CHAPTER 3

Multiple Linear Regression

Multiple Regression Models

- Suppose that the yield in pounds of conversion in a chemical process depends on temperature and the catalyst concentration. A **multiple regression model** that might describe this relationship is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \qquad (3.1)$$

- This is a multiple linear regression model in two variables.

Multiple Regression Models

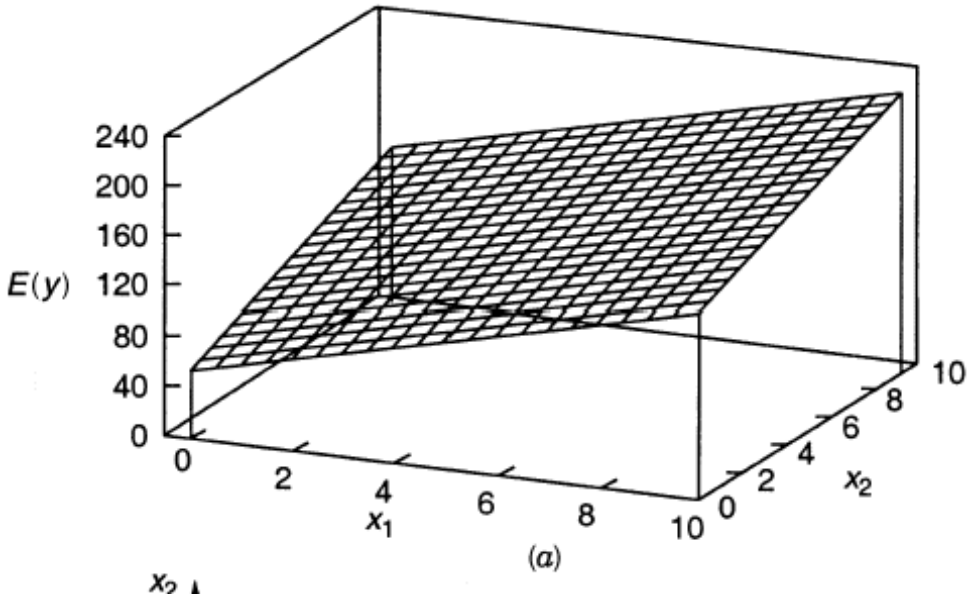


Figure 3.1 (a) The regression plane for the model E(y)= $50+10x_1+7x_2$.

Multiple Regression Models

- In general, the multiple linear regression model with *k* regressors is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

Multiple Regression Models

Models that are more complex in structure than Eq. (3.2) may often still be analyzed by multiple linear regression techniques. For example, consider the cubic polynomial model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon \tag{3.3}$$

If we let $x_1 = x$, $x_2 = x^2$, and $x_3 = x^3$, then Eq. (3.3) can be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon \tag{3.4}$$

Multiple Regression Models

- Linear regression models may also contain **interaction** effects:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon$$

- Let $x_3 = x_1 x_2$ and $\beta_3 = \beta_{12}$, the model can be written in the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

Multiple Regression Models

- A regression model that is linear in the parameters (β_i) is a linear regression model regardless of the shape of the surface it generates.

Estimation of the Model Parameters

- Least Squares Estimation of the Regression Coefficients

Notation

n – number of observations available

k – number of regressor variables

y – response or dependent variable

x_{ij} – ith observation or level of regressor j.

$$E(\varepsilon) = 0, \mathrm{Var}(\varepsilon) = \sigma^2$$

Estimation of the Model Parameters

- Least Squares Estimation of the Regression Coefficients

TABLE 3.1 Data for Multiple Linear Regression

Observation, i	Response, y	Regressors			
		x_1	x_2	$\cdots$	x_k
1	y_1	x_{11}	x_{12}	$\cdots$	x_{1k}
2	y_2	x_{21}	x_{22}	$\cdots$	x_{2k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	y_n	x_{n1}	x_{n2}	$\cdots$	x_{nk}

Estimation of the Model Parameters

- Least Squares Estimation of the Regression Coefficients

The sample regression model can be written as

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

$$= \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

Estimation of the Model Parameters

- Least Squares Estimation of the Regression Coefficients

  The least squares function is

$$S(\beta_0, \beta_1, \ldots, \beta_k) = \sum_{i=1}^{n} \varepsilon_i^2$$

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

  The function S must be minimized with respect to the coefficients.

Estimation of the Model Parameters

- Least Squares Estimation of the Regression Coefficients

  The least squares estimates of the coefficients must satisfy

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

and

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

Estimation of the Model Parameters

- Simplifying, we obtain the least squares normal equations:

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

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

The ordinary least squares estimators are the solutions to the normal equations.

Estimation of the Model Parameters

- Matrix notation is typically used:

Let $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Estimation of the Model Parameters

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

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

Note that $S(\boldsymbol{\beta})$ may be expressed as

$$\begin{aligned} S(\boldsymbol{\beta}) &= \mathbf{y}'\mathbf{y} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \end{aligned}$$

since $\boldsymbol{\beta}'\mathbf{X}'\mathbf{y}$ is a 1×1 matrix, or a scalar, and its transpose $(\boldsymbol{\beta}'\mathbf{X}'\mathbf{y})' = \mathbf{y}'\mathbf{X}\boldsymbol{\beta}$ is the same scalar. The least-squares estimators must satisfy

Estimation of the Model Parameters

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

which simplifies to

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y} \qquad (3.12)$$

- These are the **least-squares normal equations**. The solution is

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

Estimation of the Model Parameters

The vector of fitted values $\hat{y}_i$ corresponding to the observed values y_i is

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y} \qquad (3.14)$$

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

Estimation of the Model Parameters

- The n residuals can be written in matrix form as

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$$

- There will be some situations where an alternative form will prove useful

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

Example 3-1. The Delivery Time Data

The model of interest is

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

TABLE 3.2 Delivery Time Data for Example 3.1

Observation Number	Delivery Time (Minutes) y	Number of Cases x_1	Distance (Feet) x_2
1	16.68	7	560
2	11.50	3	220
3	12.03	3	340
4	14.88	4	80
5	13.75	6	150
6	18.11	7	330
7	8.00	2	110
8	17.83	7	210
9	79.24	30	1460
10	21.50	5	605
11	40.33	16	688
12	21.00	10	215
13	13.50	4	255
14	19.75	6	462
15	24.00	9	448
16	29.00	10	776
17	15.35	6	200
18	19.00	7	132
19	9.50	3	36
20	35.10	17	770
21	17.90	10	140
22	52.32	26	810
23	18.75	9	450
24	19.83	8	635
25	10.75	4	150

