Chapter 3

Multiple Linear Regression

A regression model involving two or more predictors is called a multiple regression model. In this chapter we consider multiple linear regression problems involving modeling the relationship between a response variable Y and two or more predictors $X_1, \ldots, X_k$ ($k \geq 2$). This is a generalization of simple linear regression.

3.1 Multiple Linear Regression Models

Data

Response variable	Predictors		
Y	X_1	$\ldots$	X_k
y_1	x_{11}	$\ldots$	x_{1k}
y_2	x_{21}	$\ldots$	x_{2k}
$\vdots$	$\vdots$		$\vdots$
y_n	x_{n1}	$\ldots$	x_{nk}

Multiple linear regression model

In a multiple linear regression model

$$\mathrm{E}[Y|X_1 = x_1, \ldots, X_k = x_k] = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

Thus, a multiple linear regression model with k predictors is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon,$$

where ϵ is the random error with mean 0 and variance σ^2. The parameter β_0 is the mean of Y when $x_1 = \cdots = x_k = 0$. Recall that if the range of the predictors does not include zero, then β_0 has no practical interpretation. The parameter β_1 indicates the expected change in response per unit change in X_1 when $X_2, \ldots, X_k$ are held constant. The parameters $\beta_2, \ldots, \beta_k$ are similarly interpreted.

Examples

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$
- $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$ (Polynomial regression model)
  $\Rightarrow X_1 = X,\ X_2 = X^2$
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \epsilon$ (Regression model with interaction)
  $\Rightarrow X_1 = X_1,\ X_2 = X_2,\ X_3 = X_1 X_2$

Note: The relationship between a response variable Y and k predictors $X_1, \ldots, X_k$ is described as a surface in the k-dimensional space.

The regression model in a data setting is written as

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

where $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ are unknown constants, $\mathrm{E}[\epsilon_i] = 0,\ \mathrm{Var}(\epsilon_i) = \sigma^2$ and $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$, The errors are assumed to be independently and normally distributed with a mean of 0 and variance σ^2; i.e.,

$$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2), \quad i = 1, \ldots, n.$$

Additionally we assume that the columns of

$$\begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$

are linearly independent.

3.2 Estimation and Inference in Multiple Linear Regression

3.2.1 Least-squares estimation of the regression coefficients

The least-squares estimates of $\beta_0, \beta_1, \ldots, \beta_k$, say $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$, are defined by

$$(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k) \equiv \mathrm{argmin}_{b_0, b_1, \ldots, b_k} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - \cdots - b_k x_{ik})^2.$$

Let

$$S(b_0, b_1, \ldots, b_k) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - \cdots - b_k x_{ik})^2.$$

Then, $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are given by

$$\hat{\beta}_0 = b_0, \ \hat{\beta}_1 = b_1, \ldots, \hat{\beta}_k = b_k,$$

where $b_0, b_1, \ldots, b_k$ must satisfy

$$\begin{cases} \frac{\partial S(b_0,b_1,..,b_k)}{\partial b_0} &= -2\sum_{i=1}^n (y_i - b_0 - b_1 x_{i1} - \cdots - b_k x_{ik}) = 0 \\ \frac{\partial S(b_0,b_1,..,b_k)}{\partial b_1} &= -2\sum_{i=1}^n x_{i1}(y_i - b_0 - b_1 x_{i1} - \cdots - b_k x_{ik}) = 0 \\ & \vdots \\ \frac{\partial S(b_0,b_1,..,b_k)}{\partial b_k} &= -2\sum_{i=1}^n x_{ik}(y_i - b_0 - b_1 x_{i1} - \cdots - b_k x_{ik}) = 0. \end{cases}$$

These equations are written as

$$\begin{cases} nb_0 + \sum_{i=1}^n x_{i1} b_1 + \sum_{i=1}^n x_{i2} b_2 + \cdots + \sum_{i=1}^n x_{ik} b_k = \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i1} b_0 + \sum_{i=1}^n x_{i1}^2 b_1 + \sum_{i=1}^n x_{i1} x_{i2} b_2 + \cdots + \sum_{i=1}^n x_{i1} x_{ik} b_k = \sum_{i=1}^n x_{i1} y_i \\ \sum_{i=1}^n x_{i2} b_0 + \sum_{i=1}^n x_{i1} x_{i2} b_1 + \sum_{i=1}^n x_{i2}^2 b_2 + \cdots + \sum_{i=1}^n x_{i2} x_{ik} b_k = \sum_{i=1}^n x_{i2} y_i \\ \vdots \\ \sum_{i=1}^n x_{ik} b_0 + \sum_{i=1}^n x_{i1} x_{ik} b_1 + \sum_{i=1}^n x_{i2} x_{ik} b_2 + \cdots + \sum_{i=1}^n x_{ik}^2 b_k = \sum_{i=1}^n x_{ik} y_i. \end{cases}$$

Notice that there are $p = k + 1$ normal equations, one for each of the unknown regression coefficients. Here, p is the number of parameters and k is the number of predictors.

Matrix formulation of the multiple linear regression

Note: If you don't have matrix algebra background, you must read Handout and/or Appendix C.2.

A convenient way to study the multiple linear regression is to use matrix and vector notation.

- **Model**

  Recall the multiple regression model in a data setting

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

  and observe that

  $$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Define the $n \times 1$ vector $\mathbf{y}$ and the $n \times p$ matrix $\mathbf{X}$ by

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

Also define the $p \times 1$ vector $\boldsymbol{\beta}$ of unknown regression parameters and the $n \times 1$ vector $\boldsymbol{\epsilon}$ of random errors by

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Then, we can write the multiple linear regression model in matrix notation as

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

- **Normal equation and least-squares estimates**

  Now recall the normal equations obtained from the least-squares estimation.

$$\begin{cases} nb_0 + \sum_{i=1}^n x_{i1} b_1 + \sum_{i=1}^n x_{i2} b_2 + \cdots + \sum_{i=1}^n x_{ik} b_k = \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i1} b_0 + \sum_{i=1}^n x_{i1}^2 b_1 + \sum_{i=1}^n x_{i1} x_{i2} b_2 + \cdots + \sum_{i=1}^n x_{i1} x_{ik} b_k = \sum_{i=1}^n x_{i1} y_i \\ \sum_{i=1}^n x_{i2} b_0 + \sum_{i=1}^n x_{i1} x_{i2} b_1 + \sum_{i=1}^n x_{i2}^2 b_2 + \cdots + \sum_{i=1}^n x_{i2} x_{ik} b_k = \sum_{i=1}^n x_{i2} y_i \\ \vdots \\ \sum_{i=1}^n x_{ik} b_0 + \sum_{i=1}^n x_{i1} x_{ik} b_1 + \sum_{i=1}^n x_{i2} x_{ik} b_2 + \cdots + \sum_{i=1}^n x_{ik}^2 b_k = \sum_{i=1}^n x_{ik} y_i. \end{cases}$$

