651 Empirical Economics: Assignment 1

Per Hjertstrand[*] Andrew Proctor[†]

September 2017

Abstract

The assignment gives a maximum of 5 points. Answers should be provided in English. The key to a good score is to be clear, precise and concise in your answers. Please, refer to the course webpage for further instructions regarding how the solutions are to be structured.

Part A: Revisiting basics

1. Explain in your own words:

   (a) The meaning of a random variable.

   (b) The difference between a continuous and discrete random variable.

   (c) The meaning of a probability distribution function (pdf).

2. Let X be the number of absent students from a class in 651 Empirical Economics on a Monday morning. The probability distribution for X is

x	0	1	2	3	4	5	6	7
$f(x)$	0.005	0.025	0.310	0.340	0.220	0.080	0.019	0.001

   (a) Draw the probability distribution in a graph.

   (b) Calculate the probability that $2, 3$ or 4 students are absent.

   (c) Calculate the probability that *more than* 3 students are absent.

   (d) Calculate the expected value of X and interpret the result.

   (e) Calculate the variance of X and interpret the result.

   (f) Calculate the expected value and variance of $Y = 7X + 3$.

[*] Per.Hjertstrand@ifn.se

[†] Andrew.Proctor@phdstudent.hhs.se

Part 2: Mechanics of OLS

Consider the following population model:

$$y_i = \beta_0 + \beta_1 x_i + u_i,$$

where y_i is the consumption of food for household i during one year and x_i is the disposable income during that year for household i. Both variables are measured in thousands of Swedish Krona.

Suppose we collect data from 3 households where:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 15 \\ 12 \\ 5 \end{bmatrix} \text{ and } \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 10 \\ 15 \\ 7 \end{bmatrix}.$$

(a) Look at the data on y and x. Is there something odd about this data?

(b) Calculate the OLS estimates of β_0 and β_1 by hand. Interpret your results.

(c) Calculate the variances of the OLS estimates by hand. Interpret your results.

(d) Calculate the R^2-value by hand. Interpret your result.

(e) Calculate the OLS residuals by hand.

(h) The true values are $\beta_0 = 0$ and $\beta_1 = 1$. Why are the OLS estimates different from the true values? *Hint:* Check whether MM assumptions 1 and 2 in the lecture notes are satisfied for the true values.

Part 3: Stata Exercises

For this exercise, you will be working with Swedish microdata from the "Survey of Adult Skills," which is a component of the OECD Programme for the International Assessment of Adult Competencies (PIAAC). The Survey of Adult Skills is a major international survey which combines an assessment of participants' cognitive skills with an array of questions about individuals' background and labor market experiences.

The data can be downloaded from: http://www.oecd.org/skills/piaac/publicdataandanalysis/. Go to "Download the PUF" → "CSV" → "prgswep1.csv".

Because the dataset is not available in a Stata format, you will also need the "International Codebook," available from the main page.

You will be using the PIAAC data to explore determinants of wages, similar to Example 1.2 in Wooldridge. To do so, complete the following steps, recording your work and discussion in a do-file and logging your results.

1. Import the dataset. How many observations and variables are there?

2. For this analysis, your explanatory variables will stem from the following variables in the dataset: *gender_r*, *age_r*, *pvpsl1*, and *yrsqual*.

   For each variable, first explore the structure of the variable, using the downloaded codebook along with the **codebook** command. State what the variable represents, the format of the variable in Stata (string or numeric), and whether any values are missing / unascertained. For string variables, you may want to also use the **tabulate** command to check for missing observations.

   Additionally, perform the following tasks specific to each variable:

   (a) For categorical variable *gender_r*, create a new indicator (or "dummy") variable for whether an individual is female. To do this, assign the value "0" to individuals listed as male (the omitted "reference" group) and "1" to those listed as female (be sure missing variables are coded as such). State the percentage of individuals who are female.

   (b) For *yrsqual*, ensure the variable is expressed as a numeric, then compute the mean, standard deviation, min and max.

   (c) Using the *age_r* variable, create a new "potential experience" variable, which is equal to (age - 19) if this is positive, zero otherwise. State the mean and median (50th percentile) of this variable.

