Resistant estimators in Poisson and Gamma models with missing responses

Ana Bianco

Instituto de Cálculo, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and CONICET

Graciela Boente

Departamento de Matemáticas and Instituto de Cálculo,

Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires and CONICET

Isabel Rodrigues

Departamento de Matemática and CEMAT, Instituto Superior Técnico,

Technical University of Lisbon (TULisbon), Lisboa, Portugal

Abstract

In many situations, data follow a generalized linear model in which the mean of the responses is modelled, through a link function, linearly on the covariates. In this paper, robust estimators for the regression parameter are considered when missing data occur in the responses. The estimators turn out to be consistent under mild assumptions. In particular, resistant methods for Poisson and Gamma models are introduced. A simulation study allows to compare the behaviour of the classical and robust tests, under different contamination schemes. The procedure is also illustrated analysing some real data sets.

Key Words: Fisher–consistency, Generalized Linear Models, Missing Data, Outliers, Robust Estimation

AMS Subject Classification: MSC 62F35 MSC 62F05

1 Introduction

Nelder and Wedderburn (1972) introduced the generalized linear model,GLM, which became a very popular technique for modelling a wide variety of data as an alternative to the linear model (see, McCullagh and Nelder, 1989). It assumes that the observations $(y_i, \mathbf{x}_i^{\mathrm{T}})$, $1 \leq i \leq n$, $\mathbf{x}_i \in \mathbb{R}^k$, are independent with the same distribution as $(y, \mathbf{x}^{\mathrm{T}}) \in \mathbb{R}^{k+1}$ such that the conditional distribution of $y|\mathbf{x}$ belongs to the canonical exponential family

$$\exp\left\{\left[y\theta(\mathbf{x}) - B\left(\theta(\mathbf{x})\right)\right]/A(\tau) + C(y,\tau)\right\},$$

for known functions A, B and C. In this situation, if we denote by B' the derivative of B, the mean $\mu(\mathbf{x}) = \mathbb{E}(y|(\mathbf{x})) = B'\left(\theta(\mathbf{x})\right)$ is modelled linearly through a known link function, g, i.e., $g(\mu\left(\mathbf{x}\right)) = \theta(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$.

In this setting, the classical estimators are based on the minimization of the deviance, which is equivalent to the maximum likelihood method. It is very well known that theses procedures can be affected by anomalous observations. To overcome this problem, robust procedures have been developed and among others we can cite the methods proposed by Stefanski *et al.* (1986), Künsch *et al.* (1989), Bianco and Yohai (1996), Cantoni and Ronchetti (2001, 2006), Croux and Haesbroeck (2003) and Bianco *et al.* (2005), see also, Maronna *et al.* (2006). Even when developing robust methods for GLM has been an active research area in the last decades, all these methods were designed for complete data sets. However, in practice, missing data can arise and hence these procedures are no longer a useful tool.

Indeed, missing responses may be introduced just by design, as it is the case of two-stage studies, or simply by chance. In some cases the responders may refuse to answer, for instance about some private issues, or the responses y's may be an expensive measure to be obtained. In other cases, missing data may be caused by some loss of information due to uncontrollable factors or by failure on recording the correct information. In this paper we will focus our

attention on robust inference when the response variable has missing observations but the covariate $\mathbf{x}$ is totally observed.

We introduce a robust procedure to estimate the parameter $\boldsymbol{\beta}$, under a GLM model, which includes, when there are no missing data, the family of estimators previously studied. The robust estimates of $\boldsymbol{\beta}$ are consistent under mild assumptions. The paper is organized as follows. The robust proposal is given in Section 2, consistency results are provided in Section 3. Two real data sets are analysed in Section 5 while the results of a Monte Carlo study are summarized in Section 4. Proofs are relegated to the Appendix.