  We rewrite these equations as

$$\begin{pmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} & \cdots & \sum_{i=1}^n x_{ik} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1} x_{i2} & \cdots & \sum_{i=1}^n x_{i1} x_{ik} \\ \sum_{i=1}^n x_{i2} & \sum_{i=1}^n x_{i1} x_{i2} & \sum_{i=1}^n x_{i2}^2 & \cdots & \sum_{i=1}^n x_{i2} x_{ik} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{ik} & \sum_{i=1}^n x_{i1} x_{ik} & \sum_{i=1}^n x_{i2} x_{ik} & \cdots & \sum_{i=1}^n x_{ik}^2 \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_k \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i1} y_i \\ \sum_{i=1}^n x_{i2} y_i \\ \vdots \\ \sum_{i=1}^n x_{ik} y_i \end{pmatrix}.$$

  Thus, the matrix form of the normal equations is

$$\mathbf{X}^T \mathbf{X} \hat{\boldsymbol{\beta}} = \mathbf{X}^T \mathbf{y}.$$

  Since we assumed that the matrix $\mathbf{X}$ has full rank, the least-squares estimate of $\boldsymbol{\beta}$ is

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

The fitted or predicted values are given by

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

The $n \times n$ matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is called the hat matrix. The residuals are given by

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

where $\mathbf{I}$ is the $(n \times n)$ identity matrix.

A geometrical interpretation of least squares estimates

Notes:

1) The multiple regression model can be written as

$$y_i = \underbrace{\beta_0 + (\beta_1\bar{x}_1 + \cdots + \beta_k\bar{x}_k)}_{\beta_0^*} + \beta_1(x_{i1} - \bar{x}_1) + \cdots + \beta_k(x_{ik} - \bar{x}_k) + \epsilon_i,$$

where $\bar{x}_j = \frac{1}{n}\sum_{i=1}^n x_{ij},\ j = 1, \ldots, k$. Define the centered matrix of $\mathbf{X}$ by

$$\mathbf{X}_c = \begin{pmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1k} - \bar{x}_k \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2k} - \bar{x}_k \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{nk} - \bar{x}_k \end{pmatrix}$$

and let $\boldsymbol{\beta}_* = (\beta_1, \beta_2, \ldots, \beta_k)^T$. Thus, we can rewrite the multiple linear regression model in matrix notation as

$$\mathbf{y} = \mathbf{1}\beta_0^* + \mathbf{X}_c\boldsymbol{\beta}_* + \boldsymbol{\epsilon},$$

where $\mathbf{1}$ is a $n \times 1$ vector consisting of all 1's. Note that $\mathbf{1}^T\mathbf{X}_c = \mathbf{0}$.

2) The matrix form of the normal equations is given by

$$\begin{cases} \hat{\beta}_0^* = \bar{y} \;\Rightarrow\; \hat{\beta}_0 = \bar{y} - (\hat{\beta}_1\bar{x}_1 + \cdots + \hat{\beta}_k\bar{x}_k) \\ \mathbf{X}_c^T\mathbf{X}_c\hat{\boldsymbol{\beta}}_* = \mathbf{X}_c^T\mathbf{y}. \end{cases}$$

3) The fitted values are given by

$$\hat{\mathbf{y}} = \mathbf{1}\hat{\beta}_0 + \mathbf{X}_c\hat{\boldsymbol{\beta}}_* = \mathbf{1}(\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T\mathbf{y} + \mathbf{X}_c(\mathbf{X}_c^T\mathbf{X}_c)^{-1}\mathbf{X}_c^T\mathbf{y}.$$

- **Sums of squares**
  - Total corrected sum of squares of the observations, $SS_T = \sum_{i=1}^n (y_i - \bar{y})^2$,
  - Residual sum of squares, $SS_{Res} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$,
  - Regression sum of squares, $SS_R = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$.

  Notice that $SS_T = SS_{Reg} + SS_R$. Then, the matrix form of these sums of squares is
  - $SS_T = \sum_{i=1}^n (y_i - \bar{y})^2 = \mathbf{y}^T(\mathbf{I} - \mathbf{1}(\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T)\mathbf{y}$,
  - $SS_{Res} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \mathbf{y}^T(\mathbf{I} - \mathbf{H})\mathbf{y} = \mathbf{y}^T(\mathbf{I} - \mathbf{1}(\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T - \mathbf{X}_c(\mathbf{X}_c^T\mathbf{X}_c)^{-1}\mathbf{X}_c^T)\mathbf{y}$,
  - $SS_R = SST - SS_{Res} = \mathbf{y}^T(\mathbf{H} - \mathbf{1}(\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T)\mathbf{y} = \mathbf{y}^T\mathbf{X}_c(\mathbf{X}_c^T\mathbf{X}_c)^{-1}\mathbf{X}_c^T\mathbf{y}$.

Estimation of σ^2

It can be shown that

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

is an unbiased estimate of σ^2.

Note: To show this, you may use the fact that

$$\mathrm{E}[\mathbf{y}^T\mathbf{A}\mathbf{y}] = \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu} + \mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}),$$

where $\boldsymbol{\mu} = \mathrm{E}[\mathbf{y}]$ and $\boldsymbol{\Sigma} = \mathrm{Var}(\mathbf{y})$ is the variance-covariance matrix of a random vector $\mathbf{y}$.

Example

Revisit the delivery time data in Ex 3.1 in textbook.

SAS program:

```
option ls=75 ps=80;
title 'Delivery Time Data';
proc reg data=delivery;
model Time = Cases Distance / clb clm cli;
```

```
run; quit;
proc g3d;
scatter Cases*Distance=Time;
run;
```

Output:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	5550.81092	2775.40546	261.24	<.0001
Error	22	233.73168	10.62417		
Corrected Total	24	5784.54260			

Root MSE	3.25947	R-Square	0.9596
Dependent Mean	22.38400	Adj R-Sq	0.9559
Coeff Var	14.56162		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.34123	1.09673	2.13	0.0442
Cases	1	1.61591	0.17073	9.46	<.0001
Distance	1	0.01438	0.00361	3.98	0.0006

Output Statistics

Obs	Dependent Variable	Predicted Value	Residual
1	16.6800	21.7081	-5.0281
2	11.5000	10.3536	1.1464
3	12.0300	12.0798	-0.0498
4	14.8800	9.9556	4.9244
5	13.7500	14.1944	-0.4444
6	18.1100	18.3996	-0.2896
7	8.0000	7.1554	0.8446
8	17.8300	16.6734	1.1566
9	79.2400	71.8203	7.4197
10	21.5000	19.1236	2.3764
11	40.3300	38.0925	2.2375
12	21.0000	21.5930	-0.5930
13	13.5000	12.4730	1.0270

14	19.7500	18.6825	1.0675
15	24.0000	23.3288	0.6712
16	29.0000	29.6629	-0.6629
17	15.3500	14.9136	0.4364
18	19.0000	15.5514	3.4486
19	9.5000	7.7068	1.7932
20	35.1000	40.8880	-5.7880
21	17.9000	20.5142	-2.6142
22	52.3200	56.0065	-3.6865
23	18.7500	23.3576	-4.6076
24	19.8300	24.4029	-4.5729
25	10.7500	10.9626	-0.2126

Sum of Residuals	0
Sum of Squared Residuals	233.73168
Predicted Residual SS (PRESS)	459.03931

