# A GOAL PROGRAMMING APPROACH FOR THE PROBLEMS ANALYZED USING THE METHOD OF LEAST SQUARES

BY

MAIZAH HURA AHMAD
ROBIAH ADNAN
ZALINA MOHD DAUD
LAU CHIK KONG

RESEARCH VOT NO:

75018

Universiti Teknologi Malaysia

2005

# ABSTRACT

Goal programming (GP) is one of the most promising techniques for multiple objective decision analysis. Goal programming is a powerful tool which draws upon the highly developed and tested technique of linear programming, but provides a simultaneous solution to a complex system of competing objectives. In decision analysis, the least squares method is also a popular technique. It is an approach used in the study of relations between variables, particularly for the purpose of understanding how one variable depends on one or more other variables. However, one of the main problems is that the method of least squares is biased by extreme cases. This study proposes goal programming as an alternative to analyze such problems. The analysis were done by using QM for Windows and MINITAB software package.

# ABSTRAK

Pengaturcaraan gol adalah satu kaedah yang berkesan dalam penganalisisan keputusan objektif berganda. Ia juga merupakan suatu teknik yang lebih baik berbanding pengaturcaraan linear dalam penyelesaian serentak untuk sistem kompleks. Dalam membuat keputusan, kaedah kuasa dua terkecil dalam analisis regresi juga adalah satu teknik yang terkenal. Kaedah ini mengkaji hubungan antara pembolehubah terutama dalam memahami bagaimana satu pembolehubah bersandar kepada satu atau lebih pembolehubah yang lain. Walaubagaimanapun, masalah utama bagi kaedah kuasa dua terkecil ialah pengaruh kes ekstrim. Kajian ini mencadangkan pengaturcaraan gol sebagai kaedah alternatif untuk mengatasi masalah tersebut. Kajian ini menggunakan program QM for Windows dan MINITAB dalam analisis.

## TABLE OF CONTENTS

## LIST OF TABLES

# LIST OF FIGURES

# LIST OF SYMBOLS

| | | |
|---|---|---|
| $iqr$ | - | Interquartile range |
| $p(x)$ | - | Polynomial function / fitting curve |
| $X$ | - | Independent (predictor) variable |
| $Y$ | - | Dependent (response) variable |
| $Z$ | - | Objective function |
| $\beta_x$ | - | Parameter of regression equation |
| $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ | - | Sum of the squares of the error / least squares / ordinary least squares |
| $a_{ij}$ | - | The coefficient associated with variable $j$ in the $i$th goal |
| $P_0$ | - | Super priority factor/artificial objective function |
| $P_{\mathrm{k}}$ | - | The priority factor of the $k$th goal |
| $r_{k,s}$ | - | The index number for priority $k$ under $s$th basic or nonbasic variable |
| $u_i$ | - | The function of preemptive factors and weights associated with the $i$th basic variable |
| $v_s$ | - | The function of preemptive priority factors and weights associated with the $s$th basic or nonbasic variable |
| $x_b$ | - | Basic variable |
| $x_j$ | - | The $j$th decision variable |
| $y_{i,s}$ | - | Element in the $i$th row under the $s$th basic or nonbasic variable. |
| $d_i^-$ | - | Negative deviational variable from $i$th goal (underachievement) |
| $d_i^+$ | - | Positive deviational variable from $i$th goal (overachievement) |
| $W_i^-$ | - | Positive numerical weight assigned to the negative deviational variable, $d_i^-$ of the $i$th constant |
| $W_i^+$ | - | Positive numerical weight assigned to the positive deviational variable, $d_i^+$ of the $i$th constant |
| $x_i$* | - | The optimal value of decision variables |
| $\hat{y}_i$ | - | Predicted / estimated value |

# CHAPTER 1

# RESEARCH FRAMEWORK

## 1.0 Introduction

A number of techniques have been proposed for multiple-objective decision making. One of the most promising techniques for multiple objective decision analysis is goal programming (GP). GP is a powerful tool which draws upon the highly developed and tested technique of linear programming and at the same time provides a simultaneous solution to a complex system of competing objectives (Lee,1981). GP can handle decision problems having a single goal with multiple subgoals.

GP has been widely accepted and applied technique mainly because of its underlying philosophy of "satisficing" (Lee and Shim, 1986). Nobel laureate Herbert A. Simon suggested that the satisficing approach, rather than optimizing is based on the concept of bounded rationality. This approach has emerged as a pragmatic methodology of decision making.

Another popular tool in decision making is regression. It is an approach used to study the relationships between variables, particularly for the purpose of understanding how one variable depends on one or more other variables. By identifying the relationship between variables, regression analysis helps to develop a prediction equation.

## 1.1 Research Background

A regression model is a mathematical equation that describes the relationship between two or more variables. The dependent variable is the one being explained, and the independent variables are the ones used to explain the variation in the dependent variable.

Regression techniques are associated with the fitting of straight lines, curves, or surfaces, to set of observations. The straight line is the simplest curve that can be fitted to a set of n paired observations $(x_1, y_1)$, $(x_2, y_2)$ ... $(x_n, y_n)$. The least squares method is the most frequently used procedure for obtaining a linear function. A problem of fitting occurs only if the fit is for some reason imperfect. To be a statistical problem there must be some random element present in the data which leads to this inexactitude of fit. It is the nature of this random element that determines the appropriate method of fitting, i.e. of estimating the constants or parameters in the equation.

In simple linear regression analysis, the estimated regression model is $\hat{y} = a + bx$ ($y$ denotes the predicted dependent variable and $x$ denotes the independent variable). In multiple regressions, the estimated regression model is $\hat{y} = a + \sum_{i=1}^{n} b_i x_i$ (y denotes the predicted dependent variable and $x_i$ denotes the independent variables). Although the method of least squares is one of the best known and probably widely used method employed in the analyses of making predictions of dependent variables based on independent variables, most previous efforts in this area however, suffer from several disadvantages. One of the main problems is that the method of least squares is biased by extreme cases (Campbell, 1972). The current study proposes GP as an alternative to analyze such problems.

## 1.2 Objectives of the Study

The objectives of this study are as follows:

i. To identify the types of problems analyzed by the least squares method that can be solved through the GP approach.

ii. To discuss how least squares problems can be converted into GP problems.

iii. To develop prediction equations using both the least squares and the GP methods.

iv. To compare the performances of the prediction equations obtained from both the least squares and the GP methods.

## 1.3 Importance of the Study

Prediction equations have been obtained using the least squares method. These equations have been used in various areas such as educational system planning, financial planning and economic policy analysis. This study explores GP as an alternative method to produce prediction equations.

## 1.4 Scopes of the Study

This study focuses on the use of the linear GP method to produce prediction equations in regression analysis problems. Only three data sets are considered. The first set consists of only one independent variable, the second set has two independent variables while the third set has three independent variables. The analysis are done by using QM for Windows and MINITAB software package.

## 1.5 Organization of the Report

There are six chapters. Chapter I discusses the research framework. It begins with the introduction to the goal programming and the least squares method. The objectives, importance and scope of this study are also presented.

In Chapter II, the modeling of goal programming is presented. This chapter starts with the background of goal programming. Formulation and methodology of the goal programming model are also discussed.

Chapter III reviews the least squares method. In this chapter, the least squares line and multiple regression least squares are discussed.

Chapter IV starts with the discussion of outliers in data sets. It proceeds with the analysis of data sets using both the least squares and goal programming methods.

In chapter V, comparison between the least squares and goal programming are made.

Chapter VI summarizes and concludes the whole study and makes some possible suggestions for future investigation.

## 1.6 Terminology

### Box Plot

This is a graphical display to detect outliers in a data set (Mendenball, 1993).

### Conflicting Goal

Two goals are conflicting if the level of achievement of one of the goals cannot be increased without simultaneously reducing the level of achievement of the other goal.

### Decision Variable

A decision variable, denote as $x_i$ (with $i = 1, 2, ..., j$) is a variable that is both under the control of the decision maker and one that can have an impact on the problem solution. All decision variables will be assumed nonnegative unless otherwise noted (Ignizio, 1976).

### Dependent Variable

The variable of interest in a regression equation, which is said to be functionally related to one or more independent or predictor variables (Mendenball, 1993).

### Deviational Variable

Auxiliary variables in a goal constraint equation that measure the underachievement or overachievement of the specified aspiration level. A negative deviation variable ($d_i^- \geq 0$) reflects the amount by which aspiration level $i$ is underachieved, while a positive deviational variable ($d_i^+ \geq 0$) indicates the amount by which aspiration level $i$ is exceeded, where $d_i^- \times d_i^+ = 0$.

### Feasible solution

Any set of nonnegative $x_i$, $d_i^-$ and $d_i^+$ values constitute a feasible solution (Ignizio, 1976).

### Goal Constraint

A set of constraints that corresponds to the goals expressed by the decision maker.

### Independent Variable

A nonrandom variable related to the response in a regression equation. One or more independent variables may be functionally related to the dependent variable. They are used in the regression equation to predict or estimate the value of the dependent variable (Mendenball, 1993).

### Optimal Solution

The solution ($\bar{x}$) to a given goal programming model is considered optimal if, for this solution (termed $\bar{x}$ *), the corresponding value of $\bar{g}$ (termed $\bar{g}$ *) is the same or preferred to the value of $\bar{g}$ for any other feasible solution. Note that the vector $\bar{g}$ * will be preferred to the vector $\bar{g}$ if the first nonzero component of ($\bar{g}$ * - $\bar{g}$) is negative, given that all elements of $\bar{g}$ * and $\bar{g}$ are themselves nonnegative (Ignizio, 1976).

### Pivot Element

The element of a simplex tableau occurring at the intersection of the column associated with an incoming basic variable and the pivot row.

### Pivot Row

The row of a simplex tableau in which the minimum nonnegative ratio occurs. This row is associated with the variable that will leave the basis in the next simplex iteration.

### Prediction Equation

In regression the equation used to predict the values of the dependent variable $y$ for specified values of the independent variables $x_1, x_2, \ldots, x_k$. This equation is generally obtained using the method of least squares (Mendenball, 1993).

### Preemptive Priority Factors

Priority factors $P_j$ ($j = 1, \ldots, K$; where $K$ is the number of objectives in the model) that have the following relationship

$$P_1 >>> P_2 >>> \ldots >>> P_j >>> P_{j+1}$$

where >>> implies "infinitely greater than".

### Residual

The difference between the observed value of $y$ and the value predicted ($\hat{y}$) by a model, ($y - \hat{y}$), is referred to as the error or residual (Mendenball, 1993).

# CHAPTER 2

# LINEAR GOAL PROGRAMMING

## 2.0 Introduction

Goal programming problems can be classified according to the types of mathematical programming models such as linear programming, integer programming and nonlinear programming. These goal programming problems have multiple goals instead of a single objective (Hillier and Lieberman, 2001). In this study, only the linear goal programming model is considered. These are goal programming problems that fit linear programming where each objective function is linear.

This chapter will discuss the history of goal programming, advantages and disadvantages of the goal programming, the formulation of goal programming and the solution method of goal programming

## 2.1 History of Goal Programming

Goal programming was extended from linear programming. It was first developed and introduced by A. Charnes and W.W. Cooper in 1961. It was further refined by Y. Ijiri in 1965. In 1968 B. Contini considered goal programming under conditions of uncertainty. Major applications were developed by V. Jaaskelainen, S.

Lee and J.P. Ignizio in the 1970s. Since 1968, many goal programming related studies have been published.

Goal programming has become a widely accepted and applied technique in various functional areas such as academic planning and administration, accounting analysis, advertising media scheduling, capital budgeting, decision-support system design, economic policy analysis, energy resources planning, financial planning, inventory management, marketing logistics, military strategies and planning, organizational analysis, production scheduling, quality control, urban planning and predicting student performance (Ignizio, 1976; Lee and Shim, 1986).

## 2.2 Advantages and Disadvantages of the Goal Programming

Goal programming is one of the popular and powerful methods for multiple objective decision analysis (Lee and Shim, 1986).

The following are some of the advantages of goal programming (Hughes and Grawoig, 1973):

a) Allows for an ordinal ranking of goal, where the low-priority goals are considered only after higher-priority goals have been satisfied to the fullest extent possible.

b) Useful in situations where the multiple goals are conflicting and cannot all be fully achieved.

c) Used to "satisfice" rather than to "optimize" the problem. In linear programming, what one wants is to optimize the solution. But in using the goal programming, the goal may be incorporated into the model at a value that is judged to be satisfactory, not necessarily optimal.

d) Appropriate to find a satisfactory solution where many objectives or goals are to be considered.

However there are some disadvantages of goal programming. These include the following:

a) More time and thought, is required in the construction of the model.
b) More decision-maker involvement is required, that is in the establishment of aspiration levels and weightings.
c) The subjectivity regarding the weights given to priority levels to goal deviations may be of concern.

## 2.3 Goal Programming Model Formulation

The formulation of goal programming problem is very similar to that of linear programming problems (Wu and Coppins, 1981). Goal programming extends the linear programming formulation to accommodate mathematical programming with multiple objectives (Charnes and Cooper, 1961). The major differences are an explicit consideration of goals and the various priorities associated with the different goals.

To formulate goal programming model (Ignizio, 1976), the following steps should be followed:

i. Define the decision variables.
ii. State the system constraints and goal constraints.
iii. Determine the preemptive priority factor and the relative weight (if need be).
iv. Develop the objective function.
v. State the nonnegative requirement.

The objective function in GP is always minimized and must be composed of deviational variables only. It minimizes the deviations of the compromise solution from target goals, weighted and prioritized.

In the formulation, two types of variables are used. They are decision variables and deviational variables. There are two categories of constraints, that is structural or system constraints (strict as in traditional linear programming) and goal constraints, which are expressions of the original functions with target goals set a priori and positive and negative deviational variables.

The general goal programming model can be expressed as follows:

$$\left.\begin{array}{ll} \text{Minimize } Z = \sum_{i=1}^{m} (d_i^- + d_i^+) & \\ \text{Subject to the linear constraints:} & \\ \text{Goal constraints:} & (\sum_{j=1}^{n} a_{ij}x_j) + d_i^- - d_i^+ = b_i,\ i = 1, 2, \ldots, m \\ \text{System constraints:} & \sum_{j=1}^{n} a_{ij}x_j \begin{bmatrix} \le \\ = \\ \ge \end{bmatrix} b_i,\ i = m+1, \ldots, m+p \\ \text{with} & x_j, d_i^-, d_i^+ \ge 0, \text{ for } i = 1, 2, \ldots, \text{m and } j = 1, 2, \ldots, n \end{array}\right\} \quad (2.1)$$

where there are m goals, p system constraints and n decision variables

$Z$ = objective function

$a_{ij}$ = the coefficient associated with variable $j$ in the $i$th goal

$x_j$ = the $j$th decision variable

$b_i$ = the associated right hand side value

$d_i^-$ = negative deviational variable from the $i$th goal (underachievement)

$d_i^+$ = positive deviational variable from the $i$th goal (overachievement)

Both over- and underachievement of a goal cannot occur simultaneously. Hence, either one or both of these variable must have a zero value; that is,

$$d^+ \times d^- = 0$$

Both variables apply for the nonnegativity requirement as to all other linear programming variables; that is,

$$d^+, d^- \ge 0$$

Table 2.1 shows three basic options to achieve various goals:

Table 2.1 : Procedure for Achieving a Goal

| Minimize | Goal | If goal is achieved |
|---|---|---|
| $d_i^-$ | Minimize the underachievement | $d_i^- = 0, d_i^+ \geq 0$ |
| $d_i^+$ | Minimize the overachievement | $d_i^- \geq 0, d_i^+ = 0$ |
| $d_i^- + d_i^+$ | Minimize both under- and overachievement | $d_i^- = 0, d_i^+ = 0$ |

### 2.3.1 Preemptive Goal Programming

Before solving a goal programming problem, the goals need to be ranked. Preemptive goal programming is also called non-Archimedean or lexicographic goal programming (Ignizio, 1983, 1985a). In priority goal programming, the objectives can be divided into different priority classes. Here it is assumed that no two goals have equal priority. The goals are given ordinal rankings and are called *preemptive priority factors*. These *preemptive priority factors* have the relationship

$$P_1 >>> P_2 >>> \ldots >>> P_k >>> P_{k+1}$$

where >>> means "very much greater than". This priority ranking is absolute. Therefore, the $P_1$ goal is so much more important than the $P_2$ goal and $P_2$ goal will never be attempted until the $P_1$ goal is achieved to the greatest extent possible.

The priority relationship implies that multiplication by n, however large it may be, cannot make the lower-level goal as important as the higher goal (i.e, $P_j > nP_{j+1}$). In formulating a goal programming model having prioritized goals, those preemptive priority factors are incorporated into the objective function as weights for the deviational variables.

Using equation (2.1), the preemptive goal programming model can be presented as:

$$\text{Minimize } Z = \sum_{i=1}^{m} P_k (d_i^- + d_i^+)$$

Subject to the linear constraints:

$$\left.\begin{array}{ll} \text{Goal constraints:} & \sum_{j=1}^{n} a_{ij}x_j + d_i^- - d_i^+ = b_i,\ i = 1, 2, \ldots, m \\ \text{System constraints:} & \sum_{j=1}^{n} a_{ij}x_j \begin{bmatrix} \le \\ = \\ \ge \end{bmatrix} b_i,\ i = m+1, \ldots, m+p \\ \text{with} & x_j, d_i^-, d_i^+ \ge 0,\ i = 1, 2, \ldots, m \text{ and } j = 1, 2, \ldots, n \end{array}\right\} \quad (2.2)$$

where there are m goals, p system constraints, k priority levels and n decision variables

$P_k$ = the priority factor of the $k$th goal

Here, the difference between equation (2.1) and (2.2) is the priority factor in the objective function.

