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

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

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

$$\boldsymbol{y_{n\times 1}} = \boldsymbol{X_{n\times(p+1)}}\boldsymbol{b_{(p+1)\times 1}} + \boldsymbol{\varepsilon_{n\times 1}}$$

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

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

Therefore,

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

Above gives the variance-covariance matrix of the **random errors**. So, we have $Var(y) = Var(\varepsilon) = \sigma^2 I$.

It may be noted that

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

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

and

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

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

$$X^T X \hat{b} = X^T y$$

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 $u\big((p+1) \times 1\big)$]

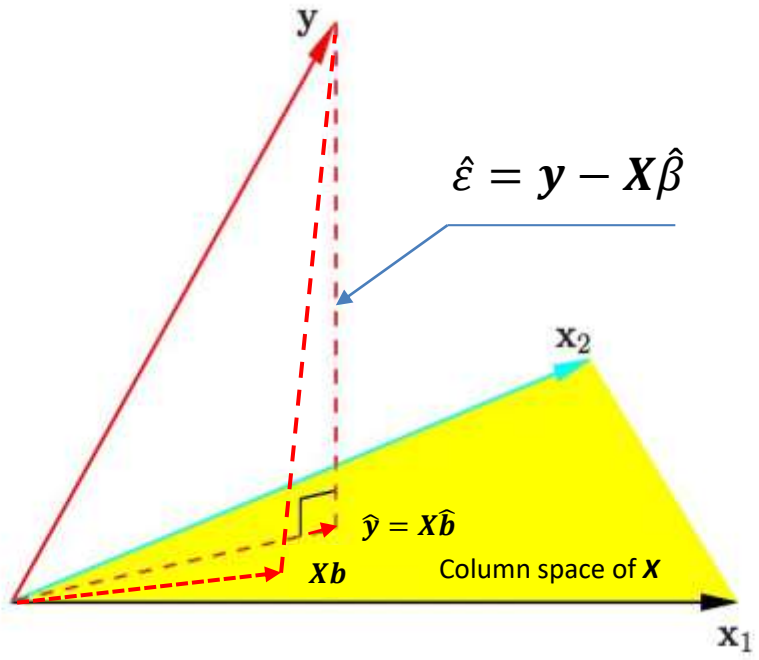
Now, we can write $u^T(X^T X)u = (Xu)^T Xu = \sum_{i=1}^{n} (Xu)_i^2$.

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

$$\boldsymbol{X}^T\boldsymbol{X}\hat{b} = X^T y \quad \Rightarrow \quad X^T\left(y - X\hat{b}\right) = \mathbf{0} \Rightarrow \begin{bmatrix} \boldsymbol{X_1^T} \\ \boldsymbol{X_2^T} \\ \vdots \\ \boldsymbol{X_{p+1}^T} \end{bmatrix}\left(y - X\hat{b}\right) = \mathbf{0}$$

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

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

$\hat{y} = X\hat{b} = X(X^TX)^{-1}X^Ty = Hy$, where $H = X(X^TX)^{-1}X^T$ 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 $h_{ij} = h_{ji}$.
  $[H^T = (X(X^TX)^{-1}X^T)^T = X(X^TX)^{-1}X^T = H]$
- H is idempotent, i.e. $H^2 = H^TH = H$.
  $[H^2 = H^TH = (X(X^TX)^{-1}X^T)^T(X(X^TX)^{-1}X^T) = X(X^TX)^{-1}X^T = H]$
- H is positive semi-definite (psd).

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

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

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

Variance of $\hat{b}$

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

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

Therefore,

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

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

$$V(\hat{b}) \Rightarrow (X^TX)^{-1}X^TE(\varepsilon\varepsilon^T)X(X^TX)^{-1}$$
$$\Rightarrow (X^TX)^{-1}X^T\{\sigma^2 I\}X(X^TX)^{-1}$$
$$\Rightarrow \sigma^2(X^TX)^{-1}\{X^TI\}X(X^TX)^{-1}$$
$$\Rightarrow \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1}$$
$$\Rightarrow \sigma^2(X^TX)^{-1}$$
$$\Rightarrow \sigma^2C, \quad \text{where } C = (X^TX)^{-1}$$

Clearly, $C = (X^TX)^{-1}$ is a symmetric matrix of order p+1 and σ^2C 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 $SS_R = \hat{b}^T X^T y - \frac{(\sum_{i=1}^{n} y_i)^2}{n}$,

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

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}	$n-1$		

Test of Individual Regression Coefficients

$$H_0: b_j = b_{j0}$$
$$H_1: b_j \neq b_{j0}$$

The test statistic for testing above hypothesis is

$$t_0 = \frac{\hat{b}_j - b_{j0}}{SE(\hat{b}_j)} = \frac{\hat{b}_j - b_{j0}}{\sqrt{MS_E C_{jj}}}, \text{ where } C = (X^T X)^{-1}$$

The null hypothesis is rejected, if $|t_0| > t_{\alpha/2,\ n-p-1}$. This is also known as **partial** or **marginal** test.

If the hypothesis is $H_0 : \hat{b}_j = 0$ against $H_1 : \hat{b}_j \neq 0$, then rejecting the null hypothesis imply that variable x_j contribute significantly to the model or vice versa.

There is another way to test the contribution of an individual or a set of regressor variables to the model. This approach determines the increase in the regression sum of squares obtained by adding a variable or a set of variables to the model given that other variables are already included in the model. The procedure used to do this is called the **partial sum of squares method**.

Suppose the full model contains p regressor variables and we are interested in determining whether the subset of regressor variables $x_1, x_2, \cdots, x_r \ \ (r < p)$ as a whole contributes significantly to the model.

Let us define

$$b(1) = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_r \end{bmatrix} \text{ and } b(2) = \begin{bmatrix} b_{r+1} \\ b_{r+2} \\ \vdots \\ b_p \end{bmatrix}, \quad \text{so that}$$

$$b = \begin{bmatrix} b_0 \\ b(1) \\ b(2) \end{bmatrix}.$$

1. Obtain the full model involving all the *p* variables.
   Calculate the values of $SS_R(Full)$ and MS_E corresponding to the full model.
2. Find the regression equation for the reduced model involving $b(2)$ and intercept i.e. by taking the columns of X corresponding to intercept and $b(2)$. Calculate resulting value of $SS_R[b(2)]$.

3. So, increase in regression sum of squares due to the inclusion of $x_1, x_2, \cdots, x_r$ given that $x_{r+1}, x_{r+2}, \cdots, x_p$ are already in the model is

$$SS_R\big(b(1)\big|b(2)\big) = SS_R(\text{Full}) - SS_R\big(b(2)\big).$$

   This sum of square has *r* degrees of freedom and is known as the increase in regression sum of squares due to inclusion of $b(1)$.

4. The null hypothesis $H_0: b(1) = 0$ is tested by the statistic

$$F_0 = \frac{SS_R[b(1)/b(2)]/r}{MS_E}$$

5. If the computed value of the test statistic $f_0 > F_{\alpha,r,n-p-1}$, we reject the null hypothesis and thereby conclude that at least one of the variables in $b(1)$ is non-zero, i.e. at least one of the variables $x_1, x_2, \cdots, x_r$ contributes significantly to the regressor model. The test statistic described above is also known as **partial F-test**.

Confidence Interval on Individual Regression Coefficients

By assumption errors $\{\varepsilon_i\}$ are distributed as $i.i.d.\ N(0, \sigma^2)$. So the observations $\{y_i\}$ are normally and independently distributed with mean $b_0 + \sum_{j=1}^{p} b_j x_{ij}$ and variance σ^2. Since the least square estimator $\hat{b}$ is a linear combination of the observations (y_i), it follows that $\hat{b}$ is normally distributed with mean vector $\boldsymbol{b}$ and the variance covariance matrix $\sigma^2(X^TX)^{-1}$, so each of the statistics

$$T = \frac{\hat{b}_j - b_j}{\sqrt{MS_E C_{jj}}} \qquad j = 0,1,2,\cdots,p$$

has a t distribution with n-p-1 degrees of freedom, where C_{jj} and MS_E are jj-th element of $(X^TX)^{-1}$matrix and estimate of error variance respectively. This leads to the following $100(1-\alpha)\%$ confidence interval for the regression coefficient $b_j, 0 \le j \le p$

$$\hat{b}_j - t_{\alpha/2,n-p-1}\sqrt{MS_E C_{jj}} \le b_j \le \hat{b}_j + t_{\alpha/2,n-p-1}\sqrt{MS_E C_{jj}}$$

Confidence Interval on the Mean Response

Let

$$x_0^T = (1, x_{01}, x_{02}, \cdots, x_{0p})$$

be the point for which we need the confidence interval on mean response. The mean response at this point is $E(y|x_0) = \mu_{y|x_0} = x_0^T b$ and is estimated by

$$\hat{\mu}_{y|x_0} = x_0^T \hat{b}$$

Since, $E(\hat{\mu}_{y|x_0}) = E(x_0^T \hat{b}) = x_0^T b = \mu_{y|x_0}$ and this implies that above estimator is unbiased. The variance of $\hat{\mu}_{y|x_0}$ is

$$V(\hat{\mu}_{y|x_0}) = V(x_0^T \hat{b}) = x_0^T V(\hat{b}) x_0 = x_0^T \sigma^2 (X^T X)^{-1} x_0 = \sigma^2 x_0^T (X^T X)^{-1} x_0.$$

A $100(1-\alpha)\%$ confidence interval can be constructed from the statistic $\frac{\hat{\mu}_{y|x_0} - \mu_{y|x_0}}{\sqrt{\hat{\sigma}^2 x_0^T (X^T X)^{-1} x_0}}$, which follows a t distribution with $n-p-1$ d. f. and the Confidence Interval is given by

$$\hat{\mu}_{y/x_0} - t_{\alpha/2,n-p-1}\sqrt{MS_E x_0^T (X^T X)^{-1} x_0} \leq \mu_{y/x_0} \leq \hat{\mu}_{y/x_0} + t_{\alpha/2,n-p-1}\sqrt{MS_E x_0^T (X^T X)^{-1} x_0}$$

Variance of Residuals

We know residual, $e = y - \hat{y} = y - Hy = (I - H)y$.

So, $V\{e\} = (I-H)^T V(y)(I-H) = (I-H)^T \sigma^2 I (I-H)$

$= \sigma^2 (I-H) I (I-H)$

$= \sigma^2 (I-H)(I-H).$

Now, $(I-H)(I-H) = I - H$, thus $I - H$ is idempotent.

So, $V\{e\} = \sigma^2 (I-H).$

The matrix $I - H$ are Idempotent and Symmetrical [$(I-H)^T = I^T - H^T = I - H$], but generally non-diagonal. So, residuals have different variances and are correlated.

Therefore, $V(e_i) = \sigma^2(1 - h_{ii})$ and $cov(e_i, e_j) = -\sigma^2 h_{ij}$ $(i \neq j)$.

Thus clearly, $1 - h_{ii} > 0 \Rightarrow h_{ii} < 1$.

