Probability & Statistics
for Engineers & Scientists

NINTH EDITION

WALPOLE | MYERS | MYERS | YE

Probability & Statistics
for Engineers & Scientists

Probability & Statistics for Engineers & Scientists

NINTH EDITION

Ronald E. Walpole
Roanoke College

Raymond H. Myers
Virginia Tech

Sharon L. Myers
Radford University

Keying Ye
University of Texas at San Antonio

Prentice Hall

Boston Columbus Indianapolis New York San Francisco Upper Saddle River
Amsterdam Cape Town Dubai London Madrid Milan Munich Paris Montréal Toronto
Delhi Mexico City São Paulo Sydney Hong Kong Seoul Singapore Taipei Tokyo

Editor in Chief: *Deirdre Lynch*
Acquisitions Editor: *Christopher Cummings*
Executive Content Editor: *Christine O'Brien*
Associate Editor: *Christina Lepre*
Senior Managing Editor: *Karen Wernholm*
Senior Production Project Manager: *Tracy Patruno*
Design Manager: *Andrea Nix*
Cover Designer: *Heather Scott*
Digital Assets Manager: *Marianne Groth*
Associate Media Producer: *Vicki Dreyfus*
Marketing Manager: *Alex Gay*
Marketing Assistant: *Kathleen DeChavez*
Senior Author Support/Technology Specialist: *Joe Vetere*
Rights and Permissions Advisor: *Michael Joyce*
Senior Manufacturing Buyer: *Carol Melville*
Production Coordination: *Lifland et al. Bookmakers*
Composition: *Keying Ye*
Cover photo: *Marjory Dressler/Dressler Photo-Graphics*

Library of Congress Cataloging-in-Publication Data

Probability & statistics for engineers & scientists/Ronald E. Walpole ... [et al.] — 9th ed.
p. cm.
ISBN 978-0-321-62911-1
1. Engineering—Statistical methods. 2. Probabilities. I. Walpole, Ronald E.

TA340.P738 2011
519.02'462–dc22

2010004857

1 2 3 4 5 6 7 8 9 10—EB—14 13 12 11 10

Prentice Hall
is an imprint of

PEARSON

www.pearsonhighered.com

ISBN 10: 0-321-62911-6
ISBN 13: 978-0-321-62911-1

Chapter 12

Multiple Linear Regression and Certain Nonlinear Regression Models

12.1 Introduction

In most research problems where regression analysis is applied, more than one independent variable is needed in the regression model. The complexity of most scientific mechanisms is such that in order to be able to predict an important response, a **multiple regression model** is needed. When this model is linear in the coefficients, it is called a **multiple linear regression model**. For the case of k independent variables $x_1, x_2, \ldots, x_k$, the mean of $Y|x_1, x_2, \ldots, x_k$ is given by the multiple linear regression model

$$\mu_{Y|x_1,x_2,\ldots,x_k} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k,$$

and the estimated response is obtained from the sample regression equation

$$\hat{y} = b_0 + b_1 x_1 + \cdots + b_k x_k,$$

where each regression coefficient β_i is estimated by b_i from the sample data using the method of least squares. As in the case of a single independent variable, the multiple linear regression model can often be an adequate representation of a more complicated structure within certain ranges of the independent variables.

Similar least squares techniques can also be applied for estimating the coefficients when the linear model involves, say, powers and products of the independent variables. For example, when $k = 1$, the experimenter may believe that the means $\mu_{Y|x}$ do not fall on a straight line but are more appropriately described by the **polynomial regression model**

$$\mu_{Y|x} = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_r x^r,$$

and the estimated response is obtained from the polynomial regression equation

$$\hat{y} = b_0 + b_1 x + b_2 x^2 + \cdots + b_r x^r.$$

Confusion arises occasionally when we speak of a polynomial model as a linear model. However, statisticians normally refer to a linear model as one in which the parameters occur linearly, regardless of how the independent variables enter the model. An example of a nonlinear model is the **exponential relationship**

$$\mu_{Y|x} = \alpha\beta^x,$$

whose response is estimated by the regression equation

$$\hat{y} = ab^x.$$

There are many phenomena in science and engineering that are inherently nonlinear in nature, and when the true structure is known, an attempt should certainly be made to fit the actual model. The literature on estimation by least squares of nonlinear models is voluminous. The nonlinear models discussed in this chapter deal with nonideal conditions in which the analyst is certain that the response and hence the response model error are not normally distributed but, rather, have a binomial or Poisson distribution. These situations do occur extensively in practice.