### 2.3.2 Weighted Goal Programming

The weighting of deviational variables at the same priority level should be considered in the goal programming model formulation. These weights show the relative importance of each deviation. Charnes and Cooper (1977) stated the weighted goal programming model as follows:

$$\text{Minimize } Z = \sum_{i=1}^{m} (W_i^- d_i^- + W_i^+ d_i^+)$$

Subject to the linear constraints:

$$\left.\begin{array}{ll} \text{Goal constraints:} & \sum_{j=1}^{n} a_{ij}x_j + d_i^- - d_i^+ = b_i,\ i = 1, 2, \ldots, m \\ \text{System constraints:} & \sum_{j=1}^{n} a_{ij}x_j \begin{bmatrix} \le \\ = \\ \ge \end{bmatrix} b_i,\ i = m+1, \ldots, m+p \\ \text{with} & x_j, d_i^-, d_i^+ \ge 0,\ i = 1, 2, \ldots, m \text{ and } j = 1, 2, \ldots, n \end{array}\right\} \quad (2.3)$$

where there are m goals, p system constraints and n decision variables

$W_i^-$ = positive numerical weight assigned to the negative deviational variable, $d_i^-$ of the *i*th constraint

$W_i^+$ = positive numerical weight assigned to the positive deviational variable, $d_i^+$ of the *i*th constraint

While Ijiri (1965) had introduced the idea of combining preemptive priorities and weighting, Charnes and Cooper (1977) suggested the goal programming model as:

$$\text{Minimize } Z = \sum_{i=1}^{m}\sum_{k=1}^{n} P_k (W_{i,k}^- d_i^- + W_{i,k}^+ d_i^+)$$

Subject to the linear constraints:

$$\left.\begin{array}{ll} \text{Goal constraints:} & \sum_{j=1}^{n} a_{ij}x_j + d_i^- - d_i^+ = b_i,\ i = 1, 2, \ldots, m \\ \text{System constraints:} & \sum_{j=1}^{n} a_{ij}x_j \begin{bmatrix} \le \\ = \\ \ge \end{bmatrix} b_i,\ i = m+1, \ldots, m+p \\ \text{with} & x_j, d_i^-, d_i^+ \ge 0,\ i = 1, 2, \ldots, m \text{ and } j = 1, 2, \ldots, n \end{array}\right\} \qquad (2.4)$$

where there are m goals, p system constraints, k priority levels and n decision variables

$z$ = objective function

$P_k$ = the priority factor of the *k*th goal

$W_i^-$ = positive numerical weight assigned to the negative deviational variable, $d_i^-$ of the *i*th constraint

$W_i^+$ = positive numerical weight assigned to the positive deviational variable, $d_i^+$ of the *i*th constraint

$d_i^-$ = negative deviational variable from the *i*th goal (underachievement)

$d_i^+$ = positive deviational variable from the *i*th goal (overachievement)

$a_{ij}$ = the coefficient associated with variable *j* in the *i*th goal

$x_j$ = the *j*th decision variable

$b_i$ = the associated right hand side value

## 2.4 Solution Method of Goal Programming

In this section, two types of goal programming solution methods will be discussed; that is, the graphical method and the modified simplex method.

### 2.4.1 The Graphical Method

The graphical method is useful for those goal programming problems which involve only two decision variables.

In goal programming, we try to minimize the deviation from the goal with the highest priority to its fullest possible extent. Then the goal with the second higher priority factor is considered, and so on. The sequential "satisficing" procedure is used in the graphical method.

According to Ignizio (1976), the steps of the graphical approach are:

1. Plot all the system and goal constraints in terms of the decision variables (these will simply be straight lines or planes in a linear model).
2. Determine the solution(s) space for the priority 1 goals.
3. Move to the set of goals having the next-highest priority and determine the "best" solution space for this set of goals, where this "best" solution cannot degrade the achievement values already obtained for higher-priority goals.
4. If, at any time in the process, the solution space is reduced to a single point, terminate the procedure since no further improvement is possible.
5. Repeat steps 3 and 4 until either we converge to a single point or we have evaluated all the priority levels.

### 2.4.2 The Modified Simplex (Multiphase) Method

The modified simplex method is a general solution technique for all types of goal programming problems. It is an iterative algorithm just like the regular simplex method for linear programming. Because of the unique features of the goal programming model, a number of modifications are necessary in the simplex operation.

To apply this method, the first thing we need to do is to develop the initial modified simplex tableau. The general initial modified simplex tableau is shown in Table 2.2.

**Table 2.2 : The General Initial Modified Simplex Tableau**

| $C_j$ | | | $v_1$ | $v_2 \dots v_n$ | $v_{n+1}\ v_{n+2} \dots v_{n+m}$ | $v_{n+m+1}\ v_{n+m+2} \dots v_{n+2m}$ |
|---|---|---|---|---|---|---|
| $C_b$ | Basic variables $x_b$ | Solution **b** | $x_1$ | $x_2 \dots x_n$ | $d_1^-\ d_2^- \dots d_m^-$ | $d_1^+\ d_2^+ \dots d_m^+$ |
| $u_1$ | $d_1^-$ | $b_1$ | $y_{1,1}$ | $y_{1,2} \cdots y_{1,n}$ | $y_{1,n+1} \cdots y_{1,n+m}$ | $y_{1,n+m+1} \cdots y_{1,n+2m}$ |
| $u_2$ | $d_2^-$ | $b_2$ | $y_{2,1}$ | $y_{2,2} \cdots y_{2,n}$ | $y_{2,n+1} \cdots y_{2,n+m}$ | $y_{2,n+m+1} \cdots y_{2,n+2m}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $u_m$ | $d_m^-$ | $b_m$ | $y_{m,1}$ | $y_{m,2} \cdots y_{m,n}$ | $y_{m,n+1} \cdots y_{m,n+m}$ | $y_{m,n+m+1} \cdots y_{m,n+2m}$ |
| $Z_j - C_j$ | $P_K$ | $g_K$ | $r_{K,1}$ | $r_{K,2} \cdots r_{K,n}$ | $r_{K,n+1} \cdots r_{K,n+m}$ | $r_{K,n+m+1} \cdots r_{K,n+2m}$ |
| | $P_{K-1}$ | $g_{K-1}$ | $r_{K-1,1}$ | $r_{K-1,2} \cdots r_{K-1,n}$ | $r_{K-1,n+1} \cdots r_{K-1,n+m}$ | $r_{K-1,n+m+1} \cdots r_{K-1,n+2m}$ |
| | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| | $P_2$ | $g_2$ | $r_{2,1}$ | $r_{2,2} \cdots r_{2,n}$ | $r_{2,n+1} \cdots r_{2,n+m}$ | $r_{2,n+m+1} \cdots r_{2,n+2m}$ |
| | $P_1$ | $g_1$ | $r_{1,1}$ | $r_{1,2} \cdots r_{1,n}$ | $r_{1,n+1} \cdots r_{1,n+m}$ | $r_{1,n+m+1} \cdots r_{1,n+2m}$ |

where

$j$ = 1, 2, …, n

$i$ = 1, 2, …, m

$k$ = 1, 2, …, K

$s$ = 1, 2, …, S

$x_j$ = the initial set of nonbasic variable

$d_1^+$ = the initial set of nonbasic variable

$d_1^-$ = the initial set of basic variable

$v_s$ = the function of preemptive priority factors and weights associated with the *s*th basic or nonbasic variable

$u_i$ = the function of preemptive priority factors and weights associated with the *i*th basic variable

*bi* = the right hand side value of the *i*th goal

$y_{i,s}$ = element in the *i*th row under the *s*th basic or nonbasic variable. That is, the coefficient of the *s*th basic or nonbasic variable in goal *i*.

$P_k$ = *k*th priority level

$g_k$ = level of achievement of the goals in priority *k*, where $\mathbf{g} = (g_1, g_2, \ldots, g_k)$

$r_{k,s}$ = the index number for priority *k* under *s*th basic or nonbasic variable

All the elements in the initial tableau, except for $r_{k,s}$ and $g_k$ are simply obtained from the mathematical model (2.1). However, $r_{k,s}$ and $g_k$ must be computed as follows:

$$r_{k,s} = \mathbf{u}_k^T \mathbf{y}_s - v_s$$

or

$$r_{k,s} = \sum_{i=1}^{m} (y_{i,s} \cdot u_i) - v_s \tag{2.5}$$

and

$$g_k = \mathbf{u}_k^T \mathbf{b}$$

or

$$g_k = \sum_{i=1}^{m} (b_i \cdot u_i) \tag{2.6}$$

If the system constraints exist in the goal programming model, some further steps have to be taken. The system constraint can exist in the three ways as follow:

1. If the system constraint is $\sum_{j=1}^{n} a_{ij} x_j \le b_i$, a slack variable, $S_i$ will be added to this equation. The equation will become

$$\sum_{j=1}^{n} a_{ij} x_j + S_i = b_i$$

The slack variable, $S_i$ will be defined as the initial basic variable.

2. If the system constraint is $\sum_{j=1}^{n} a_{ij}x_j = b_i$, an artificial variable, $A_i$ will be added to this equation. The equation will become

   $$\sum_{j=1}^{n} a_{ij}x_j + A_i = b_i$$

   The artificial variable, $A_i$ will be used as the initial basic variable.

3. If the system constraint is $\sum_{j=1}^{n} a_{ij}x_j \geq b_i$, an excess or surplus variable, $E_i$ and an artificial variable, $A_i$ will be added to this equation. The equation will become

   $$\sum_{j=1}^{n} a_{ij}x_j - E_i + A_i = b_i$$

   The artificial variable, $A_i$ will be used as the initial basic variable.

Then, a new priority factor, $P_0$ must be introduced. The $P_0$ is defined as the super priority factor, which is the highest priority factor among all the priority factors where $P_0 >>> P_1 >>> P_2 >>> \ldots >>> P_k >>> P_{k+1}$. $P_0$ also represents the artificial objective function. The initial simplex tableau when the system constraints exist is shown in Table 2.3.

**Table 2.3 : The General Initial Simplex Tableau When the System Constraints Exist**

| $C_j$ | | | $v_1 \ldots v_n$ | $\ldots$ | $v_{n+2m+q+r+t}$ |
|---|---|---|---|---|---|
| $C_b$ | Basic variables<br>$x_b$ | Solution<br>**b** | $x_1 \ldots x_n\ d_1^- \ldots d_m^-\ d_1^+ \ldots d_m^+\ S_1 \ldots S_q\ E_1 \ldots E_r\ A_1 \ldots A_t$ | | |
| $u_1$ | $d_1^-$ | $b_1$ | $y_{1,1}\ \cdots\ y_{1,n}$ | $\ldots\ldots$ | $y_{1,n+2m+q+r+t}$ |
| $u_2$ | $\vdots$ | $\vdots$ | $y_{2,1}\ \cdots\ y_{2,n}$ | $\ldots\ldots$ | $y_{2,n+2m+q+r+t}$ |
| | $d_m^-$ | $b_m$ | | | |
| | $S_1$ | $b_{m+1}$ | $\vdots$ | | $\vdots$ |
| $\vdots$ | $\vdots$ | $\vdots$ | | $\ddots$ | |
| | $S_q$ | $b_{m+q}$ | | $\ddots$ | |
| $\vdots$ | $A_1$ | $b_{m+q+1}$ | $\vdots$ | | $\vdots$ |
| | $\vdots$ | $\vdots$ | | | |
| $u_m$ | $A_t$ | $b_{m+q+t}$ | $y_{m+q+t,1}$ | $\ldots\ldots$ | $y_{m+q+t,n+2m+q+r+t}$ |
| $Z_j - C_j$ | $P_K$ | $g_K$ | $r_{K,1}\ \cdots\ r_{K,n}$ | $\ldots\ldots$ | $r_{K,n+2m}$ |
| | $P_{K-1}$ | $g_{K-1}$ | $r_{K-1,1}$ | | $\ldots r_{K-1,n+2m}$ |
| | $\vdots$ | $\vdots$ | $\vdots$ | | $\vdots$ |
| | $P_1$ | $g_1$ | $r_{1,1}$ | $\ldots\ldots$ | $r_{1,n+2m}$ |
| | $P_0$ | $g_0$ | $r_{0,1}\ \cdots\ r_{0,n}$ | $\ldots\ldots$ | $r_{0,n+2m}$ |

The element within the Table 2.4 can be defined as follows:

$i$ = $m$+1, $m$+2, …, $m$+$p$

$k$ = 0, 1, 2, …, K

$S_i$ = slack variable for $i$th goal

$E_i$ = excess or surplus variable for the $i$th goal

$A_i$ = artificial variable for the $i$th goal

$P_0$ = the super priority factor which assigned to the artificial variable, $A_i$ in the objective function

By following the steps given below, the optimal solution to the goal programming model may be derived (Ignizio, 1976).

Step 1: Initialization. Establish the initial modified simplex tableau and the index row for priority level 1 only. Set k = 1 and proceed to Step 2.

Step 2: Check for optimality. Examine $g_k$. If $g_k$ is zero go to Step 7. Otherwise, examine each positive valued index number $r_{k,s}$ in the *k*th index row. Select the largest, positive $r_{k,s}$ for which there are no negative valued index numbers, at a higher priority, in the same column. Designate this column as s'. Ties in the selection of $r_{k,s}$ may be broken arbitrary. If no such $r_{k,s}$ may be found, go to Step 7. Otherwise, go to Step 3.

Step 3: Determining the pivot column and incoming nonbasic variable.

Step 4 Determining the pivot row and outgoing basic variable. Determine the row associated with the minimum nonnegative value of

$$b_i / y_{is'}$$

In the event of ties, select that row having the basic variable with the higher priority level. Designate this row as i'. The basic variable associated with row i' is the outgoing basic variable.

Step 5: Establishment of the new tableau.

(i) Set up a new tableau with all $y_{i,s}$, $b_i$, $r_{k,s}$ and $g_k$ elements empty. Exchange the positions of the basic variable heading in row *i*' (of the previous tableau) with the nonbasic variable heading in column *s*' (of the previous tableau).

(ii) Row *i*' of the new tableau (except for $y_{i',s'}$) is obtained by dividing row *i*' of the previous tableau by $y_{i',s'}$.

(iii) Column *s*' of the new tableau (except for $y_{i',s'}$) is obtained by dividing column *s*' of the previous tableau by $(-y_{i',s'})$.

(iv) The remaining element are computed as follows:

$$\hat{y}_{i,s} = y_{i,s} - \frac{(y_{i',s})(y_{i,s'})}{y_{i',s'}} \qquad (2.7)$$

$$\hat{b}_i = b_i - \frac{(b_{i'})(y_{i,s'})}{y_{i',s'}} \qquad (2.8)$$

where $\hat{b}_i$ and $\hat{y}_{i,s}$ represent the new set of elements to be computed and $b_i$ and $y_{i,s}$ represent the previous values for these element (from previous tableau).

(v) The new values for $r_{k,s}$ and $g_k$ are then established. These values must be computed for the $k$th priority level and all higher priority levels. These can be obtained simply through the use of equations (2.5) and (2.6).

(vi) Return to Step 2.

Step 6: Check the optimality for the new solution.

Step 7: Evaluate the next-lower priority level. Set $k = k+1$. If k exceeds $K$ (the total number of priority levels) then stop as the solution is optimal. If $k \leq K$, establish the index row for $P_k$ and go to Step 2.

In the next section, the relative between goal programming and least square method will be presented.

## 2.6 Regression Analysis for Determining Relative Weighting or Goal Constraint Parameter Estimation

Goal programming in the form of a constrained regression model was used quite some time ago by Charnes, Cooper and Ferguson (1955). By minimizing deviation, the goal programming model can generate decision variable values that are the same as the beta values in some types of multiple regression models. In Charnes, Cooper and Sueyoshi (1986, 1988) it was suggested that their goal programming model serves a valuable purpose of cross checking answers from other methodologies. Likewise, multiple regression models can also be used to more accurately combine multiple criteria measures that can be used in goal programming model parameter (Schniederjans, 1995).

## 2.7 Summary

In this chapter the goal programming method was presented. The relative between goal programming and least square method was also made.

# CHAPTER 3

# LEAST SQUARES METHOD

## 3.0 Introduction

The method of least squares is a powerful technique for regression in statistics (Wonnacott and Wonnacott, 1981). This chapter will first discuss the basics of regression followed by the least squares method.

## 3.1 Regression

Regression analysis is an approach or a research tool in statistics that is used to study the relationships between variables, especially for the purpose of understanding how one variable relates or depends on one or more of other variables (Wittink, 1988).

### 3.1.1 Definition of Regression

'Regression' is often used to indicate "the return to a mean or average value" (Wittink, 1988). More than one hundred years ago, the term regression was introduced to statistics by Francis Galton in a series of paper. The most famous being Galton (1886) is to describe a hereditary phenomenon. In these papers, he reported that the average height of sons with tall fathers is less than the father's height (both measured at adult ages). Similarly, the average height of son with short fathers was reported to be greater than their fathers' height. In his data, Galton emphasized the "regression toward the mean" phenomenon. He also found a positive relationship between the height of fathers and the height of their son which lay approximately on a straight line.

Galton's paper is well worth reading as an example of the many practical considerations that have to be kept in mind in collecting and interpreting data. For Galton, the important point that justified his calling this a regression line was the slope was less then unity (implying the regression, or movement towards the population mean).

Today, any study of relations between variables is often accomplished and referred through regression analysis (Wittink, 1988). The technique is used heavily in business and government activities, and social sciences, especially in economics and related disciplines. It is also a technique for quantifying the relationship between a criterion variable (dependent variable) and one or more predictor variables (independent variables). In particular, the quality of decisions often depends on the quantification of relationships between variables.

### 3.1.2 The Purposes and Benefits of Regression Analysis

Regression may be used for two main purposes. They are

(i) to predict the criterion variable based on specified values for the predictor variable(s), and

(ii) to understand how the predictor variable(s) influence or relate to the criterion variable.

The following are the benefits of using regression analysis. This tool

(i) suggests and quantifies the nature of relations between variables,

(ii) provides consistent predictions,

(iii) may provide superior predictions, and

(iv) may save time or allow a decision maker to focus more time and energy on nonquantifiable aspects.

## 3.2 Simple Regression

**Definition 3.1**

The simple linear regression model assumes that there is a line with vertical or $y$ intercept $a$ and slope $b$, called the true or population regression line. When a value of the independent variable $x$ is fixed and an observation on the dependent variable $y$ is made,

$$y = a + bx + e$$

Without the random deviation $e$, all observed $(x, y)$ points would fall exactly on the population regression line. The inclusion of $e$ in the model equation allows points to deviate from the line by random amounts.

Simple regression has only two variables that is a criterion variable and one predictor variable.

### 3.2.1 Possible Criteria for Fitting a Line

What is a good fit? A Good fit is a fit that makes the total error small (Wannacott and Wannacott, 1981). One typical error (deviation) is shown in Figure 3.1. It can be defined as the vertical distance from the observed $Y_i$ to the fitted value $\hat{Y}_i$ on the line, that is, $Y_i$ - $\hat{Y}_i$. The error is positive if the observed $Y_i$ is above the line and negative when the observed $Y_i$ is below the line.

