Chapter 3

Multiple Linear Regression Model

We consider the problem of regression when the study variable depends on more than one explanatory or independent variables, called a multiple linear regression model. This model generalizes the simple linear regression in two ways. It allows the mean function $E(y)$ to depend on more than one explanatory variables and to have shapes other than straight lines, although it does not allow for arbitrary shapes.

The linear model:

Let y denotes the dependent (or study) variable that is linearly related to k independent (or explanatory) variables $X_1, X_2, ..., X_k$ through the parameters $\beta_1, \beta_2, ..., \beta_k$ and we write

$$y = X_1\beta_1 + X_2\beta_2 + ... + X_k\beta_k + \varepsilon.$$

This is called the multiple linear regression model. The parameters $\beta_1, \beta_2, ..., \beta_k$ are the regression coefficients associated with $X_1, X_2, ..., X_k$ respectively and ε is the random error component reflecting the difference between the observed and fitted linear relationship. There can be various reasons for such difference, e.g., the joint effect of those variables not included in the model, random factors which can not be accounted for in the model etc.

Note that the j^{th} regression coefficient β_j represents the expected change in y per unit change in the j^{th} independent variable X_j. Assuming $E(\varepsilon) = 0$,

$$\beta_j = \frac{\partial E(y)}{\partial X_j}.$$

Linear model:

A model is said to be linear when it is linear in parameters. In such a case $\frac{\partial y}{\partial \beta_j}$ (or equivalently $\frac{\partial E(y)}{\partial \beta_j}$) should not depend on any $\beta's$. For example,

i) $y = \beta_0 + \beta_1 X$ is a linear model as it is linear in the parameters.

ii) $y = \beta_0 X^{\beta_1}$ can be written as

$$\log y = \log \beta_0 + \beta_1 \log X$$

$$y^* = \beta_0^* + \beta_1 x^*$$

which is linear in the parameter β_0^* and β_1, but nonlinear is variables $y^* = \log y, x^* = \log x$. So it is a linear model.

iii) $y = \beta_0 + \beta_1 X + \beta_2 X^2$

is linear in parameters β_0, β_1 and β_2 but it is nonlinear is variables X. So it is a linear model

iv) $y = \beta_0 + \dfrac{\beta_1}{X - \beta_2}$

is nonlinear in the parameters and variables both. So it is a nonlinear model.

v) $y = \beta_0 + \beta_1 X^{\beta_2}$

is nonlinear in the parameters and variables both. So it is a nonlinear model.

vi) $y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$

is a cubic polynomial model which can be written as

$$y = \beta_0 + \beta_1 X + \beta_2 X_2 + \beta_3 X_3$$

which is linear in the parameters $\beta_0, \beta_1, \beta_2, \beta_3$ and linear in the variables $X_1 = X, X_2 = X^2, X_3 = X^3$.

So it is a linear model.

Example:

The income and education of a person are related. It is expected that, on average, a higher level of education provides higher income. So a simple linear regression model can be expressed as

$$\text{income} = \beta_0 + \beta_1 \text{ education} + \varepsilon.$$

Not that β_1 reflects the change in income with respect to per unit change in education and β_0 reflects the income when education is zero as it is expected that even an illiterate person can also have some income.

Further, this model neglects that most people have higher income when they are older than when they are young, regardless of education. So β_1 will over-state the marginal impact of education. If age and education are positively correlated, then the regression model will associate all the observed increase in income with an increase in education. So a better model is

$$\text{income} = \beta_0 + \beta_1 \text{ education} + \beta_2 \text{ age} + \varepsilon.$$

Often it is observed that the income tends to rise less rapidly in the later earning years than is early years. To accommodate such a possibility, we might extend the model to

$$\text{income} = \beta_0 + \beta_1 \text{education} + \beta_2 \text{age} + \beta_3 \text{age}^2 + \varepsilon$$

This is how we proceed for regression modeling in real-life situation. One needs to consider the experimental condition and the phenomenon before making the decision on how many, why and how to choose the dependent and independent variables.

Model set up:

Let an experiment be conducted n times, and the data is obtained as follows:

Observation number	Response y	Explanatory variables $X_1 \quad X_2 \quad \cdots \quad X_k$
1	y_1	$x_{11} \quad x_{12} \cdots x_{1k}$
2	y_2	$x_{21} \quad x_{22} \cdots x_{2k}$
$\vdots$	$\vdots$	$\vdots \quad \vdots \quad \ddots \quad \vdots$
n	y_n	$x_{n1} \quad x_{n2} \cdots x_{nk}$

Assuming that the model is

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon,$$

the n-tuples of observations are also assumed to follow the same model. Thus they satisfy

$$\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + ... + \beta_k x_{1k} + \varepsilon_1 \\
y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + ... + \beta_k x_{2k} + \varepsilon_2 \\
&\vdots \qquad\qquad\qquad\qquad \vdots \\
y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + ... + \beta_k x_{nk} + \varepsilon_n.
\end{aligned}$$

These n equations can be written as

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} \cdots x_{1k} \\ 1 & x_{21} & x_{22} \cdots x_{2k} \\ \vdots & \vdots & \vdots \quad \vdots \\ 1 & x_{n1} & x_{n2} \cdots x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

or $y = X\beta + \varepsilon.$

In general, the model with k explanatory variables can be expressed as

$$y = X\beta + \varepsilon$$

where $y = (y_1, y_2, ..., y_n)'$ is a $n \times 1$ vector of n observation on study variable,

$$X = \begin{pmatrix} x_{11} & x_{12} \cdots x_{1k} \\ x_{21} & x_{22} \cdots x_{2k} \\ \vdots & \vdots \quad \ddots \quad \vdots \\ x_{n1} & x_{n2} \cdots x_{nk} \end{pmatrix}$$

is a $n \times k$ matrix of n observations on each of the k explanatory variables, $\beta = (\beta_1, \beta_2, ..., \beta_k)'$ is a $k \times 1$ vector of regression coefficients and $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)'$ is a $n \times 1$ vector of random error components or disturbance term.

If intercept term is present, take first column of X to be (1,1,…,1)'.

Assumptions in multiple linear regression model

Some assumptions are needed in the model $y = X\beta + \varepsilon$ for drawing the statistical inferences. The following assumptions are made:

(i) $E(\varepsilon) = 0$

(ii) $E(\varepsilon\varepsilon') = \sigma^2 I_n$

(iii) $Rank(X) = k$

(iv) X is a non-stochastic matrix

(v) $\varepsilon \sim N(0, \sigma^2 I_n)$.

These assumptions are used to study the statistical properties of the estimator of regression coefficients. The following assumption is required to study, particularly the large sample properties of the estimators.

(vi) $\lim_{n\to\infty}\left(\frac{X'X}{n}\right) = \Delta$ exists and is a non-stochastic and nonsingular matrix (with finite elements).

The explanatory variables can also be stochastic in some cases. We assume that X is non-stochastic unless stated separately.

We consider the problems of estimation and testing of hypothesis on regression coefficient vector under the stated assumption.

Estimation of parameters:

A general procedure for the estimation of regression coefficient vector is to minimize

$$\sum_{i=1}^{n} M(\varepsilon_i) = \sum_{i=1}^{n} M(y_i - x_{i1}\beta_1 - x_{i2}\beta_2 - ... - x_{ik}\beta_k)$$

for a suitably chosen function M.

Some examples of choice of M are

$M(x) = |x|$

$M(x) = x^2$

$M(x) = |x|^p$, in general.

We consider the principle of least square which is related to $M(x) = x^2$ and method of maximum likelihood estimation for the estimation of parameters.

Principle of ordinary least squares (OLS)

Let B be the set of all possible vectors β. If there is no further information, the B is k-dimensional real Euclidean space. The object is to find a vector $b' = (b_1, b_2, ..., b_k)$ from B that minimizes the sum of squared deviations of ε_i's, i.e.,

$$S(\beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon'\varepsilon = (y - X\beta)'(y - X\beta)$$

for given y and X. A minimum will always exist as $S(\beta)$ is a real-valued, convex and differentiable function. Write

$$S(\beta) = y'y + \beta'X'X\beta - 2\beta'X'y .$$

Differentiate $S(\beta)$ with respect to β

$$\frac{\partial S(\beta)}{\partial \beta} = 2X'X\beta - 2X'y$$

$$\frac{\partial^2 S(\beta)}{\partial \beta^2} = 2X'X \quad \text{(atleast non-negative definite).}$$

The normal equation is

$$\frac{\partial S(\beta)}{\partial \beta} = 0$$

$$\Rightarrow X'Xb = X'y$$

where the following result is used:

Result: If $f(z) = Z'AZ$ is a quadratic form, Z is a $m \times 1$ vector and A is any $m \times m$ symmetric matrix then $\frac{\partial}{\partial z} F(z) = 2Az$.

Since it is assumed that $\text{rank}(X) = k$ (full rank), then $X'X$ is a positive definite and unique solution of the normal equation is

$$b = (X'X)^{-1}X'y$$

which is termed as **ordinary least squares estimator** (OLSE) of β.

Since $\frac{\partial^2 S(\beta)}{\partial \beta^2}$ is at least non-negative definite, so b minimize $S(\beta)$.

