迴歸分析 (一)

Kutner's Applied Linear Statistical Models (5/E)

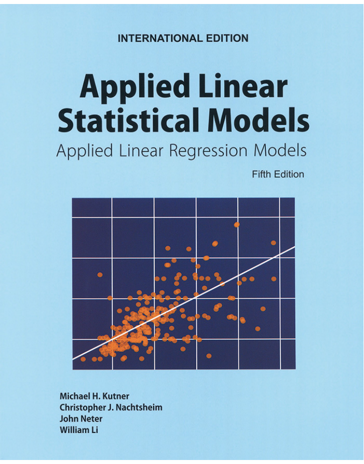
授課教師: 吳漢銘 國立政治大學統計學系

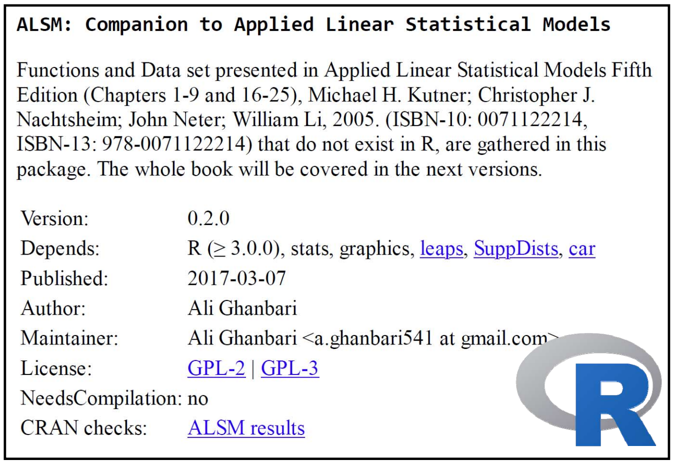
開課單位: 統計二

科目代碼: 304008001

教學網站: http://www.hmwu.idv.tw

系級: ________________ 學號: ________________ 姓名: ______________________


INTERNATIONAL EDITION
Applied Linear Statistical Models
Applied Linear Regression Models
Fifth Edition
Michael H. Kutner
Christopher J. Nachtsheim
John Neter
William Li



ALSM: Companion to Applied Linear Statistical Models

Functions and Data set presented in Applied Linear Statistical Models Fifth Edition (Chapters 1-9 and 16-25), Michael H. Kutner; Christopher J. Nachtsheim; John Neter; William Li, 2005. (ISBN-10: 0071122214, ISBN-13: 978-0071122214) that do not exist in R, are gathered in this package. The whole book will be covered in the next versions.

Version:	0.2.0
Depends:	R (≥ 3.0.0), stats, graphics, leaps, SuppDists, car
Published:	2017-03-07
Author:	Ali Ghanbari
Maintainer:	Ali Ghanbari <a.ghanbari541 at gmail.com>
License:	GPL-2 | GPL-3
NeedsCompilation:	no
CRAN checks:	ALSM results

111 學年度第 2 學期

Contents

叮嚀

A. 平常就要唸書，做習題。

B. 考過的題目，要主動訂正。

C. 上課以「互相尊重」為最高原則並盡到「告知老師」的義務。

D. 上課可小聲討論、上廁所安靜去回、不鼓勵飲食。(請一定要維護教室整潔)

E. 四不一要: 「上課不聊天，睡覺不趴著，手機不要滑，考試不作弊，要認真。」

Regression Analysis (I)

Kutner's Applied Linear Statistical Models (5/E)

Chapter 1: Linear Regression with One Predictor Variable

Thursday 09:10-12:00, 商館 260205

Han-Ming Wu

Department of Statistics, National Chengchi University

http://www.hmwu.idv.tw

Overview

1. Regression analysis (迴歸分析) is a statistical methodology that utilizes the relation between two or more quantitative variables so that a response or outcome variable can be predicted from the other, or others.

2. Examples: general form of a regression model $Y = \hat{f}(X_1, X_2, \cdots, X_p)$:

   (a) Y: the sales of a product, X: the amount of advertising expenditures (支出).

   (b) Y: the performance of an employee on a job, X: a battery of aptitude tests (能力傾向成套測驗, 性向測驗).

   (c) Y: the size of the vocabulary of a child, X_1: age of the child, X_2: amount of education of the parents.

   (d) Y: the length of hospital stay of a surgical patient, X_1: the time in the hospital, X_2: the severity of the operation.

3. In this chapter, we consider the basic ideas of regression analysis and discuss the estimation of the parameters of regression models containing a single predictor variable.

1.1 Relations between Variables

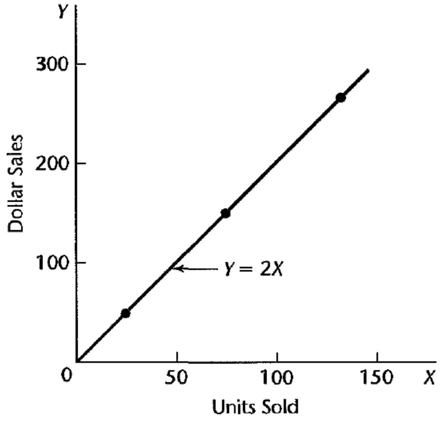
Functional Relation between Two Variables

1. A functional relation between two variables is expressed by a mathematical formula. If X denotes the independent variable and Y the dependent variable, a functional relation is of the form:

$$Y = f(X)$$

2. **Example**: Y: dollar sales of a product sold at a fixed price, X: the number of units sold. If the selling price is $2 per unit, the relation is expressed by the equation: $Y = 2X$.

FIGURE 1.1 **Example of Functional Relation.**



Statistical Relation between Two Variables

1. In general, the observations for a statistical relation do not fall directly on the curve of relationship.

2. **Example 1**: Performance evaluations

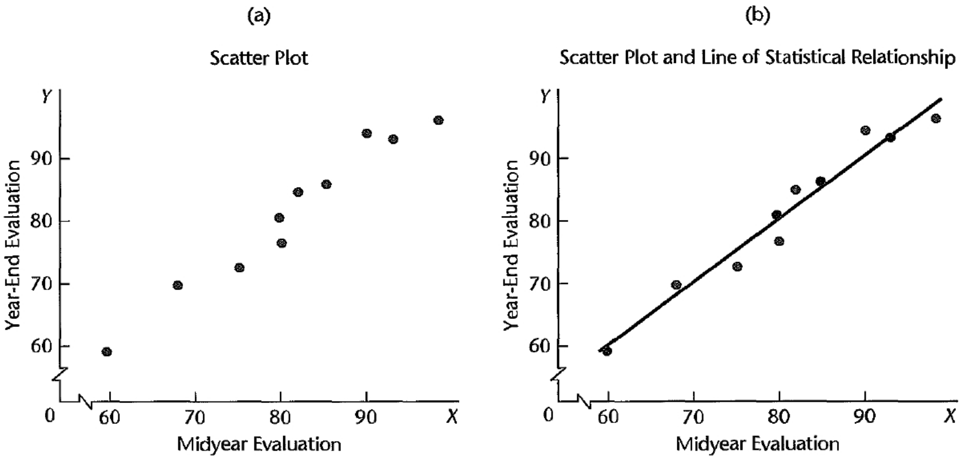
   (a) Performance evaluations for 10 employees were obtained at midyear (X) and at year-end (Y).

   (b) Figure 1.2a: the higher the midyear evaluation, the higher tends to be the year-end evaluation.

   (c) Figure 1.2b: a line of relationship that describes the statistical relation between midyear and year-end evaluations.

(d) Note: that most of the points do not fall directly on the line of statistical relationship. This scattering of points around the line represents variation in year-end evaluations that is not associated with midyear performance evaluation and that is usually considered to be of a random nature.

FIGURE 1.2 Statistical Relation between Midyear Performance Evaluation and Year-End Evaluation.



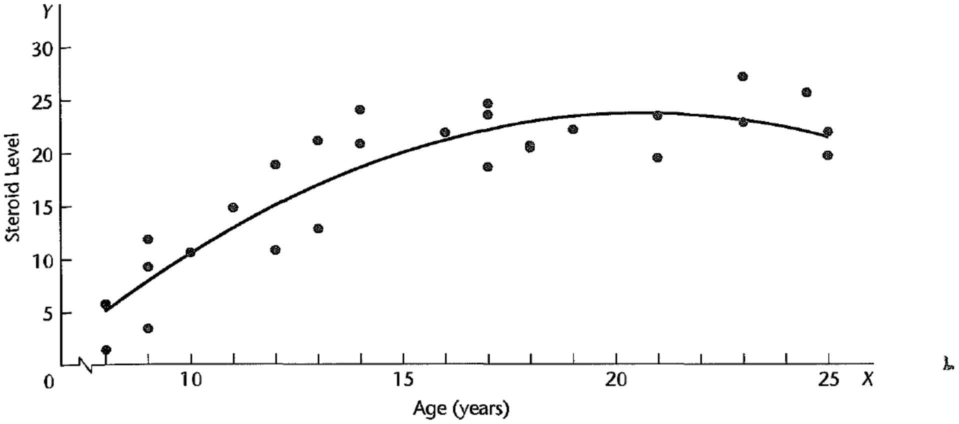
3. **Example 2**:

(a) The data on age and level of a steroid (類固醇) in plasma (血漿) for 27 healthy females between 8 and 25 years old. (Figure 1.3)

(b) The data strongly suggest that the statistical relationship is curvilinear (not linear).

(c) As age increases, steroid level increases up to a point and then begins to level off.

FIGURE 1.3 Curvilinear Statistical Relation between Age and Steroid Level in Healthy Females Aged 8 to 25.



1.2 Regression Models and Their Uses

Historical Origins

1. Regression analysis was first developed by Sir Francis Galton in the latter part of the 19th century.

2. Galton had studied the relation between heights of parents and children and noted that the heights of children of both tall and short parents appeared to ”revert” (回復) or ”regress” (回歸) to the mean of the group.

3. He considered this tendency to be a regression to ”mediocrity.”

4. Galton developed a mathematical description of this regression tendency, the precursor of today's regression models.

5. The term regression persists to this day to describe statistical relations between variables.

☺ 行銷資料科學: 小時了了，大未必佳 迴歸均值的有趣現象: `https://medium.com/marketingdatascience/d5f8e5e73163`.

☺ 均值迴歸 (regression toward the mean) 現象: 當一個特性的極端傾向發生時，會有返回這項特性的平均值 (regression toward mediocrity)。

☺ 例子: 身高較高的父母，其子女的平均身高，要低於他們父母的平均身高，不會長得更高；相對的，身高比較矮的父母，其子女的平均身高，要高於他們父母的平均身高，不會變得更矮。

Basic Concepts

1. A regression model is:
   (a) A tendency of the response variable Y to vary with the predictor variable X in a systematic fashion.
   (b) A scattering of points around the curve of statistical relationship.

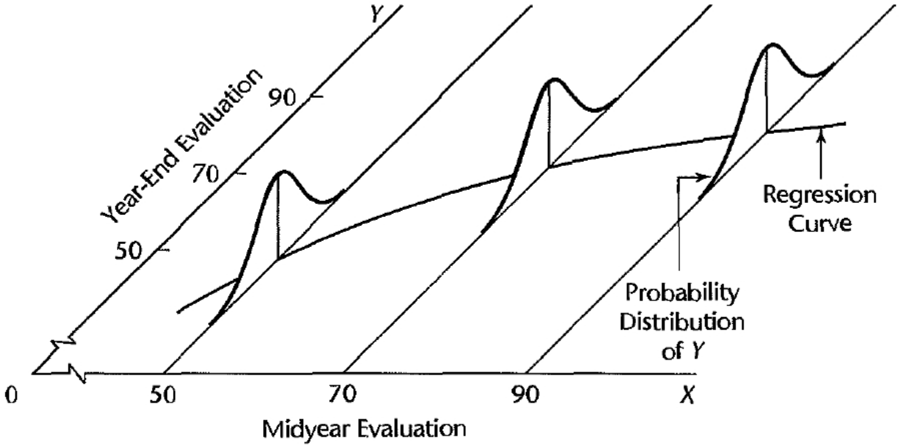
2. Assumptions for a regression model:
   (a) There is a probability distribution (機率分佈) of Y for each level of X.

(b) The means of these probability distributions vary in some systematic fashion with X. ($E(Y|X)$)

3. **Example**: Performance evaluation (Figure 1.2)

(a) The year-end evaluation Y is treated in a regression model as a random variable. For each level of midyear performance evaluation X, there is postulated a probability distribution of Y.

FIGURE 1.4 **Pictorial Representation of Regression Model.**



(b) Figure 1.4: shows probability distributions of Y for midyear evaluation levels at $X = 50$, $X = 70$ and $X = 90$. Note that the means of the probability distributions have a systematic relation to the level of X.

(c) This systematic relationship is called the regression function of Y on X. The graph of the regression function is called the regression curve.

(d) The regression curve, which describes the relation between the means of the probability distributions of Y and the level of X, is the counterpart to the general tendency of Y to vary with X systematically in a statistical relation.

Construction of Regression Models

1. **Selection of Predictor Variables**:

(a) Choosing a limited number of explanatory or predictor variables that is "good" in some sense for the purposes of the analysis.

(b) Other considerations: the importance of the variable; the degree to which observations on the variable can be obtained more accurately, or quickly, or economically than on competing variables; and the degree to which the variable can be controlled.

2. **Functional Form of Regression Relation**:

   (a) The functional form of the regression relation is not known in advance and must be decided upon empirically once the data have been collected.

   (b) The linear or quadratic regression functions are often used as satisfactory first approximations to regression functions of unknown nature.

3. **Scope of Model**:

   (a) In formulating a regression model, we usually need to restrict the coverage of the model to some interval or region of values of the predictor variable(s).

   (b) **Example**: a company studying the effect of price on sales volume investigated six price levels, ranging from \$4.95 to \$6.95. Here, the scope of the model is limited to price levels ranging from near \$5 to near \$7. The shape of the regression function substantially outside this range would be in serious doubt because the investigation provided no evidence as to the nature of the statistical relation below \$4.95 or above \$6.95.

Uses of Regression Analysis

1. Regression analysis serves three major purposes: (1) description, (2) control, and (3) prediction.

2. The several purposes of regression analysis frequently overlap in practice.

Regression and Causality (因果關係)

1. The existence of a statistical relation between the response variable Y and the explanatory or predictor variable X does not imply in any way that Y depends causally on X.

2. No matter how strong is the statistical relation between X and Y, no cause-and-effect pattern is necessarily implied by the regression model.

3. **Example**: data on size of vocabulary (X) and writing speed (Y) for a sample of young children aged 5-10 will show a positive regression relation. This relation does not imply, however, that an increase in vocabulary causes a faster writing speed. Here, other explanatory variables, such as age of the child and amount of education, affect both the vocabulary (X) and the writing speed (Y). Older children have a larger vocabulary and a faster writing speed.

4. Regression analysis by itself provides no information about causal patterns and must be supplemented by additional analyses to obtain insights about causal relations.

Use of Computers

1. Regression analysis often entails lengthy and tedious calculations, computers are usually utilized to perform the necessary calculations.

2. Almost every statistics package for computers contains a regression component: BMDP, MINITAB, SAS, SPSS, SYSTAT, JMP, S-Plus, MATLAB, and R.

1.3 Simple Linear Regression Model with Distribution of Error Terms Unspecified

Formal Statement of Model

1. A simple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \cdots, n \tag{1.1}$$

   where:

   (a) Y_i: the value of the response variable in the ith trial.

   (b) β_0 and β_1: parameters to be estimated.

(c) X_i: the value of the predictor variable in the ith trial

(d) ϵ_i: a random error term with mean $E(\epsilon_i) = 0$ and variance $\sigma^2(\epsilon_i) = \sigma^2$.

(e) ϵ_i and ϵ_j are uncorrelated so that their covariance is zero (i.e., $\sigma(\epsilon_i, \epsilon_j) = 0$ for all $i, j; i \neq j$) $i = 1, \cdots, n$.

2. Regression model (1.1) is said to be

(a) simple: there is only one predictor variable

(b) linear in the parameters: no parameter appears as an exponent or is multiplied or divided by another parameter

(c) linear in the predictor variable: because this variable appears only in the first power.

3. A model that is linear in the parameters and in the predictor variabie is also called first-order model.

Important Features of Model

1. The response Y_i in the ith trial is the sum of two components: (1) the constant term $\beta_0 + \beta_1 X_i$ and (2) the random term ϵ_i. Hence, Y_i is a random variable.

2. Since $E(\epsilon_i) = 0$, it follows that:

$$E(Y_i) = E(\beta_0 + \beta_1 X_i + \epsilon_i) = \beta_0 + \beta_1 X_i + E(\epsilon_i) = \beta_0 + \beta_1 X_i.$$

Thus, the response Y_i, when the level of X in the ith trial is X_i, comes from a probability distribution whose mean is:

$$E(Y_i) = \beta_0 + \beta_1 X_i.$$

The regression function for model (1.1) is:

$$E(Y) = \beta_0 + \beta_1 X$$

since the regression function relates the means of the probability distributions of Y for given X to the level of X.

3. The response Y_i in the ith trial <u>exceeds or falls short</u> of the value of the regression function (<u>$E(Y_i)$</u>) by the error term amount <u>ϵ_i</u>.

4. The error terms ϵ_i are assumed to have constant variance <u>σ^2</u>.It therefore follows that the responses Y_i have the same constant variance:

$$\sigma^2(Y_i) = \sigma^2$$

   Thus, regression model (1.1) assumes that the probability distributions of Y have the same variance <u>σ^2</u>, regardless of the level of the predictor variable X.

5. Since the error terms ϵ_i and ϵ_j are assumed to be uncorrelated, so are the responses <u>Y_i and Y_j</u>.

6. **Summary**: regression model <u>$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$</u> implies that the responses Y_i come from probability distributions whose means are <u>$E(Y_i) = \beta_0 + \beta_1 X_i$</u> and whose variances are <u>σ^2</u>, the same for all levels of X. Further, any two responses Y_i and Y_j are <u>uncorrelated</u>.

7. **Example**: Electrical distributor (Figure 1.6)
   A consultant for an electrical distributor is studying the relationship between the number of bids (<u>X</u>) requested by construction contractors (承包商) for basic lighting equipment during a week and the number of hours (<u>Y</u>) required to prepare the bids.

   (a) Suppose that regression model (1.1) is:

$$Y_i = 9.5 + 2.1X_i + \epsilon_i$$

   (b) The regression function is:

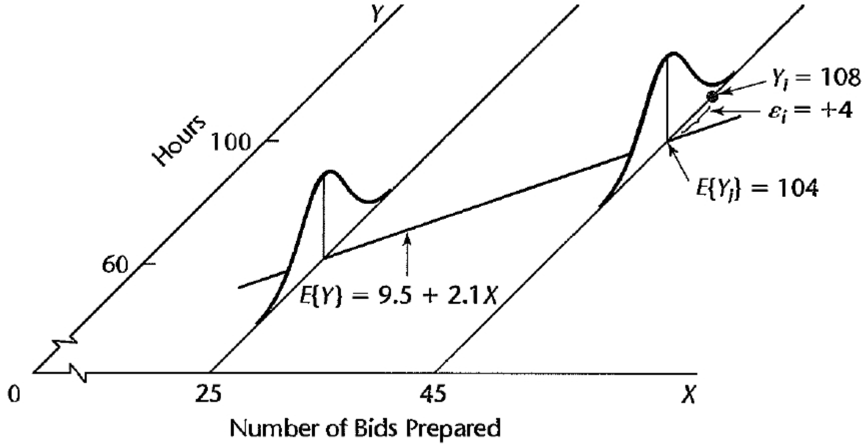
$$\underline{E(Y) = 9.5 + 2.1X.}$$

   (c) Suppose that in the ith week, $X_i = 45$ bids are prepared and the actual number of hours required is $Y_i = 108$. We have

$$E(Y_i) = \underline{9.5 + 2.1(45) = 104} \quad \text{and} \quad \epsilon_i = \underline{Y_i - E(Y_i) = 108 - 104 = 4}$$

   (d) The error term ϵ_i is simply the <u>deviation</u> of Y_i from its mean value $E(Y_i)$.

FIGURE 1.6 Illustration of Simple Linear Regression Model (1.1).



Meaning of Regression Parameters

1. The parameters β_0 and β_1, in regression model (1.1) are called regression coefficients .

   (a) The parameter β_0 is the Y intercept of the regression line. β_1, is the slope of the regression line.

   (b) β_1 indicates the change in the mean of the probability distribution of Y per unit increase in X.

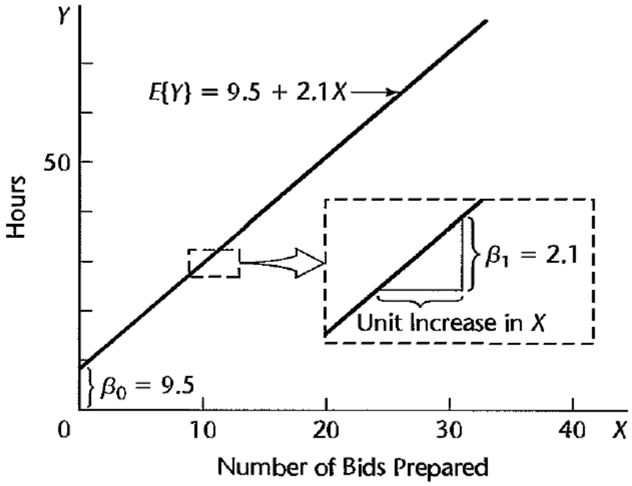
   (c) When the scope of the model includes $X = 0$, β_0 gives the mean of the probability distribution of Y at $X = 0$. When the scope of the model does not cover $X = 0$, β_0 does not have any particular meaning as a separate term in the regression model.

2. **Example**: Electrical distributor (Figure 1.7)

   (a) The regression function: $E(Y) = 9.5+2.1X$. The slope $\beta_1 = 2.1$ indicates that the preparation of one additional bid in a week leads to an increase in the mean of the probability distribution of Y of 2.1 hours.

   (b) The intercept $\beta_0 = 9.5$ indicates the value of the regression function at $X = 0$. Since the linear regression model was formulated to apply to weeks where the number of bids prepared ranges from 20 to 80, $\beta_0 = 9.5$ does not have any intrinsic meaning of its own here.

FIGURE 1.7 Meaning of Parameters of Simple Linear Regression Model (1.1).



Alternative Versions of Regression Model

1. Let X_0 be a constant identically equal to 1. Then, we can write (1.1) as follows:

$$Y_i = \beta_0 X_0 + \beta_1 X_i + \epsilon_i \qquad \text{where} \quad X_0 \equiv 1$$

This version of the model associates an X variable with each regression coefficient.

2. An alternative modification is to use for the predictor variable the deviation $X_i - \bar{X}$ rather than X_i:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1(X_i - \bar{X}) + \beta_1\bar{X} + \epsilon_i \\ &= (\beta_0 + \beta_1\bar{X}) + \beta_1(X_i - \bar{X}) + \epsilon_i \\ &= \beta_0^* + \beta_1(X_i - \bar{X}) + \epsilon_i, \end{aligned}$$

where

$$\beta_0^* = \beta_0 + \beta_1\bar{X}$$

1.4 Data for Regression Analysis*

1.5 Overview of Steps in Regression Analysis*

1.6 Estimation of Regression Function

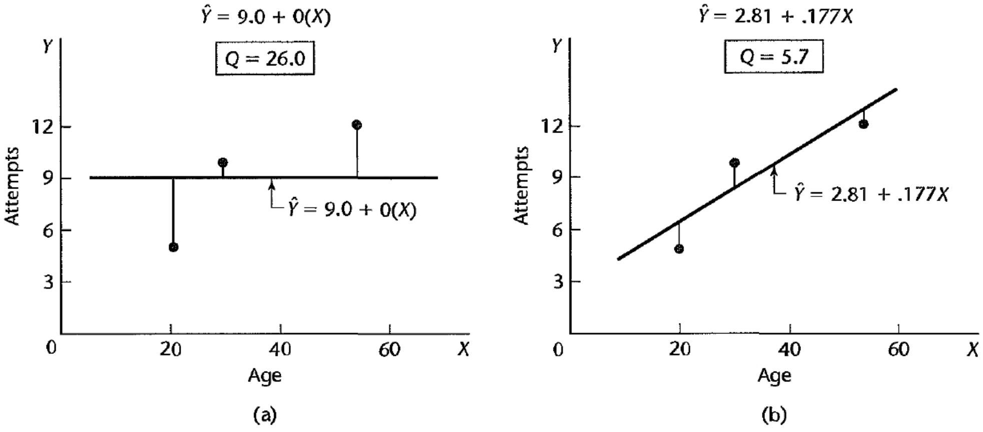
Method of Least Squares

1. For the observations (X_i, Y_i) for each case, the method of least squares considers the sum of the n squared deviation of Y_i from its expected value $E(Y_i)$:

$$Q = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2 \qquad (1.8)$$

2. According to the method of least squares, the estimators of β_0 and β_1 are those values b_0 and b_1 respectively, that minimize the criterion Q for the given sample observations $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$.

FIGURE 1.9 **Illustration of Least Squares Criterion Q for Fit of a Regression Line—Persistence Study Example.**



3. **Example**: (Figure 1.9)

   (a) Figure 1.9a: $Y = 9.0 + 0 \cdot X$. This regression line is not a good fit. The sum of the squared deviations for the three cases is:

$$Q = (5 - 9.0)^2 + (12 - 9.0)^2 + (10 - 9.0)^2 = 26.0$$

(b) Figure 1.9b: $Y = 2.81 + 0.177X$ (the least squares regression line). The criterion Q is much reduced:

$$Q = (5 - 6.35)^2 + (12 - 12.55)^2 + (10 - 8.12)^2 = 5.7$$

Thus, a better fit of the regression line to the data corresponds to a smaller sum Q.

4. **Least Squares Estimators**:

(a) For given sample observations (X_i, Y_i), the quantity Q in (1.8) is a function of β_0 and β_1. The values of β_0 and β_1, that minimize Q can tie derived by differentiating (1.8) with respect to β_0 and β_1:

$$\begin{aligned}\frac{\partial Q}{\partial \beta_0} &= \underline{-2\sum(Y_i - \beta_0 - \beta_1 X_i)}\\ \frac{\partial Q}{\partial \beta_1} &= \underline{-2\sum X_i(Y_i - \beta_0 - \beta_1 X_i)}\end{aligned}$$

(b) Set these partial derivatives equal to zero, using b_0 and b_1 (or $\underline{\hat{\beta}_0 \text{ and } \hat{\beta}_1}$) to denote the particular values of β_0 and β_1, that minimize Q:

$$\begin{aligned}-2\sum(Y_i - b_0 - b_1 X_i) = 0 &\Rightarrow \underline{\sum Y_i - nb_0 - b_1\sum X_i = 0}\\ -2\sum X_i(Y_i - b_0 - b_1 X_i) = 0 &\Rightarrow \underline{\sum X_iY_i - b_0\sum X_i - b_1\sum X_i^2 = 0}.\end{aligned}$$

(c) Normal equations:

$$\begin{aligned}\underline{\sum Y_i = nb_0 + b_1\sum X_i}\\ \underline{\sum X_iY_i = b_0\sum X_i + b_1\sum X_i^2},\end{aligned}$$

b_0 and b_1 are called point estimators of β_0 and β_1, respectively.

NOTE:

(d) The normal equations can be solved simultaneously for b_0 and b_1:

$$b_0 = \frac{1}{n}\left(\sum Y_i - b_1 \sum X_i\right) = \bar{Y} - b_1\bar{X}$$

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

where $\bar{X}$ and $\bar{Y}$ are the means of the X_i and the Y_i observations, respectively.

5. **Properties of Least Squares Estimators**:

(a) **Gauss-Markov theorem**: Under the conditions of regression model (1.1), the least squares estimators b_0 and b_1 in (1.10) are unbiased and have minimum variance among all unbiased linear estimators.

$$E(b_0) = \beta_0 \quad \text{and} \quad E(b_1) = \beta_1,$$

so that neither estimator tends to overestimate or underestimate systematically.

(b) The theorem states that the estimators b_0 and b_l are more precise (i.e., their sampling distributions are less variable) than any other estimators belonging to the class of unbiased estimators that are linear functions of the observations $Y_1, \cdots, Y_n$.

(c) The estimators b_0 and b_1 are such linear functions of the Y_i.

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

This expression is equal to:

$$b_1 = \frac{\sum(X_1 - \bar{X})Y_i}{\sum(X_i - \bar{X})^2} = \sum k_i Y_i$$

where:

$$k_i = \frac{X_i - \bar{X}}{\sum(X_i - \bar{X})^2}$$

Since the k_i are known constants (because the X_i are known constants), b_1 is a linear combination of the Y_i and hence is a linear estimator.

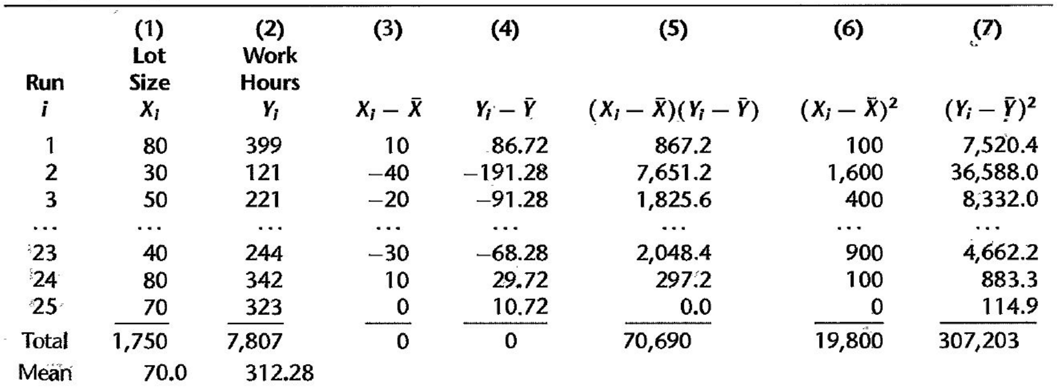
(d) In the same fashion, it can be shown that b_0 is a linear estimator.

6. **Example**: The Toluca Company Manufactures Refrigeration Equipment

   In the past, one of the replacement parts has been produced periodically in lots of varying sizes. When a cost improvement program was undertaken, company officials wished to determine the optimum lot size (X_i) for producing this part. The production of this part involves setting up the production process and machining and assembly operations. One key input for the model to ascertain the optimum lot size was the relationship between lot size and labor hours required to produce the lot. To determine this relationship, data on lot size and work hours (Y_i) for 25 recent production runs were utilized. The production conditions were stable during the six-month period in which the 25 runs were made and were expected to continue to be the same during the next three years, the planning period for which the cost improvement program was being conducted.

   (a) (Table 1.1) All lot sizes are multiples of 10, a result of company policy to facilitate the administration of the parts production.

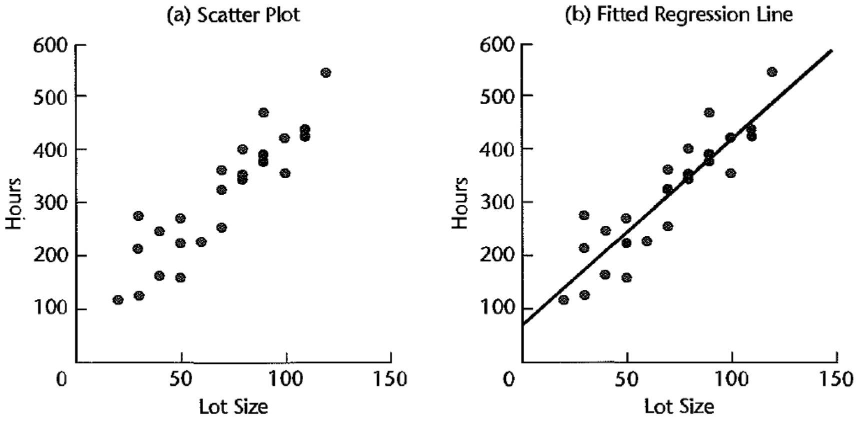
TABLE 1.1 Data on Lot Size and Work Hours and Needed Calculations for Least Squares Estimates—Toluca Company Example.



Run i	(1) Lot Size X_i	(2) Work Hours Y_i	(3) $X_i - \bar{X}$	(4) $Y_i - \bar{Y}$	(5) $(X_i - \bar{X})(Y_i - \bar{Y})$	(6) $(X_i - \bar{X})^2$	(7) $(Y_i - \bar{Y})^2$
1	80	399	10	86.72	867.2	100	7,520.4
2	30	121	−40	−191.28	7,651.2	1,600	36,588.0
3	50	221	−20	−91.28	1,825.6	400	8,332.0
···	···	···	···	···	···	···	···
23	40	244	−30	−68.28	2,048.4	900	4,662.2
24	80	342	10	29.72	297.2	100	883.3
25	70	323	0	10.72	0.0	0	114.9
Total	1,750	7,807	0	0	70,690	19,800	307,203
Mean	70.0	312.28					

   (b) (Figure 1.10a) shows a SYSTAT scatter plot of the data. The scatter plot indicates that the relationship between lot size and work hours is reasonably linear. We also see that no observations on work hours are unusually small or large, with reference to the relationship between lot size and work hours.

FIGURE 1.10 SYSTAT Scatter Plot and Fitted Regression Line—Toluca Company Example.

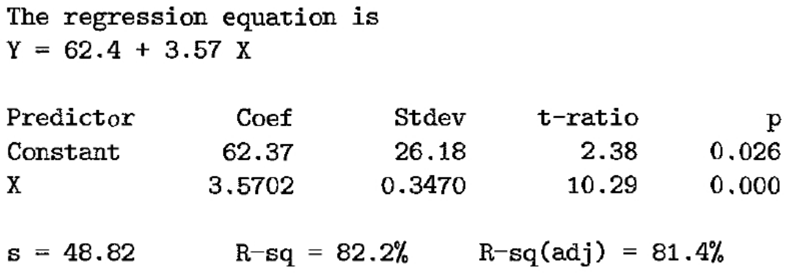


(c) Calculate the least squares estimates:

$$
\begin{aligned}
b_1 &= \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2} = \frac{70690}{19800} = 3.5702 \\
b_0 &= \bar{Y} - b_1\bar{X} = 312.28 - 3.5702(70.0) = 62.37
\end{aligned}
$$

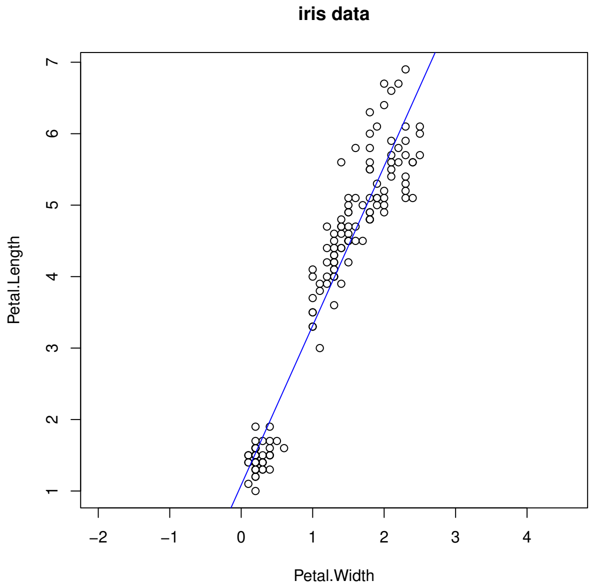
(d) We estimate that the mean number of work hours increases by 3.57 hours for each additional unit produced in the lot. This estimate applies to the range of lot sizes (from about 20 to about 120) in the data from which the estimates were derived.

FIGURE 1.11 Portion of MINITAB Regression Output—Toluca Company Example.



```
The regression equation is
Y = 62.4 + 3.57 X

Predictor        Coef       Stdev     t-ratio        p
Constant        62.37       26.18        2.38    0.026
X              3.5702      0.3470       10.29    0.000

s = 48.82        R-sq = 82.2%     R-sq(adj) = 81.4%
```

☺ *R code example*:


iris data
Petal.Length
1
2
3
4
5
6
7
−2
−1
0
1
2
3
4
Petal.Width

```
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
> str(iris)
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
> attach(iris)
> plot(Petal.Width, Petal.Length, main = "iris data", asp = 1)
> iris.lm <- lm(Petal.Length ~ Petal.Width)
> summary(iris.lm)

Call:
lm(formula = Petal.Length ~ Petal.Width)

Residuals:
     Min       1Q   Median       3Q      Max
-1.33542 -0.30347 -0.02955  0.25776  1.39453

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.08356    0.07297   14.85   <2e-16 ***
Petal.Width  2.22994    0.05140   43.39   <2e-16 ***
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ‘ ’  1

Residual standard error: 0.4782 on 148 degrees of freedom
Multiple R-squared:  0.9271,    Adjusted R-squared:  0.9266
F-statistic:  1882 on 1 and 148 DF,  p-value: < 2.2e-16

> abline(iris.lm, col = "blue")
```

Point Estimation of Mean Response

1. **Estimated Regression Function**

   (a) Given sample estimators b_0 and b_1 of the parameters in the regression function:

   $$E(Y) = \beta_0 + \beta_1 X$$

   we estimate the regression function as follows:

   $$\hat{Y} = b_0 + b_1 X$$

   where $\hat{Y}$ (read Y hat) is the value of the estimated regression function at the level X of the predictor variable.

   (b) We call a value of the response variable a response and $E(Y)$ the mean response.

   (c) The mean response stands for the mean of the probability distribution of Y corresponding to the level X of the predictor variable.

   (d) $\hat{Y}$ then is a point estimator of the mean response when the level of the predictor variable is X.

   (e) An extension of the Gauss-Markov theorem: $\hat{Y}$ is an unbiased estimator of $E(Y)$, with minimum variance in the class of unbiased linear estimators.

   (f) For the cases in the study, we will call $\hat{Y}_i$:

   $$\hat{Y}_i = b_0 + b_1 X_i, \quad i = 1, \ldots, n$$

   the fitted value for the ith case. Thus, the fitted value $\hat{Y}_i$ is to be viewed in distinction to the observed value Y_i.

2. **Example**: The Toluca Company Example

   (a) (Figure 1.10b) The estimated regression function:

   $$\hat{Y} = 62.37 + 3.5702X$$

   It appears to be a good description of the statistical relationship between lot size and work hours.

(b) Suppose that we estimate the mean number of work hours (mean response) required when the lot size is $X = 65$ units:

$$\hat{Y} = \underline{\quad 62.37 + 3.5702(65) = 294.4 \quad} \text{ hours}$$

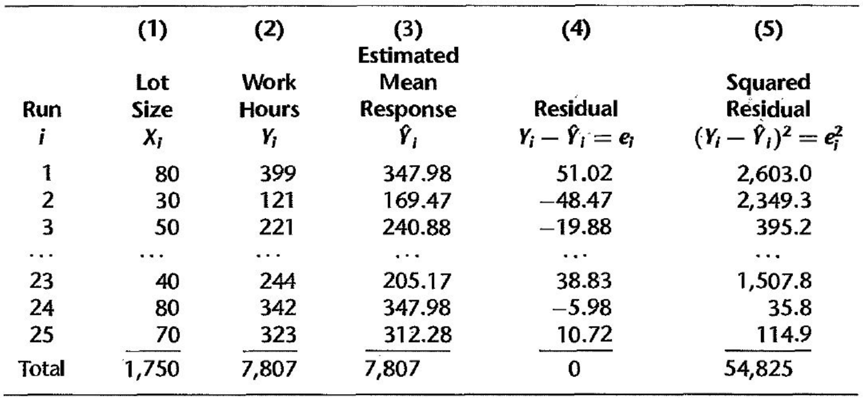
(c) Interpretation: if many lots of 65 units are produced under the conditions of the 25 runs on which the estimated regression function is based, the mean labor time for these lots is about 294 hours.

(d) $\boxed{\text{NOTE}}$ Of course, the labor time for anyone lot of size 65 is likely to fall above or below the mean response because of inherent variability in the production system, as represented by the error term in the model.

(e) (Table 1.2) The fitted value for the first case $X_1 = 80$ is:

$$\hat{Y}_1 = \underline{\quad 62.37 + 3.5702(80) = 347.98 \quad} \text{ hours}$$

TABLE 1.2 Fitted Values, Residuals, and Squared Residuals—Toluca Company Example.



	(1)	(2)	(3)	(4)	(5)
			Estimated		
	Lot	Work	Mean		Squared
Run	Size	Hours	Response	Residual	Residual
i	X_i	Y_i	$\hat{Y}_i$	$Y_i - \hat{Y}_i = e_i$	$(Y_i - \hat{Y}_i)^2 = e_i^2$
1	80	399	347.98	51.02	2,603.0
2	30	121	169.47	−48.47	2,349.3
3	50	221	240.88	−19.88	395.2
···	···	···	···	···	···
23	40	244	205.17	38.83	1,507.8
24	80	342	347.98	−5.98	35.8
25	70	323	312.28	10.72	114.9
Total	1,750	7,807	7,807	0	54,825

☺ *R code example*:

```
> predict(iris.lm, list(Petal.Width = c(0.2, 0.4)))
       1        2
1.529546 1.975534
> data.frame(iris.lm$fitted.values, iris.lm$residuals)
    iris.lm.fitted.values iris.lm.residuals
1                1.529546      -0.129546132
2                1.529546      -0.129546132
3                1.529546      -0.229546132
...
8                1.529546      -0.029546132
9                1.529546      -0.129546132
10               1.306552       0.193447918
...
```

3. **Alternative Model**

   (a) When the alternative regression model (1.6) is to be utilized:

   $$\underline{Y_i = \beta_0^* + \beta_1(X_i - \bar{X}) + \epsilon_i},$$

   the least squares estimator b_1 of β_1 remains the same as before.

   (b) The least squares estimator of $\beta_0^* = \beta_0 + \beta_1\bar{X}$ becomes

   $$b_0^* = \underline{b_0 + b_1\bar{X} = (\bar{Y} - b_1\bar{X}) + b_1\bar{X} = \bar{Y}}$$

   Hence, the estimated regression function for alternative model (1.6) is:

   $$\underline{\hat{Y} = \bar{Y} + b_1(X - \bar{X})}$$

4. In the Toluca Company example, $\bar{Y} = 312.28$ and $\bar{X} = 70.0$. Hence, the estimated regression function in alternative form is:

   $$\hat{Y} = 312.28 + 3.5702(X - 70.0)$$

   For the first lot in our example, $X_1 = 80$; hence, we estimate the mean response to be:

   $$\hat{Y}_1 = 312.28 + 3.5702(80 - 70.0) = 347.98$$

   which, of course, is identical to our earlier result.

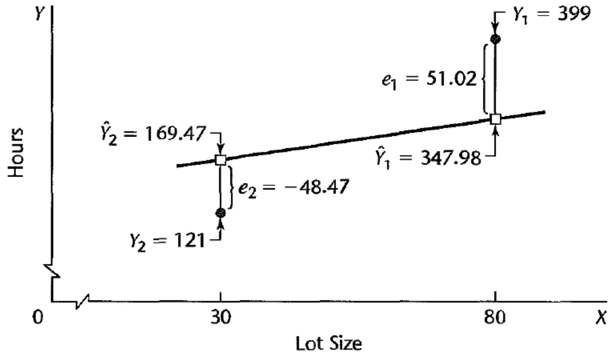


FIGURE 1.12 Illustration of Residuals—Toluca Company Example (not drawn to scale).

Residuals (殘差)

1. The ith residual is the difference between the observed value Y_i and the corresponding fitted value $\hat{Y}_i$. This residual is denoted by e_i:

$$e_i = Y_i - \hat{Y}_i$$

2. For regression model (1.1), the residual e_i becomes:

$$e_i = Y_i - (b_0 + b_1 X_i) = Y_i - b_0 - b_1 X_i$$

3. (Figure 1.12) The magnitude of a residual is represented by the vertical deviation of the Y_i observation from the corresponding point on the estimated regression function (i.e., from the corresponding fitted value $\hat{Y}_i$).

NOTE We need to distinguish between the model error term value $\epsilon_i = Y_i - E(Y_i)$ and the residual $e_i = Y_i - \hat{Y}_i$. The former involves the vertical deviation of Y_i from the unknown true regression line and hence is unknown. On the other hand, the residual is the vertical deviation of Y_i from the fitted value $\hat{Y}_i$ on the estimated regression line, and it is known.

4. Residuals are highly useful for studying whether a given regression model is appropriate for the data at hand.

Properties of Fitted Regression Line

1. The sum of the residuals is zero:

$$\sum_{i=1}^{n} e_i = 0$$

$$\sum e_i = \sum (Y_i - b_0 - b_1 X_i) = \sum Y_i - nb_0 - b_1 \sum X_i$$

$\boxed{\text{NOTE}}$ Rounding errors may, of course, be present in any particular case, resulting in a sum of the residuals that does not equal zero exactly.

2. The sum of the squared residuals, $\underline{\sum e_i^2}$, is a minimum. This was the requirement to be satisfied in deriving the least squares estimators of the regression parameters.

3. The sum of the observed values Y_i equals the sum of the fitted values $\hat{Y}_i$:

$$\underline{\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i}$$

NOTE:

4. The sum of the weighted residuals is zero when the residual in the ith trial is weighted by the level of the predictor variable in the ith trial:

$$\underline{\sum_{i=1}^{n} X_i e_i = 0}$$

NOTE:

5. The sum of the weighted residuals is zero when the residual in the ith trial is weighted by the fitted value of the response variable for the ith trial:

$$\underline{\sum_{i=1}^{n} \hat{Y}_i e_i = 0}$$

NOTE:

6. The regression line always goes through the point $(\bar{X},\bar{Y})$.

$$\hat{Y} = \underline{\bar{Y} + b_1(X - \bar{X}) = \bar{Y} + b_1(\bar{X} - \bar{X})} = \bar{Y}$$

NOTE:

1.7 Estimation of Error Terms Variance σ^2

Point Estimator of σ^2

1. The variance σ^2 of the error terms ϵ_i in regression model (1.1) needs to be estimated to obtain an indication of the variability of the probability distributions of Y. A variety of inferences (推論) concerning the regression function and the prediction of Y require an estimate of σ^2.

2. **Single Population**: The estimator of the variance σ^2 is the sample variance s^2:

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}$$

which is an unbiased estimator of the variance σ^2 of an infinite population. The sample variance is often called a mean square, because a sum of squares has been divided by the appropriate number of degrees of freedom.

3. **Regression Model**

   (a) We need to calculate a sum of squared deviations, but must recognize that the Y_i now come from different probability distributions with different means that depend upon the level X_i. The deviations are the residuals:

$$Y_i - \hat{Y}_i = e_i$$

   and the appropriate sum of squares, denoted by SSE, is:

$$\text{SSE} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

   where SSE stands for error sum of squares or residual sum of squares.

(b) The sum of squares SSE has $n-2$ degrees of freedom associated with it. Two degrees of freedom are lost because both β_0 and β_1 had to be estimated in obtaining the estimated means $\hat{Y}_i$. Hence, the appropriate mean square, denoted by MSE or s^2, is:

$$s^2 = \text{MSE} = \frac{SSE}{n-2} = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum e_i^2}{n-2}$$

where MSE stands for error mean square or residual mean square.

(c) It can be shown that MSE is an unbiased estimator of σ^2 for regression model (1.1): $E(\text{MSE}) = \sigma^2$.

4. **Example**: The Toluca Company Example

(a) (Table 1.2) we obtain: $SSE = 54,825$ and

$$s^2 = \text{MSE} = \frac{54,825}{23} = 2,384$$

A point estimate of σ, the standard deviation of the probability distribution of Y for any X, is $s = \sqrt{2,384} = 48.8$ hours.

(b) Consider again the case where the lot size is $X = 65$ units. We found earlier that the mean of the probability distribution of Y for this lot size is estimated to be 294.4 hours. Now, we have the additional information that the standard deviation of this distribution is estimated to be 48.8 hours.

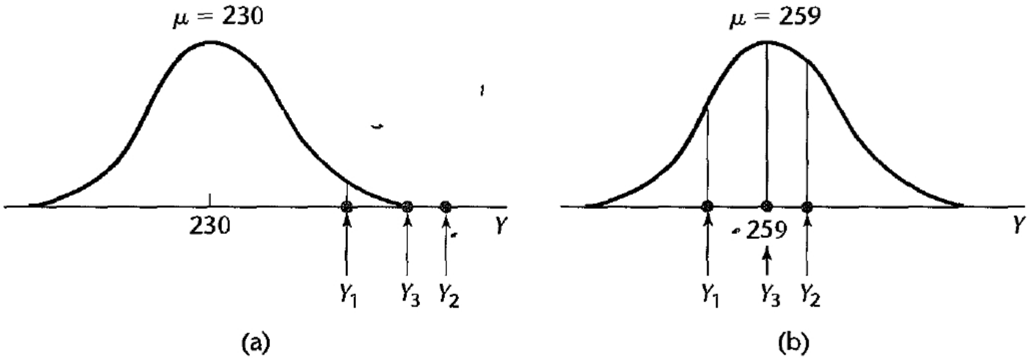


FIGURE 1.13 **Densities for Sample Observations for Two Possible Values of μ: $Y_1 = 250$, $Y_2 = 265$, $Y_3 = 259$.**

1.8 Normal Error Regression Model

Model

1. To set up interval estimates and make tests, however, we need to make an assumption about the form of the distribution of the error terms ϵ_i: they are normally distributed.

2. The normal error regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \cdots, n, \quad (1.24)$$

   (a) Y_i: the observed response in the ith trial.

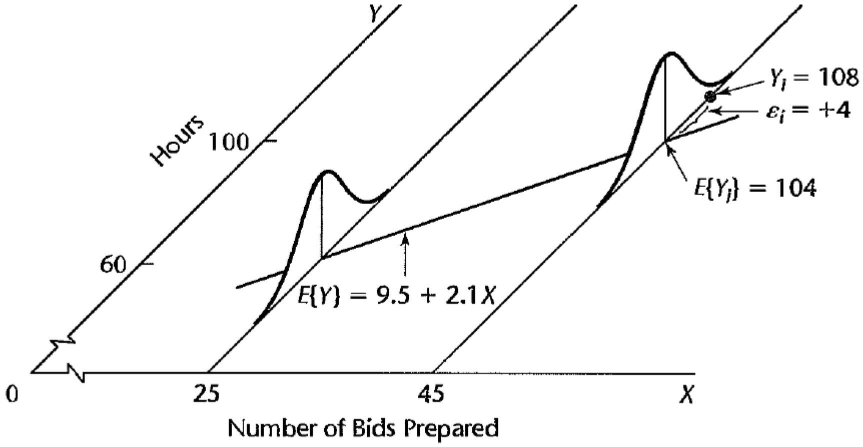
   (b) X_i: known constant, the level of the predictor variable in the ith trial.

   (c) β_0 and β_1: parameters to be estimated.

   (d) ϵ_i: independent normally distributed, with mean 0 and variance σ^2 ($N(0, \sigma^2)$).

3. (Figure 1.6) Regression model (1.24) implies that the Y_i are independent normal random variables, with mean $E(Y_i) = \beta_0 + \beta_1 X_i$ and variance σ^2.

FIGURE 1.6 **Illustration of Simple Linear Regression Model (1.1).**



4. The normality assumption for the error terms is justifiable in many situations because

   (a) the error terms frequently represent the effects of factors omitted from the model that affect the response to some extent and that vary at random without reference to the variable X.

(b) the estimation and testing procedures are based on the t distribution and are usually only sensitive to large departures from normality. Thus, unless the departures from normality are serious, particularly with respect to skewness, the actual confidence coefficients and risks of errors will be close to the levels for exact normality.

Estimation of Parameters by Method of Maximum likelihood

1. **Single Population***

2. **Regression Model**

(a) For the normal error regression model (1.24), each Y_i observation is normally distributed with mean $\beta_0 + \beta_1 X_i$ and standard deviation σ.

(b) The density of an observation Y_i for the normal error regression model (1.24) is:

$$f_i = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma}\right)^2\right]$$

(c) The likelihood function (可能性函數) for n observations $Y_1, Y_2, \cdots, Y_n$ is the product of the individual densities. Since the variance σ^2 of the error terms is usually unknown, the likelihood function is a function of three parameters, β_0, β_1 and σ^2.

$$\begin{aligned} L(\beta_0, \beta_1, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 X_i)^2\right] \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2\right] \end{aligned}$$

(d) The values of β_0, β_1, and σ^2 that maximize this likelihood function are the maximum likelihood estimators (MLE) (最大概估計量) and are denoted by $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\sigma}^2$, respectively.

(e) We find the values of β_0, β_1 and σ^2 that maximize the logarithm of likelihood function $\log L$:

$$\log L = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma^2 - \frac{1}{2\sigma^2}\sum (Y_i - \beta_0 - \beta_1 X_i)^2 .$$

(f) Partial differentiation of the logarithm of the likelihood function:

$$\frac{\partial(\log L)}{\partial\beta_0} = \frac{1}{\sigma^2}\sum(Y_i - \beta_0 - \beta_1 X_i) = 0$$

$$\frac{\partial(\log L)}{\partial\beta_1} = \frac{1}{\sigma^2}\sum X_i(Y_i - \beta_0 - \beta_1 X_i) = 0$$

$$\frac{\partial(\log L)}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum(Y_i - \beta_0 - \beta_1 X_i)^2 = 0$$

(g) Set these partial derivatives equal to zero, replacing, β_0 , β_1 and σ^2 by the estimators $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\sigma}^2$:

$$\begin{aligned}
\sum(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) &= 0 \\
\sum X_i(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) &= 0 \\
\frac{\sum(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}{n} &= \hat{\sigma}^2
\end{aligned}$$

(h) The MLE of β_0 and β_1 are the same estimators as those provided by the method of least squares :

$$\begin{aligned}
\hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1\bar{X} = b_0 \\
\hat{\beta}_1 &= \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = b_1 \\
\hat{\sigma}^2 &= \frac{\sum(Y_i - \hat{Y}_i)^2}{n}
\end{aligned}$$

(i) The maximum likelihood estimator $\hat{\sigma}^2$ is biased, and ordinarily the unbiased estimator $\text{MSE} = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-2}$ is used.

NOTE The unbiased estimator MSE or s^2 differs but slightly from the maximum likelihood estimator $\hat{\sigma}^2$, especially if n is not small:

$$s^2 = MSE = \frac{n}{n-2}\hat{\sigma}^2 .$$

3. **Properties**:

   Since the maximum likelihood estimators $\hat{\beta}_0$ and $\hat{\beta}_1$, are the same as the least squares estimators b_0 and b_1 they have the properties of all least squares estimators:

   (a) They are unbiased.

   (b) They have minimum variance among all unbiased linear estimators.

   (c) In addition, the maximum likelihood estimators $\hat{\beta}_0$ and $\hat{\beta}_1$, for the normal error regression model have other desirable properties: consistent (A. 52), sufficient (A.53) and the minimum variance unbiased estimators (linear or otherwise).

Question .... (p72)

Assume the normal error regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i.$$

Find the estimation of parameters using method of maximum likelihood.

sol:

TA Class

- **Problems**: 1.6, 1.7, 1.18, 1.20, 1.24
- **Exercises**: 1.32, 1.33, 1.35, 1.36, 1.41
- **Projects**: 1.43

"少花點時間去取悅別人，多花些時間來經營自己。"
"Spend a little more time trying to make something of yourself and a little less time trying to impress people."

— 早餐俱樂部 (Breakfast Club, 1985)

Regression Analysis (I)

Kutner's Applied Linear Statistical Models (5/E)

Chapter 2: Inferences in Regression and Correlation Analysis

Thursday 09:10-12:00, 商館 260205

Han-Ming Wu

Department of Statistics, National Chengchi University

http://www.hmwu.idv.tw

Overview

1. Take up inferences (interval estimation and tests) concerning the regression parameters β_0 and β_1.

2. Discuss interval estimation of the mean $E(Y)$ of the probability distribution of Y, for given X, prediction intervals for a new observation Y, confidence bands for the regression line, the analysis of variance approach to regression analysis, the general linear test approach, and descriptive measures of association.

3. Assume that the normal error regression model (1.24) is applicable:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

where β_0 and β_1, are parameters, X_i are known constants, ϵ_i are independent $N(0, \sigma^2)$.

2.1 Inferences Concerning β_1

1. Testing whether or not $\beta_1 = 0$ is that, when $\beta_1 = 0$, there is no linear association between Y and X.

$$E(Y) = \beta_0 + 0 \cdot X = \beta_0$$

2. For normal error regression model (2.1), the condition $\beta_1 = 0$ follows that the probability distributions of Y are identical. There is no relation of any type between Y and X.

Sampling Distribution of $\hat{\beta}_1$

✎ Question .... (p42)

For normal error regression model (2.1), show that b_1, the point estimator of β_1, is a linear combination of the observation Y_i. That is

$$b_1 = \sum k_i Y_i, \quad \text{where} \quad k_i = \frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2}.$$

sol:

✎ Question .... (p42)

For normal error regression model (2.1), if b_1 is expressed as $b_1 = \sum k_i Y_i$, show that

$$\sum k_i = 0, \ \sum k_i X_i = 1, \text{and} \ \sum k_i^2 = \frac{1}{\sum (X_i - \bar{X})^2}.$$

sol:

Question .... (p41)

For normal error regression model (2.1), show that the sampling distribution of b_1, the point estimator of β_1, is normal, with mean and variance:

$$E(b_1) = \beta_1, \quad \text{and} \quad \sigma^2(b_1) = \frac{\sigma^2}{\sum(X_i - \bar{X})^2} = \sum k_i^2 \sigma^2.$$

sol:

Question .... (p43)

Show that b_1 has minimum variance among all unbiased linear estimator of the form:

$$\hat{\beta}_1 = \sum c_i Y_i,$$

where the c_i are arbitrary constants.

sol:

Sampling Distribution of $(b_1 - \beta_1)/s(b_1)$

1. Since b_1 is normally distributed, we know that the standardized statistic $(b_1 - \beta_1)/\sigma(b_1)$ is a standard normal variable.

2. We need to estimate $\sigma(b_1)$ by $s(b_1)$, and hence are interested in the distribution of the statistic $(b_1 - \beta_1)/s(b_1)$.

3. When a statistic is standardized but the denominator is an estimated standard deviation rather than the true standard deviation, it is called a studentized statistic.

✎ Question .... (p44)

Show the studentized statistic $\dfrac{b_1 - \beta_1}{s(b_1)}$ is distributed as $t_{(n-2)}$ for regression model (2.1).

sol:

Confidence Interval for β_1

✎ Question .... (p45)

Find the $(1-\alpha)\%$ confidence interval for β_1.

sol:

✎ Question .... (p45)

(Toluca Company Example) Management wishes an estimate of β_1, with 95 percent confidence coefficient.

sol:

Obtain

$$s^2(b_1) = \frac{MSE}{\sum(X_i - \bar{X})^2} = \frac{2,384}{19,800} = 0.12040, \quad s(b_1) = 0.3470.$$

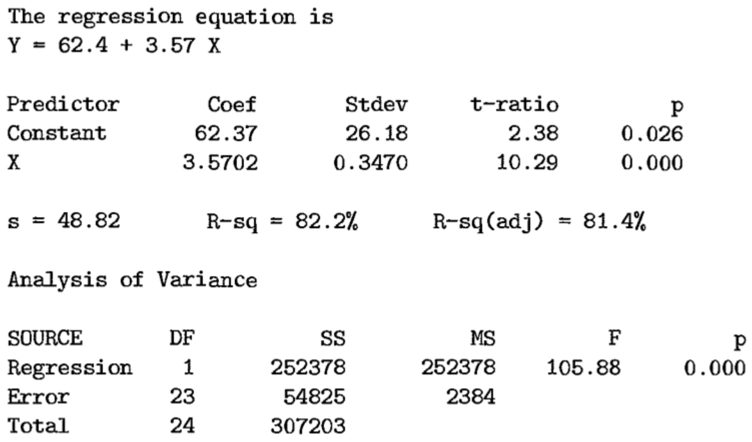
For a 95 percent confidence coefficient, we find $t_{(0.975;23)} = 2.069$. The 95 percent confidence interval:

$$3.5702 - 2.069(0.3470) \leq \beta_1 \leq 3.5702 + 2.069(0.3470)$$

$$\Rightarrow 2.85 \leq \beta_1 \leq 4.29$$

Thus, with confidence coefficient .95, we estimate that the mean number of work hours increases by somewhere between 2.85 and 4.29 hours for each additional unit in the lot.

