ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics

Robustness of Evaluation Metrics for Predicting Probability Estimates of Binary Outcomes

Master Thesis Business Analytics and Quantitative Marketing

Author: Erine de Leeuw
Student ID number: 416499el

Supervisor: Dr. M. Zhelonkin
Second assessor: Dr. A. Alfons

Abstract

In many applications such as in financial and medical areas it is valuable to have probability estimates of binary outcomes. In this work we study the robustness of evaluation metrics for these estimates using strictly proper scoring rules. This is done by testing on both simulated data and real-world data using the following probabilistic binary classification methods: (robust) logistic regression (Cantoni and Ronchetti, 2001), support vector machine with Platt scaling and random forest with isotonic regression or Platt scaling. For the simulation the focus is on the effect of outliers and/or heavy-tailed error contamination on the (asymptotic) average and variance of the strictly proper scoring rules as well as the reliability curve of the associated model (where outlier contamination occurs in either the training sample or the training and test sample). From the simulation study we confirm that at the model the scores from classic and robust logistic regression are close. In the application we conclude that robust logistic regression is very reliable in many applications and gives competitive scores compared to machine learning methods, whilst enabling statistical inference.

Keywords: Prediction, Binary Outcome, Strictly Proper Scoring Rules, Calibration, Robustness, Generalized Linear Models, Asymptotic Relative Efficiency, Sensitivity Analysis, von Mises expansion.

October 24, 2019

Contents

1 Introduction

In many applications it is valuable to have predictions of the probability of an observation being in one of the categories of a binary outcome model. For instance, the probability that an insurance claim is fraudulent. First, one might be interested to predict whether a claim is fraudulent or not. Nonetheless, with limited resources it can be convenient to give more priority to cases with higher fraudulence probability, making the probability estimates more important. Another example is the prediction whether or not someone has a certain (incurable) disease, such as cancer or diabetes. Evidently, we want to predict whether a patient has the disease or not. However, at what point can such a sensitive decision be made on whether a patient has the disease; this uncertainty highlights the importance of probability estimates, as it gives a measure between 0 and 1.

A probabilistic prediction of a binary outcome takes the form of a predictive probability distribution over a future event of interest (Gneiting and Katzfuss, 2014), which can be done using certain (probabilistic) classification methods that give a posterior density. These methods use information from the training sample, which consists of observations with a known dependent variable. There are many different methods capable of producing probability estimates for binary outcomes. These methods can be roughly divided into four categories: generalized linear models, kernel methods, ensemble methods and deep learning methods (Zhou, 2015). The benchmark method for obtaining probabilistic estimates is logistic regression, which is a member of a class of generalized linear models (GLM). Traditionally, logistic regression is estimated in practice by a maximum likelihood estimator (MLE). It is well-known that it is sensitive to data contamination (Pregibon, 1981) and because of this it is often dismissed. Cantoni and Ronchetti (2001) proposed the robust estimators for generalized linear models, and for logistic regression in particular. We refer to it as robust logistic regression or robust GLM and use it in our study. This robust method is based on weighting the likelihood score function and thus falls into the Mallows class. Note that there is another robust alternative proposed by Bianco and Yohai (1996), which is not considered in this paper. The benchmark for kernel methods is support vector machine (SVM). However, as SVM solely results in a decision boundary, this method is combined with Platt scaling (Platt, 1999) to obtain the resulting probability estimates. This scaling is chosen for its well-calibrated posterior, which they claim has at least the same prediction accuracy for a correctly specified model. Random forest is viewed as the benchmark within ensemble methods (Liaw and Wiener, 2002). Zadrozny and Elkan (2001) showed that the random forest method needs calibration, as without calibration low quality predictions are obtained. We therefore choose to calibrate this method with either Platt scaling or isotonic regression. This further allows us to interpret differences between calibrated ensemble methods and non-calibrated ones. Finally, deep learning methods perform worse when using proper scoring rules then the other methods (Elliott and Lieli, 2013), hence this category of classification models will not be accounted for in this paper.

After the probability estimates are obtained, they need to be evaluated to find which method is reliable and superior. Many researchers make use of broad evaluation metrics, such as accuracy and precision, to assess the performance of each method. Nonetheless, these evaluation metrics can be lacking, as a good value of the metric does not equate to an honest and unbiased prediction. In probabilistic forecasting, as advocated by Gneiting and Raftery (2007), one should maximize sharpness subject to calibration. Calibration is related to consistency and sharpness refers to the concentration of the predictive distribution. Savage (1971) proved that for strictly proper scoring rules of binary events only the Brier score (Brier, 1950), the logarithmic score and the spherical score are mathematically optimal. Therefore, in this research we intend to evaluate the probability estimates of binary outcomes in a more tailored manner by using strictly proper scoring rules, as these do ensure honest and unbiased predictions through calibrated and sharp posterior probability estimates. Gneiting and Katzfuss (2014) argue that this is only correctly assessed when probabilistic predictions are calibrated, and additionally are evaluated by proper scoring rules.

In this work the robustness of the strictly proper scoring rules for evaluation of probability estimates of binary outcomes is studied. The focus is on the effect of contaminated observations in either the training sample or the training and test sample on the strictly proper scoring rule of the associated model for the probability estimate. Von Mises (von Mises, 1947) expansions of the functional form of the three most common strictly proper scoring rules for probabilistic binary outcomes explain the effect of contamination on the bias of these scoring rules under contamination. We conclude that when the model is correctly specified the robust counterpart of logistic regression obtains the same score in an uncontaminated data set whilst also being calibrated and its score is superior to classic logistic regression when there is contamination, which implies that robust logistic regression should be the new benchmark for probability estimates of binary outcomes.

The robustness of a binary classification, which is a point forecast in the binary outcome setup, was studied by Croux et al. (2008a,b) using GLM with a logit link. They also quantified the behaviour of the error rate of the logistic discrimination rule when outliers were added to the training sample. Nonetheless, this evaluation metric is usually for a binary point forecast and not a probabilistic one. Hence, in this study we present an analysis for strictly proper scoring rules, an evaluation metric of probabilistic predictions of binary outcomes. To the best of our knowledge, such an analysis has not been done yet.

The structure of this thesis is organized as follows: Section 2 gives an overview of the methods used to generate probability estimates and subsequently motivates the evaluation metrics used. In Section 3 an expression for the bias of the three population versions of three strictly proper scoring for probabilistic predictions of binary outcomes is given. Also, the asymptotic variance of this bias is obtained using the results derived from the von Mises expansion of the functional of the three scoring rules. In Section 4 simulation methodology is presented in order to compare the performance of all methods that give probabilistic predictions for binary outcomes, whereupon the results are presented in Section 6.

2 Methodology

As we intend to study the robustness of the evaluation for probability estimates of binary outcomes, firstly it is necessary to present the binary classification methods that give probabilistic outcomes. This is discussed in Section 2.1. Afterwards, in Section 2.2 the evaluation metrics are described.

2.1 Binary Classification Methods

Binary probabilistic classifiers have a goal to predict with what probability an observation will have one of two outcomes $y_i \in \{0, 1\}$. In all the experiments we compare the following models: (robust) logistic regression with a logit link for generalized linear models, calibrated support vector machine for kernel method and calibrated random forest for ensemble methods. This allows us to encompass the most popular binary prediction methods.

In general we are interested in explaining the $[n \times 1]$ dependent variable vector $\mathbf{y}$ through explanatory variables. The g explanatory variables are contained in the $[n \times g]$ matrix $\mathbf{X}$. Furthermore, to this $\mathbf{x}_i$ vector a scalar of one is added as the new first element of the vector of explanatory variables to allow for the intercept (β_0).

2.1.1 (Robust) Logistic Regression

Logistic Regression is a particular form of a generalized linear model (glm). A general reference is a book by McCullagh and Nelder (1989). It models a function of the expected value of the random response variable $\mathrm{E}\left(y_i\right) = \mu_i$ for $i = 1, \ldots, N_1$, with a linear combination of g predictors $\mathbf{x}_i \in \mathbb{R}^g$ and parameters $\boldsymbol{\beta} \in \mathbb{R}^g$, including an intercept, as follows:

$$\boldsymbol{\eta}_i = f\left(\mu_i\right) = \mathbf{x}_i^\top \boldsymbol{\beta}\,, \tag{1}$$

where $f\left(\cdot\right)$ is a link function. The parameter $\boldsymbol{\beta}$ is estimated using maximum likelihood. As this method is not robust, Cantoni and Ronchetti (2001) proposed a robust alternative based on Mallow's quasi-likelihood estimation with the following estimating equations:

$$\sum_{i=1}^{n} \boldsymbol{\Psi}\left(y_i, \mu_i\right) = \sum_{i=1}^{n} \left[\nu\left(y_i, \mu_i\right) w\left(\mathbf{x}_i\right) \mu_i' - \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}_{y|\mathbf{x}}\{\nu\left(y_i, \mu_i\right)\} w\left(\mathbf{x}_i\right) \mu_i' \right]\,, \tag{2}$$

where $\boldsymbol{\Psi}(\cdot, \cdot)$ are the estimating equations (and is solved by setting them equal to zero), $\nu\left(\cdot, \cdot\right)$ and $w(\mathbf{x})$ are weight functions and $\mu_i' = \frac{\partial \mu_i}{\partial \beta}$. The last term between the square brackets is added to ensure Fisher consistency, and is abbreviated to $a(\boldsymbol{\beta})$.

This type of M-estimator has an asymptotic normal distribution with variance $\Omega = M\left(\boldsymbol{\Psi}, F\right)^{-1} Q\left(\boldsymbol{\Psi}, F\right) M\left(\boldsymbol{\Psi}, F\right)^{-1}$, where $Q\left(\boldsymbol{\Psi}, F\right) = \mathrm{E}\{\boldsymbol{\Psi}\left(y, \mu\right) \boldsymbol{\Psi}\left(y, \mu\right)^\top\}$ and $M\left(\boldsymbol{\Psi}, F\right) = \mathrm{E}\left\{\frac{\partial}{\partial \beta} \boldsymbol{\Psi}\left(y, \mu\right)\right\}$. The function $\boldsymbol{\Psi}$ can regulate the robustness of the estimator by reducing the influence of outlying observations. As Ψ is defined by two weight functions, Cantoni and Ronchetti (2001) propose $\nu\left(y_i, \mu_i\right) = \psi_c\left(r_i\right) \frac{1}{V^{1/2}(\mu_i)}$. In this function r_i are the Pearson residuals equal to $\frac{y_i \mu_i}{V^{1/2}(\mu_i)}$, $V(\mu_i)$ is the variance of μ_i for $i = 1, \ldots, n$, such that $V^{1/2}\left(\mu_i\right)$ is the standard deviation. Finally, ψ_c is the Huber function defined by

$$\psi_c(r) = \begin{cases} r & |r| \leq c\,, \\ c\,\mathrm{sign}(r) & |r| > c\,. \end{cases} \tag{3}$$

If we then also take $w\left(\mathbf{x}_i\right) = 1$ and $a(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[\psi_c(r_i)\frac{1}{V^{1/2}(\mu_i)}\mu_i']$, it leads to the Huber quasi-likelihood estimator that estimates the parameters by solving

$$\sum_{i=1}^{n}\{\psi_c\left(r_i\right)\frac{1}{V^{1/2}\left(\mu_i\right)}\mu_i' - a\left(\boldsymbol{\beta}\right)\} = \mathbf{0}\,. \tag{4}$$

Cantoni and Ronchetti (2001) derive $M\left(\boldsymbol{\Psi}, F\right)$ and $Q\left(\boldsymbol{\Psi}, F\right)$ for both binomial and Poisson models. As logistic regression with binary outcomes is equal to a binomial model with logit link, these derivations can be used to find the Huber quasi-likelihood estimates and subsequently the probabilistic posterior distributions.

In this work both the non-robust and the robust glm model are implemented in `R`. The former using the `glm`() function in `stats` package (R Core Team, 2013). The latter using the `glmrob`() function in the `robustbase` package (Maechler et al., 2019).

2.1.2 Support Vector Machine

Support vector machine is a kernel method as it uses kernels as a nonlinear higher-dimensional mapping tool to make previously nonlinear classification become linear. It minimizes the structural risk instead of the empirical risk, where the latter leads to minimizing the mean square error (MSE). A SVM method takes a training data set of $\mathbf{x}_i \in \mathbb{R}^g$ and $y_i \in \{1, -1\}$ for $i = 1, \ldots, N_1$ and maps it as follows: $\rho(\cdot) : \mathbb{R}^g \mapsto \mathbb{R}^h$ where $h > g$. A general SVM then solves the following minimization problem:

$$\begin{aligned}
\min_{\mathbf{w},b,\xi} \quad & \frac{1}{2}\mathbf{w}^\top\mathbf{w} + C\sum_{i=1}^{n}\xi_i\,, \\
\text{s.t.} \quad & \mathbf{y}_i(\mathbf{w}^\top\rho(\mathbf{x}_i) + a) \geq 1 - \xi_i\,, \\
& \xi_i \geq 0, \quad i = 1, \ldots, N_1\,,
\end{aligned}$$

where $\mathbf{w}$ is a (not necessarily normalised) normal vector to the hyperplane and a a real constant. As the first constraint can be violated in case a discriminating hyperplane does not exist in $\mathbb{R}^h$, slack variables ξ_i are introduced.

Using the calibration by Platt (1999), the raw estimates of the SVM before subsequent classification (d_i), are then transformed and used to fit a posterior predictive cumulative density function (CDF) as follows:

$$\pi_i = \mathbb{P}\left(y_i = 1|t_i\right) = \frac{1}{1 + \exp(At_i + B)}\,, \tag{5}$$

where d_i is transformed to values closer to probabilities (t_i) through $t_i = \frac{d_i+1}{2} \in [0, 1]$. The expression above can be optimized by minimizing the following expression:

$$\min_{A,B} \quad \sum_{i=1}^{N}\{y_i \log\left(\pi_i\right) + \left(1 - y_i\right)\log\left(1 - \pi_i\right)\}\,.$$

A linear kernel for determining d_i with subsequent posterior probabilities through Platt scaling corresponds to regular glm with a logit link, as it assumes that the binary estimates are proportional to the log odds of a positive sample (Platt, 1999). The regularization term in the SVM method makes it less prone to outliers. Furthermore, this method allows for alternative kernels, such as the frequently used radial kernel.

This method is also implemented in `R` using the `svm`() function in the `e1071` package Meyer et al. (2019). The data set input is a subset of the training set, as a percentage is left out for calibration. The model belonging to this subset is tuned using a a basic search for the optimization of the hyperparameters. This is done by using the `tune` function and ranging the values for the cost $C \in \{0.001, 0.01, 0.1, 1, 5, 10, 100\}$ and of $\gamma \in \{0.1, 1, 10, 100\}$, However, this search is only done for one repetition at the start of the simulation, and those values are used throughout all the other repetitions. Platt scaling is then computed through the `optim_Platt` function from Platt (1999). This function takes as inputs the predicted transformed decision values t_i of the calibration hold-out set and their real outcome values of y_i, together with the number of positive instances N^+ and negative instances N^- in the calibration set.

2.1.3 Random Forest

Random forest (RF) is an ensemble method, as it is composed of singular trees. The mechanism behind RF can be used for both supervised classification and regression. In this work we only study binary classification, meaning we are interested in $p(y_i = 1|\mathbf{x}_i) = p_i$. As such, we will discuss random forest as a classification method. The input information for random forests is a random training sample $\mathcal{A}$, defined as follows: $\mathcal{A} = (\mathbf{x}_1, y_1), \dots, (\mathbf{x}_{N_1}, y_{N_1})$.

Using the notation from Scornet (2015) we define the aforementioned probability in a random forest as a collection of M classification rules or trees (m_n) through $\frac{1}{M}\sum_{j=1}^{M} m_n(\mathbf{x}; \Theta_j, \mathcal{A})$. The goal of one m_n is to predict y_i by estimating $m(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}] = p(y = 1|\mathbf{x})$ via the information from $\mathbf{x}$ and $\mathcal{A}$. Each tree takes as input a subset of $\mathcal{A}$ denoted $\mathcal{A}^j$, which is chosen by resampling $\mathcal{A}$ by means of a random variable Θ_j. This gives:

$$m_n(\mathbf{x}; \Theta_j, \mathcal{A}) = \sum_{\mathbf{x}_j \in \mathcal{A}^j(\Theta_j)} \frac{\mathbb{1}_{\{\mathbf{x}_i \in C_n(\mathbf{x}; \Theta_j, \mathcal{A})\}} y_i}{N_n(\mathbf{x}; \Theta_j, \mathcal{A})}\,, \tag{6}$$

where $C_n(\mathbf{x}; \Theta_j, \mathcal{A})$ is the cell that contains $\mathbf{x}$ and $N_n(\mathbf{x}; \Theta_j, \mathcal{A})$ is the number of observations in the cell. The estimate m_n is a Borel measurable function of $\mathbf{x}$ and $\mathcal{A}$ and is consistent if:

$$\lim_{n\to\infty} \mathbb{E}\{m_n(\mathbf{x}; \Theta_j, \mathcal{A}) - m(\mathbf{x})\} \to 0\,. \tag{7}$$

Each tree has multiple splits, and each of these splits are made maximizing the CART-decision rule. For binary outcomes this rule is based on the Gini impurity measure. Let C be a certain cell and $N_n(C)$ the number of data points in C, where $p_{0,n}(C)$ is the probability of 0 in C and $p_{1,n}(C)$ the probability of 1 in C. A cut splits the $j^{\text{th}} \in \{1, \dots, g\}$ predictor at z. This splits C into two parts: $C_L = \{\mathbf{x} \in C : \mathbf{x}^{(j)} < z\}$ and $C_R = \{\mathbf{x} \in C : \mathbf{x}^{(j)} \geq z\}$. This brings a certain function $L_{class,n}$ that can be maximized based on the empirical probabilities of the cell for all cuts z in j and for all j:

$$L_{class,n}(j, z) = p_{0,n}(C)p_{1,n}(C) - \frac{N_n(C_L)}{N_n(C)} \times p_{0,n}(C_L)p_{1,n}(C_L) - \frac{N_n(C_R)}{N_n(C)} \times p_{0,n}(C_R)p_{1,n}(C_R)\,. \tag{8}$$

Random forest needs calibration, as it tends to be conservative in its posterior probabilities, meaning that the majority of predicted probabilities will lie in the middle of the interval between $[0, 1]$. This would automatically mean a lower scoring rule, as errors further from these two outcomes are squared for the Brier score and logarithmic increase for the logarithmic score. To compensate for this, Zadrozny and Elkan (2002) introduced isotonic regression-based calibration applied to a calibration data set. This calibration set corresponds to a hold-out set of the training data. Calibration is done by means of isotonic regression instead of Platt scaling for RF, as the former decreases the variance of single trees and the latter cannot, therefore obtaining better calibration (Niculescu-Mizil and Caruana, 2005).

