Exercises EE325: 1/2015 (Aj. Wasin)

Question 1

$$\widehat{colgpa} = 1.392 - .0135\ hsperc + .00148\ sat$$
$$n = 4{,}137,\ R^2 = .273,$$

where *colgpa* is measured on a four-point scale, *hsperc* is the percentile in the high school graduating class (defined so that, for example, *hsperc* = 5 means the *top* 5% of the class), and *sat* is the combined math and verbal scores on the student achievement test.

(i) Why does it make sense for the coefficient on *hsperc* to be negative?

(ii) What is the predicted college GPA when *hsperc* = 20 and *sat* = 1,050?

(iii) Suppose that two high school graduates, A and B, graduated in the same percentile from high school, but Student A's SAT score was 140 points higher (about one standard deviation in the sample). What is the predicted difference in college GPA for these two students? Is the difference large?

(iv) Holding *hsperc* fixed, what difference in SAT scores leads to a predicted *colgpa* difference of .50, or one-half of a grade point? Comment on your answer.

Solution

3.1 (i) *hsperc* is defined so that the smaller it is, the lower the student's standing in high school. Everything else equal, the worse the student's standing in high school, the lower his/her expected college GPA.

(ii) Just plug these values into the equation

$$\widehat{colgpa} = 1.392 - .0135(20) + .00148(1050) = 2.676.$$

(iii) The difference between A and B is simply 140 times the coefficient on *sat*, because *hsperc* is the same for both students. So A is predicted to have a score .00148(140) ≈ .207 higher.

(iv) With *hsperc* fixed, $\Delta\widehat{colgpa} = .00148\Delta sat$. Now, we want to find Δsat such that $\Delta\widehat{colgpa} = .5$, so $.5 = .00148(\Delta sat)$ or $\Delta sat = .5/(.00148) \approx 338$. Perhaps not surprisingly, a large ceteris paribus difference in SAT score – almost two and one-half standard deviations – is needed to obtain a predicted difference in college GPA or a half a point.

Question 2

3.3 The following model is a simplified version of the multiple regression model used by Biddle and Hamermesh (1990) to study the tradeoff between time spent sleeping and working and to look at other factors affecting sleep:

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + u,$$

where *sleep* and *totwrk* (total work) are measured in minutes per week and *educ* and *age* are measured in years. (See also Computer Exercise C2.3.)

(i) If adults trade off sleep for work, what is the sign of β_1?

(ii) What signs do you think β_2 and β_3 will have?

(iii) Using the data in SLEEP75.RAW, the estimated equation is

$$\widehat{sleep} = 3{,}638.25 - .148\ totwrk - 11.13\ educ + 2.20\ age$$
$$n = 706,\ R^2 = .113.$$

If someone works five more hours per week, by how many minutes is *sleep* predicted to fall? Is this a large tradeoff?

(iv) Discuss the sign and magnitude of the estimated coefficient on *educ*.

(v) Would you say *totwrk*, *educ*, and *age* explain much of the variation in *sleep*? What other factors might affect the time spent sleeping? Are these likely to be correlated with *totwrk*?

Solution

3.3 (i) If adults trade off sleep for work, more work implies less sleep (other things equal), so $\beta_1 < 0$.

(ii) The signs of β_2 and β_3 are not obvious, at least to me. One could argue that more educated people like to get more out of life, and so, other things equal, they sleep less ($\beta_2 < 0$). The relationship between sleeping and age is more complicated than this model suggests, and economists are not in the best position to judge such things.

(iii) Since *totwrk* is in minutes, we must convert five hours into minutes: $\Delta totwrk = 5(60) = 300$. Then *sleep* is predicted to fall by .148(300) = 44.4 minutes. For a week, 45 minutes less sleep is not an overwhelming change.

(iv) More education implies less predicted time sleeping, but the effect is quite small. If we assume the difference between college and high school is four years, the college graduate sleeps about 45 minutes less per week, other things equal.

(v) Not surprisingly, the three explanatory variables explain only about 11.3% of the variation in *sleep*. One important factor in the error term is general health. Another is marital status and whether the person has children. Health (however we measure that), marital status, and number and ages of children would generally be correlated with *totwrk*. (For example, less healthy people would tend to work less.)

Question 3

Consider an equation to explain salaries of CEOs in terms of annual firm sales, return on equity (*roe*, in percentage form), and return on the firm's stock (*ros*, in percentage form):

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + \beta_2 roe + \beta_3 ros + u.$$

(i) In terms of the model parameters, state the null hypothesis that, after controlling for *sales* and *roe*, *ros* has no effect on CEO salary. State the alternative that better stock market performance increases a CEO's salary.

