České vysoké učení technické v Praze
Fakulta jaderná a fyzikálně inženýrská

Czech Technical University in Prague
Faculty of Nuclear Sciences and Physical Engineering

Ing. Tomáš Hobza, Ph.D.

Robustní odhady parametrů zobecněných lineárních modelů

Robust estimators in generalized linear models

Summary

Generalized linear models and particularly logistic regression have become one of the most used statistical procedures employed by statisticians and researchers for the analysis of binary, proportional or count response variables in various fields, including epidemiologic research, machine learning, biomedical research, social science, computer science, electronics and electrical engineering, etc. In most of the areas the data sets to be analyzed usually contain atypical observations, so called outliers. Thus, one of the most important issues is the estimation of parameters of the models and testing hypotheses about these parameters in the presence of outliers. The most widely used method for parameter estimation in the frame of generalized linear models is the maximum likelihood estimator which is well known to be extremely sensitive to 'contaminated' data. For this reason some interesting robust M-estimators have been introduced in the statistical literature to overcome the above mentioned problem.

In this habilitation lecture we introduce and study a new robust estimator, called modified median estimator, as well as a Wald-type test based on it. Their theoretic asymptotic properties are also discussed. The efficiency and robustness of the modified median estimator as well as the corresponding test is studied on the basis of a simulation experiment. The results point out their good behavior in some concrete situations and in comparison with already existing methods.

Souhrn

Zobecněné lineární modely a speciálně logistická regrese se staly jedněmi z nejvíce používaných metod aplikovaných statistiky a vědeckými pracovníky na analýzu binárních, proporčních nebo četnostních vysvětlovaných proměnných v mnoha oborech, jako jsou např. epidemiologický výzkum, strojové učení, biomedicínský výzkum, sociální vědy, počítaçové vědy, elektronika a elektrické inženýrství apod. Ve většině těchto oblastí však sady dat, které mají být analyzovány, obvykle obsahují atypická pozorování, tzv. outliery. Jedním z nejůležitějších problémů je tedy odhadování parametrů modelů a testování hypotéz o těchto parametrech za přítomnosti odlehlých pozorování. V oblasti zobecněných lineárních modelů je nejvíce aplikovanou metodou pro odhad parametrů modelů metoda maximální věrohodnosti, o které je známo, že je velmi citlivá na 'zašuměná' data. Z tohoto důvodu bylo ve statistické literatuře navrženo několik zajímavých robustních odhadů, s cílem odstranit zmíněný problém.

V této habilitační přednášce zavádíme a studujeme nový robustní odhad, zvaný modifikovaný mediánový odhad, stejně jako na něm založenou testovací statistiku Waldova typu. Jsou také diskutovány jejich teoretické asymptotické vlastnosti. Eficience a robustnost modifikovaného mediánového odhadu a příslušného testu je studována pomocí simulačních experimentů. Výsledky ukazují na jejich dobré vlastnosti v některých konkrétních situacích a ve srovnání s již existujícími metodami.

Klíčová slova: zobecněné lineární modely, logistická regrese, robustnost, eficience, mediánový odhad, robustní testy hypotéz, testovací statistika Waldova typu

Keywords: generalized linear models, logistic regression, maximum likelihood estimators, robustness, efficiency, median estimator, robust hypothesis testing, Wald-type test statistic

Contents

1 Introduction

Generalized linear models (GLMs) are nowadays widely used for analysis of data in many various fields. They are in fact generalization of the model of classical linear regression which allows to assume non-normally distributed response variables, heteroscedasticity and non-linear relationship between the expectation of the response variable and the explanatory variables. GLMs were first introduced in [16] and received a lot of attention in the recent past. The class of GLMs provides an unifying framework and contains as special cases models such as linear regression, ANOVA, logistic regression, Poisson regression, log-linear models, and many others. In this lecture we focus our attention on logistic regression which is one of the most used GLMs for binary or proportional response variables. Nevertheless, the presented ideas and methods can be generalized also to other GLMs for discrete responses.

Estimation and testing procedures in logistic regression are usually based on maximum likelihood estimators (MLEs) and inherit the sensitivity of these estimators in the presence of atypical observations. A small amount of these kind of data can seriously affect the level or the power of the corresponding tests. Therefore it is important to consider robust estimators in order to be able to get robust tests, i.e., to get testing procedures in such a way that in the presence of atypical observations the level as well as the power function will be stable. It is important to note that the problem of getting robust estimators has been more developed that the problem of considering robust tests for the logistic regression model. The papers [12], [6], [2] and [1] study the problem associated with finding test statistics with stable level and power under atypical observations. In [12] robust tests for a general parametric model including logistic regression is introduced. In [6] the authors define robust deviances based on generalizations of quasi–likelihood functions and propose a family of test statistics for model selection in generalized linear models. They also investigate the stability of the asymptotic level under contamination. A Wald-type test statistic based on a weighted Bianco and Yohai estimator is proposed and studied in [2]. In [1] the problem is considered under the assumption of random covariates and family of robust Wald type tests is introduced, where the minimum density power divergence estimator is used instead of the maximum likelihood estimator. It is theoretically established that the level as well as the power of the Wald-type tests are stable against

contamination, while the classical Wald test breaks down in this scenario.

Our contribution to this topic consists in introducing a new estimator using the median function, which is known to be robust, and defining a version of a Wald-type test based on the proposed estimator. In order to describe the estimator and the corresponding test let us start with the definition of the problem of estimation in logistic regression.

We are interested in estimation of the parameter $\boldsymbol{\beta} \in \mathbb{R}^d$ in the binary regression based on independent observations $Y_1, ..., Y_n$ with Bernoulli distribution,

$$Y_i \sim Be\left(\pi_i\right), \qquad i = 1, \ldots, n,$$

where the Bernoulli parameters $\pi_i = P(Y_i = 1)$ depend on $\boldsymbol{\beta}$ and vectors of explanatory variables (regressors) $\boldsymbol{x}_i \in \mathbb{R}^d$,

$$\pi_i = \pi_i\left(\boldsymbol{\beta}\right) = \pi\left(\boldsymbol{x}_i^T \boldsymbol{\beta}\right) .$$

Here and in the sequel, $\boldsymbol{x}^T \boldsymbol{\beta}$ denotes the scalar product of vectors $\boldsymbol{x}$ and $\boldsymbol{\beta}$ and $\pi(t) : \mathbb{R} \mapsto (0, 1)$ is strictly monotone and infinitely differentiable.

