Introductory Econometrics

Tutorial 8

PART A: To be done before you attend the tutorial. The solutions will be made available at the end of the week.

1. Assume an OLS regression of a variable y on k regressors collected in $\boldsymbol{X}$ (excluding the intercept term). The sample size is equal to n. Using the formulae for R^2 and $\bar{R}^2$:

   (a) Prove that
   $$\bar{R}^2 = 1 - \left(1 - R^2\right)\left(\frac{n-1}{n-k-1}\right). \tag{1}$$

   (b) Compare R^2 and $\bar{R}^2$ when $k = 0$ and when $k > 0$.

   (c) What is the use of $\bar{R}^2$?

2. The following model is estimated using the quarterly international visitor arrivals in Victoria (the quarterly version of the data set used in the lecture last week).

Dependent Variable: LOG(VIC)
Method: Least Squares
Sample: 1991Q1 2018Q2
Included observations: 110

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	11.56726	0.019943	580.0061	0.0000
T	0.016191	0.000234	69.08998	0.0000
Q1	-0.028685	0.021049	-1.362796	0.1759
Q2	-0.364213	0.021047	-17.30455	0.0000
Q3	-0.302542	0.021239	-14.24460	0.0000

R-squared	0.980364	Mean dependent var	12.29157
Adjusted R-squared	0.979616	S.D. dependent var	0.546551
S.E. of regression	0.078032	Akaike info criterion	-2.218994
Sum squared resid	0.639352	Schwarz criterion	-2.096245
Log likelihood	127.0447	Hannan-Quinn criter.	-2.169207
F-statistic	1310.585	Durbin-Watson stat	0.539978
Prob(F-statistic)	0.000000		

In this regression T is a time trend (i.e. a non-random variable that starts from 1 and goes up by one unit each time period, here its values will be $1, 2, 3, \ldots, 110$), Q1 is a dummy variable for quarter 1 (i.e. it is equal to one when the observation is from quarter 1 of each year and is zero otherwise), and similarly Q2 and Q3 are dummy variables for quarter 2 and quarter 3.

(a) Why do we not have a dummy variable for Q4 in this regression? What happens if we add a dummy variable for Q4 as well?

(b) On a time series plot (a plot that has T on the x-axis) the predictions of this model for $\log(VIC)$ in each quarter lie on a separate line. How do these lines differ, in particular do they have different intercepts, different slopes, or both? Do a rough hand sketch of these lines given the estimation results.

(c) How would the estimation results change if we dropped Q1 and added Q4 instead? How about if we dropped Q2 and added Q4? And if we dropped Q3 and added Q4? [Yes, this is repetitive, but repetition sometimes helps to cement the idea.] After doing these by a calculator, check your calculations by running these regressions using the victouristquarterly.wf1 file on Moodle.

3. Use the data in hprice1.wf1 uploaded on Moodle for this exercise.

   (a) Estimate the model

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

   and report the results in the usual form, including the standard error of the regression. Obtain the predicted price when we plug in $lotsize = 10,000$, $sqrft = 2,300$, and $bdrms = 4$; round this price to the nearest dollar.

   (b) Run a regression that allows you to put a 95% confidence interval around the predicted value in a). Note that your prediction will differ somewhat due to rounding error.

Do not forget to bring your answers to PART A and a copy of the tutorial questions to your tutorial.

<u>Part B:</u> This part will be covered in the tutorial. It is still a good idea to attempt these questions before the tutorial.

The purpose of this tutorial is to practice running regressions with transformed variables, using binary variables and also forming prediction intervals. It adds one thing to the material covered in the lecture, and that is an alternative way to calculate the value of the F test statistic for the special case that we want to allow both the intercept and all the slope parameters to be different between different groups. This is covered in section 7.4 of the textbook. In this special case, the F test is also called the Chow test, named after Gregory Chow https://en.wikipedia.org/wiki/Gregory_Chow. *Unless otherwise specified, use the 5% level of significance for all tests.*

1. *Logarithmic and quadratic model):*

   We will be using **wage1tute4.wf1**, from Tutorial 4 for this question.

   (a) Use OLS to estimate the equation

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

   and report the results using the usual format. Name this **eq01**.

   [Note that in most other statistical software, you have to generate log(*wage*) and *exper^2* first, give them names like *lwage* and *expersq*, then use them in the regression command. Eviews allows you to do this inside the regression command, which is a great advantage].

   (b) Is $exper^2$ statistically significant at the 1% level?

   (c) Using the approximation

   $$\%\Delta\widehat{wage} \approx 100\left(\hat{\beta}_2 + 2\hat{\beta}_3 exper\right)\Delta exper,$$

   find the approximate return to the fifth year of experience. What is the approximate return to the twentieth year of experience?

   (d) At what value of *exper* does additional experience actually lower predicted log (*wage*)? How many people have more experience in this sample?

