Robust Regression Methods for Insurance Risk Classification

Esteban Flores

A Thesis
for
The Department
of
Mathematics and Statistics

Presented in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy at
Concordia University
Montréal, Québec, Canada

March, 2002

0-612-85274-1

Abstract

Robust Regression Methods for
Insurance Risk Classification

Esteban Flores, Ph.D.
Concordia University, 2002

Risk classification is an important actuarial process for Insurance companies. It allows for the underwriting of the best risks, through an appropriate choice of classification variables, and helps set fair premiums in rate-making.

Currently, insurance companies mainly use ad-hoc methods for risk classification, more often based on the type of expenses covered than on the distribution of the corresponding losses. The selection of classification variables is also, in general, based on rate-making variables rather than on an optimal choice criteria based on statistical methods.

It is known that logistic regression is among the many sophisticated statistical methods used by the banking industry in order to select credit rating variables. Extending the method to insurance risks seems only natural.

Insurance risks are not usually classified in only two categories, good and bad, as can be the case in credit rating, but in a larger number of classes. Here we consider the generalization of the model to extend the use of logistic regression to insurance risk classification.

Since insurance data presents catastrophic losses and heavy tailed claim distributions, a robust estimation analysis is very important. It is carefully studied here.

Acknowledgments

I wish to thank my supervisor Doctor José Garrido for his insightful guidance, comments, clarifications and suggestions throughout the preparation of this thesis.

I am particularly thankful to Doctor Andrew Luong for his valuable comments and suggestions that enabled me to improve the final version of my thesis. As well, I would like to thank Professor Y. P. Chaubey for his endless encouragement and significant contributions to my work. I also wish to thank Doctor Fassil Nebebe and Doctor Vincent Goulet, members of the committee for their participation in the defense of my thesis.

I would also like to express gratitude to the Concordia University and the department of Mathematics and Statistics for their endless support.

I must point out that this project would not have gotten off the ground without the financial support of the University of Talca, Chile. I would also like to acknowledge the important contribution of the Ph.D. Grant that I was awarded from the Casualty Actuarial Society (CAS) and the Society of Actuaries (SOA).

Finally, I wish to thank my family and friends to whom I am infinitely grateful for their assistance and encouragement throughout the development of this academic pursuit.

Contents

List of Figures

List of Tables

Introduction

A minimum distance method based on a quadratic distance was introduced by Luong and Thompson (1987). Following the same idea a minimum quadratic distance estimator (QDE) was defined by Luong (1991) for the simple linear regression model. An extension to multiple linear regression was studied by Luong and Garrido (1992), where the asymptotic properties of this QDE were derived. They show that the QDE is fully efficient, for special choices of odd functions h_i in the distance definition, and robust for other appropriate choices of h_i.

Chapter 1 reviews the main concepts of logistic regression and its application to binary classification, when risks are either "good" or "bad".

Chapter 2 extends the review to logistic regression when the response is multinomial.

Chapter 3 discusses the main robust regression estimators encoutered in the Statistics literature and introduces to robust logistic regression.

Chapter 4 defines the minimum quadratic distance estimator for the multinomial logistic regression model (QDM). The asymptotic properties of this QDM are derived, where consistency, asymptotic normality and robustness properties are established.

Finally, Chapter 5 illustrates the proposed method with an application to the classification of a Householder data set.

Chapter 1

Models for Binary Responses

Regression methods have become an integral component of any data analysis describing the relationship between a response variable and one or more explanatory variables. It is often the case that the outcome variable is discrete, taking on two or more possible values. This chapter focuses on binary responses, that is response variables having only two categories.

The first section of the chapter introduces a family of generalized linear models. This family contains important models for categorical data as well as standard regression.

Section 1.2 introduces generalized linear models for binary response variables. The most important model of this type is the logistic regression model, based on the logit transformation of a proportion.

1.1 Generalized Linear Models

We will use the theory of generalized linear models (GLMs) introduced by Nelder and Wedderburn (1972) and detailed in McCullagh and Nelder (1989). GLMs are an extension of classical linear models. GLMs are specified by three components: a *random component*, which identifies the probability distribution of the response variable; a *systematic component*, which specifies a linear function of the explanatory variables that is used as a predictor; and a *link* describing the functional relationship between the systematic component and the expected value of the random component.

1.1.1 Components of a Generalized Linear Model

The first component of a GLM, the *random component*, refers to the response variable, Y. Suppose the N observations on Y are independent, and denote their values by $(y_1, \ldots, y_N)$. We assume that each component y_i has a distribution in the exponential family. That is, the probability density function or mass function for y_i has the form

$$f(y; \theta_i, \phi) = \exp\{\frac{[y\,\theta_i - v(\theta_i)]}{k(\phi)} + c(y, \phi)\} \quad , \quad y \epsilon \mathbb{R} \quad . \tag{1.1}$$

The parameter θ_i is called the natural parameter. The function $k(\phi)$ often has the form $k(\phi) = \frac{\phi}{w_i}$ for known weight w_i, and ϕ is called the dispersion parameter.

When ϕ is a known constant, (1.1) can be written in the following form

$$f(y; \theta_i) = \exp\{y\,Q(\theta_i) - a(\theta_i) + b(y)\} \quad , \quad y \epsilon \mathbb{R} \quad . \tag{1.2}$$

We identify $Q(\theta_i)$ in (1.2) with $\frac{\theta_i}{k(\phi)}$ in (1.1), $a(\theta_i)$ with $\frac{v(\theta_i)}{k(\phi)}$ and $b(y)$ with $c(y, \phi)$.

Formula (1.1) is useful for two-parameter families such as the normal or gamma, for which ϕ is a nuisance parameter. It is not needed for one-parameter families such as the binomial and Poisson.

General expressions for the first two moments of Y_i use terms in (1.2). Let $l(\theta_i; y) = \ln f(y; \theta_i)$ denote the natural logarithm of the likelihood function, then

$$l(\theta_i; y) = [y\, Q(\theta_i) - a(\theta_i) + b(y)]$$

and

$$\frac{\partial l}{\partial \theta_i} = y\, Q'(\theta_i) - a'(\theta_i), \qquad \frac{\partial^2 l}{\partial \theta_i^2} = y\, Q''(\theta_i) - a''(\theta_i) \quad ,$$

where $Q'(\theta_i)$ and $Q''(\theta_i)$ denote the first two derivatives of Q evaluated at θ_i.

We apply the likelihood results for $\theta = \theta_i$

$$\mathbb{E}\left(\frac{\partial l}{\partial \theta}\right) = 0 \quad \text{and} \quad -\,\mathbb{E}\left(\frac{\partial^2 l}{\partial \theta^2}\right) = \mathbb{E}\left(\frac{\partial l}{\partial \theta}\right)^2 \quad ,$$

which hold under regularity conditions satisfied by the exponential family.

From the first formula

$$\mu_i = \mathbb{E}(Y_i) = [Q'(\theta)]^{-1} a'(\theta) \quad .$$

The second formula implies

$$Var(Y_i) = [Q'(\theta)]^{-1} \left[\frac{a'(\theta)}{Q'(\theta)}\right]' \quad .$$

Let $x_{i1}, \ldots, x_{ip}$ denote values of p explanatory variables for the ith observation. The *systematic component*, the second component of a GLM, relates parameters $\{\eta_i\}$ to the explanatory variables using a linear predictor

$$\eta_i = \sum_{j=1}^{p} \beta_j x_{ij} \quad , \quad i = 1, \ldots, N \quad .$$

In matrix form

$$\boldsymbol{\eta} \;=\; \mathbf{X}\boldsymbol{\beta} \quad ,$$

where $\boldsymbol{\eta}= (\eta_1, \dots , \eta_N)^T$, $\boldsymbol{\beta}= (\beta_1, \dots , \beta_p)^T$ are model parameters, and $\mathbf{X}$ is the $N \times p$ model matrix (sometimes called design matrix).

The *link function*, the third component of a GLM, connects the expectation of Y_i to the linear predictor by

$$\eta_i = g(\mu_i) \quad , \quad i = 1, \dots , N \quad ,$$

where g is a monotone, differentiable function. Thus, a GLM links the expected value of the response to the explanatory variables through the equation

$$g(\mu_i) \;=\; \sum_{j=1}^{p} \beta_j x_{ij} \quad , \quad i = 1, \cdots , N \quad .$$

Observe that the function $g(\mu) = \mu$ gives the identity link $\eta_i = \mu_i$, specifying a linear model for the mean response.

The function g for which $g(\mu) = Q(\theta)$ in (1.2) is called the canonical link. For it, there is the direct relationship

$$Q(\theta) = \sum_{j=1}^{p} \beta_j x_{ij} \quad , \quad i = 1, \dots , N \quad ,$$

between the natural parameter and the linear predictor.

In summary, a GLM is a linear model for a transformed mean of a variable having a distribution in the natural exponential family. To illustrate the three components of a GLM, we now introduce some important GLMs for categorical response variables.

1.1.2 Logit Models

Frequently the categorical response variables have only two categories [see Cox and Snell (1989)]. In this case, the observation for each subject might be classified as a "success" or a "failure" and regularly these possible outcomes are represented by 1 and 0. The Bernoulli distribution for binary random variables specifies probabilities $\mathbb{P}(Y=1)=\pi$ and $\mathbb{P}(Y=0)=1-\pi$ for the two outcomes, for which $\mathbb{E}(Y)=\pi, \quad 0<\pi<1$.

If Y_i has a Bernoulli distribution with parameter π_i, the probability mass function is

$$f(y;\pi_i) \quad = \quad \exp\left[y\ln\left(\frac{\pi_i}{1-\pi_i}\right)+\ln(1-\pi_i)\right] \quad , \quad y=0,1 \quad , \qquad (1.3)$$

where ln is the natural logarithm. Observe that (1.3) is in the exponential family form given by (1.2). The natural parameter $Q(\pi)=\ln\left(\frac{\pi}{1-\pi}\right)$, the *ln odds* of response 1, is called the logit of π, i.e., $\text{logit}(\pi)=\ln(\frac{\pi}{1-\pi}), \quad 0<\pi<1$.

Using the component defined in (1.2) it is easy to verify that the mean and variance for Y are given by

$$\mathbb{E}(Y)=[Q'(\pi)]^{-1}a'(\pi)=\pi \quad \text{and} \quad Var(Y)=[Q'(\pi)]^{-1}\left[\frac{a'(\pi)}{Q'(\pi)}\right]'=\pi(1-\pi)\ .$$

GLMs that use the logit link $g(\pi)=\text{logit}(\pi)=\ln\left(\frac{\pi}{1-\pi}\right)$ are called logit models.

1.2 Logistic Regression

1.2.1 Introduction

There are many important research topics for which the dependent variable is qualitative. Researchers often want to analyze whether some event occurred or not, such as voting, participation in a public program, business success or failure, granting a credit card, the occurrence of mortality or of a hurricane.

The qualitative data with which we are dealing, the binary response variable, can always be coded as having two values, 0 and 1. Before predicting response values we model the probabilities that the response takes one of these two values.

Firstly, we consider the standard linear regression model with only one predictor to model this probability [see Hosmer and Lemeshow (2000)]. Let $\pi(x)$ denote the probability that $Y = 1$ when $X = x$, i.e., $\pi(x) = \mathbb{P}(Y = 1|X = x)$, then

$$\pi(x) = \mathbb{P}(Y = 1|X = x) = \beta_0 + \beta_1 x + \epsilon \quad , \tag{1.4}$$

where X is the predictor variable, β_0 and β_1 are the unknown parameters and ϵ the error, assumed normally distributed with mean zero and finite variance σ^2.

Now, since $\pi(x)$ is a probability it must lie between 0 and 1. The linear function given in (1.4) is unbounded, and cannot be used to model probabilities. There is another reason why ordinary least squares method is unsuitable. The response variable Y is a binomial random variable, consequently its variance will be a function of $\pi(x)$ and depends on X. The assumption of equal variance (homoscedasticity) does not hold. Hence, the model given by (1.4) does not apply to binary dependent variables.

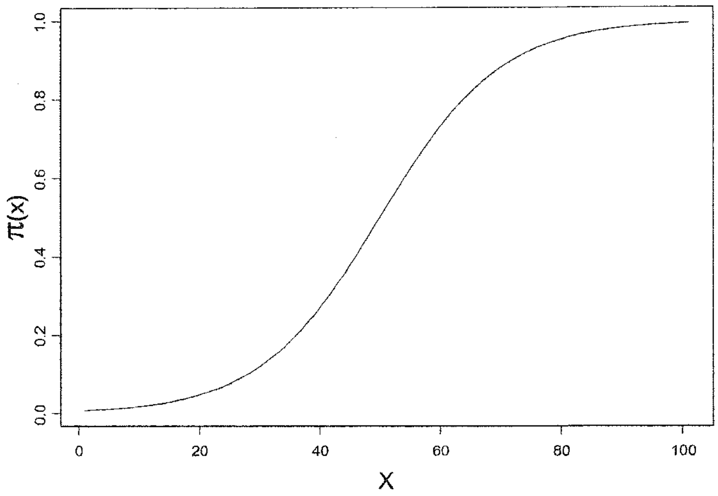


Figure 1.1: Logistic response function

The relationship between the probability $\pi(x)$ and X can often be represented by a logistic response function. It has an S-shape curve, sketched in Figure 1.1, obtained by modeling the probabilities as follows

$$\pi(x) = \mathbb{P}(Y = 1|X = x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \quad . \tag{1.5}$$

The logistic model can be generalized directly to the situation where we have several predictor variables. For instance, suppose that we have p predictor or explanatory variables. Thus, the probability $\pi(\underline{x})$ is denoted π for convenience and modeled as

$$\pi = \mathbb{P}(Y = 1|X = \underline{x}) = \frac{\exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)} \quad . \tag{1.6}$$

The equation in (1.6) is called the logistic regression function. It is nonlinear in the parameters $\beta_0, \beta_1, \ldots, \beta_p$. However, it can be linearized by the logit link. That is, instead of working directly with π we work with a transformed value of π. If π is the probability of an event happening, the ratio $\frac{\pi}{1-\pi}$ is called the odds ratio for the event. Therefore, applying the logit link discussed in the previous section, we obtain that

$$g(\pi) \;=\; \ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \quad . \tag{1.7}$$

Note that, the logit link produces a linear function of the parameters $\beta_0, \beta_1, \ldots, \beta_p$, and also that while the range of values of π in (1.4) is between 0 and 1, the range of values of $\ln\left(\frac{\pi}{1-\pi}\right)$ in (1.7) is between $-\infty$ and $+\infty$, which makes the logit more appropriate for linear regression fitting.

1.2.2 Estimation of Parameters

In linear regression models the method used most often for estimating unknown parameters is least squares. Under the usual assumptions for linear regression the method of least squares yields estimators with a number of desirable statistical properties. Unfortunately, when the method of least squares is applied to a model with a binary outcome the estimators no longer have these properties. The most commonly used method for parameter estimation of a logistic regression model is the method of maximum likelihood (MLE).

[i] The General Case

Following the general ideas developed by Pregibon (1981), we consider a single binomial response $Y \sim B(n, \pi)$. If we let $\theta = \text{logit}(\pi) = \ln\left(\frac{\pi}{1-\pi}\right)$, the probability mass function of Y can be written as

$$f(y;\theta) = \exp\{y\theta - a(\theta) + b(y)\} \quad , \quad y = 0, 1, \ldots, n$$

with $a(\theta) = n\ln(1+\exp(\theta)), \quad b(y) = \ln\binom{n}{y} \quad .$

The score and information functions for a single observation are given by

$$s(\theta; y) = \frac{\partial}{\partial\theta} l(\theta; y) = y - a'(\theta) = y - n\pi \tag{1.8}$$

and

$$v(\theta; y) = -\frac{\partial}{\partial\theta} s(\theta; y) = a''(\theta) = n\pi(1-\pi) \quad .$$

Standard results yield $\mathbb{E}[s(\theta; y)] = 0$ (or equivalently $\mathbb{E}(Y) = n\pi = a'(\theta)$) and $\text{Var}(Y) = n\pi(1-\pi) = a''(\theta)$. Also, since $s(\hat{\theta}; y) = 0$ at the maximum likelihood estimate (MLE) $\hat{\theta}$, we have $\hat{\theta} = a'^{-1}(y) = \text{logit}(\frac{y}{n})$ as the MLE of θ based on a single binomial observation $0 < y < n$.

Observe that $\text{logit}(\frac{y}{n}) = \ln(\frac{y}{n-y})$ is not defined when $y = 0$ or n. The adjusted value

$$\ln\left(\frac{y+\frac{1}{2}}{n-y+\frac{1}{2}}\right)$$

called an empirical logit, is a less biased estimator of the true logit [see Cox and Snell (1989)].

Given a sample of N independent binomial responses $Y_i \sim B(n_i, \pi_i)$, the log likelihood function for the sample is the sum of individual log likelihood contributions

$$l(\boldsymbol{\theta}; \mathbf{y}) \;=\; \sum_{i=1}^{N} l(\theta_i; y_i) = \sum_{i=1}^{N} \{y_i\theta_i - a(\theta_i) + b(y_i)\} \quad . \tag{1.9}$$

[ii] The Logistic Regression Model

The likelihood function $l(\boldsymbol{\theta}; \mathbf{y})$ given in (1.9) is over-specified, since there are as many parameters as observations.

Given a set of p explanatory variables $X_1, \ldots, X_p$, $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_N)^T$ the $N \times p$ model matrix, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ unknown parameters, the logistic regression model utilizes the relationship

$$\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}$$

as the description of the systematic component of the response $\mathbf{Y}$.

In terms of the p-dimensional parameter $\boldsymbol{\beta}$, we have the log likelihood function

$$l(\mathbf{X}\boldsymbol{\beta};\mathbf{y}) \;=\; \sum_{i=1}^{N} l(x_i^T\boldsymbol{\beta};\mathbf{y}) = \sum_{i=1}^{N}\{y_i x_i^T\boldsymbol{\beta} - a(x_i^T\boldsymbol{\beta}) + b(y_i)\} \quad . \qquad (1.10)$$

The MLE $\hat{\boldsymbol{\beta}}$ maximizes (1.10) and is solution to $\frac{\partial}{\partial\boldsymbol{\beta}} l(\mathbf{X}\boldsymbol{\beta};\mathbf{y}) = \mathbf{0}$.

In particular, $\hat{\boldsymbol{\beta}}$ satisfies the system of equations

$$\sum_{i=1}^{N} x_{ij}(y_i - a'(x_i^T\hat{\boldsymbol{\beta}})) = 0 \quad , \quad j = 1,\ldots,p \quad . \qquad (1.11)$$

Using this equation we get

$$a'(x_i^T\hat{\boldsymbol{\beta}}) = n_i \frac{\exp(x_i^T\hat{\boldsymbol{\beta}})}{1+\exp(x_i^T\hat{\boldsymbol{\beta}})} = n_i\hat{\pi}_i \quad , \quad i = 1,\ldots,N \; ,$$

and Eq. (1.8) becomes

$$\mathbf{s} = \mathbf{y} - a'(\mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{y} - \mathbf{n}\hat{\boldsymbol{\pi}}, \quad \text{where } \mathbf{n}\hat{\boldsymbol{\pi}} = (n_1\hat{\pi}_1,\ldots,n_N\hat{\pi}_N)^T.$$

Therefore, the matrix formulation of the likelihood equations (1.11) is

$$\mathbf{X}^T\mathbf{s} = \mathbf{X}^T(\mathbf{y} - \mathbf{n}\hat{\boldsymbol{\pi}}) = \mathbf{X}^T(\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{0} \quad .$$

These equations, although very similar to their normal theory counterparts, are nonlinear in $\hat{\boldsymbol{\beta}}$, and iterative methods are required to solve them.

Generally the Newton-Raphson method is employed. This involves first determining $\frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{X}^T\mathbf{s}$, which is equivalent to computing $-\frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{X}^T a'(x_i^T\boldsymbol{\beta})$.

But $-\frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{X}^T a'(x_i^T\boldsymbol{\beta}) = -[\frac{\partial}{\partial\boldsymbol{\beta}}a'(x_i^T\boldsymbol{\beta})]\mathbf{X}$ and $\frac{\partial}{\partial\boldsymbol{\beta}}a'(x_i^T\boldsymbol{\beta}) = \mathbf{X}^T\mathbf{V}$ where $\mathbf{V}$ is a diagonal matrix with elements $a''(x_i^T\boldsymbol{\beta})$, usually denoted by $\mathbf{V} = \text{diag}\{a''(x_i^T\boldsymbol{\beta})\}$. Thus, $-\frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{X}^T\mathbf{s} = \mathbf{X}^T\mathbf{V}\mathbf{X}$.

This leads to the iterative scheme

$$\boldsymbol{\beta}^{t+1} = \boldsymbol{\beta}^t + (\mathbf{X}^T\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{s} \quad , \quad t = 0, 1, \ldots \quad , \tag{1.12}$$

where both $\mathbf{V}$ and $\mathbf{s}$ are evaluated at $\boldsymbol{\beta}^t$. The value of convergence of this process is denoted by $\hat{\boldsymbol{\beta}}$, and the fitted values $n_i\hat{\pi}_i$ by $\hat{y}_i$. The estimated variance of y_i is then $v_{ii} = n_i\hat{\pi}_i(1 - \hat{\pi}_i)$.

A most useful way to view the iterative process described above is by the method of iteratively reweighted least-squares.

The value $\boldsymbol{\beta}^{t+1}$ obtained in equation (1.12) can also be expressed as

$$\boldsymbol{\beta}^{t+1} = (\mathbf{X}^T\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}(\mathbf{X}\boldsymbol{\beta}^t + \mathbf{V}^{-1}\mathbf{s}) \quad , \quad t = 0, 1, \ldots \quad . \tag{1.13}$$

Defining the pseudo observation vector $\mathbf{z}^t = \mathbf{X}\boldsymbol{\beta}^t + \mathbf{V}^{-1}\mathbf{s}$, equation (1.13) becomes

$$\boldsymbol{\beta}^{t+1} = (\mathbf{X}^T\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}\mathbf{z}^t \quad , \quad t = 0, 1, \ldots \quad .$$

At convergence, we have $\mathbf{z} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{V}^{-1}\mathbf{s}$. Thus, we can write the MLE of $\boldsymbol{\beta}$ as $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}\mathbf{z}$. This form of the estimate provides the basis for the extension of the general theory of standard linear models.

Remark 1.

In general we will assume that each explanatory variable is at least interval scaled. If some of the explanatory variables are discrete, nominal scaled variables such as race, sex, treatment group, and so forth, then it is inappropriate to include them in the model as if they were interval scaled. This is because the numbers used to represent the various levels are merely identifiers, and have no numerical significance. In this situation the method of choice is to use a collection of dummy variables defined as follows.

A categorical explanatory variable X with k possible categories $1, \ldots, k$ will generally be coded by a dummy vector with $q = k - 1$ component $X_{(1)}, \ldots, X_{(q)}$. If 0-1 dummies are used, which is shortly referred to as dummy coding, then $X_{(j)}$ is defined by

$$X_{(j)} = \begin{cases} 1, & \text{if category } j \text{ is observed} \\ 0, & \text{otherwise} \end{cases} \quad , \quad j = 1, \ldots, q \quad .$$

If the k-th category, the reference category, is observed, then $\mathbf{X}$ is the zero vector.

An alternative coding scheme, which is referred to as effect coding, is defined by

$$X_{(j)} = \begin{cases} 1, & \text{if category } j \text{ is observed} \\ -1, & \text{if category } k \text{ is observed} \\ 0, & \text{otherwise} \end{cases} \quad , \quad j = 1, \ldots, q \quad .$$

In the case of effect coding, the reference category k is given by the vector $(-1, \ldots, -1)$ instead of the zero vector.

[iii] Other Methods of Estimation

The method of maximum likelihood described above is the estimation method used in the logistic regression routines of most software packages. However, two other methods have been and may still be used for estimating the coefficients. These methods are: (1) noniterative weighted least squares, and (2) discriminant function analysis.

A linear model approach to the analysis of categorical data was proposed by Grizzle, Starmer, and Koch (1969), which uses estimators based on noniterative weighted least squares. They have proved that the logistic regression model is an example of a very general class of models that can be handled with their methods. We have shown above that the maximum likelihood estimators are usually calculated using an iterative reweighted least squares algorithm, and thus are also least squares estimators.

The main limitation of this method is that we must have an estimate of $\pi = \mathbb{P}(Y = 1|X = x)$ which is not zero or 1 for most values of x in the data set. With many explanatory variables, or even a few continuous variables this is not likely to be true.

The discriminant function approach to estimation of the logistic coefficients is based on the assumption that the distribution of the explanatory variables, given the value of the outcome variable, is normal.

Two points should be kept in mind: (1) the assumption of normality will rarely, if ever, be satisfied because of the frequent occurrence of binary explanatory variables, and (2) the discriminant function estimators of the coefficients for nonnormally distributed explanatory variables, especially binary variables, will be biased away from zero when the true coefficient is nonzero. For these reasons its use is not recommend. However, these estimators are of some historical importance as they were used in a number of classical papers in the applied literature, such as Cornfield (1962). These estimators are easily computed and should be adequate for a preliminary examination of our data.

1.2.3 Logistic Regression Diagnostics

Once the logistic regression model has been fitted, that is, we have the MLE $\hat{\boldsymbol{\beta}}$, certain diagnostic measures can be examined for the detection of outliers, high leverage points, influential observations, and other model deficiencies.

We want useful and informative diagnostic measures. These measures should readily identify observations that are not well explained by the model, as well as those dominating some important aspect of the fit. In some cases, this analysis may reveal systematic departures of the data from the model, though, in general this is not expected.