If we use the logistic function

$$\pi(t) = \frac{e^t}{1 + e^t}, \qquad t \in \mathbb{R},$$

the problem reduces to the *classical logistic regression* with binary observations $Y_i \sim Be\left(\pi_i\right)$, $i = 1, \ldots, n$, and

$$\operatorname{logit}(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \boldsymbol{x}_i^T \boldsymbol{\beta} .$$

In the classical logistic regression the MLE $\widehat{\boldsymbol{\beta}}_n = \widehat{\boldsymbol{\beta}}_n\left(Y_1, ..., Y_n\right)$ of $\boldsymbol{\beta}$ minimizes the sum of deviances (negative scores)

$$\mathcal{D}_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} d_i\left(\boldsymbol{\beta}\right)$$

of the sample $(Y_1, ..., Y_n)$, where

$$d_i\left(\boldsymbol{\beta}\right) = -Y_i \ln \pi_i\left(\boldsymbol{\beta}\right) - \left(1 - Y_i\right) \ln\left(1 - \pi_i\left(\boldsymbol{\beta}\right)\right) \tag{1}$$

are the deviances of individual observations Y_i. Thus

$$\widehat{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\beta}} \mathcal{D}_n(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} d_i\left(\boldsymbol{\beta}\right) .$$

Consistency and asymptotic normality with the variances at the Cramér-Rao lower bound can be proved for this estimator.

However, it is known that maximum likelihood estimates of the parameters $\boldsymbol{\beta}$ are sensitive to contamination of the data $(\boldsymbol{x}_1, Y_1), ..., (\boldsymbol{x}_n, Y_n)$ by outliers or leverage points. Typical outliers are

$$Y_i = 0 \text{ when } \pi\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right) \approx 1 \quad \text{or} \quad Y_i = 1 \text{ when } \pi\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right) \approx 0.$$

Such outlying values may lead to large deviances $d_i(\boldsymbol{\beta})$ (cf. (1)), thus pushing the MLE's $\widehat{\boldsymbol{\beta}}_n$ far away from the true value $\boldsymbol{\beta}$.

In order to restrict the undesired influence of large deviances resulting from contamination of data, previous authors replaced the deviances $d_i(\boldsymbol{\beta})$ by appropriate functions $\varrho(d_i(\boldsymbol{\beta}))$ of deviances, or even by more general expressions $\phi\left(Y_i, \pi\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)\right)$. This lead to M-estimators $\boldsymbol{\beta}_n$ of the type

$$\boldsymbol{\beta}_n = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \varrho(d_i(\boldsymbol{\beta})) \tag{2}$$

or

$$\boldsymbol{\beta}_n = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \phi\left(Y_i, \pi\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)\right) \tag{3}$$

for $\varrho : (0,\infty) \to \mathbb{R}$ and $\phi : (0,\infty)\times(0,1) \to \mathbb{R}$. Some estimators of this form were studied in [17], [3] and [9]. Other interesting robust estimators of the parameters in the logistic regression model have been presented for instance in [7], [8], [4] and [5].

2 Median based estimators

In this section we continue the line sketched above and we present another two estimators of the type (3). It is known (cf. e. g. [11], [18], [15], [19]) that the median estimator of parameters of linear and non-linear regression is robust with respect to contamination of observations. The idea is to generalize this concept to our model with the hope that the median estimator for the logistic regression will be robust too.

2.1 Median estimator

In [14] we propose a new robust M-estimator of the logistic regression parameter $\boldsymbol{\beta} \in \mathbb{R}^d$ which is based on median function and which is expected to

be more robust than the MLE. It can be defined by the formula

$$\widehat{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left| Y_i - m\left(\pi\left(\boldsymbol{x}_i^T \boldsymbol{\beta}\right)\right)\right| ,$$

where $m(\pi)$ is for every $\pi \in (0,1)$ the median function

$$m(\pi) = F_\pi^{-1}(1/2) = \inf\{y \in \mathbb{R} : F_\pi(y) \geq 1/2\}.$$

The condition of applicability of this estimator is sensitivity of the median function $m(\pi)$ to the change of parameter $\pi \in (0,1)$ (strict monotonicity of $m(\pi)$ on $(0,1)$). Unfortunately, in the discrete Bernoulli model the median function has the form

$$m(\pi) = I(\pi > 1/2) = \begin{cases} 0 & \text{if } \pi \leq 1/2 \\ 1 & \text{if } \pi > 1/2 \end{cases}$$

and cannot be used directly, since it is piecewise constant and it is thus not sensitive to small changes of the parameter π. This conclusion remains valid also for any other generalized linear model with discrete responses. To overcome this problem, the main and basic idea of the mentioned paper is to assume a transformation, called *statistical smoothing*, of the discrete observations $Y_1, \ldots, Y_n$. This transformation consists in adding independent and uniformly on $(0,1)$ distributed random variables U_i to the observations Y_i, i.e. it considers the continuous data

$$Z_i = Y_i + U_i, \qquad i = 1, \ldots, n,$$

where $U_i \overset{iid}{\sim} U(0,1)$. Let us note that the introduced transformation is statistically sufficient since the original Y_i may be recovered completely by applying the integer-part operation to Z_i,

$$Y_i = [Z_i] \text{ a.s.}, \qquad i = 1, \ldots, n.$$

Statistical smoothing goes in fact in the opposite direction to the statistical quantization frequently applied to continuous data. The quantization is usually accompanied by the loss of information so that it is not statistically sufficient.

Using the smoothed continuous observations $Z_1, \ldots, Z_n$, the *median estimator* $\widehat{\boldsymbol{\beta}}^{\text{Med}}$ is then defined as

$$\widehat{\boldsymbol{\beta}}^{\text{Med}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left| Z_i - m\left(\pi(\boldsymbol{x}_i^T \boldsymbol{\beta})\right) \right| , \tag{4}$$

where

$$\pi(\boldsymbol{x}_i^T \boldsymbol{\beta}) = \frac{\exp(x_i^T \boldsymbol{\beta})}{1 + \exp(x_i^T \boldsymbol{\beta})}$$

and $m(p)$ is the median function

$$m\left(p\right) = F_p^{-1}\left(1/2\right) = \inf\left\{z \in \mathbb{R} : F_p\left(z\right) \geq 1/2\right\} \tag{5}$$

corresponding to the class of distribution functions F_p of the random variables

$$Z = \text{Be}\left(p\right) + \text{U}(0,1)$$

when the parameter p varies in the closed interval $[0,1]$. It can be shown that the median function (5) has the explicit form

$$m(p) = 1 + \frac{p - 1/2}{\max\{p, 1-p\}}, \quad 0 \leq p \leq 1, \tag{6}$$

and is strictly increasing in p (cf. Figure 1). Since the logistic function is strictly increasing too, the argument $m\left(\pi(\boldsymbol{x}^T \boldsymbol{\beta})\right)$ in (4) detects every change of the product $\boldsymbol{x}^T \boldsymbol{\beta}$.

