Econometrics A HW2

Xinyu Liu

January 23, 2021

1.

Read carefully Handout 0 (Large sample theory) and answer the following questions.

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

According to the assumption, the outcome of product quality X follows a Bernoulli distribution with probability of being defective p for $X_i = 1$. Denote the whole sample as X. The distribution can be written as: $X_i \sim B(p), \forall i$. The method of moments suggest that:

$$E(X) = p$$

The estimator is constructed using the sample moments to replace the population moments, therefore:

$$\hat{p} = \frac{\sum_{i=1}^{N} X_i}{N}$$

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

Given the data:

$$p(\hat{X}) = \frac{106}{10000} = 1.06\%$$

3. Define an estimator of the variance of the random variable X_i. Using your estimate of p, provide an estimate for the variance of X_i.

First of all, we know that Bernoulli distribution has variance $p(1-p)$, Therefore:

$$Var(X_i) = p(1-p)$$

Intuitively, once we have the estimate of p, we can also provide an estimate for $Var(X_i)$ as:

$$\hat{S} = \hat{p}(1-\hat{p})$$

I show that this aligns with what method of moments suggests. We can rewrite the variance as:

$$var(X_i) = E(X_i^2) - E(X_i)^2$$

Using method of moments, denote the estimator as $\hat{S}$:

$$\begin{aligned}\hat{S} &= \frac{\sum_{i=1}^{N} X_i^2}{N} - (\frac{\sum_{i=1}^{N} X_i}{N})^2 \\ &= \frac{\sum_{i=1}^{N} X_i}{N} - (\frac{\sum_{i=1}^{N} X_i}{N})^2 \\ &= \hat{p}(1-\hat{p})\end{aligned}$$

Therefore the estimate for variance is:

$$\begin{aligned}\hat{S}(X) &= \hat{p}(X)(1-\hat{p}(X)) \\ &= 1.06\%(1-1.06\%) \\ &= 0.0105\end{aligned}$$

4. Use the Law of Large Numbers to describe the limit of the estimator provided in (1). Is this estimator consistent?

The weak law of large numbers implies that the sample mean of a random variable converges in probability to its expectation as $N \to \infty$. I.e.:

$$\lim_{N\to\infty} Pr((\hat{p} - p) < \varepsilon) = 1, \forall \varepsilon > 0$$

As $\hat{p} \xrightarrow{p} p$, it is a consistent estimator.

5. Use the Central Limit theorem to describe the asymptotic distribution of the estimator in (1).

The CLT claims that:

$$\lim_{N\to\infty} \sqrt{N}(\hat{p} - p) \sim N(0, p(1-p))$$

2.

Read carefully Handout 0 (hypothesis testing section). Using the data in the previous exercise construct an (asymptotic) test of hypotheses at the 5% level (i.e., $\alpha = .05$) to determine whether the production process works well or not (Remember that it works well if $p < 0.1$).

1. Carefully define the meaning of α (type I error); also define the meaning of the type II error?

A type one error happens if we reject the null hypothesis when the null is actually true. A type two error happens if we accept the null when it is actually false. α represents the size of the test. It is the probability of having type one error.

2. Describe the null and the alternative hypotheses.

The null hypothesis H_0 says: $p < 1\%$, i.e. the production process of a good is considered to work satisfactorily; the alternative H_1 says $p \geq 1\%$. For that the test is uniformly the most efficient test. And we essentially care the most whether p is possible to be 1. I can rewrite the hypothesis as the following:

The null hypothesis H_0 says: $p = 1\%$, i.e. the production process of a good is considered to work satisfactorily; the alternative H_1 says $p > 1\%$.

3. Describe the test statistic that you can use to test those hypotheses.

The most common statistic is t statistic, which is constructed as follows:

$$t = \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})}/\sqrt{N}} \sim T(N-1)$$

Because N is big enough, we can approximate it with the standard normal distribution:

$$t = \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})}/\sqrt{N}} \sim N(0,1)$$

Substitute in the estimates obtained previously, and let $p = p_0 = 1\%$:

$$t = \frac{0.0106 - 0.01}{0.1024} * 100 \approx 0.59$$

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

Note that this is a one-sided test, where $\alpha = 0.05$ means that $P(t > t_c) \leq 0.05$ mull hold for rejection:

$$t_c = \Phi^{-1}(95\%) = 1.645$$

Therefore the critical region of t is where

$$t > 1.645$$

5. compute the value of the test using the data the problem.

See (3).

6. Can you reject H_0? Clearly justify your answer.

No I cannot, because t falls outside the critical region of rejection and is much smaller than the critical value.

7. Define the concept of p-value.

P-value is defined to by the smallest size of the test that can be chosen in order to reject the null.

