CHAPTER 3

MULTIPLE LINEAR REGRESSION

3.1 INTRODUCTION

In this chapter the general multiple linear regression model is presented. The presentation serves as a review of the standard results on regression analysis. The standard theoretical results are given without mathematical derivations, but illustrated by numerical examples. Readers interested in mathematical derivations are referred to the bibliographic notes at the end of Chapter 2, where a number of books that contain a formal development of multiple linear regression theory is given.

3.2 DESCRIPTION OF THE DATA AND MODEL

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. 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 $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$ are constants referred to as the model *partial* regression coefficients (or simply as the *regression coefficients*) and ε is a random disturbance

Regression Analysis by Example, Fifth Edition. By Samprit Chatterjee and Ali S. Hadi

Table 3.1 Notation for Data Used in Multiple Regression Analysis

Observation	Response	Predictors			
Number	Y	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}

or 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.

Multiple linear regression is an extension (generalization) of simple linear regression. Thus, the results given here are essentially extensions of the results given in Chapter 2. One can similarly think of simple regression as a special case of multiple regression because all simple regression results can be obtained using the multiple regression results when the number of predictor variables $p = 1$. For example, when $p = 1$, (3.1) and (3.2) reduce to (2.9) and (2.10), respectively.

3.3 EXAMPLE: SUPERVISOR PERFORMANCE DATA

Throughout this chapter 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 items 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

As 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. We refer to this data set as the Supervisor Performance data. The data set can also be found at the book's Website.[1]

A linear model of the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_6 X_6 + \varepsilon, \tag{3.3}$$

relating Y and the six explanatory variables, is assumed. Methods for the validation of this and other assumptions are presented in Chapter 4.

3.4 PARAMETER ESTIMATION

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

[1] http://www.aucegypt.edu/faculty/hadi/RABE5

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

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

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

By a direct application of calculus, it can be shown that the least squares estimates $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$, which minimize $S(\beta_0, \beta_1, \cdots, \beta_p)$, are given by the solution of a system of linear equations known as the *normal equations*.[2] The estimate $\hat{\beta}_0$ is usually referred to as the *intercept* or *constant*, and $\hat{\beta}_j$ as the *estimate* of the (partial) regression coefficient of the predictor X_j.

We assume that the system of equations is solvable and has a unique solution. A closed-form formula for the solution is given in the Appendix at the end of this chapter for readers who are familiar with matrix notation. We shall not say anything more about the actual process of solving the normal equations. We assume the availability of computer software that gives a numerically accurate solution.

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

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

For each observation in our data we can compute

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

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

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \cdots, n. \tag{3.8}$$

An unbiased estimate of σ^2 is given by

$$\hat{\sigma}^2 = \frac{\text{SSE}}{n - p - 1}, \tag{3.9}$$

where

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2, \tag{3.10}$$

is the *sum of squared residuals*. The number $n - p - 1$ in the denominator of (3.9) is called the *degrees of freedom* (df). It is equal to the number of observations minus the number of estimated regression coefficients.

[2] For readers who are familiar with matrix notation, the normal equations and the least squares estimates are given in the Appendix to this chapter as (A.2) and (A.3), respectively.

When certain assumptions hold, the least squares estimators have several desirable properties. Chapter 4 is devoted entirely to validation of the assumptions. We should note, however, that we have applied these validation procedures on the Supervisor Performance data that we use as illustrative numerical examples in this chapter and found no evidence for model misspecification. We will, therefore, continue with the presentation of multiple regression analysis in this chapter knowing that the required assumptions are valid for the Supervisor Performance data.

The properties of least squares estimators are presented in Section 3.7. Based on these properties, one can develop proper statistical inference procedures (e.g., confidence interval estimation, tests of hypothesis, and goodness-of-fit tests). These are presented in Sections 3.8–3.11.

3.5 INTERPRETATIONS OF REGRESSION COEFFICIENTS

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 $X_1 = X_2 = \cdots = X_p = 0$, as in simple regression. The regression coefficient $\beta_j, j = 1, 2, \cdots, p$, 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.

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.

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

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 $e_{Y \cdot X_1}$. 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 $e_{X_2 \cdot X_1}$. The fitted regression equation is
$$\hat{X}_2 = 18.9654 + 0.513032 X_1. \tag{3.14}$$
The residuals, $e_{Y \cdot X_1}$ and $e_{X_2 \cdot X_1}$ 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 $e_{Y \cdot X_1}$ and the predictor variable is $e_{X_2 \cdot X_1}$. The fitted regression equation is
$$\hat{e}_{Y \cdot X_1} = 0 - 0.0502 e_{X_2 \cdot X_1}. \tag{3.15}$$

The interesting result here is that the coefficient of $e_{X_2 \cdot X_1}$ 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 (see also Exercise 3.5).

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 for the reader.

