Capstone Advice for Your Outline Submissions

ECON 490

Taylor Mackay || Email: tmackay@fullerton.edu

Slides Overview

In these slides, we’ll discuss:

- General advice on using R and other data analysis tools
- Suggestions for your capstone analysis with examples in R

Don’t Reinvent the Wheel

Data analysis activities and R handouts are all written very intentionally

- Goal of these activities is to prepare you for capstone analysis!
- 90+ percent of coding issues I see have been covered in class previously

Review old coding activities to remind yourself what you know how to do!

If you’re using R (and not proficient with it) pull out copies of the following:

- Introduction to Tidyverse handout
- Regression Review Pt. II

A Very Gentle Bit of Advice (Pt. 1)

Students often ask if they can use Excel / Python / whatever for their project

- If general, yes, you can (especially if you use it for work)
- Doing so means I can't directly share code (but tradeoffs might be worth it)

In my experience, students who have difficulty using R also tend to have difficulty with other programs

Disclaimer: If you can regularly / confidently use PivotTables or Excel macros, this doesn't apply to you! More generally, if you've had a "Neo in the Matrix" moment

A Very Gentle Bit of Advice (Pt. 2)

We are using R at a relatively basic, entry level

It is very likely that if you are having trouble, the issue isn't with using R specifically

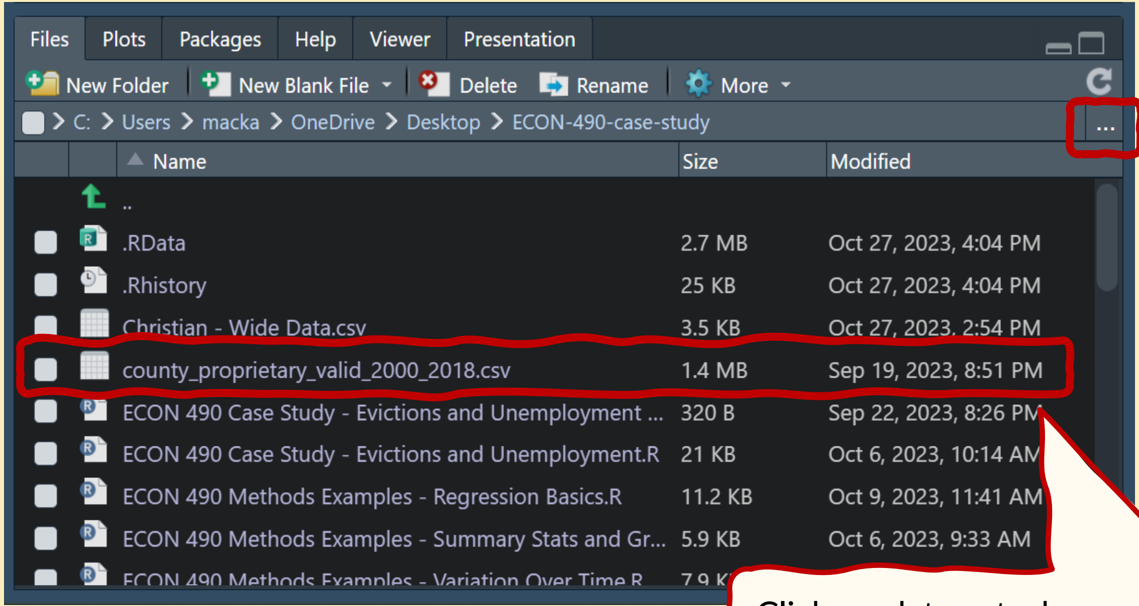
- Instead, it's often a conceptual issue – can you precisely state your goal?
- Does the regression you're trying to run make sense?
- Does the data set you're trying to put together let you run that regression?

If your goal isn't clear, and you swap to Python or Excel… you still don't have a clear goal

My advice – 1) get clear on your goals and 2) make sure you're leveraging class resources

