Text as data

Section 8

Sima Biondi

Spring 2025

Gov 51: Data Analysis and Politics

Overview

1. Housekeeping
2. Missing data implementation
3. Text as data

Housekeeping

- Remember to sign-up for office hours with me or Pranav as a group to discuss projects

Housekeeping

- Remember to sign-up for office hours with me or Pranav as a group to discuss projects
- Upcoming deadlines:

Housekeeping

- Remember to sign-up for office hours with me or Pranav as a group to discuss projects
- Upcoming deadlines:
    - April 3 (today): Problem Set II due

Housekeeping

- Remember to sign-up for office hours with me or Pranav as a group to discuss projects
- Upcoming deadlines:
  - April 3 (today): Problem Set II due
  - April 4 (tomorrow): 1-pager due, see `https://naijialiu.github.io/Gov_51/final.html` for more information

Housekeeping

- Remember to sign-up for office hours with me or Pranav as a group to discuss projects
- Upcoming deadlines:
  - April 3 (today): Problem Set II due
  - April 4 (tomorrow): 1-pager due, see `https://naijialiu.github.io/Gov_51/final.html` for more information
  - April 11: Initial result due, within one page write up

Missing data implementation

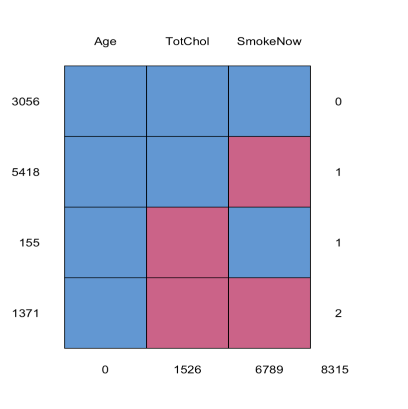
- ‘mice‘ package - ”Multiple Imputation by Chained Equations”

```r
library(mice) # recall install.packages("mice") for first use
library(NHANES)

data(NHANES)
nhanes <- NHANES[c("Age", "SmokeNow", "TotChol")]
```

Patterns in the data’s missingness

```
md.pattern(nhanes)
```



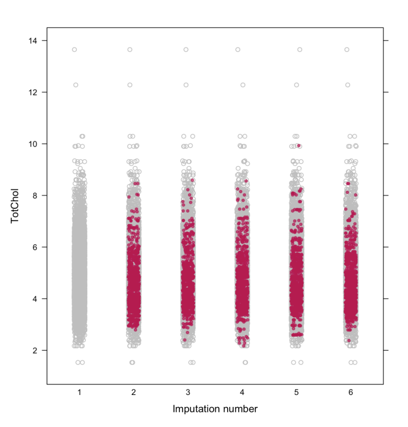
Multiple imputation

```r
# m specifies the number of imputation cycles
nhanes2MI5 <- mice(nhanes, m = 5)

# exports complete data with imputations
df <- complete(nhanes2MI5, 5)
```

Evaluating imputation

```r
stripplot(nhanes2MI5, TotChol ~ .imp,
          col = c("grey", mdc(2)),
          pch = c(1,20))
```



Regressions with imputed data

```
model1 <- lm(TotChol ~ Age + SmokeNow, data = df)
    library(modelsummary)
modelsummary(model1, stars = T,
             gof_omit =
                 'DF|Deviance|AIC|BIC|F|Log.Lik.|RMSE')
```

	(1)
(Intercept)	4.157***
	(0.028)
Age	0.018***
	(0.000)
SmokeNowYes	-0.009
	(0.022)
Num.Obs.	10000
R2	0.141
R2 Adj.	0.141

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Comparing with listwise deletion

```r
model2 <- lm(TotChol ~ Age + SmokeNow, data = nhanes)
```

	(1)
(Intercept)	4.775***
	(0.072)
Age	0.006***
	(0.001)
SmokeNowYes	-0.022
	(0.042)
Num.Obs.	3056
R2	0.010
R2 Adj.	0.009

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Multiple imputation

- Questions?

Text as data

- A large goal of this course is to give you a framework to understand different methodologies and the challenges they tackle.

