Chapter 3. Multiple Linear Regression

Outline

3.1. Description of the Data and Model

3.1 Description of the Data and Model

Introduction

*We chapter the **general multiple** linear regression model is presented. The presentation serves as a review of the standard results on regression analysis.*

Description of the Data

*The data consist of **n** observations on a dependent or **response** variable Y and **p predictor** or explanatory variables, $X_1, X_2, \cdots, X_p$. The **observations** are usually represented as in Table 3.1.*

Table 3.1 Notation for Data Used in Multiple Regression Analysis

Observation Number	Response Y	Predictors X_1	X_2	$\cdots$	X_p
1	y_1	x_{11}	x_{12}	$\cdots$	x_{1p}
2	y_2	x_{21}	x_{22}	$\cdots$	x_{2p}
3	y_3	x_{31}	x_{32}	$\cdots$	x_{3p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_n	x_{n1}	x_{n2}	$\cdots$	x_{np}

3.1 Description of the Data and Model

Model

Description of the Model

*The relationship between Y and $X_1, X_2, \cdots, X_p$ is formulated as a **linear** model*

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon, \tag{3.1}$$

*where $\beta_1, \beta_2, \cdots, \beta_p$ are **constants** referred to as the model **partial** regression coefficients (or simply as the regression coefficient) and ε is a **random** disturbance or error. error. It is assumed that for any set of fixed values of $X_1, X_2, \cdots, X_p$ that fall within the range of the data, the linear equation (3.1) provides an **acceptable** approximation of the true relationship between Y and the X's (Y is **approximately** a linear function of the X's, and ε measures the **discrepancy** in that approximation). In particular, ε contains **no** systematic information for determining Y that is not already captured by the X's.*

According to (3.1), each observation in Table 3.1 can be written as

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, 2, \cdots, n, \tag{3.2}$$

where y_i represents the ith value of the response variable Y, $x_{i1}, x_{i2}, \cdots, x_{ip}$ represent values of the predictor variables for the ith unit (the ith row in Table 3.1), and ε_i represents the error in the approximation of y_i.

3.1 Description of the Data and Model

Example

Supervisor Performance Data

We use data from a study in industrial psychology (management) to illustrate some of the standard regression results. A recent survey of the clerical employees of a large financial organization included questions related to employee satisfaction with their supervisors. There was a question designed to measure the ***overall performance*** *of a supervisor, as well as questions that were related to specific activities involving interaction between supervisor and employee. An exploratory study was undertaken to try to explain the* ***relationship*** *between specific supervisor characteristics and overall satisfaction with supervisors as perceived by the employees. Initially,* ***six questionnaire item****s were chosen as possible* ***explanatory*** *variables. Table 3.2 gives the description of the variables in the study.*

Table 3.2 Description of Variables in Supervisor Performance Data

Variable	Description
Y	Overall rating of job being done by supervisor
X_1	Handles employee complaints
X_2	Does not allow special privileges
X_3	Opportunity to learn new things
X_4	Raises based on performance
X_5	Too critical of poor performance
X_6	Rate of advancing to better jobs

next page

3.1 Description of the Data and Model

Example

Supervisor Performance Data

*We can be seen from the list, there are **two** broad types of variables included in the study. Variables X_1, X_2 and X_5 relate to **direct** interpersonal **relationships** between employee and supervisor, whereas variables X_3 and X_4 are of a **less personal nature** and relate to the job as a whole. Variable X_6 is not a direct evaluation of the supervisor but serves more as a **general measure** of how the employee perceives his or her own progress in the company.*

*The data for the analysis were generated from the **individual** employee response to the items on the survey questionnaire. The response on any item ranged from 1 through 5, indicating very **satisfactory** to very **unsatisfactory**, respectively. A dichotomous index was created to each item by collapsing the response scale to two categories: {1,2}, to be interpreted as a **favorable** response, and {3,4,5}, representing an **unfavorable** response. The data were collected in **30** departments selected at random from the organization. Each department had approximately **35** employees and **one** supervisor. The data to be used in the analysis, given in Table 3.3, were obtained by **aggregating** responses for departments to get the proportion of favorable responses for each item for each department. The resulting data therefore consist of 30 observations on seven variables, one observation for each department.*

A linear model of the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_6 X_6 + \varepsilon, \qquad (3.3)$$

relating Y and the six explanatory variables, is assumed.

Table 3.3

3.1 Description of the Data and Model

Example

Supervisor Performance Data

Table 3.3 Supervisor Performance Data

Row	Y	X_1	X_2	X_3	X_4	X_5	X_6
1	43	51	30	39	61	92	45
2	63	64	51	54	63	73	47
3	71	70	68	69	76	86	48
4	61	63	45	47	54	84	35
5	81	78	56	66	71	83	47
6	43	55	49	44	54	49	34
7	58	67	42	56	66	68	35
8	71	75	50	55	70	66	41
9	72	82	72	67	71	83	31
10	67	61	45	47	62	80	41
11	64	53	53	58	58	67	34
12	67	60	47	39	59	74	41
13	69	62	57	42	55	63	25
14	68	83	83	45	59	77	35
15	77	77	54	72	79	77	46
16	81	90	50	72	60	54	36
17	74	85	64	69	79	79	63
18	65	60	65	75	55	80	60
19	65	70	46	57	75	85	46
20	50	58	68	54	64	78	52
21	50	40	33	34	43	64	33
22	64	61	52	62	66	80	41
23	53	66	52	50	63	80	37
24	40	37	42	58	50	57	49
25	63	54	42	48	66	75	33
26	66	77	66	63	88	76	72
27	78	75	58	74	80	78	49
28	48	57	44	45	51	83	38
29	85	85	71	71	77	74	55
30	82	82	39	59	64	78	39

3.2. Parameter Estimation and Interpretation

3.2 Parameter Estimation and Interpretation

*Based on the available data, we wish to estimate the parameters $\beta_0, \beta_1, \cdots, \beta_p$. As in the case of simple regression presented in Chapter 2, we use the **least squares method**, that is, we minimize the sum of squares of the errors. From (3.2), the errors can be written as*

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip}, \quad i = 1, 2, \cdots, n. \tag{3.4}$$

The sum of squares of these errors is

$$S(\beta_0, \beta_1, \cdots, \beta_p) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2. \tag{3.5}$$

The least squares estimates $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$ are defined by minimizing $S(\beta_0, \beta_1, \cdots, \beta_p)$.

Use Matrix Notations to derive LSE

bold-face upper-case letter denote a matrix

bold-face lower-case letter denote a vector

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

3.2 Parameter Estimation and Interpretation

Derive the LSE

Use Matrix Notations to derive LSE

Then the observed data can be written as

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

The sum of squares of error can then be written as

$$S(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

$$\nabla_\beta S(\beta) = 2\mathbf{X}^\top(\mathbf{X}\beta - \mathbf{y})$$

*The LSE is the solution such that the **derivative** of $S(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ is **zero**.*

As a result, the LSE $\hat{\boldsymbol{\beta}}$ is solved by the following **normal** equations

$$\begin{bmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} & \cdots & \sum_{i=1}^n x_{i(p-1)} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} & \cdots & \sum_{i=1}^n x_{i1}x_{i(p-1)} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum_{i=1}^n x_{ip} & \sum_{i=1}^n x_{ip}\,x_{i1} & \sum_{i=1}^n x_{ip}\,x_{i2} & \cdots & \sum_{i=1}^n x_{ip}^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i1}y_i \\ \vdots \\ \sum_{i=1}^n x_{ip}y_i \end{bmatrix}$$

*The normal equations can be written much more **succinctly** in **matrix** notation,*

$$X^\top X\beta = X^\top y.$$

$\mathbf{X}^\top\mathbf{X}$ means $\mathbf{X}'\mathbf{X}$

3.2 Parameter Estimation and Interpretation

Derive the LSE

Least Squares Estimator

We can then solve this expression by multiplying both sides by the inverse of $\mathbf{X}'\mathbf{X}$ *, which exists, provided the columns of* X *are* ***linearly independent****. Then as always, we denote our solution with a* ***hat****.*

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

$(p+1)\times(p+1)$ $(p+1)\times n$ $n\times 1$

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix}$$

Using the estimated regression coefficients $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$ *, we write the* ***fitted least squares regression equation*** *as*

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p. \qquad (3.6)$$

For each observation in our data we can compute

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}, \quad i = 1, 2, \cdots, n. \qquad (3.7)$$

These are called the ***fitted*** *values. The corresponding ordinary least squares* ***residuals*** *are given by*

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \cdots, n.$$

They serve as the estimates of $\varepsilon_i, i = 1, 2, \cdots, n$

3.2 Parameter Estimation and Interpretation

If $\varepsilon_1, \cdots, \varepsilon_n$ are i.i.d. and with variance σ^2

An unbiased estimate of σ^2 is given by

$$\hat{\sigma}^2 = \frac{\text{SSE}}{n-p-1}, \quad (3.9) \qquad \textit{where} \qquad \text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2, \quad (3.10)$$

*is the **sum of squared residuals**. The number n-p-1 in the denominator is called the degrees of freedom (**df**).It is equal to the number of observations minus the number of estimated regression coefficients.*

Interpretation of Regression Coefficient

The interpretation of the regression coefficients in a multiple regression equation is a source of common confusion. The simple regression equation represents a line, while the multiple regression equation represents a plane (in cases of two predictors) or a hyperplane (in cases of more than two predictors). In multiple regression, the coefficient β_0 *, called the **constant coefficient**, is the value of Y when* $X_1 = \cdots = X_p = 0$*, as in simple regression.*

The regression coefficient $\beta_1, \cdots, \beta_p$ *, has **several** interpretations. It may be interpreted as the change in Y corresponding to a unit change in* X_j *when all **other** predictor variables are held **constant**. Magnitude of the change is not dependent on the values at which the other predictor variables are fixed. In practice, however, the predictor variables may be **inherently related**, and holding some of them constant while varying the others may not be possible.*

3.2 Parameter Estimation and Interpretation

Interpretation of Regression Coefficient

The regression coefficient β_j is also called the *partial regression coefficient* because β_j represents the contribution of X_j to the response variable Y after it has been adjusted for the other predictor variables. What does "adjusted for" mean in multiple regression? Without loss of any generality, we address this question using the simplest multiple regression case where we have two predictor variables. When $p = 2$, the model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon. \tag{3.11}$$

We use the variables X_1 and X_2 from the Supervisor data to illustrate the concepts. A statistical package gives the estimated regression equation as

$$\hat{Y} = 15.3276 + 0.7803 X_1 - 0.0502 X_2. \tag{3.12}$$

The coefficient of X_1 suggests that each unit of X_1 adds 0.7803 to Y when the value of X_2 is held fixed. As we show below, this is also the effect of X_1 after adjusting for X_2. Similarly, the coefficient of X_2 suggests that each unit of X_2 subtracts about 0.0502 from Y when the value of X_1 is held fixed. This is also the effect of X_2 after adjusting for X_1.

This interpretation can be easily understood when we consider the fact that the multiple regression equation can be obtained from a series of simple regression equations. For example, the coefficient of X_2 in (3.12) can be obtained as follows:

1. Fit the simple regression model that relates Y to X_1. Let the residuals from this regression be denoted by $\mathbf{e_{Y \cdot X_1}}$. This notation indicates that the variable that comes before the dot is treated as a response variable and the variable that comes after the dot is considered as a predictor. The fitted regression equation is
$$\hat{Y} = 14.3763 + 0.754610 X_1. \tag{3.13}$$
2. Fit the simple regression model that relates X_2 (considered temporarily here as a response variable) to X_1. Let the residuals from this regression be denoted by $\mathbf{e_{X_2 \cdot X_1}}$. The fitted regression equation is
$$\hat{X}_2 = 18.9654 + 0.513032 X_1. \tag{3.14}$$
The residuals, $\mathbf{e_{Y \cdot X_1}}$ and $\mathbf{e_{X_2 \cdot X_1}}$ are given in Table 3.4.
3. Fit the simple regression model that relates the above two residuals. In this regression, the response variable is $\mathbf{e_{Y \cdot X_1}}$ and the predictor variable is $\mathbf{e_{X_2 \cdot X_1}}$. The fitted regression equation is
$$\hat{\mathbf{e}}_{Y \cdot X_1} = 0 - 0.0502 \mathbf{e_{X_2 \cdot X_1}}. \tag{3.15}$$

The interesting result here is that the coefficient of $\mathbf{e_{X_2 \cdot X_1}}$ in this last regression is the same as the multiple regression coefficient of X_2 in (3.12). The two coefficients are equal to -0.0502. In fact, their standard errors are also the same. What's the intuition here? In the first step, we found the linear relationship between Y and X_1. The residual from this regression is Y after taking or partialling out the linear effects of X_1. In other words, the residual is that part of Y that is not linearly related to X_1. In the second step we do the same thing, replacing Y by X_2, so the residual is the part of X_2 that is not linearly related to X_1. In the third step we look for the linear relationship between the Y residual and the X_2 residual. The resultant regression coefficient represents the effect of X_2 on Y after taking out the effects of X_1 from both Y and X_2.