In case, X **is not of full rank**, then

$$b = (X'X)^{-}X'y + \left[I - (X'X)^{-}X'X\right]\omega$$

where $(X'X)^{-}$ is the generalized inverse of $X'X$ and ω is an arbitrary vector. The generalized inverse $(X'X)^{-}$ of $X'X$ satisfies

$$X'X(X'X)^{-}X'X = X'X$$
$$X(X'X)^{-}X'X = X$$
$$X'X(X'X)^{-}X' = X'$$

Theorem:

(i) Let $\hat{y} = Xb$ be the empirical predictor of y. Then $\hat{y}$ has the same value for all solutions b of $X'Xb = X'y.$

(ii) $S(\beta)$ attains the minimum for any solution of $X'Xb = X'y.$

Proof:

(i) Let b be any member in

$$b = (X'X)^{-}X'y + \left[I - (X'X)^{-}X'X\right]\omega.$$

Since $X(X'X)^{-}X'X = X,$ so then

$$Xb = X(X'X)^{-}X'y + X\left[I - (X'X)^{-}X'X\right]\omega$$
$$= X(X'X)^{-}X'y$$

which is independent of ω . This implies that $\hat{y}$ has the same value for all solution b of $X'Xb = X'y.$

(ii) Note that for any β,

$$\begin{aligned}
S(\beta) &= \left[y - Xb + X(b-\beta)\right]'\left[y - Xb + X(b-\beta)\right] \\
&= (y-Xb)'(y-Xb) + (b-\beta)'X'X(b-\beta) + 2(b-\beta)'X'(y-Xb) \\
&= (y-Xb)'(y-Xb) + (b-\beta)'X'X(b-\beta) \qquad (\text{Using } X'Xb = X'y) \\
&\geq (y-Xb)'(y-Xb) = S(b) \\
&= y'y - 2y'Xb + b'X'Xb \\
&= y'y - b'X'Xb \\
&= y'y - \hat{y}'\hat{y}.
\end{aligned}$$

Fitted values:

If $\hat{\beta}$ is any estimator of β for the model $y = X\beta + \varepsilon,$ then the fitted values are defined as

$\hat{y} = X\hat{\beta}$ where $\hat{\beta}$ is any estimator of β.

In the case of $\hat{\beta} = b,$

$$\begin{aligned}\hat{y} &= Xb \\ &= X(X'X)^{-1}X'y \\ &= Hy\end{aligned}$$

where $H = X(X'X)^{-1}X'$ is termed as **Hat matrix** which is

(i) symmetric

(ii) idempotent (i.e., $HH = H$) and

(iii) $tr\,H = tr\;X(X'X)^{-1}X' = tr\;X'X(X'X)^{-1} = tr\,I_k = k$.

Residuals

The difference between the observed and fitted values of the study variable is called as residual. It is denoted as

$$\begin{aligned}e &= y \sim \hat{y} \\ &= y - \hat{y} \\ &= y - Xb \\ &= y - Hy \\ &= (I - H)y \\ &= \bar{H}y\end{aligned}$$

where $\bar{H} = I - H$.

Note that

(i) $\bar{H}$ is a symmetric matrix

(ii) $\bar{H}$ is an idempotent matrix, i.e.,

$\bar{H}\bar{H} = (I - H)(I - H) = (I - H) = \bar{H}$ and

(iii) $tr\bar{H} = trI_n - trH = (n - k).$

Properties of OLSE

(i) Estimation error:

The estimation error of b is

$$\begin{aligned} b-\beta &= (X'X)^{-1}X'y-\beta \\ &= (X'X)^{-1}X'(X\beta+\varepsilon)-\beta \\ &= (X'X)^{-1}X'\varepsilon \end{aligned}$$

(ii) Bias

Since X is assumed to be nonstochastic and $E(\varepsilon)=0$

$$\begin{aligned} E(b-\beta) &= (X'X)^{-1}X'E(\varepsilon) \\ &= 0. \end{aligned}$$

Thus OLSE is an unbiased estimator of β.

(iii) Covariance matrix

The covariance matrix of b is

$$\begin{aligned} V(b) &= E(b-\beta)(b-\beta)' \\ &= E\left[(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}\right] \\ &= (X'X)^{-1}X'E(\varepsilon\varepsilon')X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}X'IX(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}. \end{aligned}$$

(iv) Variance

The variance of b can be obtained as the sum of variances of all $b_1, b_2, ..., b_k$ which is the trace of covariance matrix of b. Thus

$$\begin{aligned} Var(b) &= tr\left[V(b)\right] \\ &= \sum_{i=1}^{k} E(b_i-\beta_i)^2 \\ &= \sum_{i=1}^{k} Var(b_i). \end{aligned}$$

Estimation of σ^2

The least-squares criterion can not be used to estimate σ^2 because σ^2 does not appear in $S(\beta)$. Since $E(\varepsilon_i^2) = \sigma^2$, so we attempt with residuals e_i to estimate σ^2 as follows:

$$\begin{aligned} e &= y - \hat{y} \\ &= y - X(X'X)^{-1}X'y \\ &= [I - X(X'X)^{-1}X']y \\ &= \bar{H}y. \end{aligned}$$

Consider the residual sum of squares

$$\begin{aligned} SS_{res} &= \sum_{i=1}^{n} e_i^2 \\ &= e'e \\ &= (y - Xb)'(y - Xb) \\ &= y'(I - H)(I - H)y \\ &= y'(I - H)y \\ &= y'\bar{H}y. \end{aligned}$$

Also

$$\begin{aligned} SS_{res} &= (y - Xb)'(y - Xb) \\ &= y'y - 2b'X'y + b'X'Xb \\ &= y'y - b'X'y \quad (\text{Using } X'Xb = X'y) \end{aligned}$$

$$\begin{aligned} SS_{res} &= y'\bar{H}y \\ &= (X\beta + \varepsilon)'\bar{H}(X\beta + \varepsilon) \\ &= \varepsilon'\bar{H}\varepsilon \quad (\text{Using } \bar{H}X = 0) \end{aligned}$$

Since $\varepsilon \sim N(0, \sigma^2 I)$, so $y \sim N(X\beta, \sigma^2 I)$. Hence $y'\bar{H}y \sim \chi^2(n-k)$.

Thus $E[y'\bar{H}y] = (n-k)\sigma^2$

or $E\left[\dfrac{y'\bar{H}y}{n-k}\right] = \sigma^2$

or $E\left[MS_{res}\right] = \sigma^2$

where $MS_{res} = \dfrac{SS_{res}}{n-k}$ is the mean sum of squares due to residual.

Thus an unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = MS_{res} = s^2 \text{ (say)}$$

which is a model-dependent estimator.

Variance of $\hat{y}$

The variance of $\hat{y}$ is

$$\begin{aligned} V(\hat{y}) &= V(Xb) \\ &= XV(b)X' \\ &= \sigma^2 X(X'X)^{-1}X' \\ &= \sigma^2 H. \end{aligned}$$

Gauss-Markov Theorem:

The ordinary least squares estimator (OLSE) is the best linear unbiased estimator (BLUE) of β.

Proof: The OLSE of β is

$$b = (X'X)^{-1}X'y$$

which is a linear function of y. Consider the arbitrary linear estimator

$$b^* = a'y$$

of linear parametric function $\ell'\beta$ where the elements of a are arbitrary constants.

Then for b^*,

$$E(b^*) = E(a'y) = a'X\beta$$

and so b^* is an unbiased estimator of $\ell'\beta$ when

$$\begin{aligned} E(b^*) &= a'X\beta = \ell'\beta \\ &\Rightarrow a'X = \ell'. \end{aligned}$$

Since we wish to consider only those estimators that are linear and unbiased, so we restrict ourselves to those estimators for which $a'X = \ell'$.

Further

$$\begin{aligned} Var(a'y) &= a'Var(y)a = \sigma^2 a'a \\ Var(\ell'b) &= \ell'Var(b)\ell \\ &= \sigma^2 a'X(X'X)^{-1}X'a. \end{aligned}$$

Consider

$$\begin{aligned} Var(a'y) - Var(\ell'b) &= \sigma^2\left[a'a - a'X(X'X)^{-1}X'a\right] \\ &= \sigma^2 a'\left[I - X(X'X)^{-1}X'\right]a \\ &= \sigma^2 a'(I-H)a. \end{aligned}$$

Since $(I-H)$ is a positive semi-definite matrix, so

$$Var(a'y) - Var(\ell'b) \geq 0 .$$

This reveals that if b^* is any linear unbiased estimator then its variance must be no smaller than that of b. Consequently b is the best linear unbiased estimator, where 'best' refers to the fact that b is efficient within the class of linear and unbiased estimators.

Maximum likelihood estimation:

In the model, $y = X\beta + \varepsilon,$ it is assumed that the errors are normally and independently distributed with constant variance σ^2 or $\varepsilon \sim N(0, \sigma^2 I).$

The normal density function for the errors is

$$f(\varepsilon_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}\varepsilon_i^2\right] \quad i = 1, 2, ..., n..$$

The likelihood function is the joint density of $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ given as

$$\begin{aligned} L(\beta, \sigma^2) &= \prod_{i=1}^{n} f(\varepsilon_i) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\varepsilon_i^2\right] \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\varepsilon'\varepsilon\right] \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right]. \end{aligned}$$

Since the log transformation is monotonic, so we maximize $\ln L(\beta, \sigma^2)$ instead of $L(\beta, \sigma^2)$.

$$\ln\ L(\beta, \sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta).$$

The maximum likelihood estimators (m.l.e.) of β and σ^2 are obtained by equating the first-order derivatives of $\ln L(\beta, \sigma^2)$ with respect to β and σ^2 to zero as follows:

$$\frac{\partial \ln L(\beta, \sigma^2)}{\partial \beta} = \frac{1}{2\sigma^2} 2X'(y - X\beta) = 0$$

$$\frac{\partial \ln L(\beta, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(y - X\beta)'(y - X\beta).$$

The likelihood equations are given by

$$X'X\beta = X'y$$
$$\sigma^2 = \frac{1}{n}(y - X\beta)'(y - X\beta).$$

Since $\text{rank}(X) = k$, so that the unique m.l.e. of β and σ^2 are obtained as

$$\tilde{\beta} = (X'X)^{-1}X'y$$
$$\tilde{\sigma}^2 = \frac{1}{n}(y - X\tilde{\beta})'(y - X\tilde{\beta}).$$

Further to verify that these values maximize the likelihood function, we find

$$\frac{\partial^2 \ln L(\beta,\sigma^2)}{\partial\beta^2} = -\frac{1}{\sigma^2}X'X$$
$$\frac{\partial^2 \ln L(\beta,\sigma^2)}{\partial^2(\sigma^2)^2} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}(y - X\beta)'(y - X\beta)$$
$$\frac{\partial^2 \ln L(\beta,\sigma^2)}{\partial\beta\partial\sigma^2} = -\frac{1}{\sigma^4}X'(y - X\beta).$$

Thus the Hessian matrix of second-order partial derivatives of $\ln L(\beta,\sigma^2)$ with respect to β and σ^2 is

$$\begin{pmatrix} \frac{\partial^2 \ln L(\beta,\sigma^2)}{\partial\beta^2} & \frac{\partial^2 \ln L(\beta,\sigma^2)}{\partial\beta\partial\sigma^2} \\ \frac{\partial^2 \ln L(\beta,\sigma^2)}{\partial\sigma^2\partial\beta} & \frac{\partial^2 \ln L(\beta,\sigma^2)}{\partial^2(\sigma^2)^2} \end{pmatrix}$$

which is negative definite at $\beta = \tilde{\beta}$ and $\sigma_2 = \tilde{\sigma}^2$. This ensures that the likelihood function is maximized at these values.

