Regression Analysis (I)

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

Chapter 6: Multiple Regression (I)

Thursday 09:10-12:00, 資訊 140301

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 ________________.
3. The matrix expressions for multiple regression are the __________ as for SLR.
4. An example to illustrate a variety of ______________ and ____________________ 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 _______________ description since a number of ________________ affect the response variable.
2. Predictions of the response variable based on a model containing only a single predictor variable are too ______________ to be useful.
3. Multiple regression analysis is highly useful in experimental situations where the experimenter can ________________________________.

4. The multiple regression models can be utilized for either ________________ data or for ________________ data from a completely randomized design.

First-Order Model with Two Predictor Variables

1. When there are two predictor variables X_l and X_2, the regression model:

$$\underline{\hspace{6cm}} \qquad (6.1)$$

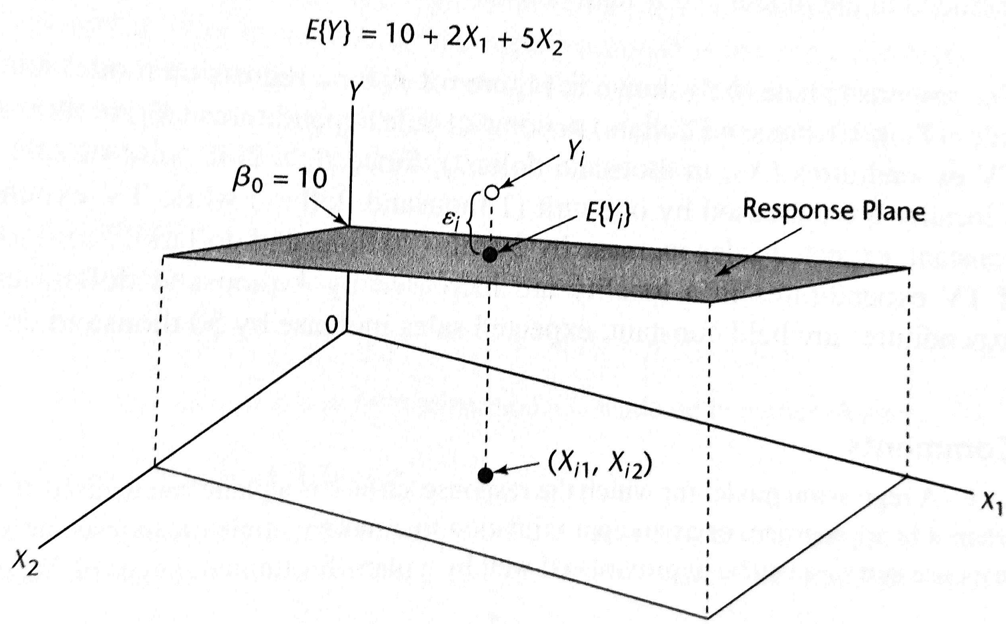
is called a ______________ model with two predictor variables.

2. Assuming that ____________, the regression function for model (6.1) is a _________:

$$\underline{\hspace{6cm}} \qquad (6.2)$$

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 ______________.

(b) The error term __________________: 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 ____________________ or a ____________________. In Figure 6.1, the response surface is a __________, but in other cases the response surface may be more ____________ in nature.

5. **Meaning of Regression Coefficients**

   (a) The parameter β_0 is the _______________ of the regression plane.

   (b) If the scope of the model includes ___________________, then β_0 represents the mean response $E(Y)$ at $X_1 = 0$, $X_2 = 0$. Otherwise, β_0 _____________ have any particular meaning as a separate term in the regression model.

   (c) The parameter β_1 (β_2) indicates the ___________ in the mean response $E(Y)$ per unit increase in ____________ when ____________ 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 __________________ or not to ____________.

   (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 ______________________________ because they reflect the partial effect of one predictor variable when the other predictor variable is included in the model and is _________________.

