Robust techniques for regression models with minimal assumptions

M.M. van der Westhuizen

12977640

Dissertation submitted in partial fulfilment of the requirements for the degree *Master of Science* at the Potchefstroom campus of the North-West University

Supervisor: Prof. H.A. Krüger

Co-supervisor: Prof. J.M. Hattingh

January 2011

ABSTRACT

Good quality management decisions often rely on the evaluation and interpretation of data. One of the most popular ways to investigate possible relationships in a given data set is to follow a process of fitting models to the data. Regression models are often employed to assist with decision making. In addition to decision making, regression models can also be used for the optimization and prediction of data. The success of a regression model, however, relies heavily on assumptions made by the model builder. In addition, the model may also be influenced by the presence of outliers; a more robust model, which is not as easily affected by outliers, is necessary in making more accurate interpretations about the data. In this research study robust techniques for regression models with minimal assumptions are explored. Mathematical programming techniques such as linear programming, mixed integer linear programming, and piecewise linear regression are used to formulate a nonlinear regression model. Outlier detection and smoothing techniques are included to address the robustness of the model and to improve predictive accuracy. The performance of the model is tested by applying it to a variety of data sets and comparing the results to those of other models. The results of the empirical experiments are also presented in this study.

Keywords: robust regression, outlier detection, piecewise linear regression, linear programming, smoothing techniques, optimization.

OPSOMMING

ROBUUSTE TEGNIEKE VIR MODELLE MET MINIMALE AANNAMES

Om hoë kwaliteit bestuursbesluite te maak hang dikwels af van die evaluering en interpretasie van data. Een van die mees algemene en gewilde maniere om moontlike verwantskappe in ‘n gegewe datastel te ondersoek, is om ‘n proses te volg wat ‘n model op die data pas. ‘n Regressiemodel word dikwels aangewend om die besluitnemingsproses te ondersteun. Behalwe vir besluitneming kan ‘n regressiemodel ook gebruik word in optimering en voorspelling. Die sukses van ‘n regressiemodel hang egter grootliks af van die aannames wat deur die modelbouer gemaak word. Die regressiemodel kan ook maklik beïnvloed word deur die teenwoordigheid van uitskieters. ‘n Meer robuuste model, wat nie maklik deur uitskieters beïnvloed word nie, is nodig om meer akkurate interpretasies oor die data te maak. In hierdie navorsingstudie word robuuste tegnieke vir regressiemodelle met minimale aannames ondersoek. Wiskundige programmeringstegnieke, o.a. lineêre programmering, gemengde heeltallige lineêre programmering en stuksgewyse lineêre regressie, word gebruik om die robuustheid van die model aan te spreek en die akkuraatheid van voorspellings te verbeter. Die model is getoets deur dit op verskillende datastelle toe te pas en die resultate te evalueer. Die resultate van die empiriese eksperimente word ook in hierdie studie voorgehou.

Sleutelwoorde: robuuste regressie, opspoor van uitskieters, stuksgewyse lineêre regressie, lineêre programmering, gladmakingstegnieke, optimering.

ACKNOWLEDGEMENTS

I hereby want to thank and acknowledge my supervisor, Prof. Krüger and my co-supervisor Prof. Hattingh for their help and advice throughout this study. I would also like to wish Prof. Hattingh a swift recovery from his operation.

I appreciate the support of my friends and family during this study and I want to give glory and honour to God for giving me the ability to do research for His glory.

CONTENTS

Chapter 1

Introduction and problem statement

1.1 Introduction

The successes or failures that managers experience in business are largely dependent upon the quality of the decisions that they make. The difference between a good and a bad decision is, to a great extent, based on the evaluation and interpretation of data. A good decision is one that is based on logic, that considers all of the available data and, in many cases, that applies a quantitative approach. One of the most popular and valuable techniques that complies with these requirements is regression analysis. Its purpose is to understand the relationship between different variables and to predict the value of one variable based on the others. Results can then be used to guide the process of decision-making and to enable managers to make more appropriate and informed decisions.

The classical linear regression model is represented as follows:

$$y = \mathbf{X}\beta + \varepsilon, \tag{1.1}$$

where y is an $n \times 1$ vector of observed values, $\mathbf{X}$ is an $n \times k$ given matrix of values where each column vector corresponds to a predictor, β is an $k \times 1$ vector of unknown parameters and ε is an $n \times 1$ vector of (random) errors, ε_i.

It is assumed that the error terms are independently distributed continuous random variables, with $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \sigma^2 > 0$. β is usually estimated by employing the least squares error criterion.

A good exposition of the technical detail concerning how to construct and test linear regression models can be found in Kutner *et al*. (2005). Two specific challenges that researchers and decision makers have to deal with when developing and using linear regression models are: the various assumptions on which the models are based and the influence of outliers on the final model. These challenges are the basis of this study. In the problem statement below, these two issues are further described.

The purpose of this chapter is to guide the reader through the research study by explaining the problem statement, the objectives of the study and the methodology employed. A layout of the study, explaining the purpose of each chapter is also presented.

1.2 Problem statement

The success of a regression model relies heavily on assumptions made by the model builder. There are a large number of literature resources that deal in great detail with these assumptions which include: the non-stochastic and uncorrelated nature of independent variables, the normal distribution of error variables and the linear and adequate nature of the regression function. The second issue regarding outliers is associated with the robustness of a model. Outliers can be defined as observations that do not follow the same model as the rest of the data (Hoeting *et al*., 1996) while robust regression tries to devise estimators that are not strongly affected by outliers (Rousseeuw & Leroy, 2003). The presence of outliers may lead to models that are not reliable as they cause so-called "masking problems" wherein multiple outliers in a data set may conceal the presence of additional outliers.

To address the two abovementioned problem areas, this study will use an existing minimal assumption regression model (Wagner, 1962) and add certain extensions to it to improve the model's robustness. The extensions are implemented through the use of linear and mixed integer linear programming techniques and include outlier detection and smoothing techniques.

1.3 Objectives of the study

The primary objective of this study is to investigate robust techniques for regression models with minimal assumptions by using linear programming techniques. This will be accomplished by addressing the following secondary research objectives:

- gain a clear understanding of and present an introductory overview of linear regression, outliers and linear and integer linear programming;
- perform an exploratory investigation into robust techniques for regression models with minimal assumptions;
- address robustness by introducing an adapted minimal assumption mixed integer linear programming model that is able to deal with possible outliers and the smoothing of functions; and
- apply the adapted model to different data sets in order to evaluate its performance.

1.4 Research methodology

The research study can be divided into three sections, a literature study, a model development phase and an empirical study. The general literature survey gives an overview of linear regression, linear programming, robust regression, outliers and mixed integer linear programming techniques. The model development phase consists of the minimal assumption

regression model used in this study as well as the extensions that are added to refine the model. This will be followed by empirical experiments using mathematical programming techniques to formulate and illustrate the effectiveness of the minimal assumption regression model using real world data.

1.5 Chapter outline

This section explains the purpose of each chapter and how it is structured.

Chapter 2 presents an overview of linear regression, outliers and linear programming. The most important types of model will be briefly reviewed and, where appropriate, the mathematical formulation will also be provided.

Chapter 3 introduces the minimal assumption regression model that is used as the basis of this study. The model will be thoroughly described and a data set will be used to illustrate how the model can be applied to data. A brief overview of other researchers who referred to this approach is also included in Chapter 3.

Chapter 4 introduces an adapted minimal assumption regression model which is used to address issues of robustness. Outlier detection is incorporated into the model through the use of a mixed integer linear programming technique. Smoothing techniques are also included in the model. Finally, a piecewise linear regression model is introduced for comparative purposes.

Chapter 5 applies the adapted model to a variety of data sets from the literature and the results of the empirical study are evaluated and discussed.

Finally, Chapter 6 summarises the objectives set forth for the study and how these were achieved. Opportunities for further studies will also be pointed out.

1.6 Chapter summary

Chapter 1 served as an introduction to the research study and explained the problem statement, objectives of the study and the methodology to be followed for the rest of the study. A layout of the study, explaining the purpose of each chapter, was also presented.

Chapter 2

Linear regression modelling and robustness

2.1 Introduction

The primary objective of this study is to investigate robust techniques for regression models with minimal assumptions. To provide sufficient background and to gain a sound understanding of techniques that will be used, this chapter presents an introductory overview of the concepts used in subsequent chapters.

The chapter starts with a review of linear regression models and will describe three well known methods that are commonly used to estimate regression parameters: the least squares (L_2-norm), the least sum of absolute deviation (L_1-norm) and the Chebychev (L_∞-norm) methods. Next, a definition of outliers and their influence on regression models will be presented while robust regression methods will also be discussed. Finally, the basic theory of a linear programming model will be explained. Aspects such as the formulation and solving of a linear programming model will be briefly reviewed.

2.2 Linear regression

Regression analysis is a quantitative technique that estimates relationships between dependent variable(s) and other variables, often called predictor or explanatory variables (Kutner *et al*., 2005). The predictor variables are also known as independent variables, but according to Chatterjee and Hadi (2006) this name is preferred least because the independence of predictor variables is rarely a proper assumption in practice. Regression techniques are widely used in areas such as business, biological, social and behavioural sciences, and are normally used for the prediction, description and optimization of variables.

A linear regression function is referred to as a simple linear regression model when only one predictor variable is used to estimate values of the dependent variable, y (see Appendix A, section A.1). Multiple linear regression is used when two or more predictor variables are made use of to predict values of the dependent variable. The parameters of the regression model can be estimated using the L_1-, L_2- or L_∞-norm and will be further discussed in subsequent sections.

2.2.1 Multiple linear regression (L_2-norm)

Often one variable in a regression model does not explain the dependent variable satisfactorily. For such cases the simple linear regression model can be extended to a multiple linear regression model by introducing additional predictor variables. A regression model that employs more than one predictor variable is termed a multiple linear regression model. The general form of such a model is defined by Bowerman *et al.* (2005) as follows:

The linear regression model relating y to $x_1, x_2, \ldots, x_k$ is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon, \tag{2.1}$$

where

$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$ is the mean value of the dependent variable y when the values of the predictor variables are $x_1, x_2, \ldots, x_k$;
$\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ are unkown regression parameters relating the mean value of y to $x_1, x_2, \ldots, x_k$; and
ε is an error term that describes the effects on y of all factors other than the values of the predictor variables $x_1, x_2, \ldots, x_k$.

For equation (2.1) it is assumed that n observations exist, with each observation consisting of an observed value of y and corresponding observed values of $x_1, x_2, \ldots, x_k$.

As is the case with the simple linear regression model, the important assumptions for the multiple linear regression model can be summarized as follows: the error terms are assumed to be independently and identically distributed normal random variables each with a mean of zero and constant variance, σ^2. The implied assumptions are given by Bowerman *et al.* (2005) as:

- *Independence assumption.* Any one value of the error term ε is statistically independent of any other value of ε. That is, the value of the error term ε corresponding to an observed value of y is statistically independent of the value of the error term corresponding to any other observed value of y;
- *Normality assumption*. At any given combination of values of $x_1, x_2, \ldots, x_k$, the population of potential error term values has a normal distribution;
- At any given combination of values of $x_1, x_2, \ldots, x_k$, the population of potential error term values has a mean equal to zero; and
- *Constant variance assumption*. At any given combination of values of $x_1, x_2, \ldots, x_k$, the population of potential error term values has a variance that does not depend on the combination of values of $x_1, x_2, \ldots, x_k$. That is, the different populations of potential error

term values corresponding to different combination of values of $x_1, x_2, \ldots, x_k$, have equal variances. The constant variance is denoted by σ^2.

According to Kutner *et al*. (2005) the multiple linear regression model defined in (2.1) can also be expressed in matrix terms

$$\underset{n \times 1}{\mathbf{Y}} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \tag{2.2}$$

$$\underset{n \times (k+1)}{\mathbf{X}} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \tag{2.3}$$

$$\underset{(k+1) \times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \tag{2.4}$$

$$\underset{n \times 1}{\boldsymbol{\varepsilon}} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}. \tag{2.5}$$

Note that the $\mathbf{X}$ matrix contains a column of 1s to allow for β_0, the intercept, as well as a column of the n observations for each of the k variables in the regression model (therefore the dimensions are different from the classical model presented in (1.1)). The row subscript for each element x_{ij} in the $\mathbf{X}$ matrix identifies the trail or case, while the column subscript identifies the predictor variable. In matrix terms, the general linear regression model can be described as

$$\underset{n \times 1}{\mathbf{Y}} = \underset{n \times (k+1)}{\mathbf{X}} \quad \underset{(k+1) \times 1}{\boldsymbol{\beta}} + \underset{n \times 1}{\boldsymbol{\varepsilon}}, \tag{2.6}$$

where

$\mathbf{Y}$ is a vector of responses;

$\boldsymbol{\beta}$ is a vector of parameters;

$\mathbf{X}$ is a matrix of constants;

$\boldsymbol{\varepsilon}$ is a vector of independent normal random variables with an expectation of $\mathbf{E}\{\boldsymbol{\varepsilon}\} = \mathbf{0}$; and with a variance-covariance matrix of

$$\underset{n \times n}{\boldsymbol{\sigma}^2\{\boldsymbol{\varepsilon}\}} = \begin{bmatrix} \sigma^2 & 0 & \ldots & 0 \\ 0 & \sigma^2 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I}. \tag{2.7}$$

Consequently, the random vector $\mathbf{Y}$ has an expectation of

$$\underset{n \times 1}{\mathbf{E}\{\mathbf{Y}\}} = \mathbf{X}\boldsymbol{\beta}, \tag{2.8}$$

and the variance-covariance matrix of $\mathbf{Y}$ is the same as that of $\boldsymbol{\varepsilon}$

$$\underset{n \times n}{\boldsymbol{\sigma}^2\{\mathbf{Y}\}} = \sigma^2\mathbf{I}. \tag{2.9}$$

Once a relationship is established, the strength of the model must be described. This is undertaken by estimating the regression coefficients first and then looking at the significance of the coefficients by making inferences. The regression coefficients $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ are usually unknown and must be estimated. The method of least squares (L_2-norm) considers the deviations of y_i from its expected value

$$y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}), \tag{2.10}$$

where $i = 1,2, \ldots, n$ denotes the n observations. The sum of the n squared deviations, can be denoted by Q and the least square estimators, denoted by $b_0, b_1, \ldots, b_k$, are those values of $\beta_0, \beta_1, \ldots, \beta_k$ that minimize Q. Set

$$Q(b_0, b_1, \ldots, b_k) = \sum_{i=1}^{n}(y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \ldots - b_k x_{ik})^2$$

$$= \sum_{i=1}^{n}(y_i - b_0 - \sum_{j=1}^{k} b_j x_{ij})^2. \tag{2.11}$$

Q is minimized by setting $\frac{\delta Q}{\delta b_0} = 0$ and $\frac{\delta Q}{\delta b_j} = 0$ for $j = 1, 2, \ldots, k$. That is

$$\frac{\delta Q}{\delta b_0} = -2\sum_{i=1}^{n}(y_i - b_0 - \sum_{j=1}^{k} b_j x_{ij}) = 0, \tag{2.12}$$

and

$$\frac{\delta Q}{\delta b_j} = -2\sum_{i=1}^{n}(y_i - b_0 - \sum_{j=1}^{k} b_j x_{ij}) x_{ij} = 0, \tag{2.13}$$

for $j = 1, 2, \ldots, k$. Solving b_0 and b_j, for $j = 1, 2, \ldots, k$ results in the following least squares normal equations

$$\begin{bmatrix} nb_0 & + & b_1\sum_{i=1}^{n} x_{i1} & + & b_2\sum_{i=1}^{n} x_{i2} & +\cdots+ & b_k\sum_{i=1}^{n} x_{ik} & = & \sum_{i=1}^{n} y_i \\ b_0\sum_{i=1}^{n} x_{i1} & + & b_1\sum_{i=1}^{n} x_{i1}^{\,2} & + & b_2\sum_{i=1}^{n} x_{i1}x_{i2} & +\cdots+ & b_k\sum_{i=1}^{n} x_{i1}x_{ik} & = & \sum_{i=1}^{n} x_{i1}y_i \\ \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\ b_0\sum_{i=1}^{n} x_{ik} & + & b_1\sum_{i=1}^{n} x_{ik}x_{i1} & + & b_2\sum_{i=1}^{n} x_{ik}x_{i2} & +\cdots+ & b_k\sum_{i=1}^{n} x_{ik}^{\,2} & = & \sum_{i=1}^{n} x_{ik}y_i \end{bmatrix}. \quad (2.14)$$

The solutions to these $k+1$ normal equations are the least squares estimators $b_0, b_1, \ldots, b_k$, which can be denoted as $\mathbf{b}$, where

$$\underset{(k+1)\times 1}{\mathbf{b}} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix}. \quad (2.15)$$

Using matrix notation is a convenient way of representing multiple linear regression models. Applying the method of least squares (L_2-norm) requires finding the vector $\mathbf{b}$ that will minimize

$$\begin{aligned} Q(\mathbf{b}) &= \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} \\ &= (\mathbf{Y}-\mathbf{Xb})'(\mathbf{Y}-\mathbf{Xb}) \\ &= \mathbf{Y}'\mathbf{Y}-\mathbf{b}'\mathbf{X}'\mathbf{Y}-\mathbf{Y}'\mathbf{Xb}+\mathbf{b}'\mathbf{X}'\mathbf{Xb} \\ &= \mathbf{Y}'\mathbf{Y}-2\mathbf{b}'\mathbf{X}'\mathbf{Y}+\mathbf{b}'\mathbf{X}'\mathbf{Xb}. \end{aligned} \quad (2.16)$$

Therefore

$$\frac{\delta Q}{\delta \mathbf{b}} = -2\mathbf{X}'\mathbf{Y}+2\mathbf{X}'\mathbf{Xb}, \quad (2.17)$$

which simplifies to the least squares normal equations for the multiple linear regression model

$$\mathbf{X}'\mathbf{Xb} = \mathbf{X}'\mathbf{Y}, \quad (2.18)$$

while the least squares estimators are

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \quad (2.19)$$

The multiple linear regression model plays an important role in this research project and the remainder of this section will therefore present a brief overview of the most important techniques used to judge overall model quality. This concise survey is based on the work of Bowerman *et al.* (2005) and some of the definitions and descriptions are quoted from this source without referencing it again.

In order to compute intervals and test hypotheses when using a multiple linear regression model, it is necessary to calculate point estimates of σ^2 and σ (the constant variance and standard deviation of the different error term populations).

Suppose that the multiple linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

utilizes k predictor variables and thus has (k+1) parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$. Then, if the regression assumptions are satisfied and if SSE denotes the sum of squared residuals for the model, and n is equal to the number of observations

- a point estimate of σ^2 can be denoted by s^2 as follows

$$s^2 = \frac{SSE}{n-(k+1)}; \tag{2.20}$$

- and a point estimate of σ can be denoted by s as follows

$$s = \sqrt{\frac{SSE}{n-(k+1)}}. \tag{2.21}$$

To assess the utility of a multiple linear regression model, a quantity called the multiple coefficient of determination, denoted by R^2, is often calculated. This coefficient is computed using the following formulas:

Total variation = $\sum(y_i - \bar{y})^2$;
Explained variation = $\sum(\hat{y} - \bar{y})^2$;
Unexplained variation = $\sum(y_i - \hat{y})^2$;
Total variation = Explained variation + Unexplained variation; and
The multiple coefficient of determination is then given by

$$R^2 = \frac{Explained\ variation}{Total\ variation}. \tag{2.22}$$

R^2 is the proportion of the total variation in the n-observed values of the dependent variable that is explained by the overall regression model.

The multiple correlation coefficient is denoted by $R = \sqrt{R^2}$.

Many analysts recommend the use of an adjusted multiple coefficient of determination to avoid overestimating the importance of the predictor variables. The adjusted multiple coefficient of determination, R^2_{adj}, is given as

$$R^2_{adj} = \left(R^2 - \frac{k}{n-1}\right)\left(\frac{n-1}{n-(k+1)}\right), \tag{2.23}$$

where R^2 is the multiple coefficient of determination, n is the number of observations, and k is the number of predictor variables in the model under consideration.

Another way to assess the utility of a regression model is to test the significance of the regression relationship between y and $x_1, x_2, \ldots, x_k$. This is called an F-test and is performed as follows:

Suppose that the regression assumptions hold and that the multiple linear regression model contains $(k+1)$ parameters; the test is

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0 \tag{2.24}$$

versus

$$H_1: \text{At least one of } \beta_1, \beta_2, \ldots, \beta_k \text{ does not equal } 0. \tag{2.25}$$

The overall F-statistic is defined to be

$$F(model) = \frac{Explained\ variation/k}{Unexplained\ variation/[n-(k+1)]}. \tag{2.26}$$

Also the p-value related to $F(model)$ is defined to be the area under the curve of the F-distribution (having k and $[n-(k+1)]$ degrees of freedom) to the right of $F(model)$. Then, H_0 is rejected in favour of H_1 at level of significance α if either of the following equivalent conditions holds:

1. $F(model) > F_{[\alpha]}$; or
2. p-value < α.

The point $F_{[\alpha]}$ is based on k numerator and $n-(k+1)$ denominator degrees of freedom.

In addition to the above techniques, it is also possible to construct confidence intervals for means and prediction intervals for individual values. A comprehensive discussion and technical details of these aspects can be found in Bowerman *et al.* (2005).

To conclude this section, it should be noted that the linear regression model and the use of the least squares (L_2-norm) technique have been studied for more than 200 years (Giloni & Padberg, 2002). The theory behind the model is highly developed, as shown in the above discussion, and goodness of fit, statistical properties and quality of the regression coefficients are some of the aspects that have been developed over the years. The next sections briefly look at the L_1- and L_∞-norm.

2.2.2 Least sum of absolute deviations regression (L_1-norm)

The least sum of absolute deviations method is an alternative technique to the least squares method to estimate regression parameters for a linear regression model. This method minimizes the sum of the absolute errors (or deviations), rather than the squared errors, as is the case with the least squares method.

The problem of minimizing the sum of absolute deviations can be handled, according to Gass (1958), as follows:

Let $x_{ij},\ i = 1, \dots, n$ and $j = 1, \dots, k+1$ denote a set of n observational measurements on k predictor variables. Let $y_i,\ i = 1, \dots, n$ denote the associated measurements on the dependent variable. The problem is to find the regression coefficients b_j such that

$$\sum_i \left| \sum_j b_j x_{ij} - y_i \right| \tag{2.27}$$

is minimized. This means that values must be found for the regression coefficients such that the sum of the absolute differences (2.28) is a minimum.

$$\left| \sum_j b_j x_{ij} - y_i \right| \tag{2.28}$$

Let

$$z_i' - z_i'' = y_i - \sum_j b_j x_{ij}, \tag{2.29}$$

$$z_i', z_i'' \geq 0. \tag{2.30}$$

Since the expression $y_i - \sum_j b_j x_{ij}$ for any set of b_j can be positive or negative, the difference can be represented as the difference of two nonnegative members. The problem can then be rewritten as follows:

$$\text{Minimize} \sum_i (z_i' + z_i'') \tag{2.31}$$

$$\text{subject to} \sum_j b_j x_{ij} + z_i' - z_i'' = y_i, \tag{2.32}$$

$$z_i', z_i'' \geq 0, \tag{2.33}$$

with the variables b_j being unrestricted in sign.

Since z_i' and z_i'' in a basic and feasible solution cannot both be positive, the optimum basic solution will select a set of b_j which minimizes the sum of the absolute differences.

Although the L_1-norm regression problem has been studied since the 18th century (Harter, 1974), the computational complexity of this technique was only overcome in the 1960s with the advent of modern computers. Little is known about the error distribution of this technique and the statistical theory for the L_1-norm regression problem is not as extensive as the L_2-norm regression problem, but Giloni and Padberg (2002) proved the unbiased nature of the L_1-norm estimators under certain assumptions. The dependence of the L_1-norm estimator on the errors is also more complicated than it is with the L_2-norm regression problem. In the last 50 years or so, a renewed interest in the L_1-norm regression problem has developed and more attention has been given to the abovementioned problems (Bassett & Koenker, 1978; Giloni & Padberg, 2002).

2.2.3 Chebychev regression (L_∞-norm)

The Chebychev regression technique uses polynomials in the process of approximating a function. The minimization of the maximum residual error, the minimax principal, is used to estimate parameters. The Chebychev problem can be described as follows (Gass, 1958):

Let x_{ij}, $i = 1, \dots, n$ and $j = 1, \dots, k+1$ denote a set of n observational measurements of k predictor variables. Let y_i, $i = 1, \dots, n$ denote the associated measurements of the dependent variable.

The Chebychev criterion is to find a set of coefficients b_j such that the following is true

$$\text{Minimize}\left\{\text{Maximum}\left|\sum_j b_j x_{ij} - y_i\right|\right\}. \tag{2.34}$$

This means that a set of b_j must be found such that the maximum deviation of the estimates of the y_i is a minimum.

Consider the constraints $z \geq \left|\sum_j b_j x_{ij} - y_i\right|$ for each i. The variable z is nonnegative and the aim is to have z as a minimum. This inequality in absolute terms can be rewritten for each i as two inequalities, in other words the value of $\sum_j b_j x_{ij} - y_i$ can lie between z and $-z$, or $-z \leq \sum_j b_j x_{ij} - y_i \leq z$.

The problem can now be stated as follows:

$$\text{Minimize} \quad \text{z} \tag{2.35}$$

$$\text{subject to} \sum_j b_j x_{ij} - y_i - z \leq 0, \tag{2.36}$$

$$\sum_j b_j x_{ij} - y_i + z \geq 0, \tag{2.37}$$

$$z \geq 0, \tag{2.38}$$

with the variables b_j being unrestricted in sign.

The objective function is non-differentiable and the unique nature of an optimal solution cannot be guaranteed. Although the L_∞-norm regression problem is the preferred method in cases where the sample midrange estimator of centrality is more effective than the sample mean or sample median, statistical literature on the L_∞-norm regression problem is scarce (Giloni & Padberg, 2002).

2.3 Outliers

Outlier detection is an important aspect of this study, and therefore this section will present a definition and overview of outliers, an explanation of their occurence, and why it is important to detect outliers and how to do so.

Outliers can be defined as observations that do not follow the same model as the rest of the data (Hoeting *et al*., 1996) or as data which are different from the majority (Ortiz *et al*., 2006). When an observation is removed from the data set and the features of the regression analysis (for example, point estimates of σ^2 and σ) change considerably, this observation is considered influential. According to Bowerman *et al.* (2005) an observation can be an outlier because of its y values or its x values or both, but an outlier is not necessarily influential even though it may be.

As stated by Kutner *et al*. (2005), outliers can create great difficulty in regression problems. When the least squares method is applied to data this difficulty can be explained particularly well: the sum of the squared deviations is minimized and the fitted line may be pulled toward the outlying observation in a disproportionate way. If this outlying observation is due to a mistake or irrelevant cause it could cause a misleading fit and explanation of the model. This problem might also influence predictions in such a way that they cannot be trusted.

The presence of outliers can be attributed to a variety of irregularities. Human error may influence the recording or transcription of data, the malfunction of measuring instruments might lead to measurement error and fraudulent behaviour or even natural deviation in populations could also be the cause of outliers.

With respect to the y values of outliers there exist several measures to detect outlying cases. In the case of simple linear regression it is sometimes possible to spot potential outliers through scatter plots, box plots and stem-and-leaf plots, but for multiple variables this may become a difficult task. For the detection of outliers in multiple linear regression, the following measures can be employed: residuals, studentized residuals, deleted- and studentized deleted residual and Cook's distance measure (Bowerman *et al.,* 2005). With respect to the x values, the leverage value of outliers can be used as a method of detection. In the rest of this section these measures will be discussed.

2.3.1 Leverage values

Bowerman *et al.* (2005) define the leverage value as a measure of the distance between the observation's x values and the centre of the experimental region. When this value is large, an observation is considered outlying with respect to its x values. When a leverage value is twice the average of all the leverage values, it is considered to be large.

2.3.2 Residuals and semistudentized residuals

To identify outliers with respect to their y values, residuals (2.39) or semistudentized residuals (2.40) may be considered. MSE denotes the mean square error (or residual) of the model. Any residual that is substantially different from the rest is suspect (Kutner *et al*., 2005).

$$e_i = y_i - \hat{y}_i \tag{2.39}$$

$$e_i^* = \frac{e_i}{\sqrt{MSE}} \tag{2.40}$$

To identify outliers with respect to their y values more effectively, some refinements to the analysis can be made. To do so it is necessary to introduce the hat matrix.

Let the vector of the fitted (or expected) values, $\hat{y}_i$, be denoted by $\widehat{\mathbf{Y}}$ and the vector of the residual terms $e_i = y_i - \hat{y}_i$ be denoted by $\boldsymbol{e}$. According to Kutner *et al*. (2005) the fitted values are represented by

$$\underset{n \times 1}{\widehat{\mathbf{Y}}} = \mathbf{Xb}, \tag{2.41}$$

and the residual terms by

$$\underset{n \times 1}{\boldsymbol{e}} = \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{Xb}. \tag{2.42}$$

The vector of the fitted values $\widehat{\mathbf{Y}}$ can be expressed in terms of the hat matrix as follows:

$$\underset{n \times 1}{\widehat{\mathbf{Y}}} = \mathbf{HY}, \tag{2.43}$$

where

$$\underset{n \times n}{\mathbf{H}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'. \tag{2.44}$$

The residuals e_i can also be represented as a linear combination of the y_i observations using the hat matrix

$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{Y}. \tag{2.45}$$

The variance-covariance matrix of the residuals is

$$\boldsymbol{\sigma^2}\{\mathbf{e}\} = \sigma^2(\mathbf{I} - \mathbf{H}), \tag{2.46}$$

and the variance of residual e_i, indicated by $\sigma^2\{e_i\}$, is

$$\sigma^2\{e_i\} = \sigma^2(1 - h_{ii}), \tag{2.47}$$

where h_{ii} is the ith element on the main diagonal of the hat matrix.

The covariance between residuals e_i and e_j $(i \neq j)$ is

$$\sigma\{e_i, e_j\} = \sigma^2(0 - h_{ij}) = -h_{ij}\sigma^2, \quad i \neq j, \tag{2.48}$$

where h_{ij} is the element in the ith row and jth column of the hat matrix. These variances and covariances are estimated by using MSE as the estimator of the error variance σ^2

$$s^2\{e_i\} = MSE(1 - h_{ii}), \tag{2.49}$$

$$s\{e_i, e_j\} = -h_{ij}(MSE), \quad i \neq j. \tag{2.50}$$

2.3.3 Studentized residuals

To improve the effectiveness of the identification of outliers with respect to their y values using residuals, it must be considered that the residuals e_i may have substantially different variances $\sigma^2\{e_i\}$. When the magnitude of each e_i relative to its estimated standard deviation is considered, the differences in the sampling errors of the residuals are recognized. Kutner *et al*. (2005) derive the standard deviation of e_i from (2.49) as

$$s\{e_i\} = \sqrt{MSE(1 - h_{ii})}. \tag{2.51}$$

The ratio of e_i to $s\{e_i\}$ is called the studentized residual, denoted by r_i

$$r_i = \frac{e_i}{s\{e_i\}}. \tag{2.52}$$

The studentized residuals r_i have constant variance when the model is appropriate.

2.3.4 Omitted data points and residuals

Another improvement upon residuals, to more effectively identify outliers with respect to their y values, is to determine the ith residual $e_i = y_i - \hat{y}_i$ when the fitted regression uses all the data points except the ith one (Kutner *et al*., 2005). The reason for this improvement is that if observation i is an outlier with respect to its y value and it is included in the computation of the least squares point estimates the point prediction $\hat{y}_i$ might be "drawn" towards y_i causing the resulting residual to be small. On the other hand, if the ith observation is excluded before the least squares point estimates are calculated the point prediction $\hat{y}_i$ is not influenced by the ith observation. This will cause the resulting residual to be larger, and therefore more likely to disclose the outlying observation with respect to its y value.

This improvement can be made by deleting the ith case and fitting the regression function to the rest of the data. Thus the estimate of the expected value for the ith case $\hat{y}_{i(i)}$ can be determined. The deleted residual for the ith case d_i, is the difference between the observed value y_i and the estimated expected value $\hat{y}_{i(i)}$

$$d_i = y_i - \hat{y}_{i(i)}. \tag{2.53}$$

The following expression can be used without recalculating the regression function for each ith observation that is omitted (Kutner *et al*., 2005)

$$d_i = \frac{e_i}{1 - h_{ii}}, \tag{2.54}$$

where e_i is the usual residual for the ith case and h_{ii} is the ith diagonal element in the hat matrix.

Deleted residuals will sometimes reveal outlying observations with respect to their y values when ordinary residuals would not have revealed them.

2.3.5 Studentized deleted residuals

The improvements in section 2.3.3 and 2.3.4 can be combined, utilizing the deleted residual, d_i, in (2.54) and studentize it by dividing it by its estimated standard deviation. This results in the studentized deleted residual, denoted by t_i

$$\begin{aligned} t_i &= \frac{d_i}{s\{d_i\}} \\ &= \frac{d_i}{\sqrt{MSE_{(i)}(1 - h_{ii})}}. \end{aligned} \tag{2.55}$$

According to Kutner *et al.* (2005) a simple relationship between MSE and $MSE_{(i)}$ can be used to express the studentized deleted residuals, t_i, in terms of the residuals e_i, the error sum of squares, SSE, and the hat matrix values h_{ii} for all n observations. This will result in the equivalent expression for t_i

$$t_i = e_i \left[\frac{n - p - 1}{SSE(1 - h_{ii}) - e_i^2}\right]^{1/2}. \tag{2.56}$$

This expression can be calculated without having to fit new regression functions each time a different observation is omitted.

2.3.6 Cook's distance measure

Following the identification of outliers with respect to their y values and/or their x values, the next step is to determine whether the observations are influential. As noted earlier, an observation is regarded as influential if its exclusion causes major changes in the features of the regression analysis.

Cook's distance measure, denoted by D_i, can be used to determine whether an observation is influential or not. When D_i is large, classifying observation i as influential, it indicates that there is a substantial difference in the least squares point estimates calculated by using all n observations and the least squares point estimates calculated by using all n observations except for observation i. Cook's distance measure can be described as follows (Kutner *et al*., 2005):

$$D_i = \frac{\sum_{j=1}^{n}\left(\hat{y}_j - \hat{y}_{j(i)}\right)^2}{(k+1)MSE}, \tag{2.57}$$

where k denotes the number of variables in the model, $k+1$ denotes the number of parameters to be estimated.

According to Bowerman *et al.* (2005) D_i can be classified as large when it is compared to two F-distribution points – the 20th percentile of the F-distribution, $F_{[0.80]}$, and the 50th percentile of the F-distribution, $F_{[0.50]}$ – based on $(k+1)$ numerator and $[n-(k+1)]$ denominator degrees of freedom. The ith observation exerts little apparent influence and should not be considered influentail if D_i is less than $F_{[0.80]}$. On the other hand, if D_i is close to or greater than $F_{[0.50]}$ the ith observation could be considered influential.

D_i can be expressed in terms of the residuals e_i, the mean error sum of squares, MSE, and the hat matrix values h_{ii} for all n observations (Kutner *et al*., 2005)

$$D_i = \frac{e_i^2}{(k+1)MSE}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right]. \tag{2.58}$$

This is useful because the least squares point estimate does not have to be recalculated each time an observation is deleted.

2.3.7 Treatment of outlying and influential observations

Once outliers with respect to their y values and/or their x values have been identified and classified as influential or not, Bowerman *et al.* (2005) suggest dealing with outliers in terms of their y values first, because other problems will often diminish or disappear. According to Bowerman *et al.* (2005), there could be several reasons for the presence of outliers; each case should be evaluated to decide what should be done with the outliers.