2 Robust inference

2.1 The robust estimators

Suppose we obtain a random sample of incomplete data $\left(y_i, \mathbf{x}_i^{\mathrm{T}}, \delta_i\right)$, $1 \leq i \leq n$, of a generalized linear model where $\delta_i = 1$ if y_i is observed, $\delta_i = 0$ if y_i is missing and $(y_i, \mathbf{x}_i) \in \mathbb{R}^{k+1}$ are such that $y_i|\mathbf{x}_i \sim F(\cdot, \mu_i, \tau)$ with $\mu_i = H(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})$ and $\mathrm{VAR}(y_i|\mathbf{x}_i) = A^2(\tau)V^2(\mu_i) = A^2(\tau)B''\left(\theta(\mathbf{x}_i)\right)$ with B'' the second derivative of B. Let $(\boldsymbol{\beta}, \tau) \in \mathbb{R}^{k+1}$ denote the true parameter values and $\mathbb{E}_F$ the expectation under the true model, thus $\mathbb{E}_F(y|\mathbf{x}) = H(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$. In a more general situation, we will think of τ as a nuisance parameter such as the tuning constant for the score function to be considered.

Let $(y, \mathbf{x}^{\mathrm{T}}, \delta)$ be a random vector with the same distribution as $\left(y_i, \mathbf{x}_i^{\mathrm{T}}, \delta_i\right)$. As mentioned in the Introduction our aim is to define the robust estimators of the regression parameter when missing responses occur. For that purpose, an ignorable missing mechanism will be imposed by assuming that y is missing at random (MAR), that is, δ and y are conditionally independent given $\mathbf{x}$, i.e.,

$$P\left(\delta = 1|(y, \mathbf{x})\right) = P\left(\delta = 1|\mathbf{x}\right) = p\left(\mathbf{x}\right) \ . \tag{1}$$

Usually, it is assumed that $\inf_{\mathbf{x}} p(\mathbf{x}) > 0$ which means that at any value of the covariate response variables are observed. This assumption can be avoided by introducing a weight function with bounded support at the cost of some loss of efficiency.

Let $w_1 : \mathbb{R}^k \to \mathbb{R}$ be a weight function to control leverage points on the carriers $\mathbf{x}$ and $\rho : \mathbb{R}^3 \to \mathbb{R}$ a loss function. Define

$$S_n(\mathbf{b}, t) = \frac{1}{n} \sum_{i=1}^{n} \delta_i \rho\left(y_i, \mathbf{x}_i^{\mathrm{T}}\mathbf{b}, t\right) w_1(\mathbf{x}_i) , \tag{2}$$

$$S(\mathbf{b}, t) = \mathbb{E}_F\left[\delta \rho\left(y, \mathbf{x}^{\mathrm{T}}\mathbf{b}, t\right) w_1(\mathbf{x})\right] = \mathbb{E}_F\left[p(\mathbf{x}) \rho\left(y, \mathbf{x}^{\mathrm{T}}\mathbf{b}, t\right) w_1(\mathbf{x})\right] . \tag{3}$$

Let us assume that $w_1(\cdot)$ and $\rho(\cdot)$ are such that, $S(\boldsymbol{\beta}, \tau) = \min_{\mathbf{b}} S(\mathbf{b}, \tau)$, then in order to estimate $\boldsymbol{\beta}$ one can minimize $S_n(\mathbf{b}, \tau)$ that provides, a consistent estimator of $S(\mathbf{b}, \tau)$. Note that τ plays the role of a nuisance parameter.

Throughout the paper, we will assume Fisher–consistency, i.e., that $S(\boldsymbol{\beta}, \tau) = \min_{\mathbf{b}} S(\mathbf{b}, \tau)$, $\boldsymbol{\beta}$ being the unique minimum (see Remark 2.1 below). The estimators can thus be defined as follows.

Let $\widehat{\tau} = \widehat{\tau}_n$ be robust consistent estimators of τ, the robust simplified estimator $\widehat{\boldsymbol{\beta}}$ of the regression parameter is defined as

$$\widehat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\mathbf{b}} S_n(\mathbf{b}, \widehat{\tau}) . \tag{4}$$

When ρ is continuously differentiable, if we denote by $\Psi(y, u, t) = \partial \rho(y, u, t)/\partial u$, $\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}}$ satisfy the differentiated equations $S^{(1)}(\boldsymbol{\beta}, \tau) = 0$ and $S_n^{(1)}(\mathbf{b}, \widehat{\tau}) = 0$, respectively, where

$$S^{(1)}(\mathbf{b}, t) = \mathbb{E}_F\left(\Psi\left(y, \mathbf{x}^{\mathrm{T}}\mathbf{b}, t\right) w_1(\mathbf{x}) p(\mathbf{x}) \mathbf{x}\right) ,$$

$$S_n^{(1)}(\mathbf{b}, t) = \frac{1}{n} \sum_{i=1}^{n} \delta_i \Psi\left(y_i, \mathbf{x}_i^{\mathrm{T}}\mathbf{b}, t\right) w_1(\mathbf{x}_i) \mathbf{x}_i .$$