Model Adequacy Checking

Coefficient of Multiple Determinations

The coefficient of multiple determinations is defined by

$$R^2 = \frac{SS_R}{S_{yy}} = 1 - \frac{SS_E}{S_{yy}}$$

The R^2 statistic should be used with caution, because of the following problems:

Note: **It can be shown that** $\boldsymbol{F = \frac{MS_{Reg}}{MS_E} = \left(\frac{n-p-1}{p}\right)\left(\frac{R^2}{1-R^2}\right)}$

Problem 1: Every time you add a predictor to a model, the R-squared increases, even if due to chance alone. It never decreases. Consequently,

a model with more terms may appear to have a better fit simply because it has more terms.

Problem 2: If a model has too many predictors and higher order polynomials, it begins to model the random noise in the data. This condition is known as over fitting the model and it produces misleadingly high R-squared values and a lessened ability to make predictions.

The **adjusted R-squared** is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases ***only if the new term improves the model SS by more than that would be expected by chance***. *It decreases* when a predictor improves the model by less than that expected by chance. The adjusted R-squared can be negative, but it's usually not. It is always lower than the R-squared. This procedure equivalently means that *the model will be a better one if the resulting error mean square is smaller than the earlier one*.

This has led to the modification of R^2 that accounts for the number of predictor variables, p, in the model and uses error mean square. This statistic is called the adjusted R^2 and is defined as

$$R^2_{\text{adj}} = 1 - \frac{MS_E}{MS_y} = 1 - \frac{SS_E/(n-p-1)}{S_{yy}/(n-1)}$$

R^2_{adj} can also be expressed as:

$$R^2_{\text{adj}} = 1 - \left[(1-R^2)\left(\frac{n-1}{n-1-p}\right)\right]$$

In general, R^2_{adj} will increase with increase in R^2, i.e. by addition of regressor variables. But after addition of certain number of regressor variable R^2_{adj} may start decreasing though R^2 may still be increasing. This

will happen in situations when $\left[(1-R^2)\left(\frac{n-1}{n-1-p}\right)\right]$ starts increasing. It may be noted that as number of regressor variables (p) increases:

1) $1-R^2$ decreases,
2) $\frac{n-1}{n-p-1}$ increases, and
3) their product will start increasing when the increase in the second term is not compensated by the decrease in the first term.

In general, $\boldsymbol{R^2_{adj} < R^2}$.

The experimenter would usually select the model with maximum value of $\boldsymbol{R^2_{adj}}$.

In the following output, one can see that the adjusted R-squared peaks in the beginning, and then declines. Meanwhile, the R-squared continues to increase. [Example of Best Subset Regression]. So, as long as R^2 increases significantly, increase in p will result in increase of R^2_{adj}. $[n = 20]$

# of variables (p)	R^2	$1-R^2$	$\frac{n-1}{n-1-p}$	$(1-R^2)\times\frac{n-1}{n-1-p}$	R^2_{adj}
1	0.721	0.279	1.0556	0.2945	0.7055
2	0.859	0.141	1.1176	0.1576	0.8424
3	**0.874**	**0.126**	**1.1875**	**0.1496**	**0.8504**
4	0.879	0.121	1.2667	0.1533	0.8467
5	0.884	0.116	1.3571	0.1574	0.8426

Thus, one might want to include only three predictors in this model. **Generally, it is not advisable to include more terms in the model than necessary**.

Note: R squared adjusted has been written as

$$R^2_{\text{adj}} = 1 - \left[(1 - R^2)\left(\frac{n-1}{n-1-p}\right)\right]$$

So, adjusted R squared will be negative, if

$$(1 - R^2)\frac{n-1}{n-p-1} > 1$$

$$i.e. \quad \frac{n-p-1}{n-1} < (1 - R^2)$$

$$\Rightarrow R^2 - \frac{p}{n-1} < 0$$

Thus, small value of R^2 and a high variable-to-sample ratio may lead to R^2_{adj} becoming negative.

For example, if p = 5 and n = 11, then R^2 must be more than 0.5 in order to R^2_{adj} remain positive.

Residual Analysis: Scaled Residuals

1. The **Standardized Residuals** are residuals scaled w.r.t. MS_E and defined as

$$d_i = \frac{e_i}{\sqrt{\hat{\sigma}^2}} = \frac{e_i}{\sqrt{MS_E}}, \quad i = 1,2,\cdots,n$$

and are often more useful than ordinary residual while assessing residual magnitude. Such residuals have mean zero and approximately unit variance. So, a large standardized residual potentially indicates an outlier.

2. The **Studentized Residuals** are improved residual scaling where ith residuals are divided by its own standard deviation and is defined as

$$r_i = \frac{e_i}{se(e_i)} = \frac{e_i}{\sqrt{\hat{\sigma}^2(1-h_{ii})}} = \frac{e_i}{\sqrt{MS_E(1-h_{ii})}}, \quad i = 1,2,\cdots,n$$

This residual also helps us in **identifying outliers** ($|r_i| \geq 3$).

Standardized Regression

Sometimes it is helpful to work with scaled explanatory and response variables that produce dimensionless regression coefficients. These dimensionless regression coefficients are called as **standardized regression coefficients.** Standardization of the coefficient is usually done to answer the question, which of the independent variables have a greater effect on the dependent variable in a multiple regression analysis when the variables are measured in different units of measurement.

For example, $\hat{y} = 10 + x_1 + 1000x_2$, where y and x_2 are measured in kg and x_1 is measured in gram. Clearly in this case, though $\hat{b}_2 \gg \hat{b}_1$, still 1 kg change in one of the two variables keeping the other fixed has the same impact on the response y.

There are two popular approaches for scaling which gives standardized regression coefficients.

Unit Normal Scaling

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

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, \qquad i = 1,2,\cdots,n \text{ and } j = 1,2,\cdots,p$$

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

where $s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2$ and $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$ are the sample variances of *j*-th explanatory variable and response variable respectively. It may be noted that all scaled explanatory variables and scaled response variable have sample mean equal to 0 and sample variance equal to 1.

Using these new variables, the regression model becomes

$$y_i^* = \gamma_1 z_{i1} + \gamma_2 z_{i2} + \cdots + \gamma_p z_{ip} + e_i', \quad i = 1, 2, \cdots, n \text{ with } \gamma_i = \hat{b}_i \frac{s_i}{s_y}$$

The least squares estimate of $\gamma = \left[\gamma_1, \gamma_2, \cdots, \gamma_p\right]^T$ is

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

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.

Unit Length Scaling

In unit length scaling, we define

$$w_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{SS_j}}, \quad i = 1,2, \cdots, n \qquad j = 1,2, \cdots, p$$

$$y_i^0 = \frac{y_i - \bar{y}}{\sqrt{SS_y}}$$

where $SS_j = \sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2$ is corrected SS for j-th explanatory variable x_j and $SS_y = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the corrected SS for response variable y.

In this scaling, each new explanatory variable w_j has a mean 0 and length unity.

$$\left[\bar{w}_j = \frac{\sum_{i=1}^{n} w_{ij}}{n} = 0; \quad \sqrt{\sum_{i=1}^{n} \left(w_{ij} - \bar{w}_j\right)^2} = 1, \;\; j = 1,2,\cdots,p\right]$$

In terms of these variables, regression model is

$$y_i^0 = \sum_{j=1}^{p} \delta_j w_{ij} + e_i', \;\; i = 1,2,\cdots,n \quad \text{with} \quad \delta_j = \hat{b}_j \sqrt{\frac{SS_j}{SS_y}}$$

The least squares estimate of $\delta = \left[\delta_1, \delta_2, \cdots, \delta_p\right]^T$ is

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

In unit length scaling, the matrix is in the form of correlation matrix, i.e.

$$W^T W = \begin{bmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1p} \\ r_{12} & 1 & r_{23} & \cdots & r_{2p} \\ r_{13} & r_{23} & 1 & \cdots & r_{3p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{1p} & r_{2p} & r_{3p} & \cdots & 1 \end{bmatrix}$$

where $r_{ij} = \sum_{u=1}^{n} w_{ui} w_{uj} = \frac{\sum_{u=1}^{n} (x_{ui} - \bar{x}_i)(x_{uj} - \bar{x}_j)}{\sqrt{SS_i}\sqrt{SS_j}} = \frac{SS_{ij}}{\sqrt{SS_i SS_j}}$ is the simple correlation coefficient between explanatory variables x_i and x_j.

Similarly,

$$W^T y^0 = \begin{bmatrix} \sum w_{i1} y_i^0 \\ \sum w_{i2} y_i^0 \\ \vdots \\ \sum w_{ip} y_i^0 \end{bmatrix} = \begin{bmatrix} r_{1y} \\ r_{2y} \\ \vdots \\ r_{py} \end{bmatrix}$$

where, $r_{jy} = \frac{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)(y_i-\bar{y})}{\sqrt{SS_j SS_y}} = \frac{SS_{jy}}{\sqrt{SS_j SS_y}}$ is the simple correlation coefficient between x_j and y.

It may be noted that $Z^T Z$ matrix is closely related to $W^T W$; in fact

$$Z^T Z = (n-1) W^T W.$$

So the estimates of regression coefficient in unit normal scaling $(\hat{\gamma})$ and unit length scaling $(\hat{\delta})$ are identical. So it does not matter which scaling is used. The regression coefficients obtained after such scaling, viz, $\hat{\gamma}$ or $\hat{\delta}$, are usually called standardized regression coefficients.

The relationship between the original and standardized regression coefficients is

$$\hat{b}_j = \hat{\delta}_j \sqrt{\frac{SS_y}{SS_j}}, \quad j = 1,2,\cdots,p \quad \text{and} \quad \hat{b}_0 = \bar{y} - \sum_{j=1}^{p} \hat{b}_j \bar{x}_j$$

where $\hat{b}_0$ and $\hat{b}_j$, $j = 1,2,\cdots,p$ are respectively OLS estimate of the intercept and slope parameters.