FIGURE 2.2 **Portion of MINITAB Regression Output—Toluca Company Example.**



```
The regression equation is
Y = 62.4 + 3.57 X

Predictor        Coef       Stdev     t-ratio        p
Constant        62.37       26.18        2.38    0.026
X              3.5702      0.3470       10.29    0.000

s = 48.82       R-sq = 82.2%      R-sq(adj) = 81.4%

Analysis of Variance

SOURCE       DF         SS          MS        F          p
Regression    1     252378      252378   105.88      0.000
Error        23      54825        2384
Total        24     307203
```

Tests Concerning β_1

Question .... (p47)

Two-Sided Test A cost analyst in the Toluca Company is interested in testing, using regression model (2.1), whether or not there is a linear association between work hours and lot size, i.e., whether or not, $\beta_1 = 0$. Please conduct the Two-Sided Test for this problem and control the risk of a Type I error at $\alpha = 0.05$.

sol:

✎ Question .... (p47)

One-Sided Test Suppose the analyst in the Toluca Company had wished to test whether or not , β_1, is positive, controlling the level of significance at $\alpha = 0.05$. Please conduct the One-Sided Test for this problem.

sol:

Comments:

1. The P-value is sometimes called the observed level of significance.
2. Many scientific publications commonly report the P-value together with the value of the test statistic. In this way, one can conduct a test at any desired level of significance a by comparing the P-value with the specified level α.
3. Users of statistical calculators and computer packages need to be careful to ascertain whether one-sided or two-sided P-values are reported.
4. It is desired to test whether or not β_1 equals some specified nonzero value β_{10}. The alternatives are:

$$H_0 : \beta_1 = \beta_{10} \quad \text{versus} \quad H_a : \beta_1 \neq \beta_{10}$$

and the appropriate test statistic is:

$$t^* = \frac{b_1 - \beta_{10}}{s(b_1)}$$

2.2 Inferences Concerning β_0

1. The point estimator b_0: $b_0 = \bar{Y} - b_1\bar{X}$.

2. The sampling distribution of b_0 is normal, with mean and variance:

$$E(b_0) = \beta_0, \quad \sigma^2(b_0) = \sigma^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right]$$

3. An estimator of $\sigma^2(b_0)$ is obtained by replacing σ^2 by its point estimator MSE:

$$s^2(b_0) = MSE\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right]$$

4. The sampling distribution of $(b_0 - \beta_0)/s(b_0)$ is $t_{(n-2)}$ for regression model (2.1)

5. The confidence intervals for β_0 is $\beta_0 \pm t_{(1-\alpha/2;n-2)}s(b_0)$.

2.3 Some Considerations on Making Inferences Concerning β_0 and β_1

Effects of Departures from Normality

1. If the probability distributions of Y are not exactly normal but do not depart seriously, the sampling distributions of b_0 and b_1 will be approximately normal, and the use of the t distribution will provide approximately the specified confidence coefficient or level of significance.

2. Even if the distributions of Y are far from normal, the estimators b_0 and b_1 generally have the property of asymptotic normality - their distributions approach normality under very general conditions as the sample size increases.

Interpretation of Confidence Coefficient and Risks of Errors

1. Since regression model (2.1) assumes that the X_i are known constants, the confidence coefficient and risks of errors are interpreted with respect to taking repeated samples in which the X observations are kept at the same levels as in the observed sample.

2. (Toluca Company Example) The meaning of a confidence interval (CI) for β_1, with confidence coefficient 0.95: if many independent samples are taken where the levels of X (the lot sizes) are the same as in the data set and a 95 percent confidence interval is constructed for each sample, 95 percent of the intervals will contain the true value of β_1.

Spacing of the X levels

1. For given n and σ^2, the variances of b_1 and b_0 are affected by the spacing of the X levels in the observed data.

2. The greater is the spread in the X levels, the larger is the quantity $\sum(X_i - \bar{X})^2$ and the smaller is the variance of b_1.

Power of Tests

(NOTE: The power of tests on β_0 and β_1, can be obtained from Appendix Table B.5.)

1. The general test concerning β_1:

$$H_0 : \beta_1 = \beta_{10} \quad \text{versus} \quad H_a : \beta_1 \neq \beta_{10}$$

2. Test statistic: $t^* = \dfrac{b_1 - \beta_{10}}{s(b_1)}$.

3. Decision rule for level of significance α:

$$\text{If } |t^*| \leq t_{(1-\alpha/2;n-2)}, \text{ conclude } H_0.$$

$$\text{If } |t^*| > t_{(1-\alpha/2;n-2)}, \text{ conclude } H_a.$$

4. The power of this test is the probability that the decision rule will lead to conclusion H_a when H_a in fact holds:

$$\text{Power} = P(|t^*| > t_{(1-\alpha/2;n-2)} |\ \delta)$$

where δ is the noncentrality measure - i.e., a measure of how far the true value of β_1, is from β_{10}:

$$\delta = \frac{|\beta_1 - \beta_{10}|}{\sigma(b_1)}$$

✎ Question . . . . (p51)

In Toluca Company example, conduct the test for:

$$H_0 : \beta_1 = \beta_{10} = 0, \quad \text{versus} \quad H_a : \beta_1 \neq \beta_{10} = 0.$$

Calculate the power of the test when $\beta_1 = 1.5$.

sol:

2.4 Interval Estimation of $E(Y_h)$

1. Let $\underline{\quad X_h \quad}$ denote the level of X for which we wish to estimate the mean response.
2. X_h may be a value which occurred in the sample, or it may be some other value of the predictor variable within the scope of the model.
3. The mean response when $X = X_h$ is denoted by $\underline{\quad E(Y_h) \quad}$. The point estimator Y_h of $E(Y_h)$ is $\underline{\quad \hat{Y}_h = b_0 + b_1 X_h \quad}$.

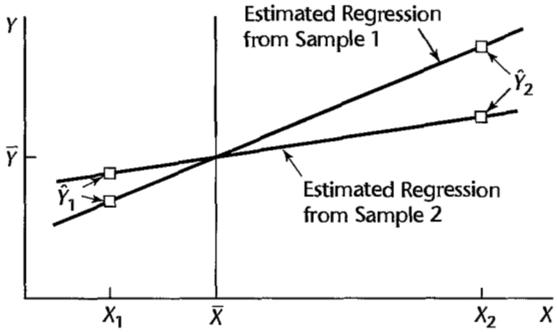
✎ Question . . . . (p52)

For normal error regression model, show that the sampling distribution of $\hat{Y}_h$ is normal, with mean and variance:

$$E(\hat{Y}_h) = E(Y_h) \quad \text{and} \quad \sigma^2(\hat{Y}_h) = \sigma^2 \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right].$$

sol:

FIGURE 2.3 Effect on $\hat{Y}_h$ of Variation in b_1 from Sample to Sample in Two Samples with Same Means $\bar{Y}$ and $\bar{X}$.



The variability of the sampling distribution of $\hat{Y}_h$ is affected by how far X_h is from $\bar{X}$ through the term $(X_h - \bar{X})^2$.

Sampling Distribution of $(\hat{Y}_h - E(Y_h))/s(\hat{Y}_h)$

1. $\dfrac{\hat{Y}_h - E(Y_h)}{s(\hat{Y}_h)}$ is distributed as $t_{(n-2)}$ for regression model (2.1).

Confidence Interval for $E(Y_h)$

1. A $(1-\alpha)\%$ confidence interval for $E(Y_h)$ is

$$\hat{Y}_h \pm t_{(1-\alpha/2;n-2)} s(\hat{Y}_h), \quad s(\hat{Y}_h) = MSE\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right].$$

Question .... (p54)

In the Toluca Company example, find a 90% CI for $E(Y_h)$ when the lot size is $X_h = 65$ units.

sol:

2.5 Prediction of New Observation

The new observation on Y to be predicted is viewed as the result of a new trial, independent of the trials on which the regression analysis is based. We denote the level of X for the new trial as $\underline{\quad X_h \quad}$ and the new observation on Y as $\underline{\quad Y_{h(new)} \quad}$.

Prediction Interval for $Y_{h(new)}$ when Parameters Known

In general, when the regression parameters of normal error regression model (2.1) are known, the $(1-\alpha)\%$ prediction limits for $Y_{h(new)}$ are:

$$E(Y_h) \pm z_{(1-\alpha/2)}\sigma$$

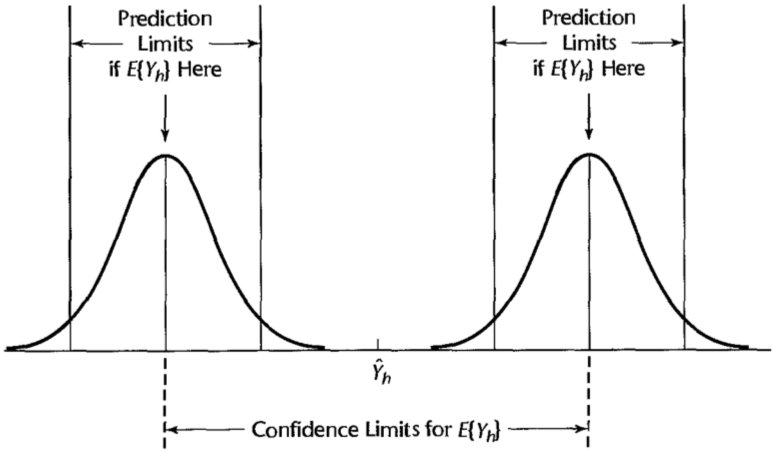
Prediction Interval for $Y_{h(new)}$ when Parameters Unknown

✎ Question .... (p58)

As we know, $\dfrac{Y_{h(new)} - \hat{Y}_h}{s(\text{pred})}$ is distributed as $t_{(n-2)}$ for a normal error regression model. Find the prediction limits for a new observation $Y_{h(new)}$ at a given level X_h.

sol:

FIGURE 2.5 **Prediction of $Y_{h(\text{new})}$ when Parameters Unknown.**


Prediction Limits if $E\{Y_h\}$ Here
Prediction Limits if $E\{Y_h\}$ Here
$\hat{Y}_h$
Confidence Limits for $E\{Y_h\}$

Question .... (p59)

The Toluca Company studied the relationship between lot size and work hours primarily to obtain information on the mean work hours required for different lot sizes for use in determining the optimum lot size. The company was also interested, however, to see whether the regression relationship is useful for predicting the required work hours for individual lots. Find a 90 percent prediction interval for the number of work hours for the next production runs of $X_h = 100$ units.

sol:

Prediction of Mean of m New Observations for Given X_h

1. Denote the mean of m new Y observations to be predicted as $\bar{Y}_{h(new)}$. The $1-\alpha$ prediction limits are, assuming that the new m Y observations are independent:

$$\hat{Y}_h \pm t_{(1-\alpha/2;n-2)} s(\text{predmean})$$

where

$$s^2(\text{predmean}) = \frac{MSE}{m} + s^2(\hat{Y}_h)$$

or equivalently:

$$s^2(\text{predmean}) = MSE\left[\frac{1}{m} + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right].$$

✎ Question .... (p61)

In the Toluca Company example, find the 90 percent prediction interval for the mean number of work hours $\bar{Y}_{h(new)}$ in three new production runs, each for $X_h = 100$ units.

sol:

2.6 Confidence-Band for Regression Line

1. A confidence band for the entire regression line $E(Y) = \beta_0 + \beta_1 X$ enables us to see the region in which the entire regression line lies. It is particularly useful for determining the appropriateness of a fitted regression function.

2. The Working-Hotelling $(1-\alpha)\%$ confidence band for the regression line for regression model (2.1) has the following two boundary values at any level X_h:

$$\underline{\hat{Y}_h \pm W s(\hat{Y}_h)}, \quad \text{where} \quad \underline{W^2 = 2F_{(1-\alpha;2,n-2)}}.$$

Question .... (p62)

Find the 90 percent confidence band for the regression line to determine how precisely we have been able to estimate the regression function for the Toluca Company example.

sol:

2.7 Analysis of Variance Approach to Regression Analysis

Partitioning of Total Sum of Squares

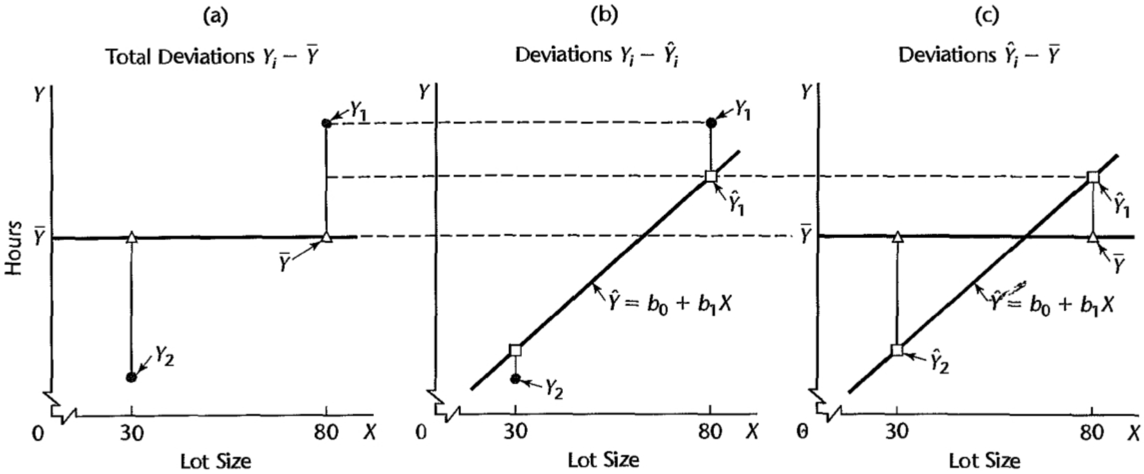
1. The variation is measured in terms of the deviations of the Y_i around their mean $\bar{Y}$: $\underline{Y_i - \bar{Y}}$.

2. $SSTO$ (total sum of squares): the measure of total variation is the sum of the squared deviations: $\underline{SSTO = \sum(Y_i - \bar{Y})^2}$.

3. SSE (error sum of squares): the measure of variation in Y_i that is present when the predictor variable X is taken into account: $\underline{SSE = \sum(Y_i - \hat{Y}_i)^2}$.

4. SSR (regression sum of squares): $\underline{SSR = \sum(\hat{Y}_i - \bar{Y})^2}$.

FIGURE 2.7 Illustration of Partitioning of Total Deviations $Y_i - \bar{Y}$—Toluca Company Example (not drawn to scale; only observations Y_1 and Y_2 are shown).



✐ Question .... (p65)

Show that $SSTO = SSR + SSE$. That is

$$\sum(Y_i - \bar{Y})^2 = \sum(\hat{Y}_i - \bar{Y})^2 + \sum(Y_i - \hat{Y}_i)^2$$

sol:

Breakdown of Degrees of Freedom

1. Corresponding to the partitioning of the total sum of squares $SSTO$, there is a partitioning of the associated degrees of freedom (df).

2. SSTO has $\underline{\ n-1\ }$ degrees of freedom associated with it. One degree of freedom is lost because the deviations $\underline{\ Y_i - \bar{Y}\ }$ are subject to one constraint: they must sum to $\underline{\text{ zero }}$. Equivalently, one degree of freedom is lost because the sample mean $\bar{Y}$ is used to estimate the population mean.

3. SSE has $\underline{\ n-2\ }$ degrees of freedom associated with it. Two degrees of freedom are lost because the two parameters $\underline{\ \beta_0 \text{ and } \beta_1\ }$ are estimated in obtaining the fitted values $\hat{Y}_i$.

4. SSR has $\underline{\text{ one }}$ degree of freedom associated with it. Although there are n deviations $\underline{\ \hat{Y}_i - \bar{Y}\ }$, all fitted values $\hat{Y}_i$ are calculated from the same estimated regression line.

Mean Squares

1. A sum of squares divided by its associated degrees of freedom is called a $\underline{\text{ mean square }}$ (MS).

2. The regression mean square: $\underline{MSR = \dfrac{SSR}{1} = SSR}$.

3. The error mean square: $\underline{MSE = \dfrac{SSE}{n-2}}$.

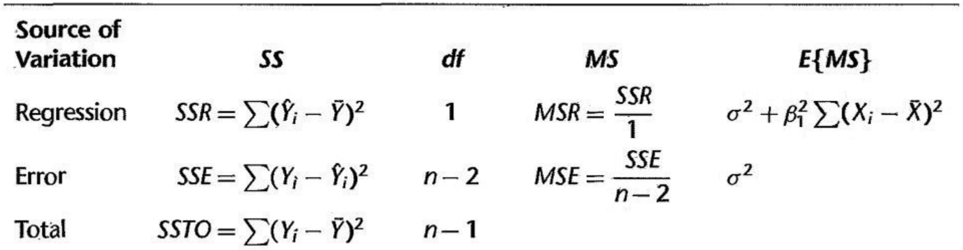
Analysis of Variance Table

1. **Basic Table**:

   (a) The breakdowns of the total sum of squares and associated degrees of freedom are displayed in the form of an analysis of variance table $\underline{\text{ (ANOVA table) }}$ in Table 2.2.

   (b) The ANOVA table contains a column of $\underline{\text{ expected mean squares }}$ that will be utilized.

TABLE 2.2 ANOVA Table for Simple Linear Regression.



Source of Variation	SS	df	MS	E{MS}
Regression	$SSR = \sum(\hat{Y}_i - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$	$\sigma^2 + \beta_1^2 \sum(X_i - \bar{X})^2$
Error	$SSE = \sum(Y_i - \hat{Y}_i)^2$	$n-2$	$MSE = \frac{SSE}{n-2}$	σ^2
Total	$SSTO = \sum(Y_i - \bar{Y})^2$	$n-1$		

2. **Modified Table**:

(a) The modified ANOVA table is based on the fact that the total sum of squares can be decomposed into two parts:

$$SSTO = \underline{\sum(Y_i - \bar{Y})^2} = \underline{\sum Y_i^2 - n\bar{Y}^2}$$

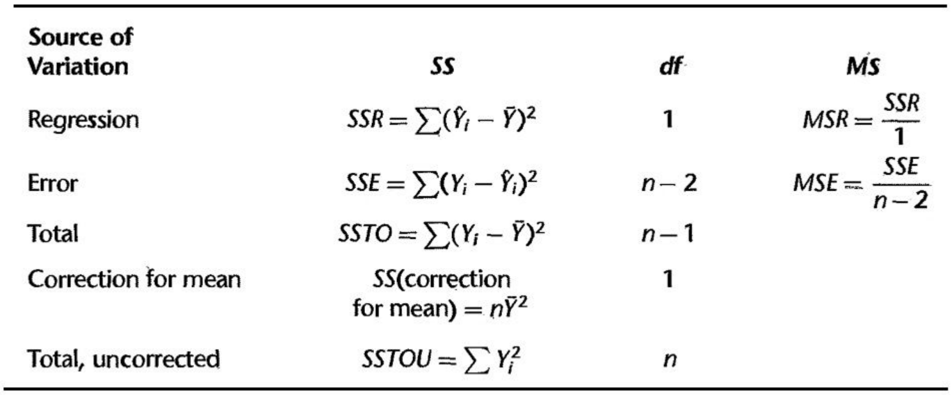
(b) In the modified ANOVA table, the total uncorrected sum of squares, denoted by SSTOU, is defined as:

$$SSTOU = \underline{\sum Y_i^2}$$

and the correction for the mean sum of squares, denoted by SS(correction for mean), is defined as:

$$SS(\text{correction for mean}) = \underline{n\bar{Y}^2}$$

TABLE 2.3 Modified ANOVA Table for Simple Linear Regression.



Source of Variation	SS	df	MS
Regression	$SSR = \sum(\hat{Y}_i - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$
Error	$SSE = \sum(Y_i - \hat{Y}_i)^2$	$n-2$	$MSE = \frac{SSE}{n-2}$
Total	$SSTO = \sum(Y_i - \bar{Y})^2$	$n-1$	
Correction for mean	$SS(\text{correction for mean}) = n\bar{Y}^2$	1	
Total, uncorrected	$SSTOU = \sum Y_i^2$	n	

Expected Mean Squares

Question .... (p68)

Show that

$$\begin{aligned} E(MSE) &= \sigma^2, \quad \text{and} \\ E(MSR) &= \sigma^2 + \beta_1^2 \sum (X_i - \bar{X})^2. \end{aligned}$$

sol:

F Test of $\beta_1 = 0$ versus $\beta_1 \neq 0$

1. The analysis of variance provides us with a test for:

   $$\underline{\quad H_0 : \beta_1 = 0 \quad} \quad \text{versus} \quad \underline{\quad H_a : \beta_1 \neq 0 \quad}.$$

2. **Test Statistic**: The test statistic for the analysis of variance approach is denoted by F^*:

   $$F^* = \underline{\quad \frac{MSR}{MSE} \quad}$$

3. Large values of F^* support $\underline{\quad H_a \quad}$ and values of F^* near $\underline{\quad 1 \quad}$ support H_0.

Question .... (p70)

Show that if H_0 holds, F^* follows the $F_{(1,n-2)}$ distribution.

sol:

1. **Construction of Decision Rule**: Since the test is upper-tail and F^* is distributed as $F_{(1,n-2)}$ when H_0 holds, the decision rule is as follows when the risk of a Type I error is to be controlled at α:

$$\text{If } \underline{\quad F^* \leq F_{(1-\alpha;1,n-2)} \quad}, \text{ conclude } H_0,$$

$$\text{If } F^* > F_{(1-\alpha;1,n-2)}, \text{ conclude } H_a$$

✎ Question .... (p71)

For the Toluca Company example, conduct a F test for $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$.

sol:

✎ Question .... (p71)

Show that for a given α level, the F test of $\beta_1 = 0$ versus $\beta_1 \neq 0$ is equivalent algebraically to the two-sided t test.

sol:

2.8 General Linear Test Approach

Full Model

1. For the simple linear regression case, the full model or unrestricted model is the normal error regression model:

$$\underline{Y_i = \beta_0 + \beta_1 X_i + \epsilon_i}.$$

2. The error sum of squares for the full model:

$$SSE(F) = \underline{\sum(Y_i - (b_0 + b_1 X_i))^2} = \underline{\sum(Y_i - \hat{Y})^2} = \underline{SSE}.$$

3. $SSE(F)$ measures the variability of the Y_i observations around the fitted regression line.

Reduced Model

1. Consider $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$, the model when H_0 holds is called the reduced or restricted model:

$$\underline{Y_i = \beta_0 + \epsilon_i}.$$

2. The error sum of squares for the reduced model:

$$SSE(R) = \underline{\sum(Y_i - b_0)^2} = \underline{\sum(Y_i - \bar{Y})^2} = \underline{SSTO}.$$

Test Statistic

1. It can be shown that $SSE(F)$ never is greater than $SSE(R)$:

$$\underline{SSE(F) \leq SSE(R)}.$$

2. The actual test statistic is a function of $SSE(R) - SSE(F)$,

$$F^* = \underline{\left(\frac{SSE(R) - SSE(F)}{df_R - df_F}\right) / \left(\frac{SSE(F)}{df_F}\right)},$$

which follows the F distribution when H_0 holds.

3. The decision rule therefore is:

$$\text{If } \underline{F^* \le F_{(1-\alpha; df_R - df_F, df_F)}}, \text{ conclude } H_0$$

$$\text{If } F^* > F_{(1-\alpha; df_R - df_F, df_F)}, \text{ conclude } H_a$$

4. For testing whether or not $\beta_1 = 0$, we therefore have:

$$SSE(R) = SSTO, \quad SSE(F) = SSE, \quad df_R = n-1, \quad df_F = n-2,$$

so that we obtain

$$F^* = \underline{\left(\frac{SSTO - SSE}{(n-1)-(n-2)}\right) \Big/ \left(\frac{SSE}{n-2}\right)} = \underline{\frac{SSR}{1} \Big/ \frac{SSE}{n-2}} = \underline{\frac{MSR}{MSE}}$$

which is identical to the analysis of variance test statistic.

2.9 Descriptive Measures of Linear Association between X and Y

Coefficient of Determination

1. The coefficient of determination R^2 is defined to measure the effect of X in reducing the variation in Y. It is expressed as the reduction in variation $\underline{(SSTO - SSE = SSR)}$ as a proportion of the total variation:

$$R^2 = \underline{\frac{SSR}{SSTO}} = \underline{1 - \frac{SSE}{SSTO}}.$$

2. We may interpret R^2 ($\underline{0 \le R^2 \le 1}$) as the proportionate reduction of total variation associated with the use of the predictor variable X.

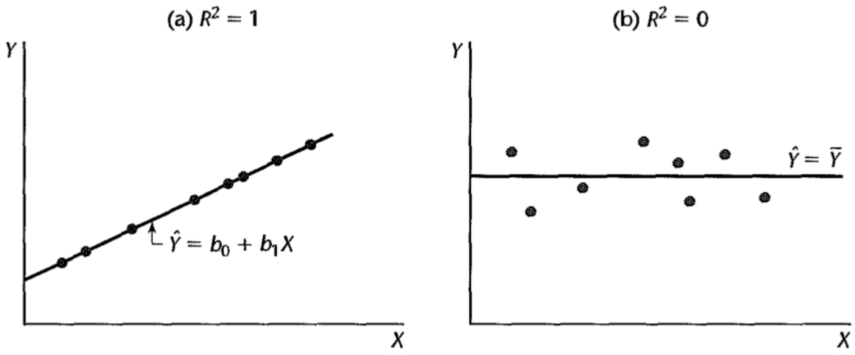
3. The larger R^2 is, the more the total <u>variation</u> of Y is reduced by introducing the predictor variable X.

4. The limiting values of R^2 may occur:

   (a) When all observations fall on the fitted regression line, then $\underline{SSE = 0}$ and $\underline{R^2 = 1}$. The predictor variable X accounts for <u>all variation</u> in the observations Y_i

(b) When the fitted regression line is horizontal so that $\underline{\quad b_1 = 0 \quad}$ and $\underline{\quad \hat{Y}_i = \bar{Y} \quad}$, then $\underline{\quad SSE = SSTO \quad}$ and $\underline{\quad R^2 = 0 \quad}$. There is no linear association between X and Y in the sample data.

FIGURE 2.8 Scatter Plots when $R^2 = 1$ and $R^2 = 0$.



limitations of R^2: three common misunderstandings

1. **Misunderstanding 1**: A high R^2 indicates that ___useful predictions___ can be made. (not necessarily correct)

   (a) (Toluca Company Example) the coefficient of determination was high ($R^2 = 0.82$). Yet the 90 percent prediction interval for the next lot, consisting of 100 units, was wide (332 to 507 hours) and not precise enough to permit management to schedule workers effectively.

   (b) Misunderstanding 1 arises because R^2 measures only a ___relative reduction___ from $SSTO$ and provides no information about absolute precision for estimating a mean response or predicting a new observation.

2. **Misunderstanding 2**: A high R^2 indicates that the estimated regression line is a ___good fit___. (not necessarily correct)

   (a) (Figure 2.9a) a scatter plot where R^2 is high ($R^2 = 0.69$). Yet a linear regression function would not be a good fit since the regression relation is curvilinear.

3. **Misunderstanding 3**: A R^2 near zero indicates that X and Y are not related. (not necessarily correct).

(a) (Figure 2.9b) a scatter plot where R^2 between X and Y is $R^2 = 0.02$. Yet X and Y are strongly related; however, the relationship between the two variables is curvilinear.

(b) Misunderstandings 2 and 3 arise because R^2 measures the degree of linear association between X and Y, whereas the actual regression relation may be curvilinear.

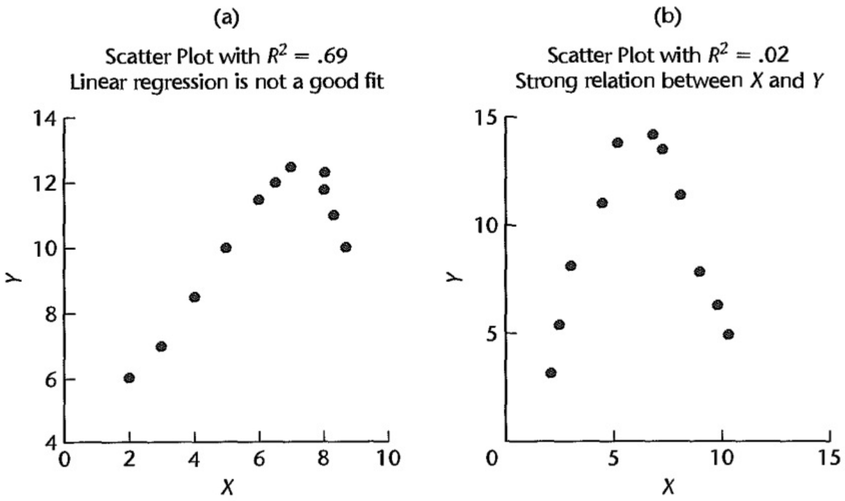


FIGURE 2.9 **Illustrations of Two Misunderstandings about Coefficient of Determination.**

Coefficient of Correlation

1. A measure of linear association between Y and X when both Y and X are random is the coefficient of correlation. This measure is the signed square root of R^2:

$$r = \pm\sqrt{R^2}$$

2. A plus or minus sign is attached to this measure according to whether the slope of the fitted regression line is positive or negative. Thus, the range of r is: $-1 \le r \le 1$.

2.10 Considerations in Applying Regression Analysis*

2.11 Normal Correlation Models*

😍 TA Class

- **Problems**: 2.5, 2.8, 2.10, 2.14, 2.17, 2.24, 2.30, 2.31, 2.32
- **Exercises**: 2.50, 2.55
- **Projects**: 2.62

“永遠不要讓別人的冷漠，影響了你對這世界的熱情。”
“Never allow the indifference of others to affect your passion for this world.”
— *魔女宅急便 (Kiki’s Delivery Service, 1989)*

Regression Analysis (I)

Kutner's Applied Linear Statistical Models (5/E)

Chapter 3: Diagnostics and Remedial Measures

Thursday 09:10-12:00, 商館 260205

Han-Ming Wu

Department of Statistics, National Chengchi University

http://www.hmwu.idv.tw

Overview

1. The features of the model, such as linearity of the regression function or normality of the error terms, may not be appropriate for the particular data.

2. It is important to examine the aptness of the model for the data before inferences based on that model are undertaken.

3. Use some simple graphic methods to study the appropriateness of a model, as well as some formal statistical tests.

4. Consider some remedial techniques that can be helpful when the data are not in accordance with the conditions of regression model (2.1).

3.1 Diagnostics for Predictor Variable

1. Diagnostic for the predictor variable to see if there are any outlying X values that could influence the appropriateness of the fitted regression function.

2. **Example**: Toluca Company Example

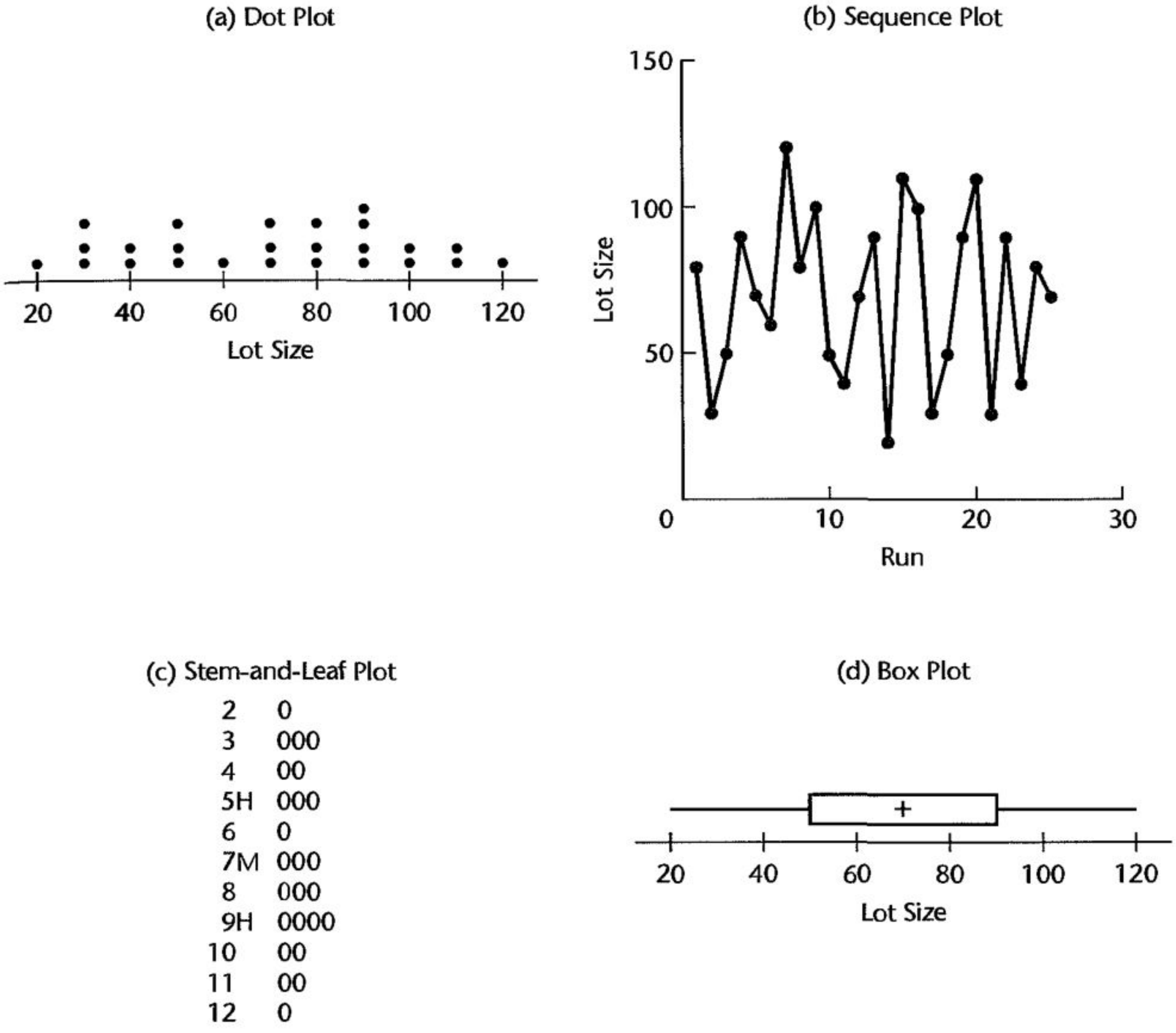
   (a) (Figure 3.1a) The dot plot: the minimum and maximum lot sizes are 20 and 120, respectively, that the lot size levels are spread throughout this interval, and that there are no lot sizes that are far outlying.

(b) (Figure 3.1b) The sequence plot : lot size is plotted against production run (i.e., against time sequence). The plot had shown that smaller lot sizes had been utilized early on and larger lot sizes later on.

(c) (Figure 3.1c) The stem-and-leaf plot : provides information similar to a frequency histogram . The letter M denotes the median, and the letter H denotes the first and third quartiles.

(d) (Figure 3.1d) The box plot : the middle half of the lot sizes range from 50 to 90, and that they are fairly symmetrically distributed because the median is located in the middle of the central box.

FIGURE 3.1 MINITAB and SYGRAPH Diagnostic Plots for Predictor Variable—Toluca Company Example.



3.2 Residuals

1. Diagnostics for the response variable are usually carried out indirectly through an examination of the residuals.

2. The residual e_i is the difference between the observed value Y_i and the fitted value $\hat{Y}_i$: $e_i = Y_i - \hat{Y}_i$.

3. The residual may be regarded as the observed error, in distinction to the unknown true error ϵ_i in the regression model: $\epsilon_i = Y_i - E(Y_i)$.

4. For regression model (2.1), the error terms ϵ_i are assumed to be independent normal random variables, with mean 0 and constant variance σ^2.If the model is appropriate for the data at hand, the observed residuals e_i should then reflect the properties assumed for the ϵ_i.

Properties of Residuals

1. **Mean**

   (a) The mean of the n residuals e_i for the simple linear regression model (2.1) is always 0: $\bar{e} = \sum e_i/n = 0$.

   (b) It provides no information as to whether the true errors ϵ_i have expected value $E(\epsilon_i) = 0$.

2. **Variance**

   (a) The variance of the n residuals e_i for regression model is

   $$s^2 = \frac{\sum(e_i - \bar{e})^2}{n-2} = \frac{\sum e_i^2}{n-2} = \frac{SSE}{n-2} = MSE.$$

   (b) If the model is appropriate, MSE is an unbiased estimator of the variance of the error terms σ^2.

3. **Nonindependence**

   (a) The residuals e_i are not independent random variables because they involve the fitted values $\hat{Y}_i$ which are based on the same fitted regression function.

(b) The residuals for regression model (2.1) are subject to two constraints. These are constraint (1.17) - $\sum e_i = 0$ - and constraint (1.19) - $\sum X_i e_i = 0$.

(c) When the sample size is large in comparison to the number of parameters in the regression model, the dependency effect among the residuals e_i is relatively unimportant and can be ignored for most purposes.

Semistudentized Residuals

1. Since the standard deviation of the error terms ϵ_i is σ, which is estimated by $\sqrt{MSE}$, it is natural to consider the semistudentized residuals:

$$e_i^* = \frac{e_i - \bar{e}}{\sqrt{MSE}} = \frac{e_i}{\sqrt{MSE}}$$

2. Both semistudentized residuals and studentized residuals can be very helpful in identifying outlying observations. (details in Chapter 10)

Departures from Model to Be Studied by Residuals

1. We shall consider the use of residuals for examining six important types of departures from the simple linear regression model (2.1) with normal errors:

   (a) The regression function is not linear.

   (b) The error terms do not have constant variance.

   (c) The error terms are not independent.

   (d) The model fits all but one or a few outlier observations.

   (e) The error terms are not normally distributed.

   (f) One or several important predictor variables have been omitted from the model.

3.3 Diagnostics for Residuals

1. Some informal diagnostic plots of residuals to provide information on whether any of the six types of departures from the simple linear regression model (2.1)

   (a) Plot of residuals against __predictor__ variable.

   (b) Plot of __absolute__ or __squared__ residuals against predictor variable.

   (c) Plot of residuals against __fitted values__. (the most important)

   (d) Plot of residuals against __time__ or other sequence.

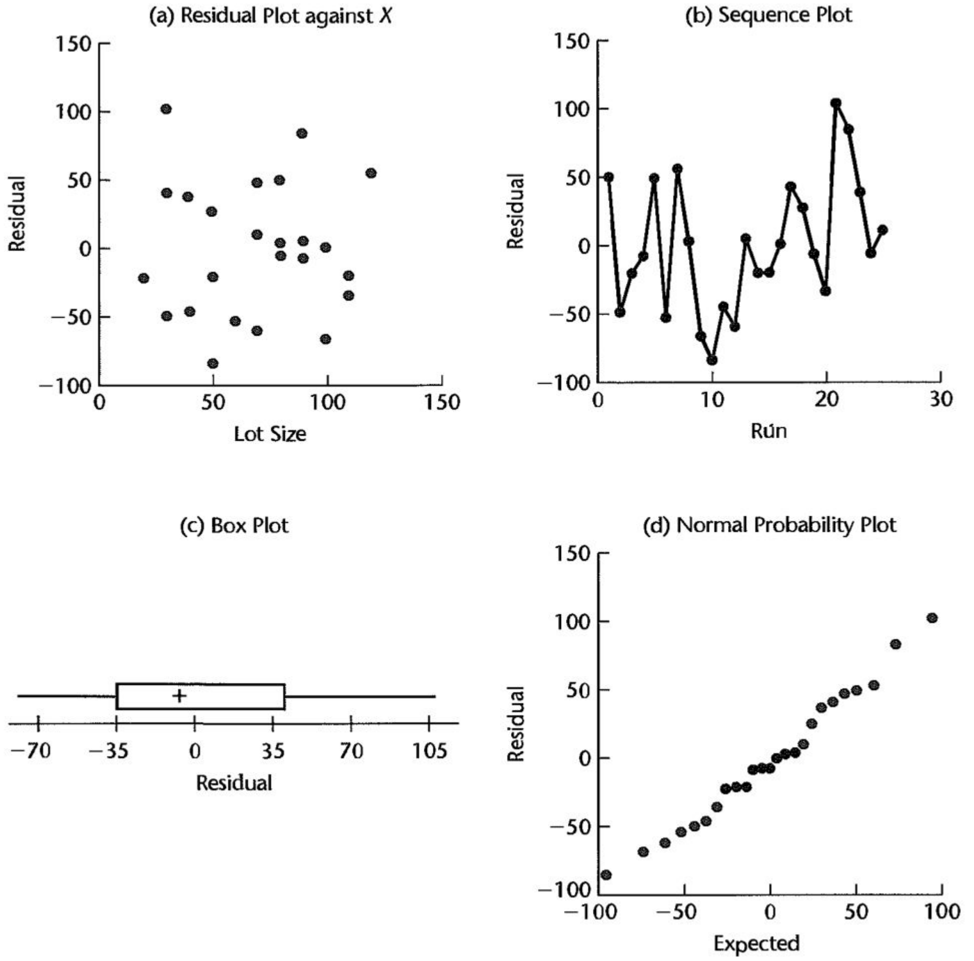
   (e) Plots of residuals against __omitted predictor__ variables.

   (f) Box plot of residuals.

   (g) __Normal probability plot__ of residuals.

2. (Figure 3.2) Toluca Company example: plots of the residuals against the predictor variable and against time, a box plot, and a normal probability plot.

FIGURE 3.2 MINITAB and SYGRAPH Diagnostic Residual Plots—Toluca Company Example.



(a) Residual Plot against X

(b) Sequence Plot

(c) Box Plot

(d) Normal Probability Plot

Nonlinearity of Regression Function

1. **Residual plot**: whether a linear regression function is appropriate for the data being analyzed can be studied from a residual plot against the fitted values.

2. Nonlinearity of the regression function can also be studied from a scatter plot of X and Y, but this plot is not always as effective as a residual plot.

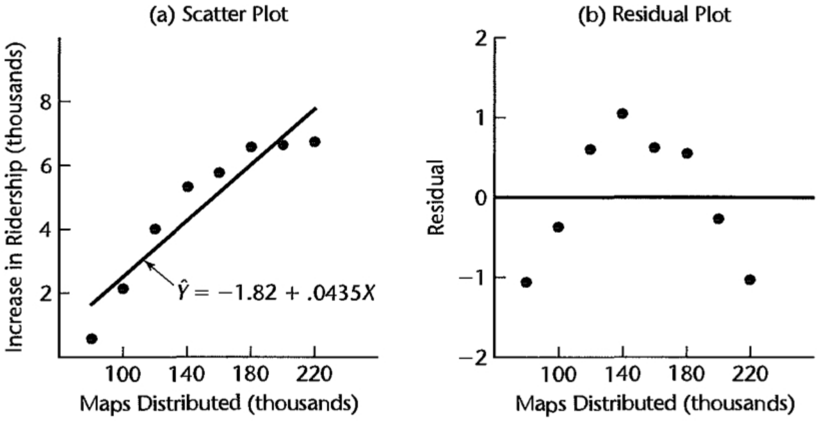
3. $\boxed{\text{Example}}$ Ridership - Transit Example (Figure 3.3)(TABLE 3.1)

   (a) One would like to study the relation between maps distributed and bus ridership in eight test cities. Let X be the number of bus transit maps distributed free to residents of the city at the beginning of the test period and Y be the increase during the test period in average daily bus ridership during nonpeak hours.

(b) (Figures 3.3) the lack of linearity of the regression function.

(c) In general, the residual plot is to be preferred. It can clearly show any systematic pattern in the deviations around the fitted regression line.

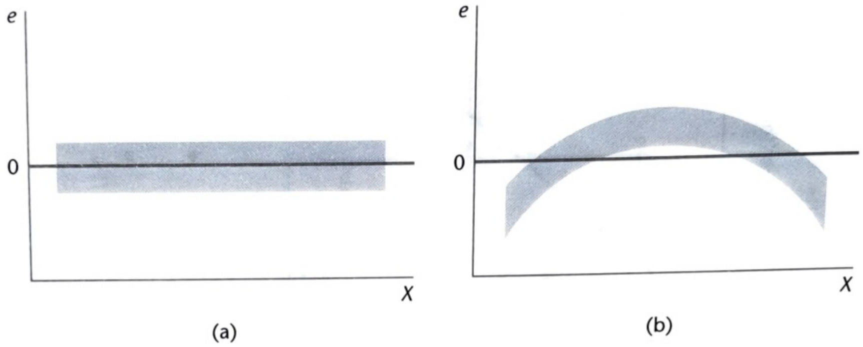
FIGURE 3.3 Scatter Plot and Residual Plot Illustrating Nonlinear Regression Function—Transit Example.



4. (Figure 3.4a) the residual plot against X when a linear regression model is appropriate. The residuals then fall within a horizontal band centered around 0, displaying no systematic tendencies to be positive and negative.

5. (Figure 3.4b) a departure from the linear regression model that indicates the need for a curvilinear regression function. Here the residuals tend to vary in a systematic fashion between being positive and negative.

FIGURE 3.4 Prototype Residual Plots.



Nonconstancy of Error Variance

1. The residuals plot is also helpful to examine whether the variance of the error terms is constant.

2. Plots of the absolute values of the residuals or of the squared residuals against the predictor variable X or against the fitted values $\hat{Y}$ are also useful for diagnosing nonconstancy of the error variance since the signs of the residuals are not meaningful for examining the constancy of the error variance.

3. Example Blood Pressure - Age Example

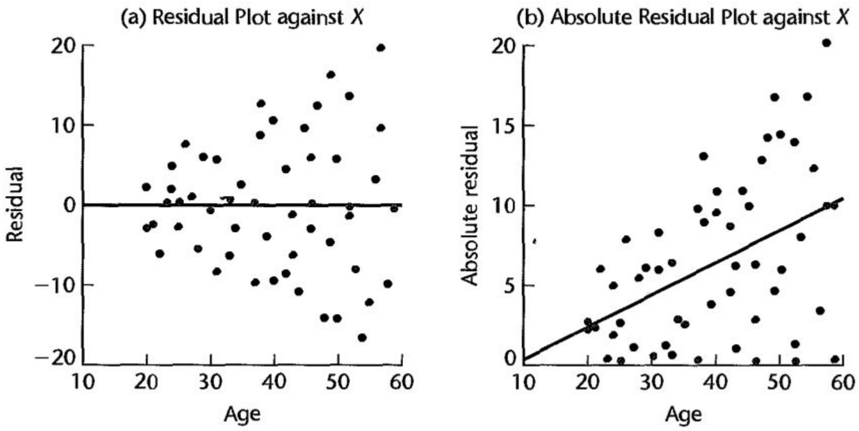
   (a) A study of the relation between diastolic blood pressure of healthy, adult women (Y) and their age (X).

   (b) (Figure 3.5) The residual plot suggests that the older the woman is, the more spread out the residuals are.

   (c) Since the relation between blood pressure and age is positive, this suggests that the error variance is larger for older women than for younger ones.

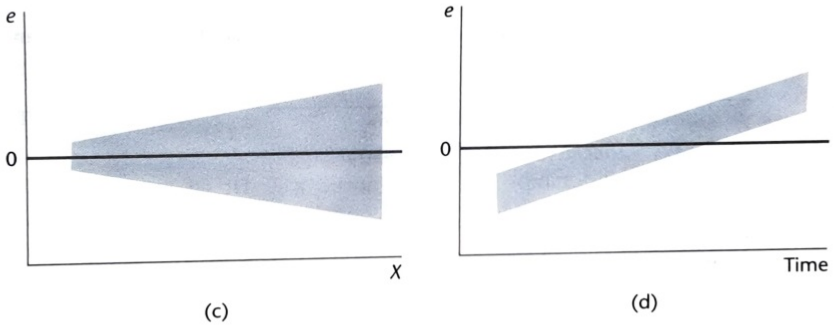
   (d) (Figure 3.5b) a plot of the absolute residuals against age for the blood pressure shows more clearly that the residuals tend to be larger in absolute magnitude for older-aged women.

FIGURE 3.5 **Residual Plots Illustrating Nonconstant Error Variance.**



4. (Figure 3.4c) a residual plots when the error variance increases with X. One can also encounter error variances decreasing with increasing levels of the predictor variable and occasionally varying in some more complex fashion.

FIGURE 3.4 **Prototype Residual Plots.**



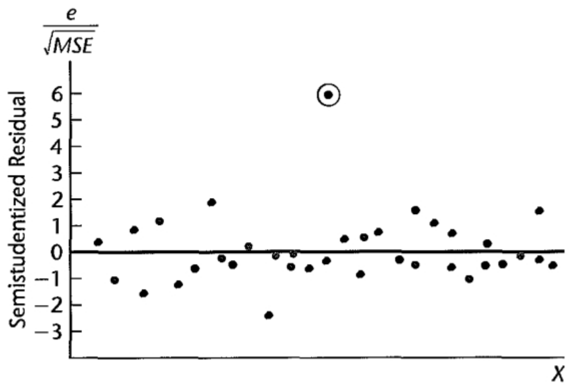
Presence of Outliers

1. Residual outliers (extreme observations) can be identified from residual plots against X or Y, as well as from box plots, stem-and-leaf plots, and dot plots of the residuals.

2. A rough rule of thumb when the number of cases is large is to consider semistudentized residuals with absolute value of four or more to be outliers. (details in Chapter 10).

3. (Figure 3.6) The residual plot in presents semistudentized residuals and contains one outlier, which is circled.

FIGURE 3.6 **Residual Plot with Outlier.**



4. How to deal with outliers:

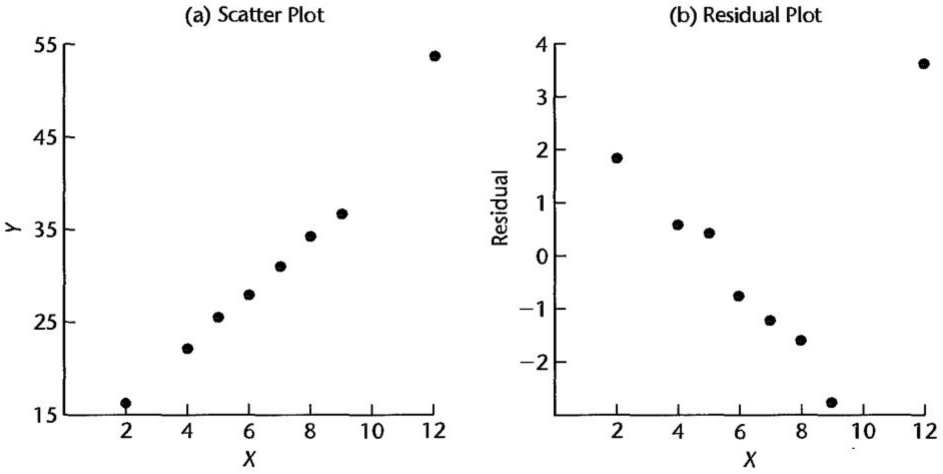
   (a) A safe rule frequently suggested is to discard an outlier only if there is direct evidence that it represents an error in recording, a miscalculation, a malfunctioning of equipment, or a similar type of circumstance.

(b) Under the least squares method, a fitted line may be pulled disproportionately toward an outlying observation because the sum of the squared deviations is minimized.

(c) This could cause a misleading fit if indeed the outlying observation resulted from a mistake or other extraneous cause.

5. (Figure 3.7) The fitted regression is distorted by the outlier that the residual plot suggest a lack of fit of the linear regression model.

FIGURE 3.7 Distorting Effect on Residuals Caused by an Outlier When Remaining Data Follow Linear Regression.



Nonindependence of Error Terms

1. **A sequence plot of the residuals**: the purpose of plotting the residuals against time or in some other type of sequence is to see if there is any correlation between error terms that are near each other in the sequence.

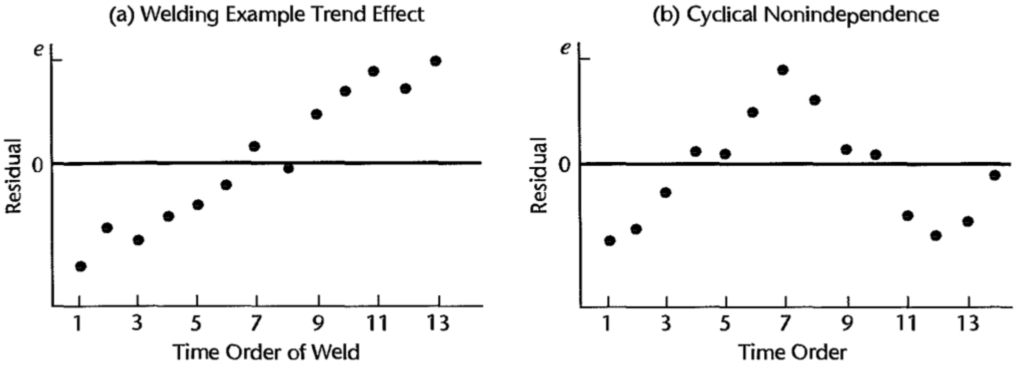
2. Example *Linear Time-related Trend Effect*

(a) (Figure 3.8a) contains a time sequence plot of the residuals in an experiment to study the relation between the diameter of a weld (X) and the shear strength of the weld (Y).

(b) An evident correlation between the error terms stands out. Negative residuals are associated mainly with the early trials, and positive residuals with the later trials.

(c) It is sometimes useful to view the problem of nonindependence of the error terms as one in which an important variable (in this case, time) has been omitted from the model.

FIGURE 3.8 **Residual Time Sequence Plots Illustrating Nonindependence of Error Terms.**



3. Example **Cyclical Nonindependent**

   (a) (Figure 3.8b) the adjacent error terms are also related, but the resulting pattern is a cyclical one with no trend effect present.

   (b) When the error terms are independent, we expect the residuals in a sequence plot to fluctuate in a more or less random pattern around the base line 0.

Nonnormality of Error Terms

1. Small departures from normality do not create any serious problems.

2. The normality of the error terms can be studied informally by examining the residuals in a variety of graphic ways.

3. **Distribution Plots** A box plot, histogram, dot plot, or stem-and-leaf plot of the residuals can be helpful for detecting gross departures from normality. Note that the number of cases in the regression study must be reasonably large for any of these plots to convey reliable information about the shape of the distribution of the error terms.

4. **Comparison of Frequencies** Another possibility when the number of cases is reasonably large is to compare actual frequencies of the residuals against

expected frequencies under normality. For example, one can determine whether, say, about 68 percent of the residuals e_i fall between $\pm\sqrt{MSE}$ or about 90 percent fall between $\pm 1.645\sqrt{MSE}$.

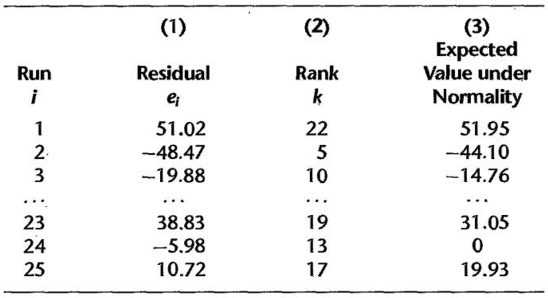
5. **Normal Probability Plot of the residuals** Each residual is plotted against its expected value under normality. A plot that is nearly linear suggests agreement with normality, whereas a plot that departs substantially from linearity suggests that the error distribution is not normal.

Question .... (p111)

In Toluca Company example, find the expected values of the ordered residuals under normality.

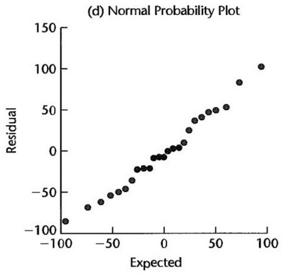
sol:

TABLE 3.2 Residuals and Expected Values under Normality—Toluca Company Example.



Run i	(1) Residual e_i	(2) Rank k	(3) Expected Value under Normality
1	51.02	22	51.95
2	−48.47	5	−44.10
3	−19.88	10	−14.76
···	···	···	···
23	38.83	19	31.05
24	−5.98	13	0
25	10.72	17	19.93

FIGURE 3.2 MINITAB and SYGRAPH Diagnostic Residual Plots —Toluca Company Example.



5. Three normal probability plots when the distribution of the error terms departs substantially from normality.

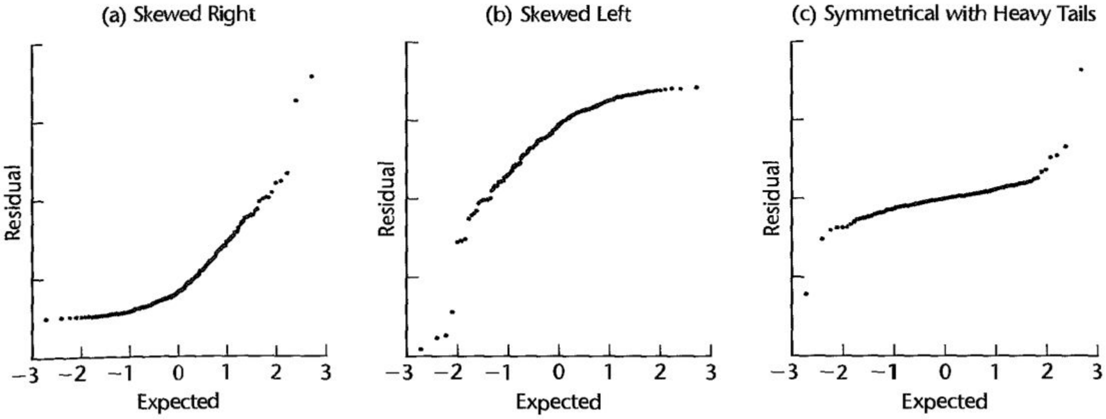
   (a) (Figure 3.9a) shows a normal probability plot when the error term distribution is highly skewed to the right. Note the concave-upward shape of the plot.

   (b) (Figure 3.9b) shows a normal probability plot when the error term distribution is highly skewed to the left. Here, the pattern is concave downward.

   (c) (Figure 3.9c) shows a normal probability plot when the distribution of the error tenus is symmetrical but has heavy tails; in other words, the distribution has higher probabilities in the tails than a normal distribution.

```
https://www.ucd.ie/ecomodel/Resources/QQplots_WebVersion.html
```

FIGURE 3.9 **Normal Probability Plots when Error Term Distribution Is Not Normal.**



6. **Difficulties in Assessing Normality**

   (a) The analysis for model departures with respect to normality is, in many respects, more difficult than that for other types of departures.

   (b) It is usually a good strategy to investigate these other types of departures first, before concerning oneself with the normality of the error terms.

Omission of Important Predictor Variables

1. Residuals should also be plotted against variables omitted from the model that might have important effects on the response.

2. Example One would like to study the relation between output (Y) and age (X) of worker in an assembling operation for a sample of employees. In this study, the machines produced by two companies (A and B) are used in the assembling operation.

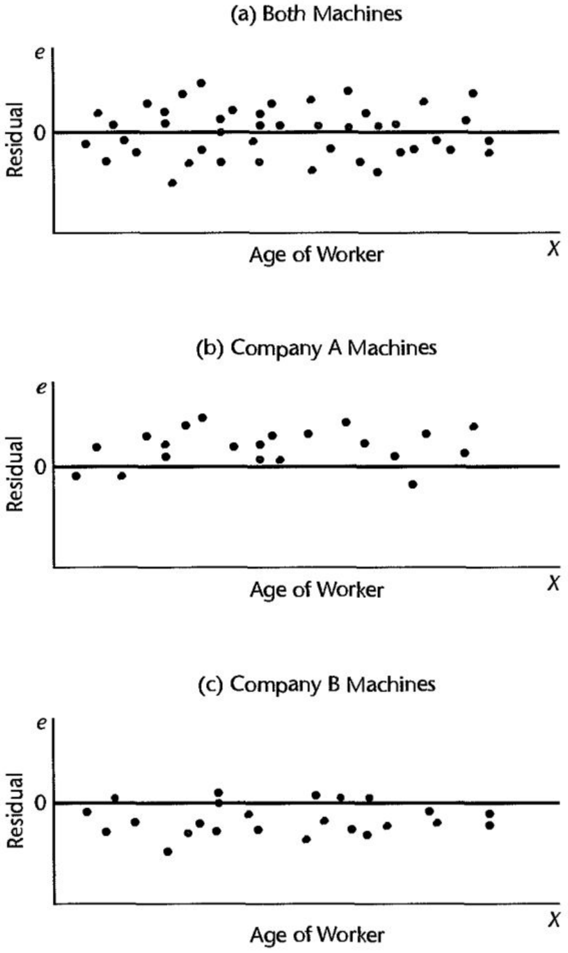


FIGURE 3.10 Residual Plots for Possible Omission of Important Predictor Variable—Productivity Example.

(a) (Figure 3.10a) no ground for suspecting the appropriateness of the linearity of the regression function or the constancy of the error variance.

(b) (Figure 3.10b, 3.l0c) The residuals for Company A machines tend to be positive: while those for Company B machines tend to be negative.

(c) Type of machine appears to have a definite effect on productivity, and output predictions may turn out to be far superior when this variable is added to the model.

Some Final Comments[1]

1. Several types of departures may occur together.
2. Although graphic analysis of residuals is only an informal method of analysis, in many cases it suffices for examining the aptness of a model.
3. The basic approach to residual analysis explained here applies not only to simple linear regression but also to more complex regression and other types of statistical models.
4. Model misspecification due to either nonlinearity or the omission of important predictor variables tends to be serious, leading to biased estimates of the regression parameters and error variance.
5. Nonconstancy of error variance tends to be less serious, leading to less efficient estimates and invalid error variance estimates.
6. The presence of outliers can be serious for smaller data sets when their influence is large.
7. The nonindependence of error terms results in estimators that are unbiased but whose variances are seriously biased.

3.4 Overview of Tests Involving Residuals

1. Graphic analysis of residuals is inherently subjective.
2. Most statistical tests require independent observations. The residuals are dependent. The dependencies become quite small for large samples, so that one can usually then ignore them.

Tests for Randomness

1. A runs test is frequently used to test for lack of randomness in the residuals arranged in time order.

[1] Some will be discussed in other Chapters.

2. Durbin-Watson test: designed for lack of randomness in least squares residuals. (Chapter 12).

Tests for Constancy of Variance

1. When a residual plot gives the impression that the variance may be increasing or decreasing in a systematic manner related to X or $E(Y)$, a simple test is based on the rank correlation between the absolute values of the residuals and the corresponding values of the predictor variable.

