Robust predictive distribution and bias-calibration in Linear Models

Branca, Mattia

Robust Predictive Distribution and Bias-Calibration in Linear Models

by

Mattia Branca

A thesis submitted to the
Geneva School of Economics and Management,
University of Geneva, Switzerland,
in fulfillment of the requirements for the degree of
PhD in Statistics

Members of the thesis committee:
Prof. Davide La Vecchia, Chair, University of Geneva
Prof. Eva Cantoni, Co-adviser, University of Geneva
Prof. Elvezio Ronchetti, Co-adviser, University of Geneva
Prof. Davide Ferrari, University of Bolzano

Thesis No. 74
December 2019

La Faculté d'économie et de management, sur préavis du jury, a autorisé l'impression de la présente thèse, sans entendre, par-là, émettre aucune opinion sur les propositions qui s'y trouvent énoncées et qui n'engagent que la responsabilité de leur auteur.

Genève, le 8 janvier 2020

Dean
Marcelo OLARREAGA

Impression d'après le manuscrit de l'auteur.

Acknowledgements

First of all, I would like to thank Professor Eva Cantoni and Professor Elvezio Ronchetti for giving me this great opportunity and allowing me to achieve this work with their help, support, patient guidance and availability during the thesis period.

I would also like to thank Professor Davide La Vecchia for accepting to be the president of the thesis committee and for giving constructive and detailed comments on the manuscript. Additionally, I thank Professor Davide Ferrari that accepted to be the external member of the jury and revised and critically judged this work during the private defense.

I am also very grateful for the opportunity of being a teaching assistant during my PhD years. It was a very rewarding experience in particular the chance to work with Dr. Ilir Roko and Professor Diego Kuonen for over four years.

Special thanks to all my colleagues that I met during my studies in Geneva. It was really efficient and fun to study during the Master, work and enjoy student life with people like Marie, Julien, Sanda, Antonio, Laura and Roberto. Bonds that developed over the years in families and lifelong friendships.

During the PhD years, I had the chance to learn and swap ideas with many PhD colleagues to which I am grateful for their support, in particular Laura, Linda, Mark, Pierre-Yves, Daniel, Marco and many others. A sincere thank you goes to Setareh, my officemate, that helped me a lot to go through this academic journey. I am likewise grateful to Samuel and Haotian for all the badminton matches and all their help and encouragements for this thesis. It was a pleasure to have some recreation time with the "dîner de cons" team, born during the coffee breaks on the third floor of Unimail with Kush, Marc-Olivier, Samuel, Elise, Roberto, Sezen and Rose.

I am particularly grateful for the patience, help and advices of Roberto that was always available to offer great suggestions, critical feedbacks and be very supportive during all these years.

Furthermore, I would also like to thank my colleagues at CTU Bern. Since I have started working at CTU Bern, I enjoyed the positive and rewarding environment we created in our offices. Many thanks to Poorya, Arnaud, Enrico, Sylvain, Dik, Mikael, Andi, Mark, Sheila, Sereina, (Ciao) Phil, Sarah, Selina, Muriel, Lea and Stefan.

Likewise I extend my thanks to my longstanding friends from Ticino (Dingo, Berto, Yle, Fabio, Chicco, Dem) for all their moral support during these years.

Finally, I would like to express my deep gratitude to my parents, grandmother and sister for all their support for my studies and career choices. I would also like to extend

thanks to my in laws for all their encouragement along the way. I am profoundly grateful to my wife Sanda for her love, encouragement and serenity during all this period to help me succeed. Last, but not least, a remarkable emotional support was offered by my two little dogs, Pinga and Pingu.

Abstract

This thesis focuses on the concept of predictive distributions and bias calibration. At first, an extension of the concept of predictive distributions under contamination is studied in the case of Generalized Linear Models. A sensitivity analysis of the impact of contamination on the predictive distribution is studied making use of the class of M-estimators. In a second step, based on the available literature on bias-calibrated estimation in linear regression, the bases to implement bias-calibration for predictive distributions are studied and developed. This development is based on the finite-sample setting and an important aspect of the reasoning behind this contribution is the distinction between representative and non-representative outliers. As a result of this distinction, the use of the bias-calibration approach allows to integrate information from representative outliers within the predictive distribution.

Résumé

Cette thèse se concentre autour du concept de distributions prédictives et de la calibration du biais. Dans un premier temps, le concept de distributions prédictives pour les Modèles Linéaires Généralisés est étendu dans le cadre où les observations sont sujettes à contamination. Une analyse de sensibilité de la contamination sur la distribution prédictive est étudiée en utilisant la classe des M-estimateurs. Dans un deuxième temps, la calibration du biais pour ces distributions prédictives est développée sur la base de la littérature disponible dans le cas de la régression linéaire. Ce développement est basé sur le cadre de l'échantillon fini et un aspect important du raisonnement derrière cette contribution est la distinction entre les valeurs aberrantes représentatives et non représentatives. En raison de cette distinction, l'utilisation de l'approche par étalonnage des biais permet d'intégrer des informations provenant de valeurs aberrantes représentatives dans la distribution prédictive.

Contents

To my family.

Introduction

Statistical analysis has different purposes that can non-exhaustively be summarized in the tasks of description and prediction. While description aims at detecting patterns and variables that significantly contribute to the explanation of particular phenomena, prediction focuses on estimating and inferring on unobserved values of this phenomena based on a chosen model. To date, prediction has focused mainly on point forecasting, which, although accompanied by confidence intervals, is not able to offer the same level of inference that is available for model parameter estimation. For this reason, we are currently witnessing a change in the way predictions are obtained. In fact, the common approach based on point forecasting is being replaced by the so-called probabilistic forecasting, where the entire distribution of the prediction is required in order to make inference on predictions and forecasts; see Gneiting and Katzfuss [2014]. The distribution of the predictions is commonly known as "predictive distribution" and it is defined as the distribution of a future random variable from the same model; see e.g. Geisser [1971], Aitchison [1975], Harris [1989] and Basu and Harris [1994].

Within this growing field of statistical research, there is only a marginal level of attention directed towards the impact that contaminated data can have on the predictive distribution. Indeed, it is widely known that statistical models are at best an approximation of reality and that observations contain almost certainly an amount of data which does not come from the distribution of the bulk of the data (i.e. contaminated data). In terms of distribution, we can say that there is a small proportion of the data which comes from an arbitrary distribution. In these situations, it is widely known that even a small proportion of contaminated observations can considerably influence the estimation procedure leading to biased parameter estimates and testing. It is reasonable to think that a predictive distribution can also be influenced by contamination, thereby delivering an inaccurate framework for inference on predictions.

The statistical field that focuses on these situations is robust statistics. A lot of research has already been done in this branch of statistics. The seminal papers are Tukey [1960], Huber [1964] and Hampel [1968]. Book-length presentations are Huber [1981] (with the second edition by Huber and Ronchetti [2009]), and Hampel et al. [1986], where the focus is centered on the methodology based on influence functions. Furthermore, an overview of robustness in linear regression can be found in Maronna et al. [2006]. A more recent book that covers a literature review of newly developed robust methods is Heritier et al. [2009].

The concept of predictive distribution covered in this thesis follows the work of Harris [1989] and Basu and Harris [1994] where the integral over the sampling distribution is used to obtain the predictive distribution, i.e. they estimate the sampling distribution of the parameters and use it in the formula for the predictive distribution. These authors

estimate the parameters using different methods where, among others, they use a robust estimator such as the Minimum Hellinger Distance Estimator. The analysis of these particular cases was done for the Binomial, the Normal and the Poisson distributions in the univariate case with the aim to compare the results obtained against the true distribution by using the Kullback-Leibler divergence (see Kullback and Leibler [1951]). On the other hand, Geisser [1971] suggests deriving the form of the predictive distribution of a future observation using a Bayesian framework underlining the essential fact that a prediction cannot be based only on plugging-in the estimated parameters in the estimated model or by using the posterior distribution of the parameter set. The difference between Geisser's predictive approach and Harris [1989], consists in obtaining the predictive distribution by taking the expectation of the distribution function of the observations over the prior distribution of the parameters and not over the distribution of the estimator.

Overview of this Thesis

Nowadays, there is a high demand for statistical tools that are essential in forecasting. The presence of distortion in the data can influence the results and induce unreliable predictions that can consequently give inaccurate interpretations. That is why a tool to treat and reduce the impact of this distortion is highly suitable in such situations and robust statistics is used in this context. Robustness plays a key role in this and the focus on it is fundamental to the end results. The main terminology and definitions of robustness and predictive distribution are explained in Chapter 1 as these are prerequisites to the subsequent chapters.

One of the purposes of the thesis is to further extend the predictive distribution described by Aitchison [1975] and Harris [1989] where the predictive distribution has been derived in the univariate setting. In this thesis, we extend the predictive distribution to the general case of M-estimators and we generalize it to a multivariate setting for both linear regression and Generalized Linear Models (GLM) class (see Nelder and Wedderburn [1972] and McCullagh and Nelder [1989]) in the presence of an arbitrary contamination. Therefore, it is necessary to investigate the robustness properties of the estimated predictive distribution and to analyze the behavior of the predictive distribution in the presence of deviations from the assumed model. To address robustness, we build on the results presented in Cantoni and Ronchetti [2001], where the robust estimation and inference are developed for the Generalized Linear Models settings. In the latter paper, M-estimators are proposed for the entire GLM class and their properties are derived.

As mentioned, the first step is the derivation of the predictive distribution keeping a general multivariate setting. We then perform a sensitivity analysis of the robust predictive distribution for the specific case of GLM when the underlying distribution is contaminated. The goal is to understand the advantages and disadvantages of computing the predictive distribution based on M-estimators that bound the impact of model deviations. We build our predictive distribution based on the existing work for robust estimation in this setting, see Cantoni and Ronchetti [2001], Lô and Ronchetti [2009] and Heritier et al. [2009]. Furthermore, a focus on the numerical methods to solve the problem of multiple integration when obtaining the final predictive distribution is also discussed and analyzed in Chapter 2. The results show the high sensitivity of the classical methods in GLM.

It is known that a robust estimator guarantees a bounded bias when there is a contamination in the data. But, as underlined by Chambers [1986], the fact that there is the presence of representative outliers can introduce a bias in the robust estimation, due to the fact that these representative outliers are down-weighted. A representative outlier can be defined as an outlying observation relevant in the sample and that cannot be considered as incorrect. Consequently, these observations are presumed to be important to describe the data. That is why we apply in Chapter 3 a bias-corrected estimator for the particular case of linear models, which was developed for the estimation of total population and quantile functions in survey for finite populations in Welsh and Ronchetti [1998]. In fact, the main contribution of this thesis focuses on the bias-calibrated estimator applied to the predictive distribution in the context of linear regression. The main idea of bias-calibration comes from Chambers [1986] where the author introduces this type of estimators and he focuses on the concept of representative and non-representative outliers. In fact, it is based on the work of bias-calibration of Chambers [1986] that Welsh and Ronchetti [1998] concentrate their work on sample survey containing outliers.

In this thesis, we develop a method to select the value of the calibration tuning constant that allows to get a result that is less biased compared to the predictive distribution based on the robust estimator and less variable with respect to an unbounded estimator (minimization of the Mean Squared Error), such as the Maximum Likelihood estimator.

To summarize, in Chapter 1, a brief theoretical review of both robustness and predictive distribution is covered. Chapter 2 presents the sensitivity analyses of the predictive distribution in the GLM setting. In Chapter 3 we develop the Mean Squared Error of the predictive distribution applied to the bias-calibrated estimator for linear regression.

Chapter 1

Some Basic Concepts of Robust Statistics and Prediction

In this chapter, we first briefly introduce the main concepts of robustness that are presented in the literature in Section 1.1. The review of the theory covers the topics that are important for this thesis, all other important concepts are available in the references made. Secondly, in Section 1.2, a brief overview of the literature and the important concept of predictive distributions is covered in order to introduce the main contribution of this work.

1.1 General Overview of Robust Statistics

Robust statistics, as it is known nowadays, was mainly introduced and discussed in the work of Tukey [1960], Huber [1964] and Hampel [1968]. The book by Huber [1981] covers the mathematical background behind the concept of robustness, where an updated and expanded version can be found in Huber and Ronchetti [2009]. Since then many authors have developed new methodological tools in this field. An important reference that delivers a general overview can be found in Hampel et al. [1986] which covers robust statistics based on the approach of the influence function. Other authors contributed to the development of robust statistics in different fields and an example is given by Maronna et al. [2006] where a wide description of the work in robustness applied to regression is given. Another example is Heritier et al. [2009] where the topic of robust statistics is tackled within the perspective of Biostatistics. In the latter book the authors give a general overview of the work that has been developed in different fields such as Mixed Linear Models and Generalized Linear Models.

Considering the wide extension of the literature about robust statistics, in the next paragraphs we focus on basic concepts that will be used in this thesis. We give the definition of some important concepts in robust statistics and, for this purpose, we follow the notation in Heritier et al. [2009]. The underlying idea to robust statistics consists in the assumption that the data is issued from a miss-specified model, e.g. some observations are considered as coming from a different generating distribution. This concept is formalized by assuming that the data generating process is

$$F_{\epsilon,\theta} = (1-\epsilon)F_\theta + \epsilon H \ , \tag{1.1}$$

where F_θ corresponds to the assumed model for the majority of the data, θ is a set of parameters, ϵ is the amount of contamination and H is an unknown contaminating

distribution. The goal is to make inference about F_θ, even though the sample comes from $F_{\epsilon,\theta}$. A special case is when the contaminating distribution H is a point mass distribution at $\tilde{z}$. In this case, $H = \Delta_{\tilde{z}}$, and $\tilde{F}_{\epsilon,\theta} = (1-\epsilon)F_\theta + \epsilon\Delta_{\tilde{z}}$. where $\Delta_{\tilde{z}}$ is the point mass distribution. The point mass distribution is relevant to introduce some important concepts in robustness in the following paragraphs.

Key Concepts

There are different measures that are used to understand how a contamination influences the estimation. A first measure is the so-called sensitivity curve (SC, Hampel et al. [1986]). The SC allows to evaluate the robustness of an estimator by measuring the effect of an observation $\tilde{z}$ on the estimator of interest, say T_n, in finite samples. The SC is defined as

$$\mathrm{SC}(\tilde{z}; y_1, \ldots, y_{n-1}, T_n) = n\left[T_n(y_1, \ldots, y_{n-1}, \tilde{z}) - T_{n-1}(y_1, \ldots, y_{n-1})\right] ,$$

where y_i are observations coming from the underlying distribution F_θ. In this way, it is possible to show the impact of an extreme value on T_n.

The limitation of the SC is that it applies only to finite samples. In fact, Hampel [1968] and Hampel [1974] introduced the concept of Influence Function (IF) to overcome this difficulty. In the case of the IF we define the general functional as T. The IF can be generally interpreted as the asymptotic version of the sensitivity curve, although this is not necessarily always the case. The IF measures the impact of an infinitesimal contamination at a point $\tilde{z}$ on the asymptotic bias. More specifically, it describes the normalized influence on the estimation of an infinitesimal observation at $\tilde{z}$ and it gives information about the robustness (stability) and efficiency properties of the functional T under contamination, when the contaminated distribution is a point mass distribution $\Delta_{\tilde{z}}$. We can see the IF as the Gâteaux derivative of T at F_θ (in the direction of $\Delta_{\tilde{z}}$) that is

$$\mathrm{IF}(\tilde{z}; T, F_\theta) = \lim_{\epsilon \to 0} \frac{T((1-\epsilon)F_\theta + \epsilon\Delta_{\tilde{z}}) - T(F_\theta)}{\epsilon} ,$$

or alternatively,

$$\mathrm{IF}(\tilde{z}; T, F_\theta) = \frac{\partial}{\partial\epsilon} T((1-\epsilon)F_\theta + \epsilon\Delta_{\tilde{z}})\Big|_{\epsilon=0} = \frac{\partial}{\partial\epsilon} T(\tilde{F}_{\epsilon,\theta})\Big|_{\epsilon=0} .$$

A bounded IF guarantees that an estimator is robust. Generally speaking, the IF measures the asymptotic bias of T caused by an infinitesimal deviation from the postulated model:

$$\mathrm{bias}(T, F_\theta, \epsilon) = \sup_H \|T((1-\epsilon)F_\theta + \epsilon H) - T(F_\theta)\| \simeq \epsilon \mathrm{GES}(T, F_\theta) , \tag{1.2}$$

where $\mathrm{GES}(T, F_\theta) = \sup_{\tilde{z}} \|\mathrm{IF}(\tilde{z}; T, F_\theta)\|$ is the gross-error sensitivity, sup is the supremum and $\|\cdot\|$ is the Euclidean norm. From (1.2) we can say that when the IF is bounded, consequently, the bias of the estimator is bounded.

The derivative of the (asymptotic) variance of an estimator, when considering the point mass distribution as contamination, gives the so-called Change-of-Variance function (CVF). The CVF measures how an amount of contamination at $\tilde{z}$ can influence the asymptotic variance of the estimator. Its form is as follows:

$$\frac{\partial}{\partial\epsilon} V(T; \tilde{F}_{\epsilon,\theta})\Big|_{\epsilon=0} = \mathrm{CVF}(\tilde{z}; T, F_\theta) ,$$

where $V(T; \tilde{F}_{\epsilon,\theta})$ is the (asymptotic) variance of an estimator.

The properties of the asymptotic variance of an estimator are measured by the change-of-variance sensitivity, that is $k^\star = \sup_{\tilde{z}} \operatorname{tr}\{\mathrm{CVF}(\tilde{z}; T, F_\theta)\} / \operatorname{tr}\{V(T; F_\theta)\}$, where $k^\star$ measures the worst variability change under an infinitesimal contamination. We have that an estimator is defined as V-robust if the value of $k^\star$ is finite, meaning that the CVF is bounded. The property of V-robustness for an estimator is stronger than that of B-robustness (bias robust) that occurs when the IF is bounded (Hampel et al. [1986]).

Robust Estimation

In robust statistics, a general class of estimators that is widely used are the so-called M-estimators. Let's assume we have $y_1, \ldots, y_n$ i.i.d. observations coming from F_θ. The definition of a M-estimator is a minimization problem, such as

$$\min_\theta \sum_{i=1}^{n} \rho(y_i; \theta) \ ,$$

or, alternatively, as the solution for θ of

$$\sum_{i=1}^{n} \Psi(y_i; \theta) = 0 \ ,$$

for suitable $\rho(\cdot)$ and $\Psi(\cdot)$ functions where $\Psi(y_i; \theta) = \partial \rho(y_i; \theta)/\partial\theta$. As an example, if we define $\rho = -\log(f_\theta)$ we get the particular case of the maximum likelihood estimator, where f_θ is the density function of F_θ. For a M-estimator, the functional is defined such that

$$T(F) : \mathrm{E}_F\left[\Psi(y; T(F))\right] = 0 \ . \tag{1.3}$$

which explicitly depends on the Ψ-function. Also, in the particular case of M-estimators, the IF is given by

$$\mathrm{IF}(\tilde{z}; \Psi, F_\theta) = M(\Psi, F_\theta)^{-1} \Psi(\tilde{z}, T) \ , \tag{1.4}$$

where $M(\Psi, F_\theta)$ is a matrix of dimensions $q \times q$ defined as

$$M(\Psi, F_\theta) = -\int \frac{\partial}{\partial t} \Psi(y; t)\Big|_{t=\theta} dF_\theta(y) \ , \tag{1.5}$$

where q is the dimension of the parameter vector $\theta \in \mathbb{R}^q$. From (1.4) we see that the IF of M-estimators is proportional to the Ψ-function. Consequently, if Ψ is bounded, the IF of the estimator will also be bounded.

We know that under some regularity conditions on Ψ (Huber [1981]), we get that $\hat{\theta}$ is asymptotically normal, such that

$$\sqrt{n}(\hat{\theta} - \theta) \to N\Big(0, V(\Psi; F_\theta)\Big) \ ,$$

where the asymptotic variance of M-estimators can be written as follows

$$V(\Psi, F_\theta) = M^{-1}(\Psi, F_\theta) Q(\Psi, F_\theta) M^{-\top}(\Psi, F_\theta) \ , \tag{1.6}$$

where $M(\Psi, F_\theta)$ is defined in (1.5) and $Q(\Psi, F_\theta)$ is defined as

$$Q(\Psi, F_\theta) = \int \Psi(y; \theta) \Psi(y; \theta)^\top dF_\theta(y) \ . \tag{1.7}$$

For large n the distribution $G(t; F_\theta(y))$ of M-estimator is given approximately by

$$N\left(T(F_\theta), \frac{1}{n}V(\Psi; F_\theta)\right) , \tag{1.8}$$

with density function

$$g(t; F_\theta(y)) = (2\pi)^{-\frac{q}{2}}\left|\frac{1}{n}V(\Psi; F_\theta)\right|^{-\frac{1}{2}} \exp\left\{-\frac{n}{2}(t - T(F_\theta))^\top V(\Psi; F_\theta)^{-1}(t - T(F_\theta))\right\} . \tag{1.9}$$

The weights that are explicitly or implicitly assigned to observations via the bounded Ψ-function defining M-estimators can affect the true underlying distribution of the scores (residuals) thereby introducing bias in the parameter estimation (for example when the underlying distribution is asymmetric). A property which is therefore desirable for an M-estimator is for it to be Fisher consistent such that, when applied to the true distribution, the M-estimator will return the desired parameter value (see e.g. Huber [1967]). Indeed, Fisher consistency is satisfied if

$$\int \Psi(y; \theta) dF_\theta(y) = 0 .$$

For example, a correction factor is often included in the Ψ-function for this purpose, as in the case of most Generalized Linear Models.

