ECON 308: Econometrics
Assignment 3

Complete each problem to the best of your ability and submit in class on Thursday, November 4. You are encouraged to collaborate with other students, but you should turn in the problem solutions individually. Your writeup should include 1) written/typed responses to the questions, including regression tables where needed, 2) the code you ran to generate them (your do-file), and 3) any graphs produced.

1. This exercise involves data on 299 eruptions of the Old Faithful geyser in Yellowstone National Park. The variable `duration` is the length of the eruption (in minutes), while the variable `waiting` is the length of time until the next eruption.

   (a) Load the dataset `geyser.dta` from the Data folder on Blackboard.

   (b) Using the `twoway scatter` command, create a scatterplot relating `duration` to the following `waiting` time. Discuss any visual evidence of measurement error in either variable.

   (c) Regress waiting time on duration. On average, how does an extra minute of duration relate to average waiting time? Is this statistically significant?

   (d) Use the Breusch-Pagan test to determine if this regression exhibits heteroskedasticity.

   (e) Rerun the model with robust standard errors.

   (f) Report the R^2 for this regression. Do you think that duration is a useful variable for predicting waiting time?

   (g) Generate a regression table and include the results from both regressions in your writeup.

2. A fundamental question of interest to economists is *intergenerational mobility* - how does the socioeconomic status of families evolve across generations? To put it another way, to what extent does the socioeconomic status of a child's family influence their ability to succeed? If there is little relationship between a father's income and his son's future income, we would say that intergenerational mobility is high. If there is a strong relationship, that suggests that intergenerational mobility is low.

   To study this, one needs to measure an individual's characteristics late in life and compare them to those of their parents when they were young. One way to do this is to link individuals across censuses - find the census record for a person when they were a child, record their parent's information, and then match that information to the child's record in a future census. A problem with this approach is that, historically, the census has not assigned unique identifiers to individuals. To get around this, economic historians have developed methods to link individuals based on factors which should remain fixed over their life course. Commonly used linking variables include given name, surname, race, year of birth, and place of birth. So, for example, if we look at the 1920 census and find Wassily Leontief, a 14-year old white male born in Russia, and we find a similar man aged 34 in the 1940 census, we can link those two records. In this exercise, we'll examine intergenerational mobility using samples of linked individuals provided by IPUMS; in particular, individuals linked between the 1880 and 1920 U.S. censuses.

   (a) Create a do-file and set it to load the dataset `linked_1880_1920_males.dta`, which you can download here: `https://www.dropbox.com/s/xy8gnicisa6y0z7/linked_1880_1920_males.dta?dl=1`.

(b) This dataset has some of the same variables from both censuses, so the variable names are distinguished by a number: _1 variables are the 1880 records, while _2 variables are from the 1920 records. Add code to keep only those who were under the age of 16 in 1880 (using `age_1`).

(c) Another limitation of historical data is a lack of information about income; the census didn't ask about this until 1950. Instead, this dataset includes two measures of income for the children and their fathers based on occupation. The first, `ln_occscore_hh` and `ln_occscore_child`, reports the log of median 1950 income associated with the father's and son's occupation, respectively.[1] The second, `ln_adj_occscore_hh` and `ln_adj_occscore_child`, is also in log terms and is based on 1950 occupational income, but it additionally accounts for heterogeneity in income based on demographics, region, and industry. Using these variables, estimate the following regression separately for both measures of income:

$$log(income)_i^{\text{son}} = \beta_0 + \beta_1 log(income)_i^{\text{father}} + \epsilon_i \tag{1}$$

What are the estimated coefficients in both cases?

(d) What is the interpretation of the estimated $\widehat{\beta}_1$ in numerical terms? Which measure of income yields a higher level of intergenerational mobility?

(e) Estimate both regressions again, but include indicator variables for region (using `region_1`) and interactions between father's income and region. What is the omitted region? Which regions have the highest and lowest intergenerational mobility?

(f) Replace the region indicators and region/father's income interaction with an indicator for race, and interact that with father's income (using `race_1`). What is the intergenerational mobility elasticity for each group using each income measure?

(g) Are the differences you found in the previous question statistically significant in either case? At what confidence level?

(h) Identify which regions have a sufficiently large number of African-Americans to estimate a black/white mobility gap.

(i) Now, estimate the regression from part (f) separately for each of the regions identified in part (h). Which have the highest/lowest intergenerational mobility for African-Americans?

(j) Where is the gap largest? Is it statistically significant?

(k) Using the `nativity_1` variable, generate an indicator variable that equals 1 if a child has at least one immigrant parent or is an immigrant themselves, and zero otherwise. Replicate the regression from part (e) for just those who are immigrants or have immigrant parents. Which regions have the highest and lowest intergenerational mobility for this group?

(l) Historical studies of linked individuals often ignore women. Why might it be harder to reliably link women across censuses? What biases might this introduce to our data?

3. Robberies in Chicago: The data for this problem contains information on robberies for 19,330 Chicago street segments over the period 2008-2013. Figure 1 provides a representative example of the unit of observation. Predictors include the number of bus stops, the percentage of the area that is commercial, an indicator for the presence of a street intersection, the number of late-hour bars, population, and a measure of walkability from Walkscore.com. It also includes the community area (neighborhood) containing the observation.

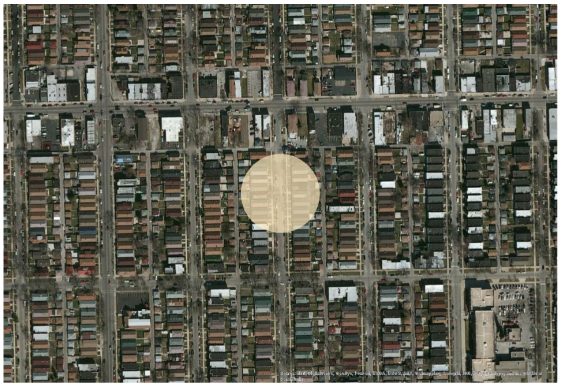
   (a) Load the dataset `chicago.dta`, which you can download here: `https://www.dropbox.com/s/8pmjh2jxgty3kvf/chicago.dta?dl=1`.

   (b) What is mean, standard deviation, and range of robberies in the data?

   (c) How many observations contain a street intersection?

[1] The "hh" denotes head of household.

Figure 1: Sample Unit of Observation


Source: Esri, DigitalGlobe, GeoEye, i-cubed, USDA, USGS, AEX, Getmapping, Aerogrid, IGN, IGP, swisstopo, and the GIS User Community

(d) How many contain a late-hour bar?

(e) How many community areas are represented?

(f) What is the correlation between bus stops and robbery counts?

(g) Generate a variable that represents robbery rates per 1000 residents.

(h) Generate standardized versions of population and Walkscore.

(i) Generate an indicator for % commercial (you can choose what threshold you'd like to separate commercial/non-commercial areas).

(j) Fit a linear model to predict robbery rates using the covariates. Report and interpret your results.

(k) Suppose you wanted to test for differences in the relationship between commercial uses and robberies across observations with different population densities. What would you include in the model to accomplish this? Do so and report/interpret your results.

(l) Suppose you believe that the relationship between population density and robbery rates is nonlinear. How might you modify the model to test for this? Do so and report/interpret your results.