When $S_n(\mathbf{b}, \widehat{\tau})$ has only one critical point, i.e., when the equation $S_n^{(1)}(\mathbf{b}, \widehat{\tau}) = 0$ has only one root, corresponding to the minimum of $S_n(\mathbf{b}, \widehat{\tau})$, the estimator $\widehat{\boldsymbol{\beta}}$ can be computed using a Newton–Raphson approach.

To improve the bias caused in the estimation by the missing mechanism, robust propensity score estimators may be considered using an estimator of the missingness probability. Denote by $\widehat{p}(\mathbf{x})$ any estimator of $p(\mathbf{x})$. For instance, if we assume that the missingness probability is given by the logistic model, i.e., that $p(\mathbf{x}) = G_{\mathrm{L}}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\lambda}_0)$ where $G_{\mathrm{L}}(s) = (1+e^{-s})^{-1}$ is the logistic distribution function, we only need to estimate the parameter $\boldsymbol{\lambda}$ to define the estimator $\widehat{p}(\mathbf{x})$. Let $\mathcal{P} = \{q : \mathbb{R}^k \to \mathbb{R} \text{ such that } 0 < q(\mathbf{x}) \leq 1\}$, $S_{\mathrm{P},n} : \mathbb{R}^{k+1} \times \mathcal{P} \to \mathbb{R}$ and its related functional $S_{\mathrm{P}} : \mathbb{R}^{k+1} \times \mathcal{P} \to \mathbb{R}$ as

$$
\begin{aligned}
S_{\mathrm{P},n}(\mathbf{b}, t, q) &= \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i}{q(\mathbf{x}_i)} \rho\left(y_i, \mathbf{x}_i^{\mathrm{T}}\mathbf{b}, t\right) w_1(\mathbf{x}_i) \,, \qquad (5) \\
S_{\mathrm{P}}(\mathbf{b}, t, q) &= \mathbb{E}_F\left[\frac{\delta}{q(\mathbf{x})} \rho\left(y, \mathbf{x}^{\mathrm{T}}\mathbf{b}, t\right) w_1(\mathbf{x})\right] = \mathbb{E}_F\left[\frac{p(\mathbf{x})}{q(\mathbf{x})} \rho\left(y, \mathbf{x}^{\mathrm{T}}\mathbf{b}, t\right) w_1(\mathbf{x})\right] \,. \qquad (6)
\end{aligned}
$$

The *robust propensity score estimator* $\widehat{\boldsymbol{\beta}}_{\mathrm{P}}$ is defined as

$$
\widehat{\boldsymbol{\beta}}_{\mathrm{P}} = \underset{\mathbf{b}}{\operatorname{argmin}}\, S_{\mathrm{P},n}(\mathbf{b}, \widehat{\tau}_{\mathrm{P}}, \widehat{p}) \,, \qquad (7)
$$

where $\widehat{\tau}_{\mathrm{P}}$ is a robust consistent estimator of τ, possible different than the one previously considered. Note that now τ and $q(\mathbf{x})$ play the role of nuisance parameters. Moreover, it is worth noticing that $S_{\mathrm{P}}(\mathbf{b}, t, p) = \mathbb{E}_F\left[\rho\left(y, \mathbf{x}^{\mathrm{T}}\mathbf{b}, t\right) w_1(\mathbf{x})\right]$, i.e., it corresponds to the objective function when the sample contains no missing responses. Throughout the paper, we will assume Fisher–consistency, i.e., that $S_{\mathrm{P}}(\boldsymbol{\beta}, \tau, p) = \min_{\mathbf{b}} S_{\mathrm{P}}(\mathbf{b}, \tau, p)$, $\boldsymbol{\beta}$ being the unique minimum.