Example 3-1. The Delivery Time Data

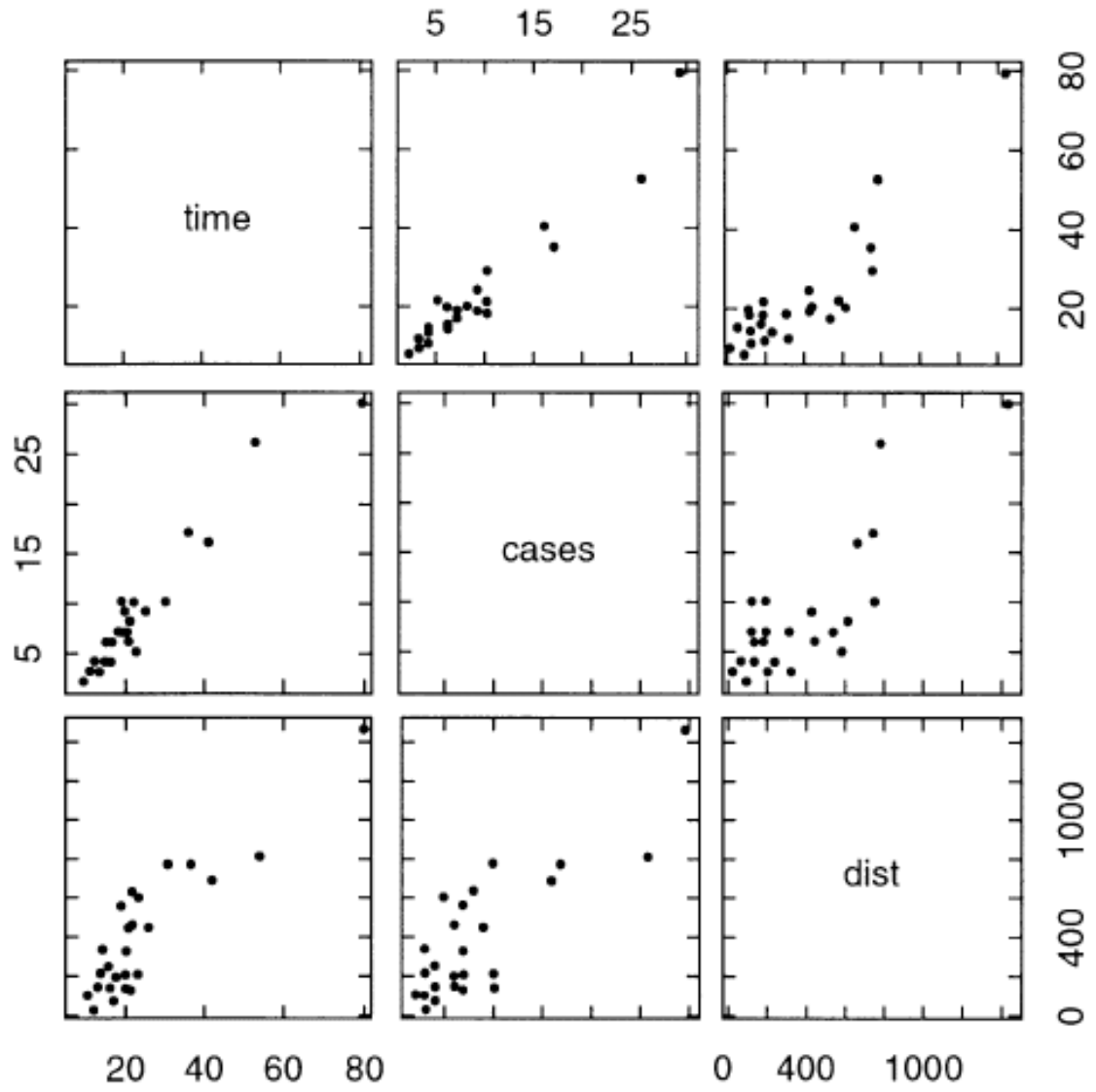


Figure 3.4 Scatterplot matrix for the delivery time data from Example 3.1.

Inadequacy of Scatter Diagrams in Multiple Regression

- Scatter diagrams of the regressor variable(s) against the response may be misleading
- Interdependency between two or more regressor variables, can mask the true relationship between x_i and y

Example 3-1 The Delivery Time Data

$$\mathbf{X} = \begin{bmatrix} 1 & 7 & 560 \\ 1 & 3 & 220 \\ 1 & 3 & 340 \\ 1 & 4 & 80 \\ 1 & 6 & 150 \\ 1 & 7 & 330 \\ 1 & 2 & 110 \\ 1 & 7 & 210 \\ 1 & 30 & 1460 \\ 1 & 5 & 605 \\ 1 & 16 & 688 \\ 1 & 10 & 215 \\ 1 & 4 & 255 \\ 1 & 6 & 462 \\ 1 & 9 & 448 \\ 1 & 10 & 776 \\ 1 & 6 & 200 \\ 1 & 7 & 132 \\ 1 & 3 & 36 \\ 1 & 17 & 770 \\ 1 & 10 & 140 \\ 1 & 26 & 810 \\ 1 & 9 & 450 \\ 1 & 8 & 635 \\ 1 & 4 & 150 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} 16.68 \\ 11.50 \\ 12.03 \\ 14.88 \\ 13.75 \\ 18.11 \\ 8.00 \\ 17.83 \\ 79.24 \\ 21.50 \\ 40.33 \\ 21.00 \\ 13.50 \\ 19.75 \\ 24.00 \\ 29.00 \\ 15.35 \\ 19.00 \\ 9.50 \\ 35.10 \\ 17.90 \\ 52.32 \\ 18.75 \\ 19.83 \\ 10.75 \end{bmatrix}$$

Example 3-1 The Delivery Time Data

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

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 7 & 3 & \cdots & 4 \\ 560 & 220 & \cdots & 150 \end{bmatrix} \begin{bmatrix} 1 & 7 & 560 \\ 1 & 3 & 220 \\ \vdots & \vdots & \vdots \\ 1 & 4 & 150 \end{bmatrix}$$

$$= \begin{bmatrix} 25 & 219 & 10{,}232 \\ 219 & 3{,}055 & 133{,}899 \\ 10{,}232 & 133{,}899 & 6{,}725{,}688 \end{bmatrix}$$

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

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 7 & 3 & \cdots & 4 \\ 560 & 220 & \cdots & 150 \end{bmatrix} \begin{bmatrix} 16.68 \\ 11.50 \\ \vdots \\ 10.75 \end{bmatrix} = \begin{bmatrix} 559.60 \\ 7{,}375.44 \\ 337{,}072.00 \end{bmatrix}$$

Example 3-1 The Delivery Time Data

The least-squares estimator of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 25 & 219 & 10{,}232 \\ 219 & 3{,}055 & 133{,}899 \\ 10{,}232 & 133{,}899 & 6{,}725{,}688 \end{bmatrix}^{-1} \begin{bmatrix} 559.60 \\ 7{,}375.44 \\ 337{,}072.00 \end{bmatrix}$$

$$= \begin{bmatrix} .11321518 & -.00444859 & -.00008367 \\ -.00444859 & .00274378 & -.00004786 \\ -.00008367 & -.00004786 & .00000123 \end{bmatrix} \begin{bmatrix} 559.60 \\ 7{,}375.44 \\ 337{,}072.00 \end{bmatrix}$$

$$= \begin{bmatrix} 2.34123115 \\ 1.61590712 \\ 0.01438483 \end{bmatrix}$$

$$\hat{y} = 2.34123 + 1.61591x_1 + 0.01438x_2$$

TABLE 3.3 Observations, Fitted Values, and Residuals for Example 3.1

Observation Number	y_i	$\hat{y}_i$	$e_i = y_i - \bar{y}_i$
1	16.68	21.7081	−5.0281
2	11.50	10.3536	1.1464
3	12.03	12.0798	−0.0498
4	14.88	9.9556	4.9244
5	13.75	14.1944	−0.4444
6	18.11	18.3996	−0.2896
7	8.00	7.1554	0.8446
8	17.83	16.6734	1.1566
9	79.24	71.8203	7.4197
10	21.50	19.1236	2.3764
11	40.33	38.0925	2.2375
12	21.00	21.5930	−0.5930
13	13.50	12.4730	1.0270
14	19.75	18.6825	1.0675
15	24.00	23.3288	0.6712
16	29.00	29.6629	−0.6629
17	15.35	14.9136	0.4364
18	19.00	15.5514	3.4486
19	9.50	7.7068	1.7932
20	35.10	40.8880	−5.7880
21	17.90	20.5142	−2.6142
22	52.32	56.0065	−3.6865
23	18.75	23.3576	−4.6076
24	19.83	24.4028	−4.5728
25	10.75	10.9626	−0.2126

