Research and R

EC 607, Set 01

Edward Rubin

Prologue

Today

- Welcome, check in, and admin
- Research basics: Why are we here? MHE: Preface & Ch. 1
- Our class: What are we doing?
- R: Part of our how in this class: Basics.

Upcoming

- Learn more R: First assignment!
- Review metrics and building intution for causality and inference.
- Build momentum.

Long run

Goal: Deepen understandings/intuitions for causality and inference.

Why are we here?

• Econ. research

Why are we here?

• Econ. research: Understand human, social, and/or economic behaviors.

- Econ. research: Understand human, social, and/or economic behaviors.
- PhD

- Econ. research: Understand human, social, and/or economic behaviors.
- PhD: Learn methods, tools, skills, and intution required for research.

- Econ. research: Understand human, social, and/or economic behaviors.
- PhD: Learn methods, tools, skills, and intution required for research.
- (Applied) econometrics

- Econ. research: Understand human, social, and/or economic behaviors.
- PhD: Learn methods, tools, skills, and intution required for research.
- **(Applied) econometrics:** Build a toolbox of *empirical* methods, tools, and skills to that combine data and statistical insights to test and/or measure theories and policies.

- Econ. research: Understand human, social, and/or economic behaviors.
- PhD: Learn methods, tools, skills, and intution required for research.
- (Applied) econometrics: Build a toolbox of empirical methods, tools, and skills to that combine data and statistical insights to test and/or measure theories and policies.
- You

- Econ. research: Understand human, social, and/or economic behaviors.
- PhD: Learn methods, tools, skills, and intution required for research.
- (Applied) econometrics: Build a toolbox of empirical methods, tools, and skills to that combine data and statistical insights to test and/or measure theories and policies.
- You: You should be thinking about this question throughout your program/work/life.

- Econ. research: Understand human, social, and/or economic behaviors.
- PhD: Learn methods, tools, skills, and intution required for research.
- (Applied) econometrics: Build a toolbox of empirical methods, tools, and skills to that combine data and statistical insights to test and/or measure theories and policies.
- You: You should be thinking about this question throughout your program/work/life. Self awareness and mental health are important.

This class

For many of people, this course marks a big shift in how school works.

- You don't have a metrics qualifying exam.
- Grades are not super important.

This class

For many of people, this course marks a big shift in how school works.

- You don't have a metrics qualifying exam.
- Grades are not super important.

The material and tools are pivotal for **a lot** of what you will do in the future.

This class

For many of people, this course marks a big shift in how school works.

- You don't have a metrics qualifying exam. 🦥
- Grades are not super important.

The material and tools are pivotal for **a lot** of what you will do in the future.

Take responsibility for your education and career.

- Commit to spending the necessary time.
- Be proactive and curious.
- Go down rabbit holes.
- Ask questions.
- Learn.

What are we doing?

Q What is the difference between econometrics and data science?

What are we doing?

- **Q** What is the difference between econometrics and data science?
- \mathbf{Q}_{\vee} Is there anything special about *econometrics*?

What are we doing?

- What is the difference between econometrics and data science?
- \mathbf{Q}_{v2} Is there anything special about econometrics?

A_{1/∞} Causality. [™]

Sources for this Q and A: Dan Hammer and Max Auffhammer.

What are we doing?

- What is the difference between econometrics and data science?
- \mathbf{Q}_{v2} Is there anything special about econometrics?

A_{1/∞} Causality. [™]

Note: There are large parts of econometrics that focus on *prediction* rather than *causality* (*e.g.*, forecasting and prediction—see Jeremy Piger).[†]

Sources for this Q and A: Dan Hammer and Max Auffhammer.

† Also: Machine learning (e.g., my ML and econometrics course here at UO)

What are we doing?

- **Q** What is the difference between econometrics and data science?
- \mathbf{Q}_{v2} Is there anything special about econometrics?

A_{1/∞} Causality. [™]

Note: There are large parts of econometrics that focus on *prediction* rather than *causality* (*e.g.*, forecasting and prediction—see Jeremy Piger).[†]

Causality plays a *huge* role in modern applied econometrics (esp. in micro).

