ECONOMETRICS

SOLUTION

Part 1: Multiple choice questions (20 points)

Circle the correct answer clearly.

1. In the simple linear regression model, the regression slope

a. indicates by how many percent *Y* increases, given a one percent increase in *X*.

b. when multiplied with the explanatory variable will give you the predicted *Y*.

c. indicates by how many units *Y* increases, given a one unit increase in *X*.

d. represents the elasticity of *Y* on *X*.

2. The OLS estimator is derived by

a. connecting the *Yi* corresponding to the lowest *Xi* observation with the *Yi* corresponding to the highest *Xi* observation.

b. making sure that the standard error of the regression equals the standard error of the slope estimator.

c. minimizing the sum of absolute residuals.

d. minimizing the sum of squared residuals.

3. Which of the following is true of the OLS t statistics?

a. The heteroskedasticity-robust t statistics are justified only if the sample size is large.

b. The heteroskedasticty-robust t statistics are justified only if the sample size is small.

c. The usual t statistics do not have exact t distributions if the sample size is large.

d. In the presence of homoskedasticity, the usual t statistics do not have exact t distributions if the sample size is small.

4. Consider the following simple regression model: $y = \beta_0 + \beta_1 x_1 + u$. Suppose *z* is an instrument for *x*. Which of the following conditions denotes instrument exogeneity?

a. Cov(z,u) > 0

b. Cov(z,x) > 0

c. Cov(z,u) = 0

d. Cov(z,x) = 0

5. Consider the equation, $Y = \beta1 + \beta2X2 + u$. A null hypothesis, $H0: \beta2 = 0$ states that:

a. X2 has no effect on the expected value of $\beta2$.

b. X2 has no effect on the expected value of Y.

c. $\beta2$ has no effect on the expected value of Y.

d. Y has no effect on the expected value of X2.

6. The following simple model is used to determine the annual savings of an individual based on his annual income and education: $Savings = \beta0+\partial0*Edu + \beta1*Inc+u$

The variable ‘Edu’ takes a value of 1 if the person is educated and the variable ‘Inc’ measures the income of the individual. The benchmark group in this model is: a. the group of educated people

b. the group of uneducated people

c. the group of individuals with a high income

d. the group of individuals with a low income

7. The significance level of a test is:

a. the probability of rejecting the null hypothesis when it is false.

b. one minus the probability of rejecting the null hypothesis when it is false.

c. the probability of rejecting the null hypothesis when it is true.

d. one minus the probability of rejecting the null hypothesis when it is true.

8. To decide whether or not the slope coefficient is large or small,

a. you should analyze the economic importance of a given increase in *X*.

b. the slope coefficient must be larger than one.

c. the slope coefficient must be statistically significant.

d. you should change the scale of the *X* variable if the coefficient appears to be too small.

9. If an independent variable in a multiple linear regression model is an exact linear combination of other independent variables, the model suffers from the problem of______.

a. perfect collinearity

b. homoskedasticity

c. heteroskedasticty

d. omitted variable bias

10. Which of the following is an example of a binary response model? a. MA model

b. ARCH model

c. GARCH model

d. Logit model

Part 2: True/false questions (20 points)

Indicate whether the statement below is true or false, no need for explanation.

1. Increasing the sample size can lead to a more precise estimate. **(T)**
2. The key assumption for the general multiple regression model is that all factors in the unobserved error term be correlated with the explanatory variables**. (F)**
3. A negative t-statistic indicates that the coefficient is not significant. **(F)**
4. The dummy variable coefficient for a particular group represents the estimated difference in intercepts between that group and the base group. **(T)**
5. The exclusion restriction for the IV requires that the instrument is significantly correlated with the endogenous variable. **(F)**

Part 3: Conceptual Questions (20 points)

Answer in detail all question:

1. Explain what happens to estimated parameters if we include an irrelevant variable in the model.

   ANSWER: If we include an irrelevant variable in the model such regression does not affect the unbiasedness of the OLS estimators but it only increases the variances of estimators. So, estimated β's will be generally inefficient, their variances will be greater than those of the true model. If we have high correlation between the irrelevant and relevant variables, then we will see that the estimated coefficient for the relevant variables have greater variance

2. Describe the Linear Probability Model and state its advantages and disadvantages compared to a probit/logit model.

   ANSWER:

   ADVANTAGES:

   Interpretability: it is very simple to estimate and use, since it is basically a multiple regression model. In LPM we can interpret the magnitude level of the estimated coefficient, which equals to the marginal effects. However, in the probit and logit models we cannot interpret the estimated coefficient and if we would like to interpret the magnitude effect then we should estimate marginal effects (**note**: signs of marginal effects and the coefficients are the same, but the size/magnitude is different).