Text as data

- A large goal of this course is to give you a framework to understand different methodologies and the challenges they tackle.
- Missing data

Text as data

- A large goal of this course is to give you a framework to understand different methodologies and the challenges they tackle.
- Missing data

Text as data

- A large goal of this course is to give you a framework to understand different methodologies and the challenges they tackle.
- Missing data $\rightarrow$ types of missingness drives our solutions.
- Text as data

Text as data

- A large goal of this course is to give you a framework to understand different methodologies and the challenges they tackle.
- Missing data $\rightarrow$ types of missingness drives our solutions.
- Text as data

Text as data

- A large goal of this course is to give you a framework to understand different methodologies and the challenges they tackle.
- Missing data $\rightarrow$ types of missingness drives our solutions.
- Text as data $\rightarrow$?

Examples of text as data in research

- Ban, Grimmer, Kaslovsky, and West (2022) - committee hearings and the effect of women in Congress

Examples of text as data in research

- Ban, Grimmer, Kaslovsky, and West (2022) - committee hearings and the effect of women in Congress
    - Finds that less interruptions with more women, more substantive conversations.

Examples of text as data in research

- Ban, Grimmer, Kaslovsky, and West (2022) - committee hearings and the effect of women in Congress
  - Finds that less interruptions with more women, more substantive conversations.
- King, Pan, and Roberts (2013) - social media in censored contexts

Examples of text as data in research

- Ban, Grimmer, Kaslovsky, and West (2022) - committee hearings and the effect of women in Congress
    - Finds that less interruptions with more women, more substantive conversations.
- King, Pan, and Roberts (2013) - social media in censored contexts
    - Censorship in China does not target individual criticisms, but instead attempts at collective action.

Examples of text as data in research

- Ban, Grimmer, Kaslovsky, and West (2022) - committee hearings and the effect of women in Congress
  - Finds that less interruptions with more women, more substantive conversations.
- King, Pan, and Roberts (2013) - social media in censored contexts
  - Censorship in China does not target individual criticisms, but instead attempts at collective action.
- Communication relies on or can be represented by text

Examples of text as data in research

- Ban, Grimmer, Kaslovsky, and West (2022) - committee hearings and the effect of women in Congress
  - Finds that less interruptions with more women, more substantive conversations.
- King, Pan, and Roberts (2013) - social media in censored contexts
  - Censorship in China does not target individual criticisms, but instead attempts at collective action.
- Communication relies on or can be represented by text

Examples of text as data in research

- Ban, Grimmer, Kaslovsky, and West (2022) - committee hearings and the effect of women in Congress
  - Finds that less interruptions with more women, more substantive conversations.
- King, Pan, and Roberts (2013) - social media in censored contexts
  - Censorship in China does not target individual criticisms, but instead attempts at collective action.
- Communication relies on or can be represented by text $\rightarrow$ many political applications!

Data generating process (DGP)

Key question: how do humans come up with sentences?

- One theory is that given a topic (or multiple!), there is a certain probability that words appear

Data generating process (DGP)

Key question: how do humans come up with sentences?

- One theory is that given a topic (or multiple!), there is a certain probability that words appear
- In a given *document*, there could be a number of topics

Data generating process (DGP)

Key question: how do humans come up with sentences?

- One theory is that given a topic (or multiple!), there is a certain probability that words appear
- In a given *document*, there could be a number of topics
- Those topics then dictate the likelihood of which words appear

Data generating process (DGP)

Key question: how do humans come up with sentences?

- One theory is that given a topic (or multiple!), there is a certain probability that words appear
- In a given *document*, there could be a number of topics
- Those topics then dictate the likelihood of which words appear

Data generating process (DGP)

Key question: how do humans come up with sentences?

- One theory is that given a topic (or multiple!), there is a certain probability that words appear
- In a given *document*, there could be a number of topics
- Those topics then dictate the likelihood of which words appear

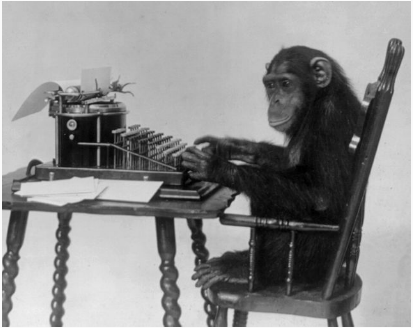


Figure 1: Opposite of this!