When dealing with outliers the first step is to check if the y value was correctly recorded: if this is not the case, the value should be corrected and the regression rerun. If it is not possible to correct the value, the observation should be discarded and the regression should be rerun

again. If the presence of the outlier(s) is not due to incorrect recording, other possible reasons should be investigated.

Sometimes the y value is caused by an effect that the regression model is not required to describe, such as a natural disaster. If this is the case, the observation can be discarded. Outliers can also occur because of inefficiency, for example, when the profit of one of ten similar businesses is significantly lower than the rest. Investigation may show that this is due to a manager who lacks basic business skills. This might possibly be corrected by training, but the observation should be removed from the data set, because the model should not be based on data from an inefficient source. Another explanation for the presence of outliers could be that the predictor variable, which would explain the seemingly large value of y, is not included in the model. This could be rectified by the re-evaluation of the predictor variables which are included in the model.

Section 2.3 described diagnostic measures based on the deletion of single observations, which are useful to identify outliers and influential observations in regression analysis. According to Rousseeuw and van Zomeren (1990) it is more difficult to detect multiple outliers, especially when more than two predictor variables are included in a model, because the data cannot be visually presented and evaluated. Classical diagnostic measures do not detect the outliers either, because the bases of these measures, the sample mean and covariance matrix, are also influenced by the outliers. In this way the outliers become masked.

Deleting one outlying observation at a time, when multiple outliers are present in the data, may prove to be inefficient and incorrect because accurate observations could inadvertently be deleted when real outliers have been masked. In the following section the robustness of a model will be addressed.

2.4 Robustness of a model

As mentioned earlier, outliers are observations which are different from the majority of the data which has been collected. This can cause great difficulty in regression analysis because such irregularities may distort the least squares point estimates, causing the incorrect prediction and interpretation of the model. Regression analysis cannot explain a model accurately unless all of the outliers can be deleted beforehand. Usually, not all of the outliers can be deleted in advance because they are often masked. Therefore another approach is needed to deal with multiple outlying observations.

According to Rousseeuw and Leroy (2003) robust regression techniques can be defined as methods that try to devise estimators that are not strongly affected by outliers. Therefore the

results or the estimators remain reasonably stable and reliable even in the presence of multiple outlying observations. In contrast to ordinary regression analysis, which detects and deletes outliers before the model is developed, robust regression techniques first develop a model which explains the bulk of the data. After this model has been developed, the outlying observations are identified by their residuals.

Two approaches to improve the robustness of a model will be discussed in the following sections. The first approach is used to perform residual analysis while the second is to use more robust methods.

2.4.1 Residual analysis

Direct diagnostic plots for the dependent variable are often not useful in regression analysis because the values of the observations of the dependent variable are a function of the level of the predictor variable(s). Indirect diagnostics for the dependent variable can be made by examining the residuals. The assumptions of the error terms are stated in section 2.2; that is, the error terms are assumed to be independent normal random variables, with a mean of zero and constant variance, σ^2.

According to Kutner *et al.* (2005) some important deviations from these assumptions can be noticed by examining the residuals (denoted by e). These include the regression function not being linear, the error terms not having a constant variance, the error terms not being independent, the model fitting all but one or a few outlying observations and the error terms not being normally distributed.

In the case of a residual plot against the predictor variable, when the residuals fall within a horizontal band centred around zero, a linear regression model seems appropriate (see figure 2.1). Figure 2.2 depicts a situation in which a linear regression function is not appropriate and a curvilinear function is more so. Plots of the residuals against the predictor variable(s) are not only helpful to study whether a linear regression function is appropriate or not, but also to examine whether the variance of the error terms is constant. Figure 2.1 displays a constant variance, while figure 2.3 shows the nonconstancy of the error variance. The error variance increases with x in a megaphone type of manner. The nonindependence of the error terms over time is displayed in figure 2.4 while residual outliers can be identified from residual plots as indicated in figure 2.5.

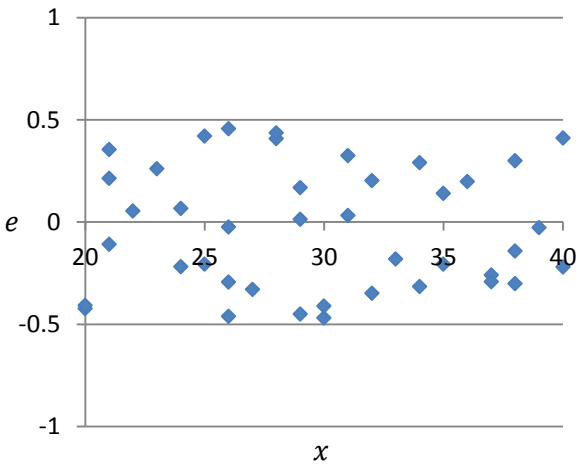

Figure 2.1 – Linearity assumption seems appropriate

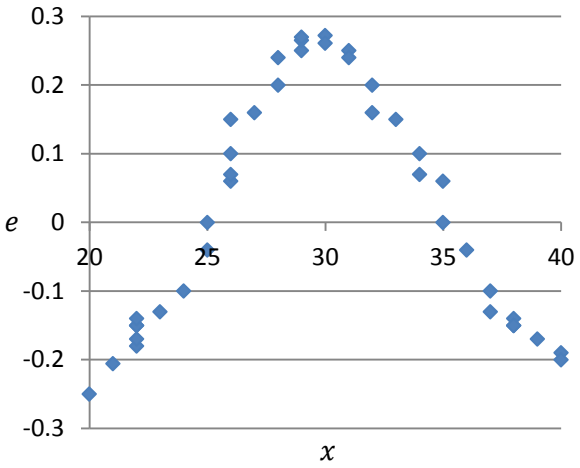

Figure 2.2 – Linearity assumption not appropriate

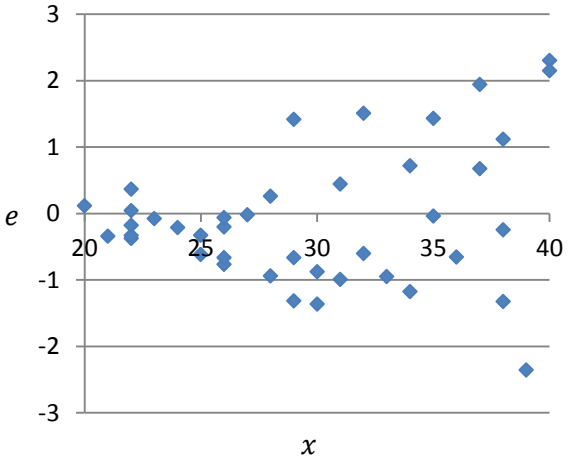

Figure 2.3 – Nonconstant variance of the error terms

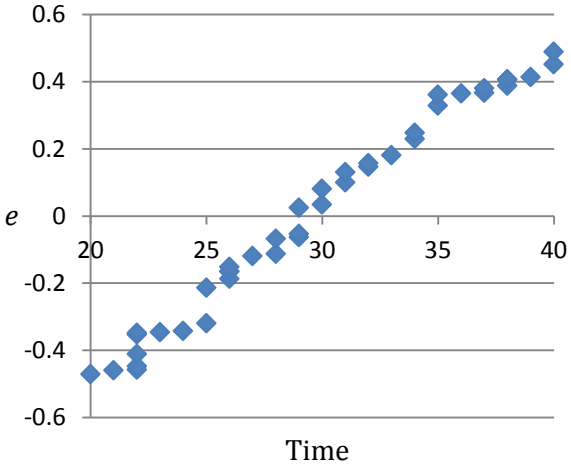

Figure 2.4 – Nonindependence of the error terms

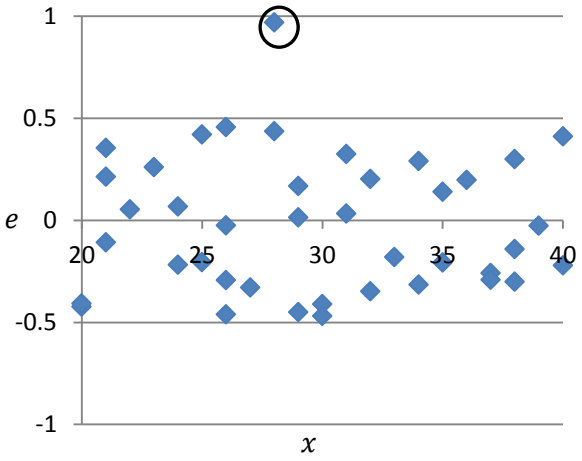

Figure 2.5 – Residual outlier identified

2.4.2 Robust methods

To measure the effectiveness of different robust estimators, the number of outliers that the estimators can deal with can be compared, for example, how many outliers can be present in

the data before an estimator breaks down (when the bulk of the data can no longer be explained). Thus, the breakdown points can be compared. Although low breakdown points are desirable attributes for a method, it must be noted that this alone is not sufficient.

Rousseeuw and Leroy (2003) show that one outlier can cause the least squares regression method to break down. For a sample size of n, its breakdown is $1/n$ which tends to 0% when n increases. The breakdown point for least absolute deviation regression is also 0%, because, although this method is more robust regarding outlying observations with respect to their y values, one influential leverage value (an outlying observation with respect to its x value) may cause the method to break down.

Two high-breakdown regression methods are introduced by Rousseeuw and Leroy (2003): the least median of squares and the least trimmed squares; these will be briefly described in the following two subsections.

2.4.2.1 Least median squares regression

By replacing the summation sign, $\sum$, of the least sum of squares by the median, which is very robust, Rousseeuw and Leroy (2003) proposed the least median of squares, which is given by

$$\underset{i}{\text{Minimize median}}\,(y_i - \hat{y}_i)^2. \tag{2.59}$$

The technical details of this method are described by Rousseeuw and Leroy (2003) who show that the breakdown point of this method is 50%, this being very good. This is the maximum value for a breakdown point because if more than 50% of the observations are outliers it is not possible to detect the 'correct' part of the sample anymore.

2.4.2.2 Least trimmed squares regression

Let $r = y_i - \hat{y}_i$, then the least trimmed squares regression can be formulated as

$$\text{Minimize} \sum_{i=1}^{h} (r^2)_{i:n}, \tag{2.60}$$

where the residuals are first squared and then ordered, $(r^2)_{1:n} \leq \cdots \leq (r^2)_{n:n}$, and h is the number of observations not trimmed from the model.

As a result the largest squared residuals are not used in the summation and a breakdown point of 50% can be achieved. The properties of this estimator are considered in Rousseeuw and Leroy (2003).

In chapter 4 another robust outlier detection method will be introduced to assist in the development of a robust regression model with minimal assumptions. In the next section the subject of linear programming will be broached.

2.5 Linear programming

Managers often have to make decisions regarding the production or quantities of different products with different profit margins, bearing in mind the available resources such as labour, materials, time and money. This and many other problems that are accompanied by their own intricacies regarding the most effective use of available resources can be solved by a widely used mathematical modelling technique called linear programming.

The objective function of any linear programming problem is to minimize or maximize a certain quantity, such as profit or cost. Another requirement for linear programming problems is the presence of constraints which limit the extent to which the problem can be minimized or maximized; for example having a limited amount of money available for marketing, or, a machine only being able to produce a limited quantity of items per hour. Therefore, a linear programming problem can be defined as a model consisting of linear relationships representing a decision, or decisions with objectives and resource constraints. The general mathematical representation of such a model can be defined as follows (Moore & Weatherford, 2001):

$$\text{Maximize (or minimize) } f(x_1, \dots, x_n) \tag{2.61}$$

$$\text{subject to the constraints}$$

$$g_1(x_1, \dots, x_n) \begin{matrix} \leq \\ = \\ \geq \end{matrix} b_1, \tag{2.62}$$

$$\vdots$$

$$g_m(x_1, \dots, x_n) \begin{matrix} \leq \\ = \\ \geq \end{matrix} b_m. \tag{2.63}$$

Although there are different types and extensions of general linear programming problems, according to Bazaraa *et al*. (2005) all of these variations can be manipulated into the following form of linear programming problem

$$
\begin{aligned}
\text{Minimize} \quad & c_1x_1 + c_2x_2 + \cdots + c_nx_n && (2.64)\\
\text{subject to} \quad & a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \geq b_1, && (2.65)\\
& a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \geq b_2, && (2.66)\\
& \vdots \quad + \quad \vdots \quad + \cdots + \quad \vdots \quad \quad \vdots && \\
& a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \geq b_m, && (2.67)\\
& x_1, \quad x_2, \quad x_n \geq 0, && (2.68)
\end{aligned}
$$

where $c_1x_1 + c_2x_2 + \cdots + c_nx_n$ is the objective function to be minimized. The $c_1, c_2, \cdots, c_n$ coefficients are the (known) cost coefficients while $x_1, x_2, \cdots, x_n$ are the (unknown) decision variables. The inequality $\sum_{j=1}^{n} a_{ij}x_j \geq b_i$ denotes the ith constraint and the right-hand-side vector is represented by $b_1, b_2, \cdots, b_m$.

When the decision variables do not take on negative values, a non-negativity constraint, $x_1, x_2, \cdots, x_n \geq 0$, is added to the formulation. A feasible solution is obtained when a set of values for the variables $x_1, x_2, \cdots, x_n$ satisfies all of the constraints. Thus, a linear programming problem aims to find, among all feasible solutions, the one that minimizes (or maximizes) the objective function.

For ease of illustration, the linear program can be formulated in matrix notation. The row vector $(c_1, c_2, \cdots, c_n)$ can be denoted by $\mathbf{c}$. The column vectors $\mathbf{x}$ and $\mathbf{b}$ and the $m \times n$ matrix $\mathbf{A}$ can be denoted by

$$
\underset{n \times 1}{\mathbf{x}} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \qquad \underset{m \times 1}{\mathbf{b}} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}, \qquad \underset{m \times n}{\mathbf{A}} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}. \qquad (2.69)
$$

The problem can now be written as

$$
\begin{aligned}
\text{Minimize} \quad & \mathbf{cx} && (2.70)\\
\text{subject to} \quad & \mathbf{Ax} \geq \mathbf{b}, && (2.71)\\
& \mathbf{x} \geq \mathbf{0}. && (2.72)
\end{aligned}
$$

Every model employs several assumptions. When the model is used it is important to take note of the assumptions and make sure that it can endure the given situation. The inherent assumptions of linear programming are given below (Bazaraa *et al*., 2005):

- *Proportionality*. If the value of x_j doubles, the contribution of x_j to cost c_jx_j doubles and the contribution of x_j to the ith constraint $a_{ij}x_j$ also doubles. No savings are realized through the usage of more of x_j;

- *Additivity*. The sum of the individual costs forms the total cost while the total contribution to the ith restriction is the sum of the individual contributions of the individual activities. There are no interaction effects among the activities;
- *Divisibility.* Non-integer values for the decision variables are permitted such that decision variables with fractional levels can be interpreted; and
- *Deterministic*. The coefficients c_j, a_{ij} and b_j are known deterministically and are approximations of any probabilistic or stochastic elements.

Although these assumptions seem restrictive, linear programming certainly helps to solve a very wide range of problems. By adjusting the program it can often be used to approximate nonlinear problems and help solve linear problems with integer restrictions on some or all of the variables.

There are different methods to solve linear programming problems. A problem with two decision variables can be solved by using graphical methods or, for larger problems, the simplex method can be employed. These methods are explained and illustrated in Appendix A, sections A.2 to A.4.

2.6 Integer programming

One of the assumptions of linear programming, mentioned earlier in section 2.5, is that of divisibility, which means that non-integer values for the decision variables are permitted. However, a large amount of problems can only be solved if the variables have integer values: for example, a company cannot hire 2.33 labourers or purchase 3.88 machines; the values must be exactly 2, 3, 4 or another integer amount.

Integer linear programming models possess the same constraint and objective functions as ordinary linear programming models and they are also formulated in the same way; the only difference is that there are one or more predictor variables that have to take on integer values in the final solution. There are cases, however, in which all of the predictor variables are required to have integer values; these problems are pure integer linear programming problems. When some, but not all, of the predictor variables are required to take on integer values, this is called a mixed integer linear programming problem. Sometimes all the predictor variables must have values of either 0 or 1; this is termed a zero-one integer linear programming problem.

According to Salkin and Mathur (1989) a mixed integer linear program can be written in the following way

$$\text{Minimize} \quad \mathbf{cx} + \mathbf{dy} \tag{2.73}$$

$$\text{subject to} \quad \mathbf{Ax} + \mathbf{Dy} \leq \mathbf{b}, \tag{2.74}$$

$$\mathbf{x} \geq \mathbf{0}, \boldsymbol{y} \geq \mathbf{0}, \tag{2.75}$$

$$\mathbf{x} \text{ integer}, \tag{2.76}$$

where

c is an n row vector;

d is an n' row vector;

A is an m by n matrix;

D is an m by n' matrix;

b is an m column vector of constants (the right-hand side);

x is an n vector of integer variables; and

y is an n' vector of continuous variables.

When $n' = 0$, the continuous variable **y** disappears and an integer program is left. If $n = 0$, there are no integer variables **x** and the problem reduces to a linear program.

Many mathematical programs can be converted to problems with integer variables. For example, suppose a variable z_j is allowed to take only one of several values, say $v_{11}, v_{12}, \ldots, v_{1n}$. This is equivalent to setting

$$z_j = x_1 v_{11} + x_2 v_{12} + \cdots + x_n v_{1n}, \tag{2.77}$$

with

$$x_1 + x_2 + \cdots + x_n = 1, \tag{2.78}$$

and

$$x_j = 0 \text{ or } 1, \qquad j = 1,2, \ldots, n. \tag{2.79}$$

To solve an integer linear programming problem is much more difficult than solving a linear programming problem. If the predictor variables take on fractional values in the solution of a linear programming problem, the simplest approach would be to round the values off, but this approach produces two problems. Firstly, the new integer solution may be outside of the feasible region and thus not a viable solution, and secondly, even if the rounded values result in a feasible solution it may not be the optimal feasible one.

Salkin and Mathur (1989) state that the principal approaches for solving mixed integer (or integer) programs are cutting plane techniques, enumerative methods, partitioning algorithms and group theoretic approaches.

The general intent of cutting plane algorithms is to deduce supplementary inequalities or "cuts" from the integrality and constraint requirements which, when added to the existing constraints, eventually produce a linear program whose optimal solution is an integer in the integer constrained variables.

The basic approach for the integer program involves the following steps:

Step 1: Starting with an all-integer tableau, solve the integer program as a linear one. If it is infeasible, so is the integer problem and thus one must terminate the problem. If the optimal solution is all-integer, the integer program is solved and thus one must again terminate the problem. If none of this step applies, go to Step 2.

Step 2: Derive a new inequality constraint (or "cut") from the integrality and other current constraint requirements which "cuts off" the (current) optimal point but does not eliminate any integer solution. Add the new inequality to the bottom of the simplex tableau which then exhibits primal infeasibility. Go to Step 3.

Step 3: Reoptimize the new linear program using the dual simplex method. If the new linear program is infeasible, the integer problem has no solution and the problem must be terminated. If the new optimal solution is in integers, the integer program is solved and the problem must be terminated. If this does not apply, go to Step 2.

The Beale tableau and Gomory cut is often used to solve mixed integer (or integer) problems in this manner. For a detailed explanation see Salkin and Mathur (1989).

The aim of enumerative methods is to enumerate, either explicitly or implicitly, all possible solution candidates to the mixed integer (or integer) program. The feasible solution which maximizes the objective function is optimal.

To solve the mixed integer (or integer) problem explicitly, one must list all of the feasible solutions and compute the objective value for each solution; the solution with the best objective function is the optimal solution. This method is applicable to small data sets, but is daunting and often impossible to apply to larger data sets.

Another enumerative method is the well known branch-and-bound method. This is an implicit enumerative method. Branching only takes place on variables that are required to take on integer values; the feasible region is divided and subproblems are formed and solved.

Bounding is used to develop bounds for the different subproblems. By comparing the objective values (or bounds) of the subproblems it is possible to eliminate certain subproblems from consideration (thus, certain feasible solutions cannot improve the current solution and do not have to be investigated further; these points are enumerated implicitly). Dakin's variation (Salkin & Mathur, 1989) of the branch-and-bound method is explained and illustrated in Appendix A, section A.5.

A comprehensive discussion and the technical details of partitioning algorithms and group theoretic algorithms can be found in Salkin and Mathur (1989).

2.7 Chapter summary

The aim of this chapter was to provide a sufficient background to, and gain a good understanding of, techniques and concepts that will be used in the subsequent chapters. An introductory overview of the concepts of linear regression models and the three associated techniques used to estimate regression parameters, the least squares (L_2-norm), least sum of absolute deviation (L_1-norm) and Chebychev (L_∞-norm) methods, were presented. This was followed by a discussion regarding outliers, outlier detection and robust regression methods. The chapter was concluded with the basic theory of linear programming models.

Chapter 3 will furnish a description of the specific linear programming model which forms the basis of this research study, followed by an example illustrating the model's application.

Chapter 3

A minimal assumption regression model

3.1 Introduction

In the previous chapter the basic concepts of linear regression models, outliers and linear programming were discussed. The aim of this chapter is to introduce the minimal assumption regression model that was used as a basis for this research study. The use of linear programming techniques, to solve least absolute deviation regression problems, will briefly be presented. This will be followed by an explanation and illustrative example of the minimal assumption regression model. The chapter will then be concluded with a brief literature review of other researchers who have referred to or used the minimal assumption regression model.

3.2 Absolute value regression using a linear programming technique

Certain problems involving absolute value terms can be transformed into a standard linear programming formulation. The absolute deviation (L_1-norm) technique for estimating regression parameters plays a central role in this study and has already been discussed in chapter 2, section 2.2.2. For this reason, the problem of minimizing the sum of absolute deviations is briefly recapitulated here.

Wagner (1959) supposes a set of n observational measurements of k predictor variables $x_{ij},\ i = 1,2,\dots,n$ and $j = 1,2,\dots k$ and dependent variables $y_i,\ i = 1,2,\dots n$ is given. Find the regression coefficients b_j that will

$$\underset{b_j}{\text{Minimize}} \sum_i \left| \sum_j b_j x_{ij} - y_i \right|. \tag{3.1}$$

As explained in chapter 2, section 2.2.3, this problem can be transformed and reduced to

$$\text{Minimize} \sum_i \varepsilon_{1i} + \sum_i \varepsilon_{2i} \tag{3.2}$$

$$\text{subject to} \sum_j b_j x_{ij} + \varepsilon_{1i} - \varepsilon_{2i} = y_i, \tag{3.3}$$

for $i = 1,2,\dots,n$, b_j unrestricted in sign and ε_{1i} and ε_{2i} nonnegative.

The variables ε_{1i} and ε_{2i} can be interpreted as vertical deviations "above" or "below" the fitted plane for the ith observation. The absolute difference between the estimate $\sum_j b_j x_{ij}$ and y_i is given by $\varepsilon_{1i} + \varepsilon_{2i}$ in an optimal solution. From linear programming theory it is known that ε_{1i} and ε_{2i} cannot both be strictly positive in an optimal solution.

3.3 A minimal assumption regression model

During 1962, Harvey M. Wagner published a linear programming model that provides a fit for regression functions according to the criteria of minimal sum of absolute deviations but without specifying a mathematical form for the functions to be estimated (Wagner, 1962). The only restrictive assumption needed is one of monotonicity of the functions, that is, the regression functions are assumed to be monotonically non-increasing or non-decreasing. These are the only assumptions that have to be made and in this sense the model employs minimal assumptions.

The model entails the following:

Using Wagner's notation, assume an additive regression model of the form

$$y = \sum_{j=1}^{k} f_j(x_j) + error \tag{3.4}$$

is applicable with y the dependent variable and $x_{j,}\ j = 1, 2, \cdots, k$, the predictor variables. Assume that n observations on the variables y and x_j are available given by $(y_i, x_{i1}, x_{i2}, \ldots, x_{ik})$ for $i = 1, 2, \cdots, n$. Wagner's model now aims to determine estimators of function values $f_j(x_{ij})$, which are abbreviated as f_{ij}, from this data, such that estimates $\hat{y}_i = \sum_j f_j(x_{ij})$ of the response are optimal in the L_1-norm sense.

Each function f_j need not be linear and no mathematical form needs to be specified. Wagner categorized this model as curvilinear regression. The moderate restrictions applicable to the behaviour of the functions are restrictions of monotonicity. Thus, a given function f_j must be monotonically non-increasing or non-decreasing.

Wagner argued that there are a number of situations in which it is difficult to *a priori* specify a mathematical form for the function f_j, and where it appears suitable to require only mild restrictions on the functions. An example from an economic viewpoint is that of diminishing marginal productivity or return. That is, after a certain point, each extra unit of variable input (for example, man-hours) produces smaller increases in outputs, and therefore reduce each

worker's mean productivity. In this case, the form of the function f_j is not known exactly. What is known, however, is that f_j will probably be monotonically non-decreasing.

Linear programming methods are used to estimate the function variables, f_{ij}, using only minimal assumptions to constrain the required shape. The least sum of absolute deviation regression (L_1-norm) is used to estimate the parameters.

Starting with a simple special case, the fundamental nature of the model will be explained. In doing so, complex notation is avoided and the important aspects of the model are highlighted. The additive constraints can be formulated as follows:

$$\sum_{j=1}^{k} f_{ij} + \varepsilon_{1i} - \varepsilon_{2i} = y_i, \tag{3.5}$$

$$\varepsilon_{1i}, \varepsilon_{2i} \geq 0, \tag{3.6}$$

for $i = 1, 2, \cdots, n$.

Monotonically non-increasing or non-decreasing constraints are imposed on the functions f_j. For illustrative purposes, Wagner assumed that the observations of variable x_j are sorted as follows: $x_{1j} \leq x_{2j} \leq x_{3j} \leq \cdots \leq x_{nj}$. The constraints in the case of non-decreasing functions are

$$f_{ij} \leq f_{i+1,j}, \tag{3.7}$$

for $i = 1,2, \ldots, n-1$ and $j = 1,2, \ldots, k$. To constrain a monotonically non-increasing function the inequalities are reversed.

In the case of a more general approach, the values x_{ij} for each j need not be distinct and are not necessarily ordered. If there are values of x_{ij} that are identical, the corresponding relevant functions variables must also have the same values: that is, if $x_{ij} = x_{kj}$, then $f_{ij} = f_{kj}$ when $i \neq k$. To simplify the inequality constraints, the x_{ij} values are ranked. A dense ranking function is defined wherein $r_{1j} = rank(x_{1j})$ denotes the rank for each value of the x_j variables. In other words, when the variables are sorted, equal x_{ij} values receive the same ranking number and the following x_{ij} value receives the ranking number that immediately follows it. The ranking can be done in increasing or decreasing order, depending on the specified monotonicity. If the function is non-decreasing, a non-decreasing ranking order will be used, on the other hand, a non-increasing ranking order will be used if the function is non-increasing.

For a given j a monotonically non-decreasing function constraint can be rewritten in the following way, using the rank values

$$f_{tj} \leq f_{lj} \quad \text{if } r_{tj} \leq r_{lj}, \tag{3.8}$$

and

$$f_{tj} = f_{lj} \quad \text{if } r_{tj} = r_{lj}, \tag{3.9}$$

for $t, l = 1,2, \ldots, n$ with $t \neq l$.

A constraint for a monotonically non-increasing function can be created by reversing the inequality relations.

The objective function for the minimal assumption regression model is to find values for f_{ij} that will

$$\text{Minimize} \left[\sum_{i=1}^{n} (\varepsilon_{1i} + \varepsilon_{2i}) \right], \tag{3.10}$$

subject to the abovementioned linear constraints.

Even when the number of variables is high the method stays feasible because current hardware and software are powerful enough to solve large linear programs. When the model is solved, the values of the function variables, f_{ij}, can be used as they are, or they can be plotted against x_{ij} to investigate the mathematical form of each f_j function. To estimate the parameters for the mathematical form a least squares or least absolute deviation method can be followed.

Below is the formulation of the minimal assumption regression model as it is used in this study

$$\text{Minimize} \quad \sum_{i=1}^{n} (\varepsilon_{1i} + \varepsilon_{2i}) \tag{3.11}$$

$$\text{subject to} \quad \sum_{j=1}^{k} f_{ij} + \varepsilon_{1i} - \varepsilon_{2i} = y_i, \text{for } i = 1, \cdots, n, \tag{3.12}$$

$$f_{tj} \leq f_{lj}, \quad \text{if } r_{tj} \leq r_{lj}, \text{and} \tag{3.13}$$

$$f_{tj} = f_{lj}, \quad \text{if } r_{tj} = r_{lj}, \text{for } t, l = 1,2, \ldots, n \text{ with } t \neq l \text{ and } j = 1,2, \ldots, k, \tag{3.14}$$

$$\varepsilon_{1i}, \varepsilon_{2i} \geq 0, \quad \text{for } i = 1, \cdots, n, \tag{3.15}$$

where f_{ij} is unrestricted in sign for all i and j.

(Note that not all constraints in (3.13) are necessary when the model is implemented since it need only be considered when $r_{lj} = r_{tj} + 1$.)

Although the minimal assumption regression model is designed to make the minimum assumptions, it is still necessary to decide the direction of monotonicity. One way to approach this problem is by performing a multiple regression beforehand and using the signs of the estimated coefficients to determine whether a function should be restricted to be non-increasing or non-decreasing.

3.4 Illustrative example

In this section a data set from the literature will be used to illustrate how the model can be applied. The data set chosen is called the delivery time data set (Montgomery & Peck, 1992) and consists of 25 observations. Suppose that several outlets have vending machines that should be serviced at regular intervals. The time spent at each outlet, called the delivery time, is measured in minutes and depends on the number of vending machines that need to be serviced and stocked in that outlet and the distance walked there, which is measured in feet. Table 3.1 shows the data set.

The delivery time is the dependent variable, y, while the predictor variables are:

x_1: number of products; and
x_2: distance.

i	1	2	3	4	5	6	7	8	9	10
y_i	16.68	11.50	12.03	14.88	13.75	18.11	8.00	17.83	79.24	21.50
x_{i1}	7	3	3	4	6	7	2	7	30	5
x_{i2}	560	220	340	80	150	330	110	210	1460	605

i	11	12	13	14	15	16	17	18	19	20
y_i	40.33	21.00	13.50	19.75	24.00	29.00	15.35	19.00	9.50	35.10
x_{i1}	16	10	4	6	9	10	6	7	3	17
x_{i2}	688	215	255	462	448	776	200	132	36	770

i	21	22	23	24	25
y_i	17.90	52.32	18.75	19.83	10.75
x_{i1}	10	26	9	8	4
x_{i2}	140	810	450	635	150

Table 3.1 – Delivery time data set (Montgomery & Peck, 1992)

The minimal assumption regression model will now be applied to the delivery time data set.

The application consists of the following steps:

- determine monotonicity for each variable;
- rank the variables;
- create inequality constraints;
- formulate the model; and
- obtain the model solution.

Each of these steps will be discussed in the forthcoming sections.

3.4.1 Determining monotonicity

Before the model can be applied to the data, the direction of monotonicity for each variable must be estimated. To determine this, a multiple regression was done using Microsoft Excel 2007. The results are given in table 3.2. The coefficients of x_1 and x_2 are positive; therefore both f_1 and f_2 will be constrained as monotonically non-decreasing functions.

	Coefficients
Intercept	2.341
x_1	1.616
x_2	0.014

Table 3.2 – Multiple regression coefficients

3.4.2 Assign ranks and create inequality constraints

The next step is to assign a rank to each x_{ij}, wherein i denotes the data point and j denotes the predictor variable, and also to create the inequality constraints. Table 3.3 indicates the delivery time data set with assigned ranks and associated function variables. The first column, i, indicates the data point, the second column contains the y values while the next three columns are the x_1 variable values, the assigned rank values and the associated function variables, f_{i1}. The last three columns are the x_2 variable values, the assigned rank values and the associated function variables, f_{i2}.

For this example, both variables are monotonically non-decreasing functions, and thus the first ranking value, 1, will be assigned to the smallest x_{ij} value for each variable; therefore $x_{7,1}$ and $x_{19,2}$ both have ranks with the value of 1. The values of variables $x_{2,1}$ and $x_{3,1}$ are equal and therefore both have rank values of 2. By using the associated function variables, the inequality constraint for variables $x_{7,1}$, $x_{2,1}$ and $x_{3,1}$ can be set up as follows: the rank of $x_{7,1}$ is less than the rank of $x_{2,1}$, and therefore $f_{7,1} \leq f_{2,1}$. The rank of $x_{2,1}$ is equal to the rank of $x_{3,1}$ and therefore the function variables are also equal, $f_{2,1} = f_{3,1}$. This can be done for all the function

variables, using the rank to impose the inequality constraints. (Note that a comma is used to separate i and j and therefore $x_{23,1}$ is the 23rd observation of the first variable.)

i	y_i	x_{i1}	r_{i1}	f_{i1}	x_{i2}	r_{i2}	f_{i2}
1	16.68	7	6	$f_{1,1}$	560	17	$f_{1,2}$
2	11.50	3	2	$f_{2,1}$	220	10	$f_{2,2}$
3	12.03	3	2	$f_{3,1}$	340	13	$f_{3,2}$
4	14.88	4	3	$f_{4,1}$	80	2	$f_{4,2}$
5	13.75	6	5	$f_{5,1}$	150	6	$f_{5,2}$
6	18.11	7	6	$f_{6,1}$	330	12	$f_{6,2}$
7	8.00	2	1	$f_{7,1}$	110	3	$f_{7,2}$
8	17.83	7	6	$f_{8,1}$	210	8	$f_{8,2}$
9	79.24	30	13	$f_{9,1}$	1460	24	$f_{9,2}$
10	21.50	5	4	$f_{10,1}$	605	18	$f_{10,2}$
11	40.33	16	10	$f_{11,1}$	688	20	$f_{11,2}$
12	21.00	10	9	$f_{12,1}$	215	9	$f_{12,2}$
13	13.50	4	3	$f_{13,1}$	255	12	$f_{13,2}$
14	19.75	6	5	$f_{14,1}$	462	16	$f_{14,2}$
15	24.00	9	8	$f_{15,1}$	448	14	$f_{15,2}$
16	29.00	10	9	$f_{16,1}$	776	22	$f_{16,2}$
17	15.35	6	5	$f_{17,1}$	200	7	$f_{17,2}$
18	19.00	7	6	$f_{18,1}$	132	4	$f_{18,2}$
19	9.50	3	2	$f_{19,1}$	36	1	$f_{19,2}$
20	35.10	17	11	$f_{20,1}$	770	21	$f_{20,2}$
21	17.90	10	9	$f_{21,1}$	140	5	$f_{21,2}$
22	52.32	26	12	$f_{22,1}$	810	23	$f_{22,2}$
23	18.75	9	8	$f_{23,1}$	450	15	$f_{23,2}$
24	19.83	8	7	$f_{24,1}$	635	19	$f_{24,2}$
25	10.75	4	3	$f_{25,1}$	150	6	$f_{25,2}$

Table 3.3 – x_{ij} values, ranks and the associated f_{ij}

3.4.3 Model formulation

The model formulation starts with the objective function to be minimized, followed by the constraints.

$$\text{Minimize } \varepsilon_{1,1} + \varepsilon_{2,1} + \varepsilon_{1,2} + \varepsilon_{2,2} + \cdots + \varepsilon_{1,24} + \varepsilon_{2,24} + \varepsilon_{1,25} + \varepsilon_{2,25} \tag{3.16}$$

$$\text{subject to } f_{1,1} + f_{1,2} + \varepsilon_{1,1} - \varepsilon_{2,1} = 16.68, \tag{3.17}$$

$$f_{2,1} + f_{2,2} + \varepsilon_{1,2} - \varepsilon_{2,2} = 11.50, \tag{3.18}$$

$$f_{3,1} + f_{3,2} + \varepsilon_{1,3} - \varepsilon_{2,3} = 12.03, \tag{3.19}$$

$$f_{4,1} + f_{4,2} + \varepsilon_{1,4} - \varepsilon_{2,4} = 14.88, \tag{3.20}$$

$$\vdots$$

$$f_{25,1} + f_{25,2} + \varepsilon_{1,25} - \varepsilon_{2,25} = 10.75, \tag{3.21}$$

$$f_{7,1} - f_{2,1} \leq 0, \tag{3.22}$$

$$f_{2,1} - f_{3,1} = 0, \tag{3.23}$$

$$f_{3,1} - f_{19,1} = 0, \tag{3.24}$$

$$f_{19,1} - f_{4,1} \leq 0, \tag{3.25}$$

$$\vdots$$

$$f_{22,1} - f_{9,1} \leq 0, \tag{3.26}$$

$$f_{19,2} - f_{4,2} \leq 0, \tag{3.27}$$

$$f_{4,2} - f_{7,2} \leq 0, \tag{3.28}$$

$$f_{7,2} - f_{18,2} \leq 0, \tag{3.29}$$

$$\vdots$$

$$f_{22,2} - f_{9,2} \leq 0, \tag{3.30}$$

$$\varepsilon_{1i}, \varepsilon_{2i} \geq 0, \qquad \text{for } i = 1, \cdots, 25, \tag{3.31}$$

$$f_{ij} \text{ is unrestricted in sign for all } i \text{ and } j. \tag{3.32}$$

3.4.4 Model solution

The data was imported from a text file into a C++ program which connects with CPLEX (version 10.1) using Concert Technology from ILOG (ILOG, 2006). The model was solved using this software to find solution values for the function variables, f_{ij}, and the objective value thereof (minimization of the sum of the error variables). Other information, such as shadow prices, the right-hand-side range, the range of insignificance and the range of optimality can be obtained from CPLEX to make further calculations and to determine the model's sensitivity.

