A Handbook of Statistical Analyses Using R — 3rd Edition

Torsten Hothorn and Brian S. Everitt

CHAPTER 16

Missing Values: Lowering Blood Pressure During Surgery

16.1 Introduction

It is sometimes necessary to lower a patient's blood pressure during surgery, using a hypotensive drug. Such drugs are administered continuously during the relevant phase of the operation; because the duration of this phase varies so does the total amount of drug administered. Patients also vary in the extent to which the drugs succeed in lowering blood pressure. The sooner the blood pressure rises again to normal after the drug is discontinued, the better. The data in Table 16.1 (a missing-value version of the data presented by **?**) relate to a particular hypotensive drug and give the time in minutes before the patient's systolic blood pressure returned to 100mm of mercury (the recovery time), the logarithm (base 10) of the dose of drug in milligrams, and the average systolic blood pressure achieved while the drug was being administered. The question of interest is how is the recovery time related to the other two variables? For some patients the recovery time was not recorded and the missing values are indicated as NA in Table 16.1.

Table 16.1: `bp` data. Blood pressure data.

logdose	bloodp	recovtime	logdose	bloodp	recovtime
2.26	66	7	2.70	73	39
1.81	52	10	1.90	56	28
1.78	72	18	2.78	83	12
1.54	67	NA	2.27	67	60
2.06	69	10	1.74	84	10
1.74	71	13	2.62	68	NA
2.56	88	21	1.80	64	22
2.29	68	12	1.81	60	21
1.80	59	9	1.58	62	14
2.32	73	NA	2.41	76	4
2.04	68	20	1.65	60	27
1.88	58	31	2.24	60	26
1.18	61	23	1.70	59	NA
2.08	68	22	2.45	84	15
1.70	69	13	1.72	66	8

Table 16.1: `bp` data (continued).

logdose	bloodp	recovtime	logdose	bloodp	recovtime
1.74	55	9	2.37	68	46
1.90	67	50	2.23	65	24
1.79	67	NA	1.92	69	NA
2.11	68	11	1.99	72	25
1.72	59	8	1.99	63	45
1.74	68	NA	2.35	56	72
1.60	63	16	1.80	70	25
2.15	65	23	2.36	69	28
2.26	72	7	1.59	60	10
1.65	58	NA	2.10	51	25
1.63	69	NA	1.80	61	44
2.40	70	NA			

16.2 Analyzing Multiply Imputed Data

From the analysis of each data set we need to look at the estimates of the quantity of interest, say Q, and the variance of the estimates. We let $\hat{Q}_i$ be the estimate from the ith data set and S_i its corresponding variance. The combined estimate of the quantity of interest is

$$\bar{Q} = \frac{1}{m}\sum_{i=1}^{m}\hat{Q}_i.$$

To find the combined variance involves first calculating the within-imputation variance,

$$\bar{S} = \frac{1}{m}\sum_{i=1}^{m}S_i$$

followed by the between-imputation variance,

$$B = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{Q}_i - \bar{Q})^2$$

then the required total variance can now be found from

$$T = \bar{S} + (1 + m^{-1})B$$

This total variance is made up of two components; the first which preserves the natural variability, $\bar{S}$, is simply the average of the variance estimates for each imputed data set and is analogous to the variance that would be suitable if we did not need to account for missing data; the second component, B, estimates uncertainty caused by missing data by measuring how the point estimates

vary from data set to data set. More explanation of how the formula for T arises is given in **?**.