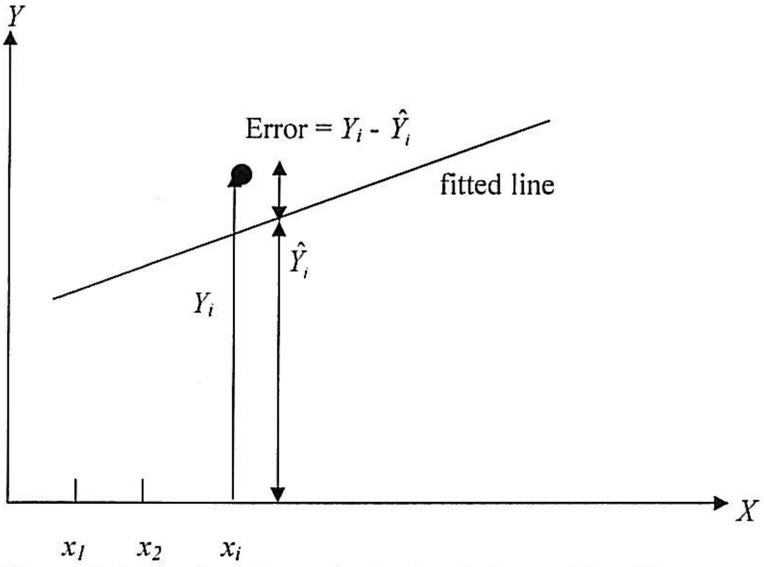

**Figure 3.1: Typical Error in Fitting Points with a Line**

To minimize the total error, the following criteria should be considered (Wannacott and Wannacott, 1981):

(a) A fitted line that minimizes the sum of all these errors can be presented as

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

Using this criterion, two type of fit lines are shown in Figure 3.2 which fit the observations equally well. The fit in panel (a) is intuitively a good one and the fit in panel (b) is a bad one. The problem is concerned with sign's where in both cases, positive error just offset negative errors and leaving their sum equals to zero. This criterion must be rejected because no distinction between bad fits and good ones.

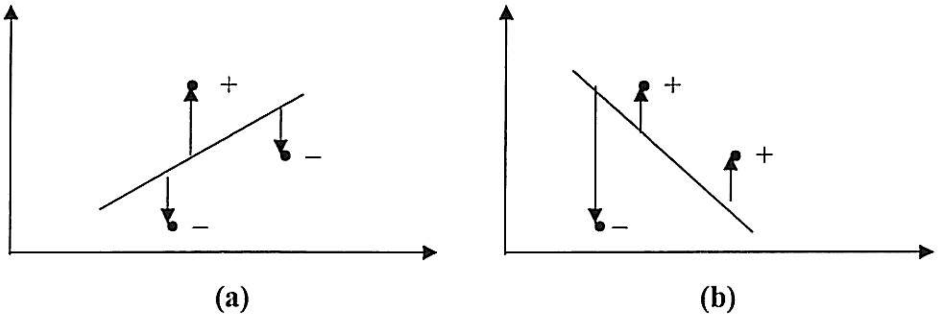

**Figure 3.2 : The Weakness of Using $\sum(Y_i - \hat{Y}_i)$ to Fit a Line**

(b) One of the ways to overcome the sign problem is to minimize the sum of the squares of the errors, that is:

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

This criterion is called least squares, or ordinary least squares (OLS). Its advantages are:

(i) In overcoming the sign problem by squaring the errors, least squares produces very manageable algebra to the geometric theorem of Pythagoras.

(ii) There are two theoretical justifications for least squares, that is the Gauss-Markov theorem and the maximum likelihood criterion for a normal regression model.

### 3.2.2 Using Residuals to Test the Assumptions of the Regression Model

One of the major uses of residual analysis is to test some of the assumptions underlying regression. The following are the assumptions of simple regression analysis.

a) The model is linear.

b) The error terms have constant variance.

c) The error terms are independent.

d) The error terms are normally distributed.

### 3.3 Multiple Regression

Multiple regression is the extension of simple regression. It takes account of more than one independent variable *X*. The appropriate technique should be used when we want to investigate the effect on *Y* of several variables simultaneously. Many times, we wish to include the other variables influencing *Y* in a multiple regression analysis. The reason is:

(i) To reduce stochastic error and hence reduce the residual variance $s^2$. This makes confidence intervals more precise.

(ii) To eliminate bias that might occur if we just ignore a variable that substantially affects *Y*.

### 3.3.1 The Mathematical Model

$Y$ is now to be regressed on the two independent variables $X_1$ and $X_2$. Our model which includes $X_2$ as a predictor variable is

$$Y = a + b_1X_1 + b_2X_2 \tag{3.1}$$

where $b_1$ is geometrically interpreted as the slope of the plane as we move in the $X_I$-direction, keeping $X_2$ constant. Thus $b_1$ is the marginal effect of $X_1$ on $Y$. Similarly $b_2$ is the slope of the plane as we move in the $X_2$-direction, keeping $X_I$ constant; thus $b_2$ is the marginal effect of $X_2$ on $Y$. More generally,

$a$ = the increase in $Y$ if $X$ is increased one unit, while all other regressions are held constant (3.2)

**Proof:**

Suppose that, in addition to $X_1$, there is only one other regressor $X_2$; that is

$$Y = a + b_1X_1 + b_2X_2$$

To establish (3.2), take the partial derivative of $Y$ with respect to $X_1$ in the equation above, i.e.,

$$\frac{\partial Y}{\partial X_1} = b_1$$

We can easily confirm that this simple interpret of $b$ is valid because the regression is linear. If it is not, then $\frac{\partial Y}{\partial X} \neq b$.

For example, if the regression is of the non-linear form

$$Y = a + b_IX + b_2X^2 + cZ$$

Then
$$\frac{\partial Y}{\partial X} = b_1 + 2b_2X$$

To establish (3.2) without calculus, hold $Z$ constant at its initial value $Z_0$, and increase $X$ from its initial $X_0$ to $(X_0 + 1)$. Substituting into the equation above, we may write

$$\begin{array}{l} \text{Initial } Y = a + bX_0 + cZ_0 \\ \text{New } \quad Y = a + b(X_0 + 1) + cZ_0 \\ \hline \text{Difference} = \text{increase in } Y = b \end{array}$$

It is easy to confirm that this is still true if there are several *Z* variables.

To generalize the regression model for problems involving any number of predictor variables, we use

$$Y = a + b_1X_1 + b_2X_2 + \ldots + b_iX_i + e_i \qquad (3.3)$$

where

$Y$ = dependent or response variable,

$X_1, X_2, \ldots, X_i$ = independent or predictor variables,

$e_i$ is the random component of the model and is called the random error.

This model is often referred to as the general linear model. It is general because it allows for an arbitrary number, *i*, of predictor variables. And, for each of the *i* predictor variables specified, the effects are assumed to be linear.

## 3.4 The Method of Least Squares

Least squares method is a computational technique for determining the 'best' equation describing a set of points, $(x_1, y_1)$, $(x_2, y_2)$,... ,and $(x_n, y_n)$, where best is defined geometrically (Larsen and Marx, 2001). It assumes that the best-fit curve of a given type is the curve that has the minimal sum of the deviations squared (least squares error) from a given set of data.

Given data that are relevant to the problem on *Y* and *X*, the most common procedure for computing the intercept, *a*, and the slope coefficient, *b*, is called least squares method. When the values of *a* and *b* are obtained, we can compute a predicted value for *Y*.

Suppose the data points are $(x_1, y_1)$, $(x_2, y_2)$, ..., $(x_n, y_n)$ where *x* is the independent variable and *y* is the dependent variable. The fitting curve or the desired polynomial, *p(x)*, can be written as

$$p(x) = \sum_{k=0}^{m} \beta_k x^k \qquad (3.4)$$

where $\beta_0$, $\beta_1$, …, $\beta_m$ are to be determined. The method of least squares will choose as 'solution' those $\beta_k$'s that minimize the sum of squares of the vertical distances from the data points to the presumed polynomial. It means that the fitting curve $p(x)$ has the deviation (error) $d$ from each data point, i.e., $d_1 = y_1 - p(x_1)$, $d_2 = y_2 - p(x_2)$, …, $d_n = y_n - p(x_n)$. The label 'best' is given to the polynomial $p(x)$ whose coefficients minimize the function $L$, where

$$L = d_1^2 + d_2^2 + ... + d_n^2 = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} [y_i - p(x_i)]^2 = \min \qquad (3.5)$$

### 3.4.1 Polynomials Least Squares Fitting

Polynomials are one of the most commonly used types of curves in regression. The applications of the method of least squares curve fitting using polynomials are briefly discussed as follows:

**The Least Squares Line**

The least squares line method uses a straight line $y = a + bx$ to approximate a given set of data, $(x_1, y_1)$, $(x_2, y_2)$, …, $(x_n, y_n)$, where $n \geq 2$.

**The Least Squares Parabola**

The least squares parabola methods uses a second degree curve $y = a + bx + cx^2$ to approximate a given set of data, $(x_1, y_1)$, $(x_2, y_2)$, …, $(x_n, y_n)$, where $n \geq 3$.

**The Least Squares $m^{th}$ degree Polynomials**

The least squares $m^{th}$ degree polynomials method uses $m^{th}$ degree polynomials $y = a_0 + a_1x + a_2x^2 + ... a_mx^m$ to approximate a given set of data, $(x_1, y_1)$, $(x_2, y_2)$, …, $(x_n, y_n)$, where $n \geq m+1$.

**Multiple Regression Least Squares**

Multiple regression estimates the outcomes which may be affected by more than one control parameters or there may be more than one control parameter being changed at the same time, e.g., $y = a + b_1x_1 + b_2x_2$.

In the next section, linear least squares and multiple regression least squares are discussed in more detail.

## 3.5 The Least-squares Line

The method of least squares can be applied to a special case where *p(x)* is a linear polynomial. In the least squares line, it involves one dependent variable, *Y* and one independent variable, *X*.

The least squares line uses a straight line

$$y = a + bx + e \tag{3.6}$$

where

$a = y$ intercept of the line,

$b$ = slope of the line, and

$e$ = error term

to approximate the given set of data, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, where $n \geq 2$.

**Theorem 3.1**

Given $n$ points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, the straight line $y = a + bx$ minimizing

$$L = \sum_{i=1}^{n}[y_i - p(x_i)]^2 = \sum_{i=1}^{n}[y_i - (a + bx_i)]^2$$

has slope

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}$$

and $y$-intercept

$$a = \frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n} = \overline{y} - b\overline{x}$$

Note that $a$ and $b$ are unknown coefficients while all $x_i$ and $y_i$ are given. To obtain the least squares error, the unknown coefficients $a$ and $b$ must yield zero first derivatives.

**Proof:**

The proof is accomplished by the usual device of taking the partial derivatives of $L$ with respect to $a$ and $b$, setting the resulting expressions equal to zero, and solving. By the first step we get

$$\frac{\partial L}{\partial a} = (-2)\sum_{i=1}^{n}[y_i - (a + bx_i)] = 0$$

and

$$\frac{\partial L}{\partial b} = (-2)\sum_{i=1}^{n} x_i[y_i - (a + bx_i)] = 0$$

Expanding the above equation, we have:

$$\sum_{i=1}^{n} y_i = a\sum_{i=1}^{n} 1 + b\sum_{i=1}^{n} x_i \qquad \text{or}$$

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i \tag{3.7}$$

and

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \tag{3.8}$$

From (3.7);

$$a = \frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n} \tag{3.9}$$

Then, substitute (3.9) into (3.8). We will get

$$\sum_{i=1}^{n} x_i y_i = (\frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n})\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2$$

$$n\sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i - b(\sum_{i=1}^{n} x_i)^2 + nb\sum_{i=1}^{n} x_i^2$$

$$b[(\sum_{i=1}^{n} x_i)^2 - n\sum_{i=1}^{n} x_i^2] = \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i - n\sum_{i=1}^{n} x_i y_i$$

$$b = \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i - n\sum_{i=1}^{n} x_i y_i}{(\sum_{i=1}^{n} x_i)^2 - n\sum_{i=1}^{n} x_i^2} \qquad \text{or}$$

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2} \tag{3.10}$$

(3.9) and (3.10) gives the solution for $a$ and $b$ which are stated in Theorem 3.1.

A line that fits the data well makes the residuals small. Requiring that the sum of residuals, $\sum_{i=1}^{n} e_i$, be small is futile, since large negative residuals can offset large positive ones. Indeed, any line through the point $(\bar{x}, \bar{y})$ has $\sum_{i=1}^{n} e_i = 0$.

## 3.6 Residuals of Least-squares Line

Residuals are also called "goodness of fit". The difference between an observed or dependent variable $y_i$ and the value of the least-squares line when $x = x_i$ is called the $i$th residual. In other words, residual is the difference between an actual value ($Y_i$) in the sample and the fitted value ($\hat{Y}_i$). With the sample data for $Y$ and $X$, we can obtain $a$ and $b$. With these estimates we can obtain fitted values for $Y$ using the sample data. Its magnitude reflects the failure of the least-squares line to 'model' that particular point.

**Definition 3.2**

Let $a$ and $b$ be the least-squares coefficients associated with the sample $(x_1, y_1)$, $(x_2, y_2)$, ..., $(x_n, y_n)$. For any value of $x$, the quantity $\hat{y} = a + bx$ is known as the predicted value of $y$. For each $i$, $i = 1, 2, \ldots, n$, the difference $y_i - \hat{y}_i = y_i - (a + bx_i)$ is called a residual.

A residual plot is a graph of the $i$th residual versus $x_i$, for all $i$. Applied statisticians find residual plots to be very helpful in assessing the appropriateness of fitting a straight line through a set of points.

**Theorem 3.2**

The sum of the residuals equals zero. Using the definition for the simple linear model applying the least squares method

$$\begin{aligned}(y_i - \hat{y}_i) &= Y_i - (a + bX_i)\\ &= Y_i - (\overline{Y} - b\overline{X} + bX_i)\end{aligned}$$

then

$$\begin{aligned}\Sigma(y_i - \hat{y}_i) &= \Sigma Y_i - \Sigma\overline{Y} + b\sum\overline{X} - b\sum X_i\\ &= \Sigma Y_i - \mathrm{n}\overline{Y} + bn\overline{X} - b\Sigma X_i\\ &= \Sigma Y_i - \mathrm{n}\frac{\sum Y_i}{n} + bn\frac{\sum X_i}{n} - b\Sigma X_i\\ &= \Sigma Y_i - \Sigma Y_i + b\Sigma X_i - b\Sigma X_i\\ &= 0\end{aligned}$$

## 3.7 Linear Multiple Regression Least-squares

Linear multiple regression predicts the outcome (dependent variables) which may be affected by more than one control parameter (independent variables) or there may be more than one control parameter being changed at the same time.

In this section, only the multiple regression least-squares with two predictor variables will be discussed. The model for one dependent variable, $Y$, and two independent variables $X_1$ and $X_2$, is

$$y = a + b_1x_1 + b_2x_2 + e \tag{3.11}$$

for a given data set $(y_1, x_{11}, x_{21})$, $(y_2, x_{12}, x_{22})$, ... , $(y_n, x_{1n}, x_{2n})$, where $n \geq 3$. The best fitting curve *P(x)* has the least squares error

$$L = \sum_{i=1}^{n}[y_i - P(x_{1i}, x_{2i})]^2 = \sum_{i=1}^{n}[y_i - (a + b_1x_{1i} + b_2x_{2i})]^2 = \min \tag{3.12}$$

Note that $a$, $b_1$, and $b_2$ are unknown coefficients while $x_{1i}$, $x_{2i}$, and $y_i$ are given. To obtain the least squares error, the unknown coefficients $a$, $b_1$, and $b_2$ must yield zero first derivatives. That is

$$\frac{\partial L}{\partial a} = (-2)\sum_{i=1}^{n}[y_i - (a + b_1x_{1i} + b_2x_{2i})] = 0 \ ,$$

$$\frac{\partial L}{\partial b_1} = (-2)\sum_{i=1}^{n}x_{1i}[y_i - (a + b_1x_{1i} + b_2x_{2i})] = 0 \ ,$$

and

$$\frac{\partial L}{\partial b_2} = (-2)\sum_{i=1}^{n}x_{2i}[y_i - (a + b_1x_{1i} + b_2x_{2i})] = 0 \ .$$

Expanding the above equations, we have

$$\sum_{i=1}^{n}y_i = a\sum_{i=1}^{n}1 + b_1\sum_{i=1}^{n}x_{1i} + b_2\sum_{i=1}^{n}x_{2i} \ , \tag{3.13}$$

$$\sum_{i=1}^{n}x_{1i}y_i = a\sum_{i=1}^{n}x_{1i} + b_1\sum_{i=1}^{n}x_{1i}^2 + b_2\sum_{i=1}^{n}x_{1i}x_{2i} \ , \tag{3.14}$$

and

$$\sum_{i=1}^{n}y_ix_{2i} = a\sum_{i=1}^{n}x_{2i} + b_1\sum_{i=1}^{n}x_{1i}x_{2i} + b_2\sum_{i=1}^{n}x_{2i}^2 \ . \tag{3.15}$$

The unknown coefficients $a$, $b_1$, and $b_2$ can hence be obtained by solving the above linear equations simultaneously. This is a system of three linear equations in three unknowns, so it usually provides a unique solution for the least-squares regression coefficient, $a$, $b_1$ and $b_2$. These value $\hat{a}$, $\hat{b}_1$ and $\hat{b}_2$ are called the least squares estimates of the coefficients.

The formula for $b_1$ estimates the effect of $X_1$ on $Y$, holding $X_2$ constant. Similarly, the formula for $b_2$ estimates the effect of $X_2$ on $Y$, holding $X_1$ constant. Finally, $a$ is the intercept, the estimated value for the criterion variable when both $X_1$ and $X_2$ are zero.

One of the differences between fitting a straight-line regression and a multiple regression is the computational difficulty. One needs to solve (*i*+1) linear equations simultaneously and this will be vary cumbersome working with a calculator.

The slope coefficients for the explanatory variables in the multiple regression are partial coefficients, while the slope coefficient in simple regression gives the marginal relationship between the response variable and a single explanatory variable. That is, each slope in multiple regression represents the 'effect' on the response variable of a one-unit increment in the corresponding explanatory variable holding the value of the other explanatory variable. The simple-regression slope effectively ignores the other explanatory variable (Fox, 2004).

## 3.8 Residuals of Multiple Regression Least-Squares

Residuals for multiple regression least-squares are actually the same as those for least-squares line. That is,

$$y_i = a + b_1 x_{1i} + b_2 x_{2i} + e_i$$

or

$$\hat{y}_i = a + b_1 x_{1i} + b_2 x_{2i}$$

where $\hat{y}$ is the predicted value of $y$.

Then, the difference

$$e_i = y_i - \hat{y}_i = y_i - (a + b_1 x_i + b_2 z_i) \tag{3.16}$$

is called a residual.

## 3.9 Summary

In this chapter, the regression analysis for simple and multiple have been presented. Then, the method of least squares method for simple and multiple regression were also discussed. Finally, residuals were explained.

# CHAPTER 4

# DATA ANALYSIS

## 4.0 Introduction

In this chapter, the least squares method and the goal programming method will be used to analyze the same data sets so that conclusions concerning the relationship of these two methods can be made. This chapter will begin with the description of the data.