Isotonic regression assumes monotonicity of the uncalibrated posterior probabilities and the Pool Adjacent Violators Algorithm (PAVA) is generally used for isotonic regression, which works as follows: firstly all observations are sorted based on the uncalibrated posterior probabilities to $p_1 \leq p_2 \leq \ldots \leq p_N$, whereupon all the calibrated probabilities are first taken equal to the true outcome $p_i^\star = y_i \in \{0, 1\}$. Passing through all the ordered values of p_i and their $p_i^\star$, whenever for each consecutive pair $(p_i^\star, p_{i+1}^\star)$ $p_i^\star > p_{i+1}^\star$ occurs, they are placed in a new set. Their new calibrated values will become $(p_i^\star, p_{i+1}^\star) = \frac{p_i^\star + p_{i+1}^\star}{2}$. Each consecutive pair is again trasversed and the whole process is repeated until no violation of ordering occurs.

The random forest is implemented in `R` using the `randomForest` package with probability predictions of the outcome variable equal to 1 (Liaw and Wiener, 2002). Calibration through isotonic regression is done by deleting all the duplicates and ordering the calibration set, then applying isotonic regression through the `isoreg` function. To be able to compare the logarithmic scores when an observation is predicted in the wrong 0 or 1 bins , they are transformed to 1e$-$8 and $1 -$ 1e$-$8 respectively such that the logarithmic score is not -∞ (or `NaN` in `R`). As Platt scaling does not have this problem, this method of calibration is also added in the same manner as for the support vector machines.

2.2 Evaluation Metrics

The goal of probabilistic predictions is to maximize the sharpness subject to calibration (Gneiting and Katzfuss, 2014). To then evaluate how close each model comes towards this goal, correct evaluation metrics need to be used.

As mentioned by Gneiting and Raftery (2007), for proper comparisons of predictions based on posterior probability models, the optimal evaluation metrics are proper scoring rules. The most used proper scoring rules for binary probabilistic predictions are the Brier score, logarithmic score and spherical score (Savage, 1971). These scoring rules evaluate the model by giving a certain value to each observation in the evaluation set, based on the predictive posterior density or cumulative density function and the actual realization. A sharp predictive distribution results from a model having high confidence in prediction thus corresponds to efficiency of a model and calibration occurs when the observations are statistically equal to random draws from this predictive posterior distribution, thus corresponds to the consistency of a model. Calibration needs to be evaluated separately. This is done by use of a calibration diagram/reliability curve, further explained in Section 2.2.1.

During evaluation of the models one compares a realization/ outcome (y_i) with the predicted probability (p_i) obtained from inputting the information set available $(\mathcal{A})$ in the modeled CDF distribution, and predicting the realization y_i. Therefore, Gneiting (2012) introduced the prediction space which is a tuple of the two $(\mathbf{p}, \mathbf{y})$, which can be used to estimate the scoring rules at each observation.

2.2.1 Calibration diagram/Reliability curve

The probabilistic predictions of binary outcomes are conditionally calibrated if: for each value of the probabilistic prediction, the ratio of $y_i = 1$ of all the outcomes belonging to the probabilistic predictions is also equal to p. So for instance for all observations that have a probabilistic prediction equal to $p = 0.5$, 50% of the observations have as outcome $y = 1$. This can be examined by means of a calibration diagram or reliability curve. This curve plots the predicted probabilities against the observed relative frequencies.

For probabilistic predictions of binary models, it is done by partitioning all predictions into K bins (for example $p^{k=1} \in [0, 0.05]$ to $p^{k=20} \in [0.95, 1]$) and then computing the relative frequency of occurrence of true outcome $y_i = 1$ divided by total number of observations in the bin, against the average of the associated predicted probabilities $(\bar{p}^k)$. These two values should ideally be equal to each other. When the curve differs from the diagonal, the model is not conditionally calibrated; if part of the curve lies above the diagonal then there is overestimation present, whereas parts below the diagonal signify underestimation.

2.2.2 Brier score

This score was first introduced by Brier (1950) for the evaluation of meteorological phenomena. Suppose for a binary event that we have a probabilistic prediction of the i^{th} observation being equal to 1 p_i, where the outcome is y_i. We obtain the Brier score (BS) by:

$$BS(p_i) = \frac{1}{N_2} \sum_{i=1}^{N_2} (y_i - p_i)^2 \ , \tag{9}$$

where N_2 is the number of observations in the test set. This score weights the over-estimation of the true probability corresponding to under-estimation, meaning that $p_i - q$ gets an equal score as $p_i + q$ (Machete, 2013). This type of score also squares the error between the outcome $y_i \in \{0, 1\}$, which leads to moderate models with few predictions in the tails get a lower score. The values of BS lie between $[0, 1]$ and the lower the score the better. A coin toss prediction has an expected BS equal to 0.25. When the true outcome is $y_i = 0$, the Brier score behaves as shown in Figure 2.1.

2.2.3 Logarithmic score

This strictly proper scoring rule was first introduced by Good (1953). This is a local scoring rule in the sense that it only depends on the value the model attains at the observation. This also makes The logarithmic scoring rule be related to Shannon's entropy. For the computation of LS the same terms apply as for BS; for a given y_i and p_i one obtains LS by:

$$LS(p_i) = \frac{1}{N_2} \sum_{i=1}^{N_2} \{ y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \} \, . \tag{10}$$

This scoring rule gives values in between $(-\infty, 0]$, and the higher the score the better. A coin toss prediction that uses zero information has a score equal to $\log(0.5)$. Furthermore, this scoring rule assigns higher values to extreme predictions in the right directions. However, it is hypersensitive to wrong predictions Selten (1998). This occurs when $y_i = 1$ and the estimated probability $p_i = 0$ the score assigned is equal to -∞, despite many other observations having good scores. Therefore, it cannot be compared to any other method as it will always perform worse. He also argues that it is prone to non-robust behaviour, as one completely wrong point makes this scoring rule ineffective for evaluation. Moreover, extremely small probabilities of for example 10^{-5} or 10^{-8} can easily be affected by the accuracy of the observations and the computing system, making the logarithmic score differs quite some when compared to the true values. When the true outcome is $y_i = 0$, the logarithmic score behaves as seen in Figure 2.1.

2.2.4 Spherical score

The geometrical representation of this strictly proper scoring rule is the cosine of the angle between the predicted probability vector $(p_i, 1 - p_i)$ and the true outcome vector $(y_i, 1 - y_i)$. When the true outcome is $y_i = 0$, the spherical score behaves as seen in Figure 2.1.

$$SP(p_i) = \frac{1}{N_2} \sum_{i=1}^{N_2} \frac{y_i p_i + (1 - y_i)(1 - p_i)}{\sqrt{p_i^2 + (1 - p_i)^2}} . \tag{11}$$

The scores per observation lie between $[0, 1]$, where larger scores are preferred. When no information is used, which would always give $p_i = 0.5$, then the SP score is equal to $0.5\sqrt{2}$.

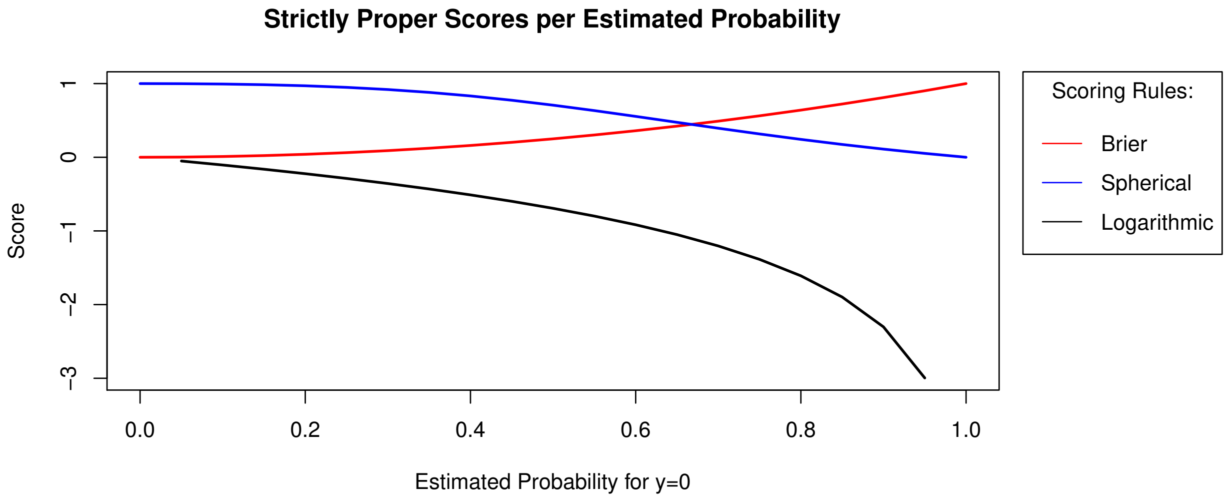

Figure 2.1: Scoring values when $y_i = 0$ for all strictly proper scoring rules

3 Robustness with Strictly Proper Scoring Rules

In Section 3.1 we introduce the necessary tools and notation. In Section 3.2 we derive the asymptotic biases of the estimator scoring rules due to contamination. The relative efficiency between the two functionals is derived in Section 3.3. Finally, the sensitivity of the two models in the presence of outliers is shown in Section 3.4, where the outliers have increasingly higher leverage.

3.1 Notation

Data for the true value in the predictive performance evaluation (y_i) come from the model distribution F also labeled as test data, whereas (training) data from model estimation (p_i) can come from either F_ϵ if contamination is present, or F if no contamination is present. To study the effect of deviations from the ideal model on strictly proper scoring rules, the gross error model by Huber (1964) used:

$$F_\epsilon = (1-\epsilon)F + \epsilon\Delta_z\,, \tag{12}$$

where z is a statistical unit on (y,x) and Δ_z is a Dirac measure with all its mass at z. Denote the statistical functional $T(F_\theta)$ and $\tau(F_\theta)$, the former estimates the parameters θ of a model F using the training set $(\mathbf{X}_1, \mathbf{y}_1)$, whilst the latter is a strictly proper scoring rule evaluating the predictions of the testing set $(\mathbf{X}_2, \mathbf{y}_2)$ and is equal to $\int \phi\left(y_2, p\{x_2, T(F)\}\right) dF$. Where $p\{x_2, T(F)\}$ is the estimated probability of an observation in the testing set $(x_2 \in \mathbf{X}_2)$.

A key tool for studying the (local) robustness is the Influence Function (Hampel, 1974). The Influence Function (IF) of an estimator can be viewed as a description of the local stability of a functional. This is the first order derivative with respect to ϵ and is defined as follows:

$$IF(z;T,F) = \lim_{\epsilon\downarrow 0} \frac{T\{(1-\epsilon)F + \epsilon\Delta_z\} - T(F)}{\epsilon} = \frac{\partial T(F_\epsilon)}{\partial \epsilon}\,. \tag{13}$$

The IF is a convenient heuristic tool to study (local) robustness properties. If it is not bounded then it means that the asymptotic bias is unbounded, and in presence of contamination the estimator can be arbitrarily biased.

3.2 Bias due to Contamination

If there is a (small) amount of contamination in the data due to the presence of a possible contamination, then the evaluation measure $\tau(F_\epsilon)$ can be approximated using a first order von Mises expansion:

$$\tau(F_\epsilon) = \tau(F) + \epsilon \left.\frac{\partial \tau(F_\epsilon)}{\partial \epsilon}\right|_{\epsilon=0} + o(\epsilon)\,. \tag{14}$$

La Vecchia et al. (2012) prove that the first-order Von Mises expansion is sufficient only if ϵ is small and that higher-order terms need to be considered, as they can give additional insights about the local robustness properties of the statistical functional, this is further studied in Section 3.2.2. The bias b of the evaluation of the contaminated model $\tau(F_\epsilon)$ with respect to score at the model $\tau(F)$ can be written as:

$$b = \tau(F_\epsilon) - \tau(F) = \epsilon \left. \frac{\partial \tau(F_\epsilon)}{\partial \epsilon} \right|_{\epsilon=0} + o(\epsilon)\,. \tag{15}$$

Nonetheless, this type of expansion is only valid when the statistical functional is differentiable, the remainder term $(o(\epsilon))$ converges to: $\sqrt{\epsilon} o(F_\epsilon - F) \overset{P}{\rightarrow} 0$ and ϵ is sufficiently small. $\tau(F_\epsilon)$ is contaminated through $T(F_\epsilon)$, which occurs when a classic method is used to estimate the coefficients under a (training set) contaminated true model. Under the same conditions, $\tau(F)$ is not contaminated by $T(F_\epsilon)$ and hence can be viewed (in a simulation setting) as the scoring rule of the predictions using estimated coefficients from a robust method.

3.2.1 First Order Influence Function

When all assumptions are met, we know that the von Mises expansion holds. This means that the bias can be calculated by evaluating the first (and second) order derivatives of $\tau(F_\epsilon)$ with respect to ϵ, when $\epsilon = 0$. We evaluate the scoring rules on the test set $(\mathbf{X}_2, \mathbf{y}_2)$, meaning that each predicted probability uses $x_2 \in \mathbf{X}_2$ for p_i and $y_2 \in \mathbf{y}_2$, such that the functional version becomes equal to $\phi\left(y_2, p\{x_2, T(F)\}\right)$. Decomposing the first order differentiation of $\tau(F_\epsilon)$ in (15) leads to:

$$\begin{aligned}
\left. \frac{\partial \tau(F_\epsilon)}{\partial \epsilon} \right|_{\epsilon=0} &= \left. \frac{\partial}{\partial \epsilon} \int \phi\left[y_2, p\{x_2, T(F_\epsilon)\}\right] d\{(1-\epsilon)F + \epsilon \Delta_z\} \right|_{\epsilon=0}, \qquad (16)\\
&= \left. \frac{\partial}{\partial \epsilon} \int (1-\epsilon)\phi\left[y_2, p\{x_2, T(F_\epsilon)\}\right] dF \right|_{\epsilon=0} + \left. \frac{\partial}{\partial \epsilon} \left(\epsilon \phi\left[y_2, p\{x_2, T(F_\epsilon)\}\right]\right) \right|_{\epsilon=0},\\
&= - \int \phi\left[y_2, p\{x_2, T(F)\}\right] dF + \int \frac{\partial \phi\left(y_2, \omega\right)}{\partial \omega} \cdot \frac{\partial p\{x_2, \eta\}}{\partial \eta} \cdot \left. \frac{\partial T(F_\epsilon)}{\partial \epsilon} dF \right|_{\epsilon=0}\\
&\quad + \phi\left[y_2, p\{x_2, T(F)\}\right] + \epsilon \frac{\partial \phi\left(y_2, \omega\right)}{\partial \omega} \cdot \frac{\partial p\{x_2, \eta\}}{\partial \eta} \cdot \left. \frac{\partial T(F_\epsilon)}{\partial \epsilon} \right|_{\epsilon=0},\\
&= -\tau(F) + \int \frac{\partial \phi\left(y_2, \omega\right)}{\partial \omega} \cdot \frac{\partial p\{x_2, \eta\}}{\partial \eta} dF \cdot \mathrm{IF}(z; T, F) + \phi\left[y, p\{x_2, T(F)\}\right],
\end{aligned}$$

where η is evaluated at $T(F)$ whilst ω at $p\{x_2, T(F)\}$, and the derivative of T with respect to ϵ is the influence function of the estimator. This means that the bias of a scoring rule because of contamination depends on the scoring rule $\tau(F)$ itself, the function in the population version integral $\phi\left[y, p\{x, T(F)\}\right]$ and an integral times the influence function of the introduced contamination on the statistical functional $T(F)$.

Furthermore, if there is no contamination in the training set, such that $\mathbf{X}_1, \mathbf{y}_1$ are uncontaminated by the statistical unit z, the term $\mathrm{IF}(z; T, F)$ drops out. When there are outliers in the training set, the estimated probability $(p\{x_2, T(F)\})$ changes as $T(F_\epsilon)$ becomes biased and thus $\mathrm{IF}(z; T, F)$ might not be negligible anymore. Finally, when there is contamination in the test set, some observations $(x_2, y_2) \in (\mathbf{X}_2, \mathbf{y}_2)$ change, affecting all the terms that depend on these values.

The evaluation of the bias for the strictly proper scoring rules of probabilistic binary outcome predictions when contamination is introduced can be obtained using the expression in (16). For both the Brier score and the logarithmic score this is a straightforward expression, making each term interpretable. This allows us to assess their behaviour. The spherical rule on the other hand gives a less interpretable expression.

This then gives the following first order derivative of the Brier score:

$$\begin{aligned}\left.\frac{\partial BS(F_\epsilon)}{\partial \epsilon}\right|_{\epsilon=0} =& - BS(F) + \int 2\left(p\{x_2, T(F)\} - y_2\right) \cdot \frac{\partial p\{x_2,\eta\}}{\partial \eta} dF \cdot \mathrm{IF}(z; T, F) \\ &+ \left(p\{x_2, T(F)\} - y_2\right)^2 \,,\end{aligned} \tag{17}$$

where a negative Brier score arises from the first term ($\tau(F)$) in (16). The last term is the squared error of the predicted probability of an observation in the test set and will lie between $[0, 1]$. The middle term is the most interesting as it includes the influence function. For the bias to become small, the integral term before the influence function should cancel IF out (as it might not be bounded). When the estimated distribution function is close to the true distribution, the estimated probabilities are close to the true outcomes, making the integral of $(p\{x_2, T(F)\} - y_2) \cdot \frac{\partial p\{x_2,\eta\}}{\partial \eta}$ small.

The same first order derivative can be found for the logarithmic score:

$$\begin{aligned}\left.\frac{\partial LS(F_\epsilon)}{\partial \epsilon}\right|_{\epsilon=0} =& \int \left[\frac{y_2 - p\{x_2, T(F)\}}{p\{x_2, T(F)\}} \cdot \frac{\partial p\{x_2,\eta\}}{\partial \eta} - \frac{1-y_2}{1-p\{x_2, T(F)\}} \cdot \frac{\partial p\{x_2,\eta\}}{\partial \eta}\right] dF \cdot \mathrm{IF}(z; T, F) \\ &+ y_2 \log\left(p\{x_2, T(F)\}\right) + (1-y_2)\log\left(1 - p\{x_2, T(F)\}\right) - LS(F)\,,\end{aligned} \tag{18}$$

where the last term is the logarithmic score of the uncontaminated model. Moreover, the integral term is simply the population version of the estimating equations of logistic regression. This inherently means that if this term is evaluated over the functional, the value attained is equal to 0. As a result, the influence function exerts no effect on the bias. Nonetheless, this is never fully the case in a finite set, as the population version is always approximated by summing the value inside the integral over all observations.

Finally, for the spherical score the expression below is found:

$$\begin{aligned}\left.\frac{\partial SP(F_\epsilon)}{\partial \epsilon}\right|_{\epsilon=0} =& \int \frac{(2y_2 - 1)}{\sqrt{p\{x_2, T(F)\}^2 + (1 - p\{x_2, T(F)\})^2}} \cdot \frac{\partial p\{x_2,\eta\}}{\partial \eta} dF \cdot \mathrm{IF}(z; T, F) \\ &- \int \frac{\frac{(y_2 p\{x_2,T(F)\} + (1-y_2)(1-p\{x_2,T(F)\}))(-2p\{x_2,T(F)\}+1)}{\sqrt{p\{x_2,T(F)\}^2 + (1-p\{x_2,T(F)\})^2}}}{p\{x_2, T(F)\}^2 + (1 - p\{x_2, T(F)\})^2} \cdot \frac{\partial p\{x_2,\eta\}}{\partial \eta} dF \cdot \mathrm{IF}(z; T, F) \\ &- SP(F) + \frac{y_2 p\{x_2, T(F)\} + (1-y_2)(1 - p\{x_2, T(F)\})}{\sqrt{p\{x_2, T(F)\}^2 + (1 - p\{x_2, T(F)\})^2}}\,.\end{aligned} \tag{19}$$

From this equation we see that the integral terms are not easily interpretable. We would need simulation to study how each term in the first order derivative behaves. Henceforward, we will focus on the more comprehensive proper scoring rules, and not include spherical scores in the results. Another reason to not study the spherical score further is that it is strictly proper with respect to the same class as the Brier score ($\mathcal{L}_2$) (Gneiting, 2012).