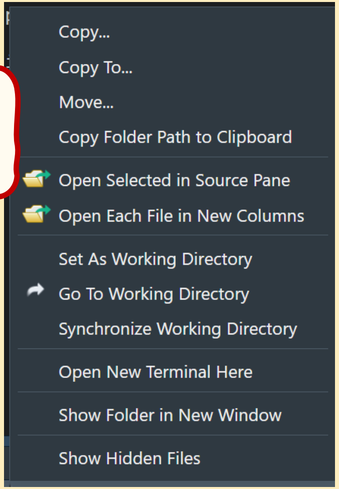
Loading Data into R

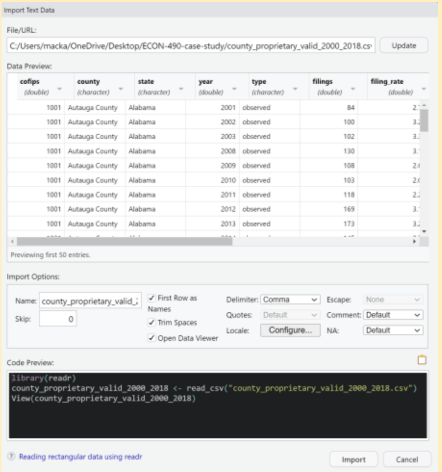
Click …, load folder with your data and set working directory



C: > Users > macka > OneDrive > Desktop > ECON-490-case-study

Name	Size	Modified
..		
.RData	2.7 MB	Oct 27, 2023, 4:04 PM
.Rhistory	25 KB	Oct 27, 2023, 4:04 PM
Christian - Wide Data.csv	3.5 KB	Oct 27, 2023, 2:54 PM
county_proprietary_valid_2000_2018.csv	1.4 MB	Sep 19, 2023, 8:51 PM
ECON 490 Case Study - Evictions and Unemployment ...	320 B	Sep 22, 2023, 8:26 PM
ECON 490 Case Study - Evictions and Unemployment.R	21 KB	Oct 6, 2023, 10:14 AM
ECON 490 Methods Examples - Regression Basics.R	11.2 KB	Oct 9, 2023, 11:41 AM
ECON 490 Methods Examples - Summary Stats and Gr...	5.9 KB	Oct 6, 2023, 9:33 AM
ECON 490 Methods Examples - Variation Over Time.R	7.9 K	


Copy...
Copy To...
Move...
Copy Folder Path to Clipboard
Open Selected in Source Pane
Open Each File in New Columns
Set As Working Directory
Go To Working Directory
Synchronize Working Directory
Open New Terminal Here
Show Folder in New Window
Show Hidden Files


Import Text Data

File/URL: C:/Users/macka/OneDrive/Desktop/ECON-490-case-study/county_proprietary_valid_2000_2018.cs

Data Preview:

cofips (double)	county (character)	state (character)	year (double)	type (character)	filings (double)	filing_rate (double)
1001	Autauga County	Alabama	2001	observed	84	2.
1001	Autauga County	Alabama	2002	observed	100	3.2
1001	Autauga County	Alabama	2003	observed	102	3.3
1001	Autauga County	Alabama	2008	observed	130	3.1
1001	Autauga County	Alabama	2009	observed	108	2.6
1001	Autauga County	Alabama	2010	observed	103	2.6
1001	Autauga County	Alabama	2011	observed	118	2.2
1001	Autauga County	Alabama	2012	observed	169	3.1
1001	Autauga County	Alabama	2013	observed	173	3.2

Previewing first 50 entries.

Import Options:

Name: county_proprietary_valid_2

Skip: 0

☑ First Row as Names

☑ Trim Spaces

☑ Open Data Viewer

Delimiter: Comma

Quotes: Default

Locale: Configure...

Escape: None

Comment: Default

NA: Default

Code Preview:

```
library(readr)
county_proprietary_valid_2000_2018 <- read_csv("county_proprietary_valid_2000_2018.csv")
View(county_proprietary_valid_2000_2018)
```

Reading rectangular data using readr

Click on data set, choose "Import Data," to preview & load data

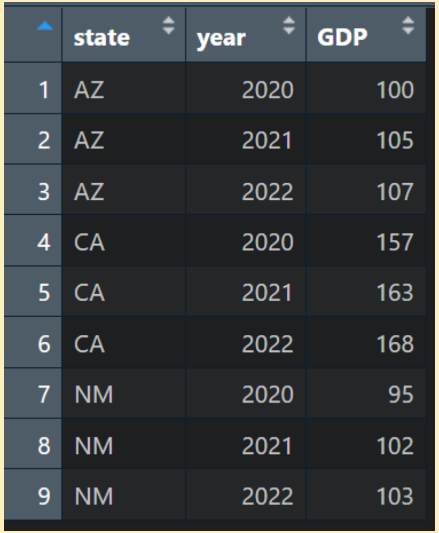
Wide vs. Long-Formatted Data

Two types of data structures:

- ***Wide*** data has same variable in multiple columns (here, GDP is split by year)
- ***Long*** data has each variable in one column

