Advanced Methods

Handling missing data

Magdalene Silberberger

11/5/2021

Contents

Introduction

By default, most of the regression models in R work with the complete cases of the data. This means that they exclude the cases in which there is at least one NA.

Illustration: Air Quality in NYC

We will use the air quality data set from base R.

```
data(airquality)
head(airquality)
```

```
##   Ozone Solar.R Wind Temp Month Day
## 1    41     190  7.4   67     5   1
## 2    36     118  8.0   72     5   2
## 3    12     149 12.6   74     5   3
## 4    18     313 11.5   62     5   4
## 5    NA      NA 14.3   56     5   5
## 6    28      NA 14.9   66     5   6
```

Summarize the data: Which variables have missing observations?

```
summary(airquality)
```

```
##      Ozone           Solar.R           Wind             Temp
##  Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00
##  1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:72.00
##  Median : 31.50   Median :205.0   Median : 9.700   Median :79.00
##  Mean   : 42.13   Mean   :185.9   Mean   : 9.958   Mean   :77.88
##  3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00
##  Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00
##  NA's   :37       NA's   :7
##      Month            Day
```

```
##  Min.   :5.000   Min.   : 1.0
##  1st Qu.:6.000   1st Qu.: 8.0
##  Median :7.000   Median :16.0
##  Mean   :6.993   Mean   :15.8
##  3rd Qu.:8.000   3rd Qu.:23.0
##  Max.   :9.000   Max.   :31.0
##
```

Let's check how many observations are actually complete:

```
comp <- complete.cases(airquality)
mean(comp)
```

```
## [1] 0.7254902
```

Only 72.55% of cases are actually complete.

Now let's add more NAs.

```
set.seed(123456)
airquality$Solar.R[runif(nrow(airquality)) < 0.7] <- NA
airquality$Day[runif(nrow(airquality)) < 0.1] <- NA
```

```
comp_na <- complete.cases(airquality)
mean(comp_na)
```

```
## [1] 0.1568627
```

Now only 15.67% of cases are complete.

Let's look at those:

```
head(airquality[comp_na, ])
```

```
##    Ozone Solar.R Wind Temp Month Day
## 1     41     190  7.4   67     5   1
## 2     36     118  8.0   72     5   2
## 9      8      19 20.1   61     5   9
## 13    11     290  9.2   66     5  13
## 14    14     274 10.9   68     5  14
## 15    18      65 13.2   58     5  15
```

Task

1. Run a linear model that looks at the determinants of ozone in NYC.
2. Identify how many observations are actually missing.
3. Drop the variable that has the most missing observations and re-run the regression.
4. Compare the results.

What to do with missing variables?

There is no "one size fits all" solution for missing data. It depends on the missing data pattern (at random or a systematic lack of data) and the approach to impute the data (parametric, nonparametric, Bayesian, etc). There are three straightforward ways to begin with:

1. Use complete cases only, i.e. restrict the analysis to the set of fully-observed observations. The advantage of this solution is that it can be implemented very easily by using the complete.cases or na.exclude functions. The cost of this approach is a substantial loss of data and therefore the precision of the

estimators will be lower. In addition, it may lead to a biased representation of the original data (if the missing process is associated with the values of the response or predictors).

2. Remove predictors with many missing data. This simple solution is useful in case most of the missing data is concentrated in one predictor.
3. Use imputation for the missing values. The idea is to replace the missing observations on the response or the predictors with artificial values that try to preserve the dataset structure: When the response is missing, we can use a predictive model to predict the missing response (possibly using sample means), then create a new fully-observed dataset containing the predictions instead of the missing values, and finally re-estimate the predictive model in this expanded dataset. However, be aware that you are in a way messing with the original data and be careful of too much manipulation when using imputation for missing values.