Quantitative Techniques:

Laboratory 9

24 November 2005

Linear regression

Overview

In this lab your tasks are:

1. to draw inferences on the estimated parameters;
2. to obtain a measure of goodness of fit;
3.

On the Excel menu bar, click on Tools, Add-ins…., then make sure that Analysis ToolPak is ticked. Download *DatasetProblemSet9.xls* from http://www.staff.city.ac.uk/a.banal-estanol/ and save it in your directory.

As last week, we have a true model:

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, *i= 1, 2… n*

Under the OLS assumptions, using our sample we can calculate

$$\hat{\beta}_1 = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2}, \ \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

Task 1: Parameter estimates

Use the data from Exercise 2, Chapter 10 of Ashenfelter et al.

1. In Excel, choose 'Tools, Data Analysis' and then 'Regression';
2. A window will pop-up:
   a. Select the cells corresponding to Input Y range (B5:B14) and Input X range (C5:C14);[1]
   b. Click on 'New Worksheet Ply' and 'Residuals';
3. The output will be displayed in a new worksheet and is reproduced at the end of this handout;
4. The estimated coefficients coincide with those obtained by hand using the formula above;
5. Now focus on the slope coefficient (β1 hat):

[1] Note that Excel can also be used for multivariate regressions. In this case, you will need to specify the Input X range by selecting the cells where your explanatory variables are (in columns).

a. The standard error of β1 hat is given by $s.e.(\hat{\beta}_1) = \sqrt{\dfrac{s_Y^2}{\sum (X_i - \overline{X})}}$ = 0.071

You can estimate s_Y^2 by $\dfrac{\sum (Y_i - \hat{Y})^2}{n-2}$ (i.e. the sum of squared residuals divided it by *n-2*).

b. The *t-statistic* is given by $t = \dfrac{\hat{\beta}_1 - 0}{s.e.(\hat{\beta}_1)}$. A large t-statistic in absolute value leads to the rejection of the null hypothesis that the parameter is zero. Therefore we can conclude that the slope is significantly different from zero.

c. The P-value of the observed *t Stat* provided by Excel has to be compared with α, the significance level of the test that the parameter is zero. In this case, the p-value is less than 10%, 5% and 1%. Therefore, we can reject the hypothesis that the coefficient is zero.

d. The confidence interval (in the table below, it is calculated for α=0.05) is given by:

$$P(\hat{\beta}_1 - t_{(n-2,1-\alpha/2)} s_{\hat{\beta}_1} \le \beta_1 \le \hat{\beta}_1 + t_{(n-2,1-\alpha/2)} s_{\hat{\beta}_1}) = 1 - \alpha$$

If zero is outside the interval, as in our example (see below), then we can conclude that the coefficient is significantly different from zero. Note that the conclusion is consistent with the results of test above.

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	5.145	0.648	7.942	0.000	3.651	6.638
X Variable 1	-0.506	0.071	-7.169	0.000	-0.669	-0.343

Task 2: Goodness of fit – R^2 and F

Now focus on the following table from the Excel output:

	df	*SS*
Regression	1	0.118
Residual	8	0.018
Total	9	0.137

1. Verify that you obtain the same results as Excel, by calculating from your data:
   a. the regression sum of squares (or explained sum of squares);
   b. the residuals sum of squares (or unexplained sum of squares);
   c. the total sum of squares.
2. Reminder from lecture:

$$\underbrace{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}_{\substack{\text{Total sum} \\ \text{of squares} \\ \text{(TSS)}}} = \underbrace{\sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2}_{\substack{\text{Explained sum} \\ \text{of squares} \\ \text{(ESS)}}} + \underbrace{\sum_{i=1}^{n} (Y_i - \hat{Y})^2}_{\substack{\text{Unexplained sum} \\ \text{of squares} \\ \text{(USS)}}}$$

3. Using the definition of R-square as the ratio between the explained sum of squares and the total sum of squares, verify your results;

4. Verify that the value you obtained before for s_Y^2 is correct (see table below). Note that Var(β1) is smaller if the variance of Y is smaller. The smaller the variance of Y, the less likely we are to observe extreme samples.

Regression Statistics		
Multiple R	0.930	*Correlation between X and Y in absolute value*
R Square	0.865	*R-sq. = ESS/TSS*
Adjusted R Square	0.848	*R-sq. adjusted for degrees of freedom*
Standard Error	0.048	s_Y^2
Observations	10	*n*

5. Excel also reports the F test, i.e. a test of the joint significance of the regression coefficients. In simple regression, there is only one explanatory variable and only one coefficient to test. Therefore, for simple regression the F test for overall significance is testing the same thing as the t test. Large values of the F test (and small P-values associated to it) lead to conclude that at least one of the independent variables has significant predictability for the dependent variable.

Task 3: Fitted values and residuals

We obtain the fitted values for Y_i as $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ while the residuals are

$e_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 + \hat{\beta}_1 X_i$.

Use the table below from the Excel output to verify your results.

RESIDUAL OUTPUT

Observation	*Predicted Y*	*Residuals*
1	0.487	0.053
2	0.639	0.031
3	0.386	-0.026
4	0.487	0.063
5	0.689	-0.079
6	0.386	0.004
7	0.487	-0.017
8	0.335	-0.045
9	0.588	0.032
10	0.537	-0.017

Excel output for Exercise 2, Chapter 10 of Ashenfelter et al.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.930
R Square	0.865
Adjusted R Square	0.848
Standard Error	0.048
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.118	0.118	51.401	0.000
Residual	8	0.018	0.002		
Total	9	0.137			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	5.145	0.648	7.942	0.000	3.651	6.638
X Variable 1	-0.506	0.071	-7.169	0.000	-0.669	-0.343

RESIDUAL OUTPUT

Observation	*Predicted Y*	*Residuals*
1	0.487	0.053
2	0.639	0.031
3	0.386	-0.026
4	0.487	0.063
5	0.689	-0.079
6	0.386	0.004
7	0.487	-0.017
8	0.335	-0.045
9	0.588	0.032
10	0.537	-0.017