## 4.1 Background of Data

Three data sets were chosen for analysis. All of the data sets contained outliers. Set 1 relates one dependent variable ($Y$) with one independent variable ($X$), set 2 relates one dependent variable ($Y$) with two independent variable ($X_1$, $X_2$) while set 3 relates one dependent variable ($Y$) with three independent variable ($X_1$, $X_2$, $X_3$). The data set are as follow:

### Data Set 1

Carbon aerosols have been identified as a contributing factor in a number of air quality problems. In a chemical analysis of diesel engine exhaust, $X$ = mass ($\mu g/cm^2$) and $Y$ = elemental carbon ($\mu g/cm^2$) were recorded ("Comparison of Solvent

Extraction and Thermal Optical Carbon Analysis Methods: Application to Diesel Vehicle Exhaust Aerosol" Environment Science Technology (1984): 231 – 234).

**Table 4.1 : Data Set I**

| **Observation number** | ***X*, mass** | ***Y*, elemental carbon** | **Observation number** | ***X*, mass** | ***Y*, elemental carbon** |
|---|---|---|---|---|---|
| 1 | 164.2 | 181 | 16 | 78.9 | 86 |
| 2 | 156.9 | 156 | 17 | 387.8 | 310 |
| 3 | 109.8 | 115 | 18 | 135.0 | 141 |
| 4 | 111.4 | 132 | 19 | 82.9 | 90 |
| 5 | 87.0 | 96 | 20 | 117.9 | 130 |
| 6 | 161.8 | 170 | 21 | 108.1 | 102 |
| 7 | 230.9 | 193 | 22 | 89.4 | 91 |
| 8 | 106.5 | 110 | 23 | 76.4 | 97 |
| 9 | 97.6 | 94 | 24 | 131.7 | 128 |
| 10 | 79.7 | 77 | 25 | 100.8 | 88 |
| 11 | 118.7 | 106 | | | |
| 12 | 248.8 | 204 | | | |
| 13 | 102.4 | 98 | | | |
| 14 | 64.2 | 76 | | | |
| 15 | 89.4 | 89 | | | |

Carbon aerosol is dangerous to our health because it influences the number of air quality. This set of data has 25 pairs, $(x_i, y_i)$ of observations as tabulated in Table 4.1. In this set, mass ($X$) is an independent variable while elemental carbon ($Y$) is a dependent variable. So, this is a simple linear regression problem.

## Data Set 2

The administrator for an organization that conducts management seminar programs is interested in examining the relationship between seminar enrollments ($Y$), the number of mailings ($X_1$), and the lead time of mailings ($X_2$) of seminar announcements. Data were obtained from a sample of $n$ = 25 management seminars offered by the organization and are listed in Table 4.2.

**Table 4.2 : Data Set II**

| **Observation number** | **Enrollment, $Y$** | **Num. of Mailings, $X_1$ (× 1,000)** | **Lead Time, $X_2$ (weeks)** | **Observation number** | **Enrollment, $Y$** | **Num. of Mailings, $X_1$ (× 1,000)** | **Lead Time, $X_2$ (weeks)** |
|---|---|---|---|---|---|---|---|
| 1 | 27 | 6.5 | 3 | 16 | 19 | 3.7 | 6 |
| 2 | 29 | 6.5 | 2 | 17 | 36 | 9.1 | 12 |
| 3 | 41 | 13.0 | 15 | 18 | 43 | 23.0 | 13 |
| 4 | 36 | 8.1 | 13 | 19 | 40 | 23.5 | 10 |
| 5 | 22 | 4.0 | 6 | 20 | 38 | 9.0 | 9 |
| 6 | 40 | 11.5 | 13 | 21 | 40 | 7.0 | 12 |
| 7 | 52 | 18.0 | 17 | 22 | 42 | 12.5 | 16 |
| 8 | 39 | 10.0 | 12 | 23 | 21 | 5.0 | 6 |
| 9 | 27 | 7.1 | 4 | 24 | 29 | 6.8 | 12 |
| 10 | 28 | 6.5 | 10 | 25 | 35 | 7.2 | 14 |
| 11 | 24 | 7.0 | 5 | | | | |
| 12 | 29 | 7.3 | 11 | | | | |
| 13 | 33 | 7.5 | 12 | | | | |
| 14 | 35 | 7.5 | 12 | | | | |
| 15 | 27 | 4.9 | 9 | | | | |

The second set of data is about management seminar program. 25 pair ($y_i$, $x_{1i}$, $x_{2i}$) of observations were recorded. The relationship between enrollment and number of mailings and lead time is deterministic if the value of enrollment is completely determined, with no uncertainty, once values of the number of mailings and lead time have been specified.

**Data Set 3**

The U.S. Bureau of Mines produces data on the price of minerals. Table 4.3 shows the average prices per year for several minerals over a decade.

**Table 4.3 : Data Set III**

| **Observation number** | ***Y*, Gold (*$* per *oz*)** | ***$X_1$*, Copper (cent per *lb*)** | ***$X_2$*, Silver (*$* per *oz*)** | ***$X_3$*, Aluminium (cents per *lb*)** |
|---|---|---|---|---|
| 1 | 161.1 | 64.2 | 4.4 | 39.8 |
| 2 | 308.0 | 93.3 | 11.1 | 61.0 |
| 3 | 613.0 | 101.3 | 20.6 | 71.6 |
| 4 | 460.0 | 84.2 | 10.5 | 76.0 |
| 5 | 376.0 | 72.8 | 8.0 | 76.0 |
| 6 | 424.0 | 76.5 | 11.4 | 77.8 |
| 7 | 361.0 | 66.8 | 8.1 | 81.0 |
| 8 | 318.0 | 67.0 | 6.1 | 81.0 |
| 9 | 368.0 | 66.1 | 5.5 | 81.0 |
| 10 | 448.0 | 82.5 | 7.0 | 72.3 |
| 11 | 438.0 | 120.5 | 6.5 | 110.1 |
| 12 | 382.6 | 130.9 | 5.5 | 87.8 |

There are four variables (minerals) – gold, copper, silver and aluminium in this data set. Gold and silver are measured by *$* per *oz* while copper and aluminium are measured by cents per *lb*. The objective here is to predict the average price of gold. Here, gold is the dependent variable denoted by $Y$ while copper, silver and aluminium are independent variable denoted by $X_1$, $X_2$ and $X_3$.

## 4.2 Outliers

**Definition 4.1**

The outlier is an unusually small or large data value (Devore and Peck, 2001).

**Definition 4.2**

Outliers are data points that lie apart from the rest points, or are data points that are apart, or far, from the mainstream of the other data (Black, 2001).

**Definition 4.3**

Outliers are observations with a unique combination of characteristics identifiable as distinctly different from the other observations (Hair et al, 1998).

Outliers can be classified into four classes. The first class arises from a procedural error, such as a data entry error or a mistake in coding. These outliers should be identified in the data cleaning stage, but if overlooked, they should be eliminated or recorded as missing value. The second class of outlier is the observation that occurs as the result of an extraordinary event, which then is an explanation for the uniqueness of the observation. The researcher must decide whether the extraordinary event should be represented in the sample. If so, the outlier should be retained in the analysis; if not, it should be deleted. The third class of outlier comprises extraordinary observations for which the researcher has no explanation. Although these are the outliers most likely to be omitted, they may be retained if the researcher feels they represent a valid segment of the population. The fourth and final class of outlier contains observations that fall within the ordinary range of values on each of the variables but are unique in their combination of values across the variables (Hair et al, 1998).

In linear regression, an outlier is defined as an observation for which the studentized residual ($r_i$ or $r_i^*$) is large in magnitude compared to other observations in the data set. Observations are judged as outliers on the basis of how unsuccessful the fitted regression equation is in accommodating them (Chatterjee and Hadi, 1988).

Potential outliers are observations that have extremely large residuals. They do not fit in with the pattern of the remaining data points and are not at all typical of the rest of the data. As a result, outliers are given careful attention in regression analysis in order to determine the reasons for the large fluctuations between the observed and predicted responses (Richard & Robert, 1980).

Both the predictor and dependent variables will have their parts to play in deciding whether an observation is unusual. The predictor variables determine whether a point has high leverage. The value of the dependent variable, $Y$, for a given set of $X$ values will determine whether the point is an outlier.

As every data point has an influence on the regression model, outliers can exert an overly important influence on the model because of their distance from other points. Thus an examination of outliers is worth considering before a set of data is analyzed. In the next section, box plot which is a simple technique to identify outliers in a data set will be presented.

### 4.2.1 Box Plot (Box and Whisker Plots)

A box plot is a diagram that utilizes the upper and lower quartiles along with the median and the two most extreme values to depict a distribution graphically. It is one technique to detect an outlier in data set. This technique is used in many statistics and management texts book. There are two types of box plot: the skeletal and the modified box plot (Devore and Peck, 2001).

**Definition 4.4**

Lower quartile = median of the lower half of the sample
Upper quartile = median of the upper half of the sample
The interquartile range (*iqr*), a resistant measure of variability, is given by

$$iqr = \text{upper quartile} - \text{lower quartile}$$

**Definition 4.5**

An observation is an outlier if it is more than 1.5 *iqr* away from the closest end of the box (the closest quartile). An outlier is extreme if it is more than 3 *iqr* from the closest end of the box, and it is mild otherwise. A modified box plot represents mild outliers by shaded circles and extreme outliers by open circles. Whiskers extend on each end to the most extreme observations that are not outliers.

The box plot is determined from the following:

1. The median ($Q_2$).
2. The lower quartile ($Q_l$).
3. The upper quartile ($Q_3$).
4. The smallest value in the distribution.
5. The largest value in the distribution.

The box of the plot is determined by locating the median and the lower and upper quartiles on a continuum. A box is drawn around the median with the lower and upper quartiles ($Q_1$ and $Q_3$) as the box endpoints. These box endpoints ($Q_1$ and $Q_3$) are referred to as the hinges of the box.

At a distance of $1.5 \cdot iqr$ outward from the lower and upper quartiles are what are referred to as inner fences. A whisker, a line segment, is drawn from the lower hinge of the box outward to the smallest data value. A second whisker is drawn from the upper hinge of the box outward to the largest data value. The inner fences are established as follows.

$$Q_1 - 1.5 \cdot iqr$$
$$Q_3 + 1.5 \cdot iqr$$

If there are data beyond the inner fences, then outer fences can be constructed:

$$Q_1 - 3 \cdot iqr$$
$$Q_3 + 3 \cdot iqr$$

Figure 4.1 shows the features of a box plots.

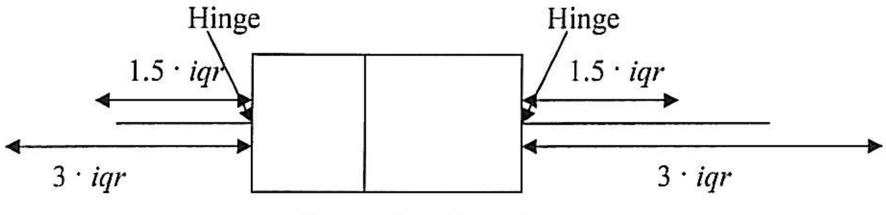

**Figure 4.1 : Box plots**

Data values that are outside the mainstream of values in a distribution are viewed as outliers. Outliers can be merely the more extreme values of a data set. Values in the data distribution that are outside the inner fences but within the outer fences are referred to as mild outliers. Values outside the outer fences are indicated by zero on the graph. These values are extreme outliers.

### 4.2.2 Existence of Outliers

Using the box plot technique, data set 1, 2 and 3 will be shown to contain outliers. It only tests the independent variable, *X*.

**Data Set 1**

First, the data need to be arranged from the smallest value to the largest value or vice versa as follows:

$x_i$ : 64.2, 76.4, 78.9, 79.7, 82.9, 87.0, 89.4, 89.4, 97.6, 100.8, 102.4, 106.5, 108.1, 109.8, 111.4, 117.9, 118.7, 131.7, 135.0, 156.9, 161.8, 164.2, 230.9, 248.8, 387.8

The quantities needed for constructing the modified box plot are as follows:

Median = 108.1, Lower quartile = 88.2, Upper quartile = 145.95

$iqr$ = upper quartile – lower quartile

$= 145.95 - 88.2 = 57.75$

$1.5 \cdot iqr = 1.5 \cdot 57.75 = 86.625$ and $3 \cdot iqr = 3 \cdot 57.75 = 173.25$

Thus,

Upper edge of box (upper quartile) + $1.5 \cdot iqr = 145.95 + 86.625 = 232.575$

Lower edge of box (lower quartile) – $1.5 \cdot iqr = 88.2 - 86.625 = 1.575$

So 248.8 and 387.8 are both outliers at the upper end, and there are no outliers at the lower end.

Since,

Upper edge of box + $3 \cdot iqr = 145.95 + 173.25 = 319.2$

387.8 is an extreme outlier and 248.8 is only a mild outlier.

The MINITAB box plot is presented in Figure 4.2.

```
       ---------
   ----I  +    I----                *                       O
       ---------

  +---------+---------+---------+---------+---------+------X
 60        120       180       240       300       360
```

**Figure 4.2 : Comparative Box Plot for Mass (µg/cm$^2$)**

**Data Set 2**

Here the data of independent variable, $X_1$ (number of mailings/×1000) and $X_2$ (lead time/weeks) would be tested.

For independent variable number of mailings,

$x_{1i}$ : 3.7, 4.0, 4.9, 5.0, 6.5, 6.5, 6.5, 6.8, 7.0, 7.0, 7.1, 7.2, 7.3, 7.5, 7.5, 8.1, 9.0, 9.1, 10.0, 11.5, 12.5, 13.0, 18.0, 23.0, 23.5

The quantities needed for constructing the modified box plot are as follows:

Median = 7.3, Lower quartile = 6.5, Upper quartile = 10.75, $iqr$ = 10.75 – 6.5 = 4.25

1.5 · $iqr$ = 1.5 · 4.25 = 6.375 and 3 · $iqr$ = 3 · 4.25 = 12.75

Thus, Upper edge of box + 1.5 · $iqr$ = 10.75 + 6.375 = 17.125

Lower edge of box – 1.5 · $iqr$ = 6.5 – 6.375 = 0.125

So, 18.0, 23.0 and 23.5 are both outliers at the upper end.

Since, Upper edge of box + 3 · $iqr$ = 10.75 + 12.75 = 23.5

There are no an extreme outlier for this data set.

The MINITAB box plot is presented in Figure 4.3.

```
          ----------
   -------I +      I--------           *           OO
          ----------
----+---------+---------+---------+---------+---------+--x1
  4.0       8.0      12.0      16.0      20.0      24.0
```

**Figure 4.3 : Comparative Box Plot for Number of Mailings (× 1000)**

For independent variable lead time of mailings,

$x_{2i}$ : 2, 3, 4, 5, 6, 6, 6, 9, 9, 10, 10, 11, 12, 12, 12, 12, 12, 12, 13, 13, 13, 14, 15, 16, 17

The quantities needed for constructing the modified box plot are as follows:

Median = 12, Lower quartile = 6, Upper quartile = 13, *iqr* = 13 – 6 = 7

1.5 · *iqr* = 1.5 · 7 = 10.5 and 3 · *iqr* = 3 · 7 = 21

Thus, Upper edge of box + 1.5 · *iqr* = 13 + 10.5 = 23.5

Lower edge of box – 1.5 · *iqr* = 6 – 10.5 = -4.5

So, there are not outliers for lead time of mailings, $X_2$.

**Data Set 3**

The data of independent variable, $X_1$ (copper/cent per *lb*), $X_2$ (silver/$ per *oz*) and $X_3$ (aluminium/cents per *lb*) would be tested.

For independent variable copper,

$x_{1i}$ : 66.1, 64.2, 66.8, 67.0, 72.8, 76.5, 82.5, 84.2, 93.3, 101.3, 120.5, 130.9

The quantities needed for constructing the modified box plot are as follows:

Median = 79.5, Lower quartile = 66.9, Upper quartile = 97.3

*iqr* = 97.3 – 66.9 = 30.4, 1.5 · *iqr* = 1.5 · 30.4 = 45.6 and 3 · *iqr* = 3 · 30.4 = 91.2

Thus, Upper edge of box + 1.5 · *iqr* = 97.3 + 45.6 = 142.9

Lower edge of box – 1.5 · *iqr* = 66.9 – 45.6 = 21.3

There are not outliers for independent $X_1$.

For independent variable silver,

$x_{2i}$ : 4.4, 5.5, 5.5, 6.1, 6.5, 7.0, 8.0, 8.1, 10.5, 11.1, 11.4, 20.6

The quantities needed for constructing the modified box plot are as follows:

Median = 7.5, Lower quartile = 5.8, Upper quartile = 10.8, *iqr* = 10.8 – 5.8 = 5.0

1.5 · *iqr* = 1.5 · 5.0 = 7.5 and 3 · *iqr* = 3 · 5.0 = 15.0

Thus, Upper edge of box + 1.5 · *iqr* = 10.8 + 7.5 = 18.3

Lower edge of box – 1.5 · *iqr* = 5.8 – 7.5 = -1.7

So, 20.6 is an outlier at the upper end.

Since, Upper edge of box + 3 · *iqr* = 10.8 + 15.0 = 25.8, 20.6 is a mild outlier.

The MINITAB box plot is presented in Figure 4.4.

```
         ------------------
     ----I      +         I--                          *
         ------------------
    ------+---------+---------+---------+---------+---------+x2
         6.0       9.0      12.0      15.0      18.0      21.0
```

**Figure 4.4 : Comparative Box Plot for Silver (*$* per *oz*)**

For independent variable aluminium,

$x_{3i}$ : 39.8, 61.0, 71.6, 72.3, 76.0, 76.0, 77.8, 81.0, 81.0, 81.0, 87.8, 110.1

The quantities needed for constructing the modified box plot are as follow:

Median = 76.9, Lower quartile = 71.95, Upper quartile = 81.0

*iqr* = 81.0 – 71.95 = 9.05, 1.5 · *iqr* = 1.5 · 9.05 = 13.575  and3 · *iqr* = 3 · 9.05 = 27.15

Thus, Upper edge of box + 1.5 · *iqr* = 81.0 + 13.575 = 94.575

Lower edge of box – 1.5 · *iqr* = 71.95 – 13.575 = 58.375

So, 110.1 is an outlier at the upper end and 39.8 is an outlier at the lower end. Since

Upper edge of box + 3 · *iqr* = 81.0 + 27.15 = 108.15

Lower edge of box – 3 · *iqr* = 71.95 – 27.15 = 44.8

39.8 and 110.1 is an extreme outlier.

The MINITAB box plot is presented in Figure 4.5.

```
                         -------
         O        -------I  +  I-----          O
                         -------
    --------+---------+---------+---------+---------+--------x3
           45        60        75        90       105
```

**Figure 4.5 : Comparative Box Plot for Aluminium (cents per *lb*)**

## 4.3 Analysis Using Least Squares Method

In this section, the three data sets will be analyzed using the least squares method to produce the best polynomial.

### 4.3.1 Analysis on Data Set 1

In this data set, only the first 20, 19 and 18 pairs ($y_i$, $x_i$) of observations will be used for analyzing the data in the following conditions: outliers included, mild outlier or extreme outlier removed, and both mild and extreme outliers removed. The last five observations will be used to predict or estimate using the least squares line equations obtained from this data set.

