Chapter 2: Multiple Linear Regression

In **multiple regressions**, the mean of the response variable is a function of *two or more explanatory variables*.

In Chapter 1 we examined the relationship between HSGPA and College GPA. There are some other possible factors that may be related to College GPA, such as ACT Scores, Rank in high school class, etc.

The corresponding multiple linear regression model in this case would look like this:

College GPA = β_0 + β_1 (HSGPA) + β_2 (ACT Scores) + β_3 (Rank) + ε

where the parameters β_0, β_1, β_2 and β_3 would be estimated from the data.

2.1 <u>Multiple Linear Regression Model and Assumptions</u>

The multiple linear regression model with k regressors or predictor variables is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i$$

where

1. y_i is the response variable that we want to predict.
2. x_{i1}, x_{i2}, …, x_{ik} are k predictor variables.
3. β_0, β_1, …, β_k are unknown parameters.
4. β_0 is the intercept – the average value of Y when X_1, X_2 and X_k are all zeros.
5. β_j's are called the (partial) regression coefficients which represent the expected change in the response y_i per unit change in x_j when all the remaining regressor variables are held constant.
6. ε_i is the random error

Assume that $\varepsilon_i \sim NID(0, \sigma^2)$ and thus the mean of y_i is $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$.

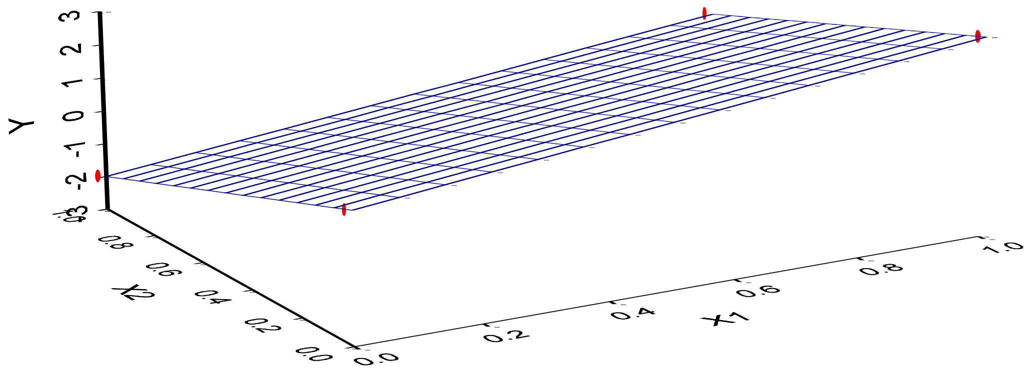
<u>Example 2.1:</u>

Suppose you want to relate a random variable Y to two independent variables X_1 and X_2. The multiple regression model is $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$.

With the mean of Y is given as $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$.

This equation is a three–dimensional extension of the line of means from Chapter 1 and trace a plane in three dimensional space (see Figure below).

Figure 2.1



Note:

(i) Any point on the response plane corresponds to the mean response $E(Y)$ at the given combination of levels of X_1 and X_2.

(ii) Correspond to the levels (x_{i1}, x_{i2}) of the two predictor variables, the vertical rule between y_i and the response plane represents the difference between y_i and the mean $E(y_i)$ of the probability distribution of Y for the given (x_{i1}, x_{i2}) combination. Hence, the vertical distance from y_i to the response plane represents the error term $\varepsilon_i = y_i - E(y_i)$.

To see this, suppose the mean $E(Y)$ of a response Y is related to two quantitative independent variables, X_1 and X_2, by the first–order model (i.e. linear in predictor variables)

$$E(Y) = 1 + 2X_1 + X_2$$

In other words, $\beta_0 = 1$, $\beta_1 = 2$ and $\beta_2 = 1$.

When $X_2 = 0$, $E(Y) = 1 + 2X_1$; $X_2 = 1$, $E(Y) = 2 + 2X_1$; $X_2 = 2$, $E(Y) = 3 + 2X_1$.

Figure 2.2:

Graphs of $E(Y) = 1 + 2X_1 + X_2$ for $X_2 = 0, 1, 2$.

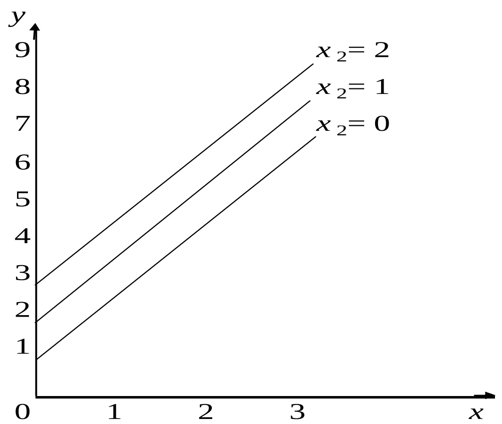


Figure 2.2 exhibits a characteristic of all first–order models: For any graph of $E(Y)$ versus any one variable – say X_1 – with fixed values of the other variables, the result will always be a straight line with slope equal to β_1. If you repeat the process for other values of the fixed independent variables, you will obtain a set of parallel straight lines. This indicates that the

effect of the independent variable X_j on $E(Y)$ is *independent* of all the other independent variables in the model, and this effect is measured by the slope β_j.

