Chapter 1

Introduction

1.1 The problem of missing data

1.1.1 Current practice

The mean of the numbers 1, 2 and 4 can be calculated in R as

```
> y <- c(1, 2, 4)
> mean(y)
[1] 2.3
```

where y is a vector containing three numbers, and where mean(y) is the R expression that returns their mean. Now suppose that the last number is missing. R indicates this by the symbol NA, which stands for "not available":

```
> y <- c(1, 2, NA)
> mean(y)
[1] NA
```

The mean is now undefined, and R informs us about this outcome by setting the mean to NA. It is possible to add an extra argument na.rm = TRUE to the function call. This removes any missing data before calculating the mean:

```
> mean(y, na.rm = TRUE)
[1] 1.5
```

This makes it possible to calculate a result, but of course the set of observations on which the calculations are based has changed. This may cause problems in statistical inference and interpretation.

Similar problems occur in multivariate analysis. Many users of R will have seen the following error message:

```
> lm(Ozone ~ Wind, data = airquality)
Error in na.fail.default(list(Ozone = c(41, 36, 12, 18, NA,
  missing values in object
```

This code calls function `lm()` to fit a linear regression to predict daily ozone concentration (ppb) from wind speed (mph) using the built-in dataset `airquality`. The program cannot continue because the value of `Ozone` is unknown for some days. It is easy to omit any incomplete records by specifying the `na.action = na.omit` argument to `lm()`. The regression weights can now be obtained as

```
> fit <- lm(Ozone ~ Wind, data = airquality,
      na.action = na.omit)
> coef(fit)
(Intercept)        Wind
       96.9        -5.6
```

The `R` object `fit` stores the results of the regression analysis, and the `coef()` function extract the regression weights from it. In practice, it is cumbersome to supply the `na.action()` function each time. We can change the factory-fresh setting of the `options` as

```
> options(na.action = na.omit)
```

This command eliminates the error message once and for all. Users of other software packages like `SPSS`, `SAS` and `Stata` enjoy the "luxury" that this deletion option has already been set for them, so the calculations can progress silently. The procedure is known as *listwise deletion* or *complete case analysis*, and is widely used.

Though listwise deletion allows the calculations to proceed, it may cause problems in interpretation. For example, we can find the number of deleted cases in the fitted model as

```
> deleted <- na.action(fit)
> naprint(deleted)
[1] "37 observations deleted due to missingness"
```

The `na.action()` function finds the cases that are deleted from the fitted model. The `naprint()` function echoes the number of deleted cases. Now suppose we fit a better predictive model by including solar radiation (`Solar.R`) in the model. We obtain

```
> fit2 <- lm(Ozone ~ Wind + Solar.R, data = airquality)
> naprint(na.action(fit2))
[1] "42 observations deleted due to missingness"
```

where the previous three separate statements have been combined into one line. The number of deleted days increased from 37 to 42 since some of the days had no value for `Solar.R`. Thus, changing the model altered the sample.

There are methodological and statistical issues associated with this procedure. Some questions that come to mind are:

- Can we compare the regression coefficients from both models?
- Should we attribute differences in the coefficients to changes in the model or to changes in the subsample?
- Do the estimated coefficients generalize to the study population?
- Do we have enough cases to detect the effect of interest?
- Are we making the best use of the costly collected data?

Getting the software to run is one thing, but this alone does not address the questions caused by the missing data. This book discusses techniques that allow us to consider the type of questions raised above.

1.1.2 Changing perspective on missing data

The standard approach to missing data is to delete them. It is illustrative to search for missing values in published data. Hand et al. (1994) published a highly useful collection of small datasets across the statistical literature. The collection covers an impressive variety of topics. Only 13 out of the 510 datasets in the collection actually had a code for the missing data. In many cases, the missing data problem has probably been "solved" in some way, usually without telling us how many missing values there were originally. It is impossible to track down the original data for most datasets in Hand's book. However, we can easily do this for dataset number 357, a list of scores of 34 athletes in 10 sport events at the 1988 Olympic decathlon in Seoul. The table itself is complete, but a quick search on the Internet revealed that initially 39 instead of 34 athletes participated. Five of them did not finish for various reasons, including the dramatic disqualification of the German favorite Jürgen Hingsen because of three false starts in the 100-meter sprint. It is probably fair to assume that deletion occurred silently in many of the other datasets.

The inclination to delete the missing data is understandable. Apart from the technical difficulties imposed by the missing data, the occurrence of missing data has long been considered a sign of sloppy research. It is all too easy for a referee to write:

> This study is weak because of the large amount of missing data.

Publication chances are likely to improve if there is no hint of missingness. Orchard and Woodbury (1972, p. 697) remarked:

> Obviously the best way to treat missing data is not to have them.

Though there is a lot of truth in this statement, Orchard and Woodbury realized the impossibility of attaining this ideal in practice.

The prevailing scientific practice is to downplay the missing data. Reviews

on reporting practices are available in various fields: clinical trials (Wood et al., 2004), cancer research (Burton and Altman, 2004), educational research (Peugh and Enders, 2004), epidemiology (Klebanoff and Cole, 2008), developmental psychology (Jeliĉić et al., 2009), general medicine (Mackinnon, 2010) and developmental pediatrics (Aylward et al., 2010). The picture that emerges from these studies is quite consistent:

- The presence of missing data is often not explicitly stated in the text;
- Default methods like listwise deletion are used without mentioning them;
- Different tables are based on different sample sizes;
- Model-based missing data methods, such as direct likelihood, full information maximum likelihood and multiple imputation, are notably underutilized.

