Econometrics A: Problemset 1

LAURA MAYORAL
Institute for Economic Analysis and Barcelona GSE

January-February 2020

Deadline: January 24rd before 14:00. Please submit your answers in paper or electronically –scanned or typed, as you prefer–, to your TA, sanghyun.park@insead.edu.

Exercises

1. The production process of a good is considered to work satisfactorily if less than 1% of the produced units are defective. To determine whether this is the case, a team has gathered data on the quality of the products. More specifically, 10,000 products were examined and it was found that 106 of them were defective. Define the random variables X$_i$=1 if product i is defective, and 0 otherwise, for product $i = 1 \dots N$, and assume that these variables are i.i.d.

(1) Provide an estimator of the probability that one product is defective.

$\bar{X_N} = \sum_{i=1}^{N} X_i/N$

(2) Using the collected data, provide an estimate of that probability.

$\bar{X_N} = \sum_{i=1}^{N} X_i/N = 106/10,000 = 0.0106$

(3) Provide an estimate of the variance of the random variable X_i. (Hint: notice that X_i is a Bernouilli random variable).

$var(X_i) = p(1-p)$ where p is the probability that one product is defective.
$\widehat{var(X_i)} = \hat{p}(1-\hat{p})$ where $\hat{p} = \bar{X_N}$
$\widehat{var(X_i)} = 0.0106 * (1 - 0.0106) \approx 0.0105$

(4) Use the Law of Large Numbers to describe the limit of the estimator provided in a)

Since X_i follows i.i.d., we can apply Law of Large Numbers (LLN).
In other words, $\bar{X_N} \overset{p}{\to} E[X] = p$

(5) Use the Central Limit theorem to describe the asymptotic distribution of the estimator in a)

Since X_i follows i.i.d., we can apply Central Limit Theorem (CLT).
In other words, $\sqrt{N}(\bar{X_N} - p)/(\sqrt{p*(1-p)}) \xrightarrow{d} N(0,1)$ where $E[X] = p$

(6) Construct an (asymptotic) test of hypotheses at the 5% level (i.e., $\alpha = .05$) to determine whether the production process works well or not. To do that carefully describe

(a) the null and the alternative hypotheses

Null hypothesis: $H_0 : p \leq 0.01$
The alternative hypothesis: $H_1 : p > 0.01$

(b) the test statistic that you can use to test those hypotheses

We can use the t-test as the test statistic

(c) describe the critical region (i.e., the values of the test-statistic for which you reject the null hypothesis)

Under the 5% significance level,
we reject the null if$\sqrt{N}(\hat{p} - p)/\sqrt{\hat{p} * (1 - \hat{p})} > 1.64$.

In other words, the critical region satisfies
$\sqrt{10,000}(\hat{p} - 0.01)/\sqrt{\hat{p} * (1 - \hat{p})} > 1.64$

(d) compute the value of the test using the data of the problem.

$\sqrt{10,000}(0.0106 - 0.01)/\sqrt{0.0106 * (1 - 0.0106)} = 0.585885$

(e) Can you reject H_0? Clearly justify your answer.

Since $0.585885 < 1.64$, we cannot reject null hypothesis.

(f) Compute the p-value associated to the test-statistic you computed in (c)