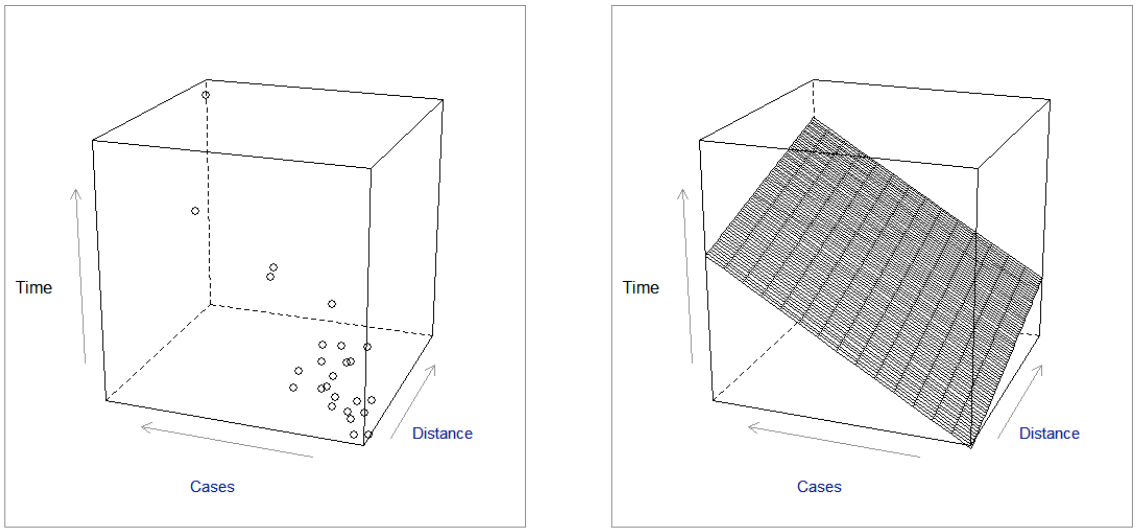

Figure 3.1: Three dimensional scatterplot of the delivery time data and the fitted regression plane.

Figure 3.1 shows a 3-dimensional scatterplot of the delivery time data with the least-squares regression plane. The least-squares fit is

$$\hat{Y} = 2.3412 + 1.6159X_1 + .0144X_2$$

or

$$\widehat{\text{Time}} = 2.3412 + 1.6159 * \text{Cases} + .0144 * \text{Distance}.$$

Also, $\hat{\sigma}^2 = MS_{Res} = 10.62417$.

Properties of least-squares estimates

Consider the regression model

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

with $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$.

Recall that the least-squares estimators are given by

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

- $\mathrm{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$; i.e., $\mathrm{E}[\hat{\beta}_i] = \beta_i,\ i = 0, 1, \ldots, k$,
- $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$,
- $\mathrm{Cov}(\hat{\boldsymbol{\beta}}_*, \bar{y}) = \mathbf{0}$; i.e., $\mathrm{Cov}(\hat{\beta}_i, \bar{y}) = 0,\ i = 1, \ldots, k$,
- The least-squares estimate $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimate of $\boldsymbol{\beta}$ (Gauss-Markov theorem, Read Appendix C.4),
- $\mathrm{E}[\hat{\sigma}^2] = \sigma^2$.

Note: Under the normality assumption of the errors, $\hat{\boldsymbol{\beta}}$ is identical to the maximum-likelihood estimate of $\boldsymbol{\beta}$ as in simple linear regression model.

Special case: Simple linear regression

Recall the multiple linear regression model

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

For simple linear regression the matrix $\mathbf{X}$ is given by

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

Thus,

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}$$

and

$$\mathbf{X}^T\mathbf{y} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix}.$$

Taking the inverse of $\mathbf{X}^T\mathbf{X}$ gives

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix} = \frac{1}{S_{xx}} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}.$$

So, it can be easily shown that

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \begin{pmatrix} \bar{y} - \frac{S_{xy}}{S_{xx}}\bar{x} \\ \frac{S_{xy}}{S_{xx}} \end{pmatrix}$$

and

$$\mathrm{Var}(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right), \quad \mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

which agree to the results in Chapter 2.

3.2.2 Inference in multiple linear regression

Inference about individual regression coefficients

Under the model assumptions, it can be shown that

$$\frac{\hat{\beta}_i - \beta_i}{\mathrm{se}(\hat{\beta}_i)} \sim t_{n-p}, \quad i = 0, 1, \ldots, k,$$

where $\mathrm{se}(\hat{\beta}_i) = \sqrt{MS_{Res}C_{ii}}$ with C_{ii} the ith diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$.

We can use this to construct confidence intervals for β_i and to test hypothesis $H_{0,i} : \beta_i = \beta_{i0}$ versus $H_{1,i} : \beta_i \neq \beta_{i0}$.

- A $100(1-\alpha)\%$ CI for β_i is
$$\hat{\beta}_i \pm t_{\alpha/2,n-p}\mathrm{se}(\hat{\beta}_i).$$
- The test statistic for the hypothesis $H_{0,i} : \beta_i = \beta_{i0}$ versus $H_{1,i} : \beta_i \neq \beta_{i0}$ is
$$t_{0,i} = \frac{\hat{\beta}_i - \beta_{i0}}{\mathrm{se}(\hat{\beta}_i)}.$$
The null hypothesis is rejected if $|t_{0,i}| > t_{\alpha/2,n-p}$. Note that this is **a test of the contribution of X_i given the other predictors in the model**.

Example (Delivery time data)

Suppose we wish to assess the value of the predictor X_2 (distance) given that the predictor X_1 (cases) is in the model. In addition, we wish to construct 95% CI for the parameter β_2.

SAS Output

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	2.34123	1.09673	2.13	0.0442	0.06675	4.61571
Cases	1	1.61591	0.17073	9.46	<.0001	1.26182	1.96999
Distance	1	0.01438	0.00361	3.98	0.0006	0.00689	0.02188

Estimation of mean response and prediction of new observations

Let $\mathbf{x}_i^T$ denote the ith row of the matrix $\mathbf{X}$. Then,

$$\mathbf{x}_i^T = (1, x_{i1}, x_{i2}, \ldots, x_{ik})$$

is a $1 \times p$ row vector which allows us to write

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} = \mathbf{x}_i^T \boldsymbol{\beta}.$$

- Confidence interval for the mean response

  Let

  $$\mu_0 = \mathrm{E}[Y|X_1 = x_{01}, \ldots, X_k = x_{0k}] = \beta_0 + \beta_1 x_{01} + \cdots + \beta_k x_{0k}.$$

  Then, we can write μ_0 as

  $$\mu_0 = \mathbf{x}_0^T \boldsymbol{\beta},$$

  where $\mathbf{x}_0 = (1, x_{01}, \ldots, x_{0k})^T$. Then, the point estimate is

  $$\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \cdots + \hat{\beta}_k x_{0k} = \mathbf{x}_0^T \hat{\boldsymbol{\beta}}$$

  and the variance of $\hat{\mu}_0$ is

  $$\mathrm{Var}(\hat{\mu}_0) = \sigma^2 \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0.$$

  Thus, under the model assumptions, it can be shown that

  $$\frac{\hat{\mu}_0 - \mu_0}{\sqrt{\sigma^2 \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0}} \sim N(0, 1)$$

and replacing σ^2 by MS_{Res} result in

$$\frac{\hat{\mu}_0 - \mu_0}{\text{se}(\hat{\mu}_0)} \sim t_{n-p}$$

where

$$\text{se}(\hat{\mu}_0) = \sqrt{MS_{Res}\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0}.$$

Hence a $100(1-\alpha)\%$ confidence interval for $\text{E}[Y|X_1 = x_{01}, \ldots, X_k = x_{0k}]$ is

$$\hat{\mu}_0 \pm t_{\alpha/2,n-p}\text{se}(\hat{\mu}_0).$$

Example (Delivery time data)

Suppose that the soft drink bottler 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.

