Econometrics A: Problemset 3

LAURA MAYORAL
Institute for Economic Analysis and Barcelona GSE

Winter 2020

Deadline: January 30th, before 15:00. Please submit your answers in paper or electronically to your TA, sanghyun.park@insead.edu.

1. Let y and z be random scalars, and let $\mathbf{x}$ be a $1\times$k random vector, such that $\mathbf{x}_1 = 1$. Consider the population model:

$$E(y|\mathbf{x}, z) = \mathbf{x}\beta + \delta z$$

$$Var(y|\mathbf{x}, z) = \sigma^2$$

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

$y = x\beta + \delta z + u$

ii) Under the assumptions above, compute the conditional mean and conditional variance of u. Is u homoscedastic? justify your answer.

Yes, it is homoscedastic as its conditional variance is constant (i.e. $Var(u|x, z) = \sigma^2$).

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

Yes. By the law of iterated expectations, $E(u|x, z) = 0$ implies $E((xz)'u) = 0$

iv) Suppose that $\mathbf{x}_2$ in equation above is a variable measured in km. Suppose that another researcher specifies a similar model but with variable $\mathbf{x}_2$ measured in cm, $\mathbf{x}_2^*$. Let $\beta{*}_2$ denote the coefficient associated to $\mathbf{x}_2^*$. How are β_2 and $\beta{*}_2$ related? Suppose now that you estimate both equations by OLS. What is the relationship between $\hat{\beta_2}$ and $\hat{\beta{*}_2}$?

We have $y = \beta_0 + \beta_1 x_1 + \beta_2^* x_2^* + u$. If we replace x_2^* with x_2,
then we will get $y = \beta_0 + \beta_1 x_1 + \beta_2^* * 1000 * x_2 + u$
$y = \beta_0 + \beta_1 x_1 + (\beta_2/1000) * 1000 * x_2 + u$
From the above equations, $\beta_2^* = \beta_2/1000$.

2. Consider again the model in the previous exercise. Assume now that z is unobservable so that you estimate:

$$y = \mathbf{x}\beta + \epsilon$$

a) Notice that $\epsilon = \delta z + u$. Is $\hat{\beta}$ (the OLS estimator of β) consistent in this case?

In many cases, no. This is so because $E(x'\epsilon) \neq 0$ in general.

b) Assume now that $E(\mathbf{x}'z) = 0$ (i.e., these variables are orthogonal). Is $\hat{\beta}$ consistent in this case. Proof mathematically your answer.

In this case, $E(x'\epsilon) = \delta E(x'z) + E(x'u) = 0$

c) Under the assumptions above and b), which estimator of β would have largest variance, that obtained in b) or another one that also includes z in the regression?. Clearly justify your answer.

Remember that under the assumptions above, the estimators from both models are unbiased. In addition, the estimator of β obtained in the model containing both x and z is BLUE. Notice that the estimator obtained in the reduced model is a particular case and thus, it has to have a higher variance.

3. Consider once more the model in Problem 1. To simplify the notation, assume that δ=0 and that we estimate β in the model $y = \mathbf{x}\beta + u$.

a) How is $\hat{\beta}$ distributed under the assumptions of that problem? Provide a consistent estimator of the variance of $\hat{\beta}$.

By CLT, $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \sigma^2 E(x'x)^{-1})$

b) Now consider the case where $Var(y|\mathbf{x})$ is a function of $\mathbf{x}$ (i.e., it is not constant with respect to $\mathbf{x}$ as it was in Problem 1). How does the distribution of $\hat{\beta}$ change in this case with respect to case a) above?. Provide a consistent estimator for the variance-covariance matrix of $\hat{\beta}$.

$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, E(x'x)^{-1}E(u^2x'x)E(x'x)^{-1})$

c) Consider again the general case where δ is different from zero but z is not included in the equation (and can be correlated with $\mathbf{x}$. Does $\sqrt{n}(\hat{\beta} - \beta)$ converge or diverge as n tends to ∞?

In this case, $\hat{\beta}$ is not consistent since $E(x'\epsilon) \neq 0$. In other words, $\hat{\beta} - \beta$ does not converge to 0. As N increases, $\sqrt{N}$ diverges to ∞. If we multiply two, then the product will diverge $+\infty$ or $-\infty$ (the sign will depend on the sign of $\hat{\beta} - \beta$).

