CHAPTER 25

Missing-data imputation

Missing data arise in almost all serious statistical analyses. In this chapter we discuss a variety of methods to handle missing data, including some relatively simple approaches that can often yield reasonable results. We use as a running example the Social Indicators Survey, a telephone survey of New York City families conducted every two years by the Columbia University School of Social Work. Nonresponse in this survey is a distraction to our main goal of studying trends in attitudes and economic conditions, and we would like to simply clean the dataset so it could be analyzed as if there were no missingness. After some background in Sections 25.1–25.3, we discuss in Sections 25.4–25.5 our general approach of random imputation. Section 25.6 discusses situations where the missing-data process must be modeled (this can be done in Bugs) in order to perform imputations correctly.

Missing data in R and Bugs

In R, missing values are indicated by NA's. For example, to see some of the data from five respondents in the data file for the Social Indicators Survey (arbitrarily picking rows 91–95), we type

R code

```
cbind (sex, race, educ_r, r_age, earnings, police)[91:95,]
```

and get

R output

```
      sex race educ_r r_age earnings police
[91,]   1    3      3    31       NA      0
[92,]   2    1      2    37   135.00      1
[93,]   2    3      2    40       NA      1
[94,]   1    1      3    42     3.00      1
[95,]   1    3      1    24     0.00     NA
```

In classical regression (as well as most other models), R automatically excludes all cases in which any of the inputs are missing; this can limit the amount of information available in the analysis, especially if the model includes many inputs with potential missingness. This approach is called a complete-case analysis, and we discuss some of its weaknesses below.

In Bugs, missing *outcomes* in a regression can be handled easily by simply including the data vector, NA's and all. Bugs explicitly models the outcome variable, and so it is trivial to use this model to, in effect, impute missing values at each iteration.

Things become more difficult when predictors have missing values. For example, if we wanted to model attitudes toward the police, given earnings and demographic predictors, then the model would *not* automatically account for the missing values of earnings. We would have to remove the missing values, impute them, or model them. In Bugs, regression predictors are typically unmodeled and so Bugs does not know how to draw from a predictive distribution for them. To handle missing data in the predictors, Bugs regression models such as those in Part IIB need to be extended by modeling (that is, supplying distributions for) the input variables.

25.1 Missing-data mechanisms

To decide how to handle missing data, it is helpful to know why they are missing. We consider four general "missingness mechanisms," moving from the simplest to the most general.

1. *Missingness completely at random.* A variable is *missing completely at random* if the probability of missingness is the same for all units, for example, if each survey respondent decides whether to answer the "earnings" question by rolling a die and refusing to answer if a "6" shows up. If data are missing completely at random, then throwing out cases with missing data does not bias your inferences.

2. *Missingness at random.* Most missingness is *not* completely at random, as can be seen from the data themselves. For example, the different nonresponse rates for whites and blacks (see Exercise 25.1) indicate that the "earnings" question in the Social Indicators Survey is not missing completely at random.

   A more general assumption, *missing at random*, is that the probability a variable is missing depends only on available information. Thus, if sex, race, education, and age are recorded for all the people in the survey, then "earnings" is missing at random if the probability of nonresponse to this question depends only on these other, fully recorded variables. It is often reasonable to model this process as a logistic regression, where the outcome variable equals 1 for observed cases and 0 for missing.

   When an outcome variable is missing at random, it is acceptable to exclude the missing cases (that is, to treat them as NA's), as long as the regression controls for all the variables that affect the probability of missingness. Thus, any model for earnings would have to include predictors for ethnicity, to avoid nonresponse bias.

   This missing-at-random assumption (a more formal version of which is sometimes called the ignorability assumption) in the missing-data framework is the basically same sort of assumption as ignorability in the causal framework. Both require that sufficient information has been collected that we can "ignore" the assignment mechanism (assignment to treatment, assignment to nonresponse).

3. *Missingness that depends on unobserved predictors.* Missingness is no longer "at random" if it depends on information that has not been recorded and this information also predicts the missing values. For example, suppose that "surly" people are less likely to respond to the earnings question, surliness is predictive of earnings, and "surliness" is unobserved. Or, suppose that people with college degrees are less likely to reveal their earnings, having a college degree is predictive of earnings, and there is also some nonresponse to the education question. Then, once again, earnings are not missing at random.

   A familiar example from medical studies is that if a particular treatment causes discomfort, a patient is more likely to drop out of the study. This missingness is not at random (unless "discomfort" is measured and observed for all patients).

   If missingness is not at random, it must be explicitly modeled, or else you must accept some bias in your inferences.

4. *Missingness that depends on the missing value itself.* Finally, a particularly difficult situation arises when the probability of missingness depends on the (potentially missing) variable itself. For example, suppose that people with higher earnings are less likely to reveal them. In the extreme case (for example, all persons earning more than $100,000 refuse to respond), this is called *censoring*, but even the probabilistic case causes difficulty.