The regression coefficient β_j is the partial regression coefficient because it represents the contribution of X_j to the response variable Y after both variables have been linearly adjusted for the other predictor variables

Note that the estimated intercept in the regression equation in (3.15) is zero because the two sets of residuals have a mean of zero (they sum up to zero). The same procedures can be applied to obtain the multiple regression coefficient of X_1 in (3.12). Simply interchange X_2 by X_1 in the above three steps. This is left as an exercise

→

Table 3.4

3.2 Parameter Estimation and Interpretation

Interpretation of Regression Coefficient

Table 3.4 Partial Residuals

Row	$e_{Y \cdot X_1}$	$e_{X_2 \cdot X_1}$	Row	$e_{Y \cdot X_1}$	$e_{X_2 \cdot X_1}$
1	−9.8614	−15.1300	16	−1.2912	−15.1383
2	0.3287	−0.7995	17	−4.5182	1.4269
3	3.8010	13.1224	18	5.3471	15.2527
4	−0.9167	−6.2864	19	−2.1990	−8.8776
5	7.7641	−2.9819	20	−8.1437	19.2787
6	−12.8799	1.8178	21	5.4393	−6.4867
7	−6.9352	−11.3385	22	3.5925	1.7397
8	0.0279	−7.4428	23	−11.1806	−0.8255
9	−4.2543	10.9660	24	−2.2969	4.0524
10	6.5925	−5.2604	25	7.8748	−4.6691
11	9.6294	6.8439	26	−6.4813	7.5311
12	7.3471	−2.7473	27	7.0279	0.5572
13	7.8379	6.2266	28	−9.3891	−4.2082
14	−9.0089	21.4529	29	6.4818	8.4269
15	4.5187	−4.4689	30	5.7457	−22.0340

3.2 Parameter Estimation and Interpretation

LSE on Supervisor Dataset

```
############   Example on Lecture Slides -- Supervisor Dataset
supervisor_dat<-read.table('data/P060.txt',header=TRUE)   ## read the data
```

```
> supervisor_dat
    Y X1 X2 X3 X4 X5 X6
1  43 51 30 39 61 92 45
2  63 64 51 54 63 73 47
3  71 70 68 69 76 86 48
4  61 63 45 47 54 84 35
5  81 78 56 66 71 83 47
6  43 55 49 44 54 49 34
7  58 67 42 56 66 68 35
8  71 75 50 55 70 66 41
9  72 82 72 67 71 83 31
10 67 61 45 47 62 80 41
11 64 53 53 58 58 67 34
12 67 60 47 39 59 74 41
13 69 62 57 42 55 63 25
14 68 83 83 45 59 77 35
15 77 77 54 72 79 77 46
16 81 90 50 72 60 54 36
17 74 85 64 69 79 79 63
18 65 60 65 75 55 80 60
19 65 70 46 57 75 85 46
20 50 58 68 54 64 78 52
21 50 40 33 34 43 64 33
22 64 61 52 62 66 80 41
23 53 66 52 50 63 80 37
24 40 37 42 58 50 57 49
25 63 54 42 48 66 75 33
26 66 77 66 63 88 76 72
27 78 75 58 74 80 78 49
28 48 57 44 45 51 83 38
29 85 85 71 71 77 74 55
30 82 82 39 59 64 78 39
```

```
y<-supervisor_dat$Y
X<-as.matrix(supervisor_dat[,-1])
X<-cbind(rep(1,30),X)
colnames(X)<-c("Const.","X1","X2","X3","X4","X5","X6")
hat_beta <- solve(t(X) %*% X) %*% t(X) %*% y          ## LSE by direct computation using the formula
```

```
> hat_beta
              [,1]
Const. 10.78707639
X1      0.61318761
X2     -0.07305014
X3      0.32033212
X4      0.08173213
X5      0.03838145
X6     -0.21705668
```

3.3. Centering and Scaling

3.3 Centering and Scaling

No-intercept Model

*The magnitudes of the regression coefficients in a regression equation depend on the **unit** of measurements of the variables. For example, if the regression coefficient of income, when measured in dollars, is 5.123, this coefficient will change to 5123 if income were measured in $1000 instead.*
*To make the regression coefficients **unitless**, one may first **center** and/or **scale** the variables before performing the regression computations. There are other situations when centering and scaling the variables are desirable as in the case when dealing with the problem of collinearity in later chapters.*

We have been mainly dealing with regression models of the

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon, \qquad (3.16)$$

which are models with a constant term β_0 *. But there are also situations where fitting the **no-intercept** model*

$$Y = \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon \qquad (3.17)$$

*necessary. When dealing with **constant** term models, it is convenient to **center** and **scale** the variables, but when dealing with a **no-intercept model**, we need only to scale the variables.*

3.3 Centering and Scaling

Centering and Scaling in Intercept Model

Centering

*When fitting an intercept model as in (3.16), we can **center** and scale the **variables**. A **centered** variable is obtained by **subtracting** from each observation the mean of all observations. For example, the centered **response** variable is $Y - \bar{y}$ and the centered jth **predictor** variable is $X_j - \bar{x}_j$. The mean of a centered variable is **zero**.*

The centered variables can also be scaled. Two types of scaling are usually performed: unit-length scaling and standardizing.

Unit-length Scaling

*Unit length scaling of the **response** variable Y and the jth **predictor** variable X_j is obtained as follows:*

$$\begin{aligned} \tilde{Z}_y &= (Y - \bar{y})/L_y\,, \\ \tilde{Z}_j &= (X_j - \bar{x}_j)/L_j\,, \quad j = 1, \cdots, p, \end{aligned} \tag{3.18}$$

where $\bar{y}$ is the mean of Y, $\bar{x}_j$ is the mean of X_j, and

$$L_y = \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad \text{and} \quad L_j = \sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}, \quad j = 1, \cdots, p. \tag{3.19}$$

*The quantities L_y is referred to as the **length** of the centered variable $Y - \bar{y}$ because it measures the size or the **magnitudes** of the observations in $Y - \bar{y}$. Similarly, L_j measures the **length** of the variable $X_j - \bar{x}_j$.The variables $\tilde{Z}_y$ and $\tilde{Z}_j$ in (3.18) have **zero** means and **unit** lengths, hence this type of scaling is called **unit length scaling**.*

3.3 Centering and Scaling

Centering and Scaling in Intercept Model

Unit-Length Scaling

In addition, unit length scaling has the following property:

$$\text{Cor}(X_j, X_k) = \sum_{i=1}^{n} \tilde{z}_{ij}\tilde{z}_{ik} \qquad (3.20)$$

*That is, the **correlation coefficient** between the original variables, X_j and X_k can be computed easily as the sum of the products of the **scaled versions** $\tilde{Z}_j$ and $\tilde{Z}_k$.*

Standardizing

The second type of scaling is called standardizing, which is defined by

$$\begin{aligned} \tilde{Y} &= \frac{Y-\bar{y}}{s_y}, \\ \tilde{X}_j &= \frac{X_j-\bar{x}_j}{s_j}, \quad j=1,\cdots,p, \end{aligned} \qquad \Longrightarrow \frac{y_1-\bar{y}}{s_y}, \frac{y_2-\bar{y}}{s_y}, \cdots, \frac{y_n-\bar{y}}{s_y} \qquad (3.21)$$

where

$$s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}} \quad \text{and} \quad s_j = \sqrt{\frac{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}{n-1}}, \quad j=1,\cdots,p, \qquad (3.22)$$

are standard deviations of the response and jth predictor variable, respectively. The standardized variables $\tilde{Y}$ and $\tilde{X}_j$ in (3.21) have means zero and unit standard deviations.

*Since **correlations** are **unaffected** by centering and/or scaling the data, it is both **sufficient** and **convenient** to deal with either the **unit length scaled** or the **standardized** versions of the variables.*

3.3 Centering and Scaling

Scaling in No-Intercept Model

Do not Center

*If we are fitting a **no-intercept** model as in (3.17), we do **not** center the data because centering has the effect of including a **constant** term in the model. This can be seen from*

$$Y - \bar{y} = \beta_1(X_1 - \bar{x}_1) + \cdots + \beta_p(X_p - \bar{x}_p) + \varepsilon. \qquad (3.23)$$

Rearranging terms, we obtain

$$\begin{aligned} Y &= \bar{y} - (\beta_1\bar{x}_1 + \cdots + \beta_p\bar{x}_p) + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon \\ &= \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon, \qquad (3.24) \end{aligned}$$

where $\beta_0 = \bar{y} - (\beta_1\bar{x}_1 + \cdots + \beta_p\bar{x}_p)$. Although a constant term does not appear in an explicit form in (3.23), it is clearly seen in (3.24). Thus, when we deal with no-intercept models, we need only to scale the data. The scaled variables are defined by

$$\begin{aligned} \tilde{Z}_y &= Y/L_y, \\ \tilde{Z}_j &= X_j/L_j, \quad j = 1, \cdots, p, \end{aligned} \qquad (3.25)$$

where

$$L_y = \sqrt{\sum_{i=1}^{n} y_i^2} \quad \text{and} \quad L_j = \sqrt{\sum_{i=1}^{n} x_{ij}^2}, \quad j = 1, 2, \cdots, p. \qquad (3.26)$$

The scaled variables in (3.25) have unit lengths but do not necessarily have means zero. Nor do they satisfy (3.20) unless the original variables have zero means.

3.3 Centering and Scaling

Remarks

We should mention here that centering (when appropriate) and/or scaling can be done without loss of generality because the regression coefficients of the original variables can be recovered from the regression coefficients of the transformed variables. For example, if we fit a regression model to centered data, the obtained regression coefficients $\hat{\beta}_1, \cdots, \hat{\beta}_p$ are the same as the estimates obtained from fitting the model to the original data. The estimate of the constant term when using the centered data will always be zero. The estimate of the constant term for an intercept model can be obtained from

$$\hat{\beta}_0 = \bar{y} - (\hat{\beta}_1 \bar{x}_1 + \cdots + \hat{\beta}_p \bar{x}_p).$$

Scaling, however, will change the values of the estimated regression coefficients. For example, the relationship between the estimates, $\hat{\beta}_1, \cdots, \hat{\beta}_p$, obtained from using the original data and those obtained using the standardized data are given by

$$\begin{aligned} \hat{\beta}_j &= (s_y/s_j)\hat{\theta}_j, \qquad j = 1, 2, \cdots, p, \\ \hat{\beta}_0 &= \bar{y} - \sum_{j=1}^{p} \hat{\beta}_j \bar{x}_j, \end{aligned} \tag{3.27}$$

where $\hat{\beta}_j$ and $\hat{\theta}_j$ are the jth estimated regression coefficients obtained when using the original and standardized data, respectively. Similar formulas can be obtained when using unit length scaling instead of standardizing.

3.3 Centering and Scaling

On Supervisor Dataset

Original Dataset

```
> supervisor_dat
    Y X1 X2 X3 X4 X5 X6
1  43 51 30 39 61 92 45
2  63 64 51 54 63 73 47
3  71 70 68 69 76 86 48
4  61 63 45 47 54 84 35
5  81 78 56 66 71 83 47
6  43 55 49 44 54 49 34
7  58 67 42 56 66 68 35
8  71 75 50 55 70 66 41
9  72 82 72 67 71 83 31
10 67 61 45 47 62 80 41
11 64 53 53 58 58 67 34
12 67 60 47 39 59 74 41
13 69 62 57 42 55 63 25
14 68 83 83 45 59 77 35
15 77 77 54 72 79 77 46
16 81 90 50 72 60 54 36
17 74 85 64 69 79 79 63
18 65 60 65 75 55 80 60
19 65 70 46 57 75 85 46
20 50 58 68 54 64 78 52
21 50 40 33 34 43 64 33
22 64 61 52 62 66 80 41
23 53 66 52 50 63 80 37
24 40 37 42 58 50 57 49
25 63 54 42 48 66 75 33
26 66 77 66 63 88 76 72
27 78 75 58 74 80 78 49
28 48 57 44 45 51 83 38
29 85 85 71 71 77 74 55
30 82 82 39 59 64 78 39
```

```
supervisor_center <- apply(supervisor_dat,2,function(x) x-mean(x))    ## center the variables
```

```
> supervisor_center
                Y    X1          X2          X3          X4          X5          X6
 [1,] -21.6333333 -15.6 -23.1333333 -17.3666667  -3.6333333  17.2333333   2.066667
 [2,]  -1.6333333  -2.6  -2.1333333  -2.3666667  -1.6333333  -1.7666667   4.066667
 [3,]   6.3666667   3.4  14.8666667  12.6333333  11.3666667  11.2333333   5.066667
 [4,]  -3.6333333  -3.6  -8.1333333  -9.3666667 -10.6333333   9.2333333  -7.933333
 [5,]  16.3666667  11.4   2.8666667   9.6333333   6.3666667   8.2333333   4.066667
 [6,] -21.6333333 -11.6  -4.1333333 -12.3666667 -10.6333333 -25.7666667  -8.933333
 [7,]  -6.6333333   0.4 -11.1333333  -0.3666667   1.3666667  -6.7666667  -7.933333
 [8,]   6.3666667   8.4  -3.1333333  -1.3666667   5.3666667  -8.7666667  -1.933333
 [9,]   7.3666667  15.4  18.8666667  10.6333333   6.3666667   8.2333333 -11.933333
[10,]   2.3666667  -5.6  -8.1333333  -9.3666667  -2.6333333   5.2333333  -1.933333
[11,]  -0.6333333 -13.6  -0.1333333   1.6333333  -6.6333333  -7.7666667  -8.933333
[12,]   2.3666667  -6.6  -6.1333333 -17.3666667  -5.6333333  -0.7666667  -1.933333
[13,]   4.3666667  -4.6   3.8666667 -14.3666667  -9.6333333 -11.7666667 -17.933333
[14,]   3.3666667  16.4  29.8666667 -11.3666667  -5.6333333   2.2333333  -7.933333
[15,]  12.3666667  10.4   0.8666667  15.6333333  14.3666667   2.2333333   3.066667
[16,]  16.3666667  23.4  -3.1333333  15.6333333  -4.6333333 -20.7666667  -6.933333
[17,]   9.3666667  18.4  10.8666667  12.6333333  14.3666667   4.2333333  20.066667
[18,]   0.3666667  -6.6  11.8666667  18.6333333  -9.6333333   5.2333333  17.066667
[19,]   0.3666667   3.4  -7.1333333   0.6333333  10.3666667  10.2333333   3.066667
[20,] -14.6333333  -8.6  14.8666667  -2.3666667  -0.6333333   3.2333333   9.066667
[21,] -14.6333333 -26.6 -20.1333333 -22.3666667 -21.6333333 -10.7666667  -9.933333
[22,]  -0.6333333  -5.6  -1.1333333   5.6333333   1.3666667   5.2333333  -1.933333
[23,] -11.6333333  -0.6  -1.1333333  -6.3666667  -1.6333333   5.2333333  -5.933333
[24,] -24.6333333 -29.6 -11.1333333   1.6333333 -14.6333333 -17.7666667   6.066667
[25,]  -1.6333333 -12.6 -11.1333333  -8.3666667   1.3666667   0.2333333  -9.933333
[26,]   1.3666667  10.4  12.8666667   6.6333333  23.3666667   1.2333333  29.066667
[27,]  13.3666667   8.4   4.8666667  17.6333333  15.3666667   3.2333333   6.066667
[28,] -16.6333333  -9.6  -9.1333333 -11.3666667 -13.6333333   8.2333333  -4.933333
[29,]  20.3666667  18.4  17.8666667  14.6333333  12.3666667  -0.7666667  12.066667
[30,]  17.3666667  15.4 -14.1333333   2.6333333  -0.6333333   3.2333333  -3.933333
```

