Introduction to Data Science: Handling Missing Data

Héctor Corrada Bravo

University of Maryland, College Park, USA

2020-03-23

Handling Missing Data

We can now move on to a very important aspect of data preparation and transformation: how to deal with missing data?

Values that are unrecorded, unknown or unspecified in a dataset.

Handling Missing Data

```
## # A tibble: 22 x 35
##    id      year month element    d1    d2    d3    d4    d5    d6    d7
##    <chr>  <dbl> <dbl> <chr>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1 MX17…   2010     1 tmax       NA  NA    NA       NA  NA      NA    NA
##  2 MX17…   2010     1 tmin       NA  NA    NA       NA  NA      NA    NA
##  3 MX17…   2010     2 tmax       NA  27.3  24.1     NA  NA      NA    NA
##  4 MX17…   2010     2 tmin       NA  14.4  14.4     NA  NA      NA    NA
##  5 MX17…   2010     3 tmax       NA  NA    NA       NA  32.1    NA    NA
##  6 MX17…   2010     3 tmin       NA  NA    NA       NA  14.2    NA    NA
##  7 MX17…   2010     4 tmax       NA  NA    NA       NA  NA      NA    NA
##  8 MX17…   2010     4 tmin       NA  NA    NA       NA  NA      NA    NA
##  9 MX17…   2010     5 tmax       NA  NA    NA       NA  NA      NA    NA
```

Handling Missing Data

Temperature observations coded as `NA` are considered *missing*.

(a) measurement failed in a specific day for a specific weather station, or

(b) certain stations only measure temperatures on certain days of the month, or

(c) measurement fails if the temperature is too high or too low

Handling Missing Data

Temperature observations coded as `NA` are considered *missing*.

(a) measurement failed in a specific day for a specific weather station, or

(b) certain stations only measure temperatures on certain days of the month, or

(c) measurement fails if the temperature is too high or too low

Knowing which of these applies can change how we approach this missing data.

Handling Missing Data

Treatment of missing data depends highly on how the data was obtained,

The more you know about a dataset, the better decision you can make.

Handling Missing Data

Central question with missing data is:

Should we *remove* observations with missing values, or should we *impute* missing values?

Handling Missing Data

Central question with missing data is:

Should we *remove* observations with missing values, or should we *impute* missing values?

In fact, can we do anything with a dataset that has missing data?

Handling Missing Data

Central question with missing data is:

Should we *remove* observations with missing values, or should we *impute* missing values?

In fact, can we do anything with a dataset that has missing data?

Answering this requires us to think **why** the data is missing.

Mechanisms of missing data

Some preliminaries

Let's assume we have the following attributes:

- y that contains missing data, (e.g., temperature measurement)
- a binary attribute r that encodes if observation in y is missing (this is not in our example dataset),
- other attributes x in our dataset (day, month, etc.)

Mechanisms of missing data

We will make statements like *depend* or *not depend*,

e.g., value of r_i does not depend on value of y_i.

Mechanisms of missing data

We will make statements like *depend* or *not depend*,

e.g., value of r_i does not depend on value of y_i.

i.e., the fact that a value is missing $(r_i = 1)$ *does not depend* on (missing) temperature value (y_i).

Mechanisms of missing data

We will make statements like *depend* or *not depend*,

e.g., value of r_i does not depend on value of y_i.

i.e., the fact that a value is missing $(r_i = 1)$ *does not depend* on (missing) temperature value (y_i).

For now:

properties of the distribution of r *do not change based on values of* y.

Mechanisms of missing data

Missing completely at random (MCAR)

Def: Missingness r_i does not depend on the (unobserved) value y_i or on observed values x_i.

Weather ex. (a): stations failed for no discernible reason.

Mechanisms of missing data

Missing completely at random (MCAR)

Def: Missingness r_i does not depend on the (unobserved) value y_i or on observed values x_i.

Weather ex. (a): stations failed for no discernible reason.

Removal: Entities with missing data can be removed from the analysis safely.