**(i) Outliers included**

Recall Table 4.1. It is calculated that

$$\sum_{i=1}^{20} x_i = 2731.8 \qquad \sum_{i=1}^{20} x_i^2 = 484531.16$$

$$\sum_{i=1}^{20} y_i = 2654 \qquad \sum_{i=1}^{20} x_i y_i = 444011.2$$

Using equations (3.9) and (3.10), we have

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}$$

$$= \frac{20(444011.2) - 2731.8(2654)}{20(484531.16) - (2731.8)^2}$$

$$= 0.7316$$

and

$$a = \frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n}$$

$$= \frac{2654 - 0.7316(2731.8)}{20}$$

$$= 32.7708$$

Then, the least-squares line is

$$y_i = a + bx_i + e_i$$

$$y_i = 32.7708 + 0.7316x_i + e_i \text{ or } \hat{y}_i = 32.7708 + 0.7316x_i \qquad (4.1)$$

The scatter plot for this best straight line is shown in Figure 4.6.

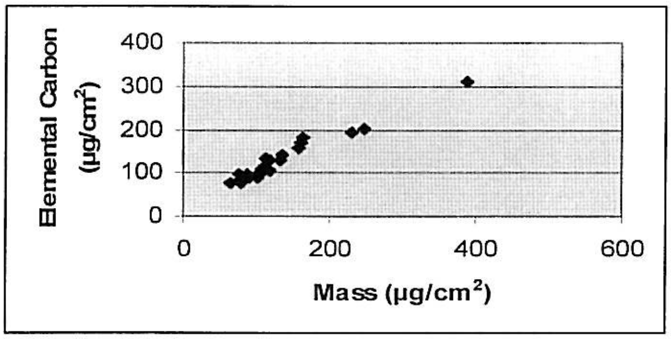

**Figure 4.6 : The Scatter Plot for Data Set I**

### (ii) Remove mild outlier

Here, the 12th observation (upper end mild outlier), $(y_{12}, x_{12}) = (204, 248.8)$, will be removed. Only the first 19 pairs of observations will be used for analysis.

Recall Table 4.1. It is calculated that

$$\sum_{i=1}^{19} x_i = 2483 \ , \ \sum_{i=1}^{19} x_i^2 = 422629.72 \ , \ \sum_{i=1}^{19} y_i = 2450 \, , \ \sum_{i=1}^{19} x_i y_i = 393256$$

Using equations (3.9) and (3.10), we have $b = 0.7446$ and $a = 31.6345$.

Then, the least-squares line is $\hat{y}_i = 31.6345 + 0.7446x_i$ (4.2)

### (iii) Remove extreme outlier

Now, the 17th observation (upper end extreme outlier), $(y_{17}, x_{17}) = (310, 387.8)$, will be removed.

Recall Table 4.1. It is calculated that

$$\sum_{i=1}^{19} x_i = 2344\,,\ \sum_{i=1}^{19} x_i^2 = 334142.32\,,\ \sum_{i=1}^{19} y_i = 2344\,,\ \sum_{i=1}^{19} x_i y_i = 323793.2$$

Using equations (3.9) and (3.10), we have $b = 0.7698$ and $a = 28.3933$.

Then, the least-squares line is $\hat{y}_i = 28.3933 + 0.7698x_i$ (4.3)

### (iv) Remove both mild and extreme outliers

Now, the 12th and 17th observation (mild and extreme outlier), will be removed.

Recall Table 4.1. It is calculated that

$$\sum_{i=1}^{18} x_i = 2095.2\,,\ \sum_{i=1}^{18} x_i^2 = 272240.88\,,\ \sum_{i=1}^{18} y_i = 2140\,,\ \sum_{i=1}^{18} x_i y_i = 273038$$

Using equations (3.9) and (3.10), we have $b = 0.8442$ and $a = 20.6206$.

Then, the least-squares line is $\hat{y}_i = 20.6206 + 0.8442x_i$ (4.4)

### 4.3.2 Analysis on Data Set 2

For this set of data only 20 triplets ($y_i$, $x_{1i}$, $x_{2i}$) of observations will be used for analysis. Then 19 and 17 triplets of observations will be used to analyze the data when the first, second or third mild outlier and all of the three mild outliers are removed from the data set using MINITAB software package.

### (i) Outliers included

Recall Table 4.2. It is calculated that

$$\sum_{i=1}^{20} x_{1i} = 193.7\ ,\ \sum_{i=1}^{20} x_{1i}^2 = 2481.57\,,\ \sum_{i=1}^{20} x_{2i} = 194\,,\ \sum_{i=1}^{20} x_{2i}^2 = 2206$$

$$\sum_{i=1}^{20} y_i = 665\,,\quad \sum_{i=1}^{20} x_{1i} y_i = 7127.2\,,\ \sum_{i=1}^{20} x_{1i} x_{2i} = 2111.5\,,\ \sum_{i=1}^{20} x_{2i} y_i = 6959$$

Using equations (3.13) to (3.15), we have

$$an + b_1\sum_{i=1}^{n} x_{1i} + b_2\sum_{i=1}^{n} x_{2i} = \sum_{i=1}^{n} y_i$$

$$20a + 193.7b_1 + 194b_2 = 665 \qquad (4.5)$$

$$a\sum_{i=1}^{n} x_{1i} + b_1\sum_{i=1}^{n} x_{1i}^2 + b_2\sum_{i=1}^{n} x_{1i}x_{2i} = \sum_{i=1}^{n} x_{1i}y_i$$

$$193.7a + 2481.57b_1 + 2111.5b_2 = 7127.2 \qquad (4.6)$$

$$a\sum_{i=1}^{n} x_{2i} + b_1\sum_{i=1}^{n} x_{1i}x_{2i} + b_2\sum_{i=1}^{n} x_{2i}^2 = \sum_{i=1}^{n} x_{2i}y_i$$

$$194a + 2111.5b_1 + 2206b_2 = 6959 \qquad (4.7)$$

Solving equations (4.5), (4.6) and (4.7) simultaneously, we get

$a = 22.2$, $b_1 = 1.1$ and $b_2 = 0.0167$.

Thus, the least-squares line is

$$y_i = a + b_1x_{1i} + b_2x_{2i} + e_i$$

$$y_i = 22.2 + 1.1x_{1i} + 0.0167x_{2i} + e_i \text{ or}$$

$$\hat{y}_i = 22.2 + 1.1x_{1i} + 0.0167x_{2i} \qquad (4.8)$$

**(ii) Remove first mild outlier**

The first mild outlier is $(y_7, x_{1,7}, x_{2,7}) = (52, 18.0, 17)$. Only 19 triplets of observations will be used for analysis. Using MINITAB software package, the least-squares line is $\hat{y}_i = 23.1 + 0.932x_{1i} + 0.02x_{2i}$ (4.9)

**(iii) Remove second mild outlier**

The second mild outlier is $(y_{18}, x_{1,18}, x_{2,18}) = (43, 23.0, 13)$. Only 19 triplets of observations will be used for analysis. Using MINITAB software package, the least-squares line is $\hat{y}_i = 21.0 + 1.27x_{1i} + 0.0134x_{2i}$ (4.10)

### (iv) Remove third mild outlier

The third mild outlier is $(y_{19}, x_{1,19}, x_{2,19}) = (40, 23.5, 10)$. Only 19 triplets of observations will be used for analysis. Using MINITAB software package, the least-squares line is $\hat{y}_i = 19.9 + 1.42x_{1i} + 0.0101x_{2i}$ (4.11)

### (v) Remove all the three mild outliers

Here, only 17 triplets of observations will be used for analysis. Using MINITAB software package, the least-squares line is

$$\hat{y}_i = 10.3 + 2.80x_{1i} - 0.0176x_{2i} \tag{4.12}$$

### 4.3.3 Analysis on Data Set 3

For this data set, only outliers in independent variable $X_3$, aluminium, would be analyzed. MINITAB, a statistical software package is used to perform all the multiple regression analysis. 10, 9 and 8 pairs $(y_i, x_{1i}, x_{2i}, x_{3i})$ of observations were used for analyzing the data in the following conditions: outliers included, first or second extreme outlier removed, and both extreme outliers removed.

### (i) Outliers included

From the MINITAB output,the least-squares line is

$$\hat{y}_i = -40.9 - 0.23x_{1i} + 18.6x_{2i} + 3.83x_{3i} \tag{4.13}$$

### (ii) Remove first extreme outlier

The first extreme outlier (lower end), $x_{3,1} = 39.8$, will be removed from data set 3. From the MINITAB output, the least-squares line is

$$\hat{y}_i = 30 - 0.13x_{1i} + 17.1x_{2i} + 3.03x_{3i} \tag{4.14}$$

### (iii) Remove second extreme outlier

The second extreme outlier (upper end), $x_{3,11} = 110.1$, will be removed from data set 3. From the MINITAB output, the least-squares line is

$$\hat{y}_i = -100 - 0.102x_{1i} + 17.7x_{2i} + 4.56x_{3i} \quad (4.15)$$

### (iv) Remove both extreme outlier

From the MINITAB output, the least-squares line is

$$\hat{y}_i = -125 - 0.03x_{1i} + 18.1x_{2i} + 4.86x_{3i} \quad (4.16)$$

## 4.4 Converting Least Squares Problem into a Goal Programming Model

From the least-squares method, we have

$$\hat{y}_i = a + bx_i \quad (4.17)$$

and the error/residual is $y_i - \hat{y}_i = e_i$ or

$$\hat{y}_i = y_i - e_i \quad (4.18)$$

(4.17) = (4.18);

$$y_i - e_i = a + bx_i$$

$$y_i = a + bx_i + e_i \quad (4.19)$$

From the goal programming model, we have

$$\text{Minimize} \qquad Z = \sum_{i=1}^{n} (d_i^+ + d_i^-)$$

Subject to the linear constraints:

$$\left.\begin{array}{ll}\text{Goal constraints:} & (\sum_{j=1}^{n} a_{ij}x_j) + d_i^- - d_i^+ = b_i\,,\ i = 1, 2, \ldots, m \\ \text{System constraints:} & \sum_{j=1}^{n} a_{ij}x_j \begin{bmatrix} \leq \\ = \\ \geq \end{bmatrix} b_i\,,\ i = m+1, \ldots, m+p \\ \text{with} & x_j, d_i^-, d_i^+ \geq 0, \text{for } i = 1, 2, \ldots, m \text{ and } j = 1, 2, \ldots, n \end{array}\right\} \quad (4.20)$$

From (4.19) and (4.20),

$$e_i = d_i^- - d_i^+ \quad (4.21)$$

## 4.4 Analysis Using Goal Programming

In this section, the data analysis were performed using goal programming. The solutions were obtained using QM for Windows.

QM for Windows is a package for quantitative methods, management science or operational research. This package is a user friendly software package available in its fields. The software can be used to either solve problems or check answers (http://www.prenhall.com).

### 4.4.1 Analysis on Data Set 1

**(i) Outliers included**

The formulation of data set 1 is presented as follows:

$$\begin{array}{ll}\text{Minimize} & Z = P_1 \sum_{i=1}^{20} (d_i^+ + d_i^-) \\ \text{Subject to} & 164.2x_1 + x_2 + d_1^- - d_1^+ = 181 \\ & 156.9x_1 + x_2 + d_2^- - d_2^+ = 156 \end{array}$$

$$109.8x_1 + x_2 + d_3^- - d_3^+ = 115$$

$$111.4x_1 + x_2 + d_4^- - d_4^+ = 132$$

$$87x_1 + x_2 + d_5^- - d_5^+ = 96$$

$$161.8x_1 + x_2 + d_6^- - d_6^+ = 170$$

$$230.9x_1 + x_2 + d_7^- - d_7^+ = 193$$

$$106.5x_1 + x_2 + d_8^- - d_8^+ = 110$$

$$97.6x_1 + x_2 + d_9^- - d_9^+ = 94$$

$$79.7x_1 + x_2 + d_{10}^- - d_{10}^+ = 77$$

$$118.7x_1 + x_2 + d_{11}^- - d_{11}^+ = 106$$

$$248.8x_1 + x_2 + d_{12}^- - d_{12}^+ = 204$$

$$102.4x_1 + x_2 + d_{13}^- - d_{13}^+ = 98$$

$$64.2x_1 + x_2 + d_{14}^- - d_{14}^+ = 76$$

$$89.4x_1 + x_2 + d_{15}^- - d_{15}^+ = 89$$

$$78.9x_1 + x_2 + d_{16}^- - d_{16}^+ = 86$$

$$387.8x_1 + x_2 + d_{17}^- - d_{17}^+ = 310$$

$$135x_1 + x_2 + d_{18}^- - d_{18}^+ = 141$$

$$82.9x_1 + x_2 + d_{19}^- - d_{19}^+ = 90$$

$$117.9x_1 + x_2 + d_{20}^- - d_{20}^+ = 130$$

with $x_i, d_i^-, d_i^+ \geq 0, i = 1,2,...,20$

Here, we only have one goal that is $P_1$ for data set 1. The goal, $P_1$ is to predict elemental carbon. This goal programming model is solved by using QM for Windows software. The results are $x_1 = 0.7215$ and $x_2 = 30.1839$.

From these results, we can write the predicted equation as

$$\hat{y}_i = 0.7215x_i + 30.1839 \qquad (4.21)$$

**(ii) Remove mild outlier**

The formulation of data set 1 without mild outlier is presented as follows:

Minimize $Z = P_1 \sum_{i=1}^{20} (d_i^+ + d_i^-)$

Subject to $164.2x_1 + x_2 + d_1^- - d_1^+ = 181$

$$156.9x_1 + x_2 + d_2^- - d_2^+ = 156$$

$$109.8x_1 + x_2 + d_3^- - d_3^+ = 115$$

$$111.4x_1 + x_2 + d_4^- - d_4^+ = 132$$

$$87x_1 + x_2 + d_5^- - d_5^+ = 96$$

$$161.8x_1 + x_2 + d_6^- - d_6^+ = 170$$

$$230.9x_1 + x_2 + d_7^- - d_7^+ = 193$$

$$106.5x_1 + x_2 + d_8^- - d_8^+ = 110$$

$$97.6x_1 + x_2 + d_9^- - d_9^+ = 94$$

$$79.7x_1 + x_2 + d_{10}^- - d_{10}^+ = 77$$

$$118.7x_1 + x_2 + d_{11}^- - d_{11}^+ = 106$$

$$102.4x_1 + x_2 + d_{13}^- - d_{13}^+ = 98$$

$$64.2x_1 + x_2 + d_{14}^- - d_{14}^+ = 76$$

$$89.4x_1 + x_2 + d_{15}^- - d_{15}^+ = 89$$

$$78.9x_1 + x_2 + d_{16}^- - d_{16}^+ = 86$$

$$387.8x_1 + x_2 + d_{17}^- - d_{17}^+ = 310$$

$$135x_1 + x_2 + d_{18}^- - d_{18}^+ = 141$$

$$82.9x_1 + x_2 + d_{19}^- - d_{19}^+ = 90$$

$$117.9x_1 + x_2 + d_{20}^- - d_{20}^+ = 130$$

with $x_i, d_i^-, d_i^+ \geq 0, i = 1,2,\ldots,20$

This goal programming model is solved by using QM for Windows. The results are $x_1 = 0.7215$ and $x_2 = 30.1837$

Thus, the predicted equation is $\hat{y}_i = 0.7215x_i + 30.1837$. (4.22)

### (iii) Remove extreme outlier

The formulation of data set 1 without extreme outlier is presented as follows:

$$
\begin{array}{ll}
\text{Minimize} & Z = P_1 \sum_{i=1}^{20} (d_i^+ + d_i^-) \\
\text{Subject to} & 164.2x_1 + x_2 + d_1^- - d_1^+ = 181 \\
 & 156.9x_1 + x_2 + d_2^- - d_2^+ = 156 \\
 & 109.8x_1 + x_2 + d_3^- - d_3^+ = 115 \\
 & 111.4x_1 + x_2 + d_4^- - d_4^+ = 132 \\
 & 87x_1 + x_2 + d_5^- - d_5^+ = 96 \\
 & 161.8x_1 + x_2 + d_6^- - d_6^+ = 170 \\
 & 230.9x_1 + x_2 + d_7^- - d_7^+ = 193 \\
 & 106.5x_1 + x_2 + d_8^- - d_8^+ = 110 \\
 & 97.6x_1 + x_2 + d_9^- - d_9^+ = 94 \\
 & 79.7x_1 + x_2 + d_{10}^- - d_{10}^+ = 77 \\
 & 118.7x_1 + x_2 + d_{11}^- - d_{11}^+ = 106 \\
 & 248.8x_1 + x_2 + d_{12}^- - d_{12}^+ = 204 \\
 & 102.4x_1 + x_2 + d_{13}^- - d_{13}^+ = 98 \\
 & 64.2x_1 + x_2 + d_{14}^- - d_{14}^+ = 76 \\
 & 89.4x_1 + x_2 + d_{15}^- - d_{15}^+ = 89 \\
 & 78.9x_1 + x_2 + d_{16}^- - d_{16}^+ = 86 \\
 & 135x_1 + x_2 + d_{18}^- - d_{18}^+ = 141 \\
 & 82.9x_1 + x_2 + d_{19}^- - d_{19}^+ = 90 \\
 & 117.9x_1 + x_2 + d_{20}^- - d_{20}^+ = 130 \\
\text{with} & x_i, d_i^-, d_i^+ \geq 0, i = 1,2,\ldots,20
\end{array}
$$

The results are $x_1 = 0.8038$ and $x_2 = 24.3972$.

Thus, the predicted equation is $\hat{y}_i = 0.8038x_i + 24.3972$. (4.23)

### (iv) Remove both mild and extreme outliers

The formulation of data set 1 without both mild and extreme outliers is presented as follows:

Minimize $Z = P_1 \sum_{i=1}^{20} (d_i^+ + d_i^-)$

Subject to

$$164.2x_1 + x_2 + d_1^- - d_1^+ = 181$$
$$156.9x_1 + x_2 + d_2^- - d_2^+ = 156$$
$$109.8x_1 + x_2 + d_3^- - d_3^+ = 115$$
$$111.4x_1 + x_2 + d_4^- - d_4^+ = 132$$
$$87x_1 + x_2 + d_5^- - d_5^+ = 96$$
$$161.8x_1 + x_2 + d_6^- - d_6^+ = 170$$
$$230.9x_1 + x_2 + d_7^- - d_7^+ = 193$$
$$106.5x_1 + x_2 + d_8^- - d_8^+ = 110$$
$$97.6x_1 + x_2 + d_9^- - d_9^+ = 94$$
$$79.7x_1 + x_2 + d_{10}^- - d_{10}^+ = 77$$
$$118.7x_1 + x_2 + d_{11}^- - d_{11}^+ = 106$$
$$102.4x_1 + x_2 + d_{13}^- - d_{13}^+ = 98$$
$$64.2x_1 + x_2 + d_{14}^- - d_{14}^+ = 76$$
$$89.4x_1 + x_2 + d_{15}^- - d_{15}^+ = 89$$
$$78.9x_1 + x_2 + d_{16}^- - d_{16}^+ = 86$$
$$135x_1 + x_2 + d_{18}^- - d_{18}^+ = 141$$
$$82.9x_1 + x_2 + d_{19}^- - d_{19}^+ = 90$$
$$117.9x_1 + x_2 + d_{20}^- - d_{20}^+ = 130$$

with $x_i, d_i^-, d_i^+ \geq 0, i = 1,2,...,20$

The results are shown $x_1 = 0.8919$ and $x_2 = 16.0622$.

