Data Translation Challenge

Nick Huntington-Klein

Updated 2023-12-20

Drawing Insights from Data

This document will describe the Data Translation Challenge for the term. This is a group project that you will complete with 3-4 other people. In the course of the project you will:

- Explore a data set
- Design analyses to answer questions of business importance
- Evaluate different research designs and models
- Write up and present your results

The Task

You work for a company in the retail sector. Your company knows how well they're weathering the pandemic, but they are having difficulty figuring out what the effect has been on the rest of the retail sector and the rest of the economy. Everyone is being tight with their numbers!

So, they've brought you on and pointed you towards the IPUMS release of the Current Population Survey, which the government often uses to understand changes in employment.

Why employment data? **They figure that revenues can be misrepresented, or might be affected by government aid. But employment in an industry can tell you a lot about how well that industry is doing!**

They don't know exactly what analyses they want you to run or variables they want you to use - figuring that out is your job. But they know the general questions they'd like to be answered, each of which will probably require more than one regression analysis. 2020 - 2022.04; EMPSTAT; IND exclude not in labor force data

1. How has COVID affected the health of the retail industry, as measured by employment? dummy variable 'After'
2. How has retail fared relative to other industries? Regress EMPSTAT on dummy variable of IsRetail
3. Retail needs to worry about who has money to spend - what has changed about *who* is working and earning money? Age; Sex; maybe race regress EMPSTAT on Age and Sex

Your Job

1. Explore the data on the IPUMS CPS website and see what is in there. Only use data up to April 2022.
2. Figure out which analyses to run that will help answer these questions
3. Run those analyses
4. Write up your results in an Quarto document with your analysis code inside, and prepare a slide presentation of your results (Quarto slides work here too, or Google Slides or Powerpoint or whatever you like)
5. Your entire project should be kept track of as an R Project (.Rproj), in a folder with files kept in a standard file structure (code/, data/, results/, etc), and uploaded to a GitHub repository. **You don't need to upload the data files if they're too large**

You may bring in outside information aside from your data (for example, "lockdowns strengthened in month X, so our analysis will…") but this is not expected/required.

How Do You Do That?

How can you take a broad generic question and figure out an analysis that will help to answer it?

Let's take an example broad question not on the list: how has the "Christmas bump" changed in the retail sector?

First, we can ask: What is this question really trying to get at? Well, we know that every year, retail sees a huge boost because of Christmas. So your firm wants to know if the size of that boost is getting bigger than it used to be, smaller, or staying the same over time.

compare employment in April relative to other months

Then, we can ask: What kind of result would answer this question? Well, first we'd need to measure the Christmas bump - the relative amount of employment in retail in December relative to the months of January-October (dropping November since it could go either way) in a given year could be called a Christmas bump. So we want to know if "the amount of employment in retail in December relative to other months" is getting bigger or smaller in recent years.

Next, we ask: How can we get that result from a regression model? This analysis is asking us to compare December to other months, and then compare *that* comparison across years. Seeing how a comparison changes over time is the realm of interaction terms. So we want to get some sort of "December effect" and interact that with year, to see if the December effect is getting bigger or smaller.

Next, a regression equation! There are a few ways we could put this together (do we collapse January-October into one time period or treat them separately? Do we include fixed effects for all the months? Do we include controls for how the demographic makeup of the country is changing? Do we include change-over-years as a linear or quadratic term, or just compare this year vs. previous years? And so on...) but one way is this:

$$ RetailEmployment_{y,m} = \beta_0 + \beta_1December_m + \beta_2Year_y + \beta_3December_mYear_y + \varepsilon_{y,m} $$

where m is month and y is year. Now that we have an equation, we should think carefully about if it's the right one! Should we include November in the data or not? Or include it and give it its own control? Do we need heteroskedasticity-robust standard errors? Clustered standard errors? Should we log that dependent variable? Lots of things to think about.

Once we have our regression model down, we can think about how to run it! This regression seems to want one observation per month, so we'd need to create a year-month variable, use `group_by(year, month, yearmo) %>% summarize(RetailEmployment = sum(indname == 'Retail Trade'))` to calculate retail employment in each month (and preserve those `year` and `month` variables so we can include them in our regression model). Look at your data afterwards to make sure you did it right!

Then, we run the analysis and it's on to writing up the results...

In the Writeup

Your writeup should be in an Quarto document. Don't worry about length - if you're hitting all the points below for each of your analyses, that is what you want. Format the writeup as a report with sections for the main questions, not a list of bullet points. Information to include:

1. Why you are running the analyses you are running
2. How the analyses answer the question being asked, and what the result is
3. Carefully interpreting the results
4. Presenting the results in an appealing way. Graphs are great, `sumtable()` is great, `etable()` is great - put a little effort into formatting tables and figures to make them look nice! **At the very least, variable names should be in English rather than statistics-package** ('Education' not 'EDUC'). You will get marked down if variable names in your tables/graphs aren't in English. If you aren't comfortable enough with `ggplot` to make its visualizations look nice, feel free to make

graphics in Excel or Tableau or anything you like, and include them in your Quarto doc as images. Econometric analyses should be in R, and code should be visible.
5. Acknowledging the assumptions you are making in each analysis, how plausible those assumptions are in the context of your data, and any evidence you can provide backing up those assumptions. **Don't forget this part, you will be marked down if you don't do it.**
6. After doing all analyses related to a given question, provide a generalized answer to the main questions.

