Two statistical problems related to credit scoring

Tanja de la Rey (B.Com, M.Sc)

Thesis submitted for the degree Philosophiae Doctor in

Risk analysis at the North-West University

Promoter: Prof. P.J. de Jongh Co-promoter: Prof. F. Lombard

2007

Potchefstroom

Acknowledgements

I want to thank everyone who has assisted me in completing this thesis, specifically the following people:

- Prof. Freek Lombard, University of Johannesburg,
- Prof. Riaan de Jongh, North-West University,
- Prof. Hennie Venter, North-West University.

These people were truly my mentors and without their help, it would have been impossible for me to complete this thesis. I also want to thank all my family and friends for their love and support during the time I was busy with my Ph.D. I want to specifically thank my husband, Arno, for all his encouragement and his unconditional love. Lastly, all the honor to Jesus Christ, my one true Mentor.

Abstract

This thesis focuses on two statistical problems related to credit scoring. In credit scoring of individuals, two classes are distinguished, namely low and high risk individuals (the so-called "good" and "bad" risk classes). Firstly, we suggest a measure which may be used to study the nature of a classifier for distinguishing between the two risk classes. Secondly, we derive a new method DOUW (detecting outliers using weights) which may be used to fit logistic regression models robustly and for the detection of outliers.

In the first problem, the focus is on a measure which may be used to study the nature of a classifier. This measure transforms a random variable so that it has the same distribution as another random variable. Assuming a linear form of this measure, three methods for estimating the parameters (slope and intercept) and for constructing confidence bands are developed and compared by means of a Monte Carlo study. The application of these estimators is illustrated on a number of datasets. We also construct statistical hypothesis to test this linearity assumption.

In the second problem, the focus is on providing a robust logistic regression fit and the identification of outliers. It is well-known that maximum likelihood estimators of logistic regression parameters are adversely affected by outliers. We propose a robust approach that also serves as an outlier detection procedure and is called DOUW. The approach is based on associating high and low weights with the observations as a result of the likelihood maximization. It turns out that the outliers are those observations to which low weights are assigned. This procedure depends on two tuning constants. A simulation study is presented to show the effects of these constants on

the performance of the proposed methodology. The results are presented in terms of four benchmark datasets as well as a large new dataset from the application area of retail marketing campaign analysis.

In the last chapter we apply the techniques developed in this thesis on a practical credit scoring dataset. We show that the DOUW method improves the classifier performance and that the measure developed to study the nature of a classifier is useful in a credit scoring context and may be used for assessing whether the distribution of the good and the bad risk individuals is from the same translation-scale family.

Keywords: credit scoring; quantile comparison function; method of moments; method of quantiles; estimation; asymptotic theory; test of linearity; logistic regression; outliers; robust estimators; trimming; down weighting.

Uittreksel

In hierdie proefskrif word gefokus op twee statistiese probleme wat betrekking het op kredietkeuring (*"credit scoring"*). In kredietkeuring van individue word daar tussen twee risikoklasse onderskei, naamlik lae- en hoë-risiko individue (die sogenaamde "goeie" en "slegte" risikoklasse). Eerstens stel ons ’n maatstaf voor wat gebruik kan word om die aard van ’n klassifiseerder vir die twee risikoklasse te bestudeer. Tweedens stel ons ’n nuwe metode DOUW ("detecting outliers using weights") voor wat gebruik kan word om robuuste passings vir logistiese regressie te bied asook om uitskieters te identifiseer.

In die eerste probleem is die fokus op ’n maatstaf wat die aard van ’n klassifiseerder bestudeer. Hierdie maatstaf transformeer ’n stogastiese veranderlike sodat dit dieselfde verdeling het as ’n ander stogastiese veranderlike. Met die aanname dat hierdie maatstaf ’n lineêre vorm het, word drie metodes ontwikkel vir die beraming van die parameters (helling en afsnit) en vir die konstruksie van vertrouensintervalle. Die beramers word vergelyk deur middel van ’n Monte Carlo studie en hulle toepassing word geïllustreer aan die hand van ’n aantal datastelle. Statistiese hipotese-toetse word gekonstrueer om die aanname dat hierdie maatstaf lineêr is, te toets.

Die fokus in die tweede probleem is op die ontwikkeling van ’n robuuste logistiese regressiepassing en die identifisering van uitskieters. Dit is alombekend dat maksimum aanneemlikheidsberamers van logistiese regressie nadelig beïnvloed word deur uitskieters. Ons bied ’n robuuste metodologie aan wat ook as ’n uitskieteridentifiseringsprosedure dien, bekend as DOUW. Die benadering is daarop gebaseer dat tydens die aanneemlikheids-maksimering proses, hoë en lae gewigte aan elke waarneming

toegeken word. Die uitskieters is dié waarnemings waaraan lae gewigte toegeken word. Die prosedure maak gebruik van twee verstelbare konstantes. ’n Simulasie-studie word aangebied om die uitwerking van hierdie konstantes op die effektiwiteit van die prosedure aan te toon. Die resultate word geïllustreer aan die hand van vier maatstaf-datastelle asook ’n groot nuwe datastel uit die toepassingsveld van klein-handelbemarkingsveldtog-analise.

In die laaste hoofstuk pas ons die tegnieke wat ons in die proefskrif ontwikkel het op ’n praktiese kredietkeuringsdatastel toe. Ons wys dat die DOUW-metode die klassifiseerder se prestasie verbeter, dat die maatstaf wat ontwikkel is om die aard van ’n klassifiseerder te bestudeer, nuttig is in ’n kredietkeuringskonteks en dat dit ook gebruik kan word om te toets of die verdeling van die goeie en die slegte risiko individue van dieselfde translasie-skaal familie is.

Sleutelwoorde: kredietkeuring; kwantielvergelykingsfunksie; metode van momente; metode van kwantiele; beraming; asimptotiese teorie; toets vir lineariteit; logistiese regressie; uitskieters; robuuste beramers; snoeiing; afweging.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

In this thesis we focus on two statistical problems related to retail credit scoring. Thomas et al. (2002) consider retail credit scoring to be one of the most successful applications of statistical modelling in finance and banking. In this chapter we introduce the reader to credit scoring and provide the motivation for the statistical applications we consider. These applications are not exclusive to the credit scoring field and we will comment on other application areas as well. In particular we focus on a measure to study the nature of a classifier and on the identification of outliers when fitting logistic regression models. In Section 1.1 we introduce credit scoring and in Section 1.2 we motivate the need for studying the nature of a classifier. In Section 1.3 we motivate the need for identifying outliers and of fitting logistic regression models robustly.

1.1 Introduction to credit scoring

The term "credit" is used to describe the loan of an amount of money to a customer by a financial institution for a period of time. In such a transaction, the lender wants to be as confident as possible that the money will be repaid in due course. In addition most borrowers would not want to borrow money if there was little chance of them being able to repay it. Thus, there is a need to distinguish between low risk and high risk applicants for credit, both from the lender's and borrower's perspectives. In credit scoring slang the low and high risk classes are frequently referred to as the 'goods' (those individuals posing low risks) and the 'bads' (those individuals posing high risks). Risk, in a credit scoring context, may be described as the ability of the customer to repay the credit granted. In reality there are not simply two well-defined classes: goods and bads. An indeterminate class might also exist. In industry, objective statistical meth-

ods of allocating individuals to risk classes are known as credit scoring methods (see e.g. Thomas et al., 2002 and Hand, 1997). The term "credit scoring" refers to a wide area. Hand (2004) describes credit scoring as the collection of formal statistical and mathematical models used to assist in running financial credit-granting operations, primarily to individual applicants in the personal or retail consumer sectors. The range of applicability of such tools is vast, covering areas such as bank loans, credit cards, mortgages, car finance, hire purchase, mail orders, customer relationship management (CRM) and others.

In a traditional scorecard each response on an application form is assigned a value and the sum of these values for an individual is that individual's overall score (Hand, 1997). This score is then compared to a threshold to produce a classification. Such a score is an application score since it measures the propensity of a new individual to default.

The objective of an application credit scoring model or credit scorecard is therefore to classify new individuals into low (good) and high (bad) credit risk classes with a high degree of accuracy. Often banks or other credit granting institutions use the knowledge obtained through the behaviour of existing customers to aid them in finding classification rules that can be used to decide whether credit may be granted to new applicants or not. In order to construct such a classification rule it is important to analyse the characteristics of existing customers, which may be used to distinguish whether the prospective client pose a good or bad risk to the bank. Typical characteristics are for example the individual's income or demographic information like age or geographic location. Credit scoring models, based on these characteristics, are then developed to classify individuals into good and bad risk classes with a certain degree of accuracy.

The characteristics used in these credit scoring models are frequently referred to as classifiers. Note that in this thesis what we refer to as a "classifier" is often called an independent, an explanatory or an input variable.

In order to build a credit scoring model, a sample of existing customers' credit behaviour is observed over a certain fixed period after which they are classified as good and bad risk individuals. Definitions of good and bad risk individuals may vary between companies and products within companies, but often good risk individuals are those who have never missed a payment and bad risk individuals are those who have defaulted, i.e. missed three consecutive payments or more. This dataset is usually augmented by data on the individuals obtained from credit rating agencies. Suppose a number of individuals have been classified into a good risk and a bad risk class and that we have available a number of characteristics of these individuals which may be used as potential classifiers. Logistic regression models are popular in developing scorecards (or classification rules) and are frequently fitted to such datasets in order to estimate the probability of default given a set of characteristics of the individuals. The best model, that emerges after the variable selection and model building process have been completed, contains those characteristics which may be considered to be the best classifiers.

The above-mentioned model-building process is flawed in that only those "good" applicants, who have been granted credit previously, are now existing customers and only those are now classified as good or bad on the basis of their credit behaviour. This bias problem is referred to as "reject bias" (see e.g. Hand and Henley, 1997). A number of bias correcting techniques have been proposed in the literature under the heading "reject inference". For more detail see e.g. Thomas (2000), Hand (2001),

Hand and Henley (1997), Mays (2004) and Siddiqi (2006). Up to and until now the focus has been on application scoring. An application score can be contrasted with a behavioural score which is a score based on existing borrowers' repayment behaviour and which can be used for such things as deciding what kind of action to take to pursue a mildly delinquent loan or deciding whether to offer a borrower a new loan. In this thesis our focus will not be on the credit scoring model building process. Rather our focus will be to study the nature of the classifiers which emerge from the model building process and to identify outliers and erroneous observations in credit scoring datasets.

Datasets in credit application are large; there may be hundreds of variables (characteristics of applicants) and tens of thousands or even a million cases (applicants). Most of the many variables are typically categorical. If not, the practice until recently was usually to categorise them prior to analysis. However, nowadays the tendency seems to have shifted towards using the uncategorised continuous variables especially when fitting logistic regression models, neural networks or support vector machines to the design set. Of course, in large datasets the number of outliers is also large, and an important pre-processing step is to purge the data of those outliers that are erroneous observations. Therefore, the identification of outlying observations is an important step when building scorecards.

In this thesis we are concerned with two aspects of credit scoring. In Section 1.2 we motivate the need for studying the nature of classifiers. We assume that potential classifiers have been defined by for example a variable selection step and that we need to understand the way in which these classifiers discriminate between the distribution of the goods and the bads. In Section 1.3 we motivate the development of a robust

logistic regression method which may be used for outlier identification. Both these ideas are applicable in other areas as well. These applications will be discussed in more detail later in this thesis.

1.2 Determining the nature of a classifier

Consider a particular classifier X which may be used to classify individuals into good and bad risk classes. Assume that the random variable V represents X for the good risk individuals and that the random variable W represents X for the bad risk individuals. Suppose further that $V \sim F_X$ and $W \sim G_X$ and that F_X and G_X are from the same translation-scale family, say H, i.e. $F_X(x) = H\left(\frac{x-\mu_V}{\sigma_V}\right)$ and $G_X(x) = H\left(\frac{x-\mu_W}{\sigma_W}\right)$. In this context one might expect this assumption to be true since the distribution of X for low and high risk individuals should not differ too much except for location and scale differences. This assumption will be discussed again when credit scoring datasets are studied later. After some mathematical manipulation it follows that

$$G_X^{-1}(F_X(v)) = \mu_W - \frac{\sigma_W}{\sigma_V}\mu_V + \frac{\sigma_W}{\sigma_V}v.$$

Now set $\alpha_0 = \mu_W - \frac{\sigma_W}{\sigma_V}\mu_V$ and $\alpha_1 = \frac{\sigma_W}{\sigma_V}$ then we have that

$$G_X^{-1}(F_X(v)) = \alpha_0 + \alpha_1 v.$$

Note that F_X and G_X are equivalent when $\alpha_0 = 0$ and $\alpha_1 = 1$ and that the difference between F_X and G_X will be indicated by deviations from these values of the alphas. For instance, if $\alpha_1 = 1$ ($\sigma_W = \sigma_V$), then the difference between F_X and G_X are determined by a location difference ($\mu_W - \mu_V$). Obviously if F_X and G_X are not from the same translation-scale family the linear relationship $G_X^{-1}(F_X(v)) = \alpha_0 + \alpha_1 v$

will not hold and some non-linearities will be introduced which will result in a more complex relationship. Ignoring the assumption that F_X and G_X are from the same translation-scale family it follows from the well-known probability integral transformation that $G_X^{-1}(F_X(v)) =_D W$, where $=_D$ denotes equality in distribution. Dropping the subscript X for the sake of notational simplicity we have that the transformation q ($q(v) = G^{-1}(F(v))$) will transform V so that it has the same distribution as W. Doksum (1974) and Doksum and Sievers (1976) have investigated this q function in a medical context, where the objective was to investigate the differences between a control group and a treatment group (see Doksum, 1974 and Doksum and Sievers, 1976). In a similar study Lombard (2005) used it in an application in the coal industry where the objective was for example to distinguish between two methods of measuring abrasive qualities of coal. In our context, we will be more interested in the linear form of q ($q(v) = G^{-1}(F(v)) = \alpha_0 + \alpha_1 v$), because we expect F and G to be from the same translation-scale family.

We now consider two examples whereby we will illustrate the nature of a classifier in distinguishing between two groups. In order to illustrate this graphically we consider random variables V and W and we assume that V represents the good risk class and W the bad risk class as observed through some characteristic. In our first example (Case 1) we assume that $V \sim N(0,1)$ and $W \sim N(2,2^2)$ and that X is the underlying characteristic, and in the second example (Case 2) we assume that $V \sim N(0,1)$ and $W \sim N(10,2^2)$ and that Y is the underlying characteristic. In the left panel of Figure 1.1 we plot the two theoretical densities as observed through characteristic X and in the right panel of Figure 1.1 the two theoretical densities as observed through characteristic Y. Clearly Y distinguishes much better between the good and bad risk

classes than does characteristic X, so Y is the better classifier.

Note that this is if we restrict attention to a single classifier at a time. However, in most practical applications we have many classifiers and then it may be misleading to evaluate the usefulness of a classifier in isolation. This is especially true if the classifiers are highly correlated. In such cases a classifier having the same distribution in both groups, and therefore being useless as a classifier when viewed on its own, may in fact be very useful when combined with other classifiers. We now return to our examples of viewing a single classifier at a time.

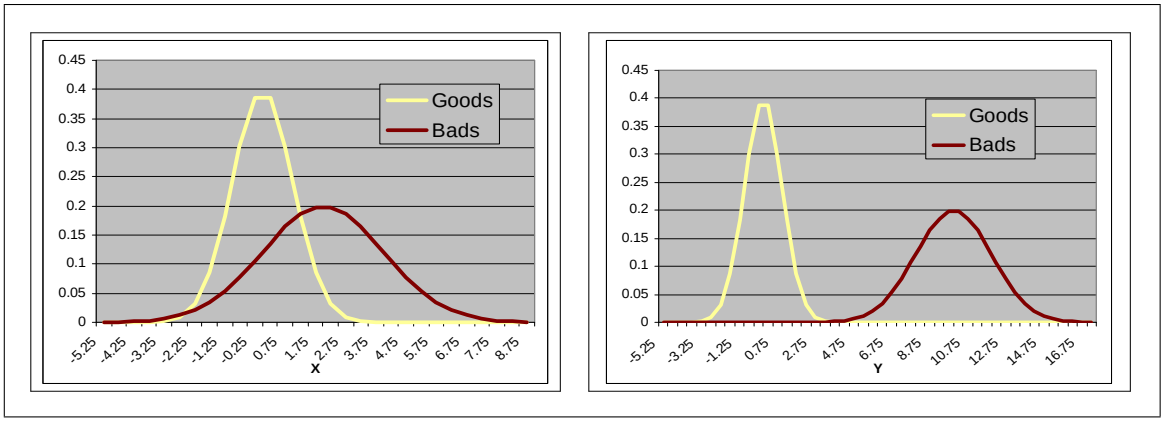


Figure 1.1: Plot of probability densities f (good risks) and g (bad risks) as observed through characteristic X (left) and Y (right)

A diagnostic tool that is frequently used to compare distributions is a QQ-plot where the empirical quantiles of the observed distribution of F are plotted against the empirical quantiles of the observed distribution of G. Deviations from the 45 degree straight line through the origin $q(v) = v$ or $F(v) = G(v)$ indicate that the empirical distribution of F deviates from that of G. We will refer to this 45 degree straight line through the origin as the equal distribution line, or the ED line for short. In order to illustrate this graphically we again study the two random variables V and W in the two above-

mentioned examples and draw a sample of 1000 observations from each distribution. In the left two graphs of Figure 1.2 we plot the frequency distributions and on the right the corresponding quantile plots. In this case, because we have equal sample sizes, our QQ plot is effectively a plot of the ordered observations of V against the ordered observations of W.

In both QQ plots the deviation from the 45 degree line through the origin is clear. In fact, the plotted points in both cases resemble a straight line, but with different slopes and intercepts. Further inspection reveals that the plotted observations in the upper QQ plot in Figure 1.2 (top right panel) resemble a line having an intercept of about 2 and slope of about 2.

The plotted observations in the QQ plot in Figure 1.2 (bottom right panel) again resemble a line but with an intercept of about 10 and a slope of about 2. Therefore the QQ-plot reveals the underlying relationship between the theoretical distributions. From the above examples it should be clear that the QQ-plot may be used to study the nature of the classifiers X and Y. The QQ-plot clearly identifies Y as the better classifier because the plotted observations resemble a line which is further removed than that of X from the line indicating equality between distributions. Also the straight line suggests that the distributions are from the same translation-scale family and may be used to study the magnitudes of location and scale differences. The obvious question is whether this plot may also be useful when a typical credit scoring dataset is used. Two credit scoring datasets will now be considered which will be referred to as Case 3 and Case 4.

The first dataset, Public.xls, was obtained from the CD accompanying the book by Thomas et al. (2002) and we selected the applicant's income ($DAINC$) characteristic

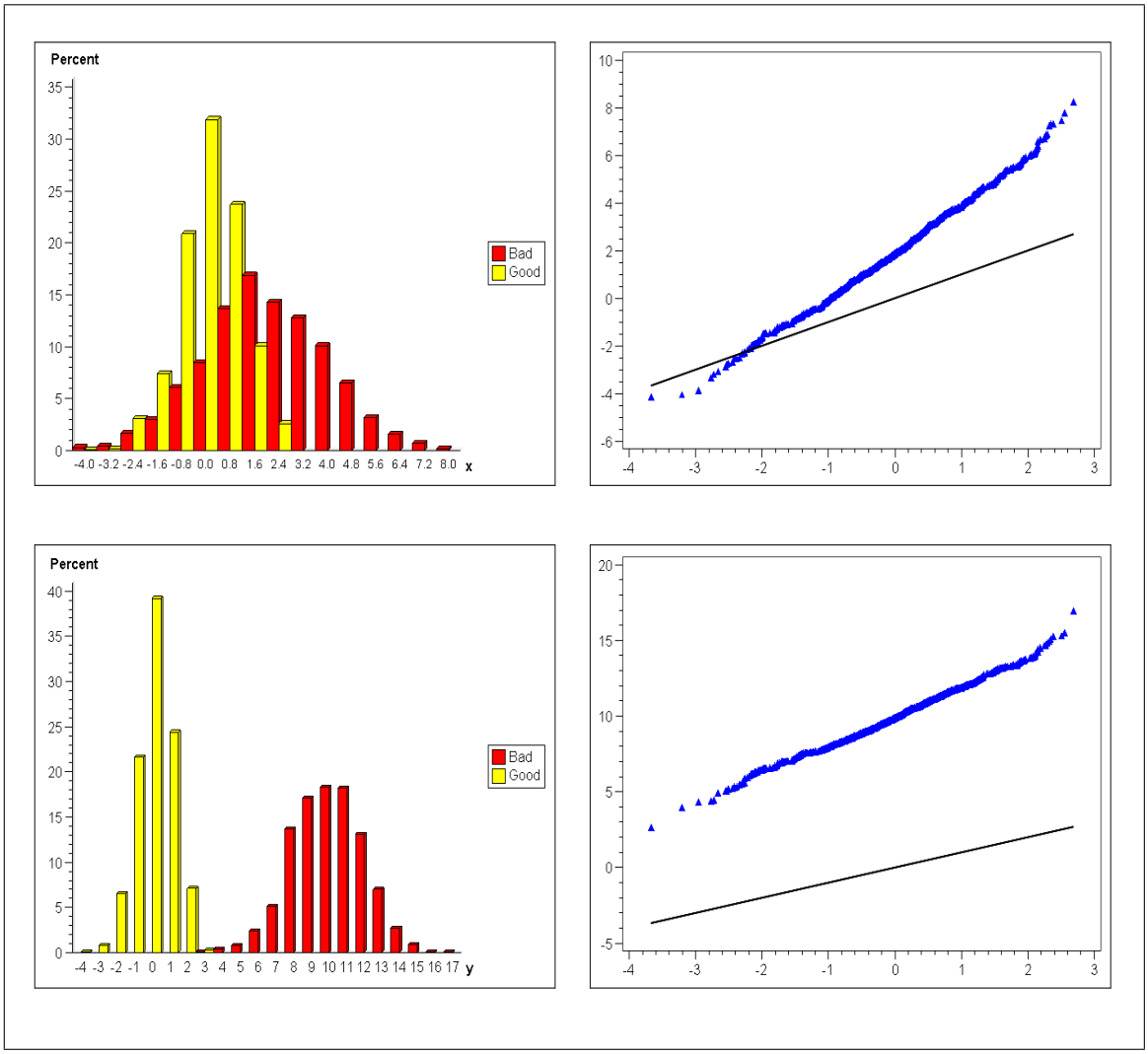


Figure 1.2: Frequency diagrams (left) and QQ plots (right) for Case 1 (top) and for Case 2 (bottom)

(Case 3). Note that the Public.xls dataset has 792 individuals in the good risk class and 227 in the bad risk class.

In the left top graph of Figure 1.3 we plot the frequency distributions and in the right top graph the corresponding quantile plot. In this case, because we have unequal sample sizes, our QQ plot is effectively a plot of $V_{(\lceil mi/n \rceil)}$ against $W_{(i)}$ where $V_{(1)} < \ldots < V_{(m)}$ are the order statistics of $V_1, \ldots, V_m$ and $W_{(1)} < \ldots < W_{(n)}$ are the order statistics of $W_1, \ldots, W_n$. Here $m >= n$ and $\lceil t \rceil$ indicates the ceiling of t.

The second dataset, HMEQ.xls, was obtained from the SAS Course Notes (see e.g. Wielanga et al., 1999). The HMEQ.xls dataset has 4234 individuals in the good risk class and 1045 observations in the bad risk class. We selected the loan amount requested ($LOAN$) as our characteristic (Case 4) and depicted the results in the two bottom panels of Figure 1.3. Note that in both cases the missing values of $DAINC$ and $LOAN$ were excluded.

When inspecting Figure 1.3 (bottom right panel), the plotted observations seem to follow a straight line, although one could argue that non-linearities are present and therefore that the distributions of the goods and the bads are not from the same translation-scale family. In both QQ plots there is very little deviation from the 45 degree line through the origin (ED line). The introduction of confidence bands for this q-function may be used to test the null hypothesis of equality between distributions. In this thesis we will consider various methods for estimating the linear form of q as well as construct associated confidence bands for the estimators.

From the above it is clear that q may be used as an ad hoc measure to study the nature of a classifier in discriminating between two populations. Furthermore it may be useful as a diagnostic measure for assessing whether the distribution of the good and bad

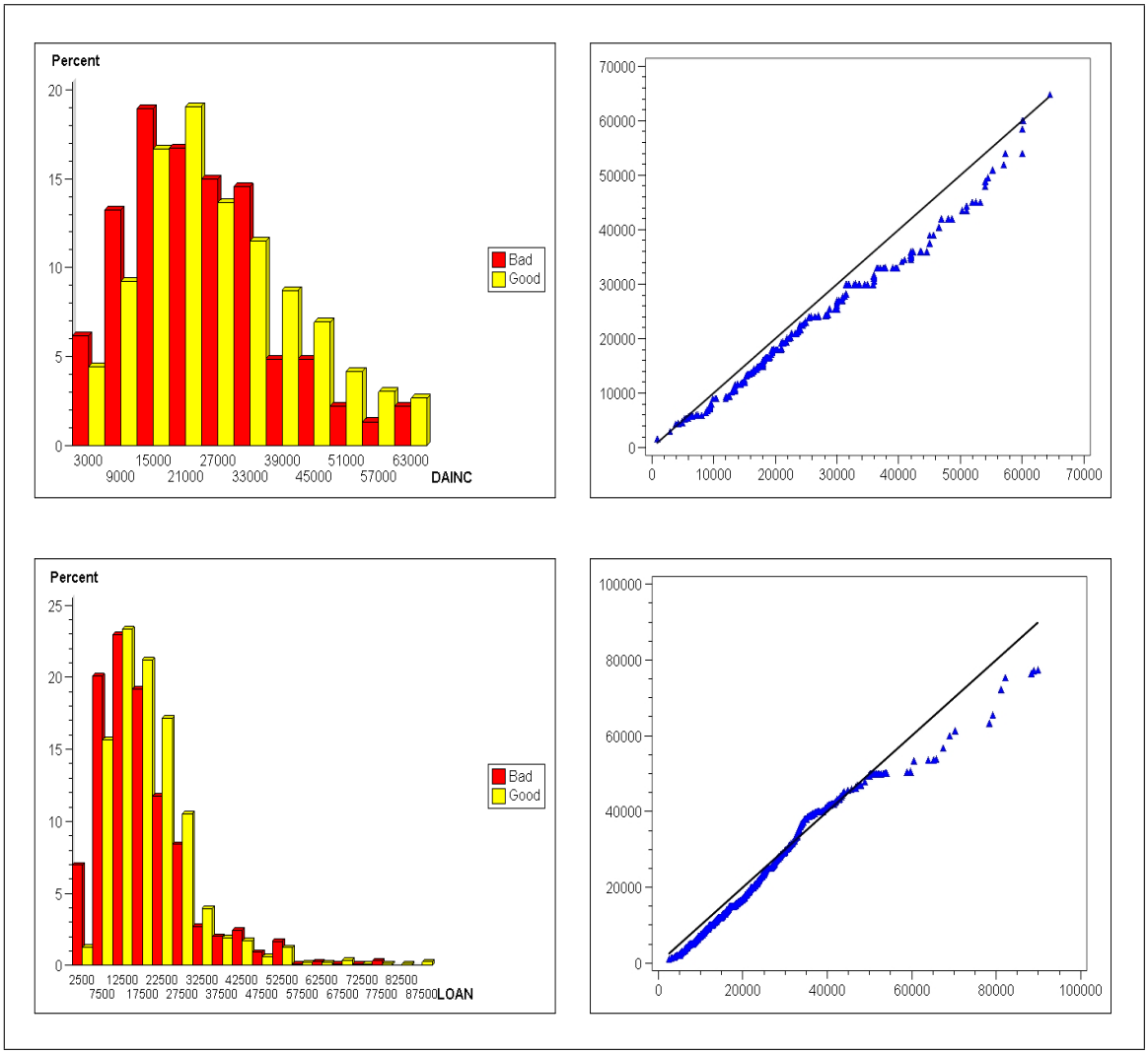


Figure 1.3: Frequency diagrams (left panels) and QQ plots (right panels) for $DAINC$ (Case 3) top panels and for $LOAN$ (Case 4) bottom panels

risk individuals is from the same translation-scale family. Since the assumption of linearity (the distributions are from the same translation-scale family) is central to this study, statistical tests for testing this assumption will be constructed and studied.

As a last observation one should note that this type of analysis is more applicable to continuous variables. Also, an informed reader might ask whether alternatives to the measure proposed here are available. Examples of statistics that are frequently used as measures of separation between two distributions are e.g., Kolmogorov-Smirnov (KS), the c-statistic (see e.g. Siddiqi, 2006) and the Receiver Operating Characteristic (ROC) curve (see e.g. Mays, 2004 or McNab and Wynn, 2000). Note that according to Siddiqi (2006) the c-statistic is equivalent to the area under the ROC curve, Gini coefficient and the Wilcoxon-Mann-Whitney statistic. The measure which we propose here should not be seen as a competitor to the above-mentioned statistics, but rather as another tool for determining the nature of a classifier in distinguishing between distributions.

The QQ plot based measure that we want to investigate is analogous to well-known measures originating from PP plots (see e.g. Mushkudiani and Einmahl, 2007 and Holmgren, 1995). A standard PP plot compares the empirical cumulative distribution function of a variable to a specified theoretical cumulative distribution function such as the normal, but in our case a PP plot will be the empirical cumulative distribution function of the one distribution plotted against the empirical cumulative distribution function of the other distribution. In a PP plot a 45^o straight line indicates that the two distributions are identical. Deviations from this line indicate, as in a QQ plot, that differences between distributions exist. Measures that have been derived based on this include the ROC curve and the Gini coefficient.

1.3 Detecting outliers using weights in logistic regression

Logistic regression (LR) is frequently used in the development of credit scoring models and is concerned with predicting a binary variable (Y) that can take the values 1 (bad risk class) or 0 (good risk class) given a number of independent explanatory variables, or classifiers, say $x_1, ..., x_K$, e.g. the responses on the application form. However, as with most model fitting techniques, logistic regression based on maximum likelihood (ML) is adversely affected by outliers. Outliers occur in most datasets and credit scoring provides no exception. In fact, large datasets often contain many outliers, even after data cleaning procedures have been carried out. In this section we define outliers in a logistic regression context and illustrate how logistic regression maximum likelihood fits are severely affected by outliers.

As stated in the introduction, the good and bad risk classes are usually assigned on the following basis: A sample of existing customers is taken and their behaviour in a particular year recorded. Good is defined as no payment in arrears for that period, while bad is having three or more payments in arrears. Alternatively good may be referred to as the group containing no defaulters and bad the group containing defaulters. Obviously there is an indeterminate class, but as we have noted in Section 1.1, only the good and bad risk classes are used in designing a credit scoring model. We then fit a logistic regression model to determine the default probability. Let $\mathbf{x}^\top = (1, x_1, ..., x_K)$. Then the logistic regression model is given by

$$P(Y = 1) = p(\mathbf{x}, \boldsymbol{\beta}) = 1/\left(1 + \exp\left(-\boldsymbol{\beta}^\top \mathbf{x}\right)\right) \tag{1.1}$$

where $\boldsymbol{\beta}^\top = (\beta_0, \beta_1, ..., \beta_K)$ is a vector of parameters (see e.g. Hosmer and Lemeshow, 1989 and Kleinbaum, 1994). $l(\mathbf{x}) = \boldsymbol{\beta}^\top \mathbf{x}$ is often referred to as the logit value of $\mathbf{x}$ and

$p(\mathbf{x}, \boldsymbol{\beta})$, a function of $l(\mathbf{x})$, as the default probability function or curve of the model we are fitting to the data. Assume that we have N observations, where the n^{th} observation is $(y_n, \mathbf{x}_n^\top)$, with y_n the observed value of Y and $\mathbf{x}_n^\top = (1, x_{n,1}, ..., x_{n,K})$ the vector of observed values of the K regressors. The log likelihood of the N observations is given by

$$\sum_{n=1}^{N} D_n(\boldsymbol{\beta}) \tag{1.2}$$

where

$$D_n(\boldsymbol{\beta}) = y_n \log p(\mathbf{x}_n, \boldsymbol{\beta}) + (1 - y_n) \log(1 - p(\mathbf{x}_n, \boldsymbol{\beta})) \tag{1.3}$$

and the maximum likelihood estimates of $\boldsymbol{\beta}$ are obtained by maximising this expression over $\boldsymbol{\beta}$.

Using an artificial dataset we now investigate outliers for the case of one regressor and illustrate how outliers may affect a logistic regression maximum likelihood fit. The concept of outliers is easiest illustrated by a graphical example, after which a formal definition of outliers will follow.

We construct our dataset by setting $K = 1$ and $\boldsymbol{\beta} = (1, 2)^\top$ in (1.1). Therefore

$$P(Y = 1) = p(x_n, (1, 2)^\top) = 1/\left(1 + \exp\left(-1 - 2x_n\right)\right) \tag{1.4}$$

A sample of 50 (x, y) observations is now constructed by generating x_n, $n = 1, ..., 50$ from a $N(0, 1)$ distribution and the corresponding y_n's are obtained as

$$y_n = \begin{cases} 1, \text{ if } u_n \leq p_n \\ 0, \text{ if } u_n > p_n \end{cases} \tag{1.5}$$

where $p_n = p(x_n, (1, 2)^\top)$ and u_n is independently drawn from a $U(0, 1)$ distribution. The maximum likelihood fit yields $\widehat{\boldsymbol{\beta}} = (1.16, 2.20)^\top$, which is close to the true value of $\boldsymbol{\beta} = (1, 2)^\top$. The true probability curve, $p_n = p(x_n, \boldsymbol{\beta})$ and the estimated probability

curve, $\widehat{p}_n{=}p(x_n, \widehat{\boldsymbol{\beta}})$ are virtually identical and are depicted in the left panel of Figure 1.4. In order to illustrate the effect of outliers we now change the last two observations and set $(x_{n-1}, y_{n-1}) = (4.5, 0)$ and $(x_n, y_n) = (-4.5, 1)$. Note that we have significantly increased the value of x and switched the y-value, in the unexpected direction, in both cases.

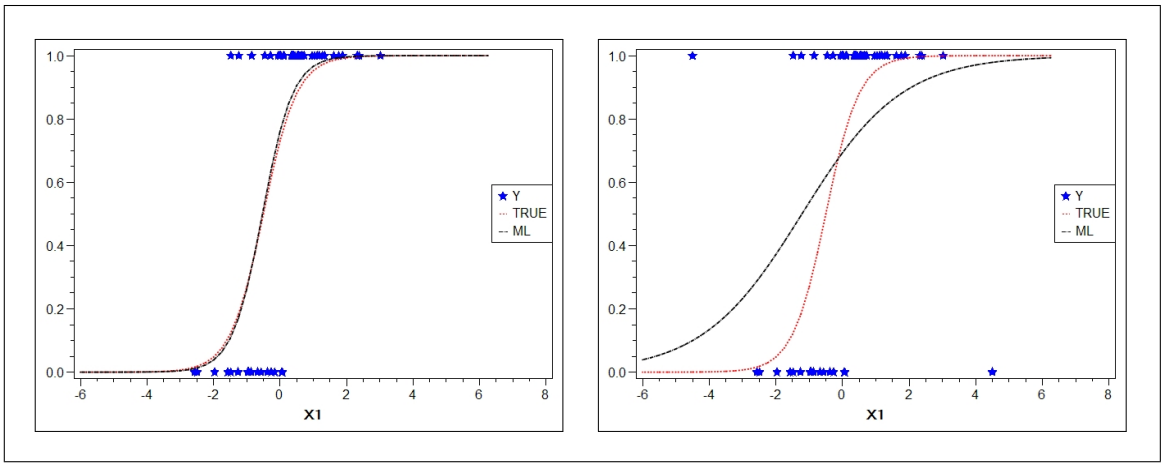


Figure 1.4: True probability curve and estimated probability curve without outliers (left) and with outliers (right)

We now explain this in more detail. In this dataset a high positive x-value causes $p(x_n, \boldsymbol{\beta})$ to be close to 1 and is therefore associated with $y = 1$, while a low x-value is similarly associated with $y = 0$. Both observations, (x_{n-1}, y_{n-1}) and (x_n, y_n), are therefore outliers in the sense that the y-values do not conform to what is expected under the true model. The maximum likelihood (ML) fit on this new dataset yields $\widehat{\boldsymbol{\beta}} = (0.81, 0.67)^\top$ which is quite different from the true parameter value, $\boldsymbol{\beta} = (1, 2)^\top$. In the right panel of Figure 1.4 we show again the true probability curve as well as the fitted maximum likelihood probability curve. It is clear from the two graphs that the introduction of the two outliers severely affected the fit. The true curve distinguishes

well between the y's equal to 1 and the y's equal to 0 as indicated by the steep gradient. However, the fitted curve distinguishes less well between the two populations as indicated by the flatter gradient. The objective of Chapter 3 is to fit a curve that achieves maximum separation between 0's and 1's and is not severely affected by outliers.

Up to now we have not yet given a formal definition of outliers. One can distinguish between outliers in the x-space and in the y-space (or y-direction). We use the artificial dataset from Rousseeuw and Christmann (2003) to graphically illustrate the concept of outliers in a logistic regression context. In Figure 1.5 we graphically depict the dataset. Observations which lie inside the bulk of the x-values (in this example between 1 and 10) are x-inliers, while those outside this area are x-outliers, specifically datapoints (a), (b), (c) and (d). For illustrative purposes we have also added two contours $p(x_n, \beta) = d$ and $p(x_n, \beta) = 1 - d$ (with d small, we expect that if $p(x_n, \beta) < d$ then $y = 0$ and if $p(x_n, \beta) > 1 - d$ we expect $y = 1$). Observations outside of these contours with inappropriate y-values are y-outliers. Copas (1988) calls an observation with $y = 1$ and p close to 0 an "uplier" (for example datapoint (a)) and an observation with $y = 0$ but p close to 1 a "downlier" (for example datapoint (d)). To summarise, datapoint (a) is an uplier and (d) is a downlier, and both are bad leverage points as they adversely affect the fitted curve. On the other hand, datapoints (b) and (c) are good leverage points which reinforce the fit.

To explain outliers in a credit scoring context, we recall that $y = 1$ for customers in the bad risk class (defaulters) and $y = 0$ for customers in the good risk class (non-defaulters). Assume that we only have one explanatory variable, namely the number of delinquent accounts (NDA). Suppose that the higher the NDA, the more likely the

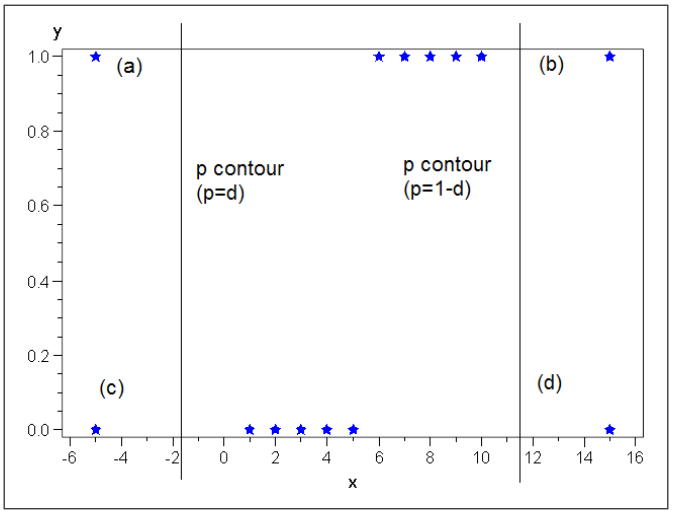


Figure 1.5: x- and y-outliers (one dimension)

customer is in the bad risk class (i.e. the higher the probability of default). Note that in this example, p is the probability of default.

An uplier will then be a customer with a low probability of default, p, (a customer with a low NDA) that is nonetheless a defaulter (i.e. in the bad risk class). This might indicate that the NDA has been captured incorrectly or maybe not all the delinquent accounts are captured in this variable. A downlier is a customer who is in the good risk class although having a high probability of default (i.e. high NDA). Again this might indicate incorrectly captured data or the customer might have a false delinquent account on his/her record. Any customer with a NDA at the extremities of the NDA range will be considered a leverage point. Leverage points which are up- and downliers may be considered to be bad leverage points.

It sometimes happens that the 0's and the 1's can be clearly separated in the x-space. For example, consider the top left panel of Figure 1.6 (the same dataset as in Figure 1.5 was used, with the leverage points removed). It is clear that for $x > 5.5$ all y's equal 1 and for $x < 5.5$ all y's equal 0. In this case the maximum likelihood estimator

(MLE) does not exist. This will be explained in more detail in Chapter 3.

Rousseeuw and Christmann (2003) overcame this non-existence problem by introducing the hidden logistic regression model with an associated estimator referred to as the maximum estimated likelihood (MEL) estimator which always exists even when the y's are perfectly separated. Rousseeuw and Christmann (2003) also proposed a robustified form of the MEL estimator, called the weighted maximum estimated likelihood (WEMEL) estimator. The WEMEL estimator downweights leverage points, where the choice of leverage points is based on robust distances in the regressor space.

We now illustrate the behaviour of the MEL and WEMEL estimators on the datasets of Rousseeuw and Christmann (2003). The estimated probability curves with respect to the MEL and WEMEL estimators are given in Figure 1.6.

The uncontaminated clearly separated dataset is given in the top left hand panel. The other three panels show data that contains outliers. The top right panel contains a downlier; the bottom left a more extreme downlier and the last panel a downlier as well as an uplier. As expected both the MEL and WEMEL estimated curves fit the uncontaminated data well (the fitted curves are indistinguishable). In all other cases the MEL estimated curve is much more affected than the WEMEL estimated curve by the outliers. As seen in Figure 1.6, the WEMEL procedure performs very well as a robust procedure compared to the MEL procedure. However, the WEMEL procedure does not take outliers in the response direction into account and is not really an outlier detection procedure in the sense that it produces a subset of the observations that may be labelled as outliers.