2. Tests for constancy of the error variance: the Brown-Forsythe test and the Breusch-Pagan test (Section 3.6.)

Tests for Outliers

1. A simple test for identifying an outlier observation: detail in (Chapter 10).

2. Many other tests to aid in evaluating outliers have been developed (Reference 3.1.)

Tests for Normality

1. Goodness of fit tests (the chi-square test, the Kolmogorov-Smirnov test and its modification, the Lilliefors test) can be employed for testing the normality of the error terms by analyzing the residuals.

2. A simple test based on the normal probability plot of the residuals (Section 3.5.)

3.5 Correlation Test for Normality

1. A formal test for normality of the error terms can be conducted by calculating the coefficient of correlation between the residuals e_i and their expected values under normality.

2. A high value of the correlation coefficient is indicative of normality.

3. (Table B.6) (Looney and Gulledge) (Ref. 3.2) contains critical values (percentiles) for various sample sizes for the distribution of the coefficient of correlation between the ordered residuals and their expected values under normality when the error terms are normally distributed.

4. If the observed coefficient of correlation is at least as large as the tabled value, for a given a level, one can conclude that the error terms are reasonably normally distributed.

5. Example Toluca Company Example: the coefficient of correlation between the ordered residuals and their expected values under normality is 0.991. Controlling the a risk at 0.05, we find from Table B.6 that the critical value for $n = 25$ is 0.959. Since the observed coefficient exceeds this level, we have support for our earlier conclusion that the distribution of the error terms does not depart substantially from a normal distribution.

☺ Normality test: `https://en.wikipedia.org/wiki/Normality_test`.

3.6 Tests for Constancy of Error Variance

Brown-Forsythe Test

1. *Assumption*: the sample size needs to be large enough so that the dependencies among the residuals can be ignored.

2. The Brown-Forsythe test is based on the variability of the residuals. The larger the error variance, the larger the variability of the residuals will tend to be.

3. The Brown-Forsythe test then consists simply of the two-sample t test based on test statistic (A.67)

$$t^* = \frac{\bar{Y} - \bar{Z}}{s(\bar{Y} - \bar{Z})}$$

to determine whether the mean of the absolute deviations for one group differs significantly from the mean absolute deviation for the second group. Steps:

   (a) Divide the data set into two groups, according to the level of X, so that one group consists of cases where the X level is comparatively low and the other group consists of cases where the X level is comparatively high.

(b) If the error variance is either increasing or decreasing with X, the residuals in one group will tend to be more variable than those in the other group.

(c) Equivalently, the absolute deviations of the residuals around their group mean will tend to be larger for one group than for the other group.

(d) In order to make the test more robust, we utilize the absolute deviations of the residuals around the median for the group (Ref. 3.5).

4. Although the distribution of the absolute deviations of the residuals is usually not normal, it has been shown that the t^* test statistic still follows approximately the t distribution when the variance of the error terms is constant and the sample sizes of the two groups are not extremely small.

5. Notations: the ith residual for group 1 (2) by e_{i1} (e_{i2}), the sample sizes of the two groups by n_1 and n_2, the medians of the residuals in the two groups by $\tilde{e}_1$ and $\tilde{e}_2$.

6. The Brown-Forsythe test uses the absolute deviations of the residuals around their group median, to be denoted by d_{i1} and d_{i2}:

$$d_{i1} = |e_{i1} - \tilde{e}_1| \quad \text{and} \quad d_{i2} = |e_{i2} - \tilde{e}_2|$$

7. The two-samplet test statistic (called the Brown-Forsythe test statistics t^*_{BF}) becomes:

$$t^*_{BF} = \frac{\bar{d}_1 - \bar{d}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where $\bar{d}_1$ and $\bar{d}_2$ are the sample means of the d_{i1} and d_{i2} respectively, and the pooled variance s^2 becomes:

$$s^2 = \frac{\sum(d_{i1} - \bar{d}_1)^2 + \sum(d_{i2} - \bar{d}_2)^2}{n - 2}$$

8. If the error terms have constant variance and n_1 and n_2 are not extremely small, t^*_{BF} follows approximately the t distribution with $n - 2$ degrees of freedom.

9. Large absolute values of t^*_{BF} indicate that the error terms do not have constant variance.

Question .... (p117)

Use the Brown-Forsythe test for the Toluca Company example to determine whether or not the error term variance varies with the level of X. (Note that since the X levels are spread fairly uniformly, you can divide the 25 cases into two groups with approximately equal X ranges. The first group consists of the 13 runs with lot sizes from 20 to 70. The second group consists of the 12 runs with lot sizes from 80 to 120. ($\alpha = 0.05, t_{0.975,23} = 2.069$)

sol:

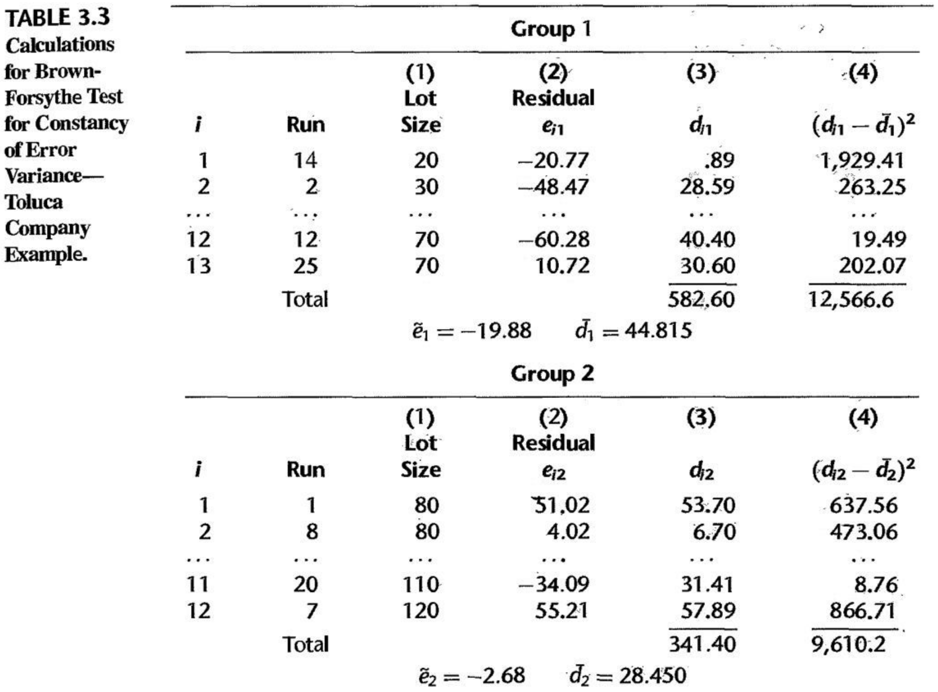


TABLE 3.3 Calculations for Brown-Forsythe Test for Constancy of Error Variance—Toluca Company Example.

Group 1

		(1)	(2)	(3)	(4)
		Lot	Residual		
i	Run	Size	e_{i1}	d_{i1}	$(d_{i1} - \bar{d}_1)^2$
1	14	20	−20.77	.89	1,929.41
2	2	30	−48.47	28.59	263.25
...	...	...	...	...	...
12	12	70	−60.28	40.40	19.49
13	25	70	10.72	30.60	202.07
	Total			582.60	12,566.6

$\tilde{e}_1 = -19.88 \qquad \bar{d}_1 = 44.815$

Group 2

		(1)	(2)	(3)	(4)
		Lot	Residual		
i	Run	Size	e_{i2}	d_{i2}	$(d_{i2} - \bar{d}_2)^2$
1	1	80	51.02	53.70	637.56
2	8	80	4.02	6.70	473.06
...	...	...	...	...	...
11	20	110	−34.09	31.41	8.76
12	7	120	55.21	57.89	866.71
	Total			341.40	9,610.2

$\tilde{e}_2 = -2.68 \qquad \bar{d}_2 = 28.450$

Breusch-Pagan Test*

3.7 F Test for Lack of Fit

Assumptions

1. F test for lack of fit is a formal test for determining whether a specific type of regression function adequately fits the data.

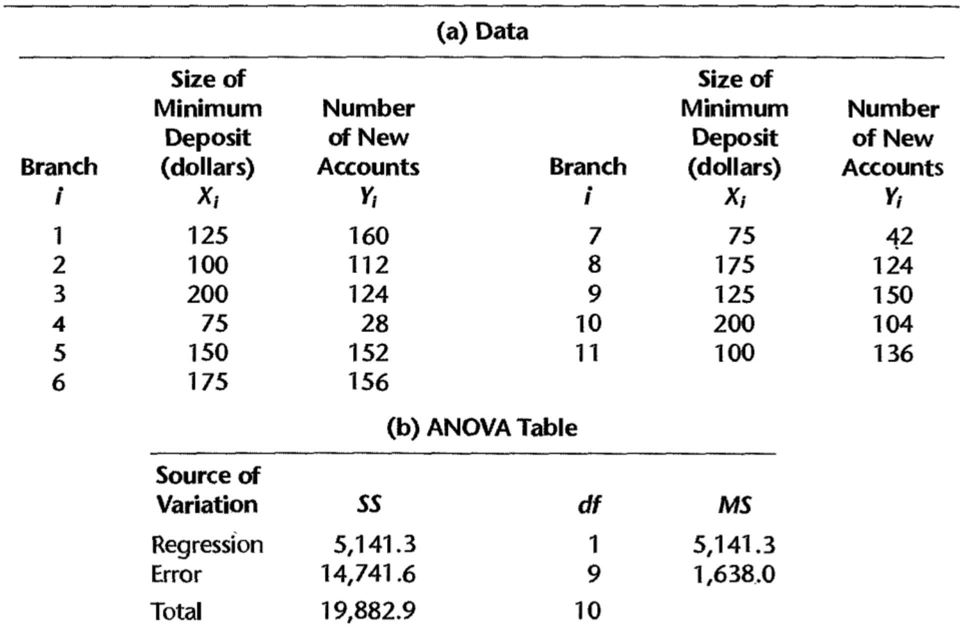
2. The lack of fit test assumes that the observations Y for given X are (1) independent and (2) normally distributed, and that (3) the distributions of Y have the same variance σ^2.

3. **Replications, Replicates**: the lack of fit test requires repeat observations at one or more X levels. Repeat trials for the same level of the predictor variable, of the type described, are called replications. The resulting observations are called replicates.

4. Example Bank Example

   (a) In an experiment involving 12 similar but scattered suburban branch offices of a commercial bank, holders of checking accounts at the offices were offered gifts for setting up money market accounts. Minimum initial deposits in the new money market account were specified to qualify for the gift. The value of the gift was directly proportional to the specified minimum deposit. Various levels of minimum deposit and related gift values were used in the experiment in order to ascertain the relation between the specified minimum deposit and gift value, on the one hand, and number of accounts opened at the office, on the other. Altogether, six levels of minimum deposit and proportional gift value were used, with two of the branch offices assigned at random to each level. One branch office had a fire during the period and was dropped from the study. Table 3.4a contains the results, where X is the amount of minimum deposit and Y is the number of new money market accounts that were opened and qualified for the gift during the test period.

TABLE 3.4 Data and Analysis of Variance Table—Bank Example.



(a) Data

Branch i	Size of Minimum Deposit (dollars) X_i	Number of New Accounts Y_i	Branch i	Size of Minimum Deposit (dollars) X_i	Number of New Accounts Y_i
1	125	160	7	75	42
2	100	112	8	175	124
3	200	124	9	125	150
4	75	28	10	200	104
5	150	152	11	100	136
6	175	156			

(b) ANOVA Table

Source of Variation	*SS*	*df*	*MS*
Regression	5,141.3	1	5,141.3
Error	14,741.6	9	1,638.0
Total	19,882.9	10	

(b) A linear regression function was fitted:

$$\hat{Y} = 50.72251 + 0.48670X$$

(Table 3.4b): The analysis of variance table.

(c) (Figure 3.11) A scatter plot, together with the fitted regression line, indicates that a linear regression function is inappropriate. We use the general linear test approach to do a formal test.

FIGURE 3.11 Scatter Plot and Fitted Regression Line—Bank Example.

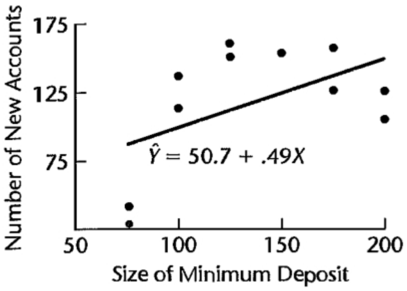
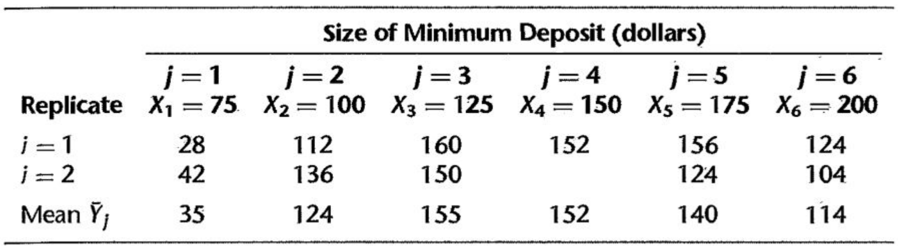


TABLE 3.5 Data Arranged by Replicate Number and Minimum Deposit—Bank Example.



	Size of Minimum Deposit (dollars)					
Replicate	$j=1$ $X_1=75$	$j=2$ $X_2=100$	$j=3$ $X_3=125$	$j=4$ $X_4=150$	$j=5$ $X_5=175$	$j=6$ $X_6=200$
$i=1$	28	112	160	152	156	124
$i=2$	42	136	150		124	104
Mean $\bar{Y}_j$	35	124	155	152	140	114

Notation

1. (Table 3.5) presents the same data but in an arrangement that recognizes the replicates. We shall denote the different X levels in the study, whether or not replicated observations are present, as $X_1, \cdots, X_c$.

2. There are six minimum deposit size levels in the study ($c = 6$), for five of which there are two observations and for one there is a single observation. We shall let $X_1 = 75$ (the smallest minimum deposit level), $X_2 = 100$, $\cdots$, $X_6 = 200$.

3. Denote the number of replicates for the jth level of X as n_j; for our example, $n_1 = n_2 = n_3 = n_5 = n_6 = 2$ and $n_4 = 1$. Thus, the total number of observations n is given by: $n = \sum_{j=1}^{c} n_j$.

4. Denote the observed value of the response variable for the ith replicate for the jth level of X by Y_{ij}, where $i = 1, \cdots, n_j$, $j = 1, \cdots, c$.

5. (Table 3.5), $Y_{11} = 28, Y_{21} = 42, Y_{12} = 112$, and so on. Denote the mean of the Y observations at the level $X = X_j$ by $\bar{Y}_j$. Thus, $\bar{Y}_1 = (28 + 42)/2 = 35$ and $\bar{Y}_4 = 152/1 = 152$.

Full model

1. The full model used for the lack of fit test makes the same assumptions as the simple linear regression model (2.1) except for assuming a linear regression relation, the subject of the test.

$$\underline{Y_{ij} = \mu_j + \epsilon_{ij}},$$

where μ_j are parameters $j = 1, \cdots, c$, ϵ_{ij} are independent $\underline{N(0, \sigma^2)}$.

2. Since the error terms have expectation zero, it follows that:

$$E(Y_{ij}) = \underline{\mu_j}.$$

Thus, the parameter μ_j ($j = 1, \cdots, c$) is the mean response when $X = X_j$.

3. The full model states that each response Y is made up of two components: the mean response when $X = X_j$ and a random error term.

4. The difference between the two models is that in the full model (3.13) there are no restrictions on the means μ_j, whereas in the regression model (2.1) the mean responses are linearly related to X (i.e., $E(Y) = \beta_0 + \beta_1 X$).

5. The least squares or maximum likelihood estimators for the parameters μ_j: $\hat{\mu}_j = \bar{Y}_j$.

6. The estimated expected value for observation Y_{ij} is $\bar{Y}_j$, and the error sum of squares (also called the pure error sum of squares, SSPE) for the full model:

$$SSE(F) = \sum_j \sum_i (Y_{ij} - \bar{Y}_j)^2 = SSPE$$

7. Note that $SSPE$ is made up of the sums of squared deviations at each X level. At level $X = X_j$, this sum of squared deviations is:

$$\sum_i (Y_{ij} - \bar{Y}_j)^2$$

These sums of squares are then added over all of the X levels ($j = 1, \cdots, c$).

8. Example For the bank example, we have:

$$SSPE = (28 - 35)^2 + (42 - 35)^2 + (112 - 124)^2 + \cdots + (104 - 114)^2 = 1,148$$

Note that any X level with no replications makes no contribution to $SSPE$ because $\bar{Y}_j = Y_{1j}$ for $j = 4$.

9. The degrees of freedom associated with $SSPE$ can be obtained by recognizing that the sum of squared deviations (3.17) at a given level of X is like an ordinary total sum of squares based on n observations, which has $n - 1$ degrees of freedom associated with it. Here, there are n_j observations when $X = X_j$; hence the degrees of freedom are $n_j - 1$.

10. Just as $SSPE$ is the sum of the sums of squares (3.17), so the number of degrees of freedom associated with $SSPE$ is the sum of the component degrees of freedom:

$$df_F = \sum_j (n_j - 1) = \sum_j n_j - c = n - c$$

Reduced Model

1. For testing the appropriateness of a linear regression relation, the alternatives are:

$$\begin{aligned} H_0 &: \underline{\quad E(Y) = \beta_0 + \beta_1 X \quad} \\ H_a &: \underline{\quad E(Y) \neq \beta_0 + \beta_1 X \quad} \end{aligned}$$

   Thus, H_0 postulates that μ_j in the full model (3.13) is linearly related to X_j

$$\underline{\quad \mu_j = \beta_0 + \beta_1 X_j \quad}$$

   The reduced model under H_0 therefore is:

$$\underline{\quad Y_{ij} = \beta_0 + \beta_1 X_j + \epsilon_{ij} \quad}$$

2. Note that the reduced model is the ordinary simple linear regression model (2.1), with the subscripts modified to recognize the existence of replications.

3. We know that the estimated expected value for observation Y_{ij} with regression model (2.1) is the fitted value $\hat{Y}_{ij}$

$$\underline{\quad \hat{Y}_{ij} = b_0 + b_1 X_j \quad}$$

   Hence, the error sum of squares for the reduced model is the usual error sum of squares SSE:

$$SSE(R) = \underline{\quad \sum\sum(Y_{ij} - (b_0 + b_1 X_j))^2 = \sum\sum(Y_{ij} - \hat{Y}_{ij})^2 = SSE \quad}$$

   We also know that the degrees of freedom associated with $SSE(R)$ are: $\underline{\quad df_R = n - 2 \quad}$.

4. Example For the bank example, we have from Table 3.4b: $SSE(R) = SSE = 14741.6, \ df_R = 9$

Test Statistic

1. The general linear test statistic (2.70):

$$F^* = \underline{\quad \frac{SSE(R) - SSE(F)}{df_r - df_F} \div \frac{SSE(F)}{df_F} \quad}$$

   here becomes:

$$F^* = \underline{\quad \frac{SSE - SSPE}{(n-2) - (n-c)} \div \frac{SSPE}{n-c} \quad}$$

2. The difference between the two error sums of squares is called the lack of fit sum of squares ($SSLF$):

$$SSLF = SSE - SSPE$$

3. We can then express the test statistic as follows:

$$F^* = \frac{SSLF}{c-2} \div \frac{SSPE}{n-c} = \frac{MSLF}{MSPE}$$

where $MSLF$ denotes the lack of fit mean square and $MSPE$ denotes the pure error mean square.

4. We know that large values of F^* lead to conclusion H_a in the general linear test. Decision rule (2.71) here becomes:

$$\text{If } F^* \leq F_{(1-\alpha;c-2,n-c)}, \text{ conclude } H_0$$

$$\text{If } F^* > F_{(1-\alpha;c-2,n-c)}, \text{ conclude } H_a$$

5. Example For the bank example, the test statistic:

$$\begin{aligned}
SSPE &= 1148.0, \quad n-c = 11-6 = 5 \\
SSE &= 14741.6, \\
SSLF &= 14741.6 - 1,148.0 = 13,593.6, \quad c-2 = 6-2 = 4 \\
F^* &= \frac{13,593.6}{4} \div \frac{1148.0}{5} = \frac{3,398.4}{229.6} = 14.80
\end{aligned}$$

If the level of significance is to be $\alpha = 0.01$, we require $F_{(0.99;4,5)} = 11.4$. Since $F^* = 14.80 > 11.4$, we conclude H_a, that the regression function is not linear. The P-value for the test is 0.006.

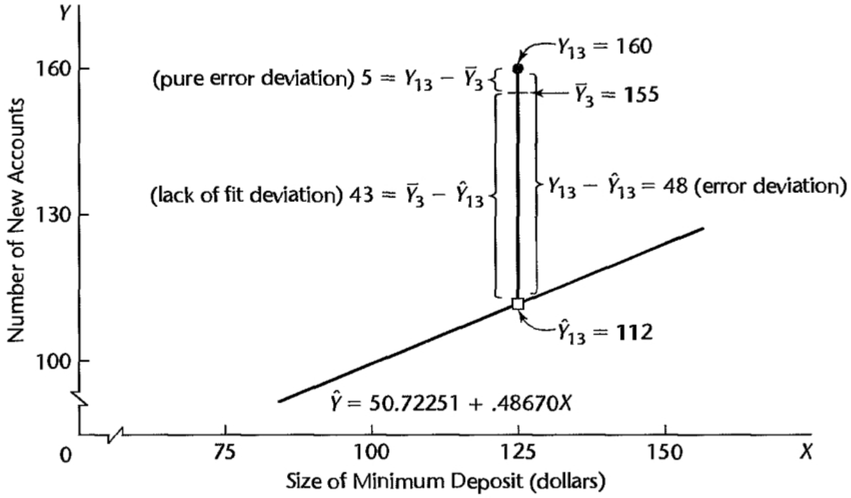
ANOVA Table

1. The error deviations in SSE are made up of a pure error component and a lack of fit component: $SSE = SSPE + SSLF$.

$$\begin{aligned}
Y_{ij} - \hat{Y}_{ij} &= (Y_{ij} - \bar{Y}_j) + (\bar{Y}_j - \hat{Y}_{ij}) \\
\text{Error deviation} &= \text{Pure error deviation} + \text{Lack of fit deviation}
\end{aligned}$$

2. $\boxed{\text{Example}}$ (Figure 3.12) illustrates this partitioning for the case $Y_{13} = 160, X_3 = 125$ in the bank example.

FIGURE 3.12 **Illustration of Decomposition of Error Deviation $Y_{ij} - \hat{Y}_{ij}$— Bank Example.**



3. When (3.28) is squared and summed over all observations, we obtain (3.27) since the cross-product sum equals zero:

$$\begin{aligned} \sum\sum(Y_{ij} - \hat{Y}_{ij})^2 &= \underline{\sum\sum(Y_{ij} - \bar{Y}_j)^2 + \sum\sum(\bar{Y}_j - \hat{Y}_{ij})^2} \\ SSE &= SSPE + SSLF \end{aligned}$$

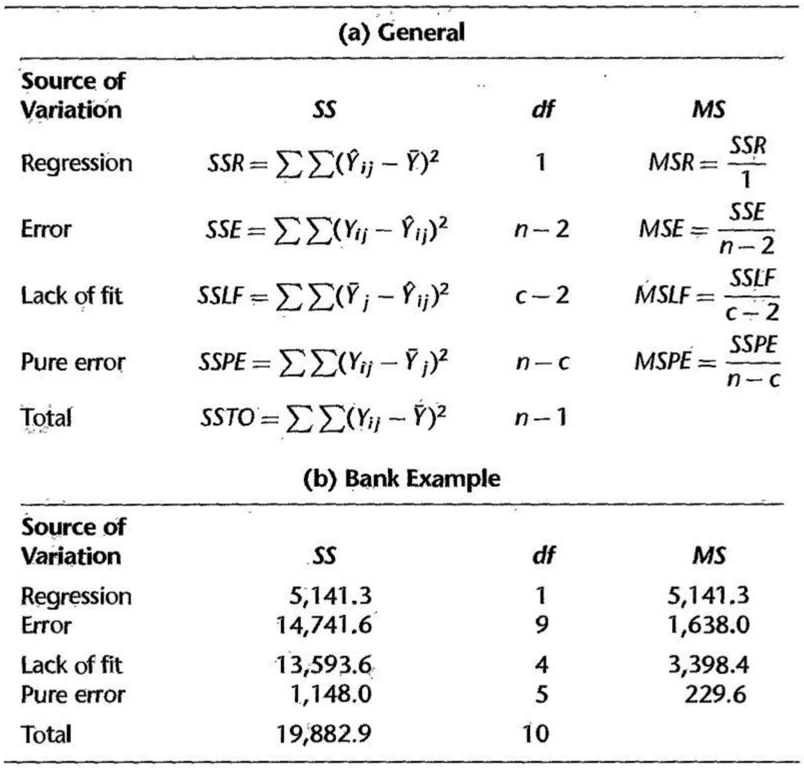
4. *Why $SSLF$ measures lack of fit?* If the linear regression function is appropriate, then the __means $\bar{Y}_j$__ will be near the __fitted values $\hat{Y}_j$__ calculated from the estimated linear regression function and $SSLF$ will be __small__.

5. On the other hand, if the linear regression function is not appropriate, the means $\bar{Y}_j$ will not be near the fitted values calculated from the estimated linear regression function and $SSLF$ will be large.

6. SSLF has $c-2$ degrees of freedom: there are __c__ means $\bar{Y}_j$ in the sum of squares, and __two__ degrees of freedom are lost in estimating the parameters β_0 and β_1, of the linear regression function to obtain the fitted values $\hat{Y}_j$ ·

7. (Table 3.6) contains the ANOVA decomposition for the bank example.

TABLE 3.6 General ANOVA Table for Testing Lack of Fit of Simple Linear Regression Function and ANOVA Table—Bank Example.



(a) General

Source of Variation	SS	df	MS
Regression	$SSR = \sum\sum(\hat{Y}_{ij} - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$
Error	$SSE = \sum\sum(Y_{ij} - \hat{Y}_{ij})^2$	$n-2$	$MSE = \frac{SSE}{n-2}$
Lack of fit	$SSLF = \sum\sum(\bar{Y}_j - \hat{Y}_{ij})^2$	$c-2$	$MSLF = \frac{SSLF}{c-2}$
Pure error	$SSPE = \sum\sum(Y_{ij} - \bar{Y}_j)^2$	$n-c$	$MSPE = \frac{SSPE}{n-c}$
Total	$SSTO = \sum\sum(Y_{ij} - \bar{Y})^2$	$n-1$	

(b) Bank Example

Source of Variation	SS	df	MS
Regression	5,141.3	1	5,141.3
Error	14,741.6	9	1,638.0
Lack of fit	13,593.6	4	3,398.4
Pure error	1,148.0	5	229.6
Total	19,882.9	10	

Comments

1. Not all levels of X need have repeat observations for the F test for lack of fit to be applicable. Repeat observations at only one or some levels of X are sufficient.

2. Suppose that prior to any analysis of the appropriateness of the model, we had fitted a linear regression model and wished to test whether or not $\beta_1 = 0$. For the bank example (Table 3Ab), test statistic (2.60) would be:

$$F^* = \frac{MSR}{MSE} = \frac{5141.3}{1638.0} = 3.14$$

For $\alpha = .10$, $F_{(0.90;1,9)} = 3.36$, and we would conclude H_0, that $\beta_1 = 0$ or that there is no linear association between minimum deposit size (and value of gift) and number of new accounts. A conclusion that there is no relation between these variables would be improper, however. Such an inference requires that regression model (2.1) be appropriate. Here, there is a definite relationship, but the regression function is not linear. This illustrates the importance of *always examining the appropriateness of a model before any inferences are drawn.*

3. The alternative H_a in (3.19) includes all regression functions other than a __linear__ one. For instance, it includes a quadratic regression function or a logarithmic one. If H_a is concluded, a study of __residuals__ can be helpful in identifying an appropriate function.

4. When no replications are present in a data set, an approximate test for lack of fit can be conducted if there are some cases at adjacent X levels for which the mean responses are quite close to each other. Such adjacent cases are grouped together and treated as __pseudo replicates__, and the test for lack of fit is then carried out using these groupings of adjacent cases. (Reference 3.8.)

3.8 Overview of Remedial Measures

1. If the simple linear regression model (2.1) is not appropriate for a data set, there are two basic choices:
   (a) Abandon regression model (2.1) and develop and use a __more appropriate model__.
   (b) Employ some __transformation__ on the data so that regression model (2.1) is appropriate for the transformed data.

Nonlinearity of Regression Function

Section 3.9, Section 3.10. Chapter 7.

Nonconstancy of Error Variance

Section 3.9, Chapter 11.

Nonindependence of Error Terms

Chapter 12.

Nonnormality of Error Terms

Section 3.9.

Omission of Important Predictor Variables

Chapter 6.

Outlying Observations

Chapter 11.

3.9 Transformations

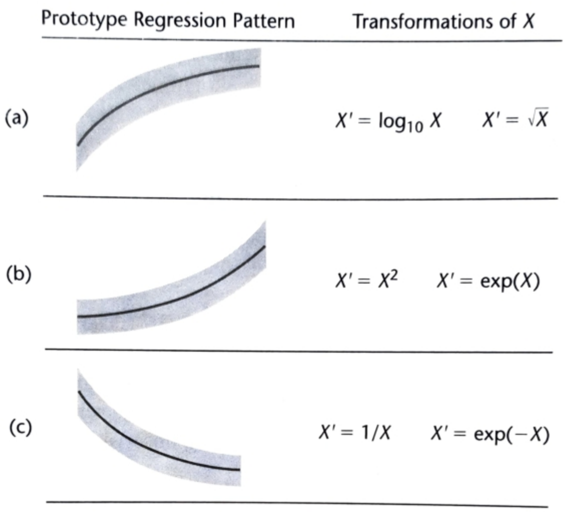
Simple transformations of either the response variable Y or the predictor variable X, or of both, are often sufficient to make the simple linear regression model appropriate for the transformed data.

Transformations for Nonlinear Relation Only

1. We first consider transformations for linearizing a nonlinear regression relation when the distribution of the error terms is reasonably close to a normal distribution and the error terms have approximately constant variance.

2. In this situation, transformations on X should be attempted. Transformation on Y may materially change the shape of the distribution of the - error terms from the normal distribution and may also lead to substantially differing error term variances.

FIGURE 3.13 **Prototype Nonlinear Regression Patterns with Constant Error Variance and Simple Transformations of *X*.**



3. (Figure 3.13) some prototype nonlinear regression relations with constant error variance and also presents some simple transformations on X that may be helpful to linearize the regression relationship without affecting the distributions of Y.

4. Example A battery of simulated sales

   (a) Data from an experiment on the effect of number of days of training received (X) on performance (Y) in a battery of simulated sales situations are presented in Table 3.7, columns 1 and 2, for the 10 participants in the study.

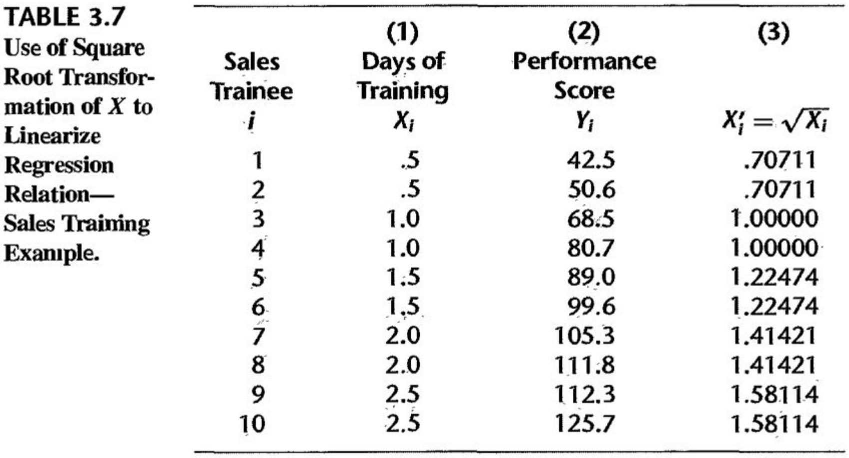
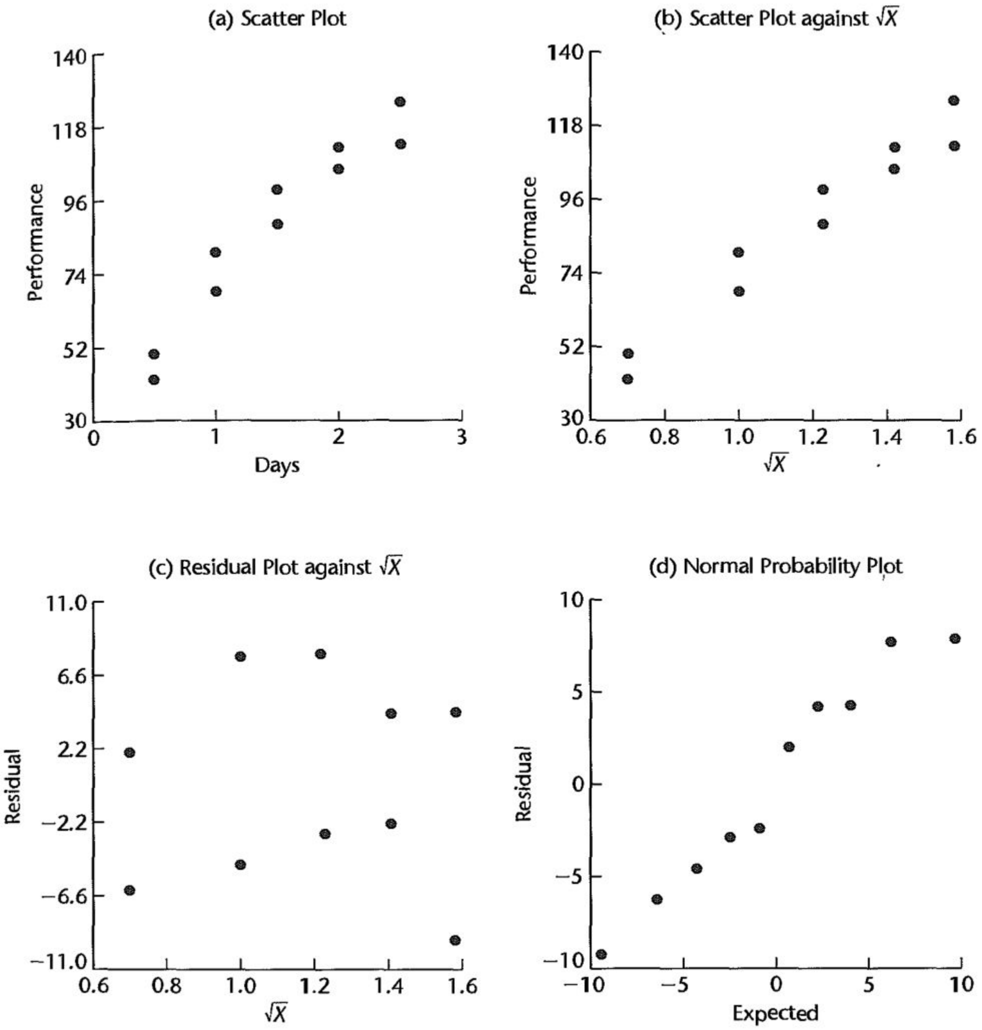


TABLE 3.7 Use of Square Root Transformation of X to Linearize Regression Relation—Sales Training Example.

Sales Trainee i	(1) Days of Training X_i	(2) Performance Score Y_i	(3) $X_i' = \sqrt{X_i}$
1	.5	42.5	.70711
2	.5	50.6	.70711
3	1.0	68.5	1.00000
4	1.0	80.7	1.00000
5	1.5	89.0	1.22474
6	1.5	99.6	1.22474
7	2.0	105.3	1.41421
8	2.0	111.8	1.41421
9	2.5	112.3	1.58114
10	2.5	125.7	1.58114

   (b) (Figure 3.14a) Clearly the regression relation appears to be curvilinear, so the simple linear regression model (2.1) does not seem to be appropriate. Since the variability at the different X levels appears to be fairly constant, we shall consider a transformation on X. Based on Figure 3.13a, consider initially the square root transformation $X' = \sqrt{X}$.

FIGURE 3.14 **Scatter Plots and Residual Plots—Sales Training Example.**



(c) (Figure 3.14b), the same data are plotted with the predictor variable transformed to $X' = \sqrt{X}$. Note that the scatter plot now shows a reasonably linear relation. The variability of the scatter at the different X levels is the same as before, since we did not make a transformation on Y.

(d) To examine further whether the simple linear regression model (2.1) is appropriate now, we fit it to the transformed X data:

$$\hat{Y} = -10.33 + 83.45X'.$$

(e) (Figure 3.14c) the plot of residuals against X' shows no evidence of lack of fit or of strongly unequal error variances.

(f) (Figure 3.14d) a normal probability plot of the residuals. No strong indications of substantial departures from normality. This conclusion is supported by the high correlation coefficient between the ordered residuals and their expected values under normality, 0.979.

(g) For $\alpha = 0.01$, Table B.6 shows that the critical value is 0.879, so the observed coefficient is substantially larger and supports the reasonableness of normal error terms. Thus, the simple linear regression model (2.1) appears to be appropriate here for the transformed data.

(h) The fitted regression function in the original units of X can easily be obtained, if desired:

$$\hat{Y} = -10.33 + 83.45\sqrt{X}$$

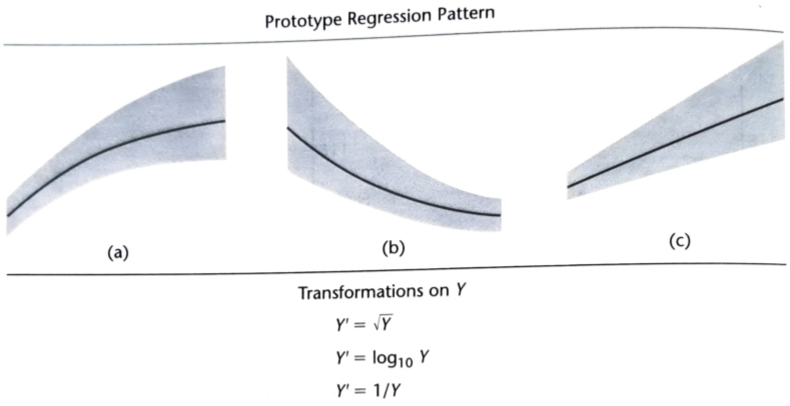
Transformations for Nonnormality and Unequal Error Variances

1. Unequal error variances and nonnormality of the error terms frequently appear together. To remedy these departures from the simple linear regression model (2.1), we need a transformation on Y, since the shapes and spreads of the distributions of Y need to be changed.

2. A simultaneous transformation on X may be needed to obtain or maintain a linear regression relation.

3. (Figure 3.15) Frequently, the nonnormality and unequal variances departures from regression model (2.1) take the form of increasing skewness and increasing variability of the distributions of the error terms as the mean response $E(Y)$ increases.

FIGURE 3.15 Prototype Regression Patterns with Unequal Error Variances and Simple Transformations of *Y*.



Note: A simultaneous transformation on X may also be helpful or necessary.

4. Scatter plots and residual plots should be prepared to determine the most effective transformations.

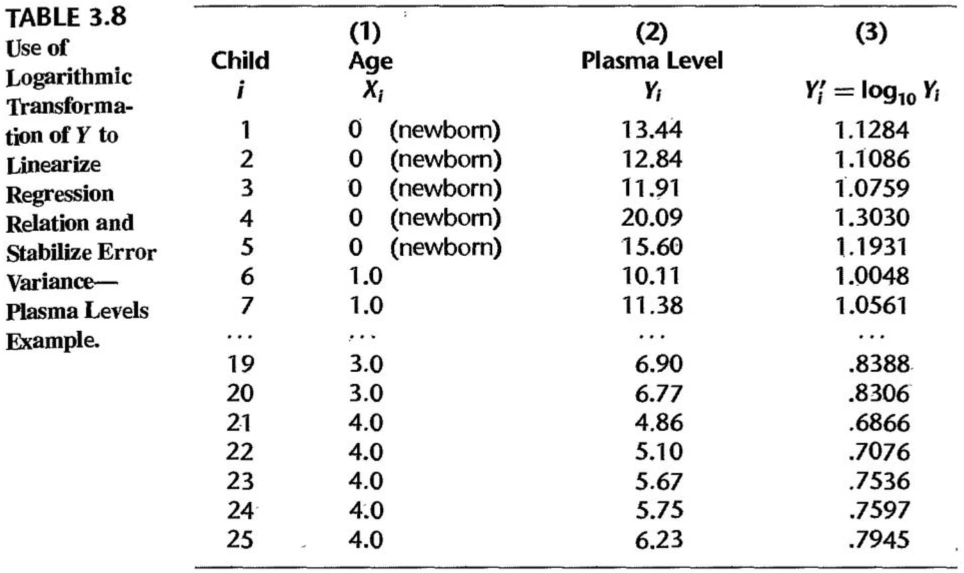


TABLE 3.8 Use of Logarithmic Transformation of Y to Linearize Regression Relation and Stabilize Error Variance—Plasma Levels Example.

Child i	(1) Age X_i	(2) Plasma Level Y_i	(3) $Y_i' = \log_{10} Y_i$
1	0 (newborn)	13.44	1.1284
2	0 (newborn)	12.84	1.1086
3	0 (newborn)	11.91	1.0759
4	0 (newborn)	20.09	1.3030
5	0 (newborn)	15.60	1.1931
6	1.0	10.11	1.0048
7	1.0	11.38	1.0561
···	···	···	···
19	3.0	6.90	.8388
20	3.0	6.77	.8306
21	4.0	4.86	.6866
22	4.0	5.10	.7076
23	4.0	5.67	.7536
24	4.0	5.75	.7597
25	4.0	6.23	.7945

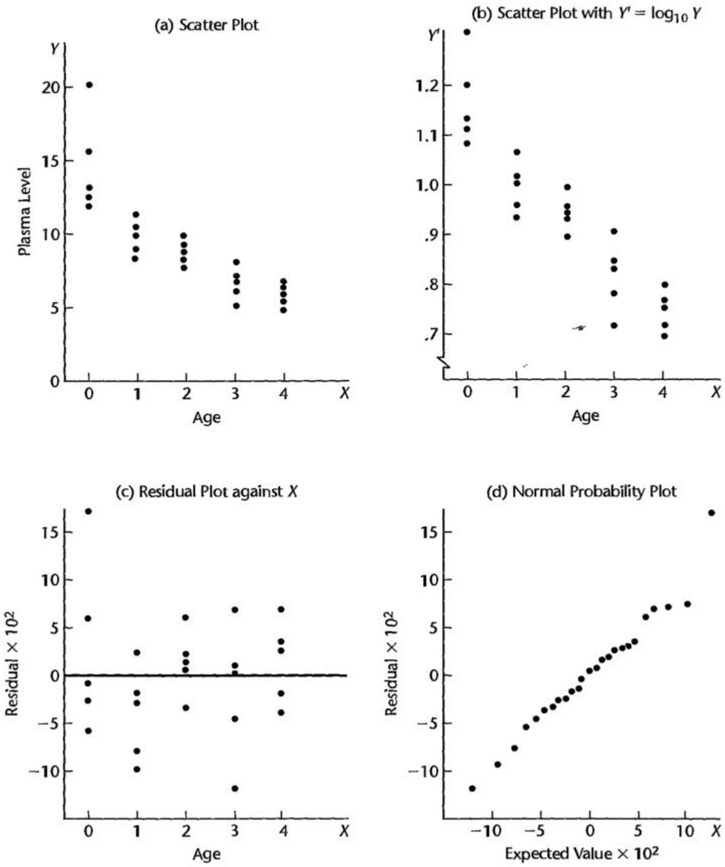
5. Example Plasma Level Example

   (a) (Table 3.8) Data on age (X) and plasma (血漿) level of a polyamine (多元胺) (Y) for a portion of the 25 healthy children in a study.

   (b) (Figure 3.16a) a scatter plot shows the distinct curvilinear regression relationship, as well as the greater variability for younger children than for older ones.

   (c) (Figure 3.16b) the scatter plot of the logarithmic transformation $Y' = \log_{10} Y$. The transformation not only has led to a reasonably linear regression relation, but the variability at the different levels of X also has become reasonably constant.

FIGURE 3.16 Scatter Plots and Residual Plots—Plasma Levels Example.



(a) Scatter Plot

(b) Scatter Plot with $Y' = \log_{10} Y$

(c) Residual Plot against X

(d) Normal Probability Plot

(d) To further examine the reasonableness of the transformation $Y' = \log_{10} Y$, we fitted the simple linear regression model (2.1) to the transformed Y data and obtained:

$$\hat{Y}' = 1.135 - 0.1023X$$

(e) (Figure 3.16c, d) the evidence supports the appropriateness of regression model (2.1) for the transformed Y data: (i) A plot of the residuals against X, and a normal probability plot of the residuals. (ii) The coefficient of correlation between the ordered residuals and their expected values under normality is 0.981. (iii) For $\alpha = 0.05$, Table B.6 indicates that the critical value is 0.959 so that the observed coefficient supports the assumption of normality of the error terms.

(f) NOTE: When Y is negative, the logarithmic transformation to shift the origin in Y and make all Y observations positive would be $Y' = \log_{10}(Y + k)$, where k is an appropriately chosen constant.

(g) NOTE: When unequal error variances are present but the regression relation is linear, a transformation on Y may not be sufficient while such a transformation may stabilize the error variance, it will also change the linear relationship to a curvilinear one. A transformation on X may therefore also be required.

Box-Cox Transformations

1. The Box-Cox procedure (Ref. 3.9) automatically identifies a transformation from the family of power transformations on Y. The family of power transformations is of the form:

$$Y' = Y^{\lambda}$$

where λ is a parameter to be determined from the data.

2. Note that this family encompasses the following simple transformations:

$$\begin{aligned}
\lambda &= 2 && Y' = Y^2 \\
\lambda &= 0.5 && Y' = \sqrt{Y} \\
\lambda &= 0 && Y' = \log_e(Y) \qquad \text{(by definition)} \\
\lambda &= -0.5 && Y' = \frac{1}{\sqrt{Y}} \\
\lambda &= -1.0 && Y = \frac{1}{Y}
\end{aligned}$$

☺ Power transform (Box-Cox transformation) - Wikipedia: `https://en.wikipedia.org/wiki/Power_transform`.

$$Y'(\lambda) = \begin{cases} \frac{Y^{\lambda}-1}{\lambda} & \lambda \neq 0 \\ \log_e Y & \lambda = 0 \end{cases}$$

3. The normal error regression model with the response variable a member of the family of power transformations becomes:

$$Y^{\lambda} = \beta_0 + \beta_1 X_i + \epsilon_i$$

Note that above regression model includes an additional parameter, λ, which needs to be estimated.

4. The Box-Cox procedure uses the method of maximum likelihood to estimate λ, as well as the other parameters β_0, β_1, and σ^2.

5. A simple procedure for obtaining $\hat{\lambda}$:

   (a) search in a range of potential λ values; for example, $\lambda = -2, \lambda = -1.75, \cdots, \lambda = 1.75, \lambda = 2$. For each λ value, the Y_i^λ observations are first standardized so that the magnitude of the error sum of squares does not depend on the value of λ.

   (b) Once the standardized observations have been obtained for a given λ value, they are regressed on the predictor variable X - and the error sum of square SSE is obtained.

   (c) It can be shown that the maximum likelihood estimate $\hat{\lambda}$ is that value of λ for which SSE is a minimum.

6. After a transformation has been tentatively selected, residual plots and other analyses described earlier need to be employed to ascertain that the simple linear regression model (2.1) is appropriate for the transformed data.

3.10 Exploration of Shape of Regression Function*

lowess Method*

Use of Smoothed Curves to Confirm Fitted Regression Function*

3.11 Case Example – Plutonium Measurement

1. *Background Description*: Some environmental cleanup work requires that nuclear materials, such as plutonium 238 (鈽-238), be located and completely removed from a restoration site. When plutonium has become mixed with other materials in very small amounts, detecting its presence can be a difficult task. Even very small amounts can be traced, however, because plutonium emits subatomic particles – alpha particles – that can be detected. Devices that are used to detect plutonium record the intensity of alpha particle strikes in counts per second (#/sec). The regression relationship between alpha counts per second (the response variable) and

plutonium activity (the explanatory variable) is then used to estimate the activity of plutonium in the material under study.

2. *Data Description*: (Table 3.10) In a study to establish the regression relationship for a particular measurement device, four plutonium standards were used. These standards are aluminum/plutonium rods containing a fixed, known level of plutonium activity. The levels of plutonium activity in the four standards were 0.0, 5.0, 10.0, and 20.0 picocuries (皮克居禮，衡量幅射的單位) per gram (pCi/g). Each standard was exposed to the detection device from 4 to 10 times, and the rate of alpha strikes, measured as counts per second, was observed for each replication.

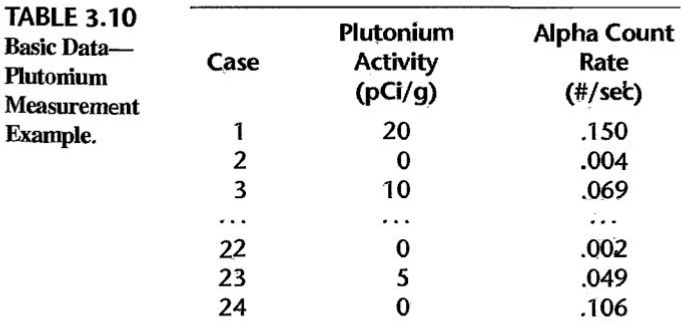


TABLE 3.10 Basic Data—Plutonium Measurement Example.

Case	Plutonium Activity (pCi/g)	Alpha Count Rate (#/sec)
1	20	.150
2	0	.004
3	10	.069
...	...	...
22	0	.002
23	5	.049
24	0	.106

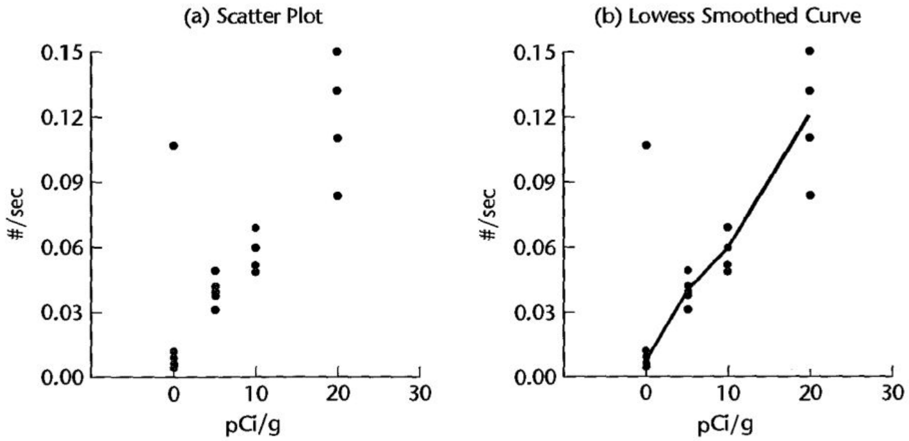
3. *Goal*: The task here is to estimate the regression relationship between alpha counts per second (Y) and plutonium activity (X).

4. *Assumption Before Doing Analysis*: the level of alpha counts increases with plutonium activity, but the exact nature of the relationship is generally unknown.

5. *Exploratory Data Analysis, EDA*:

   (a) *Scatter plot*: (Figure 3.20a) The strike rate tends to increase with the activity level of plutonium. Notice also that nonzero strike rates are recorded for the standard containing no plutonium. This results from background radiation and indicates that a regression model with an intercept term is required here.

FIGURE 3.20 SAS-JMP Scatter Plot and Lowess Smoothed Curve—Plutonium Measurement Example.

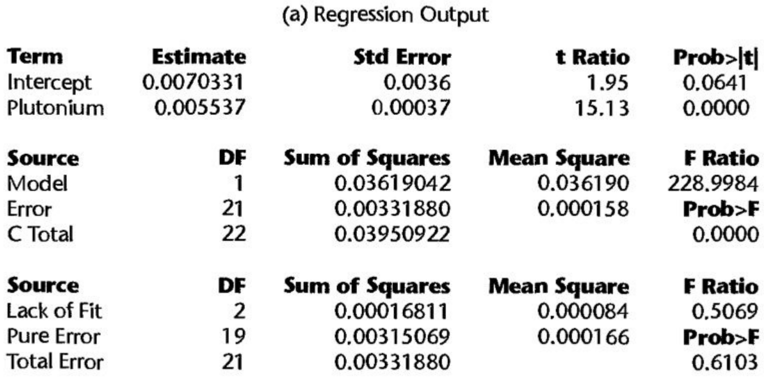


(b) *Investigate Relationship*: The regression relationship may be linear or slightly curvilinear in the range of the plutonium activity levels included in the study.

(c) *Outlier Detection*: An examination of laboratory records revealed that the experimental conditions were not properly maintained for the last case, and it was therefore decided that case 24 should be discarded. A linear regression function was fitted next, based on the remaining 23 cases.

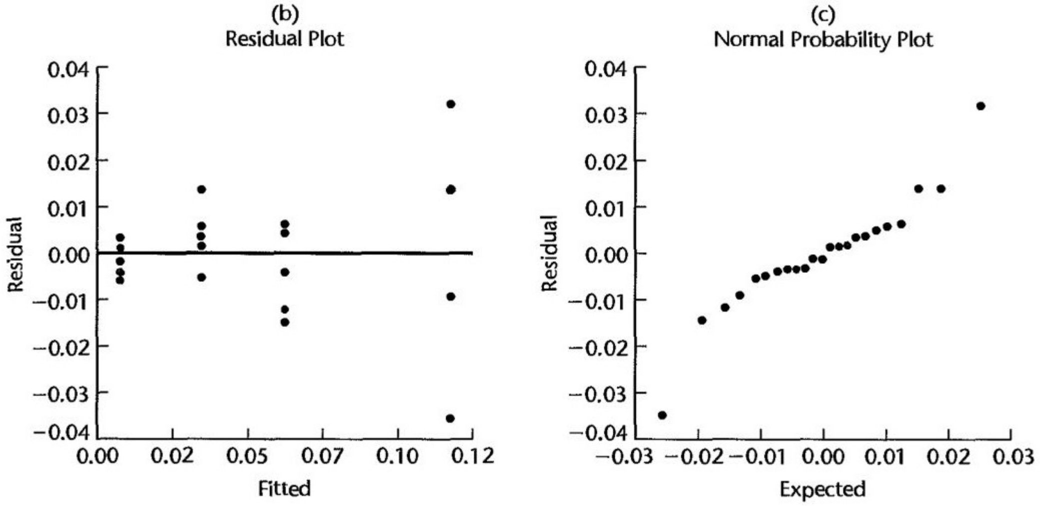
6. *Parameters Estimation and ANOVA*: (Figure 3.21a) the slope of the regression line is not zero ($F^* = 228.9984$, P-value$= 0.0000$) so that a regression relationship exists.

FIGURE 3.21 SAS-JMP Regression Output and Diagnostic Plots for Untransformed Data—Plutonium Measurement Example.



(a) Regression Output

```
Term          Estimate     Std Error   t Ratio   Prob>|t|
Intercept    0.0070331        0.0036      1.95     0.0641
Plutonium     0.005537       0.00037     15.13     0.0000

Source            DF   Sum of Squares   Mean Square    F Ratio
Model              1       0.03619042      0.036190   228.9984
Error             21       0.00331880      0.000158     Prob>F
C Total           22       0.03950922                   0.0000

Source            DF   Sum of Squares   Mean Square    F Ratio
Lack of Fit        2       0.00016811      0.000084     0.5069
Pure Error        19       0.00315069      0.000166     Prob>F
Total Error       21       0.00331880                   0.6103
```



7. *Model Diagnostic*:

   (a) *Residuals Plot*: (Figure 3.21b) the flared, megaphone shape of the residual plot shows that the error variance appears to be increasing with the level of plutonium activity.

   (b) *The Normal Probability plot*: (Figure 3.21c) suggests non-normality (heavy tails), but the nonlinearity of the plot is likely to be related (at least in part) to the unequal error variances.

   (c) *Breusch-Pagan Test*: the existence of nonconstant variance is confirmed by the Breusch-Pagan Test statistic:

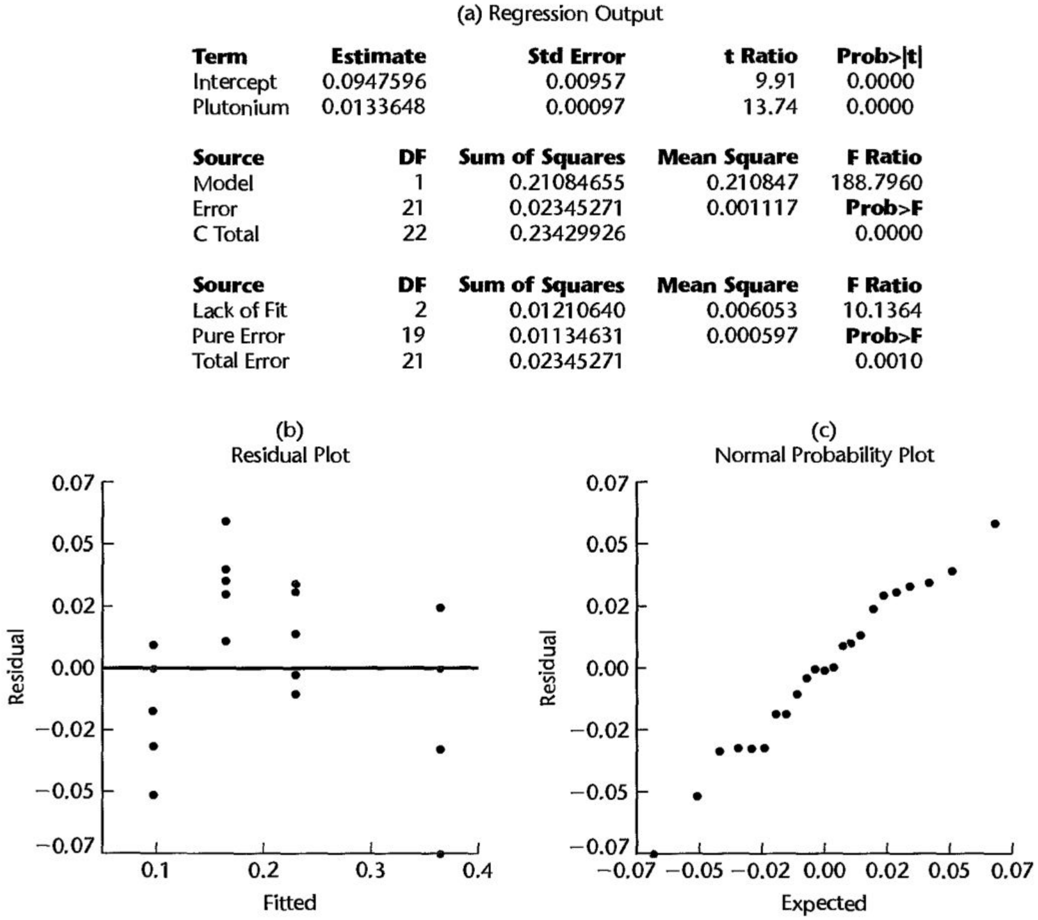
$$\chi^2_{BP} = 23.29 > \chi^2_{(0.95;1)} = 3.84$$

8. *Re-analysis After Data Transformation on Y*:

   (a) *Box-Cox transformation*: using the standardized variable, the maximum likelihood estimate of λ to be $\hat{\lambda} = 0.65$. The Box-Cox procedure supports the use of the square root transformation (i.e., use of $\lambda = 0.5$).

   (b) *Parameters Estimation and ANOVA*: (Figure 3.22a) The results of fitting a linear regression function when the response variable is $Y' = \sqrt{Y}$. The Lack of Fit Test statistic is $F^* = 10.1364$ with P-value $= 0.0010$.

FIGURE 3.22 SAS-JMP Regression Output and Diagnostic Plots for Transformed Response Variable—Plutonium Measurement Example.



(a) Regression Output

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.0947596	0.00957	9.91	0.0000
Plutonium	0.0133648	0.00097	13.74	0.0000

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	0.21084655	0.210847	188.7960
Error	21	0.02345271	0.001117	**Prob>F**
C Total	22	0.23429926		0.0000

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack of Fit	2	0.01210640	0.006053	10.1364
Pure Error	19	0.01134631	0.000597	**Prob>F**
Total Error	21	0.02345271		0.0010

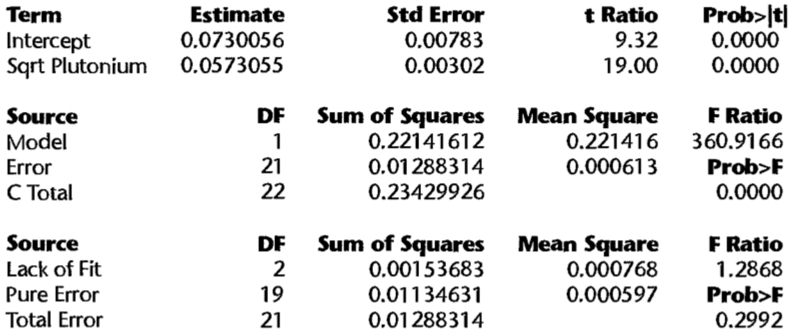
(c) *Diagnostic Plots*: (Figure 3.22b, c) the residual plot shows that the error variance appears to be more stable, it also suggests the Y' is nonlinearly related to X. The points in the normal probability plot fall roughly on a straight line.