The notion of outlying, leverage and influential points, originally proposed in classical linear regression has been extended to the logistic regression. Loosely, an outlying case in logistic regression refers to a badly predicted observation, whereas a leverage point represents an 'extreme' value in the X space. An influential observation is by definition one having a great impact on the fitted model.

For the logistic regression model, the basic fundamental element for the identification of outlying and influential points are a residual vector and a projection matrix.

Before explainning residuals and projection matrices we define the important concept of Deviance.

Definition 1.2.1. [Deviance]

In logistic regression, comparisons of observed to predicted values are based on the log likelihood function defined in (1.10). To better understand this comparison, it is helpful conceptually if we think of an observed value of the response variable as also being a predicted value resulting from a saturated model. A saturated model is one that contains as many parameters as there are data points.

The comparison of observed to predicted values using the likelihood function is based on the following expression

$$D = -2\ln\left[\frac{\text{(likelihood of the current model)}}{\text{(likelihood of the saturated model)}}\right] \quad . \tag{1.14}$$

The quantity inside the large brackets in the expression above is called the likelihood ratio. The reason for using minus twice its ln is mathematical and is necessary to obtain a quantity whose distribution is known and thus can be used for hypothesis testing purposes. Such a test is called the likelihood ratio test. Using (1.10), (1.14) becomes

$$\begin{aligned} D &= -2\{l(\mathbf{X}\hat{\boldsymbol{\beta}};\mathbf{y}) - l(\hat{\boldsymbol{\theta}};\mathbf{y})\} \\ &= \sum_{i=1}^{N} 2\{y_i \ln\Big(\frac{y_i}{n_i\hat{\pi}(x_i)}\Big) + (n_i - y_i)\ln\Big(\frac{n_i - y_i}{n_i(1-\hat{\pi}(x_i))}\Big)\} \quad , \qquad (1.15) \end{aligned}$$

where $l(\hat{\boldsymbol{\theta}};\mathbf{y})$ refers to the maximum of the log likelihood function based on fitting each point exactly, i.e., $\hat{\theta}_i = \text{logit}\big(\frac{y_i}{n_i}\big)$, for $0 < y_i < n_i$.

The statistic D in (1.15) is called as the *deviance.* The deviance for logistic regression plays the same role that the residual sum of squares plays in linear regression.

It is often claimed that D is asymptotically or approximately distributed as a chi-square with $N - p$ degrees of freedom. But such a comparison is based on the assumption that $n_i \to \infty$ for each i. That is, each n_i needs to be sufficiently large before D can be assumed to have approximately a chi-square distribution. Obviously D should not be used in this manner when, in particular, there are no repeated observations with the same combinations of explanatory values. A discussion more detailed can be found in Hosmer and Lemeshow (2000), and McCullagh and Nelder (1989).

[i] Residuals

Residuals measures the departure of fitted values from observed values of the response variable. They can be used to detect model misspecification; to detect outliers, or observations with poor fit. Residual analysis, particularly visual analysis, can potentially indicate the nature of misspecification and ways in which it may be corrected, as well as provide a feel for the magnitude of the effect of the misspecification.

The *deviance residuals* are defined as the components of the deviance $D = \sum_{i=1}^{N} d_i^2$ given by (1.15), that is

$$d_i = \begin{cases} \pm[2\{y_i \ln\left(\frac{y_i}{n_i\hat{\pi}(x_i)}\right) + (n_i - y_i)\ln\left(\frac{n_i - y_i}{n_i(1-\hat{\pi}(x_i))}\right)\}]^{\frac{1}{2}} & \text{for} \quad 0 < y_i < n_i \quad , \\ -\{2n_i[-\ln(1-\hat{\pi}(x_i))]\}^{\frac{1}{2}} & \text{for} \quad y_i = 0 \quad , \\ \{2n_i[-\ln(\hat{\pi}(x_i))]\}^{\frac{1}{2}} & \text{for} \quad y_i = n_i \quad . \end{cases}$$

So, d_i measures the disagreement between the ith component of the log likelihood of the fitted model and the corresponding component of the log likelihood that would result if each point were fitted exactly.

The *Pearson residual* is defined as

$$r_i = \frac{s_i}{\sqrt{v_{ii}}} = \frac{y_i - n_i\hat{\pi}(x_i)}{\sqrt{n_i\hat{\pi}(x_i)(1-\hat{\pi}(x_i))}} \quad . \tag{1.16}$$

Each residual divides the difference between an observed and fitted value by the estimated standard deviation of the observed count y_i. When the index n_i is large, the Pearson residual r_i has an approximate normal distribution [see Pierce and Schafer (1986)]. If the number of parameters is small compared to the number of sample logits, Pearson residuals are treated like standard normal deviates, with absolute values larger than 2 indicating possible lack of fit.

McCullagh and Nelder (1989) express a preference for the deviance residuals because they are closer to being normally distributed than the Pearson residuals. A more compelling argument was given by Pregibon (1981), who noted that the Pearson residuals are unstable when $\hat{\pi}(x_i)$ is close to either 0 or 1.

Although Pregibon (1981) and McCullagh and Nelder (1989) prefer the deviance residual over the Pearson residual, Hosmer and Lemeshow (2000) advocate plotting a function of the square of each against $\hat{\pi}(x_i)$. Especially, the latter authors recommend plotting r_{is}^2 against $\hat{\pi}(x_i)$, where $r_{is} = \frac{r_i}{\sqrt{1-h_{ii}}}$ is called the standardized Pearson residual, r_i is the Pearson residual given in (1.16) and h_{ii} is the ith leverage value. The motivation for this plot is that $\frac{r_i^2}{1-h_{ii}}$ is the approximate change in the Pearson chi-square statistic that would result from deleting x_i. In the same manner, $\frac{d_i^2}{1-h_{ii}}$ estimates the change in the deviance that would result from the deletion of x_i, then the value of this statistic would also be plotted against $\hat{\pi}(x_i)$. Despite the fact that Pregibon (1981) noted that these estimates are not very precise, they may be adequate for detecting outliers.

[ii] Projection Matrix

For the linear regression model the well known measure of leverage is given by the diagonal elements of the projection matrix

$$\mathbf{M} = \mathbf{I} - \mathbf{H} = \mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \quad ,$$

where $\mathbf{I}$ is the $N \times N$ identity matrix and the $N \times N$ idempotent, symmetric matrix $\mathbf{H}$, with trace equal to its rank p (the number of explanatory variables), is known as the hat matrix. The ith element of $\mathbf{M}$ is given by

$$m_{ii} = 1 - h_{ii} = 1 - x_i^T(\mathbf{X}^T\mathbf{X})^{-1}x_i \quad .$$

Influential points will tend to have small values of m_{ii}, much smaller than the average value $1 - \frac{p}{N}$. Hoaglin and Welsch (1978) suggest using $m_{ii} \leq 1 - \frac{2p}{N}$ as a rough guide for determining whether a point is influential or not.

The analogue of the projection matrix for the logistic regression model will also be denoted by $\mathbf{M}$, which in its general form is given as

$$\mathbf{M} = \mathbf{I} - \mathbf{H} = \mathbf{I} - \mathbf{V}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}^T\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{\frac{1}{2}} \quad ,$$

with $\mathbf{V}$ diagonal matrix defined as $\mathbf{V} = \text{diag}\{a''(x_i^T\beta)\}$.

The usefulness of $\mathbf{M}$ appears as a consequence of the iteratively reweighted least squares formulation described earlier. In particular, as $\hat{\beta} = (\mathbf{X}^T\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}\mathbf{z}$, the vector of pseudo residuals obtained in (1.13) is given by

$$\mathbf{z} - \mathbf{X}\hat{\beta} = \{\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}\}\mathbf{z} = \mathbf{V}^{-\frac{1}{2}}\mathbf{M}\mathbf{V}^{\frac{1}{2}}\mathbf{z} \quad .$$

Using the fact that $\mathbf{z} = \mathbf{X}\hat{\beta} + \mathbf{V}^{-1}\mathbf{s}$, this can be written as $\mathbf{V}^{-1}\mathbf{s} = \mathbf{V}^{-\frac{1}{2}}\mathbf{M}\mathbf{V}^{-\frac{1}{2}}\mathbf{s}$. Premultiplication by the diagonal matrix $\mathbf{V}^{\frac{1}{2}}$ yields $\mathbf{r}_p = \mathbf{M}\mathbf{r}_p$, where $\mathbf{r}_p = \mathbf{V}^{-\frac{1}{2}}\mathbf{s}$. Thus, as in the linear model case, $\mathbf{M}$ is symmetric, idempotent and spans the residual $(\mathbf{r}_p)$ space. This suggests that small m_{ii} should be useful in detecting extreme points.

In most cases, the examination of r_{ip}, d_i and m_{ii} will call attention to outlying and influential points. For displaying these quantities, index plots are generally suggested: that is, plots of r_{ip} against i, d_i against i and m_{ii} against i. In particular cases, plots of these measures against the fitted values could prove useful.

1.2.4 Determination of Variables to Retain

In the logistic regression model, the usual method for selecting the explanatory variables that better explain the relation between the binary response variable is based on the log likelihood function.

This is analogous to the problem of variable selection in linear models. Instead of looking at the reduction in the error sum of squares we will look at the change in the log likelihood for two fitted models. The main reason for this is that in the logistic regression model, the fitting criterion is the log likelihood, whereas in least squares it is the sum of squares.

Let $l(p)$ denote the logarithm of the likelihood when we have a model with p explanatory variables. Similarly, let $l(p+q)$ be the logarithm of the likelihood for a model in which we have $p+q$ explanatory variables.

To see whether the q additional explanatory variables contribute significantly we look at $2[l(p+q) - l(p)]$. This quantity is twice the difference between the log likelihood for the two models. This difference is distributed as a chi-square variable with q degrees of freedom.

The size of this difference determines the significance of the test. A small chi-square value would lead to the conclusion that the q explanatory variables do not significantly improve the prediction of the logits, and is therefore not necessary in the model. A large chi-square value would call for the retention of the q explanatory variables in the model. The critical value is determined by the significance level of the test. This test procedure is valid when N, the number of observations available for fitting the model, is large.

The procedure described above enables to test any nested model. A set of models is said to be nested if these can be obtained as special cases of a larger model. The methodology is similar to that used in analyzing nested models in linear regression. The only difference is that here the test statistic is based on the log likelihood instead of the sum of squares.

1.2.5 Assessing the Fit of a Logistic Regression

In logistic regression there does not exist any satisfactory measure to judge the fitted model.

Some ad hoc measures have been proposed which are based on likelihood ratios [see Hosmer and Lemeshow (2000)]. Most of these are functions of the ratio of the likelihood for the model and the likelihood of the data under a binomial distribution. These measures are not particularly informative and we will consider a different procedure.

Logistic regression tries to model probabilities for the two values of Y $(0 \,\text{or}\, 1)$. To judge how well the model fits we determine the number of observations in the sample that the model is classifying correctly. This procedure requires fitting the logistic regression model to the data, and calculate the fitted logits, i.e., $\text{logit}(\hat{\pi}) = \ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right)$.

From the fitted logits we check if this quantity is positive, negative or zero. If the quantity is positive or zero, i.e., $\mathbb{P}(Y_i = 1|X_i = x_i) \geq 0.5$, for some $i = 1, \ldots, N$, we classify it in group 1, whereas if the quantity is negative, i.e., $\mathbb{P}(Y_i = 1|X_i = x_i) < 0.5$, for some $i = 1, \ldots, N$, we classify it in group 0. We then determine what proportion of the data is classified correctly.

A high proportion of correct classifications indicates that the logistic regression model works well. A low proportion of correct classifications indicates poor performance.

The observed correct classification rate should be treated with caution. In practice, if the fitted logistic regression model were applied to a new set of observations from the same population, it would likely produce a lower classification rate. The classification probability has an upward bias. The bias arises due to the fact that the same data used to fit the model, is also used to judge the performance of the model. To avoid this problem, we will simulate a large number of sample data with the same law as the original data. For each sample data generated, we calculate the correct classification rate to estimate the true correct classification rate.

The classical approach to estimate the bias of the correct classification rate is using resampling methods, such as jackknife or bootstrap.

Chapter 2

Multinomial Logistic Regression

Logistic regression is most frequently used to model the relationship between a dichotomous response variable and a set of explanatory variables. More generally, the response variable may take more than two values. Logistic regression can still be employed, by means of a multinomial logistic regression model.

Data with multinomial response variables are frequently encountered in social, biomedical and actuarial sciences.

The multinomial logistic regression model is an extension of the logistic analysis discussed in Section 1.2 where the response variable has two possible outcomes.

Various complexities arise from the extension, but the basic modeling ideas and diagnostics discussed so far carry over.

2.1 Introduction to the Model and Estimation of the Parameters

We consider the problem of estimating the probability $\mathbb{P}(Y = j|\mathbf{x})$ that an individual characterized by a vector $\mathbf{x}^T = (x_1, \dots, x_p)$, with p explanatory variables, belongs to one of g groups $G_1, \dots, G_g$.

We say that an observation $\mathbf{x}$ satisfies the logistic assumption if

$$\ln\left[\frac{\mathbb{P}(Y = j|\mathbf{x})}{\mathbb{P}(Y = g|\mathbf{x})}\right] = \mathbf{x}^T\boldsymbol{\beta}_j \quad , \quad j = 1, \dots, g-1 \quad , \tag{2.1}$$

or equivalently

$$\pi_j(\mathbf{x}) = \mathbb{P}(Y = j|\mathbf{x}) = \frac{\exp(\mathbf{x}^T\boldsymbol{\beta}_j)}{1 + \sum_{l=1}^{g-1}\exp(\mathbf{x}^T\boldsymbol{\beta}_l)} \quad , \quad j = 1, \dots, g \quad , \tag{2.2}$$

with $\boldsymbol{\beta}_j^T = (\beta_{1j}, \dots, \beta_{pj})$ a vector of unknown parameters and $\boldsymbol{\beta}_g = \mathbf{0}$ for convenience.

Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \dots, \boldsymbol{\beta}_{g-1}^T)^T$ represent the $p(g-1)$ dimensional column vector of unknown parameters.

The multinomial logistic model can be viewed as a multivariate generalized linear model under a multinomial random component with m trials and cell probabilities $\pi_1(\mathbf{x}), \dots, \pi_g(\mathbf{x})$.

Thus, the joint probability mass function is given as

$$
\begin{aligned}
f(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}) = & \exp\left\{ y_1 \ln\left[\frac{\pi_1(\mathbf{x})}{\pi_g(\mathbf{x})}\right] + \ldots + y_{g-1}\ln\left[\frac{\pi_{g-1}(\mathbf{x})}{\pi_g(\mathbf{x})}\right] + m\ln[\pi_g(\mathbf{x})] + \right. \\
& \left. \ln\left(\frac{m!}{y_1!\cdots y_g!}\right)\right\} \\
= & \exp\left\{ y_1\theta_1 + \ldots + y_{g-1}\theta_{g-1} - m\ln[1+\sum_{l=1}^{g-1}\exp(\theta_l)] + \ln\left(\frac{m!}{y_1!\cdots y_g!}\right)\right\} \\
= & \exp\{\mathbf{y}^T\boldsymbol{\theta} - a(\boldsymbol{\theta}) + b(\mathbf{y})\} \quad ,
\end{aligned}
$$

where

$$
\mathbf{y}^T = (y_1,\ldots,y_{g-1}), \quad \boldsymbol{\theta}^T = (\theta_1,\ldots,\theta_{g-1}), \quad \pi_g(\mathbf{x}) = 1 - \sum_{j=1}^{g-1}\pi_j(\mathbf{x}),
$$

$$
\theta_j = \ln\left[\frac{\pi_j(\mathbf{x})}{\pi_g(\mathbf{x})}\right], \text{ for } j = 1,\ldots,g-1, \quad a(\boldsymbol{\theta}) = m\ln[1+\sum_{l=1}^{g-1}\exp(\theta_l)],
$$

$$
\text{and} \quad b(\mathbf{y}) = \ln\left(\frac{m!}{\prod_{j=1}^{g} y_j!}\right), \quad \text{with} \quad y_g = m - \sum_{j=1}^{g-1} y_j \quad .
$$

The multinomial logistic model link function is

$$
g[\boldsymbol{\pi}(\mathbf{x})] = g[\pi_1(\mathbf{x}),\ldots,\pi_g(\mathbf{x})] = \left(\ln\left[\frac{\pi_1(\mathbf{x})}{\pi_g(\mathbf{x})}\right],\ldots,\ln\left[\frac{\pi_{g-1}(\mathbf{x})}{\pi_g(\mathbf{x})}\right]\right)^T \quad . (2.3)
$$

Assumption (2.1) is then translated into

$$
g[\boldsymbol{\pi}(\mathbf{x})] = \boldsymbol{\theta} \quad , \quad \text{where} \quad \theta_j = \sum_{z=1}^{p}\beta_{zj}x_z \quad , \quad j = 1,\ldots,g-1 \quad .
$$

Thus $\theta_j = \mathbf{x}^T\boldsymbol{\beta}_j$ expresses the systematic component of the generalized linear model.

Consider now a random sample from populations $G_1,\ldots,G_g$ and denote by N the number of distinct vectors $\mathbf{x}_i$. Let n_i be the number of observations at $\mathbf{x}_i$, for $i = 1,\ldots,N$.

To estimate β_{zj}, for $z = 1, \dots, p; \quad j = 1, \dots, g-1$, we maximize the log likelihood function

$$l(\boldsymbol{\beta}) \;=\; \sum_{i=1}^{N} \sum_{k=1}^{n_i} \ln[\pi_{j(k)}(\mathbf{x}_i)] \quad , \tag{2.4}$$

where $j(k)$ indicates the group of the kth individual with vector $\mathbf{x}_i$.

Let $\hat{\boldsymbol{\beta}}$ be the MLE estimate of $\boldsymbol{\beta}$ obtained by a Newton-Raphson iterative procedure.

In the following, q^t represents the current estimate of a particular component q of the model at iteration t, $\hat{q}$ the corresponding MLE solution and $q_i = q(\mathbf{x}_i)$, with $i = 1, \dots, N$.

Therefore, let $\mathbf{r}^T = (\mathbf{r}_1^T, \dots, \mathbf{r}_N^T)$ be the residual vector of observations with components $\mathbf{r}_i^T = (r_{1i}, \dots, r_{g-1,i})$ where $r_{ji} = y_{ji} - n_i \pi_j(\mathbf{x}_i)$, for $j = 1, \dots, g$, with y_{ji} is the number of G_j observations at x_i and $n_i = \sum_{j=1}^{g} y_{ji}$.

The Newton-Raphson scheme can then be expressed as

$$\boldsymbol{\beta}^{t+1} \;=\; \boldsymbol{\beta}^t + (\mathbf{X}^T \mathbf{V}^t \mathbf{X})^{-1} \mathbf{X}^T \mathbf{r}^t \quad , \quad t = 0, 1, \dots \quad , \tag{2.5}$$

where

$$\mathbf{X}^T = (\mathbf{X}_1^T, \dots, \mathbf{X}_N^T) \qquad \text{and} \qquad \mathbf{X}_i^T = \begin{pmatrix} \mathbf{x}_i^T & \mathbf{0}^T & \cdots & \mathbf{0}^T \\ \mathbf{0}^T & \mathbf{x}_i^T & \cdots & \mathbf{0}^T \\ . & . & \cdots & . \\ \mathbf{0}^T & \mathbf{0}^T & \cdots & \mathbf{x}_i^T \end{pmatrix} \quad ,$$

is a $(g-1) \times (g-1)p$ matrix and the diagonal matrix $\mathbf{V} = \text{diag}(\mathbf{V}_1, \dots, \mathbf{V}_N)$, where $\mathbf{V}_i = n_i[\pi_j(\mathbf{x}_i)(\delta_{jt} - \pi_t(\mathbf{x}_i))]_{jt}$ is a $(g-1) \times (g-1)$ square matrix, with δ_{jt} being Kronecker's delta.

The usual inverse of matrix $(\mathbf{X}^T \mathbf{V}^t \mathbf{X})$ may not always exist here. A generalization of the usual inverse to singular or rectangular matrices, called generalized inverse or G-inverse [see Rao and Mitra (1971), Albert (1972) and Searle (1982)], is used.

Provided that the cell probabilities $\pi_j(\mathbf{x}_i)$ are positive, the simplest generalized inverse of $\mathbf{V}_i$ is

$$\mathbf{V}_i^- \;=\; \mathrm{diag}\left\{\frac{1}{n_i\pi_j(\mathbf{x}_i)}\right\} \quad . \tag{2.6}$$

This is not the Moore-Penrose inverse [see Albert (1972)], but for most statistical calculations the choice of generalized inverse is unimportant and $\mathbf{V}_i^-$ (and hence $\mathbf{V}^-$) given by (2.6) is perhaps the simplest such inverse.

Equation (2.5) can be viewed as an iterative reweighted least squares method, where

$$\beta^{t+1} \;=\; (\mathbf{X}^T\mathbf{V}^t\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^t\mathbf{z}^t \quad , \quad t = 0, 1, \ldots \quad , \tag{2.7}$$

with working vector

$$\mathbf{z}^t \;=\; \mathbf{X}\beta^t + (\mathbf{V}^t)^{-1}\mathbf{r}^t \quad .$$

Iterative schemes (2.5) and (2.7) are identical with those obtained in univariate generalized linear models.

2.2 Multinomial Logistic Regression Diagnostics

In this section we review some multinomial logistic regression diagnostics from Lesaffre and Albert (1989a) and Williams (1987) by extending the concepts reviewed in Section 1.2.3 for more than two groups. The results are given for the general case where $n_i > 1$ at $\mathbf{x}_i$.

The principal ideas are the notion of outlying, leverage and influential points when considering several groups simultaneously.

2.2.1 Goodness of Fit Measures

The fact that there is a multinomial distribution at each $\mathbf{x}_i$ yields a first diagnostic for outlying cases, namely the criterion

$$\chi_i^2 = \hat{\mathbf{r}}_i^T \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{r}}_i = \sum_{j=1}^{g} \frac{[y_{ji} - n_i \hat{\pi}_j(\mathbf{x}_i)]^2}{n_i \hat{\pi}_j(\mathbf{x}_i)} \quad ,$$

which measures the agreement between the observed and estimated frequencies at $\mathbf{x}_i$. A large value suggests a poor fit at $\mathbf{x}_i$.

Moreover, since a multinomial logistic model is constituted of N independent multinomial vectors, the first goodness of fit *statistic* χ^2 will be defined as

$$\chi^2 = \hat{\mathbf{r}}^T \hat{\mathbf{V}}^{-1} \hat{\mathbf{r}} = \sum_{i=1}^{N} \hat{\mathbf{r}}_i^T \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{r}}_i = \sum_{i=1}^{N} \chi_i^2 = \sum_{i=1}^{N} \chi_i^T \chi_i = \boldsymbol{\chi}^T \boldsymbol{\chi} \quad ,$$

where $\boldsymbol{\chi} = \hat{\mathbf{V}}^{-\frac{1}{2}} \hat{\mathbf{r}}$ is the standardized residual vector.

A second goodness of fit statistic, *the deviance*

$$\begin{aligned} D &= -2\{l(\mathbf{X}\hat{\boldsymbol{\beta}}; \mathbf{y}) - l(\hat{\boldsymbol{\theta}}; \mathbf{y})\} \\ &= 2 \sum_{i=1}^{N} \sum_{j=1}^{g} y_{ji} \ln\Big(\frac{y_{ji}}{n_i \hat{\pi}_j(\mathbf{x}_i)}\Big) \quad , \end{aligned} \tag{2.8}$$

compares the log likelihood of the fitted model $l(\mathbf{X}\hat{\boldsymbol{\beta}}; \mathbf{y})$ against that of the complete model $l(\hat{\boldsymbol{\theta}}; \mathbf{y})$ matching the data perfectly.

D is the sum of N components, $D = \sum_{i=1}^{N} d_i^2$, where each d_i^2 measures the agreement between the observed and fitted log likelihood at $\mathbf{x}_i$ and can therefore be used for diagnosing outlying cases. When $y_{ji} = n_i$, $d_i^2 = 2\ln(1 + \chi_i^2)$, as for the binomial model [see Pregibon (1981)].

The multinomial logistic regression model *hat matrix* is given by

$$\mathbf{H} = \hat{\mathbf{V}}^{\frac{1}{2}} \mathbf{X} (\mathbf{X}^T \hat{\mathbf{V}} \mathbf{X})^{-1} \mathbf{X}^T \hat{\mathbf{V}}^{\frac{1}{2}} \quad .$$

The square matrices $\mathbf{H}$ and $\mathbf{M} = \mathbf{I} - \mathbf{H}$ are $N(g-1)$ projection block matrices, where each block $\mathbf{H}_{is}$ or $\mathbf{M}_{is}$ for $i, s = 1, \ldots, N$ is $(g-1)$ dimensional. Furthermore, $\boldsymbol{\chi} = \mathbf{M}\boldsymbol{\chi}$.

In particular, for the logistic model with $g = 2$ groups, the diagonal elements $\mathbf{M}_{ii}$ can be used to define the concept of *leverage* points. When $g = 2$, m_{ii} is a function of the explanatory variables and the fitted probabilities. Thus m_{ii} does not, as in linear regression, measure how extreme $\mathbf{x}_i$ is, in the Euclidean sense. However, in linear regression, a leverage point is also characterized by the fact that it greatly increases the variability of the estimates when omitted from the sample. This property can be exploited to define the concept of a leverage point in an multinomial logistic regression model.