The overall standard error is simply the square root of T. A significance test for Q and a confidence interval is found from the usual test statistic, $(Q-$ hypothesized value of $Q)/\sqrt{T}$, the value of which is referred to a Student's t-distribution. The question arises however as to what is the appropriate value for the degrees of freedom of the test, say v_0? **?** suggests that the answer to this question is given by;

$$v_0 = (m-1)(1+1/r^2)$$

where

$$r = \frac{B + B/m}{\bar{S}}$$

But **?** noted that using this value of v_0 can produce values that are larger than the degrees of freedom in the complete data, a result which they considered 'clearly inappropriate'. Consequently they developed an adapted version that does not lead to the same problem. Barnard and Rubin's revised value for the degrees of freedom of the t-test in which we are interested is v_1 given by;

$$v_1 = \frac{v_0 v_2}{v_0 + v_2}$$

where

$$v_2 = \frac{n(n-1)(1-\lambda)}{n+2}$$

and

$$\lambda = \frac{r}{\sqrt{r^2+1}}.$$

The quantity v_1 is always less than or equal to the degrees of freedom of the test applied to the hypothetically complete data. (For more details see **?**).

16.3 Analysis Using R

To begin we shall analyze the blood pressure data in Table 16.1 using the complete-case approach, i.e., by simply removing the data for patients where the recovery time is missing. To begin we might simply count the number of missing values using the sapply function as follows:

```
R> sapply(bp, function(x) sum(is.na(x)))
```

```
  logdose    bloodp recovtime
        0         0        10
```

So there are ten missing values of recovery time but no missing values amongst the other two variables. Now we use the `summary` function to look at some basic statistics of the complete data for recovery time:

```
R> summary(bp$recovtime, na.rm = TRUE)
```

```
R> layout(matrix(1:3, nrow = 1))
R> plot(bloodp ~ logdose, data = bp)
R> plot(recovtime ~ bloodp, data = bp)
R> plot(recovtime ~ logdose, data = bp)
```

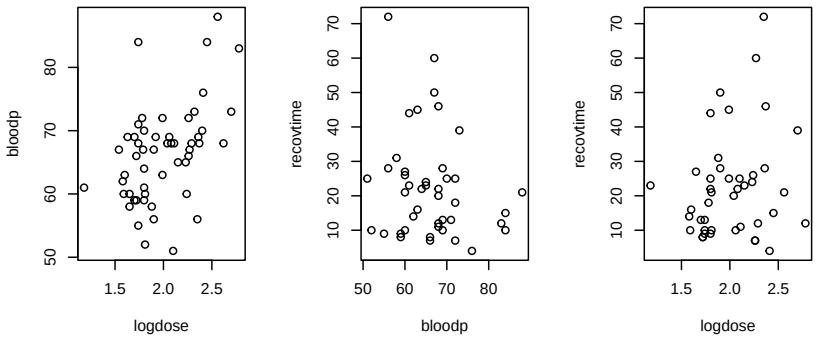

Figure 16.1 Scatterplots of the complete cases of the `bp` data.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    4.0    10.5    21.0    22.4    26.5    72.0      10
```

And next we can calculate the complete data estimate of the standard deviation of recover time

```
R> sd(bp$recovtime, na.rm = TRUE)
```

```
[1] 15.1
```

The final numerical results we might be interested in are the correlations of recovery time with blood pressure and of recovery time with logdose. These can be found as follows:

```
R> with(bp, cor(bloodp, recovtime, use = "complete.obs"))
```

```
[1] -0.189
```

```
R> with(bp, cor(logdose, recovtime, use = "complete.obs"))
```

```
[1] 0.21
```

And a useful graphic of the data is a scatterplot matrix which we can construct using `pairs`. The scatterplot matrix is given in Figure 16.1.

To investigate how recovery time is related to blood pressure and logdose we might begin by fitting a multiple linear regression model (see Chapter **??**). The relevant command and the summary of the results is shown in Figure 16.2. Note that this summary output reports that ten observations with missing values were removed prior to the analysis; this is default for many models in R.

Now let us see what happens when we impute the missing values of the

```
R> summary(lm(recovtime ~ bloodp + logdose, data = bp))

Call:
lm(formula = recovtime ~ bloodp + logdose, data = bp)