SAS Program

```
data new;
input Time Cases Distance;
datalines;
. 8  275
; run;
data delivery2;
set delivery new;
run;
proc reg data=delivery2;
model Time = Cases Distance / clm cli;
run; quit;
```

Output

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		Residual
1	16.6800	21.7081	1.0400	19.5513	23.8649	-5.0281
2	11.5000	10.3536	0.8667	8.5562	12.1510	1.1464
3	12.0300	12.0798	1.0242	9.9557	14.2038	-0.0498
4	14.8800	9.9556	0.9524	7.9805	11.9308	4.9244
5	13.7500	14.1944	0.8927	12.3430	16.0458	-0.4444
6	18.1100	18.3996	0.6749	17.0000	19.7991	-0.2896

7	8.0000	7.1554	0.9322	5.2221	9.0887	0.8446
8	17.8300	16.6734	0.8228	14.9670	18.3798	1.1566
9	79.2400	71.8203	2.3009	67.0486	76.5920	7.4197
10	21.5000	19.1236	1.4441	16.1287	22.1185	2.3764
11	40.3300	38.0925	0.9566	36.1086	40.0764	2.2375
12	21.0000	21.5930	1.0989	19.3141	23.8719	-0.5930
13	13.5000	12.4730	0.8059	10.8018	14.1442	1.0270
14	19.7500	18.6825	0.9117	16.7916	20.5733	1.0675
15	24.0000	23.3288	0.6609	21.9582	24.6994	0.6712
16	29.0000	29.6629	1.3278	26.9093	32.4166	-0.6629
17	15.3500	14.9136	0.7946	13.2657	16.5616	0.4364
18	19.0000	15.5514	1.0113	13.4541	17.6486	3.4486
19	9.5000	7.7068	1.0123	5.6075	9.8061	1.7932
20	35.1000	40.8880	1.0394	38.7324	43.0435	-5.7880
21	17.9000	20.5142	1.3251	17.7661	23.2623	-2.6142
22	52.3200	56.0065	2.0396	51.7766	60.2365	-3.6865
23	18.7500	23.3576	0.6621	21.9845	24.7306	-4.6076
24	19.8300	24.4029	1.1320	22.0553	26.7504	-4.5729
25	10.7500	10.9626	0.8414	9.2175	12.7076	-0.2126
26	.	19.2243	0.7572	17.6539	20.7947	.

When $x_{10} = 8$ and $x_{20} = 275$,

$$\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{10} + \hat{\beta}_2 x_{20} = 2.34123 + (1.61591)(8) + (.01438)(275) = 19.2243 \text{ min}$$

and

$$\text{se}(\hat{\mu}_0) = \sqrt{MS_{Res}\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0} = .7572$$

with $\mathbf{x}_0 = (1, 8, 275)^T$. The 95% CI for $\mu_0 = \text{E}[Y|X_1 = 8, X_2 = 275] = \beta_0 + 8\beta_1 + 275\beta_2$ is

$$\hat{\mu}_0 \pm t_{.025,22}\text{se}(\hat{\mu}_0) = 19.2243 \pm (2.074)(.7572) = (17.6539, 20.7947).$$

- Prediction interval for new observations

  A new observation Y at $X_1 = x_{01}, \ldots, X_k = x_{0k}$ is given by

  $$Y_0 = \beta_0 + \beta_1 x_{01} + \cdots + \beta_k x_{0k} + \epsilon_0 = \mu_0 + \epsilon_0,$$

  where a random error ϵ_0 is assumed to be normally distributed with mean 0 and variance σ^2 and to be independent of the errors $\epsilon_1, \ldots, \epsilon_n$.

  The point estimate of Y_0 is

  $$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \cdots + \hat{\beta}_k x_{0k} = \mathbf{x}_0^T\hat{\boldsymbol{\beta}}.$$

  Now,observe that $\text{E}[Y_0 - \hat{y}_0] = 0$ and

  $$\text{Var}(Y_0 - \hat{y}_0) = \sigma^2\left(1 + \mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0\right).$$

Replacing σ^2 by MS_{Res} result in

$$\frac{Y_0 - \hat{y}_0}{\text{se}(Y_0 - \hat{y}_0)} \sim t_{n-p},$$

where

$$\text{se}(Y_0 - \hat{y}_0) = \sqrt{MS_{Res}\,(1 + \mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0)}.$$

Thus, a $100(1-\alpha)\%$ prediction interval for Y_0 is

$$\hat{y}_0 \pm t_{\alpha/2,n-p}\text{se}(Y_0 - \hat{y}_0).$$

Example (Delivery time data)

Suppose that the soft drink bottler would like to construct a 95% PI on the delivery time for an outlet requiring $x_1 = 8$ cases and where the distance $x_2 = 275$ feet.

Output

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Predict		Residual
1	16.6800	21.7081	1.0400	14.6126	28.8036	-5.0281
2	11.5000	10.3536	0.8667	3.3590	17.3482	1.1464
3	12.0300	12.0798	1.0242	4.9942	19.1654	-0.0498
4	14.8800	9.9556	0.9524	2.9133	16.9980	4.9244
5	13.7500	14.1944	0.8927	7.1857	21.2031	-0.4444
6	18.1100	18.3996	0.6749	11.4965	25.3027	-0.2896
7	8.0000	7.1554	0.9322	0.1246	14.1861	0.8446
8	17.8300	16.6734	0.8228	9.7016	23.6452	1.1566
9	79.2400	71.8203	2.3009	63.5461	80.0945	7.4197
10	21.5000	19.1236	1.4441	11.7301	26.5171	2.3764
11	40.3300	38.0925	0.9566	31.0477	45.1373	2.2375
12	21.0000	21.5930	1.0989	14.4595	28.7266	-0.5930
13	13.5000	12.4730	0.8059	5.5097	19.4363	1.0270
14	19.7500	18.6825	0.9117	11.6633	25.7017	1.0675
15	24.0000	23.3288	0.6609	16.4315	30.2261	0.6712
16	29.0000	29.6629	1.3278	22.3639	36.9620	-0.6629
17	15.3500	14.9136	0.7946	7.9559	21.8713	0.4364
18	19.0000	15.5514	1.0113	8.4738	22.6290	3.4486
19	9.5000	7.7068	1.0123	0.6286	14.7850	1.7932
20	35.1000	40.8880	1.0394	33.7929	47.9831	-5.7880
21	17.9000	20.5142	1.3251	13.2172	27.8112	-2.6142
22	52.3200	56.0065	2.0396	48.0324	63.9807	-3.6865
23	18.7500	23.3576	0.6621	16.4598	30.2553	-4.6076
24	19.8300	24.4029	1.1320	17.2471	31.5586	-4.5729
25	10.7500	10.9626	0.8414	3.9812	17.9439	-0.2126
26	.	19.2243	0.7572	12.2846	26.1641	.