The variability of $\hat{\boldsymbol{\beta}}$ is given by the volume of the asymptotic confidence ellipsoid for $\boldsymbol{\beta}$ which, up to a constant factor equal $\left|(\mathbf{X}^T\hat{\mathbf{V}}\mathbf{X})^{-1}\right|^{\frac{1}{2}}$. If $\mathbf{x}_i$ is deleted, then the volume becomes $\left|(\mathbf{X}_{(i)}^T\hat{\mathbf{V}}_{(i)}\mathbf{X}_{(i)})^{-1}\right|^{\frac{1}{2}}$, where the subscript (i) indicates that the $\mathbf{x}_i$ contribution to the corresponding matrix has been removed.

A small value of

$$\frac{\left|\mathbf{X}_{(i)}^T\hat{\mathbf{V}}_{(i)}\mathbf{X}_{(i)}\right|}{\left|\mathbf{X}^T\hat{\mathbf{V}}\mathbf{X}\right|} \approx \left|\mathbf{M}_{ii}\right| < 1$$

indicates that the deletion of $\mathbf{x}_i$ substantially increases the volume.

Thus, a point with a value of $\left|\mathbf{M}_{ii}\right|$ close to zero has a stabilizing effect on the estimated coefficients and will therefore be considered as a leverage point for the multinomial logistic regression model.

Additionally, $\left|\mathbf{M}_{ii}\right| = m_{ii}$ if $g = 2$ and $\left|\mathbf{M}_{ii}\right|$ satisfy several similar properties

(a) $0 \leq \left|\mathbf{M}_{ii}\right| < 1$, a value close to zero (unity) indicates a possible large (small) impact of x_i on the MLE,

(b) if each replicate at $\mathbf{x}_i$ is considered separately, then $|\mathbf{M}_{ii}| > (1 - \frac{1}{n_i})^{g-1}$,

(c) $N - (g-1)p \leq |\mathbf{M}_{ii}| \leq N - p$ with equality if and only if $g = 2$.

Numerical results performed by Lesaffre and Albert (1989b) suggest that $\sum_{i=1}^{N} |\mathbf{M}_{ii}|$ is always close to $N - p(g-1)$. Once more, following the same idea of Hoaglin and Welsch (1978) we obtain the following practical rule to fix or determine with accuracy leverage points:

$$|\mathbf{M}_{ii}| \leq 1 - \frac{2p(g-1)}{N} \quad .$$

2.2.2 Influence Diagnostics

An indicator for the influence of the ith observation $\mathbf{x}_i$ on the vector $\hat{\boldsymbol{\beta}}$ can be calculated by the difference $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}$, where $\hat{\boldsymbol{\beta}}_{(i)}$ is the MLE obtained from the sample without observation $\mathbf{x}_i$ and $\hat{\boldsymbol{\beta}}$ is the MLE from all observations.

If $\hat{\boldsymbol{\beta}}_{(i)}$ is substantially different from $\hat{\boldsymbol{\beta}}$, observation $\mathbf{x}_i$ may be considered influential. Measures of this type have been given by Cook (1977) for linear regression models. Since the estimation of unknown parameters requires an iterative procedure, it is computationally expensive to subsequently delete each observation and fit again the model.

Pregibon (1981) suggested the use of an approximate value $\hat{\boldsymbol{\beta}}^1_{(i)}$, called the one-step estimate. A one-step approximation of $\hat{\boldsymbol{\beta}}_{(i)}$ is obtained from equation (2.5) by leaving out $\mathbf{x}_i$, using $\hat{\boldsymbol{\beta}}^0_{(i)} = \hat{\boldsymbol{\beta}}$ as a starting point and terminating after one step. Proceeding as in Pregibon (1981), a form of $\hat{\boldsymbol{\beta}}^1_{(i)}$ is given by

$$\hat{\boldsymbol{\beta}}^1_{(i)} = \hat{\boldsymbol{\beta}} - (\mathbf{X}^T\hat{\mathbf{V}}\mathbf{X})^{-1}\mathbf{X}_i^T\hat{\mathbf{V}}_i^{\frac{1}{2}}\mathbf{M}_{ii}^{-1}\hat{\mathbf{V}}_i^{-\frac{1}{2}}\hat{\mathbf{r}}_i \quad ,$$

where $\mathbf{M}_{ii}$ is the ith block diagonal of the projection matrix evaluated at $\hat{\boldsymbol{\beta}}$.

The difference between the one-step estimate $\hat{\boldsymbol{\beta}}^1_{(i)}$ and the original estimate $\hat{\boldsymbol{\beta}}$ may be used as an indicator for the impact of observation $\mathbf{x}_i$ on the estimated parameter.

To determine the influence of observations on the estimate $\hat{\boldsymbol{\beta}}$ one has to consider all the components of $\hat{\boldsymbol{\beta}}$. Therefore it is often useful to have an overall measure, as considered by Cook (1977). An asymptotic confidence region for $\boldsymbol{\beta}$ is given by the log likelihood distance

$$-2\{l(\boldsymbol{\beta}) - l(\hat{\boldsymbol{\beta}})\} = c \quad ,$$

and is based on its asymptotic chi-square distribution with p degrees of freedom, χ^2_p. Approximating $l(\boldsymbol{\beta})$ by a second order Taylor expansion yields an approximate confidence region given by

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T[\mathrm{Cov}(\hat{\boldsymbol{\beta}})]^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \approx c \quad .$$

If $\boldsymbol{\beta}$ is replaced by the one-step estimate $\hat{\boldsymbol{\beta}}^1_{(i)}$ one gets

$$\begin{aligned} c^1(i) &= \hat{\mathbf{r}}_i^T[\hat{\mathbf{V}}_i^{-\frac{1}{2}}]^T\mathbf{M}_{ii}^{-1}[\hat{\mathbf{V}}_i^{\frac{1}{2}}]^T\mathbf{X}_i(\mathbf{X}^T\hat{\mathbf{V}}\mathbf{X})^{-1}\mathbf{X}_i^T\hat{\mathbf{V}}_i^{\frac{1}{2}}\mathbf{M}_{ii}^{-1}\hat{\mathbf{V}}_i^{-\frac{1}{2}}\hat{\mathbf{r}}_i \quad , \\ &= \boldsymbol{\chi}_i^T\mathbf{M}_{ii}^{-1}\mathbf{H}_{ii}\mathbf{M}_{ii}^{-1}\boldsymbol{\chi}_i \quad . \end{aligned} \tag{2.9}$$

Clearly, a leverage point which is outlying will be an influential case.

Note that $c^1(i)$ is composed from previously defined diagnostic elements. The generalized hat matrix $\mathbf{H}_{ii}$ for the ith observation plays an important role. Large values in $\mathbf{H}_{ii}$ will produce large values $c^1(i)$. The same holds for the residuals $\boldsymbol{\chi}_i = \hat{\mathbf{V}}_i^{-\frac{1}{2}}\hat{\mathbf{r}}_i$.

2.3 Estimation and Graphical Diagnostics

The multinomial regression diagnostics can be determined by appropriate matrix manipulations, once the quantities $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}^1_{(i)}, \boldsymbol{\chi}_i, \mathbf{H}_{ij}$ and $\mathbf{M}_{ii}$ are available.

Alternatively, the log likelihood $l(\boldsymbol{\beta})$, given by (2.4), can be expressed as

$$l(\boldsymbol{\beta}) \quad = \quad \sum_{i=1}^{N} \sum_{j=1}^{g} y_{ji} \ln[\pi_j(\mathbf{x}_i)] \quad . \tag{2.10}$$

The covariance matrix of $\mathbf{Y}^T = (\mathbf{y}_1^T, \dots, \mathbf{y}_N^T)$ with $\mathbf{y}_i^T = (y_{1i}, \dots, y_{gi})$ is given by $\Sigma = \text{diag}(\Sigma_1, \dots, \Sigma_N)$, where $\Sigma_i = n_i[\pi_j(\mathbf{x}_i)(\delta_{jt} - \pi_t(\mathbf{x}_i))]_{jt}$ with $j, t = 1, \dots, g$.

Take $\bar{\boldsymbol{\Sigma}}$, any generalized inverse of $\boldsymbol{\Sigma}$, e.g.

$$\bar{\boldsymbol{\Sigma}} = \text{diag}\Big(\frac{1}{n_1\boldsymbol{\pi}_1(x_1)}, \dots, \frac{1}{n_1\boldsymbol{\pi}_g(x_1)}, \dots, \frac{1}{n_N\boldsymbol{\pi}_1(x_N)}, \dots, \frac{1}{n_N\boldsymbol{\pi}_g(x_N)}\Big) \quad .$$

Then, an alternative expression for the Newton-Raphson scheme in (2.5) is given by

$$\boldsymbol{\beta}^{t+1} \quad = \quad \boldsymbol{\beta}^t + ([\boldsymbol{\Pi}^t]^T \bar{\boldsymbol{\Sigma}}^t \boldsymbol{\Pi}^t)^{-1} [\boldsymbol{\Pi}^t]^T \bar{\boldsymbol{\Sigma}}^t \mathbf{r}^t \quad , \quad t = 0, 1, \dots \quad , \tag{2.11}$$

where $\boldsymbol{\Pi}^t$ is the matrix of derivatives $\frac{\partial \pi^t}{\partial \boldsymbol{\beta}}$ evaluated at $\boldsymbol{\beta}^t$ and $\mathbf{r}^t = \mathbf{y} - \boldsymbol{\pi}^t$. Therefore, the diagnostics of Section 2.2 can be redefined according to (2.11) and the redefined vectors $\mathbf{Y}, \boldsymbol{\pi}$ and $\mathbf{r}$.

Suppose that the residual vector χ_i^0 is given by $\bar{\boldsymbol{\Sigma}}_i^{\frac{1}{2}} \hat{\mathbf{r}}_i$, then $[\chi_i^0]^T \chi_i^0 = \chi_i^T \chi_i$.

The hat matrix becomes

$$\mathbf{H}^0 = \bar{\boldsymbol{\Sigma}}^{\frac{1}{2}} \boldsymbol{\Pi}([\boldsymbol{\Pi}^t]^T \bar{\boldsymbol{\Sigma}}^t \boldsymbol{\Pi}^t)^{-1} \boldsymbol{\Pi}^T \bar{\boldsymbol{\Sigma}}^{\frac{1}{2}} = \bar{\boldsymbol{\Sigma}}^{\frac{1}{2}} \hat{\mathbf{Q}} \mathbf{X} (\mathbf{X}^T \hat{\mathbf{V}} \mathbf{X})^{-1} \mathbf{X}^T \hat{\mathbf{Q}}^T \bar{\boldsymbol{\Sigma}}^{\frac{1}{2}} \quad ,$$

with $\hat{\mathbf{Q}} = \text{diag}(\hat{\mathbf{Q}}_1, \dots, \hat{\mathbf{Q}}_N)$ and $\hat{\mathbf{Q}}_i = n_i[\pi_j(\mathbf{x}_i)(\delta_{jt} - \pi_t(\mathbf{x}_i))]_{jt}$ with $j = 1, \dots, g$, $t = 1, \dots, g-1$.

The projection matrix denoted by $\mathbf{M}^0$ is defined as $\mathbf{I} - \mathbf{H}^0$ and it can be proved that $|\mathbf{M}^0_{ii}| = |\mathbf{M}_{ii}|$, with

$$\mathbf{M}^0_{ii} = \mathbf{I} - \bar{\mathbf{\Sigma}}_i^{\frac{1}{2}} \hat{\mathbf{Q}}_i \mathbf{X}_i (\mathbf{X}^T \hat{\mathbf{V}} \mathbf{X})^{-1} \mathbf{X}_i^T \hat{\mathbf{Q}}_i^T \bar{\mathbf{\Sigma}}_i^{\frac{1}{2}} \quad .$$

Furthermore,

$$\hat{\boldsymbol{\beta}}^1_{(i)} \;=\; \hat{\boldsymbol{\beta}} - (\mathbf{X}^T \hat{\mathbf{V}} \mathbf{X})^{-1} \mathbf{X}_i^T \hat{\mathbf{Q}}_i^T \bar{\mathbf{\Sigma}}_i^{\frac{1}{2}} [\mathbf{M}^0_{ii}]^{-1} \boldsymbol{\chi}^0_i \quad .$$

Using a matrix identity [Rao (1965), p. 24], it is possible to see that these diagnostics and those of Section 2.2 are identical. The diagnostics can therefore be calculated more easily using the alternative expression given here.

Graphical plots are particularly interesting to highlight peculiar observations. Index plots have proved to be useful in the two group logistic model [Pregibon (1981)].

2.4 Assessing the Fit of a Multinomial Logistic Regression

A multinomial logistic regression model explains the probability that an individual characterized by a vector $\mathbf{x}$ belongs to one of g groups, with $g > 2$. To judge how well the model fits we determine the number of observations in the sample that the model is classifying correctly.

Using the procedures reviewed in previous sections we estimate the multinomial logistic model from the data and calculate the fitted multinomial logits, that is

$$g(\hat{\boldsymbol{\pi}}) = g(\hat{\pi}_1, \ldots, \hat{\pi}_g) = \left(\ln\left(\frac{\hat{\pi}_1}{\hat{\pi}_g}\right), \ldots, \ln\left(\frac{\hat{\pi}_{g-1}}{\hat{\pi}_g}\right)\right)^T \quad .$$

The simplest classification rule is to assign an individual characterized by a vector $\mathbf{x}$ to group G_j, $j = 1, \ldots, g-1$ if and only if

$$\mathbf{x}^T \hat{\boldsymbol{\beta}}_j > 0 \quad \text{and} \quad \mathbf{x}^T(\hat{\boldsymbol{\beta}}_j - \hat{\boldsymbol{\beta}}_l) > 0 \quad , \quad l = 1, \ldots, g-1 \quad .$$

In any other case we classify the individual in group G_g.

Following the ideas of Section 1.2.5 we generate a large number of sample data. For each sample data generated, we calculate the correct classification rate to estimate the true correct classification rate.

Chapter 3

Robust Regression

The linear regression model is one the most used tools in statistical analysis and the least squares method is a very popular estimation technique for this model. Unfortunately, outliers and other aberrations which appear to conflict with the model can arise and weaken the levels for confidence intervals and tests. Thus it is desirable to have robust methods which are still highly efficient in the presence of these aberrations.

Robust regression is an alternative to ordinary least squares that can be appropriately used when there is evidence that the distribution of the error term is non-normal, and/or there are outliers that affect the model.

Section 3.3 gives a brief review of the development of M-estimators or Huber estimators for linear regression and their influence function. By means of the basic concepts of influence function, we show that more refined estimators are required. General M-estimators and their influence function are also presented.

Section 3.4 presents many of the robust regression estimators that have been discussed extensively in the literature; the least trimmed squares regression, the least median squares regression and S-estimators.

Finally, Section 3.5 briefly introduces the robust logistic regression model.

3.1 Classical Least Squares Estimation

Consider the following linear model: let $\{(\mathbf{x}_i, y_i) : i = 1, 2, \ldots, N\}$ be a sequence of independent identically distributed random variables such that

$$y_i \;=\; \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i \quad , \quad i = 1, \ldots, N \quad ,$$

where

$y_i \in \mathbb{R}$ is the ith observation,
$\mathbf{x}_i \in \mathbb{R}^p$ is the ith row of the random design matrix $X_{N \times p}$,
$\boldsymbol{\beta} \in \boldsymbol{\mathcal{B}} \subset \mathbb{R}^p$ is a p-vector of unknown parameters ($p \geq 1$),
$\varepsilon_i \in \mathbb{R}$ is the ith error.

Suppose that $\boldsymbol{\mathcal{B}}$ is open and convex and that ε_i is independent of $\mathbf{x}_i$ and has a symmetric distribution U, such that $\mathbb{P}(\varepsilon_i \leq \frac{\varepsilon}{\sigma}) = U(\frac{\varepsilon}{\sigma})$, where $\sigma > 0$ is a scale parameter. Denote the corresponding density by u, defined with respect to Lebesgue's measure.

Let K be the distribution function of $\mathbf{x}_i$, with the density k with respect to Lebesgue's measure. We denote by $f_{\boldsymbol{\beta}}(\mathbf{x}, y)$ the joint density of $(\mathbf{x}_i, y_i)$, that is,

$$f_{\boldsymbol{\beta}}(\mathbf{x}, y) \;=\; \sigma^{-1} u \left(\frac{y - \mathbf{x}^T \boldsymbol{\beta}}{\sigma} \right) k(\mathbf{x}) \quad , \quad \mathbf{x} \in \mathbb{R}^p \, , \; y \in \mathbb{R} \quad ,$$

and $F_{\boldsymbol{\beta}}(\mathbf{x}, y)$ the corresponding distribution function value.

Classical estimation and test procedures in linear models are based on the well known method of least squares. Consider for a moment σ as fixed.

A least squares (LS) estimate $\mathbf{T}_N^{LS}$ of $\boldsymbol{\beta}$ is any statistic that minimizes the Euclidean norm of the residuals

$$\Gamma(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}}{\sigma} \right)^2 \quad , \quad \boldsymbol{\beta} \in \boldsymbol{\mathcal{B}} \quad ,$$

that is, $\mathbf{T}_N^{LS}$ is defined by

$$T_N^{LS} = \arg\min\{\Gamma(\boldsymbol{\beta}) | \boldsymbol{\beta} \in \boldsymbol{\mathcal{B}}\}.$$

Theorem 3.1.1. [Gauss-Markov]

Under the assumptions

(i) $\mathbb{E}(\varepsilon_i) = 0, \quad i = 1, \ldots, N,$

(ii) $\mathrm{Cov}(\varepsilon_i, \ldots, \varepsilon_N) = \sigma^2 I,$ where I is the identity matrix,

every estimable function $\mathbf{b}^T\boldsymbol{\beta}$ has a unique linear estimator which has minimum variance in the class of all unbiased linear estimator. It is given by $\mathbf{b}^T\mathbf{T}_N^{LS}$, where $\mathbf{T}_N^{LS}$ is any least squares estimator. If in addition the errors are normally distributed, then this estimator has minimum variance among all unbiased estimators.

Remark 1.

The LS estimator is optimal in the class of all unbiased estimators, only if the errors are normally distributed. Hence, the restriction to linear estimators can be justified only by normality. But many maximum likelihood estimators (MLE) [e.g., under the logistic model or for all student's t-distributions of errors, including the Cauchy] are not linear.

Remark 2.

The normal model is never exactly true and in the presence of small departures from the normality assumption on the errors, the least squares procedure loses efficiency drastically. Thus, one would prefer procedures which are only closely optimal at the normal model but which behave well under certain neighboring distributions.

3.2 Goals of Robust Regression

In regression analysis we may want to test an hypothesis regarding the unknown vector of coefficients $\boldsymbol{\beta}$, which requires estimates and their standard errors. A number of model assumptions may be violated: the distributional assumption that each error $\varepsilon_i \sim N(0, \sigma^2)$; the independence of the errors; the linear dependence on the explanatory variables. It is known that a small violation in one or more of these assumptions can lead to a large change in the least squares estimator $\hat{\boldsymbol{\beta}}$ or the estimates of standard errors based on the assumed covariance structure. Robust estimators of $\boldsymbol{\beta}$ ideally should satisfy the following goals:

1. Consistency, asymptotic normality and high efficiency of the estimators, if there are no model violations.

2. Methods for forming confidence intervals for the unknown parameters and for testing hypothesis about them.

3. Relative insensitivity of the properties in 1. and results in 2. to slight violations of the model.

4. Simplicity of the theory and ease of computation.

3.3 M-Estimators

In this section we assume the model of Section 3.1 with normal errors ($U = \Phi$).

The class of M-estimators was defined by Huber (1964, 1968) for the location model and extended by him to the regression model in 1973. A detailed investigation of M-estimators can be found in Huber (1981) and in Hampel et al. (1986).

Fundamentally, Huber (1973) proposed to compute weighted LS estimates with weights of the form

$$w_i = \min\{1, \tfrac{c}{|r_i|}\} \quad , \quad i = 1, \ldots, N \quad , \tag{3.1}$$

where $r_i = y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}$ is the ith residual and c is a positive constant. The weights thus are not fixed, but depend on the estimate. More generally, Huber proposed M-estimators $\mathbf{T}_N$ defined by

$$\mathbf{T}_N = \arg\min\{\Gamma(\boldsymbol{\beta}) | \boldsymbol{\beta} \in \boldsymbol{\mathcal{B}}\} \quad ,$$

where

$$\Gamma(\boldsymbol{\beta}) = \sum_{i=1}^{N} \rho\Big(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}}{\sigma}\Big) \tag{3.2}$$

and ρ is a convex function. When ρ is differentiable, the minimum of (3.2) can often be found by setting the partial derivatives of ρ to zero and solving for $\boldsymbol{\beta}$, i.e.

$$\sum_{i=1}^{N} \psi\Big(\frac{y_i - \mathbf{x}_i^T \mathbf{T}_N}{\sigma}\Big)\mathbf{x}_i = 0 \quad , \tag{3.3}$$

with $\psi = \rho'$ and for a fixed σ. Solutions $\mathbf{T}_N$ of (3.3) are called classical M-estimators or Huber estimators because they were the first extensions of the location M-estimators to the regression case.

Remark 1.

The least squares estimator can be defined by the function $\rho(r) = \frac{r^2}{2}$ and the least absolute deviations (LAD) by $\rho(r) = |r|$. The main advantage of LAD estimates over LS estimates is that they are not so sensitive to outliers. When there are no outliers, however, LS estimates may be more accurate.

Huber's estimator, $\mathbf{T}^H$, defined by the weights in (3.1) may be obtained by setting $\rho(r) = \rho_c(r)$ in (3.2), where $\rho'(r) = \psi(r) = \psi_c(r) = r \cdot \min\{1, \frac{c}{|r_i|}\}$, for $c > 0$ [see Figure 3.1]. It reproduces the MLE when the errors are distributed according to the distribution with density proportional to $\exp[-\rho_c(r)]$.

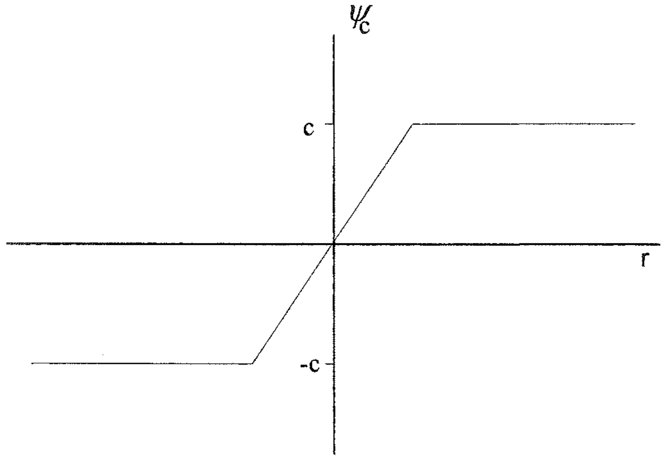


Figure 3.1: ψ_c function defining Huber's estimator

Remark 2.

We have assumed that the scale parameter $\sigma > 0$ is fixed, but in practice this parameter could be estimated.

A possible way to do this is, for a given τ, to minimize

$$\sum_{i=1}^{N} \left[\rho \left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}}{\sigma} \right) + \tau \right] \sigma \quad , \tag{3.4}$$

with respect to $\boldsymbol{\beta}$ and σ.

Setting the derivatives of (3.4) to zero and solving for $\boldsymbol{\beta}$ and σ, we obtain the following equations

$$\sum_{i=1}^{N} \psi \left(\frac{y_i - \mathbf{x}_i^T \mathbf{T}_N}{\hat{\sigma}} \right) \mathbf{x}_i \;=\; 0 \quad , \tag{3.5}$$

$$\sum_{i=1}^{N} \chi \left[\frac{y_i - \mathbf{x}_i^T \mathbf{T}_N}{\hat{\sigma}} \right] \;=\; 0 \quad , \tag{3.6}$$

where $\psi(r) = \rho'(r)$ and $\chi(r) = r\psi(r) - \rho(r) - \tau$ with $r = \frac{y_i - \mathbf{x}_i^T \mathbf{T}_n}{\hat{\sigma}}$.

Several choices of ψ and χ have been proposed, among of these $\psi(r) = \psi_c(r)$ and $\chi(r) = [\psi_c(r)]^2 - \tau$ corresponds to the proposal of Huber (1981).

Definition 3.3.1. [Influence Function]

The *influence function* (IF) of a functional T at $x \in \mathbb{R}$ under a true distribution F_κ is

$$\mathrm{IF}(x; T, F_\kappa) \;=\; \lim_{\epsilon \to 0} \frac{T[(1-\epsilon)F_\kappa + \epsilon\delta^x] - T(F_\kappa)}{\epsilon}$$

when the limit exists. δ^x is the "degenerate" probability function that assigns all its probability to x, that is, $\delta^x(x) = 1$ and $\delta^x(y) = 0$ for any $y \neq x$.

Theorem 3.3.1.

Let T_N be an M-estimator and $IF(x; T, F_\kappa)$ the influence function of T at F_κ then

$$IF(x;T,F_\kappa) \;=\; \frac{\psi(x;T(F_\kappa))}{\mathbf{M}(\psi,F_\kappa)} \quad , \quad x \in \mathbb{R} \quad ,$$

where the $p \times p$ matrix $\mathbf{M}$ is given by

$$\mathbf{M}(\psi,F_\kappa) \;:=\; -\int \left[\frac{\partial}{\partial t}\psi(x;t)\right]_{t=T(F_\kappa)} dF_\kappa(x) \quad .$$

[For proof, see Birkes and Dodge (1993)].

In order to study the robustness properties of Huber's estimator $\mathbf{T}^H$, it is necessary to compute its influence function at the model distribution $F_{\boldsymbol{\beta}}$ with density

$$f_{\boldsymbol{\beta}}(\mathbf{x},y) = \phi(y - \mathbf{x}^T\boldsymbol{\beta})k(\mathbf{x}) \quad ,$$

where the scale factor is ignored.