We conclude this section by presenting two of the most common functions that have been proposed and that are useful to obtain the Ψ function defined above. We show here Huber's and Tukey's biweight ψ functions that are defined respectively as:

$$\psi_{[Hub]}(s; c) = \begin{cases} s & \text{for } |s| \leq c \\ c\,\text{sign}(s) & \text{for } |s| > c \end{cases} , \tag{1.10}$$

and

$$\psi_{[bi]}(s; c) = \begin{cases} \left(\left(\frac{s}{c}\right)^2 - 1\right)^2 s & \text{for } |s| \leq c \\ 0 & \text{for } |s| > c \end{cases} ,$$

where $c \in \mathbb{R}^+$ is a tuning constant and $s \in \mathbb{R}$ and where $\psi_{[Hub]}$ (or $\psi_{[bi]}$) will be used in the Ψ function defined in Section 2.1, formula (2.1). The value of c defines the degree of robustness of the estimator and consequently its efficiency. The lower the value of c, the less efficient the estimator will be. The trade-off between robustness and efficiency is one of the crucial points for this type of estimators.

1.2 Prediction and Predictive Distribution

The concept of prediction is widely considered in modern statistics. The book by Hastie et al. [2009], in particular in Chapters 2 and 3, provides an analysis of the concept of prediction. Statistical forecasting is present in everyday life and precise prediction can become fundamental in certain settings, e.g. meteorological forecasts and medical settings. More technically, the prediction of a future observation is strongly based on the inference on the collected data. Indeed, various types of model inference are present in the literature and the corresponding prediction can be defined in two ways. The first one is called point prediction which refers to the forecasts made based on inference coming from a model estimated from the data thereby providing an estimated value for a future observation and its confidence interval (see Gneiting [2011]).

The second type of prediction consists in probabilistic forecasting (or predictive distribution). In this case the aim is to obtain the distribution of future observations. Furthermore, proceeding in this way, we are able to obtain a better idea of the entire distribution form for a future set of data that we observe. In particular, it is possible to see how the inference of the model influences the entire distribution, as opposed to influencing only a single point, like in point forecasting.

Classical Methods of Estimation

The predictive density of a future random variable Z observed at z is

$$f_\theta^\star(z) = \int f_t(z) dG(t; F_\theta(z)) \,, \tag{1.11}$$

where f_t is the density of the observation z and G is the distribution of the estimator $\hat{\theta}$. Let's suppose we have n observations $y_1, \dots, y_n$ that are distributed $y_i \sim F_\theta$ with density function f_θ. As previously introduced, in this specific context we consider estimators belonging to the family of M-estimators.

The definition of predictive distribution in (1.11) is based on the defintion of bootstrap predictive distribution given in Harris [1989] and Basu and Harris [1994]. Indeed, in these two papers the authors propose to estimate the predictive distribution by using three different approaches: estimative, bootstrap and Bayesian. The estimative approach consists in estimating the unknown parameters based on the data via Maximum-Likelihood (ML) estimation to then plug the estimator (and its asymptotic distribution) in the predictive distribution. On the other hand, the bootstrap method that the authors develop consists in obtaining the sampling distribution G as the distribution of the estimated parameters where the latter is obtained via parametric bootstrap (based on which the predictive distribution is successively obtained). As a result, this predictive distribution is called the bootstrap predictive distribution that is represented in (1.11) and which represents the starting point for this work. Furthermore, in the second paper, the authors compute the predictive distribution by replacing the ML estimator with the Minimum Hellinger Distance Estimator (MHDE) due to the fact that when outliers are present in the data, the ML may perform poorly. In this case, the derivation of the predictive distribution is obtained by replacing the parameter estimated via ML with the parameter estimated via MHDE within G (and this can be done considering the asymptotic equivalence of the two different estimations).

The third approach consists in a Bayesian approach where prior information on the unknown parameter θ would be used where, in this case, G would be considered as a cumulative posterior distribution in order to obtain the predictive distribution. In fact, we can also see equation (1.11) in a Bayesian framework, where $f^\star$ is the predictive distribution obtained via Bayes theorem and where G is the cumulative posterior distribution.

Chapter 2

Sensitivity Analysis of the Predictive Distribution in (Generalized) Linear Models

In the first section of Chapter 2, we briefly define the GLM setting. We focus on the basic terms that are necessary to build our final version of the predictive distribution adapted to the specific case of GLM. For an overview we consider Heritier et al. [2009] and Cantoni and Ronchetti [2001]. In the second section, we derive the predictive distribution in this setting. Finally, we perform a simulation study to analyse the impact of the contamination on the predictive distribution in the GLM (Poisson) setting.

2.1 Generalized Linear Models

Generalized Linear Models (GLM) are a class of models that go beyond normal response variables, allowing for discrete and continuous distributions in the response variable. Indeed, the need for a link function to link the predictors and the responses is fundamental in this type of models.

GLM were introduced by Nelder and Wedderburn [1972] and a complete overview is given by McCullagh and Nelder [1989]. Other interesting books on the subject are: Dobson and Barnett [2008] who gives a detailed introduction to GLM and discusses the most common distributions belonging to the exponential family; Faraway [2016] covers some applications on real data using `R`.

The theory of GLM is built upon the exponential family which includes common distributions such as Normal, Binomial, Poisson, Gamma and Exponential. Considering n independent random variables, $y_1, \ldots, y_n$ are said to follow an exponential family if their density or probability mass function can be written as

$$f(y_i; x_i, \tau_i, \phi) = \exp\left[\frac{y_i\tau_i - b(\tau_i)}{d_i(\phi)} + j(y_i, \phi)\right] ,$$

where $d_i(\cdot)$, $b(\cdot)$ and $j(\cdot)$ are some specific functions, τ_i is a function of μ_i (that depends on x_i that are the fixed covariates) called the natural parameter and $\phi \in \mathbb{R}^+$ is an additional scale (or dispersion) parameter. The expectation and the variance are denoted by

$$\mu_i = \mathrm{E}\left[y_i\right] \quad \text{and} \quad \mathrm{Var}(y_i) = \phi v_{\mu_i} ,$$

where $\upsilon_{\mu_i} \in \mathbb{R}^+$ depends on the distributional assumption on $y_i|x_i$.

In the GLM setting, the link function defines the relationship between the mean of the response variable and the assumed linear predictor. The linear predictor is $\eta_i = x_i^\top \beta$, where $x_i^\top = (1, x_{i1}, x_{i2}, \ldots, x_{i\tilde{q}})$ are $\tilde{q}$ explanatory variables for each individual $i = 1, \ldots, n$ and $\beta^\top = (\beta_0, \beta_1, \ldots, \beta_{\tilde{q}})$ is a set of $q = \tilde{q}+1$ parameters. The link function $l(\mu_i)$ is defined as

$$l(\mu_i) = \eta_i = x_i^\top \beta \ ,$$

which is a monotone function linking the random and the systematic components of the model. Moreover, it defines the form of the relationship between the mean μ_i of the response variable and η_i, the linear predictor. The natural link function directly relates the natural parameter to the linear predictor,

$$\tau_i = \eta_i = x_i^\top \beta \ ,$$

and we can write that the expectation of y_i as $\mathrm{E}[y_i] = \mu_i = l^{-1}(x_i^\top \beta)$.

GLM are usually estimated by ML. Here we also introduce the robust GLM estimators of Cantoni and Ronchetti [2001] (see also Cantoni and Ronchetti [2006]) upon which we will build our robust predictive distribution. The M-estimator is the solution of the following estimating equations

$$\sum_{i=1}^{n} \Psi(y_i, x_i; \beta, \phi, c) = 0 \ ,$$

where β is a set of parameters, $x_i^\top$ are the covariates, ϕ is the dispersion parameter and c is the robustness tuning constant. Furthermore,

$$\Psi(y_i, x_i; \beta, \phi, c) = \frac{\psi(r_i)}{\sqrt{\phi \upsilon_{\mu_i}}} w(x_i) \mu_i' - a(\beta) \ , \tag{2.1}$$

where $r_i = (y_i - \mu_i)/\sqrt{\phi \upsilon_{\mu_i}}$ are the Pearson residuals and $\psi(r_i)$ refers to the type of function that we want to apply to the residuals. In our case we refer to the Huber-function defined in (1.10). Furthermore,

$$a(\beta) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}_{F_\beta} \left[\frac{\psi(r_i)}{\sqrt{\phi \upsilon_{\mu_i}}} \right] w(x_i) \mu_i' \ , \tag{2.2}$$

is the correction factor to ensure Fisher consistency and the expectation is taken over the conditional distribution of $y_i|x_i$. The term $w(x_i)$ represents the weights on the design matrix X that here are considered equal to 1 due to the fact that we do not consider outliers in the x-space. The design matrix X (of dimension $n \times q$) is

$$X = \begin{pmatrix} x_1^\top \\ x_2^\top \\ \vdots \\ x_n^\top \end{pmatrix} \ .$$

Nevertheless, if we do not consider contamination in the x-space the weights are put equal to 1. Moreover, $\mu_i = \mu_i(\beta) = l^{-1}(x_i^\top \beta)$ and $\mu_i' = \frac{\partial}{\partial \beta} \mu_i$.

2.2 Robust Predictive Distribution

We derive the predictive distribution under contamination. We start by defining all the elements necessary to obtain its final derivation. Afterwards, we tackle in detail all the computations and derivations that are useful to understand all the steps.

2.2.1 Derivation of the Predictive Distribution under Contamination

The predictive distribution of a random variable Z is defined in (1.11), where z is a realization of Z. In the following, we consider the predictive distribution of Z that depends also on the covariates x, such that we have $f^{\star}_{\theta}(z,x)$, evaluated in (z,x). Furtermore, let us suppose that the assumed model is contaminated, as presented in (1.1). In the context of this definition, here we assume that H is absolutely continuous w.r.t. Lebesgue measure and Radon-Nikodym derivative h, meaning that the class of contamination is restricted. In this new case, the predictive distribution becomes

$$\begin{aligned} f^{\star}_{\epsilon,\theta}(z,x) &= \int f_{\epsilon,t}(z,x)dG(t;F_{\epsilon,\theta}(z,x)) \\ &= \int \Big[(1-\epsilon)f_t(z,x)+\epsilon h(z,x)\Big]g(t;F_{\epsilon,\theta}(z,x))dt\ . \end{aligned} \tag{2.3}$$

From the latter expression, we notice that the predictive distribution depends on $F_{\epsilon,\theta}$, and consequently on h. An approximation to $f^{\star}_{\epsilon,\theta}(z,x)$ by Taylor expansion around $\epsilon=0$ gives:

$$f^{\star}_{\epsilon,\theta}(z,x) \cong f^{\star}_{\theta}(z,x) + \epsilon \frac{\partial}{\partial \epsilon} f^{\star}_{\epsilon,\theta}(z,x)\Big|_{\epsilon=0}\ . \tag{2.4}$$

From the latter expression, we see that the predictive distribution depends on the non-contaminated predictive distribution according to the distribution of the postulated model (first term on the right hand side of equation (2.4)) and on the level of the contamination (through the second term in the right hand side of equation (2.4)). The level of the contamination is defined as the percentage of the contaminated data points. We see below the details of this second term, by analyzing each element, to understand its impact on equation (2.4) and consequently on the predictive distribution.

From expression (2.4) we develop the derivative of $f^{\star}_{\epsilon,\theta}(z,x)$ w.r.t to ϵ evaluated at 0. Therefore, we rewrite the second term of equation (2.4) as

$$\frac{\partial}{\partial \epsilon} f^{\star}_{\epsilon,\theta}(z,x)\Big|_{\epsilon=0} = \frac{\partial}{\partial \epsilon}\left[\int \left((1-\epsilon)f_t(z,x)+\epsilon h(z,x)\right) g(t;F_{\epsilon,\theta}(z,x))dt\right]_{\epsilon=0}\ , \tag{2.5}$$

where the density $g(t;F_{\epsilon,\theta}(z,x))$ is given by equation (1.9) with F_{θ} replaced by $F_{\epsilon,\theta}(z,x)$.

The analysis of expression (2.5) will provide insights on the robustness properties of the predictive distribution. More specifically, we want to identify the elements that can influence the predictive distribution in order to reduce the impact of the observations presumably coming from the unknown distribution H. We will study the approximation (2.4) including (2.5).

To study expression (2.5), we divide the integral in the following way

$$\frac{\partial}{\partial\epsilon}(1-\epsilon)\Big|_{\epsilon=0}\int f_t(z,x)g(t;F_\theta(z,x))dt+\int f_t(z,x)\frac{\partial}{\partial\epsilon}g(t;F_{\epsilon,\theta}(z,x))dt\Big|_{\epsilon=0}+ \\ +\ \frac{\partial}{\partial\epsilon}\epsilon\ \Big|_{\epsilon=0}h(z,x)\int g(t;F_\theta(z,x))dt\ . \tag{2.6}$$

According to expression (2.6), we rewrite the complete form of the predictive distribution as

$$f^\star_{\epsilon,\theta}(z,x)\ \cong\ f^\star_\theta(z,x)+ \tag{2.7}$$
$$+\epsilon\left[-f^\star_\theta(z,x)+\int f_t(z,x)\frac{\partial}{\partial\epsilon}g(t;F_{\epsilon,\theta}(z,x))\Big|_{\epsilon=0}dt+h(z,x)\right].$$

In expression (2.7) we need to compute the derivative of $g(t;F_{\epsilon,\theta}(z,x))$ w.r.t ϵ, evaluated at $\epsilon=0$, which is:

$$\frac{\partial}{\partial\epsilon}g(t;F_{\epsilon,\theta}(z,x))\Big|_{\epsilon=0}=\left(\frac{2\pi}{n}\right)^{-\frac{q}{2}}\frac{\partial}{\partial\epsilon}\Big[p_1(\epsilon)\exp\{p_2(\epsilon)\}\Big]_{\epsilon=0}\ . \tag{2.8}$$

where

$$p_1(\epsilon)=\Big|V(\Psi;F_{\epsilon,\theta})\Big|^{-\frac{1}{2}}\quad\text{and}\quad p_2(\epsilon)=-\frac{n}{2}(t-T(F_{\epsilon,\theta}))^\top V(\Psi;F_{\epsilon,\theta})^{-1}(t-T(F_{\epsilon,\theta}))\ ,$$

such that,

$$\begin{aligned}&\frac{\partial}{\partial\epsilon}g(t;F_{\epsilon,\theta}(z,x))\Big|_{\epsilon=0}\\ &=\ -\frac{n^{\frac{q}{2}}}{2(2\pi)^{\frac{q}{2}}}\Big|V(\Psi;F_\theta)\Big|^{-\frac{3}{2}}\left[\ \mathrm{tr}\left(\mathrm{adj}(V(\Psi;F_\theta))\frac{\partial}{\partial\epsilon}V(\Psi;F_{\epsilon,\theta})\Big|_{\epsilon=0}\right)\right]\\ &\quad\exp\left\{-\frac{n}{2}(t-T(F_\theta))^\top V(\Psi;F_\theta)^{-1}(t-T(F_\theta))\right\}+\\ &\quad-\frac{n^{1+\frac{q}{2}}}{2(2\pi)^{\frac{q}{2}}}\Big|V(\Psi;F_\theta)\Big|^{-\frac{1}{2}}\Bigg\{-2\frac{\partial}{\partial\epsilon}T(F_{\epsilon,\theta})^\top\Big|_{\epsilon=0}V(\Psi;F_\theta)^{-1}(t-T(F_\theta))+\\ &\quad-(t-T(F_\theta))^\top\left[V(\Psi;F_\theta)^{-1}\frac{\partial}{\partial\epsilon}V(\Psi;F_{\epsilon,\theta})\Big|_{\epsilon=0}V(\Psi;F_\theta)^{-1}\right](t-T(F_\theta))\Bigg\}\\ &\quad\exp\left\{-\frac{n}{2}(t-T(F_\theta))^\top V(\Psi;F_\theta)^{-1}(t-T(F_\theta))\right\}\ .\end{aligned} \tag{2.9}$$

The details of the derivation of expression (2.8) and (2.9) are shown in Appendix A.1.

We finally summarize the result of the predictive distribution in (2.7). To obtain it, we need to consider $f^\star_\theta(z,x)$, $h(z,x)$ and the derivative of the density of the M-estimator, that can be found in equation (2.9). From equation (2.7) we can notice how the contamination ϵ influences the predictive distribution by multiplying the second part of the expression on the right hand side of (2.7). In particular, the derivatives of the functional and the variance of the estimator play an important role (see definitions of IF and CVF in Section 1.1), as seen in (2.9).

Considering the predictive distribution function in further detail, we notice from expression (2.7) that an important impact of the contamination is coming from the term

$$\int f_t(z,x)\frac{\partial}{\partial\epsilon}g(t;F_{\epsilon,\theta}(z,x))\Big|_{\epsilon=0} dt\ , \tag{2.10}$$

that is the derivative of the density function of the M-estimator.

The stability and robustness of the predictive distribution highly depend on the derivative of the functional and the derivative of the variance thereby suggesting that an estimator with bounded Ψ-function and bounded CVF is needed to obtain a robust predictive distribution.

Special Case when $H(z,x) = \Delta_{(\tilde{z},\tilde{x})}(z,x)$

Until now we have assumed that contamination is delivered by an unknown distribution $H(z,x)$. However, we could consider a special case where this distribution can be defined as a point mass function such that we have $H(z,x) = \Delta_{(\tilde{z},\tilde{x})}$, where $\Delta_{(\tilde{z},\tilde{x})}$ is the point mass in $(\tilde{z},\tilde{x})$ (point mass distribution).

The IF is defined as the derivative of the general functional T when using the point mass distribution (see Chapter 1). On the other hand, the derivative of the variance, given in the Appendix A.1, when plugging-in $H(z,x) = \Delta_{(\tilde{z},\tilde{x})}(z,x)$ gives us the so-called Change-of-Variance Function. The definition of the CVF has been introduced in Chapter 1. Here, we refer to the CVF that has been derived in Ferrari and La Vecchia [2012], in Section 3.2 p. 240-241, and also a special case is covered in Zhelonkin et al. [2012], on p. 729. In both cases - IF and CVF-, the integral over the distribution H becomes a point evaluated at $\tilde{z}$, that is

$$\int \frac{\partial}{\partial t}\Psi(z;t)\Big|_{t=T(F_\theta)} dH(z) = \frac{\partial}{\partial t}\Psi(\tilde{z};t)\Big|_{t=T(F_\theta)}\ . \tag{2.11}$$

When $H(z,x) = \Delta_{(\tilde{z},\tilde{x})}(z,x)$, the form of the predictive distribution becomes

$$\begin{aligned} f^\star_{\epsilon,\theta}(z,x;\tilde{z},\tilde{x}) \ &\cong\ f^\star_\theta(z,x) + \\ &+\epsilon\left[-f^\star_\theta(z,x) + \int f_t(z,x)\frac{\partial}{\partial\epsilon}g(t;\tilde{F}_{\epsilon,\theta}(z,x))\Big|_{\epsilon=0} dt + \Delta_{(\tilde{z},\tilde{x})}(z,x)\right]. \end{aligned}$$

2.3 Computation of the Predictive Distribution

To implement the results obtained on the predictive distribution, we will use the statistical software `R`. The main package to carry out robust GLM estimation is `robustbase`. One difficulty in computing the predictive distribution in equation (2.7) is the computation of multiple integrals. From the form of the expression of the predictive distribution in expression (2.7) we see that it is possible to apply a Laplace approximation to compute this type of multiple integrals. In fact, the Laplace approximation is

$$\tilde{I} = \underbrace{\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty}}_{q-\text{times}} e^{-\xi m(t)}dt \approx e^{-\xi m(\hat{t})}\left(\frac{2\pi}{\xi}\right)^{q/2}|\Sigma|^{1/2}\ , \tag{2.12}$$

where

$$\Sigma = \left(\frac{\partial^2 m(\hat{t})}{\partial t^2} \right)^{-1} ,$$

with q the dimension of t, the element to be integrated, ξ being the sample size (e.g. see Shun and McCullagh [1995]) and $m(t)$ a twice-differentiable function. Knowing that our predictive distribution is based on distributions coming from the exponential family, we remark that it is possible to use the Laplace approximation expressed in (2.12). The closed form expression to be minimized by Newton-Raphson and used in the Laplace approximation for the Poisson GLM can be found in Appendix B.1. The Laplace approximation obtained in equation (2.12) will be used as a term of comparison with other approaches in the simulation study.

In order to find a faster implementation we used the package called `TMB` also made available in the `R`-environment. This package implements efficiently a Laplace approximation by automatic differentiation to compute multiple integrals. The `TMB` package details can be found in Kristensen et al. [2015]. In the simulation study, we will consider the three different numerical solutions mentioned which we can compare in order to obtain the robust predictive distribution.

A second alternative to compute multiple integrals consists in applying the Monte Carlo approximation. The main issue concerning the Monte Carlo approximation regards the large number of repetitions that are necessary to get the numerical results. A comparison between the three approaches is presented in Appendix B.2

2.4 Simulation Study for the Poisson GLM

In this section, we present a simulation study in which we analyze the behavior of the predictive distribution under contamination. We first introduce the setting of the Poisson GLM, that is our distribution of interest for this study. The goal is twofold: show whether and how a robust estimator coupled with the predictive distribution gives a more reliable result of the forecasting distribution for a future observation and, concurrently, study the impact of different levels of contamination.

For a Poisson GLM we have the observations $y_i \sim P(\mu_i)$ and the probability mass function f_β is

$$f_\beta(y_i, x_i) = \bar{f}_\beta(y_i|x_i)k_n(x_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}k_n(x_i) .$$

In this case, our parameter of interest is $\theta = \beta$. We have $\mathrm{E}[y_i] = \mathrm{Var}[y_i] = \mu_i$ for $i = 1, \ldots, n$, where $\beta \in \mathbb{R}^q$ is the vector of parameters, $x_i \in \mathbb{R}^q$, and $l(\mu_i)$ is the log-link function. Finally, in this context, we denote the distribution of the covariates x (defined as $K(x)$) as the empirical distribution $K_n(x)$, with $k_n(x_i)$ being its empirical density function. In the following paragraphs, we assume that the joint distribution $F_\theta(y, x)$ of (Y,x) can be rewritten as $\bar{F}_\theta(y|x)K(x)$.