We use a different form of downweighting to introduce a procedure that may be thought of as both a robust logistic regression estimation procedure and an outlier detection

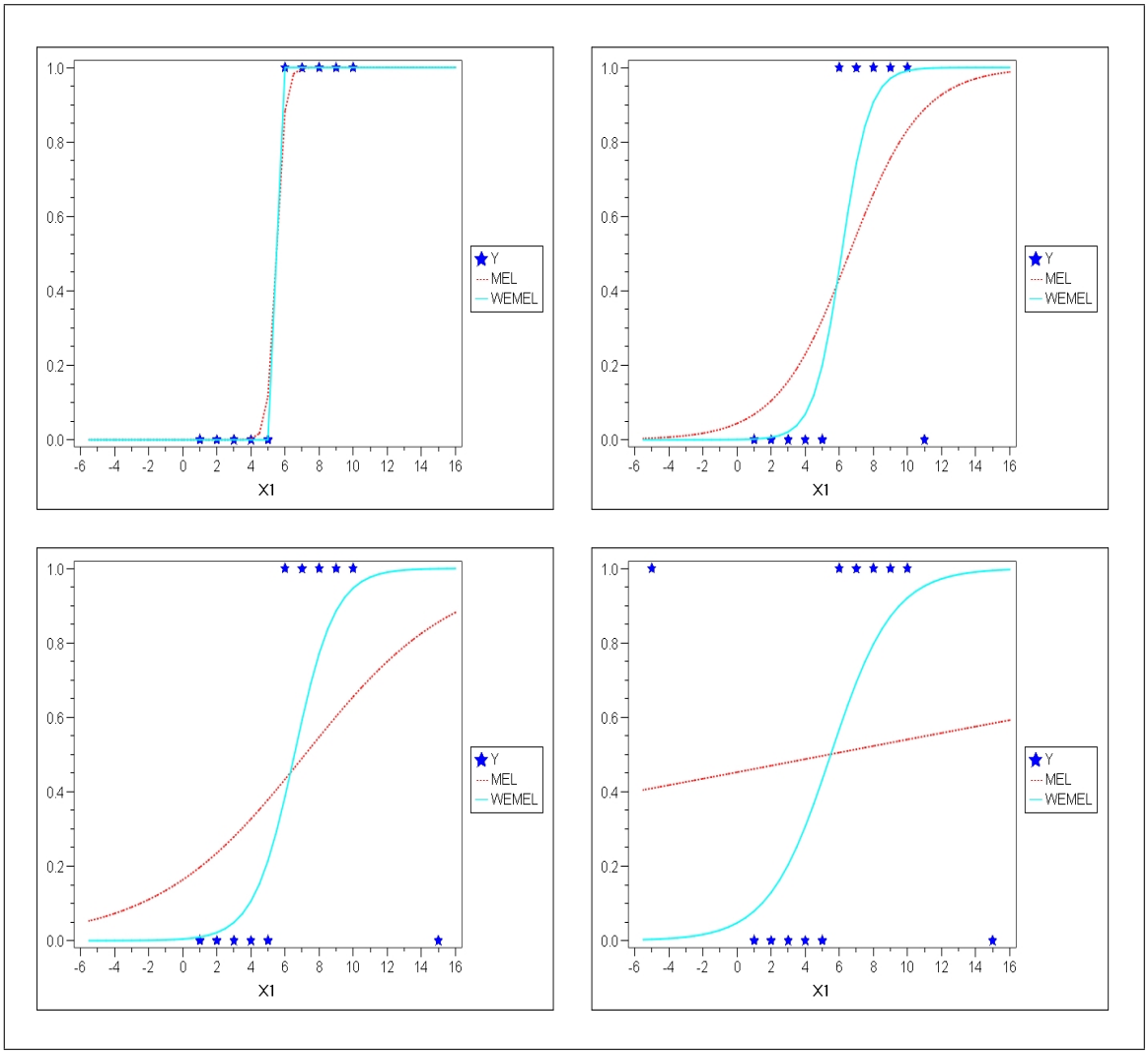


Figure 1.6: Probability curves of MEL and WEMEL

method. The procedure selects two sets of weights, namely high and low weights and then splits the data optimally into two subsets to which the high and the low weights are attached, the subset with the low weights containing the observations that are more likely to be outliers. A corresponding weighted maximum likelihood estimator of the regression coefficients is computed. This is used to estimate the response probabilities of the individual observations. Observations with a $y = 1$ response but low probability for this response and observations with a $y = 0$ response but high probability of this response can then be classified as outliers. The procedure is called detecting outliers using weights (DOUW) and in Chapter 3 we formulate the basic DOUW procedure and list a number of more elaborate versions that can also be used. We also report the result of a simulation study that evaluates the DOUW procedure. Further in Chapter 3 we discuss the application of the DOUW procedure to a number of standard datasets in the literature as well as a new large dataset relating to success probabilities in sales promotion campaigns. Note that the multivariate case ($K > 1$) will be discussed in Chapter 3.

1.4 Summary

In this chapter we have introduced the reader to credit scoring and provided some motivation for the statistical methods developed in this thesis. Specifically, in Section 1.2 we motivated the need for determining the nature of a classifier in credit scoring. The details of this measure will be discussed in Chapter 2. In Section 1.3 we motivated the need for identifying outliers which will be discussed in Chapter 3. In Chapter 4 we will apply the techniques developed in Chapters 2 and 3 to a practical credit scoring dataset.

CHAPTER 2

Determining the nature of a classifier

As stated in Chapter 1 our focus in this chapter will be to find estimators for q and to construct confidence bands based on these estimators. Again we assume that $V \sim F$ and $W \sim G$ where F and G are two unknown distribution functions both assumed to be continuous and strictly increasing. We want to find the transformation q, which will transform the random variable V so that it has the same distribution as W, i.e. $q(V) =_D W$. From the well known probability integral transformation, we know that $F(V) =_D U =_D G(W)$, where U denotes a uniformly distributed random variable. From this we have that $G^{-1}F(V) =_D W$. Therefore we can write $q(v) = G^{-1}F(v)$ as the generic form of q. Lehmann (1974) proposed a non-parametric estimator for the general form of q and Doksum (1974) derived the asymptotic distribution of this estimator. Confidence bands for q based on this estimator were derived by Doksum and Sievers (1976). We provide an overview of these results in Section 2.1.

As mentioned in Chapter 1, Section 1.2, one possibility might be to assume that q has a linear form, e.g. $q(v) = \alpha_0 + \alpha_1 v$. The method of moments estimator for estimating the linear form of q is introduced in Section 2.2 and we derive the asymptotic distribution of the estimator and construct $100(1-\alpha)\%$ confidence bands for q based on the asymptotic results. We compare the non-parametric estimator with the method of moments estimator by means of a Monte Carlo study and illustrate its application in a number of datasets (Section 2.3). As one would expect, the method of moments estimator, because of its semi-parametric nature, leads to a narrower confidence band than does the fully non-parametric estimator. We introduce a further two estimators for the linear form of q namely the method of quantiles estimator and the regression estimator (introduced by Hsieh, 1995). The method of quantiles estimator is defined in Section 2.4 and the asymptotic distribution derived as well as the confidence bands.

We compare the method of moments and the method of quantiles estimator (and associated confidence bands) in Section 2.5 by means of a Monte Carlo study and illustrate their application to a number of datasets. In Section 2.6 the regression estimator will be discussed. Then, we compare the three estimators (method of moments, method of quantiles and regression) in Section 2.7 again by means of a Monte Carlo study and illustrate their application in a number of datasets. Tests of the linearity assumption on q are discussed and analysed in Section 2.8 and some concluding remarks are given in Section 2.9.

2.1 Non-parametric estimator for the generic q-function

As stated in Chapter 1 we want to estimate the general form $q(v) = G^{-1}F(v)$. Lehmann (1974) proposed a non-parametric estimator for $q(v) = G^{-1}F(v)$, replacing G and F by their empirical distribution functions G_n and F_m where n and m denote the respective sample sizes. Thus,

$$\widehat{q}(v) = G_n^{-1}F_m(v). \tag{2.1}$$

Note that $\widehat{q}(V_{(i)}) = G_n^{-1}\left(\frac{i}{m}\right)$. When $m = n$, we have $\widehat{q}(V_{(i)}) = W_{(i)}$, where $V_{(1)} < \ldots < V_{(m)}$ and $W_{(1)} < \ldots < W_{(n)}$ are the order statistics of $V_1, \ldots, V_m$ and $W_1, \ldots, W_n$ respectively. In general, we can calculate $G_n^{-1}\left(\frac{i}{m}\right)$ as follows where $\#(A)$ denotes the cardinality of the set A:

$$\begin{aligned} G_n^{-1}(i/m) &= \inf\left\{t : G_n(t) \geq \frac{i}{m}\right\} \\ &= \inf\left\{t : \#(W \leq t)/n \geq \frac{i}{m}\right\} \end{aligned} \tag{2.2}$$

$$= \inf\{t : \#(W \leq t) \geq ni/m\}$$

$$= W_{(\lceil ni/m \rceil)}$$

where $\lceil t \rceil$ indicates the ceiling of t.

2.1.1 Confidence bands for q based on the non-parametric estimator

Doksum and Sievers (1976) proposed two $100(1-\alpha)\%$ confidence bands for q based on this non-parametric estimator. They referred to these confidence bands as the S-band and the W-band. The S-band is given by the following expression:

$$\left[S_{*(v)}, S^*_{(v)}\right) = \left[W_{(l_{m,n})}, W_{(u_{m,n})}\right) \tag{2.3}$$

where $v \in [V_{(i)}, V_{(i+1)})$; $V_{(0)} = -\infty$ and $V_{(m+1)} = \infty$; $i = 0, \ldots, m$, with

$$\begin{aligned} l_{m,n} &= \left\lfloor n\left(\frac{i}{m} - K_{S,\alpha}/M^{1/2}\right)\right\rfloor \\ u_{m,n} &= \left\lceil n\left(\frac{i}{m} + K_{S,\alpha}/M^{1/2}\right)\right\rceil + 1 \end{aligned} \tag{2.4}$$

where $M = \frac{mn}{m+n}$; $W_{(j)} = -\infty$ $(j < 0)$ and $W_{(j)} = \infty$ $(j \geq n+1)$ and where $\lfloor t \rfloor$ indicates the floor of t. $K_{S,\alpha}$ is chosen from the Kolmogorov-Smirnov tables (Pearson and Hartley, 1972, Table 55), i.e.

$$P(D_{n+m} \leq K_{S,\alpha}) = 1 - \alpha, \tag{2.5}$$

where

$$D_{n+m} = \sqrt{\frac{mn}{m+n}} \sup_v |F_m(v) - G_n(v)|. \tag{2.6}$$

The W-band is given by

$$\left[W_{*(v)}, W^*_{(v)}\right) = \left[W_{(l_{m,n})}, W_{(u_{m,n})}\right) \tag{2.7}$$

where $v \in [V_{(i)}, V_{(i+1)})$; $i = 0, \ldots, m$ with

$$\begin{aligned} l_{m,n} &= \lfloor nh^{-}(u) \rfloor \\ u_{m,n} &= \lceil nh^{+}(u) \rceil \end{aligned} \tag{2.8}$$

where

$$h^{\pm}(u) = \{u + 1/2c(1-\lambda)(1-2\lambda u) \pm \frac{1}{2}\left\{c^2(1-\lambda)^2 + 4cu(1-u)\right\}^{1/2}\}/\left(1 + c(1-\lambda)^2\right),$$

with $c = K_\alpha^2/M$, $u = F_m(v) = i/m$ and $G_n(v) = i/n$. K_α is chosen to satisfy the probability statement

$$P(W_{n+m} \leq K_\alpha) = 1 - \alpha \tag{2.9}$$

where

$$W_{n+m} = \sqrt{\frac{mn}{m+n}} \sup_v \frac{|F_m(v) - G_n(v)|}{\sqrt{\Psi(v)\left[1 - \Psi(v)\right]}} \tag{2.10}$$

and

$$\Psi(v) = \frac{m}{n+m} F_m(v) + (1 - \frac{m}{n+m}) G_n(v). \tag{2.11}$$

Canner (1975) provides Monte Carlo estimates of K_α. We now use an artificial dataset to illustrate the behaviour of the S- and W-bands. The dataset is created by sampling 100 observations from a $N(0,1)$ distribution and 100 from a $N(0,4)$ distribution. The estimate $\widehat{q}$ and the S- and W bands are depicted in Figure 2.1. As in Doksum and Sievers (1976) we find that the W-band is narrower than the S-band. This corresponds with the theoretical results by Doksum and Sievers. The ED (equal distribution) line is not contained fully in either of the confidence bands, which confirms that the distribution of the characteristic, X, is different between the two populations. The S- and W-bands are both narrower in the centre of the distributions (of V and W) than in the tails. In the tails both bands become wider and the bands do not entirely

cover the tails of the distributions. Note that the S-band has the advantage of being simpler to construct than the W-band and that its critical values are more extensively tabulated (Doksum and Sievers, 1976). Note also that the plotted observations seem to follow a straight line and this therefore suggests that the distributions are from the same translation-scale family.

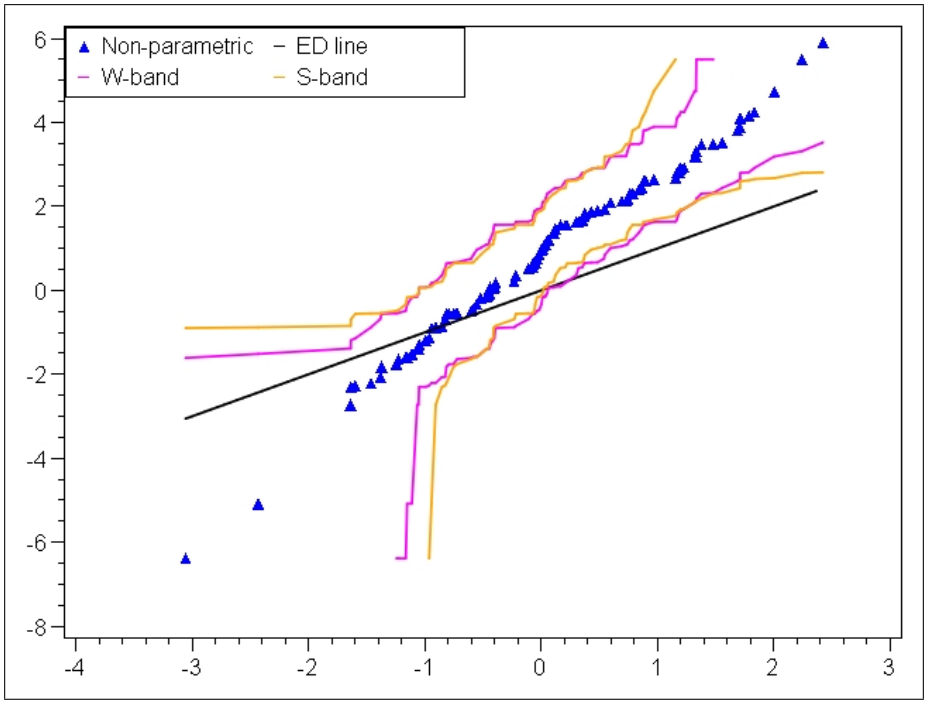


Figure 2.1: Non-parametric estimate with S- and W-bands

It seems plausible from the plot of $\widehat{q}$ in Figure 2.1 that q could be a linear function as was the case with the four examples (Cases 1 - 4) considered in Section 1.2. As we have mentioned in Section 1.2 we also expect that an estimator based on the linear form of q will lead to a narrower confidence band than the non-parametric estimator.

2.2 Method of moments estimator for q

In this section we focus on estimating the linear form of the q-function and we use the method of moments to derive an estimator for q. We first describe the estimator, then derive its asymptotic distribution, and based on this, derive confidence bands for q. As an alternative to the confidence bands based on the asymptotic distribution we propose bootstrap confidence bands and compare the two sets of bands by means of a Monte Carlo study. In this section we assume that the distributions of V and W have finite moments up to order four.

There are many ways in which the distributions of V and W can differ. Taking $q(v) = \alpha_0 + \alpha_1 v$ amounts to considering only the possibility that V and W are related by a location and scale change, which is one of the simplest ways in which they can differ. Per definition $q(V) =_D W$, therefore,

$$W =_D \alpha_0 + \alpha_1 V. \tag{2.12}$$

Let μ_W and μ_V denote the means, and σ_W and σ_V the standard deviations, of W and V respectively. We see from (2.12) that

$$\mu_W = \alpha_0 + \alpha_1 \mu_V \text{ and } \sigma_W^2 = \alpha_1^2 \sigma_V^2. \tag{2.13}$$

Therefore

$$\alpha_1 = \frac{\sigma_W}{\sigma_V} \text{ and } \alpha_0 = \mu_W - \frac{\sigma_W}{\sigma_V}\mu_V. \tag{2.14}$$

We can estimate α_0 and α_1 by the method of moments by substituting $\overline{W}$ for μ_W, $\overline{V}$ for μ_V, s_W^2 for σ_W^2 and s_V^2 for σ_V^2 in (2.14), so that

$$\widehat{\alpha}_1 = \frac{s_W}{s_V} \tag{2.15}$$

$$\widehat{\alpha}_0 = \overline{W} - \frac{s_W}{s_V}\overline{V}. \tag{2.16}$$

The method of moments estimator for q is then

$$\begin{aligned} \widehat{q}_{MOM}(v) &= \widehat{\alpha}_0 + \widehat{\alpha}_1 v & (2.17)\\ &= \overline{W} - \frac{s_W}{s_V}\overline{V} + \frac{s_W}{s_V}v. & (2.18) \end{aligned}$$

2.2.1 Asymptotic distribution of $\widehat{q}_{MOM}$

Theorem 1 *The asymptotic distribution of $\widehat{q}_{MOM}$ is given by the expression*

$$\sqrt{m+n}(\widehat{q}_{MOM}(v) - q(v)) \sim N(0, \tau(v)^2) \tag{2.19}$$

where

$$\tau(v)^2 = \sigma_0^2 + \sigma_1^2\widetilde{v}^2 + 2\sigma_{0,1}\widetilde{v}, \tag{2.20}$$

$$\widetilde{v} = v - \overline{V} \tag{2.21}$$

and

$$\sigma_0^2 = \frac{\sigma_W^2}{\lambda(1-\lambda)}, \tag{2.22}$$

$$\sigma_{0,1} = \frac{\kappa_3(V)\sigma_W^2}{2\lambda\sigma_V^4} + \frac{\kappa_3(W)}{2(1-\lambda)\sigma_V\sigma_W}, \tag{2.23}$$

$$\sigma_1^2 = \frac{\kappa_4(V)\sigma_W^2}{4\lambda\sigma_V^6} + \frac{\sigma_W^2}{2\lambda(1-\lambda)\sigma_V^2} + \frac{\kappa_4(W)}{4(1-\lambda)\sigma_V^2\sigma_W^2}. \tag{2.24}$$

κ_i indicates the i^{th} cumulant and $\lambda = m/(m+n)$.

The proof of Theorem 1 is given in Appendix A.1.1.

2.2.2 Confidence band for q based on the method of moments estimator

In this section we construct a simultaneous confidence band for q based on the asymptotic distribution of $\widehat{q}_{MOM}$. We denote the asymptotic variance of $\widehat{q}_{MOM}$ by

$$\sigma^2(v) = \frac{\tau(v)^2}{m+n} \tag{2.25}$$

and estimate this by

$$\widehat{\sigma}^2(v) = \frac{\widehat{\tau}(v)^2}{m+n}. \tag{2.26}$$

Using (2.20), we estimate $\tau(v)^2$ by

$$\widehat{\tau}(v)^2 = \widehat{\sigma}_0^2 + \widehat{\sigma}_1^2 \left(v - \overline{V}\right)^2 + 2\widehat{\sigma}_{0,1}\left(v - \overline{V}\right) \tag{2.27}$$

where

$$\widehat{\sigma}_0^2 = \frac{s_W^2}{\lambda(1-\lambda)} \tag{2.28}$$

$$\widehat{\sigma}_1^2 = \frac{\widehat{\kappa}_4(V)s_W^2}{4\lambda s_V^6} + \frac{s_W^2}{2\lambda(1-\lambda)s_V^2} + \frac{\widehat{\kappa}_4(W)}{4(1-\lambda)s_V^2 s_W^2} \tag{2.29}$$

$$\widehat{\sigma}_{0,1} = \frac{\widehat{\kappa}_3(V)s_W^2}{2\lambda s_V^4} + \frac{\widehat{\kappa}_3(W)}{2(1-\lambda)s_V s_W}. \tag{2.30}$$

A $100(1-\alpha)\%$ confidence band

$$\widehat{q}_{MOM}(v) - c_{\alpha,m+n}\widehat{\sigma}(v) \le q(v) \le \widehat{q}_{MOM}(v) + c_{\alpha,m+n}\widehat{\sigma}(v) \quad \forall \ v \tag{2.31}$$

can now be obtained if we can find the constant $c_{\alpha,m+n}$ satisfying the probability statement

$$P\left(\sup_v \left|\frac{\widehat{q}_{MOM}(v) - q(v)}{\widehat{\sigma}(v)}\right| \le c_{\alpha,m+n}/\sqrt{m+n}\right) = 1-\alpha. \tag{2.32}$$

Theorem 2 *The asymptotic value of* $c_{\alpha,m+n}$ *in (2.32) is* $c_\alpha = \sqrt{-2\log_e(\alpha)}$.

The proof of Theorem 2 is in Appendix A.1.2. Obviously, the continuous nature of this confidence band (2.31) will ensure that the entire distributions (of V and W) are covered while this is not the case for the non-parametric bands.

2.3 Empirical study

We investigate the confidence band for q based on the method of moments by means of a Monte Carlo study and then illustrate its application in a number of datasets. Two versions of the band (2.31) will be considered, namely where the critical value, $c_{\alpha,m+n}$, is estimated by $\sqrt{-2\log_e(\alpha)}$, the asymptotic value, and alternatively where $c_{\alpha,m+n}$ is estimated using the bootstrap. The first band will be referred to as the asymptotic confidence band (the A-band) and the second as the bootstrap confidence band (the B-band). In the Monte Carlo study we investigate whether the estimated coverage probability of the confidence band is close to the nominal coverage probability. Then the behaviour of the confidence bands will be illustrated by using the four examples discussed in Chapter 1 (Section 1.2) and compared with the S- and W-bands of the non-parametric estimator.

2.3.1 Monte Carlo study

In this section we investigate by means of a Monte Carlo study whether the coverage probability of the confidence band (2.31) for q based on the method of moments estimator, is close to the nominal coverage probability. We expect that the estimated coverage probability will be close to the nominal value in large samples but not necessarily in small samples. It is shown in Theorem 2 (Appendix A.1.2) that the inequality

$$\sup_v \left| \frac{\widehat{q}_{MOM}(v) - q(v)}{\widehat{\sigma}(v)} \right| \leq c_{\alpha,m+n}/\sqrt{m+n} \tag{2.33}$$

is equivalent to the inequality

$$(m+n) \begin{bmatrix} \widetilde{\gamma}_1 \\ \widetilde{\gamma}_0 \end{bmatrix}^\top \begin{bmatrix} \widehat{\sigma}_1^2 & \widehat{\sigma}_{0,1} \\ \widehat{\sigma}_{0,1} & \widehat{\sigma}_0^2 \end{bmatrix}^{-1} \begin{bmatrix} \widetilde{\gamma}_1 \\ \widetilde{\gamma}_0 \end{bmatrix} \leq c_{\alpha,m+n}^2. \tag{2.34}$$

The two gamma parameters are defined in (A.56) and (A.57) in Theorem 2, Appendix A. Denote the left hand side of (2.34) by S. Then equation (2.32) can equivalently be expressed as

$$P(S \leq c^2_{\alpha,m+n}) = 1 - \alpha. \tag{2.35}$$

Therefore, in the two Monte Carlo studies below we estimate the coverage probability of (2.31) by estimating the left hand side of (2.35).

In the Monte Carlo studies we assume some distribution for F and G and then generate equal size samples ($m = n$) from each of those distributions and repeat the process L times. The estimated coverage probability is obtained as

$$(1/L) \sum_{l=1}^{L} I\left(S_l \leq c^2\right) \tag{2.36}$$

where S_l denotes the value of S obtained in the l^{th} sample. We take $c = \sqrt{-2\log_e(\alpha)}$ (the asymptotic critical value) or $c = c^*$ (a bootstrap estimated critical value). The algorithm to obtain the estimated coverage probability is given in Algorithm 1 (Appendix A.2) and the algorithm for obtaining a bootstrap estimated critical value is given in Algorithm 2 (Appendix A.2). Throughout the thesis we used the smooth bootstrap (see Calculation 1, Appendix A.3) and used the relevant subroutines in PROC IML of SAS (SAS Institute, 2003) to generate random numbers.

In our first Monte Carlo study we assume $F \sim N(1, 2^2)$ and $G \sim N(2, 2^2)$ and generate our samples from these distributions, while in the second study, we generate samples from the empirical distributions of the good and bad risk classes of the income ($DAINC$) variable (taken from the Public.xls dataset, discussed in Chapter 1, Section 1.2). In both Monte Carlo studies we consider sample sizes of $100, 200, 500$ and 1000 and $L = 1000$ simulation runs.

The results of the first study are given in Tables 2.1 and 2.2. Table 2.1 shows the estimated coverage probability using the asymptotic critical value and Table 2.2 the estimated coverage probability using the bootstrap estimate of the critical value.

Nominal coverage	**Estimated coverage probability**			
probability	$m = n = 100$	$m = n = 200$	$m = n = 500$	$m = n = 1000$
90%	86.61%	87.23%	88.16%	88.27%
95%	91.95%	92.72%	93.26%	93.32%
97.50%	94.98%	95.45%	96.18%	96.38%

Table 2.1: Coverage probabilities for normal data when using the asymptotic critical value

Nominal coverage	**Estimated coverage probability**			
probability	$m = n = 100$	$m = n = 200$	$m = n = 500$	$m = n = 1000$
90%	90.50%	90.48%	89.81%	89.75%
95%	95.03%	95.12%	95.23%	94.85%
97.50%	97.68%	97.53%	97.37%	97.04%

Table 2.2: Coverage probabilities for normal data when using the bootstrap estimate of the critical value

We see that, especially in small samples, the estimated coverage probabilities are less than the nominal coverage probabilities when the asymptotic critical value is used. Also, as expected, the estimated coverage probabilities are closer to the nominal values as the sample size increases. However, the estimated coverage probabilities using the bootstrap estimate of the critical value, are much closer to the nominal coverage probabilities at all sample sizes considered. The improvement is especially

clear in small samples ($m = n = 100, 200$).

The results of the second study are given in Tables 2.3 and 2.4. As before, Table 2.3 contains the estimated coverage probabilities using the asymptotic critical value and Table 2.4 the estimated coverage probabilities using the bootstrap estimate of the critical value. As in the previous study, the estimated coverage probabilities are found to be less than the nominal coverage probabilities when using the asymptotic critical value while the estimated coverage probabilities obtained using the bootstrap estimated value are close to the nominal coverage probabilities.

Nominal coverage	**Estimated coverage probability**			
probability	$m = n = 100$	$m = n = 200$	$m = n = 500$	$m = n = 1000$
90%	85.53%	84.71%	89.51%	87.77%
95%	89.37%	93.02%	93.28%	94.75%
97.50%	96.60%	96.74%	96.92%	97.12%

Table 2.3: Coverage probabilities for the income variable when using the asymptotic critical value

Nominal coverage	**Estimated coverage probability**			
probability	$m = n = 100$	$m = n = 200$	$m = n = 500$	$m = n = 1000$
90%	91.26%	90.28%	90.36%	89.72%
95%	95.23%	95.16%	94.96%	95.02%
97.50%	98.16%	97.72%	97.38%	97.39%

Table 2.4: Coverage probabilities for the income variable when using the bootstrap estimate of the critical value

Because the bootstrap confidence bands yield estimated coverage probabilities which are close to the nominal values over a range of sample sizes and a range of nominal coverage probabilities, the bootstrap confidence band should be preferred to the asymptotic confidence band.

An area for future work is to derive a correction term for the asymptotic value of c to make it perform better for modest sample sizes.

2.3.2 Examples

Given the above-mentioned results, we will focus our attention on the bootstrap confidence band. The behaviour of the confidence bands will now be practically illustrated by using the four examples discussed in Chapter 1 (Section 1.2).

Recall, in the first two examples, the standard normal is compared with a $N(2, 2^2)$ (Case 1) and then with $N(10, 2^2)$ (Case 2). Note that in Case 1 and Case 2 the standard normal distribution represents the good risk class, while the bad risk class is represented by the $N(2, 2^2)$ distribution in Case 1 and by the $N(10, 2^2)$ distribution in Case 2. In each case 1000 observations were generated from each of these distributions and $\widehat{q}_{MOM}$ and $\widehat{q}$ and associated confidence bands estimated. The results are depicted in Figures 2.2 and 2.3 for Case 1 and Case 2, respectively. As expected, in both examples, the S-band is wider than the W-band (see Doksum and Sievers 1976) and both bands tend to be wide in the tails of the distributions and also do not cover the distributions entirely. It is also clear from the graphs that the 95% B-band is always narrower than the S- and W-band in the tails of the distribution. This is to be expected due to the semi-parametric nature of $\widehat{q}_{MOM}$. Note that in both examples considered, the ED lines are not contained in the confidence bands, so that it can be concluded

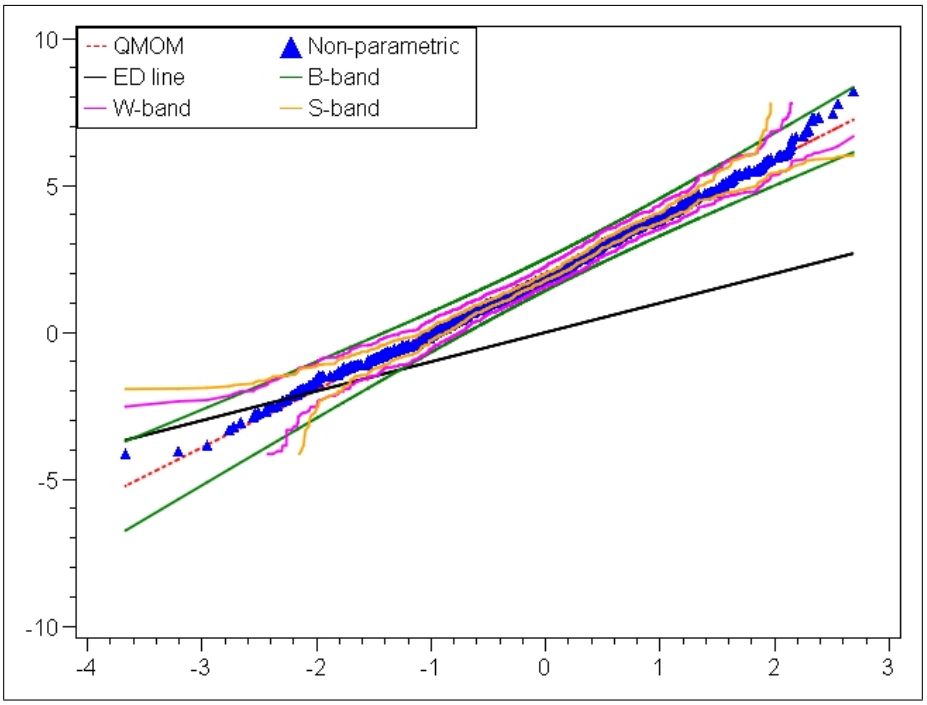


Figure 2.2: MOM and non-parametric estimate (with B-, S- and W-bands), Case 1

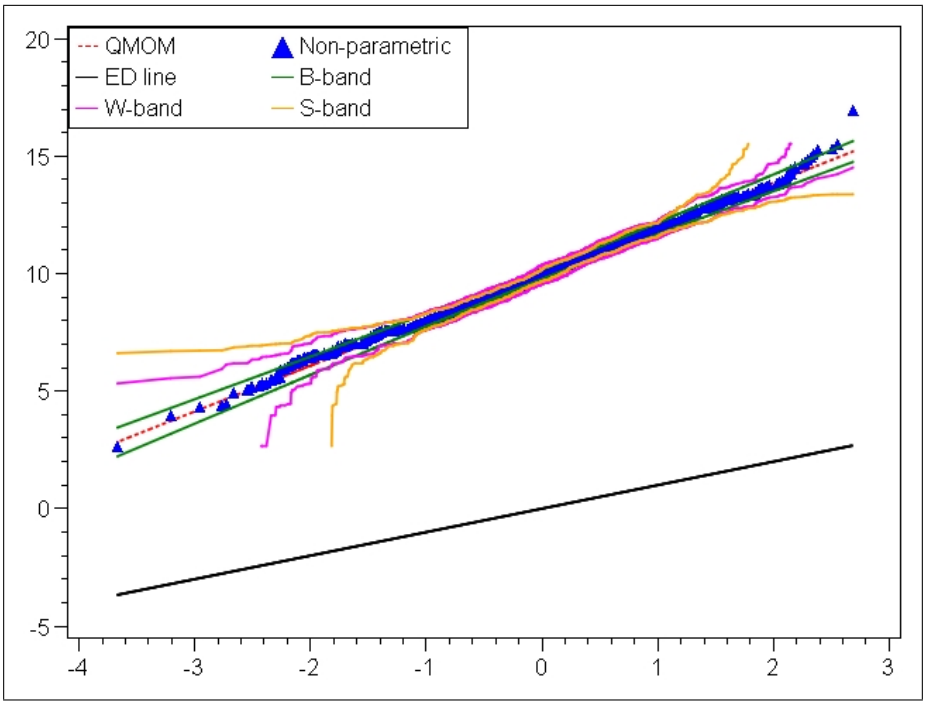


Figure 2.3: MOM and non-parametric estimate (with B-, S- and W-bands), Case 2

that there are clear differences between the distributions considered. See the remark at the end of this section.

Our next two examples are based on the "credit scoring" datasets where the $DAINC$ (Case 3) and the $LOAN$ (Case 4) classifiers are used to distinguish between good and bad risk classes. Note again that 792 observations are in the good risk class and 227 in the bad risk class in the dataset where $DAINC$ is used as a classifier and 4234 observations in the good risk class and 1045 in the bad risk class in the dataset where $LOAN$ is used. In this case we only estimate $\widehat{q}_{MOM}$ and the associated bootstrap confidence band for each example. The results are depicted in Figures 2.4 and 2.5. Note that in the last two examples the ED lines are included in the 95% bootstrap confidence bands so that it can be concluded that the classifiers do not distinguish well between the good and bad risk classes.

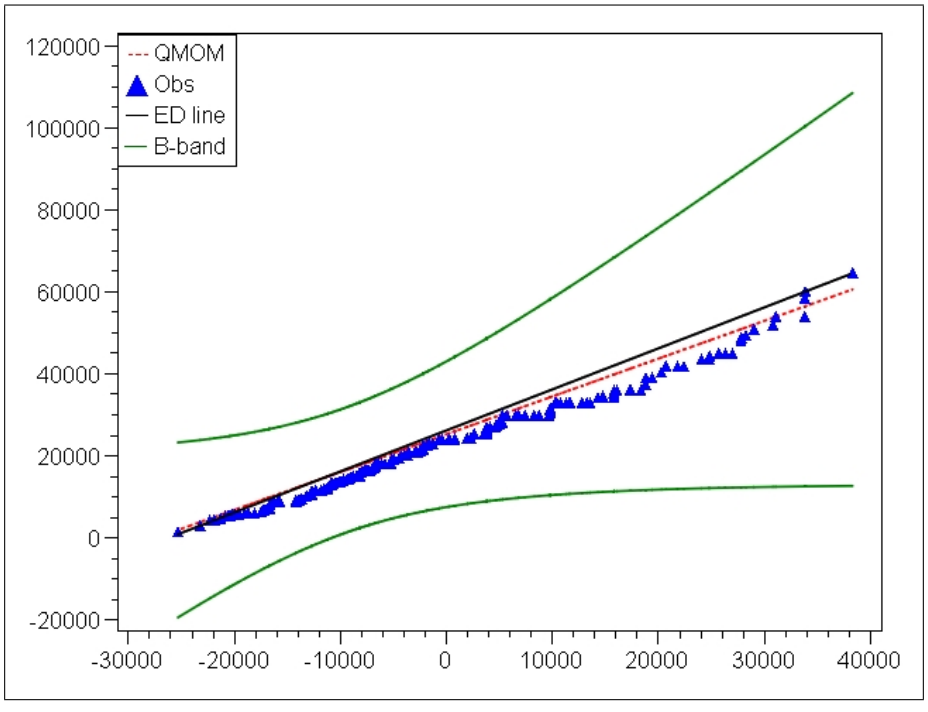


Figure 2.4: Method of moments estimate (with B-band) for $DAINC$ (Case 3)

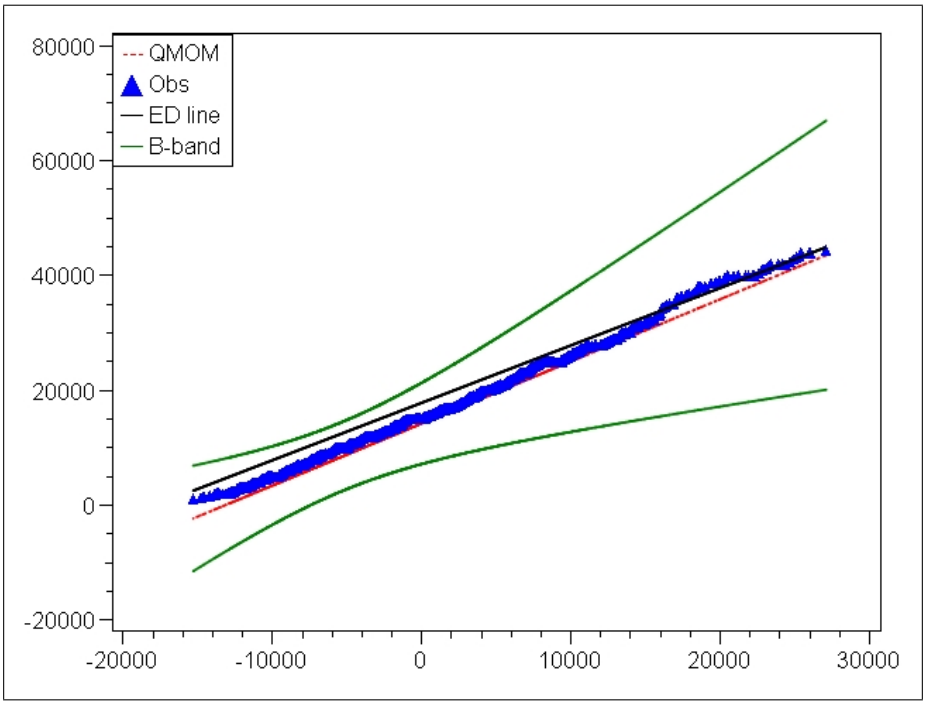


Figure 2.5: Method of moments estimate (with B-band) for $LOAN$ (Case 4)

To conclude we want a confidence band that gives the required coverage probability and a confidence band that is narrow and therefore we recommend that the B-band be used in practice, as it gives the required coverage probability over the distributions considered and a narrow confidence band, especially in the tails of the distributions in the examples considered.

In Chapter 1 we claimed that the QQ plot may be used to study the nature of classifiers. Considering the four examples just analysed, it should be clear at this stage that the estimated q function may be used to study the performance of classifiers in discriminating between the good and the bad risk classes as well as to suggest whether the two distributions (goods and bads) come from the same translation-scale family.

Remark

From the previous discussion it should be clear that $\widehat{q}_{MOM}$ and associated bootstrap confidence band may be used to test the null hypothesis $H_0 : F = G$ against the alternative hypothesis $H_a : F \neq G$. The null hypothesis will not be rejected if the ED line $q(v) = v$ is contained within the confidence bands, however, if the ED line is not fully contained within the confidence bands, we reject the null hypothesis. In Figure 2.2 and in Figure 2.3 the ED line is not fully contained in the B-bands and the hypothesis that the two distributions are identical is rejected at a significance level of 5%.

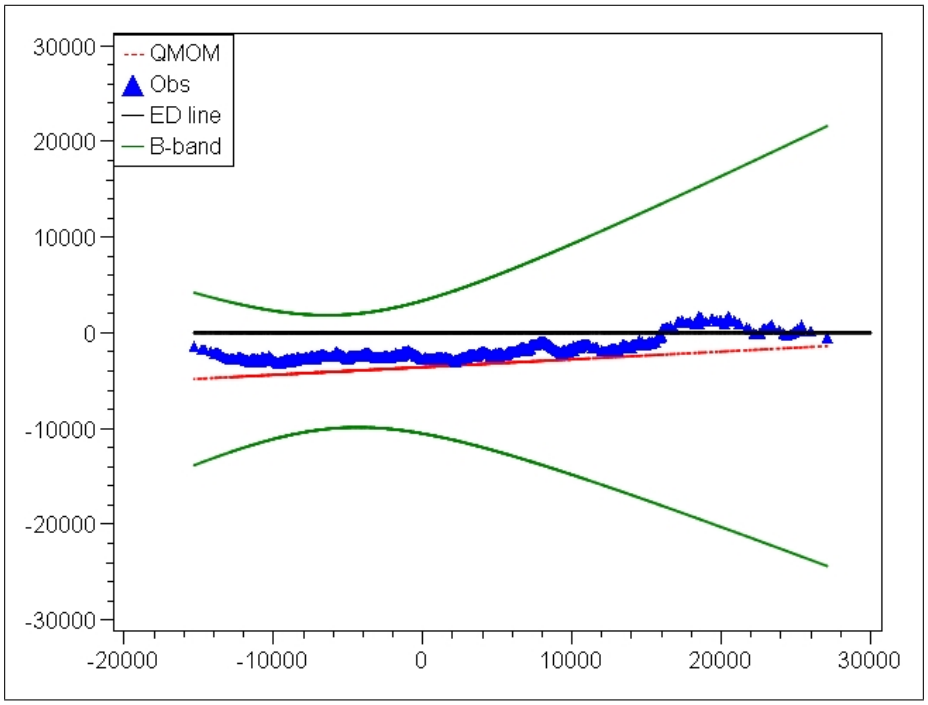


Figure 2.6: Method of moments estimate with associated confidence bands for $LOAN$ (alternative plot for Case 4)

In Figure 2.4 and Figure 2.5 the ED lines are contained in the B-bands indicating that the variables, $DAINC$ and $LOAN$, do not distinguish well between the good and the

bad risk classes. In order words, the hypothesis that the two distributions are identical cannot be rejected on a 5% significance level. In order to simplify interpretation these plots may be transformed so that the ED line becomes the $y = 0$ line. This is obtained by plotting $\widehat{q}(v) - v$ against $\frac{v-\overline{V}}{s_V}$ as shown in Figure 2.6. The ED line in Figure 2.5 was the 45^o line and now becomes the line $y = 0$ in Figure 2.6.

2.4 Method of quantiles estimator for q

From (2.23) and (2.24) it is clear that the asymptotic variance of the method of moments estimator is based on the third and fourth moments of V and W. It is well known that higher moments (3^{rd} and 4^{th}) have a tendency to become unstable (see e.g. van der Vaart, 1998 or Lehmann, 1999). A more robust alternative might be to replace the mean in the method of moments estimator with the median and the standard deviation with the interquartile range, which will lead to a more stable covariance matrix. We refer to this alternative estimator as the method of quantiles estimator, $\widehat{q}_{MOQ}(v)$.

In this section we introduce the method of quantiles estimator for q and as in the previous section we first describe the estimator, then derive the asymptotic distribution thereof and based on these, propose confidence bands for q. As an alternative to the confidence bands based on the asymptotic distribution we propose bootstrap confidence bands and compare the two sets of bands by means of a Monte Carlo study. We also illustrate the application of the method of moments estimator and the method of quantiles estimator in a number of datasets.

The method of quantiles estimator is

$$\widehat{q}_{MOQ}(v) = \widehat{\alpha}_0 + \widehat{\alpha}_1 v \tag{2.37}$$

where

$$\widehat{\alpha}_1 = \frac{\widehat{i}_W}{\widehat{i}_V} \tag{2.38}$$

$$\widehat{\alpha}_0 = \widehat{m}_W - \frac{\widehat{i}_W}{\widehat{i}_V}\widehat{m}_V. \tag{2.39}$$

$\widehat{m}_V$, $\widehat{m}_W$, $\widehat{i}_V$ and $\widehat{i}_W$ denote the sample equivalents of m_V, m_W, i_V and i_W. The p-quantiles of F and G will be denoted by ξ_p and η_p, respectively. Then

$$m_V = \xi_{\frac{1}{2}}, \tag{2.40}$$

$$m_W = \eta_{\frac{1}{2}}, \tag{2.41}$$

$$i_V = \xi_{\frac{3}{4}} - \xi_{\frac{1}{4}} \tag{2.42}$$

and

$$i_W = \eta_{\frac{3}{4}} - \eta_{\frac{1}{4}}. \tag{2.43}$$

2.4.1 The asymptotic distribution of $\widehat{q}_{MOQ}$

Theorem 3 *The asymptotic distribution of $\widehat{q}_{MOQ}$ is given by the expression*

$$\sqrt{m+n}(\widehat{q}_{MOQ}(v) - q(v)) \sim N(0, \tau_{MOQ}(v)^2) \tag{2.44}$$

where

$$\tau_{MOQ}(v)^2 = C_1 + C_2\,(v - m_V) + C_3\,(v - m_V)^2 \tag{2.45}$$

and

$$C_1 = \frac{i_W^2}{4\lambda i_V^2 f_V^2(m_V)} + \frac{1}{4\,(1-\lambda)\,f_W^2(m_W)} \tag{2.46}$$

$$C_2 = -\frac{i_W}{2\sqrt{\lambda}\sqrt{1-\lambda} f_W^2(m_W)} + \frac{1}{4\,(1-\lambda)\,i_V f_W(m_W)}\left(\frac{3}{4f_W(\eta_{\frac{3}{4}})}\right. \tag{2.47}$$

$$\left. -\frac{1}{4f_W(\eta_{\frac{1}{4}})}\right) \tag{2.48}$$

$$C_3 = \frac{i_W^2}{4\lambda f_W^2(m_W)} - \frac{i_W}{4\sqrt{\lambda}\sqrt{1-\lambda} i_V f_W(m_W)}\left(\frac{3}{4f_W(\eta_{\frac{3}{4}})} - \frac{1}{4f_W(\eta_{\frac{1}{4}})}\right) + \frac{1}{(1-\lambda)\, i_V^2}\left(\frac{3}{4f_W(\eta_{\frac{3}{4}})} - \frac{1}{4f_W(\eta_{\frac{1}{4}})}\right)^2 \quad (2.49)$$

where f_V and f_W are the probability density functions of V and W respectively. Again $\lambda = m/(m+n)$.