From the above discussion we see that the simple and the multiple regression coefficients are not the same unless the predictor variables are uncorrelated. In non-experimental or observational data, the predictor variables are rarely uncorrelated. In an experimental setting, in contrast, the experimental design is often set up to produce uncorrelated explanatory variables because in an experiment the researcher sets the values of the predictor variables. So in samples derived from experiments it may be the case that the explanatory variables are uncorrelated and hence the simple and multiple regression coefficients in that sample would be the same.

3.6 CENTERING AND SCALING

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 Chapters 9 and 10. This section describes the centering and scaling process.

We have been mainly dealing with regression models of the form

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon, \tag{3.16}$$

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

$$Y = \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon \tag{3.17}$$

is necessary (see, e.g., Chapters 3 and 7). 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.

3.6.1 Centering and Scaling in Intercept Models

If we are fitting an intercept model as in (3.16), we need to 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 of the response variable Y and the jth predictor variable X_j is obtained as follows:

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

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. In addition, unit length scaling has the following property:

$$\text{Cor}(X_j, X_k) = \sum_{i=1}^{n} z_{ij} z_{ik}. \tag{3.20}$$

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 Z_j and Z_k.

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

$$\begin{aligned} \tilde{Y} &= \frac{Y - \bar{y}}{s_y}, \\ \tilde{X}_j &= \frac{X_j - \bar{x}_j}{s_j}, \quad j = 1, \cdots, p, \end{aligned} \tag{3.21}$$

where

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

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.

3.6.2 Scaling in No-Intercept Models

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

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

Rearranging terms, we obtain

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

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

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

where

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

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.

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.

The regression coefficients obtained using the standardized version of the variables are often referred to as the *beta coefficients*. They represent marginal effects of the predictor variables in standard deviation units. For example, θ_j measures the change in standardized units of Y corresponding to an increase of one standard deviation unit in X_j.

We shall make extensive use of the centered and/or scaled variables in Chapters 9 and 10.

3.7 PROPERTIES OF THE LEAST SQUARES ESTIMATORS

Under certain standard regression assumptions (to be stated in Chapter 4), the least squares estimators have the properties listed below. A reader familiar with matrix algebra will find concise statements of these properties employing matrix notation in the Appendix at the end of the chapter.

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*).

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. These are discussed in Section 3.9.

3.8 MULTIPLE CORRELATION COEFFICIENT

After fitting the linear model to a given data set, an assessment is made of the adequacy of fit. The discussion given in Section 2.9 applies here. All the material extend naturally to multiple regression and will not be repeated here.

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.

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. As we outline in Section 3.10, a more detailed analysis is needed to ensure that the model adequately describes the data.

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

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

which is obtained from R^2 in (3.29) after dividing SSE and SST by their respective degrees of freedom. From (3.30) and (3.29) it follows 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. (This is described in Chapter 11.) 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.

3.9 INFERENCE FOR INDIVIDUAL REGRESSION COEFFICIENTS

Using the properties of the least squares estimators discussed in Section 3.7, one can make statistical inference regarding the regression coefficients. The statistic for testing $\mathrm{H}_0 : \beta_j = \beta_j^0$ versus $\mathrm{H}_1 : \beta_j \neq \beta_j^0$, where β_j^0 is a constant chosen by the investigator, is

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

which has a Student's t-distribution with $n - p - 1$ degrees of freedom. The test is carried out by comparing the observed value with the appropriate critical value $t_{(n-p-1,\alpha/2)}$, which is obtained from the t-table given in the Appendix to this book (see Table A.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

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

where $|t_j|$ denotes the absolute value of t_j. 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); see Figure 2.6. The p-value is usually computed and supplied as part of the regression output by many statistical packages.

The usual test is for $\mathrm{H}_0 : \beta_j^0 = 0$, in which case the t-Test reduces to

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

which is the ratio of $\hat{\beta}_j$ to its standard error, s.e.$(\hat{\beta}_j)$ given in the Appendix at the end of this chapter, in (A.10). The standard errors of the coefficients are computed by the statistical packages as part of their standard regression output.

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.

As another example of statistical inference, the confidence limits for β_j with confidence coefficient α are given by

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

where $t_{(n-p-1,\alpha)}$ is the $1-\alpha$ percentile point of the t-distribution with $n-p-1$ degrees of freedom. The confidence interval in (3.36) is for the individual coefficient β_j. A joint confidence region of all regression coefficients is given in the Appendix at the end of this chapter in (A.15).

Note that when $p = 1$ (simple regression), the t-Test in (3.35) and the criteria in (3.33) and (3.34) reduce to the t-Test in (2.26) and the criteria in (2.27) and (2.28), respectively, illustrating the fact that simple regression results can be obtained from the multiple regression results by setting $p = 1$.

Many other statistical inference situations arise in practice in connection with multiple regression. These will be considered in the following sections.

Example: Supervisor Performance Data (Cont.)

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

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

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.

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

3.10 TESTS OF HYPOTHESES IN A LINEAR MODEL

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.

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

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

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 of β_j), is not rejected. This procedure is described formally as follows.

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

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

This ratio is referred to as the F-Test. Note that we divide SSE(RM) – SSE(FM) and SSE(FM) in the above 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.