First-Order Model with More than Two Predictor Variables

1. The regression model:

$$Y_i = \underline{\hspace{8cm}} \qquad (6.5)$$

$$= \underline{\hspace{4cm}} \qquad (6.5a)$$

$$= \underline{\hspace{3cm}} \quad \text{where } X_{i0} \equiv 1 \qquad (6.5b)$$

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:

$$E(Y) = \underline{\hspace{6cm}} \tag{6.6}$$

3. This response function is a ____________, which is a plane in more than two dimensions.

4. The parameter β_k indicates the ____________________________ 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 __________ 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:

$$Y_i = \underline{\hspace{7cm}} \tag{6.7}$$

where:

(a) $\beta_0, \beta_1, \cdots, \beta_{p-1}$ are ____________.

(b) $X_{ii}, \cdots, X_{i,p-1}$ are _________ constants (predictors, explanatory variables).

(c) ε_i are independent __________, $i = 1, \cdots, n$.

2. The response function for regression mode1 (6.7) is:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1} \tag{6.8}$$

3. Thus, the general linear regression model with normal error terms implies that the observations Y_i are independent ________________, with mean ________ as given by (6.8) and with constant variance ______.

4. **Qualitative Predictor Variables**

(a) The general linear regression model (6.7) encompasses not only quantitative predictor variables but also ____________ ones, such as gender (male, female) or disability status (not disabled, partially disabled, fully disabled).

(b) Use ___________ variables that take on the values __________ to identify the classes of a qualitative variable.

(c) 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:

$$Y_i = \underline{\hspace{6cm}} \qquad (6.9)$$

$$X_{i1} = i\text{th patient's age}$$

$$X_{i2} = \underline{\hspace{6cm}}$$

The response function for regression model (6.9) is:

$$E(Y) = \underline{\hspace{5cm}} \qquad (6.10)$$

For male patients, $X_2 = 0$ and response function (6.10) becomes:

$$E(Y) = \underline{\hspace{3cm}}, \qquad \text{Male patients} \qquad (6.10a)$$

For female patients, $X_2 = 1$ and response function (6.10) becomes:

$$E(Y) = \underline{\hspace{4cm}}, \qquad \text{Female patients} \qquad (6.10b)$$

These two response functions represent _______________ lines with different intercepts.

(d) In general, we represent a qualitative variable with c classes by means of ________ indicator variables. (details in Chapter 8)

5. Example The first-order model with age, gender (male, female) or disability status (not disabled, partially disabled, fully disabled) as predictor variables then is:

$$Y_i = \underline{\hspace{9cm}} \qquad (6.11)$$

where:

$$X_{i1} = i\text{th patient's age}$$

$$X_{i2} = \underline{\hspace{6cm}}$$

$$X_{i3} = \underline{\hspace{7cm}}$$

$$X_{i4} = \underline{\hspace{6.5cm}}$$

6. **Polynomial Regression**

   (a) Polynomial regression models are special cases of the general linear regression model. They contain ____________ and ________________ terms of the predictor variable(s), making the response function ______________.

   (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 ______________ 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 _____________ 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 ____________________.

   (b) $\boxed{\text{Example}}$ 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 ________________. It is a special case of the general linear regression model. (detail in Chapter 8)