Using Theorem 3.3.1 with $F_\kappa = F_{\boldsymbol{\beta}}$ and $\psi_c(y - \mathbf{x}^T\boldsymbol{\beta})$, the expression for the influence function $IF(\mathbf{x}, y; \mathbf{T}^H, F_{\boldsymbol{\beta}})$ is obtained as

$$IF(\mathbf{x},y;\mathbf{T}^H,F_{\boldsymbol{\beta}}) \;=\; \frac{\psi_c(y-\mathbf{x}^T\boldsymbol{\beta})\mathbf{x}}{\mathbf{M}(\psi,F_{\boldsymbol{\beta}})} \quad , \quad \mathbf{x} \in \mathbb{R}^p \ , \ y \in \mathbb{R} \quad ,$$

where

$$\mathbf{M}(\psi,F_{\boldsymbol{\beta}}) \;=\; (\mathbb{E}\psi_c')(\mathbb{E}\mathbf{x}\mathbf{x}^T) = \left(\int \psi_c'(r)d\Phi(r)\right)\left(\int \mathbf{x}\mathbf{x}^T dK(\mathbf{x})\right) .$$

Note that this influence function depends on y only through $r := y - \mathbf{x}^T\boldsymbol{\beta}$.

Following Hampel (1973) it is possible to rewrite IF as a product of two factors, namely the (scalar) influence of the residual (IR) and the (vector valued) influence of position in factor space (IP):

$$IF(\mathbf{x}, \mathbf{x}^T\boldsymbol{\beta} + r; \mathbf{T}^H, F_{\boldsymbol{\beta}}) \quad = \quad IR(r; \mathbf{T}^H, \Phi) \cdot IP(\mathbf{x}; \mathbf{T}^H, K) \quad , \tag{3.7}$$

where

$$\begin{aligned} IR(r; \mathbf{T}^H, \Phi) &= \frac{\psi_c(r)}{(\mathbb{E}\psi_c')} \quad , \\ IP(\mathbf{x}; \mathbf{T}^H, K) &= (\mathbb{E}\mathbf{x}\mathbf{x}^T)^{-1}\mathbf{x} \quad . \end{aligned}$$

The factorization (3.7) is unique if IR is defined as the influence function of the corresponding M-estimator of location.

The influence of the residual $IR(r; \mathbf{T}^H, \Phi)$ is bounded. This is an improvement over LS estimators from the robustness point of view. But still, the influence of position $IP(\mathbf{x}; \mathbf{T}^H, K)$ in factor space is unbounded.

Thus, for a judicious choice of the function ψ, the M-estimator of $\boldsymbol{\beta}$ may attain high efficiency, relative to the LS estimator, and also be robust against large residuals. However, M-estimators are not robust to outliers in the design space. This is not a problem if the $\mathbf{x}$ matrices are in fact chosen by design to have equileverage, or at least a large proportion of points with the maximum leverage. But if the $\mathbf{x}$ matrices themselves are random or otherwise subject to errors, then the classical M-estimators may be unreliable, for the equations in (3.3) are directly affected by $\mathbf{x}$. Thus several statisticians have proposed alternatives to (3.3) of the form

$$\sum_{i=1}^{N} \eta\Big(\mathbf{x}_i, \frac{y_i - \mathbf{x}_i^T\mathbf{T}_N}{\sigma}\Big)\mathbf{x}_i \quad = \quad 0 \quad . \tag{3.8}$$

In other words the domain of the function ψ has been enlarged to include the design points, as well as the residuals. Solutions to (3.8) are also called generalized M-estimators. See Hampel, et al. (1986) for further references.

The form (3.8) restricts the M-estimators, as compared with the general formula $\sum_{i=1}^{N} \psi(\mathbf{x}_i, y_i; \mathbf{T}_N) = 0$ in two ways: $\psi(\mathbf{x}_i, y_i; \boldsymbol{\beta})$ must have the same direction as $\mathbf{x}$ and the scalar $\eta(\mathbf{x}, r)$ depends on $\boldsymbol{\beta}$ only through $r = y - \mathbf{x}^T \boldsymbol{\beta}$.

In order to study the robustness of the estimator $\mathbf{T}_N$, investigate the functional $T(F)$ corresponding to the M-estimator obtained from (3.8) is the solution of

$$\int \eta(\mathbf{x}, y - \mathbf{x}^T T(F))\, \mathbf{x}\, dF(\mathbf{x}, y) \quad = \quad 0 \quad .$$

Define

$$\mathbf{M}(\eta, F) := \int \eta'(\mathbf{x}, y - \mathbf{x}^T T(F))\, \mathbf{x}\mathbf{x}^T\, dF(\mathbf{x}, y) \quad ,$$

where $\eta' := \frac{\partial}{\partial r} \eta(\mathbf{x}, r)$.

Then the influence function of T at a distribution F is given by

$$IF(\mathbf{x}, y; T, F) \quad = \quad \eta(x, y - \mathbf{x}^T T(F)) \mathbf{M}^{-1}(\eta, F)\, \mathbf{x} \quad .$$

Maronna and Yohai (1981) prove that these estimators are consistent and asymptotically normal.

3.4 Other Robust Estimators

Definition 3.4.1. [**The Breakdown Point**]

Consider any sample of N data points $Z = \{(\mathbf{x}_1^T, y_1), \ldots, (\mathbf{x}_N^T, y_N)\}$ and let T be a regression estimator such that $T(Z) = \hat{\boldsymbol{\beta}}$.

Now consider all possible corrupted samples Z^* that are obtained by replacing any m of the original data points by arbitrary values. Denote by bias$(m; T, Z)$

the maximum bias that can be caused by such a contamination:

$$\text{bias}(m; T, Z) = \sup_{Z^*} \|T(Z^*) - T(Z)\| \quad ,$$

where the supremum is over all possible Z^*. Clearly, the function bias is non-decreasing in m. If bias$(m; T, Z)$ is infinite, this means that m outliers can have an arbitrarily large effect on T. Therefore, the (finite-sample) *breakdown point* of the estimator T at the sample Z is defined as

$$\epsilon_N^*(T, Z) = \min\{\tfrac{m}{N}; \text{bias } (m; T, Z) \text{ is infinite } \} \quad .$$

3.4.1 Least Trimmed Squares Regression

Least trimmed squares regression (LTS), introduce by Rousseeuw (1984), is a highly robust method for fitting a linear regression model. The LTS estimator $\hat{\boldsymbol{\beta}}_{LTS}$ minimizes the sum of the q smallest squared residuals, that is,

$$\hat{\boldsymbol{\beta}}_{LTS} = \arg\min_{\hat{\boldsymbol{\beta}}} \sum_{i=1}^{q} r_{(i)}^2 \quad ,$$

where $r_{(1)}^2, r_{(2)}^2, \ldots, r_{(N)}^2$ are the ordered squared residuals, from smallest to largest, and q must be determined. The value of q is often set to be slightly larger than half of N. Rousseeuw and Leroy (1987) suggest that q may be selected as $q = [N(1-\alpha)] + 1$, with α being a proportion.

The LTS estimator has the highly attractive robustness property that its breakdown point is approximately $\frac{1}{2}$ (if q is the right fraction of N). The breakdown point of a regression estimator is the largest fraction of data which may be replaced by arbitrarily large values without making the Euclidean norm $||\hat{\boldsymbol{\beta}}||$ of the resulting estimate tend to ∞, where

$$||\hat{\boldsymbol{\beta}}||^2 = \sum_{i=1}^{p} \hat{\beta}_i^2 \quad .$$

Any estimator with breakdown point approximately $\frac{1}{2}$ is called a *high breakdown* point estimator. Thus, the LTS estimator is a high breakdown point regression estimator.

The high breakdown point of the LTS estimator means that the values $\mathbf{x}_i^T\hat{\boldsymbol{\beta}}_{LTS}$, $i = 1, \ldots, N$, fit the bulk of the data well, even when the bulk of the data may consist of only somewhat more than 50% of the data. Correspondingly, the residuals $r_i = y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}_{LTS}$ will reveal the outliers quite clearly.

On the other hand, this high breakdown point estimator is highly inefficient for regression models with normal errors, but provides very robust starting values to be followed by more efficient re-weighted procedures.

3.4.2 Least Median Squares Regression

An idea quite similar to LTS regression is least median squares (LMS). Rather than minimizing the sum of the squared residuals, as in least squares regression, least median squares [Rousseeuw (1984)] minimizes the median of the squared residuals.

Least median of squares regression has a very high breakdown point of almost 50%. That is, almost half of the data can be corrupted in an arbitrary fashion and the LMS estimator continue to explain the majority of the data. At the present time this property is virtually unique among the robust regression methods available. However, least median of squares is statistically very inefficient.

3.4.3 S-Estimators

Both the LTS and the LMS are defined by minimizing a robust measure of the scatter of the residuals. Generalizing this Rousseeuw and Yohai (1984) introduced another class of high breakdown point estimators based on the minimization of the dispersion of the residuals, so called S-estimators, corresponding to

$$\underset{\boldsymbol{\beta}}{\text{Minimize}}\, s[r_1(\boldsymbol{\beta}), \ldots, r_N(\boldsymbol{\beta})] \quad ,$$

where $r_1(\boldsymbol{\beta}), \ldots, r_N(\boldsymbol{\beta})$ denote the N residuals for a given parameter $\boldsymbol{\beta}$.

The dispersion $s[r_1(\boldsymbol{\beta}), \ldots, r_N(\boldsymbol{\beta})]$ is defined implicitely, for a given K, as the solution of

$$\frac{1}{N}\sum_{i=1}^{N} \rho\left(\tfrac{r_i}{s}\right) = K \quad ,$$

where the function ρ must satisfy the following conditions:

1. ρ is symmetric and continuously differentiable, and $\rho(0) = 0$.
2. There exists $c > 0$ such that ρ is strictly increasing on $[0, c]$ and constant on $[c, \infty)$.

A function ρ must be selected; Rousseeuw and Yohai (1984) suggest

$$\rho(x) = \begin{cases} \frac{x^2}{2} - \frac{x^4}{2c^2} + \frac{x^6}{6c^4} & \text{for } |x| \le c \\ \frac{c^2}{6} & \text{for } |x| > c \end{cases} \quad . \tag{3.9}$$

The selection of c, which determines K, involves a tradeoff between breakdown point and efficiency.

Table 19 of Rousseeuw and Leroy (1987, p. 142) gives the asymptotic efficiency of the S-estimators corresponding to the function ρ defined in (3.9) for different values of the breakdown point.

A detailed study of the properties of LTS, LMS and S-estimators briefly reviewed in this Chapter can be founded in Rousseeuw and Leroy (1987).

3.5 Introduction to Robust Logistic Regression

Consider a generalized linear model with a Bernoulli independent variable Y, (i.e., Y takes values 0 and 1), p explanatory variables $X_1, \ldots, X_p$, which may be discrete or continuous and a link function G. Then, if $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_N)^T$ is the $N \times p$ design matrix, for any $\mathbf{x}^T = (x_1, \ldots, x_p) \in \mathbb{R}^p$, we have

$$\mathbb{P}(Y = 1 | \mathbf{X} = \mathbf{x}) \quad = \quad G(\mathbf{x}^T \boldsymbol{\beta}) \quad , \tag{3.10}$$

for some $\boldsymbol{\beta} \in \mathbb{R}^p$. We will suppose that G is a continuous and strictly increasing distribution function. An example of G may be the logistic distribution

$$G(t) \quad = \quad \frac{\exp(t)}{1 + \exp(t)} \quad , \quad t \in \mathbb{R} \quad , \tag{3.11}$$

or the probit function Φ corresponding to the standard normal distribution function.

Consider a sample $\mathbf{z}_1 = (\mathbf{x}_1^T, y_1), \ldots, \mathbf{z}_N = (\mathbf{x}_N^T, y_N)$ of N independent observations, where $\mathbf{x}_i^T = (x_{i1}, \ldots, x_{ip})$ for $i = 1, \ldots, N$. The $\mathbf{x}_i$'s may be fixed or random. The maximum likelihood estimate (MLE) which is obtained by maximizing

$$\sum_{i=1}^{N} \{ y_i \ln[G(\mathbf{x}_i^T \boldsymbol{\beta})] + (1 - y_i) \ln[1 - G(\mathbf{x}_i^T \boldsymbol{\beta})] \} \quad , \tag{3.12}$$

gives an asymptotically efficient procedure for estimating $\boldsymbol{\beta}$. The usual procedure to estimate $\boldsymbol{\beta}$ is an iterative process, such as Newton-Raphson, as used in Section 1.2.

Considering the deviances

$$D_i(\boldsymbol{\beta}) = y_i\{-\ln[G(\mathbf{x}_i^T\boldsymbol{\beta})]\} + (1-y_i)\{-\ln[1-G(\mathbf{x}_i^T\boldsymbol{\beta})]\} \quad ,$$

the MLE may be defined, alternatively, by the minimizing

$$\sum_{i=1}^{N} D_i(\boldsymbol{\beta}) \quad . \tag{3.13}$$

However, the MLE is extremely sensitive to the presence of anomalous data in the sample. Pregibon (1981, 1982) has exemplified this sensitivity. Pregibon (1981), Cook and Weisberg (1982), Johnson (1985), Hosmer and Lemeshow (2000) and McCullagh and Nelder (1989), have proposed procedures to identify observations which are influential for estimating $\boldsymbol{\beta}$.

Pregibon (1982) defined robust estimates by modifying the MLE goal function given in (3.13). He proposed to minimize

$$\sum_{i=1}^{N} \rho(D_i(\boldsymbol{\beta})) \quad , \tag{3.14}$$

where ρ is a suitable monotone loss function with a slower increase than the identity. He suggested to use for ρ a Huber type function in the family given by

$$\rho(t) \quad = \quad \begin{cases} t & \text{if } t \leq c \quad , \\ 2(tc)^{\frac{1}{2}} - c & \text{if } t > c \quad . \end{cases} \tag{3.15}$$

However, the estimators defined by the minimization of (3.14) are not consistent. Actually, they are asymptotically biased when the model (3.10) holds.

Copas (1988) found an approximate expression for the bias of these estimates, and he showed that it may be non-negligible. Moreover, if the ρ function is chosen in the family (3.15), the resulting estimate is not robust against outliers with high leverage explanatory variables.

Stefanski, Carroll and Ruppert (1986) proposed a class of robust estimates for generalized linear models. They obtained Hampel's optimal bounded influence estimators: these minimize some functional of the asymptotic covariance matrix, subject to a bound of some norm of the influence curve. Künsch, Stefanski and Carroll (1989) found similar optimal estimators but with the additional constraint of conditional unbiasness. These two classes of bounded influence estimates require the estimation of a robust covariance matrix of **X**. Carroll and Pederson (1993) studied the bias of these estimators for small uncontaminated samples. They found that there are estimates in the Mallows subclass [Hampel et al. (1986), Section 6.3] which have a bias comparable with the MLE.

Copas (1988) proposed a robust estimation procedure which is based on a logistic regression model which contains the outlier generation mechanism. The estimation procedure is maximum likelihood applied to this model.

Bianco and Yohai (1996) have proposed a class of M-estimators which can be thought as a Fisher consistent version of the estimators given by (3.14). These estimators are defined by the minimization of

$$\sum_{i=1}^{N}\{\rho[D_i(\boldsymbol{\beta})] + J[G(\mathbf{x}_i^T\boldsymbol{\beta})] + J[1 - G(\mathbf{x}_i^T\boldsymbol{\beta})]\} \quad , \qquad (3.16)$$

where ρ is a bounded, differentiable and nondecreasing function and

$$J(t) \quad = \quad \int_0^t \psi(-\ln u)du \quad , \qquad (3.17)$$

where $\psi(t) = \rho'(t)$. Bianco and Yohai (1996) have shown that this choice of J makes the M-estimator Fisher consistent.

The fact that ρ is bounded makes these estimators qualitatively robust in Hampel's sense; that is, a small fraction of arbitrary outliers in the sample has a small effect on the estimate. An advantage of these estimators is that they do not need a robust estimate of the scatter matrix of the **X**'s. The estimates defined by the minimization of (3.16) are a natural generalization of the M-estimates with bounded ρ function used for regression. Since these M-estimates for regression have good robustness properties [e.g., a high breakdown point, see Yohai (1987)].

In particular, the family of ρ functions studied by Bianco and Yohai (1996) is given by

$$\rho_c(t) \quad = \quad \begin{cases} t - \frac{t^2}{2c} & \text{if } t \leq c \\ \frac{c}{2} & \text{if } t > c \end{cases} \quad , \tag{3.18}$$

where c is a positive number.

The ρ function corresponding to the MLE is just the identity, which is an unbounded function. The identity function could be robustified by truncating it for values larger than a constant c . The correction term $\frac{t^2}{2c}$ in (3.18) makes the truncation process smoother. It is clear that when c tends to infinity the corresponding ρ_c function converges to the identity and therefore the corresponding M-estimate approaches the MLE. This family is mainly used because of the simplicity of the corresponding J functions. However, similar results could be obtained with other ρ functions using different truncation schemes.

Chapter 4

Robust Logistic Regression Model

The principal result in this chapter is the proposal of a new robust regression estimator for the logistic regression model, both in the binary and multinomial response cases. Its asymptotic properties are also studied.

This chapter is composed of two sections. Section 4.1 extends the quadratic distance estimators (QDE) for the multiple linear regression, suggested by Luong and Garrido (1992), to the logistic regression model, where the response variable is binary. The asymptotic properties and the influence function have also been derived for this particular model.

An extension of the QDE to the multinomial logistic regression model is presented in the last section.

4.1 Robust Quadratic Distance Estimators for Logistic Regression

A minimum distance method based on the quadratic distance (QD) for transforms was introduced by Luong and Thompson (1987). Following the same idea the minimum distance estimator based on the QD was introduced in the simple linear regression model by Luong (1991). An extension to multiple linear regression was studied by Luong and Garrido (1992), where the asymptotic properties of this QD estimator were derived. The QD estimator was shown to be efficient for a specific choice of odd function h and also shown to be robust for appropriate choices of h.

In this section, minimum distance estimators based on the QD will be introduced for the logistic regression model.

Let $\mathbf{x}_i^T = (x_{i1}, \ldots, x_{ip}) \in \mathbb{R}^p$ be a vector of p explanatory variables which may be discrete or continuous and $\mathbf{X} = (\mathbf{X}_1^T, \ldots, \mathbf{X}_N^T)^T$ be a $N \times p$ design matrix with rank $p \leq N$, where $\mathbf{x}_i \neq 0$, for $i = 1, \ldots, N$. Denote by $n^* = \lfloor \frac{N}{2} \rfloor + \lfloor \frac{p}{2} \rfloor$, where $\lfloor z \rfloor$ stands for the largest integer less than or equal to $z \in \mathbb{R}$.

We consider the logistic regression model for binary responses and N independent random variables Y_i, which have a binomial distribution with index n_i and probability $\pi(\mathbf{x}_i)$. These are denoted by $Y_i \sim \text{Binomial}(n_i, \pi(\mathbf{x}_i))$, where n_i is a known positive integer, $\pi(\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1+\exp(\mathbf{x}_i^T \boldsymbol{\beta})}$, and $\boldsymbol{\beta}$ is a vector of p unknown parameters.

From Christmann (1994) we can define relative frequencies P_i as follows.

Definition 4.1.1.

Let y_i be observations from not necessarily independent random variables Y_i, where Y_i have a binomial distribution with index n_i and probability $\pi(\mathbf{x}_i)$, then the relative frequencies P_i, for $i = 1, \ldots, N$, are defined as

$$P_i \;=\; \begin{cases} \frac{1}{2n_i} & \text{if } \; Y_i = 0 \quad , \\ \frac{Y_i}{n_i} & \text{if } \; 1 \leq Y_i \leq n_i - 1 \quad , \\ 1 - \frac{1}{2n_i} & \text{if } \; Y_i = n_i \quad . \end{cases} \tag{4.1}$$

Assumptions:

Under the above definition it is assumed that

(a) there exist $\pi(\mathbf{x}_i) \in (0,1)$, for $1 \leq i \leq N$, such that if $\min\limits_{1 \leq i \leq N} n_i \longrightarrow \infty$

$$(P_1, \ldots, P_N) \longrightarrow (\pi(\mathbf{x}_1), \ldots, \pi(\mathbf{x}_N)) \;, \quad \text{almost surely} \;, \tag{4.2}$$

(b) there exists exactly one vector $\boldsymbol{\beta}^* \in \mathbb{R}^p$ such that, for all $\boldsymbol{\beta} \neq \boldsymbol{\beta}^*$,

$$\left| \left\{ i; \pi(\mathbf{x}_i) = \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}^*}}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}^*}} \right\} \right| \geq n^* > \left| \left\{ i; \pi(\mathbf{x}_i) = \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}} \right\} \right| \quad . \tag{4.3}$$

It is obvious for the logistic regression model that the strong law of large numbers guarantees the validity of (4.2). Then (4.3) holds by definition of $\pi(\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta}^*)}{1+\exp(\mathbf{x}_i^T \boldsymbol{\beta}^*)}$ and rank $(\mathbf{X}) = p$.

Observe that (4.2) guarantees the convergence of the relative frequencies P_i to some constant $\pi(\mathbf{x}_i)$, for $i = 1, \ldots, N$, which can be interpreted as true success probabilities. The inequality (4.3) assures that at least $n^* > \frac{N}{2}$ of the quantities $\pi(\mathbf{x}_i)$ can be modeled according to a logistic regression model and that its parameter vector $\boldsymbol{\beta}^*$ is identifiable. However, the distribution of the corresponding n^* random variables Y_i need not to be binomial, whether (4.2) is valid or not.

In what follows it is assumed that all values of n_i are reasonably large, in the sense that the results are asymptotic for $n. = \sum_{i=1}^N n_i \to \infty$ such that $\frac{n_i}{n.} \to c_i \in (0, 1)$, but N and p remain fixed.

Theorem 4.1.1. [Normality of Empirical Logit Transform]

If the logistic model holds true and n_i is large, then the empirical logit transform $\ln\left(\frac{P_i}{1-P_i}\right)$ is approximately normally distributed with mean $\mathbf{x}_i^T\boldsymbol{\beta}$ and variance $\{n_i\pi(\mathbf{x}_i)(1-\pi(\mathbf{x}_i))\}^{-1}$. That is, for $i = 1, \ldots, N$,

$$\ln\left(\frac{P_i}{1-P_i}\right) \quad \approx \quad N(\mathbf{x}_i^T\boldsymbol{\beta}, \{n_i\pi(\mathbf{x}_i)(1-\pi(\mathbf{x}_i))\}^{-1}) \quad .$$

Proof.

Consider the logit transform

$$\ln\left(\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}\right) \quad = \quad \mathbf{x}_i^T\boldsymbol{\beta} \quad . \tag{4.4}$$

Under the condition that neither the number of successes nor the number of failures is too small, expression (4.4) is reasonably estimated by

$$\ln\left(\frac{Y_i}{n_i - Y_i}\right) \quad = \quad \ln\left(\frac{\frac{Y_i}{n_i}}{1-\frac{Y_i}{n_i}}\right) \quad ,$$

which we call the empirical logit transform.

More generally, if the parametric function of interest is $L[\pi(\mathbf{x}_i)]$, then we consider $L\left(\frac{Y_i}{n_i}\right)$. Now provided that the variation in $\frac{Y_i}{n_i}$ is relatively small we can write

$$L\left(\frac{Y_i}{n_i}\right) \approx L[\pi(\mathbf{x}_i)] + \left(\frac{Y_i}{n_i} - \pi(\mathbf{x}_i)\right) L'[\pi(\mathbf{x}_i)] \quad ,$$

from which it follows that $L\left(\frac{Y_i}{n_i}\right)$ is approximately normally distributed with mean $L[\pi(\mathbf{x}_i)]$ and variance

$$[L'(\pi(\mathbf{x}_i))]^2 \mathrm{Var}\left(\frac{Y_i}{n_i}\right) \quad = \quad [L'(\pi(\mathbf{x}_i))]^2 \frac{\pi(\mathbf{x}_i)(1-\pi(\mathbf{x}_i))}{n_i} \quad .$$

Now consider $L(t) = \ln\left(\frac{t}{1-t}\right)$ and $L'(t) = [t(1-t)]^{-1}$, then for $i = 1, \ldots, N$,

$$\ln\left(\frac{\frac{Y_i}{n_i}}{1 - \frac{Y_i}{n_i}}\right) \quad \approx \quad N(\mathbf{x}_i^T \boldsymbol{\beta}, \{n_i \pi(\mathbf{x}_i)[1 - \pi(\mathbf{x}_i)]\}^{-1}) \quad .$$

■

Definition 4.1.2.

(a) Let $\mathbf{X} = (\mathbf{X}_1^T, \ldots, \mathbf{X}_N^T)^T$ be the $N \times p$ matrix of explanatory variables $X_1, \ldots, X_p$, which may be discrete or continuous, with $\mathbf{x}_i^T = (x_{i1}, \ldots, x_{ip})$, for $i = 1, \ldots, N$.

(b) Let $\tilde{\mathbf{Y}} = (\tilde{Y}_1, \ldots, \tilde{Y}_N)^T$ be the $N \times 1$ vector of the empirical logit transform, where $\tilde{Y}_i = \ln\left(\frac{P_i}{1-P_i}\right)$, for $i = 1, \ldots, N$.

(c) By means of (a) and (b) for any $\boldsymbol{\beta} \in \mathbb{R}^P$ we can define the "residual" as

$$\tilde{r}_i = \tilde{y}_i - \mathbf{x}_i^T \boldsymbol{\beta}, \quad \text{for } i = 1, \ldots, N \quad . \tag{4.5}$$

Under the definition of residuals in (4.5), the logistic regression model can be considered as a particular case of the multiple linear model studied by Luong and Garrido (1992) in the context of quadratic distance estimation.