We consider the following data generating mechanism in the simulations

$$f_{\epsilon,\beta}(y, x) = (1-\epsilon)f_\beta(y, x) + \epsilon h(y, x) = (1-\epsilon)\frac{e^{-\mu}\mu^y}{y!}k_n(x) + \epsilon\frac{e^{-\tilde{\mu}}\tilde{\mu}^y}{y!}k_n(x) ,$$

where $\tilde{\mu}$ is the parameter of the contaminating distribution (Poisson). The choice of the robust M-estimator was introduced in Section 2.1 and we use the Huber ψ-function.

2.4.1 Simulation Setting

To show the impact of the contamination on the predictive distribution and to compare the different choices of tuning constants a simulation study is set up as follows: the response vector of size $n = 1000$ is simulated from a Poisson distribution with mean parameter $\mu_i = \exp(x_i^\top \beta)$ where $\beta = [1.2, -0.6, 0.4, -0.7, 0.3]$, for $i = 1, \ldots, n$. The matrix of the explanatory variables is defined as follows: x_{i1} is 1, $x_{i2} \sim Bin(2, 0.5)$, $x_{i3} \sim N(0, 1)$, $x_{i4} \sim Gamma(1)$ and $x_{i5} \sim Bin(2, 0.1)$. The values of β and x are chosen such that we get the overall mean close to $\mu = 1.4$. Finally, we fix a vector $x = [1, 2, 1, 0.5, 1]$ for the future observation z. The mean of the contaminated distribution is fixed to $\tilde{\mu} = 6$, chosen to be about four times the mean of μ. The results are generated over 200 replications. The variables that we use and analyze in different scenarios during the simulation study are the following:

- The amount of the contamination ϵ. We analyze different amounts of contamination to study the impact of the distortion on the predictive distribution.
- We use two values of the tuning constant c to determine the degree of robustness. In our case, we compare a robust M-estimator ($c = 1.345$, chosen to deliver a reasonable level of efficiency) versus a non-robust M-estimator ($c = 10$).
- A further element is the sample size. We briefly mention the differences when the sample size is larger, in particular focusing on the impact that the sample size has on the computation time of the predictive distribution.
- We also consider different values of the dimension q (the number of explanatory variables). The aim is to observe if there is a difference when q is changing.
- The last comparison performed is between the three different approaches of estimation: the two approaches using the Laplace approximation and the Monte Carlo approximation (with `1e6` replications).

2.4.2 Results of the Simulation Study

In this section, we present the results of the simulations. Different schemes are analyzed, in which we study the behavior of the robust predictive distribution for Poisson GLM.

We begin the sequence of results by considering $q = 5$. In this study, we analyze the impact of the contamination and the difference when using the robust M-estimator versus the MLE (or non-robust M-estimator). Two graphs are presented here (see Figure 2.1 and Figure 2.2) to show the results of the simulation when the level of the contamination is $\epsilon = 0.05$. Successively, the results of other schemes with different levels of contamination are shown in two tables (Table 2.1 and Table 2.2). In Table 2.1 we study the predictive distributions, while in Table 2.2, the results of the Kullback-Leibler divergence (K-L) are analyzed.

In Figure 2.1, we see how the contamination causes a distortion on the predictive distribution of a future observation z compared to $f^\star_\beta(z, x)$. We notice how the predictive

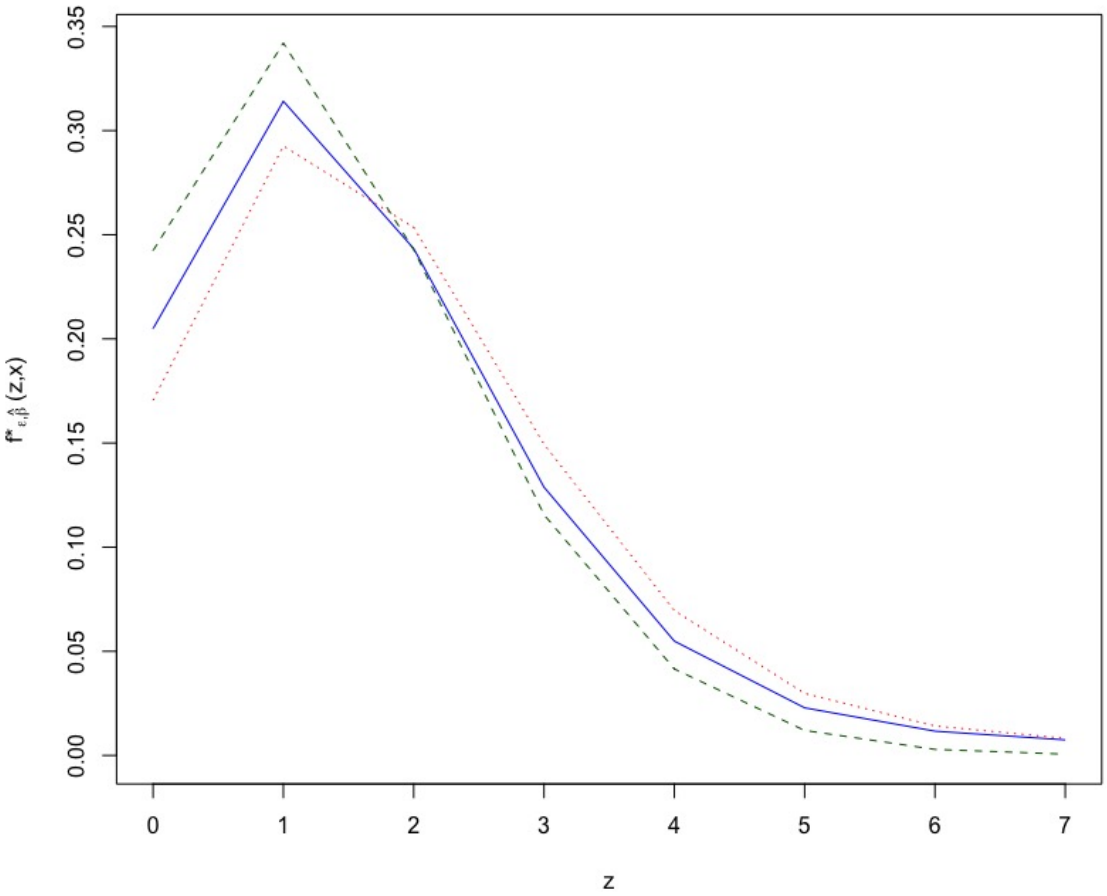

Figure 2.1: Estimation of the predictive distribution $f^{\star}_{\epsilon,\hat{\beta}}(z,x)$ for $\epsilon = 0.05$. The dashed-line (green) is the predictive distribution when using the true set of parameters β and $\epsilon = 0$. The continuous line (blue) represents the median of 200 replications of the estimated predictive distribution when using the estimated parameters with $c = 1.345$. Finally, the dotted line (red) represents the median of 200 replications of the estimated predictive distribution when using $c = 10$. The sample size is $n = 1000$.

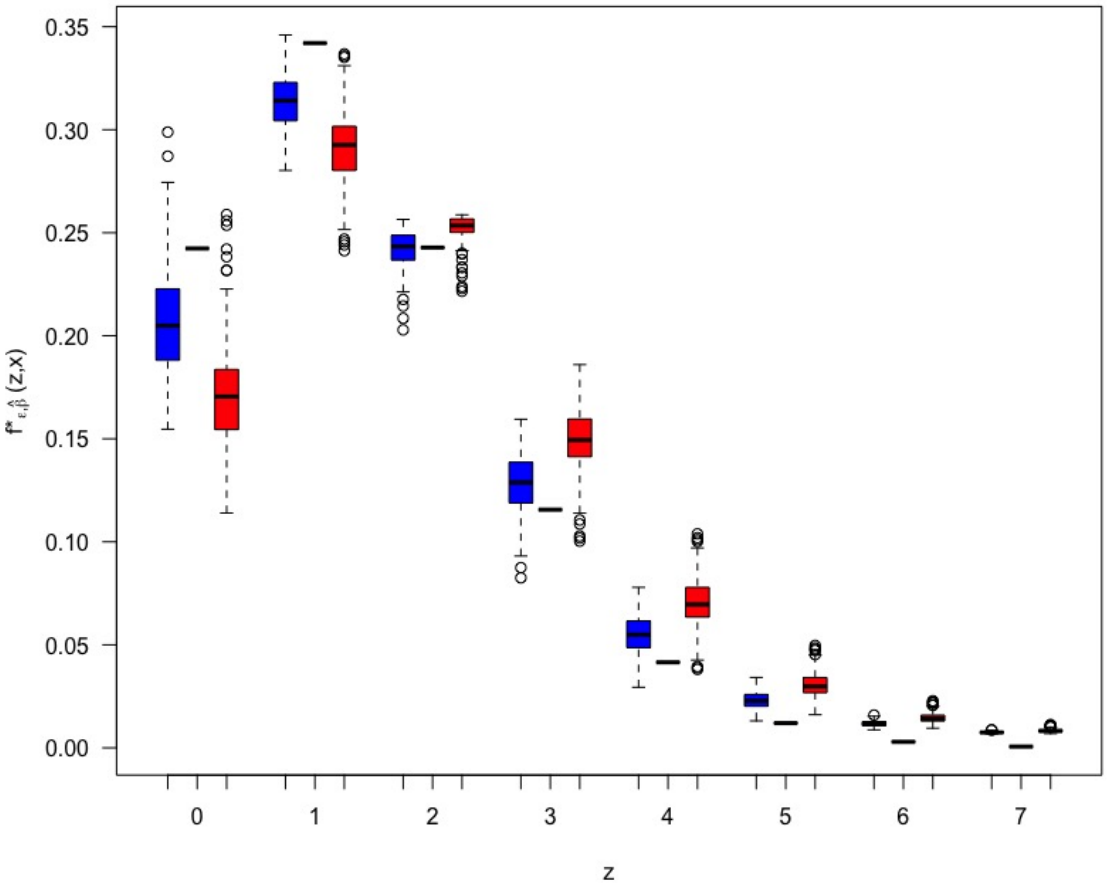

Figure 2.2: Boxplots for $z = 0, \ldots, 7$ of 200 replications of $f^{\star}_{\epsilon,\hat{\beta}}(z, x)$ obtained with a robust M-estimator with $c = 1.345$ (in blue) and a non-robust M-estimator with $c = 10$ (in red). The dashed line represents the predictive distribution when using the true set of parameters β and $\epsilon = 0$. The simulation setting is: $\epsilon = 0.05$ and $n = 1000$.

distribution based on the robust M-estimator (blue) gives a less distorted curve with respect to the reference (green-dashed line), compared to the MLE. This fact confirms the importance of using a robust M-estimator with a bounded Ψ-function. Moreover, we see that the estimated predictive distribution based on M-estimators is closer to $f^{\star}_{\beta}(z, x)$ when c is smaller. In Figure 2.2, we have the results of the 200 replications of $f^{\star}_{\epsilon,\hat{\beta}}(z, x)$ for $\epsilon = 0.05$ and for the two M-estimators. Generally, we remark the difference between the two levels of c. Furthermore, the variability when $c = 10$ is higher, in particular for the values of $z = 4, 5, 6$. We notice that the curves tend to shift towards the right, having lower probability values for $z < 2$ and higher for $z \geq 2$ for c increasing.

We now analyze three different levels of contamination, $\epsilon = 0, 0.05, 0.1$ for the two M-estimators, robust and non-robust. In Table 2.1, we see the relationship between the results of the predictive distribution, the level of the contamination and the level of c. In fact, the predictive distribution is strongly dependent on the value of c, i.e. the degree of robustness of the M-estimator. Thus, the more robust the estimator, the less distorted the results. In addition, the results in Table 2.1 are confirmed by those in Table 2.2.

z	$f^{\star}_{\beta}$	$f^{\star}_{0,\hat{\beta}_{1.345}}$	$f^{\star}_{0,\hat{\beta}_{10}}$	$f^{\star}_{0.05,\hat{\beta}_{1.345}}$	$f^{\star}_{0.05,\hat{\beta}_{10}}$	$f^{\star}_{0.1,\hat{\beta}_{1.345}}$	$f^{\star}_{0.1,\hat{\beta}_{10}}$
0	24.19	24.11	23.89	20.49	17.05	17.52	11.10
1	34.33	34.15	34.08	31.41	29.26	28.86	23.35
2	24.36	24.33	24.43	24.34	25.35	24.13	24.78
3	11.52	11.63	11.73	12.88	14.94	13.95	18.06
4	4.09	4.19	4.25	5.49	6.96	6.71	10.39
5	1.16	1.22	1.24	2.29	2.98	3.31	5.39
6	0.27	0.30	0.30	1.16	1.41	1.98	2.87
7	0.06	0.06	0.06	0.75	0.82	1.38	1.69
8	0.01	0.01	0.01	0.51	0.53	0.97	1.06

Table 2.1: Comparison of $f^{\star}_{\epsilon,\hat{\beta}}(z, x)$ for different values of c and ϵ. The first column represents the value of z. The second column shows the values of the predictive distribution when β is known and $\epsilon = 0$. The third, fifth and seventh columns represent the results when $c = 1.345$ and respectively $\epsilon = 0, 0.05$ and 0.10. The fourth, sixth and eighth columns represent the results for $c = 10$. The sample size is $n = 1000$ and the values represented are the medians of 200 replications and are multiplied by 100.

	$\epsilon = 0$	$\epsilon = 0.05$	$\epsilon = 0.1$
$c = 1.345$	0.00	8.53	19.75
$c = 10$	0.00	11.94	35.04

Table 2.2: Kullback-Leibler divergence comparison between $f^{\star}_{\epsilon,\hat{\beta}}(z, x)$ and $f^{\star}_{\beta}(z, x)$ for different combinations of c and ϵ and sample size $n = 1000$. The values are multiplied by 100.

Briefly, we comment on the sample size: the smaller the sample size, the higher the variability of the M-estimators and, consequently, of the predictive distribution. On the other hand, when using a bigger sample size, more time is necessary to perform the simulation.

Analysis of the Predictive Distribution

We hereby present a study in which we observe the impact of the contamination in detail. The study shows that the main difference between $f^{\star}_{\beta}(z,x)$ and $f^{\star}_{\epsilon,\hat{\beta}}(z,x)$ results primarily from the amount of contamination ϵ, rather than the size of the sample, where $\hat{\beta}$ is estimated considering the non-robust estimation. The setting of the simulation is the same as introduced before. This section is structured in three different parts, divided as follows:

1. In the first part we focus on the general form of the predictive distribution $f^{\star}_{\epsilon,\hat{\beta}}(z,x)$, and we look at $f^{\star}_{\epsilon,\hat{\beta}}(z,x) - f^{\star}_{\beta}(z,x)$.

2. The second term of the analysis is the difference between $f^{\star}_{\epsilon,\hat{\beta}}(z,x) - f^{\star}_{\epsilon,\beta}(z,x)$. The focus is on the impact that n and ϵ have on the estimation of the parameters and consequently, on the predictive distribution.

3. The last result concerns the difference between $f^{\star}_{\epsilon,\beta}(z,x) - f^{\star}_{\beta}(z,x)$, where the only element of interest is the effect of ϵ on the predictive distribution.

The first part is the sum of the other two. This decomposition is useful to analyze the different effects (of ϵ and n) in the predictive distribution.

Part I

In this part we focus on $f^{\star}_{\epsilon,\hat{\beta}}(z,x) - f^{\star}_{\beta}(z,x)$. In Figure 2.3 we notice a clear trend caused by the contamination: when increasing the contamination, the differences increase. To further explain, the estimated predictive distribution for $z = 0, 1, 2$ has an under estimation compared to $f^{\star}_{\beta}(z,x)$, while the successive values increase (over estimation). Table 2.3 and Table 2.4 represent the numerical results of the medians, summarizing the results in the graph. Considering that the graphs and tables for the sample size $n = 100$ show the same effect as $n = 1000$, we show in this Chapter only the results of $n = 1000$.

Size	$\epsilon = 0$	$\epsilon = 0.05$	$\epsilon = 0.10$
$n = 100$	0.19	10.36	27.36
$n = 1000$	0.01	9.29	25.20

Table 2.3: Kullback-Leibler divergence comparison between the median of 200 replications of $f^{\star}_{\epsilon,\hat{\beta}}(z,x) - f^{\star}_{\beta}(z,x)$ for three different values of contamination, $\epsilon = 0, 0.05, 0.1$, and two sample sizes, $n = 100, 1000$. The values are multiplied by 100.

Part II

The second part concentrates on $f^{\star}_{\epsilon,\hat{\beta}}(z,x) - f^{\star}_{\epsilon,\beta}(z,x)$. The results are shown in Figure 2.4 and in Table 2.5.

Part III

The third part focuses on the difference $f^{\star}_{\epsilon,\beta}(z,x) - f^{\star}_{\beta}(z,x)$, and the results are shown in Table 2.6 and Table 2.7.

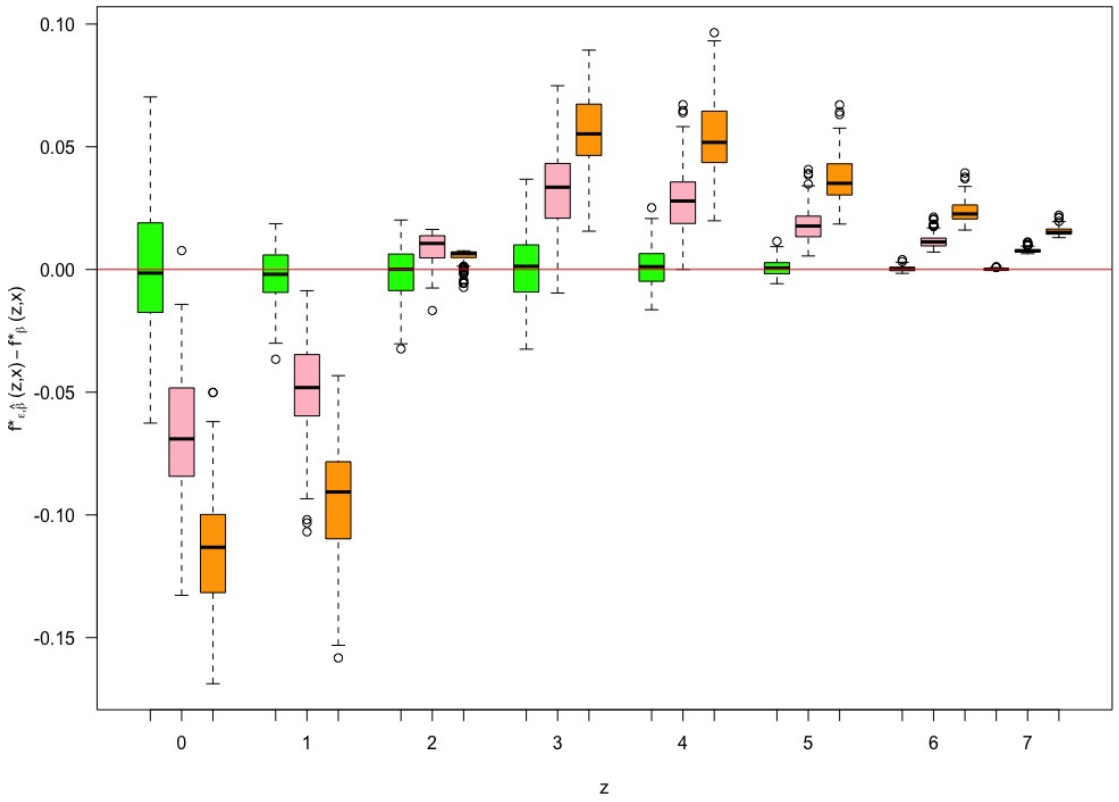

Figure 2.3: Boxplots of the differences between the estimated predictive distribution $f^{\star}_{\epsilon,\hat{\beta}}(z,x)$ and $f^{\star}_{\hat{\beta}}(z,x)$ for three different levels of contamination and values of z between 0 and 7. The first boxplots (in green) correspond to $f^{\star}_{\epsilon,\hat{\beta}}(z,x)$ when $\epsilon = 0$. The second boxplots (in pink) correspond to $\epsilon = 0.05$ and finally the last $\epsilon = 0.1$. The sample size is $n = 1000$ and the replications are 200.

z	$\epsilon = 0$	ϵ=0.05	ϵ=0.1
0	-0.15	-6.9	-11.32
1	-0.2	-4.81	-9.07
2	0.01	1.06	0.65
3	0.13	3.35	5.53
4	0.12	2.79	5.18
5	0.06	1.77	3.51
6	0.02	1.12	2.26
7	0.01	0.76	1.51

Table 2.4: Median of 200 replications of $f^{\star}_{\epsilon,\hat{\beta}}(z,x) - f^{\star}_{\beta}(z,x)$ for each point z between 0 and 7 with sample size of $n = 1000$. The values are multiplied by 100.

Size	$\epsilon = 0$	$\epsilon = 0.05$	$\epsilon = 0.10$
$n = 100$	0.19	2.40	6.77
$n = 1000$	0.01	2.31	6.62

Table 2.5: Kullback-Leibler divergence comparison between the median of 200 replications of $f^{\star}_{\epsilon,\hat{\beta}}(z,x) - f^{\star}_{\epsilon,\beta}(z,x)$ for two different levels of contamination, $\epsilon = 0.05, 0.1$, and two sample sizes, $n = 100, 1000$. The values are multiplied by 100.