TABLE 1. OLS Regression Results

	(1) lwage b/se	(2) lwage b/se	(3) lwage b/se	(4) lwage b/se	(5) lwage b/se	(6) lwage b/se	(7) lwage b/se
educ	0.060***	0.039***	0.031***	0.034***	0.046***	0.046***	0.049***
	[0.006]	[0.007]	[0.007]	[0.007]	[0.008]	[0.008]	[0.008]
iq		0.006***	0.004***	0.005***	0.005***	0.005***	0.003***
		[0.001]	[0.001]	[0.001]	[0.001]	[0.001]	[0.001]
kww			0.010***	0.006***	0.006***	0.007***	0.004*
			[0.002]	[0.002]	[0.002]	[0.002]	[0.002]
age				0.017***	0.008	0.008	0.008
				[0.005]	[0.005]	[0.005]	[0.005]
exper					0.013***	0.013***	0.012***
					[0.004]	[0.004]	[0.004]
hours						-0.005**	-0.006***
						[0.002]	[0.002]
married							0.200***
							[0.039]
black							-0.148***
							[0.040]
south							-0.086***
							[0.028]
urban							0.174***
							[0.027]
_cons	5.973***	5.658***	5.563***	5.011***	5.012***	5.216***	5.232***
	[0.082]	[0.094]	[0.093]	[0.180]	[0.177]	[0.190]	[0.188]
R_squared	0.097	0.130	0.155	0.168	0.178	0.186	0.263
Observations	935	935	935	935	935	935	935

Computer Practise

4. You would like to estimate the effect of an additional year of education on wages. Use the dataset NLS80.dta, which gathers data collected from a population of men.
A) Using the stata command *estout* produce a regression table

B) Using the results in column 7:

i) provide an interpretation of the direction of the relationship of educ and wages, as well as of the magnitude of the effect.

Education has a positive effect on lwage under 1% significance level. As education increases 1 year, wage increases about 5%.

ii) Provide an interpretation for the R^2 statistic and for the F-test of joint statistical significance. Do you reject the null hypothesis of that test?

$R^2 = 0.263$ means that our model explains about 26.3% of variation in lwage. Our F statistic is 38.79 with a p-value smaller than 0.01. In other words, we can reject the null under 1% significance level.

iii) Test the hypothesis that black, south, urban are all equal to zero using an F-test.

Our F statistic is 24.21 with a p-value smaller than 0.01. In other words, we can reject the null under 1% significance level.

iv) test the hypothesis that the coefficient of education is 0.5 using a two-tailed t-test, $\alpha = 0.05$. Repeat the same exercise using where the alternative hypothesis is $H_1 : \beta > 0.5$

Our t statistic is $\sqrt{3326.77} = 57.68$ (or $\frac{beta-0.5}{se} \approx \frac{0.049-0.5}{0.0078}$) with a p-value smaller than 0.01. In other words, we can reject the null under 5% significance level. For the one tailed test, we cannot reject the null under 5% significance level, since p-value is $1 - 0.5 * p'$ where p' is the p-value from the previous test.

v) Use the command predict to obtain the predicted values for lwage. Also obtain the residuals.

vi) Use the command estat hettest to test for heteroskedasticity in the residuals. What do you conclude from this test?

We have p-value 0.1842, which means that we cannot reject the null hypothesis under 5% significance level. In other words, we cannot say that our model violates homoskedasticity assumption.

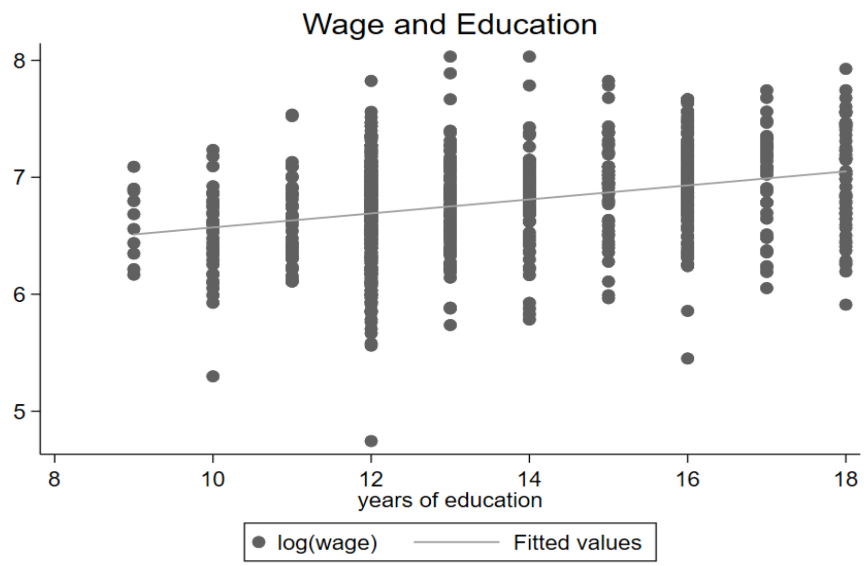
C) Introduce conditions in your regression: compute again the regression in column 7 in two different cases: for married men and for unmarried men (to do this, use the command if at the end of your regression, before the ",r" option). What do you observe?