Censoring and related missing-data mechanisms can be modeled (as discussed in Section 18.5) or else mitigated by including more predictors in the missing-data model and thus bringing it closer to missing at random. For example, whites and persons with college degrees tend to have higher-than-average incomes, so controlling for these predictors will somewhat—but probably only somewhat—correct for the higher rate of nonresponse among higher-income people. More generally, while it can be possible to predict missing values based on the other variables in your dataset, just as with other missing-data mechanisms, this situation can be more complicated in that the nature of the missing-data mechanism may force these predictive models to extrapolate beyond the range of the observed data.

General impossibility of proving that data are missing at random

As discussed above, missingness at random is relatively easy to handle—simply include as regression inputs all variables that affect the probability of missingness. Unfortunately, we generally cannot be sure whether data really are missing at random, or whether the missingness depends on unobserved predictors or the missing data themselves. The fundamental difficulty is that these potential "lurking variables" are unobserved—by definition—and so we can never rule them out. We generally must make assumptions, or check with reference to other studies (for example, surveys in which extensive follow-ups are done in order to ascertain the earnings of nonrespondents).

In practice, we typically try to include as many predictors as possible in a model so that the "missing at random" assumption is reasonable. For example, it may be a strong assumption that nonresponse to the earnings question depends only on sex, race, and education—but this is a lot more plausible than assuming that the probability of nonresponse is constant, or that it depends only on one of these predictors.

25.2 Missing-data methods that discard data

Many missing data approaches simplify the problem by throwing away data. We discuss in this section how these approaches may lead to biased estimates (one of these methods tries to directly address this issue). In addition, throwing away data can lead to estimates with larger standard errors due to reduced sample size.

Complete-case analysis

A direct approach to missing data is to exclude them. In the regression context, this usually means *complete-case analysis*: excluding all units for which the outcome or any of the inputs are missing. In R, this is done automatically for classical regressions (data points with any missingness in the predictors or outcome are ignored by the regression). In Bugs, missing values in unmodeled data are not allowed, so these cases must be excluded in R before sending the data to Bugs, or else the variables with missingness must be explicitly modeled (see Section 25.6).

Two problems arise with complete-case analysis:

1. If the units with missing values differ systematically from the completely observed cases, this could bias the complete-case analysis.
2. If many variables are included in a model, there may be very few complete cases, so that most of the data would be discarded for the sake of a simple analysis.

Available-case analysis

Another simple approach is *available-case analysis*, where different aspects of a problem are studied with different subsets of the data. For example, in the 2001 Social Indicators Survey, all 1501 respondents stated their education level, but 16% refused to state their earnings. We could thus summarize the distribution of education levels of New Yorkers using all the responses and the distribution of earnings using the 84% of respondents who answered that question. This approach has the problem that different analyses will be based on different subsets of the data and thus will not necessarily be consistent with each other. In addition, as with complete-case analysis, if the nonrespondents differ systematically from the respondents, this will bias the available-case summaries. For example in the Social Indicators Survey, 90% of African Americans but only 81% of whites report their earnings, so the "earnings" summary represents a different population than the "education" summary.

Available-case analysis also arises when a researcher simply excludes a variable or set of variables from the analysis because of their missing-data rates (sometimes called "complete-variables analyses"). In a causal inference context (as with many prediction contexts), this may lead to omission of a variable that is necessary to satisfy the assumptions necessary for desired (causal) interpretations.

Nonresponse weighting

As discussed previously, complete-case analysis can yield biased estimates because the sample of observations that have no missing data might not be representative of the full sample. Is there a way of reweighting this sample so that representativeness is restored?

Suppose, for instance, that only one variable has missing data. We could build a model to predict the nonresponse in that variable using all the other variables. The inverse of predicted probabilities of response from this model could then be used as survey weights to make the complete-case sample representative (along the dimensions measured by the other predictors) of the full sample. This method becomes more complicated when there is more than one variable with missing data. Moreover, as with any weighting scheme, there is the potential that standard errors will become erratic if predicted probabilities are close to 0 or 1.

25.3 Simple missing-data approaches that retain all the data

Rather than removing variables or observations with missing data, another approach is to fill in or "impute" missing values. A variety of imputation approaches can be used that range from extremely simple to rather complex. These methods keep the full sample size, which can be advantageous for bias and precision; however, they can yield different kinds of bias, as detailed in this section.

Whenever a single imputation strategy is used, the standard errors of estimates tend to be too low. The intuition here is that we have substantial uncertainty about the missing values, but by choosing a single imputation we in essence pretend that we know the true value with certainty.