Table 3.4 records the distinct x_1 values with their associated f_1 values after the model was solved.

x_1	2	3	4	5	6	7	8	9	10	16	17	26	30
$f_1(x_1)$	-7.08	-6.33	-4.33	0	0	0	0	0	2.82	15.27	15.27	26.14	53.06

Table 3.4 – Function values for f_1

The complete data set is contained in table 3.5. The first column, i, indicates the data point, the second column contains the y_i values, followed by the values of the first variable, x_{i1}, and the associated function values, $f_1(x_{i1})$ in the third column, followed by the values of the second variable, x_{i2}, and the associated function values, $f_2(x_{i2})$. The second last column, $\hat{y}_i$, is the estimated value, calculated by adding $f_1(x_{i1})$ and $f_2(x_{i2})$, and the last column contains the absolute deviations, $|y_i - \hat{y}_i|$, (the difference between the observed and estimated y value).

i	y_i	x_{i1}	$f_1(x_{i1})$	x_{i2}	$f_2(x_{i2})$	$\hat{y}_i$	$\lvert y_i - \hat{y}_i \rvert$
1	16.68	7	0	560	19.75	19.75	3.07
2	11.50	3	-6.33	220	17.83	11.50	0
3	12.03	3	-6.33	340	18.36	12.03	0
4	14.88	4	-4.33	80	15.08	10.75	4.13
5	13.75	6	0	150	15.08	15.08	1.33
6	18.11	7	0	330	18.11	18.11	0
7	8.00	2	-7.08	110	15.08	8.00	0
8	17.83	7	0	210	17.83	17.83	0
9	79.24	30	53.06	1460	26.18	79.24	0
10	21.50	5	0	605	19.83	19.83	1.67
11	40.33	16	15.27	688	19.83	35.10	5.23
12	21.00	10	2.82	215	17.83	20.65	0.35
13	13.50	4	-4.33	255	17.83	13.50	0
14	19.75	6	0	462	19.75	19.75	0
15	24.00	9	0	448	18.75	18.75	5.25
16	29.00	10	2.82	776	26.18	29.00	0
17	15.35	6	0	200	15.35	15.35	0
18	19.00	7	0	132	15.08	15.08	3.92
19	9.50	3	-6.33	36	15.08	8.75	0.75
20	35.10	17	15.27	770	19.83	35.10	0
21	17.90	10	2.82	140	15.08	17.90	0
22	52.32	26	26.14	810	26.18	52.32	0
23	18.75	9	0	450	18.75	18.75	0
24	19.83	8	0	635	19.83	19.83	0
25	10.75	4	-4.33	150	15.08	10.75	0

Table 3.5 – Data, function values and residuals

Figure 3.1 depicts the absolute residuals for each data point graphically. It can be observed in this figure that data points 11 and 15 indicate the largest deviation, followed by points 4 and 18. This deviation may be due to possible outliers. In the next chapter this possibility will be further investigated.

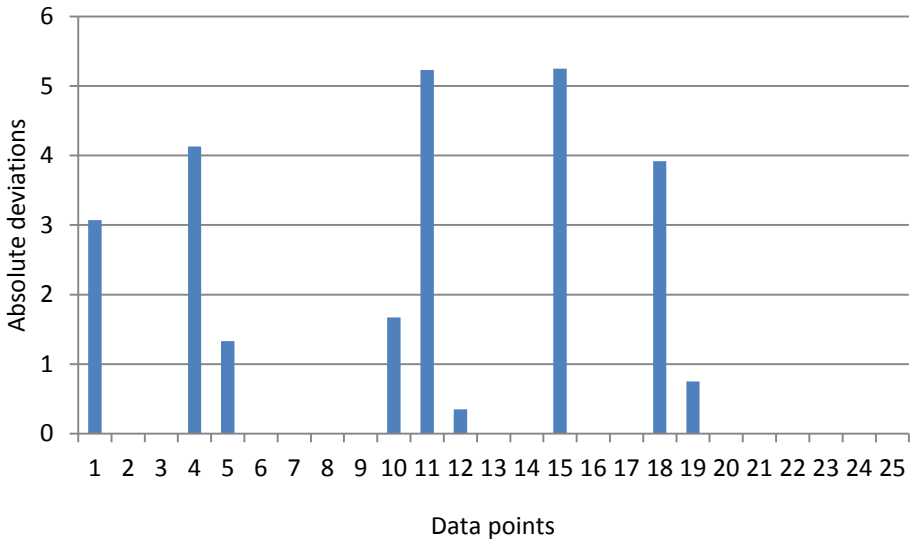

Figure 3.1 – The absolute deviation for each data point

In figures 3.2 and 3.3 the function values of f_1 and f_2 are plotted against the x_1 and x_2 values respectively. In both figures it is easy to perceive that the functions are monotonically non-decreasing, as specified by the multiple regression performed beforehand and the inequality constraints in the model.

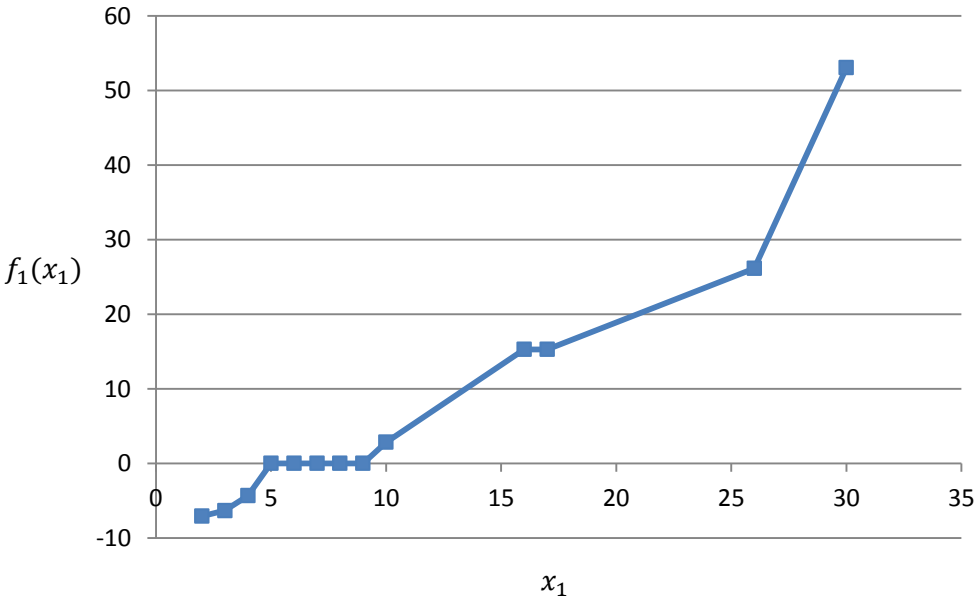

Figure 3.2 – f_1 values plotted against x_1 values

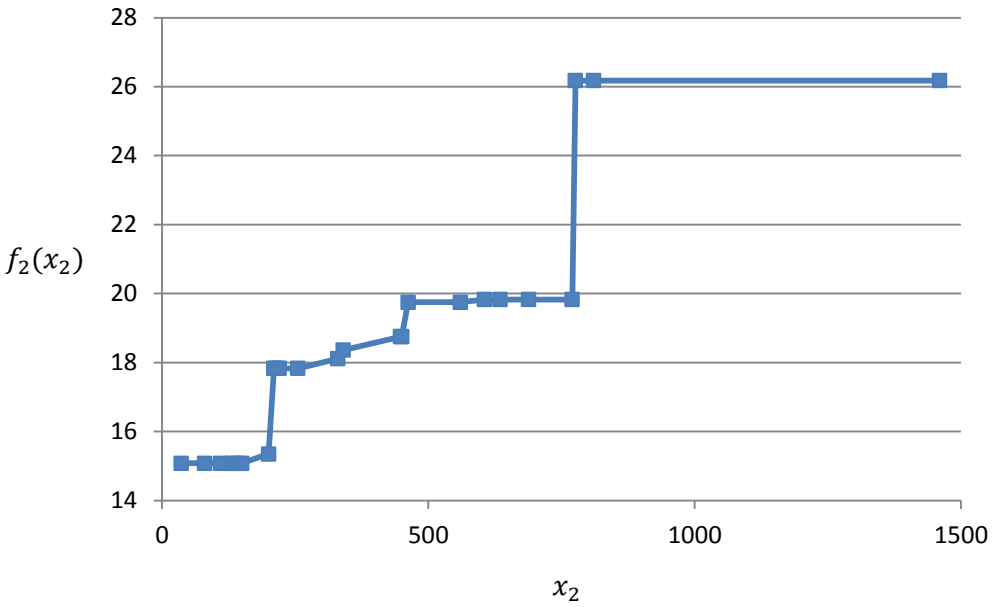

Figure 3.3 – f_2 values plotted against x_2 values

In this section the minimal assumption regression model was applied to the delivery time data set. Firstly, the direction of monotonicity for the two predictor variables was estimated, followed by the assignment of ranks and the creation of inequality constraints to enforce monotonicity. The problem was formulated as a linear programming problem and solved with CPLEX. The form of each function can be seen by plotting the function values against the values of the predictor variables (figures 3.2 and 3.3).

One of the problems of this model is that it may be difficult to estimate a value for f_{ij} if x_{ij} is not in the data set. Therefore it is sometimes necessary to use interpolation or extrapolation to determine the value of a certain f_{ij}. An extrapolation method will be discussed in the following section.

3.5 Extrapolation

To obtain a function value, f_{i1}, of a distinct x_{i1} value, table 3.4 (section 3.4.4) can be used to read the value as the solution of the model was given in table form. For example: if the x value is 2 then the function value is -7.08; if the x value is 8 the function value is 0; and if the x value is 30 the function value is 53.06. There is, however, no specific function to determine the f_{ij} value of a x value that is not specified in the table. To solve this problem, interpolation and extrapolation can be used. The formula for interpolation or extrapolation is given below (Cho & Skidmore, 2006)

$$y_2 = \frac{(x_2 - x_1)}{(x_3 - x_1)}(y_3 - y_1) + y_1. \tag{3.33}$$

When extrapolation is utilized, it may be difficult to decide which two points should be used to determine the third point. To illustrate this, consider variable x_1 and suppose a function value for f_1 must be determined where $x_1 = 35$. Using the extrapolation formula (3.33) above, the extrapolated function value will typically be calculated as follows using the last two data points (B and C in figure 3.4) (see table 3.4 for the values used in the extrapolation formula)

$$f_1 = \frac{(35-26)}{(30-26)}(53.06 - 26.14) + 26.14$$
$$= 86.71$$

This point is indicated as E in figure 3.4.

It is, however, also possible to use the first and the last available data points (A and C in figure 3.4). In this case the extrapolation function value is

$$f_1 = \frac{(35-2)}{(30-2)}\big(53.06 - (-7.08)\big) + (-7.08)$$
$$= 63.80$$

This point is indicated as D in figure 3.4.

In this example, it appears as if it may be desirable to use the first and last data points (A and C) instead of the usual last two available points (B and C).

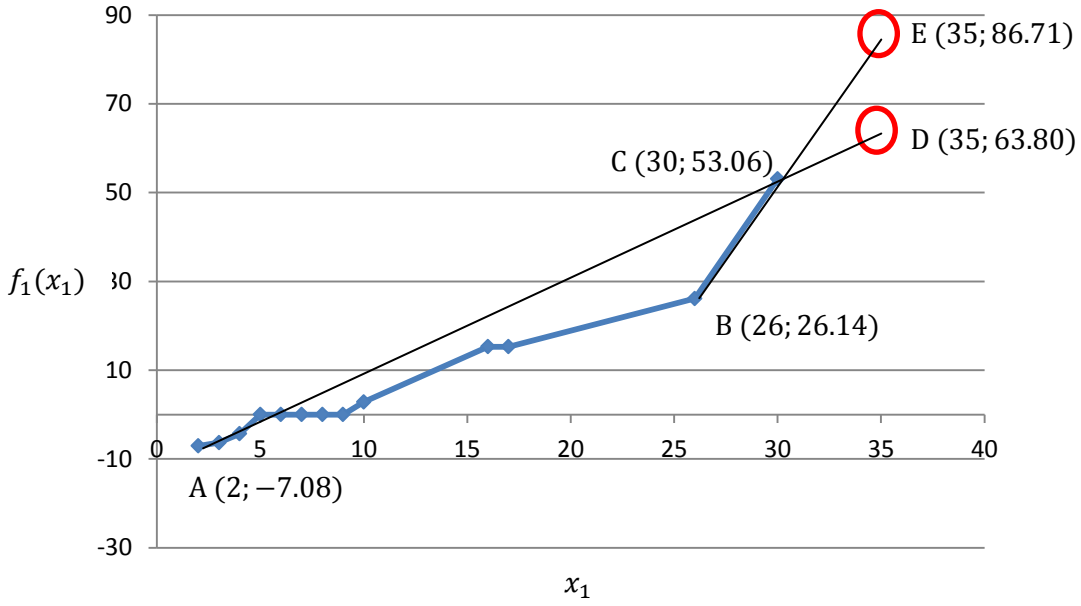

Figure 3.4 – Extrapolation techniques

To conclude the discussion on Wagner's model, the next section will present a brief literature review of other researchers who have referred to or used the minimal assumption regression model.

3.6 Literature review of other research using Wagner's model

In this section examples of other researchers who referred to or used the minimal assumption regression model developed by Harvey M. Wagner in 1962 will be furnished. Most of the references are dated with the exception of one new reference from 2005. This may be attributed to the fact that in the 1960s computers were not readily available and the computations in this method were perceived to be difficult and laborious.

An interesting paper, "Utility functions for test performance" (Dyer *et al*., 1973), discusses a utility function estimation procedure to assist elementary school principals with curriculum planning. The goal was to determine the priorities of parents, communities, and the school. The results helped in decision making and prioritisation.

The utility function is based on the scores of a group of students in different areas of the curriculum. To establish the priorities, different principals were given a questionnaire. For each question a scenario was described, with certain criteria which could be improved upon. The principal had to choose which criteria he/she preferred to improve. Dyer *et al*. (1973) mentioned the following question as an example of those in the questionnaire:

"Your students have just taken a nationally standardized test in creativity. The test has two parts, A and B, and they represent two aspects of a subject which are equally important to you. Test results are in percentile scores for school norms. Your school averages were:

Part A 50 percentile

Part B 70 percentile

Which increase would be worth more to you –

Part A from 50 to 60 percentile

or

Part B from 70 to 85 percentile?"

The principal had to choose whether he/she preferred to improve criterion A, for example, from a 50 percent to a 60 percent average, or criterion B from a 70 percent to an 85 percent average. If the principal indicated that an increase in Part A was worth more than an increase in Part B, the constraint can be written as follows

$$f(60) - f(50) \geq f(85) - f(70), \tag{3.34}$$

or

$$f(x_1) - f(x_2) \geq f(x_3) - f(x_4). \tag{3.35}$$

By using this technique it was possible to state that the difference between $f(x_1)$ and $f(x_2)$ was more important than the difference between $f(x_3)$ and $f(x_4)$. This can be done for several criteria, so that from these results information in relation to curriculum planning could be derived. Dyer *et al.* (1973) state that "this curve-fitting technique is similar in spirit to the approach proposed by Wagner."

A couple of other researchers only mentioned the approach Wagner used as an alternative regression method (Schlossmacher, 1973; Fama, 1965a; Fama 1965b; Walsh, 1963). Schlossmacher (1973) noted that absolute deviations curve fitting can be carried out using linear programming as Wagner proposed. He also noted that the program could become very big and that linear programming algorithms are not always accessible parts of statistical packages. Fama (1965a) used absolute value regression as an alternative estimation procedure to develop a portfolio analysis model for a stable paretian market. Fama (1965b) noted that an alternative technique, such as absolute-value regression, can be used to approach the stable paretian process. Walsh (1963) also briefly mentioned that least absolute deviation could be used to develop a regression method that possesses substantial curve-fitting flexibilities without substantial changes in the desirable computational properties. The non-linear regression problem that Diewert and Wales (2005) suggest for a single turning point smoothing procedure is very closely related to the nonparametric regression model by Wagner.

Another survey attempted to collect and classify different applications of linear programming to numerical analysis (Rabinowitz, 1968). Rabinowitz classified five different approximation problems: discrete linear Chebyshev approximation, L_1 approximation, fitting by spline curves, fitting by rational functions and lastly, general regression which used Wagner's paper as a basis.

There are a few other sources which mention the problems of using a specific method, but these are also only very brief in their explanation of the possibility of using the minimal assumption regression model to overcome these obstacles.

3.7 Chapter summary

In this chapter the minimal assumption regression model formulated by Harvey M. Wagner in 1962 was introduced. This model constitutes the basis of this research study and an example

was used to demonstrate its features. A brief literature review of other research that referenced the work of Wagner concluded the chapter.

The next chapter will be devoted to a discussion about robust model development which is implemented by adapting the minimal assumption regression model proposed by Wagner.

Chapter 4

Model development

4.1 Introduction

Following the literature and background reviews in chapters 2 and 3, Chapter 4 will concentrate on the research design and model development techniques employed to ensure greater model robustness and improved predictive accuracy. The chapter starts with two extensions made to the minimal assumption regression model introduced by Wagner (1962).

The first extension is intended to detect possible outliers by implementing mixed integer linear programming techniques. The second extension addresses the potential problem of overfitting the model by using constrained second derivatives to smooth the functions. These extensions are included in the model to ensure that the identification of outliers, smoothing and model development can be carried out in order to improve its robustness.

The second part of the chapter is dedicated to the problem of specifying mathematical forms for the functions $f_j(x_{ij})$. This is performed by using a piecewise linear regression model. This model, implemented through a mathematical programming approach, is used for comparative purposes when evaluating the results obtained by the minimal assumption regression model and the suggested robust extensions.

4.2 Robust model development

4.2.1 Identification of outliers for linear models

As discussed in Chapter 2, outliers can seriously distort a model and it is often difficult to detect these outliers if there are more than two predictor variables. To detect outliers a number of useful techniques were presented in Chapter 2 (section 2.3). In addition to these techniques, it is also possible to do so by using mixed integer linear programming techniques (Hattingh *et al*., 2005). This method is applicable to linear models.

Hattingh *et al*. (2005) developed a mixed integer linear programming model to simultaneously detect outlying data points and select certain predictor variables. A part of the motivation for this approach is that one-at-a-time case diagnostics may not be suitable for cases which are jointly but not individually outlying or where multiple outliers conceal the presence of additional outliers (the masking problem). Another reason for this approach is that predictor variables may

appear to be candidates for inclusion because of outliers and thus, when outliers and predictor variables are selected simultaneously, this need no longer be a problem.

Hattingh *et al.* (2005) used the following model to eliminate data and discard predictors at the same time

$$\text{Minimize} \quad \sum_{i=1}^{n} (\varepsilon_{1i} + \varepsilon_{2i}) \tag{4.1}$$

$$\text{subject to} \quad y_i - b_0 - b_1 x_{1i} - \cdots - b_s x_{si} + \varepsilon_{1i} - \varepsilon_{2i} + 2K\gamma_i - K\psi_i = 0, \;\; i = 1, 2, \ldots, n, \tag{4.2}$$

$$-L(1 - \delta_i) \leq b_i \leq L(1 - \delta_i), \;\; i = 1, 2, \ldots, s, \tag{4.3}$$

$$\sum_{i=1}^{s} \delta_i = r, \tag{4.4}$$

$$\sum_{i=1}^{n} \psi_i = p, \tag{4.5}$$

$$\gamma_i - \psi_i \leq 0, \tag{4.6}$$

$$\delta_i \in \{0, 1\}, \qquad i = 1, 2, \ldots, s, \tag{4.7}$$

$$\psi_i \in \{0, 1\}, \qquad i = 1, 2, \ldots, n, \tag{4.8}$$

$$\varepsilon_{1i}, \varepsilon_{2i}, \gamma_i \geq 0, \qquad i = 1, 2, \ldots, n, \tag{4.9}$$

where p is the number of data points to be deleted from the model and r is the number of predictor variables not included in the model. For a detailed account Hattingh *et al*. (2005) can be consulted.

Although the abovementioned problem is applicable to linear models, the same principles and techniques will be used in this study to detect outliers for nonlinear models.

4.2.2 Identification of outliers for nonlinear models

To try and identify possible outliers, similar mixed integer linear programming techniques are incorporated into Wagner's model. The model is used in 4.10 – 4.18 wherein α_i is an unrestricted slack variable. The absolute value of α_i is constrained by Mz_i where M is a large number and z_i is a binary variable. In experiments a value of M larger than the span of the y_i values proved sufficient. If z_i is zero, α_i is also constrained to zero and the ith absolute residual

contributes to the objective, but if z_i is one, the optimization process will choose the ith residual to be zero, since α_i takes up the slack.

In this manner the absolute residual for data point i is omitted from the model and from the objective function. The data point that will be omitted from the model is the point that will cause the greatest decrease in the objective function when it is omitted. The variable p specifies the number of data points omitted from the model. In this study the value of p will be determined by experimentation.

$$\text{Minimize} \quad \sum_{i=1}^{n} (\varepsilon_{1i} + \varepsilon_{2i}) \tag{4.10}$$

$$\text{subject to} \quad \sum_{j=1}^{k} f_{ij} + \varepsilon_{1i} - \varepsilon_{2i} - \alpha_i = y_i, \qquad \text{for } i = 1, \cdots, n, \tag{4.11}$$

$$f_{tj} \leq f_{lj}, \qquad \text{if } r_{lj} = r_{tj} + 1, \text{and} \tag{4.12}$$

$$f_{tj} = f_{lj}, \qquad \text{if } r_{tj} = r_{lj} \text{ for } t, l = 1,2, \dots, n \text{ with } t \neq l \text{ and } j = 1,2, \dots, k, \tag{4.13}$$

$$-Mz_i \leq \alpha_i \leq Mz_i, \qquad \text{for } i = 1, \cdots, n, \tag{4.14}$$

$$\sum_{i=1}^{n} z_i = p, \tag{4.15}$$

$$z_i \in \{0,1\}, \qquad \text{for } i = 1, \cdots, n, \tag{4.16}$$

$$\varepsilon_{1i}, \varepsilon_{2i} \geq 0, \qquad \text{for } i = 1, \cdots, n, \tag{4.17}$$

$$f_{ij} \text{ and } \alpha_i \text{ are unrestricted in sign for all } i \text{ and } j. \tag{4.18}$$

The delivery time data set introduced in Chapter 3 will now be used to illustrate how the above model can be applied to a data set. The first step is to decide how many data points should be omitted; therefore a value for p should be determined.

4.2.2.1 Determination of p

In this study the value of p is determined experimentally as follows: the model is solved for $p = 0$ (no data points are omitted) and the value of the objective function is recorded. The model is then solved repeatedly and the value of p is incremented by 1 each time the model is solved, with the respective objective values also being recorded. The different values of p (the

number of points omitted) are then plotted against the relevant recorded values to observe how the objective value has changed.

In figure 4.1 the number of points omitted (p) is plotted against the respective objective values. The values in the brackets shown on the graph indicate the data points that are omitted. One can read from the graph that when 0 points are omitted the objective value is 25.7 but when 6 points (points 1, 4, 18, 20, 23 and 24) are omitted the objective value decreases to 1.1. It is important not to omit too many data points as the data set may become too small to construct a meaningful model.

The value of p can be based on the change in the objective value. When a certain number of points, for instance m, are omitted, the change in the objective value can be determined when an additional point, $m+1$, is omitted. When the change in the objective value becomes smaller, the data point omitted does not reduce the objective value as much as the previous data points that were omitted. This may indicate that this data point is not an outlier and it will not be included in the value of p.

However, if there is no definite change in the decrease of the objective value, p can be selected as a value that represents 10-20% of the data points. This prevents the data set from becoming too small. In this case the objective function decreases very gradually, thus, for illustrative purposes, the value of p will be chosen as 3 (approximately ten percent of the data).

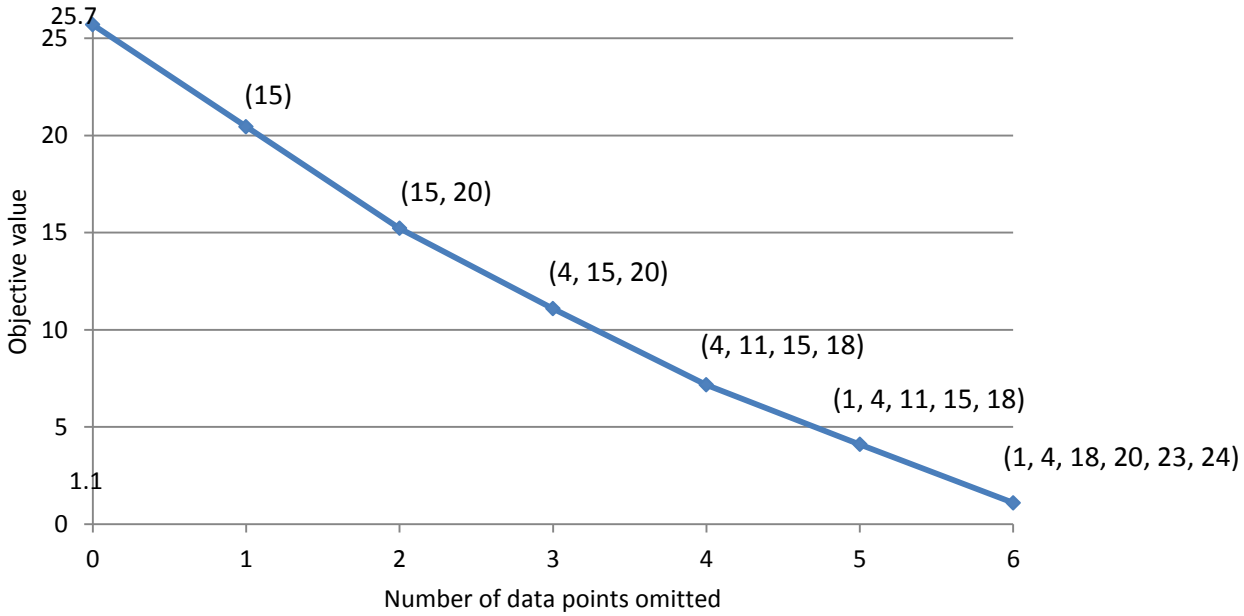

Figure 4.1 – Number of data points to omit

Du Plessis (2010) developed a heuristic to determine the number of data points to omit. This heuristic is based on the concept of the rate of change in the adjusted R^2 (or the absolute

version thereof). Using this heuristic, the rate of change (improvement) in the adjusted R^2 remains relatively high while outliers are eliminated. When all the outliers have been eliminated and "good" points are being eliminated, the rate of change in the adjusted R^2 becomes smaller. See Du Plessis (2010) for a more detailed description and implementation of this heuristic.

4.2.3 Smoothing

The second extension to the minimal assumption regression model addresses the problem of possible overfitting of the model. As mentioned, this occurs when a function fits a data set "too well", when the number of parameters estimated from the data set is too large and thus error estimation becomes unreliable. Overfitting is a very serious problem because it affects the prediction capability of a model and makes it less reliable (Hitchcock & Sober, 2004).

Some of the important goals of a model are to generate accurate predictions and describe relationships "correctly". In this research study it is important to formulate a model which makes accurate predictions. One way to prevent overfitting is by using a polynomial of a lower degree which still fits the data well, but not as well as a higher degree polynomial (Hitchcock & Sober, 2004). Another way to prevent a model from overfitting the data is to smooth the functions, or in other words to adjust the functions by making them less sensitive to the data without reducing the appropriateness of the model.

The smoothing effect is also incorporated into the model because of fluctuations, especially in small data sets. While large data sets may reveal relatively smooth functions $f_j(x_{ij})$, small data sets may show sudden, large fluctuations. To illustrate this, consider the following two variables from different data sets on which the minimal assumption regression model has been applied.

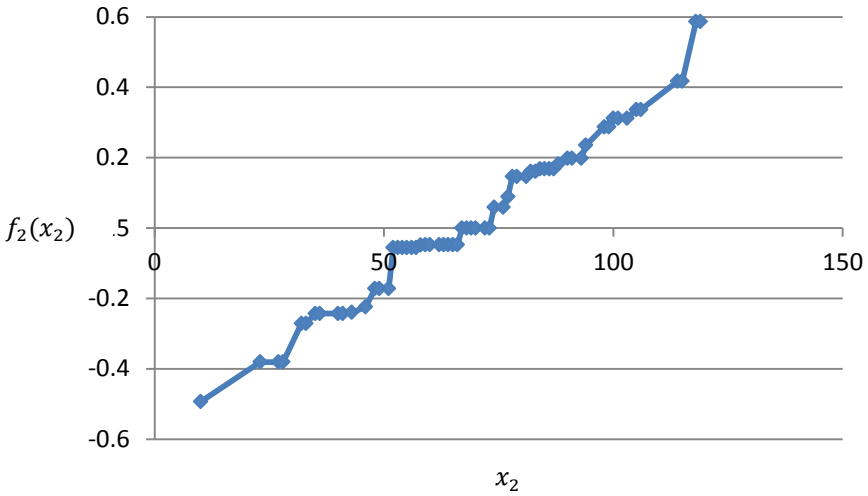

Figure 4.2 – Second variable of a liver surgery data set (Neter *et al*., 1990)

Figure 4.2 is a representation of the function of the second variable of a liver surgery data set (Neter *et al*., 1990) that consists of 108 data points. The function is quite smooth without any

large, sudden changes and the relationship is clearly observable. Figure 4.3 is a representation of the function of the second variable of the stack loss data set (Brownlee, 1965) that consists of 21 data points. This function is clearly not as smooth as the one in figure 4.2 and it also illustrates some fluctuations. Appropriate smoothing on the function can lessen this effect.

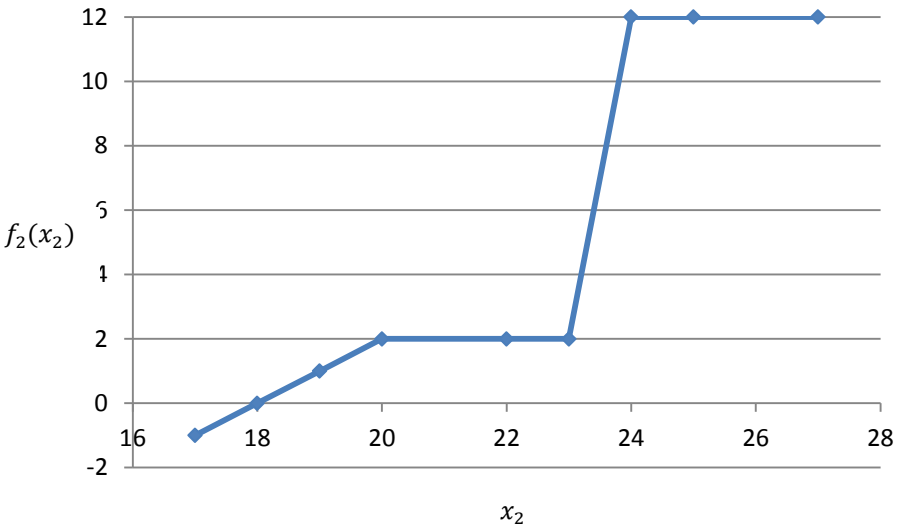

Figure 4.3 – Second variable of the stack loss data set (Brownlee, 1965)

To incorporate smoothing into the minimal assumption regression model some constraints are added to the model. The smoothing technique used in the minimal assumption regression model is intended to constrain the second derivative of the function, or in other words, the rate of direction change. The slope of a function cannot change more than a specified value and this constrains sudden large fluctuations in the slope.

To implement the smoothing of a function, constrained second derivatives are employed and can be described as follows:

Set $f_j(x_j) = f_j$ and consider

$$\frac{\partial f_j}{\partial x_j}|_{x_{i,j}} \approx \frac{f_j(x_{i+1,j}) - f_j(x_{i,j})}{x_{i+1,j} - x_{i,j}},$$

$$\frac{\partial^2 f_j}{\partial x_j^2}|_{x_{i,j}} \approx \frac{\frac{\partial f_j}{\partial x_j}|_{x_{i,j}} - \frac{\partial f_j}{\partial x_j}|_{x_{i-1,j}}}{x_{i,j} - x_{i-1,j}}, \text{and}$$

$$-\beta \leq \frac{\frac{f_j(x_{i+1,j}) - f_j(x_{i,j})}{x_{i+1,j} - x_{i,j}} - \frac{f_j(x_{i,j}) - f_j(x_{i-1,j})}{x_{i,j} - x_{i-1,j}}}{x_{i,j} - x_{i-1,j}} \leq \beta, \qquad (4.19)$$

where $\approx$ denotes an approximation.

The absolute rate of change in direction (the second derivative) is now constrained by the parameter β. This parameter is estimated by experimentation in this study.

It should be noted that extra constraints usually restrict the model further and thus the objective value may also become affected. In the case of a minimization problem, the optimal objective value will not decrease if additional restrictive constraints are added to the model.

Before the steps to determine the value of β are explained, another concept must be introduced. This is a method for carrying out cross-validation which is proposed in the next section.

4.2.3.1 Cross-validation

Cross-validation is a technique for estimating the performance of a predictive model. To test the performance a data set can be divided into two subsets: the model is applied to one set and the accuracy of that model is then tested against the other set.

Various cross-validation methods can be used to test the forecast accuracy of a model (Browne, 2000; Geisser, 1975; Stone, 1974). In this study a "leave one out" jack-knife approach (Efron & Gong, 1983) was used to determine how well the model predicts. This method consists of the following steps:

- delete the points x_i from the data set one at a time;
- recalculate the prediction rule on the basis of the remaining $n - 1$ points;
- see how well the recalculated rule predicts the deleted point; and
- average these predictions over all n deletions of x_i.