9. *Re-analysis Again After Transformation on X*

(a) *Parameters Estimation and ANOVA*: (Figure 3.23a) The Lack of Fit Test ($F^* = 1.2868$ with P-value $= 0.2992$) supports the linearity of the regression relating $Y' = \sqrt{Y}$ to $X' = \sqrt{X}$.

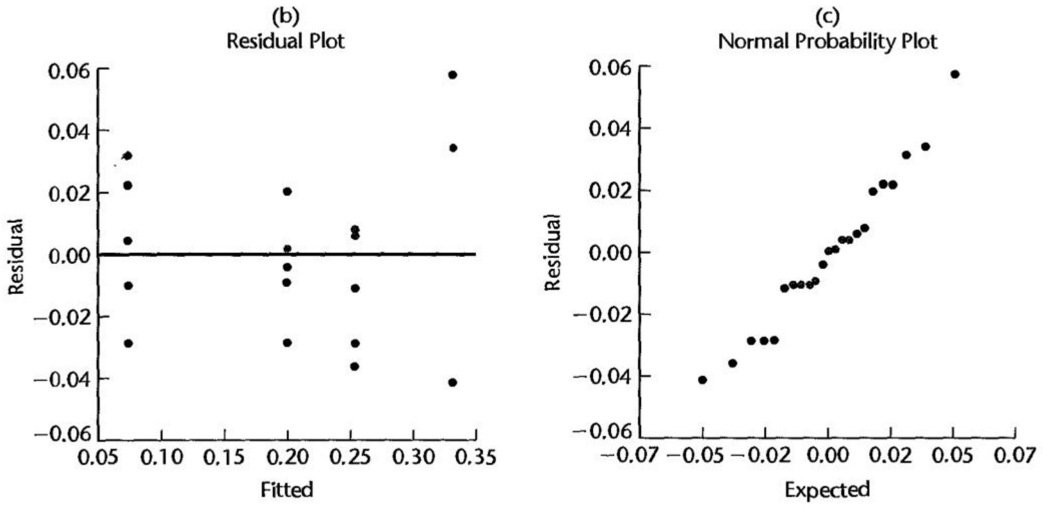
FIGURE 3.23 SAS-JMP Regression Output and Diagnostic Plots for Transformed Response and Predictor Variables—Plutonium Measurement Example.

(a) Regression Output



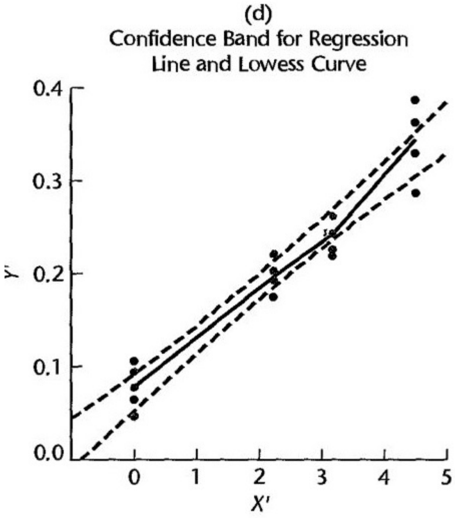
```
Term            Estimate      Std Error    t Ratio    Prob>|t|
Intercept       0.0730056       0.00783       9.32      0.0000
Sqrt Plutonium  0.0573055       0.00302      19.00      0.0000

Source               DF  Sum of Squares  Mean Square   F Ratio
Model                 1      0.22141612     0.221416  360.9166
Error                21      0.01288314     0.000613     Prob>F
C Total              22      0.23429926                 0.0000

Source               DF  Sum of Squares  Mean Square   F Ratio
Lack of Fit           2      0.00153683     0.000768    1.2868
Pure Error           19      0.01134631     0.000597     Prob>F
Total Error          21      0.01288314                 0.2992
```

(b) *Diagnostic Plots* (Figure 3.23b, c) the residual plot shows that the square root transformation of the predictor variable has eliminated the lack of fit. It also suggests that some nonconstancy of the error variance may still remain; but if so, it does not appear to be substantial. The normal probability plot of the residuals in Figure 3.23c appears to be satisfactory.



(c) *Diagnostic Tests*: the Correlation Test ($r = 0.986$) supports the assumption of normally distributed error terms (the interpolated critical value in Table B.6 for $\alpha = 0.05$ and $n = 23$ is 0.9555). The Breusch-Pagan Test ($X^2_{BP} = 3.85$ with a P-value $= 0.05$) supports the conclusion from the residual plot that the nonconstancy of the error variance is not substantial.

(d) *Additional Results*: (Figure 3.23d) the scatter plot of X and Y with the confidence band for the fitted regression line: $\hat{Y}' = 0.0730 + 0.0573X'$. The regression line has been estimated fairly precisely. The lowess curve falls entirely within the confidence band, supporting the reasonableness of a linear regression relation between Y' and X'.



😍 TA Class

- **Problems**: 3.4, 3.9, 3.13, 3.15, 3.17
- **Exercises**: 3.20, 3.21
- **Projects**: 3.25

“很多時候我們缺的不是機會，而是決心與勇氣。”
“Often times we lack is not the opportunity, but courage and determination.”
— *心靈補手 (Good Will Hunting, 1997)*

Regression Analysis (I)

Kutner's Applied Linear Statistical Models (5/E)

Chapter 5: Matrix Approach to Simple Linear Regression Analysis

Thursday 09:10-12:00, 商館 260205

Han-Ming Wu

Department of Statistics, National Chengchi University

http://www.hmwu.idv.tw

Overview

1. The matrix approach is practically a necessity in multiple regression analysis, since it permits extensive systems of equations and large arrays of data to be denoted compactly and operated upon efficiently.

2. This chapter gives a brief introduction to matrix algebra .

3. Then we apply matrix methods to the simple linear regression model.

5.1 Matrices

Definition of Matrix

1. A matrix is a rectangular array of elements arranged in rows and columns.

2. A matrix with r rows and c columns will be represented either in full:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1c} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2c} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ij} & \cdots & a_{ic} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{r1} & a_{r2} & \cdots & a_{rj} & \cdots & a_{rc} \end{bmatrix}$$

or in abbreviated form:

$$\mathbf{A} = \underline{\quad [a_{ij}] \quad}, \; i = 1, \cdots, r; j = 1, \cdots, c$$

or simply by a boldface symbol, such as **A**.

Square Matrix

1. A matrix is said to be square if the number of rows __equals__ the number of columns.

Vector

1. A matrix containing only one column is called a __column__ vector or simply a vector.

$$\mathbf{C} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \end{bmatrix}$$

the vector **C** is a __5×1 matrix__.

2. A matrix containing only one row is called a __row vector__: e.g., $\mathbf{B}' = [15\ 25\ 50]$. We use the prime symbol (__transpose__) for row vectors. Note that the row vector **B**' is a __1×3__ matrix.

Transpose

1. The transpose of a matrix **A** is another matrix, denoted by __**A**'__, that is obtained by interchanging corresponding columns and rows of the matrix **A**.

$$\mathbf{A} = \begin{bmatrix} 2 & 5 \\ 7 & 10 \\ 3 & 4 \end{bmatrix}$$

then the transpose **A**' is:

$$\mathbf{A}' = \underline{\begin{bmatrix} 2 & 7 & 3 \\ 5 & 10 & 4 \end{bmatrix}}$$

2. The transpose of a column vector is a row vector, and vice versa. This is the reason why we used the symbol **B**' earlier to identify a row vector, since it may be thought of as the transpose of a column vector **B**. In general, we have:

$$\mathbf{A} = [a_{ij}], \quad \underline{\mathbf{A}' = [a_{ji}]}$$

Equality of Matrices

1. Two matrices **A** and **B** are said to be equal if they have the same dimension and if all corresponding <u>elements are equal</u>.

Regression Examples

1. In regression analysis, one basic matrix is the vector **Y**, consisting of the n observations on response variable

$$\mathbf{Y} = \underline{\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}}$$

2. Another basic matrix in regression analysis is the **X** matrix, which is defined as follows for simple linear regression analysis:

$$\mathbf{X} = \underline{\begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}}$$

The matrix **X** consists of a column of 1s and a column containing the n observations on the predictor variable X. The **X** matrix is often referred to as the design matrix.

5.2 Matrix Addition and Subtraction

1. Adding or subtracting two matrices requires that they have the same dimension. The sum, or difference, of two matrices is another matrix whose elements each consist of the sum, or difference, of the corresponding elements of the two matrices.

2.

$$\text{if } \mathbf{A}_{r\times c} = [a_{ij}], \quad \mathbf{B}_{r\times c} = [b_{ij}], \text{ then } \mathbf{A} \pm \mathbf{B} = \underline{\quad [a_{ij}] \pm [b_{ij}] \quad}$$

3. The regression model: $Y_i = E(Y_i) + \varepsilon_i, \; i = 1, \cdots, n$ can be written in matrix notation:

$$\underline{\quad \mathbf{Y} = E(\mathbf{Y}) + \boldsymbol{\varepsilon} \quad}$$

4. The observations vector $\mathbf{Y}$ equals the sum of two vectors, a vector containing the expected values and another containing the error terms.

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} E(Y_1) \\ E(Y_2) \\ \vdots \\ E(Y_n) \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \begin{bmatrix} E(Y_1) + \varepsilon_1 \\ E(Y_2) + \varepsilon_2 \\ \vdots \\ E(Y_n) + \varepsilon_n \end{bmatrix}$$

5.3 Matrix Multiplication

Multiplication of a Matrix by a Scalar

1. A scalar is an ordinary number or a symbol representing a number. In multiplication of a matrix by a scalar, every element of the matrix is multiplied by the scalar.

2. If $\mathbf{A} = [a_{ij}]$ and k is the scalar, then

$$k\mathbf{A} = \mathbf{A}k = \underline{\quad [ka_{ij}] \quad}$$

Multiplication of a Matrix by a Matrix

1. In general, the product $\mathbf{AB}$ is defined only when the number of columns in $\mathbf{A}$ equals the number of rows in $\mathbf{B}$ so that there will be corresponding terms in the cross products.

2. Note that the dimension of the product $\mathbf{AB}$ is given by the number of rows in $\mathbf{A}$ and the number of columns in $\mathbf{B}$. Note also that in the second case the product $\mathbf{BA}$ would not be defined since the number of columns in $\mathbf{B}$ is not equal to the number of rows in $\mathbf{A}$.

3. In general, if $\mathbf{A} = [a_{ik}]$ has dimension $r \times c$ and $\mathbf{B} = [b_{kj}]$ has dimension $c \times s$, the product $\mathbf{AB}$ is a matrix of dimension $r \times s$ whose element in the ith row and jth column is:

$$\mathbf{AB} = \underline{\left[\sum_{k=1}^{c} a_{ik}b_{kj}\right]}$$

Regression Examples

1. A product frequently needed is **Y'Y**, where **Y** is the vector of observations on the response variable

$$\mathbf{Y}'\mathbf{Y} = [Y_1\ Y_2\ \cdots\ Y_n]\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \underline{Y_1^2 + Y_2^2 + \cdots + Y_n^2} = \underline{\textstyle\sum_{i=1}^{n} Y_i^2}$$

2. **X'X** is a 2×2 matrix:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_1 & X_2 & \cdots & X_n \end{bmatrix}\begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} = \underline{\begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix}}$$

3. **X'Y** is a 2×1 matrix:

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_1 & X_2 & \cdots & X_n \end{bmatrix}\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \underline{\begin{bmatrix} \sum Y_i \\ \sum X_iY_i \end{bmatrix}}$$

5.4 Special Types of Matrices

Certain special types of matrices arise regularly in regression analysis. We consider the most important of these.

Symmetric Matrix

1. If $\underline{\ \mathbf{A} = \mathbf{A}'\ }$, **A** is said to be symmetric.

2. A symmetric matrix necessarily is square.

3. Symmetric matrices arise typically in regression analysis when we premultiply a matrix, say, **X**, by its transpose, **X**'. The resulting matrix, $\mathbf{X}'\mathbf{X}$, is symmetric.

Diagonal Matrix

1. A diagonal matrix is a square matrix whose off-diagonal elements are all zeros.

2. We will often not show all zeros for a diagonal matrix, presenting it in the form:

$$\mathbf{B} = \begin{bmatrix} 4 & & & \\ & 1 & & \\ & & 10 & \\ & & & 5 \end{bmatrix}$$

3. **Identity Matrix** The identity matrix or unit matrix is denoted by $\mathbf{I}$. It is a diagonal matrix whose elements on the main diagonal are all 1s.

4. Premultiplying or postmultlying any $r \times r$ matrix $\mathbf{A}$ by the $r \times r$ identity matrix $\mathbf{I}$ leaves $\mathbf{A}$ unchanged.

$$\mathbf{AI} = \mathbf{IA} = \mathbf{A}$$

5. A **scalar matrix** is a diagonal matrix whose main-diagonal elements are the same. A scalar matrix can be expressed as $k\mathbf{I}$, where k is the scalar.

6. Multiplying an $r \times r$ matrix $\mathbf{A}$ by the $r \times r$ scalar matrix $k\mathbf{I}$ is equivalent to multiplying $\mathbf{A}$ by the scalar k.

Vector and Matrix with All Elements Unity

1. A column vector with all elements 1 will be denoted by $\mathbf{1}$ and a square matrix with all elements 1 will be denoted by $\mathbf{J}$.

2. Note that for an $n \times 1$ vector $\mathbf{1}$ we obtain:

$$\mathbf{1}'\mathbf{1} = [1\ 1\ \cdots\ 1]\begin{bmatrix}1\\1\\\vdots\\1\end{bmatrix} = \underline{\quad n \quad}$$

and

$$\mathbf{1}\mathbf{1}' = \begin{bmatrix}1\\1\\\vdots\\1\end{bmatrix}[1\ 1\ \cdots\ 1] = \begin{bmatrix}1 & \cdots & 1\\ \vdots & & \vdots\\ 1 & \cdots & 1\end{bmatrix} = \underline{\quad \mathbf{J}_{n\times n} \quad}$$

Zero Vector

1. A zero vector is a vector containing only zeros. The zero column vector will be denoted by $\underline{\quad \mathbf{0} \quad}$.

5.5 Linear Dependence and Rank of Matrix

Linear Dependence

1. Consider the following matrix:

$$\mathbf{A} = \begin{bmatrix}1 & 2 & 5 & 1\\ 2 & 2 & 10 & 6\\ 3 & 4 & 15 & 1\end{bmatrix}$$

We view $\mathbf{A}$ as being made up of four column vectors. Note that the third column vector is a multiple of the first column vector.

$$\begin{bmatrix}5\\10\\15\end{bmatrix} = 5\begin{bmatrix}1\\2\\3\end{bmatrix}$$

We say that the columns of $\mathbf{A}$ are __linearly dependent__. They contain __redundant__ information, since one column can be obtained as a linear combination of the others.

2. We define the set of c column vectors $\mathbf{C}_1, \cdots, \mathbf{C}_c$ in an $r \times c$ matrix to be linearly dependent if one vector can be expressed as a linear combination of the others. If no vector in the set can be so expressed, we define the set of vectors to be linearly independent.

3. When c scalars $k_1, \cdots, k_c$, not all zero, can be found such that:

$$k_l\mathbf{C}_1 + k_2\mathbf{C}_2 + \cdots + k_c\mathbf{C}_c = \mathbf{0}$$

   where $\mathbf{0}$ denotes the zero column vector, the c column vectors are linearly dependent. If the only set of scalars for which the equality holds is $k_1 = 0, \cdots, k_c = 0$, the set of c column vectors is linearly independent.

4. For our example, $k_1 = 5, k_2 = 0, k_3 = -1, k_4 = 0$ leads to:

$$5\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} + 0\begin{bmatrix} 2 \\ 2 \\ 4 \end{bmatrix} - 1\begin{bmatrix} 5 \\ 10 \\ 15 \end{bmatrix} + 0\begin{bmatrix} 1 \\ 6 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

   Hence, the column vectors are linearly dependent. Note that some of the k_j equal zero here. For linear dependence, it is only required that not all k_j be zero.

Rank of Matrix

1. The rank of a matrix is defined to be the maximum number of linearly independent columns in the matrix.

2. The rank of a matrix is unique and can equivalently be defined as the maximum number of linearly independent rows.

3. It follows that the rank of an $r \times c$ matrix cannot exceed $\min(r, c)$, the minimum of the two values r and c.

4. When a matrix is the product of two matrices, its rank cannot exceed the smaller of the two ranks for the matrices being multiplied. Thus, if $\mathbf{C} = \mathbf{AB}$, the rank of $\mathbf{C}$ cannot exceed $\min(\text{rank}(\mathbf{A}), \text{rank}(\mathbf{B}))$.

5.6 Inverse of a Matrix

1. In matrix algebra, the inverse of a matrix **A** is another matrix, denoted by $\underline{\ \mathbf{A}^{-1}\ }$, such that

$$\underline{\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}}$$

where **I** is the identity matrix.

Finding the Inverse

1. An inverse of a square $r \times r$ matrix exists if the __rank__ of the matrix is $\underline{\ r\ }$. Such a matrix is said to be nonsingular or of full rank.

2. An $r \times r$ matrix with rank less than r is said to be __singular__ or __not of full rank__, and does not have an inverse. The inverse of an $r \times r$ matrix of full rank also has rank r.

3. Finding the inverse of a matrix can often require a large amount of computing. We shall take the approach that the inverse of a 2×2 matrix and a 3×3 matrix can be calculated by hand. For any larger matrix, one ordinarily uses a computer to find the inverse.

4. If

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

then

$$\mathbf{A}^{-1} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \underline{\begin{bmatrix} d/D & -b/D \\ -c/D & a/D \end{bmatrix}}$$

where $\underline{\ D = ad - bc\ }$, D is called the __determinant__ of the matrix **A**.

5. If **A** were singular, its determinant would equal __zero__ and no inverse of **A** would exist.

Regression Example

1. The principal inverse matrix encountered in regression analysis is the inverse of the matrix $\mathbf{X}'\mathbf{X}$.

✎ Question .... (p191)

Find the inverse of the matrix $\mathbf{X}'\mathbf{X}$:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix}$$

sol:

Uses of Inverse Matrix

1. In matrix algebra, if we have an equation:

$$\mathbf{A}\mathbf{Y} = \mathbf{C}.$$

We correspondingly premultiply both sides by $\mathbf{A}^{-1}$, assuming $\mathbf{A}$ has an inverse

$$\underline{\quad \mathbf{A}^{-1}\mathbf{A}\mathbf{Y} \quad} = \underline{\quad \mathbf{A}^{-1}\mathbf{C} \quad}$$

we obtain the solution:

$$\mathbf{Y} = \underline{\quad \mathbf{A}^{-1}\mathbf{C} \quad}.$$

5.7 Some Basic Results for Matrices

We list here, without proof, some basic results for matrices which we will utilize in later work.

$$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$$

$$
\begin{aligned}
(\mathbf{A}+\mathbf{B})+\mathbf{C} &= \mathbf{A}+(\mathbf{B}+\mathbf{C}) \\
(\mathbf{A}\mathbf{B})\mathbf{C} &= \mathbf{A}(\mathbf{B}\mathbf{C}) \\
\mathbf{C}(\mathbf{A}+\mathbf{B}) &= \mathbf{C}\mathbf{A}+\mathbf{C}\mathbf{B} \\
k(\mathbf{A}+\mathbf{B}) &= k\mathbf{A}+k\mathbf{B} \\
(\mathbf{A}')' &= \mathbf{A} \\
(\mathbf{A}+\mathbf{B})' &= \mathbf{A}'+\mathbf{B}' \\
(\mathbf{A}\mathbf{B})' &= \underline{\quad \mathbf{B}'\mathbf{A}' \quad} \\
(\mathbf{A}\mathbf{B}\mathbf{C})' &= \underline{\quad \mathbf{C}'\mathbf{B}'\mathbf{A}' \quad} \\
(\mathbf{A}\mathbf{B})^{-1} &= \underline{\quad \mathbf{B}^{-1}\mathbf{A}^{-1} \quad} \\
(\mathbf{A}\mathbf{B}\mathbf{C})^{-1} &= \underline{\quad \mathbf{C}^{-1}\mathbf{B}^{-1}\mathbf{A}^{-1} \quad} \\
(\mathbf{A}^{-1})^{-1} &= \mathbf{A} \\
(\mathbf{A}')^{-1} &= \underline{\quad (\mathbf{A}^{-1})' \quad}
\end{aligned}
$$

5.8 Random Vectors and Matrices

Expectation of Random Vector or Matrix

1. A random vector or a random matrix contains elements that are __random variables__. Thus, the observations vector **Y** in (5.4) is a random vector since the Y_i elements are random variables.

2. The expected value of **Y** is a vector, denoted by $E(\mathbf{Y})$, that is defined as follows:

$$E(\mathbf{Y}) = \underline{\quad [E(Y_i)] \quad}, \ i = 1, \cdots, n.$$

3. For the error terms in regression model, we have

$$\underline{\quad E(\boldsymbol{\varepsilon}) = \mathbf{0} \quad}.$$

Variance-Covariance Matrix of Random Vector

1. The variance-covariance matrix of $\mathbf{Y}$, denoted by $\sigma^2(\mathbf{Y})$:

$$\begin{aligned}\sigma^2(\mathbf{Y}) &= \underline{E[(\mathbf{Y}-E(\mathbf{Y}))(\mathbf{Y}-E(\mathbf{Y}))']} \\ &= \begin{bmatrix} \sigma^2(Y_1) & \sigma^2(Y_1,Y_2) & \cdots & \sigma^2(Y_1,Y_n) \\ \\ \sigma^2(Y_2,Y_1) & \sigma^2(Y_2) & \cdots & \sigma^2(Y_2,Y_n) \\ \vdots & \vdots & & \vdots \\ \sigma^2(Y_n,Y_1) & \sigma^2(Y_n,Y_2) & \cdots & \sigma^2(Y_n,Y_n) \end{bmatrix}\end{aligned}$$

2. Note that the $\underline{\text{variances } \sigma^2(Y_i)}$ are on the main diagonal, and the $\underline{\text{covariance } \sigma^2(Y_i,Y_j)}$ is found in the ith row and jth column of the matrix.

3. The error terms in regression model have constant variance:

$$\sigma^2(\boldsymbol{\varepsilon}) = \underline{\sigma^2\mathbf{I}}.$$

Some Basic Results

1. Frequently, we shall encounter a random vector $\mathbf{W}$ that is obtained by premultiplying the random vector $\mathbf{Y}$ by a constant matrix $\mathbf{A}$ (a matrix whose elements are fixed): $\mathbf{W} = \mathbf{AY}$. Some basic results for this case are:

$$\begin{aligned} E(\mathbf{A}) &= \underline{\mathbf{A}} \\ E(\mathbf{W}) &= E(\mathbf{AY}) = \underline{\mathbf{A}E(\mathbf{Y})} \\ \sigma^2(\mathbf{W}) &= \sigma^2(\mathbf{AY}) = \underline{\mathbf{A}\sigma^2(\mathbf{Y})\mathbf{A}'}, \end{aligned}$$

where $\sigma^2(\mathbf{Y})$ is the variance-covariance matrix of $\mathbf{Y}$.

✐ Question .... (p42)

Suppose that a random vector $\mathbf{W}$ that is obtained by premultiplying the random vector $\mathbf{Y}$ by a constant matrix $\mathbf{A}$, that is $\mathbf{W} = \mathbf{A}\mathbf{Y}$. Find the expected value and the variance-covariance matrix of $\mathbf{W}$.

sol:

Multivariate Normal Distribution

1. The density function of the multivariate normal distribution can now be stated as follows:

$$f(\mathbf{Y}) = \underline{\frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{Y}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{Y}-\boldsymbol{\mu})\right]},$$

where $\mathbf{Y}$ containing an observation on each of the p Y variables

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{bmatrix}.$$

2. The mean vector $E(\mathbf{Y})$, denoted by $\underline{\boldsymbol{\mu}}$, contains the expected values for each of the p Y variables:

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}.$$

3. The variance-covariance matrix $\sigma^2(\mathbf{Y})$ is denoted by $\underline{\ \boldsymbol{\Sigma}\ }$: and contains as always the variances and covariances of the p Y variables:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{bmatrix}$$

   σ_i^2 denotes the variance of Y_1, σ_{ij} denotes the covariance of Y_i and Y_j.

4. The multivariate normal density function has properties that correspond to the ones described for the __bivariate__ normal distribution.

5. For instance, if $Y_1, \cdots, Y_p$ are jointly normally distributed (i.e., they follow the multivariate normal distribution), the marginal probability distribution of each variable Y_k is normal, with mean μ_k and standard deviation σ_k.

5.9 Simple Linear Regression Model in Matrix Terms

1. The normal error regression model (2.1):

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \cdots, n$$

2. The normal error regression model in matrix terms:

$$\underline{\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times 2}\boldsymbol{\beta}_{2\times 1} + \boldsymbol{\varepsilon}_{n\times 1}},$$

   where

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

   $\boldsymbol{\varepsilon}$ is a vector of independent normal random variables with $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\sigma^2(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$

5.10 Least Squares Estimation of Regression Parameters

Normal Equations

✎ Question . . . . (p200)

Express the normal equations (1.9),

$$\begin{aligned} nb_0 + b_1 \sum X_i &= \sum Y_i \\ b_0 \sum X_i + b_1 \sum X_i^2 &= \sum X_i Y_i \end{aligned}$$

in the matrix form

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y}$$

where **b** is the vector of the least squares regression coefficients:

$$\mathbf{b}_{2\times 1} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$$

sol:

✎ Question .... (p201)

Derive the normal equations by the method of least squares in matrix notation.

sol:

Estimated Regression Coefficients

1. Obtain the estimated regression coefficients from the normal equations (5.59) by matrix methods, We premultiply both sides by

$$\underline{\quad (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \quad}$$

We then find, since $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{I}$ and $\mathbf{I}\mathbf{b} = \mathbf{b}$,

$$\mathbf{b} = \underline{\quad (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \quad}$$

Question .... (p200)

Use matrix methods to obtain the estimated regression coefficients for the Toluca Company example.

sol:

5.11 Fitted Values and Residuals

Fitted Values

1. Let the vector of the fitted values Y_i be denoted by $\hat{\mathbf{Y}}$, then

$$\hat{\mathbf{Y}} = \underline{\quad \mathbf{Xb} \quad}$$

$$\begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} b_0 + b_1 X_1 \\ b_0 + b_1 X_2 \\ \vdots \\ b_0 + b_1 X_n \end{bmatrix}$$

2. **Hat Matrix** We can express the matrix result for $\hat{\mathbf{Y}}$ as follows by using the expression for $\mathbf{b}$ in (5.60):

$$\hat{\mathbf{Y}} = \underline{\quad \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \quad}$$

or, equivalently:

$$\hat{\mathbf{Y}} = \underline{\quad \mathbf{HY} \quad}$$

where

$$\mathbf{H}_{n\times n} = \underline{\quad \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \quad}$$

3. The fitted values $\hat{Y}_i$ can be expressed as linear combinations of the response variable observations Y_i, with the coefficients being elements of the matrix **H**.

4. The **H** matrix involves only the observations on the predictor variable **X**. The square $n \times n$ matrix **H** is called the **Hat matrix**. It plays an important role in diagnostics for regression analysis (Chapter 10) when we consider whether regression results are unduly influenced by one or a few observations.

5. The matrix **H** is symmetric and has the special property (called idempotency):

   $$\underline{\quad \mathbf{HH} = \mathbf{H} \quad}$$

   In general, a matrix **M** is said to be idempotent if **MM** = **M**.

Residuals

1. Let the vector of the residuals $e_i = Y_i - \hat{Y}_i$ be denoted by **e**:

   $$\mathbf{e}_{n\times 1} = \underline{\quad \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{Xb} \quad}$$

2. **Variance-Covariance Matrix of Residuals**. The residuals e_i, like the fitted values $\hat{Y}_i$, can be expressed as linear combinations of the response variable observations Y_i , using the result in (5.73) for $\hat{\mathbf{Y}}$:

   $$\mathbf{e} = \underline{\quad \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{HY} = (\mathbf{I} - \mathbf{H})\mathbf{Y} \quad}$$

   We thus have the important result:

   $$\mathbf{e} = \underline{\quad (\mathbf{I} - \mathbf{H})\mathbf{Y} \quad}$$

   where **H** is the hat matrix defined in (5.53a). The matrix $\mathbf{I} - \mathbf{H}$, like the matrix **H**, is symmetric and idempotent.

3. The variance-covariance matrix of the vector of residuals $\mathbf{e}$ involves the matrix $\mathbf{I}-\mathbf{H}$:

$$\sigma^2(\mathbf{e}) = \underline{\quad \sigma^2(\mathbf{I}-\mathbf{H}) \quad}$$

and is estimated by:

$$s^2(\mathbf{e}) = \underline{\quad MSE(\mathbf{I}-\mathbf{H}) \quad}$$

Question .... (p204)

Show that the variance-covariance matrix of $\mathbf{e}$ is $\sigma^2(\mathbf{e}) = \sigma^2(\mathbf{I}-\mathbf{H})$.

sol:

5.12 Analysis of Variance Results

Sums of Squares

Question .... (p204)

Express the sums of squares, $SSTO$, SSE and SSR in matrix notation.

sol:

Sums of Squares as Quadratic Forms

1. In general, a quadratic form is defined as:

$$\mathbf{Y}'\mathbf{A}\mathbf{Y}_{1\times 1} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}Y_iY_j \quad , \quad \text{where} \quad a_{ij} = a_{ji}.$$

2. $\mathbf{A}$ is a symmetric $n \times n$ matrix and is called the matrix of the quadratic form.

3. The ANOVA sums of squares $SSTO$, SSE, and SSR are all quadratic forms, as can be seen by reexpressing $\mathbf{b}'\mathbf{X}'$.

Question .... (p206)

Show that the ANOVA sums of squares $SSTO$, SSE, and SSR are all quadratic forms.

sol:

5.13 Inferences in Regression Analysis

Regression Coefficients

✎ Question .... (p42)

(a) Derive the variance-covariance matrix of the simple linear regression coefficients, **b** by matrix methods. (b) Obtain the estimated variance-covariance matrix of **b**.

sol:

Mean Response*

Prediction of New Observation*

😍 TA Class

- **Problems**: 5.5, 5.16, 5.22, 5.24, 5.26
- **Exercises**: 5.31

“會讓人後悔的從來都不是失敗，而是當機會出現時你沒有全力以赴。”
“Regrets don’t come from failure, they come from moments you failed to give your best.”
— *墊底辣妹 (Flying Colors, 2015)*

Regression Analysis (I)

Kutner's Applied Linear Statistical Models (5/E)

Chapter 6: Multiple Regression (I)

Thursday 09:10-12:00, 商館 260205

Han-Ming Wu

Department of Statistics, National Chengchi University

http://www.hmwu.idv.tw

Overview

1. Discuss a variety of multiple regression models. (more than one predictors)
2. Present the basic statistical results for multiple regression in matrix form.
3. The matrix expressions for multiple regression are the same as for SLR.
4. An example to illustrate a variety of inferences and residual analyses in multiple regression analysis.

6.1 Multiple Regression Models

Need for Several Predictor Variables

1. A single predictor variable in the model would have provided an inadequate description since a number of key variables affect the response variable.
2. Predictions of the response variable based on a model containing only a single predictor variable are too imprecise to be useful.

3. Multiple regression analysis is highly useful in experimental situations where the experimenter can control the predictor variables .

4. The multiple regression models can be utilized for either observational data or for experimental data from a completely randomized design.

First-Order Model with Two Predictor Variables

1. When there are two predictor variables X_1 and X_2, the regression model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \qquad (6.1)$$

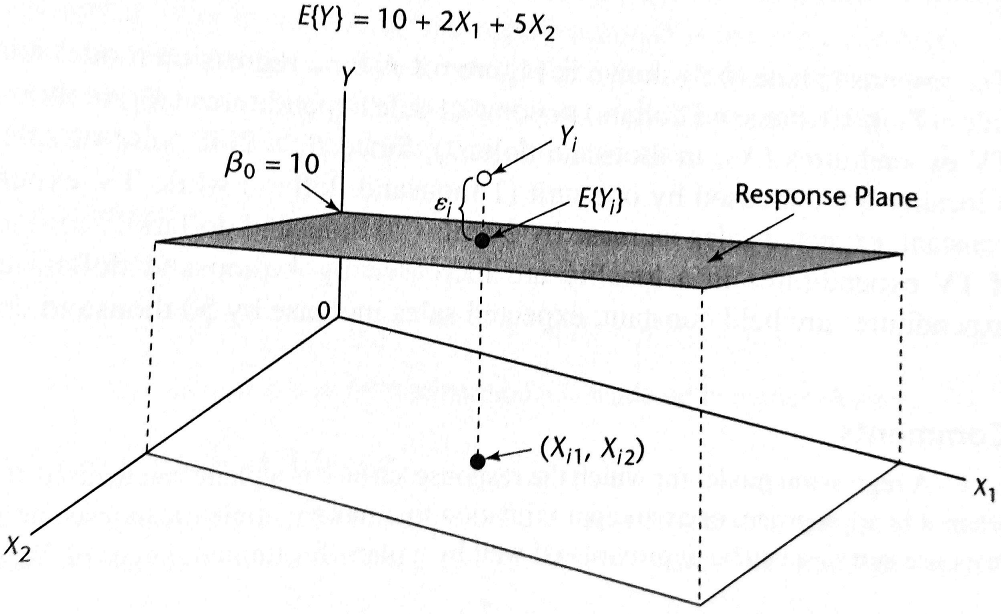
is called a first-order model with two predictor variables.

2. Assuming that $E(\varepsilon) = 0$, the regression function for model (6.1) is a plane:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \qquad (6.2)$$

3. (Figure 6.1) The response plane: $E(Y) = 10 + 2X_1 + 5X_2$ (6.3).

FIGURE 6.1 **Response Function is a Plane—Sales Promotion Example.**



(a) Any point on the response plane (6.3) corresponds to the mean response $E(Y)$ at the given combination of levels of X_1 and X_2 .

(b) The error term $\underline{\varepsilon_i = Y_i - E(Y_i)}$: the vertical rule between Y_i and the response plane represents the difference between Y_i and the mean $E(Y_i)$ of the probability distribution of Y for the given (X_{i1}, X_{i2}) combination.

4. The regression function in multiple regression is called a $\underline{\text{regression surface}}$ or a $\underline{\text{response surface}}$. In Figure 6.1, the response surface is a $\underline{\text{plane}}$, but in other cases the response surface may be more $\underline{\text{complex}}$ in nature.

5. **Meaning of Regression Coefficients**

(a) The parameter β_0 is the $\underline{Y \text{ intercept}}$ of the regression plane.

(b) If the scope of the model includes $\underline{X_1 = 0,\ X_2 = 0}$, then β_0 represents the mean response $E(Y)$ at $X_1 = 0$, $X_2 = 0$. Otherwise, β_0 $\underline{\text{does not}}$ have any particular meaning as a separate term in the regression model.

(c) The parameter β_1 (β_2) indicates the $\underline{\text{change}}$ in the mean response $E(Y)$ per unit increase in $\underline{X_1\ (X_2)}$ when $\underline{X_2\ (X_1)}$ is held constant.

(d) When the effect of X_1 on the mean response does not depend on the level of X_2, and correspondingly the effect of X_2 does not depend on the level of X_1, the two predictor variables are said to have $\underline{\text{additive effects}}$ or not to $\underline{\text{interact}}$.

(e) Thus, the first-order regression model (6.1) is designed for predictor variables whose effects on the mean response are additive or do not interact.

6. The parameters β_1 and β_2 are sometimes called $\underline{\text{partial regression coefficients}}$ because they reflect the partial effect of one predictor variable when the other predictor variable is included in the model and is $\underline{\text{held constant}}$.

First-Order Model with More than Two Predictor Variables

1. The regression model:

$$Y_i = \underline{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i} \qquad (6.5)$$

$$= \underline{\beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik} + \varepsilon_i} \qquad (6.5a)$$

$$= \underline{\sum_{k=0}^{p-1} \beta_k X_{ik} + \varepsilon_i} \qquad \text{where } X_{i0} \equiv 1 \qquad (6.5b)$$

is called a first-order model with $p-1$ predictor variables.

2. Assuming that $E(\varepsilon_i) = 0$, the response function for regression model (6.5) is:

$$E(Y) = \underline{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1}} \qquad (6.6)$$

3. This response function is a __hyperplane__, which is a plane in more than two dimensions.

4. The parameter β_k indicates the __change in the mean response $E(Y)$__ with a unit increase in the predictor variable X_k when all other predictor variables in the regression model are held constant.

5. The first-order regression model (6.5) is designed for predictor variables whose effects on the mean response are __additive__ and therefore do not interact.

General Linear Regression Model

1. Define the general linear regression model, with normal error terms, simply in terms of X variables:

$$Y_i = \underline{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i} \qquad (6.7)$$

where:

   (a) $\beta_0, \beta_1, \cdots, \beta_{p-1}$ are __parameters__.

   (b) $X_{i1}, \cdots, X_{i,p-1}$ are __known__ constants (predictors, explanatory variables).

   (c) ε_i are independent __$N(0, \sigma^2)$__, $i = 1, \cdots, n$.

2. The response function for regression mode1 (6.7) is:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1} \qquad (6.8)$$

3. Thus, the general linear regression model with normal error terms implies that the observations Y_i are independent __normal variables__, with mean __$E(Y)$__ as given by (6.8) and with constant variance __σ^2__.

4. **Qualitative Predictor Variables**

   (a) The general linear regression model (6.7) encompasses not only quantitative predictor variables but also qualitative ones, such as gender (male, female) or disability status (not disabled, partially disabled, fully disabled).

   (b) Use indicator variables that take on the values 0 and 1 to identify the classes of a qualitative variable.

   (c) $\boxed{\text{Example}}$ Consider a regression analysis to predict the length of hospital stay (Y) based on the age (X_1) and gender (X_2) of the patient. The first-order regression model is:

$$Y_i = \underline{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i} \qquad (6.9)$$

$$X_{i1} = i\text{th patient's age}$$

$$X_{i2} = \begin{cases} 1 & \text{if } i\text{th patient female} \\ 0 & \text{if } i\text{th patient male} \end{cases}$$

   The response function for regression model (6.9) is:

$$E(Y) = \underline{\beta_0 + \beta_1 X_1 + \beta_2 X_2} \qquad (6.10)$$

   For male patients, $X_2 = 0$ and response function (6.10) becomes:

$$E(Y) = \underline{\beta_0 + \beta_1 X_1}, \qquad \text{Male patients} \qquad (6.10a)$$

   For female patients, $X_2 = 1$ and response function (6.10) becomes:

$$E(Y) = \underline{(\beta_0 + \beta_2) + \beta_1 X_1}, \qquad \text{Female patients} \qquad (6.10b)$$

   These two response functions represent parallel straight lines with different intercepts.

   (d) In general, we represent a qualitative variable with c classes by means of $c - 1$ indicator variables. (details in Chapter 8)

5. $\boxed{\text{Example}}$ The first-order model with age, gender (male, female) or disability status (not disabled, partially disabled, fully disabled) as predictor variables then is:

$$Y_i = \underline{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i} \qquad (6.11)$$

where:

$$X_{i1} = i\text{th patient's age}$$

$$X_{i2} = \underline{\begin{cases} 1 & \text{if } i\text{th patient female} \\ 0 & \text{if } i\text{th patient male} \end{cases}}$$

$$X_{i3} = \underline{\begin{cases} 1 & \text{if } i\text{th patient not disabled} \\ 0 & \text{otherwise} \end{cases}}$$

$$X_{i4} = \underline{\begin{cases} 1 & \text{if } i\text{th patient partially disabled} \\ 0 & \text{otherwise} \end{cases}}$$

6. **Polynomial Regression**

   (a) Polynomial regression models are special cases of the general linear regression model. They contain __squared__ and __higher-order__ terms of the predictor variable(s), making the response function __curvilinear__.

   (b) $\boxed{\text{Example}}$ A polynomial regression model with one predictor variable:

   $$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i \qquad (6.12)$$

   (c) If we let $X_{i1} = X_i$ and $X_{i2} = X_i^2$; we can write (6.12) as

   $$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

   which is in the form of general linear regression model (6.7). (detail in Chapter 8).

7. **Transformed Variables**

   (a) Models with transformed variables involve complex, curvilinear response functions, yet still are special cases of the general linear regression model.

   (b) $\boxed{\text{Example}}$ A model with a transformed __$Y_i' = \log Y_i$__ variable:

   $$Y_i' = \log Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i.$$

   (c) $\boxed{\text{Example}}$ A model with a transformed __$Y_i' = 1/Y_i$__ variable:

   $$Y_i' = 1/Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i.$$

8. **Interaction Effects**

   (a) When the effects of the predictor variables on the response variable are not additive, the effect of one predictor variable depends on the levels of the other predictor variables. The general linear regression model (6.7) encompasses regression models with nonadditive or interacting effects.

   (b) Example An example of a nonadditive regression model with two predictor variables X_1 and X_2:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i \\ &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i \end{aligned}$$

   The response function is complex because of the interaction term $X_{i3} = X_{i1} X_{i2}$. It is a special case of the general linear regression model. (detail in Chapter 8)

9. **Combination of Cases**

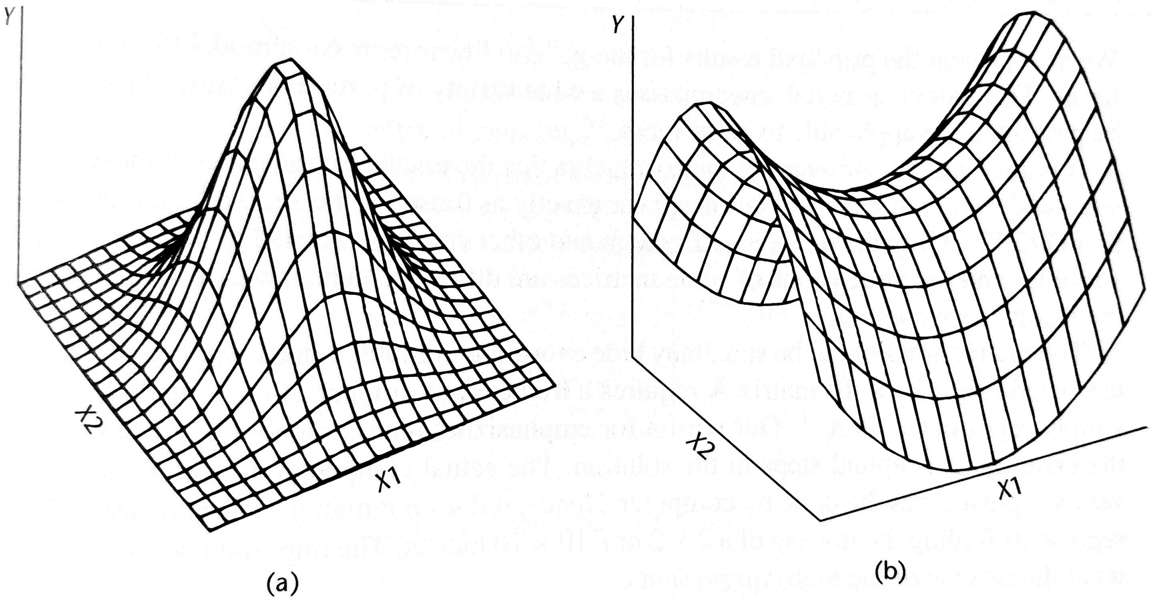
   (a) A regression model may combine several of the elements we have just noted and still be treated as a general linear regression model.

   (b) Example Consider the following regression model containing linear and quadratic terms for each of two predictor variables and an interaction term represented by the cross-product term:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1}^2 + \beta_3 X_{i2} + \beta_4 X_{i2}^2 + \beta_5 X_{i1} X_{i2} + \varepsilon_i \\ &= \beta_0 + \beta_1 Z_{i1} + \beta_2 Z_{i2} + \beta_3 Z_{i3} + \beta_4 Z_{i4} + \beta_5 Z_{i5} + \varepsilon_i. \end{aligned}$$

   (c) (Figure 6.2) Two complex response surfaces.

FIGURE 6.2 Additional Examples of Response Functions.



10. **Meaning of Linear in General Linear Regression Model**

(a) It should be clear from the various examples that general linear regression model (6.7) is not restricted to linear response surfaces.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \qquad (6.7)$$

The term linear model refers to the fact that model (6.7) is linear in the parameters ; it does-not refer to the shape of the response surface .

(b) We say that a regression model is linear in the parameters when it can be written in the form:

$$Y_i = c_{i0}\beta_0 + c_{i1}\beta_1 + c_{i2}\beta_2 + \cdots c_{i,p-1}\beta_{p-1} + \varepsilon_i ,$$

where the terms c_{i0}, c_{i1}, etc., are coefficients involving the predictor variables .

(c) An example of a nonlinear regression model is the following:

$$Y_i = \beta_0 \exp(\beta_1 X_i) + \varepsilon_i$$

This is a nonlinear regression model because it cannot be expressed in the form of (6.17). (nonlinear regression models in Part III)

6.2 General Linear Regression Model in Matrix Terms

1. We now present the principal results for the general linear regression model (6.7) in matrix terms. The matrix notation may hide enormous computational complexities.

2. The actual computations will, in all but the very simplest cases, be done by computer.

3. Express general linear regression model (6.7):

$$Y_i = \underline{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i} \qquad (6.7)$$

in matrix terms:

$$\underline{\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}},$$

where

$$\mathbf{Y} = \underline{\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}}, \quad \mathbf{X} = \underline{\begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & & & \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix}},$$

$$\boldsymbol{\beta} = \underline{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}}, \quad \boldsymbol{\varepsilon} = \underline{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}},$$

4. $\boldsymbol{\varepsilon}$ is a vector of independent normal random variables with $\underline{E(\boldsymbol{\varepsilon}) = \mathbf{0}}$ and $\underline{\sigma^2(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}}$.

5. The random vector $\mathbf{Y}$ has expectation: $\underline{E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}}$, and the variance-covariance matrix of $\mathbf{Y}$ is the same as that of $\boldsymbol{\varepsilon}$: $\underline{\sigma^2(\mathbf{Y}) = \sigma^2\mathbf{I}}$.

6.3 Estimation of Regression Coefficients

1. The least squares criterion (1.8) is generalized as follows for general linear regression model (6.7):

$$Q = \underline{\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_{p-1}X_{i,p-1})^2} \tag{6.22}$$

2. The least squares estimators are those values of $\beta_0, \beta_1, \cdots, \beta_{p-1}$ that $\underline{\text{minimize } Q}$.

3. The least squares normal equations for the general linear regression model (6.19) are:

$$\underline{\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y}} \tag{6.24}$$

4. The least squares estimators are:

$$\mathbf{b} = \underline{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}} \tag{6.25}$$

5. The method of maximum likelihood leads to the same estimators for normal error regression model (6.19) as those obtained by the method of least squares in (6.25).

6. The likelihood function in (1.26) generalizes directly for multiple regression:

$$L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_{p-1}X_{i,p-1})^2\right\} \tag{6.26}$$

7. Maximizing this likelihood function with respect to $\beta_0, \beta_1, \cdots, \beta_{p-1}$ leads to the estimators in (6.25). These estimators are least squares and maximum likelihood estimators and have all the properties mentioned in Chapter 1: they are $\underline{\text{minimum variance unbiased}}$ $\underline{\text{consistent}}$, and $\underline{\text{sufficient}}$.

6.4 Fitted Values and Residuals

1. Let the vector of the fitted values $\hat{Y}_i$ be denoted by $\hat{\mathbf{Y}}$ and the vector of the residual terms $e_i = Y_i - \hat{Y}_i$ be denoted by $\mathbf{e}$:

$$\hat{\mathbf{Y}} = \underline{\begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix}}, \qquad \mathbf{e} = \underline{\begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}},$$

2. The fitted values: $\underline{\hat{\mathbf{Y}} = \mathbf{Xb}}$.

3. The vector of the fitted values $\hat{\mathbf{Y}}$ can be expressed in terms of the hat matrix $\mathbf{H}$ as follows:

$$\hat{\mathbf{Y}} = \underline{\mathbf{HY}}, \quad \text{where} \quad \underline{\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'} \qquad (6.30)$$

4. The residual terms: $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \underline{\mathbf{Y} - \mathbf{Xb}}$.

5. Similarly, the vector of residuals can be expressed: $\mathbf{e} = \underline{(\mathbf{I} - \mathbf{H})\mathbf{Y}}$.

6. The variance-covariance matrix of the residuals is: $\sigma^2(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H})$ which is estimated by:

$$s^2(\mathbf{e}) = \underline{MSE(\mathbf{I} - \mathbf{H})} \qquad (6.33)$$

6.5 Analysis of Variance Results

Sums of Squares and Mean Squares

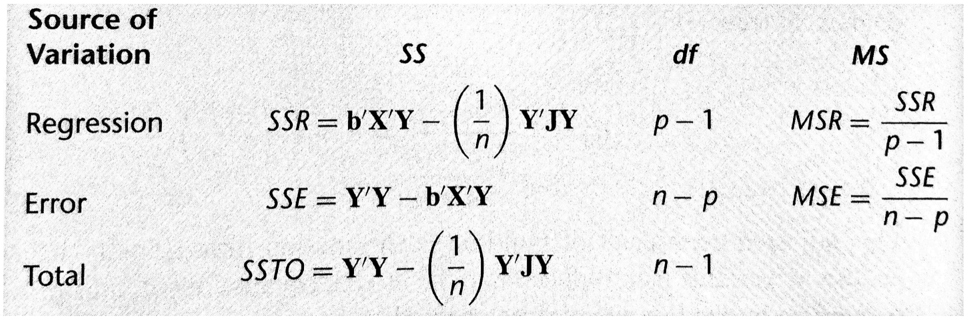
1. The sums of squares for the analysis of variance in matrix terms are, from (5.89):

$$\begin{aligned} SSTO &= \underline{\mathbf{Y}'\mathbf{Y} - \left(\frac{1}{n}\right)\mathbf{Y}'\mathbf{JY} = \mathbf{Y}'\left[\mathbf{I} - \left(\frac{1}{n}\right)\mathbf{J}\right]\mathbf{Y}} \\ SSE &= \underline{\mathbf{e}'\mathbf{e} = (\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb}) = \mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y} = \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y}} \\ SSR &= \underline{\mathbf{b}'\mathbf{X}'\mathbf{Y} - \left(\frac{1}{n}\right)\mathbf{Y}'\mathbf{JY} = \mathbf{Y}'\left[\mathbf{H} - \left(\frac{1}{n}\right)\mathbf{J}\right]\mathbf{Y}} \end{aligned}$$

where $\mathbf{J}$ is an $n \times n$ matrix of 1s defined in (5.18) and $\mathbf{H}$ is the hat matrix defined in (6.30a).

2. (Table 6.1) ANOVA Table for general linear regression:

TABLE 6.1 ANOVA Table for General Linear Regression Model (6.19).



Source of Variation	SS	df	MS
Regression	$SSR = \mathbf{b'X'Y} - \left(\frac{1}{n}\right)\mathbf{Y'JY}$	$p-1$	$MSR = \frac{SSR}{p-1}$
Error	$SSE = \mathbf{Y'Y} - \mathbf{b'X'Y}$	$n-p$	$MSE = \frac{SSE}{n-p}$
Total	$SSTO = \mathbf{Y'Y} - \left(\frac{1}{n}\right)\mathbf{Y'JY}$	$n-1$	

F Test for Regression Relation

1. To test whether there is a regression relation between the response variable Y and the set of X variables $X_1, \cdots, X_p$,

$$H_0 \;:\; \underline{\beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0}$$
$$H_a \;:\; \underline{\text{not all } \beta_k, (k = 1, \cdots, p-1) \text{ equal zero}}$$

2. The test statistic:

$$F^* = \underline{\frac{MSR}{MSE}}.$$

3. The decision rule to control the Type I error at α:

$$\underline{\text{If } F^* > F_{(1-\alpha;p-1,n-p)}, \text{ reject } H_0.}$$

Coefficient of Multiple Determination

1. The coefficient of multiple determination, denoted by R^2, is defined as

$$R^2 = \underline{\frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}} \qquad (6.40)$$

2. It measures the proportionate reduction of total variation in Y associated with the use of the set of X variables $X_1, \cdots, X_{p-1}$.

3. $0 \leq R^2 \leq 1$ assumes the value 0 when all $b_k = 0\ (k = 1, \cdots, p-1)$, and the value 1 when all Y observations fall directly on the fitted regression surface, ie., when $Y_i = \hat{Y}_i$ for all i.

4. Adding more X variables to the regression model can only increase R^2 and never reduce it, because SSE can never become larger with more X variables and $SSTO$ is always the same for a given set of responses.

5. Since R^2 usually can be made larger by including a larger number of predictor variables, it is sometimes suggested that a modified measure be used that adjusts for the number of X variables in the model.

6. The adjusted coefficient of multiple determination, denoted by R_a^2, adjusts R^2 by dividing each sum of squares by its associated degrees of freedom:

$$R_a^2 = 1 - \frac{SSE/(n-p)}{SSTO/(n-1)} = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE}{SSTO}$$

This adjusted coefficient of multiple determination may actually become smaller when another X variable is introduced into the model, because any decrease in SSE may be more than offset by the loss of a degree of freedom in the denominator $n-p$.

7. Comments A large value of R^2 does not necessarily imply that the fitted model is a useful one. For instance, observations may have been taken at only a few levels of the predictor variables. Despite a high R^2 in this case, the fitted model may not be useful if most predictions require extrapolations outside the region of observations. Again, even though R^2 is large, MSE may still be too large for inferences to be useful when high precision is required.

Coefficient of Multiple Correlation

1. The coefficient of multiple correlation R is the positive square root of

$$R = \sqrt{R^2}$$

6.6 Inferences about Regression Parameters

1. The least squares and maximum likelihood estimators in **b** are unbiased:

$$E\{\mathbf{b}\} = \boldsymbol{\beta} \qquad (6.44)$$

2. The variance-covariance matrix (dimension $p \times p$):

$$\sigma^2\{\mathbf{b}\} = \underline{\ \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\ } \qquad (6.46)$$

3. The estimated variance-covariance matrix (dimension $p \times p$):

$$s^2\{\mathbf{b}\} = \underline{\ MSE(\mathbf{X}'\mathbf{X})^{-1}\ } \qquad (6.48)$$

Interval Estimation of β_k

1. For the normal error regression model (6.19), we have:

$$\underline{\ \frac{b_k - \beta_k}{s\{b_k\}} \sim t_{(n-p)}\ }\ , \quad k = 0, 1, ..., p-1 \qquad (6.49)$$

2. The confidence limits for β_k with $1-\alpha$ confidence coefficient are:

$$\underline{\ b_k \pm t_{(1-\alpha/2;n-p)} s\{b_k\}\ } \qquad (6.50)$$

Tests for β_k

1. The test hypothesis:

$$\underline{\ H_0 : \beta_k = 0 \quad \text{against} \quad H_a : \beta_k \neq 0\ }$$

2. The test statistic:

$$\underline{\ t^* = \frac{b_k}{s\{b_k\}}\ }$$

3. The decision rule:

$$\underline{\ \text{If } |t^*| \geq t_{(1-\alpha/2;n-p)}, \quad \text{reject} \quad H_0\ }.$$

4. The __power__ of the t test can be obtained as explained in Chapter 2, with the degrees of freedom modified to $n-p$. As with simple linear regression, an __F test__ can also be conducted to determine whether or not $\beta_k = 0$ in multiple regression models. (details in Chapter 7).

Joint Inferences*

6.7 Estimation of Mean Response and Prediction of New Observation*

Interval Estimation of $E\{Y_h\}$

Confidence Region for Regression Surface

Simultaneous Confidence Intervals for Several Mean Responses

Prediction of New Observation $Y_{h(new)}$

Prediction of Mean of m New Observations at X_h

Predictions of g New Observations

Caution about Hidden Extrapolations

6.8 Diagnostics and Remedial Measures

1. Diagnostics play an important role in the development and evaluation of multiple regression models.

2. Most of the diagnostic procedures for SLR (Chapter 3) carry over directly to multiple regression.

3. Many specialized diagnostics and remedial procedures for multiple regression have also been developed (details in Chapters 10 and 11.)

Scatter Plot Matrix

1. *Univariate plots*: Box plots, sequence plots, stem-and-leaf plots, and dot plots for each of the predictor variables and for the response variable can provide helpful,

preliminary univariate information about these variables.

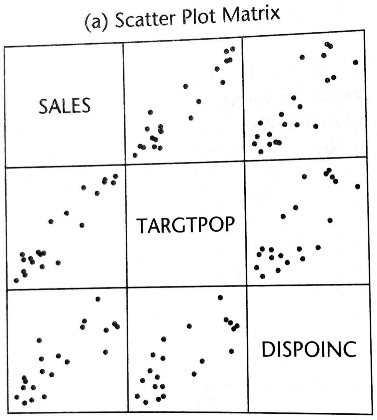
2. *Bivariate plots: Scatter plots*

   (a) Scatter plots of the response variable against each predictor variable can aid in determining the nature and strength of the bivariate relationships between each of the predictor variables and the response variable and in identifying gaps in the data points as well as outlying data points.

   (b) Scatter plots of each predictor variable against each of the other predictor variables are helpful for studying the bivariate relationships among the predictor variables and for finding gaps and detecting outliers.

3. *Multiivariate plots: Scatter plot matrix*

FIGURE 6.4 SYGRAPH Scatter Plot Matrix and Correlation Matrix—Dwaine Studios Example.



(b) Correlation Matrix

	SALES	TARGTPOP	DISPOINC
SALES	1.000	.945	.836
TARGTPOP		1.000	.781
DISPOINC			1.000

   (a) (Figure 6.4) the Y variable for anyone scatter plot is the name found in its row, and the X variable is the name found in its column.

   (b) The scatter plot matrix in Figure 6.4 shows in the first row the plots of Y (SALES) against X_1 (TARGETPOP) and X_2 (DISPOINC), of X_1 against Y and X_2 in the second row, and of X_2 against Y and X_1 in the third row. (These variables are described on page 236.)

   (c) Scatter plot matrix facilitates the study of the relationships among the variables by comparing the scatter plots within a row or a column.

4. A complement to the scatter plot matrix that may be useful at times is the correlation matrix. This matrix contains the coefficients of simple correlation $r_{Y1}, r_{Y2}, \cdots, r_{Y,p-1}$

between Y and each of the predictor variables $X_i, i = 1, \cdots, p-1$, as well as all of the coefficients of simple correlation among the predictor variables: r_{12} between X_1 and X_2, r_{13} between X_1 and X_3, etc.

5. Note that the correlation matrix is symmetric and that its main diagonal contains 1s because the coefficient of correlation between a variable and itself is 1.

Three-Dimensional Scatter Plots

1. Some interactive statistics packages provide three-dimensional scatter plots or point clouds, and permit spinning of these plots to enable the viewer to see the point cloud from different perspectives or patterns. (Figure 6.6)

Residual Plots

1. *plot($e_i \sim \hat{Y}_i$)*: A plot of the residuals against the fitted values is useful for assessing the appropriateness of the multiple regression function and the constancy of the variance of the error terms, as well as for providing information about outliers, just as for simple linear regression.

2. *plot($e_i \sim$ time)*: A plot of the residuals against time or against some other sequence can provide diagnostic information about possible correlations between the error terms in multiple regression.

3. *boxplot(e_i), qqnorm(e_i)*: Box plots and normal probability plots of the residuals are useful for examining whether the error terms are reasonably normally distributed.

4. *plot($e_i \sim X_i$)*: The plot of the residuals against each of the predictor variables can provide further information about the adequacy of the regression function with respect to that predictor variable (e.g., whether a curvature effect is required for that variable) and about possible variation in the magnitude of the error variance in relation to that predictor variable.

5. *plot($e_i \sim X_{omit}$)*: Plot the residuals against important predictor variables that were omitted from the model, to see if the omitted variables have substantial ad-

ditional effects on the response variable that have not yet been recognized in the regression model.

6. *plot(*$e_i \sim X_iX_j$*)*: Plot the residuals against interaction terms for potential interaction effects not included in the regression model, such as against X_1X_2, X_1X_3, and X_2X_3, to see whether some or all of these interaction terms are required in the model.

7. *plot(*$|e_i| \sim \hat{Y}_i$*), plot(*$e_i^2 \sim \hat{Y}_i$*)*: A plot of the absolute residuals or the squared residuals against the fitted values is useful for examining the constancy of the variance of the error terms.

8. *plot(*$|e_i| \sim X_i$*), plot(*$e_i^2 \sim X_i$*)*: If nonconstancy is detected, a plot of the absolute residuals or the squared residuals against each of the predictor variables may identify one or several of the predictor variables to which the magnitude of the error variability is related.

Correlation Test for Normality*

1. The correlation test for normality described in Chapter 3 carries forward directly to multiple regression.

Brown-Forsythe Test for Constancy of Error Variance

1. The Brown-Forsythe test statistic (3.9) for assessing the constancy of the error variance can be used readily in multiple regression when the error variance increases or decreases with one of the predictor variables.

