Capstone Data Analysis

ECON 490

Email: tmackay@fullerton.edu

Overview

In these slides, begin by talking about:

- The Iron Law of data cleaning and data analysis
- A quick recap of the capstone introduction slides

Then we'll talk about:

- Specific requirements for your capstone data analysis
- Several examples of data analysis strategies

Goal – set you up for completing Capstone Proposals

The Iron Law of Data Cleaning and Data Analysis

*Actually doing data cleaning and analysis is **always** more difficult (time-consuming, complex, etc.) than thinking about doing data cleaning or analysis*

Learning how to do data analysis is all about recognizing limitations:

- What can my data really tell me?
- What can I actually say with a particular regression?
- Does this data analysis really let me address my research question?

Something to remember – you can ***always*** revise your research question later!

Common Comments from Students

Number 1 (by far): “I wish I left myself more time for cleaning up my data.”

Other Common Comments:

- “The analysis I proposed seemed a lot easier in my head…”
- “I wish I paid closer attention to variable definitions / documentation.”
- “I had to adjust my research question once I started trying to run regressions.”

Capstone Research and Data Proposal

First capstone assignment is due Week 9 (the week of Monday, 3/17)

Steps to complete this assignment:

1) Decide if you want to work alone or with a partner
2) Identify a clear research question
3) Identify the data you'll use to answer your research question
4) Complete the proposal worksheet on Canvas

NOTE: *Before moving forwards with your project, you will need approval on the topic and data source(s) – I may ask you to submit a revised proposal.*

Working Backwards

Do ***not*** start by thinking of a research question – instead, ***work backwards***

Start thinking about general topics of interest (i.e., the environment, crime, etc.)

- Then ask, “What kind of data analysis can I do?”
- Your data analysis requires (1) data and (2) an analysis strategy

Start by (1) identifying and exploring data that’s available related to your topic

- Then, think about (2) possible analysis strategies
- Finally, develop a research question that ***sets the stage*** for your analysis

General Suggestions on Research Questions

Your first draft of a research question will likely be too superficial

- Remember, the goal is to write a 12+ page paper
- As part of this, you'll spend 2-3 pages talking about your data analysis output

If your research question is something like, "What's the correlation between GDP and unemployment?" you're not going to have much to say!

Getting your research question right is a trial-and-error process

- Your research question will likely change as your project evolves
- That is not a problem (and is likely a good sign!)

Two Components for Your Capstone Data Analysis

As part of your capstone paper, you'll include the following output:

Summary Output – *what does your data look like?*

- Summary stats table for outcome + explanatory variable(s)
- At least 2 graphs or maps for your outcome + explanatory variable(s)

Results Output – *results in (at least) two stages*

- 1st stage is a "first pass" answer to your research question
- 2nd stage builds on and develops 1st stage along some dimension
- For most projects, this entails 2+ regression tables (some projects may differ)

Deliverables for Final Capstone Project

In addition to the output described on the last slide, you'll also need to write about your analysis output in the "Results" section of your final paper

We'll talk about writing later in the semester – key point now is ***telling a story***

- Your "Results" section will be 2-3 pages long
- Benchmark your planned analysis against this – *Do I have enough to talk about?*

Can you say something more interesting than, "A 1-unit change in X…"

- What do you results tell us about the world? What context do we need?

Data Analysis Starting Point – Your Working Data Set

This is a ***general*** guide (some papers will have a different setup)

- Download data set(s) you want to use ("raw data") and clean it up
- Create a ***single*** working data set for your regressions (what you use in `lm()`)

Your ***working data set*** should have:

- Outcome variable + key explanatory variable(s)
- "Control variables" (if needed) – X variables you include but don't emphasize
- ID variables – uniquely ID each row (might be a combination of variables)

Stages of Your Analysis

Proposal asks for 2 regressions you'd like to run to answer your research question

- Starting points for what will become the 2 stages of your analysis
- As mentioned before, this is likely to change!

The 1st regression serves as your starting point – its your "first pass" answer

- The 2nd regression builds on the 1st and provides a more thorough answer
- What this means in practice depends on your specific topic, data, etc.

For your final project, each stage will likely feature ***multiple*** regressions

Thinking Descriptively

How to go beyond, “What’s the correlation between X and Y?”

- We’ll talk about some analysis strategy examples
- Your specific topic, interests, data, etc. will dictate best approaches

In general, think about ***exploring variation*** in either:

1. The level of one economic variable as a function of other variables
2. A relationship between economic variables

In both cases, you can explore variation across time, places, economic factors, etc.

