Foundation of Data Science and Analytics

Multiple Linear Regression

Arun K. Timalsina

Material Adaptation : Applied Statistics and Probability for Engineers, by Montgomery and Runger. John Wiley & Sons,

12-1: Multiple Linear Regression Models

12-1.1 Introduction

- Many applications of regression analysis involve situations in which there are more than one regressor variable.
- A regression model that contains more than one regressor variable is called a **multiple regression model**.

12-1: Multiple Linear Regression Models

12-1.1 Introduction

- For example, suppose that the effective life of a cutting tool depends on the cutting speed and the tool angle. A possible multiple regression model could be

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

where
Y – tool life
x_1 – cutting speed
x_2 – tool angle

12-1: Multiple Linear Regression Models

12-1.1 Introduction

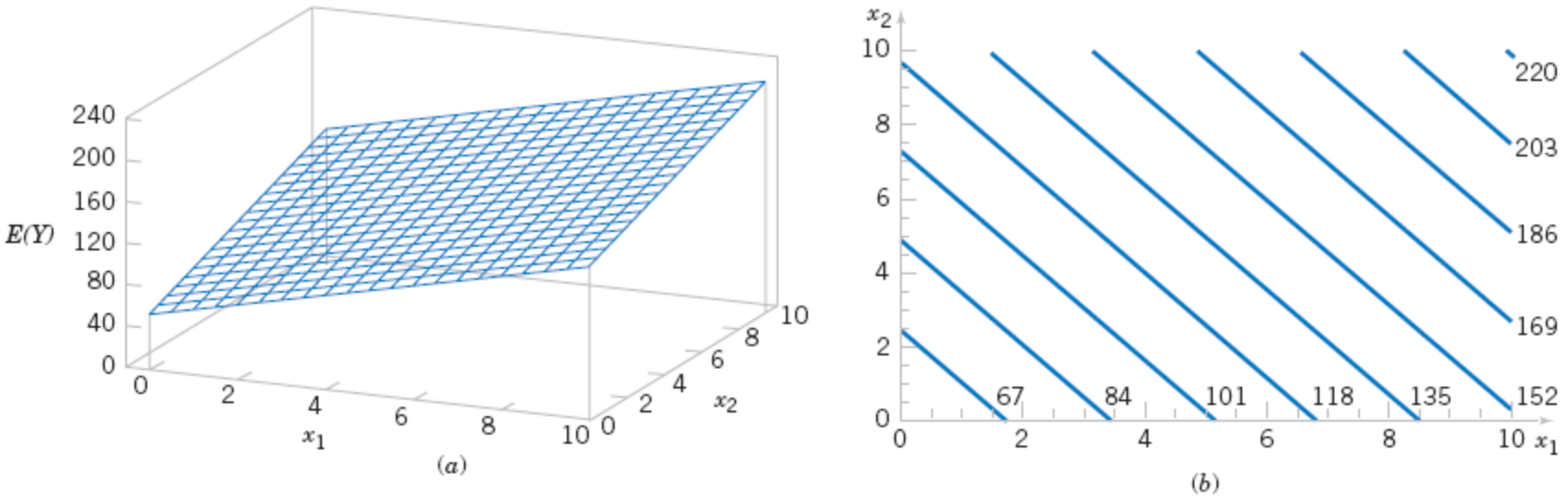


Figure 12-1 (a) The regression plane for the model $E(Y) = 50 + 10x_1 + 7x_2$. (b) The contour plot.

Figure 12-1 (a) The regression plane for the model $E(Y) = 50 + 10x_1 + 7x_2$. (b) The contour plot

12-1: Multiple Linear Regression Models

12-1.1 Introduction

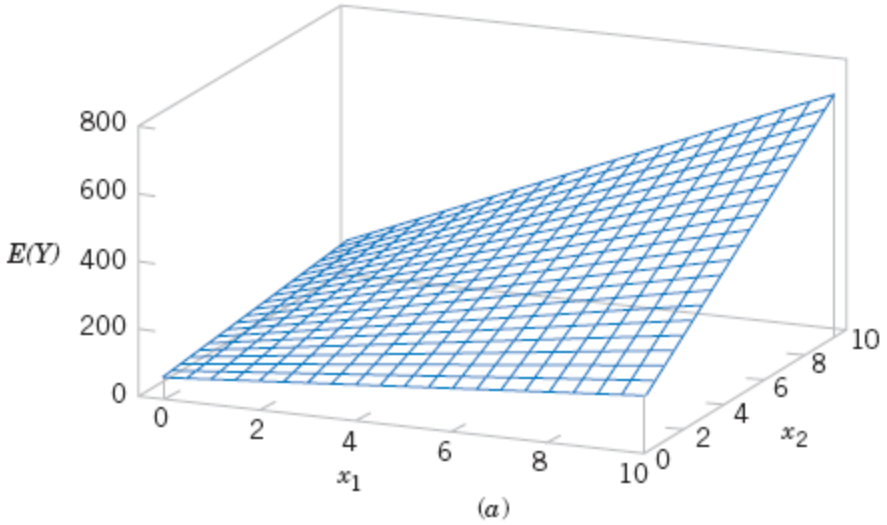
In general, the **dependent variable** or **response** Y may be related to k **independent** or **regressor variables.** The model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon \qquad (12\text{-}2)$$

is called a multiple linear regression model with k regressor variables. The parameters β_j, $j = 0, 1, \ldots, k$, are called the regression coefficients. This model describes a hyperplane in the k-dimensional space of the regressor variables $\{x_j\}$. The parameter β_j represents the expected change in response Y per unit change in x_j when all the remaining regressors x_i $(i \neq j)$ are held constant.

12-1: Multiple Linear Regression Models

12-1.1 Introduction



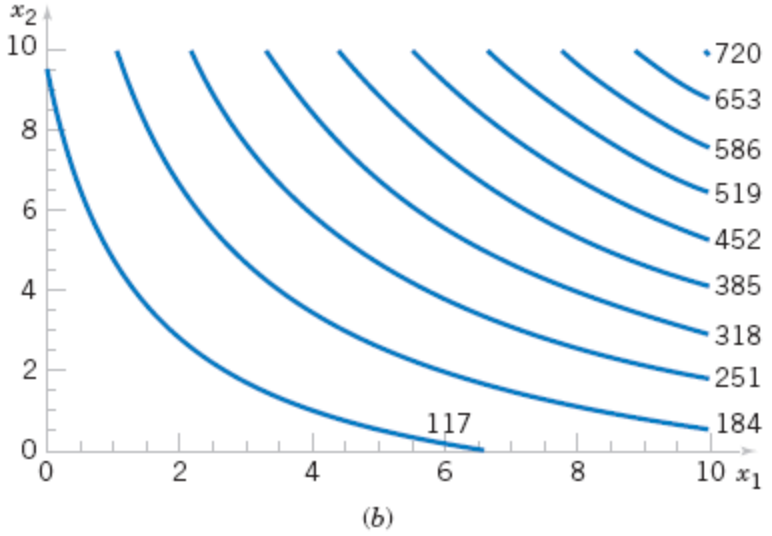


Figure 12-2 (a) Three-dimensional plot of the regression model $E(Y) = 50 + 10x_1 + 7x_2 + 5x_1x_2$. (b) The contour plot

12-1: Multiple Linear Regression Models

12-1.1 Introduction

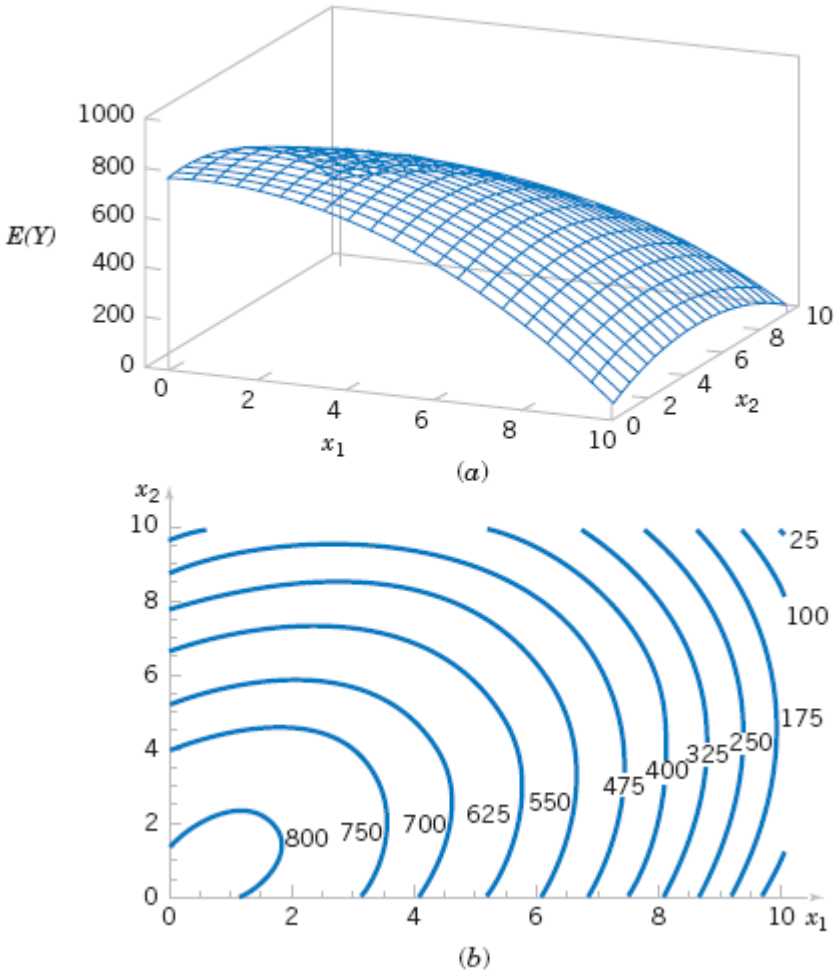


Figure 12-3 (a) Three-dimensional plot of the regression model $E(Y) = 800 + 10x_1 + 7x_2 - 8.5x_1^2 - 5x_2^2 + 4x_1x_2$. (b) The contour plot

12-1: Multiple Linear Regression Models

12-1.2 Least Squares Estimation of the Parameters

The **method of least squares** may be used to estimate the regression coefficients in the multiple regression model, Equation 12-2. Suppose that $n > k$ observations are available, and let x_{ij} denote the ith observation or level of variable x_j. The observations are

$$(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i), \qquad i = 1, 2, \ldots, n \quad \text{and} \quad n > k$$

It is customary to present the data for multiple regression in a table such as Table 12-1.

