Introductory Econometrics

Tutorial 8 Solutions

PART A:

1. Assume an OLS regression of a variable y on k regressors collected in $\mathbf{X}$ (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. 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 is:

$$\begin{aligned} price &= \underset{(29.475)}{-21.7703} + \underset{(0.0006)}{0.0021} lotsize + \underset{(0.0132)}{0.1228} sqrft + \underset{(9.0101)}{13.8525} bdrms \\ n &= 88,\ R^2 = 0.672,\ \bar{R}^2 = 0.661,\ \hat{\sigma} = 59.833. \end{aligned}$$

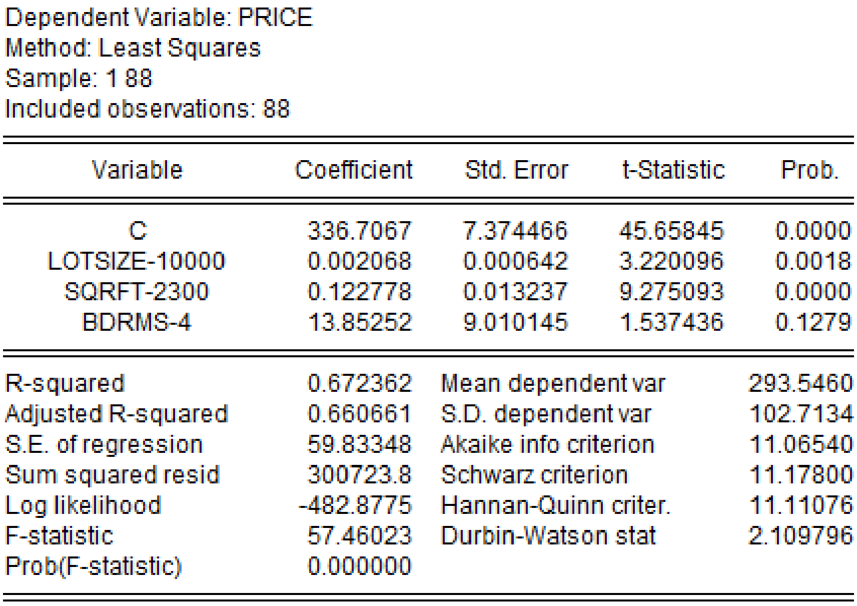
The predicted price at $lotsize = 10,000$, $sqrft = 2,300$, and $bdrms = 4$ is about $-21.7703 + 0.0021 * 10000 + 0.1228 * 2300 + 13.8525 * 4 = 337.08$ thousand dollars

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?

Answer: The regression is $price_i$ on $(lotsize_i - 10,000)$, $(sqrft_i - 2,300)$, and $(bdrms_i - 4)$. In this regression, the estimate of the intercept will be the predicted price of a house with $lotsize = 10,000$, $sqrft = 2,300$, and $bdrms = 4$, and its standard error will be the standard error for $\widehat{price}$:



```
Dependent Variable: PRICE
Method: Least Squares
Sample: 1 88
Included observations: 88

        Variable          Coefficient    Std. Error    t-Statistic    Prob.

           C               336.7067      7.374466      45.65845      0.0000
     LOTSIZE-10000         0.002068      0.000642      3.220096      0.0018
       SQRFT-2300          0.122778      0.013237      9.275093      0.0000
        BDRMS-4            13.85252      9.010145      1.537436      0.1279

R-squared                  0.672362    Mean dependent var        293.5460
Adjusted R-squared         0.660661    S.D. dependent var        102.7134
S.E. of regression         59.83348    Akaike info criterion     11.06540
Sum squared resid          300723.8    Schwarz criterion         11.17800
Log likelihood            -482.8775    Hannan-Quinn criter.      11.11076
F-statistic                57.46023    Durbin-Watson stat        2.109796
Prob(F-statistic)          0.000000
```

Taking 2.000 from t_{60} as a conservative approximation for the 97.5 percentile of t_{84}, the 95% confidence interval for the conditional mean of price is $336.7067 \pm 2.000 \times 7.3745$, which is about 321.96 to 351.46 thousand dollars. Of course the textbook's datasets are US data and are pretty old. No, I would not be 95% confident that the price of a random house with these characteristics will be in this interval. This interval is an interval estimate for the population mean of all houses with these given characteristics.