Consistency and asymptotic normality of the median estimator defined in (4) are proved in the paper [14]. Simulation studies are carried out to study the sensitivity of the median estimators to outlying and leverage points and to compare it with the sensitivity of some robust estimators previously introduced in the literature. The median estimators seem to be more robust for larger sample sizes and higher levels of contamination. Unfortunately, the increased robustness of median estimator is usually accompanied by a loss of efficiency. In the next section we will describe a method for suppressing the inefficiency.

2.2 Modified median estimator

In this section we are going to propose a modification of the median estimator, defined in (4), with the aim to improve its behavior. In fact, we have made the

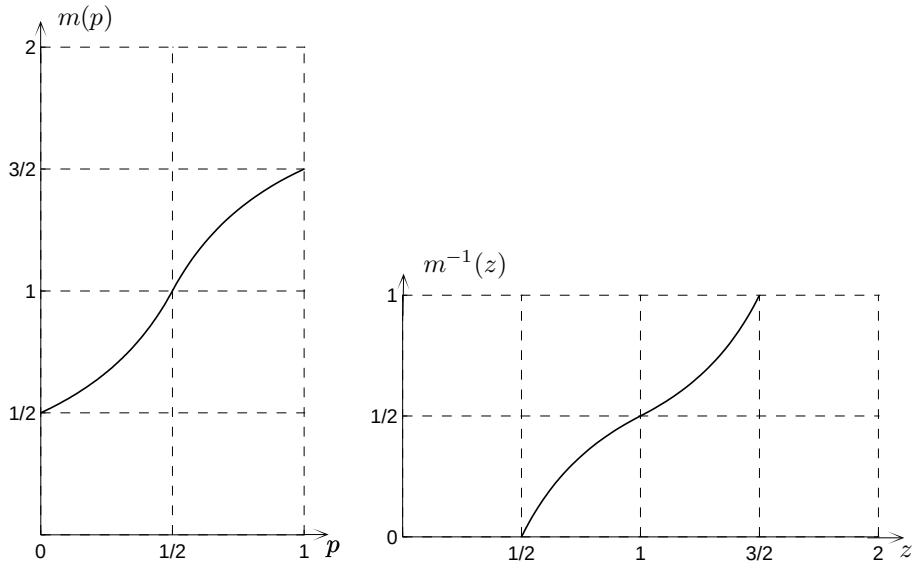

Figure 1. Median function $m(p)$ and its inverse $m^{-1}(z)$.

first attempt already in [14], where the method of enhancing of the median estimator was introduced. This method increases efficiency of the median estimator in some cases and it consists in replacing the set of statistically smoothed data $Z_i = Y_i + U_i$, $1 \leq i \leq n$, by the expanded set obtained by considering for $k > 1$ the matrix of data

$$Z_{ij} = Y_i + U_{ij}, \qquad 1 \leq i \leq n, \qquad 1 \leq j \leq k, \tag{7}$$

where U_{ij} are $U(0,1)$-distributed and mutually as well as on $Y_1, \ldots, Y_n$ independent random variables, and applying the median estimator to this expanded set. In other words the *k-enhanced median estimator* can be defined by

$$\widehat{\boldsymbol{\beta}}^{k\mathrm{Med}} = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^{d+1}} \frac{1}{k} \sum_{i=1}^{n} \sum_{j=1}^{k} \left| Y_i + U_{ij} - m\left(\pi(\boldsymbol{x}_i^T \boldsymbol{\beta})\right) \right| . \tag{8}$$

It seems that the idea can still be improved upon. If we let $k \to \infty$ in (8), we get the formula

$$\widehat{\boldsymbol{\beta}}^{\mathrm{MMe}} = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^{d+1}} \sum_{i=1}^{n} \int_0^1 \left| Y_i + u - m\left(\pi(\boldsymbol{x}_i^T \boldsymbol{\beta})\right) \right| du \tag{9}$$

defining a deterministic estimate (i.e. it does not depend on any additionally generated random sample used for statistical smoothing), which would conceivably inherit the good properties of the original estimate plus a smaller

variance. The estimator defined by the formula (9) for the median function (6) will be called *modified median estimator*, or briefly MMe-estimator, in the sequel. It is obvious that the MMe-estimator is an M-estimator, since

$$\widehat{\boldsymbol{\beta}} = \arg\min \sum_{i=1}^{n} \phi\left(Y_i, \pi\left(\boldsymbol{x}_i^T \boldsymbol{\beta}\right)\right)$$

with

$$\phi\left(Y_i, \pi\left(\boldsymbol{x}_i^T \boldsymbol{\beta}\right)\right) = \int_0^1 \left|Y_i + u - m\left(\pi\left(\boldsymbol{x}_i^T \boldsymbol{\beta}\right)\right)\right| du\,.$$

So general asymptotic theory for M-estimators can be applied and asymptotic normality of the modified median estimator can be established. For more details see [13]. In the next section we carry out a simulation study in order to asses the behavior of the proposed modified median estimator.

2.3 Simulation experiment - robustness

Properties of the median estimator and some other well known estimators, tailor-made for robust estimation in logistic regression, were compared by an extensive simulation study done in [14]. The results show the robustness of the median estimators by demonstrating that they outperform the above mentioned classical robust estimators in certain special situations (e. g. heavy contaminations and large sample sizes).

Now we present a simulation study in order to see if the conclusions concerning robustness obtained in the above quoted work for the median estimator remain valid also for the modified median estimator defined in (9). For the sake of completeness, we have included in the comparisons also the results for the *weighted Bianco and Yohai estimator* (WBY-*estimator*) defined in [9] and MLE.

The robustness is compared by means of simulated performances of all selected estimators in a logit model ε-contaminated at the levels $0 \leq \varepsilon \leq 0.1$ by an alternative data source generating outliers.

The simulated data $Y_1, \ldots, Y_n$ are generated by the contaminated logit model

$$Y_i \sim (1-\varepsilon)\, Be\left(\pi\left(\boldsymbol{x}_i^T \boldsymbol{\beta}_0\right)\right) + \varepsilon Be\left(1 - \pi\left(\boldsymbol{x}_i^T \boldsymbol{\beta}_0\right)\right), \tag{10}$$

where $\boldsymbol{x}_i$ are the concrete regressors

$$\boldsymbol{x}_i = \left(x_{i0} \equiv 1,\ x_{i1} \sim N(0,1)\right)^T \tag{11}$$

and $\boldsymbol{\beta}_0 = (\beta_{00}, \beta_{01})^T = (-2.82,\ 2.82)^T$ are the true parameters leading to the probability

$$\Pr(Y_i = 1) \equiv \pi\left(\boldsymbol{x}_i^T \boldsymbol{\beta}_0\right) = 0.2\,. \tag{12}$$

Let us note that the considered model is the same as used in [3] for demonstration of robustness of their estimator.

