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. Use the data in hprice1.wf1 uploaded on Moodle for this exercise. We assume that all assumptions of the Classical Linear Model are satisfied for the model used in this question.

   (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 compute the 95% confidence interval of
   $$E\left(price \mid lotsize = 10000,\ sqrft = 2300,\ bdrms = 4\right)$$
   Note that your prediction may differ somewhat due to rounding error. Compute this confidence interval. If you were going to an auction of a house with $lotsize = 10,000$, $sqrft = 2,300$, and $bdrms = 4$, based on this data, would you be 95% confident that the price will be in this interval?

   (c) Compute a 95% prediction interval for the price of house with $lotsize = 10,000$, $sqrft = 2,300$, and $bdrms = 4$. If you were going to an auction of a house with $lotsize = 10,000$, $sqrft = 2,300$, and $bdrms = 4$, based on this data, would you be 95% confident that the price will be in this prediction interval?

Notes:

- In order to get the standard error for predicted value of price given lot size, square footage and number of bedrooms, 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 you rerun the regression with $lotsize - 10000$, $sqrft - 2300$, and $bdrms - 4$ as explanatory variables, then the constant term will be the predicted price of a house with $lotsize = 10,000$, $sqrft = 2,300$, and $bdrms = 4$. The calculation of the prediction 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 parameters and using the formula for the variance of a linear combination of these to compute the variance of the prediction. With this reparameterisation trick, you get the standard error of $\widehat{price}$ directly. This is a very useful trick.

- It is important to note the distinction between the confidence interval for

$$E\left(price \mid lotsize = 10000,\ sqrft = 2300,\ bdrms = 4\right) = \beta_0 + 10000\beta_1 + 2300\beta_2 + 4\beta_3$$

and the prediction interval for *price* conditional on $lotsize = 10000$, $sqrft = 2300$, $bdrms = 4$. Our estimate of the mean of price given house characteristics varies in different samples because the estimates of the intercept and slope parameters vary, that is, it only varies because of "estimation uncertainty". The *price* itself, however, includes u, a source of uncertainty that we cannot explain with the three observed characteristics, so the prediction interval for *price* is much wider, because it allows for the variation in u in addition to the variation in the estimated coefficients. 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.

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.

1. *Logarithmic and quadratic model with dummy variables:* We have a data set that includes data for a random sample of 526 individuals (this is quite an old data set from the mid-eighties and is used for educational purposes only. The conclusions made here give a picture of the labour market in the mid-eighties). The variables in the data set are:

Variable	Description
wage	hourly wage in dollars
educ	years of education
exper	years of experience
female	=1 if the person is female, 0 otherwise
married	=1 if the person is married, 0 otherwise
urban	=1 if the person lives in an urban area, 0 otherwise

We have estimated the following model using OLS:

Dependent Variable: LOG(WAGE)
Method: Least Squares
Sample: 1 526
Included observations: 526

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.3354	0.1012	3.3136	0.0010
EDUC	0.0762	0.0070	10.8122	0.0000
EXPER	0.0360	0.0052	6.9935	0.0000
EXPER^2	-0.0006	0.0001	-5.7341	0.0000
FEMALE	-0.3319	0.0361	-9.2063	0.0000
MARRIED	0.0812	0.0415	1.9546	0.0512
URBAN	0.1773	0.0409	4.3410	0.0000
R-squared	0.4232	Mean dependent var		1.6233

[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]

(a) By referring to their p-values only, determine if each of these dummy variables is statistically significant at the 5% level (no need to write all the steps of hypothesis testing).

(b) Interpret each of the estimated parameters.

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) How would the estimation results (in particular coefficients of each regressor, the R^2, SSR and standard error of the regression) change if we dropped Q1 and added Q4 instead? After answering this question using a calculator, check your calculations by running the regression using the victouristquarterly.wf1 file on Moodle. [To make sure that you have understood this, in your own time outside of the tutorial, answer these questions: 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.]

(c) 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.

(d) Test the hypothesis that the true intercepts for Q2 and Q3 are equal, versus the alternative that they are not equal, at the 5% level of significance using a t-test.

(e) Test for a structural break due to the global financial crisis. To do that, generate a dummy variable called gfc that is 0 before 2008Q3 and 1 in and after 2008Q3, and then test that all coefficients of the above regression have been the same before and after the GFC versus the alternative that at least some have changed. Perform the test at the 5% level of significance. From the unrestricted regression, what does the data reveal about the effect of the GFC on international tourism in Victoria? (Note: in EViews, series $gfc = @after("2008Q3")$ in the command window or $gfc = @after("2008Q3")$ in the generate series window creates the GFC dummy).