Quantitative Methods for Economics

Tutorial 11

Katherine Eyal

TUTORIAL 11

18 October 2010

ECO3021S

Part A: Problems

1. An equation explaining chief executive officer salary is (standard errors in parentheses):

$$\begin{aligned} \log(salary) &= \underset{(0.30)}{4.59} + \underset{(0.032)}{0.257}\ \log(sales) + \underset{(0.004)}{0.011}\ roe + \underset{(0.089)}{0.158}\ finance \\ &\quad + \underset{(0.085)}{0.181}\ consprod - \underset{(0.099)}{0.283}\ utility \\ n &= 209, \quad R^2 = 0.357 \end{aligned}$$

where *salary* is the CEO's 1990 salary in thousands of Rands, *sales* is the firm's 1990 sales in millions of Rands, *roe* is the firm's 1988–1990 average return on equity, *finance*, *consprod* and *utility* are binary variables indicating the financial, consumer products, and utilities industries. The omitted industry is transportation.

   (a) Compute the approximate percentage difference in estimated salary between the utility and transportation industries, holding *sales* and *roe* fixed. Is the difference statistically significant at the 1% level?

   (b) Use equation (7.10) in Wooldridge (p. 233) to obtain the exact percentage difference in estimated salary between the utility and transportation industries and compare this with the answer obtained in part (a).

   (c) What is the approximate percentage difference in estimated salary between the consumer products and finance industries? Write an equation that would allow you to test whether the difference is statistically significant.

2. Suppose you collect data from a survey on wages, education, experience, and gender. In addition, you ask for information about marijuana usage. The original question is: "On how many separate occasions last month did you smoke marijuana?"

   (a) Write an equation that would allow you to estimate the effects of marijuana usage on wage, while controlling for other factors. You should be able to make statements such as, "Smoking marijuana five more times per month is estimated to change wage by x%."

(b) Write a model that would allow you to test whether drug usage has different effects on wages for men and women. How would you test that there are no differences in the effects of drug usage for men and women?

(c) Suppose you think it is better to measure marijuana usage by putting people into one of four categories: nonuser, light user (1 to 5 times per month), moderate user (6 to 10 times per month), and heavy user (more than 10 times per month). Now, write a model that allows you to estimate the effects of marijuana usage on wage.

(d) Using the model in part (c), explain in detail how to test the null hypothesis that marijuana usage has no effect on wage. Be very specific and include a careful listing of degrees of freedom.

(e) What are some potential problems with drawing causal inference using the survey data that you collected?

Part B: Computer Exercises

1. Use the data in CHARITY.DTA to answer this question. The variable *respond* is a dummy variable equal to one if a person responded with a contribution on the most recent mailing sent by a charitable organisation. The variable *resplast* is a dummy variable equal to one if the person responding to the previous mailing, $avggift$.is the average of past gits (in Dutch guilders), and *propresp* is the proportion of times the person has responded to past mailings.

   (a) Estimate a linear probability model relating *respond* to *resplast* and $avggift$. Interpret the coefficient on *resplast*.

   (b) Does the average value of past gifts seem to affect the probability of responding?

   (c) Add the variable *propresp* to the model and interpret its coefficient. (Be careful here: an increase of one in *propresp* is the largest possible change.)

   (d) What happened to the coefficient on *resplast* when *propresp* was added to the regression? Does this make sense?

   (e) Add *mailsyear*, the number of mailings per year, to the model. How big is its estimated effect? Why might this not be a good estimate of the causal effect of mailings on responding?

TUTORIAL 11 SOLUTIONS

18 October 2010

ECO3021S

Part A: Problems

1. An equation explaining chief executive officer salary is (standard errors in parentheses):

$$\begin{aligned} \log(salary) &= \underset{(0.30)}{4.59} + \underset{(0.032)}{0.257}\ \log(sales) + \underset{(0.004)}{0.011}\ roe + \underset{(0.089)}{0.158}\ finance \\ &\quad + \underset{(0.085)}{0.181}\ consprod - \underset{(0.099)}{0.283}\ utility \\ n &= 209, \quad R^2 = 0.357 \end{aligned}$$

where *salary* is the CEO's 1990 salary in thousands of Rands, *sales* is the firm's 1990 sales in millions of Rands, *roe* is the firm's 1988–1990 average return on equity, *finance*, *consprod* and *utility* are binary variables indicating the financial, consumer products, and utilities industries. The omitted industry is transportation.

   (a) Compute the approximate percentage difference in estimated salary between the utility and transportation industries, holding *sales* and *roe* fixed. Is the difference statistically significant at the 1% level?

   (b) Use equation (7.10) in Wooldridge (p. 233) to obtain the exact percentage difference in estimated salary between the utility and transportation industries and compare this with the answer obtained in part (a).

   (c) What is the approximate percentage difference in estimated salary between the consumer products and finance industries? Write an equation that would allow you to test whether the difference is statistically significant.

SOLUTION:

   (a) The approximate difference is just the coefficient on *utility* times 100, or -28.3%. The t statistic is $-.283/.099 \approx -2.86$, which is very statistically significant.

   (b) $100[\exp(-.283) - 1) \approx -24.7\%$, and so the estimate is somewhat smaller in magnitude.

(c) The proportionate difference is $.181 - .158 = .023$, or about 2.3%. One equation that can be estimated to obtain the standard error of this difference is

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + \beta_0 roe + \delta_1 consprod + \delta_2 utility + \delta_3 trans + u,$$

where *trans* is a dummy variable for the transportation industry. Now, the base group is $finance$, and so the coefficient directly measures the difference between the consumer products and finance industries, and we can use the t statistic on *consprod*.