Example 2>

Sr. No.	Original		
	y	x_1	x_2
1	293	1.6	851
2	230	15.5	816
3	172	22	1058
4	91	43	1201
5	113	33	1357
6	125	40	1115
Average	170.6666667	25.85	1066.333333
SS	30245.33333	1255.475	214095.3333
Variance	6049.066667	251.095	42819.06667

Unit Normal Scaling

$$\textbf{Scaled } y_i = \frac{y_i - \bar{y}}{s_y} \qquad \textbf{Scaled } x_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

Sr. No.	Scaled data		
	y	x_1	x_2
1	1.572898	-1.53036	-1.04062119
2	0.762877	-0.65316	-1.209762405
3	0.017143	-0.24296	-0.040271718
4	-1.02431	1.082294	0.650790961
5	-0.74145	0.451219	1.404677519
6	-0.58716	0.892971	0.235186832

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	4.924309443	2.462155	97.58766	0.001862548
Residual	3	0.075690557	0.02523		
Total	5	5			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	3.05966E-16	0.06484621	4.72E-15	1
X1	**-0.74122164**	0.115409648	-6.42253	0.007651
X2	**-0.29713644**	0.115409648	-2.57462	0.082162

R^2	Adj R^2	Standard Error
98.49%	97.48%	0.1588

Unit Length Scaling

Scaled $y_i = \frac{y_i - \bar{y}}{\sqrt{SS_y}}$ **Scaled** $x_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{SS_j}}$

Sr. No.	Scaled data		
	y	x_1	x_2
1	0.70342147	-0.684396387	-0.465379944
2	0.341169	-0.2921032	-0.541022195
3	0.00766672	-0.108656746	-0.01801006
4	-0.45808646	0.484016414	0.291042565
5	-0.3315856	0.2017911	0.628190884
6	-0.26258513	0.39934882	0.105178749

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	0.984861889	0.492431	97.58766	0.001862548
Residual	3	0.015138111	0.005046		
Total	5	1			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	1.2624E-16	0.029000107	4.35E-15	1
X1	**-0.74122164**	0.115409648	-6.42253	0.007651
X2	**-0.29713644**	0.115409648	-2.57462	0.082162

R^2	Adj R^2	Standard Error
98.49%	97.48%	0.0710

Multicollinearity

Multicollinearity occurs when a strong linear relationship exists among the *independent* variables. A strong relationship among the independent variables implies one cannot realistically change one variable without changing other independent variables as well. Moreover, strong relationships between the independent variables make it increasingly difficult to determine the contributions of individual variables.

Multicollinearity is often manifested by one or more nonsensical regression coefficients (e.g. parameter estimates with signs that defy prior knowledge i.e. a model coefficient with a negative sign when a positive sign is expected). In some cases, multiple regression results may seem paradoxical. For instance, the model may fit the data well (significant F-Test), even though none of the X variables has a statistically significant impact on explaining Y. In general, multicollinearity makes interpretations of coefficients very difficult and often impossible.

How is this possible? When two X variables are highly correlated, they both convey essentially the same information. When this happens, the X variables are *collinear* and the results show *multicollinearity*. In case of perfect multicollinearity, $X^T X$ becomes singular and OLS estimator does not exist. Some amount of multicollinearity among the variables often remain due to correlations inherent in the system being studied.

Suppose that there are only two regressor variables, x_1 and x_2. The model, assuming that x_1, x_2 and y are scaled to unit length, is

$$y = \beta_1 w_1 + \beta_2 w_2 + \varepsilon$$

and the least-squared normal equations are

$$\left(W^T W\right)\hat{\beta} = W^T y$$

$$\begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix}\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} r_{1y} \\ r_{2y} \end{bmatrix}$$

where r_{12} is the correlation coefficient between x_1 and x_2, whereas r_{jy} is the same between x_j and y. Now, the inverse of $W^T W$ is

$$C^* = \left(W^T W\right)^{-1} = \begin{bmatrix} \frac{1}{\left(1-r_{12}^2\right)} & \frac{-r_{12}}{\left(1-r_{12}^2\right)} \\ \frac{-r_{12}}{\left(1-r_{12}^2\right)} & \frac{1}{\left(1-r_{12}^2\right)} \end{bmatrix}$$

Therefore, the estimates of the regression coefficients can be obtained as $\left[\text{as } \hat{\beta} = (W^T W)^{-1} W^T y\right]$

$$\hat{\beta}_1 = \frac{r_{1y} - r_{12} r_{2y}}{1 - r_{12}^2}, \quad \hat{\beta}_2 = \frac{r_{2y} - r_{12} r_{1y}}{1 - r_{12}^2}$$

If there is a strong multicollinearity between x_1 and x_2, then the correlation coefficient r_{12} will be large and consequently,

$$Var\left(\hat{\beta}_j\right) = C_{jj}^* \sigma^2 \to \infty \quad and \; Cov\left(\hat{\beta}_1, \hat{\beta}_2\right) = C_{12}^* \sigma^2 \to \pm\infty$$

depending upon whether r_{12} is negative or positive.

Thus, strong multicollinearity between x_1 and x_2 results in large variances and covariances of the least square estimators of the regression coefficients.

Why is multicollinearity a problem?

If the goal is simply to predict Y from a set of X variables, then multicollinearity is not a problem as the regression coefficient estimators are still **unbiased**. The predictions will still be accurate, and the overall R^2 (or $R^2_{adjusted}$ /$R^2_{predicted}$) will quantify how well the model predicts the Y values and will be close to each other.

But, if the goal is how the various X variables impact Y, then multicollinearity is a big problem. One problem, as discussed earlier, is that multicollinearity increases the standard errors of the coefficients. Increased standard errors may lead to an important predictor to become insignificant, whereas without multicollinearity and with lower standard errors, these same coefficients would have been significant.

The other problem is that due to the presence of multicollinearity, confidence intervals on the regression coefficients becomes very wide. The confidence intervals may even include zero, which means one can't even be confident whether an increase in the X value is associated with an increase, or a decrease, in Y.

Detecting multicollinearity

Multicollinearity can be detected by looking at the correlations among pairs of predictor variables. If they are large, we can conclude that the variables are collinear.

Looking at correlations only among pairs of predictors, however, is limiting. It is possible that the pair wise correlations are small, and yet a linear dependence exists among three or even more variables. That's why many regression analysts often rely on what are called **variance inflation factors** (*VIF*) to help detect multicollinearity, which are basically the diagonal elements of C^*.

It can be shown that, if some of the predictors are correlated with the predictor x_k then the variance of b_k is inflated and the same is given by

$$\text{Var}(\hat{b}_k) = \sigma^2 C_{kk}^* = \sigma^2 \times \frac{1}{1 - R_k^2}$$

where R_k^2 is the R^2-value of the model obtained by regressing the k^{th} predictor on the remaining (p-1) predictors. Above shows that the variance of b_k is inflated by the factor $1/\left(1-R_k^2\right)$ because of x_k's linear dependence with other predictors and hence the name. So, formally VIF is defined as

$$VIF(b_k) = \frac{1}{1 - R_k^2}$$

Note that, the greater the linear dependence of the predictor x_k with other predictors, the larger the R_k^2 value. And, as the above formula suggests, the larger the R_k^2 value, the larger will be the corresponding VIF. If $R_k^2 = 0$, then corresponding VIF will be 1, which is the minimum possible value of VIF. It may be noted that VIF exists for each of the predictor variables in a multiple regression model.

The general rule of thumb is that VIFs exceeding 4 (i.e. $R_k^2 > 0.75$) warrant further investigations, while VIF exceeding 10 (i.e. $R_k^2 > 0.9$) is sign of serious multicollinearity and taken as an indication that the multicollinearity may be unduly influencing the least squares estimates.

Dealing with Multicollinearity

There are multiple ways to overcome the problem of multicollinearity.

- ✓ One may use ridge regression or principal component regression or partial least squares regression.
- ✓ The alternate way could be to drop off variables which are resulting in multicollinearity. One may drop of variables which have VIF more than 10.
- ✓ If two or more variables have multicollinearity in the range of 10 or more, remove those variables that have least/low correlation/ impact with the response variable.

Influential Observations

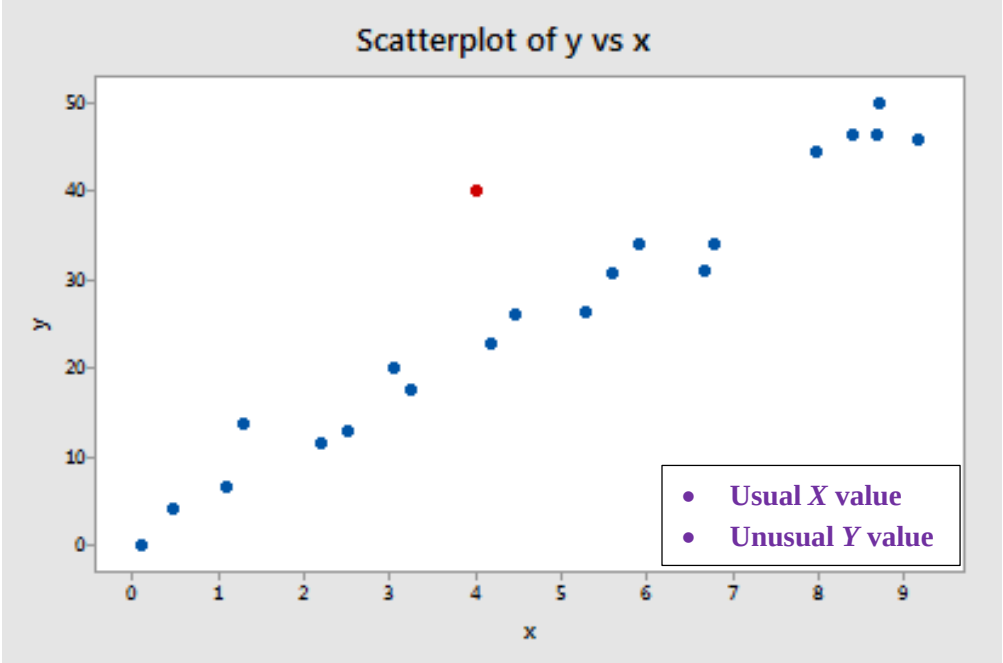
The influence of an observation can be thought of in terms of how much the predicted values for other observations would differ if the observation in question were not included. If the predictions are the same with or without the observation in question, then the observation has no influence on the regression model. If the predictions differ greatly when the observation is not included in the analysis, then the observation is influential.

Outliers

An **outlier** is a data point whose response y does not follow the general trend of the rest of the data.

- An observation whose response value is unusual given its values on the predictor variables (X), resulting in large residual, or error in prediction.

- An outlier may indicate a sample peculiarity or may indicate a data entry error or other problem.

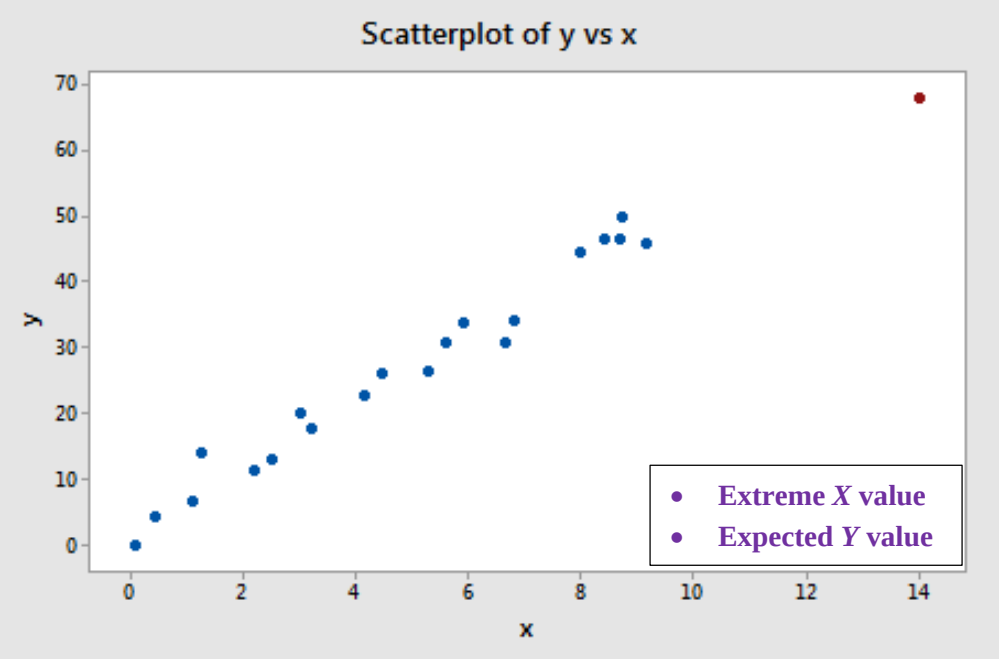


In this case, the red data point, though have a usual X value, but have an unusual Y value and hence will result in a large residual.