Mean imputation. Perhaps the easiest way to impute is to replace each missing value with the mean of the observed values for that variable. Unfortunately, this strategy can severely distort the distribution for this variable, leading to complications with summary measures including, notably, underestimates of the standard deviation. Moreover, mean imputation distorts relationships between variables by "pulling" estimates of the correlation toward zero.

Last value carried forward. In evaluations of interventions where pre-treatment measures of the outcome variable are also recorded, a strategy that is sometimes used is to replace missing outcome values with the pre-treatment measure. This is often thought to be a conservative approach (that is, one that would lead to underestimates of the true treatment effect). However, there are situations in which this strategy can be *anti*conservative. For instance, consider a randomized evaluation of an intervention that targets couples at high risk of HIV infection. From the regression-to-the-mean phenomenon (see Section 4.3), we might expect a reduction in risky behavior even in the absence of the randomized experiment; therefore, carrying the last value forward will result in values that look worse than they truly are. Differential rates of missing data across the treatment and control groups will result in biased treatment effect estimates that are anticonservative.

Using information from related observations. Suppose we are missing data regarding the income of fathers of children in a dataset. Why not fill these values in with mother's report of the values? This is a plausible strategy, although these imputations may propagate measurement error. Also we must consider whether there is any incentive for the reporting person to misrepresent the measurement for the person about whom he or she is providing information.

Indicator variables for missingness of categorical predictors. For unordered categorical predictors, a simple and often useful approach to imputation is to add an extra category for the variable indicating missingness.

Indicator variables for missingness of continuous predictors. A popular approach in the social sciences is to include for each continuous predictor variable with missingness an extra indicator identifying which observations on that variable have missing data. Then the missing values in the partially observed predictor are replaced by zeroes or by the mean (this choice is essentially irrelevant). This strategy is prone to yield biased coefficient estimates for the other predictors included in the model because it forces the slope to be the same across both missing-data groups. Adding interactions between an indicator for response and these predictors can help to alleviate this bias (this leads to estimates similar to complete-case estimates).

Imputation based on logical rules. Sometimes we can impute using logical rules: for example, the Social Indicators Survey includes a question on "number of months worked in the previous year," which all 1501 respondents answered. Of the persons who refused to answer the earnings question, 10 reported working zero months during the previous year, and thus we could impute zero earnings to them. This type of imputation strategy does not rely on particularly strong assumptions since, in effect, the missing-data mechanism is known.

25.4 Random imputation of a single variable

When more than a trivial fraction of data are missing, however, we prefer to perform imputations more formally. In order to understand missing-data imputation, we start with the relatively simple setting in which missingness is confined to a single variable, y, with a set of variables X that are observed on all units. We shall consider the case of imputing missing earnings in the Social Indicators Survey.

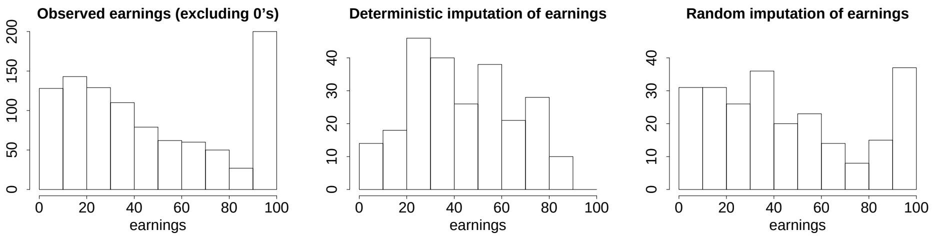

Figure 25.1 *Histogram of earnings (in thousands of dollars) in the Social Indicators Survey: (a) for the 988 respondents who answered the question and had positive earnings, (b) deterministic imputations for the 241 missing values from a regression model, (c) random imputations from that mode. All values are topcoded at 100, with zero values excluded.*

Simple random imputation

The simplest approach is to impute missing values of earnings based on the observed data for this variable. We can write this as an R function:

R code

```
random.imp <- function (a){
  missing <- is.na(a)
  n.missing <- sum(missing)
  a.obs <- a[!missing]
  imputed <- a
  imputed[missing] <- sample (a.obs, n.missing, replace=TRUE)
  return (imputed)
}
```

(To see how this function works, take a small dataset and evaluate the function line by line.) We use `random.imp` to create a *completed data* vector of earnings:

R code

```
earnings.imp <- random.imp (earnings)
```

imputing into the missing values of the original `earnings` variable. This approach does not make much sense—it ignores the useful information from all the other questions asked of these survey responses—but these simple random imputations can be a convenient starting point. A better approach is to fit a regression to the observed cases and then use that to predict the missing cases, as we show next.

Zero coding and topcoding

We begin with some practicalities of the measurement scale. We shall fit the regression model to those respondents whose earnings were observed and positive (since, as noted earlier, the respondents with zero earnings can be identified from their zero responses to the "months worked" question). In addition, we shall "topcode" all earnings at $100,000—that is, all responses above this value will be set to $100,000—before running the regression. Figure 25.1a shows the distribution of positive earnings after topcoding.

