Exam1 I like your other sections but Data section needs more work

Vinish Shrestha

10/18/2021

Please note that I would require your R code for section 1 and your work should be typed.

# Section 1: Data work

In this data work we want to explore the relationship between mother’s education and infant health more closely and issues surrounding such relationships.

Open the data set “NCHS\_birthweight2000\_sample.csv” located in the [**gitpage**.](https://github.com/vinishshrest/Health-Economics-Course) Note that this is a csv file. You will need to save the file in a particular location in your computer and state the path before you open the file. For example, if you store your file in this path: /home/desktop, then use the command: **data** <- read.csv(“/home/desktop/NCHS\_birthweight2000\_sample.csv”) to read the file.

Note that data consists of six variables: i) person\_id (person number unique to each person), ii) dbirwt

(infant’s birthweight), iii) dmeduc (mother’s education), iv) race\_white (an indicator for if the mother is White), v) cigar6 (if mom smoked cigarettes during pregnancy), and vi) dmage (mothers age).

Next, open the data set “NCHS\_income.csv” similarly as you opened the birthweight data. You can find this data in blackboard’s data tab. After saving this file, here is how you do this.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| datainc <- **read.csv**( "/home/user1/Dropbox/health economics teaching/data/NCHS\_income.csv")  *#hist(data$annualincome)* **head**(data) | | | | | |
| ## person\_id dbirwt dmeduc race\_white cigar6 dmage  ## 1 1 3997 12 1 0 26 | | | | | |
| ## 2 2 | 3827 | 17 | 1 | 0 | 41 |
| ## 4 4 | 2124 | 14 | 1 | 0 | 24 |
| ## 5 5 | 3459 | 6 | 1 | 6 | 23 |
| ## 6 6 | 3204 | 14 | 1 | 0 | 25 |
| ## 7 7 | 3430 | 12 | 1 | 3 | 35 |
| **head**(datainc) |  |  |  |  |  |

## annualincome person\_id

## 1 29439.93 1

## 2 57753.68 2

## 3 44600.88 3

## 4 11181.03 4

## 5 54090.66 5

## 6 34249.53 6

1. Plot a histogram for income in dataframe datainc.
2. In this question, we would want to merge two files **data** (birthweight file) and **datainc** (income file) by using person id. Note that each person has a unique id in both files and they can be merged together using *person\_id* variable. For example, if person\_id 1 refers to Maya in data file then person\_id of 1

will also refer to Maya in income file. We want to bring Maya’s birthweight observation and her income

observation together. To do so use:

datanew <- **merge**(data, datainc, by = "person\_id", all.x = T) **head**(datanew)

## person\_id dbirwt dmeduc race\_white cigar6 dmage annualincome

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ## 1 | 1 | 3997 | 12 | 1 | 0 | 26 | 29439.93 |
| ## 2 | 2 | 3827 | 17 | 1 | 0 | 41 | 57753.68 |
| ## 3 | 4 | 2124 | 14 | 1 | 0 | 24 | 11181.03 |
| ## 4 | 5 | 3459 | 6 | 1 | 6 | 23 | 54090.66 |
| ## 5 | 6 | 3204 | 14 | 1 | 0 | 25 | 34249.53 |
| ## 6 | 7 | 3430 | 12 | 1 | 3 | 35 | 52639.11 |

The table above gives a snapshot of 6 observations following the merge after using head command. You can obtain this by doing *head(datanew)*.

1. Run a univariate regression of the form: *birthweight* = *α* + *βmother*\_*education* + . Report the coefficient on *β*. Does this show a causal relationship between mother’s education and infant health. **Note that you should delete all observations with mother’s education of 99 (not reported values) before you proceed with this.**

datanew <- **subset**(datanew, dmeduc **<** 99)

Comment on education-gradient using your results from this regression.

1. Next, generate an indicator variable for low birthweight called “low\_bwgt” if *birthweight <* 2*,*500 grams.

datanew**$**low\_bwgt <- 0

datanew**$**low\_bwgt[datanew**$**dbirwt **<** 2500] <- 1

1. Calculate the proportion of infants with low birthweight by mother’s education. For instance, what proportion of infants belonging to mothers with 12 years of schooling have low birthweight. Do this for all education values.

**mean**(datanew**$**low\_bwgt[datanew**$**dmeduc **==** 12]) *#.. this gives the proportion of infants with low birthweig* ## [1] 0.07935223

1. Using your calculations from 3 plot the relationship between mother’s years of schooling and proportion of infants with low birthweight. *hint:* type help(plot) in the console.
2. Now construct an indicator that represents mothers with more than high school education; “aboveHS” by using the following hint.

datanew**$**aboveHS[datanew**$**dmeduc**>**12] <- 1.

Follow similar approach to assign the value “0” for mothers with high school or less than high school.

1. Now run the regression of the model specification: *low*\_*birthweight* = *α* + *βabove*\_*HighSchool* + . Comment on the coefficient of *β*.
2. Next, generate an indicator of race: black vs. white. Then run the following specification: *race* = *α* + *γabove*\_*HighSchool* + . Comment about the point estimate of *γ*ˆ.
3. Using your realizations from 9, comment on the case of omitted variable bias when estimating *low*\_*birthweight* = *α* + *βabove*\_*HighSchool* + in 6.
4. Run the regression specified as *low*\_*birthweight* = *α* + *βabove*\_*HighSchool* + *κannualincome* + . Compare the estimate of *β* with the specification *low*\_*birthweight* = *α* + *βabove*\_*HighSchool* + . Comment on which specification you’d prefer and why.

