Econometrics Problem Set Due 4/24

Felicia Cowley[1]

[1]George Mason University

April 24, 2018

Chapter 7 Problem 1:

Using the data in SLEEP75, we obtain the estimated equation:

$$\hat{sleep} = 3840.83 - .163totwrk - 11.71educ - 8.70age + .128age^2 + 87.75male$$

(235.11) (.018) (5.86) (11.21) (.134) (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?

Looking at the coefficient in front of "male," the coefficient is 87.75 which is a close to an hour and a half more sleep for men compared to women. To test the evidence, we can use a t test to see if the coefficient is statistically significant. $t = \frac{E(m)-\mu_m)}{Se(m)} = \frac{87.75-0}{34.33} = 2.56$

To find the critical value, $n - k - 1 = 706 - 5 - 1 = 700$ which means that for a two tailed test the t is close to 2.576 which is at the 1% level. This means that the evidence is highly statistically significant that men sleep more than women.

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

The coefficient with working is .163 and using the t test, $t = \frac{E(m)-\mu_m}{Se(m} = \frac{-0.163}{0.018} = -9.06$. The absolute value is obviously larger than the critical value found in (i) meaning that working is a statistically significant. The trade off is that an extra hour of work is .163*60 = 9.8 minutes of less sleep.

(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?

I would need to determine the R-squared from both running a restricted regression and unrestricted regression so that I can compare the significance with the F-test. In this case, the unrestricted model would have all the variables except the one we are trying to test, *age* (and *age* squared as well). This means that if *age* (*age* squared) is in the model, age will be statistically insignificant or no effect if the parameters on both terms equal zero.

Chapter 7 Problem 2:

The following equations were estimated using the data in BWGHT:

$\hat{log(bwght)} = 4.66 - 0.0044cigs + 0.0093log(faminc) + 0.016parity + 0.027male + 0.055white$

0.22 0.0009 0.0059 0.006 0.010

$n = 1388, R^2 = 0.0472$

and

$\hat{log(bwght)} = 4.65 - 0.0052cigs + 0.0110log(faminc) + 0.017parity + 0.034male + 0.045white - 0.0030motheduc + 0.0032fatheduc$

(0.38) (0.0010) (0.0085) (0.006) (0.011)

$n = 1191, R^2 = 0.0493$

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?

Because bwght is a log function, the coefficient will be in a percentage change. If cigs or the consumption of cigarettes increase by 5 per day, then $5(-0.0044)(100) = -2.22.2\% lowerbirthweight$.

(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 significance?

There is a 5.5% (100*0.055) difference in weight between a white child and nonwhite child. To find significance, we use the t test:

$t = \frac{E(m)-\mu_m}{Se(m)} = \frac{0.055}{0.013} = 4.23$ Knowing that n-k-1=1388-5-1=1382 which is greater than 2.576 on the t table for infinite degrees of freedom means that with a two tailed test, the difference between white and nonwhite babies is statistically significant.

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

Using the t test for the coefficient and standard error of *motheduc* we see that $t = \frac{E(m)-\mu_m}{Se(m)} = \frac{0.0030}{0.0030} = 1$. This is less than than any critical value on the t table for infinite degrees of freedom and means that the mother education variable is not statistically significant.

(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?

We cannot compute the statistic because the regressions are different in observations as motheduc and fatheduc are missinging. In order to compute the F statistic, we need to reestimate the first equation using the same observations used to estimate the second equation.

Chapter 8 Problem 4:

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

$\hat{trmgpa} = -2.12 + .900crsgpa + .193cumgpa + .0014tothrs + .0018sat - .0039hsperc + .351female - .157season$

(.55) (.175) (.064) (.0012) (.0002) (.0018)

(.55) (.166) (.074) (.0012) (.0002) (.0019)

$n = 269, R^2 = .465$

(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 what standard errors are used?

The variables do have the expected estimated effects since you would expect that the weighted average of overall GPA in courses taken, GPA prior to the current semester, and total credit hours to the semester would be positively correlated with term GPA. You would expect that how well a student does in the current term would generally be similar to his or her performance in prior terms.

To find statistical significance, we use the t-test on each variable using the standard error and heteroskedasticity-robust standard error.

$t = \frac{E(m)-\mu_m}{Se(m)} = \frac{.900}{.175} = 5.14$ and $t = \frac{E(m)-\mu_m}{Se(m)} = \frac{0.900}{0.166} = 5.42$ for *crsgpa*

$t = \frac{E(m)-\mu_m}{Se(m)} = \frac{.193}{.064} = 3.02$ and $t = \frac{E(m)-\mu_m}{Se(m)} = \frac{0.193}{0.074} = 2.61$ for *cumgpa*

$t = \frac{E(m)-\mu_m}{Se(m)} = \frac{.0014}{.0012} = 1.17$ and $t = \frac{E(m)-\mu_m}{Se(m)} = \frac{0.0014}{0.0012} = 1.17$ for *tothrs*

The critical value is 1.96 since we have n-k-1=269-7-1=261 degrees of freedom and a 5% level. Since the t values for *crsgpa* and *cumgpa* are greater than the critical value, they are statistically significant values whereas *tothrs* is not at the 5% significance level. In this case it does not matter which standard errors are used to determine statistical significance since the use of one or the other doesn't change whether the variables are statistically significant.

