# POLS-Y 575

## Political Data Analysis I

Caleb Lucas

August 28, 2024

Indiana University

# Introductions

- Your instructor

# Introductions

- Your instructor
  - Me (Caleb Lucas)

# Introductions

- Your instructor
  - Me (Caleb Lucas)
- Your colleagues

# Introductions

- Your instructor
  - Me (Caleb Lucas)
- Your colleagues
  - Name

# Introductions

- Your instructor
  - Me (Caleb Lucas)
- Your colleagues
  - Name
  - Brief research interests

# Introductions

- Your instructor
  - Me (Caleb Lucas)
- Your colleagues
  - Name
  - Brief research interests
  - Favorite food

# Introductions

- Your instructor
  - Me (Caleb Lucas)
- Your colleagues
  - Name
  - Brief research interests
  - Favorite food
  - Briefly: if you could have access to any data you can think of for your research, what would you choose?

# Orientation

- First course in quantitative methods sequence

# Orientation

- First course in quantitative methods sequence
  - **575** $\rightarrow$ **576** $\rightarrow$ MLE $\rightarrow$ Topical$_1$ $\rightarrow$ Topical$_2$

# Orientation

- First course in quantitative methods sequence
  - **575** $\rightarrow$ **576** $\rightarrow$ MLE $\rightarrow$ Topical$_1$ $\rightarrow$ Topical$_2$
- We will cover:

# Orientation

- First course in quantitative methods sequence
  - **575** $\rightarrow$ **576** $\rightarrow$ MLE $\rightarrow$ Topical$_1$ $\rightarrow$ Topical$_2$
- We will cover:
  - Basic probability

# Orientation

- First course in quantitative methods sequence
  - **575** $\rightarrow$ **576** $\rightarrow$ MLE $\rightarrow$ Topical$_1$ $\rightarrow$ Topical$_2$
- We will cover:
  - Basic probability
  - Variety of statistical tests

# Orientation

- First course in quantitative methods sequence
  - **575** $\rightarrow$ **576** $\rightarrow$ MLE $\rightarrow$ Topical$_1$ $\rightarrow$ Topical$_2$
- We will cover:
  - Basic probability
  - Variety of statistical tests
  - Linear regression

# Orientation

- First course in quantitative methods sequence
  - **575** $\rightarrow$ **576** $\rightarrow$ MLE $\rightarrow$ Topical$_1$ $\rightarrow$ Topical$_2$
- We will cover:
  - Basic probability
  - Variety of statistical tests
  - Linear regression
  - Causality and observational data

# Orientation

- First course in quantitative methods sequence
  - **575** $\rightarrow$ **576** $\rightarrow$ MLE $\rightarrow$ Topical$_1$ $\rightarrow$ Topical$_2$
- We will cover:
  - Basic probability
  - Variety of statistical tests
  - Linear regression
  - Causality and observational data
  - Statistical programming

# Orientation

- First course in quantitative methods sequence
  - **575** $\rightarrow$ **576** $\rightarrow$ MLE $\rightarrow$ Topical$_1$ $\rightarrow$ Topical$_2$
- We will cover:
  - Basic probability
  - Variety of statistical tests
  - Linear regression
  - Causality and observational data
  - Statistical programming
- How to succeed in this course

# How to succeed in this course

- My musings on winning at grad classes

# How to succeed in this course

- My musings on winning at grad classes
- More broadly: treat your PhD like a typical professional job