Leverage

A data point has high leverage if it has an extreme **predictor value, i.e. X-values**.

- Leverage is a measure of how far a predictor variable deviates from its mean value.
- These leverage points may or may not have an effect on the estimate of regression coefficients.



In this case, the red data point does follow the general trend of the rest of the data. Therefore, it is not deemed an outlier here. However, this point does have an extreme x value, so it does have high leverage.

- Regression Equation **excluding** the High Leverage point:

$$y = 0.0513 + 5.2624\,x$$

- Regression Equation **including** the High Leverage point:

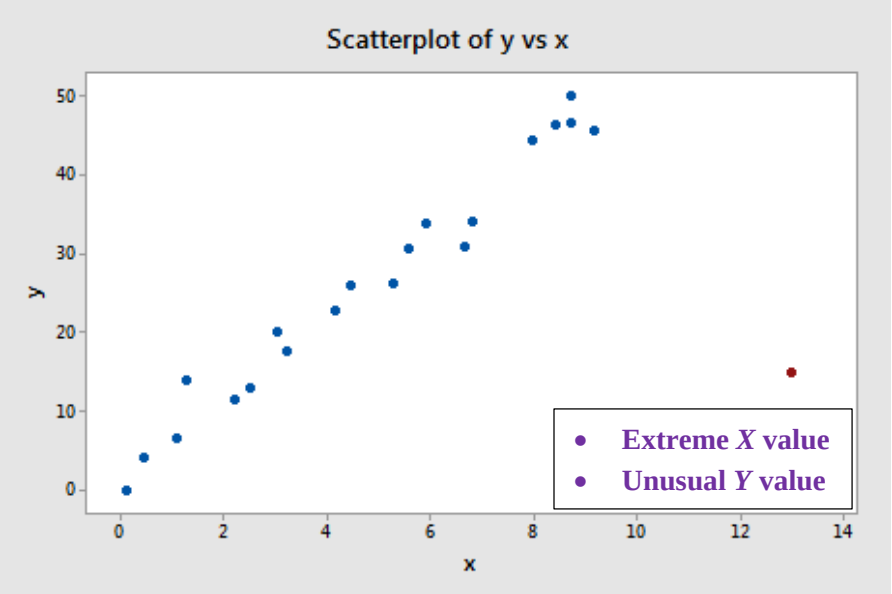
$$y = 0.9092 + 5.0523\,x$$

Influence

When an observation has ***high leverage*** (in terms of X-value) and is an ***outlier*** (in terms of Y-value) it will strongly influence the regression line.

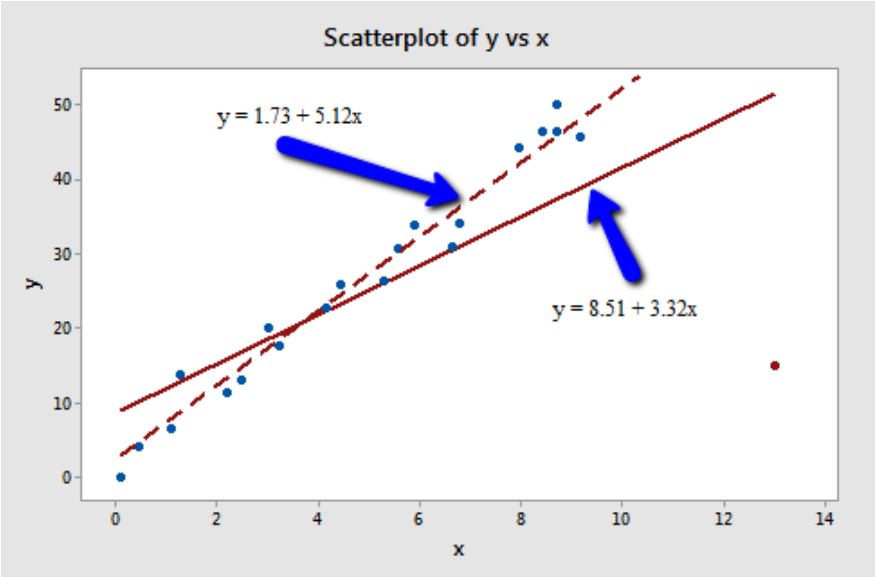
In other words, it must have an unusual X-value with an unusual Y-value *given* its X-value. In such cases both the intercept and slope are affected, as ***the line chases the observation***.

- Influence can be thought of as the product of leverage and error in prediction. **Influence = Leverage X Residual**.
- Removing the observation substantially changes the estimate of coefficients.



In this case, the red data point is most certainly an outlier and has high leverage! The **red** data point does not follow the general trend of the rest of the data and it also has an extreme x value. And, in this case the red data point is influential.

The two best fitting lines — one obtained when the red data point is included and the other obtained when the red data point is excluded:



Effect of Leverage

The greater an observation's leverage, the more potential it has to be an influential observation. For example, an observation with *X*-value equal to the mean on the predictor variable has no influence on the slope of the regression line. On the other hand, an observation that has an unusual *X* value has the potential to affect the slope greatly.

A data point that has an unusual *X* value is known as a **Leverage Point**. The **diagonal elements** $\boldsymbol{h_{ii}}$ of the hat matrix have some useful property: their values are always between 0 and 1, i.e. $0 \leq h_{ii} \leq 1$ and **their sum is** $\boldsymbol{P}$, the number of parameters estimated (including the intercept), i.e. $P = p+1$.

These H values are functions only of the dependent variable (X) values; h_{ii} measures the distance between the X values for the *i*-th data point, i.e. $(X_{i1}, X_{i2}, \cdots\cdots, X_{ip})$ to the mean of all *X* values for all n data points, called the "centroid", i.e. $(\bar{X}_1, \bar{X}_2, \cdots\cdots, \bar{X}_p)$. Each is also called the "**leverage**"; the larger the leverage the point is further away from the centroid.

The fitted value $\hat{y} = Hy$ is linear combination of the observed Y values, where h_{ii} is the weight corresponding to the observation y_i . We can express $\hat{y}_i$ as

$$\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \cdots + h_{ii}y_i + \cdots + h_{in}y_n$$

the **leverage**, h_{ii} quantifies the effect that the observed response y_i has on its predicted value $\hat{y}_i$. That is, if h_{ii} is small, then the observed response y_i plays only a small role in determining the predicted response $\hat{y}_i$. On the other hand, if h_{ii}, is large, then the observed response y_i plays a large role in determining the predicted response $\hat{y}_i$. It's for this reason that h_{ii} are called the "**leverages"**.

Also, since $\sigma^2(e_i) = (1 - h_{ii})\sigma^2$, large h_{ii} will result in small residual variation and will force the fitted value to be closer to the observed value. A leverage value is usually considered to be **large** if it is more than **twice** the mean leverage value (which is $2P/n$).

Data points with high leverage have the potential of moving the regression line up or down as the case may be. Recall that the regression line represents the regression equation in a graphic form, and is represented by the b coefficients. High leverage points make our estimation of b coefficients inaccurate. In such a situation, any

conclusions drawn about the response variable could be misleading. Similarly, any predictions made on the basis of the regression model could be wrong.

Measure of Influence

Cook's Distance

If leverage gives us a warning about data points that have the potential of influencing the regression line, then Cook's Distance indicates how much actual influence each case has on the slope of the regression line.

Cook's Distance is a good measure of the influence of an observation and is proportional to the sum of the squared differences between predictions made with all observations in the analysis and predictions made leaving out the observation in question.

If the predictions are the same with or without the observation in question, then the observation has no influence on the regression model. If the predictions differ greatly when the observation is not included in the analysis, then the observation is influential.

Cook's *Distance* is thus a way of identifying data points that actually do exert too big an influence.

$$D_i = \frac{\sum_{j=1}^{n}\left[\hat{y}_j - \hat{y}_{j(i)}\right]^2}{P \times MS_E}$$

where

- $\hat{y}_j =$ prediction for observation j from the full model,
- $\hat{y}_{j(i)} =$ prediction for j-th observation from the model in which i-th observation has been removed,
- $P = p + 1$ = number of coefficients in the full model, and

- MS_E = mean square error for the full model

Above expression can be algebraically simplified to

$$D_i = \frac{r_i^2}{P} \times \frac{h_{ii}}{1-h_{ii}}, \qquad i = 1, 2, \cdots, n \text{ and } r_i \text{ = Studentized residual.}$$

It may be noted that first component measures how well the model fits the i-th observation y_i (since smaller value of r_i implies better fit) whereas the second component gives the impact of the leverage of the i-th observation.

It may also be noted that D_i is large, if

i) Studentized residual is large, i.e. i-th observation is unusual w.r.t. y-values and

ii) the point is far from the centroid of the *X*-space, that is, if h_{ii} is large, or i-th observation is unusual w.r.t. x-values. In that case i-th data point will have substantial pull on the fit and the second term will be large.

Large values for Cook's Distance signify unusual observations. $D_i > 1$ require careful checking; whereas $D_i > 4$ would indicate that the point has a high influence.

[*Ref*: Cook, R. Dennis (February 1977). "Detection of Influential Observations in Linear Regression", *Technometrics*, **19** (1), pp 15–18]

Difference in Fits (DFFITS)

We may also investigate the deletion influence of the ith observation on the predicted of fitted value. This leads to following diagnostic proposed by Belsley, Kuh and Welsch.

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{S^2_{(i)} h_{ii}}}, \qquad i = 1, 2, \cdots, n$$

where $\hat{y}_{(i)}$ is the fitted value of y_i and $S^2_{(i)}$ is the estimate of σ^2 based on a data set with the i-th observation removed. The denominator is just used for standardization, since

$$\text{var}(\hat{\text{y}}) = \text{var}(\text{Hy}) = \text{H var}(\text{y}) \text{ H}^{\text{T}} = (\sigma^2 \text{I})\text{H} \Rightarrow \text{var}(\hat{\text{y}}_{\text{i}}) = \sigma^2 \text{h}_{\text{ii}}.$$

Thus, $DFFITS_i$ is the change in the fitted value $\hat{y}_i$, in standard deviation term, if the observation i is removed. Belsley, Kuh and Welsch suggest that any observation for which $|DFFITS_i| > 2\sqrt{\frac{p}{n}}$ warrants attention.