Table 12-1 Data for Multiple Linear Regression

y	x_1	x_2	$\ldots$	x_k
y_1	x_{11}	x_{12}	$\ldots$	x_{1k}
y_2	x_{21}	x_{22}	$\ldots$	x_{2k}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
y_n	x_{n1}	x_{n2}	$\ldots$	x_{nk}

12-1: Multiple Linear Regression Models

12-1.2 Least Squares Estimation of the Parameters

- The least squares function is given by

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2$$

- The least squares estimates must satisfy

$$\left. \frac{\partial L}{\partial \beta_0} \right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) = 0$$

and

$$\left. \frac{\partial L}{\partial \beta_j} \right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) x_{ij} = 0 \quad j = 1, 2, \ldots, k$$

12-1: Multiple Linear Regression Models

12-1.2 Least Squares Estimation of the Parameters

- The **least squares normal Equations** are

$$
\begin{aligned}
n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik} &= \sum_{i=1}^{n} y_i \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1}x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{i1}x_{ik} &= \sum_{i=1}^{n} x_{i1}y_i \\
\vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots & \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \hat{\beta}_1 \sum_{i=1}^{n} x_{ik}x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{ik}x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik}^2 &= \sum_{i=1}^{n} x_{ik}y_i
\end{aligned}
$$

- The solution to the normal Equations are the **least squares estimators** of the regression coefficients.

12-1: Multiple Linear Regression Models

Example 12-1

EXAMPLE 12-1 Wire Bond Strength

In Chapter 1, we used data on pull strength of a wire bond in a semiconductor manufacturing process, wire length, and die height to illustrate building an empirical model. We will use the same data, repeated for convenience in Table 12-2, and show the details of estimating the model parameters. A three-dimensional scatter plot of the data is presented in Fig. 1-15. Figure 12-4 shows a matrix of two-dimensional scatter plots of the data. These displays can be helpful in visualizing the relationships among variables in a multivariable data set. For example, the plot indicates that there is a strong linear relationship between strength and wire length.

12-1: Multiple Linear Regression Models

Example 12-1

Table 12-2 Wire Bond Data for Example 12-1

Observation Number	Pull Strength y	Wire Length x_1	Die Height x_2	Observation Number	Pull Strength y	Wire Length x_1	Die Height x_2
1	9.95	2	50	14	11.66	2	360
2	24.45	8	110	15	21.65	4	205
3	31.75	11	120	16	17.89	4	400
4	35.00	10	550	17	69.00	20	600
5	25.02	8	295	18	10.30	1	585
6	16.86	4	200	19	34.93	10	540
7	14.38	2	375	20	46.59	15	250
8	9.60	2	52	21	44.88	15	290
9	24.35	9	100	22	54.12	16	510
10	27.50	8	300	23	56.63	17	590
11	17.08	4	412	24	22.13	6	100
12	37.00	11	400	25	21.15	5	400
13	41.95	12	500				

12-1: Multiple Linear Regression Models

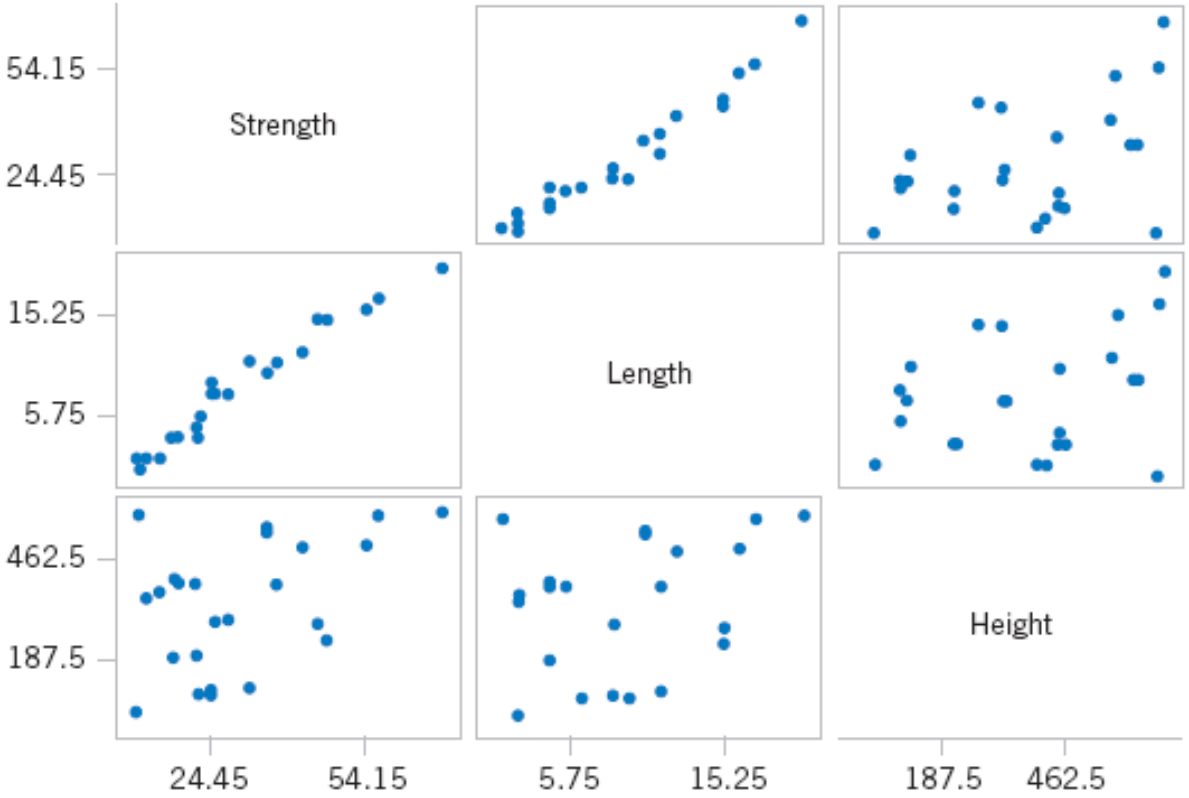


Figure 12-4 Matrix of scatter plots (from Minitab) for the wire bond pull strength data in Table 12-2.

Figure 12-4 Matrix of scatter plots (from Minitab) for the wire bond pull strength data in Table 12-2.

12-1: Multiple Linear Regression Models

Example 12-1

Specifically, we will fit the multiple linear regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

where Y = pull strength, x_1 = wire length, and x_2 = die height. From the data in Table 12-2 we calculate

$$n = 25, \sum_{i=1}^{25} y_i = 725.82$$

$$\sum_{i=1}^{25} x_{i1} = 206, \sum_{i=1}^{25} x_{i2} = 8{,}294$$

$$\sum_{i=1}^{25} x_{i1}^2 = 2{,}396, \sum_{i=1}^{25} x_{i2}^2 = 3{,}531{,}848$$

$$\sum_{i=1}^{25} x_{i1} x_{i2} = 77{,}177, \sum_{i=1}^{25} x_{i1} y_i = 8{,}008.47,$$

$$\sum_{i=1}^{25} x_{i2} y_i = 274{,}816.71$$

12-1: Multiple Linear Regression Models

Example 12-1

For the model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, the normal equations 12-10 are

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1}x_{i2} = \sum_{i=1}^{n} x_{i1}y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{i2} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}x_{i2} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2}^2 = \sum_{i=1}^{n} x_{i2}y_i$$

Inserting the computed summations into the normal equations, we obtain

$$25\hat{\beta}_0 + 206\hat{\beta}_1 + 8294\hat{\beta}_2 = 725.82$$

$$206\hat{\beta}_0 + 2396\hat{\beta}_1 + 77{,}177\hat{\beta}_2 = 8{,}008.47$$

$$8294\hat{\beta}_0 + 77{,}177\hat{\beta}_1 + 3{,}531{,}848\hat{\beta}_2 = 274{,}816.71$$

12-1: Multiple Linear Regression Models

Example 12-1

The solution to this set of equations is

$$\hat{\beta}_0 = 2.26379, \quad \hat{\beta}_1 = 2.74427, \quad \hat{\beta}_2 = 0.01253$$

Therefore, the fitted regression equation is

$$\hat{y} = 2.26379 + 2.74427x_1 + 0.01253x_2$$

Practical Interpretation: This equation can be used to predict pull strength for pairs of values of the regressor variables wire length (x_1) and die height (x_2). This is essentially the same regression model given in Section 1-3. Figure 1-16 shows a three-dimensional plot of the plane of predicted values $\hat{y}$ generated from this equation.