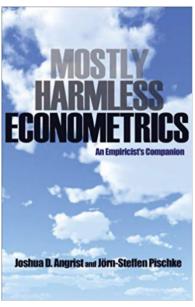
Sources for this Q and A: Dan Hammer and Max Auffhammer.

† Also: Machine learning (e.g., my ML and econometrics course here at UO)

Toward this end—causality—we will use two books (favoring MHE).

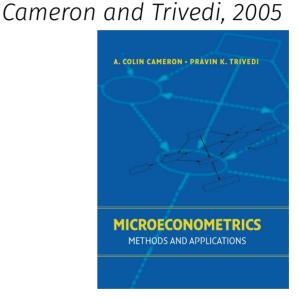
Mostly Harmless Econometrics

Angrist and Pischke, 2008



MHE Buy now. **Read this book.**The standard for causal metrics.
Could use an update.

Microeconometrics: Methods and Applications

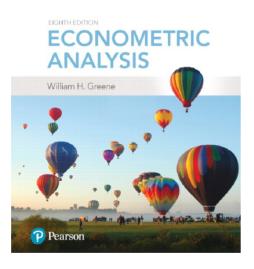


We will use more C&T than Greene.

While you're at it, buy one or two more...

Econometric Analysis

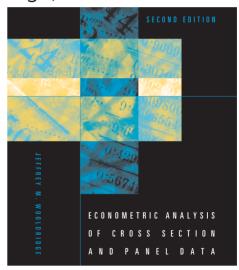
Greene, 2018



Encyclopedic reference.

Econometric Analysis of Cross Section and Panel Data

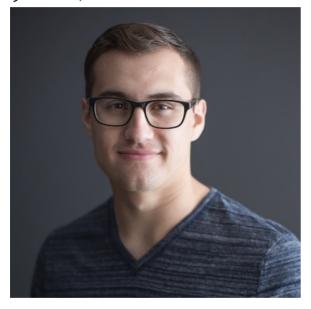
Wooldridge, 2010



This book has some great sections.

Two more "free" books...

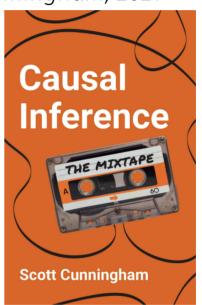
Introduction to Causal Inference Brady Neal, 2020



Under development but great. Targets folks from prediction.

Causal Inference: The Mixtape

Scott Cunningham, 2021



Relatively new.

Includes R, Stata, and Python code.

Motivation

First, we believe that empirical research is most valuable when it uses data to answer specific causal questions, as if in a randomized clinical trial. This view shapes our approach to most research questions. In the absence of a real experiment, we look for well-controlled comparisons and/or natural quasi-experiments. Of course, some quasi-experimental research designs are more convincing than others, but the econometric methods used in these studies are almost always fairly simple.

Mostly Harmless Econometrics, p. xii (color added)

Motivation

First, we believe that empirical research is most valuable when it uses data to answer specific causal questions, as if in a randomized clinical trial. This view shapes our approach to most research questions. In the absence of a real experiment, we look for well-controlled comparisons and/or natural quasi-experiments. Of course, some quasi-experimental research designs are more convincing than others, but the econometric methods used in these studies are almost always fairly simple.

Mostly Harmless Econometrics, p. xii (color added)

1. This ideology inherently compares research to "gold-standard" RCTs.

Motivation

First, we believe that empirical research is most valuable when it uses data to answer specific causal questions, as if in a randomized clinical trial. This view shapes our approach to most research questions. In the absence of a real experiment, we look for well-controlled comparisons and/or natural quasi-experiments. Of course, some quasi-experimental research designs are more convincing than others, but the econometric methods used in these studies are almost always fairly simple.

Mostly Harmless Econometrics, p. xii (color added)

- 1. This ideology inherently compares research to "gold-standard" RCTs.
- 2. The methods are usually (relatively) straightforward (after training).

Angrist and Pischke's FAQs[†]

- 1. What is the **causal relationship of interest**?
- 2. How would an **ideal experiment** capture this causal effect of interest?
- 3. What is your **identification strategy**?
- 4. What is your **mode of inference**?