centered dataset

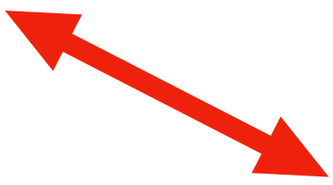
3.3 Centering and Scaling

On Supervisor Dataset

LSE on centered dataset

```
supervisor_center <- apply(supervisor_dat,2,function(x) x-mean(x))    ## center the variables
y_center<-supervisor_center[,1]
X_center<-as.matrix(supervisor_center[,-1])
X_center<-cbind(rep(1,30),X_center)
colnames(X_center)<-c("Const.","X1","X2","X3","X4","X5","X6")
hat_beta_center <- solve(t(X_center) %*% X_center) %*% t(X_center) %*% y_center        ## LSE by direct computation using the formula
```

```
> hat_beta_center
                [,1]
Const. -9.214851e-15
X1      6.131876e-01
X2     -7.305014e-02
X3      3.203321e-01
X4      8.173213e-02
X5      3.838145e-02
X6     -2.170567e-01
```



LSE on original dataset

```
> hat_beta
                [,1]
Const. 10.78707639
X1      0.61318761
X2     -0.07305014
X3      0.32033212
X4      0.08173213
X5      0.03838145
X6     -0.21705668
```

3.3 Centering and Scaling

On Supervisor Dataset

centered and standardized dataset

```
> supervisor_center_stand <- apply(supervisor_dat,2,function(x) (x-mean(x))/sd(x))
> supervisor_center_stand
                Y          X1          X2          X3          X4          X5          X6
 [1,] -1.77722105 -1.17163233 -1.89068413 -1.47964962 -0.34945220  1.74163662  0.2008675
 [2,] -0.13418156 -0.19527206 -0.17435704 -0.20164131 -0.15709319 -0.17854302  0.3952554
 [3,]  0.52303424  0.25535576  1.21505061  1.07636700  1.09324037  1.13526410  0.4924494
 [4,] -0.29848551 -0.27037669 -0.66473621 -0.79804519 -1.02270873  0.93313993 -0.7710720
 [5,]  1.34455398  0.85619286  0.23429227  0.82076534  0.61234285  0.83207784  0.3952554
 [6,] -1.77722105 -0.87121379 -0.33781676 -1.05364685 -1.02270873 -2.60403309 -0.8682660
 [7,] -0.54494143  0.03004185 -0.90992579 -0.03124020  0.13144532 -0.68385345 -0.7710720
 [8,]  0.52303424  0.63087895 -0.25608690 -0.11644076  0.51616334 -0.88597762 -0.1879083
 [9,]  0.60518621  1.15661140  1.54197006  0.90596589  0.61234285  0.83207784 -1.1598478
[10,]  0.19442634 -0.42058597 -0.66473621 -0.79804519 -0.25327270  0.52889158 -0.1879083
[11,] -0.05202958 -1.02142306 -0.01089731  0.13916090 -0.63799071 -0.78491554 -0.8682660
[12,]  0.19442634 -0.49569060 -0.50127648 -1.47964962 -0.54181121 -0.07748093 -0.1879083
[13,]  0.35873029 -0.34548133  0.31602213 -1.22404796 -0.92652923 -1.18916388 -1.7430115
[14,]  0.27657831  1.23171604  2.44099853 -0.96844630 -0.54181121  0.22570533 -0.7710720
[15,]  1.01594609  0.78108822  0.07083255  1.33196866  1.38177888  0.22570533  0.2980615
[16,]  1.34455398  1.75744850 -0.25608690  1.33196866 -0.44563170 -2.09872266 -0.6738781
[17,]  0.76949016  1.38192531  0.88813116  1.07636700  1.38177888  0.42782950  1.9503586
[18,]  0.03012239 -0.49569060  0.96986102  1.58757032 -0.92652923  0.52889158  1.6587768
[19,]  0.03012239  0.25535576 -0.58300635  0.05396035  0.99706086  1.03420202  0.2980615
[20,] -1.20215723 -0.64589988  1.21505061 -0.20164131 -0.06091369  0.32676741  0.8812252
[21,] -1.20215723 -1.99778334 -1.64549455 -1.90565239 -2.08068328 -1.08810180 -0.9654599
[22,] -0.05202958 -0.42058597 -0.09262718  0.47996312  0.13144532  0.52889158 -0.1879083
[23,] -0.95570130 -0.04506278 -0.09262718 -0.54244353 -0.15709319  0.52889158 -0.5766841
[24,] -2.02367697 -2.22309725 -0.90992579  0.13916090 -1.40742675 -1.79553640  0.5896433
[25,] -0.13418156 -0.94631842 -0.90992579 -0.71284464  0.13144532  0.02358115 -0.9654599
[26,]  0.11227437  0.78108822  1.05159089  0.56516368  2.24739442  0.12464324  2.8251042
[27,]  1.09809806  0.63087895  0.39775199  1.50236977  1.47795839  0.32676741  0.5896433
[28,] -1.36646118 -0.72100451 -0.74646607 -0.96844630 -1.31124724  0.83207784 -0.4794902
[29,]  1.67316188  1.38192531  1.46024019  1.24676811  1.18941987 -0.07748093  1.1728070
[30,]  1.42670596  1.15661140 -1.15511538  0.22436146 -0.06091369  0.32676741 -0.3822962
```

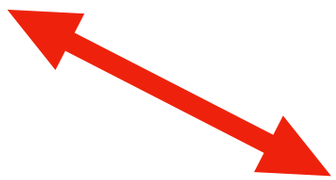
3.3 Centering and Scaling

On Supervisor Dataset

LSE on centered and standardized dataset

```
supervisor_center_stand <- apply(supervisor_dat,2,function(x) (x-mean(x))/sd(x))    ## center and standardize the variables
y_center_stand<-supervisor_center_stand[,1]
X_center_stand<-as.matrix(supervisor_center_stand[,-1])
X_center_stand<-cbind(rep(1,30),X_center_stand)
colnames(X_center_stand)<-c("Const.","X1","X2","X3","X4","X5","X6")
hat_beta_center_stand <- solve(t(X_center_stand) %*% X_center_stand) %*% t(X_center_stand) %*% y_center_stand       ## LSE by direct computation using the formula
```

```
> hat_beta_center_stand
                [,1]
Const. -7.147061e-16
X1      6.707252e-01
X2     -7.342743e-02
X3      3.088702e-01
X4      6.981172e-02
X5      3.119975e-02
X6     -1.834645e-01
```



LSE on original dataset

```
> hat_beta
                [,1]
Const. 10.78707639
X1      0.61318761
X2     -0.07305014
X3      0.32033212
X4      0.08173213
X5      0.03838145
X6     -0.21705668
```

```
> sd_all<-apply(supervisor_dat,2,sd)
> sd_all
        Y        X1        X2        X3        X4        X5        X6
12.172562 13.314757 12.235430 11.737013 10.397226  9.894908 10.288706
```

Sample standard deviation of each variable

verify the formula $\hat{\beta}_j = (s_y/s_j)\hat{\theta}_j$ for $j = 1, \cdots, p$

3.4. Properties of Least Square Estimators and Multiple Correlation Coefficient

3.4 Properties of Least Square Estimators and Multiple Correlation Coefficient

Core Assumptions

Core Assumption 1

For every fixed values of $X_1, \cdots, X_p$, the error $\varepsilon's$ are **independent normal** random variables with mean zero and variance σ^2.

$$\varepsilon_1, \cdots, \varepsilon_n \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$$

Core Assumption 2

The $(p+1) \times (p+1)$ matrix $\mathbf{X}'\mathbf{X}$ is invertible so that $(\mathbf{X}'\mathbf{X})^{-1}$ exists.

This guarantees the existence of the LSE: $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$

3.4 Properties of Least Square Estimators and Multiple Correlation Coefficient

Properties of LSE

Let $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \cdots, \hat{\beta}_p)'$ be the least squares estimator. If the core assumptions hold, then

1. The estimator $\hat{\beta}_j, j = 0, 1, \cdots, p$, is an unbiased estimate of β_j and has a variance of $\sigma^2 c_{jj}$, where c_{jj} is the jth diagonal element of the inverse of a matrix known as the *corrected sums of squares and products* matrix. The covariance between $\hat{\beta}_i$ and $\hat{\beta}_j$ is $\sigma^2 c_{ij}$, where c_{ij} is the element in the ith row and jth column of the inverse of the corrected sums of squares and products matrix. For all unbiased estimates that are linear in the observations the least squares estimators have the smallest variance. Thus, the least squares estimators are said to be BLUE (*best linear unbiased estimators*).

$$\mathbf{C} = (c_{ij})_{i,j=1}^{p+1} = (\mathbf{X}'\mathbf{X})^{-1}$$

2. The estimator $\hat{\beta}_j, j = 0, 1, \cdots, p$, is normally distributed with mean β_j and variance $\sigma^2 c_{jj}$.

3. $W = \text{SSE}/\sigma^2$ has a χ^2 distribution with $n - p - 1$ degrees of freedom, and $\hat{\beta}_j$'s and $\hat{\sigma}^2$ are distributed independently of each other.

4. The vector $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p)$ has a $(p+1)$-dimensional normal distribution with mean vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)$ and variance-covariance matrix with elements $\sigma^2 c_{ij}$.

*The results above enable us to test various **hypotheses** about individual regression parameters and to construct **confidence intervals**.*

3.4 Properties of Least Square Estimators and Multiple Correlation Coefficient

Example on Supervisor Dataset

```
supervisor_dat<-read.table('data/P060.txt',header=TRUE)   ## read the data
y<-supervisor_dat$Y
X<-as.matrix(supervisor_dat[,-1])
X<-cbind(rep(1,30),X)
colnames(X)<-c("Const.","X1","X2","X3","X4","X5","X6")
hat_beta <- solve(t(X) %*% X) %*% t(X) %*% y          ## LSE by direct computation using the formula
```

```
> hat_beta
              [,1]
Const. 10.78707639
X1      0.61318761
X2     -0.07305014
X3      0.32033212
X4      0.08173213
X5      0.03838145
X6     -0.21705668
```

obtain $\hat{\sigma}^2$

```
> n<-30
> p<-6
> hat_sig2<-sum((hat_y-y)^2)/(n-p-1)
> hat_sig2
[1] 49.95654
```

obtain the matrix $(\mathbf{X}'\mathbf{X})^{-1}$

```
> solve(t(X) %*% X)
              Const.           X1           X2           X3           X4           X5           X6
Const.  2.6885547839 -2.543299e-03 -3.633619e-03 -6.153588e-03 -4.329621e-03 -2.266050e-02  6.562191e-04
X1     -0.0025432993  5.187622e-04 -1.636348e-04 -1.638338e-04 -3.717724e-04  4.677579e-07  2.309832e-04
X2     -0.0036336194 -1.636348e-04  3.687444e-04 -6.251148e-05  4.616729e-05 -9.156320e-06 -8.936448e-05
X3     -0.0061535881 -1.638338e-04 -6.251148e-05  5.684761e-04 -1.539315e-04  9.702610e-05 -2.087411e-04
X4     -0.0043296211 -3.717724e-04  4.616729e-05 -1.539315e-04  9.819008e-04 -1.827734e-04 -3.373815e-04
X5     -0.0226604991  4.677579e-07 -9.156320e-06  9.702610e-05 -1.827734e-04  4.325292e-04 -6.705033e-05
X6      0.0006562191  2.309832e-04 -8.936448e-05 -2.087411e-04 -3.373815e-04 -6.705033e-05  6.357249e-04
```

3.4 Properties of Least Square Estimators and Multiple Correlation Coefficient

Multiple Correlation Coefficient

*After fitting the linear model to a given data set, an assessment is made of the **adequacy of fit**.*
*All the material, discussed in the **simple** linear regression, **extend** naturally to **multiple** regression.*

Multiple Correltion Coefficient

The strength of the linear relationship between Y and the set of predictors $X_1, X_2, \cdots, X_p$ can be assessed through the examination of the scatter plot of Y versus $\hat{Y}$ and the correlation coefficient between Y and $\hat{Y}$, which is given by

$$\text{Cor}(Y, \hat{Y}) = \frac{\sum(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum(y_i - \bar{y})^2 \sum(\hat{y}_i - \bar{\hat{y}})^2}}, \tag{3.28}$$

where $\bar{y}$ is the mean of the response variable Y and $\bar{\hat{y}}$ is the mean of the fitted values. As in the simple regression case, the coefficient of determination $R^2 = [\text{Cor}(Y, \hat{Y})]^2$ is also given by

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}, \tag{3.29}$$

as in (2.46). Thus, R^2 may be interpreted as the proportion of the total variability in the response variable Y that can be accounted for by the set of predictor variables $X_1, X_2, \cdots, X_p$. In multiple regression, $R = \sqrt{R^2}$ is called the *multiple correlation coefficient* because it measures the relationship between one variable Y and a set of variables $X_1, X_2, \cdots, X_p$.

