MULTIPLE LINEAR REGRESSION

A regression model that involves more than one regressor variable is called a **multiple regression model.**
Suppose that the yield in pounds of conversion in a chemical process depends on temperature and the catalyst concentration. A multiple regression model that might describe this relationship is

$y= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$…….Eq. 1

where *y* denotes the yield, *x*1 denotes the temperature, and *x*2 denotes the catalyst concentration. This is a **multiple linear regression model** with two regressor variables. The term **linear** is used because Eq. 1 is a linear function of the unknown parameters $\beta_0, \beta_1, \beta_2$.

In general, the **response** *y* may be related to *k* **regressor** or **predictor variables**. The model

$y= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots \ldots \ldots \beta_k x_k + \epsilon$ Eq. 2

is called a **multiple linear regression model** with *k* regressors.
Here y is called **dependent or response variables** and $x_1, x_2, \ldots \ldots \ldots x_k$ are **called regressors or independent variables or predictors.**
The parameters β_j, j = 0, 1,…, *k*, are called the **parameters or regression coefficients**. The parameter β_j represents the expected change in the response *y* per unit change in *xj* **when all of the remaining regressor variables** *xi*($i \neq j$) **are held constant**. For this

reason the parameters β_j, j = 1, 2,…, k, are often called **partial regression coefficients**.

ϵ is called the statistical error.

Multiple linear regression models are often used as **empirical models** or approximating functions. That is, the true functional relationship between y and $x1$, $x2$,…, xk is unknown, but over certain ranges of the regressor variables the linear regression model is an adequate approximation to the true unknown function.

If we let $x_1 = x, x_2 = x^2, \ldots \ldots x_k = x^k$, then Eq. 2 can be written as

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots \ldots . + \beta_k x^k + \epsilon$$

Which is called polynomial regression model.

Least - Squares Estimation of the Parameters:

Suppose that n > k observations are available, and let yi denote the i th observed response.

The method of ordinary least squares is used to estimate β_0, β_1, β_k. This OLS method is attributed to Carl Friedrich Gauss, a German Mathematician. The parameters β_0, β_1,……β_k are unknown and must be estimated using sample data. Suppose that we have n pairs of data, say (y1, x1), (y2, x2), …, (yn, xn). These data may result either from a controlled experiment designed specifically to collect the data, from an observational study, or from existing historical records (a retrospective study).

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots \ldots \ldots \beta_k x_k + \epsilon$ Eq. 2

We may write the sample regression model corresponding to Eq. 2 as

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots \ldots \ldots \beta_k x_{ik} + \epsilon_i$ Eq.3

i = 1, 2, ……n

Let each of the k predictor variables, x_1, x_2... x_k, have n levels. Then x_{ij} represents the i th level of the j th predictor variable x_j. For example, x_{51} represents the fifth level of the first predictor variable x_1, while x_{19} represents the first level of the ninth predictor variable, x_9. Observations, y_1, y_2... y_n, recorded for each of these n levels can be expressed in the following way:

$$y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + ... + \beta_k x_{1k} + \epsilon_1$$
$$y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + ... + \beta_k x_{2k} + \epsilon_2$$
$$..$$
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \epsilon_i$$
$$..$$
$$y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + ... + \beta_k x_{nk} + \epsilon_n$$

In matrix notation, the model given by Eq. 3 is

$Y = X\beta + \epsilon$ Eq. 4

Where $Y = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ . \\ y_n \end{bmatrix}$ $X = \begin{bmatrix} 1 & x_{11} & x_{12} & . & . & . & x_{1k} \\ 1 & x_{21} & x_{22} & . & . & . & x_{2k} \\ . & . & . & . & . & . & . \\ 1 & x_{n1} & x_{n2} & . & . & . & x_{nk} \end{bmatrix}$

$$\epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ . \\ . \\ \epsilon_n \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ . \\ \beta_k \end{bmatrix}$$

In general, **y** is an n × 1 vector of the observations, **X** is an n × (k +1) matrix of the levels of the regressor variables, β is a (k+1) × 1 vector of the regression coefficients, and **ε** is an n × 1 vector of random errors.

Assumptions:

1. For any set of values of $x_1, x_2, \ldots\ldots x_k$, the statistical or random error has a Normal probability distribution with mean zero and variance σ^2.

$$E(\epsilon_i) = 0 \; and \; V(\epsilon_i) = \sigma^2 \; for \; all \; i = 1,2, \ldots \ldots n.$$

2. The errors associated with any two observations is zero.

$$E\left(\epsilon_i, \epsilon_j\right) = 0$$

Combining points 1 and 2 we get,

$\mathrm{E}(\epsilon) = 0 \qquad V(\epsilon) = \sigma^2 I, where \; I \; is \; identity \; matrix.$