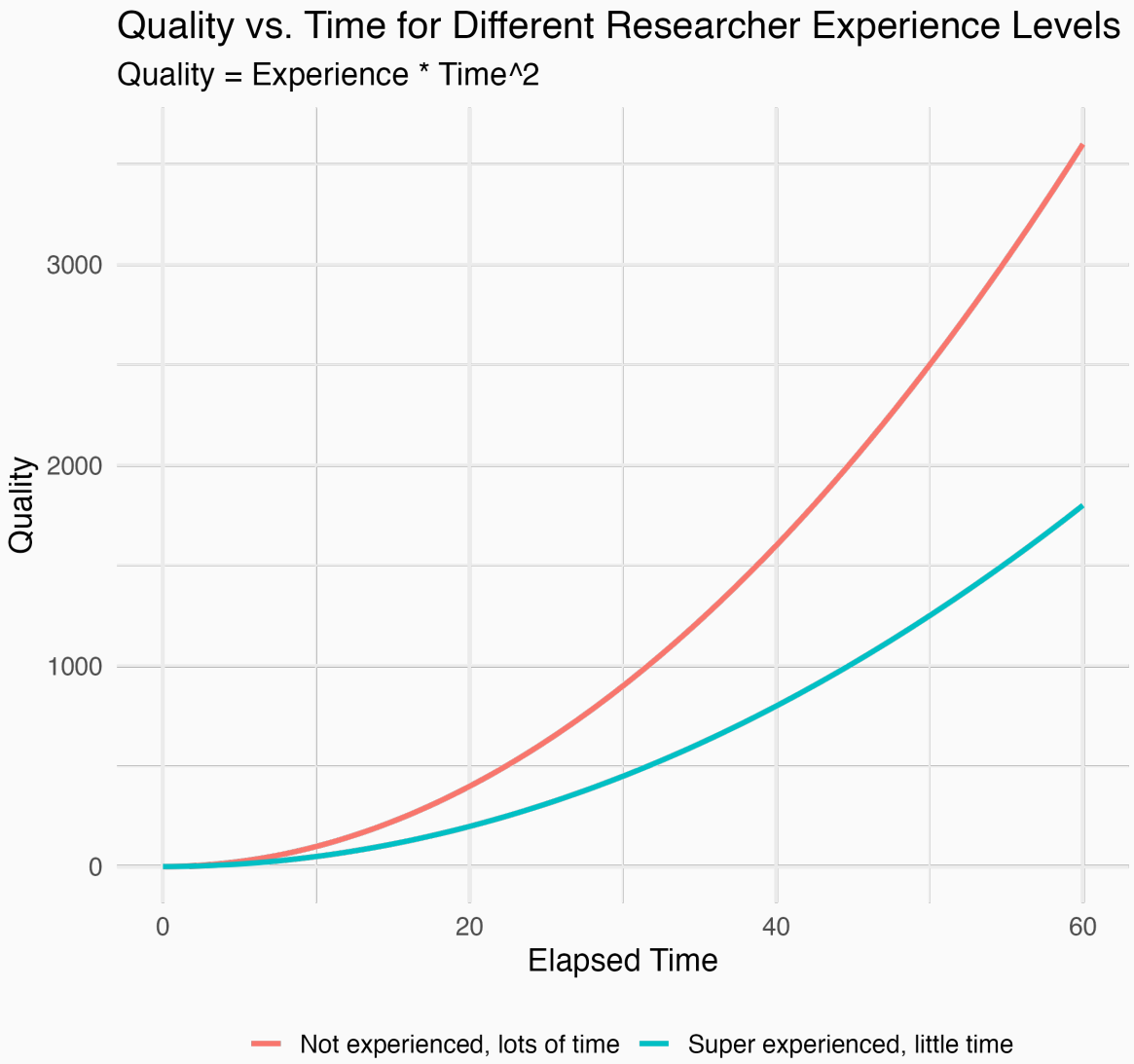
# How to succeed in this course

- My musings on winning at grad classes
- More broadly: treat your PhD like a typical professional job
- Balance mental health with ballooning academic demands (true of any professional ‘thinking’ job)

# How to succeed in this course

- My musings on winning at grad classes
- More broadly: treat your PhD like a typical professional job
- Balance mental health with ballooning academic demands (true of any professional ‘thinking’ job)
- Professionalization

# Importance of Effort for Research (and PhD-getting)



# Why Statistics

- Science without data analysis is impossible

# Why Statistics

- Science without data analysis is impossible
  - Observational studies

# Why Statistics

- Science without data analysis is impossible
  - Observational studies
  - Surveys, lab and field experiments

# Why Statistics

- Science without data analysis is impossible
  - Observational studies
  - Surveys, lab and field experiments
  - Text, audio, video, social networks

# Why Statistics

- Science without data analysis is impossible
  - Observational studies
  - Surveys, lab and field experiments
  - Text, audio, video, social networks
- Probability and statistics provide a standard language for formalizing quantitative data analysis, making claims about the world, and weighing evidence

# Why Statistics

- Science without data analysis is impossible
  - Observational studies
  - Surveys, lab and field experiments
  - Text, audio, video, social networks
- Probability and statistics provide a standard language for formalizing quantitative data analysis, making claims about the world, and weighing evidence
  - Rigorous measurement

# Why Statistics

- Science without data analysis is impossible
  - Observational studies
  - Surveys, lab and field experiments
  - Text, audio, video, social networks
- Probability and statistics provide a standard language for formalizing quantitative data analysis, making claims about the world, and weighing evidence
  - Rigorous measurement
  - Explicitly stated assumptions

# Why Statistics

- Science without data analysis is impossible
  - Observational studies
  - Surveys, lab and field experiments
  - Text, audio, video, social networks
- Probability and statistics provide a standard language for formalizing quantitative data analysis, making claims about the world, and weighing evidence
  - Rigorous measurement
  - Explicitly stated assumptions
  - Principled statements about certainty of conclusions