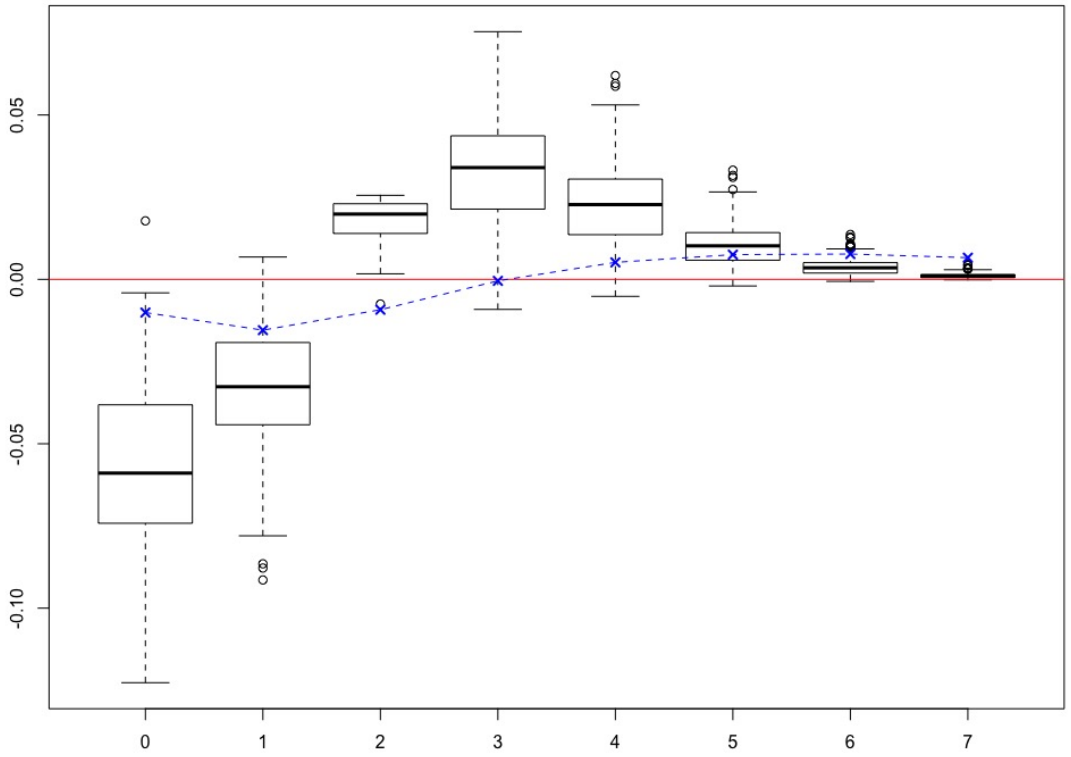

Figure 2.4: Boxplots of $f^{\star}_{\epsilon,\hat{\beta}}(z,x) - f^{\star}_{\epsilon,\beta}(z,x)$ for $\epsilon = 0.05$, $n = 1000$ and 200 replications. The (blue) crosses represent the median of $f^{\star}_{\epsilon,\beta}(z,x) - f^{\star}_{\beta}(z,x)$.

The comparison between Part II and Part III shows that the impact on the predictive distribution of ϵ is relevant on the predictive distribution itself (i.e. Part III) and is

	$\epsilon = 0$	$\epsilon = 0.05$	$\epsilon = 0.10$
K-L	0.00	5.04	12.24

Table 2.6: Kullback-Leibler divergence comparison between the median of 200 replications of $f^{\star}_{\epsilon,\beta}(z,x) - f^{\star}_{\beta}(z,x)$ for two levels of contamination, $\epsilon = 0.05, 0.1$. The values in the table are multiplied by 100.

z	$\epsilon = 0$	$\epsilon = 0.05$	$\epsilon = 0.1$
0	0.05	-1.01	-1.97
1	-0.12	-1.54	-2.85
2	-0.08	-0.92	-1.70
3	0.03	-0.05	-0.12
4	0.06	0.52	0.94
5	0.04	0.75	1.41
6	0.02	0.77	1.46
7	0	0.66	1.26

Table 2.7: Median of 200 replications of $f^{\star}_{\epsilon,\beta}(z,x) - f^{\star}_{\beta}(z,x)$ for each point z between 0 and 7. The values in the table are multiplied by 100.

even more important when it also affects the estimation of the M-estimators (i.e. Part II, joint impact of n and ϵ). The contamination causes a distortion on the predictive distribution (see (2.7)) and the estimator, while the impact of n is relevant concerning the variability of the M-estimators. Indeed, when using a small number of observations, the predictive distribution is affected by the high variability of the results of the estimation. In Figure 2.4, we notice the difference in the tail of the distribution that is coming entirely from ϵ. The numerical results are shown in Table 2.5-2.7.

2.4.3 Discussion

The main conclusion that we draw from this sensitivity study is that the impact of the M-estimator is crucial on the result of the predictive distribution. We see that when using a robust M-estimator the predictive distribution is less distorted, compared to the non-robust M-estimator and consequently, closer to the distribution of the assumed model f (see Section 2.4.2). In this work, we have defined the general form of the predictive distribution and we applied it, as an example, to GLM and we have looked at the particular instance of Huber-function applied to the Poisson GLM. Further work to compare different distributions and/or different estimators can be attained by modifying these formulas for any specific context.

Chapter 3

Robust Predictive Distribution and Bias-Calibration for Linear Models

In Chapter 2 we performed the sensitivity analysis of the predictive distribution for M-estimators in the particular case of GLMs. Based on this we observed that robust M-estimators can down-weight the influence of outliers in order to reduce their influence on the resulting estimated predictive distribution. Nevertheless, this down-weighting procedure can deliver bias if (some) outliers can be considered as "representative" outliers. This phenomenon has been underlined by Chambers [1986] where, for this reason, a bias-calibrated estimator was developed to reduce the bias of the robust estimator by calibrating on the importance of the representative outliers. In Welsh and Ronchetti [1998] the bias-calibrated estimator was used to estimate the quantile and the total population in finite survey samples for linear models. In this chapter, based on the results of the sensitivity analysis in the previous chapter, the objective is to to obtain a predictive distribution by making use of this bias-calibrated estimator. In this thesis, considering the technicality and complexity of the development of the predictive distribution and the notion of bias-calibration, the development of the predictive distribution is based on the bias-calibrated estimator in the particular case of linear models. This development will allow us to set the basis for future work in the direction of predictive distributions based on bias-calibrated estimators in the GLM case.

3.1 The Bias-Calibration Estimator Applied to the Predictive Distribution

As introduced at the beginning of this thesis, the idea is to apply the bias-calibration estimator to the predictive distribution in linear regression. This comes from the fact that we would like to correct the bias of the robust estimator that performs better compared to a non-robust estimator, but it is biased in the context of sample survey containing representative outliers that are relevant observations and not errors (as described by Chambers [1986]). Furthermore, an overview of the most important points about bias-calibration is given in Appendix C.1.

For the linear regression model, when considering the Normal distribution, the parameter θ corresponds to the set of parameters (β, σ). In our case, to derive the predictive distribution in Section 3.2.1, we consider σ as known and we write $f^{\star}_{\epsilon,\theta}(z,x) = f^{\star}_{\epsilon,\beta}(z,x)$.

3.2 The Bias-Calibrated Estimator

To derive the bias-calibrated estimator we consider the regression model

$$Y_i = X_i\beta + \sigma e_i, \qquad i = 1, \dots, N \ , \tag{3.1}$$

where e_i are *iid* random variables, β unknown parameters, σ a known parameter (replaced by the value of $\hat{\sigma}_R$) and we consider the errors with expectation 0 and variance σ_e^2. The bias-calibration estimator for linear regression (see Chambers [1986] and Welsh and Ronchetti [1998]) takes the following form:

$$\hat{\beta}_{cal}(c_2) = \hat{\beta}_R + \left(\sum_{i=1}^{n} x_i x_i^\top \right)^{-1} \sum_{i=1}^{n} \hat{\sigma}_R \psi_{c_2}\{(y_i - x_i^\top \hat{\beta}_R)/\hat{\sigma}_R\} x_i \ , \tag{3.2}$$

where $\hat{\beta}_R$ is a robust M-estimator of β. The function ψ_{c_2} is a Huber-function defined by the constant c_2 greater than the constant used to obtain $\hat{\beta}_R$. Considering that with the calibration the aim is to include more information coming from the outliers, the value of c_2 is set to be greater than the constant to obtain $\hat{\beta}_R$, called c_1, because it should not put a weight on the calibration that is stronger than the weight given by the M-estimator. Finally, the value of $\hat{\sigma}_R$ is obtained using the MAD (Median Absolute Deviation, see Huber and Ronchetti [2009]) in order to bound the impact of the outliers also in the estimation of $\hat{\sigma}_R$.

3.2.1 Properties of the Predictive Distribution

In this section, we focus on the properties of the predictive distribution in order to study the impact of the bias-calibrated estimator. In this way, we show that depending on the contamination, a different value of c_2 can be selected in order to obtain the best trade-off between variability and bias. To do the latter, we first compute the variance and the bias of the predictive distribution. When computing the MSE of the predictive distribution, we consider the value of σ as known. In practice, we plug-in the value of $\hat{\sigma}_R$ for σ.

Computation of the MSE of the Predictive Distribution

The predictive distribution for a future observation is

$$f_\beta^\star(z, x) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f_t(z, x) g(t; F_\beta(z, x)) dt \ , \tag{3.3}$$

where

$$f_t(z, x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(z - x^\top t)^2 \right) ,$$

is the distribution of the response, and

$$g(t; F_\beta(z, x)) = \frac{1}{(2\pi)^{q/2}} \left| \frac{1}{n} V(\Psi; F_\beta) \right|^{-1/2} \exp\left(-\frac{n}{2}(t - T(F_\beta))^\top V(\Psi; F_\beta)^{-1}(t - T(F_\beta)) \right) ,$$

is the asymptotic density function of the M-estimator, where $T(F_\beta)$ and $V(\Psi; F_\beta)$ are defined in Chapter 1, in (1.3) and (1.6) respectively.

We have seen in Chapter 2 (see section 2.3) that a way to approximate the multiple integration in $f^{\star}_{\beta}(z,x)$ is to use the Laplace approximation. Therefore, we apply the same approach presented in Appendix B.1 for the Poisson GLM. Its form is

$$
\begin{aligned}
f^{\star}_{\beta}(z,x) \;\approx\; & \frac{1}{\sigma\sqrt{2\pi}}\Big|V(\Psi;F_{\beta})\Big|^{-1/2}\Big|\Big[\frac{1}{n\sigma^{2}}xx^{\top}+V(\Psi;F_{\beta})^{-1}\Big]^{-1}\Big|^{1/2} \\
& \exp\Bigg(-n\Big[\underbrace{\frac{1}{2n\sigma^{2}}(z-x^{\top}\hat{t})^{2}+\frac{1}{2}(\hat{t}-T(F_{\beta}))^{\top}V(\Psi;F_{\beta})^{-1}(\hat{t}-T(F_{\beta}))}_{m(\hat{t})}\Big]\Bigg)\,,
\end{aligned}
\tag{3.4}
$$

where $\hat{t}=\operatorname{argmin}_{t} m(t)$. In order to compute the MSE of the predictive distribution, we first find the variance, which is given by the following approximation

$$
\operatorname{Var}_{\hat{\beta}}(f^{\star}_{\beta}(z,x)) \approx \frac{\partial f^{\star}_{\beta}(z,x)}{\partial \hat{t}^{\top}} V(\Psi;F_{\beta}) \frac{\partial f^{\star}_{\beta}(z,x)}{\partial \hat{t}}\,, \tag{3.5}
$$

where $V(\Psi;F)$ is the variance of the corresponding estimator $\hat{\beta}_{cal}$ and

$$
\begin{aligned}
\frac{\partial f^{\star}_{\beta}(z,x)}{\partial \hat{t}^{\top}} = & \frac{1}{\sigma\sqrt{2\pi}}\Big|V(\Psi;F_{\beta})\Big|^{-1/2}\Big|\Big[\frac{1}{n\sigma^{2}}xx^{\top}+V(\Psi;F_{\beta})^{-1}\Big]^{-1}\Big|^{1/2} \\
& \Bigg(\Big[\frac{1}{\sigma^{2}}(z-x^{\top}\hat{t})x^{\top}-n(\hat{t}-T(F_{\beta}))^{\top}V(\Psi;F_{\beta})^{-1}\Big]\Bigg) \\
& \exp\Bigg(-n\Big[\frac{1}{2n\sigma^{2}}(z-x^{\top}\hat{t})^{2}+\frac{1}{2}(\hat{t}-T(F_{\beta}))^{\top}V(\Psi;F_{\beta})^{-1}(\hat{t}-T(F_{\beta}))\Big]\Bigg)\,.
\end{aligned}
\tag{3.6}
$$

The expected value of the predictive distribution with respect to the distribution of the M-estimator is given by the following approximation

$$
\begin{aligned}
& \mathrm{E}_{\hat{\beta}}\Big[f^{\star}_{\beta}(z,x)\Big] \approx \mathrm{E}_{\hat{\beta}}\Bigg[\frac{1}{\sigma\sqrt{2\pi}}\Big|V(\Psi;F_{\beta})\Big|^{-1/2}\Big|\Big[\frac{1}{n\sigma^{2}}xx^{\top}+V(\Psi;F_{\beta})^{-1}\Big]^{-1}\Big|^{1/2} \\
& \exp\Bigg(-n\Big[\frac{1}{2n\sigma^{2}}(z-x^{\top}\hat{t})^{2}+\frac{1}{2}(\hat{t}-T(F_{\beta}))^{\top}V(\Psi;F_{\beta})^{-1}(\hat{t}-T(F_{\beta}))\Big]\Bigg)\Bigg] \\
& = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{1}{\sigma\sqrt{2\pi}}\Big|V(\Psi;F_{\beta})\Big|^{-1/2}\Big|\Big[\frac{1}{n\sigma^{2}}xx^{\top}+V(\Psi;F_{\beta})^{-1}\Big]^{-1}\Big|^{1/2} \\
& \exp\Bigg(-n\Big[\frac{1}{2n\sigma^{2}}(z-x^{\top}t_{E})^{2}+\frac{1}{2}(t_{E}-T(F_{\beta}))^{\top}V(\Psi;F_{\beta})^{-1}(t_{E}-T(F_{\beta}))\Big]\Bigg) \\
& \frac{1}{(2\pi)^{q/2}}\Big|\frac{1}{n}V(\Psi;F_{\beta})\Big|^{-1/2}\exp\Big(-\frac{n}{2}(t_{E}-T(F_{\beta}))^{\top}V(\Psi;F_{\beta})^{-1}(t_{E}-T(F_{\beta}))\Big)dt_{E} \\
& = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{1}{\sigma\sqrt{2\pi}}\Big|V(\Psi;F_{\beta})\Big|^{-1/2}\Big|\Big[\frac{1}{n\sigma^{2}}xx^{\top}+V(\Psi;F_{\beta})^{-1}\Big]^{-1}\Big|^{1/2}\frac{1}{(2\pi)^{q/2}} \\
& \Big|\frac{1}{n}V(\Psi;F_{\beta})\Big|^{-1/2}\exp\Bigg(\Big[-\frac{1}{2\sigma^{2}}(z-x^{\top}t_{E})^{2}-n(t_{E}-T(F_{\beta}))^{\top}V(\Psi;F_{\beta})^{-1}(t_{E}-T(F_{\beta}))\Big]\Bigg)dt_{E}.
\end{aligned}
$$

To find an analytic approximation of the expectation of the predictive distribution we

apply again the Laplace method, that leads to the following approximation

$$\begin{aligned}\mathrm{E}_{\hat{\beta}}\left[f^{\star}_{\beta}(z,x)\right] \approx \frac{1}{\sigma^2\sqrt{2\pi}}\left|V(\Psi;F_\beta)\right|^{-1}\left|\left[\frac{1}{n\sigma^2}xx^\top + V(\Psi;F_\beta)^{-1}\right]^{-1}\right|^{1/2} \\ \left|\left[\frac{1}{n\sigma^2}xx^\top + 2V(\Psi;F_\beta)^{-1}\right]^{-1}\right|^{1/2} \\ \exp\left(\underbrace{-n\left[\frac{1}{2n\sigma^2}(z-x^\top\hat{t}_E)^2 + (\hat{t}_E - T(F_\beta))^\top V(\Psi;F_\beta)^{-1}(\hat{t}_E - T(F_\beta))\right]}_{\tilde{m}(\hat{t}_E)}\right),\end{aligned} \tag{3.7}$$

where $\hat{t}_E = \operatorname{argmin}_{t_E}\tilde{m}(t_E)$ and $\tilde{m}$ is the m function derived from the second Laplace approximation. Finally, the bias is:

$$\operatorname{bias}(f^{\star}_{\beta}(z,x)) := \mathrm{E}_{\hat{\beta}}\left[f^{\star}_{\beta}(z,x)\right] - f_\beta(z,x)\,, \tag{3.8}$$

where $f_\beta(z,x)$ is the distribution of the responses. Considering that the Laplace approximation is of order $\mathrm{O}(n^{-1})$ for fixed values of q (or under the condition that q is smaller than n with a certain rate (see Shun and McCullagh [1995])) the application of a second Laplace approximation would give a composition of approximations of order $\mathrm{O}(n^{-1})$. Considering the fact that the constant terms can be ignored, we end up having approximately the same rate of convergence as the first Laplace approximation, considering fixed q and increasing n. Furthermore, the approximations obtained by Laplace have been compared with the Monte Carlo simulation to evaluate their accuracy. As found in Chapter 2, the difference between the two different methods is at level of `1e-5`, comparing the results either with the Kullback-Leibler divergence or obtaining the maximum difference in absolute value. In order to compute the total MSE of the predictive distribution, we need to integrate the MSE over z, such that we can get the MISE (Mean Integrated Squared Error), that is

$$\mathrm{MISE}(f^\star) = \int_{-\infty}^{\infty}\left[\mathrm{Var}_{\hat{\beta}}(f^{\star}_{\beta}(z,x)) + \operatorname{bias}(f^{\star}_{\beta}(z,x))^2\right]dz\,. \tag{3.9}$$

In the simulation study, we focus on this theoretical result about the MISE of the predictive distribution. But, in order to derive the variance of the predictive distribution (3.5), we need to obtain the variance of the bias-calibrated estimator in (3.2). The details of the derivation of the variance of the bias-calibrated estimator can be found in Appendix C.2.

3.2.2 Methods to Evaluate the Predictive Distribution

There are different measures that can be used to analyze the influence of the bias-calibrated estimator on the predictive distribution. As a first example, if we want to concentrate on the predictive distribution forecasting as close as possible the distribution of the bulk of the data (as discussed in Chapter 2) we would use the Kullback-Leibler divergence.

Instead, it could be possible to use the functional boxplots of Sun and Genton [2011] that give a summary of all the predictive distributions that are obtained. In this way, it is obtained the band of the central 50% of the curves (like the IQR in the boxplot) and the curves of the outlying densities. Besides, these functional boxplots could give a

better idea about the sharpness (that can be seen as the width of the bands) of the overall predictive distribution of each level of calibration applied to the robust M-estimator.

Moreover, in this chapter we derived the MISE of the predictive distribution. In this way, we can analyze the trade-off between the bias of the predictive distribution and its variability. The aim of the simulation study is to consider the analysis of the theoretical MISE. Thus, we can use the MISE as measure of selection, so that we choose the value of c_2 that minimizes the MISE.

3.3 Simulation Study

In this section, we analyze the selection of the value of c_2 that minimizes the MISE. We obtain the MISE considering the trade-off between variability and bias, but keeping in mind the fact that in applied data sets we do not know the value of ϵ, the distribution H and the true parameter β.

3.3.1 Selection of the Best Value of the Tuning Constant via MISE

We consider the set of observations $Y = (y_1, \ldots, y_{100})$ generated as follows:

$$y_i = x_i^\top \beta + e_i, \quad i = 1, \ldots, 100 \ ,$$

where

$$\beta = [1.1, -2.2, 1.5, -1.6, 2]^\top \ ,$$

and x is the vector of the explanatory variables. The elements of the matrix of the explanatory variables are defined as follows: x_{i1} is 1, $x_{i2} \sim N(2, 0.7)$, $x_{i3} \sim U[1, 4]$, $x_{i4} \sim U[0, 3]$, $x_{i5} \sim N(1, 0.5)$. Moreover, the distribution of e_i is:

$$e_i \sim (1 - \epsilon)N(0, 3^2) + \epsilon H \ ,$$

where H is the contaminating distribution defined for three scenarios as $N(9, 3^2)$, $N(12, 3^2)$ and $N(15, 3^2)$ and $\epsilon = 5\%$ is the level of the contamination present in the data. The estimation is defined by using a M-estimator $\hat{\beta}_R$ computed via function `rlm` in `R` using the Huber-function with tuning constant $c_1 = 1.345$. The value of $\hat{\sigma}_R$ is estimated via MAD and obtained from `rlm` when estimating $\hat{\beta}_R$. $\hat{\beta}_{cal}$ is obtained via equation (3.2). The estimation of the MISE is obtained via equations: MISE formulation (3.9), with variance (3.5) and bias (3.8). The x's values of the future observations are set to: $x = [1, 2, 2, 1, 0.95]$, for a value in the center of the data, $x = [1, 1.3, 1, 1, -0.1]$ and $x = [1, -0.5, 1.5, 2, 1.5]$, for two observations in the tail, in order to observe possible differences between the three points.

