Multiple Linear Regressions

Instead of one predictor variable, when there are at least two predictor variables, we use multiple linear regressions. In case of p regressor variables, multiple linear regression models are given by

$$y_i = b_0 + \sum_{j=1}^{p} b_j x_{ij} + \varepsilon_i, \quad i = 1,2,\cdots,n \text{ and } n > p$$

Errors $\varepsilon_i,\ i = 1,2,\cdots,n$ are assumed independent $N(0,\sigma^2)$, as in simple linear regression.

We wish to find the vector of least square estimators, $\hat{b}$, that minimizes

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij} \right)^2$$

Just as in simple linear regression, model is fit by minimizing with respect to $b_0, b_1, \cdots, b_p$. The least square estimators say, $\hat{b}_0, \hat{b}_1, \cdots, \hat{b}_p$ must satisfy

$$\left.\frac{\partial L}{\partial b_0}\right|_{\hat{b}_0,\hat{b}_1,\cdots,\hat{b}_p} = -2 \sum_{i=1}^{n} \left(y_i - \hat{b}_0 - \sum_{j=1}^{p} \hat{b}_j x_{ij} \right) = 0,$$

and

$$\left.\frac{\partial L}{\partial b_j}\right|_{\hat{b}_0,\hat{b}_1,\cdots,\hat{b}_p} = -2 \sum_{i=1}^{n} \left(y_i - \hat{b}_0 - \sum_{j=1}^{p} \hat{b}_j x_{ij} \right) x_{ij} = 0, \quad j = 1,2,\cdots,p$$

Above can be written as

$$
\begin{aligned}
n\hat{b}_0 + \hat{b}_1 \sum_{i=1}^{n} x_{i1} + \hat{b}_2 \sum_{i=1}^{n} x_{i2} + \cdots + \hat{b}_p \sum_{i=1}^{n} x_{ip} &= \sum_{i=1}^{n} y_i \\
\hat{b}_0 \sum_{i=1}^{n} x_{i1} + \hat{b}_1 \sum_{i=1}^{n} x_{i1}^2 + \hat{b}_2 \sum_{i=1}^{n} x_{i1}x_{i2} + \cdots + \hat{b}_p \sum_{i=1}^{n} x_{i1}x_{ip} &= \sum_{i=1}^{n} x_{i1}y_i \\
&\vdots \\
&\vdots \\
\hat{b}_0 \sum_{i=1}^{n} x_{ip} + \hat{b}_1 \sum_{i=1}^{n} x_{i1}x_{ip} + \hat{b}_2 \sum_{i=1}^{n} x_{i2}x_{ip} + \cdots + \hat{b}_p \sum_{i=1}^{n} x_{ip}^2 &= \sum_{i=1}^{n} x_{ip}y_i
\end{aligned}
$$

These are called the **least square normal equations**. Note that there are p+1 normal equations, one for each of the unknown regression coefficients.

Matrix Approach to Multiple Linear Regressions

In matrix notation the p variable regression model can be written as

$$\boldsymbol{y}_{\boldsymbol{n}\times\mathbf{1}} = \boldsymbol{X}_{\boldsymbol{n}\times(\boldsymbol{p}+\mathbf{1})}\boldsymbol{b}_{(\boldsymbol{p}+\mathbf{1})\times\mathbf{1}} + \boldsymbol{\varepsilon}_{\boldsymbol{n}\times\mathbf{1}}$$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \quad b = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{bmatrix} \quad and\ \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

where y is an (n×1) vector of responses, X is an [n×(p+1)] design matrix of the model, b is a column vector of order p+1, and ε is an (n×1) vector of uncorrelated random errors with $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \sigma^2$. Further, it is assumed that X is a non-stochastic and is of full rank.

Since $E(\varepsilon_i) = 0$, $i = 1, 2, \cdots, n$ so, $E(\varepsilon) = 0$. Also, $E(\varepsilon_i^2) = \sigma^2$. Moreover as ε_i's are uncorrelated, $E(\varepsilon_i\varepsilon_j) = 0$, for $i \neq j$.

Therefore,

$$Var(\varepsilon) = E\left[\left(\varepsilon - E(\varepsilon)\right)\left(\varepsilon - E(\varepsilon)\right)^T\right] = E(\varepsilon\varepsilon^T) = \sigma^2\boldsymbol{I}.$$

Above gives the variance-covariance matrix of the **random errors**.

