Introductory Econometrics
Tutorial 8 Solutions

PART A:

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

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

Answer: We have that:

$$R^2 = 1 - \frac{RSS}{TSS},$$

or

$$\frac{RSS}{TSS} = 1 - R^2.$$

Also,

$$\bar{R}^2 = 1 - \frac{RSS/\left(n-k-1\right)}{TSS/(n-1)},$$

or

$$\bar{R}^2 = 1 - \left(1 - R^2\right)\left(\frac{n-1}{n-k-1}\right),$$

as required.

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

Answer: When $k = 0$, then from (1) we have that $\bar{R}^2 = R^2$. When $k > 0$, then using the same expression we have that $\bar{R}^2 < R^2$.

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

Answer: $\bar{R}^2$ is used for selecting among competing models for the same dependent variable. We cannot use R^2 for that purpose because as we add explanatory variables to a model, RSS always decreases which makes R^2 increase even when the additional explanatory variables have no predictive power for explaining the dependent variable. $\bar{R}^2$ has the number of explanatory variables k in its formula in a way that if we make k larger, RSS has to decrease by a large amount for $\bar{R}^2$ to improve.

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?
Answer: Because we have a constant and three dummies, it would be redundant to have a Q4 dummy as well. If the other 3 dummies are zero, we know that we must be in Q4. So, the implied model for Q4 is $11.5676 + 0.16191T$. Technically, if we add Q4 then we will have exact multicollinealrity in the X matrix. Our X matrix will be:

$$\begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 2 & 0 & 1 & 0 & 0 \\ 1 & 3 & 0 & 0 & 1 & 0 \\ 1 & 4 & 0 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

which will have column1=column3+column4+column5+column6 exactly. This means columns of X will be linearly dependent, hence $X'X$ will not be invertible and we cannot compute the OLS estimator.

(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.
They will have the same slope but different intercepts.

$$\begin{aligned} Q1 &: \widehat{\log(VIC)} = (11.56726 - 0.028685) + 0.016191 \times T = 11.53858 + 0.016191 \times T \\ Q2 &: \widehat{\log(VIC)} = (11.56726 - 0.364213) + 0.016191 \times T = 11.20305 + 0.016191 \times T \\ Q3 &: \widehat{\log(VIC)} = (11.56726 - 0.302542) + 0.016191 \times T = 11.26472 + 0.016191 \times T \\ Q4 &: \widehat{\log(VIC)} = 11.56726 + 0.016191 \times T \end{aligned}$$

Your sketch must show four upward sloping parallel straight lines with Q4 line being above all four, then Q1 line below Q4 but by not much distance, then Q3 a larger distant below Q1, and Q2 below Q3 but by not much.

(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.

If we dropped Q1 and added Q4, then the intercept will be the intercept for Q1, which is 11.53858. The coefficients of the three dummies will be the difference between the intercept of that quarter from the intercept of Q1. So the coefficient of Q2 will be $11.20305 - 11.53858 = -0.33553$, the coefficient of Q3 will be $11.26472 - 11.53858 = -0.27386$, and the coefficient of Q4 will be $11.56726 - 11.53858 = 0.02868$. The coefficient of T, the R^2, the sum of squared residuals will all stay exactly the same.

If we dropped Q2 and added Q4, then the intercept will be the intercept for Q2, which is 11.20305. The coefficients of the three dummies will be the difference between the intercept of that quarter from the intercept of Q2. So the coefficient of Q1 will be $11.53858 - 11.20305 = 0.33553$, the coefficient of Q3 will be $11.26472 - 11.20305 = 0.06167$, and the coefficient of Q4 will be $11.56726 - 11.20305 = 0.36421$. The coefficient of T, the R^2, the sum of squared residuals will all stay exactly the same.

Finally, if we dropped Q3 and added Q4, then the intercept will be the intercept for Q3, which is 11.26472. The coefficients of the three dummies will be the difference between the intercept of that quarter from the intercept of Q3. So the coefficient of Q1 will be $11.53858 - 11.26472 = 0.27386$, the coefficient of Q2 will be $11.20305 - 11.26472 = -0.06167$, and the coefficient of Q4 will be $11.56726 - 11.26472 = 0.30254$. The coefficient of T, the R^2, the sum of squared residuals will all stay exactly the same.