8. Compute the p-value associated to the test-statistic you computed.

p-value is calculated using the CDF of normal:

$$\text{p-value} = Pr(t > 0.59) = \Phi(0.59) \approx 0.28$$

Therefore, α must be as large as 0.28 to reject the null.

3.

Let y and z be random scalars, and let $\boldsymbol{x}$ be a $1k$ random vector, such that $x_1 = 1$. Consider the population model:

$$\begin{aligned} E(y|\boldsymbol{x}, z) &= \boldsymbol{x}\gamma + \delta z \\ Var(y|\boldsymbol{x}, z) &= \sigma^2 \end{aligned}$$

1. Write a probabilistic model of y as a function of the conditional expectation specified above and a random disturbance u.

In class we have proved that the best statistical predictor of y given $\boldsymbol{x}, z$ is its conditional expectation. Therefore, the model can be written as follows:

$$\begin{aligned} y &= E(y|\boldsymbol{x}, z) + u \\ &= \boldsymbol{x}\gamma + \delta z \end{aligned}$$

2. Under the assumptions above, compute the conditional mean and the conditional variance of u. Is u conditionally homoscedastic? (that is, is the (conditional) variance unrelated the values of x or z). Justify your answer.

$$\begin{aligned} E(y|\boldsymbol{x}, z) &= E(E(y|\boldsymbol{x}, z)|\boldsymbol{x}, z) + E(u|\boldsymbol{x}, z) \\ &= E(y|\boldsymbol{x}, z) + E(u|\boldsymbol{x}, z) \\ \Rightarrow E(u|\boldsymbol{x}, z) &= 0 \\ Var(y|\boldsymbol{x}, z) &= 0 + Var(u|\boldsymbol{x}, z) \\ &= Var(u|\boldsymbol{x}, z) \\ \Rightarrow Var(u|\boldsymbol{x}, z) &= \sigma^2 \end{aligned}$$

This indicates that the error term satisfy the mean zero conditional expectation condition, and that it is homoscedastic because its conditional variance does not depend on other variables.

3. Compute the unconditional mean and the unconditional variance of u.

$$\begin{aligned} E(u) &= E(E(u|\boldsymbol{x}, z)) = E(0) = 0 \\ Var(u) &= Var(E(u|\boldsymbol{x}, z)) + E(Var(u|\boldsymbol{x}, z)) \\ &= Var(0) + E(\sigma^2) \\ &= \sigma^2 \end{aligned}$$

4. The main assumption that we need for identification of γ and δ is that $E((\boldsymbol{x}z)'u) = 0$. Under the assumptions above, is this condition met? clearly justify your answer.

Yes it is satisfied. I show it by using the LIE:

$$\begin{aligned} E((\boldsymbol{x}z)'u) &= E[E((\boldsymbol{x}z)'u|\boldsymbol{x}, z)] \\ &= E[(\boldsymbol{x}z)'E(u|\boldsymbol{x}, z)] \\ &= E[(\boldsymbol{x}z)'0] = 0 \end{aligned}$$

5. Assume now that the variable z cannot be observed, and that you are considering this model instead.

$$y = \boldsymbol{x}\gamma + u^*$$

Yes we can identify γ. First, we can set the mean of z to zero, i.e. to move its mean to the constant part in x_1. Therefore for the new error term u^*, it satisfies:

$$E(\boldsymbol{x}'u^*) = 0$$

As

$$\begin{aligned} E(\boldsymbol{x}'(u+\delta z)) &= E(\boldsymbol{x}'u) + \delta E(\boldsymbol{x}'z) \\ &= 0 + \delta Cov(\boldsymbol{x}', z) = 0 \end{aligned}$$

To identify γ, we use method of moments and get the same result in class:

$$\gamma = E[\boldsymbol{x}'\boldsymbol{x}]^{-1}E[\boldsymbol{x}'y]$$

6. Will your answer in v) change if x and z were correlated? Justify your answer.

It will change and we will no longer be able to identify the true γ. This is caused by omitted variable (unobservables that correlate with $\boldsymbol{x}$) problem. Suppose we continue to use the previous result as γ^*, what we eventually get is as follows:

$$\begin{aligned} \gamma^* &= E[\boldsymbol{x}'\boldsymbol{x}]^{-1}E[\boldsymbol{x}'y] \\ &= E[\boldsymbol{x}'\boldsymbol{x}]^{-1}E[\boldsymbol{x}'(\boldsymbol{x}\gamma + u^*)] \\ &= \gamma + E[\boldsymbol{x}'u^*] \\ &= \gamma + E[\boldsymbol{x}'(u+\delta z)] \\ &= \gamma + \delta E[\boldsymbol{x}'z] \\ &= \gamma + \delta Cov[\boldsymbol{x}', z] \neq \gamma \end{aligned}$$

4. Computer Practise

Binscatter is a useful STATA command for data visualization, that provides a non-parametric estimation of the conditional expectation.