2. The data set marks.wf1 contains data on students' performance in 2015 in Introductory Econometrics. The variables are:

$exam$	:	mark on final exam (%)
$asgnmt$	:	mark on assignments (%)
$etc3440$	:	=1 for students in ETC3440, =0 for students in ETC2410.

There were no students taking Introductory Econometrics under any other code in that year. The goal is to make a predictive model that uses assignment mark to predict final exam mark.

(a) "Look at the data." Do you see any anomalies in the final exam marks? If so, discuss what you would do about it. Your tutor will moderate the discussions and whatever decision the tutorial group makes will be maintained throughout the rest of this exercise.

(b) By using the dummy variable $etc3440$ in a regression, test if there is any difference between the expected exam mark in ETC2410 and ETC3440.

(c) Estimate a regression of $exam$ on a constant and $asgnmt$, along with the scatter plot with regression line added. Based on this regression, provide predictions for the exam marks of a student who has obtained 80% on the assignment. Provide 95% prediction intervals for the exam mark, once ignoring estimation uncertainty, and once properly accounting for estimation uncertainty. The point here is that incorporating estimation uncertainty widens the prediction interval but by very little. In practice, it is often ignored. But we have to be careful here because we had only two parameters to estimate and more than 100 observations. The effect of estimation uncertainty can be larger if we had a large number of independent variables.

(d) Specify and estimate a regression model to test whether *both the intercept and slope* in the regression in part (c) is different for ETC2410 and ETC3440. From the unrestricted regression results, obtain the estimate of the intercept and slope for ETC2410 and for ETC3440 regression lines.

(e) Estimate two separate models to predict exam mark based on assignment mark, one for ETC2410 students and another for ETC3440 students. Compare the estimates of the intercept and slope you obtain here with what you obtained in part (d) and discuss. Also add the sum of squared residuals of these two separate models and compare that to the SSR of your unrestricted model in part (d). Can you explain what you see intuitively? The goal here is to learn that when you want to test if all parameters (i.e. the entire conditional expectation function) is different for different groups, you can compute the SSR of the unrestricted model by running separate regressions for each group. The practical usefulness of this result is for models with many explanatory variables.

(f) Consider the estimated intercept and slope for ETC2410 students in part (d) and part (e). While the numerical values are the same, their standard errors are not. Discuss why they are different, and which one you would prefer to use to construct a 95% confidence interval for the slope parameter for ETC2410 students.

Notes:

- I am assuming you know how to get a scatter plot of two variables in Eviews. To show the regression line on a scatter plot, in the Graph Options, when you choose Scatter, you will get as an option "Fit lines" with the default value of "None". In the drop down menu in front of "Fit lines" you can choose "Regression Line" and then press OK. You will get a scatter plot with the OLS estimated regression line.

- You can exclude certain observations from all analysis by changing your Sample. Click on Sample, and in the "IF condition" window enter a logical expression that will only include the observations that you want to use. For example, entering "exam > 0" will only include observations whose final exam mark is strictly positive. You can have more than one condition by connecting logical expressions with "and" or "or" to get what you want. For example, if you want to consider only ETC2410 student with strictly positive exam marks, in the "IF condition" window you can enter "exam > 0 and etc3440=0".

- If you want to exclude some observations from a specific regression, not from all analysis, then you can do that within the equation window. For example, to get a regression of exam marks on a constant and assignment marks for ETC2410 students who had strictly positive exam marks, following Quick/Estimate Equation and entering "exam c asgnmt" in the equation specification window, in the "Sample" window that gives you the range of data, say, "1 118", add "if exam>0 and etc3440=0" after 118. Remember that here you have to type "if" also, whereas in the previous bullet point in the "IF condition window", you only needed to enter the logical expressions.

- In order to get the standard error for predicted value of exam given a particular assignment mark, we use the property that OLS results do not change qualitatively when we add or subtract a constant from an explanatory variable. Only the interpretation of the constant term changes. So, if we rerun the regression with $asgnmt - 80$ as the x variable instead of $asgnmt$, then the constant term will be the prediction for the final exam when $asgnmt - 80 = 0$, that is, when $asgnmt = 80$. The calculation of the prediction when $asgnmt = 80$ is not a big deal, but getting its standard error would have required using the estimated variance and covariances of the estimated intercept and slope and using the formula for the variance of a linear combination of the intercept and the slope. With this trick, we get the standard error of $\widehat{exam}$ directly. This is a very useful trick.

- It is important to note the distinction between the confidence interval for $\widehat{exam}$ and the prediction interval for $exam$ conditional on $asgnmt = 80$. Given $asgnmt = 80$, $\widehat{exam}$ varies in different samples because the estimates of the intercept and slope vary, that is, it only varies because of "estimation uncertainty". The actual exam mark, however, includes u, a source of uncertainty that we cannot explain with assignment marks, so the prediction interval for $exam$ is much wider, because it allows for the variation in u in addition to the variation in the estimated conditional mean. In fact, the variation in u dominates and as we get larger and larger samples, the estimation uncertainty becomes smaller and smaller while the variation due to u does not change. Your tutor will emphasise this in the tutorial.