Tutorial 8

Hong Xiang Yue

25/04/2019

Part A

Question 1a

We have

$$R^2 = 1 - \frac{SSR}{SST}$$

and

$$\bar{R}^2 = 1 - \frac{SSR/(n-k-1)}{SST/(n-1)}$$

.

Prove that

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

Question 1c

Problems with R^2

- Adding more variables will never decrease R^2, regardless of how rubbish they are
- You can have a model which is 99% junk but still appear to have a good fit
- This is problematic when you want to choose the most important variables for your model
- The model with the most variables will always have the highest R^2

Advantages of $\bar{R}^2$

- A penalty is added for each variable included in the model
- $\bar{R}^2$ will only increase if the new variables adds more explanatory power than the penalty

Question 2

The estimated equation should be

$$\textit{price} = \underset{(29.475)}{-21.7703} + \underset{(0.0006)}{0.0021} \times \textit{lotsize} + \underset{(0.0132)}{0.1228} \times \textit{sqrft} + \underset{(9.0101)}{13.852} \textit{bdrms}$$

$$n = 88, R^2 = 0.672, \bar{R}^2 = 0.661, \hat{\sigma} = 59.833$$

Question 2a

The predicted price (in thousands of dollars) is
$\neg 21.7703 + 0.0021 \times 10000 + 0.1228 \times 2300 + 13.8525 \times 4 = 337.08$

- Note, this is the same as estimating the conditional mean of a house with these characteristics
- The only difference between $E[price|lotsize = 10000, sqrft = 2300, bdrms = 4]$ and $y_i|lotsize = 10000, sqrft = 2300, bdrms = 4$ is the error term u_i
- Since this cannot be predicted, our point estimate will be the same for both
- Things change when we have an interval estimate

The linear regression model

$$y = \beta_0 + \beta_1 x + u$$

- Two components:
- Deterministic component $E(y|x) = \beta_0 + \beta_1 x$, aka conditional mean
- Random errors: u where $u \sim N(0, \sigma^2)$ (intrinsic variability)
- Running a regression gives us $\hat{y}$ which estimates $E(y|x)$
- $\hat{y}$ also estimates y

Confidence interval for the conditional mean

- We need some sort of measure of how dispersed our regression lines will be around the true conditional mean
- This is different from $\hat{\sigma}^2$ which we use to estimate σ^2, the variance of the errors
- What we want is the variance of the estimation error or $Var(\hat{y})$
- I.e. how much will $\hat{y}$ move around if we took different samples
- Also called the variance of the sampling error

Question 2b

To calculate the confidence interval for $E[price|lotsize = 10000, sqrft = 2300, bdrms = 4]$ we need to run a regression of $price_i$ on $(lotsize_i - 10000)$,$(sqrft_i - 2300)$ and $(bdrms_i - 4)$.

- The point estimate for the conditional mean will be given by the intercept of the regression
- We can also use the standard error on the intercept to get our confidence interval

Question 2b

CI for $E[price|lotsize = 10000, sqrft = 2300, bdrms = 4]$

$$= [337.08 \pm t_{60}(0.975) \times 7.374466] = [321.96, 351.46]$$

- We are 95% confident that the conditional mean for a house with these characteristics would lie in this interval
- Would we be 95% sure that the price of a house with these characteristics would lie in this interval? NO
- This only accounts for *estimation* uncertainty and not uncertainty due to the error term u_i

Question 2cv

Now we calculate a prediction interval for a house with these characteristics.

- The point estimate will be the same, but the standard error of the prediction will have two components: estimation uncertainty and ‘intrinsic’ uncertainty
- There is estimation uncertainty because we don’t know what β is, only what $\hat{\beta}$ is
- There is intrinsic uncertainty, because even if we did know the true values of the βs, each observation has a random component u
- The standard error of prediction $var(\hat{e}_i) = \hat{\sigma}^2 + [se(\hat{y}_i)]^2$
- Most of the time estimation uncertainty is small, so we can ignore $se(\hat{y}_i)$

Part B

Question 1a

We have a data set on the characteristics of individuals in the labour force and estimated a model of the form

$$\ln(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2$$
$$+\beta_3 female + \beta_4 married + \beta_5 urban + u$$

where female, married and urban are dummy variables.

Question 1 Output

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

Figure 1:

- Most of these variables are statistically significant at the 5% level because their p-values are close to zero

Question 1b

How do we interpret the estimated coefficients of the model?

$$ln(y) = \beta_0 + \beta_1 x + u$$

- For a one unit change in x $\%\Delta y \approx \beta_1 \times 100$, approximation only valid if $|\beta_1| < 0.1$
- The exact expression is $\%\Delta y = (e^{\beta_1} - 1) \times 100\%$

Question 2

Open up the victouristquarterly.wf1 file in EViews and estimate a model of the form:

$$\ln(VIC) = \beta_0 + \beta_1 T + \beta_2 Q1 + \beta_3 Q2 + \beta_4 Q3 + u$$

- T is the deterministic time trend
- Q1, Q2 and Q3 are dummy variables for each quarter of the year

Why do we not include a dummy variable for Q4? What would happen if we did?

https://flux.qa/LDPPHD

Question 2b

If we dropped the dummy variable Q1 and included a dummy variable for Q4, how would that affect:

- the coefficient estimates?
- R^2 of the model?
- SSR?

Question 2c

Plot (by hand) the predictions of $\ln(VIC)$ across time for each of the quarters.

- These lines should all be parallel but with different intercepts

Now, generate a new series by taking the natural logarithm of the variable VIC and create a seasonal plot

- Go to view and select graph
- In the graph options click on Seasonal Graph and press Enter

Question 2d

Use a t-test to test the hypothesis that the true intercepts for Q2 and Q3 are equal vs the alternative they are not with $\alpha = 0.05$.

- What model should we estimate to test this hypothesis?
- https://flux.qa/LDPPHD

Question 2e

At the 5% significance level, test whether or not there exists a structural break in the intercepts for each quarter and the time trend due to the 2008 global financial crisis.

- To create the dummy variable, go to generate series and type “@after(”2008Q3“)”
- Run a regression with the gfc dummy variable now included and 4 extra variables $GFC * T$, $GFC * Q1$, $GFC * Q2$ and $GFC * Q3$
- What is the null hypothesis that we are testing?