Multiple Regression

The linear model discussed so far has been **very simple** as it contains **just one explanatory variable**:

$$y = \alpha + \beta x + \varepsilon$$

Using just a **single x variable** limits the **functionality of the regression model** since in practice, analysis requires **examination of multiple independent x variables**

Multiple Regression Analysis

Instead, we can formulate a model that contains multiple x variables, which causes hypothesis tests to be run slightly differently

The **total number** of independent variables used (i.e. **all x variables and the error variable, ε**) is represented by the letter **'k'**, creating the following **linear multiple regression equation**:

$$y = \alpha + \beta_1 x_1 + \ldots + \beta_{k-1} x_{k-1} + \varepsilon$$

The OLS estimator subtly changes accordingly, since we now need to find the **value of alpha and all the independent beta variables to find the lowest RSS**:

$$Min_{\alpha,\beta_1,\ldots,\beta_{k-1}} \sum {\varepsilon_i}^2$$

A **new condition** is also required to use OLS **to formulate this multiple regression model**:

6. **There is no exact linear relationship between any of the explanatory independent x variables**
    - **If so, they must be combined into one variable**

The other 5 conditions discussed in earlier sections **remain**

An **unbiased sample variance**, $\underline{\sigma^2}$, is also required to **proxy the population variance, using a similar equation to before with a different degrees of freedom calculation:**

$$\underline{\sigma^2} = \frac{\sum {\varepsilon_i}^2}{n-k} = \frac{RSS}{n-k}$$

Here, k is equal to the **number of parameters used** (i.e. number of x variables plus error term)

Hypothesis Testing

To test a hypothesis on a multiple regression line, an **individual parameter can be tested** to assess the significance of **one of the explanatory variables on the dependent variable (y)**

Testing a hypothesis follows the same method as with a single x variable model, with the exception of **replacing the 2 with the letter 'k' as noted**, and the coefficient for the variable in question must be identified

For instance, previously a hypothesis test on a **single x variable model may have been the following:**

$$H_0: \beta = 0$$

$$H_1: \beta \neq 0$$

Now, say we are testing the **significance of x_2 on y**, the hypothesis test would be **set out as follows**:

$$H_0: \beta_2 = 0$$

$$H_1: \beta_2 \neq 0$$

With:

$$\frac{\underline{\beta_2}}{\underline{Se(\beta_2)}} \sim t_{(n-k)}$$

A **suitable hypothesis test may now be run like previously**

The F Distribution

A **multiple variable regression** also allows for **entire population variances to be compared** (e.g. ${\sigma_1}^2$ and ${\sigma_2}^2$), since a model can be run **with and without** certain variables to see how this **effects overall variance**

This uses two estimators ($\underline{{\sigma_1}^2}$ and $\underline{{\sigma_2}^2}$) from **independent data sample of x variables** drawn from **normally distributed populations**

On **repeated sampling**, given the populations are **normally distributed**, the following holds:

$$\frac{\underline{{\sigma_1}^2}/{\sigma_1}^2}{\underline{{\sigma_2}^2}/{\sigma_2}^2} \sim F_{df1,df}$$

This means the **ratio of one populations sample variance to its population variance** divided by the **ratio of the other populations sample variance to its population variance** follows an **F distribution**

The F distribution is a ratio of two augmented Chi distributions (χ)

An F distribution test can be used to **test the significance of all variables in a multiple regression together**, to test whether or not the model **describes anything at all:**

$$H_0: \beta_1 = \beta_2 = \ldots = \beta_{k-1} = 0$$

$$H_1: at\ least\ one\ beta \neq 0$$

This hypothesis' test statistic is calculated as follows:

$$F = \frac{(TSS - RSS) \div (k-1)}{RSS \div (n-k)}$$

Note that the RSS and TSS are the **residual sum of squares** and the **total sum of squares** respectively, as discussed in the second section.