(ii) Using the data in CEOSAL1.RAW, the following equation was obtained by OLS:

$$\widehat{\log(salary)} = 4.32 + .280 \log(sales) + .0174\, roe + .00024\, ros$$
$$(.32)\ \ (.035)\qquad\qquad (.0041)\qquad (.00054)$$
$$n = 209,\ R^2 = .283.$$

By what percentage is *salary* predicted to increase if *ros* increases by 50 points? Does *ros* have a practically large effect on *salary*?

(iii) Test the null hypothesis that *ros* has no effect on *salary* against the alternative that *ros* has a positive effect. Carry out the test at the 10% significance level.

(iv) Would you include *ros* in a final model explaining CEO compensation in terms of firm performance? Explain.

Solution

4.2 (i) $H_0: \beta_3 = 0$. $H_1: \beta_3 > 0$.

(ii) The proportionate effect on $\widehat{salary}$ is .00024(50) = .012. To obtain the percentage effect, we multiply this by 100: 1.2%. Therefore, a 50-point ceteris paribus increase in *ros* is predicted to increase salary by only 1.2%. Practically speaking, this is a very small effect for such a large change in *ros*.

(iii) The 10% critical value for a one-tailed test, using $df = \infty$, is obtained from Table G.2 as 1.282. The *t* statistic on *ros* is .00024/.00054 ≈ .44, which is well below the critical value. Therefore, we fail to reject H_0 at the 10% significance level.

(iv) Based on this sample, the estimated *ros* coefficient appears to be different from zero only because of sampling variation. On the other hand, including *ros* may not be causing any harm; it depends on how correlated it is with the other independent variables (although these are very significant even with *ros* in the equation).

Question 4

Consider the multiple regression model with three independent variables, under the classical linear model assumptions MLR.1 through MLR.6:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u.$$

You would like to test the null hypothesis H_0: $\beta_1 - 3\beta_2 = 1$.

(i) Let $\hat{\beta}_1$ and $\hat{\beta}_2$ denote the OLS estimators of β_1 and β_2. Find $\text{Var}(\hat{\beta}_1 - 3\hat{\beta}_2)$ in terms of the variances of $\hat{\beta}_1$ and $\hat{\beta}_2$ and the covariance between them. What is the standard error of $\hat{\beta}_1 - 3\hat{\beta}_2$?

(ii) Write the t statistic for testing H_0: $\beta_1 - 3\beta_2 = 1$.

(iii) Define $\theta_1 = \beta_1 - 3\beta_2$ and $\hat{\theta}_1 = \hat{\beta}_1 - 3\hat{\beta}_2$. Write a regression equation involving β_0, θ_1, β_2, and β_3 that allows you to directly obtain $\hat{\theta}_1$ and its standard error.

Solution

4.8 (i) We use Property VAR.3 from Appendix B: $\text{Var}(\hat{\beta}_1 - 3\hat{\beta}_2) = \text{Var}(\hat{\beta}_1) + 9\,\text{Var}(\hat{\beta}_2) - 6\,\text{Cov}(\hat{\beta}_1, \hat{\beta}_2)$.

$$\text{se}\left(\hat{\beta}_1 - 3\hat{\beta}_2\right) = \left\{\left[\text{se}\left(\hat{\beta}_1\right)\right]^2 + 9\left[\text{se}\left(\hat{\beta}_2\right)\right]^2 - 6\text{Cov}\left(\hat{\beta}_1, \hat{\beta}_2\right)\right\}^{1/2}$$

(ii) $t = (\hat{\beta}_1 - 3\hat{\beta}_2 - 1)/\text{se}(\hat{\beta}_1 - 3\hat{\beta}_2)$, so we need the standard error of $\hat{\beta}_1 - 3\hat{\beta}_2$.

(iii) Because $\theta_1 = \beta_1 - 3\beta_2$, we can write $\beta_1 = \theta_1 + 3\beta_2$. Plugging this into the population model gives

$$\begin{aligned} y &= \beta_0 + (\theta_1 + 3\beta_2)x_1 + \beta_2 x_2 + \beta_3 x_3 + u \\ &= \beta_0 + \theta_1 x_1 + \beta_2(3x_1 + x_2) + \beta_3 x_3 + u. \end{aligned}$$

This last equation is what we would estimate by regressing y on x_1, $3x_1 + x_2$, and x_3. The coefficient and standard error on x_1 are what we want.

Question 5

In Problem 3.3, we estimated the equation

$$\widehat{sleep} = \underset{}{3{,}638.25} - \underset{(112.28)}{.148}\ totwrk - \underset{(5.88)}{11.13}\ educ + \underset{(1.45)}{2.20}\ age$$
$$\underset{(.017)}{}$$
$$n = 706,\ R^2 = .113,$$

where we now report standard errors along with the estimates.

(i) Is either *educ* or *age* individually significant at the 5% level against a two-sided alternative? Show your work.

