ECO3121 INTRODUCTORY ECONOMETRICS
Problem Set #2 Solution

TA Group

November 4, 2024

Question 1

Download from the Blackboard site and load into Stata the Indonesia education and earning data `inpresdata.dta`. The main data we use is the 1995 intercensal survey of Indonesia (SUPAS), which is a nationally representative large cross section of men born between 1950 and 1972 from the 1995 intercensal survey of Indonesia. It is the main dataset used in the influential paper "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment" by Esther Duflo.

Since this will be a multiple linear regression model, we'll be estimating regressions of the form (model 1):

$$wage_{ij} = \alpha + \beta_1 education_i + \beta_2 nin_j + \mu_i$$

First, let's analyze the effect of **years of education** of each individual i and province-level (indexed by j) **number of inpres school per children** on **monthly wages**.

1. Run the regression and report the estimated coefficients, and their standard errors, and comment on the statistical significance of the regression estimates and R^2. (2 points)

Ans:

The regression result of model 1 is (1 points)

$$\widehat{wage}_{ij} = \underset{(2519.881)}{72117.21} + \underset{(193.9603)}{17826.42}\ education_i \underset{(714.7795)}{-6880.068}\ nin_j$$

$$R^2 = 0.1238$$

Because the p-value for $\hat{\beta}_1$ and $\hat{\beta}_2$ are smaller than 0.000, they are significantly different from 0 at the 1% significance level. (0.5 points, t-test or CI are also correct)

$R^2 = 0.1238$ implies 12.38% of the sample variation in wages is explained by individual education and province-level number of inpres schools per child. (0.5 points)

2. Let's assume the specification on question 1 is the true model, and you misspecified the model as model 2:

$$\widetilde{wage}_{ij} = \tilde{\alpha} + \tilde{\beta}_1 education_i$$

Comparing the estimates from these two models (specifications), What would be the bias and in which direction of the misspecified model? Be as precise as possible to explain the comparison. (1 point)

Ans:
The regression result of model 2 is (0.5 points)

$$\begin{array}{rcccl} \widetilde{wage}_{ij} = & 57405.06 & + & 17910.03 & education_i \\ & (2004.897) & & (193.9109) & \end{array}$$

$\hat{\beta}_1 < \tilde{\beta}_1$, so there is upward bias. (0.5 points)

3. Based on the estimates and comparison from these two models, what's the direction of $cov(education_i, nin_j)$ from theoretical judgment? Based on the dataset itself, what's the direction of $cov(education_i, nin_j)$? Are these two results consistent with each other? (3 points)

Ans: By the omitted variable bias formula: (1 point)

$$\text{Bias of } \tilde{\beta}_1 = \tilde{\beta}_1 - \hat{\beta}_1 = \hat{\beta}_2 \dot{\delta}_1$$

where $\dot{\delta}_1$ comes from the regression: $\dot{nin}_j = \dot{\delta}_0 + \dot{\delta}_1 education_i$
Since the bias of $\tilde{\beta}_1 > 0$, and $\hat{\beta}_2 < 0$, we think $cov(education_i, nin_j) < 0$.
Based on the data, you can choose one of the following methods to verify the direction of $cov(education_i, nin_j)$:

(a) Run a regression of *nin* on *education*, and the result is

$$\begin{array}{rccl} \dot{nin}_j = & 2.24942 & -0.017078 & education_i \\ & (0.0063333) & (0.0007015) & \end{array}$$

Since $\dot{\delta}_1 = -0.017078 < 0$, which implies $cov(education_i, nin_j) < 0$.

(b) Use the data to calculate the $corr(education_i, nin_j) = -0.0621$, which also implies $cov(education_i, nin_j) < 0$.

Thus, the direction of $cov(education_i, nin_j)$ is consistent with our theoretical judgment. (1 point)

4. A professor thinks about adding an additional variable "en71" to model 1, and asks that does the enrollment rate in 1971 of each province also change the wage outcome? Run the regressions recommended by the professor. Write down the new model (specification) as model 3. (1 point)

Ans: The specification of model 3 is (0.5 points)

$$wage_{ij} = \alpha + \beta_1 education_i + \beta_2 nin_j + \beta_3 en71_j + \mu_i$$

Based on the data, the regression result of model 3 is (0.5 points)

$$\begin{array}{rcccccccc} \dot{wage}_{ij} = & 68233.96 & + & 17704.64 & education_i & -7513.557 & nin_j + & 35557.16 & en71_j \\ & (2736.359) & & (197.781) & & (735.1522) & & (9684.733) & \end{array}$$