Comparing with OLSEs, we find that

(i) OLSE and m.l.e. of β are same. So m.l.e. of β is also an unbiased estimator of β.

(ii) OLSE of σ^2 is s^2 which is related to m.l.e. of σ^2 as $\tilde{\sigma}^2 = \frac{n-k}{n}s^2$. So m.l.e. of σ^2 is a biased estimator of σ^2.

Consistency of estimators

(i) Consistency of b :

Under the assumption that $\lim_{n\to\infty}\left(\frac{X'X}{n}\right)=\Delta$ exists as a nonstochastic and nonsingular matrix (with finite elements), we have

$$\begin{aligned}\lim_{n\to\infty}V(b) &= \sigma^2 \lim_{n\to\infty}\frac{1}{n}\left(\frac{X'X}{n}\right)^{-1} \\ &= \sigma^2 \lim_{n\to\infty}\frac{1}{n}\Delta^{-1} \\ &= 0.\end{aligned}$$

This implies that OLSE converges to β in quadratic mean. Thus OLSE is a consistent estimator of β. This holds true for maximum likelihood estimators also.

The same conclusion can also be proved using the concept of convergence in probability.

An estimator $\hat{\theta}_n$ converges to θ in probability if

$$\lim_{n\to\infty}P\left[\left|\hat{\theta}_n-\theta\right|\geq\delta\right]=0 \quad \text{for any} \quad \delta>0$$

and is denoted as $\text{plim}(\hat{\theta}_n)=\theta.$

The consistency of OLSE can be obtained under the weaker assumption that

$$\text{plim}\left(\frac{X'X}{n}\right)=\Delta_*.$$

exists and is a nonsingular and nonstochastic matrix such that

$$\text{plim}\left(\frac{X'\varepsilon}{n}\right)=0.$$

Since

$$\begin{aligned}b-\beta &= (X'X)^{-1}X'\varepsilon \\ &= \left(\frac{X'X}{n}\right)^{-1}\frac{X'\varepsilon}{n}.\end{aligned}$$

So

$$\begin{aligned}\text{plim}(b-\beta) &= \text{plim}\left(\frac{X'X}{n}\right)^{-1}\text{plim}\left(\frac{X'\varepsilon}{n}\right) \\ &= \Delta_*^{-1}.0 \\ &= 0.\end{aligned}$$

Thus b is a consistent estimator of β. Same is true for m.l.e. also.

(ii) Consistency of s^2

Now we look at the consistency of s^2 as an estimate of σ^2 as

$$\begin{aligned}
s^2 &= \frac{1}{n-k} e'e \\
&= \frac{1}{n-k} \varepsilon' \bar{H} \varepsilon \\
&= \frac{1}{n}\left(1-\frac{k}{n}\right)^{-1}\left[\varepsilon'\varepsilon - \varepsilon' X (X'X)^{-1} X'\varepsilon\right] \\
&= \left(1-\frac{k}{n}\right)^{-1}\left[\frac{\varepsilon'\varepsilon}{n} - \frac{\varepsilon' X}{n}\left(\frac{X'X}{n}\right)^{-1}\frac{X'\varepsilon}{n}\right].
\end{aligned}$$

Note that $\frac{\varepsilon'\varepsilon}{n}$ consists of $\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2$ and $\{\varepsilon_i^2, i = 1,2,...,n\}$ is a sequence of independently and identically distributed random variables with mean σ^2. Using the law of large numbers

$$\begin{aligned}
&\operatorname{plim}\left(\frac{\varepsilon'\varepsilon}{n}\right) = \sigma^2 \\
&\operatorname{plim}\left[\frac{\varepsilon' X}{n}\left(\frac{X'X}{n}\right)^{-1}\frac{X'\varepsilon}{n}\right] = \left(\operatorname{plim}\frac{\varepsilon'\mathrm{X}}{\mathrm{n}}\right)\left[\operatorname{plim}\left(\frac{X'X}{n}\right)^{-1}\right]\left(\operatorname{plim}\frac{X'\varepsilon}{n}\right) \\
&\qquad\qquad = 0.\Delta_*^{-1}.0 \\
&\qquad\qquad = 0 \\
&\Rightarrow \operatorname{plim}(s^2) = (1-0)^{-1}\left[\sigma^2 - 0\right] \\
&\qquad\qquad = \sigma^2.
\end{aligned}$$

Thus s^2 is a consistent estimator of σ^2. The same holds true for m.l.e. also.

Cramer-Rao lower bound

Let $\theta = (\beta, \sigma^2)'$. Assume that both β and σ^2 are unknown. If $E(\hat{\theta}) = \theta$, then the Cramer-Rao lower bound for $\hat{\theta}$ is grater than or equal to the matrix inverse of

$$
\begin{aligned}
I(\theta) &= -E\left[\frac{\partial^2 \ln L(\theta)}{\partial\theta\partial\theta'}\right] \\
&= \begin{bmatrix} -E\left[\frac{\partial \ln L(\beta,\sigma^2)}{\partial\beta^2}\right] & -E\left[\frac{\partial \ln L(\beta,\sigma^2)}{\partial\beta\partial\sigma^2}\right] \\ -E\left[\frac{\partial \ln L(\beta,\sigma^2)}{\partial\sigma^2\partial\beta}\right] & -E\left[\frac{\partial \ln L(\beta,\sigma^2)}{\partial^2(\sigma^2)^2}\right] \end{bmatrix} \\
&= \begin{bmatrix} -E\left[-\frac{X'X}{\sigma^2}\right] & -E\left[\frac{X'(y-X\beta)}{\sigma^4}\right] \\ -E\left[\frac{(y-X\beta)'X}{\sigma^4}\right] & -E\left[\frac{n}{2\sigma^4}-\frac{(y-X\beta)'(y-X\beta)}{\sigma^6}\right] \end{bmatrix} \\
&= \begin{bmatrix} \frac{X'X}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}.
\end{aligned}
$$

Then

$$
\left[I(\theta)\right]^{-1} = \begin{bmatrix} \sigma^2(X'X)^{-1} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}
$$

is the Cramer-Rao lower bound matrix of β and σ^2.

The covariance matrix of OLSEs of β and σ^2 is

$$
\sum_{OLS} = \begin{bmatrix} \sigma^2(X'X)^{-1} & 0 \\ 0 & \frac{2\sigma^4}{n-k} \end{bmatrix}
$$

which means that the Cramer-Rao have bound is attained for the covariance of b but not for s^2.

Standardized regression coefficients:

Usually, it is difficult to compare the regression coefficients because the magnitude of $\hat{\beta}_j$ reflects the units of measurement of j^{th} explanatory variable X_j. For example, in the following fitted regression model

$$\hat{y} = 5 + X_1 + 1000X_2,$$

y is measured in litres, X_1 in litres and X_2 in millilitres. Although $\hat{\beta}_2 >> \hat{\beta}_1$ but the effect of both explanatory variables is identical. One litre change in either X_1 and X_2 when another variable is held fixed produces the same change is $\hat{y}$.

Sometimes it is helpful to work with scaled explanatory variables and study variable that produces dimensionless regression coefficients. These dimensionless regression coefficients are called as **standardized regression coefficients.**

There are two popular approaches for scaling, which gives standardized regression coefficients. We discuss them as follows:

1. Unit normal scaling:

Employ unit normal scaling to each explanatory variable and study variable. So define

$$z_{ij} = \frac{x_{ij} - \overline{x}_j}{s_j}, \; i = 1,2,...,n, \; j = 1,2,...,k$$

$$y_i^* = \frac{y_i - \overline{y}}{s_y}$$

where $s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2$ and $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$ are the sample variances of j^{th} explanatory variable and study variable, respectively.

All scaled explanatory variable and scaled study variable has mean zero and sample variance unity, i.e., using these new variables, the regression model becomes

$$y_i^* = \gamma_1 z_{i1} + \gamma_2 z_{i2} + ... + \gamma_k z_{ik} + \varepsilon_i, \; i = 1,2,...,n.$$

Such centering removes the intercept term from the model. The least-squares estimate of $\gamma = (\gamma_1, \gamma_2, ..., \gamma_k)'$ is

$$\hat{\gamma} = (Z'Z)^{-1}Z'y^*.$$

This scaling has a similarity to standardizing a normal random variable, i.e., observation minus its mean and divided by its standard deviation. So it is called as a unit normal scaling.