3.2.2 Second Order Influence Function

As stated by La Vecchia et al. (2012), when ϵ becomes too large (or if the first order expression is zero in all cases, as is the case in Croux et al. (2008a)) higher-order terms in the Von Mises expansion improve the approximation of the bias. Similarly to the first order of the Von Mises expansion, the second order term gives the following expression:

$$\begin{aligned}
\left.\frac{\partial^2 \tau(F_\epsilon)}{\partial \epsilon^2}\right|_{\epsilon=0} &= \left.\frac{\partial^2}{\partial \epsilon^2}\int \phi\left[y, p\{x, T(F_\epsilon)\}\right] d\{(1-\epsilon)F + \epsilon\Delta_z\}\right|_{\epsilon=0}, \\
&= \left.\frac{\partial^2}{\partial \epsilon^2}\int (1-\epsilon)\phi\left[y, p\{x, T(F_\epsilon)\}\right] dF\right|_{\epsilon=0} + \left.\frac{\partial^2}{\partial \epsilon^2}\left(\epsilon\phi\left[y, p\{x, T(F_\epsilon)\}\right]\right)\right|_{\epsilon=0}, \\
&= \left.\int \frac{\partial^2}{\partial \epsilon^2}\phi\left[y, p\{x, T(F_\epsilon)\}\right] dF\right|_{\epsilon=0} - \left.\frac{\partial^2}{\partial \epsilon^2}\left(\epsilon\int \phi\left[y, p\{x, T(F_\epsilon)\}\right] dF\right)\right|_{\epsilon=0} \\
&\quad + \left.\frac{\partial}{\partial \epsilon}\left(\phi\left[y, p\{x, T(F_\epsilon)\}\right] + \epsilon\frac{\partial \phi\left[y, p\{x, T(F_\epsilon)\}\right]}{\partial \epsilon}\right)\right|_{\epsilon=0}, \\
&= \left.\int \frac{\partial \phi\left(y, \omega\right)}{\partial \omega}\cdot\frac{\partial p(x,\eta)}{\partial \eta}\cdot\frac{\partial^2 T(F_\epsilon)}{\partial \epsilon^2} dF\right|_{\epsilon=0} \\
&\quad + \left.\int \left(\frac{\partial^2 \phi\left(y, \omega\right)}{\partial \omega^2}\cdot\frac{\partial p(x,\eta)}{\partial \eta} + \frac{\partial \phi\left(y, \omega\right)}{\partial \omega}\cdot\frac{\partial^2 p(x,\eta)}{\partial \eta^2}\right)\cdot\frac{\partial T(F_\epsilon)}{\partial \epsilon} dF\right|_{\epsilon=0} \\
&\quad - 2\left.\int \frac{\partial \phi\left(y, \omega\right)}{\partial \omega}\cdot\frac{\partial p(x,\eta)}{\partial \eta}\cdot\frac{\partial T(F_\epsilon)}{\partial \epsilon} dF\right|_{\epsilon=0} - \epsilon\left.\int \frac{\partial^2 \phi\left[y, p\{x, T(F_\epsilon)\}\right]}{\partial \epsilon^2} dF\right|_{\epsilon=0} \\
&\quad + \left.\frac{\partial}{\partial \epsilon}\phi\left[y, p\{x, T(F_\epsilon)\}\right]\right|_{\epsilon=0} + \left.\frac{\partial \phi\left[y, p\{x, T(F_\epsilon)\}\right]}{\partial \epsilon}\right|_{\epsilon=0} + \epsilon\left.\frac{\partial^2 \phi\left[y, p\{x, T(F_\epsilon)\}\right]}{\partial \epsilon^2}\right|_{\epsilon=0}, \\
&= \left.\int \frac{\partial \phi\left(y, \omega\right)}{\partial \omega}\cdot\frac{\partial p(x,\eta)}{\partial \eta}\cdot\frac{\partial^2 T(F_\epsilon)}{\partial \epsilon^2} dF\right|_{\epsilon=0} \\
&\quad + \left.\int \left(\frac{\partial^2 \phi\left(y, \omega\right)}{\partial \omega^2}\cdot\frac{\partial p(x,\eta)}{\partial \eta} + \frac{\partial \phi\left(y, \omega\right)}{\partial \omega}\cdot\frac{\partial^2 p(x,\eta)}{\partial \eta^2}\right)\cdot\frac{\partial T(F_\epsilon)}{\partial \epsilon} dF\right|_{\epsilon=0} \\
&\quad - 2\left.\int \frac{\partial \phi\left(y, \omega\right)}{\partial \omega}\cdot\frac{\partial p(x,\eta)}{\partial \eta}\cdot\frac{\partial T(F_\epsilon)}{\partial \epsilon} dF\right|_{\epsilon=0} + 2\left.\frac{\partial}{\partial \epsilon}\phi\left[y, p\{x, T(F_\epsilon)\}\right]\right|_{\epsilon=0}, \\
&= \int \frac{\partial \phi\left(y, \omega\right)}{\partial \omega}\cdot\frac{\partial p(x,\eta)}{\partial \eta} dF\cdot\{\mathrm{IF2}(z;T,F) - 2\cdot\mathrm{IF}(z;T,F)\} \\
&\quad + \int \left(\frac{\partial^2 \phi\left(y, \omega\right)}{\partial \omega^2}\cdot\frac{\partial p(x,\eta)}{\partial \eta} + \frac{\partial \phi\left(y, \omega\right)}{\partial \omega}\cdot\frac{\partial^2 p(x,\eta)}{\partial \eta^2}\right) dF\cdot\mathrm{IF}(z;T,F) \\
&\quad + 2\left.\frac{\partial}{\partial \epsilon}\phi\left[y, p\{x, T(F_\epsilon)\}\right]\right|_{\epsilon=0}.
\end{aligned} \tag{20}$$

In this equation many of the same parameters are present as in the first order term. The difference lies in the inclusion of a second order influence function ($\mathrm{IF2}(z;T,F)$), the exclusion of the scoring rule itself, and the new terms $\frac{\partial^2\phi(y,\omega)}{\partial\omega^2}$ and $\frac{\partial^2 p(x,\eta)}{\partial\eta^2}$. The first new term is the second order differentiation of the population version of the scoring rule evaluated at $p\{x_2, T(F)\}$. The second new term is the first order derivative of the density function evaluated at $T(F)$ (or the second order derivative of the distribution function evaluated at

$T(F)$). We see again that the influence functions (IF and $IF2$) are multiplied by an integral. To see how all these terms behave, further simulation would be needed.

3.3 Asymptotic Relative Efficiency

To compare the scoring rules of two estimators and determine asymptotic equivalence, the (asymptotic) relative efficiency between two proper scoring rules from two different estimators $\tau_1(F)$ to $\tau_2(F)$ is computed. This can be done as follows:

$$ARE\{\tau_1(F), \tau_2(F)\} = \frac{AVar\{\tau_2(F)\}}{AVar\{\tau_1(F)\}}\,, \tag{21}$$

where $AVar\{\tau(F)\}$ is the asymptotic variance of a scoring rule. Under suitable regularity conditions (Fernholz, 1983), the asymptotic variance of the scoring rules is calculated by:

$$AVar\{\tau(F)\} = \int \mathrm{IF}(z; T, F)\mathrm{IF}(z; T, F)^\top dF\,. \tag{22}$$

Let $\tau(F)$ be the scores from predictions of an estimator $T(F)$, then: $\sqrt{n}\{\tau(F_m) - \tau(F)\} \rightarrow N(0, AVar\{\tau(F)\})$. This asymptotic variance can be approximated by simulating the scoring rule a sufficient number of times. For both classic and robust logistic regression an analytical expression exists for $\mathrm{IF}(z; T, F)$. For the former $\mathrm{IF}_C(z; T, F) = x - \mu$ and for the latter expression is $IF_R(z; \boldsymbol{\psi}, F) = M(\boldsymbol{\psi}, F)^{-1}\psi(z, \mu)$ (recall $\boldsymbol{\psi}$ are the estimating equation of the robust estimator and $\mu = \mathrm{E}(y)$) (Cantoni and Ronchetti, 2001).

3.4 Sensitivity Analysis

To illustrate the effect contamination has on the bias of one of the aforementioned scoring rules, we analyze the more bound strictly proper scoring rule: the Brier score. The effect on the bias of the Brier score is quantified by contaminating only one observation in increasing quantity (in either the training set or in the testing set).

This approach allows us to compare how close the first order von Mises approximation of the bias is to simulated values of the true bias. For the sensitivity analysis of contamination in the training set, the simulated bias is computed M times using the difference between the Brier score of the increasing-quantity contaminated model $BS(F_\epsilon)$ and the uncontaminated Brier score $BS(F)$:

$$N \cdot (BS(F_\epsilon) - BS(F))\,, \tag{23}$$

where N is the number of observations per simulation. These simulated Brier scores are calculated from predictions of a univariate model obtained as follows:

$$\mathbf{y} \sim B\left(N, \frac{\exp{(3x_i)}}{1 + \exp{(3x_i)}}\right)\,, \tag{24}$$

where $x_i \sim \mathbb{N}(0, 1)$ for $i = 1, \ldots, N$ and the contaminated values are created by contaminating the first observation with $u \in \{1, 2, \ldots, 10\}$ to obtain: $(x_{1,c}, y_{1,c}) = (u - 10, 1)$ for sensitivity of $y_{1,c} = 1$ and $(x_{1,c}, y_{1,c}) = (u, 0)$ for $y_{1,c} = 0$. Simulating this bias $M = 200$ times also gives an approximation of the bias.

The outcome of the sensitivity analysis of the Brier score is shown in Figure 3.2, where the first order approximation of the bias of the Brier score for $y = 1$ and $y = 0$ are the red lines and the simulated values per contamination level u of x_1 are the boxplots. We see that the first order approximation of the bias is close to the true value from the simulations, meaning that the first order suffices for a good approximation of the bias.

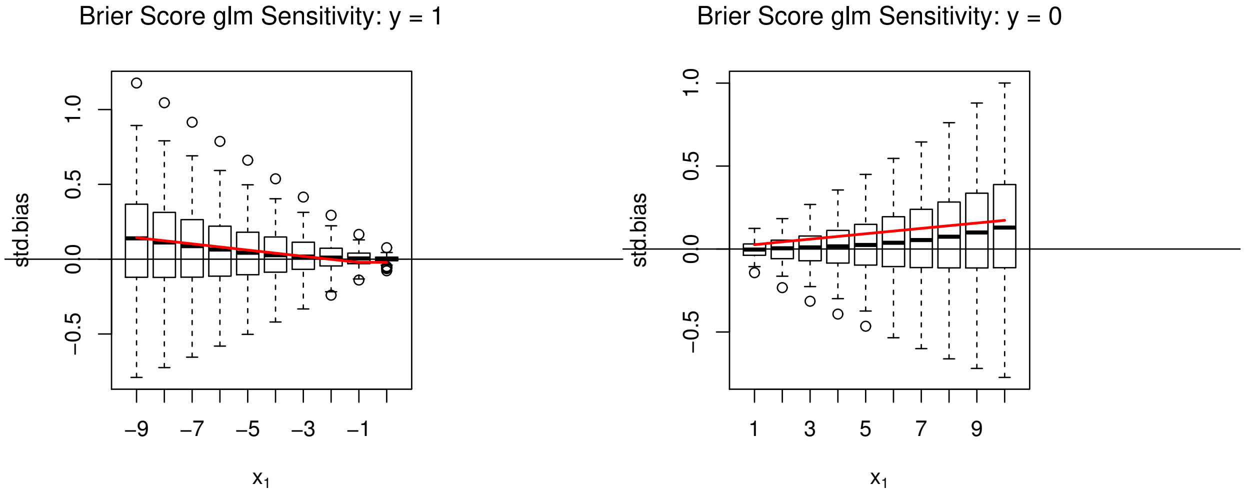

Figure 3.2: Sensitivity of the Brier Score for the training set contamination using glm predictions

4 Simulation

To study the behavior of scoring rules using different prediction methods we carry out a Monte Carlo simulation study. We simulate data from Binomial distribution, hence the canonical link function is logit (McCullagh and Nelder, 1989). We firstly consider two linear predictors. Both have three explanatory variables, which are generated from zero mean multivariate normal distribution with covariances equal to $cov(x_i, x_j) = (1/2)^{|i-j|}$ (variant of a Toeplitz matrix). The first linear predictor (LP1) is $g(x) = x_1 + 0.5x_2 + 0.25x_3$. In this case the data is noisy and good discrimination is not possible, i.e. the predicted probabilities close to zero and one are rare. The second linear predictor (LP2) is $g(x) = 2x_1 + x_2 + 0.05x_3$. In this case the discrimination is better and many probabilities are close to zero and one. The case with first linear predictor would mimic applications in Economics and Marketing and the second linear predictor is more sensitive to heavy disturbances in the true model, i.e. mislabeling occurs more often in the presence of heavy-tailed disturbances. The sample sizes are $N_1 = 1000$ in the training set and $N_2 = 2000$ in the testing set. We repeat the experiment $R = 500$ times. Next, to study the robustness issues we add contamination and study the effects of contamination when these different contaminations are in training set and both in training and testing set:

A We firstly study two levels of contamination equal to $\epsilon = 0.02$ and $\epsilon = 0.04$, as we are interested in finding enough differences in both the predictions and the coefficients under contamination. As well as the two different linear predictors LP1 and LP2 and which has a more critical effect on the predictions. We also briefly study the breakdown of predictions and coefficient estimates of both robust and classic logistic regression when $\epsilon = 0.08$ and when the more critical linear predictor is chosen. These two contamination levels chosen add moderate outliers with probability $\epsilon = 0.02$ or $\epsilon = 0.04$, replacing the original observations by outliers with the point mass $(y_1, x_1, x_2, x_3) = (1, -2, -2, 0)$. This generates leverage outliers in the design space. Notice that, it is not straightforward to detect them, since the outliers are only two standard deviations away from the mean and the third predictor is not contaminated. In the second contamination scenario we study heavy outliers. The contamination mechanism is the same as above, but the point mass is at $(y_1, x_1, x_2, x_3) = (1, -5, -5, -5)$. This contamination has more dramatic consequences for both estimation and prediction. Although, it is not difficult to detect these outliers in our particular setup, in a higher dimensional setting such drastic outliers are possible and hardly detectable (without careful diagnostics).

B The second scenario is contamination by heavy-tailed errors. With probability $\zeta = 0.05$ (or 0.1) we replace the original observations by $y = \mathcal{I}\{g(x) + e \geq 0\}$, where $e \sim 5 \cdot t(2)$, where 2 is the number of degrees of freedom.

C The third scenario is misspecification of the link function. We use the same linear predictors, but the data is generated from the probit model, i.e., $y = \mathcal{I}\{g(x) + e \geq 0\}$, where $e \sim N(0, 1)$.

5 Applications

To enhance the simulation study (findings) we also test each proposed method on real data. For each data set we investigate the reliability and Brier score of the predictions from each method. As the methods used are similar to the ones used by Croux et al. (2008b), namely logistic regression, we investigate similar well-known data sets, as well as data sets that are more applicable to probabilistic prediction.

5.1 Similar Research Data Sets

Croux et al. (2008b) investigated four data sets to see if their new diagnosing measure for detecting influential points in the training data set found the well-documented outliers. Three of them are used in this experiment, as they also documented which observations were labeled as influential. This can be used to explain the reliability curves and scores:

i. The vaso constriction data set (Pregibon, 1981) consists of the dependent binary variable: the presence or absence of vaso constriction of the skin of the digits after air inspiration, and two explanatory variables: volume of air inspired and the inspiration rate. Pregibon (1981) log-transforms both explanatory variables and found two outliers at observation numbers 7 and 13.

ii. The goal in the foodstamp data set (Stefanski et al., 1986) is to predict the dependent variable, the participation of US citizens in the US Food Stamp Program, by using three economic factors as explanatory variables: tenancy, supplemental income and monthly income. In total there are 150 observations, whereof 24 participated in the program. The benchmark model for this data set is logistic regression. There is one outlier in the income data set with a much higher value of income than the other participants

iii. The last data set used is about leukemia (Cook and Weisberg, 1982), where the cancer survival time of 33 patients is recorded. The dependent variable is 1 when the patient survives more than 52 weeks. This variable is explained by the number of white blood cells of each patient and whether they have a certain characteristic in these blood cells.

5.2 Open Source Probabilistic Prediction Application Data Sets

Next to the well-documented data sets, it is also of interest to see calibration and scores of real-life data sets. For that we choose three open-source data sets where probabilistic prediction is of more importance than discrimination. Furthermore, as we are also interested in imbalance, we have one data set where this occurs.

i. Credit card fraud data set from a payment service provider in Belgium is composed of credit card transactions. The dependent variable is whether fraud has occurred or not and in 492 out of 284,807 transactions that is the case, meaning this data set is heavily imbalanced at 0.172%. The explanatory variables include 28 PCA transformed variables for privacy measures, as well as variables such as transaction amount and time of transaction.

ii. Banknote authentification data set by Lohweg (2012), online available through Dua and Graff (2017), tries to explain the falsification of banknote by $1,373$ instances of extracted images. Four continuous variables are available from the extracted images: variance, skewness, kurtosis and entropy of the Wavelet Transformed image.

iii. Breastcancer Wisconsin data set, which is provided by Wolberg et al. (1992) is also made available by Dua and Graff (2017) on the repository. The goal is to explain the malignancy of breastcancer tumors of 699 patients by use of 30 continuous variables such as the radius, texture and area of the tumor. This data set is also slightly imbalanced, as 212 out of 699 patients have a malignant tumor.

6 Results

The performance of proper scoring rules and the reliability of the predictions is studied by the simulations described in Section 4. This is done to see if the scores from classic logistic regression predictions at the true model are close to those of robust logistic regression predictions. For this, we firstly study the average and the asymptotic relative efficiency of the scoring rules per simulation for all the method and also compare the asymptotic relative efficiency of the estimators for both classic and robust logistic regression. Finally, calibration per method is also researched as predictions should in all cases be reliable.

An overview of the results is given here. The results of the simulation study encompass those at the model explained in Section 6.1 where outlier contamination is also added, Heavy tailed error contamination is studied in Section 6.2. In Section 6.3 the results for a misspecified link function can be found. Next, The outcome of the application of proper scoring rules and calibration on real-world examples is explained in Section 6.4 and Section 6.5. In these results we discern between the similar research applications and open source applications, respectively.

6.1 Outlier Contamination

At the model. The strictly proper scoring rules of predictions from classic logistic regression are close to the predictions from robust logistic regression. This is firstly seen in (17) and (18) in Section 3, where the approximation of the bias of the Brier score and logarithmic score indicates that at the model the scores are close to each other. This closeness is also visible in Table 6.2 first (and for other data generating processes in Table A.1), where the average Brier scores and logarithmic scores for both linear predictors at the model are similar for classic and robust logistic regression.