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.

Answer: The estimated equation (when *price* is in dollars) is:

$$\begin{aligned} price &= \underset{(29,475)}{-21,770.3} + \underset{(0.642)}{2.068} lotsize + \underset{(13.24)}{122.78} sqrft + \underset{(9,010.1)}{13,852.5} bdrms \\ n &= 88,\ R^2 = 0.672,\ \bar{R}^2 = 0.661,\ \hat{\sigma} = 59,833. \end{aligned}$$

The predicted price at $lotsize = 10,000$, $sqrft = 2,300$, and $bdrms = 4$ is about \$336,714.

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.

Answer: The regression is $price_i$ on $(lotsize_i - 10,000)$, $(sqrft_i - 2,300)$, and $(bdrms_i - 4)$. We want the intercept estimate and the associated 95% CI from this regression. The CI is approximately $336,706.7 \pm 14,665$, or about \$322,042 to \$351,372 when rounded to the nearest dollar.

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

PART B: You do not need to hand this part in. It will be covered in the tutorial. It is still a good idea to attempt these questions before the tutorial.

1. *(Logarithmic and quadratic model):* Chapter 6, problem C2, page 199.

   (a) The estimated equation is

$$\begin{aligned} \widehat{\log(wage)} &= \underset{(0.106)}{0.128} + \underset{(0.0075)}{0.0904}educ + \underset{(0.0052)}{0.0410}\ exper - \underset{(0.0001)}{0.0007}\ exper^2 \\ n &= 526,\ R^2 = 0.30 \end{aligned}$$

   (b) **Please remind students on how to answer a test of significance appropriately by stating the null and the alternative, the test statistics, the distribution of the test statistic under the null, and using the pvalue when reported or comparing the test statistic with the critical value of a certain significance level obtained from the appropriate tables.** The t statistic on exper2 is about –6.16, which has a p-value of essentially zero. So exper2 is significant at the 1% level (and much smaller significance levels).

   (c) To estimate the return to the fifth year of experience, we start at exper = 4 and increase exper by one, so Δexper = 1:

$$\%\Delta\widehat{wage} = 100(0.0410 - 2 \times 0.0007 \times 4) \approx 3.5\%$$

   Similarly, for the 20th year of experience,

$$\%\Delta\widehat{wage} = 100(0.0410 - 2 \times 0.0007 \times 19) \approx 1.4\%$$

   (d) The turnaround point is about $0.041/(2 \times 0.0007) \approx 29$ years of experience. In the sample, there are 121 people with at least 29 years of experience. This is a fairly sizeable fraction of the sample. *Some students may be worried about slight variation in results depending on the number of significant digits that they may use. Please assure them that such rounding will not make qualitative difference and in the exam, if they are asked to do any calculations, they should use the numbers at any precision that are presented and they are allowed to round the numbers to the extent that they think is reasonable. For example, here we rounded the % change in wage to one decimal point, and the years of experience to the nearest integer because we thought that those were the levels of precision that made sense.*

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

exam	:	mark on final exam (%)
asgnmt	:	mark on assignments (%)
*etc*3440	:	=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.