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. The reader interested in the proofs of the statements above is referred to Graybill (1976), Rao (1973), Searle (1971), or Seber and Lee (2003).

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 given in the Appendix to this book (see Tables A.4 and A.5), 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). The p-value is usually computed and supplied as part of the regression output by many statistical packages.

In the rest of this section, we give several special cases of the general F-Test in (3.40) with illustrative numerical examples using the Supervisor Performance data.

3.10.1 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

$$\text{RM:} \quad H_0 : Y = \beta_0 \qquad\qquad\qquad\qquad + \varepsilon, \tag{3.43}$$

$$\text{FM:} \quad H_1 : Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon. \tag{3.44}$$

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 SSE(RM) $= \sum(y_i - \bar{y})^2 =$ 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.

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

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

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	$n-p-1$	$\text{MSE} = \frac{\text{SSE}}{n-p-1}$	

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.

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, $p(F)$.

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

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

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

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 $\text{SSE(FM)} = \text{SSE} = 1149$. Under the null hypothesis, where all the β'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	

Note that this is the same quantity obtained by $\sum(y_i - \bar{y})^2$. The observed F-ratio is 10.5. In our present example the numerical equivalence of (3.46) and (3.47) is easily seen for

$$F = \frac{R_p^2/p}{(1 - R_p^2)/(n - p - 1)} = \frac{0.7326/6}{(1 - 0.7326)/23} = 10.50.$$

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 in Table A.5 to be 3.71. (Note that the value of 3.71 is obtained in this case by interpolation.) Since the observed F-value is larger than this value, the null hypothesis is rejected; not all the β's can be taken as zero. This, of course, comes as no surprise, because of the large values of some of the t-Tests.

If any of the t-Tests for the individual regression coefficients prove significant, the F for testing all the regression coefficients zero will usually be significant. A more puzzling case can, however, arise when none of the t-values for testing the regression coefficients are significant, but the F-Test given in (3.47) is significant. This implies that although none of the variables individually have significant explanatory power, the entire set of variables taken collectively explain a significant part of the variation in the dependent variable. This situation, when it occurs, should be looked at very carefully, for it may indicate a problem with the data analyzed, namely, that some of the explanatory variables may be highly correlated, a situation commonly called *collinearity*. We discuss this problem in Chapters 9 and 10.

3.10.2 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.

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

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.

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.

We conclude this section with a few 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

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

   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.

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. [Check the t- and F-Tables A.2, A.4, and A.5 in the Appendix to this book to see that $F(1, v) = t^2(v)$.]

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 Section 2.6 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	$n-2$	$\text{MSE} = \frac{\text{SSE}}{n-2}$	

3.10.3 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

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

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 $R_1^2 = 0.6685$. 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.[3] 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.

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

$$H_0' : \beta_1 = \beta_3, \beta_2 = \beta_4 = \beta_5 = \beta_6 = 0. \tag{3.61}$$

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,

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

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?)

3.10.4 Estimating and Testing of Regression Parameters Under Constraints

Sometimes in fitting regression equations to a given body of data it is desired to impose some constraints on the values of the parameters. A common constraint is that the regression coefficients sum to a specified value, usually unity. The constraints often arise because of some theoretical or physical relationships that may connect the variables. Although no such relationships are obvious in our present example, we consider $\beta_1 + \beta_3 = 1$ for the purpose of demonstration. Assuming that the model in (3.58) has already been accepted, we may further argue

[3] The s.e.$(\hat{\beta}_i - \hat{\beta}_j) = \sqrt{\text{Var}(\hat{\beta}_i) + \text{Var}(\hat{\beta}_j) - 2\text{Cov}(\hat{\beta}_i, \hat{\beta}_j)}$. These quantities are defined in the Appendix to this chapter.

that if each of X_1 and X_3 is increased by a fixed amount, Y should increase by that same amount. Formally, we are led to the null hypothesis H_0 which states that

$$\mathrm{H}_0 : \beta_1 + \beta_3 = 1 \mid (\beta_2 = \beta_4 = \beta_5 = \beta_6 = 0). \tag{3.62}$$

Since $\beta_1 + \beta_3 = 1$, or equivalently, $\beta_3 = 1 - \beta_1$, then under H_0 the reduced model is

$$\mathrm{H}_0 : Y = \beta_0 + \beta_1 X_1 + (1 - \beta_1) X_3 + \varepsilon.$$

Rearranging terms we obtain

$$\mathrm{H}_0 : Y - X_3 = \beta_0 + \beta_1 (X_1 - X_3) + \varepsilon,$$

which can be written as

$$\mathrm{H}_0 : Y' = \beta_0 + \beta_1 V + \varepsilon,$$

where $Y' = Y - X_3$ and $V = X_1 - X_3$. The least squares estimates of the parameters, β_1 and β_3 under the constraint are obtained by fitting a regression equation with Y' as response variable and V as the predictor variable. The fitted equation is

$$\hat{Y}' = 1.166 + 0.694\ V,$$

from which it follows that the fitted equation for the reduced model is

$$\hat{Y} = 1.166 + 0.694\ X_1 + 0.306\ X_3$$

with $R^2 = 0.6905$.