That is, $\mathrm{E}(\epsilon) = E \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ . \\ \epsilon_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ . \\ . \\ . \\ 0 \end{bmatrix}$

$$\mathrm{V}(\epsilon) = E(\epsilon\,\epsilon') = E\begin{bmatrix}\epsilon_1\\ \epsilon_2\\ .\\ .\\ .\\ \epsilon_n\end{bmatrix}[\epsilon_1 \quad \epsilon_2 \quad . \quad . \quad . \quad \epsilon_n]$$

$$= \begin{pmatrix} E(\varepsilon_1^2) & E(\varepsilon_1\varepsilon_2) \ldots E(\varepsilon_1\varepsilon_n) \\ E(\varepsilon_2) & E(\varepsilon_2^2) \ldots E(\varepsilon_2\varepsilon_n) \\ | & | \quad | \\ E(\varepsilon_n\varepsilon_1) & E(\varepsilon_n\varepsilon_2) \ldots E(\varepsilon_n^2) \end{pmatrix}$$

$$= \begin{pmatrix} \mathrm{Var}(\varepsilon_1) & \mathrm{Cov}(\varepsilon_1\varepsilon_2) \ldots \mathrm{Cov}(\varepsilon_1\varepsilon_n) \\ \mathrm{Cov}(\varepsilon_2\varepsilon_1) & \mathrm{Var}(\varepsilon_2) \ldots \quad \mathrm{Cov}(\varepsilon_2\varepsilon_n) \\ | & | \quad | \\ \mathrm{Cov}(\varepsilon_n\varepsilon_1) & \mathrm{Cov}(\varepsilon_n\varepsilon_2) \ldots \mathrm{Var}(\varepsilon_n) \end{pmatrix}$$

$$= \begin{pmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ | & | & | \\ 0 & 0 & \sigma^2 \end{pmatrix} = \sigma^2 \mathrm{I}$$

All covariances vanishes (zeros) as all ε's are independent and

(*a*) Each "ε" distribution has the same variance (Homosedasticity).

(*b*) All disturbances are pair wise uncorrelated. (No autocorrelation).

3. $E(x_i, \epsilon_i) = 0 \quad for\ all\ i = 1,2, \ldots \ldots n.$

4. There exists no linear relationship between any two of the regresors. That is there is no multicollinearity between the regressors.

5. The ϵ vector follows Multivariate Normal Distribution, that is, $\epsilon \sim N(0, \sigma^2 I)$

We wish to find the vector of least-squares estimators $\hat{\beta}$, that minimizes

$$f = \sum_{i=1}^{n} \epsilon_i^2 = \epsilon' \epsilon = (Y - X\beta)'(Y - X\beta)$$

$$= (Y' - \beta' X')(Y - X\beta)$$

$$= Y' Y - \beta' X' Y - Y' X\beta + \beta' X' X\beta$$

$$= Y' Y - 2\beta' X' Y + \beta' X' X\beta$$

since $\beta' X' Y$ is a 1×1 matrix, or a scalar, and its transpose $(\boldsymbol{\beta' X' Y})' = \boldsymbol{Y' X \beta}$ is the same scalar. The least-squares estimators must satisfy

$$\frac{\partial f}{\partial \beta} = -2X'Y + 2X'X\hat{\beta} = 0$$

$$\Rightarrow X'X\hat{\beta} = X'Y$$

$$\Rightarrow \hat{\beta} = (X'X)^{-1}X'Y \qquad \text{Eq. 5}$$

Thus, the **least-squares estimator** of **β** is given in Eq. 5 provided that the inverse matrix $(\mathbf{X'X})^{-1}$ exists. The $(\mathbf{X'X})^{-1}$ matrix will always exist if the regressors are **linearly independent**, that is, if no column of the **X** matrix is a linear combination of the other columns.

The fitted regression model corresponding to the observed values of

y_i is $\hat{y} = X\hat{\beta}$

$$\Rightarrow \hat{y} = X\,(X'X)^{-1}X'Y \quad = HY \qquad \text{From Eq. 5}$$

The n × n matrix $H = X\,(X'X)^{-1}X'$ is called the hat matrix.

The hat matrix and its properties play a central role in regression analysis.

H is a symmetric and idempotent matrix: HH=H

The difference between the observed value yi and the corresponding

fitted value $\hat{y}_1$ is the **residual** ei = yi – $\hat{y}_1$. The n residuals may be conveniently written in matrix notation as

$e = Y - \hat{Y} = Y - X\hat{\beta}$

There are several other ways to express the vector of residuals e that will prove useful, including

$e = Y - HY = (I - H)Y$

Examples:

1.

Obtain the least squares regression equation by matrix approach of the form $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, Given

$n = 6, \quad \Sigma X_1 = 48, \quad \Sigma X_2 = 42, \quad \Sigma X_1 X_2 = 236$

$\Sigma Y = 300, \quad \Sigma X_1 Y = 1818, \quad \Sigma X_2 Y = 2820$

$\Sigma X_1^2 = 474, \quad \Sigma X_2^2 = 434, \quad \Sigma Y^2 = 19008$

Sol. We know that $\hat{\beta} = (X'X)^{-1} X'Y$ ..... (1)

Where $X'X = \begin{pmatrix} n & \Sigma X_1 & \Sigma X_2 \\ \Sigma X_1 & \Sigma X_1^2 & \Sigma X_1 X_2 \\ \Sigma X_2 & \Sigma X_1 X_2 & \Sigma X_2^2 \end{pmatrix} = \begin{pmatrix} 6 & 48 & 42 \\ 48 & 474 & 236 \\ 42 & 236 & 434 \end{pmatrix}$

and $X'Y = \begin{pmatrix} \Sigma Y \\ \Sigma X_1 Y \\ \Sigma X_2 Y \end{pmatrix} = \begin{pmatrix} 300 \\ 1818 \\ 2820 \end{pmatrix}$