The proof of Theorem 3 is given in Appendix A.1.3.

2.4.2 Confidence band for q based on the method of quantiles estimator

In this section we construct a simultaneous confidence band for q based on the asymptotic distribution of $\widehat{q}_{MOQ}$. The derivation of the confidence bands follows along the same lines as the derivation of the confidence bands of the method of moments. We denote the asymptotic variance of $\widehat{q}_{MOQ}$ by

$$\sigma^2_{MOQ}(v) = \frac{\tau_{MOQ}(v)^2}{m+n} \quad (2.50)$$

and estimate this by

$$\widehat{\sigma}^2_{MOQ}(v) = \frac{\widehat{\tau}_{MOQ}(v)^2}{m+n}. \quad (2.51)$$

Using (2.45), we estimate $\tau_{MOQ}(v)^2$ by

$$\widehat{\tau}_{MOQ}(v)^2 = \widehat{C}_1 + \widehat{C}_2\,(v - \widehat{m}_V) + \widehat{C}_3\,(v - \widehat{m}_V)^2 \quad (2.52)$$

where $\widehat{\tau}_{MOQ}(v)^2$ is obtained by substituting all the quantities in (2.45) with the sample equivalents. $\widehat{f}_V$ and $\widehat{f}_W$ are kernel estimates of the probability density functions, f_V and f_W (see Calculation 3, Appendix A.3).

A $100(1-\alpha)\%$ confidence band

$$\widehat{q}_{MOQ}(v) - d_{\alpha,m+n}\widehat{\sigma}_{MOQ}(v) \le q(v) \le \widehat{q}_{MOQ}(v) + d_{\alpha,m+n}\widehat{\sigma}_{MOQ}(v) \quad \forall\ v \quad (2.53)$$

can be obtained if we can find a constant $d_{\alpha,m+n}$ satisfying the following probability statement

$$P\left(\sup_{v}\left|\frac{\widehat{q}_{MOQ}(v)-q(v)}{\widehat{\tau}_{MOQ}(v)}\right|\leq d_{\alpha,m+n}/\sqrt{m+n}\right)=1-\alpha \tag{2.54}$$

Theorem 4 *The asymptotic value of* $d_{\alpha,m+n}$ *in (2.54) is* $d_\alpha=\sqrt{-2\log_e(\alpha)}$.

Given the result in (2.44), it follows exactly as in the case of $\widehat{q}_{MOM}$ that the asymptotic value of $d_{\alpha,m+n}$ is $d_\alpha=\sqrt{-2\log_e(\alpha)}$.

2.5 Empirical study

We investigate the confidence band for q based on the method of quantiles by means of a Monte Carlo study and then illustrate its application in a number of datasets. In the Monte Carlo study we investigate whether the estimated coverage probability of the confidence band is close to the nominal coverage probability. As before, we investigate the asymptotic band (2.53) as well as the bootstrap band.

2.5.1 Monte Carlo study

In this section we investigate by means of a Monte Carlo study whether the coverage probability of the confidence band (2.53) for q based on the method of quantiles estimator, is close to the nominal coverage probability. The Monte Carlo study here follows along the same lines as the Monte Carlo study used to investigate the coverage probability of the confidence band of q based on the method of moments estimator. As in Theorem 2 (Appendix A.1.2), it can be shown in Theorem 4 (Appendix A.1.4) that the inequality

$$\sup_{v}\left|\frac{\widehat{q}_{MOQ}(v)-q(v)}{\widehat{\tau}_{MOQ}(v)}\right|\leq d_{\alpha,m+n}/\sqrt{m+n} \tag{2.55}$$

is equivalent to the inequality

$$(m+n)\begin{bmatrix}\widetilde{\gamma}_1^* \\ \widetilde{\gamma}_0^*\end{bmatrix}^{\top}\begin{bmatrix}\widehat{C}_1 & \frac{\widehat{C}_2}{2} \\ \frac{\widehat{C}_2}{2} & \widehat{C}_3\end{bmatrix}^{-1}\begin{bmatrix}\widetilde{\gamma}_1^* \\ \widetilde{\gamma}_0^*\end{bmatrix} \leq d^2_{\alpha,m+n}. \tag{2.56}$$

The two gamma parameters are given in Theorem 4, Appendix A. Denote the left hand side of (2.56) by S_{MOQ}. Then equation (2.54) can equivalently be expressed as

$$P\left(S_{MOQ} \leq d^2_{\alpha,m+n}\right) = 1-\alpha. \tag{2.57}$$

In the two Monte Carlo studies below we estimate the coverage probability of (2.53) by estimating the left hand side of (2.57).

Our Monte Carlo studies are designed as follows. We assume some distribution for F and G and then generate equal size samples ($m = n$) from each of those distributions and repeat the process L times. The estimated coverage probability is obtained as

$$(1/L)\sum_{l=1}^{L} I\left(S_{MOQ,l} \leq d^2\right) \tag{2.58}$$

where $S_{MOQ,l}$ denotes the value of S_{MOQ} obtained in the l^{th} sample, and where d can take the value $\sqrt{-2\log_e(\alpha)}$ (the asymptotic critical value) or d^* (a bootstrap estimated critical value). The algorithm to obtain the coverage probability is the same as in the case for the method of moments (Algorithm 1, Appendix A.2), except that S has to be substituted by S_{MOQ} and the algorithm for obtaining a bootstrap estimated critical value is also the same, see Algorithm 2 (Appendix A.2).

In our first Monte Carlo study we assume $F \sim N(1, 2^2)$ and $G \sim N(2, 2^2)$ and generate our samples from these distributions, while in the second study we generate samples from the empirical distributions of the good and bad risk classes of the income ($DAINC$) variable (taken from the Public.xls dataset, discussed in Chapter 1). In both

Monte Carlo studies we consider sample sizes of $100, 200, 500$ and 1000 and $L = 1000$ simulation runs.

The results of the first study are given in Tables 2.5 and 2.6. Table 2.5 contains the estimated coverage probability using the asymptotic critical value and Table 2.6 the estimated coverage probability using the bootstrap estimate of the critical value. We see that, especially in small samples, the estimated coverage probabilities are less than the nominal coverage probability when using the asymptotic critical value. Also, as expected, the estimated coverage probabilities are closer to the nominal coverage probabilities as the sample size increases. However, the estimated coverage probabilities using the bootstrap estimate of the critical value, are much closer to the nominal coverage probabilities in all sample sizes considered. The improvement is especially clear in small samples ($m = n = 100, 200$).

Nominal coverage	**Estimated coverage probability**			
probability	$m = n = 100$	$m = n = 200$	$m = n = 500$	$m = n = 1000$
90%	85.6%	86.1%	86.8%	86.9%
95%	91.7%	92.5%	92.8%	93.8%
97.50%	93.1%	94.9%	93.7%	95.3%

Table 2.5: Coverage probabilities for normal data when using the asymptotic critical value

The results of the second study are given in Tables 2.7 and 2.8. As previously, Table 2.7 contains the estimated coverage probabilities using the asymptotic critical value and Table 2.8 the estimated coverage probabilities using the bootstrap estimate of the critical value. Again, the estimated coverage probabilities are found to be less than the

Nominal coverage	**Estimated coverage probability**			
probability	$m = n = 100$	$m = n = 200$	$m = n = 500$	$m = n = 1000$
90%	88.6%	89.1%	91.4%.	91.5%
95%	94.0%	94.6%	94.9%	95.1%
97.50%	98.7%	97.2%	97.3%	98.0%

Table 2.6: Coverage probabilities for normal data when using the bootstrap estimate of the critical value

nominal coverage probabilities when using the asymptotic critical value, even in large sample sizes ($m = n = 1000$). We can therefore assume that the asymptotic theory of MOQ estimation only holds with very large samples. As in the previous study, the estimated coverage probabilities obtained using the bootstrap estimated value, are very close to the nominal coverage probabilities at all the sample sizes considered. Because the bootstrap confidence band provides much better estimates of the nominal coverage probabilities than the asymptotic confidence bands, it is recommended for practical applications.

Nominal coverage	**Estimated coverage probability**			
probability	$m = n = 100$	$m = n = 200$	$m = n = 500$	$m = n = 1000$
90%	73.3%	73.0%	74.5%	79.1%
95%	72.3%	82.8%	83.7%	88.6%
97.50%	76.3%	86.2%	86.2%	87.6%

Table 2.7: Coverage probabilities for the income variable when using the asymptotic critical value

Nominal coverage	Estimated coverage probability			
probability	$m = n = 100$	$m = n = 200$	$m = n = 500$	$m = n = 1000$
90%	88.6%	92.6%	91.4%	90.9%
95%	94.3%	94.6%	94.3%	95.3%
97.50%	91.9%	97.6%	97.3%	97.6%

Table 2.8: Coverage probabilities for the income variable when using the bootstrap estimate of the critical value

Comparing the results obtained in this section with the corresponding results obtained for method of moments, we observe that the B-band is preferred in both cases, as it gives the required coverage probability and in both cases the A-band is less than the nominal coverage probabilities. However, note that the A-band based on the method of quantiles estimator is less than the nominal coverage probability to a greater extent due to the fact that the asymptotic theory of method of quantiles estimation only holds in very large samples. A further Monte Carlo comparison will be made between the method of moments (MOM) and the method of quantiles (MOQ) estimators in Section 2.7.1 using bias and mean squared error.

2.5.2 Examples

The behaviour of the bootstrap confidence bands for MOM and MOQ will now be practically illustrated by again using the four examples discussed in Section 1.2. Again, in the first two examples, the standardised normal is compared with a $N(2, 2^2)$ (Case 1) and then with $N(10, 2^2)$ (Case 2). Recall that in Case 1 and Case 2 the standardised normal distribution represents the good risk class, while the bad risk class is repre-

sented by the $N(2, 2^2)$ distribution in Case 1 and by the $N(10, 2^2)$ distribution in Case 2. In each case 1000 observations were generated from each of these distributions and $\widehat{q}_{MOM}$ and $\widehat{q}_{MOQ}$ and associated confidence bands estimated. The results are depicted in Figures 2.7 and 2.8 for Case 1 and Case 2, respectively. As expected, in both cases the 95% B-band of the method of quantiles is wider than the 95% B-band of the method of moments estimate. This could be explained by the fact that in the case of normally distributed data, the variance will be higher when estimating a quantile as opposed to a moment (van der Vaart, 1998, Chapter 4 and Chapter 21).

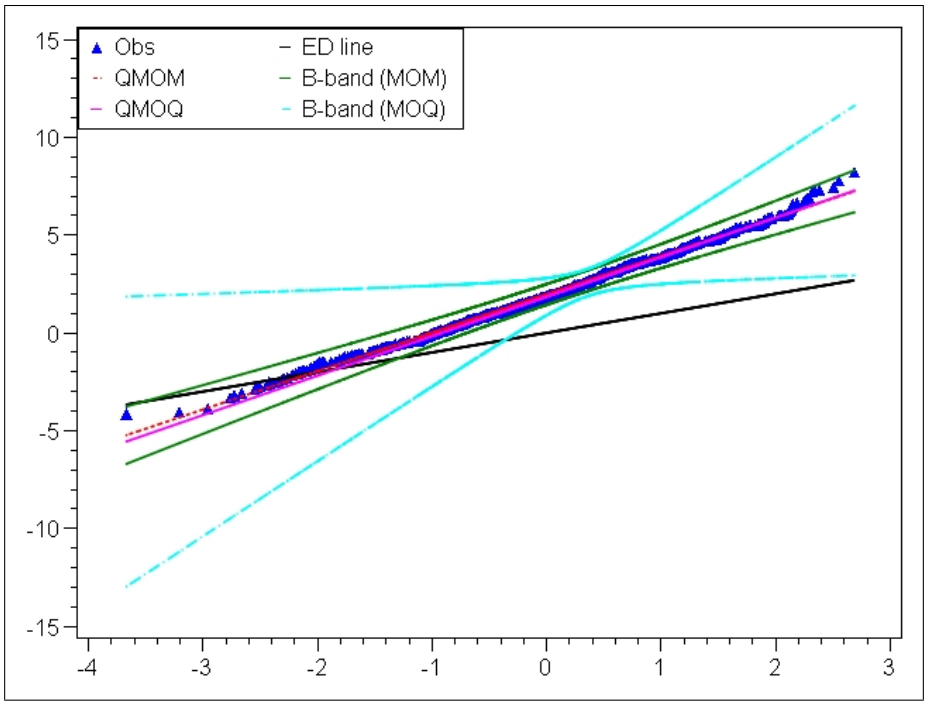


Figure 2.7: MOM and MOQ estimates (with B-bands) for Case 1

The two estimates, $\widehat{q}_{MOM}$ and $\widehat{q}_{MOQ}$, are similar and almost indistinguishable in Figures 2.7 and 2.8. As previously, the last two examples are based on the "credit scoring" datasets where the $DAINC$ (Case 3) and the $LOAN$ (Case 4) classifiers are used to distinguish between good and bad risk classes. The MOM estimate for q, the

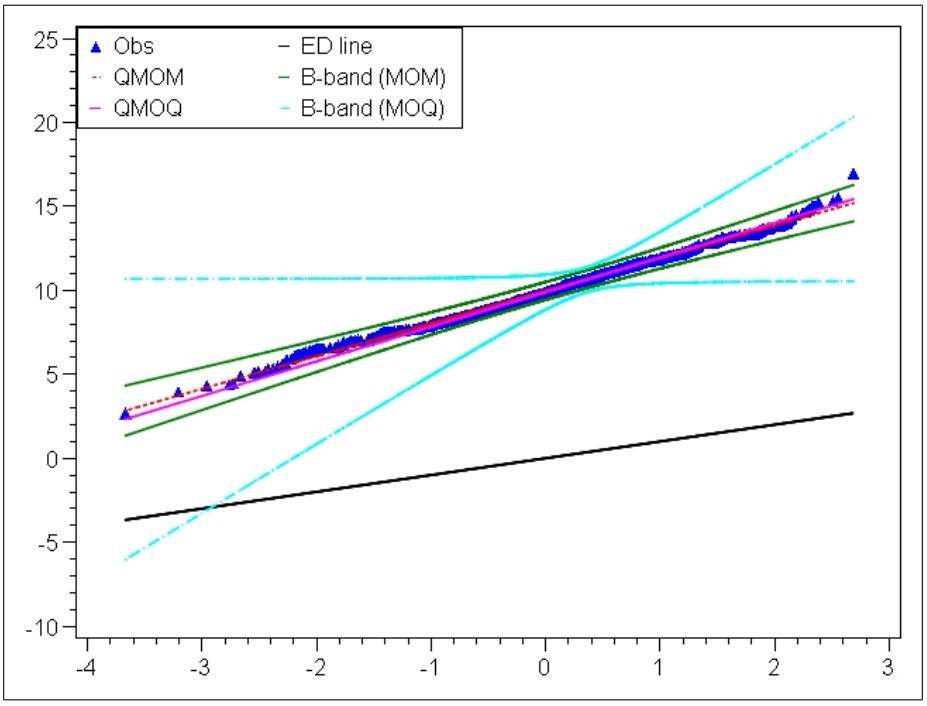


Figure 2.8: MOM and MOQ estimates (with B-bands) for Case 2

MOQ estimate for q, the associated 95% bootstrap confidence bands and the ED lines are plotted in Figures 2.9 and 2.10, respectively. Again the B-band of the method of quantiles estimate is wider than the B-band of the method of moments estimate and again the two estimates, $\widehat{q}_{MOM}$ and $\widehat{q}_{MOQ}$, are similar and almost indistinguishable. To conclude, in the four examples considered, the bootstrap confidence bands based on the MOQ estimator are wider than the bootstrap confidence bands based on the MOM estimator. Recall that a further Monte Carlo comparison will be made between the MOM and the MOQ estimator in Section 2.7.1 using bias and mean squared error.

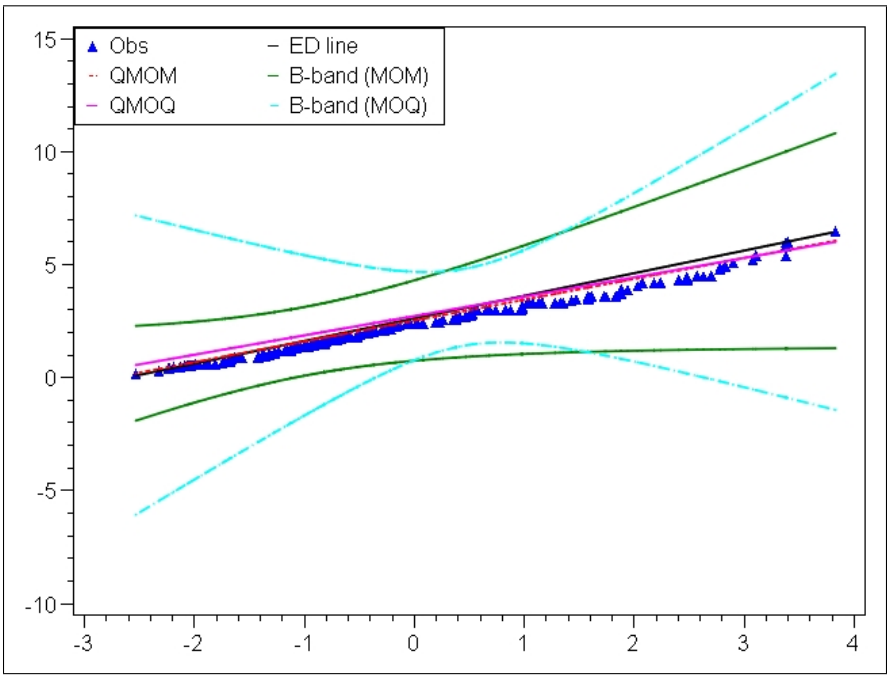


Figure 2.9: Method of moments and method of quantiles estimates (with B-bands) for $DAINC$, Case 3 (values in R'000)

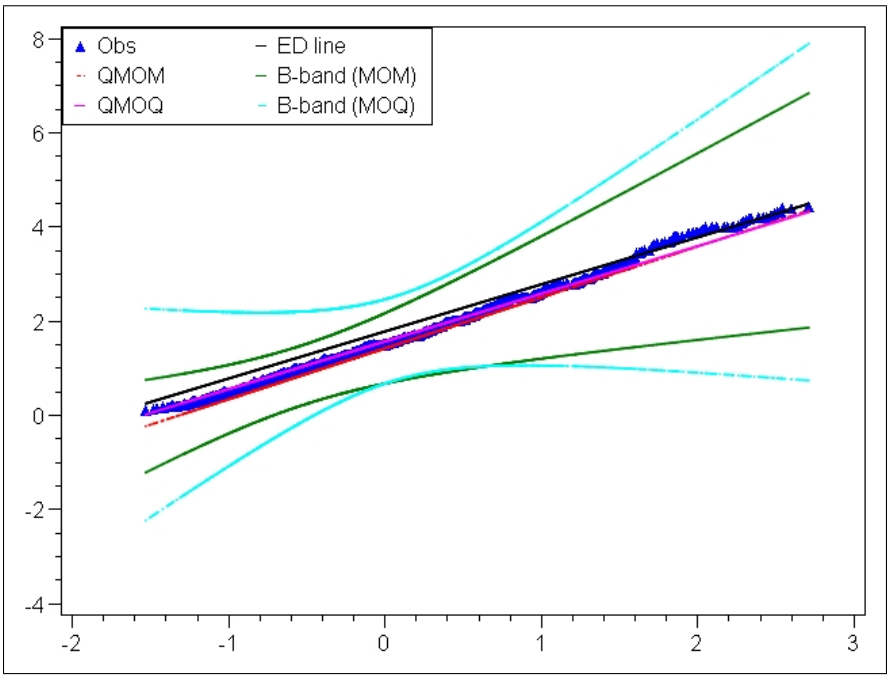


Figure 2.10: Method of moments and method of quantiles estimates (with B-bands) for $LOAN$, Case 4 (values in R'000)

2.6 Regression method estimator for q

An alternative estimator for q was proposed by Hsieh (1995). We refer to this as the regression method estimator, $\widehat{q}_{RM}$. The methodology described by Hsieh (1995) is a weighted least squares regression that estimates the linear form of q. Let $\boldsymbol{\theta} = [\alpha_0, \alpha_1]^\top$. The model that Hsieh assumes is:

$$G_n^{-1}(\mathbf{t}) = \overline{\overline{X}}_{k(n)}\boldsymbol{\theta} + \sigma D_{k(n)}\left[\frac{B_2(\mathbf{t})}{\sqrt{n}} - \frac{B_1(\mathbf{t})}{\sqrt{m}}\right] \tag{2.59}$$

with

$$\overline{\overline{X}}_{k(n)}^\top = \begin{bmatrix} 1 & ... & 1 \\ F_m^{-1}(t_1) & ... & F_m^{-1}(t_{k(n)}) \end{bmatrix} \tag{2.60}$$

where $B_1(\mathbf{t})$ and $B_2(\mathbf{t})$ are two independent Brownian bridges and $D_{k(n)}$ denotes the diagonal matrix of $1/f(F^{-1}(t_{k(n)}))$. The covariance matrix is given by

$$\boldsymbol{\Sigma} = cov\left[\frac{B_2(\mathbf{t})}{\sqrt{n}} - \frac{B_1(\mathbf{t})}{\sqrt{m}}\right] \tag{2.61}$$

where the ij^{th} element of $\boldsymbol{\Sigma}$ is given by

$$\boldsymbol{\Sigma}_{ij} = \frac{\left(\frac{1}{m} + \frac{1}{n}\right)\left(\frac{i}{m+1} - \frac{i \cdot j}{(m+1)^2}\right)}{\widehat{g}\left(F_n^{-1}\left(\frac{i}{m+1}\right)\right)\widehat{g}\left(F_n^{-1}\left(\frac{j}{m+1}\right)\right)} \tag{2.62}$$

for $i \leq j$. The regression estimates for $\widehat{\alpha}_0$ and $\widehat{\alpha}_1$ are obtained as the weighted least squares regression estimates

$$\widehat{\boldsymbol{\theta}} = [\widehat{\alpha_0}, \widehat{\alpha_1}]^\top = (X_*^\top X_*)^{-1} X_*^\top Y_* \tag{2.63}$$

where

$$Y_* = \left(D_{k(n)}\boldsymbol{\Sigma}^{1/2}\right)^{-1} G_n^{-1}(\mathbf{t}) \tag{2.64}$$

and

$$X_* = \left(D_{k(n)}\boldsymbol{\Sigma}^{1/2}\right)^{-1} \overline{\overline{X}}_{k(n)}. \tag{2.65}$$

For each m and n, an integer $k(n)$ and a vector $\mathbf{t} = (t_1, ..., t_{k(n)})^\top$, are chosen, where $0 < t_1 < ... < t_{k(n)} < 1$. The vector $\mathbf{t}$ refers to the percentiles corresponding to the data points used in fitting the model. Note that the asymptotic theory of Hsieh requires that m/n tends to a positive value when m and n increase. Therefore asymptotically $k(n, m)$ can be written as $k(n)$.

To summarise, the regression method of Hsieh (1995), is basically a weighted regression of the interpolated order statistics (at $t_1 < ... < t_{k(n)}$) of the Y-values calculated on the associated interpolated order statistics of the X-values.. Hsieh (1995) mentioned that the number of datapoints to be used, is critical. Yet, his article is unclear on how the datapoints should be selected.

2.6.1 Asymptotic distribution of $\widehat{\boldsymbol{\theta}} = [\widehat{\alpha}_0, \widehat{\alpha}_1]^\top$

The asymptotic distribution of $\widehat{\boldsymbol{\theta}} = [\widehat{\alpha}_0, \widehat{\alpha}_1]^\top$ is given by the expression

$$(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})\sqrt{\frac{mn}{m+n}} \cong N\left\{\mathbf{0}, \sigma^2 \left(\overline{\overline{X}}_{k(n)}^\top \boldsymbol{\Sigma}_{k(n)}^{-1} \overline{\overline{X}}_{k(n)}\right)^{-1}\right\}. \tag{2.66}$$

For details on the derivation, see Hsieh (1995). Note that confidence bands for the regression method estimator can be constructed by the usual method used when applying weighted least squares.

2.7 Empirical study

In this section the three estimators, $\widehat{q}_{MOM}$, $\widehat{q}_{MOQ}$, and $\widehat{q}_{RM}$ of the linear form of q are firstly compared by means of a Monte Carlo study and secondly their application is illustrated by the datasets considered in Chapter 1 (Section 1.2).

2.7.1 Monte Carlo study

The objective of this study is to determine the accuracy of the three estimators in estimating the parameters α_0 and α_1. We use the root mean squared error and the bias as measures of how well the method of moments, the method of quantiles and the regression method estimate the true parameters.

Our Monte Carlo study is designed as follows. We assume some distribution for F and G and then generate equal size samples ($m = n = 100, 200, 500$ and 1000) from those distributions and repeat the process L times ($L = 1000$). For each sample, we calculate $\widehat{\alpha}_{0,l}$ and $\widehat{\alpha}_{1,l}$ ($l = 1, ..., L$) for each of the three estimators. Define the estimated bias of $\widehat{\alpha}_{j,l}$ as

$$\frac{1}{L}\sum_{l=1}^{L}\left(\widehat{\alpha}_{j,l} - \alpha_j\right), \tag{2.67}$$

and the estimated root mean squared error (RMSE) of $\widehat{\alpha}_{j,l}$ as

$$\sqrt{\frac{1}{L}\sum_{l=1}^{L}\left(\widehat{\alpha}_{j,l} - \alpha_j\right)^2}, \tag{2.68}$$

for $j = 0, 1$. Without loss of generality, we take $\alpha_0 = 0$ and $\alpha_1 = 1$ which implies that $F = G$. For F and G we generate samples from four distributions: the standard normal, exponential, t distribution (with 5 degrees of freedom) and the standard Cauchy distribution. Note that the exponential distribution is defined by

$$\begin{aligned} F(x, \lambda) &= 1 - e^{-\lambda x} \text{ for } x \geq 0 \qquad (2.69)\\ F(x, \lambda) &= 0 \text{ otherwise} \end{aligned}$$

and the Cauchy distribution is defined by

$$F(x; x_0, \gamma) = \frac{1}{\pi}\arctan\left(\frac{x - x_0}{\gamma}\right) + \frac{1}{2}. \tag{2.70}$$

In this study we take $\lambda = 1$ (for the exponential distribution) and $x_0 = 0$ and $\gamma = 1$ (for the Cauchy distribution).

The formula for the cumulative distribution function of the t distribution is complicated and is not included here. It is given in the book of Evans et al., 2000.

The results of the Monte Carlo study are given in Table 2.9 (F and G are normally distributed), Table 2.10 (F and G are exponentially distributed), Table 2.11 (F and G are t distributed) and Table 2.12 (F and G are distributed according to the Cauchy distribution). Because the moments of the Cauchy distribution do not exist, we omitted the methods of moments from this Monte Carlo study.

Note that throughout the thesis we used kernel density estimation to estimate f. Specifically we used the bandwidth in (A.98) for the normally, exponentially and the t distributed data and the bandwidth in (A.104) for the Cauchy distributed data (see Calculation 3, Appendix A.3). Note also that to calculate the regression method estimates we need estimated values for f, g, F^{-1} and G^{-1} (refer to Calculations 2 and 3, Appendix A.3, for more detail). We used eight evenly spaced datapoints for the regression method. We will now discuss the four tables containing the results of the Monte Carlo study separately and make some concluding remarks at the end.

When considering the results in Table 2.9 it is clear that the method of moments and the regression method (using eight datapoints) are the best performers with respect to the RMSE and the bias criterion for both parameters and over all sample sizes considered. As far as RMSE is concerned the regression method (using eight datapoints) performed the best in most cases, while the method of moments performed the best in most cases when bias is considered.

The simulation results in Table 2.10 (exponentially distributed data) again shows that

Normal data	$m = n$	**RMSE**		**Bias**	
		α_0	α_1	α_0	α_1
Method of moments	100	0.1443	0.1059	0.0002	0.0118
	200	0.0991	0.0748	−0.0058	0.0017
	500	0.0627	0.0451	0.0028	0.0012
	1000	0.0459	0.0327	0.0013	0.0015
Method of quantiles	100	0.1828	0.1753	−0.0011	0.0177
	200	0.1239	0.1211	−0.0113	0.0155
	500	0.0806	0.0741	0.0018	0.0026
	1000	0.0559	0.0529	0.0010	−0.0004
Regression method (used all datapoints)	100	0.1920	0.1452	0.0021	0.0298
	200	0.1494	0.1024	−0.0020	0.0209
	500	0.0900	0.0705	0.0020	0.0224
	1000	0.0629	0.0519	0.0022	0.0197
Regression method (used 8 datapoints)	100	0.1392	0.1009	−0.0039	0.0092
	200	0.1004	0.0689	−0.0074	0.0049
	500	0.0601	0.0425	0.0032	0.0013
	1000	0.0432	0.0281	0.0015	−0.0002

Table 2.9: Bias and root mean squared error for the three estimators (normal data)

the regression method (using eight datapoints) is the best performing method in terms of the RMSE criterion and also does well in terms of the bias criterion. As far as bias is concerned the closest competitors are the method of quantiles and the method of moments, while in terms of RMSE, the method of moments, the method of quantiles and the regression method (using all datapoints) perform similarly and not too much worse than the regression method (using eight datapoints).

The simulation results in Table 2.11 (t distributed data) again shows the regression method (using eight datapoints) as the best performer in terms of RMSE. As far as bias is concerned all four methods perform similarly.

The results in Table 2.12 (Cauchy distributed data) show that the method of quantiles is the best performing method in terms of RMSE and bias over most of the samples sizes considered, with the regression method (using eight datapoints) a close contender especially in larger sample sizes. The regression method (using all datapoints) performs very poorly in all cases considered.

When considering all the results the regression method (using eight datapoints) seems to be the best performer while the regression method (using all the datapoints) seems to be the worst performer. The method of quantiles and the method of moments perform reasonably well when compared to the best performing method, with the method of quantiles having the advantage that it is a robust method which can be applied in extreme cases (like the Cauchy). However, as seen in Section 2.5.2, this method will yield slightly wider confidence bands than the method of moments.

An important question that we need to answer is what estimator should be used in practice on real datasets? Although the regression method (using eight datapoints) comes out as the clear winner in this simulation study it is unclear whether it will keep

Exponential data	$m = n$	**RMSE**		**Bias**	
		α_0	α_1	α_0	α_1
Method of moments	100	0.1376	0.2000	−0.0084	0.0162
	200	0.1009	0.1429	−0.0065	0.0062
	500	0.0635	0.0912	−0.0058	0.0045
	1000	0.0452	0.0633	−0.0003	0.0021
Method of quantiles	100	0.1624	0.2260	−0.0142	0.0265
	200	0.1045	0.1476	0.0003	0.0025
	500	0.0673	0.0923	−0.0029	−0.0010
	1000	0.0485	0.0669	0.0005	0.0018
Regression method (used all datapoints)	100	0.0339	0.2075	0.0281	0.0579
	200	0.0163	0.1577	0.0137	0.0532
	500	0.0064	0.1026	0.0056	0.0452
	1000	0.0032	0.0771	0.0028	0.0390
Regression method (used 8 datapoints)	100	0.0230	0.1511	0.0067	0.0152
	200	0.0120	0.1001	0.0025	0.0038
	500	0.0045	0.0604	0.0009	−0.0005
	1000	0.0027	0.0439	0.0013	0.0024

Table 2.10: Bias and root mean squared error for the three estimators (Exponential data)

t distribution (5)	$m = n$	**RMSE**		**Bias**	
		α_0	α_1	α_0	α_1
Method of moments	100	0.1852	0.1748	0.0081	0.0097
	200	0.1325	0.1255	−0.0031	0.0028
	500	0.0791	0.0796	0.0006	−0.0016
	1000	0.0576	0.0608	0.0018	0.0021
Method of quantiles	100	0.1930	0.1830	0.0062	0.0094
	200	0.1384	0.1305	0.0012	0.0066
	500	0.0804	0.0756	0.0006	−0.0057
	1000	0.0586	0.0552	0.0031	0.0011
Regression method (used all datapoints)	100	0.2789	0.2288	0.0054	0.0707
	200	0.1960	0.1729	−0.0027	0.0587
	500	0.1112	0.1019	−0.0001	0.0347
	1000	0.0793	0.0834	0.0021	0.0315
Regression method (used 8 datapoints)	100	0.1767	0.1229	0.0024	0.0076
	200	0.1259	0.0896	−0.0030	0.0038
	500	0.0730	0.0466	0.0022	−0.0041
	1000	0.0531	0.0343	0.0032	0.0029

Table 2.11: Bias and root mean squared error for the three estimators (t(5) data)

Cauchy data	$m = n$	**RMSE**		**Bias**	
		α_0	α_1	α_0	α_1
Method of quantiles	100	0.2210	0.2329	−0.0068	0.0246
	200	0.1561	0.1669	0.0006	0.0148
	500	0.0980	0.1008	0.0020	0.0042
	1000	0.0710	0.0720	−0.0004	−0.0009
Regression method (used all datapoints)	100	59.7400	18.7939	−1.8921	2.5151
	200	23.9294	55.0324	−1.3648	3.5878
	500	127.5930	22.8143	−4.6896	2.9019
	1000	93.7040	111.9280	−1.8882	6.5011
Regression method (used 8 datapoints)	100	0.2530	0.2660	−0.0083	0.0577
	200	0.1600	0.1753	0.0025	0.0312
	500	0.0969	0.0993	−0.0049	0.0118
	1000	0.0666	0.0662	0.0001	0.0061

Table 2.12: Bias and root mean squared error for the two estimators (Cauchy data)

up this good performance on real datasets. The choice of the number of datapoints to use in the construction of the estimator remains a practical problem. In his paper Hsieh (1995) showed that the regression method is asymptotically efficient in the semi-parametric location-scale model but also stated that the choice of the datapoints to be used in the weighted regression is problematic. In his paper he provides no guidelines on how the datapoints should be selected and also uses a rather arbitrary selection of eight datapoints in one of his examples.

A practical recommendation is to use the method of quantiles since it is a robust method, which can be applied in extreme cases (like the Cauchy). However, if initial plots of the data do not exhibit heavy tails or any outlying observations, the method of moments should be used, since this method has narrower confidence bands than the method of quantiles.

2.7.2 Examples

In order to study and compare the behaviour of the methods of moments, the method of quantiles and the regression method (using eight datapoints) further, we again consider the four examples discussed in Chapter 1, Section 1.2. The three estimators were computed for the normally distributed datasets (Case 1 and Case 2) and the resulting fits are depicted in Figure 2.11. In both cases the lines depicting the fits are almost indistinguishable. This is not surprising since the simulation results for the normally distributed data (Table 2.9) showed that the three methods performed rather similarly. Figure 2.12 contains the fits of the same three estimators to the credit scoring datasets (Case 3 and Case 4). Again the fitted lines of the method of moments and the method of quantiles are closely similar, while the fitted line of the regression

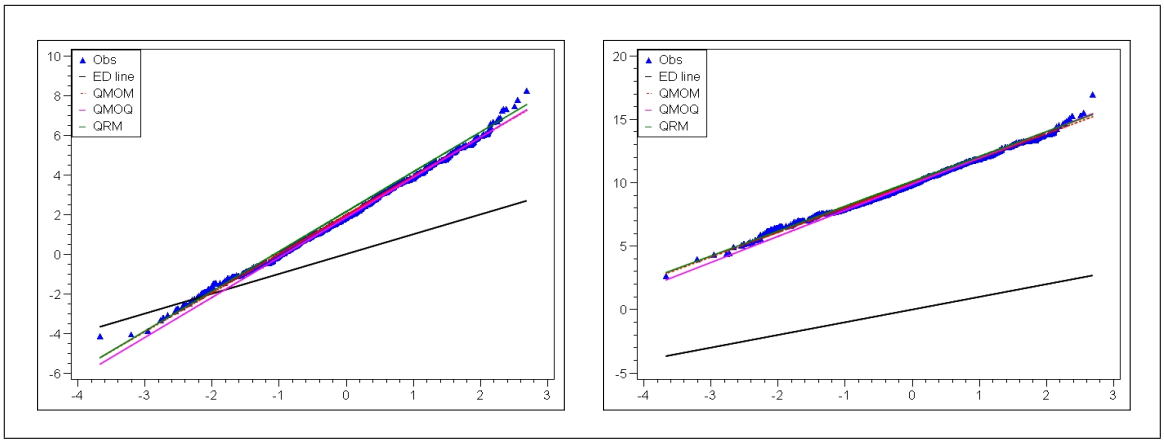


Figure 2.11: Method of moments estimator, method of quantiles estimator and regression method estimator for Case 1 (left) and Case 2 (right)

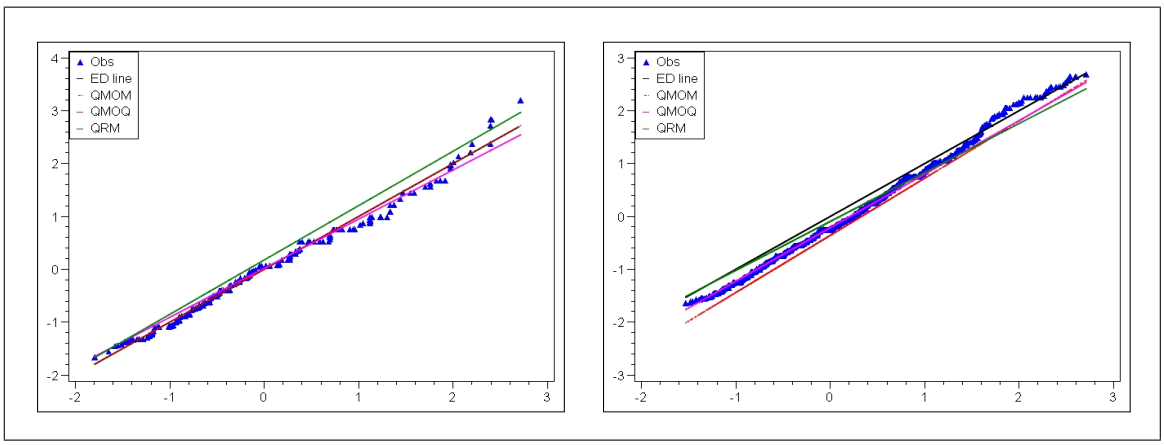


Figure 2.12: Method of moments estimator, method of quantiles estimator and regression method estimator for $DAINC$, Case 3 (left) and $LOAN$, Case 4 (right), values in R‘000

method is slightly removed from the latter fits as well as from the bulk of the data.

Based on the resulting fits in the last four examples, we would recommend that the method of quantiles or the method of moments be used, since these two fitted lines are similar. However, the method of moments has the advantage that it has narrower confidence bands than the method of quantiles, and is therefore the better option in these four examples considered. We will discuss this in more detail when a practical credit scoring dataset is analysed in Chapter 4.

2.8 Tests of linearity

In the previous sections, we assumed that q is linear. In this section we propose statistics for testing this assumption. The test for linearity in a credit scoring context could also be used to test whether the two distributions, the goods and the bads, are from the same translation-scale family.

The null hypothesis is

$$H_0 : q(v) = \alpha_0 + \alpha_1 v. \tag{2.71}$$

Recall that $q(v) = G^{-1}(F(v))$. One could formulate the null hypothesis in various alternative ways. For example, the hypothesis that F and G belong to the same translation-scale family, is equivalent to testing whether q is linear. The null hypothesis is then

$$H_0 : F_0 = G_0 \tag{2.72}$$

and the alternative hypothesis is

$$H_a : F_0 \neq G_0 \tag{2.73}$$

where

$$F_0(x) = F\left(\frac{x-\mu_V}{\sigma_V}\right) \tag{2.74}$$

and

$$G_0(x) = G\left(\frac{x-\mu_W}{\sigma_W}\right). \tag{2.75}$$

A possible test statistic is given by the following expression

$$\frac{1}{\sqrt{m+n}}\sum_{i=1}^{n}\left(\widetilde{V}_{(im/n)} - \widetilde{W}_{(i)}\right)^2 \tag{2.76}$$

where $\widetilde{V}_{(i)}$ ($\widetilde{W}_{(i)}$) indicates the i^{th} ordered value of $\widetilde{V}$ ($\widetilde{W}$) and where $\widetilde{V}$ and $\widetilde{W}$ indicate the standardised versions of V and W, i.e.

$$\widetilde{V}_i = \frac{V_i - \overline{V}}{s_V} \text{ for } i = 1, ..., m \tag{2.77}$$

$$\widetilde{W}_j = \frac{W_j - \overline{W}}{s_W} \text{ for } j = 1, ..., n. \tag{2.78}$$

Note that throughout this thesis we assumed that $m \geq n$.

Another possible test statistic is the two-sample Kolmogorov-Smirnov test statistic applied to the standardised values $\widetilde{V}_i$ and $\widetilde{W}_j$, i.e.

$$K_1 = \sqrt{m+n}\sup_y\left|\widetilde{F}_m(y) - \widetilde{G}_n(y)\right| \tag{2.79}$$

where $\widetilde{F}_m$ and $\widetilde{G}_n$ denote the EDF's (empirical distribution functions) of the $\widetilde{V}$ and $\widetilde{W}$ values.

The previous test statistics may be used to test the null hypothesis of linearity against the broad alternative that the relationship is not linear. To create a more powerful test, we would need to restrict this broad alternative. One could, for instance, replace the alternative hypothesis by a special form such as that the relationship is a polynomial. We do not develop such tests here and leave the details of their construction for further

research. The null hypothesis remains the same, i.e.

$$H_0 : q(v) = \alpha_0 + \alpha_1 v \tag{2.80}$$

but the alternative hypothesis changes to

$$H_a : q(v) = h(v) \tag{2.81}$$

where $h(v)$ is a polynomial of degree > 1. Then another potentially useful test statistic is the well known t test that is used in regression to test whether a regression coefficient is zero (Neter et al., 1985, e.g. p. 51 or p. 118-121). In our case, we use the weighted regression method of Hsieh (1995), where the regression model is

$$Y_* = X_* \boldsymbol{\theta} + \varepsilon \tag{2.82}$$

This weighted regression is shown in Section 2.6, in (2.59). Y_* is given by (2.64), X_* is now similar to (2.65) just adding a quadratic term, $P_2(\cdot)$, i.e.

$$X_* = \boldsymbol{\Sigma}^{-1/2} \begin{bmatrix} 1 & F_m^{-1}\left(\frac{1}{n+1}\right) & P_2(F_m^{-1}\left(\frac{1}{n+1}\right)) \\ \vdots & \vdots & \vdots \\ 1 & F_m^{-1}\left(\frac{n}{n+1}\right) & P_2(F_m^{-1}\left(\frac{n}{n+1}\right)) \end{bmatrix} \tag{2.83}$$

where $\boldsymbol{\Sigma}$ is given in (2.61) and now $\boldsymbol{\theta} = [\alpha_0, \alpha_1, \alpha_2]^\top$. Furthermore, ε is a vector of i.i.d. (independent and identically distributed) random variables. The test statistic is then

$$K_2 = \frac{|\widehat{\alpha}_2|}{s(\widehat{\alpha}_2)} = \frac{|\widehat{\alpha}_2|}{\sqrt{MSE\left((X_*^\top X_*)^{-1}\right)_{[3,3]}}} \tag{2.84}$$

where $(X_*^\top X_*)^{-1}_{[3,3]}$ denotes the $[3,3]$ element of $(X_*^\top X_*)^{-1}$.