(ii) Dropping *educ* and *age* from the equation gives

$$\widehat{sleep} = \underset{}{3{,}586.38} - \underset{(38.91)}{.151}\ totwrk$$
$$\underset{(.017)}{}$$
$$n = 706,\ R^2 = .103.$$

Are *educ* and *age* jointly significant in the original equation at the 5% level? Justify your answer.

(iii) Does including *educ* and *age* in the model greatly affect the estimated tradeoff between sleeping and working?

(iv) Suppose that the sleep equation contains heteroskedasticity. What does this mean about the tests computed in parts (i) and (ii)?

Solution:

4.9 (i) With $df = 706 - 4 = 702$, we use the standard normal critical value ($df = \infty$ in Table G.2), which is 1.96 for a two-tailed test at the 5% level. Now $t_{educ} = -11.13/5.88 \approx -1.89$, so $|t_{educ}| = 1.89 < 1.96$, and we fail to reject H_0: $\beta_{educ} = 0$ at the 5% level. Also, $t_{age} \approx 1.52$, so *age* is also statistically insignificant at the 5% level.

(ii) We need to compute the *R*-squared form of the *F* statistic for joint significance. But $F = [(.113 - .103)/(1 - .113)](702/2) \approx 3.96$. The 5% critical value in the $F_{2,702}$ distribution can be obtained from Table G.3b with denominator $df = \infty$: $cv = 3.00$. Therefore, *educ* and *age* are jointly significant at the 5% level ($3.96 > 3.00$). In fact, the *p*-value is about .019, and so *educ* and *age* are jointly significant at the 2% level.

(iii) Not really. These variables are jointly significant, but including them only changes the coefficient on *totwrk* from –.151 to –.148.

(iv) The standard *t* and *F* statistics that we used assume homoskedasticity, in addition to the other CLM assumptions. If there is heteroskedasticity in the equation, the tests are no longer valid.

Question 6

4.7 In Example 4.7, we used data on nonunionized manufacturing firms to estimate the relationship between the scrap rate and other firm characteristics. We now look at this example more closely and use all available firms.

(i) The population model estimated in Example 4.7 can be written as

$$\log(scrap) = \beta_0 + \beta_1 hrsemp + \beta_2 \log(sales) + \beta_3 \log(employ) + u.$$

Using the 43 observations available for 1987, the estimated equation is

$$\widehat{\log(scrap)} = \underset{(4.57)}{11.74} - \underset{(.019)}{.042}\, hrsemp - \underset{(.370)}{.951}\, \log(sales) + \underset{(.360)}{.992}\, \log(employ)$$
$$n = 43,\ R^2 = .310.$$

Compare this equation to that estimated using only the 29 nonunionized firms in the sample.

(ii) Show that the population model can also be written as

$$\log(scrap) = \beta_0 + \beta_1 hrsemp + \beta_2 \log(sales/employ) + \theta_3 \log(employ) + u,$$

where $\theta_3 = \beta_2 + \beta_3$. [*Hint:* Recall that $\log(x_2/x_3) = \log(x_2) - \log(x_3)$.] Interpret the hypothesis H_0: $\theta_3 = 0$.

(iii) When the equation from part (ii) is estimated, we obtain

$$\widehat{\log(scrap)} = \underset{(4.57)}{11.74} - \underset{(.019)}{.042}\, hrsemp - \underset{(.370)}{.951}\, \log(sales/employ) + \underset{(.205)}{.041}\, \log(employ)$$
$$n = 43,\ R^2 = .310.$$

Controlling for worker training and for the sales-to-employee ratio, do bigger firms have larger statistically significant scrap rates?

(iv) Test the hypothesis that a 1% increase in *sales/employ* is associated with a 1% drop in the scrap rate.

Solution

4.7 (i) While the standard error on *hrsemp* has not changed, the magnitude of the coefficient has increased by half. The *t* statistic on *hrsemp* has gone from about –1.26 to –2.21, so now the

coefficient is statistically less than zero at the 5% level. (From Table G.2 the 5% critical value with 40 *df* is –1.684. The 1% critical value is –2.423, so the *p*-value is between .01 and .05.)

(ii) If we add and subtract $\beta_2 \log(employ)$ from the right-hand-side and collect terms, we have

$$\begin{aligned} \log(scrap) &= \beta_0 + \beta_1 hrsemp + [\beta_2 \log(\text{sales}) - \beta_2 \log(employ)] \\ &\quad + [\beta_2 \log(employ) + \beta_3 \log(employ)] + u \\ &= \beta_0 + \beta_1 hrsemp + \beta_2 \log(sales/employ) \\ &\quad + (\beta_2 + \beta_3)\log(employ) + u, \end{aligned}$$

where the second equality follows from the fact that $\log(sales/employ) = \log(sales) - \log(employ)$. Defining $\theta_3 \equiv \beta_2 + \beta_3$ gives the result.