   (d) Using *pvpsl1*, create a new variable equal to the individuals' sample percentile of cognitive ability.[1] To do this, use **egen** to create a variable equal to **rank** of each observation's *pvpsl1* value, and another variable equal to the number of observations (**count**) of *pvpsl1*. The percentile rank is then equal to $\frac{rank-1}{count-1}$.

3. The dependent variable you will use for this analysis is "monthly income percentile rank," *monthlyincpr* in the dataset.

[1] An individuals percentile of ability in the sample could of course vary significantly from it's percentile in the general population, which is the more relevant consideration for labor market outcomes. But if we are willing to assume random sampling, the sample percentile should closely approximate the population values.

Unfortunately, income is interval-censored.[2] To approximate the true percentile of income, create a new variable equal to the midpoint of the percentile interval in which they belong (in range 0-100). If the value of *monthlyincpr* is missing and an individual is unemployed according to *c_d05*, assign the value of 0. Drop individuals whose labor force status is "out of the labor force" or "not known" according to *c_d05*.

Report the number of missing observations for both *monthlyincpr* and your revised income variable.

4. Regression Analysis:

   (a) For individuals aged 30-65, use linear regression (OLS) to model the relationship:

   $$Income_i = \alpha + \beta_1 CognAbility_i + \beta_2 PotentExper_i + \beta_3 Female_i + u_i$$

   Interpret the parameter estimates of the model (be specific).

   (b) Add *Age* and $PotentExper^2$ to the model above.

   How do the results differ? Discuss which of the MLRM assumptions discussed by Wooldridge are implicated by your results for these variables.

   Also compare the R^2 obtained from the two regressions. Will the R^2 ever decrease with the inclusion of additional variables? Why or why not?

   (c) Now compare R^2 in the model from 4(a) to the following:

   $$Income_i = \alpha + \beta_1 SchoolYrs_i + \beta_2 PotentExper_i + \beta_3 Female_i + u_i$$

   Which model explains more of the total variation in income percentile? Can you think of any possible problems with using either of the two models?

[2] Interval censoring is when the true value of a variable is not observed, only that it lies within a certain interval. Censoring, along with the fact that percentiles are constrained between 0 and 100, make inference using OLS somewhat more problematic. More advanced econometric methods, such as interval regression, have been developed for these purposes. We will ignore these subtleties for now.

5. (For three person groups only:) You are now interested in whether continuing use of mathematics in one's profession might affect cognitive ability and thus income. Specifically, you are interested in the effects of using advanced math at work, captured by *g_q03h* in the dataset.

   (a) Perform the same basic exploratory analysis on *g_q03h* as in step 2.

   (b) Suitably reformatting *g_q03h* for quantitative analysis, provide the mean and standard of this variable according to its original values scheme.

   (c) Now using *g_q03h*, creating a variable equal to the # of times each month an individual uses advanced math or statistics at work.[3] Provide summary statistics for the recoded variable.

   (d) With your 'advanced math use at work' frequency variable, use simple linear regression to estimate the following relationships:[4]
   (1) the effect of math use at work on cognitive ability percentile.
   (2) the effect of cognitive ability percentile on earnings percentile.
   (3) the effect of math use at work on earnings percentile.
   Interpret the results of the regressions and comment on what they suggest about your initial hypothesis.

   (e) You get dizzy for a minute and think maybe math use at work also affects an individual's years of schooling.[5] Now run the following regressions:
   (1) the effect of math use at work on years of schooling.
   (2) the effect of years of schooling on earnings percentile.
   Interpret the results of the regressions. Comment on whether the first regression make sense from a causal standpoint and give your opinion of why you obtain the estimates you do in light of this analysis. Demonstrate your explanation by running an additional regression, if you can.

   (f) What is a potential concern with inference from part (d) in light of your results from part (e)? What general concerns might you have about inference from the regressions in part (d).

Good luck!

[3] Note: Since the original variable is not given in directly interpretable frequencies, you will need to decide (and comment) on how to equate the original frequency descriptions with 'times per month').

[4] For regressions in question 5, assume that we are only interested in the relationship for working age (age 30 to 65) individuals who are employed.

[5] Of course, some individuals might have attained additional years of schooling after starting at their current job, in which case this regression would make more sense. For the purposes of this exercise, instead assume that schooling strictly predates math use in one's current job.