Imputation: Go for it (but see later)

Mechanisms of missing data

Missing at random (MAR)

Def: missingness r_i does not depend on the value of y_i, but may depend on the value of x_i.

Weather ex. (b): measurements are not taken on specific days of the month (where "day of the month" serves the role of x).

Mechanisms of missing data

Missing at random (MAR)

Def: missingness r_i does not depend on the value of y_i, but may depend on the value of x_i.

Weather ex. (b): measurements are not taken on specific days of the month (where "day of the month" serves the role of x).

Removal: No!, it will bias analysis since you would drop values of x based on missingness and potentially change the distribution of x.

Imputation: Go for it (but see later)

Mechanisms of missing data

Not missing at random (NMAR)

Def: missingness r_i depends on y_i.

Weather ex. (c): measurements fail when the temperature is too hot or cold.

Mechanisms of missing data

Not missing at random (NMAR)

Def: missingness r_i depends on y_i.

Weather ex. (c): measurements fail when the temperature is too hot or cold.

The worst case! Usually means that we want to go back to our collaborator and tell them that we are in a bind.

Removal: No.

Imputation No.

Mechanisms of missing data

Summary

The **first step** when dealing with missing data is to understand *why* and *how* data may be missing.

I.e., talk to collaborator, or person who created the dataset.

Handling missing data

Removing missing data

(MCAR) Not a lot of entities with missing data:

```
tidy_weather_nomissing <-
  tidy_weather %>%
    tidyr::drop_na(tmax, tmin)
```

```
## # A tibble: 33 x 6
##    id       year month day    tmax  tmin
##    <chr>   <dbl> <dbl> <chr> <dbl> <dbl>
##  1 MX17004  2010     1 d30    27.8  14.5
##  2 MX17004  2010     2 d11    29.7  13.4
##  3 MX17004  2010     2 d2     27.3  14.4
##  4 MX17004  2010     2 d23    29.9  10.7
##  5 MX17004  2010     2 d3     24.1  14.4
```

Handling missing data

Encoding as missing

(MCAR or MAR) For categorical attributes: encode the fact that a value is missing as a new category and use in subsequent modeling.

```
## # A tibble: 4 x 6
##   iso2     year sex   age       n iso2_missing
##   <chr>   <dbl> <chr> <chr> <dbl> <lgl>
## 1 missing  1985 m     04       NA TRUE
## 2 missing  1986 m     04       NA TRUE
## 3 AD       1989 m     04       NA FALSE
## 4 AD       1990 m     04       NA FALSE
```

Handling missing data

Imputation (MCAR)

(Also for MAR but not ideal) Numeric values, replace missing values of y with, e.g., the mean of non-missing y

```r
library(nycflights13)

flights %>%

  tidyr::replace_na(list(dep_delay=mean(.$dep_delay, na.rm=TRUE)))
```

Categorical attributes, replace missing y with most common category in non-missing y.

Handling missing data

Imputation (MAR)

Replace missing y predicting from other variables x (we will see linear regression using the `lm` and `predict` functions later on)

```
dep_delay_fit <- flights %>% lm(dep_delay~origin, data=.)

flights %>%

  modelr::add_predictions(dep_delay_fit, var="pred_delay") %>%

  mutate(dep_delay_fixed =

           ifelse(!is.na(dep_delay), dep_delay, pred_delay))
```

(categorical, use logistic regression)

Handling missing data

Imputation

After imputation it is useful to add an additional indicator attribute stating if a missing value was imputed

```
flights %>%
  mutate(dep_delay_missing = is.na(dep_delay))
```

Implications of imputation

Imputing missing values as discussed has two effects.

Central tendency of data is retained

If we impute missing data using the mean of a numeric variable, the mean after imputation will not change.

This is a good reason to impute based on estimates of central tendency.

Implications of imputation

The spread of the data will change

After imputation, the spread of the data will be smaller relative to spread if we ignore missing values.

This could be problematic as underestimating the spread of data can yield over-confident inferences in downstream analysis.