# Missing data

## Section 7

Sima Biondi

Spring 2025

Gov 51: Data Analysis and Politics

# Overview

1. Housekeeping
2. Back to basics with R
3. Hypothesis testing review
4. Fixed effects
5. Missing data

# Housekeeping

- Take a deep breath - you’re through the midterm

# Housekeeping

- Take a deep breath - you’re through the midterm
- Only 30% of the class is completed thus far (20% midterm, 10% Problem Set)

# Housekeeping

- Take a deep breath - you're through the midterm
- Only 30% of the class is completed thus far (20% midterm, 10% Problem Set)
    - Even if you didn't do the best, lots of the course left

# Housekeeping

- Take a deep breath - you're through the midterm
- Only 30% of the class is completed thus far (20% midterm, 10% Problem Set)
    - Even if you didn't do the best, lots of the course left
- Upcoming deadlines:

# Housekeeping

- Take a deep breath - you're through the midterm
- Only 30% of the class is completed thus far (20% midterm, 10% Problem Set)
    - Even if you didn't do the best, lots of the course left
- Upcoming deadlines:
    - Problem Set II: released today and due 4/3

# Housekeeping

- Take a deep breath - you're through the midterm
- Only 30% of the class is completed thus far (20% midterm, 10% Problem Set)
    - Even if you didn't do the best, lots of the course left
- Upcoming deadlines:
    - Problem Set II: released today and due 4/3
    - 1-pager: due 4/4

# Housekeeping: final project deadlines

- April 4th $\rightarrow$ one-page memo

More information on the details of each submission can be found here: `https://naijialiu.github.io/Gov_51/final.html`

# Housekeeping: final project deadlines

- April 4th $\rightarrow$ one-page memo
- April 10th $\rightarrow$ preliminary results draft due

More information on the details of each submission can be found here: `https://naijialiu.github.io/Gov_51/final.html`

# Housekeeping: final project deadlines

- April 4th $\rightarrow$ one-page memo
- April 10th $\rightarrow$ preliminary results draft due
- April 18th $\rightarrow$ first draft of poster

More information on the details of each submission can be found here: `https://naijialiu.github.io/Gov_51/final.html`

# Housekeeping: final project deadlines

- April 4th $\rightarrow$ one-page memo
- April 10th $\rightarrow$ preliminary results draft due
- April 18th $\rightarrow$ first draft of poster
- April 24th $\rightarrow$ final poster deadline

More information on the details of each submission can be found here: `https://naijialiu.github.io/Gov_51/final.html`

# Housekeeping: final project deadlines

- April 4th $\rightarrow$ one-page memo
- April 10th $\rightarrow$ preliminary results draft due
- April 18th $\rightarrow$ first draft of poster
- April 24th $\rightarrow$ final poster deadline
- April 29th $\rightarrow$ poster session

More information on the details of each submission can be found here: `https://naijialiu.github.io/Gov_51/final.html`

# Back to basics with R: working directories

Setting working directory

- You must tell R where you want it to find files

# Back to basics with R: working directories

Setting working directory

- You must tell R where you want it to find files
- You can do this with `setwd()` or `here()`

# Back to basics with R: working directories

Setting working directory

- You must tell R where you want it to find files
- You can do this with `setwd()` or `here()`
- If you are saving things to your Downloads, then you must tell R to look there

# Back to basics with R: working directories

Setting working directory

- You must tell R where you want it to find files
- You can do this with `setwd()` or `here()`
- If you are saving things to your Downloads, then you must tell R to look there

# Back to basics with R: working directories

Setting working directory

- You must tell R where you want it to find files
- You can do this with `setwd()` or `here()`
- If you are saving things to your Downloads, then you must tell R to look there

```
library(here) ## or setwd("~/path/to/your/project/root")
ed3_visits = read.csv(here("data", "processed",
    "rulers", "edwardiii_visits3.csv"))
```

# Back to basics with R: installing packages

Installing and loading packages

- If you want to use `tidyverse`, you must load `tidyverse`

# Back to basics with R: installing packages

Installing and loading packages

- If you want to use `tidyverse`, you must load `tidyverse`
- If you have loaded it, but closed R, you must load it again

# Back to basics with R: installing packages

Installing and loading packages

- If you want to use `tidyverse`, you must load `tidyverse`
- If you have loaded it, but closed R, you must load it again