The mean absolute deviation calculated in the last step gives an indication of how well the model explains the data. Generally a small mean absolute deviation is preferred because this will provide more accurate predictions.

The mean absolute deviation is used as a measurement to determine the value of β. This approach is explained in the following section.

4.2.3.2 Determination of β

In this study it was established by experimentation that a smoothing factor between 1 and 400 will smooth the model sufficiently. To determine a smoothing factor β (4.19), set β equal to 25 and solve the model. Determine the mean absolute deviation and record the value. Increase β by 25, and solve the model again. Record the mean absolute deviation for each iteration until

$\beta = 400$. Plot the recorded mean absolute deviations against the β values. Select the β value which results in the lowest mean absolute deviation.

In figure 4.4 the mean absolute deviation values were calculated for $\beta = 25$ to $\beta = 400$ with increments of 25. The mean absolute deviation of the delivery time data set was the lowest for $\beta = 25$ and therefore the rest of the calculations in this illustrative example will be performed with $\beta = 25$.

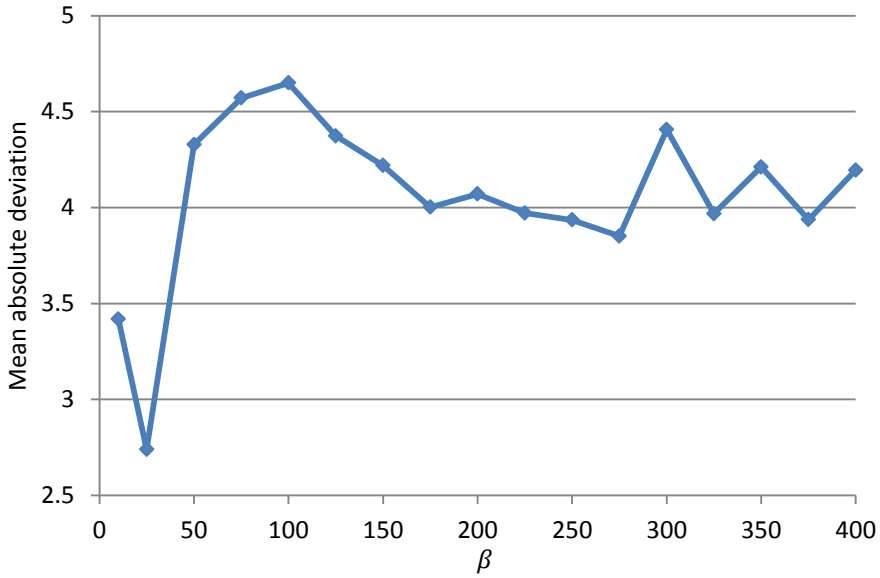

Figure 4.4 – Mean absolute deviation for different values of β

To illustrate the change in the form of the functions a smoothing factor of $\beta = 25$ is applied to the delivery time data set. In figures 4.5 and 4.6 the representations of the functions of the minimal assumption regression model are shown, indicated as $f_1(x_1)$ and $f_2(x_2)$ respectively.

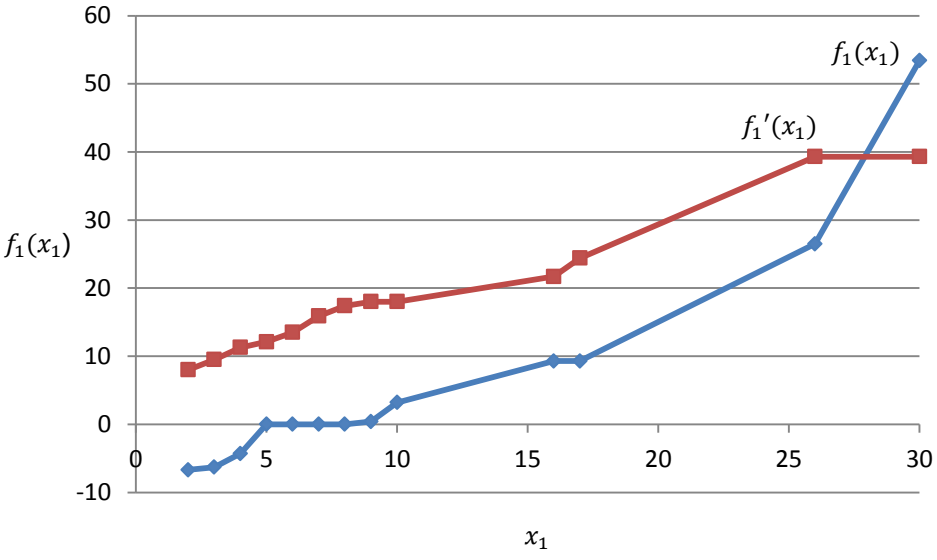

Figure 4.5 – Change in function f_1 after smoothing with β = 25

The smoothed functions with $\beta = 25$ are also shown and are indicated as $f_1'(x_1)$ and $f_2'(x_2)$ in figures 4.5 and 4.6 respectively. In these two figures it is noticeable that the slope of the smoothed functions changes more gradually from one data point to the next.

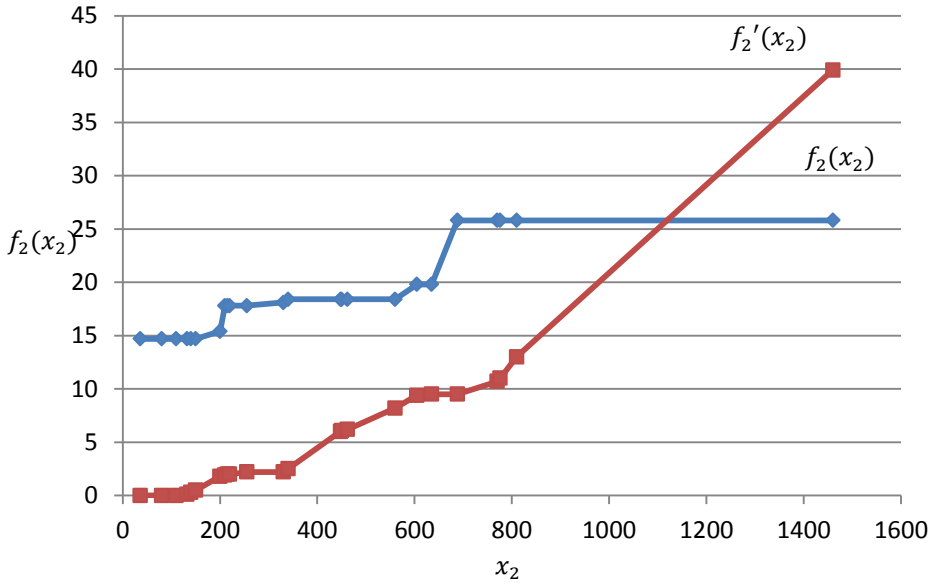

Figure 4.6 – Change in function f_2 after smoothing with β = 25

The extensions to the minimal assumption regression model can be explained as follows. A model is made more robust by omitting outliers, but when too many data points are omitted the data set may become too small to estimate the relationship between the dependent variable and the predictor variables. By smoothing the functions the model is prevented from overfitting the data, but when a function is smoothed too much it might only indicate a general trend and predictive accuracy is compromised. It is therefore important to find the balance between omitting outliers and smoothing in order to develop the model.

The next section will introduce another mathematical programming approach, piecewise linear regression, to specify mathematical forms for the functions $f_j(x_{ij})$. This model will also be used for comparative purposes when evaluating the results obtained by the minimal assumption regression model and the suggested robust extensions.

4.3 Piecewise linear regression

Another method to approach nonlinear functions is by using a piecewise linear regression model. In this section a model of this type is developed in order to compare the results of this model with the results of the minimal assumption regression model.

To be able to obtain alternative mathematical models, a piecewise linear approach is used. Piecewise linear regression is a form of regression that allows multiple linear models to be fitted

to data for different ranges of x (Ryan & Porth, 2007). Breakpoints are the values of x where the slope of the linear function changes. The value of a breakpoint may or may not be known before the analysis, but typically it is unknown and must be estimated. Data sets in this study are either modelled as one linear regression model or as piecewise linear continuous segments, each represented by a linear model.

A model using two breakpoints and therefore producing three linear models of the form $y = a + bx$ is illustrated below. In this model, Q_{1j} and Q_{2j} represent the two breakpoints which were chosen to be the 33rd and 66th percentiles. The variables a_{sj} and b_{sj} $(s = 1, 2, 3 \text{ and } j = 1, \cdots, k)$ are the coefficients of the different linear models.

$$\text{Minimize} \quad \sum_{i=1}^{n} (\varepsilon_{1i} + \varepsilon_{2i}) \tag{4.20}$$

$$\text{subject to} \quad \sum_{j=1}^{k} f_{ij} + \varepsilon_{1i} - \varepsilon_{2i} = y_i, \qquad \text{for } i = 1, \cdots, n, \tag{4.21}$$

$$f_{ij} = a_{1j} + b_{1j} x_{ij}, \qquad \text{if } x_{ij} < Q_{1j}, \tag{4.22}$$

$$a_{2j} + b_{2j} x_{ij}, \qquad \text{if } Q_{1j} \leq x_{ij} < Q_{2j}, \tag{4.23}$$

$$a_{3j} + b_{3j} x_{ij}, \qquad \text{if } Q_{2j} \leq x_{ij}, \text{for } j = 1, \cdots, k, \tag{4.24}$$

$$a_{1j} + b_{1j} Q_{1j} = a_{2j} + b_{2j} Q_{1j}, \qquad \text{for } j = 1, \cdots, k, \tag{4.25}$$

$$a_{2j} + b_{2j} Q_{2j} = a_{3j} + b_{3j} Q_{2j}, \qquad \text{for } j = 1, \cdots, k, \tag{4.26}$$

$$\varepsilon_{1i}, \varepsilon_{2i} \geq 0, \qquad \text{for } i = 1, \cdots, n, \tag{4.27}$$

$$a_{1j}, a_{2j}, a_{3j}, b_{1j}, b_{2j}, b_{3j} \text{ unrestricted for } j = 1, \ldots, k, \tag{4.28}$$

$$f_{ij} \text{ unrestricted for all } i \text{ and } j. \tag{4.29}$$

The purpose of this model is to simultaneously fit piecewise linear models to the (additive) regression model. Outliers can be omitted from this model in the same way as described in section 4.2.2.

To illustrate the performance of the piecewise linear regression model, the model specified above is applied to the delivery time data set (Montgomery & Peck, 1992). Figures 4.7 to 4.9

depict the function f_2 with zero, one and two breakpoints respectively. Figure 4.7, f_2 with no breakpoint, is simply a L_1-norm regression with the linear model $f_2(x_2) = 0.0 + 0.0161x_2$.

For this study, in the case of one breakpoint (figure 4.8), the 50th percentile was chosen as breakpoint Q_{2j} while in the case of two breakpoints (figure 4.9), the 33rd and 67th percentiles were chosen as breakpoints Q_{1j} and Q_{2j}. It should be noted that the 50th percentile (Q_{12}) in figure 4.8 is not necessarily located in the middle of the data on the graph. This is due to the fact that there are duplicate data values for x_2 and consequently the corresponding $f_2(x_2)$ values also take on similar values. The same explanation can be provided when two breakpoints, the 33rd and 67th percentiles, are specified in figure 4.9.

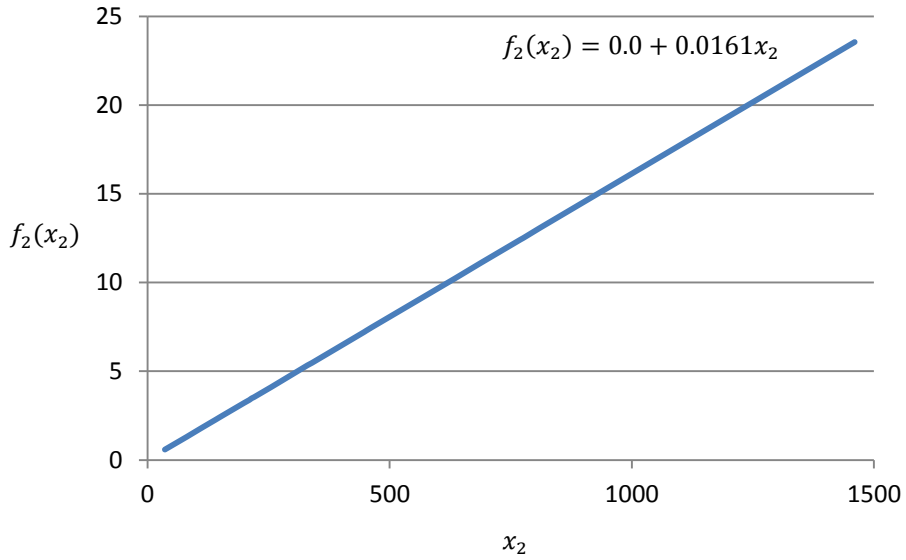

Figure 4.7 – x_2 with no breakpoint

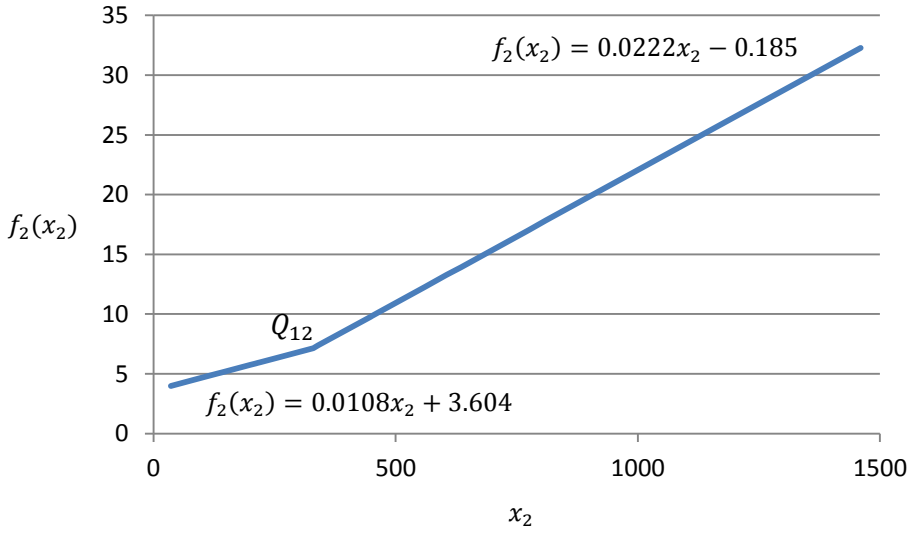

Figure 4.8 – x_2 with one breakpoint

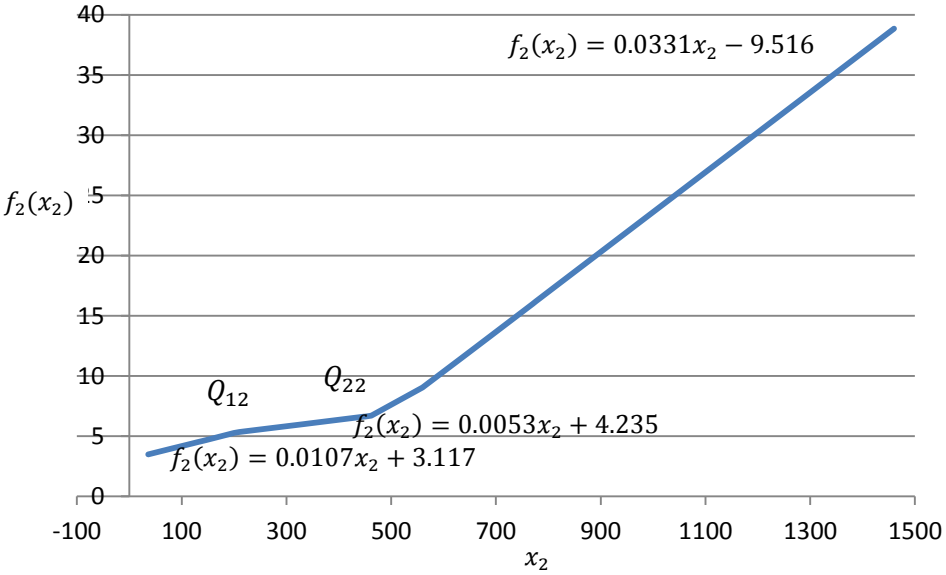

Figure 4.9 – x_2 with two breakpoints

The two results of the models introduced in this chapter, the minimal assumption regression model with extensions, and the piecewise linear regression model, can now be compared. Since the mean absolute deviation measures the accuracy of the predictive capability of a model, this measure will be used to compare the models with their different extensions. This comparison is discussed in the following section.

4.4 Model comparison

It is often difficult to weigh up different models; however, in this study the predictive capability of the different models will be compared. This is done by employing a simple form of cross-validation, the "leave one out" jack-knife procedure, explained in section 4.2.3.1. The mean absolute deviation measurement, obtained from the jack-knife procedure, is used for comparing the different models. The difficulty in this approach is determining which data points must be used to compute the mean absolute deviation. A regression procedure is not supposed to predict outlying data points well. Therefore, when data points are omitted, these data points are not estimated and they do not contribute to the mean absolute deviation because they are considered as outliers.

In this section and in Chapter 5, models that do not omit outliers will be compared separately because these models still have to estimate possible outliers. Models that omit outliers and incorporate other extensions do not estimate omitted data points.

Consider table 4.1 which shows the comparison for the L_2-norm regression, L_1-norm regression and the original minimal assumption regression model introduced by Wagner. No extensions (outlier detection and smoothing) were implemented and from the mean absolute deviation it will be noticed that the L_2- and L_1-norm regression outperformed the original Wagner model.

Model	Mean absolute deviation
L_2-norm regression	2.878
L_1-norm regression	2.536
Original minimal assumption regression model (Wagner, 1962)	3.885

Table 4.1 – Comparison of models that do not omit outliers

To illustrate the effect of the proposed extensions and the piecewise linear regression technique, consider table 4.2 below. In all cases 3 data points (outliers) were omitted (in section 4.2.2.1 the value of p was chosen as 3). By introducing a smoothing factor, $\beta = 25$ (as proposed in section 4.2.3.2), to the minimal assumption regression model, a significant improvement in the mean absolute deviation was observed (from 3.75 to 2.739). However, for this specific data set, the L_1-norm piecewise linear regression model (no breakpoint) outperformed all the other models with a mean absolute deviation of 1.652.

Model	Mean absolute deviation
Wagner's model omitting 3 data points	3.75
Wagner's model omitting 3 data points and a smoothing factor of $\beta = 25$	2.739
L_1-norm regression omitting 3 data points and no breakpoint	1.652
Piecewise L_1-norm regression omitting 3 data points and one breakpoint	2.732
Piecewise L_1-norm regression omitting 3 data points and two breakpoints	3.826

Table 4.2 – Comparison of models that omit outliers

The methods and techniques that were introduced in this chapter will be investigated in more detail in Chapter 5, in which the models are applied to several other data sets.

4.5 Chapter summary

In this chapter, two extensions (outlier detection and smoothing) to the minimal assumption regression model were suggested in order to improve robustness and to enhance the predictive capabilities of the model. A linear programming model to incorporate piecewise linear regression techniques was also introduced, and all the suggested techniques were explained and illustrated using a data set from the literature.

The next chapter will present empirical tests and results obtained by applying the models to different data sets.

Chapter 5

Empirical experiments and results

5.1 Introduction

In chapter 3 the minimal assumption regression model, introduced by Wagner (1962) was presented. This research study is based on Wagner's model and in Chapter 4 some extensions were made to improve the predictive accuracy of the model. The objective in adding these extensions (outlier detection and smoothing of functions) is to make the model more robust. Also included in Chapter 4 is a piecewise linear regression model. This model specifies mathematical forms for the $f_j(x_{ij})$ functions. The results of this model are compared with the results of the minimal assumption regression model.

This chapter is dedicated to empirical experiments and results. Five data sets are used to test and evaluate the model that has been developed. The minimal assumption regression model will be applied to each data set, as described in section 3.4.1 to 3.4.4. To determine the smoothing factor, or the number of data points to omit, some further experiments will be carried out (explained in section 4.2). The piecewise linear regression model will also be applied to each data set as described in section 4.3. The results for each data set will then be compared and the end results will be used to draw some conclusions.

The minimal assumption regression model has some specific features, such as the specification of the function form. The function is restricted, to be only monotonically non-increasing or non-decreasing. To observe how this model will perform when a data set with specific elements is used, different data sets were simulated and the results of the different models were compared.

In the following section, the five data sets that will be used in this study are introduced.

5.2 Data sets

5.2.1 Stack loss

The stack loss data set (Brownlee, 1965) was also examined by Hoeting *et al*. (1996). They developed a method for simultaneous variable selection and outlier identification based on the computation of posterior model probabilities. Their model appears as if it could identify masked outliers successfully.

The relationship between the dependent variable, the percentage of unconverted ammonia that escapes from a plant, and three predictor variables was inspected over 21 days. The data set is presented in table 5.1. The three predictor variables are:

x_1: air flow, which measures the rate of operation of a plant;
x_2: inlet temperature of cooling water circulating through coils in a tower; and
x_3: a value proportional to the concentration of acid in the tower.

i	1	2	3	4	5	6	7	8	9	10	11
y_i	42	37	37	28	18	18	19	20	15	14	14
x_{i1}	80	80	75	62	62	62	62	62	58	58	58
x_{i2}	27	27	25	24	22	23	24	24	23	18	18
x_{i3}	89	88	90	87	87	87	93	93	87	80	89

i	12	13	14	15	16	17	18	19	20	21
y_i	13	11	12	8	7	8	8	9	15	15
x_{i1}	58	58	58	50	50	50	50	50	56	70
x_{i2}	17	17	19	18	18	19	19	20	20	20
x_{i3}	88	82	93	89	86	72	79	80	82	91

Table 5.1 – Stack loss data (Brownlee, 1965)

5.2.2 Scottish hill racing

The second data set comprises data about Scottish hill runners (Atkinson, 1986). The relationship between the dependent variable, record time in minutes, and the two predictor variables was evaluated for 35 hill races. The two predictor variables are:

x_1: distance, the total length of the race, measured in miles; and
x_2: climb, the total elevation gained in the race, measured in feet.

The longer the race and the higher the climb, the longer one can expect the record time to be. The data set is shown in table 5.2.

i	1	2	3	4	5	6	7	8	9	10
y_i	16.083	48.350	33.650	45.600	62.267	73.217	204.617	36.367	29.750	39.750
x_{i1}	2.5	6	6	7.5	8	8	16	6	5	6
x_{i2}	650	2500	900	800	3070	2866	7500	800	800	650

i	11	12	13	14	15	16	17	18	19	20
y_i	192.667	43.050	65.00	44.133	26.933	72.250	98.417	78.620	17.417	32.567
x_{i1}	28	5	9.5	6	4.5	10	14	3	4.5	5.5
x_{i2}	2100	2000	2200	500	1500	3000	2200	350	1000	600

i	21	22	23	24	25	26	27	28	29	30
y_i	15.950	27.900	47.633	17.933	18.683	26.217	34.433	28.567	50.500	20.950
x_{i1}	3	3.5	6	2	3	4	6	5	6.5	5
x_{i2}	300	1500	2200	900	600	2000	800	950	1750	500

i	31	32	33	34	35
y_i	85.583	32.383	170.250	28.100	159.833
x_{i1}	10	6	18	4.5	20
x_{i2}	4400	600	5200	850	5000

Table 5.2 – Scottish hill racing data (Atkinson, 1986)

5.2.3 Weisberg fuel consumption

The third data set contains data about fuel consumption in different states in America (Weisberg, 2005). The relationship between the dependent variable, fuel consumption in gallons per person, and the four predictor variables was evaluated for 48 states (see table 5.3 below). The four predictor variables for each state are:

x_1: tax, the 1972 amount of tax per gallon, measured in cents;
x_2: income, the 1972 per-capita income in thousands of dollars;
x_3: road, the 1971 thousands of miles of primary highway; and
x_4: licence, the percentage of the population with a driver's licence.

i	1	2	3	4	5	6	7	8	9	10
y_i	554	628	632	524	457	587	540	574	631	635
x_{i1}	7	7.5	7	7	10	7	8	8	7.5	7
x_{i2}	3.333	3.357	4.300	5.002	5.342	4.449	4.983	4.188	3.846	4.318
x_{i3}	6.594	4.121	3.635	9.794	1.333	4.639	0.602	5.975	9.061	10.340
x_{i4}	51.3	54.7	60.3	59.3	57.1	62.6	60.2	56.3	57.9	58.6

i	11	12	13	14	15	16	17	18	19	20
y_i	648	471	580	649	534	487	414	464	541	525
x_{i1}	8.5	7.5	8	7	9	8	7.5	9	9	7
x_{i2}	3.635	5.126	4.391	4.593	3.601	3.528	4.870	4.897	3.571	4.817
x_{i3}	3.274	14.186	5.939	7.834	4.650	3.495	2.351	2.449	1.976	6.930
x_{i4}	66.3	52.5	53	66.3	49.3	48.7	52.9	51.1	52.5	57.4

i	21	22	23	24	25	26	27	28	29	30
y_i	566	603	577	704	566	714	640	524	467	699
x_{i1}	7	7	8	7	9	7	8.5	9	8	7
x_{i2}	4.332	4.206	3.063	3.897	3.721	3.718	4.341	4.092	5.126	3.656
x_{i3}	8.159	8.508	6.524	6.385	4.746	4.725	6.010	1.250	2.138	3.985
x_{i4}	60.8	57.2	57.8	58.6	54.4	54	67.7	57.2	55.3	56.3

i	31	32	33	34	35	36	37	38	39	40
y_i	782	344	498	644	610	464	410	577	865	571
x_{i1}	6	8	7	6.58	7	8	8	8	7	7
x_{i2}	5.215	5.319	4.512	3.802	4.296	4.447	4.399	3.448	4.716	3.640
x_{i3}	2.302	11.868	8.507	7.834	4.083	8.577	0.431	5.399	5.915	6.905
x_{i4}	67.2	45.1	55.2	62.9	62.3	52.9	54.4	54.8	72.4	51.8

i	41	42	43	44	45	46	47	48
y_i	640	591	547	561	508	510	460	968
x_{i1}	5	7	9	9	7	9	8.5	7
x_{i2}	4.045	3.745	4.258	3.865	4.207	4.476	4.574	4.345
x_{i3}	17.782	2.611	4.686	1.586	6.580	3.942	2.619	3.905
x_{i4}	56.6	50.8	51.7	58	54.5	57.1	55.1	67.2

Table 5.3 – Weisberg fuel data (Weisberg, 2005)

5.2.4 Gross national product (GNP)

The fourth data set was obtained from Roux (1994). He performed a regression study relating the gross national product (GNP) to 10 factors for 43 different countries. Using linear model selection Hattingh *et al*. (2005) simultaneously selected data points and variables that could be omitted from the data set. According to this research, variables 8, 9 and 10 can be omitted from the data set.

Based on the study undertaken by Hattingh *et al*. (2005), 7 of the 10 variables were selected to be included in the data set, which is presented in table 5.4. The predictor variables are described as follows:

x_1: Nett exports per capita. If the value is negative the country imports more than it exports whereas if the value is positive the country exports more than it imports;

x_2: Change in inflation. If the value is negative it means there was an improvement in the inflation (the inflation decreased);

x_3: Agriculture as a percentage of the gross household product (GHP). A high percentage points to a greater dependency on agriculture in the GHP of the country;

x_4: Political situation. A 1 symbolizes a more autocratic rule, whereas 5 represents a more democratic rule;

x_5: Average illiteracy of the population in the country. A high percentage suggests a high level of illiteracy;

x_6: Growth in life expectancy of the inhabitants of the country. If the percentage is high, there is a positive growth in the life expectancy of the population and the people live longer; and

x_7: Growth in the population of the country.

i	1	2	3	4	5	6	7	8	9	10	11
y_i	105	169	202	281	276	306	317	349	542	597	687
x_{i1}	-14	-29	47	8	-42	-4	-17	-23	60	-77	61
x_{i2}	12	7	0.5	35	2	0.2	6	-4	27	-2	32
x_{i3}	60	41	38	25	54	26	15	38	25	31	39
x_{i4}	1	2	2	2	2	2	2	2	3	2	2
x_{i5}	38	59	58	39	41	57	31	70	26	72	24
x_{i6}	20	2	4	13	7	13	9	6	15	12	10
x_{i7}	42	35	31	30	31	30	32	27	23	22	13

i	12	13	14	15	16	17	18	19	20	21	22
y_i	697	901	972	1180	1239	1345	1141	1467	1883	1715	1639
x_{i1}	-50	20	-88	-103	-76	17	88	-278	81	168	9
x_{i2}	3	-1	2	3	20	14	140	-3	1	256	-5
x_{i3}	28	39	26	22	24	40	24	33	7	17	34
x_{i4}	2	2	2	2	2	3	3	2	2	3	3
x_{i5}	14	41	67	9	26	12	15	12	10	5	50
x_{i6}	3	12	9	6	6	5	7	4	1	1	16
x_{i7}	24	21	16	15	17	29	8	11	16	14	6

i	23	24	25	26	27	28	29	30	31	32	33
y_i	2083	2085	2189	2310	2359	3475	3699	9379	10113	12135	12983
x_{i1}	-98	272	109	136	11	-875	-606	-692	-525	-793	-52
x_{i2}	6	4	196	1	60	8	7	-3	91	-5	1
x_{i3}	28	28	26	7	29	8	5	5	18	1	3
x_{i4}	3	3	3	3	3	3	4	4	4	4	4
x_{i5}	25	29	22	5	10	8	16	5	5	3	3
x_{i6}	1	1	5	3	5	4	6	5	6	4	3
x_{i7}	6	9	11	9	16	9	5	3	9	2	8

i	34	35	36	37	38	39	40	41	42	43
y_i	15053	16889	16523	18867	20440	20559	20306	22297	23217	25758
x_{i1}	-153	-244	-314	32	1162	-582	-269	809	550	-1016
x_{i2}	-1	-2	-2	-1	-3	-3	-4	-2	-6	-2
x_{i3}	1	17	5	11	1	11	4	2	6	1
x_{i4}	3	4	4	4	4	4	4	4	4	4
x_{i5}	3	3	3	3	3	3	3	3	3	3
x_{i6}	4	7	4	4	3	3	3	3	4	4
x_{i7}	4	4	3	7	2	2	6	3	3	4

Table 5.4 – Gross national product data (Roux, 1994)

5.2.5 Financial ratios

The last data set, contained in table 5.5, was taken from the book "Regression analysis by example" (Chatterjee & Hadi, 2006). To identify any financial decline in an organization is a crucial aspect in the control and management of a business. Chatterjee and Hadi (2006) point out that failure to identify poor performance can lead to severe difficulties, such as the savings-and-loan fiasco of the 1980s in the United States of America. The data set consists of specific information ratios of 66 firms, of which 33 went bankrupt after two years while 33 remained solvent during the same period. The dependent variable is defined as

$$y = \begin{cases} 0 & \text{if bankrupt after 2 years} \\ 1 & \text{if solvent after 2 years} \end{cases}$$

The three predictor variables are operating financial ratios and are described as follows:

$x_1 = \dfrac{\text{Retained earnings}}{\text{Total assets}}$;

$x_2 = \dfrac{\text{Earnings before interest and taxes}}{\text{Total assets}}$; and

$x_3 = \dfrac{\text{Sales}}{\text{Total assets}}$.

i	1	2	3	4	5	6	7	8	9	10	11
y_i	0	0	0	0	0	0	0	0	0	0	0
x_{i1}	-62.8	3.3	-120.8	-18.1	-3.8	-61.2	-20.3	-194.5	20.8	-106.1	-39.4
x_{i2}	-89.5	-3.5	-103.2	-28.8	-50.6	-56.2	-17.4	-25.8	-4.3	-22.9	-35.7
x_{i3}	1.7	1.1	2.5	1.1	0.9	1.7	1	0.5	1	1.5	1.2

i	12	13	14	15	16	17	18	19	20	21	22
y_i	0	0	0	0	0	0	0	0	0	0	0
x_{i1}	-164.1	-308.9	7.2	-118.3	-185.9	-34.6	-27.9	-48.2	-49.2	-19.2	-18.1
x_{i2}	-17.7	-65.8	-22.6	-34.2	-280	-19.4	6.3	6.8	-17.2	-36.7	-6.5
x_{i3}	1.3	0.8	2	1.5	6.7	3.4	1.3	1.6	0.3	0.8	0.9

i	23	24	25	26	27	28	29	30	31	32	33
y_i	0	0	0	0	0	0	0	0	0	0	0
x_{i1}	-98	-129	-4	-8.7	-59.2	-13.1	-38	-57.9	-8.8	-64.7	-11.4
x_{i2}	-20.8	-14.2	-15.8	-36.3	-12.8	-17.6	1.6	0.7	-9.1	-4	4.8
x_{i3}	1.7	1.3	2.1	2.8	2.1	0.9	1.2	0.8	0.9	0.1	0.9

i	34	35	36	37	38	39	40	41	42	43	44
y_i	1	1	1	1	1	1	1	1	1	1	1
x_{i1}	43	47	-3.3	35	46.7	20.8	33	26.1	68.6	37.3	59
x_{i2}	16.4	16	4	20.8	12.6	12.5	23.6	10.4	13.8	33.4	23.1
x_{i3}	1.3	1.9	2.7	1.9	0.9	2.4	1.5	2.1	1.6	3.5	5.5

i	45	46	47	48	49	50	51	52	53	54	55
y_i	1	1	1	1	1	1	1	1	1	1	1
x_{i1}	49.6	12.5	37.3	35.3	49.5	18.1	31.4	21.5	8.5	40.6	34.6
x_{i2}	23.8	7	34.1	4.2	25.1	13.5	15.7	-14.4	5.8	5.8	26.4
x_{i3}	1.9	1.8	1.5	0.9	2.6	4	1.9	1	1.5	1.8	1.8

i	56	57	58	59	60	61	62	63	64	65	66
y_i	1	1	1	1	1	1	1	1	1	1	1
x_{i1}	19.9	17.4	54.7	53.5	35.9	39.4	53.1	39.8	59.5	16.3	21.7
x_{i2}	26.7	12.6	14.6	20.6	26.4	30.5	7.1	13.8	7	20.4	-7.8
x_{i3}	2.3	1.3	1.7	1.1	2	1.9	1.9	1.2	2	1	1.6

Table 5.5 – Financial ratio data (Chatterjee & Hadi, 2006)

5.3 Model application

As explained in Chapter 3, section 3.4.4, the programming and data preparation for all of the models was done in C++ and CPLEX (version 10.1) using Concert Technology from ILOG (ILOG, 2006). A CD is included in this dissertation which contains the programs which were developed to implement the models as well as the data sets that are used in the study.

In this section the results of the models, applied to the five data sets, will be presented. To ensure that a comprehensive explanation is provided and that the practical limitations of the written study report are adhered to, the first application to the stack loss data set will be

described in detail. For the remaining four data sets only the main results and associated graphs will be presented.

5.3.1 Stack loss

The minimal assumption regression model will now be applied to the stack loss data set. Before this model can be applied to the data, the direction of monotonicity for each variable must be estimated. To determine this, a multiple regression was performed using Microsoft Excel 2007. The results are given in table 5.6. The coefficients of x_1 and x_2 are positive and the coefficient of x_3 is negative; therefore both f_1 and f_2 will be constrained as monotonically non-decreasing functions while f_3 will be constrained as a monotonically non-increasing function.