It may be noted that

$$X^TX = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}^T \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

$$= \begin{bmatrix} n & \sum x_{i1} & \sum x_{i2} & \cdots & \sum x_{ip} \\ \sum x_{i1} & \sum x_{i1}^2 & \sum x_{i1}x_{i2} & \cdots & \sum x_{i1}x_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum x_{ip} & \sum x_{ip}x_{i1} & \sum x_{ip}x_{i2} & \cdots & \sum x_{ip}^2 \end{bmatrix}_{(p+1)\times(p+1)}$$

and

$$X^Ty = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}^T \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_{i1}y_i \\ \vdots \\ \sum x_{ip}y_i \end{bmatrix}$$

So, clearly the **least square normal equations** can be expressed in matrix form as

$$X^TX\hat{b} = X^Ty$$

Alternatively, we can obtain the **least square normal equations** by differentiating ESS and equating the same to zero. We have

$$L = \sum e_i^2 = e^T e = \left(y - X\hat{b}\right)^T \left(y - X\hat{b}\right)$$
$$= y^T y - \hat{b}^T X^T y - y^T X\hat{b} + \hat{b}^T X^T X\hat{b}$$
$$= y^T y - 2\hat{b}^T X^T y + \hat{b}^T X^T X\hat{b}$$

as the transpose of a scalar is also the same scalar.

It may be noted that, both b and $X^T y$ are column vectors of order p+1.

So, we get

$$\frac{\partial L}{\partial \hat{b}} = -2X^T y + 2X^T X\hat{b} = 0$$
$$\Rightarrow X^T X\hat{b} = X^T y$$

Therefore, the regression coefficients can be estimated by

$\hat{b} = (X^T X)^{-1} X^T y$, provided $X^T X$ is invertible.

Moreover, $\frac{\partial^2 L}{\partial \hat{b}^2} = 2X^T X$. Now, $X^T X$ is positive definite, hence $\hat{b}$ minimizes the normal equation.

Let u be a non-zero column vector of order $(p+1)$. So, clearly Xu will be a column vector of order n. [Since, $X\big(n \times (p+1)\big)$ and $u\big((p+1) \times 1\big)$]

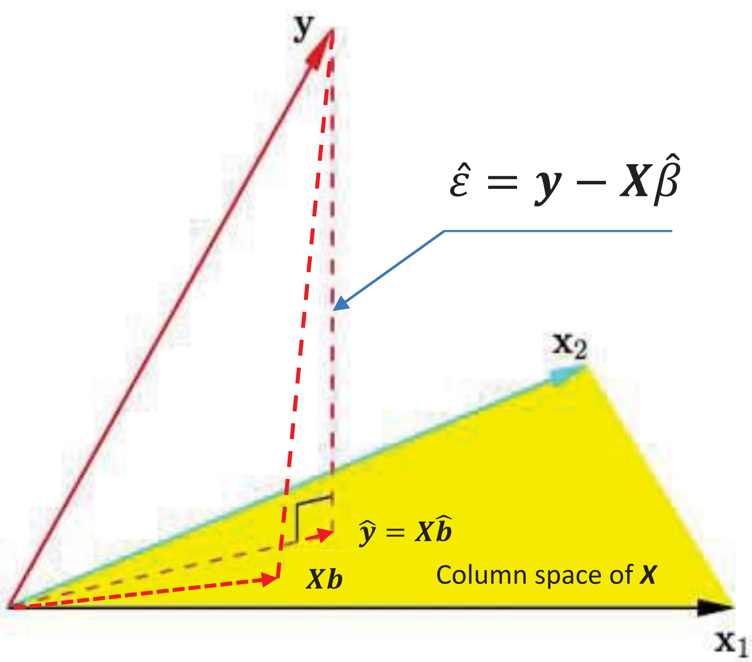
Now, we can write $u^T (X^T X) u = (Xu)^T Xu = \sum_{i=1}^{n} (Xu)_i^2$.

Since columns of X are assumed to be a matrix of full rank, so X cannot be a null matrix. Therefore, as $u \neq 0$, $Xu \neq 0$.

Above implies, $u^T (X^T X) u > 0$ and hence $X^T X$ is positive definite.

Geometrical Interpretation of Regression

A geometric interpretation of linear regression is, perhaps, more intuitive. The column vectors of $\boldsymbol{X}$ span a subspace, and minimizing the residuals amounts to making an orthogonal projection of y onto this subspace, as seen in the figure below.



Thus the output vector y is orthogonally projected onto the hyperplane spanned by input vectors x_1 and x_2. The projection $\hat{y}$ represents the vector of the least squares predictions.

Mathematically, from Normal Equations

$$\boldsymbol{X}^T\boldsymbol{X}\hat{b} = X^T y \quad \Rightarrow \quad X^T(y - X\hat{b}) = \mathbf{0} \Rightarrow \begin{bmatrix} \boldsymbol{X}_1^T \\ \boldsymbol{X}_2^T \\ \vdots \\ \boldsymbol{X}_{p+1}^T \end{bmatrix}(y - X\hat{b}) = \mathbf{0}$$

or, $\quad \boldsymbol{X}_i^T(y - X\hat{b}) = 0, \quad i = 1, 2, \cdots, p+1$

Thus, $y - X\hat{b}$, i.e. residuals are orthogonal to every column vector in X, i.e. the space spanned by column vectors of X. It may also be noted that, out of all vectors in the space spanned by column vectors of X, the one that minimizes the length $||\hat{\varepsilon}||$ is the orthogonal projection of $\hat{y}$.

