Dummy variables

- Dummy variables are variables taking the value 1 or 0, to represent a particular observation either having or not having a particular property.

- Examples include:
  - A dummy for gender – e.g. 1 for female, 0 for male.
  - A dummy for years in which there was some unusual circumstance, e.g. war – would equal 1 in war years, 0 otherwise.
  - A set of seasonal dummies – e.g. dummies for the four quarters of the year, each equal to 1 in its own quarter, 0 otherwise.
  - A set of category dummies – e.g. for different industries – Manufacturing dummy = 1 for manufacturing firms, 0 otherwise, retail dummy =1 for retail firms, 0 otherwise, etc.

- A model involving a dummy variable or dummy variables essentially allows different intercept terms for the different categories represented by the variables.

- E.g. an individual earnings equation might take the form

  $$\text{Earnings} = \beta_0 + \beta_1\text{Education} + \beta_2\text{Age} + \beta_3\text{Age}^2 + \beta_4\text{Gender} + u$$

  Where Gender = 1 for females and 0 for males. We can break this down into two separate equations:

  $$\text{Males: Earnings} = \beta_0 + \beta_1\text{Education} + \beta_2\text{Age} + \beta_3\text{Age}^2 + u$$
  $$\text{Females: Earnings} = (\beta_0 + \beta_4) + \beta_1\text{Education} + \beta_2\text{Age} + \beta_3\text{Age}^2 + u$$

- Thus, the slope coefficients on the variables are the same for the two groups, but the intercept term is $(\beta_0 + \beta_4)$ for women instead of just β_0 for men. (Would generally expect β_4 to be negative in this case.)

- If we were to estimate a Cobb-Douglas production function across a range of firms,

  $$\text{Ln}(Q) = \beta_0 + \beta_1\text{Ln}(L) + \beta_2\text{Ln}(K) + u$$

  Where Q is output, L is labour and K is capital, we could add dummies for different industries, e.g. manufacturing, retail, communications etc. These would allow for a different β_0 term for each industry.

- Similarly, seasonal dummy would capture the effect of seasonal differences.

- When using seasonal or category dummies, must be careful to avoid dummy variable trap. E.g. if you had a dummy for each quarter, say variables Q1, Q2, Q3 and Q4, which were 1 in their corresponding quarter and 0 otherwise, then you would have Q1+Q2+Q3+Q4=1, which would give perfect multicollinearity in a model with a constant term. Therefore, must leave out one category, which then becomes the baseline.

- A dummy covering a single observation – e.g. an anomalous year – effectively takes that observation out of the sample from the point of view of calculating the coefficient estimates. In minimising the RSS, can choose coefficients of the other variables that minimise the RSS for the other observations, then choose coefficient on dummy to give error 0 on that observation.

- However, coefficient and significance of dummy variable will tell if that observation is behaving significantly differently from the rest, and by how much.

- Should be very careful about using dummies to account for apparent outliers – should look for an explanation that allows them to be kept in the model first. But sometimes no better option.

Slope dummies

- Dummy variables allow the intercept to vary between different groups of observations. But sometimes we might want to allow the slope coefficients on other variables to vary between groups.

- We can do this by creating a slope dummy, equal to a dummy variable times another variable.

- E.g. suppose in our earnings model,

  $\text{Earnings} = \beta_0 + \beta_1\text{Education} + \beta_2\text{Age} + \beta_3\text{Age}^2 + \beta_4\text{Gender} + u$

We want to allow women and men to get a differing return to education. We would create a slope dummy, say GenEdu = Gender*Education (where Gender=1 for females, 0 for males).

Thus, GenEdu = Education (women)
GenEdu = 0 (men)

We now estimate the model

$\text{Earnings} = \beta_0 + \beta_1\text{Education} + \beta_2\text{Age} + \beta_3\text{Age}^2 + \beta_4\text{Gender} + \beta_5\text{GenEdu} + u$

This can be rewritten as

$$\text{Earnings} = \beta_0 + \beta_1\text{Education} + \beta_2\text{Age} + \beta_3\text{Age}^2 + u \qquad \text{(Men)}$$

$$\text{Earnings} = (\beta_0 + \beta_4) + (\beta_1+\beta_5)*\text{Education} + \beta_2\text{Age} + \beta_3\text{Age}^2 + u \quad \text{(Women)}$$

- If β_4 is significant, then the interept term varies significantly between the genders. If β_5 is significant, then the slope coefficient for education varies significantly between genders.