Missing data are there, whether we like it or not. In the social sciences, it is nearly inevitable that some respondents will refuse to participate or to answer certain questions. In medical studies, attrition of patients is very common. Allison (2002, p. 1) begins by observing:

> Sooner or later (usually sooner), anyone who does statistical analysis runs into problems with missing data.

Even the most carefully designed and executed studies produce missing values. The really interesting question is how we deal with incomplete data.

The theory, methodology and software for handling incomplete data problems have been vastly expanded and refined over the last decades. The major statistical analysis packages now have facilities for performing the appropriate analyses. This book aims to contribute to a better understanding of the issues involved, and provides a methodology for dealing with incomplete data problems in practice.

1.2 Concepts of MCAR, MAR and MNAR

Before we review a number of simple fixes for the missing data in Section 1.3 let us take a short look at the terms MCAR, MAR and MNAR. A more detailed definition of these concepts will be given later in Section 2.2.3. Rubin (1976) classified missing data problems into three categories. In his theory every data point has some likelihood of being missing. The process that governs these probabilities is called the *missing data mechanism* or *response mechanism*. The model for the process is called the *missing data model* or *response model*.

If the probability of being missing is the same for all cases, then the data are said to be missing completely at random (MCAR). This effectively implies that causes of the missing data are unrelated to the data. We may consequently ignore many of the complexities that arise because data are missing, apart from the obvious loss of information. An example of MCAR is a weighing scale that ran out of batteries. Some of the data will be missing simply because of bad luck. Another example is when we take a random sample of a population, where each member has the same chance of being included in the sample. The (unobserved) data of members in the population that were not included in the sample are MCAR. While convenient, MCAR is often unrealistic for the data at hand.

If the probability of being missing is the same only within groups defined by the *observed* data, then the data are missing at random (MAR). MAR is a much broader class than MCAR. For example, when placed on a soft surface, a weighing scale may produce more missing values than when placed on a hard surface. Such data are thus not MCAR. If, however, we know surface type and if we can assume MCAR *within* the type of surface, then the data are MAR. Another example of MAR is when we take a sample from a population, where the probability to be included depends on some known property. MAR is more general and more realistic than MCAR. Modern missing data methods generally start from the MAR assumption.

If neither MCAR nor MAR holds, then we speak of missing not at random (MNAR). In the literature one can also find the term NMAR (not missing at random) for the same concept. MNAR means that the probability of being missing varies for reasons that are unknown to us. For example, the weighing scale mechanism may wear out over time, producing more missing data as time progresses, but we may fail to note this. If the heavier objects are measured later in time, then we obtain a distribution of the measurements that will be distorted. MNAR includes the possibility that the scale produces more missing values for the heavier objects (as above), a situation that might be difficult to recognize and handle. An example of MNAR in public opinion research occurs if those with weaker opinions respond less often. MNAR is the most complex case. Strategies to handle MNAR are to find more data about the causes for the missingness, or to perform what-if analyses to see how sensitive the results are under various scenarios.

Rubin's distinction is important for understanding why some methods will not work. His theory lays down the conditions under which a missing data method can provide valid statistical inferences. Most simple fixes only work under the restrictive and often unrealistic MCAR assumption. If MCAR is implausible, such methods can provide biased estimates.

1.3 Simple solutions that do not (always) work

1.3.1 Listwise deletion

Complete case analysis (listwise deletion) is the default way of handling incomplete data in many statistical packages, including `SPSS`, `SAS` and `Stata`. The function `na.omit()` does the same in `S-PLUS` and `R`. The procedure eliminates all cases with one or more missing values on the analysis variables.

The major advantage of complete case analysis is convenience. If the data are MCAR, listwise deletion produces unbiased estimates of means, variances and regression weights. Under MCAR, listwise deletion produces standard errors and significance levels that are correct for the reduced subset of data, but that are often larger relative to all available data.

A disadvantage of listwise deletion is that it is potentially wasteful. It is not uncommon in real life applications that more than half of the original sample is lost, especially if the number of variables is large. King et al. (2001) estimated that the percentage of incomplete records in the political sciences exceeded 50% on average, with some studies having over 90% incomplete records. It will be clear that a smaller subsample could seriously degrade the ability to detect the effects of interest.

If the data are not MCAR, listwise deletion can severely bias estimates of means, regression coefficients and correlations. Little and Rubin (2002, pp. 41–44) showed that the bias in the estimated mean increases with the difference between means of the observed and missing cases, and with the proportion of the missing data. Schafer and Graham (2002) reported an elegant simulation study that demonstrates the bias of listwise deletion under MAR and MNAR. However, complete case analysis is not always bad. The implications of the missing data are different depending on where they occur (outcomes or predictors), and the parameter and model form of the complete data analysis. In the context of regression analysis, listwise deletion possesses some unique properties that make it attractive in particular settings. There are cases in which listwise deletion can provide better estimates than even the most sophisticated procedures. Since their discussion requires a bit more background than can be given here, we defer the treatment to Section 2.6.

Listwise deletion can introduce inconsistencies in reporting. Since listwise deletion is automatically applied to the active set of variables, different analyses on the same data are often based on different subsamples. In principle, it is possible to produce one global subsample using all active variables. In practice, this is unattractive since the global subsample will always have fewer cases than each of the local subsamples, so it is common to create different subsets for different tables. It will be evident that this complicates their comparison and generalization to the study population.

In some cases, listwise deletion can lead to nonsensical subsamples. For example, the rows in the `airquality` dataset used in Section 1.1.1 correspond

to 154 consecutive days between May 1, 1973 and September 30, 1973. Deleting days affects the time basis. It would be much harder, if not impossible, to perform analyses that involve time, e.g., to identify weekly patterns or to fit autoregressive models that predict from previous days.