2. Unit length scaling:

In unit length scaling, define

$$\omega_{ij} = \frac{x_{ij} - \overline{x}_j}{S_{jj}^{1/2}}, \; i = 1,2,...,n; \; j = 1,2,...,k$$

$$y_i^0 = \frac{y_i - \overline{y}}{SS_T^{1/2}}$$

where $S_{jj} = \sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2$ is the corrected sum of squares for j^{th} explanatory variables X_j and $S_T = SS_T = \sum_{i=1}^{n}(y_i - \overline{y})^2$ is the total sum of squares. In this scaling, each new explanatory variable W_j has mean $\overline{\omega}_j = \frac{1}{n}\sum_{i=1}^{n}\omega_{ij} = 0$ and length $\sqrt{\sum_{i=1}^{n}(\omega_{ij} - \overline{\omega}_j)^2} = 1.$

In terms of these variables, the regression model is

$$y_i^o = \delta_1\omega_{i1} + \delta_2\omega_{i2} + ... + \delta_k\omega_{ik} + \varepsilon_i, \; i = 1,2,...,n.$$

The least-squares estimate of the regression coefficient $\delta = (\delta_1, \delta_2, ..., \delta_k)'$ is

$$\hat{\delta} = (W'W)^{-1}W'y^0.$$

In such a case, the matrix $W'W$ is in the form of the correlation matrix, i.e.,

$$W'W = \begin{pmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1k} \\ r_{12} & 1 & r_{23} & \cdots & r_{2k} \\ r_{13} & r_{23} & 1 & \cdots & r_{3k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{1k} & r_{2k} & r_{3k} & \cdots & 1 \end{pmatrix}$$

where

$$r_{ij} = \frac{\sum_{u=1}^{n}(x_{ui} - \overline{x}_i)(x_{uj} - \overline{x}_j)}{(S_{ii}S_{jj})^{1/2}} = \frac{S_{ij}}{(S_{ii}S_{jj})^{1/2}}$$

is the simple correlation coefficient between the explanatory variables X_i and X_j. Similarly

$$W'y^o = (r_{1y}, r_{2y}, ..., r_{ky})'$$

where

$$r_{jy} = \frac{\sum_{u=1}^{n}(x_{uj} - \overline{x}_j)(y_u - \overline{y})}{(S_{jj}SS_T)^{1/2}} = \frac{S_{iy}}{(S_{jj}SS_T)^{1/2}}$$

is the simple correlation coefficient between the j^{th} explanatory variable X_j and study variable y.

Note that it is customary to refer r_{ij} and r_{jy} as correlation coefficient though X_i's are not random variable.

If unit normal scaling is used, then

$$Z'Z = (n-1)W'W.$$

So the estimates of regression coefficient in unit normal scaling (i.e., $\hat{\gamma}$) and unit length scaling (i.e., $\hat{\delta}$) are identical. So it does not matter which scaling is used, so $\hat{\gamma} = \hat{\delta}$.

The regression coefficients obtained after such scaling, viz., $\hat{\gamma}$ or $\hat{\delta}$ usually called standardized regression coefficients.

The relationship between the original and standardized regression coefficients is

$$b_j = \hat{\delta}_j \left(\frac{SS_T}{S_{jj}} \right)^{1/2}, \quad j = 1, 2, ..., k$$

and

$$b_0 = \overline{y} - \sum_{j=1}^{k} b_j \overline{x}_j$$

where b_0 is the OLSE of intercept term and b_j are the OLSE of slope parameters.

The model in deviation form

The multiple linear regression model can also be expressed in the deviation form.

First, all the data is expressed in terms of deviations from the sample mean.

The estimation of regression parameters is performed in two steps:

- **First step:** Estimate the slope parameters.
- **Second step** : Estimate the intercept term.

The multiple linear regression model in deviation form is expressed as follows:

Let

$$A = I - \frac{1}{n} \ell\ell'$$

where $\ell = (1, 1, ..., 1)'$ is a $n \times 1$ vector of each element unity. So

$$A=\begin{bmatrix}1 & 0 & \cdots & 0\\ 0 & 1 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 1\end{bmatrix}-\frac{1}{n}\begin{bmatrix}1 & 1 & \cdots & 1\\ 1 & 1 & \cdots & 1\\ \vdots & \vdots & \ddots & \vdots\\ 1 & 1 & \cdots & 1\end{bmatrix}.$$

Then

$$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i=\frac{1}{n}(1,1,...,1)\begin{pmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{pmatrix}=\frac{1}{n}\ell' y$$

$$Ay=y-\ell\bar{y}=\left(y_1-\bar{y},y_2-\bar{y},...,y_n-\bar{y}\right)'.$$

Thus pre-multiplication of any column vector by A produces a vector showing those observations in deviation form:

Note that

$$\begin{aligned}A\ell&=\ell-\frac{1}{n}\ell\ell'\ell\\ &=\ell-\frac{1}{n}\ell.n\\ &=\ell-\ell\\ &=0\end{aligned}$$

and A is a symmetric and idempotent matrix.

In the model

$$y=X\beta+\varepsilon,$$

the OLSE of β is

$$b=\left(X'X\right)^{-1}X'y$$

and the residual vector is

$$e=y-Xb.$$

Note that $Ae=e.$

If the $n\times k$ matrix is partitioned as

$$X=\begin{bmatrix}X_1 & X_2^*\end{bmatrix}$$

where $X_1 = (1,1,...,1)'$ is $n \times 1$ vector with all elements unity, X_2^* is $n \times (k-1)$ matrix of observations of $(k-1)$ explanatory variables $X_2, X_3, ..., X_k$ and OLSE $b = (b_1, b_2^{*\prime})$ is suitably partitioned with OLSE of intercept term β_1 as b_1 and b_2 as a $(k-1) \times 1$ vector of OLSEs associated with $\beta_2, \beta_3, ..., \beta_k$.

Then

$$y = X_1 b_1 + X_2^* b_2^* + e.$$

Premultiply by A,

$$\begin{aligned} Ay &= AX_1 b_1 + AX_2^* b_2^* + Ae \\ &= AX_2^* b_2^* + e. \end{aligned}$$

Premultiply by X_2^* gives

$$\begin{aligned} X_2^{*\prime} Ay &= X_2^{*\prime} AX_2^* b_2^* + X_2^{*\prime} e \\ &= X_2^{*\prime} AX_2^* b_2^*. \end{aligned}$$

Since A is symmetric and idempotent,

$$(AX_2^*)'(Ay) = (AX_2^*)'(AX_2^*) b_2^* \; ..$$

This equation can be compared with the normal equations $X'y = X'Xb$ in the model $y = X\beta + \varepsilon$. Such a comparison yields the following conclusions:

- b_2^* is the sub vector of OLSE.
- Ay is the study variables vector in deviation form.
- AX_2^* is the explanatory variable matrix in deviation form.
- This is the normal equation in terms of deviations. Its solution gives OLS of slope coefficients as

$$b_2^* = \left[(AX_2^*)'(AX_2^*) \right]^{-1} (AX_2^*)'(Ay).$$

The estimate of the intercept term is obtained in the second step as follows:

Premultiplying $y = Xb + e$ by $\frac{1}{n}\ell'$ gives

$$\frac{1}{n}\ell' y = \frac{1}{n}\ell' Xb + \frac{1}{n}\ell' e$$

$$\bar{y} = \begin{bmatrix} 1 & \bar{X}_2 & \bar{X}_3 & \dots & \bar{X}_k \end{bmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{pmatrix} + 0$$

$$\Rightarrow b_1 = \bar{y} - b_2\bar{X}_2 - b_3\bar{X}_3 - \dots - b_k\bar{X}_k.$$

Now we explain various sums of squares in terms of this model.

The expression of the total sum of squares (TSS) remains the same as earlier and is given by

$$TSS = y'Ay.$$

Since

$$Ay = AX_2^*b_2^* + e$$

$$y'Ay = y'AX_2^*b_2^* + y'e$$

$$= \left(Xb + e\right)'AX_2^*b_2^* + y'e$$

$$= \left(X_1b_1 + X_2^*b_2^* + e\right)'AX_2^*b_2^* + \left(X_1b_1 + X_2^*b_2^* + e\right)'e$$

$$= b_2^{*\prime}X_2^{*\prime}AX_2^*b_2^* + e'e$$

$$TSS = SS_{reg} + SS_{res}$$

where the sum of squares due to regression is

$$SS_{reg} = b_2^{*\prime}X_2^{*\prime}AX_2^*b_2^*$$

and the sum of squares due to residual is

$$SS_{res} = e'e.$$

Testing of hypothesis:

There are several important questions which can be answered through the test of hypothesis concerning the regression coefficients. For example

1. What is the overall adequacy of the model?
2. Which specific explanatory variables seem to be important?

etc.

In order the answer such questions, we first develop the test of hypothesis for a general framework, viz., general linear hypothesis. Then several tests of hypothesis can be derived as its special cases. So first, we discuss the test of a general linear hypothesis.

Test of hypothesis for $H_0 : R\beta = r$

We consider a general linear hypothesis that the parameters in β are contained in a subspace of parameter space for which $R\beta = r$, where R is $(J \times k)$ a matrix of known elements and r is a $(J \times 1)$ vector of known elements.

In general, the null hypothesis

$$H_0 : R\beta = r$$

is termed as general linear hypothesis and

$$H_1 : R\beta \neq r$$

is the alternative hypothesis.

We assume that $\text{rank}(R) = J$, i.e., full rank so that there is no linear dependence in the hypothesis.

Some special cases and interesting example of $H_0 : R\beta = r$ are as follows:

(i) $H_0 : \beta_i = 0$

Choose $J = 1,\ r = 0,\ R = [0,0,...,0,1,0,...,0]$ where 1 occurs at the i^{th} position is R.

This particular hypothesis explains whether X_i has any effect on the linear model or not.