In what follows assume that the random errors

$$\tilde{r}_i = \tilde{y}_i - \mathbf{x}_i^T \boldsymbol{\beta}_0, \quad \text{for } i = 1, \dots, N \quad ,$$

where $\boldsymbol{\beta}_0 = (\beta_{01}, \dots, \beta_{0p})^T$ is the vector of unknown parameters, are independent and identically distributed. Their common distribution function, F_0, is unknown (non-parametric model) but assumed to be absolutely continuous with density function f_0, symmetric around zero. In fact, using Theorem 4.1.1 it is simple to show that the expected value and the index of skewness of the random errors are both equal to zero.

Define, for any $\boldsymbol{\beta} \in \mathbb{R}^p$

$$\hat{F}_j^{\boldsymbol{\beta}}(y) \;=\; \sum_{i=1}^{N} w_{ij} I(\tilde{y}_i - \mathbf{x}_i^T \boldsymbol{\beta} \leq y), \quad \text{for } j = 1, \dots, p \quad , \tag{4.6}$$

where I denotes the indicator function and w_{ij} are known weights. Similarly, define

$$F_j^0(y) \;=\; \sum_{i=1}^{N} w_{ij} F_0(y), \quad \text{for } j = 1, \dots, p \quad . \tag{4.7}$$

Note that $\hat{F}_j^{\boldsymbol{\beta}}$ are empirical processes based on the residuals in (4.5) and the known weights $w_{1j}, \dots, w_{Nj}$, while F_j^0 are the corresponding theoretical distributions.

Also define for $j = 1, \dots, p$

$$\begin{aligned}
\mathbf{Z}_j^{\boldsymbol{\beta}} \;&=\; \left[\int_{-\infty}^{\infty} h_1(x) d\hat{F}_j^{\boldsymbol{\beta}}(x), \dots, \int_{-\infty}^{\infty} h_k(x) d\hat{F}_j^{\boldsymbol{\beta}}(x)\right]^T \\
&=\; \left[\sum_{i=1}^{N} w_{ij} h_1(\tilde{y}_i - \mathbf{x}_i^T \boldsymbol{\beta}), \dots, \sum_{i=1}^{N} w_{ij} h_k(\tilde{y}_i - \mathbf{x}_i^T \boldsymbol{\beta})\right]^T \quad , \\
\text{and } \mathbf{Z}_j^0 \;&=\; \left[\int_{-\infty}^{\infty} h_1(x) dF_j^0(x), \dots, \int_{-\infty}^{\infty} h_k(x) dF_j^0(x)\right]^T \quad ,
\end{aligned}$$

where $h_1, \dots, h_k$ is a fixed choice of odd functions, i.e. $h_i(x) = -h_i(-x)$, for $x \neq 0$ and $h(0) = 0$.

The QD estimator is the vector $\hat{\boldsymbol{\beta}}$ which minimizes the following sum of quadratic forms

$$d(\boldsymbol{\beta}) \quad = \quad (\mathbf{Z}_1^{\boldsymbol{\beta}} - \mathbf{Z}_1^0)^T \mathbf{Q} (\mathbf{Z}_1^{\boldsymbol{\beta}} - \mathbf{Z}_1^0) + \cdots + (\mathbf{Z}_p^{\boldsymbol{\beta}} - \mathbf{Z}_p^0)^T \mathbf{Q} (\mathbf{Z}_p^{\boldsymbol{\beta}} - \mathbf{Z}_p^0) \quad , \tag{4.8}$$

where $\mathbf{Q}$ denotes a $k \times k$ constant symmetric positive-definite matrix.

Furthermore, since $\mathbf{Z}_j^0 = 0$ for $j = 1, \ldots, p$, when h is odd, minimizing (4.8) with respect to $\boldsymbol{\beta}$, is reduced to minimizing

$$d(\boldsymbol{\beta}) \quad = \quad [\mathbf{Z}_1^{\boldsymbol{\beta}}]^T \mathbf{Q} \mathbf{Z}_1^{\boldsymbol{\beta}} + \cdots + [\mathbf{Z}_p^{\boldsymbol{\beta}}]^T \mathbf{Q} \mathbf{Z}_p^{\boldsymbol{\beta}} \quad . \tag{4.9}$$

Using Kronecker's product notation [see, Graham (1981) for a brief introduction to this theory] and calling $\mathbf{Z}^{\boldsymbol{\beta}} = ([\mathbf{Z}_1^{\boldsymbol{\beta}}]^T, \ldots, [\mathbf{Z}_p^{\boldsymbol{\beta}}]^T)^T$, then (4.9) can be expressed more concisely as

$$d(\boldsymbol{\beta}) \quad = \quad [\mathbf{Z}^{\boldsymbol{\beta}}]^T (\mathbf{I}_p \otimes \mathbf{Q}) \mathbf{Z}^{\boldsymbol{\beta}} \quad , \tag{4.10}$$

where $\mathbf{I}_p$ denotes the identity matrix of order p.

The QD estimator $\hat{\boldsymbol{\beta}}$ is the vector which minimizes (4.10) with respect to $\boldsymbol{\beta}$.

4.1.1 Asymptotic Properties of the QD Estimator

In this section we derive the asymptotic properties of the QD estimators such as consistency and asymptotic normality.

The derivation is based on the results of Luong and Garrido (1992) for multiple linear regression, adapted here to logistic regression. We need to impose the following regulatory conditions.

Definition 4.1.3.

Let $\mathbf{W} = (\mathbf{w}_1^T, \dots, \mathbf{w}_N^T)^T$ be the $N \times p$ matrix of weights used in (4.6), where $\mathbf{w}_i^T = (w_{i1}, \dots, w_{ip})$, for $i = 1, \dots, N$.

Assumptions for Asymptotic Properties

($\boldsymbol{a1}$) $\lim_{N \to \infty} \sum_{i=1}^N w_{ij}^2 = 0$, for each $j = 1, \dots, p$,

($\boldsymbol{a2}$) $\lim_{N \to \infty} \mathbf{W}^T \mathbf{X}$ exists and is invertible,

($\boldsymbol{a3}$) $\lim_{N \to \infty} \sum_{i=1}^N w_{ij}^2 x_{il}^2 = 0$, for each $j = 1, \dots, p;\ l = 1, \dots, p$,

($\boldsymbol{a4}$) $\lim_{N \to \infty} \sum_{i=1}^N \left| w_{ij} x_{il} \right|$ exists for each $j = 1, \dots, p;\ l = 1, \dots, p$,

($\boldsymbol{a5}$) $\dot{h}_i(x) = \frac{\partial}{\partial x} h_i(x)$ is uniformily continuous and $\mathrm{Var}[\dot{h}(\tilde{r})] < \infty$,

($\boldsymbol{a6}$) the x_{ij} values belong to a compact set,

($\boldsymbol{a7}$) $\max_{1 \le i \le N} \{\mathbf{w}_i \Sigma \mathbf{w}_i^T\}$ is bounded for all N,

($\boldsymbol{a8}$) $\underline{\lambda}(\mathbf{W}^T\mathbf{W} \otimes \Sigma) \longrightarrow \infty$ if $N \longrightarrow \infty$, where $\underline{\lambda}(M)$ represents the smallest eigenvalue of matrix $\mathbf{M}$ and $\boldsymbol{\Sigma}$ is the variance covariance matrix of

$$h(\tilde{r}) = [h_1(\tilde{r}), \dots, h_k(\tilde{r})]^T \quad .$$

Theorem 4.1.2. [Consistency]

Consider the $N \times p$ matrix of weights $\mathbf{W}$ defined above. Let $\mathbf{X}$ be the $N \times p$ matrix given in Definition 4.1.2 (a). Matrices $\mathbf{W}$ and $\mathbf{X}$ are assumed to have rank p. If the weights matrix $\mathbf{W}$ satisfies assumption **($\boldsymbol{a1}$)**, then the QD estimator $\hat{\boldsymbol{\beta}}$, obtained minimizing the function $d(\boldsymbol{\beta})$, is consistent.

Proof.

Using Chebyshev's inequality and assumption $(\boldsymbol{a1})$, we have that $\mathbf{Z}^{\boldsymbol\beta_0} \xrightarrow{P} 0$ provided that the density function of the random errors, f_0, is symmetric. This implies that both

$$d(\boldsymbol\beta_0) \xrightarrow{P} 0 \quad \text{and} \quad d(\hat{\boldsymbol\beta}) \xrightarrow{P} 0, \quad \text{as} \quad N \longrightarrow \infty \quad .$$

Therefore, the consistency of $\hat{\boldsymbol\beta}$ is guaranteed as long as $\mathbb{E}(\mathbf{Z}^{\boldsymbol\beta}) = 0$ at, and only at, $\boldsymbol\beta = \boldsymbol\beta_0$ when the parametric space is compact. ■

Theorem 4.1.3. [Asymptotic Normality]

Under assumptions $(\boldsymbol{a2})$ to $(\boldsymbol{a8})$, the central limit property of the QD estimator $\hat{\boldsymbol\beta}$ gives

$$(\mathbf{W}^T\mathbf{W})^{-\frac{1}{2}}(\hat{\boldsymbol\beta} - \boldsymbol\beta_0) \quad \xrightarrow{L} \quad \mathrm{N}\left(\mathbf{0}, \boldsymbol\Sigma_2\right) \quad , \tag{4.11}$$

where $\boldsymbol\Sigma_2 = [(\mathbf{W}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{W})^{-1}](\mathbf{S}_0^T\mathbf{Q}\mathbf{S}_0)^{-2}(\mathbf{S}_0^T\mathbf{Q}\boldsymbol\Sigma\mathbf{Q}\mathbf{S}_0)$.

Proof.

The proof follows from the derivation of the asymptotic variance-covariance matrix of $\hat{\boldsymbol\beta}$ and the use of the multivariate central limit theorem.

Let $\mathbf{S}_0^T = [\mathbb{E}(\dot{h}_1(\tilde{r})), \ldots, \mathbb{E}(\dot{h}_k(\tilde{r}))]$, where $\dot{h}_i(x) = \frac{\partial}{\partial x} h_i(x)$ and assume that the function d, given by (4.10), is differentiable. Then $\hat{\boldsymbol\beta}$ satisfies the following p-system of equations

$$\frac{\partial}{\partial \boldsymbol\beta}[\mathbf{Z}^{\hat{\boldsymbol\beta}}]^T(\mathbf{I}_p \otimes \mathbf{Q})\mathbf{Z}^{\hat{\boldsymbol\beta}} \quad = \quad \mathbf{0} \quad . \tag{4.12}$$

Under assumptions $(\boldsymbol{a3})$ to $(\boldsymbol{a6})$ and using the properties of Kronecker's product, we obtain that

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{Z}^{\hat{\boldsymbol{\beta}}} &= \frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{Z}^{\boldsymbol{\beta}_0} + o_p(1) \quad , & (4.13)\\
\frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{Z}^{\boldsymbol{\beta}_0} &= -\mathbf{W}^T\mathbf{X} \otimes \mathbf{S}_0 + o_p(1) \quad , & (4.14)\\
\frac{\partial}{\partial \boldsymbol{\beta}} [\mathbf{Z}^{\boldsymbol{\beta}_0}]^T (\mathbf{I} \otimes \mathbf{Q}) \frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{Z}^{\boldsymbol{\beta}_0} &= (\mathbf{X}^T\mathbf{W})(\mathbf{W}^T\mathbf{X}) \otimes (\mathbf{S}_0^T\mathbf{Q}\mathbf{S}_0) + o_p(1) \\
&= (\mathbf{X}^T\mathbf{W})(\mathbf{W}^T\mathbf{X})(\mathbf{S}_0^T\mathbf{Q}\mathbf{S}_0) + o_p(1) \quad , & (4.15)
\end{aligned}$$

where $o_p(1)$ stands for a random infinitesimal term converging in probability.

Substituting (4.14) and (4.15) in (4.12) and using a Taylor's series expansion, we have

$$(\mathbf{S}_0^T\mathbf{Q}\mathbf{S}_0)(\mathbf{X}^T\mathbf{W})(\mathbf{W}^T\mathbf{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = -(\mathbf{X}^T\mathbf{W} \otimes \mathbf{S}_0^T)(\mathbf{I} \otimes \mathbf{Q})\mathbf{Z}^{\boldsymbol{\beta}_0} + o_p(1) \,. \quad (4.16)$$

Since $\mathbf{Z}^{\boldsymbol{\beta}_0}$ is a vector of sums of independent variables, then under assumptions $(\boldsymbol{a7})$, $(\boldsymbol{a8})$ and the multivariate central limit theorem, we obtain that

$$(\mathbf{W}^T\mathbf{W} \otimes \boldsymbol{\Sigma})^{-\frac{1}{2}} \mathbf{Z}^{\boldsymbol{\beta}_0} \xrightarrow{L} \mathrm{N}\,(\mathbf{0}, \mathbf{I}) \quad . \quad (4.17)$$

Using (4.17) and (4.16), we have that

$$\mathrm{Var}[-(\mathbf{X}^T\mathbf{W} \otimes \mathbf{S}_0^T)(\mathbf{I} \otimes \mathbf{Q})\mathbf{Z}^{\boldsymbol{\beta}_0}] = (\mathbf{X}^T\mathbf{W} \otimes \mathbf{S}_0^T)(\mathbf{W}^T\mathbf{W} \otimes \mathbf{Q}\boldsymbol{\Sigma}\mathbf{Q})(\mathbf{W}^T\mathbf{X} \otimes \mathbf{S}_0)\,.$$

Then $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ is asymptotically normal with asymptotic variance-covariance matrix given by

$$\boldsymbol{\Sigma}_1 = \mathbf{A}(\mathbf{W}^T\mathbf{W} \otimes \mathbf{Q}\boldsymbol{\Sigma}\mathbf{Q})\mathbf{A}^T \quad ,$$

where $\mathbf{A} = [(\mathbf{S}_0^T\mathbf{Q}\mathbf{S}_0)(\mathbf{X}^T\mathbf{W})(\mathbf{W}^T\mathbf{X})]^{-1}[\mathbf{X}^T\mathbf{W} \otimes \mathbf{S}_0^T]$.

Finally, $\boldsymbol{\Sigma}_1$ can be expressed as

$$\boldsymbol{\Sigma}_1 = (\mathbf{W}^T\mathbf{X})^{-1}(\mathbf{W}^T\mathbf{W})(\mathbf{X}^T\mathbf{W})^{-1}(\mathbf{S}_0^T\mathbf{Q}\mathbf{S}_0)^{-2}(\mathbf{S}_0^T\mathbf{Q}\boldsymbol{\Sigma}\mathbf{Q}\mathbf{S}_0) \quad , \quad (4.18)$$

or equivalently,

$$(\mathbf{W}^T\mathbf{W})^{-\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{L} \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_2) \quad ,$$

where $\boldsymbol{\Sigma}_2 = (\mathbf{W}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{W})^{-1}(\mathbf{S}_0^T\mathbf{Q}\mathbf{S}_0)^{-2}(\mathbf{S}_0^T\mathbf{Q}\boldsymbol{\Sigma}\mathbf{Q}\mathbf{S}_0)$. ■

Corollary 4.1.1.

The minimum asymptotic variance of the QD estimator $\hat{\boldsymbol{\beta}}$, $\boldsymbol{\Sigma}_1$, is reached when the weights matrix $\mathbf{W} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$ and the $k \times k$ matrix $\mathbf{Q} = \boldsymbol{\Sigma}^{-1}$. That is, $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{S}_0^T\boldsymbol{\Sigma}^{-1}\mathbf{S}_0)^{-1}$.

Proof.

The "optimal weights" can be chosen to minimize (4.18). Using a generalized Cauchy-Schwartz inequality it is easy to verify that $\mathbf{W} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$. Also, if $\boldsymbol{\Sigma}$ is invertible, the optimal choice of $\mathbf{Q}$, in the sense of minimizing the variance-covariance matrix $\boldsymbol{\Sigma}_1$, is $\mathbf{Q} = \boldsymbol{\Sigma}^{-1}$. ■

4.1.2 Influence Function of the QD Estimator

Let $\hat{G}_i$ be the degenerate distribution at $\tilde{y}_i$ and define $G_i(\tilde{y}_i) = F_0(\tilde{y}_i - \mathbf{x}_i^T\boldsymbol{\beta}_0)$. Then the QD estimator, $\hat{\boldsymbol{\beta}}$, can be considered as the statistical functional $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}(\hat{G}_1, \ldots, \hat{G}_N)$, where $\boldsymbol{\beta}(G_1, \ldots, G_N)$ is defined implicitly as a solution of the p-system of equations

$$\frac{\partial}{\partial\boldsymbol{\beta}}[\mathbf{Z}^{\boldsymbol{\beta}}]^T(\mathbf{I}_p \otimes \mathbf{Q})\mathbf{Z}^{\boldsymbol{\beta}} = \mathbf{0} \quad ,$$

with $\mathbf{Z}^{\boldsymbol{\beta}} = ([\mathbf{Z}_1^{\boldsymbol{\beta}}]^T, \ldots, [\mathbf{Z}_p^{\boldsymbol{\beta}}]^T)^T$, and

$$\mathbf{Z}_j^{\boldsymbol{\beta}} = \left[\sum_{i=1}^{N}\int_{-\infty}^{\infty} w_{ij}h_1(\tilde{y} - \mathbf{x}_i^T\boldsymbol{\beta})dG_i(\tilde{y}), \ldots, \sum_{i=1}^{N}\int_{-\infty}^{\infty} w_{ij}h_k(\tilde{y} - \mathbf{x}_i^T\boldsymbol{\beta})dG_i(\tilde{y})\right]^T \quad ,$$

for $j = 1, \ldots, p$.

Proposition 4.1.1.

Let $G_{l,\lambda} = (1-\lambda)G_l + \lambda\delta^{\eta_l}$, where δ^{η_l} denotes the usual degenerate distribution at η_l and $\lambda \in (0,1)$. Let $H(\boldsymbol{\beta},\lambda) = \frac{\partial}{\partial\boldsymbol{\beta}}[\mathbf{Z}^{\boldsymbol{\beta}}]^T(\mathbf{I}_p \otimes \mathbf{Q})\mathbf{Z}^{\boldsymbol{\beta}}$, if G_l in $\mathbf{Z}^{\boldsymbol{\beta}}$ is replaced by $G_{l,\lambda}$ then the influence function of an observation η_l at x_l^T is given by

$$\mathrm{IF}(\eta_l, x_l^T) = -\left[\frac{\partial H}{\partial\boldsymbol{\beta}}\right]^{-1}\left[\frac{\partial H}{\partial\lambda}\right] \quad ,$$

evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ and $\lambda = 0$.

Proof.

Under the assumption that $G_{l,\lambda} = (1-\lambda)G_l + \lambda\delta^{\eta_l}$, the influence function of an observation η_l at x_l^T can be written as

$$\mathrm{IF}(\eta_l, x_l^T; \boldsymbol{\beta}, G_{l,\lambda}) = \frac{\partial\boldsymbol{\beta}(G_1,\dots,G_{l,\lambda},\dots,G_N)}{\partial\lambda}\bigg|_{\lambda=0} \quad .$$

Now, if G_l in $\mathbf{Z}^{\boldsymbol{\beta}}$ is replaced by $G_{l,\lambda}$, we have that

$$H(\boldsymbol{\beta}(G_1,\dots,G_{l,\lambda},\dots,G_N)) = 0 \quad .$$

Thus

$$\frac{\partial H}{\partial\boldsymbol{\beta}}\bigg|_{\substack{\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}(G_1,\dots,G_{l,\lambda},\dots,G_N) \\ \lambda = 0}} \times \frac{\partial\boldsymbol{\beta}(G_1,\dots,G_{l,\lambda},\dots,G_N)}{\partial\lambda}\bigg|_{\lambda=0} + \frac{\partial H}{\partial\lambda}\bigg|_{\lambda=0} = 0 \quad .$$

Hence

$$\mathrm{IF}(\eta_l, x_l^T) = -\left[\frac{\partial H}{\partial\boldsymbol{\beta}}\right]^{-1}\left[\frac{\partial H}{\partial\lambda}\right] \quad .$$

■

Corollary 4.1.2.

If the conditions of Proposition 4.1.1 hold, then the vector of influence functions of $\hat{\boldsymbol{\beta}}$ can be expressed as

$$\mathrm{IF}(\eta_l, x_l^T) = (\mathbf{S}_0^T\mathbf{Q}\mathbf{S}_0)^{-1}[(\mathbf{W}^T\mathbf{X})(\mathbf{X}^T\mathbf{W})]^{-1}[\mathbf{W}^T\mathbf{X} \otimes \mathbf{S}_0^T\mathbf{Q}][\mathbf{w}_l^T \otimes h(\eta_l - \mathbf{x}_l^T\boldsymbol{\beta}_0)] \, .$$

4.2 Robust Quadratic Distance Estimators for Multinomial Logistic Regression

In this section we propose an extension of QD estimation to the multinomial logistic regression model. We use ideas similar to those developed in Section 4.1 for the case of binary responses.

Consider again an individual characterized by a vector $\mathbf{x}_i^T = (x_{i1}, \ldots, x_{ip}) \in \mathbb{R}^p$, with p explanatory variables, which may be discrete or continuous. Let $G_1, \ldots, G_g$ be all the possible groups in which this individual can be classified. We are interested in estimating the probability $\mathbb{P}(Y_i = j|\mathbf{x}_i)$ that the individual $\mathbf{x}_i$ belongs to one of the g groups $G_1, \ldots, G_g$.

Assume a random sample from populations $G_1, \ldots, G_g$ and denote by N the number of different vectors $\mathbf{x}_i$. Then let n_i be the number of observations at $\mathbf{x}_i$ for $i = 1, \ldots, N$ and y_{ji} the number of G_j observations at $\mathbf{x}_i$, with $n_i = \sum_{j=1}^{g} y_{ji}$.

Fixing the last classification group G_g and comparing to it the inclusion probabilities of every other class, we say that an observation $\mathbf{x}_i$ satisfies the logistic assumptions if

$$\ln\left[\frac{\mathbb{P}(Y_i = j|\mathbf{x}_i)}{\mathbb{P}(Y_i = g|\mathbf{x}_i)}\right] = \mathbf{x}_i^T \boldsymbol{\beta}_j \quad , \quad j = 1, \ldots, g-1 \quad ,$$

or correspondingly,

$$\pi_j(\mathbf{x}_i) = \mathbb{P}(Y_i = j|\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta}_j)}{1 + \sum\limits_{l=1}^{g-1} \exp(\mathbf{x}_i^T \boldsymbol{\beta}_l)} \quad , \quad j = 1, \ldots, g \quad ,$$

where $\boldsymbol{\beta}_j^T = (\beta_{1j}, \ldots, \beta_{pj})$ is a vector of unknown parameters, with $\boldsymbol{\beta}_g = \mathbf{0}$, and the corresponding $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_{g-1}^T)^T$ be the $p(g-1)$ dimensional column vector of unknown parameters.

Note that the random variables $(Y_{1i}, \ldots, Y_{gi})$ have a multinomial distribution with n_i trials and cell probabilities $\pi_1(\mathbf{x}_i), \ldots, \pi_g(\mathbf{x}_i)$. The joint probability mass function of $(Y_{1i}, \ldots, Y_{gi})$ is

$$f(y_{1i}, \ldots, y_{gi}) = \frac{n_i!}{y_{1i}! \cdots y_{gi}!} \pi_1(\mathbf{x}_i)^{y_{1i}} \cdots \pi_g(\mathbf{x}_i)^{y_{gi}} \quad ,$$

with $n_i = \sum_{j=1}^{g} y_{ji}$, for $i = 1, \ldots, N$.

Observe that the marginal random variables Y_{ji} for $j = 1, \ldots, g$, obtained from the multinomial distribution have a binomial distribution with index n_i and probability $\pi_j(\mathbf{x}_i)$, that is

$$Y_{ji} \sim \text{Binomial}\,(n_i, \pi_j(\mathbf{x}_i)), \quad j = 1, \ldots, g \quad ; \quad i = 1, \ldots, N \quad .$$

An extension of Definition 4.1.1 to the multinomial case can be given as follows.

Definition 4.2.1.

For $i = 1, \ldots, N$ fixed, let $(y_{1i}, \ldots, y_{gi})$ be observations from $(Y_{1i}, \ldots, Y_{gi}) \sim$ Multinomial$[n_i, (\pi_1(\mathbf{x}_i), \ldots, \pi_g(\mathbf{x}_i))]$. Then the relative frequencies P_{ji} are defined, for each $j = 1, \ldots, g$, as

$$P_{ji} = \begin{cases} \frac{1}{2n_i} & \text{if } \; Y_{ji} = 0 \\ \frac{Y_{ji}}{n_i} & \text{if } \; 1 \leq Y_{ji} \leq n_i - 1 \\ 1 - \frac{1}{2n_i} & \text{if } \; Y_{ji} = n_i \end{cases} \quad , \quad i = 1, \ldots, N \quad . \tag{4.19}$$

Assumptions:

Under the above definitions it is assumed that

(a) there exist $\pi_j(\mathbf{x}_i) \in (0,1)$, for $j = 1, \ldots, g$; $i = 1, \ldots, N$, such that if $\min\limits_{1 \leq i \leq N} (n_i) \longrightarrow \infty$, then

$$(P_{11}, \ldots, P_{gN}) \longrightarrow (\pi_1(\mathbf{x}_1), \ldots, \pi_g(\mathbf{x}_N)) \; , \quad \text{almost surely} \; , \tag{4.20}$$

(b) there exists exactly one vector $\boldsymbol{\beta}_j^* \in \mathbb{R}^p$, for each $j = 1, \ldots, g$, such that, for all $\boldsymbol{\beta}_j \neq \boldsymbol{\beta}_j^*$

$$\left|\left\{ i; \pi_j(\mathbf{x}_i) = \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}_j^*}}{\sum\limits_{l=1}^{g} e^{\mathbf{x}_i^T \boldsymbol{\beta}_l^*}} \right\}\right| \geq n_j^* > \left|\left\{ i; \pi_j(\mathbf{x}_i) = \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}_j}}{\sum\limits_{l=1}^{g} e^{\mathbf{x}_i^T \boldsymbol{\beta}_l}} \right\}\right| , \qquad (4.21)$$

where $n_j^* = \lfloor \frac{N}{2} \rfloor + \lfloor \frac{p}{2} \rfloor$, with $\lfloor z \rfloor$ being the largest integer less than or equal to z.