Examples of multiple linear regression models:

1. Polynomial models: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + ... + \beta_k x^k + \varepsilon$.
2. Models with interaction effects: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon$.
3. Second–order model with interaction:
   $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon$.

From these examples, note that the multiple linear regression models are linear in the parameters.

2.1.1 Matrix Form of Multiple Linear Regression Model

The multiple linear regression model can be written in matrix notation as $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$

$$\text{where } \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

In general, $\mathbf{y}$ is an $n \times 1$ vector of the observations, $\mathbf{X}$ is an $n \times p$ matrix of the levels of the regressor's variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector of the regression coefficients, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of random errors. Note that $p = k + 1$.

2.2 Estimation of the Model Parameter

2.2.1 Least Squares Estimation of the Regression Coefficients

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

$$\begin{aligned} S(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X\beta})'(\mathbf{y} - \mathbf{X\beta}) \\ &= \mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \end{aligned}$$

The least–squares estimators must satisfy

$$\left.\frac{\partial \mathbf{S}}{\partial \boldsymbol{\beta}}\right|_{\hat{\boldsymbol{\beta}}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{0}$$

$$\begin{aligned} &\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y} \\ &\therefore \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \end{aligned}$$

The fitted multiple linear regression model is

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is an $n \times n$ matrix and is called the **hat matrix**.

Note that

$$\mathbf{X'X} = \begin{bmatrix} n & \Sigma x_{i1} & \Sigma x_{i2} & \cdots & \Sigma x_{ik} \\ \Sigma x_{i1} & \Sigma x_{i1}^2 & \Sigma x_{i1}x_{i2} & \cdots & \Sigma x_{i1}x_{ik} \\ \Sigma x_{i2} & \Sigma x_{i1}x_{i2} & \Sigma x_{i2}^2 & \cdots & \Sigma x_{i2}x_{ik} \\ \vdots & \vdots & \vdots & & \vdots \\ \Sigma x_{ik} & \Sigma x_{i1}x_{ik} & \Sigma x_{i2}x_{ik} & \cdots & \Sigma x_{ik}^2 \end{bmatrix} \qquad \mathbf{X'y} = \begin{bmatrix} \Sigma y_i \\ \Sigma x_{i1}y_i \\ \Sigma x_{i2}y_i \\ \vdots \\ \Sigma x_{ik}y_i \end{bmatrix}$$

LSE Properties:

1. $E(\hat{\boldsymbol{\beta}}) = E[(\mathbf{X'X})^{-1}\mathbf{X'y}] = \boldsymbol{\beta}$.
2. $\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'y}$ is the BLUE of $\boldsymbol{\beta}$.
3. $\mathbf{cov}(\hat{\boldsymbol{\beta}}) = E\left\{\left[\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})\right]\left[\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})\right]'\right\} = Var(\hat{\boldsymbol{\beta}}) = Var[(\mathbf{X'X})^{-1}\mathbf{X'y}] = \sigma^2(\mathbf{X'X})^{-1}$

   which is a $p \times p$ symmetric matrix whose jth diagonal element is the variance of $\hat{\beta}_j$ and (ij)th off–diagonal element is the covariance between $\hat{\beta}_i$ and $\hat{\beta}_j$.

   Let $\mathbf{C} = (\mathbf{X'X})^{-1}$ then $Cov(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 C_{i-1,j-1}$.

Example 2.2:

In a small-scale experimental study of the relation between degree of brand liking (Y) and moisture content (X_1) and sweetness (X_2) of the product, the following results were obtained.

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
x_{i1}	4	4	4	4	6	6	6	6	8	8	8	8	10	10	10	10
x_{i2}	2	4	2	4	2	4	2	4	2	4	2	4	2	4	2	4
y_i	64	73	61	76	72	80	71	83	83	89	86	93	88	95	94	100

$$\mathbf{X} = \begin{bmatrix} 1 & 4 & 2 \\ 1 & 4 & 4 \\ \vdots & \vdots & \vdots \\ 1 & 10 & 4 \end{bmatrix}, \ \mathbf{y} = \begin{bmatrix} 64 \\ 73 \\ \vdots \\ 100 \end{bmatrix}, \ \mathbf{X'X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 4 & 4 & \cdots & 10 \\ 2 & 4 & \cdots & 4 \end{bmatrix} \begin{bmatrix} 1 & 4 & 2 \\ 1 & 4 & 4 \\ \vdots & \vdots & \vdots \\ 1 & 10 & 4 \end{bmatrix} = \begin{bmatrix} 16 & 112 & 48 \\ 112 & 864 & 336 \\ 48 & 336 & 160 \end{bmatrix}$$