(ii) $H_0 : \beta_3 = \beta_4$ or $H_0 : \beta_3 - \beta_4 = 0$

Choose $J = 1, r = 0, R = [0,0,1,-1,0,...,0]$

(iii) $H_0 : \beta_3 = \beta_4 = \beta_5$

or $H_0 : \beta_3 - \beta_4 = 0,\ \beta_3 - \beta_5 = 0$

Choose $J = 2, r = (0,0)',\ R = \begin{bmatrix} 0 & 0 & 1 & -1 & 0 & 0 & ... & 0 \\ 0 & 0 & 1 & 0 & -1 & 0 & ... & 0 \end{bmatrix}$.

(iv) $H_0 : \beta_3 + 5\beta_4 = 2$

Choose $J = 1, r = 2, R = [0,0,1,5,0...0]$

(v) $H_0 : \beta_2 = \beta_3 = ... = \beta_k = 0$

$J = k - 1$

$r = (0,0,...,0)'$

$$R = \begin{bmatrix} 0 & 1 & 0 & ... & 0 \\ 0 & 0 & 1 & ... & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & ... & 1 \end{bmatrix}_{(k-1)\times k} = \begin{bmatrix} 0 & \\ 0 & I_{k-1} \\ \vdots & \\ 0 & \end{bmatrix}.$$

This particular hypothesis explains the goodness of fit. It tells whether β_i has a linear effect or not and are they of any importance. It also tests that $X_2, X_3, ..., X_k$ have no influence in the determination of y . Here $\beta_1 = 0$ is excluded because this involves additional implication that the mean level of y is zero. Our main concern is to know whether the explanatory variables help to explain the variation in y around its mean value or not.

We develop the likelihood ratio test for $H_0 : R\beta = r.$

Likelihood ratio test:

The likelihood ratio test statistic is

$$\lambda = \frac{\max L(\beta, \sigma^2 \mid y, X)}{\max L(\beta, \sigma^2 \mid y, X, R\beta = r)} = \frac{\hat{L}(\Omega)}{\hat{L}(\omega)}$$

where Ω is the whole parametric space and ω is the sample space.

If both the likelihoods are maximized, one constrained, and the other unconstrained, then the value of the unconstrained will not be smaller than the value of the constrained. Hence $\lambda \geq 1.$

First, we discuss the likelihood ratio test for a more straightforward case when $R = I_k$ and $r = \beta_0$, *i.e.*, $\beta = \beta_0$. This will give us a better and detailed understanding of the minor details, and then we generalize it for $R\beta = r$, in general.

Likelihood ratio test for $H_0 : \beta = \beta_0$

Let the null hypothesis related to $k \times 1$ vector β is

$$H_0 : \beta = \beta_0$$

where β_0 is specified by the investigator. The elements of β_0 can take on any value, including zero. The concerned alternative hypothesis is

$H_1 : \beta \neq \beta_0.$

Since $\varepsilon \sim N(0, \sigma^2 I)$ in $y = X\beta + \varepsilon$, so $y \sim N(X\beta, \sigma^2 I).$ Thus the whole parametric space and sample space are Ω and ω respectively given by

$$\Omega : \left\{(\beta, \sigma^2) : -\infty < \beta_i < \infty,\ \sigma^2 > 0,\ i = 1, 2, ..., k\right\}$$

$$\omega : \left\{(\beta, \sigma^2) : \beta = \beta_0, \sigma^2 > 0\right\}.$$

The unconstrained likelihood under Ω.

$$L(\beta,\sigma^2 \mid y,X)=\frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left[-\frac{1}{2\sigma^2}(y-X\beta)'(y-X\beta)\right].$$

This is maximized over Ω when

$$\tilde{\beta}=(X'X)^{-1}X'y$$

$$\tilde{\sigma}^2=\frac{1}{n}(y-X\tilde{\beta})'(y-X\tilde{\beta}).$$

where $\tilde{\beta}$ and $\tilde{\sigma}^2$ are the maximum likelihood estimates of β and σ^2 which are the values maximizing the likelihood function.

$$\hat{L}(\Omega)=\max L\left(\beta,\sigma^2 \mid y,X\right)$$

$$=\frac{1}{\left[\frac{2\pi}{n}(y-X\tilde{\beta})'(y-X\tilde{\beta})\right]^{\frac{n}{2}}}\exp\left[-\frac{(y-X\tilde{\beta})'(y-X\tilde{\beta})}{\left(\frac{2(y-X\tilde{\beta})'(y-X\tilde{\beta})}{n}\right)}\right]$$

$$=\frac{n^{n/2}\exp\left(-\frac{n}{2}\right)}{(2\pi)^{n/2}\left[(y-X\tilde{\beta})'(y-X\tilde{\beta})\right]^{\frac{n}{2}}}.$$

The constrained likelihood under ω is

$$\hat{L}(\omega)=\max L(\beta,\sigma^2 \mid y,X,\beta=\beta_0)=\frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left[-\frac{1}{2\sigma^2}(y-X\beta_0)'(y-X\beta_0)\right].$$

Since β_0 is known, so the constrained likelihood function has an optimum variance estimator

$$\tilde{\sigma}_\omega^2=\frac{1}{n}(y-X\beta_0)'(y-X\beta_0)$$

$$\hat{L}(\omega)=\frac{n^{n/2}\exp\left(-\frac{n}{2}\right)}{(2\pi)^{n/2}\left[(y-X\beta_0)'(y-X\beta_0^{)}\right]^{n/2}}.$$

The likelihood ratio is

$$\frac{\hat{L}(\Omega)}{\hat{L}(\omega)} = \frac{\left(\frac{n^{n/2}\exp(-n/2)}{(2\pi)^{n/2}\left[(y-X\tilde{\beta})'(y-X\tilde{\beta})\right]^{n/2}} \right)}{\left(\frac{n^{n/2}\exp(-n/2)}{(2\pi)^{n/2}\left[(y-X\tilde{\beta}_0)'(y-X\tilde{\beta}_0)\right]^{n/2}} \right)}$$

$$= \left[\frac{(y-X\beta_0)'(y-X\beta_0)}{(y-X\tilde{\beta})'(y-X\tilde{\beta})} \right]^{n/2}$$

$$= \left(\frac{\tilde{\sigma}_\omega^2}{\tilde{\sigma}^2} \right)^{n/2} = (\lambda)^{n/2}$$

where $\lambda = \dfrac{(y-X\beta_0)'(y-X\beta_0)}{(y-X\tilde{\beta})'(y-X\tilde{\beta})}$ is the ratio of the quadratic forms.

Now we simplify the numerator in λ as follows:

$$(y-X\beta_0)'(y-X\beta_0) = \left[(y-X\tilde{\beta}) + X(\tilde{\beta}-\beta_0)\right]' \left[(y-X\tilde{\beta}) + X(\tilde{\beta}-\beta_0)\right]$$

$$= (y-X\tilde{\beta})'(y-X\tilde{\beta}) + 2y'\left[I - X(X'X)^{-1}X'\right]X(\tilde{\beta}-\beta_0) + (\tilde{\beta}-\beta_0)'X'X(\tilde{\beta}-\beta_0)$$

$$= (y-X\tilde{\beta})'(y-X\tilde{\beta}) + (\tilde{\beta}-\beta_0)'X'X(\tilde{\beta}-\beta_0).$$

Thus

$$\lambda = \frac{(y-X\tilde{\beta})'(y-X\tilde{\beta}) + (\tilde{\beta}-\beta_0)'X'X(\tilde{\beta}-\beta_0)}{(y-X\tilde{\beta})'(y-X\tilde{\beta})}$$

$$= 1 + \frac{(\tilde{\beta}-\beta_0)'X'X(\tilde{\beta}-\beta_0)}{(y-X\tilde{\beta})'(y-X\tilde{\beta})}$$

or $\lambda - 1 = \lambda_0 = \dfrac{(\tilde{\beta}-\beta_0)'X'X(\tilde{\beta}-\beta_0)}{(y-X\tilde{\beta})'(y-X\tilde{\beta})}$

where $0 \le \lambda_0 < \infty$.

Distribution of ratio of quadratic forms

Now we find the distribution of the quadratic forms involved is λ_0 to find the distribution of λ_0 as follows:

$$(y-X\tilde{\beta})'(y-X\tilde{\beta}) = \tilde{e}'\tilde{e}$$

$$= y'\left[I - X(X'X)^{-1}X'\right]y$$

$$= y'\bar{H}y$$

$$= (X\beta+\varepsilon)'\bar{H}(X\beta+\varepsilon)$$

$$= \varepsilon'\bar{H}\varepsilon \qquad (\text{using } \bar{H}X = 0)$$

$$= (n-k)\hat{\sigma}^2$$

Result: If Z is a $n\times 1$ random vector that is distributed as $N(0,\sigma^2 I_n)$ and A is any symmetric idempotent $n\times n$ matrix of rank, p then $\frac{Z'AZ}{\sigma^2}\sim\chi^2(p)$. If B is another $n\times n$ symmetric idempotent matrix of rank q, then $\frac{Z'BZ}{\sigma^2}\sim\chi^2(q)$. If $AB=0$ then $Z'AZ$ is distributed independently of $Z'BZ$.