The algorithm is the following:

- At each replication (called k) we generate X_k and Y_k, and estimate $\hat{\beta}_R^k$ and $\hat{\sigma}_R^k$. In order to minimize the MISE, we compute $\hat{\beta}_{cal}^{c_2^k} = \hat{\beta}_{cal}^{c_2^k}(\hat{\beta}_R^k, \hat{\sigma}_R^k)$ for each chosen value of c_2. At each iteration, we select the value $\hat{c}_2^k$ that minimizes the MISE for a specific value of x.

- The grid of values of c_2 is: $c_2 = 0$ and $c_2 = 1.6, \ldots, 7$ with steps of 0.3. The starting value of $c_2 = 1.6$ is chosen to avoid a too strong shrinkage of the variance of the bias-calibrated estimator and to allow a small difference from the value of $c_1 = 1.345$.
- Repeat the procedure for $k = 1, \ldots, 100$, the three contamination schemes and the three different values of x.

Figure 3.1 shows that the selected value of c_2 is influenced by the type of contamination present in the data. In fact, we notice that the more the distortion in the data, the greater the selection of a higher value of c_2. In particular, when the contamination does not have an important influence, the calibration does not improve the results compared to the robust estimator, and the value of c_2 tends to be the smallest calibrating value or the robust estimator itself (i.e $c_2 = 0$ or $c_2 = 1.6$). With the increasing size of the introduced outliers, the chances to select a higher value of c_2 increase as well.

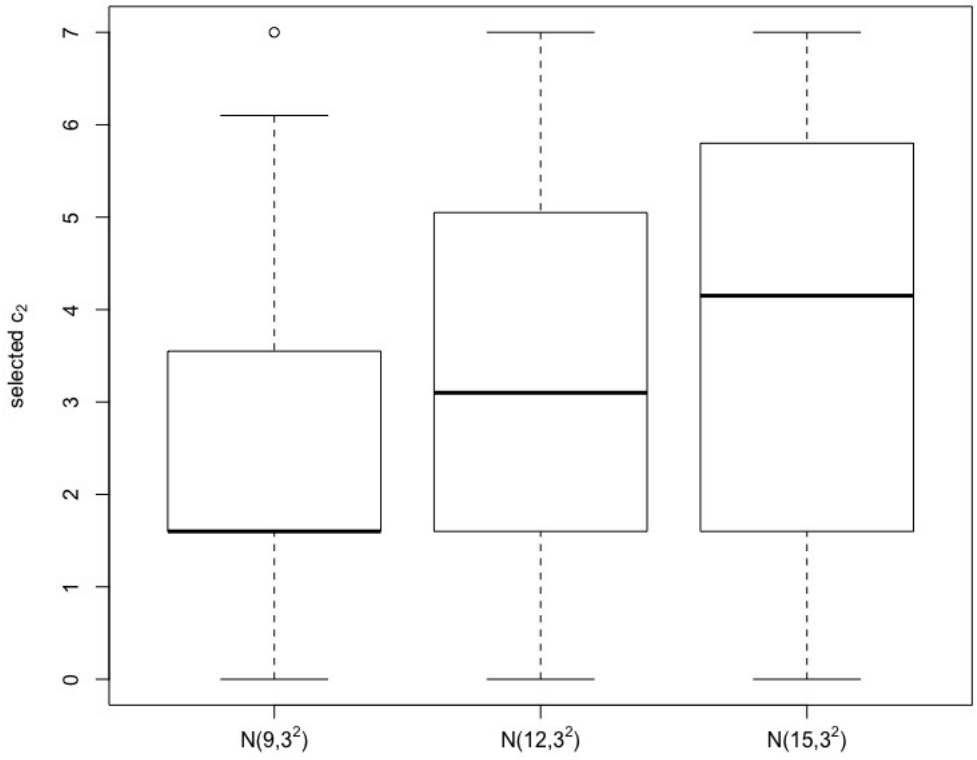

Figure 3.1: Results of the selected values of c_2 when minimizing the MISE of the predictive distribution for three different contamination. The size of the sample is $n = 100$ and the future observation is $x = [1, 2, 2, 1, 0.95]$.

Based on the graphical results in Figure 3.1 we choose to display four different estimators that are respectively the robust estimator $\hat{\beta}_R$ and three examples of bias-calibrated estimators, with $c_2 = 2.2, 4, \infty$. The choice of the first two values is to show intermediary points in the scale of c_2, the first one close to the best choice of c_2 when the contamination is low (first and second boxplot on the left), and $c_2 = 4$ is the median value of c_2 when the contamination is higher. Additionally, a further result corresponds to the results when selecting the best value of c_2 (i.e. the value that minimizes the MISE in each replication). Table 3.1 shows an example of numerical results of the median of 100 replications of the MISE of the predictive distribution for the three different scenarios of contamination for $x = [1, 2, 2.5, 1.5, 0.95]$. We notice that the choice of the value of c_2 that minimizes the MISE gives a relevant difference with respect to the other values of c_2. Moreover, in Table 3.2-3.4 we have further results to understand the behavior of the MISE for three different x's (chosen to be in the middle of the distribution and in the tails). We can notice that decomposing the MISE into the variance and the bias squared, we attain that the variance has a greater impact on the MISE than the bias squared, but also that the value of the bias is increasing when the contamination is increasing (Table 3.2). In addition, we have similar results when minimizing the MISE selecting the best value of c_2 predicting the distribution for an observation in the tails (Table 3.3 and Table 3.4). The main difference, as it would have been expected, is that in these situations the variance has even a stronger impact than an observation in the center of the data and the values are around 10 times bigger than in Table 3.2.

contamination	Robust	$c_2 = 2.2$	$c_2 = 4$	$c_2 = \infty$	min c_2
$N(9, 3^2)$	0.136	0.129	0.130	0.132	0.106
$N(12, 3^2)$	0.150	0.142	0.146	0.146	0.097
$N(15, 3^2)$	0.229	0.196	0.170	0.180	0.122

Table 3.1: Median of the MISE of the predictive distribution over 100 replications for three different scenarios for $x = [1, 2, 2, 1, 0.95]$. The values are multiplied by 100.

	contamination	Robust	$c_2 = 2.2$	$c_2 = 4$	$c_2 = \infty$	min c_2
Var	$N(9, 3^2)$	0.097	0.086	0.089	0.089	0.079
bias^2	$N(9, 3^2)$	0.049	0.048	0.043	0.045	0.031
Var	$N(12, 3^2)$	0.096	0.085	0.087	0.089	0.081
bias^2	$N(12, 3^2)$	0.058	0.054	0.053	0.052	0.015
Var	$N(15, 3^2)$	0.121	0.118	0.113	0.117	0.112
bias^2	$N(15, 3^2)$	0.104	0.079	0.053	0.063	0.005

Table 3.2: Median of the variance and bias squared of the predictive distribution over 100 replications for three different contamination for $x = [1, 2, 2, 1, 0.95]$. The values are multiplied by 100.

	contamination	Robust	$c_2 = 2.2$	$c_2 = 4$	$c_2 = \infty$	min c_2
MISE	$N(12, 3^2)$	1.867	1.859	1.777	1.866	1.457
Var	$N(12, 3^2)$	1.357	1.172	1.162	1.161	1.050
bias^2	$N(12, 3^2)$	0.477	0.434	0.561	0.526	0.219

Table 3.3: Median of the MISE, the variance and the bias squared of the predictive distribution over 100 replications for $x = [1, 1.3, 1, 1, -0.1]$. The values are multiplied by 100.

	contamination	Robust	$c_2 = 2.2$	$c_2 = 4$	$c_2 = \infty$	min c_2
MISE	$N(12, 3^2)$	3.972	3.711	4.198	4.406	3.082
Var	$N(12, 3^2)$	3.233	3.140	3.430	3.460	2.612
bias^2	$N(12, 3^2)$	0.586	0.572	0.552	0.543	0.568

Table 3.4: Median of the MISE, the variance and the bias squared of the predictive distribution over 100 replications for $x = [1, -0.5, 1.5, 2, 1.5]$. The values are multiplied by 100.

Additionally, we want to show the behavior of the predictive distribution by using the functional boxplots, where the predictive distribution is obtained via the `TMB` package. The functional boxplots of the 100 replications of the simulation study are presented in Figure 3.2 and in Figure 3.3 we represent the difference between the predictive distribution and the true distribution. The (green) vertical line represents the mean of the future observation with covariates corresponding to x (computation of $(1-\epsilon)x^\top\beta + \epsilon\tilde{\mu}$). In the graphs representing the functional boxplots, the black curve is the median and the dashed (yellow) lines represent the distributions that are considered as outlier candidates detected by 1.5 times the 50%. Moreover, the thick (blue) lines are the limits of the intervals of the distributions (50% of central region, and minimum and maximum of the range of non-outlying curves) and the filled (green) area is the indication of the 50% spread of the central curves.

In Figure 3.2 and Figure 3.3 we show an example when considering an observation in the tail. We can remark the high variability of the results and consequently the shape of the functional boxplot is not smooth. Furthermore, the green areas are consistently wide for this case and there are bumps in the middle of the limiting (blue) lines. Due to the high variability of the predictive distributions, the shift of the curves towards the direction of the contamination is less evident as it would be in the case of an observation centered in 0. Furthermore, for these type of observations (i.e. in the tail of the distribution), we select more often lower values of c_2 (in order to avoid a too strong effect of the outliers). Generally, we can conclude that for observations in the tail the results are too variable and not easy to distinguish the differences, as it could have been expected. Finally, in Figure 3.3 it is slightly clearer how the variability of the predictive distributions is lower when selecting the best value of c_2 that minimizes the MISE.

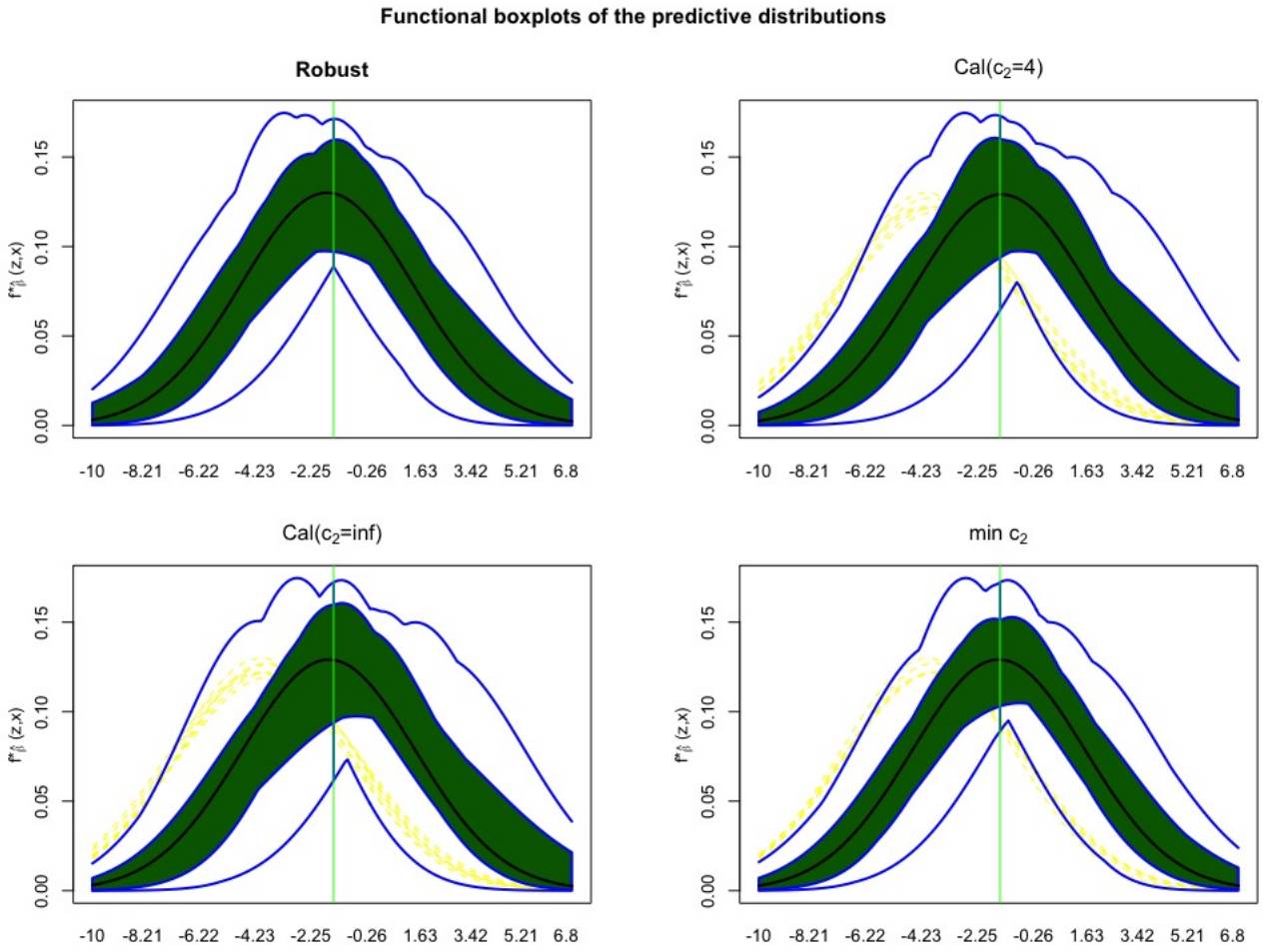

Figure 3.2: Functional boxplots of the 100 replications to compute the predictive distribution $f^{\star}_{\hat{\beta}}(z, x)$ considering a robust estimator, three different calibrated estimators $c_2 = 4, \infty$ and the minimum of c_2 selected at each replication. The value of the future observation is $x = [1, 1.3, 1, 1, -0.1]$. The contaminating distribution is $N(12, 3^2)$.

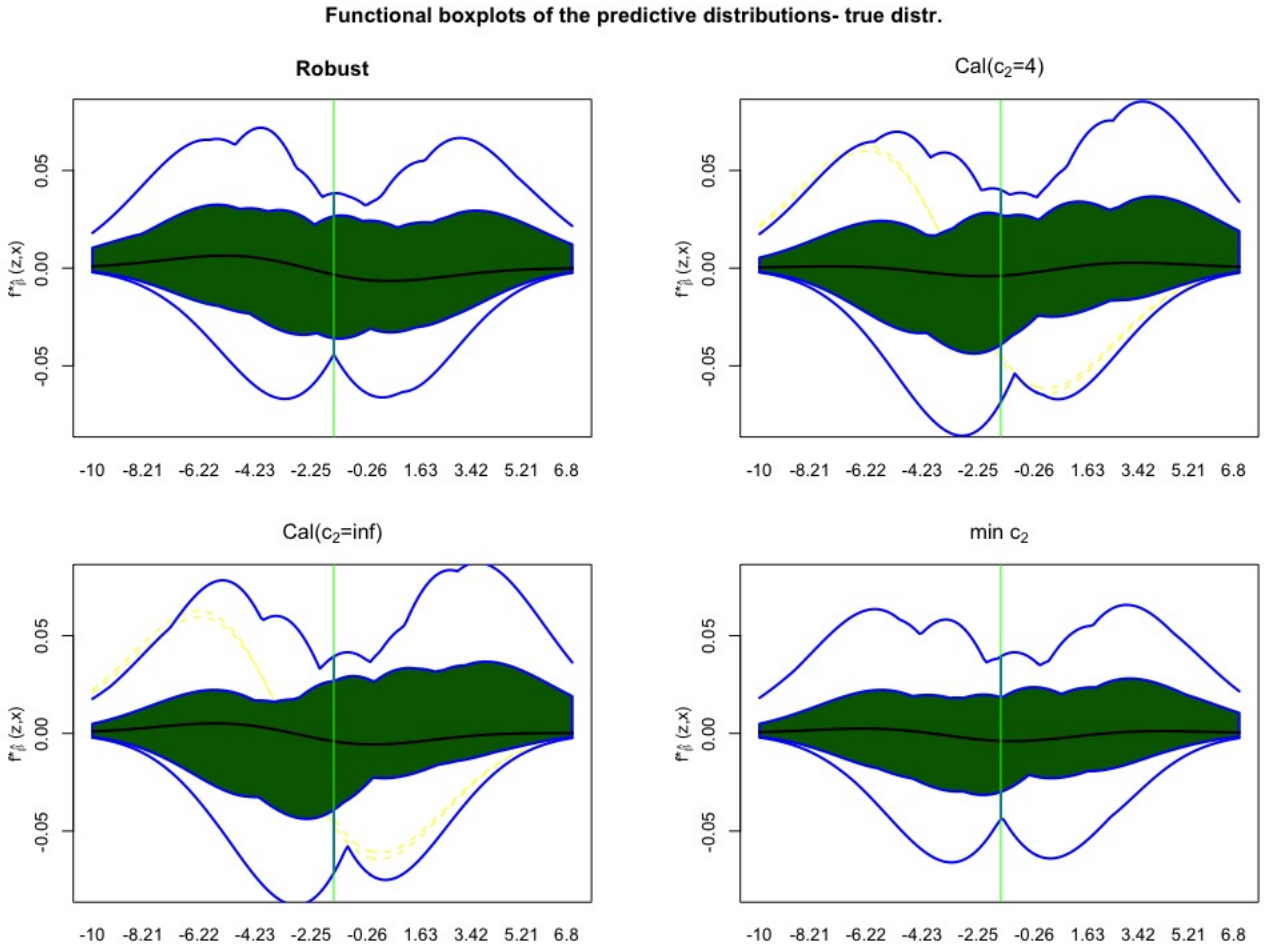

Figure 3.3: Functional boxplots of the 100 replications of the predictive distribution $f^{\star}_{\hat{\beta}}(z,x)$ minus the true distribution when considering a robust estimator, three different calibrated estimators $c_2 = 4, \infty$ and the minimum of c_2 selected at each replication. The value of the future observation is $x = [1, 1.3, 1, 1, -0.1]$. The contaminating distribution is $N(12, 3^2)$.

3.4 Data Example: Prostate Cancer

To conclude this chapter, we apply our approach to a real data set. The aim of this example is to present how the predictive distribution can be used in practical situations. The chosen data are available on the `R` package `Brq`, the reference is a study done by Stamey et al. [1989], and this dataset has been also used in the book Hastie et al. [2009]. The data set consists of 97 observations about the correlation between the level of prostate specific antigen (PSA) and a number of clinical measures. The goal of the study was to predict the value of PSA based on different measurements, such as: prostate weight, age, benign prostatic hyperplasia amount, semina vesicle invasion, capsular penetration, Gleason score and percent of Gleason scores 4 or 5. Table 3.5 and Table 3.6 include the description of the variable and their summary statistics.

Name	Description
cavol	Cancer volume
weight	Prostate weight
age	Age
bph	Amount of benign prostatic hyperplasia
svi	Seminal vesicle invasion
cp	Capsular penetration
gleason	Gleason score
pgg45	Percentage Gleason scores 4 or 5
psa	Value of prostate-specific antigen (response variable)

Table 3.5: Variable's description.

Variable	N	Mean	St. Dev.	Min	Max
cavol	97	7.001	7.885	0.260	45.650
weight	97	45.478	45.610	10.750	449.250
age	97	63.866	7.445	41	79
bph	97	2.645	2.937	0.250	10.240
svi	97	0.216	0.414	0	1
cp	97	2.362	3.720	0.250	18.250
gleason	97	6.753	0.722	6	9
pgg45	97	24.381	28.204	0	100
psa	97	23.740	40.825	0.650	265.850

Table 3.6: Variable's statistic.

We visualize the response through a boxplot in Figure 3.4. We remark that the response variable contains observations that can be considered as contamination (i.e. outliers). The amount of the observations that can be detected as outliers is about 8-10%, there are 9 observations (9.3%) that have a value of PSA greater than 50. This fact, and considering the results in the simulation study, could mean that the median value of the selected calibrating constant is around 4 and 5. Figure 3.4 shows the boxplot of the response variable and the density of the standardized residuals obtained from the

regression model estimated via robust M-estimator with tuning constant $c = 1.345$ and with $c = \infty$ as an example, where we can see the difference between the robust and classical residuals. Comparing the graphical results in Figure 3.4 and the estimations in Table 3.7, we can see the differences between the robust and the ML estimation and more in particular how the ML is influenced by the presence of outliers in the data.

