

Problem Set 2. Properties of Least Squares Estimates

The idea behind these questions is to help you become more familiar with the concepts, formula and intuition that underlie the issues of goodness, of fit, bias and efficiency

1. Show that the R^2 (coefficient of determination) can be written as the square of the correlation coefficient between the actual and predicted value of y .

- *Hint: see lecture notes on Properties of Least Squares*

2. Given $y = a + bx + u$, show that the least squares estimate of the intercept gives an unbiased estimate of the true intercept.

3. Given the following set of estimates taken from 10 random samples of a population using 2 alternative estimation methods, work out if the estimates are unbiased estimates of the true population value 6

Estimate	1	2	3	4	5	6	7	8	9	10
$\hat{\beta}$	1	4	4	6	10	10	9	8	6	5
\tilde{b}	1	2	6	8	13	9	6	10	3	2

What can you say about the (sample) variance of the two estimators?

4. Consider an alternative estimator of the slope in $y = a + bx + u$, given by

$$\tilde{b} = \frac{y_2 - y_1}{x_2 - x_1}$$

where (x_1, y_1) is the pair of values from the first observation and (x_2, y_2) is the pair of values from the 2nd observation.

Sketch the fitted regression line implied by this estimator.

Find the expected value of this estimator

How might you decide whether to use this estimator of that derived from least squares?

Turn over

5. Given the following regression output

$$\widehat{Consumption} = 5000 + 0.90 * Income$$

$R^2=0.89$ TSS = 1000 RSS = 102 Var(Income)=1

estimated over the period 1990-2001

i) What factors influence the precision of the OLS estimate of the slope and why?

$$Var(\hat{\beta}) = \frac{s_u^2}{NVar(X)}$$

ii) Hence find the standard error on the estimate of income in model I

6. Read in the data set on QM2 marks and seminar attendance, (*marks.dta*), from the course website.

Run a regression of mark on attendance for

- a) the 1st 10 observations in the sample
- b) the 1st 30 observations in the sample
- c) the 1st 100
- d) the full sample

What do you find happens to

- i) the estimated coefficient on attendance
- ii) the standard error of this estimate

Why?

(Hint to do a regression on a sub-sample of data use the command regress y x if _n<=a

where y is the name of the dependent variable, x is the name of the explanatory variable and a is the number of observations you wish to run the regression on)