The value of R^2 for the Supervisor Performance data is 0.73, showing that about 73% of the total variation in the overall rating of the job being done by the supervisor can be accounted for by the six variables.

3.4 Properties of Least Square Estimators and Multiple Correlation Coefficient

Obtain R-squared on Supervisor Dataset

using the original dataset

```
supervisor_dat<-read.table('data/P060.txt',header=TRUE)   ## read the data
y<-supervisor_dat$Y
X<-as.matrix(supervisor_dat[,-1])
X<-cbind(rep(1,30),X)
colnames(X)<-c("Const.","X1","X2","X3","X4","X5","X6")
hat_beta <- solve(t(X) %*% X) %*% t(X) %*% y          ## LSE by direct computation using the formula
hat_y <- X %*% hat_beta
Rsquared <- sum((hat_y-mean(y))^2)/sum((y-mean(y))^2)
```

```
> Rsquared
[1] 0.732602
```

using the centered and standardized dataset

```
> supervisor_center_stand <- apply(supervisor_dat,2,function(x) (x-mean(x))/sd(x))    ## center and standardize the variables
> y_center_stand<-supervisor_center_stand[,1]
> X_center_stand<-as.matrix(supervisor_center_stand[,-1])
> X_center_stand<-cbind(rep(1,30),X_center_stand)
> colnames(X_center_stand)<-c("Const.","X1","X2","X3","X4","X5","X6")
> hat_beta_center_stand <- solve(t(X_center_stand) %*% X_center_stand) %*% t(X_center_stand) %*% y_center_stand       ## LSE by direct computation using the formula
> hat_y <- X_center_stand %*% hat_beta_center_stand
> Rsquared <- sum((hat_y-mean(y_center_stand))^2)/sum((y_center_stand-mean(y_center_stand))^2)
> Rsquared
[1] 0.732602
```

3.4 Properties of Least Square Estimators and Multiple Correlation Coefficient

Multiple Correlation Coefficient

When the model fits the data well, it is clear that the value of R^2 is close to unity. With a good fit, the observed and predicted values will be close to each other, and $\sum(y_i - \hat{y}_i)^2$ will be small. Then R^2 will be near unity. On the other hand, if there is no linear relationship between Y and the predictor variables, $X_1, \cdots, X_p$, the linear model gives a poor fit, the best predicted value for an observation y_i would be $\bar{y}$; that is, in the absence of any relationship with the predictors, the best estimate of any value of Y is the sample mean, because the sample mean minimizes the sum of squared deviations. So in the absence of any linear relationship between Y and the X's, R^2 will be near zero. The value of R^2 is used as a summary measure to judge the fit of the linear model to a given body of data. As pointed out in Chapter 2, a large value of R^2 does not necessarily mean that the model fits the data well.

3.4 Properties of Least Square Estimators and Multiple Correlation Coefficient

Multiple Correlation Coefficient

Drawbacks of R^2

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

R^2 measures how good the linear combinations of columns of $\mathbf{X}$ can approximate the response $\mathbf{y}$

However, suppose we find a "new" variable X_{p+1} and we add this into our model

$$y = \alpha_0 + \alpha_1 X_1 + \cdots + \alpha_p X_p + \alpha_{p+1} X_{p+1} + \varepsilon$$

Then, the new predictor matrix $\mathbf{X}_{\text{new}}$ will have one more column than $\mathbf{X}$:

$$\mathbf{X}_{\text{new}} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} & x_{1(p+1)} \\ 1 & x_{21} & \cdots & x_{2p} & x_{2(p+1)} \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} & x_{n(p+1)} \end{pmatrix}$$

Clearly, the R^2 by this new model is always larger than or equal to R^2 of the original model, even if this new variable X_{p+1} is an *irrelevant* and *completely* random predictor. In other words, generally speaking, *more variables will yield a larger* R^2. Therefore, a fair measure of goodness-of-fit should be re-balanced by the number of variables.

3.4 Properties of Least Square Estimators and Multiple Correlation Coefficient

Multiple Correlation Coefficient

Adjusted R-squared

A quantity related to R^2, known as the **adjusted R-squared** R_a^2, is also used for judging the goodness of fit

$$R_a^2 = 1 - \frac{\text{SSE}/(n-p-1)}{\text{SST}/(n-1)}, \tag{3.30}$$

implying that

$$R_a^2 = 1 - \frac{n-1}{n-p-1}(1-R^2). \tag{3.31}$$

*R_a^2 is sometimes used to **compare** models having **different** numbers of predictor variables. In comparing the goodness of fit of models with **different** numbers of explanatory variables, R_a^2 tries to "**adjust**" for the **unequal** number of variables in the different models. Unlike R^2, R_a^2 **cannot** be interpreted as the proportion of total variation in Y **accounted** for by the predictors. Many regression packages provide values for **both** R^2 and R_a^2 .*

$$\frac{\text{SSE}/(n-p-1)}{\text{SST}/(n-1)}$$

By adding a new predictor variable, p increases to $p+1$ so that $n-p-1$ decreases to $n-p-2$. While SSE also decreases by adding this new variable, their ratio $\frac{SSE}{n-p-1}$ can either increase or decrease depending only how much this new predictor variable decreases SSE.

3.4 Properties of Least Square Estimators and Multiple Correlation Coefficient

Adjusted R-squared on Supervisor Dataset

using the original dataset

```
supervisor_dat<-read.table('data/P060.txt',header=TRUE)   ## read the data
y<-supervisor_dat$Y
X<-as.matrix(supervisor_dat[,-1])
X<-cbind(rep(1,30),X)
colnames(X)<-c("Const.","X1","X2","X3","X4","X5","X6")
hat_beta <- solve(t(X) %*% X) %*% t(X) %*% y          ## LSE by direct computation using the formula
hat_y <- X %*% hat_beta
Rsquared <- sum((hat_y-mean(y))^2)/sum((y-mean(y))^2)
n<-30
p<-6
R2adj<-1-sum((hat_y-y)^2)/sum((y-mean(y))^2)*(n-1)/(n-p-1)
```

```
> R2adj
[1] 0.662846
```

using the centered and standardized dataset

```
> supervisor_center_stand <- apply(supervisor_dat,2,function(x) (x-mean(x))/sd(x))    ## center and standardize the variables
> y_center_stand<-supervisor_center_stand[,1]
> X_center_stand<-as.matrix(supervisor_center_stand[,-1])
> X_center_stand<-cbind(rep(1,30),X_center_stand)
> colnames(X_center_stand)<-c("Const.","X1","X2","X3","X4","X5","X6")
> hat_beta_center_stand <- solve(t(X_center_stand) %*% X_center_stand) %*% t(X_center_stand) %*% y_center_stand       ## LSE by direct computation using the formula
> hat_y <- X_center_stand %*% hat_beta_center_stand
> Rsquared <- sum((hat_y-mean(y_center_stand))^2)/sum((y_center_stand-mean(y_center_stand))^2)
> R2adj<-1-sum((hat_y-y_center_stand)^2)/sum((y_center_stand-mean(y_center_stand))^2)*(n-1)/(n-p-1)
> R2adj
[1] 0.662846
```

3.5. Inference for Individual Regression Coefficients

3.5 Inference for Individual Regression Coefficients

Testing β_j

Using the properties of the least squares estimators discussed, one can make statistical inference regarding the regression coefficients.

$H_0 : \beta_j = \beta_j^0$ versus $H_1 : \beta_j \neq \beta_j^0$ where β_j^0 is a constant chosen by the investigator

The test statistic is

$$t_j = \frac{\hat{\beta}_j - \beta_j^0}{\text{s.e.}(\hat{\beta}_j)}, \qquad (3.32)$$

$(c_{ij})_{i,j=1}^{p} = \mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}$

where $s.e.(\hat{\beta}_j) = \hat{\sigma}\sqrt{c_{jj}}$

*which has a Student's t-distribution with n-p-**1degrees** of freedom. The test is carried out by **comparing** the **observed** value with the appropriate **critical value** $t_{(n-p-1,\alpha/2)}$, where α is the significance level. Note that we divide the significance level α by 2 because we have a **two-sided** alternative hypothesis.*

Accordingly, H_0 is to be rejected at the significance level α if

$$|t_j| \geq t_{(n-p-1,\alpha/2)}, \qquad (3.33)$$

where $|t_j|$ denotes the absolute value of t_j.

3.5 Inference for Individual Regression Coefficients

Testing β_j

Using p-value A criterion equivalent to that in (3.33) is to compare the p-value of the test with α and reject H_0 if

$$p(|t_j|) \leq \alpha, \tag{3.34}$$

where $p(|t_j|)$, is the *p-value* of the test, which is the probability that a random variable having a Student t-distribution, with $n-p-1$, is greater than $|t_j|$ (the absolute value of the observed value of the t-Test); The p-value is usually computed and supplied as part of the regression output by many statistical packages.

A special case The usual test is for $H_0 : \beta_j^0 = 0$, in which case the t-Test reduces to

$$t_j = \frac{\hat{\beta}_j}{\text{s.e.}(\hat{\beta}_j)}, \tag{3.35}$$

Note that the rejection of H_0: $\beta_j = 0$ would mean that β_j is likely to be different from 0, and hence the predictor variable X_j is a statistically significant predictor of the response variable Y after adjusting for the other predictor variables.

Confidence interval:

$$\hat{\beta}_j \pm t_{(n-p-1,\alpha/2)} \times \text{s.e.}(\hat{\beta}_j), \tag{3.36}$$

3.5 Inference for Individual Regression Coefficients

Example on Supervisor Dataset

$\hat{\beta}$

```
> hat_beta
              [,1]
Const. 10.78707639
X1      0.61318761
X2     -0.07305014
X3      0.32033212
X4      0.08173213
X5      0.03838145
X6     -0.21705668
```

```
> n<-30
> p<-6
> hat_sig2<-sum((hat_y-y)^2)/(n-p-1)
> hat_sig2
[1] 49.95654
```

$\hat{\sigma}^2$

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

```
> solve(t(X) %*% X)
              Const.            X1            X2            X3            X4            X5            X6
Const.  2.6885547839 -2.543299e-03 -3.633619e-03 -6.153588e-03 -4.329621e-03 -2.266050e-02  6.562191e-04
X1     -0.0025432993  5.187622e-04 -1.636348e-04 -1.638338e-04 -3.717724e-04  4.677579e-07  2.309832e-04
X2     -0.0036336194 -1.636348e-04  3.687444e-04 -6.251148e-05  4.616729e-05 -9.156320e-06 -8.936448e-05
X3     -0.0061535881 -1.638338e-04 -6.251148e-05  5.684761e-04 -1.539315e-04  9.702610e-05 -2.087411e-04
X4     -0.0043296211 -3.717724e-04  4.616729e-05 -1.539315e-04  9.819008e-04 -1.827734e-04 -3.373815e-04
X5     -0.0226604991  4.677579e-07 -9.156320e-06  9.702610e-05 -1.827734e-04  4.325292e-04 -6.705033e-05
X6      0.0006562191  2.309832e-04 -8.936448e-05 -2.087411e-04 -3.373815e-04 -6.705033e-05  6.357249e-04
```

computing $\frac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)}$ **directly**

```
> digX<-diag(solve(t(X)%*% X))
> hat_beta/sqrt(hat_sig2*digX)
              [,1]
Const.  0.9307824
X1      3.8090182
X2     -0.5382229
X3      1.9008516
X4      0.3690310
X5      0.2611064
X6     -1.2179862
```

3.5 Inference for Individual Regression Coefficients

Example on Supervisor Dataset

Use R to automatically obtain the confidence intervals of coefficient parameters

```
> ##### Confidence intervals of regression coefficients
> supervisor_dat<-read.table('data/P060.txt',header=TRUE)   ## read the data
> fmodel<-lm(Y~.,data=supervisor_dat)
> confint(fmodel,level=0.95)
                   2.5 %     97.5 %
(Intercept) -13.18712881 34.7612816
X1            0.28016866  0.9462066
X2           -0.35381806  0.2077178
X3           -0.02827872  0.6689430
X4           -0.37642935  0.5398936
X5           -0.26570179  0.3424647
X6           -0.58571106  0.1515977
```

Can you derive the above results by direction computation using the formulas?