2. To conduct the Brown-Forsythe test, we divide the data set into two groups, as for simple linear regression, where one group consists of cases where the level of the predictor variable is relatively low and the other group consists of cases where the level of the predictor variable is relatively high.

3. The Brown-Forsythe test then proceeds as for simple linear regression.

Breusch-Pagan Test for Constancy of Error Variance*

F Test for Lack of Fit

1. The lack of fit F test (Chapter 3) for SLR can be carried over to test whether the multiple regression response function:

$$\underline{E[Y] = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}}$$

is an appropriate response surface.

2. Repeat observations in multiple regression are replicate observations on Y corresponding to levels of each of the X variables that are constant from trial to trial.

3. With two predictor variables, repeat observations require that X_1 and X_2 each remain at given levels from trial to trial.

4. Once the ANOVA table (Table 6.1), has been obtained, SSE is decomposed into the pure error sum of squares (SSPE) and the lack of fit sum of squares (SSLF).

5. SSPE is obtained by first calculating for each replicate group the sum of squared deviations of the Y observations around the group mean, where a replicate group has the same values for each of the X variables.

6. Let c denote the number of groups with distinct sets of levels for the X variables, and let the mean of the Y observations for the jth group be denoted by $\bar{Y}_j$. Then the pure error sum of squares is $\sum_j \sum_i (Y_{ij} - \bar{Y}_j)$. The lack of fit sum of squares $SSLF$ equals the difference $SSE - SSPE$.

7. *Test hypothesis*:

$$H_0 : E\{Y\} = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}$$

$$H_a : E\{Y\} \neq \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}$$

8. *Test statistic*:

$$F^* = \frac{SSLF}{c-p} \div \frac{SSPE}{n-c} = \frac{MSLF}{MSPE}$$

9. *Decision rule*:

$$\text{If } F^* > F_{(1-\alpha; c-p, n-p)}, \text{ reject } H_0.$$

Remedial Measures

1. The remedial measures described in Chapter 3 are also applicable to multiple regression.

2. When a more complex model is required to recognize curvature or interaction effects, the multiple regression model can be expanded to include these effects.

3. Transformations on the response variable Y may be helpful when the distributions of the error terms are quite skewed and the variance of the error terms is not constant.

4. Transformations of some of the predictor variables may be helpful when the effects, of these variables are curvilinear.

5. Transformations on Y and/or the predictor variables may be helpful in eliminating or substantially reducing interaction effects.

6. The usefulness of potential transformations needs to be examined by means of residual plots and other diagnostic tools to determine whether the multiple regression model for the transformed data is appropriate.

7. Box-Cox Transformations is also applicable to multiple regression models.

6.9 An Example - Multiple Regression with Two Predictor Variables

Setting

1. (Figure 6.5a) Dwaine Studios, Inc., operates portrait studios in 21 cities ($n = 21$) of medium size. These studios specialize in portraits of children. The company is considering an expansion into other cities of medium size and wishes to investigate whether sales (Y or SALES, in thousands of dollars) in a community can be predicted from the number of persons aged 16 or younger in the community (X_1 or TARGTPOP for target population) and the per capita disposable (平均每人可支配所得) personal income in the community (X_2 or DISPOINC for disposable income in thousands of dollars).

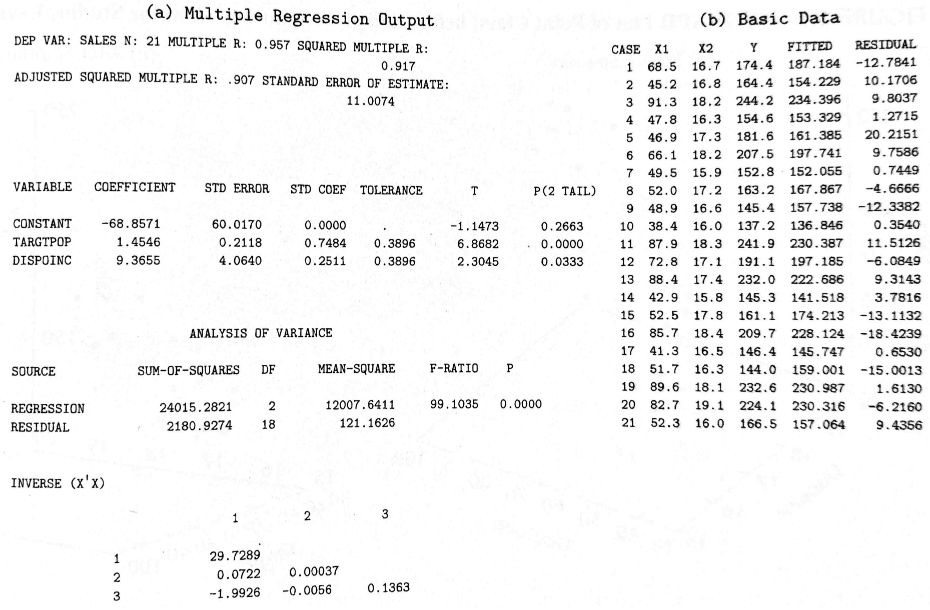


FIGURE 6.5 SYSTAT Multiple Regression Output and Basic Data—Dwaine Studios Example.

(a) Multiple Regression Output

```
DEP VAR: SALES N: 21 MULTIPLE R: 0.957 SQUARED MULTIPLE R:
                                                   0.917
ADJUSTED SQUARED MULTIPLE R: .907 STANDARD ERROR OF ESTIMATE:
                                              11.0074



VARIABLE   COEFFICIENT    STD ERROR   STD COEF  TOLERANCE       T        P(2 TAIL)

CONSTANT     -68.8571       60.0170     0.0000      .        -1.1473      0.2663
TARGTPOP       1.4546        0.2118     0.7484    0.3896      6.8682      0.0000
DISPOINC       9.3655        4.0640     0.2511    0.3896      2.3045      0.0333



                      ANALYSIS OF VARIANCE

SOURCE           SUM-OF-SQUARES   DF     MEAN-SQUARE     F-RATIO     P

REGRESSION          24015.2821     2      12007.6411     99.1035   0.0000
RESIDUAL             2180.9274    18        121.1626


INVERSE (X'X)

                      1         2          3

         1       29.7289
         2        0.0722    0.00037
         3       -1.9926   -0.0056     0.1363
```

(b) Basic Data

CASE	X1	X2	Y	FITTED	RESIDUAL
1	68.5	16.7	174.4	187.184	-12.7841
2	45.2	16.8	164.4	154.229	10.1706
3	91.3	18.2	244.2	234.396	9.8037
4	47.8	16.3	154.6	153.329	1.2715
5	46.9	17.3	181.6	161.385	20.2151
6	66.1	18.2	207.5	197.741	9.7586
7	49.5	15.9	152.8	152.055	0.7449
8	52.0	17.2	163.2	167.867	-4.6666
9	48.9	16.6	145.4	157.738	-12.3382
10	38.4	16.0	137.2	136.846	0.3540
11	87.9	18.3	241.9	230.387	11.5126
12	72.8	17.1	191.1	197.185	-6.0849
13	88.4	17.4	232.0	222.686	9.3143
14	42.9	15.8	145.3	141.518	3.7816
15	52.5	17.8	161.1	174.213	-13.1132
16	85.7	18.4	209.7	228.124	-18.4239
17	41.3	16.5	146.4	145.747	0.6530
18	51.7	16.3	144.0	159.001	-15.0013
19	89.6	18.1	232.6	230.987	1.6130
20	82.7	19.1	224.1	230.316	-6.2160
21	52.3	16.0	166.5	157.064	9.4356

2. The first-order regression model:

$$\underline{\quad Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \quad}$$

with normal error terms is expected to be appropriate, on the basis of the scatter plot matrix in Figure 6.4a.

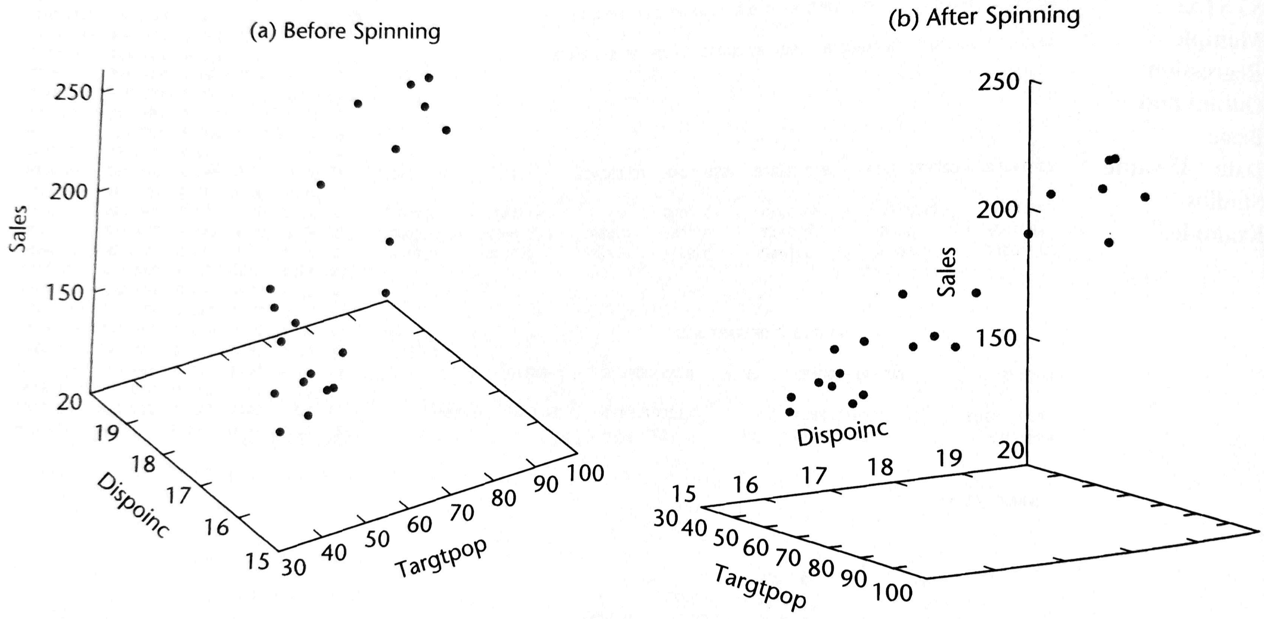
3. Note the __linear relation__ between target population and sales and between disposable income and sales.

4. Also note that there is __more scatter__ between disposable income and sales relationship.

5. Finally note that there is also some __linear__ relationship between the two predictor variables.

6. (Figure 6.6) A 3D plot of the point cloud supports the tentative conclusion that a response plane may be a reasonable regression function to utilize here.

FIGURE 6.6 SYGRAPH Plot of Point Cloud before and after Spinning—Dwaine Studios Example.



Basic Calculations

1. The X and $\mathbf{Y}$ matrices for the Dwaine Studios example:

$$\mathbf{X} = \begin{bmatrix} 1 & 68.5 & 16.7 \\ 1 & 45.2 & 16.8 \\ \vdots & \vdots & \vdots \\ 1 & 52.3 & 16.0 \end{bmatrix} \qquad \mathbf{Y} = \begin{bmatrix} 174.4 \\ 164.4 \\ \vdots \\ 166.5 \end{bmatrix}$$

2.

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 29.7289 & 0.0722 & -1.9926 \\ 0.0722 & 0.00037 & -0.0056 \\ -1.9926 & -0.0056 & 0.1363 \end{bmatrix}$$

3.

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} 3.820 \\ 249.643 \\ 66.073 \end{bmatrix}$$

Estimated Regression Function

1. The least squares estimates **b** are readily obtained by

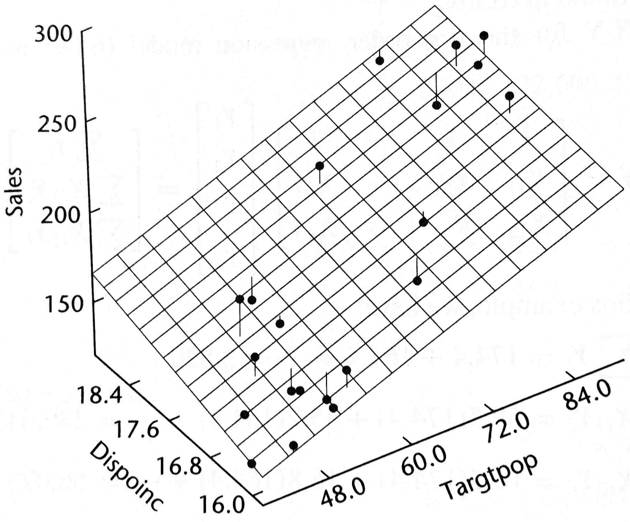
$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{bmatrix} -68.857 \\ 1.455 \\ 9.366 \end{bmatrix}$$

2. The estimated regression function is:

$$\hat{Y} = -68.857 + 1.455X_1 + 9.366X_2$$

3. (Figure 6.7) A 3D plot of the estimated regression function, with the responses super-imposed. The residuals are represented by the small vertical lines connecting the responses to the estimated regression surface.

FIGURE 6.7 **S-Plus Plot of Estimated Regression Surface—Dwaine Studios Example.**



4. This estimated regression function indicates that mean sales are expected to increase by 1.455 thousand dollars when the target population increases by 1 thousand persons aged 16 years or younger, holding per capita disposable personal income constant, and that mean sales are expected to increase by 9.366 thousand dollars when per capita income increases by 1 thousand dollars, holding the target population constant.

5. (Figure 6.5a) Software output for the Dwaine Studios example.

Fitted Values and Residuals

1. The fitted values

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b} = \begin{bmatrix} 187.2 \\ 154.2 \\ \vdots \\ 157.1 \end{bmatrix}$$

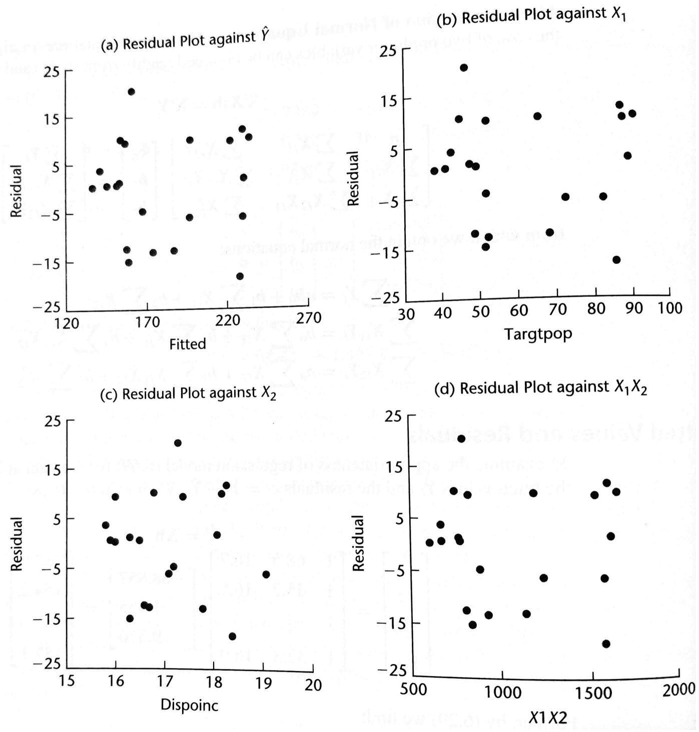
2. The residuals

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \begin{bmatrix} -12.8 \\ 10.2 \\ \vdots \\ 9.4 \end{bmatrix}$$

Analysis of Appropriateness of Model

1. (Figure 6.8a) Begin analysis of the appropriateness of regression model by considering the plot of the residuals e_i against the fitted values Y in Figure 6.8a. This plot does not suggest any systematic deviations from the response plane nor that the variance of the error terms varies with the level of $\hat{Y}$.

FIGURE 6.8 **SYGRAPH Diagnostic Plots—Dwaine Studios Example.**



2. (Figures 6.8b, 6.8c) Plots of the residuals e against X_1 and X_2 are entirely consistent with the conclusions of good fit by the response function and constant variance of the error terms.

3. If a plot of the residuals e against the interaction term X_1X_2 shows a systematic pattern, that means an interaction effect may be present, so that a response function of the type

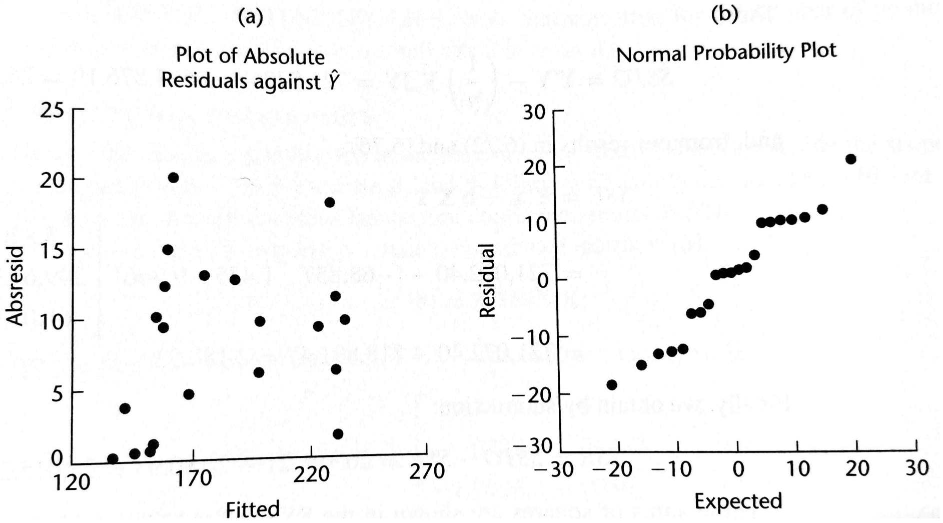
$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

might be more appropriate.

4. (Figure 6.8d) Plot does not exhibit any systematic pattern; hence, no interaction effects reflected by the model term X_1X_2 appear to be present.

5. (Figure 6.9a) A plot of the absolute residuals against the fitted values. There is no indication of nonconstancy of the error variance.

FIGURE 6.9 Additional Diagnostic Plots—Dwaine Studios Example.



6. (Figure 6.9b) A normal probability plot of the residuals shows a moderately linear pattern.

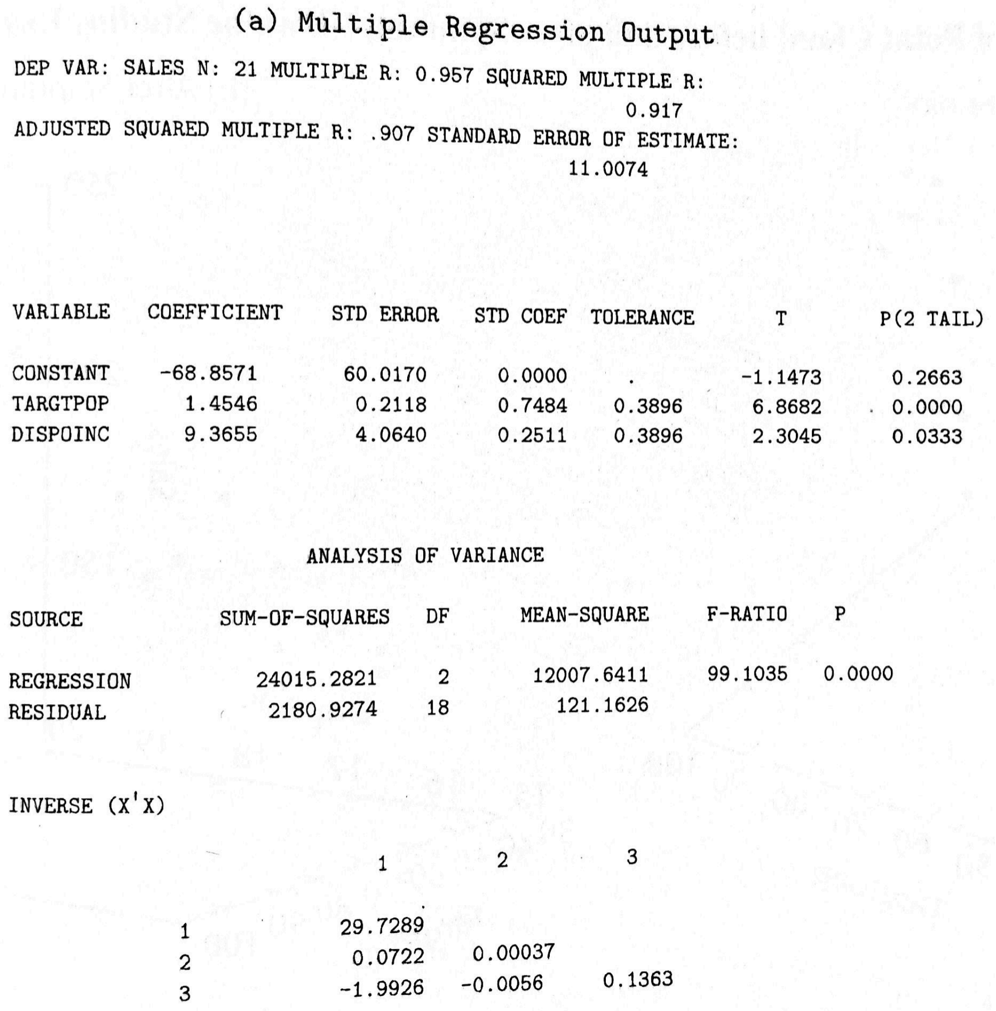
7. The coefficient of correlation between the ordered residuals and their expected values under normality is 0.980. This high value helps to confirm the reasonableness of the conclusion that the error terms are fairly normally distributed.

8. Since the Dwaine Studios data are cross-sectional and do not involve a time sequence, a time sequence plot is not relevant here. Thus, all of the diagnostics support the use of regression model (6.69) for the Dwaine Studios example.

Analysis of Variance

1. To test whether sales are related to target population and per capita disposable income, we require the ANOVA table.

FIGURE 6.5 SYSTAT Multiple Regression Output and Basic Data—Dwaine Studios Example.



(a) Multiple Regression Output

```
DEP VAR: SALES N: 21 MULTIPLE R: 0.957 SQUARED MULTIPLE R:
                                                      0.917
ADJUSTED SQUARED MULTIPLE R: .907 STANDARD ERROR OF ESTIMATE:
                                                   11.0074


VARIABLE    COEFFICIENT     STD ERROR   STD COEF  TOLERANCE       T       P(2 TAIL)

CONSTANT     -68.8571        60.0170     0.0000       .        -1.1473     0.2663
TARGTPOP       1.4546         0.2118     0.7484    0.3896       6.8682     0.0000
DISPOINC       9.3655         4.0640     0.2511    0.3896       2.3045     0.0333


                        ANALYSIS OF VARIANCE

SOURCE           SUM-OF-SQUARES   DF      MEAN-SQUARE    F-RATIO    P

REGRESSION          24015.2821     2      12007.6411     99.1035   0.0000
RESIDUAL             2180.9274    18        121.1626


INVERSE (X'X)

                          1          2          3

            1          29.7289
            2           0.0722    0.00037
            3          -1.9926   -0.0056     0.1363
```

2. **Test of Regression Relation**. To test whether sales are related to target population and per capita disposable income:

$$H_0 : \beta_1 = 0 \text{ and } \beta_2 = 0$$

$$H_a: \text{ not both } \beta_1 \text{ and } \beta_2 \text{ equal zero}$$

Test statistic:

$$F^* = 99.1$$

For $\alpha = 0.05$, we require $F_{(0.95;2.18)} = 3.55$. Since $F^* = 99.1 > 3.55$, we conclude H_a (reject H_0), that sales are related to target population and per capita disposable income. The P-value for this test is 0.0000.

3. **Coefficient of Multiple Determination**.

$$R^2 = 0.917$$

Thus, when the two predictor variables, target population and per capita disposable income, are considered, the variation in sales is reduced by 91.7 percent. The adjusted coefficient of multiple determination $R^2 = 0.907$.

Estimation of Regression Parameters*

Estimation of Mean Response*

Prediction Limits for New Observations*

😍 TA Class

- **Problems**: 6.5 (a-d, f), 6.6 (a, b), 6.9, 6.10 (a-d)
- **Exercises**: 6.22

"不要畏懼失敗，你應該要擔心沒有機會嘗試，但你有的是機會嘗試!"
"Don't fear failure. Be afraid of not having the chance, you have the chance!"
— *汽車總動員 3: 閃電再起 (Cars 3, 2017)*

Regression Analysis (I)

Kutner's Applied Linear Statistical Models (5/E)

Chapter 7: Multiple Regression (II)

Thursday 09:10-12:00, 商館 260205

Han-Ming Wu

Department of Statistics, National Chengchi University

`http://www.hmwu.idv.tw`

Overview

1. Some specialized topics that are unique to multiple regression: (1) extra sums of squares, (2) the standardized version of the multiple regression model, and (3) multicollinearity.

7.1 Extra Sums of Squares

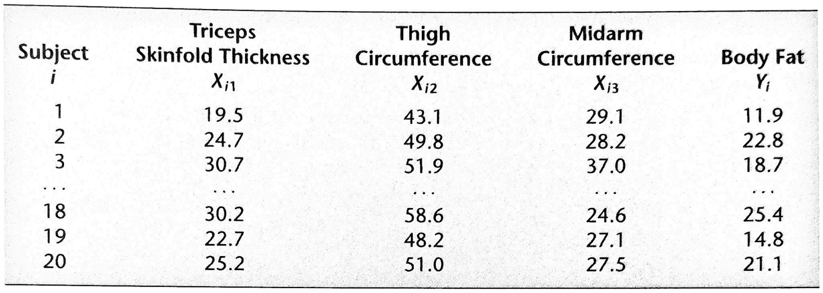
Basic Ideas

1. An extra sum of squares measures the marginal reduction in the error sum of squares when one or several predictor variables are added to the regression model, given that other predictor variables are already in the model.

2. Equivalently, one can view an extra sum of squares as measuring the marginal increase in the regression sum of squares when one or several predictor variables are added to the regression model.

3. Example (Table 7.1) A portion of the data for a study of the relation of amount of body fat (Y) to several possible predictor variables, based on a sample of 20 healthy females $25 - 34$ years old. The possible predictor variables are triceps skinfold thickness (X_1)(三頭肌皮下脂肪厚度), thigh circumference (X_2)(大腿圍), and midarm circumference (X_3) (中臂圍).

TABLE 7.1 Basic Data—Body Fat Example.



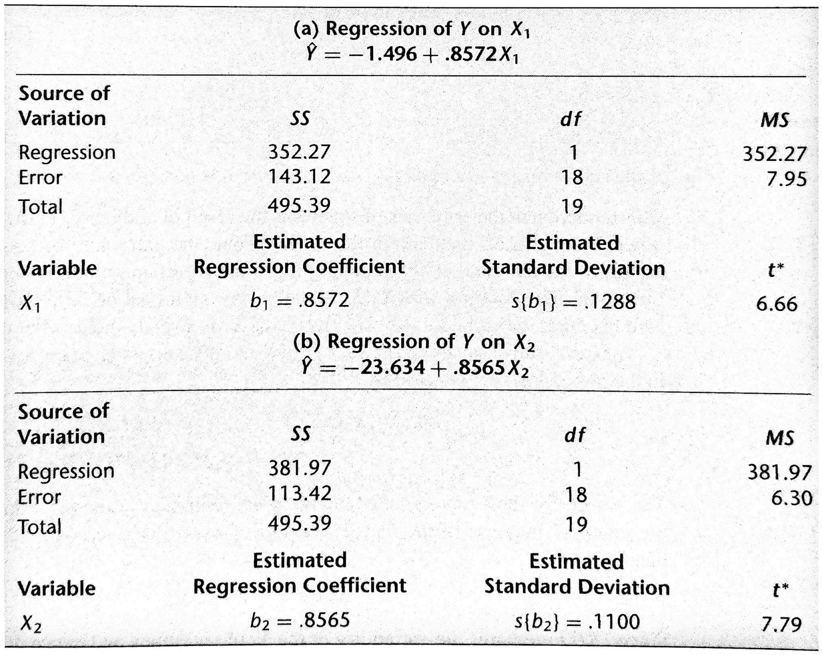
Subject i	Triceps Skinfold Thickness X_{i1}	Thigh Circumference X_{i2}	Midarm Circumference X_{i3}	Body Fat Y_i
1	19.5	43.1	29.1	11.9
2	24.7	49.8	28.2	22.8
3	30.7	51.9	37.0	18.7
···	···	···	···	···
18	30.2	58.6	24.6	25.4
19	22.7	48.2	27.1	14.8
20	25.2	51.0	27.5	21.1

4. *Background and goal*: The amount of body fat in Table 7.1 for each of the 20 persons was obtained by a cumbersome and expensive procedure requiring the immersion of the person in water. It would therefore be very helpful if a regression model with some or all of these predictor variables could provide reliable estimates of the amount of body fat since the measurements needed for the predictor variables are easy to obtain.

5. (Table 7.2) Conduct four regression results when body fat (Y) is regressed on triceps skinfold thickness (X_1) alone, (2) on thigh circumference (X_2) alone, (3) on X_1, and X_2 only, and (4) on all three predictor variables. The total sum of squares is $SSTO = 495.39$.

   (a) (Table 7.2a) The regression sum of squares when X_1, only is in the model is, $SSR(X_1) = 352.27$. The error sum of squares for this model is $SSE(X_1) = 143.12$.

TABLE 7.2 Regression Results for Several Fitted Models—Body Fat Example.



(a) Regression of Y on X_1

$\hat{Y} = -1.496 + .8572X_1$

Source of Variation	SS	df	MS
Regression	352.27	1	352.27
Error	143.12	18	7.95
Total	495.39	19	

Variable	Estimated Regression Coefficient	Estimated Standard Deviation	t^*
X_1	$b_1 = .8572$	$s\{b_1\} = .1288$	6.66

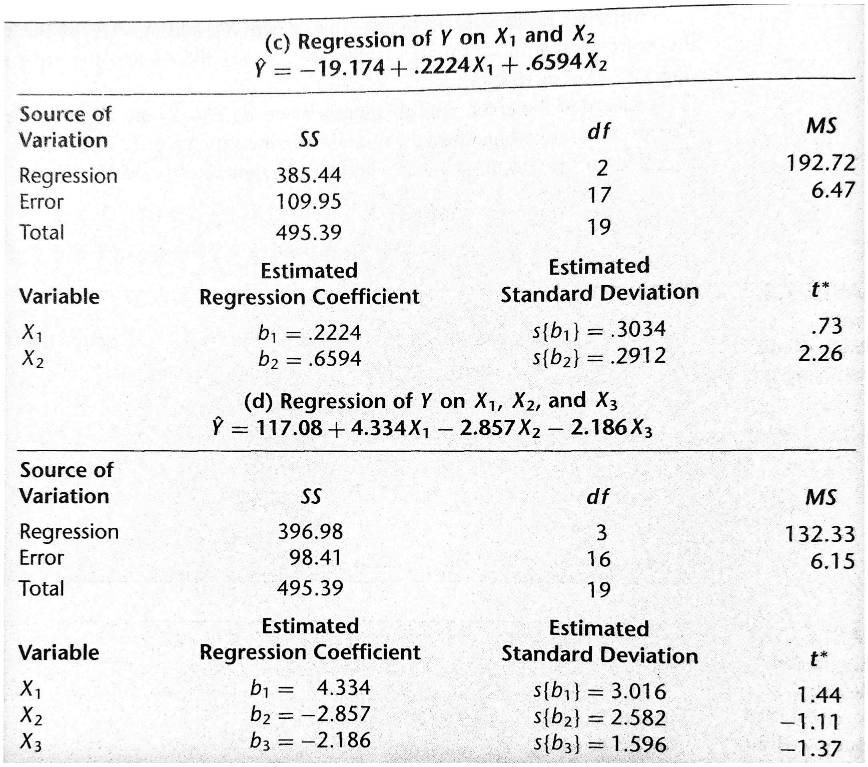
(b) Regression of Y on X_2

$\hat{Y} = -23.634 + .8565X_2$

Source of Variation	SS	df	MS
Regression	381.97	1	381.97
Error	113.42	18	6.30
Total	495.39	19	

Variable	Estimated Regression Coefficient	Estimated Standard Deviation	t^*
X_2	$b_2 = .8565$	$s\{b_2\} = .1100$	7.79

TABLE 7.2
(Continued).



(c) Regression of Y on X_1 and X_2
$\hat{Y} = -19.174 + .2224X_1 + .6594X_2$

Source of Variation	SS	df	MS
Regression	385.44	2	192.72
Error	109.95	17	6.47
Total	495.39	19	

Variable	Estimated Regression Coefficient	Estimated Standard Deviation	t^*
X_1	$b_1 = .2224$	$s\{b_1\} = .3034$	.73
X_2	$b_2 = .6594$	$s\{b_2\} = .2912$	2.26

(d) Regression of Y on X_1, X_2, and X_3
$\hat{Y} = 117.08 + 4.334X_1 - 2.857X_2 - 2.186X_3$

Source of Variation	SS	df	MS
Regression	396.98	3	132.33
Error	98.41	16	6.15
Total	495.39	19	

Variable	Estimated Regression Coefficient	Estimated Standard Deviation	t^*
X_1	$b_1 = \ \ 4.334$	$s\{b_1\} = 3.016$	1.44
X_2	$b_2 = -2.857$	$s\{b_2\} = 2.582$	−1.11
X_3	$b_3 = -2.186$	$s\{b_3\} = 1.596$	−1.37

(b) (Table 7.2c) When X_1 and X_2 are in the regression model, the regression sum of squares is $SSR(X_1, X_2) = 385.44$ and the error sum of squares is $SSE(X_1, X_2) = 109.95$.

(c) Notice that the error sum of squares when X_1 and X_2 are in the model, $SSE(X_1, X_2) = 109.95$, is smaller than when the model contains only X_1, $SSE(X_1) = 143.12$.

(d) The difference is called an extra sum of squares and will be denoted by $SSR(X_2|X_1)$:

$$\begin{aligned}
SSR(X_2|X_1) &= SSR(X_1, X_2) - SSR(X_1) \\
&= 385.44 - 352.27 = 33.17 \\
&= (SSTO - SSE(X_1, X_2)) - (SSTO - SSE(X_1)) \\
&= SSE(X_1) - SSE(X_1, X_2) \\
&= 143.12 - 109.95 = 33.17
\end{aligned}$$

This reduction in the error sum of squares is the result of adding X_2 to the regression model when X_1, is already included in the model.

(e) Thus, the extra sum of squares $SSR(X_2|X_1)$ measures the marginal effect (additional or extra reduction) of adding X_2 to the regression model when X_1, is already in the model.

(f) The reason for the equivalence of the marginal reduction in the error sum of squares and the marginal increase in the regression sum of squares is the basic analysis of variance identity:

$$SSTO = SSR + SSE$$

Since SSTO measures the variability of the Y_i observations and hence does not depend on the regression model fitted, any reduction in SSE implies an identical increase in SSR.

6. (Tables 7.2c, 7.2d) We can consider other extra sums of squares, such as the marginal effect of adding X_3 to the regression model when X_1, and X_2 are already in the model.

$$SSR(X_3|X_1, X_2) = SSE(X_1, X_2) - SSE(X_1, X_2, X_3) = 109.95 - 98.41 = 11.54$$

or, equivalently:

$$SSR(X_3|X_1, X_2) = SSR(X_1, X_2, X_3) - SSR(X_1, X_2) = 396.98 - 385.44 = 11.54.$$

7. (table 7.2a, 7.2d) We can even consider the marginal effect of adding several variables, such as adding both X_2 and X_3 to the regression model already containing X_1.

$$SSR(X_2, X_3|X_1) = SSE(X_1) - SSE(X_1, X_2, X_3) = 143.12 - 98.41 = 44.71$$

or, equivalently:

$$SSR(X_2, X_3|X_1) = SSR(X_1, X_2, X_3) - SSR(X_1) = 396.98 - 352.27 = 44.71$$

Definitions

1. An extra sum of squares always involves the difference between the error sum of squares for the regression model containing the X variable(s) already in the model and the error sum of squares for the regression model containing both the original X variable(s) and the new X variable(s).

2. Equivalently, an extra sum of squares involves the difference between the two corresponding regression sums of squares.

3. Thus, we define:

$$SSR(X_1|X_2) = \underline{SSE(X_2) - SSE(X_1, X_2)} \qquad (7.1a)$$

or, equivalently:

$$SSR(X_1|X_2) = \underline{SSR(X_1, X_2) - SSR(X_2)} \qquad (7.1b)$$

4. If X_2 is the extra variable, We define:

$$SSR(X_2|X_1) = \underline{SSE(X_1) - SSE(X_1, X_2)} \qquad (7.2a)$$

or, equivalently:

$$SSR(X_2|X_1) = \underline{SSR(X_1, X_2) - SSR(X_1)} \qquad (7.2b)$$

5. Extensions for three or more variables are straightforward:

$$SSR(X_3|X_1, X_2) = \underline{SSE(X_1, X_2) - SSE(X_1, X_2, X_3)} \qquad (7.3a)$$

or:

$$SSR(X_3|X_1, X_2) = \underline{SSR(X_1, X_2, X_3) - SSR(X_1, X_2)} \qquad (7.4b)$$

and

$$SSR(X_2, X_3|X_1) = \underline{SSE(X_1) - SSE(X_1, X_2, X_3)} \qquad (7.4a)$$

or:

$$SSR(X_2, X_3|X_1) = \underline{SSR(X_1, X_2, X_3) - SSR(X_1)} \qquad (7.4b)$$

Decomposition of SSR into Extra Sums of Squares

1. In multiple regression, we can obtain a variety of decompositions of SSR into extra sums of squares.

2. Consider the multiple regression model with two X variables:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i, \ i = 1, \cdot, n$$

3. Begin with the identity for X_1:

$$\underline{SSTO = SSR(X_1) + SSE(X_1)} \tag{7.5}$$

when X_1 is the X variable in th model. Replacing $SSE(X_1)$ by its equivalent in (7.2a): $\underline{SSR(X_2|X_1) = SSE(X_1) - SSE(X_1, X_2)}$, we obtain:

$$SSTO = \underline{SSR(X_1) + SSR(X_2|X_1) + SSE(X_1, X_2)} \tag{7.6}$$

4. Use the same identity for multiple regression with two X variables as in (7.5) for a single X variable:

$$SSTO = \underline{SSR(X_1, X_2) + SSE(X_1, X_2)} \tag{7.7}$$

Solving (7.7) for $SSE(X_1, X_2)$ and using this expression in (7.6) lead to:

$$\underline{SSR(X_1, X_2) = SSR(X_1) + SSR(X_2|X_1)} \tag{7.8}$$

5. We have decomposed $SSR(X_1, X_2)$ into two marginal components:

   (a) $\underline{SSR(X_1)}$: measuring the contribution by including X_1 alone in the model.

   (b) $\underline{SSR(X_2|X_1)}$: measuring the additional contribution when X_2 is included, given that X_1 is already in the model.

6. The order of the X variables is arbitrary:

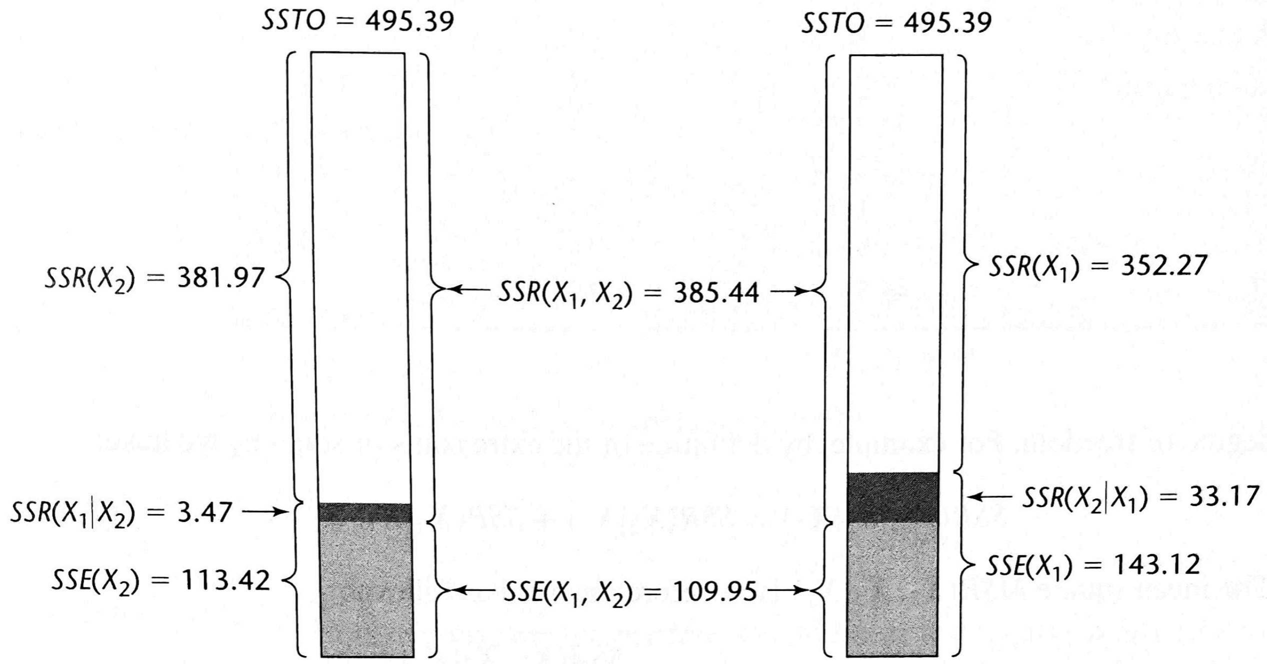
$$SSR(X_1, X_2) = \underline{SSR(X_2) + SSR(X_1|X_2)} \tag{7.9}$$

7. $\boxed{\text{Example}}$ Body Fat Example

   (a) A sample of $n = 20$ healthy females $25 - 34$ years old; Y: amount of body fat; X_1: triceps skinfold thickness; X_2: thigh circumference; X_3: midarm circumference.

(b) (Figure 7.1): The extra sum of squares can be viewed either as a reduction in SSE or as an increase in SSR when the second predictor variable is added to the regression model.

FIGURE 7.1 Schematic Representation of Extra Sums of Squares—Body Fat Example.



8. When the regression model contains three X variables, a variety of decompositions of $SSR(X_1, X_2, X_3)$ can be obtained. We illustrate three of these:

$$SSR(X_1, X_2, X_3) = SSR(X_1) + SSR(X_2|X_1) + SSR(X_3|X_1, X_2) \quad (7.10a)$$

$$SSR(X_1, X_2, X_3) = SSR(X_2) + SSR(X_3|X_2) + SSR(X_1|X_2, X_3) \quad (7.10b)$$

$$SSR(X_1, X_2, X_3) = SSR(X_1) + SSR(X_2, X_3|X_1) \quad (7.10e)$$

9. The number of possible decompositions becomes vast as the number of X variables in the regression model increases.

ANOVA Table Containing Decomposition of SSR

1. (Table 7.3, 7.4) ANOVA tables can be constructed containing decompositions of the regression sum of squares into extra sums of squares.

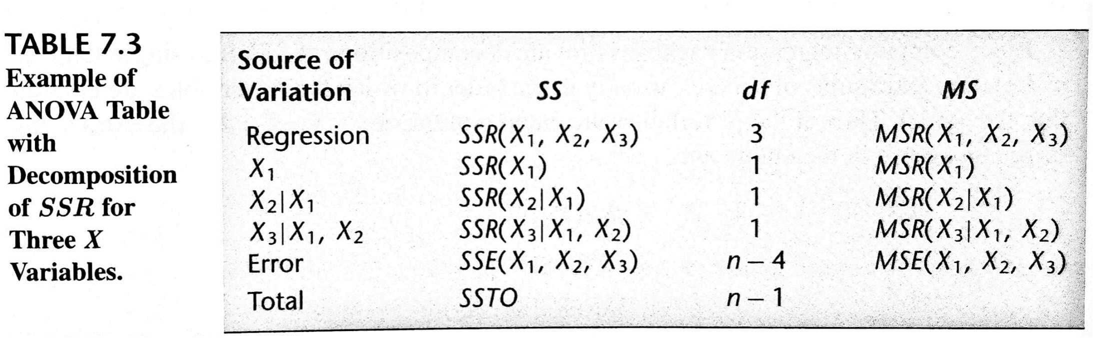


TABLE 7.3 Example of ANOVA Table with Decomposition of *SSR* for Three *X* Variables.

Source of Variation	*SS*	*df*	*MS*
Regression	$SSR(X_1, X_2, X_3)$	3	$MSR(X_1, X_2, X_3)$
X_1	$SSR(X_1)$	1	$MSR(X_1)$
$X_2 \mid X_1$	$SSR(X_2 \mid X_1)$	1	$MSR(X_2 \mid X_1)$
$X_3 \mid X_1, X_2$	$SSR(X_3 \mid X_1, X_2)$	1	$MSR(X_3 \mid X_1, X_2)$
Error	$SSE(X_1, X_2, X_3)$	$n-4$	$MSE(X_1, X_2, X_3)$
Total	$SSTO$	$n-1$	

2. Note that each extra sum of squares involving a single extra X variable has associated with it one degree of freedom.

3. Extra sums of squares involving two extra X variables, such as $SSR(X_2, X_3|X_1)$, have two degrees of freedom associated with them: an extra sum of squares as a sum of two extra sums of squares, each associated with one degree of freedom.

4. Many computer regression packages provide decompositions of SSR into single-degree-of-freedom extra sums of squares, usually in the order in which the X variables are entered into the model.

5. If the X variables are entered in the order X_1, X_2, X_3, the extra sums of squares given in the output are:

$$SSR(X_1), \quad SSR(X_2|X_1) \quad SSR(X_3|X_1, X_2)$$

6. If an extra sum of squares involving several extra X variables is desired, it can be obtained by summing appropriate single-degree-of-freedom extra sums of squares. For instance, to obtain $SSR(X_2, X_3|X_1)$:

$$SSR(X_2, X_3|X_1) = SSR(X_2|X_1) + SSR(X_3|X_1, X_2).$$

7. The reason why extra sums of squares are of interest is that they occur in a variety of tests about regression coefficients where the question of concern is whether certain X variables can be dropped from the regression model.

7.2 Uses of Extra Sums of Squares in Tests for Regression Coefficients

Test whether a Single $\beta_k = 0$

1. Test whether the term $\beta_k X_k$ can be dropped from a multiple regression model,

$$H_0 : \underline{\ \beta_k = 0\ } \qquad H_a : \underline{\ \beta_k \neq 0\ },$$

the test statistic: $\underline{\ t^* = \dfrac{b_k}{s(b_k)}\ }$ is appropriate for this test.

2. Use the general linear test approach : consider the first-order regression model with three predictor variables:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i \qquad \text{Full model} \qquad (7.12)$$

To test the alternatives:

$$H_0 : \beta_3 = 0 \quad H_a : \beta_3 \neq 0. \qquad (7.13)$$

3. The error sum of squares $SSE(F)$ for the full model:

$$SSE(F) = \underline{\ SSE(X_1, X_2, X_3)\ }, \qquad df_F = n - 4.$$

4. (*Reduced Model*) The reduced model when H_0 in (7.13) holds:

$$\underline{\ Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i\ } \qquad \text{Reduced model} \qquad (7.14)$$

The error sum of squares $SSE(E)$ for the reduced model:

$$SSE(R) = \underline{\ SSE(X_1, X_2)\ }, \qquad df_R = n - 3.$$

5. The general linear test statistic:

$$\begin{aligned}
F^* &= \frac{SSE(R) - SSE(F)}{df_R - df_F} \div \frac{SSE(F)}{df_F} \\
&= \frac{SSE(X_1, X_2) - SSE(X_1, X_2, X_3)}{(n-3) - (n-4)} \div \frac{SSE(X_1, X_2, X_3)}{n-4} \\
&= \frac{SSR(X_3|X_1, X_2)}{1} \div \frac{SSE(X_1, X_2, X_3)}{n-4} \\
&= \frac{MSR(X_3|X_1, X_2)}{MSE(X_1, X_2, X_3)} \qquad (7.15)
\end{aligned}$$

6. The test whether or not $\beta_3 = 0$ is a marginal test, given that X_1 and X_2 are already in the model.

7. Test statistic (7.15) shows that we do not need to fit both the full model and the reduced model to use the general linear test approach here. A single computer run can provide a fit of the full model and the appropriate extra sum of squares.

8. Example Body Fat Example

   (a) To test for the model with all three predictor variables whether midarm circumference (X_3) can be dropped from the model.

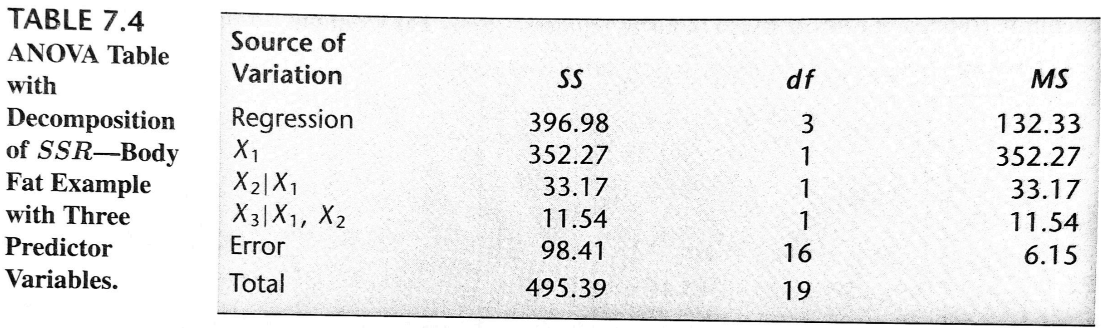


TABLE 7.4 ANOVA Table with Decomposition of *SSR*—Body Fat Example with Three Predictor Variables.

Source of Variation	*SS*	*df*	*MS*
Regression	396.98	3	132.33
X_1	352.27	1	352.27
$X_2 \mid X_1$	33.17	1	33.17
$X_3 \mid X_1, X_2$	11.54	1	11.54
Error	98.41	16	6.15
Total	495.39	19	

   (b) (Table 7.4) ANOVA results of the full regression model (7.12), including the extra sums of squares when the predictor variables are entered in the order X_1, X_2, X_3. Hence, test statistic (7.15) is:

$$F^* = \frac{SSR(X_3|X_1, X_2)}{1} \div \frac{SSE(X_1, X_2, X_3)}{n-4}$$

$$= \frac{11.54}{1} \div \frac{98.41}{16} = 1.88$$

   For $\alpha = 0.01$, we require $F(0.99; 1, 16) = 8.53$. Since $F^* = 1.88 \leq 8.53$, we conclude H_0, that X_3 can be dropped from the regression model that already contains X_1 and X_2.

   (c) (Table 7.2d) the t^* test statistic:

$$t^* = \frac{b_3}{s(b_3)} = \frac{-2.186}{1.596} = -1.37$$

   Since $(t^*)^2 = (-1.37)^2 = 1.88 = F^*$, we see that the two test statistics are equivalent, just as for simple linear regression.

9. The F^* test statistic (7.15) to test whether or not $\beta_3 = 0$ is called a partial F test statistic to distinguish it from the F^* statistic in (6.39b) for testing whether all $\beta_k = 0$, i.e., whether or not there is a regression relation between Y and the set of X variables. The latter test is called the overall F test.

Test whether Several $\beta_k = 0$

1. To know whether both $\beta_2 X_2$ and $\beta_3 X_3$ can be dropped from the full model (7.12). The alternatives here are:

$$H_0: \beta_2 = \beta_3 = 0 \qquad H_a: \text{ not both } \beta_2 \text{ and } \beta_3 \text{ equal zero} \qquad (7.16)$$

2. With the general linear test approach, the reduced model under H_0 is:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i \qquad \text{Reduced model (7.17)}$$

and the error sum of squares for the reduced model is:

$$SSE(R) = SSE(X_1) \qquad df_R = n - 2$$

3. The general linear test statistic:

$$\begin{aligned} F^* &= \frac{SSE(X_1) - SSE(X_1, X_2, X_3)}{(n-2) - (n-4)} \div \frac{SSE(X_1, X_2, X_3)}{n-4} \\ &= \frac{SSR(X_2, X_3|X_1)}{2} \div \frac{SSE(X_1, X_2, X_3)}{n-4} \\ &= \frac{MSR(X_2, X_3|X_1)}{MSE(X_1, X_2, X_3)} \end{aligned}$$

4. Example Body Fat Example

   (a) To test in the body fat example for the model with all three predictor variables whether both thigh circumference (X_2) and midarm circumference (X_3) can be dropped from the full regression model (7.12):

   $$SSR(X_2, X_3|X_1) = SSR(X_2|X_1) + SSR(X_3|X_1, X_2) = 33.17 + 11.54 = 44.71$$

   (b) Test statistic (7.18) therefore:

   $$F^* = \frac{SSR(X_2, X_3|X_1)}{2} \div MSE(X_1, X_2, X_3) = \frac{44.71}{2} \div 6.15 = 3.63$$

(c) For $\alpha = 0.05$, we require $F(0.95; 2, 16) = 3.63$. Since $F^* = 3.63$ is at the boundary of the decision rule (the P-value of the test statistic is 0.05), we may wish to make further analyses before deciding whether X_2 and X_3 should be dropped from the regression model that already contains X_1.

7.3 Summary of Tests Concerning Regression Coefficients*

7.4 Coefficients of Partial Determination

1. Extra sums of squares are not only useful for tests on the regression coefficients of a multiple regression model, but they are also encountered in descriptive measures of relationship called coefficients of partial determination.
2. Recall: the coefficient of multiple determination, R^2, measures the proportionate reduction in the variation of Y achieved by the introduction of the entire set of X variables considered in the model.
3. A coefficient of partial determination measures the marginal contribution of one X variable when all others are already included in the model.

Two Predictor Variables

1. Consider a first-order multiple regression model with two predictor variables:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i.$$

   (a) $SSE(X_2)$: measures the variation in Y when X_2 is included in the mode1.

   (b) $SSE(X_1, X_2)$ measures the variation in Y when both X_1 and X_2 are included in the model.

2. (Recall) Coefficient of determination: $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$.

3. The relative marginal reduction in the variation in Y associated with X_1 when X_2 is already in the model is:

$$R^2_{Y1|2} = \frac{SSE(X_2) - SSE(X_1, X_2)}{SSE(X_2)} = \frac{SSR(X_1|X_2)}{SSE(X_2)}$$

This measure is the coefficient of partial determination between Y and X_1, given that X_2 is in the mode1. Denoted by $R^2_{Y1|2}$.

4. $R^2_{Y1|2}$ measures the proportionate reduction in the variation in Y remaining after X_2 is included in the model that is gained by also including X_1 in the model.

5. The coefficient of partial determination between Y and X_2, given that X_1 is in the model, is defined correspondingly:

$$R^2_{Y2|1} = \frac{SSR(X_2|X_1)}{SSE(X_1)}$$

General Case

1. The generalization of coefficients of partial determination to three or more X variables in the model:

$$R^2_{Y1|23} = \frac{SSR(X_1|X_2, X_3)}{SSE(X_2, X_3)} \tag{7.37}$$

$$R^2_{Y2|13} = \frac{SSR(X_2|X_1, X_3)}{SSE(X_1, X_3)} \tag{7.38}$$

$$R^2_{Y3|12} = \frac{SSR(X_3|X_1, X_2)}{SSE(X_1, X_2)} \tag{7.39}$$

$$R^2_{Y4|123} = \frac{SSR(X_4|X_1, X_2, X_3)}{SSE(X_1, X_2, X_3)} \tag{7.40}$$

2. $\boxed{\text{Example}}$ Body Fat Example

   (a) Example: we can obtain a variety of coefficients of partial determination. (Tables 7.2 and 7.4):

$$R^2_{Y2|1} = \frac{SSR(X_2|X_1)}{SSE(X_1)} = \frac{33.17}{143.12} = 0.232$$

$$R^2_{Y3|12} = \underline{\frac{SSR(X_3|X_1,X_2)}{SSE(X_1,X_2)} = \frac{11.54}{109.95} = 0.105}$$

$$R^2_{Y1|2} = \underline{\frac{SSR(X_1|X_2)}{SSE(X_2)} = \frac{3.47}{113.42} = 0.031}$$

(b) When X_2 is added to the regression model containing X_1, the __error__ sum of squares __$SSE(X_1)$__ is reduced by __23.2 percent__.

(c) SSE for the model containing both X_1 and X_2 is only reduced by another __10.5__ percent when X_3 is added to the model.

(d) If the regression model already contains X_2, adding X_1 reduces __$SSE(X_2)$__ by only __3.1 percent__.

Coefficients of Partial Correlation

1. The __square root__ of a coefficient of partial determination is called a __coefficient of partial correlation__.

2. One use of partial correlation coefficients is in computer routines for finding the __best predictor variable__ to be selected next for inclusion in the regression model.

3. For the body fat example, we have:

$$\begin{aligned} r_{Y2|1} &= \sqrt{0.232} = 0.482 \\ r_{Y3|12} &= -\sqrt{0.105} = -0.324 \\ r_{Y1|2} &= \sqrt{0.031} = 0.176 \end{aligned}$$

4. The coefficients $r_{Y2|1}$ and $r_{Y1|2}$ are positive because we see from Table 7.2c that $b_2 = 0.6594$ and $b_1 = 0.2224$ are __positive__. Similarly, $r_{Y3|12}$ is negative because we see from Table 7.2d that $b_3 = -2.186$ is __negative__.

7.5 Standardized Multiple Regression Model*

7.6 Multicollinearity and Its Effects

1. In multiple regression analysis, some questions frequently asked:

(a) What is the relative importance of the effects of the different predictor variables?

(b) What is the magnitude of the effect of a given predictor variable on the response variable?

(c) Can any predictor variable be dropped from the model because it has little or no effect on the response variable?

(d) Should any predictor variables not yet included in the model be considered for possible inclusion ?

2. In many nonexperimental situations in business, economics, and the social and biological sciences, the predictors tend to be correlated among themselves and with other variables that are related to the response variable but are not included in the model.

3. Example In a regression of family food expenditures on the explanatory variables family income, family savings, and age of head of household, the explanatory variables will be correlated among themselves. Further, they will also be correlated with other socioeconomic variables not included in the model that do affect family food expenditures, such as family size.

4. When the predictor variables are correlated among themselves, intercorrelation or multicollinearity among them is said to exist.

Uncorrelated Predictor Variables

1. (Table 7.6) The data for a small-scale experiment on the effect of work crew size (X_1) and level of bonus pay (X_2) on crew productivity (Y). The predictor variables X_1 and X_2 are uncorrelated ($r_{12}^2 = 0$).

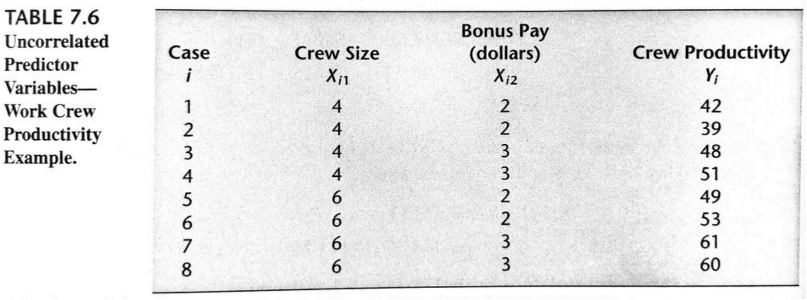


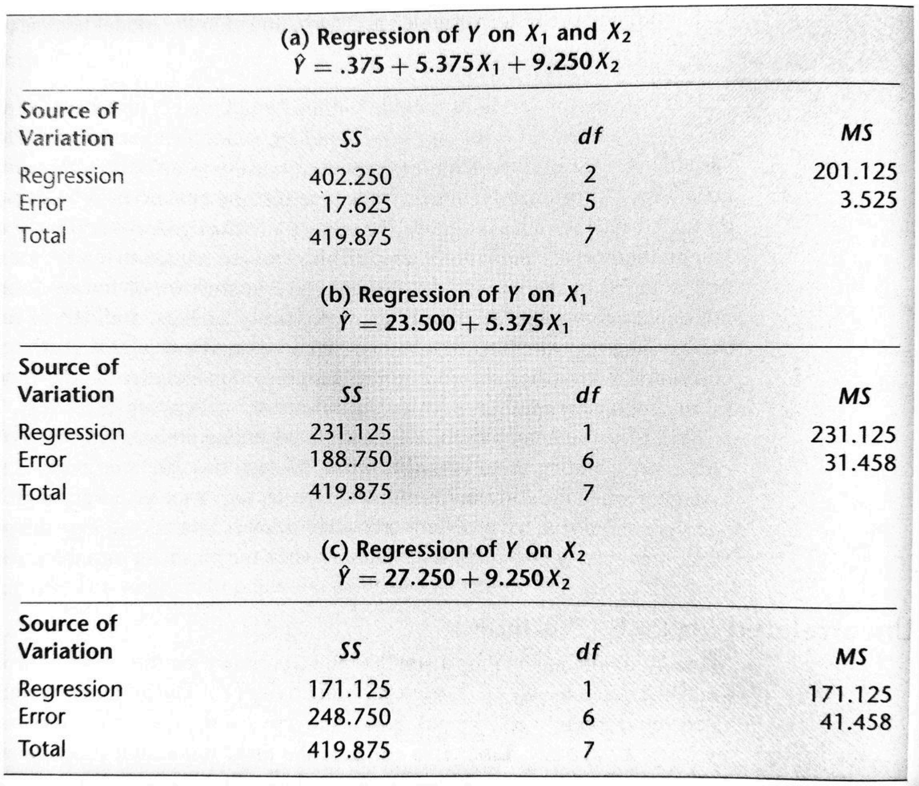
TABLE 7.6 Uncorrelated Predictor Variables—Work Crew Productivity Example.