$$\mathbf{X'y} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 4 & 4 & \cdots & 10 \\ 2 & 4 & \cdots & 4 \end{bmatrix} \begin{bmatrix} 64 \\ 73 \\ \vdots \\ 100 \end{bmatrix} = \begin{bmatrix} 1308 \\ 9510 \\ 3994 \end{bmatrix}$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'y}$$

$$= \begin{bmatrix} 16 & 112 & 48 \\ 112 & 864 & 336 \\ 48 & 336 & 160 \end{bmatrix}^{-1} \begin{bmatrix} 1308 \\ 9510 \\ 3994 \end{bmatrix} = \begin{bmatrix} 1.2375 & -0.0875 & -0.1875 \\ -0.0875 & 0.0125 & 0 \\ -0.1875 & 0 & 0.0625 \end{bmatrix} \begin{bmatrix} 1308 \\ 9510 \\ 3994 \end{bmatrix} = \begin{bmatrix} 37.65 \\ 4.425 \\ 4.375 \end{bmatrix}$$

The estimated regression equation is given by

When $x_1 = 1$ and $x_2 = 2$.

How would you interpret the values of $\hat{\beta}_1$ and $\hat{\beta}_2$?

Example 2.3

How do real estate agents decide on the asking price for a newly listed property? A computer database in a small community contains the listed selling price Y (in RM), the area of living space on the property X_1, the appraised home value on the property X_2, and appraised land value of the property X_3, for $n = 20$ randomly selected properties currently on the market. The data are shown below.

SalePrice, y_i	Area, x_{i1} (sf)	HomeValue, x_{i2} (RM)	LandValue, x_{i3} (RM)
68900	1873	44967	5960
48500	928	27860	9000
55500	1126	31439	9500
.	.	.	.
.	.	.	.
.	.	.	.
22400	962	5779	1500

The multiple regression model is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i \quad \text{or} \quad E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

The fitted regression equation is

$$\hat{Y} = 1470.28 + 13.53\,X_1 + 0.8204\,X_2 + 0.8145\,X_3 \quad \text{or}$$

predicted price = 1470.28 + 13.53∗*area* + 0.8204∗*HomeValue* + 0.8145***LandValue*

(a) How would you interpret the values of $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$?

$\hat{\beta}_1$= 13.53: The sale price of a property is expected to increase by RM13.53 for each additional in square foot of living area (X_1) when both appraised home value (X_2) and land value (X_3) are held constant.

$\hat{\beta}_2$= 0.8204: The sale price of a property is expected to increase by RM0.8204 for every RM1 increase in appraised home value (X_2) when both area (X_1) and appraised land value (X_3) are held constant.

$\hat{\beta}_3$=0.8145: The sale price of a property is expected to increase by RM0.8145 for every RM1 increase in appraised in land value (X_3) when both area (X_1) and appraised home value (X_2) are held constant.

(b) Predict the sale price for a property with X_1 = 1,800 square feet, X_2 = RM50,000, and X_3 = RM15,000.

2.2.2 Estimation of σ^2

As in simple linear regression, we may develop an estimator of σ^2 from the residual sum of squares

$$SS_E = \sum_{i=1}^{n} e_i^2 = \mathbf{e}'\mathbf{e} = \left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)'\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right) = \mathbf{y}'\mathbf{y}-\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}$$

This residual sum of squares has $n-k-1$ degrees of freedom associate with it since $k+1$ parameters are estimated in the regression model.

The residual mean square is

$$\hat{\sigma}^2 = MS_E = \frac{SS_E}{n-k-1}$$

Note: MS_E is an unbiased estimator.

Example 2.4

Estimate σ^2, variance of $\hat{\beta}_2$, and covariance between $\hat{\beta}_1$ and $\hat{\beta}_2$ in *Ex.2.2*.

2.3 Hypothesis Testing in Multiple Linear Regression

2.3.1 The Analysis of Variance (ANOVA) for Multiple Linear Regression

$SS_T = \sum\left(y_i - \bar{y}\right)^2 = \mathbf{y}'\mathbf{y}-\left(\dfrac{1}{n}\right)\mathbf{y}'\mathbf{J}\mathbf{y}$ is measure of "how good" $\bar{y}$ does.

where $\mathbf{J}$ is an $n\times n$ square matrix with all elements 1.

Computational formula : $SS_T = \mathbf{y}'\mathbf{y}-\dfrac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$.

$$\begin{aligned} SS_E &= \mathbf{e}'\mathbf{e} = \sum\left(y_i - \hat{y}_i\right)^2 \\ &= \left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)'\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right) \\ &= \mathbf{y}'\mathbf{y}-\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} \end{aligned}$$

is a measure of "how good" $\hat{y}$ does

$$\begin{aligned} SS_R &= SS_T - SS_E \\ &= \sum (y_i - \bar{y})^2 - \sum (y_i - \hat{y}_i)^2 \\ &= (\mathbf{y}'\mathbf{y} - \left(\frac{1}{n}\right)\mathbf{y}'\mathbf{J}\mathbf{y}) - (\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}) \\ &= \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \left(\frac{1}{n}\right)\mathbf{y}'\mathbf{J}\mathbf{y} \end{aligned}$$

is the amount "gained" by doing the regression.