3.5 Inference for Individual Regression Coefficients

Example: Supervisor Performance Data

Let us now illustrate the above t-Tests using the Supervisor Performance data set described earlier in this chapter. The results of fitting a linear regression model relating Y and the six explanatory variables are given in Table 3.5. The fitted regression equation is

$$\hat{Y} = 10.787 + 0.613X_1 - 0.073X_2 + 0.320X_3 + 0.081X_4 + 0.038X_5 - 0.217X_6. \quad (3.37)$$

The t-values in Table 3.5 test the null hypothesis $H_0 : \beta_j = 0, j = 0, 1, \cdots, p$, against an alternative $H_1 : \beta_j \neq 0$. From Table 3.5 it is seen that only the regression coefficient of X_1 is significantly different from zero and X_3 has a regression coefficient that approach being significantly different from zero. The other variables have insignificant t-Tests. The construction of confidence intervals for the individual parameters is left as an exercise for the reader.

It should be noted here that the constant in the above model is statistically not significant (t-value of 0.93 and p-value of 0.3616). In any regression model, unless there is strong theoretical reason, a constant should always be included even if the term is statistically not significant. The constant represents the base or background level of the response variable. Insignificant predictors should not be in general retained but a constant should be retained.

```
### use lm function to obtain the regression table
supervisor_model<-lm(Y ~ . , data = supervisor_dat)
```

```
> summary(supervisor_model)

Call:
lm(formula = Y ~ ., data = supervisor_dat)

Residuals:
     Min       1Q   Median       3Q      Max
-10.9418  -4.3555   0.3158   5.5425  11.5990

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.78708   11.58926   0.931 0.361634
X1           0.61319    0.16098   3.809 0.000903 ***
X2          -0.07305    0.13572  -0.538 0.595594
X3           0.32033    0.16852   1.901 0.069925 .
X4           0.08173    0.22148   0.369 0.715480
X5           0.03838    0.14700   0.261 0.796334
X6          -0.21706    0.17821  -1.218 0.235577
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.068 on 23 degrees of freedom
Multiple R-squared:  0.7326,    Adjusted R-squared:  0.6628
F-statistic:  10.5 on 6 and 23 DF,  p-value: 1.24e-05
```

Table 3.5 Regression Output for Supervisor Performance Data

Variable	Coefficient	s.e.	t-Test	p-value
Constant	10.787	11.5890	0.93	0.3616
X_1	0.613	0.1610	3.81	0.0009
X_2	−0.073	0.1357	−0.54	0.5956
X_3	0.320	0.1685	1.90	0.0699
X_4	0.081	0.2215	0.37	0.7155
X_5	0.038	0.1470	0.26	0.7963
X_6	−0.217	0.1782	−1.22	0.2356
$n = 30$	$R^2 = 0.73$	$R_a^2 = 0.66$	$\hat{\sigma} = 7.068$	df = 23

Regression

3.6. Test of Hypothesis in Linear Model and Prediction

3.6 Test of Hypothesis in Linear Model and Prediction

General Framework

In addition to looking at hypotheses about individual β's, several different hypotheses are considered in connection with the analysis of linear models. The most commonly investigated hypotheses are

1. All the regression coefficients associated with the predictor variables are zero.
2. Some of the regression coefficients are zero.
3. Some of the regression coefficients are equal to each other.
4. The regression parameters satisfy certain specified constraints.

The different hypotheses about the regression coefficients can all be tested in the same way by a unified approach. Rather than describing the individual tests, we first describe the general unified approach, then illustrate specific tests using the Supervisor Performance data.

3.6 Test of Hypothesis in Linear Model and Prediction

Example on Supervisor Dataset

Table 3.2 Description of Variables in Supervisor Performance Data

Variable	Description
Y	Overall rating of job being done by supervisor
X_1	Handles employee complaints
X_2	Does not allow special privileges
X_3	Opportunity to learn new things
X_4	Raises based on performance
X_5	Too critical of poor performance
X_6	Rate of advancing to better jobs

$$H_0 : Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \text{ V.S. } H_1 : Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 x_4 + \beta_5 X_5 + \beta_6 X_6 + \varepsilon$$

$$H_0 : Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 x_4 + \beta_5 X_5 + \varepsilon \text{ V.S. } H_1 : Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 x_4 + \beta_5 X_5 + \beta_6 X_6 + \varepsilon$$

In these two examples, H_0 is a reduced model and H_1 is a full model.

3.6 Test of Hypothesis in Linear Model and Prediction

General Framework

*The model given in (3.1) will be referred to as the **full model** (FM). The null hypothesis to be tested specifies values for **some** of the regression coefficients. When these values are substituted in the full model, the resulting model is called the **reduced model** (RM). The number of distinct parameters to be estimated in the reduced model is **smaller** than the number of parameters to be estimated in the **full** model. Accordingly, we wish to test*

$$\mathrm{H}_0 : \text{Reduced model is adequate} \quad \text{against} \quad \mathrm{H}_1 : \text{Full model is adequate.}$$

*Note that the reduced model is **nested**. A set of models are said to be nested if they can be obtained from a **larger** model as special cases. The test for these nested hypotheses involves a comparison of the goodness of fit that is obtained when using the full model, to the **goodness of fit** that results using the reduced model specified by the **null hypothesis**. If the reduced model gives as good a fit as the full model, the **null** hypothesis, which defines the reduced model (by specifying some values β_j), is not **rejected**. This procedure is described formally as follows.*

Radj^2

$$\text{Model 1: } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$
$$\text{Model 2: } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$
$$\text{Model 3: } Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

Next page

3.6 Test of Hypothesis in Linear Model and Prediction

General Framework

F test

Let $\hat{y}_i$ and $\hat{y}_i^*$ be the values predicted for y_i by the full model and the reduced model, respectively. The lack of fit in the data associated with the full model is the sum of the squared residuals obtained when fitting the full model to the data. We denote this by SSE(FM), the sum of squares due to error associated with the full model,

$$\text{SSE(FM)} = \sum (y_i - \hat{y}_i)^2. \tag{3.38}$$

Similarly, the lack of fit in the data associated with the reduced model is the sum of the squared residuals obtained when fitting the reduced model to the data. This quantity is denoted by SSE(RM), the sum of squares due to error associated with the reduced model,

$$\text{SSE(RM)} = \sum (y_i - \hat{y}_i^*)^2. \tag{3.39}$$

In the full model there are $p + 1$ regression parameters $(\beta_0, \beta_1, \beta_2, \cdots, \beta_p)$ to be estimated. Let us suppose that for the reduced model there are k distinct parameters. Note that SSE(RM) $\geq$ SSE(FM) because the additional parameters (variables) in the full model cannot increase the residual sum of squares. Note also that the difference SSE(RM) – SSE(FM) represents the increase in the residual sum of squares due to fitting the reduced model. If this difference is large, the reduced model is inadequate. To see whether the reduced model is adequate, we use the ratio

$$F = \frac{[\text{SSE(RM)} - \text{SSE(FM)}]/(p + 1 - k)}{\text{SSE(FM)}/(n - p - 1)}. \tag{3.40}$$

*The ratio is called the **F Test.***

3.6 Test of Hypothesis in Linear Model and Prediction

General Framework

F ratio

Note that we divide SSE(RM) - SSE(FM) and SSE(FM) in the ratio by their respective degrees of freedom to ***compensate*** *for the different number of parameters involved in the two models as well as to* ***ensure*** *that the resulting test statistic has a standard statistical* ***distribution****. The* ***full*** *model has p+1 parameters, hence SSE(FM) has n-p-1 degrees of freedom. Similarly, the* ***reduced*** *model has k parameters and SSE(RM) has n-k degrees of freedom. Consequently, the* ***difference*** *SSE(RM) - SSE(FM) has (n-k)-(n-p-1)= p + 1 - k degrees of freedom. Therefore, the observed* ***F-ratio*** *in (3.40) has* ***F-distribution*** *with p+1-k and n-p-1 degrees of freedom.*

F test

If the observed F-value is large in comparison to the tabulated value of F with $p+1-k$ and $n-p-1$ degrees of freedom, the result is significant at level α; that is, the reduced model is unsatisfactory and the null hypothesis, with its suggested values of β's in the full model is rejected.

Accordingly, H_0 is rejected if

$$F \geq F_{(p+1-k,n-p-1;\alpha)}, \tag{3.41}$$

or, equivalently, if

$$p(F) \leq \alpha, \tag{3.42}$$

where F is the observed value of the F-Test in (3.40), $F_{(p+1-k,n-p-1;\alpha)}$ is the appropriate critical value obtained from the F table, α is the significance level, and $p(F)$ is the *p- value for the F -Test, which is the probability that a random variable having an F -distribution, with p + 1 - k and n - p - 1 degrees of freedom, is greater than the observed F-test in (3.40)*

3.6 Test of Hypothesis in Linear Model and Prediction

Intuitive Understanding of F-test

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \underbrace{1}_{\mathbf{x}_1} & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

In the full model, all columns of $\mathbf{X}$ are used for regression while in the reduced model only a subset of columns are used. For simplicity, suppose only the first k columns are used. Denote the first k columns of $\mathbf{X}$ by $\mathbf{X}_1$.

By the results in Chapter 3-proof.pdf, we know

$$\text{SSE(RM)} = \mathbf{y}'(\mathbf{I} - \mathbf{H}_1)\mathbf{y} \quad \text{where } \mathbf{H}_1 = \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$$

and

$$\text{SSE(FM)} = \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y} \quad \text{where } \mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

so $\text{SSE(RM)} - \text{SSE(FM)} = \mathbf{y}'(\mathbf{H} - \mathbf{H}_1)\mathbf{y}$. If null hypothesis is true, we have $\text{SSE(RM)}/\sigma^2$ follows a chi-squared distribution with degrees of freedom $n - k$.

The ratio of scaled and independent chi-squared random variables follows an F distribution, See Chapter 1.

3.6 Test of Hypothesis in Linear Model and Prediction

Example on Supervisor Dataset

We test $H_0 : Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ versus $H_1 : Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_6 X_6 + \varepsilon$ on the *supervisor dataset*.

```
######################## Test of Hypothesis for Linear Models
##  H0: Y ~ X_1 + X_2
##  H1: full model
supervisor_dat<-read.table('data/P060.txt',header=TRUE)   ## read the data
y<-supervisor_dat$Y
X<-as.matrix(supervisor_dat[,-1])
X<-cbind(rep(1,30),X)
colnames(X)<-c("Const.","X1","X2","X3","X4","X5","X6")
hat_beta <- solve(t(X) %*% X) %*% t(X) %*% y
haty <- X %*% hat_beta
SSE_fm  <- sum((y-haty)^2)     ## SSE for full model
# now fit the reduced model
X1<-X[,c(1,2,3)]
hat_beta_rm <- solve(t(X1) %*% X1) %*% t(X1) %*% y
haty_rm <- X1 %*% hat_beta_rm
SSE_rm <- sum((y-haty_rm)^2)
n<-30;p<-6;k<-3;
F_val <- (SSE_rm-SSE_fm)/SSE_fm*(n-p-1)/(p+1-k)
```

```
> F_val
[1] 1.065244
```

By direct computation

```
### automatically using anova
full_mod <- lm(Y~., data=supervisor_dat)
red_mod <- lm(Y~X1+X2, data=supervisor_dat)
anova(red_mod,full_mod)
```

Use anova function in R

```
> anova(red_mod,full_mod)
Analysis of Variance Table

Model 1: Y ~ X1 + X2
Model 2: Y ~ X1 + X2 + X3 + X4 + X5 + X6
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     27 1361.9
2     23 1149.0  4    212.86 1.0652 0.3962
```

3.6 Test of Hypothesis in Linear Model and Prediction

Class Discussion

When we fit a multiple linear model, say $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$, we obtain the SSE and SSR whose d.f. is $n - p - 1$ and p, respectively. Then, we can define MSE $= \text{SSE}/n - p - 1$ and MSR $= \text{SSR}/p$, respectively. We can use the F-statistic: $\frac{\text{MSR}}{\text{MSE}}$ to test the hypothesis

$$H_0 : Y = \beta_0 + \varepsilon \quad \text{v.s.} \quad H_1 : Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

Why is it for testing this hypothesis ?

