Andrew ElHabr
ECON 7022
9/12/17

Problem Set 1

I. Reading Assignments

First Readings of Hansen

Chapter 2

I found that subchapters 2.21 (Regression Sub-Vectors), 2.22 (Coefficient Decomposition), and 2.28 (Random Coefficient Model) were the most difficult sections for me to understand.

Subchapter 2.21 was the first subchapter in which I thought the notation began to be difficult to interpret. I am generally very comfortable with matrix notation, which is why I liked the matrix subchapters in Chapter 3 (3.9-3.11), but when I saw something like $\boldsymbol{Q_{11 \cdot 2}}$, I felt overwhelmed by the notation. After a few minutes, I was able to wrap my mind around what it meant, but I was left wondering whether or not using notation like this was necessary. Besides the notation, I understood what the intent of this subchapter was.

After reading subchapter 2.22 a couple of times, I was able to understand the goal of this subchapter. However, the concept of iterated projections is fairly complex and not something that had occurred to me could be done (although it seems to clearly be a valid concept). Fortunately, I ended up being able to comprehend this subchapter fairly well.

Finally, I thought subchapter 2.28 was the most unique subchapter that I read in Chapter 2 because the idea of a random coefficient model was completely new to me, so it took me a while to comprehend. I am so used to seeing regression models in which the regression coefficients correspond to the marginal effect of a predictor on the population average that I had to think hard about how a regression model could be created in which the coefficients corresponded more directly to individuals instead of a population average.

Chapter 3

I would not say that I struggled too much with any of the subchapter of Chapter 3, but if I had to pick one, it would be subchapter 3.12 (Estimation of Error Variance). It was not the notation or derivations that I struggled to understand in this subchapter, but rather the fact that the feasible estimator of error variance is smaller than idealized estimator of error variance. Besides clearly seeing how this is true from the mathematical steps taken in this subchapter, I have rationalized this by thinking that the feasible estimator cannot be as large as the idealized estimator because the feasible estimator makes use of the residuals of the regression model, and these residuals are likely "overfit" estimations of the true error of the model since they are dependent on the fitted values of the regression model.

First Readings of Angrist

FUQs

A FUQ is a fundamentally unidentified question. In other words, it is a research question that cannot be answered by an experiment. The problem with FUQs with respect to economic estimation is that there are certain effects among the predictors and some immeasurable/uncontrollable outside variables that cannot

be "untangled," making it impossible to properly answer the proposed research question. For example, with the example proposed in the "Questions" section of the book, it is impossible to examine the effects of school start age on elementary test scores because of the complications that arise due to maturity and time-in-school effects.

I think an example of a FUQ is the question, "What effect does cloud seeding (manually inducing rain) have on the sea water level in California?" This would be a FUQ because there are many outside factors, such as atmospheric pressure, the orbit of the moon orbit, etc., that we would not be able to control in an experiment that sought to answer this question. Thus, it would be impossible to know what truly caused changes (if changes were even observed) in the sea water level if we performed cloud seeding in California.

A simple example of an identified question is the question, "What effect does the amount of study time for an exam on a previously unknown language have on the performance on this exam?" One can relatively control the amount of time that students study for this exam if desired by monitoring over the students while they study. Also, the fact that this exam is about a previously unknown language means that no student would be more apt to be successful than any other student. This is different from the previous example because all effects are measurable in some way and all variables can be controlled.

Book Impressions

This book (AP) seems like it reads much more like a "quest for knowledge" than the Hansen lecture notes. In AP, it looks like one can get a decent feel of the authors' voice, and the text appears to be written like a typical novel instead of like a textbook. It appears to still be relatively mathematically rigorous like the Hansen lecture notes while adding extra commentary.

I am not sure if my opinion will stand with the majority, but the Hansen lecture notes seem more appealing to me. After taking numerous math and science related classes, I have grown accustomed to and fond of the textbook-style nature of the Hansen lecture notes. Although it looks like AP may provide some useful exposition, I worry that the extra amount of text may drag on and prevent me from wanting to read more.