Angrist and Pischke's FAQs[†]

- 1. What is the **causal relationship of interest**?
- 2. How would an **ideal experiment** capture this causal effect of interest?
- 3. What is your **identification strategy**?
- 4. What is your **mode of inference**?

Note: Other questions also matter for developing quality research, e.g., ††

- Why is your question important/interesting?
- Why is the current literature lacking or nonexistant?
- How do you propose to advance the literature?

[†] See MHE, chapter 1. †† Credit for these questions goes to Reed Walker.

FAQ₁: What is the causal relationship of interest?

Descriptive exercises can be very interesting and important, but in modern applied econometrics, causality is king.

Why?

FAQ₁: What is the causal relationship of interest?

Descriptive exercises can be very interesting and important, but in modern applied econometrics, causality is king.

Why?

- Causal relationships directly test theories of how the world works.
- Causal relationships provide us with *counterfactuals*—how the world would have looked with different sets of policies/circumstances.

FAQ₁: What is the causal relationship of interest?

Descriptive exercises can be very interesting and important, but in modern applied econometrics, causality is king.

Why?

- Causal relationships directly test theories of how the world works.
- Causal relationships provide us with *counterfactuals*—how the world would have looked with different sets of policies/circumstances.

If you can't clearly and succinctly name the causal relationship of interest, then you may not actually have a research project.

FAQ₁: What is the causal relationship of interest?

Some classic examples...

Labor and Education

How does an additional year of schooling affect wages?

Political Economy and Development

How do democratic institutions affect economic development?

Environment and Urban

Do the poor receive substantive benefits from environmental clean ups?

Health, Crime, and Law

Do gun-control laws actually reduce gun violence?

FAQ₂: What is the ideal experiment for this setting?

Describing the ideal experiment helps us formulate

- the exact causal question(s)
- the dimensions we want to manipulate
- the factors we need to hold constant

FAQ₂: What is the ideal experiment for this setting?

Describing the ideal experiment helps us formulate

- the exact causal question(s)
- the dimensions we want to manipulate
- the factors we need to hold constant

These *ideal experiments* are generally hypothetical, but if you can't describe the ideal, it will probably be hard to come up with data and plausible research designs in real life.

FAQ₂: What is the ideal experiment for this setting?

Describing the ideal experiment helps us formulate

- the exact causal question(s)
- the dimensions we want to manipulate
- the factors we need to hold constant

These *ideal experiments* are generally hypothetical, but if you can't describe the ideal, it will probably be hard to come up with data and plausible research designs in real life.

Angrist and Pischke call questions without ideal experiments fundamentally unanswerable questions (FUQs).

FAQ₂: What is the ideal experiment for this setting?

Examples of potentially answerable questions...

- The effect of education on wages: Randomize scholarships or incentives to remain in school.
- **Democracy and development:** Arbitrarily assign institutional types to countries as they receive independence.
- Environmental cleanups: Ask EPA to randomly clean toxic sites.
- **Gun laws:** Randomly assign gun restrictions to jurisdictions.

FAQ₂: What is the ideal experiment for this setting?

Examples of potentially answerable questions...

- The effect of education on wages: Randomize scholarships or incentives to remain in school.
- **Democracy and development:** Arbitrarily assign institutional types to countries as they receive independence.
- Environmental cleanups: Ask EPA to randomly clean toxic sites.
- **Gun laws:** Randomly assign gun restrictions to jurisdictions.

Examples of challenging questions to answer (potentially unanswerable?)...

- How does gender affect eventual career paths?
- What role does race play in one's wages?

FAQ₂: What is the ideal experiment for this setting?

Sometimes even simple-sounding policy questions turn out to be fundamentally unanswerable.

FAQ₂: What is the ideal experiment for this setting?

Sometimes even simple-sounding policy questions turn out to be fundamentally unanswerable.

Example of a fundamentally unanswerable question:

Do children perform better by starting school at an older age?

FAQ₂: What is the ideal experiment for this setting?

Sometimes even simple-sounding policy questions turn out to be fundamentally unanswerable.

Example of a fundamentally unanswerable question:

Do children perform better by starting school at an older age?

Proposed ideal experiment

- 1. Randomize kids to start 1st grade at age 6 or 7.
- 2. Compare 2nd grade test scores.