9. **Combination of Cases**

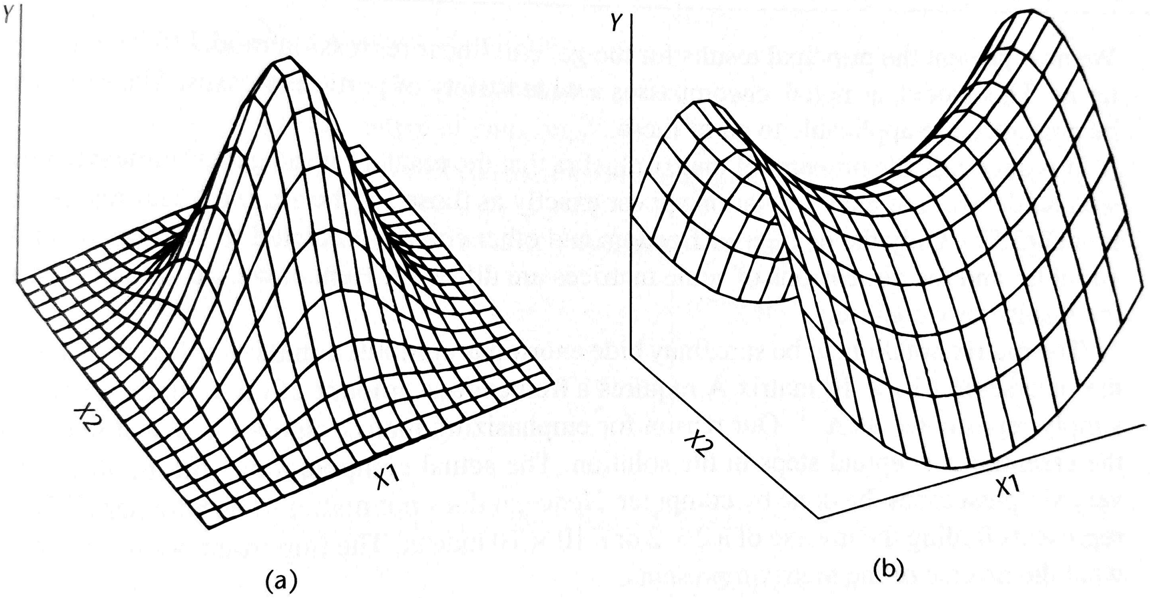
   (a) A regression model may combine several of the elements we have just noted and still be treated as a general linear regression model.

   (b) $\boxed{\text{Example}}$ 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 ______________ refers to the fact that model (6.7) is linear in the ______________; it does-not refer to the ____________________________.

(b) We say that a regression model is linear in the parameters when it can be written in the form:

$$Y_i = \underline{\hspace{6cm}},$$

where the terms c_{i0}, c_{i1}, etc., are coefficients involving the ____________________.

(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 ____________ 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):

$$Y_i = \underline{\hspace{7cm}} \tag{6.7}$$

   in matrix terms:

$$\underline{\hspace{4.5cm}},$$

   where

$$\mathbf{Y} = \qquad\qquad\qquad, \quad \mathbf{X} = \qquad\qquad\qquad\qquad\qquad\qquad,$$

$$\underline{\hspace{2cm}} \qquad \underline{\hspace{5.5cm}}$$

$$\boldsymbol{\beta} = \qquad\qquad\qquad, \quad \boldsymbol{\varepsilon} = \qquad\qquad\qquad,$$

$$\underline{\hspace{2.2cm}} \qquad \underline{\hspace{2cm}}$$

4. $\boldsymbol{\varepsilon}$ is a vector of independent normal random variables with ____________ and ______________.

5. The random vector $\mathbf{Y}$ has expectation: ______________, and the variance-covariance matrix of $\mathbf{Y}$ is the same as that of $\boldsymbol{\varepsilon}$: ______________.

6.3 Estimation of Regression Coefficients

1. The least squares criterion (1.8) is generalized as follows for general linear regression model (6.7):

$$Q = \underline{\hspace{8cm}} \qquad (6.22)$$

2. The least squares estimators are those values of $\beta_0, \beta_1, \cdots, \beta_{p-1}$ that ______________.

3. The least squares normal equations for the general linear regression model (6.19) are:

$$\underline{\hspace{3cm}} \qquad (6.24)$$

4. The least squares estimators are:

$$\mathbf{b} = \underline{\hspace{3.5cm}} \qquad (6.25)$$

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:

$$L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_{p-1}X_{i,p-1})^2\right\} \qquad (6.26)$$

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 ____________________________ ______________, and ____________.