Second, the asymptotic relative efficiency of robust logistic regression compared to the classic counterpart are close to 1. For LP1 it is 1.00 and LP2 0.995, shown in Table 6.1 and Table A.2 respectively. This means that for these two linear predictors the scores from the predictions of robust logistic regression are at least as efficient. This is unforeseen, as it is well known that classic logistic regression using MLE is asymptotically efficient at the model, whilst robust logistic regression using quasi-maximum likelihood estimation is less efficient. This is also confirmed by our results of the coefficient estimates at the model in Figure 6.3 for LP1, as the average coefficient estimates are $(0.0015, 1.0, 0.51, 0.87)$ for both classic and robust logistic regression, whilst the average standard errors of the coefficient estimates from robust logistic regression lie higher than for classic logistic regression: at $(0.77, 1.0, 1.0, 0.89) \cdot 10^{-1}$ versus $(0.75, 0.99, 1.0, 0.87) \cdot 10^{-1}$.

All these scores are conditional on the predictions being calibrated. The classic and robust logistic regression predictions at the model are both calibrated and their status is seen in Table 6.3. These statusses are obtained using the reliability curves in Appendix A.3 use a similar data generating process, but with many more observations and one repetition.

We also see in Table 6.2, Table 6.1 and Table A.2 that SVM and the three different variations of random forest are slightly biased and less efficient than the logistic regression. For both linear predictors SVM performs best among the machine learning methods. Additionally, calibrating random forest by means of isotonic regression or Platt scaling has a positive effect on the Brier score and no negative effect on the reliability of the predictions, as shown in Table 6.3.

The logarithmic score is negative towards isotonic regression calibration visible in Table 6.2. Isotonic regression calibrates predicted probabilities close to the boundaries of 0 and 1, completely at these bounds. Therefore, if one prediction has a probability that is near the incorrect true outcome (e.g. $p_i = 0.9$ whilst $y_i = 0$), isotonic regression has a high chance of calibrating it to be at the incorrect outcome (e.g. $p_i = 1$). It is thus recommended not to use isotonic regression as a calibration method in combination with the logarithmic score.

Table 6.1: Relative efficiencies of Brier score for LP1 at the model

	Classic	Missp.	Robust	Rand. Forest	RF iso.	RF Platt.	SVM
Classic glm	1	0.973	1.00	2.82	1.47	1.15	1.14
Missp. glm		1	1.03	2.90	1.51	1.19	1.17
Robust glm			1	2.81	1.46	1.15	1.14
Rand. Forest				1	0.520	0.409	0.405
RF iso. cal.					1	0.786	0.778
RF Platt cal.						1	0.990
SVM Platt							1

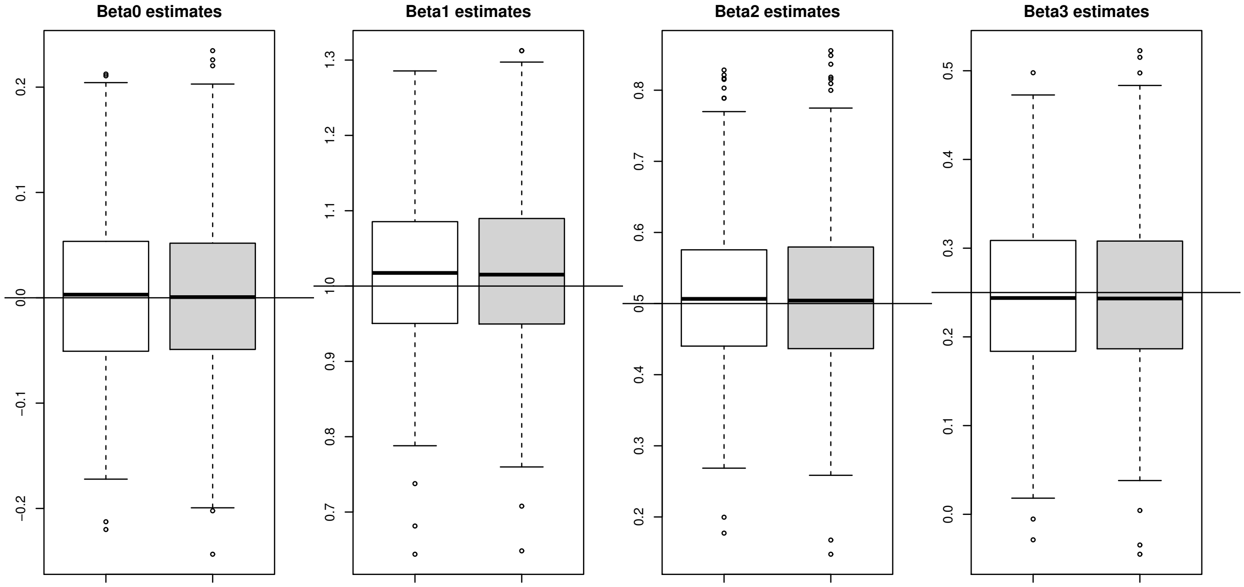

Figure 6.3: Boxplots of all estimated coefficients of classic and robust logistic regression per predictor at the true model of LP1, where the grey boxplots are the robust glm coefficient estimates over all runs per predictor and the horizontal line is the true value of the coefficient

Table 6.2: Average of Brier scores and logarithmic scores for all methods of the runs with outlier contamination at the model, where Uncnt'n means uncontaminated, Tr. Cont'n means training set outlier contamination and T&T Cont'n training and testing set outlier contamination

	Brier Score			**Logarithmic Score**		
	Uncont'n	*Tr. Cont'n*	*T&T Cont'n*	*Uncont'n*	*Tr. Cont'n*	*T&T Cont'n*
LP2, $\epsilon = 0.02$, moderate outliers						
Clas. glm	***0.126***	0.131	0.139	***-0.392***	-0.412	-0.445
Miss. glm	0.125	0.131	0.131	-0.392	-0.410	-0.410
Rob. glm	***0.126***	0.126	0.135	***-0.392***	-0.393	-0.447
Rand. For.	0.137	0.138	0.140	-0.450	-0.455	-0.459
RF iso.	0.138	0.139	0.141	-0.522	-0.516	-0.520
RF Platt	0.136	0.137	0.139	-0.432	-0.435	-0.440
SVM	0.128	0.130	0.137	-0.403	-0.411	-0.429
LP2, $\epsilon = 0.04$, moderate outliers						
Clas. glm	***0.126***	0.143	0.157	***-0.391***	-0.447	-0.490
Miss. glm	0.125	0.141	0.141	-0.391	-0.445	-0.445
Rob. glm	***0.126***	0.126	0.144	***-0.392***	-0.393	-0.499
Rand. For.	0.137	0.140	0.140	-0.448	-0.461	-0.461
RF iso.	0.138	0.141	0.141	-0.522	-0.522	-0.520
RF Platt	0.136	0.139	0.138	-0.430	-0.439	-0.438
SVM	0.128	0.133	0.144	-0.402	-0.420	-0.447
LP1, $\epsilon = 0.02$, moderate outliers						
Clas. glm	***0.181***	0.184	0.190	***-0.538***	-0.546	-0.562
Miss. glm	0.182	0.185	0.185	-0.542	-0.549	-0.549
Rob. glm	***0.181***	0.181	0.188	***-0.538***	-0.539	-0.562
Rand. For.	0.200	0.201	0.202	-0.642	-0.646	-0.647
RF iso.	0.192	0.193	0.195	-0.652	-0.651	-0.652
RF Platt	0.189	0.191	0.191	-0.562	-0.566	-0.567
SVM	0.184	0.185	0.189	-0.548	-0.551	-0.561
LP1, $\epsilon = 0.04$, moderate outliers						
Class. glm	***0.181***	0.190	0.199	***-0.538***	-0.562	-0.584
Miss. glm	0.182	0.191	0.191	-0.542	-0.566	-0.566
Rob. glm	***0.181***	0.182	0.195	***-0.538***	-0.541	-0.582
Rand. For.	0.200	0.203	0.201	-0.642	-0.649	-0.644
RF iso.	0.192	0.195	0.194	-0.652	-0.656	-0.652
RF Platt	0.189	0.192	0.191	-0.562	-0.570	-0.566
SVM	0.184	0.187	0.193	-0.548	-0.557	-0.570

Outlier contamination in training set. When adding moderate outliers in the training set, we see that for both linear predictors the predictions of robust logistic regression outperform the classic logistic regression predictions. The average Brier score and logarithmic score for classic logistic regression becomes biased, whilst the average scores for robust logistic regression remain unbiased, as shown in Table 6.2. A possible reason for the biased classic logistic regression predictions is the bias from its coefficient estimates when one outlier is present within the data set, presented in Figure 6.4. The average coefficient estimates when the contamination level is at $\epsilon = 0.04$ are equal to $(0.16, 0.66, 0.13, 0.40)$ for classic logistic regression. Contrarily, we see that robust logistic regression gives close to unbiased average coefficient estimates equal to $(0.035, 0.89, 0.39, 0.28)$. Robust logistic regression is also more efficient for both predictors, as the asymptotic relative efficiency that is less than 1, shown in Table A.3 for LP1 and in Table A.4 for LP2.

Unlike classic logistic regression, machine learning methods are not heavily affected by outlier contamination in the training set, indicating some bias-robustness in our simulation design. This can be explained in random forest by the sampling of the subset biasing the results, and the averaging of all the classification trees from the subsets moderating this bias. The robustness of SVM is mainly due to the support vectors being the only necessary uncontaminated observations for a clear discrimination rule. However, the predictions from machine learning methods (still) have worse average scores than robust logistic regression.

When good discrimination is possible, calibrating random forest has no effect on the average scores in Table 6.2 and the calibration status in Table 6.3. On the other hand calibration has much added value when there is no good discrimination. Also, Platt scaling has a slightly negative effect on the reliability curve.

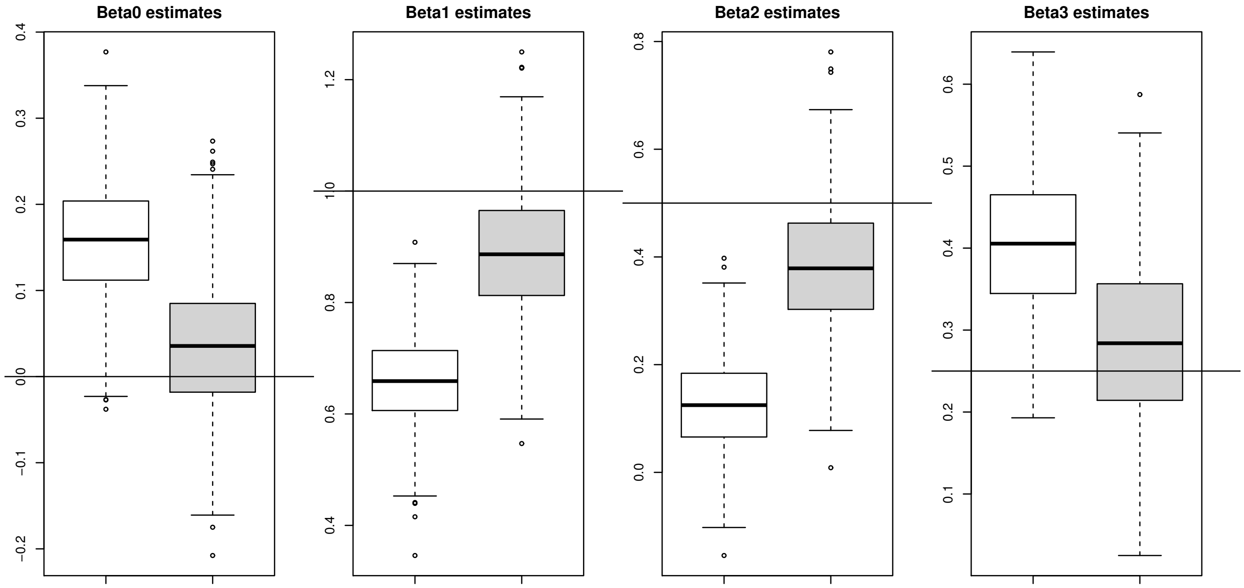

Figure 6.4: Boxplots of all estimated coefficients of classic and robust logistic regression per predictor for LP1 with $\epsilon = 0.04$ contamination level of moderate outliers in the training set, where the grey boxplots are the robust glm coefficient estimates over all runs per predictor and the horizontal line is the true value of the coefficient

Outlier contamination in training and testing set. When outliers are present in the training and testing set, robust logistic regression has a better average Brier score than classic logistic regression in Table 6.2. On the other hand, robust logistic regression obtains a worse average logarithmic score. This holds for both linear predictors and the effect is more pronounced when the contamination level is increased.

Table 6.3: Overview of Brier score and calibration status for LP1 and LP2 with two contamination levels $\epsilon \in \{0.02, 0.04\}$ of moderate outliers, where + indicates the method's predictions are calibrated, $\sim$ somewhat calibrated and $-$ uncalibrated. Actual plots are in Appendix A.3

	Uncontaminated		**Training Cont'n**		**T&T Cont'n**	
	Score	*Calibrated*	*Score*	*Calibrated*	*Score*	*Calibrated*
LP2, $\epsilon = 0.02$**, moderate outliers**						
Classic glm	0.125	+	0.131	$\sim$	0.157	-
Missp. glm	0.125	+	0.130	$\sim$	0.156	-
Robust glm	0.125	+	0.125	+	0.143	+
Rand. Forest	0.141	+	0.141	$\sim$	0.147	$\sim$
RF iso.	0.139	+	0.140	$\sim$	0.148	$\sim$
RF Platt	0.137	$\sim$	0.138	$\sim$	0.143	-
SVM	0.127	+	0.129	+	0.134	$\sim$
LP2, $\epsilon = 0.04$**, moderate outliers**						
Classic glm	0.125	+	0.142	-	0.185	$\sim$
Missp. glm	0.125	+	0.140	-	0.187	-
Robust glm	0.125	+	0.125	+	0.162	+
Rand. Forest	0.141	+	0.144	$\sim$	0.143	+
RF iso.	0.139	+	0.143	$\sim$	0.141	$\sim$
RF Platt	0.137	$\sim$	0.141	$\sim$	0.139	$\sim$
SVM	0.127	+	0.127	$\sim$	0.137	-
LP1, $\epsilon = 0.02$**, moderate outliers**						
Classic glm	0.180	+	0.183	$\sim$	0.198	-
Missp. glm	0.182	+	0.184	$\sim$	0.201	$\sim$
Robust glm	0.180	+	0.180	+	0.194	+
Rand. Forest	0.205	$\sim$	0.206	$\sim$	0.209	-
RF iso.	0.195	$\sim$	0.195	$\sim$	0.199	$\sim$
RF Platt	0.188	$\sim$	0.188	$\sim$	0.191	-
SVM	0.183	$\sim$	0.184	$\sim$	0.188	-
LP1, $\epsilon = 0.04$**, moderate outliers**						
Classic glm	0.180	+	0.189	$\sim$	0.213	$\sim$
Missp. glm	0.182	+	0.190	$\sim$	0.219	-
Robust glm	0.180	+	0.181	+	0.207	$\sim$
Rand. Forest	0.205	$\sim$	0.216	$\sim$	0.210	$\sim$
RF iso.	0.195	+	0.200	$\sim$	0.198	$\sim$
RF Platt	0.188	$\sim$	0.193	$\sim$	0.191	$\sim$
SVM	0.183	$\sim$	0.187	-	0.190	-

6.2 Heavy-tailed Error Contamination

Heavy-tailed error contamination. The average Brier scores and logarithmic scores from the predictions of classic and robust logistic regression in Table 6.5 are still close when heavy-tailed errors of $5t(2)$ are introduced at a level of $\zeta \in \{0.05, 0.1\}$. However, the asymptotic relative efficiency in Table 6.4 (and Table B.6 in Appendix B.2) is greater than 1. This indicates that the Brier scores from classic logistic regression predictions are more efficient than their robust counterpart. The scores can thus not be considered equal. Next, the predictions from classic and robust logistic regression for both linear predictors remain calibrated, as seen in Table 6.6. This means that although mislabeling occurs, the predictions will still be reliable, more so than the machine learning methods. These other methods become somewhat calibrated or even uncalibrated for LP1.

The boxplots of the coefficient estimates of all repetitions are biased for both classic and robust logistic regression in Figure 6.5 and Figure 6.6. From these two figures we also see that the coefficient estimates for LP1 are more biased than for LP2, although in both cases robust logistic regression shows a smaller bias. Hence, for statistical inference for these two linear predictors, the coefficient estimates of robust logistic regression will be closer to the true values than for classic logistic regression.

The averages of the Brier scores of the machine learning techniques predictions all show a bias in Table 6.5, with SVM obtaining the lowest bias for LP1 and LP2. In terms of efficiency the scores of SVM also outperform the other machine learning techniques. Although, logistic regression in general still has a higher efficiency for LP1 with heavy-tailed error contamination.

The SVM predictions also remain at most somewhat reliable for both linear predictors, unlike Platt scaled random forest predictions that become unreliable for LP1. A possible reason for this is that classic logistic regression is used to obtain the coefficients that calibrate the predictions. This might be biased under heavy-tailed error contamination.

Table 6.4: Asymptotic relative efficiencies of Brier score of LP1 for $\zeta = 0.1$ contamination level of heavy-tailed error of $5t(2)$

	Classic	Missp.	Robust	Rand. Forest	RF iso.	RF Platt.	SVM
Classic glm	1	0.978	1.09	2.12	2.10	1.72	1.41
Missp. glm		1	1.11	2.17	2.14	1.76	1.44
Robust glm			1	1.95	1.93	1.58	1.30
Rand. Forest				1	0.989	0.813	0.666
RF iso. cal.					1	0.821	0.673
RF Platt cal.						1	0.819
SVM Platt							1

Table 6.5: Average of Brier score and logarithmic score for all methods of the runs with contamination by heavy-tailed errors, where Uncnt'n means only contamination by heavy-tailed errors, Tr. Cont'n means training set outlier contamination as well as heavy-tailed error contamination and T&T Cont'n training and testing set contamination with heavy-tailed error contamination.