- Once you’ve done this, ***then*** use research question to set the stage for analysis

Data Analysis Examples – Pt. 1

Suppose you're interested in housing construction and home prices

- Use ACS state-level + Zillow home price data as a starting point
- Merge home construction data from St Louis FRED ("housing units started")

1st Stage: $Home\ Price_{st} = \beta_0 + \beta_1 Housing\ Starts_{st} + \beta_2 X_{st} + u_{st}$

2nd Stage: $Home\ Price_{st} = \beta_0 + \beta_1 Housing\ Starts_{st} + \beta_2 Housing\ Starts_{st} \times Region_s + X_{st} + \gamma_s + \theta_t + u_{st}$

In 2nd stage, explore variation in relationship across regions of the country using region interaction, in a regression with state and year fixed effects

Example 1 – *Home Prices and Housing Starts*

Two regressions from prior slide fulfills proposal requirements

- Are they sufficiently interesting for our final paper?
- Maybe … but probably not! Next steps?

One option – think about additional explanatory variables

- Does public policy matter? Changes to building-related regulations?
- Can you proxy for changes in demand? I.e., migration data from ACS?

Another option – spend some time talking about what other papers have found

- Helpful when you've found an interesting pattern, but you feel "stuck"
- Add a "Discussion" section after "Results" – put your results in context

Data Analysis Examples – Pt. 2

Suppose you're interested in how the Pandemic affected wages

- Let's focus on the kinds of jobs you're most likely to work as an econ major
- Use individual-level CPS data as a starting point and identify occupations of interest

1st Stage: $Earnings_{it} = \beta_0 + \sum_{k \neq 1}^{N} \beta_{k-1} I(occupation = k) + X_{it} + u_{it}$

2nd Stage: $Earnings_{it} = \beta_0 + \beta_1 I(Post - Pandemic_t) + \sum_{k \neq 1}^{N} \beta_{k-1} I(Post - Pandemic_i) \times I(occupation = k) + X_{it} + u_{it}$

Start by calculating average earnings across occupations of interest, then see how they changed post-Pandemic via interaction in 2nd stage

Example 2 – *COVID and Employment Outcomes*

Just like 1st example, we've got a starting point here… but what else can we do?

- One easy option = consider other outcome variables
- E.g., do people work remotely more? Are they more likely to move?

Big advantage of ACS + CPS + other microdata sets = lots of potential outcomes!

Remember we need to merge external data (or use an advanced analysis strategy)

- We could gather data on COVID cases (or fatalities, etc.)
- Use that as a variation of 2nd regression to control for pandemic intensity

Things to Note (Pt. 1)

Examples from prior slides are ***starting points***

- Your proposal asks for two regressions
- What if you run them and they don't say anything interesting?

Try to give yourself room for adjustment – what does this mean?

- More explanatory variables = easier to explore variation = find interesting patterns
- More observations ***(across groups)*** = helps you find ***statistically significant*** results

More time spent collecting and cleaning data now = much easier analysis later

Things to Note (Pt. 2)

Common question – what if I add an X variable (or swap Y)? Is that a new stage?

- Suppose your first stage `wage ~ education`,
- Second stage is then `wage ~ education + age`
- This doesn't count as a new stage** – and it's not making your life easier!

Remember, examples here refer to single regressions for each stage

- More output = more to talk about in "Results" section = your life is easier
- That likely means having ***multiple*** regressions per stage
- Look out for easy extensions (swap outcomes, use interactions, etc.)

****NOTE:** Obviously, context dependent (maybe new X variable is from a new / messy data set, etc.)

Canvas Proposal Example

Suppose you're interested in migration – *why do people live where they live?*

From the data sets on Canvas, CPS data is best-suited for this question

- Individual-level data on migration choices – *Did you move last year? From where?*

Possible analysis strategies

- Baseline regression – *relationship between migration and education + income*
- Exploring variation – *differences across regions of the country? During Pandemic*?

After doing the above, tailor your research question to the analysis you're envisioning

Several Additional Examples

Relationship b/w state public health spending health and labor market outcomes

- Use individual-level CPS data – health status, missed work due to illness, etc.
- Merge state-level public health spending data from SHADAC (link on Canvas)

Relationship b/w state-level Pandemic assistance provided by federal government and state-level economic outcomes

- Use state-level ACS data + merge Pandemic spending data from USASpending.gov

Relationship between state-level criminal justice policies and crime rates

- Use state-level ACS data + merge supplemental state crime data
- Identify a ***specific*** policy implemented in ***multiple*** states – try starting here

An Example Using Data Sets on Canvas

Suppose we're interested in the relationship between unemployment and crime

- We can estimate this general relationship using our state-level data…
- How do we explore things further?

One option – see how this relationship differs across regions of the country

- Is crime more responsive to economic conditions in certain places?

NOTE: *I picked this example to setup some practice with R – for your capstone analysis, you'll want to be able to "tell a story" about your results!*