The 95% PI for $Y_0 = \beta_0 + 8\beta_1 + 275\beta_2 + \epsilon_0$ is

$$\hat{y}_0 \pm t_{.025,22}\sqrt{MS_{Res}\,(1 + \mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0)} = 19.2243 \pm 2.074\sqrt{10.614 + .7572^2} = (12.2846, 26.1641).$$

Therefore, an outlet requiring $x_1 = 8$ cases and where the distance $x_2 = 275$ feet could reasonably be expected to have the delivery time between 12.2846 min and 26.1641 min.

Tests of hypothesis in the regression model

- **Analysis of variance to test for significance of regression**

  The test for significance of regression is to determine if there is a linear relationship between Y and any of the predictors $X_1, \ldots, X_k$. This procedure is often referred as an overall test of model adequacy. We wish to test

  $$H_0 : \beta_1 = \cdots = \beta_k = 0 \quad \text{versus} \quad H_1 : \text{at least one of } \beta_i \neq 0.$$

  Rejection of this null hypothesis implies that at least one of the predictors $X_1, \ldots, X_k$ contributes significantly to the model.

  Now, recall the following terminology:

  - Total corrected sum of squares of the observations, $SS_T = \sum_{i=1}^n (y_i - \bar{y})^2$,
  - Residual sum of squares, $SS_{Res} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$,
  - Regression sum of squares, $SS_R = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$.

  Note: SS_{Res} and SS_R are independent,

  $$\mathrm{E}[MS_{Res}] = \sigma^2, \quad \mathrm{E}[MS_R] = \sigma^2 + \frac{\boldsymbol{\beta}_*^T \mathbf{X}_c^T \mathbf{X}_c \boldsymbol{\beta}_*}{k}.$$

  To test $H_0 : \beta_1 = \cdots = \beta_k = 0$ against $H_1 :$ at least one of $\beta_i \neq 0$ we use the test statistic

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

  and under the assumption that $\epsilon_1, \ldots, \epsilon_n$ are independent and normally distributed, it can be shown that

  $$F_0 \sim F_{k,n-k-1}$$

  when H_0 is true since

  i. $\frac{SS_R}{\sigma^2} \sim \chi^2_k$ under H_0,

  ii. $\frac{SS_{Res}}{\sigma^2} \sim \chi^2_{n-k-1}$,

iii. SS_R and SS_{Res} are independent.

The null hypothesis is rejected at level α if $F_0 > F_{\alpha,k,n-k-1}$ or $P - value < \alpha$.

Source of variation	Degree of freedom (df)	Sum of squares (SS)	Mean squares (MS)	F
Regression	k	SS_R	$MS_R = SS_R/k$	$F_0 = \frac{MS_R}{MS_{Res}}$
Residual	$n-k-1(=n-p)$	SS_{Res}	$MS_{Res} = SS_{Res}/(n-k-1)$	
Total	$n-1$	SS_T		

Table 3.1: Analysis of variance for testing significance of regression in multiple regression

Notes:

1. R^2, the coefficient of determination of the regression line, is defined as the proportion of the total sample variability in the observations explained by the regression model, that is,
$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T}.$$
Adding irrelevant predictors to the regression equation often increases R^2 since R^2 increases whenever an additional predictor is included. So, it is difficult to judge whether an increase in R^2 is really telling us anything important. To compensate for this one can define an adjusted R^2 defined as
$$R^2_{adj} = 1 - \frac{SS_{Res}/(n-p)}{SS_T/(n-1)}.$$
The adjusted R^2 is a version that includes a penalty for unnecessary predictor variables. Since $MS_{Res} = \frac{SS_{Res}}{n-p}$ is the residual mean square and $SS_T/(n-1)$ is constant regardless of how many variables are in the model, R^2_{adj} will only increase on adding a variable to the model if the addition of the variable reduces the residual mean square. Thus, when comparing models with different numbers of predictors one should use R^2_{adj}. Note that adjusted R^2 is useful for casual assessment of improvement of fit, but it does not have the simple summarizing interpretation that R^2 has.
2. The F-test is always used first to test for the existence of a linear relationship between Y and any of the predictors $X_1, \ldots, X_k$. If the F-test is significant then a natural question to ask is
"For which of the predictors is there evidence of a linear relationship with Y?"
To answer this question we could perform separate t-tests of $H_{0i} : \beta_i = 0$ for $i = 1, \ldots, k$.

Example (Delivery time data)

Suppose that we wish to test for significance of regression using the delivery time data.

Output

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	5550.81092	2775.40546	261.24	<.0001
Error	22	233.73168	10.62417		
Corrected Total	24	5784.54260			

Root MSE	3.25947	R-Square	0.9596
Dependent Mean	22.38400	Adj R-Sq	0.9559
Coeff Var	14.56162		

From the above analysis of variance table, the test statistic for testing $H_0 : \beta_1 = \beta_2 = 0$ is

$$F_0 = \frac{MS_R}{MS_{Res}} = \frac{2775.40546}{10.62417} = 261.24.$$

Since $F_0 = 261.24 > F_{.05,2,22} = 3.44$, H_0 is rejected at the significance level .05. Also, since $P - value < .0001$ from SAS Output, we reject the null hypothesis and conclude that delivery time is related to delivery volume and/or distance. However, this does not necessarily imply that the relationship found is an appropriate one for predicting delivery time as a function of volume and distance.

- **Test of significance of some of the predictors**

  Suppose that we are interested in testing

  $$H_0 : \beta_{q+1} = \cdots = \beta_k = 0 \quad \text{versus} \quad H_1 : H_0 \text{ is not true}$$

  for $q < k$. This can be achieved using an F-test.

  a. Full model (the initial model under consideration):

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

  where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

  b. Reduced model (the model under H_0):

  $$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_q x_{iq} + \epsilon_i, \quad i = 1, \ldots, n,$$

  where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

Let $SS_{Res,Full}$ be the residual sum of squares under the full model and $SS_{Res,Reduced}$ be the residual sum of squares under the reduced model. Then, the test statistic is

$$F_0 = \frac{(SS_{Res,Reduced} - SS_{Res,Full})/(df_{Reduced} - df_{Full})}{SS_{Res,Full}/df_{Full}} = \frac{(SS_{Res,Reduced} - SS_{Res,Full})/(k-q)}{SS_{Res,Full}/(n-k-1)}.$$

The null hypothesis is rejected if $F_0 > F_{\alpha,k-q,n-k-1}$.

This is called a **partial F-test**. Note that a partial F-test measures the contribution of the predictors $X_{q+1}, \ldots, X_k$ given the other predictors $X_1, \ldots, X_q$ are in the model.