The argument generalizes that given in Theorem 4.1.1 for a multinomial logistic regression model. Again assume that all values of n_i are reasonably large, such that the results are asymptotic for $n. = \sum_{i=1}^{N} n_i \to \infty$, where $\frac{n_i}{n.} \to c_i \in (0,1)$, with N and p remaining fixed.

Theorem 4.2.1. [Normality of Multinomial Logit Transform]

Consider the above multinomial logistic model and suppose that n_i is large. Then the multinomial logit transform $[\ln(\frac{P_{1i}}{P_{gi}}), \ldots, \ln(\frac{P_{(g-1)i}}{P_{gi}})]^T$ is approximately multinormally distributed, with vector of mean $[\mathbf{x}_i^T\boldsymbol{\beta}_1, \ldots, \mathbf{x}_i^T\boldsymbol{\beta}_{g-1}]^T$ and variance-covariance matrix given by

$$n_i^{-1}\boldsymbol{\Sigma}_0 \;=\; n_i^{-1} \begin{pmatrix} \frac{1}{\pi_1(\mathbf{x}_i)} + \frac{1}{\pi_g(\mathbf{x}_i)} & \cdots & \frac{1}{\pi_g(\mathbf{x}_i)} \\ \vdots & \ddots & \vdots \\ \frac{1}{\pi_g(\mathbf{x}_i)} & \cdots & \frac{1}{\pi_{g-1}(\mathbf{x}_i)} + \frac{1}{\pi_g(\mathbf{x}_i)} \end{pmatrix} . \qquad (4.22)$$

Proof. Consider the multinomial logistic model link function given by

$$\begin{aligned} l(\boldsymbol{\pi}(\mathbf{x}_i)) &= \left[\ln\left(\frac{\pi_1(\mathbf{x}_i)}{\pi_g(\mathbf{x}_i)}\right), \ldots, \ln\left(\frac{\pi_{g-1}(\mathbf{x}_i)}{\pi_g(\mathbf{x}_i)}\right)\right]^T , \qquad (4.23) \\ &= [\mathbf{x}_i^T\boldsymbol{\beta}_1, \ldots, \mathbf{x}_i^T\boldsymbol{\beta}_{g-1}]^T , \end{aligned}$$

where $\boldsymbol{\pi}(\mathbf{x}_i) = (\pi_1(\mathbf{x}_i), \ldots, \pi_g(\mathbf{x}_i))$, for $i = 1, \ldots, N$.

Suppose that $\pi_j(\mathbf{x}_i)$ and $\pi_g(\mathbf{x}_i)$ are never "too small". We can estimate (4.23) reasonably by $l(\frac{Y_{1i}}{n_i}, \dots, \frac{Y_{gi}}{n_i}) = \left[\ln\left(\frac{Y_{1i}}{Y_{gi}}\right), \dots, \ln\left(\frac{Y_{(g-1)i}}{Y_{gi}}\right)\right]^T$, which we call the empirical logit transform. In general if the function of interest is $l[\boldsymbol{\pi}(\mathbf{x}_i)]$, then we consider $l(\frac{Y_{1i}}{n_i}, \dots, \frac{Y_{gi}}{n_i})$.

Define

$$\dot{l}_j[\boldsymbol{\pi}(\mathbf{x}_i)] = \frac{\partial}{\partial z_j} l(z_1, \dots, z_g)\Big|_{z_1=\pi_1(\mathbf{x}_i), \dots, z_g=\pi_g(\mathbf{x}_i)} .$$

The first-order Taylor series expansion of l about $\boldsymbol{\pi}(\mathbf{x}_i)$ is

$$l\left(\frac{Y_{1i}}{n_i}, \dots, \frac{Y_{gi}}{n_i}\right) \approx l[\boldsymbol{\pi}(\mathbf{x}_i)] + \sum_{i=1}^{g} \dot{l}_j[\boldsymbol{\pi}(\mathbf{x}_i)]\left(\frac{Y_{ji}}{n_i} - \pi_j(\mathbf{x}_i)\right)$$

$$\begin{pmatrix} \ln\left(\frac{Y_{1i}}{Y_{gi}}\right) \\ \vdots \\ \ln\left(\frac{Y_{(g-1)i}}{Y_{gi}}\right) \end{pmatrix} \approx \begin{pmatrix} \ln\left(\frac{\pi_1(\mathbf{x}_i)}{\pi_g(\mathbf{x}_i)}\right) \\ \vdots \\ \ln\left(\frac{\pi_{g-1}(\mathbf{x}_i)}{\pi_g(\mathbf{x}_i)}\right) \end{pmatrix} + \begin{pmatrix} \frac{\left(\frac{Y_{1i}}{n_i} - \pi_j(\mathbf{x}_i)\right)}{\pi_1(\mathbf{x}_i)} - \frac{\left(\frac{Y_{gi}}{n_i} - \pi_g(\mathbf{x}_i)\right)}{\pi_g(\mathbf{x}_i)} \\ \vdots \\ \frac{\left(\frac{Y_{(g-1)i}}{n_i} - \pi_j(\mathbf{x}_i)\right)}{\pi_{g-1}(\mathbf{x}_i)} - \frac{\left(\frac{Y_{gi}}{n_i} - \pi_g(\mathbf{x}_i)\right)}{\pi_g(\mathbf{x}_i)} \end{pmatrix} . \tag{4.24}$$

Now, take expectations on both sides of (4.24) to get

$$\mathbb{E}\begin{pmatrix} \ln\left(\frac{Y_{1i}}{Y_{gi}}\right) \\ \vdots \\ \ln\left(\frac{Y_{(g-1)i}}{Y_{gi}}\right) \end{pmatrix} \approx \begin{pmatrix} \mathbf{x}_i^T \boldsymbol{\beta}_1 \\ \vdots \\ \mathbf{x}_i^T \boldsymbol{\beta}_{g-1} \end{pmatrix} .$$

Define

$$\mathbf{R}^T = \left(\frac{\left(\frac{Y_{1i}}{n_i} - \pi_j(\mathbf{x}_i)\right)}{\pi_1(\mathbf{x}_i)} - \frac{\left(\frac{Y_{gi}}{n_i} - \pi_g(\mathbf{x}_i)\right)}{\pi_g(\mathbf{x}_i)}, \dots, \frac{\left(\frac{Y_{(g-1)i}}{n_i} - \pi_j(\mathbf{x}_i)\right)}{\pi_{g-1}(\mathbf{x}_i)} - \frac{\left(\frac{Y_{gi}}{n_i} - \pi_g(\mathbf{x}_i)\right)}{\pi_g(\mathbf{x}_i)} \right)$$

We can now approximate the variance of $\left[\ln\left(\frac{Y_{1i}}{Y_{gi}}\right), \dots, \ln\left(\frac{Y_{(g-1)i}}{Y_{gi}}\right)\right]^T$ as

$$\mathrm{Var}\begin{pmatrix} \ln\left(\frac{Y_{1i}}{Y_{gi}}\right) \\ \vdots \\ \ln\left(\frac{Y_{(g-1)i}}{Y_{gi}}\right) \end{pmatrix} \approx \mathbb{E}\mathbf{R}\mathbf{R}^T = n_i^{-1} \begin{pmatrix} \frac{1}{\pi_1(\mathbf{x}_i)} + \frac{1}{\pi_g(\mathbf{x}_i)} & \cdots & \frac{1}{\pi_g(\mathbf{x}_i)} \\ \vdots & \ddots & \vdots \\ \frac{1}{\pi_g(\mathbf{x}_i)} & \cdots & \frac{1}{\pi_{g-1}(\mathbf{x}_i)} + \frac{1}{\pi_g(\mathbf{x}_i)} \end{pmatrix} .$$

■

Definition 4.2.2.

(a) Let $\underline{\mathbf{X}} = (\underline{\mathbf{X}}_1^T, \dots, \underline{\mathbf{X}}_N^T)^T$ be the $N(g-1) \times p(g-1)$ matrix of (discrete or continuous) explanatory variables $X_1, \dots, X_p$, where

$$\underline{\mathbf{X}}_i^T = (\underline{\mathbf{x}}_{1i}^T, \underline{\mathbf{x}}_{2i}^T, \dots, \underline{\mathbf{x}}_{(g-1)i}^T)^T = \begin{pmatrix} \mathbf{x}_i^T & \mathbf{0}^T & \cdots & \mathbf{0}^T \\ \mathbf{0}^T & \mathbf{x}_i^T & \cdots & \mathbf{0}^T \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}^T & \mathbf{0}^T & \cdots & \mathbf{x}_i^T \end{pmatrix} \quad ,$$

with $\mathbf{x}_i^T = (x_{i1}, \dots, x_{ip})$, for $i = 1, \dots, N$.

(b) Let $\underline{\mathbf{Y}} = (\underline{\mathbf{Y}}_1^T, \dots, \underline{\mathbf{Y}}_N^T)^T$ be the $N(g-1) \times 1$ vector of multinomial logit transforms, where $\underline{\mathbf{Y}}_i^T = (\underline{Y}_{1i}, \dots, \underline{Y}_{(g-1)i})$, with $\underline{Y}_{ji} = \ln\left(\frac{P_{ji}}{P_{gi}}\right)$, for $j = 1, \dots, g-1 \;\; ; \;\; i = 1, \dots, N$.

(c) By means of (a) and (b) we define the "residual" as

$$\underline{r}_{ji} = \underline{y}_{ji} - \underline{\mathbf{x}}_{ji}^T \boldsymbol{\beta}, \quad \text{for } j = 1, \dots, g-1 \;\; ; \;\; i = 1, \dots, N \quad . \tag{4.25}$$

In the multinomial logistic model the radom errors are defined by

$$\underline{r}_{ji} = \underline{y}_{ji} - \underline{\mathbf{x}}_{ji}^T \boldsymbol{\beta}_0^*, \quad \text{for} \quad j = 1, \dots, g-1 \;\; ; \;\; i = 1, \dots, N \quad ,$$

where $\boldsymbol{\beta}_0^* = (\boldsymbol{\beta}_{01}^T, \dots, \boldsymbol{\beta}_{0(g-1)}^T)$ is the $p(g-1)$ dimensional column vector of unknown parameters. Their common multivariate distribution function, F_0^*, is unknown but assumed to be absolutely continuous, with a density function f_0^* symmetric around zero. Applying Theorem 4.2.1 one easily checks that the expected value and the index of skewness of these random errors are both equal to zero.

Define

$$\hat{F}_t^{\boldsymbol{\beta}}(y) = \sum_{i=1}^{N}\sum_{j=1}^{g-1} w_{ji[t]} I(\underline{y}_{ji} - \underline{\mathbf{x}}_{ji}^T\boldsymbol{\beta} \leq y), \quad \text{for } t = 1, \ldots, p(g-1) \quad , \tag{4.26}$$

where I denotes an indicator function and $w_{ji[t]}$ are known weights. Similarly, define

$$F_t^0(y) \;=\; \sum_{i=1}^{N}\sum_{j=1}^{g-1} w_{ji[t]} F_{0,t}^*(y), \quad \text{for } t = 1, \ldots, p(g-1) \quad , \tag{4.27}$$

where $F_{0,t}^*$ is the marginal distribution of $F_0^* = [F_{0,1}^*, \ldots, F_{0,p(g-1)}^*]$.

Observe that the functions $\hat{F}_t^{\boldsymbol{\beta}}$ are empirical processes based on the residuals and weights $w_{j1[t]}, \ldots, w_{jN[t]}$, while F_t^0 are the corresponding theoretical distributions.

Now we define for $t = 1, \ldots, p(g-1)$

$$\begin{aligned} \mathbf{Z}_t^{\boldsymbol{\beta}} &= \left[\int_{-\infty}^{\infty} h_1(x) d\hat{F}_t^{\boldsymbol{\beta}}(x), \ldots, \int_{-\infty}^{\infty} h_k(x) d\hat{F}_t^{\boldsymbol{\beta}}(x)\right]^T \quad , \\ &= \left[\sum_{i=1}^{N}\sum_{j=1}^{g-1} w_{ji[t]} h_1(\underline{y}_{ji} - \underline{\mathbf{x}}_{ji}^T\boldsymbol{\beta}), \ldots, \sum_{i=1}^{N}\sum_{j=1}^{g-1} w_{ji[t]} h_k(\underline{y}_{ji} - \underline{\mathbf{x}}_{ji}^T\boldsymbol{\beta})\right]^T \quad , \\ \text{while } \mathbf{Z}_t^0 &= \left[\int_{-\infty}^{\infty} h_1(x) dF_t^0(x), \ldots, \int_{-\infty}^{\infty} h_k(x) dF_t^0(x)\right]^T \quad , \end{aligned}$$

where $h_1, h_2, \ldots, h_k$ is a fixed choice of odd functions, that is $h_l(x) = -h_l(-x)$, for $x \neq 0$, and $h_l(0) = 0$.

The QD estimator for the multinomial logistic model (QDM) is the vector $\hat{\boldsymbol{\beta}}_M$ which minimizes the following sum of quadratic forms

$$\begin{aligned} d_M(\boldsymbol{\beta}) \;=\; & (\mathbf{Z}_1^{\boldsymbol{\beta}} - \mathbf{Z}_1^0)^T \underline{\mathbf{Q}} (\mathbf{Z}_1^{\boldsymbol{\beta}} - \mathbf{Z}_1^0) + \cdots + \\ & (\mathbf{Z}_{p(g-1)}^{\boldsymbol{\beta}} - \mathbf{Z}_{p(g-1)}^0)^T \underline{\mathbf{Q}} (\mathbf{Z}_{p(g-1)}^{\boldsymbol{\beta}} - \mathbf{Z}_{p(g-1)}^0) \quad , \end{aligned} \tag{4.28}$$

where $\underline{\mathbf{Q}}$ denotes a $k \times k$ constant, symmetric, positive-definite matrix.

Furthermore, since $\mathbf{Z}_t^0 = 0$ for $t = 1, \dots, p(g-1)$, when h is odd, minimizing (4.28), with respect to $\boldsymbol{\beta}$, is reduced to minimizing

$$d_M(\boldsymbol{\beta}) \;=\; [\mathbf{Z}_1^{\boldsymbol{\beta}}]^T \underline{\mathbf{Q}} \mathbf{Z}_1^{\boldsymbol{\beta}} + \cdots + [\mathbf{Z}_{p(g-1)}^{\boldsymbol{\beta}}]^T \underline{\mathbf{Q}} \mathbf{Z}_{p(g-1)}^{\boldsymbol{\beta}} \quad . \tag{4.29}$$

Using Kronecker's product notation and calling $\mathbf{Z}^{\boldsymbol{\beta}} = ([\mathbf{Z}_1^{\boldsymbol{\beta}}]^T, \dots, [\mathbf{Z}_{p(g-1)}^{\boldsymbol{\beta}}]^T)^T$, then (4.29) can be expressed more concisely as

$$d_M(\boldsymbol{\beta}) \;=\; [\mathbf{Z}^{\boldsymbol{\beta}}]^T (\mathbf{I}_{p(g-1)} \otimes \underline{\mathbf{Q}}) \mathbf{Z}^{\boldsymbol{\beta}} \quad , \tag{4.30}$$

where $\mathbf{I}_{p(g-1)}$ denotes the identity matrix of order $p(g-1)$.

The QDM estimator $\hat{\boldsymbol{\beta}}_M$ is the vector which minimizes (4.30) with respect to $\boldsymbol{\beta}$.

4.2.1 Asymptotic Properties of the QDM Estimator

In this section we derive the asymptotic properties of the QDM estimators, such as consistency and asymptotic normality.

The derivation is obtained on the results of Section 4.1.1 which we have adapted to multinomial logistic regression.

Definition 4.2.3.

Let $\underline{\mathbf{W}} = (\underline{\mathbf{W}}_1^T, \dots, \underline{\mathbf{W}}_N^T)^T$ be the $N(g-1) \times p(g-1)$ matrix of weights used in (4.26), where

$$\underline{\mathbf{W}}_i^T = \begin{pmatrix} w_{1i[1]} & \cdots & w_{1i[p]} & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & \cdots & w_{(g-1)i[p(g-2)+1]} & \cdots & w_{(g-1)i[p(g-1)]} \end{pmatrix} ,$$

for $i = 1, \dots, N$.

In order to prove this asymptotic properties we need to consider the following regulatory conditions.

Assumptions for Asymptotic Properties

(*b*1) $\lim_{N\to\infty}\sum_{i=1}^{N}\sum_{j=1}^{g-1} w_{ji[t]}^2 = 0$, for each $t = 1, \dots, p(g-1)$,

(*b*2) $\lim_{N\to\infty}(\underline{\mathbf{X}}^T\underline{\mathbf{W}})(\underline{\mathbf{W}}^T\underline{\mathbf{X}})$ exists and is invertible,

(*b*3) $\lim_{N\to\infty}\sum_{i=1}^{N}\sum_{j=1}^{g-1} w_{ji[t]}^2 x_{ti}^2 = 0$, for each $t = 1, \dots, p(g-1)$,

(*b*4) $\lim_{N\to\infty}\sum_{i=1}^{N}\sum_{j=1}^{g-1}\left|w_{ji[t]}x_{ti}\right|$ exists for each $t = 1, \dots, p(g-1)$,

(*b*5) $\dot{h}_i(x) = \frac{\partial}{\partial x}h_i(x)$ is uniformily continuous and $\mathrm{Var}[\dot{h}(\underline{r})] < \infty$,

(*b*6) the x_{ti} values belong to a compact set,

(*b*7) $\max_{1\le i\le N}\{\mathbf{w}_i\Sigma^*\mathbf{w}_i^T\}$ is bounded for all N,

(*b*8) $\underline{\lambda}(\underline{\mathbf{W}}^T\underline{\mathbf{W}} \otimes \Sigma^*) \longrightarrow \infty$ if $N \longrightarrow \infty$, where $\underline{\lambda}(M)$ represents the smallest eigenvalue of matrix $\mathbf{M}$ and $\Sigma^* = n_i^{-1}\Sigma_0^*$ is the variance-covariance matrix of the function

$$h(\underline{r}) \quad = \quad [h_1(\underline{r}), \dots, h_k(\underline{r})]^T \quad ,$$

where Σ_0^* is a function of (4.22) and $\underline{r}$ are asymptotically normal distributed with variance-covariance matrix given in Theorem 4.2.1.

Theorem 4.2.2. [Consistency]

Consider the matrix of weights $\underline{\mathbf{W}}$ defined above and the $N(g-1) \times p(g-1)$ matrix $\underline{\mathbf{X}}$ given in Definition 4.2.2. Matrices $\underline{\mathbf{W}}$ and $\underline{\mathbf{X}}$ are assumed to have rank $p(g-1)$, respectively. Suppose that the weights matrix $\underline{\mathbf{W}}$ satisfies assumption **(*b*1)**, then the QDM estimator $\hat{\boldsymbol{\beta}}_M$, obtained minimizing the function $d_M(\boldsymbol{\beta})$, is consistent.

Proof.

Using Chebyshev's inequality and assumption **(b1)**, we have that $\mathbf{Z}^{\boldsymbol{\beta}_0^*} \xrightarrow{P} 0$ provided that the density function of the random errors, f_0^*, is symmetric. This implies that both

$$d_M(\boldsymbol{\beta}_0^*) \xrightarrow{P} 0 \quad \text{and} \quad d_M(\hat{\boldsymbol{\beta}}_M) \xrightarrow{P} 0, \quad \text{as} \quad N \longrightarrow \infty \quad .$$

Therefore, the consistency of $\hat{\boldsymbol{\beta}}_M$ is guaranteed as long as $\mathbb{E}(\mathbf{Z}^{\boldsymbol{\beta}}) = 0$ at, and only at $\boldsymbol{\beta} = \boldsymbol{\beta}_0^*$, when the parametric space is compact. ■

Theorem 4.2.3. [Asymptotic Normality]

Under assumptions **(b2)** to **(b8)**, the asymptotic distribution of the QDM estimator $\hat{\boldsymbol{\beta}}_M$ is given by

$$(\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_0^*) \xrightarrow{L} \mathrm{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_3\right) \quad , \tag{4.31}$$

with variance-covariance matrix $\boldsymbol{\Sigma}_3 = \mathbf{A}_3(\underline{\mathbf{W}}^T\underline{\mathbf{W}})\mathbf{A}_3^T(\underline{\mathbf{S}}_0^T\underline{\mathbf{Q}}\boldsymbol{\Sigma}_0^*\underline{\mathbf{Q}}\underline{\mathbf{S}}_0)$, where
$\mathbf{A}_3 = (\underline{\mathbf{S}}_0^T\underline{\mathbf{Q}}\underline{\mathbf{S}}_0)^{-1}[(\underline{\mathbf{X}}^T\underline{\mathbf{W}})(\underline{\mathbf{W}}^T\underline{\mathbf{X}})]^{-1}(\underline{\mathbf{X}}^T\underline{\mathbf{W}})$.

Proof.

The proof is based on the derivation of the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\beta}}_M$ and the use of the multivariate central limit theorem.

Consider $\underline{\mathbf{S}}_0^T = [\mathbb{E}(\dot{h}_1(\underline{r})), \ldots, \mathbb{E}(\dot{h}_k(\underline{r}))]$, where $\dot{h}_i(x) = \frac{\partial}{\partial x}h_i(x)$ and assume that the function d_M, given by (4.30), is differentiable. Then $\hat{\boldsymbol{\beta}}_M$ satisfies the following $p(g-1)$-system of equations

$$\frac{\partial}{\partial \boldsymbol{\beta}}[\mathbf{Z}^{\hat{\boldsymbol{\beta}}_M}]^T(\mathbf{I}_{p(g-1)} \otimes \mathbf{Q})\mathbf{Z}^{\hat{\boldsymbol{\beta}}_M} \;=\; \mathbf{0} \quad . \tag{4.32}$$

Under assumptions $(\boldsymbol{b3})$ to $(\boldsymbol{b6})$ and using the properties of Kronecker's product, we obtain that

$$\frac{\partial}{\partial\boldsymbol{\beta}}\,\mathbf{Z}^{\hat{\boldsymbol{\beta}}_M} = \frac{\partial}{\partial\boldsymbol{\beta}}\,\mathbf{Z}^{\boldsymbol{\beta}_0^*} + o_p(1) \quad , \tag{4.33}$$

$$\frac{\partial}{\partial\boldsymbol{\beta}}\,\mathbf{Z}^{\boldsymbol{\beta}_0^*} = -\underline{\mathbf{W}}^T\underline{\mathbf{X}} \otimes \underline{\mathbf{S}}_0 + o_p(1) \quad , \tag{4.34}$$

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{\beta}}\,[\mathbf{Z}^{\boldsymbol{\beta}_0^*}]^T(\mathbf{I}_{\mathbf{p(g-1)}} \otimes \underline{\mathbf{Q}})\frac{\partial}{\partial\boldsymbol{\beta}}\,\mathbf{Z}^{\boldsymbol{\beta}_0^*} &= (\underline{\mathbf{X}}^T\underline{\mathbf{W}} \otimes \underline{\mathbf{S}}_0^T)(\mathbf{I}_{\mathbf{p(g-1)}} \otimes \underline{\mathbf{Q}})(\underline{\mathbf{W}}^T\underline{\mathbf{X}} \otimes \underline{\mathbf{S}}_0^T) + o_p(1)\\ &= (\underline{\mathbf{X}}^T\underline{\mathbf{W}})(\underline{\mathbf{W}}^T\underline{\mathbf{X}})(\underline{\mathbf{S}}_0^T\underline{\mathbf{Q}}\underline{\mathbf{S}}_0) + o_p(1) \quad , \end{aligned}\tag{4.35}$$

where $o_p(1)$ stands for a random infinitesimal term converging in probability.