Example 3-1 Excel Output

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.980
R Square	0.960
Adjusted R Square	0.956
Standard Error	3.259
Observations	25

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	5550.811	2775.405	261.235	4.68742E-16
Residual	22	233.732	10.624		
Total	24	5784.543			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	2.341	1.097	2.135	0.044	0.067	4.616	-0.750	5.433
Number of Cases, x_1	1.616	0.171	9.464	3.25E-09	1.262	1.970	1.135	2.097
Distance, x_2 (ft)	0.014	0.004	3.981	0.001	0.007	0.022	0.004	0.025

R code

```
rm(list=ls())
# read data
delivery <- read.csv("eg3.1.delivery.csv",h=T)
names(delivery)
cor(delivery)
# visualize data
pairs (delivery,pch=20)
# fit linear regression
model1 <- lm(DeliveryTime ~ NumberofCases+Distance, data=delivery)
summary(model1)
anova(model1)
# generate confidence interval for the mean response at each x
predict(model1,delivery,level=.95,interval="confidence")
# generate prediction interval for a new observation y at NumberofCaese=15, Distance=1500
predict(model1,list(NumberofCases=15, Distance=1500),interval="pred")
# check VIF
library(car)
vif(model1)
# obtain beta estimate
model1$coefficients
# a manual way to calculate beta estimate
X=as.matrix(cbind(1,delivery[,c("NumberofCases","Distance")]))
Y=delivery$DeliveryTime
beta_hat=solve(t(X)%*%X)%*%t(X)%*%Y
# you can see beta_hat is the same as model1$coefficients
```

Properties of Least-Squares Estimators

- Statistical Properties

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$$

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1} = \sigma^2 \mathrm{C}$$

- Variances/Covariances

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_j) = \sigma^2 \mathrm{C}_{jj}$$

$$\mathrm{Cov}(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 \mathrm{C}_{ij}$$

Estimation of σ^2

- The residual sum of squares can be shown to be:

$$SS_{\text{Res}} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}$$

- The residual mean square for the model with *p* parameters is:

$$MS_{\text{Res}} = \frac{SS_{\text{Res}}}{n-p} = \hat{\sigma}^2$$

Estimation of σ^2

- Recall that the estimator of σ^2 is **model dependent**
- Change the form of the model and the estimate of σ^2 will change
- Note that the variance estimate is a function of the errors: “unexplained noise about the fitted regression line”

Example 3.2 Delivery Time Data

$$\mathbf{y}'\mathbf{y} = \sum_{i=1}^{25} y_i^2 = 18{,}310.6290$$

$$\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} = \begin{bmatrix} 2.34123115 & 1.61590721 & 0.01438483 \end{bmatrix} \begin{bmatrix} 559.60 \\ 7{,}375.44 \\ 337{,}072.00 \end{bmatrix}$$

$$= 18{,}076.90304$$

Example 3.2 Delivery Time Data

$$SS_{\text{Res}} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}$$
$$= 18{,}310.6290 - 18{,}076.9030 = 233.7260$$

$$\hat{\sigma}^2 = \frac{SS_{\text{Res}}}{n - p} = \frac{233.7260}{25 - 3} = 10.6239$$

Example 3-1 Excel Output

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.980
R Square	0.960
Adjusted R Square	0.956
Standard Error	3.259
Observations	25

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	5550.811	2775.405	261.235	4.68742E-16
Residual	22	233.732	10.624		
Total	24	5784.543			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	2.341	1.097	2.135	0.044	0.067	4.616	-0.750	5.433
Number of Cases, x_1	1.616	0.171	9.464	3.25E-09	1.262	1.970	1.135	2.097
Distance, x_2 (ft)	0.014	0.004	3.981	0.001	0.007	0.022	0.004	0.025

R code

- # below we estimate sigma square
- # first obtain sample size
- n=length(Y)
- # obtain SSRes
- SSRes=t(Y)%*%Y-t(beta_hat)%*%t(X)%*%Y
- SSRes
- # Obtain estimate of sigma square, note here 3=2+1, 2 is the number of variable, 1 represents the intercept. Usually, we use k to denote number of variables, and p to denote number of parameters, which is usually p=k+1.
- sigma2_hat=SSRes/(n-3)
- # display our estimate of sigma square.
- sigma2_hat

Hypothesis Testing in Multiple Linear Regression

- Once we have estimated the parameters in the model, we face two immediate questions:
  - What is the overall adequacy of the model?
  - Which specific regressors seem important?

Hypothesis Testing in Multiple Linear Regression

- Test for Significance of Regression (sometimes called the global test of model adequacy)
- Tests on Individual Regression Coefficients (or groups of coefficients)
- Special Case of Hypothesis Testing with Orthogonal Columns in **X**
- Testing the General Linear Hypothesis

Test for Significance of Regression

- The test for significance is a test to determine if there is a linear relationship between the response and **any** of the regressor variables
- The hypotheses are

  H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$

  H_1: $\beta_j \neq 0$ for at least one j

Test for Significance of Regression

- As in Chapter 2, the total sum of squares can be partitioned in two parts:

$$SS_T = SS_R + SS_{Res}$$

- This leads to an ANOVA procedure with the test (F) statistic

$$F_0 = \frac{SS_R / k}{SS_{Res} /(n-k-1)} = \frac{MS_R}{MS_{Res}}$$

Test for Significance of Regression

- ANOVA Table:

TABLE 3.4 Analysis of Variance for 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_{Res}
Residual	SS_{Res}	$n-k-1$	MS_{Res}	
Total	SS_T	$n-1$		

Reject H_0 if $F_0 > F_{\alpha, k, n-k-1}$

Example 3-1 Excel Output

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.980
R Square	0.960
Adjusted R Square	0.956
Standard Error	3.259
Observations	25

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	5550.811	2775.405	261.235	4.68742E-16
Residual	22	233.732	10.624		
Total	24	5784.543			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	2.341	1.097	2.135	0.044	0.067	4.616	-0.750	5.433
Number of Cases, x_1	1.616	0.171	9.464	3.25E-09	1.262	1.970	1.135	2.097
Distance, x_2 (ft)	0.014	0.004	3.981	0.001	0.007	0.022	0.004	0.025

R code

```
# perform F test
summary(model1)
anova(model1)

# a manual way obtain F test statistic
# first obtain SST
SST=sum((delivery$DeliveryTime-mean(delivery$DeliveryTime))^2)
# then obtain SSRes
SSRes=sum((delivery$DeliveryTime-model1$fitted.values)^2)
# and obtain SSR
SSR=sum((model1$fitted.values-mean(delivery$DeliveryTime))^2)
# check to see if SSR+SSRes is the same as SST
SSR+SSRes
SST
# obtain F test statistic
F=SSR/2 / (SSRes/(n-2-1))
# note here 2 is the number of variables, i.e., k=2.
F
```

Test for Significance of Regression

- R^2
  - R^2 is calculated exactly as in simple linear regression

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

  - R^2 can be inflated simply by adding more terms to the model (even insignificant terms)

- Adjusted R^2
  - Penalizes for added terms to the model

$$R^2_{Adj} = 1 - \frac{SS_{Res}/(n-p)}{SS_T/(n-1)} = 1 - \left(\frac{n-1}{n-p}\right)\frac{SS_{Res}}{SS_T}$$

R code

- # obtain R square and adjusted R square
- summary(model1)$r.square
- summary(model1)$adj.r.squared

