixed models with fixed effects terms for time and BMI but with varying variance/covariance structures. You might try a random slopes model along with first-, second-, and third-order autoregressive (AR1–AR3). Do you reach similar conclusions about changes in SBP over time (given by the coefficient for `visit`) in these models?

Note: For this data, you can specify the covariance in `xtmixed` with the options `|| pptid: visit, cov(un)` for random slopes and `|| pptid:, residuals(ar 1, t(visit))` for the AR1 model, with AR2 and AR3 defined similarly.

(b) Repeat the model fits in (a) restricted to available data (`miss_mar` equal to 0) under simulated missingness. Do you reach similar conclusions about changes in SBP over time across these models? How do they compare to the corresponding complete data results in Problem 11.10? Discuss how this might affect choice of variance–covariance structure for mixed models with missing data. Would you prefer a more parsimonious structure (like random intercepts) or a richer one (like third-order autoregressive)? Explain.

11.15 Learning Objectives

(1) Define the different types of missing data mechanisms (MCAR, CD-MCAR, MAR, MNAR).
(2) Explain why complete case analysis may lead to biased and/or inefficient analyses.
(3) Explain the drawbacks of LOCF as an imputation method.
(4) Identify situations in which ICEs multiple imputation is to be preferred over MVN multiple imputation.
(5) Use ICEs multiple imputation and MVN multiple imputation to analyze datasets with missing predictor information.
(6) Explain why maximum-likelihood methods for longitudinal data can be considered methods for handling missing data.
(7) Explain how multiple imputation can be used as a sensitivity analysis when data are MNAR.
(8) Use pattern mixture models to analyze datasets with missing outcome data.