3.6 Test of Hypothesis in Linear Model and Prediction

Testing All Regression Coefficients Equal to Zero

*An important **special** case of the F-Test in (3.40) is obtained when we test the hypothesis that **all** predictor variables under consideration have no explanatory power and that all their regression coefficients are **zero**. In this case, the **reduced** and **full** models become*

$$\text{RM:} \quad \text{H}_0 : Y = \beta_0 + \varepsilon, \tag{3.43}$$

$$\text{FM:} \quad \text{H}_1 : Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon. \tag{3.44}$$

The residual sum of squares from the full model is SSE(FM) = SSE. Because the least squares estimate of β_0 in the reduced model is $\bar{y}$, the residual sum of squares from the reduced model is $\text{SSE(RM)} = \sum(y_i - \bar{y})^2 = \text{SST}$. The reduced model has one regression parameter and the full model has $p + 1$ regression parameters. Therefore, the F-Test in (3.40) reduces to

$$\begin{aligned} F &= \frac{[\text{SSE(RM)} - \text{SSE(FM)}]/(p + 1 - k)}{\text{SSE(FM)}/(n - p - 1)} \\ &= \frac{[\text{SST} - \text{SSE}]/p}{\text{SSE}/(n - p - 1)}. \end{aligned} \tag{3.45}$$

Because SST = SSR + SSE, we can replace SST − SSE in the above formula by SSR and obtain

$$F = \frac{\text{SSR}/p}{\text{SSE}/(n - p - 1)} = \frac{\text{MSR}}{\text{MSE}}, \tag{3.46}$$

where MSR is the *mean square due to regression* and MSE is the *mean square due to error*. The F-Test in (3.46) can be used for testing the hypothesis that the regression coefficients of all predictor variables (excluding the constant) are zero.

3.6 Test of Hypothesis in Linear Model and Prediction

Testing All Regression Coefficients Equal to Zero

The F-Test in (3.46) can also be expressed directly in terms of the sample multiple correlation coefficient. The null hypothesis which tests whether all the population regression coefficients are zero is equivalent to the hypothesis that states that the population multiple correlation coefficient is zero. Let R_p denote the sample multiple correlation coefficient, which is obtained from fitting a model to n observations in which there are p predictor variables (i.e., we estimate p regression coefficients and one intercept). The appropriate F for testing

$$\mathrm{H}_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

$R_p = \mathrm{Cor}(Y, \hat{Y})$

in terms of R_p is

$$F = \frac{R_p^2/p}{(1 - R_p^2)/(n - p - 1)}, \tag{3.47}$$

with p and $n - p - 1$ degrees of freedom.

$$F = \frac{\mathrm{SSR}/p}{\mathrm{SSE}/(n-p-1)} = \frac{(\mathrm{SSR}/\mathrm{SST})/p}{(\mathrm{SSE}/\mathrm{SST})/(n-p-1)} \overset{\text{definition of } R^2}{=} \frac{R^2/p}{(1-R^2)/(n-p-1)}$$

3.6 Test of Hypothesis in Linear Model and Prediction

Testing All Regression Coefficients Equal to Zero

$H_0 : \beta_1 = \cdots = \beta_p = 0$

Table 3.6 Analysis of Variance (ANOVA) Table in Multiple Regression

Source	Sum of Squares	df	Mean Square	F-Test
Regression	SSR	p	$\text{MSR} = \frac{\text{SSR}}{p}$	$F = \frac{\text{MSR}}{\text{MSE}}$
Residuals	SSE	$n-p-1$	$\text{MSE} = \frac{\text{SSE}}{n-p-1}$	

The values involved in the above F-Test are customarily computed and compactly displayed in a table called the *analysis of variance* (ANOVA) table. The ANOVA table is given in Table 3.6. The first column indicates that there are two sources of variability in the response variable Y. The total variability in Y, $\text{SST} = \sum(y_i - \bar{y})^2$, can be decomposed into two sources: the *explained* variability, $\text{SSR} = \sum(\hat{y}_i - \bar{y})^2$, which is the variability in Y that can be accounted for by the predictor variables, and the *unexplained* variability, $\text{SSE} = \sum(y_i - \hat{y}_i)^2$. This is the same decomposition $\text{SST} = \text{SSR} + \text{SSE}$. This decomposition is given under the column heading Sum of Squares. The third column gives the degrees of freedom (df) associated with the sum of squares in the second column. The fourth column is the Mean Square (MS), which is obtained by dividing each sum of squares by its respective degrees of freedom. Finally, the F-Test in (3.46) is reported in the last column of the table. Some statistical packages also give an additional column containing the corresponding p-value, $p(F)$.

3.6 Test of Hypothesis in Linear Model and Prediction

Testing All Regression Coefficients Equal to Zero

Example (Supervisor Performance Data)

*Returning now to the Supervisor Performance data, although the t-Tests for the **regression** coefficients have already indicated that some of the regression coefficients (β_1 and β_3) are significantly **different** from zero, we will, for illustrative purposes, test the hypothesis that **all six** predictor variables have no explanatory power, that is, $\beta_1 = \beta_2 = \cdots = \beta_6 = 0$ In this case, the **reduced** and **full** models in (3.43) and (3.44) become*

$$\text{RM:} \quad \text{H}_0 : Y = \beta_0 + \varepsilon, \tag{3.48}$$

$$\text{FM:} \quad \text{H}_1 : Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_6 X_6 + \varepsilon. \tag{3.49}$$

*For the full model we have to estimate seven parameters, six regression coefficients and an intercept term β_0 . The **ANOVA** table is given in Table 3.7. The sum of squares due to error in the full model is SSE(PM) = SSE = 1149. Under the **null** hypothesis, where all the $\beta's$ are zero, the number of parameters estimated for the **reduced** model is therefore 1 (β_0). Consequently, the sum of squares of the residuals in the **reduced** model is*

$$\text{SSE(RM)} = \text{SST} = \text{SSR} + \text{SSE} = 3147.97 + 1149 = 4296.97.$$

Table 3.7 Supervisor Performance Data: Analysis of Variance (ANOVA) Table

Source	Sum of Squares	df	Mean Square	F-Test
Regression	3147.97	6	524.661	10.5
Residuals	1149.00	23	49.9565	

*This F-value has an F-distribution with 6 and 23 degrees of freedom. The 1% F-value with 6 and 23 degrees of freedom is found to be 3.71. the observed F -value is larger than this value, the null hypothesis is **rejected**; not all the $\beta's$ can be taken as zero.*

3.6 Test of Hypothesis in Linear Model and Prediction

Use R to derive the results of previous slide

```
##  H0: Y = beta_0
##  H1: full model
supervisor_dat<-read.table('data/P060.txt',header=TRUE)   ## read the data
y<-supervisor_dat$Y
X<-as.matrix(supervisor_dat[,-1])
X<-cbind(rep(1,30),X)
colnames(X)<-c("Const.","X1","X2","X3","X4","X5","X6")
hat_beta <- solve(t(X) %*% X) %*% t(X) %*% y
haty <- X %*% hat_beta
SSE_fm  <- sum((y-haty)^2)     ## SSE for full model
# now fit the reduced model
X1<-X[,1]
hat_beta_rm <- solve(t(X1) %*% X1) %*% t(X1) %*% y
haty_rm <- X1 %*% hat_beta_rm
SSE_rm <- sum((y-haty_rm)^2)
n<-30;p<-6;k<-1;
F_val <- (SSE_rm-SSE_fm)/SSE_fm*(n-p-1)/(p+1-k)
```

```
> F_val
[1] 10.50235
```

By direct computation

```
### automatically using anova
full_mod <- lm(Y~., data=supervisor_dat)
red_mod <- lm(Y~1, data=supervisor_dat)
anova(red_mod,full_mod)
```

Use anova function in R

Note that the arrangement in this table is slightly different from the canonical form of anova table

```
> anova(red_mod,full_mod)
Analysis of Variance Table

Model 1: Y ~ 1
Model 2: Y ~ X1 + X2 + X3 + X4 + X5 + X6
  Res.Df  RSS Df Sum of Sq      F   Pr(>F)
1     29 4297
2     23 1149  6      3148 10.502 1.24e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

3.6 Test of Hypothesis in Linear Model and Prediction

Testing a Subset of Regression Coefficients Equal to Zero

We have so far attempted to explain Y in the Supervisor Performance data, in terms of six variables, $X_1, X_2, \cdots X_6$. The F-Test in (3.46) indicates that all the regression coefficients cannot be taken as zero, hence one or more of the predictor variables is related to Y. The question of interest now is: Can Y be explained adequately by fewer variables? An important goal in regression analysis is to arrive at adequate descriptions of observed phenomenon in terms of as few meaningful variables as possible. This economy in description has two advantages. First, it enables us to isolate the most important variables, and second, it provides us with a simpler description of the process studied, thereby making it easier to understand the process. *Simplicity of description* or the *principle of parsimony*, as it is sometimes called, is one of the important guiding principles in regression analysis.

3.6 Test of Hypothesis in Linear Model and Prediction

Testing a Subset of Regression Coefficients Equal to Zero

To examine whether the variable Y can be explained in terms of fewer variables, we look at a hypothesis that specifies that some of the regression coefficients are zero. If there are no overriding theoretical considerations as to which variables are to be included in the equation, preliminary t-Tests, like those given in Table 3.5, are used to suggest the variables. In our current example, suppose it was desired to explain the overall rating of the job being done by the supervisor by means of two variables, one taken from the group of personal employee interaction variables X_1, X_2, X_5, and another taken from the group of variables X_3, X_4, X_6, which are of a less personal nature. From this point of view X_1 and X_3 suggest themselves because they have significant t-Tests. Suppose then that we wish to determine whether Y can be explained by X_1 and X_3 as adequately as the full set of six variables. The reduced model in this case is

$$\text{RM: } Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon. \tag{3.50}$$

This model corresponds to hypothesis

$$\text{H}_0 : \beta_2 = \beta_4 = \beta_5 = \beta_6 = 0. \tag{3.51}$$

The regression output from fitting this model is given in Table 3.8, which includes both the ANOVA and the coefficients tables.

Table 3.8

3.6 Test of Hypothesis in Linear Model and Prediction

Testing a Subset of Regression Coefficients Equal to Zero

Table 3.8 Regression Output from the Regression of Y on X_1 and X_3

ANOVA Table

Source	Sum of Squares	df	Mean Square	F-Test
Regression	3042.32	2	1521.1600	32.7
Residuals	1254.65	27	46.4685	

Coefficients Table

Variable	Coefficient	s.e.	t-Test	p-value
Constant	9.8709	7.0610	1.40	0.1735
X_1	0.6435	0.1185	5.43	< 0.0001
X_3	0.2112	0.1344	1.57	0.1278
$n = 30$	$R^2 = 0.708$	$R_a^2 = 0.686$	$\hat{\sigma} = 6.817$	df = 27

This anova table is to test

$H_0 : Y = \beta_0 + \varepsilon$

v.s. $H_1 : Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon$

The residual sum of squares in this output is the residual sum of squares for the reduced model, which is SSE(RM) = 1254.65. From Table 3.7, the residual sum of squares from the full model is SSE(FM) = 1149.00. Hence the F-Test in (3.40) is

$$F = \frac{[1254.65 - 1149]/4}{1149/23} = 0.528, \tag{3.52}$$

with 4 and 23 degrees of freedom.

The corresponding tabulated value for this test is $F_{(4,23,0.05)} = 2.8$. The value of F is not significant and the null hypothesis is not rejected. The variables X_1 and X_3 together explain the variation in Y as adequately as the full set of six variables. We conclude that the deletion of X_2, X_4, X_5, X_6 does not adversely affect the explanatory power of the model.

3.6 Test of Hypothesis in Linear Model and Prediction

Use R to derive the results on last slide

```
##########################################  Example of Table 3.8 on lecture slides
##  H0: Y = beta_0
##  H1: Y ~ 1+X1+X3
########## obtain the anova table by direct computation
supervisor_dat<-read.table('data/P060.txt',header=TRUE)   ## read the data
y<-supervisor_dat$Y
X<-as.matrix(supervisor_dat[,c(2,4)])
X<-cbind(rep(1,30),X)
colnames(X)<-c("Const.","X1","X3")
hat_beta <- solve(t(X) %*% X) %*% t(X) %*% y
haty <- X %*% hat_beta
SSE <- sum((y-haty)^2)
SST <- sum((y-mean(y))^2)
SSR <- SST-SSE
Fval<-SSR/SSE*27/2
```

```
> Fval
[1] 32.73528
```

```
> ### automaticlally obtain the anova table
> red_mod<-lm(Y~X1+X3, data=supervisor_dat)
> null_mod<-lm(Y~1,data=supervisor_dat)
> anova(null_mod,red_mod)
Analysis of Variance Table

Model 1: Y ~ 1
Model 2: Y ~ X1 + X3
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     29 4297.0
2     27 1254.6  2    3042.3 32.735 6.058e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> summary(red_mod)

Call:
lm(formula = Y ~ X1 + X3, data = supervisor_dat)

Residuals:
     Min       1Q   Median       3Q      Max
-11.5568  -5.7331   0.6701   6.5341  10.3610