FAQ₂: What is the ideal experiment for this setting?

Sometimes even simple-sounding policy questions turn out to be fundamentally unanswerable.

Example of a fundamentally unanswerable question:

Do children perform better by starting school at an older age?

Proposed ideal experiment

- 1. Randomize kids to start 1st grade at age 6 or 7.
- 2. Compare 2nd grade test scores.

Problem

FAQ₂: What is the ideal experiment for this setting?

Sometimes even simple-sounding policy questions turn out to be fundamentally unanswerable.

Example of a fundamentally unanswerable question:

Do children perform better by starting school at an older age?

Proposed ideal experiment

- 1. Randomize kids to start 1st grade at age 6 or 7.
- 2. Compare 2nd grade test scores.

Problem Kids who started later are older in 2nd grade. Older kids do better. Do we want the effect of starting later or just being older?

FAQ₂: What is the ideal experiment for this setting?

Sometimes even simple-sounding policy questions turn out to be fundamentally unanswerable.

Example of a fundamentally unanswerable question:

Do children perform better by starting school at an older age?

Proposed ideal experiment_{2.0}

- 1. Randomize kids to start 1st grade at age 6 or 7.
- 2. Control for age. Compare test scores when kids are age 8.

FAQ₂: What is the ideal experiment for this setting?

Sometimes even simple-sounding policy questions turn out to be fundamentally unanswerable.

Example of a fundamentally unanswerable question:

Do children perform better by starting school at an older age?

Proposed ideal experiment_{2.0}

- 1. Randomize kids to start 1st grade at age 6 or 7.
- 2. Control for age. Compare test scores when kids are age 8.

Problem_{2.0}

FAQ₂: What is the ideal experiment for this setting?

Sometimes even simple-sounding policy questions turn out to be fundamentally unanswerable.

Example of a fundamentally unanswerable question:

Do children perform better by starting school at an older age?

Proposed ideal experiment_{2.0}

- 1. Randomize kids to start 1st grade at age 6 or 7.
- 2. Control for age. Compare test scores when kids are age 8.

Problem_{2.0} The two groups will have been in school for different numbers of years (1 vs. 2). More school should mean better scores.

FAQ₂: What is the ideal experiment for this setting?

Sometimes even simple-sounding policy questions turn out to be fundamentally unanswerable.

Example of a fundamentally unanswerable question:

Do children perform better by starting school at an older age?

Central problem: Mechanical links between ages and time in school.

(Start Age) = (Current Age) – (Time in School)

No experiment can separate these effects (for school-age children).

FAQ₃: What's your identification strategy?

This question describes how you plan to recover/observe as good as random assignment of your variable of interest (approximating your ideal experiment) in real life.

Examples

- Compulsory school-attendance laws interacted with quarter of birth
- Vietnam War draft
- Thresholds for the Clean Air Act violations
- Notches in income-tax policies
- Judge assignments
- Randomly assigned characteristics on résumés

FAQ₃: What's your identification strategy?

A brief history

The term "identification strategy" goes back to Angrist and Krueger (1991).

However, the comparison of *ideal* and *natural* experiments goes back much farther to Haavelmo (1944)...

A design of experiments... is an essential appendix to any quantitative theory. And we usually have some such experiment in mind when we construct the theories, although-unfortunatelymost economists do not describe their design of experiments explicitly. If they did, they would see that the experiments they have in mind may be grouped into two different classes, namely, (1) experiments that we should like to make to see if certain real economic phenomena—when artificially isolated from "other influences"—would verify certain hypotheses, and (2) the stream of experiments that Nature is steadily turning out from her own enormous laboratory, and which we merely watch as passive observers. In both cases the aim of the theory is the same, to become master of the happenings of real life.

Haavelmo, 1944 (color added)

FAQ₄: What is your mode of inference?

Historically, inference—standard errors, confidence intervals, hypothesis tests, *etc.*—has received much less attention than point estimates. It's becoming more important (more than an afterthought).

- Which population does your sample represent?
- How much **noise** (error) exists in your estimator (and estimates)?
- How much variation do you actually have in your variable of interest?

FAQ₄: What is your mode of inference?

Historically, inference—standard errors, confidence intervals, hypothesis tests, *etc.*—has received much less attention than point estimates. It's becoming more important (more than an afterthought).