Now $|X'X| = \begin{vmatrix} 6 & 48 & 42 \\ 48 & 474 & 236 \\ 42 & 236 & 434 \end{vmatrix} = 15600$

Then $(X'X)^{-1} = \dfrac{\text{Adj}\,(X'X)}{|X'X|}$

Where adj$(X'X)$ = Transpose of cofactor matrix $(X'X)$

$$= \begin{pmatrix} 9.617 & -0.700 & -0.550 \\ -0.700 & 0.054 & 0.038 \\ -0.55 & 0.038 & 0.035 \end{pmatrix}$$

and $$\hat{\beta} = \begin{pmatrix} 9.617 & -0.700 & -0.550 \\ -0.700 & 0.054 & 0.038 \\ -0.55 & 0.038 & 0.035 \end{pmatrix} \begin{pmatrix} 300 \\ 1818 \\ 2820 \end{pmatrix} = \begin{pmatrix} 61.600 \\ -3.646 \\ 2.538 \end{pmatrix}$$

Thus $\hat{Y} = 61.60 - 3.646X_1 + 2.538X_2$

2.

Delivery Time, y (min)	Number of Cases, x_1	Distance, x_2 (ft)
16.68	7	560
11.50	3	220
12.03	3	340
14.88	4	80
13.75	6	150
18.11	7	330
8.00	2	110
17.83	7	210
79.24	30	1460
21.50	5	605

Example

An analyst studying a chemical process expects the yield to be affected by the levels of two factors, x_1 and x_2. Observations recorded for various levels of the two factors ar following table. The analyst wants to fit a first order regression model to the data. Interaction between x_1 and x_2 is not expected based on knowledge of similar processes. U levels and the yield are ignored for the analysis.

Observation Number	Factor 1 (x_{i1})	Factor 2 (x_{i2})	Yield (y_i)
1	41.9	29.1	251.3
2	43.4	29.3	251.3
3	43.9	29.5	248.3
4	44.5	29.7	267.5
5	47.3	29.9	273.0
6	47.5	30.3	276.5
7	47.9	30.5	270.3
8	50.2	30.7	274.9
9	52.8	30.8	285.0
10	53.2	30.9	290.0
11	56.7	31.5	297.0
12	57.0	31.7	302.5
13	63.5	31.9	304.5
14	65.3	32.0	309.3
15	71.1	32.1	321.7
16	77.0	32.5	330.7
17	77.8	32.9	349.0

Polynomial Regression

Problem: Data: average claims paid per policy for automobile insurance in New Brunswick in the years 1971-1980:

Year 1971 1972 1973 1974 1975

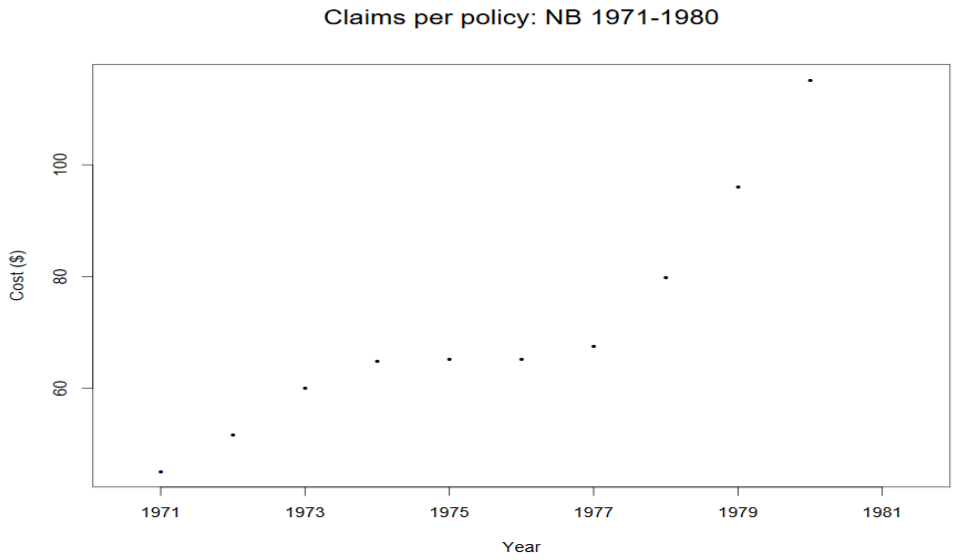
Cost 45.13 51.71 60.17 64.83 65.24

Year 1976 1977 1978 1979 1980

Cost 65.17 67.65 79.80 96.13 115.19

Sol.