The test for H_0 is given by

$$F = \frac{(0.7080 - 0.6905)/1}{0.2920/27} = 1.62, \quad \mathrm{df} = 1 \text{ and } 27,$$

which is not significant. The data support the proposition that the sum of the partial regression coefficients of X_1 and X_3 equal unity.

Recall that we have now tested two separate hypotheses about β_1 and β_3, one which states that they are equal and the other that they sum to unity. Since both hypotheses hold, it is implied that both coefficients can be taken to be 0.5. A test of this null hypothesis, $\beta_1 = \beta_3 = 0.5$, may be performed directly by applying the methods we have outlined.

The previous example, in which the equality of β_1 and β_3 was investigated, can be considered as a special case of a constrained problem in which the constraint is $\beta_1 - \beta_3 = 0$. The tests for the full set or subsets of regression coefficients being zero can also be thought of as examples of testing regression coefficients under constraints.

From the above discussion it is clear that several models may describe a given body of data adequately. Where several descriptions of the data are available, it is important that they all be considered. Some descriptions may be more meaningful than others (meaningful being judged in the context of the application and considerations of subject matter), and one of them may be finally adopted. Looking at alternative descriptions of the data provides insight that might be overlooked in focusing on a single description.

The question of which variables to include in a regression equation is very complex and is taken up in detail in Chapter 11. We make two remarks here that will be elaborated on in later chapters.

1. The estimates of regression coefficients that do not significantly differ from zero are most commonly replaced by zero in the equation. The replacement has two advantages: a simpler model and a smaller prediction variance.

2. A variable or a set of variables may sometimes be retained in an equation because of their theoretical importance in a given problem, even though the sample regression coefficients are statistically insignificant. That is, sample coefficients which are not significantly different from zero are not replaced by zero. The variables so retained should give a meaningful process description, and the coefficients help to assess the contributions of the X's to the value of the dependent variable Y.

3.11 PREDICTIONS

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 Appendix to this chapter, in (A.12) for readers who are familiar with matrix notation. The standard error is usually computed by many statistical packages. Confidence limits for $\hat{y}_0$ with confidence coefficient α are

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

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 Appendix to this chapter, in (A.14) for readers who are familiar with matrix notation. Confidence limits for $\hat{\mu}_0$ with confidence coefficient α are

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

3.12 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.

EXERCISES

3.1 Using the Supervisor data, verify that the coefficient of X_1 in the fitted equation $\hat{Y} = 15.3276 + 0.7803X_1 - 0.0502X_2$ in (3.12) can be obtained from a series of simple regression equations, as outlined in Section 3.5 for the coefficient of X_2.

3.2 Construct a small data set consisting of one response and two predictor variables so that the regression coefficient of X_1 in the following two fitted equations are equal: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1$ and $\hat{Y} = \hat{\alpha}_0 + \hat{\alpha}_1 X_1 + \hat{\alpha}_2 X_2$. Hint: The two predictor variables should be uncorrelated.

3.3 Table 3.10 shows the scores in the final examination F and the scores in two preliminary examinations P_1 and P_2 for 22 students in a statistics course. The data can be found at the book's Website.

(a) Fit each of the following models to the data:

$$
\begin{aligned}
\text{Model 1:} \quad & F = \beta_0 + \beta_1 P_1 && + \varepsilon \\
\text{Model 2:} \quad & F = \beta_0 && + \beta_2 P_2 + \varepsilon \\
\text{Model 3:} \quad & F = \beta_0 + \beta_1 P_1 && + \beta_2 P_2 + \varepsilon
\end{aligned}
$$

(b) Test whether $\beta_0 = 0$ in each of the three models.

(c) Which variable individually, P_1 or P_2, is a better predictor of F?

(d) Which of the three models would you use to predict the final examination scores for a student who scored 78 and 85 on the first and second preliminary examinations, respectively? What is your prediction in this case?

3.4 Find or construct a simple or multiple regression data set such that the resulting adjusted R_a^2 a is negative.

Table 3.10 Examination Data: Scores in Final (F), First Preliminary (P_1), and Second Preliminary (P_2) Examinations

Row	F	P_1	P_2	Row	F	P_1	P_2
1	68	78	73	12	75	79	75
2	75	74	76	13	81	89	84
3	85	82	79	14	91	93	97
4	94	90	96	15	80	87	77
5	86	87	90	16	94	91	96
6	90	90	92	17	94	86	94
7	86	83	95	18	97	91	92
8	68	72	69	19	79	81	82
9	55	68	67	20	84	80	83
10	69	69	70	21	65	70	66
11	91	91	89	22	83	79	81