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 $\mathbf{e}$:

$$\hat{\mathbf{Y}} = \underline{\hspace{4cm}}, \qquad \mathbf{e} = \underline{\hspace{4cm}},$$

2. The fitted values: ____________.
3. The vector of the fitted values $\hat{\mathbf{Y}}$ can be expressed in terms of the hat matrix $\mathbf{H}$ as follows:

$$\hat{\mathbf{Y}} = \underline{\hspace{2.5cm}}, \quad \text{where} \quad \underline{\hspace{6cm}} \tag{6.30}$$

4. The residual terms: $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \underline{\hspace{3cm}}$.
5. Similarly, the vector of residuals can be expressed: $\mathbf{e} = \underline{\hspace{3cm}}$.
6. The variance-covariance matrix of the residuals is: $\sigma^2(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H})$ which is estimated by:

$$s^2(\mathbf{e}) = \underline{\hspace{5cm}} \tag{6.33}$$

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):

$$SSTO = \underline{\hspace{10cm}}$$

$$SSE = \underline{\hspace{14cm}}$$

$$SSR = \underline{\hspace{11cm}}$$

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 $X_1, \cdots, X_p$,

$$H_0 : \underline{\hspace{6cm}}$$
$$H_a : \underline{\hspace{9cm}}$$

2. The test statistic:

$$F^* = \underline{\hspace{2cm}}.$$

3. The decision rule to control the Type I error at α:

$$\underline{\hspace{7cm}}$$

Coefficient of Multiple Determination

1. The coefficient of multiple determination, denoted by $R2$, is defined as

$$R^2 = \underline{\hspace{5cm}} \qquad (6.40)$$

2. It measures the ____________________ 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 ____________________, and the value 1 when all Y observations fall directly on the fitted regression surface, ie., when ________ for all i.

4. Adding more X variables to the regression model can only ___________ 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 ___________ coefficient of multiple determination, denoted by R_a^2, adjusts R^2 by dividing each sum of squares by its associated degrees of freedom:

$$R_a^2 = \underline{\qquad\qquad\qquad\qquad\qquad\qquad}$$

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. $\boxed{\text{Comments}}$ 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

$$R = \underline{\qquad\qquad}$$

6.6 Inferences about Regression Parameters

1. The least squares and maximum likelihood estimators in $\mathbf{b}$ are ___________:

$$E\{\mathbf{b}\} = \underline{\qquad} \tag{6.44}$$

2. The variance-covariance matrix (dimension $p \times p$):

$$\sigma^2\{\mathbf{b}\} = \underline{\hspace{6em}} \tag{6.46}$$

3. The estimated variance-covariance matrix (dimension $p \times p$):

$$s^2\{\mathbf{b}\} = \underline{\hspace{7em}} \tag{6.48}$$

Interval Estimation of β_k

1. For the normal error regression model (6.19), we have:

$$\underline{\hspace{8em}}, \quad k = 0, 1, ..., p-1 \tag{6.49}$$

2. The confidence limits for β_k with $1-\alpha$ confidence coefficient are:

$$\underline{\hspace{10em}} \tag{6.50}$$

Tests for β_k

1. The test hypothesis:

$$\underline{\hspace{15em}}$$

2. The test statistic:

$$\underline{\hspace{6em}}$$

3. The decision rule:

$$\underline{\hspace{15em}}.$$

4. The __________ 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 __________ 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 ______________ and ____________ of multiple regression models.
2. Most of the diagnostic procedures for ________ (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*: ____
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*

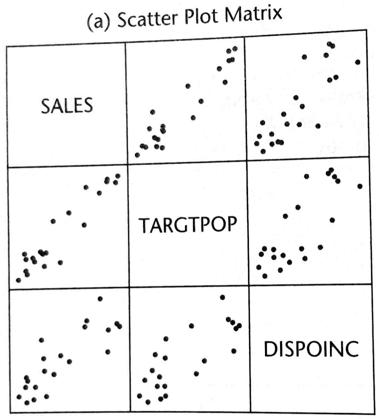
   (a) Scatter plots of the ____________ variable against each ____________ variable can aid in determining the nature and strength of the ________________________ between each of the predictor variables and the response variable and in identifying gaps in the data points as well as ____________ data points.