Case i	Crew Size X_{i1}	Bonus Pay (dollars) X_{i2}	Crew Productivity Y_i
1	4	2	42
2	4	2	39
3	4	3	48
4	4	3	51
5	6	2	49
6	6	2	53
7	6	3	61
8	6	3	60

2. (Table 7.7a (7.7b) (7.7c)) The fitted regression function and the analysis of variance table when both X_1 and X_2 are (only (X_1) (X_2) is) included in the model.

3. (Table 7.7) The regression coefficient for X_1, $b_1 = 5.375$, is the same whether only X_1 is included in the model or both predictor variables are included. The same holds for $b_2 = 9.250$.

TABLE 7.7 Regression Results when Predictor Variables Are Uncorrelated—Work Crew Productivity Example.



(a) Regression of Y on X_1 and X_2

$\hat{Y} = .375 + 5.375X_1 + 9.250X_2$

Source of Variation	*SS*	*df*	*MS*
Regression	402.250	2	201.125
Error	17.625	5	3.525
Total	419.875	7	

(b) Regression of Y on X_1

$\hat{Y} = 23.500 + 5.375X_1$

Source of Variation	*SS*	*df*	*MS*
Regression	231.125	1	231.125
Error	188.750	6	31.458
Total	419.875	7	

(c) Regression of Y on X_2

$\hat{Y} = 27.250 + 9.250X_2$

Source of Variation	*SS*	*df*	*MS*
Regression	171.125	1	171.125
Error	248.750	6	41.458
Total	419.875	7	

4. When the predictor variables are uncorrelated, the effects ascribed to them by a first-order regression model are the same no matter which other of these predictor variables are included in the model.

5. The extra sum of squares $SSR(X_1|X_2)$ equals the regression sum of squares $SSR(X_1)$ when only X_1, is in the regression model:

$$\begin{aligned} SSR(X_1|X_2) &= SSE(X_2) - SSE(X_1, X_2) \\ &= 248.750 - 17.625 = 231.125 \\ SSR(X_1) &= 231.125 \end{aligned}$$

6. Similarly, the extra sum of squares $SSR(X_2|X_1)$ equals $SSR(X_2)$, the regression sum of squares when only X_2 is in the regression model:

$$\begin{aligned} SSR(X_2|X_1) &= \underline{SSE(X_1) - SSE(X1_1, X_2)} \\ &= \underline{188.750 - 17.625 = 171.125} \\ SSR(X_2) &= \underline{171.125} \end{aligned}$$

7. In general, when two or more predictor variables are uncorrelated, the marginal contribution of one predictor variable in reducing the error sum of squares when the other predictor variables are in the model is exactly the same as when this predictor variable is in the model alone.

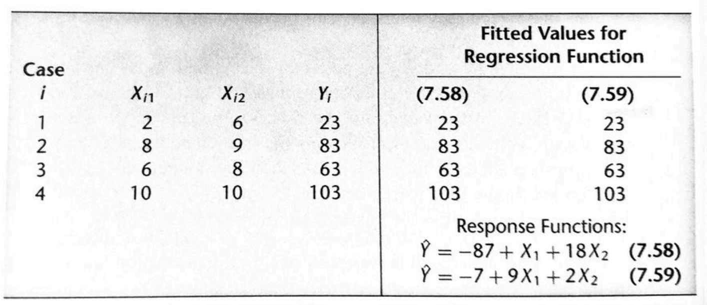
8. See **Comment** on page 281 for the proof: when X_1 and X_2 are uncorrelated, adding X_2 to the regression model does not change the regression coefficient for X_1; correspondingly, adding X_1 to the regression model does not change the regression coefficient for X_2.

Nature of Problem when Predictor Variables Are Perfectly Correlated

1. (Table 7.8) $\boxed{\text{Example}}$ The data refer to four sample observations on a response variable and two predictor variables. The first-order multiple regression function fit:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2.$$

TABLE 7.8 Example of Perfectly Correlated Predictor Variables.



Case i	X_{i1}	X_{i2}	Y_i	Fitted Values for Regression Function (7.58)	Fitted Values for Regression Function (7.59)
1	2	6	23	23	23
2	8	9	83	83	83
3	6	8	63	63	63
4	10	10	103	103	103

Response Functions:

$\hat{Y} = -87 + X_1 + 18X_2$ **(7.58)**

$\hat{Y} = -7 + 9X_1 + 2X_2$ **(7.59)**

$$\text{Mr. A} : \hat{Y} = -87 + X_1 + 18X_2 \quad \text{(perfect fit)} \quad (7.58)$$

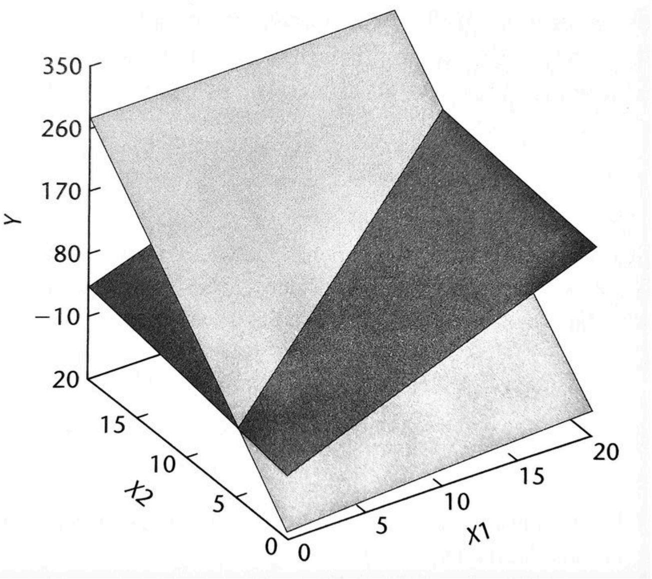
$$\text{Mr. B} : \hat{Y} = -7 + 9X_1 + 2X_2 \quad \text{(perfect fit)} \quad (7.59)$$

2. It can be shown that infinitely many response functions will fit the data in Table 7.8 perfectly. The reason is that the predictor variables X_1, and X_2 are perfectly related:

$$X_2 = 5 + 0.5X_1 \qquad (7.60)$$

3. (Figure 7.2) The fitted response functions (7.58) and (7.59) are entirely different response surfaces. The two response surfaces have the same fitted values only when they intersect.

FIGURE 7.2 Two Response Planes That Intersect when $X_2 = 5 + .5X_1$.



4. Two key implications of this example are:

   (a) The perfect relation between X_1, and X_2 did not inhibit our ability to obtain a good fit to the data.

   (b) Since many different response functions provide the same good fit, we cannot interpret anyone set of regression coefficients as reflecting the effects of the different predictor variables.

Effects of Multicollinearity

1. The fact that some or all predictor variables are correlated among themselves (a) does not, in general, inhibit our ability to obtain a good fit (b) nor does it tend

to affect inferences about mean responses or predictions of new observations, provided these inferences are made within the region of observations.

2. The estimated regression coefficients tend to have large sampling variability when the predictor variables are highly correlated. Thus, the estimated regression coefficients tend to vary widely from one sample to the next when the predictor variables are highly correlated.

3. Many of the estimated regression coefficients individually may be statistically not significant even though a definite statistical relation exists between the response variable and the set of predictor variables.

4. The common interpretation of a regression coefficient as measuring the change in the expected value of the response variable when the given predictor variable is increased by one unit while all other predictor variables are held constant is not fully applicable when multicollinearity exists.

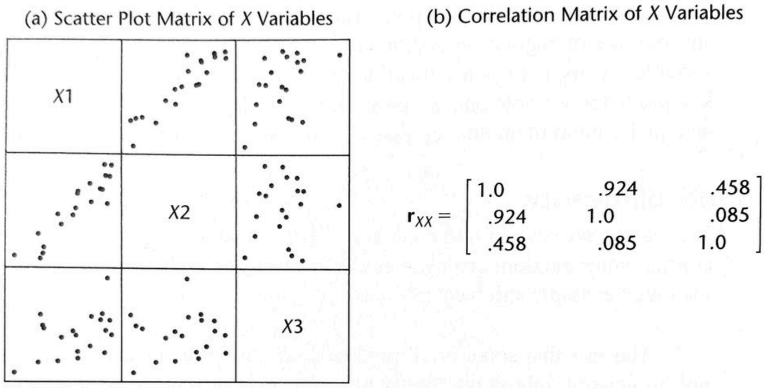
5. $\boxed{\text{Example}}$ The Body Fat Example

   (a) (Table 7.1): A sample of 20 healthy females $25-34$ years old, Y: amount of body fat, X_1: triceps skinfold thickness, X_2: thigh circumference, X_3: midarm circumference. (Table 7.2): The regression results for different fitted models.

   (b) (Figure 7.3) The scatter plot matrix and the correlation matrix of the predictor variables: predictor variables X_1 and X_2 are highly correlated $(r_{12} = 0.924)$.

   (c) $r_{13} = 0.458$ and $r_{23} = 0.085$.

   (d) The coefficient of multiple determination when X_3 is regressed on X_1 and X_2 is 0.998: X_3 is highly correlated with X_1 and X_2 together.

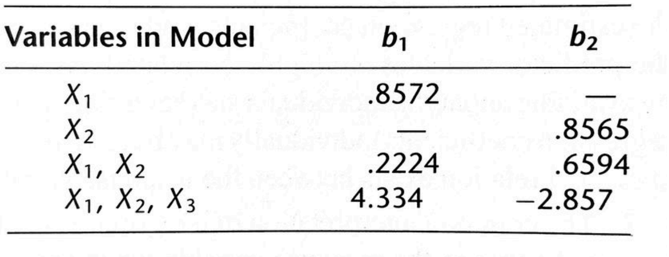


$$\mathbf{r}_{XX} = \begin{bmatrix} 1.0 & .924 & .458 \\ .924 & 1.0 & .085 \\ .458 & .085 & 1.0 \end{bmatrix}$$

FIGURE 7.3 Scatter Plot Matrix and Correlation Matrix of the Predictor Variables—Body Fat Example.

6. **Effects on Regression Coefficients.**

   (a) The regression coefficient for X_1, triceps skinfold thickness, varies markedly depending on which other variables are included in the model.



Variables in Model	b_1	b_2
X_1	.8572	—
X_2	—	.8565
X_1, X_2	.2224	.6594
X_1, X_2, X_3	4.334	−2.857

   (b) The story is the same for the regression coefficient for X_2. The regression coefficient b_2 even changes sign when X_3 is added to the model that includes X_1 and X_2.

   (c) *Important conclusion*: When predictor variables are correlated, the regression coefficient of anyone variable depends on which other predictor variables are included in the model and which ones are left out. Thus, a regression coefficient does not reflect any inherent effect of the particular predictor variable on the response variable but only a marginal or partial effect, given whatever other correlated predictor variables are included in the model.

7. **Effects on Extra Sums of Squares.**

   (a) When predictor variables are correlated, the marginal contribution of anyone predictor variable in reducing the error sum of squares varies, depending on which other variables are already in the regression model, just as for regression coefficients.

   (b) (Table 7.2) Consider the following extra sums of squares for X_1:

$$SSR(X_1) = 352.27 \qquad SSR(X_1|X_2) = 3.47.$$

   The reason why $SSR(X_1|X_2)$ is so small compared with $SSR(X_1)$ is that X_1 and X_2 are highly correlated with each other and with the response variable.

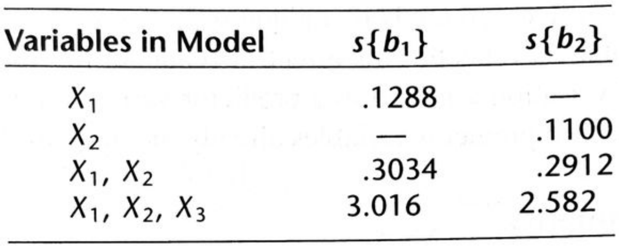
   (c) When X_2 is already in the regression model, the marginal contribution of X_1 in reducing the error sum of squares is comparatively small because X_2 contains much of the same information as X_1.

(d) The same story is found in Table 7.2 for X_2. Here $SSR(X_2|X_1) =$ 33.17, which is much smaller than $SSR(X_2) =$ 381.97.

(e) *Important conclusion*: When predictor variables are correlated, there is no unique sum of squares that can be ascribed to anyone predictor variable as reflecting its effect in reducing the total variation in Y. The reduction in the total variation ascribed to a predictor variable must be viewed in the context of the other correlated predictor variables already included in the model.

8. **Effects on** $s(b_k)$.

(a) (Table 7.2 for the body fat example) how much more imprecise the estimated regression coefficients b_1 and b_2 become as more predictor variables are added to the regression model:

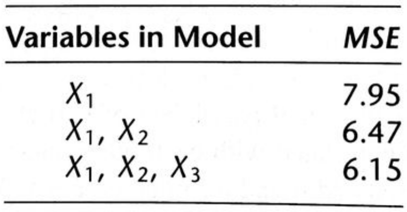


Variables in Model	$s\{b_1\}$	$s\{b_2\}$
X_1	.1288	—
X_2	—	.1100
X_1, X_2	.3034	.2912
X_1, X_2, X_3	3.016	2.582

(b) The high degree of multicollinearity among the predictor variables is responsible for the inflated variability of the estimated regression coefficients.

9. **Effects on Fitted Values and Predictions**.

(a) (Table 7.2 for the body fat example) the high multicollinearity among the predictor variables does not prevent the mean square error, measuring the variability of the error terms, from being steadily reduced as additional variables are added to the regression model:



Variables in Model	MSE
X_1	7.95
X_1, X_2	6.47
X_1, X_2, X_3	6.15

(b) The precision of fitted values within the range of the observations on the predictor variables is not eroded with the addition of correlated predictor variables into the regression model.

(c) Example Consider the estimation of mean body fat when the only predictor variable in the model is triceps skinfold thickness (X_1) for $X_{h1} = 25.0$. The fitted value and its estimated standard deviation are (calculations not shown):

$$\hat{Y}_h = 19.93, \quad s(\hat{Y}_h) = 0.632$$

When the highly correlated predictor variable thigh circumference (X_2) is also included in the model, the estimated mean body fat and its estimated standard deviation are as follows for $X_{h1} = 25.0$ and $X_{h2} = 50.0$:

$$\hat{Y}_h = 19.36 \quad s(\hat{Y}_h) = 0.624$$

Thus, the precision of the estimated mean response is equally good as before, despite the addition of the second predictor variable that is highly correlated with the first one.

(d) The essential reason for the stability is that the covariance between b_1 and b_2 is negative, which plays a strong counteracting influence to the increase in $s^2(b_1)$, in determining the value of $s^2(\hat{Y}_h)$ as given in (6.79).

$$s^2\{\hat{Y}_h\} = s^2\{b_0\} + X_{h1}^2 s^2\{b_1\} + X_{h2}^2 s^2\{b_2\} + 2X_{h1}s\{b_0, b_1\} \\ + 2X_{h2}s\{b_0, b_2\} + 2X_{h1}X_{h2}s\{b_1, b_2\} \qquad \textbf{(6.79)}$$

10. **Effects on Simultaneous Tests of β_k.** Paradox of t-test and F-test:

(a) (The Body Fat Example) test whether $\beta_1 = 0$ and $\beta_2 = 0$. Controlling the family level of significance at 0.05, we require with the Bonferroni method that each of the two t tests be conducted with level of significance 0.025.

(b) Hence, we need $t_{(.975;17)} = 2.46$. Since both t^* statistics in Table 7.2c have absolute values that do not exceed 2.46, we would conclude from the two separate tests that $\beta_1 = 0$ and that $\beta_2 = 0$.

(c) (Table 7.2c) Yet the proper F test for $H_0 : \beta_1 = \beta_2 = 0$ would lead to the conclusion H_a that not both coefficients equal zero. We find $F^* = MSR/MSE = 192.72/6.47 = 29.8$, which far exceeds $F_{(0.95;2,17)} = 3.59$.

(d) The reason for this apparently paradoxical result is that each t^* test is a marginal test, as we have seen in (7.15) from the perspective of the general linear test approach.

(e) Thus, a small $SSR(X_1|X_2)$ here indicates that X_1, does not provide much additional information beyond X_2, which already is in the model; hence, we are led to the conclusion that $\beta_1 = 0$.

(f) Similarly, we are led to conclude $\beta_2 = 0$ here because $SSR(X_2|X_1)$ is small, indicating that X_2 does not provide much more additional information when X_1 is already in the model.

(g) But the two tests of the marginal effects of X_1 and X_2 together are not equivalent to testing whether there is a regression relation between Y and the two predictor variables.

(h) The reason is that the reduced model for each of the separate tests contains the other predictor variable, whereas the reduced model for testing whether both $\beta_1 = 0$ and $\beta_2 = 0$ would contain neither predictor variable. The proper F test shows that there is a definite regression relation here between Y and X_1 and X_2.

Need for More Powerful Diagnostics for Multicollinearity

1. The diagnostic tool for identifying multicollinearity: the pairwise coefficients of simple correlation between the predictor variables is frequently helpful.

2. (Chapter 10) more powerful tool for identifying the existence of serious multicollinearity.

3. (Chapter 11) Some remedial measures for lessening the effects of multicollinearity.

😍 TA Class

- **Problems**: 7.2, 7.3, 7.6, 7.11, 7.24.
- **Exercises**: 7.31

“你無法改變別人的長相，但我們可以改變我們看人的方式。”
“You can not change someone’s looks, but we can change the way we look.”

— *奇蹟男孩 (Wonder, 2017)*

Regression Analysis (I)

Kutner's Applied Linear Statistical Models (5/E)

Chapter 8: Regression Models for Quantitative and Qualitative Predictors

Thursday 09:10-12:00, 商館 260205

Han-Ming Wu

Department of Statistics, National Chengchi University

http://www.hmwu.idv.tw

Overview

1. We consider in greater detail standard modeling techniques for quantitative predictors, for qualitative predictors, and for regression models containing both quantitative and qualitative predictors.

2. These techniques include the use of interaction and polynomial terms for quantitative predictors, and the use of indicator variables for qualitative predictors.

8.1 Polynomial Regression Models

1. The polynomial regression models for quantitative predictor variables are among the most frequently used curvilinear response models in practice because they are handled easily as a special case of the general linear regression model (6.7).

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

2. We discuss several commonly used polynomial regression models.

3. Then we present a case to illustrate some of the major issues encountered with polynomial regression models.

Uses of Polynomial Models

1. Polynomial regression models have two basic types of uses:

   (a) When the true curvilinear response function is indeed a polynomial function.

   (b) When the true curvilinear response function is unknown (or complex) but a polynomial function is a good approximation to the true function.

One Predictor Variable - Second Order

1. Polynomial regression models may contain one, two, or more than two predictor variables. Further, each predictor variable may be present in various powers.

2. Considering a polynomial regression model (called a second-order model with one predictor variable):

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i \qquad (8.1)$$

   or

$$Y_i = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2 + \varepsilon_i \qquad (8.2)$$

   where: $x_i = X_i - \bar{X}$.

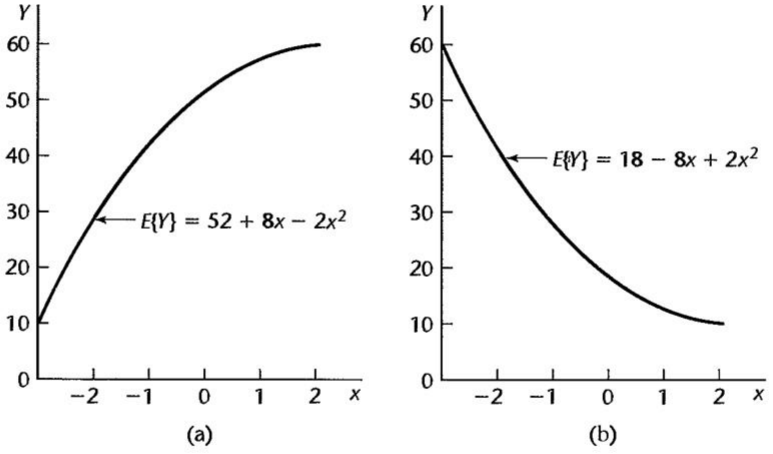
3. Note that the predictor variable is centered-in other words, expressed as a deviation around its mean $\bar{X}$ - and that the ith centered observation is denoted by x_i.

4. The reason for using a centered predictor variable in the polynomial regression model is that X and X^2 often will be highly correlated. Centering the predictor variable often reduces the multicollinearity substantially.

5. The response function for regression model (8.2) is (called a quadratic response function):

$$E\{Y\} = \beta_0 + \beta_1 x + \beta_{11} x^2 \qquad (8.3)$$

FIGURE 8.1 Examples of Second-Order Polynomial Response Functions.



6. The regression coefficient β_0 represents the mean response of Y when $x = 0$, i.e., when $X = \bar{X}$. The regression coefficient β_1 is called the linear effect coefficient, and β_{11} is called the quadratic effect coefficient.

One Predictor Variable - Third Order

1. The regression model is called a third-order model with one predictor variable

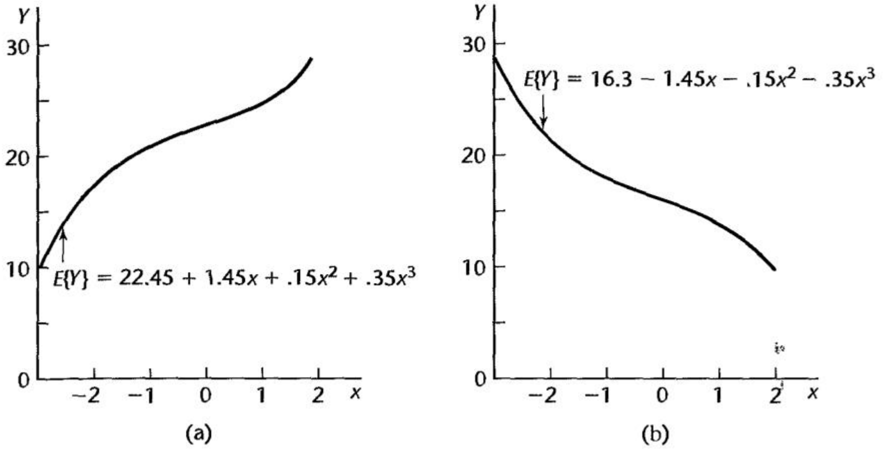
$$Y_i = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2 + \beta_{111} x_i^3 + \varepsilon_i \tag{8.5}$$

where $x_i = X_i - \bar{X}$

2. The response function for regression model (8.5) is:

$$E\{Y\} = \beta_0 + \beta_1 x + \beta_{11} x^2 + \beta_{111} x^3 \tag{8.6}$$

FIGURE 8.2 Examples of Third-Order Polynomial Response Functions.



One Predictor Variable - Higher Orders

1. Polynomial models with the predictor variable present in higher powers than the third should be employed with special caution. The interpretation of the coefficients becomes difficult for such models.

Two Predictor Variables - Second Order

1. The regression model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2 + \beta_{12} x_{i1} x_{i2} + \varepsilon_i \tag{8.7}$$

   where $x_{i1} = X_{i1} - \bar{X}_1$, $x_{i2} = X_{i2} - \bar{X}_2$.

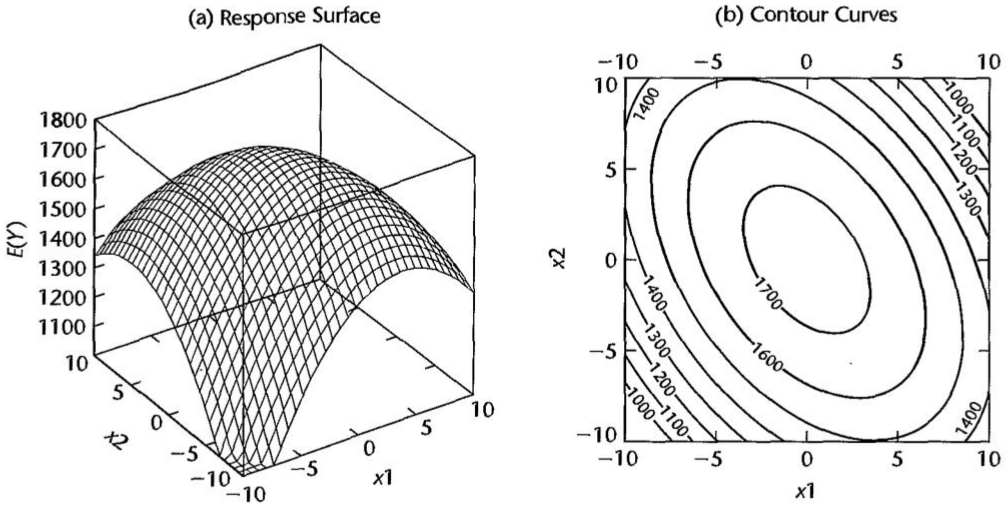
2. The response function is:

$$E\{Y\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 \tag{8.8}$$

3. Note that regression model (8.7) contains separate linear and quadratic components for each of the two predictor variables and a cross-product term.

4. The latter represents the interaction effect between X_1 and X_2. The coefficient β_{12} is often called the interaction effect coefficient.

FIGURE 8.3 **Example of a Quadratic Response Surface—$E\{Y\} = 1{,}740 - 4x_1^2 - 3x_2^2 - 3x_1x_2$.**



(a) Response Surface

(b) Contour Curves

Three Predictor Variables - Second Order*

Implementation of Polynomial Regression Models*

Case Example

1. **Setting**. A researcher studied the effects of the charge rate and temperature on the life of a new type of power cell in a preliminary small-scale expetiment. The charge rate (X_1) was controlled at three levels (0.6, 1.0, and 1.4 amperes (安培)) and the ambient temperature (X_2) was controlled at three levels (l0, 20, 30○C). Factors pertaining to the discharge of the power cell were held at fixed levels. The life of the power cell (Y) was measured in terms of the number of discharge - charge cycles that a power cell underwent before it failed.

2. **Model to be Considered**. (Table 8.1) The data obtained in the study are contained in Table 8.1, columns 1-3. The researcher was not sure about the nature of the response function in the range of the factors studied. Hence, the researcher decided to fit the second-order polynomial regression model (8.7):

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2 + \beta_{12} x_{i1} x_{i2} + \varepsilon_i \qquad (8.13)$$

for which the response function is:

$$E\{Y\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 \qquad (8.14)$$

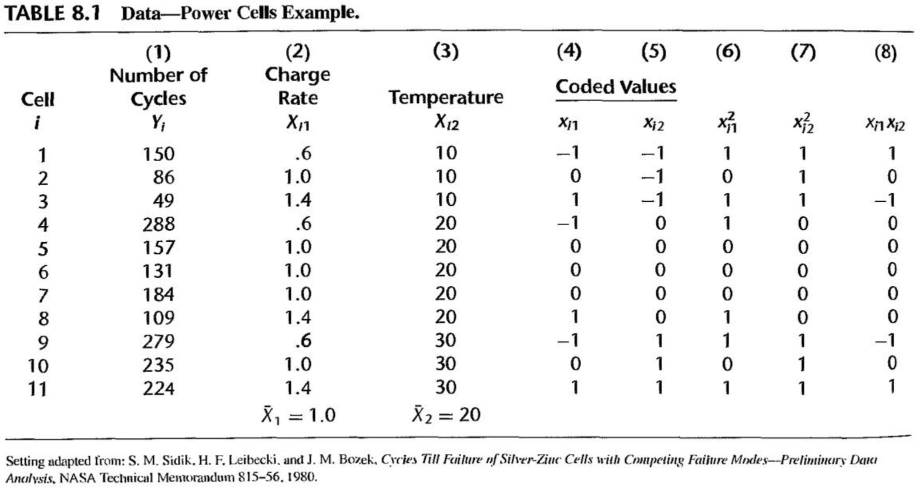


TABLE 8.1 Data—Power Cells Example.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Number of	Charge		Coded Values				
Cell	Cycles	Rate	Temperature					
i	Y_i	X_{i1}	X_{i2}	x_{i1}	x_{i2}	x_{i1}^2	x_{i2}^2	$x_{i1}x_{i2}$
1	150	.6	10	−1	−1	1	1	1
2	86	1.0	10	0	−1	0	1	0
3	49	1.4	10	1	−1	1	1	−1
4	288	.6	20	−1	0	1	0	0
5	157	1.0	20	0	0	0	0	0
6	131	1.0	20	0	0	0	0	0
7	184	1.0	20	0	0	0	0	0
8	109	1.4	20	1	0	1	0	0
9	279	.6	30	−1	1	1	1	−1
10	235	1.0	30	0	1	0	1	0
11	224	1.4	30	1	1	1	1	1
		$\bar{X}_1 = 1.0$	$\bar{X}_2 = 20$					

Setting adapted from: S. M. Sidik, H. F. Leibecki, and J. M. Bozek, *Cycles Till Failure of Silver-Zinc Cells with Competing Failure Modes—Preliminary Data Analysis*, NASA Technical Memorandum 815-56, 1980.

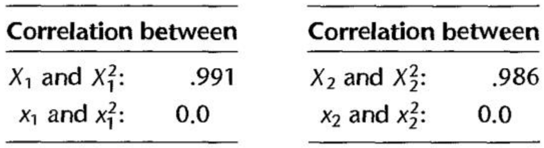
3. **Coded Variables**. Because of the balanced nature of the X_1 and X_2 levels studied, the researcher not only centered the variables X_1 and X_2 around their respective means but also scaled them in convenient units, as follows:

$$\begin{aligned} x_{i1} &= \frac{X_{i1} - \bar{X}_1}{0.4} = \frac{X_{i1} - 1.0}{0.4} \\ x_{i2} &= \frac{X_{i2} - \bar{X}_2}{10} = \frac{X_{i1} - 20}{10} \end{aligned}$$

(a) Here, the denominator used for each predictor variable is the absolute difference between adjacent levels of the variable.

(b) These centered and scaled variables are shown in columns 4 and 5 of Table 8.1. Note that the codings defined in (8.15) lead to simple coded values, -1, 0, and 1. The squared and cross-product terms are shown in columns 6-8 of Table 8.1.

(c) Use of the coded variables x_1 and x_2 rather than the original variables X_1 and X_2 reduces the correlations between the first power and second power terms markedly. Low levels of multicollinearity can be helpful in avoiding computational inaccuracies.



Correlation between	
X_1 and X_1^2:	.991
x_1 and x_1^2:	0.0

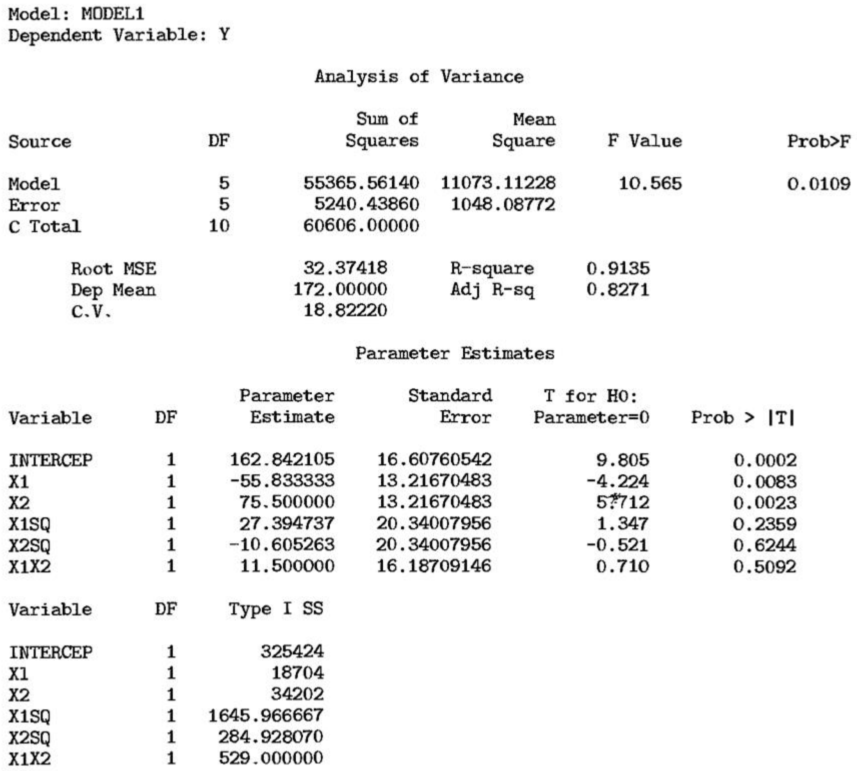
Correlation between	
X_2 and X_2^2:	.986
x_2 and x_2^2:	0.0

(d) The researcher was particularly interested in whether interaction effects and curvature effects are required in the model for the range of the X variables considered.

4. **Fitting of Model**. (Figure 8.4) contains the basic regression results for the fit of model (8.13) with the SAS regression package. The estimated regression function :

$$\hat{Y} = 162.84 - 55.83x_1 + 75.50x_2 + 27.39x_1^2 - 10.61x_2^2 + 11.50x_1x_2 \tag{8.16}$$

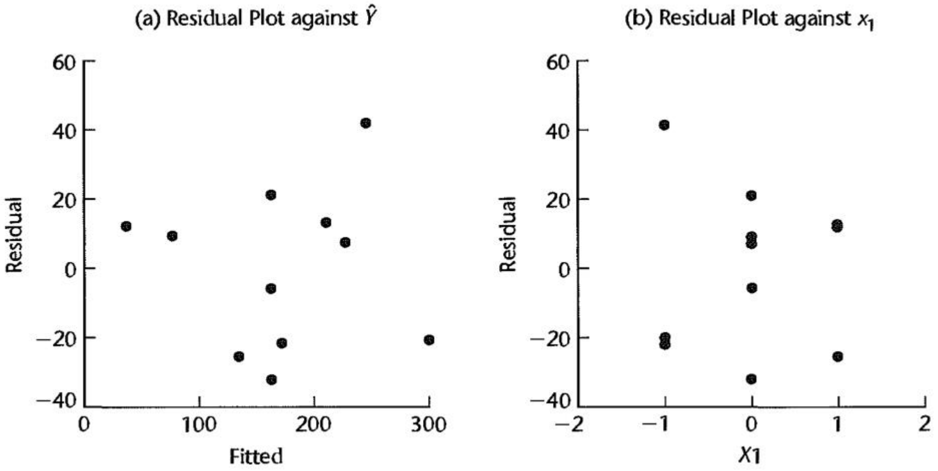
FIGURE 8.4 SAS Regression Output for Second-Order Polynomial Model (8.13)—Power Cells Example.

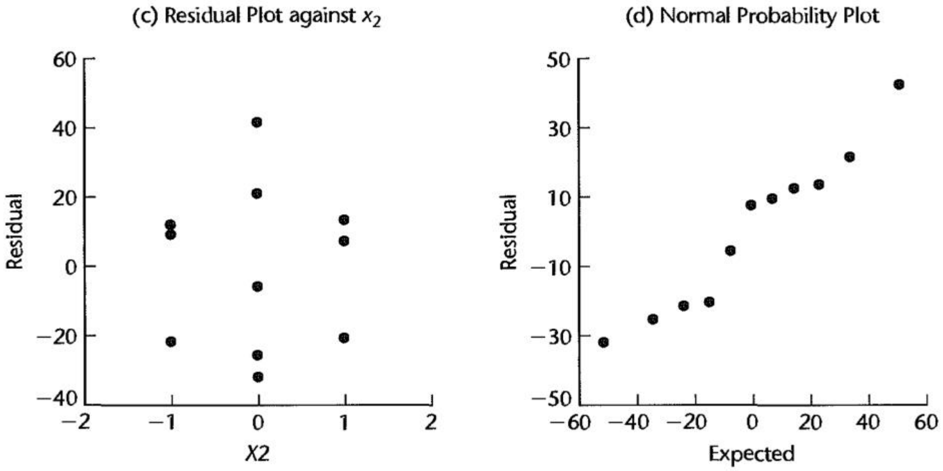


```
Model: MODEL1
Dependent Variable: Y

                          Analysis of Variance

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

Model              5     55365.56140   11073.11228       10.565        0.0109
Error              5      5240.43860    1048.08772
C Total           10     60606.00000

      Root MSE             32.37418      R-square      0.9135
      Dep Mean            172.00000      Adj R-sq      0.8271
      C.V.                 18.82220

                          Parameter Estimates

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

INTERCEP       1     162.842105      16.60760542         9.805       0.0002
X1             1     -55.833333      13.21670483        -4.224       0.0083
X2             1      75.500000      13.21670483         5.712       0.0023
X1SQ           1      27.394737      20.34007956         1.347       0.2359
X2SQ           1     -10.605263      20.34007956        -0.521       0.6244
X1X2           1      11.500000      16.18709146         0.710       0.5092

Variable      DF      Type I SS

INTERCEP       1         325424
X1             1          18704
X2             1          34202
X1SQ           1    1645.966667
X2SQ           1     284.928070
X1X2           1     529.000000
```

5. **Residual Plots**. (Figure 8.5) None of these plots suggest any gross inadequacies of regression model (8.13). The coefficient of correlation between the ordered residuals and their expected values under normality is 0.974, which supports the assumption of normality of the error terms.

FIGURE 8.5 Diagnostic Residual Plots—Power Cells Example.





6. **Test of Fit**. Since there are three replications at $x_1 = 0$, $x_2 = 0$, another indication of the adequacy of regression model (8.13) can be obtained by the formal test in (6.68) of the goodness of fit of the regression function (8.14).

   (a) The pure error sum of squares (3.16):

   $$SSPE = (157 - 157.33)^2 + (131 - 157.33)^2 + (184 - 157.33)^2 = 1,404.67$$

   Since there are $c = 9$ distinct combinations of levels of the X variables here, there are $n - c = 11 - 9 = 2$ degrees of freedom associated with SSPE.

   (b) (Figure 8.4) $SSE = 5,240.44$. Hence the lack of fit sum of squares (3.24) is:

   $$SSLF = \underline{SSE - SSPE = 5,240.44 - 1,404.67} = 3,835.77$$

   with which $c - p = 9 - 6 = 3$ degrees of freedom are associated. ($p = 6$ regression coefficients in model (8.13) had to be estimated.)

   (c) Hence, test statistic (6.68b) for testing the adequacy of the regression function (8.14) is:

   $$F^* = \underline{\frac{SSLF}{c-p} \div \frac{SSPE}{n-c} = \frac{3,835.77}{3} \div \frac{1,404.67}{2} = 1.82}$$

   (d) For $\alpha = 0.05$, we require $F_{(0.95;3,2)} = 19.2$. Since $F^* = 1.82 \leq 19.2$, we conclude according to decision rule (6.68c) that the second-order polynomial regression function (8.14) is a good fit.

7. **Coefficient of Multiple Determination**. (Figure 8.4) $R^2 = 0.9135$: the variation in the lives of the power cells is reduced by about 91 percent when the first-order and second-order relations to the charge rate and ambient temperature are utilized. The adjusted $R^2 = 0.8271$.

8. **Partial F Test**. Whether a first-order model would be sufficient? The test alternatives are:

$$H_0 : \beta_{11} = \beta_{22} = \beta_{12} = 0, \qquad H_a : \text{ not all } \beta\text{s in } H_0 \text{ equal zero}$$

   (a) The partial F test statistic (7.27) here is:

$$F^* = \frac{SSR(x_1^2, x_2^2, x_1x_2|x_1, x_2)}{3} \div MSE$$

   (b) (Figure 8.4) $SSR(x_1) = 18,704$, $SSR(x_2|x_1) = 34,202$. The required extra sum of squares is therefore obtained:

$$\begin{aligned} SSR(x_1^2, x_2^2, x_1x_2|x_1, x_2) &= SSR(x_1^2|x_1, x_2) + SSR(x_2^2|x_1, x_2, x_1^2) \\ &\quad +SSR(x_1x_2|x_1, x_2, x_1^2, x_2^2) \\ &= 1,646.0 + 284.9 + 529.0 = 2,459.9. \end{aligned}$$

   (c) (Figure 8.4) $MSE = 1048.1$. Hence the test statistic is:

$$F^* = \frac{2459.9}{3} \div 1048.1 = 0.78$$

   (d) For level of significance $\alpha = 0.05$, we require $F_{(0.95;3.5)} = 5.41$. Since $F^* = 0.78 \leq 5.41$, we conclude H_0, that no curvature and interaction effects are needed, so that a first-order model is adequate for the range of the charge rates and temperatures considered.

9. **First-Order Model**. On the basis of this analysis, the researcher decided to consider the first-order model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i \qquad (8.17)$$

   (a) A fit of this model yielded the estimated response function:

$$\hat{Y} = 172.00 - 55.83x_1 + 75.50x_2 \qquad (8.18) \quad s(b_1) = 12.67, s(b_2) = 12.67.$$

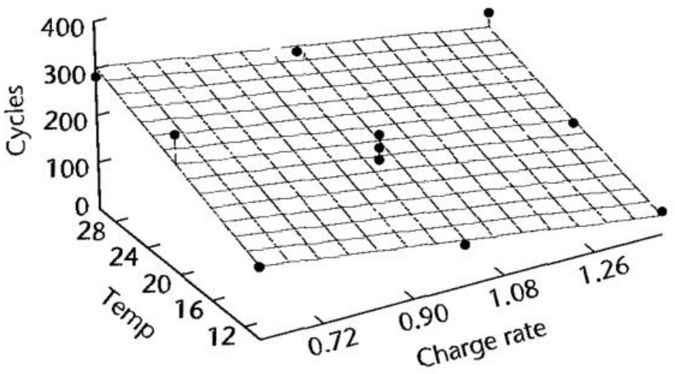
(b) A variety of residual plots for this first-order model were made and analyzed by the researcher (not shown here), which confirmed the appropriateness of first-order model (8.17).

10. **Fitted First-Order Model in Terms of X.** The fitted first-order regression function (8.18) can be transformed back to the original variables by utilizing (8.15). We obtain:

$$\hat{Y} = 160.58 - 139.58X_1 + 7.55X_2 \qquad (8.19)$$

(Figure 8.6) contains an S-Plus regression-scatter plot of the fitted response plane. The researcher used this fitted response surface for investigating the effects of charge rate and temperature on the life of this new type of power cell.

FIGURE 8.6 S-Plus Plot of Fitted Response Plane (8.19)—Power Cells Example.



11. **Estimation of Regression Coefficients.** The researcher wished to estimate the linear effects of the two predictor variables in the first-order model, with a 90 percent family confidence coefficient, by means of the Bonferroni method.

(a) *Joint Inferences* (page 228) The Bonferroni joint confidence intervals can be used to estimate several regression coefficients simultaneously. If g parameters are to be estimated jointly (where $g \leq p$), the confidence limits with family confidence coefficient $1 - \alpha$ are:

$$b_k \pm B\ s\{b_k\}, \text{ where } B = t_{(1-\alpha/2g;n-p)} \qquad (6.52)$$

(b) Here, $g = 2$ statements are desired; hence, by (6.52a), we have:

$$B = t_{(1-0.10/2(2)),8} = t_{(0.975;8)} = 2.306$$

(c) The estimated standard deviations of b_1 and b_2 in (8.18) apply to the model in the coded variables. Since only first-order terms are involved in this fitted model, we obtain the estimated standard deviations of b_1' and b_2' for the fitted model (8.19) in the original variables:

$$\begin{aligned} s\{b_1'\} &= \frac{1}{0.4}s\{b_1\} = \frac{12.67}{0.4} = 31.68 \\ s\{b_2'\} &= \frac{1}{10}s\{b_2\} = \frac{12.67}{10} = 1.267 \end{aligned}$$

(d) The Bonferroni confidence limits by (6.52) therefore are $-139.58 \pm 2.306(31.68)$ and $7.55 \pm 2.306(1.267)$, yielding the confidence limits:

$$-212.6 \leq \beta_1 \leq -66.5, \qquad \text{and} \qquad 4.6 \leq \beta_2 \leq 10.5$$

(e) With confidence 90%, we conclude that the mean number of charge/discharge cycles before failure decreases by 66 to 213 cycles with a unit increase in the charge rate for given ambient temperature, and increases by 5 to 10 cycles with a unit increase of ambient temperature for given charge rate.

(f) The researcher was satisfied with the precision of these estimates for this initial small-scale study.

Some Further Comments on Polynomial Regression

8.2 Interaction Regression Models*

8.3 Qualitative Predictors

1. Examples of qualitative predictor variables are gender (male, female), purchase status (purchase, no purchase), and disability status (not disabled, partly disabled, fully disabled).

2. Example In a study of innovation in the insurance industry, an economist wished to relate the speed with which a particular insurance innovation is adopted (Y) to the size of the insurance firm (X_1) and the type of firm (X_2).

   (a) Y: the number of months elapsed between the time the first firm adopted the innovation and the time the given firm adopted the innovation.

(b) X_1: size of firm, is quantitative, and is measured by the amount of total assets of the firm.

(c) X_2: type of firm, is qualitative and is composed of two classes – stock companies and mutual companies.

In order that such a qualitative variable can be used in a regression model, quantitative indicators for the classes of the qualitative variable must be employed.

Qualitative Predictor with Two Classes

1. We shall use indicator variables that take on the values 0 and 1 to quantify a qualitative variable.

2. $\boxed{\text{Example}}$ For the insurance innovation example, where the qualitative predictor variable has two classes, we might define two indicator variables X_2 and X_3:

$$X_2 = \begin{cases} 1 & \text{if stock company} \\ 0 & \text{otherwise} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{if mutual company} \\ 0 & \text{otherwise} \end{cases}$$

3. A first-order model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i \qquad (8.31)$$

4. This intuitive approach of setting up an indicator variable for each class of the qualitative predictor variable unfortunately leads to computational difficulties: $\mathbf{X'X}$ matrix does not have an inverse, and no unique estimators of the regression coefficient can be found (see details at page 314.)

5. Principle: A qualitative variable with c classes will be represented by $c-1$ indicator variables, each taking on the values 0 and 1.

6. Indicator variables are frequently also called dummy variables or binary variables.

Interpretation of Regression Coefficients

1. $\boxed{\text{Example}}$ Returning to the insurance innovation example, suppose that we drop the indicator variable X_3 from regression model (8.31) so that the model becomes:

$$\underline{\quad Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \quad} \qquad (8.33)$$

where:

$$\begin{aligned} X_1 &= \text{size of firm} \\ X_2 &= \begin{cases} 1 & \underline{\quad \text{if stock company} \quad} \\ 0 & \underline{\quad \text{if mutual company} \quad} \end{cases} \end{aligned}$$

2. The response function for this regression model is:

$$\underline{\quad E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \quad} \qquad (8.34)$$

(a) (Figure 8.11) Consider first the case of a mutual firm. For such a firm, $X_2 = 0$ and response function (8.34) becomes:

$$\underline{\quad E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2(0) = \beta_0 + \beta_1 X_1 \quad} \quad \text{Mutual firms} \quad (8.34a)$$

Thus, the response function for mutual firms is a straight line, with Y intercept $\underline{\quad \beta_0 \quad}$ and slope $\underline{\quad \beta_1 \quad}$.

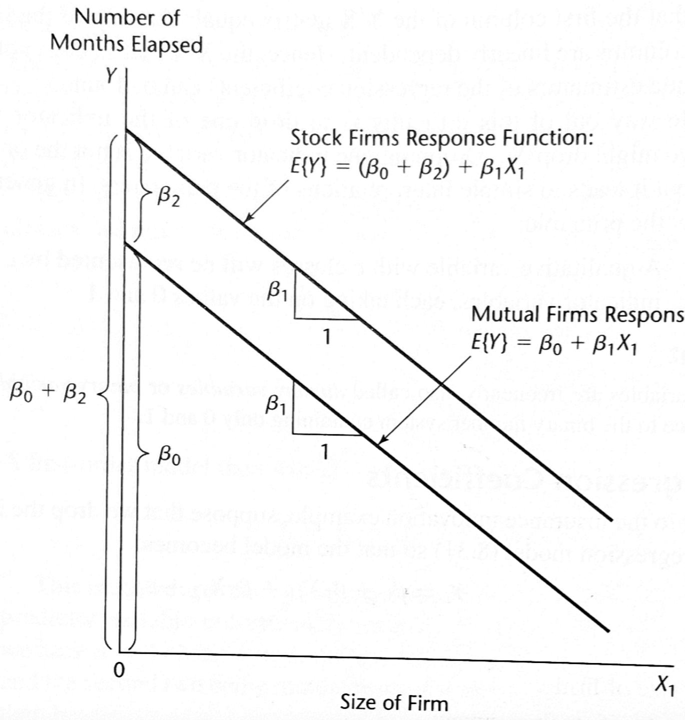
(b) For a stock firm, $X_2 = 1$ and response function (8.34) becomes:

$$\underline{\quad E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2(1) = (\beta_0 + \beta_2) + \beta_1 X_1 \quad} \quad \text{Stock firms} \quad (8.34b)$$

This also is a straight line, with the same slope $\underline{\quad \beta_1 \quad}$ but with Y intercept $\underline{\quad \beta_0 + \beta_2 \quad}$.

316 Part Two *Multiple Linear Regression*

FIGURE 8.11 Illustration of Meaning of Regression Coefficients for Regression Model (8.33) with Indicator Variable X_2—Insurance Innovation Example.



3. The meaning of the regression coefficients in response function (8.34)

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \qquad (8.34)$$

(a) The mean time elapsed before the innovation is adopted, $E\{Y\}$, is a linear function of size of firm (X_1), with the same slope β_1 for both types of firms.

(b) β_2 indicates how much higher (lower) the response function for stock firms (coded 1) is than the one for mutual firms (coded 0), for any given size of firm.

(c) β_2 measures the differential effect of type of firm.

4. Why we did not simply fit separate regressions for stock firms and mutual firms in our example, and instead adopted the approach of fitting one regression with an indicator variable. There are two reasons:

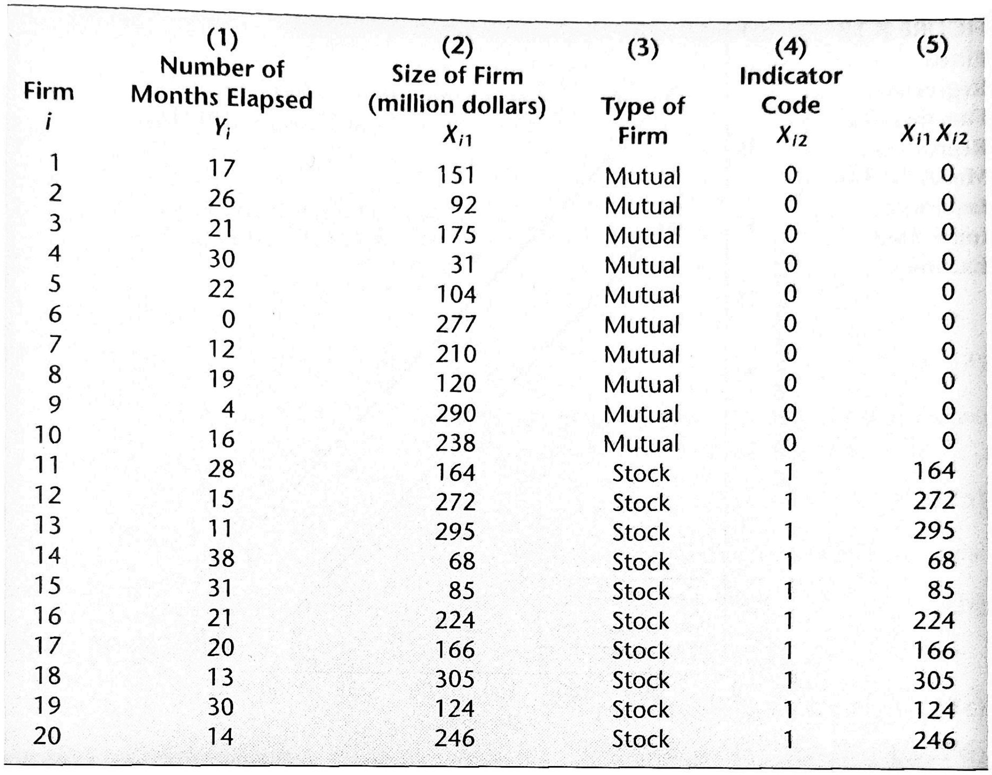
(a) Since the model assumes equal slopes and the same constant error term variance for each type of firm, the common slope β_1 can best be estimated by pooling the two types of firms.

(b) Also, other inferences, such as for β_0 and β_2, can be made more precisely by working with one regression model containing an indicator variable since more degrees of freedom will then be associated with MSE.

Example: the insurance innovation example

1. (Table 8.2) In the insurance innovation example, the economist studied 10 mutual firms and 10 stock firms Note that $X_2 = 1$ for each stock firm and $X_2 = 0$ for each mutual firm.

TABLE 8.2 Data and Indicator Coding—Insurance Innovation Example.



Firm i	(1) Number of Months Elapsed Y_i	(2) Size of Firm (million dollars) X_{i1}	(3) Type of Firm	(4) Indicator Code X_{i2}	(5) $X_{i1}X_{i2}$
1	17	151	Mutual	0	0
2	26	92	Mutual	0	0
3	21	175	Mutual	0	0
4	30	31	Mutual	0	0
5	22	104	Mutual	0	0
6	0	277	Mutual	0	0
7	12	210	Mutual	0	0
8	19	120	Mutual	0	0
9	4	290	Mutual	0	0
10	16	238	Mutual	0	0
11	28	164	Stock	1	164
12	15	272	Stock	1	272
13	11	295	Stock	1	295
14	38	68	Stock	1	68
15	31	85	Stock	1	85
16	21	224	Stock	1	224
17	20	166	Stock	1	166
18	13	305	Stock	1	305
19	30	124	Stock	1	124
20	14	246	Stock	1	246

2. (Table 8.3) The fitted response function is:

$$\hat{Y} = 33.87407 - 0.10174X_1 + 8.05547X_2$$

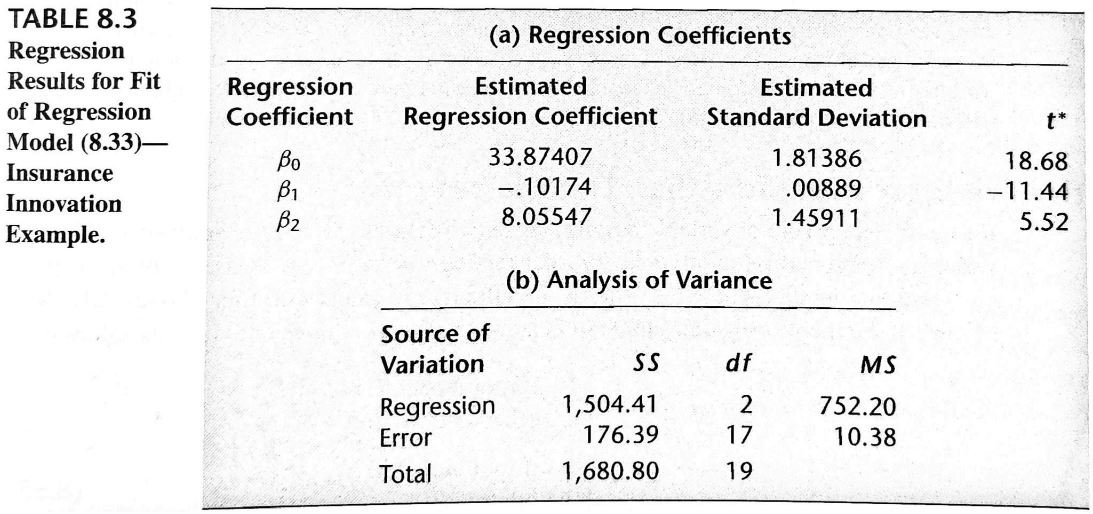


TABLE 8.3 Regression Results for Fit of Regression Model (8.33)—Insurance Innovation Example.

(a) Regression Coefficients

Regression Coefficient	Estimated Regression Coefficient	Estimated Standard Deviation	t^*
β_0	33.87407	1.81386	18.68
β_1	−.10174	.00889	−11.44
β_2	8.05547	1.45911	5.52

(b) Analysis of Variance

Source of Variation	SS	df	MS
Regression	1,504.41	2	752.20
Error	176.39	17	10.38
Total	1,680.80	19	

3. (Figure 8.12) contains the fitted response function for each type of firm, together with the actual observations.

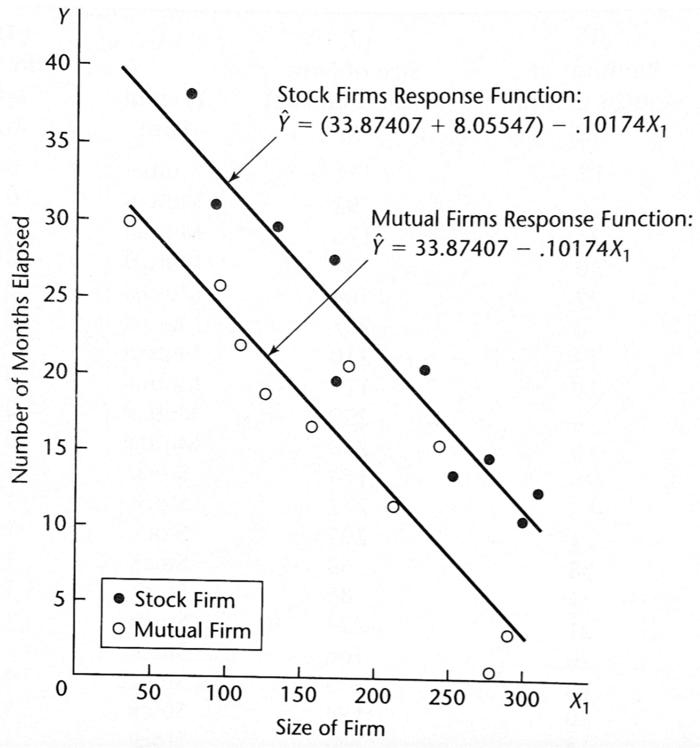


FIGURE 8.12 Fitted Regression Functions for Regression Model (8.33)—Insurance Innovation Example.

4. The economist was most interested in the effect of type of firm (X_2) on the elapsed time for the innovation to be adopted and wished to obtain a 95 percent confidence interval for β_2.

(a) We require $t_{(0.975;17)} = 2.110$ and obtain from the results in Table 8.3 the confidence limits $\underline{\quad 8.05547 \pm 2.110(1.45911) \quad}$.

(b) The confidence interval for β_2 therefore is:

$$4.98 \leq \beta_2 \leq 11.13$$

Thus, with 95 percent confidence, we conclude that stock companies tend to adopt the innovation somewhere between __5 and 11 months later__, on the average, than mutual companies __for any given size of firm__.

(c) A formal test of:

$$H_0 : \beta_2 = 0 \quad H_a : \beta_2 \neq 0$$

with level of significance 0.05 would lead to __H_a__, that type of firm has an effect, since the 95 percent confidence interval for β_2 __does not include zero__.

Qualitative Predictor with More than Two Classes

1. $\boxed{\text{Example}}$ Consider the regression of tool wear (Y) on tool speed (X_1) and tool model, where the latter is a qualitative variable with four classes ($M1, M2, M3, M4$):

$$X_2 = \begin{cases} 1 & \text{if tool model } M1 \\ 0 & \text{otherwise} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{if tool model } M2 \\ 0 & \text{otherwise} \end{cases}$$

$$X_4 = \begin{cases} 1 & \text{if tool model } M3 \\ 0 & \text{otherwise} \end{cases}$$

2. A first-order regression model:

$$\underline{\quad Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i \quad} \tag{8.36}$$

3. For this model, the data input for the X variables would be as follows:

Tool Model	X_1	X_2	X_3	X_4
$M1$	X_{i1}	1	0	0
$M2$	X_{i1}	0	1	0
$M3$	X_{i1}	0	0	1
$M4$	X_{i1}	0	0	0

4. The response function for regression model (8.36) is:

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \qquad (8.37)$$

(a) To understand the meaning of the regression coefficients, consider first what response function (8.37) becomes for tool models $M4$ for, which $X_2 = 0$, $X_3 = 0$, and $X_4 = 0$:

$$\underline{E\{Y\} = \beta_0 + \beta_1 X_1} \qquad \text{Tool models } M4 \qquad (8.37a)$$

(b) For tool models $M1$, $X_2 = 1$, $X_3 = 0$, and $X_4 = 0$, and response function (8.37) becomes:

$$\underline{E\{Y\} = (\beta_0 + \beta_2) + \beta_1 X_1} \qquad \text{Tool models } M1 \qquad (8.37b)$$

(c) Similarly, response functions (8.37) becomes for tool models M2 and M3:

$$\underline{E\{Y\} = (\beta_0 + \beta_3) + \beta_1 X_1} \qquad \text{Tool models } M2 \qquad (8.37c)$$

$$\underline{E\{Y\} = (\beta_0 + \beta_4) + \beta_1 X_1} \qquad \text{Tool models } M3 \qquad (8.37d)$$

(d) Response function (8.37) implies that the regression of tool wear on tool speed is linear, with the same slope for all four tool models.