   DISADVANTAGES:

   - *Heteroskedasticity*, however it can be fixed by using the white/clustered/robust standard errors.
   - *Possible to get predicted y < 0 or predicted y > 1*. This is the main problem with the LPM that we can't overcome with linear probability distribution function. Solution: Use the logit or probit model. These models are specifically made for binary dependent variables and always result in 0 < predicted y < 1.
   - Marginal probability effects sometimes logically impossible. It assumes *constant marginal effects.*

3. Describe the concept and the use of *p-value*.

ANSWER: The p-value is the probability that if the null hypothesis were true sampling variation would produce an estimate that is further away from the hypothesized value than our data estimate. The p-value tells us how likely it is to get a result like if the null hypothesis is true.

The p-value is used null hypothesis testing in order to measure the statistical significance. Having decided the significance level of 5%, if the p-value is lower than significance level we will reject the null.

4. Describe when/why we use the instrumental variables and state the necessary conditions that valid instruments should satisfy.

Answer: **IV Regression: Why/When?**

Three important threats to internal validity are:

- Omitted variable bias from a variable that is correlated with X but is unobserved (so cannot be included in the regression) and for which there are inadequate control variables;

- Simultaneous causality bias (X causes Y, Y causes X);

- Measurement error (X is measured with error) All three problems result in $E(u|X) \neq 0$. Instrumental variables regression can eliminate bias when $E(u|X) \neq 0$ – using an instrumental variable.

Necessary conditions that valid instruments should satisfy:

1.Instrument relevance: corr(Zi,Xi) ≠0, instruments must significantly be correlated with the endogenous variable;

2. Instrument exogeneity: corr(Zi,ui) = 0, instrument should only affect the outcome via the endogeneous variable, in other words, instrument must be uncorrelated with error term (note that in outcome variable we have two parts: regressor X which is endogeneous in this case and e (error term))

Part 4: Solve the problem (40 points)

Earnings functions attempt to find the determinants of earnings, using both continuous and binary variables. One of the central questions analyzed in this relationship is the returns to education.

a) Collecting data from 253 individuals, you estimate the following relationship

$$\begin{array}{l} \ln(Earn_i) = 0.54 + 0.083 * Educ,\ R^2 = 0.20 \\ \qquad\qquad\quad (0.14)\ \ (0.011) \end{array}$$

where *Earn* is average hourly earnings and *Educ* is years of education. What is the effect of an additional year of schooling?

- additional year of schooling will increase earnings by 8.3%

If you had a strong belief that years of high school education were different from a college education, how would you modify the equation? What if your theory suggested that there was a “diploma effect”?

ANSWER:

- D1 = 1, if highest degree from high school, 0 otherwise;
- D2 = 1, if college degree, 0 otherwise.

Include D1 and D2 as dummy predictors in a regression model, **by exluding EDUCATION!**

The intercept β0 corresponds to the baseline (D1 = 0, D2 = 0).

Baseline (all dummies 0): high school dropout (diploma matters)!!!

OR

Educ is years of education which is the same to current grade completed

we can answer the same question using margins. For instance, we can estimate expected hourly wages for different values of the independent variables. what is the expected hourly wage if we set grade/educ at values ranging from 12 to 16 years of education?

Taking the difference between these values, say, the difference between the expected value when grade=16 and when grade=12, gives us the effect of having a college education instead of a high school education on hourly wages.

b) You read in the literature that there should also be returns to on-the-job training. To approximate on-the-job training, researchers often use the so-called Mincer or potential experience variable, which is defined as *Exper = Age – Educ – 6*. Explain the reasoning behind this approximation.

Answer: In most surveys data on years of schooling and age in years is available, but not the number of years of actual labor market experience. So if researcher is interesting in the effect of Experience on Earnings, then s/he should construct experience variable based on age and education. In order to approximate *experience* we assume that people have been in the labor force continuously since leaving school that's why we subtract Education from Age, and -6 stands for the age at school entry (in some countries it is at age 7, however this is approximation so not a big deal) .

c) You incorporate the experience variable into your original regression

$$\begin{array}{l} \ln(Earn_i) = \underset{(0.16)}{-0.01} + \underset{(0.012)}{0.101} * Educ + \underset{(0.006)}{0.033} * Exper - \underset{(0.0001)}{0.0005} * Exper^2,\ R^2 = 0.34 \end{array}$$

What is the effect of an additional year of experience for a person who is 40 years old and had 12 years of education? What about for a person who is 60 years old with the same education background?

Answer:

In order to calculate marginal effects of experience (with specific characteristics) we take the derivative of equation with respect to experience:

0.033-0.0005*2*(Experience=40-12-6=22) =0.011 An additional year of experience for a person who is 40 years old and had 12 years of education will **increase** earnings by 1.1%

0.033-0.0005*2*(Experience=60-12-6=42) =-0.009 An additional year of experience for a person who is 60 years old and had 12 years of education will **decrease** earnings by 0.9%

d) Test for the significance of each of the coefficients in the equation from part c). Why has the coefficient on education changed so little?