In the remainder of this section, the focus will be on the test statistics, K_1 (2.79) and K_2 (2.84). The other test statistic (2.76) was also investigated by means of a Monte Carlo study, but the results obtained were unsatisfactory. Some comments on this will be made at the end of Section 2.8.1.

2.8.1 Monte Carlo study

In this section we use a Monte Carlo study to investigate the finite sample significance levels and power attained by K_1 and K_2.

Our Monte Carlo study is designed as follows. Assume some common distribution type for F and G and then generate equal-sized samples ($m = n = 20,\ 50,\ 100$ and 500) from each of those distributions and repeat the process $L = 1000$ times. Let $K_{s,l}$ denote test statistic K_s calculated in the l^{th} Monte Carlo sample, $s = 1, 2$; $l = 1, 2, ..L$. The bootstrap estimate of the critical value for each test statistic is denoted by $K^*_{\alpha,s,l}$. The estimated bootstrap significance level is given by

$$(1/L)\sum_{l=1}^{L} I\left(K_{s,l} > K^*_{\alpha,s,l}\right). \tag{2.85}$$

When F and G above are not of the same type, (2.85) gives the estimated power of the tests.

In summary, to compare the estimated significance levels with the nominal significance levels, we assume the same distribution type for F and G. Note that we can assume without loss of generality that $\mu_V = \mu_W$ and $\sigma_V = \sigma_W$ as both the test statistics are based on the standardised values of V and W. To estimate the power of the tests of linearity, we assume some distribution type for F and another distribution type for G.

Significance levels Firstly, we assume that F and G are the standard normal distribution functions. The estimates of the bootstrap significance levels from this Monte Carlo study are given in Table 2.13. The results in Table 2.13 were obtained by Algorithm 3 (Appendix A.2) with $B = 1000$ and $L = 1000$. In all cases, the attained significance levels are very close to the nominal value of 5%. However, the estimated

significance levels for K_1 is constantly lower than the nominal value of 5% and the opposite is true for K_2.

	Estimated bootstrap significance levels		
$m = n$	K_1	K_2 (using all datapoints)	K_2 (using eight datapoints)
20	4.90%	5.80%	5.10%
50	4.10%	5.70%	5.00%
100	4.60%	5.40%	5.10%
500	4.70%	5.20%	5.20%

Table 2.13: Estimates of the significance levels (nominal significance level of 0.05), standard normal distribution

Note that for K_2, which is based on the method of Hsieh (1995), the choice of which datapoints to use for the weighted regression is problematic, as mentioned in Section 2.7.1. As in the previous section, we use all the datapoints as well as eight evenly spaced datapoints. The calculation of K_1 is shown in Appendix A.3, Calculation 4.

Secondly, we assume that F and G are the CDF (cumulative distribution function) of the logarithm ($\ln$) of an exponential random variable, namely

$$F(x) = 1 - e^{-e^x}, \ -\infty < x < \infty.$$

The natural logarithm of an exponential random variable is a standard Gumbel random variable. The estimates of the bootstrap significance levels are given in Table 2.14. For K_1 and K_2 the attained significance levels are very close to or slightly lower than the nominal value of 5%.

	Estimated bootstrap significance levels		
$m = n$	K_1	K_2 (using all datapoints)	K_2 (using eight datapoints)
20	4.10%	4.40%	4.60%
50	4.90%	5.10%	3.80%
100	4.50%	4.80%	5.10%
500	5.10%	5.10%	5.20%

Table 2.14: Estimates of the significance levels (nominal significance level of 0.05), standard Gumbel distribution

Thirdly, we assume that F and G are standard Cauchy distribution functions. Since the Cauchy distribution does not have a mean and a standard deviation, we cannot standardise by sample mean and sample standard deviation. Rather we standardise by subtracting the sample median and dividing by the sample interquartile range. The estimates of the bootstrap significance levels are given in Table 2.15. The attained significance levels of K_2 are close to the nominal value of 5%, especially in larger sample sizes ($m = n > 20$). The significance levels of K_1 are lower than the nominal value of 5%.

Fourthly, we assume that F and G are both mixtures of the standard normal CDF and the CDF of the standard Gumbel random variable (Mixture 1). The estimates of the bootstrap significance levels are given in Table 2.16. The attained significance levels of K_2 (using eight datapoints) are close to the nominal value of 5%. The significance levels of K_1 are higher than the nominal value of 5% and for K_2 (using all datapoints)

	Estimated bootstrap significance levels		
$m = n$	K_1	K_2 (using all datapoints)	K_2 (using eight datapoints)
20	2.70%	6.20%	3.40%
50	2.90%	5.80%	5.50%
100	3.30%	5.40%	5.30%
500	4.30%	5.00%	5.10%

Table 2.15: Estimates of the significance levels (nominal significance level of 0.05), Cauchy distribution

lower than the nominal value of 5%.

Lastly, we assume that F and G are both mixtures of the standard normal CDF and the standard Cauchy CDF (Mixture 2). The estimates of the bootstrap significance levels are given in Table 2.17. The attained significance levels of K_2 (using eight datapoints) are close to the nominal value of 5%. Again the attained significance levels of K_1 are higher than the nominal value of 5% and for K_2 (using all datapoints) lower than the nominal value of 5%.

To summarise, the attained significance levels of K_2 (using eight datapoints) are close to the nominal value of 5% over all the distributions considered. The attained significance levels of K_2 (using all datapoints) are slightly lower than the nominal value of 5% over all the distributions considered and the significance levels of K_1 are slightly lower for the first three distributions considered, and slightly higher for the last two mixture distributions considered. We now investigate the power of these tests.

	Estimated bootstrap significance levels		
$m = n$	K_1	K_2 (using all datapoints)	K_2 (using eight datapoints)
20	6.20%	3.90%	5.10%
50	7.00%	4.30%	4.80%
100	6.30%	4.20%	4.70%
500	6.10%	4.80%	5.00%

Table 2.16: Estimates of the significance levels (nominal significance level of 0.05),

Mixture 1

	Estimated bootstrap significance levels		
$m = n$	K_1	K_2 (using all datapoints)	K_2 (using eight datapoints)
20	6.40%	3.80%	4.00%
50	5.70%	4.20%	4.50%
100	6.90%	4.00%	4.50%
500	6.00%	4.40%	4.90%

Table 2.17: Estimates of the significance levels (nominal significance level of 0.05),

Mixture 2

Power Note that in the previous section, where the significance levels are calculated, we assume the same distribution type for F and G, so it does not matter whether bootstrap samples are taken from F or from G. In this section, where the power is calculated, F and G do not have the same distribution type, therefore one is not sure whether to sample from F or G to determine the bootstrap critical value. To investigate whether to sample from F or G, a sample will firstly be taken from F, secondly from G and thirdly from a mixture of F and G, denoted by H. These answers will then be compared.

Note that, because we used bootstrap critical values, it is difficult to ensure that the type I error (significance level) is exactly the same when comparing the power of the tests. This fact should be taken into account when interpreting the results that follow, however, the estimated bootstrap significance levels were close to the nominal value of 5% (see previous section).

Firstly, we assume the standard normal distribution function for F and the CDF of the standard Gumbel random variable for G (Mixture 1). The estimates of the power of K_1 and K_2 are given in Figures 2.13, 2.14 and 2.15.

For K_1 and K_2 the estimated power increases with the sample size, as expected. The power is higher for K_1 than K_2 in large samples. This might be due to the choice of which datapoints to use for the weighted regression, as mentioned previously. The power for K_2 (using all datapoints) is higher than K_2 (using eight datapoints) for all the sample sizes considered. The results are very similar whether the samples were taken from F, G or H to calculate the bootstrap critical values. This was also confirmed with two way analysis of variance tests. From these three graphs it is clear that sample size has a considerable effect on the power of the test statistics. Whether samples

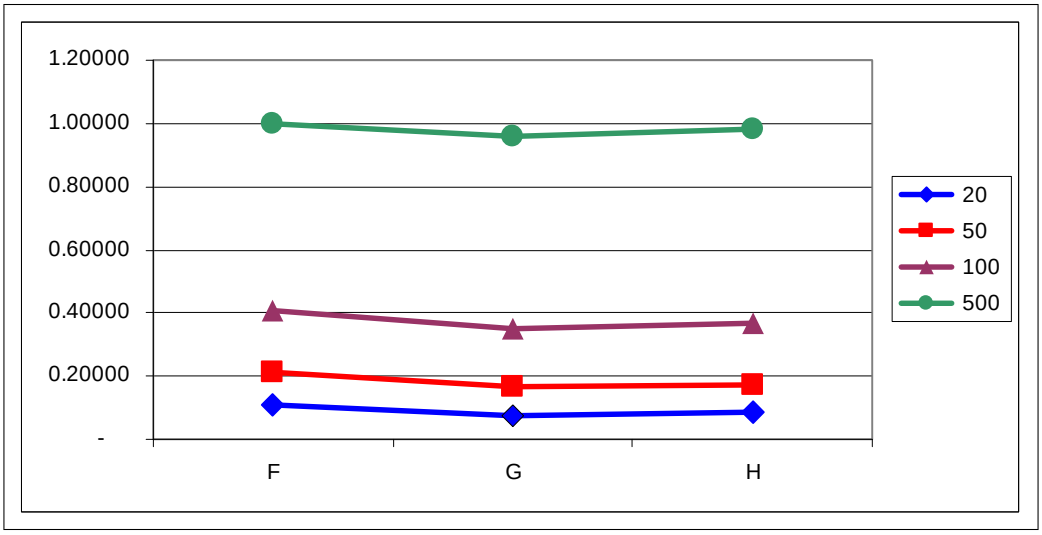


Figure 2.13: Estimates of the power of K_1 (Mixture 1)

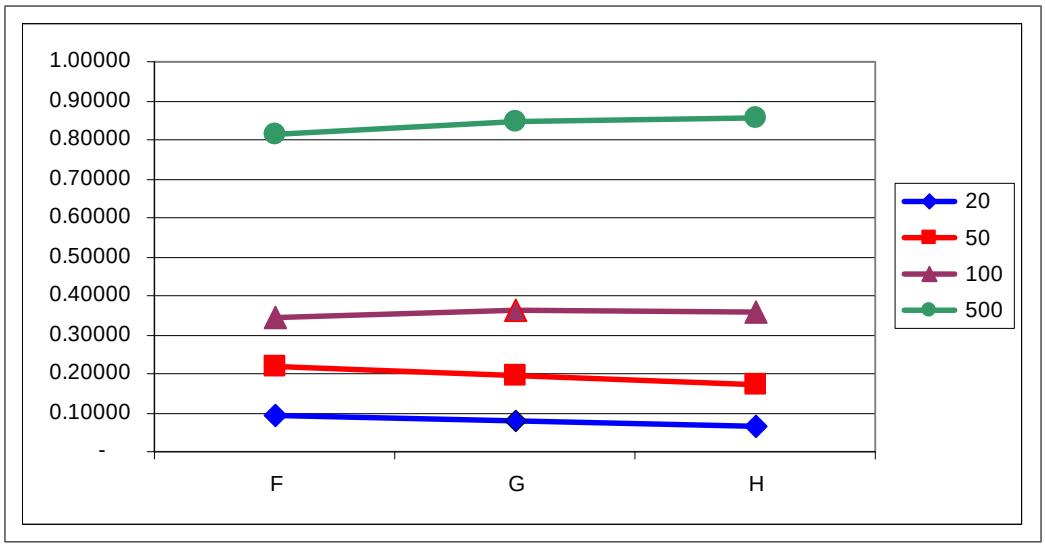


Figure 2.14: Estimates of the power of K_2 (using all datapoints), Mixture 1

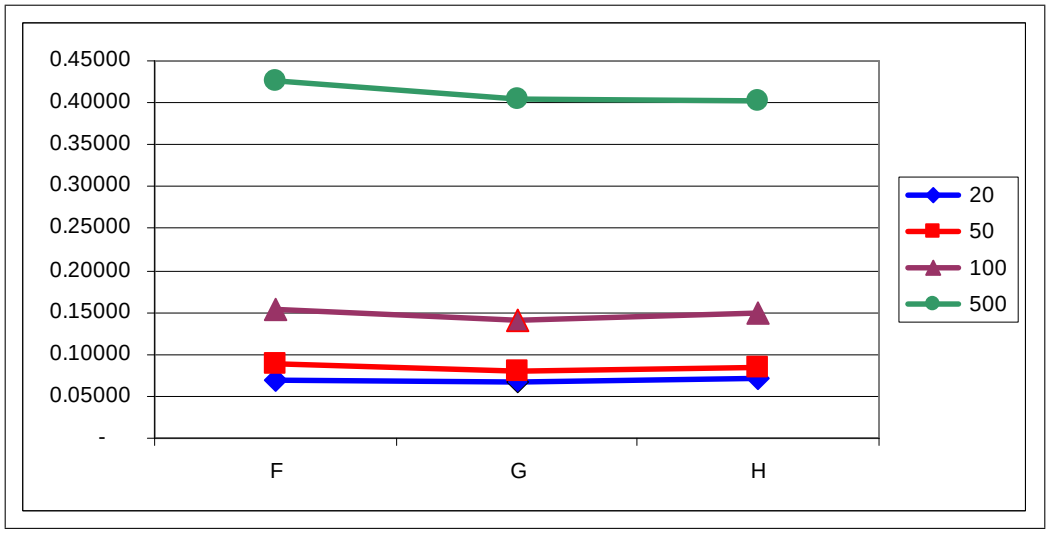


Figure 2.15: Estimates of the power of K_2 (using eight datapoints), Mixture 1

were taken from F, G or H to calculate the bootstrap critical values, have little effect on the power of the test statistics.

Secondly, we assume the standard normal distribution function for F and the Cauchy distribution function for G (Mixture 2). The estimates of the power of these tests are given in Figures 2.16, 2.17 and 2.18.

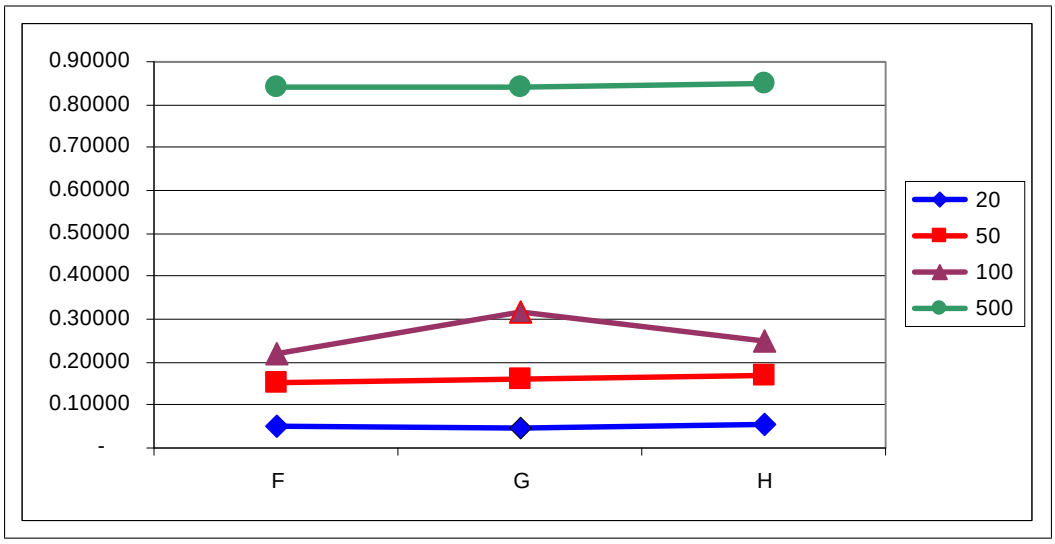


Figure 2.16: Estimates of the power of K_1 (Mixture 2)

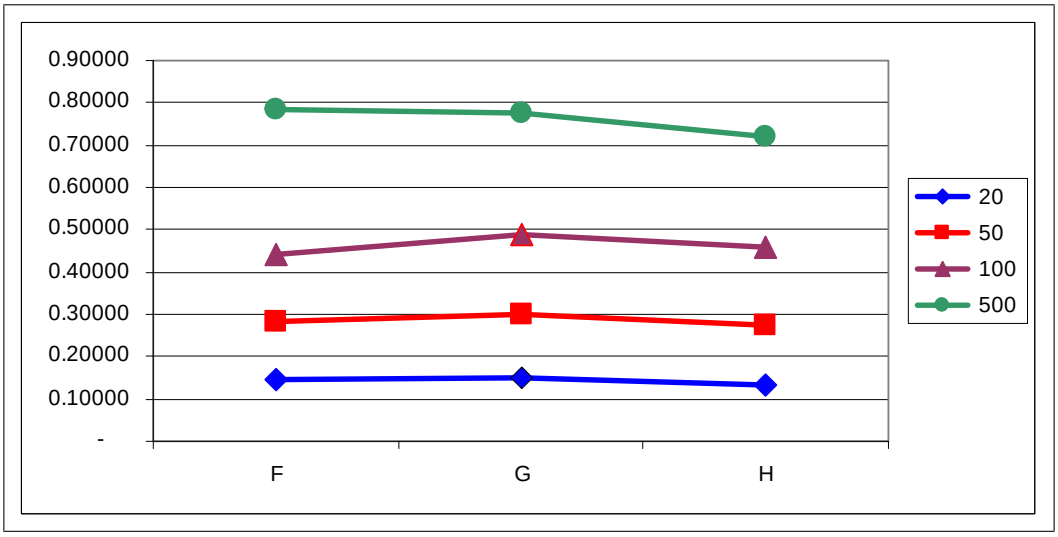


Figure 2.17: Estimates of the power of K_2 (using all datapoints), Mixture 2

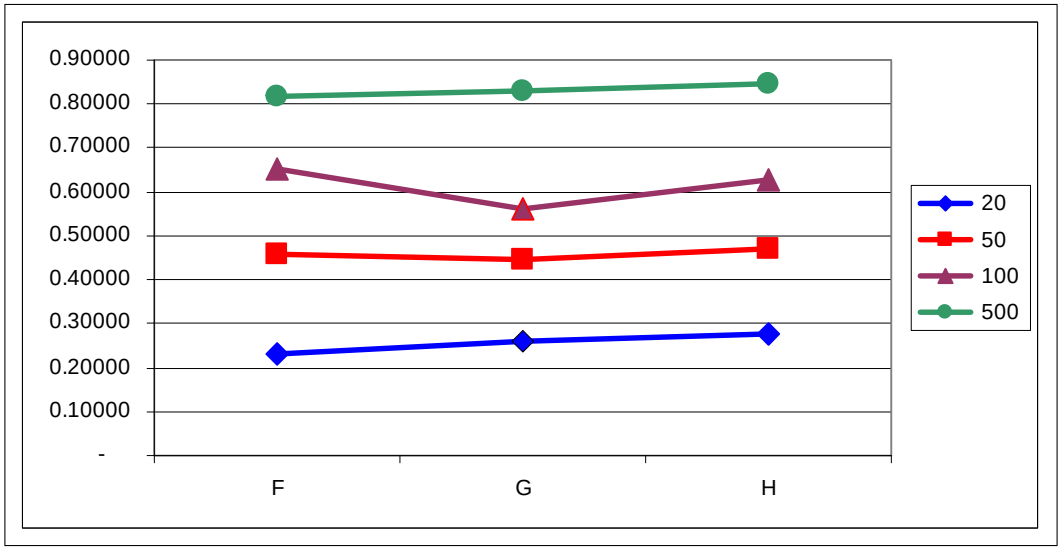


Figure 2.18: Estimates of the power of K_2 (using eight datapoints), Mixture 2

Again the estimated power increases with the sample size, as it should. The power is very high, especially in large samples. The power of K_2 (using eight datapoints) is slightly higher than the power of K_2 (using all the datapoints). Again, the results are very similar whether the samples were taken from F, G or H to calculate the bootstrap critical values. This was again confirmed with two way analysis of variance tests. From these three graphs it is again clear that sample size has a considerable effect on the power of these tests. The difference between the power of K_1 and K_2 when samples were taken from F, G or H to calculate the bootstrap critical values, is as in Mixture 1, very small.

Taking into account the results of the attained significance levels as well as the power, it is recommended to use K_1 (2.79) to test the null hypothesis that q is linear. K_2 (2.84) is not recommended due to the problematic choice of the number of datapoints to use. Another reason for not using the test K_2 is that it is testing against a very restricted alternative.

Remarks

1. One of the test statistics mentioned, (2.76), has not been considered in this section. The reason is that preliminary simulation studies showed poor estimated significance levels for this test statistic. Potgieter (2006) explained that in its current unstandardised form, the statistic does not have a proper limiting distribution.

2. Note that the value of the regression test statistic, K_2, depends on whether W is regressed on V or V is regressed on W. In all other test statistics V and W can be switched around without changing the value of the test statistic.

3. Note that as an alternative to the regression test statistic, K_2, an unweighted regression could also be used. Preliminary studies showed that the weighted regression gives better results than the unweighted regression and therefore only the weighted regression was studied.

2.9 Summary and conclusion

In Chapter 2 our aim was to find methods for estimating q and to construct confidence bands for q.

As a departure point we provided an overview of a non-parametric estimator for the general form of q proposed by Lehmann (1974) and associated confidence bands derived by Doksum and Sievers (1976).

Then we introduced the method of moments estimator for estimating the linear form of q and we derived the asymptotic distribution of the estimator and constructed $100(1-\alpha)\%$ confidence bands for q based on the asymptotic results. We compared the non-parametric estimator with the method of moments estimator by means of a Monte Carlo study and illustrate its application in a number of datasets. As one would expect, the method of moments estimator, because of its semi-parametric nature, yields narrower confidence bands in the tails of the distribution than did the non-parametric estimator.

The method of moments estimator has certain deficiencies, for example the asymptotic covariance matrix of the method of moments estimator is based on the third and fourth moments. It is well known that higher moments (3^{rd} and 4^{th}) have a tendency to become unstable (see e.g. van der Vaart, 1998 or Lehmann, 1999). A more robust alternative might be to replace the mean in the method of moments estimator

with the median and the standard deviation with the interquartile range, which will have a more stable covariance matrix. This method, referred to as the method of quantiles estimator was defined and the asymptotic distribution derived as well as the confidence bands. We compared the method of moments and the method of quantiles estimator (and associated confidence bands) by means of a Monte Carlo study and illustrate their applications in a number of datasets. In both cases (method of moments and method of quantiles) the bootstrap confidence band was preferred above the asymptotic confidence band as the required coverage probability was obtained.

Another alternative to estimate the linear form of q, is the regression method proposed by Hsieh (1995). We compared the three estimators (method of moments, method of quantiles and regression method) by means of a Monte Carlo study where we measured their performance using mean squared error. All three methods performed reasonably well over the distributions considered. However, the choice of the number of datapoints to use in the regression method presented a problem. The method of quantiles was the more robust choice, but gave slightly larger confidence bands.

To conclude this chapter we constructed statistical hypothesis tests of the linearity assumption on q and did some Monte Carlo investigations of these tests. Taking into account the results of the attained significance levels as well as the power, some recommendations were made on which test statistic to use.

CHAPTER 3

Detecting outliers using weights in logistic regression

As we have mentioned in Chapter 1, Section 1.3, logistic regression (LR) is frequently used in the development of credit scoring models and is concerned with estimating the probability of an event occuring. Pregibon (1981) states that the estimated LR relationship may be severely affected by outliers; this motivates the need for robust logistic regression procedures. Studies in this direction have been reported by Pregibon (1981), Copas (1988), Rousseeuw and Christmann (2003), Huber (1973), Rousseeuw and Leroy (1987) and Yohai (1987). Trimming is a broad approach towards robustifying statistical procedures. It allows one to identify outliers and remove them from the data used in the estimation process. Trimming has been developed extensively by a number of authors in least squares regression, multivariate analysis and other fields (see e.g. Rousseeuw, 1984, Rousseeuw and Van Driessen ,1999a,b, where further references can be found). At first thought it seems attractive to use trimming also in LR to identify outliers and to limit their effects. When trimming, a subset of the data that is highly likely to be free from outliers is needed and a method is required to select such a subset. One possibility is to use ML considerations, but this approach tends to run into the separation problem. As we mentioned in Section 1.3, the maximum likelihood estimator (MLE) does not exist if there is separation in the data. The problem is that those observations that are considered as outliers are usually the same observations that will provide some overlap in the data. Therefore, as pointed out by Christmann and Rousseeuw (2001), trimming these observations removes the overlap and may lead to non-existence (indeterminacy) of the ML estimator (MLE) applied to the remaining data. These authors produced methodology to measure this overlap, enabling the user to judge the closeness to indeterminacy. In a further contribution Rousseeuw and Christmann (2003) overcame the non-existence problem by introdu-

cing the hidden logistic regression model with an associated estimator referred to as the maximum estimated likelihood (MEL) estimator which always exists even when there is no overlap in the data. They also proposed a robustified form of the MEL estimator, called the weighted maximum estimated likelihood (WEMEL) estimator. The WEMEL estimator does not trim, but downweights leverage points, where the choice of leverage points is based on the robust distances in the regressor space. Using a simulation study they show that the WEMEL estimator performs very well as a robust procedure compared to its competitors. However, the WEMEL estimator does not take outliers in the response direction into account and is not really an outlier detection procedure because it does not produce a subset of the observations that may be labelled as outliers. In this chapter we use a different form of downweighting to introduce a procedure that may be thought of as both a robust LR estimation procedure and an outlier detection method.

Our procedure may be described as a method that "Detects Outliers Using Weights" and is referred to below as the DOUW method. The DOUW method begins by selecting two sets of weights, namely high and low weights and then splits the data optimally into two subsets to which the high and the low weights are attached, the subset with the high weights including the observations that are more likely not to be outliers. A corresponding weighted ML estimator of the regression coefficients is computed. This is used to estimate the response probabilities of the individual observations. Observations with success response ($= 1$) but low estimated success response probability and observations with failure response ($= 0$) but high estimated success response probability are then classified as outliers. A final weighted MLE can then be computed by redistributing the high and low weights according to the

degree of "outlyingness" of the observations in question. The method depends on the specification of some quantities, such as the size of the initial high weight subset and the levels of the high and low weights as well as the threshold according to which outlyingness is decided. We study the effects of the choices of these items below and also compare the DOUW estimator with the ML, MEL and WEMEL estimators.

The layout of the present chapter is as follows. Section 3.1 introduces the notation and terminology used here and reviews briefly some notions regarding outliers relevant to LR. Section 3.2 formulates the basic DOUW procedure and lists a number of more elaborate versions that can also be used. Section 3.3 reports the results of a simulation study that evaluates the cost/benefit balance that has to be taken into account when specifying the tuning parameters of the procedure. Section 3.4 discusses the application of the DOUW procedure to a number of standard datasets in the literature as well as a new large dataset relating to success probabilities in sales promotion campaigns. Section 3.5 gives a summary of, and main conclusions drawn from, the work in this chapter. Technical details are provided in Appendix B.1.

3.1 Notation and terminology

As mentioned in Chapter 1, Section 1.3, in a linear logistic regression (LR) setup we have a dichotomous response variable Y that can take the values 1 (bad risk class) or 0 (good risk class), and we have K regressors $x_1, ..., x_K$. Let $\mathbf{x}^\top = (1, x_1, ..., x_K)$ with $\top$ denoting transpose. We fit the LR model

$$P(Y = 1) = p(\mathbf{x}, \boldsymbol{\beta}) = 1/\left(1 + \exp\left(-\boldsymbol{\beta}^\top \mathbf{x}\right)\right) \tag{3.1}$$

where $\boldsymbol{\beta}^\top = (\beta_0, \beta_1, ..., \beta_K)$ is the vector of LR coefficients (see e.g. Hosmer and Lemeshow, 1989 and Kleinbaum, 1994). $l(\mathbf{x}) = \boldsymbol{\beta}^\top \mathbf{x}$ is often referred to as the logit value of $\mathbf{x}$ and $p(\mathbf{x}, \boldsymbol{\beta})$ considered as a function of $l(\mathbf{x})$ as the probability function or curve of the model we are fitting to the data. In the credit scoring context we refer to $p(\mathbf{x}, \boldsymbol{\beta})$ as the default probability curve, but in the general context, we see $Y = 1$ as a success and therefore refer to $p(\mathbf{x}, \boldsymbol{\beta})$ as the success probability curve.

Assume that we have N observations, where the n^{th} observation is $(y_n, \mathbf{x}_n^\top)$, with y_n the observed value of Y and $\mathbf{x}_n^\top$ the vector of observed values of the K regressors. Under the independence assumption the log likelihood of the N observations is given by

$$\sum_{n=1}^{N} D_n(\boldsymbol{\beta}) \text{ with } D_n(\boldsymbol{\beta}) = y_n \log p(\mathbf{x}_n, \boldsymbol{\beta}) + (1 - y_n) \log(1 - p(\mathbf{x}_n, \boldsymbol{\beta})) \tag{3.2}$$

and the MLE of $\boldsymbol{\beta}$ is obtained by maximising this expression over $\boldsymbol{\beta}$.

It will become clear later on that a link exists between outliers and separation. To understand this link, separation will first be discussed. For datasets in which there is no overlap between the 0 and 1 responses (i.e. the $\mathbf{x}_n$'s corresponding to $y_n = 0$ can be separated by a hyperplane from the $\mathbf{x}_n$'s corresponding to $y_n = 1$), the MLE does not exist since the likelihood function achieves its maximum when some of the components of $\boldsymbol{\beta}$ are +∞ or -∞. This can be argued as follows. If there is separation, there exists a choice for $\boldsymbol{\beta}$, say $\overline{\beta}_0, \overline{\beta}_1, ..., \overline{\beta}_K$ so that

$$\begin{aligned} \overline{\beta}_0 + \overline{\beta}_1 x_{n,1} + ... + \overline{\beta}_K x_{n,K} &> 0 \text{ for all cases where } y_n = 1 \\ \overline{\beta}_0 + \overline{\beta}_1 x_{n,1} + ... + \overline{\beta}_K x_{n,K} &< 0 \text{ for all cases where } y_n = 0. \end{aligned} \tag{3.3}$$

Therefore, the choice, $c\overline{\beta}_0, c\overline{\beta}_1, ..., c\overline{\beta}_K$ with c a positive constant leads to the log likeli-

hood which is equal to

$$\sum_{\{y_n=1\}} \log\left[\frac{1}{1+e^{-c\left(\overline{\beta}_0+\overline{\beta}_1 x_{n,1}+\ldots+\overline{\beta}_K x_{n,K}\right)}}\right]+\sum_{\{y_n=0\}} \log\left[1-\frac{1}{1+e^{-c\left(\overline{\beta}_0+\overline{\beta}_1 x_{n,1}+\ldots+\overline{\beta}_K x_{n,K}\right)}}\right]. \tag{3.4}$$

Since

$$\overline{\beta}_0+\overline{\beta}_1 x_{n,1}+\ldots+\overline{\beta}_K x_{n,K}>0$$

for all cases where $y_n=1$, each term in the first sum tends to $log(1)=0$ as $c\to\infty$. Again, since

$$\overline{\beta}_0+\overline{\beta}_1 x_{n,1}+\ldots+\overline{\beta}_K x_{n,K}<0$$

for all cases where $y_n=0$, each term in the second sum also tends to $log(1)=0$ as $c\to\infty$. Therefore the log likelihood then tends to 0 as $c\to\infty$.

Thus the log likelihood which is strictly negative for all finite $\boldsymbol{\beta}$ can be made 0 by letting $\boldsymbol{\beta}\to\infty$ in the above explained fashion and in this sense the MLE does not exist. In practice, using PROC NLP of SAS (SAS Institute, 2003) on separated data, we still find a solution, but typically some of the components are very large, reflecting the situation discussed above. It will become apparent below, that when observations that are considered to be outliers are excluded from the dataset, separation is often the result. Thus the MLE does not exist.

Rousseeuw and Christmann (2003) introduced the MEL estimator to overcome this difficulty when the MLE does not exist. The MEL estimator may be summarised as follows. Set $\delta=0.01$. Define

$$\overline{\pi}=\frac{1}{N}\sum_{n=1}^{N} y_n,$$

$$\widehat{\pi}=\max(\delta,\min(1-\delta,\overline{\pi})),$$

$$\delta_0=\widehat{\pi}\delta/\left(1+\delta\right)$$

and

$$\delta_1 = (1 + \widehat{\pi}\delta) / (1 + \delta).$$

Transform the y_n's to $\widetilde{y}_n = (1 - y_n)\delta_0 + y_n\delta_1$. Then the MEL estimator chooses $\boldsymbol{\beta}$ to maximise the "estimated" log likelihood

$$\sum_{n=1}^{N} \widetilde{D}_n(\boldsymbol{\beta}) \text{ with } \widetilde{D}_n(\boldsymbol{\beta}) = \widetilde{y}_n \log p(\mathbf{x}_n, \boldsymbol{\beta}) + (1 - \widetilde{y}_n) \log(1 - p(\mathbf{x}_n, \boldsymbol{\beta})). \tag{3.5}$$

Unlike the classical MLE, Rousseeuw and Christmann (2003) show that when $0 < \delta_0 < \delta_1 < 1$ and the dataset has a design matrix of full column rank, the MEL estimator always exists and is unique. The argument we used previously to show that the MLE does not exist if there is no overlap, no longer holds, because the equivalent of (3.4) does not exist in this case.

The related robust WEMEL estimator is defined as the maximiser over $\boldsymbol{\beta}$ of the weighted estimated log likelihood $\sum_{n=1}^{N} w_n \widetilde{D}_n(\boldsymbol{\beta})$. The weights only depend on how far away $\mathbf{x}_n$ is from the bulk of the data. They use

$$w_n = M / \max\left\{RD^2(\mathbf{x}_n^*), M\right\}, \tag{3.6}$$

where $\mathbf{x}_n^* = (x_{n,1}, ..., x_{n,K})^\top$ and $RD(\mathbf{x}_n^*)$ is its robust distance and M is the 75^{th} percentile of all the $RD^2(\mathbf{x}_n^*)$ values. Rousseeuw and Christmann (2003) used the robust distances that come out of the minimum covariance determinant (MCD) estimator of Rousseeuw (1984). In their article, Rousseeuw and Christmann (2003) also provide other properties and results on the performance of the MEL and WEMEL estimators.

As mentioned in Chapter 1, outliers may severely affect the fitted model (3.1). This motivates the need for robust LR procedures (of which the WEMEL estimator is an example). In Chapter 1 we mentioned that one can distinguish between outliers in the x-space and in the y-space (or also referred to as the $y-$direction). Many methods

have been developed to deal with outliers in the x-space. Perhaps the most prominent of these is the fast minimum covariance determinant (FAST-MCD) methodology due to Rousseeuw and Van Driessen (1999b) which is used in the WEMEL procedure. In the present chapter our emphasis is more on outliers in the y-direction. Figure 3.1 illustrates the situation in the case of two regressors.

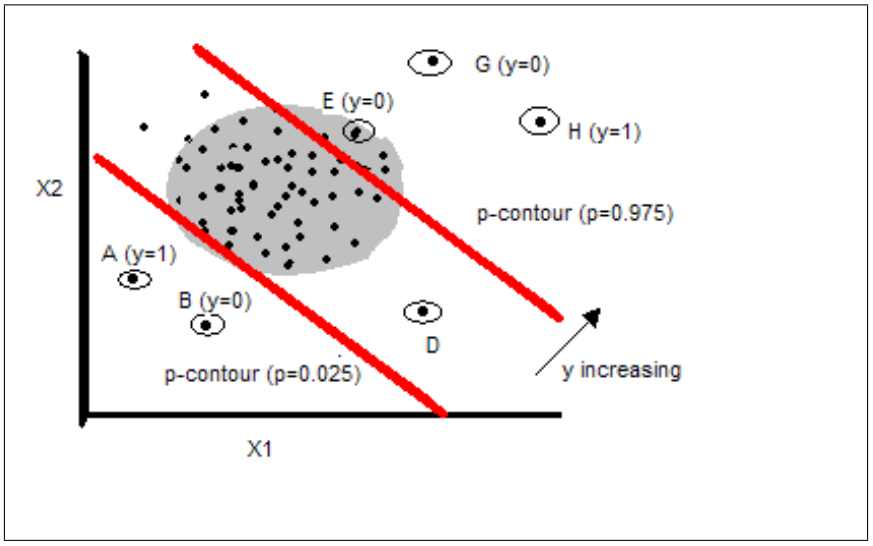


Figure 3.1: x- and y-outliers (two dimensions)

The shaded area which includes most of the (x_1, x_2) pairs may be thought of as containing the x-inliers while the complementary area contains the x-outliers, for example A, B, D, G and H. In Figure 3.1 we have added two contours $p(\mathbf{x}, \boldsymbol{\beta}) = d$ and $p(\mathbf{x}, \boldsymbol{\beta}) = 1 - d$ (with d small). Observations outside the region between the contours with inappropriate y-values may be thought of as possible y-outliers, especially those that are far from the contours, for example A and G. A is an observation with $y = 1$ and p close to 0 and will be called an "uplier", while G and E are observations with $y = 0$ but p close to 1 and will be called "downliers". Again upliers and downliers may also be thought of as bad leverage points, in the sense that they are likely to influence

the estimated regression coefficients seriously in directions different from that implied by the other observations. By contrast observations such as those corresponding to B and H may be called good leverage points in the sense that although they are outlying in the x-space, their y-values are consistent with what is to be expected in the x-region where they lie.

3.2 Detecting outliers using weights

The methodology introduced here has some parallels with the least trimmed squares (LTS) methodology of Rousseeuw and Van Driessen (1999a, b) in ordinary regression; hence we briefly review the LTS methodology. The LTS methodology starts with a lower bound, g_1, on the number of good observations (inliers). We then look for a subset $\mathbf{G}$ of the observations with size $\#(\mathbf{G}) = g_1$, which has smallest residual sum of squares among all subsets of size g_1. The optimal $\mathbf{G_1}$ found in this way is thought most likely to be free of outliers and should therefore result in an estimate of the regression coefficients that is least influenced by possible outliers. The estimated regression coefficients obtained in this way are the LTS estimates based on g_1 observations. It may be that the choice of g_1 is conservative in the sense that there could be observations outside of $\mathbf{G}_1$ which are also good. One could use the LTS estimates to calculate residuals for the observations outside of $\mathbf{G}_1$ and use these residuals to decide which observations to add to $\mathbf{G}_1$ to obtain a larger subset $\mathbf{G}_2$ containing $g_2 \geq g_1$ good observations and then base the final trimmed regression estimate on $\mathbf{G_2}$. The LTS optimisation is computationally difficult and if done exactly, requires a number of steps that grows combinatorially with the number of observations and is practically infeasible especially in large problems. Rousseeuw and Van Driessen (1999a, b)

handle the computational procedure by starting with an initial random choice of $\mathbf{G}$, where $\#(\mathbf{G}) = g_1$, and then improves this choice of $\mathbf{G}$ iteratively until convergence, using a so called C-step procedure. This is repeated many times and the best solution is kept and taken to represent the optimal $\mathbf{G}_1$ required. More detail can be found in Rousseeuw and Van Driessen (1999b).

The LTS procedure is quite attractive and successful in ordinary regression and our initial aim was to formulate an analogue for the LR case, using trimmed likelihood instead of trimmed least squares. This would mean that we want to select a subset $\mathbf{G}$ containing g_1 observations (together with an associated $\boldsymbol{\beta}$) which maximises $\sum_{n\in\mathbf{G}} D_n(\boldsymbol{\beta})$. Again this is computationally difficult but the equivalent to the C-step is already available from the work of Neykov and Muller (2002). However, when implementing the procedure we found that the "optimal" $\mathbf{G}_1$ usually tends to get bogged down among subsets with no overlap for which the corresponding MLE does not exist. This especially happens when one starts with a conservative g_1 which is much smaller than N. Replacing the MLE with the MEL estimator avoids the non-existence issue but does not eliminate the possibility that the "optimal" $\mathbf{G}_1$ may be chosen poorly. To circumvent these problems we decided to follow a downweighting approach rather than a trimming approach. For this purpose let $0 < \epsilon < 1$ and for a given subset $\mathbf{G}$ define a corresponding weighted log likelihood by

$$l_\epsilon(\boldsymbol{\beta}, \mathbf{G}) = \sum_{n\in\mathbf{G}} D_n(\boldsymbol{\beta}) + \epsilon \sum_{n\notin\mathbf{G}} D_n(\boldsymbol{\beta}). \tag{3.7}$$

This expression is a weighted log likelihood function with $w_n = 1$ for $n \in \mathbf{G}$ and $w_n = \epsilon$ for $n \notin \mathbf{G}$. Thus the observations in $\mathbf{G}$ are associated with the higher weight 1 and the observations outside of $\mathbf{G}$ with the lower weight ϵ. In order to make (3.7) large, an

observation that is good in the sense of making a large contribution $D_n(\boldsymbol{\beta})$ to the log likelihood function should be in $\mathbf{G}$ while ones that are bad in the sense of making low contributions should be outside $\mathbf{G}$. As in LTS, to capture the good observations in $\mathbf{G}$ we now look for that subset $\mathbf{G}$ with size $\#(\mathbf{G}) = g_1$ and associated $\boldsymbol{\beta}$ which maximise $l_\epsilon(\boldsymbol{\beta}, \mathbf{G})$ among all subsets of given size g_1. For any given set $\mathbf{G}$ it is relatively easy to calculate the corresponding optimiser, $\boldsymbol{\beta}^*(\mathbf{G})$ of $l_\epsilon(\boldsymbol{\beta}, \mathbf{G})$ over $\boldsymbol{\beta}$ using, for example, a Newton-Raphson method (see Hastie et al., 2001). Maximising $l_\epsilon(\boldsymbol{\beta}, \mathbf{G})$ over both $\boldsymbol{\beta}$ and $\mathbf{G}$ can be handled by a more general form of the C-step procedure of Neykov and Muller (2002), which is presented in Appendix B.1, C-step lemma. Once this has been done we have an optimal $\mathbf{G_1}$ with an associated estimator $\boldsymbol{\beta}^*(\mathbf{G}_1)$. Next we must set up a criterion that can be used to identify outliers. To do this, we use a small number c with $0 < c < 1$ and then declare observation n to be an outlier if $y_n = 1$ but $p(\mathbf{x}_n, \boldsymbol{\beta}^*(\mathbf{G}_1)) \leq c$ or if $y_n = 0$ but $p(\mathbf{x}_n, \boldsymbol{\beta}^*(\mathbf{G}_1)) > 1 - c$. The reasoning here is that if $y_n = 1$, and $p(\mathbf{x}_n, \boldsymbol{\beta}^*(\mathbf{G}_1))$ is small it is probable that observation n is an uplier and therefore should be downweighted (given the weight ϵ). Similarly if $y_n = 0$, and $p(\mathbf{x}_n, \boldsymbol{\beta}^*(\mathbf{G}_1)))$ is large it is probable that observation n is a downlier and should therefore be downweighted. The remaining observations are given weights 1. We could now let $\mathbf{G_2}$ be the set of observations that are given the high weights in this classification step and then compute a final weighted estimate for the regression coefficients as $\boldsymbol{\beta}^*(\mathbf{G}_2)\dot{}$ which by definition maximises $l_\epsilon(\boldsymbol{\beta}, \mathbf{G_2})$. This is the DOUW procedure.