5. Formally write down the corresponding hypothesis testing (t-test) to verify the professor's question step by step (5 steps). And please make your conclusion to the professor's question. (3 points)

Ans:
Step 1: H_0: $\beta_3 = 0$. (0.5 points)
Step 2: H_1: $\beta_3 \neq 0$. (0.5 points)
Step 3: We choose the 1% significance level, and the critical value is about 2.56. (0.5 points)
Step 4: The t-statistic is (0.5 points)

$$t = \frac{\hat{\beta}_3}{se(\hat{\beta}_3)} = \frac{35557.16}{9684.733} = 3.67$$

Step 5: Since 3.67>2.56, we reject the H_0 in favor of H_1. (0.5 points)
Conclusion: the enrollment rate in 1971 of each province significantly changes the wage outcome at the 1% significance level. (0.5 points)

6. Comparing model 3 and model 2, the professor wants to test the joint significance of variables "nin" and "$en71$". What kind of test shall we use? Please offer the detailed procedure for the test step by step and conclude. (3 points)

Ans:
F test. The procedures are as follows:
Step 1: $H_0 : \beta_2 = 0, \beta_3 = 0$. (0.5 points)
Step 2: $H_1 : H_0$ is not true. (0.5 points)
Step 3: We choose the 1% significance level, and the critical value is $F_{1\%}(2, 60932) \approx 4.605$. (0.5 points)
Step 4: The F statistic is (0.5 points)

$$F = \frac{(R^2_{ur} - R^2_r)/q}{(1 - R^2_{ur})/(n - k - 1)} = \frac{(0.1239 - 0.1225)/2}{(1 - 0.1239)/(60936 - 3 - 1)} = 48.67872$$

Step 5: Since 48.67872>4.605, we reject the H_0 in favor of H_1. (0.5 points)
Conclusion: variables "nin_j" and "$en71$" are jointly significant at the 1% significance level (0.5 points).

Question 2: Linear and non-linear probability model

Now a professor is thinking about introducing dummy variables in the regression to better understand the contribution of education on wage.

First, we introduce a dummy variable to the dependent variable by dividing wages into high and low levels. Specifically we define $high\_wage = 1$ if $wage >= 192000$ and **is not a missing value**, and $high\_wage = 0$ if $wage < 192000$.

1. Please run a regression of Y: *high_wage* on X_i: years of education (*yedu*) and Z_j: number of inpres school per children (*nin*) using 1) linear probability model, and 2) Probit model, respectively.

   Summarize these estimates in a table with 3 columns (the format of the table can refer to the table on page 24 of the "week 8-2" slide, but the information in the last row of that table does not need to be displayed), and only interpret the coefficient of *yedu* in the linear probability model. (3 points)

   Ans:
   The results are in Table.1. (2 points for two models in Table.1)
   Interpretation: holding *nin* and other factors fixed, one more year of education increases the probability of being in the high-wage group by 4.9%. (1 point)

Table 1: Results for question 2.1

Dependent variable:	*high_wage*	
Regression model	**LPM**	**Probit**
yedu	0.049***	0.137***
	(0.000)	(0.001)
nin	-0.012***	-0.035***
	(0.002)	(0.005)
constant	0.060***	-1.234***
	(0.006)	(0.018)

Note: Standard errors in parentheses.

3. The professor uses the following Probit model,

$$P(Y_i = 1|X_i, Z_j) = \Phi(\beta_0 + \beta_1 X_i + \beta_2 Z_j)$$

   where i denotes an individual. Discuss 2-3 potential drawbacks of using a linear probability model instead of a non-linear model when Y is a binary variable. (2 points)

   Ans:

   (a) Predicted probabilities may be larger than one or smaller than zero. (0.5 points)

   (b) The estimated partial effects are constant throughout all the possible values of independent variables. (0.5 points)

   (c) The linear probability model must exhibit heteroskedasticity. (1 point)

4. Use the Probit model and the results from the table you offer in question (1), and calculate the **change** in probability when $Z = 3$ and X increases from 8 to 12. (Note: you can keep $\Phi(\cdot)$ in your answers without solving for it.) (1 point)

Ans:

$$\begin{aligned} change &= P(Y=1|X=12, Z=3) - P(Y=1|X=8, Z=3) \\ &= \Phi(-1.234 + 0.137 \times 12 - 0.035 \times 3) - \Phi(-1.234 + 0.137 \times 8 - 0.035 \times 3) \\ &= \Phi(0.305) - \Phi(-0.243) \end{aligned}$$