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?

$$\begin{aligned}
\text{Variance of the prediction error} &= 7.3745^2 + 59.8335^2 = 3634.4 \\
\text{standard error of the sum of } u \text{ and estimation error} &= \sqrt{3634.4} = 60.286 \\
\text{95\% prediction error for price} &= 336.7067 \pm 2.000 \times 60.286 \\
&= [216.13, 457.28] \text{ thousand dollars}
\end{aligned}$$

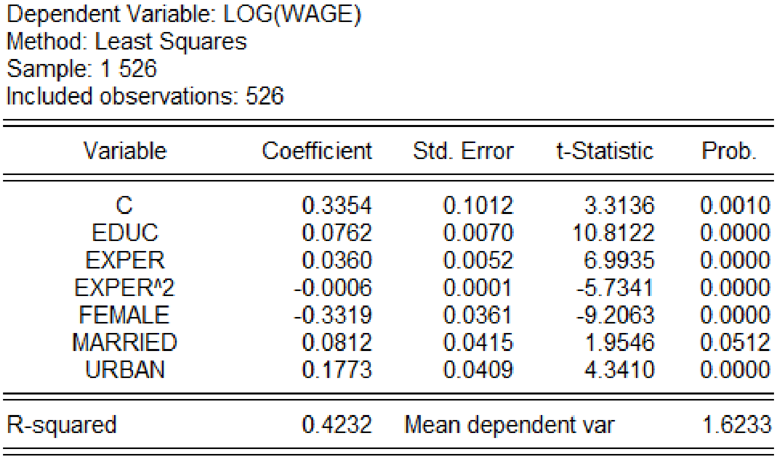
Yes, I will be 95% confident that the price of a random house with these given characteristics will be in this prediction interval. Note that the addition of estimation uncertainty only raised the standard error from 59.8335 to 60.286. That is why in practice the estimation uncertainty is often ignored and the standard error of regression (the estimate of standard error of u) is used for forming prediction intervals.

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 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

(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)
The p-values show that the coefficients of FEMALE and URBAN are statistically significant at the 5% level, but the coefficient of MARRIED is not (well it is borderline).

(b) Interpret each of the estimated parameters.
Since the dependent variable is the logarithm of wage, and the coefficients of two of the dummy variables are larger than 0.10 in absolute value, we use the $100\left(e^{\hat{\beta}}-1\right)$ transformation to explain what they tell us:

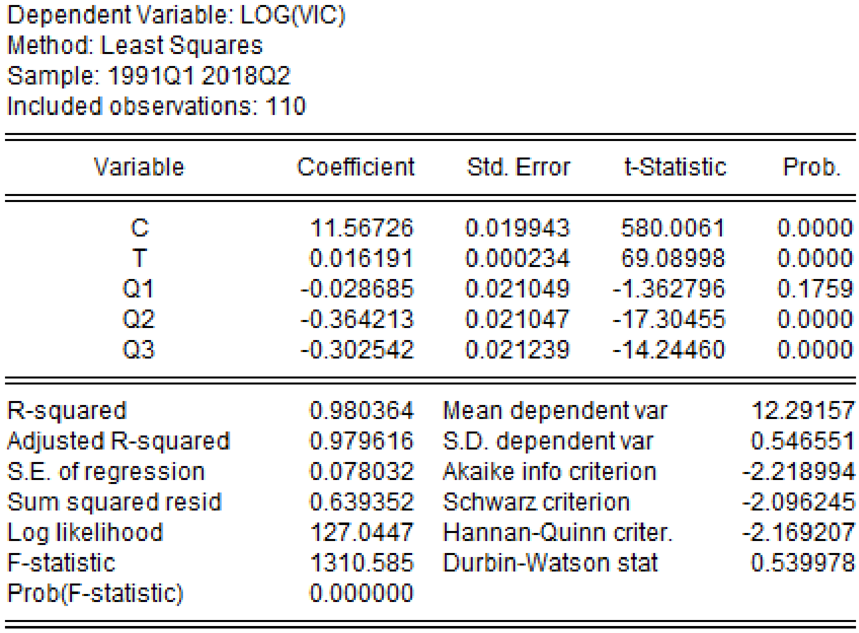
- The estimated coefficient of FEMALE tells us that for two people of opposite sexes with the same education, experience, marital status and area of residence, the female person's wage is predicted to be 28% less than the male person's wage. This is because $100\left(e^{-0.3319}-1\right)=-28.244$
- The estimated coefficient of MARRIED tells us that for two people of the same sex, education, experience and area of residence, the married person's wage is predicted to be 8% higher than the wage of the single person. However, this is not precisely estimated and is not statistically significant at the 5% level. Here, if we used $100\left(e^{0.0812}-1\right)=8.4588$, we would still get 8%.