So using this result, we have

$$\frac{y'\bar{H}y}{\sigma^2}=\frac{(n-k)\hat{\sigma}^2}{\sigma^2}\sim\chi^2(n-k).$$

Further, if H_0 is true, then $\beta=\beta_0$ and we have the numerator in λ_0. Rewriting the numerator in λ_0, in general, we have

$$\begin{aligned}(\tilde{\beta}-\beta)'X'X(\tilde{\beta}-\beta)&=\varepsilon'X(X'X)^{-1}X'X(X'X)^{-1}X'\varepsilon\\&=\varepsilon'X(X'X)^{-1}X'\varepsilon\\&=\varepsilon'H\varepsilon\end{aligned}$$

where H is an idempotent matrix with rank k. Thus using this result, we have

$$\frac{\varepsilon'H\varepsilon}{\sigma^2}=\frac{\varepsilon'X'(X'X)^{-1}X'\varepsilon}{\sigma^2}\sim\chi^2(k).$$

Furthermore, the product of the quadratic form matrices in the numerator $(\varepsilon'\bar{H}\varepsilon)$ and denominator $(\varepsilon'H\varepsilon)$ of λ_0 is

$$\left[I-X(X'X)^{-1}X'\right]X(X'X)^{-1}X'=X(X'X)^{-1}X'-X(X'X)^{-1}X'X(X'X)^{-1}X'=0$$

and hence the χ^2 random variables in the numerator and denominator of λ_0 are independent. Dividing each of the χ^2 random variables by their respective degrees of freedom

$$\begin{aligned}\lambda_1&=\left(\frac{\frac{(\tilde{\beta}-\beta_0)'X'X(\tilde{\beta}-\beta_0)}{\sigma^2}}{\left(\frac{\frac{(n-k)\hat{\sigma}^2}{\sigma^2}}{n-k}\right)}\right)\\&=\frac{(\tilde{\beta}-\beta_0)'X'X(\tilde{\beta}-\beta_0)}{k\hat{\sigma}^2}\\&=\frac{(y-X\beta_0)'(y-X\beta_0)-(y-X\tilde{\beta})'(y-X\tilde{\beta})}{k\hat{\sigma}^2}\\&\sim F(k,n-k)\text{ under }H_0.\end{aligned}$$

Note that

$(y - X\beta_0)'(y - X\beta_0)$: Restricted error sum of squares

$(y - X\tilde{\beta})'(y - X\tilde{\beta})$: Unrestricted error sum of squares

Numerator in λ_1 : Difference between the restricted and unrestricted error sum of squares.

The decision rule is to reject $H_0 : \beta = \beta_0$ at α level of significance whenever

$$\lambda_1 \geq F_\alpha(k, n-k)$$

where $F_\alpha(k, n-k)$ is the upper critical points on the central F -distribution with k and $n-k$ degrees of freedom.

Likelihood ratio test for $H_0 : R\beta = r$

The same logic and reasons used in the development of the likelihood ratio test for $H_0 : \beta = \beta_0$ can be extended to develop the likelihood ratio test for $H_0 : R\beta = r$ as follows.

$$\Omega = \left\{(\beta, \sigma^2) : -\infty < \beta_i < \infty,\ \sigma^2 > 0,\ i = 1, 2, ..., k\right\}$$
$$\omega = \left\{(\beta, \sigma^2) : -\infty < \beta_i < \infty,\ R\beta = r,\ \sigma^2 > 0\right\}.$$

Let $\tilde{\beta} = (X'X)^{-1}X'y..$

Then

$$E(R\tilde{\beta}) = R\beta$$
$$V(R\tilde{\beta}) = E\left[R(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)'R'\right]$$
$$= RV(\tilde{\beta})R'$$
$$= \sigma^2\ R(X'X)^{-1}R'.$$

Since $\tilde{\beta} \sim N\left[\beta, \sigma^2(X'X)^{-1}\right]$

so $R\tilde{\beta} \sim N\left[R\beta, \sigma^2 R(X'X)^{-1}R'\right]$

$$R\tilde{\beta} - r = R\tilde{\beta} - R\beta = R(\tilde{\beta} - \beta) \sim N\left[0, \sigma^2 R(X'X)^{-1}R'\right].$$

There exists a matrix Q such that $\left[R(X'X)^{-1}R'\right]^{-1} = QQ'$ and then

$\xi = QR(b - \beta) \sim N(0, \sigma^2 I_n)$. Therefore under $H_0 : R\beta - r = 0$, so

$$\frac{\xi\xi'}{\sigma^2} = \frac{(R\tilde{\beta}-r)'QQ'(R\tilde{\beta}-r)}{\sigma^2}$$
$$= \frac{(R\tilde{\beta}-r)'\left[R(X'X)^{-1}R'\right]^{-1}(R\tilde{\beta}-r)}{\sigma^2}$$
$$= \frac{(\tilde{\beta}-\beta)'R'\left[R(X'X)^{-1}R'\right]^{-1}R(\tilde{\beta}-\beta)}{\sigma^2}$$
$$= \frac{\varepsilon'X(X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}R(X'X)^{-1}X'\varepsilon}{\sigma^2}$$
$$\sim \chi^2(J).$$

which is obtained as $X(X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}R(X'X)^{-1}X'$ is an idempotent matrix, and its trace is J which is the associated degrees of freedom.

Also, irrespective of whether H_0 is true or not,

$$\frac{\tilde{e}'\tilde{e}}{\sigma^2} = \frac{(y-X\tilde{\beta})'(y-X\tilde{\beta})}{\sigma^2} = \frac{y'\bar{H}y}{\sigma^2} = \frac{(n-k)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-k).$$

Moreover, the product of quadratic form matrices of $\tilde{e}'\tilde{e}$ and $(\tilde{\beta}-\beta)'R'\left[R(X'X)^{-1}R'\right]^{-1}R(\tilde{\beta}-\beta)$ is zero implying that both the quadratic forms are independent. So in terms of likelihood ratio test statistic

$$\lambda_1 = \frac{\left(\frac{(R\tilde{\beta}-r)'\left[R(X'X)^{-1}R'\right]^{-1}(R\tilde{\beta}-r)}{\frac{\sigma^2}{J}}\right)}{\frac{\left(\frac{(n-k)\hat{\sigma}^2}{\sigma^2}\right)}{n-k}}$$
$$= \frac{\left(R\tilde{\beta}-r)'\left[R(X'X)^{-1}R'\right]^{-1}\left(R\tilde{\beta}-r\right)\right)}{J\hat{\sigma}^2}$$
$$\sim F(J,n-k) \text{ under } H_0.$$

So the decision rule is to reject H_0 whenever

$$\lambda_1 \geq F_\alpha(J,n-k)$$

where $F_\alpha(J,n-k)$ is the upper critical points on the central F distribution with J and $(n-k)$ degrees of freedom.

Test of significance of regression (Analysis of variance)

If we set $R = [0 \quad I_{k-1}]$, $r = 0$, then the hypothesis $H_0 : R\beta = r$ reduces to the following null hypothesis:

$$H_0 : \beta_2 = \beta_3 = ... = \beta_k = 0$$

against the alternative hypothesis

$$H_1 : \beta_j \neq 0 \text{ for at least one } j = 2,3,...,k$$

This hypothesis determines if there is a linear relationship between y and any set of the explanatory variables $X_2, X_3, ..., X_k$. Notice that X_1 corresponds to the intercept term in the model and hence $x_{i1} = 1$ for all $i = 1,2,...,n$.

This is an **overall** or **global test of model adequacy**. Rejection of the null hypothesis indicates that at least one of the explanatory variables among $X_2, X_3, ..., X_k$. contributes significantly to the model. This is called as **analysis of variance.**

Since $\varepsilon \sim N(0, \sigma^2 I)$,

so $y \sim N(X\beta, \sigma^2 I)$

$$b = (X'X)^{-1}X'y \sim N\left[\beta, \sigma^2 (X'X)^{-1}\right].$$

Also $$\hat{\sigma}^2 = \frac{SS_{res}}{n-k}$$

$$= \frac{(y-\hat{y})'(y-\hat{y})}{n-k}$$

$$= \frac{y'\left[I - X(X'X)^{-1}X'\right]y}{n-k} = \frac{y'\bar{H}y}{n-k} = \frac{y'y - b'X'y}{n-k}.$$

Since $(X'X)^{-1}X'\bar{H} = 0$, so b and $\hat{\sigma}^2$ are independently distributed.

Since $y'\bar{H}y = \varepsilon'\bar{H}\varepsilon$ and $\bar{H}$ is an idempotent matrix, so

$$SS_{res} \sim \chi^2_{(n-k)},$$

i.e., central χ^2 distribution with $(n-k)$ degrees of freedom.

Partition $X = [X_1, \quad X_2^*]$ where the submatrix X_2^* contains the explanatory variables $X_2, X_3, ..., X_k$ and partition $\beta = [\beta_1, \quad \beta_2^*]$ where the subvector β_2^* contains the regression coefficients $\beta_2, \beta_3, ..., \beta_k$.

Now partition the total sum of squares due to $y's$ as

$$SS_T = y'Ay$$
$$= SS_{reg} + SS_{res}$$

where $SS_{reg} = b_2^{*}{}'X_2^{*}{}'AX_2^{*}b_2^{*}$ is the sum of squares due to regression and the sum of squares due to residuals is given by

$$SS_{res} = (y - Xb)'(y - Xb)$$
$$= y'\bar{H}y$$
$$= SS_T - SS_{reg}.$$

Further

$\frac{SS_{reg}}{\sigma^2} \sim \chi^2_{k-1}\left(\frac{\beta_2^{*}{}'X_2^{*}{}'AX_2^{*}\beta_2^{*}}{2\sigma^2}\right)$, i.e., non-central χ^2 distribution with non – centrality parameter $\frac{\beta_2^{*}{}'X_2^{*}{}'AX_2^{*}\beta_2^{*}}{2\sigma^2}$,

$\frac{SS_T}{\sigma^2} \sim \chi^2_{n-1}\left(\frac{\beta_2^{*}{}'X_2^{*}{}'AX_2^{*}\beta_2^{*}}{2\sigma^2}\right)$, i.e., non-central χ^2 distribution with non – centrality parameter $\frac{\beta_2^{*}{}'X_2^{*}{}'AX_2^{*}\beta_2^{*}}{2\sigma^2}$.