- # a manual way to obtain R square and adjusted R square
- R_square=1-SSRes/SST
- Adj_R_square=1-SSRes/(n-3)/(SST/(n-1))

Tests on Individual Regression Coefficients

- Hypothesis test on any single regression coefficient:

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

- Test Statistic: $t_0 = \dfrac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} = \dfrac{\hat{\beta}_j}{se(\hat{\beta}_j)}$

  - Reject H_0 if $|t_0| > t_{\alpha/2,n-k-1}$

  - This is a **partial** or **marginal** test

Example 3-1 Excel Output

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.980
R Square	0.960
Adjusted R Square	0.956
Standard Error	3.259
Observations	25

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	5550.811	2775.405	261.235	4.68742E-16
Residual	22	233.732	10.624		
Total	24	5784.543			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	2.341	1.097	2.135	0.044	0.067	4.616	-0.750	5.433
Number of Cases, x_1	1.616	0.171	9.464	3.25E-09	1.262	1.970	1.135	2.097
Distance, x_2 (ft)	0.014	0.004	3.981	0.001	0.007	0.022	0.004	0.025

R code

- # obtain results of testing of individual regression coefficients
- summary(model1)
- summary(model1)$coef

Tests on Individual Regression Coefficients

- Extra Sum of Squares
  - Measures the marginal reduction in the error sum of squares when one or more predictor variables are added to the regression model
  - Considered a partial F test

Tests on Individual Regression Coefficients

- Generalized Test
  - Test a Single $\beta_k = 0$

$$H_0: \beta_k = 0$$

$$H_1: \beta_k \neq 0$$

$$F_0 = \frac{SS_{\mathrm{R}}(\beta_1, \cdots, \beta_p) - SS_{\mathrm{R}}(\beta_1, \cdots, \beta_{k-1}, \beta_{k+1}, \cdots, \beta_p)}{MS_{\mathrm{Res}}}$$

$$= \frac{SS_{\mathrm{R}}(\beta_k | \beta_1, \cdots, \beta_{k-1}, \beta_{k+1}, \cdots, \beta_p)}{MS_{\mathrm{Res}}}$$

$$F^* \sim F(1, n-p)$$

R code

- # use F test to compare two models
- # reduced is a simpler model with one covariate, Distance
- reduced = lm(DeliveryTime ~ Distance, data=delivery) # Reduced model
- # full is a more complicated model with two covariates, NumberofCases and Distance
- full = model1 # full model
- # use F test to compare two models
- anova(reduced, full)
- summary(model1)

- # we could repeat the above procedure for another reduced model with only NumberofCases as covariate
- reduced = lm(DeliveryTime ~ NumberofCases, data=delivery) # Reduced model
- full = model1 # full model
- anova(reduced, full)

- # we can even let the reduced model be a model with no covariate, note that 1 here represents the linear regression only has intercept and no covariates.
- reduced = lm(DeliveryTime ~ 1, data=delivery) # Reduced model
- full = model1 # full model
- anova(reduced, full)
- summary(model1)

Tests on Individual Regression Coefficients

- Special Case of Orthogonal Columns in X
  - If the columns **X** are **orthogonal** to each of the other columns in **X**, the sum of squares due to β_j is free of any dependence on the other regressors in **X**.

Special Case of Orthogonal Columns in X Example

- Consider a dataset with four regressor variables and a single response.
- Fit the equation with all regressors and find that:

$$y = -19.9 + 0.0123x_1 + 27.3x_2 - 0.0655x_3 - 0.196x_4$$

- Considering the t-tests, suppose that x_3 is insignificant. So it is removed. What is the equation now?
- Generally, it is ***not***

$$y = -19.9 + 0.0123x_1 + 27.3x_2 - 0.196x_4$$

Special Case of Orthogonal Columns in X Example

- The model must be refit with the insignificant regressors left out of the model.
- The regression equation is

$$y = -24.9 + 0.0117x_1 + 31.0x_2 - 0.217x_4$$

- The refitting must be done since the coefficient estimates for an individual regressor depend on all of the regressors, x_j
- However, if the columns are **orthogonal** to each other, then there is no need to refit.

R code

```
# generate a linear regression model
x1=rnorm(1000)
x2=rnorm(1000)
y=1+2*x1+3*x2+rnorm(1000)*0.01
# visualize the data
pairs(cbind(y,x1,x2),pch=20)
# build two linear regressions and compare their coefficient estimates.  Since x1 and x2 are orthogonal, x1's coefficient estimate doesn't change too much across different models.
model2=lm(y~x1+x2)
model2
model3=lm(y~x1)
model3

# generate another linear regression model with x1 and x2 correlated (i.e., not orthogonal any more).
x1=rnorm(1000)
x2=2*x1+rnorm(1000)
y=1+2*x1+3*x2+rnorm(1000)*0.01
# visualize the data, we see the correlation between x1 and x2
pairs(cbind(y,x1,x2),pch=20)
# build two regressions.  Since x1 and x2 are not orthogonal, the coefficients of x1 from these two regressions are quite different
model2=lm(y~x1+x2)
model2
model3=lm(y~x1)
model3
```

Confidence Intervals on the Regression Coefficients

A 100(1-α) percent C.I. for the regression coefficient, β_j is:

$$\hat{\beta}_j - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}} \le \beta_j \le \hat{\beta}_j + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}}$$

Or,

$$\hat{\beta}_j - t_{\alpha/2,n-p}se(\hat{\beta}_j) \le \beta_j \le \hat{\beta}_j + t_{\alpha/2,n-p}se(\hat{\beta}_j)$$

Example 3.8 The Delivery Time Data

We now find a 95% CI for the parameter β_1 in Example 3.1. The point estimate of β_1 is $\hat{\beta}_1 = 1.61591$, the diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$ corresponding to β_1 is $C_{11} = 0.00274378$, and $\hat{\sigma}^2 = 10.6239$ (from Example 3.2). Using Eq. (3.45), we find that

$$\hat{\beta}_1 - t_{0.025,22}\sqrt{\hat{\sigma}^2 C_{11}} \le \beta_1 \le \hat{\beta}_1 + t_{0.025,22}\sqrt{\hat{\sigma}^2 C_{11}}$$

$$1.61591 - (2.074)\sqrt{(10.6239)(0.00274378)} \le \beta_1 \le 1.61591 + (2.074)\sqrt{(10.6239)(0.00274378)}$$

$$1.61591 - (2.074)(0.17073) \le \beta_1 \le 1.61591 + (2.074)(0.17073)$$

and the 95% CI on β_1 is

$$1.26181 \le \beta_1 \le 1.97001$$

Example 3-1 Excel Output

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.980
R Square	0.960
Adjusted R Square	0.956
Standard Error	3.259
Observations	25

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	5550.811	2775.405	261.235	4.68742E-16
Residual	22	233.732	10.624		
Total	24	5784.543			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	2.341	1.097	2.135	0.044	0.067	4.616	-0.750	5.433
Number of Cases, x_1	1.616	0.171	9.464	3.25E-09	1.262	1.970	1.135	2.097
Distance, x_2 (ft)	0.014	0.004	3.981	0.001	0.007	0.022	0.004	0.025

Confidence Interval Estimation of the Mean Response

- 100(1-α) percent CI on the mean response at the point $x_{01}, x_{02}, \ldots, x_{0k}$ is

$$\hat{y}_0 - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} \le E(y \mid \mathrm{x}_0) \le \hat{y}_0 + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}$$