Example (Delivery time data)

Suppose that we wish to investigate the contribution of the variable distance (X_2) to the model.

SAS Program

```
proc reg data=delivery;
model Time = Cases Distance;
test Distance;
model Time = Cases;
run; quit;
```

Output

```
/* Reduced model */
                          Analysis of Variance

                                   Sum of          Mean
Source                     DF       Squares        Square   F Value   Pr > F

Model                       1    5382.40880    5382.40880    307.85   <.0001
Error                      23     402.13380      17.48408
Corrected Total            24    5784.54260


          Root MSE              4.18140    R-Square     0.9305
          Dependent Mean       22.38400    Adj R-Sq     0.9275
          Coeff Var            18.68029
```

The appropriate hypotheses are

$$H_0 : \beta_2 = 0 \quad \text{versus} \quad H_1 : \beta_2 \neq 0.$$

We have $SS_{Res,Full} = 233.73168$ and $SS_{Res,Reduced} = 402.13380$. Thus, the test statistic is

$$F_0 = \frac{(402.13380 - 233.73168)/1}{233.73168/22} = 15.85.$$

Since $F_{.05,1,22} = 4.30$, we reject H_0 at the level of significance .05 and conclude that distance contributes significantly to the model. Notice that this partial F-test is equivalent to the t-test. Observe that $t_0^2 = (3.98)^2 = F_0$.

- **Testing equality of regression coefficients**

  Suppose that we are interested in testing

  $$H_0 : \beta_{r-1} = \beta_r \quad \text{versus} \quad H_1 : H_0 \text{ is not true.}$$

  We use a partial F-test to testing $H_0 : \beta_{r-1} = \beta_r$.

  The reduced model is

  $$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{r-1}(x_{i,r-1} + x_{i,r}) + \beta_{r+1} x_{i,r+1} + \cdots + \beta_k x_{ik} + \epsilon_i, \quad i = 1, \ldots, n,$$

  where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

3.3 Standardized Regression Coefficients

It is usually difficult to directly compare regression coefficients because the magnitude of $\hat{\beta}_j$ reflects the units of measurement of the predictor x_j. For example, suppose that the regression model is

$$\hat{Y} = 5 + X_1 + 1000 X_2$$

and Y is measured in liters, X_1 is measured in liters, and X_2 is measured in milliliters. Note that although $\hat{\beta}_2$ is considerably larger than $\hat{\beta}_1$, the effect of both predictors on $\hat{Y}$ is identical, since a 1-liter change in either X_1 or X_2 when the other variable is held constant produces the same change in $\hat{Y}$. Thus, it is sometimes helpful to work with standardized predictors and response variable that produce standardized regression coefficients.

Recall the multiple linear regression model with k predictors. For $i = 1, \ldots, n$,

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

Now, let $y_i^s = \frac{y_i - \bar{y}}{s_y}$ with s_y the sample standard deviation of the y_i's and $x_{ij}^s = \frac{x_{ij} - \bar{x}_j}{s_{x_j}}$ with s_{x_j} the sample standard deviation of the x_{ij}'s, $j = 1, \ldots, k$. Then, the regression model can be written as

$$y_i^s = \alpha_0 + \alpha_1 x_{i1}^s + \alpha_2 x_{i2}^s + \cdots + \alpha_k x_{ik}^s + \epsilon_i,$$

where $\alpha_0 = \frac{\beta_0+\beta_1\bar{x}_1+\cdots+\beta_k\bar{x}_k-\bar{y}}{s_y}$ and $\alpha_j = \beta_j\frac{s_{x_j}}{s_y}, j = 1,\ldots,k$. We can now observe that $\hat{\alpha}_0 = \bar{y}^s - \hat{\alpha}_1\bar{x}_1^s - \cdots - \hat{\alpha}_k\bar{x}_k^s = 0$ and the least-square estimates of $\boldsymbol{\alpha}_* = (\alpha_1,\ldots,\alpha_k)^T$ are given by

$$\hat{\boldsymbol{\alpha}}_* = (\mathbf{X}_s^T\mathbf{X}_s)^{-1}\mathbf{X}_s^T\mathbf{y}_s,$$

where

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

We notice that

$$(n-1)^{-1}\mathbf{X}_s^T\mathbf{X}_s = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1k} \\ r_{12} & 1 & \cdots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1k} & r_{2k} & \cdots & 1 \end{pmatrix}$$

is a correlation matrix. Here, r_{jl} is the correlation coefficient between predictors X_j and X_l.

The regression coefficients $\hat{\alpha}_j$ are usually called standardized regression coefficients. The relationship between the original and standardized regression coefficients is

$$\hat{\beta}_j = \hat{\alpha}_j\frac{s_y}{s_{x_j}},\ j = 1,\ldots,k,$$

where s_y is the standard deviation in y_i and s_{x_j} is the standard deviation in x_{ij}.
Note: The least-squares estimates of the regression coefficients will not be changed with the standardized variables

$$y_i^s = \frac{y_i - \bar{y}}{S_{yy}}, \quad x_{ij}^s = \frac{x_{ij} - \bar{x}_j}{S_{jj}},\ j = 1,\ldots,k,$$

where $S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = (n-1)s_y^2$ and $S_{jj} = \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 = (n-1)s_{x_j}^2$.

Example

Revisit the delivery time data example.

SAS program

```
proc reg data=delivery;
model Time = Cases Distance / stb; /* displays standardized parameter estimates*/
run; quit;
```

Output

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	5550.81092	2775.40546	261.24	<.0001
Error	22	233.73168	10.62417		
Corrected Total	24	5784.54260			

Root MSE	3.25947	R-Square	0.9596
Dependent Mean	22.38400	Adj R-Sq	0.9559
Coeff Var	14.56162		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	2.34123	1.09673	2.13	0.0442	0
Cases	1	1.61591	0.17073	9.46	<.0001	0.71627
Distance	1	0.01438	0.00361	3.98	0.0006	0.30131

The fitted regression line is

$$\hat{Y}^* = .71627X_1^s + .30131X_2^s.$$

Thus, increasing the standardized values of cases X_1^s by one unit increases the standardized value of time $\hat{Y}^*$ by .71627 unit. Furthermore, increasing the standardized value of distance X_2^s by one unit increases $\hat{Y}^*$ by .3013 unit. Therefore, it seems that the volume of product delivered is more important than the distance in that it has a larger effect on delivery time in terms of the standardized variables.

3.4 Multicollinearity

Recall that we assumed that the column vectors of the matrix $\mathbf{X}$ are linearly independent in the multiple linear regression model. When the model includes more than one predictor it is important to assess whether strong correlations exist among the predictors (often referred to as multicollinearity). Multicollinearity can seriously disturb the least-squares fit and in some situations render the regression model almost useless.

Brief introduction and effects of multicollinearity will be introduced throughout this section.

- **Effect of multicollinearity**

  Multicollinearity implies near-linear dependence among the predictors. This would result in a singular $\mathbf{X}^T\mathbf{X}$ (equivalently, $\mathbf{X}_s^T\mathbf{X}_s$). Hence, the presence of near-linear dependencies can dramatically impact the ability to estimate regression coefficients. For example, regression coefficients can have wrong sign or many of the predictors are not statistically significant when the overall F-test is highly significant. Read p.113 in textbook.