Residuals:
   Min     1Q Median     3Q    Max
-20.06 -10.49  -1.77   5.92  36.46

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   32.310     18.414    1.75    0.087
bloodp        -0.688      0.301   -2.28    0.028
logdose       17.773      7.497    2.37    0.023

Residual standard error: 14.2 on 40 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.154,         Adjusted R-squared:  0.112
F-statistic: 3.65 on 2 and 40 DF,  p-value: 0.0349
```

Figure 16.2 R output of the complete-case linear model for the `bp` data.

recovery time variable simply by the mean of the complete case; for this we will use the **mice** (**?**) package;

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

We begin by creating a new data set, `imp`, which will contain the three variables log-dose, blood pressure, and recovery time with the missing values in the latter replaced by the mean recovery time of the complete cases;

```
R> imp <- mice(bp, method = "mean", m = 1, maxit = 1)

 iter imp variable
  1   1  recovtime
```

So now we can find the summary statistics of recovery time to compare with those given previously

```
R> with(imp, summary(recovtime))

call :
with.mids(data = imp, expr = summary(recovtime))

call1 :
mice(data = bp, m = 1, method = "mean", maxit = 1)

nmis :
  logdose    bloodp recovtime
        0         0        10

analyses :
[[1]]
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    4.0    12.0    22.4    22.4    25.0    72.0
```

Making the comparison we see that only the values of the first and third

quantile and the median have changed. The minimum and maximum values are the same and so, of course, is the mean. But of more interest is what happens to the sample standard deviation; its value for the imputed data can be found using:

```
R> with(imp, sd(recovtime))
```

```
call :
with.mids(data = imp, expr = sd(recovtime))

call1 :
mice(data = bp, m = 1, method = "mean", maxit = 1)

nmis :
  logdose    bloodp recovtime
        0         0        10

analyses :
[[1]]
[1] 13.6
```

The value for the imputed data, 13.56 is, as we would expect, lower than that for the complete data, 15.09. What about the correlations?

```
R> with(imp, cor(bloodp, recovtime))
```

```
call :
with.mids(data = imp, expr = cor(bloodp, recovtime))

call1 :
mice(data = bp, m = 1, method = "mean", maxit = 1)

nmis :
  logdose    bloodp recovtime
        0         0        10

analyses :
[[1]]
[1] -0.183
```

```
R> with(imp, cor(logdose, recovtime))
```

```
call :
with.mids(data = imp, expr = cor(logdose, recovtime))

call1 :
mice(data = bp, m = 1, method = "mean", maxit = 1)

nmis :
  logdose    bloodp recovtime
        0         0        10

analyses :
```

```
R> layout(matrix(1:2, nrow = 1))
R> plot(recovtime ~ bloodp, data = complete(imp),
+       pch = is.na(bp$recovtime) + 1)
R> plot(recovtime ~ logdose, data = complete(imp),
+       pch = is.na(bp$recovtime) + 1)
R> legend("topleft", pch = 1:2, bty = "n",
+         legend = c("original", "imputed"))
```

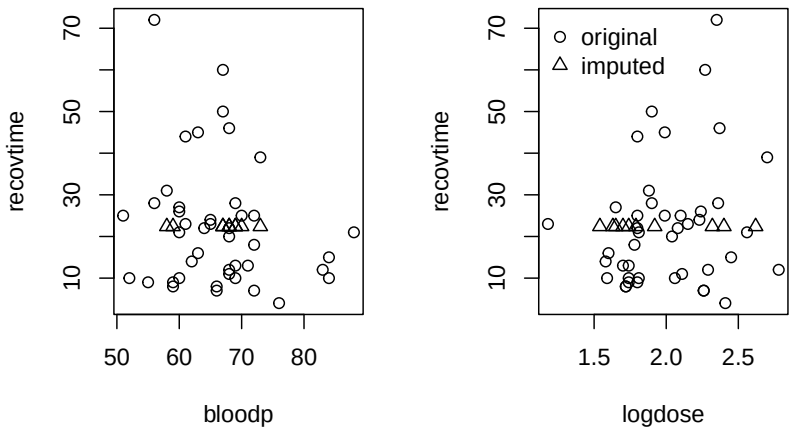

Figure 16.3 Scatterplots of the imputed `bp` data. Imputed observations are depicted as triangles.