Since $X_2\bar{H} = 0$, so SS_{reg} and SS_{res} are independently distributed. The mean squares due to regression is

$$MS_{reg} = \frac{SS_{reg}}{k-1}$$

and the mean square due to error is

$$MS_{res} = \frac{SS_{res}}{n-k}.$$

Then

$$\frac{MS_{reg}}{MS_{res}} \sim F_{k-1,n-k}\left(\frac{\beta_2^{*}{}'X_2^{*}{}'AX_2^{*}\beta_2^{*}}{2\sigma^2}\right)$$

which is a non-central F-distribution with $(k-1, n-k)$ degrees of freedom and noncentrality parameter $\frac{\beta_2^{*}{}'X_2^{*}{}'AX_2^{*}\beta_2^{*}}{2\sigma^2}$.

Under $H_0 : \beta_2 = \beta_3 = ... = \beta_k$,

$$F = \frac{MS_{reg}}{MS_{res}} \sim F_{k-1,n-k}.$$

The decision rule is to reject at α level of significance whenever

$$F \geq F_{\alpha}(k-1, n-k).$$

The calculation of F-statistic can be summarized in the form of an analysis of variance (ANOVA) table given as follows:

Source of variation	Sum of squares	Degrees of freedom	Mean squares	F
Regression	SS_{reg}	$k-1$	$MS_{reg} = SS_{reg} / k-1$	F
Error	SS_{res}	$n-k$	$MS_{res} = SS_{res} /(n-k)$	
Total	SS_T	$n-1$		

Rejection of H_0 indicates that it is likely that atleast one $\beta_i \neq 0 \quad (i = 1, 2, ..., k)$.

Test of hypothesis on individual regression coefficients

In case, if the test in analysis of variance is rejected, then another question arises is that which of the regression coefficients is/are responsible for the rejection of the null hypothesis. The explanatory variables corresponding to such regression coefficients are important for the model.

Adding such explanatory variables also increases the variance of fitted values $\hat{y}$, so one needs to be careful that only those regressors are added that are of real value in explaining the response. Adding unimportant explanatory variables may increase the residual mean square, which may decrease the usefulness of the model.

To test the null hypothesis

$$H_0 : \beta_j = 0$$

versus the alternative hypothesis

$$H_1 : \beta_j \neq 0$$

has already been discussed is the case of a simple linear regression model. In the present case, if H_0 is accepted, it implies that the explanatory variable X_j can be deleted from the model. The corresponding test statistic is

$$t = \frac{b_j}{se(b_j)} \sim t(n-k-1) \text{ under } H_0$$

where the standard error of OLSE b_j of β_j is

$se(b_j) = \sqrt{\hat{\sigma}^2 C_{jj}}$ where C_{jj} denotes the j^{th} diagonal element of $(X'X)^{-1}$ corresponding to b_j.

The decision rule is to reject H_0 at α level of significance if

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

Note that this is only a **partial or marginal test** because $\hat{\beta}_j$ depends on all the other explanatory variables $X_i (i \neq j$ that are in the model. This is a test of the contribution of X_j given the other explanatory variables in the model.

Confidence interval estimation

The confidence intervals in a multiple regression model can be constructed for individual regression coefficients as well as jointly. We consider both of them as follows:

Confidence interval on the individual regression coefficient:

Assuming ε_i's are identically and independently distributed following $N(0,\sigma^2)$ in $y = X\beta + \varepsilon$, we have

$$y \sim N(X\beta, \sigma^2 I)$$
$$b \sim N(\beta, \sigma^2 (X'X)^{-1}).$$

Thus the marginal distribution of any regression coefficient estimate

$$b_j \sim N(\beta_j, \sigma^2 C_{jj})$$

where C_{jj} is the j^{th} diagonal element of $(X'X)^{-1}$.

Thus

$$t_j = \frac{b_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \sim t(n-k) \text{ under } H_0,\ j = 1,2,...$$

where $\hat{\sigma}^2 = \frac{SS_{res}}{n-k} = \frac{y'y - b'X'y}{n-k}$.

So the $100(1-\alpha)\%$ confidence interval for $\beta_j (j = 1,2,...,k)$ is obtained as follows:

$$P\left[-t_{\frac{\alpha}{2},n-k} \leq \frac{b_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \leq t_{\frac{\alpha}{2},n-k} \right] = 1-\alpha$$

$$P\left[b_j - t_{\frac{\alpha}{2},n-k}\sqrt{\hat{\sigma}^2 C_{jj}} \leq \beta_j \leq b_j + t_{\frac{\alpha}{2},n-k}\sqrt{\hat{\sigma}^2 C_{jj}} \right] = 1-\alpha.$$

Thus the confidence interval is

$$\left(b_j - t_{\frac{\alpha}{2},n-k}\sqrt{\hat{\sigma}^2 C_{jj}}, b_j + t_{\frac{\alpha}{2},n-k}\sqrt{\hat{\sigma}^2 C_{jj}} \right).$$

Simultaneous confidence intervals on regression coefficients:

A set of confidence intervals that are true simultaneously with probability $(1-\alpha)$ are called simultaneous or joint confidence intervals.

It is relatively easy to define a joint confidence region for β in multiple regression model.

Since

$$\frac{(b-\beta)'X'X(b-\beta)}{k\,MS_{res}} \sim F_{k,n-k}$$

$$\Rightarrow P\left[\frac{(b-\beta)'X'X(b-\beta)}{k\,MS_{res}} \le F_{\alpha}(k,n-k)\right] = 1-\alpha.$$

So a $100(1-\alpha)\%$ joint confidence region for all of the parameters in β is

$$\frac{(b-\beta)'X'X(b-\beta)}{k\,MS_{res}} \le F_{\alpha}(k,n-k)$$

which describes an elliptically shaped region.

Coefficient of determination (R^2) and adjusted R^2

Let R be the multiple correlation coefficient between y , and $X_1, X_2, ..., X_k$. Then square of multiple correlation coefficient (R^2) is called a coefficient of determination. The value of R^2 commonly describes how well the sample regression line fits the observed data. This is also treated as a measure of **goodness of fit** of the model.

Assuming that the intercept term is present in the model as

$$y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + ... + \beta_k X_{ik} + u_i,\ i = 1,2,...,n$$

then

$$R^2 = 1 - \frac{e'e}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$= 1 - \frac{SS_{res}}{SS_T} = \frac{SS_{reg}}{SS_T}$$

where

SS_{res} : sum of squares due to residuals,

SS_T : total sum of squares

SS_{reg} : sum of squares due to regression.

R^2 measures the explanatory power of the model, which in turn reflects the goodness of fit of the model. It reflects the model adequacy in the sense of how much is the explanatory power of the explanatory variables.

Since

$$e'e = y'\left[I - X(X'X)^{-1}X'\right]y = y'\bar{H}y,$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2,$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\ell' y$ with $\ell = (1,1,...,1)'$, $y = (y_1, y_2, ..., y_n)'$

Thus

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \bar{y})^2 &= y'y - n\left(\frac{1}{n^2}\ell' y y'\ell\right)\\ &= y'y - y'\ell\frac{1}{n}\ell' y\\ &= y'y - y'\ell(\ell'\ell)^{-1}\ell' y\\ &= y'\left[I - \ell(\ell'\ell)^{-1}\ell'\right]y\\ &= y'Ay\end{aligned}$$

where $A = I - \ell(\ell'\ell)^{-1}\ell'$.

So $R^2 = 1 - \frac{y'\bar{H}y}{y'Ay}$.

The limits of R^2 are 0 and 1, i.e.,

$$0 \le R^2 \le 1.$$

$R^2 = 0$ indicates the poorest fit of the model.

$R^2 = 1$ indicates the best fit of the model

$R^2 = 0.95$ indicates that 95% of the variation in y is explained by R^2. In simple words, the model is 95% good.

Similarly, any other value of R^2 between 0 and 1 indicates the adequacy of the fitted model.

Adjusted R^2

If more explanatory variables are added to the model, then R^2 increases. In case the variables are irrelevant, then R^2 will still increase and gives an overly optimistic picture.

With a purpose of correction in the overly optimistic picture, adjusted R^2, denoted as $\bar{R}^2$ or adj R^2 is used which is defined as

$$\begin{aligned}\bar{R}^2 &= 1-\frac{SS_{res}/(n-k)}{SS_T/(n-1)} \\ &= 1-\left(\frac{n-1}{n-k}\right)(1-R^2).\end{aligned}$$

We will see later that $(n-k)$ and $(n-1)$ are the degrees of freedom associated with the distributions of SS_{res} and SS_T. Moreover, the quantities $\frac{SS_{res}}{n-k}$ and $\frac{SS_T}{n-1}$ are based on the unbiased estimators of respective variances of e and y in the context of analysis of variance.

The adjusted R^2 will decline if the addition if an extra variable produces too small a reduction in $(1-R^2)$ to compensate for the increase in $\left(\frac{n-1}{n-k}\right)$.

Another limitation of adjusted R^2 is that it can be negative also. For example, if $k=3, n=10, R^2=0.16$, then

$$\bar{R}^2 = 1-\frac{9}{7}\times 0.97 = -0.25 < 0$$

which has no interpretation.

Limitations

1. If the constant term is absent in the model, then R^2 can not be defined. In such cases, R^2 can be negative. Some ad-hoc measures based on R^2 for regression line through origin have been proposed in the literature.