Computationally we may find above, after simplification, as

$$\boldsymbol{DFFITS_i = \sqrt{\frac{h_{ii}}{1 - h_{ii}}} \times \frac{e_i}{\sqrt{S^2_{(i)}(1 - h_{ii})}}, \qquad i = 1, 2, \cdots, n}$$

The second term in above expression is known as ***R-Student***, another scaled residual.

[*Ref*: D. A. Belsley, K. Kuh and R. E. Welsch. (1980) John Wiley & Sons, New York: *Regression diagnostics: Identifying influential data and sources of collinearity*]

PRESS Residuals

PRESS residuals are defined as

$$e_{(i)} = y_i - \hat{y}_{(i)} \qquad i = 1, 2, \cdots, n$$

where $\hat{y}_{(i)}$ is the estimated value of y_i based on a model where i-th observation is ignored. This prediction error calculation is repeated for all n observations. The i-th press residual can be simplified to

$$e_{(i)} = \frac{e_i}{1 - h_{ii}} \qquad i = 1, 2, \cdots, n$$

It may be noted that observations having large h_{ii} values will have large PRESS residuals. These observations will in general be **high influence** points.

Note that h_{ii} is always between 0 and 1. If h_{ii} is larger (close to 1), even a small value of residual e_i could result in a larger value of the PRESS residual. On the other hand, if h_{ii} is small (close to 0), even a large value of the ordinary residual e_i may result in a relatively small value of the PRESS residual. Thus, an influential observation is determined not only by the magnitude of residual but also by the corresponding value of leverage h_{ii}.

Example

Table 1 shows the leverage, Studentized residual, and influence for each of the five observations in a small dataset.

Table 1. Example Data.

ID	X	Y	h	r	D
A	1	2	0.39	-1.02	0.40
B	2	3	0.27	-0.56	0.06
C	3	5	0.21	0.89	0.11
D	4	6	0.20	1.22	0.19
E	8	7	0.73	-1.68	8.86

h is the leverage, *r* is the Studentized residual, and *D* is Cook's measure of influence.

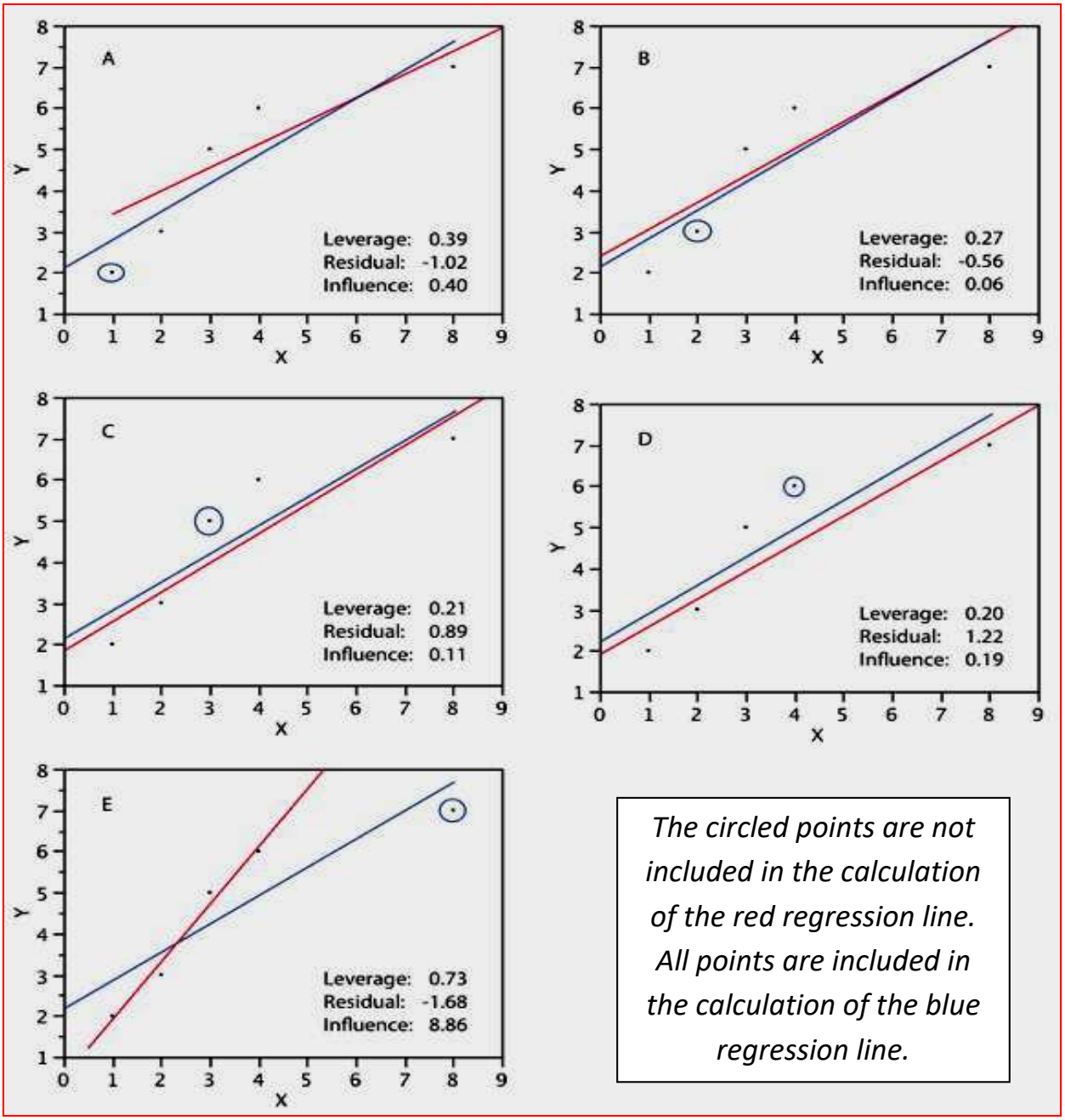
Observation A has fairly high leverage, a relatively high residual and moderately high influence.

Observation B has small leverage and a relatively small residual. It has very little influence.

Observation C has small leverage and a relatively high residual. The influence is relatively low.

Observation D has the lowest leverage and the second highest residual. Although its residual is much higher than Observation A, its influence is much less because of its low leverage.

Observation E has by far the largest leverage and the largest residual. This combination of high leverage and high residual makes this observation extremely influential.



The circled points are not included in the calculation of the red regression line. All points are included in the calculation of the blue regression line.

Selection of variables and Model building

An important problem in many application of regression analysis involves selecting the set of regressor variables to be used in the model. Sometimes, domain knowledge may help the analyst to specify the set of regressor variables to be used in a particular situation. Usually, however, the problem consists of selecting an appropriate set of regressor variables that adequately models the response variable and provides a reasonably good fit. In such a situation, we are interested in **variable selection** that is, screening the candidate variables to obtain a regression model that contains the "**best**" subset of regressor variables.

All Possible regression

This approach requires that the analyst fit all the regression equations involving one candidate variable, all regression equations involving two candidate variables, and so on. Then these equations are evaluated according to some suitable criteria to select the "best" regression model. If there are K candidate regressor, there are 2^K total equations to be examined out of which one is that with no regressor. For example, if K = 4, there are $2^4 = 16$ possible regression equations (1 with no regressor, 4 with one regressor, 6 with 2 regressors, 4 with 3 regressors and 1 with all the regressors); while if $K = 10$, there are $2^{10} = 1024$ possible regression equations. Hence, the number of equations to be examined increases rapidly as the number of candidate variables increases. However, there are some very efficient computing algorithms for all possible regressions available and they are widely implemented in statistical software, so it is a very practical procedure unless the number of candidate regressor is fairly large. Look for a menu choice such as "**Best Subsets**" regression.

Several criteria may be used for evaluating and comparing the different regression models obtained. A commonly used criterion is based on the

value of R^2 or R^2_{adj}. Basically, the analyst continues to increase the number of variables in the model until the increase in R^2 or R^2_{adj} is small. Often, we will find that R^2_{adj} will stabilize and actually begin to decrease as the number of variables in the model increases. Usually, the model that maximizes R^2_{adj} is considered to be a good candidate for the best regression equation. Because we can write $R^2_{adj} = 1 - [MS_E/\{SS_T/(n-1)\}]$ and $SS_T/(n-1)$ is constant, the model that maximizes the R^2_{adj} value also minimizes the mean square error, so this is a very attractive criterion.

Another criterion used to evaluate regression models is the Mallow's C_p statistics that is related to the mean square error of a fitted value and is defined as

$$C_P = \frac{SS_E(P)}{MS_E} - n + 2P$$

where MS_E is the mean square error corresponding to the full $P = p + 1$ term model [see Montgomery, Peck and Vining or Myers]. Generally small values of C_P are desirable, i.e. a model with smaller value of C_P is considered to be better among the candidate regression models. For the full model involving $P = p+1$ coefficients, $C_P = P$.

The **PRESS** statistic can also be used to evaluate competing regression models. PRESS is an acronym for **Prediction Error Sum of Squares,** and it is defined as the sum of the squares of the differences between each observation y_i and the corresponding predicted value based on a model fit by ignoring ith observation, say $\hat{y}_{(i)}$(i.e. PRESS residuals). So PRESS provides a measure of how well the model is likely to perform when predicting *new* data, i.e. a data that was not used to fit the regression model.

The computing formula for PRESS is

$$\text{PRESS} = \sum_{i=1}^{n} \left[y_i - \hat{y}_{(i)} \right]^2 = \sum_{i=1}^{n} \left(\frac{e_i}{1 - h_{ii}} \right)^2$$

where $e_i = y_i - \hat{y}_i$ is the usual residual. Thus PRESS is easy to calculate from the standard least squares regression results.

A better regression model should be less sensitive to each individual observation. In other words, a better regression model should be less impacted by excluding one observation, that is should have a small value of $(y_i - \hat{y}_i)$ for all i. Therefore, a regression model with a smaller value of the PRESS statistic should be a preferred model.

The PRESS statistic can be used to compute an R^2-like statistic for prediction that would give the predictive capability of the model while predicting new observations.

$$R^2_{\text{prediction}} = 1 - \frac{\text{PRESS}}{\text{SS}_\text{T}}$$

$R^2_{\text{prediction}}$ value of, say, 0.9209 would mean that we expect the model to explain about 92.09% of the variability in predicting new observations.

Stepwise Regression

Stepwise Regression is probably the most widely used variable selection technique. The procedure iteratively constructs a sequence of regression models by adding or removing variables at each step. The criterion for adding or removing a variable at any step is usually expressed in terms of a partial F-test. Let f_{in} be the value of the F-random variable for adding a variable to the model, and let f_{out} be the value of the *F*-random variable for removing a variable from the model. We must have $f_{in} \geq f_{out}$, and usually $f_{in} = f_{out}$.

