ABSTRACT

Title of dissertation: ON ROBUSTNESS
IN SOME
EXTENDED REGRESSION MODELS

Gabriela V. Cohen Freue, Doctor of Philosophy, 2004

Dissertation directed by: Professor Paul J. Smith
Department of Mathematics

Generalized Linear Models extends classical regression models to non-normal response variables and allows a non-linear relation between the mean of the responses and the predictors. In addition, when the responses are correlated or show overdispersion, one can add a linear combination of random components to the linear predictor. The resulting models are known as Generalized Linear Mixed Models. Traditional estimation methods in these classes of models rely on distributional assumptions about the random components, as well as the implicit assumption that the explanatory variables are uncorrelated with the error term. In Chapters 2 and 3 we investigate, using the Change-of-Variance Function, the behavior of the asymptotic variance-covariance matrix of the class of M-estimators when the distribution of the random components is slightly contaminated. In Chapter 4 we study a different concept of robustness for classical models that contain explanatory variables correlated with the error term. For these models we propose an instrumental variables estimator and study its robustness by means of its Influence Function.

We extend the definitions of Change-of-Variance Function to Generalized Linear Models and Generalized Linear Mixed Models. We use them to analyze in detail the sensitivity of the asymptotic variance of the maximum likelihood estimator. For the first class of models, we found

that, in general, a contamination of the distribution can seriously affect the asymptotic variance of the estimators. For the second class, we focus on the Poisson-Gamma model and two mixed-effects Binomial models. We found that the effect of a contamination in the mixing distribution on the asymptotic variance of the maximum likelihood estimator remain bounded for both models. A simulation study was performed in all cases to illustrate the relevance of our results.

Finally, we propose a robust instrumental variables estimator based on high breakdown point S-estimators of location and scatter. The resulting estimator has bounded Influence Function and satisfies the usual asymptotic properties for suitable choices of the S-estimator used. We also derive an estimate for the asymptotic covariance matrix of our estimator which is robust against outliers and leverage points. We illustrate our results using a real data example.

ON ROBUSTNESS IN SOME EXTENDED REGRESSION MODELS

by

Gabriela V. Cohen Freue

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2004

Advisory Commmittee:

Associate Professor Paul J. Smith, Chair/Advisor
Professor Grace L. Yang
Professor Abraham M. Kagan
Professor Benjamin J. Kedem
Professor Francis Alt

DEDICATION

In the memory of my grandparents

Maria and Dante

ACKNOWLEDGMENTS

I am very grateful to my advisor, Professor Paul Smith for giving me an invaluable opportunity to work on interesting projects over the past four years. He has always made himself available for help and provided good insights. I would also like to thank Professor Grace Young. She was there from the very beginning, when I was searching for a dissertation topic. She was always very enthusiastic about my work, encourage me and give me invaluable advice throughout these years. Special thanks are extended to the remaining members of my committee: Professor Abraham Kagan and Professor Benjamin Kedem, for their useful comments and suggestions that enhanced the quality of this work.

Professor Ruben Zamar deserves a special mention and my sincere appreciation. I would like to thank him for his support and helpful comments. It has been a pleasure to work with and learn from such an extraordinary individual.

I would also like to acknowledge the staff members of the Mathematical department for creating a friendly environment and being always available for help. In particular I thank Timothy Strobell and Tony Zhang for their help and support with computer softwares.

Many thanks go to my colleague and great friend Ru for being a tremendous source of strength and enriching my graduate life. I thank my friends at Graduate Hills, who have been a crucial factor in my finishing smoothly, specially to Maria Jose and Juan Gabriel for their help with computer problems.

I owe my deepest thanks to my family, especially my parents, Josefina and Jaime, my sister, Silvina, and my brother, Guillermo, who have always stood by me and give me unwavering support and love. I am very grateful to my family in law for always being there and for showing support. I would also like to thank Boris for believing in me and always encouraging me to pursue my dreams.

Words cannot express the gratitude I owe to my husband, Hernan, who has guided me

through my career and has pulled me through against impossible odds at times. I want to thank him for sharing the happiness of my achievements, for his understanding in hard times and his unrelenting love. A special thank you goes to my little son Lucas, for his patience during the last months of work and for bringing so much light to my life.

Lastly, thank you all and thank God!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Literature review

1.1 Introduction

This thesis is focused on the theory of robust estimation in Generalized Linear Models (GLMs), Generalized Linear Mixed Models (GLMMs) and Linear Models with endogeneity. Traditional estimation methods in GLMs and GLMMs rely on distributional assumptions about the random components, as well as the implicit assumption that the explanatory variables are uncorrelated with the error term. In most real world applications these distributional assumptions are only approximately valid, and some covariates may be endogenous. During the past decades researchers have become increasingly aware that some statistical procedures can be extremely sensitive to small deviations from the model assumptions, hence questioning their empirical usefulness. Although there is extensive work on robust inference in the context of linear regression models, its extension to GLMs, GLMMs and linear models with endogeneity remains limited. In the following paragraphs I briefly describe these models and address the main focus of this thesis.

Many regression problems involve response variables that have a distribution other than the Normal distribution. There are a variety of models commonly used in these cases. Logistic and Probit regressions are used to model binary response variables. Poisson regression is often used to model count data, and proportional hazard and accelerated failure time models are well known models for survival times. Nelder and Wedderburn (1972) demonstrate the unity of these and other methods using the idea of Generalized Linear Models (GLMs). Generalized Linear Models extend classical linear regressions in two main directions. First, GLMs accommodate non-normally distributed responses. Second, they allow a non-linear relation (the link function) between the mean of the response variable and a linear function of the predictors. These models are described in detail in Section 1.2.

A natural extension of GLMs is to add a linear combination of random components to the linear predictor. The resulting models are known as Generalized Linear Mixed Models (GLMMs). These new models are widely used when the responses are correlated, such as those coming from longitudinal data studies, or when the data show overdispersion, such as the Poisson-Gamma model. A description of them is given in Section 1.3.

Most of the methods of estimation considered for both GLMs and GLMMs rely on strong distributional assumptions (e.g., Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989; McGilchrist and Yau, 1995; Lee and Nelder, 1996). However, their sensitivity to small deviations from these assumptions has not been extensively studied. By the 1960's statisticians were concerned by the fact that the performance of some estimators was very unstable under small deviations from idealized distributional assumptions. Various measures of robustness were introduced for the location estimation problem, such as the Influence Function (Hampel, 1968, 1974) and the Change-of-Variance Function (Rousseeuw, 1981). Section 1.5 will briefly describe these concepts. Their extensions to estimators of multiple linear regression parameters have been developed (Hampel, 1973; Huber, 1973 ; Ronchetti and Rousseeuw, 1985). However, there are still many concepts that have not been explored for GLMs and GLMMs, for example the Change-of-Variance Function (CVF).

In Chapters 2 and 3 we investigate, using the CVF, the behavior of the asymptotic variance-covariance matrix of the class of M-estimators of GLMs and GLMMs respectively, when the distribution of the random components is slightly contaminated. Given that the notion of CVF had not been extended to these models before, we extend its definition and analyze its behavior for some existing M-estimators in the literature.

Another generally implicit assumption in GLMs is that the covariates are uncorrelated with the error term. In practice, however, this may not be true. That is, some explanatory variables may be endogenous. Neglecting this endogeneity may cause a severe bias in the parameter estimates. In particular, linear models with endogeneity problems have been widely studied. Section 1.4 includes a brief review of these models. However, most of the existing estimation methods for

these models are extremely sensitive to outliers and influential points. In Chapter 4, we propose robust instrumental variables estimators for linear models with endogenous covariates. These estimators are constructed using high breakdown point S-estimators of multivariate location and scatter matrix. Their robustness is investigated by means of the Influence Function (IF). Moreover, diagnostic techniques to identify outliers and influential points in the sample are developed.

The remaining of this Chapter consists of a brief survey of the literature on GLMs (Section 1.2), GLMMs (Section 1.3) and Linear Models with endogeneity (Section 1.4) together with a description of some elements of robustness theory that are going to be used throughout this thesis (Section 1.5).

1.2 Generalized Linear Models

Nelder and Wedderburn (1972) introduced Generalized Linear Models (GLMs) as a unified framework for models that had previously been studied in the literature such as linear regression, introduced over two hundred years ago by Gauss and Legendre (e.g., Stigler, 1981), logit (e.g., Berkson, 1944) and probit models (e.g., Bliss, 1934). Even though all pieces had already existed, these authors were the first to show the similarities between seemingly disparate methods. Using a methodology analogous to that developed for linear models, GLMs can be used to model response variables having a distribution belonging to an exponential family of distributions. Furthermore, the relationship between the response and the explanatory variables does not need to be linear. Section 1.2.1 presents the general setup of GLMs and Section 1.2.2 some of its methods of estimation.

1.2.1 The model

GLMs are built under the following set of assumptions:

- Consider an n-dimensional vector of independent random responses $\boldsymbol{Y}$. Conditionally, given that the vector of explanatory variables $\boldsymbol{X}_i = \boldsymbol{x}_i$, each response variable Y_i has a distribution

in the exponential family, taking the form

$$f_{Y_i|\boldsymbol{X}_i}(y_i|\boldsymbol{x}_i,\theta_i,\phi)=\exp\left\{\frac{y_i\theta_i-b(\theta_i)}{a(\phi)}+h(y_i,\phi)\right\},\quad i=1,\ldots,n, \tag{1.1}$$

for some specific functions $a(.)$, $b(.)$ and $h(.)$. If the dispersion parameter ϕ is known, this density belongs to an exponential family with canonical parameter θ_i.

- An arbitrary function of the conditional mean of the response is modelled as linear in the predictors, i.e.,

$$g(E[Y_i|\boldsymbol{x}_i])=\eta_i=\boldsymbol{x}_i\boldsymbol{\beta} \tag{1.2}$$

where $\boldsymbol{\beta}$ is the vector of unknown parameters that we want to estimate. The function $g(.)$ is called the link function and η_i is called the linear predictor. If $\eta_i=\theta_i$, $g(.)$ is called the canonical link.

- Assume that $\boldsymbol{X}_1,\boldsymbol{X}_2,\ldots,\boldsymbol{X}_n$ are iid random vectors with a marginal density given by $u(\boldsymbol{x})$ which does not depend on the unknown vector of parameters $\boldsymbol{\beta}$.

It can be easily shown that

$$E[Y_i|\boldsymbol{x}_i]=\mu_i=b'(\theta_i)\quad\text{and}\quad V[Y_i|\boldsymbol{x}_i]=b''(\theta_i)a(\phi) \tag{1.3}$$

where primes denote differentiation with respect to θ_i. Thus the variance of Y_i is the product of two functions: the variance function, $b''(.)$, depending only on θ_i (and hence on the mean, μ_i), and another function depending on the dispersion parameter ϕ.

Moreover, from (1.1), (1.2) and (1.3) one can derive the score statistic

$$S_j=\sum_{i=1}^{n}\left[\frac{\partial l(\theta_i;y_i)}{\partial\theta_i}\frac{d\theta_i}{d\mu_i}\frac{d\mu_i}{d\eta_i}\frac{\partial\eta_i}{\partial\beta_j}\right]=\sum_{i=1}^{n}\left[\frac{(y_i-\mu_i)}{a(\phi)}\frac{1}{V}\frac{d\mu_i}{d\eta_i}x_{ij}\right] \tag{1.4}$$

where $l(\theta_i;y_i)$ is the log-likelihood function of each component of $\boldsymbol{Y}$.

1.2.2 Estimation

Estimates of parameters for GLMs can be obtained using methods based on maximum likelihood (Nelder and Wedderburn, 1972). These estimates are generally computed not as global maximizers

of the log-likelihood function, but as the roots of the score statistics (1.4) which correspond to local maxima. For many important models, however, global and local maxima coincide. Explicit mathematical expressions for estimators can be found only in some special cases (such as the Normal or the exponential distribution), but in general, numerical methods such as Newton-Raphson or Fisher's scoring method will be needed. It can be shown that the latter is equivalent to an iterative weighted least squares algorithm. GLMs were incorporated in the GENSTAT statistics package and in the GLIM software. Now, most major statistical packages, such as SAS, S-Plus and R, have facilities for GLMs.

Conditions for uniqueness and existence of MLE have been studied for various models based on concavity of the log-likelihood (Haberman, 1974; Wedderburn, 1976; Silvapulle, 1981; Kaufmann, 1988). However, these conditions are difficult to check in practice. Moreover, under regularity conditions, asymptotic existence and uniqueness, consistency and asymptotic normality of the MLE have been proved (e.g., Haberman, 1977; Fahrmeir and Kaufmann, 1985).

An important extension of GLMs is the approach known as Quasi-likelihood models introduced by Wedderburn (1974). Noting that the score function defined in (1.4) depends on the parameters only through the mean, μ_i and the variance, $b(\theta_i)a(\phi)$, the full distributional assumption about the random component was replaced by a weaker assumption in which only the first and the second moments have to be specified. The estimators derived from the score equations in this manner are called maximum quasi-likelihood estimators (MQLE). When the distribution of Y_i belongs to an exponential family, the MQLE are the MLE. Under general conditions, Fahrmeir (1990) proved the consistency and asymptotic normality of the MQLE.

In the past decade, Bayesian methods have been developed for analyzing GLMs (Dey et al., 1999; Fahrmeir and Kaufmann, 1985). Bayesian models assume that $\boldsymbol{\beta}$ is a random vector with prior density $p(\boldsymbol{\beta})$. It is well known that an optimal estimator for $\boldsymbol{\beta}$ under quadratic loss is the posterior mean. However, its computation requires solving integrals having the dimension of $\boldsymbol{\beta}$, which in general is not feasible. Some methods based on numerical or Monte Carlo integration have been proposed (e.g., Naylor and Smith, 1982; Smith et al., 1985). In general, application of these

methods is limited to models having a low dimensional parameter vector $\boldsymbol{\beta}$. For higher dimensions, MCMC simulation techniques are commonly used (e.g., Dellaportas and Smith, 1993; Clayton, 1996). Based on samples drawn from the posterior density, posterior means and variances can be approximated using their sample analog. Moreover, posterior mode estimation is an alternative to full posterior mean estimation (e.g., Laird, 1978; Duffy and Santner, 1989). As the posterior mode estimator maximizes the posterior density, it is not required to solve any problem of integration.

Kedem and Fokianos (2001) extends the generalized linear models methodology to time series where the data and the covariates are time dependent. For more details on a GLM and its extensions see McCullagh and Nelder (1989), Fahrmeir and Tutz (1997) and Dobson (2002).

1.3 Generalized Linear Mixed Models

If the linear predictor of a Generalized Linear Model includes one or more random components in addition to the usual fixed effects, the resulting model is known as the Generalized Linear Mixed Model (GLMM). Examples include the Poisson-Gamma model used to account for the overdispersion often observed in count data or the Binomial-Beta model for binary data with correlated responses inherent in longitudinal or repeated measures designs (Lee and Nelder, 1996; Breslow and Clayton, 1993). In Section 1.3.1 we present the general setup of GLMMs and in Section 1.3.2 we describe some methods of estimations proposed for these models.

1.3.1 The model

Let $\boldsymbol{Y}$ be the vector of n observations and $\boldsymbol{U}$ a vector of random effects. Assume that

- Conditionally, given the vector of random effects, $\boldsymbol{U} = \boldsymbol{u}$ and $\boldsymbol{X} = \boldsymbol{x}$, the variables $Y_1, Y_2, ..., Y_n$ are independent and each one has a distribution belonging to the exponential family, talking the form

$$f_{Y_i|\boldsymbol{X}_i,\boldsymbol{u}}(y_i|\boldsymbol{X}_i, \boldsymbol{u}, \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + h(y_i, \phi)\right\}$$

  for some specific functions $a(.),\ b(.)$ and $h(.)$.

- Let $E[Y_i|\boldsymbol{X}_i, \boldsymbol{u}] = \mu_i = b'(\theta_i)$ and $g(\mu_i) = \eta_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{u}$, where $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of unknown parameters, $\boldsymbol{u} \in \mathbb{R}^q$ is a vector of random effects, $\boldsymbol{X}_i$ is a row vector of explanatory variables and $\boldsymbol{Z}_i$ is the model vector for the random effects.

- $\boldsymbol{U} \sim F_{\boldsymbol{u}}(\boldsymbol{u}|\boldsymbol{D})$, where $\boldsymbol{D}$ is the covariance matrix. We may assume later that $\boldsymbol{D} = \boldsymbol{D}(\boldsymbol{\tau})$, for some unknown vector $\boldsymbol{\tau}$. The mixing distribution $F_{\boldsymbol{u}}$ is often assumed to be normal (McGilchrist, 1994; Breslow and Clayton, 1993) but we are not going to make that assumption in this thesis.

- Let $\boldsymbol{\gamma} = (\boldsymbol{\beta}^T, \boldsymbol{\tau}^T)^T \in \mathbb{R}^{(p+q)}$ be the vector of unknown parameters that we want to estimate.

1.3.2 Estimation

A major difficulty in making inference about GLMMs is computational. Provided that the model is correctly specified and that the usual regularity conditions hold, the (marginal) maximum likelihood estimator of $\boldsymbol{\gamma}$ is consistent and asymptotically normal (White, 1982). However, this estimator requires the evaluation of high-dimensional integrals as the likelihood function of the model is given by:

$$L(\boldsymbol{\beta}, \phi, \boldsymbol{D}|\boldsymbol{Y}) = \int \prod f_{Y_i|\boldsymbol{u}}(y_i|\boldsymbol{u}, \boldsymbol{X}_i, \boldsymbol{\beta}, \phi) dF_{\boldsymbol{u}}(\boldsymbol{u}).$$

The integral has dimension equal to q, which makes the problem practically intractable, except for some particular cases (Anderson and Aitkin, 1985; Crouch and Siegelman, 1990).

To overcome this difficulty, alternative methods of estimation have been proposed. One such method was developed by McCulloch (1997) who used EM-like algorithms to obtain full maximum likelihood estimators. Three algorithms were proposed. First, he constructed a Monte Carlo version of the EM algorithm, called MCEM. Second, he proposed a new procedure, called Monte Carlo Newton-Raphson (MCNR). Finally, he adapted simulated maximum likelihood (SML) to this class of models. Some simulation studies were performed to investigate the convergence of these procedures.

Other methods of estimation that avoid integration of the random effects have been proposed (Schall, 1991; McGilchrist, 1994; Kuk, 1995; Lee and Nelder, 1996). Instead of using the marginal

density to construct the likelihood function, the idea is to maximize the joint density to obtain approximate maximum likelihood estimates (or REMLE) of the fixed effects and the dispersion components.

Breslow and Clayton (1993) used maximum quasi-likelihood estimation, introduced by Wedderburn (1974) in GLMs (see Section 1.2.2), for GLMs with random effects. They proposed two different methods of estimation for GLMMs: the penalized quasi-likelihood (PQL) method and the marginal quasi-likelihood (MQL). The essential difference between the PQL and the MQL estimating equations is that the former incorporates the random effect terms $\boldsymbol{Z}_i\boldsymbol{u}$ in the linear predictor while the other specifies the model in terms of the marginal mean $h(E[Y_i|\boldsymbol{X}_i]) = \boldsymbol{X}_i\boldsymbol{\beta}$.

An alternative method of estimation in GLMMs is based on the maximization of the joint distribution of the observed data and the random effects with respect to the parameters and the random effects (Lee and Nelder, 1996; McGilchrist et al., 1995). The idea is to extend the mixed-models equations of Henderson (1950) to models with more general distributional assumptions. This method is of particular interest when the estimation (or prediction) of the random effects is desired. In plant variety trials, for example, it is sometimes realistic to consider the variety as a random effect. The objective of variety trials is generally to find the best variety or to estimate the yield of each variety.

Other methods have recently been proposed by Jiang (1998, 1999, 2001) for estimating the fixed effect and variance components in a GLMM. Jiang (1998) used simulated moments in a method of moments approach to avoid the evaluation of high dimensional integrals. He called it the "method of simulated moments" (MSM). Under suitable conditions the MSM estimators are consistent as the total number of observations and the number of simulated random vectors go to infinity. However, simulation shows that these estimators can be inefficient. Jiang (2001) developed a new method that improves the efficiency of previous estimators as well as weakens some assumptions about the model. In addition, Jiang (1999) proposed a method of inference for GLMMs which relies on weak distributional assumptions about the random effects. He generalized the well-known method of weighted least squares (WLS) and proposed the penalized generalized

WLS (PGWLS) estimate of both the fixed and the random effects. In practice, one may not have sufficient information to estimate the random effects adequately. In those cases we may have to integrate out the random effects and estimate only the fixed effects and the variance components. The author remarked that it might be possible to estimate a subset of the random effects. This requires distributional assumptions only about the random effects that can not be estimated with adequacy, and only those are integrated out. He derived the likelihood function conditional on a subset of the random effects and maximized it to obtain the maximum conditional likelihood (MCL) estimates.

1.4 Endogenous covariates

A key condition for the ordinary least squares (OLS) estimator to be consistent is that the error term is uncorrelated with each of the regressors. However, there are many situations in which this assumption is not satisfied, i.e., the model contains "endogenous" covariates. In such a situation, the OLS estimator yields biased and inconsistent parameter estimates, even when not all the covariates are endogenous. Moreover, when some covariates are endogenous, other covariates with theoretical coefficient zero in the regression may appear as significant in an ordinary estimation. The "endogeneity" problem arises very often due to three main reasons: omitted variables, measurement error, and simultaneity.

Omitted variables appear when we would like to control for one or more additional variables but, usually because of data unavailability, we cannot include them in the regression model. In such a case the omitted covariate becomes part of the disturbance term. If such covariate is correlated with any of the covariates included in the regression, then the correlation between the regressors and the error term is different from zero.

Measurement errors appear when instead of observing the true explanatory variables $\boldsymbol{x}_i^*$, one observes $\boldsymbol{X}_i = \boldsymbol{x}_i^* + \boldsymbol{\nu}_i$. Then, the measurement error $\boldsymbol{\nu}_i$ becomes part of the disturbance term inducing a correlation between the observed covariates and the error term.

Simultaneity arises when at least one of the explanatory variables is determined simultane-

ously along with the response. If one covariate is determined partly as a function of the response, then that covariate and the error term are generally correlated.

An equation can have more than one source of endogeneity. For example, in looking at the effect of alcohol consumption on worker productivity measured by wages, we usually think that alcohol usage is correlated with unobserved factors such us family background, that also affect wage. In addition, alcohol demand would generally depend on income, which is largely determined by wage. Finally, measurement error in alcohol usage is always a possibility. For a discussion of the three kinds of endogeneity as they arise in a particular field, see Deaton's (1995) survey chapter on econometric issues in development economics. For a classical reference on measurement error models see Fuller (1987). Endogeneity problems are clearly explained in Amemiya (1985). A complete survey on endogeneity can be found in Anderson (1984).

A common approach to address this problem is to use additional information contained in instrumental variables, which are variables that do not belong to the original equation, are correlated with the existing explanatory variables but uncorrelated with the error term. These new variables can be used to construct ordinary instrumental variables (OIV) estimators that yield consistent parameter estimates. Although the use of instrumental variables dated to the late twenties, Sargan (1958) provided a general treatment to the IV method and established its asymptotic properties. For a review of Sargan's work see Arellano (2002).

1.4.1 Linear model with endogeneity

Consider the following model where some covariates are correlated with the error term, i.e., the model contains endogenous covariates.

$$\boldsymbol{Y} = \beta_0 + \boldsymbol{X}\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon} \tag{1.5}$$

$$\text{with} \quad E(\boldsymbol{\varepsilon}) = \mathbf{0} \quad \text{and} \quad Cov(\boldsymbol{X}, \boldsymbol{\varepsilon}) \neq \mathbf{0}$$

where $\boldsymbol{Y}$ is the n-dimensional column vector of observations of the response, β_0 is the regression intercept, $\boldsymbol{\beta}_1$ is a p-dimensional column vector of parameters, $\boldsymbol{X}$ is an $n \times p$ matrix of observable random covariates, and $\boldsymbol{\varepsilon}$ is the n-dimensional column vector of i.i.d. unobserved disturbances

with zero mean. We can also write this model as

$$Y_i = \beta_0 + \boldsymbol{X}_i\boldsymbol{\beta}_1 + \varepsilon_i$$

$$\text{with} \quad E(\varepsilon_i) = 0 \quad \text{and} \quad Cov(\boldsymbol{X}_i, \varepsilon_i) \neq 0$$

for $i = 1, \ldots, n$, where Y_i, and ε_i are the ith elements of $\boldsymbol{Y}$ and $\boldsymbol{\varepsilon}$, respectively, and $\boldsymbol{X}_i$ is the ith row of $\boldsymbol{X}$.

If the covariates are uncorrelated with the disturbances, then β_0 and $\boldsymbol{\beta}_1$ can be estimated consistently by ordinary least squares (OLS). However, when $Cov(\boldsymbol{X}, \boldsymbol{\varepsilon}) \neq \boldsymbol{0}$, OLS produces inconsistent estimates. For example, assuming normality in a classical error-in-variables model with only one explanatory variable and uncorrelated error terms, it can be shown using the properties of the bivariate normal distribution that

$$E[\hat{\beta}_1] = \beta_1(\sigma_{x^*x^*} + \sigma_{uu})^{-1}\sigma_{x^*x^*}. \tag{1.6}$$

where σ_{uu} and $\sigma_{x^*x^*}$ are the variances of the measurement error and the true covariates, respectively. Dropping the normality assumption, the RHS of (1.6) represents the probability limit of $\hat{\beta}_1$ as n tends to infinity. In both cases, the OLS is inconsistent and it is usually said that it has been attenuated by the measurement error in X.

1.4.2 Estimation

The method of ordinary instrumental variables provides a general solution to the problem of endogenous covariates. To use this approach, we need a q-dimensional row vector of instrumental variables $\boldsymbol{W}_i$, such that $q \geqslant p$. In this thesis, however, we focus on the case where the model is exactly identified, i.e., $p = q$. For the instruments to be valid, $\boldsymbol{W}_i$ needs to be correlated with the endogenous covariates, but uncorrelated with the disturbance term. More formally, $\boldsymbol{W}_i$ must satisfy the following two conditions

$$E(\boldsymbol{W}_i^T\varepsilon_i) = \boldsymbol{0} \tag{1.7}$$

$$\text{rank } E(\boldsymbol{W}_i^T\boldsymbol{X}_i) = p. \tag{1.8}$$

Note that the rank condition (1.8) means that $\boldsymbol{W}_i$ is sufficiently linearly related to $\boldsymbol{X}_i$ so that $E(\boldsymbol{W}_i^T \boldsymbol{X}_i)$ has full rank.

If all covariates are endogenous, then $\boldsymbol{W}_i$ is a list of p variables not contained in the original equation. When the model contains s exogenous and r endogenous variables (with $s + r = p$), each exogenous variable is already uncorrelated with the disturbance term and thus serves as an instrument for itself. In this case the vector $\boldsymbol{W}_i = (X_1, \ldots, X_s, I_1, \ldots, I_r)$, where $I_1, \ldots, I_r$ are also uncorrelated with the disturbance but are not included in the original equation.

The ordinary instrumental variables (OIV) estimator is defined as

$$\hat{\boldsymbol{\beta}}_{OIV} = (\hat{\beta}_0, \hat{\boldsymbol{\beta}}_1) = (\bar{Y} - \bar{\boldsymbol{X}}\hat{\boldsymbol{\beta}}_1, (\boldsymbol{W}^T \boldsymbol{X})^{-1} \boldsymbol{W}^T \boldsymbol{Y}) \tag{1.9}$$

where $\bar{Y} = n^{-1} \sum_{i=1}^{n} Y_i$, $\bar{\boldsymbol{X}} = n^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i$, and $\boldsymbol{W}$ is the $(n \times p)$ matrix of observations on the p instruments. This estimator is consistent provided that the data contains no extreme observations. However, it is well known that it has an unbounded Influence Function (Krasker and Welsch, 1985) and that a single aberrant observation can break it down (i.e., it has zero breakdown point). Thus, in Chapter 4 we present a robust version of this estimator.

1.5 Robustness

The problem of robustness was addressed by a number of eminent statisticians many years before a mathematical theory of robust estimation was developed. By the 1960's statisticians were concerned by the fact that the performance of some estimators was very unstable under small deviations from idealized distributional assumptions. This motivated the search of "robust" procedures which still behave fairly well under deviations from the assumed model. There have been several approaches to robust estimation of population parameters, including minimax asymptotic variance (Huber 1964), qualitative robustness (Hampel 1971), Influence Function (Hampel 1974), and Change-of-Variance Function (Hampel et al. 1981), among others.

In this Section we review the definitions of M-estimates, Influence Function and Change-of-Variance Function of one-dimensional estimators and its extensions to classical linear models.

1.5.1 M-estimates

Let $X_1, ..., X_n$ be a set of independent and identically distributed observations belonging to some sample space $\mathfrak{X}$. Consider the parametric model F_θ, where the unknown parameter θ belongs to some parameter space Θ.

Huber (1964) proposed to estimate the parameter θ using T_n, defined by solving

$$\sum \rho(X_i; T_n) = \min_{T_n}! \tag{1.10}$$

When $\rho(x;\theta) = \ln f(x;\theta)$, the estimator T_n is the maximum likelihood estimator. Thus, estimators satisfying equation (1.10) are called "M-estimator", which comes from "generalized maximum likelihood estimator". When ρ has derivative $\psi(x,\theta) = \partial\rho(x,\theta)/\partial\theta$, the estimate T_n satisfies the implicit equation

$$\sum \psi(X_i; T_n) = 0.$$

We will often identify ψ with the M-estimator it defines. If F_n is the empirical distribution function of $\boldsymbol{X}$, the M-estimator is also defined as $T_n = T(F_n)$, where T is the functional given by

$$\int \psi(x; T(F))F(dx) = 0. \tag{1.11}$$

1.5.2 Influence Function

The Influence Function (IF) was first introduced by Hampel (1974) in order to investigate the behavior of the asymptotic value of a one-dimensional estimator under small perturbations of the underlying distribution. More precisely, let $F_\varepsilon = (1-\varepsilon)F + \varepsilon\Delta_{\boldsymbol{x}}$ denote a neighborhood of the nominal distribution of the observations, F, contaminated by $\Delta_{\boldsymbol{x}}$, the point mass at $\boldsymbol{x}$. Then,

Definition 1.5.1. *The Influence Function of the estimator defined by the functional T at a distribution F is given by*

$$IF(\boldsymbol{x}; T, F) = \lim_{\varepsilon \downarrow 0} \frac{T(F_\varepsilon) - T(F)}{\varepsilon}.$$

for those $\boldsymbol{x}$ where the limit exists.

Remark 1.5.1. *Note that if the limit exists, then* $IF(\boldsymbol{x}; T, F) = (\partial/\partial\varepsilon)T(F_\varepsilon)|_{\varepsilon=0}$*, which is the directional (Gâteaux) derivative of* T *at* F*, in the direction of* $\Delta_{\boldsymbol{x}}$*. See Hampel et al. (1986) for further discussion.*

Heuristically, the IF describes the effect of an infinitesimal contamination at the point $\boldsymbol{x}$ on the estimate, standardized by the mass of the contamination.

Replacing F by F_ε in equation (1.11), differentiating with respect to ε and assuming that integration and differentiation may be interchanged, one can derive the IF of the M-estimator defined by ψ; that is,

$$IF(x; \psi, F) = \frac{\partial}{\partial\varepsilon} T(F_\varepsilon)|_{\varepsilon=0} = \frac{\psi(x; T(F))}{-\int(\partial/\partial\theta)[\psi(y;\theta)]_{T(F)} dF(y)}. \tag{1.12}$$

An important summary of the IF is the *gross-error sensitivity* of T at F, which can be thought as a measure of the worst influence that an infinitesimal contamination can have on the estimate.

Definition 1.5.2. *The gross-error sensitivity of* T *at* F *is given by*

$$\gamma^* = \sup_x |IF(x; T, F)|,$$

where the supremum is taken over all x *where the* $IF(x; T, F)$ *exists. Moreover, we say that* T *is* B*-robust at* F *if* γ^* *is finite.*

Therefore, an M-estimator defined by a function ψ in (1.11), is B-robust at F if and only if $\psi(., T(F))$ is bounded.

The IF in Linear Models

Huber (1973) extended his results on robust estimation of a location parameter to the case of linear regression. In the framework of M-estimation, he proposed using T_n defined by

$$T_n = \min_{\theta\in\Theta} \sum_{i=1}^{n} \rho((y_i - \boldsymbol{x}_i\theta)/\sigma), \tag{1.13}$$

for some function $\rho : \mathbb{R} \to \mathbb{R}^+$ and for a fixed σ. If ρ has derivative ψ, then T_n satisfies the system of equations

$$\sum_{i=1}^{n} \psi((y_i - \boldsymbol{x}_i T_n)/\sigma)\boldsymbol{x}_i = \mathbf{0}. \tag{1.14}$$

The functional $T(F)$ corresponding to the M-estimator defined by 1.14 is the solution of

$$\int \psi((y - \boldsymbol{x}T(F))/\sigma)\boldsymbol{x}dF(y, \boldsymbol{x}) = \mathbf{0}.$$

Using (1.12) it can be shown that $IF((\boldsymbol{x}, y); T, F)$ is unbounded in the $\boldsymbol{x}$-space. Thus, these estimators are sensitive to high leverage points. Other estimators addressing this problem have been proposed.

Definition 1.5.3. *(Maronna and Yohai, 1981) A GM-estimator T_n for linear models is defined implicitly by*

$$\sum_{i=1}^{n} \delta(\boldsymbol{x}_i, (y_i - \boldsymbol{x}_i T_n)/\sigma)\boldsymbol{x}_i = \mathbf{0}. \tag{1.15}$$

where the function $\delta : \mathbb{R}^p \times \mathbb{R} \to \mathbb{R}$ is continuous up to a finite set $C(\boldsymbol{x}; \delta)$, odd in the second argument and positive. Moreover, it is assumed that the set of points where it is continuous but $(\partial/\partial r)\delta(\boldsymbol{x}, r)$ is not defined or not continuous, denoted by $D(\boldsymbol{x}; \delta)$, is finite for all $\boldsymbol{x}$.

All known proposals for δ may be written in the form

$$\delta(\boldsymbol{x}, r) = w(x)\psi(r\nu(\boldsymbol{x})).$$

Note that Huber's proposal, defined in (1.13), uses $w(x) = \nu(\boldsymbol{x}) = 1$. Mallows' and Schweppe's proposals use $\nu(\boldsymbol{x}) = 1$ and $\nu(\boldsymbol{x}) = 1/w(x)$, respectively (see Hill, 1977 and Merrill and Schweppe, 1971).