R code

```
topcode <- function (a, top){
  return (ifelse (a>top, top, a))
}
earnings.top <- topcode (earnings, 100)   # earnings are in $thousands
hist (earnings.top[earnings>0])
```

The topcoding reduces the sensitivity of the results to the highest values, which in this survey go up to the millions. By topcoding we lose information, but the main use of earnings in this survey is to categorize families into income quantiles, for which purpose topcoding at $100,000 has no effect.

Similarly, we topcoded number of hours worked per week at 40 hours. The purpose of topcoding was not to correct the data—we have no particular reason to disbelieve the high responses—but rather to perform a simple transformation to improve the predictive power of the regression model.

Using regression predictions to perform deterministic imputation

A simple and general imputation procedure that uses individual-level information uses a regression to the nonzero values of earnings. We begin by setting up a data frame with all the variables we shall use in our analysis:

R code

```
sis <- data.frame (cbind (earnings, earnings.top, male, over65, white,
  immig, educ_r, workmos, workhrs.top, any.ssi, any.welfare, any.charity))
```

and then fit a regression to positive values of earnings:

R code

```
lm.imp.1 <- lm (earnings ~ male + over65 + white + immig + educ_r +
  workmos + workhrs.top + any.ssi + any.welfare + any.charity,
  data=SIS, subset=earnings>0)
```

We shall describe these predictors shortly, but first we go through the steps needed to create deterministic and then random imputations. We first get predictions for all the data:

R code

```
pred.1 <- predict (lm.imp.1, SIS)
```

To get predictions for the entire data vector, we must include the data frame, `sis`, in the `predict()` call. Simply writing `predict(lm.imp.1)` would give predictions only for the data used in the fitting, which in this case are the subset of cases for which earnings are positive and for which none of the variables used in the regression are missing.

Next we write a little function to create a completed dataset by imputing the predictions into the missing values:

R code

```
impute <- function (a, a.impute){
  ifelse (is.na(a), a.impute, a)
}
```

and use this to impute missing earnings:

R code

```
earnings.imp.1 <- impute (earnings, pred.1)
```

Transforming and topcoding. For the purpose of predicting incomes in the low and middle range (where we are most interested), we can do better by working on the square root scale of income, topcoded to 100 (in thousands of dollars):

R code

```
lm.imp.2.sqrt <- lm (I(sqrt(earnings.top)) ~ male + over65 + white +
  immig + educ_r + workmos + workhrs.top + any.ssi + any.welfare +
  any.charity, data=SIS, subset=earnings>0)
display (lm.imp.2.sqrt)
pred.2.sqrt <- predict (lm.imp.2.sqrt, SIS)
pred.2 <- topcode (pred.2.sqrt^2, 100)
earnings.imp.2 <- impute (earnings.top, pred.2)
```

Here is the fitted model:

R output

```
               coef.est coef.se
(Intercept)    -1.67    0.44
male            0.32    0.13
over65         -1.44    0.58
white           0.96    0.15
immig          -0.62    0.14
educ_r          0.79    0.07
workmos         0.33    0.03
workhrs.top     0.06    0.01
any.ssi        -0.97    0.55
any.welfare    -1.35    0.37
any.charity    -1.17    0.60
  n = 988, k = 11
  residual sd = 1.96, R-Squared = 0.44
```

Figure 25.1b shows the deterministic imputations:

R code

```
hist (earnings.imp.2[is.na(earnings)])
```

From this graph, it appears that most of the nonrespondents have incomes in the middle range (compare to Figure 25.1a). Actually, the central tendency of Figure 25.1b is an artifact of the deterministic imputation procedure. One way to see this is through the regression model: its R^2 is 0.44, which means that the explained variance from the regression is only 44% of the total variance. Equivalently, the explained standard deviation is $\sqrt{0.44} = 0.66 = 66\%$ of the data standard deviation. Hence, the predicted values from the regression will tend to be less variable than the original data. If we were to use the resulting deterministic imputations, we would be falsely implying that most of these nonrespondents had incomes in the middle of the scale.

Random regression imputation

We can put the uncertainty back into the imputations by adding the prediction error into the regression, as discussed in Section 7.2. For this example, this involves creating a vector of random predicted values for the 241 missing cases using the normal distribution, and then squaring, as before, to return to the original dollar scale:

R code

```
pred.4.sqrt <- rnorm (n, predict (lm.imp.2.sqrt, SIS),
  sigma.hat (lm.imp.2.sqrt))
pred.4 <- topcode (pred.4.sqrt^2, 100)
earnings.imp.4 <- impute (earnings.top, pred.4)
```

Figure 25.1c shows the resulting imputed values from a single simulation draw. Compared to Figure 25.1b, these random imputations are more appropriately spread across the range of the population.