Computational formula : $SS_R = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \dfrac{\left(\sum_{i=1}^n y_i\right)^2}{n}$

Note that:

$$\frac{SS_R}{\sigma^2} \sim \chi_k^2; \qquad \frac{SS_E}{\sigma^2} \sim \chi_{n-k-1}^2; \; F = \frac{SS_R / k}{SS_E / (n-k-1)} = \frac{MS_R}{MS_E} \sim F_{k,\,n-k-1}$$

The ANOVA Table

Source	DF	Sum of Squares	Mean Square	F^*
Regression	k	SS_R	MS_R	$F = \dfrac{MS_R}{MS_E}$
Error	$n-k-1$	SS_E	MS_E	
Corrected Total	$n-1$	SS_T		

Notes:

1. SS_R has k degrees of freedom. (Note: When there is only 1 independent variable, the degree of freedom is 1).
2. SS_E has $n-k-1$ degrees of freedom.

2.3.2 Testing for Significance of Regression (Global Test of Model Adequacy)

Is the regression equation that uses information provided by the predictor variables $X_1, X_2, ..., X_k$ substantially better than the simple predictor $\bar{y}$ that does not rely on any of the $X-$values? Is there a linear relationship between the response Y and any of the regressor variables? These questions can be answered using an overall $F-$test.

The null and alternative hypothesis are

$$H_0 : \beta_0 = \beta_1 = ... = \beta_k = 0$$
$$H_1 : \beta_j \neq 0 \quad \textit{for at least one } j.$$

Test statistics:

$$F = \frac{MS_R}{MS_E} \sim F_{k,\,n-k-1}; \qquad \text{Reject } H_0 \text{ when } F_0 > F_{\alpha;\,k,\,n-k-1}$$

Reject H_0 implies that at least one of the regressors contributes significantly to the model.

Example 2.5:
Refer to *Ex. 2.2*, determine if the model found is significant. Use $\alpha = 0.01$ in the hypothesis test.

2.3.3 Testing the Significance of the Partial Regression Coefficients (Partial / Marginal Tests)

Once you have determined that the model is useful for predicting Y, you should explore the nature of the "usefulness" in more detail. Do all the predictor variables add important information for prediction in **the presence of other predictors already in the model**?
The null and alternative hypothesis are

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

Test statistics:

$$t = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} = \frac{\hat{\beta}_j}{se\left(\hat{\beta}_j\right)} \sim t_{n-k-1};$$

C_{jj} is the $(j+1)th$ diagonal element of the $(\mathbf{X}'\mathbf{X})^{-1}$ matrix.
If H_0 is not rejected, this indicates that the regressor x_j does not contribute significantly to the model, in another word, x_j can be deleted from the model.

Example 2.6: (Refer to *Ex.2.2*)
Is moisture content (X_1) significantly related to the degree of brand liking in the model, given that sweetness (X_2) is already in the model? Use $\alpha = 0.01$ for the hypothesis test.

<u>Recommendation for checking a model's variables usefulness in estimating Y: 2 steps</u>

(1) Do a $F-$test for regression relation
   (a) If H_0 is not rejected : Stop.
   Model variables may not be useful in estimating Y.
   (b) If H_0 is rejected: There is at least one variable that is useful in estimating Y. Go to step 2.

(2) Do $t-$ tests on individual β_j s to determine <u>which</u> variables are useful in estimating Y
   (a) If H_0 is rejected for a β_j: The corresponding regressor X_j is useful in estimating Y.
   (b) If H_0 is not rejected for a β_j: The corresponding regressor X_j may not be useful in estimating Y.

2.3.4 <u>Testing Sets of Regression Coefficient (Partial $F-$test or "Extra Sum of Squares")</u>

Suppose a company suspects that the demand Y for a product could be related to as many as five predictor variables, X_1, X_2, X_3, X_4, X_5. The cost of obtaining measurements on the variables X_3, X_4 and X_5 is very high. If, in a small pilot study, the company could show that these three variables contribute little or no information for predicting Y, they can be removed for the study at great savings to the company.

If X_3, X_4 and X_5 contribute little or no information for predicting Y, then you want to test

$$H_0 : \beta_3 = \beta_4 = \beta_5 = 0$$

$$H_1 : \textit{At least one of } \beta_3, \beta_4 \textit{ or } \beta_5 \textit{ differs from } 0.$$

This is a test to investigate the contribution of a subset of the regressor variables to the model.

