Chapter 7: Multiple Regression Analysis with Qualitative Information: Binary (or Dummy) Variables Problems
Book Title: Introductory Econometrics
Printed By: Wanwiphang Manachotipong (wanwiphang@econ.tu.ac.th)

Chapter Review

Problems

1. Using the data in SLEEP75 (see also Problem 3 in Chapter 3), we obtain the estimated equation

$$\begin{aligned}\widehat{sleep} &= 3{,}840.83 - .163\,totwrk - 11.71\,educ - 8.70\,age \\ &\qquad (235.11)\ \ (.018)\qquad\quad (5.86)\qquad (11.21) \\ &\quad + .128\,age^2 + 87.75\,male \\ &\qquad\ (.134)\qquad\ (34.33) \\ n &= 706,\ R^2 = .123,\ \overline{R}^2 = .117.\end{aligned}$$

The variable *sleep* is total minutes per week spent sleeping at night, *totwrk* is total weekly minutes spent working, *educ* and *age* are measured in years, and *male* is a gender dummy.

   i. All other factors being equal, is there evidence that men sleep more than women? How strong is the evidence?

   ii. Is there a statistically significant tradeoff between working and sleeping? What is the estimated tradeoff?

   iii. What other regression do you need to run to test the null hypothesis that, holding other factors fixed, age has no effect on sleeping?

2. The following equations were estimated using the data in BWGHT:

$$\begin{aligned}\widehat{\log(bwght)} &= 4.66 - .0044\,cigs + .0093\log(faminc) + .016\,parity \\ &\qquad (.22)\ \ (.0009)\qquad (.0059)\qquad\quad (.006) \\ &\quad + .027\,male + .055\,white \\ &\qquad (.010)\qquad (.013) \\ n &= 1{,}388,\ R^2 = .0472\end{aligned}$$

and

$$\widehat{\log(bwght)} = 4.65 - .0052\,cigs + .0110\log(faminc) + .017\,parity$$
$$(.38)\ (.0010)\qquad (.0085)\qquad (.006)$$
$$+ .034\,male + .045\,white - .0030\,motheduc + .0032\,fatheduc$$
$$(.011)\qquad (.015)\qquad (.0030)\qquad (.0026)$$
$$n = 1{,}191,\ R^2 = .0493.$$

The variables are defined as in Example 4.9, but we have added a dummy variable for whether the child is male and a dummy variable indicating whether the child is classified as white.

i. In the first equation, interpret the coefficient on the variable *cigs*. In particular, what is the effect on birth weight from smoking 10 more cigarettes per day?

ii. How much more is a white child predicted to weigh than a nonwhite child, holding the other factors in the first equation fixed? Is the difference statistically significant?

iii. Comment on the estimated effect and statistical significance of *motheduc*.

iv. From the given information, why are you unable to compute the *F* statistic for joint significance of *motheduc* and *fatheduc*? What would you have to do to compute the *F* statistic?

3. Using the data in GPA2, the following equation was estimated:

$$\widehat{sat} = 1{,}028.10 + 19.30\,hsize - 2.19\,hsize^2 - 45.09\,female$$
$$(6.29)\ (3.83)\qquad (.53)\qquad (4.29)$$
$$- 169.81\,black + 62.31\,female{\cdot}black$$
$$(12.71)\qquad (18.15)$$
$$n = 4{,}137,\ R^2 = .0858.$$

The variable *sat* is the combined SAT score; *hsize* is size of the student's high school graduating class, in hundreds; *female* is a gender dummy variable; and *black* is a race dummy variable equal to one for blacks, and zero otherwise.

i. Is there strong evidence that $hsize^2$ should be included in the model? From this equation, what is the optimal high school size?

ii. Holding *hsize* fixed, what is the estimated difference in SAT score between nonblack females and nonblack males? How statistically significant is this estimated difference?

iii. What is the estimated difference in SAT score between nonblack males and black males? Test the null hypothesis that there is no difference between their scores, against the alternative that there is a difference.

iv. What is the estimated difference in SAT score between black females and nonblack females? What would you need to do to test whether the difference is statistically significant?

4. An equation explaining chief executive officer salary is

$$\widehat{\log(salary)} = 4.59 + .257\log(sales) + .011\,roe + .158\,finance$$
$$(.30)\ \ (.032)\qquad\qquad (.004)\qquad (.089)$$
$$+ .181\,consprod - .283\,utility$$
$$(.085)\qquad\qquad (.099)$$
$$n = 209,\ R^2 = .357.$$

The data used are in CEOSAL1, where *finance*, *consprod*, and *utility* are binary variables indicating the financial, consumer products, and utilities industries. The omitted industry is transportation.

i. 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?

ii. Use equation (7.10) to obtain the exact percentage difference in estimated salary between the utility and transportation industries and compare this with the answer obtained in part (i).

iii. 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.

5. In Example 7.2, let *noPC* be a dummy variable equal to one if the student does not own a PC, and zero otherwise.

i. If *noPC* is used in place of *PC* in equation (7.6), what happens to the intercept in the estimated equation? What will be the coefficient on *noPC*? (*Hint:* Write $PC = 1 - noPC$ and plug this into the equation $\widehat{colGPA} = \hat{\beta}_0 + \hat{\delta}_0 PC + \hat{\beta}_1 hsGPA + \hat{\beta}_2 ACT$.)

ii. What will happen to the *R*-squared if *noPC* is used in place of *PC*?

iii. Should *PC* and *noPC* both be included as independent variables in the model? Explain.