There are further issues that have to be dealt with to complete the specification of the DOUW method. Among these are the initial choice of g_1, the number of iterations in the C-steps, the choice of ϵ and the choice of cut-off c. The choices of ϵ and c will be

dealt with after we have reported the results of a simulation study in Section 3.3. The choice of g_1 and the number of iterations are spelled out in the form of the following **pseudocode for the DOUW procedure** and was programmed in PROC IML of SAS (SAS Institute, 2003).

1. Select $g_1 = \max\{[(N+K+1)/2], K+1\}$ where $[x]$ is the integer part of x. This is in line with the suggestion of Rousseeuw and van Driessen (1999a) for the FAST-LTS method.

   (a) Repeat 50 times:

      i. Select a starting subset $\mathbf{H} \subset \{1, ..., N\}$ at random with $\#(\mathbf{H}) = K + 1 = g_1$. Calculate $\boldsymbol{\beta}^*(\mathbf{H})$, the $D_n(\boldsymbol{\beta}^*(\mathbf{H}))$'s and find the π_i's so that $D_{\pi_1}(\boldsymbol{\beta}^*(\mathbf{H})) \geq ... \geq D_{\pi_N}(\boldsymbol{\beta}^*(\mathbf{H}))$. Put $\mathbf{G} = \{\pi_1, ..., \pi_{g_1}\}$. Carry out two C-steps starting with this $\mathbf{G}$ and ending with $\mathbf{G}''$ say.

   (b) For all these 50 $\mathbf{G}''$, store the five best results, in terms of the highest values of $l_\epsilon(\boldsymbol{\beta}^*(\mathbf{G}''), \mathbf{G}'')$.

   (c) For each of these five best results, repeat the C-step iteration until convergence.

   (d) Retain the overall best subset $\mathbf{G_1}$ among the last five.

2. Put $\mathbf{G_2} = \{n : (y_n = 1$ and $p(\mathbf{x}_n, \boldsymbol{\beta}^*(\mathbf{G_1})) \geq c\ )$ or $(y_n = 0$ and $p(\mathbf{x}_n, \boldsymbol{\beta}^*(\mathbf{G_1})) \leq 1 - c)\}$ and then calculate the final weighted estimate $\boldsymbol{\beta}^*(\mathbf{G}_2\dot{)}$ which maximises $l_\epsilon(\boldsymbol{\beta}, \mathbf{G_2})$. The outliers are the observations outside of $\mathbf{G}_2$.

This basic DOUW procedure can be varied in a number of ways as indicated in the following list:

- We varied the number of repetitions in step 2 from 50 to 500 and found that it made little difference in most cases and feel that 50 repetitions are generally sufficient.

- The number of C-steps in step 2 (a) and subsets in step 2 (b) were also varied. For most of the results reported in this chapter two C-steps and five subsets were found to be suitable.

- Since $\mathbf{G_1}$ is likely to be outlier free by construction we could restrict the choice of $\mathbf{G}_2$ in step 3 to observations outside of $\mathbf{G_1}$ only, i.e. we could define $\mathbf{G_2} = \mathbf{G_1} \cup \{n \notin \mathbf{G_1} : (y_n = 1 \text{ and } p(\mathbf{x}_n, \boldsymbol{\beta}^*(\mathbf{G_1})) \geq c \ ) \text{ or } (y_n = 0 \text{ and } p(\mathbf{x}_n, \boldsymbol{\beta}^*(\mathbf{G_1})) \leq 1 - c)\}$. However there is no guarantee that $\mathbf{G_1}$ will be outlier free in the strict sense that a high response observation has estimated probability larger than c and the low response observations has estimated success probabilities less than $(1 - c)$. It seems unreasonable to treat such observations differently from those outside of $\mathbf{G_1}$. Nonetheless, in our experience this variation makes little difference to the properties of this procedure.

- For a small to moderate sample size N, the C-step algorithm does not take much time, but when N grows the computation time increases. Nested extensions similar to those of Rousseeuw and Van Driessen (1999a) for the LTS methodology can be used to limit the computational effort required for large datasets.

- We could follow step 3 by a fourth step which would identify the outliers using $\boldsymbol{\beta}^*(\dot{\mathbf{G}_2})$ as the current estimate, i.e. we define $\mathbf{G_3} = \{n : (y_n = 1 \text{ and } p(\mathbf{x}_n, \boldsymbol{\beta}^*(\mathbf{G_2})) \geq c) \text{ or } (y_n = 0 \text{ and } p(\mathbf{x}_n, \boldsymbol{\beta}^*(\mathbf{G}_2)) \leq 1 - c)\dot{\}}$ and declare the observations outside of $\mathbf{G_3}$ as outliers. This can be repeated until convergence

is obtained. However, this declaration of current outlier candidates was seldom worthwhile in our simulation study.

- Another variation is to use $g_2 = \#(\mathbf{G}_2)$ obtained in step 3 only as an estimate of the number of good observations. In other words, after step 3, we repeat steps 2 and 3 to identify an optimal subset of size g_2 to which to assign the high weights while the rest get the low weights. Again we found that this additional computational effort seldom appeared to be worthwhile.

- We could use different choices for ϵ in step 3. For example since we feel more confident that the observations outside of $\mathbf{G}_2$ are indeed outliers, we could use $\epsilon/2$, say, and compute the final estimate $\boldsymbol{\beta}^*(\dot{\mathbf{G}_2})$ by maximising $l_{\epsilon/2}(\boldsymbol{\beta}, \mathbf{G_2})$. We could even choose $\epsilon = 0$ thereby trimming the outliers in the computation of the final estimate. However we found that the results are quite similar when different values of ϵ are used.

- We could replace the MLE with the MEL estimate throughout the procedure, i.e. work with the weighted estimated log likelihood (3.7) with $D_n(\boldsymbol{\beta})$ replaced by $\widetilde{D}_n(\boldsymbol{\beta})$. We comment on this after presenting the results in Section 3.3.

As a first illustration we compare the MEL, WEMEL and DOUW procedures on the four small artificial datasets of Rousseeuw and Christmann (2003). We also used these examples in Chapter 1, Section 1.3. The estimated success probability curves with respect to the MEL, WEMEL and DOUW procedures (with choices $\epsilon = 0.2$ and $c = 0.05$) are given in Figure 3.2 and the estimates are given in Table 3.1. We included ML estimates in the table. Case (a) is in the top left panel of Figure 3.2, case (b) in the top right, case (c) in the bottom left and case (d) in the bottom right panel.

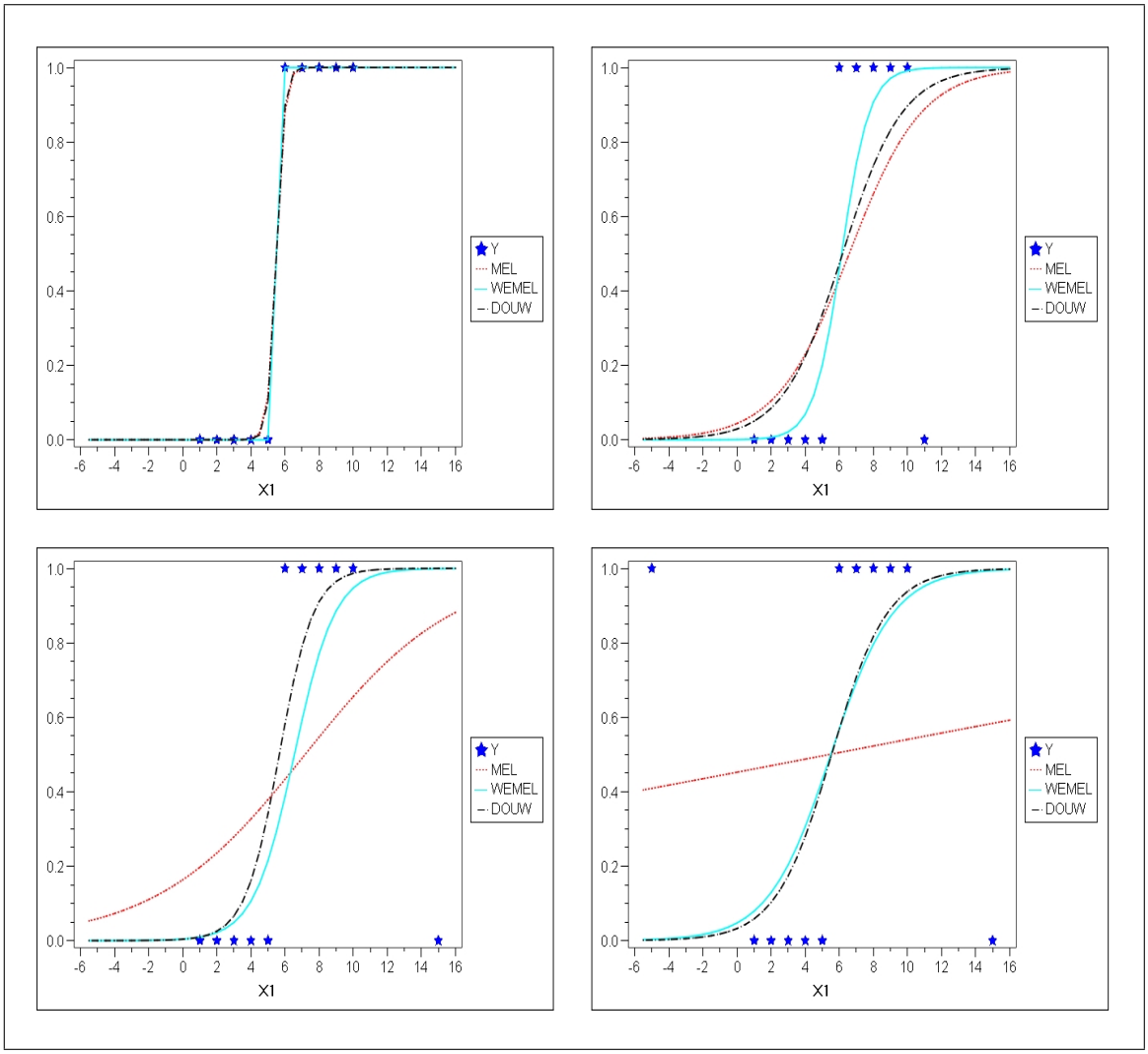


Figure 3.2: Probability success curves of MEL, WEMEL and DOUW

We assume that the data in case (a) reflects the true situation in the sense that successes and failures are practically separated at $x = 5.5$ while the additional observations in (b), (c) and (d) represent outliers. It is clear that all procedures perform equally well in case (a) where no outliers are present. Consider the large values for the ML coefficients in case (a) being due to separation. For cases (b), (c) and (d) the WEMEL and DOUW estimates outperform the MEL estimates. The DOUW estimates decrease systematically (in absolute value) across the three cases. If we take $\epsilon = 0$ it may happen that the procedure chooses only the observations corresponding to $y_n = 0$ as the best subset in the first phase and then trim the $y_n = 1$ observations in the second phase (or vice versa). With ϵ positive this wrong conclusion is avoided.

Case (a)	$\widehat{\beta}_0$	$\widehat{\beta}_1$	**Case (b)**	$\widehat{\beta}_0$	$\widehat{\beta}_1$
ML	-183.8074	33.4195	ML	-3.1389	0.4772
MEL	-22.2688	4.0489	MEL	-3.0880	0.4691
WEMEL	-23.9615	4.3566	WEMEL	-3.5180	0.5682
DOUW	-146.8575	26.7010	DOUW	-7.5155	1.2249
Case (c)	$\widehat{\beta}_0$	$\widehat{\beta}_1$	**Case (d)**	$\widehat{\beta}_0$	$\widehat{\beta}_1$
ML	-1.6526	0.2307	ML	-0.1931	0.0356
MEL	-1.6333	0.2276	MEL	-0.1911	0.0353
WEMEL	-5.6379	0.9953	WEMEL	-3.3921	0.6112
DOUW	-5.4845	0.8373	DOUW	-3.0009	0.5457

Table 3.1: ML, MEL, WEMEL and DOUW estimates

We also refer back to the first example used in Chapter 1, Section 1.3. Again a sample of 50 (x, y) observations is constructed by generating x_n, $n = 1, ..., 50$ from

a $N(0,1)$ distribution. As in Chapter 1, we again show the true probability curve, $p_n = p(x,(1,2)^\top)$ as well as the estimated probability curve, $\widehat{p}_n = p(x,\widehat{\boldsymbol{\beta}})$ but this time we show the estimated probability curves when using the ML, MEL, WEMEL and DOUW algorithms, see Figure 3.3. In the left panel we show the data without outliers

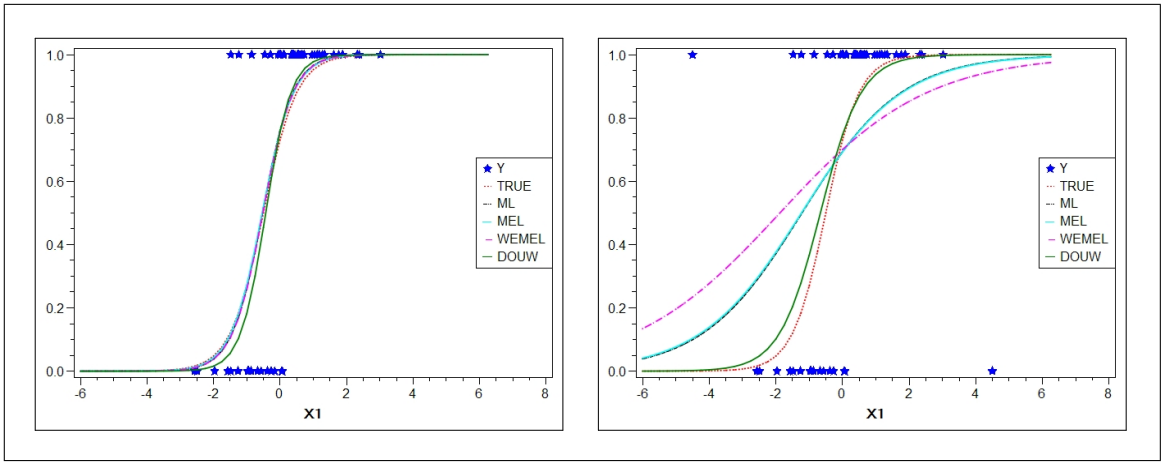


Figure 3.3: Probability success curves of ML, MEL, WEMEL and DOUW compared with the true probability success curve

and in the right panel with outliers (as in Chapter 1, Section 1.3). In the left panel the ML, MEL, WEMEL and DOUW curves are very similar and almost indistinguishable. In the right panel we observe that the ML and MEL curves are almost indistinguishable and the true and DOUW curves are also almost indistinguishable. The DOUW curves outperform the ML, MEL and WEMEL curves. We also added the estimates in Table 3.2. It seems that the DOUW procedure performs very well if there are outliers present in the data. We investigate this more thoroughly using simulation studies.

Data (no outliers)	$\widehat{\beta}_0$	$\widehat{\beta}_1$	**Data (with outliers)**	$\widehat{\beta}_0$	$\widehat{\beta}_1$
ML	1.1604	2.1958	ML	0.8150	0.6701
MEL	1.1404	2.1261	MEL	0.8129	0.6601
WEMEL	1.1323	2.1914	WEMEL	0.8483	0.4521
DOUW	1.1353	2.6400	DOUW	1.0825	1.6321

Table 3.2: ML, MEL, WEMEL and DOUW estimates

3.3 Simulation studies

3.3.1 Design

To study the properties of the DOUW procedure and to get some guidance on the effects of the choices of the tuning parameters, we report the results of simulation studies in this section. As in Rousseeuw and Christmann (2003) we use $K = 2$ regressors x_1 and x_2 and we take $\boldsymbol{\beta}^\top = (1, 1, 2)$. We study the following **cases**:

1. $N = 100$ with x_1 and x_2 independently $N(0, 1)$ distributed,
2. $N = 100$ with x_1 and x_2 independently $N(0, 4)$ distributed,
3. $N = 100$ with x_1 and x_2 independently t_3 distributed and
4. $N = 20, 50, 100, 200$ with x_1 and x_2 independently $N(0, 1)$ distributed.

These specifications give the true values of the model $p(\mathbf{x}, \boldsymbol{\beta})$ of (3.1) to be fitted to the data (the "**approximating family**" in the terms of Linhart and Zucchini, 1986). Of course data generated under this model may contain apparent up- and downliers since there is always positive probability to get $y_n = 1$ even though $p(\mathbf{x}, \boldsymbol{\beta})$ may be small and vice versa. One way to study the sensitivity of procedures to outliers is to manually

add a number of additional up- and downliers into the data. Instead of following this approach we choose to select a data generating model $q(\mathbf{x})$ (the "**operating model**" in the terms of Linhart and Zucchini, 1986) with a flatter success probability curve than $p(\mathbf{x}, \boldsymbol{\beta})$. Data generated from such a model will then contain more up- and downliers than expected under $p(\mathbf{x}, \boldsymbol{\beta})$. An example of such a $q(\mathbf{x})$ is

$$q(\mathbf{x}) = q(\mathbf{x}, \boldsymbol{\beta}, \alpha) = \begin{cases} \alpha & \text{if } p(\mathbf{x}, \boldsymbol{\beta}) < \alpha \\ p(\mathbf{x}, \boldsymbol{\beta}) & \text{if } \alpha \leq p(\mathbf{x}, \boldsymbol{\beta}) \leq 1 - \alpha \\ 1 - \alpha & \text{if } p(\mathbf{x}, \boldsymbol{\beta}) > 1 - \alpha \end{cases} \tag{3.8}$$

which will be used extensively in this section. Here α is a parameter ranging between 0 and 1/2. Figure 3.4 panel (a) illustrates the corresponding success probability curves for the choice $\alpha = 0.2$.

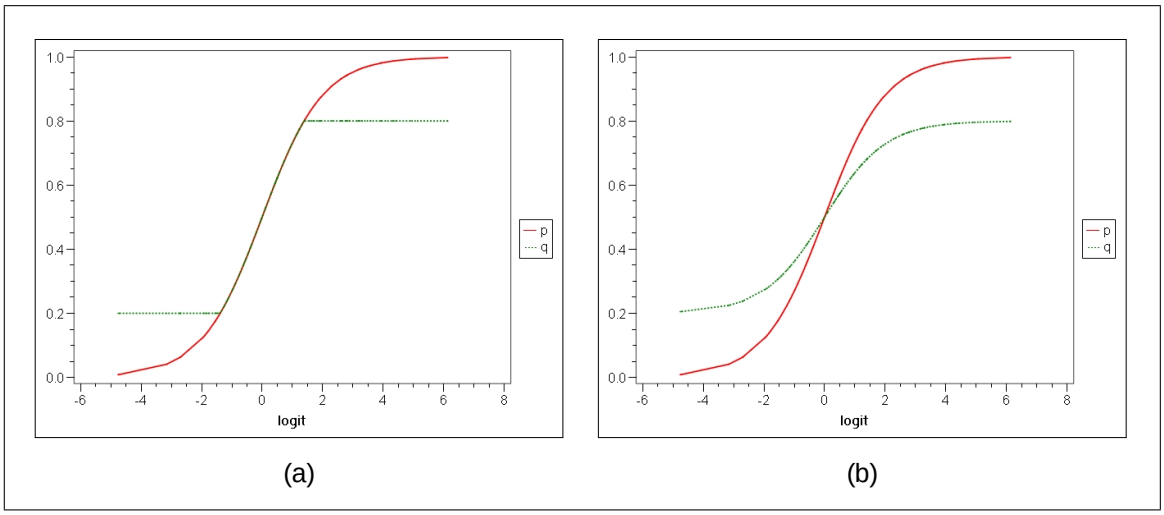


Figure 3.4: Examples of success probability curves of $p(\mathbf{x}, \boldsymbol{\beta})$ and $q(\mathbf{x}, \boldsymbol{\beta}, \alpha)$ with $\alpha = 0.2$. Panel (a) $q(\mathbf{x}, \boldsymbol{\beta}, \alpha)$ given by (3.8) and panel (b) by HLR

For $p(\mathbf{x}, \boldsymbol{\beta})$ between α and $1 - \alpha$ the two curves coincide. For $p(\mathbf{x}, \boldsymbol{\beta}) < \alpha$ the probability of getting $y = 1$ (an uplier) is larger under the data generating model $q(\mathbf{x})$ so that

relatively more upliers will tend to be present in the actual data than expected under the fitting model $p(\mathbf{x}, \boldsymbol{\beta})$. For $p(\mathbf{x}, \boldsymbol{\beta}) > 1 - \alpha$ the probability of getting $y = 0$ (a downlier) is larger under the data generating model $q(\mathbf{x})$ so that relatively more downliers will tend to be present in the actual data than expected under $p(\mathbf{x}, \boldsymbol{\beta})$.

If $\alpha = 0$ the two models do not differ but as α increases the number of outliers produced under the data generating model increases. Many other examples of $q(\mathbf{x})$ are possible, for example the hidden logistic regression (HLR) model used in Copas (1988) and Rousseeuw and Christmann (2003). For our purposes we may take this HLR model to be given by $q(\mathbf{x}, \boldsymbol{\beta}, \alpha) = \alpha + (1 - 2\alpha)p(\mathbf{x}, \boldsymbol{\beta})$ and illustrate it in panel (b) of Figure 3.4. Again the severity of the number of outliers increases as α increases. The results for both these models are largely similar and to save space we report only in terms of the data generating model of (3.8).

3.3.2 Performance criteria

The first performance criterion to be used may be described as the average mean squared error of estimation of the true $\boldsymbol{\beta}$'s, i.e. if $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_K)$ is the estimator of $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_K)$ then

$$CB = (1/(K+1))\, E \sum_{k=0}^{K} (\widehat{\beta}_k - \beta_k)^2. \tag{3.9}$$

It is often also important that the success probabilities be estimated accurately. A criterion to judge this accuracy may be described as the average mean weighted squared error of estimation of the success probabilities at the actual x-values i.e. if $\widehat{p}_n = p(\mathbf{x}_n, \widehat{\boldsymbol{\beta}})$ then

$$CWP = (1/N) E \sum_{n=1}^{N} \lambda_n \left(\widehat{p}_n - p(\mathbf{x}_n, \boldsymbol{\beta})\right)^2 \tag{3.10}$$

where λ_n is a weight associated with the n^{th} observation. Often the observations with small or large response probabilities are particularly important and we can choose the weights to emphasise this. One possibility is to take

$$\lambda_n = 1/p(\mathbf{x}_n, \boldsymbol{\beta}) \left(1 - p(\mathbf{x}_n, \boldsymbol{\beta})\right), \tag{3.11}$$

but then λ_n tends to infinity when $p(\mathbf{x}_n, \boldsymbol{\beta})$ tends to 0 or 1 causing these extreme cases to dominate the others. Therefore to lessen the influence of the extreme cases on the weights we take

$$\lambda_n = 1/\left(A + p(\mathbf{x}_n, \boldsymbol{\beta}) \left(A + 1 - p(\mathbf{x}_n, \boldsymbol{\beta})\right)\right). \tag{3.12}$$

As long as A is positive the weights no longer tends to infinity when $p(\mathbf{x}_n, \boldsymbol{\beta})$ tends to 0 or 1. The choice of A is somewhat arbitrary but does not seem to matter much for our purposes and we report mostly for the choice $A = 0.01$. We could also consider an unweighted form where we take $\lambda_n = 1,$ so that the criterion becomes

$$CP = (1/N)E \sum_{n=1}^{N} \left(\widehat{p}_n - p(\mathbf{x}_n, \boldsymbol{\beta})\right)^2. \tag{3.13}$$

The expected value is estimated by averaging over the simulation results in equations (3.9), (3.10) and (3.13).

Many other criteria could be considered (for example judging bias and variance separately or using absolute deviation, etc.). There are however, other features that need to be varied (for example the sample size N, the number of regressors K, and the distribution of the $\mathbf{x}_n$'s as well as the DOUW tuning constants, etc.). It is not feasible to report on a large number of criteria in conjunction with all these other features; hence attention will be restricted to these criteria.

3.3.3 Choice of ϵ and c

We first study the effect of the choice of ϵ on the performance of the DOUW procedure. Provisionally we choose the cut-off $c = 0.05$ and will comment on its effect later on. We show that choosing ϵ very small causes the procedure to perform very poorly while choosing a moderate value for ϵ leads to good performance in the presence of outliers, at limited cost when no outliers are present.

To begin with consider **case 1**. The top two panels of Figure 3.5 show CB for the choices $\epsilon = \{0.01, 0.1, 0.2, 0.3, 0.4, 0.5\}$ as functions of α with the data generating model (3.8) based on 1000 simulation runs. In the left hand panel it is evident that for $\epsilon = 0.01$, CB is uniformly very large. If ϵ is made smaller than 0.01 CB rapidly becomes even worse. To understand the reasons for this phenomenon we studied the best subset $\mathbf{G_1}$ for the very small choices of ϵ. It turns out that most often the observations in this subset can be separated. Strictly speaking in such cases the ML estimators of the $\boldsymbol{\beta}$'s do not exist; however the optimiser that we used in this study did converge but to values rather different from the true beta's and this discrepancy causes the large values for CB shown in Figure 3.5. Therefore ϵ should not be chosen very small. The right hand top panel of Figure 3.5 shows the same curves as the left hand panel with the curve corresponding to $\epsilon = 0.01$ omitted in order to increase the vertical scale and display more details of the remaining curves. For comparison purposes we also included the curves of MEL and WEMEL. Considering first the CB curve corresponding to $\epsilon = 0.1$, the DOUW procedure outperforms both the MEL and WEMEL procedures if there are a substantial number of outliers ($\alpha > 0.15$) but this advantage comes at a cost when there are few or no outliers in that CB for the DOUW procedure is larger than CB for the MEL and WEMEL procedures when α is small.

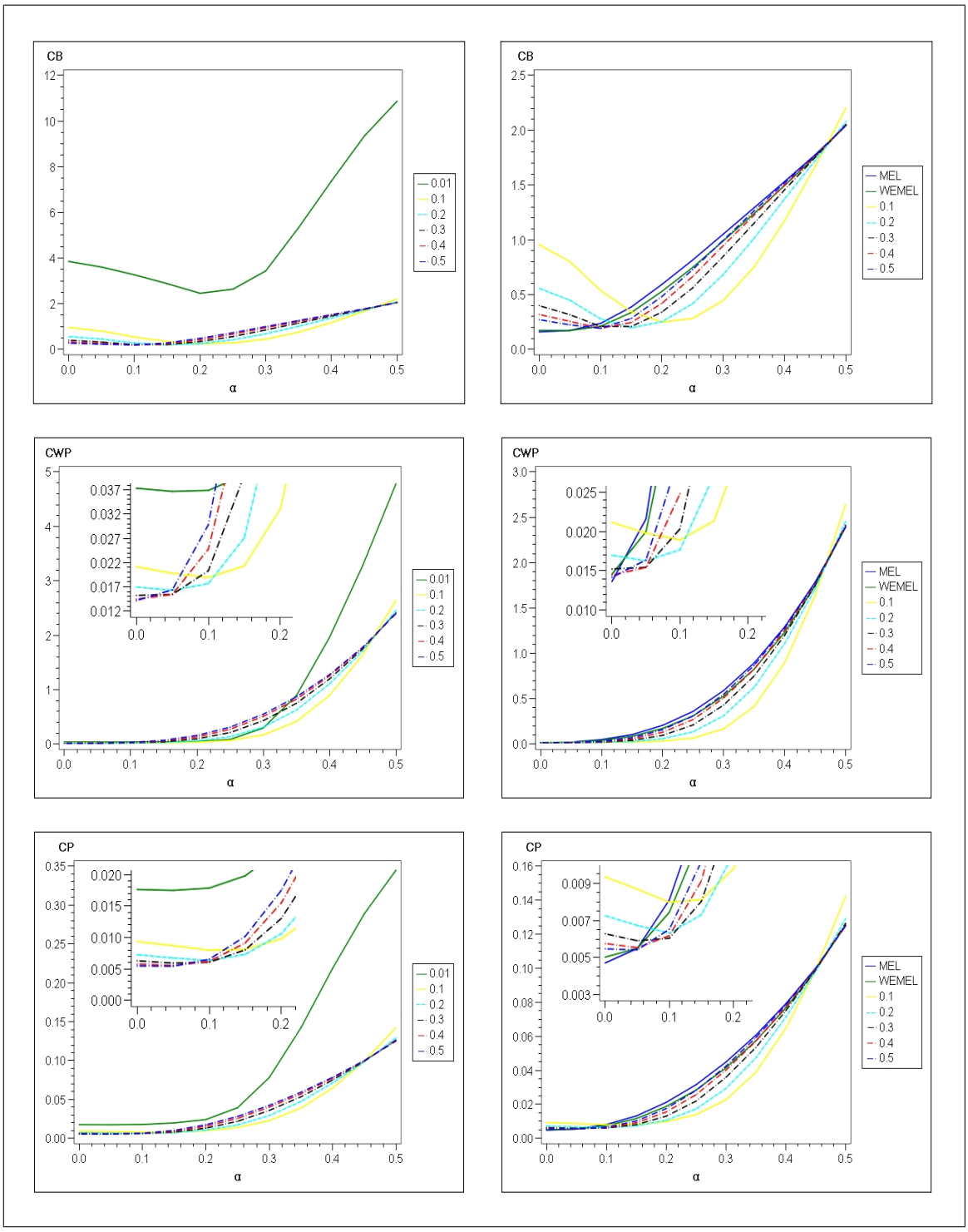


Figure 3.5: CB, CWP and CP values for $\epsilon = \{0.01, 0.1, 0.2, 0.3, 0.4, 0.5\}$ and $c = 0.05$ for case 1

If we increase ϵ to 0.2 the DOUW procedure outperforms the MEL and WEMEL procedures over a larger range of α values but not to the same extent as for $\epsilon = 0.1$. At the same time a smaller cost when few outliers are present is involved for $\epsilon = 0.2$ as compared to $\epsilon = 0.1$. With further increases in ϵ the extent of the improvement when many outliers are present diminishes and the required cost when few outliers are present also diminishes so that the CB curve resembles those of the MEL and WEMEL procedures progressively more closely. Examining the curves of $\epsilon = 0.4$ and $\epsilon = 0.5$ it would seem that choosing ϵ greater than 0.3 leaves little scope for different performance of the DOUW procedure as compared to the MEL and WEMEL procedures so that the interesting range for ϵ is from 0.1 to 0.3 and our impression at this time is that $\epsilon = 0.2$ is a reasonable compromise for the scenarios modelled.

The middle and the bottom four panels of Figure 3.5 show CWP and CP for the cases corresponding to the top panels. We have also inserted an enlargement on each of these four graphs to show details of the curves at small values of α. In the left of these four panels it is again evident that choosing $\epsilon = 0.01$ leads to very poor performance in terms of the CWP and CP criteria and we have confirmed that smaller choices of ϵ lead to even worse performance. The right hand middle and bottom panels show that the DOUW procedure with $\epsilon = 0.1$ improves substantially on the MEL and WEMEL procedures when many outliers are present (for $\alpha > 0.09$) but again this comes at a cost when few outliers are present as is shown in the enlargements. Making ϵ larger decreases the cost but also decreases the benefit when many outliers are present. We can clearly see that the results for CWP and CP are similar. This similarity was true for all the different cases and from now on we only report on CB and CP. Thus the effects of varying ϵ are quite similar for all performance criteria. The choice of ϵ

must take into account the cost/benefit balance associated with ϵ. This feature is a typical dilemma in the choice of tuning constants in statistical procedures; for example selecting a small size for a test is desirable to make the type I error rate small, but also decreases the power of the test to reject the null hypothesis when it is not true and vice versa.

In Figure 3.6 we repeat the analysis but this time for a different cut-off value c. We used $c = 0.01$ for the two left panels and $c = 0.1$ for the two right panels. The results are qualitatively similar to Figure 3.5 where we used $c = 0.05$. However with the choice of $c = 0.01$ we have to make ϵ smaller (for example 0.1) for the cost/benefit balance not to disappear, since otherwise the DOUW procedure performs similar to the MEL and WEMEL procedures. By contrast, for the choice $c = 0.1$ the benefit increases but so does the cost and to keep these in balance we need to make a larger choice of ϵ (for example 0.3). It appears that the combination of choices, $(\epsilon, c) = (0.1, 0.01)$, $(0.2, 0.05)$, $(0.3, 0.1)$ are reasonable but to some extent this combination is a judgement call based on limited experience and the issue of making sensible choices in practice is still open at this stage and will be investigated in future research. However, in the next section we consider some examples to investigate the affect that changes in these two tuning parameters have on the results. This will give some practical guidelines.

In Figure 3.7 we compared the DOUW procedure with the MLE replaced by the MEL estimate throughout (last variation in Section 3.2). Here we see that the DOUW-MEL procedure performs better than the DOUW-ML procedure when α is small (few outliers), but otherwise the two versions of the DOUW procedure perform quite similarly.

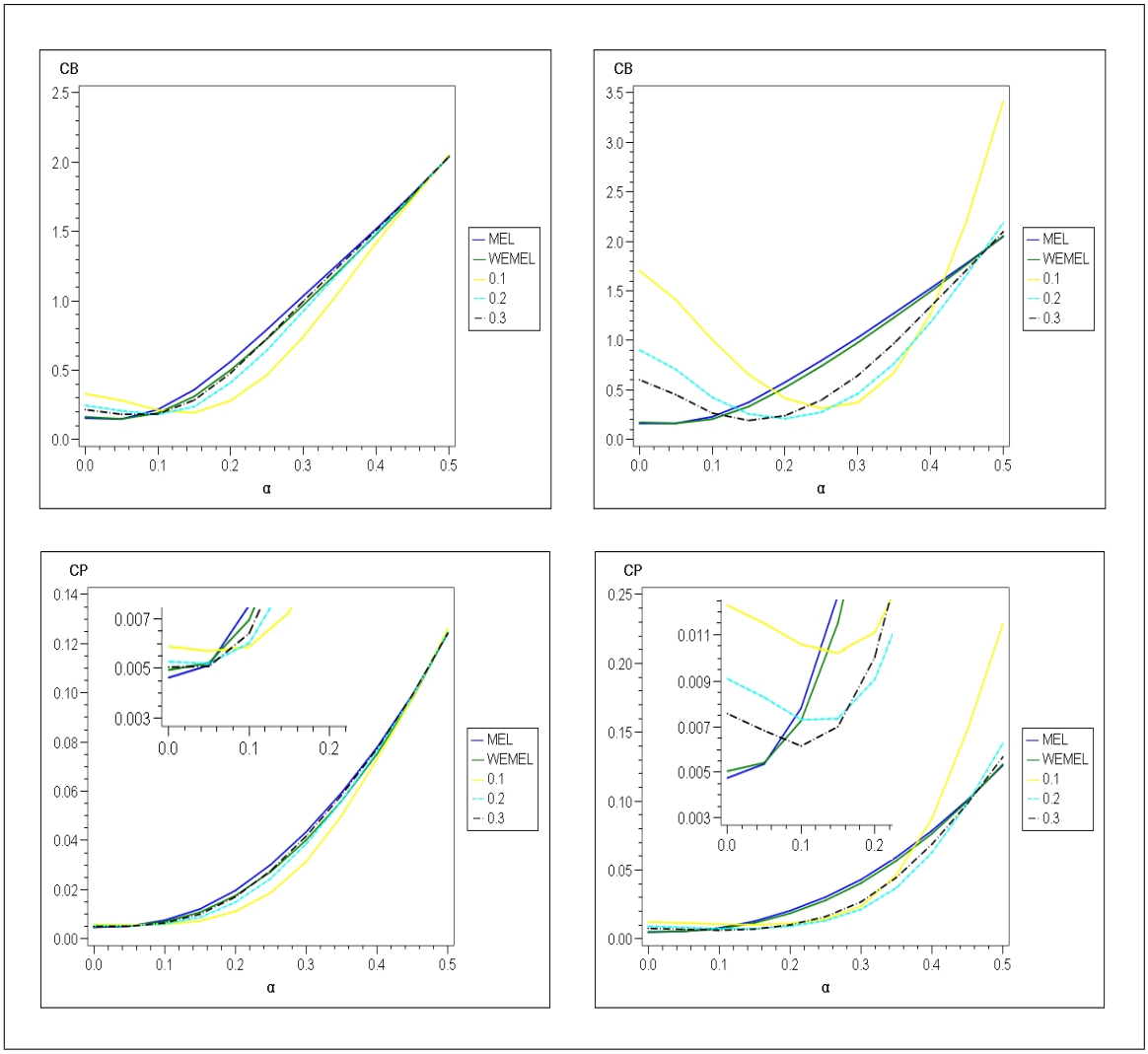


Figure 3.6: CB and CP values for $\epsilon = \{0.1, 0.2, 0.3\}$ (with $c = 0.01$ left and $c = 0.10$ right) for case 1

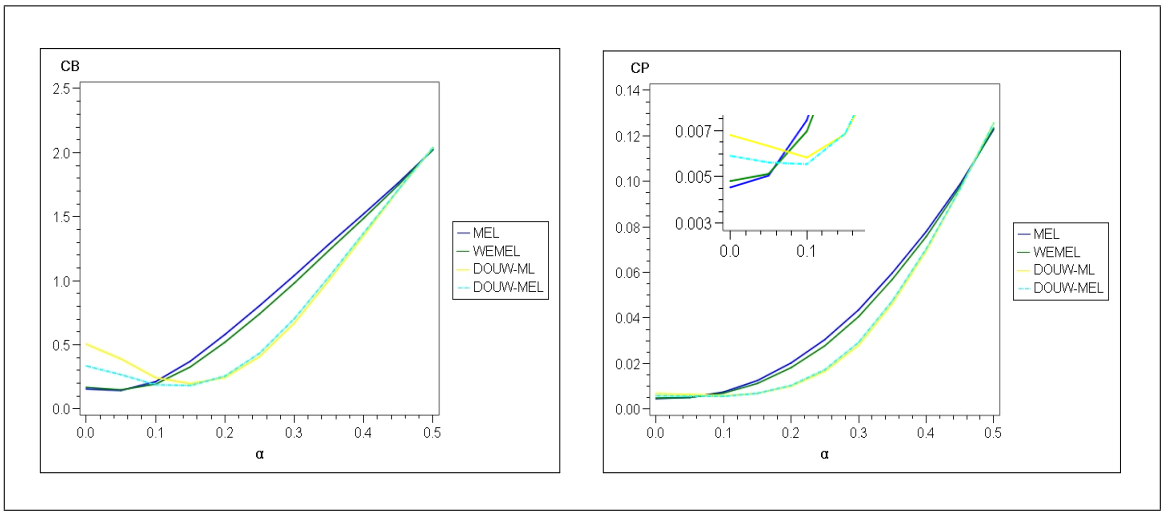


Figure 3.7: CB and CP values for DOUW when using ML and MEL for case 1 ($\epsilon = 0.2$, $c = 0.05$)

Proceeding to **cases 2** and **3**, Figures 3.8 and 3.9 show the corresponding CB and CP curves respectively. A similar cost/benefit balance effect is again evident as in case 1. In both these figures, we used $c = 0.05$ on the left and $c = 0.10$ on the right while ϵ varies over the values 0.1 to 0.3. We see that the choice of the cut-off again influences the extent of the cost/benefit balance. For illustrative purposes, we from now on use $\epsilon = 0.2$ and $c = 0.05$. We also did the same analysis with data generated under the HLR model and found similar results.

For **case 4** we have plotted the CB and CP values for $n = 50$ and $n = 200$ (Figure 3.10). These may be compared to the case $n = 100$ in the top right and bottom right panels of Figure 3.5. We can see in all these cases that we have similar behaviour with regards to the cost/benefit balance of the DOUW procedure vs the MEL and WEMEL procedures. However the DOUW procedure does even better than the MEL and WEMEL procedures as n increases. The benefit is very large and the cost is minimal if we consider $n = 200$, while the benefit of the DOUW procedure above the WEMEL

and MEL procedures, is less when $n = 50$ and the cost of the DOUW procedure is higher when $n = 50$.

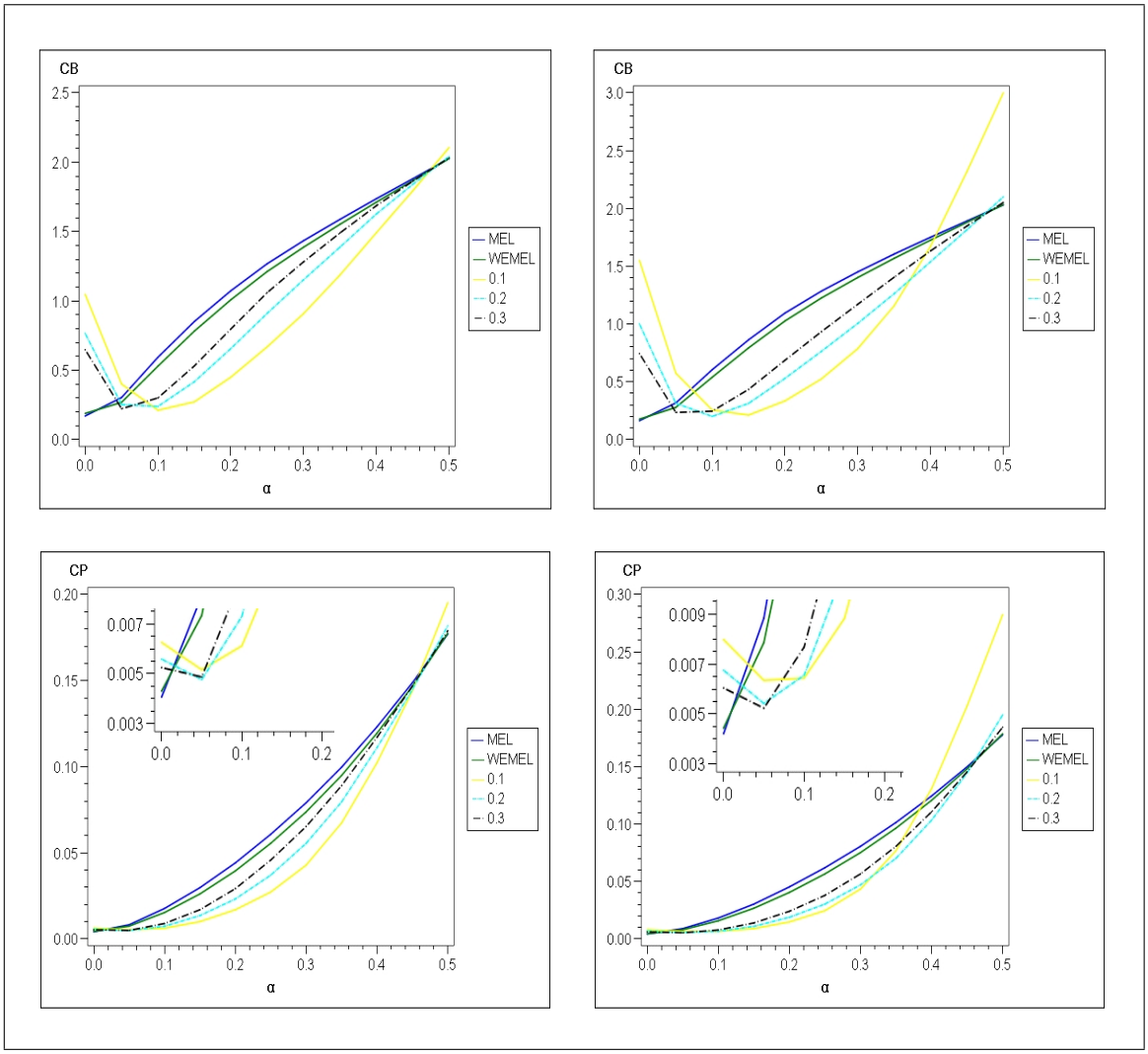


Figure 3.8: CB and CP values for $\epsilon = \{0.1, 0.2, 0.3\}$ and $c = 0.05$(left), $c = 0.1$(right) for case 2

We have also run simulations on 3 and more regressors and found similar results namely increasing ϵ diminishes both the cost and the benefit associated with the DOUW procedure and so does decreasing c. The effects are more pronounced in large samples than in small samples. The behaviour of the tuning constants, ϵ and c,

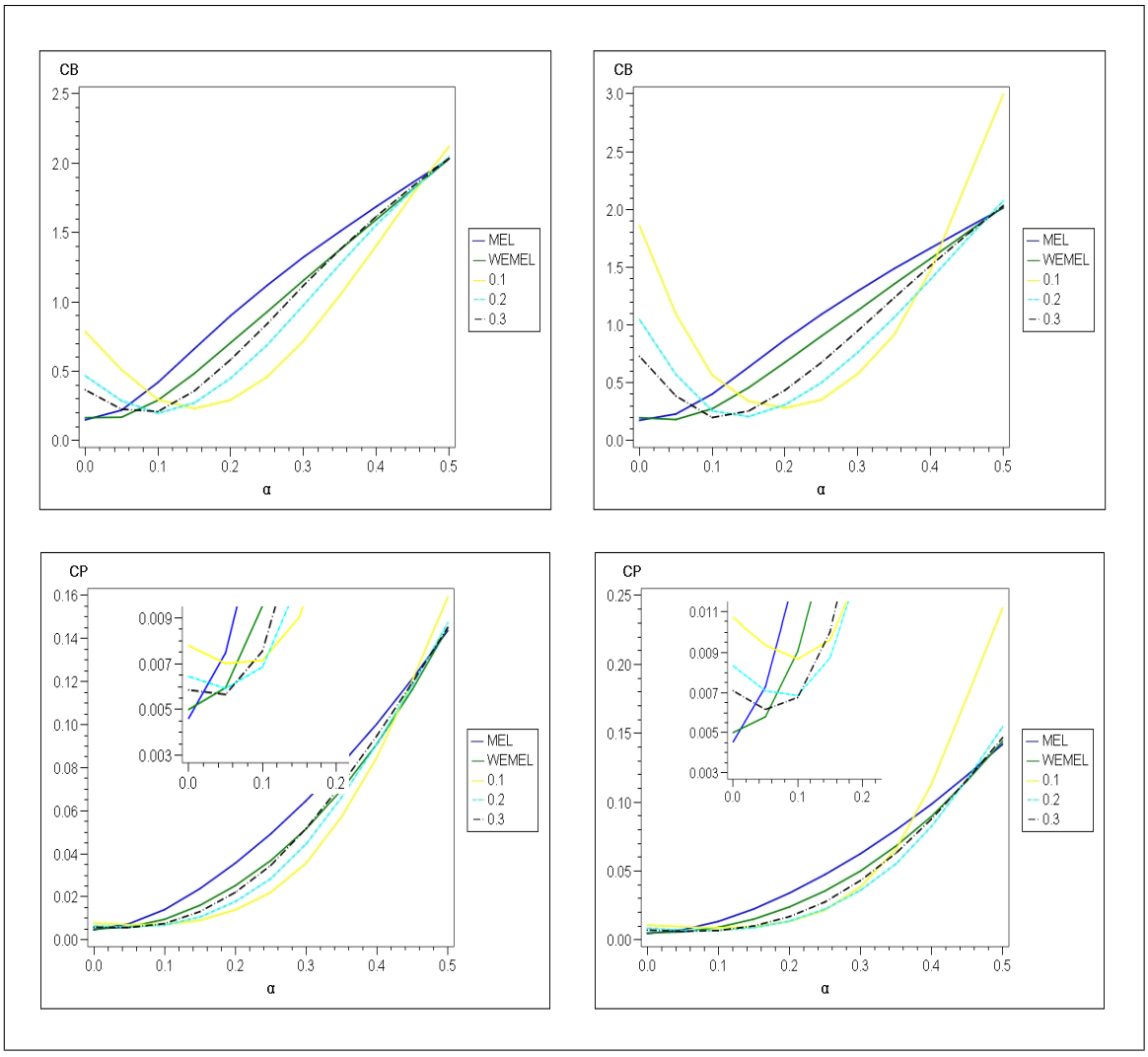


Figure 3.9: CB and CP values for $\epsilon = \{0.1, 0.2, 0.3\}$ and $c = 0.05$(left), $c = 0.1$(right)

for case 3

are similar in all simulations.