- Which population does your sample represent?
- How much **noise** (error) exists in your estimator (and estimates)?
- How much variation do you actually have in your variable of interest?

Without careful inference, we don't know the difference between

- 21% ± 2.3%
- 21% ± 20.3%

Our class

Our class

Mini-syllabus

Class Attend/participate. Read assigned readings—especially papers.

Lab Practice applying our in-class content in R with Jaichung/me. Attend.

Problem sets 3+ problem sets mixing theory and applications in R.

Other grades Project plus take-home final.

R

What is it?

The R project website:

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS.

What is it?

The R project website:

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS.

What does that mean?

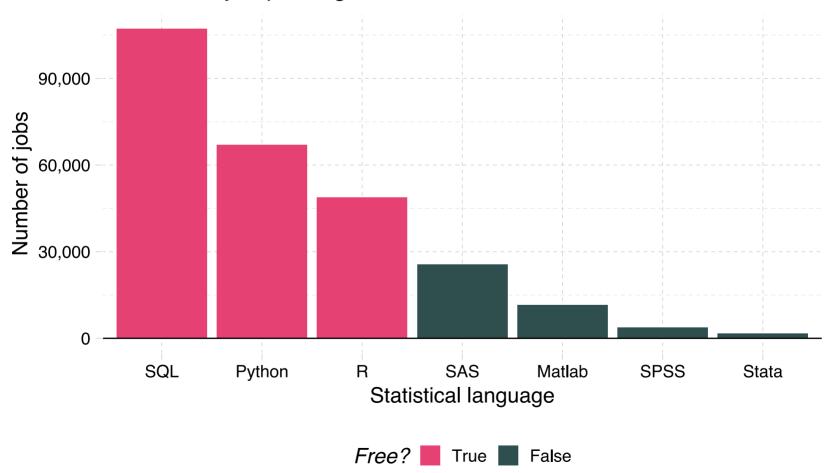
- R was created for the statistical and graphical work required by econometrics.
- R has a vibrant, thriving online community (e.g., Stack Overflow).
- Plus it's free and open source.

Why are we using R?

- 1. R is **free** and **open source**—saving both you and the university 💰 🛂 💰 .
- 2. *Related:* Outside of a small group of economists, private- and public-sector **employers favor R** over Stata and most competing softwares.
- 3. R is very **flexible and powerful**—adaptable to nearly any task, *e.g.*, 'metrics, spatial data analysis, machine learning, web scraping, data cleaning, website building, teaching. My website, the TWEEDS website, and these notes all came out of R.

Comparing statistical languages

Number of job postings on Indeed.com, 2019/01/06



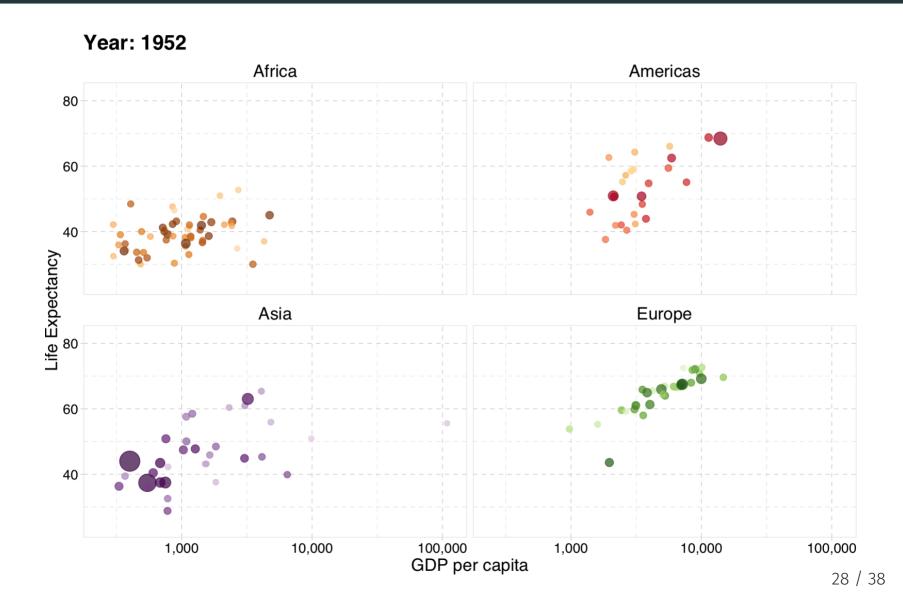
Why are we using R?