- The estimated coefficient of URBAN tells us that for two people of the same sex, education, experience and marital status, the person who lives in an urban area is predicted to earn 19% more than the person who lives in a rural area. This is because $100\left(e^{0.1773}-1\right)=19.399$
- The estimated coefficient of EDUC tells us that given years of experience, gender, marital status and place of residence, an extra year of education increases the predicted wage by approximately 8% (we get the same if we use $100\left(e^{0.0762}-1\right)=7.9178$, which is 8%)
- The coefficients of EXPER and EXPER^2 each on their own do not have an interpretation, but together they show how wage changes with experience.

  $$\frac{\partial \log(WAGE)}{\partial EXPER}=\frac{1}{WAGE}\frac{\partial WAGE}{\partial EXPER}=0.036-2\times 0.0006\ EXPER$$

  This shows that given years of education, gender, marital status and place of residence, one year of experience starts by increasing wage by approximately 4% (this is given by 0.036) and the effect of each additional year decreases with experience and eventually becomes negative after $\frac{0.036}{2\times 0.0006}=30.0$ years of experience. So, WAGE given gender, marital status and place of residence reaches its maximum when EXPER=30.

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 multicollinearity 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.

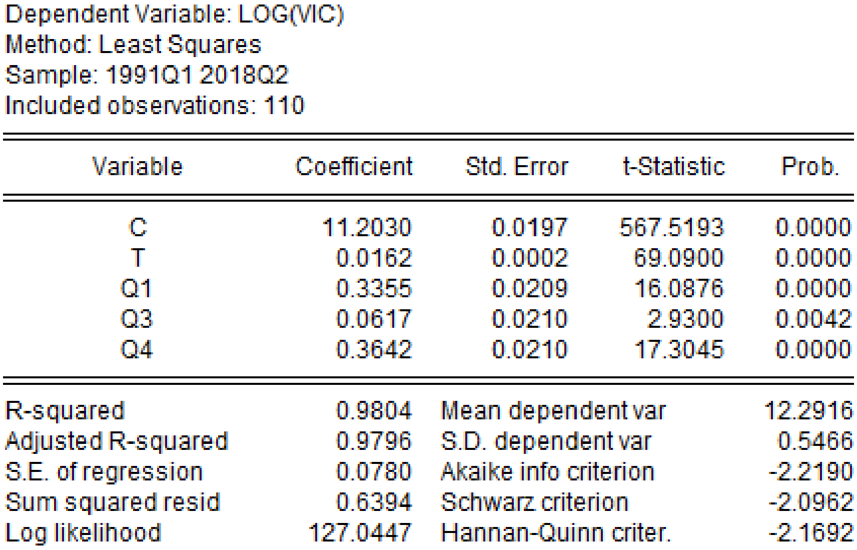
(b) How would the estimation results (in particular coefficients of each regressor, the R^2, SSR and standard error of the regression) change if we dropped Q1 and added Q4 instead? After answering this question using a calculator, check your calculations by running the regression using the victouristquarterly.wf1 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.]
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.

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

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

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.