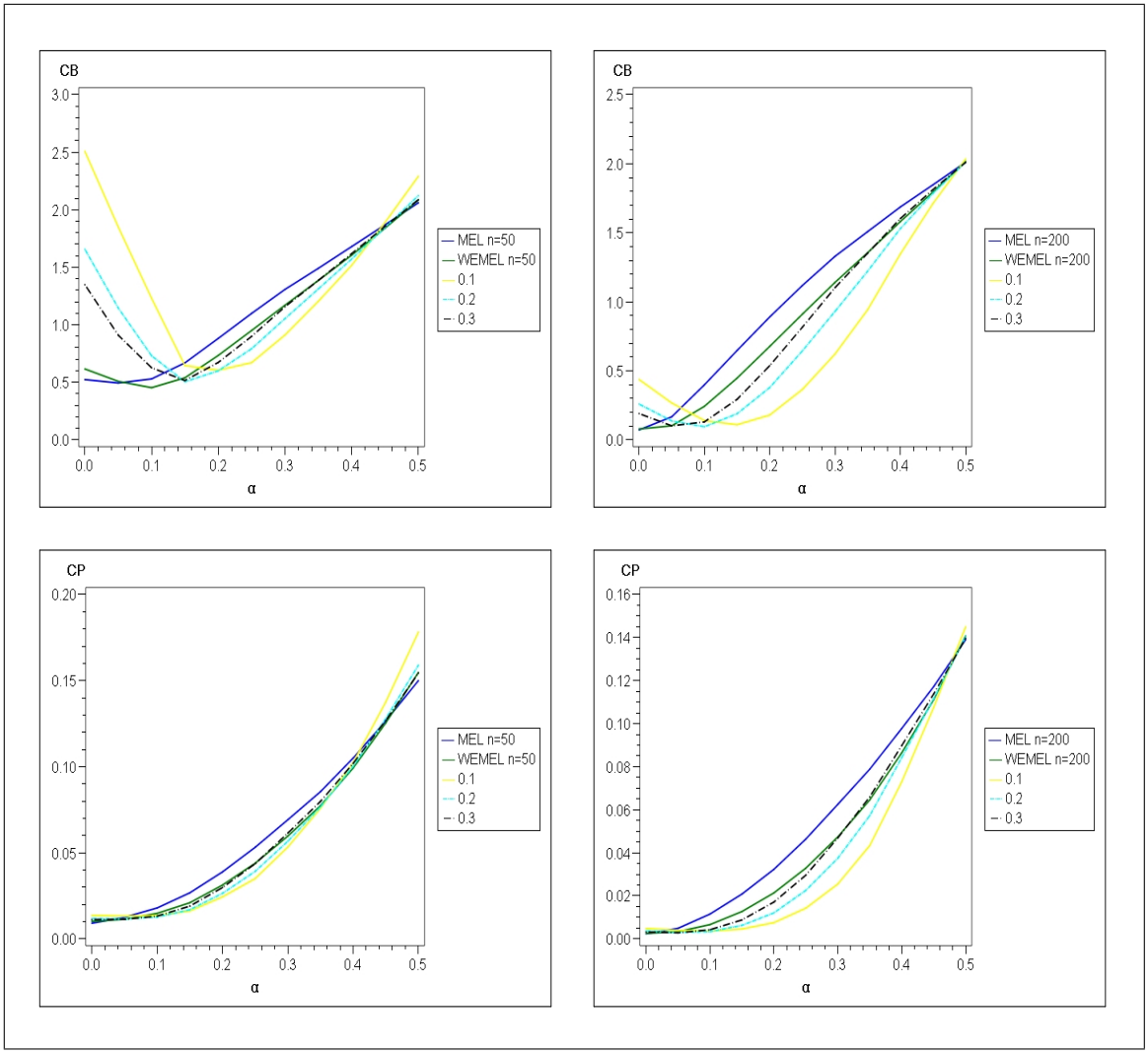


Figure 3.10: CB and CP values for $n = 50$ on left and $n = 200$ on right for $\epsilon = \{0.1, 0.2, 0.3\}$ and $c = 0.05$

3.4 Examples

In this section we apply the DOUW procedure to a credit scoring dataset as well as to a number of benchmark datasets. Note that in Chapter 4 we will apply the techniques derived in Chapters 2 and 3 on another practical credit scoring dataset.

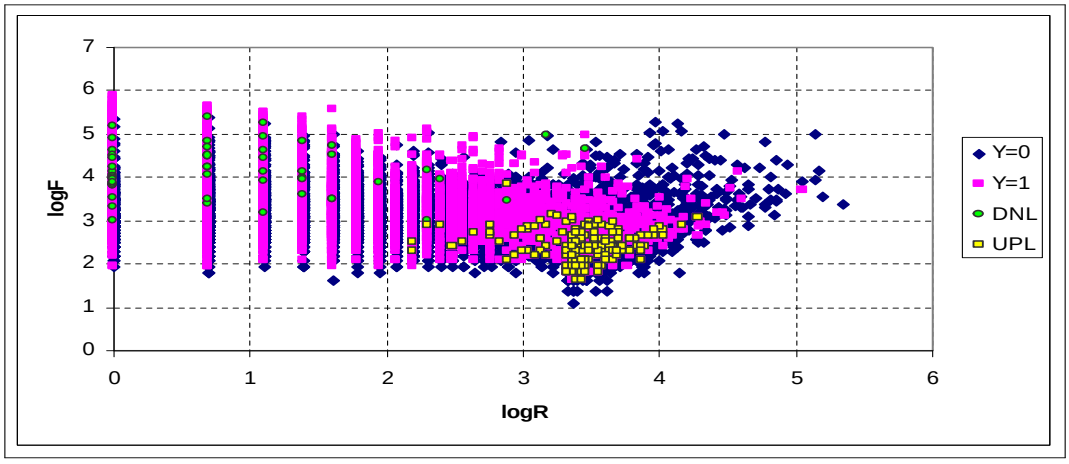
The first example relates to what is called Recency, Frequency and Monetary Analysis (RFM) in Customer Relations Management (CRM). As mentioned in Chapter 1, credit scoring refers to a wide field and CRM is often, but not always, included in credit scoring. For a detailed discussion of RFM analysis, see Nova (2000). Our data is a random sample of 10000 customers from a promotional campaign of a large retailer. The retailer would like to stay anonymous and therefore no reference is given. However, the data used in the analysis can be obtained from the author. It is known that typically customers with low R, high F and high M values are more likely to respond (actually take up the offers of the promotion). The R, F and M values for each customer before the campaign starts were recorded and the response indicators Y were observed when the campaign ended. LR may be used to predict the probability of response in terms of R, F and M and then the results used to save costs in future campaigns by not sending promotional offers to customers that are unlikely to respond. Outlier analyses are also important in this context. Here an uplier is a customer who did respond although being unlikely to do so. In the interest of good CRM practice such a customer should be "welcomed back into the fold" by a suitable thank you note or even an additional offer. A downlier is a customer who did not respond although being highly likely to do so. Such a customer may be in the process of defection and again good CRM practice requires that an effort be made to find out why and try to retain the customer. Clearly the analysis should not only provide LR model estimates but also a list of the up- and downliers so that appropriate action can be taken.

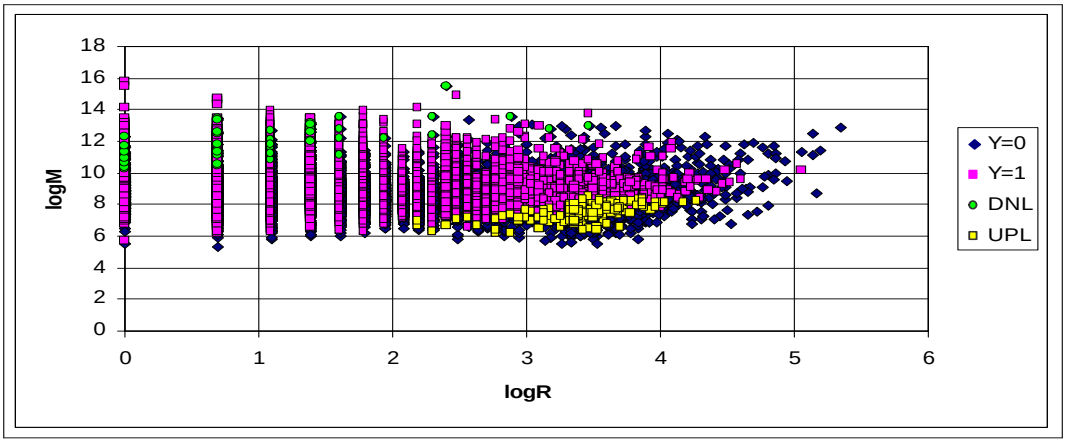
Typically the variable R is the number of days since the last transaction by the customer, F is the number of transactions over a previous period (e.g. one year) and M is the total sales to the customer over that period. The values of these variables

can vary widely and to make their ranges more reasonable we use their logs as the regressors. Table 3.3 shows the estimates of the β's and the number of outliers found for the different procedures. We took $\epsilon = 0.2$ throughout and varied c from 0.01 to 0.1. The estimated β's are quite similar for small values of c. However, as c increases the number of identified outliers increased substantially and the estimates of β's increased in absolute value. The reason why this happened is the tendency to separation becomes more pronounced as more outliers are removed from the dataset (details not shown here). In a CRM context it is prudent to use a rather high cut-off value c to identify the outliers. The 224 customers identified corresponding to the last line in Table 3.3 can be listed easily. It is instructive to see where they are located in the RFM space. The three panels in Figure 3.11 show their positions in the three

Method	β_0	β_1 (log R)	β_2 (log F)	β_3 (log M)	**#O**
ML	-4.36384	-0.33743	0.34672	0.41973	
MEL	-4.30700	-0.33322	0.34195	0.41396	
WEMEL	-4.73874	-0.29085	0.34567	0.45498	
DOUW$_{c=0.01,\epsilon=0.2}$	-4.52931	-0.34218	0.35075	0.43802	0
DOUW$_{c=0.05,\epsilon=0.2}$	-5.73153	-0.46132	0.45818	0.55445	4
DOUW$_{c=0.06,\epsilon=0.2}$	-5.93786	-0.48374	0.48041	0.57283	14
DOUW$_{c=0.07,\epsilon=0.2}$	-6.17061	-0.50961	0.49989	0.59641	36
DOUW$_{c=0.08,\epsilon=0.2}$	-6.41890	-0.53143	0.51864	0.62036	79
DOUW$_{c=0.09,\epsilon=0.2}$	-6.62722	-0.55139	0.53249	0.64092	132
DOUW$_{c=0.10,\epsilon=0.2}$	-6.86272	-0.57202	0.54757	0.66476	224

Table 3.3: RFM (N=10000,K=4)





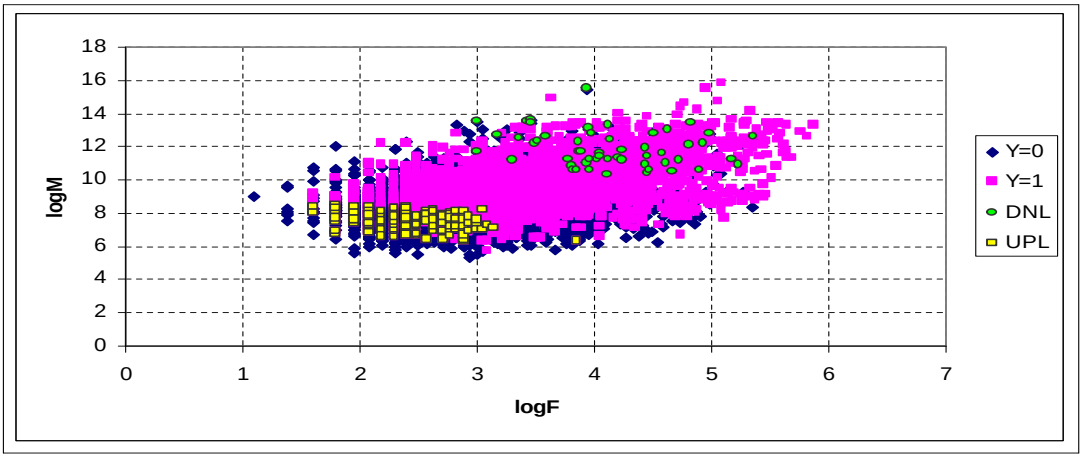


Figure 3.11: Scatterplot of the RFM dataset

pairs of two dimensional spaces of the regressors. Clearly there is large overlap between the responders and non-responders with the responders tending to be more prevalent in the direction with decreasing R values, increasing F values and higher M values. The downliers are visible in the low R, high F, high M directions, while the upliers are visible in the higher R, lower F and lower M directions. The graphs shown above indicate 5732 non-responders and 4258 responders. The labels of the responders are superimposed on that of the non-responders, thereby concealing the labels of the non-responders. Therefore, although it appears that there are far more responders than non-responders, this is not the case.

Secondly, we apply the DOUW procedure to a number of benchmark datasets. The benchmark datasets are:

- banknotes (Rousseeuw and Christmann, 2003),
- toxoplasmosis (Efron, 1986),
- vaso constriction (Finney, 1947; Pregibon, 1981) and
- food stamp (Kunsch et al. 1989).

These benchmark datasets were also used for illustration purposes by Rousseeuw and Christmann (2003).

Tables 3.4 - 3.7 present the results for the benchmark datasets and the column labelled #O indicates the number of outliers found. In these examples we used the MEL estimate in the DOUW method. The parameter estimates for ML, MEL and WEMEL are the same as those of Rousseeuw and Christmann (2003), barring the slight difference in the case of the WEMEL method which may be due to our using different software.

The **banknotes** dataset in Table 3.4 has no overlap and therefore the MLE does not exist. The DOUW procedure found no outliers and therefore has the same estimated β's as the MEL procedure, differing from the WEMEL procedure due to the radically different weighting scheme used by the latter.

Method	β_0	β_1	β_2	β_3	β_4	β_5	β_6	**#O**
ML	- does not exist -							
MEL	147.09	0.4649	−1.0204	1.3316	2.2049	2.3218	−2.3703	
WEMEL	252.55	−0.2541	−1.5791	2.0337	2.1012	2.3706	−2.1496	
DOUW$_{c=0.01,\epsilon=0.1}$	147.09	0.4649	−1.0204	1.3316	2.2049	2.3218	−2.3703	0
DOUW$_{c=0.05,\epsilon=0.2}$	147.09	0.4649	−1.0204	1.3316	2.2049	2.3218	−2.3703	0
DOUW$_{c=0.10,\epsilon=0.3}$	147.09	0.4649	−1.0204	1.3316	2.2049	2.3218	−2.3703	0

Table 3.4: Banknotes (N=200, K=7)

Method	β_0	β_1	β_2	β_3	**#O**
ML	0.09939	−0.44846	−0.18727	0.21342	
MEL	0.09882	−0.44395	−0.18536	0.21126	
WEMEL	0.09932	−0.40999	−0.16756	0.20259	
DOUW$_{c=0.01,\epsilon=0.1}$	0.13460	−0.46283	−0.27584	0.25642	0
DOUW$_{c=0.05,\epsilon=0.2}$	0.13002	−0.46045	−0.26424	0.25064	0
DOUW$_{c=0.10,\epsilon=0.3}$	0.12562	−0.45814	−0.25308	0.24508	0

Table 3.5: Toxoplasmosis (N=694, K=4)

In Table 3.5 we show the results for the **toxoplasmosis** dataset. Again no observations are identified as outliers when using the DOUW procedure and the estimated β's

Method	β_0	β_1	β_2	**#O**
ML	-2.92382	5.2205	4.6312	
MEL	-2.76789	4.9844	4.4064	
WEMEL	-2.73954	4.9487	4.3641	
DOUW$_{c=0.01,\epsilon=0.1}$	-2.76789	4.9844	4.4064	0
DOUW$_{c=0.05,\epsilon=0.2}$	-4.12743	6.8738	6.0565	2
DOUW$_{c=0.10,\epsilon=0.3}$	-6.11277	9.6801	8.5351	2

Table 3.6: Vaso constriction (N=39, K=3)

Method	β_0	β_1	β_2	β_3	**#O**
ML	0.92638	-1.85021	0.89606	-0.33275	
MEL	0.89360	-1.82665	0.88498	-0.32772	
WEMEL	5.37607	-1.75504	0.61952	-1.06607	
DOUW$_{c=0.01,\epsilon=0.1}$	1.21335	-2.14949	1.06178	-0.39777	0
DOUW$_{c=0.05,\epsilon=0.2}$	0.93637	-2.31400	1.13623	-0.35559	3
DOUW$_{c=0.10,\epsilon=0.3}$	0.51745	-3.00769	0.75962	-0.25222	6

Table 3.7: Food stamp (N=150, K=4)

are quite similar for all the procedures. In Table 3.6 we have the results for the **vaso constriction** dataset which was extensively used in the literature often reporting observations 4 and 18 as outliers. With the small cut-off $c = 0.01$ the DOUW procedure reports no outliers but with $c >= 0.05$ observations 4 and 18 are flagged as outliers here also. Note that the estimates of β's for these choices are substantially different from the estimates found by the ML, MEL and WEMEL and DOUW (with $c = 0.01$) procedures. Again this is a reflection of the substantial influence that outliers have on the estimated parameters. In Table 3.7 we show the results of the **foodstamp** data. Again the small choice of $c = 0.01$ identifies no outliers but either 3 (observations 66, 137 and 147) or 6 outliers (the former plus observations 22, 103 and 120) are declared when $c = 0.05$ and 0.10 respectively. Pregibon (1981) developed logistic regression diagnostic plots as a tool to identify outliers. One of these namely the deviance residual plot can be stated as follows: Define $Dev_n = -\sqrt{-D_n(\widehat{\boldsymbol{\beta}})}$ if $y_n = 0$ and $Dev_n = \sqrt{-D_n(\widehat{\boldsymbol{\beta}})}$ if $y_n = 1$. In Figure 3.12 we plot Dev_n against the observation number n.

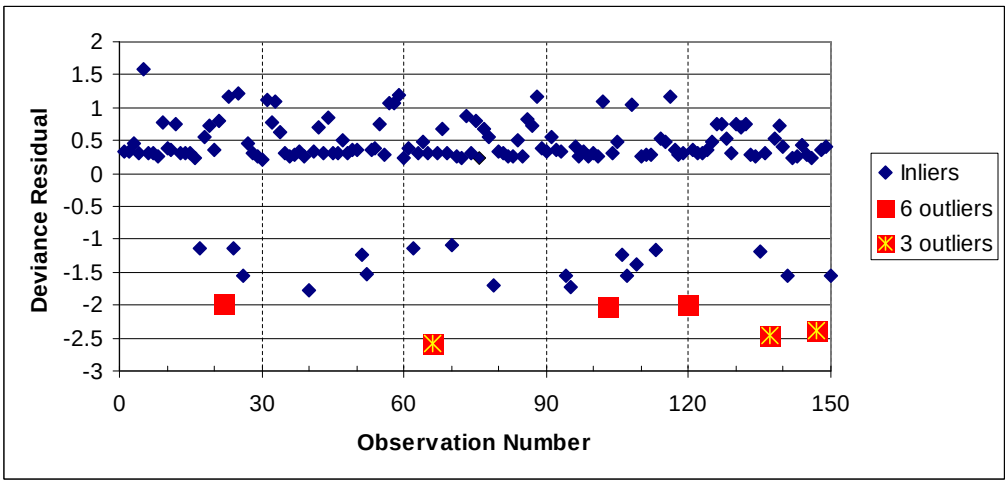


Figure 3.12: Deviance residual diagnostic plot of the foodstamp data

The downliers are represented by the extreme lower observations and the upliers by the extreme upper observations in this plot. Again observations 66, 137 and 147 show up as downliers when the cut-off level is about -2.2 and in addition also observations 22, 103 and 120 when the cut-off level is about -1.9. It is noteworthy that the estimates of the β_i's produced by the different estimators are substantially different for this dataset.

It is interesting to note that in Table 3.4 (Banknotes example) the DOUW procedure finds no outliers and the estimates from this procedure are identical over the three sets of tuning parameters used and are also identical with those from the MEL procedure. In contrast, in Table 3.5 (Toxoplasmosis example), the DOUW procedure also finds no outliers, but the estimates vary over the different sets of tuning parameters and also differ from the MEL estimates. We would expect the same estimates in Table 3.5. The reason is that the initial parameters differ in each of these cases. For the ML and the MEL procedure, we used the initial estimates as discussed in Hastie et al. (2001). For the DOUW procedure, we used the final estimates out of the first phase as initial estimates for the second phase. For future research we will investigate the stability of using the initial estimates as discussed in Hastie et al. (2001).

3.5 Summary and conclusion

Logistic regression (LR) is frequently used in the development of credit scoring models and is concerned with predicting a binary variable. Outliers in logistic regression data can be flagged by so-called deviance diagnostic (or similar) analysis. Once we have the ML estimate $\widehat{\beta}$ of the regression coefficients, we calculate the deviances of the observations and classify those with the most extreme negative deviances as the

downliers and those with the most extreme positive deviances as the upliers, using some cut-off level to express the extent of outlyingness. The problem with this approach is that the outliers (if any) were included in the observations on which the ML estimate $\widehat{\beta}$ was based to begin with and this inclusion may seriously effect the results. Among other consequences, it leaves the procedure vulnerable to the well-known dangers of masking and swamping in outlier identification. To guard against these risks one must use a procedure that does parameter estimation and outlier detection simultaneously. Trimming approaches have been used successfully for this purpose in a number of areas in statistics but in logistic regression trimming runs into the separation problem which makes it difficult to apply. In this chapter we presented an approach based on associating high and low weights with the observations in an optimal way as a result of the likelihood maximisation. This device enables the identification of the outliers as those points that are assigned the low weights. The required maximisation is handled by a search method based on repeated random starting subsets to which the high weights are assigned, followed by C-step improvements similar to that used in ordinary LTS (least trimmed squares) regression. We refer to the method as the DOUW method and its properties depend on two tuning constants, namely the ratio of the small to the large weights and the probability cut-off level used to measure outlyingness. We present a simulation study to show the effects of these constants on the performance of the DOUW procedure and illustrate the results in terms of four benchmark datasets as well as a large new dataset from the application area of retail marketing campaign analysis.

CHAPTER 4

Analysis of a credit scoring dataset

In this chapter we will apply the techniques developed in this thesis on a practical credit scoring dataset. We will first fit a logistic regression model and then, using the q-function, study the nature of the statistically significant classifiers. Then we will use the DOUW method to identify outliers and we will compare the performance of the DOUW fit with that of the logistic regression fit. As performance measures we will use the mean squared error (MSE), the so-called c-statistic and the best accuracy rate at the optimal threshold. After discarding the outliers we will again fit a logistic regression model and, using the q-function, again study the nature of the classifiers. We will expect the classifiers' performance in discriminating between goods and bads to improve. Our main objective is to illustrate that the application of the DOUW method will yield a better logistic regression fit and better classification performance and we want to show that the q-function may be used to study the classification performance of classifiers.

The layout of the chapter is as follows. In the next section the performance measures are defined. Then, in Section 4.2, we will firstly analyse an artificially generated dataset (Case1 in Chapter 1). This is necessary to prepare the reader for the discussion of the analysis of the practical credit scoring dataset that will follow in Section 4.3. Some concluding remarks will be made in Section 4.4. We conclude the thesis with ideas for future research.

4.1 Performance measures

To compare the logistic regression fits, we use three performance measures, the c-statistic, the mean squared error and the best accuracy rate at the optimal threshold. The c-statistic measures the area under the *Sensitivity vs. (1- Specificity)* curve for

the entire score range (Siddiqi, 2006). The sensitivity is equal to the total actual positives divided by the total actual positives and the specificity is equal to the total actual negatives divided by the total actual negatives where the total actual positives are those customers that are actually good customers and are predicted as being good customers and the total actual negatives are those customers that are actually bad customers and are correctly predicted as being bad customers. The mean squared error (MSE) is given by the following equation

$$MSE = \frac{1}{N}\sum_{n=1}^{N}(y_n - \widehat{p}_n)^2$$

where y_n is the target variable indicating whether a customer is a good customer ($y_n = 0$) or a bad customer ($y_n = 1$) and $\widehat{p}_n$ is the estimated probability of being a bad customer, $P(y_n = 1)$.

Another performance measure similar to the c-statistic is the best accuracy rate obtained. The accuracy rate is equal to the total actual positives plus the total actual negatives divided by the total number of customers. The best accuracy rate is then determined by calculating the accuracy rate at all possible thresholds and then choosing the highest accuracy rate. Let us refer to the threshold at which the best accuracy rate is obtained as the 'optimal threshold'. The accuracy rate obtained at the optimal threshold may be used as a third performance measure and will be reported in some of the examples discussed.

4.2 Analysis of Case 1

We firstly analyse the artificial dataset (Case 1) which contains no 'real' outliers. Then we add outliers to the artificial dataset and repeat the analysis. Using a simple and

easy to understand dataset, the objective is to illustrate jointly the behaviour of the DOUW method and the q-function on a well behaved dataset and then on a dataset containing 'known' outliers. This will prepare the reader for the discussion pertaining to the real dataset in Section 4.3.

Recall that Case 1 assumes that $V \sim N(0,1)$ (goods) and $W \sim N(2,2^2)$ (bads) with X the underlying characteristic. A sample of 1000 observations were generated from each of these distributions (see the frequency distribution in Figure 4.1) and $\widehat{q}_{MOM}$ and associated 90% and 95% bootstrap confidence bands estimated (see Figure 4.2).

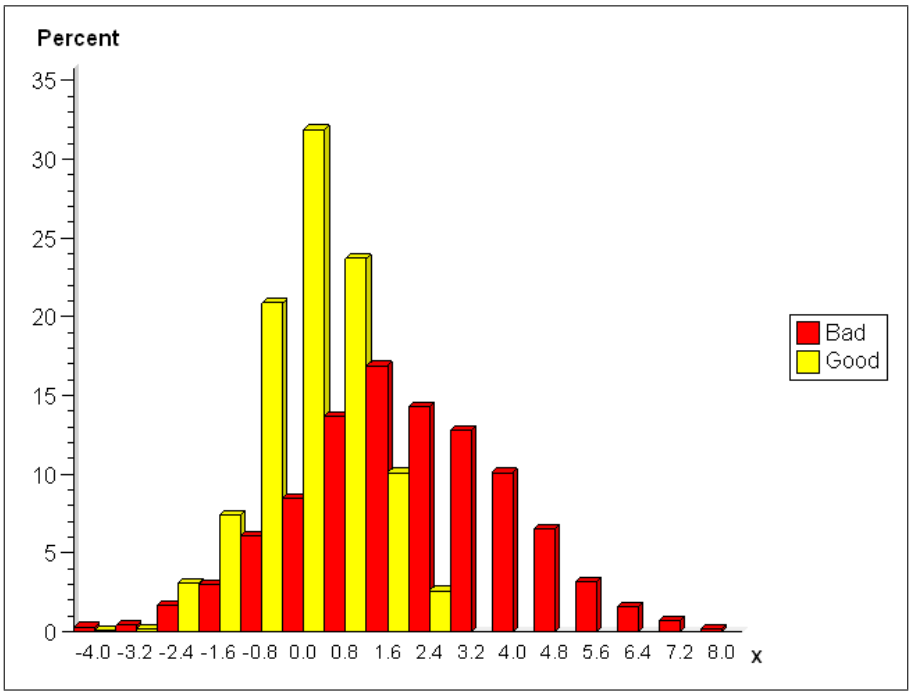


Figure 4.1: Frequency distribution of Case 1

Note that the equal distribution line is not contained in either of the confidence bands, so that it can be suggested that the classifier X distinguishes between the good and the bad risk classes. The plotted observations in the QQ plot seem to follow a linear trend which is expected because V and W are from the same translation-scale family.

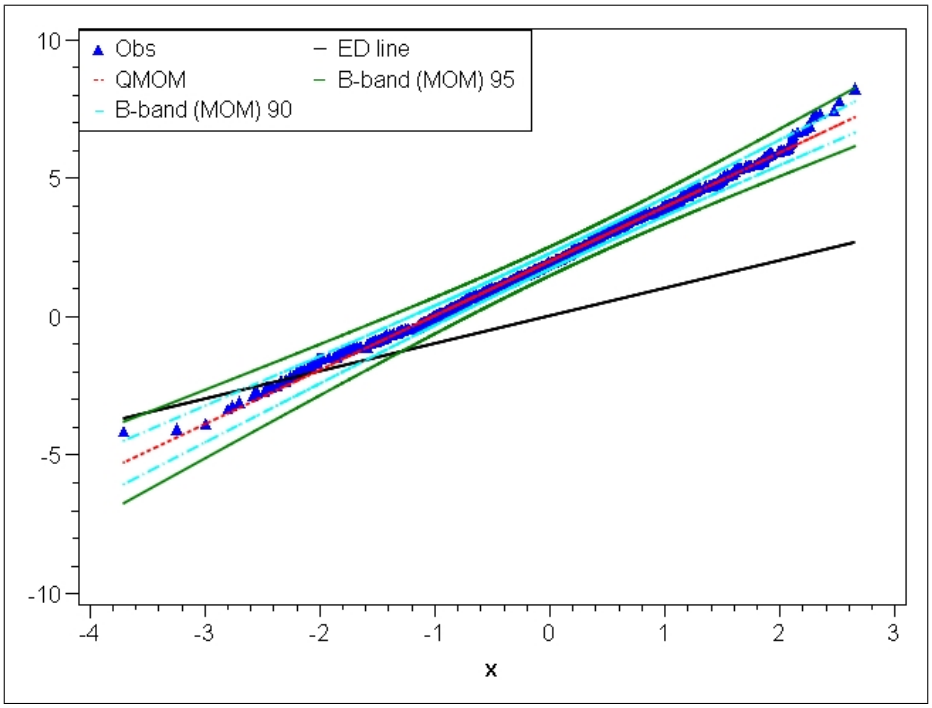


Figure 4.2: MOM (with B-bands), Case 1

The logistic regression maximum likelihood (ML) fit yields a c-statistic of 0.814 and a MSE of 0.0000711 while the DOUW fit (using $\epsilon = 0.2$ and $c = 0.05$) yields a c-statistic of 0.814 and a MSE of 0.0000664. From the above it is clear that the DOUW improves the MSE but not the c-statistic. This seems surprising that the c-statistics for the ML and the DOUW fits to the data are identical, whereas their MSE's differ. This may be explained by examining Figure 4.3 in which the ML and DOUW are depicted. The DOUW probability curve is slightly steeper than the ML curve resulting in a better MSE however the classification performance is similar resulting in the same c-statistic. This can be explained by the fact that the curves overlap at the same probability (approximately 0.5). Note that the best accuracy rate at the optimal threshold of the ML fit as well as the DOUW fit was 77.4% and the optimal thresholds were similar (in the region of 0.5).

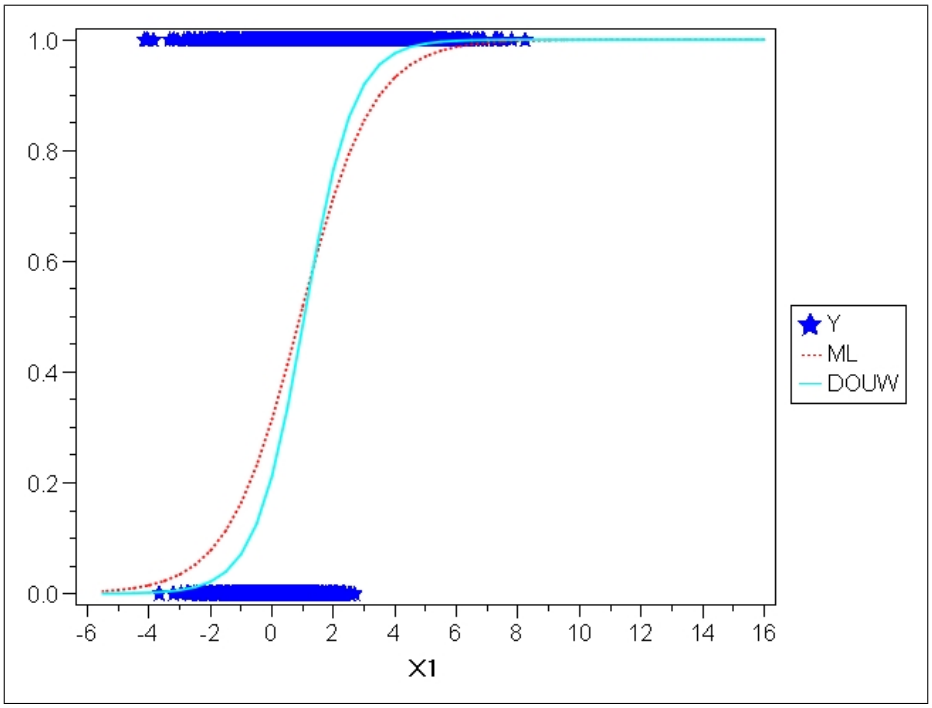


Figure 4.3: Estimated probability curves of ML and DOUW, Case 1

The DOUW fit applied to the artificial dataset identified 48 outliers. These outliers are now excluded from the dataset and the logistic regression fitted on the reduced dataset. This yields a c-statistic of 0.854 and a MSE of 0.0000662 which is a slight improvement on the fit on the full dataset.

However, the QQ plot (see Figure 4.4) now shows a non-linear relationship between the distributions of the goods and the bads. The exclusion of the outliers distorted the distributions to such an extent that the goods and the bads can no longer be regarded to come from the same translation-scale family. This is especially clear in the tails of the distributions.

Obviously in this dataset there are no "real" outliers. In order to examine the effect of real outliers we will now introduce outliers to the full dataset. We add 200 random uniform variables between -4 and 8 (see Figure 4.5) and selected 100 randomly as

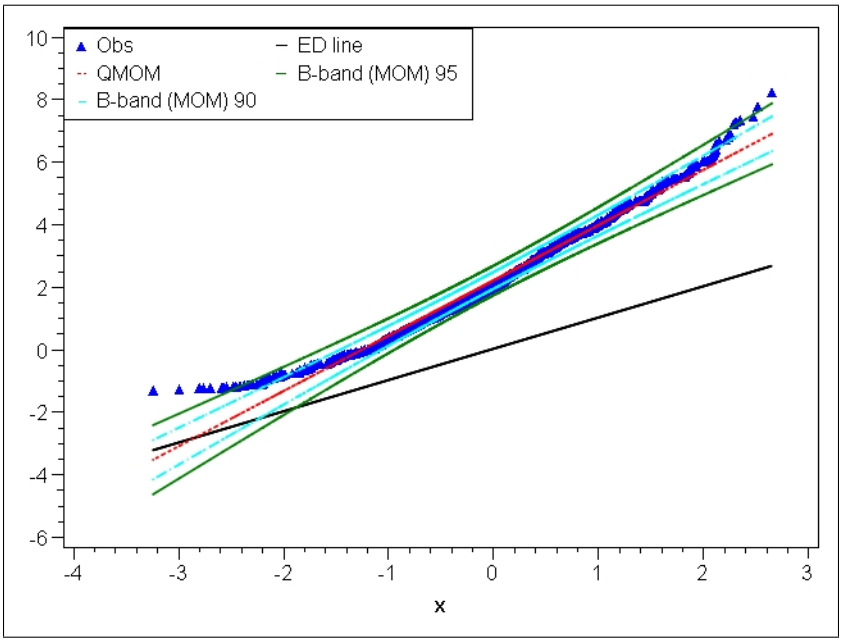


Figure 4.4: MOM (with B-bands) Case 1, excluding outliers

good customers and 100 as bad customers. Note that the 200 additional observations may not all be regarded as 'true' outliers, because some of the 'good' ('bad') outliers may by chance lie in the bulk of the data from the 'good' ('bad') distribution. The frequency distribution is shown in Figure 4.5 and the effect of the inclusion of the outliers can clearly be seen, especially in the right tail of the distribution of the goods. The associated QQ plot with $\widehat{q}_{MOM}$ and the bootstrap confidence bands are depicted in Figure 4.6. It is clear that the introduction of the outliers shifted the QQ plot closer to the equal distribution line in the left and right tails of the distribution. These additional observations give the illusion that the distribution of the very low x values and the very high x-values are similar for the goods and the bads. The QQ plot is no longer linear suggesting that the goods and the bads are no longer from the same translation-scale family. This is expected due to the way in which the outliers were included. The ML fit obtains a c-statistic of 0.768 and a MSE of 0.000615, which are both worse than in

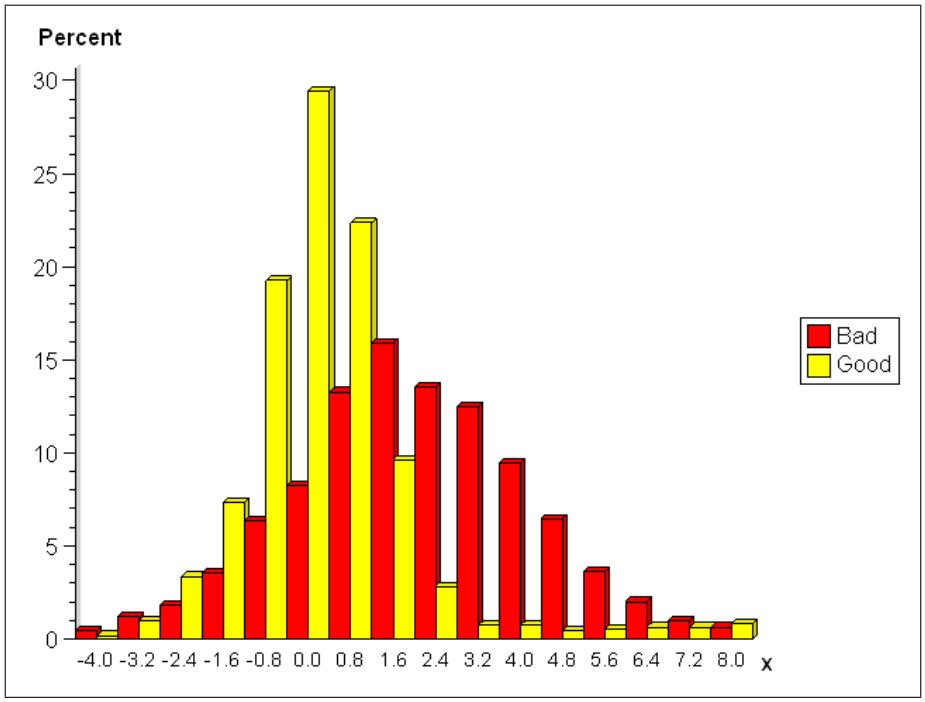


Figure 4.5: Frequency distribution of Case 1, with added noise

the corresponding fit on the original dataset.

The DOUW fit yields a c-statistic of 0.768 and a MSE of 0.000152 and identified 73 outliers. Of these, 22 are identified from the first 2000 observations and 51 are from the 200 additional observations. We depict the estimated probability curves of the ML fit as well as the DOUW fit in Figure 4.7. The DOUW fit is much steeper than the ML fit resulting in a much better MSE. The steepness of the DOUW fit is now more pronounced than in the previous examples, but again the classification performance is the same and can again be explained by the fact that the curves overlap at a probability of approximately 0.5, which is close to the optimal threshold.