4. *Related:* R imposes **no limitations** on your amount of observations, variables, memory, or processing power. (I'm looking at **you**, Stata.)

5. If you put in the work, you (and your students!) will come away with a **valuable and marketable** tool.

6. I 💚 🖪

[:] Learning R definitely requires time and effort.



The install

Installing R is fairly straightfoward, but it occasionally involves challenges for older computers.

Step 1: Download (r-project.org) and install R for your operating system.

Step 2: Download (rstudio.com) and install RStudio Desktop for your operating system.

DataCamp has a nice tutorial on installing R and RStudio for Windows, Mac, and Linux operating systems.[†]

[†] I applied for free access to DataCamp for our class. I'll let you know when I hear back.

Fundamentals

Let's get started. There are a few principals to keep in mind with R:

Fundamentals

Let's get started. There are a few principals to keep in mind with R:

1. Everything is an **object**.

Fundamentals

Let's get started. There are a few principals to keep in mind with R:

1. Everything is an **object**.

foo

Fundamentals

Let's get started. There are a few principals to keep in mind with R:

1. Everything is an **object**.

foo

2. Every object has a **name** and **value**.

Fundamentals

Let's get started. There are a few principals to keep in mind with R:

1. Everything is an **object**.

foo

2. Every object has a name and value.

foo \leftarrow 2

Fundamentals

Let's get started. There are a few principals to keep in mind with R:

1. Everything is an **object**.

foo

2. Every object has a **name** and **value**.

foo = 2

Fundamentals

Let's get started. There are a few principals to keep in mind with R:

1. Everything is an **object**.

foo

2. Every object has a name and value.

foo = 2

3. You use **functions** on these objects.

Fundamentals

Let's get started. There are a few principals to keep in mind with R:

1. Everything is an **object**. foo

2. Every object has a **name** and **value**. foo = 2

3. You use **functions** on these objects. mean(foo)

Fundamentals

Let's get started. There are a few principals to keep in mind with R:

1. Everything is an **object**.

foo

2. Every object has a name and value.

foo = 2

3. You use **functions** on these objects.

mean(foo)

4. Functions come in **libraries** (packages)

Fundamentals

Let's get started. There are a few principals to keep in mind with R:

1. Everything is an **object**. foo

2. Every object has a **name** and **value**. foo = 2

3. You use **functions** on these objects. mean(foo)

4. Functions come in **libraries** (**packages**) library(dplyr)

Fundamentals

Let's get started. There are a few principals to keep in mind with R:

1. Everything is an **object**.

foo

2. Every object has a name and value.

foo = 2

3. You use **functions** on these objects.

mean(foo)

4. Functions come in **libraries** (packages)

library(dplyr)

5. R will try to **help** you.

Fundamentals

Let's get started. There are a few principals to keep in mind with R:

1. Everything is an **object**. foo

2. Every object has a **name** and **value**. foo = 2

3. You use **functions** on these objects. mean(foo)

4. Functions come in **libraries** (**packages**) library(dplyr)

5. R will try to **help** you. ?dplyr

Fundamentals

Let's get started. There are a few principals to keep in mind with R:

1. Everything is an **object**. foo

2. Every object has a **name** and **value**. foo = 2

3. You use **functions** on these objects. mean(foo)

4. Functions come in **libraries** (**packages**) library(dplyr)

5. R will try to **help** you. ?dplyr

6. R has its quirks.

Fundamentals

Let's get started. There are a few principals to keep in mind with R:

1. Everything is an **object**. foo

2. Every object has a **name** and **value**. foo = 2

3. You use **functions** on these objects. mean(foo)

4. Functions come in **libraries** (**packages**) library(dplyr)

5. R will try to **help** you. ?dplyr

6. R has its **quirks**. NA; error; warning

Fundamentals of functions

Functions operate on objects, but they need some guidance—arguments.