(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.
Here, you need to remember that when we have m categories (here $m = 4$ for our 4 quarters), when we take one category as the base category (i.e. the one quarter than we do not have a dummy variable in the regression), the coefficients of the $m-1$ dummies will be the difference between the constant in that category and the constant in the base category. This makes it clear that to answer this question we need to estimate the regression with Q2 or Q3 as the base category:



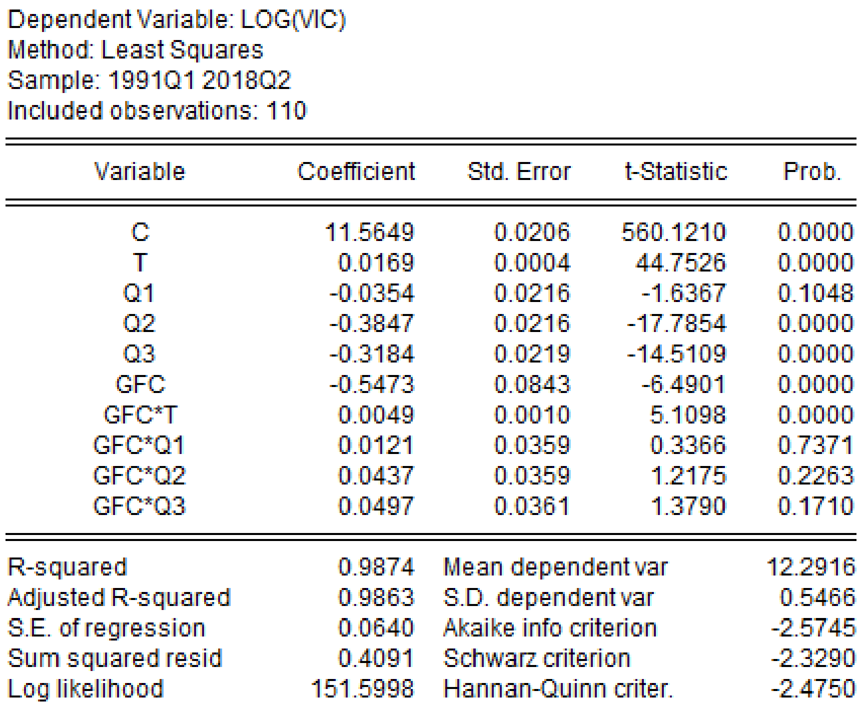
```
Dependent Variable: LOG(VIC)
Method: Least Squares
Sample: 1991Q1 2018Q2
Included observations: 110
```

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	11.2030	0.0197	567.5193	0.0000
T	0.0162	0.0002	69.0900	0.0000
Q1	0.3355	0.0209	16.0876	0.0000
Q3	0.0617	0.0210	2.9300	0.0042
Q4	0.3642	0.0210	17.3045	0.0000

R-squared	0.9804	Mean dependent var	12.2916
Adjusted R-squared	0.9796	S.D. dependent var	0.5466
S.E. of regression	0.0780	Akaike info criterion	-2.2190
Sum squared resid	0.6394	Schwarz criterion	-2.0962
Log likelihood	127.0447	Hannan-Quinn criter.	-2.1692

$$
\begin{aligned}
H_0 &: \beta_{Q3} = 0 \\
H_1 &: \beta_{Q3} \neq 0 \\
t_{\hat{\beta}_{Q3}} &= \frac{\hat{\beta}_{Q3}}{se\left(\hat{\beta}_{Q3}\right)} \sim t_{110-5} \text{ under } H_0 \\
t_{calc} &= 2.9300 \\
t_{crit} &= 1.983 \text{ using } @qtdist(0.975, 105) \text{ in EViews} \\
t_{calc} &> t_{crit} \Rightarrow \text{we reject the null} \\
\text{Conclusion} &: \text{the trend lines for Q2 and Q3 are different.}
\end{aligned}
$$

(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).



