CHAPTER II

FUNDAMENTAL CONCEPT OF DESIGN AND ANALYSIS OF EXPERIMENTAL RESULTS BY THE REGRESSION TECHNIQUE

2.1 Concept of Regression Techniques

2.1.1 Introduction

Most problems have two or more interrelated variables, and it is often very interesting to model and explore this relationship. For example, in a chemical process, if the yield of product is related to the operating temperature, the chemical engineer may need to build a model that shows relationship between product yield and temperature and then use it for prediction, process optimization and/or process control.

Suppose that there is a single dependent variable or response y that depends on k independent variables, called "regressor", i.e. $x_1, x_2, \ldots, x_k$ .The relationship among these variables is characterized by a mathematical model called the regression model. The regression model is determined by fitting a set of sample data into a desired equation with unknown variables. In some situations, the experimenter knows the exact form of the true functional relationship between y and $x_1, x_2,\ldots, x_k$, i.e. $y = \phi(x_1, x_2,\ldots, x_k)$. However, in most cases, the true functional relationship is unknown, and an appropriate function is chosen by the experimenter to approximate ϕ. Low-order polynomial models are generally used as approximate functions.

Practical regression models as representatives of the experimental results will improve understanding, interpretation, and implementation of studied variables. Since there is a strong interplay between design of experiments and regression analysis, success of quantitative expression of experimental results depends upon the basis empirical regression models.

Regression methods are frequently used to analyze data from unplanned experiments, such as those arising from observation of uncontrolled phenomena or historical records. Regression methods are also very useful in designed experiments where something has "gone wrong."

2.1.2 Linear Regression Models

In the standard linear regression model, three assumptions are made about the relation between Y and X:

1. For each selected X, there is a normal distribution of Y from which the sample value can be randomly selected. If desired, more than one Y may be drawn from each distribution.

2. The population of values of Y corresponding to a selected X has a mean μ that lies on the straight line

$$\mu = \alpha + \beta(X - \bar{X}) = \alpha + \beta X,$$

where α and β are parameters to be determined.

3. In all populations, the standard deviation of Y around its mean $\alpha + \beta X$ is equal, denoted by σ_{yx}

The mathematical model is specified concisely by the equation

$$Y = \beta_0 + \beta X + \varepsilon, \qquad (2\text{-}1)$$

Where ε is a random variable drawn from the normal distribution of $N(0, \sigma_{yx})$

In this model, Y is the sum of the random term, ε, and the statement function of X. The function of X, according to the Assumption (2) above, determines the mean of the populations for each individual X. For every X, the mean lies on the straight line represented by $\mu = \alpha + \beta X$, which is the population regression line. The parameter β_0 is the mean of the population that corresponds to $X = 0$; thus, β_0 specifies the height of the line when $X = 0$. β is the slope of the regression line, the change in Y per unit increase in X. As for the variable part of Y, ε is drawn randomly from $N(0, \sigma_{yx})$; therefore, it is independent of X and normally distributed, as the symbol N signifies.

2.1.3 Estimation of the Parameter in Linear and Multiple Linear Regression Models

2.1.3.1 Estimation of the Parameter in Linear and Multiple Linear Regression Models

The method of least squares is typically used to estimate the regression coefficients in a multiple linear regression model. Suppose that $n > k$ (n : the number of dependent variable and k : the number of independent variables) observations on the response variables are available, i.e. $y_1, y_2, \ldots, y_n$. Along with each observed response y_i , we will have an observation on each regressor variable and let x_{ij} denote the i^{th} observation or level of variable x_j . The data will appear as shown in Table (2-1). We assume that the error term ε in the model has $E(\varepsilon) = 0$ and $V(\varepsilon) = \sigma^2$ and that the $\{\varepsilon_i\}$ is the matrix of uncorrelated random variables.

Table 2.1 Data for Multiple Linear Regression

y	x_1	x_2	...	x_k
y_1	x_{11}	x_{12}	...	x_{1k}
y_2	x_{21}	x_{22}	...	x_{2k}
:	:	:		:
y_n	x_{n1}	x_{n2}	...	x_{nk}

We may write the model equation (Equation 2-1) in terms of the observations in Table (2-1) as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i \qquad (2\text{-}2)$$

$$= \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i \qquad i = 1, 2, \ldots, n$$

The method of least squares chooses the β_i ($i = 0, 1, \ldots, k$) (Equation 2-2) so that the sum of the squares of the errors, ε_i is minimized. The least squares function is

$$L = \sum_{i=1}^{n} \varepsilon_i^2 \qquad (2\text{-}3)$$

$$= \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{2} \beta_j x_{ij})^2$$