Example 3.9 The Delivery Time Data

The soft drink bottler in Example 3.1 would like to construct a 95% CI on the mean delivery time for an outlet requiring $x_1 = 8$ cases and where the distance $x_2 = 275$ feet. Therefore,

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

The fitted value at this point is found from Eq. (3.47) as

$$\hat{y}_0 = \mathbf{x}_0'\hat{\boldsymbol{\beta}} = [1 \quad 8 \quad 275]\begin{bmatrix} 2.34123 \\ 1.61591 \\ 0.01438 \end{bmatrix} = 19.22 \text{ minutes}$$

The variance of $\hat{y}_0$ is estimated by

$$\hat{\sigma}^2 \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0 = 10.6239[1 \quad 8 \quad 275]$$
$$\times \begin{bmatrix} 0.11321518 & -0.00444859 & -0.00008367 \\ -0.00444859 & 0.00274378 & -0.00004786 \\ -0.00008367 & -0.00004786 & 0.00000123 \end{bmatrix} \begin{bmatrix} 1 \\ 8 \\ 275 \end{bmatrix}$$
$$= 10.6239(0.05346) = 0.56794$$

Therefore, a 95% CI on the mean delivery time at this point is found from Eq. (3.49) as

$$19.22 - 2.074\sqrt{0.56794} \le E(y|x_0) \le 19.22 + 2.074\sqrt{0.56794}$$

which reduces to

$$17.66 \le E(y|x_0) \le 20.78$$

Ninety-five percent of such intervals will contain the true delivery time. ■

The length of the CI or the mean response is a useful measure of the quality of the regression model. It can also be used to compare competing models. To illustrate, consider the 95% CI on the the mean delivery time when $x_1 = 8$ cases and $x_2 = 275$ feet. In Example 3.9 this CI is found to be (17.66, 20.78), and the length of this interval is 20.78 – 17.16 = 3.12 minutes. If we consider the simple linear regression model with x_1 = cases as the only regressor, the 95% CI on the mean delivery time with $x_1 = 8$ cases is (18.99, 22.97). The length of this interval is 22.47 – 18.99 = 3.45 minutes. Clearly, adding cases to the model has improved the precision of estimation. However, the change in the length of the interval depends on the location of the point in the x space. Consider the point $x_1 = 16$ cases and $x_2 = 688$ feet. The 95% CI for the multiple regression model is (36.11, 40.08) with length 3.97 minutes, and for the simple linear regression model the 95% CI at $x_1 = 16$ cases is (35.60, 40.68) with length 5.08 minutes. The improvement from the multiple regression model is even better at this point. Generally, the further the point is from the centroid of the x space, the greater the difference will be in the lengths of the two CIs.

Prediction of New Observations

- A 100(1-α) percent prediction interval for a future observation is

$$\hat{y}_0 - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2(1+\mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)} \le y_0$$
$$\le \hat{y}_0 + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2(1+\mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)}$$

Example 3.12 The Delivery Time Data

Suppose that the soft drink bottler in Example 3.1 wishes to construct a 95% prediction interval on the delivery time at an outlet where $x_1 = 8$ cases are delivered and the distance walked by the deliveryman is $x_2 = 275$ feet. Note that $\mathbf{x}_0' = [1, \ 8, \ 275]$, and the point estimate of the delivery time is $\hat{y}_0 = \mathbf{x}_0'\hat{\boldsymbol{\beta}} = 19.22$ minutes. Also, in Example 3.9 we calculated $\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0 = 0.05346$. Therefore, from (3.54) we have

$$19.22 - 2.074\sqrt{10.6239(1+0.05346)} \le y_0 \le 19.22 + 2.074\sqrt{10.6239(1+0.05346)}$$

and the 95% prediction interval is

$$12.28 \le y_0 \le 26.16$$

■

R code

- # obtain confidence interval for the mean response for each observation.
- predict(model1,delivery,level=.95,interval="confidence")
- # obtain prediction interval for a new observation with NumberofCases=15, Distance=1500.
- predict(model1,list(NumberofCases=15, Distance=1500),interval="pred")
- # we can change the significane level to 95% or 99%. 95% is by default.
- predict(model1,level=0.95,list(NumberofCases=15, Distance=1500),interval="pred")
- predict(model1,level=0.99,list(NumberofCases=15, Distance=1500),interval="pred")

Hidden Extrapolation in Multiple Regression

- In prediction, exercise care about potentially extrapolating beyond the region containing the original observations.

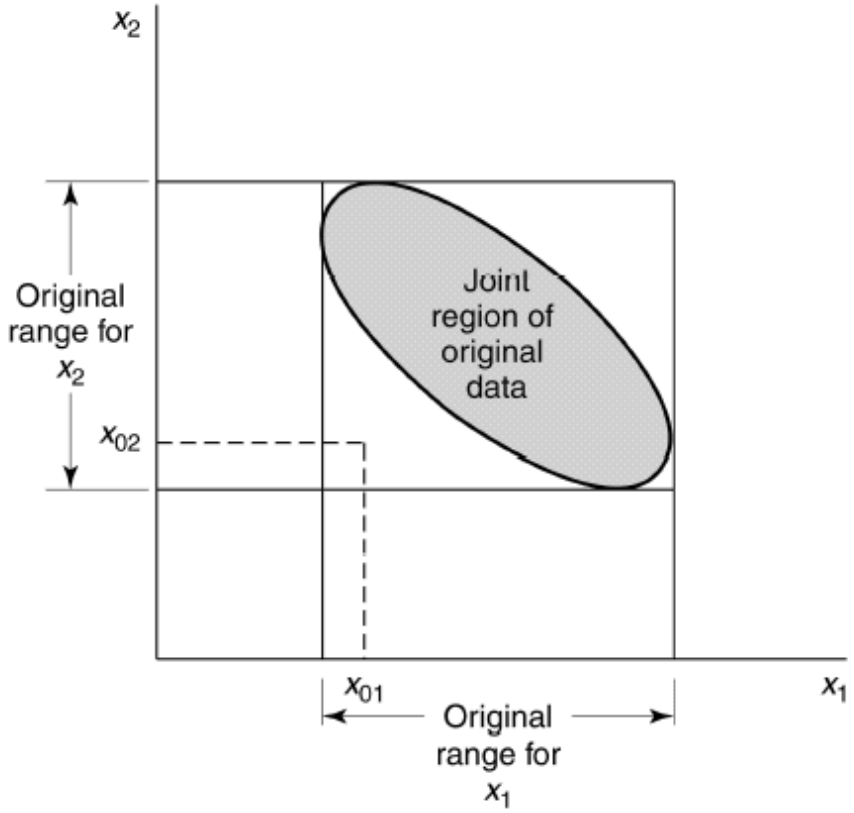


Figure 3.10 An example of extrapolation in multiple regression.

Hidden Extrapolation in Multiple Regression

- We will define the smallest convex set containing all of the original n data points $(x_{i1}, x_{i2}, \ldots x_{ik})$, $i = 1, 2, \ldots, n$, as the **regressor variable hull** RVH.

- If a point $x_{01}, x_{02}, \ldots, x_{0k}$ lies inside or on the boundary of the RVH, then prediction or estimation involves interpolation, while if this point lies outside the RVH, extrapolation is required.

Hidden Extrapolation in Multiple Regression

- Diagonal elements of the matrix **H = X(X'X)$^{-1}$X'** can aid in determining if hidden extrapolation exists:
  - The set of points **x** (not necessarily data points used to fit the model) that satisfy

$$\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x} \leq h_{\max}$$

is an ellipsoid enclosing all points inside the RVH.

