GOV 51 Section

Week 7: Missing Data

Pranav Moudgalya

Harvard College

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)
  - 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
- Questions?

Housekeeping Continued

- Reminder of deadlines

Housekeeping Continued

- Reminder of deadlines
  - April 4th → one-page memo
  - April 10th → preliminary results draft due
  - April 18th → first draft of poster
  - April 24th → final poster deadline
  - April 29th → poster session
- **By April 4th, mandatory OH with Pranav or Sima per group. This is mandatory!** We'd like to hear about your idea and dataset. Please feel free to come to my office hours (walk in).

Housekeeping Continued

- Reminder of 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
- **By April 4th, mandatory OH with Pranav or Sima per group. This is mandatory!** We'd like to hear about your idea and dataset. Please feel free to come to my office hours (walk in).
- If you are still looking for data: Run `data()` in RStudio. These are built-in and cleaned datasets!

Housekeeping Continued

- Reminder of deadlines
  - April 4th → one-page memo
  - April 10th → preliminary results draft due
  - April 18th → first draft of poster
  - April 24th → final poster deadline
  - April 29th → poster session
- **By April 4th, mandatory OH with Pranav or Sima per group. This is mandatory!** We'd like to hear about your idea and dataset. Please feel free to come to my office hours (walk in).
- If you are still looking for data: Run `data()` in RStudio. These are built-in and cleaned datasets!
- You can also check out these GOV 50 data sources.

How to set up a research project in R

- **Organize!** Have a dedicated folder for your data, code, and figures/tables.

How to set up a research project in R

- **Organize!** Have a dedicated folder for your data, code, and figures/tables.
- **Divide and conquer!** Split up coding tasks into manageable, smaller R script files. Don't use markdown!

How to set up a research project in R

- **Organize!** Have a dedicated folder for your data, code, and figures/tables.
- **Divide and conquer!** Split up coding tasks into manageable, smaller R script files. Don't use markdown!
- **Comment!** Comment on your code so that you recall what steps you took in each step of your analysis.

Hyphothesis Testing

- We've covered hypothesis testing for $\widehat{\beta}$

Hyphothesis Testing

- We've covered hypothesis testing for $\widehat{\beta}$
  - Taking a step back - it was simply a comparison of distributions with **means**

Hyphothesis Testing

- We've covered hypothesis testing for $\widehat{\beta}$
  - Taking a step back - it was simply a comparison of distributions with **means**
- We can apply hypothesis testing to compare means of quantities

Hyphothesis Testing

- We've covered hypothesis testing for $\widehat{\beta}$
  - Taking a step back - it was simply a comparison of distributions with **means**
- We can apply hypothesis testing to compare means of quantities
- Recall that the `lm` function uses the t-distribution instead of the normal

Hyphothesis Testing

- We've covered hypothesis testing for $\widehat{\beta}$
  - Taking a step back - it was simply a comparison of distributions with **means**
- We can apply hypothesis testing to compare means of quantities
- Recall that the `lm` function uses the t-distribution instead of the normal

1. Specify a null and an alternative hypothesis

Hyphothesis Testing

- We've covered hypothesis testing for $\widehat{\beta}$
  - Taking a step back - it was simply a comparison of distributions with **means**
- We can apply hypothesis testing to compare means of quantities
- Recall that the `lm` function uses the t-distribution instead of the normal

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

Hyphothesis Testing

- We've covered hypothesis testing for $\widehat{\beta}$
  - Taking a step back - it was simply a comparison of distributions with **means**
- We can apply hypothesis testing to compare means of quantities
- Recall that the `lm` function uses the t-distribution instead of the normal

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

Hyphothesis Testing

- We've covered hypothesis testing for $\widehat{\beta}$
  - Taking a step back - it was simply a comparison of distributions with **means**
- We can apply hypothesis testing to compare means of quantities
- Recall that the `lm` function uses the t-distribution instead of the normal

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

- The MLB lowered the height of the pitching mound by five inches after the 1968 season

Hypothesis Testing Example

- The MLB lowered the height of the pitching mound by five inches after the 1968 season
- Pitchers were getting too good!

Hypothesis Testing Example

- The MLB lowered the height of the pitching mound by five inches after the 1968 season
- Pitchers were getting too good!
  - Was the average number of homeruns per player different in 1968 to 1969?