Writing (1.15) as a functional equation, replacing the joint distribution H of the responses and the carriers by $H_\varepsilon = (1-\varepsilon)H + \varepsilon\Delta_{(\boldsymbol{x},y)}$, and following Definition 1.5.1, it is easy to show that the IF of T at H is a $p \times 1$ vector given by

$$IF((\boldsymbol{x}, y); T, H) = \delta(\boldsymbol{x}_i, (y_i - \boldsymbol{x}_i T(H)/\sigma)M^{-1}(\delta, H)\boldsymbol{x}$$

where $M(\delta, H) = \int(\partial/\partial r)\delta(\boldsymbol{x}_i, (y_i - \boldsymbol{x}_i T(H)/\sigma)\boldsymbol{x}\boldsymbol{x}^T dH(\boldsymbol{x}, y)$ (Hampel et al., 1986).

Two different measures of sensitivity were introduced to describe the worst possible influence of contamination by outliers on the asymptotic value of T.

Definition 1.5.4. *The unstandardized gross-error sensitivity of T is defined as*

$$\gamma_u^*(T, H) = \sup\{\|IF((\boldsymbol{x}, y); T, H)\|; \boldsymbol{x} \in \mathbb{R}^p, y \in \mathbb{R} \setminus C(\boldsymbol{x}, \delta)\},$$

and the (self-)standardized gross-error sensitivity is defined as

$$\gamma_s^*(T,H) = \sup\{[IF^T((\boldsymbol{x},y);T,H)V^{-1}(T,H)IF((\boldsymbol{x},y);T,H)];\boldsymbol{x}\in\mathbb{R}^p, y\in\mathbb{R}\setminus C(\boldsymbol{x},\delta)\},$$

where $V(T,H)$ is the asymptotic variance of T under model H. Moreover, an estimator T is B_u-(B_s-)robust when $\gamma_u^*(\gamma_s^*)$ is finite.

Krasker and Welsch (1982) noted that the *unstandardized gross-error sensitivity* is not invariant with respect to linear parameter transforms. Thus they introduced the *(self-)standardized gross-error sensitivity* to overcome this lack of invariance. For appropriate choices of functions $\delta(.)$, the GM-estimators are B_u-(B_s-)robust (Hampel et al., 1986).

Other estimators, not covered in this review, have been proposed for classical linear models such as MM-estimators, τ-estimators and S-estimators. For a survey on robust regression estimation see Maronna et al. (1993) and Rousseeuw and Leroy (1987).

1.5.3 Change-of-variance function

Other important asymptotic concepts that are also interesting to study include the asymptotic variance and the asymptotic efficiency. Rousseeuw (1981) first defined the Change-of-Variance Function (CVF) of an M-estimator of a location parameter to investigate the infinitesimal stability of its asymptotic variance in the presence of contamination of the nominal distribution, assumed to be symmetric. In this Section we briefly describe the CVF of an M-estimator of a one-dimensional location parameter, together with its extensions to linear regression models.

Let $F_\varepsilon = (1-\varepsilon)F + \varepsilon(\frac{1}{2}\Delta_x + \frac{1}{2}\Delta_{-x})$. Consider the M-estimator defined by

$$\sum_{i=1}^{n}\psi(x_i - T_n) = 0$$

which corresponds to the functional T defined by

$$\int \psi(x - T(F))dF(x) = 0.$$

Assume that ψ is a continuous, odd and positive function, with continuous derivative, ψ', up to a finite set $D(\psi)$. Let

$$0 < A(\psi) = \int \psi^2 dF < \infty \quad \text{and} \quad 0 < B(\psi) = \int \psi' dF < \infty$$

Definition 1.5.5. *The Change-of-Variance Function (CVF) of ψ at F is defined as*

$$CVF(x;\psi,F) = \frac{\partial}{\partial\varepsilon}[V(\psi,F_\varepsilon)]_{\varepsilon=0},$$

for all $x \in \mathbb{R} \setminus D(\psi)$ and where $V(\psi,F_\varepsilon)$ is the asymptotic variance of the M-estimator defined by ψ under model F_ε.

Definition 1.5.6. *The Change-of-Variance Sensitivity (CVS) of the M-estimator ψ is defined as*

$$k^*(\psi,F) = \sup\{CVF(x;\psi,F); x \in \mathbb{R} \setminus D(\psi)\}.$$

Moreover, an estimator is V-robust when its CVS, k^*, is finite.

It follows that the CVF is well-defined and continuous in $\mathbb{R} \setminus D(\psi)$, where it equals

$$CVF(x;\psi,F) = \frac{A(\psi)}{B(\psi)^2}\left[1 + \frac{\psi^2(x)}{A(\psi)} - 2\frac{\psi'(x)}{B(\psi)}\right].$$

These definitions were later extended to piecewise continuous ψ-functions (Rousseeuw, 1982) and to linear regression problems (Ronchetti and Rousseeuw, 1985). We summarize the latter below.

The CVF in Linear Models

Ronchetti and Rousseeuw (1985) extended the notion of Change-of-Variance Function to regression problems. Working in the framework of GM-estimators, presented in Definition 1.5.3, they defined

Definition 1.5.7. *The Change-of-Variance Function (CVF) of the GM-estimator corresponding to the functional T at H is defined as the $p \times p$ matrix*

$$CVF((\boldsymbol{x},y);T,H) = \frac{\partial}{\partial\varepsilon}[V(T,H_\varepsilon)]_{\varepsilon=0}$$

for all $(\boldsymbol{x},y)$ where it exists and where $V(T,H_\varepsilon)$ is the asymptotic variance of T under model H_ε.

In analogy to the gross-error-sensitivity, two different measures of sensitivity were introduced.

Definition 1.5.8. *The unstandardized Change-of-Variance Sensitivity of T is defined as*

$$k_u^*(T,H) = \sup\{tr\ CVF((\boldsymbol{x},y);T,H)/tr\ V(T,H); \boldsymbol{x} \in \mathbb{R}^p, y \in \mathbb{R} \setminus D(\boldsymbol{x},\delta)\}.$$

and the (self-)standardized Change-of-Variance Sensitivity is defined as

$$k_s^*(T,H) = \sup\{tr(CVF((\boldsymbol{x},y);T,H)V^{-1}(T,H)); \boldsymbol{x} \in \mathbb{R}^p, y \in \mathbb{R} \setminus D(\boldsymbol{x},\delta)\}. \tag{1.16}$$

Moreover, an estimator is V-robust when k^* is finite.

Maronna and Yohai (1981) showed that under certain conditions, the GM estimators are consistent and asymptotically normal with asymptotic variance covariance matrix

$$\begin{aligned} V(T,H) &= \int IF((\boldsymbol{x},y);T,H)IF^T((\boldsymbol{x},y);T,H)dH(\boldsymbol{x},y) \\ &= M^{-1}(\delta,H)Q(\delta,H)M^{-1}(\delta,H), \end{aligned}$$

where $Q(\delta,H) = \int \delta^2(\boldsymbol{x}_i,(y_i - \boldsymbol{x}_i T(H)/\sigma)\boldsymbol{x}\boldsymbol{x}^T dH(\boldsymbol{x},y)$. Replacing H with H_ε in (1.17) and following Definition 1.5.7, the CVF in regression is given by

$$\begin{aligned} CVF((\boldsymbol{x},y);T,H) &= M^{-1}QM^{-1} + \delta^2(\boldsymbol{x},y)M^{-1}\boldsymbol{x}\boldsymbol{x}^T M^{-1} \\ &\quad -\delta'(\boldsymbol{x},y)\left[M^{-1}\boldsymbol{x}\boldsymbol{x}^T M^{-1}QM^{-1} + M^{-1}QM^{-1}\boldsymbol{x}\boldsymbol{x}^T M^{-1}\right], \end{aligned}$$

where $M = M(\delta,H)$ and $Q = Q(\delta,H)$ (Ronchetti and Rousseeuw, 1985).

The CVF and the IF have many characteristics in common (Hampel et al., 1986). However, these curves are not interpreted in the same way. Both large positive and negative values of the IF are unfavorable, meaning a large asymptotic bias caused by the contamination. Unlike the case of the IF, one does not have to worry about large negative values of the CVF as much as about large positive values.

1.5.4 Extensions to more general models

The robustness of parameter estimates has been considered by several authors for more general models. Cantoni and Ronchetti (2001) derived robust estimates for GLMs based on estimating equations that are natural generalizations of quasi-likelihood functions. Fellner (1986) suggested robust estimation methods in linear mixed models. Yau and Kuk (2002) borrowed Fellner's ideas to obtain robust estimators in GLMMs. Neuhaus et al. (1992) examined mixing distribution misspecification in logistic mixed models. Gustafson (1996) investigated the magnitude of the

asymptotic bias using an approximation based on infinitesimal contamination of the mixing distribution for some conjugate GLMMs. Smith and Weems (2004) showed that the MLEs of the Poisson-lognormal models, an important class of GLMMs, are B-robust under perturbations of the mixing distribution. However, none of them examines the Change-of-Variance Function in these more general scenarios.

In Chapters 2 and 3 we extend the notion of CVF to GLMs and GLMMs respectively and use it to study the infinitesimal behavior of the asymptotic variance of some well known estimators under contamination. Chapter 4 introduces a Robust Instrumental Variables Estimator for linear models with endogeneity and analyzes its robustness properties by means of its IF. The IF is also used to estimate the variance of the coefficient estimates and to develop some diagnostic techniques.

1.5.5 Robust Multivariate Location and Scatter Matrix Estimation

In Chapter 4 we propose an estimator for Measurement Error Models based on robust multivariate location and scatter matrix estimators. It is well known that the usual sample mean ($\bar{\boldsymbol{X}}$), and sample variance covariance matrix ($\boldsymbol{S}^2$), are extremely sensitive to outliers. In this section we review some robust location and scatter matrix estimators that have been proposed in the literature.

Multivariate M-estimators

Maronna (1976) extended the univariate definition on M-estimators to the multivariate scenario. An M-estimate $\hat{\boldsymbol{\theta}} = (\boldsymbol{t}, \boldsymbol{C})$ of multivariate location and covariance $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is defined as the solution to the equations

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} v_1\left(d(\boldsymbol{x}_i, \boldsymbol{t}; \boldsymbol{C})\right)(\boldsymbol{x}_i - \boldsymbol{t}) &= \boldsymbol{0} \\ \frac{1}{n}\sum_{i=1}^{n} v_2\left(d(\boldsymbol{x}_i, \boldsymbol{t}; \boldsymbol{C})^2\right)(\boldsymbol{x}_i - \boldsymbol{t})(\boldsymbol{x}_i - \boldsymbol{t})^T &= \boldsymbol{C}. \end{aligned}$$

where v_1 and v_2 are weighting functions which control the influence of outliers on the location and covariance estimates, and, for ease of notation, we define $d(\boldsymbol{x}, \boldsymbol{t}; \boldsymbol{C}) = [(\boldsymbol{x} - \boldsymbol{t})^T \boldsymbol{C}^{-1} (\boldsymbol{x} - \boldsymbol{t})]^{1/2}$.

Note that if we let $v_1(s) = v_2(s^2) = 1$, we obtain the sample mean and covariance as our estimates. Bounded choices of v_1 and v_2 lead to robust multivariate estimates.

The M-estimate is then the solution of a system of nonlinear equations of the weighted sample moments. Maronna (1976) shows several important properties of the M-estimate under certain conditions on the weighting functions and distribution, including existence, uniqueness and convergence. A serious drawback of the M-estimators is that, as in the case of regression problems, they have a low breakdown point which decreases with increasing dimensionality of the data. This led to a search for multivariate affine equivariant estimates which possess a high breakdown independently of the dimension of the data. The following estimators are some alternative multivariate estimators with these properties.

Minimum Volume Ellipsoid (MVE)

Rousseeuw (1985) introduced an affine equivariant estimator with maximal breakdown point known as the *Minimum Volume Ellipsoid* estimator. The idea is to find an ellipsoid containing h points which is of minimum volume. The location estimator is then given by the center of the ellipsoid and the covariance estimator is defined as the shape matrix of the ellipsoid (multiplied by a suitable factor to obtain consistency). Davies (1987) and Lopuhaä and Rousseeuw (1991) proved that $h = [(n - p + 1)/2]$ leads to maximal robustness (maximal breakdown point). However, it was shown that the MVE estimator is not $\sqrt{n}$ consistent (Davies, 1992). Its low convergence rate reduces the relative efficiency of the estimator.

In most applications, it is not feasible to consider all sets of h points of the data and compute the volume of the smallest ellisoid containing them. Instead, one can obtain an approximate solution using a resampling algorithm (Rousseeuw and Leroy, 1987).

Minimum Covariance Determinant (MCD)

Another multivariate location and covariance estimator which is affine equivariant and has maximum breakdown point is the *Minimum Covariance Determinant* estimator (Rousseeuw 1983, 1984). This estimator corresponds to the sample mean and covariance from the set of h points whose sample covariance has the minimum determinant or what is the same, for which the classical confidence ellipsoid has minimum volume. Moreover, this robust estimator has the normal rate of

convergence, $\sqrt{n}$, making it more attractive than the MVE estimator.

The computational complexity is a major issue regarding the MCD. To find an exact solution requires searching the entire space of all possible subsets of h out of n data points. Thus, the computational burden grows combinatorially with the sample size. However, there are fast approximations to the MCD, the most prevalent of which is proposed by Rousseeuw and Van Driessen (2002).

Multivariate S-estimators

Davies (1987) and Lopuhaä (1989) extended regression S-estimates for multivariate location and covariance estimators. Lopuhaä defines the S-estimate $\hat{\boldsymbol{\theta}} = (\boldsymbol{t}, \boldsymbol{C})$ as the solution to the optimization problem

$$\min_{(\boldsymbol{t},\boldsymbol{C})\in(\mathbb{R}^d,\ \mathrm{PDS}(d))} \{|\boldsymbol{C}|\}$$

such that

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(d(\boldsymbol{x}_i, \boldsymbol{t}; \boldsymbol{C})\right) = b_0$$

where $PDS(d)$ is the set of all positive definite symmetric matrices of order d and both b_0 and ρ have to be chosen.

Note that as in the regression case, choosing $\rho(s) = s^2$ yields to the least squares solution for the location-covariance problem. Choosing $b_0 = d$ for appropriate scaling of the covariance matrix, the S-estimate reduces to the sample mean and covariance matrix as the unique solution (Grübel, 1988). The MVE and the MCD described in previous sections are also particular cases of S-estimators. For example, the MVE estimator can be obtained by using the discontinuous ρ-function

$$\rho(u) = \begin{cases} 0 & \text{if } \mid u \mid < (\chi^2_{d,0.5})^{1/2} \\ 1 & \text{otherwise} \end{cases}$$

and setting $b_0 = 1/2$.

Lopuhaä and Davies proved many of the same results such as existence, convergence, Fisher consistency, and asymptotic normality but under different conditions on the ρ function and the

underling distribution. Moreover, Davies demonstrated an even more significant attribute of S-estimates, which is the ability to achieve the maximal breakdown point (asymptotically 1/2) regardless of dimension for an appropriate choice of ρ.

We create an algorithm to compute this estimator in S-Plus/R which is available under request.

Coordinatewise and pairwise estimators

Much faster estimates can be computed if one drops the requirements of positive definiteness and affine equivariance. A straightforward approach for multivariate location is to simply calculate a robust location estimate for each individual variable in the dataset. In the case of the multivariate covariance matrix, one can similarly apply a robust covariance estimate to each pair of variables. Estimates of this type are called coordinatewise and pairwise, respectively. The pair and coordinatewise approach is appealing in that the resulting estimators inherit the robustness (breakdown point) of those estimators applied to each variable or pair of variables respectively and it reduces the computational complexity. However, the estimators obtained are not affine equivariant and the scatter matrix is not guaranteed to be positive definite.

Recently, Alqallaf et al. (2002) and Maronna and Zamar (2002) proposed new pairwise methods that preserve positive definiteness and are computationally inexpensive. With sequential algorithms these methods can be applied to problems with up to a few hundreds variables. Moreover, parallel algorithms can be used to scale these estimators to problems with thousands of variables (see Chilson et al., 2003).

Chapter 2

Change-of-variance function in GLMs

2.1 Introduction

In this Chapter we derive the Change-of-Variance function (CVF) and the Change-of-Variance sensitivity (CVS) of M-estimators of GLM parameters in order to examine the sensitivity of the asymptotic variance-covariance matrix of the estimates under a slight contamination of the distribution of the random components. Although there exists some previous work on robust inference in GLMs, the CVF was studied only for classical linear models. The following paragraphs contain a brief survey of the estimation methods commonly used for GLMs and some robust procedures suggested in the literature.

Generalized Linear Models have been introduced by Nelder and Wedderburn (1972) as a unifying family of models with not necessarily normal responses, which allow a nonlinear link between the mean of the response variable and the predictors. This family includes a variety of commonly used models such as Poisson regressions to model count data, proportional hazard models and accelerated failure time models to model survival times, and logistic and probit regressions to model binary response variables. A detailed description of GLMs was presented in Section 1.2.

In general, inference about these models is based on maximum likelihood procedures assuming that the model is completely and correctly specified (Nelder and Wedderburn, 1972; Fahrmeir and Kaufmann, 1985). However, slight violations of these assumptions can have a potentially large influence on the estimator. Pregibon (1982) studied the sensitivity of the MLE to outlying and influential points in logistic regression models. He proposed a resistant fitting method of estimation consisting of minimizing a modified deviance function which limits the effect of observations poorly explained by the model. Other work on robustness in GLMs includes Stefanski et al. (1986) and Künsch et al. (1989). Following Hampel's problem (1968, 1974) in the single parameter case,

they proposed optimal bounded-influence estimators. Imposing a bound on the Influence Function, they found an estimator which minimizes the asymptotic variance matrix in the strong sense of positive-definiteness. Furthermore, Cantoni and Ronchetti (2001) considered a class of Mallows-type robust estimators, where the influence on the estimators of deviations in the y-space and in the explanatory variables are bounded separately. They discuss robust estimators of a generalized linear model based on quasi-likelihood methods (see 1.2.2 for a definition of quasi-likelihood estimators.)

Most of the research on the nonrobustness of estimators in GLMs is focused on the sensitivity of the estimator to outlying and influential data points. However, a perturbation of the model assumptions may also drastically affect the asymptotic variance of the estimator, leading to decreased precision and wider confidence intervals. Ronchetti and Rousseeuw (1985) introduced the notion of Change-of-Variance Function to investigate the influence of contamination at a single data point $(\boldsymbol{x}, y)$ on the asymptotic variance of the regression parameters of a linear model. However, extensions to GLMs have not been studied.

This Chapter is organized as follows. In Section 2.2 we extend the notion of CVF and CVS of M-estimators to GLMs. In particular, the MLE is analyzed in detail in Section 2.2.1. In Section 2.3 we study the CVF of a subclass of bounded influenced M-estimators commonly used in the robustness literature. Finally, a simulation is performed for a Logistic model and the results are summarized in Section 2.4. We end this Chapter with some conclusions.

2.2 CVF of the M-estimators

In this Section we derive the CVF for GLMs to study the effect of an ε-contamination of the nominal distribution on the asymptotic variance of the M-estimators. Although Definitions 1.5.5 and 1.5.6 were made in the framework of M-estimation of a one-dimensional parameter, they can be extended to the case of multivariate parameters. In particular, Ronchetti et al. (1985) defined the CVF and the CVS for estimators of the regression coefficients of classical linear models. In this Section we extend these definitions for M-estimators of GLM parameters under a contamination

in the nominal distribution. We first introduce some notation that is going to be used throughout this Chapter.

Consider the model introduced in Section 1.2.1. Let H_0 the true distribution of the independent pairs $(\boldsymbol{X}, Y)$. Suppose that the nominal distribution H_0 is slightly contaminated by a distribution G, so that the random pairs $(\boldsymbol{X}, Y)$ are actually generated from H_ε which is an ε-contamination of the distribution H_0. That is,

$$H_\varepsilon = (1-\varepsilon)H_0 + \varepsilon G. \tag{2.1}$$

Using $G(\boldsymbol{u}, v) = \Delta_{(\boldsymbol{x},y)}(\boldsymbol{u}, v)$, the probability measure which puts all its mass at ($\boldsymbol{x}$,y), the distribution given in (2.1) may describe a mixture which contains a fraction of ε of outliers at ($\boldsymbol{u}$,v).

Notation 2.2.1. *Let E_ε be the expected value with respect to H_ε. Similarly define E_{H_0}.*

Consider the class of M-estimators, $\hat{\boldsymbol{\beta}}_\psi$, defined implicitly as the solution of

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\psi}(\boldsymbol{x}_i, y_i; \boldsymbol{\beta}) = \mathbf{0}. \tag{2.2}$$

for suitably chosen functions $\boldsymbol{\psi}$ from $\mathbb{R}^p \times \mathbb{R} \times \mathbb{R}^p$ to $\mathbb{R}^p$ such that

$$E_H[\boldsymbol{\psi}(\boldsymbol{X}, Y; \boldsymbol{\beta})] = \mathbf{0}.$$

Under regularity conditions (Huber, 1967; Stefanski et al., 1986), $\hat{\boldsymbol{\beta}}_\psi$ is consistent and asymptotically normal with asymptotic variance given by

$$V(\boldsymbol{\psi}, H_0) = M_0^{-1} Q_0 \left\{M_0^{-1}\right\}^T, \tag{2.3}$$

where

$$M_0 = -E_{H_0}\left[\nabla_{\boldsymbol{\beta}} \boldsymbol{\psi}(\boldsymbol{X}, Y; \boldsymbol{\beta})\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right], \tag{2.4}$$

$$Q_0 = E_{H_0}\left[(\boldsymbol{\psi}(\boldsymbol{X}, Y; \boldsymbol{\beta}))(\boldsymbol{\psi}(\boldsymbol{X}, Y; \boldsymbol{\beta}))^T\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right]. \tag{2.5}$$

Finally let $\boldsymbol{\beta}_\varepsilon$ the solution of the equation

$$E_\varepsilon[\boldsymbol{\psi}(\boldsymbol{X}, Y; \boldsymbol{\beta})] = 0, \tag{2.6}$$

and let M_ε and Q_ε be the matrices defined in (2.4) and (2.5) replacing H_0 and $\boldsymbol{\beta}_0$ by H_ε and $\boldsymbol{\beta}_\varepsilon$, respectively.

We extend Definition 1.5.5 for GLMs by

$$CVF(\boldsymbol{x}, y; \boldsymbol{\psi}, H) = \left.\frac{dV(\boldsymbol{\psi}, H_\varepsilon)}{d\varepsilon}\right|_{\varepsilon=0}$$

where $V(\boldsymbol{\psi}, H_\varepsilon) = M_\varepsilon^{-1} Q_\varepsilon \left\{M_\varepsilon^{-1}\right\}^T$. Assuming that interchange of expectation and differentiation is allowed, we obtain

$$CVF(\boldsymbol{x}, y; \boldsymbol{\psi}, H) = \left.\frac{dM_\varepsilon^{-1}}{d\varepsilon}\right|_{\varepsilon=0} Q_0 \left\{M_0^{-1}\right\}^T + M_0^{-1} \left.\frac{dQ_\varepsilon}{d\varepsilon}\right|_{\varepsilon=0} \left\{M_0^{-1}\right\}^T + M_0^{-1} Q_0 \left.\frac{d\left\{M_\varepsilon^{-1}\right\}^T}{d\varepsilon}\right|_{\varepsilon=0}, \tag{2.7}$$

Following natural calculations we derive the CVF:

$$\left(\left.\frac{dM_\varepsilon}{d\varepsilon}\right|_{\varepsilon=0}\right)_{ij} = \sum_{s=1}^{p} (\boldsymbol{K}_{ij})_s \psi_s(\boldsymbol{x}, y; \boldsymbol{\beta}_0) - \psi'_{ij}(\boldsymbol{x}, y; \boldsymbol{\beta}_0) - (M_0)_{ij}, \tag{2.8}$$

where $(\boldsymbol{K}_{ij})_s$ is the sth element of the p-dimensional vector given by

$$\boldsymbol{K}_{ij} = -E_H\left[\nabla_{\boldsymbol{\beta}^T} \psi'_{ij}(\boldsymbol{X}, Y; \boldsymbol{\beta})\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right] M_0^{-1},$$

and ψ'_{ij} is the $(ij)th$ element of the $(p \times p)$-matrix $\boldsymbol{\psi}'$ given by

$$\boldsymbol{\psi}'(\boldsymbol{X}, Y; \boldsymbol{\beta}) = \nabla_{\boldsymbol{\beta}} \boldsymbol{\psi}(\boldsymbol{X}, Y; \boldsymbol{\beta}).$$

We can express (2.8) in matrix notation as

$$\left.\frac{dM_\varepsilon}{d\varepsilon}\right|_{\varepsilon=0} = \boldsymbol{W}(\boldsymbol{x}, y; \boldsymbol{\beta}_0) - \boldsymbol{\psi}'(\boldsymbol{x}, y; \boldsymbol{\beta}_0) - M, \tag{2.9}$$

where $(\boldsymbol{W}(\boldsymbol{x}, y; \boldsymbol{\beta}_0))_{ij} = \sum_{s=1}^{p} (\boldsymbol{K}_{ij})_s \psi_s(\boldsymbol{x}, y; \boldsymbol{\beta}_0)$.

Similarly,

$$\begin{aligned}
\left.\frac{dQ_\varepsilon}{d\varepsilon}\right|_{\varepsilon=0} \quad = \quad & E_{H_0}[\boldsymbol{\psi}'(\boldsymbol{X}, Y; \boldsymbol{\beta}_0) M_0^{-1} \boldsymbol{\psi}(\boldsymbol{x}, y; \boldsymbol{\beta}_0) \boldsymbol{\psi}^T(\boldsymbol{X}, Y; \boldsymbol{\beta}_0)] \\
& + E_{H_0}[\boldsymbol{\psi}(\boldsymbol{X}, Y; \boldsymbol{\beta}_0) \boldsymbol{\psi}^T(\boldsymbol{x}, y; \boldsymbol{\beta}_0) \left\{M_0^{-1}\right\}^T \left\{\boldsymbol{\psi}'(\boldsymbol{X}, Y; \boldsymbol{\beta}_0)\right\}^T] \\
& + \boldsymbol{\psi}(\boldsymbol{x}, y; \boldsymbol{\beta}_0) \boldsymbol{\psi}^T(\boldsymbol{x}, y; \boldsymbol{\beta}_0) - Q_0.
\end{aligned} \tag{2.10}$$

Thus, substituting (2.9) and (2.10) into (2.7) we obtain

$$\begin{aligned} CVF(\boldsymbol{x}, y; \boldsymbol{\psi}, H_0) &= \left[-M_0^{-1}\boldsymbol{W}(\boldsymbol{x}, y; \boldsymbol{\beta}_0)M_0^{-1} + M_0^{-1}\boldsymbol{\psi}'(\boldsymbol{x}, y; \boldsymbol{\beta}_0)M_0^{-1}\right] Q_0 \left\{M_0^{-1}\right\}^T \\ &+M_0^{-1}Q_0 \left[-M_0^{-1}\boldsymbol{W}(\boldsymbol{x}, y; \boldsymbol{\beta}_0)M_0^{-1} + M_0^{-1}\boldsymbol{\psi}'(\boldsymbol{x}, y; \boldsymbol{\beta}_0)M_0^{-1}\right]^T \\ &+M_0^{-1}\left[\boldsymbol{L}(\boldsymbol{x}, y; \boldsymbol{\beta}_0) + \boldsymbol{L}^T(\boldsymbol{x}, y; \boldsymbol{\beta}_0)\right]\left\{M_0^{-1}\right\}^T \\ &+M_0^{-1}\left[\boldsymbol{\psi}(\boldsymbol{x}, y; \boldsymbol{\beta}_0)\boldsymbol{\psi}^T(\boldsymbol{x}, y; \boldsymbol{\beta}_0)\right]\left\{M_0^{-1}\right\}^T \\ &+M_0^{-1}Q_0\left\{M_0^{-1}\right\}^T, \end{aligned} \tag{2.11}$$

where $\boldsymbol{L}(\boldsymbol{x}, y; \boldsymbol{\beta}_0)$ is the first expectation in the RHS of (2.10).

Finally, the following definition generalizes the Change-of-Variance Sensitivity (CVS) of the one-dimensional location case defined in (1.16).

Definition 2.2.1. *The self-standardized Change-of-Variance Sensitivity of an M-estimators defined by a function $\boldsymbol{\psi}$ is given by*

$$k_s^*(\boldsymbol{\psi}, H) = \sup_{(\boldsymbol{x}, y)} \{tr[CVF((\boldsymbol{x}, y); \boldsymbol{\psi}, H)V^{-1}]\}. \tag{2.12}$$

where V is the asymptotic variance defined in (2.3). We say that an estimator is V-robust when k^* is finite.

Substituting (2.11) and (2.3) into (2.12) we derive the CVS of the M-estimators defined in (2.2):

$$\begin{aligned} k^*(\boldsymbol{\psi}, H) = p + \sup_{(\boldsymbol{x}, y)} \{&2tr[\boldsymbol{\psi}'(\boldsymbol{x}, y)M^{-1}] + 2\sum_{i=1}^{p}(c_i - a_i)\psi_i(\boldsymbol{x}, y) \\ &+\boldsymbol{\psi}^T(\boldsymbol{x}, y)Q^{-1}\boldsymbol{\psi}(\boldsymbol{x}, y)\}, \end{aligned} \tag{2.13}$$

where c_i is the i^{th} component of the p-dimensional row vector $\boldsymbol{c} = E_H\left[\boldsymbol{\psi}^T(\boldsymbol{X}, Y)Q^{-1}\boldsymbol{\psi}'(\boldsymbol{X}, Y)M^{-1}\right]$ and $a_i = \sum_{r=1}^{p}\sum_{k=1}^{p}(M^{-1})_{kr}(K_{rk})_i$.

Ronchetti and Rousseeuw (1985) proved that V-robustness implies B-robustness (bounded IF) for classical linear models. Unfortunately, this does not hold in general for GLMs. Note that the last term in the RHS of the supremum in (2.13) is the argument of the gross error sensitivity

of an M-estimator in GLMs (Stefanski et al., 1986; Künsch et al., 1989). However, the signs of the other two terms are not known to establish any further implication.

2.2.1 Examples

In this Section we analyzed the CVS of the MLE for GLMs and the M-estimators for linear models in detail.

1) Maximum Likelihood Estimator

As the MLE is a particular M-estimator when

$$\boldsymbol{\psi}(\boldsymbol{x}, y; \boldsymbol{\beta}) = \nabla_{\boldsymbol{\beta}} \log(f_H(\boldsymbol{x}, y; \boldsymbol{\beta})), \tag{2.14}$$

following Definition (2.13) one can derive its CVS and study the V-robustness of the MLE in GLMs. For simplicity, in this Section we omit the subindex for each observation.

As the distribution of $\boldsymbol{X}$ does not depend on the unknown parameter $\boldsymbol{\beta}$, using (1.1) the function $\boldsymbol{\psi}$ defined in (2.14) reduces to

$$\boldsymbol{\psi}(\boldsymbol{x}, y; \boldsymbol{\beta}) = [y - h(\eta)] \frac{h'(\eta)}{V(\mu)} \boldsymbol{x}^T, \tag{2.15}$$

where $V(\mu) = V[Y|\boldsymbol{x}] = b''(\theta)$ and $h(\cdot)$ is the inverse of the link function $g(\cdot)$ defined in (1.2). For simplicity in the notation, let

$$s(\eta) = \frac{h'(\eta)}{V(\mu)} \tag{2.16}$$

Thus, differentiating (2.14) with respect to $\boldsymbol{\beta}$, we obtain

$$\boldsymbol{\psi}'(\boldsymbol{x}, y; \boldsymbol{\beta}) = \big[- h'(\eta) s(\eta) + (y - h(\eta)) s'(\eta) \big] \boldsymbol{x}^T \boldsymbol{x}.$$

Then,

$$\begin{aligned} k^*(\boldsymbol{\psi}, H) \quad = \quad & p + \sup_{(\boldsymbol{x}, y)} \{ 2[-h'(\eta) s(\eta) + (y - h(\eta)) s'(\eta) \big] (\boldsymbol{x} M^{-1} \boldsymbol{x}^T) \\ & + (y - h(\eta)) s(\eta) (\boldsymbol{c} - \boldsymbol{a}) \boldsymbol{x}^T + [(y - h(\eta)) s(\eta)]^2 (\boldsymbol{x} Q^{-1} \boldsymbol{x}^T) \}. \end{aligned}$$

Note that when the canonical link function $g(\cdot)$ is used, the function $s(\cdot)$ in (2.16) reduces to unity

and the CVS is given by

$$\begin{aligned} k^*(\boldsymbol{\psi}, H) &= p + \sup_{(\boldsymbol{x},y)} \{-2h'(\eta)(\boldsymbol{x}M^{-1}\boldsymbol{x}^T) \\ &\quad -(y-h(\eta))\boldsymbol{a}\boldsymbol{x}^T + [y-h(\eta)]^2(\boldsymbol{x}Q^{-1}\boldsymbol{x}^T)\}. \end{aligned} \tag{2.17}$$

As the first term is linear in $[y - h(\eta)]\boldsymbol{x}$, we need to analyze only the last two terms of (2.17). Thus, we need to study the following quantity:

$$\sup_{(\boldsymbol{x},y)} \{-2h'(\eta)(\boldsymbol{x}M^{-1}\boldsymbol{x}^T) + [y-h(\eta)]^2(\boldsymbol{x}Q^{-1}\boldsymbol{x}^T)\}. \tag{2.18}$$

We can rewrite (2.18) as

$$\sup_{(\boldsymbol{x},y)} \{\boldsymbol{x}\boldsymbol{A}(\boldsymbol{x},y)\boldsymbol{x}^T\}, \tag{2.19}$$

where $\boldsymbol{A}(\boldsymbol{x},y) = -2h'(\eta)M^{-1} + [y-h(\eta)]^2Q^{-1}$. In this case, the $(p \times p)$ matrix M defined in (2.4) is the Fisher information matrix. Then, both Q and M are positive definite.