	Coefficients
Intercept	-38.863
x_1	0.726
x_2	1.260
x_3	-0.163

Table 5.6 – Multiple regression coefficients

As explained in Chapter 3, an unknown function value, f_{ij}, is assigned to each x_{ij} value, and for each j the x_{ij} values are ranked by using a dense ranking function, depicted as r_{ij}. Table 5.7 reports the stack loss data with x values, ranks, and associated function variables.

i	y_i	x_{i1}	r_{i1}	f_{i1}	x_{i2}	r_{i2}	f_{i2}	x_{i3}	r_{i3}	f_{i3}
1	42	80	7	$f_{1,1}$	27	9	$f_{1,2}$	89	4	$f_{1,3}$
2	37	80	7	$f_{2,1}$	27	9	$f_{2,2}$	88	5	$f_{2,3}$
3	37	75	6	$f_{3,1}$	25	8	$f_{3,2}$	90	3	$f_{3,3}$
4	28	62	4	$f_{4,1}$	24	7	$f_{4,2}$	87	6	$f_{4,3}$
5	18	62	4	$f_{5,1}$	22	5	$f_{5,2}$	87	6	$f_{5,3}$
6	18	62	4	$f_{6,1}$	23	6	$f_{6,2}$	87	6	$f_{6,3}$
7	19	62	4	$f_{7,1}$	24	7	$f_{7,2}$	93	1	$f_{7,3}$
8	20	62	4	$f_{8,1}$	24	7	$f_{8,2}$	93	1	$f_{8,3}$
9	15	58	3	$f_{9,1}$	23	6	$f_{9,2}$	87	6	$f_{9,3}$
10	14	58	3	$f_{10,1}$	18	2	$f_{10,2}$	80	9	$f_{10,3}$
11	14	58	3	$f_{11,1}$	18	2	$f_{11,2}$	89	4	$f_{11,3}$
12	13	58	3	$f_{12,1}$	17	1	$f_{12,2}$	88	5	$f_{12,3}$
13	11	58	3	$f_{13,1}$	17	1	$f_{13,2}$	82	8	$f_{13,3}$
14	12	58	3	$f_{14,1}$	19	3	$f_{14,2}$	93	1	$f_{14,3}$
15	8	50	1	$f_{15,1}$	18	2	$f_{15,2}$	89	4	$f_{15,3}$
16	7	50	1	$f_{16,1}$	18	2	$f_{16,2}$	86	7	$f_{16,3}$

17	8	50	1	$f_{17,1}$	19	3	$f_{17,2}$	72	11	$f_{17,3}$
18	8	50	1	$f_{18,1}$	19	3	$f_{18,2}$	79	10	$f_{18,3}$
19	9	50	1	$f_{19,1}$	20	4	$f_{19,2}$	80	9	$f_{19,3}$
20	15	56	2	$f_{20,1}$	20	4	$f_{20,2}$	82	8	$f_{20,3}$
21	15	70	5	$f_{21,1}$	20	4	$f_{21,2}$	91	2	$f_{21,3}$

Table 5.7 – Stack loss data with function values and ranks

The model formulation for the minimal assumption regression model starts with the objective function that should be minimized. This is followed by the constraints, first the additive constraints and then the inequality constraints and finally the range constraints.

$$\text{Minimize} \quad \varepsilon_{1,1} + \varepsilon_{2,1} + \varepsilon_{1,2} + \varepsilon_{2,2} + \cdots + \varepsilon_{1,20} + \varepsilon_{2,20} + \varepsilon_{1,21} + \varepsilon_{2,21} \tag{5.1}$$

$$\text{subject to} \quad f_{1,1} + f_{1,2} + f_{1,3} + \varepsilon_{1,1} - \varepsilon_{2,1} = 42, \tag{5.2}$$

$$f_{2,1} + f_{2,2} + f_{2,3} + \varepsilon_{1,2} - \varepsilon_{2,2} = 37, \tag{5.3}$$

$$f_{3,1} + f_{3,2} + f_{3,3} + \varepsilon_{1,3} - \varepsilon_{2,3} = 37, \tag{5.4}$$

$$f_{4,1} + f_{4,2} + f_{4,3} + \varepsilon_{1,4} - \varepsilon_{2,4} = 28, \tag{5.5}$$

$$\vdots$$

$$f_{21,1} + f_{21,2} + f_{21,3} + \varepsilon_{1,21} - \varepsilon_{2,21} = 15, \tag{5.6}$$

$$f_{15,1} - f_{16,1} = 0, \tag{5.7}$$

$$f_{16,1} - f_{17,1} = 0, \tag{5.8}$$

$$f_{17,1} - f_{18,1} = 0, \tag{5.9}$$

$$f_{18,1} - f_{19,1} = 0, \tag{5.10}$$

$$f_{19,1} - f_{20,1} \leq 0, \tag{5.11}$$

$$\vdots$$

$$f_{1,1} - f_{2,1} = 0, \tag{5.12}$$

$$f_{12,2} - f_{13,2} = 0, \tag{5.13}$$

$$f_{13,2} - f_{10,2} \leq 0, \tag{5.14}$$

$$f_{10,2} - f_{11,2} = 0, \tag{5.15}$$

$$\vdots$$

$$f_{1,2} - f_{2,2} = 0, \tag{5.16}$$

$$f_{7,3} - f_{8,3} = 0, \tag{5.17}$$

$$f_{8,3} - f_{14,3} = 0, \tag{5.18}$$

$$f_{14,3} - f_{21,3} \leq 0, \tag{5.19}$$

$$\vdots$$

$$f_{18,3} - f_{17,3} \leq 0, \tag{5.20}$$

$$\varepsilon_{1i}, \varepsilon_{2i} \geq 0, \qquad \text{for } i = 1, \cdots, 25, \tag{5.21}$$

$$f_{ij} \text{ is unrestricted in sign for all } i \text{ and } j. \tag{5.22}$$

The results obtained from the model (5.1 – 5.22) are recorded in table 5.8. For each unique x value the corresponding $f(x)$ value is given.

x_1	50	56	58	62	70	75	80
f_1	-7	-1	0	2	7	16	16

x_2	17	18	19	20	22	23	24	25	27
f_2	-1	0	1	2	2	2	12	12	12

x_3	93	91	90	89	88	87	86	82	80	79	72
f_3	6	6	9	14	14	14	14	14	14	14	14

Table 5.8 – Stack loss x values and associated function values, $f(x)$

The complete data set is shown in table 5.9. The first column, i, indicates the data point, while the second column reports the y values, followed by the variables, x_j, and the associated function values, $f_j(x_{ij})$. The second last column, $\hat{y}_i$, contains the estimated value which is calculated by adding the function values. The last column represents the absolute deviations, $|y_i - \hat{y}_i|$, which is the difference between the observed and estimated y value.

i	y_i	x_{i1}	$f_1(x_{i1})$	x_{i2}	$f_2(x_{i2})$	x_{i3}	$f_3(x_{i3})$	$\hat{y}_i$	$\lvert y_i - \hat{y}_i \rvert$
1	42	80	16	27	12	89	14	42	0
2	37	80	16	27	12	88	14	42	5
3	37	75	16	25	12	90	9	37	0
4	28	62	2	24	12	87	14	28	0
5	18	62	2	22	2	87	14	18	0
6	18	62	2	23	2	87	14	18	0
7	19	62	2	24	12	93	6	20	1
8	20	62	2	24	12	93	6	20	0
9	15	58	0	23	2	87	14	16	1
10	14	58	0	18	0	80	14	14	0
11	14	58	0	18	0	89	14	14	0
12	13	58	0	17	-1	88	14	13	0
13	11	58	0	17	-1	82	14	13	2
14	12	58	0	19	1	93	6	7	5
15	8	50	-7	18	0	89	14	7	1
16	7	50	-7	18	0	86	14	7	0
17	8	50	-7	19	1	72	14	8	0
18	8	50	-7	19	1	79	14	8	0
19	9	50	-7	20	2	80	14	9	0
20	15	56	-1	20	2	82	14	15	0
21	15	70	7	20	2	91	6	15	0

Table 5.9 – Data, function values and residuals

Figure 5.1 depicts the absolute deviation for each data point graphically. In this figure, data points 2 and 14 exhibit the largest deviation, followed by point 13. This deviation could be due to possible outliers.

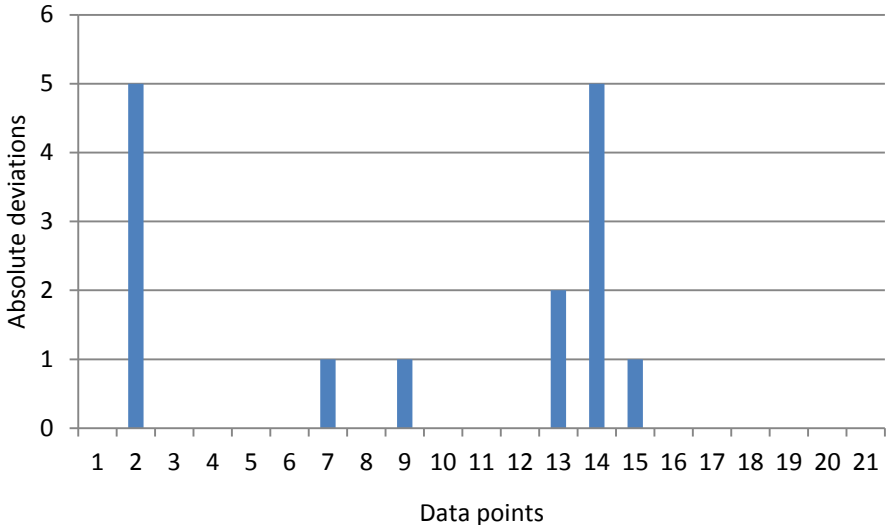

Figure 5.1 – The absolute deviation for each data point (Stack loss data)

In figures 5.2 to 5.4 the function values of f_1, f_2 and f_3 are plotted against the x_1, x_2 and x_3 values respectively. In the first two figures (5.2 and 5.3) it is easy to see that the functions are monotonically non-decreasing, as specified by the multiple regression performed beforehand and the inequality constraints in the model. The third figure (5.4) depicts the monotonically non-increasing function.

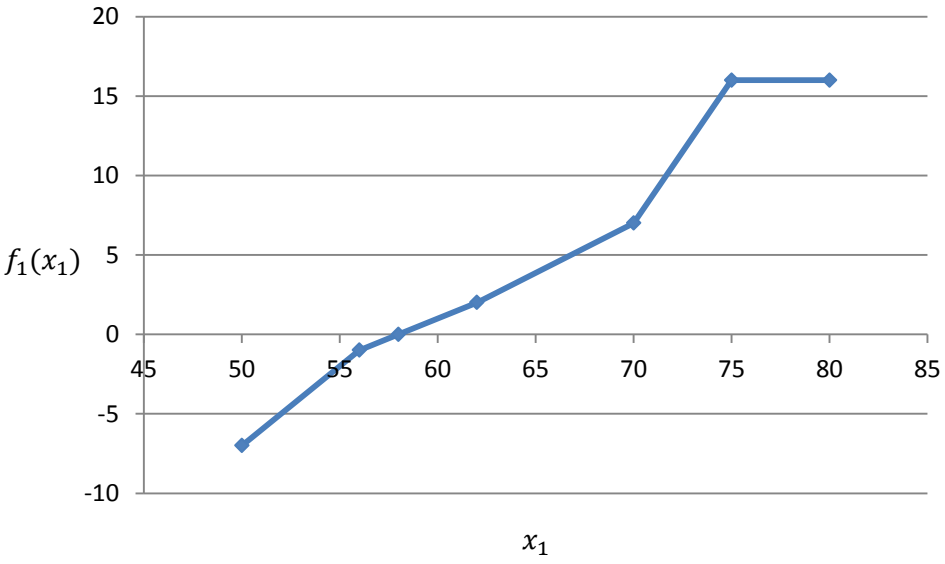

Figure 5.2 – f_1 values plotted against x_1 values (Stack loss data)

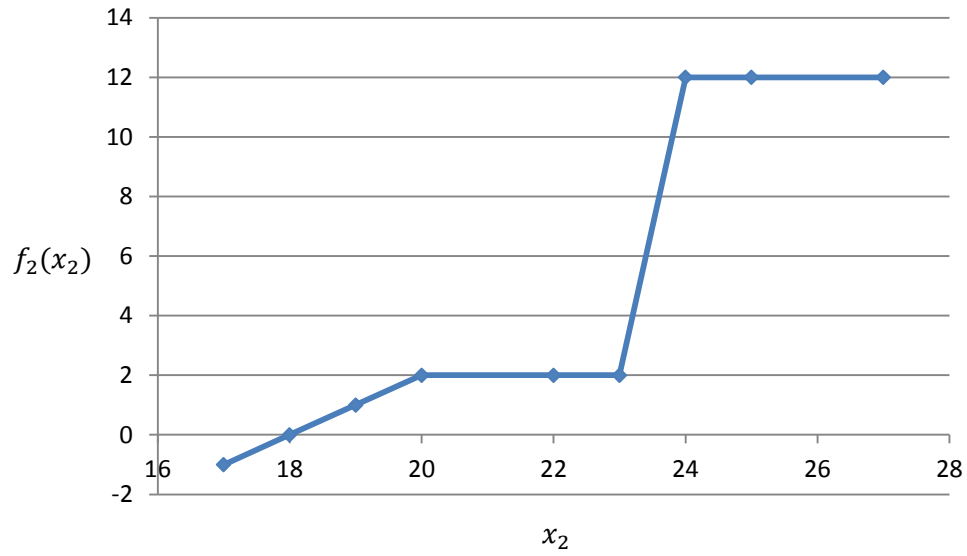

Figure 5.3 – f_2 values plotted against x_2 values (Stack loss data)

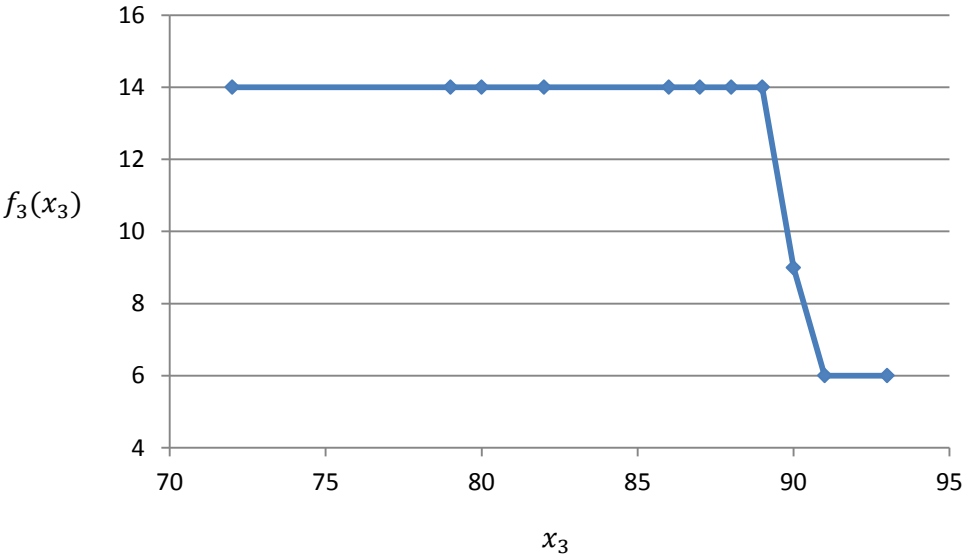

Figure 5.4 – f_3 values plotted against x_3 values (Stack loss data)

This concludes the application of the minimal assumption regression model to the stack loss data. The next steps will be to apply the various extensions introduced in Chapter 4. The first extension is to omit data points. To identify possible outliers, two other variables are introduced into the model, a binary variable, z_i, and an unrestricted slack variable, α_i, for $i = 1, \dots, n$. The slack variable is added to the additive constraints for each data point such that for data point 1 the extended additive constraint is given by

$$f_{1,1} + f_{1,2} + f_{1,3} + \varepsilon_{1,1} - \varepsilon_{2,1} - \alpha_1 = 42,$$

while the slack variable is constrained as follows

$$-Mz_1 \leq \alpha_1 \leq Mz_1,$$

where M is a large number, so that if $z_1 = 0$, α_1 is also constrained to 0 and there are no changes in the additive constraint. If $z_1 = 1$, α_1 can have any value and because α_1 is not in the objective function, it is able to take up the slack between y_1 and $\hat{y}_1$ instead of $\varepsilon_{1,1}$ or $\varepsilon_{2,1}$ taking up the slack.

In this manner a set of points that would realize the greatest decrease in the objective function will be omitted.

As explained in Chapter 4 (section 4.2.2.1), to determine how many points to omit some experimentation must be done. In figure 5.5 the number of points omitted, p, are plotted against the respective optimal objective values. The values in brackets indicate the data points that are omitted. When more than 2 points are omitted there is no longer such a large decrease in the objective function. Therefore, in this case 2 data points (points 2 and 14) are omitted.

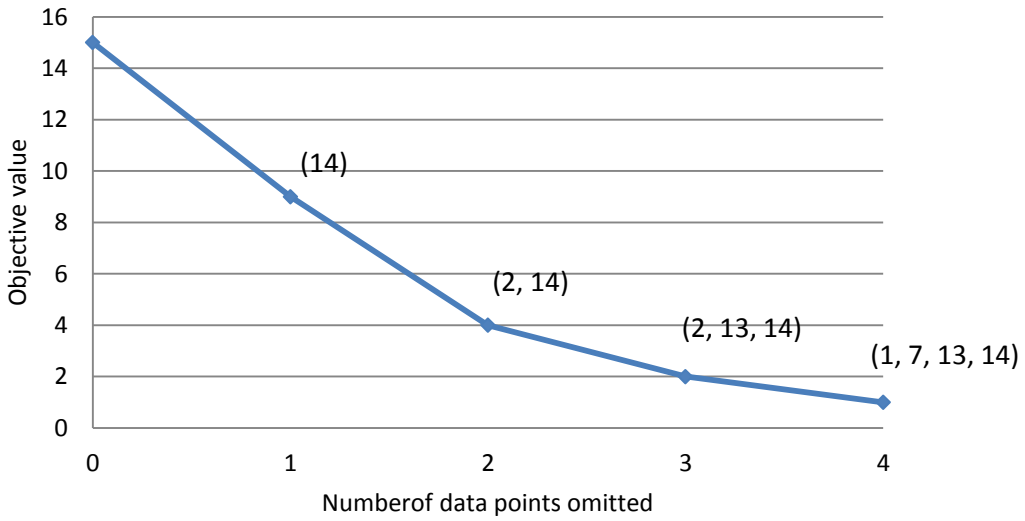

Figure 5.5 – Number of data points to omit (Stack loss data)

To prevent the model from the possibility of overfitting, the second extension (smoothing of factors) are incorporated into the model. Section 4.2.3.2 described how to choose a smoothing factor. The constrained second derivative is used to limit the change in the slope of the function by the smoothing factor β.

The influence of the smoothing factor on the form of a function is illustrated in figure 5.6. The functions of the minimal assumption regression model are shown, indicated as $f_1(x_1)$, $f_2(x_2)$ and $f_3(x_3)$. The smoothed functions with $\beta = 10$ are also depicted and are indicated as $f_1'(x_1)$, $f_2'(x_2)$ and $f_3'(x_3)$. In the case of functions $f_2(x_2)$ and $f_3(x_3)$, the smoothing effects are more pronounced.

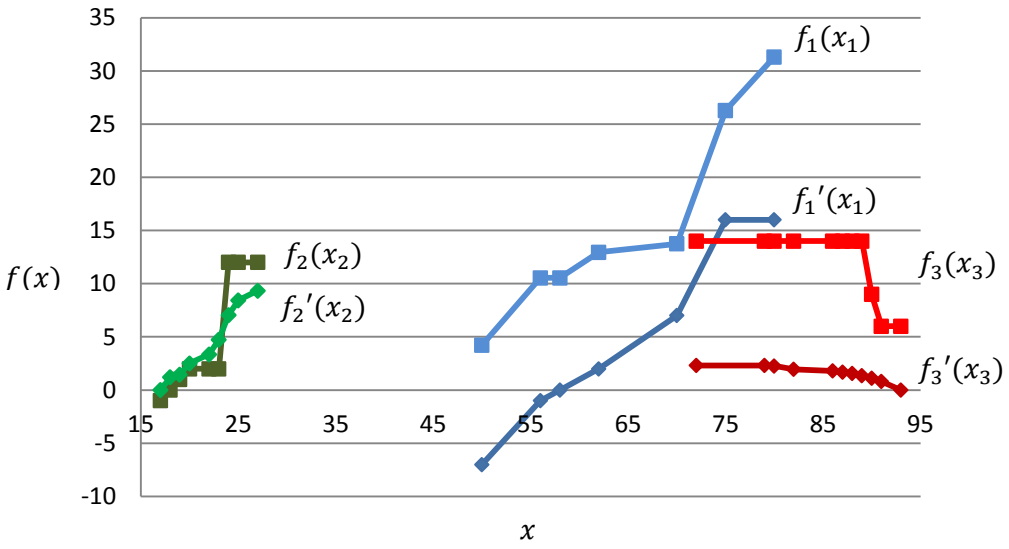

Figure 5.6 – Smoothing effect with β = 10 (Stack loss data)

The smoothing factor can be determined by the following steps:

- using the jack-knife approach;
- determining the mean absolute deviation for different values of β;
- evaluating the solutions by comparing the different mean absolute deviations; and
- choosing the solution with the β value that produce the smallest mean absolute deviation.

In figure 5.7 the value of β is incremented by 25 from 25 to 200. It is clear that $\beta = 50$ gives the lowest mean absolute deviation and it is therefore assumed that a β value of 50 will yield the best results for the stack loss data set.

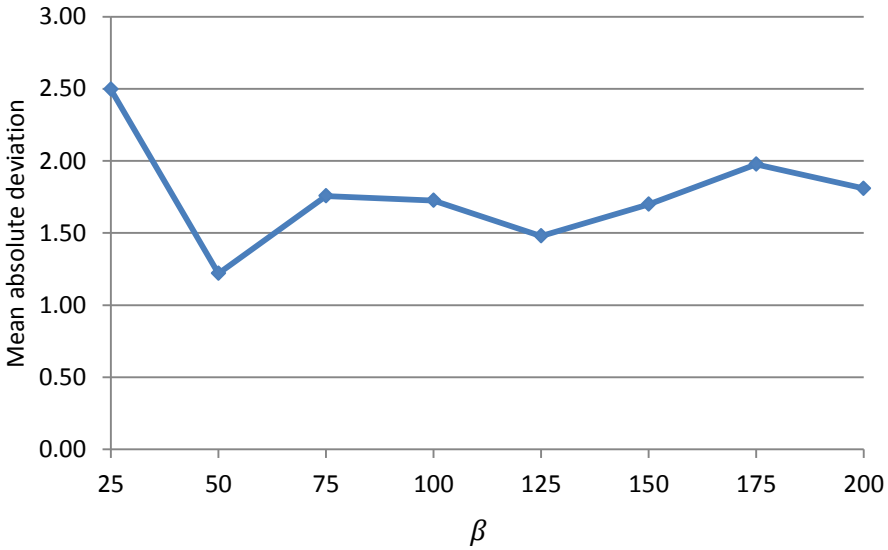

Figure 5.7 – Mean absolute deviation for different values of β (Stack loss data)

In figure 5.8 the original function and the smoothed function are shown for x_1 with $\beta = 50$. The other functions, not shown here, follow in the same way.

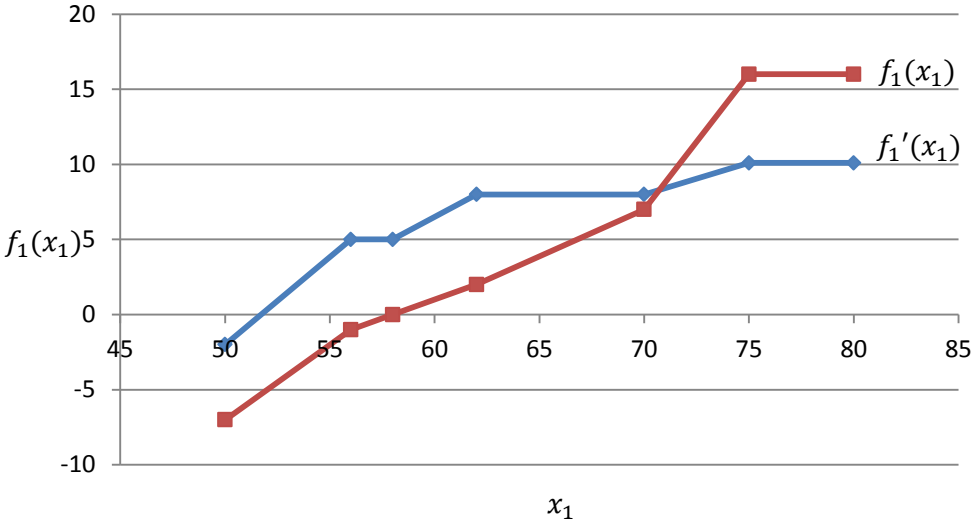

Figure 5.8 – Change in function f_1 after smoothing with β = 50 (Stack loss data)

Before the results of the extensions to the minimal assumption regression model are discussed, the piecewise linear regression model, introduced in section 4.3, will be applied to the stack loss data set. Piecewise linear regression with no breakpoints is equivalent to least absolute deviation regression (L_1-norm) where only one function is specified for the whole range of x. With one breakpoint two linear models are specified, one for the x values smaller than the 50th percentile and the other model for the x values greater than the 50th percentile. When two breakpoints are used these are specified as the 33rd and 67th percentile. For this study the breakpoints are chosen in this manner, but other choices are possible. Figure 5.9 illustrates how the function of variable x_2 changes with zero, one, and two breakpoints. The other functions, not shown here, follow in the same way.

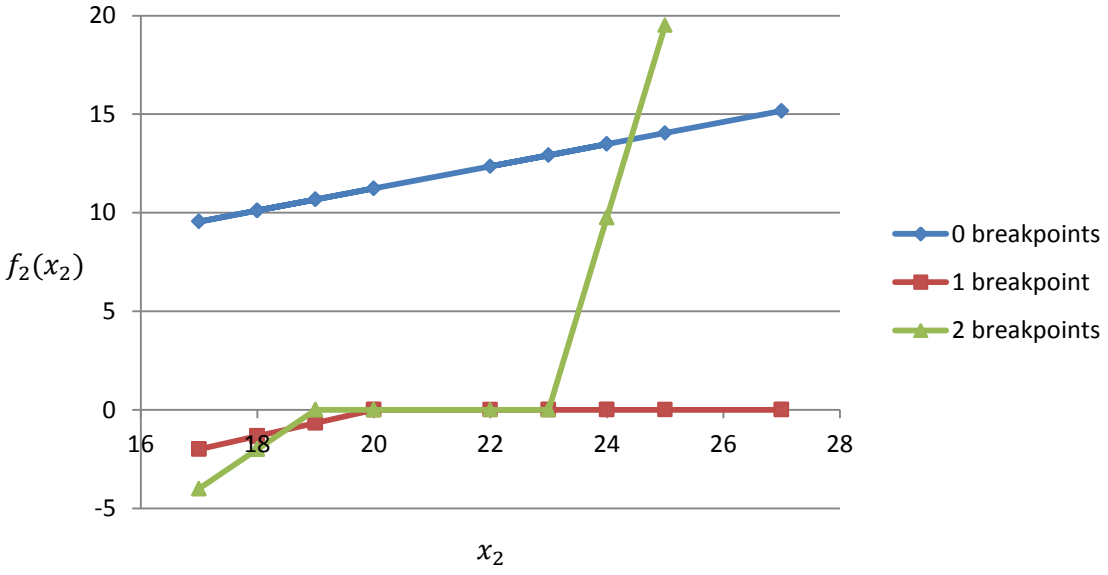

Figure 5.9 – $f_2(x_2)$ with zero, one and two breakpoints (Stack loss data)

When a model is developed its effectiveness must be tested. In this study the models are tested by their predictive accuracy and the mean absolute deviation measurement is used for comparing different models.

Consider table 5.10 which reports the comparison for L_2-norm regression, L_1-norm regression and the original minimal assumption regression model introduced by Wagner. No extensions (outlier detection and smoothing) were implemented and it can be noted from the mean absolute deviation that the original Wagner model outperformed L_2-norm regression. The mean absolute deviation of the L_1-norm regression is only 0.042 (or about 2%) less than the mean absolute deviation of Wagner's model.

Model	**Mean absolute deviation**
L_2-norm regression	2.887
L_1-norm regression	2.035
Original minimal assumption regression model (Wagner, 1962)	2.077

Table 5.10 – Comparison of models that do not omit outliers (Stack loss data)

In table 5.11 the results of the proposed extensions and the piecewise linear model are given. In all cases 2 data points (outliers) were omitted. As stated earlier, points 2 and 14 were omitted. By introducing a smoothing factor of $\beta = 50$ the minimal assumption regression model outperformed all the other models with a mean absolute deviation of 1.220. The extensions added to the minimal assumption regression model decreased the mean absolute deviation by 41% (from 2.077 to 1.220). This value is also 12% less than the second lowest mean absolute deviation of the L_1-norm regression model which omits 2 data points with no breakpoints.

Model	**Mean absolute deviation**
Wagner's model omitting 2 data points	2.194
Wagner's model omitting 2 data points and a smoothing factor of $\beta = 50$	1.220
L_1-norm regression omitting 2 data points and no breakpoint	1.394
Piecewise L_1-norm regression omitting 2 data points and one breakpoint	1.882
Piecewise L_1-norm regression omitting 2 data points and two breakpoints	2.150

Table 5.11 – Comparison of models that omit outliers (Stack loss data)

The same steps and methods that were used for examining this data set will be employed for the other data sets in the results that follow, though the next sections will only contain explanatory graphs and results.

5.3.2 Scottish hill racing

The functions of the two variables (figures 5.10 and 5.11) of the Scottish hill race data set are both monotonic non-decreasing because the longer the race and the higher the climb, the longer one can expect the record time to be. This was confirmed by regression analysis performed on the original data.

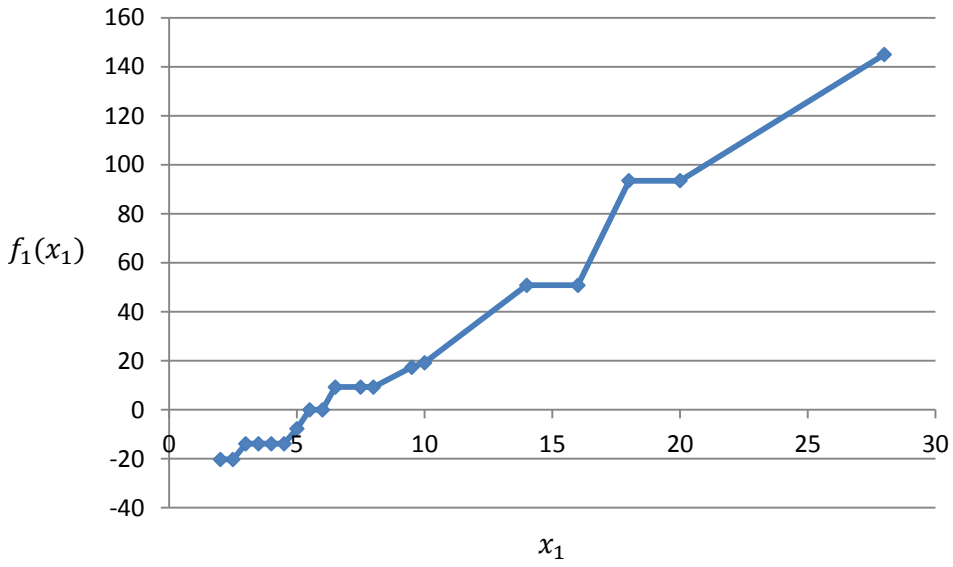

Figure 5.10 – f_1 values plotted against x_1 values (Scottish hill race data)

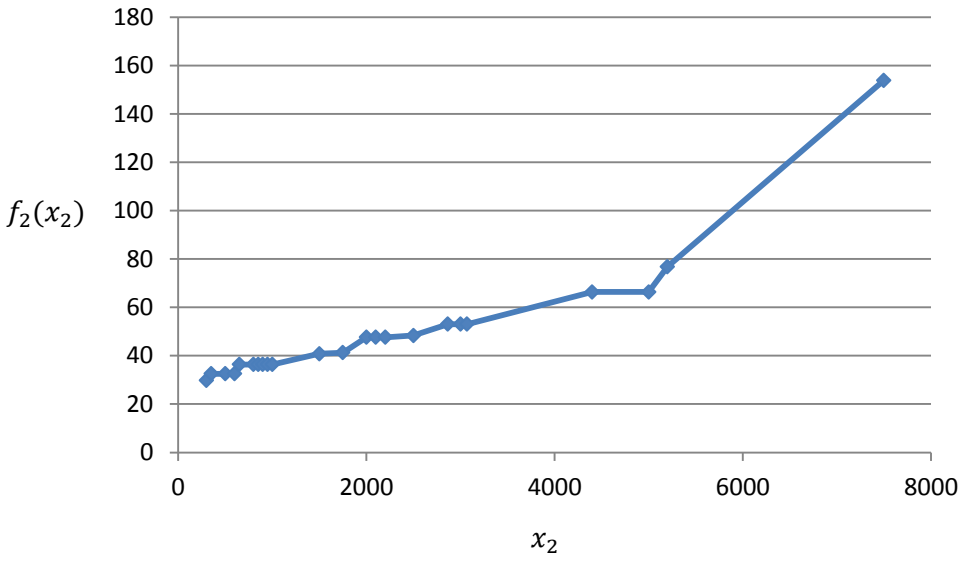

Figure 5.11 – f_2 values plotted against x_2 values (Scottish hill race data)

In figure 5.12 the objective values are plotted against the number of data points to be omitted. There is a significant drop in the objective function when data point 18 is omitted. After the omission of data point 18, a constant decrease in the objective function value can be observed (figure 5.12). According to Atkinson (1988) data point 18 is indeed an outlier because the time was incorrectly recorded; instead of 1 hour, 18 minutes, 39 seconds it should only be 18 minutes, 39 seconds. This graph does not clearly indicate how many points to omit and it was decided to omit 4 data points, which represent about ten percent of the data. The four data points omitted by the model were data points 6, 12, 14 and 18.

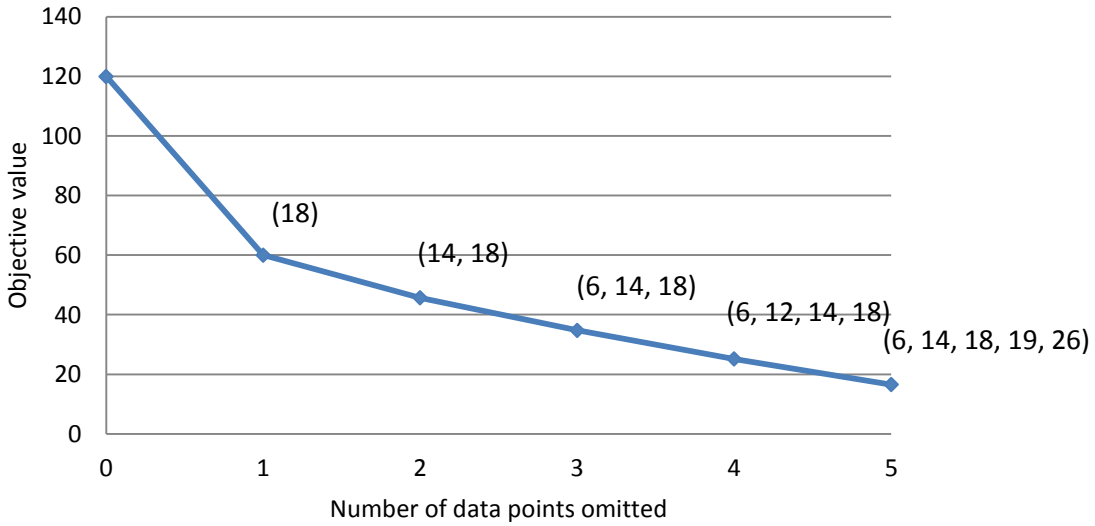

Figure 5.12 – Number of data points omitted (Scottish hill race data)

To determine the smoothing parameter, the mean absolute deviation is plotted against different values of β. Figure 5.13 indicates that a smoothing factor of $\beta = 10$ results in the lowest mean absolute deviation value.