A student who wants a more general account of nonlinear regression should consult *Classical and Modern Regression with Applications* by Myers (1990; see the Bibliography).

12.2 Estimating the Coefficients

In this section, we obtain the least squares estimators of the parameters $\beta_0, \beta_1, \ldots, \beta_k$ by fitting the multiple linear regression model

$$\mu_{Y|x_1,x_2,\ldots,x_k} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

to the data points

$$\{(x_{1i}, x_{2i}, \ldots, x_{ki}, y_i); \quad i = 1, 2, \ldots, n \text{ and } n > k\},$$

where y_i is the observed response to the values $x_{1i}, x_{2i}, \ldots, x_{ki}$ of the k independent variables $x_1, x_2, \ldots, x_k$. Each observation $(x_{1i}, x_{2i}, \ldots, x_{ki}, y_i)$ is assumed to satisfy the following equation.

Multiple Linear Regression Model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i$$

or

$$y_i = \hat{y}_i + e_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki} + e_i,$$

where ϵ_i and e_i are the random error and residual, respectively, associated with the response y_i and fitted value $\hat{y}_i$.

As in the case of simple linear regression, it is assumed that the ϵ_i are independent and identically distributed with mean 0 and common variance σ^2.

In using the concept of least squares to arrive at estimates $b_0, b_1, \ldots, b_k$, we minimize the expression

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{1i} - b_2 x_{2i} - \cdots - b_k x_{ki})^2.$$

Differentiating SSE in turn with respect to $b_0, b_1, \ldots, b_k$ and equating to zero, we generate the set of $k + 1$ **normal equations for multiple linear regression**.

Normal Estimation Equations for Multiple Linear Regression

$$\begin{aligned}
nb_0 + b_1\sum_{i=1}^n x_{1i} \quad + b_2\sum_{i=1}^n x_{2i} \quad + \cdots + b_k\sum_{i=1}^n x_{ki} &= \sum_{i=1}^n y_i\\
b_0\sum_{i=1}^n x_{1i} + b_1\sum_{i=1}^n x_{1i}^2 \quad + b_2\sum_{i=1}^n x_{1i}x_{2i} + \cdots + b_k\sum_{i=1}^n x_{1i}x_{ki} &= \sum_{i=1}^n x_{1i}y_i\\
\vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad &\vdots\\
b_0\sum_{i=1}^n x_{ki} + b_1\sum_{i=1}^n x_{ki}x_{1i} + b_2\sum_{i=1}^n x_{ki}x_{2i} + \cdots + b_k\sum_{i=1}^n x_{ki}^2 &= \sum_{i=1}^n x_{ki}y_i
\end{aligned}$$

These equations can be solved for $b_0, b_1, b_2, \ldots, b_k$ by any appropriate method for solving systems of linear equations. Most statistical software can be used to obtain numerical solutions of the above equations.

Example 12.1: A study was done on a diesel-powered light-duty pickup truck to see if humidity, air temperature, and barometric pressure influence emission of nitrous oxide (in ppm). Emission measurements were taken at different times, with varying experimental conditions. The data are given in Table 12.2. The model is

$$\mu_{Y|x_1,x_2,x_3} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,$$

or, equivalently,

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i, \quad i = 1, 2, \ldots, 20.$$

Fit this multiple linear regression model to the given data and then estimate the amount of nitrous oxide emitted for the conditions where humidity is 50%, temperature is 76°F, and barometric pressure is 29.30.

Table 12.1: Data for Example 12.1

Nitrous Oxide, y	Humidity, x_1	Temp., x_2	Pressure, x_3	Nitrous Oxide, y	Humidity, x_1	Temp., x_2	Pressure, x_3
0.90	72.4	76.3	29.18	1.07	23.2	76.8	29.38
0.91	41.6	70.3	29.35	0.94	47.4	86.6	29.35
0.96	34.3	77.1	29.24	1.10	31.5	76.9	29.63
0.89	35.1	68.0	29.27	1.10	10.6	86.3	29.56
1.00	10.7	79.0	29.78	1.10	11.2	86.0	29.48
1.10	12.9	67.4	29.39	0.91	73.3	76.3	29.40
1.15	8.3	66.8	29.69	0.87	75.4	77.9	29.28
1.03	20.1	76.9	29.48	0.78	96.6	78.7	29.29
0.77	72.2	77.7	29.09	0.82	107.4	86.8	29.03
1.07	24.0	67.7	29.60	0.95	54.9	70.9	29.37

Source: Charles T. Hare, "Light-Duty Diesel Emission Correction Factors for Ambient Conditions," EPA-600/2-77-116. U.S. Environmental Protection Agency.