Note that if the support of Y is not bounded, the CVS defined in (2.17) is infinity and the MLE is not V-robust. When the support of Y is bounded, the value of the CVS depends on $h(\cdot)$, the inverse of the link function. Thus, in general, we can not say whether the MLE is V-robust or not in this case. We now study the CVS of the MLE in three commonly used GLMs using the canonical link function.

a) Logit models: the canonical link function and its derivative are given by:

$$h(\eta) = \frac{e^\eta}{1+e^\eta}, \quad \text{and} \quad h'(\eta) = \frac{e^\eta}{(1+e^\eta)^2}$$

It is easy to see that the limits of these functions are equal to 1 and 0, respectively, as η tends to $+\infty$. Let $y = 0$ be fixed. Then

$$\lim_{\eta \to +\infty} \boldsymbol{A}(\boldsymbol{x},y) = Q^{-1}.$$

Thus,

$$\lim_{\eta \to +\infty} \boldsymbol{x}^T\boldsymbol{A}(\boldsymbol{x},y)\boldsymbol{x} = +\infty$$

Then both the supremum in (2.19) and the CVS, $k^*(\boldsymbol{\psi}, H)$, equal $+\infty$. Hence, the MLE is not V-robust. However, it is important to note that in an experimental study, where the support of

$\boldsymbol{X}$ is bounded, the MLE is V-robust.

b) Poisson models: the canonical link function and its derivative are given by:

$$h(\eta) = h'(\eta) = e^{\eta}$$

Then, for any value of y fixed,

$$\lim_{\eta \to +\infty} \boldsymbol{A}(\boldsymbol{x}, y) = +\infty,$$

Thus,

$$\lim_{\eta \to +\infty} \boldsymbol{x}^T \boldsymbol{A}(\boldsymbol{x}, y)\boldsymbol{x} = +\infty,$$

so $k^*(\boldsymbol{\psi}, H)$ equals $+\infty$.

c) Gamma models: the canonical link function and its derivative are given by:

$$h(\eta) = \frac{1}{\eta}, \quad \text{and} \quad h'(\eta) = -\frac{1}{\eta^2}.$$

In this case, the argument of (2.18) is bounded in $\boldsymbol{x}$. However, for any fixed value of $\boldsymbol{x}$,

$$\lim_{y \to +\infty} \boldsymbol{A}(\boldsymbol{x}, y) = +\infty, \tag{2.20}$$

Using (2.20), we prove that the supremum defined in (2.19) equals $+\infty$. Thus $k^*(\boldsymbol{\psi}, H)$ equals $+\infty$.

Thus, despite of the widespread use of the MLE for GLMs, in general, this estimator is not V-robust.

2) M-estimators for linear models

As the normal distribution belongs to the exponential family defined in (1.1), linear regression models are a particular case of GLMs. In this example, we show that $k^*(\boldsymbol{\psi}, H)$ reduces to equation (3.11) in Ronchetti and Rousseeuw (1985) when $\boldsymbol{\psi}$ defines an M-estimator. That is,

$$\boldsymbol{\psi}(\boldsymbol{x}, y; \boldsymbol{\beta}) = \delta(\boldsymbol{x}, (y - \boldsymbol{x}\boldsymbol{\beta}))\boldsymbol{x}^T, \tag{2.21}$$

where the function $\delta(\boldsymbol{x},\cdot)$ is continuous except on a finite set, odd and $\delta(\boldsymbol{x},v)\geq 0$ when $v\geq 0$ for all $\boldsymbol{x}$ (see Hampel et al., 1986, for detailed regularity conditions on this function).

Let $\delta'(\boldsymbol{u},v)=\partial\delta(\boldsymbol{u},v)/\partial v$. Then,

$$\boldsymbol{\psi}'(\boldsymbol{x},y;\boldsymbol{\beta})=\Big[-\delta'(\boldsymbol{u},v)\big|_{(\boldsymbol{x},(y-\boldsymbol{x}\boldsymbol{\beta}))}\Big]\boldsymbol{x}^T\boldsymbol{x}, \tag{2.22}$$

and, using the fact that $\delta(\boldsymbol{x},\cdot)$ is odd, we obtain

$$\begin{aligned}
\boldsymbol{c} &= E_H\big[\boldsymbol{\psi}^T(\boldsymbol{X},Y)Q^{-1}\boldsymbol{\psi}'(\boldsymbol{X},Y)M^{-1}\big] \\
&= -E\big[\delta(\boldsymbol{X},(Y-\boldsymbol{X}\boldsymbol{\beta}))\delta'(\boldsymbol{X},(Y-\boldsymbol{X}\boldsymbol{\beta}))\boldsymbol{X}^TQ^{-1}\boldsymbol{X}^T\boldsymbol{X}M^{-1}\big]=0, \qquad (2.23)\\
(\boldsymbol{K}_{ij})_s &= \sum_k E_H\left[\frac{\partial^2\psi_i(\boldsymbol{X},Y;\boldsymbol{\beta})}{\partial\beta_j\partial\beta_k}\right](M^{-1})_{ks} \\
&= \sum_s E\big[\delta''(\boldsymbol{X},(Y-\boldsymbol{X}\boldsymbol{\beta}))X_iX_jX_k\big](M^{-1})_{ks}=0. \qquad (2.24)
\end{aligned}$$

Thus, substituting (2.21)-(2.24) into (2.13), we obtain

$$\begin{aligned}
k^*(T,H) &= p+\sup_{(\boldsymbol{x},y)}\{-2\delta'(\boldsymbol{x},(y-\boldsymbol{x}\boldsymbol{\beta}))\boldsymbol{x}M^{-1}\boldsymbol{x}^T \\
&\qquad +\delta^2(\boldsymbol{x},(y-\boldsymbol{x}\boldsymbol{\beta}))\boldsymbol{x}Q^{-1}\boldsymbol{x}^T]\}.
\end{aligned}$$

Ronchetti and Rousseeuw (1985) proved that V-robustness implies B-robustness and that if $\delta(\boldsymbol{x},\cdot)$ is nondecreasing and QM^{-1} is nonnegative definite, then V-robustness and B-robustness are equivalent.

2.3 Robust M-estimators

In previous Section we proved the nonrobustness of the MLE in linear models. There have been several proposals for choosing the function δ in (2.21) for linear regression problems so that the resulting estimator is both B- and V-robust (eg. Krasker and Welsch, 1982; Hampel et al., 1986). In general they are of the form

$$\delta(\boldsymbol{x},r)=w(\boldsymbol{x})\psi(rv(\boldsymbol{x})), \quad \text{where} \quad r=\frac{y-\boldsymbol{x}\boldsymbol{\beta}}{\sigma}. \tag{2.25}$$

Mallows- and Schweppe-type estimators are those for which $v(\boldsymbol{x})=1$ and $v(\boldsymbol{x})=1/w(\boldsymbol{x})$ respectively.

Extensions to GLMs of both types of estimators have been proposed by Stefanski et al. (1986), Pregibon (1981), Künsch (1989), and Cantoni and Ronchetti (2001), among others. All these authors focused on the B-robustness of the proposed estimators but none of them analyzed their V-robustness. In this Section we study the V-robustness of a class of Mallows-type estimators that are B-robust. Cantoni and Ronchetti (2001) proposed these estimators as a natural extension of some robust estimators of linear models for GLMs.

2.3.1 A Class of robust M-estimators

In this Section we analyze a class of estimators defined as the solution of (2.2) when

$$\boldsymbol{\psi}(\boldsymbol{x}, y; \boldsymbol{\beta}) = \nu(r_i)\omega(\boldsymbol{x}_i)\mu_i' - d(\boldsymbol{\beta}), \tag{2.26}$$

where $r = (y - \mu)/V^{1/2}$ are the Pearson residuals, $\mu' = \nabla_{\boldsymbol{\beta}}\mu = h'(\eta)\boldsymbol{x}$, $d(\boldsymbol{\beta}) = E_H[\nu(r)\omega(\boldsymbol{x})\mu']$ and $\nu(.)$, $\omega(.)$ are weight functions.

These estimators are a natural extension of (2.25) and a robust version of (2.15) for GLMs. As in general the responses are not symmetrically distributed around their means, replacing r with $(y - \mu)/V^{1/2}(\mu)$ in (2.25) does not lead to consistent estimators. Thus, a correcting term, $d(\cdot)$, has to be added in the estimating function, which requires full knowledge of the underling distribution of $(\boldsymbol{X}, Y)$ (Stefanski et al., 1986 and Pregibon, 1981). Künsch et al. (1989) and Cantoni and Ronchetti (2001) defined a class of conditionally unbiased bounded influence estimators that assumed only the conditional distribution of Y given $\boldsymbol{X}$ known.

As the influence function of M-estimators is given by $IF(\boldsymbol{x}, y; \boldsymbol{\psi}, H) = M^{-1}\boldsymbol{\psi}(\boldsymbol{x}, y)$, a bounded function $\boldsymbol{\psi}$ ensures their B-robustness. Thus, for a bounded function ν and a down-weighting function ω in (2.26), the corresponding estimators are B-robust (Cantoni and Ronchetti, 2001). To study their V-robustness, we derive the CVS.

2.3.2 CVS

We start differentiating the function $\boldsymbol{\psi}$ defined in (2.26) with respect to $\boldsymbol{\beta}$ to derive the CVS of this class of estimators. We have

$$\boldsymbol{\psi}'(\boldsymbol{x}, y; \boldsymbol{\beta}) = \omega(\boldsymbol{x}) t(\boldsymbol{x}, y; \mu) \boldsymbol{x}^T \boldsymbol{x} - \nabla_{\boldsymbol{\beta}} d(\boldsymbol{\beta}),$$

where the scalar function t is given by

$$t(\boldsymbol{x}, y; \mu) = \left\{ \nu'(r) \left[\frac{r V'(\mu)}{2V(\mu)} - \frac{1}{V^{1/2}(\mu)} \right] (h'(\eta))^2 + \nu(r) h''(\eta) \right\}. \tag{2.27}$$

Then

$$\mathrm{tr}(\boldsymbol{\psi}' M^{-1}) = \omega(\boldsymbol{x}) t(\boldsymbol{x}, y; \mu) \boldsymbol{x} M^{-1} \boldsymbol{x}^T - \mathrm{tr}(\nabla_{\boldsymbol{\beta}} d(\boldsymbol{\beta}) M^{-1}). \tag{2.28}$$

Plugging (2.26) and (2.28) into (2.13) we obtain the CVS given by

$$\begin{aligned} k^*(\boldsymbol{\psi}, H) &= p - 2\mathrm{tr}(\nabla_{\boldsymbol{\beta}} d(\boldsymbol{\beta}) M^{-1} + d(\boldsymbol{\beta})^T Q^{-1} d(\boldsymbol{\beta})) + \sup_{(\boldsymbol{x}, y)} \{ 2\omega(\boldsymbol{x}) t(\boldsymbol{x}, y; \mu) \boldsymbol{x} M^{-1} \boldsymbol{x}^T \\ &\quad + 2[\nu(r)\omega(\boldsymbol{x}) h'(\eta) \boldsymbol{x} - d(\boldsymbol{\beta})](\boldsymbol{c} - \boldsymbol{a})^T + [\nu(r)\omega(\boldsymbol{x}) h'(\eta)]^2 \boldsymbol{x} Q^{-1} \boldsymbol{x}^T \}, \end{aligned} \tag{2.29}$$

where the constant vectors $\boldsymbol{c}$ and $\boldsymbol{a}$ were defined in (2.13).

It is interesting to see that in general, for GLMs, V-robustness does not imply B-robustness as it was proved in particular for classical linear models (Ronchetti and Rousseeuw, 1985). The function $\omega(\boldsymbol{x})$ may downweight leverage points so that $\boldsymbol{\psi}$ is bounded and still may not suffice to control the scalar function $t(\boldsymbol{X}, Y; \mu)$ defined in (2.27). As an example, consider a Poisson model with a canonical link (see Example 1-b in Section 2.2.1). In this case,

$$h(\eta) = h'(\eta) = h''(\eta) = V(\mu) = e^{\eta}$$

The following choices of $\omega(\boldsymbol{x})$ and $\nu(.)$ ensure the B-robustness of the estimators.

$$\omega(\boldsymbol{x}) = \frac{1}{\| e^{\eta} \boldsymbol{x} \|} \quad \text{and} \quad h_c(r) = \begin{cases} r & |r| \leq c \\ c\,\mathrm{sign}(r) & |r| > c \end{cases} \tag{2.30}$$

where c is a tuning constant. For this particular case, the last two terms of (2.29) are bounded as $\boldsymbol{\psi}$ is bounded. However, the function $t(\boldsymbol{X}, Y; \mu)$ defined in (2.27) contains a quadratic term $e^{2\eta}$ which is not downweighted by $\omega(\boldsymbol{x})$. Then, the first term in the supremum in (2.29) is still unbounded

with respect to $\boldsymbol{x}$. Thus, for these choices of weighting functions, the resulting estimators are not V-robust.

2.4 Simulation

We carried out a set of simulations to compare the sensitivity of the estimated variance of the MLE with the estimated variance of a robust estimator (ROB) under different levels of contamination in a Logistic model with one covariate. For each level of contamination, we generated 500 Monte Carlo replications of the response variable using S-Plus Version 6.2.1. The values of the covariate are fixed in all the simulations and range from 1.52 to 2.36 in an equally spaced grid. The MLE of the regression coefficients and the estimates of its variance matrix were obtained using the `glm` function available in S-Plus. The robust estimator is in the class of M-estimators defined in (2.26) using the Huber function defined in (2.30) for $\nu(\cdot)$ and $w(\boldsymbol{x}) = 1$. For the Huber function we used two different values for the tuning constant, $c = 1.2$ and $c = .8$. The estimates of the coefficients and the variance matrix were obtain using an algorithm written by Cantoni and Ronchetti (2001). We also report the asymptotic efficiency of the robust estimator relative to the MLE. This quantity corresponds to the ratio of the traces of the variance matrices.

2.4.1 The Logistic model

We generated clustered binary data according to the following model

$$P(Y_{ij} = 1) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}, \quad \text{for} \quad i = 1, \ldots, 29; \quad \text{and} \quad j = 1, \ldots, 20, \tag{2.31}$$

with regression coefficients given by $\beta_0 = 2$ and $\beta_1 = -2$. The binary data were grouped by covariate class so that the responses represent the number of successes in each cluster.

Table 2.1 summarizes the results of the parameter estimates obtained by the MLE and the robust estimator (ROB) for the generated data. The last two rows correspond to the Monte Carlo standard errors and the mean of the standard errors obtained using the algorithms, respectively. When there is no contamination, the biases of both the MLE and the ROB estimators of the beta coefficients are not significantly different from zero. The bias of the robust estimator is reduced

when its tuning constant c is smaller. However, the efficiency of the estimator relative to the MLE decreases as well from .9467 to .8656. The estimated standard errors are slightly overestimated using the `glm` and Cantoni and Ronchetti's algorithms.

Table 2.1: Summary of maximum likelihood (MLE) and robust (ROB) estimation for uncontaminated data in a Logistic Model.

	$c = 1.2$				$c = .8$			
	MLE		ROB		MLE		ROB	
	β_0	β_1	β_0	β_1	β_0	β_1	β_0	β_1
Min.	-0.7548	-3.3021	-1.1359	-3.3392	-0.8777	-3.3640	-0.7365	-3.7034
1st Qu.	1.3687	-2.3304	1.3228	-2.3717	1.2806	-2.3166	1.2526	-2.3580
Median	2.0231	-2.0128	2.0788	-2.0332	2.0154	-2.0234	2.0320	-2.0108
Mean	2.0195	-2.0121	2.0126	-2.0078	1.9753	-1.9898	1.9841	-1.9936
3rd Qu.	2.6667	-1.6652	2.6686	-1.6166	2.5894	-1.5970	2.7161	-1.6099
Max.	4.4567	-0.6145	4.6807	-0.4784	4.4945	-0.5340	5.0911	-0.6397
MC.sd	0.9266	0.4974	0.9519	0.5112	0.9196	0.4915	0.9615	0.5130
est.sd	0.9467	0.5068	0.9824	0.5262	0.9466	0.5065	1.0174	0.5445

2.4.2 The contaminated model

In this Section we examine the effect of different levels of contamination on the maximum likelihood and the robust estimation. In each generated dataset, one cluster i is randomly chosen and with probability $(1 - \varepsilon)$ the observations y_{ij} in the cluster that are 0 are turned into 1. The responses are again grouped by covariate class and we fit the Logistic model described in (2.31).

Tables 2.2-2.4 present the results of the simulation for the data generated under a contaminated model, using a tuning constant $c = 1.2$ and $c = .8$ to construct the robust estimator. Figure 2.1 and Figure 2.2 show that the bias of the MLE increases as the level of contamination increases, while that of the robust estimator (ROB) remains almost constant for both values of c. However, the bias of these estimators is not significantly different from zero considering the

estimated standard errors presented in last row of these tables.

Table 2.2: Summary of maximum likelihood (MLE) and robust (ROB) estimation using $c = 1.2$, for various levels of contaminated data in a Logistic Model.

	MLE		ROB		MLE		ROB	
	β_0	β_1	β_0	β_1	β_0	β_1	β_0	β_1
	5% contamination				10% contamination			
Min.	-1.7595	-3.5773	-1.8903	-3.6289	-1.1987	-3.6296	-1.1567	-3.8509
1st Qu.	1.3807	-2.3772	1.3773	-2.3863	1.3098	-2.3147	1.2777	-2.3691
Median	2.0590	-2.0280	2.0892	-2.0247	1.9862	-1.9673	2.0122	-1.9989
Mean	2.0466	-2.0254	2.0290	-2.0148	1.9453	-1.9673	1.9806	-1.9875
3rd Qu.	2.6799	-1.6664	2.7376	-1.6564	2.6289	-1.6297	2.6854	-1.6202
Max.	4.9967	-0.1254	5.1118	-0.0503	4.8443	-0.2771	5.1597	-0.2814
	20% contamination				30% contamination			
Min.	-0.8327	-3.6770	-0.8941	-3.4581	-0.9000	-3.5435	-0.5828	-3.4148
1st Qu.	1.2577	-2.3001	1.2968	-2.3304	1.0483	-2.3046	1.3224	-2.3494
Median	1.8884	-1.9313	1.9647	-1.9666	1.8508	-1.8810	1.9884	-1.9789
Mean	1.9190	-1.9366	1.9634	-1.9690	1.8661	-1.8954	2.0171	-1.9951
3rd Qu.	2.5661	-1.5923	2.6672	-1.5954	2.6417	-1.4769	2.6916	-1.6227
Max.	5.1027	-0.4582	4.6842	-0.4088	4.8628	-0.4323	4.7456	-0.6242
	40% contamination				50% contamination			
Min.	-1.4049	-3.6251	-0.9642	-3.5806	-1.9302	-3.6233	-1.2113	-3.5559
1st Qu.	0.9911	-2.2364	1.2731	-2.3223	0.9138	-2.2813	1.2958	-2.4215
Median	1.9086	-1.8786	2.0225	-1.9736	1.7519	-1.7966	1.9846	-1.9757
Mean	1.7828	-1.8389	1.9740	-1.9722	1.7683	-1.8164	1.9957	-1.9805
3rd Qu.	2.5389	-1.4124	2.6264	-1.6031	2.6694	-1.3668	2.8081	-1.5955
Max.	5.1939	-0.1428	5.0892	-0.5024	5.1219	0.0796	4.9889	-0.3342

Table 2.3: Summary of maximum likelihood (MLE) and robust (ROB) estimation using $c = .8$, for various levels of contaminated data in a Logistic Model.

	MLE		ROB		MLE		ROB	
	β_0	β_1	β_0	β_1	β_0	β_1	β_0	β_1
	5% contamination				10% contamination			
Min	-0.6247	-3.5177	-1.1913	-3.5637	-1.0991	-3.5059	-1.2039	-3.6807
1st Qu	1.3611	-2.3937	1.3877	-2.3985	1.2671	-2.2925	1.2936	-2.3186
Median	2.0484	-2.0152	2.0777	-2.0305	1.9197	-1.9419	1.9657	-1.9531
Mean	2.0594	-2.0334	2.0590	-2.0328	1.9232	-1.9507	1.9839	-1.9836
3rd Qu.	2.7384	-1.6762	2.7768	-1.6682	2.5642	-1.5879	2.6048	-1.5993
Max	4.7088	-0.6826	4.8789	-0.5246	4.9747	-0.3288	4.9895	-0.2965
	20% contamination				30% contamination			
Min	-0.9245	-3.8372	-1.3915	-3.9301	-0.8242	-3.9357	-1.0934	-3.7273
1st Qu.	1.2288	-2.2896	1.3691	-2.3765	1.1221	-2.2363	1.3032	-2.3186
Median	1.9212	-1.8963	2.0182	-1.9693	1.8223	-1.8636	2.0180	-1.9884
Mean	1.9135	-1.9306	2.0646	-2.0188	1.8454	-1.8855	2.0010	-1.9898
3rd Qu.	2.5893	-1.5614	2.8056	-1.6373	2.4620	-1.5201	2.6552	-1.6164
Max.	5.3126	-0.4582	5.5562	-0.2790	5.5506	-0.5593	5.1225	-0.4277
	40% contamination				50% contamination			
Min.	-1.2236	-4.1892	-0.7640	-3.8083	-1.6370	-4.3537	-0.9848	-4.0250
1st Qu.	1.0760	-2.2464	1.3246	-2.3371	0.8108	-2.3223	1.1498	-2.3348
Median	1.8088	-1.8266	2.0358	-1.9940	1.7012	-1.7711	1.9199	-1.9424
Mean	1.8166	-1.8535	2.0224	-1.9970	1.7622	-1.8147	1.9511	-1.9591
3rd Qu.	2.6091	-1.4548	2.6604	-1.6330	2.7166	-1.3026	2.6534	-1.5339
Max.	6.2623	-0.2642	5.5913	-0.6453	6.1654	0.0290	5.7108	-0.4373

Table 2.4: Monte Carlo and estimated standard errors of the maximum likelihood (MLE) and the robust (ROB) estimators for various levels of contaminated data in a Logistic Model.

	$c = 1.2$							
	MLE		ROB		MLE		ROB	
	β_0	β_1	β_0	β_1	β_0	β_1	β_0	β_1
	5% contamination				10% contamination			
MC.sd	0.9788	0.5233	0.9932	0.5307	0.9969	0.5325	1.0276	0.5495
est.sd	0.9465	0.5068	0.9816	0.5258	0.9425	0.5040	0.9801	0.5246
	20% contamination				30% contamination			
MC.sd	1.0044	0.5436	1.0052	0.5412	1.0844	0.5819	1.0010	0.5349
est.sd	0.9318	0.4978	0.9735	0.5207	0.9233	0.4926	0.9726	0.5205
	40% contamination				50% contamination			
MC.sd	1.1692	0.6240	1.0343	0.5531	1.3121	0.7014	1.0621	0.5657
est.sd	0.9151	0.4874	0.9719	0.5198	0.9070	0.4826	0.9703	0.5190
	$c = .8$							
	5% contamination				10% contamination			
MC.sd	0.9421	0.5075	1.0084	0.5418	0.9636	0.5157	1.0008	0.5379
est.sd	0.9474	0.5074	1.0187	0.5457	0.9390	0.5019	1.0113	0.5410
	20% contamination				30% contamination			
MC.sd	0.9813	0.5282	1.0366	0.5584	1.0334	0.5499	1.0231	0.5447
est.sd	0.9296	0.4964	1.0076	0.5393	0.9236	0.4926	1.0097	0.5401
	40% contamination				50% contamination			
MC.sd	1.1270	0.6027	1.0440	0.5534	1.3252	0.7108	1.1097	0.5999
est.sd	0.9136	0.4867	1.0074	0.5389	0.9081	0.4833	1.0071	0.5384

Figure 2.3 and Figure 2.4 illustrate that the standard errors of the MLE slightly increase

with the level of contamination while those of the ROB estimator are almost constant. Moreover, Figure 2.5 and Figure 2.6 present the estimated bias of the standard errors of both estimators for the two values of c, respectively. Again, the difference between the MC and the estimated standard errors of the MLE becomes more important with more contamination. However, for both values of c, bias of the estimated standard error of the ROB estimator remains almost unchanged.

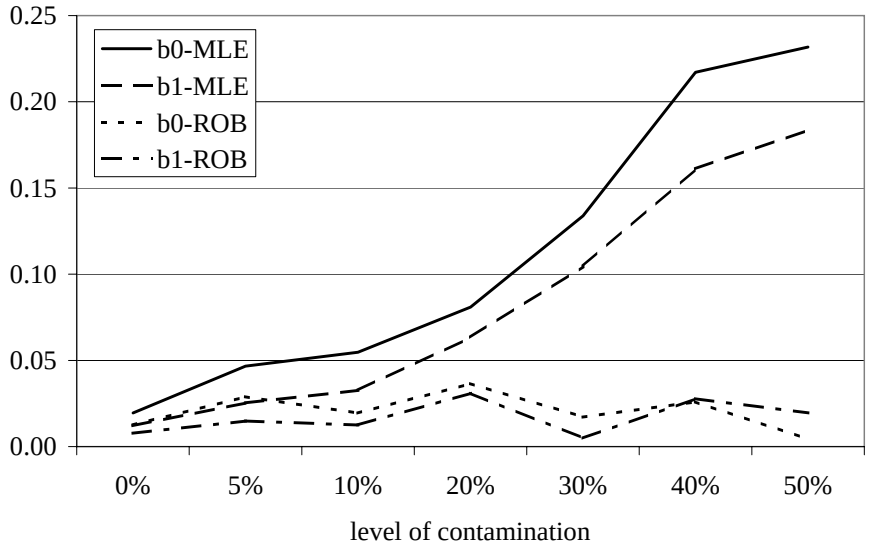

Figure 2.1: Bias of the MLE and the robust (ROB) estimator using $c = 1.2$.

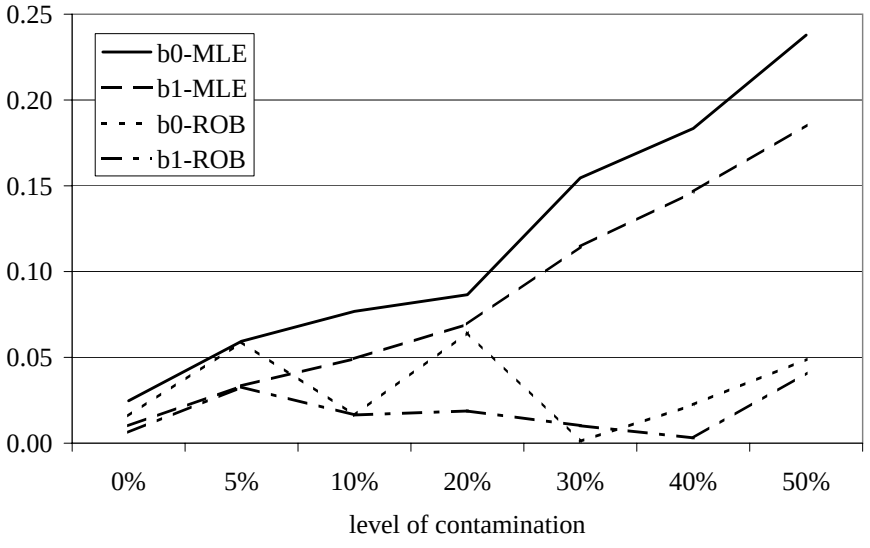

Figure 2.2: Bias of the MLE and the robust (ROB) estimator using $c = .8$.

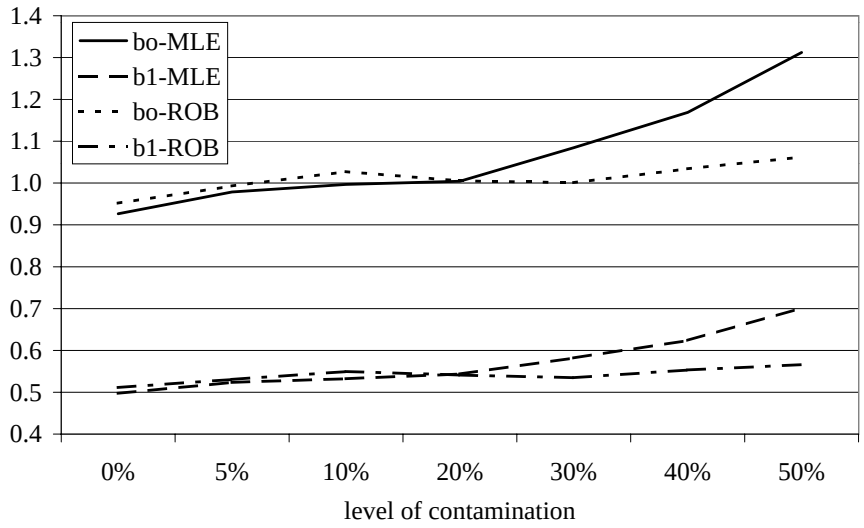

Figure 2.3: Estimated standard errors of the MLE and the robust (ROB) estimator using $c = 1.2$.

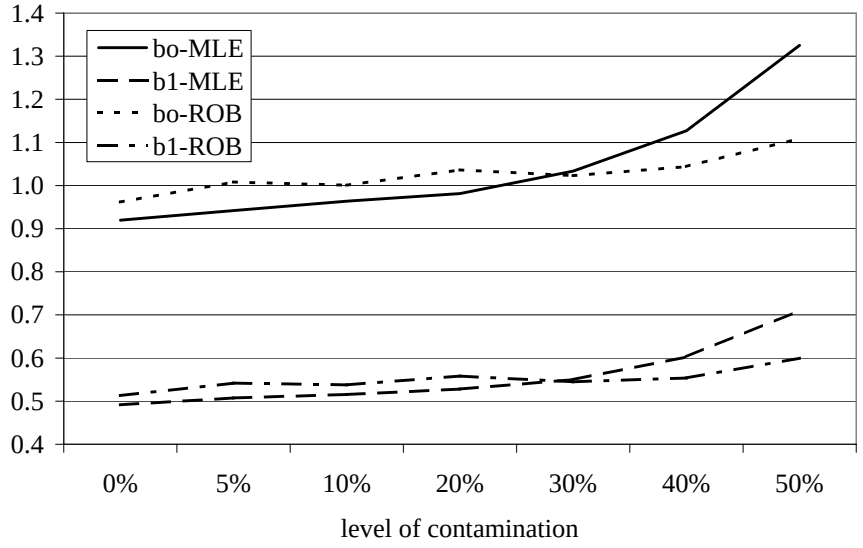

Figure 2.4: Estimated standard errors of the MLE and the robust (ROB) estimator using $c = .8$.

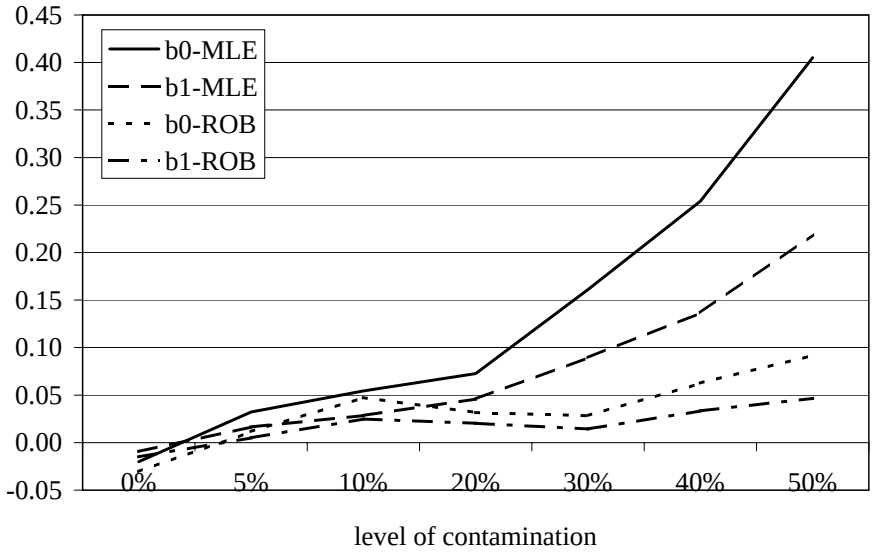

Figure 2.5: Bias of the estimated standard errors of the MLE and the robust (ROB) estimator using $c = 1.2$.

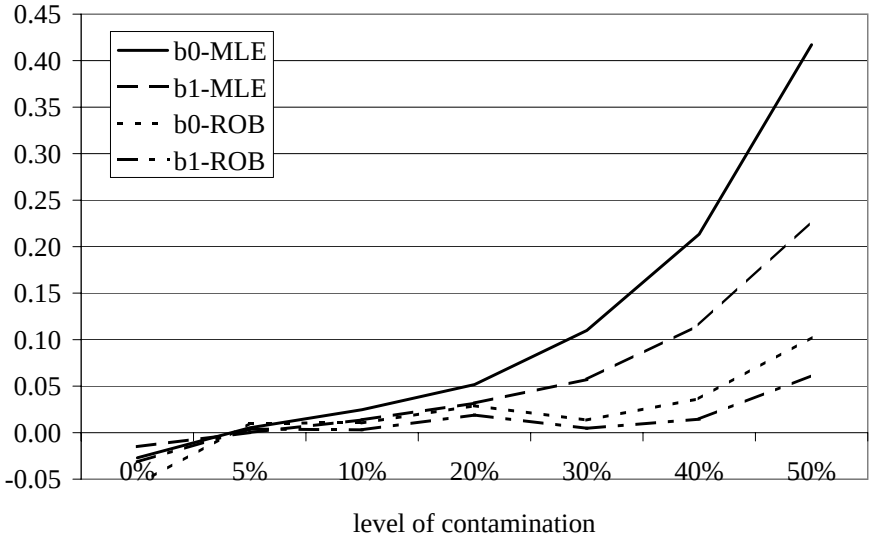

Figure 2.6: Bias of the estimated standard errors of the MLE and the robust (ROB) estimator using $c = .8$.

2.5 Conclusions

In this Chapter we extend the definitions of CVF and CVS to GLMs and study how an ε-contamination in the distribution perturbs the asymptotic variance of the estimators. We derive the CVF and the CVS for the class of M-estimators and we analyze in detail the MLE of GLMs with canonical links. In particular we derive the CVS for three commonly used GLMs: Logistic, Poisson and Gamma models. We found that, in general, the MLE is not V-robust, thus a contamination of the distribution can seriously affect its asymptotic variance. Moreover, we obtain the CVF for the class of M-estimators in the subclass of linear models, which was previously analyzed by Ronchetti and Rousseeuw (1985). We also study the CVS of a class of Mallows-type estimators and conclude that in general, for GLMs, V-robustness does not imply B-robustness as was proved for linear models (Ronchetti and Rousseeuw, 1985).

We perform a simulation study to compare the performance of the MLE with that of a robust estimator in a Logistic model. In all simulated cases, the variance of the robust estimator remains almost constant and unbiased under different levels of contamination. However, that of the MLE increases with the level of contamination as well as its bias. The bias of the MLE is also increasing while that of the robust estimators does not change significantly.