# Back to basics with R: installing packages

Installing and loading packages

- If you want to use **tidyverse**, you must load **tidyverse**
- If you have loaded it, but closed R, you must load it again

```
  # for regular expression functionality in R, use the
      stringr package
install.packages("stringr")
library(stringr)
```

# Back to basics with R: workflow

Setting up an RMarkdown workflow for the final project:

1. Establish a Rproject object

# Back to basics with R: workflow

Setting up an RMarkdown workflow for the final project:

1. Establish a Rproject object
2. Store your data in the file associated with the Rproject

# Back to basics with R: workflow

Setting up an RMarkdown workflow for the final project:

1. Establish a Rproject object
2. Store your data in the file associated with the Rproject
3. Start writing your code, but be careful about reproducibility, ESPECIALLY if you overwrite your dataframes

# Back to basics with R: workflow

Setting up an RMarkdown workflow for the final project:

1. Establish a Rproject object
2. Store your data in the file associated with the Rproject
3. Start writing your code, but be careful about reproducibility, ESPECIALLY if you overwrite your dataframes
   - ↪ If you want to use RMarkdown as a script, I ***highly suggest*** you use the ”Run All Chunks Above” feature

# Hypothesis testing

Last time: we've covered hypothesis testing for $\widehat{\beta}$

- *Conceptually*: it was simply a comparison of distributions with **means**

# Hypothesis testing

Last time: we've covered hypothesis testing for $\widehat{\beta}$

- *Conceptually*: it was simply a comparison of distributions with **means**

# Hypothesis testing

Last time: we've covered hypothesis testing for $\widehat{\beta}$

- *Conceptually*: it was simply a comparison of distributions with **means** $\rightarrow$ we can apply hypothesis testing to compare means of quantities
- *Application*: Recall that the `lm` function uses the t-distribution instead of the normal

# Hypothesis testing

Last time: we've covered hypothesis testing for $\widehat{\beta}$

- *Conceptually*: it was simply a comparison of distributions with **means** $\rightarrow$ we can apply hypothesis testing to compare means of quantities
- *Application*: Recall that the `lm` function uses the t-distribution instead of the normal

# Hypothesis testing

Last time: we've covered hypothesis testing for $\widehat{\beta}$

- *Conceptually*: it was simply a comparison of distributions with **means** $\rightarrow$ we can apply hypothesis testing to compare means of quantities
- *Application*: Recall that the `lm` function uses the t-distribution instead of the normal

Generalized steps:

1. Specify a null and an alternative hypothesis

# Hypothesis testing

Last time: we've covered hypothesis testing for $\widehat{\beta}$

- *Conceptually*: it was simply a comparison of distributions with **means** $\rightarrow$ we can apply hypothesis testing to compare means of quantities
- *Application*: Recall that the `lm` function uses the t-distribution instead of the normal

Generalized steps:

1. Specify a null and an alternative hypothesis
2. Use the null hypothesis to specify a null distribution

# Hypothesis testing

Last time: we've covered hypothesis testing for $\widehat{\beta}$

- *Conceptually*: it was simply a comparison of distributions with **means** $\rightarrow$ we can apply hypothesis testing to compare means of quantities
- *Application*: Recall that the `lm` function uses the t-distribution instead of the normal

Generalized steps:

1. Specify a null and an alternative hypothesis
2. Use the null hypothesis to specify a null distribution
3. See how likely our alternative hypothesis is given the null distribution

# Hypothesis testing example

A study of medieval English king's power: when and why does the king travel around his kingdom?

# Hypothesis testing example

A study of medieval English king's power: when and why does the king travel around his kingdom?

↪ *Potential hypothesis*: king visits places around the country to collect more information about his reign

**Figure 1:** Source material for GoT

# Hypothesis testing example

A study of medieval English king's power: when and why does the king travel around his kingdom?

↪ *Potential hypothesis*: king visits places around the country to collect more information about his reign

**Figure 1:** Source material for GoT

The Black Death was a huge shock to England's economy and society (Payling 1992)

# Hypothesis testing example

A study of medieval English king's power: when and why does the king travel around his kingdom?