***Solution*:** The solution of the set of estimating equations yields the unique estimates

$$b_0 = -3.507778,\ b_1 = -0.002625,\ b_2 = 0.000799,\ b_3 = 0.154155.$$

Therefore, the regression equation is

$$\hat{y} = -3.507778 - 0.002625\,x_1 + 0.000799\,x_2 + 0.154155\,x_3.$$

For 50% humidity, a temperature of 76°F, and a barometric pressure of 29.30, the estimated amount of nitrous oxide emitted is

$$\begin{aligned}\hat{y} &= -3.507778 - 0.002625(50.0) + 0.000799(76.0) + 0.1541553(29.30)\\ &= 0.9384 \text{ ppm}.\end{aligned}$$

Polynomial Regression

Now suppose that we wish to fit the polynomial equation

$$\mu_{Y|x} = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_r x^r$$

to the n pairs of observations $\{(x_i, y_i);\ i = 1, 2, \ldots, n\}$. Each observation, y_i, satisfies the equation

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_r x_i^r + \epsilon_i$$

or

$$y_i = \hat{y}_i + e_i = b_0 + b_1 x_i + b_2 x_i^2 + \cdots + b_r x_i^r + e_i,$$

where r is the degree of the polynomial and ϵ_i and e_i are again the random error and residual associated with the response y_i and fitted value $\hat{y}_i$, respectively. Here, the number of pairs, n, must be at least as large as $r+1$, the number of parameters to be estimated.

Notice that the polynomial model can be considered a special case of the more general multiple linear regression model, where we set $x_1 = x, x_2 = x^2, \ldots, x_r = x^r$. The normal equations assume the same form as those given on page 445. They are then solved for $b_0, b_1, b_2, \ldots, b_r$.

Example 12.2: Given the data

x	0	1	2	3	4	5	6	7	8	9
y	9.1	7.3	3.2	4.6	4.8	2.9	5.7	7.1	8.8	10.2

fit a regression curve of the form $\mu_{Y|x} = \beta_0 + \beta_1 x + \beta_2 x^2$ and then estimate $\mu_{Y|2}$.

Solution: From the data given, we find that

$$\begin{aligned}10b_0 + \quad 45\,b_1 + \quad 285\,b_2 &= 63.7,\\ 45b_0 + \quad 285b_1 + \quad 2025\,b_2 &= 307.3,\\ 285b_0 + 2025\,b_1 + 15{,}333b_2 &= 2153.3.\end{aligned}$$

Solving these normal equations, we obtain

$$b_0 = 8.698, \quad b_1 = -2.341, \quad b_2 = 0.288.$$

Therefore,

$$\hat{y} = 8.698 - 2.341x + 0.288x^2.$$

When $x = 2$, our estimate of $\mu_{Y|2}$ is

$$\hat{y} = 8.698 - (2.341)(2) + (0.288)(2^2) = 5.168.$$

Example 12.3: The data in Table 12.2 represent the percent of impurities that resulted for various temperatures and sterilizing times during a reaction associated with the manufacturing of a certain beverage. Estimate the regression coefficients in the polynomial model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{11} x_{1i}^2 + \beta_{22} x_{2i}^2 + \beta_{12} x_{1i} x_{2i} + \epsilon_i,$$

for $i = 1, 2, \ldots, 18$.

Table 12.2: Data for Example 12.3

Sterilizing	**Temperature, x_1 (°C)**		
Time, x_2 (min)	**75**	**100**	**125**
15	14.05	10.55	7.55
	14.93	9.48	6.59
20	16.56	13.63	9.23
	15.85	11.75	8.78
25	22.41	18.55	15.93
	21.66	17.98	16.44

***Solution*:** Using the normal equations, we obtain

$$b_0 = 56.4411, \quad b_1 = -0.36190, \quad b_2 = -2.75299,$$
$$b_{11} = 0.00081, \quad b_{22} = 0.08173, \quad b_{12} = 0.00314,$$

and our estimated regression equation is

$$\hat{y} = 56.4411 - 0.36190x_1 - 2.75299x_2 + 0.00081x_1^2 + 0.08173x_2^2 + 0.00314x_1x_2.$$