The opinions on the value of listwise deletion vary. Miettinen (1985, p. 231) described listwise deletion as

> ... the only approach that assures that no bias is introduced under any circumstances...

a bold statement, but incorrect. At the other end of the spectrum we find Enders (2010, p. 39):

> In most situations, the disadvantages of listwise deletion far outweigh its advantages.

Schafer and Graham (2002, p. 156) cover the middle ground:

> If a missing data problem can be resolved by discarding only a small part of the sample, then the method can be quite effective.

The leading authors in the field are, however, wary of providing advice about the percentage of missing cases below which it is still acceptable to do listwise deletion. Little and Rubin (2002) argue that it is difficult to formulate rules of thumb since the consequences of using listwise deletion depend on more than the missing data rate alone. Vach (1994, p. 113) expressed his dislike for simplistic rules as follows:

> It is often supposed that there exists something like a critical missing rate up to which missing values are not too dangerous. The belief in such a global missing rate is rather stupid.

1.3.2 Pairwise deletion

Pairwise deletion, also known as *available-case analysis*, attempts to remedy the data loss problem of listwise deletion. The method calculates the means and (co)variances on all observed data. Thus, the mean of variable X is based on all cases with observed data on X, the mean of variable Y uses all cases with observed Y-values, and so on. For the correlation and covariance, all data are taken on which both X and Y have non-missing scores. Subsequently, the matrix of summary statistics are fed into a program for regression analysis, factor analysis or other modeling procedures.

We can calculate the mean, covariances and correlations of the `airquality` data under pairwise deletion in `R` as:

```
> mean(airquality, na.rm = TRUE)
> cor(airquality, use = "pair")
> cov(airquality, use = "pair")
```

SPSS, SAS and Stata contain many procedures with an option for pairwise deletion. The method is simple, uses all available information and produces consistent estimates of mean, correlations and covariances under MCAR (Little and Rubin, 2002, p. 55). Nevertheless, when taken together these estimates have major shortcomings. The estimates can be biased if the data are not MCAR. Furthermore, there are computational problems. The correlation matrix may not be positive definite, which is requirement for most multivariate procedures. Correlations outside the range [−1,+1] can occur, a problem that comes from different subsets used for the covariances and the variances. Such problems are more severe for highly correlated variables (Little, 1992). Another problem is that it is not clear which sample size should be used for calculating standard errors. Taking the average sample size yields standard errors that are too small (Little, 1992).

The idea behind pairwise deletion is to use all available information. Though this idea is good, the proper analysis of the pairwise matrix requires sophisticated optimization techniques and special formulas to calculate the standard errors (Van Praag et al., 1985; Marsh, 1998). Pairwise deletion should only be used if the procedure that follows it is specifically designed to take deletion into account. The attractive simplicity of pairwise deletion as a general missing data method is thereby lost.

1.3.3 Mean imputation

A quick fix for the missing data is to replace them by the mean. We may use the mode for categorical data. Suppose we want to impute the mean in Ozone and Solar.R of the airquality data. SPSS, SAS and Stata have prebuilt functions that substitute the mean. This book uses the R package mice (Van Buuren and Groothuis-Oudshoorn, 2011). This software is a contributed package that extends the functionality of R. Before mice can be used, it must be installed. An easy way to do this is to type:

```
> install.packages("mice")
```

which searches the Comprehensive R Archive Network (CRAN), and installs the requested package on the local computer. After successful installation, the mice package can be loaded by

```
> library("mice")
```

Imputing the mean in each variable can now be done by

```
> imp <- mice(airquality, method = "mean", m = 1,
      maxit = 1)
 iter imp variable
  1   1  Ozone  Solar.R
```

The argument method="mean" specifies mean imputation, the argument m=1 requests a single imputed dataset, and maxit=1 sets the number of iterations

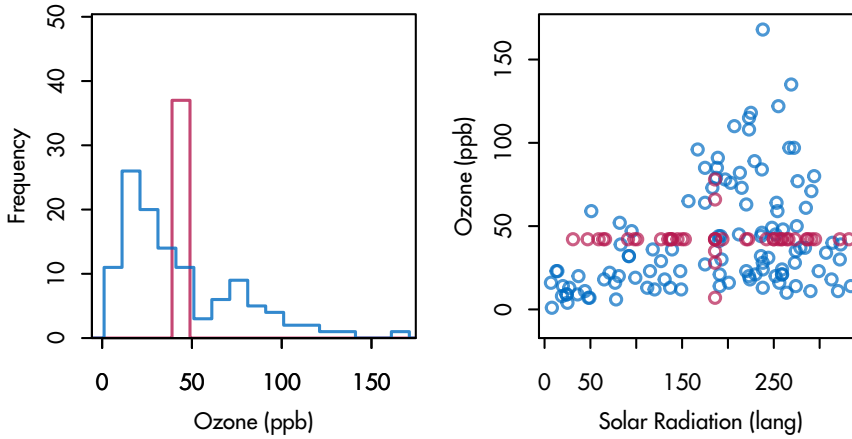

Figure 1.1: Mean imputation of `Ozone`. Blue indicates the observed data, red indicates the imputed values.

to 1 (no iteration). The latter two options options can be left to their defaults with essentially the same result.

Mean imputation distorts the distribution in several ways. Figure 1.1 displays the distribution of `Ozone` after imputation. The `mice` package adopts the Abayomi convention for the colors (Abayomi et al., 2008). Blue refers to the observed part of the data, red to the synthetic part of the data (also called the *imputed values* or *imputations*), and black to the combined data (also called the *imputed data* or *completed data*). The printed version of this book replaces blue by gray. In the figure on the left, the red bar at the mean stands out. Imputing the mean here actually creates a bimodal distribution. The standard deviation in the imputed data is equal to 28.7, much smaller than from the observed data alone, which is 33. The figure on the right-hand side shows that the relation between `Ozone` and `Solar.R` is distorted because of the imputations. The correlation drops from 0.35 in the blue points to 0.3 in the combined data.