The simulation experiment can be described by the following steps:

1. Select $n \in \{50, 100, 250, 500, 1000\}$ and $\varepsilon \in \{0, 0.05, 0.1\}$.
2. Repeat $L = 1000$ times $(l = 1, \ldots, L)$
   a) generate sample of size n
   b) for each method (MLE, Med, MMe, WBY) calculate the estimate $\widetilde{\boldsymbol{\beta}}_n^{(l)} = (\widetilde{\beta}_{n0}^{(l)}, \widetilde{\beta}_{n1}^{(l)})^T$ of the true parameter $\boldsymbol{\beta}_0$
3. Output: the mean absolute errors

$$MAE(n) = \frac{1}{2000} \sum_{l=1}^{1000} \left(\left| \widetilde{\beta}_{n0}^{(l)} - \beta_{00} \right| + \left| \widetilde{\beta}_{n1}^{(l)} - \beta_{01} \right| \right) .$$

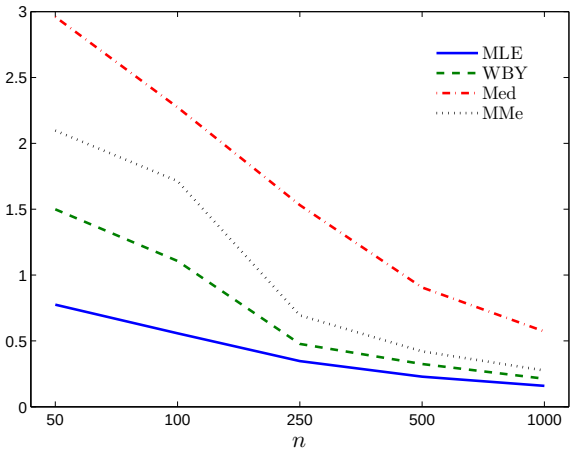

Figure 2. Mean absolute errors (MAE) for data without contamination

The obtained results are presented in Figure 2 for non-contaminated data and in Figure 3 for data with outliers. First of all one can see that there is a clear gain of precision of the MMe-estimator with respect to the original Med-estimator. This effect is visible particularly in the parts for non-contaminated data ($\varepsilon = 0$) and it is weakening with increasing sample size and level of

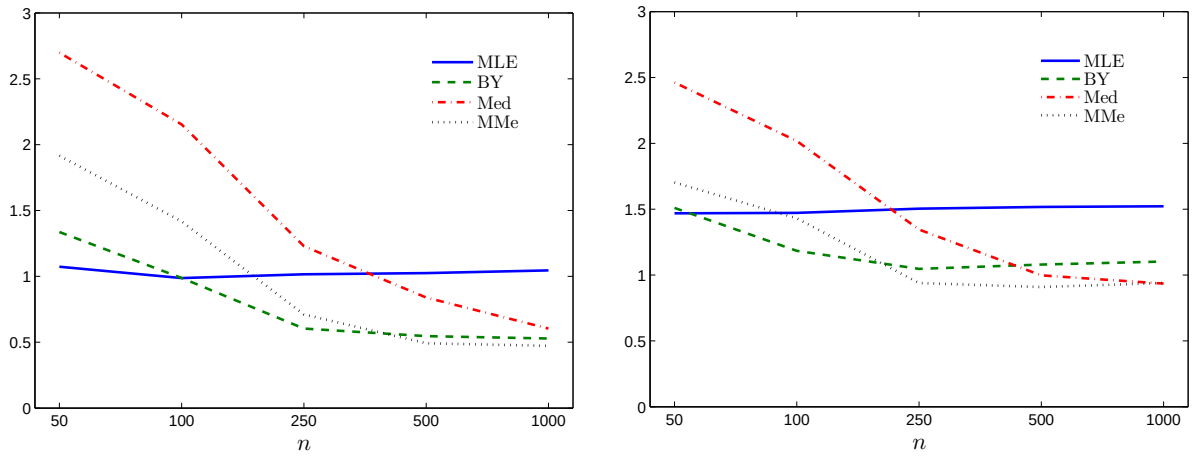

Figure 3. Mean absolute errors (MAE) for data with 5% of outliers (left) and 10% of outliers (right)

contamination. E.g. for $n \in \{500, 1000\}$ and 10% of outliers ($\varepsilon = 0.1$) the MME-estimator and Med-estimator are comparable in the sense of MAE.

Generally, if there is no contamination the MLE is clearly the best followed by WBY-estimator and MMe-estimator. This picture, however, dramatically changes with contamination of the data. For $\varepsilon = 0.05$ one can observe that for smaller sample sizes ($n = 50, 100$) the WBY-estimator shows the best behavior and for larger sample sizes ($n \geq 250$) MME-estimator has the smallest MAE's. For larger contamination ($\varepsilon = 0.1$) the median based estimators seem to be more resistent to distortion of the model than the WBY-estimator and MLE.

3 Median based tests

As mentioned in the introduction, while the problem of parameter estimation in logistic regression is widely studied in the literature, much less attention has been paid to tests about parameters of these models. In this section we propose a new statistics for testing general hypotheses about the parameter $\boldsymbol{\beta}$ of the logistic regression model.

3.1 Wald-type test

Based on the modified median estimator, we define a Wald-type test statistics for the problem of testing

$$H_0 : \boldsymbol{K}^T\boldsymbol{\beta} = \boldsymbol{m} \qquad \text{against} \qquad H_1 : \boldsymbol{K}^T\boldsymbol{\beta} \neq \boldsymbol{m}, \tag{13}$$

where $\boldsymbol{K}^T$ is a matrix of dimension $r \times d$ and $\text{rank}(\boldsymbol{K}^T) = r$ and $\boldsymbol{m}$ is a vector of order r such that $\text{rank}(\boldsymbol{K}^T, \boldsymbol{m}) = r$. For example, the considered null hypothesis could be

$$H_0 : \beta_1 = \beta_2 = ... = \beta_d = 0 \quad \text{if} \quad \boldsymbol{K}^T = (\boldsymbol{I}_{d\times d}),\ \boldsymbol{m} = \boldsymbol{0}_d,$$

or

$$H_0 : \beta_i = 0 \quad \text{if} \quad \boldsymbol{K}^T = (0, \ldots, 0, 1, 0, \ldots, 0)_d,\ \boldsymbol{m} = 0\,.$$

The proposed Wald-type test statistic for testing the null hypothesis (13) is given by

$$W_n(\widehat{\boldsymbol{\beta}}) = n(\boldsymbol{K}^T\widehat{\boldsymbol{\beta}} - \boldsymbol{m})^T \left(\boldsymbol{K}^T\widehat{V}_n(\widehat{\boldsymbol{\beta}})\boldsymbol{K}\right)^{-1} (\boldsymbol{K}^T\widehat{\boldsymbol{\beta}} - \boldsymbol{m}),$$

where $\widehat{\boldsymbol{\beta}}$ is the MMe of $\boldsymbol{\beta}$ and $\widehat{V}_n(\widehat{\boldsymbol{\beta}})$ is an estimator of the asymptotic covariance matrix of the modified median estimator. For more details see [13]. The classical Wald type test statistic based on maximum likelihood estimator, for this problem, can be seen for instance in [10].