Data Plot



The equation of the polynomial regression for the above graph data would be

$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$

$\hat{\beta} = (X'X)^{-1}X'Y$

Where Y = $\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix}$ X = $\begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \\ . & . & . \\ 1 & x_n & x_n^2 \end{bmatrix}$ $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$ $\epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ . \\ \epsilon_n \end{bmatrix}$

$$X'X = \begin{bmatrix} 1 & 1 & . & . & 1 \\ x_1 & x_2 & . & . & x_n \\ x_1^2 & x_2^2 & . & . & x_n^2 \end{bmatrix} \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \\ . & . & . \\ 1 & x_n & x_n^2 \end{bmatrix}$$

$$= \begin{bmatrix} n & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{bmatrix}$$

$$X'Y = \begin{bmatrix} 1 & 1 & . & . & 1 \\ x_1 & x_2 & . & . & x_n \\ x_1^2 & x_2^2 & . & . & x_n^2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \\ \sum x_i^2 y_i \end{bmatrix}$$

Properties of the Least-Squares Estimators:

1. E ($\hat{\beta}$) $= \beta \Rightarrow \hat{\beta}$ *is an unbiased estimator of* β.

Proof: Since E (ϵ) $= 0, E(\mathrm{Y}) =$ E(X$\beta + \epsilon$) = E(Xβ) $+ E(\epsilon) =$ Xβ

$\Rightarrow E(\hat{\beta}) = E((X'X)^{-1}X'Y)$

$= (X'X)^{-1}X'\, E(Y)$

$= (X'X)^{-1}X'$ Xβ

$= \beta$ [Since $(X'X)^{-1}X'$ X = I]

2. V ($\hat{\beta}$) $= \sigma^2\ (X'X)^{-1}$

Proof:

$\hat{\beta} = (X'X)^{-1}X'Y$

$= ((X'X)^{-1}X'(X\beta + \epsilon)$

$= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'\epsilon$

$= \beta + (X'X)^{-1}X'\epsilon$

$\Rightarrow \hat{\beta} - \beta = (X'X)^{-1}X'\epsilon$

By definition,

V ($\hat{\beta}$) = E $(\hat{\beta} - \beta)(\hat{\beta} - \beta)^{/}$

= E $((X'X)^{-1}X'\epsilon)((X'X)^{-1}X'\epsilon)')$

$= E\left((X'X)^{-1}X'\epsilon\right)\left(\epsilon'X\ (X'X)^{-1}\right)$

$= E\left((X'X)^{-1}X'\epsilon\ \epsilon'X\ (X'X)^{-1}\right)$

$= (X'X)^{-1}X'E(\epsilon\ \epsilon')X\ (X'X)^{-1})$

$= (X'X)^{-1}X'\ \sigma^2 IX\ (X'X)^{-1})$ [Since $V(\epsilon) = E(\epsilon\ \epsilon') = \sigma^2 I$]

$= \sigma^2\ [(X'X)^{-1}(X'\ X)\ (X'X)^{-1})]$

$= \sigma^2\ (X'X)^{-1}$

3. An unbiased estimator of σ^2 is given by

$$\widehat{\sigma^2} = \frac{RSS}{n-k-1} = MSR \text{ 205}$$

Proof: **residual** ei = yi – $\hat{y}_i$

RSS = Residual sum of squares

$= \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = e'e$

$= (Y - X\hat{\beta})'\ (Y - X\hat{\beta})$

$= (Y' - \hat{\beta}'X')\ (Y - X\hat{\beta})$

$= Y'Y - \hat{\beta}'X'Y - Y'\ X\hat{\beta} + \hat{\beta}'X'\ X\hat{\beta}$

$= Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'\ X\hat{\beta}$

$= Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'Y$ [Since $X'\ X\hat{\beta} = X'Y$]

$= Y'Y - \hat{\beta}'X'Y$

$e_i \sim$ NID $(0, \sigma^2)$

$\Rightarrow \frac{e_i}{\sigma} \sim$ NID (0, 1)

$\Rightarrow \left(\frac{e_i}{\sigma}\right)^2 \sim \chi^2$ distribution with (1) degrees of freedom.

$\Rightarrow \sum_{i=1}^{n}\left(\frac{e_i}{\sigma}\right)^2 = \frac{RSS}{\sigma^2} \sim \chi^2$ distribution with residual sum of squares has n – (k+1) degrees of freedom associated with it since (k+1) parameters are estimated in the regression model. The **residual mean square** is (n-k-1) degrees of freedom.

$E\left(\frac{RSS}{\sigma^2}\right) = n - k - 1$

$\Rightarrow \sigma^2 = E(\frac{RSS}{n-k-1})$

$\Rightarrow \hat{\sigma}^2 = \frac{RSS}{n-k-1} = \text{MSR}$

$\Rightarrow$ MSR is an unbiased estimator of σ^2

$\Rightarrow$ MSR = $\frac{Y'Y - \widehat{\beta}'X'Y}{n-k-1}$ is an unbiased estimator of σ^2

4. The sum of observed values of

y_i *is always equal to the sum of estimated or fitted values of* y_i.

$\Rightarrow \sum y_i = \sum \widehat{y_i} \quad \Rightarrow \sum_{i=1}^{n} e_i = 0$ (The sum of the residuals in any regression model is always zero)

5. The sum of the residuals weighted by the corresponding value of the regressor variable always equals zero, that is,

$$\sum_{i=1}^{n} x_i e_i = 0$$

6. The sum of the residuals weighted by the corresponding fitted value always equals zero, that is,

$$\sum_{i=1}^{n} \widehat{y_i}\, e_i = 0$$

7. $e'\hat{Y} = 0$

Co-efficient of Determination R^2 and adjusted R^2:

We now consider the Goodness of Fit of the fitted regression line to a set of data; i.e. we will find out how well the sample regression line fits the data. If all the observations were to lie on the regression line, we would obtain a perfect fit, but this is rarely the case. Generally there will be some positive and some

negative e_i. These residuals around the regression line should be as small as possible. The **Co-efficient of Determination $\mathbf{R^2}$** is a summary measure that tells how well the sample regression fits the data.