Mean imputation is a fast and simple fix for the missing data. However, it will underestimate the variance, disturb the relations between variables, bias almost any estimate other than the mean and bias the estimate of the mean when data are not MCAR. Mean imputation should perhaps only be used as a rapid fix when a handful of values are missing, and it should be avoided in general.

1.3.4 Regression imputation

Regression imputation incorporates knowledge of other variables with the idea of producing smarter imputations. The first step involves building a model

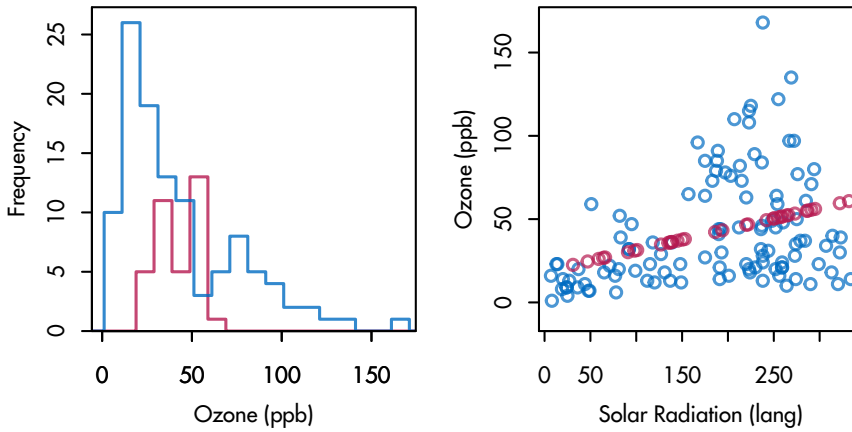

Figure 1.2: Regression imputation: Imputing `Ozone` from the regression line.

from the observed data. Predictions for the incomplete cases are then calculated under the fitted model, and serve as replacements for the missing data. Suppose that we model `Ozone` by the linear regression function of `Solar.R`.

```
> fit <- lm(Ozone ~ Solar.R, data = airquality)
> pred <- predict(fit, newdata = ic(airquality))
```

Figure 1.2 shows the result. The imputed values correspond to the most likely values under the model. However, the ensemble of imputed values vary less than the observed values. It may be that each of the individual points is the best under the model, but it is very unlikely that the real (but unobserved) values of `Ozone` would have had this distribution. Imputing predicted values also has an effect on the correlation. The red points have a correlation of 1 since they are located on a line. If the red and blue dots are combined, then the correlation increases from 0.35 to 0.39.

Regression imputation yields unbiased estimates of the means under MCAR, just like mean imputation, and of the regression weights of the imputation model if the explanatory variables are complete. Moreover, the regression weights are unbiased under MAR if the factors that influence the missingness are part of the regression model. In the example this corresponds to the situation where `Solar.R` would explain any differences in the probability that `Ozone` is missing. On the other hand, the variability of the imputed data is systematically underestimated. The degree of underestimation depends on the explained variance and on the proportion of missing cases (Little and Rubin, 2002, p. 64).

Imputing predicted values can yield realistic imputations if the prediction is close to perfection. If so, the method reconstructs the missing parts from the available data. In essence, there was not really any information missing

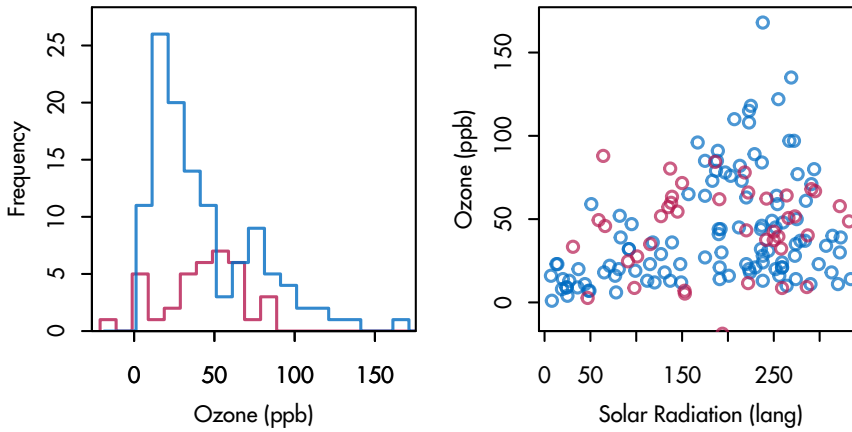

Figure 1.3: Stochastic regression imputation of `Ozone`.

in the first place, it was only coded in a different form. This type of missing data is unlikely to surface in most applications.

1.3.5 Stochastic regression imputation

Stochastic regression imputation is a refinement of regression imputation that adds noise to the predictions. This will have a downward effect on the correlation. We can impute `Ozone` by stochastic regression imputation as:

```
> imp <- mice(airquality[, 1:2], method = "norm.nob",
      m = 1, maxit = 1, seed = 1)
 iter imp variable
  1   1  Ozone  Solar.R
```

The `method="norm.nob"` argument requests a plain, non-Bayesian, stochastic regression method. This method first estimates the intercept, slope and residual variance under the linear model, then generates imputed value according to these specification. We will come back to the details in Section 3.2. The `seed` argument makes the solution reproducible. Figure 1.3 shows the results. The addition of noise to the predictions opens up the distribution of the imputed values, as intended.