Thus, the predicted equation is $\hat{y}_i = 0.8919x_i + 16.0622$. (4.24)

### 4.4.2 Analysis on Data Set 2

**(i) Outliers included**

The complete model of data set 2 can be presented as follows:

Minimize $Z = P_1 \sum_{i=1}^{20} (d_i^+ + d_i^-)$

Subject to

$$6.5x_1 + 3x_2 + x_3 + d_1^- - d_1^+ = 27$$

$$6.5x_1 + 2x_2 + x_3 + d_2^- - d_2^+ = 29$$

$$13.0x_1 + 15x_2 + x_3 + d_3^- - d_3^+ = 41$$

$$8.1x_1 + 13x_2 + x_3 + d_4^- - d_4^+ = 36$$

$$4.0x_1 + 6x_2 + x_3 + d_5^- - d_5^+ = 22$$

$$11.5x_1 + 13x_2 + x_3 + d_6^- - d_6^+ = 40$$

$$18.0x_1 + 17x_2 + x_3 + d_7^- - d_7^+ = 52$$

$$10.0x_1 + 12x_2 + x_3 + d_8^- - d_8^+ = 39$$

$$7.1x_1 + 4x_2 + x_3 + d_9^- - d_9^+ = 27$$

$$6.5x_1 + 10x_2 + x_3 + d_{10}^- - d_{10}^+ = 28$$

$$7.0x_1 + 5x_2 + x_3 + d_{11}^- - d_{11}^+ = 24$$

$$7.3x_1 + 11x_2 + x_3 + d_{12}^- - d_{12}^+ = 29$$

$$7.5x_1 + 12x_2 + x_3 + d_{13}^- - d_{13}^+ = 33$$

$$7.5x_1 + 12x_2 + x_3 + d_{14}^- - d_{14}^+ = 35$$

$$4.9x_1 + 9x_2 + x_3 + d_{15}^- - d_{15}^+ = 27$$

$$3.7x_1 + 6x_2 + x_3 + d_{16}^- - d_{16}^+ = 19$$

$$9.1x_1 + 12x_2 + x_3 + d_{17}^- - d_{17}^+ = 36$$

$$23.0x_1 + 13x_2 + x_3 + d_{18}^- - d_{18}^+ = 43$$

$$23.5x_1 + 10x_2 + x_3 + d_{19}^- - d_{19}^+ = 40$$

$$9.0x_1 + 9x_2 + x_3 + d_{20}^- - d_{20}^+ = 38$$

with $x_i, d_i^-, d_i^+ \geq 0, i = 1, 2, \ldots, 20$

Data set 2 have only one goal, P$_1$. This goal is to predict seminar enrollment. This goal programming model is solved by using QM for Windows software. The results are $x_1$ = 0.5055, $x_2$ = 1.2615 and $x_3$ = 15.5055.

Thus, the predicted equation is $\hat{y}_i = 0.5055x_{1i} + 1.2615x_{2i} + 15.5055$. (4.25)

**(ii) Remove first mild outlier**

The complete model of data set 2 without first mild outlier can be presented as follows:

$$\text{Minimize} \quad Z = P_1 \sum_{i=1}^{20} (d_i^+ + d_i^-)$$

$$\text{Subject to} \quad 6.5x_1 + 3x_2 + x_3 + d_1^- - d_1^+ = 27$$

$$6.5x_1 + 2x_2 + x_3 + d_2^- - d_2^+ = 29$$

$$13.0x_1 + 15x_2 + x_3 + d_3^- - d_3^+ = 41$$

$$8.1x_1 + 13x_2 + x_3 + d_4^- - d_4^+ = 36$$

$$4.0x_1 + 6x_2 + x_3 + d_5^- - d_5^+ = 22$$

$$11.5x_1 + 13x_2 + x_3 + d_6^- - d_6^+ = 40$$

$$10.0x_1 + 12x_2 + x_3 + d_8^- - d_8^+ = 39$$

$$7.1x_1 + 4x_2 + x_3 + d_9^- - d_9^+ = 27$$

$$6.5x_1 + 10x_2 + x_3 + d_{10}^- - d_{10}^+ = 28$$

$$7.0x_1 + 5x_2 + x_3 + d_{11}^- - d_{11}^+ = 24$$

$$7.3x_1 + 11x_2 + x_3 + d_{12}^- - d_{12}^+ = 29$$

$$7.5x_1 + 12x_2 + x_3 + d_{13}^- - d_{13}^+ = 33$$

$$7.5x_1 + 12x_2 + x_3 + d_{14}^- - d_{14}^+ = 35$$

$$4.9x_1 + 9x_2 + x_3 + d_{15}^- - d_{15}^+ = 27$$

$$3.7x_1 + 6x_2 + x_3 + d_{16}^- - d_{16}^+ = 19$$

$$9.1x_1 + 12x_2 + x_3 + d_{17}^- - d_{17}^+ = 36$$

$$23.0x_1 + 13x_2 + x_3 + d_{18}^- - d_{18}^+ = 43$$

$$23.5x_1 + 10x_2 + x_3 + d_{19}^- - d_{19}^+ = 40$$

$$9.0x_1 + 9x_2 + x_3 + d_{20}^- - d_{20}^+ = 38$$

with $x_i, d_i^-, d_i^+ \geq 0, i = 1, 2, \ldots, 20$

The results are $x_1 = 0.4698$, $x_2 = 1.0783$ and $x_3 = 18.1767$.

Thus, the predicted equation is $\hat{y}_i = 0.4698x_{1i} + 1.0783x_{2i} + 18.1767$. (4.26)

### (iii) Remove second mild outlier

The complete model of data set 2 without second mild outlier can be presented as follows:

Minimize $Z = P_1 \sum_{i=1}^{20} (d_i^+ + d_i^-)$

Subject to

$$6.5x_1 + 3x_2 + x_3 + d_1^- - d_1^+ = 27$$

$$6.5x_1 + 2x_2 + x_3 + d_2^- - d_2^+ = 29$$

$$13.0x_1 + 15x_2 + x_3 + d_3^- - d_3^+ = 41$$

$$8.1x_1 + 13x_2 + x_3 + d_4^- - d_4^+ = 36$$

$$4.0x_1 + 6x_2 + x_3 + d_5^- - d_5^+ = 22$$

$$11.5x_1 + 13x_2 + x_3 + d_6^- - d_6^+ = 40$$

$$18.0x_1 + 17x_2 + x_3 + d_7^- - d_7^+ = 52$$

$$10.0x_1 + 12x_2 + x_3 + d_8^- - d_8^+ = 39$$

$$7.1x_1 + 4x_2 + x_3 + d_9^- - d_9^+ = 27$$

$$6.5x_1 + 10x_2 + x_3 + d_{10}^- - d_{10}^+ = 28$$

$$7.0x_1 + 5x_2 + x_3 + d_{11}^- - d_{11}^+ = 24$$

$$7.3x_1 + 11x_2 + x_3 + d_{12}^- - d_{12}^+ = 29$$

$$7.5x_1 + 12x_2 + x_3 + d_{13}^- - d_{13}^+ = 33$$

$$7.5x_1 + 12x_2 + x_3 + d_{14}^- - d_{14}^+ = 35$$

$$4.9x_1 + 9x_2 + x_3 + d_{15}^- - d_{15}^+ = 27$$

$$3.7x_1 + 6x_2 + x_3 + d_{16}^- - d_{16}^+ = 19$$

$$9.1x_1 + 12x_2 + x_3 + d_{17}^- - d_{17}^+ = 36$$

$$23.5x_1 + 10x_2 + x_3 + d_{19}^- - d_{19}^+ = 40$$

$$9.0x_1 + 9x_2 + x_3 + d_{20}^- - d_{20}^+ = 38$$

with $x_i, d_i^-, d_i^+ \geq 0, i = 1, 2, \ldots, 20$

The results are $x_1 = 1.4878$, $x_2 = 0.6756$ and $x_3 = 13.7342$.

Thus, the predicted equation is $\hat{y}_i = 1.4878x_{1i} + 0.6756x_{2i} + 13.7342$. (4.27)

**(iv) Remove third mild outlier**

The complete model of data set 2 without third mild outlier can be presented as follows:

Minimize $Z = P_1 \sum_{i=1}^{20} (d_i^+ + d_i^-)$

Subject to

$$6.5x_1 + 3x_2 + x_3 + d_1^- - d_1^+ = 27$$

$$6.5x_1 + 2x_2 + x_3 + d_2^- - d_2^+ = 29$$

$$13.0x_1 + 15x_2 + x_3 + d_3^- - d_3^+ = 41$$

$$8.1x_1 + 13x_2 + x_3 + d_4^- - d_4^+ = 36$$

$$4.0x_1 + 6x_2 + x_3 + d_5^- - d_5^+ = 22$$

$$11.5x_1 + 13x_2 + x_3 + d_6^- - d_6^+ = 40$$

$$18.0x_1 + 17x_2 + x_3 + d_7^- - d_7^+ = 52$$

$$10.0x_1 + 12x_2 + x_3 + d_8^- - d_8^+ = 39$$

$$7.1x_1 + 4x_2 + x_3 + d_9^- - d_9^+ = 27$$

$$6.5x_1 + 10x_2 + x_3 + d_{10}^- - d_{10}^+ = 28$$

$$7.0x_1 + 5x_2 + x_3 + d_{11}^- - d_{11}^+ = 24$$

$$7.3x_1 + 11x_2 + x_3 + d_{12}^- - d_{12}^+ = 29$$

$$7.5x_1 + 12x_2 + x_3 + d_{13}^- - d_{13}^+ = 33$$

$$7.5x_1 + 12x_2 + x_3 + d_{14}^- - d_{14}^+ = 35$$

$$4.9x_1 + 9x_2 + x_3 + d_{15}^- - d_{15}^+ = 27$$

$$3.7x_1 + 6x_2 + x_3 + d_{16}^- - d_{16}^+ = 19$$

$$9.1x_1 + 12x_2 + x_3 + d_{17}^- - d_{17}^+ = 36$$

$$23.0x_1 + 13x_2 + x_3 + d_{18}^- - d_{18}^+ = 43$$

$$9.0x_1 + 9x_2 + x_3 + d_{20}^- - d_{20}^+ = 38$$

with $x_i, d_i^-, d_i^+ \geq 0, i = 1, 2, \ldots, 20$

The results are $x_1 = 1.4878$, $x_2 = 0.6756$ and $x_3 = 13.7342$.

Thus, the predicted equation is $\hat{y}_i = 1.4878x_{1i} + 0.6756x_{2i} + 13.7342$ (4.28)

**(v) Remove all of the three mild outliers**

The complete model of data set 2 when all the mild outliers are removed can be presented as follows:

Minimize $Z = P_1 \sum_{i=1}^{20} (d_i^+ + d_i^-)$

Subject to $6.5x_1 + 3x_2 + x_3 + d_1^- - d_1^+ = 27$

$$6.5x_1 + 2x_2 + x_3 + d_2^- - d_2^+ = 29$$

$$13.0x_1 + 15x_2 + x_3 + d_3^- - d_3^+ = 41$$

$$8.1x_1 + 13x_2 + x_3 + d_4^- - d_4^+ = 36$$

$$4.0x_1 + 6x_2 + x_3 + d_5^- - d_5^+ = 22$$

$$11.5x_1 + 13x_2 + x_3 + d_6^- - d_6^+ = 40$$

$$10.0x_1 + 12x_2 + x_3 + d_8^- - d_8^+ = 39$$

$$7.1x_1 + 4x_2 + x_3 + d_9^- - d_9^+ = 27$$

$$6.5x_1 + 10x_2 + x_3 + d_{10}^- - d_{10}^+ = 28$$

$$7.0x_1 + 5x_2 + x_3 + d_{11}^- - d_{11}^+ = 24$$

$$7.3x_1 + 11x_2 + x_3 + d_{12}^- - d_{12}^+ = 29$$

$$7.5x_1 + 12x_2 + x_3 + d_{13}^- - d_{13}^+ = 33$$

$$7.5x_1 + 12x_2 + x_3 + d_{14}^- - d_{14}^+ = 35$$

$$4.9x_1 + 9x_2 + x_3 + d_{15}^- - d_{15}^+ = 27$$

$$3.7x_1 + 6x_2 + x_3 + d_{16}^- - d_{16}^+ = 19$$

$$9.1x_1 + 12x_2 + x_3 + d_{17}^- - d_{17}^+ = 36$$

$$9.0x_1 + 9x_2 + x_3 + d_{20}^- - d_{20}^+ = 38$$

with $x_i, d_i^-, d_i^+ \geq 0,\ i = 1, 2, \ldots, 20$

The results are $x_1 = 1.875$, $x_2 = 0.6563$ and $x_3 = 11.0625$.

Thus, the predicted equation is $\hat{y}_i = 1.875x_{1i} + 0.6563x_{2i} + 11.0625$ (4.29)

### 4.4.3 Analysis on Data Set 3

As in the case of analyzing using least squares method, the analysis using goal programming is also done for independent variable, $X_3$, aluminium only.

**(i) Outliers included**

From data set 3, the complete goal programming model can be presented as follows:

Minimize $Z = P_1 \sum_{i=1}^{10} (d_i^+ + d_i^-)$

Subject to

$$64.2x_1 + 4.4x_2 + 39.8x_3 + x_4 + d_1^- - d_1^+ = 161.1$$

$$93.3x_1 + 11.1x_2 + 61x_3 + x_4 + d_2^- - d_2^+ = 308$$

$$101.3x_1 + 20.6x_2 + 71.6x_3 + x_4 + d_3^- - d_3^+ = 613$$

$$84.2x_1 + 10.5x_2 + 76x_3 + x_4 + d_4^- - d_4^+ = 460$$

$$76.5x_1 + 11.4x_2 + 77.8x_3 + x_4 + d_5^- - d_5^+ = 424$$

$$66.8x_1 + 8.1x_2 + 81x_3 + x_4 + d_6^- - d_6^+ = 361$$

$$66.1x_1 + 5.5x_2 + 81x_3 + x_4 + d_7^- - d_7^+ = 368$$

$$82.5x_1 + 7x_2 + 72.3x_3 + x_4 + d_8^- - d_8^+ = 448$$

$$120.5x_1 + 6.5x_2 + 110.1x_3 + x_4 + d_9^- - d_9^+ = 438$$

$$130.9x_1 + 5.5x_2 + 87.8x_3 + x_4 + d_{10}^- - d_{10}^+ = 382.6$$

with $x_i, d_i^-, d_i^+ \geq 0, i = 1, 2, \ldots, 10$

The only goal, $P_1$ in data set 3 is to predict value of gold. This goal programming model is solved by using QM for Windows. The results are $x_1 = 0.7848$, $x_2 = 18.9434$, $x_3 = 2.0009$ and $x_4 = 0$.

Thus, the predicted equation is $\hat{y}_i = 0.7848x_{1i} + 18.9434x_{2i} + 2.0009x_{3i}$ (4.30)

**(ii) Remove first extreme outlier**

From data set 3, the complete goal programming model without first extreme outlier can be presented as follows:

Minimize $Z = P_1 \sum_{i=1}^{10} (d_i^+ + d_i^-)$

Subject to

$$93.3x_1 + 11.1x_2 + 61x_3 + x_4 + d_2^- - d_2^+ = 308$$

$$101.3x_1 + 20.6x_2 + 71.6x_3 + x_4 + d_3^- - d_3^+ = 613$$

$$84.2x_1 + 10.5x_2 + 76x_3 + x_4 + d_4^- - d_4^+ = 460$$

$$76.5x_1 + 11.4x_2 + 77.8x_3 + x_4 + d_5^- - d_5^+ = 424$$

$$66.8x_1 + 8.1x_2 + 81x_3 + x_4 + d_6^- - d_6^+ = 361$$

$$66.1x_1 + 5.5x_2 + 81x_3 + x_4 + d_7^- - d_7^+ = 368$$

$$82.5x_1 + 7x_2 + 72.3x_3 + x_4 + d_8^- - d_8^+ = 448$$

$$120.5x_1 + 6.5x_2 + 110.1x_3 + x_4 + d_9^- - d_9^+ = 438$$

$$130.9x_1 + 5.5x_2 + 87.8x_3 + x_4 + d_{10}^- - d_{10}^+ = 382.6$$

with $x_i, d_i^-, d_i^+ \geq 0, i = 1, 2, \ldots, 10$

The results are $x_1 = 0.0434$, $x_2 = 17.2028$, $x_3 = 1.7331$ and $x_4 = 130.1294$.

Thus, the predicted equation is

$$\hat{y}_i = 0.0434x_{1i} + 17.2028x_{2i} + 1.7331x_{3i} + 130.1294 \qquad (4.31)$$

**(iii) Remove second extreme outlier**

From data set 3, the complete goal programming model without second extreme outlier can be presented as follows:

$$\begin{aligned}
&\text{Minimize} && Z = P_1\sum_{i=1}^{10}(d_i^+ + d_i^-)\\
&\text{Subject to} && 64.2x_1 + 4.4x_2 + 39.8x_3 + x_4 + d_1^- - d_1^+ = 161.1\\
& && 93.3x_1 + 11.1x_2 + 61x_3 + x_4 + d_2^- - d_2^+ = 308\\
& && 101.3x_1 + 20.6x_2 + 71.6x_3 + x_4 + d_3^- - d_3^+ = 613\\
& && 84.2x_1 + 10.5x_2 + 76x_3 + x_4 + d_4^- - d_4^+ = 460\\
& && 76.5x_1 + 11.4x_2 + 77.8x_3 + x_4 + d_5^- - d_5^+ = 424\\
& && 66.8x_1 + 8.1x_2 + 81x_3 + x_4 + d_6^- - d_6^+ = 361\\
& && 66.1x_1 + 5.5x_2 + 81x_3 + x_4 + d_7^- - d_7^+ = 368\\
& && 82.5x_1 + 7x_2 + 72.3x_3 + x_4 + d_8^- - d_8^+ = 448\\
& && 130.9x_1 + 5.5x_2 + 87.8x_3 + x_4 + d_{10}^- - d_{10}^+ = 382.6\\
&\text{with} && x_i, d_i^-, d_i^+ \geq 0,\ i = 1, 2, \ldots, 10
\end{aligned}$$

The results are $x_1 = 0$, $x_2 = 18.6781$, $x_3 = 3.1876$ and $x_4 = 0$.