  Suppose that there are only two predictors. Recall that the regression model with standardized variables. Then, the least-squares normal equations are

  $$\mathbf{X}_s^T\mathbf{X}_s\hat{\boldsymbol{\alpha}}_* = \mathbf{X}_s^T\mathbf{y}_s.$$

  It can be easily shown that normal equations become

  $$\begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix} \begin{pmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{pmatrix} = \begin{pmatrix} r_{1Y} \\ r_{2Y} \end{pmatrix}$$

  where r_{12} is the correlation coefficient between X_1 and X_2 and r_{jY} is the correlation coefficient between X_j and Y, $j = 1, 2$. Then, we observe that

  $$(\mathbf{X}_s^T\mathbf{X}_s)^{-1} = (n-1)^{-1} \begin{pmatrix} \frac{1}{1-r_{12}^2} & -\frac{r_{12}}{1-r_{12}^2} \\ -\frac{r_{12}}{1-r_{12}^2} & \frac{1}{1-r_{12}^2} \end{pmatrix}, \quad \mathbf{X}_s^T\mathbf{y}_s = (n-1)^{-1} \begin{pmatrix} r_{1Y} \\ r_{2Y} \end{pmatrix},$$

  and, hence, the estimates of the regression coefficients are

  $$\hat{\alpha}_1 = \frac{r_{1Y} - r_{12}r_{2Y}}{1 - r_{12}^2}, \quad \hat{\alpha}_2 = \frac{r_{2Y} - r_{12}r_{1Y}}{1 - r_{12}^2}.$$

  If there is strong multicollinearity between X_1 and X_2, then the correlation coefficient r_{12} will be large. Consequently, $\mathrm{Var}(\hat{\alpha}_1) = \mathrm{Var}(\hat{\alpha}_2) = \frac{1}{1-r_{12}^2} \to \infty$. Therefore, strong multicollinearity results in large variances for the least-squares estimates of the regression coefficients. This implies that different samples taken at the same levels of a predictor could lead to widely different estimates of the model parameters.

The simplest way to assess the extent of multicollinearity among the predictors is to look at scatterplot matrix of the predictors (plots of X_i versus X_j) and obtain the correlation matrix among the predictors. If the predictors X_i and X_j are nearly independent, then $|r_{ij}|$ will be near unity. Here r_{ij} is the sample correlation coefficient between X_i and X_j.

- **Measure of multicollinearity**:

  The variance inflation factor (VIF) for the predictor X_j is defined as the jth diagonal element of the inverse of the correlation matrix among predictors and it is given by

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

where R_j^2 is the coefficient of multiple determination obtained from regression X_j on the other predictors.
VIF's larger than 10 imply serious problems with multicollinearity. We can overcome the multicollinearity problem by either penalizing the regression coefficients (known as ridge regression, see Section 11.5.3 in textbook) or by defining new predictors as linear combinations of predictors that are linear dependent (e.g., principal-component (PC) regression (see Section 11.5.4 in textbook) or partial least-squares (PLS) regression).

Example

Revisit the delivery time data.

SAS Program:

```
proc reg;
model Time = Cases Distance / vif;
run;
```

Output:

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	2.34123	1.09673	2.13	0.0442	0
Cases	1	1.61591	0.17073	9.46	<.0001	3.11847
Distance	1	0.01438	0.00361	3.98	0.0006	3.11847

There is no VIF's larger than 10. The multicollinearity among two predictors is not detected.

3.5 Interaction in Multiple Regression Model

In the multiple regression models we have been considering so far, the effects of the predictors have been additive. Even if additivity is appropriate for many situations, there are times when it does not apply.

Two predictor variables are said to interact if the effect that one of them has on the mean response depends on the value of the other. In multiple regression, a predictor variable for interaction can be constructed as the product of the two predictor variables that are thought to interact.

Consider the two-predictor model including interaction effect given by

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \epsilon.$$

It can be rewritten two ways to show how the change in response with one variable depends on the other.

(1) $Y = \beta_0 + \beta_1 X_1 + (\beta_2 + \beta_{12} X_1) X_2 + \epsilon$

(2) $Y = \beta_0 + \beta_2 X_2 + (\beta_1 + \beta_{12} X_2) X_1 + \epsilon.$

Expression (1) shows the difference in Y per unit difference in X_2 when X_1 is held fixed is $(\beta_2 + \beta_{12} X_1)$. This varies with the value of X_1. Expression (2) shows the difference in Y per unit difference in X_1 when X_2 is held fixed is $(\beta_1 + \beta_{12} X_2)$. This varies with the value of X_2. The coefficient β_{12} measures the amount by which the change in response with one predictor is affected by the other predictor. If β_{12} is not statistically significant, then the data have not demonstrated the change in response with one predictor depends on the value of the other predictor.

When to include interaction terms

Interaction terms are not routinely included in regression models. Inclusion is indicated in three situations: when a question of interest pertains to interaction; when good reason exists to suspect interaction; or when interactions are proposed as a more general model for the purpose of examining whether additivity holds.

Remark A model including an interaction term between two predictors should also include terms with each of the predictors individually, even though their coefficients may not be significant. Following this rule avoids the logical inconsistency of saying that the effect of X_1 depends on the level of X_2 but that there is no effect of X_1.

Example

```
Dependent Variable: HCHOL

/* no-interaction model (additive model) */

                     Parameter   Standard    T for H0:
     Variable  DF     Estimate     Error    Parameter=0    Prob > |T|

     INTERCEPT  1     64.853       8.377       7.742       0.000
     BMI        1     -1.441       0.321      -4.488       0.000
     CHOL       1      0.068       0.027       2.498       0.014

/* interaction model */

                     Parameter   Standard    T for H0:
     Variable  DF     Estimate     Error    Parameter=0    Prob > |T|

     INTERCEPT  1    -24.990      38.234      -0.654       0.515
     BMI        1      2.459       1.651       1.489       0.139
     CHOL       1      0.498       0.181       2.753       0.007
     BMI*CHOL   1     -0.019       0.008      -2.406       0.018
```

When HDL (High Density Lipoprotein) cholesterol is regressed on total cholesterol and BMI (Body Mass Index), the model fit to the data

$$\text{HCHOL} = \beta_0 + \beta_1 * \text{CHOL} + \beta_2 * \text{BMI} + \epsilon$$

forces the effects to be additive, that is, the effect of CHOL is the same for all values of BMI and vice-versa because the model won't let it be anything else. The fitted model is

$$\widehat{\text{HCHOL}} = 64.853 - 1.441 * \text{BMI} + .068 * \text{CHOL}.$$

The expected difference in HCHOL is .068 per unit difference in CHOL when BMI is held fixed. This is true whatever the value of BMI. The difference in HCHOL is −1.441 per unit difference in BMI when CHOL is held fixed. This is true whatever the value of CHOL. The effects of CHOL and BMI are additive because the expected difference in HDL cholesterol corresponding to differences in both CHOL and BMI is obtained by adding the differences expected from CHOL and BMI determined without regard to the other's value.