Note that some new complexities arise. There are several imputations with negative values. Such values are implausible since negative `Ozone` concentrations do not exist in the real world. Also, the high end of the distribution is not well covered. The cause of this is that the relation in the observed data is somewhat heteroscedastic. The variability of `Ozone` seems to increase up to the solar radiation level of 250 langleys, and decreases after that. Though it is

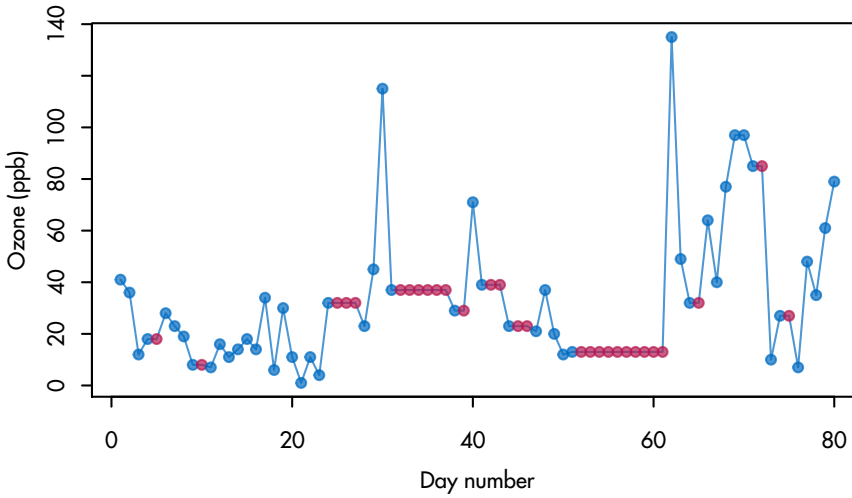

Figure 1.4: Imputation of `Ozone` by last observation carried forward (LOCF).

unclear whether this is a genuine meteorological phenomenon, the imputation model did not account for this feature.

Stochastic regression imputation is an important step forward. In particular it preserves not only the regression weights, but also the correlation between variables (cf. Exercise 3). Stochastic regression imputation does not solve all problems, and there are many subtleties that need to be addressed. However, the main idea to draw from the residuals is very powerful, and forms the basis of more advanced imputation techniques.

1.3.6 LOCF and BOFC

Last observation carried forward (LOCF) and baseline observation carried forward (BOCF) require longitudinal data. The idea is to take the last observed value as a replacement for the missing data. Figure 1.4 illustrates the method applied to the first 80 days of the `Ozone` series. The stretches of red dots indicate the imputations.

LOCF is convenient because it generates a complete dataset. The method is used in clinical trials. The U.S. Food and Drug Administration (FDA) has traditionally viewed LOCF as the preferred method of analysis, considering it conservative and less prone to selection than listwise deletion. However, Molenberghs and Kenward (2007, pp. 47–50) show that the bias can operate in both directions, and that LOCF can yield biased estimates even under MCAR. LOCF needs to be followed by a proper statistical analysis method that distinguishes between the real and imputed data. This is typically not done, however. Additional concerns about a reversal of the time direction are given in Kenward and Molenberghs (2009).

The Panel on Handling Missing Data in Clinical Trials recommends that LOCF and BOCF should not be used as the primary approach for handling missing data unless the assumptions that underlie them are scientifically justified (National Research Council, 2010, p. 77).

1.3.7 Indicator method

Suppose that we want to fit a regression, but there are missing values in one of the explanatory variables. The indicator method (Miettinen, 1985, p. 232) replaces each missing value by a zero and extends the regression model by the response indicator. The procedure is applied to each incomplete variable. The user analyzes the extended model instead of the original.

This method is popular in public health and epidemiology. An advantage is that the indicator method retains the full dataset. Also, it allows for systematic differences between the observed and the unobserved data by inclusion of the response indicator. However, the method can yield severely biased regression estimates, even under MCAR and for low amounts of missing data (Vach and Blettner, 1991; Greenland and Finkle, 1995; Knol et al., 2010).

On the other hand, White and Thompson (2005) point out that the method can be useful to estimate the treatment effect in randomized trials when a baseline covariate is partially observed. If the missing data are restricted to the covariate, if the interest is solely restricted to estimation of the treatment effect, if compliance to the allocated treatment is perfect and if the model is linear without interactions, then using the indicator method for that covariate yields an unbiased estimate of the treatment effect. This is true even if the missingness depends on the covariate itself.

The conditions under which the indicator method works are often difficult to achieve in practice. The method does not allow for missing data in the outcomes, both of which frequently occur in real data. While the indicator method may be suitable in some special cases, it falls short as a general way to treat missing data.

1.3.8 Summary

Table 1.1 provides a summary of the methods discussed in this section. The table addresses two topics: whether the method yields the correct results on average (unbiasedness), and whether it produces the correct standard error. Unbiasedness is evaluated with respect to the mean, the regression weight (of the regression with the incomplete variable as dependent) and the correlation.

The table identifies the assumptions on the missing data mechanism each method must make in order to produce unbiased estimates. Both deletion methods always require MCAR. In addition, for listwise deletion there are two MNAR special cases (cf. Section 2.6). Regression imputation and stochastic regression imputation can yield unbiased estimates under MAR. In order to

Table 1.1: Overview of assumptions made by simple methods

	Mean	Unbiased Reg Weight	Correlation	Standard Error
Listwise deletion	MCAR	MCAR	MCAR	Too large
Pairwise deletion	MCAR	MCAR	MCAR	Complicated
Mean imputation	MCAR	–	–	Too small
Regression imp	MAR	MAR	–	Too small
Stochastic imp	MAR	MAR	MAR	Too small
LOCF	–	–	–	Too small
Indicator	–	–	–	Too small

work, the model needs to be correctly specified. LOCF and the indicator method are incapable of providing consistent estimates, even under MCAR.