The new imputations certainly do not look perfect—in particular, there still seem to be too few imputations at the topcoded value of $100,000—suggesting that the linear model on the square root scale, with normal errors, is not quite appropriate for these data. (This makes sense given the spike in the data from the topcoding.) The results look much better than the deterministic imputations, however.

Figure 25.2 illustrates the deterministic and random imputations in another way. The left plot in the figure shows the deterministic imputations as a function of the predicted earnings from the regression model. By the definition of the imputation procedure, the values are identical and so the points fall along the identity line. The right plot shows the random imputations, which follow a generally increasing

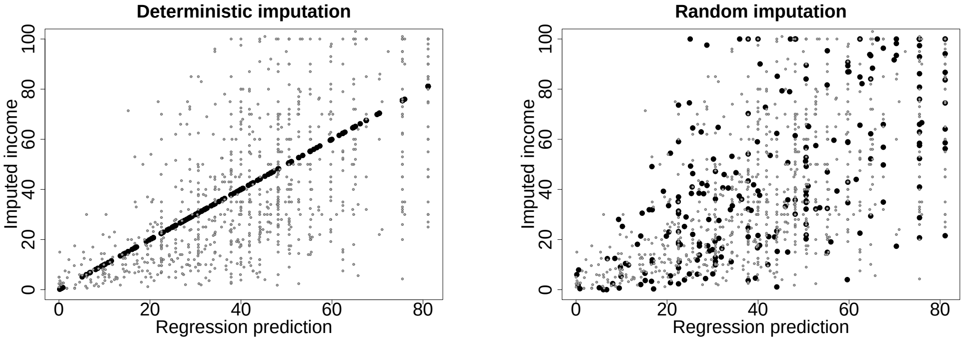

Figure 25.2 *Deterministic and random imputations for the 241 missing values of earnings in the Social Indicators Survey. The deterministic imputations are exactly at the regression predictions and ignore predictive uncertainty. In contrast, the random imputations are more variable and better capture the range of earnings in the data. See also Figure 25.1.*

pattern but with scatter derived from the unexplained variance in the model. (The increase in variance as a function of predicted value arises from fitting the model on the square root scale and squaring at the end.)

Predictors used in the imputation model

We fit a regression of earnings on sex, age, ethnicity, nationality, education, the number of months worked in the previous year and hours worked per week, and indicators for whether the respondent's family receives each of three forms of income support (from disability payments, welfare, and private charities).

It might seem strange to model earnings given information on income support—which is, in part, a consequence of earnings—but for the purposes of imputation this is acceptable. The goal here is not causal inference but simply accurate prediction, and it is acceptable to use any inputs in the imputation model to achieve this goal.

Two-stage modeling to impute a variable that can be positive or zero

In the Social Indicators Survey, we only need to impute the positive values of earnings: the "hours worked" and "months worked" questions were answered by everyone in the survey, and these variables are a perfect predictor of whether the value of earnings (more precisely, employment income) is positive. For the missing cases of `earnings`, we can impute 0 if `workhrs` $= 0$ and `workmos` $= 0$, and impute a continuous positive value when either of these is positive. This imputation process is what was described above, with the regression based on $n = 988$ data points and displayed in Figure 25.2. The survey as a whole included 1501 families, of whom 272 reported working zero hours and months and were thus known to have zero earnings. Of the 1229 persons reporting positive working hours or months, 988 responded to the earnings question and 241 did not.

Now suppose that the `workhrs` and `workmos` variables were *not* available, so that we could not immediately identify the cases with zero earnings. We would then impute missing responses to the earnings question in two steps: first, imputing an indicator for whether earnings are positive, and, second, imputing the continuous positive values of earnings.

Mathematically, we would impute earnings y given regression predictors X in a two-step process, defining

$$y = I^y y^{\text{pos}},$$

where $I^y = 1$ if $y > 0$ and 0 otherwise, and $y^{\text{pos}} = y$ if $y > 0$. The first model is a logistic regression for I^y:

$$\Pr(I_i^y = 1) = \text{logit}^{-1}(X_i\alpha),$$

and the second part is a linear regression for the square root of y^{pos}:

$$\sqrt{y_i^{\text{pos}}} \sim \text{N}(X_i\beta, \sigma^2).$$

The first model is fit to all the data for which y is observed, and the second model is fit to all the data for which y is observed and positive.

