AQA-II
Spring 2020

Lab assignment 3

Due at 6:30 pm on March 12th

Question 1. Analytical problems[1]

(a) Derive the OLS **variance** estimator for β_{ols}, assume all necessary assumptions held and the error term $\varepsilon_i \sim N(0, \sigma^2)$ and i.i.d.
(b) What could be an effective way of reducing the variance for your estimator?
(c) If you know the measurement for your variable of interest X is susceptible to a non-differentiable error term $\varepsilon \sim N(0,1)$, what influence does this have on your interpretation of the regression estimates?
(d) Explain the Conditional Independence Assumption (CIA) for using the propensity score matching method.
(e) Explain the Parallel Line Assumption (PLA) for using the ordered logit regression model.

Question 2. Empirical analysis

As a social scientist, you are getting paid (or not) to study interesting and important questions. And this assignment will take you to experience a few with real data derived from the US Census data (ACS). So, first you are asked to download and create a dataset from the https://usa.ipums.org/usa/ website as per the descriptions, and then perform a series tasks of data cleaning and analysis.

A. Creating the dataset
(a) Create an account on the impus website, and then select the sample from year 2012 – 2018 (ACS is fine, do not select the 5-year sample)
(b) Then, we move on to select variables. The list of variables you should include is as below:

Var-name	Variable label
year	Census year
sample	IPUMS sample identifier
statefip	State (FIPS code)

[1] All the of answers could be found in lab lecture slides 2 – 4, but you should try to write the answers on your own without referring to them first. Let me know if you get stuck. However, be sure to explain in your own words rather than copying verbatim.

countyfip	County (FIPS code)
density	Population-weighted density of PUMA
city	City
ssmc	Same-sex married couple
nfams	Number of families in household
pernum	Person number in sample unit
perwt	Person weight
sex	Sex
age	Age
marst	Marital status
race	Race [general version]
raced	Race [detailed version]
hispan	Hispanic origin [general version]
hispand	Hispanic origin [detailed version]
citizen	Citizenship status
hcovany	Any health insurance coverage
educ	Educational attainment [general version]
educd	Educational attainment [detailed version]
empstat	Employment status [general version]
empstatd	Employment status [detailed version]
inctot	Total personal income
poverty	Poverty status
migrate1	Migration status, 1 year [general version]
migrate1d	Migration status, 1 year [detailed version]
tranwork	Means of transportation to work
trantime	Travel time to work

Note that some variables are included automatically, while for others you need to find out by browsing categories or searching. Also that, for example, if you include the variable race, two versions (general and detailed) of the same variable will be included as well. Before you reach to the extract page, you should confirm that you have all the variables listed above, and the sample should only include **year 2012 – 2018**. Then proceed to extract, where you will select the data to be in the ".dat" format. Eventually, you will download the data and do not forget to open the command files for any language you like, copy and paste into your console/dofile/interface, run to load the data properly.

Extract Number	Date	Formatted Data	Fixed-width Text Files: Data	Fixed-width Text Files: Command Files	Codebook	Revise Extract	Resubmit Extract
6	2020-02-24	--	Download .DAT	SPSS SAS Stata R	Basic DDI	revise	

(c) Verify you have the correct version of the dataset. Below is what I have, and your dataset should match if done correctly.

```
. d,s

Contains data
  obs:     8,596,583
 vars:            29
Sorted by:
     Note: Dataset has changed since last saved.
```

(d) Examine the dataset by yourself and get ready for the next task (no need to write anything for this question).

B. Data cleaning

(a) Some variables, though providing rich information, are not in the most convenient format.

Let's recode them into convenient categories.