Hypothesis Testing Example

- The MLB lowered the height of the pitching mound by five inches after the 1968 season
- Pitchers were getting too good!
  - Was the average number of homeruns per player different in 1968 to 1969?

```
data(baseball)
baseball <- baseball[baseball$year >= 1968 &
                              baseball$year <= 1969,]
```

Hypothesis Testing Example 2

```
hr <- t.test(baseball$hr[baseball$year == 1968],
             baseball$hr[baseball$year == 1969],
             na.action = na.omit)

hr
```

```
##
##  Welch Two Sample t-test
##
## data:  baseball$hr[baseball$year == 1968] and baseball$h
## t = -1.6923, df = 463.12, p-value = 0.09126
## alternative hypothesis: true difference in means is not
## 95 percent confidence interval:
##  -3.2476715  0.2422414
## sample estimates:
## mean of x mean of y
##  4.943925  6.446640
```

Hypothesis Testing Example 3

How do we do this by hand?

```
est <- mean(baseball$hr[baseball$year == 1968]) -
  mean(baseball$hr[baseball$year == 1969])
treatSE <- var(baseball$hr[baseball$year == 1969])/
  length(baseball$hr[baseball$year == 1969])
controlSE <- var(baseball$hr[baseball$year == 1968])/
  length(baseball$hr[baseball$year == 1968])
se <- sqrt(treatSE + controlSE)

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

```
## [1] -3.2431433  0.2377132
```

Hypothesis Testing

- Remember that a difference in means is an **estimator**, which means it has some **standard error**, meaning it varies from sample to sample.

Hypothesis Testing

- Remember that a difference in means is an **estimator**, which means it has some **standard error**, meaning it varies from sample to sample.
- By the central limit theorem, with a large sample sizes, over many samples, the difference in means estimator will be approximately normal (or we can use a t distribution when the sample size is small $\rightsquigarrow$ t distribution has fatter tails).

Hypothesis Testing

- Remember that a difference in means is an **estimator**, which means it has some **standard error**, meaning it varies from sample to sample.
- By the central limit theorem, with a large sample sizes, over many samples, the difference in means estimator will be approximately normal (or we can use a t distribution when the sample size is small $\rightsquigarrow$ t distribution has fatter tails).
- Formally, for samples A and B, with small sample sizes

Hypothesis Testing

- Remember that a difference in means is an **estimator**, which means it has some **standard error**, meaning it varies from sample to sample.
- By the central limit theorem, with a large sample sizes, over many samples, the difference in means estimator will be approximately normal (or we can use a t distribution when the sample size is small $\rightsquigarrow$ t distribution has fatter tails).
- Formally, for samples A and B, with small sample sizes
  1. Diff in means estimator $\rightsquigarrow \bar{X}_A - \bar{X}_B$

Hypothesis Testing

- Remember that a difference in means is an **estimator**, which means it has some **standard error**, meaning it varies from sample to sample.
- By the central limit theorem, with a large sample sizes, over many samples, the difference in means estimator will be approximately normal (or we can use a t distribution when the sample size is small $\rightsquigarrow$ t distribution has fatter tails).
- Formally, for samples A and B, with small sample sizes
  1. Diff in means estimator $\rightsquigarrow \bar{X}_A - \bar{X}_B$
  2. Standard error $\rightsquigarrow \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$

Hypothesis Testing

- Remember that a difference in means is an **estimator**, which means it has some **standard error**, meaning it varies from sample to sample.
- By the central limit theorem, with a large sample sizes, over many samples, the difference in means estimator will be approximately normal (or we can use a t distribution when the sample size is small $\rightsquigarrow$ t distribution has fatter tails).
- Formally, for samples A and B, with small sample sizes
  1. Diff in means estimator $\rightsquigarrow \bar{X}_A - \bar{X}_B$
  2. Standard error $\rightsquigarrow \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$
  3. Critical values (where $\alpha = 0.95$) $\rightsquigarrow t_{\alpha/2}$

Hypothesis Testing

- Remember that a difference in means is an **estimator**, which means it has some **standard error**, meaning it varies from sample to sample.
- By the central limit theorem, with a large sample sizes, over many samples, the difference in means estimator will be approximately normal (or we can use a t distribution when the sample size is small $\rightsquigarrow$ t distribution has fatter tails).
- Formally, for samples A and B, with small sample sizes
  1. Diff in means estimator $\rightsquigarrow \bar{X}_A - \bar{X}_B$
  2. Standard error $\rightsquigarrow \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$
  3. Critical values (where $\alpha = 0.95$) $\rightsquigarrow t_{\alpha/2}$
  4. 95% confidence interval $\rightsquigarrow \bar{X}_A - \bar{X}_B \pm t_{\alpha/2}\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$

Reminder: Fixed Effects

- In regression, we can use something called **fixed effects** to control for unobserved characteristics such as ability level in studies of educational policy.

Reminder: Fixed Effects

- In regression, we can use something called **fixed effects** to control for unobserved characteristics such as ability level in studies of educational policy.
- We often include **time** and **unit** fixed effects to account for time-specific, but unit invariant fixed effects and unit-specific, but time invariant fixed effects, respectively.

Reminder: Fixed Effects

- In regression, we can use something called **fixed effects** to control for unobserved characteristics such as ability level in studies of educational policy.
- We often include **time** and **unit** fixed effects to account for time-specific, but unit invariant fixed effects and unit-specific, but time invariant fixed effects, respectively.
- Operationally, this means just including a factor variable in your regression that uniquely represents each time period or unit.

Reminder: Fixed Effects

- In regression, we can use something called **fixed effects** to control for unobserved characteristics such as ability level in studies of educational policy.
- We often include **time** and **unit** fixed effects to account for time-specific, but unit invariant fixed effects and unit-specific, but time invariant fixed effects, respectively.
- Operationally, this means just including a factor variable in your regression that uniquely represents each time period or unit.
- Great way to account for some unobserved potential confounding variables, but often not sufficient!

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

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

Missing Data Background

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

