Cross-sectional Analysis Project

[ECON 5753] — *University of Arkansas*

This assignment will summarize the first part of the course where we studied how to use forecasting methods for cross-sectional data. For this assignment, you will be asked to perform a lightly guided analysis of your choice of one of 5 possible datasets (or you may chose another dataset you might be excited about). I will give you some prompts to make sure you use the methods we covered in the class.

You should think of this as a project where you are *trying to learn something* about your dataset. A key takeaway from this class is to (1) know how to formulate questions about data available to you and (2) use the data to answer your question. That is, you are going to be use a dataset about a topic that (may) interest you and I want you to try and learn something new using data.

Part of your grade will be based on how professional your results look. All graphs should be labelled clearly and you should write in complete sentences in your writeup. Think of this as a report that a stakeholder will read to learn about your question. I have provided a template that will hide your code in your results document that you submit (in the real world, your boss is not looking for your code). When discussing your findings, make sure to include proper statistical inference. If the results are not statistically significant, then do not write about them as if they are iron-clad. In that vain, make sure to use proper standard errors (e.g. either `HC1` or clustered if you think that's important in your case).

Semi-guided questions

Please work through these questions and write up your results. See the template for an example of how I want this to look.

1. *Familiarizing your reader with the data:*

   Take a look through the dataset and select a few key outcome variables that you want to explore. These outcomes can be continuous or binary variables. Please tell the reader about the distribution of the key variables. For binary variables, you could tell the reader the percent of 1s in the data. Try visualizing continuous variables using a histogram.

   Note, you can *create* a binary variable using your data. For example, it is common to create a binary variable for being above some threshold (e.g. age above 55): `df$age_above_55 = df$age >= 55`.

2. *Investigating Relationships Between Two Variables:*

   Pick two continuous variables from the dataset: one outcome variable and a predictor variable. You should have a motivation for predicting one variable using another (do not just pick two random variables). You should write out a motivation stating your rationale for the variables you are choosing.

   How do the two variables relate to one another? Create a scatter plot to visualize their relationship. What sort of relationship do you observe? Do you think a linear approximation does a good job? If you feel it's appropriate, run a bivariate regression to quantify the relationship. Otherwise, fit a more flexible relationship (e.g. using 'binsreg')

3. *Adding additional controls:*

   In class we talked about the importance of multivariate regression. Specifically, we were nervous that the regression coefficient from a bivariate regression might 'pick up on' the influence of other variables that are correlated with our X variable.[1]

   I want you to reflect on the regression you ran for question 2 and discuss some potential variables that might be correlated with X *and* have an effect on the outcome variable. Include them in a new regression model and discuss how your estimates on the original variable

1. For example, we discussed regressing wages on years of schooling and talked about the challenge that arises in the this bivariate regression because years of schooling are often correlated with family income.

change. Please motivate your choice for the variables you include in your write-up.

4. *Interaction Effects:*

   Explore potential interaction effects between your predictor variables. Are the effects of one predictor variable dependent on the level of another? If so, describe and interpret these interactions. For example, if you're predicting income, does the relationship between education and income differ by gender or race?

5. *Create and evaluate a forecast model:*

   Split your data into two parts: 80% of the data will be your *training sample* and the other 20% will be your *testing sample*. Fit your forecast model on the training sample, but forecast on your testing sample. Evaluate how well your model does (e.g. test-sample mean-square prediction error). Are your prediction errors particularly large for any specific groups of units (e.g. plot your prediction error and some explanatory variable)?

Datasets

You have the choice of one of five datasets. You can see the brief overview of each as well as the *data dictionary* that describes each variable in the dataset by clicking on the links.

1. AirBnB listings in Nashville is a scrapped dataset from the website Inside Airbnb. It contains information about the listing as well as reviews received. I downloaded the current "snapshot" for Nashville, Tennessee.
2. The NBA 2K25 dataset contains information about all the players in the NBA 2K25 video game. It contains info on the player (team, salary, years in nba) and their ratings. Additionally, players 2023 – 2024 stats were merged in.
3. College Scorecard is a dataset created by the US Education Department. The dataset was created to allow for potential college students to easily understand the cost-benefit of different colleges.
4. The CPS contains survey data on workers collected by the US Government. I downloaded this from IPUMS using 2017, 2018, and 2019 data. Then I slightly cleaned it to create some useful variables.
5. The movies dataset contains data on a random sample of films released since 2010 and have more than 10 reviews on Letterboxd. The data contains information on reviews, the budget and revenue, genres, and other information.