# Why Statistics

- Science without data analysis is impossible
  - Observational studies
  - Surveys, lab and field experiments
  - Text, audio, video, social networks
- Probability and statistics provide a standard language for formalizing quantitative data analysis, making claims about the world, and weighing evidence
  - Rigorous measurement
  - Explicitly stated assumptions
  - Principled statements about certainty of conclusions
- Statistical literacy now mandatory to be informed consumer of social science

# Why R

- R is a popular computer language

# Why R

- R is a popular computer language
  - Used in statistical analyses, data science, policy analysis, and more

# Why R

- R is a popular computer language
  - Used in statistical analyses, data science, policy analysis, and more
- Free to use (open source)

# Why R

- R is a popular computer language
  - Used in statistical analyses, data science, policy analysis, and more
- Free to use (open source)
- Vibrant community with new packages that enable tons of tasks and models

# Why R

- R is a popular computer language
  - Used in statistical analyses, data science, policy analysis, and more
- Free to use (open source)
- Vibrant community with new packages that enable tons of tasks and models
- Becoming the de facto programming language of political science

# Why Programming

- Quantitative and qualitative analyses rely on data, much of which is digital/digitized

# Why Programming

- Quantitative and qualitative analyses rely on data, much of which is digital/digitized
  - Managing data is difficult, often overlooked

# Why Programming

- Quantitative and qualitative analyses rely on data, much of which is digital/digitized
  - Managing data is difficult, often overlooked
  - We will learn *best practices*

# Why Programming

- Quantitative and qualitative analyses rely on data, much of which is digital/digitized
  - Managing data is difficult, often overlooked
  - We will learn *best practices*
- Enable massive tasks with minimal code

# Why Programming

- Quantitative and qualitative analyses rely on data, much of which is digital/digitized
  - Managing data is difficult, often overlooked
  - We will learn *best practices*
- Enable massive tasks with minimal code
  - Simplify mundane procedures

# Why Programming

- Quantitative and qualitative analyses rely on data, much of which is digital/digitized
  - Managing data is difficult, often overlooked
  - We will learn *best practices*
- Enable massive tasks with minimal code
  - Simplify mundane procedures
  - Analyses impossible without computers, e.g. 10+ gbs of textual data

# Why Programming

- Quantitative and qualitative analyses rely on data, much of which is digital/digitized
  - Managing data is difficult, often overlooked
  - We will learn *best practices*
- Enable massive tasks with minimal code
  - Simplify mundane procedures
  - Analyses impossible without computers, e.g. 10+ gbs of textual data
- Build up to more sophisticated methods and data collections procedures later

# Why Programming

- Quantitative and qualitative analyses rely on data, much of which is digital/digitized
  - Managing data is difficult, often overlooked
  - We will learn *best practices*
- Enable massive tasks with minimal code
  - Simplify mundane procedures
  - Analyses impossible without computers, e.g. 10+ gbs of textual data
- Build up to more sophisticated methods and data collections procedures later
  - *Learn* R

# Why Programming

- Quantitative and qualitative analyses rely on data, much of which is digital/digitized
  - Managing data is difficult, often overlooked
  - We will learn *best practices*
- Enable massive tasks with minimal code
  - Simplify mundane procedures
  - Analyses impossible without computers, e.g. 10+ gbs of textual data
- Build up to more sophisticated methods and data collections procedures later
  - *Learn* R
- Programming scripts enable **reproducible research**

# For Today

1. Questions?
2. Syllabus Review
3. Pre-test
4. Programming Lab

# Functions in R

- Sets of statements organized together to perform a specific task
- R has built-in functions and users can create their own functions
- Incredibly useful:
  - Provide a set of inputs, receive a consistent output
  - Allows Decreases repetition in code (and possibility of errors)
  - Easy to recycle
  - Easy to edit multiple calls (relative to repetition)
  - Protects the objects within it (more on that)

# Functions in R

You have been using functions! All the dplyr verbs:

```
dplyr::group_by(country,year)
```

And those base R calls as well:

```
sum(c(1:4))
```

But what are ‘arguments’ exactly? When do we need to name them?

# Parts of Functions

1. Function Name: Actual name of function/object. Stored in R environment.
2. Arguments: Fancy placeholders. When you use a function, you pass a value to the argument for function to use. Sometimes optional (if there is default), sometimes required.
3. Function Body: The collection of statements that defines what the function does.
4. Return value: The last expression in the function body to be evaluated or what is passed explicitly

# Function Formula

```
function(arglist){body}
```

```
function_name <- function(arguments){
  computations on the arguments
  some other code
  return(value)
}

# return without using return() explicitly
function_name <- function(arguments){
  computations on the arguments
  some other code
  value
}
```