↪ *Potential hypothesis*: king visits places around the country to collect more information about his reign

**Figure 1:** Source material for GoT

The Black Death was a huge shock to England's economy and society (Payling 1992) → What happens to the king's travel after the plague?

# Hypothesis testing example

What happens to the king's travel after the plague?

# Hypothesis testing example

What happens to the king's travel after the plague? $\rightarrow$ the average number of miles traveled different in 1347 to 1350?

- $H_0$: no difference exists in the average number of miles traveled different in 1347 to 1350

# Hypothesis testing example

What happens to the king's travel after the plague? $\rightarrow$ the average number of miles traveled different in 1347 to 1350?

- $H_0$: no difference exists in the average number of miles traveled different in 1347 to 1350
- $H_1$: a difference exists in the average number of miles traveled different in 1347 to 1350

# Hypothesis testing example

What happens to the king's travel after the plague? $\rightarrow$ the average number of miles traveled different in 1347 to 1350?

- $H_0$: no difference exists in the average number of miles traveled different in 1347 to 1350
- $H_1$: a difference exists in the average number of miles traveled different in 1347 to 1350

# Hypothesis testing example

What happens to the king's travel after the plague? $\rightarrow$ the average number of miles traveled different in 1347 to 1350?

- $H_0$: no difference exists in the average number of miles traveled different in 1347 to 1350
- $H_1$: a difference exists in the average number of miles traveled different in 1347 to 1350

```
#load data frame and subset data
data(ed3_visits)
distdf <- ed3_visits[ed3_visits$year >= 1347 &
    ed3_visits$year <= 1350,]
```

# Hypothesis testing example

```r
# run ttest comparing average distances in 1347 and 1350
distance_ttest <-
    t.test(distdf$distance[ed3_visits$year == 1347],
           distdf$distance[ed3_visits$year == 1350],
           na.action = na.omit)
distance_ttest
```

```
        Welch Two Sample t-test

data:  treat$distance and control$distance
t = 2.7367, df = 36.886, p-value = 0.009489
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  5.996715 40.211449
sample estimates:
mean of x mean of y
 49.45357  26.34949
```

# Hypothesis testing example

How do we do this by hand?

```
est <- mean(distdf$distance[ed3_visits$year == 1347]) -
  mean(distdf$distance[ed3_visits$year == 1350])
treatSE <- var(distdf$distance[ed3_visits$year ==
    1350])/
  length(distdf$distance[ed3_visits$year == 1350])
controlSE <- var(distdf$distance[ed3_visits$year ==
    1347])/
  length(distdf$distance[ed3_visits$year == 1347])
se <- sqrt(treatSE + controlSE)

c(est - (se * 1.96), est + (se * 1.96))
```

# Fixed effects

Problem: your data may have observations that are just simply somewhat different from each other

- Example: do Presidents allocate federal funds to districts that supported them in the previous election?

# Fixed effects

Problem: your data may have observations that are just simply somewhat different from each other

- Example: do Presidents allocate federal funds to districts that supported them in the previous election?
    - California has different needs from Arkansas

# Fixed effects

Problem: your data may have observations that are just simply somewhat different from each other

- Example: do Presidents allocate federal funds to districts that supported them in the previous election?
    - California has different needs from Arkansas

# Fixed effects

Problem: your data may have observations that are just simply somewhat different from each other

- Example: do Presidents allocate federal funds to districts that supported them in the previous election?
  - California has different needs from Arkansas

How do we control for this in our regressions?

# Fixed effects

Problem: your data may have observations that are just simply somewhat different from each other

- Example: do Presidents allocate federal funds to districts that supported them in the previous election?
  - California has different needs from Arkansas

How do we control for this in our regressions? **Fixed effects!**

- Fixed effects are simply indicators for a particular trait of an observation or multiple observations

# Fixed effects

Problem: your data may have observations that are just simply somewhat different from each other

- Example: do Presidents allocate federal funds to districts that supported them in the previous election?
  - California has different needs from Arkansas

How do we control for this in our regressions? **Fixed effects!**

- Fixed effects are simply indicators for a particular trait of an observation or multiple observations
- If we simply ran a regression, the California observation would dominate our calculation of $\widehat{\beta}$

# Fixed effect: implementation

- Packages such as `fixest`, but can manually do it through base R

# Fixed effect: implementation

- Packages such as `fixest`, but can manually do it through base R
- Sometimes the fixed effect we want to control for is a year