We illustrate with the earnings example. First we fit the two models:

R code

```
glm.sign <- glm (I(earnings>0) ~ male + over65 + white +
  immig + educ_r + any.ssi + any.welfare + any.charity,
  data=SIS, family=binomial(link=logit))
display (glm.sign)
lm.ifpos.sqrt <- lm (I(sqrt(earnings.top)) ~ male + over65 + white +
  immig + educ_r + any.ssi + any.welfare + any.charity,
  data=SIS, subset=earnings>0)           # (same as lm.imp.2 from above)
display (lm.ifpos.sqrt)
```

Then we impute whether missing earnings are positive:

R code

```
pred.sign <- rbinom (n, 1, predict (glm.sign, data, type="response"))
pred.pos.sqrt <- rnorm (n, predict (lm.ifpos.sqrt, SIS),
  sigma.hat(lm.ifpos.sqrt))
```

and then impute the earnings themselves:

R code

```
pred.pos <- topcode (pred.pos.sqrt^2, 100)
earnings.imp <- impute (earnings, pred.sign*pred.pos)
```

Matching and hot-deck imputation

A different way to impute is through *matching*: for each unit with a missing y, find a unit with similar values of X in the observed data and take its y value. This approach is also sometimes called “hot-deck” imputation (in contrast to “cold deck” methods, where the imputations come from a previously collected data source). Matching imputation can be combined with regression by defining “similarity” as closeness in the regression predictor (for example, $0.32\cdot\texttt{male} - 1.44\cdot\texttt{over65} + 0.96\cdot\texttt{white} + \cdots$ for the model on page 536). Matching can be viewed as a nonparametric or local version of regression and can also be useful in some settings where setting up a regression model can be challenging.

For example, the New York City Department of Health has the task of assigning risk factors to all new HIV cases. The risk factors are assessed from a reading of each patient’s medical file, but for a large fraction of the cases, not enough information is available to determine the risk factors. For each of these “unresolved” cases, we proposed taking a random imputation from the risk factors of the five closest resolved cases, where “closest” is defined based on a scoring function that penalizes differences in sex, age, the clinic where the HIV test was conducted, and other information that is available on all or most cases.

More generally, one could estimate a propensity score that predicts the missingness of a variable conditional on several other variables that are fully observed, and then match on this propensity score to impute missing values.

25.5 Imputation of several missing variables

It is common to have missing data in several variables in an analysis, in which case one cannot simply set up a model for a single partially observed variable y given a set of fully observed X variables. In fact, even in the Social Indicators Survey example, some of the predictor variables (ethnicity, interest income, and the indicators for income supplements) had missing values in the data, which we crudely imputed before running the regression for the imputations. More generally, we must think of the dataset as a multivariate outcome, any components of which can be missing.

Routine multivariate imputation

The direct approach to imputing missing data in several variables is to fit a multivariate model to all the variables that have missingness, thus generalizing the approach of Section 25.4 to allow the outcome Y as well as the predictors X to be vectors. The difficulty of this approach is that it requires a lot of effort to set up a reasonable multivariate regression model, and so in practice an off-the-shelf model is typically used, most commonly the multivariate normal or t distribution for continuous outcomes, and a multinomial distribution for discrete outcomes. Software exists to fit such models automatically, so that one can conceivably "press a button" and impute missing data. These imputations are only as good as the model, and so they need to be checked in some way—but this automatic approach is easy enough that it is a good place to start, in any case.

Iterative regression imputation

A different way to generalize the univariate methods of the previous section is to apply them iteratively to the variables with missingness in the data. If the variables with missingness are a matrix Y with columns $Y_{(1)}, \dots, Y_{(K)}$ and the fully observed predictors are X, this entails first imputing all the missing Y values using some crude approach (for example, choosing imputed values for each variable by randomly selecting from the observed outcomes of that variable); and then imputing $Y_{(1)}$ given $Y_{(2)}, \dots, Y_{(K)}$ and X; imputing $Y_{(2)}$ given $Y_{(1)}, Y_{(3)}, \dots, Y_{(K)}$ and X (using the newly imputed values for $Y_{(1)}$), and so forth, randomly imputing each variable and looping through until approximate convergence.

For example, the Social Indicators Survey asks about several sources of income. It would be helpful to use these to help impute each other since they have non-overlapping patterns of missingness. We illustrate for the simple case of imputing missing data for two variables—interest income and earnings—using the same fully observed predictors used to impute earnings in the previous section.

We create random imputations to get the process started:

R code

```
interest.imp <- random.imp (interest)
earnings.imp <- random.imp (earnings)
```

and then we write a loop to iteratively impute. For simplicity in demonstrating the programming, we set up the function on the original (non-square-root) scale of the data:

R code

```
n.sims <- 10
for (s in 1:n.sims){
  lm.1 <- lm (earnings ~ interest.imp + male + over65 + white +
    immig + educ_r + workmos + workhrs.top + any.ssi + any.welfare +
      any.charity)
    pred.1 <- rnorm (n, predict(lm.1), sigma.hat(lm.1))
    earnings.imp <- impute (earnings, pred.1)

    lm.2 <- lm (interest ~ earnings.imp + male + over65 + white +
      immig + educ_r + workmos + workhrs.top + any.ssi + any.welfare +
      any.charity)
    pred.2 <- rnorm (n, predict(lm.2), sigma.hat(lm.2))
    interest.imp <- impute (interest, pred.2)
  }
```

This code could be easily elaborated to handle topcoding, transformations, and two-stage modeling for variables that could be zero or positive (see Exercise 25.4). These operations should be done within the imputation loop, not merely tacked on at the end.