Where $R^2 = \frac{ESS}{TSS}$ = 1- $\frac{RSS}{TSS}$

RSS = Residual Sum of Squares = $e'e = \left(Y - X\hat{\beta}\right)'\left(Y - X\hat{\beta}\right)$

$$= Y'Y - \hat{\beta}'X'Y - Y'X\hat{\beta} + \hat{\beta}X'X\hat{\beta}$$

$$= Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}X'X\hat{\beta}$$

$$= Y'Y - \hat{\beta}'X'Y$$

(Since $X'X\hat{\beta} = X'Y$)

ESS= Explained sum of squares = TSS – RSS

$$= Y'Y - (Y'Y - \hat{\beta}'X'Y)$$

$$= \hat{\beta}'X'Y$$

We now define

Co-efficient of Determination $\mathbf{R^2}$ as

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{RSS}{TSS} = \frac{\hat{\beta}'X'Y}{Y'Y}$$

Since TSS is a measure of the variability in y without considering the effect of the regressor variable x and ESS is a measure of the variability in y remaining after x has been considered, R^2 is often called the proportion of variation explained by the regressor x.

Properties of $\boldsymbol{R^2}$

1. Since $0 \leq ESS \leq TSS$, it follows that $0 \leq R^2 \leq 1$. Values of R^2 that are close to 1 imply that most of the variability in y is explained by the regression model.

2. Sometimes, a low value of R^2 is a result of a poorly specified model. In these cases the model can often be improved by the addition of one or more predictor or regressor variables.

3. Sometimes, a low value of R^2 results from having a lot of variability in the measurements of the response.

4. The statistic R^2 should be used with caution, since it is always possible to make R^2 large by adding enough terms to the model. This makes R^2 misleading.

We don't necessarily discard a model based on a low R-Squared value. Its a better practice to look at the AIC and prediction accuracy on validation sample when deciding on the efficacy of a model.

What about adjusted R-Squared?

As you add more *X* variables to your model, the R-Squared value of the new bigger model will always be greater than that of the smaller subset. This is because, since all the variables in the original model is also present in the super-set as well, therefore, whatever new variable we add can only add (if not significantly) to the variation that was already explained. It is here, the adjusted R-Squared value comes to help. Adj R-Squared penalizes total value for the number of terms (read predictors) in your model. Therefore it is a good practice to look at adj-R-squared value over R-squared.

$$R^2_{adj} = 1 - \frac{MSR}{MST}$$

here, *MSR* is the *mean squared error* given by MSR = $\frac{RSS}{n-k-1}$

and MST = $\frac{TSS}{n-1}$ is the *mean squared total*, where *n* is the number of observations

Therefore, by moving around the numerators and denominators, the relationship between R^2 *and* R^2_{adj} becomes:

$$R^2_{adj} = 1 - \frac{(1-R^2)(n-1)}{n-k-1}$$

Note: 1. When k > 2, $R^2_{adj} < R^2$

2. Sometimes R^2_{adj} can be negative.

Test for Significance of Regression:

The test for significance of regression is a test to determine if there is a linear relationship between the response y and any of the regressor variables x_1, x_2,…, x_k. This procedure is often thought of as an overall or global test of model adequacy. The appropriate hypotheses are

$$H_0: \beta_1 = \beta_2 = \cdots \ldots = \beta_k = 0$$

against $H_1: \beta_j \neq 0 \; for \; atleast \; one \; j = 1,2, \ldots \ldots k$

Rejection of this null hypothesis implies that at least one of the regressors x_1, x_2,…, x_k contributes significantly to the model.

This test procedure is a generalization of the analysis of variance used in simple linear regression. The total sum of squares TSS is partitioned into a sum of squares due to regression, ESS, and a residual sum of squares, RSS.

TSS = ESS + RSS

Where TSS = $Y'Y$ ESS= $\hat{\beta}'X'Y$ RSS = $Y'Y - \hat{\beta}'X'Y$

The **degree-of-freedom** breakdown is determined as follows. The total sum of squares, TSS, has df = *n* − 1 degrees of freedom. The model or regression sum

of squares, ESS, has df = k degree of freedom because ESS is completely determined by k parameters. RSS has df = n – k-1 degrees of freedom.

$\frac{RSS}{\sigma^2}$ follows χ^2 distribution with (n-k-1) degrees of freedom.

$\frac{ESS}{\sigma^2}$ follows χ^2 distribution with k degree of freedom.

RSS and ESS are independent. By the definition of an F statistic given, we get

F = $\frac{\frac{\frac{ESS}{\sigma^2}}{\mathrm{k}}}{\frac{\frac{RSS}{\sigma^2}}{\mathrm{n-k-1}}}$ follows $F_{k,n-k-1}$ distribution

The F test is equivalent to F = $\frac{\frac{R^2}{k}}{\frac{1-R^2}{n-k-1}}$

Analysis of Variance Table:

Sources of Variation	Sum of Squares	Degrees of Freedom	Mean Sum of Squares	F
Regression	ESS	K	MSE=ESS/k	F= MSE/MSR
Residual	RSS	n-k-1	MSR=RSS/(n-k-1)	
Total	TSS	n-1	MST=TSS/(n-1)	

Therefore, to test the hypothesis

$H_0: \beta_1 = \beta_2 = \cdots \ldots = \beta_k = 0$, compute the test statistic F and reject H_0 if F > $F_{\alpha,k,n-k-1}$.