Stepwise regression begins by forming a one-variable model using the regressor variable that has the highest correlation with the response variable Y. This will also be the regressor producing the largest F-statistic. For example, suppose that at this step, x_1 is selected. At the second step, the remaining K - 1 candidate variables are examined, and the variable for which the partial F-statistic

$$F_j = \frac{SS_R(\beta_j|\beta_1, \beta_0)}{MS_E(x_j, x_1)} \qquad (1)$$

is a maximum is added to the equation, provided that $f_j > f_{in}$. In equation 1, $MS_E(x_j, x_1)$ denotes the mean square for error for the model containing both x_1 and x_j. Suppose that this procedure indicates that x_2 should be added to the model. Now the stepwise regression algorithm determines whether the variable x_1 added at the first step should be removed. This is done by calculating the F-statistic

$$F_1 = \frac{SS_R(\beta_1|\beta_2, \beta_0)}{MS_E(x_1, x_2)} \qquad (2)$$

If the calculated value $f_1 < f_{out}$, the variable x_1 is removed; otherwise it is retained, and we would attempt to add a regressor to the model containing both x_1 and x_2.

In general, at each step the set of remaining candidate regressor variables are examined, and the regressor with the largest partial F-statistic is entered, provided that the observed value of F exceeds f_{in}. Then the partial F-statistic for each regressor already in the model is calculated and the regressor, with the smallest observed value of F, is deleted if the observed $f < f_{out}$. The procedure continues until no other regressor variables can be added to or removed from the model.

Stepwise regression is almost always performed using a computer program. The analyst exercises control over the procedure by the choices of f_{in} and f_{out}. Some stepwise regression computer programs require that numerical values be specified for f_{in} and f_{out}. Since the number of degrees of freedom on MS_E depends on the number of variables in the model, which changes from step to step, a fixed value of f_{in} and f_{out} causes the type I and type II error rates to vary. Some computer programs allow the analyst to specify the type I error levels for f_{in} and f_{out}. Sometimes it is useful to experiment with different values of f_{in} and f_{out} (or different type I error levels) in several different runs to see if this substantially affects the choice of the final model.

Forward Selection

The **forward selection** procedure is a variation of stepwise regression and is based on the principle that regressor variables should be added to the model one at a time until there are no remaining candidate regressor variables that produce a significant increase in the regression sum of squares. That is, variables are added one at a time as long as their partial F-value exceeds f_{in}. Forward selection is a simplification of stepwise regression that omits the partial F-test for deleting variables from the model that have been added at previous steps. This is a potential weakness of forward selection; that is, the procedure does not explore the effect that adding a regressor at the current step has on regressor variables added at earlier steps. Notice that forward selection method

will give exactly the same model, if stepwise regression terminated without deleting a variable.

Backward Elimination

The **backward elimination** algorithm begins with all K candidate regressor variables in the model. Then the regressor with the smallest partial F-statistic is deleted if this F-statistic is insignificant, that is, if $f < f_{out}$. Next, the model with K-1 regressor is fit, and the next regressor for potential elimination is found. The algorithm terminates when no further regressor can be deleted.

Some Comments on Final Model Selection

We have illustrated several different approaches to the selection of variables in multiple linear regressions. The final model obtained from any model-building procedure should be subjected to usual adequacy checks, such as residual analysis, lack-of-fit testing and examination of the effect influential points. A major criticism of variable selection methods such as stepwise regression is that the analyst may conclude there is one "best" regression equation. Generally, this is not the case, because several equally good regression models can often be obtained. One way to avoid this problem is to use several different model-building techniques and see if different models result.

If the number of candidate regressor is not too large, the all-possible regressions method is recommended. It is usually recommended using the MSE, PRESS and C_P evaluation criterion conjunction with this procedure. The all-possible regressions approach can find the "best" regression equation with respect to above criteria, while stepwise-type methods offer no such assurance. Furthermore, the all-possible regressions procedure is not distorted by multicollinearity among the regressor, as stepwise-type methods are.

Example

Following table presents data concerning the heat evolved in calories per gram of cement (y) as a function of the amount of each of four ingredients in the mix: tricalcium aluminate (x_1), tricalcium silicate (x_2), tetracalcium alumina ferrite (x_3) and dicalcium silicate (x_4).

y	x_1	x_2	x_3	x_4
78.5	7	26	6	60
74.3	1	29	15	52
104.3	11	56	8	20
87.6	11	31	8	47
95.9	7	52	6	33
109.2	11	55	9	22
102.7	3	71	17	6
72.5	1	31	22	44
93.1	2	54	18	22
115.9	21	47	4	26
83.8	1	40	23	34
113.3	11	66	9	12
109.4	10	68	8	12

Note: Analyzed using Minitab 17.

Multiple Linear Regression: y versus x_1, x_2, x_3, x_4

Analysis of Variance

Source	DF	SS	MS	F-value	P-value
Regression	4	2667.90	666.975	111.48	0.000
X_1	1	25.951	25.951	4.34	0.071
X_2	1	2.972	2.972	0.50	0.501
X_3	1	0.109	0.109	0.02	0.896
X_4	1	0.247	0.247	0.04	0.844
Error	8	47.86	5.893		
Total	12	2715.76			

Model Summary

$\sqrt{\mathbf{MSE}}$	$\mathbf{R^2}$	$\mathbf{R^2_{adjusted}}$	$\mathbf{R^2_{predicted}}$
2.44601	98.24%	97.36%	95.94%

Coefficients

Term	Coefficient	SE(Coeff)	t-value	P-value	VIF
Constant	62.4	70.1	0.89	0.399	
X_1	1.551	0.745	2.08	0.071	38.50
X_2	0.510	0.724	0.70	0.501	254.42
X_3	0.102	0.755	0.14	0.896	46.87
X_4	-0.144	0.709	-0.20	0.844	282.51

Regression Equation

$$y = 62.4 + 1.551x_1 + 0.510x_2 + 0.102x_3 - 0.144x_4$$

Note that due to the presence of multicollinearity, standard error of regression coefficients is quite large. So, the 95% confidence interval will be very wide, sometime the same would even include zero (0).

Multiple Linear Regression: y versus x_1, x_2, x_3

Analysis of Variance

Source	DF	SS	MS	F-value	P-value
Regression	3	2667.65	889.22	166.34	0.000
X_1	1	367.33	367.33	68.72	0.000
X_2	1	1178.96	1178.96	220.55	0.000
X_3	1	9.79	9.79	1.83	0.209
Error	9	48.11	5.35		
Total	12	2715.76			

Model Summary

$\sqrt{\mathbf{MSE}}$	$\mathbf{R^2}$	$\mathbf{R^2_{adjusted}}$	$\mathbf{R^2_{predicted}}$
2.31206	98.23%	97.64%	96.69%

Coefficients

Term	Coefficient	SE(Coeff)	t-value	P-value	VIF
Constant	48.19	3.91	12.32	0.000	
X_1	1.696	0.205	8.29	0.000	3.25
X_2	0.657	0.044	14.85	0.000	1.06
X_3	0.250	0.185	1.35	0.209	3.14

Regression Equation

$$y = 48.19 + 1.696x_1 + 0.657x_2 + 0.250x_3$$

Note: To begin with, variable X_4 is removed as it has highest VIF as well has least impact on the response variable (refer the full model involving all 4 regressor variables).

Multiple Linear Regression: y versus x_1, x_3, x_4

Analysis of Variance

Source	DF	SS	MS	F-value	P-value
Regression	3	2664.93	888.31	157.27	0.000
X_1	1	124.90	124.90	22.11	0.001
X_3	1	23.93	23.93	4.24	0.070
X_4	1	1176.24	1176.24	208.24	0.000
Error	9	50.84	5.65		
Total	12	2715.76			

Model Summary

$\sqrt{\mathbf{MSE}}$	$\mathbf{R^2}$	$\mathbf{R^2_{adjusted}}$	$\mathbf{R^2_{predicted}}$
2.37665	98.13%	97.50%	96.52%

Coefficients

Term	Coefficient	SE(Coeff)	t-value	P-value	VIF
Constant	111.68	4.56	24.48	0.000	
X_1	1.052	0.224	4.70	0.001	3.68
X_3	-0.410	0.199	-2.06	0.070	3.46
X_4	-0.643	0.044	-14.43	0.000	1.18

Regression Equation

$$y = 111.68 + 1.052x_1 - 0.410x_3 - 0.643x_4$$

Summary – Multiple Linear Regression

Predictors in the Model	$\sqrt{\mathbf{MSE}}$	$\mathbf{R^2}$	$\mathbf{R^2_{adjusted}}$	$\mathbf{R^2_{predicted}}$
X_1, X_2, X_3, X_4	2.44601	98.24%	97.36%	95.94%
$\boldsymbol{X_1, X_2, X_3}$	**2.31206**	**98.23%**	**97.64%**	**96.69%**
X_1, X_3, X_4	2.37665	98.13%	97.50%	96.52%

Above clearly shows that multicollinearity does not pose much problem if the goal is simply to predict *Y* for a given value of *X*.

Best Subsets Regression: y versus x_1, x_2, x_3, x_4

# of Vars	R^2	R^2 (Adj)	PRESS	R^2 (Pred)	Mallows C_P	$\sqrt{MS_E}$	x_1	x_2	x_3	x_4
1	67.5	64.5	1194.2	56.0	138.7	8.9639				√
	66.6	63.6	1202.1	55.7	142.5	9.0771		√		
2	97.9	97.4	93.9	96.5	2.7	2.4063	√	√		
	97.2	96.7	121.2	95.5	5.5	2.7343	√			√
3	98.2	97.6	85.4	96.9	3.0	2.3087	√	√		√
	98.2	97.6	90.0	96.7	3.0	2.3121	√	√	√	
4	98.2	97.4	110.3	95.9	5.0	2.4460	√	√	√	√

Note: For each variable size, summary of two best models are tabulated.

Stepwise Selection of Terms

Candidate terms: x_1, x_2, x_3, x_4

	Step 1		Step 2		Step 3		Step 4	
	Coeff	P-value	Coeff	P-value	Coeff	P-value	Coeff	P-value
Constant	117.57		103.10		71.6		52.58	
x_4	-0.738	0.001	-0.614	0.000	-0.237	**0.205**		
x_1			1.440	0.000	1.452	0.000	1.468	0.000
x_2					0.416	0.052	0.6623	0.000
Sqrt MSE	8.9639		2.7343		2.3087		2.4063	
R^2	67.45%		97.25%		98.23%		97.87%	
R^2 (Adj)	64.50%		96.70%		97.64%		97.44%	
R^2 (Pred)	56.03%		95.54%		96.86%		96.54%	
C_p	138.73		5.50		3.02		2.68	

α to enter = 0.15, α to remove = 0.15

More than α to remove, hence removed in Step 4.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F-value	P-value
Regression	2	2657.86	1328.93	229.52	0.000
Error	10	57.90	5.79		
Total	12	2715.76			

Model Summary

Sqrt MSE	R^2	R^2 (Adj)	R^2 (Pred)
2.40634	97.87 %	97.44%	96.54%

Coefficients

Term	Coefficient	SE Coeff.	t-value	P-value	VIF
Constant	52.58	2.29	23.00	0.000	
x_1	1.468	0.121	12.10	0.000	1.06
x_2	0.6623	0.0459	14.44	0.000	1.06

Regression Equation

$y = 52.58 + 1.468\ x_1 + 0.6623\ x_2$

Forward Selection of Terms

Candidate terms: x1, x2, x3, x4

	-----Step 1----		-----Step 2-----		-----Step 3----	
	Coeff.	**P**	**Coeff.**	**P**	**Coeff.**	**P**
Constant	117.57		103.10		71.6	
x4	-0.738	0.001	-0.6140	0.000	-0.237	0.205
x1			1.440	0.000	1.452	0.000
x2					0.416	0.052
S		8.96390		2.73427		2.30874
R-square		67.45%		97.25%		98.23%
R-square(Adj)		64.50%		96.70%		97.64%
Mallows' Cp		138.73		5.50		3.02

$\alpha_E = 0.15$

Note: In this case the variable with smallest *t*-test P-value less than $\alpha_E = 0.15$ is added to the model in each step.