Let the vector of regression coefficients be partitioned into 2 groups:

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \end{bmatrix} \text{ where } \boldsymbol{\beta_1} = (\beta_0, \beta_1, ..., \beta_{r-1}) \text{ and } \boldsymbol{\beta_2} = (\beta_r, \beta_{r+1}, ..., \beta_k).$$

A test of hypothesis concerning a set of parameters involves two models:

<u>Full Model</u>

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{X_1}\boldsymbol{\beta_1} + \mathbf{X_2}\boldsymbol{\beta_2} + \boldsymbol{\varepsilon}$$

$$E(Y) = \beta_0 + \beta_1 X_1 + ... + \beta_r X_r + \beta_{r+1} X_{r+1} ... + \beta_k X_k$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'y}\text{ ; } SS_R(\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}'\mathbf{X'y}\text{ , } k \text{ degrees of freedom; } MS_E = \frac{\mathbf{y'y} - \hat{\boldsymbol{\beta}}'\mathbf{X'y}}{n-k-1}.$$

where $\mathbf{X_1}$ and $\mathbf{X_2}$ are orthogonal.

<u>Hypothesis Test for sets of Parameters</u>

$$H_0 : \beta_r = \beta_{r+1} = ... = \beta_k = 0$$

$$H_1 : \textit{At least one of } \beta_r, \beta_{r+1}, ... \textit{ or } \beta_k \textit{ differs from } 0.$$

To find the contribution of the $\boldsymbol{\beta_2}$ to the regression, we assume that H_0 is true.

The Reduced Model

$$\mathbf{y} = \mathbf{X_1\beta_1} + \boldsymbol{\varepsilon}$$

$$E(Y) = \beta_0 + \beta_1 X_1 + \dots + \beta_{r-1} X_{r-1}$$

$$\hat{\boldsymbol{\beta}}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}\ ;\ SS_R(\boldsymbol{\beta}_1) = \hat{\boldsymbol{\beta}}_1'\mathbf{X}_1'\mathbf{y}\ ,\ r-1 \text{ degrees of freedom.}$$

The regression sum of squares due to $\boldsymbol{\beta}_2$ given that $\boldsymbol{\beta}_1$ is already in the model is

$$SS_R(\boldsymbol{\beta}_2 | \boldsymbol{\beta}_1) = SS_R(\boldsymbol{\beta} - SS_R(\boldsymbol{\beta}_1),\ (k-r+1) \text{ degrees of freedom.}$$

This is called "extra sum of squares" because it measures the increase in the regression sum of squares that results from adding the regressors $X_r, \dots, X_k$ to a model that already contains $X_1, \dots, X_{r-1}$.

Test statistics:

$$F = \frac{SS_R(\boldsymbol{\beta}_2 | \boldsymbol{\beta}_1)/(k-r+1)}{MS_E} \sim F_{k-r+1,\, n-k-1},$$ Reject H_0 when

$$F_0 > F_{k-r+1,\, n-k-1}$$

Remark:
The Partial $F-$test on a single variable X_j is equivalent to the $t-$ test.

Example 2.7:
Repeat *Ex. 2.6* with partial $F-$test.

2.4 Confidence Intervals in Multiple Linear Regression

2.4.1 Confidence Intervals on Regression Coefficients

$(1-\alpha)100\%$ confidence interval (C.I.) for β_j is

$$\hat{\beta}_j \pm t_{\alpha/2;\, n-k-1} se(\hat{\beta}_j) \qquad j = 0, 1, \dots, k$$

where
$se(\hat{\beta}_j) = \sqrt{MS_E * C_{jj}}$, C_{jj} is the $(j+1)th$ diagonal element of the $(\mathbf{X}'\mathbf{X})^{-1}$ matrix.

Example 2.8
Refer to *Ex. 2.2*, construct a 95% confidence interval for β_1.

2.4.2 Using the Model for Estimation and Prediction

Recall that the least squares line yielded the same value for both the estimate for $E(Y_h)$ and the prediction of some future value of y_h. The confidence interval for the mean $E(Y_h)$ is narrower than the prediction interval for y_h because of the additional uncertainty attributable to the random error ε when predicting some future value of y_h.

These same concepts carry over to multiple regression models.

The fitted value at point $\mathbf{x_h} = (1, x_{h1}, x_{h2}, ..., x_{hk})'$ is $\hat{y}_h = \mathbf{x_h'}\hat{\boldsymbol{\beta}}$.

$$E(\hat{y}_h) = \mathbf{x_h'}\boldsymbol{\beta} = y_h \qquad Var(\hat{y}_h) = \sigma^2\,\mathbf{x_h'}(\mathbf{X'X})^{-1}\mathbf{x_h}$$

Confidence Interval for $E(Y_h)$ and Prediction Interval for y_h:

$(1-\alpha)100\%$ CI on the mean response at the point $\mathbf{x_h}$, $E(Y_h)$ is

$$\hat{E}(Y_h) - t_{\alpha/2;n-k-1}se[\hat{E}(Y_h)] \le E(Y_h) \le \hat{E}(Y_h) + t_{\alpha/2;n-k-1}se[\hat{E}(Y_h)]$$

where $\hat{E}(Y_h) = \mathbf{x_h'}\hat{\boldsymbol{\beta}}$ and $se[\hat{E}(Y_h)] = \sqrt{MS_E * \mathbf{x_h'}(\mathbf{X'X})^{-1}\mathbf{x_h}}$.