Many of the principles and procedures associated with the estimation of polynomial regression functions fall into the category of **response surface methodology**, a collection of techniques that have been used quite successfully by scientists and engineers in many fields. The x_i^2 are called **pure quadratic terms**, and the $x_i x_j$ $(i \neq j)$ are called **interaction terms**. Such problems as selecting a proper experimental design, particularly in cases where a large number of variables are in the model, and choosing optimum operating conditions for $x_1, x_2, \ldots, x_k$ are often approached through the use of these methods. For an extensive exposure, the reader is referred to *Response Surface Methodology: Process and Product Optimization Using Designed Experiments* by Myers, Montgomery, and Anderson-Cook (2009; see the Bibliography).

12.3 Linear Regression Model Using Matrices

In fitting a multiple linear regression model, particularly when the number of variables exceeds two, a knowledge of matrix theory can facilitate the mathematical manipulations considerably. Suppose that the experimenter has k independent variables $x_1, x_2, \ldots, x_k$ and n observations $y_1, y_2, \ldots, y_n$, each of which can be expressed by the equation

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i.$$

This model essentially represents n equations describing how the response values are generated in the scientific process. Using matrix notation, we can write the following equation:

General Linear Model

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

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

Then the least squares method for estimation of β, illustrated in Section 12.2, involves finding $\mathbf{b}$ for which

$$SSE = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})$$

is minimized. This minimization process involves solving for $\mathbf{b}$ in the equation

$$\frac{\partial}{\partial \mathbf{b}}(SSE) = \mathbf{0}.$$

We will not present the details regarding solution of the equations above. The result reduces to the solution of $\mathbf{b}$ in

$$(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{y}.$$

Notice the nature of the $\mathbf{X}$ matrix. Apart from the initial element, the ith row represents the x-values that give rise to the response y_i. Writing

$$\mathbf{A} = \mathbf{X}'\mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{2i} & \cdots & \sum_{i=1}^{n} x_{ki} \\ \sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{1i}^2 & \sum_{i=1}^{n} x_{1i}x_{2i} & \cdots & \sum_{i=1}^{n} x_{1i}x_{ki} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum_{i=1}^{n} x_{ki} & \sum_{i=1}^{n} x_{ki}x_{1i} & \sum_{i=1}^{n} x_{ki}x_{2i} & \cdots & \sum_{i=1}^{n} x_{ki}^2 \end{bmatrix}$$

and

$$\mathbf{g} = \mathbf{X}'\mathbf{y} = \begin{bmatrix} g_0 = \sum_{i=1}^{n} y_i \\ g_1 = \sum_{i=1}^{n} x_{1i}y_i \\ \vdots \\ g_k = \sum_{i=1}^{n} x_{ki}y_i \end{bmatrix}$$

allows the normal equations to be put in the matrix form

$$\mathbf{A}\mathbf{b} = \mathbf{g}.$$

If the matrix $\mathbf{A}$ is nonsingular, we can write the solution for the regression coefficients as

$$\mathbf{b} = \mathbf{A}^{-1}\mathbf{g} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

Thus, we can obtain the prediction equation or regression equation by solving a set of $k+1$ equations in a like number of unknowns. This involves the inversion of the $k+1$ by $k+1$ matrix $\mathbf{X}'\mathbf{X}$. Techniques for inverting this matrix are explained in most textbooks on elementary determinants and matrices. Of course, there are many high-speed computer packages available for multiple regression problems, packages that not only print out estimates of the regression coefficients but also provide other information relevant to making inferences concerning the regression equation.

Example 12.4: The percent survival rate of sperm in a certain type of animal semen, after storage, was measured at various combinations of concentrations of three materials used to increase chance of survival. The data are given in Table 12.3. Estimate the multiple linear regression model for the given data.