Listwise deletion produces standard errors that are correct for the subset of complete cases, but in general too large for the entire dataset. Calculation of standard errors under pairwise deletion is complicated. The standard errors after imputation are too small since the standard calculations make no distinction between the observed data and the imputed data. Correction factors for some situations have been developed (Schafer and Schenker, 2000), but a more convenient solution is multiple imputation.

1.4 Multiple imputation in a nutshell

1.4.1 Procedure

Multiple imputation creates $m > 1$ complete datasets. Each of these datasets is analyzed by standard analysis software. The m results are pooled into a final point estimate plus standard error by simple pooling rules ("Rubin's rules"). Figure 1.5 illustrates the three main steps in multiple imputation: imputation, analysis and pooling.

The analysis starts with observed, incomplete data. Multiple imputation creates several complete versions of the data by replacing the missing values by plausible data values. These plausible values are drawn from a distribution specifically modeled for each missing entry. Figure 1.5 portrays $m = 3$ imputed datasets. In practice, m is often taken larger (cf. Section 2.7). The number $m = 3$ is taken here just to make the point that the technique creates multiple versions of the imputed data. The three imputed datasets are identical for the observed data entries, but differ in the imputed values. The magnitude of these difference reflects our uncertainty about what value to impute.

The second step is to estimate the parameters of interest from each imputed dataset. This is typically done by applying the analytic method that we would

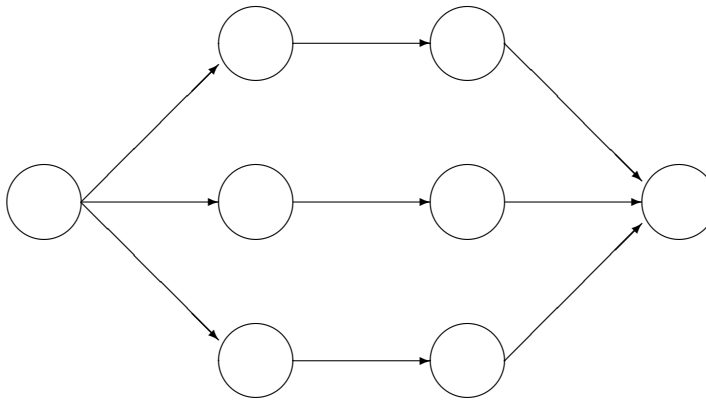

Figure 1.5: Scheme of main steps in multiple imputation.

have used had the data been complete. The results will differ because their input data differ. It is important to realize that these differences are caused only because of the uncertainty about what value to impute.

The last step is to pool the m parameter estimates into one estimate, and to estimate its variance. The variance combines the conventional sampling variance (within-imputation variance) and the extra variance caused by the missing data extra variance caused by the missing data (between-imputation variance). Under the appropriate conditions, the pooled estimates are unbiased and have the correct statistical properties.

1.4.2 Reasons to use multiple imputation

Multiple imputation (Rubin, 1987a, 1996) solves the problem of "too small" standard errors in Table 1.1. Multiple imputation is unique in the sense that it provides a mechanism for dealing with the inherent uncertainty of the imputations themselves.

Our level of confidence in a particular imputed value is expressed as the variation across the m completed datasets. For example, in a disability survey, suppose that the respondent answered the item whether he could walk, but did not provide an answer to the item whether he could get up from a chair. If the person can walk, then it is highly likely that the person will also be able to get up from the chair. Thus, for persons who can walk, we can draw a "yes" for missing "getting up from a chair" with a high probability, say 0.99, and use the drawn value as the imputed value. In the extreme, if we are really certain, we always impute the same value for that person. More generally, we

are less confident about the true value. Suppose that, in a growth study, height is missing for a subject. If we only know that this person is a woman, this provides some information about likely values, but not so much. So the range of plausible values from which we draw is much larger here. The imputations for this woman will thus vary a lot over the different datasets. Multiple imputation is able to deal with both high-confidence and low-confidence situations equally well.

Another reason to use multiple imputation is that it separates the solution of the missing data problem from the solution of the complete data problem. The missing data problem is solved first, the complete data problem next. Though these phases are not completely independent, the answer to the scientifically interesting question is not obscured anymore by the missing data. The ability to separate the two phases simplifies statistical modeling, and hence contributes to a better insight into the phenomenon of scientific study.

1.4.3 Example of multiple imputation

Continuing with the `airquality` dataset, it is straightforward to apply multiple imputation. The following code imputes the missing data five times, fits a linear regression model to predict `Ozone` in each of the imputed datasets, and pools the five sets of estimated parameters.

```
> imp <- mice(airquality, seed = 1, print = FALSE)
> fit <- with(imp, lm(Ozone ~ Wind + Temp + Solar.R))
> tab <- round(summary(pool(fit)), 3)
> tab[, c(1:3, 5)]
               est     se    t Pr(>|t|)
(Intercept) -64.31 24.614 -2.6    0.015
Wind         -3.11  0.828 -3.8    0.003
Temp          1.64  0.243  6.8    0.000
Solar.R       0.05  0.023  2.1    0.037
```

Fitting the same model to the complete cases can be done by:

```
> fit <- lm(Ozone ~ Wind + Temp + Solar.R, data = airquality,
      na.action = na.omit)
> round(coef(summary(fit)), 3)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -64.34     23.055    -2.8    0.006
Wind           -3.33      0.654    -5.1    0.000
Temp            1.65      0.254     6.5    0.000
Solar.R         0.06      0.023     2.6    0.011
```

The solutions are nearly identical here, which is due to the fact that most missing values occur in the outcome variable. The standard errors of the multiple imputation solution are slightly smaller than in the complete case analysis.

