MULTIPLE REGRESSION ANALYSIS:

This is the analysis involving more than one explanatory variable. Given a regression model as $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots . + \beta_k x_k + u_i$, where β_0 is the intercept term. As usual, it gives the mean or average effect on *Y* of all the variables excluded from the model, although its mechanical interpretation is ;the Average of *Y* when *X*1 and *X*2 etc are set equal to zero. The coefficients β1, *β*2 etc are called the **partial regression coefficients.**

Assuming a model with two explanatory variables x_1 and x_2 and the model $= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u_i$, the following classical assumptions are fulfilled if x1 and *X2* are nonstochastic;

1. Linear regression model, or *linear in the parameters.*
2. Fixed *X* values or *X* values independent of the error term. Here, this means we require zero covariance between *ui* and each *X* variables.
cov (*ui* , *X*1*i*) and cov (*ui* , *X2i*) = 0
3. Zero mean value of disturbance *ui* . *E*(*ui* |*X*1*i, X2i*) = 0 for each *i.*
4. Homoscedasticity or constant variance of *ui* .that is; var (*ui*) = σ^2
5. No autocorrelation, or serial correlation, between the disturbances. that is;
cov (*ui* , *uj*) = 0 for $i \neq j$.
6. The number of observations *n* must be greater than the number of parameters to be estimated, which is 3 in our current case.
7. There must be variation in the values of the *X* variables.
8. No exact collinearity between the *X* variables. No **exact linear relationship** between *X*1 and *X2.*
9. There is no *specification bias*. The model is correctly specified.

The Meaning of Partial Regression Coefficients

As mentioned earlier, the regression coefficients *β*1 and *β2* are known as **partial regression** or **partial slope coefficients.** The meaning of partial regression coefficient is as follows: *β*1 measures the *change* in the mean value of *Y*, *E*(*Y*), per unit change in X_1, holding the value of X_2 constant. Put differently, it gives the "direct" or the "net" effect of a unit change in X_1 on the mean value of *Y*, net of any effect that X_2 may have on mean *Y*.

Likewise, β_2 measures the change in the mean value of *Y* per unit change in X_2, holding the value of X_1 constant. That is, it gives the "direct" or "net" effect of a unit change in X_2 on the mean value of *Y*, net of any effect that X_1 may have on mean *Y*.

ESTIMATION OF PARAMETERS

When we assume that the basic assumptions hold, the least squares estimators can be determined. For two explanatory variables x_1 and x_2, the least squares estimators can be determined using partial derivatives.

- For the model $y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$, determine sums of squares and differentiate with respect to the parameters. Obtain regression equations that can be solved simultaneously for the parameters α, β_1 and β_2.

$$\sum Y = n\alpha + \beta_1 \sum x_1 + \beta_2 \sum x_2 \ldots\ldots\ldots\ldots\ldots (a)$$

$$\sum x_1 Y = \alpha \sum x_1 + \beta_1 \sum x_1^2 + \beta_2 \sum x_1 x_2 \ldots\ldots\ldots\ldots\ldots (b)$$

$$\sum x_2 Y = \alpha \sum x_2 + \beta_1 \sum x_2 x_1 + \beta_2 \sum x_2^2 \ldots\ldots\ldots\ldots\ldots (c)$$

Solve the above regression equations simultaneously.

- The coefficient of determination shows the percentage of Y explained by variations due to changes in x_1 and x_2 and is a measure of the goodness of fit. It is given by $R^2 = \frac{SSR}{SST} = \frac{\sum \hat{y}_i^2}{\sum y_i^2} = \frac{\sum(\hat{Y}_i - \overline{Y})^2}{\sum(Y_i - \bar{Y})^2}$ OR

$R^2 = \frac{\widehat{\beta}_1 \sum y_i x_{1i} + \widehat{\beta}_2 \sum y_i x_{2i}}{\sum y_i^2}$. This formula does not put into consideration the degrees of freedom left in introducing a new explanatory variable to the model. To overcome this, the adjusted coefficient of multiple determination is used. This is given by; $R^2 = 1 - \frac{(\frac{\sum e_i^2}{n-k})}{(\frac{\sum y_i^2}{n-1})}$.

- The variances of the estimators of the parameters are estimated using;

$v(\hat{\beta}_1) = \frac{\delta^2 \sum x_2^2}{[\Sigma x_1^2 \sum x_2^2 - (\sum x_1 x_2)^2]}$, $v(\hat{\beta}_2) = \frac{\delta^2 \sum x_1^2}{[\Sigma x_1^2 \sum x_2^2 - (\sum x_1 x_2)^2]}$ and

$v(\hat{\alpha}) = \delta^2 [\frac{1}{n} + \frac{\bar{X}_1^2 \sum x_2^2 + \bar{X}_2^2 \sum x_1^2 - 2\bar{X}_1 \bar{X}_2 \sum x_1 x_2}{[\Sigma x_1^2 \sum x_2^2 - (\sum x_1 x_2)^2]}]$ Where $\delta^2 = \frac{\sum e_i^2}{n-k} = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-k}$

TESTING FOR SIGNIFICANCE OF REGRESSION

a) using the t-test for the individual parameters

b) using the f-test for the overall significance of the model.

USING THE T-TEST

Given the H_o: β_i =0 [not significant] and the H_A: $\beta_i \neq 0$ [significant] and a level of significance λ, the critical region is given as a two tailed test such that; $-t_{\frac{\lambda}{2},(n-k)} \leq t \leq t_{\frac{\lambda}{2}(n-k)}$ where k are the parameters estimated. For two explanatory variables k=2 and the computed t-statistic is given as $t = \hat{\beta}_i / se(\hat{\beta}_i)$. This is used to test significance of the individual parameters.

EXAMPLE: for the data below estimate the regression equation, variances and test the significance of the parameters at 5%

Y	10	8	7	7	5	6	9	10	11	6
X_1	5	7	6	6	8	7	5	4	3	9
X_2	10	6	12	5	3	4	13	11	13	3

USING THE F-TEST

Used to test for the overall significance of the model. The null hypothesis is that, the model is not significant while the alternative is that the model is significant at a given level of probability.

The critical region is $F_C \geq F_{\gamma,(K-1,N-K)}$. Where k are the parameters to be estimated, (k-1 and N-k are the degrees of freedom),

The computed f- statistic is given by F_c =MSR/MSE.

Try: using the example above, test for significance of the model at 5% significance level.