(iii) No. We are interested in the coefficient on log(*employ*), which has a *t* statistic of .2, which is very small. Therefore, we conclude that the size of the firm, as measured by employees, does not matter, once we control for training *and* sales per employee (in a logarithmic functional form).

(iv) The null hypothesis in the model from part (ii) is $H_0: \beta_2 = -1$. The *t* statistic is $[-.951 - (-1)]/.37 = (1 - .951)/.37 \approx .132$; this is very small, and we fail to reject whether we specify a one- or two-sided alternative.

Question 7

In Section 4.5, we used as an example testing the rationality of assessments of housing prices. There, we used a log-log model in *price* and *assess* [see equation (4.47)]. Here, we use a level-level formulation.

(i) In the simple regression model

$$price = \beta_0 + \beta_1 assess + u,$$

the assessment is rational if $\beta_1 = 1$ and $\beta_0 = 0$. The estimated equation is

$$\begin{aligned} \widehat{price} &= -14.47 + .976\ assess \\ &\quad (16.27)\quad (.049) \\ n &= 88,\ \text{SSR} = 165{,}644.51,\ R^2 = .820. \end{aligned}$$

First, test the hypothesis that H_0: $\beta_0 = 0$ against the two-sided alternative. Then, test H_0: $\beta_1 = 1$ against the two-sided alternative. What do you conclude?

(ii) To test the joint hypothesis that $\beta_0 = 0$ and $\beta_1 = 1$, we need the SSR in the restricted model. This amounts to computing $\sum_{i=1}^{n}(price_i - assess_i)^2$, where $n = 88$, since the residuals in the restricted model are just $price_i - assess_i$. (No estimation is needed for the restricted model because both parameters are specified under H_0.) This turns out to yield SSR = 209,448.99. Carry out the F test for the joint hypothesis.

(iii) Now, test H_0: $\beta_2 = 0$, $\beta_3 = 0$, and $\beta_4 = 0$ in the model

$$price = \beta_0 + \beta_1 assess + \beta_2 lotsize + \beta_3 sqrft + \beta_4 bdrms + u.$$

The R-squared from estimating this model using the same 88 houses is .829.

(iv) If the variance of *price* changes with *assess*, *lotsize*, *sqrft*, or *bdrms*, what can you say about the F test from part (iii)?

Solution:

4.6 (i) With $df = n - 2 = 86$, we obtain the 5% critical value from Table G.2 with $df = 90$. Because each test is two-tailed, the critical value is 1.987. The t statistic for H_0: $\beta_0 = 0$ is about -.89, which is much less than 1.987 in absolute value. Therefore, we fail to reject $\beta_0 = 0$. The t statistic for H_0: $\beta_1 = 1$ is $(.976 - 1)/.049 \approx -.49$, which is even less significant. (Remember, we reject H_0 in favor of H_1 in this case only if $|t| > 1.987$.)

(ii) We use the SSR form of the F statistic. We are testing $q = 2$ restrictions and the df in the unrestricted model is 86. We are given $SSR_r = 209,448.99$ and $SSR_{ur} = 165,644.51$. Therefore,

$$F = \frac{(209,448.99 - 165,644.51)}{165,644.51} \cdot \left(\frac{86}{2}\right) \approx 11.37,$$

which is a strong rejection of H_0; from Table G.3c, the 1% critical value with 2 and 90 df is 4.85.

(iii) We use the R-squared form of the F statistic. We are testing $q = 3$ restrictions and there are $88 - 5 = 83$ df in the unrestricted model. The F statistic is $[(.829 - .820)/(1 - .829)](83/3) \approx 1.46$. The 10% critical value (again using 90 denominator df in Table G.3a) is 2.15, so we fail to reject H_0 at even the 10% level. In fact, the p-value is about .23.

(iv) If heteroskedasticity were present, Assumption MLR.5 would be violated, and the F statistic would not have an F distribution under the null hypothesis. Therefore, comparing the F statistic against the usual critical values, or obtaining the p-value from the F distribution, would not be especially meaningful.

Question 8

6.4 The following model allows the return to education to depend upon the total amount of both parents' education, called *pareduc*:

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 educ \cdot pareduc + \beta_3 exper + \beta_4 tenure + u.$$

(i) Show that, in decimal form, the return to another year of education in this model is

$$\Delta\log(wage)/\Delta educ = \beta_1 + \beta_2 pareduc.$$

What sign do you expect for β_2? Why?