$(1-\alpha)100\%$ prediction interval (PI) on the future observation y_h at a specified value of $\mathbf{x_h}$ is

$$\hat{y}_h - t_{\alpha/2;n-k-1}se(\hat{y}_h) \le y_h \le \hat{y}_h + t_{\alpha/2;n-k-1}se(\hat{y}_h)$$

where $se(\hat{y}_h) = \sqrt{MS_E(1 + \mathbf{x_h'}(\mathbf{X'X})^{-1}\mathbf{x_h})}$.

Note:

$$se(\hat{y}_h) = \sqrt{MS_E(1 + \mathbf{x_h'}(\mathbf{X'X})^{-1}\mathbf{x_h})} = \sqrt{MS_E + MS_E\,\mathbf{x_h'}(\mathbf{X'X})^{-1}\mathbf{x_h}} = \sqrt{MS_E + (se[\hat{E}(Y_h)])^2}$$

Example 2.9:
Reconsider *Ex. 2.2*, predict the degree of brand liking (Y) with moisture content (X_1) and sweetness (X_2) of the product are both 5.

(a) Find the corresponding 95% confidence interval for the mean response $E(Y_h)$.

(b) Find the corresponding 95% prediction interval for y_h.

Solution for *Ex. 2.9*

Example 2.10:

When performing a regression of Y on X_1 and X_2, we find that

(i) $\hat{y}_i = 20 - 1.5x_{i1} + 1.8x_{i2}$

(ii)

Source	DF	SS	MS
Regression	2	42	21
Error	3	12	4

(iii) $(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 4/3 & -1/4 & -1/3 \\ -1/4 & 1/16 & 0 \\ -1/3 & 0 & 2/3 \end{bmatrix}$

(a) Find $se(\hat{\beta}_2)$.

(b) Calculate the value of the test statistic for testing $H_0 : \beta_2 = 1$.

(c) Suppose $x_{h1} = 2$, and $x_{h2} = 3$, find $se[\hat{E}(Y_h)]$ and $se(\hat{y}_h)$.

2.5 Coefficient of determination R^2

Recall that

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

R^2 has the same interpretation as before, but with respect to k independent variables. (i.e. $R^2 100\%$ of the variation in Y can be explained by using the independent variables to predict Y)

Notes:

(1) Use R^2 as a measure of fit when the sample size is substantially larger than the number of variables in the model; otherwise, R^2 may be artificially high.

(2) As more variables are added to the model, R^2 will always increase even if the additional variables do a poor job of estimating Y (i.e. SS_E can never become larger with more predictor variables and SS_T is always the same for a given set of responses).

The Adjusted R^2, R^2_{Adj}

$$R^2_{Adj} = 1 - \frac{SS_E/(n-k-1)}{SS_T/(n-1)} = 1 - \frac{n-1}{n-k-1}(1-R^2)$$

Note:

1. Since $SS_E/(n-k-1)$ is the residual mean square and $SS_T/(n-1)$ is constant regardless of how many variables are in the model, the R^2_{Adj} will only increase on adding a variable to the model if the addition of the variable reduces the residual mean square.
2. The interpretation of R^2_{Adj} is about the same as R^2.
3. $R^2_{Adj} \le R^2$.
4. R^2_{Adj} can be less than 0.

Relationship between F and R^2

$$F = \frac{R^2/k}{(1-R^2)/(n-k-1)} = \frac{(n-k-1)R^2}{(1-R^2)k}$$

Example 2.11:

Examine what happens to R^2 and R^2_{Adj} when additional variables are added to the model

Consider the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

The data for Y, X_1, X_2, X_3, X_4 for $n = 15$ observations were analyzed. The following results were obtained.

Model	R^2	R^2_{Adj}
$\hat{Y} = 35.5 + 7.5X_1$	0.9052	0.8979
$\hat{Y} = 18.62 + 5.75X_1 + 19.49X_2$	0.9510	0.9429
$\hat{Y} = 15.81 + 6.04X_1 + 128.22X_2 - 14.66X_3$	0.9704	0.9623
$\hat{Y} = 18.76 + 6.27X_1 + 30.3X_2 - 16.2X_3 - 2.67X_4$	0.9714	0.9599

(a) Note that X_1, X_2 and X_3 were significantly related to Y in the model, but X_4 was not.
When a variable is added to the model that "may not" be useful, the R^2_{Adj} decreased. Thus, the decrease in R^2_{Adj} after X_4 is added to the model suggests that X_4 may not be useful in estimating Y.

(b) Notice that R^2 increased after each variable was added to the model.

Note:
R^2_{Adj} is mainly used to compare two or more models that use different numbers of predictor variables.

Example 2.12:
R^2 and R^2_{Adj} were calculated for all possible subsets of three independent variables. The results are as follow:

Subsets of Regression: Y versus X_1, X_2, X_3

Independent Variable	R^2	R^2_{Adj}
X_1	0.9052	0.8979
X_2	0.6948	0.6713
X_3	0.5565	0.5223
X_1, X_2	0.9510	0.9429
X_1, X_3	0.9150	0.9008
X_2, X_3	0.7565	0.7159
X_1, X_2, X_3	0.9519	0.9388

If you had to compare these models and choose the best one, which model would you choose? Explain.