$Prob(t - statistic > 0.585885|$ H0 is true)$= 0.2789$

2. Read Chapter 2 in Wooldridge and solve problem 2.3.
See the attached file.

Computer Practise

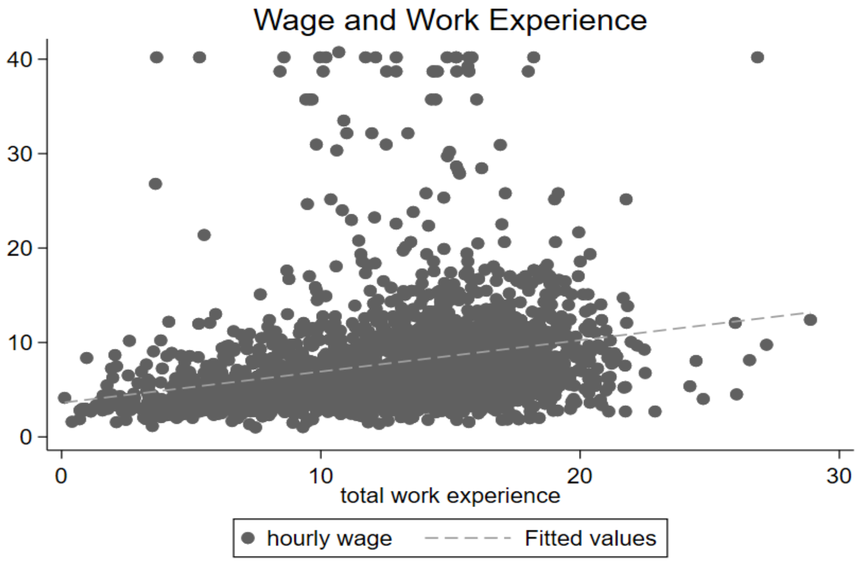
3. Binscatter is a useful STATA command for data visualization, that provides a non-parametric estimation of the conditional expectation. Read this document to understand what it does.

i) install the binscatter command in STATA.

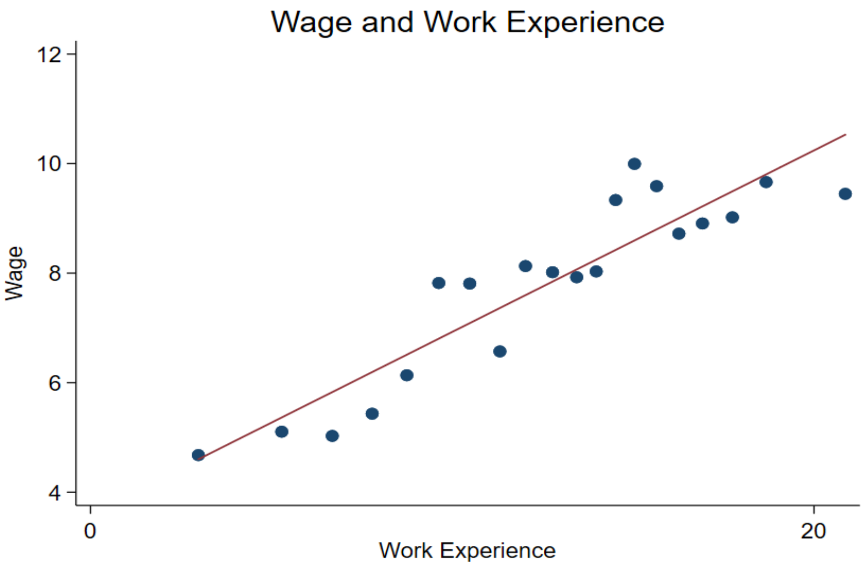
ii) Type “help binscatter” to know more about the options that this command provides.

iii) Load the dataset “nlsw88.dta” in the stata memory (hint: type “help sysuse” to learn how you can do this.)

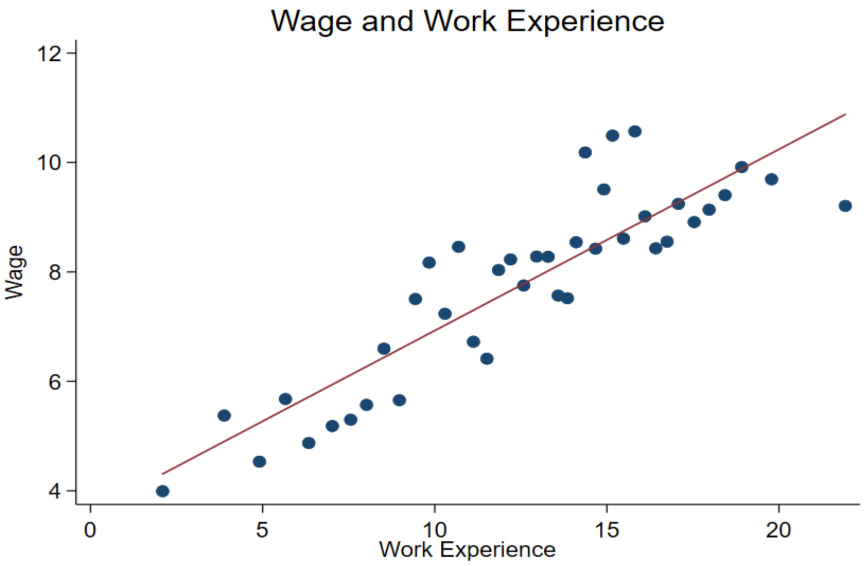
iv) Plot a scatter plot relating wages and ttl_exp (total work experience). Add a line to the plot that describes the best linear fit between these variables. Describe the plot. (Hint: you can check this page for help `https://stats.idre.ucla.edu/stata/modules/graph8/intro/introduction-to-graphs-in-stata/`



