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 $cov(e_i, e_j) = -\sigma^2 h_{ij} \quad (i \neq j)$.

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:

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

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

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,

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

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^*.