# Fixed effect: implementation

- Packages such as `fixest`, but can manually do it through base R
- Sometimes the fixed effect we want to control for is a year
    - Years are numeric, so to turn them into indicators we use `factor`

# Fixed effect: implementation

- Packages such as `fixest`, but can manually do it through base R
- Sometimes the fixed effect we want to control for is a year
    - Years are numeric, so to turn them into indicators we use `factor`
- Generally good practice to "factorize" our fixed effects

# Fixed effect: implementation

- Packages such as `fixest`, but can manually do it through base R
- Sometimes the fixed effect we want to control for is a year
    - Years are numeric, so to turn them into indicators we use `factor`
- Generally good practice to "factorize" our fixed effects

# Fixed effect: implementation

- Packages such as `fixest`, but can manually do it through base R
- Sometimes the fixed effect we want to control for is a year
  - Years are numeric, so to turn them into indicators we use `factor`
- Generally good practice to "factorize" our fixed effects

```
model1 <- lm(y ~ x1 + x2, data = df)
model2 <- lm(y ~ x1 + x2 + factor(state), data = df)
```

# Fixed effect: interpretation

Recall: we're interested in do Presidents allocate federal funds to districts that supported them in the previous election?

# Fixed effect: interpretation

Recall: we're interested in do Presidents allocate federal funds to districts that supported them in the previous election?

```
model1 <- lm(y ~ x1 + x2, data = df)
model2 <- lm(y ~ x1 + x2 + factor(state), data = df)
```

How is `model1` different than `model2`?

# Fixed effect: interpretation

Recall: we're interested in do Presidents allocate federal funds to districts that supported them in the previous election?

```
model1 <- lm(y ~ x1 + x2, data = df)
model2 <- lm(y ~ x1 + x2 + factor(state), data = df)
```

How is `model1` different than `model2`?

- Controls for state fixed effects

# Fixed effect: interpretation

Recall: we're interested in do Presidents allocate federal funds to districts that supported them in the previous election?

```
model1 <- lm(y ~ x1 + x2, data = df)
model2 <- lm(y ~ x1 + x2 + factor(state), data = df)
```

How is `model1` different than `model2`?

- Controls for state fixed effects
- Approach helps control for omitted variable bias due to unobserved state-specific characteristics that could influence the allocation of federal funds (dependent variable)

# Missing data background

- Throughout modern social science, researchers have oftentimes dropped missing data.

# Missing data background

- Throughout modern social science, researchers have oftentimes dropped missing data.
- Example command in R:

```
mean(data$variable, na.rm = TRUE)
```

# Missing data background

- Throughout modern social science, researchers have oftentimes dropped missing data.
- Example command in R:

```
mean(data$variable, na.rm = TRUE)
```

- However, simply dropping missing data can induce bias, given missingness is not always random.

# Example of non-random missingness

What if poll response is not representative?

JUN. 15, 2020, AT 5:58 AM

## How To Read Polls In 2020

By Nathaniel Rakich

Filed under 2020 Election

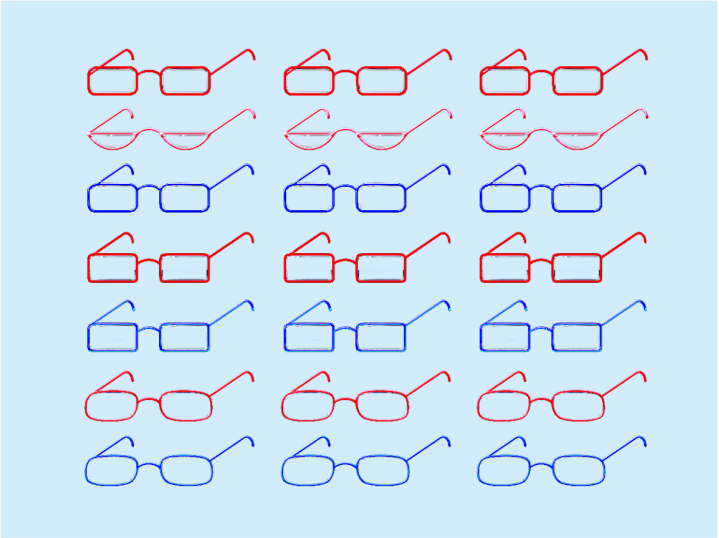

PHOTO ILLUSTRATION BY FIVETHIRTYEIGHT / GETTY IMAGES

We're about to enter the thick of general-election season, which means we're about to get a boatload of polls.