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.8709     7.0612   1.398    0.174
X1            0.6435     0.1185   5.432 9.57e-06 ***
X3            0.2112     0.1344   1.571    0.128
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.817 on 27 degrees of freedom
Multiple R-squared:  0.708,     Adjusted R-squared:  0.6864
F-statistic: 32.74 on 2 and 27 DF,  p-value: 6.058e-08
```

3.6 Test of Hypothesis in Linear Model and Prediction

Testing a Subset of Regression Coefficients Equal to Zero

Remarks

1. The F-Test in this case can also be expressed in terms of the sample multiple correlation coefficients. Let R_p denote the sample multiple correlation coefficient that is obtained when the full model with all the p variables in it is fitted to the data. Let R_q denote the sample multiple correlation coefficient when the model is fitted with q specific variables: that is, the null hypothesis states that $p - q$ specified variables have zero regression coefficients. The F-Test for testing the above hypothesis is

$$F = \frac{(R_p^2 - R_q^2)/(p-q)}{(1-R_p^2)/(n-p-1)}, \quad \text{df} = p-q \text{ and } n-p-1. \qquad (3.53)$$

In our present example, from Tables 3.7 and 3.8, we have $n = 30, p = 6, q = 2, R_6^2 = 0.7326$, and $R_2^2 = 0.7080$. Substituting these in (3.53) we get an F-value of 0.528, as before.

More remarks

3.6 Test of Hypothesis in Linear Model and Prediction

Testing a Subset of Regression Coefficients Equal to Zero

Remarks

2. When the reduced model has only one coefficient (predictor variable) less than the full model, say β_j, then the F-Test in (3.40) has 1 and $n-p-1$ degrees of freedom. In this case, it can be shown that the F-Test in (3.40) is equivalent to the t-Test in (3.33). More precisely, we have

$$F = t_j^2, \tag{3.54}$$

which indicates that an F-value with 1 and $n-p-1$ degrees of freedom is equal to the square of a t-value with $n-p-1$ degrees of freedom, a result which is well-known in statistical theory.

More remarks

3.6 Test of Hypothesis in Linear Model and Prediction

Testing a Subset of Regression Coefficients Equal to Zero

Remarks

3. In simple regression the number of predictors is $p = 1$. Replacing p by one in the multiple regression ANOVA table (Table 3.6) we obtain the simple regression ANOVA table (Table 3.9). The F-Test in Table 3.9 tests the null hypothesis that the predictor variable X_1 has no explanatory power, that is, its regression coefficient is zero. But this is the same hypothesis tested by the t_1-Test introduced in *Chapter 2* and defined in (2.26) as

$$t_1 = \frac{\hat{\beta}_1}{\text{s.e.}(\hat{\beta}_1)} . \tag{3.55}$$

Therefore in simple regression, the F and t_1 tests are equivalent, they are related by

$$F = t_1^2. \tag{3.56}$$

Table 3.9 Analysis of Variance (ANOVA) Table in Simple Regression

Source	Sum of Squares	df	Mean Square	F-Test
Regression	SSR	1	MSR = SSR	$F = \frac{\text{MSR}}{\text{MSE}}$
Residuals	SSE	$n-2$	MSE = $\frac{\text{SSE}}{n-2}$	

3.6 Test of Hypothesis in Linear Model and Prediction

Testing the Equality of Regression Coefficients

It is possible to test the equality of two or more regression coefficients in the same model. In the present example we test whether the regression coefficient of the variables X_1 and X_3 can be treated as equal. The test is performed assuming that it has already been established that the regression coefficients for X_2, X_4, X_5, and X_6 are zero. The null hypothesis to be tested is

$$\mathrm{H}_0 : \beta_1 = \beta_3 \mid (\beta_2 = \beta_4 = \beta_5 = \beta_6 = 0). \tag{3.57}$$

The full model assuming that $\beta_2 = \beta_4 = \beta_5 = \beta_6 = 0$ is

$$Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon. \tag{3.58}$$

Under the null hypothesis, where $\beta_1 = \beta_3 = \beta_1'$, say, the reduced model is

$$Y = \beta_0' + \beta_1'(X_1 + X_3) + \varepsilon. \tag{3.59}$$

A simple way to carry out the test is to fit the model given by (3.58) to the data. The resulting regression output has been given in Table 3.8. We next fit the reduced model given in (3.59). This can be done quite simply by generating a new variable $W = X_1 + X_3$ and fitting the model

$$Y = \beta_0' + \beta_1' W + \varepsilon. \tag{3.60}$$

The least squares estimates of β_0', β_1' and the sample multiple correlation coefficient (in this case it is the simple correlation coefficient between Y and W since we have only one variable) are obtained. The fitted equation is

$$\hat{Y} = 9.988 + 0.444W$$

with $R_1^2 = 0.6685$.

3.6 Test of Hypothesis in Linear Model and Prediction

Testing the Equality of Regression Coefficients

The appropriate F for testing the null hypothesis, defined in (3.53), becomes

$$F = \frac{(R_p^2 - R_q^2)/(p-q)}{(1-R_p^2)/(n-p-1)} = \frac{(0.7080 - 0.6685)/(2-1)}{(1-0.7080)/(30-2-1)} = 3.65,$$

with 1 and 27 degrees of freedom. The tabulated value is $F_{(1,27,0.05)} = 4.21$. The resulting F is not significant; the null hypothesis is not rejected. The distribution of the residuals for this equation (not given here) was found satisfactory.

The equation $\hat{Y} = 9.988 + 0.444(X_1 + X_3)$ is not inconsistent with the given data. We conclude then that X_1 and X_3 have the same incremental effect in determining employee satisfaction with a supervisor. This test could also be performed by using a t-Test, given by

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_3}{\text{s.e.}(\hat{\beta}_1 - \hat{\beta}_3)}$$

with 27 degrees of freedom. The conclusions are identical and follow from the fact that F with 1 and p degrees of freedom is equal to the square of t with p degrees of freedom.

3.6 Test of Hypothesis in Linear Model and Prediction

Use R to derive results for testing (3.57)

```
#####################################  Testing (3.57) on lecture slides
## H_0: Y=beta0+beta1 X1 + beta1 X3 +eps
## H_1: Y=beta0+beta1 X1 + beta2 X2 + eps
supervisor_dat<-read.table('data/P060.txt',header=TRUE)   ## read the data
y<-supervisor_dat$Y
X<-as.matrix(supervisor_dat[,c(2,4)])
X<-cbind(rep(1,30),X)
colnames(X)<-c("Const.","X1","X3")
hat_beta_fm <- solve(t(X) %*% X) %*% t(X) %*% y
haty_fm <- X %*% hat_beta_fm
SSE_fm <- sum((y-haty_fm)^2)

Xrm <- X[,2]+X[,3]
Xrm <- cbind(rep(1,30),Xrm)
colnames(Xrm) <- c("Const.","X1+X3")
hat_beta_rm <- solve(t(Xrm) %*% Xrm) %*% t(Xrm) %*% y
haty_rm <- Xrm %*% hat_beta_rm
SSE_rm <- sum((y-haty_rm)^2)
```

```
> hat_beta_fm
             [,1]
Const. 9.8708805
X1     0.6435176
X3     0.2111918
> hat_beta_rm
             [,1]
Const. 9.9882136
X1+X3  0.4443897
```

```
> n<-30
> p<-2
> k<-2
> Fval <- (SSE_rm-SSE_fm)/SSE_fm*(n-p-1)/(p+1-k)
> Fval
[1] 3.657224
```

3.6 Test of Hypothesis in Linear Model and Prediction

Testing the Equality of Regression Coefficients

In this example we have discussed a sequential or step-by-step approach to model building. We have discussed the equality of β_1 and β_3 under the assumption that the other regression coefficients are equal to zero. We can, however, test a more complex null hypothesis which states that β_1 and β_3 are equal and $\beta_2, \beta_4, \beta_5$, and β_6 are all equal to zero. This null hypothesis H_0' is formally stated as

$$H_0' : \beta_1 = \beta_3, \beta_2 = \beta_4 = \beta_5 = \beta_6 = 0. \qquad (3.61)$$

K = 2

The difference between (3.57) and (3.61) is that in (3.57), $\beta_2, \beta_4, \beta_5$, and β_6 are assumed to be zero, whereas in (3.61) this is under test. The null hypothesis (3.61) can be tested quite easily. The reduced model under H_0' is (3.59), but this model is not compared to the model of equation (3.58), as in the case of H_0, but with the full model with all six variables in the equation. The F-Test for testing H_0' is, therefore,

$$F = \frac{(0.7326 - 0.6685)/5}{0.2674/23} = 1.10, \quad \text{df} = 5 \text{ and } 23.$$

P = 6

The result is insignificant as before. The first test is more sensitive for detecting departures from equality of the regression coefficients than the second test. (Why?)

3.6 Test of Hypothesis in Linear Model and Prediction

Use R to derive results for testing (3.61)

```
#####################################  Testing (3.61) on lecture slides
## H_0: Y=beta0+beta1 X1 + beta1 X3 +eps
## H_1: full model
supervisor_dat<-read.table('data/P060.txt',header=TRUE)   ## read the data
y<-supervisor_dat$Y
X<-as.matrix(supervisor_dat[,-1])
X<-cbind(rep(1,30),X)
colnames(X)<-c("Const.","X1","X2","X3","X4","X5","X6")
hat_beta_fm <- solve(t(X) %*% X) %*% t(X) %*% y
haty_fm <- X %*% hat_beta_fm
SSE_fm <- sum((y-haty_fm)^2)

Xrm <- X[,2]+X[,4]
Xrm <- cbind(rep(1,30),Xrm)
colnames(Xrm) <- c("Const.","X1+X3")
hat_beta_rm <- solve(t(Xrm) %*% Xrm) %*% t(Xrm) %*% y
haty_rm <- Xrm %*% hat_beta_rm
SSE_rm <- sum((y-haty_rm)^2)
```

```
> n<-30
> p<-6
> k<-2
> Fval <- (SSE_rm-SSE_fm)/SSE_fm*(n-p-1)/(p+1-k)
> Fval
[1] 1.103336
```

use R-squared to derive the results

```
> ##### use Rsquared to obtain the results
> fmodel<-lm(Y~.,data=supervisor_dat)
> rmodel<-lm(y~.,data=as.data.frame(cbind(y,Xrm[,-1]),col.names = c("Y","X1+X3")))
> fm_r2<-summary(fmodel)$r.squared
> rm_r2<-summary(rmodel)$r.squared
> Fval<-(fm_r2-rm_r2)/(1-fm_r2)*(n-p-1)/(p+1-k)
> Fval
[1] 1.103336
```

3.6 Test of Hypothesis in Linear Model and Prediction

Prediction

The fitted multiple regression equation can be used to predict the value of the response variable using a set of specific values of the predictor variables, $\mathbf{x}_0 = (x_{01}, x_{02}, \cdots, x_{0p})$. The predicted value, $\hat{y}_0$, corresponding to $\mathbf{x}_0$ is given by

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02} + \cdots + \hat{\beta}_p x_{0p}, \tag{3.63}$$

and its standard error, s.e.$(\hat{y}_0)$, is given, *in the following red formula.* The standard error is usually computed by many statistical packages. Confidence limits for $\hat{y}_0$ with confidence coefficient α are

$$\hat{y}_0 \pm t_{(n-p-1,\alpha/2)}\ \text{s.e.}(\hat{y}_0).$$

Here, $s.e.(\hat{y}_0) = \hat{\sigma}\sqrt{1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}$

As already mentioned in connection with simple regression, instead of predicting the response Y corresponding to an observation $\mathbf{x}_0$ we may want to estimate the mean response corresponding to that observation. Let us denote the mean response at $\mathbf{x}_0$ by μ_0 and its estimate by $\hat{\mu}_0$. Then

$$\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02} + \cdots + \hat{\beta}_p x_{0p},$$

as in (3.63), but its standard error, s.e.$(\hat{\mu}_0)$, is given, *in the following red formula.* Confidence limits for $\hat{\mu}_0$ with confidence coefficient α are

$$\hat{\mu}_0 \pm t_{(n-p-1,\alpha/2)}\text{s.e.}(\hat{\mu}_0).$$

Here, $s.e.(\hat{\mu}_0) = \hat{\sigma}\sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}$

3.6 Test of Hypothesis in Linear Model and Prediction
Example of Prediction on Supervisor Dataset

```
########  Do prediction for Supervisor dataset
## at a new observation x0=(80,42,37,63,51,48)
## we use the full model
##  We begin with direct computation by using the formulas
supervisor_dat<-read.table('data/P060.txt',header=TRUE)   ## read the data
n<-30
p<-6
y<-supervisor_dat$Y
X<-as.matrix(supervisor_dat[,-1])
X<-cbind(rep(1,30),X)
colnames(X)<-c("Const.","X1","X2","X3","X4","X5","X6")
hat_beta_fm <- solve(t(X) %*% X) %*% t(X) %*% y
haty_fm <- X %*% hat_beta_fm
SSE_fm <- sum((y-haty_fm)^2)
hat_sigma<- sqrt(SSE_fm/(n-p-1))

x0<-c(1,80,42,37,63,51,48)  ## at a new observation
haty0<-sum(x0*hat_beta_fm)    ## point prediction
se_y0 <- hat_sigma*sqrt(1+t(x0) %*% solve(t(X)%*% X) %*% x0)  ## standard error
crt_val <- qt(1-0.025,n-p-1)    ## critical value at alpha=0.05
pred_interval <-c(haty0-crt_val*se_y0, haty0+crt_val*se_y0)    ## prediction interval
hatmu0<-sum(x0*hat_beta_fm)    ## point prediction
se_mu0 <- hat_sigma*sqrt(t(x0) %*% solve(t(X)%*% X) %*% x0)  ## standard error
conf_interval <-c(hatmu0-crt_val*se_mu0, hatmu0+crt_val*se_mu0)  ## confidence interval
```

Direct computation by formulas

```
> pred_interval
[1] 44.96006 85.66819
> conf_interval
[1] 51.15409 79.47416
```

Use built-in functions

```
> ###### by using built-in function
> fmodel<-lm(Y~.,data=supervisor_dat)     ## fit a full model
> new_obs <- data.frame(X1=80, X2=42,X3=37,X4=63,X5=51,X6=48)   ## create the new observation x0
> predict(fmodel,newdata=new_obs,interval="prediction",level=0.95)   ## output the prediction interval
       fit      lwr      upr
1 65.31412 44.96006 85.66819
> predict(fmodel,newdata=new_obs,interval="confidence",level=0.95)   ## output the confidence interval
       fit      lwr      upr
1 65.31412 51.15409 79.47416
```

Chapter 3

Summary

We have illustrated the testing of various hypotheses in connection with the linear model. Rather than describing individual tests we have outlined a general procedure by which they can be performed. It has been shown that the various tests can also be described in terms of the appropriate sample multiple correlation coefficients. It is to be emphasized here, that before starting on any testing procedure, the adequacy of the model assumptions should always be examined. As we shall see in Chapter 4, residual plots provide a very convenient graphical way of accomplishing this task. The test procedures are not valid if the assumptions on which the tests are based do not hold. If a new model is chosen on the basis of a statistical test, residuals from the new model should be examined before terminating the analysis. It is only by careful attention to detail that a satisfactory analysis of data can be carried out.

3.7. An Example using R

3.7 An Example Using R

We will once again discuss a dataset with information about cars. This dataset, which can be found at the UCI Machine Learning Repository contains a response variable `mpg` which stores the city fuel efficiency of cars, as well as several predictor variables for the attributes of the vehicles. We load the data, and perform some basic tidying before moving on to analysis.