	Brier Score			**Logarithmic Score**		
	Uncont'n	*Tr. Cont'n*	*T&T Cont'n*	*Uncont'n*	*Tr. Cont'n*	*T&T Cont'n*
LP2, $5t(2)$, $\zeta = 0.1$, $\epsilon = 0.04$**, moderate outliers**						
Clas. glm	***0.108***	0.136	0.150	***-0.355***	-0.437	-0.475
Miss. glm	0.108	0.135	0.135	-0.355	-0.436	-0.436
Rob. glm	***0.107***	0.109	0.127	***-0.352***	-0.359	-0.450
Rand. For.	0.114	0.117	0.117	-0.379	-0.387	-0.388
RF iso.	0.118	0.119	0.119	-0.423	-0.427	-0.428
RF Platt	0.115	0.118	0.117	-0.381	-0.387	-0.387
SVM	0.109	0.118	0.128	-0.360	-0.389	-0.413
LP2, $5t(2)$, $\zeta = 0.05$, $\epsilon = 0.04$**, moderate outliers**						
Clas. glm	***0.133***	0.150	0.163	***-0.415***	-0.465	-0.504
Miss. glm	0.133	0.148	0.148	-0.415	-0.463	-0.463
Rob. glm	***0.133***	0.133	0.151	-***0.415***	-0.416	-0.514
Rand. For.	0.145	0.147	0.147	-0.477	-0.486	-0.485
RF iso.	0.145	0.148	0.147	-0.540	-0.549	-0.547
RF Platt	0.143	0.145	0.145	-0.450	-0.456	-0.457
SVM	0.136	0.140	0.150	-0.424	-0.439	-0.464
LP1, $5t(2)$, $\zeta = 0.1$, $\epsilon = 0.04$**, moderate outliers**						
Class. glm	***0.190***	0.198	0.206	***-0.560***	-0.582	-0.600
Miss. glm	0.191	0.199	0.199	-0.563	-0.585	-0.585
Rob. glm	***0.190***	0.191	0.203	***-0.561***	-0.563	-0.598
Rand. For.	0.210	0.213	0.211	-0.672	-0.679	-0.672
RF iso.	0.201	0.204	0.202	-0.661	-0.666	-0.661
RF Platt	0.197	0.200	0.198	-0.580	-0.587	-0.583
SVM	0.192	0.195	0.201	-0.568	-0.576	-0.588

Heavy-tailed error contamination and outlier contamination in training set. When further introducing outliers in the training set we see in Table 6.5 that robust logistic regression is the method whose average scores are lowest, and remain closest to the average scores without outlier contamination in the training set. It thus shows the smallest bias compared to the average scores of only heavy-tailed error contamination. This latter phenomenon is also visible for the variations of random forest, although their average scores with only heavy-tailed error contamination are higher. The average scores from classic logistic regression predictions show the highest bias. This is in agreement with the results found with only outlier contamination in Section 6.1.

For the asymptotic relative efficiency we see that for LP1 in Table B.7 and for LP2 with the low level of heavy-tailed errors in Table B.5, that the value is smaller than 1. In these cases robust logistic regression scores are more efficient. In terms of reliability in Table 6.6 the only method whose predictions remain completely reliable is robust logistic regression.

In Figure B.2 the random forest predictions are reliable in the middle range of predictions, but become more unstable near the probability borders. Low probabilities are overestimated and high probabilities are underestimated, leading to less decisive results when employing random forest (or SVM).

Platt scaling gives in all cases studied with heavy-tailed error contamination unreliable predictions. SVM predictions are also at most somewhat calibrated. This could be again because of the usage of classic logistic regression to estimate the calibration coefficients. For both linear predictors the scores from SVM predictions and random forest predictions calibrated using Platt scaling do obtain the highest efficiency. Isotonic regression on the other hand, has no negative effect on the reliability and in case of LP1 a big positive effect on the average scores (although for LP2 it has a negative effect on the average). It also gives more efficient results for both linear predictors.

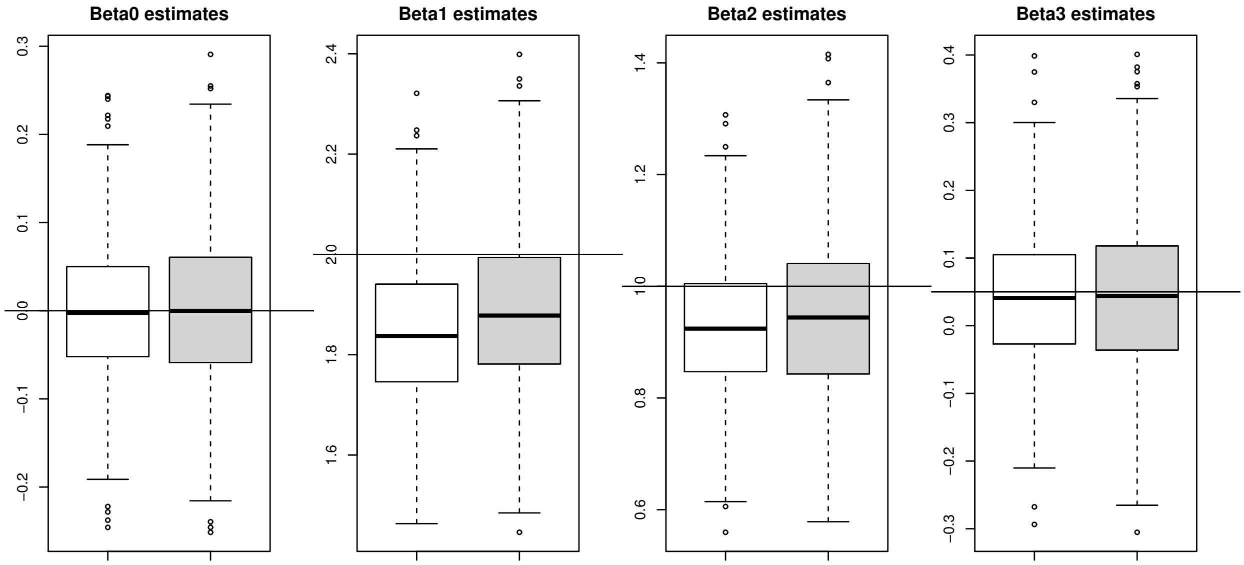

Figure 6.5: Boxplots of all estimated coefficients of classic and robust logistic regression per predictor of LP2 for $\zeta = 0.1$ contamination level of heavy-tailed error of $5t(2)$, where the grey boxplots are the robust glm coefficient estimates over all runs per predictor and the horizontal line is the true value of the coefficient

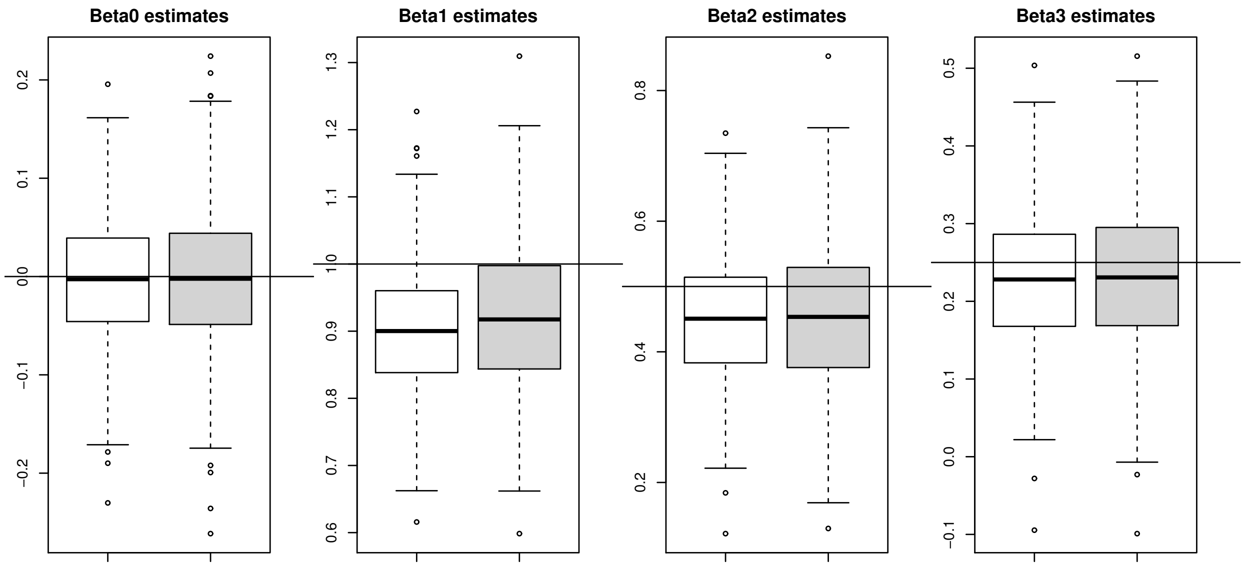

Figure 6.6: Boxplots of all estimated coefficients of classic and robust logistic regression per predictor of LP1 for $\zeta = 0.1$ contamination level of heavy-tailed error of $5t(2)$, where the grey boxplots are the robust glm coefficient estimates over all runs per predictor and the horizontal line is the true value of the coefficient

Heavy-tailed error contamination and outlier contamination in training and testing set. When outliers are also present in the testing set, we see in Table 6.5 that robust logistic regression obtains a similar biased average Brier score and logarithmic score as the classic logistic regression predictions. However, its predictions remain reliable for LP2 and somewhat reliable for LP1 in Table 6.6.

For the machine learning techniques adding outlier contamination also in the testing set has no effect on the conclusions. Random forest predictions still have the best average scores. Calibration still has a positive effect on the average scores of LP1, whilst for LP2 there is no effect. The reliability of the machine learning methods in Table 6.6 also remains unchanged compared to only outlier contamination in the training set.

Table 6.6: Overview of Brier score and calibration status of the runs with contamination by heavy-tailed errors, where $+$ indicates the method's predictions are calibrated, $\sim$ somewhat calibrated and $-$ uncalibrated. Uncontaminated means only contamination by heavy-tailed errors, Training Cont'n means training set outlier contamination as well as heavy-tailed error contamination and T&T Cont'n training and testing set contamination together with heavy-tailed error contamination. Some reliability plots found in Appendix B.3

	Uncontaminated		**Training Cont'n**		**T&T Cont'n**	
	Score	*Calibrated*	*Score*	*Calibrated*	*Score*	*Calibrated*
LP2, $5t(2)$ $\zeta = 0.1$, $\epsilon = 0.04$, **moderate outliers**						
Classic glm	0.139	$+$	0.155	-	0.182	-
Missp. glm	0.139	$+$	0.154	-	0.182	-
Robust glm	0.139	$+$	0.139	$+$	0.174	$+$
Rand. Forest	0.157	$\sim$	0.160	$\sim$	0.157	$\sim$
RF iso.	0.154	$\sim$	0.157	$\sim$	0.157	$\sim$
RF Platt	0.151	$\sim$	0.154	-	0.150	-
SVM	0.140	$+$	0.147	$\sim$	0.148	$\sim$
LP2, $5t(2)$ $\zeta = 0.05$, $\epsilon = 0.04$, **moderate outliers**						
Classic glm	0.132	$+$	0.149	-	0.177	-
Missp. glm	0.132	$+$	0.148	-	0.177	-
Robust glm	0.132	$+$	0.132	$+$	0.168	$+$
Rand. Forest	0.149	$+$	0.152	$\sim$	0.150	$\sim$
RF iso.	0.147	$\sim$	0.149	$\sim$	0.147	$\sim$
RF Platt	0.145	$\sim$	0.147	-	0.144	-
SVM	0.133	$+$	0.140	$\sim$	0.142	$\sim$
LP1, $5t(2)$ $\zeta = 0.1$, $\epsilon = 0.04$, **moderate outliers**						
Classic glm	0.189	$+$	0.198	$\sim$	0.214	-
Missp. glm	0.191	$+$	0.199	$\sim$	0.218	-
Robust glm	0.189	$+$	0.190	$+$	0.215	$\sim$
Rand. Forest	0.219	$\sim$	0.228	$\sim$	0.221	$\sim$
RF iso.	0.205	$\sim$	0.209	$\sim$	0.206	$\sim$
RF Platt	0.196	-	0.201	-	0.196	-
SVM	0.192	$\sim$	0.197	-	0.196	-

6.3 Misspecified Link Function

At the model. When a probit link is the canonical link for GLM, but instead the logit link is used for the coefficient estimates and the predictions, similar average scores and efficiencies are obtained for classic and robust logistic regression in Table 6.7 and Table 6.8 respectively. A possible reason is the similarity of the probit and logit link. This similarity of the link function is also a possible reason for the classic and robust logistic regression predictions still being calibrated in Table 6.9. Hence, the results are comparable to a correctly specified link.

On the other hand, statistical inference without the true link is impossible. The average values of the estimated coefficients for LP1 with the misspecified link is biased, seen in Figure 6.7. These values are $(0.0034, 3.6, 1.8, 0.086)$ for classic logistic regression and for robust logistic regression $(0.0053, 3.5, 1.8, 0.080)$. Recall LP1: $g(x) = 2x_1 + x_2 + 0.05x_3$ with $\beta_0 = 0$. Conclusively, a canonical link is only absolutely necessary when both statistical inference and predictions are the goal of modelling.

Equal to the results found at the model, Brier scores fro SVM predictions are the most efficient and have the best average score out of all machine learning methods, .

Table 6.7: Average of Brier score and logarithmic score for all methods of the all misspecified runs, where Uncnt'n means uncontaminated, Tr. Cont'n means training set contamination and T&T Cont'n training and testing set contamination

	Brier Score			**Logarithmic Score**		
	Uncont'n	*Tr. Cont'n*	*T&T Cont'n*	*Uncont'n*	*Tr. Cont'n*	*T&T Cont'n*
LP1, $\epsilon = 0.04$**, moderate outliers**						
Classic	***0.0807***	0.108	0.124	***-0.255***	-0.356	-0.415
Miss. glm	0.0806	0.105	0.105	-0.255	-0.353	-0.353
Rob. glm	***0.0807***	0.0809	0.0997	***-0.255***	-0.256	-0.462
Rand. For.	0.0918	0.0946	0.0954	-0.302	-0.313	-0.316
RF iso.	0.0928	0.0945	0.0951	-0.365	-0.365	-0.366
RF Platt	0.0911	0.0939	0.0938	-0.304	-0.312	-0.312
SVM	0.0841	0.0903	0.104	-0.268	-0.297	-0.336
LP3, $\epsilon = 0.04$**, moderate outliers**						
Clas. glm	***0.167***	0.169	0.170	***-0.502***	-0.507	-0.510
Miss. glm	0.231	0.235	0.235	-0.653	-0.662	-0.662
Rob. glm	***0.167***	0.168	0.170	***-0.502***	-0.504	-0.510
Rand. For.	0.188	0.189	0.185	-0.594	-0.596	-0.586
RF iso.	0.183	0.183	0.181	-0.607	-0.624	-0.616
RF Platt	0.180	0.180	0.178	-0.543	-0.543	-0.537
SVM	0.173	0.171	0.172	-0.518	-0.515	-0.517

Table 6.8: Relative efficiencies of Brier score for LP1 with a misspecified link (logit link instead of probit link) at the model

	Classic	Missp.	Robust	Rand. Forest	RF iso.	RF Platt.	SVM
Classic glm	1	1.00	0.977	1.33	1.77	1.59	1.32
Missp. glm		1	0.977	1.33	1.77	1.59	1.32
Robust glm			1	1.37	1.81	1.63	1.36
Rand. Forest				1	1.33	1.20	0.993
RF iso. cal.					1	0.901	0.748
RF Platt cal.						1	0.831
SVM Platt							1

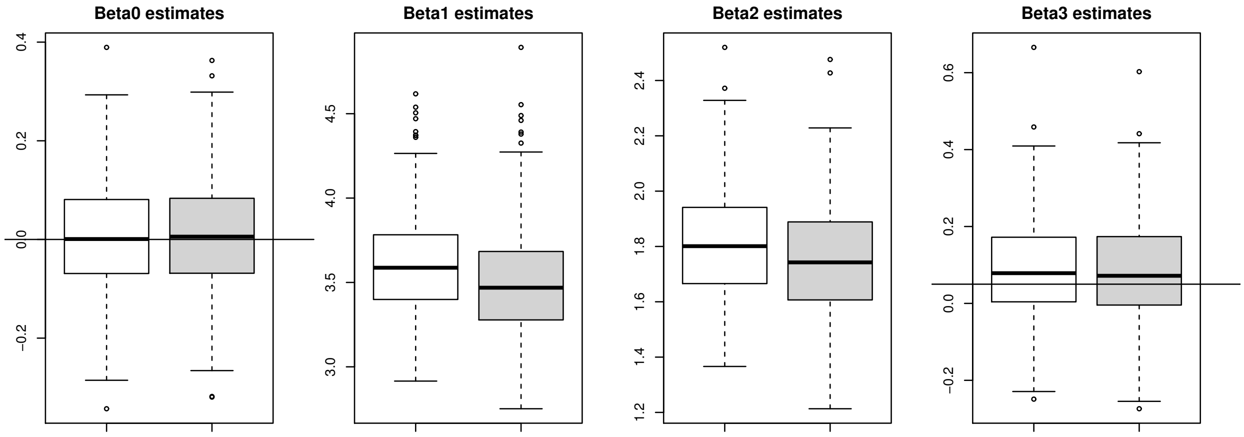

Figure 6.7: Boxplots of all estimated coefficients of classic and robust logistic regression per predictor for LP1 with a misspecified link (probit instead of logit) at the model, where the grey boxplots are the robust glm coefficient estimates over all runs per predictor and the horizontal line is the true value of the coefficient

Outlier contamination in training set. When introducing moderate outliers in the training set, we still see in Table 6.7 that robust logistic regression predictions have the lowest average scores, closely followed by SVM predictions. These are also the two methods whose predictions remain calibrated in Table 6.9.

Outlier contamination in training and testing set. The average scores of both classic and robust logistic regression predictions are much worse than at the (misspecified link) model and are thus broken down. The average logarithmic score of robust logistic regression is worse than that of classic logistic regression. This is due to an observations obtaining a probability very close to 1, whilst its true value is 0, or vice versa.

Table 6.9: Per method an overview of Brier Score and calibration status of LP1 using a misspecified link with moderate outliers at a contamination level of $\epsilon = 0.04$, where + indicates the method's predictions are calibrated, $\sim$ somewhat calibrated and $-$ uncalibrated.

	Uncontaminated		**Training Cont'n**		**T&T Cont'n**	
	Score	*Calibrated*	*Score*	*Calibrated*	*Score*	*Calibrated*
Classic glm	0.080	+	0.11	-	0.16	-
Missp. glm	0.081	+	0.10	-	0.16	-
Robust glm	0.080	+	0.080	+	0.12	+
RF	0.093	$\sim$	0.096	$\sim$	0.096	-
RF iso.	0.092	+	0.096	$\sim$	0.094	$\sim$
RF Platt	0.091	$\sim$	0.095	+	0.094	-
SVM	0.081	+	0.088	+	0.093	+

6.4 Similar Research Application Results

Firstly, the vaso constriction data set is split into 75% training set and 25% testing. Pregibon (1981) documented that observation 7 and 13 in the center of the covariance space are outliers. When both indices fall into the training set, the results from Figure 6.8 and Table 6.10 are obtained: calibrated robust logistic regression predictions and uncalibrated classic ones, with equal Brier scores of 0.11. When one of the two observations falls into the test set the same overall conclusion can be made that robust logistic regression is preferred as it is as sharp as classic logistic regression, but also calibrated. However, when both observations fall into the test set, robust logistic regression also becomes uncalibrated, as its model based on the training set becomes more similar to the classic logistic regression model. However, the robust logistic regression predictions still give slightly better Brier scores, meaning that out of the two methods robust logistic regression is still preferred.

