Problem Set I – Estimation and Inference (I) – Solution

1. The random variable $\hat{u}_i$ is not observed, even though it exist a random sample for the dependent variable y_i and the explanatory variables $x_{i1}, x_{i2}, \dots, x_{1k}$.
2. The R^2 and the $\bar{R}^2$ are equal if the dependent variable is logarithmic.
3. Under heteroscedasticity, the OLS estimator is unbiased.
4. Implies that the assumption MLR.4 does not hold, if the omitted variable is correlated with at least one explanatory variable included in the model.

5. EVIEW'S OUTPUT

Dependent Variable: ED
Method: Least Squares
Date: 10/28/16 Time: 16:43
Sample: 1 3796
Included observations: 3796

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	9.248261	0.173561	53.28529	0.0000
LDIST	-0.105777	0.020993	-5.038580	0.0000
BYTEST	0.098372	0.002972	33.10075	0.0000
TUITION	-0.225412	0.093224	-2.417964	0.0157

R-squared	0.232347	Mean dependent var	13.82929
Adjusted R-squared	0.231739	S.D. dependent var	1.813969
S.E. of regression	1.589952	Akaike info criterion	3.766338
Sum squared resid	9585.981	Schwarz criterion	3.772916
Log likelihood	-7144.510	Hannan-Quinn criter.	3.768676
F-statistic	382.5765	Durbin-Watson stat	1.873294
Prob(F-statistic)	0.000000		

a) Estimated equation

$$\widehat{ED} = 9.248261 - 0.105777ldist + 0.098372bytest - 0.225412tuition$$

Coefficients and correspondent standard errors:

$$\widehat{\beta_0} = 9.248261, \quad \widehat{\beta_1} = -0.105777ldist, \quad \widehat{\beta_2} = 0.098372, \quad \widehat{\beta_3} = -0.225412;$$

$$\widehat{\sigma_{\beta_0}} = 0.173561, \quad \widehat{\sigma_{\beta_1}} = 0.020993, \quad \widehat{\sigma_{\beta_2}} = 0.002972, \quad \widehat{\sigma_{\beta_3}} = 0.093224;$$

b)

Coefficient β_1: Regarding, all other factors fixed (*ceteris paribus*), if distance, $dist$, increases 1%, the years of completed education, ED, decrease $\frac{0.105777}{100}$, since it is a log-lin relation. Therefore, higher the distance, lower the completed education years. The negative sign in this coefficient makes sense, because it is expected that as higher is the distance that students live away from the university more difficult it would be for them to be enrolled.

Coefficient β_2: Regarding, all other factors fixed (*ceteris paribus*), if the base year test score, $bytest$, increases by 1 point the years of completed education, ED, increase 0.098372, since it is a lin-lin relation. Therefore, the higher the test score, the higher the completed education years. The positive sign in this coefficient makes sense, because the higher the score of this test we can expect the higher is the ability to understand and learn, leading to a positive effect on the completed education years.

c)

Rachel's information:

$$dist = 5 \Rightarrow ldist = \ln(5), bytest = 50, tuition = 1;$$

Hence, the predicted years of completed education for Rachel are

$$\widehat{ED} = 9.248261 - 0.105777 \times \ln(5) + 0.098372 \times 50 - 0.225412 \times 1 \approx 13.7712$$

If her completed years of education are 10 years, then Rachel is below the expected, compared with the population with the same characteristics. For Rachel, the value estimated for ED ($\widehat{ED} = 13.7712$) is over predicted.

The residual for Rachel is given by $ED - \widehat{ED} = 10 - 13.7712 = -3.7712$

d) The R^2 represents the proportion of the sample variation in the dependent variable (ED in this case) that is explained by the regression (independent variables).