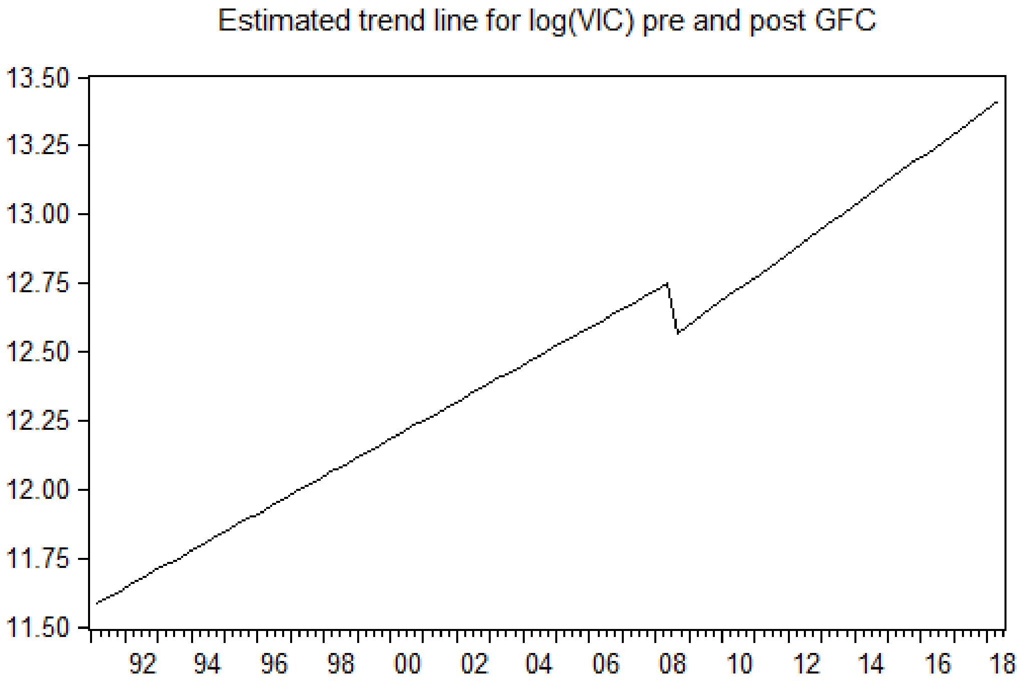
Dependent Variable: LOG(VIC)
Method: Least Squares
Sample: 1991Q1 2018Q2
Included observations: 110

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	11.5649	0.0206	560.1210	0.0000
T	0.0169	0.0004	44.7526	0.0000
Q1	-0.0354	0.0216	-1.6367	0.1048
Q2	-0.3847	0.0216	-17.7854	0.0000
Q3	-0.3184	0.0219	-14.5109	0.0000
GFC	-0.5473	0.0843	-6.4901	0.0000
GFC*T	0.0049	0.0010	5.1098	0.0000
GFC*Q1	0.0121	0.0359	0.3366	0.7371
GFC*Q2	0.0437	0.0359	1.2175	0.2263
GFC*Q3	0.0497	0.0361	1.3790	0.1710

R-squared	0.9874	Mean dependent var	12.2916
Adjusted R-squared	0.9863	S.D. dependent var	0.5466
S.E. of regression	0.0640	Akaike info criterion	-2.5745
Sum squared resid	0.4091	Schwarz criterion	-2.3290
Log likelihood	151.5998	Hannan-Quinn criter.	-2.4750

$$\begin{aligned}
H_0 &: \beta_{GFC} = \beta_{GFC*T} = \beta_{GFC*Q1} = \beta_{GFC*Q2} = \beta_{GFC*Q3} = 0 \\
H_1 &: \text{ at least one of the above is not zero} \\
F &= \frac{(SSR_r - SSR_{ur})/5}{SSR_{ur}/(105-10)} \sim F_{5,95} \text{ under } H_0 \\
F_{calc} &= \frac{(0.6394 - 0.4091)/5}{0.4091/95} = 10.696 \text{ (differences due to rounding don't matter)} \\
F_{crit} &= 2.310 \text{ (using } @qfdist(0.95,5,95) \text{ in EViews)} \\
F_{calc} &> F_{crit} \Rightarrow \text{ we reject the null} \\
\text{Conclusion} &: \text{data supports that there has been a structural break in Victorian tourism post GFC}
\end{aligned}$$

The results of the unrestricted regression reveals no significant change in the seasonal patterns, but a drop in the constant term and an increase in the trend. This means that GFC made a sudden drop in tourist numbers, but the tourist numbers caught up due to an increase in the trend.



The line is the plot of a new series $trendline = 11.5649 + 0.0169 * t - 0.5473 * gfc + 0.0049 * gfc * t$ generated using the estimated unrestricted model and suppressing the seasonal factors (to be able to see the trend clearly) in EViews.