(ii) Using the data in WAGE2.RAW, the estimated equation is

$$\widehat{\log(wage)} = 5.65 + .047\ educ + .00078\ educ \cdot pareduc +$$
$$(.13)\ \ (.010)\qquad\quad (.00021)$$
$$.019\ exper + .010\ tenure$$
$$(.004)\qquad\ (.003)$$
$$n = 722,\ R^2 = .169.$$

(Only 722 observations contain full information on parents' education.) Interpret the coefficient on the interaction term. It might help to choose two specific values for *pareduc*—for example, *pareduc* = 32 if both parents have a college education, or *pareduc* = 24 if both parents have a high school education—and to compare the estimated return to *educ*.

(iii) When *pareduc* is added as a separate variable to the equation, we get:

$$\widehat{\log(wage)} = 4.94 + .097\ educ + .033\ pareduc - .0016\ educ \cdot pareduc$$
$$(.38)\ \ (.027)\qquad\ (.017)\qquad\quad (.0012)$$
$$+ .020\ exper + .010\ tenure$$
$$(.004)\qquad\ (.003)$$
$$n = 722,\ R^2 = .174.$$

Does the estimated return to education now depend positively on parent education? Test the null hypothesis that the return to education does not depend on parent education.

Solution:

6.4 (i) Holding all other factors fixed, we have

$$\Delta \log(wage) = \beta_1 \Delta educ + \beta_2 \Delta educ \cdot pareduc = (\beta_1 + \beta_2 pareduc)\Delta educ.$$

Dividing both sides by $\Delta educ$ gives the result. The sign of β_2 is not obvious; although, $\beta_2 > 0$ if we think that the more highly educated the child's parents, the more it gets out of another year of education.

(ii) We use the values *pareduc* = 32 and *pareduc* = 24 to interpret the coefficient on *educ*·*pareduc*. The difference in the estimated return to education is .00078(32 – 24) = .0062, or about .62 percentage points.

(iii) When we add *pareduc* by itself, the coefficient on the interaction term is negative. The *t* statistic on *educ*·*pareduc* is about –1.33, which is not significant at the 10% level against a two-sided alternative. Note that the coefficient on *pareduc* is significant at the 5% level against a two-sided alternative. This provides a good example of how omitting a level effect (*pareduc* in this case) can lead to biased estimation of the interaction effect.

Question 9

6.7 The following three equations were estimated using the 1,534 observations in 401K.RAW:

$$\widehat{prate} = 80.29 + 5.44\ mrate + .269\ age - .00013\ totemp$$
$$(.78)\quad (.52)\qquad (.045)\qquad (.00004)$$
$$R^2 = .100,\ \bar{R}^2 = .098.$$

$$\widehat{prate} = 97.32 + 5.02\ mrate + .314\ age - 2.66\ \log(totemp)$$
$$(1.95)\quad (0.51)\qquad (.044)\qquad (.28)$$
$$R^2 = .144,\ \bar{R}^2 = .142.$$

$$\widehat{prate} = 80.62 + 5.34\ mrate + .290\ age - .00043\ totemp$$
$$(.78)\quad (.52)\qquad (.045)\qquad (.00009)$$
$$+ .0000000039\ totemp^2$$
$$(.0000000010)$$
$$R^2 = .108,\ \bar{R}^2 = .106.$$

Which of these three models do you prefer? Why?

Solution

6.7 The second equation is clearly preferred, as its adjusted *R*-squared is notably larger than that in the other two equations. The second equation contains the same number of estimated parameters as the first and one fewer than the third. The second equation is also easier to interpret than the third.

Question 10

Using the data in SLEEP75.RAW (see also Problem 3.3), we obtain the estimated equation

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

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?

Solution:

7.1 (i) The coefficient on *male* is 87.75, so a man is estimated to sleep almost one and one-half hours more per week than a comparable woman. Further, $t_{male} = 87.75/34.33 \approx 2.56$, which is close to the 1% critical value against a two-sided alternative (about 2.58). Thus, the evidence for a gender differential is fairly strong.

(ii) The *t* statistic on *totwrk* is $-.163/.018 \approx -9.06$, which is very statistically significant. The coefficient implies that one more hour of work (60 minutes) is associated with $.163(60) \approx 9.8$ minutes less sleep.

(iii) To obtain R_r^2, the *R*-squared from the restricted regression, we need to estimate the model without *age* and age^2. When *age* and age^2 are both in the model, *age* has no effect only if the parameters on both terms are zero.

Question 11

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

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

and

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

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?

Solution

7.2 (i) If $\Delta cigs = 10$, then $\Delta\widehat{\log(bwght)} = -.0044(10) = -.044$, which means about a 4.4% lower birth weight.

(ii) A white child is estimated to weigh about 5.5% more, other factors in the first equation fixed. Further, $t_{white} \approx 4.23$, which is well above any commonly used critical value. Thus, the difference between white and nonwhite babies is also statistically significant.

(iii) If the mother has one more year of education, the child's birth weight is estimated to be .3% higher. This is not a huge effect, and the t statistic is only one, so it is not statistically significant.