Chapter 3

Change-of-variance function in GLMMs

3.1 Introduction

This Chapter investigates how the asymptotic variance of the estimators of Generalized Linear Mixed Models (GLMMs) is affected when the conditional distribution of the responses is correctly specified, but the mixing distribution of the random effects is slightly contaminated. To study the infinitesimal stability of the asymptotic variance of the estimators, we extend the notion of Change-of-Variance Function (CVF) and the Change-of-Variance Sensitivity (CVS) to GLMMs.

Generalized Linear Mixed Models are used to model the relationship between a function of the mean of the responses and a linear predictor that include a linear combination of random components. In addition GLMMs can accommodate nonnormally distributed responses such as Gamma or Poisson random variables. The random effects included in the linear predictor allow us to account for correlation between observations and overdispersion or to make subject-specific inference. A commonly used estimator for these models is the marginal MLE. Provided that the model is correctly specified and that the usual regularity conditions hold, this estimator is consistent and asymptotically normal (White, 1982). Some authors derive the joint maximum likelihood estimator to overcome computational difficulties (see Section 1.3.2). However, a contamination in the mixing distribution does not affect the estimation of the model coefficients. Thus we are not examining this estimator here. For further details see Section 1.3.

In most practical applications, one rarely knows the true model. A natural question is what happens to the estimation if one does not assume the correct model. In particular, in this Chapter we will examine the case where the conditional distribution of the responses is correctly specified, but the mixing distribution of the random effects U is not. Gustafson (1996) studied the inconsistency of maximum likelihood estimators for certain conjugate mixture models under

misspecifications of the mixing distribution. He investigated the magnitude of the asymptotic bias using an influence function approach. Smith and Weems (2004) extended Gustafson's approach to include a regression structure in the mean. They proved that the maximum likelihood estimators are robust under perturbations of the mixing distribution for Poisson-lognormal models. Neuhaus et al. (1992) examine the performance of the mixed-effects logistic regression MLE when the mixing distribution is misspecified. By a simulation experiment, they also studied the effect of the misspecification over the estimated standard errors of the estimators. However, to the best of my knowledge, there is no previous analytical work on the local effect of a mixing distribution misspecification in the asymptotic variance of the estimators for GLMMs.

The remainder of the Chapter is organized as follows. In Section 3.2 we extend the notions of CVF and CVS for GLMMs. In particular, we derive the CVF of the (marginal) MLE. The CVS of this estimator is analyzed in detail for the Poisson-Gamma model in Section 3.3 and for two mixed-effects Binomial models in Section 3.4. A simulation study is performed for the Poisson-Gamma model and the results are summarized in Section 3.5. We end with some conclusions and some future research directions in Section 3.6.

3.2 The CVF of the M-estimators

A misspecification of the mixing distribution may affect not only the behavior of the estimator itself but also its asymptotic variance. One can investigate the infinitesimal effects of a contamination of the type (3.1) on the asymptotic variance of the estimator by studying the CVF and the CVS. Although Definitions 1.5.5 and 1.5.6 were made in the framework of M-estimation of a one-dimensional parameter, they can be extended to the case of multivariate parameters. Ronchetti et al. (1985) defined the CVF and the CVS for estimators of classical linear regression coefficients. In this Section we extend these definitions for M-estimators of GLMMs parameters under a contamination in the mixing distribution.

Consider the model introduced in Section 1.3.1. Let f_F be the marginal density of Y_i given

$\boldsymbol{X}_i = \boldsymbol{x}_i$ when the random effects are distributed according to the mixing distribution F. That is,

$$f_F(y_i) = f_{Y_i|\boldsymbol{x}_i}(y_i) = \int \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + h(y_i,\phi)\right\} f(\boldsymbol{u})d\boldsymbol{u}, \quad i = 1,\ldots,n.$$

We are interested in estimating the vector of unknown parameters $\boldsymbol{\gamma} = (\boldsymbol{\beta}^T, \boldsymbol{\tau}^T) \in \mathbb{R}^{(p+q)}$, where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of regression coefficients and $\boldsymbol{\tau} \in \mathbb{R}^q$ is the vector of unknown parameters of the mixing distribution.

Suppose that the mixing distribution F is slightly contaminated by a distribution G, so that the random variables $\boldsymbol{U}$ are actually generated from a distribution which is an ε-contamination of the nominal distribution, denoted

$$F_\varepsilon = (1-\varepsilon)F + \varepsilon G. \tag{3.1}$$

We will assume that G is any distribution having the same first two moments as F. This restriction on G ensures that $\boldsymbol{\gamma}$ is interpretable as the true parameter vector, no matter how much the true model deviates from the nominal model (Gustafson, 1996). Let Λ be the class of all such distributions G.

Notation 3.2.1. *Let E_F be the expected value taken with respect to the density f_F. Similarly define E_G and E_ε.*

Let H_F be the joint distribution of the response, the random effects and the covariates assuming the correct distribution F for the random effects and H_G be the corresponding one when the contaminating distribution G is assumed. Note that a contamination of type (3.1) in the mixing distribution induces the same kind of contamination in the joint distribution, i.e. $H_\varepsilon = (1-\varepsilon)H_F + \varepsilon H_G$.

The M-estimator $\hat{\boldsymbol{\gamma}}$ is the solution of

$$\sum_{i=1}^{n} \boldsymbol{\psi}(\boldsymbol{x}_i, y_i; \boldsymbol{\gamma}) = \boldsymbol{0} \tag{3.2}$$

for suitably chosen functions $\boldsymbol{\psi}$ from $\mathbb{R}^p \times \mathbb{R} \times \mathbb{R}^{(p+q)}$ to $\mathbb{R}^{(p+q)}$ such that

$$E_H[\boldsymbol{\psi}(\boldsymbol{X}, Y; \boldsymbol{\beta})] = \boldsymbol{0}.$$

Note that $\hat{\boldsymbol{\gamma}}$ can be also defined as $\hat{\boldsymbol{\gamma}} = \boldsymbol{T}(H_n)$, where the functional $\boldsymbol{T}$ is implicitly defined by

$$\int \boldsymbol{\psi}(\boldsymbol{x}, y; \boldsymbol{T}(H)) dH(\boldsymbol{x}, y) = \mathbf{0}. \tag{3.3}$$

Under regularity conditions, by the law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\psi}(\boldsymbol{x}_i, y_i; \boldsymbol{\gamma}) \rightarrow E_{\varepsilon}[\boldsymbol{\psi}(\boldsymbol{X}_i, Y_i; \boldsymbol{\gamma})],$$

where the expected value is taken under the true model (3.1). Let $\boldsymbol{\gamma}_{\varepsilon}$ be the solution of the equation

$$E_{\varepsilon}[\boldsymbol{\psi}(\boldsymbol{X}_i, Y_i; \boldsymbol{\gamma})] = 0. \tag{3.4}$$

Then the zeros of (3.2) and those of (3.4) should also become close as n goes to infinity. In other words, under regularity conditions (Huber, 1967) we expect $\hat{\boldsymbol{\gamma}}$ to converge to $\boldsymbol{\gamma}_{\varepsilon}$. In particular, note that when $\boldsymbol{\psi}(\boldsymbol{x}, y; \boldsymbol{\gamma}) = \nabla_{\boldsymbol{\gamma}} \log(f_F(y; \boldsymbol{x}, \boldsymbol{\gamma}))$, then $\hat{\boldsymbol{\gamma}}$ is the MLE.

Definition 3.2.1. *The Change-of-Variance Function (CVF) of $\boldsymbol{T}$ at H_F under a contamination G in the mixing distribution is defined as*

$$CVF(\boldsymbol{T}, H_F, G) = \frac{\partial}{\partial \varepsilon}[\boldsymbol{V}(\boldsymbol{T}, H_{\varepsilon})]_{\varepsilon=0}.$$

Definition 3.2.2. *The unstandardized Change-of-Variance Sensitivity of T at H_F is*

$$k^*(\boldsymbol{T}, H_F) = \sup_{G \in \Lambda} \{tr\ CVF(\boldsymbol{T}, H_F, G)/tr\ \boldsymbol{V}(\boldsymbol{T}, H_F)\}.$$

The estimator is called V-robust when its CVS, k^*, is finite.

In order to derive the CVF presented in Definition (3.2.1), we need the asymptotic variance of the M-estimates of GLMMs parameters. Huber (1967) proved asymptotic normality of M-estimators under weaker conditions than usual for a general class of models. Using a Taylor expansion approach, one can derive the asymptotic variance presented in next theorem.

Theorem 3.2.1. *Under regularity conditions, the asymptotic variance of the M-estimator $\widehat{\boldsymbol{\gamma}}$ defined by the functional $\boldsymbol{T}$ in (3.3) is given by*

$$V(\boldsymbol{T}, H_{\varepsilon}) = M_{\varepsilon}^{-1} Q_{\varepsilon} \left\{M_{\varepsilon}^{-1}\right\}^{T}, \tag{3.5}$$

where

$$M_\varepsilon = E_\varepsilon \left[-\nabla_{\boldsymbol{\gamma}} \boldsymbol{\psi}(\boldsymbol{X}, Y; \boldsymbol{\gamma}) \big|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}_\varepsilon} \right], \tag{3.6}$$

$$Q_\varepsilon = E_\varepsilon \left[\left(\nabla_{\boldsymbol{\gamma}} \boldsymbol{\psi}(\boldsymbol{X}, Y; \boldsymbol{\gamma}) \right) \left(\nabla_{\boldsymbol{\gamma}} \boldsymbol{\psi}(\boldsymbol{X}, Y; \boldsymbol{\gamma}) \right)^{\boldsymbol{T}} \big|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}_\varepsilon} \right]. \tag{3.7}$$

Applying Definition 3.2.1 to the asymptotic variance defined in (3.5), we can derive the CVF of the M-estimators of GLMMs. As the MLE is a commonly used estimator in GLMMs (Anderson and Aitkin, 1985; Crouch and Siegelman, 1990; McCulloch, 1997), in this Section we will derive its CVF instead. In particular, in Sections 3.3 and 3.4 we will examine the CVF and the CVS of the Poisson-Gamma models and two mixed effects Binomial models respectively.

3.2.1 The MLE

We introduce some notation that will be used throughout this Chapter. For simplicity, the subindex i corresponding to the ith observation is omitted in the following results.

Notation 3.2.2. *Let $l = \log f_F(y; \boldsymbol{x}, \boldsymbol{\gamma})$ be the log-likelihood function of Y given $\boldsymbol{x}$ under the nominal model.*

Notation 3.2.3. *Let $l_r = (\partial/\partial\gamma_r) \log f_F(y; \boldsymbol{x}, \boldsymbol{\gamma})\big|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}_0}$; $l_{rj} = (\partial^2/\partial\gamma_r\partial\gamma_j) \log f_F(y; \boldsymbol{x}, \boldsymbol{\gamma})\big|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}_0}$; and $l_{rjk} = (\partial^3/\partial\gamma_r\partial\gamma_j\partial\gamma_k) \log f_F(y; \boldsymbol{x}, \boldsymbol{\gamma})\big|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}_0}$, where $\gamma_r = \beta_r$ for $r = 1, \ldots, p$, and $\gamma_r = \tau_r$ for $r = p+1, ..., p+q$.*

Notation 3.2.4. *Let $I^{ks} = [I^{-1}(\boldsymbol{\gamma})]_{ks}$, where $I(\boldsymbol{\gamma})$ is the Fisher information matrix.*

Notation 3.2.5. *$E_{G-F}(\cdot) = E_G(\cdot) - E_F(\cdot)$.*

Notation 3.2.6. *Define $J_{rjk} = E_F[l_{rjk}]$, and let J_{rj} be the vector obtained by fixing the first two indices of the three-way array.*

We start by evaluating (3.6) and (3.7) at $\varepsilon = 0$:

$$\begin{aligned} M_\varepsilon|_{\varepsilon=0} &= E_F \left[-\nabla^2_{\boldsymbol{\gamma}\boldsymbol{\gamma}^{\boldsymbol{T}}} \log(f_F(Y; \boldsymbol{x}, \boldsymbol{\gamma})) \big|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}_0} \right] = I(\boldsymbol{\gamma}_0), \\ Q_\varepsilon|_{\varepsilon=0} &= E_F \left[\left(\nabla_{\boldsymbol{\gamma}} \log(f_F(Y; \boldsymbol{x}, \boldsymbol{\gamma})) \right) \left(\nabla_{\boldsymbol{\gamma}} \log(f_F(Y; \boldsymbol{x}, \boldsymbol{\gamma})) \right)^{\boldsymbol{T}} \big|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}_0} \right] = I(\boldsymbol{\gamma}_0). \end{aligned}$$

Assuming that interchange of expectation and differentiation is allowed, and after some straightforward calculations, one obtains:

$$CVF(MLE, H_F, G) = \frac{\partial V(\varepsilon)}{\partial \varepsilon}\bigg|_{\varepsilon=0} = I^{-1}(\boldsymbol{\gamma}_0)\bigg[-2\frac{\partial M(\varepsilon)}{\partial \varepsilon}\bigg|_{\varepsilon=0} + \frac{\partial Q(\varepsilon)}{\partial \varepsilon}\bigg|_{\varepsilon=0}\bigg] I^{-1}(\boldsymbol{\gamma}_0), \qquad (3.8)$$

where

$$\left(\frac{\partial M(\varepsilon)}{\partial \varepsilon}\bigg|_{\varepsilon=0}\right)_{ij} = \sum_k \left[\sum_s \left(I^{ks} E_G(l_s)\right)\right] E_F(l_{ijk}) + E_{G-F}(l_{ij}), \qquad (3.9)$$

$$\left(\frac{\partial Q(\varepsilon)}{\partial \varepsilon}\bigg|_{\varepsilon=0}\right)_{ij} = \sum_k \left[\sum_s \left(I^{ks} E_G(l_s)\right)\right] E_F(l_{ik}l_j + l_r l_{jk}) + E_{G-F}(l_r l_j). \qquad (3.10)$$

Therefore, the V-robustness of the MLE depends on the contaminating function G through the first two order derivatives of the log-likelihood function.

3.3 The Poisson-Gamma Model

Poisson models are widely used in various areas of application such as biology, reliability and environmental statistics, where the observed responses consist of the number of times an event occurs. Examples include the Gaver and O'Muircheartaigh (1987) data, which consists of the number of failures of 10 pumps (Lee and Nelder, 1996), or the number of colonies produced in the spleen of a recipient animal (Frome et al., 1973).

A common practical complication of these models is overdispersion. In most cases, count data display substantial extra variation relative to the Poisson, which is completely determined by its mean. Some authors studied the effect of overdispersion on inferences made under the Poisson model (Paul and Plackett, 1978; Cox 1983). Many models have been proposed to accommodate overdispersion in statistical analysis, including the use of GLM with random effects (Lee and Nelder, 1996).

If the distribution of multiplicative random effects applied to the mean of a Poisson model is assumed to be Gamma, then the marginal distribution of the response is Negative Binomial. This mixture of Poisson distributions is called Poisson-Gamma. In this Section we will derive the CVF when the Gamma mixing distribution is contaminated by another distribution G. For simplicity, we will examine the model with only one fixed and one random effect. That is,

- Let $Y_i|x_i, u_i$ be a Poisson random variable with mean $u_i\alpha_i$, where $\alpha_i = \exp(\beta_0 + \beta_1 x_i)$.
- Let X_i and U_i be independent random variables.
- Assume further that $U_i \sim \Gamma(1/\tau, \tau)$. Then $E(U_i) = 1$, and $V(U_i) = \tau$

Therefore the conditional distribution of $Y_i|x_i$ is Negative Binomial, so that the log-likelihood function is given by

$$l = K + \log\Gamma\left(y_i + \frac{1}{\tau}\right) - \log\Gamma\left(\frac{1}{\tau}\right) - \left(y_i + \frac{1}{\tau}\right)\log(1 + \alpha_i\tau) + y_i\log(\alpha_i) + y_i\log(\tau), \quad (3.11)$$

$$E[Y_i|x_i] = \alpha_i,\ E[Y_i^2|x_i] = \alpha_i^2(1+\tau) + \alpha_i. \quad (3.12)$$

For simplicity, the subindex i is omitted in the following results.

Notation 3.3.1. *Let* $\rho = \alpha/(1 + \tau\,\alpha)$.

Notation 3.3.2. *Let* $\Psi^{(n)}(u) = (d^{n+1}/du^{n+1})(\log(\Gamma(u))$. *These functions are known as polygamma functions. For example: the Digamma function is the function* $\Psi(u) = (d/du)(\log(\Gamma(u))$ *and the Trigamma function is* $\Psi'(u) = (d^2/(du^2)\log(\Gamma(u))$.

Notation 3.3.3. *For any* $z > 0$, *let* $\Delta\Psi(y; z) = \Psi(y + z) - \Psi(z)$, *and similarly define* $\Delta\Psi'(y; z)$ *and* $\Delta\Psi''(y; z)$.

It is easy to show that for any positive integer n

$$0 \le \Delta\Psi(n; z) = \sum_{j=0}^{n-1} \frac{1}{(z+j)} \le \frac{n}{z} \quad (3.13)$$

and

$$-\frac{n}{z^2} \le \Delta\Psi'(n; z) = -\sum_{j=0}^{n-1} \frac{1}{(z+j)^2} \le 0 \quad (3.14)$$

3.3.1 The CVF

Using (3.8)-(3.11), we can derive the CVF for $\hat{\boldsymbol{\gamma}}$, the (marginal) MLE of $\boldsymbol{\gamma} = (\beta_0, \beta_1, \tau)^T$, when the mixing distribution F is contaminated by a distribution G. For simplicity in the notation we

will write $\Delta\Psi(Y)$ instead of $\Delta\Psi(Y;1/\tau)$ throughout this Chapter. Similarly we will write $\Delta\Psi'(Y)$ and $\Delta\Psi''(Y)$. Tedious calculations give

$$I(\boldsymbol{\gamma}_0) = \begin{pmatrix} \boldsymbol{B} & \mathbf{0} \\ \mathbf{0} & -e_{33}^F \end{pmatrix},$$

where $\boldsymbol{B} = \begin{pmatrix} E_X(\rho) & E_X(\rho X) \\ E_X(\rho X) & E_X(\rho X^2) \end{pmatrix}$ and $e_{33}^F = E_F(l_{33}) = \tau^{-2}E_X(\rho) + \tau^{-4}E_F[\Delta\Psi'(Y)]$. Moreover,

$$\left.\frac{\partial M(\varepsilon)}{\partial \varepsilon}\right|_{\varepsilon=0} = \frac{-1}{e_{33}^F}\begin{pmatrix} e_3^G A & \mathbf{0} \\ \mathbf{0} & e_3^G e_{333}^F - a_{33}^G e_{33}^F \end{pmatrix}, \tag{3.15}$$

where

$$\begin{aligned}
A &= \begin{pmatrix} E_X(\rho^2) & E_X(\rho^2 X) \\ E_X(\rho^2 X) & E_X(\rho^2 X^2) \end{pmatrix}, \\
e_3^G &= E_G(l_3) = \frac{1}{\tau^2}E_X[\log(1+\tau\alpha)] - \frac{1}{\tau^2}E_G[\Delta\Psi(Y)], \\
a_{33}^G &= E_{G-F}(l_{33}) = -\frac{2}{\tau^3}E_X[\log(1+\tau\alpha)] + \frac{2}{\tau^3}E_G[\Delta\Psi(Y)] + \frac{1}{\tau^4}E_{G-F}[\Delta\Psi'(Y)], \\
e_{333}^F &= E_F(l_{333}) = \frac{1}{\tau^3}E_X\left[-\frac{\alpha(4+5\alpha\tau)}{(1+\tau\alpha)^2}\right] - \frac{6}{\tau^5}E_F[\Delta\Psi'(Y)] - \frac{1}{\tau^6}E_F[\Delta\Psi''(Y)].
\end{aligned}$$

Similarly,

$$\left.\frac{\partial Q(\varepsilon)}{\partial \varepsilon}\right|_{\varepsilon=0} = \frac{1}{e_{33}^F}\begin{pmatrix} 2e_3^G A & \boldsymbol{w} \\ \boldsymbol{w}^t & b_{3,3}^G e_{33}^F - e_3^G c_{33}^F \end{pmatrix},$$

where

$$\begin{aligned}
\boldsymbol{w} &= \begin{pmatrix} b_{1,3}^G e_{33}^F - e_3^G c_{13}^F \\ b_{2,3}^G e_{33}^F - e_3^G c_{23}^F \end{pmatrix}, \\
b_{1,3}^G &= -\frac{1}{\tau^2}E_{G-F}\left[\frac{(Y-\alpha)}{(1+\tau\alpha)}\Delta\Psi(Y)\right], \\
b_{2,3}^G &= -\frac{1}{\tau^2}E_{G-F}\left[\frac{X(Y-\alpha)}{(1+\tau\alpha)}\Delta\Psi(Y)\right], \\
b_{3,3}^G &= E_{G-F}\left[-\frac{2Y}{\tau^3(1+\tau\alpha)}\Delta\Psi(Y) + \left(\frac{2\alpha}{\tau^3(1+\tau\alpha)} - \frac{2\log(1+\tau\alpha)}{\tau^4}\right)\Delta\Psi(Y) + \frac{1}{\tau^4}\Delta\Psi(Y)^2\right].
\end{aligned}$$

and $c_{k3}^F = E_F\left[l_{k3}l_3 + l_k l_{33}\right]$, for $i = 1,2,3$. As the expectations c_{k3}^F depend only on the nominal distribution F, which is fixed in our analyzes, we will not present their detailed forms. Finally,

following (3.8) we get the CVF for the Poisson-Gamma model:

$$\begin{aligned} CVF &= \frac{\partial V(\varepsilon)}{\partial \varepsilon}\Big|_{\varepsilon=0} \\ &= \frac{1}{(e_{33}^F)^3}\begin{pmatrix} 4e_3^G(e_{33}^F)^2\boldsymbol{B}^{-1}\boldsymbol{A}\boldsymbol{B}^{-1} & -e_{33}^F\boldsymbol{B}^{-1}\boldsymbol{w} \\ -e_{33}^F\boldsymbol{w}^T\boldsymbol{B}^{-1} & e_3^G(2e_{333}^F - c_{33}^F) + (b_{3,3}^G - 2a_{33}^G)e_{33}^F \end{pmatrix}. \end{aligned} \tag{3.16}$$

3.3.2 V-Robustness of Parameter Estimates

According to Definition 3.2.2, an estimator is V-robust if its Change-of-Variance Sensitivity (CVS) is finite. In this Section we show that the diagonal entries of the CVF in (3.16) are bounded, which suffices to prove that the MLE of the Poisson-Gamma models are V-robust.

Lemma 3.3.1. *For any $G \in \Lambda$, the quantities*

(i) e_3^G,

(ii) a_{33}^G, *and*

(iii) $b_{3,3}^G$

are all bounded in $G \in \Lambda$.

Proof.

(i) Recall from (3.15) that

$$e_3^G = \frac{1}{\tau^2}E_X[\log(1+\tau\alpha)] - \frac{1}{\tau^2}E_G[\Delta\Psi(Y)]$$

Using $z = 1/\tau$ in (3.13), the law of iterated expectations and (3.12), we get

$$0 \leq E_G[\Delta\Psi(Y)] \leq \tau E_G(Y) = \tau E_X(\alpha), \text{ for all } G \in \Lambda. \tag{3.17}$$

Then

$$\frac{1}{\tau^2}E_X[\log(1+\tau\alpha) - \tau\alpha] \leq e_3^G \leq \frac{1}{\tau^2}E_X[\log(1+\tau\alpha)].$$

(ii) From (3.15)

$$a_{33}^G = -\frac{2}{\tau^3}E_X[\log(1+\tau\alpha)] + \frac{2}{\tau^3}E_G[\Delta\Psi(Y)] + \frac{1}{\tau^4}E_{G-F}[\Delta\Psi'(Y)]. \tag{3.18}$$

Again, using $z = 1/\tau$ in (3.14), the law of iterative expectations and (3.12), we get

$$-\tau^2 E_G[Y] \leq E_G[\Delta\Psi'(Y)] \leq 0. \tag{3.19}$$

Thus, using inequalities (3.17) and (3.19) in (3.18) we obtain

$$\begin{aligned} a_{33}^G &\leq -\frac{2}{\tau^3}E_X[\log(1+\tau\alpha)] + \frac{2}{\tau^2}E_X(\alpha) - \frac{1}{\tau^4}E_F[\Delta\Psi'(Y)], \text{ for all } G \in \Lambda, \\ a_{33}^G &\geq -\frac{2}{\tau^3}E_X[\log(1+\tau\alpha)] - \frac{1}{\tau^4}\left(\tau E_X(\alpha) + E_F[\Delta\Psi'(Y)]\right), \text{ for all } G \in \Lambda. \end{aligned}$$

(iii) Finally,

$$b_{3,3}^G = E_{G-F}\left[-\frac{2Y}{\tau^3(1+\tau\alpha)}\Delta\Psi(Y) + \left(\frac{2\alpha\tau - 2(1+\tau\alpha)\log(1+\tau\alpha)}{\tau^4(1+\tau\alpha)}\right)\Delta\Psi(Y) + \frac{1}{\tau^4}\Delta\Psi(Y)^2\right].$$

If $(1+\tau\alpha)\log(1+\tau\alpha) \leq \alpha\tau$, using (3.12), (3.13) and (3.17) we can show that

$$\begin{aligned} b_{3,3}^G &\leq E_F\left[\frac{2Y\Delta\Psi(Y)}{\tau^3(1+\tau\alpha)} - \left(\frac{2\alpha}{\tau^3(1+\tau\alpha)} - \frac{2\log(1+\tau\alpha)}{\tau^4}\right)\Delta\Psi(Y) - \frac{1}{\tau^4}\Delta\Psi(Y)^2\right] \\ &\quad + E_X\left[\frac{2\alpha^2}{\tau^2(1+\tau\alpha)} - \frac{2\alpha\log(1+\tau\alpha)}{\tau^3}\right] + \frac{1}{\tau^2}E_X[\alpha^2(1+\tau)+\alpha], \\ b_{3,3}^G &\geq E_F\left[\frac{2Y\Delta\Psi(Y)}{\tau^3(1+\tau\alpha)} - \left(\frac{2\alpha}{\tau^3(1+\tau\alpha)} - \frac{2\log(1+\tau\alpha)}{\tau^4}\right)\Delta\Psi(Y) - \frac{1}{\tau^4}\Delta\Psi(Y)^2\right] \\ &\quad - \frac{2}{\tau^2(1+\tau\alpha)}[\alpha + (1+\tau)\alpha^2]. \end{aligned}$$

Similarly, if $\alpha\tau < (1+\tau\alpha)\log(1+\tau\alpha)$, then

$$\begin{aligned} b_{3,3}^G &< E_F\left[\frac{2Y\Delta\Psi(Y)}{\tau^3(1+\tau\alpha)} - \left(\frac{2\alpha}{\tau^3(1+\tau\alpha)} - \frac{2\log(1+\tau\alpha)}{\tau^4}\right)\Delta\Psi(Y) - \frac{1}{\tau^4}\Delta\Psi(Y)^2\right] \\ &\quad + \frac{1}{\tau^2}E_X[\alpha^2(1+\tau)+\alpha], \\ b_{3,3}^G &> E_F\left[\frac{2Y\Delta\Psi(Y)}{\tau^3(1+\tau\alpha)} - \left(\frac{2\alpha}{\tau^3(1+\tau\alpha)} - \frac{2\log(1+\tau\alpha)}{\tau^4}\right)\Delta\Psi(Y) - \frac{1}{\tau^4}\Delta\Psi(Y)^2\right] \\ &\quad E_X\left[\frac{2\alpha^2}{\tau^2(1+\tau\alpha)} - \frac{2\alpha\log(1+\tau\alpha)}{\tau^3}\right] - \frac{2}{\tau^2(1+\tau\alpha)}[\alpha + (1+\tau)\alpha^2]. \end{aligned}$$

□

Proposition 3.3.1. *The MLE of $\boldsymbol{\gamma} = (\beta_0, \beta_1, \tau)^T$ for the Poisson-Gamma Model is V-robust.*

Proof. Note that the diagonal entries of the CVF (3.16) depend on the contaminating distribution G only through the quantities of the previous Corollary. Moreover, $V(\psi, F)$ is the asymptotic variance under the nominal model and therefore it does not depend on G. Thus, using the results of Corollary 3.3.1 and according to Definition 3.2.2, we prove that the MLE of the Poisson-Gamma parameters is V-robust. □

3.4 Mixed-Effects Binomial Models

In many applications, one needs to study the relationship between binomial responses and several explanatory variables. The response may also be a vector of binary responses per experimental unit or cluster. If the data are grouped as frequencies for each cluster, the response variable can be modelled by the binomial distribution. For example, in teratologic applications, pregnant animals are exposed to a pharmaceutical substance and they are sacrificed prior to the birth of the litter (Heagerty and Zeger, 2000). The fetuses of each litter are then examine to determine the presence or absence of a malformation. The response variable records this information for each fetus per litter (binary responses) or the number of fetuses per litter affected by the drug (binomial response).

As in the case of the Poisson models, binary or binomial data often exhibit overdispersion with respect to the nominal variance. One possible explanation for the overdispersion is that in general, there exists intracluster dependence. In other words, observations from the same individual or cluster tend to be more similar than observations from different subjects. Many models have been proposed to model clustered binary data, including the use of GLM with random effects (e.g., Stiratelli et al., 1984; Neuhaus et al., 1992; Prentice, 1988; Heagerty and Zeger, 2000). The Beta-Binomial distribution is sometimes used to model binomial data with extra variation (e.g., Crowder, 1978; Williams, 1982; McCullagh and Nelder, 1989).

In this Section we review alternative models that have been proposed in the literature to study binomial data. In particular, we examine two simple models that have an attractive marginal closed form density and hence maximum likelihood procedures are used to estimate the model

parameters. Finally, the effect of a contamination in the mixing distribution on the asymptotic variance-covariance matrix is investigated using the CVF.

3.4.1 General Mixed-Effect Binomial Models

- For $i = 1, \ldots, n$, let $\boldsymbol{X}_i$ be a p-dimensional vector of covariates independent of U_i, a random intercept.
- Assume further that $\boldsymbol{X}_i$ are independent and identically distributed for $i = 1, \ldots, n$.
- Similarly, the random effects $U_i,\ i = 1, \ldots, n$ are independent and identically distributed.
- Conditionally, given that $\boldsymbol{X}_i = \boldsymbol{x}_i$ and $U_i = u_i$, for each cluster i, we observe n_i binary responses $Y_{ij},\ j = 1, \ldots, n_i$. Let $Y_i = \sum_{j=1}^{n_i} Y_{ij}$. Then, $Y_i|\boldsymbol{x}_i, u_i$ is a binomial random variable with mean $n_i p_i$, where $g(p_i) = \nu_i + \boldsymbol{x}_i\boldsymbol{\beta}$ and $\nu_i = \nu(u_i)$.

There are several link functions g commonly used in the literature:

- The logit link function $g(\mu) = \log(\mu/(1-\mu))$.
- The identity link function $g(\mu) = \mu$.
- The probit link function $g(\mu) = \Phi^{-1}(\mu)$, where Φ is the cumulative distribution function of a standard normal random variable.

A common approach to estimate the parameters of a mixed-effects binomial model is using maximum likelihood methods (Neuhaus et al., 1992; Heagerty and Zeger, 2000; Neuhaus, 2001). The main difficulty of this method is that to obtain the likelihood, one must solve a set of integrals that are typically intractable. The marginal likelihood function is given by

$$\prod_{i=1}^{n} \int f(y_i|\boldsymbol{x}_i, u_i) dG(u_i) = \prod_{i=1}^{n} \int \begin{pmatrix} n_i \\ \\ y_i \end{pmatrix} p_i^{y_i} (1-p_i)^{n_i - y_i} dG(u_i). \tag{3.20}$$

In general, there is no closed form for the marginal likelihood (3.20). Approaches to overcome this difficulty include numerical integration (Neuhaus, 2001), approximate solutions (Stiratelli et al.,

1984; Neuhaus et al. 1992) and Monte Carlo EM algorithms (McCulloch, 1997). Two exceptions that we are going to examine later are:

(1) A model with $n_i = 1,$ for $i = 1, \ldots, n$, g the identity link function and a random intercept $U_i \sim Beta(\alpha_1, \alpha_2)$.

(2) A model with only a random intercept having Beta distribution, i.e. $U_i \sim Beta(\alpha_1, \alpha_2)$ and $\boldsymbol{\beta} = \mathbf{0}$.

The model described in item (2) is known as the Beta-Binomial model and was extensively used to model overdispersion of binomial data (e.g., Crowder, 1978; Williams, 1982; McCullagh and Nelder, 1989). A disadvantage of this model is that it does not include the relation with other explanatory variables or fixed effects. Lee and Nelder (1996) consider an alternative approach to incorporate fixed effects in this model. They maximize the hierarchical likelihood (logarithm of the joint density function) to obtain the Maximum Hierarchical Likelihood Estimates (MHLEs). Note however, that a contamination in the mixing distribution will not affect these estimators. Therefore, we are not going to study this approach. Another commonly used model consists of assuming that conditionally on the random effects $U_i = u_i$, the response variable Y_i has a binomial distribution with mean $n_i u_i$ and the random effects have a Beta distribution. Moreover, a p-dimensional vector of covariates $\boldsymbol{X}_i$ is incorporated to the model using the relation $g(E(U_i)) = \boldsymbol{X}_i\boldsymbol{\beta}$ (Williams, 1982; Kuppert et al., 1986). However, this is not a mixed-effects model so we are not going to cover its analysis in this Chapter.

A major disadvantage of maximum likelihood estimation is that it requires full specification of the mixing distribution. Neuhaus et al. studied the effect of a misspecification of the mixing distribution on the parameter estimates. The effects on the estimated standard errors were examined by a simulation study. We derive the CVF of MLE for Models (1) and (2) to study the stability of the estimated standard errors under a slight contamination of the mixing distribution.

3.4.2 V-Robustness of Parameter Estimates

Model (1)

Let's consider first a linear probability model with one observation per cluster, i.e., $n_i = 1$ and $g(\mu) = \mu$. The main disadvantage of this model is that μ is restricted to the interval $[0, 1]$, thus imposing a restriction on the parameters $\boldsymbol{\beta}$.