So, the regression model can be written as

$\hat{y} = X\hat{b} = X(X^T X)^{-1} X^T y = Hy$, where $H = X(X^T X)^{-1} X^T$ is known as the 'hat' matrix, i.e. the matrix that converts observed values of y into vector of fitted values $\hat{y}$.

Note that, (i) H is a square matrix of order n, and

(ii) both X and X^T are rectangular matrices, hence non invertible, so $H \neq I$.

- H is symmetric, i.e $H = H^T$, so that $h_{ij} = h_{ji}$.

$$[H^T = (X(X^T X)^{-1} X^T)^T = X(X^T X)^{-1} X^T = H]$$

- H is idempotent, i.e. $H^2 = H^T H = H$.

$$[H^2 = H^T H = (X(X^T X)^{-1} X^T)^T (X(X^T X)^{-1} X^T) = X(X^T X)^{-1} X^T = H]$$

- H is positive semi-definite (psd).

Statistical properties of least square estimator $\hat{b}$

$$\begin{aligned}
E(\hat{b}) &= E\left[\left(X^T X\right)^{-1} X^T y\right] \\
&= E\left[\left(X^T X\right)^{-1} X^T (Xb + \varepsilon)\right] \\
&= E\left[\left(X^T X\right)^{-1} X^T Xb + \left(X^T X\right)^{-1} X^T \varepsilon\right] \\
&= E\left[b + \left(X^T X\right)^{-1} X^T \varepsilon\right] = b
\end{aligned}$$

Since $E(\varepsilon) = 0$ and $(X^T X)^{-1} X^T X = I$, the identity matrix. Thus, $\hat{b}$ is an unbiased estimator of b.

Variance of $\hat{b}$

Since, $\hat{b} = (X^T X)^{-1} X^T y$, so replacing y by $Xb + \varepsilon$, we get

$$\hat{b} = (X^T X)^{-1} X^T (Xb + \varepsilon) \Rightarrow \hat{b} = (X^T X)^{-1} X^T X b + (X^T X)^{-1} X^T \varepsilon$$
$$\Rightarrow \hat{b} = b + (X^T X)^{-1} X^T \varepsilon$$
$$\Rightarrow \hat{b} - E(\hat{b}) = (X^T X)^{-1} X^T \varepsilon$$

Therefore,

$$V(\hat{b}) = E\left[\left(\hat{b} - E(\hat{b})\right)\left(\hat{b} - E(\hat{b})\right)^T\right] = E\left[\left(\left(X^T X\right)^{-1} X^T \varepsilon\right)\left(\left(X^T X\right)^{-1} X^T \varepsilon\right)^T\right]$$
$$= E\left[\left(X^T X\right)^{-1} X^T \varepsilon \varepsilon^T X \left(X^T X\right)^{-1}\right]$$

Since X is non-stochastic and we know that $E(\varepsilon\varepsilon^T) = \sigma^2 I$, so we have

$$V(\hat{b}) \qquad \Rightarrow \left(X^T X\right)^{-1} X^T E\left(\varepsilon\varepsilon^T\right) X \left(X^T X\right)^{-1}$$
$$\Rightarrow \left(X^T X\right)^{-1} X^T \left\{\sigma^2 I\right\} X \left(X^T X\right)^{-1}$$
$$\Rightarrow \sigma^2 \left(X^T X\right)^{-1} \left\{X^T I\right\} X \left(X^T X\right)^{-1}$$
$$\Rightarrow \sigma^2 \left(X^T X\right)^{-1} X^T X \left(X^T X\right)^{-1}$$
$$\Rightarrow \sigma^2 \left(X^T X\right)^{-1}$$
$$\Rightarrow \sigma^2 C, \quad \text{where } C = \left(X^T X\right)^{-1}$$

Clearly, $C = (X^T X)^{-1}$ is a symmetric matrix of order p+1 and $\sigma^2 C$ is known as the **Variance Covariance Matrix of the OLS estimator $\hat{b}$**.

