Missing data

Prof. Maria Tackett

Click for PDF of slides

What missing data looks like

```
##       age  bmi  hyp chl
## 1   20-39   NA <NA>  NA
## 2   40-59 22.7   no 187
## 3   20-39   NA   no 187
## 4   60-99   NA <NA>  NA
## 5   20-39 20.4   no 113
## 6   60-99   NA <NA> 184
## 7   20-39 22.5   no 118
## 8   20-39 30.1   no 187
## 9   40-59 22.0   no 238
## 10 40-59   NA <NA>  NA
```

Why is missing data an issue?

Do you have missingness in your data for the final project?

Why is missing data an issue?

Do you have missingness in your data for the final project?

Why is missing data an issue when doing an analysis?

Why is missing data an issue?

Do you have missingness in your data for the final project?

Why is missing data an issue when doing an analysis?

It is important to understand missingness in the data, because there can be large implications when you fit the model and use it to make decisions.

Example: The U.S. Census

Dealing with missingness

- Deal with missingness before doing any analysis
    - This is one of the many reasons exploratory data analysis is an important first step!
- Some things to consider if you find missing values:
    - Why are the values missing?
    - Is there a pattern of missingness? If so, what is it?
    - What is the proportion of missing values?
- The answers to these questions will help you determine how to deal with the missing data

Types of missingness

- **Missing Completely at Random (MCAR)**: Missingness does not depend on the observed data or missing data, i.e. the probability of missing is the same for each observation
  - Example: People used a die to decide whether to share their income on a survey

Types of missingness

- **Missing Completely at Random (MCAR)**: Missingness does not depend on the observed data or missing data, i.e. the probability of missing is the same for each observation
  - Example: People used a die to decide whether to share their income on a survey
- **Missing at Random (MAR)**: Missingness depends on other observed variables but is random after conditioning on those variables, i.e. the probability that a variable is missing only depends on available information
  - Example: People with a college degree are less likely to share income than people without college degree

Types of missingness

- **Missing Not at Random (MNAR)**: Missingness depends on the variable itself
  - Example: People with higher incomes are less likely to share their income on a survey

How to deal with missing data?

1. Only use observations with no missingness (complete-case analysis)

2. Only use variables with no missingness

3. Impute the missing values

Complete-case analysis

Use only complete observations in the analysis, i.e. those that have a value for each variable

What are potential disadvantages of dealing with missing data this way?

Complete-case analysis

- This may be OK if there are very few observations with missing values
- R does this automatically in its regression functions

Complete-case analysis

Potential problems:

- Could result in a model being built on very few observations
    - This is especially true if there are many variables included in the model
    - Standard errors of model coefficients increase since you're losing information from the partially complete data
- If the observations with missingness differ systematically from the complete observations, then resulting analysis could be biased
    - This is especially true if the missingness is not random

Single Imputation

Single Imputation: Replace each missing value with a single number/category

- Mean imputation
- Use information from related observations
- Indicator variable for missingness
- Logical rule

Mean Imputation

- Replace missing values of a variable with the mean calculated from the observed data
- **Advantage:** Easy and straightforward method
- **Disadvantages:**
  - Can distort the distribution of the variable
  - Standard deviation underestimated
  - Results in inaccurate regression coefficients; relationships between variables become distorted

Related observations

- Replace the missing values using information from another observation that is "similar" to the one with missingness
- The "similar" observation can come from within the same dataset (hot deck) or from an external dataset (cold deck)
- Examples:
  - Hot Deck: Mother's income can be used to fill in missing values for father's income
  - Cold Deck: Use respondents from 2009 NHANES survey to fill in missing values for the 2011 NHANES survey
- **Disadvantage:** Could expand effects of measurement error

Indicator variable: categorical predictor

- Make "missing" an additional category for the variable
  - Use this updated variable in the regression model; "missing" becomes a term for the model

What can you conclude if the term for missing is significant in the model?

Indicator variable: quantitative predictor

- Impute the missing in the original variable using the mean (or some other method) and create a new indicator variable for the missingness
- Can lead to inaccurate estimates of the coefficients of other variables, since the slope is forced to be the same for the groups with and without missingness
- Reduce some of this bias by including interactions between the missing indicator and the other predictors

Logical Rule

- Can use some logical rule to impute missing values
- Example: The Social Indicators Survey includes a question on the "number of months worked in the previous year" which was answered by all 1501 respondents. Of the people who didn't answer the question about total earnings in the previous year, 10 reported working 0 months during the previous year.

Logical Rule

- Can use some logical rule to impute missing values
- Example: The Social Indicators Survey includes a question on the "number of months worked in the previous year" which was answered by all 1501 respondents. Of the people who didn't answer the question about total earnings in the previous year, 10 reported working 0 months during the previous year.

For these 10 respondents, what is a logical value to use to impute their earnings?

Logical Rule

- Can use some logical rule to impute missing values
- Example: The Social Indicators Survey includes a question on the "number of months worked in the previous year" which was answered by all 1501 respondents. Of the people who didn't answer the question about total earnings in the previous year, 10 reported working 0 months during the previous year.

How would you impute the earnings for the other respondents who didn't share their earnings?

Acknowledgements

These slides draw material from

- Missing Data
- Handling Missing Data: An Introduction
- *Data Analysis Using Regression and Multilevel/Hierarchical Models*, "Chapter 25: Missing-data Imputation"