3.5 The relationship between the simple and the multiple regression coefficients can be seen when we compare the following regression equations:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2, \tag{3.64}$$

$$\hat{Y} = \hat{\beta}'_0 + \hat{\beta}'_1 X_1, \tag{3.65}$$

$$\hat{Y} = \hat{\beta}''_0 + \hat{\beta}'_2 X_2, \tag{3.66}$$

$$\hat{X}_1 = \hat{\alpha}_0 + \hat{\alpha}_2 X_2, \tag{3.67}$$

$$\hat{X}_2 = \hat{\alpha}'_0 + \hat{\alpha}_1 X_1. \tag{3.68}$$

Using the Examination Data in Table 3.10 with $Y = F$, $X_1 = P_1$, and $X_2 = P_2$, verify that:

(a) $\hat{\beta}'_1 = \hat{\beta}_1 + \hat{\beta}_2\hat{\alpha}_1$, that is, the simple regression coefficient of Y on X_1 is the multiple regression coefficient of X_1 plus the multiple regression coefficient of X_2 times the coefficient from the regression of X_2 on X_1.

(b) $\hat{\beta}'_2 = \hat{\beta}_2 + \hat{\beta}_1\hat{\alpha}_2$, that is, the simple regression coefficient of Y on X_2 is the multiple regression coefficient of X_2 plus the multiple regression coefficient of X_1 times the coefficient from the regression of X_1 on X_2.

3.6 Table 3.11 shows the regression output, with some numbers erased, when a simple regression model relating a response variable Y to a predictor variable X_1 is fitted based on 20 observations. Complete the 13 missing numbers, then compute Var(Y) and Var(X_1).

3.7 Table 3.12 shows the regression output, with some numbers erased, when a simple regression model relating a response variable Y to a predictor variable X_1 is fitted based on 18 observations. Complete the 13 missing numbers, then compute Var(Y) and Var(X_1).

Table 3.11 Regression Output When Y is Regressed on X_1 for 20 Observations

ANOVA Table				
Source	Sum of Squares	df	Mean Square	F-Test
Regression	1848.76	–	–	–
Residuals	–	–	–	

Coefficients Table				
Variable	Coefficient	s.e.	t-Test	p-value
Constant	−23.4325	12.74	–	0.0824
X_1	–	0.1528	8.32	< 0.0001
n = –	R^2 = –	R_a^2 = –	$\hat{\sigma}$ = –	df = –

Table 3.12 Regression Output When Y is Regressed on X_1 for 18 Observations

ANOVA Table				
Source	Sum of Squares	df	Mean Square	F-Test
Regression	–	–	–	–
Residuals	–	–	–	

Coefficients Table				
Variable	Coefficient	s.e.	t-Test	p-value
Constant	3.43179	–	0.265	0.7941
X_1	–	0.1421	–	< 0.0001
n = –	R^2 = 0.716	R_a^2 = –	$\hat{\sigma}$ = 7.342	df = –

3.8 Construct the 95% confidence intervals for the individual parameters β_1 and β_2 using the regression output in Table 3.5.

3.9 Explain why the test for testing the hypothesis H_0 in (3.57) is more sensitive for detecting departures from equality of the regression coefficients than the test for testing the hypothesis H_0' in (3.61).

3.10 Using the Supervisor Performance data, test the hypothesis $H_0 : \beta_1 = \beta_3 = 0.5$ in each of the following models:

(a) $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon$.

(b) $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + +\beta_3 X_3 + \varepsilon$.

Table 3.13 Regression Outputs for Salary Discriminating Data

Model 1: Dependent variable is Salary				
Variable	Coefficient	s.e.	t-Test	p-value
Constant	20009.5	0.8244	24271	< 0.0001
Qualification	0.935253	0.0500	18.7	< 0.0001
Gender	0.224337	0.4681	0.479	0.6329

Model 2: Dependent variable is Qualification				
Variable	Coefficient	s.e.	t-Test	p-value
Constant	−16744.4	896.4	−18.7	< 0.0001
Gender	0.850979	0.4349	1.96	0.0532
Salary	0.836991	0.0448	18.7	< 0.0001

3.11 Refer to Exercise 2.10 and the data in Table 2.11, which can also be found at the book's Website.

(a) Using your choice of the response variable Exercise 2.10(f), test the null hypothesis that both the intercept and the slope are zero.

(b) Which of the hypotheses and tests in Exercises 2.10(g), 2.10(h), and 3.11(a) would you choose to test whether people of similar heights tend to marry each other? What is your conclusion?

(c) If none of the above tests is appropriate for testing the hypothesis that people of similar heights tend to marry each other, which test would you use? What is your conclusion based on this test?

3.12 To decide whether a company is discriminating against women, the following data were collected from the company's records: Salary is the annual salary in thousands of dollars, Qualification is an index of employee qualification, and Gender (1, if the employee is a man, and 0, if the employee is a woman). Two linear models were fit to the data and the regression outputs are shown in Table 3.13. Suppose that the usual regression assumptions hold.

(a) Are men paid more than equally qualified women?

(b) Are men less qualified than equally paid women?

(c) Do you detect any inconsistency in the above results? Explain.

(d) Which model would you advocate if you were the defense lawyer? Explain.