Table 12.3: Data for Example 12.4

y (% survival)	x_1 (weight %)	x_2 (weight %)	x_3 (weight %)
25.5	1.74	5.30	10.80
31.2	6.32	5.42	9.40
25.9	6.22	8.41	7.20
38.4	10.52	4.63	8.50
18.4	1.19	11.60	9.40
26.7	1.22	5.85	9.90
26.4	4.10	6.62	8.00
25.9	6.32	8.72	9.10
32.0	4.08	4.42	8.70
25.2	4.15	7.60	9.20
39.7	10.15	4.83	9.40
35.7	1.72	3.12	7.60
26.5	1.70	5.30	8.20

***Solution*:** The least squares estimating equations, $(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{y}$, are

$$\begin{bmatrix} 13.0 & 59.43 & 81.82 & 115.40 \\ 59.43 & 394.7255 & 360.6621 & 522.0780 \\ 81.82 & 360.6621 & 576.7264 & 728.3100 \\ 115.40 & 522.0780 & 728.3100 & 1035.9600 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 377.5 \\ 1877.567 \\ 2246.661 \\ 3337.780 \end{bmatrix}.$$

From a computer readout we obtain the elements of the inverse matrix

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 8.0648 & -0.0826 & -0.0942 & -0.7905 \\ -0.0826 & 0.0085 & 0.0017 & 0.0037 \\ -0.0942 & 0.0017 & 0.0166 & -0.0021 \\ -0.7905 & 0.0037 & -0.0021 & 0.0886 \end{bmatrix},$$

and then, using the relation $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, the estimated regression coefficients are obtained as

$$b_0 = 39.1574,\ b_1 = 1.0161,\ b_2 = -1.8616,\ b_3 = -0.3433.$$

Hence, our estimated regression equation is

$$\hat{y} = 39.1574 + 1.0161x_1 - 1.8616x_2 - 0.3433x_3.$$

Exercises

12.1 A set of experimental runs was made to determine a way of predicting cooking time y at various values of oven width x_1 and flue temperature x_2. The coded data were recorded as follows:

y	x_1	x_2
6.40	1.32	1.15
15.05	2.69	3.40
18.75	3.56	4.10
30.25	4.41	8.75
44.85	5.35	14.82
48.94	6.20	15.15
51.55	7.12	15.32
61.50	8.87	18.18
100.44	9.80	35.19
111.42	10.65	40.40

Estimate the multiple linear regression equation

$$\mu_{Y|x_1,x_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

12.2 In *Applied Spectroscopy*, the infrared reflectance spectra properties of a viscous liquid used in the electronics industry as a lubricant were studied. The designed experiment consisted of the effect of band frequency x_1 and film thickness x_2 on optical density y using a Perkin-Elmer Model 621 infrared spectrometer. (*Source*: Pacansky, J., England, C. D., and Wattman, R., 1986.)

y	x_1	x_2
0.231	740	1.10
0.107	740	0.62
0.053	740	0.31
0.129	805	1.10
0.069	805	0.62
0.030	805	0.31
1.005	980	1.10
0.559	980	0.62
0.321	980	0.31
2.948	1235	1.10
1.633	1235	0.62
0.934	1235	0.31

Estimate the multiple linear regression equation

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2.$$

12.3 Suppose in Review Exercise 11.53 on page 437 that we were also given the number of class periods missed by the 12 students taking the chemistry course. The complete data are shown.

Student	Chemistry Grade, y	Test Score, x_1	Classes Missed, x_2
1	85	65	1
2	74	50	7
3	76	55	5
4	90	65	2
5	85	55	6
6	87	70	3
7	94	65	2
8	98	70	5
9	81	55	4
10	91	70	3
11	76	50	1
12	74	55	4

(a) Fit a multiple linear regression equation of the form $\hat{y} = b_0 + b_1 x_1 + b_2 x_2$.

(b) Estimate the chemistry grade for a student who has an intelligence test score of 60 and missed 4 classes.

12.4 An experiment was conducted to determine if the weight of an animal can be predicted after a given period of time on the basis of the initial weight of the animal and the amount of feed that was eaten. The following data, measured in kilograms, were recorded:

Final Weight, y	Initial Weight, x_1	Feed Weight, x_2
95	42	272
77	33	226
80	33	259
100	45	292
97	39	311
70	36	183
50	32	173
80	41	236
92	40	230
84	38	235

(a) Fit a multiple regression equation of the form

$$\mu_{Y|x_1,x_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

(b) Predict the final weight of an animal having an initial weight of 35 kilograms that is given 250 kilograms of feed.

12.5 The electric power consumed each month by a chemical plant is thought to be related to the average ambient temperature x_1, the number of days in the month x_2, the average product purity x_3, and the tons of product produced x_4. The past year's historical data are available and are presented in the following table.