Iterative regression imputation has the advantage that, compared to the full multivariate model, the set of separate regression models (one for each variable, $Y_{(k)}$) is easier to understand, thus allowing the imputer to potentially fit a reasonable model at each step. Moreover, it is easier in this setting to allow for interactions (difficult to do using most joint model specifications).

The disadvantage of the iterative approach is that the researcher has to be more careful in this setting to ensure that the separate regression models are consistent with each other. For instance, it would not make sense to impute age based on income but then to later ignore age when imputing income.

Moreover, even if such inconsistencies are avoided, the resulting specification will not in general correspond to any joint probability model for all of the variables being imputed. It is an open research project to develop methods to diagnose problems with multivariate imputations, by analogy to the existing methods such as residual plots for finding problems in regressions. In the meantime, it makes sense to examine histograms and scatterplots of observed and imputed data to check that the imputations are reasonable.

25.6 Model-based imputation

Missing data can be handled in Bugs by modeling the input variables that have missingness. This requires some work, however: with multiple missing input variables, a multivariate model is required, and this can be particularly tricky when some of the variables are discrete. So in practice it can be helpful to do some simple imputation in R, as we have described, before then analyzing completed datasets. When more is known about the missing-data mechanism (for example, with censored or truncated data; see the model on page 405), it can make more sense to explicitly model the missingness in Bugs.

Nonignorable missing-data models

Realistic censored-data problems often have particular complications. For example, in the study of death penalty appeals described in Section 6.3, we are interested in the duration of the appeals process for individual cases. For example, if a death sentence is imposed in 1983 and its final appeal is decided in 1994, then the process lasted 11 years. It is challenging to estimate the distribution of these waiting times, and to model them based on case-level predictors, because our dataset includes appeals only up to the year 1995. Figure 25.3 illustrates. The censoring model, by analogy to model (18.17) on page 404, looks like:

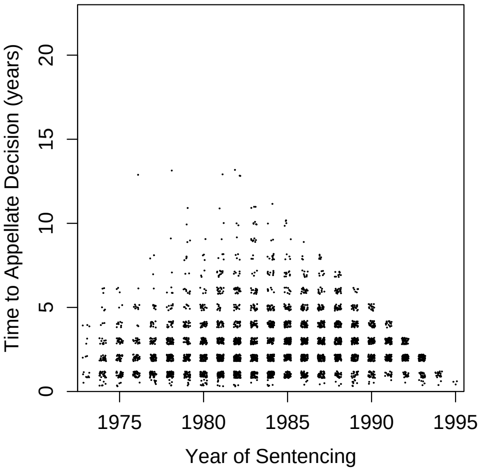

Figure 25.3 *Delays in state appeals court for death penalty cases, plotted versus year of sentencing (jittered to allow individual data points to be visible). We only have results up to the year 1995. The data show a steady increase in delay times for the first decade, but after that, the censoring makes the graph difficult to interpret directly.*

$$y_i = \begin{cases} z_i & \text{if } z_i \leq 1995 - t_i \\ \text{censored} & \text{otherwise,} \end{cases}$$

where y_i is the observed waiting time for case i, z_i is the ultimate waiting time, and t_i is the year of sentencing. We shall not analyze these data further here; we have introduced this example just to illustrate the complexities that arise in realistic censoring situations. The actual analysis for this problem is more complicated because death sentences have three stages of review, and cases can be waiting at any of these stages.

Imputation in multilevel data structures

Imputing becomes more complicated with clustered data. Suppose, for instance, that we have individual-level observations on children grouped within schools (for instance, test scores and demographics), and then measurements pertaining to the schools themselves (for instance, school policies and characteristics such as public versus private). We would not want to impute on a standard individual-level dataset where the school-level measurements are just repeated over each individual in the same school because, if a given school measurement is missing, such an approach would not be likely to impute the same value of this variable for each member of the group (as it should).

Our general advice in this situation is to create two datasets, as in Figure 11.3 on page 239, one with only individual-level data, and one with group-level data and do separate imputations within each dataset while using results from one in the other (perhaps iterating back and forth). For instance, one could first impute individual-level variables using individual-level data and observed group-level measurement. Then in the group-level dataset one could include aggregated forms of the individual-level measurements when imputing missingness at this level.