(e) The coefficients β_2, β_3, and β_4 indicate, respectively, how much higher (lower) the response functions for tool models $M1$, $M2$, and $M3$ are than the one for, tool models $M4$, for any given level of tool speed.

(f) Thus, β_2, β_3, and β_4 measure the differential effects of the qualitative variable classes on the height of the response function for any given level of X_1, always compared with the class for which $X_2 = X_3 = X_4 = 0$.

(g) (Figure 8.13) we may wish to estimate differential effects other than against tool models M_4. $\beta_4 - \beta_3$ measures how much higher (lower) the response function for tool models M_3 is than the response function for tool models M_2 for any given level of tool speed, as may be seen by comparing (8.37c) and (8.37d). The point estimator of this quantity is, of course, $b_4 - b_3$, and the estimated variance of this estimator is:

$$s^2\{b_4 - b_3\} = s^2\{b_4\} + s^2\{b_3\} - 2s\{b_4, b_3\}. \qquad (8.38)$$

The needed variances and covariance can be readily obtained from the estimated variance-covariance matrix of the regression coefficients.

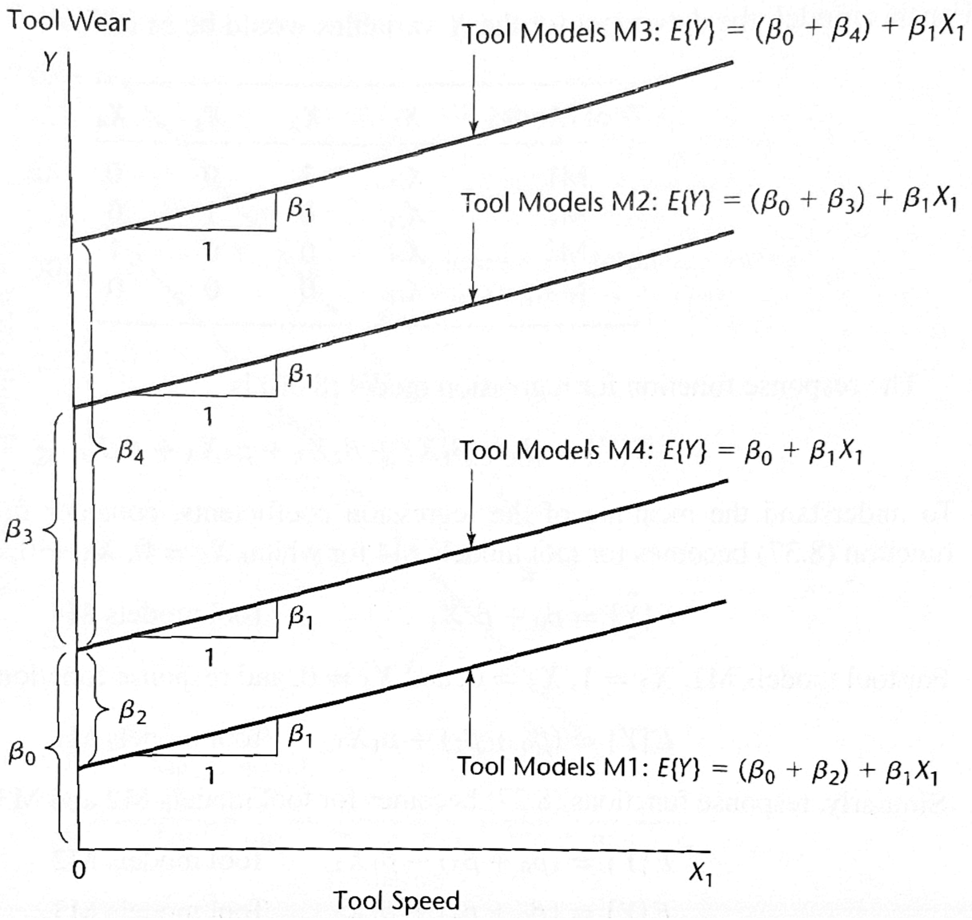

Tool Wear
Y
Tool Models M3: $E\{Y\} = (\beta_0 + \beta_4) + \beta_1 X_1$
Tool Models M2: $E\{Y\} = (\beta_0 + \beta_3) + \beta_1 X_1$
Tool Models M4: $E\{Y\} = \beta_0 + \beta_1 X_1$
Tool Models M1: $E\{Y\} = (\beta_0 + \beta_2) + \beta_1 X_1$
β_1
1
β_4
β_3
β_2
β_0
0
X_1
Tool Speed

FIGURE 8.13 Illustration of Regression Model (8.36)—Tool Wear Example.

8.4 Some Considerations in Using Indicator Variables*

8.5 Modeling Interactions between Quantitative and Qualitative Predictors

1. Example: the insurance innovation example The economist actually did not begin the analysis with regression model (8.33) because of the possibility of interaction effects between size of firm and type of firm on the response variable:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \qquad (8.33)$$

2. Even though one of the predictor variables in the regression model here is qualitative, interaction effects can still be introduced into the model in the usual manner, by including cross-product terms.

3. A first-order regression model with an added interaction term for the insurance innovation example is:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i \tag{8.49}$$

$$\begin{aligned} X_1 &= \text{size of firm} \\ X_2 &= \begin{cases} 1 & \text{if stock company} \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

4. The response function for this regression model is:

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 \tag{8.50}$$

Meaning of Regression Coefficients

1. (Figure 8.14) The meaning of the regression coefficients in response function (8.50) can best be understood by examining the nature of this function for each type of firm.

   (a) For a mutual firm, $X_2 = 0$ and hence $X_1 X_2 = 0$. Response function (8.50) therefore becomes for mutual firms:

   $$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2(0) + \beta_3(0) \qquad \text{Mutual films} \qquad (8.50a)$$

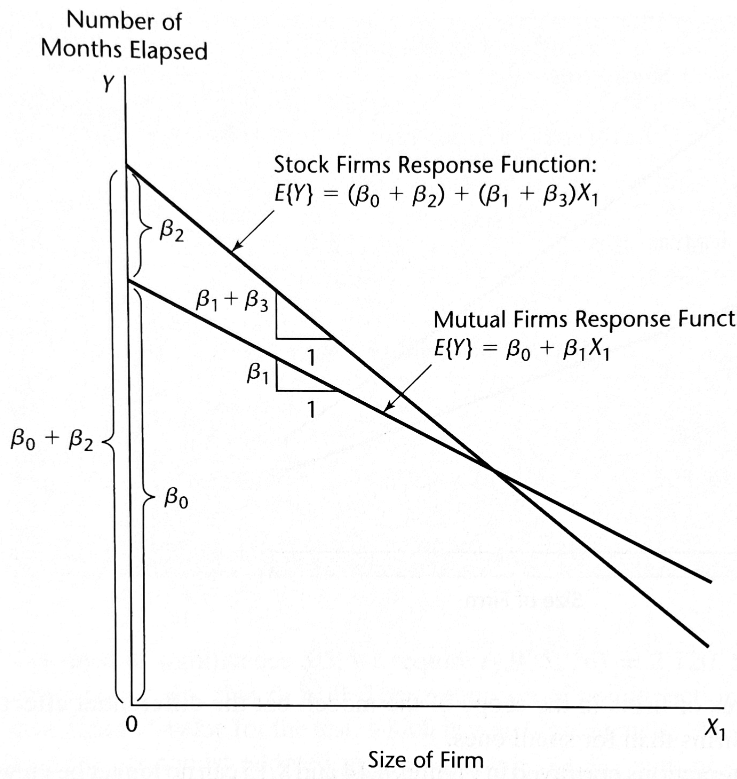
   (b) For stock firms, $X_2 = 1$ and hence $X_1 X_2 = 1$. Response function (8.50) therefore becomes for stock firms:

   $$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2(1) + \beta_3 X_1$$

   or

   $$E\{Y\} = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1 \qquad \text{Stock films} \qquad (8.50b)$$

FIGURE 8.14 Illustration of Meaning of Regression Coefficients for Regression Model (8.49) with Indicator Variable X_2 and Interaction Term—Insurance Innovation Example.



(c) β_2: indicates how much greater (smaller) is the Y intercept of the response function for the class coded 1 (stock firms) than that for the class coded 0 (mutual firms).

(d) β_3: indicates how much greater (smaller) is the slope of the response function for the class coded 1 than that for the class coded 0.

(e) (Figure 8.14) shows that the effect of type of firm with regression model (8.49) depends on X_1, the size of the firm.

i. For smaller firms, mutual firms tend to innovate more quickly.

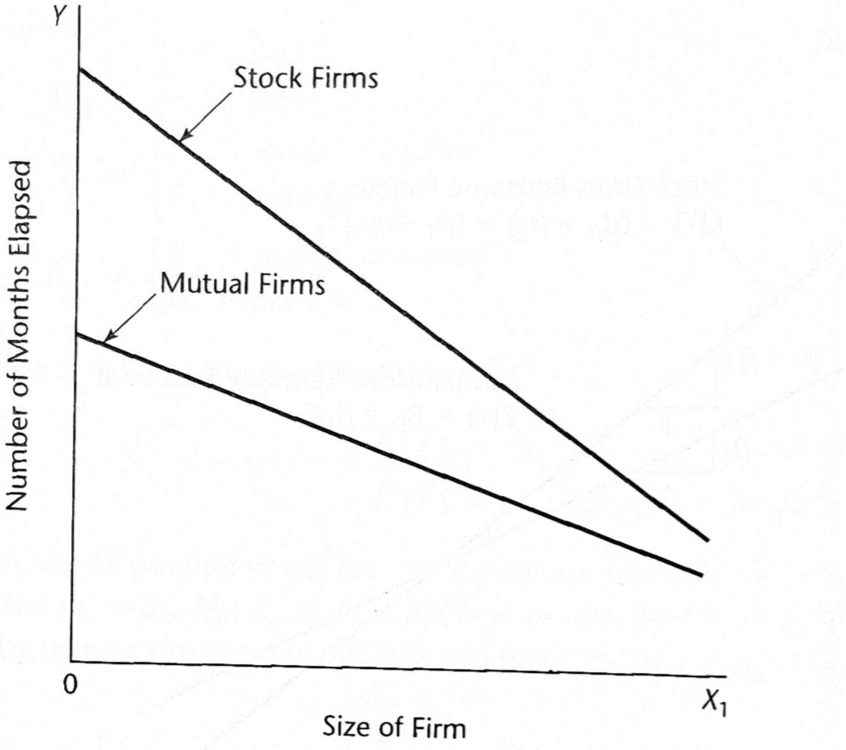
ii. For larger firms stock firms tend to innovate more quickly.

2. When interaction effects are present, the effect of the qualitative predictor variable can be studied only by comparing the regression functions within the scope of the model for the different classes of the qualitative variable.

3. (Figure 8.15) Another possible interaction pattern for the insurance innovation example. Here, mutual firms tend to introduce the innovation more quickly than stock

firms for all sizes of firms in the scope of the model. but the differential effect is much smaller for large firms than for small ones.

FIGURE 8.15 Another Illustration of Regression Model (8.49) with Indicator Variable X_2 and Interaction Term—Insurance Innovation Example.



4. When one of the predictor variables is qualitative and the other quantitative, nonparallel response functions that do not intersect within the scope of the model (as in Figure 8.15) are sometimes said to represent an ordinal interaction. When the response functions intersect within the scope of the model (as in Figure 8.14), the interaction is then said to be a disordinal interaction.

Example

1. Example: the insurance innovation example Since the economist was concerned that interaction effects between size and type of firm may be present, the initial regression model fitted was model (8.49):

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i$$

2. The values for the interaction term X_1X_2 for the insurance innovation example are shown in Table 8.2, column 5, on page 317. Note that this column contains 0 for mutual companies and X_{i1} for stock companies.

3. (Table 8.4) the regression results of Y on X_1, X_2 , and X_1X_2. To test for the presence of interaction effects:

$$H_0 : \beta_3 = 0, \quad H_a : \beta_3 \neq 0,$$

the economist used the t^* statistic from Table 8.4a:

$$t^* = \frac{b_3}{s\{b_3\}} = \frac{-0.0004171}{0.01833} = -0.02$$

4. For level of significance 0.05, we require $t_{(0.975;16)} = 2.120$. Since $|t^*| = 0.02 \le 2.120$, we conclude H_0, that $\beta_3 = 0$.

5. The conclusion of no interaction effects is supported by the two-sided p-value for the test, which is very high, 0.98.

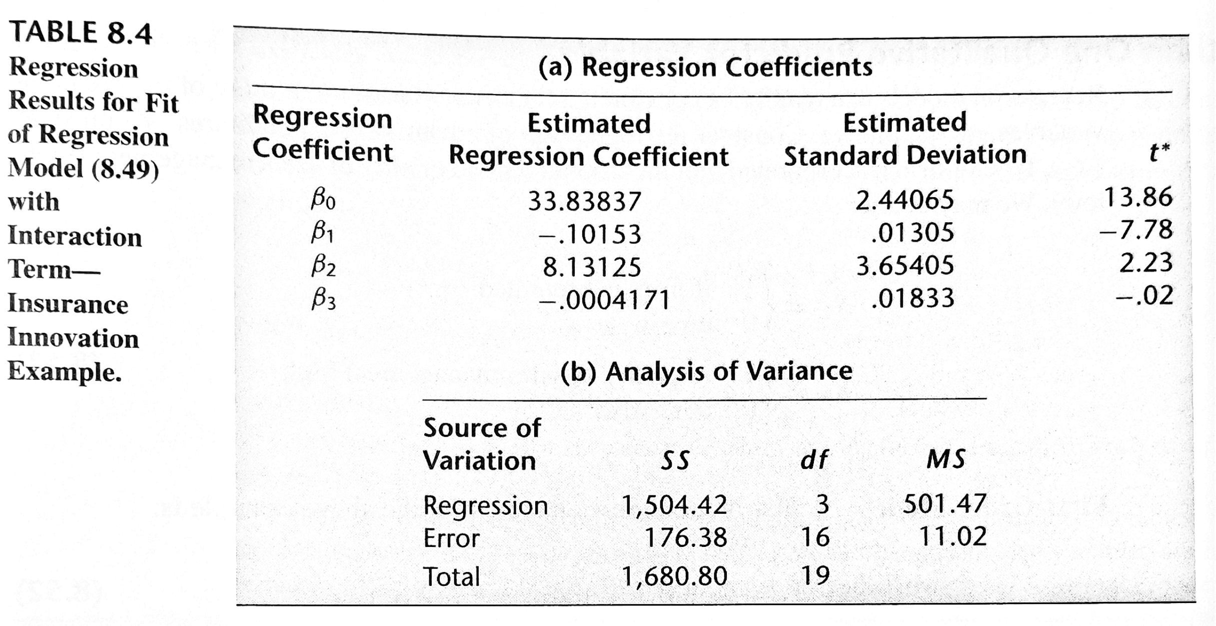


TABLE 8.4 Regression Results for Fit of Regression Model (8.49) with Interaction Term—Insurance Innovation Example.

(a) Regression Coefficients

Regression Coefficient	Estimated Regression Coefficient	Estimated Standard Deviation	t^*
β_0	33.83837	2.44065	13.86
β_1	−.10153	.01305	−7.78
β_2	8.13125	3.65405	2.23
β_3	−.0004171	.01833	−.02

(b) Analysis of Variance

Source of Variation	SS	df	MS
Regression	1,504.42	3	501.47
Error	176.38	16	11.02
Total	1,680.80	19	

8.6 More Complex Models*

8.7 Comparison of Two or More Regression Functions*

😍 TA Class

- **Problems**: 8.4, 8.5, 8.15, 8.21
- **Exercises**: 8.33, 8.34
- **Projects**: 8.39

“有時候壞事是注定要發生，而我們卻無能為力。那我們何必擔心呢?”
“Look, sometimes bad things happen —and there′ s nothing you can do about it. So why worry?”

— 獅子王 (The Lion King, 2019)

Regression Analysis (I)

Kutner's Applied Linear Statistical Models (5/E)

Chapter 9: Model Selection and Validation

Thursday 09:10-12:00, 商館 260205

Han-Ming Wu

Department of Statistics, National Chengchi University

http://www.hmwu.idv.tw

9.1 Overview of Model-Building Process

A strategy for the building of a regression model:

1. Data collection and preparation
2. Reduction of explanatory or predictor variables (for exploratory observational studies)
3. Model refinement and selection
4. Model validation

FIGURE 9.1
Strategy for Building a Regression Model.

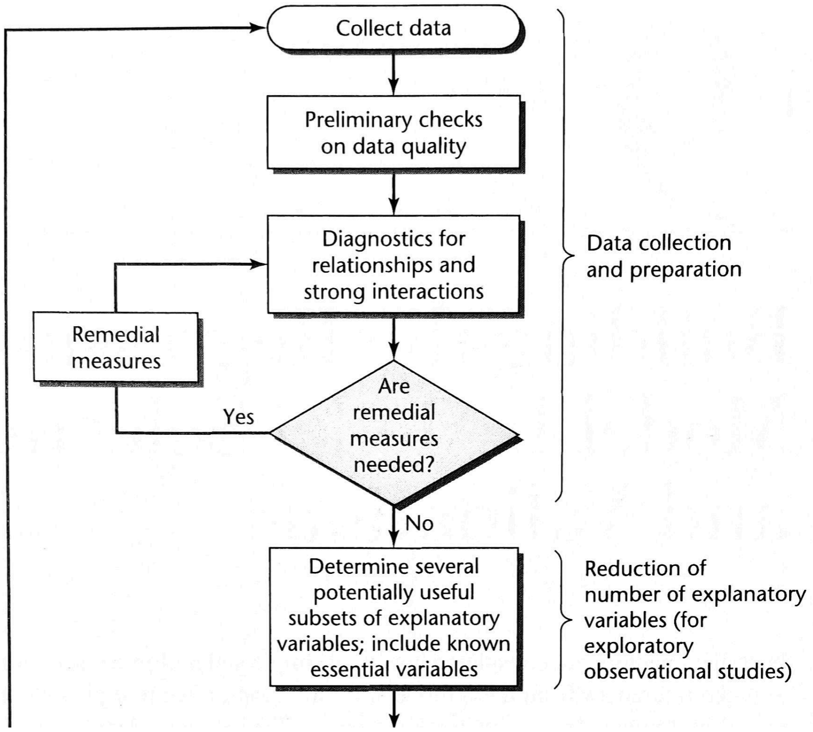
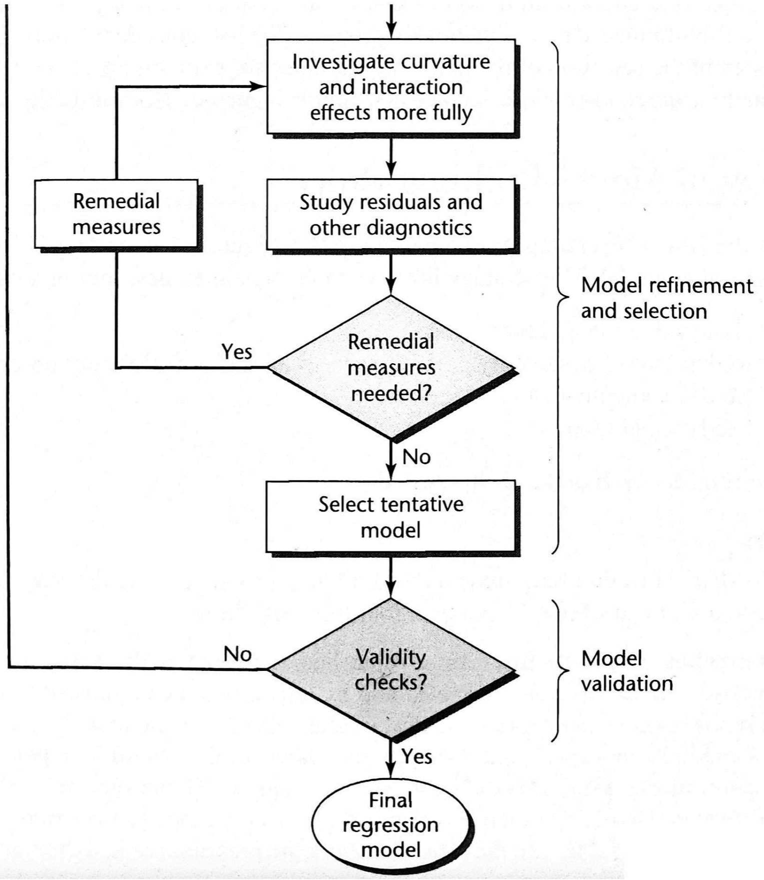

Collect data
Preliminary checks on data quality
Diagnostics for relationships and strong interactions
Remedial measures
Yes
Are remedial measures needed?
No
Data collection and preparation
Determine several potentially useful subsets of explanatory variables; include known essential variables
Reduction of number of explanatory variables (for exploratory observational studies)

FIGURE 9.1
Strategy for Building a Regression Model.


Investigate curvature and interaction effects more fully
Remedial measures
Study residuals and other diagnostics
Yes
Remedial measures needed?
No
Model refinement and selection
Select tentative model
No
Validity checks?
Model validation
Yes
Final regression model

9.2 Surgical Unit Example

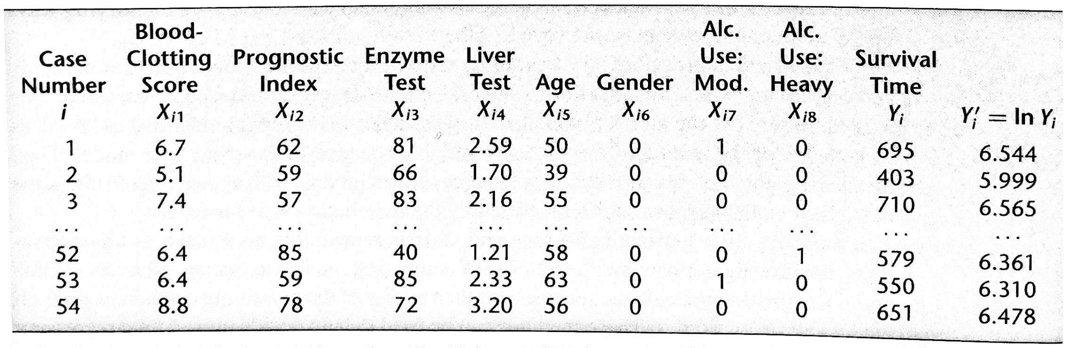
1. A hospital surgical unit was interested in predicting survival in patients undergoing a particular type of liver operation. A random selection of 108 patients was available for analysis. From each patient record, the following information was extracted from the pre-operation evaluation:

X_1	blood clotting score (血栓分數)
X_2	prognostic index (預後指數)
X_3	enzyme function test score (酶功能)
X_4	liver function test score (肝功能)
X_5	age, in years
X_6	indicator variable for gender (0 = male, 1 =female)
X_7, X_8	indicator variables for history of alcohol use: None: $X_7 = 0, X_8 = 0$, Moderate: $X_7 = 1, X_8 = 0$,Severe:$X_7 = 0, X_8 = 1$

2. These constitute the pool of potential explanatory or predictor variables for a predictive regression model.

3. (Table 9.1) The response variable Y is survival time, which was ascertained in a follow-up study. A portion of the data on the potential predictor variables and the response variable is presented in Table 9.1. These data have already been screened and properly edited for errors.

TABLE 9.1 Potential Predictor Variables and Response Variable—Surgical Unit Example.



Case Number i	Blood-Clotting Score X_{i1}	Prognostic Index X_{i2}	Enzyme Test X_{i3}	Liver Test X_{i4}	Age X_{i5}	Gender X_{i6}	Alc. Use: Mod. X_{i7}	Alc. Use: Heavy X_{i8}	Survival Time Y_i	$Y'_i = \ln Y_i$
1	6.7	62	81	2.59	50	0	1	0	695	6.544
2	5.1	59	66	1.70	39	0	0	0	403	5.999
3	7.4	57	83	2.16	55	0	0	0	710	6.565
...	...	...	...	...	...	...	...	...	...	...
52	6.4	85	40	1.21	58	0	0	1	579	6.361
53	6.4	59	85	2.33	63	0	1	0	550	6.310
54	8.8	78	72	3.20	56	0	0	0	651	6.478

4. To illustrate the model-building procedures discussed in this and the next section, we will use only the *first four explanatory variables.* We will also use only the *first 54 of the 108 patients.*

5. Since the pool of predictor variables is small, a reasonably full exploration of relationships and of possible strong interaction effects is possible at this stage of data preparation.

   (a) *Stem-and-leaf plots* for each of the predictor variables (not shown). These highlighted several cases as outlying with respect to the explanatory variables. The investigator was thereby alerted to examine later the influence of these cases.

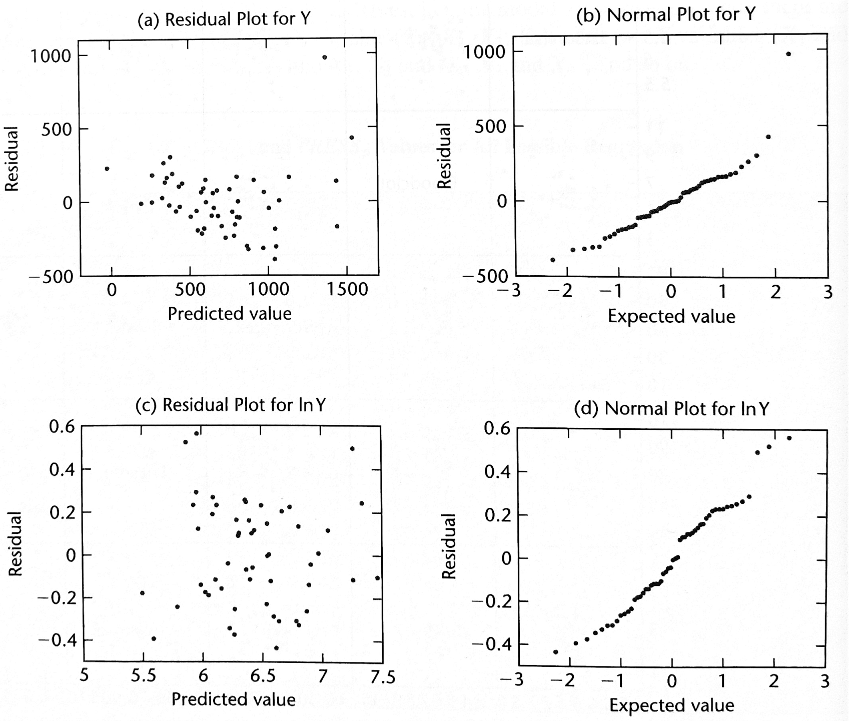
   (b) *A scatter plot matrix and the correlation matrix* (not shown)

6. *A first-order regression model* based on all predictor variables was fitted to serve as a starting point.

   (a) (Figure 9.2a) *A plot of residuals against predicted values* suggests that both curvature and nonconstant error variance are apparent.

   (b) (Figure 9.2b) *the normal probability plot* suggests some departure from normality.

FIGURE 9.2 Some Preliminary Residual Plots—Surgical Unit Example.



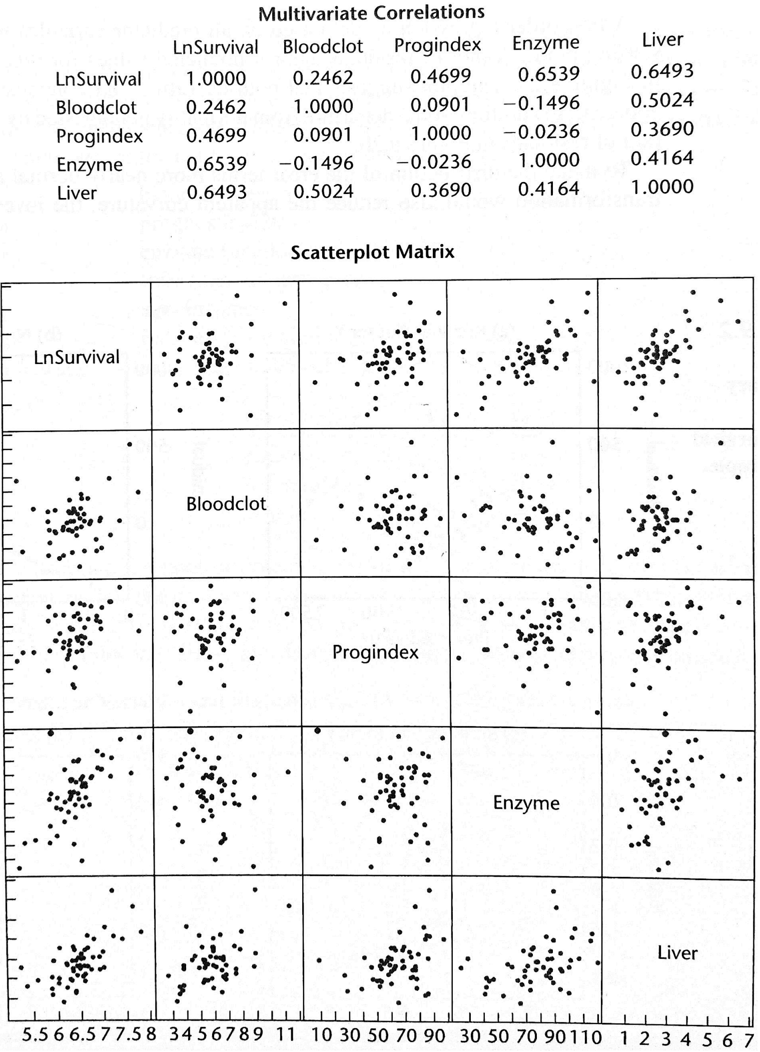
7. *Transformation*: To make the distribution of the error terms more nearly normal and to see if the same transformation would also reduce the apparent curvature, the investigator examined the logarithmic transformation $Y' = \ln(Y)$.

   (a) (Figure 9.2c) *A plot of residuals against fitted values* when Y' is regressed on all four predictor variables in a first-order model;

   (b) (Figure 9.2d) *The normal probability plot of residuals* for the transformed data shows that the distribution of the error terms is more nearly normal.

8. (Figure 9.3) *A scatter plot matrix and the correlation matrix* with the transformed Y variable.

FIGURE 9.3 **JMP Scatter Plot Matrix and Correlation Matrix when Response Variable Is Y'—Surgical Unit Example.**



Multivariate Correlations

	LnSurvival	Bloodclot	Progindex	Enzyme	Liver
LnSurvival	1.0000	0.2462	0.4699	0.6539	0.6493
Bloodclot	0.2462	1.0000	0.0901	−0.1496	0.5024
Progindex	0.4699	0.0901	1.0000	−0.0236	0.3690
Enzyme	0.6539	−0.1496	−0.0236	1.0000	0.4164
Liver	0.6493	0.5024	0.3690	0.4164	1.0000

Scatterplot Matrix

(a) Each of the predictor variables is linearly associated with Y', with X_3 and X_4 showing the highest degrees of association and X_1 the lowest.

(b) Show inter-correlations among the potential predictor variables. In particular, X_4 has moderately high pairwise correlations with X_1, X_2, and X_3

9. Various scatter and residual plots were obtained (not shown here).

10. On the basis of these analyses, the investigator concluded to use, at this stage of the model-building process, $Y' = \ln(Y)$ as the response variable, to represent the predictor variables in linear terms, and not to include any interaction terms.

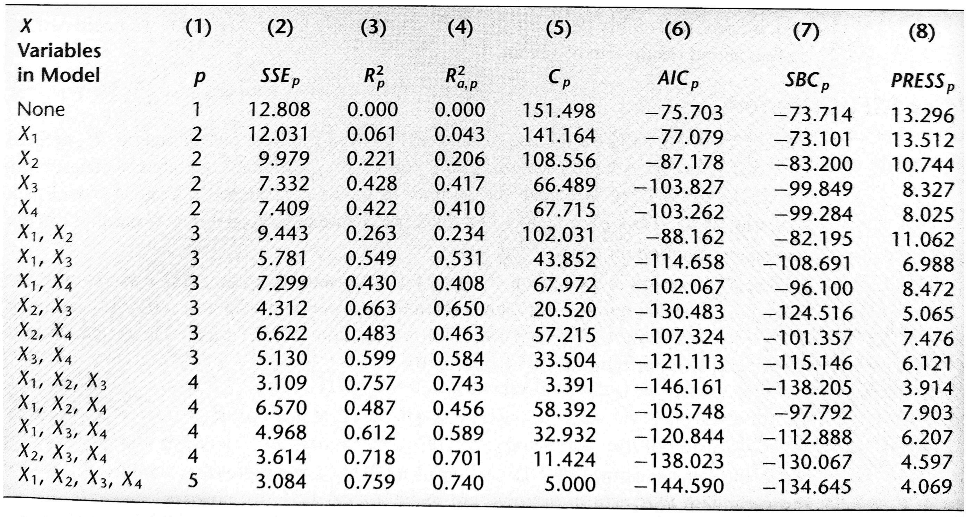
11. The next stage is to examine whether all of the potential predictor variables are needed or whether a subset of them is adequate.

9.3 Criteria for Model Selection

1. From any set of $p-1$ predictors, 2^{p-1} alternative models can be constructed. This calculation is based on the fact that each predictor can be either included or excluded from the model.

2. (Table 9.2) the $2^4 = 16$ different possible subset models that can be formed from the pool of four X variables in The Surgical Unit Example.

TABLE 9.2 SSE_p, R_p^2, $R_{a,p}^2$, C_p, AIC_p, SBC_p, and $PRESS_p$ Values for All Possible Regression Models—Surgical Unit Example.



X Variables in Model	(1) p	(2) SSE_p	(3) R_p^2	(4) $R_{a,p}^2$	(5) C_p	(6) AIC_p	(7) SBC_p	(8) $PRESS_p$
None	1	12.808	0.000	0.000	151.498	−75.703	−73.714	13.296
X_1	2	12.031	0.061	0.043	141.164	−77.079	−73.101	13.512
X_2	2	9.979	0.221	0.206	108.556	−87.178	−83.200	10.744
X_3	2	7.332	0.428	0.417	66.489	−103.827	−99.849	8.327
X_4	2	7.409	0.422	0.410	67.715	−103.262	−99.284	8.025
X_1, X_2	3	9.443	0.263	0.234	102.031	−88.162	−82.195	11.062
X_1, X_3	3	5.781	0.549	0.531	43.852	−114.658	−108.691	6.988
X_1, X_4	3	7.299	0.430	0.408	67.972	−102.067	−96.100	8.472
X_2, X_3	3	4.312	0.663	0.650	20.520	−130.483	−124.516	5.065
X_2, X_4	3	6.622	0.483	0.463	57.215	−107.324	−101.357	7.476
X_3, X_4	3	5.130	0.599	0.584	33.504	−121.113	−115.146	6.121
X_1, X_2, X_3	4	3.109	0.757	0.743	3.391	−146.161	−138.205	3.914
X_1, X_2, X_4	4	6.570	0.487	0.456	58.392	−105.748	−97.792	7.903
X_1, X_3, X_4	4	4.968	0.612	0.589	32.932	−120.844	−112.888	6.207
X_2, X_3, X_4	4	3.614	0.718	0.701	11.424	−138.023	−130.067	4.597
X_1, X_2, X_3, X_4	5	3.084	0.759	0.740	5.000	−144.590	−134.645	4.069

3. Model selection procedures, also known as subset selection or variables selection procedures, have been developed to identify a small group of regression models that are ”good” according to a specified criterion.

4. While many criteria for comparing the regression models have been developed, we will focus on six: R_p^2, $R_{a,p}^2$, C_p, AIC_p, SBC_p and $PRESS_p$.

5. We shall denote the number of potential X variables in the pool by $P-1$. We assume throughout this chapter that all regression models contain an intercept term β_0. Hence, the regression function containing all potential X variables contains P parameters, and the function with no X variables contains one parameter (β_0).

6. The number of X variables in a subset will be denoted by $p-1$, as always, so that there are p parameters in the regression function for this subset of X variables. Thus, we have: $1 \le p \le P$.

7. We will assume that the number of observations exceeds the maximum number of potential parameters: $n > p$.

R_p^2 or SSE_p Criterion

1. R_p^2 criterion calls for the use of the coefficient of multiple determination R^2:

$$R_p^2 = 1 - \frac{SSE_p}{SSTO}$$

2. R_p^2 indicates that there are p parameters, or $(p-1)$ X variables, in the regression function on which R_p^2 is based.

3. The R_p^2 criterion is equivalent to using the error sum of squares SSE_p as the criterion (we again show the number of parameters in the regression model as a subscript).

4. The R_p^2 criterion is not intended to identify the subsets that maximize this criterion.

5. We know that R_p^2 can never decrease as additional X variables are included in the model. Hence, R_p^2 will be a maximum when all $(P-1)$ potential X variables are included in the regression model.

6. The intent in using the R_p^2 criterion is to find the point where adding more X variables is not worthwhile because it leads to a very small increase in R_p^2.

7. Example The Surgical Unit Example

   (a) (Table 9.2, column 3) the R_p^2 values were obtained from a series of computer runs.

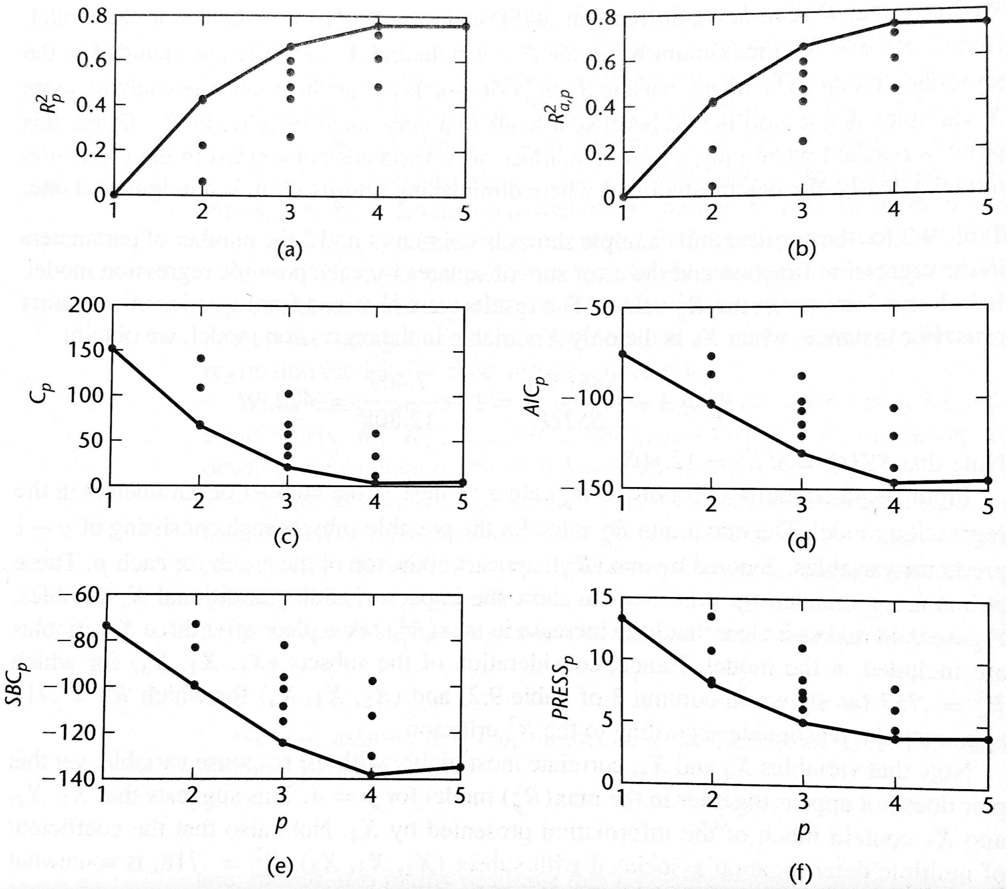
   (b) For instance, when X_4 is the only X variable in the regression model, we obtain:

$$R_2^2 = 1 - \frac{SSE(X_4)}{SSTO} = 1 - \frac{7.409}{12.808} = 0.422$$

   Note that $SSTO = SSE_1 = 12.808$

   (c) (Figure 9.4a) a plot of the R_p^2 values against p, the number of parameters in the regression model.

FIGURE 9.4 Plot of Variables Selection Criteria—Surgical Unit Example.



(d) The maximum R_p^2 value for the possible subsets each consisting of $p-1$ predictor variables, denoted by $\max(R_p)$, appears at the top of the graph for each p. These points are connected by solid lines to show the impact of adding additional X variables.

(e) (Figure 9.4a) little increase in $\max(R_p)$ takes place after three X variables are included in the model.

(f) Hence, consideration of the subsets (X_1, X_2, X_3) for which $R_4^2 = 0.757$ (as shown in column 3 of Table 9.2) and (X_2, X_3, X_4) for which $R_4^2 = 0.718$ appears to be reasonable according to the R_p^2 criterion.

(g) Note that variables X_3 and X_4, correlate most highly with the response variable, yet this pair does not appear together in the $\max(R_p^2)$ model for $p = 4$.

$R_{a,p}^2$ or MSE_p Criterion

1. Since R_p^2 does not take account of the number of parameters in the regression model and since $\max(R_p^2)$ can never decrease as p increases, the adjusted coefficient of multiple determination $R_{a,p}^2$ in (6.42) has been suggested as an alternative criterion:

$$R_{a,p}^2 = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE_p}{SSTO} = 1 - \frac{MSE_p}{SSTO/(n-1)} \tag{9.4}$$

2. It can be seeg from (9.4) that $R_{a,p}^2$ increases if and only if MSE_p decreases since $SSTO/(n-1)$ is fixed for the given Y observations. Hence, $R_{a,p}^2$ and MSE_p provide equivalent information.

3. The largest $R_{a,p}^2$ for a given number of parameters in the model, $\max(R_{a,p}^2)$, can, indeed, decrease as p increases.

4. Find a few subsets for which $R_{a,p}^2$ is at the maximum or so close to the maximum that adding more variables is not worthwhile.

5. $\boxed{\text{Example}}$ The Surgical Unit Example

   (a) (Table 9.2, column 4). For instance, we have for the regression model containing only X_4:

$$R_{a,2}^2 = 1 - \left(\frac{n-1}{n-2}\right)\frac{SSE(X_4)}{SSTO} = 1 - \left(\frac{53}{52}\right)\frac{7.409}{12.808} = 0.410$$

(b) (Figure 9.4b) The story told by the $R^2_{a,p}$ plot in Figure 9.4b is very similar to that told by the R^2_p plot in Figure 9.4a.

(c) Consideration of the subsets (X_1, X_2, X_3) and (X_2, X_3, X_4) appears to be reasonable according to the $R^2_{a,p}$ criterion.

(d) Notice that $R^2_{a,4} = 0.743$ is maximized for subset (X_1, X_2, X_3), and that adding X_4 to this subset – thus using all four predictors – decreases the criterion slightly: $R^2_{a,5} = 0.740$.

Mallows' C_p Criterion*

AIC_p and SBC_p Criteria

1. Two popular alternatives that also provide penalties for adding predictors are Akaike's (赤池) information criterion (AIC_p) and Schwarz' Bayesian criterion (SBC_p).

2. We search for models that have small values of AIC_p, or SBC_p:

$$AIC_p = n \ln SSE_p - n \ln n + 2p \tag{9.14}$$

$$SBC_p = n \ln SSE_p - n \ln n + (\ln n)p \tag{9.15}$$

3. Notice that for both of these measures, the first term is $n \ln SSE_p$ which decreases as p increases, The second term is fixed (for a given sample size n), and the third term increases with the number of parameters, p.

4. Models with small SSE_p will do well by these criteria as long as the penalties – $2p$ for AIC_p and $(\ln n)p$ for SBC_p – are not too large.

5. If $n \geq 8$ the penalty for SBC_p is larger than that for AIC_p.

6. Example The Surgical Unit Example

   (a) (Table 9.2, columns 6 and 7) When X_4 is the only X variable in the regression model:

$$\begin{aligned} AIC_2 &= n \ln SSE_2 - n \ln n + 2p \\ &= 54 \ln 7.409 - 54 \ln 54 + 2(2) = -103.262 \\ SBC_2 &= n \ln SSE_2 - n \ln n + (\ln n)p \\ &= 54 \ln 7.409 - 54 \ln 54 + (\ln 54)(2) = -99.284 \end{aligned}$$

(b) (Figures 9.4d, e) both of AIC_p and SBC_p criteria are minimized for subset (X_1, X_2, X_3).

$PRESS_p$ Criterion

1. The $PRESS_p$ (prediction sum of squares) criterion is a measure of how well the use of the fitted values for a subset model can predict the observed responses Y_i. The error sum of squares, $SSE = \sum(Y_i - \hat{Y}_i)^2$, is also such a measure.

2. The $PRESS$ measure differs from SSE in that each fitted value Y_i for the $PRESS$ criterion is obtained by deleting the ith case from the data set, estimating the regression function for the subset model from the remaining $n-1$ cases, and then using the fitted regression function to obtain the predicted value $\hat{Y}_{i(i)}$ for the ith case.

3. We use the notation $\hat{Y}_{i(i)}$ now for the fitted value to indicate, by the first subscript i, that it is a predicted value for the ith case and, by the second subscript (i), that the ith case was omitted when the regression function was fitted.

4. The $PRESS$ prediction error for the ith case then is:

$$Y_i - \hat{Y}_{i(i)} \tag{9.16}$$

   and the $PRESS_p$ criterion is the sum of the squared prediction errors over all n cases:

$$PRESS_p = \sum_{i=1}^{n}(Y_i - \hat{Y}_{i(i)})^2 \tag{9.17}$$

5. Models with small $PRESS_p$ values are considered good candidate models. The reason is that when the prediction errors $Y_i - \hat{Y}_{i(i)}$ are small, so are the squared prediction errors and the sum of the squared prediction errors.

6. Example The Surgical Unit Example

   (a) (Table 9.2, column 8)(Figure 9.4f) The message given by the $PRESS_p$ values in Table 9.2 and plot in Figure 9.4f is very similar to that told by the other criteria.

(b) We find that subsets (X_1, X_2, X_3) and (X_2, X_3, X_4) have small $PRESS$ values;

(c) The set of all X variables (X_1, X_2, X_3, X_4) involves a slightly larger $PRESS$ value than subset (X_1, X_2, X_3).

(d) The subset (X_2, X_3, X_4) involves a $PRESS$ value of 4.597, which is moderately larger than the $PRESS$ value of 3.914 for subset (X_1, X_2, X_3).

9.4 Automatic Search Procedures for Model Selection

1. The number of possible models, 2^{p-1}, grows rapidly with the number of predictors.

2. A variety of automatic computer-search procedures have been developed, e.g., ”best” subsets regression and stepwise regression.

”Best” Subsets Algorithms

1. Time-saving algorithms require the calculation of only a small fraction of all possible regression models.

2. For instance, the algorithms search for the five best subsets of X variables with the smallest C_p values using much less computational effort than when all possible subsets are evaluated. These algorithms are called ”best” subsets algorithms.

3. When the pool of potential X variables is very large, say greater than 30 or 40, even the ”best” subset algorithms may require excessive computer time.

4. As previously emphasized, our objective at this stage is not to identify a single best model; we hope to identify a small set of promising models for further study.

5. Example The Surgical Unit Example (eight predictors), we know there are $2^8 = 256$ possible models.

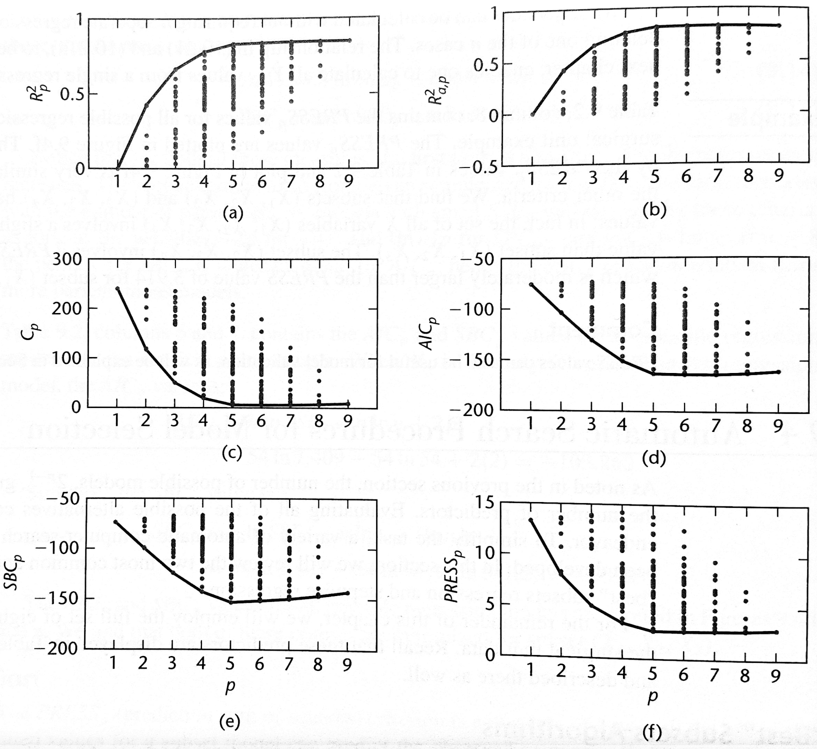


FIGURE 9.5 Plot of Variable Selection Criteria with All Eight Predictors—Surgical Unit Example.

(a) (Figure 9.5) Plots of the six model selection criteria. The best values of each criterion for each p have been connected with solid lines.

(b) (Table 9.3) The overall optimum criterion values have been underlined in each column of the table.

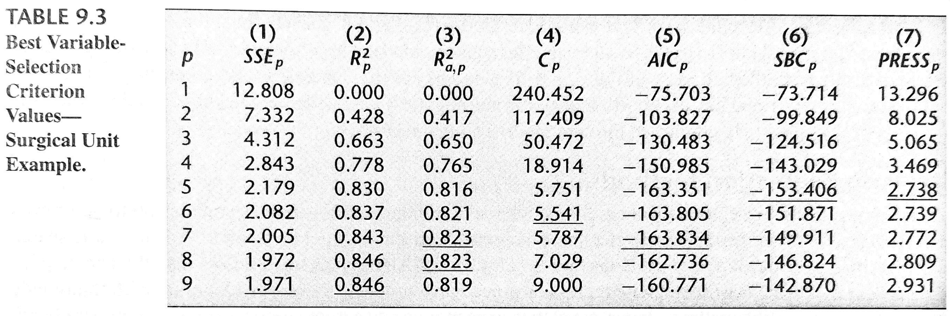


TABLE 9.3 Best Variable-Selection Criterion Values—Surgical Unit Example.

p	(1) SSE_p	(2) R_p^2	(3) $R_{a,p}^2$	(4) C_p	(5) AIC_p	(6) SBC_p	(7) $PRESS_p$
1	12.808	0.000	0.000	240.452	−75.703	−73.714	13.296
2	7.332	0.428	0.417	117.409	−103.827	−99.849	8.025
3	4.312	0.663	0.650	50.472	−130.483	−124.516	5.065
4	2.843	0.778	0.765	18.914	−150.985	−143.029	3.469
5	2.179	0.830	0.816	5.751	−163.351	−153.406	2.738
6	2.082	0.837	0.821	5.541	−163.805	−151.871	2.739
7	2.005	0.843	0.823	5.787	−163.834	−149.911	2.772
8	1.972	0.846	0.823	7.029	−162.736	−146.824	2.809
9	1.971	0.846	0.819	9.000	−160.771	−142.870	2.931

(c) For example

i. a 7-or 8-parameter model is identified as best by the $R^2_{a,p}$ criterion (both have $\max(R^2_{a,p}) = 0.823$)

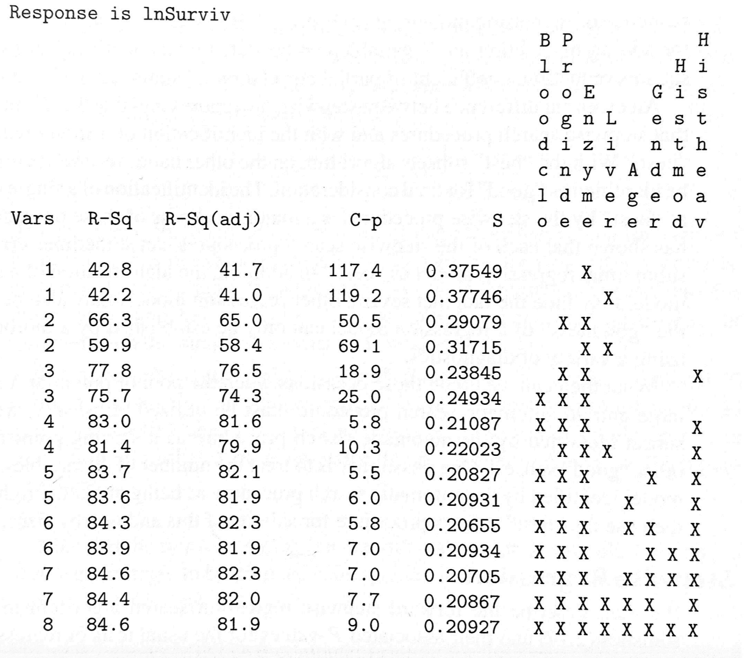
ii. a 6-parameter model is identified by the C_p criterion ($\min(C_7) = 5.541$),

iii. a 7-parameter model is identified by the AIC_p criterion ($\min(AIC_7) = -163.834$).

iv. Both the SBC_p and $PRESS_p$ criteria point to 5-parameter models ($\min(SBC_5) = -153.406$ and $\min(PRESS_5) = 2.738$).

(d) (Figure 9.6) MINITAB output for the "best" subsets algorithm. We specified that the best two subsets be identified for each number of variables in the regression model.

FIGURE 9.6 MINITAB Output for "Best" Two Subsets for Each Subset Size—Surgical Unit Example.



```
Response is lnSurviv

                                       B P           H
                                       l r         H i
                                       o o E     G i s
                                       o g n L   e s t
                                       d i z i   n t h
                                       c n y v A d m e
                                       l d m e g e o a
Vars  R-Sq  R-Sq(adj)    C-p        S  o e e r e r d v

   1  42.8       41.7  117.4  0.37549      X
   1  42.2       41.0  119.2  0.37746        X
   2  66.3       65.0   50.5  0.29079    X X
   2  59.9       58.4   69.1  0.31715      X X
   3  77.8       76.5   18.9  0.23845    X X         X
   3  75.7       74.3   25.0  0.24934  X X X
   4  83.0       81.6    5.8  0.21087  X X X         X
   4  81.4       79.9   10.3  0.22023    X X X       X
   5  83.7       82.1    5.5  0.20827  X X X     X   X
   5  83.6       81.9    6.0  0.20931  X X X   X     X
   6  84.3       82.3    5.8  0.20655  X X X   X X   X
   6  83.9       81.9    7.0  0.20934  X X X     X X X
   7  84.6       82.3    7.0  0.20705  X X X   X X X X
   7  84.4       82.0    7.7  0.20867  X X X X X X   X
   8  84.6       81.9    9.0  0.20927  X X X X X X X X
```

(e) The MINITAB algolithm uses the R_p^2 criterion, but also shows for each of the "best" subsets the $R_{a,p}^2$, C_p, and $\sqrt{MSE_p}$ (labeled S) values. The right-most columns of the tabulation show the X variables in the subset.

(f) According to the $R_{a,p}^2$ criterion, the 7-parameter model based on all predictors except Liver (X_4) and Histmod (history of moderate alcohol use X_7), or the 8-parameter model based on all predictors except Liver (X_4) are best.

(g) The $R_{a,p}^2$ criterion value for both of these models is 0.823.

6. The all-possible-regressions procedure leads to the identification of a small number of subsets that are "good" according to a specified criterion.

7. Consequently, one may wish at times to consider more than one criterion in evaluating possible subsets of X variables.

8. Once the investigator has identified a few "good" subsets for intensive examination, a final choice of the model variables must be made. This choice is aided by residual analyses (and other diagnostics to be covered in Chapter 10) and by the investigator's knowledge of the subject under study, and is finally confirmed through model validation studies.

Stepwise Regression Methods

1. When the pool of potential X variables contains 30 to 40 or even more variables, use of a "best" subsets algorithm may not be feasible.

2. An automatic search procedure that develops the "best" subset of X variables sequentially may then be helpful. The forward stepwise regression procedure is probably the most widely used of the automatic search methods.

3. Essentially, the forward stepwise search method develops a sequence of regression models, at each step adding or deleting an X variable. The criterion for adding or deleting an X variable can be stated equivalently in terms of error sum of squares reduction, coefficient of partial correlation, t^* statistic, or F^* statistic.

4. An essential difference between stepwise procedures and the "best" subsets algorithm is that stepwise search procedures end with the identification of a single regression model as "best." With the "best" subsets algorithm, several regression models can be identified as "good" for final consideration.

Forward Stepwise Regression

We shall describe the forward stepwise regression search algorithm in terms of the t^* statistics (2.17) and their associated P-values for the usual tests of regression parameters.

1. The stepwise regression routine first fits a simple linear regression model for each of the $p-1$ potential X variables. For each SLR model, the t^* statistic for testing whether or not the slope is zero is obtained:

$$t_k^* = \frac{b_k}{s\{b_k\}}$$

(a) The X with the largest t^* value is the candidate for first addition. If this t^* value exceeds a predetermined level, or if the corresponding P-value is less than a predetermined α, the X variable is added.

(b) Otherwise, the program terminates with no X variable considered sufficiently helpful to enter the regression model.

2. Assume X_7 is the variable entered at step 1. The stepwise regression routine now fits all regression models with two X variables, where X_7 is one of the pair.

(a) For each such regression model, the t^* test statistic corresponding to the newly added predictor X_k is obtained.

(b) This is the statistic for testing whether or not $\beta_k = 0$ when X_7 and X_k are the variables in the model.

(c) The X variable with the largest t^* value-or equivalently, the smallest P-value is the candidate for addition at the second stage.

(d) If this t^* value exceeds a predetermined level (i.e., the P-value falls below a predetermined level), the second X variable is added. Otherwise, the program terminates.

3. Suppose X_3 is added at the second stage. Now the stepwise regression routine examines whether any of the other X variables already in the model should be dropped.

(a) There is at this stage only one other X variable in the model, X_7 , so that only one t^* test statistic is obtained:

$$t_7^* = \frac{b_7}{s\{b_7\}}$$

(b) At later stages, there would be a number of these t^* statistics, one for each of the variables in the model besides the one last added.

(c) The variable for which this t^* value is smallest (or equivalently the variable for which the P-value is largest) is the candidate for deletion.

(d) If this t^* value falls below-or the P-value exceeds-a predetermined limit, the variable is dropped from the model; otherwise, it is retained.

4. Suppose X_7 is retained so that both X_3 and X_7 are now in the model.

   (a) The stepwise regression routine now examines which X variable is the next candidate for addition.

   (b) Then examines whether any of the variables already in the model should now be dropped.

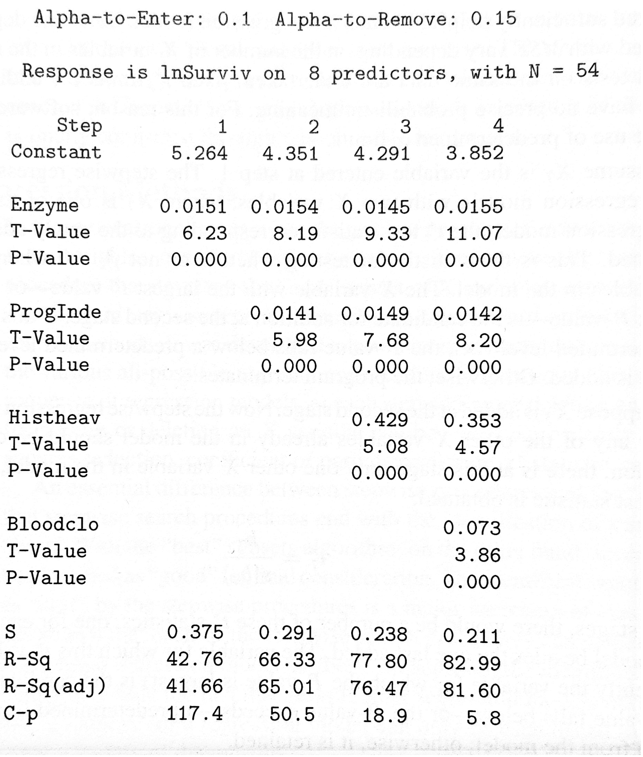
   (c) And so on until no further X variables can either be added or deleted, at which point the search terminates.

5. Note that the stepwise regression algorithm allows an X variable, brought into the model at an earlier stage, to be dropped subsequently if it is no longer helpful in conjunction with variables added at later stages.

Example

(Figure 9.7) MINITAB computer printout for the forward stepwise regression procedure for The Surgical Unit Example. The maximum acceptable a limit for adding a variable is 0.10 and the minimum acceptable a limit for removing a variable is 0.15.