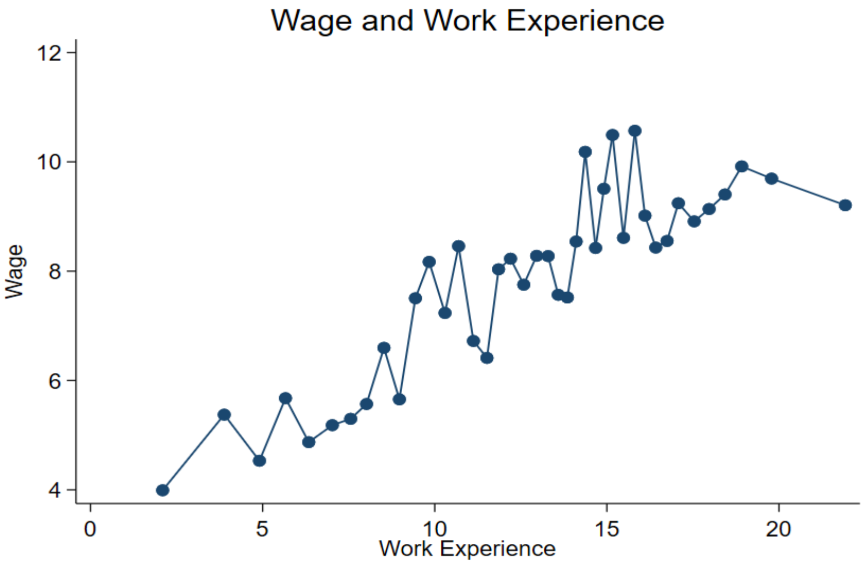
v) Now plot a binscatter plot with the same variables. Compare the two graphs.
We now have less points in the figure as we binned observations. Each point represents mean of the observations in the same bin. It is easier to see the relationship between two variables compared to the scatter plot.



v) Change the default number of bins in your binscatter plot to 40



vi) Produce a binscatter that connects the different bins (hint: use the linetype option)



4. Use Stata to answer the following questions

(1) Empirical papers usually start by presenting a **table of summary statistics**. This table typically includes the number of observations, mean, standard deviation, minimum and maximum values, kurtosis, skewness, among other statistics. Load in Stata the dataset mroz_ps0.dta; See Wooldridge p. 59 for a description of the data. Produce a table summarizing the main variables you'll find there.

Note: you can use different STATA commands to create tables of summary statistics that are easily exportable to other documents, see for instance tabstat or latabstat (for latex output).