The function L is to be minimized with respect to $\beta_0, \beta_1, \ldots, \beta_k$. The least squares estimators, say $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$, must satisfy

$$\left.\frac{\partial L}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij}) = 0 \qquad (2\text{-}4a)$$

and

$$\left.\frac{\partial L}{\partial \beta_j}\right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij})\, x_{ij} = 0 \qquad j = 1, 2, \ldots, k \qquad (2\text{-}4b)$$

Simplifying Equation (2-4), we obtain

$$
\begin{aligned}
n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} + \ldots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik} &= \sum_{i=1}^{n} y_i \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i2}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1}x_{i2} + \ldots + \hat{\beta}_k \sum_{i=1}^{n} x_{i1}x_{ik} &= \sum_{i=1}^{n} x_{i1}y_i \qquad (2\text{-}5) \\
\vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad &\quad \vdots \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \hat{\beta}_1 \sum_{i=1}^{n} x_{ik}x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{ik}x_{i2} + \ldots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik}^2 &= \sum_{i=1}^{n} x_{ik}y_i
\end{aligned}
$$

These equations are called the least squares normal equations. Note that there are $p = k + 1$ normal equations, one for each of the unknown regression coefficients. The solution to the normal equations will be the least squares estimators of the regression coefficients $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$

2.1.3.2 The estimator of σ^2 by the Method of Least Square

It is also usually necessary to estimate σ^2. To develop an estimator of this parameter, consider the sum of squares of the residuals, say

$$SS_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (2\text{-}6)$$

$$= \sum_{i=1}^{n} e_i^2 \qquad (2\text{-}7)$$

Equation 2-6 is called the error or residual sum of squares, and it has $n - p$ degrees of freedom associated with it. It can be shown that

$$E(SS_E) = \sigma^2 (n - p)$$

so an unbiased estimator of σ^2 is given by

$$\hat{\sigma}^2 = \frac{SS_E}{n - p} \qquad (2\text{-}8)$$

2.1.4 Hypothesis Testing

2.1.4.1 Hypothesis Testing

2.1.4.1.1 Hypothesis Testing of the Significance in Linear and Multiple Linear Regression Models

The test for significance of regression is a test to determine if there is a linear relationship between the response variable y and the subset of the regressor variables x_1, x_2, ...,x_k . The appropriate hypotheses are

$$H_0 : \underline{\beta}_1 = \underline{\beta}_2 = \ldots = \underline{\beta}_k = 0 \qquad (2\text{-}9)$$
$$H_1 : \underline{\beta}_j \neq 0 \qquad \text{for at least one j}$$

Rejection of H_0 in Equation (2-9) implies that at least one of the regressor variables x_1, x_2,...,x_k contributes significantly to the model. The test procedure involves an analysis of variance partitioning of the total sum of squares (SS_T) into a sum of squares due to the model (or to regression) (SS_R) and a sum of squares due to residual (or error) (SS_E)

To show the above detail, we use the separation of variation

$$y_i - \overline{y} = (\hat{y}_i - \overline{y}) + (y_i - \hat{y}_i) \qquad (2\text{-}10)$$

Evaluate the sum of square of equation (2-10)

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + 2\sum_{i=1}^{n}(\hat{y}_i - \overline{y})(y_i - \hat{y}_i)$$

$$2\sum_{i=1}^{n}(\hat{y}_i - \overline{y})(y_i - \hat{y}_i) = 2\sum_{i=1}^{n}\hat{y}_i(y_i - \hat{y}_i) - 2\overline{y}\sum_{i=1}^{n}(y_i - \hat{y}_i)$$

$$= 2\sum_{i=1}^{n}\hat{y}_i e_i - 2\overline{y}\sum_{i=1}^{n}e_i$$

$$= 0$$

Then

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (2\text{-}11)$$

According to equation (2-11), Term $\sum_{i=1}^{n}(y_i - \bar{y})^2$, the total sum of square, SS_T, is divided into two terms. Term $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ is a sum of square due to model (or regression), SS_R and Term $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is a sum of square due to residual, SS_E

$$SS_T = SS_R + SS_E \qquad (2\text{-}12)$$

Now if the null hypothesis $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$ is true, then SS_R/σ^2 is distributed as X_k^2, when the number of degrees of freedom for X^2 is equal to the number of regressor variables in the model. Also, we can show that SS_E / σ^2 is distributed as X^2_{n-k-1} and that SS_E and SS_R are independent. The test procedure for $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$ is to compute

$$F_0 = \frac{SS_R / k}{SS_E / (n-k-1)} = \frac{MS_R}{MS_E} \qquad (2\text{-}13)$$

And to reject H_0 if F_0 exceeds $F_{\alpha, k, n-k-1}$. Alternatively, we could use the *P*-value approach to hypothesis testing and, thus, reject H_0 if the *P*-value for the statistic F_0 is less than α. The test is usually summarized in an analysis of variance shown in Table (2-2).