12-1: Multiple Linear Regression Models

12-1.3 Matrix Approach to Multiple Linear Regression

Suppose the model relating the regressors to the response is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i \qquad i = 1, 2, \ldots, n$$

In matrix notation this model can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

12-1: Multiple Linear Regression Models

12-1.3 Matrix Approach to Multiple Linear Regression

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad \text{and} \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

12-1: Multiple Linear Regression Models

12-1.3 Matrix Approach to Multiple Linear Regression

We wish to find the vector of least squares estimators that minimizes:

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \boldsymbol{\epsilon}'\boldsymbol{\epsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

The resulting least squares estimate is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \qquad (12\text{-}13)$$

12-1: Multiple Linear Regression Models

12-1.3 Matrix Approach to Multiple Linear Regression

The fitted regression model is

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \qquad i = 1, 2, \ldots, n \tag{12-14}$$

In matrix notation, the fitted model is

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

The difference between the observation y_i and the fitted value $\hat{y}_i$ is a **residual,** say, $e_i = y_i - \hat{y}_i$. The $(n \times 1)$ vector of residuals is denoted by

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} \tag{12-15}$$

12-1: Multiple Linear Regression Models

Example 12-2

In Example 12-1, we illustrated fitting the multiple regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

where y is the observed pull strength for a wire bond, x_1 is the wire length, and x_2 is the die height. The 25 observations are in Table 12-2. We will now use the matrix approach to fit the regression model above to these data. The model matrix **X** and **y** vector for this model are

Example 12-2

$$\mathbf{X} = \begin{bmatrix} 1 & 2 & 50 \\ 1 & 8 & 110 \\ 1 & 11 & 120 \\ 1 & 10 & 550 \\ 1 & 8 & 295 \\ 1 & 4 & 200 \\ 1 & 2 & 375 \\ 1 & 2 & 52 \\ 1 & 9 & 100 \\ 1 & 8 & 300 \\ 1 & 4 & 412 \\ 1 & 11 & 400 \\ 1 & 12 & 500 \\ 1 & 2 & 360 \\ 1 & 4 & 205 \\ 1 & 4 & 400 \\ 1 & 20 & 600 \\ 1 & 1 & 585 \\ 1 & 10 & 540 \\ 1 & 15 & 250 \\ 1 & 15 & 290 \\ 1 & 16 & 510 \\ 1 & 17 & 590 \\ 1 & 6 & 100 \\ 1 & 5 & 400 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} 9.95 \\ 24.45 \\ 31.75 \\ 35.00 \\ 25.02 \\ 16.86 \\ 14.38 \\ 9.60 \\ 24.35 \\ 27.50 \\ 17.08 \\ 37.00 \\ 41.95 \\ 11.66 \\ 21.65 \\ 17.89 \\ 69.00 \\ 10.30 \\ 34.93 \\ 46.59 \\ 44.88 \\ 54.12 \\ 56.63 \\ 22.13 \\ 21.15 \end{bmatrix}$$

12-1: Multiple Linear Regression Models

Example 12-2

The $\mathbf{X}'\mathbf{X}$ matrix is

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 2 & 8 & \cdots & 5 \\ 50 & 110 & \cdots & 400 \end{bmatrix} \begin{bmatrix} 1 & 2 & 50 \\ 1 & 8 & 110 \\ \vdots & \vdots & \vdots \\ 1 & 5 & 400 \end{bmatrix}$$

$$= \begin{bmatrix} 25 & 206 & 8{,}294 \\ 206 & 2{,}396 & 77{,}177 \\ 8{,}294 & 77{,}177 & 3{,}531{,}848 \end{bmatrix}$$

and the $\mathbf{X}'\mathbf{y}$ vector is

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 2 & 8 & \cdots & 5 \\ 50 & 110 & \cdots & 400 \end{bmatrix} \begin{bmatrix} 9.95 \\ 24.45 \\ \vdots \\ 21.15 \end{bmatrix} = \begin{bmatrix} 725.82 \\ 8{,}008.47 \\ 274{,}816.71 \end{bmatrix}$$

The least squares estimates are found from Equation 12-13 as

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

12-1: Multiple Linear Regression Models

Example 12-2

or

$$\begin{bmatrix}\hat{\beta}_0\\ \hat{\beta}_1\\ \hat{\beta}_2\end{bmatrix} = \begin{bmatrix}25 & 206 & 8{,}294\\ 206 & 2{,}396 & 77{,}177\\ 8{,}294 & 77{,}177 & 3{,}531{,}848\end{bmatrix}^{-1}\begin{bmatrix}725.82\\ 8{,}008.37\\ 274{,}811.31\end{bmatrix}$$

$$= \begin{bmatrix}0.214653 & -0.007491 & -0.000340\\ -0.007491 & 0.001671 & -0.000019\\ -0.000340 & -0.000019 & +0.0000015\end{bmatrix}\begin{bmatrix}725.82\\ 8{,}008.47\\ 274{,}811.31\end{bmatrix}$$

$$= \begin{bmatrix}2.26379143\\ 2.74426964\\ 0.01252781\end{bmatrix}$$

Therefore, the fitted regression model with the regression coefficients rounded to five decimal places is

$$\hat{y} = 2.26379 + 2.74427x_1 + 0.01253x_2$$

This is identical to the results obtained in Example 12-1.

12-1: Multiple Linear Regression Models

Example 12-2

This regression model can be used to predict values of pull strength for various values of wire length (x_1) and die height (x_2). We can also obtain the **fitted values** $\hat{y}_i$ by substituting each observation (x_{i1}, x_{i2}), $i = 1, 2, \ldots, n$, into the equation. For example, the first observation has $x_{11} = 2$ and $x_{12} = 50$, and the fitted value is

$$\begin{aligned}\hat{y}_1 &= 2.26379 + 2.74427x_{11} + 0.01253x_{12} \\ &= 2.26379 + 2.74427(2) + 0.01253(50) \\ &= 8.38\end{aligned}$$

The corresponding observed value is $y_1 = 9.95$. The *residual* corresponding to the first observation is

$$\begin{aligned}e_1 &= y_1 - \hat{y}_1 \\ &= 9.95 - 8.38 \\ &= 1.57\end{aligned}$$

Table 12-3 displays all 25 fitted values $\hat{y}_i$ and the corresponding residuals. The fitted values and residuals are calculated to the same accuracy as the original data.

12-1: Multiple Linear Regression Models

Example 12-2

Table 12-3 Observations, Fitted Values, and Residuals for Example 12-2

Observation Number	y_i	$\hat{y}_i$	$e_i = y_i - \hat{y}_i$	Observation Number	y_i	$\hat{y}_i$	$e_i = y_i - \hat{y}_i$
1	9.95	8.38	1.57	14	11.66	12.26	−0.60
2	24.45	25.60	−1.15	15	21.65	15.81	5.84
3	31.75	33.95	−2.20	16	17.89	18.25	−0.36
4	35.00	36.60	−1.60	17	69.00	64.67	4.33
5	25.02	27.91	−2.89	18	10.30	12.34	−2.04
6	16.86	15.75	1.11	19	34.93	36.47	−1.54
7	14.38	12.45	1.93	20	46.59	46.56	0.03
8	9.60	8.40	1.20	21	44.88	47.06	−2.18
9	24.35	28.21	−3.86	22	54.12	52.56	1.56
10	27.50	27.98	−0.48	23	56.63	56.31	0.32
11	17.08	18.40	−1.32	24	22.13	19.98	2.15
12	37.00	37.46	−0.46	25	21.15	21.00	0.15
13	41.95	41.46	0.49				

Table 12-4 Minitab Multiple Regression Output for the Wire Bond Pull Strength Data

Regression Analysis: Strength versus Length, Height

The regression equation is
Strength = 2.26 + 2.74 Length + 0.0125 Height

Predictor		Coef	SE Coef	T	P	VIF
Constant	$\hat{\beta}_0$ →	2.264	1.060	2.14	0.044	
Length	$\hat{\beta}_1$ →	2.74427	0.09352	29.34	0.000	1.2
Height	$\hat{\beta}_2$ →	0.012528	0.002798	4.48	0.000	1.2

S = 2.288 R-Sq = 98.1% R-Sq (adj) = 97.9%
PRESS = 156.163 R-Sq (pred) = 97.44%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	5990.8	2995.4	572.17	0.000
Residual Error	22	115.2	5.2 ← $\hat{\sigma}^2$		
Total	24	6105.9			

Source	DF	Seq SS
Length	1	5885.9
Height	1	104.9

Predicted Values for New Observations

New Obs	Fit	SE Fit	95.0% CI	95.0% PI
1	27.663	0.482	(26.663, 28.663)	(22.814, 32.512)

Values of Predictors for New Observations

New Obs	Length	Height
1	8.00	275

12-1: Multiple Linear Regression Models

Estimating σ^2

An unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-p} = \frac{SS_E}{n-p} \qquad (12\text{-}16)$$

12-1: Multiple Linear Regression Models

12-1.4 Properties of the Least Squares Estimators

Unbiased estimators:

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}) &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}] \\ &= \boldsymbol{\beta} \end{aligned}$$

Covariance Matrix:

$$\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} C_{00} & C_{01} & C_{02} \\ C_{10} & C_{11} & C_{12} \\ C_{20} & C_{21} & C_{22} \end{bmatrix}$$

12-1: Multiple Linear Regression Models

12-1.4 Properties of the Least Squares Estimators

Individual variances and covariances:

$$V(\hat{\beta}_j) = \sigma^2 C_{jj}, \qquad j = 0, 1, 2$$

$$\text{cov}(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 C_{ij}, \qquad i \neq j$$

In general,

$$\text{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} = \sigma^2 \mathbf{C}$$

12-2.1 Test for Significance of Regression

The appropriate hypotheses are

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$$
$$H_1: \beta_j \neq 0 \quad \text{for at least one } j \qquad (12\text{-}18)$$

The test statistic is

$$F_0 = \frac{SS_R/k}{SS_E/(n-p)} = \frac{MS_R}{MS_E} \qquad (12\text{-}19)$$

12-2.1 Test for Significance of Regression

Table 12-9 Analysis of Variance for Testing Significance of Regression in Multiple Regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	SS_R	k	MS_R	MS_R/MS_E
Error or residual	SS_E	$n - p$	MS_E	
Total	SS_T	$n - 1$		

Example 12-3

EXAMPLE 12-3 Wire Bond Strength ANOVA

We will test for significance of regression (with $\alpha = 0.05$) using the wire bond pull strength data from Example 12-1. The total sum of squares is

$$SS_T = \mathbf{y}'\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = 27{,}178.5316 - \frac{(725.82)^2}{25}$$

$$= 6105.9447$$

Example 12-3

The regression or model sum of squares is computed from Equation 12-20 as follows:

$$SS_R = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = 27{,}063.3581 - \frac{(725.82)^2}{25}$$

$$= 5990.7712$$

and by subtraction

$$SS_E = SS_T - SS_R = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} = 115.1716$$

Example 12-3

The analysis of variance is shown in Table 12-10. To test H_0: $\beta_1 = \beta_2 = 0$, we calculate the statistic

$$f_0 = \frac{MS_R}{MS_E} = \frac{2995.3856}{5.2352} = 572.17$$

Since $f_0 > f_{0.05,2,22} = 3.44$ (or since the P-value is considerably smaller than $\alpha = 0.05$), we reject the null hypothesis and conclude that pull strength is linearly related to either wire length or die height, or both.

Practical Interpretation: Rejection of H_0 does not necessarily imply that the relationship found is an appropriate model for predicting pull strength as a function of wire length and die height. Further tests of model adequacy are required before we can be comfortable using this model in practice.