3.13 Table 3.14 shows the regression output of a multiple regression model relating the beginning salaries in dollars of employees in a given company to the following predictor variables:

Table 3.14 Regression Output When Salary is Related to Four Predictor Variables

ANOVA Table				
Source	Sum of Squares	df	Mean Square	F-Test
Regression	23665352	4	5916338	22.98
Residuals	22657938	88	257477	

Coefficients Table				
Variable	Coefficient	s.e.	t-Test	p-value
Constant	3526.4	327.7	10.76	0.000
Gender	722.5	117.8	6.13	0.000
Education	90.02	24.69	3.65	0.000
Experience	1.2690	0.5877	2.16	0.034
Months	23.406	5.201	4.50	0.000
$n = 93$	$R^2 = 0.515$	$R_a^2 = 0.489$	$\hat{\sigma} = 507.4$	df = 88

Gender	An indicator variable (1 = man and 0 = woman)
Education	Years of schooling at the time of hire
Experience	Number of months of previous work experience
Months	Number of months with the company

In (a)–(b) below, specify the null and alternative hypotheses, the test used, and your conclusion using a 5% level of significance.

(a) Conduct the F-Test for the overall fit of the regression.

(b) Is there a *positive* linear relationship between Salary and Experience, after accounting for the effect of the variables Gender, Education, and Months?

(c) What salary would you forecast for a man with 12 years of education, 10 months of experience, and 15 months with the company?

(d) What salary would you forecast, on average, for men with 12 years of education, 10 months of experience, and 15 months with the company?

(e) What salary would you forecast, on average, for women with 12 years of education, 10 months of experience, and 15 months with the company?

3.14 Consider the regression model that generated the output in Table 3.14 to be a full model. Now consider the reduced model in which Salary is regressed on only Education. The ANOVA table obtained when fitting this model is shown in Table 3.15. Conduct a single test to compare the full and reduced models. What conclusion can be drawn from the result of the test? (Use $\alpha = 0.05$.)

3.15 Cigarette Consumption Data: A national insurance organization wanted to study the consumption pattern of cigarettes in all 50 states and the District of

Table 3.15 ANOVA Table When the Beginning Salary is Regressed on Education

ANOVA Table				
Source	Sum of Squares	df	Mean Square	F-Test
Regression	7862535	1	7862535	18.60
Residuals	38460756	91	422646	

Table 3.16 Variables in the Cigarette Consumption Data in Table 3.17

Variable	Definition
Age	Median age of a person living in a state
HS	Percentage of people over 25 years of age in a state who had completed high school
Income	Per capita personal income for a state (income in dollars)
Black	Percentage of blacks living in a state
Female	Percentage of females living in a state
Price	Weighted average price (in cents) of a pack of cigarettes in a state
Sales	Number of packs of cigarettes sold in a state on a per capita basis

Columbia. The variables chosen for the study are given in Table 3.16. The data from 1970 are given in Table 3.17. The states are given in alphabetical order. The data can be found at the book's Website.

In (a)–(b) below, specify the null and alternative hypotheses, the test used, and your conclusion using a 5% level of significance.

(a) Test the hypothesis that the variable Female is not needed in the regression equation relating Sales to the six predictor variables.

(b) Test the hypothesis that the variables Female and HS are not needed in the above regression equation.

(c) Compute the 95% confidence interval for the true regression coefficient of the variable Income.

(d) What percentage of the variation in Sales can be accounted for when Income is removed from the above regression equation? Explain.

(e) What percentage of the variation in Sales can be accounted for by the three variables: Price, Age, and Income? Explain.

(f) What percentage of the variation in Sales that can be accounted for by the variable Income, when Sales is regressed on only Income? Explain.

3.16 Consider the two models:

$$\text{RM:} \quad \text{H}_0 : Y = \varepsilon,$$
$$\text{FM:} \quad \text{H}_1 : Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon.$$

Table 3.17 Cigarette Consumption Data (1970)

