DEALING WITH MISSING DATA IN R

MCARDLE GC–3 WORKSHOP

Yichi Zhang

AGENDA

01 Assess Missingness

02 Visualize Missingness

03 Handle Missingness

04 Exercise

PROBLEM

Have you encountered these errors?

```r
mean(linelist$age)
```

```
[1] NA
```

PROBLEM

What about this?

```{r}
lm(bmi ~ age + gender + wt_kg + ht_cm, data = linelist, na.action = na.pass)
```

```
Error in lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :
 NA/NaN/Inf in 'x'
```

Show Traceback

ASSESS MISSINGNESS

- The number of NAs in the dataset by columns

```
> colSums(is.na(linelist))
             case_id           generation       date_infection           date_onset
                   0                    0                 2087                  256
date_hospitalisation        date_outcome              outcome               gender
                   0                  936                 1323                  278
                 age             age_unit           age_years              age_cat
                  86                    0                   86                   86
```

ASSESS MISSINGNESS

- The proportion of missing values in each variable

```
> linelist %>%
+     # check each variable's missing values
+     map(is.na) %>%
+     # calculate the total sum of missing values in each variable
+     map(sum) %>%
+     # pick the sum of missing values in each variable and divide by the sample size
+     map(~ . /nrow(linelist))%>%
+     # bind multiple columns together
+     bind_cols()
# A tibble: 1 × 30
  case_id generation date_inf…¹ date_…² date_…³ date_…⁴ outcome gender    age age_u…⁵ age_y…⁶ age_cat
    <dbl>      <dbl>      <dbl>   <dbl>   <dbl>   <dbl>   <dbl>  <dbl>  <dbl>   <dbl>   <dbl>   <dbl>
1       0          0      0.354  0.0435       0   0.159   0.225 0.0472 0.0146       0  0.0146  0.0146
```

ASSESS MISSINGNESS

- Percent of all data frame values that are missing

```
> pct_miss(linelist)
[1] 6.688745
```

- Percent of rows with any value missing

```
> pct_miss_case(linelist)
[1] 69.12364
```

- Percent of rows that are complete

```
> pct_complete_case(linelist)
[1] 30.87636
```

TYPES OF MISSING DATA

1. Missing completely at random (MCAR)
    a. The probability of being missing is the same for all cases
    b. Example: a weighing scale that ran out of batteries
    c. Usually will not bias the result
2. Missing at random (MAR)
    a. The probability of being missing is systematically related to the observed but not the unobserved data
    b. Example: a weighing scale produces more missing values on a soft surface vs a hard surface
    c. Bias can be properly corrected

TYPES OF MISSING DATA

3. Missing not at random (MNAR)
    a. The probability of being missing is systematically related to the unobserved data
    b. Example: a weighing scale may wear out over time
    c. Bias usually cannot be corrected

ADDRESS MISINGNESS

1. Listwise deletion (complete cases analysis)
    a. Eliminates all cases with one or more missing values
    b. Assumes MCAR
    c. Disadvantage: data loss
2. Pairwise deletion (available cases analysis)
    a. Includes the observations with missing values, removed them when the analysis involves pairs of values that are missing
    b. Assumes MCAR + numerical data follow an approximately normal distribution

ADDRESS MISINGNESS

3. Mean Imputation
    - Replace missing data by the mean
    - Should be **avoided** in general
    - Disadvantage: **Underestimate** the variance, **distort** the relations between variables, and **biased** estimates when data are not MCAR
4. Regression Imputation
    - Build a model from the observed data and use predictions for incomplete cases to replace the missing data
    - Should be **avoided**
    - Disadvantage: **artificially strengthen** the relations in the data; **underestimate** the variability; might lead to **spurious** relations

ADDRESS MISINGNESS

	Unbiased			Standard Error
	Mean	**Reg Weight**	**Correlation**	
Listwise deletion	MCAR	MCAR	MCAR	Too large
Pairwise deletion	MCAR	MCAR	MCAR	Complicate
Mean	MCAR	-	-	Too small
Regression	MAR	MAR	-	Too small

SEE BUUREN (2021) FOR FURTHER DETAIL

ADDRESS MISINGNESS

5. Multiple Imputation
    - Create multiple datasets with the missing values imputed to plausible data values
    - Apply the statistical model for each of the imputed datasets and pool the results
    - Advantage:
        - reduce bias in MCAR and MAR
        - more accurate standard error estimates

RESOURCES

- Multiple Imputation
  - Flexible Imputation of Missing data https://stefvanbuuren.name/fimd/
  - Multilevel
    https://www.gerkovink.com/miceVignettes/Multi_level/Multi_level_data.html
- Tutorial
  - https://epirhandbook.com/en/missing-data.html
- Case study
  - https://rpubs.com/SmilodonCub/714097
- Visualizations
  - https://cran.r-project.org/web/packages/naniar/vignettes/naniar-visualisation.html

THANKS FOR COMING

McArdle GC3-Workshop