It is easy to show that this model has marginal density given by

$$f(y_i|\boldsymbol{x}_i) = \left(\frac{\alpha_1}{\alpha_1+\alpha_2} + \boldsymbol{x}_i\boldsymbol{\beta}\right)^{y_i} + \left(\frac{\alpha_2}{\alpha_1+\alpha_2} - \boldsymbol{x}_i\boldsymbol{\beta}\right)^{1-y_i}, \text{ for } y_i = 0, 1.$$

Then the log-likelihood is a linear function of the responses y_i:

$$l = l(\boldsymbol{\gamma}; \boldsymbol{x}_i, y_i) = \log\left(\frac{\alpha_2}{\alpha_1+\alpha_2} - \boldsymbol{x}_i\boldsymbol{\beta}\right) + y_i\log\left(\frac{\alpha_1 + (\alpha_1+\alpha_2)\boldsymbol{x}_i\boldsymbol{\beta}}{\alpha_2 - (\alpha_1+\alpha_2)\boldsymbol{x}_i\boldsymbol{\beta}}\right),$$

where $\boldsymbol{\gamma} = (\alpha_1, \alpha_2, \boldsymbol{\beta}^T)^T \in \mathbb{R}^{p+2}$.

Remark 3.4.1. *Note that for any quadratic function $q(.)$, $E_F[q(Y)] = E_G[q(Y)]$ for any $G \in \Lambda$. Thus, $E_{G-F}[l_j] = E_{G-F}[l_j l_k] = E_{G-F}[l_{jk}] = 0$, where $l_j = \partial l/\partial\gamma_j$, $l_{jk} = \partial^2 l/\partial\gamma_j\partial\gamma_k$ and $j, k = 1, \ldots, p+2$.*

Gustafson (1996) defined the local effect of the mixing distribution misspecification on the estimators $\hat{\boldsymbol{\gamma}}$ by

$$\boldsymbol{\gamma}_j'(0) = \int \{I^{-1}(\boldsymbol{\gamma}_0)s_j(\boldsymbol{\gamma}_0; u)\}d(G-F)(u)$$

where $I^{-1}(\boldsymbol{\gamma}_0)$ is the inverse of the Fisher information matrix and $s_j(\boldsymbol{\gamma}_0; u) = E[l_j|u]$ is the conditional score for $j = 1, \ldots, p+2$. An immediate consequence of Remark 3.4.1 is that $\boldsymbol{\gamma}_j'(0) = 0$, for all $G \in \Lambda$. Gustafson (1996) called these estimators first-order consistent.

The results of Remark 3.4.1 can also be used to analyze the CVF of the MLE for Model (1). Following (3.8)-(3.10) we can derive the CVF and note that the CVF(MLE,G) does not depend on G, for any $G \in \Lambda$. Therefore, the (marginal) maximum likelihood estimators are V-robust according to Definition 3.2.2.

Model (2)

The Beta-Binomial Model is more useful from a practical point of view. Its marginal density is given by

$$f(y_i) = \begin{pmatrix} n_i \\ \\ y_i \end{pmatrix} \frac{\Gamma(\alpha_1+\alpha_2)\Gamma(y_i+\alpha_1)\Gamma(n_i+\alpha_2-y_i)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_1+\alpha_2+n_i)},$$

the conditional first two moments of the response are given by

$$E[Y_i|u_i] = n_i u_i, \quad V[Y_i|u_i] = n_i u_i(1-u_i),$$

and the unconditional moments are

$$E[Y_i] = n_i\pi, \quad V[Y_i] = n_i\pi(1-\pi)[1+\delta(n_i-1)]. \tag{3.21}$$

where $\pi = \alpha_1/(\alpha_1+\alpha_2)$, and $\delta = 1/(\alpha_1+\alpha_2+1)$. From equation (3.21), we can see how the extra variation is added to the model.

For simplicity, we assume that $n_i = m$ for $i = 1, \ldots, n$ and the subindex i is omitted in the following results.

Notation 3.4.1. *For $k = 1, 2$, let $\Delta\Psi_k(u) = \Delta\Psi(u; \alpha_k)$ defined in Notation (3.3.2). Similarly define $\Delta\Psi_k'(u)$.*

The log-likelihood function, except for a constant that does not depend on the unknown parameters, is given by

$$l = \log\Gamma(\alpha_1+\alpha_2)+\log\Gamma(y+\alpha_1)+\log\Gamma(m+\alpha_2-y)-\log\Gamma(\alpha_1)-\log\Gamma(\alpha_2)-\log\Gamma(\alpha_1+\alpha_2+m). \tag{3.22}$$

Using (3.8)-(3.11), we can derive the CVF for $\hat{\boldsymbol{\gamma}}$, the (marginal) MLE of $\boldsymbol{\gamma} = (\alpha_1, \alpha_2)$, when the mixing distribution F is contaminated by a distribution G. Note that this matrix depends on the contaminating function G only through $E_G[l_r]$, $E_G[l_{rj}]$, $E_G[l_r l_j]$, for $r, j = 1, 2$. Moreover, the partial derivatives l_r and l_{rj} are up to a constant equal to $\Delta\Psi_k(y)$ and $\Delta\Psi_k'(m-y)$, respectively. Thus, the CVF depends on G only through the expectations analyzed in the following Lemma:

Lemma 3.4.1. *Let U be a random variable having a beta distribution with parameters (α_1, α_2). Given $U = u$, let Y be a binomial random variable with mean m times u. Let $\Psi_k(.)$ and $\Psi_k'(.)$ be the polygamma functions defined in Notation (3.4.1). Then, for any $G \in \Lambda$, the expectations*

(i) $E_G[\Delta\Psi_1(Y)]$,

(ii) $E_G[\Delta\Psi_2(m-Y)]$,

(iii) $E_G[(\Delta\Psi_1(Y))^2]$,

(iv) $E_G[(\Delta\Psi_2(m-Y))^2]$,

(v) $E_G[(\Delta\Psi_1(Y))(\Delta\Psi_2(m-Y))]$,

(vi) $E_G[\Delta\Psi_1'(Y)]$ *and*

(vii) $E_G[\Delta\Psi_2'(m-Y)]$

are all bounded.

Proof. We first compute the first two absolute moments of Y that are going to be used throughout this proof. Using conditional expectations,

$$\begin{aligned} E_G[Y] &= E_G[E[Y|U]] = E_G[mU] = m\frac{\alpha_1}{\alpha_1+\alpha_2}, \text{and} \\ E_G[Y^2] &= E_G[E[Y^2|U]] = E_G[mU(1-U)+m^2U^2] = m(m-1)E_G[U^2]+mE_G[U] \\ &= \frac{m(m-1)\alpha_1(\alpha_1+1)}{(\alpha_1+\alpha_2)(\alpha_1+\alpha_2+1)}+\frac{m\alpha_1}{\alpha_1+\alpha_2}. \end{aligned} \tag{3.23}$$

(i) Replacing $z=\alpha_1$ in (3.13), taking iterated expectations and using (3.23) we obtain

$$0 \le E_G[\Delta\Psi_1(Y)] \le \frac{1}{\alpha_1}E_G[Y] = \frac{m}{\alpha_1+\alpha_2}. \tag{3.24}$$

Hence, $E_G[\Delta\Psi_1(Y)]$ is bounded for all $G\in\Lambda$.

(ii) The proof is almost identical to that in (i) once we note that conditionally, given that $U=u$, the random variable $W=m-Y$ is a binomial random variable with mean $m(1-u)$. Then, replacing α_1 with α_2 and Y with W in (3.24) we get

$$0 \le E_G[\Delta\Psi_2(W)] \le \frac{m}{\alpha_1+\alpha_2}, \quad \text{for all} \quad G\in\Lambda.$$

(iii) From (3.13) we can prove that

$$0 \leq E_G[\Delta\Psi_1^2(Y)] \leq \frac{1}{\alpha_1^2} E_G[Y^2]. \tag{3.25}$$

for all $G \in \Lambda$. Then, using (3.23) we obtain

$$0 \leq E_G[\Delta\Psi_1^2(Y)] \leq \frac{m}{\alpha_1}\left[\frac{(m-1)(\alpha_1+1)}{(\alpha_1+\alpha_2)(\alpha_1+\alpha_2+1)} + \frac{1}{(\alpha_1+\alpha_2)}\right].$$

(iv) Again replacing Y with $W = m - Y$ in (3.25) and α_1 with α_2, we obtain

$$0 \leq E_G[\Delta\Psi_1^2(W)] \leq \frac{m}{\alpha_2^2}\left[\frac{(m-1)\alpha_1(\alpha_1+1)}{(\alpha_1+\alpha_2)(\alpha_1+\alpha_2+1)} + \frac{(1-2m)\alpha_1}{(\alpha_1+\alpha_2)} + m\right]$$

for all $G \in \Lambda$.

(v) As $\Delta\Psi_1(Y)$ and $\Delta\Psi_2(m-Y)$ are both nonnegative functions, and using the Cauchy-Schwarz inequality, we get

$$0 \leq E_G[\Delta\Psi_1(Y)\Delta\Psi_2(m-Y)] \leq E_G^{1/2}[\Delta\Psi_1(Y)]E_G^{1/2}[\Delta\Psi_2(m-Y)]$$

Then, by (iii) and (iv), $E_G[\Delta\Psi_1(Y)\Delta\Psi_2(m-Y)]$ is also bounded for all $G \in \Lambda$.

(vi) Replacing z with α_1 in (3.14), taking iterated expectations and using (3.23) we can show that

$$0 \geq E_G[\Delta\Psi_1'(Y)] \geq -\frac{1}{\alpha_1^2} E_G[Y] = -\frac{m}{\alpha_1(\alpha_1+\alpha_2)} \tag{3.26}$$

Hence, $E_G[\Delta\Psi_1'(Y)]$ is bounded for all $G \in \Lambda$.

(vii) Similarly, replacing α_1 with α_2 and Y with $W = m - Y$ in (3.26) we obtain

$$0 \geq E_G[\Delta\Psi_2'(W)] \geq -\frac{m}{\alpha_2(\alpha_1+\alpha_2)}$$

for all $G \in \Lambda$. □

Proposition 3.4.1. *The MLE of* $\boldsymbol{\gamma} = (\alpha_1, \alpha_2)^T$ *for the Beta-Binomial Model is* V*-robust.*

Proof. Note that the entries of the CVF (3.8) depend on the contaminating distribution G only through the expectations analyzed in Lemma 3.4.1. Thus, using the results of this lemma and according to Definition 3.2.2, we prove that the (marginal) MLE of the Poisson-Gamma parameters is V-robust. □

Remark 3.4.2. *Gustafson (1996) found the exact minimum and maximum first order bias of the estimators but only for* $m \leq 5$*. Using results (i) and (ii) in Lemma 3.4.1, we can get bounds for any value of* m*. As these bounds are finite, we can conclude that the estimators are* B*-robust for any value of* m*.*

3.5 Simulation

A simulation study was performed in order to assess the magnitude of the change in the variance of the estimators when the mixing distribution is contaminated in the Poisson-Gamma Model. The performance of the estimators was investigated in samples generated by S-Plus Version 6.2.1 for different choices of population parameters and different types of contaminations. The MLE of these parameters was obtained using a modified version of the `glm.nb` function available in the MASS library (Venables and Ripley, 1999). The modification of the `glm.nb` function corresponds to a reparametrization of the log-likelihood in order to estimate the variance of the gamma distribution.

3.5.1 The Poisson-Gamma model

We start examining the Poisson-Gamma model without any contamination in the mixing distribution. More precisely, we generate 1000 covariates X_i from a standard normal distribution and 1000 random effects U_i from a gamma distribution with $E(U_i) = 1$ and $V(U_i) = \tau$, for $i = 1, \ldots, 1000$. Conditionally on (x_i, u_i), a sample of 1000 random variables Y_i is generated from a Poisson distribution with $E(Y_i|x_i, u_i) = u_i \exp\{\beta_0 + \beta_1 x_i\}$ (see Section (3.3.1) for details of the model).

Different choices of $\boldsymbol{\gamma} = (\beta_0, \beta_1, \tau)^T$ are considered in order to analyze later the effect of the contamination on distributions with different characteristics. The vector of parameters

$(\beta_0, \beta_1)^T$ determines the shape and location of $\alpha = \exp(\beta_0 + \beta_1 x)$. Three different choices are used in the simulation study for this vector that describe three general positions of the curve α: $\{(0,1), (-2,1), (2,1)\}$. The Gamma distribution depends on τ. For $0 < \tau < 1$, the density has a mode at $y = 1 - \tau$ and is positively skewed. For $\tau > 1$, it decreases monotonically. For $\tau = 1$ the exponential distribution is obtained as a special case. Therefore, we consider the following set of parameter values for τ: $\{.25, .5, 1, 1.5, 2\}$.

For each choice of parameters, 1000 Monte Carlo replications of the random effects $\{U_i\}$ and the responses $\{Y_i\}$ were generated. The same sample of $\{X_i\}$ was used in all replications. The MLE of $\boldsymbol{\gamma}$ was computed for each of these random samples and the results were used to obtain estimates of the mean and variance of the estimators, which are summarized in Table 3.1 below.

When the model is correctly specified there is a small bias in the estimates of the population parameters. In almost all cases the bias is of the order of 10^{-3}, with the exception of the bias of τ when the true population parameter is $\boldsymbol{\gamma} = (2, 1, \tau)^T$, where the bias is of order 10^{-2} for all choices of τ. The estimated variance of the estimate of τ increases in all cases as the true parameter τ increases.

Table 3.1: Means and variances of MLE of Poisson-Gamma model.

True parameters	MC mean			MC variance		
$\beta_0 = 0$, $\beta_1 = 1$	β_0	β_1	τ	β_0	β_1	τ
tau=.25	-0.001778	1.000549	0.249692	0.001527	0.001263	0.001769
τ=.5	-0.003705	1.001110	0.500010	0.001915	0.001785	0.003416
τ=1	-0.003886	0.997852	0.997655	0.002604	0.002574	0.008107
τ=1.5	-0.002655	0.998346	1.491551	0.003081	0.003264	0.015580
τ=2	-0.003281	1.000541	1.999453	0.003572	0.003799	0.028664
$\beta_0 = 2$, $\beta_1 = 1$	β_0	β_1	τ	β_0	β_1	τ
τ=.25	1.998179	1.000251	0.249354	0.000427	0.000449	0.000301
τ=.5	1.999330	0.998404	0.499168	0.000709	0.000784	0.000839
τ=1	1.997411	1.001111	0.998728	0.001295	0.001311	0.002738
τ=1.5	1.996011	0.999635	1.496918	0.001638	0.001849	0.006091
τ=2	1.998376	1.001075	1.993784	0.002238	0.002227	0.010421
$\beta_0 = -2$, $\beta_1 = 1$	β_0	β_1	τ	β_0	β_1	τ
τ=.25	-2.005116	0.999984	0.239731	0.009631	0.005439	0.026945
τ=.5	-1.999489	0.996300	0.477674	0.010382	0.006407	0.047147
τ=1	-2.015171	1.005591	0.988291	0.010992	0.008519	0.083042
τ=1.5	-2.005543	0.998886	1.480759	0.012495	0.009616	0.133988
τ=2	-2.008588	0.999845	1.973542	0.013128	0.009952	0.198262

3.5.2 The contaminated model

In this Section we examine the behavior of the estimators and their estimated variances under two contaminated models with different tail behaviors. The contaminating distribution considered are the lognormal and the scaled F-distribution (cF, where c is a positive constant and F is a random variable having an F-distribution). The random effects are now generated from a mixed

distribution given by

$$F_\varepsilon = (1-\varepsilon)G + \varepsilon L$$

where G represents the Gamma distribution and L the contaminating distribution, both with expectation equal to 1 and and variance equal to τ.

Various choices of the τ parameter are considered for the lognormal contaminating distribution and results are illustrated in Figures 3.1-3.18. In the case of scaled F, the restriction on the first two moments imposes a constraint on the degrees of freedom and the constant c. Because smaller degrees of freedom of the denominator means heavier tails of the resulting distribution, we choose this parameter to be 6. Both c and the degrees of freedom of the numerator are now completely determined by τ. As we want the degrees of freedom to be an integer, this restricts the choices of τ. For this reason only one set of parameters is used in the simulation for this contaminating distribution.

For the case of a lognormal contamination, Figures 3.1, 3.3 and 3.5 show that the bias of $\hat{\tau}$ increases as the level of contamination increases. For true parameters $\beta_0 = 0$ and $\beta_0 = 2$, this bias also increases at each level of contamination, with the value of the true variance (see Figures 3.3 and 3.5). Figures 3.2, 3.4 and 3.6 show that the estimated variance of this estimate remains almost constant throughout all levels of contamination for all choices of the true $\boldsymbol{\gamma}$. Moreover, for all levels of contamination, the magnitude of the estimated variance is larger for larger values of the true variance.

A similar behavior is found in the estimated bias and variance of the parameter β_0 as can be seen in Figures 3.7-3.12. However, note that the values of the bias of the estimate of β_0 are much smaller than those of τ.

The behavior of the bias of $\hat{\beta}_1$ is not monotonic as can be seen in Figures 3.13, 3.15 and 3.17. Considering that the values of this bias are of the order of 10^{-3}, we interpret these as pure noise from the simulation. The estimated variance follows the same pattern as that of previous parameters (see Figures 3.14, 3.16 and 3.18).

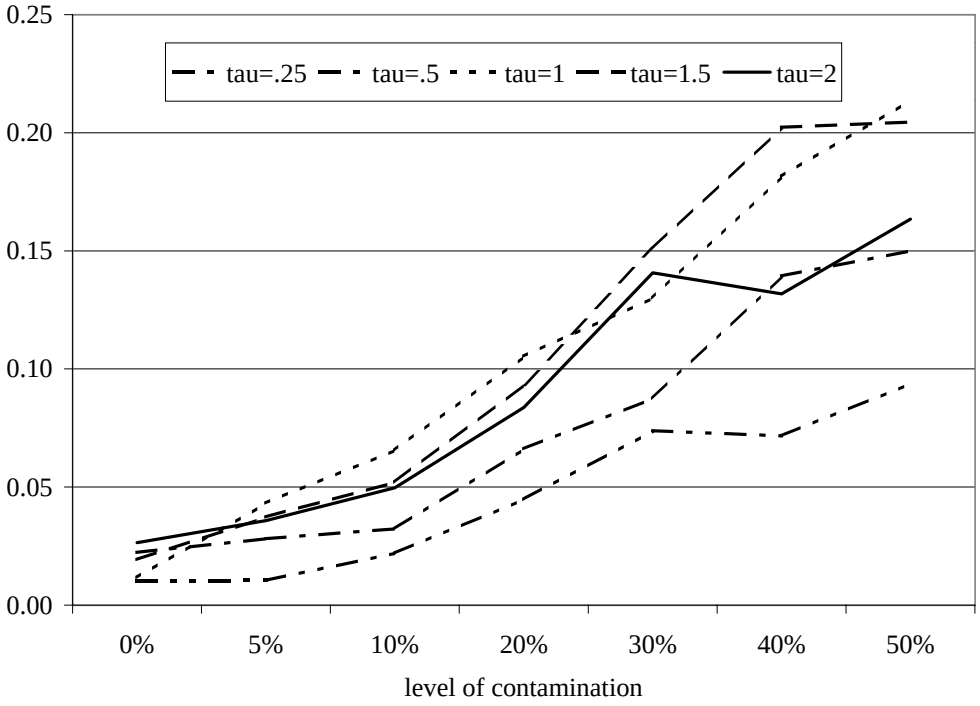

Figure 3.1: Bias in the estimation of τ for $\beta_0 = -2,\ \beta_1 = 1$, and different values of τ under different levels of a lognormal contamination.

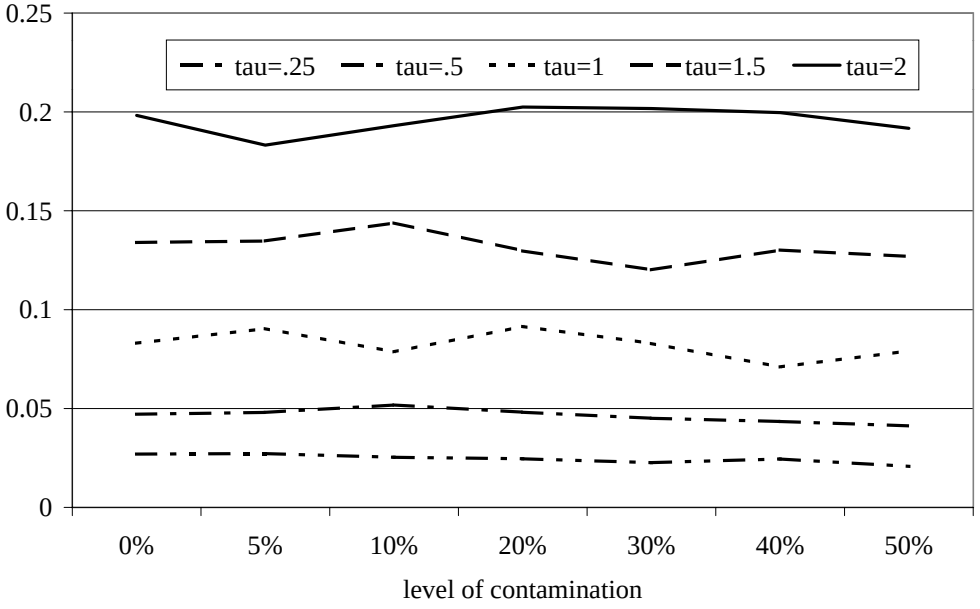

Figure 3.2: Variance of estimate of τ for $\beta_0 = -2,\ \beta_1 = 1$, and different values of τ under different levels of a lognormal contamination.

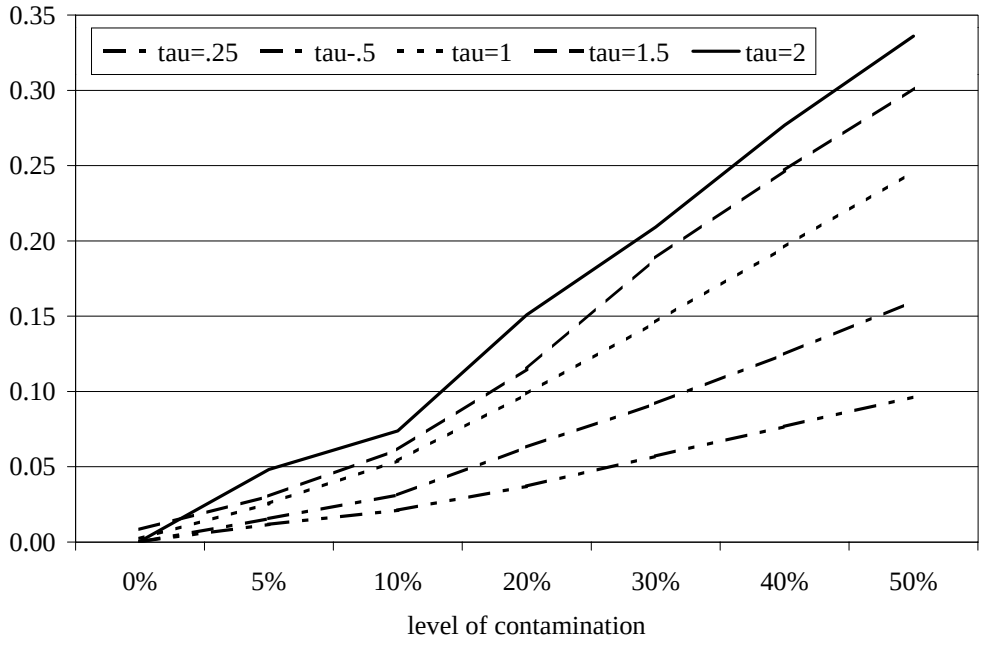

Figure 3.3: Bias in the estimation of τ for $\beta_0 = 0,\ \beta_1 = 1$, and different values of τ under different levels of a lognormal contamination.

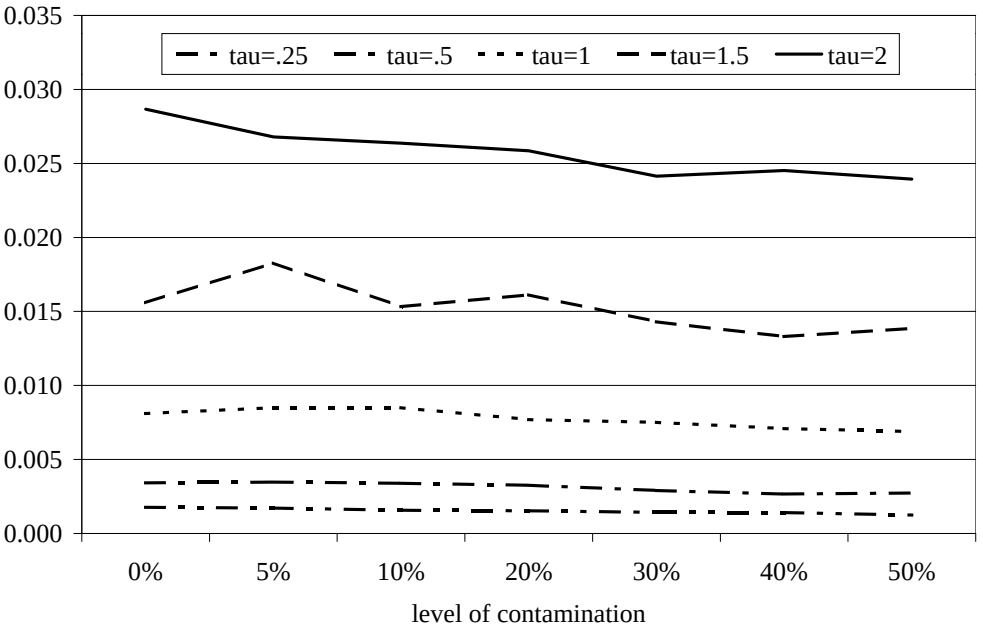

Figure 3.4: Variance of estimate of τ for $\beta_0 = 0,\ \beta_1 = 1$, and different values of τ under different levels of a lognormal contamination.

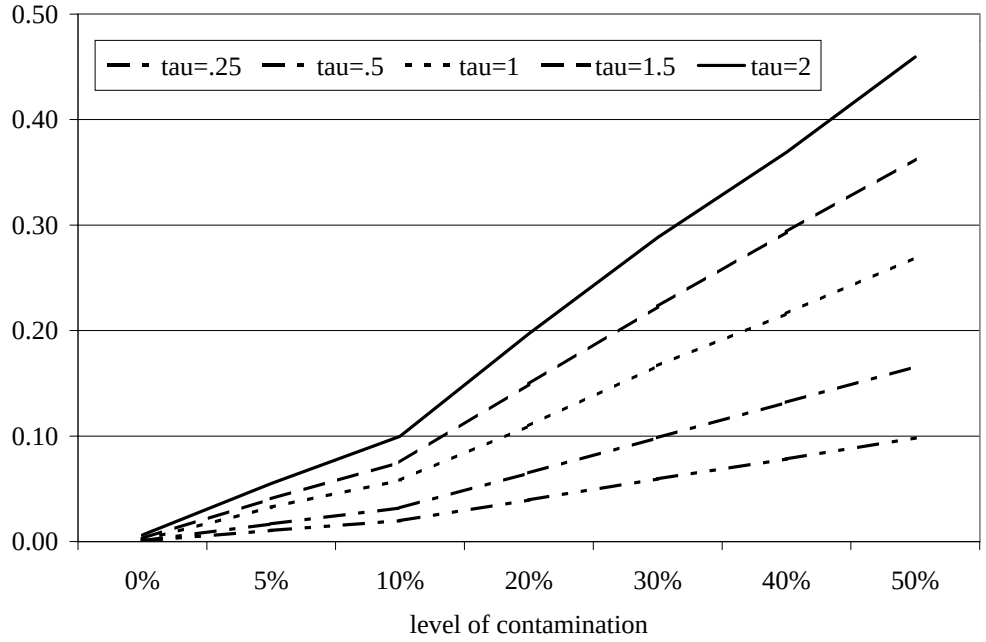

Figure 3.5: Bias in the estimation of τ for $\beta_0 = 2,\ \beta_1 = 1$, and different values of τ under different levels of a lognormal contamination.

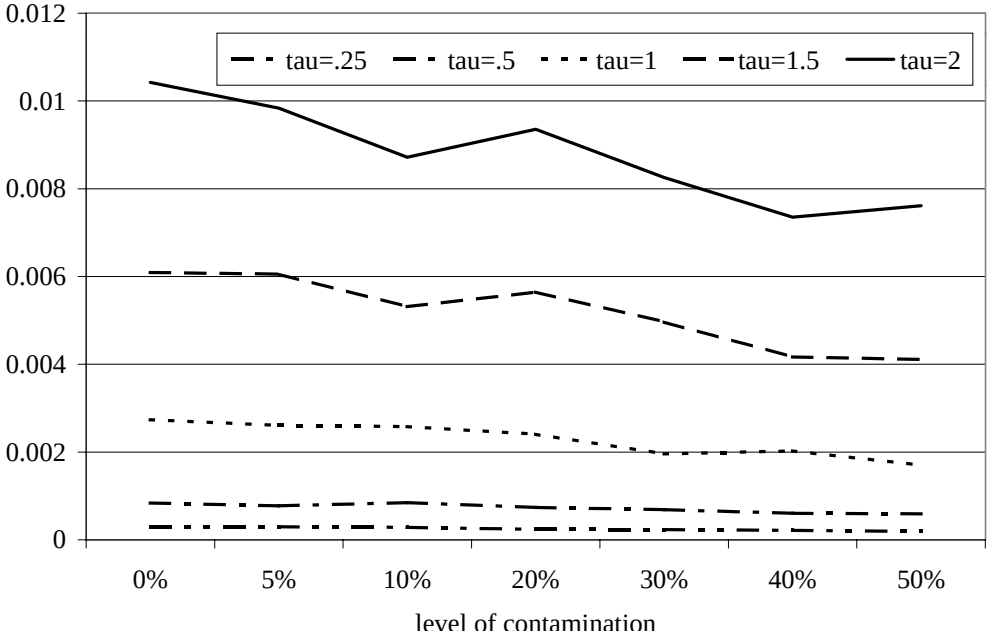

Figure 3.6: Variance of estimate of τ for $\beta_0 = 2,\ \beta_1 = 1$, and different values of τ under different levels of a lognormal contamination.

It is also interesting to note that for each choice of τ, the variances of all estimates decrease as β_0 moves from -2 (for which the curve α is flatter around 0) to 2 (for which the curve α is steeper around 0).

Similar results are found in the estimates of the parameters and its variances under a scaled F contamination. Figures 3.19, 3.21 and 3.23 show the bias of the estimates under different levels of contamination. As before, the bias of $\hat{\beta}_0$ and $\hat{\tau}$ increases with ε while that of $\hat{\beta}_1$ does not follow a monotonic behavior. Moreover, under this contamination, the magnitude of the bias of $\hat{\beta}_0$ and $\hat{\tau}$ is larger than for the lognormal contamination.

It is important to note that the variances of all the estimators, under different parameter choices and under both contaminated functions, remain bounded as it was proved theoretically in Section 3.3.1.

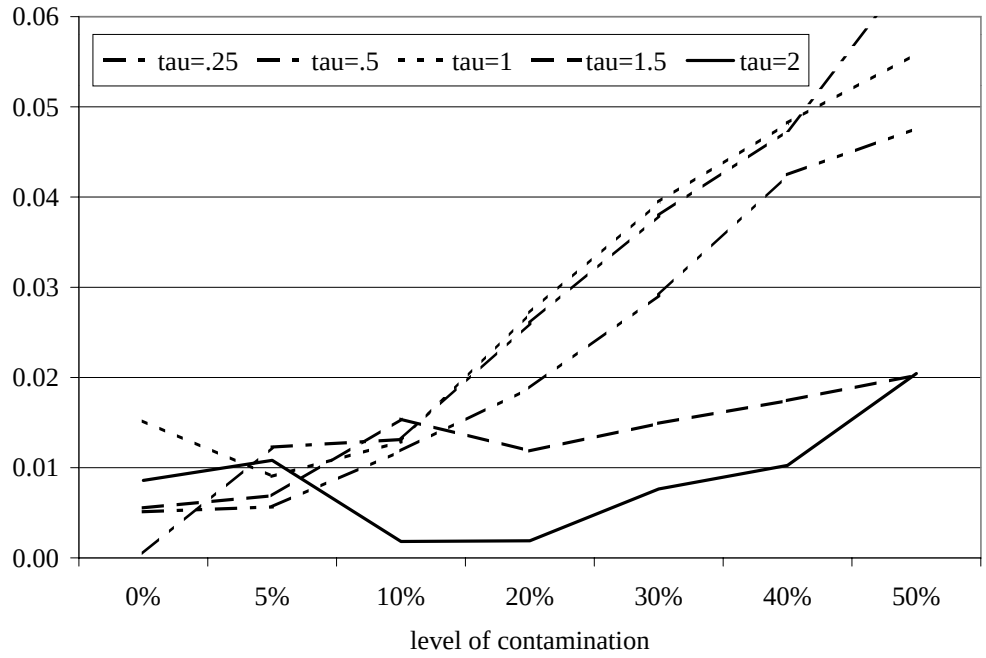

Figure 3.7: Bias in the estimation of β_0 for $\beta_0 = -2,\ \beta_1 = 1$, and different values of τ under different levels of a lognormal contamination.

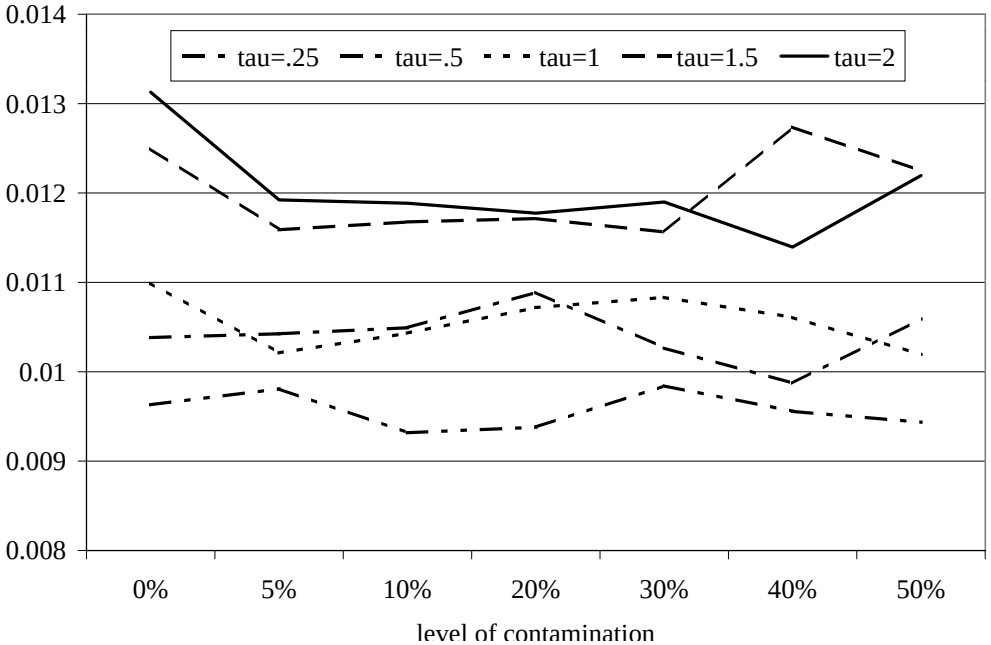

Figure 3.8: Variance of estimate of β_0 for $\beta_0 = -2,\ \beta_1 = 1$, and different values of τ under different levels of a lognormal contamination.