Tests on Individual Regression
Coefficients and Subsets of Coefficients:

Once we have determined that at least one of the regressors is important, a logical question becomes which one(s). Adding a variable to a regression model always causes the sum of squares for regression to increase and the residual sum of squares to decrease.

The addition of a regressor also increases the variance of the fitted value, so we must be careful to include only regressors that are of real value in explaining the response. Furthermore, adding an unimportant regressor may decrease the residual mean square, which may decrease the usefulness of the model.
The hypotheses for testing the significance of any individual regression coefficient, such as βj, are

$$H_0: \beta_j = 0 \; against \; H_1: \beta_j \neq 0$$

If H_0: βj = 0 is accepted, then this indicates that the regressor xj can be deleted from the model. The **test statistic** for this hypothesis is

$$t = \frac{\widehat{\beta_j}}{\sqrt{\widehat{\sigma}^2 \, C_{jj}}} \qquad (\text{where } \widehat{\sigma^2} = \frac{RSS}{n-k-1} = MSR)$$

Cjj is the diagonal element of $(X'X)^{-1}$ corresponding to $\widehat{\beta}_j$. .
The null hypothesis H_0: $\beta_j = 0$ is rejected if

$|t| > t_{\alpha/2, n-k-1}$.

Note that this is really a partial or marginal test because the regression coefficient depends on all of the other regressor variables x_i ($i \neq j$) that are in the

model. Thus, this is a test of the contribution of xj given the other regressors in the model.

We can also investigate the contribution of a subset of the regressor variables to the model.

Consider the regression model with k regressors

$Y = X\beta + \epsilon$

Where Y is an n × 1 vector of the observations, X is an n × k matrix of the levels of the regressor variables, β is a k × 1 vector of the regression coefficients, and ε is an n × 1 vector of random errors.

We would like to determine if some subset of (< k) regressors contributes significantly to the regression model. Let the vector of regression coefficients be partitioned as follows:

$$\beta = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}$$

where θ_1 is r × 1 and θ_2 is (k +1– r) × 1. We wish to test the hypotheses

$$H_0 : \theta_2 = 0 \; against \; H_1 : \theta_2 \neq 0$$

where $\theta_1 = (\beta_0, \beta_1, \beta_2, \dots\dots\dots \beta_{r-1})'$

$\theta_2 = (\beta_r, \beta_{r+1}, \dots\dots\dots \beta_k)'$

For the full model,

$\hat{\beta} = (X'X)^{-1}X'Y$

RSS = $Residual\ Sum\ of\ Squares = Y'Y - \hat{\beta}'X'Y$

MSR = $\frac{RSS}{n-k-1}$

ESS= Explained sum of squares = $\hat{\beta}'X'Y$

To find the contribution of the terms in θ_2 to the regression, fit the model assuming that the null hypothesis H0: θ_2 = 0 is true. This reduced model is

Y = $X_1\theta_1 + \epsilon$

The least-squares estimator of θ_1 in the reduced model is

$$\widehat{\theta_1} = (X_1'X_1)^{-1}(X_1'Y)$$

The regression sum of squares is

$$ESS(\theta_1) = \widehat{\theta_1}'X_1'Y$$

Where $X_1 = (1, x_1, x_2, \dots\dots\dots x_{r-1})$ and $X_2 = (x_r, x_{r+1}, \dots\dots\dots x_k)$

The test statistic for this test follows the F distribution and can be calculated as follows:

$$F = \frac{\frac{ESS(\theta_2|\theta_1)}{k+1-r}}{MSR}$$

$$where\ ESS(\theta_2|\theta_1) = \text{ESS - ESS } (\theta_1) = \hat{\beta}'X'Y - \widehat{\theta_1}'X_1'Y$$

where $ESS(\theta_2|\theta_1)$is the the increase in the regression sum of squares when the variables corresponding to the coefficients in θ_2 are added to a model already containing θ_1.

The null hypothesis, H_0, is rejected if $F \geq F_{\alpha,k+1-r,n-k-1}$. Rejection of $H_0: \theta_2 = 0$ leads to the conclusion that at least one of the variables in, $x_r, x_{r+1}, \dots x_k$ contributes significantly to the regression model.

Goodness of fit of the model:

$$H_0: R^2 = 0\ (The\ model\ is\ not\ good)$$

against $H_1: R^2 \neq 0\ (The\ model\ is\ good)$

The test statistic to be used $F = \frac{\frac{R^2}{k}}{\frac{1-R^2}{n-k-1}}$

F follows $F_{k,n-k-1}$ $distribution.$

Compute the test statistic F and reject H_0 if F > $F_{\alpha,k,n-k-1}$.