Data generating process (DGP)

Takeaways of the DGP for text data:

- Pretty rigid framework, but some ground truth

Data generating process (DGP)

Takeaways of the DGP for text data:

- Pretty rigid framework, but some ground truth
- Also undergirds the logic under Chat GPT and other large language models (LLMs)

Data generating process (DGP)

Takeaways of the DGP for text data:

- Pretty rigid framework, but some ground truth
- Also undergirds the logic under Chat GPT and other large language models (LLMs)
- Why GPTZero and other programs to detect AI usage are somewhat easily able to detect because text generated using this framework is rigid!

Bag of words model

Where do we begin analysis on text data?

Bag of words model

Where do we begin analysis on text data? One of the most common models is ***bag-of-words***

- Text is just a collection of words - the order and structure do not matter particularly when we care about topical relevancy

Bag of words model

Where do we begin analysis on text data? One of the most common models is *bag-of-words*

- Text is just a collection of words - the order and structure do not matter particularly when we care about topical relevancy
- Method **assumes** that the frequency of words can provide us information about the context in the text

Bag of words model

Where do we begin analysis on text data? One of the most common models is *bag-of-words*

- Text is just a collection of words - the order and structure do not matter particularly when we care about topical relevancy
- Method **assumes** that the frequency of words can provide us information about the context in the text

Bag of words model

Where do we begin analysis on text data? One of the most common models is ***bag-of-words***

- Text is just a collection of words - the order and structure do not matter particularly when we care about topical relevancy
- Method **assumes** that the frequency of words can provide us information about the context in the text

Process:

1. Tokenization - dividing text into individual words

Bag of words model

Where do we begin analysis on text data? One of the most common models is *bag-of-words*

- Text is just a collection of words - the order and structure do not matter particularly when we care about topical relevancy
- Method **assumes** that the frequency of words can provide us information about the context in the text

Process:

1. Tokenization - dividing text into individual words
2. Counting - counting the frequency they show up

Bag of words model

Where do we begin analysis on text data? One of the most common models is ***bag-of-words***

- Text is just a collection of words - the order and structure do not matter particularly when we care about topical relevancy
- Method **assumes** that the frequency of words can provide us information about the context in the text

Process:

1. Tokenization - dividing text into individual words
2. Counting - counting the frequency they show up
3. Vectorization - representing the text as a vector of word frequencies

Example: Congressional hearings about TikTok

Example: Congressional hearings about TikTok

"Does TikTok access the home Wi-Fi Network" - Richard Hudson (R-NC)

"Tiktok is fun" - Sima Biondi (Not in Congress-CA)

Example: Congressional hearings about TikTok

"Does TikTok access the home Wi-Fi Network" - Richard Hudson (R-NC)

"Tiktok is fun" - Sima Biondi (Not in Congress-CA)

Process:

1. Tokenization: {TikTok, network, fun, etc.}

Example: Congressional hearings about TikTok

"Does TikTok access the home Wi-Fi Network" - Richard Hudson (R-NC)

"Tiktok is fun" - Sima Biondi (Not in Congress-CA)

Process:

1. Tokenization: {TikTok, network, fun, etc.}
2. Counting frequency of text appearances: {2, 1, 1, etc.}

Example: Congressional hearings about TikTok

"Does TikTok access the home Wi-Fi Network" - Richard Hudson (R-NC)

"Tiktok is fun" - Sima Biondi (Not in Congress-CA)

Process:

1. Tokenization: {TikTok, network, fun, etc.}
2. Counting frequency of text appearances: {2, 1, 1, etc.}
3. Vectorization $\rightarrow$

i	TikTok	network	fun
Richard Hudson	1	1	0
Sima Biondi	0	0	1

Example vectorized table + applications

i	Does	TikTok	access	the	home	Wi-Fi	network	is	fun	China	...
1	1	1	1	1	1	1	1	0	0	0	...
2	0	1	0	0	0	0	0	1	1	0	...
...	...	...	...	...	...	...	...	...	...	...	...