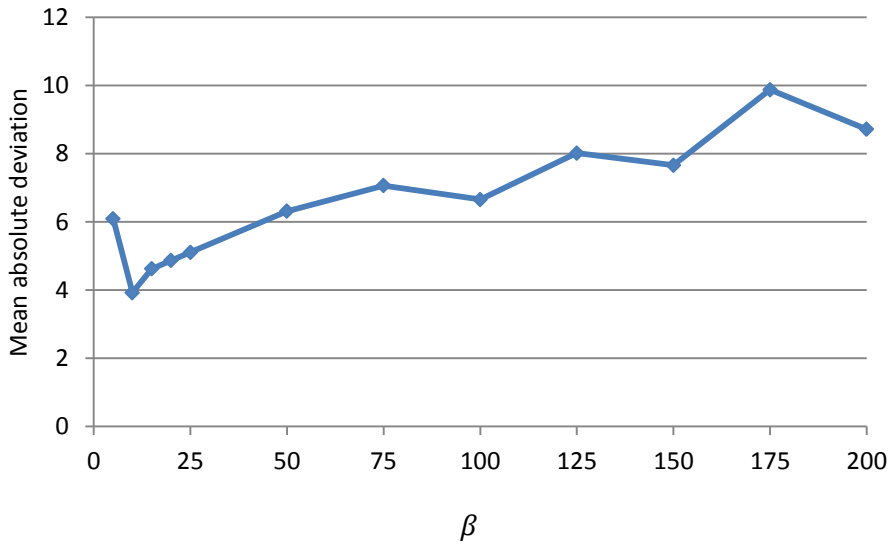

Figure 5.13 – Mean absolute deviation for different values of β (Scottish hill race data)

Figure 5.14 illustrates the change in the function of the second variable, x_2, from 0 to 2 breakpoints. Although it does not seem as if the functions alter much, there is quite a large decrease in the mean absolute deviation value from 0 to 1 breakpoints (see table 5.13). The change in the function of the first variable, x_1, follows in the same manner.

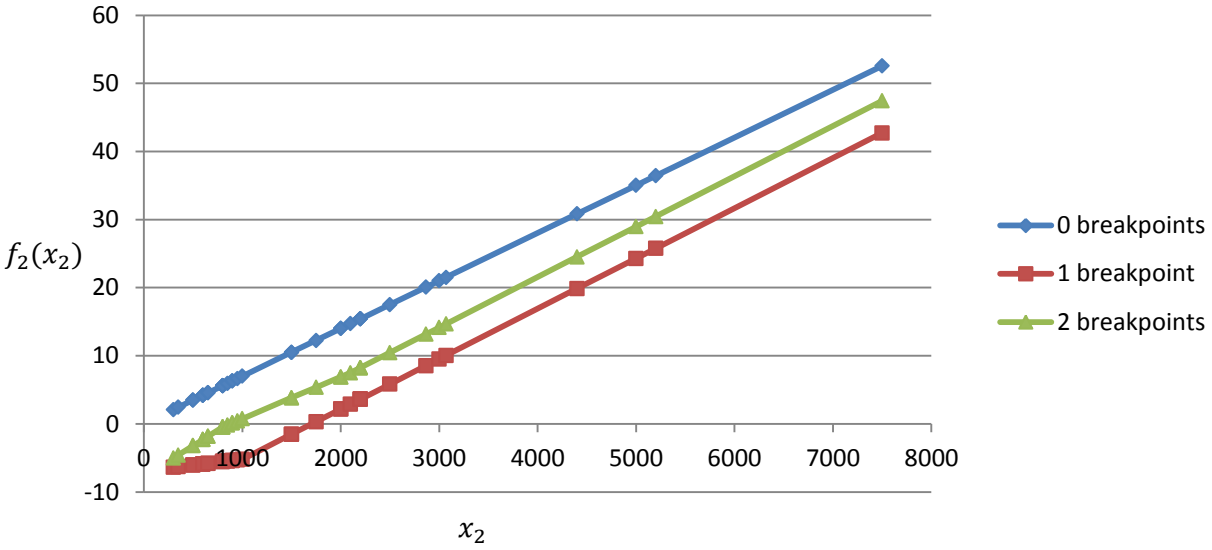

Figure 5.14 – $f_2(x_2)$ with zero, one and two breakpoints (Scottish hill race data)

By comparing the different models without any extensions (table 5.12), it can be noted that the mean absolute deviation of the minimal assumption regression model lies between the values of the mean absolute deviation of the L_2-norm and L_1-norm regression models.

Model	Mean absolute deviation
L_2-norm regression	9.367
L_1-norm regression	8.211
Original minimal assumption regression model (Wagner, 1962)	8.927

Table 5.12 – Comparison of models that do not omit outliers (Scottish hill race data)

The effects of including the proposed extensions are quite remarkable. Consider table 5.13. In all cases 4 data points (outliers) were omitted. By introducing a smoothing factor of $\beta = 10$, the mean absolute deviation decreases from 8.927 to 3.921, a reduction of 56%. The piecewise L_1-norm regression model with one breakpoint outperformed all the other models with a mean absolute deviation of 3.559. This value is about 9.2% less than the mean absolute deviation of the minimal assumption regression model with extensions (3.921).

Model	Mean absolute deviation
Wagner's model omitting 4 data points	8.469
Wagner's model omitting 4 data points and a smoothing factor of $\beta = 10$	3.921
L_1-norm regression omitting 4 data points and no breakpoint	4.253
Piecewise L_1-norm regression omitting 4 data points and one breakpoint	3.559
Piecewise L_1-norm regression omitting 4 data points and two breakpoints	4.280

Table 5.13 – Comparison of models that omit outliers (Scottish hill race data)

5.3.3 Weisberg fuel consumption

Figures 5.15 to 5.18 depict the functions of the four variables in this data set: the first three variables are monotonically non-increasing whereas the last one is monotonically non-decreasing; this has been estimated by carrying out regression analysis beforehand.

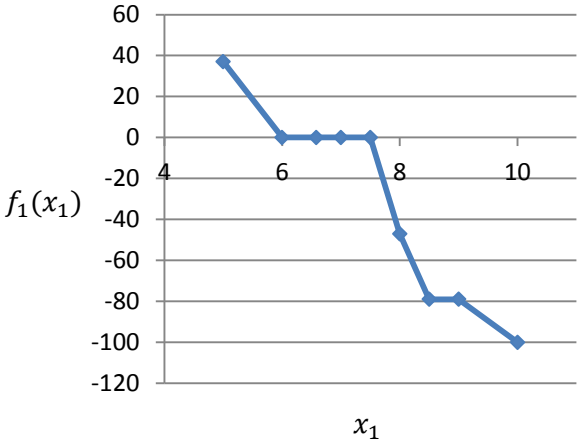

Figure 5.15 – f_1 values plotted against x_1 values (Weisberg fuel data)

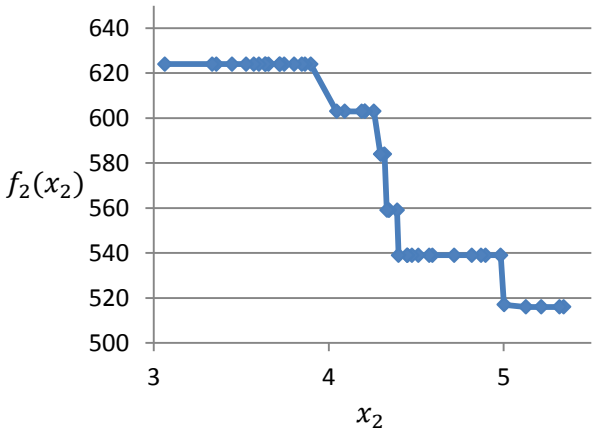

Figure 5.16 – f_2 values plotted against x_2 values (Weisberg fuel data)

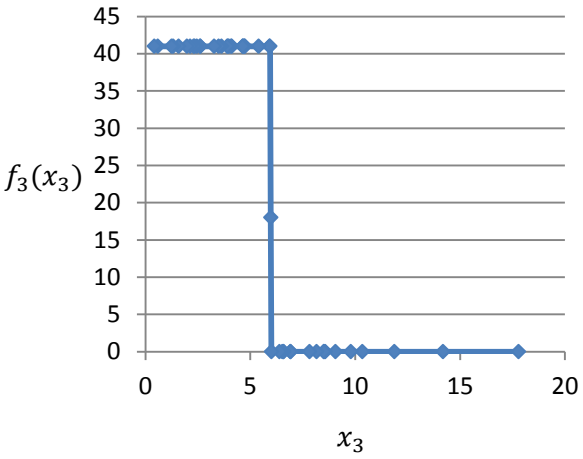

Figure 5.17 – f_3 values plotted against x_3 values (Weisberg fuel data)

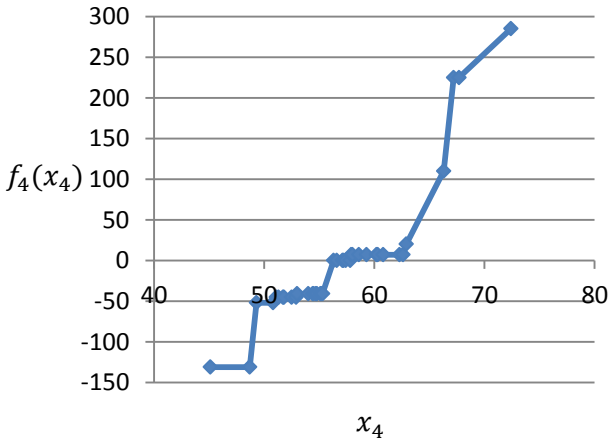

Figure 5.18 – f_4 values plotted against x_4 values (Weisberg fuel data)

It is difficult to determine the number of data points to be omitted for this data set. By looking at figure 5.19, it appears that there is a slight change in the slope of the line at 6 data points to be omitted. 6 points constitute about ten percent of the data and therefore this number is chosen as the amount of data points to be omitted. The data points omitted are points 4, 5, 18, 20, 38 and 40.

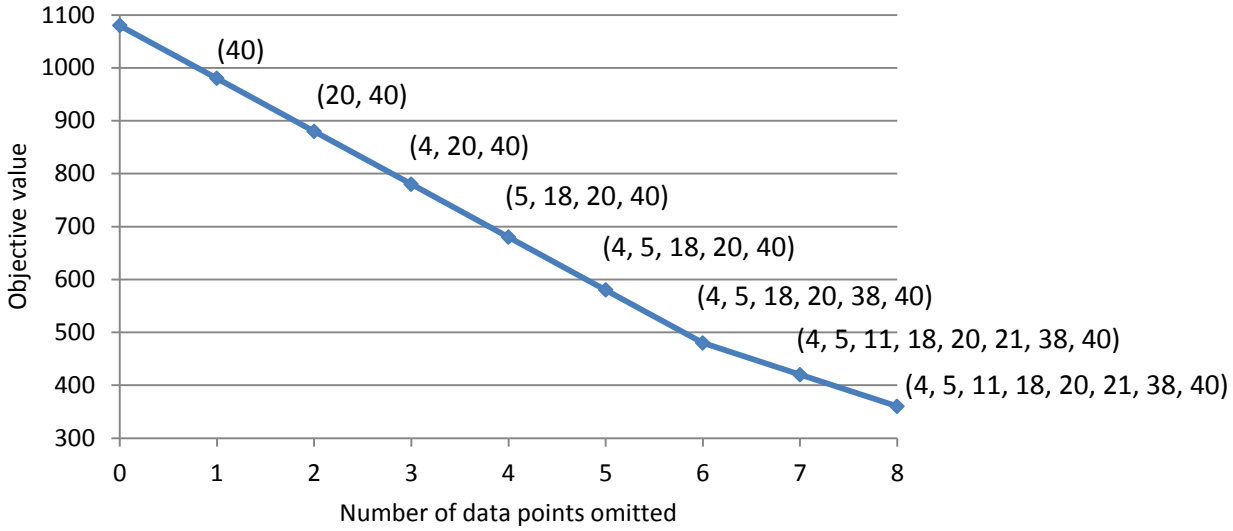

Figure 5.19 – Number of data points omitted (Weisberg fuel data)

Figure 5.20 indicates that $\beta = 1$ is a good smoothing parameter, and the next two figures (figures 5.21 and 5.22) exemplify the change in the forms of the functions of x_3 and x_4 when a smoothing factor of $\beta = 1$ is implemented. The changes in the forms of the other functions are not portrayed here, but follow in the same way.

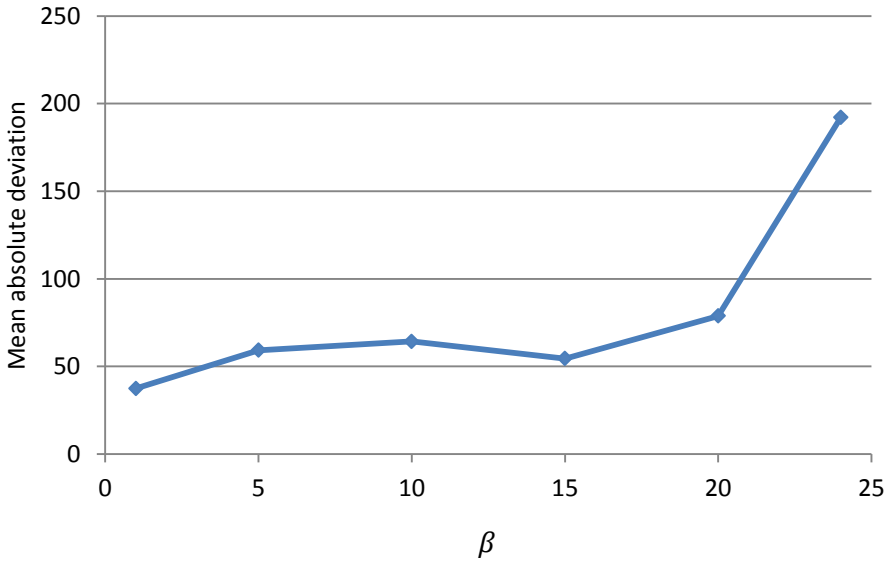

Figure 5.20 – Mean absolute deviation for different values of β (Weisberg fuel data)

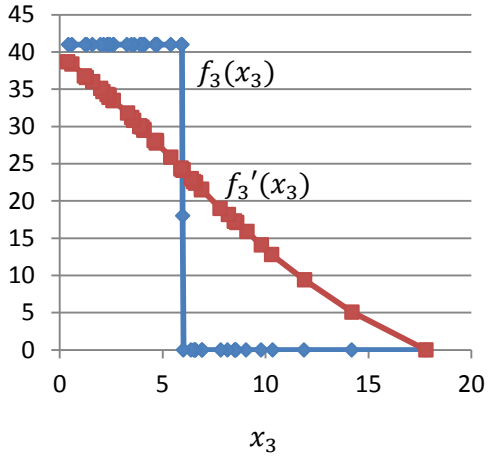

Figure 5.21 – Smoothing effect with β = 1 (Weisberg fuel data)

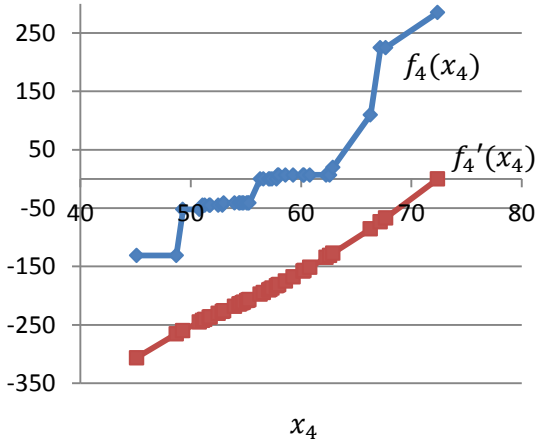

Figure 5.22 – Smoothing effect with β = 1 (Weisberg fuel data)

The last figure before the results are shown, figure 5.23, furnishes a good example of piecewise linear regression. It demonstrates how the function form changes as the breakpoints increase from zero to two. This figure represents only the change in the function form for the third variable, x_3.

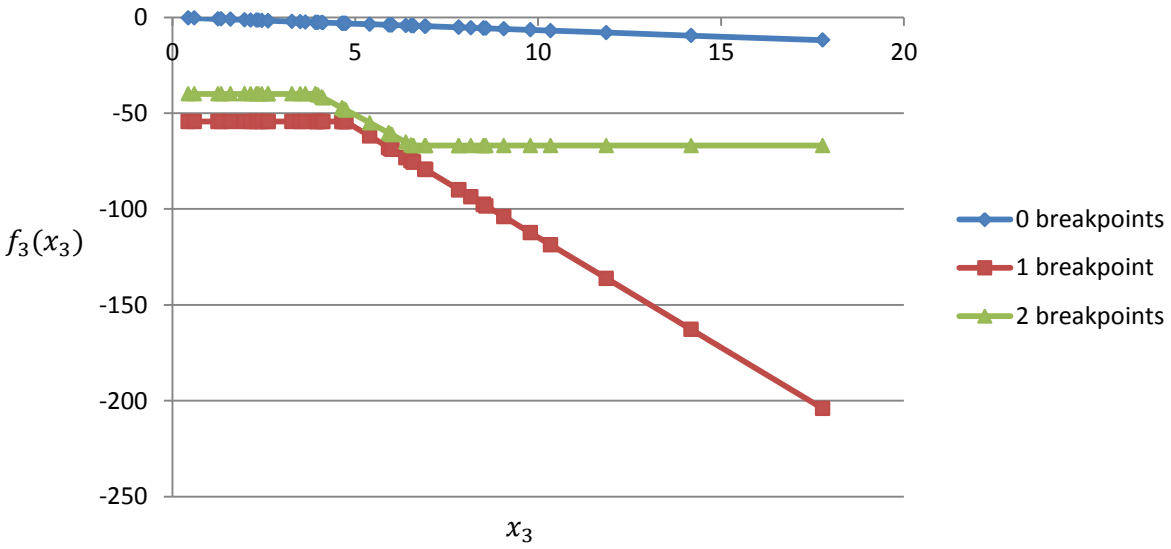

Figure 5.23 – $f_3(x_3)$ with zero, one and two breakpoints (Weisberg fuel data)

Model	Mean absolute deviation
L_2-norm regression	54.532
L_1-norm regression	49.466
Original minimal assumption regression model (Wagner, 1962)	52.726

Table 5.14 – Comparison of models that do not omit outliers (Weisberg fuel data)

Table 5.14 contains the comparison of the L_2-norm regression, L_1-norm regression and the original minimal assumption regression model. No extensions (outlier detection and smoothing) were implemented and the mean absolute deviation of the original Wagner model lies between the mean absolute deviation values of the L_2-norm and L_1-norm regression model respectively.

By omitting 6 data points (table 5.15) it is evident that the mean absolute deviation decreases considerably. It does not seem that a smoothing factor of $\beta = 1$ improves the mean absolute deviation. The piecewise L_1-norm regression model with one breakpoint outperformed all the other models with a mean absolute deviation of 33.293. This is only slightly better (about 1%) than the mean absolute deviation of the minimal assumption regression model omitting 6 data points (33.639). In this case the models all give comparable results.

Model	Mean absolute deviation
Wagner's model omitting 6 data points	33.639
Wagner's model omitting 6 data points and a smoothing factor of $\beta = 1$	37.425
L_1-norm regression omitting 6 data points and no breakpoint	35.426
Piecewise L_1-norm regression omitting 6 data points and one breakpoint	33.293
Piecewise L_1-norm regression omitting 6 data points and two breakpoints	44.320

Table 5.15 – Comparison of models that omit outliers (Weisberg fuel data)

5.3.4 Gross national product (GNP)

The functions of the seven variables that were chosen for this data set are illustrated in figures 5.24 to 5.30. Functions 1, 4, 6 and 7 are monotonically non-decreasing while the rest of the functions are monotonically non-increasing. This was confirmed by regression analysis performed on the original data.

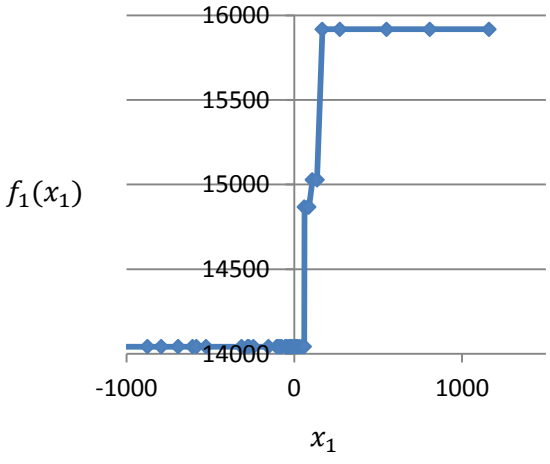

Figure 5.24 – f_1 values plotted against x_1 values (GNP data)

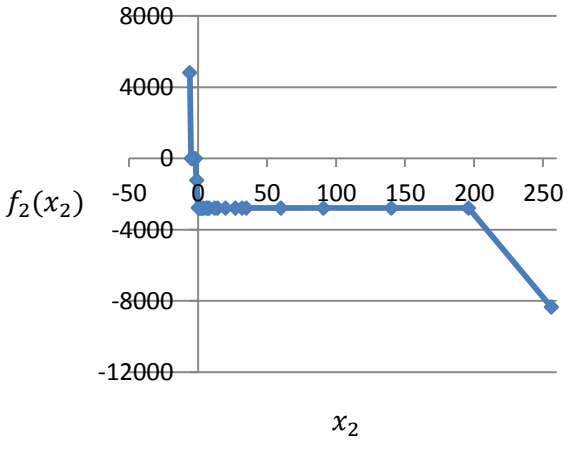

Figure 5.25 – f_2 values plotted against x_2 values (GNP data)

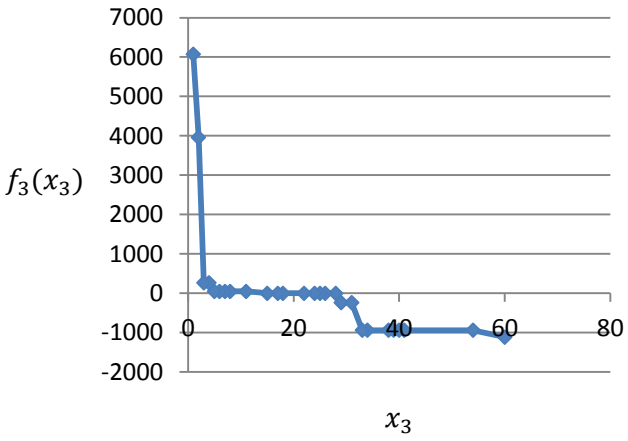

Figure 5.26 – f_3 values plotted against x_3 values (GNP data)

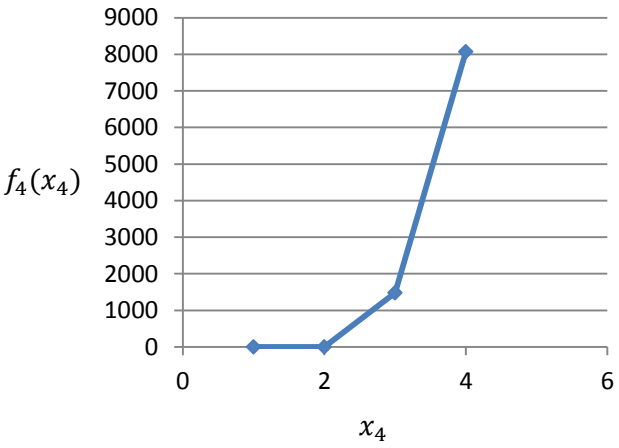

Figure 5.27 – f_4 values plotted against x_4 values (GNP data)

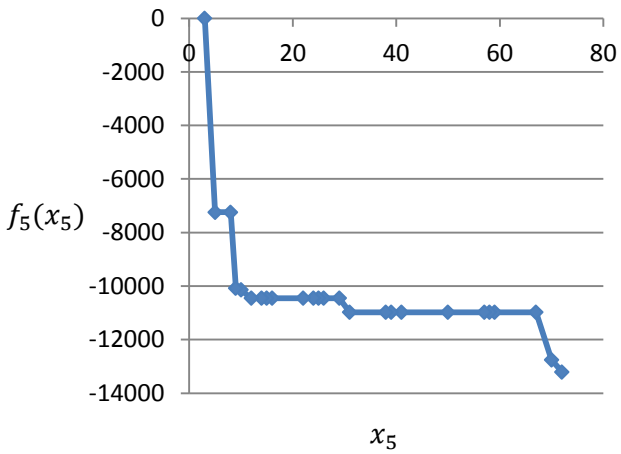

Figure 5.28 – f_5 values plotted against x_5 values (GNP data)

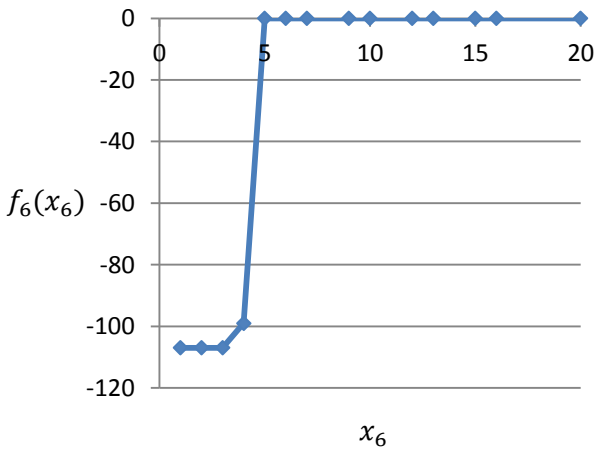

Figure 5.29 – f_6 values plotted against x_6 values (GNP data)

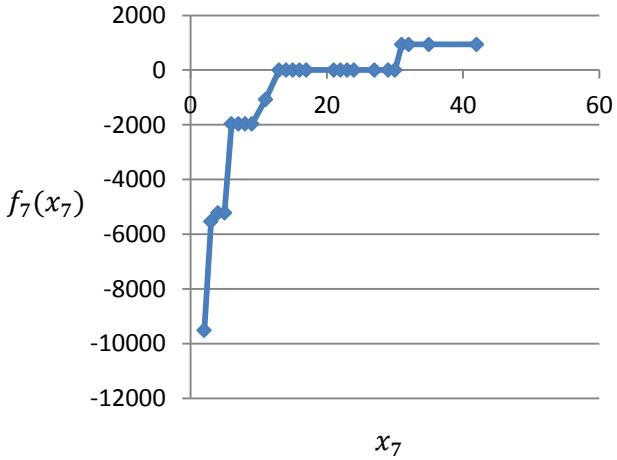

Figure 5.30 – f_7 values plotted against x_7 values (GNP data)

There is a constant decrease in the objective value (figure 5.31); therefore ten percent of the data points, which is equal to about 4 points, is chosen as the number of data points to be omitted. The points that were omitted are points 9, 26, 39 and 43.

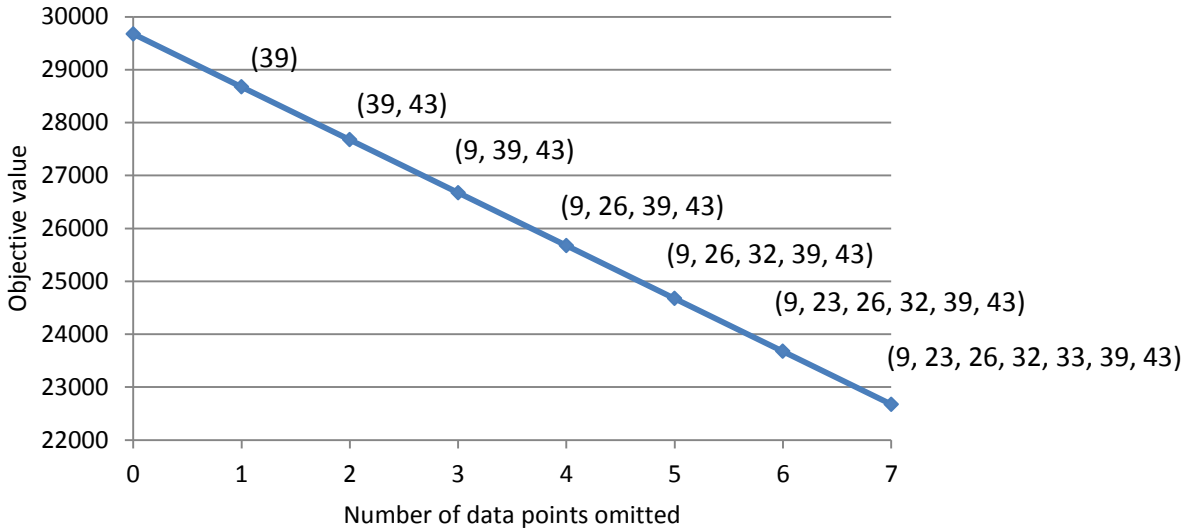

Figure 5.31 – Number of data points omitted (GNP data)

Figure 5.32 depicts the mean absolute deviation against different values of β. The smoothing factor is chosen as $\beta = 200$, because it yields the lowest mean absolute deviation.

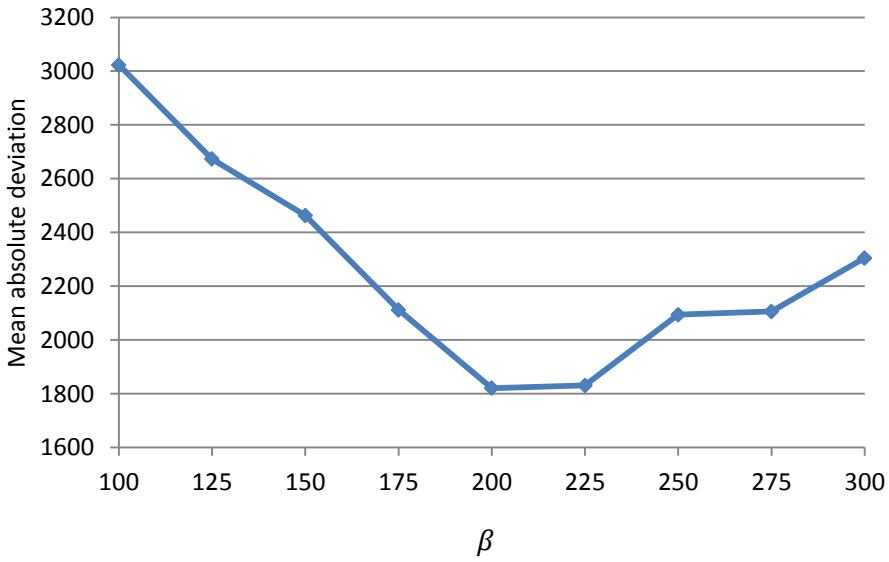

Figure 5.32 – Mean absolute deviation for different values of β (GNP data)

Once again the models that do not omit outliers are compared. Table 5.16 makes clear that Wagner's model lies between L_2-norm and L_1-norm regression, but is much closer to the lower mean absolute deviation of the L_1-norm regression model.

Model	Mean absolute deviation
L_2-norm regression	4038.425
L_1-norm regression	3215.116
Original minimal assumption regression model (Wagner, 1962)	3282.140

Table 5.16 – Comparison of models that do not omit outliers (GNP data)

All the models in table 5.17 omit 4 data points. By adding a smoothing factor of $\beta = 200$ the mean absolute deviation decreases by almost 45%, from 3282.14 to 1820.18. This value is only 6% bigger than the lowest mean absolute deviation of the piecewise L_1-norm regression model with two breakpoints, which is 1713.885.

Model	Mean absolute deviation
Wagner's model omitting 4 data points	2764.669
Wagner's model omitting 4 data points and a smoothing factor of $\beta = 200$	1820.180
L_1-norm regression omitting 4 data points and no breakpoint	2360.977
Piecewise L_1-norm regression omitting 4 data points and one breakpoint	2782.106
Piecewise L_1-norm regression omitting 4 data points and two breakpoints	1713.885

Table 5.17 – Comparison of models that omit outliers (GNP data)

5.3.5 Financial ratios

In this data set the mean absolute deviation does not seem to be improved by the omission of outliers or the smoothing of the functions. To increase the number of breakpoints from zero to two does prove to lower the mean absolute deviation, but in this case the original Wagner model seems to give the best results.

Model	Mean absolute deviation
L_2-norm regression	0.3228
L_1-norm regression	0.3297
Original minimal assumption regression model (Wagner, 1962)	0.0531
Piecewise L_1-norm regression with one breakpoint	0.1893
Piecewise L_1-norm regression with two breakpoints	0.0965

Table 5.18 – Comparison of models that do not omit outliers (Financial ratios data)

In table 5.18 the mean absolute deviation values for models which do not omit outliers or incorporate smoothing factors are evaluated. The original minimal assumption regression model obtained the lowest mean absolute deviation followed by the piecewise L_1-norm regression model with two breakpoints.

Figure 5.33 depicts the function form of variable x_1 for Wagner's model without extensions while figure 5.34 illustrates the piecewise L_1-norm regression with two breakpoints.

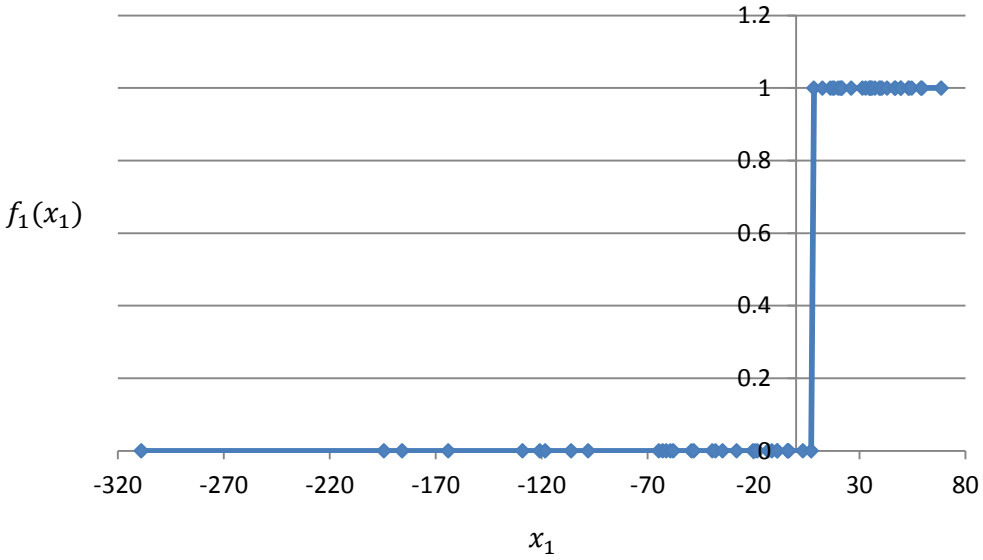

Figure 5.33 – f_1 values plotted against x_1 values (Financial ratios data)

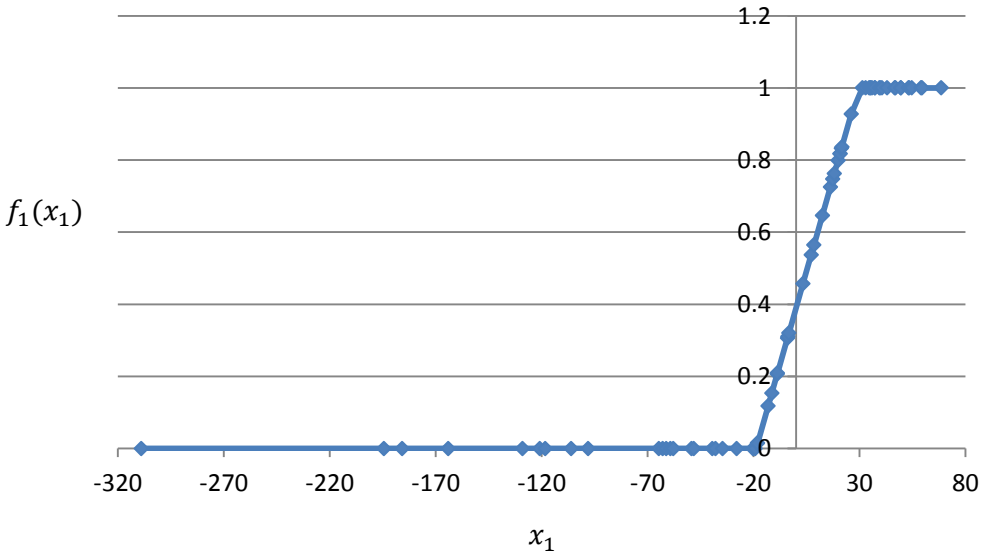

Figure 5.34 – $f_1(x_1)$ with two breakpoints (Financial ratios data)

5.4 Specific cases

In the previous section the mean prediction errors using absolute deviations for the different models were compared. L_2-norm, L_1-norm regression and the minimal assumption regression model without any extensions were compared. The other models, when outliers are omitted and a smoothing factor is incorporated into the minimal assumption regression model, can be compared with L_1-norm regression that omits outliers and piecewise L_1-norm regression with one, two or three breakpoints; each of these models omits some data points.