Diagonal elements of the variance covariance matrix are the variances of $\hat{b}_j,\ 0 \le j \le p$, whereas the off-diagonal elements are the covariance's. So that, we have

$$V(\hat{b}_j) = \sigma^2 C_{jj}, \qquad j = 0, 1, 2, \cdots, p$$
$$Cov(\hat{b}_i, \hat{b}_j) = \sigma^2 C_{ij}, \qquad i \neq j$$

Estimate of σ^2

Similar to simple linear regression, we can get an estimate of σ^2 from sum of squares of the residuals, as

$$\begin{aligned} SS_E &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} e_i^2 \quad = e^T e \end{aligned}$$

Substituting $e = y - \hat{y} = y - X\hat{b}$, we get

$$\begin{aligned} SS_E &= (y - X\hat{b})^T (y - X\hat{b}) \\ &= y^T y - \hat{b}^T X^T y - y^T X\hat{b} + \hat{b}^T X^T X\hat{b} \\ &= y^T y - 2\hat{b}^T X^T y + \hat{b}^T X^T X\hat{b}. \end{aligned}$$

Since, $X^T X\hat{b} = X^T y$ (matrix form of the least square normal equations), above equation simplifies to

$$SS_E = y^T y - \hat{b}^T X^T y. \qquad \text{(A)}$$

Above error sum of squares has $(n-1) - p = n - p - 1$ degrees of freedom associated with it. The mean square error is

$$MS_E = \frac{SS_E}{n-p-1},$$

where p is the number of regressor variables and this mean square error is taken as an unbiased estimator of σ^2, i.e. $\hat{\sigma}^2 = MS_E$.

Example> A study was performed on wear of bearing y and its relationship to x_1 = oil viscosity and x_2 = load. The following data were obtained

y	293	230	172	91	113	125
x_1	1.6	15.5	22.0	43.0	33.0	40.0
x_2	851	816	1058	1201	1357	1115

a) Fit a multiple linear regression model to this data.
b) Estimate σ^2.

Here,

$$X = \begin{bmatrix} 1 & 1.6 & 851 \\ 1 & 15.5 & 816 \\ 1 & 22 & 1058 \\ 1 & 43 & 1201 \\ 1 & 33 & 1357 \\ 1 & 40 & 1115 \end{bmatrix} \quad and \quad y = \begin{bmatrix} 293 \\ 230 \\ 172 \\ 91 \\ 113 \\ 125 \end{bmatrix}.$$

$$X^T X = \begin{bmatrix} 6 & 155.1 & 6398 \\ 155.1 & 5264.81 & 178309.6 \\ 6398 & 178309.6 & 7036496 \end{bmatrix} \quad and \quad X^T y = \begin{bmatrix} 1024 \\ 20459.8 \\ 1021006 \end{bmatrix}.$$

$$\left(X^T X\right)^{-1} = \begin{bmatrix} 8.595096 & 0.080958 & -0.0098667 \\ 0.080958 & 0.002102 & -0.0001269 \\ -0.00987 & -0.00013 & 1.2329E-05 \end{bmatrix}$$

Therefore,

$$\begin{bmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \hat{b}_1 \end{bmatrix} = \left(X^T X\right)^{-1} * X^T y = \begin{bmatrix} 383.801 \\ -3.638 \\ -0.112 \end{bmatrix}$$

Therefore, the regression equation is: $y = 383.801 - 3.638x_1 - 0.112x_2$ **.**

SS_E = 205008 – 204550.14 = 457.86, therefore MS_E= 457.86/3=152.62.

Test for Significance of Regression

$$H_0: b_1 = b_2 = \cdots = b_p = 0$$
$$H_1: b_j \neq 0 \text{ for at least one } j$$

Rejection of null hypothesis implies that at least one of the predictor variables $x_1, x_2, \cdots, x_p$ contributes significantly to the model.

We test this hypothesis using ANOVA, where total variation in the response is divided into i) variation explained by regression model, and ii) unexplained variation, i.e. $S_{yy} = SS_R + SS_E$. As usual to test the null hypothesis, we compute

$$F_0 = \frac{SS_R/p}{SS_E/(n-p-1)} = \frac{MS_R}{MS_E}$$

and reject H_0 if $f_0 > F_{\alpha,p,n-p-1}$.

We have earlier proved that [ref equation (A)], $SS_E = y^T y - \hat{b}^T X^T y.$
Now we know that

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = y^T y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}.$$

So, we may rewrite the above equation as

$$SS_E = y^T y - \frac{(\sum_{i=1}^{n} y_i)^2}{n} - \left[\hat{b}^T X^T y - \frac{(\sum_{i=1}^{n} y_i)^2}{n}\right]$$

Or, $$SS_E = S_{yy} - SS_R.$$

Therefore, the regression sum of squares is $SS_R = \hat{b}^T X^T y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$,

and total sum of squares $S_{yy} = y^T y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$.

ANOVA table

Source of variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	SS_R	p	MS_R	MS_R/MS_E
Error	SS_E	$n-p-1$	MS_E	
Total	S_{yy}	$n-1$		