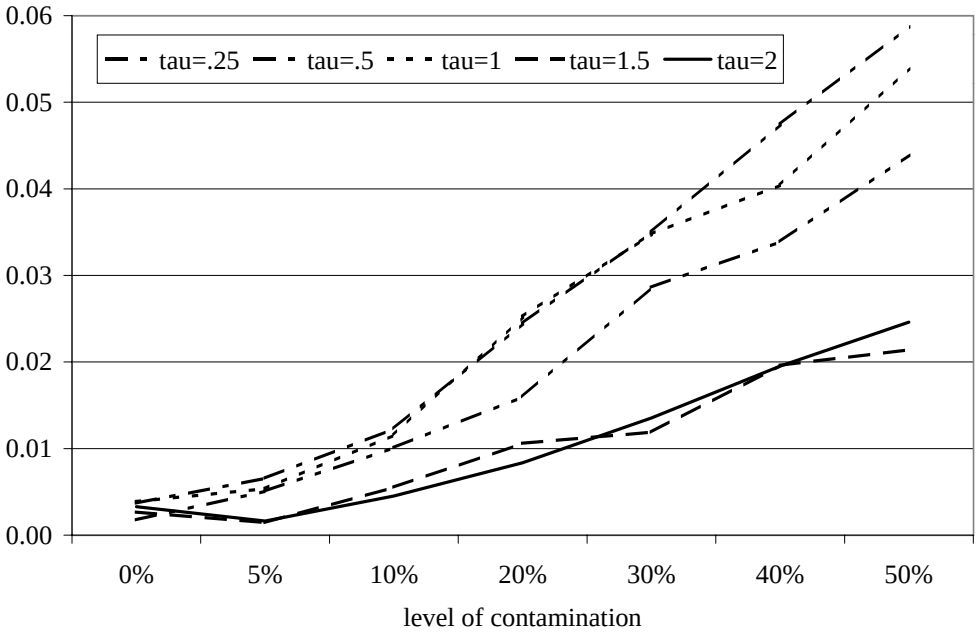

Figure 3.9: Bias in the estimation of β_0 for $\beta_0 = 0,\ \beta_1 = 1$, and different values of τ under different levels of a lognormal contamination.

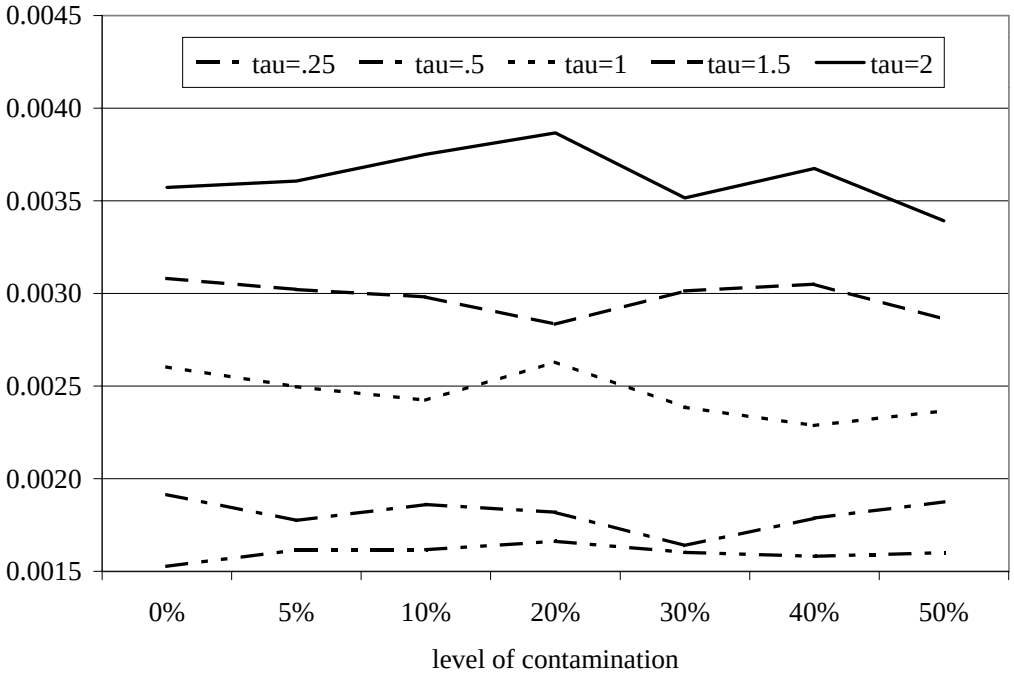

Figure 3.10: Variance of estimate of β_0 for $\beta_0 = 0,\ \beta_1 = 1$, and different values of τ under different levels of a lognormal contamination.

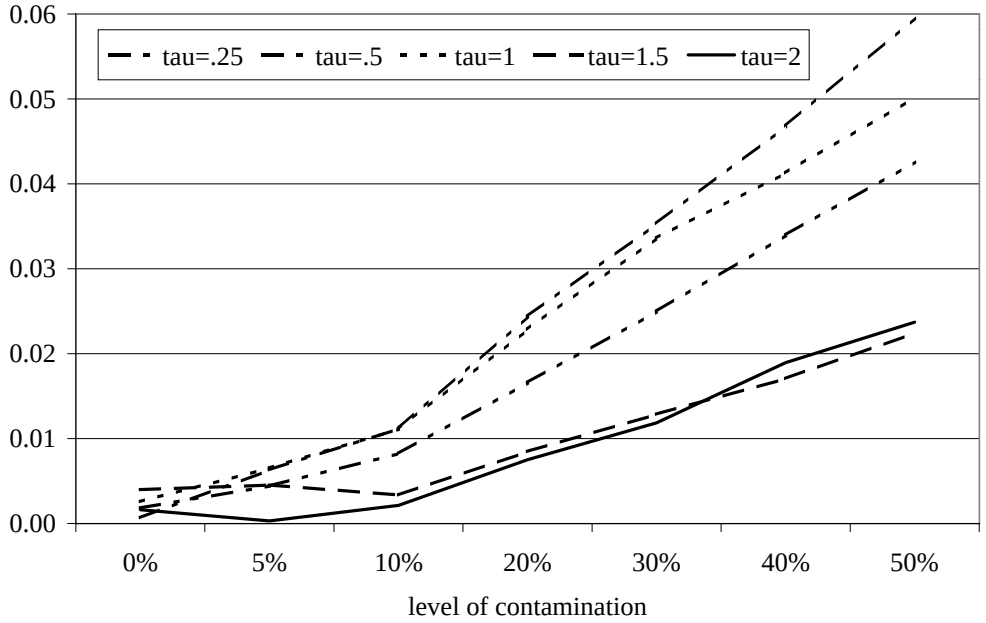

Figure 3.11: Bias in the estimation of β_0 for $\beta_0 = 2,\ \beta_1 = 1$, and different values of τ under different levels of a lognormal contamination.

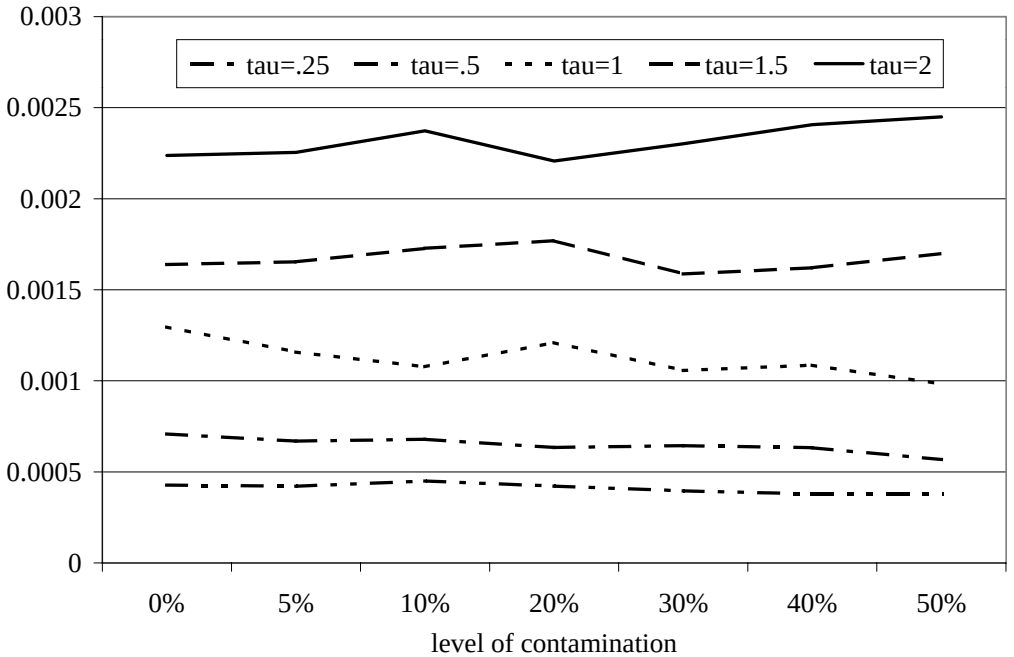

Figure 3.12: Variance of estimate of β_0 for $\beta_0 = 2,\ \beta_1 = 1$, and different values of τ under different levels of a lognormal contamination.

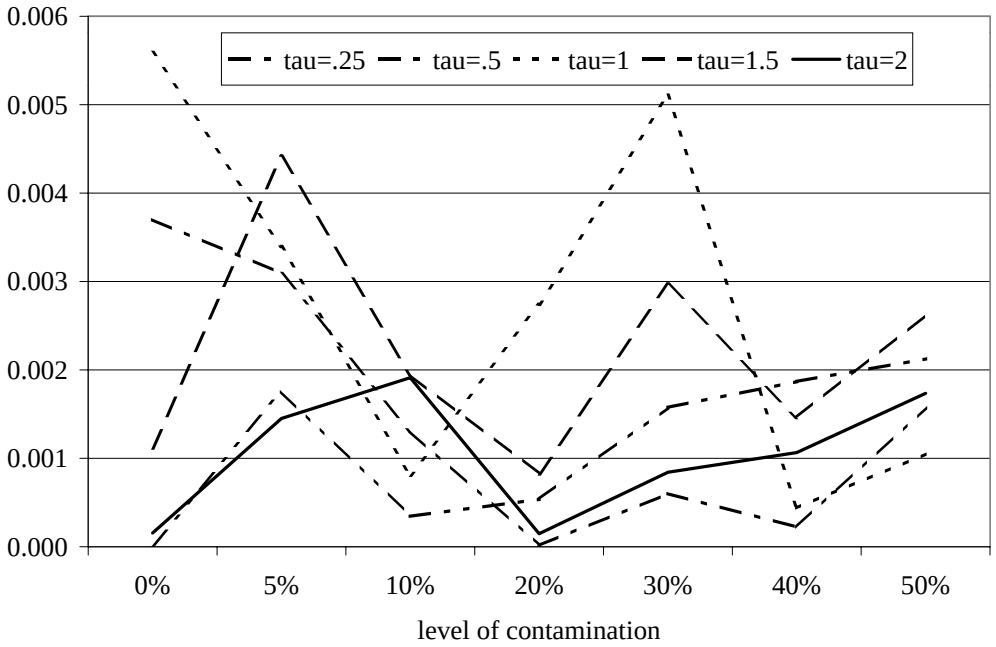

Figure 3.13: Bias in the estimation of β_1 for $\beta_0 = -2,\ \beta_1 = 1$, and different values of τ under different levels of a lognormal contamination.

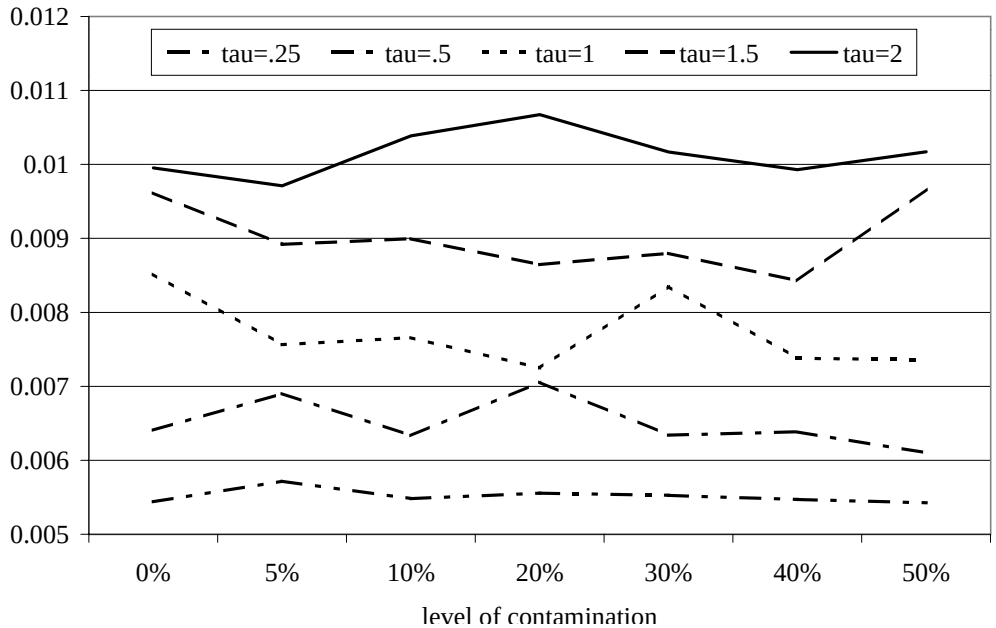

Figure 3.14: Variance of estimate of β_1 for $\beta_0 = -2,\ \beta_1 = 1$, and different values of τ under different levels of a lognormal contamination.

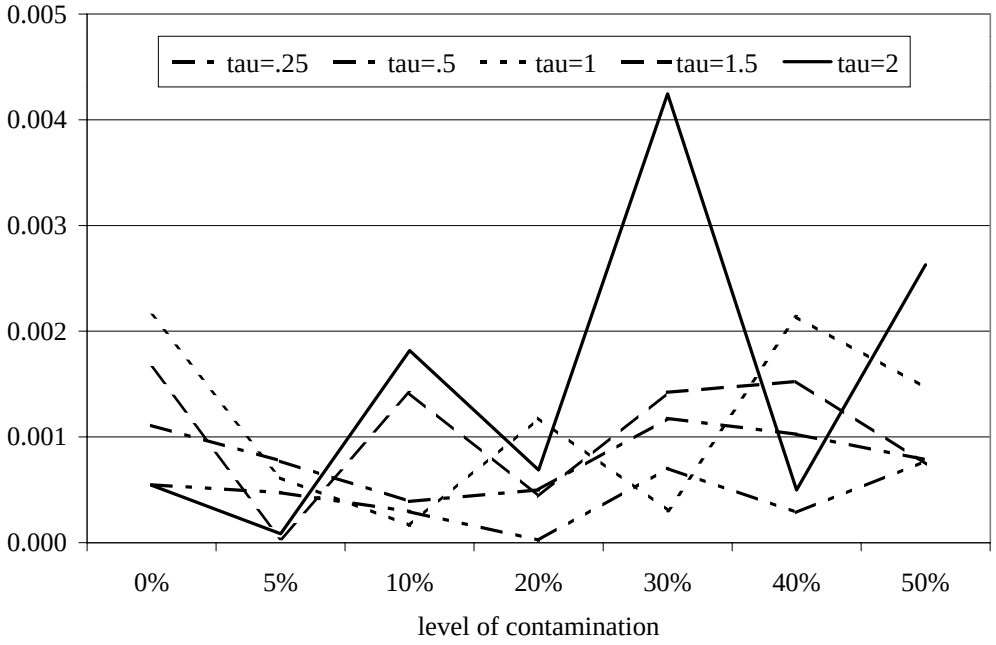

Figure 3.15: Bias in the estimation of β_1 for $\beta_0 = 0,\ \beta_1 = 1$, and different values of τ under different levels of a lognormal contamination.

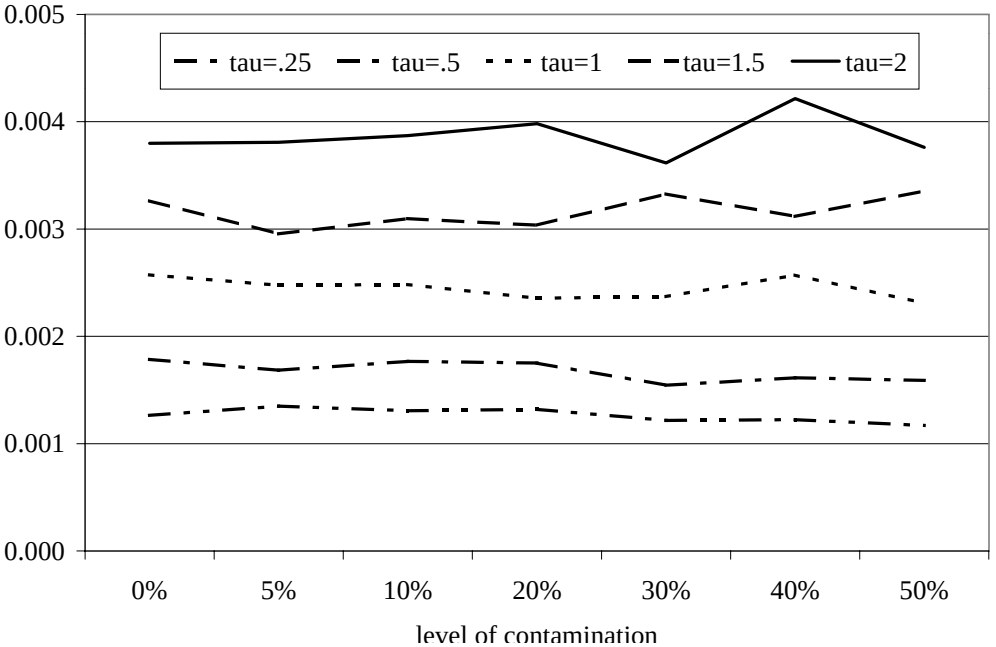

Figure 3.16: Variance of estimate of β_1 for $\beta_0 = 0,\ \beta_1 = 1$, and different values of τ under different levels of a lognormal contamination.

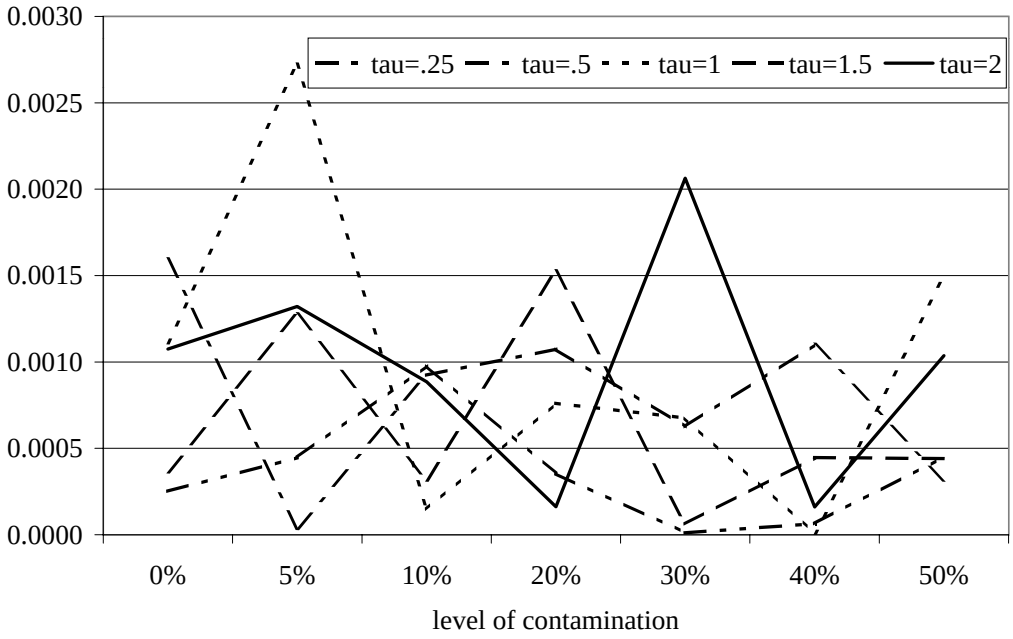

Figure 3.17: Bias in the estimation of β_1 for $\beta_0 = 2,\ \beta_1 = 1$, and different values of τ under different levels of a lognormal contamination.

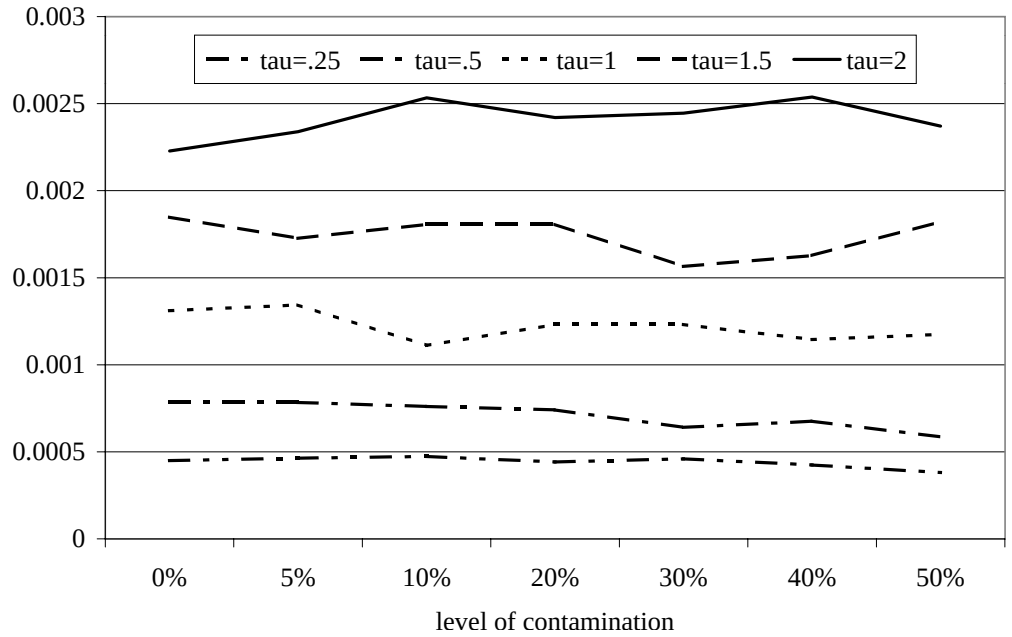

Figure 3.18: Variance of estimate of β_1 for $\beta_0 = 2,\ \beta_1 = 1$, and different values of τ under different levels of a lognormal contamination.

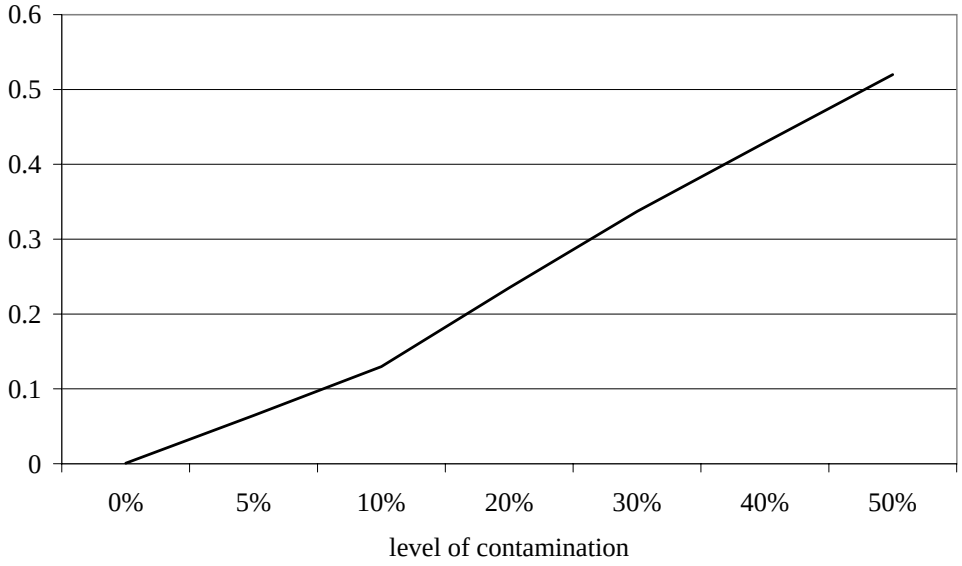

Figure 3.19: Bias in the estimation of τ for $\beta_0 = 0,\ \beta_1 = 1$, and $\tau = 2$ under different levels of a scaled $F_{6,6}$ contamination.

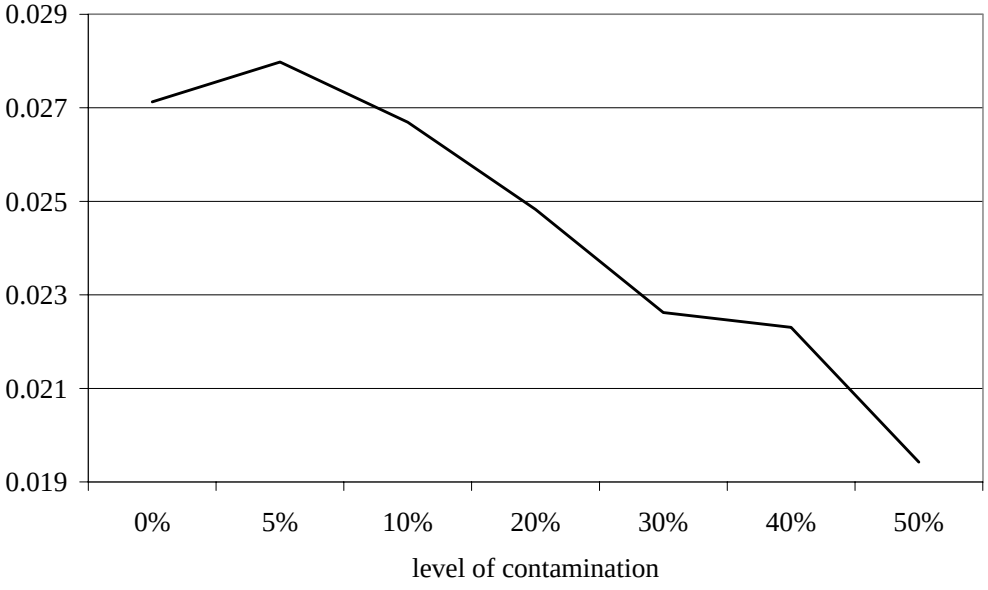

Figure 3.20: Variance of estimate of τ for $\beta_0 = 0,\ \beta_1 = 1$, and $\tau = 2$ under different levels of a scaled $F_{6,6}$ contamination.

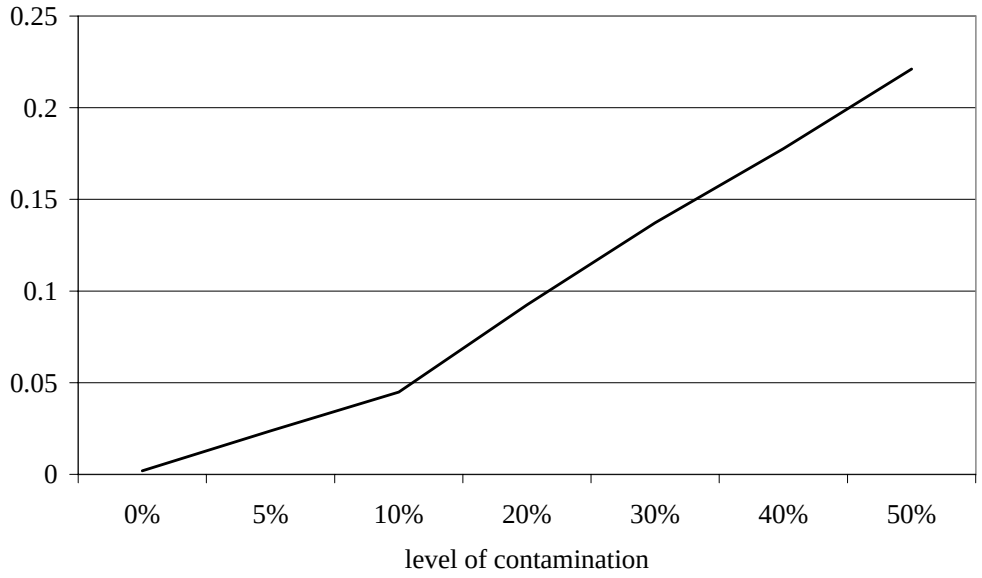

Figure 3.21: Bias in the estimation of β_0 for $\beta_0 = 0,\ \beta_1 = 1$, and $\tau = 2$ under different levels of a scaled $F_{6,6}$ contamination.

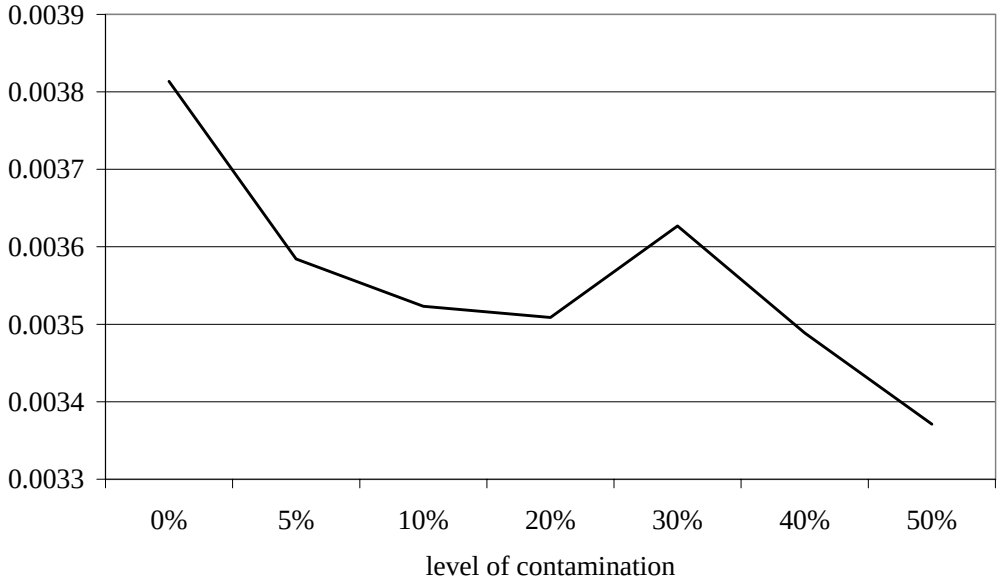

Figure 3.22: Variance of estimate of β_0 for $\beta_0 = 0,\ \beta_1 = 1$, and $\tau = 2$ under different levels of a scaled $F_{6,6}$ contamination.

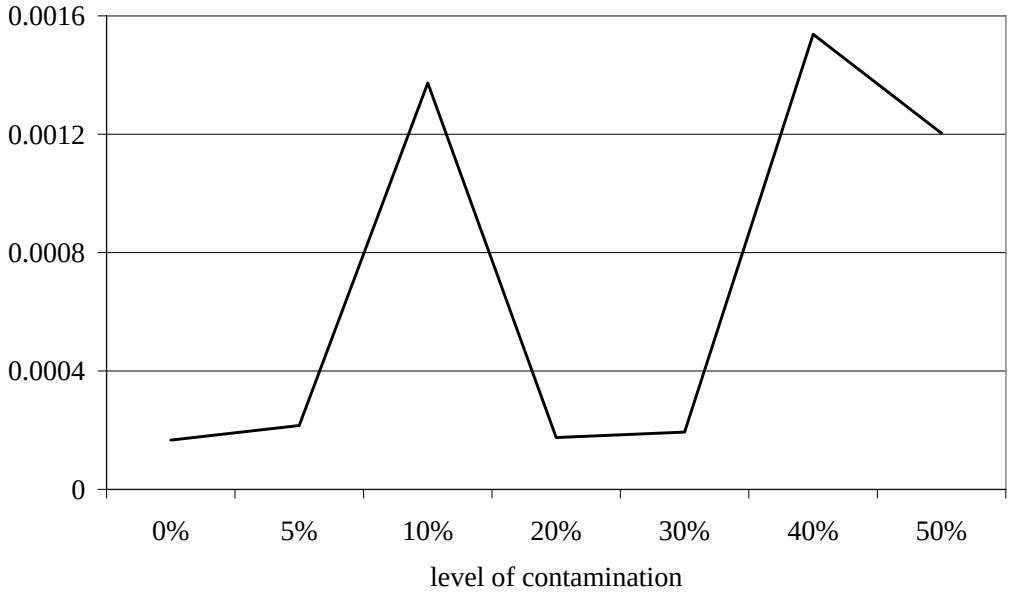

Figure 3.23: Bias in the estimation of β_1 for $\beta_0 = 0,\ \beta_1 = 1$, and $\tau = 2$ under different levels of a scaled $F_{6,6}$ contamination.

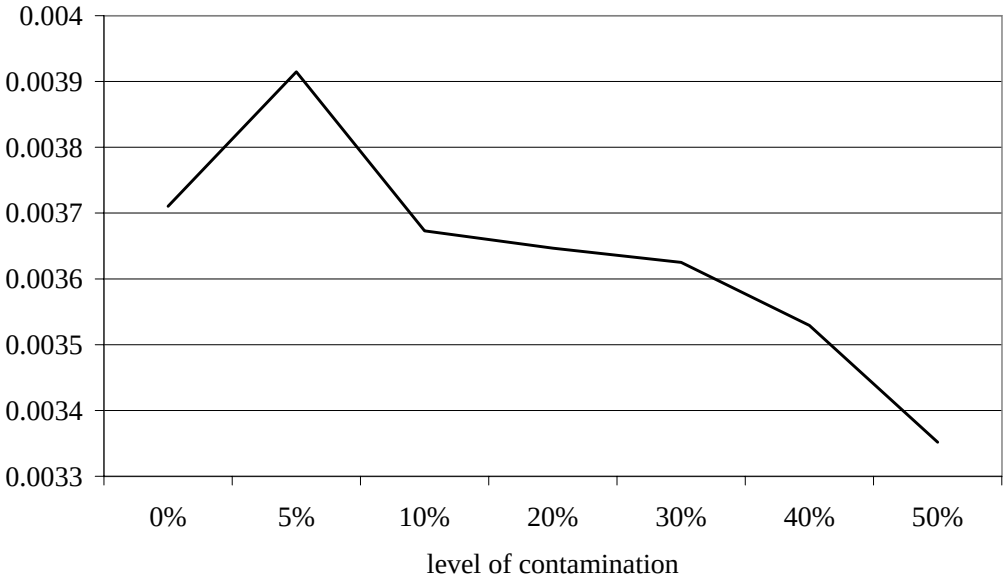

Figure 3.24: Variance of estimate of β_1 for $\beta_0 = 0,\ \beta_1 = 1$, and $\tau = 2$ under different levels of a scaled $F_{6,6}$ contamination.