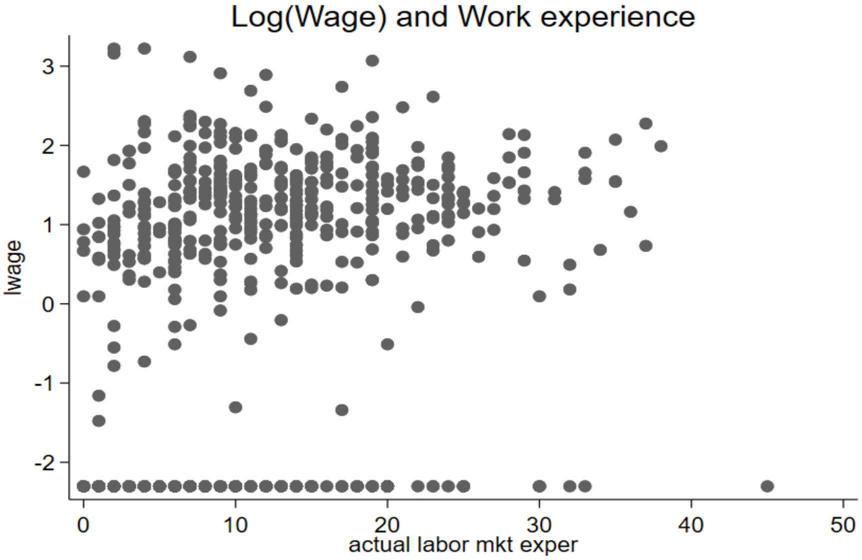
TABLE 1. Summary Statistics

stats	**wage**	**exper**	**educ**	**age**	**kidslt6**	**kidsge6**
N	753	753	753	753	753	753
mean	2.374565	10.63081	12.28685	42.53785	.2377158	1.353254
sd	3.241829	8.06913	2.280246	8.072574	.523959	1.319874
min	0	0	5	30	0	0
max	25	45	17	60	3	8
kurtosis	15.79665	3.70137	3.744087	1.981077	8.254322	3.809829
skewness	2.777771	.9605118	.021034	.150879	2.309519	.9077226

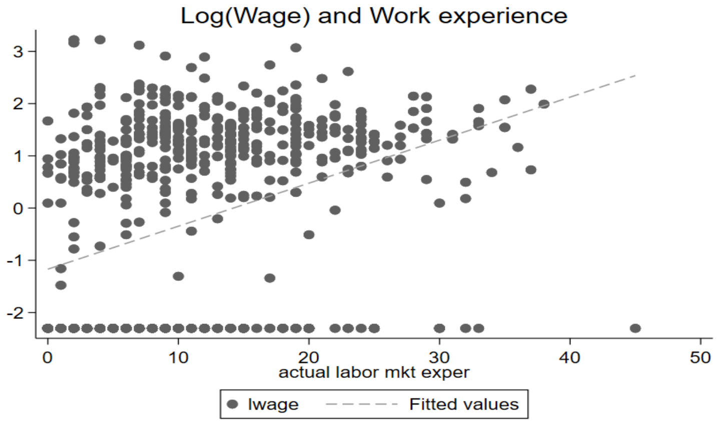
(2) For reasons that we will see later on in the course, many times we'll model variables in logs. Generate a new variable (lwage) that is the log of the variable *wage*. Note: notice that $log(0) = -\infty$. Thus, when variables contain zeros, we typically add a small quantity, .1 for instance, and then compute the log. Add a label to the new variable (for instance, label the new variable: log of (wage+0.1)).

```
gen lwage=log(wage+0.1)
```

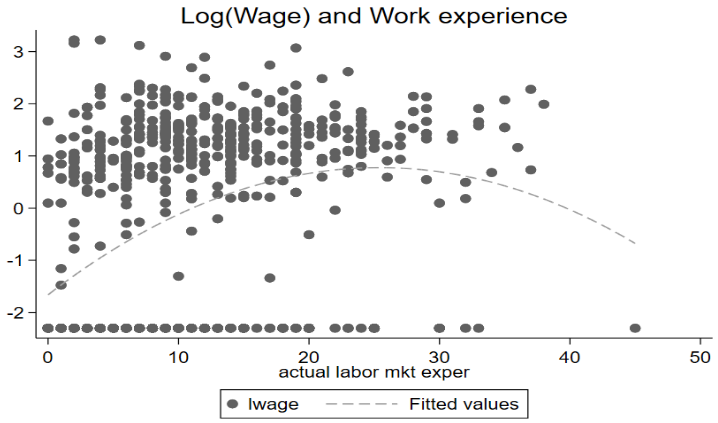
(3) Plot a scatter plot showing the relationship between lwage (Y axis) and exper (X axis). What do you observe?
It is difficult to see the relationship through scatter plot without fitted line.



(4) Add to the previous scatter the line that best fits the data. The slope of the line gives you an idea of the type of (linear) relationship between those variables. What do you see now? Is the relationship positive or negative?
According to the fitted line, experience and wage have a positive relationship.



(5) Add to the scatter plot above the quadratic fit (instead of the linear one) between these two variables. What do you observe now? How would you interpret this result?
According to the quadratic fit, there is inverted U shaped relationship between experience and wage. In other words, the wage increases with experience up to certain point, but, beyond that point, wage decreases with experience.


Log(Wage) and Work experience
actual labor mkt exper
● lwage
– – – – – Fitted values