(iv) The two regressions use different sets of observations. The second regression uses fewer observations because *motheduc* or *fatheduc* are missing for some observations. We would have to reestimate the first equation (and obtain the R-squared) using the same observations used to estimate the second equation.

Question 12

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

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

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?

Solution

7.3 (i) The *t* statistic on $hsize^2$ is over four in absolute value, so there is very strong evidence that it belongs in the equation. We obtain this by finding the turnaround point; this is the value of *hsize* that maximizes $\hat{sat}$ (other things fixed): $19.3/(2\cdot 2.19) \approx 4.41$. Because *hsize* is measured in hundreds, the optimal size of graduating class is about 441.

(ii) This is given by the coefficient on *female* (since *black* = 0); nonblack females have SAT scores about 45 points lower than nonblack males. The *t* statistic is about –10.51, so the difference is very statistically significant. (The very large sample size certainly contributes to the statistical significance.)

(iii) Because *female* = 0, the coefficient on *black* implies that a black male has an estimated SAT score almost 170 points less than a comparable nonblack male. The *t* statistic is over 13 in absolute value, so we easily reject the hypothesis that there is no ceteris paribus difference.

(iv) We plug in *black* = 1, *female* = 1 for black females and *black* = 0 and *female* = 1 for nonblack females. The difference is therefore –169.81 + 62.31 = –107.50. Because the estimate depends on two coefficients, we cannot construct a *t* statistic from the information given. The easiest approach is to define dummy variables for three of the four race/gender categories and choose nonblack females as the base group. We can then obtain the *t* statistic we want as the coefficient on the black female dummy variable.

Question13

Which of the following are consequences of heteroskedasticity?

(i) The OLS estimators, $\hat{\beta}_j$, are inconsistent.

(ii) The usual F statistic no longer has an F distribution.

(iii) The OLS estimators are no longer BLUE.

Solution

8.1 Parts (ii) and (iii). The homoskedasticity assumption played no role in Chapter 5 in showing that OLS is consistent. But we know that heteroskedasticity causes statistical inference based on the usual t and F statistics to be invalid, even in large samples. As heteroskedasticity is a violation of the Gauss-Markov assumptions, OLS is no longer BLUE.

Question 14

Consider a linear model to explain monthly beer consumption:

$$beer = \beta_0 + \beta_1 inc + \beta_2 price + \beta_3 educ + \beta_4 female + u$$

$$\mathrm{E}(u|inc,price,educ,female) = 0$$

$$\mathrm{Var}(u|inc,price,educ,female) = \sigma^2 inc^2.$$

Write the transformed equation that has a homoskedastic error term.

Solution

8.2 $\mathrm{Var}(u|inc,price,educ,female) = \sigma^2 inc^2$, $h(\mathbf{x}) = inc^2$, where $h(\mathbf{x})$ is the heteroskedasticity function defined in equation (8.21). Therefore, $\sqrt{h(\mathbf{x})} = inc$, and so the transformed equation is obtained by dividing the original equation by inc:

$$\frac{beer}{inc} = \beta_0(1/inc) + \beta_1 + \beta_2(price/inc) + \beta_3(educ/inc) + \beta_4(female/inc) + (u/inc).$$

Notice that β_1, which is the slope on inc in the original model, is now a constant in the transformed equation. This is simply a consequence of the form of the heteroskedasticity and the functional forms of the explanatory variables in the original equation.

Question 15

Using the data in GPA3.RAW, the following equation was estimated for the fall and second semester students:

$$\widehat{trmgpa} = -2.12 + .900\ crsgpa + .193\ cumgpa + .0014\ tothrs$$
$$(.55)\ (.175)\qquad (.064)\qquad (.0012)$$
$$[.55]\ [.166]\qquad [.074]\qquad [.0012]$$
$$+ .0018\ sat - .0039\ hsperc + .351\ female - .157\ season$$
$$(.0002)\qquad (.0018)\qquad (.085)\qquad (.098)$$
$$[.0002]\qquad [.0019]\qquad [.079]\qquad [.080]$$
$$n = 269,\ R^2 = .465.$$

Here, *trmgpa* is term GPA, *crsgpa* is a weighted average of overall GPA in courses taken, *cumgpa* is GPA prior to the current semester, *tothrs* is total credit hours prior to the semester, *sat* is SAT score, *hsperc* is graduating percentile in high school class, *female* is a gender dummy, and *season* is a dummy variable equal to unity if the student's sport is in season during the fall. The usual and heteroskedasticity-robust standard errors are reported in parentheses and brackets, respectively.

(i) Do the variables *crsgpa*, *cumgpa*, and *tothrs* have the expected estimated effects? Which of these variables are statistically significant at the 5% level? Does it matter which standard errors are used?