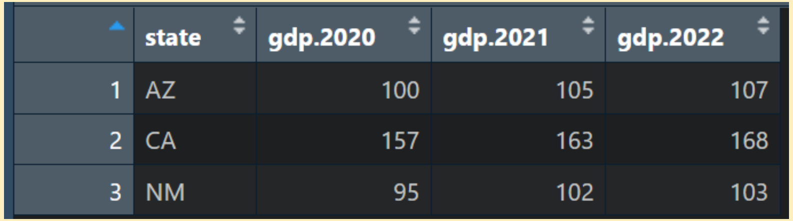
In almost all cases, data should be in long format for running regressions

- In R, use `pivot_longer()`



	state	year	GDP
1	AZ	2020	100
2	AZ	2021	105
3	AZ	2022	107
4	CA	2020	157
5	CA	2021	163
6	CA	2022	168
7	NM	2020	95
8	NM	2021	102
9	NM	2022	103

Long-formatted data



	state	gdp.2020	gdp.2021	gdp.2022
1	AZ	100	105	107
2	CA	157	163	168
3	NM	95	102	103

Wide-formatted data

Who’s in my Regression Sample?

Key question before running any regression – who gets included?

- Sometimes, this answer is straightforward (esp. with state or county data)
- Matters a lot with individual data like the CPS (or business-level data, etc.)

Data sets like the CPS have kids and retired folks included

- You might not want them in a regression exploring union membership!
- Use the `filter()` function to restrict sample to observations you want

Describing sample restrictions is a ***key*** part of interpreting regression output

Splitting Your Data

In general, your working data should be a ***single*** data set

- You can use `filter()` to explore a subsample of people in your data…
- But always make sure you're clear on ***why*** you're doing this!

General rule – splitting your data and running "symmetric" regressions probably shouldn't be your first approach

Let's say you have state-year data on gas prices and car sales for 2023 and 2024

- Your default strategy should be to use ***all*** data in `lm(sales ~ prices)`
- Worried about trends in both variables? Start by including `as.factor(year)`

R Code Example

R code file shows an example of splitting your sample using filter(), then shows how to run an interaction regression as an alternative to splitting your data by group (in the example, by state)

Outliers and “Weird” Observations

Always check your data with `summary()` and / or `table()`

- This lets you catch values that don’t look right
- Do you see outliers or repeated instances of the same random value?

Dealing with these situations requires background knowledge about your data

- Are extreme values feasible or plausible for a given variable?
- Do you have unusually coded missing values (e.g., -99, 999, 1000)?

As a general rule, be careful removing outliers if you think they’re real

- If you remove values or rows, how does this change regression interpretation?

R Code Example

R code file shows examples of using summary() and table() to check variables, then describes top-coding for income data

R-Squared (aka R^2)

Economic outcomes are high-dimensional (lots of stuff matters!)

R^2 tells us how much of the variation in Y we can explain with our Xs

- This matters if we want to predict Y…
- But less important if we’re interested in assessing how a specific X impacts Y

For most descriptive projects, it’s easy to increase R^2 without doing anything substantively interesting

- Try this yourself – add FEs, other X variables, etc. and see what happens
- Your R^2 might go up… does the answer to your research question change?

R Code Example

R code file shows an example of how adding FE’s can change both your regression interpretation and R2

“Regression-Level” Things to Include in Interpretation

We’ve talked a lot about interpreting βs – “regression-level” context matters too

Things to when include interpreting regression output:

- Sample restrictions *(Is anyone excluded?)*
- Missing values *(Are there people you couldn’t include? Why are they missing?)*
- Outliers *(Did you remove anyone for being an outlier? Why were they an outlier?)*

Things you don’t need to worry about (unless you’re doing prediction):

- R^2, other model-level “goodness of fit” tests

Is Doing X Okay?

A common question I get is, “Is it okay if I just do X for my 2nd stage”

- This is very hard to know in advance
- Recurring theme of slides – explore variation / find patterns / etc.

These are all ways of saying, “Explore your data”

- If you’re doing a descriptive project, you should be looking at a lot of results
- Try different conditional summary stats, regressions, graphs, etc.

Try running at least 3-4 variations of any regression – what differences matter?

Disclaimer: Technically, in some contexts (esp. causal inference), we want to avoid this (it’s known as multiple testing). For this class, however, the goal is to learn how to explore data and use regression. The best way to do that is (not surprisingly) exploring data and running lots of regressions!