Hidden Extrapolation in Multiple Regression

- Let $\mathbf{x}_0$ be a point at which prediction or estimation is of interest. Then

$$h_{00} = \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0$$

- If $h_{00} > h_{max}$ then the point is a point of extrapolation.

Example 3.13

Consider prediction or estimation at:

Point	Symbolism in Figure 3.11	x_{10}	x_{20}	h_{00}
a	□	8	275	0.05346
b	△	20	250	0.58917
c	+	28	500	0.89874
d	×	8	1200	0.86736

TABLE 3.6 Values of h_{ii} for the Delivery Time Data

Observation i	Cases x_{i1}	Distance x_{i2}	h_{ii}
1	7	560	0.10180
2	3	220	0.07070
3	3	340	0.09874
4	4	80	0.08538
5	6	150	0.07501
6	7	330	0.04287
7	2	110	0.08180
8	7	210	0.06373
9	30	1460	$0.49829 = h_{\max}$
10	5	605	0.19630
11	16	688	0.08613
12	10	215	0.11366
13	4	255	0.06113
14	6	462	0.07824
15	9	448	0.04111
16	10	776	0.16594
17	6	200	0.05943
18	7	132	0.09626
19	3	36	0.09645
20	17	770	0.10169
21	10	140	0.16528
22	26	810	0.39158
23	9	450	0.04126
24	8	635	0.12061
25	4	150	0.06664

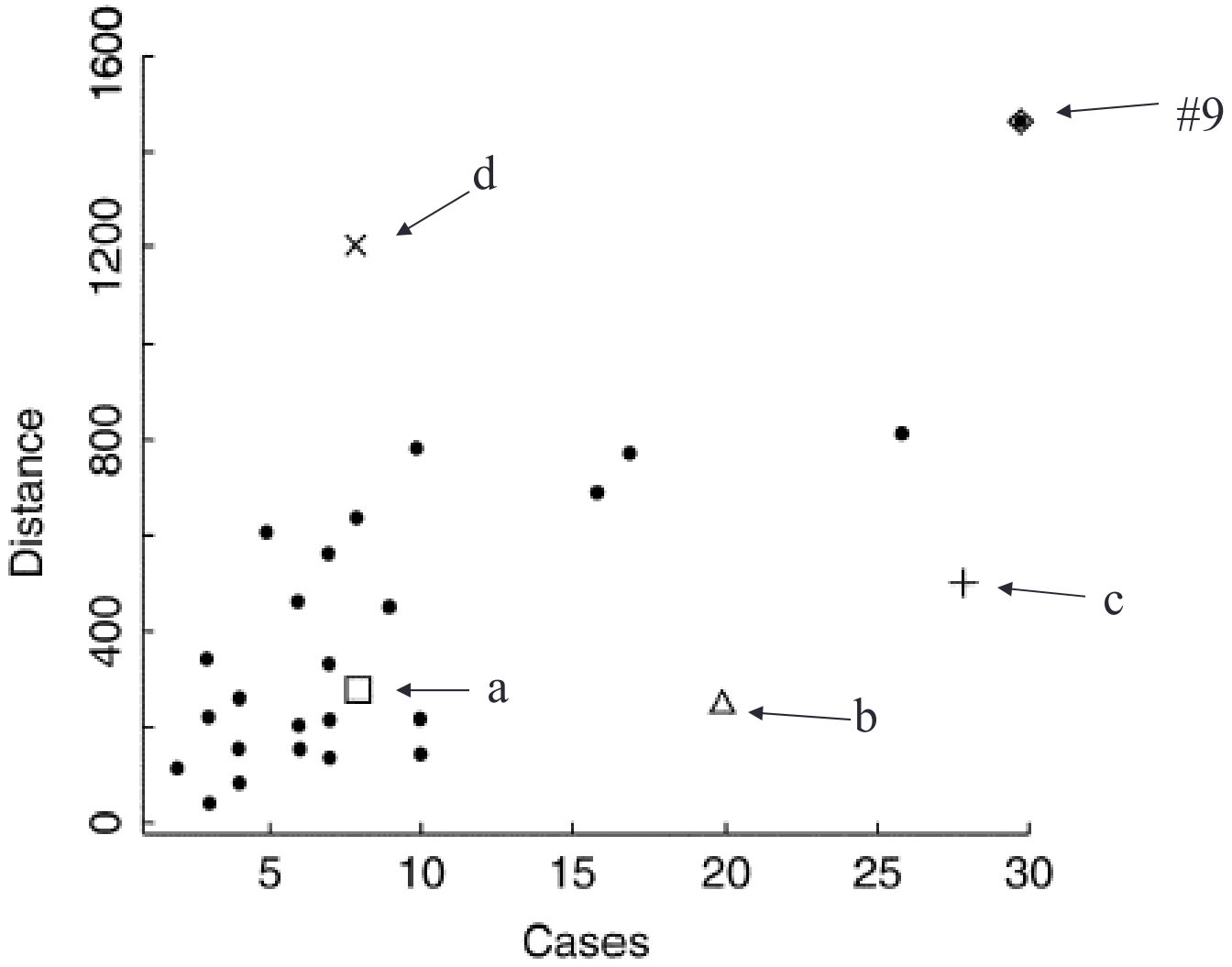


Figure 3.10 Scatterplot of cases and distance for the delivery time data.