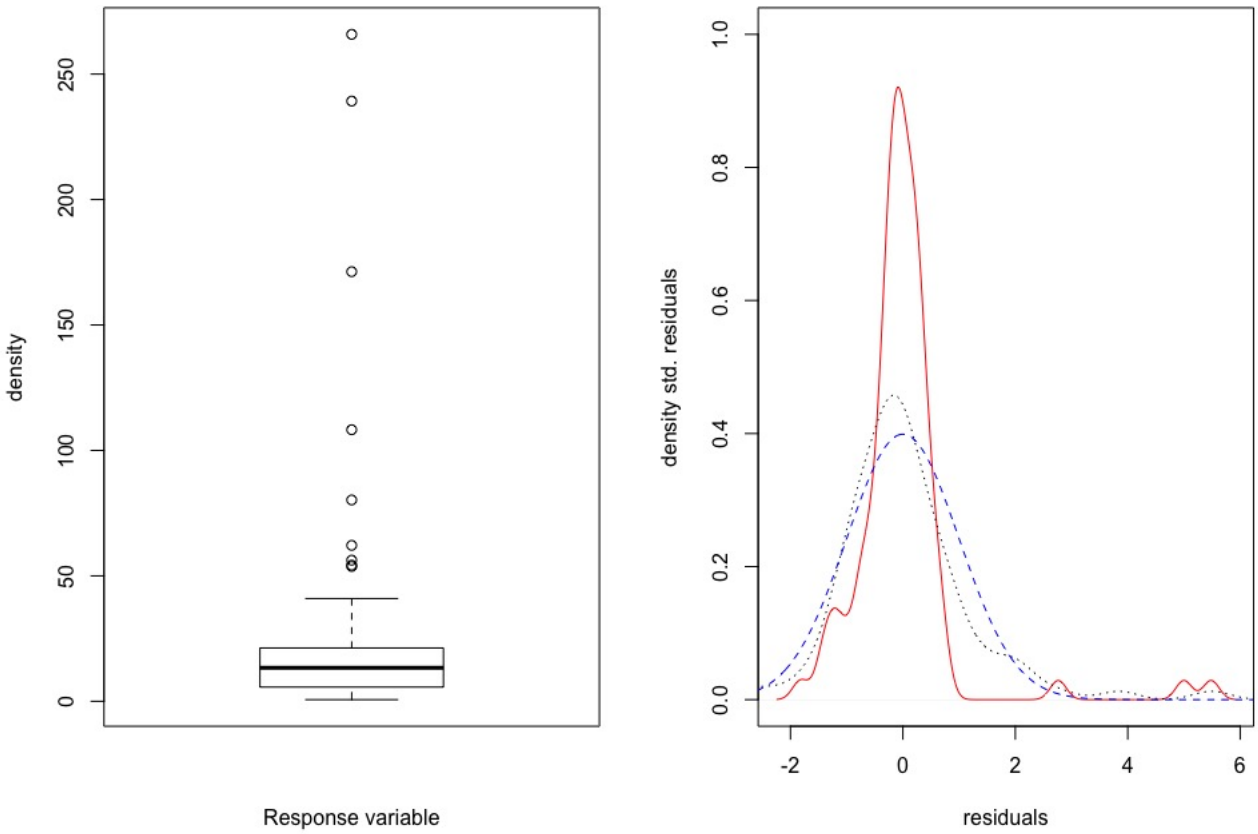

Figure 3.4: The graph on the left shows the boxplot of the response variable. The graph on the right shows the density of the standardized robust residuals (short-dashed line), the density following a $N(0, 1)$ (dashed-line) and the standardized residuals (continuous line) for the MLE.

	robust coefficient (SD)	ML coefficient (SD
cavol	1.970***	2.235***
	(0.186)	(0.584)
weight	0.016	−0.007
	(0.024)	(0.075)
age	−0.240	−0.280
	(0.156)	(0.489)
bph	0.835**	1.427
	(0.400)	(1.252)
svi	15.772***	21.774*
	(3.560)	(11.147)
cp	−1.463***	1.741
	(0.450)	(1.410)
gleason	−1.018	−4.933
	(2.195)	(6.873)
pgg45	0.079	−0.015
	(0.060)	(0.188)
Constant	22.237	47.340
	(15.940)	(49.906)
Observations	97	97
Residual Std. Error	9.250 (df = 88)	31.479 (df = 88)
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 3.7: Summary of the robust regression using `rlm` with Huber-function and tuning constant $c = 1.345$ and the estimation of Residual Std. Error via MAD.

In order to predict the distribution of PSA for future patients, to perform our analysis, we would like to use the bias-calibrated approach and find the best value of c_2 by minimizing the value of the MISE of the predictive distribution as performed in the simulation study on the available data. For this purpose, we regress the model via robust M-estimator applying the Huber-function with tuning constant $c = 1.345$ via function `rlm` and we obtain the value of $\hat{\sigma}_R$ as our value of σ (using the MAD method).

Table 3.7 shows the result of the full estimated robust model and the coefficients give us the values of $\hat{\beta}_R$, while the value of the residual standard error is our $\hat{\sigma}_R$. This table gives us the idea about the influence that each variable has on the response. In our case, we consider $\hat{\beta}_R$ with the explanatory variables that are significant in the regression model.

The variables are selected via robust variable selection, such that we take into account the relevant information from the data to compute the MISE of the predictive distribution.

As an example and in order to apply our selection of the value of c_2 that minimizes the MISE, we use an algorithm considering the leave-one-out Cross-Validation (LOOCV, or Jackknife). The idea is to repeat the selection of the best value of c_2 for N times, dividing the sample in the training data set and the testing data set. The training data set corresponds to all the observations except one and the observation left out represents the testing observation. This means, that at each iteration we compute the value of $\hat{\beta}_R^{[-i]}$ and $\hat{\sigma}_R^{[-i]}$ that correspond to the estimation of β and σ without the i-th observation. Consequently, we compute the MISE of the predictive distribution considering the observation left out of the sample as the future observation, and we perform the predictive based on the information given by the training data. The possible values of c_2 go from 1.6 until 8 with steps of 0.3 and the value of 0 for the robust M-estimator.

Alternatively, we can apply our selection of the value of c_2 that minimizes the MISE, performing the analysis only once on the available dataset and we choose an arbitrary observation as future observation to be predicted (choosing new values such as the mean or median in the descriptive statistics). Consequently, we compute the MISE of the predictive distribution considering the arbitrary future observation, and we perform the predictive based on the information given by the full data set.

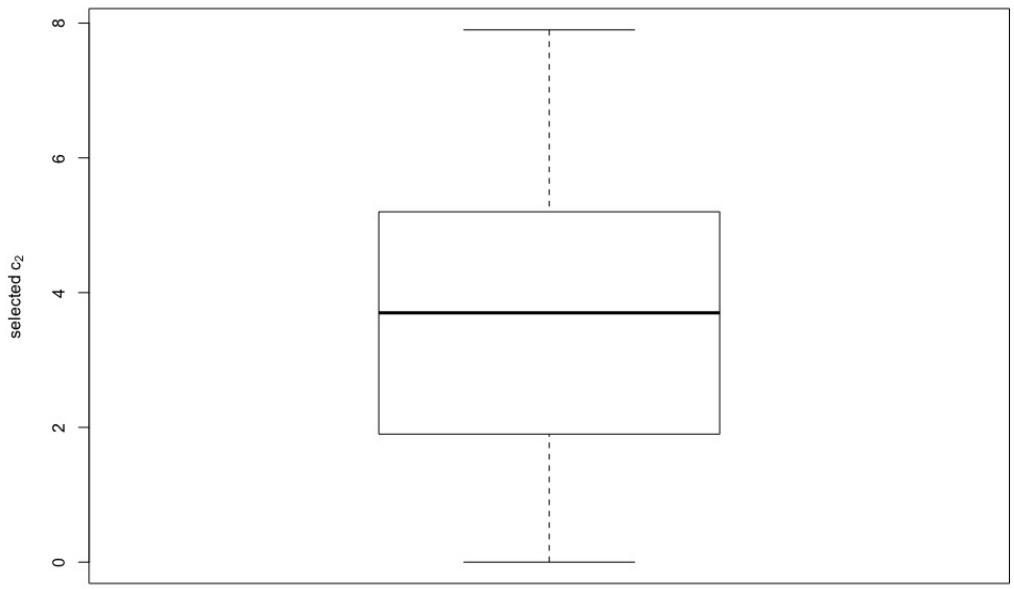

Figure 3.5: Boxplot of the selected value of c_2 using the LOOCV approach.

Figure 3.5 and Table 3.8 show the results of the analysis of the MISE, obtained in section 3.2.1. From Table 3.8 we see that the best value of c_2 that minimizes the MISE is around 4-4.5. In particular, we can see the gain in terms of MISE compared to the robust M-estimator without any calibration. This result confirms what we have seen in

	Robust	$c_2 = 1.9$	$c_2 = 2.8$	$c_2 = 3.7$	$c_2 = 4.6$	$c_2 = 8$	min c_2
Mean	6.780	5.479	4.564	3.723	3.538	4.164	1.963
Median	6.830	4.954	2.669	1.147	1.169	2.716	0.321

Table 3.8: Mean and median of the MISE of the predictive distribution using the LOOCV approach.

the simulation study that in the presence of a contamination the best value of the selected c_2 depends on the amount of contaminated data and tends to be higher when the presence of the contamination is more important. In this data set, the predictive distribution of the value of PSA gives a lower MISE most of the times for values in the range $3.5 - 5$. The robust estimator happens to be selected mostly for the values that are in the tails, as we have remarked in the simulation study.

Furthermore, in Figure 3.6 we present an example of predictive distribution comparing the robust and a non-robust estimators used to obtain the predictive distribution. From Figure 3.6 we can see how in this case the estimation of the predictive distribution based on a non-robust estimator are different and for this observation taken from the sample is more influenced by the presence of a contamination in the data. The vertical line represent the mean of the response variable that is also shown in Table 3.6.

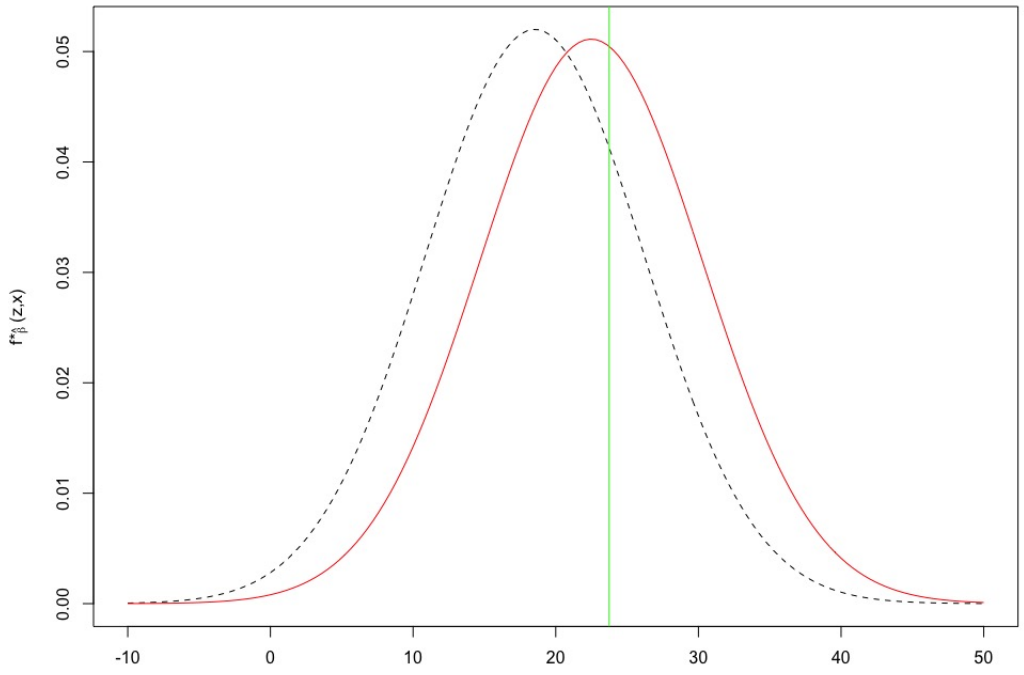

Figure 3.6: Eample of the predictive distributions for an observation from the sample considering the robust estimator a calibrated estimators and a non-robust estimator $c_2 = \infty$.

In this case, the predictive distribution could be used to predict and estimate the distribution of values of PSA for patients based on their characteristics or different groups

(e.g. different age or weight classes). Therefore they can be compared to the estimated value or the known distribution of the data from medical practices. Additionally it can be very useful when the value of PSA might be missing as outcome in a clinical trial due to lost to follow-up or other reasons and this value needs to be imputed via statistical methods. In this specific case, the use of the predictive distribution is appropriate.

Generally speaking, the use of a calibrated predictive distribution can help in performing more precise prediction of the value of PSA compared to the robust estimator itself that would give too conservative results and ignore the effect of outliers (e.g. miss some patients that have a level of PSA that might be on the borderline and be missclassified). The use of the MLE estimator would give more "high PSA level" predictions and increase medical costs because more patients would need more specific analysis to detect the cancer.

Conclusion

The work in this thesis has been mainly developed around the concepts of predictive distribution, bias-calibration and robustness. It is known that when using robust estimators in predictive distribution, the variability of the latter is lower compared to a predictive distribution based on non-robust estimators. Nevertheless, in the case of distorted data where there is the presence of representative outliers, the robust estimator has the disadvantage to be biased with respect to these representative outliers because they are down-weighted. In fact, most of the times the outliers are considered as representative for the sample or population and not to be down-weighted too heavily. Therefore, the overall objective was to correct the robust estimator bias and to improve the trade-off between variability and bias of the predictive distribution in linear regression. This peculiarity allows to adapt the prediction, reducing the robustness of the classic robust estimators, while keeping a good efficiency in terms of variability of the bias-calibrated estimator and consequently of the predictive distribution itself.

The field of predictive distribution and bias-calibration can be further developed with new methods that are partially mentioned or not yet covered in the literature. The concept of bias-calibration, based on the work in this thesis, can be extended into the GLM setting that would give a wide range of distributions. A first step in this direction would be the derivation of this bias-calibrated estimator for GLM following the same concept of the bias-calibration used in linear models.

The challenges of the development of this bias-calibrated estimator for the GLM applied to the predictive distribution lay mainly in the derivation of the variance of the bias-calibrated estimator. In fact, as it was done in this thesis for the linear regression case, the derivation of the variance can be cumbersome. The presence of the correction term in the GLM estimators increases the difficulty in obtaining a precise result of the variance of the bias-calibrated estimator. Not only the derivation would require a long and demanding process, but, in particular, the implementation of the variance in the software would drive the results to probable rounding and numerical problems due to the complexity of the analytic result of the variance. This is why more work for future research should be done in this direction to find an appropriate approximation and solution that could solve this problem.

Appendix A

Additional Material About Predictive Distribution and Robust GLM

A.1 Derivation of the Predictive Distribution

We give here the details of the computation of the predictive distribution presented in Chapter 2.

Recalling from Chapter 2

$$p_1(\epsilon) = \left|V(\Psi; F_{\epsilon,\theta})\right|^{-\frac{1}{2}} \quad \text{and} \quad p_2(\epsilon) = -\frac{n}{2}(t - T(F_{\epsilon,\theta}))^\top V(\Psi; F_{\epsilon,\theta})^{-1}(t - T(F_{\epsilon,\theta}))\ ,$$

we have the derivative of (2.8) w.r.t ϵ evaluated at $\epsilon = 0$ equivalent to

$$\left(\frac{2\pi}{n}\right)^{-\frac{q}{2}} \left[\frac{\partial}{\partial\epsilon}p_1(\epsilon)\Big|_{\epsilon=0} \exp\{p_2(0)\} + p_1(0)\ \frac{\partial}{\partial\epsilon}p_2(\epsilon)\Big|_{\epsilon=0} \exp\{p_2(0)\}\right]\ , \tag{A.1}$$

where

$$\frac{\partial}{\partial\epsilon}p_1(\epsilon)\Big|_{\epsilon=0} = -\frac{1}{2}\left|V(\Psi; F_\theta)\right|^{-\frac{3}{2}} \left[\operatorname{tr}\left(\operatorname{adj}(V(\Psi; F_\theta))\frac{\partial}{\partial\epsilon}V(\Psi; F_{\epsilon,\theta})\Big|_{\epsilon=0}\right)\right]\ , \tag{A.2}$$

and

$$\frac{\partial}{\partial e}\det A(e) = \operatorname{tr}\left(\operatorname{adj}(A(e))\frac{\partial A(e)}{\partial e}\right)\ ,$$

where adj is the adjugate matrix, that is the transpose of the cofactor matrix. In the following computation, to ease the notation, we define $D = (t - T(F_\epsilon))$. Furthermore,

$$\begin{aligned}
\frac{\partial}{\partial\epsilon}p_2(\epsilon)\Big|_{\epsilon=0} &= -\frac{n}{2}\Bigg\{\frac{\partial}{\partial\epsilon}\Big[D^\top\Big]_{\epsilon=0} V(\Psi; F_\theta)^{-1}D + D^\top\frac{\partial}{\partial\epsilon}V(\Psi; F_{\epsilon,\theta})^{-1}\Big|_{\epsilon=0} D + \\
&\quad + D^\top V(\Psi; F_\theta)^{-1}\frac{\partial}{\partial\epsilon}\Big[D\Big]_{\epsilon=0}\Bigg\} \qquad (A.3)\\
&= -\frac{n}{2}\Bigg\{-2\frac{\partial}{\partial\epsilon}T(F_{\epsilon,\theta})^\top\Big|_{\epsilon=0} V(\Psi; F_\theta)^{-1}D + D^\top\frac{\partial}{\partial\epsilon}V(\Psi; F_{\epsilon,\theta})^{-1}\Big|_{\epsilon=0} D\Bigg\}\ .
\end{aligned}$$

The details to complete the derivation of expression (2.9) are explained below. The derivative of. $T(F_{\epsilon,\theta})$ is

$$\frac{\partial}{\partial\epsilon}T(F_{\epsilon,\theta})\Big|_{\epsilon=0} = M(\Psi,F_\theta)^{-1}\int\Psi(z,x;T(F_\theta))dH(z,x)\ , \tag{A.4}$$

and to obtain the derivative of the variance of the estimator, defined in (1.6), we use the differentiation w.r.t. ϵ, at $\epsilon=0$, of

$$V(\Psi,F_{\epsilon,\theta})V(\Psi,F_{\epsilon,\theta})^{-1} = I_{q\times q}\ ,$$

we have

$$\left(\frac{\partial}{\partial\epsilon}V(\Psi,F_{\epsilon,\theta})\Big|_{\epsilon=0}\right)V(\Psi,F_\theta)^{-1} + V(\Psi,F_\theta)\frac{\partial}{\partial\epsilon}V(\Psi,F_{\epsilon,\theta})^{-1}\Big|_{\epsilon=0} = 0_{q\times q}\ ,$$

i.e.

$$\frac{\partial}{\partial\epsilon}V(\Psi,F_{\epsilon,\theta})^{-1}\Big|_{\epsilon=0} = -V(\Psi,F_\theta)^{-1}\left(\frac{\partial}{\partial\epsilon}V(\Psi,F_{\epsilon,\theta})\Big|_{\epsilon=0}\right)V(\Psi,F_\theta)^{-1}\ . \tag{A.5}$$

Therefore,

$$\begin{aligned}
&\frac{\partial}{\partial\epsilon}V(\Psi,F_{\epsilon,\theta})\Big|_{\epsilon=0}\\
&= \frac{\partial}{\partial\epsilon}\left[\underbrace{-\int\frac{\partial}{\partial t}\Psi(z,x;t)\Big|_{t=T(F_{\epsilon,\theta})}dF_{\epsilon,\theta}(z,x)}_{M(\Psi,F_\theta)}\right]_{\epsilon=0}^{-1}Q(\Psi,F_\theta)M(\Psi,F_\theta)^{-\top}\\
&\quad+M(\Psi,F_\theta)^{-1}\left[\frac{\partial}{\partial\epsilon}\underbrace{\int\Psi(z,x;T(F_{\epsilon,\theta}))\Psi(z,x;T(F_{\epsilon,\theta}))^\top dF_{\epsilon,\theta}(z,x)}_{Q(\Psi,F_\theta)}\right]_{\epsilon=0}M(\Psi,F_\theta)^{-\top}\\
&\quad+M(\Psi,F_\theta)^{-1}Q(\Psi,F_\theta)\frac{\partial}{\partial\epsilon}\left[\underbrace{-\int\frac{\partial}{\partial t}\Psi(z,x;t)\Big|_{t=T(F_{\epsilon,\theta})}dF_{\epsilon,\theta}(z,x)}_{M(\Psi,F_\theta)}\right]_{\epsilon=0}^{-\top}.
\end{aligned}\tag{A.6}$$

Subsequently we need to compute the derivative of $M(\Psi,F_{\epsilon,\theta})^{-1}$. Its expression is obtained following the same approach as the case for the variance in (A.5), for which we need the derivative of the matrix $M(\Psi,F_{\epsilon,\theta})$ itself. To compute the derivative of this matrix we define $\int\frac{\partial}{\partial t}\Psi(z,x;t)|_{t=T(F_{\epsilon,\theta})} = I_{\partial\Psi(\epsilon)}$, $\int\frac{\partial}{\partial t}\Psi(z,x;t)|_{t=T(F_\theta)} = I_{\partial\Psi}$, $\int\frac{\partial}{\partial t}\frac{\partial}{\partial t^\top}\Psi(z,x;t)|_{t=T(F_\theta)} = I_{2\partial\Psi}$ and $T_\partial = \frac{\partial}{\partial\epsilon}T(F_{\epsilon,\theta})|_{\epsilon=0}$

$$\begin{aligned}
\frac{\partial}{\partial\epsilon}M(\Psi,F_{\epsilon,\theta})\Big|_{\epsilon=0} &= \frac{\partial}{\partial\epsilon}\left[-(1-\epsilon)I_{\partial\Psi(\epsilon)}dF_\theta(z,x)\right]_{\epsilon=0} - \frac{\partial}{\partial\epsilon}\left[\epsilon I_{\partial\Psi(\epsilon)}dH(z,x)\right]_{\epsilon=0}\\
&= I_{\partial\Psi}dF_\theta(z,x) - I_{\partial\Psi}dH(z,x) - I_{2\partial\Psi}dF_\theta(z,x)T_\partial\\
&= -M(\Psi,F_\theta) - I_{\partial\Psi}dF_\theta(z,x)T_\partial - I_{\partial\Psi}dH(z,x).
\end{aligned}\tag{A.7}$$

On the other hand, the derivative of the matrix $Q(\Psi,F_{\epsilon,\theta})$ w.r.t ϵ evaluated at $\epsilon=0$ is

$$\begin{aligned}
\frac{\partial}{\partial\epsilon}Q(\Psi,F_{\epsilon,\theta})\Big|_{\epsilon=0} &= \frac{\partial}{\partial\epsilon}(1-\epsilon)I_{2\Psi(\epsilon)}dF_\theta(z,x)\Big|_{\epsilon=0} + \frac{\partial}{\partial\epsilon}\epsilon I_{2\Psi(\epsilon)}dH(z,x)\Big|_{\epsilon=0} \qquad (A.8)\\
&= -Q(\Psi,F_\theta) + I_{\partial\Psi\Psi}dF_\theta(z,x)T_\partial + T_\partial^\top I_{\partial\Psi\Psi}dF_\theta(z,x) + I_{2\Psi}dH(z,x)\ .
\end{aligned}$$

where $I_{2\Psi(\epsilon)} = \int \Psi(z,x;T(F_{\epsilon,\theta}))\Psi(z,x;T(F_{\epsilon,\theta}))^\top$, $I_{2\Psi} = \int \Psi(z,x;T(F_\theta))\Psi(z,x;T(F_\theta))^\top$ and $I_{\partial\Psi\Psi} = \int \frac{\partial}{\partial t}\Psi(z,x;t)|_{t=T(F_\theta)}\Psi(z,x;T(F_\theta))^\top$.