Table 2.2 Analysis of Variance for Significance of Regression in Multiple Regression

Source of Variation	Sum of Squares	Degree of Freedom	Mean Square	F_0
Regression	$SS_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	k	$MS_R = \frac{SS_R}{k}$	MS_R/MS_E
Error or residual	$SS_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$n-k-1$	$MS_E = \frac{SS_E}{n-k-1}$	
Total	$SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2$	$n-1$		

2.1.4.1.2 Hypothesis Testing of the Parameter in Linear and Multiple Linear Regression Models

We are frequently interested in testing hypotheses on the individual regression coefficients. Such tests would be useful in determining the value of each of the regressor variables in the regression model. For example, the model might be more effective with the inclusion of additional variables or perhaps with the deletion of one or more of the variables already in the model.

Adding a variable to the regression model always causes the sum of squares for regression to increase and the error sum of squares to decrease. We must decide whether the increase in the regression sum of squares is sufficient to warrant using the additional variable in the model. In fact, adding an unimportant variable to the model can actually increase the mean square error, thereby decreasing the usefulness of the model.

The hypotheses for testing the significance of any individual regression coefficient, say β_j, are

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0$$

If $H_0 : \beta_j = 0$ is not rejected, then it indicates that X_j can be deleted from the model. The test statistic for this hypothesis is

$$t_0 = \frac{\hat{\beta}_j}{(\hat{\sigma}^2 C_{jj})^{1/2}} \qquad (2\text{-}14)$$

If $H_0 : \beta_j = 0$ is rejected, it means $|t_0| > t_{\alpha/2,\, n-k-1}$. Note that this is really a partial or marginal test because the regression coefficient, $\hat{\beta}_j$, depends on all the other regressor variables X_i $(i \neq j)$ present in the model.

The denominator of Equation (2-14), $(\hat{\sigma}^2 C_{jj})^{1/2}$, is often called the standard error of the regression coefficient $\hat{\beta}_j$ *se* $(\hat{\beta}_j)$. That is

$$se\,(\hat{\beta}_j) = (\hat{\sigma}^2 C_{jj})^{1/2} \qquad (2\text{-}15)$$

Therefore, an equivalent way to write the test statistic in Equation (2-14) is

$$t_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \qquad (2\text{-}16)$$

Most regression computer programs provide the t test for each model parameter.

2.1.4.2 Hypothesis Testing of the Confidence Intervals

2.1.4.2.1 Confidence Intervals on the Individual Regression Coefficients

Because the least squares estimator $\hat{\beta}$ is a linear combination of the observations, it follows that $\hat{\beta}$ is normally distributed with mean vector β and covariance matrix $\sigma^2(X'X)^{-1}$. Then, each of the statistics

$$\frac{\hat{\beta}_j - \beta_j}{(\hat{\sigma}^2 C_{jj})^{1/2}} \qquad j = 0, 1, \ldots, k \qquad (2\text{-}17)$$

is distributed as t with n-p degrees of freedom, where C_{jj} is the $(jj)^{th}$ element of the $(X'X)^{-1}$ matrix, and $\hat{\sigma}^2$ is the estimate of the error variance, obtained from Equation (2-8). Therefore, a $100(1-\alpha)$ percent confidence interval for the regression coefficient β_j, $j = 0, 1, \ldots, k$, is

$$\hat{\beta}_j - t_{\alpha/2,n-p}(\hat{\sigma}^2 C_{jj})^{1/2} \leq \beta_j \leq \hat{\beta}_j + t_{\alpha/2,n-p}\ (\hat{\sigma}^2 C_{jj})^{1/2} \qquad (2\text{-}18)$$

Note that this confidence interval could also be written as

$$\hat{\beta}_j - t_{\alpha/2,n-p}\ se(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + t_{\alpha/2,n-p}\ se(\hat{\beta}_j)$$

Since $se(\hat{\beta}_j) = (\hat{\sigma}^2 C_{jj})^{1/2}$