Please kindly find following the graphs and codes in Appendix. The main difference between 'scatter' and 'binscatter' is that the latter involves a stage of nonparametric extrapolation, which is to divide the sample into equal size bins by 'x' variable, and compute the sample mean of all points in each bin, as a 'representative point of the bin'. This helps make the graph informative when there is a large sample. However, there are several caveats for using 'binscatter'. First, many functional form can be represented by the same fitted line. This also have to do with the subjective choice of bin numbers. Usually the more complicated the function, the more bins it takes to reveal certain information about the functional form. As we can also see from the density of the sample points, they are heterogeneous along the variable values, meaning we will get less accurate representation in bins where the sample are more spare. Latest, I plot in the last graph using quadratic fitted line, which apart from level difference, also shows a concave wage curve for married ones. This curvature difference is not captured in the linear line case (I did not attach the graph here but have tested it). This may suggest that married people are treated differently in the labor market, or it can also be just a sampling issue.

Figure 1

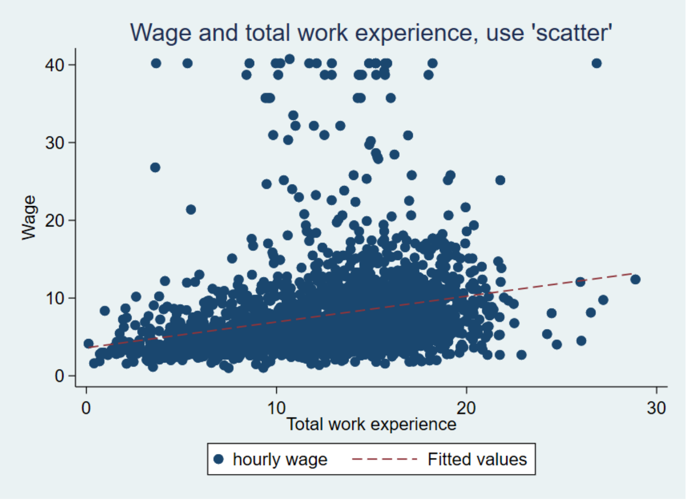


Figure 2

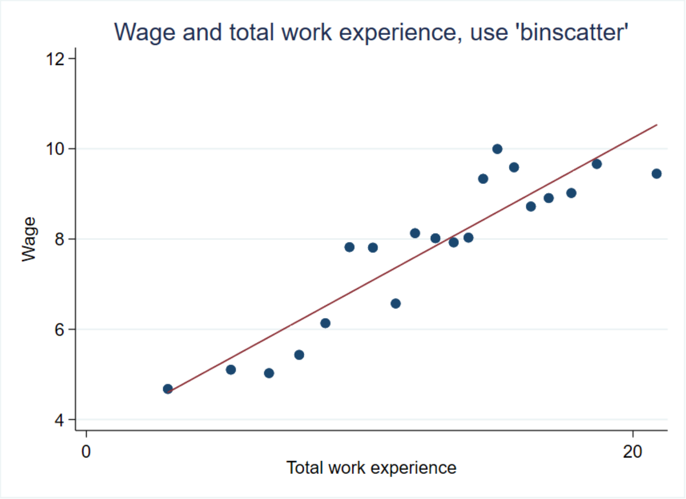


Figure 3

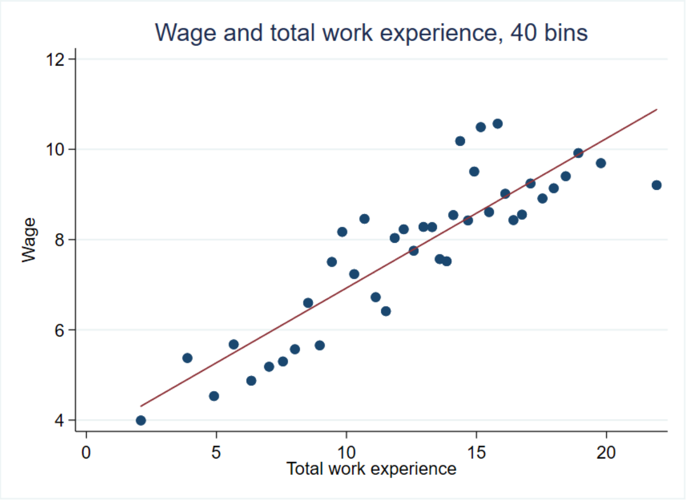


Figure 4

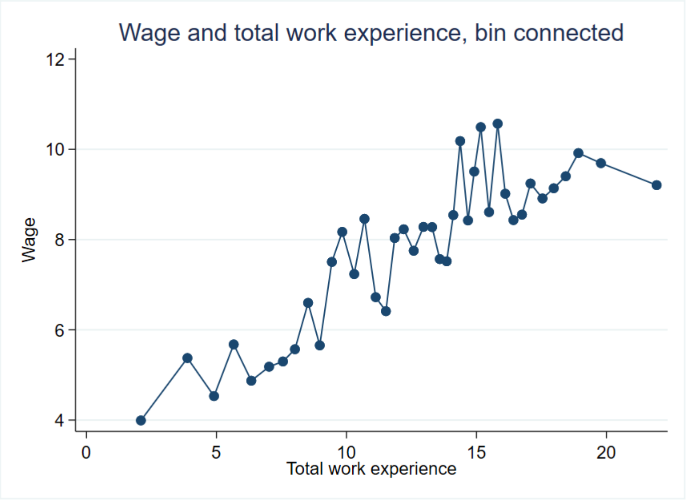
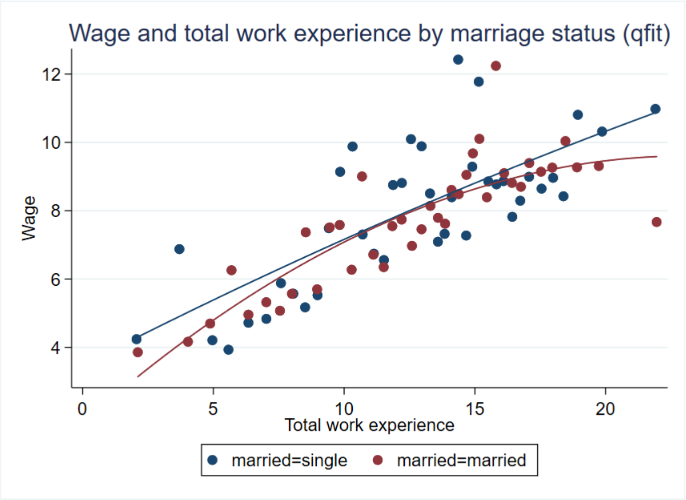


Figure 5



Appendix

```
capture log close
capture drop _all

pwd

set logtype text
log using hw_2.txt, replace

*********************************************************************************
*** Install package and check its guide notes
*********************************************************************************
ssc install binscatter, replace