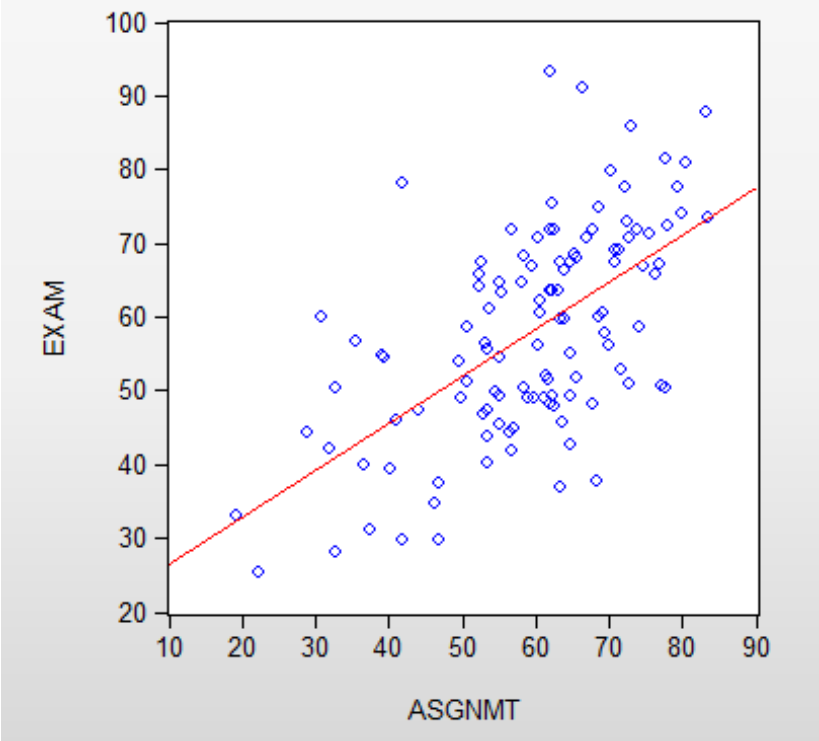
   (a) "Look at the data." Do you see any anomalies in the final exam marks? If so, discuss what you would do about it. Your tutor will moderate the discussions and whatever decision the tutorial group makes will be maintained throughout the rest of this exercise.
   There are three students with zero marks. The best way to treat them is to drop them because we know that these may have been students with very special circumstances, so we do not want these to affect our estimate of correlation between assignment mark and final.

(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.
Regress exam on c and etc3440.

$$\begin{aligned} H_0 &: \beta_{etc3440} = 0 \\ H_1 &: \beta_{etc3440} \neq 0 \\ t &= \frac{\hat{\beta}_{etc3440}}{se\left(\hat{\beta}_{etc3440}\right)} \sim t_{113} \text{ under } H_0 \\ t_{calc} &= -0.96 \text{ for this sample, } pvalue = 0.34 \end{aligned}$$

We cannot reject the null, and we conclude that there is no evidence to suggest that the expected exam mark of ETC2410 students and ETC3440 students is different.

(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 40% on the assignment, and another 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.



Dependent Variable: EXAM
Method: Least Squares
Date: 04/08/16 Time: 11:39
Sample: 1 118 IF EXAM>0
Included observations: 115

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	71.09253	1.961324	36.24721	0.0000
ASGNMT-80	0.638077	0.080088	7.967197	0.0000

R-squared	0.359687	Mean dependent var	57.96523
Adjusted R-squared	0.354021	S.D. dependent var	14.19578
S.E. of regression	11.40955	Akaike info criterion	7.724017
Sum squared resid	14710.10	Schwarz criterion	7.771755
Log likelihood	-442.1310	Hannan-Quinn criter.	7.743394
F-statistic	63.47622	Durbin-Watson stat	1.240656
Prob(F-statistic)	0.000000		

For asgnmt $= 80$, *prediction is* 71.1, *the prediction interval ignoring estimation uncertainty is* $71.1 \pm 1.98 \times 11.4 = [48.528, 93.672]$, where 1.98 is the 97.5 percentile of t_{113}. *To incorporate estimation uncertainty,* $Var(y - \hat{y} \mid x) = \sigma^2 + Var(\hat{y} \mid x)$. *Therefore,* $\widehat{Var(y - \hat{y} \mid x)} = \hat{\sigma}^2 + \widehat{Var(\hat{y} \mid x)} = 11.4^2 + 1.96^2 = 133.8$ *and its standard error is therefore* $\sqrt{11.4^2 + 1.96^2} = 11.57$, *larger but only slightly relative to* 11.4. *That is why in practice estimation uncertainty is often ignored when sample size is large. The prediction interval based on this is* $71.1 \pm 1.98 \times 11.57 = [48.191, 94.009]$ *only slightly larger than the prediction interval that ignores estimation uncertainty.*

(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.