The monotonicity constraints, specified by Wagner's minimal assumption regression model, should enable this model to fit a specific kind of non-linear function better than other models. To determine whether there are some specific cases where the minimal assumption regression model may outperform the other models, data sets with certain specific features were simulated. A range of data sets were simulated and tested: in sections 5.4.1 – 5.4.3 three cases will be presented.

The simulation of a data set was carried out as follows:

- Choose a dimension for the data;
- Randomly select some x values and sort it in an increasing or decreasing order for each decision variable;
- Determine the function values for a specific range of the x values for each variable by a straight line equation $f = mx + c$, for example $f = 13x + 40$. The equation can be different for each decision variable;
- For the rest of the x values the function values are equal to a randomly chosen constant, for example 200. The value can be different for each decision variable;
- Randomly add or subtract a certain error from the function values; and
- Calculate the y value of each data point by adding the function values of each variable for each data point.

All random selections were based on a discrete probability distribution for the specific selection. After the data is simulated the different models are used and the prediction capability is tested. The mean absolute deviation values for each model are compared to evaluate how each of these models performs with the specific set of data.

Table 5.19 constitutes an example of a simulated data set of 30 points. The function values are first determined by an equation or a chosen constant value after which a random error is added or subtracted. The dependent variable, y, is determined by adding the function values for each data point. Take, for example, data point 10. The value for $x_{10,1}$ was selected as 19. The equation that was used to determine $f_{10,1}$ is $f_{10,1} = 13x + 40$; therefore $f_{10,1} = 287$ before an

error value is added or subtracted. In this case a random error of 11.48 was added to $f_{10,1}$, and the final value is equal to 298.48. Another equation was used in the same way to determine $f_{10,2}$. The value of the dependent variable, y_{10}, was determined by adding the values of $f_{10,1}$ and $f_{10,2}$.

i	1	2	3	4	5	6	7	8	9	10
y_i	1402.36	1465.56	1473.70	1496.64	1528.90	1698.60	1714.28	1851.20	1901.00	2020.48
x_{i1}	4	4	8	11	11	13	13	15	17	19
f_{i1}	195	202.80	195	187.20	198.90	209	209	209	261	298.48
x_{i2}	200	201	219	222	228	248	256	263	268	268
f_{i2}	1207.36	1262.76	1278.70	1309.44	1330.00	1489.60	1505.28	1642.20	1640.00	1722.00

i	11	12	13	14	15	16	17	18	19	20
y_i	1949.76	2005.46	2063.78	2337.92	2398.12	2407.00	2405.28	2573.01	2541.64	2551.11
x_{i1}	21	23	23	24	24	25	26	29	29	29
f_{i1}	300.48	339	345.78	337.92	352	365	393.12	404.49	417	429.51
x_{i2}	281	281	281	328	329	335	344	349	356	363
f_{i2}	1649.28	1666.46	1718.00	2000.00	2046.12	2042.00	2012.16	2168.52	2124.64	2121.60

i	21	22	23	24	25	26	27	28	29	30
y_i	2772.20	2946.00	3134.60	3145.00	3159.85	3126.08	3173.60	3159.40	3178.24	3156.38
x_{i1}	30	30	33	35	35	36	36	37	38	38
f_{i1}	430	430	469	495	509.85	487.68	508	521	512.64	517.98
x_{i2}	393	414	448	448	451	453	464	478	485	496
f_{i2}	2342.20	2516.00	2665.60	2650.00	2650.00	2638.40	2665.60	2638.40	2665.60	2638.40

Table 5.19 – Example of simulated data

5.4.1 Case 1

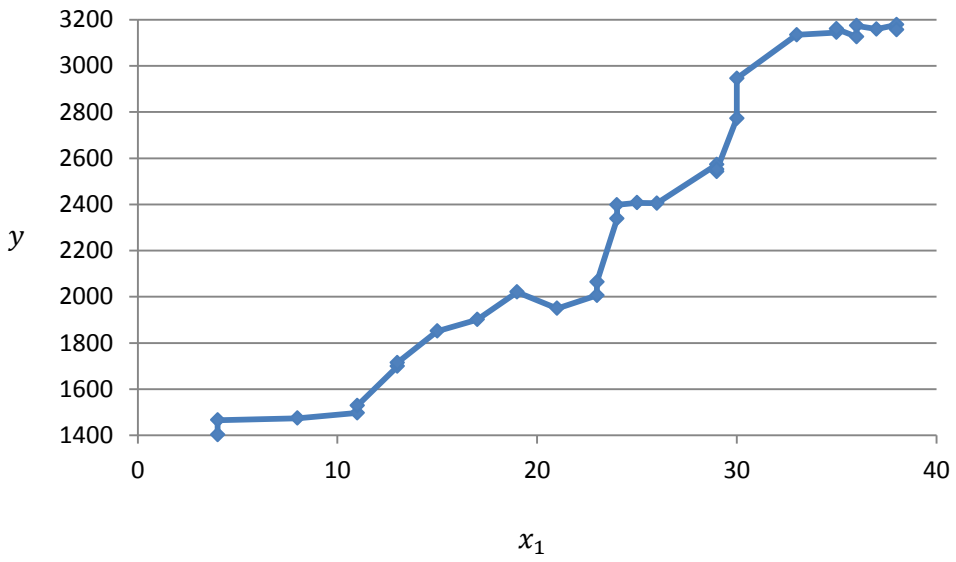

Figure 5.35 – y values plotted against x_1 values (Raw artificial data, case 1)

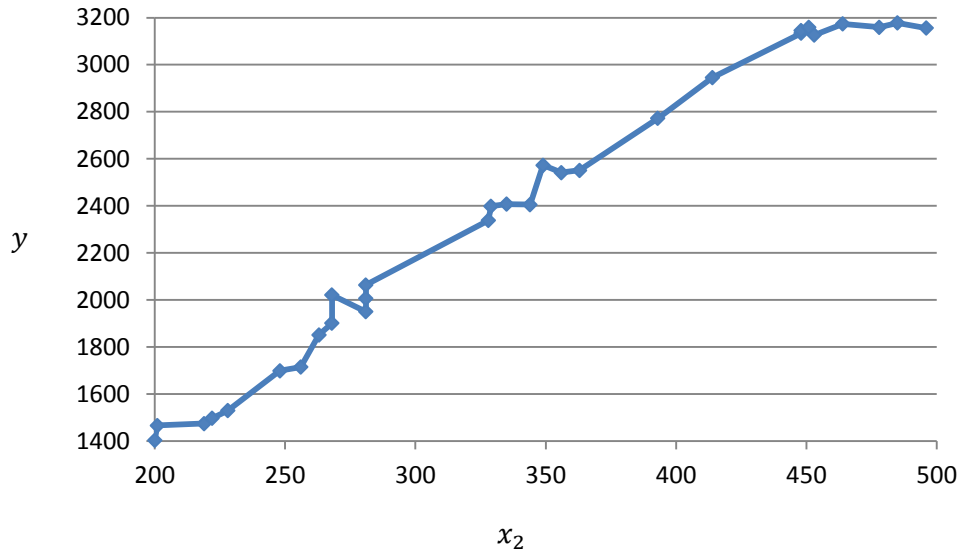

Figure 5.36 – y values plotted against x_2 values (Raw artificial data, case 1)

In the first case the data in table 5.19 was used. Figures 5.35 and 5.36 depict the y values plotted against the two decision variables, x_1 and x_2, respectively. From the graphs it can be observed that both functions are generally monotonically non-decreasing and therefore that the function variables will also be constrained as monotonically non-decreasing. The same procedure as in section 5.3.1 was followed to determine the mean absolute deviation value for each model. In this section only the summary of the experiment will be presented. The number of data points to be omitted, p, was chosen as 2; a smoothing factor of $\beta = 125$ was also incorporated into the minimal assumption regression model.

Figures 5.37 and 5.38 portray the function values, $f_1(x_1)$ and $f_2(x_2)$, plotted against the decision variables, x_1 and x_2, for the minimal assumption regression model when 2 data points are omitted and a smoothing function of $\beta = 125$ is incorporated into the model.

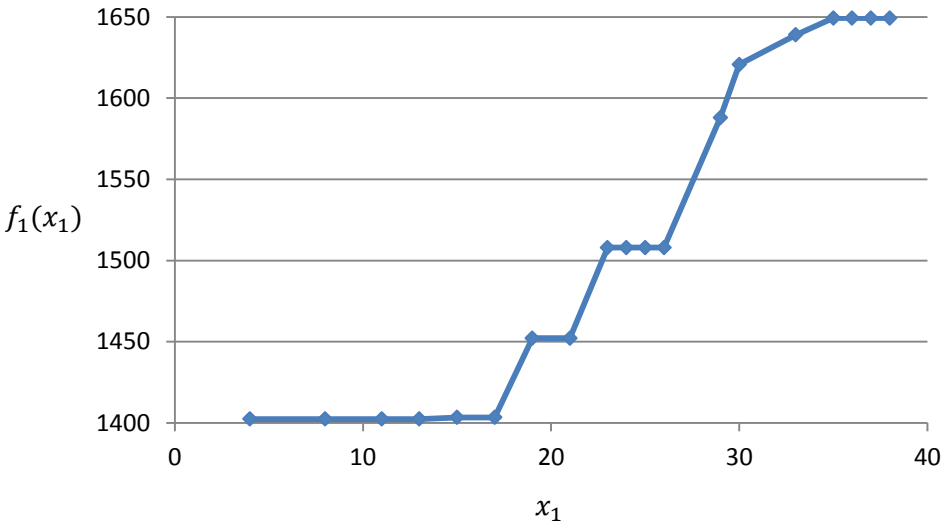

Figure 5.37 – f_1 values plotted against x_1 values (Artificial data, case 1)

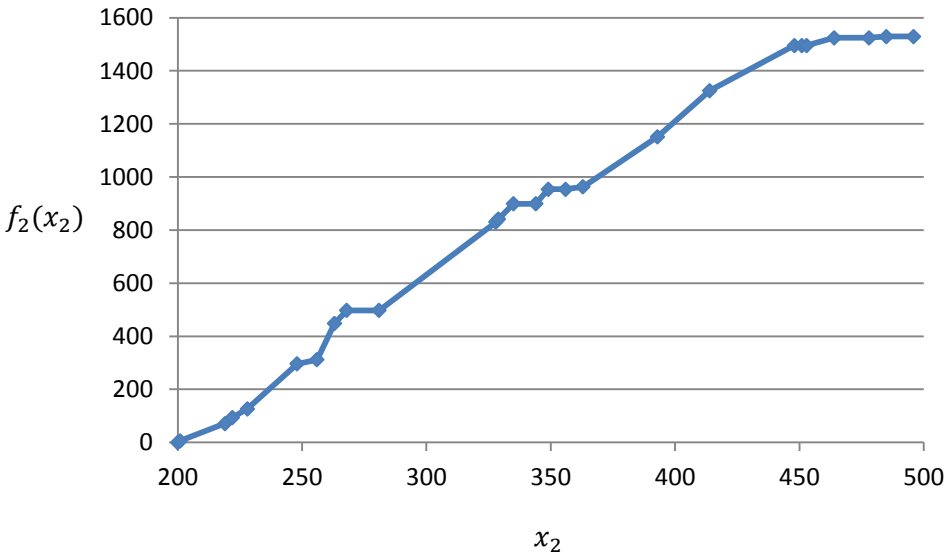

Figure 5.38 – f_2 values plotted against x_2 values (Artificial data, case 1)

Model	Mean absolute deviation
L_2-norm regression	69.025
L_1-norm regression	78.096
Original minimal assumption regression model (Wagner, 1962)	44.275

Table 5.20 – Comparison of models that do not omit outliers (Artificial data, case 1)

For the models that do not omit outliers (table 5.20), the minimal assumption regression model contributed the lowest mean absolute deviation, with a value of 44.275. This is about 36% better than the mean absolute deviation of the L_2-norm regression.

Model	Mean absolute deviation
Wagner's model omitting 2 data points	37.180
Wagner's model omitting 2 data points and a smoothing factor of $\beta = 125$	37.179
L_1-norm regression omitting 2 data points and no breakpoint	71.699
Piecewise L_1-norm regression omitting 2 data points and one breakpoint	59.580
Piecewise L_1-norm regression omitting 2 data points and two breakpoint	62.679
Piecewise L_1-norm regression omitting 2 data points and three breakpoint	39.482
Piecewise L_1-norm regression omitting 2 data points and four breakpoints	40.835

Table 5.21 – Comparison of models that omit outliers (Artificial data, case 1)

Table 5.21 reports the mean absolute deviations of the models that omit 2 data points. By implementing the omission of 2 data points the mean absolute deviation value decreases with a

further 16%. In this case, the smoothing factor does not really improve the mean absolute deviation. By increasing the breakpoints to three, the mean absolute deviation of the piecewise L_1-norm regression that omits 2 data points also improved, but it is still 6% more than Wagner's model which omits 2 data points.

5.4.2 Case 2

The data set for Case 2 was simulated in the same manner as with Case 1 in the previous section, but both functions in this case are generally monotonically non-increasing (see figures 5.39 and 5.40); therefore the function variables will also be constrained as monotonically non-increasing.

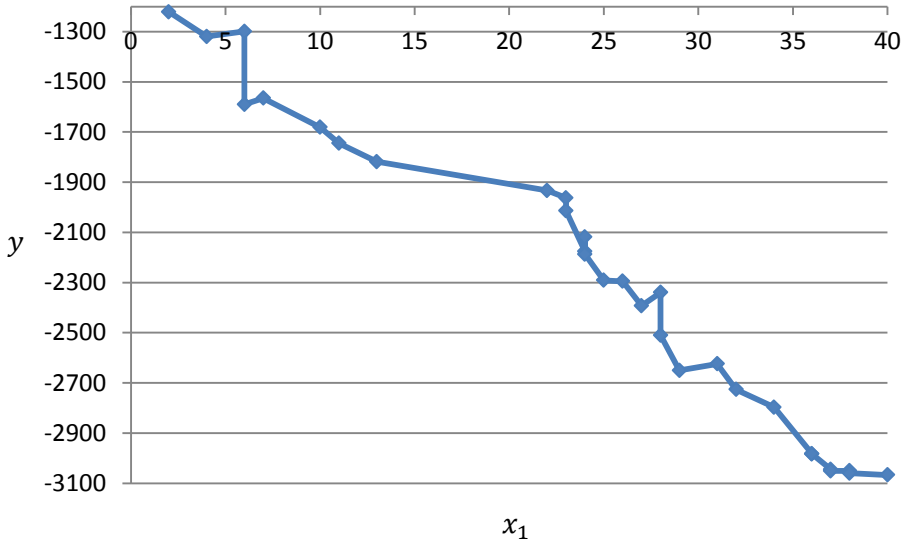

Figure 5.39 – y values plotted against x_1 values (Raw artificial data, case 2)

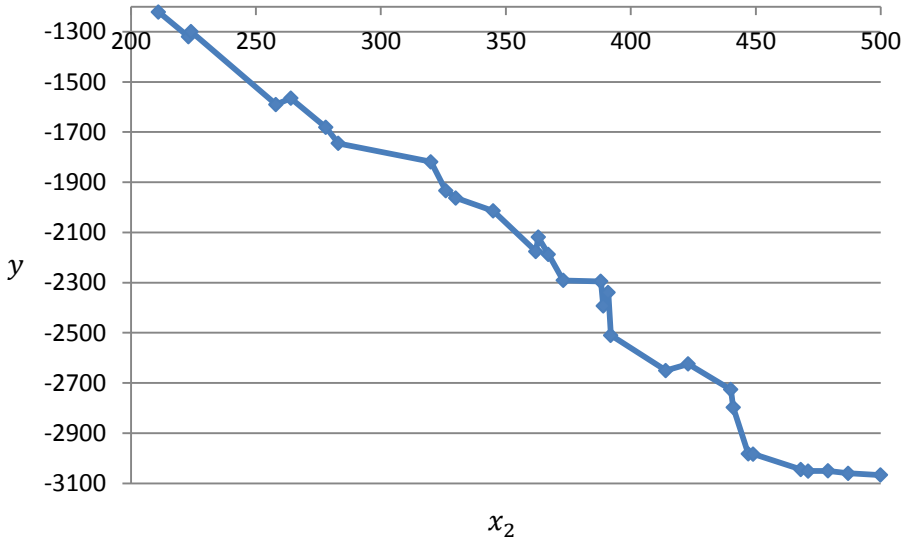

Figure 5.40 – y values plotted against x_2 values (Raw artificial data, case 2)

In this case, the number of data points to be omitted, p, was chosen as 2; a smoothing factor of $\beta = 50$ proved to enhance the predictive capability of the minimal assumption regression model further.

Figures 5.41 and 5.42 depict the function values, $f_1(x_1)$ and $f_2(x_2)$, plotted against the decision variables, x_1 and x_2, for the minimal assumption regression model when 2 data points are omitted and a smoothing function of $\beta = 50$ is incorporated into the model.

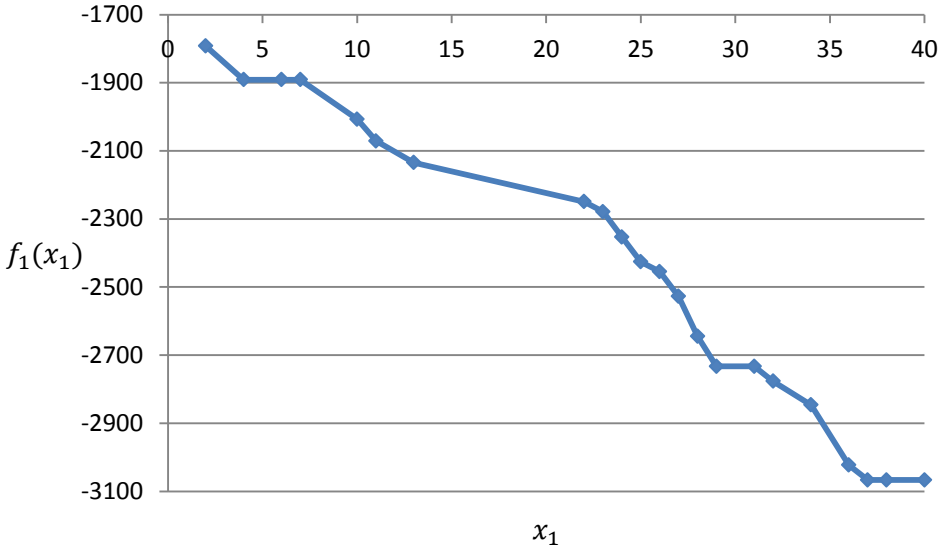

Figure 5.41 – f_1 values plotted against x_1 values (Artificial data, case 2)

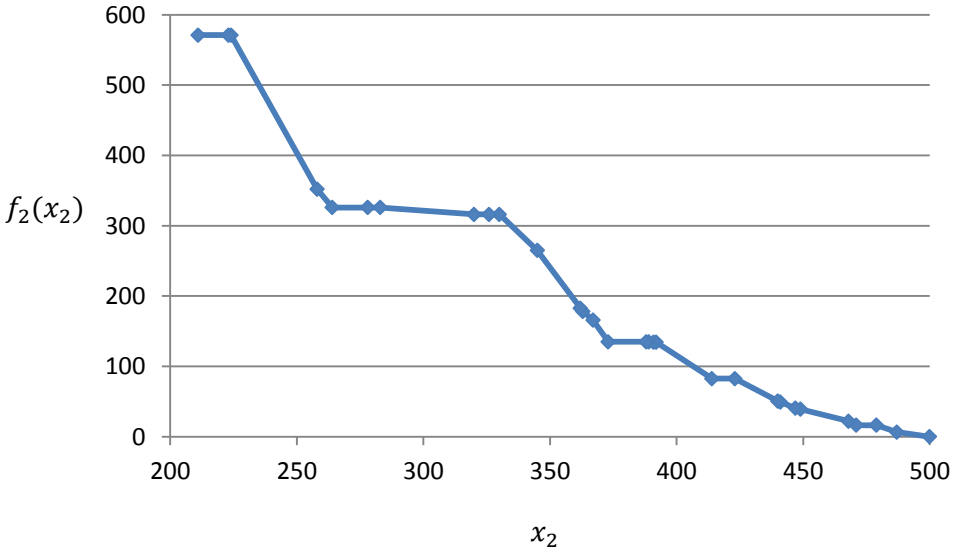

Figure 5.42 – f_2 values plotted against x_2 values (Artificial data, case 2)

From table 5.22 it is clear that the mean absolute deviation values of the L_1-norm regression model and the minimal assumption regression model are almost the same and approximately 11% lower than the mean absolute deviation of the L_2-norm regression model.

Model	Mean absolute deviation
L_2-norm regression	72.164
L_1-norm regression	64.425
Original minimal assumption regression model (Wagner, 1962)	64.602

Table 5.22 – Comparison of models that do not omit outliers (Artificial data, case 2)

The improvement in the mean absolute deviation of the minimal assumption regression model, when 2 data points are omitted and a smoothing function of $\beta = 50$ is introduced, is quite remarkable. With a 44% improvement (from 64.602 to 36.133) this model gives the best mean absolute deviation for the data set in Case 2.

Model	Mean absolute deviation
Wagner's model omitting 2 data points	52.813
Wagner's model omitting 2 data points and a smoothing factor of $\beta = 50$	36.133
L_1-norm regression omitting 2 data points and no breakpoint	60.680
Piecewise L_1-norm regression omitting 2 data points and one breakpoint	44.817
Piecewise L_1-norm regression omitting 2 data points and two breakpoint	44.817
Piecewise L_1-norm regression omitting 2 data points and three breakpoint	47.467
Piecewise L_1-norm regression omitting 2 data points and four breakpoints	40.888

Table 5.23 – Comparison of models that omit outliers (Artificial data, case 2)

5.4.3 Case 3

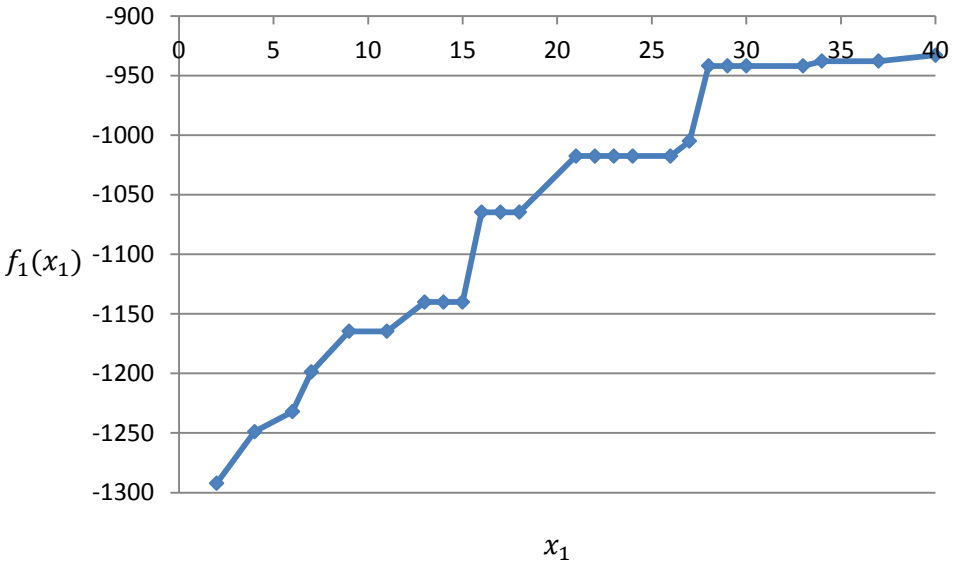

Figure 5.43 – f_1 values plotted against x_1 values (Artificial data, case 3)

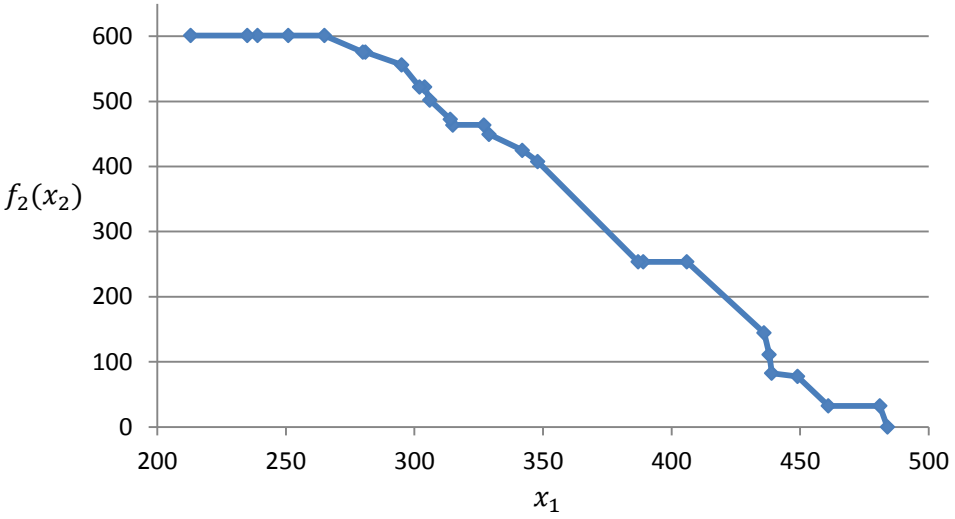

Figure 5.44 – f_2 values plotted against x_2 values (Artificial data, case 3)

For the last data set (raw data not given) the first function variable, $f_1(x_1)$, is constrained as monotonically non-decreasing while the second function variable, $f_2(x_2)$, is constrained as monotonically non-increasing. Figures 5.43 and 5.44 depict the function values, $f_1(x_1)$ and $f_2(x_2)$, plotted against the decision variables, x_1 and x_2, for the minimal assumption regression model with no extensions.

Model	Mean absolute deviation
L_2-norm regression	43.523
L_1-norm regression	41.579
Original minimal assumption regression model (Wagner, 1962)	24.685

Table 5.24 – Comparison of models that do not omit outliers (Artificial data, case 3)

The minimal assumption regression model with no extensions outperformed the other two models which do not omit outliers (table 5.24). The mean absolute deviation of this model is 43% better than that of the L_2-norm regression model.

In this case the omission of outliers does not improve the mean absolute value and neither does the incorporation of a smoothing factor (see table 5.25). By increasing the breakpoints of the piecewise L_1-norm regression omitting 1 data point to three breakpoints, the mean absolute deviation decreases to 25.268. The mean absolute deviation of the minimal assumption regression model with no extensions is still 2% smaller and thus gives the best results for the data set used in Case 3.

Model	Mean absolute deviation
Wagner's model omitting 1 data point	24.822
Wagner's model omitting 1 data point and a smoothing factor of $\beta = 350$	30.099
L_1-norm regression omitting 1 data point and no breakpoint	37.275
Piecewise L_1-norm regression omitting 1 data point and one breakpoint	25.611
Piecewise L_1-norm regression omitting 1 data point and two breakpoint	25.449
Piecewise L_1-norm regression omitting 1 data point and three breakpoint	25.268
Piecewise L_1-norm regression omitting 1 data point and four breakpoints	28.855

Table 5.25 – Comparison of models that omit outliers (Artificial data, case 3)

In all three cases the minimal assumption regression model with no extensions did better than, or was at least equal to, the other models that did not omit outliers. The minimal assumption regression model with the added extensions (omission of outliers and/or incorporation of a smoothing factor) outperformed the other model in all three cases.

If one considers these results it seems permissible to draw the conclusion that cases can be constructed for which the minimal assumption regression model, and extensions thereof, will yield better results than other models.

In the next section the results of this section will be discussed and conclusions that are reached will be explained.

5.5 Discussion and summary of results

From the results obtained in section 5.4 it is possible to arrive at some conclusions about the performance of the minimal assumption regression model and extensions thereof.

In the results of the first four data sets (indicated in table 5.26 as Stack loss, Scottish hill race, Weisberg fuel and GNP) the value of the mean absolute deviation of the minimal assumption regression model without extensions was always situated between the values of the mean absolute deviation of the L_1-norm and L_2-norm regression models. In two out of the four times (for the Stack loss and GNP data sets) the value of the mean absolute deviation of the minimal assumption regression model compared very well with the L_1-norm regression model (which gave the lowest mean absolute deviation in all four cases). In the other two cases the mean absolute deviation of the minimal assumption regression model was approximately in the middle of the results of the other two models.

Model	Stack loss	Scottish hill race	Weisberg fuel	GNP
L_2-norm regression	2.887	9.367	54.532	4038.425
L_1-norm regression	2.035	8.211	49.466	3215.116
Original minimal assumption regression model (Wagner, 1962)	2.077	8.927	52.726	3282.140

Table 5.26 – Summary of models that do not omit outliers (Four datasets)

The results obtained for the financial ratio data set (see table 5.18) indicate that the value of the mean absolute deviation of the minimal assumption regression model without extensions outperformed the L_1-norm and L_2-norm regression models, with a value more than 80% less than the values of the L_1-norm and L_2-norm regression models. In this data set the mean absolute deviation did not improve by adding extensions to the model. Omitting outliers in the piecewise linear regression models did not help either. The breakpoints in the piecewise L_1-norm regression model, without omitting outliers, did improve the mean absolute deviation, but the mean absolute deviation of the minimal assumption regression model without extensions was still the lowest.

For the other four data sets the extensions were implemented, and the results improved considerably in all of the cases. In table 5.27 the mean absolute deviation values for the different models and data sets are recorded. The number of data points omitted or the specific smoothing factor incorporated into the model is not included in this table because it differs for each data set. These values can be obtained in sections 5.3.1 – 5.3.4 where the results of the specific data set are discussed. They are also highlighted in the discussion that follows.

Model	Stack loss	Scottish hill race	Weisberg fuel	GNP
Wagner's model omitting data points	2.194	8.469	33.639	2764.669
Wagner's model omitting data points with a smoothing factor	1.220	3.921	37.425	1820.180
L_1-norm regression omitting data points and no breakpoint	1.394	4.253	35.426	2360.977
Piecewise L_1-norm regression omitting data points and one breakpoint	1.882	3.559	33.293	2782.106
Piecewise L_1-norm regression omitting data points and two breakpoints	2.150	4.280	44.320	1713.885

Table 5.27 – Summary of models that omit outliers (Four datasets)

For the stack loss data set the omission of 2 data points and a smoothing factor of $\beta = 50$ incorporated into the minimal assumption regression model resulted in the lowest mean absolute deviation of 1.220; the next mean absolute deviation is 12% larger than this value.

When the different models are applied to the Scottish hill race data set, the piecewise L_1-norm regression model omitting 4 data points with one breakpoint yields the lowest mean absolute deviation value (3.559). The value of the mean absolute deviation of the minimal assumption regression model omitting 4 data points with a smoothing factor of $\beta = 1$ differs by less than 10% from the best results in this case.

The piecewise L_1-norm regression model omitting 6 data points with one breakpoint (33.293) yielded the lowest mean absolute deviation value for the Weisberg fuel consumption data set. The minimal assumption regression model which omits 6 data points but does not include a smoothing factor follows closely with the second best result (1% difference).

The best result for the gross national product data set was produced by the L_1-norm regression model omitting 4 data points (1713.885). The results of the minimal assumption regression model omitting 4 data points with a smoothing factor of $\beta = 200$ differs by less than 6% from the best results in this case.

To summarize, the minimal assumption regression model (or extensions thereof) obtained the best mean absolute deviation value for two out of the five data sets. For the other three data sets, the minimal assumption regression model with one or both of the extensions closely followed the model with the best result.

In table 5.28 a summary of the models that do not omit outliers is given for the three cases discussed in section 5.4. In Case 2 the mean absolute deviation values of the L_1-norm regression model and the minimal assumption regression model are very close, but for the other two cases the minimal assumption regression model gave the best results.

Model	**Case 1**	**Case 2**	**Case 3**
L_2-norm regression	69.025	72.164	43.523
L_1-norm regression	78.096	64.425	41.579
Original minimal assumption regression model	44.275	64.602	24.685

Table 5.28 – Summary of models that do not omit outliers (Artificial datasets)

When the extensions were incorporated into the model the minimal assumption regression model with extensions still outperformed the other models (see table 5.29).

Model	Case 1	Case 2	Case 3
Wagner's model omitting data points	37.180	52.813	24.822
Wagner's model omitting data points with a smoothing factor	37.179	36.133	30.099
L_1-norm regression omitting data points and no breakpoint	71.699	60.680	37.275
Piecewise L_1-norm regression omitting data points and one breakpoint	59.580	44.817	25.611
Piecewise L_1-norm regression omitting data points and two breakpoints	62.679	44.817	25.449
Piecewise L_1-norm regression omitting data points and three breakpoints	39.482	47.467	25.268
Piecewise L_1-norm regression omitting data points and four breakpoints	40.835	40.888	28.855

Table 5.29 – Summary of models that omit outliers (Artificial datasets)

From table 5.26 it can clearly be seen that the minimal assumption regression model without extension compares well with the L_1-norm and L_2-norm regression models. Although it did not yield the best results, it also did not give the worst results. It is also feasible to solve these types of models with current software such as CPLEX. The conclusion that can be arrived at is that if one is unsure about the validity or applicability of the assumptions made by a model (which will not necessarily be suitable for the data that is used) it may be advisable to use the minimal assumption regression model. The results obtained should not be worse than the results obtained by a L_2-norm regression model; in other words the results will be comparable.

By implementing the robust extensions as described in this study, the predictive capabilities as measured by the mean absolute deviation values should improve significantly. It is feasible, as noted, to solve these kinds of models with current software such as CPLEX. In one of the four cases (stack loss) in which extensions were added to the models, the minimal assumption regression model with extensions provided the best results. In the other three cases the minimal assumption regression model with extensions was very competitive.

One conclusion that can be deduced from the models which implemented extensions is that these enhanced the predictive accuracy of the models. The robustness of the models is thus improved. Another conclusion is that the minimal assumption regression model with various

extensions can be implemented and give results that are competitive with other types of regression models.