2.6 Polynomial in one Variable

Polynomial regression models may contain one, two or more than two predictor variables, and each predictor variable may be present in various powers.

If you suspect that one independent variable X affects the response Y, but that relationship is curvilinear rather than linear, then you might choose to fit a polynomial model with degree k.

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_k X^k + \varepsilon_i$$

The term involving x_i^2, called a quadratic term (or second-order term).When the curve opens upward, the sign of β_2 is positive (see figure 2.3a); when the curve opens downward, the sign of β_2 is negative (See figure 2.3b). This polynomial model is a *second-order model with one predictor variable*.

The response function for regression model is

$$E(Y) = \beta_0 + \beta_1 X + \beta_2 X^2$$

Figure 2.3a: Graphs for Quadratic Models when $\beta_2 > 0$ (Concave up)

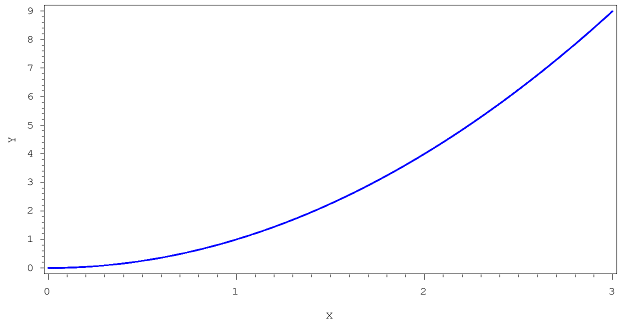
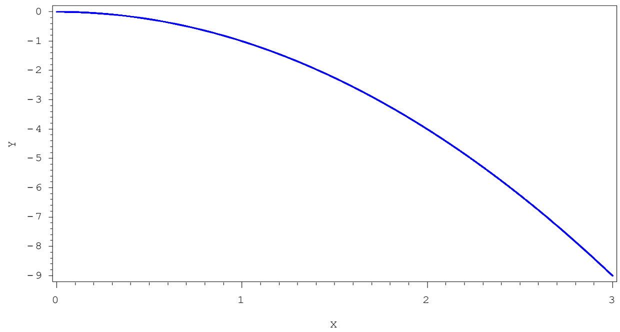


Figure 2.3b: Graphs for Quadratic Models when $\beta_2 < 0$ (Concave down)



Note:

1) Note that the predictor variable is centered - i.e. $x_i = X_i - \bar{X}$. The reason is that X and X^2 will often be highly correlated leading to multicollinearity and problems with inverting $\mathbf{X'X}$ (i.e. $(\mathbf{X'X})^{-1}$).

2) After a polynomial regression model has been developed, we could express the final model in terms of the original variables X rather than keeping it in terms of the centered variables:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_{11} x_i^2 \qquad \text{where} \quad x_i = X_i - \bar{X}$$

becomes in terms of the original variable X:

$$\hat{y}_i = \hat{\beta}'_0 + \hat{\beta}'_1 X + \hat{\beta}'_{11} X^2$$

where:

$$\hat{\beta}'_0 = \hat{\beta}_0 - \hat{\beta}_1 \bar{X} + \hat{\beta}_{11} \bar{X}^2$$

$$\hat{\beta}'_1 = \hat{\beta}_1 - 2\hat{\beta}_{11} \bar{X}$$

$$\hat{\beta}'_2 = \hat{\beta}_{11}$$

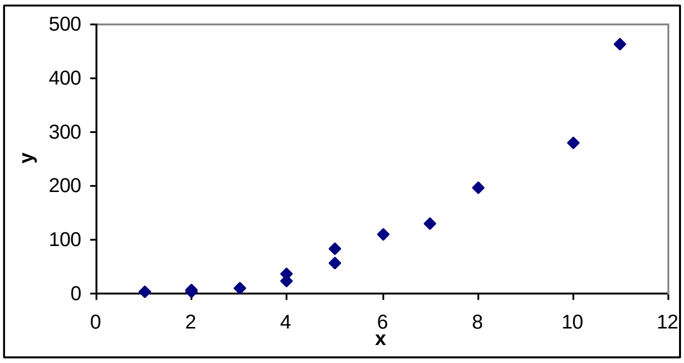
3) The fitted values and residuals for the regression function in terms of X_i are exactly the same as for the regression function in terms of the centered values x_i. However, the estimated standard deviations of the regression coefficients in terms of the centered variables x_i do not agree to the regression coefficients in terms of the original variables X_i.

Example 2.13

The following table shows the sales, y (in RM million) for 13 manufacturers with the number of manufacturer's representatives, x associated with each firm.

y	2.1	3.6	6.2	10.4	22.8	35.6	57.1	83.5	109.4	128.6	196.8	280.0	462.3
x	2	1	2	3	4	4	5	5	6	7	8	10	11

The corresponding scatter diagram is shown below.