25.7 Combining inferences from multiple imputations

Rather than replacing each missing value in a dataset with one randomly imputed value, it may make sense to replace each with several imputed values that reflect our uncertainty about our imputation model. For example, if we impute using a regression model we may want our imputations to reflect not only sampling variability (as random imputation should) but also our uncertainty about the regression coefficients in the model. If these coefficients themselves are modeled, we can draw a new set of missing value imputations for each draw from the distribution of the coefficients.

Multiple imputation does this by creating several (say, five) imputed values for each missing value, each of which is predicted from a slightly different model and each of which also reflects sampling variability. How do we analyze these data? The simple idea is to use each set of imputed values to form (along with the observed data) a *completed* dataset. Within each completed dataset a standard analysis can be run. Then inferences can be combined across datasets.

For instance, suppose we want to make inferences about a regression coefficient, β. We obtain estimates $\hat{\beta}_m$ in each of the M datasets as well as standard errors, $s_1, \ldots, s_M$. To obtain an overall point estimate, we then simply average over the estimates from the separate imputed datasets; thus, $\hat{\beta} = \frac{1}{m}\sum_{m=1}^{M}\hat{\beta}_m$. A final variance estimate V_β reflects variation within and between imputations:

$$V_\beta = W + \left(1 + \frac{1}{m}\right) B,$$

where $W = \frac{1}{m}\sum_{m=1}^{M} s_m^2$, and $B = \frac{1}{m-1}\sum_{m=1}^{M}(\hat{\beta}_m - \hat{\beta})^2$.

If missing data have been included in the main data analysis (as when variables X and y are given distributions in a Bugs model), the uncertainty about the missing-data imputations is automatically included in the Bayesian inference, and the above steps are not needed.

25.8 Bibliographic note

Little and Rubin (2002) provide an overview of methods for analysis with missing data. For more on multiple imputation in particular, see Rubin (1987, 1996). "Missing at random" and related concepts were formalized by Rubin (1976). A simple discrete-data example appears in Rubin, Stern, and Vehovar (1995). King et al. (2001) review many of the practical costs and benefits of multiple imputation.

For routine imputation of missing data, Schafer (1997) presents a method based on the multivariate normal distribution, Liu (1995) uses the t distribution, and Van Buuren, Boshuizen, and Knook (1999) use interlocking regressions. Abayomi, Gelman, and Levy (2005) discuss methods for checking the fit of imputation models, and Troxel, Ma, and Heitjan (2004) present a method to assess sensitivity of inferences to missing-data assumptions.

Software for routine imputation in R and SAS has been developed by Van Buuren and Oudshoom (2000), Raghunathan, Van Hoewyk, and Solenberger (2001), and Raghunathan, Solenberger, and Van Hoewyk (2002). An overview of some imputation software is at `www.missing-data.com`.

Specialized imputation models have been developed for particular problems, with multilevel models used to adjust for discrete predictors. Some examples include Clogg et al. (1991), Belin et al. (1993), and Gelman, King, and Liu (1998). See also David et al. (1986).

Meng (1994), Fay (1996), Rubin (1996), Clayton et al. (1998), and Robins and

Wang (2000) discuss situations in which the standard rules for combining multiple imputations have problems. Barnard and Meng (1994) and Robins and Wang (2000) propose alternative variance estimators and reference distributions.

For more on the Social Indicators Survey, see Garfinkel and Meyers (1999). The death-sentencing example is discussed by Gelman, Liebman, et al. (2004) and Gelman (2004a); see also Finkelstein et al. (2006).

25.9 Exercises

1. Based on the summaries at the very end of Section 25.2, show that the response rates for the "earnings" question in the Social Indicators Survey are statistically significantly different for whites and blacks.
2. Take a complete dataset (with no missingness) of interest to you with two variables, x and y. Call this the "full data."
   (a) Write a program in R to cause approximately half of the values of x to be missing. Design this missingness mechanism to be at random but *not* completely at random; that is, the probability that x is missing should depend on y. Call this new dataset, with missingness in x, the "available data."
   (b) Perform the regression of x on y (that is, with y as predictor and x as outcome) using complete-case analysis (that is, using only the data for which both variables are observed) and show that it is consistent with the regression on the full data.
   (c) Perform the complete-case regression of y on x and show that it is *not* consistent with the corresponding regression on the full data.
   (d) Using just the available data, fit a model in R for x given y, and use this model to randomly impute the missing x data. Perform the regression of y on x using this imputed dataset and compare to your results from (c).
3. Nonignorable missing data: in Exercise 9.13, you estimated the effects of incumbency in U.S. congressional elections, discarding uncontested elections.
   (a) Construct three "bad" imputation procedures and one "good" imputation procedure for these uncontested elections.
   (b) Define clearly how to interpret these imputations. (These election outcomes are not actually "missing"—it is known that they were uncontested.)
   (c) Fit the model to the completed dataset under each of the imputation procedures from (a) and compare the results.
4. Use iterative regression to impute missing data for all the income components in the Social Indicators Survey (data at folder `sis`).