Dummy Variables

Previous sections have shown how to **find the beta coefficient variables (β_1, β_2)** to understand the **significance of the corresponding explanatory variable** (x_1, x_2) on the **dependent y variable**

However we have assumed that the **true population beta being estimated is constant** across all x variable observations (e.g. a 1 unit change in x always gives a 2 unit change in y if β = 2) and that any deviations from this trend is due to **unavoidable variance that we have minimised** as best we can using OLS (i.e. the error terms)

> **As you may have guessed, it may not necessarily be the case that the effect of x on y is constant through all observations**

This is usually the case where **qualitative variables** not included in the regression model have an effect on y (e.g. sales for shoes may differ depending on gender)

Types of Dummy

A multiple regression model can be **adjusted** to represent scenarios where the parameters **differ for some of the sample observations** using dummy variables

Dummy variables are explanatory variables which take only a limited number of values (usually just 0 and 1) to model a situation with only two possible outcomes

There are **two** types

Intercept Dummy

An intercept dummy can be used where qualitative data **changes the value of the intercept constant** in a regression model

For instance, suppose we had the following regression model, where the **price of a house (y) depends only on the size of the house in square feet (x)**:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

This isn't a very accurate model since the effect that the size of a house has on its price is **unlikely to be constant across all sample observations** since, for instance, homes in more **desirable locations add a premium to the cost**

If this premium were to be **fixed**, **the intercept dummy variable** will model this as follows:

$$D = 1, if\ house\ in\ desirable\ area$$

$$D = 0, if\ house\ in\ undesirable\ area$$

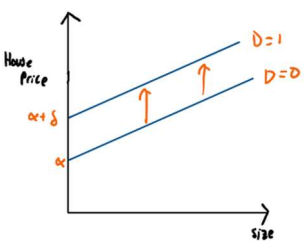
If we add D_i to the equation above using a new parameter, δ (equal to premium added), we get:

$$y_i = \alpha + \delta D_i + \beta x_i + \varepsilon_i$$

Hence the relationship will now depend on **whether or not the house is in a desirable area or not**, as the intercept dummy variable will create a **parallel shift** in the regression line by parameter δ:

$$IF\ D = 1\ THEN\ y = (\alpha + \delta) + \beta x_i$$

$$IF\ D = 0\ THEN\ y = \alpha + \beta x_i$$



Slope Dummy

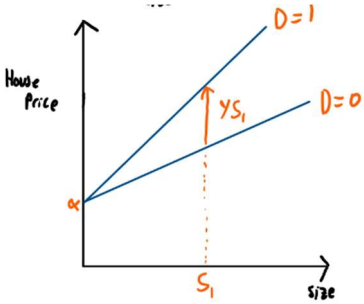
A slope dummy is used where the **qualitative data influences the data in a variate way**, and hence a new variable needs to be added that is **a product of a dummy and continuous variable**

For instance, perhaps being in a desirable area doesn't just **add a constant fixed premium to the house price**, but instead the premium **depends on the size of the house** (e.g. instead of adding £50,000 to every house in a desirable area, £100 is added onto the price for **every additional square foot**)

This situation can be modelled with a slope dummy, which is the **product of the house size (x) and the dummy (D)** with the parameter γ:

$$y_i = \alpha + \gamma(x_i D_i) + \beta x_i + \varepsilon_i$$

This will cause the adapted regression line when D = 1 to increase **further away from the original regression line with every subsequent increase in house price**



$$IF\ D = 1\ THEN\ y = \alpha + (\beta + \gamma)x_i$$

$$IF\ D = 0\ THEN\ y = \alpha + \beta x_i$$

The price of a home per square foot in a desirable area **when D = 1 is β + γ**, and in **less desirable areas** when D = 0 the price per square foot is **just β**

Hypothesis Testing with Dummy Variables

A hypothesis test can be performed to test **whether or not the qualitative factor (in this case the desirability of the area) affects the relationship (house price per square foot):**

$$H_0: \gamma = 0$$

$$H_1: \gamma \neq 0$$

The null hypothesis is that the **qualitative factor does not affect the relationship**, whilst the alternative hypothesis **asserts that it does**

This would be performed **using a simple t-test**

A **chow test** can also be used for dummy variables, discussed in next section

Combining Dummy Variables

In practice however, neither of the two dummy situations modelled so far **describe the true effect of a more desirable area on the price**

> It isnt usually the case that a **better area adds a constant fixed premium to the house price of different sizes** (usually depends on house size)
>
> Similarly, it is unusual for a house price to **increase at a constant rate at all sizes since discounts are usually enforced**

A houses price would definitely be expected to rise with size **but the relative impact of house size should depreciate as the size rises, as with the intercept dummy where the constant factor will dominate the equation except with larger houses**