In [13] we show that under some regularity assumptions the asymptotic distribution of our Wald-type test statistics is a chi-square distribution with r degrees of freedom and we derive approximation of the power function of the proposed test.

Let us now illustrate performance of the introduced test statistics by some results of a simulation study done in [13]. We compare behavior of three tests:

- the classical Wald test based on MLE
- the Wald-type test based on MMe-estimator
- the Wald-type test introduced in [2] and based on WBY-estimator.

We further consider two scenarios: sizes of the tests and powers of the tests and two types of contamination: with outliers and with leverage points.

3.2 Simulation experiment - sizes of tests

For comparison of sizes we use the same model as in [2]. In the case of outliers, the response variables Y_i are generated from the model

$$Y_i \sim (1-\varepsilon)\, Be\left(\pi\left(\boldsymbol{x}_i^T\boldsymbol{\beta}_0\right)\right) + \varepsilon Be\left(1-\pi\left(\boldsymbol{x}_i^T\boldsymbol{\beta}_0\right)\right), \tag{14}$$

with $\boldsymbol{x}_i = (1, U_{i1}, U_{i2})$, where U_{i1}, U_{i2} are independent and $N(0,1)$ distributed and $\boldsymbol{\beta}_0 = (0,2,2)^T$. In the case of leverage points, the data follow the same but non-contaminated model ($\varepsilon = 0$)

$$Y_i \sim Be\left(\pi\left(\boldsymbol{x}_i^T\boldsymbol{\beta}_0\right)\right) \tag{15}$$

and for each considered n 2% of misclassified observations are added on a hyperplane parallel to the true discriminating hyperplane $\boldsymbol{x}^T\boldsymbol{\beta}$ with a shift equal to $m \times \sqrt{2}$ and with the first covariate x_1 around 3.

We consider the null hypothesis $H_0 : \boldsymbol{\beta}_0 = (0,2,2)$, we select $n \in [50, 1000]$ and for each test we evaluate simulated sizes (frequencies of rejection) based on 5000 simulated samples.

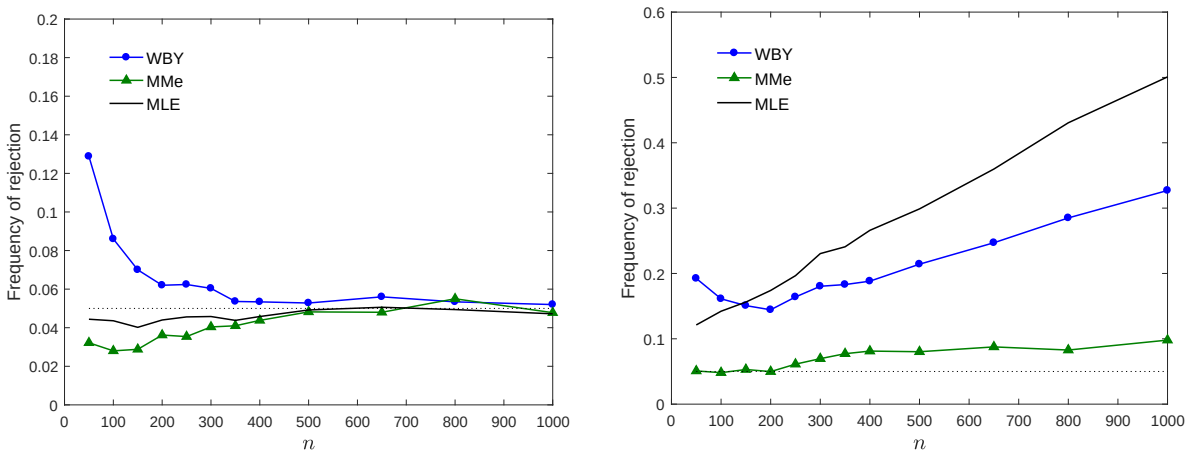

Figure 4. Simulated sizes in the non-contaminated case (left) and in the case with 2% of outliers (right)

The observed frequencies of rejection are presented in Figures 4 and 5. In Figure 4 we observe that for non-contaminated data all the test statistics have simulated size reasonably close to the nominal level 0.05 at least for sample sizes $n \geq 200$. But for smaller sample sizes the robust tests seem to be approximately reliable too. Under contamination with outliers the test based on the MMe-estimator is clearly the most stable one while the classical Wald test is the worst one. In Figure 5 we see results for contamination with

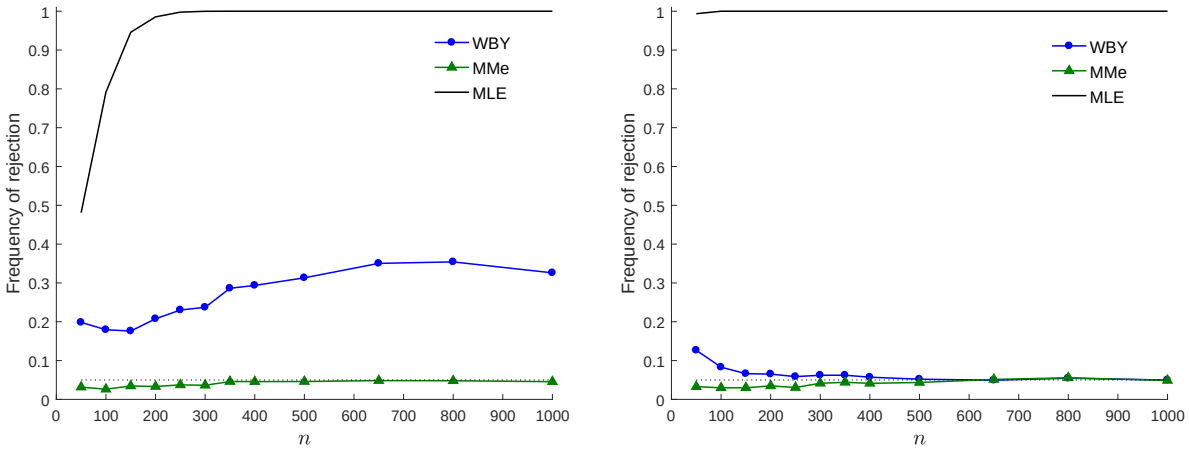

Figure 5. Simulated sizes in case with 2% of leverage points for $m = 2$ (left) and in the case with 2% of leverage points for $m = 4$. (right)

misclassified observations and the situation is quite different. For $m = 4$, when the misclassified observations are further away from the discriminating hyperplane, the levels of the tests based on the MMe and WBY-estimators remain very stable while the classical test breaks in level since the frequencies of rejection under the null hypothesis are 1 or near to 1. For $m = 2$, when the misclassified observations are in fact more close to outliers, the situation is more similar to that presented in Figure 4 and the test based on the MMe-estimator has the best behavior.