State	Age	HS	Income	Black	Female	Price	Sales
AL	27.0	41.3	2948.0	26.2	51.7	42.7	89.8
AK	22.9	66.7	4644.0	3.0	45.7	41.8	121.3
AZ	26.3	58.1	3665.0	3.0	50.8	38.5	115.2
AR	29.1	39.9	2878.0	18.3	51.5	38.8	100.3
CA	28.1	62.6	4493.0	7.0	50.8	39.7	123.0
CO	26.2	63.9	3855.0	3.0	50.7	31.1	124.8
CT	29.1	56.0	4917.0	6.0	51.5	45.5	120.0
DE	26.8	54.6	4524.0	14.3	51.3	41.3	155.0
DC	28.4	55.2	5079.0	71.1	53.5	32.6	200.4
FL	32.3	52.6	3738.0	15.3	51.8	43.8	123.6
GA	25.9	40.6	3354.0	25.9	51.4	35.8	109.9
HI	25.0	61.9	4623.0	1.0	48.0	36.7	82.1
ID	26.4	59.5	3290.0	0.3	50.1	33.6	102.4
IL	28.6	52.6	4507.0	12.8	51.5	41.4	124.8
IN	27.2	52.9	3772.0	6.9	51.3	32.2	134.6
IA	28.8	59.0	3751.0	1.2	51.4	38.5	108.5
KS	28.7	59.9	3853.0	4.8	51.0	38.9	114.0
KY	27.5	38.5	3112.0	7.2	50.9	30.1	155.8
LA	24.8	42.2	3090.0	29.8	51.4	39.3	115.9
ME	28.0	54.7	3302.0	0.3	51.3	38.8	128.5
MD	27.1	52.3	4309.0	17.8	51.1	34.2	123.5
MA	29.0	58.5	4340.0	3.1	52.2	41.0	124.3
MI	26.3	52.8	4180.0	11.2	51.0	39.2	128.6
MN	26.8	57.6	3859.0	0.9	51.0	40.1	104.3
MS	25.1	41.0	2626.0	36.8	51.6	37.5	93.4
MO	29.4	48.8	3781.0	10.3	51.8	36.8	121.3
MT	27.1	59.2	3500.0	0.3	50.0	34.7	111.2
NB	28.6	59.3	3789.0	2.7	51.2	34.7	108.1
NV	27.8	65.2	4563.0	5.7	49.3	44.0	189.5
NH	28.0	57.6	3737.0	0.3	51.1	34.1	265.7
NJ	30.1	52.5	4701.0	10.8	51.6	41.7	120.7
NM	23.9	55.2	3077.0	1.9	50.7	41.7	90.0
NY	30.3	52.7	4712.0	11.9	52.2	41.7	119.0
NC	26.5	38.5	3252.0	22.2	51.0	29.4	172.4
ND	26.4	50.3	3086.0	0.4	49.5	38.9	93.8
OH	27.7	53.2	4020.0	9.1	51.5	38.1	121.6
OK	29.4	51.6	3387.0	6.7	51.3	39.8	108.4
OR	29.0	60.0	3719.0	1.3	51.0	29.0	157.0
PA	30.7	50.2	3971.0	8.0	52.0	44.7	107.3
RI	29.2	46.4	3959.0	2.7	50.9	40.2	123.9
SC	24.8	37.8	2990.0	30.5	50.9	34.3	103.6
SD	27.4	53.3	3123.0	0.3	50.3	38.5	92.7
TN	28.1	41.8	3119.0	15.8	51.6	41.6	99.8
TX	26.4	47.4	3606.0	12.5	51.0	42.0	106.4
UT	23.1	67.3	3227.0	0.6	50.6	36.6	65.5
VT	26.8	57.1	3468.0	0.2	51.1	39.5	122.6
VA	26.8	47.8	3712.0	18.5	50.6	30.2	124.3
WA	27.5	63.5	4053.0	2.1	50.3	40.3	96.7
WV	30.0	41.6	3061.0	3.9	51.6	41.6	114.5
WI	27.2	54.5	3812.0	2.9	50.9	40.2	106.4
WY	27.2	62.9	3815.0	0.8	50.0	34.4	132.2

(a) Develop an F-Test for testing the above hypotheses.

(b) Let $p = 1$ (simple regression) and construct a data set Y and X_1 such that H_0 is not rejected at the 5% significance level.

(c) What does the null hypothesis indicate in this case?

(d) Compute the appropriate value of R^2 that relates the above two models.

Appendix: Multiple Regression in Matrix Notation

We present the standard results of multiple regression analysis in matrix notation. Let us define the following matrices:

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} x_{10} & x_{11} & \cdots & x_{1p} \\ x_{20} & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n0} & x_{n1} & \cdots & x_{np} \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \ \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

The linear model in (3.1) can be expressed in terms of the above matrices as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{A.1}$$

where $x_{i0} = 1$ for all i. The assumptions made about ε for least squares estimation are

$$E(\boldsymbol{\varepsilon}) = \mathbf{0} \quad \text{and} \quad \mathrm{Var}(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) = \sigma^2\mathbf{I}_n,$$

where $E(\boldsymbol{\varepsilon})$ is the expected value (mean) of $\boldsymbol{\varepsilon}$, $\mathbf{I}_n$ is the identity matrix of order n, and $\boldsymbol{\varepsilon}^T$ is the transpose of $\boldsymbol{\varepsilon}$. Accordingly, ε_i's are independent and have zero mean and constant variance. This implies that

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}.$$

The least squares estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is obtained by minimizing the sum of squared deviations of the observations from their expected values. Hence the least squares estimators are obtained by minimizing $S(\boldsymbol{\beta})$, where

$$S(\boldsymbol{\beta}) = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).$$

Minimization of $S(\boldsymbol{\beta})$ leads to the system of equations

$$(\mathbf{X}^T\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{Y}. \tag{A.2}$$

This is the system of *normal equations* referred to in Section (3.4). Assuming that $(\mathbf{X}^T\mathbf{X})$ has an inverse, the least squares estimates $\hat{\boldsymbol{\beta}}$ can be written explicitly as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}, \tag{A.3}$$

from which it can be seen that $\hat{\boldsymbol{\beta}}$ is a linear function of $\mathbf{Y}$. The vector of fitted values $\hat{\mathbf{Y}}$ corresponding to the observed $\mathbf{Y}$ is