Excluding the 73 outliers, the ML fit yields a c-statistic of 0.819 and a MSE of 0.0000524 which improves on both the previous fits. We construct the QQ plot with $\widehat{q}_{MOM}$ and associated confidence bands in Figure 4.8 on this reduced dataset. Although the QQ

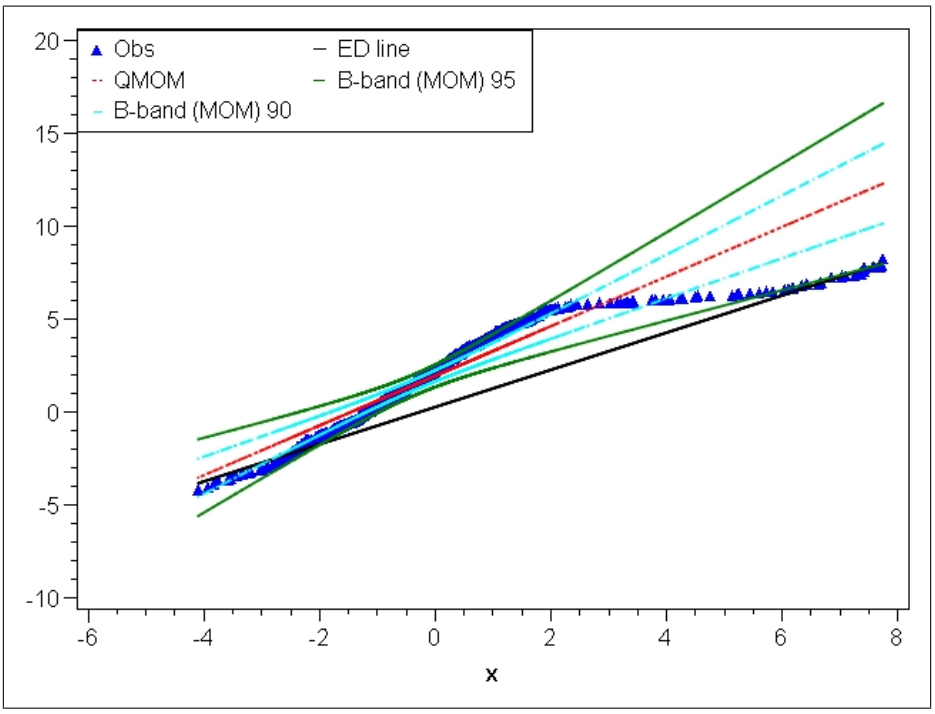


Figure 4.6: MOM (with B-bands) Case 1, with additional observations

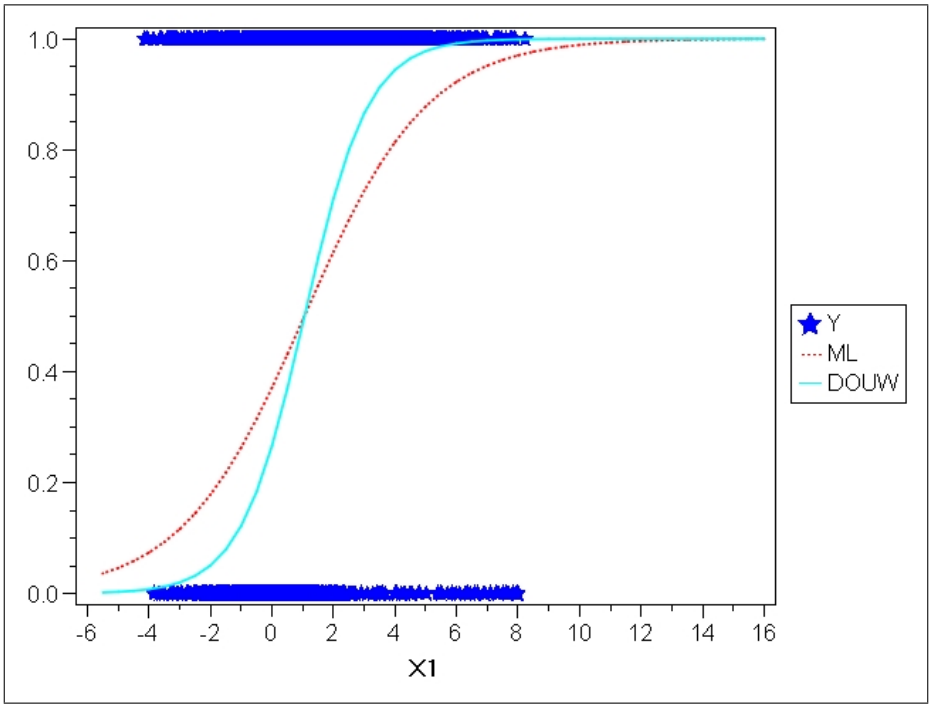


Figure 4.7: Estimated probability curves of ML and DOUW, Case 1, with additional observations

plot is still non-linear, it is much closer to linearity than the QQ plot in the full dataset.

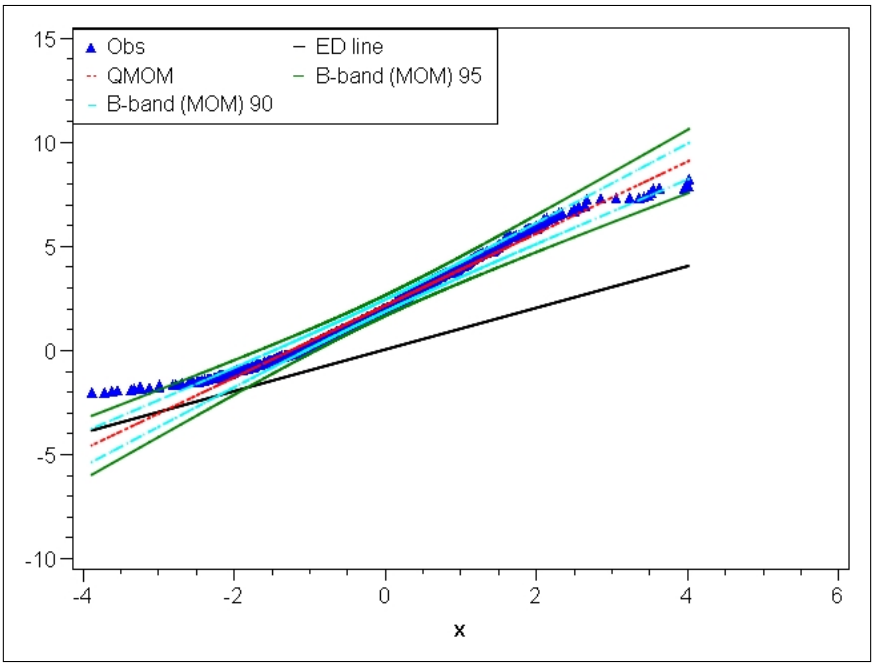


Figure 4.8: MOM (with B-bands) Case 1, with additional observations, but outliers excluded

This artificial example illustrated how we can use the techniques used in Chapter 2 and Chapter 3 in analysing a dataset.

4.3 Analysis of credit scoring dataset

Our next example is concerned with a practical credit scoring dataset and is an extract of the HMEQ.xls dataset from Wielenga, Lucas & Georges (1999). This dataset was extracted after doing validity and reliability checks on the data and has 4672 individuals in the good risk class and 1133 observations in the bad risk class. Initial analysis identified three classifiers as being statistically significant classifiers using forward stepwise logistic regression, namely $LOAN$ (loan amount requested), $MORTDUE$ (amount due to existing mortgage) and $DELINQ$ (number of delinquent trade lines).

In the analysis that follows, QQ plots will only be constructed for the two continuous variables, $LOAN$ and $MORTDUE$, since this analysis is not suitable for categorical variables.

As there is a high percentage of missing values in the dataset some method is needed to handle the missing values in the dataset: the standard median value imputation method was used in this example. Initial plots of the frequency diagrams showed that the distributions for the two continuous variables were skewed and therefore we log transformed (natural logarithm) the variables $LOAN$ and $MORTDUE$.

We plot the frequency distributions of $DELINQ$ and the natural logarithm of $LOAN$ and $MORTDUE$ in Figure 4.9. In Figure 4.10 we plot the QQ plot, $\widehat{q}_{MOM}$ and associated 90% and 95% bootstrap confidence bands for $LOAN$ (left) and $MORTDUE$ (right). When inspecting both panels in Figure 4.10 the plotted observations seem to follow approximately a straight line although some deviation is observed in the tails of the distributions. This suggests that the distributions of the good and the bad risk individuals could be from the same translation-scale family. In both QQ plots there are little deviation from the 45 degree line through the origin (ED line) and this suggests that the classifiers, $LOAN$ and $MORTDUE$, do not statistically distinguish between the goods and the bads. The ML fit yields a c-statistic of 0.707 and a MSE of 0.00000148. The DOUW fit (using $\epsilon = 0.2$ and $c = 0.05$) yields a c-statistic of 0.701 and a MSE of 0.00000073146 and identifies 86 outliers. The MSE improved substantially while the c-statistic marginally worsened. This latter reduction in the c-statistic is very small and could possibly be explained by random effects, as the best accuracy rate at the optimal threshold was 82.4% for both the ML and the DOUW fit. Note again that the best accuracy rate is determined at an optimal threshold, whereas the c-statistic

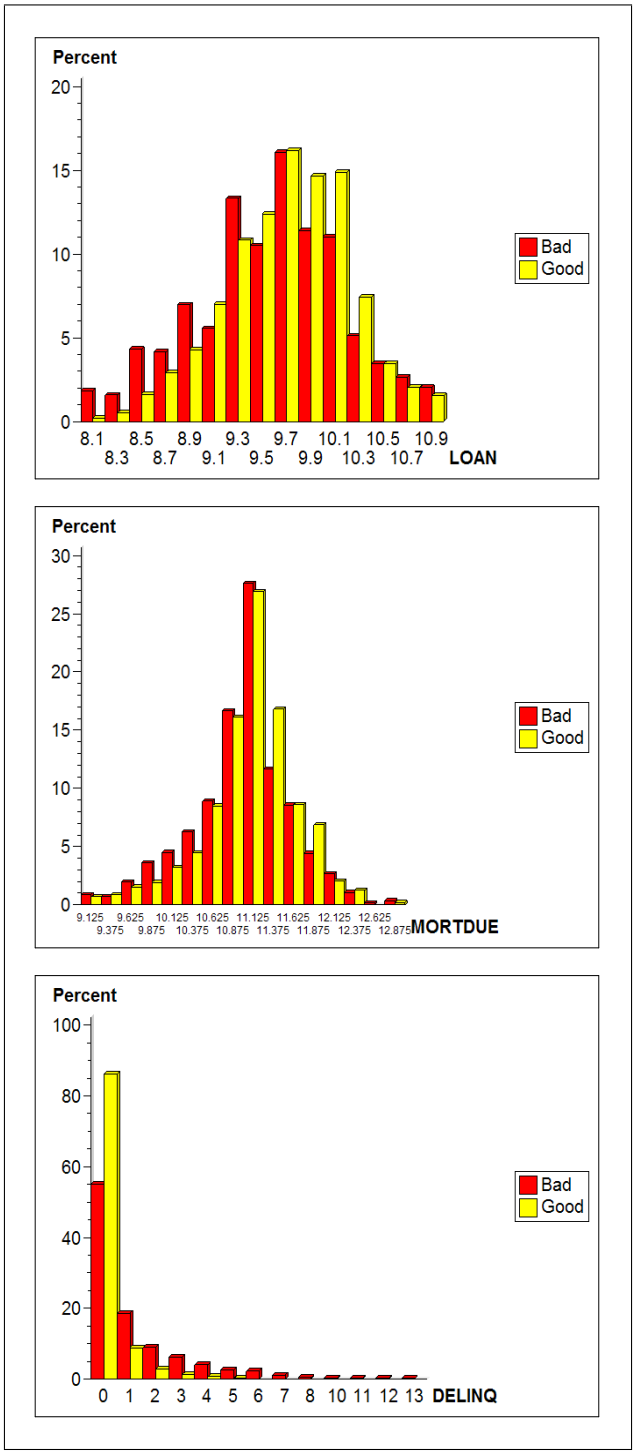


Figure 4.9: Frequency diagrams of $LOAN$ (top), $MORTDUE$ (middle) and $DELINQ$ (bottom)

measures performance over all possible thresholds. We are therefore less concerned about the slight reduction in the c-statistic, because the best accuracy rate is the same for both the ML and the DOUW fits.

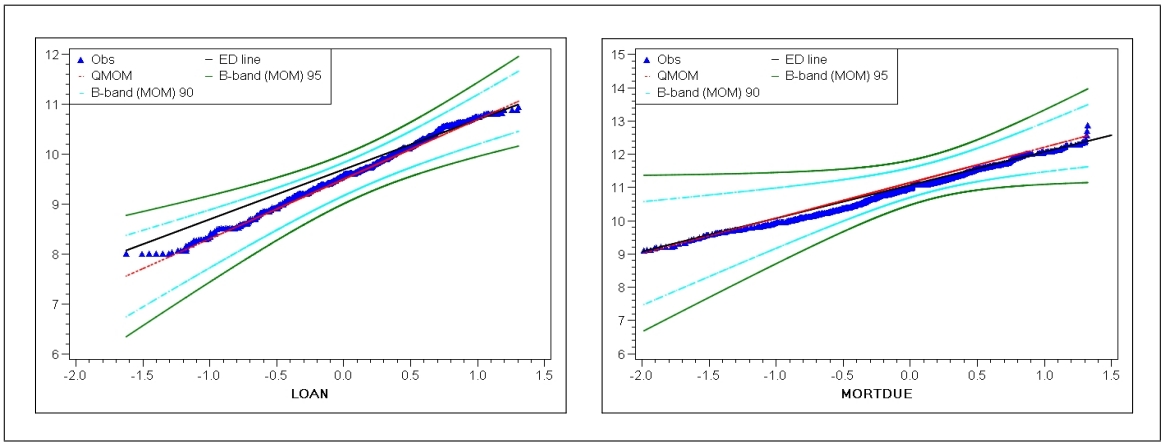


Figure 4.10: MOM (with B-bands) for $LOAN$ (left) and $MORTDUE$ (right)

Excluding these 86 outliers, the ML fit yields a c-statistic of 0.752 and a MSE of 0.0000005909. Both these values improved when compared to the original ML fit using all the data. Excluding these 86 outliers, we redraw the QQ plots in Figure 4.11 for $LOAN$ (left) and $MORTDUE$ (right).

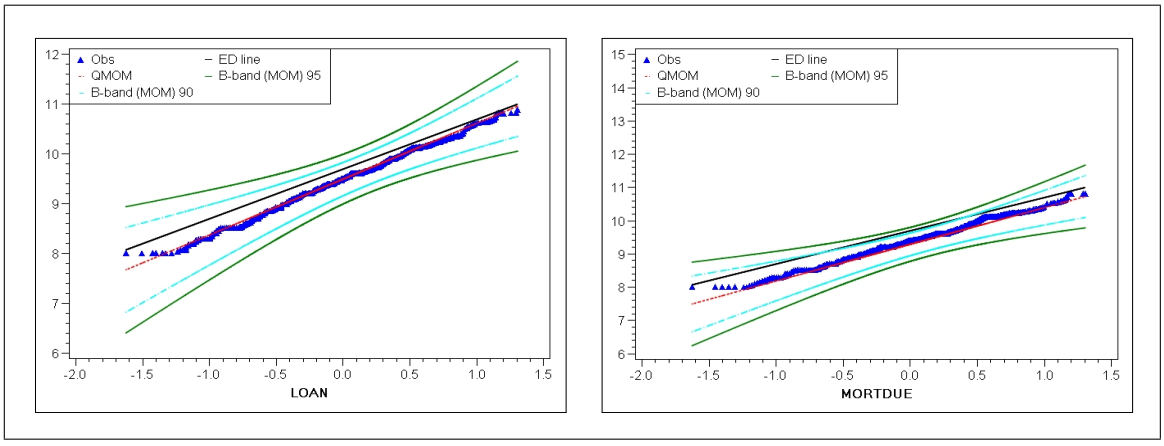


Figure 4.11: MOM (with B-bands) for $LOAN$ (left) and $MORTDUE$ (right) excluding 86 outliers

The plot for $LOAN$ is similar (compare with Figure 4.10), but for $MORTDUE$ we observe that the confidence bands are narrower and now the ED line is not contained in the 90% confidence bands, suggesting that the classifier, $MORTDUE$, distinguishes better between the good and the bad risk classes than in the original dataset.

As mentioned in Chapter 3, we could vary c from 0.01 to 0.1 and usually in bigger datasets, a value closer to 0.1 might be more appropriate. The reason for this is that for bigger datasets the probability of separation is low and even if you remove a large number of outliers from the dataset, you should not get into a break-down situation. Therefore, we increase the DOUW parameter, c, to 0.10, and this DOUW fit yields a c-statistic of 0.695 and a MSE of 0.00000003535 and identifies 512 outliers. Again the MSE improves substantially from the original ML fit, but the c-statistic marginally worsened. Note that the reduction in the c-statistic and the increase in best accuracy rate at the optimal threshold (82.5% from 82.4%) are very small.

Excluding these 512 outliers the ML fit yields a c-statistic of 0.919 and a MSE of 0.0000000025, which both improve substantially from the previous fits. In Figure 4.12 we depict the QQ plots again and estimate $\widehat{q}_{MOM}$ and the associated bootstrap confidence bands for $LOAN$ and $MORTDUE$, excluding the 512 outliers.

Clearly with the exclusion of the outliers, the ED lines are no longer contained within the confidence bands suggesting that the two classifiers, $LOAN$ and $MORTDUE$, now distinguish much better between the goods and the bads.

To complete this example, a QQ analysis was also performed on the three variables jointly ($LOAN$, $MORTDUE$ and $DELINQ$) using the regression function $\beta^T X$ (see Figure 4.13). The latter analysis revealed that the join QQ plot is different from the ED line, far more strongly so than with the individual variables.

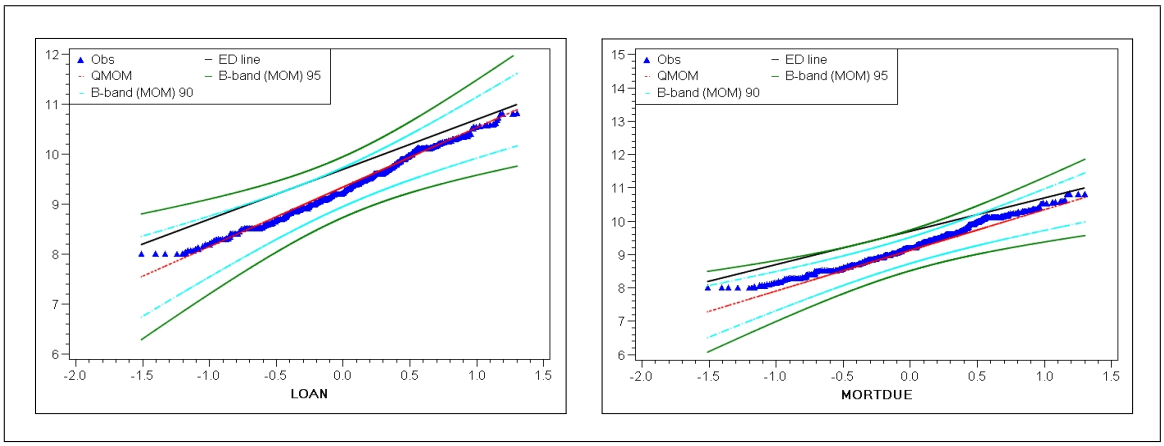


Figure 4.12: MOM (with B-bands) for $LOAN$ (left) and $MORTDUE$ (right) excluding 512 outliers

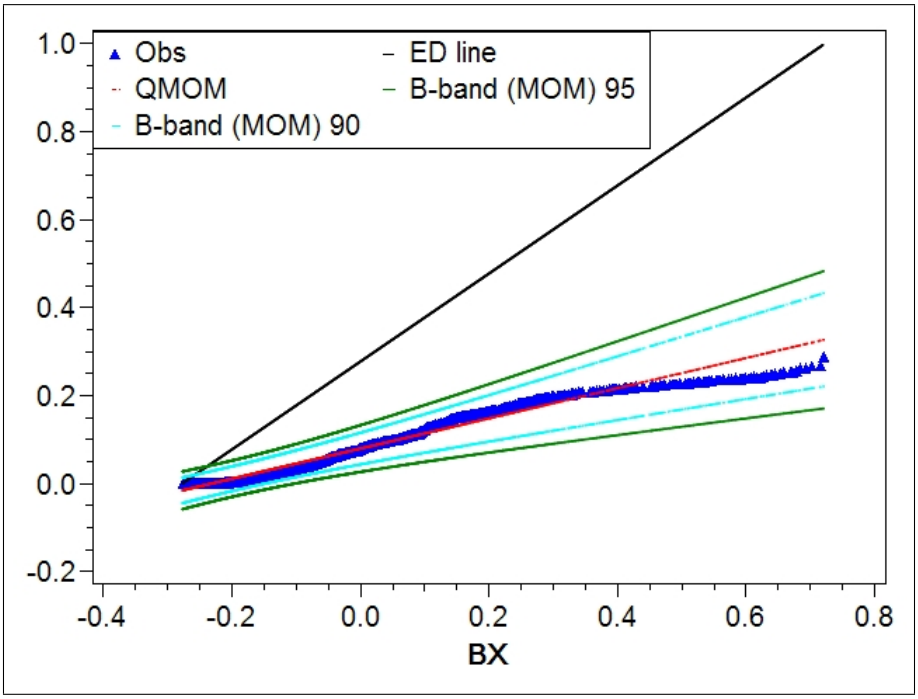


Figure 4.13: MOM (with B-bands) for $\beta^T X$ excluding 512 outliers

In this example we have shown that the DOUW method improves the classifier performance especially as measured by the MSE and from our analyses it is clear that q is useful in studying the nature of a classifier.

One might argue that a weakness in the analysis of the credit scoring datasets is that the evaluation, using the different performance measures (in particular the c-statistic and classification accuracy) has been carried out on the training datasets and not on holdout datasets, as is usual in credit scoring and other classification exercises. This is certainly true, and should be considered in a follow-up analyses. Here the purpose was to illustrate the use of the techniques.

4.4 Summary and conclusions

In this chapter we applied the techniques developed in this thesis on a practical credit scoring dataset. Firstly we analysed an artificially generated dataset (Case1 in Chapter 1) and secondly the practical credit scoring dataset. In both examples, we first fitted a logistic regression model and then, using the q-function, studied the nature of the significant classifiers. Then, we used DOUW to identify outliers and we compared the performance of the DOUW fit with that of the logistic regression fit. After discarding the outliers we again fitted a logistic regression model and, using the q-function, study the nature of the classifiers. The logistic regression fit and the classifiers' performance in discriminating between goods and bads improved after excluding the outliers identified by the DOUW methodology.

4.5 Ideas for future research

Ideas for future research were mentioned in previous chapters, and in this section, we summarise and expand on those ideas.

In Chapter 2, we made the simplifying assumption that $q(v)$ is linear, i.e. we assumed that V and W are from the same translation scale family. In some of the datasets we analysed, we noted that this is not always the case and that investigation of different shapes of $q(v)$ might prove fruitful. For example in Figure 1.3 (bottom right panel) and in Figure 4.4 the plotted observations exhibit a slightly non-linear pattern. Another example of a non-linear $q(v)$ appears in the QQ plot in Figure 4.6. Here one would postulate that the $q(v)$ function has a quadratic shape. The above-mentioned empirical evidence motivates the need for investigating different forms of $q(v)$ and for developing theory to test such assumptions.

Also, in Chapter 2, we conducted the majority of Monte Carlo studies by generating equal samples from the distributions of V and W. This is certainly an oversimplification and future research should additionally focus on considering unequal sample sizes. Comment from an external referee suggested that assessment of the accuracy of the Monte Carlo studies is required. This may be done by conducting the simulation study in blocks, and by calculating standard errors of answers obtained in each block.

In one set of these Monte Carlo studies we investigated whether the coverage probability of the asymptotic confidence band for q, based on the method of moments estimator, is close to the nominal coverage probability and found that this was true in large samples but not in small samples. Another referee suggested that a correction term for the asymptotic value in small samples should be determined. This can easily be done through a further Monte Carlo study (see e.g. de Jongh and de Wet, 1986).

In Chapter 3, we recommend some rule of thumb choices for ϵ and c. This is a judgment call based on limited experience and the issue of making sensible choices in practice is still open and could provide a fruitful area for further research. A possible approach is as follows:

- Start with a small c, increasing the value of c up until one outlier is identified (i.e. a downlier or uplier as explained in Section 1.3). The p-value of this outlier is then the c (or $1-c$) value where this observation became an outlier. We can now further increase c until we identify a second outlier, and this c (or $1-c$) is the p-value of the second outlier. With these newly defined p-values, the user can then decide how many outliers there are.

Another possible area for future research in Chapter 3 is to compare our methodology with other robust procedures for estimation of the logistic regression, for example the work of Bianco and Yohai (1996). One could compare and possibly combine these types of methodologies.

In the last chapter our purpose was to illustrate the use of the techniques developed in Chapters 2 and 3 and for this purpose we analysed a credit scoring dataset. Evaluating the performance of a classifier on the same data used to train the classifier usually leads to an optimistically biased assessment. The simplest strategy for correcting the optimistic bias is to holdout a portion of the development data for assessment (SAS, 2000). Therefore, for future research, we recommend partitioning this credit scoring dataset into training and holdout datasets.

As a final remark, credit scoring is by no means the only application area for the theoretical concepts that were developed in this thesis. As a case in point, the techniques

developed in Chapter 2 may be applied to the medical field (see Doksum, 1974 and Doksum and Sievers, 1976) as well as the coal industry (see Lombard, 2005). Similarly the techniques developed in Chapter 3 may be applied to wider application areas. Examples include banking (Rousseeuw and Christmann, 2003) and the medical field: toxoplasmosis (Efron, 1986) and vaso constriction (Finney, 1947; Pregibon, 1981).

APPENDIX A

Technical details of Chapter 2

This appendix contains the technical details of Chapter 2. In Appendix A.1 all the Theorems with associated proofs are given. In Appendix A.2 all the algorithms are given and in Appendix A.3 we have general items.

A.1 Theorems

A.1.1 Proof of Theorem 1

Lemma 1 *The asymptotic joint distribution of* $\overline{V}$, $\overline{W}$, s_V *and* s_W *is given by the following expression:*

$$\begin{bmatrix} \sqrt{m+n}(\overline{V}-\mu_V) \\ \sqrt{m+n}(\overline{W}-\mu_W) \\ \sqrt{m+n}(s_V-\sigma_V) \\ \sqrt{m+n}(s_W-\sigma_W) \end{bmatrix} \sim N_4\left(0,\Sigma^{**}\right) \tag{A.1}$$

where

$$\Sigma^{**} = \begin{bmatrix} \left(\frac{1}{\lambda}\right)\sigma_V^2 & 0 & \left(\frac{1}{\lambda}\right)\frac{\kappa_3(V)}{2\sigma_V} & 0 \\ 0 & \left(\frac{1}{1-\lambda}\right)\sigma_W^2 & 0 & \left(\frac{1}{1-\lambda}\right)\frac{\kappa_3(W)}{2\sigma_W} \\ \left(\frac{1}{\lambda}\right)\frac{\kappa_3(V)}{2\sigma_V} & 0 & \left(\frac{1}{\lambda}\right)\left(\frac{\kappa_4(V)}{4\sigma_V^2}+\frac{\sigma_V^2}{2}\right) & 0 \\ 0 & \left(\frac{1}{1-\lambda}\right)\frac{\kappa_3(W)}{2\sigma_W} & 0 & \left(\frac{1}{1-\lambda}\right)\left(\frac{\kappa_4(W)}{4\sigma_W^2}+\frac{\sigma_W^2}{2}\right) \end{bmatrix}, \tag{A.2}$$

with $\lambda = m/(m+n)$ *and* κ_i *indicating the* i^{th} *cumulant.*

Proof of Lemma 1. We know that

$$var(\sqrt{m}(\overline{V}-\mu_V)) = \sigma_V^2 \tag{A.3}$$

and

$$var(\sqrt{n}(\overline{W}-\mu_W)) = \sigma_W^2. \tag{A.4}$$

Note that we can assume, without loss of generality, that μ_V and μ_W are zero because we consider the distribution of $(\overline{V} - \mu_V)$ and $(\overline{W} - \mu_W)$.

V and W are independent, therefore:

$$\begin{aligned}
cov(\sqrt{m}(\overline{V} - \mu_V), \sqrt{n}(\overline{W} - \mu_W)) = cov(\sqrt{n}(\overline{W} - \mu_W), \sqrt{m}(\overline{V} - \mu_V)) &= 0 \\
cov(\sqrt{n}(\overline{W} - \mu_W), \sqrt{m}(s_V - \sigma_V)) = cov(\sqrt{m}(s_V - \sigma_V), \sqrt{n}(\overline{W} - \mu_W)) &= 0 \\
cov(\sqrt{n}(s_W - \sigma_W), \sqrt{m}(\overline{V} - \mu_V)) = cov(\sqrt{m}(\overline{V} - \mu_V), \sqrt{n}(s_W - \sigma_W)) &= 0 \\
cov(\sqrt{n}(s_W - \sigma_W), \sqrt{m}(s_V - \sigma_V)) = cov(\sqrt{m}(s_V - \sigma_V), \sqrt{n}(s_W - \sigma_W)) &= 0
\end{aligned} \tag{A.5}$$

and

$$\begin{aligned}
& cov(\sqrt{m}(\overline{V} - \mu_V), \sqrt{m}(s_V^2 - \sigma_V^2)) \\
= \; & \sqrt{m}\sqrt{m} cov\left(\overline{V}, \frac{1}{m}\sum_j (V_j - \overline{V})^2\right) \\
= \; & mcov\left(\overline{V}, \frac{1}{m}\sum_j V_j^2 - \overline{V})^2\right) \\
= \; & mcov\left(\overline{V}, \frac{1}{m}\sum_j V_j^2\right) + mcov\left(\overline{V}, \overline{V}^2\right) \\
= \; & A + B.
\end{aligned} \tag{A.6}$$

Now,

$$\begin{aligned}
A &= mcov\left(\overline{V}, \frac{1}{m}\sum_j V_j^2\right) \\
&= \kappa_3(V) + 2\kappa_1(V)\kappa_2(V) \\
&= \kappa_3(V)
\end{aligned} \tag{A.7}$$

because $\kappa_1(V) = 0$. Note that we use a method described by Brillinger (1965, p.19-

21) to write covariances in terms of cumulants throughout this appendix. Also,

$$\begin{aligned} B &= mcov\left(\overline{V}, \overline{V}^2\right) \\ &= \frac{\kappa_3(V)}{m} + o(1) \end{aligned} \tag{A.8}$$

as $m \to \infty$.

Thus,

$$cov(\sqrt{m}(\overline{V} - \mu_V), \sqrt{m}(s_V^2 - \sigma_V^2)) = \kappa_3(V) + o(1) \tag{A.9}$$

as $m \to \infty$. Similarly,

$$cov(\sqrt{n}(\overline{W} - \mu_W), \sqrt{n}(s_W^2 - \sigma_W^2)) = \kappa_3(W) + o(1) \tag{A.10}$$

as $n \to \infty$. Next,

$$var\left(\sqrt{m}(s_V^2 - \sigma_V^2)\right) \tag{A.11}$$

$$\begin{aligned} &= mcov(s_V^2, s_V^2) \\ &= mcov\left(\frac{1}{m}\sum_j (V_j^2 - \overline{V})^2, \frac{1}{m}\sum_j (V_j^2 - \overline{V})^2\right) \\ &= m\left[cov\left(\frac{1}{m}\sum_j V_j^2, \frac{1}{m}\sum_j V_j^2\right) - 2cov\left(\frac{1}{m}\sum_j V_j^2, \overline{V}^2\right) + cov\left(\overline{V}^2, \overline{V}^2\right)\right] \\ &= m\left[C + D + E\right] \end{aligned}$$

where C is given by

$$\begin{aligned} C &= cov\left(\frac{1}{m}\sum_j V_j^2, \frac{1}{m}\sum_j V_j^2\right) \\ &= \frac{1}{m}\left[\kappa_4(V) + 2\sigma_V^4\right], \end{aligned} \tag{A.12}$$

D is given by

$$\begin{aligned} D &= -2cov\left(\frac{1}{m}\sum_j V_j^2, \overline{V}^2\right) \\ &= \frac{-2}{m^2}\left[\kappa_4(V) + 2\sigma_V^4\right] \end{aligned} \tag{A.13}$$

and E is given by

$$\begin{aligned} E &= cov\left(\overline{V}^2, \overline{V}^2\right) \\ &= \frac{1}{m^3}\left[\kappa_4(V) + 6m\sigma_V^4\right]. \end{aligned} \tag{A.14}$$

Note that all these results, A - E, can also be found in Rao (1965, p.368).

Using (A.11), (A.12), (A.13) and (A.14), we therefore have

$$\begin{aligned} & var\left(\sqrt{m}(s_V^2 - \sigma_V^2)\right) \\ = \; & m\left(\frac{1}{m}\left[\kappa_4(V) + 2\sigma_V^4\right] + \frac{-2}{m^2}\left[\kappa_4(V) + 2\sigma_V^4\right]\frac{1}{m^3}\left[\kappa_4(V) + 6m\sigma_V^4\right]\right) \\ = \; & \left[\kappa_4(V) + 2\sigma_V^4\right] + o(1) \end{aligned} \tag{A.15}$$

as $m \to \infty$. Similarly,

$$var\left(\sqrt{n}(s_W^2 - \sigma_W^2)\right) = \left[\kappa_4(W) + 2\sigma_W^4\right] + o(1) \tag{A.16}$$

as $n \to \infty$. From the previous results together with the central limit theorem (CLT), we have, asymptotically

$$\begin{bmatrix} \sqrt{m}(\overline{V} - \mu_V) \\ \sqrt{m}(s_V^2 - \sigma_V^2) \end{bmatrix} \sim N_2\left(0, \begin{bmatrix} \sigma_V^2 & \kappa_3(V) \\ \kappa_3(V) & \kappa_4(V) + 2\sigma_V^4 \end{bmatrix}\right). \tag{A.17}$$

We require the distribution of

$$\begin{bmatrix} \sqrt{m}(\overline{V} - \mu_V) \\ \sqrt{m}(s_V - \sigma_V) \end{bmatrix}. \tag{A.18}$$

and to obtain this we use the delta method described in Ferguson (1996, p. 44-50). This method is also found in Rao (1965, p. 322). Define g by

$$\begin{bmatrix} \overline{V} \\ s_V \end{bmatrix} = g\left(\begin{bmatrix} \overline{V} \\ s_V^2 \end{bmatrix}\right), \tag{A.19}$$

so that

$$g\left(\begin{bmatrix} a_1 \\ a_2 \end{bmatrix}\right) = \begin{bmatrix} a_1 \\ \sqrt{a_2} \end{bmatrix}. \tag{A.20}$$

Then

$$g' = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{2\sqrt{a_2}} \end{bmatrix} \tag{A.21}$$

because

$$\frac{\partial g}{\partial a_1} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{A.22}$$

and

$$\frac{\partial g}{\partial a_2} = \begin{bmatrix} 0 \\ \frac{1}{2\sqrt{a_2}} \end{bmatrix}. \tag{A.23}$$

Using the delta method, we then have

$$\begin{bmatrix} \sqrt{m}(\overline{V} - \mu_V) \\ \sqrt{m}(s_V - \sigma_V) \end{bmatrix} \sim N_2\left(0, g' \begin{bmatrix} \sigma_V^2 & \kappa_3(V) \\ \kappa_3(V) & \kappa_4(V) + 2\sigma_V^4 \end{bmatrix} g'^{\top}\right). \tag{A.24}$$

Because V and W are independent, we have

$$\begin{bmatrix} \sqrt{m}(\overline{V} - \mu_V), & \sqrt{n}(\overline{W} - \mu_W), & \sqrt{m}(s_V - \sigma_V), & \sqrt{n}(s_W - \sigma_W) \end{bmatrix}^{\top} \sim N_4\left(0, \Sigma^*\right) \tag{A.25}$$

with $\top$ denoting transpose and where

$$\Sigma^* = \begin{bmatrix} \sigma_V^2 & 0 & \frac{\kappa_3(V)}{2\sigma_V} & 0 \\ 0 & \sigma_W^2 & 0 & \frac{\kappa_3(W)}{2\sigma_W} \\ \frac{\kappa_3(V)}{2\sigma_V} & 0 & \frac{\kappa_4(V)}{4\sigma_V^2} + \frac{\sigma_V^2}{2} & 0 \\ 0 & \frac{\kappa_3(W)}{2\sigma_W} & 0 & \frac{\kappa_4(W)}{4\sigma_W^2} + \frac{\sigma_W^2}{2} \end{bmatrix}. \tag{A.26}$$

Note that

$$\begin{bmatrix} \sqrt{m+n}(\overline{V}-\mu_V) \\ \sqrt{m+n}(\overline{W}-\mu_W) \\ \sqrt{m+n}(s_V-\sigma_V) \\ \sqrt{m+n}(s_W-\sigma_W) \end{bmatrix} = A \begin{bmatrix} \sqrt{m}(\overline{V}-\mu_V) \\ \sqrt{n}(\overline{W}-\mu_W) \\ \sqrt{m}(s_V-\sigma_V) \\ \sqrt{n}(s_W-\sigma_W) \end{bmatrix} \tag{A.27}$$

where

$$A = diag(\frac{\sqrt{m+n}}{\sqrt{m}}, \frac{\sqrt{m+n}}{\sqrt{n}}, \frac{\sqrt{m+n}}{\sqrt{m}}, \frac{\sqrt{m+n}}{\sqrt{n}}). \tag{A.28}$$

Therefore

$$\begin{bmatrix} \sqrt{m+n}(\overline{V}-\mu_V) \\ \sqrt{m+n}(\overline{W}-\mu_W) \\ \sqrt{m+n}(s_V-\sigma_V) \\ \sqrt{m+n}(s_W-\sigma_W) \end{bmatrix} \sim N_4\left(0, A\Sigma^* A\right) \tag{A.29}$$

where $A\Sigma^* A = \Sigma^{**}$ is given by

$$\Sigma^{**} = \begin{bmatrix} \frac{1}{\lambda}\sigma_V^2 & 0 & \frac{1}{\lambda}\frac{\kappa_3(V)}{2\sigma_V} & 0 \\ 0 & \frac{1}{1-\lambda}\sigma_W^2 & 0 & \frac{1}{1-\lambda}\frac{\kappa_3(W)}{2\sigma_W} \\ \frac{1}{\lambda}\frac{\kappa_3(V)}{2\sigma_V} & 0 & \frac{1}{\lambda}\left(\frac{\kappa_4(V)}{4\sigma_V^2}+\frac{\sigma_V^2}{2}\right) & 0 \\ 0 & \frac{1}{1-\lambda}\frac{\kappa_3(W)}{2\sigma_W} & 0 & \frac{1}{1-\lambda}\left(\frac{\kappa_4(W)}{4\sigma_W^2}+\frac{\sigma_W^2}{2}\right) \end{bmatrix}. \tag{A.30}$$

■

Lemma 2 *The joint distribution of $\overline{W}$, $\overline{V}$ and $\frac{s_W}{s_V}$ is given by the expression:*

$$\sqrt{m+n}\left[\ \overline{W}-\mu_W,\quad \overline{V}-\mu_V,\quad \frac{s_W}{s_V}-\frac{\sigma_W}{\sigma_V}\ \right]^\top \sim N_2(0,\Sigma^{***}) \tag{A.31}$$

where

$$\Sigma^{***} = \begin{bmatrix} \frac{\sigma_W^2}{(1-\lambda)} & 0 & \frac{\kappa_3(W)}{2(1-\lambda)\sigma_W\sigma_V} \\ 0 & \frac{\sigma_V^2}{\lambda} & -\frac{\kappa_3(V)\sigma_W}{2\lambda\sigma_V^3} \\ \frac{\kappa_3(W)}{2(1-\lambda)\sigma_W\sigma_V} & -\frac{\kappa_3(V)\sigma_W}{2\lambda\sigma_V^3} & \frac{\kappa_4(V)\sigma_W^2}{4\lambda\sigma_V^6}+\frac{\sigma_W^2}{2\lambda(1-\lambda)\sigma_V^2}+\frac{\kappa_4(W)}{4(1-\lambda)\sigma_W^2\sigma_V^2} \end{bmatrix}. \tag{A.32}$$

Proof of Lemma 2. Set

$$g = \left[\begin{array}{ccc} \mu_W & \mu_V & \frac{\sigma_W}{\sigma_V} \end{array} \right]^\top. \tag{A.33}$$

Then

$$\begin{aligned} \frac{\partial g}{\partial \mu_V} &= \left[\begin{array}{ccc} 0 & 1 & 0 \end{array} \right]^\top \\ \frac{\partial g}{\partial \mu_W} &= \left[\begin{array}{ccc} 1 & 0 & 0 \end{array} \right]^\top \\ \frac{\partial g}{\partial \sigma_V} &= \left[\begin{array}{ccc} 0 & 0 & -\frac{\sigma_W}{\sigma_V^2} \end{array} \right]^\top \\ \frac{\partial g}{\partial \sigma_W} &= \left[\begin{array}{ccc} 0 & 0 & \frac{1}{\sigma_V} \end{array} \right]^\top, \end{aligned} \tag{A.34}$$

so that

$$g' = \left[\begin{array}{cccc} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & -\frac{\sigma_W}{\sigma_V^2} & \frac{1}{\sigma_V} \end{array} \right]. \tag{A.35}$$

Now, algebra gives $g'\Sigma^{**}g'^\top = \Sigma^{***}$ and the result follows by the delta method. Σ^{**} is given in Lemma 1, (A.30). ■

Theorem 1 *The asymptotic distribution of $\widehat{q}_{MOM}$ is given by the expression*

$$\sqrt{m+n}(\widehat{q}_{MOM}(v) - q(v)) \sim N(0, \tau(v)^2) \tag{A.36}$$

where

$$\tau(v)^2 = \sigma_0^2 + \sigma_1^2 \widetilde{v}^2 + 2\sigma_{0,1}\widetilde{v}, \tag{A.37}$$

$$\widetilde{v} = v - \overline{V} \tag{A.38}$$

and

$$\sigma_0^2 = \frac{\sigma_W^2}{\lambda(1-\lambda)}, \tag{A.39}$$

$$\sigma_{0,1} = \frac{\kappa_3(V)\sigma_W^2}{2\lambda\sigma_V^4} + \frac{\kappa_3(W)}{2(1-\lambda)\sigma_V\sigma_W}, \tag{A.40}$$

$$\sigma_1^2 = \frac{\kappa_4(V)\sigma_W^2}{4\lambda\sigma_V^6} + \frac{\sigma_W^2}{2\lambda(1-\lambda)\sigma_V^2} + \frac{\kappa_4(W)}{4(1-\lambda)\sigma_V^2\sigma_W^2}, \tag{A.41}$$

where κ_i *indicates the* i^{th} *cumulant and* $\lambda = m/(m+n)$.