So the main message from these figures is that for huge leverage points, the proposed test is comparable with the test based on WBY-estimator and in the case of outliers, it is significantly better.

3.3 Simulation experiment - powers of tests

Now we shift our attention to the powers of the Wald-type tests and we use the same simulation scenario as used in [2] for the study of powers of their WBY test statistics. Namely, the non-contaminated data are generated from the model

$$Y_i \sim Be\left(\pi\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)\right), \quad \text{where} \quad \pi\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right) = \frac{\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)}{1+\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)},$$

$\boldsymbol{x}_i^T = (1, U_i)$, U_i are i.i.d. $N(0,1)$, $\boldsymbol{\beta} = \boldsymbol{\beta}_0 + \Delta\boldsymbol{c}$ and $\boldsymbol{\beta}_0 = (-2.82, 2.82)^T$. We consider the hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ and the direction in which we move away from H_0 is selected as $\boldsymbol{c} = (1,1)^T$ for $\Delta \in (-1, 1.5)$. We further assume

two types of contamination: with outliers and with leverage points. In the first case data are generated from the ε-contaminated source

$$Y_i \sim (1-\varepsilon)\, Be\left(\pi\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)\right) + \varepsilon Be\left(1-\pi\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)\right) .$$

In the latter case non-contaminated sample is generated for $\varepsilon = 0$ and five or ten distorted observations are added at points with $y = 0$ and $\boldsymbol{x}^T = (1, 5)$.

According to the results obtained for levels of the tests we select the sample size $n = 250$ and for each considered value of parameter Δ we generate 1000 samples and evaluate the simulated powers (frequencies of rejection) for all the statistics under consideration. Results for non-contaminated case, contamination with outliers and contamination with leverage points are presented in Figures 6, 7, and 8, respectively.

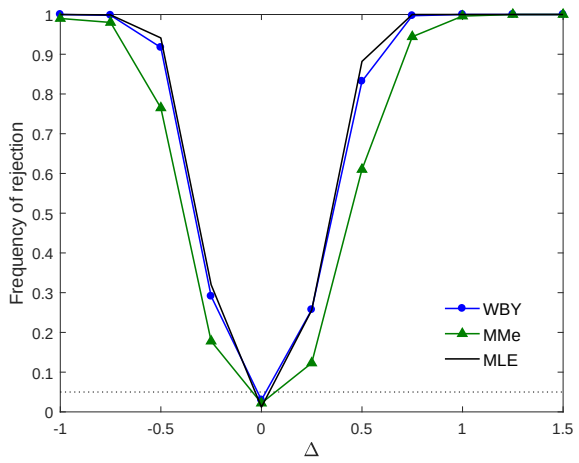

Figure 6. Simulated powers in the non-contaminated case

In Figure 6 one can observe that in the non-contaminated case the powers of the proposed test are comparable with the powers of the remaining two tests. However, with increasing level of contamination with outliers, the behavior of the test based on the MME-estimator seems to be the most stable as can be seen in Figure 7.

Under contamination with distorted observations, we see in Figure 8 that the Wald-type tests based on the MME and WBY-estimators show very stable performance with almost the same power as for the non-contaminated samples while the classical Wald test becomes non-informative since the power function equals 1 for any value of parameter Δ.

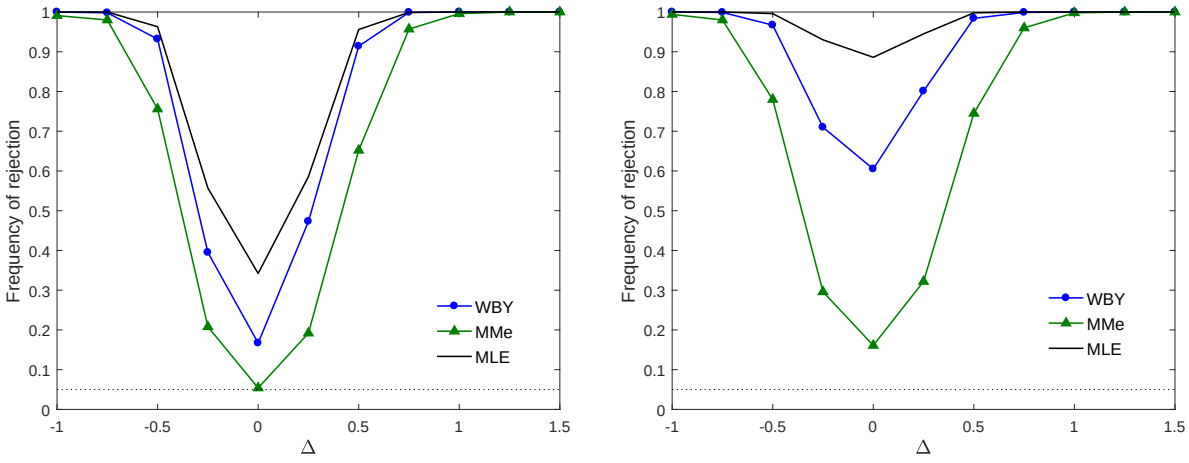

Figure 7. Simulated powers in the case with 2% of outliers (left) and in the case with 5% of outliers (right)