Example: Suppose we have the following data from a random sample of n=8 car sales at Bob's Used Car's lot:

Selling price ($1000s):y	11	15	13	14	0	19	16	8
Hours of required work:x1	0	11	11	7	4	10	5	8
Buying price ($1000s):x2	1	5	4	3	1	4	4	2

Bob thinks that he can predict a car's selling price (y) from the number of work hours the car requires (x1) and the price he pays for it (x2).

Given the data

	$Y_t - \bar{Y}$	$X_{2t} - \bar{X}_2$	$X_{3t} \quad \bar{X}_3$
$Y_t - \bar{Y}$	3450	−300	65000
$X_{2t} - \bar{X}_2$	−300	30	−5900
$X_{3t} - \bar{X}_3$	65000	−5900	15,80,000

$\bar{Y} = 80, \quad \bar{X}_2 = 6, \quad \bar{X}_3 = 800, \quad T = 10$

and the model

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t$$

Where Y_t is the quantity demanded of a certain commodity, X_2 is its price and X_3 is the consumer income.

(*a*) Estimate the model using OLS, and, Interpret β_2 and β_3.

(*b*) Compute price and income elasticities and interpret.

(*c*) Compute R^2 and test for overall goodness of fit.

Confidence Intervals on the Regression Coefficients:

To construct confidence interval estimates for the regression coefficients βj, we will continue to assume that the errors εi are normally and independently distributed with mean zero and variance σ^2. Therefore, the observations yi are

normally and independently distributed with mean $\beta_0 + \sum_{j=1}^{k} x_{ij}$ and variance σ^2. Since the least squares estimator is a linear combination of the observations, it follows that is normally distributed with mean vector β and covariance matrix $\sigma^2 (X'X)^{-1}$. This implies that the marginal distribution of any regression coefficient is normal with mean βj and variance σ^2Cjj, where Cjj is the jth diagonal element of the $(X'X)^{-1}$ matrix. Consequently, each of the statistics

$\frac{\widehat{\beta_j}-\beta_j}{\sqrt{\widehat{\sigma}^2 C_{jj}}}$ follows t distribution with df n-k-1.

where $\widehat{\sigma^2} = \frac{RSS}{n-k-1} = MSR$

So the 100(1- α)% confidence interval for β_j (j=1,2,..., k) is obtained as follows:

$$(\widehat{\beta_j} - t_{\frac{\alpha}{2},n-k-1}\sqrt{\widehat{\sigma}^2 C_{jj}} - t_{\frac{\alpha}{2},n-k-1}\sqrt{\widehat{\sigma}^2 C_{jj}}\ ,\ \ \widehat{\beta_j} + t_{\frac{\alpha}{2},n-k-1}\sqrt{\widehat{\sigma}^2 C_{jj}}$$

Weighted Least Squares:

The assumptions were: The model errors have mean zero and constant variance and are uncorrelated. Here we focus on methods and procedures for building regression models when some of the above assumptions are violated.

The method of weighted least squares is an useful method in building regression models in situations where some of the underlying assumptions are violated.

* The standard linear model assumes that $\text{Var}(\varepsilon_i) = \sigma^2$ for $i = 1, \ldots, n$.

* As we have seen, however, there are instances where

$$\text{Var}(Y \mid \boldsymbol{X} = \boldsymbol{x}_i) \;=\; \text{Var}(\varepsilon_i) \;=\; \frac{\sigma^2}{w_i}.$$

* Here $w_1, \ldots, w_n$ are known positive constants.

* Weighted least squares is an estimation technique which weights the observations proportional to the reciprocal of the error variance for that observation and so overcomes the issue of non-constant variance.

Weighted Least Squares in Simple Regression

* Suppose that we have the following model

$$Y_i \;=\; \beta_0 + \beta_1 X_i + \varepsilon_i \qquad i = 1, \ldots, n$$

where $\varepsilon_i \sim \text{N}(0, \sigma^2/w_i)$ for **known** constants $w_1, \ldots, w_n$.

* The weighted least squares estimates of β_0 and β_1 minimize the quantity

$$S_w(\beta_0, \beta_1) \;=\; \sum_{i=1}^{n} w_i (y_i - \beta_0 - \beta_1 x_i)^2$$

* Note that in this weighted sum of squares, the weights are inversely proportional to the corresponding variances; points with low variance will be given higher weights and points with higher variance are given lower weights.

* The weighted least squares estimates are then given as

$$\begin{aligned}\hat{\beta}_0 &= \overline{y}_w - \hat{\beta}_1 \overline{x}_w \\ \hat{\beta}_1 &= \frac{\sum w_i(x_i - \overline{x}_w)(y_i - \overline{y}_w)}{\sum w_i(x_i - \overline{x}_w)^2}\end{aligned}$$

where $\overline{x}_w$ and $\overline{y}_w$ are the weighted means

$$\overline{x}_w = \frac{\sum w_i x_i}{\sum w_i} \qquad \overline{y}_w = \frac{\sum w_i y_i}{\sum w_i}.$$

* Some algebra shows that the weighted least squares estimates are still unbiased.

Let the model is Y = X$\beta + \epsilon$

$\Rightarrow E(Y) = X\beta$

General Weighted Least Squares Solution

* Let $\boldsymbol{W}$ be a diagonal matrix with diagonal elements equal to $w_1, \ldots, w_n$.