12-2: Hypothesis Tests in Multiple Linear Regression

Example 12-3

Table 12-10 Test for Significance of Regression for Example 12-3

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	f_0	P-value
Regression	5990.7712	2	2995.3856	572.17	1.08E-19
Error or residual	115.1735	22	5.2352		
Total	6105.9447	24			

R^2 and Adjusted R^2

The **coefficient of multiple determination**

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

- For the wire bond pull strength data, we find that $R^2 = SS_R/SS_T = 5990.7712/6105.9447 = 0.9811$.
- Thus, the model accounts for about 98% of the variability in the pull strength response.

R^2 and Adjusted R^2

The **adjusted R^2** is

$$R^2_{\text{adj}} = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)} \qquad (12\text{-}23)$$

- The adjusted R^2 statistic penalizes the analyst for adding terms to the model.
- It can help guard against overfitting (including regressors that are not really useful)

12-2.2 Tests on Individual Regression Coefficients and Subsets of Coefficients

The hypotheses for testing the significance of any individual regression coefficient:

$$H_0: \beta_j = \beta_{j0}$$
$$H_1: \beta_j \neq \beta_{j0} \qquad (12\text{-}24)$$

12-2.2 Tests on Individual Regression Coefficients and Subsets of Coefficients

The test statistic is

$$T_0 = \frac{\hat{\beta}_j - \beta_{j0}}{\sqrt{\sigma^2 C_{jj}}} = \frac{\hat{\beta}_j - \beta_{j0}}{se(\hat{\beta}_j)} \qquad (12\text{-}25)$$

- Reject H_0 if $|t_0| > t_{\alpha/2,n-p}$.
- This is called a **partial** or marginal test

Example 12-4

EXAMPLE 12-4 Wire Bond Strength Coefficient Test

Consider the wire bond pull strength data, and suppose that we want to test the hypothesis that the regression coefficient for x_2 (die height) is zero. The hypotheses are

$$H_0: \beta_2 = 0$$
$$H_1: \beta_2 \neq 0$$

The main diagonal element of the $(\mathbf{X}'\mathbf{X})^{-1}$ matrix corresponding to $\hat{\beta}_2$ is $C_{22} = 0.0000015$, so the t-statistic in Equation 12-25 is

$$t_0 = \frac{\hat{\beta}_2}{\sqrt{\hat{\sigma}^2 C_{22}}} = \frac{0.01253}{\sqrt{(5.2352)(0.0000015)}} = 4.477$$

12-2: Hypothesis Tests in Multiple Linear Regression

Example 12-4

Note that we have used the estimate of σ^2 reported to four decimal places in Table 12-10. Since $t_{0.025,22} = 2.074$, we reject H_0: $\beta_2 = 0$ and conclude that the variable x_2 (die height) contributes significantly to the model. We could also have used a P-value to draw conclusions. The P-value for $t_0 = 4.477$ is $P = 0.0002$, so with $\alpha = 0.05$ we would reject the null hypothesis.

Practical Interpretation: Note that this test measures the marginal or partial contribution of x_2 given that x_1 is in the model. That is, the t-test measures the contribution of adding the variable x_2 = die height to a model that already contains x_1 = wire length. Table 12-4 shows the value of the t-test computed by Minitab. The Minitab t-test statistic is reported to two decimal places. Note that the computer produces a t-test for each regression coefficient in the model. These t-tests indicate that both regressors contribute to the model.

12-2: Hypothesis Tests in Multiple Linear Regression

The general regression significance test or the extra sum of squares method:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}$$

We wish to test the hypotheses:

$$H_0: \boldsymbol{\beta}_1 = \mathbf{0}$$
$$H_1: \boldsymbol{\beta}_1 \neq \mathbf{0} \qquad (12\text{-}28)$$

12-2: Hypothesis Tests in Multiple Linear Regression

A general form of the model can be written:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$$

where $\mathbf{X}_1$ represents the columns of $\mathbf{X}$ associated with $\boldsymbol{\beta}_1$ and $\mathbf{X}_2$ represents the columns of $\mathbf{X}$ associated with $\boldsymbol{\beta}_2$

12-2: Hypothesis Tests in Multiple Linear Regression

For the full model:

$$SS_R(\boldsymbol{\beta}) = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} \qquad (p = k + 1 \text{ degrees of freedom})$$

$$MS_E = \frac{\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}\mathbf{X}'\mathbf{y}}{n - p}$$

If H_0 is true, the reduced model is

$$\mathbf{y} = \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$$

$$SS_R(\boldsymbol{\beta}_2) = \hat{\boldsymbol{\beta}}_2'\mathbf{X}_2'\mathbf{y} \qquad (p - r \text{ degrees of freedom})$$

12-2: Hypothesis Tests in Multiple Linear Regression

The test statistic is:

$$F_0 = \frac{SS_R(\beta_1 | \beta_2)/r}{MS_E} \qquad (12\text{-}33)$$

Reject H_0 if $f_0 > f_{\alpha,r,n-p}$
The test in Equation (12-32) is often referred to as a **partial *F*-test**

Example 12-6

EXAMPLE 12-6 Wire Bond Strength General Regression Test

Consider the wire bond pull-strength data in Example 12-1. We will investigate the contribution of two new variables, x_3 and x_4, to the model using the partial F-test approach. The new variables are explained at the end of this example. That is, we wish to test

$$H_0: \beta_3 = \beta_4 = 0 \qquad H_1: \beta_3 \neq 0 \text{ or } \beta_4 \neq 0$$

To test this hypothesis, we need the extra sum of squares due to β_3 and β_4 or

$$\begin{aligned} SS_R(\beta_4, \beta_3|\beta_2, \beta_1, \beta_0) &= SS_R(\beta_4, \beta_3, \beta_2, \beta_1, \beta_0) - SS_R(\beta_2, \beta_1, \beta_0) \\ &= SS_R(\beta_4, \beta_3, \beta_2, \beta_1|\beta_0) - SS_R(\beta_2, \beta_1|\beta_0) \end{aligned}$$

12-2: Hypothesis Tests in Multiple Linear Regression

Example 12-6

In Example 12-3 we calculated

$$SS_R(\beta_2, \beta_1|\beta_0) = \boldsymbol{\beta}'\mathbf{X}'\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = 5990.7712 \text{ (two degrees of freedom)}$$

Also, Table 12-4 shows the Minitab output for the model with only x_1 and x_2 as predictors. In the analysis of variance table, we can see that $SS_R = 5990.8$ and this agrees with our calculation. In practice, the computer output would be used to obtain this sum of squares.

If we fit the model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$, we can use the same matrix formula. Alternatively, we can look at SS_R from computer output for this model. The analysis of variance table for this model is shown in Table 12-11 and we see that

$$SS_R(\beta_4, \beta_3, \beta_2, \beta_1|\beta_0) = 6024.0 \text{ (four degrees of freedom)}$$

Therefore,

$$SS_R(\beta_4, \beta_3|\beta_2, \beta_1, \beta_0) = 6024.0 - 5990.8 = 33.2 \text{ (two degrees of freedom)}$$

12-2: Hypothesis Tests in Multiple Linear Regression

Example 12-6

This is the increase in the regression sum of squares due to adding x_3 and x_4 to a model already containing x_1 and x_2. To test H_0, calculate the test statistic

$$f_0 = \frac{SS_R(\beta_4, \beta_3|\beta_2, \beta_1, \beta_0)/2}{MS_E} = \frac{33.2/2}{4.1} = 4.05$$

Note that MS_E from the full model using x_1, x_2, x_3 and x_4 is used in the denominator of the test statistic. Because $f_{0.05,2,20} = 3.49$, we reject H_0 and conclude that at least one of the new variables contributes significantly to the model. Further analysis and tests will be needed to refine the model and determine if one or both of x_3 and x_4 are important.

The mystery of the new variables can now be explained. These are quadratic powers of the original predictors wire length and wire height. That is, $x_3 = x_1^2$ and $x_4 = x_2^2$. A test for quadratic terms is a common use of partial F-tests. With this information and the original data for x_1 and x_2, you can use computer software to reproduce these calculations. Multiple regression allows models to be extended in such a simple manner that the real meaning of x_3 and x_4 did not even enter into the test procedure. Polynomial models such as this are discussed further in Section 12-6.

12-3.1 Confidence Intervals on Individual Regression Coefficients

Definition

A $100(1-\alpha)\%$ **confidence interval on the regression coefficient** β_j, $j = 0, 1, \ldots, k$ in the multiple linear regression model is given by

$$\hat{\beta}_j - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}} \le \beta_j \le \hat{\beta}_j + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}} \qquad (12\text{-}35)$$

12-3: Confidence Intervals in Multiple Linear Regression

Example 12-7

EXAMPLE 12-7 Wire Bond Strength Confidence Interval

We will construct a 95% confidence interval on the parameter β_1 in the wire bond pull strength problem. The point estimate of β_1 is $\hat{\beta}_1 = 2.74427$ and the diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$ corresponding to β_1 is $C_{11} = 0.001671$. The estimate of σ^2 is $\hat{\sigma}^2 = 5.2352$, and $t_{0.025,22} = 2.074$. Therefore, the 95% CI on β_1 is computed from Equation 12-35 as

$$2.74427 - (2.074)\sqrt{(5.2352)(.001671)} \le \beta_1 \le 2.74427 + (2.074)\sqrt{(5.2352)(.001671)}$$

which reduces to

$$2.55029 \le \beta_1 \le 2.93825$$

Also, computer software such as Minitab can be used to help calculate this confidence interval. From the regression output in Table 10-4, $\hat{\beta}_1 = 2.74427$ and the standard error of $\hat{\beta}_1 = 0.0935$. This standard error is the multiplier of the t-table constant in the confidence interval. That is, $0.0935 = \sqrt{(5.2352)(0.001671)}$. Consequently, all the numbers are available from the computer output to construct the interval and this is the typical method used in practice.