6. To test the effectiveness of a job training program on the subsequent wages of workers, we specify the model

$$\log(wage) = \beta_0 + \beta_1 train + \beta_2 educ + \beta_3 exper + u,$$

where *train* is a binary variable equal to unity if a worker participated in the program. Think of the error term *u* as containing unobserved worker ability. If less able workers have a greater chance of being selected for the program, and you use an OLS analysis, what can you say about the likely bias in the OLS estimator of β_1? (*Hint:* Refer back to Chapter 3.)

7. In the example in equation (7.29), suppose that we define *outlf* to be one if the woman is out of the labor force, and zero otherwise.

   i. If we regress *outlf* on all of the independent variables in equation (7.29), what will happen to the intercept and slope estimates? (*Hint:* $inlf = 1 - outlf$. Plug this into the population equation $inlf = \beta_0 + \beta_1 nwifeinc + \beta_2 educ +$… and rearrange.)

   ii. What will happen to the standard errors on the intercept and slope estimates?

   iii. What will happen to the *R*-squared?

8. 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?”

   i. 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*%.”

   ii. 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?

   iii. 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.

iv. Using the model in part (iii), 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.

v. What are some potential problems with drawing causal inference using the survey data that you collected?

9. Let *d* be a dummy (binary) variable and let *z* be a quantitative variable. Consider the model

$$y = \beta_0 + \delta_0 d + \beta_1 z + \delta_1 d \cdot z + u;$$

this is a general version of a model with an interaction between a dummy variable and a quantitative variable. [An example is in equation (7.17).]

i. Since it changes nothing important, set the error to zero, $u = 0$. Then, when $d = 0$ we can write the relationship between *y* and *z* as the function $f_0(z) = \beta_0 + \beta_1 z$. Write the same relationship when $d = 1$, where you should use $f_1(z)$ on the left-hand side to denote the linear function of *z*.

ii. Assuming that $\delta_1 \neq 0$ (which means the two lines are not parallel), show that the value of z^* such that $f_0(z^*) = f1(z^*)$ is $z^* = -\delta_0/\delta_1$. This is the point at which the two lines intersect [as in Figure 7.2 (b)]. Argue that z^* is positive if and only if δ_0 and δ_1 have opposite signs.

iii. Using the data in TWOYEAR, the following equation can be estimated:

$$\widehat{\log(wage)} = 2.289 - .357\, female + .50\, totcoll + .030\, female \cdot totcoll$$
$$(0.011)\ \ (.015)\qquad (.003)\qquad (.005)$$
$$n = 6{,}763,\ R^2 = .202,$$

where all coefficients and standard errors have been rounded to three decimal places. Using this equation, find the value of *totcoll* such that the predicted values of $\log(wage)$ are the same for men and women.

iv. Based on the equation in part (iii), can women realistically get enough years of college so that their earnings catch up to those of men? Explain.

10. For a child *i* living in a particular school district, let $voucher_i$ be a dummy variable equal to one if a child is selected to participate in a school voucher

program, and let $score_i$ be that child's score on a subsequent standardized exam. Suppose that the participation variable, $voucher_i$, is completely randomized in the sense that it is independent of both observed and unobserved factors that can affect the test score.

i. If you run a simple regression $score_i$ on $voucher_i$ using a random sample of size *n*, does the OLS estimator provide an unbiased estimator of the effect of the voucher program?

ii. Suppose you can collect additional background information, such as family income, family structure (e.g., whether the child lives with both parents), and parents' education levels. Do you need to control for these factors to obtain an unbiased estimator of the effects of the voucher program? Explain.

iii. Why should you include the family background variables in the regression? Is there a situation in which you would not include the background variables?

11. The following equations were estimated using the data in ECONMATH, with standard errors reported under coefficients. The average class score, measured as a percentage, is about 72.2; exactly 50% of the students are male; and the average of *colgpa* (grade point average at the start of the term) is about 2.81.

$$\widehat{score} = 32.31 + 14.32\, colgpa$$
$$(2.00) \quad (0.70)$$
$$n = 856, R^2 = .329, \overline{R}^2 = .328.$$

$$\widehat{score} = 29.66 + 3.83\, male + 14.57\, colgpa$$
$$(2.04) \ (0.74) \qquad (0.69)$$
$$n = 856, R^2 = .349, \overline{R}^2 = .348.$$

$$\widehat{score} = 30.36 + 2.47\, male + 14.33\, colgpa + 0.479\, male \cdot colgpa$$
$$(2.86) \ (3.96) \qquad (0.98) \qquad (1.383)$$
$$n = 856, R^2 = .349, \overline{R}^2 = .347.$$

$$\widehat{score} = 30.36 + 3.82\, male + 14.33\, colgpa + 0.479\, male \cdot (colgpa - 2.81)$$
$$(2.86) \ (0.74) \qquad (0.98) \qquad (1.383)$$
$$n = 856, R^2 = .349, \overline{R}^2 = .347.$$

i. Interpret the coefficient on *male* in the second equation and construct a 95% confidence interval for β_{male}. Does the confidence interval exclude zero?

ii. In the second equation, how come the estimate on *male* is so imprecise? Should we now conclude that there are no gender differences in *score* after controlling for *colgpa*? [*Hint*: You might want to compute an *F* statistic for the null hypothesis that there is no gender difference in the model with the interaction.]

iii. Compared with the third equation, how come the coefficient on *male* in the last equation is so much closer to that in the second equation and just as precisely estimated?