```
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

NOVEMBER 13, 2020

Understanding how 2020 election polls performed and what it might mean for other kinds of survey work

BY SCOTT KEETER, COURTNEY KENNEDY AND CLAUDIA DEANE

(Brianna Soukup/Portland Press Herald via Getty Images)

What if poll response is not representative?

Framework for Understanding Missing Data

- Problem: Our data is incomplete
- Solution: Depends on our assumptions about the missing data
- Each assumption is generally mutually exclusive and affects our strategies to address them

Assumptions

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

Missing Completely at Random (MCAR)

- Observations are missing at random
- Listwise deletion (e.g. dropping the observations with missing data) does not induce bias
- 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 Completely at Random (MCAR)

- Observations are missing at random
- Listwise deletion (e.g. dropping the observations with missing data) does not induce bias because data is missing at random

i	Gender	White	Democrat	Vote Choice
1	1	1	1	Trump
3	0	0	1	Biden
6	0	0	1	Biden

Missing at Random (MAR)

- 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 stringent assumption

Missing at Random (MAR)

- 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 stringent assumption
- Listwise deletion does induce bias because data is not missing randomly
- Example: Worry with polling in 2016 and 2020 is that conservatives are not being captured - listwise deletion would underrepresent this population, making accurate predictions difficult

Missing at Random (MAR)

- 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 stringent assumption
- Listwise deletion does induce bias because data is not missing randomly
- Example: Worry with polling in 2016 and 2020 is that conservatives are not being captured - listwise deletion would underrepresent this population, making accurate predictions difficult
- Multiple imputation as a solution
- Implementation requires using observed data to **impute** values that are missing, using linear regression for instance!

Missing Not at Random (MNAR)

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

Framework for Missing Data

- Missing data has been insufficiently addressed throughout empirical social science
- In order to address how missing data affects our results, we organize types of missing data

1. MCAR $\rightarrow$ listwise deletion
2. MAR $\rightarrow$ multiple imputation
3. MNAR $\rightarrow$ better modelling/data collection

- Gov department features leaders in research on missing data
- Professor Naijia Liu
- Professor Matthew Blackwell
- Professor Kosuke Imai

Summary

- Missing data is everywhere!
- Three possible mechanisms:
  - Missing completely at random $\rightsquigarrow$ listwise deletion
  - Missing at random $\rightsquigarrow$ multiple imputation
  - Missing not at random $\rightsquigarrow$ more careful modeling
- Dealing with missing values often leads to different study results!