II. Empirical (programming) Problems

Question 1: Coin Flips

(a) (in ECON_7022_Problem_Set_1_Q1.R file)

(b) Yes, I produce the binomial random variable, which is the number of heads appearing in 10 flips of a coin, in the first line of code.

```
coin.flips <- rbinom(n = 10, size = 1, prob = 0.5)
```

Yes, the outcome of the random variable can be found in the output of the third line of code, which happens to be 6 heads for the seed I set.

```
sum(coin.flips)
[1] 6
```

Yes, a sample can be found in the output of the second line of code. The vector of 0's and 1's is considered a sample.

```
coin.flips
[1] 0 0 1 1 0 1 1 1 1 0
```

(c) (in ECON_7022_Problem_Set_1_Q1.R file)

For the random seed that I set, the sample mean is 4.987, and the sample variance is 2.611753. These match closely with what we expected: $np = 5$ and $np(1-p) = 2.5$, respectively.

Question 2: Do the 'Grades' analysis from the slides on a different dataset

(a) (in ECON_7022_Problem_Set_1_Q2.R file)

(b) (in ECON_7022_Problem_Set_1_Q2R file)

(c) (in ECON_7022_Problem_Set_1_Q2.R file)

The 95% confidence interval for wages of people who are not married is simply the 95% confidence interval on the intercept of the line of best fit, [4.35140, 5.336427]. And unmarried people do have "significantly different" wages than married people.

Bonus Question: DIY: Simulate a multivariate distribution

(a) (in ECON_7022_Problem_Set_1_Bonus.R file)

(b) (in ECON_7022_Problem_Set_1_Bonus.R file)

(c) (in ECON_7022_Problem_Set_1_Bonus.R file)

(d) (in ECON_7022_Problem_Set_1_Bonus.R file)

III. Project

(a) Potential questions:

1. What are the most significant variables for projecting a National Football League (NFL) team's win total?
2. What are the most significant factors for winning a National Basketball Association (NBA) game?

(b) I care about my first question because every year I try to accurately predict how each NFL team will perform that season by using only my intuition. I would like to be able to better predict these win totals in order to beat my brother in estimating how well each team will perform for sake of familial competition. I would probably run a linear regression model using various predictors like previous year's win total, previous year's offensive ranking, previous year's turnover margin, starting quarterback salary, total years of experience, etc. (X variables) and determine which ones of these are statistically significant in predicting current season win total (Y variable). I could probably scrape all relevant historical data from the internet from a website like ProFootballReference. I would use ANOVA to determine if any relationship I found is statistically significant by examining the p values I got for each predictor. If this turns out to be the case, I would not just take this relationship at face value and would check how correlated this relationship is to other variables in the model. Also, if it turns out that there are many statistically significant predictors, I could use stepwise regression to generate a more parsimonious model.

I care about my second question because I would like to know what a basketball team should focus on in order to maximize their chance of winning a game. I used to coach basketball, still enjoy playing it, and also enjoy watching the NBA, so this is interesting to me on a personal level. I would use box score statistics and advanced statistics for all NBA games over the past 5-10 years as my data set and would probably scrape this data from a website like BasketballReference on the internet. I would consider using a logistic regression model on predictors like turnovers, free throw shooting percentage, offensive rebounding percentage, defensive rebounding percentage, etc. (X variables) to determine which ones of these are statistically significant in predicting the outcome of a game (Y variable). Once again, I would use ANOVA to generate p values for each predictor to determine which ones of them have significant relationships with winning a game. I could also a classification method such as k-Nearest

Neighbor (kNN) to check the results I find from a logistic regression model. Once again, I would not take any relationship I find at face value. I would check the collinearity of the predictors among each other to ensure that there are no redundant relationships.

(c) I think both projects have the potential to be promising, but at the moment, I think the first project has better potential. This is because there are so many possible ways I can imagine that I could approach the problem, and I think the data is relatively easy to gather. In addition, the second project has likely been investigated in depth by other people before, so there may be a smaller chance to discover something new.

(d) In general, both are very interesting to me because I really enjoy both the NFL and the NBA. However, I think the first project is a little more interesting I have always wanted to really hone in my prediction abilities in this area. It would be fun to dedicate my time to a project like this.

(e) I talked to Cyrus Rich about my projects, and he thought that both were interesting. He agreed that the first project has a lot of potential because of the various ways in which you could approach it. Since he thought both were generally good, he encouraged me to pick the project about which I was more passionate.