Problem Set 6. Specification Analysis and Use of Dummy Variables

1. Read in the data set ps4data.dta (from the course web site). Remember that this file contains information, among other things, on the hourly pay of 12,098 British employees, (hourpay).

Generate the log of hourly pay as in the previous problem set and regress the log of hourly pay on the part-time work dummy variable, (*parttime*), *age*, age squared (you will have to create this variable using the command

gen *variable_name* = age^2

Interpret your coefficients and the significance of each variable.

At what age are earnings maximised?

You think that part-time work may not be rewarded at the same rate as full-time work as individuals get older. Test your hypothesis by introducing a slope dummy variable into the regression above. Interpret your coefficients.

Now add 4 dummy variables which group individuals according to their highest educational level, (*postgrad, grad, highint, low*). There is a fifth category, (*none*). What happens when you include this variable? Why?

Test the hypothesis that education matters (ie a joint test of significance of the education variables)

Now change the reference category in your regression by including the variable *none* and dropping the variable *postgrad*
Interpret your results.

Now generate a male dummy variable (1 if male, 0 if female)
Regress log of hourly earnings on this dummy variable.
Now add length of time (in months) at the firm (the variable *tenure*). Compare your results. Why does the coefficient for men change?

Now add the variables age, agesquared, *numkids* and *single*. What happens to the estimated standard errors. What happens to the R^2 and the adjusted R^2 ? Why?

Now remove the variable *single* and re-run your regression.
What happens to the R^2 and the adjusted R^2 ? Why?

Test your specification using the Ramsey RESET test.

2. Read in the data set *incin.dta* from the web site. The data set contains information on house prices over 2 years in 2 areas. In the first area a waste incinerator was installed between the 2 dates. In the other area nothing happened. The idea is to use dummy variables (and their interactions) to try and assess the effect of the incinerator on house prices in the area.

Regress the house price variable on the dummy for the area with the incinerator for year==2 only.

$$\text{Houseprice} = b_0 + b_2\text{Incinerator} \qquad \text{if year==2}$$

Interpret the coefficient.

You suspect that this may not actually give you the right answer.

Do the same regression, but for year==1 data only.

$$\text{Houseprice} = b_0 + b_1\text{Year2} + b_2\text{Incinerator} \qquad \text{if year==1}$$

What do you find?

To obtain the "correct" effect regress house prices on a dummy variable for year 2, the incinerator dummy and an interaction of the year and incinerator variables.

(you will have to create this last variable)

$$\text{Houseprice} = b_0 + b_1\text{Year2} + b_2\text{Incinerator} + b_3\text{Year2*Incinerator} + e$$

The coefficient on the interaction term gives the change over the period in average house prices in the 2 areas. What do you find?

3. Given of a sample of employees, you have data on the number of years of work experience, (YEARS), its square, (YEARS2), and a dummy variable to indicate whether the employee was female or not (FEMALE), together with information on the log of hourly pay measured in pounds, (LHWAGE).

You estimate the following regressions:

(1) $\widehat{\text{LHWAGE}} = 5.00 + 0.05\text{YEARS} - 0.001\text{YEARS2} - 0.30\text{FEMALE}$
$\qquad\qquad\qquad\quad (2.00) \quad (0.01) \qquad (0.002) \qquad (0.10)$

TSS=10000 ESS=4000 N=124

(2) $\widehat{\text{LHWAGE}} = 6.00 - 0.25\text{FEMALE}$
$\qquad\qquad\qquad\quad (3.0) \quad (0.10)$

TSS=10000 ESS=3000 N=1000

where the numbers in brackets are estimated standard errors

i) Interpret the estimated coefficients.
ii) At what level of work experience is pay maximised?
iii) Test the hypothesis that years of experience (and its square) have no explanatory power in the model at the 95% level.

4. Given the following information

$Cons = 500 + 0.9\hat{Income} + 0.3Assets$ for the period 1940-2003 RSS=700

$Cons = 400 + 0.8\hat{Income} + 0.2Assets$ for the period 1940-1979 RSS=350

$Cons = 600 + 0.95\hat{Income} + 0.35Assets$ for the period 1980-2003
RSS=250

Test the hypothesis that the coefficients are the same across the 2 sub-periods

5. Given the following information from a "quasi-natural" experiment to examine the effect of the introduction of the minimum wage on the gender pay gap, interpret the effect of each of the coefficients

$$\widehat{Ln(WAGE)} = \underset{(1.00)}{2.00} + \underset{(0.01)}{0.05}YEAR2 - \underset{(0.002)}{0.250}FEMALE + \underset{(0.02)}{0.03}FEMALE*YEAR2$$

The data are pooled over 2 periods and N=500

YEAR2 =1 if data come from the 2nd period
= 0 otherwise
FEMALE = 1 if individual is female
= 0 if male
FEMALE*YEAR2 is an interaction of the 2 dummy variables

6. In April 2000 the UK government introduced the Working Families Tax Credit aimed at increasing the income in work relative to out of work for groups of traditionally low paid individuals with children. In addition financial help was also given toward child care.

If successful the scheme could have been expected to increase the hours worked of those who benefited most from the scheme- namely single parents. By comparing hours of worked for this group before and after the change with a suitable control group, it should be possible to obtain a difference in difference estimate of the policy effect.

The data set lonep.dta has data on the number of hours worked in the years 1998 and 2000 by female lone parents and by single childless women used as a control group.

Find the difference in difference estimate of the policy on hours worked.

7. Using the framework of question 5, how would you set up an evaluation of the effect of the central London congestion charge on traffic volume?