Result is a sparse matrix (lots of missing data) to analyze!

Applications

- Topic modeling

Example vectorized table + applications

i	Does	TikTok	access	the	home	Wi-Fi	network	is	fun	China	...
1	1	1	1	1	1	1	1	0	0	0	...
2	0	1	0	0	0	0	0	1	1	0	...
...	...	...	...	...	...	...	...	...	...	...	...

Result is a sparse matrix (lots of missing data) to analyze!

Applications

- Topic modeling
- Sentiment analysis

Example vectorized table + applications

i	Does	TikTok	access	the	home	Wi-Fi	network	is	fun	China	...
1	1	1	1	1	1	1	1	0	0	0	...
2	0	1	0	0	0	0	0	1	1	0	...
...	...	...	...	...	...	...	...	...	...	...	...

Result is a sparse matrix (lots of missing data) to analyze!

Applications

- Topic modeling
- Sentiment analysis
- Text classification

Example vectorized table + applications

i	Does	TikTok	access	the	home	Wi-Fi	network	is	fun	China	...
1	1	1	1	1	1	1	1	0	0	0	...
2	0	1	0	0	0	0	0	1	1	0	...
...	...	...	...	...	...	...	...	...	...	...	...

Result is a sparse matrix (lots of missing data) to analyze!

Applications

- Topic modeling
- Sentiment analysis
- Text classification
- Also a great application of Lasso, since data can get VERY large.

Example application: topic modeling

- Can be supervised or unsupervised

Example application: topic modeling

- Can be supervised or unsupervised
  - Supervised: use labeled data to guide model in identifying topics

Example application: topic modeling

- Can be supervised or unsupervised
  - Supervised: use labeled data to guide model in identifying topics
  - Unsupervised: no labeled data, based on distribution of words in document

Example application: topic modeling

- Can be supervised or unsupervised
  - Supervised: use labeled data to guide model in identifying topics
  - Unsupervised: no labeled data, based on distribution of words in document
- Oftentimes, will require some human interpretation of the generated categories

Example application: topic modeling

- Can be supervised or unsupervised
  - Supervised: use labeled data to guide model in identifying topics
  - Unsupervised: no labeled data, based on distribution of words in document
- Oftentimes, will require some human interpretation of the generated categories
- Example: Terman (2017)

Example application: topic modeling

- Can be supervised or unsupervised
  - Supervised: use labeled data to guide model in identifying topics
  - Unsupervised: no labeled data, based on distribution of words in document
- Oftentimes, will require some human interpretation of the generated categories
- Example: Terman (2017)
  - Examines portrayals of Muslim woman in American media

Example application: topic modeling

- Can be supervised or unsupervised
  - Supervised: use labeled data to guide model in identifying topics
  - Unsupervised: no labeled data, based on distribution of words in document
- Oftentimes, will require some human interpretation of the generated categories
- Example: Terman (2017)
  - Examines portrayals of Muslim woman in American media
  - Results suggest that US news media propagate the perception that Muslims are distinctly sexist

Summary

- Multiple imputation is easily implemented through the ‘mice‘ package.

Summary

- Multiple imputation is easily implemented through the ‘mice‘ package.
- Introduces a number of easy to use functions that help with descriptive statistics and modeling.

Summary

- Multiple imputation is easily implemented through the ‘mice‘ package.
- Introduces a number of easy to use functions that help with descriptive statistics and modeling.
- Text as data is incredibly popular and powerful tools in social science.

Summary

- Multiple imputation is easily implemented through the ‘mice‘ package.
- Introduces a number of easy to use functions that help with descriptive statistics and modeling.
- Text as data is incredibly popular and powerful tools in social science.
- Data generating process of text is an important foundation for understanding how to tackle text as data.

Summary

- Multiple imputation is easily implemented through the ‘mice‘ package.
- Introduces a number of easy to use functions that help with descriptive statistics and modeling.
- Text as data is incredibly popular and powerful tools in social science.
- Data generating process of text is an important foundation for understanding how to tackle text as data.
- Bag-of-words is a simple, but powerful model to analyze text.