12-3.2 Confidence Interval on the Mean Response

The mean response at a point $\mathbf{x}_0$ is estimated by

$$\hat{\mu}_{Y|\mathbf{x}_0} = \mathbf{x}_0'\hat{\boldsymbol{\beta}}$$

The variance of the estimated mean response is

$$V(\hat{\mu}_{Y|\mathbf{x}_0}) = \sigma^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0$$

12-3.2 Confidence Interval on the Mean Response

Definition

For the multiple linear regression model, a $100(1-\alpha)\%$ **confidence interval on the mean response** at the point $x_{01}, x_{02}, \ldots, x_{0k}$ is

$$\hat{\mu}_{Y|\mathbf{x}_0} - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} \leq \mu_{Y|\mathbf{x}_0} \leq \hat{\mu}_{Y|\mathbf{x}_0} + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} \quad (12\text{-}39)$$

EXAMPLE 12-8 Wire Bond Strength Confidence Interval on the Mean Response

Example 12-8

The engineer in Example 12-1 would like to construct a 95% CI on the mean pull strength for a wire bond with wire length $x_1 = 8$ and die height $x_2 = 275$. Therefore,

$$\mathbf{x}_0 = \begin{bmatrix} 1 \\ 8 \\ 275 \end{bmatrix}$$

The estimated mean response at this point is found from Equation 12-36 as

$$\hat{\mu}_{Y|\mathbf{x}_0} = \mathbf{x}_0'\hat{\boldsymbol{\beta}} = \begin{bmatrix} 1 & 8 & 275 \end{bmatrix} \begin{bmatrix} 2.26379 \\ 2.74427 \\ 0.01253 \end{bmatrix} = 27.66$$

12-3: Confidence Intervals in Multiple Linear Regression

Example 12-8

The variance of $\hat{\mu}_{Y|x_0}$ is estimated by

$$\hat{\sigma}^2 \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0 = 5.2352\,[1\ 8\ 275]$$
$$\times \begin{bmatrix} .214653 & -.007491 & -.000340 \\ -.007491 & .001671 & -.000019 \\ -.000340 & -.000019 & .0000015 \end{bmatrix} \begin{bmatrix} 1 \\ 8 \\ 275 \end{bmatrix}$$
$$= 5.2352\,(0.0444) = 0.23244$$

Therefore, a 95% CI on the mean pull strength at this point is found from Equation 12-39 as

$$27.66 - 2.074\sqrt{0.23244} \le \mu_{Y|x_0} \le 27.66 + 2.074\sqrt{0.23244}$$

which reduces to

$$26.66 \le \mu_{Y|x_0} \le 28.66$$

12-4: Prediction of New Observations

A point estimate of the future observation Y_0 is

$$\hat{y}_0 = \mathbf{x}_0'\hat{\boldsymbol{\beta}}$$

A 100(1-α)% **prediction interval** for this future observation is

A $100(1-\alpha)\%$ **prediction interval for this future observation** is

$$\hat{y}_0 - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2(1+\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)} \le Y_0 \le \hat{y}_0 + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2(1+\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)} \quad (12\text{-}41)$$

12-4: Prediction of New Observations

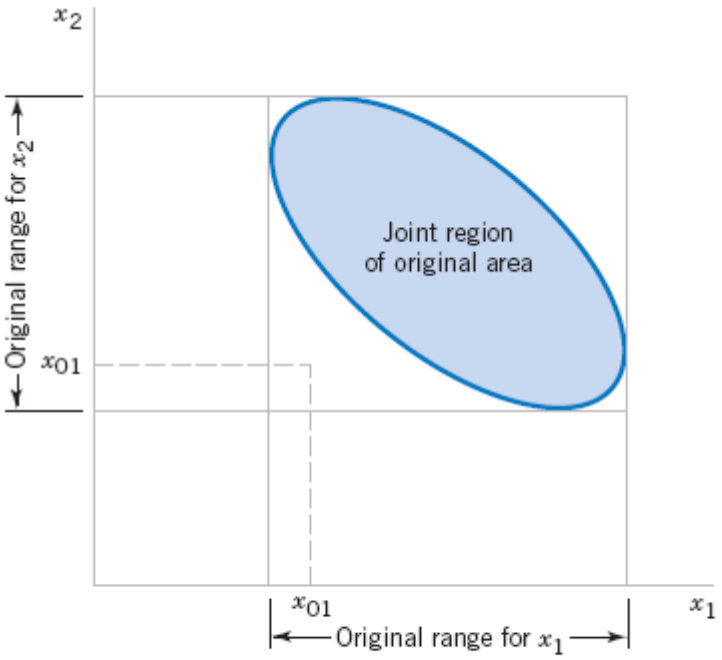


Figure 12-5 An example of extrapolation in multiple regression.

Figure 12-5 An example of extrapolation in multiple regression

12-4: Prediction of New Observations

Example 12-9

EXAMPLE 12-9 Wire Bond Strength Confidence Interval

Suppose that the engineer in Example 12-1 wishes to construct a 95% prediction interval on the wire bond pull strength when the wire length is $x_1 = 8$ and the die height is $x_2 = 275$. Note that $\mathbf{x}_0' = [1 \quad 8 \quad 275]$, and the point estimate of the pull strength is $\hat{y}_0 = \mathbf{x}_0'\hat{\boldsymbol{\beta}} = 27.66$. Also, in Example 12-8 we calculated $\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0 = 0.04444$. Therefore, from Equation 12-41 we have

$$27.66 - 2.074\sqrt{5.2352(1 + 0.0444)} \le Y_0 \le 27.66 + 2.074\sqrt{5.2352(1 + 0.0444)}$$

and the 95% prediction interval is

$$22.81 \le Y_0 \le 32.51$$

Notice that the prediction interval is wider than the confidence interval on the mean response at the same point, calculated in Example 12-8. The Minitab output in Table 12-4 also displays this prediction interval.

12-5: Model Adequacy Checking

12-5.1 Residual Analysis

Example 12-10

The residuals for the model from Example 12-1 are shown in Table 12-3. A normal probability plot of these residuals is shown in Fig. 12-6. No severe deviations from normality are obviously apparent, although the two largest residuals ($e_{15} = 5.84$ and $e_{17} = 4.33$) do not fall extremely close to a straight line drawn through the remaining residuals.

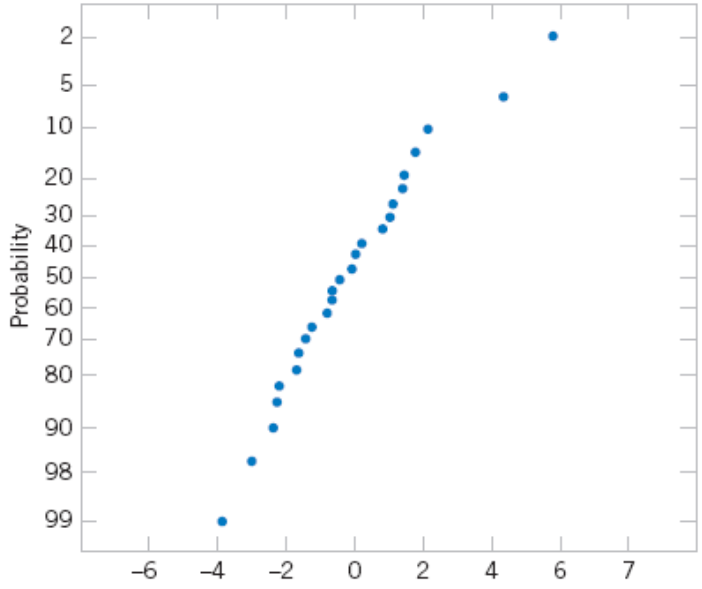


Figure 12-6 Normal probability plot of residuals.

Figure 12-6 Normal probability plot of residuals

12-5: Model Adequacy Checking

12-5.1 Residual Analysis

Example 12-10

The **standardized residuals**

$$d_i = \frac{e_i}{\sqrt{MS_E}} = \frac{e_i}{\sqrt{\hat{\sigma}^2}} \qquad (12\text{-}42)$$

are often more useful than the ordinary residuals when assessing residual magnitude. For the wire bond strength example, the standardized residuals corresponding to e_{15} and e_{17} are $d_{15} = 5.84/\sqrt{5.2352} = 2.55$ and $d_{17} = 4.33/\sqrt{5.2352} = 1.89$, and they do not seem unusually large. Inspection of the data does not reveal any error in collecting observations 15 and 17, nor does it produce any other reason to discard or modify these two points.

12-5: Model Adequacy Checking

12-5.1 Residual Analysis

Example 12-10

The residuals are plotted against $\hat{y}$ in Fig. 12-7, and against x_1 and x_2 in Figs. 12-8 and 12-9, respectively.* The two largest residuals, e_{15} and e_{17}, are apparent. Figure 12-8 gives some indication that the model underpredicts the pull strength for assemblies with short wire length ($x_1 \leq 6$) and long wire length ($x_1 \geq 15$) and overpredicts the strength for assemblies with intermediate wire length ($7 \leq x_1 \leq 14$). The same impression is obtained from Fig. 12-7. Either the relationship between strength and wire length is not linear (requiring that a term involving x_1^2, say, be added to the model), or other regressor variables not presently in the model affected the response. ___

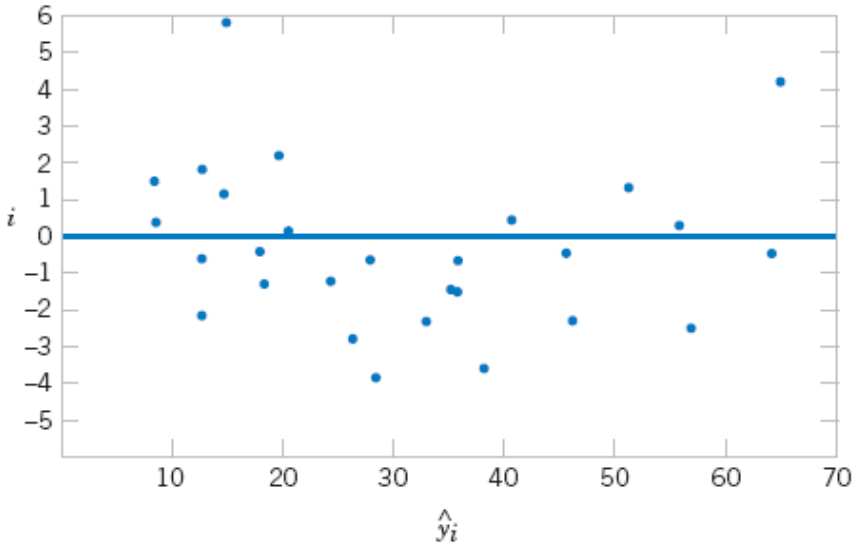


Figure 12-7 Plot of residuals against $\hat{y}$.