Substituting (4.34) and (4.35) in (4.32) and using a Taylor's series expansion, we have

$$(\underline{\mathbf{S}}_0^T\underline{\mathbf{Q}}\underline{\mathbf{S}}_0)(\underline{\mathbf{X}}^T\underline{\mathbf{W}})(\underline{\mathbf{W}}^T\underline{\mathbf{X}})(\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_0^*) = -\sqrt{n_i}\,(\underline{\mathbf{X}}^T\underline{\mathbf{W}} \otimes \underline{\mathbf{S}}_0^T\underline{\mathbf{Q}})\mathbf{Z}^{\boldsymbol{\beta}_0^*} + o_p(1) \quad . \tag{4.36}$$

Since $\mathbf{Z}^{\boldsymbol{\beta}_0^*}$ represents a vector of sums of independent variables, then under assumptions $(\boldsymbol{b7})$, $(\boldsymbol{b8})$ and using the multivariate central limit theorem, we obtain that

$$\sqrt{n_i}\,\mathbf{Z}^{\boldsymbol{\beta}_0^*} \quad \xrightarrow{L} \quad \mathrm{N}\,(\mathbf{0}, \underline{\mathbf{W}}^T\underline{\mathbf{W}} \otimes \boldsymbol{\Sigma}_{\mathbf{0}}^*) \quad . \tag{4.37}$$

From (4.37) and (4.36), we then obtain that

$$\mathrm{Var}[\sqrt{n_i}(\underline{\mathbf{X}}^T\underline{\mathbf{W}} \otimes \underline{\mathbf{S}}_0^T\underline{\mathbf{Q}})\mathbf{Z}^{\boldsymbol{\beta}_0^*}] = (\underline{\mathbf{X}}^T\underline{\mathbf{W}} \otimes \underline{\mathbf{S}}_0^T)(\underline{\mathbf{W}}^T\underline{\mathbf{W}} \otimes \underline{\mathbf{Q}}\boldsymbol{\Sigma}_{\mathbf{0}}^*\underline{\mathbf{Q}})(\underline{\mathbf{W}}^T\underline{\mathbf{X}} \otimes \underline{\mathbf{S}}_0^T)\,.$$

Thus

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_M) = \mathbf{A}_2(\underline{\mathbf{W}}^T\underline{\mathbf{W}} \otimes \underline{\mathbf{Q}}\boldsymbol{\Sigma}_{\mathbf{0}}^*\underline{\mathbf{Q}})\mathbf{A}_2^T \quad ,$$

where $\mathbf{A}_2 = (\underline{\mathbf{S}}_0^T\underline{\mathbf{Q}}\underline{\mathbf{S}}_0)^{-1}[(\underline{\mathbf{X}}^T\underline{\mathbf{W}})(\underline{\mathbf{W}}^T\underline{\mathbf{X}})]^{-1}(\underline{\mathbf{X}}^T\underline{\mathbf{W}} \otimes \underline{\mathbf{S}}_0^T)$.

Finally,

$$\mathbf{\Sigma}_3 = \mathbf{A}_3(\underline{\mathbf{W}}^T\underline{\mathbf{W}})\mathbf{A}_3^T(\underline{\mathbf{S}}_0^T\underline{\mathbf{Q}}\mathbf{\Sigma}_0^*\underline{\mathbf{Q}}\underline{\mathbf{S}}_0) \quad ,$$

where $\mathbf{A}_3 = (\underline{\mathbf{S}}_0^T\underline{\mathbf{Q}}\underline{\mathbf{S}}_0)^{-1}[(\underline{\mathbf{X}}^T\underline{\mathbf{W}})(\underline{\mathbf{W}}^T\underline{\mathbf{X}})]^{-1}(\underline{\mathbf{X}}^T\underline{\mathbf{W}})$.

Therefore $(\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_0^*)$ is asymptotically normal with asymptotic variance-covariance matrix $\mathbf{\Sigma}_3$. ■

Corollary 4.2.1.

The minimum asymptotic variance $\mathbf{\Sigma}_3$ of the QDM estimator $\hat{\boldsymbol{\beta}}_M$ is reached when the weights matrix $\underline{\mathbf{W}} = \underline{\mathbf{X}}(\underline{\mathbf{X}}^T\underline{\mathbf{X}})^{-1}$ and the $k \times k$ matrix $\underline{\mathbf{Q}} = [\mathbf{\Sigma}_0^*]^{-1}$. In that case, $\mathrm{Var}(\hat{\boldsymbol{\beta}}_M) = (\underline{\mathbf{X}}^T\underline{\mathbf{X}})^{-1}(\underline{\mathbf{S}}_0^T[\mathbf{\Sigma}_0^*]^{-1}\underline{\mathbf{S}}_0)^{-1}$.

Proof.

An argument similar to that given for Corollary 4.1.1, but applied to the variance-covariance matrix $\mathbf{\Sigma}_3$ completes the proof. ■

Examples

1. Let $\hat{\boldsymbol{\beta}}$ be the MLE estimate of $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_{g-1}^T)^T$. Consider the variance-covariance matrix of the vector $\hat{\boldsymbol{\beta}}$ given by Amemiya (1985) and the odd function $h(\underline{r}) = \underline{r}$, where $\underline{r}$ is approximately multinornal distributed.

   Using the Corollary 4.2.1 is simple to see that the multinomial quadratic estimator, $\hat{\boldsymbol{\beta}}_M$, is as efficient as the maximum likelihood estimator, $\hat{\boldsymbol{\beta}}$.

2. Consider the odd function

$$h_2(x) = \begin{cases} x & \text{if } |x| \leq M \\ \text{sign}(x)M & \text{if } |x| > M \end{cases} ,$$

where $\underline{r}$ is approximately multinormal distributed and M a constant.

The efficiency of $\hat{\boldsymbol{\beta}}_M$ now depend also the value of truncation, M. Note that if $M \rightarrow \infty$, the $\hat{\boldsymbol{\beta}}_M$ again is as efficient as the $\hat{\boldsymbol{\beta}}$.

Chapter 5

Applications

5.1 Logistic Regression - Householder Data

Consider here a logistic regression example with real data. We illustrate the robustness of the QDE when the data is contaminated with aberrant observations. Also, we show that a more reasonable classification is obtained with the QDE.

5.1.1 Description of Data and Estimation of Parameters

We calculated the QDE for the data set given in Table 5.1. This householder data set is a sample of $15,521$ house values extracted from the 1990 Census of California. House value is an important measure of the "amount at risk" in Homeowner's Insurance and, as such, can be used as a classification variable for these insurance portfolios.

The original data reports 8 explanatory variables for each house value. We have chosen the 2 best predictive variables, which are also the most relevant in our study. For each house value, the householder's income (in 10-thousand US$ units) and the house age (in years) compose the explanatory vector $\mathbf{x}_i$ (including a first dummy for the intercept term).

i	x_i^T				y_i	n_i	i	x_i^T				y_i	n_i
1	1	1	0	1.021	0	168	31	1	1	0	3.566	22	243
2	1	0	1	1.021	0	371	32	1	0	1	3.566	229	254
3	1	0	0	1.021	0	238	33	1	0	0	3.566	199	236
4	1	1	0	1.678	0	201	34	1	1	0	3.820	38	258
5	1	0	1	1.678	0	313	35	1	0	1	3.820	308	341
6	1	0	0	1.678	0	260	36	1	0	0	3.820	248	281
7	1	1	0	1.934	0	181	37	1	1	0	4.070	43	220
8	1	0	1	1.934	11	350	38	1	0	1	4.070	254	274
9	1	0	0	1.934	0	260	39	1	0	0	4.070	183	207
10	1	1	0	2.149	0	187	40	1	1	0	4.308	156	198
11	1	0	1	2.149	12	305	41	1	0	1	4.308	308	317
12	1	0	0	2.149	10	307	42	1	0	0	4.308	221	245
13	1	1	0	2.350	0	231	43	1	1	0	4.608	206	239
14	1	0	1	2.350	21	313	44	1	0	1	4.608	272	277
15	1	0	0	2.350	15	269	45	1	0	0	4.608	294	307
16	1	1	0	2.553	0	219	46	1	1	0	4.953	214	235
17	1	0	1	2.553	25	295	47	1	0	1	4.953	280	280
18	1	0	0	2.553	25	312	48	1	0	0	4.953	286	287
19	1	1	0	2.762	0	233	49	1	1	0	5.334	273	287
20	1	0	1	2.762	24	250	50	1	0	1	5.334	225	225
21	1	0	0	2.762	31	321	51	1	0	0	5.334	280	280
22	1	1	0	2.983	13	270	52	1	1	0	5.800	306	309
23	1	0	1	2.983	37	281	53	1	0	1	5.800	193	193
24	1	0	0	2.983	25	267	54	1	0	0	5.800	276	276
25	1	1	0	3.177	18	235	55	1	1	0	6.429	355	355
26	1	0	1	3.177	104	229	56	1	0	1	6.429	137	137
27	1	0	0	3.177	31	235	57	1	0	0	6.429	287	287
28	1	1	0	3.363	9	180	58	1	1	0	10.897	402	402
29	1	0	1	3.363	184	222	59	1	0	1	10.897	116	116
30	1	0	0	3.363	60	195	60	1	0	0	10.897	257	257

Table 5.1: Householder data set

House age is grouped in three categories through two additional dummy variables as follows:

$$\begin{array}{rcl} 0 \le \text{House Age} < 21 & \longleftrightarrow & 1 \quad 0 \\ 21 \le \text{House Age} < 35 & \longleftrightarrow & 0 \quad 0 \\ 35 \le \text{House Age} < 53 & \longleftrightarrow & 0 \quad 1 \quad . \end{array} \tag{5.1}$$

Different $\mathbf{x}_i$ values define different classes, $i = 1, \ldots, 60$. The corresponding number of house values in class i is denoted n_i. Then the portfolio is divided in two groups, class by class; out of the n_i houses in class i, all those with a value less than the portfolio median value, $173,600$ US\$, belong to the first group ("small houses"), the total number of which is denoted y_i.

Using a multiple logistic regression model we estimate the parameters by two methods: maximum likelihood (MLE) and our QDE (see Table 5.2). No outliers were detected in this data set using the pre-programmed function *lmsreg* (least median of squares robust regression) in S-Plus.

	MLE	QDE
$\hat{\beta}_0$(Constant)	-11.8601971	-9.8922812
$\hat{\beta}_1$(Age 1)	-2.6710579	-2.1039454
$\hat{\beta}_2$(Age 2)	0.7977853	0.7871895
$\hat{\beta}_3$(Income)	3.4529568	2.9682419

Table 5.2: MLE and QDE parameters estimates

The QDE was obtained using the optimal weight matrix $\mathbf{W} = \tilde{\mathbf{X}}(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}$. Since $\mathbf{Q} = \Sigma^{-1}$ depends on the beta parameter values, we used the MLE as initial value to obtain $\mathbf{Q}$ iteratively.

Here the minimization of $d(\boldsymbol{\beta})$ in (4.10), was done with the following choice of functions

$$h_1(x) = \begin{cases} 1 & \text{if} \quad x > 0 \\ 0 & \text{if} \quad x = 0 \\ -1 & \text{if} \quad x < 0 \end{cases} \qquad h_2(x) = \begin{cases} x & \text{if} \quad |x| \leq M \\ \text{sign}(x)M & \text{if} \quad |x| > M \end{cases} \quad . \tag{5.2}$$

Using the %-differences given in Table 5.3 as a measure of comparison between parameter estimates, we conclude that the estimators in Table 5.2 are essentially the same under the two methods when outliers are not present.

	MLE vs QDE
$\hat{\beta}_0$ (Constant)	16.5 %
$\hat{\beta}_1$ (Age 1)	21.2 %
$\hat{\beta}_2$ (Age 2)	1.3 %
$\hat{\beta}_3$ (Income)	14.0 %

Table 5.3: %-Differences between MLE and QDE estimators

5.1.2 Robustness of the QDE

Now we contaminate the data set in Table 5.1 by adding a single outlying class, here the observation $\mathbf{x}_{61}^T = (1, 0, 1, 10897)$, $y_{61} = 10$, with a count of $n_{61} = 500$. This observation is an outlier in both y and x, under the assumptions of a logistic regression model.

The re-calculated parameter estimates for the contaminated logistic regression are given for the two methods in Table 5.4.

	MLE	QDE
$\hat{\beta}_0$(Constant)	-1.6614898	-10.8703211
$\hat{\beta}_1$(Age 1)	-0.7539516	-1.9931438
$\hat{\beta}_2$(Age 2)	-0.2245816	0.6984448
$\hat{\beta}_3$(Income)	0.7415504	3.0060373

Table 5.4: MLE and QDE parameter estimates with outlier

From Tables 5.2 and 5.4 we see that the MLE's are greatly affected by the presence of a single outlying class, while the QDE remains relatively stable. We use here the %-difference as a measure of comparison between parameters estimates.

Table 5.5 clearly illustrates how the QDE is more robust to presence of outliers than the MLE.

	MLE vs $\text{MLE}_{\text{Outlier}}$	QDM vs $\text{QDM}_{\text{Outlier}}$
$\hat{\beta}_0$ (Constant)	86.0 %	9.9 %
$\hat{\beta}_1$ (Age 1)	71.8 %	5.3 %
$\hat{\beta}_2$ (Age 2)	128.2 %	11.3 %
$\hat{\beta}_3$ (Income)	78.5 %	1.3 %

Table 5.5: %-Differences between MLE and QDE estimators due to outlier

5.1.3 Binary Classification

In this section we consider a measure to judge the accuracy of the classification under the fitted model. It is based on the number of observations in the sample that the model classifies correctly.

We simulate a large number of sample data with the same law as the original data. For each re-sampled data, we calculate the correct classification rate (CCR) as an estimate the true CCR.

Consider the original data composed of $15,521$ house values. The data base was partitioned in two groups: those with a value less than $173,600$ US$ belong to the group of small houses (coded group "0"), while those with a value greater than $173,600$ belong to group "1".

The distribution function of the variables house age and income was obtained for both groups as follows.

Analysis of group 1

To find the distribution function of house age and income variables we fitted Weibull, normal and gamma distributions. Using MATLAB we estimated the parameters in each case. Figure 5.1 presents the Quantile-quantile plots (QQ-plot) of the house age variable for a Weibull, normal and gamma distribution function, as well as the normal probability plot.

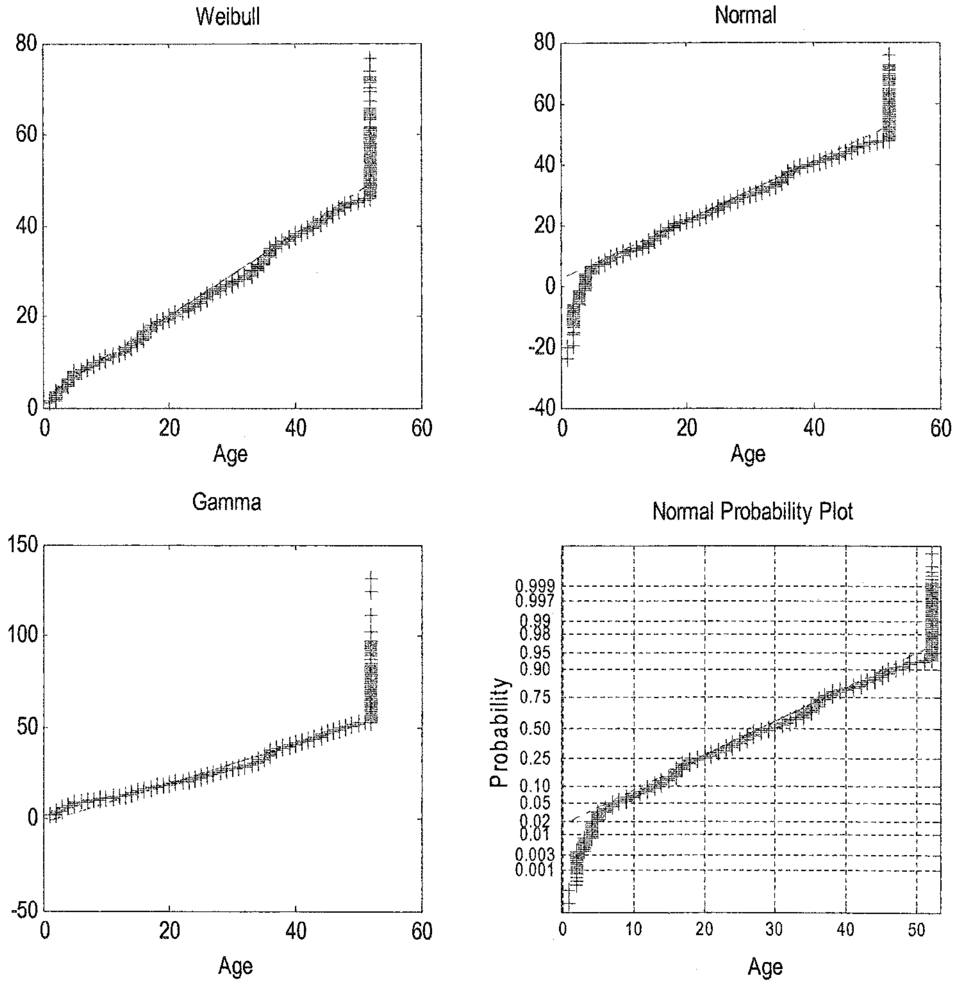


Figure 5.1: QQ-Plots of house age variable

Departures from linearity show how the sample data differs from the considered distributions. A more formal test to determine the best fitting distribution function was also used. Based on Kolmogorov-Smirnov's test we decided that the house age variable can be modeled by a normal distribution with $\mu = 29.1722$ and $\sigma = 12.6238$.

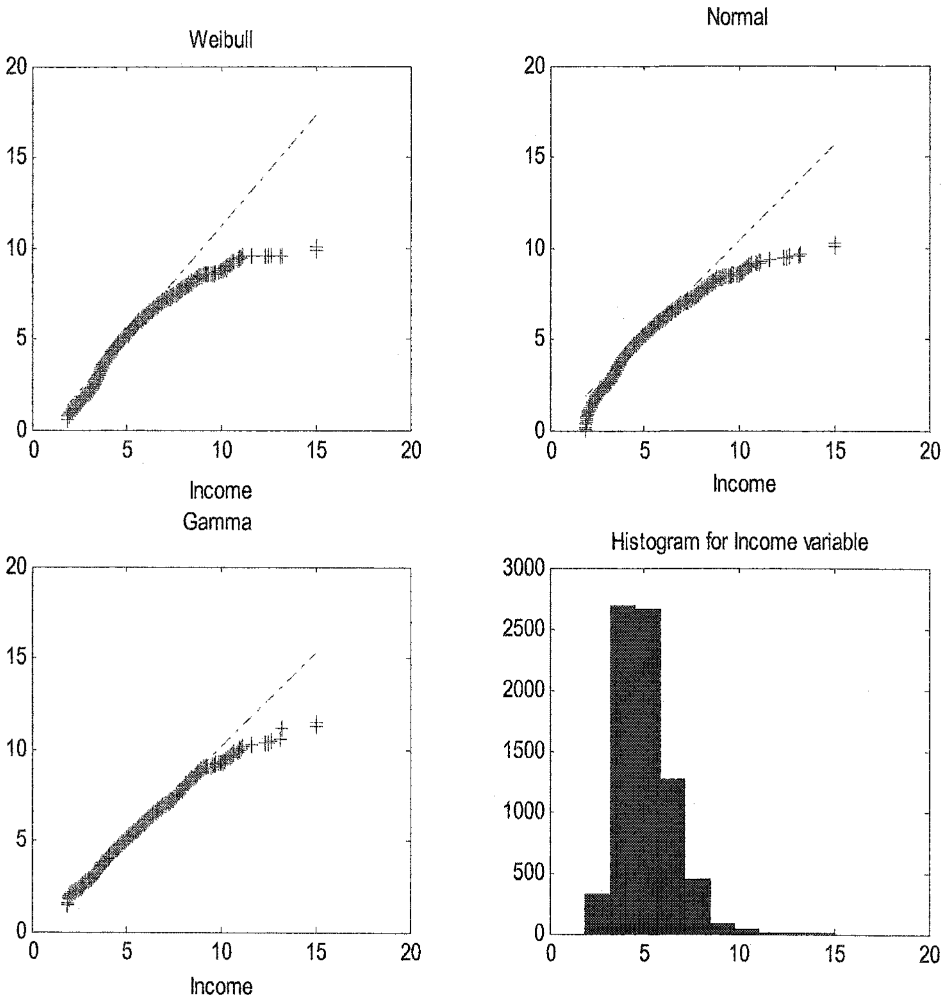
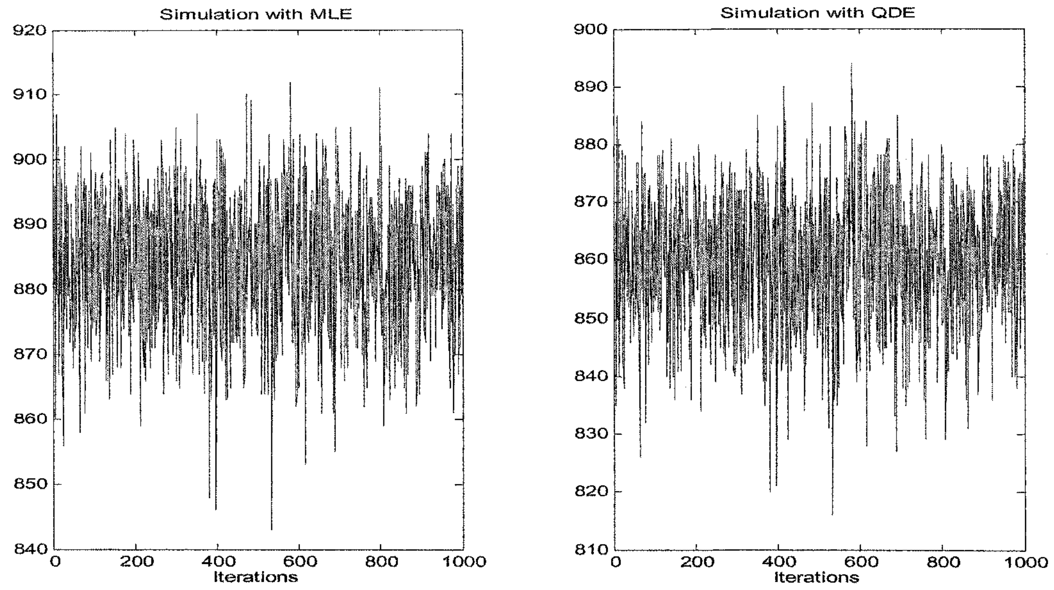


Figure 5.2: QQ-Plots of income variable

A similar procedure was used for the income variable (see Figure 5.2). We conclude that a gamma distribution with $\alpha = 14.1217$ and $\beta = 0.3565$ provides a good fit for this variable.

We now simulate $1,000$ samples of size $1,000$ from a normal($\mu = 29.1722, \sigma = 12.6238$) and a gamma($\alpha = 14.1217, \beta = 0.3565$) distributions, for the house age and income variables, respectively.

Using the MLE and QDE given in Table 5.2 we classify each sample generated using the logistic regression model given in (1.6). Then we calculate the correct classification rate (CCR) as the ratio of the number of $\pi(\mathbf{x}_i) \geq 0.5$ to the total number observations. Figure 5.3 shows the result of this simulation.


Simulation with MLE

Simulation with QDE

Figure 5.3: Simulation of 1000 samples without outliers

Table 5.6 reports the mean CCR values obtained for over the $1,000$ simulated samples. These results confirm our prior evaluation that the MLE and QDE estimators in Table 5.2 produce essentially the same classification when outliers are not present.

Classification Rate	
MLE	QDE
88.42 %	85.98 %

Table 5.6: Estimated CCR's without outlier

By contrast, when outliers are present in the simulated samples, then the resulting classifications differ. Using the MLE an QDE given in Table 5.4 we repeat the above classification of simulated samples. We can see in Figure 5.4 that under presence of outliers, the correct classification rate remains stable for the QDE but changes drastically for the MLE, as we expected.

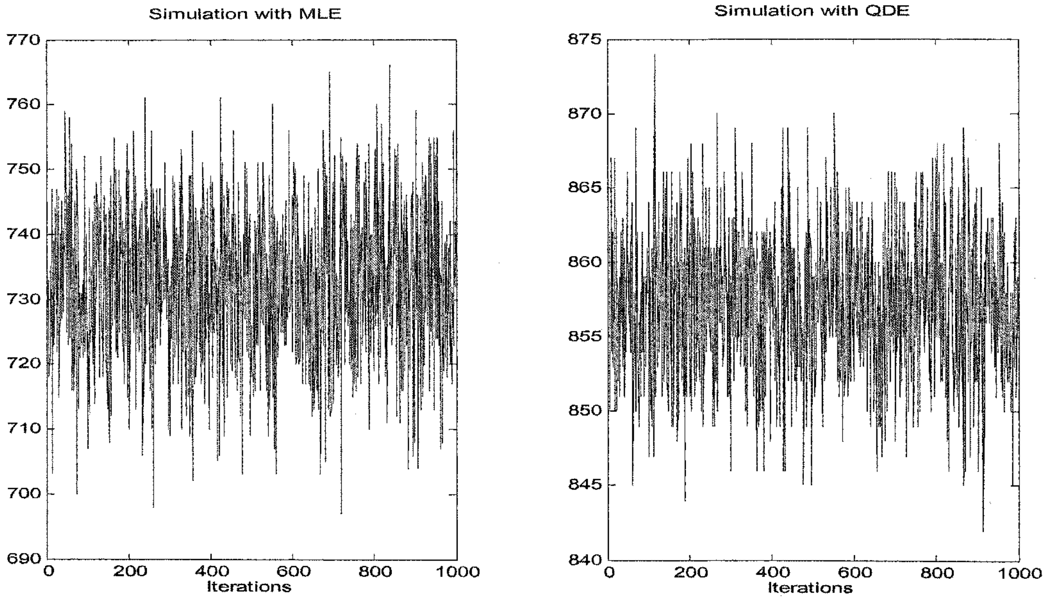


Figure 5.4: Simulation of 1000 samples with outliers

Table 5.7 reports the comparison between the CCR obtained under the presence of outliers and the CCR without outliers.

	Group 1	
	MLE	QDE
Without outliers	88.42 %	85.98 %
With outliers	73.21 %	85.73 %

Table 5.7: CCR comparison for group 1

Analysis of group 0

A similar analysis to that in the previous section was carried out for group 0. First, we found the distribution functions of the house age and income variables. In both cases the best fit was provided by a Weibull. For the house age variable the parameters are $\alpha = 0.0002$ and $\theta = 2.4101$, while for the income variable the parameters are $\alpha = 0.0282$ and $\theta = 3.4625$.