I do not expect undisputable flawless results - the data can only do so much, and we always have to rely on assumptions. However, an analysis with big flaws goes down a lot easier if you can very accurately interpret the results, point out the flaws or implausible assumptions, discuss how those flaws affect the results, and perhaps suggest an improved analysis you would run if it were feasible. Don't claim more than your results can actually show.

You can work collaboriatively on a Markdown file using GitHub, which is recommended. If you prefer, you can instead use Google Docs with the Markdown Preview add-in, or you can use Draft. These will not run the R code in the document, but they will let you work together on text and code.

In the Presentation

Your group presentation should be between 12 and 15 minutes long, plus a few minutes for questions. As many or as few of the people in your group can present as you like. There are no bonus points for being a presenter.

For your presentation:

1. Choose two of the main questions you feel you can answer best
2. Present your analyses related to those questions. Depending on how many you have you may want to select only a subset of them.

3. Make clear to the audience how your analysis works, how to interpret the results, and what the general answer to the question is
4. Be prepared to answer questions about your analyses and the assumptions behind them

Grading

Grading will be based on:

- 5% Following all requirements of the assignment
- 30% Selecting and accurately performing appropriate analyses to answer the questions
- 30% Correctly interpreting your analyses, linking them to the main questions, and discussing flaws and assumptions being made, and the plausibility of those assumptions
- 20% Clear and compelling writing
- 15% Clear and engaging presentation

Important and Helpful Notes

1. You can find the CPS data on Canvas, as well as the file linking the industry codes to their titles. You will probably want to work with the named industry, since they are more broadly defined and closer to the level you want.
2. The CPS has several *supplements* that provide additional information about a subset of the respondents. You may want to dig around in these to see if there's anything useful! The most well-known supplement is the ASEC, which has a *lot* more economic information than the typical monthly files. However, the ASEC is only released once a year, in March. So think carefully about whether your analysis can actually make use of that data before you go trying to use it.
3. Many of the variables in the data have labeled values, and these will be preserved if you read in the data with `import()` in `rio`. For example, a value of 40 in the `educ` variable means you left school after 9th grade. To see how the values relate to the labels, send a variable to `labeltable()` in **vtable** (or read the documentation).

4. Our class has focused a lot on identification error and causality, but not all questions are causal in nature! It's fine if you've picked an analysis that helps answer a main question without needing to isolate causality - often the methods are the same, but the assumptions you need to make and the interpretation of the results are different. For example, running a difference-in-difference comparing retail vs. non-retail before vs. after COVID hits won't tell you the "effect of COVID on retail" since DID assumes that the "control" group is untreated (and non-retail was definitely affected by COVID!). But the same regression you'd run for DID would tell you the *difference* between retail and other industries in how COVID affected them.
5. You will need to read the data documentation to do this project. See the .XML documentation file, or head to IPUMS CPS to read their documentation there.
6. Everything February 2020 and before is "before COVID" and everything April and after is "after COVID." But what about March? This is something you'll need to consider how to deal with in your analyses.
7. What time range are you going to include in your analysis? You don't necessarily want to include all data ever - what will data from September 2021 really tell you? (and at some point it will be too much for your computer to handle!) But also if you limit it too far you'll be leaving out important information
8. The effect of COVID on employment changes over time - it might be worse, for example, in April 2020 than in August 2020. Your analysis finds the effect of COVID *when*? Is the *when* you're getting the most useful *when* you could get? It's also not too difficult to let the effect itself change over time, if you want.
9. If you want an analysis to be at the industry-month level, you should make your data be at that level too! Use `group_by(year, month, indname) %>% summarize()` (with the appropriate `mean()` /etc. functions in `summarize()`) to collapse data to the `yearmo` / `indname` level
10. We didn't really talk about sample weights in class. Sample weights are a common tool in survey research (and CPS is a survey!) where you receive a weight based on how heavily they want to count you in the sample. For example, if 1% of their survey sample is Hispanic men aged 30-35 living in California, and they happen to know that *in the full population*, that group makes up 2.5%

of the population, then each of those men might get a weight of 2.5 to scale them up so the population is represented. When weights are available (as they are here in `wtfinl`), it's a good idea to use them. Look in your regression function's documentation - there is usually a `weights` option. Or if you're doing `group_by() %>% summarize()`, look into `weighted.mean()` rather than `mean()`.