# Framework for understanding missing data

**Problem**: Our data is incomplete, and we (probably) don't know why apriori

# Framework for understanding missing data

**Problem**: Our data is incomplete, and we (probably) don't know why apriori

Solution: Depends on our assumptions about the missing data

# Framework for understanding missing data

**Problem**: Our data is incomplete, and we (probably) don't know why apriori

Solution: Depends on our assumptions about the missing data

- Each assumption is generally mutually exclusive and affects our strategies to address them

Missing data assumptions:

# Framework for understanding missing data

**Problem**: Our data is incomplete, and we (probably) don't know why apriori

Solution: Depends on our assumptions about the missing data

- Each assumption is generally mutually exclusive and affects our strategies to address them

Missing data assumptions:

1. Missing Completely at Random (MCAR)

# Framework for understanding missing data

**Problem**: Our data is incomplete, and we (probably) don't know why apriori

Solution: Depends on our assumptions about the missing data

- Each assumption is generally mutually exclusive and affects our strategies to address them

Missing data assumptions:

1. Missing Completely at Random (MCAR)
2. Missing at Random (MAR)

# Framework for understanding missing data

**Problem**: Our data is incomplete, and we (probably) don't know why apriori

**Solution**: Depends on our assumptions about the missing data

- Each assumption is generally mutually exclusive and affects our strategies to address them

Missing data assumptions:

1. Missing Completely at Random (MCAR)
2. Missing at Random (MAR)
3. Missing Not at Random (MNAR)

# Missing Completely at Random (MCAR)

- **Problem**: Observations are missing at random

# Missing Completely at Random (MCAR)

- **Problem**: Observations are missing at random
- Listwise deletion (e.g., dropping the observations with missing data) does not induce bias *because we assume MCAR*

# Missing Completely at Random (MCAR)

- **Problem**: Observations are missing at random
- Listwise deletion (e.g., dropping the observations with missing data) does not induce bias *because we assume MCAR*
- Incredibly stringent assumption - not many real-world situations have data that is missing completely at random

# Missing Completely at Random (MCAR)

- **Problem**: Observations are missing at random
- Listwise deletion (e.g., dropping the observations with missing data) does not induce bias *because we assume MCAR*
- Incredibly stringent assumption - not many real-world situations have data that is missing completely at random

# Missing Completely at Random (MCAR)

- **Problem**: Observations are missing at random
- Listwise deletion (e.g., dropping the observations with missing data) does not induce bias *because we assume MCAR*
- Incredibly stringent assumption - not many real-world situations have data that is missing completely at random

| i | Gender | White | Democrat | Vote Choice |
|---|---|---|---|---|
| 1 | 1 | 1 | 1 | Trump |
| 2 | NA | 1 | 0 | Biden |
| 3 | 0 | 0 | 1 | Biden |
| 4 | 1 | 0 | NA | Trump |
| 5 | NA | 0 | 1 | Trump |
| 6 | 0 | 0 | 1 | Biden |

# Missing at Random (MAR)

- **Problem**: Conditional on observable covariates, observations are missing at random

# Missing at Random (MAR)

- **Problem**: Conditional on observable covariates, observations are missing at random
  - A bit of a misnomer - probably better to call it conditionally missing at random

# Missing at Random (MAR)

- **Problem**: Conditional on observable covariates, observations are missing at random
  - A bit of a misnomer - probably better to call it conditionally missing at random
- Less restrictive than MCAR, but still a stringent assumption

# Missing at Random (MAR)

- **Problem**: Conditional on observable covariates, observations are missing at random
  - A bit of a misnomer - probably better to call it conditionally missing at random
- Less restrictive than MCAR, but still a stringent assumption
  - Listwise deletion does induce bias because data is not missing randomly

# Missing at Random (MAR)

- **Problem**: Conditional on observable covariates, observations are missing at random
  - A bit of a misnomer - probably better to call it conditionally missing at random
- Less restrictive than MCAR, but still a stringent assumption
  - Listwise deletion does induce bias because data is not missing randomly
- Solution: Multiple imputation

# Missing at Random (MAR)

- **Problem**: Conditional on observable covariates, observations are missing at random
  - A bit of a misnomer - probably better to call it conditionally missing at random
- Less restrictive than MCAR, but still a stringent assumption
  - Listwise deletion does induce bias because data is not missing randomly
- Solution: Multiple imputation
  - Implementation requires using observed data to **impute** missing values.