Using (A.7) and (A.8) in (A.6) gives:

$$\begin{aligned}\frac{\partial}{\partial\epsilon}V(\Psi;F_{\epsilon,\vartheta})\Big|_{\epsilon=0} &= -M(\Psi,F_\theta)^{-1}M_\partial V(\Psi,F_\theta) - V(\Psi,F_\theta)M_\partial M(\Psi,F_\theta)^{-\top}+\\ &+ M(\Psi,F_\theta)^{-1}I_{\partial\Psi\Psi}dF_\theta(z,x)T_\partial^\top M(\Psi,F_\theta)^{-\top}+\\ &+ M(\Psi,F_\theta)^{-1}T_\partial\left[I_{\partial\Psi\Psi}^\top dF_\theta(z,x)\right]^\top M(\Psi,F_\theta)^{-\top}+\\ &+ M(\Psi,F_\theta)^{-1}I_{2\Psi}dH(z,x)M(\Psi,F_\theta)^{-\top} - V(\Psi,F_\theta)\ , \end{aligned} \tag{A.9}$$

with $M_\partial = \frac{\partial M(\Psi,F_{\epsilon,\theta})}{\partial\epsilon}|_{\epsilon=0}$.

Appendix B

The Laplace Approximation

B.1 Laplace Approximation for Multiple Integrals

In this Appendix, we explain in details the steps necessary to compute the robust predictive distribution. An approach to tackle this problem is to use the Laplace approximation defined in (2.12).

The first integral we want to approximate is

$$\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \quad f_t(z,x)(2\pi)^{-\frac{q}{2}}|n^{-1}V(\Psi;F_\theta)|^{-\frac{1}{2}} \\ \exp\left\{-\frac{n}{2}(t-T(F_\theta))^\top V(\Psi;F_\theta)^{-1}(t-T(F_\theta))\right\} dt \,, \tag{B.1}$$

and for a Poisson GLM we have

$$f_t(z,x) = \frac{\exp(-\exp(x^\top t))(\exp(x^\top t))^z}{z!} \,,$$

and we can rewrite (B.1) as

$$(2\pi)^{-\frac{q}{2}}|n^{-1}V(\Psi;F_\theta)|^{-\frac{1}{2}} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \exp\Big\{zx^\top t + \\ -\exp(x^\top t) - \log(z!) - \frac{n}{2}(t-T(F_\theta))^\top V(\Psi;F_\theta)^{-1}(t-T(F_\theta))\Big\} dt \,.$$

Rearranging some terms we can define $\xi = n$, such that we get

$$-\xi m(t) = -n\left[\frac{-zx^\top t + \exp(x^\top t) + \log(z!)}{n} + \frac{1}{2}(t-T(F_\theta))^\top V(\Psi;F_\theta)^{-1}(t-T(F_\theta))\right] \,.$$

To conclude, we need to get the first and second derivative of $m(t)$, that is

$$\frac{\partial m(t)}{\partial t} = \frac{1}{n}\Big(-zx^\top + x^\top \exp(x^\top t)\Big) + (t-T(F_\theta))^\top V(\Psi;F_\theta)^{-1} \,, \tag{B.2}$$

$$\frac{\partial^2 m(t)}{\partial t^2} = \frac{1}{n}\Big(xx^\top \exp(x^\top t)\Big) + V(\Psi;F_\theta)^{-1} \,.$$

The final step is to find $\hat{t}$ such that

$$\frac{\partial m(\hat{t})}{\partial t} = 0 \,. \tag{B.3}$$

To find the solution of $\hat{t}$ we use the Newton-Raphson method.

We tackle now the second term of the predictive distribution in (2.7) that is

$$\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} f_t(z,x)\frac{\partial}{\partial\epsilon}g(t;F_{\epsilon,\theta}(z,x))\Big|_{\epsilon=0} dt\ ,$$

where

$$\begin{aligned}
\frac{\partial}{\partial\epsilon}g(t;F_{\epsilon,\theta}(z,x))\Big|_{\epsilon=0} &= -\frac{n^{q/2}}{2(2\pi)^{q/2}}\Big|V(\Psi;F_\theta)\Big|^{-\frac{3}{2}} \\
&\left[\mathrm{tr}\left(\mathrm{adj}(V(\Psi;F_\theta))\frac{\partial}{\partial\epsilon}V(\Psi;F_{\epsilon,\theta})\Big|_{\epsilon=0}\right)\right] \\
&\exp\left\{-\frac{n}{2}(t-T(F_\theta))^\top V(\Psi;F_\theta)^{-1}(t-T(F_\theta))\right\} - \frac{n^{q/2+1}}{2(2\pi)^{q/2}}\Big|V(\Psi;F_\theta)\Big|^{-\frac{1}{2}} \\
&\Bigg\{-2\frac{\partial}{\partial\epsilon}T(F_{\epsilon,\theta})^\top\Big|_{\epsilon=0}V(\Psi;F_\theta)^{-1}(t-T(F_\theta)) - (t-T(F_\theta))^\top \\
&\left[V(\Psi;F_\theta)^{-1}\frac{\partial}{\partial\epsilon}V(\Psi;F_{\epsilon,\theta})\Big|_{\epsilon=0}V(\Psi;F_\theta)^{-1}\right](t-T(F_\theta))\Bigg\} \\
&\exp\left\{-\frac{n}{2}(t-T(F_\theta))^\top V(\Psi;F_\theta)^{-1}(t-T(F_\theta))\right\}\ . \qquad \text{(B.4)}
\end{aligned}$$

From (B.4) we can divide the integral in two different parts, such that

$$\begin{aligned}
-\frac{n^{q/2}}{2(2\pi)^{q/2}}\Big|V(\Psi;F_\theta)\Big|^{-\frac{3}{2}} &\qquad \left[\mathrm{tr}\left(\mathrm{adj}(V(\Psi;F_\theta))\frac{\partial}{\partial\epsilon}V(\Psi;F_{\epsilon,\theta})\Big|_{\epsilon=0}\right)\right] \qquad \text{(B.5)} \\
\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} &\qquad f_t(z,x)\exp\left\{-\frac{n}{2}(t-T(F_\theta))^\top V(\Psi;F_\theta)^{-1}(t-T(F_\theta))\right\}dt\ ,
\end{aligned}$$

is the first term and the second is

$$\begin{aligned}
\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} &\qquad f_t(z,x) - \frac{n^{q/2+1}}{2(2\pi)^{q/2}}\Big|V(\Psi;F_\theta)\Big|^{-\frac{1}{2}} \\
&\Bigg\{-2\frac{\partial}{\partial\epsilon}T(F_{\epsilon,\theta})^\top\Big|_{\epsilon=0}V(\Psi;F_\theta)^{-1}(t-T(F_\theta)) - (t-T(F_\theta))^\top \\
&\left[V(\Psi;F_\theta)^{-1}\frac{\partial}{\partial\epsilon}V(\Psi;F_{\epsilon,\theta})\Big|_{\epsilon=0}V(\Psi;F_\theta)^{-1}\right](t-T(F_\theta))\Bigg\} \\
&\exp\left\{-\frac{n}{2}(t-T(F_\theta))^\top V(\Psi;F_\theta)^{-1}(t-T(F_\theta))\right\}dt\ . \qquad \text{(B.6)}
\end{aligned}$$

We see that these two integrals have a similar form as in (B.1). For equation (B.5) only the term not depending on t is different (the constant term outside the integral). For (B.6) we have an additional term depending on t changing the function $m(t)$, that in this case, is

$$\begin{aligned}
m(t) &= f(z,x|t) - 2\frac{\partial}{\partial\epsilon}T^\top(F_{\epsilon,\theta})\Big|_{\epsilon=0}V(\Psi;F_\theta)^{-1}(t-T(F_\theta)) + \\
&\quad -(t-T(F_\theta))^\top\left[V(\Psi;F_\theta)^{-1}\frac{\partial}{\partial\epsilon}V(\Psi;F_{\epsilon,\theta})\Big|_{\epsilon=0}V(\Psi;F_\theta)^{-1}\right](t-T(F_\theta))\ .
\end{aligned}$$

The last term to compute is the second part of (B.4). In this case it is necessary to take the logarithm of the term not present in exponential form. Consequently, a condition to make the Laplace approximation works is that the element in the logarithm should not be negative. In this case, we have

$$-m(t) \;=\; -\Big[-zx^\top t + \exp(x^\top t) + \log(z!) - \log(P) \Big] ,$$

where

$$\begin{aligned} P \;=\; & -2\frac{\partial}{\partial\epsilon} T(F_{\epsilon,\theta})^\top \Big|_{\epsilon=0} V(\Psi; F_\theta)^{-1}(t - T(F_\theta)) - (t - T(F_\theta))^\top \\ & \Big[V(\Psi;F_\theta)^{-1} \frac{\partial}{\partial\epsilon} V(\Psi;F_{\epsilon,\theta}) \Big|_{\epsilon=0} V(\Psi;F_\theta)^{-1} \Big] (t - T(F_\theta)) \, . \end{aligned} \tag{B.7}$$

The first two derivatives of (B.7) are

$$\frac{\partial h(t)}{\partial t} = -zx^\top + x^\top \exp(x^\top t) - \frac{\partial P}{\partial t} P^{-1} \, , \tag{B.8}$$

$$\frac{\partial^2 h(t)}{\partial t^2} = xx^\top \exp(x^\top t) - \left[\frac{\partial^2 P}{\partial t^2} P - \frac{\partial P}{\partial t}^\top \frac{\partial P}{\partial t} \right] P^{-2} \, ,$$

where

$$\begin{aligned} \frac{\partial P}{\partial t} \;=\; & -2\frac{\partial}{\partial\epsilon} T(F_{\epsilon,\theta})^\top \Big|_{\epsilon=0} V(\Psi; F_\theta)^{-1} + \\ & -2(t - T(F_\theta))^\top \Big[V(\Psi;F_\theta)^{-1} \frac{\partial}{\partial\epsilon} V(\Psi;F_{\epsilon,\theta}) \Big|_{\epsilon=0} V(\Psi;F_\theta)^{-1} \Big] \, , \end{aligned}$$

and

$$\frac{\partial^2 P}{\partial t^2} = \; -2\Big[V(\Psi;F_\theta)^{-1} \tfrac{\partial}{\partial\epsilon} V(\Psi;F_{\epsilon,\theta}) \Big|_{\epsilon=0} V(\Psi;F_\theta)^{-1} \Big] \, .$$

In order to obtain the value of t such that equation (B.8) is equal to 0, we use a Newton-Raphson method. Knowing that the value in the logarithm might be negative, when computing the minimization of (B.8) it might be necessary to introduce a constraint that $P > 0$. Also in this case, by simulation, we get that $\hat{t}$ tends to $T(F_\theta)$. Applying this, the result of (B.7) tends to $-\infty$ because $P \to 0$.

B.2 Numerical Evaluations of the Approximation

Table B.1 reports the comparison between the three different approaches discussed in Section 2.3 to compute the multiple integrals (Laplace approximation via TMB, Monte Carlo approximation and analytic Laplace approximation). We notice that the results are rather similar and the order of magnitude of the difference is around `1e-4`. Furthermore, the number of the covariates does not have a relevant impact on the accuracy of the predictive distribution. We conclude that `TMB` is the best approach that most efficiently satisfies the trade-off between accuracy of the results and computational effort via simulation. The setting used to perform this comparison refers to Section 2.4 for $q = 5$. For $q = 10$ and $q = 20$ the same distributions have been considered with different parameters, that for simplicity we do not specify here. The MC method is considered as the benchmark.

# variables	TMB vs MC	TMB vs Laplace	Laplace vs MC
$q = 5$	5.2	4.1	13.3
$q = 10$	1.1	0.4	0.8
$q = 20$	3.6	1.7	0.8

Table B.1: Kullback-Leibler divergence between three approaches used to approximate the multiple integration: `TMB` package (TMB), Monte Carlo approximation (MC) and analytic Laplace approximation (Laplace). The value of q represents the number of covariates. The values represented are multiplied by `1e4`.

Appendix C

Additional Material for Chapter 3

C.1 Brief Overview of the Bias-Calibration in the Literature

The concept of bias-calibration is mainly used in sample survey theory. The main idea behind it comes from Chambers [1986] where the author introduces this type of estimators and he focuses on the concept of representative and non-representative outliers. A representative outlier can be defined as an outlying observation relevant in the sample and that cannot be considered as incorrect. Consequently, these observations are presumed to be important to describe the finite population. In fact, it is based on the work of bias-calibration of Chambers [1986] that Welsh and Ronchetti [1998] concentrate their work on sample survey containing outliers. The main focus in this introduction is to briefly review the existing bias-calibrated estimators in the literature.

The finite population cumulative distribution function is defined as

$$F(t) = N^{-1} \sum_{i=1}^{N} I(Y_i \leq t) \quad t \in \mathbb{R} \, ,$$

where $Y_1, \ldots, Y_N$ are the population values and $I(\cdot)$ is the indicator function. The finite population total

$$T = N \int_{-\infty}^{\infty} t \, dF(t) \, ,$$

the finite population mean T/N, and the finite population quantile

$$Q(\alpha) = F^{-1}(\alpha) \, , \qquad 0 \leq \alpha \leq 1 \, ,$$

where α is the quantile to estimate, are statistical properties of the finite population distribution.

The approach in Welsh and Ronchetti [1998] is model-based and it first needs to define the super-population model for the conditional distribution of Y given X, where X is the matrix of the explanatory variables. The model is defined as:

$$Y_i = X_i \beta + \sigma v^{1/2}(X_i) e_i, \qquad i = 1, \ldots, N \, , \tag{C.1}$$

where e_i are iid random variables, β and σ unknown parameters and $v(\cdot)$ a non-negative function that accounts for heteroskedasticity. The approach to estimate the population

distribution function consists in first obtaining the fit of the model in (C.1) by using an M-estimator $\hat{\beta}_R$ of β with a robust function and a chosen tuning constant c_1.

An important class of estimators of T under model (C.1) can be written as

$$\hat{T} = T_1 + \hat{\beta} \sum_{i=n+1}^{N} X_i \ , \tag{C.2}$$

where $T_1 = \sum_{i=1}^{n} Y_i$ is the sample total and $\hat{\beta}$ is the estimator of β. The bias-calibrated estimator from Chambers [1986] is defined as

$$\hat{T}_{cal}(c_2) = T_1 + \hat{\beta}_R \sum_{i=n+1}^{N} X_i + \hat{B}(c_2) \ ,$$

where

$$\hat{B}(c_2) = \Big(\sum_{i=n+1}^{N} X_i \Big) \Big\{ \sum_{j=1}^{n} X_j^2 / v(X_j) \Big\}^{-1} \sum_{j=1}^{n} \hat{\sigma} c_2 \{ X_j / v(X_j)^{1/2} \} \psi_{c_2} \{ (Y_j - \hat{\beta}_R X_j) / \hat{\sigma} c_2 v(X_j)^{1/2} \} \ .$$

with ψ_{c_2} the Huber function and $\hat{\beta}_R$ is the bi-weight estimator introduced above. Finally, the estimator in Chambers [1986] (the bias-calibrated estimator) can be written in the form (C.2) as

$$\hat{\beta}_{cal}(c_2) = \hat{\beta}_R + \Big\{ \sum_{j=1}^{n} X_j^2 / v(X_j) \Big\}^{-1} \sum_{j=1}^{n} \hat{\sigma} c_2 \{ X_j / v(X_j)^{1/2} \} \psi_{c_2} \{ (Y_j - \hat{\beta}_R X_j) / \hat{\sigma} c_2 v(X_j)^{1/2} \} \ .$$

Regarding the choice of the robust estimator of $\hat{\beta}_R$, $\hat{\sigma}_R$ and the ψ-function we refer to Welsh and Ronchetti [1998]. However, the choice of the ψ-function is far-reaching. In fact, it is necessary to consider that the more linear is the ψ-function the lower the bias of $\hat{T}_{cal}$ and the more bounded $\hat{\beta}_R$ the lower the variance of $\hat{T}_{cal}$.

Finally, it is required to estimate the population distribution function F by using the bias-calibrated estimator. The form of the estimated population function is

$$\hat{F}(t, c_2) = N^{-1} \Big\{ \sum_{i=1}^{n} I(Y_i \leq t) + (N - n) \hat{F}_2(t, c_2) \Big\} \quad t \in \mathbb{R} \ ,$$

where

$$\hat{F}_2(t, c_2) = n^{-1} (N - n)^{-1} \sum_{j=n+1}^{N} \sum_{i=1}^{n} I[\hat{\beta}_R X_j + v^{1/2}(X_j) \hat{\sigma} c_2 \psi_{c_2} \{ (Y_i - \hat{\beta}_R X_i) / \hat{\sigma} c_2 v^{1/2}(X_i) \} \leq t] \ ,$$

To conclude, the aim in Welsh and Ronchetti [1998] was to represent the quantile functions and the total population. In order to do this, they consider different levels of the tuning constant c_2 for the different quantile. As an example, they consider the following scheme of choice for c_2

$$c_2(\alpha) = \begin{cases} 0 & 0 \leq \alpha \leq 0.6 \\ 6 & 0.6 < \alpha \leq 0.85 \\ 10 & 0.85 < \alpha \leq 0.90 \\ 15 & 0.90 < \alpha \leq 0.95 \\ \infty & 0.95 < \alpha \leq 1 \end{cases} \ .$$

As mentioned in the latter paper, the values are chosen arbitrary and more research into the choice of them is required. Nevertheless, the choice of the different values of c will depend on the structure of the data and the amount of representative outliers that are included in the sample.

C.2 Variance of the Bias-calibrated Estimator

Considering σ as known, the variance of (3.2) is

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_{cal}) &= \mathrm{Var}\left(\hat{\beta}_R + \left(\sum_{i=1}^{n} x_i x_i^\top\right)^{-1} \sum_{i=1}^{n} \sigma\psi_{c_2}\{(y_i - x_i^\top\hat{\beta}_R)/\sigma\}x_i\right) \\ &= \mathrm{Var}\left(\hat{\beta}_R\right) + \mathrm{Var}\left(\left(\sum_{i=1}^{n} x_i x_i^\top\right)^{-1} \sum_{i=1}^{n} \sigma\psi_{c_2}\{(y_i - x_i^\top\hat{\beta}_R)/\sigma\}x_i\right) + \\ &\quad +2\,\mathrm{Cov}\left(\hat{\beta}_R, \left(\left(\sum_{i=1}^{n} x_i x_i^\top\right)^{-1} \sum_{i=1}^{n} \sigma\psi_{c_2}\{(y_i - x_i^\top\hat{\beta}_R)/\sigma\}x_i\right)^\top\right) .\end{aligned} \tag{C.3}$$

From (C.3), we need to compute the variance of the ψ-function and the covariance between the M-estimator and the ψ-function. To do this, we linearize $\hat{\beta}_R$ and ψ_{c_2} using the IF from Hampel et al. [1986]. To simplify the notation of the following computations, we define $\hat{r}_i = (y_i - x_i^\top\hat{\beta}_R)/\sigma$, $r_i = (y_i - x_i^\top\beta)/\sigma = e_i/\sigma$ and $\sum_{i=1}^{n} x_i x_i^\top = X^\top X$.

We define $\hat{\beta}_R$ as

$$\hat{\beta}_R \cong \beta + \frac{1}{n}\sum_{k=1}^{n} \mathrm{IF}(x_k, y_k; \hat{\beta}_R, F_\beta) ,$$

where

$$\mathrm{IF}(x_k, y_k; \hat{\beta}_R, F) = \frac{1}{E\left[\psi'_{c_1}\right]}(X^\top X)^{-1}\sigma\psi_{c_1}(r_k)x_k , \tag{C.4}$$

with c_1 representing the tuning constant used to obtain $\hat{\beta}_R$. The ψ-function depending on c_2 can be approximated by

$$\psi_{c_2}(\hat{r}_i) \cong \psi_{c_2}(r_i) + \frac{1}{n}\sum_{k=1}^{n} \mathrm{IF}(x_k, y_k; \psi_{c_2}(\hat{r}_i), F_\beta) .$$

where

$$\begin{aligned}\mathrm{IF}(y_k, x_k; \psi_{c_2}(\hat{r}_i), F) &= \frac{\partial\psi_{c_2}(r_i)}{\partial\beta}\,\mathrm{IF}(y_k, x_k; \hat{\beta}_R, F_\beta) \\ &= -\frac{1}{\sigma}\psi'_{c_2}(r_i)x_i^\top\,\mathrm{IF}(y_k, x_k; \hat{\beta}_R, F_\beta) = -\frac{1}{E\left[\psi'_{c_1}\right]}\psi'_{c_2}(r_i)\psi_{c_1}(r_k)x_i^\top(X^\top X)^{-1}x_k .\end{aligned} \tag{C.5}$$

The subsequent step is to tackle the elements of the variance of the bias-calibrated estimator in (C.3).