(ii) Why does the hypothesis H_0: $\beta_{crsgpa} = 1$ make sense? Test this hypothesis against the two-sided alternative at the 5% level, using both standard errors. Describe your conclusions.

(iii) Test whether there is an in-season effect on term GPA, using both standard errors. Does the significance level at which the null can be rejected depend on the standard error used?

Solution

8.4 (i) These coefficients have the anticipated signs. If a student takes courses where grades are, on average, higher – as reflected by higher *crsgpa* – then his/her grades will be higher. The better the student has been in the past – as measured by *cumgpa* – the better the student does (on average) in the current semester. Finally, *tothrs* is a measure of experience, and its coefficient indicates an increasing return to experience.

The t statistic for *crsgpa* is very large, over five using the usual standard error (which is the largest of the two). Using the robust standard error for *cumgpa*, its t statistic is about 2.61, which is also significant at the 5% level. The t statistic for *tothrs* is only about 1.17 using either standard error, so it is not significant at the 5% level.

(ii) This is easiest to see without other explanatory variables in the model. If *crsgpa* were the only explanatory variable, H_0: $\beta_{crsgpa} = 1$ means that, without any information about the student, the best predictor of term GPA is the average GPA in the students' courses; this holds essentially by definition. (The intercept would be zero in this case.) With additional explanatory variables, it is not necessarily true that $\beta_{crsgpa} = 1$ because *crsgpa* could be correlated with characteristics of the student. (For example, perhaps the courses students take are influenced by ability – as measured by test scores – and past college performance.) But it is still interesting to test this hypothesis.

The t statistic using the usual standard error is $t = (.900 - 1)/.175 \approx -.57$; using the hetero-skedasticity-robust standard error gives $t \approx -.60$. In either case we fail to reject H_0: $\beta_{crsgpa} = 1$ at any reasonable significance level, certainly including 5%.

(iii) The in-season effect is given by the coefficient on *season*, which implies that, other things equal, an athlete's GPA is about .16 points lower when his/her sport is competing. The t statistic using the usual standard error is about –1.60, while that using the robust standard error is about –1.96. Against a two-sided alternative, the t statistic using the robust standard error is just significant at the 5% level (the standard normal critical value is 1.96), while using the usual standard error, the t statistic is not quite significant at the 10% level ($cv \approx 1.65$). So the standard error used makes a difference in this case. This example is somewhat unusual, as the robust standard error is more often the larger of the two.

Question 16

From the annual data for the U.S. manufacturing sector for 1899–1922, Dougherty obtained the following regression results[†]:

$$\widehat{\log Y} = 2.81 - 0.53 \log K + 0.91 \log L + 0.047t \quad \textbf{(1)}$$
$$\text{se} = (1.38) \quad (0.34) \quad (0.14) \quad (0.021)$$
$$R^2 = 0.97 \quad F = 189.8$$

where Y = index of real output, K = index of real capital input, L = index of real labor input, t = time or trend.

Using the same data, he also obtained the following regression:

$$\widehat{\log (Y/L)} = -0.11 + 0.11 \log (K/L) + 0.006t \quad \textbf{(2)}$$
$$\text{se} = (0.03) \quad (0.15) \quad (0.006)$$
$$R^2 = 0.65 \quad F = 19.5$$

a. Is there multicollinearity in regression (1)? How do you know?
b. In regression (1), what is the a priori sign of log K? Do the results conform to this expectation? Why or why not?
c. How would you justify the functional form of regression (1)? (*Hint:* Cobb–Douglas production function.)
d. Interpret regression (1). What is the role of the trend variable in this regression?
e. What is the logic behind estimating regression (2)?
f. If there was multicollinearity in regression (1), has that been reduced by regression (2)? How do you know?
g. If regression (2) is a restricted version of regression (1), what restriction is imposed by the author? (*Hint:* returns to scale.) How do you know if this restriction is valid? Which test do you use? Show all your calculations.
h. Are the R^2 values of the two regressions comparable? Why or why not? How would you make them comparable, if they are not comparable in the present form?

Solution

(*a*) Given the relatively high R^2 of 0.97, the significant F value and the (economically speaking) improperly signed insignificant coefficient of log K, it may be that there is collinearity in the model.

(*b*) A priori, capital is expected to have positive impact on output. It is not in the present case probably due to collinearity in the regressors.

(*c*) It is a Cobb-Douglas type production function, as the given model can be written as:

$$Y = \beta_1 K^{\beta_2} L^{\beta_3} e^{\beta_4 t}$$

(*d*)On average, over the sample period, a 1% increase in the index of the real labor input resulted in about 0.91% increase in the index of real output. The t variable in the model represents time. Very often, time is taken as a proxy for technical change. The coefficient of 0.47 suggests that over the sample period, on average, the rate of growth of real output (as measured by the output index) was about 4.7%.