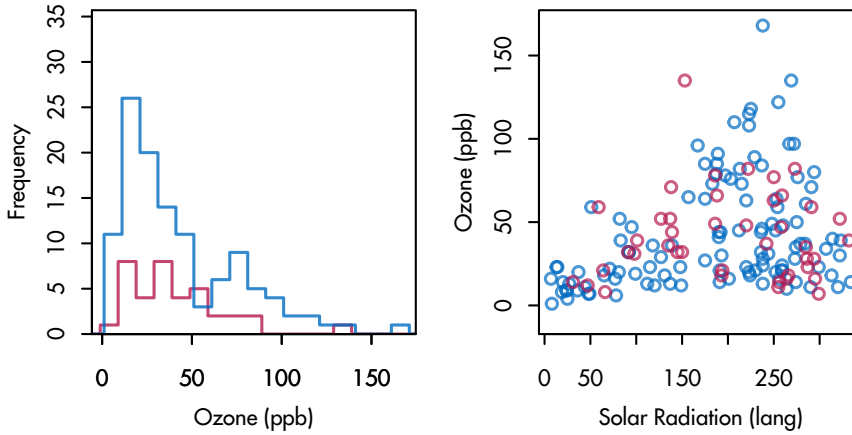

Figure 1.6: Multiple imputation of `Ozone`. Plotted are the imputed values from the first imputation.

It is often the case that multiple imputation is more efficient than complete case analysis. Depending on the data and the model at hand, the differences can be dramatic.

Figure 1.6 shows the distribution and scattergram for the observed and imputed data combined. The imputations are taken from the first completed dataset. The blue and red distributions are quite similar. Problems with the negative values as in Figure 1.3 are now gone since the imputation method used observed data as donors to fill the missing data. Section 3.4 describes the method in detail. Note that the red points respect the heteroscedastic nature of the relation between `Ozone` and `Solar.R`. All in all, the red points look as if they could have been measured if they had not been missing. The reader can easily recalculate the solution and inspect these plots for the other imputations.

Figure 1.7 plots the completed `Ozone` data. The imputed data of all five imputations are plotted for the days with missing `Ozone` scores. In order to avoid clutter, the lines that connect the dots are not drawn for the imputed values. Note that the pattern of imputed values varies substantially over the days. At the beginning of the series, the values are low and the spread is small, in particular for the cold and windy days 25–27. The small spread for days 25–27 indicates that the model is quite sure of these values. High imputed values are found around the hot and sunny days 35–42, whereas the imputations during the moderate days 52–61 are consistently in the moderate range. Note how the available information helps determine sensible imputed values that respect the relations between wind, temperature, sunshine and ozone.

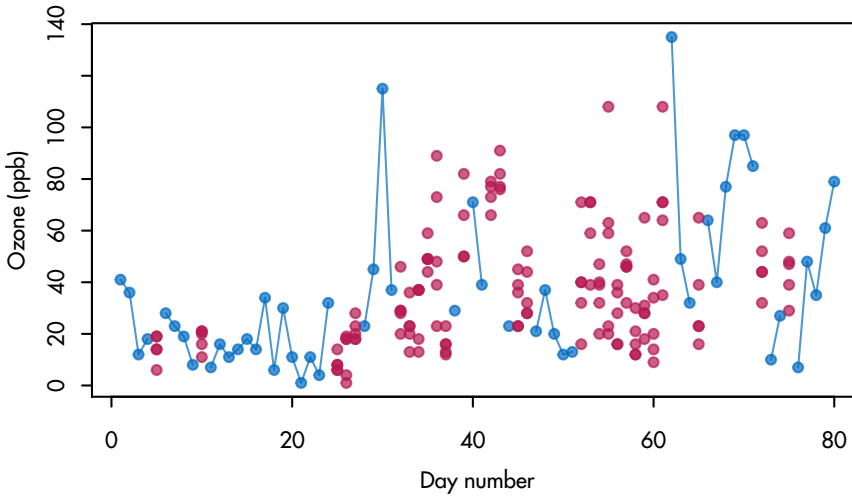

Figure 1.7: Multiple imputation of `Ozone`. Plotted are the observed values (in blue) and the multiply imputed values (in red). One red dot at (61,168) is not plotted.

1.5 Goal of the book

The main goal of this book is to add multiple imputation to the tool chest of practitioners. The text explains the ideas underlying multiple imputation, discusses when multiple imputation is useful, how to do it in practice and how to report the results of the steps taken.

The computations are done with the help of the `R` package `mice`, written by Karin Groothuis-Oudshoorn and myself (Van Buuren and Groothuis-Oudshoorn, 2011). The book thus also serves as an extended tutorial on the practical application of `mice`. Online materials that accompany the book can be found on www.multiple-imputation.com. My hope is that this hands-on approach will facilitate understanding of the key ideas in multiple imputation.

1.6 What the book does not cover

The field of missing data research is vast. This book focuses on multiple imputation. The book does not attempt cover the enormous body of literature on alternative approaches to incomplete data. This section briefly reviews three of these approaches.

1.6.1 Prevention

With the exception of McKnight et al. (2007, Chapter 4), books on missing data do not mention prevention. Yet, prevention of the missing data is the most direct attack on problems caused by the missing data. Prevention is fully in spirit with the quote of Orchard and Woodbury given on p. 5. There is a lot one could do to prevent missing data. The remainder of this section lists point-wise advice.