Since $0 \leq R^2 \leq 1$, for R^2=0.232, a small part of the variation in ED is explained by the independent variables of the model. However, this is a reasonable value for cross-sectional data.

e) Confidence Interval for $\beta_3, \alpha = 1\%$

$$CI_{(1-\alpha)\%}(\beta_3) = (\widehat{\beta_3} \pm z_{\alpha/2} \times se(\widehat{\beta_3}))$$

$\widehat{\beta_3} = -0.225412;$

$se(\widehat{\beta_3}) = 0.093224;$

$z_{0.01/2} = 2.57583;$

Therefore, $CI_{99\%}(\beta_3) = (-0.225412 \pm 2.57583 \times 0.093224)$= (-0.46547; 0.014667)

Since the value 0 is included in this interval, it is possible that β_3 is statistically equal to zero, at a significance level of 1% (confidence level of 99%). This also means that there is enough evidence to say that the variable $tuition$ is not statistically significant to explain the dependent variable ED, at a significance level of 1%.

f) $\alpha = 5\%$

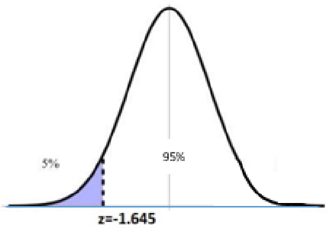
Test of hypothesis:

$H_0: \beta_1 = 0 \quad vs \quad H_1: \beta_1 < 0$

Test statistic:

$t = \frac{\widehat{\beta_1} - \beta_1}{se(\widehat{\beta_1})} \sim t_{(n-k-1)}$ (Under H_0), since the sample is large is it possible to write

$t = \frac{\widehat{\beta_1} - \beta_1}{se(\widehat{\beta_1})} \sim N(0,1)$ (Under H_0)



Observed value of the Test Statistic:

$t_{obs} = \frac{-0.105777 - 0}{0.020993} = -5.038580$

Rejection Rule:

Reject H_0 if $t_{obs} < c$, where c is the critical value.

$$\alpha = 5\% \Rightarrow c = z_{0.05} = -1.645$$

Conclusion:

Hence, $t_{obs} = -5.038580 < c$ = -1.645 ⇒ Reject H_0

Therefore, at significance level of 5% reject $H_0: \beta_1 = 0$. There is enough evidence to assume that $\beta_1 < 0$ meaning that β_1 is statistically significant (and therefore different from zero) and important to describe the dependent variable.

g) $\alpha = 10\%$

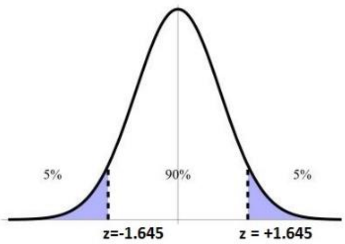
Test of hypothesis:

$H_0: \beta_1 = 0 \quad vs \quad H_1: \beta_1 \neq 0$

Test statistic:

$t = \frac{\widehat{\beta_1} - \beta_1}{se(\widehat{\beta_1})} \sim t_{(n-k-1)}$ (Under H_0) since the sample is large is it possible to write

$t = \frac{\widehat{\beta_1} - \beta_1}{se(\widehat{\beta_1})} \sim N(0,1)$ (Under H_0)



Observed value of the Test Statistic:

$t_{obs} = \frac{-0.105777 - 0}{0.020993} = -5.038580$

Rejection Rule:

Reject H_0 if $|t_{obs}| > c$, where c is the critical value.

$$\alpha = 10\% \Rightarrow c = z_{0.10/2=} z_{0.05} = 1.645$$

Conclusion:

Hence, $|t_{obs}| = 5.038580 > c$ = 1.645 ⇒ Reject H_0

The conclusion is the same as the previous question. Furthermore, we can conclude that a test of one side alternative at 5% significance level is equal to a test of two sided alternative at a 10% significance level.

h)

The estimated variance of the error term is given by $\widehat{\sigma^2} = \frac{SSR}{n-k-1}$ or by the square of the SE of the regression.

So, $\widehat{\sigma^2} = \text{SE of the regression}^2 = 1.589952^2 \approx 2.5279$

i) Under assumptions MLR.1 to MLR.4 the OLS estimator is unbiased.

If the $bytest$ is correlated with the error term the assumption MLR.4 is violated. Hence, the OLS estimator will be biased.