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{P}\mathbf{Y}, \tag{A.4}$$

where

$$\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T, \tag{A.5}$$

is known as the *hat* or *projection* matrix. The vector of residuals is given by

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{PY} = (\mathbf{I}_n - \mathbf{P})\mathbf{Y}. \tag{A.6}$$

The properties of the least squares estimators are

1. $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$ (i.e., $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$) with variance-covariance matrix Var($\hat{\boldsymbol{\beta}}$), which is

$$\text{Var}(\hat{\boldsymbol{\beta}}) = E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} = \sigma^2\mathbf{C},$$

   where

$$\mathbf{C} = (\mathbf{X}^T\mathbf{X})^{-1}. \tag{A.7}$$

   Of all unbiased estimators of $\boldsymbol{\beta}$ that are linear in the observations, the least squares estimator has minimum variance. For this reason, $\hat{\boldsymbol{\beta}}$ is said to be the *best linear unbiased estimator* (BLUE) of $\boldsymbol{\beta}$.

2. The residual sum of squares can be expressed as

$$\mathbf{e}^T\mathbf{e} = \mathbf{Y}^T(\mathbf{I}_n - \mathbf{P})^T(\mathbf{I}_n - \mathbf{P})\mathbf{Y} = \mathbf{Y}^T(\mathbf{I}_n - \mathbf{P})\mathbf{Y}. \tag{A.8}$$

   The last equality follows because $(\mathbf{I}_n - \mathbf{P})$ is a symmetric idempotent matrix.

3. An unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{\mathbf{e}^T\mathbf{e}}{n-p-1} = \frac{\mathbf{Y}^T(\mathbf{I}_n - \mathbf{P})\mathbf{Y}}{n-p-1}\,. \tag{A.9}$$

   With the added assumption that the ε_i's are normally distributed we have the following additional results:

4. The vector $\hat{\boldsymbol{\beta}}$ has a $(p+1)$-dimensional normal distribution with mean vector $\boldsymbol{\beta}$ and variance-covariance matrix $\sigma^2\mathbf{C}$. The marginal distribution of $\hat{\beta}_j$ is normal with mean β_j and variance $\sigma^2 c_{jj}$, where c_{jj} is the jth diagonal element of $\mathbf{C}$ in (A.7). Accordingly, the standard error of $\hat{\beta}_j$ is

$$\text{s.e.}(\hat{\beta}_j) = \hat{\sigma}\sqrt{c_{jj}}, \tag{A.10}$$

   and the covariance of $\hat{\beta}_i$ and $\hat{\beta}_j$ is $\text{Cov}(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 c_{ij}$.

5. The quantity $W = \mathbf{e}^T\mathbf{e}/\sigma^2$ has an χ^2 distribution with $n - p - 1$ degrees of freedom.

6. $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are distributed independently of one another.

7. The vector of fitted values $\hat{\mathbf{Y}}$ has a singular n-dimensional normal distribution with mean $E(\hat{\mathbf{Y}}) = \mathbf{X}\boldsymbol{\beta}$ and variance-covariance matrix $\text{Var}(\hat{\mathbf{Y}}) = \sigma^2\mathbf{P}$.

8. The residual vector $\mathbf{e}$ has a singular n-dimensional normal distribution with mean $E(\mathbf{e}) = \mathbf{0}$ and variance-covariance matrix $\text{Var}(\mathbf{e}) = \sigma^2(\mathbf{I}_n - \mathbf{P})$.

9. The predicted value $\hat{y}_0$ corresponding to an observation vector $\mathbf{x}_0 = (x_{00}, x_{01}, x_{02}, \cdots, x_{0p})^T$, with $x_{00} = 1$ is

$$\hat{y}_0 = \mathbf{x}_0^T \hat{\boldsymbol{\beta}} \tag{A.11}$$

and its standard error is

$$\text{s.e.}(\hat{y}_0) = \hat{\sigma}\sqrt{1 + \mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0}. \tag{A.12}$$

The mean response μ_0^T corresponding to $\mathbf{x}_0^T$ is

$$\hat{\mu}_0 = \mathbf{x}_0^T \hat{\boldsymbol{\beta}} \tag{A.13}$$

with a standard error

$$\text{s.e.}(\hat{\mu}_0) = \hat{\sigma}\sqrt{\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0}. \tag{A.14}$$

10. The $100(1-\alpha)\%$ joint confidence region for the regression parameters $\boldsymbol{\beta}$ is given by

$$\left\{\boldsymbol{\beta} : \frac{(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T(\mathbf{X}^T\mathbf{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})}{\hat{\sigma}^2(p+1)} \leq F_{(p+1, n-p-1, \alpha)}\right\}, \tag{A.15}$$

which is an ellipsoid centered at $\hat{\boldsymbol{\beta}}$.