# Section 2: Lectures

1. The demand for doctor visits can be written as:

*doc visits* = *g*(*Insurance, Education, Income, Demographic,* ),

where contains unobserved factors that a researcher cannot perceive. One example is risk preference – as a researcher you cannot observe the risk preference of an individual. Now, assuming a linear functional form, you can express the demand for doctor visits as:

*doc visits* = *α* + *βInsurance* + *γEducation* + *δincome* + *κDemographic* + .

1. After estimating the regression, say you get *β >* 0. Does this mean that having an insurance increases doctor visits compared to those without insurance? Explain.
   1. *We can say that if Insurance is present, it has a β sized effect on the number of doctor visits.*
2. Based on part a. can you say that insurance has a causal effect on increasing doc;visits? Why or why not?
   1. *We would need more information to come to this conclusion. We would need to know that nothing that directly effects Insurance also directly effects doc visits. If the variable age effects insurance, education and doc visits directly, we can say that there is no causal effect, they might just be correlated.*
3. Based on the evidence that insured individuals face lower price compared to uninsured and insured individuals are more likely to go to the doctor (*β >* 0), can you state that the demand for doctor visits is downward sloping? Why or why not?
   1. If we are looking solely at the fact that a lower price increases doctor visits, demand is clearly downward sloping.

2. This problem pertains to the randomized control trial. Say, you want to find out whether the demand for doctor visits is downward sloping among young adults. To do so, you do a lottery and randomly assign health insurance for 25 people, who are technically termed as the “treated group." The other 25 people who did not receive insurance are control group.

1. Consider the specification: *visits* = *α* + *βTreat* + V.

Where *visits* represent the total number of doctor visits, *Treat* is the group that was assigned insurance through the lottery draw, and Vis the error term. Say, after estimating this specification, you find that *β >* 0. What does this suggest?

This suggests that Treat has a positive effect on visits, and has a high likelihood of being causal because of it being an RCT.

1. Does the estimate of *β >* 0 (in this case of randomization) suggest that there is a causal relationship between getting insured and increased doctor visits? Why or why not?
   1. *It does a better job at proving causality because it removes some sorts of bias in the study*.
2. Now, in a different analysis based on the survey data, you find that those insured tend to have higher doctor visits. From this finding can you say that insurance leads to more doctor visits? Explain in resonance to your answer from part b.
   1. *From what we have learned in class, I feel like you can never say that something actually leads to another, but through a randomized control trial you have the best opportunity to do so, at least based on the trials we have reviewed.*
3. Briefly describe the Oregon Health Insurance experiment.
   1. The Oregon Health Insurance experiment was effectively an RCT that provided via lottery health insurance to a pool of applicants. They analyzed the effects of the “treatment” 1 year into providing it. Their results showed a positive causality, because it was an RCT, between insurance, and utilization of the health care available. There were other measured outcomes, like depression and financial strain as well as some others, but the main focus was on health care utilization.

# Section 3: Readings

1. Card et al. [(2008)](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2600774/) uses a natural experiment that uses eligibility criteria (age of 65) as an identification strategy to identify the effects of Medicare on mortality outcomes.

1. Explain in detail how the eligibility criteria set for Medicare is used as a quasi-natural experiment.

Talk about the approach (method) they use.

A natural experiment takes a variable not controlled by those administering the study, and sets the differentiation of the groups based on said variable. They observed that the severity of incident on either side of age 65 is seemingly similar, therefore, the only significant difference in this model is if they are covered by Medicare or not. This is also something in which the treatment is not something they will have to devise and implement, meaning that they are under much less funding pressure to rollout a treatment to observe a reaction to.

1. The analysis of Card et al.’s study is based on hospital level data such that people in the sample are those who are admitted to the hospital. This creates a selection problem if we compare people who are right above the threshold vs. those who are right below the threshold. This is because if a 64 year old plays a waiting game and goes to the doctor when she is 65, she would have waited longer compared to a 64 year old with similar illness but who does not play the waiting game (i.e., goes to the doctor when 64). Hence, comparing the health outcomes of 64 year old individuals (without Medicare) vs. 65 year old individuls (with Medicare) might mean that we are comparing two groups that are systematically different in health stock to begin with. Describe how the researchers get around this selection problem. *(hint: hospital admissions are higher in week days compared to weekends.)*
   1. They looked at conditions that have similar admission rates on weekdays and weekends.
2. By using a figure that corresponds to the eligibility criteria (age), briefly describe the findings of their study on: i) insurance coverage, ii) quality of care, and iii) mortality outcomes.
   1. Medicare coverage has an important impact on patient survival.
   2. This difference in quality of care for this subpopulation is insignificant in terms of its overall affect on patient mortality.
   3. Mortality outcomes are decreased because of the availability of the Medicare treatment, and the nearly 65 percentage point increase in coverage of population that was previously uncovered.
3. What are some potential drawbacks of their study?
   1. I think that, for what they are trying to measure, they are doing a fantastic job for controlling out many different potentially problematic variations that could occur. Maybe there is not enough