help binscatter

sysuse nlsw88.dta, clear
twoway (scatter wage ttl_exp) (lfit wage ttl_exp, lpatt(dash)), title("Wage and total work experience, use ’scatter’") scheme(s2color) plotregion(style(none)) ylabel(,angle(0)) xtitle("Total work experience") ytitle("Wage")
graph export "D:\INSEAD\Course\P3\Econometrics A\Econometrics-A-2021\HW2\q4p1.png", as(png) replace
binscatter wage ttl_exp,nq(20) title("Wage and total work experience, use ’binscatter’") scheme(s2color) plotregion(style(none)) ylabel(,angle(0)) xtitle("Total work experience") ytitle("Wage")
graph export "D:\INSEAD\Course\P3\Econometrics A\Econometrics-A-2021\HW2\q4p2.png", as(png) replace
* Change the default number of bins in your binscatter plot to 40
binscatter wage ttl_exp,nq(40) title("Wage and total work experience, 40 bins") scheme(s2color) plotregion(style(none)) ylabel(,angle(0)) xtitle("Total work experience") ytitle("Wage")
graph export "D:\INSEAD\Course\P3\Econometrics A\Econometrics-A-2021\HW2\q4p3.png", as(png) replace
*Produce a binscatter that connects the different bins (hint: use the linetype option)
binscatter wage ttl_exp,nq(40) title("Wage and total work experience, bin connected") linetype(connect) scheme(s2color) plotregion(style(none)) ylabel(,angle(0)) xtitle("Total work experience") ytitle("Wage")
graph export "D:\INSEAD\Course\P3\Econometrics A\Econometrics-A-2021\HW2\q4p4.png", as(png) replace
*Compute two different binscatters
binscatter wage ttl_exp,by(married) nq(40) title("Wage and total work experience by marriage status (qfit)") linetype(qfit) scheme(s2color) plotregion(style(none)) ylabel(,angle(0)) xtitle("Total work experience") ytitle("Wage")
graph export "D:\INSEAD\Course\P3\Econometrics A\Econometrics-A-2021\HW2\q4p5.png", as(png) replace
```