

## Problem Set 5: Multiple Regression Analysis

1. (Final exam question 2003)

You estimate the following regression model by OLS:

$$\hat{HWAGE} = 5.00 + 0.5YEARS - 0.01YEARS^2 - 3.00FEMALE - 2.5PARTIME$$

(2.00)
(0.10)
(0.002)
(2.00)
(3.25)

N = 125      TSS=1000      RSS = 200

where the numbers in brackets are estimated standard errors

- i) Interpret the regression output.
- ii) Hourly wages reach a maximum at how many years of work experience?
- iii) Test the hypothesis that the female dummy variable has no explanatory power in the model at the 95% level. Test the hypothesis that the true value of the female coefficient is -4. How do you reconcile the two results?
- iv) test the hypothesis that regression as a whole has no explanatory power.

2. Which of the following models do you prefer and why?

$$\hat{Profits} = 55.00 + 0.6Sales - 3.00employment$$

(25.00)
(0.10)
(2.00)

$$R^2=0.15 \quad \bar{R}^2 = 0.14$$

$$\hat{Profits} = 50.00 + 0.4Sales - 0.02Sales^2 - 3.15employment$$

(20.00)
(0.02)
(0.03)
(1.50)

$$R^2=0.16 \quad \bar{R}^2 = 0.13$$

$$\hat{Profits} = 60.00 + 1.6\log(Sales) - 2.70employment$$

(10.00)
(0.60)
(1.40)

$$R^2=0.15 \quad \bar{R}^2 = 0.15$$

3. (mid-Term test 2003)

You have time series data for the period 1935-2000. You are given an estimate of the effects of income, (measured in £billion) and interest rates, (measured in percentage points) on aggregate consumption expenditure (measured in £billion).

$$\hat{C}_{cons} = 10.00 + 0.90\text{Income} - 6.00\text{Interest Rate} \quad \text{TSS} = 70$$

$$(1.00) \quad (0.45) \quad (2.00) \quad \text{ESS} = 10$$

You then split the data and run 2 regressions, one for the period 1935-1970 and one for the period, 1971-2000

1935-70

$$\hat{C}_{cons} = 6.00 + 0.95\text{Income} - 2.00\text{Interest Rate} \quad \text{TSS} = 30$$

$$(1.00) \quad (0.40) \quad (1.00) \quad \text{ESS} = 10$$

1971-2000

$$\hat{C}_{cons} = 14.00 + 0.80\text{Income} - 10.00\text{Interest Rate} \quad \text{TSS} = 20$$

$$(1.00) \quad (0.50) \quad (4.00) \quad \text{ESS} = 10$$

Test the hypothesis that the data could be pooled across both time periods and estimated as a single equation

4. Given the following information, work out the estimate of the standard error of on income

$$\hat{C}_{cons} = 10.00 + 0.80\text{Income} \quad \text{N}=100$$

$$(1.00) \quad ( ) \quad \text{TSS}= 1000 \quad \text{ESS} = 902$$

$$\text{Var}(\text{Income})= 1$$

Now given the following information

	income	income <sub>t-1</sub>	interest
income	1.0000		
income <sub>t-1</sub>	0.9000	1.0000	
interest	0.1000	0.0500	1.0000

$$\hat{income} = 1.00 + 0.80\text{Income}_{t-1} - 0.50\text{Interest Rate} \quad \text{TSS}= 100$$

$$(0.50) \quad (0.20) \quad (0.35) \quad \text{ESS} = 80$$

Work out what would happen if you add the following variables

- income<sub>t-1</sub>
- interest
- both

5. Read in the data set ps4data.dta (from the same course website as the other data sets). This file contains information, among other things, on the hourly pay of 12,098 British employees, (hourpay).

Create a log of hourly pay variable, and a years of education variable (using the commands: gen <var. name> = log(hourpay)  
gen yrsed=edage-5 )

- a) Estimate a simple regression of the effect of age on log of hourly pay.
- b) Estimate a multiple regression of the impact of age and years of education on the log of hourly pay  
Interpret your results. What has happened to the coefficient on age? How do you account for this? What does it say about the correlation between age and years of education.

Now run separate regressions for those in full-time jobs and those in part-time jobs.

(use the commands  
reg ... if parttime==1

for part-time jobs (the double == sign is not a misprint)

and

reg ... if parttime==0

for full-time jobs )

- b) Explain why the standard errors on the estimate of age are different in the two regressions.
- c) Test the hypothesis that there is no difference between the estimated coefficients for full-timers and part-timers.
- d) Now introduce union status and gender as additional explanatory variables, (they are called union and female in the data set). Test the hypothesis that these variables are jointly significantly different from zero in the part-time regression.
- e) Now add the number of children, (numkids) and estimate the regression for full-time workers.  
Next add the age of youngest child, (agyung) estimate and compare your results with the previous regression.

What do you find? Why? How might you detect this problem before you did the regression? What can you do about it?

**Turn over**

6. Read in the file `miles.dta`. The file contains data on average annual number of miles travelled by car (*miles*) over a 23 year period in the United States and a list of potential explanatory variables, (*ncars* – number of cars (millions); *mpg* – average miles per gallon per car, *price* – petrol price index (year 1=1, year 2= 1.039 ie 3.9% increase in price over the year); *pripub* – public transport price index; *popsize* – population (millions); *nworkers* – employees (millions); *gdpcap* – per capita gdp (\$ 000).

Estimate the number of miles travelled as a function of *mpg*, *price*, *pripub* and *gdpcap* for the first 21 observations in your data set (use the regression command

```
reg miles mpg, price pripub gdpcap if year<22 )
```

Interpret your results

Now estimate the same regression over the full sample of 23 observations. What do you see? Why?