The first element of expression (C.3) is defined in (1.6). Hence, the second term of (C.3) is

$$\mathrm{Var}\left((X^\top X)^{-1}\sum_{i=1}^{n}\sigma\psi_{c_2}(\hat{r}_i)x_i\right) = \sigma^2(X^\top X)^{-1}\mathrm{Var}\left(\sum_{i=1}^{n}\psi_{c_2}(\hat{r}_i)x_i\right)(X^\top X)^{-1} ,$$

where the variance is

$$\begin{aligned}\mathrm{Var}\left(\sum_{i=1}^{n}\psi_{c_2}(\hat{r}_i)x_i\right) &\cong \mathrm{Var}\Bigg(\sum_{i=1}^{n}\underbrace{\left[\psi_{c_2}(r_i) - \frac{1}{n\,\mathrm{E}[\psi'_{c_1}]}\sum_{k=1}^{n}\psi'_{c_2}(r_i)\psi_{c_1}(r_k)x_i^\top(X^\top X)^{-1}x_k\right]}_{a_i}x_i\Bigg) \\ &= \sum_{i=1}^{n}\mathrm{Var}(a_i x_i) + \sum_{i\neq j}\mathrm{Cov}(a_i x_i, a_j x_j^\top) = \sum_{i=1}^{n}\mathrm{Var}(a_i)x_i x_i^\top + \sum_{i\neq j}\mathrm{Cov}(a_i, a_j)x_i x_j^\top .\end{aligned} \tag{C.6}$$

Thereby, we extend the two terms in expression (C.6). Firstly we have

$$\begin{aligned}\mathrm{Var}(a_i) &= \mathrm{Var}\Big(\psi_{c_2}(r_i) - \frac{1}{n\,\mathrm{E}[\psi'_{c_1}]}\sum_{k=1}^{n}\psi'_{c_2}(r_i)\psi_{c_1}(r_k)x_i^\top(X^\top X)^{-1}x_k\Big)\\ &= \mathrm{Var}\Big(\psi_{c_2}(r_i)\Big) + \frac{1}{n^2\,\mathrm{E}[\psi'_{c_1}]^2}\mathrm{Var}\Big(\sum_{k=1}^{n}\psi'_{c_2}(r_i)\psi_{c_1}(r_k)x_i^\top(X^\top X)^{-1}x_k\Big) +\\ &\quad - \frac{2}{n\,\mathrm{E}[\psi'_{c_1}]}\,\mathrm{Cov}\left(\psi_{c_2}(r_i), \Big(\sum_{k=1}^{n}\psi'_{c_2}(r_i)\psi_{c_1}(r_k)x_i^\top(X^\top X)^{-1}x_k\Big)^\top\right), \qquad \text{(C.7)}\end{aligned}$$

where

$$\mathrm{Var}\Big(\psi_{c_2}(r_i)\Big) = \mathrm{E}\left[\psi_{c_2}(r_i)^2\right] = 2\Phi(c_2) - 1 - 2c_2\phi(c_2) + 2c_2^2(1-\Phi(c_2)) \quad \forall i\,, \qquad \text{(C.8)}$$

due to the fact that the expectation of $\psi_{c_2}(r_i)$ is zero. The details about the solution of the expectation are shown in Appendix C. Afterwards, we have the variance of the second term in (C.7).

$$\begin{aligned}&\frac{1}{n^2\,\mathrm{E}[\psi'_{c_1}]^2}\mathrm{Var}\Big(\sum_{k=1}^{n}\psi'_{c_2}(r_i)\psi_{c_1}(r_k)x_i^\top(X^\top X)^{-1}x_k\Big)\\ &\qquad = \frac{1}{n^2 E[\psi'_{c_1}]^2}\sum_{k=1}^{n}x_i^\top(X^\top X)^{-1}x_k\mathrm{Var}\Big(\psi'_{c_2}(r_i)\psi_{c_1}(r_k)\Big)x_k^\top(X^\top X)^{-1}x_i\,. \qquad \text{(C.9)}\end{aligned}$$

We can verify that the expectation of expression (C.5) (and expression (C.4)) is zero, for the reason that the expectation of the Huber function is zero. Therefore, the variance in (C.9) is

$$\mathrm{Var}\Big(\psi'_{c_2}(r_i)\psi_{c_1}(r_k)\Big) = \mathrm{E}\left[\psi'_{c_2}(r_i)^2\psi_{c_1}(r_k)^2\right].$$

Thereupon, when $i = k$,

$$\mathrm{E}\left[\psi'_{c_2}(r_k)^2\psi_{c_1}(r_k)^2\right] = 2c_1^2\Big[\Phi(c_2) - \Phi(c_1)\Big] + 2\Phi(c_1) - 1 - 2c_1\phi(c_1)\,, \qquad \text{(C.10)}$$

and when $i \neq k$

$$\mathrm{E}\left[\psi'_{c_2}(r_i)^2\right]\mathrm{E}\left[\psi_{c_1}(r_k)^2\right] = \Big[2\Phi(c_2) - 1\Big]\Big[2c_1^2\Big[1-\Phi(c_1)\Big] + 2\Phi(c_1) - 1 - 2c_1\phi(c_1)\Big]\,, \qquad \text{(C.11)}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ represent the cumulative distribution function and the density function of a standard normal distribution, respectively, and $E[\psi'_{c_1}] = 2\Phi(c_1) - 1$. The third term in (C.7) is describable as

$$-\frac{2}{n\,\mathrm{E}[\psi'_{c_1}]}\sum_{k=1}^{n}\mathrm{Cov}\left(\psi_{c_2}(r_i), \psi'_{c_2}(r_i)\psi_{c_1}(r_k)\right)x_k^\top(X^\top X)^{-1}x_i\,,$$

where

$$\mathrm{Cov}\left(\psi_{c_2}(r_i), \psi'_{c_2}(r_i)\psi_{c_1}(r_k)\right) = \mathrm{E}\left[\psi_{c_2}(r_i)\psi'_{c_2}(r_i)\psi_{c_1}(r_k)\right],$$

for which the expectation is different from zero only when $i = k$, that is

$$\mathrm{E}\left[\psi_{c_2}(r_i)\psi'_{c_2}(r_i)\psi_{c_1}(r_i)\right] = -2c_1\Big[\phi(c_1) - \phi(c_2)\Big] + 2\Phi(c_1) - 1 - 2c\phi(c_1)\,. \qquad \text{(C.12)}$$

Afterwards, we expand the covariance of the second term in (C.6), that is

$$\operatorname{Cov}\Bigg(\underbrace{\psi_{c_2}(r_i)}_{b_i}\underbrace{-\frac{1}{n\,\mathrm{E}[\psi'_{c_1}]}\sum_{k=1}^{n}\psi'_{c_2}(r_i)\psi_{c_1}(r_k)x_i^\top(X^\top X)^{-1}x_k}_{d_{i(k)}},$$
$$\underbrace{\psi_{c_2}(r_j)}_{b_j}\underbrace{-\frac{1}{n\,\mathrm{E}[\psi'_{c_1}]}\sum_{l=1}^{n}\psi'_{c_2}(r_j)\psi_{c_1}(r_l)x_l^\top(X^\top X)^{-1}x_i}_{d_{j(l)}}\Bigg)\,. \tag{C.13}$$

Consequently, we need to obtain $b_i, b_j, d_{i(k)}$ and $d_{j(l)}$ in (C.13). Due to the independence of the residuals and $i \neq j$, the covariance of the combination $b_i b_j$ is equal to zero. The second term can be rewritten as the expectation of $b_i d_{j(l)}$ and as mentioned the expectation of these terms individually are zero. Thus, we have

$$\begin{aligned}&\mathrm{E}\Big[\psi_{c_2}(r_i)\Big(-\frac{1}{n\,\mathrm{E}[\psi'_{c_1}]}\sum_{l=1}^{n}\psi'_{c_2}(r_j)\psi_{c_1}(r_l)x_l^\top(X^\top X)^{-1}x_j\Big)\Big]\\ &\quad=-\frac{1}{n\,\mathrm{E}[\psi'_{c_1}]}\sum_{l=1}^{n}\mathrm{E}\Big[\psi_{c_2}(r_i)\psi'_{c_2}(r_j)\psi_{c_1}(r_l)\Big]x_l^\top(X^\top X)^{-1}x_j\,,\end{aligned}$$

where the latter expectation is different from zero only when $i = l$ and knowing that $i \neq j$ we obtain

$$\begin{aligned}&\mathrm{E}\Big[\psi_{c_2}(r_i)\psi'_{c_2}(r_j)\psi_{c_1}(r_i)\Big] = \\ &\quad=\Big[2c_2c_1\Big[1-\Phi(c_2)\Big]-2c_1\Big(\phi(c_1)-\phi(c_2)\Big)+2\Phi(c_1)-1-2c_1\phi(c_1)\Big]\Big[2\Phi(c_2)-1\Big]\,.\end{aligned} \tag{C.14}$$

Lastly, we consider $d_{i(k)}d_{j(l)}$. The latter element can be written as

$$\begin{aligned}&\mathrm{E}\Big[-\frac{1}{n\,\mathrm{E}[\psi'_{c_1}]}\sum_{k=1}^{n}\psi'_{c_2}(r_i)\psi_{c_1}(r_k)x_i^\top(X^\top X)^{-1}x_k\\ &\qquad\Big(-\frac{1}{n\,\mathrm{E}[\psi'_{c_1}]}\sum_{l=1}^{n}\psi'_{c_2}(r_j)\psi_{c_1}(r_l)x_l^\top(X^\top X)^{-1}x_j\Big)\Big]\\ &\quad=\frac{1}{n^2\,\mathrm{E}[\psi'_{c_1}]^2}\sum_{k=1}^{n}\mathrm{E}\Big[\psi'_{c_2}(r_i)\psi_{c_1}(r_k)^2\psi'_{c_2}(r_j)\Big]x_i^\top(X^\top X)^{-1}x_kx_k^\top(X^\top X)^{-1}x_j\,,\end{aligned}$$

and the expectation when $i = k \neq j$ and $j = k \neq i$ is

$$\begin{aligned}&\mathrm{E}\Big[\psi'_{c_2}(r_i)\psi_{c_1}(r_i)^2\Big]E\Big[\psi'_{c_2}(r_j)\Big]\\ &\quad=\Big[2c_1^2\Big[\Phi(c_2)-\Phi(c_1)\Big]+2\Phi(c_1)-1-2c_1\phi(c_1)\Big]\Big[2\Phi(c_2)-1\Big]\,.\end{aligned} \tag{C.15}$$

The last case occurs when we have $i \neq k \neq j$, that is

$$\mathrm{E}\Big[\psi'_{c_2}(r_i)\Big]E\Big[\psi_{c_1}(r_k)^2\Big]E\Big[\psi'_{c_2}(r_j)\Big] = \Big[2c_1^2\Big[1-\Phi(c_1)\Big]+2\Phi(c_1)-1-2c_1\phi(c_1)\Big]\Big[2\Phi(c_2)-1\Big]^2.$$

Finally, we need to compute the third element of expression (C.3), that is

$$
\begin{aligned}
&2\,\mathrm{Cov}\Big(\hat{\beta}_R, \Big((X^\top X)^{-1}\sum_{i=1}^{n}\sigma\psi_{c_2}(\hat{r}_i)x_i\Big)^\top\Big) \cong \qquad \text{(C.16)}\\
&\quad = 2\sigma\,\mathrm{Cov}\left(\beta + \frac{1}{n\,\mathrm{E}[\psi'_{c_1}]}\sum_{k=1}^{n}(X^\top X)^{-1}\sigma\psi_{c_1}(r_k)x_k,\right.\\
&\quad \left.\sum_{i=1}^{n}\Big(\Big(\psi_{c_2}(r_i) - \frac{1}{n\,\mathrm{E}[\psi'_{c_1}]}\sum_{l=1}^{n}\psi'_{c_2}(r_i)\psi_{c_1}(r_l)x_i^\top(X^\top X)^{-1}x_l\Big)x_i\Big)^\top\right)(X^\top X)^{-1}\,.
\end{aligned}
$$

Due to the fact that β is a constant, the covariances where β is involved are zero. Consequently, we expand the first non-zero covariance that is

$$
\begin{aligned}
&2\sigma\,\mathrm{Cov}\Big(\frac{1}{n\,\mathrm{E}[\psi'_{c_1}]}\sum_{k=1}^{n}(X^\top X)^{-1}\sigma\psi_{c_1}(r_k)x_k, \sum_{i=1}^{n}\psi_{c_2}(r_i)x_i^\top\Big)(X^\top X)^{-1}\\
&\qquad = \frac{2\sigma^2}{n\,\mathrm{E}[\psi'_{c_1}]}(X^\top X)^{-1}\Big(\sum_{k=1}^{n}x_k\,\mathrm{Cov}\Big(\psi_{c_1}(r_k),\psi_{c_2}(r_k)\Big)x_k^\top\Big)(X^\top X)^{-1}\,,
\end{aligned}
$$

where

$$
\mathrm{Cov}\Big(\psi_{c_1}(r_k),\psi_{c_2}(r_k)\Big) = 2c_2c_1\Big[1-\Phi(c_2)\Big] - 2c_1\Big[\phi(c_1)-\phi(c_2)\Big] + 2\Phi(c_1) - 1 - 2c_1\phi(c_1)\,.
$$

Conclusively, the last term is

$$
\begin{aligned}
&2\sigma\,\mathrm{Cov}\left(\frac{1}{n\,\mathrm{E}[\psi'_{c_1}]}\sum_{k=1}^{n}(X^\top X)^{-1}\sigma\psi_{c_1}(r_k)x_k,\right.\\
&\quad \Big(-\frac{1}{n\,\mathrm{E}[\psi'_{c_1}]}\sum_{i=1}^{n}\Big(\sum_{l=1}^{n}\psi_{c_2}(r_i)\psi_{c_1}(r_l)x_i(X^TX)^{-1}x_l\Big)x_i\Big)^\top\Big)(X^\top X)^{-1}\\
&\quad = \frac{2\sigma^2}{n^2\,\mathrm{E}[\psi'_{c_1}]^2}(X^\top X)^{-1}\left(\sum_{k}^{n}x_k\,\mathrm{Cov}\Big(\psi_{c_1}(r_k),\psi'_{c_2}(r_k)\psi_{c_1}(r_k)\Big)x_k^\top(X^\top X)^{-1}x_kx_k^\top+\right.\\
&\quad \left.+\sum_{k=l\neq i}^{n}x_k\,\mathrm{Cov}\Big(\psi_{c_1}(r_k),\psi'_{c_2}(r_i)\psi_{c_1}(r_k)\Big)x_k^\top(X^\top X)^{-1}x_ix_i^\top\right)(X^\top X)^{-1}\,,
\end{aligned}
$$

where

$$
\mathrm{Cov}\Big(\psi_{c_1}(r_k),\psi'_{c_2}(r_k)\psi_{c_1}(r_k)\Big) = 2c_1^2\Big[\Phi(c_2)-\Phi(c_1)\Big] + 2\Phi(c_1) - 1 - 2c_1\phi(c_1)\,,
$$

and

$$
\mathrm{Cov}\Big(\psi_{c_1}(r_k),\psi'_{c_2}(r_i)\psi_{c_1}(r_k)\Big) = \Big[2c_1^2\Big[1-\Phi(c_1)\Big] + 2\Phi(c_1) - 1 - 2c_1\phi(c_1)\Big]\Big[2\Phi(c_2)-1\Big]\,. \qquad \text{(C.17)}
$$

Summarizing, in the simulation study we will apply the MISE from expression (3.9) obtained computing the elements (3.5) and (3.7), where the variance for the bias-calibrated estimator has been computed in expressions (C.3)-(C.17).

We describe here some more details of the expectation of the ψ-function obtained above. The first result refers to the expectation of $\psi^2_{c_2}$ in (C.8), that is

$$\mathrm{E}\Big[\psi_{c_2}(r_i)^2\Big] = \int_{-\infty}^{-c_2}(-c_2)^2 dr + \int_{-c_2}^{c_2} r^2 dr + \int_{c_2}^{\infty} c_2^2 dr \tag{C.18}$$

Furthermore, we show the result in (C.10), that is

$$\mathrm{E}\Big[\psi'_{c_2}(r_k)^2\psi_{c_1}(r_k)^2\Big] = \int_{-\infty}^{-c_2} 0(-c_1)dr + \int_{-c_2}^{-c_1}(-c_1)dr + \int_{c_1}^{c_1} 1r^2 dr$$

and in (C.11) we need to obtain the first term that is

$$\mathrm{E}\Big[\psi'_{c_2}(r_k)^2\Big] = \int_{-\infty}^{-c_2} 0dr + \int_{-c_2}^{c_2} 1dr + \int_{c_2}^{\infty} 0dr\ , \tag{C.19}$$

while the second term is the same as (C.18) but for c_1. Furthermore, in (C.12) we have the following result

$$\begin{aligned}\mathrm{E}\Big[\psi_{c_2}(r_i)\psi'_{c_2}(r_i)\psi_{c_1}(r_i)\Big] &= \int_{-\infty}^{-c_2} 0(-c_1)(-c_2)dr + \int_{-c_2}^{-c_1} r(-c_1)dr + \int_{c_1}^{c_1} 1r^2 dr + \\ &+ \int_{c_1}^{c_2} 1\ r\ c_1 dr + \int_{c_2}^{\infty} 0\ c_1 c_2 dr\ .\end{aligned}$$

The two expectations in (C.14) are

$$\begin{aligned}\mathrm{E}\Big[\psi_{c_2}(r_i)\psi_{c_1}(r_i)\Big] &= \int_{-\infty}^{-c_2}(-c_1)(-c_2)dr + \int_{-c_2}^{-c_1} r(-c_1)dr + \int_{c_1}^{c_1} r^2 dr + \\ &+ \int_{c_1}^{c_2} r\ c_1 dr + \int_{c_2}^{\infty} c_1 c_2 dr\ .\end{aligned}$$

and the result of the second term is the same as (C.19). A last element that we need to define is the first element in (C.15), that is

$$\begin{aligned}\mathrm{E}\Big[\psi'_{c_2}(r_i)\psi_{c_1}(r_i)^2\Big] &= \int_{-\infty}^{-c_2} 0\ (-c_1)^2 dr + \int_{-c_2}^{-c_1} 1\ (-c_1)^2 dr + \int_{c_1}^{c_1} 1\ r^2 dr + \\ &+ \int_{c_1}^{c_2} 1\ c_1^2 dr + \int_{c_2}^{\infty} 0\ c_1^2 dr\ .\end{aligned}$$

References

J. Aitchison. Goodness of prediction fit. *Biometrika*, 62(3):547–554, 1975.

A. Basu and I. R. Harris. Robust predictive distributions for exponential families. *Biometrika*, 81(4):790–794, 1994.

E. Cantoni and E. M. Ronchetti. Robust inference for generalized linear models. *Journal of the American Statistical Association*, 96(455), 2001.

E. Cantoni and E. M. Ronchetti. A robust approach for skewed and heavy-tailed outcomes in the analysis of health care expenditures. *Journal of Health Economics*, 25(2):198–213, 2006.

R. L. Chambers. Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81(396):1063–1069, 1986.

A. J. Dobson and A. G. Barnett. *An introduction to generalized linear models*. CRC press, 2008.

J. J. Faraway. *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*, volume 124. CRC press, 2016.

D. Ferrari and D. La Vecchia. On robust estimation via pseudo-additive information. *Biometrika*, 99(1):238–244, 2012.

S. Geisser. The inferential use of predictive distributions. *Foundations of Statistical Inference*, pages 456–469, 1971.

T. Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011.

T. Gneiting and M. Katzfuss. Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151, 2014.

F. R. Hampel. *Contributions to the theory of robust estimation*. University of California, 1968.

F. R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974.

F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust statistics: the approach based on influence functions*, volume 114. John Wiley & Sons, 1986.

I. R. Harris. Predictive fit for natural exponential families. *Biometrika*, 76(4):675–684, 1989.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*, volume 2. Springer, 2nd edition, 2009.

S. Heritier, E. Cantoni, S. Copt, and M.-P. Victoria-Feser. *Robust methods in Biostatistics*, volume 825. John Wiley & Sons, 2009.

P. J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

P. J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–233, 1967.

P. J. Huber. *Robust statistics*. John Wiley & Sons, 1981.

P. J. Huber and E. M. Ronchetti. *Robust statistics*. John Wiley & Sons, 2nd edition, 2009.

K. Kristensen, A. Nielsen, C. W. Berg, H. Skaug, and B. Bell. Tmb: automatic differentiation and laplace approximation. *arXiv preprint*, 2015.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, pages 79–86, 1951.

S. N. Lô and E. M. Ronchetti. Robust and accurate inference for generalized linear models. *Journal of Multivariate Analysis*, 100(9):2126–2136, 2009.

R. A. Maronna, R. D. Martin, and V. J. Yohai. *Robust statistics. Theory and Methods*. John Wiley & Sons, 2006.

P. McCullagh and J. A. Nelder. *Generalized linear models*, volume 2. Chapman and Hall London, 1989.

J. A. Nelder and R. W. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.

Z. Shun and P. McCullagh. Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(4):749–760, 1995.

T. A. Stamey, J. N. Kabalin, J. E. McNeal, I. M. Johnstone, F. Freiha, E. A. Redwine, and N. Yang. Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. ii. radical prostatectomy treated patients. *The Journal of Urology*, 141 (5):1076–1083, 1989.

Y. Sun and M. G. Genton. Functional boxplots. *Journal of Computational and Graphical Statistics*, 20(2):316–334, 2011.

J. W. Tukey. A survey of sampling from contaminated distributions. *Contributions to Probability and Statistics*, 2:448–485, 1960.

A. H. Welsh and E. M. Ronchetti. Bias-calibrated estimation from sample surveys containing outliers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2):413–428, 1998.

M. Zhelonkin, M. G. Genton, and E. M. Ronchetti. On the robustness of two-stage estimators. *Statistics & Probability Letters*, 82(4):726–732, 2012.