3.6 Conclusions and future research

In this Chapter we analyze the sensitivity of the asymptotic variance of the M-estimators for GLMMs under a slight contamination of the mixing distribution. This article adds to previous work by presenting the CVF for this general class of estimators in GLMMs and analyzing the CVS in detail for the MLE of the Poisson-Gamma model and two mixed-effects Binomial models. In all cases, it was found that the MLE is V-robust when the distribution of the random effects is contaminated by any other distribution sharing the first two moments with the nominal distribution. A simulation study was performed to illustrate the relevance of this result for the Poisson-Gamma model. In all simulated cases, the variance of the estimators remain almost constant under different levels of contamination.

While the Poisson-Gamma model is attractive for its distributional closed form and its applicability, it might be interesting to examine other Poisson mixed models, as the Poisson-inverse Gaussian or the Poisson-lognormal. Moreover, other estimators suggested in the literature can be also studied. For example, we can derive the CVF of quasi-likelihood estimators for GLMMs or in particular, for Poisson-mixed models.

The simulation study was performed to see the relevance of our theoretical result. It would be interesting to repeat this study for small samples to see if the asymptotic results still hold in the case of finite samples.

Chapter 4

Robust Instrumental Variables Estimator

4.1 Introduction

A classical problem in linear regression arises when some of the covariates are "endogenous", that is, correlated with the error term in the equation to be estimated. In such a situation the ordinary least squares (OLS) estimator yields biased and inconsistent parameter estimates. A common approach to address this problem is to use additional information contained in variables that do not belong to the original equation but are correlated with the endogenous covariates. Under certain conditions, such "instruments" can be used to construct ordinary instrumental variables (OIV) estimators that yield consistent parameter estimates. However, despite its widespread use, the OIV estimator is highly sensitive to outliers in the response, the covariates, and even the instruments.

In this Chapter we propose a *robust* instrumental variables (RIV) estimator based on a robust multivariate location and scatter matrix estimator. Instead of estimating the regression parameters directly as a solution to a robust estimating equation, we robustify the solution of the ordinary estimating equations using high breakdown point S-estimators. We show that, when an appropriate S-estimator is chosen, our RIV estimator is bounded influence (i.e., resistant to extreme observations), consistent, and asymptotically normal.

Since the ordinary instrumental variables estimator (OIV) and its most efficient version known as two-stage least squares estimator (2SLS) are extremely sensitive to aberrant observations, some robust instrumental variables estimators have been developed. Amemiya (1982) extended the least absolute deviation (LAD) estimator as an alternative to the 2SLS estimator. Powell (1983) shows the asymtotic normality of Amemiya's estimator under weak conditions. However, like LAD, this estimator is not bounded-influence. Krasker and Welsch (1985) proposed an instrumental vari-

ables version of the earlier Krasker-Welsch estimator (Krasker and Welsch, 1982). Their estimator is a bounded-influence weighted instrumental variables estimator that downweights an observation only if its influence would otherwise exceed the maximum allowable influence. However, the estimator is complex and hard to implement. More recently, Flavin (1999) derived an instrumental variables version of the Huber (1973) estimator. Although the author claims that his estimator is easier to implement than the Krasker-Welsch instrumental variables estimator, such an estimator is not bounded-influence. Wagenvoort and Waldmann (2002) developed two bounded-influence instrumental variables estimators which are robust versions of 2SLS and generalized method of moments (GMM) estimators, respectively. These estimators are also complicated to implement and compute.

We add to previous work by providing a robust instrumental variables estimator which is less computational expensive than those currently available. As our estimator is a natural extension of the ordinary instrumental variables estimator, it is also easy to implement and interpret. In addition, it is in the class of *weighted* instrumental variables estimators, which gives a simple way to flag outliers and influential points. These properties are extremely useful, specially when using high-dimensional datasets such as those used in data mining, where it is unfeasible to identify one aberrant point at a time or to use computationally demanding algorithms. We also provide an S-Plus/R algorithm to compute both the regression coefficients and the asymptotic covariance matrix estimates (available from the author).

We also propose a diagnostic technique based on our robust covariance-based estimator. Our RIV estimator is a weighted instrumental variables estimator which downweights those points with high Mahalanobis distances. Thus, we propose to detect outliers in any of the variables by comparing the Mahalanobis distances of each data point to the quantiles of the chi-squared distribution with degrees of freedom given by the number of variables in the dataset. Equivalently, we can also look at the weights to flag outliers and leverage points.

The remainder of this Chapter is organized as follows. In Section 4.2, we introduce our RIV estimator. In Section 4.3, we discuss some of its properties. In Section 4.4, we compute

the corresponding Influence Function and show that it is bounded, and we use it to derive a covariance matrix estimator of our RIV estimator. In Section 4.5, we use a real data example with measurement errors and we artificially contaminate it to compare the performance of our RIV estimator with that of the OIV estimator. In addition, we illustrate the use of our diagnostic techniques. Section 4.6 concludes.

4.2 Robust Instrumental Variables Estimator

In this Section we propose a robust instrumental variables estimator. Instead of estimating the regression parameters directly as solutions to robust estimating equations, we robustify the solution of the ordinary estimating equations. We note that the OIV estimator, defined in (1.9), is a function of the sample mean and the sample covariance matrix. However, the sample mean and the sample covariance matrix are not robust estimators of the multivariate location and scatter matrix. Hence, the OIV estimator is extremely sensitive to outliers and influential points. Thus, we propose using a *robust* multivariate location and scatter estimator to construct a robust instrumental variables estimator (RIV) analogous to the OIV estimator.

Let $\boldsymbol{Z}_i = (\boldsymbol{X}_i, \boldsymbol{W}_i, Y_i)^T$, for $i = 1, \ldots, n$, be the $(2p+1)$-dimensional vector of observations. Let $(\boldsymbol{M}, \boldsymbol{S}) \in \mathbb{R}^{(2p+1)} \times \text{PDS}(2p+1)$ be a robust multivariate location and scatter matrix estimator, where $\text{PDS}(2p+1)$ is the set of all positive definite symmetric matrices of order $2p+1$. We can split up the vector $\boldsymbol{M}$ and the matrix $\boldsymbol{S}$ accordingly. That is,

$$\boldsymbol{M} = \left(\boldsymbol{M}_{\boldsymbol{X}}, \boldsymbol{M}_{\boldsymbol{W}}, M_Y\right)^T \quad \text{and} \quad \boldsymbol{S} = \begin{pmatrix} \boldsymbol{S}_{\boldsymbol{XX}} & \boldsymbol{S}_{\boldsymbol{XW}} & \boldsymbol{S}_{\boldsymbol{X}Y} \\ \boldsymbol{S}_{\boldsymbol{WX}} & \boldsymbol{S}_{\boldsymbol{WW}} & \boldsymbol{S}_{\boldsymbol{W}Y} \\ \boldsymbol{S}_{Y\boldsymbol{X}} & \boldsymbol{S}_{Y\boldsymbol{W}} & S_{YY} \end{pmatrix}. \tag{4.1}$$

Consider the model described in Section 1.4.1. We define the RIV estimator of the regression coefficients $\boldsymbol{\beta}$ as

$$\hat{\boldsymbol{\beta}}_{RIV} = (\hat{\beta}_0, \hat{\boldsymbol{\beta}}_1) = g(\boldsymbol{M}, \boldsymbol{S}) = (M_Y - \boldsymbol{M}_{\boldsymbol{X}}\hat{\boldsymbol{\beta}}_1, \boldsymbol{S}_{\boldsymbol{WX}}^{-1}\boldsymbol{S}_{\boldsymbol{W}Y}). \tag{4.2}$$

Note that the RIV estimator defined in (4.2) reduces to the OIV estimator defined in (1.9) when

$(\boldsymbol{M}, \boldsymbol{S})$ is the sample location and scatter estimator. Maronna and Morgenthaler (1986) and Croux et al. (2003) proposed analogous estimators for classical regression models without endogeneity.

We need a robust location and covariance matrix estimator to replace the sample ones. Many of these multivariate estimators are available, such as M-estimators (Maronna, 1976), Stahel-Donoho estimators (Stahel, 1981; Donoho, 1982), the Minimum Volume Ellipsoid and Minimum Covariance Determinant estimators (Rousseeuw, 1984), S-estimators (Davies, 1987; Lopuhaä, 1989) and componentwise and pairwise estimators (Gnanadesikan and Kettenring, 1972; Maronna and Zamar, 2002). It can be proved that the S-estimators are consistent and asymptotically normal, affine equivariant, positive definite, bounded-influence and they achieve the maximal breakdown point (asymptotically 1/2) regardless of dimension of the data for an appropriate choice of ρ (Davies, 1987; Lopuhaä, 1989). Thus, we use this family of estimators to summarize the data and construct our estimator. However, our estimator can be constructed using any other choice of multivariate location and scatter matrix estimator (see Section 1.5.5 for a further description of these estimators).

For a finite sample $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n \in \mathbb{R}^{(2p+1)}$ the S-estimator is defined as the solution $(\boldsymbol{M}, \boldsymbol{S})$ to the problem of minimizing $\det(\boldsymbol{S})$ subject to

$$\frac{1}{n}\sum_{i=1}^{n}\rho\Big(\sqrt{(\boldsymbol{Z}_i - \boldsymbol{M})^T\boldsymbol{S}^{-1}(\boldsymbol{Z}_i - \boldsymbol{M})}\Big) = b_0 \tag{4.3}$$

among all $(\boldsymbol{M}, \boldsymbol{S}) \in \mathbb{R}^{(2p+1)} \times \mathrm{PDS}(2p+1)$ (Lopuhaä, 1989). In order to obtain robust estimates and preserve asymptotic normality the function ρ must satisfy the following conditions:

(R1) ρ is symmetric, has a continuous derivative ψ and $\rho(0) = 0$.

(R2) There exists a finite constant $c_0 > 0$ such that ρ is strictly increasing on $[0, c_0]$ and constant on $[c_0, +\infty)$.

(R3) $\psi'(y)$ and $u(y) = \psi(y)/y$ are bounded and continuous.

The constant $0 < b_0 < \sup\{\rho\}$ is generally chosen to be $E_{\mathbf{0},\boldsymbol{I}}[\rho(\| \boldsymbol{U} \|)]$, where $\boldsymbol{U}$ has an elliptical distribution. If ρ satisfies these conditions, the resulting S-estimator is consistent, asymptotically

normal and has bounded Influence Function (Lopuhaä, 1989). If $b_0 = 2p + 1$, using $\rho(y) = y^2$ in (4.3) yields the sample mean and covariance matrix as a unique solution to previous problem. However, this function does not satisfy the previous conditions. In Section 4.3 we show that using S-estimators to construct our RIV estimator yields a weighted instrumental variables estimator with weights depending on ρ. Then, we add the following condition so that the weights downweight extreme points:

(R4) ρ is such that $u(y) = \psi(y)/y$ is non-increasing in $[0, +\infty)$, where $\psi(y) = \rho'(y)$.

In Section 4.5 we choose to compute our RIV estimator using the Tukey's biweight function as the ρ function in (4.3). However, our estimator can be computed using any ρ function satisfying conditions (R1)-(R4).

4.3 Properties of the RIV estimator

In this Section, we discuss some properties of our estimator.

Proposition 4.3.1. *The estimator $\hat{\boldsymbol{\beta}}_{RIV}$, defined in (4.2) using multivariate location and scatter S-estimators, is a weighted instrumental variables estimator.*

Proof. The S-estimator $(\boldsymbol{M}, \boldsymbol{S})$ satisfies the following first-order conditions (Lopuhaä, 1989):

$$\frac{1}{n}\sum_{i=1}^{n} u(d_i)(\boldsymbol{Z}_i - \boldsymbol{M}) = \boldsymbol{0} \tag{4.4}$$

$$\frac{1}{n}\sum_{i=1}^{n} pu(d_i)(\boldsymbol{Z}_i - \boldsymbol{M})(\boldsymbol{Z}_i - \boldsymbol{M})^T = \frac{1}{n}\sum_{i=1}^{n} v(d_i)\boldsymbol{S}, \tag{4.5}$$

where $u(y) = \psi(y)/y$ and $v(y) = \psi(y)y - \rho(y) + b_0$, for $\psi(y) = \rho'(y)$. Let d_i be the Mahalanobis distances of the data points to $(\boldsymbol{M}, \boldsymbol{S})$. That is, $d_i^2 = (\boldsymbol{Z}_i - \boldsymbol{M})^T \boldsymbol{S}^{-1}(\boldsymbol{Z}_i - \boldsymbol{M})$.

If $\sum_{i=1}^{n} v(d_i) \neq 0$, equation (4.5) can be written as

$$\sum_{i=1}^{n} \left(\frac{pu(d_i)}{\sum_{j=1}^{n} v(d_j)} \right) (\boldsymbol{Z}_i - \boldsymbol{M})(\boldsymbol{Z}_i - \boldsymbol{M})^T = \boldsymbol{S}. \tag{4.6}$$

Partitioning equation (4.6) according to the blocks of $\boldsymbol{S}$ (see equation (4.1)), we obtain

$$\sum_{i=1}^{n} \left(\frac{pu(d_i)}{\sum_{j=1}^{n} v(d_j)} \right) (\boldsymbol{W}_i - \boldsymbol{M}_{\boldsymbol{W}})^T (Y_i - M_Y) = \boldsymbol{S}_{\boldsymbol{W}Y}. \tag{4.7}$$

Using the definition of $\hat{\boldsymbol{\beta}}_1$ given in (4.2), we can replace $\boldsymbol{S_{WY}}$ by $\boldsymbol{S_{WX}}\hat{\boldsymbol{\beta}}_1$ in equation (4.7). Furthermore, using the block of (4.6) corresponding to $\boldsymbol{S_{WX}}$ we have

$$\sum_{i=1}^{n}\left(\frac{pu(d_i)}{\sum_{j=1}^{n}v(d_j)}\right)(\boldsymbol{W}_i-\boldsymbol{M_W})^T\left[(Y_i-M_Y)-(X_i-\boldsymbol{M_X})\hat{\boldsymbol{\beta}}_1\right]=\mathbf{0}. \tag{4.8}$$

Multiplying the previous equation by $\sum_{j=1}^{n}v(d_j)/(p\sum_{i=1}^{n}u(d_i))$ we obtain

$$\sum_{i=1}^{n}\omega_i(\boldsymbol{W}_i-\boldsymbol{M_W})^T\left[(Y_i-M_Y)-(X_i-\boldsymbol{M_X})\hat{\boldsymbol{\beta}}_1\right]=\mathbf{0},$$

where $\omega_i=u(d_i)/\sum_{i=1}^{n}u(d_i)$ and $\sum_{i=1}^{n}\omega_i=1$.

Similarly, we can rewrite equation (4.4) as

$$\sum_{i=1}^{n}\omega_i(\boldsymbol{Z}_i-\boldsymbol{M})=\mathbf{0}. \tag{4.9}$$

Combining equation (4.9) with the definition of $\hat{\beta}_0$ given in (4.2), we obtain

$$\hat{\beta}_0=\boldsymbol{M}_Y-\boldsymbol{M_X}\hat{\boldsymbol{\beta}}_1=\sum_{i=1}^{n}\omega_i\left[Y_i-\boldsymbol{X}_i\hat{\boldsymbol{\beta}}_1\right]. \tag{4.10}$$

Note that (4.8) and (4.10) can be rewritten as

$$\begin{cases} 0=\frac{1}{n}\sum_{i=1}^{n}\omega_i\left[Y_i-\hat{\beta}_0-\boldsymbol{X}_i\hat{\boldsymbol{\beta}}_1\right] \\ \mathbf{0}=\frac{1}{n}\sum_{i=1}^{n}\omega_i\boldsymbol{W}_i^T\left[Y_i-\hat{\beta}_0-\boldsymbol{X}_i\hat{\boldsymbol{\beta}}_1\right]. \end{cases}$$

Thus, our estimator belongs to the class of weighted instrumental variables estimators of the form

$$\hat{\boldsymbol{\beta}}_{RIV}=(\tilde{\boldsymbol{W}}^T\boldsymbol{\Omega}\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{W}}^T\boldsymbol{\Omega}\boldsymbol{Y}$$

where $\tilde{\boldsymbol{X}}=[\mathbf{1}\ \boldsymbol{X}]$, $\tilde{\boldsymbol{W}}=[\mathbf{1}\ \boldsymbol{W}]$, $\mathbf{1}$ is an n-dimensional column vector of ones, and $\boldsymbol{\Omega}$ is a diagonal weighting matrix with ith diagonal element

$$\omega_i=\omega(\boldsymbol{X}_i,\boldsymbol{W}_i,Y_i,\boldsymbol{M},\boldsymbol{S}).$$

□

Under conditions (R1)-(R4), ω is such that $0\leq\omega\leq 1$ and is non-increasing in $[0,+\infty)$. Thus, our estimator downweights points that are far from the bulk of the data. In Section 4.5 we discuss some diagnostic techniques based on the weights that the estimator assigns to each observation. Note

that our estimator reduces to the OIV estimator when the sample mean and sample covariance matrix is used to construct the estimator, i.e., $\omega_i = 1,$ for $i = 1, \ldots, n$. However, its corresponding ρ does not satisfy condition (R2). Finally, we present two properties that our RIV estimator inherits from the S-estimators.

Proposition 4.3.2. *The estimator $\hat{\boldsymbol{\beta}}_{RIV}$, defined in (4.2) using a multivariate location and scatter S-estimator, is consistent and asymptotic normally distributed if conditions (1.7) and (1.8) hold and the S-estimator is appropriately chosen.*

Proof. Davies (1987) and Lopuhaä (1989) proved consistency and asymptotic normality of the multivariate S-estimator of location and scatter matrix. Assuming that this estimator is consistent and asymptotically normal, using the continuous mapping theorem, it is trivial to prove that the regression parameter estimator defined in (4.2) is also consistent and asymptotically normal. □

Proposition 4.3.3. *Let $\boldsymbol{z}$ and $\tilde{\boldsymbol{z}}$ be the matrices of the original and the transformed observations with rows given by $\boldsymbol{z}_i = (\boldsymbol{x}_i, \boldsymbol{w}_i, y_i)^T$ and $\tilde{\boldsymbol{z}}_i = (\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{w}}_i, \tilde{y}_i)^T = (\boldsymbol{x}_i\boldsymbol{Q}, \boldsymbol{w}_i\boldsymbol{P}, \eta y_i + \boldsymbol{x}_i\boldsymbol{\gamma} + \delta)^T$ respectively, for $i = 1, \ldots, n$. Let $\hat{\boldsymbol{\beta}}(\boldsymbol{z})$ be the estimator based on the original data points, and similarly $\hat{\boldsymbol{\beta}}(\tilde{\boldsymbol{z}})$ be the one based on the transformed data. The estimator $\hat{\boldsymbol{\beta}}_{RIV}$, defined in (4.2), is regression and carrier equivariant. That is,*

$$\begin{aligned} \hat{\beta}_0(\tilde{\boldsymbol{z}}) &= \eta\hat{\beta}_0(\boldsymbol{z}) + \delta \\ \hat{\boldsymbol{\beta}}_1(\tilde{\boldsymbol{z}}) &= (\boldsymbol{Q}^{-1})(\eta\hat{\boldsymbol{\beta}}_1(\boldsymbol{z}) + \boldsymbol{\gamma}), \end{aligned}$$

for all $\boldsymbol{\gamma} \in \mathbb{R}^p$, $\eta, \delta \in \mathbb{R}$ and nonsingular $(p \times p)$ matrices $\boldsymbol{Q}$ and $\boldsymbol{P}$.

Proof. We can write $\tilde{\boldsymbol{z}}_i = (\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{w}}_i, \tilde{y}_i)^T = \boldsymbol{A}\boldsymbol{z}_i + \boldsymbol{b}$, where

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{Q}^T & \mathbf{0} & 0 \\ \mathbf{0} & \boldsymbol{P}^T & 0 \\ \boldsymbol{\gamma}^T & \mathbf{0} & \eta \end{pmatrix}, \quad \text{and} \quad \boldsymbol{b} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \delta \end{pmatrix}$$

Let $(\boldsymbol{M}(\boldsymbol{z}), \boldsymbol{S}(\boldsymbol{z}))$ be the multivariate location and scatter estimator based on the observations z. Similarly, define $(\boldsymbol{M}(\tilde{\boldsymbol{z}}), \boldsymbol{S}(\tilde{\boldsymbol{z}}))$ for the transformed observations $\tilde{\boldsymbol{z}}$. As the S-estimators are affine

equivariant,

$$(\boldsymbol{M}(\tilde{\boldsymbol{z}}), \boldsymbol{S}(\tilde{\boldsymbol{z}})) = (\boldsymbol{A}\boldsymbol{M}(\boldsymbol{z}) + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{S}(\boldsymbol{z})\boldsymbol{A}^T). \tag{4.11}$$

Furthermore, by (4.2), the RIV estimator based on the transformed observations is given by

$$\begin{aligned}
\hat{\beta}_0(\tilde{\boldsymbol{z}}) &= M_{\tilde{Y}}(\tilde{\boldsymbol{z}}) - \boldsymbol{M}_{\tilde{\boldsymbol{X}}}(\tilde{\boldsymbol{z}})\hat{\boldsymbol{\beta}}_1(\tilde{\boldsymbol{z}}) \\
\hat{\boldsymbol{\beta}}_1(\tilde{\boldsymbol{z}}) &= \boldsymbol{S}^{-1}_{\tilde{\boldsymbol{W}}\tilde{\boldsymbol{X}}}(\tilde{\boldsymbol{z}})\boldsymbol{S}_{\tilde{\boldsymbol{W}}\tilde{Y}}(\tilde{\boldsymbol{z}}),
\end{aligned}$$

Using (4.11), it is easy to show that

$$\begin{aligned}
M_{\tilde{Y}}(\tilde{\boldsymbol{z}}) &= \eta M_Y(\boldsymbol{z}) + \boldsymbol{M}_{\boldsymbol{X}}(\boldsymbol{z})\boldsymbol{\gamma} + \delta \\
\boldsymbol{M}_{\tilde{\boldsymbol{X}}}(\tilde{\boldsymbol{z}}) &= \boldsymbol{M}_{\boldsymbol{X}}(\boldsymbol{z})\boldsymbol{Q} \\
\boldsymbol{S}_{\tilde{\boldsymbol{W}}\tilde{Y}}(\tilde{\boldsymbol{z}}) &= \boldsymbol{P}^T(\eta\boldsymbol{S}_{\boldsymbol{W}Y}(\boldsymbol{z}) + \boldsymbol{S}_{\boldsymbol{W}\boldsymbol{X}}(\boldsymbol{z})\boldsymbol{\gamma}) \\
\boldsymbol{S}_{\tilde{\boldsymbol{W}}\tilde{\boldsymbol{X}}}(\tilde{\boldsymbol{z}}) &= \boldsymbol{P}^T\boldsymbol{S}_{\boldsymbol{W}\boldsymbol{X}}(\boldsymbol{z})\boldsymbol{Q}.
\end{aligned}$$

Plugging (4.12) into (4.12) we obtain (4.11). □

Then, we can add a linear function of the explanatory variables to the response and the estimator will change accordingly. Moreover, if the coordinate system in the space of the explanatory is linearly transformed, so is our estimator. These equivariance properties enable us to use a single combination of parameters in a simulation study without loss of generality.

4.4 Influence Function and Asymptotic Variance

Hampel (1968, 1974) introduced the Influence Function (IF) in order to investigate the behavior of the asymptotic value of a one-dimensional estimator under small perturbations of the underlying distribution. This concept was later extended to classical linear models estimation (Huber, 1973). In this section we show that the IF of our RIV estimator is bounded. In other words, $\hat{\boldsymbol{\beta}}_{RIV}$ is in the class of bounded-influence instrumental variables estimators. Thus, this is a reliable estimator even if small perturbations in the central model occur.

Let $F_\varepsilon = (1-\varepsilon)F + \varepsilon\Delta_{\boldsymbol{z}}$ denote a neighborhood of the nominal distribution of the observations, F, where $\Delta_{\boldsymbol{z}}$ is the point mass at $\boldsymbol{z}$. For an estimator T representable as a functional of

the empirical distribution, its IF is defined as

$$IF(\boldsymbol{z}; T, F) = \frac{\partial}{\partial \varepsilon} T(F_\varepsilon)|_{\varepsilon=0} = \lim_{\varepsilon \downarrow 0} \frac{T(F_\varepsilon) - T(F)}{\varepsilon}.$$

for those $\boldsymbol{z}$ where the limit exists.

To derive the IF of $\hat{\boldsymbol{\beta}}_{RIV}$, we need to extend the definition of our estimator to a functional formulation. Let $(\boldsymbol{M}(H), \boldsymbol{S}(H))$ be the functional form of the S-estimator of multivariate location and scatter (Lopuhaä, 1989), where H is the joint distribution of the observations $\boldsymbol{Z}_i = (\boldsymbol{X}_i, \boldsymbol{W}_i, Y_i)^T,\ i = 1, \ldots, n$. Similarly to equation (4.1), we can split these functionals accordingly to the components of the observations. Then, the functional corresponding to $\hat{\boldsymbol{\beta}}_{RIV}$ is defined as $\boldsymbol{b}(H) = (b_0(H), \boldsymbol{b}_1(H)^T)^T$, where

$$\boldsymbol{b}_1(H) = \boldsymbol{S}^{-1}_{\boldsymbol{WX}}(H)\boldsymbol{S}_{\boldsymbol{W}Y}(H) \tag{4.12}$$

and

$$b_0(H) = M_Y(H) - \boldsymbol{M}_{\boldsymbol{X}}(H)\boldsymbol{b}_1(H). \tag{4.13}$$

Theorem 4.4.1. *Let H be the distribution of the $(2p+1)$-dimensional vector $\boldsymbol{Z} = (\boldsymbol{X}, \boldsymbol{W}, Y)^T$ and H_0 be the distribution of $\boldsymbol{Z}_0 = (\boldsymbol{X}, \boldsymbol{W}, \varepsilon)^T$. Then, if assumptions (1.7) and (1.8) hold, the Influence Function of the functional $\boldsymbol{b}$ at H is given by*

$$IF(\boldsymbol{z}; \boldsymbol{b}, H) = \begin{pmatrix} IF(\boldsymbol{z}_0; M_Y, H_0) - \boldsymbol{M}_{\boldsymbol{X}}(H_0)\boldsymbol{S}^{-1}_{\boldsymbol{WX}}(H_0) IF(\boldsymbol{z}_0; \boldsymbol{S}_{\boldsymbol{W}Y}, H_0) \\ \boldsymbol{S}^{-1}_{\boldsymbol{WX}}(H_0) IF(\boldsymbol{z}_0; \boldsymbol{S}_{\boldsymbol{W}Y}, H_0) \end{pmatrix}. \tag{4.14}$$

Moreover, using a bounded influence S-estimator (Lopuhaä, 1989), the Influence Function defined above is bounded.

Proof. From Proposition 4.3.3, we can assume that $\beta_0 = 0$ and $\boldsymbol{\beta}_1 = \boldsymbol{0}$ in model (1.5). Then,

$$\boldsymbol{b}(H) = \begin{pmatrix} b_0(H) \\ \boldsymbol{b}_1(H) \end{pmatrix} = \begin{pmatrix} b_0(H_0) \\ \boldsymbol{b}_1(H_0) \end{pmatrix} + \begin{pmatrix} \beta_0 \\ \boldsymbol{\beta}_1 \end{pmatrix}.$$

Therefore,

$$IF(\boldsymbol{z}; \boldsymbol{b}, H) = IF(\boldsymbol{z}_0; \boldsymbol{b}, H_0) = \begin{pmatrix} IF(\boldsymbol{z}_0; b_0, H_0) \\ IF(\boldsymbol{z}_0; \boldsymbol{b}_1, H_0) \end{pmatrix}. \tag{4.15}$$

Consider the contaminated distribution $H_\varepsilon = (1-\varepsilon)H_0 + \varepsilon\Delta_{\boldsymbol{z}}$. Lopuhaä (1989) showed that under certain conditions, the multivariate S-estimator $(\boldsymbol{M}, \boldsymbol{S})$ is uniquely defined and consistent. Then $M_Y(H_0) = 0$ and $\boldsymbol{S}_{\boldsymbol{W}Y}(H_0) = \boldsymbol{0}$ by (A1). Moreover, by Proposition 4.3.2, $b_0(H_0) = 0$ and $\boldsymbol{b}_1(H_0) = 0$. Then,

$$\begin{aligned} IF(\boldsymbol{z}_0; b_0, H_0) &= \frac{d}{d\varepsilon}[M_Y(H_\varepsilon) - \boldsymbol{M}_{\boldsymbol{X}}(H_\varepsilon)\boldsymbol{b}_1(H_\varepsilon)]|_{\varepsilon=0} \\ &= IF(\boldsymbol{z}_0; M_Y, H_0) - \boldsymbol{M}_{\boldsymbol{X}}(H_0) IF(\boldsymbol{z}_0; \boldsymbol{b}_1, H_0) \end{aligned}$$

and

$$\begin{aligned} IF(\boldsymbol{z}_0; \boldsymbol{b}_1, H_0) &= \frac{d}{d\varepsilon}[\boldsymbol{S}^{-1}_{\boldsymbol{W}\boldsymbol{X}}(H_\varepsilon)\boldsymbol{S}_{\boldsymbol{W}Y}(H_\varepsilon)]|_{\varepsilon=0} \\ &= \boldsymbol{S}^{-1}_{\boldsymbol{W}\boldsymbol{X}}(H_0) IF(\boldsymbol{z}_0; \boldsymbol{S}_{\boldsymbol{W}Y}, H_0) \end{aligned}$$

Then, from (4.15), (4.16) and (4.16) we obtain equation (4.14). Moreover, as for an appropriate choice of the function ρ, the S-estimator has bounded Influence Function (Lopuhaä, 1989), then (4.14) and (1.8) imply that $\hat{\boldsymbol{\beta}}_{RIV}$ is B-robust (i.e., it has bounded Influence Function). □

Lopuhaä (1989) showed that, under regularity conditions, the asymptotic variance of the S-estimator $(\boldsymbol{M}, \boldsymbol{S})$ is given by

$$\int IF(\boldsymbol{z}; (\boldsymbol{M}, \boldsymbol{S}), H) IF^T(\boldsymbol{z}; (\boldsymbol{M}, \boldsymbol{S}), H) dH(\boldsymbol{z}).$$

As our estimator can be written as a continuous function of $(\boldsymbol{M}, \boldsymbol{S})$ (see equation (4.2)), then the asymptotic variance of $\hat{\boldsymbol{\beta}}_{RIV}$ is given by

$$AV(\boldsymbol{b}, H) = \int IF(\boldsymbol{z}; \boldsymbol{b}, H) IF^T(\boldsymbol{z}; \boldsymbol{b}, H) dH(\boldsymbol{z}), \tag{4.16}$$

where $\boldsymbol{b}$ is the functional defined at (4.12) and (4.13). Given the complexity of the calculations, we are not presenting a closed form formula for (4.16). However, replacing expectations with average and parameters with estimates we can estimate the asymptotic variance of the regression estimators by

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}[IF(\boldsymbol{z}_i; \boldsymbol{b}, H_n) IF^T(\boldsymbol{z}_i; \boldsymbol{b}, H_n)], \tag{4.17}$$

where H_n is the empirical joint distribution of Z, $\boldsymbol{b}(H_n) = \hat{\boldsymbol{\beta}}_{RIV}$ and the IF is defined in (4.14).

4.5 Example

We use the Alaskan earthquake data in Fuller (1987) to examine the performance of the RIV estimator. The dataset contains information about 62 earthquakes occurring between 1969 and 1978. Specifically, we want to see how the earthquake strength, measured in terms of the true value of the body waves, x_t, impacts on the amplitude of the surface waves of the earthquake. However, instead of observing x_t, we observe X_t, which is the logarithm of the seismogram amplitude of longitudinal body waves measured at some observation stations. In addition, we observe Y_t, the logarithm of the seismogram amplitude on 20 second surface waves and an instrumental variable W_t, which is the logarithm of maximum seismogram trace amplitude at short distance (See Table 4.1).

To compute the RIV, we use an S-estimator based on Tukey's biweight function

$$\rho_c(t) = \min(\frac{t^2}{2} - \frac{t^4}{2c^2} + \frac{t^6}{6c^4}, \frac{c^2}{6}) \tag{4.18}$$

where the constant b is set to ensure maximum breakdown point of the S-estimator and the tuning constant c is selected such that $E_H[\rho(d(\boldsymbol{z}))] = b$ for $H = N(\mathbf{0}, \boldsymbol{I})$. The S-estimator is computed using an adapted version for S-Plus of the fast and accurate algorithm of Ruppert (1992). A program to compute the RIV estimate and its asymptotic variance covariance matrix, defined in (4.17), was written in S-Plus and it is available from the author. The estimate of the asymptotic covariance matrix of the OIV estimator is computed using formula (1.4.15) in Fuller (1987).

Figure 4.1 shows the observations of the response variable and the covariate of the original dataset, together with the OIV (dashed line) and the RIV (solid line) fit. The discrepancy between both lines reflects the presence of outlying observations in the original dataset.