```
[[1]]
[1] 0.186
```

The correlations of blood pression and recovery time are very similar before (−0.19) after (−0.18) imputation. For log-dose, imputation changes the correlation from 0.21 to 0.19.

The scatterplot of the imputed data is found as given by the code displayed with Figure 16.3. For mean imputation, the imputed value of the recovery time is constant for all observations and so they appear as a series of points along the value of the mean value of the observed recovery times namely, 22.4.

Comparison of the multiple linear regression results in Figure 16.4 with those in Figure 16.2 show some interesting differences, for example, the standard errors of the regression coefficients are somewhat lower for the mean imputed data but the conclusions drawn from the results in each table would be broadly similar.

The single imputation of a sample mean is not to be recommended and so

```
R> with(imp, summary(lm(recovtime ~ bloodp + logdose)))
call :
with.mids(data = imp, expr = summary(lm(recovtime ~ bloodp +
    logdose)))

call1 :
mice(data = bp, m = 1, method = "mean", maxit = 1)

nmis :
  logdose    bloodp recovtime
        0         0        10

analyses :
[[1]]

Call:
lm(formula = recovtime ~ bloodp + logdose)

Residuals:
   Min     1Q Median     3Q    Max
-19.31  -8.19  -0.60   5.11  38.38

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   35.004     15.915    2.20    0.032
bloodp        -0.606      0.262   -2.31    0.025
logdose       13.864      5.960    2.33    0.024

Residual standard error: 12.9 on 50 degrees of freedom
Multiple R-squared:  0.128,        Adjusted R-squared:  0.0928
F-statistic: 3.66 on 2 and 50 DF,  p-value: 0.0328
```

Figure 16.4 R output of the mean imputation linear model for the `bp` data.

we will move on to using a more sophisticated multiple imputation procedure know as *predictive mean matching*. The method is described in detail in **?** who considers it both easy-to-use and versatile. And imputations outside the observed data range will not occur so that problems with meaningless imputations, for example, a negative recovery time, will not occur. The method is labeled `pmm` in the **mice** package and here we will apply it to the blood pressure data with $m = 10$ (we need to fix the seed in order to make the result reproducible):

```
R> imp_ppm <- mice(bp, m = 10, method = "pmm",
+                  print = FALSE, seed = 1)
```

The scatterplot of the imputed data is found as given by the code displayed with Figure 16.5. We only show the imputed recovery times from the first iteration ($m = 1$).The imputed recovery times now take different values.

From the resulting object we can compute the mean and standard deviations of recovery time for each of the $m = 10$ iterations. We first extract these numbers from the `analyses` element of the returned object, convert this list to a vector, and use the `summary` function to compute the usual summary statistics:

```
R> summary(unlist(with(imp_ppm, mean(recovtime))$analyses))
```

```
R> layout(matrix(1:2, nrow = 1))
R> plot(recovtime ~ bloodp, data = complete(imp_ppm),
+       pch = is.na(bp$recovtime) + 1)
R> plot(recovtime ~ logdose, data = complete(imp_ppm),
+       pch = is.na(bp$recovtime) + 1)
R> legend("topleft", pch = 1:2, bty = "n",
+         legend = c("original", "imputed"))
```

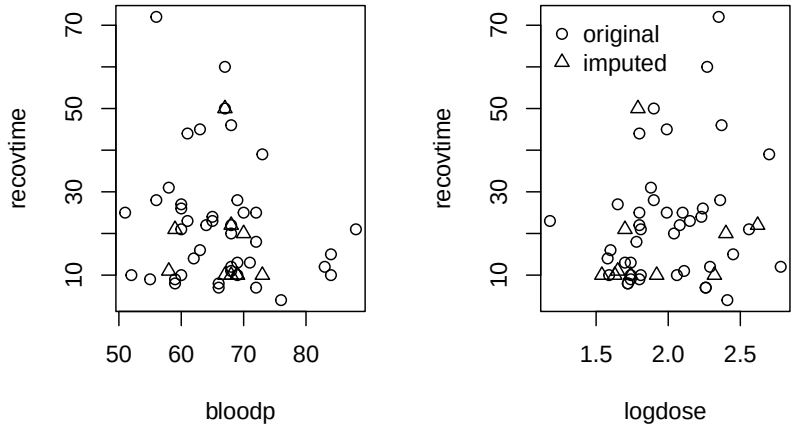

Figure 16.5 Scatterplots of the multiple imputed `bp` data (first iteration). Imputed observations are depicted as triangles.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   20.8    21.5    21.9    21.9    22.3    23.4
```

```
R> summary(unlist(with(imp_ppm, sd(recovtime))$analyses))
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   13.9    14.1    14.3    14.5    14.7    15.7
```

We do the same with the correlations as follows

```
R> summary(unlist(with(imp_ppm,
+         cor(bloodp, recovtime))$analyses))
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -0.233  -0.179  -0.176  -0.172  -0.161  -0.102
```

```
R> summary(unlist(with(imp_ppm,
+         cor(logdose, recovtime))$analyses))
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.118   0.189   0.229   0.227   0.264   0.311
```

The estimate of the mean of the blood pressure data from the multiply imputed results is 21.95, very similar to the values found previously. Similarly the estimate of the standard deviation of the data is 14.47 which lies between the complete data estimate and the *mean-imputed* value. The two correlation estimates are also very close to the previous values. The variation in the estimates of mean, standard deviation, and correlations across the ten imputation is relatively small apart from that for the correlation between log-dose and recovery time – here there is considerable variation in the values for the ten imputations.

Finally, we will fit a linear model to each of the imputed samples and then find the summary statistics for the ten sets of regression coefficients: the results are given in Figure 16.6:

```
R> fit <- with(imp_ppm, lm(recovtime ~ bloodp + logdose))
```

```
R> summary(pool(fit))
         term estimate std.error statistic   df p.value
1 (Intercept)   32.851     17.57      1.87 41.9  0.0685
2      bloodp   -0.668      0.28     -2.38 46.3  0.0213
3     logdose   16.773      6.91      2.43 34.7  0.0205
```

Figure 16.6 R output of the multiple imputed linear model for the `bp` data.

The result for blood pressure is similar to the previous complete data and mean-imputed results with the regression coefficient for this variable being highly significant NA But the result for log dose differs from those found previously; for the multiply imputed data the regression coefficient for log dose is not significant at the 5% level NA whereas in both of the previous two analyses it was significant. This finding reflects the greater variation of the value of the correlation between log dose and recovery time in the ten imputations noted above. (Remember that the standard errors in Figure 16.6 computed by `pool` arise from the formulae given in Section 16.2.)

Now suppose we wish to test the hypothesis that in the population from which the sample data in Table 16.1 arises a mean recovery time of 27 minutes. We will test this hypothesis in the usual way using Student's t-test applied to the complete-data, the singly imputed data, and the multiply imputed data:

```
R> with(bp, t.test(recovtime, mu = 27))
        One Sample t-test

data:  recovtime
t = -2, df = 42, p-value = 0.05
alternative hypothesis: true mean is not equal to 27
95 percent confidence interval:
 17.8 27.0
sample estimates:
mean of x
     22.4
```

```
R> with(imp, t.test(recovtime, mu = 27))$analyses[[1]]
```

```
        One Sample t-test

data:  recovtime
t = -2, df = 52, p-value = 0.02
alternative hypothesis: true mean is not equal to 27
95 percent confidence interval:
 18.7 26.1
sample estimates:
mean of x
     22.4
```