Figure 12-7 Plot of residuals

12-5: Model Adequacy Checking

12-5.1 Residual Analysis

Example 12-10

Either the relationship between strength and wire length is not linear (requiring that a term involving x_1^2, say, be added to the model), or other regressor variables not presently in the model affected the response.

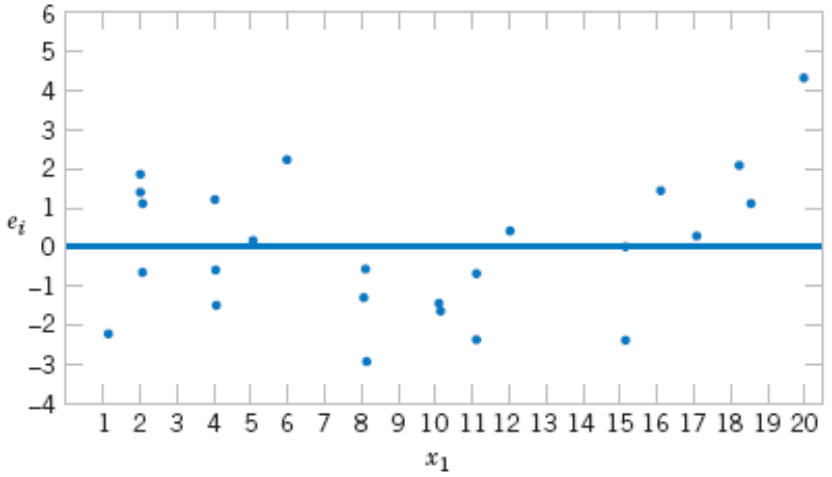


Figure 12-8 Plot of residuals against x_1.

Figure 12-8 Plot of residuals against x_1.

12-5: Model Adequacy Checking

12-5.1 Residual Analysis

Example 12-10

Figure 12-9 Plot of residuals against x_2.

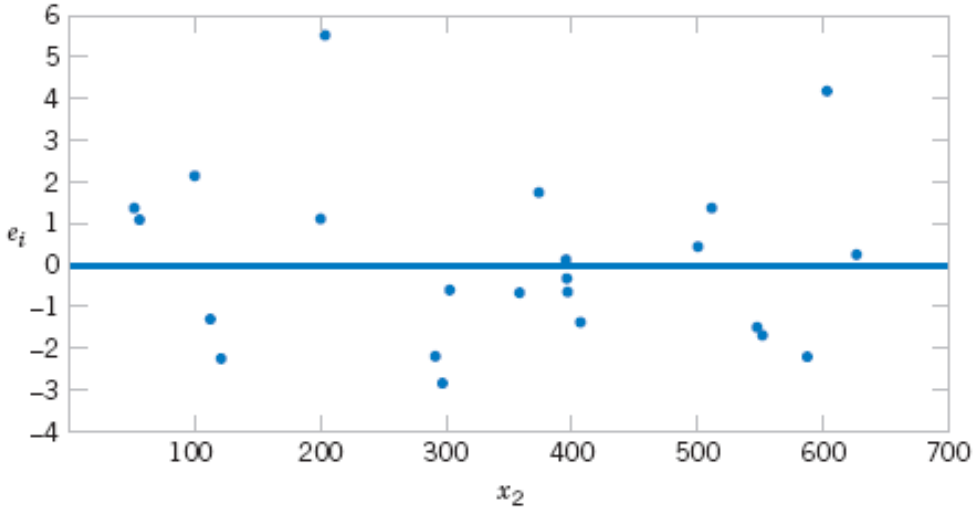


Figure 12-9 Plot of residuals against x_2.

12-5: Model Adequacy Checking

12-5.1 Residual Analysis

$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}} \qquad i = 1, 2, \ldots, n \qquad (12\text{-}43)$$

where h_{ii} is the ith diagonal element of the matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

The **H** matrix is sometimes called the **"hat" matrix,** since

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

12-5: Model Adequacy Checking

12-5.1 Residual Analysis

Since each row of the matrix $\mathbf{X}$ corresponds to a vector, say $\mathbf{x}_i' = [1, x_{i1}, x_{i2}, \ldots, x_{ik}]$, another way to write the diagonal elements of the hat matrix is

$$h_{ii} = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i \qquad (12\text{-}44)$$

The variance of the *i*th residual is

$$V(e_i) = \sigma^2(1 - h_{ii}), \qquad i = 1, 2, \ldots, n$$

12-5: Model Adequacy Checking

12-5.1 Residual Analysis

To illustrate, consider the two observations identified in the wire bond strength data (Example 12-10) as having residuals that might be unusually large, observations 15 and 17. The standardized residuals are

$$d_{15} = \frac{e_{15}}{\sqrt{\hat{\sigma}^2}} = \frac{5.84}{\sqrt{5.2352}} = 2.55 \qquad \text{and} \qquad d_{17} = \frac{e_{17}}{\sqrt{MS_E}} = \frac{4.33}{\sqrt{5.2352}} = 1.89$$

Now $h_{15,15} = 0.0737$ and $h_{17,17} = 0.2593$, so the studentized residuals are

$$r_{15} = \frac{e_{15}}{\sqrt{\hat{\sigma}^2(1 - h_{15,15})}} = \frac{5.84}{\sqrt{5.2352(1 - 0.0737)}} = 2.65$$

and

$$r_{17} = \frac{e_{17}}{\sqrt{\hat{\sigma}^2(1 - h_{17,17})}} = \frac{4.33}{\sqrt{5.2352(1 - 0.2593)}} = 2.20$$

Notice that the studentized residuals are larger than the corresponding standardized residuals. However, the studentized residuals are still not so large as to cause us serious concern about possible outliers.

12-5: Model Adequacy Checking

12-5.2 Influential Observations

Figure 12-10 A point that is remote in *x*-space.

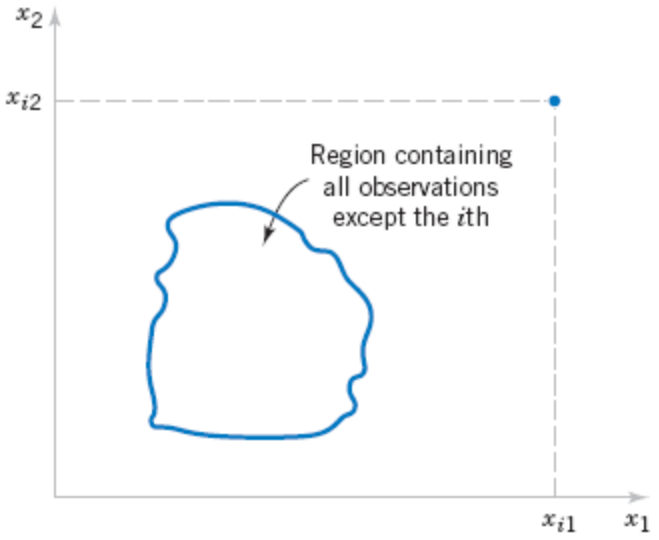


Figure 12-10 A point that is remote in *x*-space.

12-5: Model Adequacy Checking

12-5.2 Influential Observations

Cook's distance measure

$$D_i = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})' \mathbf{X}' \mathbf{X} (\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{p\hat{\sigma}^2} \qquad i = 1, 2, \ldots, n$$

$$D_i = \frac{r_i^2}{p} \frac{h_{ii}}{(1 - h_{ii})} \qquad i = 1, 2, \ldots, n \qquad (12\text{-}45)$$

12-5: Model Adequacy Checking

Example 12-11

EXAMPLE 12-11 Wire Bond Strength Cook's Distances

Table 12-12 lists the values of the hat matrix diagonals h_{ii} and Cook's distance measure D_i for the wire bond pull strength data in Example 12-1. To illustrate the calculations, consider the first observation:

$$D_1 = \frac{r_1^2}{p} \cdot \frac{h_{11}}{(1 - h_{11})}$$

$$= -\frac{[e_1/\sqrt{MS_E(1 - h_{11})}]^2}{p} \cdot \frac{h_{11}}{(1 - h_{11})}$$

$$= \frac{[1.57/\sqrt{5.2352(1 - 0.1573)}]^2}{3} \cdot \frac{0.1573}{(1 - 0.1573)}$$

$$= 0.035$$

The Cook distance measure D_i does not identify any potentially influential observations in the data, for no value of D_i exceeds unity.

12-5: Model Adequacy Checking

Example 12-11

Table 12-12 Influence Diagnostics for the Wire Bond Pull Strength Data 2

Observations i	h_{ii}	Cook's Distance Measure D_i	Observations i	h_{ii}	Cook's Distance Measure D_i
1	0.1573	0.035	14	0.1129	0.003
2	0.1116	0.012	15	0.0737	0.187
3	0.1419	0.060	16	0.0879	0.001
4	0.1019	0.021	17	0.2593	0.565
5	0.0418	0.024	18	0.2929	0.155
6	0.0749	0.007	19	0.0962	0.018
7	0.1181	0.036	20	0.1473	0.000
8	0.1561	0.020	21	0.1296	0.052
9	0.1280	0.160	22	0.1358	0.028
10	0.0413	0.001	23	0.1824	0.002
11	0.0925	0.013	24	0.1091	0.040
12	0.0526	0.001	25	0.0729	0.000
13	0.0820	0.001			

12-6: Aspects of Multiple Regression Modeling

12-6.1 Polynomial Regression Models

The linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is a general model that can be used to fit any relationship that is **linear in the unknown parameters $\boldsymbol{\beta}$.** This includes the important class of **polynomial regression models.** For example, the second-degree polynomial in one variable

$$Y = \beta_0 + \beta_1 x + \beta_{11} x^2 + \epsilon \tag{12-46}$$

and the second-degree polynomial in two variables

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon \tag{12-47}$$

are linear regression models.