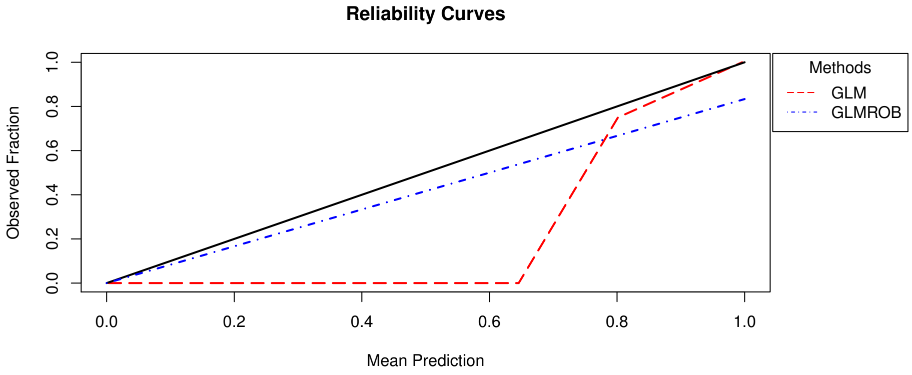

Figure 6.8: Reliability Curves for vasoconstriction data

Brier Scores	
glm	0.11
glmrob	0.11

Table 6.10: Scores for glm & glmrob

For the foodstamp data set used by Croux et al. (2008b) 1234 is taken as the seed for partitioning 80% of the data into the training set and 20% into the testing set, as the data set is relatively small.

We find from Figure 6.9 that the only method that remains at least (somewhat) calibrated is robust logistic regression. Scaling the random forest predictions by means of isotonic regression decreases the calibration of the predictions, and by means of Platt scaling has next to no effect on the calibration status. Again we observe that out of all more machine learning methods SVM is the most calibrated.

On the other hand, the Brier scores for robust and classic logistic regression from Table 6.11 are quite low and similar. Furthermore, we discern that random forest without any further calibration obtains the best Brier score, followed by random forest calibrated using Platt scaling.

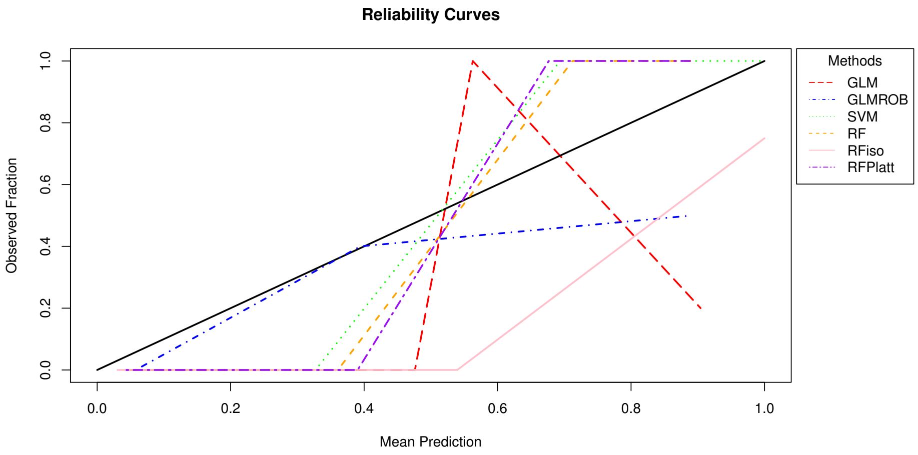

Figure 6.9: Reliability Curves for foodstamp data

Brier Scores	
glm	0.10
glmrob	0.091
SVM	0.26
RF	0.036
RFiso	0.21
RFPlatt	0.052

Table 6.11: Brier Score

In Figure 6.10 we notice that all methods have a relatively bad Brier score (especially considering that coin-toss predictions where $p_i = 0.5$ for $i = 1, \ldots, N_2$ you have a Brier score equal to 0.25) All methods are also uncalibrated except for robust logistic regression. The reason is that in this specific setting of the partitioning between the training data set and the testing data set, the influential observation found by Croux et al. (2008b) is present in the training set. This leads to robust logistic regression being calibrated and having a better Brier score, whilst classic logistic regression is uncalibrated and has a worse score. Comparing the more machine learning based methods, we view that all variations of random forest are uncalibrated, and SVM is somewhat calibrated. This is more or less in line with the Brier scores obtained, as SVM has the best out of all machine learning methods (but is still worse than the value for robust logistic regression).

This all results in robust logistic regression being preferred over all other methods.

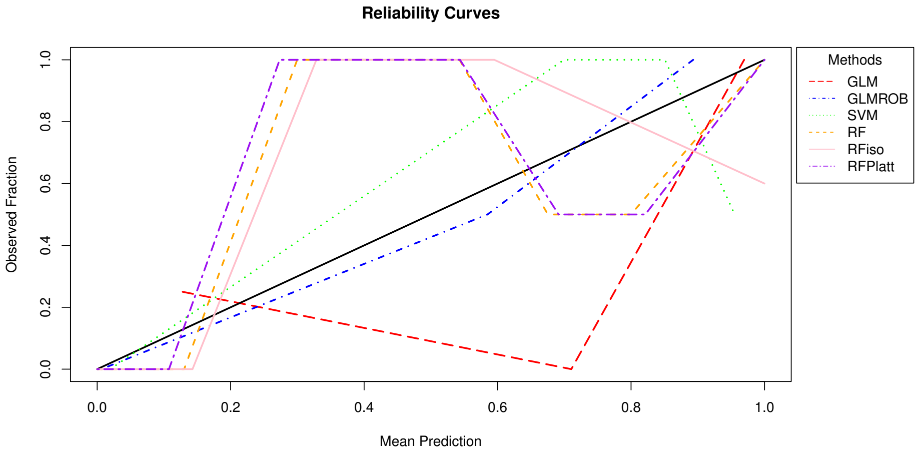

Figure 6.10: Reliability Curves for leukemia data

Brier Scores	
glm	0.20
glmrob	0.14
SVM	0.20
RF	0.22
RFiso	0.30
RFPlatt	0.21

Table 6.12: Brier Score

6.5 Open Source Probabilistic Prediction Application Results

For the credit card data set the seed used is 123, and here a 60% training set and 40% testing set partition is used. Looking at Figure 6.11 most methods overestimate the predictions, as all curves are above the black linear line corresponding to perfect calibration. Except for robust logistic regression predictions and random forest predictions calibrated using isotonic regression. These two methods are thus the ones whose predictions come closest to being reliable, as they follow the perfect calibration line more closely (although both are only somewhat reliable).

When looking at the Brier scores presented in Table 6.13, we see that the reliability of the methods reflects the scores, as the two most calibrated methods (”RFiso” and ”glmrob” in the table) obtain the lowest and thus best Brier score. Furthermore, in general these scores are very good, as a high level of imbalance is present (lowering the scores drastically, as the methods become very accurate in predicting no credit card fraud). This is also in line with the additional simulation results with imbalance in Table A.1.

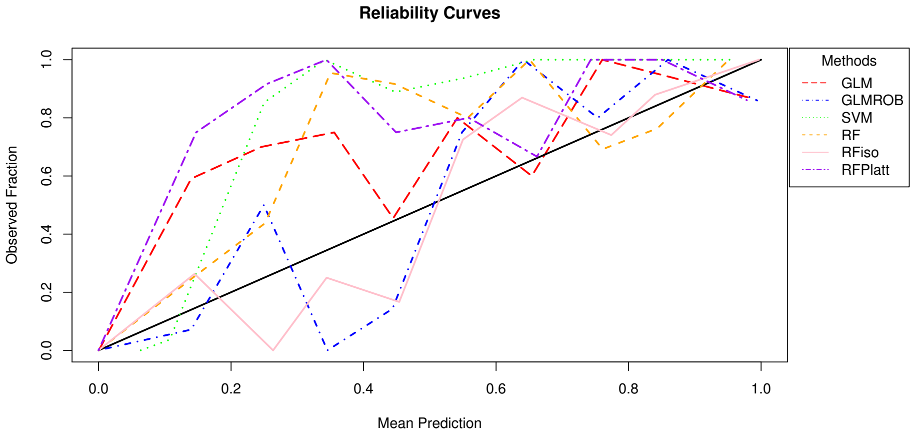

Figure 6.11: Reliability Curves for creditcard data

Brier Scores	
glm	0.00080
glmrob	0.00075
SVM	0.25
RF	0.00074
RFiso	0.00067
RFPlatt	0.00091

Table 6.13: Brier Score

The banknote authentication data set is split it into 85% training and 15% testing with seed 1234. From Figure 6.12 we identify that robust logistic regression is the only method whose predictions are calibrated. All the other methods give uncalibrated predictions. This is in line with the calibration results found in the simulation study. Also analyzing the Brier scores in Table 6.14, we conclude that classic and robust logistic regression predictions have the best Brier score, with the other methods having a Brier score that is between 3 and 7 times as high. This might mean that calibration in this data set is crucial and drives the scores.

Comparing uncalibrated random forest predictions with the isotonically calibrated and Platt scaled predictions we discover that for this data set isotonic regression improves calibration, whilst Platt scaling does not. This again indicates that the usage of classic logistic regression for calibration can have negative effects.

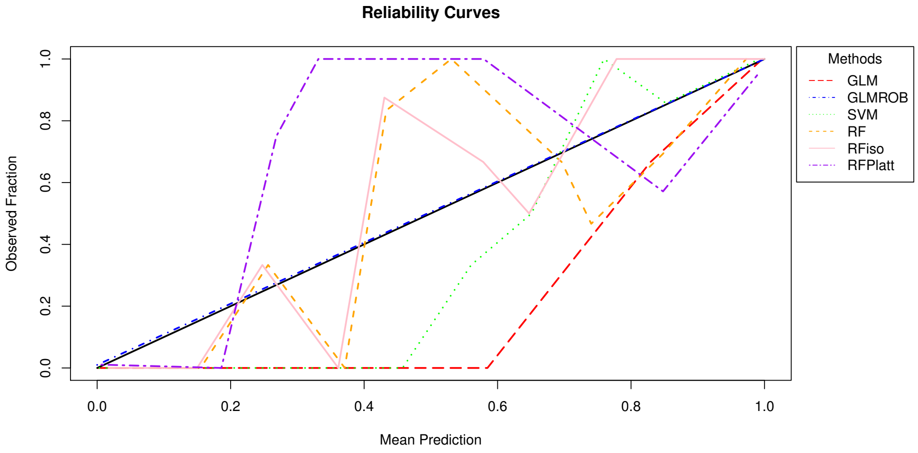

Figure 6.12: Reliability Curves for banknote data

Brier Scores	
glm	0.0058
glmrob	0.0058
SVM	0.015
RF	0.039
RFiso	0.048
RFPlatt	0.046

Table 6.14: Brier Score

For the breast cancer data we use seed 123 with 70% in training set and 30% in testing set, as this data set contains more observations, and we want to have enough observations in the testing set to obtain good Brier scores and good reliability curves. In Figure 6.13 we find that robust logistic regression gives calibrated predictions and classic logistic regression does not. When looking at the calibration status of the predictions of the other methods, we deduce that out of all of them SVM or random forest with Platt scaling give the most calibrated predictions next to robust logistic regression, although it is still uncalibrated when looking at Figure 6.13.

Adding the Brier scores in Table 6.15 to these calibration results we see that robust logistic regression predictions have the best score together with classic logistic regression predictions, although the latter ones are uncalibrated. This means that overall for this data set robust logistic regression would be the preferred method as it is the only one whose predictions are sharp subject to calibration.

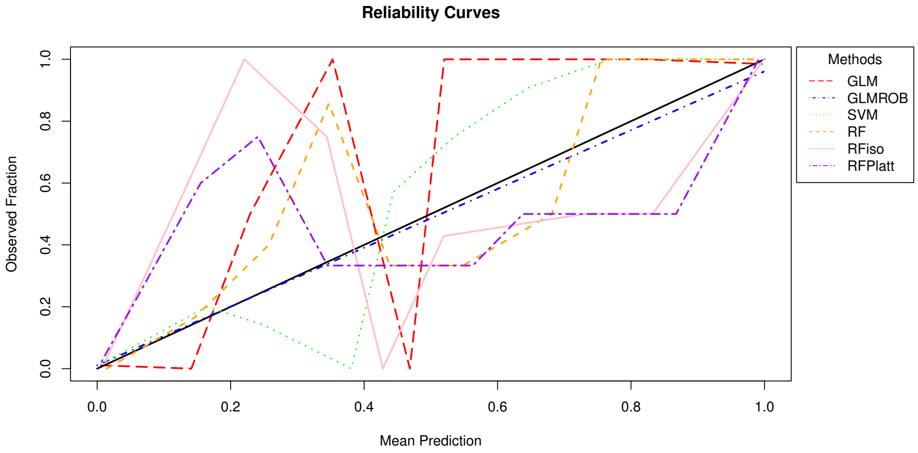

Figure 6.13: Reliability Curves for breastcancer data

Brier Scores	
glm	0.023
glmrob	0.023
SVM	0.057
RF	0.050
RFiso	0.055
RFPlatt	0.058

Table 6.15: Brier Score

7 Discussion

In this section we give a more thorough explanation of the results in Section 6 and their implications for future work in Section 7.1. Afterwards, we delve into the limitations of this thesis as well as recommendations for future work on this thesis in Section 7.2.

7.1 Discussion of Results

Robust logistic regression gives predictions that are at least equally accurate as classic logistic regression, whilst still giving almost unbiased estimates of the coefficients when contamination is present. This is supported by the proof that the strictly proper scoring rules of predictions from classic logistic regression are close to robust logistic regression predictions at the model. This is further reinforced by the average scores being close to each other and the asymptotic relative efficiency of the Brier scores of robust logistic regression predictions compared to the classic counterpart being equal to 1. This is unforeseen, as it is well known that under the true model, maximum likelihood estimation (MLE) is asymptotically efficient, whilst robust counterparts perform worse, which is also confirmed by our results of the coefficient estimates under the true model.

In this particular simulation setting, robust logistic regression outperformed classic logistic regression under training set outlier contamination in terms of average scores and reliability, since classic logistic regression already gives biased coefficient estimates (and thus inaccurate predictions) when one outlier is present within the data set. Contrarily, we know that robust logistic regression gives unbiased coefficient estimates when the contamination level is small (around $\epsilon = 0.05$). This is also reinforced in the simulation study.

These two results indicate that robust logistic regression has a wider range of usage than classic logistic regression. Nonetheless, this is not reflected in current research, where the benchmark is still classic logistic regression. Thus, when accurate probabilistic predictions are the main goal and statistical inference a close second, robust logistic regression is the superior method.

Calibration in general has a beneficial effect on the predictions from random forest, in that both the scores obtained improve and the predictions become more reliable. This is seen by the reliability curves improving in the simulation and in the applications, and the mean scores being at least as good as well. Thus, when using random forest to generate predictions, studying the reliability and adding calibration is advantageous.

Robust logistic regression is the only method out of all of them that is at least somewhat reliable in all the simulation and applications studied. This means that in a very broad setting robust logistic regression remains statistically consistent between the probabilistic predictions and the actual outcome of the dependent variable. This reliability is desirable when predictions from an unknown data set are necessary.

7.2 Limitations and Future Research

The limitations of this simulation study are firstly that Platt scaling utilizes classic logistic regression to calibrate the probabilities. However, throughout this work we see that the coefficients of classic logistic regression are easily affected by outliers. Hence, it is worthwhile to further study robust calibration. A possible way to perform this, is to first quantify the effect of the outliers in the calibration set on the scaling coefficients and afterwards comparing this effect to a robust counterparts of Platt scaling

The hyperparameters of random forest and Support Vector Machine were not tuned per repetition, but solely once at initialization. Furthermore the maximum number of nodes from the random forest hyperparameters is set high such that all kinds of trees are allowed. This might have a negative effect on the scores obtained. Despite this restricted hyperparameter tuning, the main conclusions still hold.

In terms of evaluation, we did not take the computation time into consideration, even though this is of great important with the emergence of big data. Nonetheless, predictions from classic and robust logistic regression are much quicker obtained than those from other machine learning methods.

Finally, in this thesis solely the logarithmic score and the Brier score were researched. However Gneiting and Katzfuss (2014) state that there exists a more robust counterpart of the logarithmic score and the Brier score. Nonetheless, as these two scoring methods are often used as benchmark, it will not impact our main finding, which states that robust logistic regression predictions are equally efficient as classic logistic regression predictions at the model, as they shift the scoring rule to account for contamination.

8 Conclusion

In this thesis the robustness of the strictly proper scoring rules for evaluation of probability estimates of binary outcomes was studied. The focus was on the results of the predictions and their scores at the model, and the effect outlier contamination or heavy-tailed error contamination has. Von Mises expansions of the functional form of the three most common strictly proper scoring rules for probabilistic binary outcomes explained that when the contamination level is small, the bias between the contaminated scoring rule (classic logistic regression) and uncontaminated scoring rule (robust logistic regression) is small as well, and the scores are close.

It is well known that at the model, maximum likelihood estimation is asymptotically efficient and that estimation based on the more robust counterparts derived using quasi-maximum likelihood perform worse. However, in this work we find that evaluation of probabilistic predictions from a classic logistic regression model are equally asymptotically efficient as those of its robust counterpart. Furthermore, when contamination is present, robust logistic regression performs better than classic logistic regression. As robust logistic regression performs at least equally good than classical logistic regression both for simulated and real-world data (unless breakdown of both methods has occurred), the new benchmark for calibrated unbiased and efficient probabilistic predictions should be robust logistic regression.

In real data sets the data generating process is mostly unknown and different types of contamination might be present. The method whose probabilistic predictions obtain best Brier scores (and logarithmic scores) under outlier and/or heavy-tailed error contamination is SVM.

Finally, another extremely important more industry-related conclusion is the usage of calibration to improve predictions. There are many companies nowadays using machine learning methods to obtain predictions. However, in many of these cases they focus too little on the reliability of these predictions and whether they are an accurate reflection of reality. Two quick additional steps lead to an improvement of the predictions. Firstly, reliability curves should always be computed to give an illustrative representation of how well the predictions reflect reality. If this reflection is not accurate enough, an additional second step is to add calibration through either isotonic regression or Platt scaling to improve the reliability of the predictions.

References

Bianco, A. M. and Yohai, V. J. (1996), “Robust estimation in the logistic regression model,” in *Robust statistics, data analysis, and computer intensive methods*, New York: Springer, pp. 17–34.

Brier, G. W. (1950), “Verification of forecasts expressed in terms of probability,” *Monthly Weather Review*, 78, 1–3.

Cantoni, E. and Ronchetti, E. (2001), “Robust inference for generalized linear models,” *Journal of the American Statistical Association*, 96, 1022–1030.

Cook, R. D. and Weisberg, S. (1982), *Residuals and influence in regression*, New York: Chapman and Hall.

Croux, C., Filzmoser, P., and Joossens, K. (2008a), “Classification efficiencies for robust linear discriminant analysis,” *Statistica Sinica*, 581–599.

Croux, C., Haesbroeck, G., and Joossens, K. (2008b), “Logistic discrimination using robust estimators: an influence function approach,” *Canadian Journal of Statistics*, 36, 157–174.

Dua, D. and Graff, C. (2017), “UCI Machine Learning Repository,” .

Elliott, G. and Lieli, R. P. (2013), “Predicting binary outcomes,” *Journal of Econometrics*, 174, 15–26.

Fernholz, L. T. (1983), *Von Mises calculus for statistical functionals*, New York: Springer-Verlag.

Gneiting, T. (2012), “Making and evaluating point forecasts,” *Journal of the American Statistical Association*, 106, 746–762.