(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.

This null hypothesis would make sense if crsgpa is the only explanatory variable since you would assume that weighted average of overall gpa would be correlated with term gpa however in the instance that other explanatory variables are included, then the null would no longer make sense since the added variables could be correlated with term gpa.

To test, we let $H_0 : \hat{\beta}_{crsgpa} = 1$ and $H_A : \hat{\beta}_{crshpa} \neq 1$ and use the t statistic using both the standard error and heteroskedasticity-robust standard error:

$t = \frac{\hat{\beta}-1}{SE(\hat{\beta})} = \frac{0.9-1}{0.175} = -0.57$ and $t = \frac{\hat{\beta}-1}{SE(\hat{\beta})} = \frac{0.9-1}{0.166} = -0.6$.

The critical value is 1.96 for infinite degrees of freedom and a 5% confidence interval. Both t statistics are less than the critical value meaning that we reject the null and we know that *crsgpa* is statistically insignificant at a 5% level of significance.

(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?

Using the t testfor both standard errors, $t = \frac{\hat{\beta}_{season}}{SE(\hat{\beta}_{season})} = \frac{-.157}{.098} = -1.6$ and $t = \frac{\hat{\beta}_{season}}{SE(\hat{\beta}_{season})} = \frac{-.157}{.080} = -1.96$. To find the critical value, n-k-1=269-7-1=261 degrees of freedom (basically infinity on the table), we find that at a 5% confidence interval, c=1.96. The absolute value of the heteroskedasticity-robust standard is 1.96 so using this error means that season in statistically significant at a 5% level of significance. However, for the same level of significance, using the regular standard error, season is not statistically significant. This means that the null can be rejected dependent upon which standard error is used.

Chapter 8 Problem 6:

There are different ways to combine features of the Breusch-Pagan and White tests for heteroskedasticity. One possibility not covered in the test is to run the regression: $\hat{u_i^2} on x_{i1}, x_{i2}, ..., x_{ik}, \hat{y_i^2}, i =$

$1, ..., n$, **where the $\hat{u_i}$ are the OLS residuals and the $\hat{y_i}$ are the OLS fitted values. Then, we would test joint significance of $x_{i1}, x_{i2}, ..., x_{ik}$ and $\hat{y^2}$.**

(i) What are the df associated with the proposed F test for heteroskedasticity?

The degrees of freedom for the F test is equal to (k,2) in numerator and (n-k-1,n-3) in the denominator. The degrees of freedom for the chi-squared test are (k,2). This means that if the Breusch-Pagan and White tests are significant then there might be heteroskedasticity. If the two tests have no significance, then we accept the null hypothesis of heteroskedasticity.

(ii) Explain why the R-squared from the regression above will always be at least as large as the R-squareds for the BP regression and the special case of the White test.

The hyprid test has an extra regressor of y hat squared meaning that the R-square will not be less for the BP test. For the White test the fitted values are linear of the regressors meaning that there is a restriction on how the original explanatory varibles can be presented in the regression. The R-squared will be no greater than the R-squared from the hybrid equation.

(iii) Does part (ii) imply that the new test always delivers a smaller p-value than either the BP or special case of the White statistic? Explain.

No since the F test also depends on degrees of freedom which is different in all three tests, the BP test, the White test, and the hybrid.

(iv) Suppose someone suggests also adding $\hat{y_i}$ to the newly proposed test. What do you think of this idea?

The OLS fitted values are linearly combined of the prior regressors. Since those regressors are also in the hybrid test, addind the OLS fitted values would be helpful and result in perfect collinearity.

Chapter 8 Problem 8 (i) and (ii):

(i) Compute the usual Chow statistic for testing the null hypothesis that the regression equations are the same for men and women. Find the p-value of the test.

Assuming the SSR for n=406 is 38,781.38 and SSR for n=408 is 48,029.82, this means unrestricted SSR is the sum: 86,811.20. The SSR for the third equation given is 87,128.96 when n=814. The F statistic is $F = \frac{SSR_R - SSR_{UR}}{SSR_{UR}} \frac{n-k-1}{q} = \frac{87128.96-86811.20}{86811.20} \frac{814-6}{3} = .99$ with a p value of .614.

(ii) The F statistic using the same values as in (i) which means that once again F=.99 and the p value is .614.