Fundamentals of functions

Functions operate on objects, but they need some guidance—arguments.

Example: ex_fun(arg1, arg2, arg3)

Fundamentals of functions

Functions operate on objects, but they need some guidance—arguments.

```
Example: ex_fun(arg1, arg2, arg3)
```

- Our function is named ex_fun.
- This function takes three arguments: arg1, arg2, arg3.
- You can tell R which values to assign to which arguments:
 ex_fun(arg1 = 13, arg2 = 25, arg3 = 7) (probably best practice)
- ... or R will assign the values using the arguments' defined order:
 ex_fun(13, 25, 7) (shorter/lazier but has the same result)
- You must assign a name to a function's outputted object (to keep it).

Example function: matrix

We will need to create matrices in this class.

Enter: R's matrix() function!

```
# 3×2 matrix filled w/ zeros
matrix(
  data = 0, nrow = 3, ncol = 2
)
```

```
#> [,1] [,2]

#> [1,] 0 0

#> [2,] 0 0

#> [3,] 0 0
```

```
# 3×2 matrix filled w/ zeros
matrix(
  data = 0, nrow = 3, ncol = 2
)
```

```
#> [,1] [,2]

#> [1,] 0 0

#> [2,] 0 0

#> [3,] 0 0
```

```
# 3×2 matrix filled w/ 1 to 6
matrix(
  data = 1:6, nrow = 3, ncol = 2
)
```

```
#> [,1] [,2]
#> [1,] 1 4
#> [2,] 2 5
#> [3,] 3 6
```

```
# 3×2 matrix filled w/ 1:6 by row
matrix(
  data = 1:6, nrow = 3, ncol = 2,
  byrow = T
)
```

```
#> [,1] [,2]
#> [1,] 1 2
#> [2,] 3 4
#> [3,] 5 6
```

```
# 3×2 matrix filled w/ 1:6 by row
matrix(
  data = 1:6, nrow = 3, ncol = 2,
  byrow = T
)
```

```
# 3×2 matrix filled w/ 1:3
matrix(
  data = 1:3,
  nrow = 3, ncol = 2
)
```

```
#> [,1] [,2]
#> [1,] 1 2
#> [2,] 3 4
#> [3,] 5 6
```

```
#> [,1] [,2]
#> [1,] 1 1
#> [2,] 2 2
#> [3,] 3 3
```

```
# 3×2 matrix filled w/ 1:6 by row
matrix(
  data = 1:6, nrow = 3, ncol = 2,
  byrow = T
)
```

```
#> [,1] [,2]
#> [1,] 1 2
#> [2,] 3 4
#> [3,] 5 6
```

```
# 3×2 matrix filled w/ 1:3
# Assigned to memory
our_matrix ← matrix(
  data = 1:3,
  nrow = 3, ncol = 2
)
```

```
# 3×2 matrix filled w/ 1:3
matrix(
  data = 1:3,
  nrow = 3, ncol = 2
)
```

```
#> [,1] [,2]
#> [1,] 1 1
#> [2,] 2 2
#> [3,] 3 3
```

Help and functions

Q How do we know which arguments a function requires/accepts?

Help and functions

Q How do we know which arguments a function requires/accepts?

A ?

Help and functions

Q How do we know which arguments a function requires/accepts?

A ? Meaning you can type ?matrix into your R console to find the help file associated with the functions/objects named matrix.

Help and functions

Q How do we know which arguments a function requires/accepts?

A ? Meaning you can type ?matrix into your R console to find the help file associated with the functions/objects named matrix.

Double bonus: Use ??matrix to perform a fuzzy search for the term matrix in all of the help files.

Example function: matrix

Q How do we know which arguments a function requires/accepts?

A₂ RStudio will also try to help you.

- Type a name (e.g., matrix) into the console; RStudio will show you some info about the function.
- After you type the name and parentheses (e.g., matrix()), press tab, and RStudio will show you a list of arguments for the function.

Table of contents

Admin

- 1. Schedule
- 2. Mini-syllabus

Research

- 1. Why are we here?
- 2. MHE's FAQs
 - 1. Question
 - 2. Experiment
 - 3. Identification
 - 4. Inference

R

- 1. Basics
- 2. Install
- 3. Fundamentals