   (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 ________ and detecting ___________.

3. *Multiivariate plots: Scatter plot matrix*

FIGURE 6.4 SYGRAPH Scatter Plot Matrix and Correlation Matrix— Dwaine Studios Example.

(a) Scatter Plot Matrix



(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 ________, and the X variable is the name found in its ___________.

   (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 ____________________. This matrix contains the coefficients of simple correlation ____________________

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: _______ between X_1 and X_2, _______ between X_1 and X_3, etc.

5. Note that the correlation matrix is ______________ and that its main diagonal contains _______ because the coefficient of correlation between a variable and itself is ______.

Three-Dimensional Scatter Plots

1. Some ______________ statistics packages provide three-dimensional scatter plots or point clouds, and permit ____________ 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 ____________ against the ________________ is useful for assessing the ___________________ of the multiple regression function and the ______________ of the variance of the error terms, as well as for providing information about ___________, just as for simple linear regression.

2. *plot(*$e_i \sim$ *time)*: A plot of the ____________ against _________ or against some other ____________ can provide diagnostic information about possible _______________ 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 ____________ distributed.

4. *plot(*$e_i \sim X_i$*)*: The plot of the residuals against each of the ____________ 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 ____________ in the magnitude of the error variance in relation to that predictor variable.

5. *plot(*$e_i \sim X_{omit}$*)*: Plot the residuals against ______________________ 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 ____________________ 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 ____________ residuals or the ____________ residuals against the fitted values is useful for examining the ______________ 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 ___________________ 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 ______________________ variables.

2. To conduct the Brown-Forsythe test, we divide the data set into ______________, as for simple linear regression, where one group consists of cases where the level of the predictor variable is relatively ________ and the other group consists of cases where the level of the predictor variable is relatively ________.

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:

   ____

   is an appropriate response surface.

2. Repeat observations in multiple regression are __________ 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 ____________ for each of the X variables.

6. Let c denote the number of groups with ______________________________, 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 ________________. The lack of fit sum of squares $SSLF$ equals the difference ______________.

7. *Test hypothesis*:

   $$H_0 : \rule{8cm}{0.4pt}$$

   $$H_a : E\{Y\} \neq \beta_0 + \eta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}$$

8. *Test statistic*:

   $$F^* = \rule{6cm}{0.4pt}$$

9. *Decision rule*:

   ______________________________.

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 ______________ or _______________ effects, the multiple regression model can be expanded to include these effects.

3. Transformations on the _____________ variable Y may be helpful when the distributions of the error terms are _________________ and the variance of the error terms is _________________.

4. Transformations of some of the predictor variables may be helpful when the effects, of these variables are _______________.

5. Transformations on Y and/or the predictor variables may be helpful in eliminating or substantially ______________________________.

6. The usefulness of potential transformations needs to be examined by means of __________________ and other ____________________ 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:

   ____

   with normal error terms is expected to be appropriate, on the basis of the scatter plot matrix in Figure 6.4a.

