Cross-sectional Data Analysis Project

[ECON 4753] — *University of Arkansas*

Data Analysis Project

This assignment will summarize the first part of the course where we studied how to use forecasting methods for cross-sectional data (i.e. data about individuals measured at a point in time). For this assignment, you will be asked to perform a lightly guided analysis of your choice of one of 5 possible datasets. 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 will be to learn how to formulate questions about data available to you and use data to try and 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.

You will be graded, in part, on how professional your results look. Here is some points of things you should do to make your report look professional:

- All graphs should have clear x-axis and y-axis labels. Include units in the label if releveant
- Title should be a one line takeaway (e.g. Players with higher salaries have more points per game).
- Code cells should all be hidden. 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).
- You should talk with precise statistical language (e.g. when interpreting regressions, you should not use causal language).
- Discuss statistical significances of your results.
- Write with complete sentences and clear gramma.

How to write up data analysis report

Your data analysis report should be structured to clearly communicate your findings to a non-technical audience. Think of this as presenting your analysis to a manager or stakeholder who wants to understand what you learned from the data.

Your report should include the following sections:

1) High-level Introduction

Begin your report with an introduction that establishes the context for your analysis. This section should explain what you want to understand about the data at a high level. Clearly state the research questions you will address in your analysis. This should just introduce the topic and not give very much technical details.

2) Present Your Data

Provide an overview of your dataset to familiarize readers with the information you are working with. Explain the dataset you are using. Whatis a row in that data? Describe the key columns in your dataset, focusing on the variables you want to use in your analysis. Highlight particularly important variables you might think are critical for your analysis.

Finally, explain your sample selection decisions. Did you drop any units from the data and does this change who your 'population of interest' is? Did you create any new variables? Justify these decisions by explaining how they improve the quality or focus of your analysis.

3) Summary Figures to Get Reader Acquainted with Your Data

Create descriptive visualizations that introduce readers to the key features of your dataset. The goal here is to help orient the individuals about the data and give high-level summaries of important variables. These could include:

- A histogram or other visualization showing key variables in your data
- Summary statistics (e.g. yearly-averages plotted on a graph)
- A table or text containing averages of a variable for different subgroups.

For the latter two, you could present these as tables, the output of a regression, as figures (using 'iplot'), or in plain text. These figures should be clearly labeled and help readers understand the basic patterns in your data before you present more complex analyses.

4) Main Analyses

Present your core analytical findings in this section. Structure your analysis around clear research questions and you should tell the reader what statistical methods you use to answer them. Your analysis should:

- Pose clear questions about the relationships in your data
- Describe the methods you used to answer these questions (e.g., regression analysis, correlation analysis)
- Discuss your results with appropriate statistical detail, including discussing statistical significance

Remember to interpret your findings carefully, avoiding causal language when your analysis is correlational. Explain what your results mean in practical terms that a non-technical reader can understand.

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. *Generate Questions:*

   Take a look through your dataset and formula a question or two (e.g. What kinds of colleges have high tuition? What kinds of players score a lot of points.)

   Choose some variable from the dataset and explore their distribution. These should be variables relevant to your questions. What can you tell me about its distribution? Try visualizing the data using a histogram and/or calculate the mean and variance.

   Write this up as (1) a paragraph motivating why you want to explore this dataset to ”your boss” and then (2) write a paragraph of what questions you want to answer.

2. *Run initial regression*

   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? C reate a scatter plot to visualize their relationship and/or estimate the correlation. 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. What insights can you draw?

3. *Reflecting on previous question*

   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. 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.

   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. Please discuss how you think that other variable might influence your bivariate regression coefficient.

4. *Analyzing a Binary Indicator Variable:*

   Identify a binary indicator variable in your dataset (e.g., gender, yes/no decision) and a continuous outcome variable to predict. How does the outcome variable differ based on this indicator? Compare the means by value of the indicator (0 or 1) using a bivariate regression where the outcome variable is regressed on the indicator variable. What conclusions can you draw?

   In this exercise, you can *create* a binary variable using your data. For example, it is common to create a binary variable for being above or below the median value:

   ```
   df$above_median_x = df$x >= median(df$x, na.rm = TRUE).
   ```

5. *Analyzing a Discrete Variable with Multiple Categories:*

   Select a discrete variable with multiple categories (e.g., education level or income bracket). How does the outcome variable vary across these categories? You can create multiple histograms (one per value of the variable) or run a regression using separate indicator variables for each category. What differences, if any, do you observe between categories?

6. *Explore!*

   Do not let this limit you. Consider interactions between discrete variables, interact a continuous variable with a discrete variable. Answer more than one question! I encourage you to follow this project whereeever it may take you! The quality of your analysis will make up a large portion of your grade, so ask and try to answer interesting questions!

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.
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. iMDB collects user reviews of movies as well as some basic info of the film.
5. The CPS contains survey data on workers collected by the US Government. This has characteristics of workers and their wages.