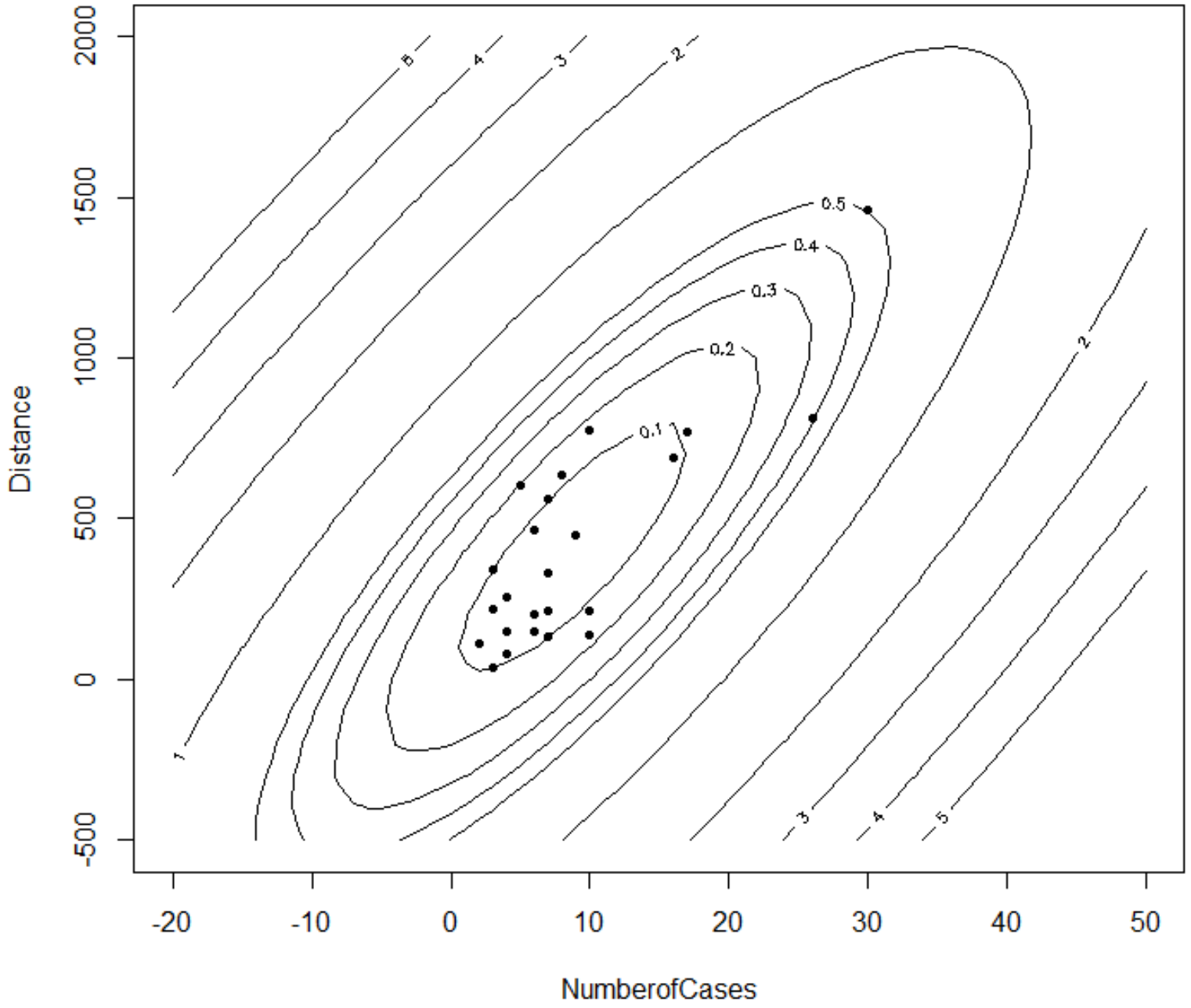
R code

- # obtain the hii for a new observation with NumberofCases=8 and Distance =275
- # first define a vector containing the covariate information. Note that we put 1 in the front of this vector. This 1 is for the intercept. So this 1 never changes.
- x0=c(1,8,275)
- # use this vector to obtain hii for this new observation
- t(x0)%*%solve(t(X)%*%X)%*%x0

- # similarly, we could calculate the hii for new observations with NumberofCases=20, Distance 250 (or NumberofCases=28, Distance 500, or NumberofCases=8, Distance 1200).
- x0=c(1,20,250)
- t(x0)%*%solve(t(X)%*%X)%*%x0
- x0=c(1,28,500)
- t(x0)%*%solve(t(X)%*%X)%*%x0
- x0=c(1,8,1200)
- t(x0)%*%solve(t(X)%*%X)%*%x0
- H=X%*%solve(t(X)%*%X)%*%t(X)
- diag(H)
- max(diag(H))

R code

```
# the program below generates the figure on the next page.
x01_ls=seq(-20,50,1)
x02_ls=seq(-500,2000,100)
h_mat=matrix(NA,length(x01_ls),length(x02_ls))
for (i in 1:length(x01_ls))
{
  for (j in 1:length(x02_ls))
  {
    x0=c(1,x01_ls[i],x02_ls[j])
    h_mat[i,j]=t(x0)%*%solve(t(X)%*%X)%*%x0
  }
}
par(mar=c(4,4,1,1))
contour(x=x01_ls,y=x02_ls,h_mat,levels=c(0.1,0.2,0.3,0.4,0.5,1,2,3,4,
5),xlab="NumberofCases",ylab="Distance")
points(delivery$NumberofCases,delivery$Distance,pch=20)
```


Distance
NumberofCases
2000
1500
1000
500
0
-500
-20
-10
0
10
20
30
40
50
0.5
0.4
0.3
0.2
0.1
1
2
3
4
5
6

Standardized Regression Coefficients

- It is often difficult to directly compare regression coefficients due to possible varying dimensions.
- It may be beneficial to work with dimensionless regression coefficients.
- Dimensionless regression coefficients are often referred to as **standardized regression coefficients.**
- Two common methods of scaling:
  1. Unit normal scaling
  2. Unit length scaling

Standardized Regression Coefficients

- Unit Normal Scaling

  The first approach employs **unit normal scaling** for the regressors and the response variable. That is,

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, k$$

$$y_i^* = \frac{y_i - \bar{y}}{s_y}, \quad i = 1, 2, \ldots, n$$

Standardized Regression Coefficients

- Unit Normal Scaling

where

$$s_j^2 = \frac{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2}{n - 1}$$

$$s_y^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n - 1}$$

Standardized Regression Coefficients

- Unit Normal Scaling
  - All of the scaled regressors and the scaled response have sample mean equal to zero and sample variance equal to 1.
  - The model becomes

$$y_i^* = b_1 z_{i1} + b_2 z_{i2} + \cdots + b_k z_{ik} + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

  - The least squares estimator:

$$\hat{\mathbf{b}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}^*$$

Standardized Regression Coefficients

- Unit Length Scaling

  In unit length scaling:

$$w_{ij} = \frac{x_{ij} - \bar{x}_j}{S_{jj}^{1/2}}, \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, k$$

$$y_i^0 = \frac{y_i - \bar{y}}{SS_T^{1/2}}, \quad i = 1, 2, \ldots, n$$

$$S_{jj} = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$$

Standardized Regression Coefficients

- Unit Length Scaling
  - Each regressor has mean 0 and length

$$\sqrt{\Sigma_{i=1}^{n}(w_{ij}-\overline{w}_j)^2}=1$$

  - The regression model becomes

$$y_i^0 = b_q w_{i1} + b_2 w_{i2} + \cdots + b_k w_{ik} + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

  - The vector of least squares regression coefficients:

$$\hat{\mathbf{b}} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{y}^0$$

Standardized Regression Coefficients

- Unit Length Scaling

  In unit length scaling, the **W'W** matrix is in the form of a **correlation matrix**:

$$\mathbf{W'W} = \begin{bmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1k} \\ r_{12} & 1 & r_{23} & \cdots & r_{2k} \\ r_{13} & r_{23} & 1 & \cdots & r_{34} \\ \vdots & \vdots & \vdots & & \vdots \\ r_{1k} & r_{2k} & r_{3k} & \cdots & 1 \end{bmatrix}$$

  where r_{ij} is the simple correlation between x_i and x_j.

Standardized Regression Coefficients

The regression coefficients $\hat{\mathbf{b}}$ are usually called **standardized regression coefficients**. The relationship between the original and standardized regression coefficients is

$$\hat{\beta}_j = \hat{b}_j \left(\frac{SS_T}{S_{jj}} \right)^{1/2}, \quad j = 1, 2, \ldots, k \tag{3.63}$$

and

$$\hat{\beta}_0 = \bar{y} - \sum_{j=1}^{k} \hat{\beta}_j \bar{x}_j \tag{3.64}$$

Example 3.14

We will find the standardized regression coefficients for the delivery time data in Example 3.1. Since

$$SS_T = 5784.5426 \qquad S_{11} = 1136.5600$$

$$S_{1y} = 2473.3440 \qquad S_{22} = 2{,}537{,}935.0330$$

$$S_{2y} = 108{,}038.6019 \qquad S_{12} = 44{,}266.6800$$

Example 3.14

we find (using the unit length scaling) that

$$r_{12} = \frac{S_{12}}{(S_{11}S_{22})^{1/2}} = \frac{44{,}266.6800}{\sqrt{(1136.5600)(2{,}537{,}935.0303)}} = 0.824215$$

$$r_{1y} = \frac{S_{1y}}{(S_{11}SS_{\mathrm{T}})^{1/2}} = \frac{2473.3440}{\sqrt{(1136.5600)(5784.53426)}} = 0.964615$$

$$r_{2y} = \frac{S_{2y}}{(S_{22}SS_{\mathrm{T}})^{1/2}} = \frac{108{,}038.6019}{\sqrt{(2{,}537{,}935.0330)(5784.5426)}} = 0.891670$$