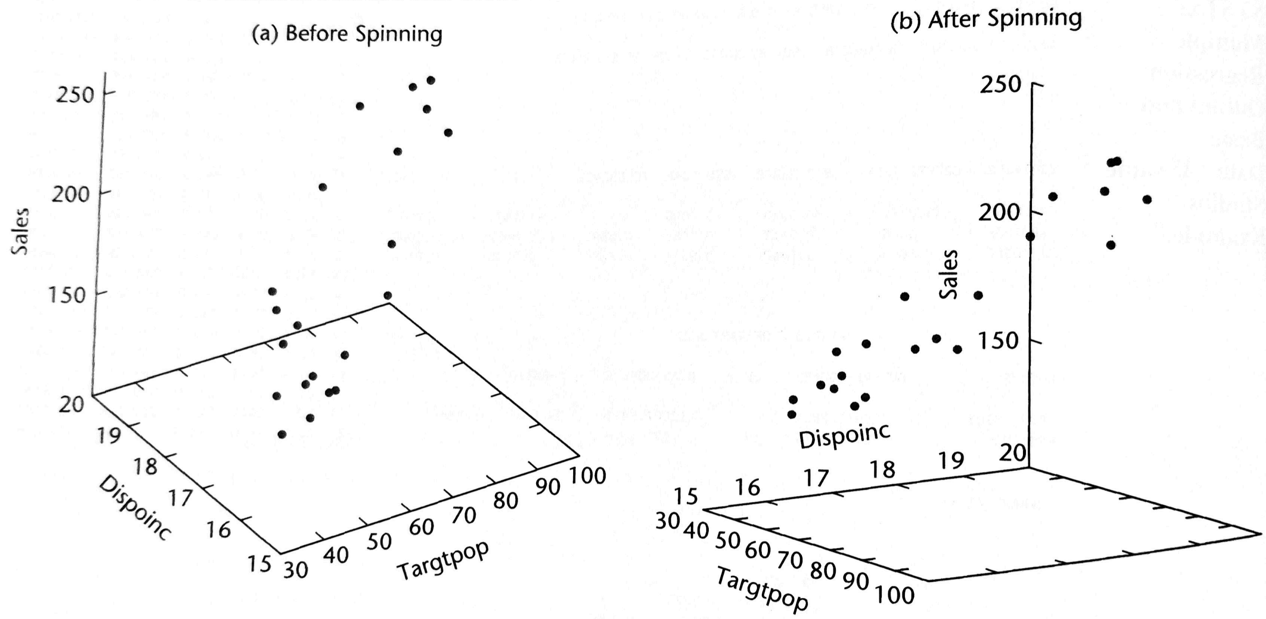
3. Note the ________________ between target population and sales and between disposable income and sales.

4. Also note that there is _______________ between disposable income and sales relationship.

5. Finally note that there is also some _________ 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

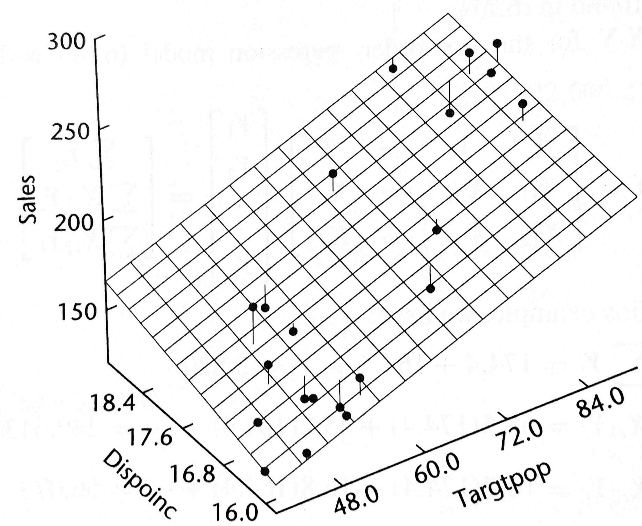
$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{bmatrix} -68.857 \\ 1.455 \\ 9.366 \end{bmatrix}$$

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 ______ 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 ______ 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

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b} = \begin{bmatrix} 187.2 \\ 154.2 \\ \vdots \\ 157.1 \end{bmatrix}$$