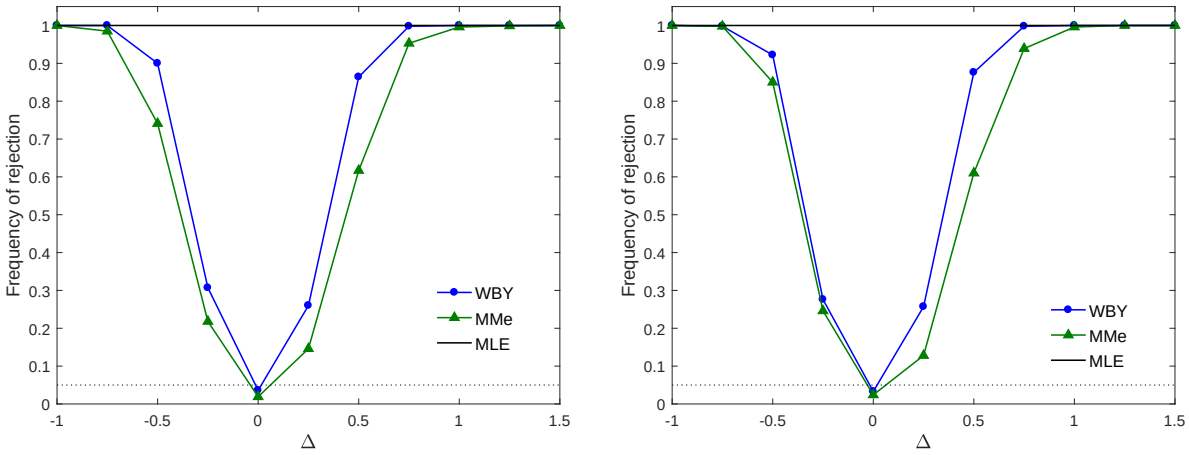

Figure 8. Simulated powers in the case with 5 leverage points (left) and in the case with 10 leverage points (right)

4 Conclusion

To summarize the lecture into a few main points, we can say the following.

- The new modified median estimator appears to be more robust than compared tests for larger sample sizes and higher level of contamination.
- For huge leverage points, the proposed test is comparable with the test based on WBY.
- In the case of outliers, the proposed test is significantly better than WBY-based test.

- Statistical smoothing can also be applied to integer valued observations Y in other discrete models. This makes the statistical methods, developed for continuous models, more widely applicable in discrete statistics than just in the particular situation studied in this lecture.

References

[1] Basu, A., Ghosh, A., Mandal, A., Martin, N., Pardo, L. (2017). A Wald-type test statistic for testing linear hypothesis in logistic regression models based on minimum density power divergence estimator. arXiv:1609.07452.

[2] Bianco, A. M., Martínez, E. (2009). Robust testing in the logistic regression model. *Computational Statistics and Data Analysis*, 53, 4095–4105.

[3] Bianco, A. M., Yohai, V. J. (1996). Robust estimation in the logistic regression model. in: H. Rieder (Ed.), *Robust Statistics, Data Analysis, and Computer Intensive Methods*, Lecture Notes in Statistics, vol. 109, Springer Verlag, New York , 17–34.

[4] Bondell, H. D. (2005). Minimum distance estimation for the logistic regression model. *Biometrika*, 92, 724–731.

[5] Bondell, H. D. (2008). A characteristic function approach to the biased sampling model, with application to robust logistic regression. *Journal of Statistical Planning and Inference*, 138, 742–755.

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

[7] Carroll, R. J., Pederson, S. (1993). On Robustness in the logistic regression model. *Journal of the Royal Statistical Society*, Series B, 55, 669–706.

[8] Christmann, A. (1994). Least Median of Weighted Squares in Logistic Regression with Large Strata. *Biometrika*, 81, 413–417.

[9] Croux, C., Haesbroeck, G. (2003). Implementing the Bianco and Yohai estimator for logistic regression. *Computational Statistics and Data Analysis*, 44, 273–295.

[10] Greene, W. H. (2003). *Econometric Analysis.* Prentice Hall, Upper Saddle River, New Jersey.

[11] Hampel, F. R., Roncheti, E.M., Rousseeuw, P. J., Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions.* John Wiley & Sons, New York.

[12] Heritier, S., Ronchetti, E. (1994). Robust bounded–influence tests in general parametric models. *Journal of the American Statistical Association*, 89, 897–904.

[13] Hobza, T., Martín, N., Pardo, L. (2017). A Wald-type test statistic based on robust modified median estimator in logistic regression models. *Journal of Statistical Computation and Simulation*,87(12), 2309–2333.

[14] Hobza, T., Pardo, L., Vajda, I. (2008). Robust median estimator in logistic regression. *Journal of Statistical Planning and Inference*, 138, 3822–3840.

[15] Jurečková, J., Sen, P. K. (1996), *Robust Statistical Procedures.* Wiley, New York.

[16] Nelder, J. A., Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, 135(3), 370–384.

[17] Morgenthaler, S. (1992). Least-absolute-deviations fits for generalized linear models. *Biometrika,* 79, 747–754.

[18] Yohai, V. J. (1987). High breakdown point high efficiency robust estimates for regression. *The Annals of Statistics,* 15(2), 692–656.

[19] Zwanzig, S. (1997). On L_1-norm Estimators in Nonlinear Regression and in Nonlinear Error-in-Variables Models. *IMS Lecture Notes*, 31, 101–118, Hayward.

Ing. Tomáš Hobza, Ph.D. - Curriculum Vitae

Personal data:

Name: **Tomáš Hobza**

Birth: **18. 1. 1975**

Nationality: **Czech**

Mail address: **FJFI ČVUT, Trojanova 13, CZ-12000 Praha 2**

Home address: **Halouny 68, 267 28 Svinaře**

Phone.: +**420 224 358 547**

E-mail: **hobza@fjfi.cvut.cz**

Education:

1993-1998	Czech Technical University in Prague, Faculty of Nuclear Sciences and Physical Engineering, field of study: Mathematical Engineering, graduated on 8th June 1998 (**M.Sc. degree**); Thesis: *Modeling of Density Estimates*
1998-2003	Czech Technical University in Prague, Faculty of Nuclear Sciences and Physical Engineering, Ph.D. student, field of study: Mathematical Engineering; Thesis defended on December 12, 2003 (**Ph.D. degree**); Thesis: *Asymptotics of Some Histogram-based Density Estimates*

Affiliation:

Sept. 1998 - Dec. 2011:	Academy of Sciences of the Czech Republic, half-time employee in the Institute of Information Theory and Automation (ÚTIA AV ČR), Department of Stochastic Informatics. Position: Young researcher
Since 2004:	Czech Technical University in Prague, Faculty of Nuclear Sciences and Physical Engineering (ČVUT FJFI), Department of Mathematics. Present position: Assistant professor
Teaching:	Combinatorics and Probability; Mathematical Statistics; Applied Statistical Methods; Information Theory; Generalized Linear Models. Czech Technical University, since 2004
Research areas:	Small Area Estimation: methods based on linear and generalized linear models, Divergence measures and their application in statistics

Participation in research grants:

1998-1999	participant of the grant EU - *Copernicus 579: Research of ATM networks*, coordinated by Igor Vajda, ÚTIA AV ČR
1999-2001	participant of the grant GA ČR n. 102/99/1137, *Estimation and optimization in telecommunication networks*, coordinated by Igor Vajda, ÚTIA AV ČR