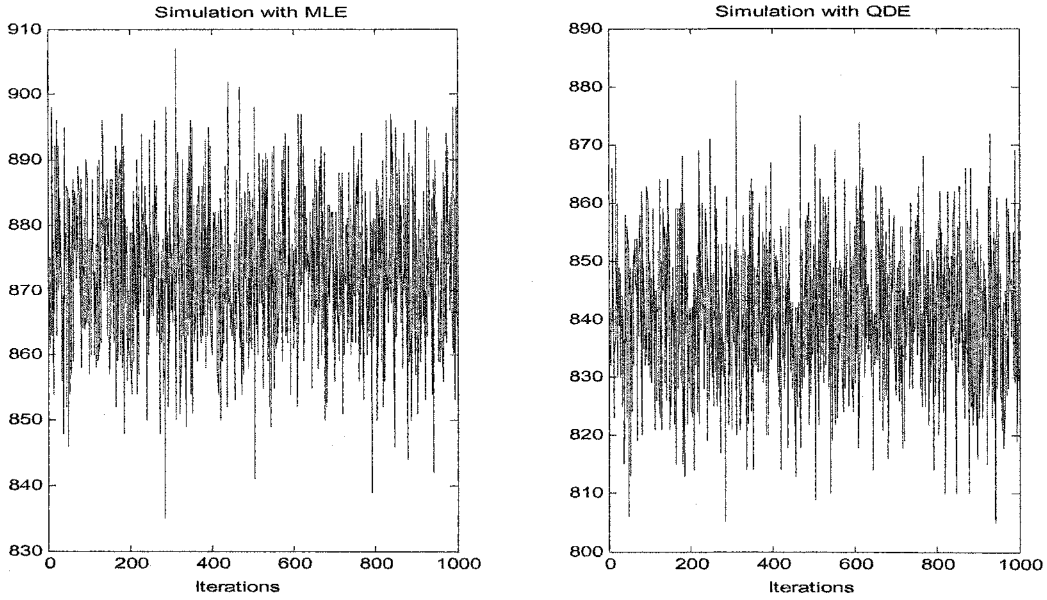


Figure 5.5: Simulation of 1000 samples without outliers

The conclusions are same as for group 1. When the simulated samples are not contaminated by outliers, the correct classification rate is similar for both estimators, the MLE and QDE. Plots in Figure 5.5 confirm this conclusion.

However, the presence of outliers in the generated samples greatly affects the CCR under the MLE, while the CCR under the QDE remains relatively stable.

Graphical and numerical results are shown in Figure 5.6 and Table 5.8.

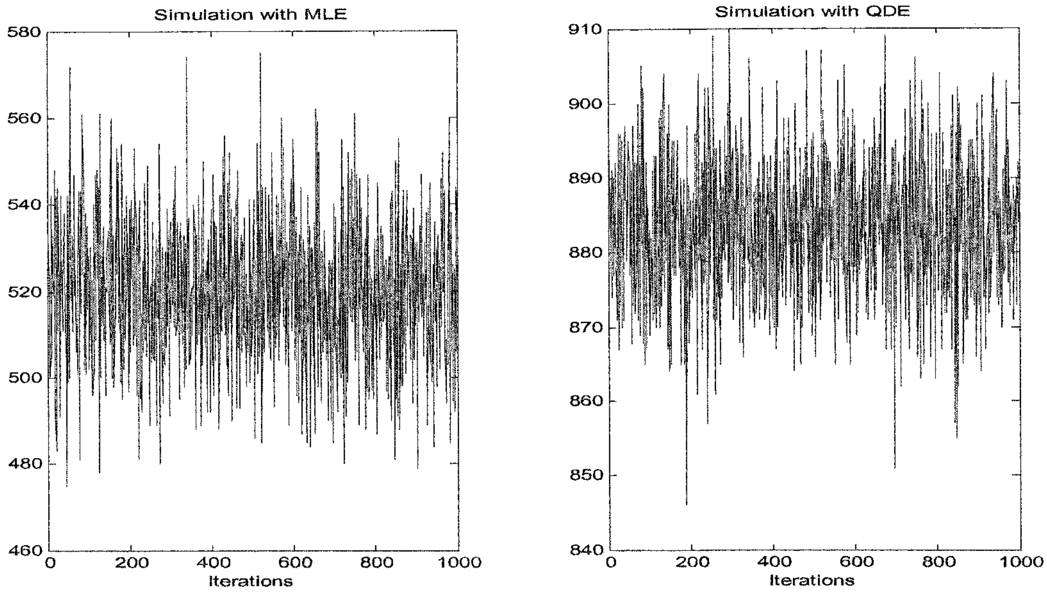


Figure 5.6: Simulation of 1000 samples with outliers

	Group 0	
	MLE	QDE
Without	87.33 %	84.11 %
With	51.95 %	88.29 %

Table 5.8: CCR comparison for group 0

5.2 Multinomial Logistic Regression Example

This section gives an illustration of the classification method with a multinomial logistic regression model. The method works for any number $g \geq 2$ of groups, but to simplify the notation, consider three categories in finding the multinomial MLE and QDM estimators.

The QDM estimator shows to be a robust estimator under the presence of outliers. The classification obtained using the QDM estimator is also more accurate than with the MLE when the data are contaminated. This suggests that the QDM will become an important alternative to the MLE for classification purposes.

5.2.1 Data and Estimation of Parameters

Consider the data used in Section 5.1 and define three categories for the variable house value, as follows:

$$\begin{aligned}
14,999 \leq \text{House Value} < 127,100 \quad &\longleftrightarrow \quad 0 \\
127,100 \leq \text{House Value} < 225,400 \quad &\longleftrightarrow \quad 1 \\
225,400 \leq \text{House Value} < 499,101 \quad &\longleftrightarrow \quad 2
\end{aligned}$$

The above boundary points correspond to the 33% and 67% sample percentiles, respectively.

For each house value, the variables house age and householder's income (in 10-thousand US$ units) are again used in the explanatory vector $\mathbf{x}_i$ (including a first dummy for the intercept term). The variable house age is used in categorical form, as given in (5.1) of the previous section. Different $\mathbf{x}_i$ values define different classes, $i = 1, \ldots, 180$. As before, the number of house values in class i is denoted

n_i, while the total number out of these in a group is denoted y_i. The data is reported in full detail in Table 5.13 at end the section.

Using multinomial logistic regression with these three categories, we again estimate the parameters by two methods: MLE and our QDM. The estimated parameters are reported in Table 5.9. No outliers were detected using the pre-programmed function *lmsreg* (least median of squares robust regression) in S-Plus.

Here the QDM was obtained using the optimal weight matrix $\underline{\mathbf{W}} = \underline{\mathbf{X}}(\underline{\mathbf{X}}^T\underline{\mathbf{X}})^{-1}$. Since $\underline{\mathbf{Q}} = [\Sigma^*]^{-1}$ depends on the beta parameters values, again we used the MLE as an initial value to obtain $\underline{\mathbf{Q}}$ iteratively.

Finally, the minimization of $d_M(\boldsymbol{\beta})$ in (4.30) was done with the choice of functions given in (5.2).

	MLE	QDM
$\hat{\beta}_{00}$	9.0023521	9.2730634
$\hat{\beta}_{10}$	1.2299739	1.2965890
$\hat{\beta}_{20}$	-0.8731412	-0.7370614
$\hat{\beta}_{30}$	-2.5493312	-2.5350713
$\hat{\beta}_{01}$	5.0019578	5.1912438
$\hat{\beta}_{11}$	0.7963158	0.8490555
$\hat{\beta}_{21}$	-0.5658412	-0.4702797
$\hat{\beta}_{31}$	-1.1578578	-1.1828727

Table 5.9: MLE and QDM parameters estimates

Based on the %-differences given in Table 5.10 as a measure of comparison between parameter estimates, we conclude that here also, the estimators in Table 5.9 are essentially the same under the two methods when outliers are not present.

	MLE vs QDM
$\hat{\beta}_{00}$	3.01 %
$\hat{\beta}_{10}$	5.42 %
$\hat{\beta}_{20}$	15.59 %
$\hat{\beta}_{30}$	0.56 %
$\hat{\beta}_{01}$	3.78 %
$\hat{\beta}_{11}$	6.62 %
$\hat{\beta}_{21}$	16.89 %
$\hat{\beta}_{31}$	2.16 %

Table 5.10: %-Differences between MLE and QDM estimators

5.2.2 Robustness of the QDM

We introduce an outlier in each group of the multinomial classification given in Table 5.13. These outliers are given by

	Age 1	Age 2	Income	y_i	n_i
Group 0	1	0	15.5	300	410
Group 1	1	0	15.5	10	410
Group 2	1	0	15.5	100	410

The re-calculated parameter estimates for the contaminated multinomial logistic regression are given, for each of the two methods, in Table 5.11.

	MLE	QDM
$\hat{\beta}_{00}$	3.0556782	7.8330769
$\hat{\beta}_{10}$	1.0331448	0.9001147
$\hat{\beta}_{20}$	-0.3373748	-0.5296588
$\hat{\beta}_{30}$	-0.8565958	-2.2446234
$\hat{\beta}_{01}$	1.1709443	5.1914493
$\hat{\beta}_{11}$	0.8080413	0.8383701
$\hat{\beta}_{21}$	-0.2007203	-0.4498738
$\hat{\beta}_{31}$	-0.3049192	-1.1819892

Table 5.11: MLE and QDM parameters estimates with outliers

From Tables 5.9 and 5.11 we see clearly that the MLE's are extremely influenced by the presence of outliers, while the QDM's remain relatively stable. Table 5.12 reports the %-differences between parameter estimates.

	MLE vs $\text{MLE}_{\text{Outlier}}$	QDM vs $\text{QDM}_{\text{Outlier}}$
$\hat{\beta}_{00}$	66.06 %	15.53 %
$\hat{\beta}_{10}$	16.00 %	30.58 %
$\hat{\beta}_{20}$	61.36 %	28.14 %
$\hat{\beta}_{30}$	66.40 %	11.46 %
$\hat{\beta}_{01}$	76.59 %	0.004 %
$\hat{\beta}_{11}$	1.47 %	1.26 %
$\hat{\beta}_{21}$	64.53 %	4.34 %
$\hat{\beta}_{31}$	73.67 %	0.07 %

Table 5.12: %-Differences between MLE and QDM estimators due to outliers

5.3 Summary and Conclusion

This chapter illustrates the robustness of the QDE and QDM against outliers. Estimations using the Householder data set clearly show that even if MLE's are the standard estimators in the logistic regression literature, it can be useful to have alternative estimators like the QDE and its multinomial version, the QDM.

MLE's are simple to calculate since pre-programmed functions for it are available with various statistical softwares. In addition, MLE's satisfy desirable optimality properties such as consistency, sufficiency and asymptotic normality.

By contrast, the QDM calculations require special algorithms, which are less direct than for the MLE. Also, its properties are all asymptotic. However, if the data presents outlying observations, the QDE definitely provides more stable results. While using the QDM over the MLE implies a certain cost in optimality properties, it offers a clear gain in robustness.

In risk classification for Insurance portfolios, where outlying values are common, the QDM should provide more reasonable groupings.

i	x_i^T				y_i	n_i	i	x_i^T				y_i	n_i
1	1	1	0	1.021	130	168	36	1	0	0	2.149	4	307
2	1	1	0	1.021	38	168	37	1	1	0	2.350	168	231
3	1	1	0	1.021	0	168	38	1	1	0	2.350	63	231
4	1	0	1	1.021	322	371	39	1	1	0	2.350	0	231
5	1	0	1	1.021	49	371	40	1	0	1	2.350	188	313
6	1	0	1	1.021	0	371	41	1	0	1	2.350	114	313
7	1	0	0	1.021	211	238	42	1	0	1	2.350	11	313
8	1	0	0	1.021	27	238	43	1	0	0	2.350	180	269
9	1	0	0	1.021	0	238	44	1	0	0	2.350	85	269
10	1	1	0	1.678	173	201	45	1	0	0	2.350	4	269
11	1	1	0	1.678	28	201	46	1	1	0	2.553	141	219
12	1	1	0	1.678	0	201	47	1	1	0	2.553	78	219
13	1	0	1	1.678	268	313	48	1	1	0	2.553	0	219
14	1	0	1	1.678	45	313	49	1	0	1	2.553	166	295
15	1	0	1	1.678	0	313	50	1	0	1	2.553	119	295
16	1	0	0	1.678	219	260	51	1	0	1	2.553	10	295
17	1	0	0	1.678	41	260	52	1	0	0	2.553	172	312
18	1	0	0	1.678	0	260	53	1	0	0	2.553	132	312
19	1	1	0	1.934	150	181	54	1	0	0	2.553	8	312
20	1	1	0	1.934	31	181	55	1	1	0	2.762	162	233
21	1	1	0	1.934	0	181	56	1	1	0	2.762	71	233
22	1	0	1	1.934	253	350	57	1	1	0	2.762	0	233
23	1	0	1	1.934	91	350	58	1	0	1	2.762	103	250
24	1	0	1	1.934	6	350	59	1	0	1	2.762	134	250
25	1	0	0	1.934	203	260	60	1	0	1	2.762	13	250
26	1	0	0	1.934	57	260	61	1	0	0	2.762	175	321
27	1	0	0	1.934	0	260	62	1	0	0	2.762	133	321
28	1	1	0	2.149	134	187	63	1	0	0	2.762	13	321
29	1	1	0	2.149	53	187	64	1	1	0	2.983	145	270
30	1	1	0	2.149	0	187	65	1	1	0	2.983	122	270
31	1	0	1	2.149	197	305	66	1	1	0	2.983	3	270
32	1	0	1	2.149	100	305	67	1	0	1	2.983	118	281
33	1	0	1	2.149	8	305	68	1	0	1	2.983	148	281
34	1	0	0	2.149	225	307	69	1	0	1	2.983	15	281
35	1	0	0	2.149	78	307	70	1	0	0	2.983	135	267

Table 5.13: Householder multinomial data set

i	x_i^T				y_i	n_i	i	x_i^T				y_i	n_i
71	1	0	0	2.983	121	267	106	1	0	0	3.820	10	281
72	1	0	0	2.983	11	267	107	1	0	0	3.820	157	281
73	1	1	0	3.177	129	235	108	1	0	0	3.820	114	281
74	1	1	0	3.177	97	235	109	1	1	0	4.070	65	220
75	1	1	0	3.177	9	235	110	1	1	0	4.070	134	220
76	1	0	1	3.177	55	229	111	1	1	0	4.070	21	220
77	1	0	1	3.177	118	229	112	1	0	1	4.070	7	274
78	1	0	1	3.177	56	229	113	1	0	1	4.070	106	274
79	1	0	0	3.177	111	235	114	1	0	1	4.070	161	274
80	1	0	0	3.177	109	235	115	1	0	0	4.070	7	207
81	1	0	0	3.177	15	235	116	1	0	0	4.070	92	207
82	1	1	0	3.363	99	180	117	1	0	0	4.070	108	207
83	1	1	0	3.363	77	180	118	1	1	0	4.308	10	198
84	1	1	0	3.363	4	180	119	1	1	0	4.308	102	198
85	1	0	1	3.363	11	222	120	1	1	0	4.308	86	198
86	1	0	1	3.363	98	222	121	1	0	1	4.308	1	317
87	1	0	1	3.363	113	222	122	1	0	1	4.308	118	317
88	1	0	0	3.363	78	195	123	1	0	1	4.308	198	317
89	1	0	0	3.363	89	195	124	1	0	0	4.308	5	245
90	1	0	0	3.363	28	195	125	1	0	0	4.308	133	245
91	1	1	0	3.566	98	243	126	1	0	0	4.308	107	245
92	1	1	0	3.566	133	243	127	1	1	0	4.608	8	239
93	1	1	0	3.566	12	243	128	1	1	0	4.608	133	239
94	1	0	1	3.566	10	254	129	1	1	0	4.608	98	239
95	1	0	1	3.566	114	254	130	1	0	1	4.608	2	277
96	1	0	1	3.566	130	254	131	1	0	1	4.608	98	277
97	1	0	0	3.566	15	236	132	1	0	1	4.608	177	277
98	1	0	0	3.566	126	236	133	1	0	0	4.608	2	307
99	1	0	0	3.566	95	236	134	1	0	0	4.608	130	307
100	1	1	0	3.820	90	258	135	1	0	0	4.608	175	307
101	1	1	0	3.820	152	258	136	1	1	0	4.953	3	235
102	1	1	0	3.820	16	258	137	1	1	0	4.953	108	235
103	1	0	1	3.820	8	341	138	1	1	0	4.953	124	235
104	1	0	1	3.820	135	341	139	1	0	1	4.953	0	280
105	1	0	1	3.820	198	341	140	1	0	1	4.953	85	280

Table 5.13: Householder multinomial data set (...continued)

i	x_i^T				y_i	n_i	i	x_i^T				y_i	n_i
141	1	0	1	4.953	195	280	161	1	0	0	5.800	41	276
142	1	0	0	4.953	0	287	162	1	0	0	5.800	235	276
143	1	0	0	4.953	106	287	163	1	1	0	6.429	0	355
144	1	0	0	4.953	181	287	164	1	1	0	6.429	35	355
145	1	1	0	5.334	2	287	165	1	1	0	6.429	320	355
146	1	1	0	5.334	126	287	166	1	0	1	6.429	0	137
147	1	1	0	5.334	159	287	167	1	0	1	6.429	4	137
148	1	0	1	5.334	0	225	168	1	0	1	6.429	133	137
149	1	0	1	5.334	46	225	169	1	0	0	6.429	0	287
150	1	0	1	5.334	179	225	170	1	0	0	6.429	18	287
151	1	0	0	5.334	0	280	171	1	0	0	6.429	269	287
152	1	0	0	5.334	82	280	172	1	1	0	10.897	0	402
153	1	0	0	5.334	198	280	173	1	1	0	10.897	12	402
154	1	1	0	5.800	1	309	174	1	1	0	10.897	390	402
155	1	1	0	5.800	83	309	175	1	0	1	10.897	0	116
156	1	1	0	5.800	225	309	176	1	0	1	10.897	3	116
157	1	0	1	5.800	0	193	177	1	0	1	10.897	113	116
158	1	0	1	5.800	17	193	178	1	0	0	10.897	0	257
159	1	0	1	5.800	176	193	179	1	0	0	10.897	4	257
160	1	0	0	5.800	0	276	180	1	0	0	10.897	253	257

Table 5.13: Householder multinomial data set (...continued)

Conclusion

Risk classification is an important part of the selection and underwriting process in Insurance companies.

For the most part, companies currently use ad-hoc methods for risk classification, more often based on the type of expenses covered than on the distribution of the corresponding losses. The selection of classification variables is also, in general, based on rate-making variables rather than on an optimal choice criteria based on statistical methods.

This thesis reviews the use of statistical methods to classify risks. Two issues arise, which make this actuarial classification problem different from applications to other fields.

First, Insurance risks are not usually classified in only two categories, good and bad, as can be the case in credit rating, but in a larger number of groups.

But most importantly, Insurance data presents catastrophic losses and heavy tailed claim distributions, forcing the use of a robust estimation analysis.

The main result in this thesis is the generalization of logistic regression models to multinomial groups in a robust estimation framework.

Chapters 3 and 4 review the main robust regression estimators encoutered in the

Statistics literature and discuss robust logistic regression. We define the minimum quadratic distance estimator for the multinomial logistic regression model (QDM), deriving the asymptotic properties of our QDM, like consistency, asymptotic normality and robustness.

Finally, Chapter 5 illustrates the proposed method with an application to the classification of the Householder data set. It clearly shows that even if MLE's are a standard in the logistic regression literature, it is useful to have alternative estimators like the QDM.

In risk classification for Insurance portfolios, where outlying values are common, the QDM should provide more reasonable groupings.

Bibliography

[1] Albert, A. (1972), *Regression and the Moore-Penrose Inverse.* Academic Press, New-York.

[2] Amemiya, T. (1985), *Advanced Econometrics.* Boston: Harvard Press.

[3] Bianco, A. M. and Yohai, V. J. (1996), "Robust estimation in the logistic regression model". *Lecture Notes in Statistics*, Vol. 109, Springer-Verlag, New-York.

[4] Birkes, D. and Dodge, Y. (1993), *Alternative Methods of Regression.* Wiley, New-York.

[5] Carroll, R. J. and Pederson, S. (1993), "On robustness in the logistic regression model". *Biometrika*, **55**, 693-706.

[6] Christmann, A. (1994), "Least median of weighted squares in logistic regression with large strata". *Biometrika*, **81**, 413-417.

[7] Cook, R. D. (1977), "Detection of influential observations in linear regression". *Technometrics*, **19**, 15-18.

[8] Cook, R. D. and Weisberg, S. (1982), *Residuals and Influence in Regression.* Chapman & Hall, London.

[9] Copas, J. B. (1988), "Binary regression models for contaminated data". *Journal Royal Statistical Society B*, **50**, 225-265.

[10] Cornfield, J. (1962), "Joint dependence of the risk of coronary heart disease on serum cholesterol and systolic blood pressure: A discriminant function analysis". *Federation Proceedings*, **21**, 58-61.

[11] Cox, D. R. and Snell, E. J. (1989), *The Analysis of Binary Data*, 2nd edition. Chapman & Hall, London.

[12] Graham, A. (1981), *Kronecker Products and Matrix Calculus: with Applications.* Halsted Press, Chichester.

[13] Grizzle, J., Starmer, F. and Koch, G. (1969), "Analysis of categorical data by linear models". *Biometrics*, **25**, 489-504.

[14] Hampel, F. R. (1973), "Robust estimation: A condensed partial survey". *Z. Wahrsch. Verw. Geb.*, **27**, 87-104.

[15] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. J. (1986), *Robust Statistics, the Approach Based on Influence Functions.* Wiley, New-York.

[16] Hoaglin, D. C. and Welsch, R. E. (1978), "The hat matrix in regression and ANOVA". *The American Statistician*, **32**, 17-22.

[17] Hosmer, D. W. and Lemeshow, S. (2000), *Applied Logistic Regression*, 2nd edition. Wiley, New-York.

[18] Huber, P. J. (1964), "Robust estimation of a location parameter". *Annals of Mathematical Statistics*, **35**, 73-101.

[19] Huber, P. J. (1968), "Robust confidence limits". *Z. Wahrsch. Verw. Geb.*, **10**, 269-278.

[20] Huber, P. J. (1973), "Robust regression: Asymptotics, conjectures, and Monte Carlo". *Annals of Statistics*, **1**, 799-821.

[21] Huber, P. J. (1981), *Robust Statistics.* Wiley, New-York.

[22] Johnson, W. (1985), "Influence measures for logistic regression: another point of view". *Biometrika*, **72**, 59-66.

[23] Künsch, H. R., Stefanski, L. A. and Carroll, R. J. (1989), "Conditionally unbiased influence estimation in general regression models, with applications to generalized linear models". *Journal of the American Statistical Association*, **84**, 460-466.

[24] Lesaffre, E. and Albert, A. (1989a), "Multiple-group logistic regression diagnostics". *Applied Statistics*, **38**, 425-440.

[25] Lesaffre, E. and Albert, A. (1989b), "Partial separation in logistic discrimination". *Journal Royal Statistical Society B*, **51**, 109-116.

[26] Luong, A. (1991), " Minimum distance methods based on quadratic distances for transforms in simple linear regression models". *Journal Royal Statistical Society B*, **53**, 465-471.

[27] Luong, A. and Garrido, J. (1992), "Nonparametric estimation based on minimum quadratic distances for the multiple linear regression model". *Cuadernos Aragoneses de Economia*, **2**, 69-78, (in Spanish).

[28] Luong, A. and Thompson, M. E. (1987), "Minimum distance methods based on quadratic distances for transforms". *Canadian Journal of Statistics*, **15**, 239-251.

[29] Maronna, R. A., and Yohai, V. J. (1981), "Asymptotic behaviour of general M-estimates for regression and scale with random carriers". *Z. Wahrsch. Verw. Geb.*, **58**, 7-20.

[30] McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*, 2nd edition. Chapman & Hall, London.

[31] Nelder, J. A. and Wedderburn, R. W. (1972), "Generalized linear models". *Journal of the Royal Statistical Society A*, **135**, 370-384.

[32] Pierce, D. A. and Schafer, D. W. (1986), "Residuals in generalized linear models". *Journal of the American Statistical Association*, **81**, 977-986.

[33] Pregibon, D. (1981), "Logistic regression diagnostics". *Annals of Statistics*, **9**, 705-724.

[34] Pregibon, D. (1982), "Resistant fits for some commonly used logistic models with medical applications". *Biometrics*, **38**, 485-498.

[35] Rao, C. R. (1965), *Linear Statistical Inference and its Applications.* Wiley, New-York.

[36] Rao, C. R. and Mitra, S. K. (1971), *Generalized Inverse of Matrices with Applications.* Wiley, New-York.

[37] Rousseeuw, P. J. (1984), "Least median of squares regression". *Journal of the American Statistical Association*, **79**, 871-881.

[38] Rousseeuw, P. J. and Leroy, A. (1987), *Robust Regression and Outlier Detection.* Wiley, New-York.

[39] Rousseeuw, P. J., and Yohai, V. J. (1984), "Robust regression by means of S-estimators". *Lecture Notes in Statistics*, **26**, 256-272, Springer-Verlag, Berlin.

[40] Searle, S. R. (1982), *Matrix Algebra Useful for Statistics.* Wiley, New-York.

[41] Stefanski, L. A., Carroll, R. J. and Ruppert, D. (1986), "Optimally bounded score functions for generalized linear models with applications to logistic regression". *Biometrika*, **73**, 413-425.

[42] Williams, D. A. (1987), "Generalized linear model diagnostics using the deviance and single ease deletions ". *Applied Statistics*, **36**, 181-191.

[43] Yohai, V. J. (1987), "High breakdown point and high efficiency robust estimates for regression". *Annals Statistics*, **15**, 642-665.