You may notice that the **denominator of this test statistic is equal to the sample variance estimator** ($\underline{\sigma^2}$)

If the null hypothesis is true, namely that all variables are insignificant in the regression model, then the numerator is also an **unbiased estimator of the sample variance**

On repeated sampling the distribution of the test statistic $\boldsymbol{F_{df1,df2} = F_{n-k,k-1}}$, which are the values used when **calculating critical values for this test**

If the null hypothesis is rejected, the test statistic will always be **biased upwards in the F distribution**, and hence the reject decision is based on the right hand tail of the test statistic (i.e. **more than positive critical value**)

The co-efficient of determination, R^2 will also be **biased upwards with increasing numbers of independent variables** used in the multiple regression model, irrespective of whether the variables are irrelevant or not. This will give a **skewed estimation** of a model's true accuracy, and hence an **adjusted measure, aR^2, is used** to **compensate** for the increasing use of independent variables:

$$aR^2 = R^2 \times \frac{n-1}{n-k}$$

We can rewrite the F test above in terms of R^2, creating a formula with the **exact same F distribution properties as previously**:

$$F = \frac{R^2 \div k}{(1 \div R^2) \div (n-k-1)}$$

This allows a hypothesis test to be run with **H_0 being $R^2 = 0$** and that none of the independent variables can explain changes in the dependent variable

Using an F Distribution to Assess the Significance of Additional Variables

Another way in which an F distribution can be used is to **test the significance of additional variables in a regression model**

> Important since including **irrelevant additional variables hinders efficiency** whilst **excluding relevant additional variables creates bias**

A test can determine if additional variables need to be **excluded or included in a model**, dividing a model into **two separate versions**: one with **a full unrestricted version including additional variables and one restricted version excluding additional variables**

For instance a restricted model may look like this:

$$y = \alpha + \beta_1 x_1 + \ldots + \beta_m x_m + \varepsilon_i$$

A corresponding unrestricted model, including additional variables would be represented like so:

$$y = \alpha + \beta_1 x_1 + \ldots + \beta_m x_m + \beta_{m+1} x_{m+1} + \ldots + \beta_{k-1} x_{k-1} + \varepsilon_i$$

The number of **removed variables, R, is equal to k-1-m**

A test can now be performed to see whether or not the additional variables, R, **make a significant change to the model:**

$$H_0: \beta_{m+!} = \beta_{m+2} = \ldots = \beta_{k-1} = 0$$

$$H_1: not\ all\ extra\ variables\ equal\ 0$$

An unbiased estimator of variance can be found with the **unrestricted model based on the RSS of the unrestricted (uRSS) model using the same sample variance equation**:

$$\underline{\sigma}^2 = \frac{uRSS}{n-k}$$

If the null hypothesis were **true and all extra beta variables did equal 0**, then **additional reduction in RSS would be due to statistical error**

Hence a **further unbiased variance estimator may be used** as follows, where **rRSS is the RSS of the restricted model and r is the number of additional variables that it does not include**:

$$\underline{\sigma^2} = \frac{rRSS - uRSS}{r}$$

If the null hypothesis was not true, then this above equation would be an **upwardly biased estimator of variance since the residual sum will be greater if the extra variables play a significant role** on the data in question

This gives the following test statistic which is the **ratio of the two above variance estimators in an F distribution**, where k is equal to m + number of additional variables:

$$F = \frac{(rSS - uSS) \div r}{uRSS \div (n-k)}$$

The distribution of the F test statistic is $F_{r,n-k}$ where **D1 equals r and D2 equals n-k**

AREA IN THE RIGHT TAIL OF DISTRIBUTION = 0.05

D_2 \ D_1	1	2	3	4	5	6	7	8	9	
1	161.448	199.500	215.707	224.583	230.162	233.986	236.768	238.883	240.543	241.882
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385	19.396
3	10.128	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.786
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637
8	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438	3.388	3.347
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544

Say for instance our restricted regression model had 3 independent variables whilst our unrestricted regression model had an additional 3 variables, if we took 20 data samples then:

R = 3
N = 20
K = 7 (since we add one to the total number of independent variables in the unrestricted model)

Hence the critical value would be **3.411**, so if the test statistic calculated using the above formula is **above this value**, then the **additional variables have a significant relevance to the model and we must reject the null hypothesis**