Minimize the use of intrusive measures, like blood samples. Visit the subject at home. Use incentives to stimulate response, and try to match up the interviewer and respondent on age and ethnicity. Adapt the mode of the study (telephone, face to face, web questionnaire, and so on) to the study population. Use a multi-mode design for different groups in your study. Quickly follow-up for people that do not respond, and where possible try to retrieve any missing data from other sources.

In experimental studies, try to minimize the treatment burden and intensity where possible. Prepare a well-thought-out flyer that explains the purpose and usefulness of your study. Try to organize data collection through an authority, e.g., the patient's own doctor. Conduct a pilot study to detect and smooth out any problems.

Economize on the number of variables collected. Only collect the information that is absolutely essential to your study. Use short forms of measurement instruments where possible. Eliminate vague or ambivalent questionnaire items. Use an attractive layout of the instruments. Refrain from using blocks of items that force the respondent to stay on a particular page for a long time. Use computerized adaptive testing where feasible. Do not allow other studies to piggy-back on your data collection efforts.

Do not overdo it. Many Internet questionnaires are annoying because they force the respondent to answer. Do not force your respondent. The result will be an apparently complete dataset with mediocre data. Respect the wish of your respondent to skip items. The end result will be more informative.

Use double coding in the data entry, and chase up any differences between the versions. Devise nonresponse forms in which you try to find out why people they did not respond, or why they dropped out.

Last but not least, consult experts. Many academic centers have departments that specialize in research methodology. Sound expert advice may turn out to be extremely valuable for keeping your missing data rate under control.

Most of this advice can be found in books on research methodology and data quality. Good books are Shadish et al. (2001), De Leeuw et al. (2008), Dillman et al. (2008) and Groves et al. (2009).

1.6.2 Weighting procedures

Weighting is a method to reduce bias when the probability to be selected in the survey differs between respondents. In sample surveys, the responders

are weighted by design weights, which are inversely proportional to their probability of being selected in the survey. If there are missing data, the complete cases are re-weighted according to design weights that are adjusted to counter any selection effects produced by nonresponse. The method is widely used in official statistics. Relevant pointers include Cochran (1977) and Särndal et al. (1992) and Bethlehem (2002).

The method is relatively simple in that only one set of weights is needed for all incomplete variables. On the other hand, it discards data by listwise deletion, and it cannot handle partial response. Expressions for the variance of regression weights or correlations tend to be complex, or do not exist. The weights are estimated from the data, but are generally treated as fixed. The implications for this are unclear (Little and Rubin, 2002, p. 53).

There has been interest recently in improved weighting procedures that are "double robust" (Scharfstein et al., 1999; Bang and Robins, 2005). This estimation method requires specification of three models: Model A is the scientifically interesting model, Model B is the response model for the outcome and model C is the joint model for the predictors and the outcome. The dual robustness property states that: if either Model B or Model C is wrong (but not both), the estimates under Model A are still consistent. This seems like a useful property, but the issue is not free of controversy (Kang and Schafer, 2007).

1.6.3 Likelihood-based approaches

Likelihood-based approaches define a model for the observed data. Since the model is specialized to the observed values, there is no need to impute missing data or to discard incomplete cases. The inferences are based on the likelihood or posterior distribution under the posited model. The parameters are estimated by maximum likelihood, the EM algorithm, the sweep operator, Newton–Raphson, Bayesian simulation and variants thereof. These methods are smart ways to skip over the missing data, and are known as direct likelihood and full information maximum likelihood (FIML).

Likelihood-based methods are, in some sense, the "royal way" to treat missing data problems. The estimated parameters nicely summarize the available information under the assumed models for the complete data and the missing data. The model assumptions can be displayed and evaluated, and in many cases it is possible to estimate the standard error of the estimates.

Multiple imputation is an extension of likelihood-based methods. It adds an extra step in which imputed data values are drawn. An advantage of this is that it is generally easier to calculate the standard errors for a wider range of parameters. Moreover, the imputed values created by multiple imputation can be inspected and analyzed, which helps us to gauge the effect of the model assumptions on the inferences.

The likelihood-based approach receives excellent treatment in the book by Little and Rubin (2002). A less technical account that should appeal to

social scientists can be found in Enders (2010, chapters 3–5). Molenberghs and Kenward (2007) provide a hands-on approach of likelihood-based methods geared toward clinical studies, including extensions to data that are MNAR.

1.7 Structure of the book

This book consists of three main parts: basics, case studies and extensions. Chapter 2 reviews the history of multiple imputation and introduces the notation and theory. Chapter 3 provides an overview of imputation methods for univariate missing data. Chapter 4 distinguishes three approaches to attack the problem of multivariate missing data. Chapter 5 discusses issues that may arise when applying multiple imputation to multivariate missing data. Chapter 6 reviews issues pertaining to the analysis of the imputed datasets.

Chapters 7–9 contain case studies of the techniques described in the previous chapters. Chapter 7 deals with "problems with the columns," while Chapter 8 addresses "problems with the rows." Chapter 9 discusses studies on problems with both rows and columns.

Chapter 10 concludes the main text with a discussion of limitations and pitfalls, reporting guidelines, alternative applications and future extensions. The appendix discusses software options for multiple imputation.

1.8 Exercises

1. *Reporting practice*. What are the reporting practices in your field? Take a random sample of articles that have appeared during the last 10 years in the leading journal in your field. Select only those that present quantitative analyses, and address the following topics:
   (a) Did the authors report that there were missing data?
   (b) If not, can you infer from the text that there must have been missing data?
   (c) Did the authors discuss how they handled the missing data?
   (d) Were the missing data properly addressed?
   (e) Can you detect a trend over time in reporting practice?
   (f) Would the editors of the journal be interested in your findings?

2. *Loss of information*. Suppose that a dataset consists of 100 cases and 10