Backward Elimination of Terms

Candidate terms: x1, x2, x3, x4

	-----Step 1----		-----Step 2----		-----Step 3----	
	Coeff	**P**	**Coeff**	**P**	**Coeff**	**P**
Constant	62.4		71.6		52.58	
x1	1.551	0.071	1.452	0.000	1.468	0.000
x2	0.510	0.501	0.416	0.052	0.6623	0.000
x3	0.102	**0.896**				
x4	-0.144	0.844	-0.237	**0.205**		
S		2.44601		2.30874		2.40634
R-square		98.24%		98.23%		97.87%
R-square(Adj)		97.36%		97.64%		97.44%
Mallows' Cp		5.00		3.02		2.68

$\alpha_R = 0.15$

Note: In this case the variable with largest *t*-test P-value greater than $\alpha_R = 0.15$ is removed from the model in each step.

Forward Selection and Backward Elimination

Method	Variables added/removed	Regression Equation	R^2 (%)	R^2_{adj} (%)	R^2_{pred} (%)
Forward Selection	x_4 x_4, x_1 x_4, x_1, x_2	$y = 71.6 + 1.452x_1 + 0.416x_2 - 0.237x_4$	98.23	97.64	96.86
Backward Elimination	x_1, x_2, x_3, x_4 x_1, x_2, x_4 x_1, x_2	$y = 52.58 + 1.468x_1 + 0.662x_2$	97.87	97.44	96.54

Dummy Variables in Regression

A dummy (indicator) variable in is an artificial variable created to represent an attribute with two or more distinct categories / levels.

Why used

Regression analysis treats all independent variables (X) in the analysis as numerical. Numerical variables are interval or ratio scale variables whose values are directly comparable, e.g. '10 is twice as much as 5' or '3 minus 1 equals 2'. Often however, one might want to include an attribute or nominal scale variable such as "Product Band' or 'Type of Defect' in his/her analysis. Say one may have three types defect, numbered '1', '2' and '3'. In this case '3 minus 1' doesn't mean anything. Here the numbers are used merely to indicate or identify the different types of defect and hence do not have any intrinsic meaning of their own. Dummy variables are created in such situation to 'trick' the regression algorithm to correctly analyze attribute variables.

Example> For expressing the categorical variable "Gender" (male or female), one requires only one dummy variable:

Gender	G
Male	0
Female	1

Example> To express the categorical variable "Education", where possible outcome could be –Secondary, Higher Secondary, Graduate and Post Graduate, one need to consider three dummy variables:

Education	Z_1	Z_2	Z_3
Post Graduate	1	0	0
Graduate	0	1	0
Higher Secondary	0	0	1
Secondary	0	0	0

Thus, the number of dummy variables necessary to represent a single attribute variable is equal to **number of categories in that variable – 1**.

Moreover, the interactions of two attribute variables (e.g. Gender and Marital status) is represented by a third dummy variable which is simply the product of the two individual dummy variables.

Suppose the regression model involving *income*, age (X_1), gender (G) and education, with categories as stated above, is:

$$Y = b_0 + b_1 X_1 + b_2 G + b_3 Z_1 + b_4 Z_2 + b_5 Z_3 + \varepsilon$$

Now let us study above relationship under different conditions:

Gender	Education	Derived Model
Male	PG	$Y = (b_0 + b_3) + b_1 X_1 + \varepsilon$
	Graduate	$Y = (b_0 + b_4) + b_1 X_1 + \varepsilon$
	HS	$Y = (b_0 + b_5) + b_1 X_1 + \varepsilon$
	Secondary	$Y = b_0 + b_1 X_1 + \varepsilon$
Female	PG	$Y = (b_0 + b_2 + b_3) + b_1 X_1 + \varepsilon$
	Graduate	$Y = (b_0 + b_2 + b_4) + b_1 X_1 + \varepsilon$
	HS	$Y = (b_0 + b_2 + b_5) + b_1 X_1 + \varepsilon$
	Secondary	$Y = (b_0 + b_2) + b_1 X_1 + \varepsilon$

It may be noted that all the above models are parallel to each other with different intercept, i.e. they have common slope b_1 and different intercepts. So, the slope b_1 does not depend on the categorical variable, whereas the categorical variable does affect the intercept.

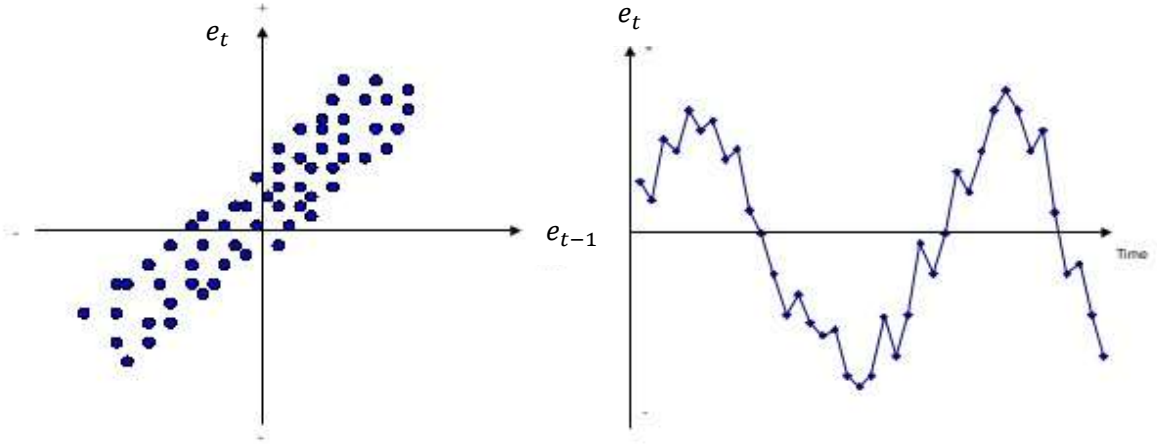
Autocorrelation

The fundamental assumptions in linear regression are that the error terms ε_i have mean zero and constant variance and uncorrelated $\left[E(\varepsilon_i) = 0,\ Var(\varepsilon_i) = \sigma^2\ and\ E(\varepsilon_i\varepsilon_j) = 0, i \neq j\right]$. For purposes of testing hypotheses and constructing confidence intervals we often add the assumption of normality, so that the ε_i's are $NID(0, \sigma^2)$. Some applications of regression involve regressor and response variables that have a natural sequential order over time. Such data are called **time series data**. Regression models using time series data occur quite often in economics, business, and some fields of engineering. The assumption of uncorrelated or independent errors for time series data is often not appropriate. Usually the errors in time series data exhibit **serial correlation**, that is, $E(\varepsilon_i\varepsilon_j) = 0,\ i \neq j$. Such error terms are said to be **auto correlated**. Because time series data occur frequently in business and economics, much of the basic methodology appears in the economics literature.

Residual plots can be useful for the detection of autocorrelation. The most meaningful display is the plot of residuals versus time.

Positively auto correlated residuals

If autocorrelation is present, positive autocorrelation is the most likely outcome. *Positive autocorrelation* occurs when an error of a given sign tends to be followed by an error of the same sign. For example, positive errors are usually followed by positive errors, and negative errors are usually followed by negative errors. So, if there is positive autocorrelation, residuals of identical sign occur in clusters. That is, there is not enough changes of sign in the pattern of residuals.

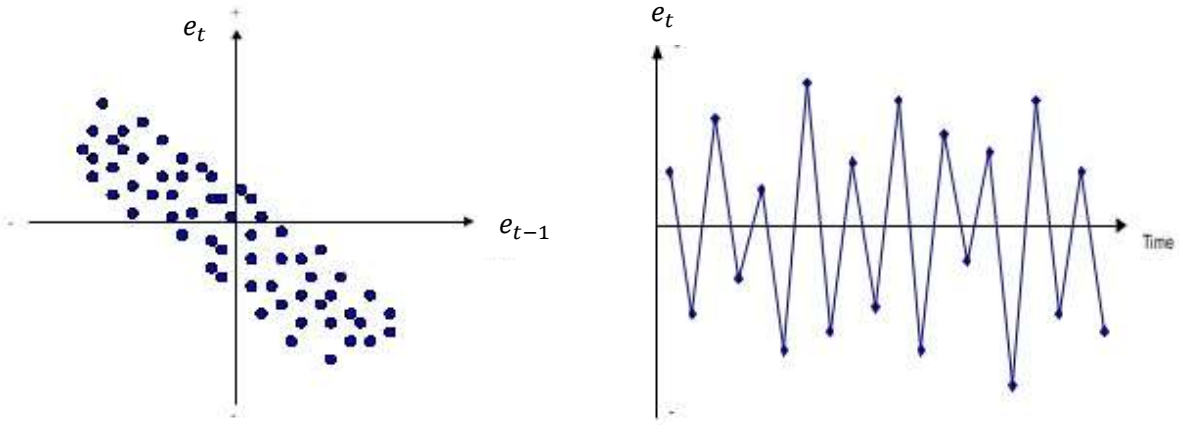


Positive autocorrelation is indicated by a cyclical residual plot over time. So, here difference between two successive errors will tend to be small.

For example, the temperatures on different days in a month are positively auto correlated. The temperature the next day tends to rise when it's been increasing and tends to drop when it's been decreasing during the previous days.

Negatively auto correlated residuals

On the other hand, if there is negative autocorrelation although unlikely, an error of a given sign tends to be followed by an error of the opposite sign, that is, the residuals will alternate signs too rapidly.



Negative autocorrelation is indicated by an alternating pattern where the residuals cross the time axis more often that if they were distributed randomly. So, here difference between two successive errors will tend to be large.

Various **statistical tests** can be used to detect the presence of autocorrelation. The test developed by Durbin and Watson is widely used. This test is based on the assumption that the errors in the regression model are generated by a **first-order autoregressive process** observed at equally spaced time periods, that is,

$$\varepsilon_t = \rho\varepsilon_{t-1} + a_t \qquad (4.1)$$

where ε_t is the error term in the model at time period t, a_t is an $NID(0, \sigma_a^2)$ random variable and ρ $(|\rho| < 1)$ is the **autocorrelation parameter**. Thus, a simple linear regression model with **first-order autoregressive errors** would be

$$y_t = b_0 + b_1 x_t + \varepsilon_t$$
$$\varepsilon_t = \rho\varepsilon_{t-1} + a_t \qquad (4.2)$$

Where y_t and x_t are the observations on the response and regressor variables at time period t. The white noise a_t is assumed to be

independently and identically distributed with zero mean and constant variance so that $E(a_t) = 0,\ E(a_t^2) = \sigma_a^2$ and $E(a_t a_{t+u}) = 0$ for all $u \neq 0$.