5.6 Chapter summary

In this chapter several data sets were used to undertake empirical experiments. The first data set was used to explain the process of applying Wagner's minimal assumption regression model and of incorporating extensions such as the omission of outliers and the smoothing of functions. The data set was also used to introduce piecewise linear regression. The other data sets were discussed by showing explanatory graphs and the results they illustrated. For each data set the predictive capabilities of each model were compared.

The minimal assumption regression model with extensions compared fairly well with the other models. Although this model did not contain the lowest mean absolute deviation in all the data sets investigated, the results were always comparable to the results of the other models.

In section 5.4 data was simulated to test whether there are instances in which the minimal assumption regression model will perform better than other models. The data were simulated with specific features to test the monotonicity constraints of the model. These data sets showed that this is possible.

The next chapter furnishes some final remarks.

Chapter 6

Summary and conclusions

6.1 Introduction

Chapter 6 presents the final comments and concluding remarks of the study. The objectives of the study and their achievement will be summarised. The new problems and opportunities for further study that were presented during the research project will also be outlined.

6.2 Objectives of the study

Chapter 1 stated that the primary objective of this study was to investigate robust techniques for regression models with minimal assumptions by using linear programming and integer linear programming techniques. To accomplish this, a list of four secondary research objectives was defined in order to achieve the primary objective. These were to:

- gain a clear understanding of, and present an introductory overview of linear regression, outliers and linear and integer linear programming;
- perform an exploratory investigation into robust techniques for regression models with minimal assumptions;
- address robustness by introducing an adapted minimal assumption mixed integer linear programming model that is able to deal with possible outliers as well as the smoothing of functions; and
- apply the adapted model to different data sets in order to evaluate its performance.

A summary of how these objectives were addressed follows below:

Gain a clear understanding of and present an introductory overview of linear regression, outliers and linear and integer linear programming.

This objective was addressed by describing linear regression models and three different ways of estimating parameters for such models by using the L_1-norm, L_2-norm and L_∞-norm methods (Chapter 2, section 2.2). The influence of outliers and the detection thereof (ordinary and robust methods of detection) was then discussed (Chapter 2, section 2.3). Sections 2.5 and 2.6 of Chapter 2 presented some fundamentals of linear and integer linear programming.

Perform an exploratory investigation into robust techniques for regression models with minimal assumptions.

Harvey M. Wagner proposed a different approach to solving regression problems and introduced a model that made minimal assumptions about the form of a regression function. To address this objective a detailed explanation of the minimal assumption regression model was presented (Chapter 3, section 3.3). An example of how the model was applied to data was given (Chapter 3, section 3.4) and a brief literature review of other researchers who referred to the minimal assumption regression model was provided (Chapter 3, section 3.5).

Address robustness by introducing an adapted minimal assumption mixed integer linear programming model that is able to deal with possible outliers as well as the smoothing of functions.

This objective was addressed by offering a comprehensive explanation of the development of a robust model (Chapter 4, section 4.2) which includes the concepts of outlier identification and smoothing techniques (Chapter 4, section 4.2.1 – 4.2.2). These extensions were incorporated to make the model more robust.

Apply the adapted model to different data sets in order to evaluate its performance.

To test whether the suggested robust model compares well with other models, a piecewise linear regression model was introduced (Chapter 4, section 4.3). This objective was addressed by applying the different models to five different data sets and the results were compared (Chapter 5, section 5.3). In addition, data sets with specific features were simulated to investigate the performance of the minimal assumption regression model (Chapter 5, section 5.4).

To summarize, all objectives set forth in Chapter 1 were addressed. Based on the results and discussion presented in chapter 5, it was concluded that:

- It is feasible to solve regression problems with the minimal assumption regression model;
- The minimal assumption regression model compares favourably with two other classical techniques;

- The extensions towards robustness that were incorporated into the minimal assumption regression model seem to improve the predictive capability of the model;
- It is feasible to solve regression problems by the minimal assumption regression model with extensions. This model also compares favourably with other techniques and in some cases outperforms them;
- There are specific cases where the minimal assumption regression model with or without extensions will perform better than other regression models;
- Modern optimization software, such as CPLEX, seems powerful enough to solve regression problems using the minimal assumption approach. The software can also handle models for the omission of outliers and smoothing of regression functions; and
- By using the minimal assumption regression model, small to medium sized problems can be solved in a relatively short time.

6.3 Problems experienced

Model selection and robustness issues remain a challenge in regression analysis. Measures to test the suitability of models are also by no means an easy task. Many researchers are making contributions to this field and this study merely begins to explore the merits of the Wagner method.

6.4 Possibilities for further research

The estimation of the parameters p (the number of data points to omit) and β (the smoothing factor) could be explored further. Experiments with different parameter values of p and β might be performed on a large number of data sets. By applying the model to more data sets a general guideline may be set up in order to specify p and β. This may lead to further refinement of the model.

This model could be extended by incorporating simultaneous variable selection, as suggested by some researchers, which may improve model performance. Techniques to handle larger data sets more effectively might also be investigated.

6.5 Chapter summary

Chapter 6 is the final chapter of this study. The chapter presented a summary of the initial objectives and how the objectives were addressed. In conclusion, the problems experienced and possible future research opportunities were outlined.

Appendix A

A.1 Simple linear regression

A simple linear regression model assumes that a straight line can approximate the relationship between the dependent variable, which is denoted y, and the predictor variable, denoted x. Bowerman *et al.* (2005) formally defines a simple linear regression model as follows:

$$y = \beta_0 + \beta_1 x + \varepsilon, \tag{A.1}$$

where

β_0 is the y-intercept. β_0 is the mean value of y when x equals 0;
β_1 is the slope. β_1 is the change (amount of increase or decrease) in the mean value of y associated with a one-unit increase (or decrease) in x. If β_1 is positive the mean value of y increases as x increases. If β_1 is negative, the mean value of y decreases as x increases; and
ε is an error term that describes the effects on y of all factors other than the value of the predictor variable x.

The main assumptions about the simple linear regression model are summarized as follows: the error terms are assumed to be independently and identically distributed (iid) normal random variables each with a mean of zero and constant variance, σ^2. This statement implies four assumptions which are explained by Bowerman *et al.* (2005) as:

- *Independence assumption*. Any one value of the error term ε is statistically independent of any other value of ε. That is, the value of the error term ε corresponding to an observed value of y is statistically independent of the value of the error term corresponding to any other observed value of y;
- *Normality assumption*. At any given value of x, the population of potential error term values has a normal distribution;
- At any given value of x, the population of potential error term values has a mean equal to zero; and
- *Constant variance assumption*. At any given value of x, the population of potential error term values has a variance that does not depend on the value of x. That is, the different populations of potential error term values corresponding to different values of x have equal variances. The constant variance is denoted by σ^2.

The point estimates b_0 and b_1 of the parameters β_0 and β_1 can be calculated using the least squares point estimate (L_2-norm) for the simple linear regression model. The formulas are given by Taylor (2001) as

$$b_1 = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}, \tag{A.2}$$

and

$$b_0 = \bar{y} - b_1\bar{x}, \tag{A.3}$$

where

$n =$ number of observations;

$\bar{x} = \frac{\sum x_i}{n}$; and

$\bar{y} = \frac{\sum y_i}{n}$.

Following the construction of a simple linear regression model, it is possible to test the significance of the predictor variables to calculate a confidence interval for the mean value of the dependent variable. It is also possible to calculate a prediction interval for an individual value of the dependent variable. A technical discussion and examples of such calculations can be found in Bowerman *et al.* (2005) and will not be presented here. This section is concluded with a brief reference to a measure of the usefulness of a simple linear regression model, called the simple coefficient of determination, as well as a measure of the relationship between the two variables y and x, termed the simple correlation coefficient.

The simple coefficient of determination for a simple linear regression model is defined as

$$r^2 = \frac{Explained\ variation}{Total\ variation}, \tag{A.4}$$

where

Total variation = $\sum(y_i - \bar{y})^2$;

Explained variation = $\sum(\hat{y} - \bar{y})^2$;

Unexplained variation = $\sum(y_i - \hat{y})^2$; and

Total variation = Explained variation + Unexplained variation.

r^2 is the proportion of the total variation in the n observed values of the dependent variable that is explained by the simple linear regression model.

The simple correlation coefficient, which measures the strength of the linear relationship between the variables y and x, is defined as

$$r = +\sqrt{r^2} \text{ if } b_1 \text{ is positive, and} \tag{A.5}$$

$$r = -\sqrt{r^2} \text{ if } b_1 \text{ is negative,} \tag{A.6}$$

where

b_1 is the slope of the least squares line relating y to x.

The correlation coefficient r can take on values between -1 and 1, because the value of r^2 is always between 0 and 1. A value of r close to 0 indicates little or no linear relationship between y and x, while a value of r close to 1 or -1 indicates a strong linear relationship between y and x. A value of r close to 1 means that y and x are highly related and are positively correlated whereas a value of r close to -1 means that y and x are highly related and negatively correlated. When $r = 1$, y and x have a perfect linear relationship with a positive slope and when $r =$ -1, y and x have a perfect linear relationship with a negative slope.

A.2 Graphical methods for linear programming problems

As explained in section 2.5 of Chapter 2, the objective function of any linear programming problem is to minimize or maximize a certain quantity such as profit or cost. Another requirement for linear programming problems is the presence of constraints or restrictions; these constraints limit the extent to which the problem can be minimized or maximized. When there are only two variables to consider, a graphical representation is the easiest way to solve the problem. The isoprofit and corner point methods are two ways to solve a two-variable problem graphically. Each method consists of four steps, which will be presented below, followed by an example.

A.2.1 Isoprofit method

According to Moore and Weatherford (2001) the isoprofit line can be defined as a contour of a profit function. A contour of the function $f(x_1, x_2)$ is the set of all combinations of values for the variables (x_1, x_2) such that the function $f(x_1, x_2)$ takes on a specified constant value. The steps for the isoprofit method are as follows:

1. Graph all constraints and find the feasible region;
2. Select a specific profit (or cost) line and graph it to find the slope;
3. Move the profit (or cost) line in the direction of increasing profit (or decreasing cost) while maintaining the slope. The last point it touches in the feasible region is the optimal solution; and
4. Find the values of the decision variables at this last point and compute the profit (loss).

A.2.2 Corner point method

Bazaraa *et al.* (2005) demonstrate that if an optimal solution for a problem exists, then an optimal extreme point (or corner point) also exists. Hence, it is only necessary to evaluate all the corner points, because the optimal solution will be found at one of them. For a two-variable problem this is fairly easy; the steps for the corner point method are listed below:

1. Graph all constraints and find the feasible region;
2. Find the corner points of the feasible region;
3. Compute the profit (cost) at each of the feasible corner points; and
4. Select the corner point with the best value of the objective function found in step 3. This is the optimal solution.

To illustrate these concepts a two-variable example is furnished:

Example A.1 Suppose a company can manufacture two products, x_1 and x_2. Each unit of product x_1 yields a profit of R5 and each unit of product x_2 is sold for a profit of R3. To manufacture one unit of product x_1 requires 3 labour hours and 3 units of material. One unit of product x_2 requires 1 labour hour and 2 units of material. For the current production period there are 120 labour hours available and 210 units of material at hand. The company is interested in the best possible combination of products x_1 and x_2 to manufacture in order to maximize the profit. This situation can be formulated as the following linear programming problem:

$$\begin{aligned} \text{Maximize} \quad & 5x_1 + 3x_2 \\ \text{subject to} \quad & 3x_1 + \quad x_2 \leq 120 \\ & 3x_1 + \ 2x_2 \leq 210 \\ & x_1, x_2 \geq 0 \end{aligned}$$

To find a solution for the linear program, values for the decision variables need to be chosen which will maximize the objective function. These values should be located in the feasible region. Graphically the objective function (or isoprofit line) can be moved from the origin, parallel to itself, until it reaches the extreme point within the feasible region. Outside of the feasible region the decision variables will take on values which will violate the constraints.

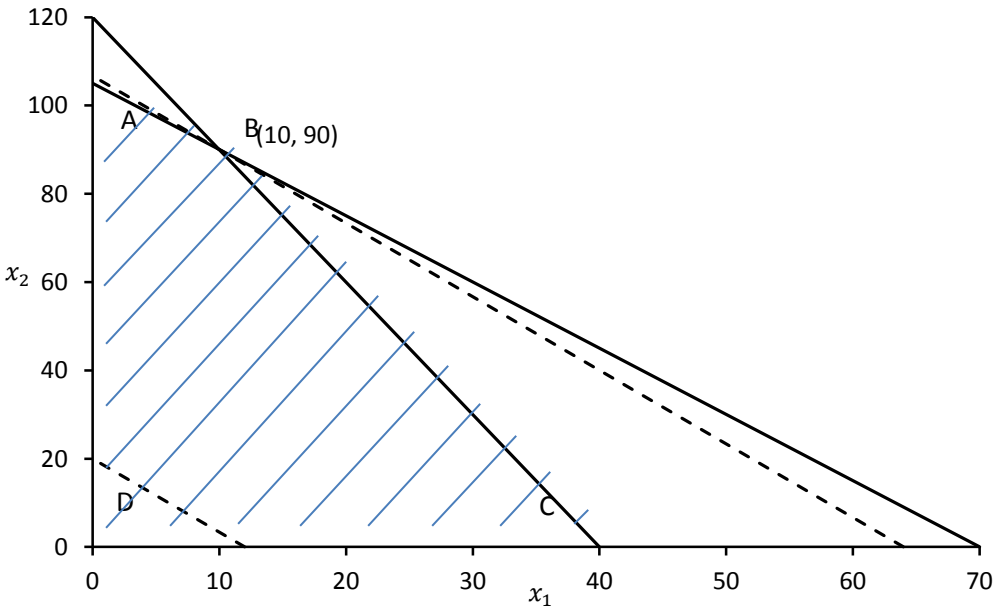

Figure A.1 – Graphical solution for example A.1

Figure A.1 depicts the feasible region as the region within the corner points ABCD. The dotted line is the objective function (or isoprofit line) whereas the solid lines represent the constraints. The optimum point is B, with x_1 equal to 10 and x_2 equal to 90. The maximum value, when the

solution values are substituted into the equation, is 320. Any point outside of the feasible region will result in an infeasible solution.

The other corner points are A (0, 105) which gives an objective value of 315, C (40, 0) which gives an objective value of 200 and D (0, 0) which, consequently, gives an objective value of 0.

The solution of a linear program can either be unique, as in this example, or alternative solutions can be found. They can be found when there is a constraint that is parallel to the objective function, which implies that any point on the constraint line is considered optimal. When there is no constraint that restricts the objective function the objective function can be increased indefinitely; this is an unbounded solution scenario. If no feasible solution exists, the problem is infeasible.

Graphical methods provide a good conceptual basis for solving linear programming problems, but to solve real life problems with numerous variables and constraints, a solution procedure called the simplex method is usually used. This method is described in appendix A, section A.3.

A.3 The Simplex method

Example A.1, in section A.2, contains only two decision variables and it is possible to solve the problem graphically by using the methods described in the previous section. However, business problems may contain hundreds, even thousands, of variables and for these cases a graphing technique will not suffice. For these instances the simplex method, a more powerful technique, can be used. It was developed by George B. Dantzig in 1947, and is a popular and effective tool for solving optimization problems (Bazaraa *et al*., 2005). It is a simple concept; examine the corner points in an iterative, systematic manner until an optimal solution is reached. The optimum solution will lie at a corner point of the many-sided, many-dimensional figure in the area of the feasible solutions. For each iteration the objective value obtains a higher value and is always moving closer to the optimal solution.

The setup of an initial simplex method tableau will be illustrated by using the linear programming model presented in example A.1. Firstly the inequality constraints must be converted into equations. This can be achieved by adding slack variables, S_1 and S_2 to the constraints

$$3x_1 + \quad x_2 + S_1 = 120,$$
$$3x_1 + 2x_2 + S_2 = 210.$$

Thus, the unused resources are represented by the slack variables S_1 and S_2. To find a solution for the tableau, an initial solution must be obtained. A basic feasible solution can be established by setting the decision variables equal to 0, if $x_1 = x_2 = 0$, then $S_1 = 120$ and $S_2 = 210$.

The initial tableau is set up as in table A.1, where x_1 and x_2 are the decision variable columns and S_1 and S_2 the slack variable columns. The 'Solution Mix' column shows which variables are in the production mix while the C_j column depicts the profit per unit. The 'Quantity' column is the constant column. The second row reports the profit per unit and in the following two rows each of the constraint equations is given. Z_j is the gross profit row, while $C_j - Z_j$ represents the net profit.

	Solution Mix	x_1	x_2	S_1	S_2	Quantity
C_j		R5	R3	R0	R0	
R0	S_1	3	1	1	0	120
R0	S_2	3	2	0	1	210
	Z_j	R0	R0	R0	R0	R0
	$C_j - Z_j$	R5	R3	R0	R0	

Table A.1 – The initial simplex tableau

After an initial tableau has been completed a series of five steps is followed to compute all the values needed for the next tableau (Render *et al*., 2009):

1. Determine which variable to enter into the solution mix next. One way of doing this is by identifying the column (pivot column), and hence the variable, with the largest positive number in the $C_j - Z_j$ row of the preceding tableau. Producing this variable will contribute the greatest additional profit per unit;
2. Determine which variable to replace. A basic variable must be chosen to make room for the new variable chosen in step 1. Divide each amount in the quantity column by the corresponding number in the column selected in step 1. The row (pivot row) with the smallest nonnegative number calculated in this way will be replaced in the next tableau. The pivot number is the number at the intersection of the pivot row and pivot column;
3. Compute new values for the pivot row. To do this, divide every number in the row by the pivot number;
4. Compute the new values for each remaining row. All remaining row(s) are calculated as follows

   $$(\text{new row numbers}) = (\text{numbers in old row}) - \left[\begin{pmatrix}\text{number above}\\ \text{or below}\\ \text{pivot number}\end{pmatrix} \times \begin{pmatrix}\text{corresponding number in}\\ \text{the new row, that is, the}\\ \text{row replaced in step 3}\end{pmatrix}\right];$$

5. Compute the Z_j and $C_j - Z_j$ rows, as demonstrated in the initial tableau. If all numbers in the $C_j - Z_j$ row are 0 or negative, an optimal solution has been reached. If this is not the case, return to step 1.

To apply these steps, the pivot column, -row and -number must be identified in table A.1. In this case x_1 is the pivot column, with the largest positive $C_j - Z_j$ value. S_1 provides the smallest nonnegative number and is therefore the pivot row. The pivot number, 3, is located at the intersection of the pivot column and -row. After establishing this information the second tableau can be completed. Table A.2 provides the second simplex tableau.

	Solution Mix	x_1	x_2	S_1	S_2	Quantity
C_j		R5	R3	R0	R0	
R5	x_1	1	1/3	1/3	0	40
R0	S_2	0	1	-1	1	90
	Z_j	R5	R1.666	R1.666	R0	R200
	$C_j - Z_j$	R0	R1.333	-R1.666	R0	

Table A.2 – The second simplex tableau

The same procedure is followed to obtain the last tableau. In this case x_2 is the pivot column, S_2 the pivot row and 1 the pivot number. Table A.3 presents the third and last tableau which

contains the optimal solution for this problem, with $x_1 = 10$ and $x_2 = 90$ to obtain a profit of R320.

	Solution Mix	x_1	x_2	S_1	S_2	Quantity
C_j		R5	R3	R0	R0	
R5	x_1	1	0	2/3	-1/3	10
R3	x_2	0	1	-1	1	90
	Z_j	R5	R3	R0.333	R1.333	R320
	$C_j - Z_j$	R0	R0	-R0.333	-R1.333	

Table A.3 – The third simplex tableau

The procedure for solving linear programming maximizations problems is summarized by (Render *et al*., 2009) as follows:

I. Formulate the linear programming problem's objective function and constraints;
II. Add slack variables to each less-than-or-equal-to constraint and to the problem's objective function;
III. Develop an initial simplex tableau with the slack variables in the basis and the decision variables equal to 0. Compute the Z_j and $C_j - Z_j$ values for this tableau;
IV. Follow these five steps until an optimal solution has been reached:
    1. Choose the variable with the greatest positive $C_j - Z_j$ value to enter the solution. This is the pivot column;
    2. Determine the solution mix variable to be replaced and the pivot row by selecting the row with the smallest (nonnegative) ratio of the quantity-to-pivot column substitution rate. This row is the pivot row;
    3. Calculate the new values for the pivot row;
    4. Calculate the new values for the other row(s); and
    5. Calculate the Z_j and $C_j - Z_j$ values for this tableau. If there are any $C_j - Z_j$ values greater than 0, return to step 1. If there is no $C_j - Z_j$ values that are greater than 0, an optimal solution has been reached.

Note that the procedure for solving linear programming minimizations problems is similar to the abovementioned procedure and that the details can be found in Render *et al*. (2009).

A.4 Sensitivity analysis

When an optimal solution for a linear programming problem is found, it is important to know how sensitive the solution is to changes in the data. For example, will a small change in resources or profit per product cause the optimal solution to change?

Linear programming problems are often used to determine quantities for production for a future time period. In real-world situations it is not always possible to know the exact values of, for example, profit per product or available material. Estimated values can be used to solve the problems, but when a solution is found it would be helpful to know how sensitive the optimal solution is to any inexact data.

The basis of sensitivity analysis is the proposition that all parameter values, except for one number, in the model are held fixed (Moore & Weatherford, 2001). By determining a range for each parameter which will not affect the optimal solution, the sensitivity of the solution values can be considered. Sensitivity analysis can be used to determine how much the objective function coefficient of a parameter can change before the objective value changes. A change in the right-hand-side values (or resources) can cause the feasible region to change and it may lead to a different optimal solution; sensitivity analysis can indicate how much these values can change without influencing the optimal solution. Most of these answers can be derived from the final simplex tableau.

A detailed discussion of sensitivity analysis can be found in Render *et al.* (2009).

A.5 The branch-and-bound method

As explained in section 2.6 of Chapter 2, the branch-and-bound method is an implicit enumerative method that can be used to solve integer linear programming problems. Branching only takes place on variables that are required to take on integer values; the feasible region is divided and subproblems are formed and solved. Bounding is used to develop bounds for the different subproblems. By comparing the objective values (or bounds) of the subproblems it is possible to eliminate certain subproblems from consideration (thus, certain feasible solutions cannot improve the current solution and do not have to be investigated further; these points are enumerated implicitly).

Render *et al.* (2009) list the following steps to solve an integer linear programming maximization problem using the branch-and-bound method (in a minimization problem the roles of the upper and lower bounds are reversed):

1. Find an initial solution by solving the linear programming relaxation of the integer programming problem. If integer values are assigned to the required predictor variables, an optimal solution has been found. If not, the objective value provides an initial upper bound;
2. Any feasible solution can be used for a lower bound;
3. Select any predictor variable that is constrained to be integer, but does not have an integer value. Branch the problem into two subproblems based on the integer values that are immediately above and below the non-integer value;
4. Create nodes at the end of the new branches by solving the new problems;
5. a) A branch can be terminated if it yields an infeasible solution;
   b) If a feasible solution for the linear programming problem is found, but it is not an integer solution, go to step 6;
   c) When a feasible integer solution is found, evaluate the objective function. An optimal solution is reached when this value is equal to the upper bound. If this value is less than the upper bound, but exceeds the lower bound, set it as the new lower bound and go to step 6. The branch can be terminated when this value is less than the lower bound; and
6. Inspect the branches and find the maximum value of all the objective functions at the final nodes; set the upper bound equal to this value. If the upper bound is equal to the lower bound, stop. If not, go back to step 3.

To illustrate Dakin's variation (Salkin & Mathur, 1989) of the branch-and-bound method, an example will be presented.

Example A.2 Consider the integer linear programming model in which the objective is to

$$\begin{aligned} \text{Maximize } z &= 4x_1 + 5x_2 \\ \text{subject to} \quad 2x_1 + 3x_2 &\leq 12 \\ 5x_1 + 4x_2 &\leq 20 \\ x_1, x_2 &\geq 0 \\ x_1, x_2 &\in \mathbb{N}_0 \end{aligned}$$

Figure A.2 shows the branch-and-bound tree for this example. Each node refers to a subproblem with the constraint indicated on the arc added to the linear programming problem of the parent node. The values of the variables, x_1 and x_2 are denoted by a vector, $\boldsymbol{x} = (x_1; x_2)$. The linear programming relaxation of the integer programming problem is solved in the root node, node 0, with an optimal solution when $\boldsymbol{x} = (1{,}71; 2{,}86)$ which results in an objective value of $z = 21{,}14$. This value will serve as an initial upper bound. Rounding down gives $\boldsymbol{x} = (1; 2)$ and objective value $z = 14$, which is feasible and can be used as a lower bound. Although $\boldsymbol{x} = (1{,}71; 2{,}86)$ is an optimal solution for the linear programming problem, it is not optimal for the integer linear programming problem. As a result it is necessary to branch on either x_1 or x_2. If branching on x_1 takes place, the region of x_1 is split into two regions, $0 \leq x_1 \leq 1$ and $x_1 \geq 2$. In node 1 the constraint $x_1 \leq 1$ is considered and included into the subproblem. The optimal solution is $\boldsymbol{x} = (1; 3{,}333)$ with $z = 20{,}6667$ (new upper bound). Branching takes place on x_2, splitting the region of x_2 into $x_2 \leq 3$ and $x_2 \geq 4$. Node 2 deals with the constraint $x_2 \leq 3$ resulting in a feasible integer solution with $\boldsymbol{x} = (1; 3)$ and $z = 19$ (new lower bound). The solution at node 3 is also optimal with $\boldsymbol{x} = (0; 4)$ and $z = 20$ (new lower bound).

Returning to node 4, $x_1 \geq 2$ is added to the initial linear programming problem resulting in $\boldsymbol{x} = (2; 2{,}5)$ with an associated objective function of $z = 20{,}5$ (new upper bound). This is still not an integer solution, and therefore branching takes place, resulting in nodes 5 and 6. At node 5, x_2 is constrained to $x_2 \leq 2$, resulting in a solution with $\boldsymbol{x} = (2{,}4; 2)$ and $z = 19{,}6$. This objective function value is lower than the current lower bound and this branch will not be explored further. Node 6, with $x_2 \geq 3$, produces a solution that is infeasible and therefore it can be discarded. Node 3 produced the best feasible integer solution and $\boldsymbol{x} = (0; 4)$ is consequently the optimal solution with an objective value of $z = 20$.

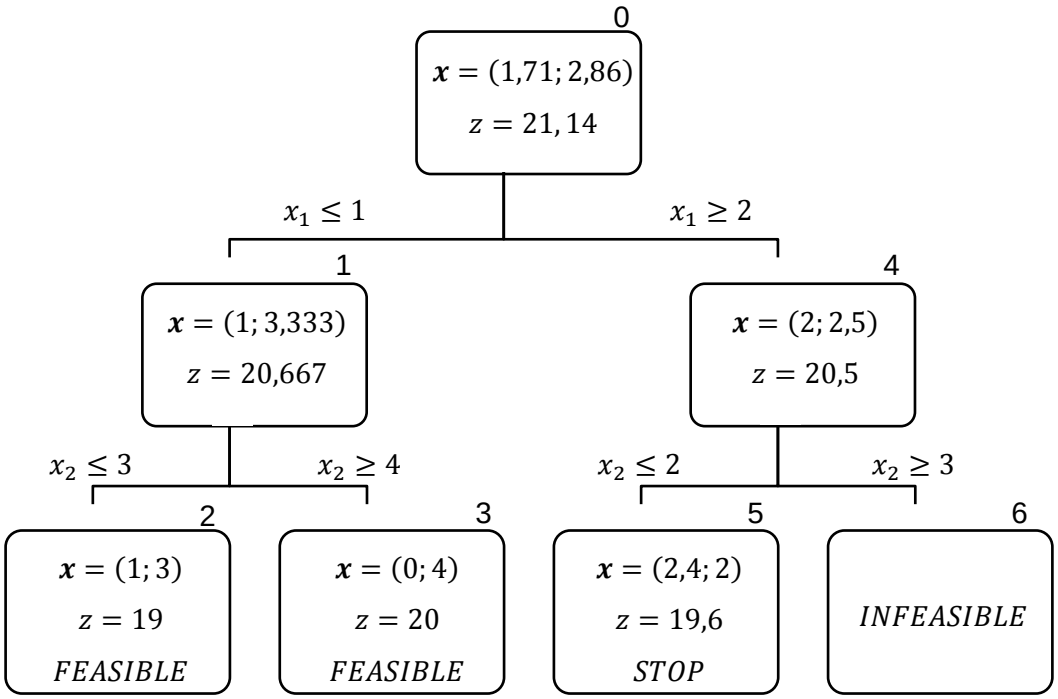

Figure A.2 – The branching tree for the branch-and-bound method in example A.2.

Appendix B

The following paper was presented at the 2nd International Conference on Applied Operational Research held at Turku, Finland on 25 – 27 August 2010. The paper was subjected to a double blind peer review process and was published in 'Lecture Notes on Management Science', August 2010, ISSN: 2008-0050, pp. 34-44 (van der Westhuizen *et al*., 2010).

Bibliography

ATKINSON, A.C. 1986. [Influential observations, high leverage points, and outliers in linear regression]: Comment: aspects of diagnostic regression analysis. *Statistical Science,* 1(3):397-402.

ATKINSON, A.C. 1988. Transformations unmasked. *Technometrics,* 30(3):311-318.

BASSETT, G. & KOENKER, R. 1978. Asymptotic theory of least absolute error regression. *Journal of the American Statistical Association,* 73(363):618-622.

BAZARAA, M.S., JARVIS, J.J. & SHERALI, H.D. 2005. Linear programming and network flows. 3rd ed. Hoboken, NJ: Wiley-Interscience.

BOWERMAN, B.L., O'CONNELL, R.T. & KOEHLER, A.B. 2005. Forecasting, time series, and regression: an applied approach. 4th ed. Belmont, CA: Thomson Brooks/Cole.

BROWNE, M.W. 2000. Cross-validation methods. *Journal of Mathematical Psychology,* 44(1):108-132.

BROWNLEE, K.A. 1965. Statistical theory and methodology in science and engineering. 2nd ed. New York, NY: Wiley.

CHATTERJEE, S. & HADI, A.S. 2006. Regression analysis by example. 4th ed. Hoboken, NJ: Wiley-Interscience.

CHO, M.A. & SKIDMORE, A.K. 2006. A new technique for extracting the red edge position from hyperspectral data: the linear extrapolation method. *Remote Sensing of Environment,* 101(2):181-193.

DIEWERT, W.E. & WALES, T.J. 2005. A 'new' approach to the smoothing problem. (*In* BELONGIA, M.T. & BINNER, J.M., *eds*. Money, measurement and computation. New York: Palgrave Macmillan. p. 104-144.)

DU PLESSIS, P.M. 2010. Contributions to robust regression using the L_1-norm criterion and mixed integer linear programming. Potchefstroom: NWU. (Thesis - Ph.D.)

DYER, J.S., FARRELL, W. & BRADLEY, P. 1973. Utility functions for test performance. *Management Science,* 20(4):507-519.

EFRON, B. & GONG, G. 1983. A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician,* 37(1):36-48.

FAMA, E.F. 1965a. Portfolio analysis in a stable paretian market. *Management Science,* 11(3):404-419.

FAMA, E.F. 1965b. The behavior of stock-market prices. *The Journal of Business,* 38(1):34-105.

GASS, S.I. 1958. Linear programming: methods and aplications. New York: McGraw-Hill.

GEISSER, S. 1975. The predictive sample reuse method with applications. *Journal of the American Statistical Association,* 70(350):320-328.

GILONI, A. & PADBERG, M. 2002. Alternative methods of linear regression. *Mathematical and Computer Modelling,* 35(3-4):361-374.

HARTER, W.L. 1974. The method of least squares and some alternatives: part I. *International Statistical Review,* 42(2):147-174.

HATTINGH, J.M., KRUGER, H.A. & DU PLESSIS, P.M. 2005. Linear model selection: towards a framework using a mixed integer linear programming approach. *South African Statistical Journal,* 39(2):197-220.

HITCHCOCK, C. & SOBER, E. 2004. Prediction versus accommodation and the risk of overfitting. *The British Journal for the Philosophy of Science,* 55(1):1-34.

HOETING, J., RAFTERY, A.E. & MADIGAN, D. 1996. A method for simultaneous variable selection and outlier identification in linear regression. *Computational Statistics & Data Analysis,* 22(3):251-270.

ILOG. 2006. *ILOG CPlex 10.1 User's manual.* France: ILOG.

KUTNER, M.H., NACHTSHEIM, C.J., NETER, J. & LI, W. 2005. Applied linear statistical models. 5th ed. Boston, MA: McGraw-Hill Irwin.

MONTGOMERY, D.C. & PECK, E.A. 1992. Introduction to linear regression analysis. 2nd ed. New York: Wiley.

MOORE, J.H. & WEATHERFORD, L.R. 2001. Decision modeling with Microsoft Excel. 6th ed. Upper Saddle River, NJ: Prentice-Hall.

NETER, J., WASSERMAN, W. & KUTNER, M.H. 1990. Applied linear statistical models. 3rd ed. Homewood, IL: Irwin.

ORTIZ, M.C., SARABIA, L.A. & HERRERO, A. 2006. Robust regression techniques: a useful alternative for the detection of outlier data in chemical analysis. *Talanta,* 70(3):499-512.

RABINOWITZ, P. 1968. Applications of linear programming to numerical analysis. *SIAM Review,* 10(2):121-159.

RENDER, B., STAIR, R.M. & HANNA, M.E. 2009. Quantitative analysis for management. 10th ed. Upper Saddle River, NJ: Pearson Prentice-Hall.

ROUSSEEUW, P.J. & LEROY, A.M. 2003. Robust regression and outlier detection. Hoboken, NJ: Wiley-Interscience.

ROUSSEEUW, P.J. & VAN ZOMEREN, B.C. 1990. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association,* 85(411):633-639.

ROUX, T.P. 1994. ’n Rekenaargebaseerde stelsel om kwantifseerbare aspekte van sosio-ekonomiese en sosio-politiese faktore van lande te ontleed. Potchefstroom: PU vir CHO. (Dissertation - M.Comm.)

RYAN, S.E. & PORTH, L.S. 2007. A tutorial on the piecewise regression approach applied to bedload transport data. *General Technical Report RMRS-GTR-189*:1-41.

SALKIN, H. M. & MATHUR, K. 1989. Foundations of integer programming. New York: North-Holland.

SCHLOSSMACHER, E.J. 1973. An iterative technique for absolute deviations curve fitting. *Journal of the American Statistical Association,* 68(344):857-859.

STONE, M. 1974. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological),* 36(2):111-147.

TAYLOR, B.W. 2001. Introduction to management science. 7th ed. Upper Saddle River, NJ: Prentice-Hall.

VAN DER WESTHUIZEN, M.M., HATTINGH, J.M. & KRUGER, H.A. 2010. Experiments to improve forecasting accuracy of regression models with minimal assumptions. (*In* COLLAN, M., *ed*. Lecture notes in management science: Proceedings of the 2nd International Conference on Applied Operational Research, Turku, Finland. Turku: Uniprint. p. 34-44.)

WAGNER, H.M. 1959. Linear programming techniques for regression analysis. *Journal of the American Statistical Association,* 54(285):206-212.

WAGNER, H.M. 1962. Non-linear regression with minimal assumptions. *Journal of the American Statistical Association,* 57(299):572-578.

WALSH, J.E. 1963. Use of linearized nonlinear regression for simulations involving Monte Carlo. *Operations Research,* 11(2):228-235.

WEISBERG, S. 2005. Applied linear regression. 3rd ed. Hoboken, NJ: Wiley.