5. Use a graph of cumulative standard normal distribution to demonstrate such change in probability in question (3), and label the direction of the change. (Note: no need to spend time drawing a very neat figure. The information conveyed by this figure is more important.) (2 points)

Ans:
The graph for the probability density function (pdf) of the standard normal distribution is shown on the left of Figure.1. You can also draw a graph for the cumulative distribution function (cdf) as the right one of Figure.1. (2 points for either graph in Figure.1. It is also correct to replace z with x=8 and 12 and draw a graph like the right one of Figure.1.)

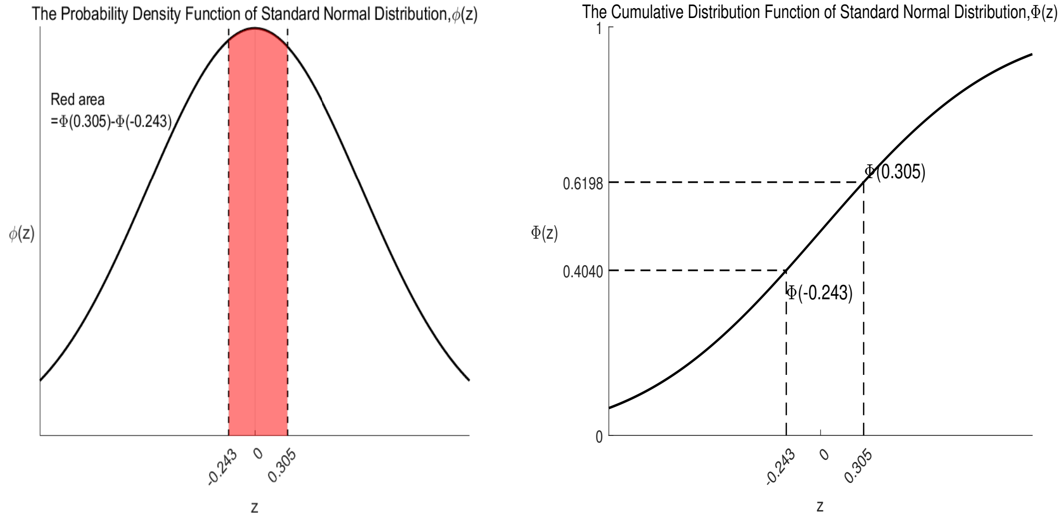

Figure 1: Example figure for question 2.4

Question 3: Dummy variable as the independent variable

The professor is also interested in studying whether high-education groups of people have higher wages than low-education groups of people. Therefore, the professor divides education level into high-level ($high\_level = 1$) if years of education are >= 9 and there is no missing value in years of education. And the variable of $high\_level = 0$ if years of education are < 9.

1. Run a regression of monthly wages on $high\_level$. Demonstrate the coefficient of $high\_level$ and interpret the estimate. (2 points)

Ans:
The regression result is (1 point)

$$\widehat{wage}_i = \underset{(1327.543)}{156631.9} + \underset{(1669.638)}{112764.3}\ high\_level_i$$

Interpretation: holding other factors fixed, the wage of high-education groups of people is 112764.3 more than that of low-education groups of people. (1 point)

2. The professor wants to study whether urban as the birthplace ($urban = 1$ in the data) can affect the slope of $high\_level$ on wage by introducing the interaction term.

1) Please write down the model (specification) and report your estimates. (Don't forget to include $urban$ as an independent variable in your model.) (1 point)

Ans:
The model specification is (0.5 points)

$$wage_i = \beta_0 + \beta_1 high\_level_i + \beta_2 urban_i + \beta_3 urban_i \times high\_level_i + \mu_i$$

The regression result is (0.5 points)

$$\widehat{wage}_i = \underset{(1446.661)}{152485.9} + \underset{(1914.016)}{99802.76}\ high\_level_i + \underset{(3553.394)}{25014.22}\ urban_i + \underset{(4131.893)}{23411.8}\ urban_i \times high\_level_i$$

2) Conduct your hypothesis testing using p-value, and conclude. (1 point)

Ans:
The p-value for $\hat{\beta}_3$ is 0.000, so we can reject $H_0 : \beta_3 = 0$ at the 1% significance level (0.5 points). That means people born in urban have significantly higher returns on education than those born in rural. (0.5 points)