* The the Weighted Residual Sum of Squares is defined by

$$S_w(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i(y_i - \boldsymbol{x}_i^t\boldsymbol{\beta})^2 = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^t\boldsymbol{W}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}).$$

* Weighted least squares finds estimates of $\boldsymbol{\beta}$ by minimizing the weighted sum of squares.

* The general solution to this is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{W}\boldsymbol{Y}.$$

$$\Rightarrow V(Y) = \begin{bmatrix} 1/w_1 & 0 & 0 & \cdots & 0 \\ 0 & 1/w_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1/w_n \end{bmatrix} \sigma^2$$

$$\Rightarrow W = \begin{bmatrix} w_1 & 0 & 0 & . & . & . & 0 \\ 0 & w_2 & 0 & . & . & . & 0 \\ . & . & . & . & . & . & . \\ 0 & 0 & 0 & . & . & . & w_n \end{bmatrix}$$

Variable selection and model building

In most practical problems, especially those involving historical data, the analyst has a rather large pool of possible **regressors**, of which only a few are likely to be important. Finding an appropriate subset of regressors for the model is often called the **variable selection and model building problem.**

(1) We would like the model to include as many regressors as possible so that the information content in these factors can influence the predicted value of y.

(2) We want the model to include as few regressors as possible because the variance of the prediction increases as the number of regressors increases. Also the more regressors there are in a model, the greater the costs of data collection and model maintenance.

The process of finding a model that is a compromise between these two objectives is called selecting the **"best" regression equation**.

1. Coefficient of Multiple Determination A measure of the adequacy of a regression model that has been widely used is the coefficient of multiple determination, R^2.

Let us denote the coefficient of multiple determination for a subset regression model with r terms by R_r^2.

$$R_r^2 = \frac{ESS(r)}{TSS} = 1 - \frac{RSS(r)}{TSS}$$

where ESS(r)and RSS(r) denote the regression sum of squares and the residual sum of squares, respectively, for a r-term subset model.

Now r increases as ESS(r) increases and is a maximum when r = k + 1.

Therefore, the analyst uses this criterion by adding regressors to the model up to the point where an additional variable is not useful in that it provides only a small increase in.

2. Adjusted R^2 To avoid the difficulties of interpreting R^2, some analysts prefer to use the adjusted R^2 statistic, defined as

$$R_{adj}^2 = 1 - \frac{MSR}{MST}$$

where MSRand MST denote the residual mean sum of squares and the mean total sum of squares.

3. Residual Mean Square The residual mean square for a subset regression model may also used as a model evaluation criterion. RSS (p) always decreases as p increases, MSR (p) initially decreases, then stabilizes, and eventually may increase. The subset regression model that minimizes MSR (p) should be selected.

When we fit a multiple regression model, we use the *p*-value in the ANOVA table to determine whether the model, as a whole, is significant. A natural next question to ask is which predictors, among a larger set of all potential predictors, are important. We could use the individual *p*-values and refit the model with only significant terms. But, remember that the *p*-values are adjusted for the other terms in the model. So, picking out the subset of significant

predictors can be somewhat challenging. This task of identifying the best subset of predictors to include in the model, among all possible subsets of predictors, is referred to as *variable selection*

These methods are generally referred to as stepwise-type procedures. They can be classified into three broad categories:
(1) forward selection, (2) backward elimination,
and (3) stepwise regression,

Forward Selection

- This procedure begins with the assumption that there are no regressors in the model other than the intercept. The first regressor selected for entry into the equation is the one that has the largest simple correlation with the response variable y. Suppose that this regressor is $x1$. This is also the regressor that will produce the largest value of the F statistic for testing significance of regression.

The second regressor chosen for entry is the one that now has the largest correlation with y after adjusting for the effect of the first regressor entered ($x1$) on y. We refer to these correlations as partial correlations.

Every time we always choose from the rest of the variables the one that yields the best accuracy in prediction when added to the pool of already selected variables. This accuracy can be measured by the F-statistic, LRT, AIC, BIC, etc.

For example, if we have 10 predictor variables, first we would approximate y with a constant, and then use one variable out of the 10 (I would perform 10 regressions, each time using a different predictor variable; for every regression I have a residual sum of squares; the variable that yields the minimum residual sum of squares is chosen and put in the pool of selected variables). We then proceed to choose the next variable from the 9 left, etc.

Backward Elimination

Forward selection begins with no regressors in the model and attempts to insert variables until a suitable model is obtained. Backward elimination attempts to find a good model by working in the opposite direction. That is, we begin with a model that includes all k regressors. Then the partial F statistic (or equivalently, a t statistic) is computed for each regressor as if it were the last variable to enter the model. The smallest of these partial F (or t) statistics is compared with a preselected value, that regressor is removed from the model. Now a regression model with $k - 1$ regressors is fit, the partial F (or t) statistics for this new model calculated, and the procedure repeated.

Stepwise Regression

The two procedures described above suggest a number of possible combinations. One of the most popular is the stepwise regression algorithm of Efroymson [1960]. Stepwise regression is a modification of forward selection in which at each step all regressors entered into the model previously are reassessed via their partial F (or t) statistics. A regressor added at an earlier step may now be redundant because of the relationships between it and regressors now in the equation. If the partial F (or t) statistic for a variable is less than critical values, that variable is dropped from the model.

Stepwise regression requires two cutoff values, one for entering variables and one for removing them.