Dependent Variable: EXAM
Method: Least Squares
Date: 04/07/16 Time: 18:04
Sample: 1 118 IF EXAM>0
Included observations: 115

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	13.72129	5.657110	2.425495	0.0169
ETC3440	21.87665	10.38634	2.106291	0.0374
ASGNMT	0.768005	0.093077	8.251292	0.0000
ASGNMT*ETC3440	-0.420274	0.170347	-2.467161	0.0151

R-squared	0.404915	Mean dependent var	57.96523
Adjusted R-squared	0.388831	S.D. dependent var	14.19578
S.E. of regression	11.09788	Akaike info criterion	7.685548
Sum squared resid	13671.08	Schwarz criterion	7.781024

$$
\begin{aligned}
H_0 &: \beta_{etc3440} = \beta_{asgnmt*etc3440} = 0 \\
H_1 &: \text{at least one of the above is not zero} \\
F &= \frac{(SSR_r - SSR_{ur})/2}{SSR_{ur}/(115-3-1)} \sim F_{2,111} \text{ under } H_0 \\
F_{calc} &= \frac{(14710.10 - 13671.08)/2}{13671.08/(115-3-1)} = 4.2181 \text{ for this sample} \\
& \quad 5\% \; cv \text{ using Eviews is } 3.078
\end{aligned}
$$

We reject the null, and conclude that there is evidence that the conditional expectation of exam mark given assignment mark is different for ETC2410 and ETC3440 students. For ETC2410 students we have $\widehat{exam} = 13.72129 + 0.768005 \times asgnmt$, *and for ETC3440 students we have* $\widehat{exam} = 35.59794 + 0.347731 \times asgnmt$.

(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.

Dependent Variable: EXAM
Method: Least Squares
Date: 04/08/16 Time: 12:55
Sample: 1 118 IF EXAM>0 AND ETC3440=0
Included observations: 67

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	13.72129	5.918047	2.318550	0.0236
ASGNMT	0.768005	0.097370	7.887477	0.0000

R-squared	0.489043	Mean dependent var	59.03932
Adjusted R-squared	0.481182	S.D. dependent var	16.11819
S.E. of regression	11.60977	Akaike info criterion	7.770968
Sum squared resid	8761.143	Schwarz criterion	7.836779

Dependent Variable: EXAM
Method: Least Squares
Date: 04/08/16 Time: 12:55
Sample: 1 118 IF EXAM>0 AND ETC3440=1
Included observations: 48

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	35.59794	8.108920	4.389973	0.0001
ASGNMT	0.347731	0.132817	2.618119	0.0119

R-squared	0.129687	Mean dependent var	56.46599
Adjusted R-squared	0.110767	S.D. dependent var	10.95598
S.E. of regression	10.33139	Akaike info criterion	7.549025
Sum squared resid	4909.934	Schwarz criterion	7.626992

Why the parameter estimates are the same may be difficult for students to understand, so please only provide an intuitive answer and do not try to show or expect them to be able

to understand the reason. Basically the 2410 parameters only affect the sum or squares of residuals of the 2410 group, and similarly the parameters of ETC3440 dummy and its interaction with assignment only enters the sum of squares of residuals of ETC3440 group, and there is no common parameter that enters both. So obviously the minimised sum of squares is achieved when the sum of squares of each group is minimised which can also be achieved with running a separate regression for each group. Doing to sum of squares of residuals, we see $8761.143 + 4909.934 = 13671.077$, *which is the same as the* SSR_{ur}. *The important thing is to understand that* SSR_{ur} *could be computed by running separate regressions for each group and adding the two SSRs.*

(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.
The difference is due to using different $\hat{\sigma}$ *for the two groups when we estimate them separately. Given the evidence that the conditional expectation functions are different, there is no compelling reason to insist that the error variance for the two groups (which is the conditional variance of exam) should be the same. So I would use the standard error of the separate model, although it is really too much ado over nothing since these are very close!*

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.