```
# read the data from the web
autompg = read.table(
  "http://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-mpg.data",
  quote = "\"",
  comment.char = "",
  stringsAsFactors = FALSE)
# give the dataframe headers
colnames(autompg) = c("mpg", "cyl", "disp", "hp", "wt", "acc", "year", "origin", "name")
# remove missing data, which is stored as "?"
autompg = subset(autompg, autompg$hp != "?")
# remove the plymouth reliant, as it causes some issues
autompg = subset(autompg, autompg$name != "plymouth reliant")
# give the dataset row names, based on the engine, year and name
rownames(autompg) = paste(autompg$cyl, "cylinder", autompg$year, autompg$name)
# remove the variable for name, as well as origin
autompg = subset(autompg, select = c("mpg", "cyl", "disp", "hp", "wt", "acc", "year"))
# change horsepower from character to numeric
autompg$hp = as.numeric(autompg$hp)
# check final structure of data
str(autompg)
```

3.7 An Example Using R

R outputs from last slide

```
## 'data.frame':    390 obs. of  7 variables:
##  $ mpg : num  18 15 18 16 17 15 14 14 14 15 ...
##  $ cyl : int  8 8 8 8 8 8 8 8 8 8 ...
##  $ disp: num  307 350 318 304 302 429 454 440 455 390 ...
##  $ hp  : num  130 165 150 150 140 198 220 215 225 190 ...
##  $ wt  : num  3504 3693 3436 3433 3449 ...
##  $ acc : num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
##  $ year: int  70 70 70 70 70 70 70 70 70 70 ...
```

For now we will focus on using two variables, `wt` and `year`, as predictor variables. That is, we would like to model the fuel efficiency (`mpg`) of a car as a function of its weight (`wt`) and model year (`year`). To do so, we will define the following linear model,

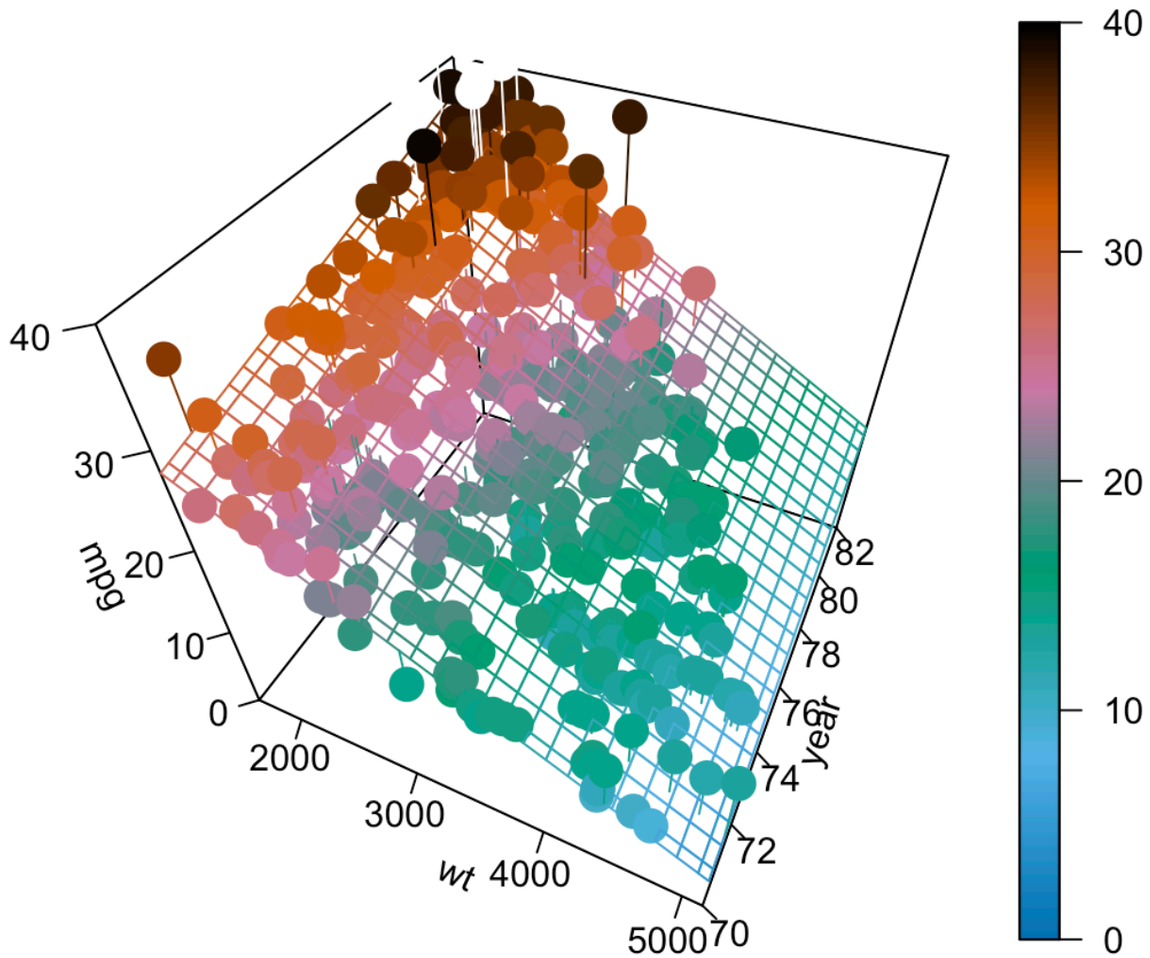
$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \qquad i = 1, 2, \ldots, n$$

where $\epsilon_i \sim N(0, \sigma^2)$. In this notation we will define:

- x_{i1} as the weight (`wt`) of the ith car.
- x_{i2} as the model year (`year`) of the ith car.

3.7 An Example Using R

The picture below will visualize what we would like to accomplish. The data points (x_{i1}, x_{i2}, y_i) now exist in 3-dimensional space, so instead of fitting a line to the data, we will fit a plane. (We'll soon move to higher dimensions, so this will be the last example that is easy to visualize and think about this way.)



3.7 An Example Using R

LSE by 1 R command

```
mpg_model = lm(mpg ~ wt + year, data = autompg)
coef(mpg_model)
```

```
##   (Intercept)            wt          year
## -14.637641945  -0.006634876   0.761401955
```

$$\hat{y} = -14.6376419 + -0.0066349x_1 + 0.761402x_2$$

Here we have once again fit our model using `lm()`, however we have introduced a new syntactical element. The formula `mpg ~ wt + year` now reads: "model the response variable `mpg` as a linear function of `wt` and `year`". That is, it will estimate an intercept, as well as slope coefficients for `wt` and `year`. We then extract these as we have done before using `coef()`.

3.7 An Example Using R

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

```
n = nrow(autompg)
p = length(coef(mpg_model))
X = cbind(rep(1, n), autompg$wt, autompg$year)
y = autompg$mpg

(beta_hat = solve(t(X) %*% X) %*% t(X) %*% y)
```

```
##                [,1]
## [1,] -14.637641945
## [2,]  -0.006634876
## [3,]   0.761401955
```

```
coef(mpg_model)
```

```
##   (Intercept)            wt          year
## -14.637641945  -0.006634876   0.761401955
```

3.7 An Example Using R

MLR output by R

As we can see in the output below, the results of calling `summary()` are similar to SLR, but there are some differences, most obviously a new row for the added predictor variable.

```
summary(mpg_model)
```

```
##
## Call:
## lm(formula = mpg ~ wt + year, data = autompg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.852 -2.292 -0.100  2.039 14.325
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.464e+01  4.023e+00  -3.638 0.000312 ***
## wt          -6.635e-03  2.149e-04 -30.881  < 2e-16 ***
## year         7.614e-01  4.973e-02  15.312  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.431 on 387 degrees of freedom
## Multiple R-squared:  0.8082, Adjusted R-squared:  0.8072
## F-statistic: 815.6 on 2 and 387 DF,  p-value: < 2.2e-16
```

3.7 An Example Using R

Single Parameter Test by R

Then the test

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_1 : \beta_1 \neq 0$$

can be found in the `summary()` output, in particular:

```
summary(mpg_model)$coef
```

```
##                  Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept) -14.637641945 4.0233913563  -3.638135  3.118311e-04
## wt           -0.006634876 0.0002148504 -30.881372 1.850466e-106
## year          0.761401955 0.0497265950  15.311765  1.036597e-41
```

The estimate (`Estimate`), standard error (`Std. Error`), test statistic (`t value`), and p-value (`Pr(>|t|)`) for this test are displayed in the second row, labeled `wt` . Remember that the p-value given here is specifically for a two-sided test, where the hypothesized value is 0.

Also note in this case, by hypothesizing that $\beta_1 = 0$ the null and alternative essentially specify two different models:

- H_0: $Y = \beta_0 + \beta_2 x_2 + \epsilon$
- H_1: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$

This is important. We are not simply testing whether or not there is a relationship between weight and fuel efficiency. We are testing if there is a relationship between weight and fuel efficiency, given that a term for year is in the model. (Note, we dropped some indexing here, for readability.)

3.7 An Example Using R

Confidence Intervals
Test by R

Since $\hat{\beta}_j$ is our estimate for β_j and we have

$$\mathrm{E}[\hat{\beta}_j] = \beta_j$$

Here s_e just means $\hat{\sigma}$

as well as the standard error,

$$\mathrm{SE}[\hat{\beta}_j] = s_e\sqrt{C_{jj}}$$

and the sampling distribution of $\hat{\beta}_j$ is Normal, then we can easily construct confidence intervals for each of the $\hat{\beta}_j$.

$$\hat{\beta}_j \pm t_{\alpha/2,n-p} \cdot s_e\sqrt{C_{jj}}$$

We can find these in `R` using the same method as before. Now there will simply be additional rows for the additional β.

```
confint(mpg_model, level = 0.99)
```

```
##                    0.5 %       99.5 %
## (Intercept) -25.052563681 -4.222720208
## wt           -0.007191036 -0.006078716
## year          0.632680051  0.890123859
```

3.7 An Example Using R

Nested Models Test by R

For example, the `autompg` dataset has a number of additional variables that we have yet to use.

```
names(autompg)
```

```
## [1] "mpg"  "cyl"  "disp" "hp"   "wt"   "acc"  "year"
```

We'll continue to use `mpg` as the response, but now we will consider two different models.

- Full: `mpg ~ wt + year + cyl + disp + hp + acc`
- Null: `mpg ~ wt + year`

Note that these are nested models, as the null model contains a subset of the predictors from the full model, and no additional predictors. Both models have an intercept β_0 as well as a coefficient in front of each of the predictors. We could then write the null hypothesis for comparing these two models as,

$$H_0 : \beta_{\texttt{cyl}} = \beta_{\texttt{disp}} = \beta_{\texttt{hp}} = \beta_{\texttt{acc}} = 0$$

The alternative is simply that at least one of the β_j from the null is not 0.

To perform this test in `R` we first define both models, then give them to the `anova()` commands.

```
null_mpg_model = lm(mpg ~ wt + year, data = autompg)
#full_mpg_model = lm(mpg ~ wt + year + cyl + disp + hp + acc, data = autompg)
full_mpg_model = lm(mpg ~ ., data = autompg)
anova(null_mpg_model, full_mpg_model)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt + year
## Model 2: mpg ~ cyl + disp + hp + wt + acc + year
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    387 4556.6
## 2    383 4530.5  4     26.18 0.5533 0.6967
```

More on next slide

3.7 An Example Using R

Nested Models
Test by R

Here we have used the formula `mpg ~ .` to define to full model. This is the same as the commented out line. Specifically, this is a common shortcut in `R` which reads, “model `mpg` as the response with each of the remaining variables in the data frame as predictors.”

Here we see that the value of the F statistic is 0.553, and the p-value is very large, so we fail to reject the null hypothesis at any reasonable α and say that none of `cyl`, `disp`, `hp`, and `acc` are significant with `wt` and `year` already in the model.