Example 3.14

the correlation matrix for this problem is

$$\mathbf{W}'\mathbf{W} = \begin{bmatrix} 1 & 0.824215 \\ 0.824215 & 1 \end{bmatrix}$$

The normal equations in terms of the standardized regression coefficients are

$$\begin{bmatrix} 1 & 0.824215 \\ 0.824215 & 1 \end{bmatrix} \begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \end{bmatrix} = \begin{bmatrix} 0.964615 \\ 0.891670 \end{bmatrix}$$

Example 3.14

the standardized regression coefficients are

$$\begin{bmatrix}\hat{b}_1\\ \hat{b}_2\end{bmatrix} = \begin{bmatrix}1 & 0.824215\\ 0.824215 & 1\end{bmatrix}^{-1}\begin{bmatrix}0.964615\\ 0.891670\end{bmatrix}$$

$$= \begin{bmatrix}3.11841 & -2.57023\\ -2.57023 & 3.11841\end{bmatrix}\begin{bmatrix}0.964615\\ 0.891670\end{bmatrix}$$

$$= \begin{bmatrix}0.716267\\ 0.301311\end{bmatrix}$$

The fitted model is

$$\hat{y}^0 = 0.716267w_1 + 0.301311w_2$$

R code

```
# standardized regression coefficient
Delivery
# transform the data using unit normal scaling
delivery_unit_normal=as.data.frame(apply(delivery,2,function(x){(x-mean(x))/sd(x)}))
# redo regression
model1_unit_normal <- lm(DeliveryTime ~ NumberofCases+Distance, data=delivery_unit_normal)
# check the coefficient
model1_unit_normal
# transform the data using unit length scaling
delivery_unit_length=as.data.frame(apply(delivery,2,function(x){(x-mean(x))/sqrt(sum((x-mean(x))^2))}))
# redo regression
model1_unit_length <- lm(DeliveryTime ~ NumberofCases+Distance, data=delivery_unit_length)
# check the coefficient
model1_unit_length
# compare the coefficient from unit normal scaling and unit length scaling. They are the same.
model1_unit_normal

summary(model1_unit_normal)
summary(model1_unit_length)

# obtain correlation
cor(delivery)
# compare correlation with W'W, where W is the transformed data using unit length scaling.
t(as.matrix(delivery_unit_length))%*%as.matrix(delivery_unit_length)
```

Multicollinearity

- A serious problem that may dramatically impact the usefulness of a regression model is **multicollinearity**, or **near-linear dependence** among the regression variables.
- Multicollinearity implies near-linear dependence among the regressors. The regressors are the columns of the **X** matrix, so clearly an **exact linear dependence** would result in a **singular X'X**.
- The presence of multicollinearity can dramatically impact the ability to estimate regression coefficients and other uses of the regression model.

Multicollinearity

- Issues with Multicollinearity
  - Adding or deleting a predictor variable changes the regression coefficients
  - Estimated standard deviation of the regression coefficients become large when the predictor variables are highly correlated
    - (IMPORTANT!!! A large wiggle room in 3D or higher dimesnion)
  - Estimated regression coefficients individually may not be significant but a statistical relationship exists for the set of predictor variables and the response variable

Multicollinearity

- Informal Diagnostics
  - Large changes in the estimated regression coefficients when a predictor variable is added or deleted
  - Nonsignificant results in individual tests of regression coefficients for important predictors
  - Regression coefficients with a sign that is opposite of that expected
  - Large coefficients of simple correlation between pairs of predictor variables
  - Wide confidence intervals for the regression coefficients for important predictors

Multicollinearity

- The main diagonal elements of the inverse of the **X'X** matrix in correlation form $(\mathbf{W'W})^{-1}$ are often called **variance inflation factors** VIFs, and they are an important multicollinearity diagnostic.
- For the soft drink delivery data,

$$\mathrm{VIF}_1 = \mathrm{VIF}_2 = 3.11841$$

Multicollinearity

- The variance inflation factors can also be written as:

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2}$$

where R^2_j is the coefficient of multiple determination obtained from regressing x_j on the other regressor variables.

- If x_j is highly correlated with any other regressor variable, then R^2_j will be large.

Multicollinearity

- Variance Inflation Factor
  - Diagnostic Uses
    - Largest *VIF* used as an indicator of multicollinearity
      - Guideline: Greater than 10
    - Mean *VIF* used as an indicator of multicollinearity
      - Guideline: Greater than 1

R code

- # obtain VIF
- # first load/install R package car
- library(car)
- # obtain VIF for each variable
- vif(model1)

Other Reasons Regression Coefficients Have The Wrong Sign

- Regression coefficients may have the wrong sign for the following reasons:
  1. The range of some of the regressors is too small.
  2. Important regressors have not been included in the model.
  3. Computational errors have been made.

Variance Inflation

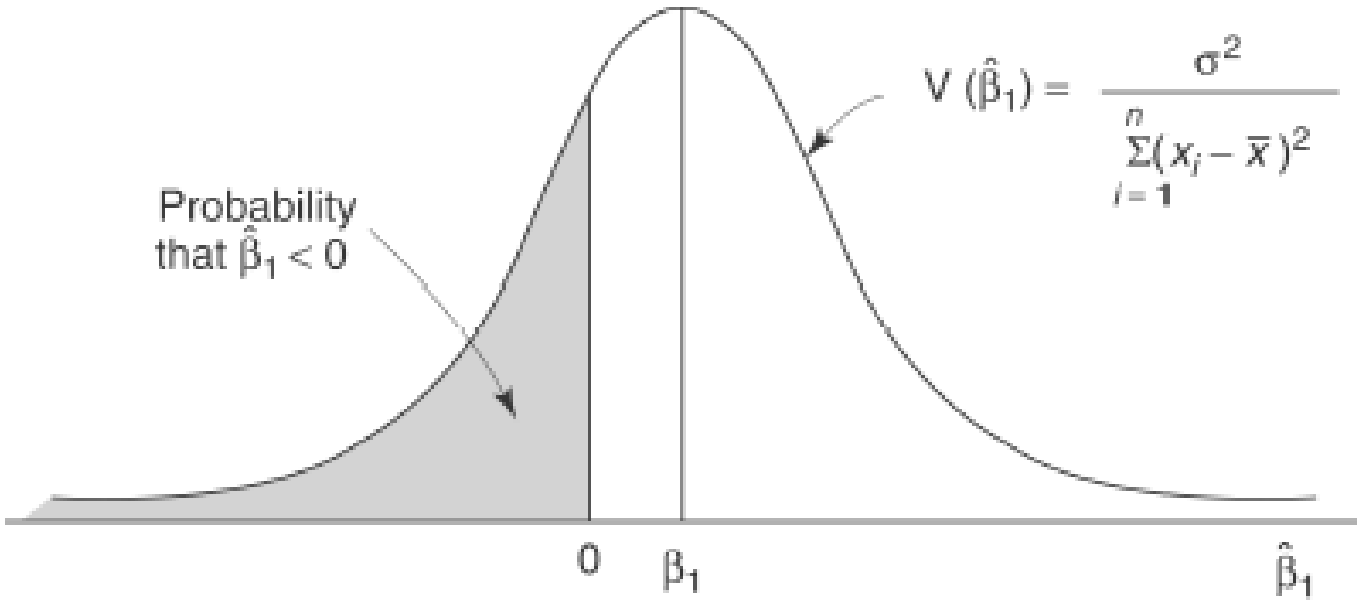


Figure 3.13 Sampling distribution of $\hat{\beta}_1$.