1. Marital status
   1 Married, 0 Not married (value 3 - 6)
2. Race
   1 White, 2 Black, 3 Asian (value 3 - 6), 4 other (value 7- 9)
3. Ethnicity (hispan_lbl)
   0 Not Hispanic, 1 Hispanic (value 1-4)
4. Health insurance (hcovany)
   0 Do not have health insurance, 1 Have health insurance
5. Employment status (empstat)
   1 Employed, 0 Unemployed, 2 Not in labor force (value 0 and 3)
6. Migration status
   0 No moving, 1 Moved (value 2 - 4)
7. Means of transportation
   0 Motor vehicles (value 10, 20, 35), 1 Public transportation (value 31 – 34, 36), 2 other (40, 50, 60, 70), missing (value 0)
8. Sex
   0 Female, 1 Male
9. Number of families in household
   1 Single member household, 2 Two-member household, 3 Three or more members (value 3-20)
10. Same-sex couple indicator
    0 No same-sex couple, 1 Same-sex couple (value 1, 2), missing (value ".")
11. Education
    No need to change but remember to use "educ" not "educd".

(b) Use a single paragraph to describe the dataset, do not provide a screenshot of the describe command, but rather think of this as a chance for your to concisely describe in words the dataset you have to your reader. Be sure to mention the name, year coverage, variables

available, sampling method, observations, etc. (You will always have to do this in any research paper you write)

(c) Investigate and explain why the employment status variable (the original one) has more than a million "N/A". (Hint: think about the age requirement for work, and probably state law variations)

C. Data analysis

Theme 1. Health insurance in the US

First, let's drop any observations if age <18, and this should give you a total of 6,832,338 observations.

(a) What are the predicting factors for having a health insurance in the US? Run an OLS regression on health insurance status. Include all the relevant variables in their proper forms (continuous or categorical), **ignore any weight variables and city, county indicator, citizenship** for now.

(b) Now add county (as indicated by the country fips code) fixed effects to your model in (a). Is there a change in degree of freedom? Why? (hint: it's easiest to use the "areg" command, type "help areg" to find out)

(c) Save your predicted outcome from (b) and separate by state, plot the predicted probability of having insurance for NY, CA, FL across years of 2012 – 2018. Interpret what you observe.

(d) Now several states want to pass laws to add a working requirement for residents to receive government funded health insurance. From the Census data, which state do you think might be impacted the most if the law was to pass? (Hint: which state has the highest proportion of residents having insurance but unemployed and poor)

Theme 2. Marital status and same-sex marriage in the US

(e) Use table or tabulate, show which state has the most same-sex couple in terms of percentage? Which has the least?

(f) Compare the age, income, race of the most and least state from question (e)

(g) What might be some factors that affect where these same-sex couple live? (hint: cultural/political/religious tolerance/self-selections?)

(h) Run a Probit model predicting the factors of being a same-sex couple. Interpret your model. (Use the same regressors as in theme 1)

(i) Save your predicted probabilities from question (h), draw a plot where Y-axis is the probability and X-axis is the year, and across the different states.

(j) Repeat (h) and (i) but this time the outcome is the marital status. Report your findings and compare to those in the previous question.

Theme 3. Domestic migration and economic opportunities

(k) Preserve your data first, and then collapse income into years and state, draw a yearly trend of income across each state. What do you observe?
(l) The election year is 2012 and 2016. Did you observe any fluctuations around these two time points?
(m) Restore your data. Which state has the highest rate of residents moving out? What about before and after 2016?
(n) Economy is subject to a delayed effect from policies while people also anticipate and strategize their behaviors in expectance of incoming policy changes. Create a variable that is the lagged income of previous year for each year on the county level, i.e., starting from year 2013, each county in 2013 has a column of last year's mean personal income. (hint: " bysort county year: egen mean_income = mean (income variable)", then to construct the lag, you will have to work with "[_n]")
(o) Create a variable that is the difference in mean income between the current year and last year for each year. This variable is called the lagged difference in income.
(p) Regress the moving status on the variable you created in (m) and (n) respectively. What do you observe?
(q) Is it a good idea to add state or country fixed effects to the models in (o)? Explain your answer either way.
(r) Finally, run a final model of moving status against the variable you created in (o) along with other controls. (hint: if you notice variables being omitted, just exclude them)
(s) A reporter from NPR asked for your opinion on the status of migration for economic incentives after Trump being elected. What would you say if you can only say 4 sentences?