Under 5%significance level, the positive effect of education on lwage for married men is significant, while that for unmarried men is insignificant (Table 2).

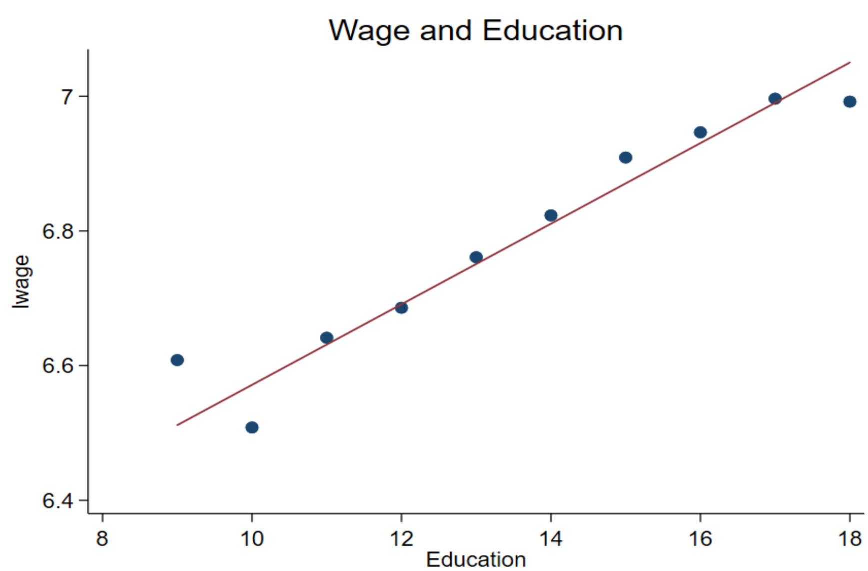
D) Use the binscatter command to produce a plot of lwage versus education, controlling by all the remaining variables.

i) Plot the scatter plot of lwage versus education. Also compute the binscatter of the same variables. Finally, plot a new binscatter, controlling this time for all the regressors in column 7.

- The scatter plot



- Binscatter without controls



- Binscatter with controls

ii) Use the commands pwcorr and graphic matrix to compute the correlation matrix of the regressors in column 7.

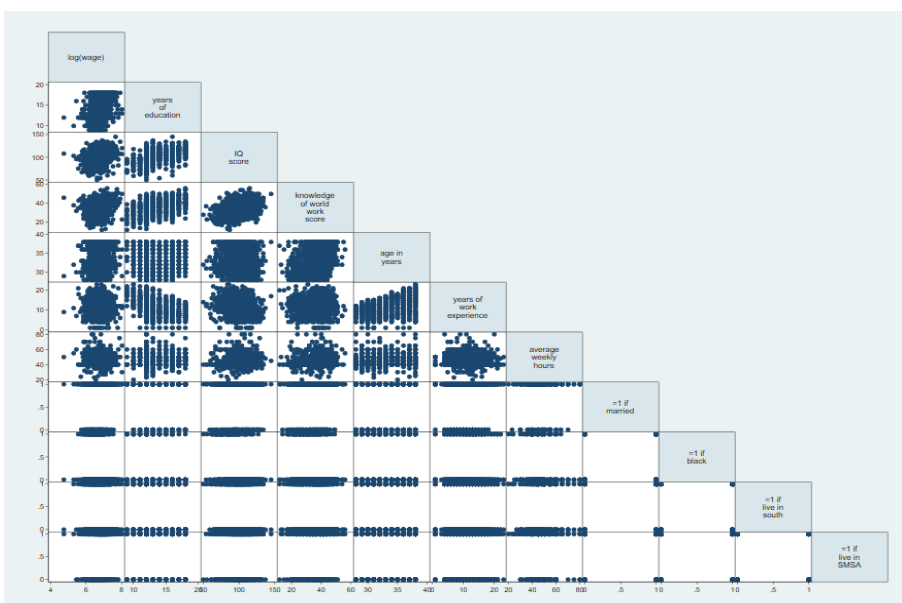


TABLE 2. OLS Regression Results

	unmarried lwage b/se	married lwage b/se
educ	0.032	0.052***
	[0.029]	[0.008]
iq	0.004	0.003***
	[0.003]	[0.001]
kww	0.002	0.004*
	[0.007]	[0.002]
age	0.005	0.008
	[0.016]	[0.006]
exper	0.008	0.013***
	[0.013]	[0.004]
hours	-0.009*	-0.006***
	[0.005]	[0.002]
black	-0.244**	-0.131***
	[0.105]	[0.044]
south	-0.113	-0.086***
	[0.082]	[0.030]
urban	0.255***	0.164***
	[0.094]	[0.028]
_cons	5.683***	5.363***
	[0.505]	[0.200]
R_squared	0.251	0.249
Observations	100	835