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

Gneiting, T. and Raftery, A. E. (2007), “Strictly proper scoring rules, prediction, and estimation,” *Journal of the American Statistical Association*, 102, 359–378.

Good, I. J. (1953), “The population frequencies of species and the estimation of population parameters,” *Biometrika*, 40, 237–264.

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

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

La Vecchia, D., Ronchetti, E., and Trojani, F. (2012), “Higher-order infinitesimal robustness,” *Journal of the American Statistical Association*, 107, 1546–1557.

Liaw, A. and Wiener, M. (2002), “Classification and Regression by randomForest,” *R News*, 2, 18–22.

Lohweg, V. (2012), “Banknote Authentication Data Set,” Online Data Set, https://archive.ics.uci.edu/ml/datasets/banknote+authentication.

Machete, R. L. (2013), “Contrasting probabilistic scoring rules,” *Journal of Statistical Planning and Inference*, 143, 1781–1790.

Maechler, M., Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M., Conceicao, E. L. T., and Anna di Palma, M. (2019), *robustbase: Basic Robust Statistics*, r package version 0.93-5.

McCullagh, P. and Nelder, J. A. (1989), *Generalized linear models*, London, United Kingdom: Chapman & Hall, 2nd ed.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2019), *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, r package version 1.7-2.

Niculescu-Mizil, A. and Caruana, R. (2005), “Predicting good probabilities with supervised learning,” in *Proceedings of the 22nd international conference on Machine learning*, ACM, pp. 625–632.

Platt, J. (1999), “Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods,” *Journal of Advances in Large Margin Classifiers*, 10, 61–74.

Pregibon, D. (1981), “Logistic regression diagnostics,” *The Annals of Statistics*, 9, 705–724.

R Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Rousseeuw, P. J. and Driessen, K. V. (1999), “A fast algorithm for the minimum covariance determinant estimator,” *Technometrics*, 41, 212–223.

Savage, L. J. (1971), “Elicitation of personal probabilities and expectations,” *Journal of the American Statistical Association*, 66, 783–801.

Scornet, E. (2015), “Learning with random forests,” Ph.D. thesis, Université Pierre et Marie Curie-Paris VI.

Selten, R. (1998), “Axiomatic characterization of the quadratic scoring rule,” *Experimental Economics*, 1, 43–61.

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

von Mises, R. (1947), "On the Asymptotic Distribution of Differentiable Statistical Functions," *The Annals of Mathematical Statistics*, 18, 309–348.

Wolberg, W. H., Street, W. N., and Mangasarian, O. L. (1992), "Breast Cancer Wisconsin (Diagnostic) Data Set," Online Data Set, https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original).

Zadrozny, B. and Elkan, C. (2001), "Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers," in *International Conference on Machine Learning*, vol. 1, pp. 609–616.

— (2002), "Transforming classifier scores into accurate multiclass probability estimates," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM.

Zhou, S. K. (2015), *Medical image recognition, segmentation and parsing: machine learning and multiple object approaches*, Amsterdam: Academic Press.

A Additional Results at The Model

A.1 Additional Simulations for Average Scores

Table A.1: Average of Brier score and logarithmic score for all methods of the additional runs at the model, where Uncnt'n means uncontaminated, Tr. Cont'n means training set outlier contamination and T&T Cont'n training and testing set outlier contamination

	Brier Score			**Logarithmic Score**		
	Uncont'n	*Tr. Cont'n*	*T&T Cont'n*	*Uncont'n*	*Tr. Cont'n*	*T&T Cont'n*
LP2 with 3:1 imbalance, $\epsilon = 0.04$**, moderate outliers**						
Clas. glm	***0.108***	0.122	0.135	***-0.350***	-0.398	-0.434
Miss. glm	0.109	0.122	0.122	-0.352	-0.400	-0.400
Rob. glm	***0.108***	0.109	0.126	***-0.351***	-0.353	-0.438
Rand. For.	0.122	0.125	0.123	-0.556	-0.624	-0.615
RF iso.	0.117	0.122	0.120	-0.512	-0.581	-0.571
RF Platt	0.116	0.119	0.117	-0.385	-0.393	-0.387
SVM	0.295	0.305	0.300	-0.786	-0.809	-0.799
LP2 with 10:1 imbalance, $\epsilon = 0.04$**, moderate outliers**						
Clas. glm	***0.073***	0.086	0.099	***-0.253***	-0.310	-0.344
Miss. glm	0.073	0.086	0.086	-0.254	-0.312	-0.312
Rob. glm	***0.073***	0.074	0.092	***-0.253***	-0.257	-0.352
Rand. For.	0.081	0.084	0.082	-0.443	-0.501	-0.493
RF iso.	0.080	0.083	0.082	-0.419	-0.484	-0.474
RF Platt	0.080	0.082	0.080	-0.294	-0.296	-0.290
SVM	0.286	0.295	0.290	-0.766	-0.788	-0.777
LP2, $\epsilon = 0.08$**, moderate outliers**						
Clas. glm	***0.181***	0.205	0.214	***-0.538***	-0.597	-0.615
Miss. glm	0.182	0.208	0.208	-0.542	-0.605	-0.605
Rob. glm	***0.181***	0.189	0.208	***-0.538***	-0.560	-0.607
Rand. For.	0.200	0.203	0.196	-0.642	-0.641	-0.622
RF iso.	0.192	0.197	0.191	-0.652	-0.654	-0.637
RF Platt	0.189	0.195	0.188	-0.562	-0.574	-0.560
SVM	0.184	0.194	0.198	-0.548	-0.571	-0.581
LP2, $\epsilon = 0.04$, $cov(\mathbf{x}_i) = I_p$						
Clas. glm	***0.198***	0.209	0.217	***-0.580***	-0.605	-0.625
Miss. glm	0.201	0.211	0.211	-0.586	-0.611	-0.611
Rob. glm	***0.198***	0.199	0.212	***-0.580***	-0.582	-0.624
Rand. For.	0.216	0.220	0.217	-0.643	-0.652	-0.645
RF iso.	0.211	0.214	0.212	-0.650	-0.660	-0.655
RF Platt	0.208	0.211	0.209	-0.604	-0.611	-0.606
SVM	0.201	0.205	0.209	-0.587	-0.596	-0.606

In Table A.1 extra simulations for completeness purposes are given. This includes runs where the effect of imbalance is tested. As well as run with no correlated explantory variables, a higher contamination level equal to $\epsilon = 0.04$.

A.2 Asymptotic Relative Efficiencies

Table A.2: Asymptotic relative efficiency of Brier scores of LP2 at the model

	Classic	Missp.	Robust	Rand. Forest	RF iso.	RF Platt	SVM
Classic glm	1	1.008	0.995	1.801	1.602	1.464	1.146
Missp. glm		1	0.987	1.786	1.590	1.452	1.137
Robust glm			1	1.809	1.610	1.471	1.152
Rand. Forest				1	0.890	0.813	0.637
RF iso.					1	0.913	0.715
RF Platt						1	0.783
SVM							1

Table A.3: Relative efficiencies of Brier score of LP1 with a contamination level $\epsilon = 0.04$ of moderate outliers in the training set

	Classic	Missp.	Robust	Rand. Forest	RF iso.	RF Platt.	SVM
Classic glm	1	0.970	0.908	2.89	1.66	1.25	1.09
Missp. glm		1	0.933	2.97	1.70	1.28	1.12
Robust glm			1	2.89	1.65	1.24	1.09
Rand. Forest				1	0.588	0.442	0.387
RF iso. cal.					1	0.850	0.744
RF Platt cal.						1	0.947
SVM Platt							1

Table A.4: Asymptotic relative efficiency of Brier scores of LP2 with a contamination level $\epsilon = 0.04$ of moderate outliers in the training set

	Classic	Missp.	Robust	Rand. Forest	RF iso.	RF Platt	SVM
Classic glm	1	1.382	0.941	1.972	1.888	1.607	1.160
Missp. glm		1	0.933	1.956	1.873	1.594	1.151
Robust glm			1	1.981	1.897	1.615	1.166
Rand. Forest				1	1.049	0.893	0.644
RF iso.					1	1.003	0.724
RF Platt						1	0.793
SVM							1

A.3 Reliability Curves

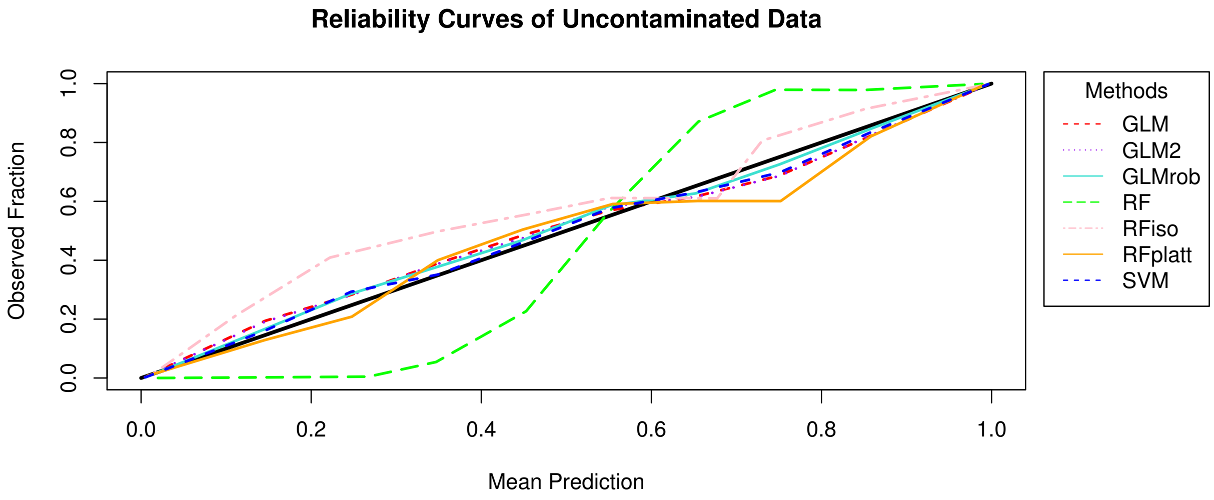

Figure A.1: Reliability curves of the predictions of LP1 at the model, where the "Mean Prediction" label is equal to the Quantile

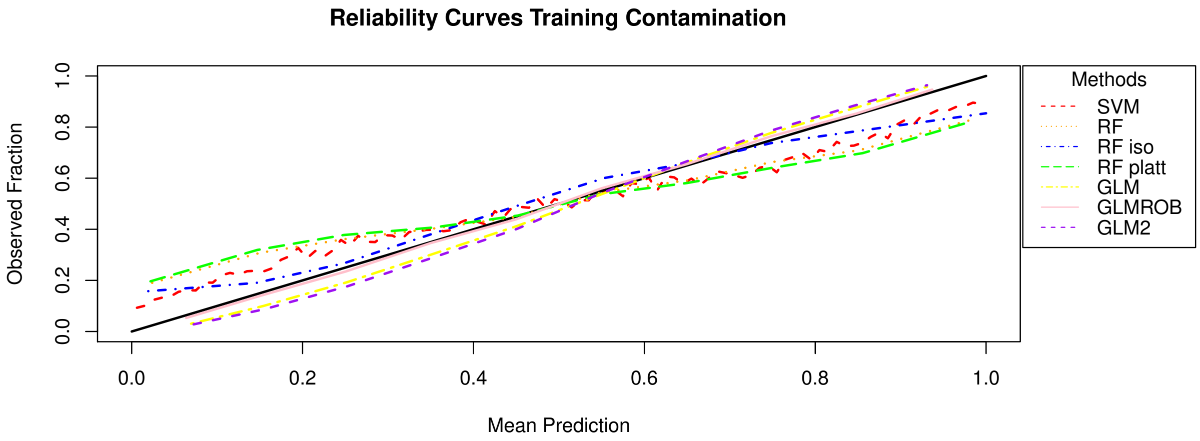

Figure A.2: Reliability curves of the predictions of LP1 with $\epsilon = 0.02$ level of moderate outliers in the training set, where the "Mean Prediction" label is equal to the Quantile

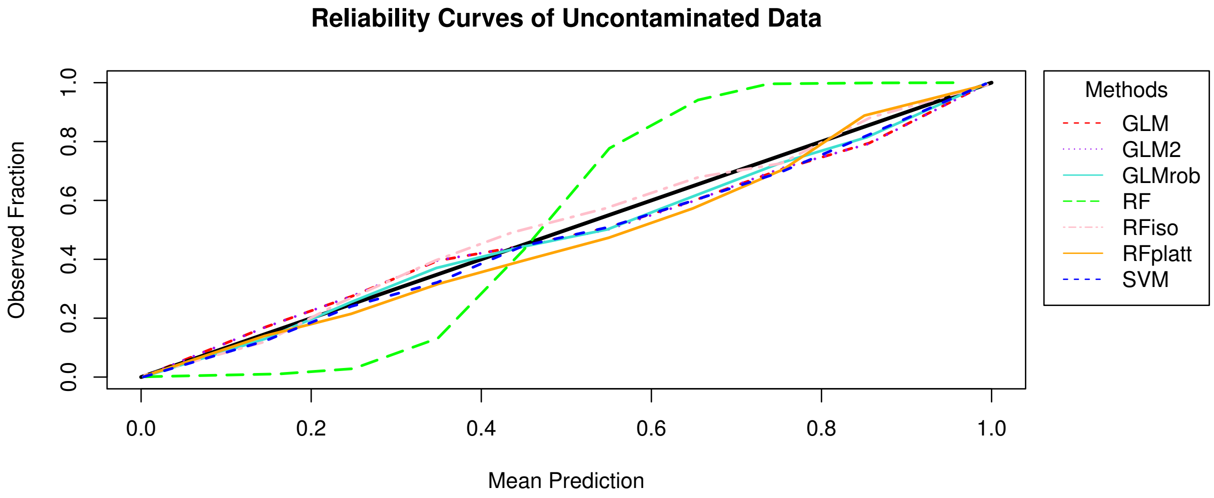

Figure A.3: Reliability curves of the predictions of LP2 at the model, where the ”Mean Prediction” label is equal to the Quantile

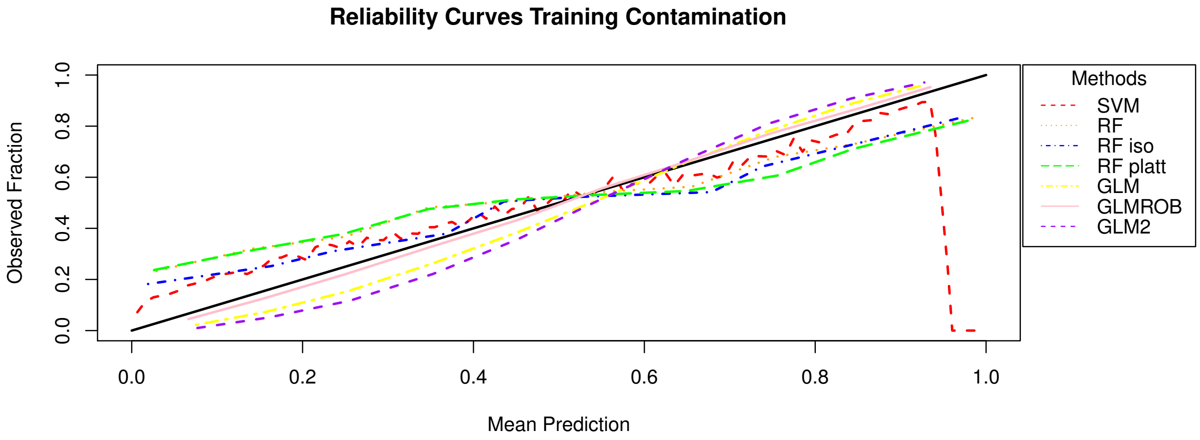

Figure A.4: Reliability curves of the predictions of LP1 with $\epsilon = 0.04$ level of moderate outliers in the training set, where the ”Mean Prediction” label is equal to the Quantile

B Additional Results of Heavy-tailed Error Contamination

B.1 Average Brier scores and Logarithmic Scores

The first important conclusion that can be drawn from Table B.1 is that when the model is contaminated by heavy-tailed errors, that the average an variance of the scores obtained from predictions of classic and robust logistic regression are still similar. One would expect this to not be the case as heavy-tailed disturbances add more mislabeling errors to the model and in general robust logistic regression is less sensitive to outliers.

Table B.1: Averages and variances of Brier score and logarithmic score for classic and robust logistic regression of LP1 with contamination by heavy-tailed errors and two levels of outlier contamination $\epsilon \in \{0.02, 0.04\}$ (variances are the values in between brackets), where Uncnt'n means only contamination by heavy-tailed errors, Tr. Cont'n means training set outlier contamination as well as heavy-tailed error contamination and T&T Cont'n training and testing set contamination with heavy-tailed error contamination

	Uncontaminated		**Training Cont'n**		**T&T Cont'n**	
	classic	*robust*	*classic*	*robust*	*classic*	*robust*
LP1, $t(5)$, $\zeta = 0.05$, $\epsilon = 0.04$**, moderate outliers**						
BS	0.178	0.177	0.186	0.178	0.195	0.191
	(2.3E-9)	(1.2E-9)	(4.7E-6)	(2.1E-7)	(7.5E-6)	(4.7E-6)
LS	-0.529	-0.528	-0.554	-0.531	-0.576	-0.571
	(1.5E-8)	(7.4E-9)	(3.0E-5)	(2.2E-6)	(4.0E-5)	(3.2E-5)
LP1, $t(5)$, $\zeta = 0.05$, $\epsilon = 0.02$**, moderate outliers**						
BS	0.178	0.177	0.180	0.177	0.186	0.184
	(2.3E-9)	(1.2E-9)	(9.8E-7)	(1.7E-8)	(3.9E-6)	(2.4E-6)
LS	-0.529	-0.528	-0.538	-0.528	-0.554	-0.551
	(1.5E-8)	(7.4E-9)	(8.7E-6)	(1.9E-7)	(2.6E-5)	(2.3E-5)

In Table B.2 we display the rest of the simulation runs with the more critical contamination level equal to $\epsilon = 0.04$. Again the most important detail between classic and robust logistic regression is that at the model with heavy-tailed error contamination robust logistic regression obtains a score similar to that of classic logistic regression. Unlike the correctly specified simulations, the efficiency of the two for error mixture simulations is not equal. In most runs here the scores of robust logistic regression predictions are more efficient that classic logistic regression scores, even when no outlier contamination is added.

Furthermore, when the tails of the t-distributed mixed error are fatter, there are more observations that have errors with tail values. For classic logistic regression this influences the predictions slightly for all contamination cases. On the other hand, robust logistic regression is only influenced when there is no contamination, and the effect of even fatter tails diminishes when there are outliers in the data set.