FIGURE 9.7 MINITAB Forward Stepwise Regression Output— Surgical Unit Example.



```
  Alpha-to-Enter: 0.1  Alpha-to-Remove: 0.15

 Response is lnSurviv on  8 predictors, with N = 54

     Step          1       2       3       4
 Constant      5.264   4.351   4.291   3.852

 Enzyme       0.0151  0.0154  0.0145  0.0155
 T-Value        6.23    8.19    9.33   11.07
 P-Value       0.000   0.000   0.000   0.000

 ProgInde             0.0141  0.0149  0.0142
 T-Value                5.98    7.68    8.20
 P-Value               0.000   0.000   0.000

 Histheav                      0.429   0.353
 T-Value                        5.08    4.57
 P-Value                       0.000   0.000

 Bloodclo                              0.073
 T-Value                                3.86
 P-Value                               0.000

 S             0.375   0.291   0.238   0.211
 R-Sq          42.76   66.33   77.80   82.99
 R-Sq(adj)     41.66   65.01   76.47   81.60
 C-p           117.4    50.5    18.9     5.8
```

1. At the start of the stepwise search, no X variable is in the model so that the model to be fitted is $Y_i = \beta_0 + \epsilon_i;$.

   (a) (Step 1), the t^* statistics and corresponding P-values are calculated for each potential X variable, and the predictor having the smallest P-value (largest t^* value) is chosen to enter the equation.

   (b) Enzyme (X_3) had the largest test statistic:

   $$t_3^* = \frac{b_3}{s\{b_3\}} = \frac{0.015124}{0.002427} = 6.23.$$

   (c) The P-value for this test statistic is 0.000, which falls below the maximum acceptable α-to-enter value of 0.10; hence Enzyme (X_3) is added to the model.

   (d) The current regression model contains Enzyme (X_3), "Step 1": the regression coefficient for Enzyme (0.0151).

   (e) At the bottom of column 1, a number of variables-selection criteria, including $R_1^2(42.76)$, $R_{a,1}^2(41.66)$, and $C_1(117.4)$ are also provided.

2. Next, all regression models containing X_3 and another X variable are fitted, and the t^* statistics calculated:

   $$t_k^* = \sqrt{\frac{MSR(X_k|X_3)}{MSE(X_3, X_k)}}, \quad \text{since} \quad F^* = \frac{MSR}{MSE}, \quad F^* = (t^*)^2$$

   Progindex (X_2) has the highest t^* value, and its P-value (0.000) falls below 0.10, so that X_2 now enters the model.

3. Enzyme and Progindex (X_3 and X_2) are now in the model. At this point, a test whether Enzyme (X_3) should be dropped is undertaken, but because the P-value (0.000) corresponding to X_3 is not above 0.15, this variable is retained.

4. Next, all regression models containing X_2, X_3, and one of the remaining potential X variables are fitted. The appropriate t^* statistics:

   $$t_k^* = \sqrt{\frac{MSR(X_k|X_2, X_3)}{MSE(X_2, X_3, X_k}}$$

   The predictor labeled Histheavy (X_8) had the largest t^* value, (P-value = 0.000) and was next added to the model. X_2, X_3 , and X_8 are now in the model.

5. Next, a test is undertaken to determine whether X_2 or X_3 should be dropped. Since both of the corresponding P-values are less than 0.15, neither predictor is dropped from the model.

6. (Step 4) Bloodclot (X_1) is added, and no terms previously included were dropped. The right-most column of Figure 9.7 summarizes the addition of variable X_1 into the model containing variables X_2, X_3, and X_8.

7. Next, a test is undertaken to determine whether either X_2 , X_3 , or X_8 should be dropped. Since all P-values are less than 0.15 (all are 0.0(0), all variables are retained.

8. Finally, the stepwise regression routine considers adding one of X_4 , X_5 , X_6 , or X_7 to the model containing X_1, X_2, X_3, and X_8. In each case, the P-values are greater than 0.10 (not shown); therefore, no additional variables can be added to the model and the search process is terminated.

9. Thus, the stepwise search algorithm identifies (X_1, X_2, X_3, X_8) as the ”best” subset of X variables. This model also happens to be the model identified by both the SBC_p and $PRESS_p$ criteria in our previous analyses based on an assessment of ”best” subset selection.

Other Stepwise Procedures

1. *Forward Selection*. The forward selection search procedure is a simplified version of forward stepwise regression, omitting the test whether a variable once entered into the model should be dropped.

2. *Backward Elimination*. The backward elimination search procedure is the opposite of forward selection.

   (a) It begins with the model containing all potential X variables and identifies the one with the largest P-value.

   (b) If the maximum P-value is greater than a predetermined limit, that X variable is dropped.

(c) The model with the remaining $(P-2)$ X variables is then fitted, and the next candidate for dropping is identified.

(d) This process continues until no further X variables can be dropped.

9.5 Some Final Comments on Automatic Model Selection Procedures*

9.6 Model Validation

1. The final step in the model-building process is the ___validation___ of the selected regression models.

2. Model validation usually involves checking a ___candidate model___ against ___independent data___. Three basic ways of validating a regression model are:

   (a) Collection of ___new data___ to check the model and its predictive ability.

   (b) ___Comparison___ of results with theoretical expectations, earlier empirical results, and simulation results.

   (c) Use of a ___holdout sample___ to check the model and its ___predictive ability___.

3. What is difference between: training set, testing set and hold-out set: (The training set is for ___model-building___)

   (a) A observed data set (100%): e.g, training set (75%), testing set (25%).

   (b) A observed data set (100%): k-fold cross validation: e.g, $k = 4$ (25%, 25%, 25%, 25%), in turns "testing set (25%), training set (75%)" 4 times.

   (c) A observed data set (100%): hold-out set (20%), Not hold-out set (80% for 4-fold CV)

Collection of New Data to Check Model

1. The ___best___ means of model validation is through the ___collection of new data___. The purpose of collecting new data is to be able to examine whether the regression model developed from the earlier data is still ___applicable for the new data___. If

so, one has assurance about the applicability of the model to data beyond tho,se on which the model is based.

Methods of Checking Validity. A means of measuring the actual predictive capability of the selected regression model is to use this model to predict each case in the new data set and then to calculate the mean of the squared prediction errors, to be denoted by $MSPR$, which stands for mean squared prediction error:

$$MSPR = \frac{\sum_{i=1}^{n^*}(Y_i - \hat{Y}_i)^2}{n^*}$$

where:

- Y_i is the value of the response variable in the ith validation case.
- $\hat{Y}_i$ is the predicted value for the ith validation case based on the model-building dataset.
- n^* is the number of cases in the validation data set.

2. If the mean squared prediction error $MSPR$ is fairly close to MSE based on the regression fit to the model-building data set, then the error mean square MSE for the selected regression model is not seriously biased and gives an appropriate indication of the predictive ability of the model.

3. If the mean squared prediction error is much larger than MSE, one should rely on the mean squared prediction error as an indicator of how well the selected regression model will predict in the future.

😍 TA Class

- **Problems**: 9.6, 9.11, 9.18, 9.21
- **Exercises**: none
- **Projects**: none

“對一個不滿意的人生你只有兩種選擇，強迫自己接受，或說服自己改變。”
“You can only do one of two things to an unsatisfying life: force yourself to accept it, or convince yourself to change.”

— *媽的多重宇宙 (Everything Everywhere All at Once, 2022)*

Regression Analysis (I)

Kutner's Applied Linear Statistical Models (5/E)

Chapter 14: Logistic Regression

Thursday 09:10-12:00, 商館 260205

Han-Ming Wu

Department of Statistics, National Chengchi University

http://www.hmwu.idv.tw

14.1 Regression Models with.Binary Response Variable*

14.2 Sigmoidal Response Functions for Binary Responses*

14.3 Simple Logistic Regression

1. If X is a random variable with Bernoulli distribution, then

$$P(X=1)=\pi=1-P(X=0)$$

and the probability mass function of this distribution

$$f_X(k,\pi)=\pi^k(1-\pi)^{1-k}, k\in\{0,1\}.$$

2. The logit is the logarithm of the odds, where π = probability of a positive outcome (e.g., survived Titanic sinking)

$$\text{logit}(\pi)=\log\left(\frac{\pi}{1-\pi}\right).$$

3. A formal statement of the simple logistic regression model : recall that when the response variable is binary , taking on the values 1 and 0 with probabilities π and $1-\pi$, respectively, Y is a Bernoulli random variable with parameter $E\{Y\}=\pi$.

4. We could state the simple logistic regression model in the usual form:

$$Y_i = E\{Y_i\} + \varepsilon_i$$

5. Since the distribution of the error term ε_i depends on the Bernoulli distribution of the response Y_i, it is preferable to state the simple logistic regression model as: Y_i are independent Bernoulli random variables with expected values:

$$E\{Y_i\} = \pi_i = \frac{\exp(\beta_0+\beta_1 X_i)}{1+\exp(\beta_0+\beta_1 X_i)} \quad . \qquad (14.20)$$

6. The X observations are assumed to be known constants . Alternatively, if the X observations are random, $E\{Y_i\}$ is viewed as a conditional mean , given the value of X_i.

Likelihood Function

1. Since each Y_i observation is an ordinary Bernoulli random variable, where:

$$P(Y_j = 1) = \pi_i; \quad P(Y_j = 0) = 1 - \pi_i; \quad i = 1, \cdots, n.$$

we can represent its probability distribution as follows:

$$f_i(Y_i) = \pi_i^{Y_i}(1-\pi_i)^{1-Y_i} \quad , \quad Y_i = 0, 1; \quad i = 1, \cdots, n. \qquad (14.21)$$

Note that $f_i(1) = \pi_i$ and $f_i(0) = 1 - \pi_i$. Hence, $f_i(Y_i)$ simply represents the probability that $Y_i = 1$ or 0.

2. Since the Y_i observations are independent, their joint probability function is:

$$g(Y_1, \cdots, Y_n) = \prod_{i=1}^{n} f_i(Y_i) = \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i} \quad . \tag{14.22}$$

3. Find the maximum likelihood estimates by working with the logarithm of the joint probability function:

$$\begin{aligned} \ln g(Y_1, \cdots, Y_n) &= \ln \prod_{i=1}^{n} f_i(Y_i) \\ &= \sum_{i=1}^{n} [Y_i \ln \pi_i + (1-Y_i)\ln(1-\pi_i)] \\ &= \sum_{i=1}^{n} \left[Y_i \ln \left(\frac{\pi_i}{1-\pi_i} \right) \right] + \sum_{i=1}^{n} \ln(1-\pi_i) \quad . \end{aligned}$$

4. Since $E\{Y_i\} = \pi_i$; for a binary variable, it follows from (14.20) that:

$$1 - \pi_i = [1 + \exp(\beta_0 + \beta_1 X_i)]^{-1} \tag{14.24}$$

5. Furthermore, from (14.l8a), we obtain:

$$\ln \left(\frac{\pi_i}{1-\pi_i} \right) = \beta_0 + \beta_1 X_i \tag{14.25}$$

6. Hence, log likelihood (14.23) can be expressed as follows:

$$\ln L(\beta_0, \beta_1) = \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 X_i) - \sum_{i=1}^{n} \ln[1 + \exp(\beta_0 + \beta_1 X_i)] \tag{14.26}$$

where $L(\beta_0, \beta_1)$ replaces $g(Y_1, \cdots, Y_n)$ to show explicitly that we now view this function as the likelihood function of the parameters to be estimated, given the sample observations.

Maximum Likelihood Estimation

1. The maximum likelihood estimates of β_0 and β_1 in the simple logistic regression model are those values of β_0 and β_1 that maximize the log-likelihood function in (14.26).

2. No closed-form solution exists for the values of β_0 and β_1, in (4.26) that maximize the log-likelihood function. Computer-intensive numerical search procedures are therefore required to find the maximum likelihood estimates b_0 and b_1.

3. Once the maximum likelihood estimates b_0 and b_1 are found, we substitute these values into the response function in (14.20) to obtain the fitted response function. We shall use π_i to denote the fitted value for the ith case:

$$\hat{\pi}_i = \frac{\exp(b_0 + b_1 X_i)}{1 + \exp(b_0 + b_1 X_i)}.$$

4. The fitted logistic response function is as follows:

$$\hat{\pi} = \frac{\exp(b_0 + b_1 X)}{1 + \exp(b_0 + b_1 X)}$$

5. If we utilize the logit transformation in (14.18), we can express the fitted response function in (14.28) as follows:

$$\hat{\pi}' = b_0 + b_1 X, \qquad \hat{\pi}' = \ln\left(\frac{\hat{\pi}}{1 - \hat{\pi}}\right) \tag{14.29}$$

   We call (14.29) the fitted logit response function.

6. Once the fitted logistic response function has been obtained, the usual next steps are to examine the appropriateness of the fitted response function and, if the fit is good, to make a variety of inferences and predictions.

7. We shall postpone a discussion of how to examine the goodness of fit of a logistic response function and how to make inferences and predictions until we have considered the multiple logistic regression model with a number of predictor variables.

Example

1. A systems analyst studied the effect of computer programming experience on ability to complete within a specified time a complex programming task, including debugging. Twenty-five persons were selected for the study. They had varying amounts of programming experience (measured in months of experience), as shown in Table 14.1a column 1.

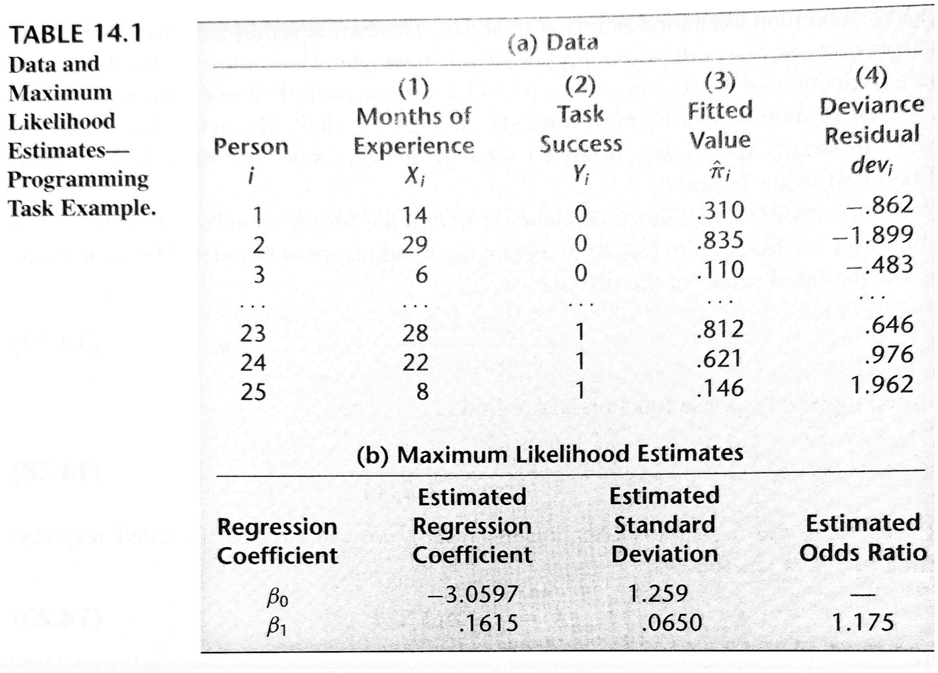


TABLE 14.1 Data and Maximum Likelihood Estimates—Programming Task Example.

(a) Data

Person i	(1) Months of Experience X_i	(2) Task Success Y_i	(3) Fitted Value $\hat{\pi}_i$	(4) Deviance Residual dev_i
1	14	0	.310	−.862
2	29	0	.835	−1.899
3	6	0	.110	−.483
...	...	...	...	...
23	28	1	.812	.646
24	22	1	.621	.976
25	8	1	.146	1.962

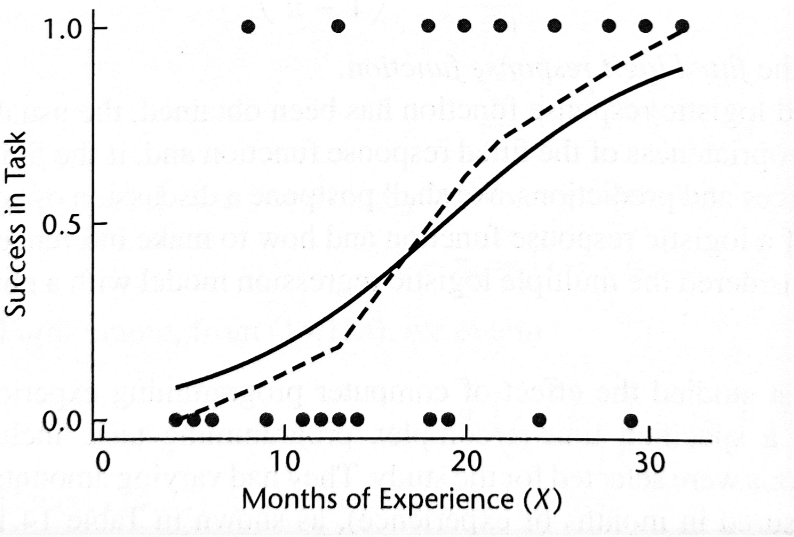
(b) Maximum Likelihood Estimates

Regression Coefficient	Estimated Regression Coefficient	Estimated Standard Deviation	Estimated Odds Ratio
β_0	−3.0597	1.259	—
β_1	.1615	.0650	1.175

2. All persons were given the same programming task, and the results of their success in the task are shown in column 2. The results are coded in binary fashion: $Y = 1$ if the task was completed successfully in the allotted time, and $Y = 0$ if the task was not complete d successfully.

3. (Figure 14.5) contains a scatter plot of the data. This plot is not too informative because of the nature of the response variable, other than to indicate that ability to complete the task successfully appears to increase with amount of experience. A lowess nonparametric response curve was fitted to the data and is also shown in Figure 14.5.

FIGURE 14.5 Scatter Plot, Lowess Curve (dashed line), and Estimated Logistic Mean Response Function (solid line)—Programming Task Example.



4. A sigmoidal S-shaped response function is clearly suggested by the nonparametric lowess fit. It was therefore decided to fit the logistic regression model (14.20).

5. A standard logistic regression package was run on the data. The results are contained in Table 14.1b. Since $b_0 = -3.0597$ and $b_1 = 0.1615$, the estimated logistic regression function:

$$\hat{\pi} = \frac{\exp(-3.0597 + 0.1615X)}{1 + \exp(-3.0597 + 0.1615X)}.$$

6. This fitted value is the estimated probability that a person with 14 months experience ($X_1 = 14$) will successfully complete the programming task.

7. In addition to the lowess fit, Figure 14.5 also contains a plot of the fitted logistic response function, $\hat{\pi}(x)$.

Interpretation of b_1

1. The interpretation of the estimated regression coefficient b_1 in the fitted logistic response function (14.30) is not the straightforward interpretation of the slope in a linear regression model.

2. The reason is that the effect of a unit increase in X varies for the logistic regression model according to the location of the starting point on the X scale.

3. An interpretation of b_1 is found in the property of the fitted logistic function that the estimated odds $\hat{\pi}/(1-\hat{\pi})$ are multiplied by $\exp(b_1)$ for any unit increase in X.

   (a) Consider the value of the fitted logit response function (14.29) at $X = X_j$:

   $$\hat{\pi}'(X_j) = b_0 + b_1 X_j .$$

   The notation $\hat{\pi}'(X_j)$ indicates specifically the X level associated with the fitted value.

   (b) We also consider the value of the fitted logit response function at $X = X_j + 1$: The difference between the two fitted values is simply:

   $$\hat{\pi}'(X_j+1) - \hat{\pi}'(X_j) = b_1 .$$

   (c) Now according to (14.29a), $\hat{\pi}'(X_j)$ is the logarithm of the estimated odds when $X = X_j$; we shall denote it by $\log_e(\text{odds}_1)$. Similarly, $\hat{\pi}'(X_j+1)$ is the logarithm of the estimated odds when $X = X_j + 1$; we shall denote it by $\log_e(\text{odds}_2)$.

   $$\hat{\pi}'(X_j) = \log_e(\text{odds}_1) = \ln\left(\frac{\pi(\hat{X}_j)}{1-\pi(\hat{X}_j)}\right) = b_0 + b_1 X_j .$$

   (d) Hence, the difference between the two fitted logit response values can be expressed as follows:

   $$\log_e(\text{odds}_2) - \log_e(\text{odds}_1) = \log_e \frac{\text{odds}_2}{\text{odds}_1} = b_1$$

   (e) Taking antilogs of each side, we see that the estimated ratio of the odds, called the odds ratio and denoted by $\widehat{OR}$, equals $\exp(b_1)$:

   $$\widehat{OR} = \frac{\text{odds}_2}{\text{odds}_1} = \exp(b_1) \qquad (14.31)$$

4. Example The programming task example.

   (a) We see from Figure 14.5 that the probability of success increases sharply with experience.

   (b) Specifically, Table 14.1b shows that the odds ratio is

   $$\widehat{OR} = \exp(b_1) = \exp(0.1615) = 1.175,$$

   so that the odds of completing the task increase by 17.5 percent with each additional month of experience.

(c) Since a unit increase of one month is quite small, the estimated odds ratio of 1.175 may not adequately show the change in odds for a longer difference in time. In general, the estimated odds ratio when there is a difference of c units of X is $\exp(cb_1)$.

(d) For example, should we wish to compare individuals with relatively little experience to those with extensive experience, say 10 months versus 25 months so that $c = 15$, then the odds ratio would be estimated to be $\exp[15(0.1615)] = 11.3$. This indicates that the odds of completing the task increase over 11-fold for experienced persons compared to relatively inexperienced persons.

Supplementary

1. The 6 Assumptions of Logistic Regression

   (a) The response variable is binary .

   (b) The observations are independent .

   (c) There is no multicollinearity among explanatory variables.

   (d) There are no extreme outliers .

   (e) There is a linear relationship between explanatory variables and the logit of the response Variable.

   (f) The sample size is sufficiently large .

2. Assumptions of Logistic Regression vs. Linear Regression: In contrast to linear regression, logistic regression does not require:

   (a) A linear relationship between the explanatory variable(s) and the response variable.

   (b) The residuals of the model to be normally distributed.

   (c) The residuals to have constant variance , also known as homoscedasticity .

😍 TA Class

- **Problems**: 14.7
- **Exercises**: none
- **Projects**: none

“不管你再怎麼努力，還是有人會忽略你的付出；就為了自己而奮鬥吧。”
“Your efforts will always be neglected no matter how hard you try; so fight for yourself.”
— *緊急迫降 (Emergency Declaration, 2022)*

國立政治大學 110 學年度第 1 學期 小考 (1) 考試命題紙

考試科目： Regression Analisis (I) 開課班別： 商院選修 命題教授: 吳漢銘

<table>
<tr><td>考試日期：10 月 21 日 (四) 11:10-12:00</td><td colspan="4">※准帶項目打「O」，否則打「×」</td><td rowspan="3">1. 需加發計算紙或答案紙請備註。
2. 為環保節能減碳，試題一律採雙面印刷，如有特殊印製需求，請註記：</td></tr>
<tr><td>本試題共 2 頁，印刷份數: 36 份</td><td>Calculator</td><td>Book
Notes</td><td>Dictionary</td><td>Cell phone
Laptop</td></tr>
<tr><td>備註：注意事項要看!! (§1~§2)</td><td>O</td><td>×</td><td>×</td><td>×</td></tr>
</table>

Note: (1) Fill in your name and student ID。(2) Answer the questions in English。(3) Answer the questions in the order in which they appear。(4) Pencils are permitted for use。(5) Hand in the question, the answer sheets and the sketch papers。(6) The calculation process is required.

1. (20%) Explain the following:

   (a) What is the ”Regression Analysis”?

   (b) Let α be the level of the significance. What is the so-called ”$(1-\alpha)\%$ Confidence Interval” for a parameter θ of the population.

   (c) What is the ”Coefficient of Determination” for a regression model? How to interpret this number?

   (d) What is the ”ANOVA table” for simple linear regression? What is it used for?

2. (15%) For the given sample observations $\{(X_i, Y_i), i = 1, \cdots, n\}$, we assume a simple linear regression model with distribution of error term unspecified as $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$. Find the least squares estimators of the parameters β_0 and β_1.

3. (20%) For the given sample observations $\{(X_i, Y_i), i = 1, \cdots, n\}$, we assume a normal error regression model as $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, where ϵ_i are independent normally distributed with mean 0 and variance σ^2. Find the MLEs of the parameters β_0 and β_1.

4. (10%) Given a random sample of data, $\{(X_i, Y_i), i = 1, \cdots, n\}$, and the level of the significance α, describe how to conduct the two-sided test concerning whether or not there is a linear association between X and Y for a normal error regression model. (State the null hypothesis, alternative hypothesis, test statistics (in terms of data), and decision rule.)

5. **Grade point average**. The director of admissions of a small college selected 120 students at random from the new freshman class in a study to determine whether a student's grade point average (GPA) at the end of the freshman year (Y) can be predicted from the ACT test score (X). The results of the study follow. Assume that a simple linear regression model is appropriate.

i:	1	2	3	$\cdots$	118	119	120
X_i:	21	14	28	$\cdots$	28	16	28
Y_i:	3.897	3.885	3.778	$\cdots$	3.914	1.860	2.948

國立政治大學 110 學年度第 1 學期 小考 (1) 考試命題紙

考試科目： Regression Analsis (I)　　開課班別： 商院選修　　命題教授: 吳漢銘

考試日期：10 月 21 日 (四) 11:10-12:00	※准帶項目打「O」，否則打「×」				1. 需加發計算紙或答案紙請備註。 2. 為環保節能減碳，試題一律採雙面印刷，如有特殊印製需求，請註記：
本試題共 2 頁，印刷份數: 36 份	Calculator	Book Notes	Dictionary	Cell phone Laptop	
備註：注意事項要看!! (§1~§2)	O	×	×	×	

The regression analysis report conducted by R is given in Table 1.

(b) (10%) Obtain a 95 percent confidence interval for β_1. Interpret your confidence interval. Does it include zero? Why might the director of admissions be interested in whether)he confidence interval includes zero? ($t_{0.025,120} = -1.97993$, $t_{0.05,120} = -1.657651$, $t_{0.025,119} = -1.9801$, $t_{0.05,119} = -1.657759$, $t_{0.025,118} = -1.980272$, $t_{0.05,118} = -1.65787$)

(c) (10%) Test, using the test statistic $t*$, whether or not a linear association exists between student's ACT score (X) and GPA at the end of the freshman year (Y). Use a level of significance of $0.O5$. State the alternatives, decision rule, and conclusion.

(d) (5%) What is the P-value of your test in part (b)? How does it support the conclusion reached in part (b)?

(e) (5%) How do you interpret R-squared in this analysis?

(f) (5%) The ANOVA table is shown in Table 2. How to you interpret ANOVA results?

注意： 1、考試求公平及公正，請同學務必自律，維護學校與學生之榮譽。

2、考試時不得有交談、窺視、夾帶、抄襲、傳遞、代考或其它作弊等舞弊行為，考畢務必交卷，不得攜卷出場，違者依考場規則議處。

Table 1: Regression analysis for Grade point average data

```
Call:
lm(formula = GPA ~ ACT, data = ex2.4.data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.74004 -0.33827  0.04062  0.44064  1.22737

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.11405    0.32089   6.588  1.3e-09 ***
ACT          0.03883    0.01277   3.040  0.00292 **
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ‘ ’  1

Residual standard error: 0.6231 on 118 degrees of freedom
Multiple R-squared:  0.07262,Adjusted R-squared:  0.06476
F-statistic:  9.24 on 1 and 118 DF,  p-value: 0.002917
```

Table 2: Analysis of Variance Table for Grade point average data

```
Analysis of Variance Table

Response: GPA
           Df Sum Sq Mean Sq F value   Pr(>F)
ACT         1  3.588  3.5878  9.2402 0.002917 **
Residuals 118 45.818  0.3883
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ‘ ’  1
```

國立政治大學 110 學年度第 1　學期　小考 (2)　考試命題紙

考試科目： Regression Analsis (I)　　開課班別： 商院選修　　命題教授: 吳漢銘

考試日期：12 月 02 日 (四) 11:10-12:00	※准帶項目打「O」，否則打「×」				1. 需加發計算紙或答案紙請備註。 2. 為環保節能減碳，試題一律採雙面印刷，如有特殊印製需求，請註記：
本試題共 1頁，印刷份數: 30 份	Calculator	Book Notes	Dictionary	Cell phone Laptop	
備註：注意事項要看!! (範圍: §5)	O	×	×	×	

Note: (1) Fill in your name and student ID。(2) Answer the questions in English。(3) Answer the questions in the order in which they appear。(4) Pencils are permitted for use。(5) Hand in the question, the answer sheets and the sketch papers。(6) The calculation process is required. (7) Use $\underset{\sim}{\beta}$ or $\underset{\sim}{X}$ to represent a vector $\boldsymbol{\beta}$ or a matrix $\mathbf{X}$.

1. One would like to fit the simple linear regression (SLR) model to a given dataset $\{(Y_i, X_i), i = 1, \cdots, n\}$.

   (a) (10%) Write down the normal error regression model for SLR in terms of (Y_i, X_i).

   (b) (10%) Express variables and regression coefficient by column vectors or a matrix first. And then Express the model in matrix terms (boldface symbols).

   (c) (20%) Derive the normal equations (in matrix notation) by the method of least squares:

   $$Q = \sum [Y_i - (\beta_0 + \beta_1 X_i)]^2 .$$

   (d) (10%) Obtain the estimated regression coefficients (denoted by $\mathbf{b}$) from normal equations by matrix methods.

2. (20%) Use matrix methods to obtain the estimated regression coefficients for the following data:

i	1	2	3	4	5	6	7	8	9	10
X_i	1	0	2	0	3	1	0	1	2	0
Y_i	16	9	17	12	22	13	8	15	19	11

NOTE: If $\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ then $\mathbf{A}^{-1} = \begin{bmatrix} d/D & -b/D \\ -c/D & a/D \end{bmatrix}$, where $D = ad - bc$.

3. ANOVA results from SLR.

   (a) (15%) There are three sums of squares in ANOVA results, write down their formulas (definitions) and derive their corresponding matrix representation. (Not just express them in matrix terms directly.)

   (b) (15%) Show that these three sums of squares are all quadratic forms.

注意： 1、考試求公平及公正，請同學務必自律，維護學校與學生之榮譽。

2、考試時不得有交談、窺視、夾帶、抄襲、傳遞、代考或其它作弊等舞弊行為，考畢務必交卷，不得攜卷出場，違者依考場規則議處。

國立政治大學 110 學年度第 1 學期 小考 (3) 考試命題紙

考試科目： Regression Analsis (I) 開課班別：商院選修 命題教授: 吳漢銘

考試日期：12 月 23 日 (四) 11:10-12:00	※准帶項目打「O」，否則打「×」				1. 需加發計算紙或答案紙請備註。 2. 為環保節能減碳，試題一律採雙面印刷，如有特殊印製需求，請註記：
本試題共 4頁，印刷份數: 30 份	Calculator	Book Notes	Dictionary	Cell phone Laptop	
備註：注意事項要看!! (範圍: §6~7)	O	×	×	×	

Note: (1) Fill in your name and student ID。(2) Answer the questions in English。(3) Answer the questions in the order in which they appear。(4) Pencils are permitted for use。(5) Hand in the question, the answer sheets and the sketch papers。(6) The calculation process is required. (7) Use $\underset{\sim}{\beta}$ or $\underset{\sim}{X}$ to represent a vector $\boldsymbol{\beta}$ or a matrix $\mathbf{X}$.

1. (10%) Consider the multiple linear regression model for a given data $\{Y_i, X_{i1}, X_{i2}, \cdots, X_{ip}\}_{i=1}^{n}$, someone would like to perform a F-test for Lack of Fit for this model. Please state (a) the general (multiple) linear regression model for this data; (b) the mean response function; (c) the test hypothesis (H_0, H_a); (d) the test statistic; and (e) the decision rule.

2. (5%) What is the extra sums of squares and what does it measure?

3. (10%) When the regression model contains three X variables, a variety of decompositions of $SSR(X_1, X_2, X_3)$ into extra sums of squares can be obtained. Please give three examples.

4. (10%) Consider the first-order regression model with three predictor variables, someone would like to use extra sums of squares in testing whether both $\beta_2 X_2$ and $\beta_3 X_3$ can be dropped from the full model. Please state (a) the test hypothesis (H_0, H_a); (b) the full model and the reduced model; (c) the general linear test statistics; and (d) the decision rule.

5. (5%) What is the definition of the coefficient of partial determination (take $R^2_{Y1|2}$ as an example and express it in terms of the extra sum of squares) and what does it measure?

6. (20%) Consider the multiple regression analysis, what is the multicollinearity problem? What are the effects of multicollinearity when conduct the multiple regression analysis? (Hint: you cannot just say that the multicollinearity has effects on the regression coefficients, for example, you need to describe what does it result in on the regression coefficients.)

國立政治大學 110 學年度第 1 學期 小考 (3) 考試命題紙

考試科目： Regression Analasis (I) 開課班別： 商院選修 命題教授: 吳漢銘

考試日期：12 月 23 日 (四) 11:10-12:00	※准帶項目打「O」，否則打「×」				1. 需加發計算紙或答案紙請備註。 2. 為環保節能減碳，試題一律採雙面印刷，如有特殊印製需求，請註記：
本試題共 4頁，印刷份數: 30 份	Calculator	Book Notes	Dictionary	Cell phone Laptop	
備註：注意事項要看!! (範圍: §6∼7)	O	×	×	×	

7. **Commercial properties**. A commercial real estate company evaluates vacancy rates, square footage, rental rates, and operating expenses for commercial properties in a large metropolitan area in order to provide clients with quantitative information upon which to make rental decisions. The data below are taken from 81 suburban commercial properties that are the newest, best located, most attractive, and expensive for five specific geographic areas. Shown here are the age (X_1), operating expenses and taxes (X_2), vacancy rates (X_3), total square footage (X_4), and rental rates (Y).

   (a) (10%) Obtain the analysis of variance table that decomposes the regression sum of squares into extra sums of squares associated with X_4; with X_1 given X_4; with X_2 , given X_1, and X_4; and with X_3, given X_1, X_2 and X_4. (Hint: $SSR(X_4), SSR(X_1|X_4), \cdots$)

   (b) (10%) Test whether X_3 can be dropped from the regression model given that X_1, X_2 and X_4 are retained. Use the F^* test statistic and level of significance 0.01. State the alternatives, decision rule, and conclusion. (Hint: $F(0.99; 1, 76) = 6.980578; F(0.99; 2, 76) = 4.89584; F(0.99; 3, 76) = 4.050282; F(0.99; 1, 75) = 6.985359; F(0.99; 2, 75) = 4.899877; F(0.99; 3, 75) = 4.054022$)

   (c) (10%) Test whether both X_2 and X_3 can be dropped from the regression model given that X_1 and X_4 are retained; use $\alpha = 0.01$. State the alternatives, and decision rule. (Hint: specify $df1$ and $df2$ in $F(0.99; df1, df2)$ as a critical value.)

   (d) (10%) Using the given R report sheet below, calculate the coefficient of partial determination $R^2_{Y2|14}$ and interpret. (Hint: Answer ”There was not sufficient information provided.” if the information provided was not sufficient to calculate $R^2_{Y2|14}$.)

注意： 1、考試求公平及公正，請同學務必自律，維護學校與學生之榮譽。

2、考試時不得有交談、窺視、夾帶、抄襲、傳遞、代考或其它作弊等舞弊行為，考畢務必交卷，不得攜卷出場，違者依考場規則議處。

```
> summary(m4)
lm(formula = Y ~ X4)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.378e+01  2.903e-01  47.482  < 2e-16 ***
X4          8.437e-06  1.498e-06   5.632 2.63e-07 ***
---
Residual standard error: 1.462 on 79 degrees of freedom
Multiple R-squared:  0.2865,    Adjusted R-squared:  0.2775
F-statistic: 31.72 on 1 and 79 DF,  p-value: 2.628e-07

> summary(m14)
lm(formula = Y ~ X1 + X4)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.436e+01  2.771e-01  51.831  < 2e-16 ***
X1          -1.145e-01  2.242e-02  -5.105 2.27e-06 ***
X4           1.045e-05  1.363e-06   7.663 4.23e-11 ***
---
Residual standard error: 1.274 on 78 degrees of freedom
Multiple R-squared:  0.4652,    Adjusted R-squared:  0.4515
F-statistic: 33.93 on 2 and 78 DF,  p-value: 2.506e-11

> summary(m124)
lm(formula = Y ~ X1 + X2 + X4)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.237e+01  4.928e-01  25.100  < 2e-16 ***
X1          -1.442e-01  2.092e-02  -6.891 1.33e-09 ***
X2           2.672e-01  5.729e-02   4.663 1.29e-05 ***
X4           8.178e-06  1.305e-06   6.265 1.97e-08 ***
---
Residual standard error: 1.132 on 77 degrees of freedom
Multiple R-squared:  0.583,     Adjusted R-squared:  0.5667
F-statistic: 35.88 on 3 and 77 DF,  p-value: 1.295e-14

> summary(m1234)
lm(formula = Y ~ X1 + X2 + X3 + X4)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.220e+01  5.780e-01  21.110  < 2e-16 ***
X1          -1.420e-01  2.134e-02  -6.655 3.89e-09 ***
X2           2.820e-01  6.317e-02   4.464 2.75e-05 ***
X3           6.193e-01  1.087e+00   0.570     0.57
X4           7.924e-06  1.385e-06   5.722 1.98e-07 ***
---
Residual standard error: 1.137 on 76 degrees of freedom
Multiple R-squared:  0.5847,    Adjusted R-squared:  0.5629
F-statistic: 26.76 on 4 and 76 DF,  p-value: 7.272e-14
```

```
> anova(m4)
Analysis of Variance Table
Response: Y
          Df  Sum Sq Mean Sq F value    Pr(>F)
X4         1  67.775  67.775  31.723 2.628e-07 ***
Residuals 79 168.782   2.136
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ‘ ’  1

> anova(m124)
Analysis of Variance Table
Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 14.819  14.819  11.566  0.001067 **
X2         1 72.802  72.802  56.825 7.841e-11 ***
X4         1 50.287  50.287  39.251 1.973e-08 ***
Residuals 77 98.650   1.281
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ‘ ’  1

> anova(m14)
Analysis of Variance Table
Response: Y
          Df  Sum Sq Mean Sq F value    Pr(>F)
X1         1  14.819  14.819  9.1365  0.003389 **
X4         1  95.231  95.231 58.7160 4.225e-11 ***
Residuals 78 126.508   1.622
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ‘ ’  1

> anova(m1234)
Analysis of Variance Table
Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 14.819  14.819 11.4649  0.001125 **
X2         1 72.802  72.802 56.3262 9.699e-11 ***
X3         1  8.381   8.381  6.4846  0.012904 *
X4         1 42.325  42.325 32.7464 1.976e-07 ***
Residuals 76 98.231   1.293
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ‘ ’  1
```

國立政治大學 110 學年度第 1 學期 期中考 考試命題紙

考試科目： Regression Analysis (I) 開課班別： 商院選修 命題教授: 吳漢銘

考試日期：11 月 11 日 (四) 9:10-10:30	※准帶項目打「O」，否則打「×」				1. 需加發計算紙或答案紙請備註。 2. 為環保節能減碳，試題一律採雙面印刷，如有特殊印製需求，請註記：
本試題共 3 頁，印刷份數: 36 份	Calculator	Book Notes	Dictionary	Cell phone Laptop	
備註：注意事項要看!! (§1~§3)	O	×	×	×	

Note: (1) Fill in your name and student ID on the answer sheet。(2) Answer the questions in English。(3) Answer the questions in the order in which they appear。(4) Pencils are permitted for use。(5) Hand in the question, the answer sheets and the sketch papers。(6) The calculation process is required.

1. (10%) For the given sample observations $\{(X_i, Y_i), i = 1, \cdots, n\}$, we assume a normal error regression model as $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, where ϵ_i are independent normally distributed with mean 0 and variance σ^2. Find the MLEs of the parameters β_0 and β_1.

2. Let $\{(X_i, Y_i), i = 1, \cdots, n\}$ be the observed data and we would like to perform a simple linear regression analysis. Please answer the following questions.

   (a) (8%) Which plots can be used to conduct the diagnostics for predictor variable?

   (b) (12%) The residuals can be used to examine six important types of departures from the simple linear regression model with normal errors. What are those six important types of departures?

   (c) (10%) Describe the Brown-Forsythe Test with a level of significant α (including at least the assumption for the data, the null hypothesis, the test statistics and the decision rule.)

3. (25%) In the textbook, we have already learned some transformations for X and/or Y to ensure that the assumptions for a simple linear regression normal error model are adequate. The transformations are

$$\log_{10}(X), 1/X, \sqrt{X}, X^2, \exp(X), \exp(Y), \log_{10}(Y), 1/Y, \sqrt{Y}, Y^{\lambda}.$$

   Four real world cases given below are analyzed each by a simple linear regression normal error model.

   (a) A research would like to study the regression relationship between alpha counts per second (Y) and plutonium activity (X) to estimate the activity of plutonium in the material under study.

   (b) A chemist studied the concentration of a solution (Y) over time (X). Fifteen identical solutions were prepared. The 15 solutions were randomly divided into five sets of three, and the five sets were measured, respectively, after 1, 3, 5, 7, and 9 hours.

   (c) A marketing researcher studied annual sales of a product that had been introduced 10 years ago. The data is collected, where X is the year (coded) and Y is sales in thousands of units.

   (d) In a manufacturing study, the production times for 111 recent production runs were obtained. The data consists of records for each run the production time in hours (Y) and the production lot size (X).

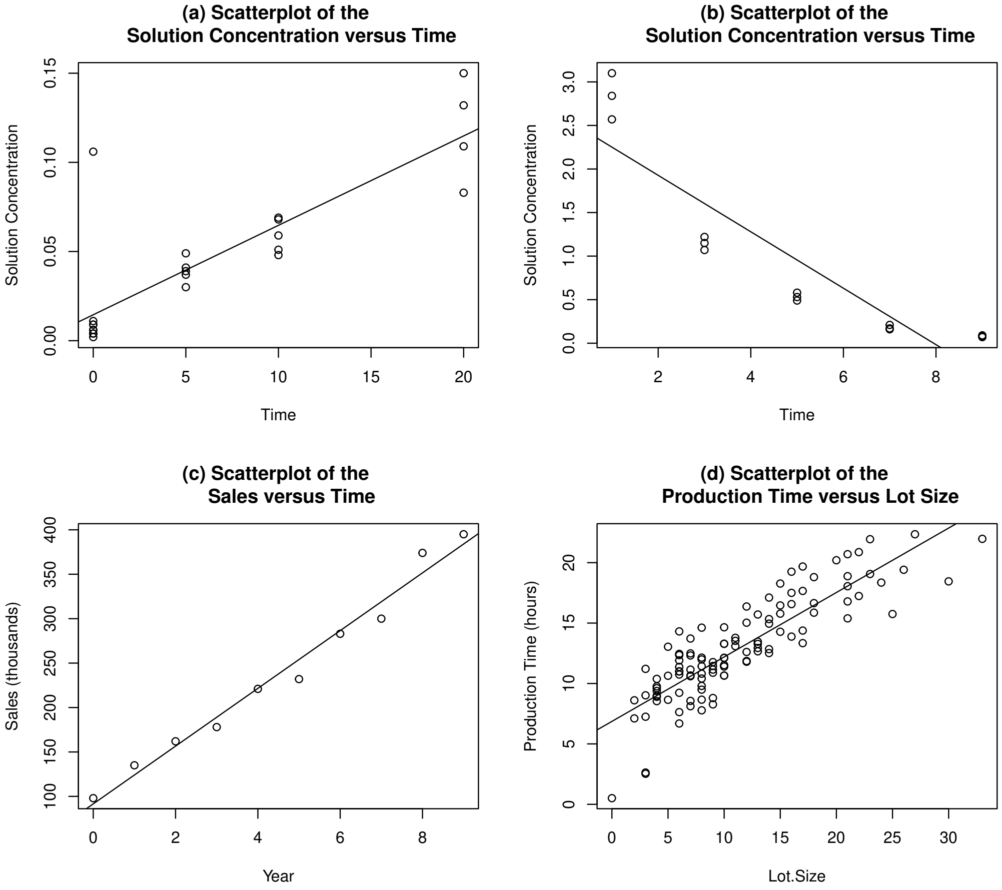
國立政治大學 110 學年度第 1 學期 期中考 考試命題紙

考試科目： Regression Analysis (I) 開課班別： 商院選修 命題教授: 吳漢銘

考試日期：11 月 11 日 (四) 9:10-10:30	※准帶項目打「O」，否則打「×」				1. 需加發計算紙或答案紙請備註。 2. 為環保節能減碳，試題一律採雙面印刷，如有特殊印製需求，請註記：
本試題共 3 頁，印刷份數: 36 份	Calculator	Book Notes	Dictionary	Cell phone Laptop	
備註：注意事項要看!! (§1∼§3)	O	×	×	×	

Based on the scatterplots of Y versus X with a regression line, please indicates whether the transformations are needed for Y and/or X and conclude which transformations are possible for each case. That is, fill in the blank spaces with the transformation methods in the following table in the answer sheet. Mark the blank by ”×” if the transformation is not necessary. You don’t have to specify the λ value when you think the Box-Cox transformation is appropriate.

Case	Transformation of X	Transformation of Y
(a)		
(b)		
(c)		
(d)		


(a) Scatterplot of the Solution Concentration versus Time

(b) Scatterplot of the Solution Concentration versus Time

(c) Scatterplot of the Sales versus Time

(d) Scatterplot of the Production Time versus Lot Size

國立政治大學 110 學年度第 1 學期 期中考 考試命題紙

考試科目： Regression Analysis (I) 開課班別：商院選修 命題教授: 吳漢銘

考試日期：11 月 11 日 (四) 9:10-10:30	※准帶項目打「O」，否則打「×」				1. 需加發計算紙或答案紙請備註。 2. 為環保節能減碳，試題一律採雙面印刷，如有特殊印製需求，請註記：
本試題共 3 頁，印刷份數: 36 份	Calculator	Book Notes	Dictionary	Cell phone Laptop	
備註：注意事項要看!! (§1~§3)	O	×	×	×	

4. Suppose that we obtain a data set that can be expressed in the form:

$$\{(X_j, Y_{ij}) : i = 1, \cdots, n_j; j = 1, \cdots, c\}, \text{ where } c > 2.$$

Someone would like to use F test for lack of fit to determine whether a simple linear regression model adequately fits the data, where X is the predictor variables and Y is the response.

(a) (5%) What are the assumptions of the lack of fit test?

(b) (5%) What is the full model used for the lack of fit test?

(c) (5%) What is the reduced model used for the lack of fit test?

(d) (5%) What is the null hypothesis for the lack of fit test?

(e) (10%) The Growth rate data are available on the effect of dietary supplement on the growth rates of rats. Here X = dose of dietary supplement and Y = growth rate. The following table presents the data in a form suitable for the analysis ($c = 6, n = 12$). Construct a general ANOVA Table (including Source of Variation, Sum of Square (SS), Degree of Freedom (df), Mean Square (MS) and F statistics) for testing lack of fit of a simple linear regression function.

Data		$j = 1$	$j = 2$	$j = 3$	$j = 4$	$j = 5$	$j = 6$
Replicate		$X_1 = 10$	$X_2 = 15$	$X_3 = 20$	$X_4 = 25$	$X_5 = 30$	$X_6 = 35$
Y_{ij}	$i = 1$	73	85	90	87	75	65
	$i = 2$	78	88	91	86		63
	$i = 3$			91			

(f) (5%) State the test statistics, decision rule and conclusion. (for all j at 5% level of significance)

(Some numbers: error sum of squares for the reduced model ($SSE(R)$) = 891.73, regression sum of squares (SSR) = 204.27, total sum of squares ($SSTO$) = 1096.00, $F(0.95; 5, 5) = 5.050, F(0.95; 6, 4) = 6.163, F(0.95; 4, 6) = 4.534, F(0.95; 1, 10), F(0.95; 10, 1) = 241.881, F(0.95; 2, 10) = 4.103, F(0.95; 2, 9) = 4.256, F(0.95; 2, 8) = 4.459$; $\hat{Y}_{ij} = 92.003 - 0.498X_j$)

注意： 1、考試求公平及公正，請同學務必自律，維護學校與學生之榮譽。

2、考試時不得有交談、窺視、夾帶、抄襲、傳遞、代考或其它作弊等舞弊行為，考畢務必交卷，不得攜卷出場，違者依考場規則議處。

國立政治大學 110 學年度第 1　學期　期末考　考試命題紙

考試科目：Regression Analysis (I)　　開課班別：商院選修　　命題教授: 吳漢銘

考試日期：01 月 13 日 (四) 9:10-10:40	※准帶項目打「O」，否則打「×」				1. 需加發計算紙或答案紙請備註。 2. 為環保節能減碳，試題一律採雙面印刷，如有特殊印製需求，請註記：
本試題共 5頁，印刷份數: 36 份	Calculator	Book Notes	Dictionary	Cell phone Laptop	
備註：注意事項要看!! (§5∼§14)	O	×	×	×	

Note: (1) Fill in your name and student ID on the answer sheet。(2) Answer the questions in English。(3) Answer the questions in the order in which they appear。(4) Pencils are permitted for use。(5) Hand in the question, the answer sheets and the sketch papers。(6) The calculation process is required. (7) The total is 100 points.

1. (15%) For SLR, there are three sums of squares in ANOVA results, write down their formulas (definitions) and derive their corresponding matrix representation. (Do not just express them in matrix terms directly.)

2. (5%) What is the four main steps for building a regression model?

3. (10%) Describe the "Forward Stepwise Regression" procedure to a hypothesized data set with variables $\{Y, X_1, X_2, X_3, X_4\}$ for selecting a good model.

see next page...

國立政治大學 110 學年度第 1　學期　 期末考　考試命題紙

考試科目： Regression Analysis (I)　　開課班別： 商院選修　　命題教授: 吳漢銘

考試日期：01 月 13 日 (四) 9:10-10:40	※准帶項目打「O」，否則打「×」				1. 需加發計算紙或答案紙請備註。 2. 為環保節能減碳，試題一律採雙面印刷，如有特殊印製需求，請註記：
本試題共 5頁，印刷份數: 36 份	Calculator	Book Notes	Dictionary	Cell phone Laptop	
備註：注意事項要看!! (§1∼§3)	O	×	×	×	

4. (20%) **Patient satisfaction**. A hospital administrator wished to study the relation between patient satisfaction (Y) and patient's age (X_1, in years), severity (嚴重性) of illness (X_2, an index), and anxiety (焦慮) level (X_3, an index). The administrator randomly selected 46 patients and collected the data presented below (not shown), where larger values of Y, X_2, and X_3 are, respectively, associated with more satisfaction, increased severity of illness, and more anxiety.

   (a) (5%) Obtain the analysis of variance table that decomposes the regression sum of squares into extra sums of squares associated with X_2; with X_1 given X_2; with X_3 , given X_2, and X_1.

   (b) (5%) Test whether X_3 can be dropped from the regression model given that X_1, and X_2 are retained. Use the F^* test statistic and level of significance 0.025 State the alternatives, decision rule, and conclusion. (Hint: (lower.tail) $F(0.975, 1, 41) = 5.4136, F(0.975; 1, 42) = 5.4039, F(0.975, 2, 41) = 4.0416, F(0.975, 2, 42) = 4.0327$)

   (c) (5%) Test whether both X_2 and X_3 can be dropped from the regression model given that X_1 are retained; use $\alpha = 0.01$. State the alternatives, and decision rule. (Hint: specify $df1$ and $df2$ in $F(0.99; df1, df2)$ as a critical value. Since the value of $F(0.99; df1, df2)$ is not given, you don't have to draw a conclusion.)

   (d) (5%) Using the given R report sheet below, calculate the coefficient of partial determination $R^2_{Y1|23}$ and interpret. (Hint: Answer "There was not sufficient information provided." if the information provided was not sufficient to calculate $R^2_{Y1|23}$.)

```
> anova(m2)
Analysis of Variance Table
Response: Y
          Df Sum Sq Mean Sq F value   Pr(>F)
X2         1 4860.3  4860.3  25.132 9.23e-06 ***
Residuals 44 8509.0   193.4
> anova(m12)
Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 8275.4  8275.4 77.1389 3.802e-11 ***
X2         1  480.9   480.9  4.4828   0.04006 *
Residuals 43 4613.0   107.3
> anova(m123)
Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 8275.4  8275.4 81.8026 2.059e-11 ***
X2         1  480.9   480.9  4.7539   0.03489 *
X3         1  364.2   364.2  3.5997   0.06468 .
Residuals 42 4248.8   101.2
```

國立政治大學 110 學年度第 1　學期　 期末考　考試命題紙

考試科目： Regression Analysis (I)　　開課班別： 商院選修　　命題教授: 吳漢銘

考試日期：01 月 13 日 (四) 9:10-10:40	※准帶項目打「O」，否則打「×」				1. 需加發計算紙或答案紙請備註。 2. 為環保節能減碳，試題一律採雙面印刷，如有特殊印製需求，請註記：
本試題共 5頁，印刷份數: 36 份	Calculator	Book Notes	Dictionary	Cell phone Laptop	
備註：注意事項要看!! (§1~§3)	O	×	×	×	

5. (20%) **Assessed valuations** Assessed valuations. A tax consultant studied the current relation between selling price and assessed valuation of one-family residential dwellings in a large taX district by obtaining data for a random sample of 16 recent ”arm’s-length” sales transactions of one-family dwellings located on comer lots and for a random sample of 48 recent sales of one-family dwellings not located on corger lots. In the data that follow, both selling price (Y) and assessed valuation (X_1 are expressed in thousand dollars, whereas lot location (X_2) is coded 1 for comer lots and 0 for non-comer lots. Assume that the error variances in the two populations are equal and that a first-order regression model with an added interaction term is appropriate.

   (a) State the estimated regression function.

   (b) Explain the meaning of all regression coefficients in the model.

   (c) Test whether the interaction term can be dropped from the model; use $\alpha = 0.05$. State the alternatives. decision rule, and conclusion. If the interaction term cannot be dropped from the model, describe the nature of the interaction effect.

   (d) What is the predicted selling price $\hat{Y}$ when the assessed valuation X_1 is 77.1 (thousand dollars) for corner lots.

```
Call:
lm(formula = Y ~ X1 * X2, data = AssessedValuations)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.8470  -2.1639   0.0913   1.9348   9.9836 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -126.9052    14.7225  -8.620 4.33e-12 ***
X1             2.7759     0.1963  14.142  < 2e-16 ***
X21           76.0215    30.1314   2.523  0.01430 *  
X1:X21        -1.1075     0.4055  -2.731  0.00828 ** 
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ″  1

Residual standard error: 3.893 on 60 degrees of freedom
Multiple R-squared:  0.8233,	Adjusted R-squared:  0.8145 
F-statistic: 93.21 on 3 and 60 DF,  p-value: < 2.2e-16
```

6. (10%) **Peruvian Blood Pressure Data** This dataset consists of variables possibly relating to blood pressures of $n = 39$ Peruvians (秘魯人) who have moved from rural high altitude areas to urban lower altitude areas. The variables in this dataset are: Y = Systolic blood pressure (`Systol`), X_1 = `Age`, X_2 = `Years` in urban area, X_3 = `Weight` (kg), X_4 = `Calf` (小腿肚) skinfold, and X_5 = resting `Pulse` rate. Using only first-order terms for predictor variables, the various criteria values using R for all possible regression models is given below. (a) What are the formulas of the following critera: $R_{a,p}, AIC_p, SBC_p$ and $PRESS_p$ (b) Find just one best subset regression model according to the above criteria and state your reasons.

```
  p 1 2 3 4 5     SSEp     r2  r2.adj      Cp     AICp     SBCp   PRESSp
1 2 0 0 1 0 0 4756.056 0.2718  0.2521  6.9711 191.3409 194.6680 5182.089
1 2 0 0 0 1 0 6120.640 0.0629  0.0376 19.0132 201.1785 204.5057 6744.847
1 2 0 0 0 0 1 6411.558 0.0184 -0.0082 21.5805 202.9895 206.3167 7521.225
1 2 0 1 0 0 0 6481.452 0.0077 -0.0192 22.1973 203.4124 206.7395 7579.467
1 2 1 0 0 0 0 6531.213 0.0000 -0.0270 22.6364 203.7107 207.0378 7866.668
2 3 0 1 1 0 0 3783.157 0.4208  0.3886  0.3855 184.4154 189.4060 4549.213
2 3 1 0 1 0 0 4370.331 0.3309  0.2937  5.5671 190.0423 195.0329 5470.343
2 3 0 0 1 1 0 4739.383 0.2744  0.2341  8.8239 193.2039 198.1946 5424.335
2 3 0 0 1 0 1 4750.751 0.2726  0.2322  8.9242 193.2974 198.2880 5663.745
2 3 0 1 0 1 0 6070.340 0.0706  0.0190 20.5693 202.8567 207.8474 7341.404
2 3 0 0 0 1 1 6073.444 0.0701  0.0185 20.5967 202.8767 207.8673 7389.767
2 3 1 0 0 1 0 6120.302 0.0629  0.0109 21.0102 203.1764 208.1671 7662.934
2 3 0 1 0 0 1 6312.616 0.0335 -0.0202 22.7073 204.3830 209.3737 8276.004
2 3 1 0 0 0 1 6411.285 0.0184 -0.0361 23.5781 204.9879 209.9786 8753.436
2 3 1 1 0 0 0 6448.660 0.0127 -0.0422 23.9079 205.2146 210.2053 8350.733
3 4 1 1 1 0 0 3755.255 0.4250  0.3758  2.1392 186.1266 192.7809 4933.377
3 4 0 1 1 1 0 3772.562 0.4224  0.3729  2.2920 186.3060 192.9602 4708.035
3 4 0 1 1 0 1 3782.245 0.4209  0.3713  2.3774 186.4059 193.0602 4955.800
3 4 1 0 1 0 1 4359.345 0.3326  0.2754  7.4702 191.9441 198.5983 5986.127
3 4 1 0 1 1 0 4370.329 0.3309  0.2735  7.5671 192.0422 198.6965 5727.281
3 4 0 0 1 1 1 4731.979 0.2755  0.2134 10.7586 195.1429 201.7972 5904.062
3 4 0 1 0 1 1 5992.029 0.0826  0.0040 21.8782 204.3503 211.0046 8040.035
3 4 1 1 0 1 0 6035.794 0.0759 -0.0033 22.2644 204.6341 211.2884 7986.608
3 4 1 0 0 1 1 6073.440 0.0701 -0.0096 22.5967 204.8766 211.5309 8554.830
3 4 1 1 0 0 1 6269.788 0.0401 -0.0422 24.3294 206.1175 212.7718 9148.679
4 5 1 1 1 1 0 3740.114 0.4274  0.3600  4.0056 187.9691 196.2869 5112.528
4 5 1 1 1 0 1 3755.138 0.4251  0.3574  4.1382 188.1254 196.4432 5373.499
4 5 0 1 1 1 1 3770.654 0.4227  0.3548  4.2751 188.2862 196.6040 5105.065
4 5 1 0 1 1 1 4359.281 0.3326  0.2540  9.4696 193.9435 202.2613 6255.259
4 5 1 1 0 1 1 5950.595 0.0889 -0.0183 23.5126 206.0797 214.3975 8804.993
5 6 1 1 1 1 1 3739.478 0.4275  0.3407  6.0000 189.9624 199.9438 5545.949
```

國立政治大學 110 學年度第 1 學期 期末考 考試命題紙

考試科目： Regression Analysis (I) 開課班別： 商院選修 命題教授: 吳漢銘

考試日期：01 月 13 日 (四) 9:10-10:40	※准帶項目打「O」，否則打「×」				1. 需加發計算紙或答案紙請備註。 2. 為環保節能減碳，試題一律採雙面印刷，如有特殊印製需求，請註記：
本試題共 5頁，印刷份數: 36 份	Calculator	Book Notes	Dictionary	Cell phone Laptop	
備註：注意事項要看!! (§1~§3)	O	×	×	×	

7. (20%) **Toxicity experiment**. In an experiment testing the effect of a toxic substance, 1,500 experimental insects were divided at random into six groups of 250 each. The insects in each group were exposed to a fixed dose of the toxic substance. A day later, each insect was observed. Death from exposure was scored 1, and survival was scored 0. The results are shown below; X_j denotes the dose level (on a logarithmic scale) administered to the insects in group j and $Y_{.j}$ denotes the number of insects that died out of the 250 (n_j) in the group. The estimated proportions is denoted by $p_j = Y_{.j}/n_j$.

j:	1	2	3	4	5	6
X_j:	1	2	3	4	5	6
n_j:	250	250	250	250	250	250
$Y_{.j}$	28	53	93	126	172	197

Simple Logistic regression model is assumed to be appropriate. The R output for the logistic regression is given below.

(a) State the fitted logistic response function.

(b) Obtain $\exp(b_1)$ and interpret this number.

(c) What is the estimated probability that an insect dies when the dose level is $X = 3.5$?

(d) What is the estimated median lethal dose-that is, the dose for which 50 percent of the experimental insects are expected to die?

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.67466    0.08285  -32.28 5.49e-06 ***
X            0.67908    0.02128   31.92 5.74e-06 ***
---
Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  "  1

Residual standard error: 0.089 on 4 degrees of freedom
Multiple R-squared:  0.9961,        Adjusted R-squared:  0.9951
F-statistic:  1019 on 1 and 4 DF,  p-value: 5.743e-06
```

注意： 1、考試求公平及公正，請同學務必自律，維護學校與學生之榮譽。

2、考試時不得有交談、窺視、夾帶、抄襲、傳遞、代考或其它作弊等舞弊行為，考畢務必交卷，不得攜卷出場，違者依考場規則議處。

“堅持做對的事，永遠不會錯。”
“You are never wrong to do the right thing.”

— *高年級實習生 (The intern, 2015)*