Perhaps the way HDL cholesterol varies with BMI depends on total cholesterol. One way to investigate this is by including an interaction term in the model. The fitted model incorporating the interaction is

$$\widehat{\text{HCHOL}} = -24.99 + 2.459 * \text{BMI} + .498 * \text{CHOL} - .019 * \text{BMI} * \text{CHOL}.$$

This can be rewritten

$$\widehat{\text{HCHOL}} = -24.99 + .498 * \text{CHOL} + (2.459 - .019 * \text{CHOL}) * \text{BMI}.$$

The coefficient of BMI is $(2.459 - .019 * \text{CHOL})$. This would be interpreted as saying that among those with a given total cholesterol level less than 129.42, those with greater BMIs are expected to have greater HDL levels whereas among those with a given total cholesterol level greater than 129.42, those with greater BMIs are expected to have lower HDL levels.

3.6 Polynomial Regression

We consider an important special case of multiple regression, known as polynomial regression. Polynomial models are useful in situations where the analyst knows that curvilinear effects are present in the true response function. They are also useful as approximating functions to unknown and possibly very complex nonlinear relationships. The kth-order polynomial model in a single predictor, X, is

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_k X^k + \epsilon.$$

If we set $X_j = X^j$, $j = 1, \ldots, k$, then this polynomial model becomes a multiple linear regression model in the k predictors $X_1, \ldots, X_k$. In the kth-order polynomial regression in a single predictor, we can display the result of our multiple regression on a 2-dimensional graph.

Note: The 2nd-order polynomial model in two predictors is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{12} X_1 X_2 + \epsilon.$$

Example

The Hardwood Data in Ex 7.1 on p.205 in the textbook.

Nineteen observations on the strength of kraft paper and the percentage of hardwood in the batch of pulp from which the paper was produced have been collected.

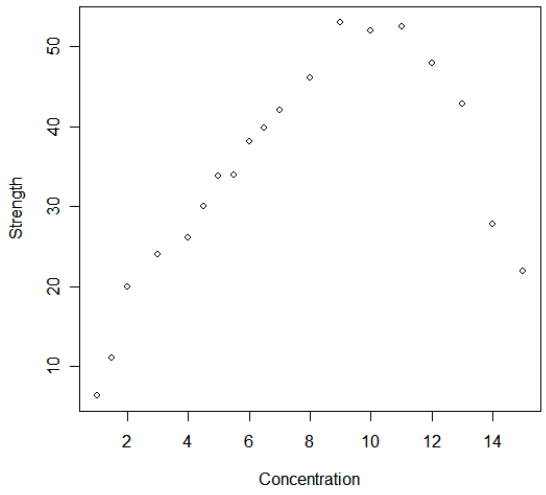

Figure 3.2: A scatterplot of the hardwood data.

SAS Program:

```
option ls=90 ps=75;
title 'Hardwood Data';
data hardwood;
infile 'Hardwood.dat';
input Concentration Strength;
Concentration2 = Concentration**2;
run;
proc reg;
model Strength = Concentration;
plot Strength*Concentration;
run;
proc reg;
model Strength = Concentration Concentration2 / clb clm cli;
run;
```

Output:

```
/* Simple Linear Regression */

                          Analysis of Variance

                                    Sum of          Mean
Source                   DF       Squares        Square   F Value   Pr > F

Model                     1    1043.42743    1043.42743      7.47   0.0141
Error                    17    2373.45783     139.61517
Corrected Total          18    3416.88526

          Root MSE             11.81589    R-Square     0.3054
          Dependent Mean       34.18421    Adj R-Sq     0.2645
          Coeff Var            34.56533


                        Parameter Estimates

                       Parameter       Standard
Variable         DF     Estimate          Error    t Value    Pr > |t|

Intercept         1     21.32126        5.43018       3.93      0.0011
Concentration     1      1.77099        0.64781       2.73      0.0141


/* Polynomial Regression */

                          Analysis of Variance

                                    Sum of          Mean
Source                   DF       Squares        Square   F Value   Pr > F

Model                     2    3104.24697    1552.12349     79.43   <.0001
Error                    16     312.63829      19.53989
Corrected Total          18    3416.88526

          Root MSE              4.42040    R-Square     0.9085
          Dependent Mean       34.18421    Adj R-Sq     0.8971
          Coeff Var            12.93110


                        Parameter Estimates

                       Parameter       Standard
Variable         DF     Estimate          Error    t Value    Pr > |t|

Intercept         1     -6.67419        3.39971      -1.96      0.0673
Concentration     1     11.76401        1.00278      11.73      <.0001
Concentration2    1     -0.63455        0.06179     -10.27      <.0001
```

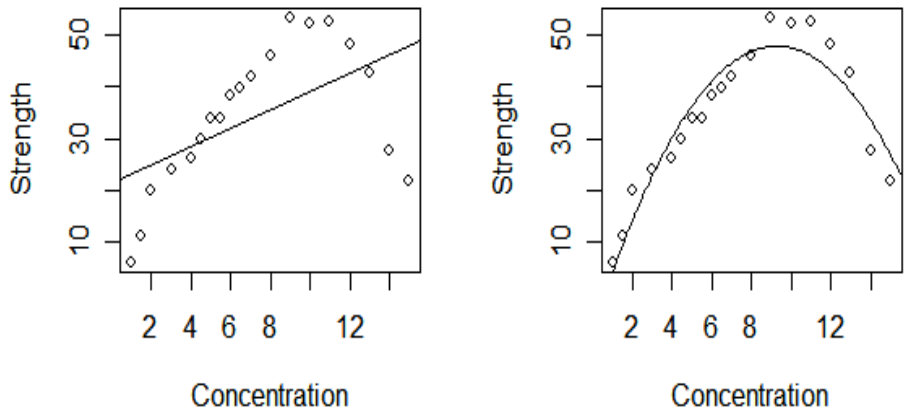

Figure 3.3: Plots of tensile strength against hardwood concentration with simple linear regression fit and a quadratic fit

Recommendations when fitting a polynomial model

- It is important to keep the order of the model as low as possible. When the response function appears to be curvilinear, transformations (Chapter 7) should be tried to linearize the model. If this fails, a second-order polynomial should be tried. As a general rule the use of high-order polynomials ($k > 2$) should be avoided unless they can justified for reasons outside the data. A low-order model in a transformed variable is almost always preferable to a high-order model in the original metric. Arbitrary fitting of high-order polynomials is a serious abuse of regression analysis. One should always maintain a sense of parsimony, that is, use the simplest possible model that is consistent with the data and knowledge of the problem environment. Remember that in extreme case it is always possible to pass a polynomial of order $n-1$ through n points so that a polynomial of sufficiently high degree can always be found that provides a "good" fit to the data. Such a model would do nothing to enhance understanding of the unknown function, nor will it likely be a good predictor.

- To choose the order of the polynomial model, one approach is to successively fit models of increasing order until the t-test for the highest order term is nonsignificant (forward selection). An alternate procedure is to appropriately fit the highest order model and then delete terms one at a time, starting with the highest order, until the highest order remaining term has a significant t statistic (backward elimination).

- Extrapolation with polynomial models can be extremely hazardous.