Table B.2: Average of Brier score and logarithmic score for all methods of the runs with contamination by heavy-tailed errors, where Uncnt'n means only contamination by heavy-tailed errors, Tr. Cont'n means training set outlier contamination as well as heavy-tailed error contamination and T&T Cont'n training and testing set contamination with heavy-tailed error contamination

	Brier Score			**Logarithmic Score**		
	Uncont'n	*Tr. Cont'n*	*T&T Cont'n*	*Uncont'n*	*Tr. Cont'n*	*T&T Cont'n*
LP2, $5t(2)$, $\zeta = 0.1$, $tcc = 1.0$ $\epsilon = 0.04$, **moderate outliers**						
Clas. glm	***0.108***	0.136	0.150	***-0.355***	-0.437	-0.475
Miss. glm	0.108	0.135	0.135	-0.355	-0.436	-0.436
Rob. glm	***0.107***	0.109	0.127	***-0.352***	-0.359	-0.450
Rand. For.	0.114	0.117	0.117	-0.379	-0.387	-0.388
RF iso.	0.118	0.119	0.119	-0.423	-0.427	-0.428
RF Platt	0.115	0.118	0.117	-0.381	-0.387	-0.387
SVM	0.109	0.118	0.128	-0.360	-0.389	-0.413
LP2, $t(10)$, $\zeta = 0.05$, $\epsilon = 0.04$, **moderate outliers**						
Clas. glm	***0.166***	0.177	0.187	***-0.501***	-0.531	-0.557
Miss. glm	0.168	0.179	0.179	-0.506	-0.537	-0.537
Rob. glm	***0.166***	0.167	0.182	***-0.502***	-0.504	-0.557
Rand. For.	0.182	0.185	0.183	-0.593	-0.603	-0.599
RF iso.	0.177	0.180	0.180	-0.621	-0.627	-0.624
RF Platt	0.175	0.178	0.177	-0.528	-0.536	-0.533
SVM	0.169	0.172	0.180	-0.511	-0.521	-0.538
LP2, t(5), $\zeta = 0.05$, $\epsilon = 0.04$, **moderate outliers**						
Class. glm	***0.180***	0.189	0.198	***-0.535***	-0.559	-0.581
Miss. glm	0.181	0.190	0.190	-0.539	-0.564	-0.564
Rob. glm	***0.180***	0.181	0.194	***-0.535***	-0.538	-0.580
Rand. For.	0.199	0.201	0.199	-0.640	-0.645	-0.640
RF iso.	0.191	0.194	0.193	-0.648	-0.651	-0.647
RF Platt	0.188	0.191	0.189	-0.559	-0.566	-0.563
SVM	0.182	0.186	0.192	-0.544	-0.553	-0.567
LP2, $t(5)$, $\zeta = 0.05$, $\epsilon = 0.02$, **moderate outliers**						
Clas. glm	***0.180***	0.182	0.188	***-0.535***	-0.543	-0.559
Miss. glm	0.181	0.184	0.184	-0.539	-0.546	-0.546
Rob. glm	***0.180***	0.180	0.187	***-0.535***	-0.536	-0.559
Rand. For.	0.199	0.200	0.200	-0.640	-0.643	-0.644
RF iso.	0.191	0.192	0.193	-0.648	-0.650	-0.651
RF Platt	0.188	0.189	0.190	-0.559	-0.562	-0.564
SVM	0.182	0.183	0.188	-0.544	-0.548	-0.557

Conclusively, in all these runs we found that for an heavy-tailed error level of $\zeta = 0.05$ using $t(5)$, the probability that the dependent variable becomes mislabeled is lower, at 0.015 (as well as multiple other runs found in Table B.2). To increase this percentage of mislabeling to get a significant difference between scores of robust logistic regression predictions and scores of classic logistic regression predictions, we increase the scale of the t-distribution by multiplying it by a factor as well as increasing the contamination level to $\zeta = 0.1$. Furthermore, to distinguish the outliers from the true values, robust logistic regression is calculated using the weights found by `covMcd` (Rousseeuw and Driessen, 1999).

B.2 Asymptotic Relative Efficiencies

Table B.3: Asymptotic relative efficiencies of Brier score for LP2 with $\zeta = 0.1$ heavy-tailed error contamination level of $5t(2)$ and outlier contamination in the training set of moderate outliers with a level of $\epsilon = 0.04$

	Classic	Missp.	Robust	Rand. Forest	RF iso.	RF Platt.	SVM
Classic glm	1	2.34	1.19	2.41	2.31	2.05	1.98
Missp. glm		1	1.21	2.47	2.37	2.10	2.02
Robust glm			1	2.22	2.13	1.88	1.82
Rand. Forest				1	1.09	0.967	0.932
RF iso. cal.					1	0.978	0.942
RF Platt cal.						1	1.15
SVM Platt							1

Table B.4: Asymptotic relative efficiencies of Brier score for LP2 with $\zeta = 0.1$ heavy-tailed error contamination level of $5t(2)$

	Classic	Missp.	Robust	Rand. Forest	RF iso.	RF Platt.	SVM
Classic glm	1	1.00	1.02	1.65	1.51	1.25	1.20
Missp. glm		1	1.03	1.66	1.51	1.25	1.20
Robust glm			1	1.61	1.47	1.22	1.17
Rand. Forest				1	0.913	0.755	0.725
RF iso. cal.					1	0.828	0.794
RF Platt cal.						1	0.959
SVM Platt							1

Table B.5: Asymptotic relative efficiencies of Brier score for LP2 with $\zeta = 0.05$ heavy-tailed error contamination level of $5t(2)$ and outlier contamination in the training set of moderate outliers with a level of $\epsilon = 0.04$

	Classic	Missp.	Robust	RF	RF iso.	RF Platt.	SVM
Classic glm	1	1.29	0.969	1.77	1.61	1.48	1.21
Missp. glm		1	0.973	1.78	1.62	1.48	1.22
Robust glm			1	1.73	1.57	1.44	1.19
Rand. Forest				1	0.976	0.894	0.736
RF iso. cal.					1	0.980	0.806
RF Platt cal.						1	0.974
SVM Platt							1

Table B.6: Asymptotic relative efficiencies of Brier score for LP1 with $\zeta = 0.1$ heavy-tailed error contamination level of $5t(2)$

	Classic	Missp.	Robust	Rand. Forest	RF iso.	RF Platt.	SVM
Classic glm	1	0.979	1.03	3.38	1.81	1.21	1.13
Missp. glm		1	1.06	3.46	1.85	1.23	1.15
Robust glm			1	3.27	1.75	1.17	1.09
Rand. Forest				1	0.535	0.356	0.334
RF iso. cal.					1	0.666	0.624
RF Platt cal.						1	0.937
SVM Platt							1

Table B.7: Asymptotic relative efficiencies of Brier score for LP1 with $\zeta = 0.1$ heavy-tailed error contamination level of $5t(2)$ and outlier contamination in the training set of moderate outliers with a level of $\epsilon = 0.04$

	Classic	Missp.	Robust	Rand. Forest	RF iso.	RF Platt.	SVM
Classic glm	1	0.941	0.928	3.60	1.98	1.33	1.04
Missp. glm		1	0.948	3.68	2.02	1.35	1.07
Robust glm			1	3.48	1.91	1.28	1.01
Rand. Forest				1	0.584	0.392	0.308
RF iso. cal.					1	0.731	0.575
RF Platt cal.						1	0.864
SVM Platt							1

B.3 Reliability Curves

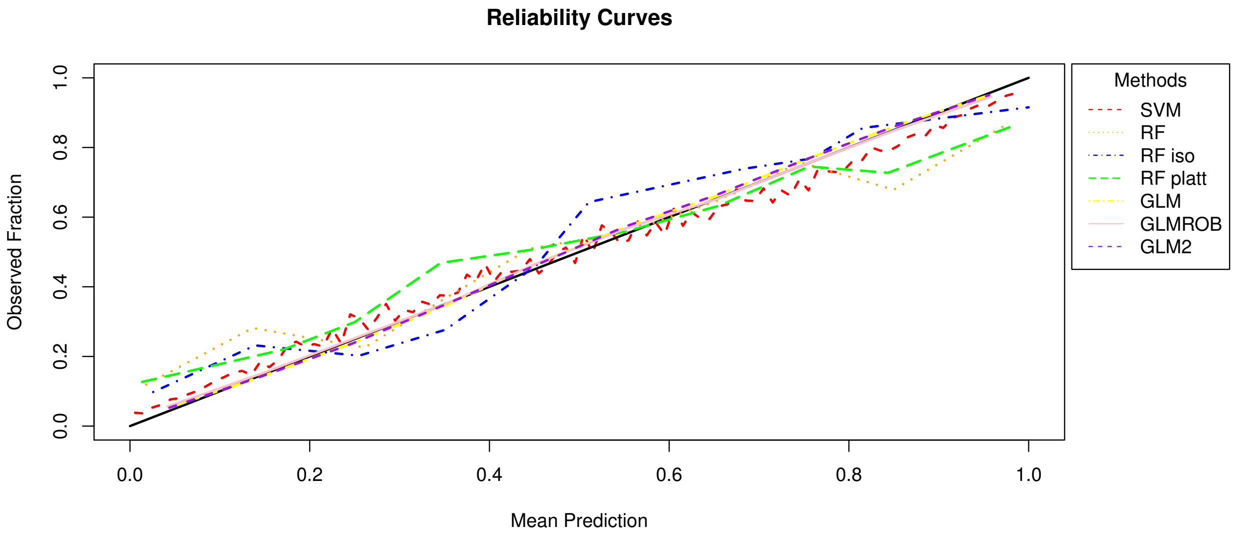

Figure B.1: Reliability curves for LP2 with $\zeta = 0.1$ heavy-tailed error contamination level of $5t(2)$, where the "mean prediction" label is equal to the Quantile

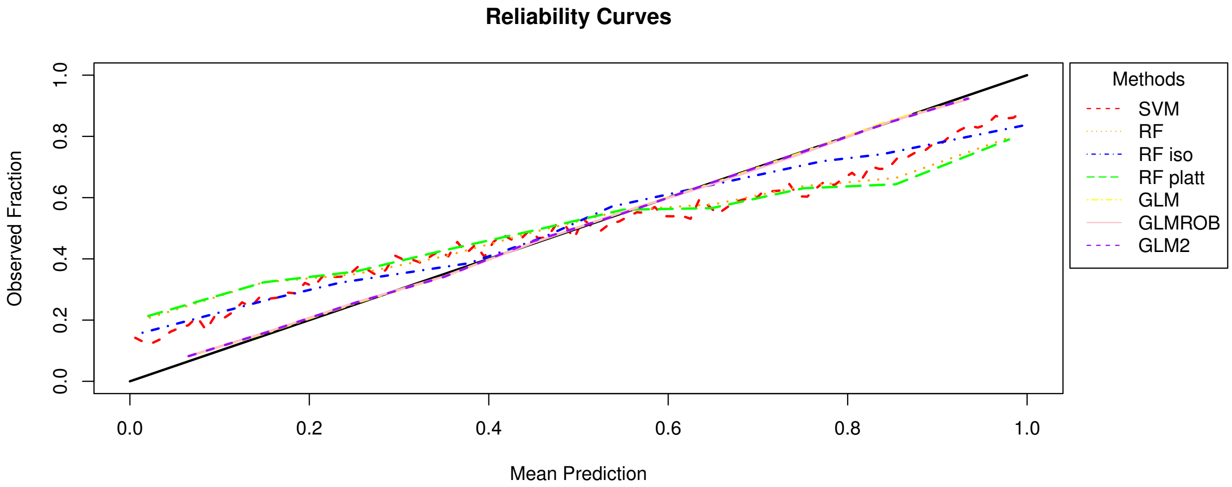

Figure B.2: Reliability curves for LP1 with $\zeta = 0.1$ heavy-tailed error contamination level of $5t(2)$, where the "mean prediction" label is equal to the Quantile

C Additional Results of Misspecified Link Function

C.1 Asymptotic Relative Efficiencies

Again the most asymptotically efficient Brier score are obtained from the predictions of robust logistic regression, meaning that although it does not have the best average Brier score, we are more sure of its value (as the variance is smallest out of all methods). This is also seen in Table C.1, as the row of ”Robust glm” has only values greater than 1 and the column of ”Robust” has all values smaller than 1.

Table C.1: Relative efficiencies of Brier score for 4% training set contaminated data set of LP1 with misspecified link function

	Classic	Missp.	Robust	Rand. Forest	RF iso.	RF Platt.	SVM
Classic glm	1	0.848	0.284	0.391	0.515	0.487	0.462
Missp. glm		1	0.336	0.462	0.608	0.575	0.545
Robust glm			1	1.38	1.81	1.71	1.62
Rand. Forest				1	1.32	1.24	1.18
RF iso. cal.					1	0.946	0.896
RF Platt cal.						1	0.948
SVM Platt							1

C.2 Reliability Curves

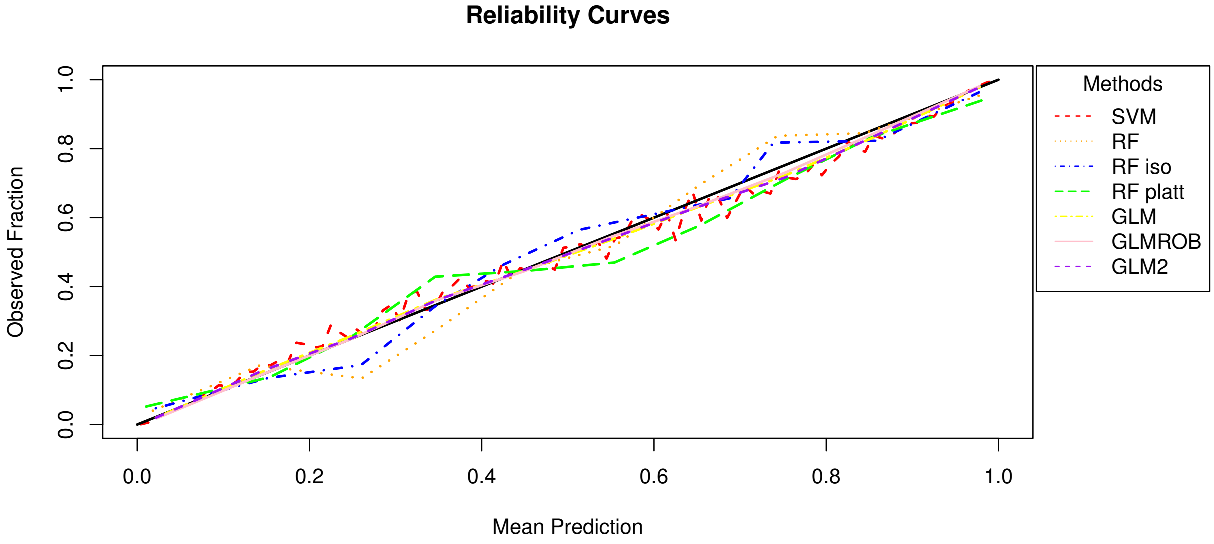

Figure C.1: Reliability curves of LP1 at the model with a misspecfied link function, where the ”Mean Prediction” label is equal to the Quantile

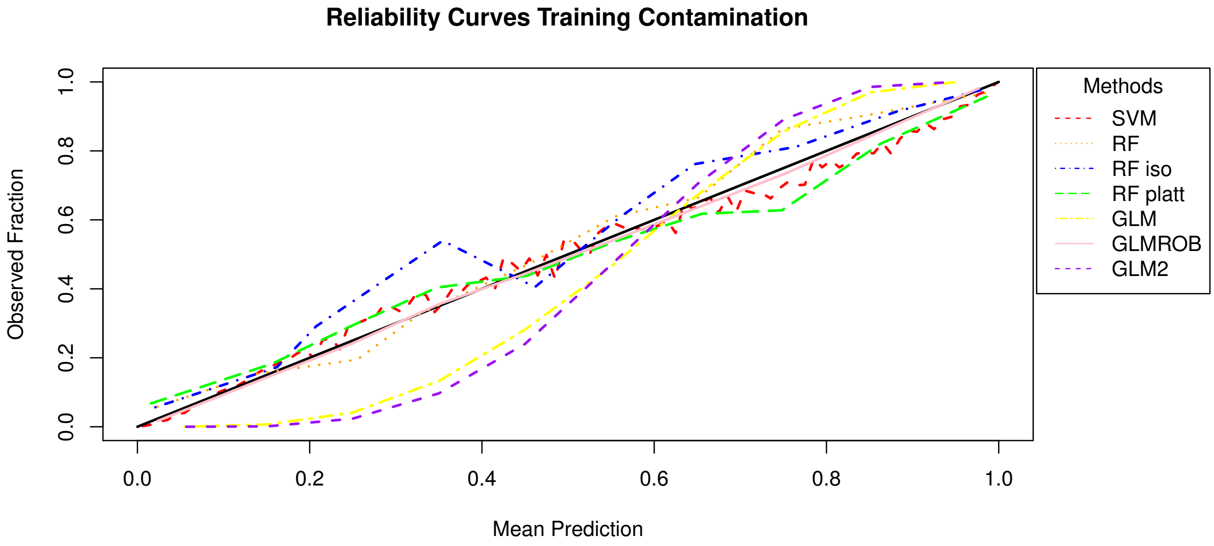

Figure C.2: Reliability curves of LP1 with training set contamination level $\epsilon = 0.04$ of moderate outliers with a misspecfied link function, where the ”Mean Prediction” label is equal to the Quantile

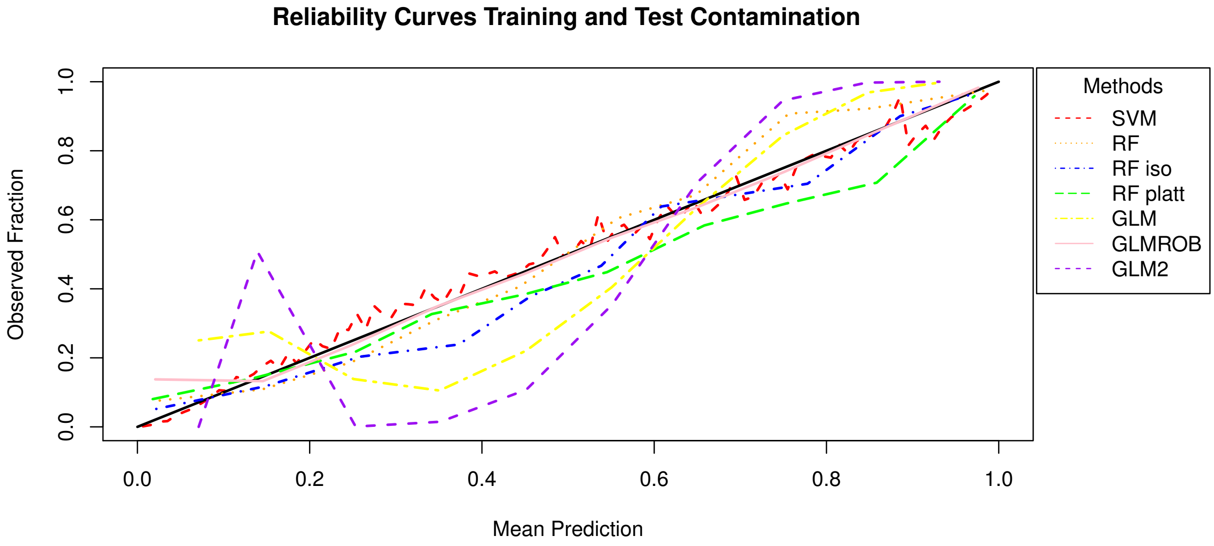

Figure C.3: Reliability curves of LP1 with training and test set contamination level $\epsilon = 0.04$ of moderate outliers with a misspecfied link function, where the ”Mean Prediction” label is equal to the Quantile