(*e*) This equation implicitly assumes that there are constant returns to scale, that is, $(\beta_2 + \beta_3) = 1$. An incidental advantage of the transformation may be to reduce the collinearity problem.

(*f*) Given that the capital-labor ratio coefficient is statistically insignificant, it appears that the collinearity problem has not been resolved.

(*g*) As mentioned in (*e*), the author is trying to find out if there are constant returns to scale. One could use the *F* test discussed in Chapter 8 to find out if the restriction is valid. But since the dependent variables in the two models are different, we cannot use the R^2 version of the *F* test. We need the restricted and unrestricted residual sums of squares to use the *F* test.

(*h*) As noted in (*g*) the two R^2's are not comparable. One could follow the procedure discussed in Chapter 7 to render the two R^2 values comparable.

Question 17

For pedagogic purposes Hanushek and Jackson estimate the following model:

$$C_t = \beta_1 + \beta_2 \text{GNP}_t + \beta_3 D_t + u_i \qquad \textbf{(1)}$$

where C_t = aggregate private consumption expenditure in year t, GNP_t = gross national product in year t, and D = national defense expenditures in year t, the objective of the analysis being to study the effect of defense expenditures on other expenditures in the economy.

Postulating that $\sigma_t^2 = \sigma^2(\text{GNP}_t)^2$, they transform (1) and estimate

$$C_t/\text{GNP}_t = \beta_1 (1/\text{GNP}_t) + \beta_2 + \beta_3 (D_t/\text{GNP}_t) + u_t/\text{GNP}_t \qquad \textbf{(2)}$$

The empirical results based on the data for 1946–1975 were as follows (standard errors in the parentheses)*:

$$\hat{C}_t = 26.19 + 0.6248\ \text{GNP}_t - 0.4398\ D_t$$
$$(2.73) \qquad (0.0060) \qquad (0.0736) \qquad R^2 = 0.999$$

$$\widehat{C_t/\text{GNP}_t} = 25.92\,(1/\text{GNP}_t) + 0.6246 - 0.4315\,(D_t/\text{GNP}_t)$$
$$(2.22) \qquad (0.0068) \qquad (0.0597) \qquad R^2 = 0.875$$

a. What assumption is made by the authors about the nature of heteroscedasticity? Can you justify it?

b. Compare the results of the two regressions. Has the transformation of the original model improved the results, that is, reduced the estimated standard errors? Why or why not?

c. Can you compare the two R^2 values? Why or why not? (*Hint:* Examine the dependent variables.)

Solution

(a)The assumption made is that the error variance is proportional to the square of GNP, as is described in the postulation. The authors make this assumption by looking at the data over time and observing this relationship.

(b) The results are essentially the same, although the standard errors for two of the coefficients are lower in the second model; this may be taken as empirical justification of the transformation for heteroscedasticity.

(c) No. The R^2 terms may not be directly compared, as the dependent variables in the two models are not the same.

Question 17

Given the following regression:

$$\ln \hat{C}_t = -1.500 + 0.468 \ln I_t + 0.279 \ln L_t + -0.005 \ln H_t + 0.441 \ln A_t$$

$$se = (1.003)\ (0.166)\quad (0.115)\quad (0.143)\quad (0.107)$$

$$t = (-1.496)(2.817)\quad (2.436)\quad (-0.036)\quad (4.415)$$

$$R^2 = 0.936; \bar{R}^2 = 0.926; F = 91.543; d = 0.955$$

n=30

where

C = 12-month average U.S. domestic price of copper (cents per pound)
G = annual gross national product ($, billions)
I = 12-month average index of industrial production
L = 12-month average London Metal Exchange price of copper (pounds sterling)
H = number of housing starts per year (thousands of units)
A = 12-month average price of aluminum (cents per pound)

Question:

Estimate the Durbin-Watson d statistic and comment on the nature of autocorrelation present in the data.

(c)As shown in the regression output given in (a) above, the d statistic is 0.955. Now for $n = 30$, $k' = 4$ and $\alpha = 5\%$, the lower limit of d is 1.138.Since the computed d value is below this critical d value, there is evidence of positive first-order autocorrelation.

Question 18

The regression results are:

$$\hat{Y}_t = 246.240 + 15.182X_t$$
$$se = (\ 5.849)\quad (0.643)$$
$$t\ = (42.104)\quad (23.603)$$
$$r^2 = 0.977; d = 0.4148$$

n=15

Is there positive serial correlation in the disturbances?

Solution:

(b) Yes. For $n = 15$, $k' = 1$ and $\alpha = 0.05$, $d_L = 1.077$. Since the computed d is smaller than d_L, there is evidence of positive first-order autocorrelation.

$$

I will be in my office this Monday and Tuesday for answering your questions.

Best,
Aj.Wasin