Thus, the predicted equation is $\hat{y}_i = 18.9434x_{2i} + 2.0009x_{3i}$ (4.32)

**(iv) Remove both extreme outliers**

From data set 3, the complete goal programming model when both extreme outliers are removed can be presented as follows:

$$\begin{aligned}
&\text{Minimize} && Z = P_1\sum_{i=1}^{10}(d_i^+ + d_i^-)\\
&\text{Subject to} && 93.3x_1 + 11.1x_2 + 61x_3 + x_4 + d_2^- - d_2^+ = 308
\end{aligned}$$

$$101.3x_1 + 20.6x_2 + 71.6x_3 + x_4 + d_3^- - d_3^+ = 613$$

$$84.2x_1 + 10.5x_2 + 76x_3 + x_4 + d_4^- - d_4^+ = 460$$

$$76.5x_1 + 11.4x_2 + 77.8x_3 + x_4 + d_5^- - d_5^+ = 424$$

$$66.8x_1 + 8.1x_2 + 81x_3 + x_4 + d_6^- - d_6^+ = 361$$

$$66.1x_1 + 5.5x_2 + 81x_3 + x_4 + d_7^- - d_7^+ = 368$$

$$82.5x_1 + 7x_2 + 72.3x_3 + x_4 + d_8^- - d_8^+ = 448$$

$$130.9x_1 + 5.5x_2 + 87.8x_3 + x_4 + d_{10}^- - d_{10}^+ = 382.6$$

with $x_i, d_i^-, d_i^+ \geq 0, i = 1, 2, \ldots, 10$

The results are $x_1 = 0$, $x_2 = 18.6781$, $x_3 = 3.1876$ and $x_4 = 0$.

Thus, the predicted equation is $\hat{y}_i = 18.6781x_{2i} + 3.1876x_{3i}$ (4.33)

## 4.5 Concluding Remarks

This chapter described the data analysis using least squares method and goal programming method. The analysis were done for data sets in the following conditions: outliers included and outlier(s) removed. Outliers and the box plot technique that can be used to identify outliers were also discussed.

# CHAPTER 5

# COMPARISON BETWEEN THE LEAST SQUARES AND GOAL PROGRAMMING METHOD

## 5.0 Introduction

In chapter 4, the least squares and the goal programming methods have been used to analyze data sets that contain outliers and without outlier(s). The objectives of the analysis were to obtain prediction equations. This chapter will compare the prediction values obtained using the least squares and the goal programming methods. Mean absolute percentage error (MAPE) will be used to analyze the comparison.

## 5.1 Mean Absolute Percentage Error (MAPE)

When choosing between competing models or when evaluating an existing model, we need to use measures that summarize the overall accuracy provided by the model(s) (Mendenball et al, 1993). Generally, the closer the estimates $\hat{y}_i$ are to the actual $y_i$ of the series, the more accurate the f model is. Thus, the quality of a model can be evaluated by examining the series of errors $(y_i - \hat{y}_i)$.

Since MAPE is measured as a percentage, it is particularly useful for comparing the performance of a model on many different time series. The mean absolute percentage error (MAPE) is the average of the absolute values of the percentage errors.

The formula for MAPE is as follow:

$$MAPE = \frac{\sum_{i=1}^{n}[\frac{|e_i|}{y_i}(100)]}{n} \tag{5.1}$$

where $n$ is number of predictions. A large value of MAPE means that the value of error is large.

## 5.2 MAPE for Data Set 1

### (i) Outliers included

Predictions will be calculated for the 21st to 25th observations.

For the 21st observation; $x_{21} = 108.1$, $y_{21} = 102$

From (4.1), we have $\hat{y}_{i,LS} = 32.7708 + 0.7316x_i$

Substitute $x_{21}$ = 108.1 into this equation, $\hat{y}_{21} = 32.7708 + 0.7316(108.1) = 111.86$

From (4.21), we have $\hat{y}_{i,GP} = 0.7215x_i + 30.1839$

Substitute $x_{21}$ = 108.1 into this equation, $\hat{y}_{21} = 0.7215(108.1) + 30.1839 = 108.18$

The complete predicted values and errors/residuals are shown in Table 5.1.

**Table 5.1 : The Predicted Values and Errors for Set 1**

| Observation number | $x_i$ | $y_i$ | $\hat{y}_{i,LS}$ | $e_{i,LS}$ | $\hat{y}_{i,GP}$ | $e_{i,GP}$ |
|---|---|---|---|---|---|---|
| 21 | 108.1 | 102 | 111.86 | -9.86 | 108.18 | -6.18 |
| 22 | 89.4 | 91 | 98.18 | -7.18 | 94.69 | -3.69 |
| 23 | 76.4 | 97 | 88.67 | 8.33 | 85.31 | 11.69 |
| 24 | 131.7 | 128 | 129.12 | -1.12 | 125.21 | 2.79 |
| 25 | 100.8 | 88 | 106.52 | -18.52 | 102.91 | -14.91 |

$\text{MAPE}_{\text{LS}}$ =

$$\frac{\sum_{i=21}^{25}[(\frac{|-9.86|}{102}(100)) + (\frac{|-7.18|}{91}(100)) + (\frac{|8.33|}{97}(100)) + (\frac{|-1.12|}{128}(100)) + (\frac{|-18.52|}{88}(100))]}{5}$$

= 9.62 %

$\text{MAPE}_{\text{GP}}$ =

$$\frac{\sum_{i=21}^{25}[(\frac{|-6.18|}{102}(100)) + (\frac{|-3.69|}{91}(100)) + (\frac{|11.69|}{97}(100)) + (\frac{|2.79|}{128}(100)) + (\frac{|-14.91|}{88}(100))]}{5}$$

= 8.26 %

**(ii) Remove mild outlier**

The complete predicted values and errors/residuals with mild outlier removed are shown in Table 5.2.

**Table 5.2 : The Predicted Values and Errors for Set 1 with Mild Outlier Removed**

| Observation number | $x_i$ | $y_i$ | $\hat{y}_{i,LS}$ | $e_{i,LS}$ | $\hat{y}_{i,GP}$ | $e_{i,GP}$ |
|---|---|---|---|---|---|---|
| 21 | 108.1 | 102 | 112.13 | -10.13 | 108.18 | -6.18 |
| 22 | 89.4 | 91 | 98.20 | -7.20 | 94.69 | -3.69 |
| 23 | 76.4 | 97 | 88.52 | 8.48 | 85.31 | 11.69 |
| 24 | 131.7 | 128 | 129.70 | -1.70 | 125.21 | 2.79 |
| 25 | 100.8 | 88 | 106.69 | -18.69 | 102.91 | -14.91 |

$\text{MAPE}_{\text{LS}}$ = 9.83 % and $\text{MAPE}_{\text{GP}}$ = 8.26 %

**(iii) Remove extreme outlier**

The complete predicted values and errors/residuals with extreme outlier removed are shown in Table 5.3.

**Table 5.3 : The Predicted Values and Errors for Set 1 with Extreme Outlier Removed**

| Observation number | $x_i$ | $y_i$ | $\hat{y}_{i,LS}$ | $e_{i,LS}$ | $\hat{y}_{i,GP}$ | $e_{i,GP}$ |
|---|---|---|---|---|---|---|
| 21 | 108.1 | 102 | 111.61 | -9.61 | 111.29 | -9.29 |
| 22 | 89.4 | 91 | 97.21 | -6.21 | 96.26 | -5.26 |
| 23 | 76.4 | 97 | 87.21 | 9.79 | 85.81 | 11.19 |
| 24 | 131.7 | 128 | 129.78 | -1.78 | 130.26 | -2.26 |
| 25 | 100.8 | 88 | 105.99 | -17.99 | 105.42 | -17.42 |

$\text{MAPE}_{LS} = 9.63$ % and $\text{MAPE}_{GP} = 9.60$ %

**(iv) Remove both mild and extreme outliers**

The complete predicted values and errors/residuals with both mild and extreme outliers removed are shown in Table 5.4.

**Table 5.4 : The Predicted Values and Errors for Set 1 with Both Mild and Extreme Outliers Removed**

| Observation number | $x_i$ | $y_i$ | $\hat{y}_{i,LS}$ | $e_{i,LS}$ | $\hat{y}_{i,GP}$ | $e_{i,GP}$ |
|---|---|---|---|---|---|---|
| 21 | 108.1 | 102 | 111.88 | -9.88 | 112.48 | -10.48 |
| 22 | 89.4 | 91 | 96.09 | -5.09 | 95.80 | -4.80 |
| 23 | 76.4 | 97 | 85.12 | 11.88 | 84.20 | 12.80 |
| 24 | 131.7 | 128 | 131.80 | -3.80 | 133.53 | -5.53 |
| 25 | 100.8 | 88 | 105.72 | -17.72 | 105.97 | -17.97 |

$\text{MAPE}_{LS} = 10.13$ % and $\text{MAPE}_{GP} = 10.70$ %

### 5.2.1 Discussion of Data Set 1 Results

**Table 5.5 : MAPE for Data Set 1**

| | $MAPE_{LS}$ (%) | $MAPE_{GP}$ (%) |
|---|---|---|
| With outliers | 9.62 | 8.26 |
| Mild outlier removed | 9.83 | 8.26 |
| Extreme outlier removed | 9.63 | 9.60 |
| Both mild and extreme outliers removed | 10.13 | 10.70 |

From Table 5.5, MAPE of the least squares method is always higher than the goal programming model except when both mild and extreme outliers are removed. The case when both outliers are removed, MAPE of goal programming model is 10.70% and 10.13% for least squares method. $MAPE_{GP}$ is 0.57% higher than $MAPE_{LS}$. For cases without extreme outlier and both outliers, $MAPE_{LS}$ and $MAPE_{GP}$ is closer. That is, 9.63% versus 9.60% and 10.13% versus 10.70%.

From the analysis, the average of MAPE for all cases is 9.50%. This means that the error in data set 1 is high. One of the reasons is the number of data points is very small. In data set 1, only 20 pairs of observations are used for analysis.

However, we can conclude that goal programming model is better than least squares method when outliers are included in the data sets. If outlier(s) is removed, sometimes least squares method is better than goal programming model, or vice versa.

## 5.3 MAPE for Data Set 2

### (i) Outliers included

For this data set, the 21st to 25th observations will be analyzed. The calculations for the 21st observation is shown as follows:

For the 21st observation;

$x_{1,21} = 7.0$, $x_{2,21} = 12$, $y_{21} = 40$

From (4.8), we have $\hat{y}_{i,LS} = 22.2 + 1.1x_{1i} + 0.0167x_{2i}$

Substitute $x_{1,21} = 7.0$ and $x_{2,21} = 12$ into this equation,

$\hat{y}_{21} = 22.2 + 1.1(7.0) + 0.0167(12) = 30.10$

From (4.25), we have $\hat{y}_{i,GP} = 0.5055x_{1i} + 1.2615x_{2i} + 15.5055$

Substitute $x_{1,21} = 7.0$ and $x_{2,21} = 12$ into this equation,

$\hat{y}_{21} = 0.5055(7.0) + 1.2615(12) + 15.5055 = 34.18$

The complete predicted values and errors/residuals are shown in Table 5.6.

**Table 5.6 : The Predicted Values and Errors for Set 2**

| **Observation number** | $x_{1i}$ | $x_{2i}$ | $y_i$ | $\hat{y}_{i,LS}$ | $e_{i,LS}$ | $\hat{y}_{i,GP}$ | $e_{i,GP}$ |
|---|---|---|---|---|---|---|---|
| 21 | 7.0 | 12 | 40 | 30.10 | 9.90 | 34.18 | -4.08 |
| 22 | 12.5 | 16 | 42 | 36.22 | 5.78 | 42.01 | -0.01 |
| 23 | 5.0 | 6 | 21 | 27.80 | -6.80 | 25.60 | -4.60 |
| 24 | 6.8 | 12 | 29 | 29.88 | -0.88 | 34.08 | -5.08 |
| 25 | 7.2 | 14 | 35 | 30.35 | 4.65 | 36.81 | -1.81 |

MAPE$_{LS}$ =

$$\frac{\sum_{i=21}^{25}[(\frac{|9.90|}{40}(100)) + (\frac{|5.78|}{42}(100)) + (\frac{|-6.80|}{21}(100)) + (\frac{|-0.88|}{29}(100)) + (\frac{|4.65|}{35}(100))]}{5}$$

= 17.44 %

MAPE$_{GP}$ =

$$\frac{\sum_{i=21}^{25}[(\frac{|-4.08|}{40}(100)) + (\frac{|-0.01|}{42}(100)) + (\frac{|-4.6|}{21}(100)) + (\frac{|-5.08|}{29}(100)) + (\frac{|-1.81|}{35}(100))]}{5}$$

= 10.96 %

### (ii) Remove first mild outlier

The complete predicted values and errors/residuals with the first mild outlier removed are shown in Table 5.7.

**Table 5.7 : The Predicted Values and Errors for Set 2 with First Mild Outlier Removed**

| **Observation number** | $x_{1i}$ | $x_{2i}$ | $y_i$ | $\hat{y}_{i,LS}$ | $e_{i,LS}$ | $\hat{y}_{i,GP}$ | $e_{i,GP}$ |
|---|---|---|---|---|---|---|---|
| 21 | 7.0 | 12 | 40 | 29.86 | 10.14 | 34.40 | 5.60 |
| 22 | 12.5 | 16 | 42 | 35.07 | 6.93 | 41.30 | 0.70 |
| 23 | 5.0 | 6 | 21 | 27.88 | -6.88 | 27.00 | -6.00 |
| 24 | 6.8 | 12 | 29 | 29.68 | -0.68 | 34.31 | -5.31 |
| 25 | 7.2 | 14 | 35 | 30.09 | 4.91 | 36.66 | -1.66 |

$\text{MAPE}_{\text{LS}} = 18.20$ % and $\text{MAPE}_{\text{GP}} = 13.46$ %

### (iii) Remove second mild outlier

The complete predicted values and errors/residuals with the second mild outlier removed are shown in Table 5.8.

**Table 5.8 : The Predicted Values and Errors for Set 2 with Second Mild Outlier Removed**

| **Observation number** | $x_{1i}$ | $x_{2i}$ | $y_i$ | $\hat{y}_{i,LS}$ | $e_{i,LS}$ | $\hat{y}_{i,GP}$ | $e_{i,GP}$ |
|---|---|---|---|---|---|---|---|
| 21 | 7.0 | 12 | 40 | 30.05 | 9.95 | 32.26 | 7.74 |
| 22 | 12.5 | 16 | 42 | 37.09 | 4.91 | 43.14 | -1.14 |
| 23 | 5.0 | 6 | 21 | 27.43 | -6.43 | 25.23 | -4.23 |
| 24 | 6.8 | 12 | 29 | 29.80 | -0.80 | 31.96 | -2.96 |
| 25 | 7.2 | 14 | 35 | 30.33 | 4.67 | 33.90 | 1.10 |

$\text{MAPE}_{\text{LS}} = 16.66$ % and $\text{MAPE}_{\text{GP}} = 11.11$ %

**(iv) Remove third mild outlier**

The complete predicted values and errors/residuals with the third mild outlier removed are shown in Table 5.9.

**Table 5.9 : The Predicted Values and Errors for Set 2 with Third Mild Outlier Removed**

| Observation number | $x_{1i}$ | $x_{2i}$ | $y_i$ | $\hat{y}_{i,LS}$ | $e_{i,LS}$ | $\hat{y}_{i,GP}$ | $e_{i,GP}$ |
|---|---|---|---|---|---|---|---|
| 21 | 7.0 | 12 | 40 | 29.86 | 10.14 | 34.40 | 5.60 |
| 22 | 12.5 | 16 | 42 | 35.07 | 6.93 | 41.30 | 0.70 |
| 23 | 5.0 | 6 | 21 | 27.88 | -6.88 | 27.00 | -6.00 |
| 24 | 6.8 | 12 | 29 | 29.68 | -0.68 | 34.31 | -5.31 |
| 25 | 7.2 | 14 | 35 | 30.09 | 4.91 | 36.66 | -1.66 |

$\text{MAPE}_{\text{LS}}$ = 15.96 % and $\text{MAPE}_{\text{GP}}$ = 11.11 %

**(v) Remove all of the outliers**

The complete predicted values and errors/residuals with all mild outliers removed are shown in Table 5.10.

**Table 5.10 : The Predicted Values and Errors for Data Set 2 with All the Mild Outliers Removed**

| Observation number | $x_{1i}$ | $x_{2i}$ | $y_i$ | $\hat{y}_{i,LS}$ | $e_{i,LS}$ | $\hat{y}_{i,GP}$ | $e_{i,GP}$ |
|---|---|---|---|---|---|---|---|
| 21 | 7.0 | 12 | 40 | 29.69 | 10.31 | 32.06 | 7.94 |
| 22 | 12.5 | 16 | 42 | 45.02 | -3.02 | 45.00 | -3.00 |
| 23 | 5.0 | 6 | 21 | 24.19 | -3.19 | 24.38 | -3.38 |
| 24 | 6.8 | 12 | 29 | 29.13 | -0.13 | 31.69 | -2.69 |
| 25 | 7.2 | 14 | 35 | 30.21 | 4.71 | 33.75 | 1.25 |

$\text{MAPE}_{\text{LS}}$ = 12.41 % and $\text{MAPE}_{\text{GP}}$ = 11.19 %

### 5.3.1 Discussion of Data Set 2 Results

Table 5.11 : MAPE for Data Set 2

| | $MAPE_{LS}$ (%) | $MAPE_{GP}$ (%) |
|---|---|---|
| With outliers | 17.44 | 10.96 |
| First mild outlier removed | 18.20 | 13.46 |
| Second mild outlier removed | 16.60 | 11.11 |
| Third mild outlier removed | 15.96 | 11.11 |
| All outliers removed | 12.41 | 11.19 |

From the table, we get that all the MAPE is over than 10%. From this result, the average of MAPE for all cases is 13.84%. This means that the error in data set 2 is very high. One of the reasons is the number of data points is very small. In data set 2, only 20 triplets of observations are used for analysis.

$MAPE_{LS}$ is decreased when the outliers are removed one by one until all the three mild outliers are thrown away. This shows that the least squares line is better when the outlier is removed. But, MAPE of goal programming model with outliers is better than the case without outlier.

In this analysis, we can conclude that goal programming model is better than the least squares method in all cases. All of the MAPE of goal programming model are less than the MAPE for least squares.