2. Suppose you collect data from a survey on wages, education, experience, and gender. In addition, you ask for information about marijuana usage. The original question is: "On how many separate occasions last month did you smoke marijuana?"

(a) Write an equation that would allow you to estimate the effects of marijuana usage on wage, while controlling for other factors. You should be able to make statements such as, "Smoking marijuana five more times per month is estimated to change wage by x%."

(b) Write a model that would allow you to test whether drug usage has different effects on wages for men and women. How would you test that there are no differences in the effects of drug usage for men and women?

(c) Suppose you think it is better to measure marijuana usage by putting people into one of four categories: nonuser, light user (1 to 5 times per month), moderate user (6 to 10 times per month), and heavy user (more than 10 times per month). Now, write a model that allows you to estimate the effects of marijuana usage on wage.

(d) Using the model in part (c), explain in detail how to test the null hypothesis that marijuana usage has no effect on wage. Be very specific and include a careful listing of degrees of freedom.

(e) What are some potential problems with drawing causal inference using the survey data that you collected?

SOLUTION:

(a) We want to have a constant semi-elasticity model, so a standard wage equation with marijuana usage included would be

$$\log(wage) = \beta_0 + \beta_1 usage + \beta_2 educ + \beta_3 exper + \beta_4 exper^2 + \beta_5 female + u.$$

Then $100 \cdot \beta_1$ is the approximate percentage change in wage when marijuana usage increases by one time per month.

(b) We would add an interaction term in female and usage:

$$\begin{aligned}\log(wage) &= \beta_0 + \beta_1 usage + \beta_2 educ + \beta_3 exper + \beta_4 exper^2 + \beta_5 female \\ &\quad + \beta_6 female \cdot usage + u.\end{aligned}$$

The null hypothesis that the effect of marijuana usage does not differ by gender is $H_0 : \beta_6 = 0$.

(c) We take the base group to be nonuser. Then we need dummy variables for the other three groups: *lghtuser*, *moduser*, and *hvyuser*. Assuming no interactive effect with gender, the model would be

$$\begin{aligned}\log(wage) &= \beta_0 + \delta_1 lghtuser + \delta_2 moduser + \delta_3 hvyuser + \beta_2 educ + \beta_3 exper \\ &\quad + \beta_4 exper^2 + \beta_5 female + u\end{aligned}$$

(d) The null hypothesis is $H_0 : \delta_1 = 0, \delta_2 = 0, \delta_3 = 0$, for a total of $q = 3$ restrictions. If n is the sample size, the *df* in the unrestricted model – the denominator *df* in the F distribution – is $n - 8$. So we would obtain the critical value from the $F_{q,n-8}$ distribution.

(e) The error term could contain factors, such as family background (including parental history of drug abuse) that could directly affect wages and also be correlated with marijuana usage. We are interested in the effects of a person's drug usage on his or her wage, so we would like to hold other confounding factors fixed. We could try to collect data on relevant background information.

Part B: Computer Exercises

1. Use the data in CHARITY.DTA to answer this question. The variable *respond* is a dummy variable equal to one if a person responded with a contribution on the most recent mailing sent by a charitable organisation. The variable *resplast* is a dummy variable equal to one if the person responding to the previous mailing, $avggift$.is the average of past gits (in Dutch guilders), and *propresp* is the proportion of times the person has responded to past mailings.

   (a) Estimate a linear probability model relating *respond* to *resplast* and $avggift$. Interpret the coefficient on *resplast*.

   (b) Does the average value of past gifts seem to affect the probability of responding?

   (c) Add the variable *propresp* to the model and interpret its coefficient. (Be careful here: an increase of one in *propresp* is the largest possible change.)

   (d) What happened to the coefficient on *resplast* when *propresp* was added to the regression? Does this make sense?

(e) Add *mailsyear*, the number of mailings per year, to the model. How big is its estimated effect? Why might this not be a good estimate of the causal effect of mailings on responding?

SOLUTION:

(a) The estimated LPM is

$$\begin{aligned} \widehat{respond} &= \underset{(.009)}{.282} + \underset{(.015)}{.344}\ resplast + \underset{(.00009)}{.00015}\ avggift \\ n &= 4,268, \quad R^2 = .110 \end{aligned}$$

Holding the average gift fixed, the probability of a current response is estimated to be .344 higher if the person responded most recently.

(b) Once we control for responding most recently, the effect of $avggift$ is very small. Even if $avggift$ is 100 guilders more (the mean is about 18.2 with standard deviation 78.7), the probability of responding this period is only .015 higher. Plus, the t statistic has a two-sided p-value of about .09, so it is only marginally statistically significant.

(c) The coefficient on *propresp* is about .747 (standard error = .034). If propresp increases by .1 (for example, from .4 to .5), the probability of responding is about .075 higher.

(d) When *propresp* is added to the regression, the coefficient on *resplast* falls to about .095 (although it is still very statistically significant). This makes sense, because the relationship between responding currently and responding most recently should be weaker once the average response is controlled for. Certainly *resplast* and *propresp* are positively correlated.

(e) The coefficient on *mailsyear* is about .062($t = 6.18$). This is a reasonably large effect: each new mailing is estimated to increase the probability of responding by .062. Unfortunately, we do not know how the charitable organization determines the mailings sent. To the extent that it depends only on past gift giving, as controlled for by the average gift, the most recent response, and the response rate, the estimate could be a good (consistent) estimate of the causal effect. But if mailings are determined by other factors that are necessarily in the error term – such as income – then the estimate would be systematically biased. If, say, more mailings are sent to people with higher incomes, and higher income people are more likely to respond, then the regression that omits income produces an upward bias for the *mailsyear* coefficient.