Proof of Theorem 1. The proof of Theorem 1 is now completed by using the results of Lemma 1 and Lemma 2. We can rewrite $\widehat{q}_{MOM}(v) - q(v)$ as follows

$$\widehat{q}_{MOM}(v) - q(v) \tag{A.42}$$

$$\begin{aligned}
&= \overline{W} - \frac{s_W}{s_V}\overline{V} + \frac{s_W}{s_V}v - \mu_W + \frac{\sigma_W}{\sigma_V}\mu_V - \frac{\sigma_W}{\sigma_V}v \\
&= \left(\overline{W} - \mu_W\right) - \frac{\sigma_W}{\sigma_V}\overline{V} + \left(\frac{\sigma_W}{\sigma_V} - \frac{s_W}{s_V}\right)\overline{V} + \left(\frac{s_W}{s_V} - \frac{\sigma_W}{\sigma_V}\right)v + \frac{\sigma_W}{\sigma_V}\mu_V \\
&= (\overline{W} - \mu_W) + \left(\frac{s_W}{s_V} - \frac{\sigma_W}{\sigma_V}\right)(v - \overline{V}) - (\frac{\sigma_W}{\sigma_V})(\overline{V} - \mu_V) \qquad \text{(A.43)} \\
&= (\overline{W} - \mu_W) - \frac{\sigma_W}{\sigma_V}(\overline{V} - \mu_V) + \left(\frac{s_W}{s_V} - \frac{\sigma_W}{\sigma_V}\right)(v - \overline{V}).
\end{aligned}$$

Since $(\overline{V} - \mu_V) = O_p(n^{-1/2})$ by the CLT and $\left(\frac{s_W}{s_V} - \frac{\sigma_W}{\sigma_V}\right) = o_p(1)$ by consistency, we have

$$\left(\frac{s_W}{s_V} - \frac{\sigma_W}{\sigma_V}\right)(\overline{V} - \mu_V) = o_p(n^{-1/2}). \tag{A.44}$$

Thus,

$$\widehat{q}_{MOM}(v) - q(v) \tag{A.45}$$

$$= \left(\overline{W} - \mu_W\right) - \frac{\sigma_W}{\sigma_V}\left(\overline{V} - \mu_V\right) + \left(\frac{s_W}{s_V} - \frac{\sigma_W}{\sigma_V}\right)\left(v - \overline{V}\right) + o_p(n^{-1/2}) \tag{A.46}$$

$$= \begin{bmatrix} 1 & -\frac{\sigma_W}{\sigma_V} & \left(v - \overline{V}\right) \end{bmatrix} \begin{bmatrix} \left(\overline{W} - \mu_W\right) \\ \left(\overline{V} - \mu_V\right) \\ \left(\frac{s_W}{s_V} - \frac{\sigma_W}{\sigma_V}\right) \end{bmatrix} + o_p(n^{-1/2}). \tag{A.47}$$

The asymptotic variance of $\sqrt{m+n}\,(\widehat{q}_{MOM}(v) - q(v))$ will therefore be

$$\begin{bmatrix} 1 & -\frac{\sigma_W}{\sigma_V} & v-\overline{V} \end{bmatrix} \cdot \Sigma^{***} \cdot \begin{bmatrix} 1 & -\frac{\sigma_W}{\sigma_V} & v-\overline{V} \end{bmatrix}^\top \tag{A.48}$$

$$= \left(\frac{\sigma_W^2}{\lambda(1-\lambda)}\right) + \left(\frac{\kappa_4(V)\sigma_W^2}{4\lambda\sigma_V^6} + \frac{\sigma_W^2}{2\lambda(1-\lambda)\sigma_V^2} + \frac{\kappa_4(W)}{4(1-\lambda)\sigma_W^2\sigma_V^2}\right)(v-\overline{V})^2 + \left(\frac{\kappa_3(W)}{(1-\lambda)\sigma_W\sigma_V} + \frac{\kappa_3(V)\sigma_W^2}{\lambda\sigma_V^4}\right)(v-\overline{V}) \tag{A.49}$$

$$= \sigma_0^2 + \sigma_1^2(v-\overline{V})^2 + 2\sigma_{0,1}(v-\overline{V})$$

$$= \tau(v)^2 \tag{A.50}$$

where Σ^{***} is given in Lemma 2, (A.32). ■

Note also that

$$\begin{bmatrix} \sigma_0^2 & \sigma_{0,1} \\ \sigma_{0,1} & \sigma_1^2 \end{bmatrix} = \begin{bmatrix} \frac{\sigma_W^2}{\lambda(1-\lambda)} & \frac{\kappa_3(V)\sigma_W^2}{2\lambda\sigma_V^4} + \frac{\kappa_3(W)}{2(1-\lambda)\sigma_V\sigma_W} \\ \frac{\kappa_3(V)\sigma_W^2}{2\lambda\sigma_V^4} + \frac{\kappa_3(W)}{2(1-\lambda)\sigma_V\sigma_W} & \frac{\kappa_4(V)\sigma_W^2}{4\lambda\sigma_V^6} + \frac{\sigma_W^2}{2\lambda(1-\lambda)\sigma_V^2} + \frac{\kappa_4(W)}{4(1-\lambda)\sigma_V^2\sigma_W^2} \end{bmatrix} \tag{A.51}$$

is the asymptotic covariance matrix of the vector

$$\left[(\overline{W}-\mu_W) - \frac{\sigma_W}{\sigma_V}(\overline{V}-\mu_V), \left(\frac{s_W}{s_V} - \frac{\sigma_W}{\sigma_V}\right)\right]. \tag{A.52}$$

This fact will be used in the proof of the next theorem. Then, a $100(1-\alpha)\%$ confidence band for q is based on the probability statement

$$P\left(\sup_v \left|\frac{\widehat{q}_{MOM}(v) - q(v)}{\widehat{\tau}(v)}\right| \le c_{\alpha,m+n}/\sqrt{m+n}\right) = 1-\alpha. \tag{A.53}$$

A.1.2 Proof of Theorem 2

Theorem 2 *The asymptotic value of $c_{\alpha,m+n}$ in (A.53) is $c_\alpha = \sqrt{-2\log_e(\alpha)}$.*

Proof of Theorem 2. From (A.46) we have

$$\sup_v \left|\frac{\widehat{q}_{MOM}(v) - q(v)}{\widehat{\tau}(v)}\right| \le \frac{c_{\alpha,m+n}}{\sqrt{m+n}} \tag{A.54}$$

if and only if for all v,

$$\left|\left(\overline{W}-\mu_W\right)-\frac{\sigma_W}{\sigma_V}\left(\overline{V}-\mu_V\right)+\left(\frac{s_W}{s_V}-\frac{\sigma_W}{\sigma_V}\right)\left(v-\overline{V}\right)\right| \leq \frac{c_{\alpha,m+n}\widehat{\tau}(v)}{\sqrt{m+n}}. \tag{A.55}$$

Set

$$\widetilde{\gamma}_0=\left(\overline{W}-\mu_W\right)-\frac{\sigma_W}{\sigma_V}\left(\overline{V}-\mu_V\right) \tag{A.56}$$

$$\widetilde{\gamma}_1=\left(\frac{s_W}{s_V}-\frac{\sigma_W}{\sigma_V}\right). \tag{A.57}$$

Then (A.55) becomes

$$\left|\widetilde{\gamma}_0+\widetilde{\gamma}_1\left(v-\overline{V}\right)\right| \leq c_{\alpha,m+n}\widehat{\tau}(v)\,/\sqrt{m+n}. \tag{A.58}$$

Now, (A.58) is true, if and only if,

$$(\widetilde{\gamma}_0+\widetilde{\gamma}_1\left(v-\overline{V}\right))^2 \leq \frac{c^2_{\alpha,m+n}}{m+n}(\widehat{\sigma}_0^2+\widehat{\sigma}_1^2)\left(v-\overline{V}\right)^2+2\widehat{\sigma}_{0,1}\left(v-\overline{V}\right)$$

which, after some algebra, is seen to be equivalent to

$$(\widetilde{\gamma}_1^2-\frac{c^2_{\alpha,m+n}\widehat{\sigma}_1^2}{m+n})\left(v-\overline{V}\right)^2+(2\widetilde{\gamma}_0\widetilde{\gamma}_1-2\frac{c^2_{\alpha,m+n}\widehat{\sigma}_{0,1}}{m+n})\left(v-\overline{V}\right)+(\widetilde{\gamma}_0^2-\frac{c^2_{\alpha,m+n}\widehat{\sigma}_0^2}{m+n}) \leq 0. \tag{A.59}$$

For (A.59) to be true for all v, two conditions must hold. The first condition is that the coefficient of $\left(v-\overline{V}\right)^2$ must be negative, i.e.

$$\left(\widetilde{\gamma}_1^2-\frac{c^2_{\alpha,m+n}\widehat{\sigma}_1^2}{m+n}\right)<0. \tag{A.60}$$

The second condition is that the maximum of (A.59) must be smaller than or equal to 0, i.e.

$$(2\widetilde{\gamma}_0\widetilde{\gamma}_1-2\frac{c^2_{\alpha,m+n}\widehat{\sigma}_{0,1}}{m+n})^2 \leq 4(\widetilde{\gamma}_1^2-\frac{c^2_{\alpha,m+n}\widehat{\sigma}_1^2}{m+n})(\widetilde{\gamma}_0^2-\frac{c^2_{\alpha,m+n}\widehat{\sigma}_0^2}{m+n}), \tag{A.61}$$

or, equivalently,

$$S:=(m+n)\left(\frac{\widetilde{\gamma}_1^2\widehat{\sigma}_0^2+\widetilde{\gamma}_0^2\widehat{\sigma}_1^2-2\widetilde{\gamma}_0\widetilde{\gamma}_1\widehat{\sigma}_{0,1}}{\widehat{\sigma}_0^2\widehat{\sigma}_1^2-\widehat{\sigma}_{0,1}^2}\right) \leq c^2_{\alpha,m+n}. \tag{A.62}$$

The left hand side of (A.62) is

$$S = (m+n) \begin{bmatrix} \widetilde{\gamma}_1 \\ \widetilde{\gamma}_0 \end{bmatrix}^\top \begin{bmatrix} \widehat{\sigma}_0^2 & -\widehat{\sigma}_{0,1} \\ -\widehat{\sigma}_{0,1} & \widehat{\sigma}_1^2 \end{bmatrix} \begin{bmatrix} \widetilde{\gamma}_1 \\ \widetilde{\gamma}_0 \end{bmatrix} / (\widehat{\sigma}_0^2 \widehat{\sigma}_1^2 - \widehat{\sigma}_{0,1}^2) \quad \text{(A.63)}$$

$$= (m+n) \begin{bmatrix} \widetilde{\gamma}_1 \\ \widetilde{\gamma}_0 \end{bmatrix}^\top \begin{bmatrix} \widehat{\sigma}_1^2 & \widehat{\sigma}_{0,1} \\ \widehat{\sigma}_{0,1} & \widehat{\sigma}_0^2 \end{bmatrix}^{-1} \begin{bmatrix} \widetilde{\gamma}_1 \\ \widetilde{\gamma}_0 \end{bmatrix}. \quad \text{(A.64)}$$

From the remark preceding Theorem 2, it follows that S has an asymptotic chi-squared distribution with 2 degrees of freedom, i.e. $S \sim \chi^2(2)$ (see e.g. Rice, 1995, page 318), i.e.

$$1 - \alpha = \lim_{m,n \to \infty} P\left(\sup_v \left| \frac{\widehat{q}_{MOM}(v) - q(v)}{\widehat{\tau}(v)} \right| \leq c_{\alpha,m+n} / \sqrt{m+n} \right) \quad \text{(A.65)}$$

$$= \lim_{m,n \to \infty} P\left((m+n) \begin{bmatrix} \widetilde{\gamma}_1 \\ \widetilde{\gamma}_0 \end{bmatrix}^\top \begin{bmatrix} \widehat{\sigma}_1^2 & \widehat{\sigma}_{0,1} \\ \widehat{\sigma}_{0,1} & \widehat{\sigma}_0^2 \end{bmatrix}^{-1} \begin{bmatrix} \widetilde{\gamma}_1 \\ \widetilde{\gamma}_0 \end{bmatrix} \leq c_{\alpha,m+n}^2 \right) \quad \text{(A.66)}$$

$$= P(\chi^2(2) \leq c_\alpha^2). \quad \text{(A.67)}$$

Thus, $c_\alpha = \sqrt{-2 \log_e(\alpha)}$. ■

A.1.3 Proof of Theorem 3

Lemma 3 *The joint asymptotic distribution of $\widehat{m}_V$, $\widehat{m}_W$, $\widehat{i}_V$ and $\widehat{i}_W$ is:*

$$\sqrt{m+n} \left[\widehat{m}_V - m_V, \widehat{m}_W - m_W, \frac{\widehat{i}_W}{\widehat{i}_V} - \frac{i_W}{i_V} \right] \sim N_3\left(0, C \Sigma C'\right) \quad \text{(A.68)}$$

where

$$\Sigma = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \quad \text{(A.69)}$$

and

$$\Sigma_1 = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} \quad \text{(A.70)}$$

with

$$\sigma_{11} = \frac{1}{4f_V^2(m_V)} \tag{A.71}$$

$$\sigma_{22} = \left(\frac{3}{4f_V(\xi_{\frac{3}{4}})} - \frac{1}{4f_V(\xi_{\frac{1}{4}})}\right)^2 \tag{A.72}$$

$$\sigma_{12} = \frac{1}{8f_V(m_V)}\left(\frac{3}{4f_V(\xi_{\frac{3}{4}})} - \frac{1}{4f_V(\xi_{\frac{1}{4}})}\right) \tag{A.73}$$

$$C = \begin{bmatrix} \frac{1}{\sqrt{\lambda}} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{1-\lambda}} & 0 \\ 0 & -\frac{i_W}{\sqrt{\lambda}(i_V)^2} & 0 & \frac{1}{\sqrt{1-\lambda}i_V} \end{bmatrix}. \tag{A.74}$$

Σ_1 *is the asymptotic covariance matrix of* $(\sqrt{m}\,(\widehat{m}_V - m_V), \sqrt{m}\left(\widehat{i}_V - i_V\right))$ *and similarly* Σ_2 *is the asymptotic covariance matrix of* $(\sqrt{n}\,(\widehat{m}_W - m_W), \sqrt{n}\left(\widehat{i}_W - i_W\right))$.

Proof of Lemma 3. The joint distribution of sample quantiles is known (see e.g. van der Vaart, 1998 or Lehmann, 1999) and therefore we have that $\left(z_{\widehat{m}_V}, z_{\widehat{i}_V}\right)$ is asymptotically bivariate normal with mean zero and covariance matrix

$$\Sigma_1 = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} \tag{A.75}$$

where σ_{11}, σ_{12} and σ_{22} are given in (A.71), (A.72) and (A.73), and where

$$\begin{aligned} z_{\widehat{m}_V} &= \sqrt{m}\,(\widehat{m}_V - m_V) \qquad (A.76) \\ z_{\widehat{m}_W} &= \sqrt{n}\,(\widehat{m}_W - m_W) \\ z_{\widehat{i}_V} &= \sqrt{m}\left(\widehat{i}_V - i_V\right) \\ z_{\widehat{i}_W} &= \sqrt{n}\left(\widehat{i}_W - i_W\right). \end{aligned}$$

Similarly $(z_{\widehat{m}_W}, z_{\widehat{i}_W})$ is asymptotically bivariate normal with mean zero and covariance matrix Σ_2. Also $(z_{\widehat{m}_V}, z_{\widehat{i}_V})$ and $(z_{\widehat{m}_W}, z_{\widehat{i}_W})$ are statistically independent. We prove in

Lemma 4 below that

$$\frac{\widehat{i}_W}{\widehat{i}_V} = \frac{i_W}{i_V} + \frac{z_{\widehat{i}_W}}{\sqrt{n}\widehat{i}_V} - \frac{i_W}{\sqrt{m}(i_V)^2} z_{\widehat{i}_V} + o_p((n+m)^{-1/2}). \tag{A.77}$$

Then it follows that

$$\sqrt{m+n}\left(\widehat{m}_V - m_V, \widehat{m}_W - m_W, \frac{\widehat{i}_W}{\widehat{i}_V} - \frac{i_W}{i_V}\right)$$

$$= \sqrt{m+n}\left(\frac{z_{\widehat{m}_V}}{\sqrt{n}}, \frac{z_{\widehat{m}_W}}{\sqrt{m}}, \frac{z_{\widehat{i}_W}}{\sqrt{n}i_V} - \frac{i_W}{\sqrt{m}(i_V)^2} z_{\widehat{i}_V}\right) + o_p(1) \tag{A.78}$$

$$= \left(\frac{z_{\widehat{m}_V}}{\sqrt{\lambda}}, \frac{z_{\widehat{m}_W}}{\sqrt{1-\lambda}}, \frac{z_{\widehat{i}_W}}{\sqrt{1-\lambda}i_V} - \frac{i_W}{\sqrt{\lambda}(i_V)^2} z_{\widehat{i}_V}\right) + o_p(1)$$

$$= \begin{bmatrix} \frac{1}{\sqrt{\lambda}} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{1-\lambda}} & 0 \\ 0 & -\frac{i_W}{\sqrt{\lambda}(i_V)^2} & 0 & \frac{1}{\sqrt{1-\lambda}i_V} \end{bmatrix} \begin{bmatrix} z_{\widehat{m}_V} \\ z_{\widehat{i}_V} \\ z_{\widehat{m}_W} \\ z_{\widehat{i}_W} \end{bmatrix} + o_p(1)$$

$$= C \begin{bmatrix} z_{\widehat{m}_V} & z_{\widehat{i}_V} & z_{\widehat{m}_W} & z_{\widehat{i}_W} \end{bmatrix}^\top + o_p(1) \tag{A.79}$$

where $\lambda = m/(m+n)$. $\left(z_{\widehat{m}_V}, z_{\widehat{i}_V}, z_{\widehat{m}_W}, z_{\widehat{i}_W}\right)$ is asymptotically 4-variate normal with mean zero and covariance matrix Σ where Σ is given in (A.69). Thus,

$$\sqrt{m+n}\left[\widehat{m}_V - m_V, \widehat{m}_W - m_W, \frac{\widehat{i}_W}{\widehat{i}_V} - \frac{i_W}{i_V}\right] \sim N_3\left(0, C\Sigma C'\right). \tag{A.80}$$

This completes the proof. ■

Lemma 4 *The following expression holds*

$$\frac{\widehat{i}_W}{\widehat{i}_V} = \frac{i_W}{i_V} + \frac{z_{\widehat{i}_W}}{\sqrt{n}\widehat{i}_V} - \frac{i_W}{\sqrt{m}(i_V)^2} z_{\widehat{i}_V} + o_p((n+m)^{-1/2}). \tag{A.81}$$

Proof of Lemma 4.

$$\begin{aligned}
\frac{\widehat{i}_W}{\widehat{i}_V} &= \frac{z_{\widehat{i}_W}/\sqrt{n}+i_W}{z_{\widehat{i}_V}/\sqrt{m}+i_V} && \text{(A.82)}\\
&= \left(z_{\widehat{i}_W}/\sqrt{n}+i_W\right)\frac{1}{i_V\left(1+\frac{z_{\widehat{i}_V}}{\sqrt{m}i_V}\right)}\\
&= \left(z_{\widehat{i}_W}/\sqrt{n}+i_W\right)\frac{1-z_{\widehat{i}_V}/\sqrt{m}i_V}{i_V}+o_p((n+m)^{-1/2}) && \text{(A.83)}\\
&= \frac{z_{\widehat{i}_W}}{\sqrt{n}i_V}+\frac{i_W}{i_V}-\frac{i_W}{\sqrt{m}(i_V)^2}z_{\widehat{i}_V}-\frac{z_{\widehat{i}_W}z_{\widehat{i}_V}/\sqrt{m}i_V}{\sqrt{n}\sqrt{m}(i_V)^2}+o_p((n+m)^{-1/2})\\
&= \frac{z_{\widehat{i}_W}}{\sqrt{n}i_V}-\frac{i_W}{\sqrt{m}(i_V)^2}z_{\widehat{i}_V}+\frac{i_W}{i_V}+o_p((n+m)^{-1/2}).
\end{aligned}$$

■

Theorem 3 *The asymptotic distribution of $\widehat{q}_{MOQ}$ is given by the expression*

$$\sqrt{m+n}(\widehat{q}_{MOQ}(v)-q(v))\sim N(0,\tau_{MOQ}(v)^2) \tag{A.84}$$

where

$$\tau_{MOQ}(v)^2 = C_1+C_2\left(v-\widehat{m}_V\right)+C_3\left(v-\widehat{m}_V\right)^2 \tag{A.85}$$

and

$$\begin{aligned}
C_1 &= \frac{i_W^2}{4\lambda i_V^2 f_V^2(m_V)}+\frac{1}{4\left(1-\lambda\right)f_W^2(m_W)} && \text{(A.86)}\\
C_2 &= \frac{-i_W}{2\sqrt{\lambda}\sqrt{1-\lambda}f_W^2(m_W)}+ && \text{(A.87)}\\
&\quad \frac{1}{4\left(1-\lambda\right)i_V f_W(m_W)}\left(\frac{3}{4f_W(\eta_{\frac{3}{4}})}-\frac{1}{4f_W(\eta_{\frac{1}{4}})}\right) && \text{(A.88)}\\
C_3 &= \frac{i_W^2}{4\lambda f_W^2(m_W)}-\frac{i_W}{4\sqrt{\lambda}\sqrt{1-\lambda}i_V f_W(m_W)}\left(\frac{3}{4f_W(\eta_{\frac{3}{4}})}-\frac{1}{4f_W(\eta_{\frac{1}{4}})}\right) && \text{(A.89)}\\
&\quad +\frac{1}{\left(1-\lambda\right)i_V^2}\left(\frac{3}{4f_W(\eta_{\frac{3}{4}})}-\frac{1}{4f_W(\eta_{\frac{1}{4}})}\right)^2.
\end{aligned}$$

Proof of Theorem 3. The proof of Theorem 3 can now be completed by using the results of Lemma 3 and Lemma 4. The proof follows along exactly the same lines as that of Theorem 1, and is therefore omitted.

A.1.4 Proof of Theorem 4

Theorem 4 *The asymptotic value of* $d_{\alpha,m+n}$ *in (2.54) is* $d_\alpha = \sqrt{-2\log_e(\alpha)}$.

Proof of Theorem 4. The proof follows along exactly the same lines as that of Theorem 2 and is therefore omitted. ■

A.2 Algorithms

Algorithm 1: Determine coverage probability given any critical value, c

1. Generate $V_1, \ldots, V_m$, from F and generate $W_1, \ldots, W_n$ from G.

2. For each sample calculate S_l, where S_l denotes the value of S in the l^{th} sample and S is given in (A.63).

3. Repeat steps 1 and 2 L times, therefore calculate $S_1, ..., S_L$.

4. Calculate the coverage probability, i.e. $(1/L)\sum_{l=1}^{L} I\left(S_l \leq c^2\right)$, where $c = \sqrt{-2\log_e(\alpha)}$ (the asymptotic critical value) or c^* (a bootstrap estimated critical value).

Algorithm 2: Calculate the bootstrap critical value

1. Generate a bootstrap sample, $V_1^*, \ldots, V_m^*$, from F, and a bootstrap sample, W_1^*, $\ldots, W_n^*$, from G.

2. For each sample calculate S_b^*, where S_b^* denotes the value of S in the b^{th} bootstrap sample, with $b = 1, ..., B$ and S is given in (A.63).

3. Repeat steps 1 and 2 B times. Keep a record of all the S_b^*-values.

4. The $100(1-\alpha)\%$ percentile of the S_b^*-values is the bootstrap estimated critical value, c_α^{*2}.

Algorithm 3: Determine significance levels and power of tests

1. Repeat the following steps L times:

   (a) Assume two independent samples from F and G.

   (b) Calculate the $K_{s,l}$ statistic, $s = 1, 2$ and $l = 1, 2, ..L$.

   (c) Repeat the next three steps B times:

      i. For significance levels, we generate bootstrap samples from F, G **or** H. For power, we generate bootstrap samples from F, G **and** H (see also the comments in Section 2.8.1 under the heading "Power").

      ii. Compute $K^*_{b,s,l}$, $b = 1, .., B$.

      iii. The $(1-\alpha)^{th}$ percentile of the $K^*_{b,s,l}$, is the critical value, $K^*_{\alpha,s,l}$.

   (d) Record $I\left(K_{s,l} > K^*_{\alpha,s,l}\right)$.

2. The estimated significance level (or power) is given by

$$(1/L)\sum_{l=1}^{L} I\left(K_{s,l} > K^*_{\alpha,s,l}\right). \tag{A.90}$$

A.3 General

Calculation 1: Smooth bootstrap Throughout the thesis, we used a smooth bootstrap. Smooth bootstrap gives a continuous estimation for the distribution and is therefore usually the preferred method of bootstrap if the underlying data is continuous (Davison & Hinkley, 1997). Given a sample $V_1, \ldots, V_m$, a smooth bootstrap sample, V^*, is given by

$$V_i^* = X_i + \epsilon_i,\ i = 1, ..., m \tag{A.91}$$

where X_i is sampled with replacement from V, $\epsilon_{1,}..., \epsilon_m$ is from $N(0, h^2)$ and h is the bandwidth (see e.g. Davison and Hinkley, 1997, p.79). We used the optimal bandwidth of Wand and Jones (1995, p. 60), $h = \left[\frac{4}{3m}\right]^{1/5} s_V$, assuming that the data is normally distributed. The density function of V^* is precisely the kernel estimator of the density function of V .

Calculation 2: Calculations for G^{-1} and F^{-1}: We know that (see e.g. Rice, 1995),

$$F_m^{-1}(\frac{1}{m}) = \widetilde{V}_{(1)}, ..., F_m^{-1}(\frac{m}{m}) = \widetilde{V}_{(m)} \tag{A.92}$$

$$G_n^{-1}(\frac{1}{n}) = \widetilde{W}_{(1)}, ..., G_n^{-1}(\frac{n}{n}) = \widetilde{W}_{(n)} \tag{A.93}$$

where $\widetilde{V}_i = \frac{V_i - \overline{V}}{s_V}$ and $\widetilde{W}_j = \frac{W_j - \overline{W}}{s_W}$. Therefore, in general we have

$$F_m^{-1}(t) = \widetilde{V}_{(\lceil mt \rceil)} \tag{A.94}$$

$$G_n^{-1}(t) = \widetilde{W}_{(\lceil nt \rceil)}. \tag{A.95}$$

Calculation 3: Calculate the density estimator, $\widehat{f}$: Throughout the thesis, we used kernel density estimation to estimate f and chose the normal kernel and the optimal bandwidth given by Wand and Jones (1995, p.60), unless stated otherwise. Note that the optimal bandwidth given by Wand and Jones (1995) assume that the underlying data is normally distributed. The density estimate of $V_1, \ldots, V_m$ will therefore be

$$\widehat{f}_b(x) = \frac{1}{mb} \sum_{i=1}^{m} K_0 \left(\frac{x - \widetilde{V}_i}{b} \right) \tag{A.96}$$

where

$$K_0(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} z^2 \right) \tag{A.97}$$

and

$$b = \left[\frac{4}{3m} \right]^{1/5} s_{\widetilde{V}}. \tag{A.98}$$

We now find the optimal bandwidth for Cauchy distributed data. According to Silverman (1986) the optimal bandwidth is.

$$b_{opt} = \sigma^{-2/5} \left\{ \int_{-\infty}^{\infty} K_0^2(t) dt \right\}^{1/5} \left\{ \int_{-\infty}^{\infty} f''(x)^2 dx \right\}^{-1/5} n^{-1/5} \tag{A.99}$$

where

$$\int_{-\infty}^{\infty} K_0^2(t) dt = \int_{-\infty}^{\infty} \frac{1}{2\pi} \exp\left(-z^2 \right) = \frac{1}{2\sqrt{\pi}}. \tag{A.100}$$

The Cauchy probability density function is given by

$$f(x) = \frac{1}{\gamma} \frac{1}{\pi \left(1 + \left(\frac{x}{\gamma} \right)^2 \right)} = \frac{\gamma}{\pi \left(x^2 + \gamma^2 \right)}. \tag{A.101}$$

The second derivative of $f(x)$ is given by

$$\begin{aligned} f''(x) &= \frac{2\gamma 2 \left(x^2 + \gamma^2 \right) 2x}{\pi \left(x^2 + \gamma^2 \right)^4} \\ &= \frac{8\gamma}{\pi} \frac{1}{\left(x^2 + \gamma^2 \right)^3}. \end{aligned} \tag{A.102}$$

Therefore we have

$$\begin{aligned}\int_{-\infty}^{\infty} f''(x)^2 dx &= \frac{64\gamma^2}{\pi^2}\int_{-\infty}^{\infty}\frac{dx}{(x^2+\gamma^2)^6} \\ &= \frac{64\gamma^2}{\pi^2}\frac{\gamma}{\gamma^{12}}\int_{-\infty}^{\infty}\frac{d\left(\frac{x}{\gamma}\right)}{\left(1+\left(\frac{x}{\gamma}\right)^2\right)^6} \\ &= \frac{64\gamma^2}{\pi^2\gamma^9}\int_{-\infty}^{\infty}\frac{dy}{(1+y^2)^6} \\ &= \frac{64\gamma^2}{\pi^2\gamma^9}\frac{63}{256}\pi.\end{aligned} \tag{A.103}$$

The optimal bandwidth is therefore

$$b_{opt} = \sigma^{-2/5}\left\{\frac{1}{2\sqrt{\pi}}\right\}^{1/5}\left\{\frac{64\gamma^2}{\pi^2\gamma^9}\frac{63}{256}\pi\right\}^{-1/5} n^{-1/5}. \tag{A.104}$$

Calculation 4: Calculation of $\sup_y\left|\widetilde{F}_m(y)-\widetilde{G}_n(y)\right|$ This supremum will be at one of the $m+n$ datapoints, and therefore the distribution functions F and G have to be evaluated at these points only. In other words, calculate

$$\widetilde{F}_m(\widetilde{V}_1),...,\widetilde{F}_m(\widetilde{V}_m),\widetilde{F}_m(\widetilde{W}_1),...,\widetilde{F}_m(\widetilde{W}_n) \tag{A.105}$$

and

$$\widetilde{G}_n(\widetilde{V}_1),...,\widetilde{G}_n(\widetilde{V}_m),\widetilde{G}_n(\widetilde{W}_1),...,\widetilde{G}_n(\widetilde{W}_n) \tag{A.106}$$

where $\widetilde{V}_1,...,\widetilde{V}_m$ are the ordered values of $\widetilde{V}$ and $\widetilde{W}_1,...,\widetilde{W}_n$ are the ordered values of $\widetilde{W}$. We calculate these as follow

$$\widetilde{F}_m(\widetilde{V}_1)=\frac{1}{m},...,\widetilde{F}_m(\widetilde{V}_m)=\frac{m}{m} \tag{A.107}$$

$$\widetilde{G}_n(\widetilde{W}_1)=\frac{1}{n},...,\widetilde{G}_n(\widetilde{W}_n)=\frac{n}{n} \tag{A.108}$$

$$\widetilde{F}_m(\widetilde{W}_1)=\frac{1}{m}\sum_{j=1}^{m}I(\widetilde{V}_j\le\widetilde{W}_1),...,\widetilde{F}_m(\widetilde{W}_n)=\frac{1}{m}\sum_{j=1}^{m}I(\widetilde{V}_j\le\widetilde{W}_n) \tag{A.109}$$

$$\widetilde{G}_n(\widetilde{V}_1)=\frac{1}{n}\sum_{j=1}^{n}I(\widetilde{W}_j\le\widetilde{V}_1),...,\widetilde{G}_n(\widetilde{V}_m)=\frac{1}{n}\sum_{j=1}^{n}I(\widetilde{W}_j\le\widetilde{V}_m). \tag{A.110}$$

APPENDIX B

Technical details of Chapter 3

This appendix contains the technical details of Chapter 3.

B.1 Proof of C-Step Lemma

C-step lemma *Let*

$$D_{\pi_1}(\overline{\boldsymbol{\beta}}(\mathbf{w})) \geq ... \geq D_{\pi_N}(\overline{\boldsymbol{\beta}}(\mathbf{w})) \tag{B.1}$$

and g *be an integer with* $1 \leq g \leq N$. *Suppose*

$$\mathbf{w}' = \{w'_1, ..., w'_N\} \tag{B.2}$$

is another set of weights such that

$$w'_{\pi_i} \geq w_{\pi_i} \text{ for } i = 1, ..., g \tag{B.3}$$

and

$$w'_{\pi_i} \leq w_{\pi_i} \text{ for } i = g+1, ..., N \tag{B.4}$$

and also

$$\sum_n w'_n = \sum_n w_n. \tag{B.5}$$

Then

$$\sum_n w'_n D_n(\overline{\boldsymbol{\beta}}(\mathbf{w}') \geq \sum_n w_n D_n(\overline{\boldsymbol{\beta}}(\mathbf{w})). \tag{B.6}$$

Proof. For any sets of weights, $\mathbf{w} = \{w_1, ..., w_N\}$, let $\overline{\boldsymbol{\beta}}(\mathbf{w})$ be the maximiser over $\boldsymbol{\beta}$ of $\sum_n w_n \, D_n(\boldsymbol{\beta})$ where $D_n(\boldsymbol{\beta})$ is given by (3.2). Then

$$\sum_n w'_n D_n(\overline{\boldsymbol{\beta}}(\mathbf{w}')) = \max_{\boldsymbol{\beta}} \sum_n w'_n D_n(\boldsymbol{\beta}) \tag{B.7}$$

$$
\begin{aligned}
&\geq \sum_n w'_n D_n(\overline{\boldsymbol{\beta}}(\mathbf{w})) \\
&= \sum_i w'_{\pi_i} D_{\pi_i}(\overline{\boldsymbol{\beta}}(\mathbf{w})) \\
&\geq \sum_i w_{\pi_i} D_{\pi_i}(\overline{\boldsymbol{\beta}}(\mathbf{w})) \\
&= \sum_n w_n D_n(\overline{\boldsymbol{\beta}}(\mathbf{w}))
\end{aligned}
$$

where the inequality in the second last line may be argued as follows. To simplify notation, write $D_{\pi_i} = D_{\pi_i}(\overline{\boldsymbol{\beta}}(\mathbf{w}))$. Then we have

$$
\begin{aligned}
\sum_{i=1}^{g}(w'_{\pi_i} - w_{\pi_i})D_{\pi_i} &\geq D_{\pi_{g+1}} \sum_{i=1}^{g}(w'_{\pi_i} - w_{\pi_i}) \qquad \text{(B.8)} \\
&= D_{\pi_{g+1}} \left[(\sum_{n=1}^{N} w'_n - \sum_{i=g+1}^{N} w'_{\pi_i}) - (\sum_{n=1}^{N} w_n - \sum_{i=g+1}^{N} w_{\pi_i}) \right] \\
&= D_{\pi_{g+1}} \sum_{i=g+1}^{N} (w_{\pi_i} - w'_{\pi_i}) \\
&\geq \sum_{i=g+1}^{N} (w_{\pi_i} - w'_{\pi_i}) D_{\pi_i}.
\end{aligned}
$$

Hence,

$$
\sum_{i=1}^{N}(w'_{\pi_i} - w_{\pi_i})D_{\pi_i} \geq 0 \qquad \text{(B.9)}
$$

or

$$
\sum_i w'_{\pi_i} D_{\pi_i} \geq \sum_i w_{\pi_i} D_{\pi_i} \qquad \text{(B.10)}
$$

which proves the lemma. ■

Comments on the lemma: The Neykov and Muller (2002) C-step proposition is a special case where the weights w_n are restricted to the values 0 and 1. To apply this lemma to our procedure, for a given $\mathbf{G}$ define

$$
w_n = 1 \text{ for } n \in \mathbf{G} \qquad \text{(B.11)}
$$

and

$$w_n = \epsilon \text{ for } n \notin \mathbf{G} \tag{B.12}$$

so that

$$\overline{\boldsymbol{\beta}}(\mathbf{w}) = \boldsymbol{\beta}^*(\mathbf{G}) \tag{B.13}$$

is the corresponding maximiser over $\boldsymbol{\beta}$ of $l_\epsilon(\boldsymbol{\beta}, \mathbf{G})$. Then get the π_i's so that

$$D_{\pi_1}(\overline{\boldsymbol{\beta}}(\mathbf{w})) \geq ... \geq D_{\pi_N}(\overline{\boldsymbol{\beta}}(\mathbf{w})) \tag{B.14}$$

and let

$$w'_{\pi_i} = 1 \text{ for } i = 1, ..., g_1 \tag{B.15}$$

and

$$w'_{\pi_i} = \epsilon \text{ for } i = g_1 + 1, ..., N. \tag{B.16}$$

By the lemma $\mathbf{G}' = \{\pi_1, ..., \pi_{g_1}\}$ is a better subset in the sense that

$$l_\epsilon(\boldsymbol{\beta}^*(\mathbf{G}'), \mathbf{G}') \geq l_\epsilon(\boldsymbol{\beta}^*(\mathbf{G}), \mathbf{G}). \tag{B.17}$$

This is the C-step iteration of the DOUW procedure.

It is interesting to note that the specific form of $D_n(\boldsymbol{\beta})$ does not influence the result of this lemma. It can therefore be applied to maximisation or minimisation of expressions of the form $\sum_n w_n \; D_n(\boldsymbol{\beta})$ in many other situations.

REFERENCES

1. BIANCO, A.M. and YOHAI, V.J. 1996. Robust Estimation in the Logistic Regression Model. In H.Rieder, editor, *Robust Statistics, Data Analysis, and Computer Intensive Methods*, pp. 17-34; *Lecture Notes in Statistics* **109**, Springer Verlag: New York.

2. BRILLINGER, D.R. 1965. *Time series: data analysis and theory*. New York : Holt, Rinehart and Winston, INC.

3. CANNER, P.L. 1975. A simulation study of one- and two-sample Kolmogorov-Smirnov Statistics with a particular weight Function. *Journal of the American Statistical Association*, **70**(349):209-211.

4. CHRISTMANN, A. and ROUSSEEUW, P.J. 2001. Measuring overlap in binary regression. *Computational Statistics and Data Analysis*, **37**(1):65-75.

5. COPAS, J.B. 1988. Binary regression models for contaminated data. With discussion. *Journal of the Royal Statistical Society. Series B (Methodological)*, **50**(2):225-265.

6. DAVISON, A.C. and HINKLEY, D.V. 1997. *Bootstrap Methods and their Application*. Cambridge: Cambridge University Press.

7. DE JONGH, P.J. and DE WET, T. 1986. Confidence Intervals for regression parameters based on trimmed means. *South African Statistical Journal*, **20**(2):137-164.

8. DOKSUM, K. 1974. Empirical probability Plots and Statistical Inference for Nonlinear Models in the Two-Sample Case. *The Annals of Statistics*, **2**(2):267 – 277.

9. DOKSUM, K.A. and SIEVERS, G.L. 1976. Plotting with confidence: Graphical Comparisons of Two populations. *Biometrika*, **63**(3):421–434.

10. EFRON, B. 1986. Double exponential families and their use in generalized linear regression. *Journal of American Statistical Association*, **81**(395):709-721.

11. EVANS, M., HASTINGS, H and PEACOCK, B. 2000. *Statistical Distributions*, 3rd. Ed., John Wiley and Sons.

12. FERGUSON, T.S. 1996. *A course in large sample theory*. London: Chapman and Hall.

13. FINNEY, D.J. 1947. The estimation from individual records of the relationship between dose and quantal response. *Biometrika*, **32**(1):320-334.

14. HAND, D.J. 1997. *Construction and assessment of Classification Rules*. New York: Wiley.

15. HAND, D. J. 2001. Modeling Consumer Credit Risk. *IMA Journal of Management Mathematics,* **12**(1):139-155.

16. HAND, D.J. 2004. Credit scoring. *Encyclopedia of Actuarial Science*, **1**:410-441.

17. HAND, D.J. and HENLEY, W.E. 1997. Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society, Series A*, **160**:523-541.

18. HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. 2001. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer.

19. HSIEH, F. 1995. The Empirical Process Approach for Semiparametric Two-Sample Models with Heterogeneous Treatment Effect. *Journal of the Royal Statistical Society, Series B (Methodological),* **57**(2):735-748.

20. HOLMGREN, E.B. 1995. The P-P plot as a means of comparing treatment effects. *Journal of the American Statistical Association*, **90**(429):360-365.

21. HOSMER, D.W. and LEMESHOW, S. 1989. *Applied logistic regression*. New York: Wiley.

22. HUBER, P.J. 1973. Robust Regression: asymptotics, Conjectures and Monte Carlo. *The Annals of Statistics*, **1**(5):799-821.

23. KLEINBAUM, D.G. 1994. *Logistic regression: a self-learning text*. New York: Springer.

24. KUNSCH, H.R., STEFANSKI, L.A. and CARROLL, R.J. 1989. Conditionally unbiased bounded influence estimation in general regression models, with applications to generalized linear models. *Journal of the American Statistical Association*, **84**(406):460-466.

25. LEHMANN, E.L. 1974. *Statistical Methods Based on Ranks*. San Francisco: Holden Day.

26. LEHMANN, E.L. 1999. *Elements of large-sample theory*. New York: Springer.

27. LINHART, H. and ZUCCHINI, W. 1986. *Model Selection*. New York: John Wiley and Sons.

28. LOMBARD, F. 2005. Nonparametric Confidence Bands for a Quantile Comparison Function. *Technometrics*, **47**(3):348-378.

29. MAYS, E. 2004. *Credit Scoring for Risk Managers: The Handbook for Lenders*. Mason: Thomson South-Western.

30. MCNAB, H. and WYNN, A. 2000. *Principles and Practice of Consumer Credit Risk Management*. Canterbury: Financial World Publishing.

31. MUSHKUDIANI, N.A. and EINMAHL, J.H.J. 2007. Generalized probability–probability plots. *Journal of Statistical Planning and Inference*, **137**(3):738-752.

32. NETER, J., WASSERMAN, W. and KUTNER, M.H. 1985. *Applied Linear Statistical Models (Second Edition)*. Illinois: Richard D Irwin, INC.

33. NEYKOV, N.M. and MULLER, C.H. 2002. Breakdown Point and Computation of Trimmed Likelihood Estimators in GLMs. In: R. Dutter et al., editors, *Developments in robust statistics*, Physica Verlag, Heidelberg.

34. NOVA, J. 2000. *Drilling Down: Turning Customer Data into Profits with a Spreadsheet*. Florida: Deep South Publishing Company.

35. PEARSON, E.S. and HARTLEY, H.O. 1972. *Biometrika Tables for Statisticians, Vol. 2*. Cambridge: Cambridge University Press.

36. POTGIETER, C.J. 2006. *Estimation and testing of linear treatment effects from matched pair data*. Johannesburg: UJ. (M.Sc.- Dissertation, University of Johannesburg).

37. PREGIBON, D. 1981. Logistic Regression Diagnostics. *The Annals of Statistics*, **9**(4):705-724.

38. RAO, C.R. 1965. *Linear statistical inference and its applications*. New York: Wiley.

39. RICE, J.A. 1995. *Mathematical Statistics and Data Analysis. 2nd Ed*. Belmont: Duxbury Press.

40. ROUSSEEUW, P.J. 1984. Least Median of Squares Regression. *Journal of the American Statistical Association*, **79**(388):871-880.

41. ROUSSEEUW, P.J. and CHRISTMANN, A. 2003. Robustness against separation and outliers in logistic regression. *Computational Statistics and Data Analysis*, **43**(3):315-332.

42. ROUSSEEUW, P.J. and LEROY, A.M. 1987. *Robust regression and outlier detection*. New York: John Wiley and Sons.

43. ROUSSEEUW, P.J. and VAN DRIESSEN, K. 1999a. *Computing LTS Regression for Large data sets*. Technical report, University of Antwerp.

44. ROUSSEEUW, P.J. and VAN DRIESSEN, K. 1999b. A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, **41**(3):212 -223.

45. SAS INSTITUTE INC. 2000. *SAS Institute Course Notes. Predictive Modeling Using Logistic Regression*. Cary, NC.

46. SAS INSTITUTE INC. 2003. *SAS software version 9.1.3*. Cary, NC.

47. SIDDIQI, N. 2006. *Credit Risks Scorecards: Developing and Implementing Intelligent Credit Scoring*. New Jersey: John Wiley and Sons.

48. SILVERMAN, B.W. 1986. *Density estimation for statistics and data analysis*. London: Chapman and Hall.

49. THOMAS, L.C. 2000. A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, **16**(2):149-172.

50. THOMAS, L.C., EDELMAN, D.B. and CROOK, J.N. 2002. *Credit Scoring and Its Applications*. Philadelphia: SIAM.

51. VAN DER VAART, A.W. 1998. *Asymptotic Statistics*. Cambridge: Cambridge University Press.

52. WAND, M.P. and JONES, M.C. 1995. Kernel Smoothing. *Vol. 60 of Monographs on Statistics and Applied Probability*. London: Chapman and Hall.

53. WIELANGA, D., LUCAS, B. and GEORGES. J. 1999. *Enterprise Miner: Applying Data Mining Techniques Course Notes*. SAS Institute Inc., Cary, NC, United States of America.

54. YOHAI, V.J. 1987. High Break-Point and High Efficiency Robust Estimates for Regression. *The Annals of Statistics*, **15**(2):642-656.