12-6: Aspects of Multiple Regression Modeling

Example 12-12

EXAMPLE 12-12 Airplane Sidewall Panels

Sidewall panels for the interior of an airplane are formed in a 1500-ton press. The unit manufacturing cost varies with the production lot size. The data shown below give the average cost per unit (in hundreds of dollars) for this product (y) and the production lot size (x). The scatter diagram, shown in Fig. 12-11, indicates that a second-order polynomial may be appropriate.

y	1.81	1.70	1.65	1.55	1.48	1.40
x	20	25	30	35	40	50
y	1.30	1.26	1.24	1.21	1.20	1.18
x	60	65	70	75	80	90

12-6: Aspects of Multiple Regression Modeling

Example 12-11

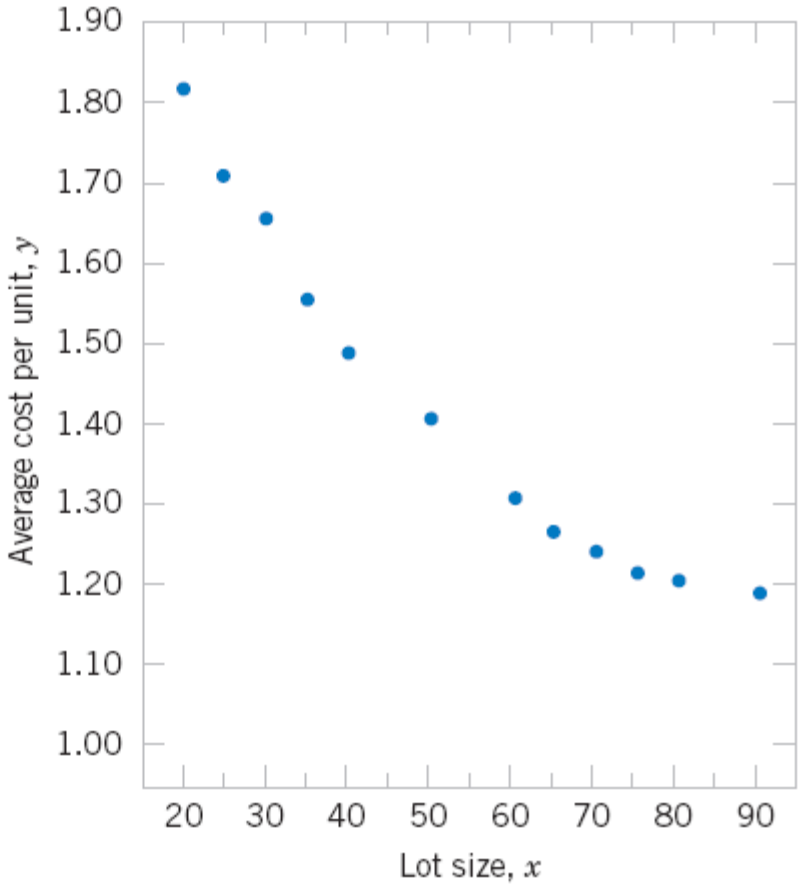


Figure 12-11 Data for Example 12-11.

Figure 12-11 Data for Example 12-11.

Example 12-12

We will fit the model

$$Y = \beta_0 + \beta_1 x + \beta_{11} x^2 + \epsilon$$

The **y** vector, the model matrix **X** and the **β** vector are as follows:

$$\mathbf{y} = \begin{bmatrix} 1.81 \\ 1.70 \\ 1.65 \\ 1.55 \\ 1.48 \\ 1.40 \\ 1.30 \\ 1.26 \\ 1.24 \\ 1.21 \\ 1.20 \\ 1.18 \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & 20 & 400 \\ 1 & 25 & 625 \\ 1 & 30 & 900 \\ 1 & 35 & 1225 \\ 1 & 40 & 1600 \\ 1 & 50 & 2500 \\ 1 & 60 & 3600 \\ 1 & 65 & 4225 \\ 1 & 70 & 4900 \\ 1 & 75 & 5625 \\ 1 & 80 & 6400 \\ 1 & 90 & 8100 \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_{11} \end{bmatrix}$$

12-6: Aspects of Multiple Regression Modeling

Example 12-12

Solving the normal equations $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$ gives the fitted model

$$\hat{y} = 2.19826629 - 0.02252236x + 0.00012507x^2$$

Conclusions: The test for significance of regression is shown in Table 12-13. Since $f_0 = 1762.3$ is significant at 1%, we conclude that at least one of the parameters β_1 and β_{11} is not zero. Furthermore, the standard tests for model adequacy do not reveal any unusual behavior, and we would conclude that this is a reasonable model for the sidewall panel cost data.

Table 12-13 Test for Significance of Regression for the Second-Order Model in Example 12-12

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	f_0	P-value
Regression	0.52516	2	0.26258	1762.28	2.12E-12
Error	0.00134	9	0.00015		
Total	0.5265	11			

12-6.2 Categorical Regressors and Indicator Variables

- Many problems may involve **qualitative** or **categorical** variables.
- The usual method for the different levels of a qualitative variable is to use **indicator** variables.
- For example, to introduce the effect of two different operators into a regression model, we could define an indicator variable as follows:

$$x = \begin{cases} 0 \text{ if the observation is from operator 1} \\ 1 \text{ if the observation is from operator 2} \end{cases}$$

12-6: Aspects of Multiple Regression Modeling

Example 12-13

EXAMPLE 12-13 Surface Finish

A mechanical engineer is investigating the surface finish of metal parts produced on a lathe and its relationship to the speed (in revolutions per minute) of the lathe. The data are shown in Table 12-15. Note that the data have been collected using two different types of cutting tools. Since the type of cutting tool likely affects the surface finish, we will fit the model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

where Y is the surface finish, x_1 is the lathe speed in revolutions per minute, and x_2 is an indicator variable denoting the type of cutting tool used; that is,

$$x_2 = \begin{cases} 0, \text{ for tool type 302} \\ 1, \text{ for tool type 416} \end{cases}$$

12-6: Aspects of Multiple Regression Modeling

Example 12-13

Table 12-15 Surface Finish Data for Example 12-13

Observation Number, i	Surface Finish y_i	RPM	Type of Cutting Tool	Observation Number, i	Surface Finish y_i	RPM	Type of Cutting Tool
1	45.44	225	302	11	33.50	224	416
2	42.03	200	302	12	31.23	212	416
3	50.10	250	302	13	37.52	248	416
4	48.75	245	302	14	37.13	260	416
5	47.92	235	302	15	34.70	243	416
6	47.79	237	302	16	33.92	238	416
7	52.26	265	302	17	32.13	224	416
8	50.52	259	302	18	35.47	251	416
9	45.58	221	302	19	33.49	232	416
10	44.78	218	302	20	32.29	216	416

12-6: Aspects of Multiple Regression Modeling

Example 12-13

The parameters in this model may be easily interpreted. If $x_2 = 0$, the model becomes

$$Y = \beta_0 + \beta_1 x_1 + \epsilon$$

which is a straight-line model with slope β_1 and intercept β_0. However, if $x_2 = 1$, the model becomes

$$Y = \beta_0 + \beta_1 x_1 + \beta_2(1) + \epsilon = (\beta_0 + \beta_2) + \beta_1 x_1 + \epsilon$$

which is a straight-line model with slope β_1 and intercept $\beta_0 + \beta_2$. Thus, the model $Y = \beta_0 + \beta_1 x + \beta_2 x_2 + \epsilon$ implies that surface finish is linearly related to lathe speed and that the slope β_1 does not depend on the type of cutting tool used. However, the type of cutting tool does affect the intercept, and β_2 indicates the change in the intercept associated with a change in tool type from 302 to 416.

Example 12-12

The model matrix **X** and **y** vector for this problem are as follows:

$$\mathbf{X} = \begin{bmatrix} 1 & 225 & 0 \\ 1 & 200 & 0 \\ 1 & 250 & 0 \\ 1 & 245 & 0 \\ 1 & 235 & 0 \\ 1 & 237 & 0 \\ 1 & 265 & 0 \\ 1 & 259 & 0 \\ 1 & 221 & 0 \\ 1 & 218 & 0 \\ 1 & 224 & 1 \\ 1 & 212 & 1 \\ 1 & 248 & 1 \\ 1 & 260 & 1 \\ 1 & 243 & 1 \\ 1 & 238 & 1 \\ 1 & 224 & 1 \\ 1 & 251 & 1 \\ 1 & 232 & 1 \\ 1 & 216 & 1 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} 45.44 \\ 42.03 \\ 50.10 \\ 48.75 \\ 47.92 \\ 47.79 \\ 52.26 \\ 50.52 \\ 45.58 \\ 44.78 \\ 33.50 \\ 31.23 \\ 37.52 \\ 37.13 \\ 34.70 \\ 33.92 \\ 32.13 \\ 35.47 \\ 33.49 \\ 32.29 \end{bmatrix}$$

12-6: Aspects of Multiple Regression Modeling

Example 12-13

The fitted model is

$$\hat{y} = 14.27620 + 0.14115x_1 - 13.28020x_2$$

Conclusions: The analysis of variance for this model is shown in Table 12-16. Note that the hypothesis $H_0: \beta_1 = \beta_2 = 0$ (significance of regression) would be rejected at any reasonable level of significance because the P-value is very small. This table also contains the sums of squares

$$\begin{aligned} SS_R &= SS_R(\beta_1,\beta_2|\beta_0) \\ &= SS_R(\beta_1|\beta_0) + SS_R(\beta_2|\beta_1,\beta_0) \end{aligned}$$

so a test of the hypothesis $H_0: \beta_2 = 0$ can be made. Since this hypothesis is also rejected, we conclude that tool type has an effect on surface finish.

12-6: Aspects of Multiple Regression Modeling

Example 12-13

Table 12-16 Analysis of Variance for Example 12-13

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	f_0	P-value
Regression	1012.0595	2	506.0297	1103.69	1.02E-18
$SS_R(\beta_1\|\beta_0)$	130.6091	1	130.6091	284.87	4.70E-12
$SS_R(\beta_2\|\beta_1,\beta_0)$	881.4504	1	881.4504	1922.52	6.24E-19
Error	7.7943	17	0.4585		
Total	1019.8538	19			

12-6: Aspects of Multiple Regression Modeling

12-6.3 Selection of Variables and Model Building

$$C_p = \frac{SS_E(p)}{\hat{\sigma}^2} - n + 2p \qquad (12\text{-}48)$$

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

12-6: Aspects of Multiple Regression Modeling

12-6.3 Selection of Variables and Model Building

All Possible Regressions – Example 12-14

EXAMPLE 12-14 Wine Quality

Table 12-17 presents data on taste-testing 38 brands of pinot noir wine (the data were first reported in an article by Kwan, Kowalski, and Skogenboe in an article in the *Journal of Agricultural and Food Chemistry,* Vol. 27, 1979, and it also appears as one of the default data sets in Minitab). The response variable is y = quality, and we wish to find the "best" regression equation that relates quality to the other five parameters.