If simple linear model is proposed, the results obtained are:

(i) $\hat{y}_i = -107.029 + 41.026x$

(ii)

Source	df	SS
Regression	1	192396.416
Error	11	28721.452

(iii) $R^2 = 0.87$ and $R^2_{Adj} = 0.858$

If the quadratic model is proposed, the results obtained are:

(i) $\hat{y}_i = 18.067 - 15.7232\,x + 4.75x^2$

(ii)

Source	df	SS
Regression	1	192396.416
Error	11	28721.452

(iii) $R^2 = 0.973$ and $R^2_{Adj} = 0.967$

Refer to the quadratic model obtained.

(a) Explain why the value $\hat{\beta}_0 = 18.067$ has no practical interpretation.

Since the range of x does not include 0 (a manufacturer with 0 representative), thus the interpretation of $\hat{\beta}_0$ is not meaningful.

(b) Explain why the value $\hat{\beta}_1 = -15.723$ should not be interpret as a slope.

$\hat{\beta}_1 = -15.723$ is no longer representing a slope in the presence of the quadratic term x^2. The estimated coefficient of the first-order x will not have a meaningful interpretation in the quadratic model.

(c) Examine the value of $\hat{\beta}_2$ to determine the nature of the curvature (concave upward or downward) in the sample data.

$\hat{\beta}_2 = 4.75$. The positive sign of $\hat{\beta}_2$ indicates that the curve is concave upward.

(d) Test whether or not there is a regression relation; use $\alpha = 0.01$.

(e) Which models is preferred?

(f) Refer to the quadratic model, interpret the value of R^2

(g) Predict the sales for a manufacturer with 5 representatives.

2.7 Inferences Concerning Linear Functions of the Model Parameters

Further discussion on making inferences about a single β_j or linear combinations of the model parameters $\beta_0, \beta_1, ..., \beta_k$.

Consider the linear function $a_0\beta_0 + a_1\beta_1 + ... + a_k\beta_k$, where $a_0, a_1, ..., a_k$ are constants (some of which may equal zero).

Define the $1\times(k+1)$ matrix: $\mathbf{a}' = [a_0 \quad a_1 \quad ... \quad a_k]$

The linear function can be written in matrix form as $\mathbf{a}'\boldsymbol{\beta} = a_0\beta_0 + a_1\beta_1 + ... + a_k\beta_k$.

Let the estimator $\mathbf{a}'\hat{\boldsymbol{\beta}} = a_0\hat{\beta}_0 + a_1\hat{\beta}_1 + ... + a_k\hat{\beta}_k$.

Properties:

1. $E(\mathbf{a}'\hat{\boldsymbol{\beta}}) = \mathbf{a}'E[\hat{\boldsymbol{\beta}}] = \mathbf{a}'\boldsymbol{\beta}$.
2. $V(\mathbf{a}'\hat{\boldsymbol{\beta}}) = \mathbf{a}'V(\hat{\boldsymbol{\beta}})\mathbf{a} = \mathbf{a}'\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a} = \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}\,\sigma^2$
3. $\mathbf{a}'\hat{\boldsymbol{\beta}} \sim N(\mathbf{a}'\boldsymbol{\beta}, \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}\,\sigma^2)$

2.7.1 Testing for Linear Function of the Model Parameters

The null and alternative hypothesis are

$$H_0: \mathbf{a}'\boldsymbol{\beta} = (\mathbf{a}'\boldsymbol{\beta})_0$$
$$H_1: \mathbf{a}'\boldsymbol{\beta} \neq (\mathbf{a}'\boldsymbol{\beta})_0$$

Test statistics:

$$t = \frac{\mathbf{a}'\hat{\boldsymbol{\beta}} - (\mathbf{a}'\boldsymbol{\beta})_0}{\sqrt{\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}MS_E}} \sim t_{n-k-1}$$

2.7.2 Confidence Interval on Linear Function of the Model Parameters

A $(1-\alpha)100\%$ confidence interval (C.I.) for $\mathbf{a}'\boldsymbol{\beta}$ is

$$\mathbf{a}'\hat{\boldsymbol{\beta}} \pm t_{\alpha/2;n-k-1}\sqrt{\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}MS_E}$$

Special Cases

1. If vector $\mathbf{a}$ is defined by $a_j = \begin{cases} 1 & j = i \\ 0 & j \neq i \end{cases}$, then $\mathbf{a}'\boldsymbol{\beta} = \beta_i$.
2. If $\mathbf{a}' = [1 \quad x_1^* \quad ... \quad x_k^*]$, then $\mathbf{a}'\boldsymbol{\beta} = \beta_0 + \beta_1 x_1^* + ... + \beta_k x_k^*$ which is the mean response with $\mathbf{x}_h' = [x_1^* \quad ... \quad x_k^*]$

Example 2.14

Refer to *Ex.2.2*, find a 95% confidence interval for $\beta_1 + 2\beta_2$.