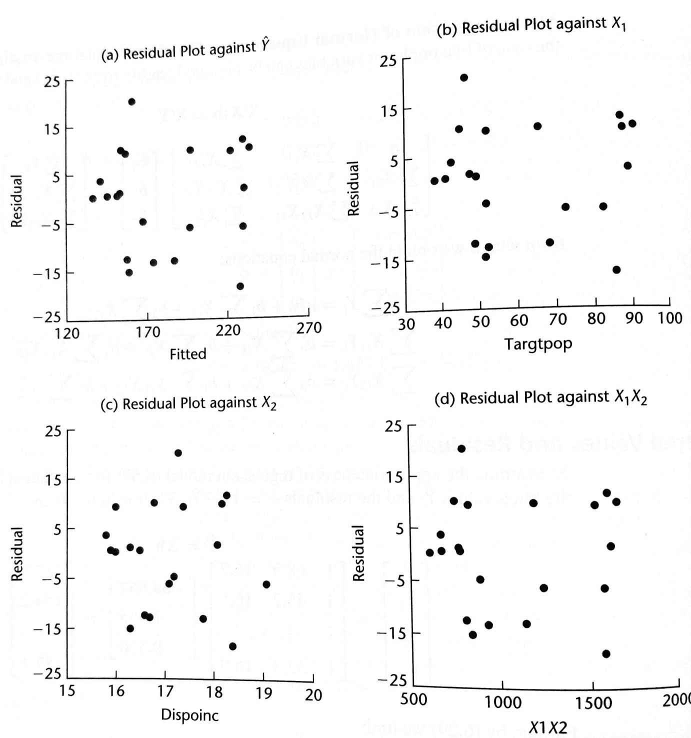
2. The residuals

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \begin{bmatrix} -12.8 \\ 10.2 \\ \vdots \\ 9.4 \end{bmatrix}$$

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 against the fitted values Y in Figure 6.8a. This plot does not suggest any ____________________ 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 ___________ by the response function and ___________________ of the error terms.

3. If a plot of the residuals e against the interaction term X_1X_2 shows a ____________________, that means an interaction effect may be present, so that a response function of the type

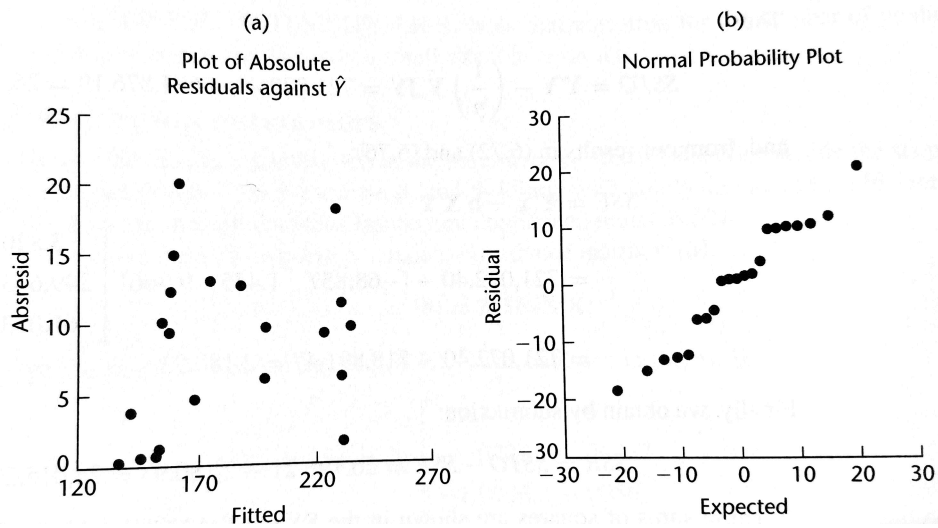
   ____

   might be more appropriate.

4. (Figure 6.8d) Plot does not exhibit any ____________________; 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 ________________ 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 ___________________ pattern.

7. The coefficient of ______________ between the ordered residuals and their expected values under normality is _________. 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 ____________ 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 _______________. 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