Figure 12-12 is the matrix of scatter plots for the wine quality data, as constructed by Minitab. We notice that there are some indications of possible linear relationships between quality and the regressors, but there is no obvious visual impression of which regressors would be appropriate. Table 12-18 lists the all possible regressions output from Minitab. In this analysis, we asked Minitab to present the best three equations for each subset size. Note that Minitab reports the values of R^2, R^2_{adj}, C_p, and $S = \sqrt{MS_E}$ for each model.

12-6: Aspects of Multiple Regression Modeling

12-6.3 Selection of Variables and Model Building

All Possible Regressions – Example 12-14

From Table 12-18 we see that the three-variable equation with x_2 = aroma, x_4 = flavor, and x_5 = oakiness produces the minimum C_p equation, whereas the four-variable model, which adds x_1 = clarity to the previous three regressors, results in maximum R^2_{adj} (or minimum MS_E).

The three-variable model is

$$\hat{y} = 6.47 + 0.580x_2 + 1.20x_4 - 0.602x_5$$

and the four-variable model is

$$\hat{y} = 4.99 + 1.79x_1 + 0.530x_2 + 1.26x_4 - 0.659x_5$$

12-6: Aspects of Multiple Regression Modeling

12-6.3 Selection of Variables and Model Building

All Possible Regressions – Example 12-14

Figure 12-12 A matrix of Scatter plots from Minitab for the Wine Quality Data.

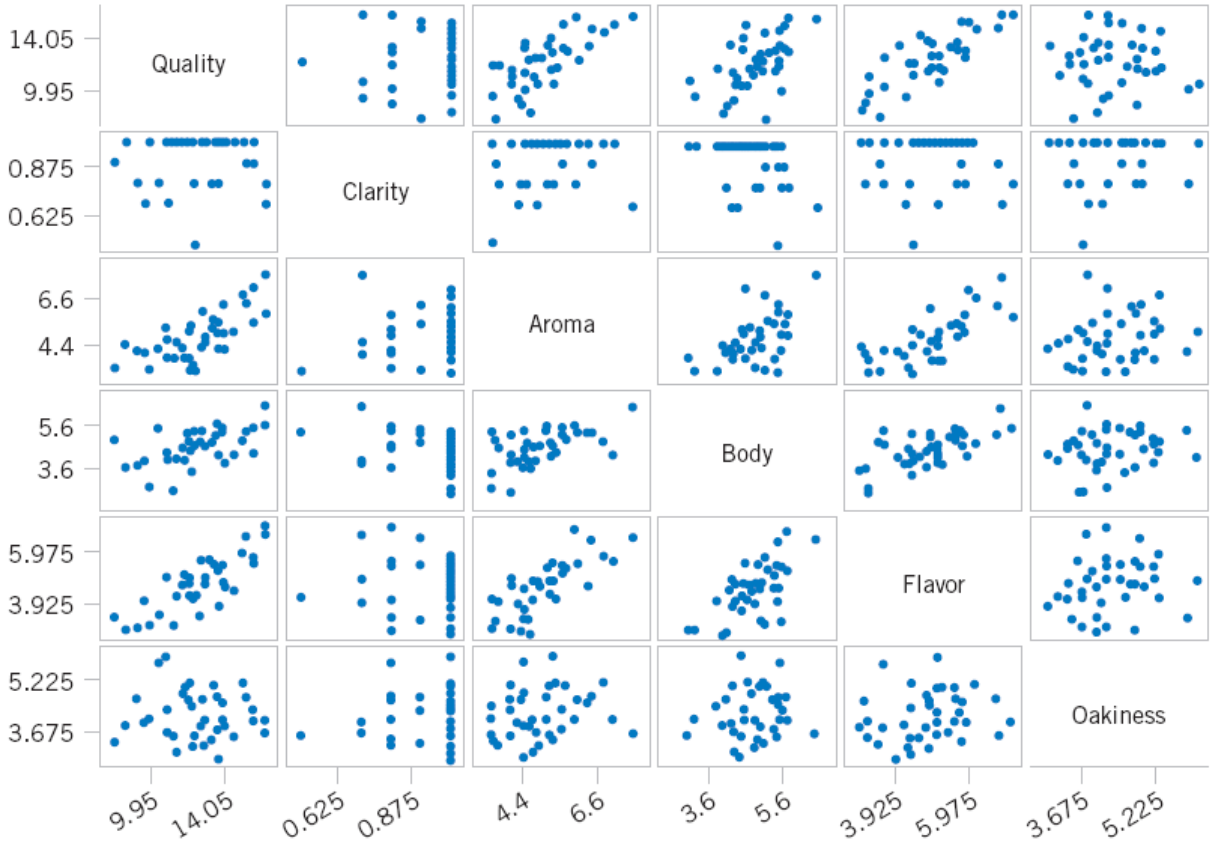


Figure 12-12 A matrix of scatter plots from Minitab for the wine quality data.

Table 12-18 Minitab All Possible Regressions Output for the Wine Quality Data

Best Subsets Regression: Quality versus Clarity, Aroma, . . .

Response is Quality

Vars	R-Sq	R-Sq (adj)	C–p	S	Clarity	Aroma	Body	Flavor	Oakiness
1	62.4	61.4	9.0	1.2712				X	
1	50.0	48.6	23.2	1.4658		X			
1	30.1	28.2	46.0	1.7335			X		
2	66.1	64.2	6.8	1.2242				X	X
2	65.9	63.9	7.1	1.2288		X		X	
2	63.3	61.2	10.0	1.2733	X			X	
3	70.4	67.8	3.9	1.1613		X		X	X
3	68.0	65.2	6.6	1.2068	X			X	X
3	66.5	63.5	8.4	1.2357			X	X	X
4	71.5	68.0	4.7	1.1568	X	X		X	X
4	70.5	66.9	5.8	1.1769		X	X	X	X
4	69.3	65.6	7.1	1.1996	X		X	X	X
5	72.1	67.7	6.0	1.1625	X	X	X	X	X

12-6.3: Selection of Variables and Model Building - Stepwise Regression

Example 12-14

Table 12-19 Minitab Stepwise Regression Output for the Wine Quality Data

Stepwise Regression: Quality versus Clarity, Aroma, . . .

Alpha-to-Enter: 0.15 Alpha-to-Remove: 0.15

Response is Quality on 5 predictors, with N = 38

Step	1	2	3
Constant	4.941	6.912	6.467
Flavor	1.57	1.64	1.20
T-Value	7.73	8.25	4.36
P-Value	0.000	0.000	0.000
Oakiness		−0.54	−0.60
T-Value		−1.95	−2.28
P-Value		0.059	0.029
Aroma			0.58
T-Value			2.21
P-Value			0.034
S	1.27	1.22	1.16
R-Sq	62.42	66.11	70.38
R-Sq(adj)	61.37	64.17	67.76
C–p	9.0	6.8	3.9

12-6.3: Selection of Variables and Model Building - Backward Regression

Example 12-14

Table 12-20 Minitab Backward Elimination Output for the Wine Quality Data

Stepwise Regression: Quality versus Clarity, Aroma, ...

Backward elimination. Alpha-to-Remove: 0.1

Response is Quality on 5 predictors, with N = 38

Step	1	2	3
Constant	3.997	4.986	6.467
Clarity	2.3	1.8	
T-Value	1.35	1.12	
P-Value	0.187	0.269	
Aroma	0.48	0.53	0.58
T-Value	1.77	2.00	2.21
P-Value	0.086	0.054	0.034
Body	0.27		
T-Value	0.82		
P-Value	0.418		
Flavor	1.17	1.26	1.20
T-Value	3.84	4.52	4.36
P-Value	0.001	0.000	0.000
Oakiness	−0.68	−0.66	−0.60
T-Value	−2.52	−2.46	−2.28
P-Value	0.017	0.019	0.029
S	1.16	1.16	1.16
R-Sq	72.06	71.47	70.38
R-Sq(adj)	67.69	68.01	67.76
C–p	6.0	4.7	3.9

12-6: Aspects of Multiple Regression Modeling

12-6.4 Multicollinearity

Variance Inflation Factor (VIF)

$$VIF(\beta_j) = \frac{1}{(1 - R_j^2)} \qquad j = 1, 2, \ldots, k \qquad (12\text{-}51)$$

12-6: Aspects of Multiple Regression Modeling

12-6.4 Multicollinearity

The presence of multicollinearity can be detected in several ways. Two of the more easily understood of these are:

1. The **variance inflation factors,** defined in equation 12-50, are very useful measures of multicollinearity. The larger the variance inflation factor, the more severe the multicollinearity. Some authors have suggested that if any variance inflation factor exceeds 10, multicollinearity is a problem. Other authors consider this value too liberal and suggest that the variance inflation factors should not exceed 4 or 5. Minitab will calculate the variance inflation factors. Table 12-4 presents the Minitab multiple regression output for the wire bond pull strength data. Since both VIF_1 and VIF_2 are small, there is no problem with multicollinearity.
2. If the F-test for significance of regression is significant, but tests on the individual regression coefficients are not significant, multicollinearity may be present.

Important Terms & Concepts

All possible regressions

Analysis of variance test in multiple regression

Categorical variables

Confidence intervals on the mean response

Cp statistic

Extra sum of squares method

Hidden extrapolation

Indicator variables

Inference (test & intervals) on individual model parameters

Influential observations

Model parameters & their interpretation in multiple regression

Multicollinearity

Multiple regression

Outliers

Polynomial regression model

Prediction interval on a future observation

PRESS statistic

Residual analysis & model adequacy checking

Significance of regression

Stepwise regression & related methods

Variance Inflation Factor (VIF)