By successively substituting for $\varepsilon_{t-1}, \varepsilon_{t-2}, \cdots$ on the right hand side of equation (4.1), we obtain

$$\varepsilon_t = \sum_{u=0}^{\infty} \rho^u a_{t-u}$$

Thus, the error term for period *t* is just a linear combination of all current and previous realizations of the $NID(0, \sigma_a^2)$ random variables a_t. Furthermore, we can also show that

$$E(\varepsilon_t) = 0$$

$$Var(\varepsilon_t) = \sigma_a^2 \left(\frac{1}{1-\rho^2} \right)$$

$$Cov(\varepsilon_t, \varepsilon_{t+u}) = \rho^u Var(\varepsilon_t) = \rho^u \sigma_a^2 \left(\frac{1}{1-\rho^2} \right)$$

That is, the errors have zero mean and constant variance but are auto correlated unless $\rho = 0$.

Because most regression problems involving time series data exhibit positive autocorrelation, the hypotheses usually considered in the Durbin-Watson test are

$$H_0: \rho = 0$$
$$H_1: \rho > 0$$

The test statistics used is

$$d = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

where the e_t, t = 1,2,….,n are the residuals from an ordinary least-squares analysis applied to the (y_t, x_t) data. It may be noted that, d becomes smaller as the serial correlation increases. It can be shown that $d \cong 2(1-\rho)$. So, $d \cong 2$ indicates no autocorrelation. Since ρ can take values between -1 and +1, the value of d lies between 0 and 4.

Small values of d indicate successive error terms are, on average, close in value to each other, or positively correlated. Thus, if the Durbin–Watson statistic is substantially less than 2, there is evidence of positive serial correlation. On the other hand, if $d > 2$, successive error terms are, on average, much different in value from each other, i.e., negatively correlated.

Testing positive autocorrelation

We have shown under null hypothesis, $d \cong 2$, otherwise $d < 2$ for positive autocorrelation. So, decision rule could be

i) $d = 2$: no autocorrelation, and
ii) $0 < d < 2$: positive autocorrelation. Smaller the value, larger is the probability of positive autocorrelation.

The exact distribution of d depends on ρ, which is unknown, as well as on the observations on the X-variable. Durbin and Watson in their paper [*"Testing for serial correlation in least square regression II", Biometrika, 38, 159-178, 1951*] showed that d lies between two bounds, say d_L and d_U, such that if d is outside these limits, a conclusion regarding the hypothesis can be reached. The decision procedure is as follows

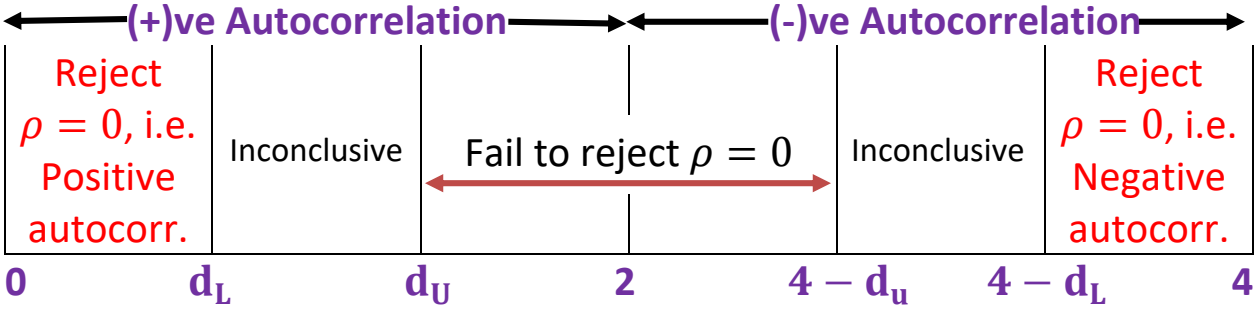
If $d < d_L$	reject H_0: $\rho = 0$	⇒ presence of positive autocorrelation
If $d > d_U$	fail to reject H_0: $\rho = 0$	⇒ no autocorrelation
If $d_L < d < d_U$	test is inconclusive.	

Testing negative autocorrelation

Situations where negative autocorrelation occurs are not often encountered. However, if a test for negative autocorrelation is desired, one can use the statistic 4 - d. From earlier discussion, it is apparent that for negative autocorrelation we must get $2 < d < 4$ and so, $4 - d$ will lie in the interval $(0, 2)$. Also larger the value of d greater is the possibility of negative autocorrelation.

Thus the decision rules for $H_0: \rho = 0$ versus $H_1: \rho < 0$ are the same as those used in testing for positive autocorrelation.

Graphically, the testing procedure can be depicted as:



It is also possible to conduct a two-side test ($H_0: \rho = 0$ versus $H_1: \rho \neq 0$) by using both one-side tests simultaneously. If this is done, the two-sided procedure has Type I error 2α, where α is the Type I error used for each one-sided test.

Durbin-Watson statistic: Significance points of d_L and d_u at 5% level of significance

k' = number of explanatory variables excluding the constant term

obs	$k'=1$		$k'=2$		$k'=3$		$k'=4$		$k'=5$	
N	d_L	d_u	d_L	d_u	d_L	d_u	d_L	d_u	d_L	d_u
6	0.610	1.400	-	-	-	-	-	-	-	-
7	0.700	1.356	0.467	1.896	-	-	-	-	-	-
8	0.763	1.332	0.559	1.777	0.368	2.287	-	-	-	-
9	0.724	1.320	0.629	1.699	0.455	2.128	0.296	2.588	-	-
10	0.879	1.320	0.697	1.641	0.525	2.016	0.376	1.414	0.243	2.822
11	0.927	1.324	0.658	1.604	0.595	1.928	0.444	2.283	0.316	2.645
12	0.971	1.331	0.812	1.579	0.658	1.864	0.512	2.177	0.379	2.506
13	1.010	1.340	0.861	1.562	0.715	1.816	0.574	1.094	0.445	2.390
14	1.045	1.350	0.905	1.551	0.767	1.779	0.632	2.030	0.505	2.296
15	1.077	1.361	0.946	1.543	0.814	1.750	0.685	1.977	0.562	2.220
16	1.106	1.371	0.982	1.539	0.857	1.728	0.734	1.935	0.615	2.157
17	1.133	1.381	1.015	1.536	0.897	1.710	0.779	1.900	0.664	2.104
18	1.158	1.391	1.046	1.535	0.933	1.696	0.820	1.872	0.710	2.060
19	1.180	1.401	1.074	1.536	0.967	1.685	0.859	1.848	0.752	2.023
20	1.201	1.411	1.100	1.537	0.998	1.676	0.894	1.828	0.792	1.991
obs	$k'=1$		$k'=2$		$k'=3$		$k'=4$		$k'=5$	
N	d_L	d_u	d_L	d_u	d_L	d_u	d_L	d_u	d_L	d_u
21	1.221	1.420	1.125	1.538	1.026	1.669	0.927	1.812	0.829	1.964
22	1.239	1.429	1.147	1.541	1.053	1.664	0.958	1.797	0.863	1.940
23	1.257	1.437	1.168	1.543	1.078	1.660	0.986	1.785	0.895	1.920
24	1.273	1.446	1.188	1.546	1.101	1.656	1.013	1.775	0.925	1.902
25	1.288	1.454	1.206	1.550	1.123	1.654	1.038	1.767	0.953	1.886
26	1.302	1.461	1.224	1.553	1.143	1.652	1.062	1.759	0.979	1.873
27	1.316	1.469	1.240	1.556	1.162	1.651	1.084	1.753	1.004	1.861
28	1.328	1.476	1.255	1.560	1.181	1.650	1.104	1.747	1.028	1.850
29	1.341	1.483	1.270	1.563	1.198	1.650	1.124	1.743	1.050	1.841
30	1.352	1.489	1.284	1.567	1.214	1.650	1.143	1.739	1.071	1.833
31	1.363	1.496	1.297	1.570	1.229	1.650	1.160	1.735	1.090	1.825
32	1.373	1.502	1.309	1.574	1.244	1.650	1.177	1.732	1.109	1.819
33	1.383	1.508	1.321	1.577	1.258	1.651	1.193	1.730	1.127	1.813
34	1.993	1.514	1.333	1.580	1.271	1.652	1.208	1.728	1.144	1.808
35	1.402	1.519	1.343	1.584	1.283	1.653	1.222	1.726	1.160	1.803
36	1.411	1.525	1.354	1.587	1.295	1.654	1.236	1.724	1.175	1.799
obs	$k'=1$		$k'=2$		$k'=3$		$k'=4$		$k'=5$	
N	d_L	d_u	d_L	d_u	d_L	d_u	d_L	d_u	d_L	d_u
37	1.419	1.530	1.364	1.590	1.307	1.655	1.249	1.723	1.190	1.795
38	1.427	1.535	1.373	1.594	1.318	1.656	1.261	1.722	1.204	1.792
39	1.435	1.540	1.382	1.597	1.328	1.658	1.273	1.722	1.218	1.789
40	1.442	1.544	1.391	1.600	1.338	1.659	1.285	1.721	1.230	1.786
45	1.475	1.566	1.430	1.615	1.383	1.666	1.336	1.720	1.287	1.776
50	1.503	1.585	1.462	1.628	1.421	1.674	1.378	1.721	1.335	1.771
55	1.528	1.601	1.490	1.641	1.452	1.681	1.414	1.724	1.374	1.768
60	1.549	1.616	1.514	1.652	1.480	1.689	1.444	1.727	1.408	1.767
65	1.567	1.629	1.536	1.662	1.503	1.696	1.471	1.731	1.438	1.767
70	1.583	1.641	1.554	1.672	1.525	1.703	1.494	1.735	1.464	1.768
75	1.598	1.652	1.571	1.680	1.543	1.709	1.515	1.739	1.487	1.770
80	1.611	1.662	1.586	1.688	1.560	1.715	1.534	1.743	1.507	1.772
85	1.624	1.671	1.600	1.696	1.575	1.721	1.550	1.747	1.525	1.774
90	1.635	1.679	1.612	1.703	1.589	1.726	1.566	1.751	1.542	1.776
95	1.645	1.687	1.623	1.709	1.602	1.732	1.579	1.755	1.557	1.778
100	1.654	1.694	1.634	1.715	1.613	1.736	1.592	1.758	1.571	1.780
150	1.720	1.746	1.706	1.760	1.693	1.774	1.679	1.788	1.665	1.802
200	1.758	1.778	1.748	1.789	1.738	1.799	1.728	1.810	1.718	1.820