Answer:

Degree of freedom: DF= 253-4=249

Critical value for two-tailed test at 5% significance level is = 1.96

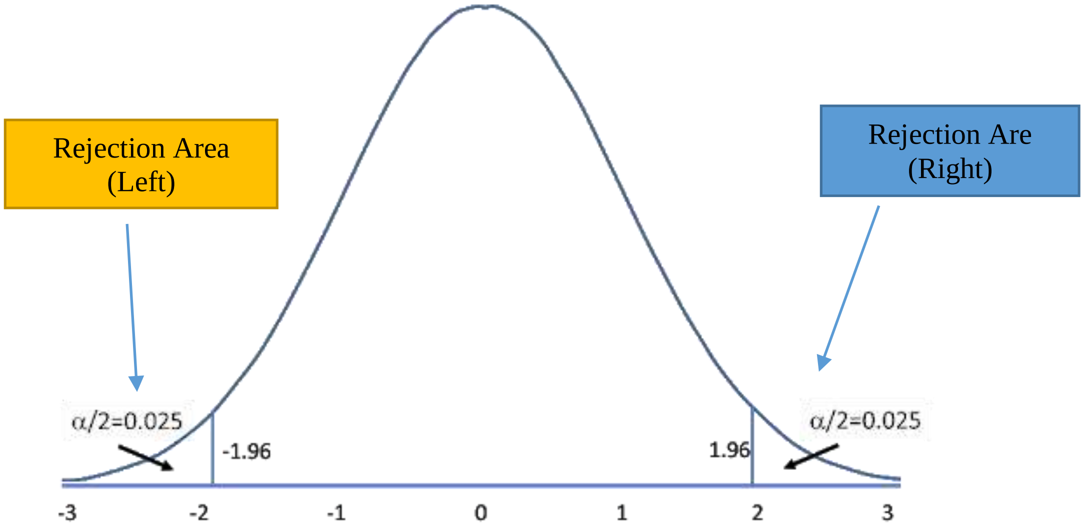
T-Statistics are:

Constant/intercept=–0.01/0.16 = -0.0625 > -1.96 (insignificant at 5%)
(left: outside the rejection area)

Educ=0.101/0.012= 8.417>1.96 (significant at 5%)
(right: falls in the rejection area)

Experience=0.033/0.006=5.5 >1.96 (significant at 5%)
(right: falls in the rejection area)

Expereience^2= -0.0005/0.0001=-5<-1.96 (significant at 5%)
(left: falls in the rejection area)



There are very little changes in the remaining coefficients, and their standard errors, when experience is omitted. The equation goodness-of-fit statistic decreases slightly from 0.34 to 0.20, as expected when variables are omitted. Based on these casual observations the consequences of omitting experience and experience squared are negligible. In other words, experience we add is perfectly correlated with education which is already in the model. Of course it is not the same, but it is additively, by constant. That will not change coefficient on education, because more or less we include redundant information.

e) Suppose you expect that returns to education is different for man and woman. Explain carefully how you would adjust the model to check this hypothesis (in case you need to expand the model, clearly define variable that you wish to include).

- If we expect that the returns to education is different for man and woman, we must incorporate two new components in the model.

1) The gender dummy, D=1 if person is male, and D=0 if person is female

2) **The interaction term D*Education**; intersection of gender and education. If there is gender-based discrimination in the labor market, it could take the form of a different intercept for men and women, or a different return to education for men and women

Our new model is:

$$\ln(Earn_i) = \alpha_1 + \alpha_2 * Educ + \alpha_3 * Exper - \alpha_4 * \text{Exper}^2 + \alpha_5 * \text{D} + \boldsymbol{\alpha_6} \ \mathbf{(D*Educ)} + \varepsilon$$

f) Suppose your friend suggests that the variable *Educ* is correlated to the regression error (that is the variable is endogenous), therefore OLS estimates are not unbiased. Explain her how would you address this issue and ensure that the variable(s) you use satisfy desired conditions

- By introducing the instrument for the education and estimate the model with two stage least square method
- TSLS: **FIRSRT stage**: regress education on instrument, and save predicted education

SECOND stage: regress Wage on predicted education and other regressors (in our case experience and experience squared)

Potential instruments for the education in wage equation could be

- distance to schools
- parental education
- spousal education
- number of siblings

Necessary conditions that valid instruments should satisfy:

1.Instrument relevance: corr(Zi,Xi) ≠0, instruments must significantly be correlated with the endogenous variable;

2. Instrument exogeneity: corr(Zi,ui) = 0, instrument should only affect the outcome via the endogeneous variable, in other words, instrument must be uncorrelated with error term (note that in outcome variable we have two parts: regressor X which is endogeneous in this case and e (error term))