2000	principle investigator of the CTU grant n. 300009704, *Optimization of binwidth in the generalized piecewise linear histogram.*
2002-2004	participant of the grant GA ČR n. 201/02/1391, *Asymptotic Properties of Information Contained in Quantized Observations*, coordinated by Igor Vajda, ÚTIA AV ČR
2004-2005	participant of the grant AV ČR n. A1075403, *New results in testing the goodness-of-fit based on Pearson-type statistics*, coordinated by Igor Vajda, ÚTIA AV ČR
2007-2009	participant of the grant GA ČR n. 102/07/1131, *Theoretical information in stochastic data and its empirical approximation and application in processes of detection and identification*, coordinated by Igor Vajda, ÚTIA AV ČR
2010-2012	participant of the grant GA ČR n. P202/10/0618, *Bregman distances, divergence of distributions, information retrieval, optimal decisions, machine learning*, coordinated by Igor Vajda, ÚTIA AV ČR
2010-2012	participant of the grant MTM2009-06997, *Fitting marginal models for longitudinal data*, coordinated by María del Carmen Pardo, UCM, Spain
2015-present	participant of the grant GA ČR n. P103/15/15049S, *Detection of stochastic universalities in non-equilibrium states of socio-physical systems by means of Random Matrix Theory*, coordinated by Milan Krbálek, ČVUT FJFI.

Long-term stays abroad (in the frame of the EU Socrates-Erasmus program)**:**

2001:	6 months research stay at the Miguel Hernández University, Elche, Spain
2003:	1 month research stay at the Miguel Hernández University, Elche, Spain
2007:	3 months research stay at the Miguel Hernández University, Elche, Spain
2011:	1 month research stay at the Miguel Hernández University, Elche, Spain

Other:

- 2007 - 2010, 2010 - 2013, 2013 - 2016, 2016 - present, member of the Academic Senate of the Faculty of Nuclear Sciences and Physical Engineering, since May 2014 chair of the Academic Senate.
- Language skills: English (postgraduate exam), Spanish (advanced), German (postgraduate exam)
- Scientometric data (as of Sep 30, 2017); citations (WoS, Scopus, . . .): 27 (17 without self-citation);

Prague, September 2017 Tomáš Hobza

List of publications

Articles in impacted journals

1. Hobza, T., Martín, N., Pardo, L. (2017). A Wald-type test statistic based on robust modified median estimator in logistic regression models. Journal of Statistical Computation and Simulation, 87(12), pp. 2309-2333.

2. Hobza, T., Morales, D., Santamaría, L. (2017). Small area estimation of poverty proportions under unit-level temporal binomial-logit mixed models. TEST (in print), DOI: 10.1007/s11749-017-0545-3

3. Hobza, T., Morales, D. (2016). Empirical best prediction under unit-level logit mixed models. Journal of Official Statistics, 32(3), pp. 661-692.

4. Krbálek, M., Hobza, T. (2016). Inner structure of vehicular ensembles and random matrix theory. Physics Letters A, 380(21), pp. 1839-1847.

5. Pardo, M.C., Hobza, T. (2014). Outlier detection method in GEEs, Biometrical Journal, Vol. 56(5), pp. 838-850.

6. Hobza, T., Morales, D., Pardo, L. (2014). Divergence-based tests of homogeneity for spatial data. Statistical Papers, 55(4), pp. 1059-1077.

7. Hobza, T. and Morales, D. (2013). Small area estimation under random regression coefficient models. Journal of Statistical Computation and Simulation, 83(11), pp. 2160-2177.

8. Esteban, M.D., Herrador, M., Hobza, T., Morales, D. (2013). A modified nested-error regression model for small area estimation. Statistics: A Journal of Theoretical and Applied Statistics, 47(2), pp. 258-273.

9. Hobza, T., Pardo, L. and Vajda, I. (2012). Robust median estimator for generalized linear models with binary responses. Kybernetika, 48(4), pp. 768-794.

10. Esteban, M.D., Herrador, M., Hobza, T., Morales, D. (2011). A Fay-Herriot model with different random effect variances. Communications in Statistics - Theory and Methods, 40(5), pp. 785-797.

11. Hobza T., Morales D., Pardo L. (2009). Rényi statistics for testing equality of autocorrelation coefficients. Statistical Methodology, 6(45), pp. 424-436.

12. Hobza T., Molina, I., Morales D. (2009). Multi-sample Rényi test statistics. Brazilian Journal of Probability and Statistics, 23(2), pp. 196-215.

13. Hobza T., Pardo L., Vajda, I. (2008). Robust Median Estimator in Logistic Regression. Journal of Statistical Planning and Inference 138(12), pp. 3822-3840.

14. Esteban, M.D., Hobza, T., Morales, D., Marhuenda, Y. (2008). Divergence-based tests for model diagnostic. Statistics and Probability Letters, 78(13), pp. 1702-1710.

15. Hobza, T., Molina, I., Vajda, I. (2005). On convergence of Fisher Informations in continuous models with quantized observation spaces. TEST, 14(1), pp. 151-179.

16. Hobza, T., Molina, I., Morales, D. (2003). Likelihood divergence statistics for testing hypothesis in familial data. Communications in Statistics - Theory and Methods, 32(2), pp. 415-434.

17. Berlinet, A., Hobza, T., Vajda, I. (2002). Generalized piecewise linear histograms. Statistica Neerlandica, 56(3), pp. 301-313.

18. Hobza, T., Vajda, I. (2001). On the Newcomb-Benford law in models of statistical data. Revista Matematica Complutense, 14(2), pp. 1-13.

Reviewed articles in books

1. Herrador, M., Esteban, M. D., Hobza, T., Morales, D. (2011). An Area-Level Model with Fixed or Random Domain Effects in Small Area Estimation Problems. Modern Mathematical Tools and Techniques in Capturing Complexity - Understanding Complex Systems, Springer Berlin, pp. 303 - 314.

2. Hobza, T., Morales, D. (2011). Small Area Estimation of Poverty Proportions under Random Regression Coefficient Models. Modern Mathematical Tools and Techniques in Capturing Complexity - Understanding Complex Systems, Springer Berlin, pp. 315 - 328.

Articles in international reviewed journals

1. Berlinet, A., Hobza, T., Vajda, I. (2002). Asymptotics for generalized piecewise linear histograms. Annals de l'Institute de Statistique de l'Université de Paris, 46(3), pp. 3-19.

Other publications

1. Hobza, T. (2003). Asymptotics of some histogram-based density estimates. PhD dissertation, Czech Technical University in Prague, Czech Republic.