Reason that why R^2 is valid only in linear models with intercept term:

In the model $y = X\beta + \varepsilon$, the ordinary least squares estimator of β is $b = (X'X)^{-1}X'y$. Consider the fitted model as

$$\begin{aligned} y &= Xb + (y - Xb) \\ &= Xb + e \end{aligned}$$

where e is the residual. Note that

$$\begin{aligned} y - \bar{l}\bar{y} &= Xb + e - \bar{l}\bar{y} \\ &= \hat{y} + e - \bar{l}\bar{y} \end{aligned}$$

where $\hat{y} = Xb$ is the fitted value and $l = (1,1,...,1)'$ is a $n \times 1$ vector of elements unity. The total sum of squares $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is then obtained as

$$\begin{aligned} TSS &= (y - \bar{l}\bar{y})'(y - \bar{l}\bar{y}) = [(\hat{y} - \bar{l}\bar{y}) + e]'[(\hat{y} - \bar{l}\bar{y}) + e] \\ &= (\hat{y} - \bar{l}\bar{y})'(\hat{y} - \bar{l}\bar{y}) + e'e + 2(\hat{y} - \bar{l}\bar{y})'e \\ &\qquad \downarrow \qquad\quad \downarrow \qquad \downarrow \\ &= \quad SS_{reg} \quad + \quad SS_{res} + 2(Xb - \bar{l}\bar{y})'e \quad \text{(because } \hat{y} = Xb\text{)} \\ &= SS_{reg} + SS_{res} - 2\bar{y}l'e \quad \text{(because } X'e = 0\text{).} \end{aligned}$$

The Fisher Cochran theorem requires $TSS = SS_{reg} + SS_{res}$ to hold true in the context of analysis of variance and further to define the R^2. In order that $TSS = SS_{reg} + SS_{res}$ holds true, we need that $l'e$ should be zero, i.e. $l'e = l'(y - \hat{y}) = 0$ which is possible only when there is an intercept term in the model. We show this claim as follows:

First, we consider a no intercept simple linear regression model $y_i = \beta_1 x_i + \varepsilon_i$, $(i = 1,2,...,n)$ where the parameter β_1 is estimated as $b_1^* = \dfrac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$. Then $l'e = \sum_{i=1}^{n} e_i = \sum_{i=1}^{n}(y_i - \hat{y}_i) = \sum_{i=1}^{n}(y_i - b_1^* x_i) \neq 0$, in general.

Similarly, in a no intercept multiple linear regression model $y = X\beta + \varepsilon$, we find that

$l'e = l'(y - \hat{y}) = l'(X\beta + \varepsilon - Xb) = -l'X(b - \beta) + l'\varepsilon \neq 0$, in general.

Next, we consider a simple linear regression model with intercept term $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $(i = 1, 2, ..., n)$ where the parameters β_0 and β_1 are estimated as $b_0 = \bar{y} - b_1 \bar{x}$ and $b_1 = \dfrac{s_{xy}}{s_{xx}}$ respectively, where $s_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), \quad s_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2, \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$ We find that

$$\begin{aligned} l'e &= \sum_{i=1}^{n} e_i = \sum_{i=1}^{n}(y_i - \hat{y}_i) \\ &= \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) \\ &= \sum_{i=1}^{n}(y_i - \bar{y} + b_1\bar{x} - b_1 x_i) \\ &= \sum_{i=1}^{n}[(y_i - \bar{y}) - b_1(x_i - \bar{x})] \\ &= \sum_{i=1}^{n}(y_i - \bar{y}) - b_1\sum_{i=1}^{n}(x_i - \bar{x}) \\ &= 0. \end{aligned}$$

In a multiple linear regression model with an intercept term $y = \beta_0 l + X\beta + \varepsilon$ where the parameters β_0 and β are estimated as $\hat{\beta}_0 = \bar{y} - b\bar{x}$ and $b = (X'X)^{-1}X'y$, respectively. We find that

$$\begin{aligned} l'e &= l'(y - \hat{y}) \\ &= l'(y - \hat{\beta}_0 - Xb) \\ &= l'(y - \bar{y} + \bar{X}b - Xb) \quad , \\ &= l'(y - \bar{y}) + l'(X - \bar{X})b \\ &= 0. \end{aligned}$$

Thus we conclude that for the Fisher Cochran to hold true in the sense that the total sum of squares can be divided into two orthogonal components, viz., the sum of squares due to regression and sum of squares due to errors, it is necessary that $l'e = l'(y - \hat{y}) = 0$ holds and which is possible only when the intercept term is present in the model.

2. R^2 is sensitive to extreme values, so R^2 lacks robustness.

3. R^2 always increases with an increase in the number of explanatory variables in the model. The main drawback of this property is that even when the irrelevant explanatory variables are added in the model, R^2 still increases. This indicates that the model is getting better, which is not really correct.

4. Consider a situation where we have the following two models:

$$y_i = \beta_1 + \beta_2 X_{i2} + ... + \beta_k X_{ik} + u_i, \quad i = 1, 2, .., n$$
$$\log y_i = \gamma_1 + \gamma_2 X_{i2} + ... + \gamma_k X_{ik} + v_i$$

The question is now which model is better?

For the first model,

$$R_1^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

and for the second model, an option is to define R^2 as

$$R_2^2 = 1 - \frac{\sum_{i=1}^{n} (\log y_i - \log \hat{y}_i)^2}{\sum_{i=1}^{n} (\log y_i - \log \bar{y})^2}.$$

As such R_1^2 and R_2^2 are not comparable. If still, the two models are needed to be compared, a better proposition to define R^2 can be as follows:

$$R_3^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - anti \log \hat{y}_i^*)}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

where $y_i^* = \widehat{\log y_i}$. Now R_1^2 and R_3^2 on the comparison may give an idea about the adequacy of the two models.

Relationship of analysis of variance test and coefficient of determination

Assuming the β_1 to be an intercept term, then for $H_0 : \beta_2 = \beta_3 = ... = \beta_k = 0,$ the $F-$statistic in analysis of variance test is

$$\begin{aligned} F &= \frac{MS_{reg}}{MS_{res}} \\ &= \frac{(n-k)}{(k-1)} \frac{SS_{reg}}{SS_{res}} \\ &= \left(\frac{n-k}{k-1}\right) \frac{SS_{reg}}{SS_T - SS_{reg}} \\ &= \left(\frac{n-k}{k-1}\right) \frac{\frac{SS_{reg}}{SS_T}}{1 - \frac{SS_{reg}}{SS_T}} = \left(\frac{n-k}{k-1}\right) \frac{R^2}{1-R^2} \end{aligned}$$

where R^2 is the coefficient of determination. So F and R^2 are closely related. When $R^2 = 0,$ then $F = 0.$ In the limit, when $R^2 = 1, F = \infty$. So both F and R^2 vary directly. Larger R^2 implies greater F value. That is why the F test under the analysis of variance is termed as the measure of the overall significance of estimated regression. It is also a test of significance of R^2. If F is highly significant, it implies that we can reject H_0, i.e. y is linearly related to X's.

Prediction of values of study variable

The prediction in the multiple regression model has two aspects

1. Prediction of the average value of study variable or mean response.
2. Prediction of the actual value of the study variable.

1. Prediction of average value of y

We need to predict $E(y)$ at a given $x_0 = (x_{01}, x_{02}, ..., x_{0k})'$.

The predictor as a point estimate is

$$p = x_0' b = x_0' (X'X)^{-1} X'y$$
$$E(p) = x_0' \beta.$$

So p is an unbiased predictor for $E(y)$.

Its variance is

$$Var(p) = E\left[p - E(y)\right]'\left[p - E(y)\right]$$
$$= \sigma^2 x_0' (X'X)^{-1} x_0$$

Then

$$E(\hat{y}_0) = x_0' \beta = E(y \mid x_0)$$
$$Var(\hat{y}_0) = \sigma^2 x_0' (X'X)^{-1} x_0$$

The confidence interval on the mean response at a particular point, such as $x_{01}, x_{02}, ..., x_{0k}$ can be found as follows:

Define $x_0 = (x_{01}, x_{02}, ..., x_{0k})'$. The fitted value at x_0 is $\hat{y}_0 = x_0' b.$

Then

$$P\left[-t_{\frac{\alpha}{2},n-k} \le \frac{\hat{y}_0 - E(y \mid x_0)}{\sqrt{\hat{\sigma}^2 x_0'(X'X)^{-1}x_0}} \le t_{\frac{\alpha}{2},n-k}\right] = 1-\alpha$$

$$P\left[\hat{y}_0 - t_{\frac{\alpha}{2},n-k} \le \sqrt{\hat{\sigma}^2 x_0'(X'X)^{-1}x_0} \le E(y \mid x_0) \le \hat{y}_0 + t_{\frac{\alpha}{2},n-k} \le \sqrt{\hat{\sigma}^2 x_0'(X'X)^{-1}x_0}\right] = 1-\alpha.$$

The $100(1-\alpha)\%$ confidence interval on the mean response at the point $x_{01}, x_{02}, ..., x_{0k}$, i.e., $E(y / x_0)$ is

$$\left[\hat{y}_0 - t_{\frac{\alpha}{2},n-k}\sqrt{\hat{\sigma}^2 x_0'(X'X)^{-1}x_0}, \hat{y}_0 + t_{\frac{\alpha}{2},n-k} \le \sqrt{\hat{\sigma}^2 x_0'(X'X)^{-1}x_0}\right].$$

2. Prediction of actual value of y

We need to predict y at a given $x_0 = (x_{01}, x_{02}, ..., x_{0k})'$.

The predictor as a point estimate is

$$p_f = x_0'b$$

$$E(p_f) = x_0'\beta$$

So p_f is an unbiased for y. It's variance is

$$Var(p_f) = E\left((p_f - y)(p_f - y)'\right)$$

$$= \sigma^2\left[1 + x_0'(X'X)^{-1}x_0\right].$$

The $100(1-\alpha)\%$ confidence interval for this future observation is

$$\left(p_f - t_{\frac{\alpha}{2},n-k}\sqrt{\hat{\sigma}^2[1 + x_0'(X'X)^{-1}x_0]},\ \ p_f + t_{\frac{\alpha}{2},n-k}\sqrt{\hat{\sigma}^2[1 + x_0'(X'X)^{-1}x_0]}\right).$$