For the multiply imputed data we need to use the `lm` function to get the equivalent of the t-test by modeling recovery time minus 27 with an intercept only and testing for zero intercept. So the code needed is:

```
R> fit <- with(imp_ppm, lm(I(recovtime - 27) ~ 1))
R> summary(pool(fit))
```

```
         term estimate std.error statistic   df p.value
1 (Intercept)    -5.05      2.14     -2.36 39.9   0.023
```

Looking at the results of the three analyses we see that the complete-case analysis fails to reject the hypothesis at the 5% level whereas the other two analyses lead to results that are statistically significant at the level. This simple (and perhaps rather artificial) example demonstrates that different conclusions can be reached by the different approaches.

16.4 Summary of Findings

The estimated standard deviation of the blood pressure is lower when computed from the mean-imputed data than from the complete data. The corresponding value from the multiply imputed data lies between these two values.

The estimate of the mean from the multiply imputed data is very similar to the value obtained in the complete data analysis. (The value from the singly imputed data is, of course, the same as from the complete data.)

The estimates of the correlations between blood pressure and recovery time and log dose and recovery time are very similar in all three analyses but the variation in the latter across the ten multiple imputations is considerable and this results in the regression coefficient for log dose being less significant than in the other two analyses.

Testing the hypothesis that the population mean of recovery time is 27 minutes using complete-case analysis leads to a different conclusion than is arrived at by the two multiple imputations approaches.

16.5 Final Comments

Missing values are an ever-present possibility in all types of studies although everything possible should be done to avoid them. But when data contain

missing values multiple imputation can be used to provide valid inferences for parameter estimates from the incomplete data. If carefully handled, multiple imputation can cope with missing data in all types of variables. In this chapter we have given only a brief account of dealing with missing values; a detailed account is available in the issue of *Statistical Methods in Medical Research entitled Multiple Imputation: Current Perspectives* (Volume 16, Number 3, 2007) and in **?**.

Exercises

Ex. 16.1 The data in Table 16.2 give the lowest temperatures (in Fahrenheit) recorded in various months for cities in the US; missing values are indicated by NA. Calculate the correlation matrix of the data using

1. the complete-case approach,
2. the available-data approach, and
3. a multiple-imputation approach.

Find the principal components of the data using each of three correlation matrices and plot the cities in the space of the first two components of each solution.

Table 16.2: `UStemp` data. Lowest temperatures in Fahrenheit recorded in various months for cities in the US.

	January	April	July	October
Atlanta	-8	26	53	28
Baltimore	-7	20	NA	25
Bismark	-44	-12	35	5
Boston	-12	16	54	28
Chicago	-27	7	40	17
Dallas	4	NA	59	29
Denver	-25	-2	43	3
ElPaso	-8	23	57	NA
Honolulu	53	57	67	NA
Houston	12	31	62	33
Juneau	-22	6	36	11
LosAngeles	23	39	49	NA
Miami	30	46	69	51
Nashville	-17	23	51	26
NewYork	-6	12	52	28
Omaha	-23	5	44	13
Phoenix	NA	32	61	34
Portland	-26	8	40	15
Reno	-16	NA	33	8
SanFrancisco	24	31	43	NA

Table 16.2: `UStemp` data (continued).

	January	April	July	October
Seattle	NA	29	43	28
Washington	-5	24	55	29

Ex. 16.2 Find 95% confidence intervals for the population means of the lowest temperature in each month using

1. the complete-case approach,
2. the mean value imputation, and
3. a multiple-imputation approach.

Ex. 16.3 Find the correlation matrix for the four months in Table 16.2 using complete-case analysis, listwise deletion, and multiple imputation.