## 5.4 MAPE for Data Set 3

### (i) Outliers included

For this data set, the 5th and 8th observations will be analyzed. The calculations for observations number 5 are as follows:

For the 5th observation;

$x_{1,5} = 72.8, x_{2,5} = 8.0, \quad x_{3,5} = 76.0, y_{1,5} = 376$

From (4.13), we have $\hat{y}_{i,LS} = -40.9 - 0.23x_{1i} + 18.6x_{2i} + 3.83x_{3i}$

Substitute $x_{1,5} = 72.8,\ x_{2,5} = 8.0,\ x_{3,5} = 76.0$ into this equation,

$$\hat{y}_5 = -40.9 - 0.23(72.8) + 18.6(8.0) + 3.83(76.0) = 382.24$$

From (4.30), we have $\hat{y}_{i,GP} = 0.7848x_{1i} + 18.9434x_{2i} + 2.0009x_{3i}$

Substitute $x_{1,5} = 72.8,\ x_{2,5} = 8.0,\ x_{3,5} = 76.0$ into this equation,

$$\hat{y}_5 = 0.7848(72.8) + 18.9434(8.0) + 2.0009(76.0) = 360.75$$

The complete predicted values and errors/residuals are shown in Table 5.12.

**Table 5.12 : The Predicted Values and Errors for Set 3**

| **Observation number** | $x_{1i}$ | $x_{2i}$ | $x_{3i}$ | $y_i$ | $\hat{y}_{i,LS}$ | $e_{i,LS}$ | $\hat{y}_{i,GP}$ | $e_{i,GP}$ |
|---|---|---|---|---|---|---|---|---|
| 5 | 72.8 | 8.0 | 76.0 | 376 | 382.24 | -6.24 | 360.75 | 15.25 |
| 8 | 67.0 | 6.1 | 81.0 | 318 | 367.38 | -49.38 | 330.21 | -12.21 |

$$\text{MAPE}_{\text{LS}} = \frac{\sum_{i=21}^{25}[(\frac{|-6.24|}{376}(100)) + (\frac{|-49.38|}{318}(100))]}{2}$$

$$= 8.59\ \%$$

$$\text{MAPE}_{\text{GP}} = \frac{\sum_{i=21}^{25}[(\frac{|15.25|}{376}(100)) + (\frac{|-12.21|}{318}(100))]}{2}$$

$$= 3.95\ \%$$

**(ii) Remove first extreme outlier**

The complete predicted values and errors/residuals with the first extreme outlier removed are shown in Table 5.13.

**Table 5.13 : The Predicted Values and Errors for Set 3 with First Extreme Outlier Removed**

| Observation number | $x_{1i}$ | $x_{2i}$ | $x_{3i}$ | $y_i$ | $\hat{y}_{i,LS}$ | $e_{i,LS}$ | $\hat{y}_{i,GP}$ | $e_{i,GP}$ |
|---|---|---|---|---|---|---|---|---|
| 5 | 72.8 | 8.0 | 76.0 | 376 | 387.62 | -11.62 | 402.63 | -26.63 |
| 8 | 67.0 | 6.1 | 81.0 | 318 | 371.03 | -53.03 | 378.36 | -60.36 |

$\text{MAPE}_{\text{LS}}$ = 9.88 % and $\text{MAPE}_{\text{GP}}$ = 13.03 %

**(iii) Remove second extreme outlier**

The complete predicted values and errors/residuals with the second extreme outlier removed are shown in Table 5.14.

**Table 5.14 : The Predicted Values and Errors for Set 3 with Second Extreme Outlier Removed**

| Observation number | $x_{1i}$ | $x_{2i}$ | $x_{3i}$ | $y_i$ | $\hat{y}_{i,LS}$ | $e_{i,LS}$ | $\hat{y}_{i,GP}$ | $e_{i,GP}$ |
|---|---|---|---|---|---|---|---|---|
| 5 | 72.8 | 8.0 | 76.0 | 376 | 386.70 | -10.70 | 391.68 | -15.68 |
| 8 | 67.0 | 6.1 | 81.0 | 318 | 375.99 | -57.99 | 372.13 | -54.13 |

$\text{MAPE}_{\text{LS}}$ = 10.54 % and $\text{MAPE}_{\text{GP}}$ = 10.60 %

**(iv) Remove both extreme outliers**

The complete predicted values and errors/residuals with both extreme outliers removed are shown in Table 5.15.

**Table 5.15 : The Predicted Values and Errors for Set 3 with both Extreme Outliers Removed**

| **Observation number** | $x_{1i}$ | $x_{2i}$ | $x_{3i}$ | $y_i$ | $\hat{y}_{i,LS}$ | $e_{i,LS}$ | $\hat{y}_{i,GP}$ | $e_{i,GP}$ |
|---|---|---|---|---|---|---|---|---|
| 5 | 72.8 | 8.0 | 76.0 | 376 | 386.98 | -10.98 | 391.68 | -15.68 |
| 8 | 67.0 | 6.1 | 81.0 | 318 | 377.06 | -59.06 | 372.13 | -54.13 |

$\text{MAPE}_{\text{LS}} = 10.75$ % and $\text{MAPE}_{\text{GP}} = 10.60$ %

### 5.4.1 Discussion of Data Set 3 Results

**Table 5.16 : MAPE for Data Set 3**

| | $\text{MAPE}_{\text{LS}}$ (%) | $\text{MAPE}_{\text{GP}}$ (%) |
|---|---|---|
| With outliers | 8.95 | 3.95 |
| First extreme outlier removed | 9.88 | 13.03 |
| Second extreme outlier removed | 10.54 | 10.60 |
| Both extreme outliers removed | 10.75 | 10.60 |

From Table 5.16, MAPE of least squares method is increased from 8.59% (with outliers) to 10.75% (both extreme outliers removed). For goal programming model, MAPE increases from 3.95% to 13.03% then decreases to 10.60%. Without first and second extreme outlier, $\text{MAPE}_{\text{LS}}$ is less than $\text{MAPE}_{\text{GP}}$. When both extreme outliers are removed, $\text{MAPE}_{\text{LS}}$ is higher 0.15% than $\text{MAPE}_{\text{GP}}$.

From this result, the average of MAPE for all cases is 9.74%. This means that the error in data set 3 is high. One of the reasons is the number of data points is very small. In data set 3, only 10 quadraplets of observations are used for analysis.

However, we can conclude that goal programming model is better than least squares method when outliers exist in the data sets. If outlier is removed, sometimes least squares method is better than goal programming model, or vice versa.

## 5.5 Discussion of the Overall Results

An examination of the mean absolute percentage error (MAPE) may give an idea of how well the least squares method and goal programming are in analyzing data points. The equations of the least squares line and linear goal programming model are influenced by every data points used in its calculation in a manner similar to the arithmetic mean. Table 5.17 shows the MAPE of least squares method and goal programming model for all the three data sets with outliers.

**Table 5.17 : MAPE for Data Set Containing Outliers**

| | $MAPE_{LS}$ (%) | $MAPE_{GP}$ (%) | $MAPE_{LS\text{-}GP}$ (%) |
|---|---|---|---|
| Data set 1 | 9.62 | 8.26 | 1.36 |
| Data set 2 | 17.44 | 10.96 | 6.48 |
| Data set 3 | 8.59 | 3.95 | 4.64 |

For data set one, there are two outliers. One is a mild outlier and the other one is an extreme outlier. Recall section 5.2, MAPE of least square method is 9.62% and 8.26% for goal programming. The percentage of MAPE analyzed using least square method is higher than using goal programming model. The higher percentage implies that error is higher. In this analysis, we can conclude that goal programming is better than least squares method.

Data set two have three outliers. All 18.0, 23.0 and 23.5 are mild outliers at the upper end. MAPE of least squares method is 17.44% and 10.96% for goal programming model. The difference in the value of MAPE between these two methods is 6.48%. This is a quite large difference. This means that $\hat{y}_{i,GP} = 0.5055x_{1i} + 1.2615x_{2i} + 15.5055$ is more accurate than $\hat{y}_{i,LS} = 22.2 + 1.1x_{1i} + 0.0167x_{2i}$ to be used for prediction for this data set.

Data set three have two outliers for independent variable $X_3$, aluminium. That is, 39.8 is an extreme outlier at the lower end while 110.1 is an extreme outlier at the upper end. After analysis, percentage of MAPE for least squares is 8.59% and

3.95% for goal programming. MAPE for least square is 4.64% higher than goal programming. So, $\hat{y}_{i,GP} = 0.7848x_{1i} + 18.9434x_{2i} + 2.0009x_{3i}$ is more accurate than $\hat{y}_{i,LS} = 0.7848x_{1i} + 18.9434x_{2i} + 2.0009x_{3i}$ for the prediction in this set of data points.

When an error point is removed, some MAPE of least squares method and goal programming is greater than the MAPE when outliers exist. In this case, the error point cannot be thrown away from the data set because it presents a valid segment of the population. If we look at the overall analysis result, we can conclude that these outlier data points play a key role in the data set. So, the researcher must decide whether these extraordinary events should be presented in the sample, or can be eliminated.

## 5.6 Concluding Remarks

In all the three cases, it has been shown that MAPE for predictions using goal programming are lower than those produced using least squares method. It can be concluded that the prediction equations obtained from goal programming are more accurate than those obtained from least squares method when using data sets that contain outliers.

# CHAPTER 6

# CONCLUSIONS AND SUGGESTIONS FOR FUTURE INVESTIGATION

## 6.0 Introduction

In this chapter the conclusion and summary of the study will be presented. Finally, suggestions for future work are made.

## 6.1 Conclusions

The least squares method is used to describe the approximate relationship between a criterion and a predictor variable, based on a sample of data. The resulting equation can be used to predict the criterion variable based on a specified value for the predictor variable. Prediction is one of the uses of a regression equation. For example, given a sample of values ($x_i$, $y_i$), the regression of $Y$ on $X$ being linear, we can predict the value of $Y$ corresponding to a further observed $X$ value.

The least squares method is affected by outliers in data sets. Outliers are data points that lie apart from the rest of the points. In other word, outliers are observations that have extremely large residuals or errors. They do not fit in with the pattern of the remaining data points. The equation of the regression line is influenced by every data point. Therefore, outliers can unduly influence the least squares equation. As a result, the least squares equation developed from data sets which contains outliers cannot predict the future values of $Y$ very well. Something need to be done with these outliers. We can either delete them, or find other techniques to create a new prediction equation. It depends on the researcher whether these error points should be represented in the sample.

The current study proposes the goal programming approach to solve the problem of outliers when developing prediction equations. Goal programming is one of the techniques to solve multiple-objective decision making problems. Decision making in modern organizations often involves more than one goal or objective. Generally, goal programming deals with decision problems involving conflicting objectives. The basic idea of the technique is to transform the multi objective problem into one or more problems with one objective each. The goal approach is not the ultimate solution for all managerial decision problems. It requires that the decision maker be capable of defining, quantifying and ordering objectives. The technique simply provides the best solution under the given constraints and priority structure of goals. The quality of the final solution is influenced by the decision maker's ranking of the different objectives as well as by the "tightness" of the limits set for the goals. In this regard, goal programming seeks an efficient solution that attempts to meet all the goals of the problem.

The first step in problem solving using the goal programming method is to formulate the complete goal programming model. This formulation include the development of an objective function, goal and system constraints, deviational variables and decision variables. In this context, deviation is the failure to achieve a particular goal which will result in a positive or negative deviation from the goal. In the goal programming formulation, the deviational variable is the same as the error term in the least squares equation. That is $e_i = d_i^- - d_i^+$ (refer equation (4.2)). By comparing these two equations, it can be concluded that the constraints $a$ and $b$, in the least squares model, are equal to $X_1$ and $X_2$, in the goal programming formulation (with only one predictor variable). For example, given a pair observation $(x_1, y_1) = (2, 4)$, the least squares equation is $4 = a + 2b$, while the goal programming formulation/equation is $2x_1 + x_2 + d_1^- - d_1^+ = 4$.

In the current study, Quantitative Method or QM for Window package was used in solving the goal programming problems. QM for Window is a friendly software which is easy to use. The least squares model was either solved by manual or using MINITAB software. In the case with one or two independent variables, the problem was solved manually while MINITAB software was used when there are at least two independent/predictor variables. Other software packages can also be used, such as S-Plus, Microsoft Excel and SPSS. From these analysis, predicted equations for both the least squares and goal programming were obtained.

In this study, after obtaining predicted equations from each method, estimates for some dependent variables were calculated. For example, the last five observations from data set 1 were predicted from each equation. Then, these five predicted values ($\hat{y}_{21}$ to $\hat{y}_{25}$) were compared with the actual values ($y_{21}$ to $y_{25}$). The main purpose for computing the predicted value was to obtain the errors or residuals. Here, we did not compare the error between the least squares and goal programming one by one for each data point. This is because the single residual cannot show which method is more superior for that data set. For example, from data set 1, three predicted values from the least squares were closer to the actual values while only two for goal programming. In data set 2, four predicted values from the goal programming were closer to the actual values while only one for the least squares.

In this study, the mean absolute percentage error (MAPE) was used to compare the errors obtained using the least squares and goal programming. MAPE is the average of the absolute value of the percentage errors of the percentage error.

The results in chapter 5 showed that the goal programming technique is a better method in developing a prediction equation from a set of data points with outliers included. If the outlier(s) is removed, sometimes least squares method is better than goal programming model or vice versa. Although the least squares is a powerful method in regression, the outliers in the data sets will make the equation lose its accuracy. This is clearly shown in the scatter and residual plots. The outliers in the scatter and residual plots are far from the normal pattern. So, we can conclude that the goal programming approach can be an alternative way for the problems analyzed using the method of least squares.

## 6.2 Suggestions for Future Investigation

The current discusses two popular approaches in decision making: goal programming and least squares. The scope of this study is focused on the linear goal programming. However, there are other aspects in linear goal programming that are not covered such as the duality in linear goal programming, the primal-dual algorithm for linear goal programming, a reordering and/or permutation of the original priority levels for sensitivity analysis in linear goal programming and the continuous variations over a range or parametric linear goal programming. A study in these areas might improve the level of decision making made.

A number of important advances have also been made in the area of goal programming. Some of the prominent new developments include zero-one programming, integer programming, interative systems, decomposition goal programming, nonlinear goal programming, interval goal programming and fuzzy linear programming. Therefore, further study can also be done on such topics.

Future study should also be undertaken to cover other multiple regression models such as polynomial regression, nonlinear multiple regression mode, interaction between variables and multicollinearity. This is because, in the current study, the least squares method discussed is only focused on the linear least squares and multiple least squares analysis. A comparison between other aspects in goal programming and also other multiple regression models can be explored.

## REFERENCES

Brook, Richard J. and Arnold, Gregory C. (1985). "Applied Regression Analysis and Experimental Design." USA: Marcel Dekker Inc. 1 – 50.

Bunday, Brian D. and Garside, Gerald R. (1987). "Linear Programming in Pascal." London: Edward Arnold. 54 – 55.

Chatterjee, Samprit and Hadi, Ali S. (1988). "Sensitivity Analysis in Linear Regression." New York: John Wiley & Sons. 106 – 107.

Cook, R. Dennis and Weisberg, Sanford. (1982). "Residuals and Influence in Regression." New York and London: Chapman and Hall. 32 – 33.

Cooke, William P. (1985). "Quantitative Methods for Management Decisions." New York: McGraw-Hill. 111 – 118.

Daellenbach, Hans G. (1983). "Introduction to Operational Research Techniques." 2nd. Ed. USA: Allyn and Bacon. 616 – 634.

Dantzig, George B. and Thapa, Mukund N. (1997). "Linear Programming 1: Introduction." New York: Springer-Verlag. 150 – 152.

Devore, Jay and Peck, Roxy. (2001). "Statistics: The Exploration and Analysis of Data." 4th. Ed. USA: Duxbury Thomson Learning. 497 – 593.

Dinkel, John J, et al. (1978). "Management Science: Text and Applications." USA: Richard D. Irwin. 162 – 185.

Goicoechea, Ambrose, et al. (1982). "Multiobjective Decision Analysis with Engineering and Business Applications." Canada: John Wiley & Son. 99 – 160.

Green, Paul E. and Carroll, J. Douglas. (1976). "Mathematical Tools for Applied Multivariate Analysis." New York: Academic Press Inc. 259 – 270.

Gunst, Richard F. and Mason, Robert L. (1980). "Regression Analysis & Its Application: A Data-Oriented Approach." New York and Basel: Marcel Dekker Inc. 52 – 132.

Hair, Joseph F, et al. (1998). "Multivariate Data Analysis." 5th. Ed. New Jersey: Prentice Hall.

Hillier, Frederick S. and Lieberman, Gerald J. (2001). "Introduction to Operational Research." 7th. Ed. New York: McGraw-Hill. 332 – 340.

Holzman, Albert G. (Ed.) (1981). "Mathematical Programming for Operations Researchers and Computer Scientists." New York: Marcel Dekker Inc. 101 – 122.

Hughes, Ann J. and Grawoig, Dennis E. (1973). "Linear Programming: An Emphasis on Decision Making." London: Addison-Wesley Publishing Company Inc. 300 – 316.

Ignizio, James P. (1976). "Goal Programming and Extensions." Canada: D. C. Heath and Company. 1 – 114.

Ignizio, James P. (1982). "Linear Programming In single- & Multiple- Objective Systems." Englewood Cliffs, New Jersey: Prentice-Hall. 372 – 473.

Ignizio, James P. and Cavalier, Tom M. (1994). "Linear Programming." New Jersey: Prentice Hall Inc..

Larsen, Richard J. and Marx, Morris L. (2001). "An Introduction to Mathematical Statistics and Its Applications." USA: Prentice Hall International Inc. 558 – 565.

Lee, Sang M. and Shim, Jung. (1986). "Micro Management Science: Microcomputer Applications of Management Science." Dubuque, Iowa: Wm. C. Brown Publishers. 169 – 185.

Markland, Robert E. and Sweigart, James R. (1987). "Quantitative Methods: Applications to Managerial Decision Making." New York: John Wiley & Sons. 314 – 340.

Mendenball, William, et al. (1993). "Statistics for Management And Economics." USA: Duxbury Press. 668 – 669.

Schniederjans, Marc J. (1995). "Goal Programming: Methodology and Applications." USA: Kluwer Academic Publishers. 1 – 40.

Sprent, Peter. (1969). "Model in Regression and Related Topics." Methuen & Co Ltd.

Taha, Hamdy A. (1997). "Operational Research: An Introduction." 6th. Ed. Upper Saddle river, New Jersey: Prentice-Hall. 349 – 363.

Wittink, Dick R. (1988). "The Application of Regression Analysis." Allyn and Bacon. 1 – 97.

Wonnacott, Thomas H. and Wonnacott, Ronald J. (1981). "Regression: A Second Course In Statistics." Canada: John Wiley & Sons. 13 – 148.

Wu, Nesa and Coppins, Richard. (1981). "Linear Programming and Extensions." New York: McGraw-Hill. 358 – 393.