# Missing Not at Random (MNAR)

- **Problem**: Unobserved covariates are influencing missingness

# Missing Not at Random (MNAR)

- **Problem**: Unobserved covariates are influencing missingness
- Least restrictive assumption, but difficult to address given the unobserved nature of the bias

# Missing Not at Random (MNAR)

- **Problem**: Unobserved covariates are influencing missingness
- Least restrictive assumption, but difficult to address given the unobserved nature of the bias
  - $\rightarrow$ Listwise deletion would induce bias because data is not missing randomly

# Missing Not at Random (MNAR)

- **Problem**: Unobserved covariates are influencing missingness
- Least restrictive assumption, but difficult to address given the unobserved nature of the bias
  - $\rightarrow$ Listwise deletion would induce bias because data is not missing randomly
  - $\rightarrow$ Multiple imputation relies on observed covariates - cannot impute with unobserved covariates

# Missing Not at Random (MNAR)

- **Problem**: Unobserved covariates are influencing missingness
- Least restrictive assumption, but difficult to address given the unobserved nature of the bias
  - $\rightarrow$ Listwise deletion would induce bias because data is not missing randomly
  - $\rightarrow$ Multiple imputation relies on observed covariates - cannot impute with unobserved covariates
- **Solution**: better modeling and/or data collection

# Framework for missing data

- Missing data has been insufficiently addressed throughout empirical social science

# Framework for missing data

- Missing data has been insufficiently addressed throughout empirical social science
- Organize types of missing data for resolution:

# Framework for missing data

- Missing data has been insufficiently addressed throughout empirical social science
- Organize types of missing data for resolution:
  - MCAR $\rightarrow$ listwise deletion.

# Framework for missing data

- Missing data has been insufficiently addressed throughout empirical social science
- Organize types of missing data for resolution:
  - MCAR $\rightarrow$ listwise deletion.
  - MAR $\rightarrow$ multiple imputation.

# Framework for missing data

- Missing data has been insufficiently addressed throughout empirical social science
- Organize types of missing data for resolution:
  - MCAR $\rightarrow$ listwise deletion.
  - MAR $\rightarrow$ multiple imputation.
  - MNAR $\rightarrow$ better modeling/data collection.

# Framework for missing data

- Missing data has been insufficiently addressed throughout empirical social science
- Organize types of missing data for resolution:
  - MCAR $\rightarrow$ listwise deletion.
  - MAR $\rightarrow$ multiple imputation.
  - MNAR $\rightarrow$ better modeling/data collection.
- Gov department features leaders in research on missing data:

# Framework for missing data

- Missing data has been insufficiently addressed throughout empirical social science
- Organize types of missing data for resolution:
  - MCAR $\rightarrow$ listwise deletion.
  - MAR $\rightarrow$ multiple imputation.
  - MNAR $\rightarrow$ better modeling/data collection.
- Gov department features leaders in research on missing data:
  - Professor Naijia Liu

# Framework for missing data

- Missing data has been insufficiently addressed throughout empirical social science
- Organize types of missing data for resolution:
  - MCAR $\rightarrow$ listwise deletion.
  - MAR $\rightarrow$ multiple imputation.
  - MNAR $\rightarrow$ better modeling/data collection.
- Gov department features leaders in research on missing data:
  - Professor Naijia Liu
  - Professor Matthew Blackwell

# Framework for missing data

- Missing data has been insufficiently addressed throughout empirical social science
- Organize types of missing data for resolution:
  - MCAR $\rightarrow$ listwise deletion.
  - MAR $\rightarrow$ multiple imputation.
  - MNAR $\rightarrow$ better modeling/data collection.
- Gov department features leaders in research on missing data:
  - Professor Naijia Liu
  - Professor Matthew Blackwell
  - Professor Kosuke Imai