Chapter 9 Problem 1:

There exists functional form misspecification with the population parameters (since $\beta_6 \neq 0, \beta_7 \neq 0$) on ceoten^2 and comten^2. We have to test joint significance of the variables using the F-test. $F = \frac{.375-.353}{1-.375} \frac{177-8}{2} = 2.97$

For 2 degrees of freedom in the numerator and 169 degrees of freedom in the denominator, at 5% level of significance, the F statistic is 3.0. Since the estimated F statistic is less than the critical F statistic, the two variables are not jointly significant at a 5% level and there is no mis-speification. But at a 10% level of siginifcance and 2 degrees of freedom in the numerator and 169 degrees of freedom in the denominator, the critical F statistic changes to 2.3. The estimated F-statistic is now greater than the estimated F statistic at 10% level of significance and we see the variables are jointly correlated signalling mis-specification.

Chapter 9 Problem 2:

(i) Interpret the coefficient on *voteA88* and discuss its statistical significance.

As provided in question, the coefficient for *voteA88* is 0.067 which means that increasing the vote by one percent of Candidate A amounts to a 0.067% point increase in the cote during 1990. We can test significance using the t test.

$t = \frac{\hat{\beta_{voteA88}} - \beta_{voteA88}}{SE_{voteA88}} = \frac{.067-0}{.053} = 1.2641$. The critical value from the t table for 186-5-1=180 degrees of freedom at a 5% level of significance level is 1.645. Because the t statistic is less than the critical value, *voteA88* is not statistically significance.

(ii) Does adding voteA88 have much effect on the other coefficients?

Due to the statistical insignificance of voteA88, including it in the model does not affect the signs or statistical significance of the coefficients of other the explanatory variables. However, this could have a magnitude effect on the other variables since prtystrA changes from .312 to .282, democA changes from 4.93 to 4.52, log(expendA) changes from -.929 to -.839, and log(expendB) changes from -1.95 to -1.846.

Chapter 9 Problem 3:

(i) The variable *lnchprg* is the percentage of students eligible for the federally funded school lunch program. Why is this a sensible proxy variable for poverty?

This might be a sensible proxy since children on the school lunch program are generally below the poverty line and can represent the number of children in poverty.

(ii) The table that follows... Explain why the effect of expenditures on *math10* is lower in column (2) than in column (1). Is the effect in column (2) still statistically greater than zero?

The variables log(expand) and lnchprg are negatively correlated since school districts with poorer children spend less on schools. $\beta_3 < 0$ with omitted lnchprg from the regression causes an unpward biased estimate of beta 1. Controlling for the poverty rate means the effect of spending falls.

(iii) The pass rate for math10 is lower at larger schools all else equal since the coefficient of log(enroll) is -1.26 in column 2 which indicates that an increase of 10% in enroll means a decrease in math10 by 0.126% points.

(iv) The coefficient of lnchprg in column 2 is -0.324 meaning that a 1% increase in lnchprg results in a 0.324% point decrease in math10, all else equal.

(v) The R-squared value in column 1 indicates that the model without variable *lnchprg* could explain a 2.97% variation in math10 where the R-squared value in column 2 indicates the model with *lnchprg* could explain 18.93% of the variation in *math10*. This means that including *lnchprg* in the model increases the explanatory power of the model meaning that *lnchprg* is more important a determinant of *math10* than *log(enroll)* or *log(expend)*.

Chapter 10 Problem 1

(i) Most time series observations are correlated over time, unlike cross-sectional observations. This means that there may be a natural trend or common tendency of growing over time in the time series data so the statement is false.

(ii) I agree with this statement as it follows from the theorem 10.1. We do not need homoskedasticity and no serial correlation assumptions.

(iii) This is false. Trending variables are often used as dependent variables in regression models. Interpreting the results is tricky because we might find a inaccurate association between yt and explanatory variables. Including a trend regression is helpful for trending dependent or independent variables.

(iv) This is true. With annual data, time periods represent a year and not associated with a season.

Chapter 10 Problem 4

To find joint significance we use the F statistic.

$F = \frac{.305-.281}{1-.305}\frac{124}{3} = 1.43$ The critical value for a 10% significance level is 2.13 given degrees of freedom is 120 (since 124-3-1=120). The F statistic is below the critical value meaning the variables are jointly insignificant at the 10% level.

Chapter 10 Problem 5

An example of a model would be $log(housingstarts) = \alpha_0 + \alpha_1 t + \delta_1 Q2_t + \delta_2 Q3_t + \delta_3 Q4_t + \beta_1 int_t + \beta_2 log(rcpinc_t) + u_t$

where Q2t,Q3t, and Q4t are quarterly dummy variables and the remaining variables are explanatory. The linear time trend gives the dependent variable and log(rcpinc) to trend over time and the quarterly dummy variables allow all variables to demonstrate seasons. The beta 2 is elasticity and beta time 100 is semi-elasticity.