Thus, we examine the Mahalanobis distances of each point to a robust S-estimator and the weights that our estimator gives to each observation. Figure 4.2 contains these distances and, as described in Section (4.1), they are compared with the 95% quantiles of the chi-squared distribution. Observations 16, 25, 28, 45, 54 and 60 exceed this cutoff point. Figure 4.3 shows that the RIV estimator downweights these points. We identify these observations with filled circles in

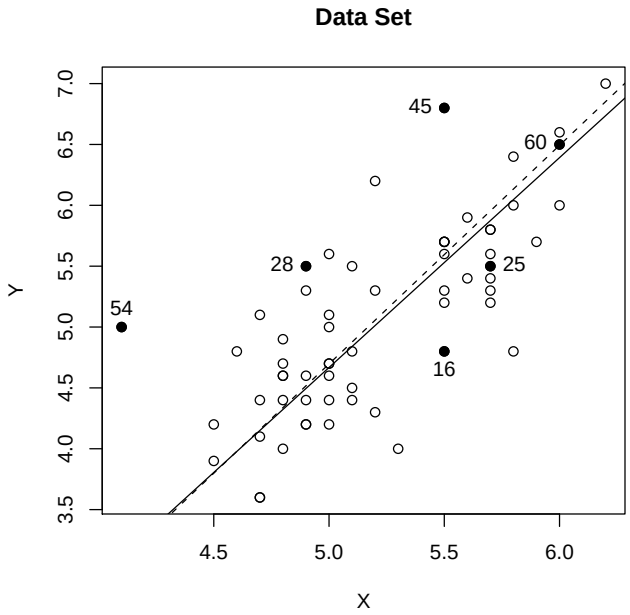

Figure 4.1: Measures of strength for 62 Alaskan earthquakes with the OIV (dashed line) and the RIV (solid line) fit.

Figure 4.1. It is also important to see that observations $\{28, 54, 60\}$ have a Mahalanobis distance larger than the tunning constant c of equation (4.18). Figure 4.3 shows that the RIV estimator completely downweightes these three observations.

It is interesting to note that Fuller (1987) identified only observation 54 using residual plots based on the OIV estimator. However, using our robust estimator and diagnostic techniques enables us to detect additional outlying observations such as observation 28, which has also zero weight in the robust method. It may result strange that observation 60 appears as an outlier according to our diagnostic techniques, while this point lies very close to the robust line and has exact fit for the other. This can be explained analyzing the distribution of the instrumental variable. The value of the instrument corresponding to this observation is the maximum one and differs from that of the other observations.

In addition, we create a clean dataset deleting those points which are identified as outliers or leverage points by our estimator (see Figure 4.1-4.3).

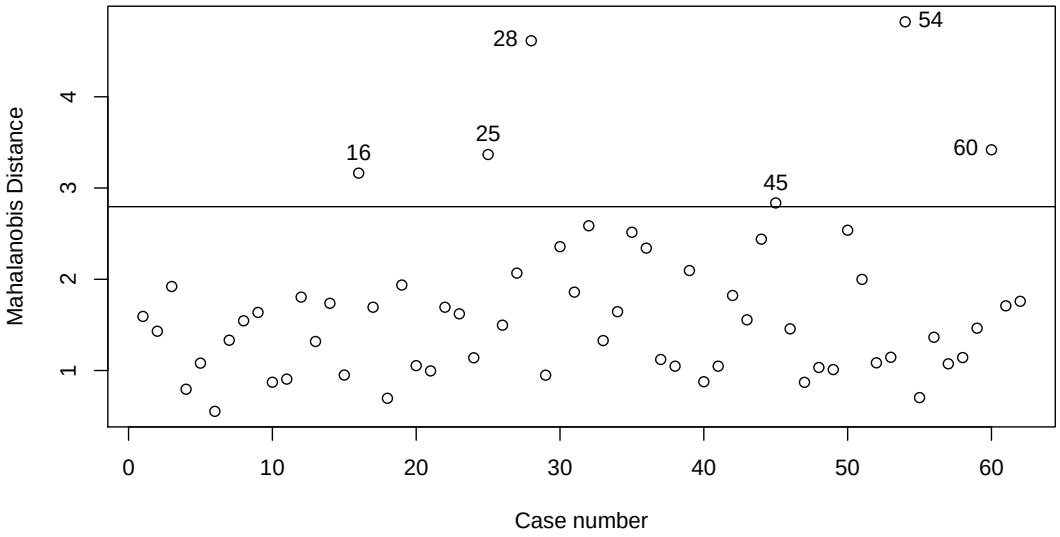

Figure 4.2: Mahalanobis Distances for the original dataset.

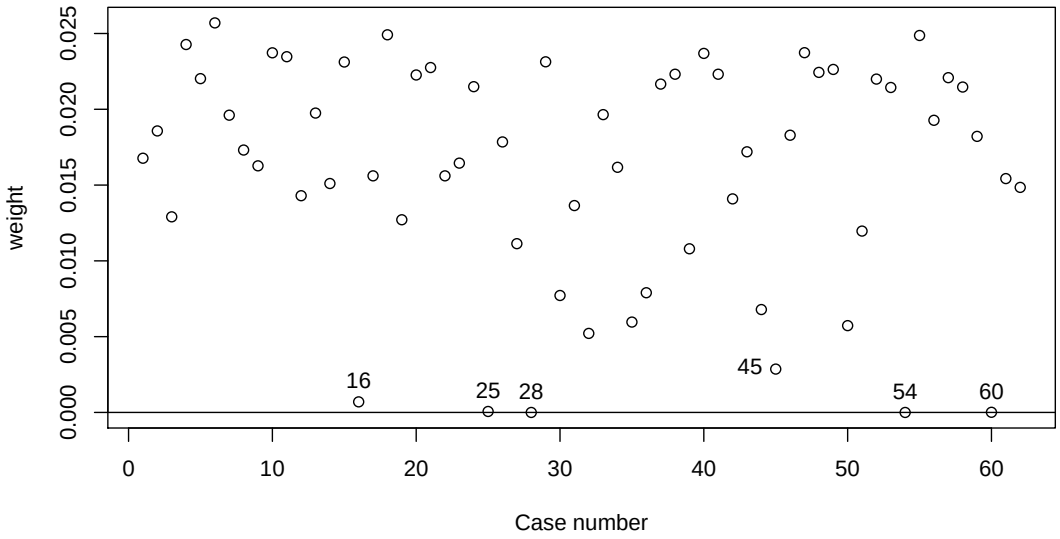

Figure 4.3: Weights assigned by the RIV estimator to each point of the original dataset.

The first graph of Figure 4.4 contains the remaining observations and the robust (RIV) and non-robust (OIV) regression lines. We can see that when there are no extreme points in the data, the lines are almost identical. Finally, we artificially contaminate the original data set in the different subspaces of the data to compare the performance of the RIV estimator with the OIV estimator. We illustrate the results of the both estimations in Figure 4.4. The last seven graphs of Figure 4.4 correspond to the contaminated datasets. We emphasize the effect that an extreme observation in the instrumental variables have on the OIV using the third graph of Figure 4.4. In this dataset, the instrument of an observation over the line has been highly contaminated. In all cases, the classical solution, represented by the dashed line (OIV estimator), is pulled away by the extreme observations. Graphs 1, 2, 3, 5 and 7 show that the effect of these observations may even change the sign of the slope. Note that a negative slope does not make any sense in this problem. On the contrary, our robust estimator remains almost unchanged in all cases. Tables 4.1 and 4.2 report the original and the created datasets, respectively. Table 4.3 summarizes the estimates found using both estimators for all the datasets together with its estimated covariance matrices.

We see that both the estimate and the estimated asymptotic covariance matrix of the OIV estimator are highly influenced by the outliers and leverage points, while those of the RIV estimator appear to be very stable throughout all the datasets.

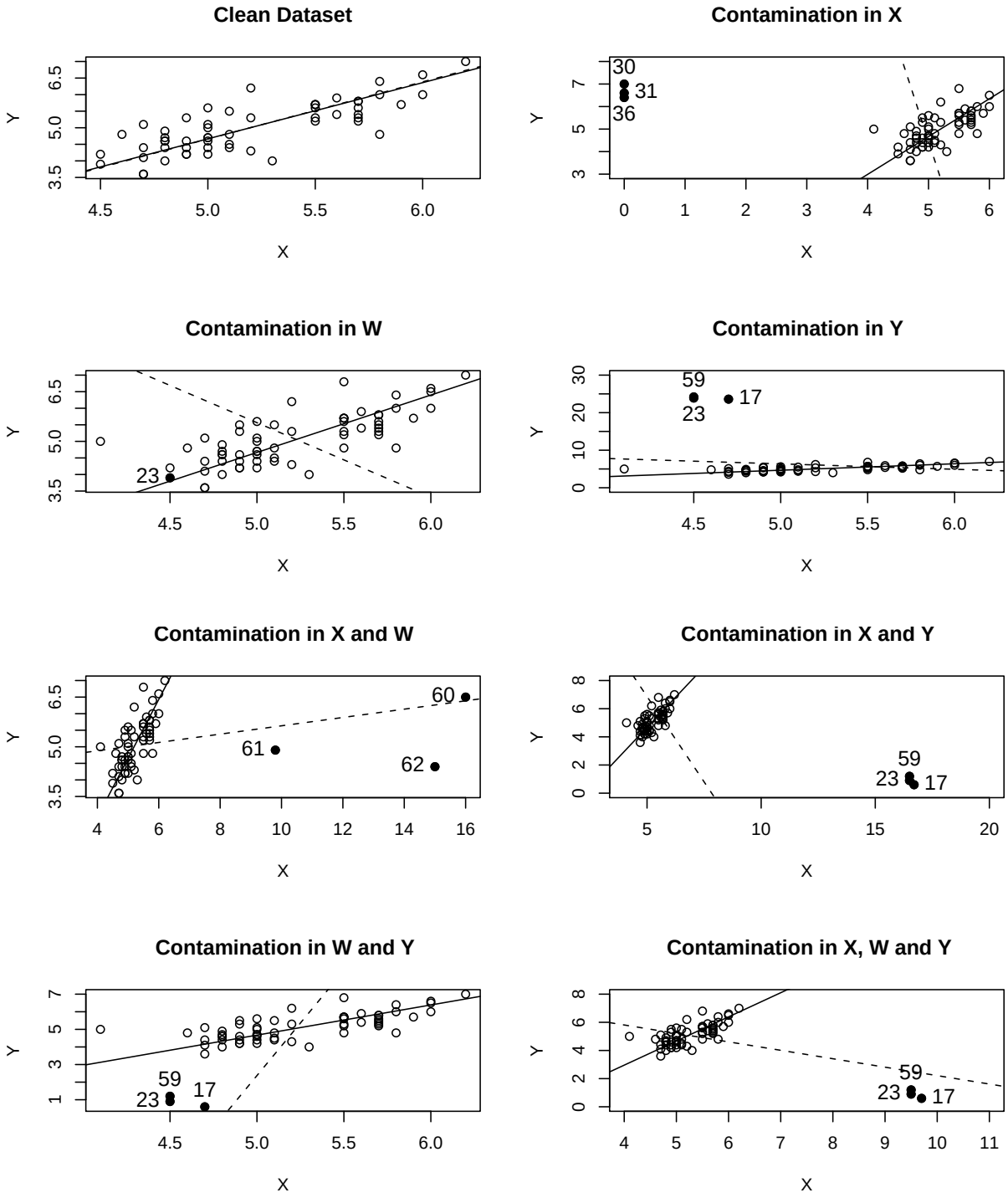

Figure 4.4: Clean and Contaminated Datasets with the OIV (dashed line) and the RIV (solid line) fit. Solid points identify those that have been artificially contaminated.

Table 4.1: Three measures of strength for 62 Alaskan earthquake (from Fuller, 1987; source Meyers and von Hake, 1976).

t	Y_t	X_t	W_t	t	Y_t	X_t	W_t	t	Y_t	X_t	W_t
1	5.5	5.1	5.6	22	3.6	4.7	4.3	43	4.8	5.1	5.5
2	5.7	5.5	6	23	3.9	4.5	4.6	44	6.2	5.2	5.8
3	6	6	6.4	24	4	4.8	4.6	45	6.8	5.5	6.2
4	5.3	5.2	5.2	25	5.5	5.7	4.9	46	6	5.8	5.8
5	5.2	5.5	5.7	26	5.6	5.7	5.5	47	4.6	4.9	4.7
6	4.7	5	5.1	27	5.1	4.7	4.7	48	4.1	4.7	4.5
7	4.2	5	5	28	5.5	4.9	4.1	49	4.4	4.9	4.6
8	5.2	5.7	5.5	29	4.4	4.8	4.9	50	4	5.3	5.2
9	5.3	4.9	5	30	7	6.2	6.5	51	5	5	5.6
10	5.1	5	5.2	31	6.6	6	6.3	52	5.9	5.6	5.9
11	5.6	5.5	5.8	32	5.4	5.7	5.1	53	5.7	5.5	5.9
12	4.8	4.6	4.9	33	5.3	5.7	5.7	54	5	4.1	5.3
13	5.4	5.6	5.9	34	5.7	5.9	5.8	55	5.3	5.5	5.5
14	4.3	5.2	4.7	35	4.8	5.8	5.7	56	5.7	5.5	5.4
15	4.4	5.1	4.9	36	6.4	5.8	5.7	57	4.7	4.8	4.7
16	4.8	5.5	4.6	37	4.2	4.9	4.9	58	4.6	4.8	4.6
17	3.6	4.7	4.3	38	5.8	5.7	5.9	59	4.2	4.5	4.6
18	4.6	5	4.8	39	4.6	4.8	4.3	60	6.5	6	7.1
19	4.5	5.1	4.5	40	4.7	5	5.2	61	4.9	4.8	4.6
20	4.2	4.9	4.6	41	5.8	5.7	5.9	62	4.4	5	5.3
21	4.4	4.7	4.6	42	5.6	5	5.4				

Table 4.2: Contaminated Datasets.

t	Y_t	X_t	W_t	t	Y_t	X_t	W_t
Contamination in XWY				Contamination in XW			
17	9.7	11.3	0.6	60	16	17.1	6.5
23	9.5	11.6	0.9	61	9.8	9.6	4.9
59	9.5	11.6	1.2	62	15	15.3	4.4
Contamination in Y				Contamination in XY			
17	4.7	4.3	23.6	17	16.7	4.3	0.6
23	4.5	4.6	23.9	23	16.5	4.6	0.9
59	4.5	4.6	24.2	59	16.5	4.6	1.2
Contamination in X				Contamination in WY			
30	0	6.5	7	17	4.7	16.3	0.6
31	0	6.3	6.6	23	4.5	16.6	0.9
36	0	5.7	6.4	59	4.5	16.6	1.2
Contamination in W							
23	4.5	24.6	3.9				

Table 4.3: Estimates of the regression coefficients, standard errors (in parentheses) and variance-covariance matrix of each estimator.

RIV		$\hat{\boldsymbol{V}}_{RIV}$			OIV		$\hat{\boldsymbol{V}}_{OIV}$		
$\hat{\beta}_0$	$\hat{\beta}_1$		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$		$\hat{\beta}_0$	$\hat{\beta}_1$
Original Dataset									
-3.954	1.724	$\hat{\beta}_0$	0.943		-4.283	1.796	$\hat{\beta}_0$	1.240	
(.971)	(.184)	$\hat{\beta}_1$	-0.179	0.034	(1.113)	(.212)	$\hat{\beta}_1$	-0.237	0.045
Clean Dataset									
-3.811	1.695	$\hat{\beta}_0$	1.059		-3.966	1.725	$\hat{\beta}_0$	0.848	
(1.029)	(.195)	$\hat{\beta}_1$	-0.201	0.038	(.921)	(.176)	$\hat{\beta}_1$	-0.162	0.031
Contamination in X									
-3.722	1.678	$\hat{\beta}_0$	1.045		45.937	-8.297	$\hat{\beta}_0$	6552.581	
(1.022)	(.197)	$\hat{\beta}_1$	-0.201	0.039	(80.948)	(16.437)	$\hat{\beta}_1$	-1330.372	270.171
Contamination in W									
-4.00	1.733	$\hat{\beta}_0$	1.016		16.803	-2.248	$\hat{\beta}_0$	125806.90	
(1.008)	(.197)	$\hat{\beta}_1$	-0.193	0.037	(354.69)	(68.02)	$\hat{\beta}_1$	-24126.28	4626.753
Contamination in Y									
-3.902	1.715	$\hat{\beta}_0$	0.993		13.554	-1.439	$\hat{\beta}_0$	59.138	
(.996)	(.189)	$\hat{\beta}_1$	-0.188	0.036	(7.69)	(1.47)	$\hat{\beta}_1$	-11.289	2.165

Table 4.4: Estimates of the regression coefficients, standard errors (in parentheses) and variance-covariance matrix of each estimator.

RIV		$\hat{\boldsymbol{V}}_{RIV}$			OIV		$\hat{\boldsymbol{V}}_{OIV}$		
$\hat{\beta}_0$	$\hat{\beta}_1$		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$		$\hat{\beta}_0$	$\hat{\beta}_1$
Contamination in XW									
-4.235	1.778	$\hat{\beta}_0$	0.876		4.376	0.126	$\hat{\beta}_0$	0.092	
(.936)	(.179)	$\hat{\beta}_1$	-0.167	0.032	(.303)	(.054)	$\hat{\beta}_1$	-0.015	0.003
Contamination in XY									
-3.902	1.715	$\hat{\beta}_0$	0.993		18.986	-2.424	$\hat{\beta}_0$	140.541	
(.996)	(.189)	$\hat{\beta}_1$	-0.188	0.036	(11.85)	(2.042)	$\hat{\beta}_1$	-24.172	4.172
Contamination in WY									
-3.902	1.715	$\hat{\beta}_0$	0.993		-57.360	11.947	$\hat{\beta}_0$	2674.348	
(.996)	(.189)	$\hat{\beta}_1$	-0.188	0.036	(51.714)	(9.916)	$\hat{\beta}_1$	-512.794	98.334
Contamination in XWY									
-3.902	1.715	$\hat{\beta}_0$	0.993		8.208	-0.599	$\hat{\beta}_0$	0.535	
(.996)	(.189)	$\hat{\beta}_1$	-0.188	0.036	(.731)	(.131)	$\hat{\beta}_1$	-0.095	0.017

4.6 Conclusions

In this Chapter we propose a robust instrumental variables estimator based on high breakdown point S-estimators of location and scatter. The resulting estimator has bounded Influence Function and satisfies the usual asymptotic properties for suitable choices of the S-estimator used. Moreover, it is a weighted instrumental variables estimator with weights depending on the Mahalanobis distances of the data points. In particular, when these weights are one, the estimator reduces to the ordinary instrumental variables estimator. We also derive an estimate for the asymptotic covariance matrix of our estimator which is robust against outliers and leverage points.

In addition, we build a diagnostic technique using the RIV estimator and we use it in an example to flag outliers in all the dimensions of the data. We compare the RIV estimator with the OIV estimator in many datasets. Both the estimate and the estimated covariance matrix of the RIV estimator remains almost unchanged, while those of the OIV are highly affected by the introduced contamination.

References

Alqallaf, F. A., Konis, K. P., Martin, R. D. and Zamar, R. H. (2002). "Scalable robust covariance and correlation estimataes for data mining." *Proceedings of the Seventh ACM SIGKDD*, 14-23.

Amemiya, T. (1982). "Two Stage Least Absolute Deviation Estimators." *Econometrica* **50**, 689-711.

Amemiya Y. and Fuller W. A. (1984). "Estimation for the Multivariate Errors-in-Variables Model with Estimated Error Covariance Matrix." *Ann. Statist.* **12**, 497-509.

Amemiya, T. (1985). *Advanced Econometrics*. Cambridge, MA: Harvard University Press.

Anderson, D. A. and Aitkin, M. A. (1985). "Variance component models with binary response: interviewer variability." *J. Roy. Statist. Soc. Ser. B* **47**, 203-210.

Anderson, T. W. (1984). "Estimating Linear Statistical Relationships." *Ann. Statist.* **12**, 1-45.

Arellano, M. (2002). "Sargan's instrumental variables estimation and the generalized method of moments." *J. Bus. Econom. Statist.* **20**, 450-459.

Berkson, J. (1944). "Application of the logistic function to bio-assay." *J. Amer. Statist. Assoc.* **39**, 357-365.

Bliss, C. (1934), "The Method of Probits." *Science* **79**, 38-39.

Breslow, N. E. and Clayton, D. G. (1993). "Approximate Inference in Generalized Linear Mixed Models." *J. Amer. Statist. Assoc.* **88**, 9-25.

Cantoni, E. and Ronchetti, E. (2001). "Robust Inference in Generalize Linear Models." *J. Amer. Statist. Assoc.* **96**, 1022-1030.

Carrol, R. J., Ruppert, D. and Stefanski, L.A. (1995). *Measurement Error in Nonlinear Models*. London: Chapman and Hall.

Cheng, C. L. and Van Ness, J. W. (1997). "Robust Calibration." *Technometrics* **39**, 401-411.

Chilson, J.; Ng, R.; Wagner, A. and Zamar, R. H. (2003). "Parallel computation of high dimensional robust correlation and covariance matrix." To appear in *Proceedings of the Seventh ACM SIGKDD*,

Clayton, D. G. (1996). "Generalized linear mixed models." In *Markov Chain Monte Carlo in Practice*. Gilks, W. R. ; Richardson S. and Spiegelhalter, D. J. eds. London: Chapman and Hall. 275-301.

Cox, D. R. (1983). "Some remarks on overdispersion." *Biometrika* **70**, 269-274.

Crouch, E. and Siegelman, D. (1990). "The evaluation of integrals of the form $\int f(t)\exp(-t^2)dt$: application to logistic-normal models." *J. Amer. Statist. Assoc.* **85**, 464-469.

Croux, C., Van Aelst, S., and Dehon, C. (2003). "Bounded influence regression using high breakdown scatter matrices." *Ann. Inst. Statist. Math.* **55**, 265-285.

Crowder, M. J. (1978). "Beta-Binomial anova for proportions." *Appl. Statist.* **27**, 34-37.

Davies, P. L. (1987). "Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices." *Ann. Statist.* **15**, 1269-1292.

Dellaportas, P. and Smith, A. F. M. (1993). "Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling." *J. Roy. Statist. Soc. Ser. C* **42**, 443-459.

Dey, D. K., Ghosh, S. and Mallick, B. (1999). *Generalized Linear Models: a Bayesian Perspective*. New York: Marcel Dekker.

Dobson, A. J. (2002). *An Introduction to Generalized Linear Models*. (2nd ed.). New York: Chapman and Hall.

Donoho, D. L. (1982). "Breakdown properties of multivariate location estimators." Qualifying paper, Harvard University.

Duffy, D. E. and Santner, T. J. (1989). "On the small sample properties of norm-restricted maximum likelihood estimators for logistic regression models." *Comm. Statist. Theory Methods* **18**, 959-980.

Fahrmeir, L. (1990). "Maximum likelihood estimation in misspecified generalized linear models." *Statistics* **21**, 487-502.

Fahrmeir, L. and Kaufmann, H. (1985). "Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models." *Ann. Statist.* **13**, 342-368.

Fahrmeir, L. and Tutz, G. (1997). *Multivariate Statistical Modelling Based on Generalized Linear Models.* (3rd ed.) New York: Springer-Verlag.

Fellner, W. H. (1986). "Robust estimation of variance components." *Technometrics* **28**, 51-60.

Flavin, M. (1999), "Robust Estimation of the Joint Consumption/Asset Demand Decision." National Bureau of Economics Research Working Paper 7011.

Frome, E. L., Kutner, M. H. and Beauchamp, J. J. (1973). "Regression analysis of Poisson-distributed data." *J. Amer. Statist. Assoc.* **68**, 935-940.

Fuller, W. A. (1980). "Properties of some estimators for the error-in-variables model." *Ann. Statist.* **8**, 407-422.

Fuller, W. A. (1987). *Measurement Error Models.* New York: Wiley.

Gaver, D. P. and O'Muircheartaigh, I. G. (1987). "Robust empirical Bayes analysis of event rates." *Technometrics* **29**, 1-15.

Gleser, L. J. (1981). "Estimation in a multivariate errors in variables regression model: large sample results." *Ann. Statist.* **9** , 24-44.

Gnanadesikan, R. and Kettenring J. R. (1972). "Robust estimates, residuals, and outlier detection with multiresponse data." *Biometrics* **28**, 81-124.

Gustafson, P. (1996). "The effect of mixing-distribution misspecification in conjugate mixture models." *The Canadian Journal of Statistics* **24**, 307-318.

Grübel, R. (1988). "A minimal characterization of the covariance matrix." *Metrika* **35**, 49-52.

Haberman, S. J. (1974). "Log-linear models for frequency tables with ordered classifications." *Biometrics* **30**, 589-600.

Haberman, S. J. (1977). "Maximum likelihood estimates in exponential response models." *Ann. Statist.* **5**, 815-841.

Hampel, F. R. (1968) *Contributions to the Theory of Robust Estimation.* Ph.D. thesis. University of California, Berkeley.

Hampel, F. R. (1971). "A general qualitative definition of robustness." *Ann. Math. Statist.* **42**, 1887-1896.

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

Hampel, F. R. (1974). "The influence curve and its role in robust estimation." *J. Amer. Statist. Assoc.* **69**, 383-393.

Hampel, F. R., Rousseeuw and P. J., Ronchetti, E. (1981). "The change-of-variance curve and optimal redescending M-estimators." *J. Amer. Statist. Assoc.* **76**, 643-648.

Hampel, F. R., Ronchetti, E., Rousseeuw, P. J. and Stahel, A. (1986). *Robust Statistics: The Approach Based on Influence Functions.* New York: Wiley.

Heagerty, P. J. and Zeger, S. L. (2000). "Marginalized multilevel models and likelihood inference." *Statist. Sci.* **15**, 1-26.

Henderson, C. R. (1950). "Estimation of genetic parameters (abstract)." *Ann. Math. Statist.* **21**, 309-310.

Hill, R. W. (1977). "Robust regression when there are outliers in the carriers." Ph.D. thesis. Harvard University, Cambridge, Mass.

Huber, P. J. (1964). "Robust estimation of location parameter." *Ann. Math. Statist.* **35**, 73-101.

Huber, P. J. (1967). "The behavior of maximum likelihood estimates under nonstandard conditions." *Proc. Fifth Berkeley Sympos. Math. Statist. Prob.* **1**, 221-233. University of California Press.

Huber, P. J. (1973). "Robust regression: Asymptotics, conjectures, and Monte Carlo." *Ann. Statist.* **1**, 799-821.

Jiang, J. (1998). "Consistent estimators in generalized linear mixed models." *J. Amer. Statist. Assoc.* **93**, 720-729.

Jiang, J. (1999). "Conditional inference about generalized linear mixed models." *Ann.. Statist.* **27**, 1974-2007.

Jiang, J. (2001). "Robust estimation in GLMM." *Biometrika* **88**, 753-765.

Kaufmann, H. (1988). "On existence and uniqueness of maximum likelihood estimates in quantal and ordinal response models." *Metrika* **35**, 291-313.

Kedem, B. and Fokianos, K. (2002). *Regression models for time series analysis.* Hoboken: Wiley.

Kendall, M. G. and Stuart, A. (1979). *The advanced theory of Statistics.* Vol. 2, *Inference and Relationship.* (4th ed.). London: Griffin.

Krasker, W. S. and Welsch, R. E. (1982). "Efficient bounded-influence regression estimation." *J. Amer. Statist. Assoc.* **77**, 595-604.

Krasker, W. S. and Welsch, R. E. (1985). "Resistant estimation for simultaneous-equations models using weighted instrumental variables." *Econometrica* **53**, 1475-1488.

Kuk, A. Y. C. (1995). "Asymptotically unbiased estimation in generalized linear models with random effects." *J. Roy. Statist. Soc. Ser. B* **57**, 395-407.

Künsch, H. R., Stefanski, L. A. and Carroll, R. J. (1989), "Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models." *J. Amer. Statist. Assoc.* **84**, 460-466.

Kupper L. L., Portier C., Hogan M. D., Yamamoto, E. (1986). "The impact of litter effects on dose-response modeling in teratology." *Biometrics* **42**, 85-98.

Laird N. M. (1978). "Empirical Bayes methods for two-way contingency tables." *Biometrika* **65**, 581-590.

Lee, Y. and Nelder, J. A. (1996). "Hierarchical generalized linear models." With discussion. *J. Roy. Statist. Soc. Ser. B* **58**, 619-678.

Lopuhaä, H. P. (1989). "On the relation between S-estimators and M-estimators of multivariate location and covariance." *Ann. Statist.* **17**, 1662-1683.

Lopuhaä, H. P. and Rousseeuw, P. J. (1991). "Breakdown points of affine equivariant estimators of multivariate location and covariance matrices." *Ann. Statist.* **19**, 229–248.

Maronna, R. A. (1976). "Robust M-estimators of multivariate location and scatter." *Ann. Statist.* **4**, 51-67.

Maronna, R. A. and Morgenthaler, S. (1986). "Robust regression through robust covariances." *Comm. Statist. A–Theory Methods* **15**, 1347-1365.

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

Maronna, R. A., Yohai, V. J. and Zamar, R. H. (1993). "Bias-robust regression estimation: a partial survey." In *New Directions in Statistical Data Analysis and Robustness.* Morgenthaler, S., Ronchettii, E. and Stahel, W. A. eds. Birkhauser Verlag: Basel-Boston-Berlin.

Maronna, R. A. and Zamar, R. H. (2002). "Robust estimates of location and dispersion for high-dimensional datasets." *Technometrics* **44**, 307-317.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models.* London: Chapman and Hall.

McCulloch, C. E. (1997). "Maximum likelihood algorithms for generalized linear mixed models." *J. Amer. Statist. Assoc.* **92**, 162-170.

McGilchrist, C. A. (1994). "Estimation in generalized mixed models." *J. Roy. Statist. Soc. Ser. B* **56**, 61-69.

McGilchrist, C. A. and Yau, K. K. W. (1995). "The derivation of BLUP, ML, REML estimation methods for generalized linear mixed models." *Comm. Statist. Theory Methods* **24**, 2963-2980.

Merrill, H. M. and Schweppe, F. C. (1971). "Bad data suppression in power system static state estimation." *IEEE Trans. Power App. Syst.* **PAS-90**, 2718-2725.

Morton R. (1981). "Efficiency of estimating equations and the use of pivots." *Biometrika* **68**, 227-233.

Naylor, J. C. and Smith, A. F. M. (1982). "Applications of a method for the efficient computation of posterior distributions." *Appl. Statist.* **31**, 214-225.

Nelder, J. A. and Wedderburn, R. W. M. (1972). "Generalized linear models." *J. Roy. Statist. Soc. Ser. A* **135**, 370-384.

Neuhaus, W. W., Hauck, J. M. and Kalbfleisch, J. D. (1992). "The effects of mixture distribution misspecification when fitting mixed-effects logistic models." *Biometrika* **79**, 755-762.

Neuhaus, W. W. (2001). "Assessing change with longitudinal and clustered binary data." *Annu. Rev. Public Health.* **22**, 115-128.

Paul, S. R. and Plackett, R. L. (1978). "Inference sensitivity for Poisson mixtures." *Biometrika* **65**, 591-602.

Powell, J. L. (1983). "The asymptotic normality of two-stage least absolute deviations estimators." *Econometrica* **51**, 1569-1576.

Pregibon, D. (1981). "Logistic regression diagnostic." *Ann. Statist.* **9**, 705-724.

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

Prentice, R. L. (1988). "Correlated binary regression with covariates specific to each binary observation." *Biometrics* **44**, 1033-1048.

Ronchetti, E. and Rousseeuw, P. J. (1985). "Change-of-variance sensitivities in regression analysis." *Z. Wahrsch. Verw. Geb.* **68**, 503-519.

Rousseeuw, P. J. (1981). "A new infinitesimal approach to robust estimation." *Z. Wahrsch. Verw. Geb.* **56**, 127-132.

Rousseeuw, P. J. (1985). "Multivariate estimation with high breakdown point." In *Mathematical Statistics and Applications.* **B**, Grossmann, W., Pflug, G., Vincze, I. and Wertz, W. eds. Dordrecht, The Netherlands: Reidel, 283-297.

Rousseeuw, P. J. and Leroy, A. (1987). *Robust Regression and Outliers Detection.* New York: Wiley.

Rousseeuw, P. J. and Van Driessen, K. (2002). "Computing LTS regression for large data sets." *Estadistica* **54**, 163-190.

Ruppert, D. (1992). "Computing S-estimators for regression and multivariate location and dispersion." *J. Comput. Graph. Statist.* 1, 253-270.

Sargan, J. D. (1958). "The estimation of economic relationships using instrumental variables." *Econometrica* **26**, 393-415.

Schall, R. (1991). "Estimation in generalized linear models with random effects." *Biometrika* **78**, 719-727.

Shalabh, (1998). "Improved estimation in measurement error models through stein rule procedure." *J. Multivariate Anal.* **67** , 35-48.

Silvapulle, M. J. (1981) "On the existence of maximum likelihood estimators for the binomial response models." *J. Roy. Statist. Soc. Ser. B* **43**, 310-313.

Smith, A. F. M., Skene, A. M., Shaw, J. E. H., Naylor, J. C. and Dransfield, M. (1985). "The implementation of the Bayesian paradigm." *Comm. Statist. A-Theory Methods* **14**, 1079-1102.

Smith, P. J. and Weems, K. S. (2004). "On robustness of maximum likelihood estimates for Poisson-lognormal models." *Statistics and Probability Letters* **66**, 189-196.

Stahel, W. A. (1981). *Robust Estimation: Infinitesimal Optimality and Covariance Matrix Estimators.* PhD. Thesis, ETH, Zurich.

Stefanski, L. A., Carroll, R. J., and Ruppert, D. (1986). "Optimally bounded score functions for generalized linear regression models with applications to logistic regression." *Biometrika* **73**, 413-424.

Stigler, S. N. "Gauss and the invention of least squares." *Ann. Statist.* **9**, 465-474.

Stiratelli, R., Laird, N. and Ware, J. H. (1984). "Random-effects models for serial observations with binary response." *Biometrics* **40**, 961-971.

Venables, W. N. and Ripley, B. D. (1999). *Modern Applied Statistics with S-Plus.* (3rd ed.). New York: Springer.

Wagenvoort, R. and Waldmann, R. (2002). "On B-robust instrumental variable estimation of the linear model with panel data." *Journal of Econometrics* **106**, 297-324.

Wedderburn, R. W. M. (1974). "Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method." *Biometrika* **61**, 439-448.

Wedderburn, R. W. M. (1976). "On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models." *Biometrika* **63**, 27-32.

White, H. A. (1982). "Maximum likelihood estimation of misspecified models." *Econometrica* **50**, 1-25.

Williams, D. A. (1982). "Extrabinomial variation in logistic linear models." *Appl. Statist.* **31**, 144-148.

Yau, K. K. W. and Kuk, A. Y. C. (2002). "Robust estimation in generalized linear mixed models." *J. Roy. Statist. Soc. Ser. B Stat. Methodol.* **64**, 101-117.

Zamar, R. H. (1989). "Robust estimation in the errors-in-variables model." *Biometrika* **76**, 149-160.