As above, when ρ is continuously differentiable, if we denote by $\Psi(y, u, t) = \partial\rho(y, u, t)/\partial u$, $\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}}$ satisfy the differentiated equations $S_{\mathrm{P}}^{(1)}(\boldsymbol{\beta}, \tau, p) = 0$ and $S_{\mathrm{P},n}^{(1)}(\mathbf{b}, \widehat{\tau}, \widehat{p}) = 0$, respectively, where

$$
\begin{aligned}
S_{\mathrm{P}}^{(1)}(\mathbf{b}, t, q) &= \mathbb{E}_F\left(\Psi\left(y, \mathbf{x}^{\mathrm{T}}\mathbf{b}, t\right) w_1(\mathbf{x}) \frac{p(\mathbf{x})}{q(\mathbf{x})}\mathbf{x}\right) \,, \\
S_{\mathrm{P},n}^{(1)}(\mathbf{b}, t, q) &= \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i}{q(\mathbf{x}_i)} \Psi\left(y_i, \mathbf{x}_i^{\mathrm{T}}\mathbf{b}, t\right) w_1(\mathbf{x}_i)\mathbf{x}_i \,.
\end{aligned}
$$

2.2 The loss functions used in the estimation procedure

Two classes of loss functions ρ have been considered in the literature. The first one aims to bound the deviances, while the second one introduced by Cantoni and Ronchetti (2001) bounds the Pearson residuals. For the sake of completeness, we recall their definition.

For families of distributions that can be transformed to avoid an extra parameter in the model, the first class of loss function takes the form of

$$\rho(y,u,t) = \phi_t[-\ln\, f(y,H(u)) + D(y)] + G(H(u))\,, \tag{8}$$

where ϕ_t is a bounded nondecreasing function with continuous derivative φ_t, t being the tuning constant and $f(\cdot,s)$ is the density of the distribution function $F(\cdot,s)$ with $y|\mathbf{x} \sim F\left(\cdot, H\left(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\right)\right)$. To avoid triviality, it is assumed that ϕ_t is non–constant in a positive probability set. Typically, ϕ_t is a function performing like the identity function in a neighbourhood of 0. The function $D(y)$ is typically used to remove a term from the log–likelihood that is independent of the parameter, and can be defined as $D(y) = \ln(f(y,y))$ in order to get the deviance. The correction term G is used to guarantee the Fisher–consistency, and satisfies

$$\begin{aligned} G'(s) &= \int \varphi_t[-\ln\, f(y,s) + A(y)]\, f'(y,s) d\mu(y) \\ &= \mathbb{E}_s\left(\varphi_t[-\ln\, f(y,s) + A(y)]\, f'(y,s)/f(y,s)\right)\,, \end{aligned}$$

where $\mathbb{E}_s$ indicates expectation taken under $y \sim F(\cdot,s)$ and $f'(y,s)$ is shorthand for $\partial\, f(y,s)/\partial s$. Note that, when considering generalized linear models, the maximum likelihood estimator corresponds to the choice $\phi(s) = s$, $D(y) = \ln(f(y,y))$, $G(u) = 0$ and $w_1 \equiv 1$.

In a logistic regression setting $A(\tau) \equiv 1$ and the tuning constant does not need to be estimated and, in order to guarantee existence of solution, Croux and Haesbroeck (2003)

proposed using the score function

$$\phi_c(s) = \begin{cases} s\exp(-\sqrt{c}) & \text{if } s \leq c \\ -2(1+\sqrt{s})\exp(-\sqrt{s}) + (2(1+\sqrt{c})+c)\exp(-\sqrt{c}) & \text{otherwise.} \end{cases}$$

It is worth noting that, when considering the deviance and a continuous family of distributions with strongly unimodal density function, the correction term G can be avoided, as discussed in Bianco *et al.* (2005). For regression models with asymmetric errors, such as the transformed Gamma model, τ plays the role of the tuning constant depending on the shape parameter of the Gamma distribution and so, initial estimators need to be considered. In the case of the Poisson model $\tau = 1$.

The second class of loss functions is based on the proposal given by Cantoni and Ronchetti (2001) for generalized linear models, where they consider a general class of $M-$estimators of Mallows type, by bounding separately the influence of deviations on y and $(\mathbf{x})$. Their approach is based on robustifying the quasi–likelihood, which is an alternative to the generalizations given for generalized linear regression models by Stefanski *et al.* (1986) and Künsch *et al.* (1989). Let $r(y,\mu,\tau) = (y-\mu)\,/\,(V(\mu)A(\tau))$ be the Pearson residuals with $\text{VAR}\,(y_i|\mathbf{x}_i) = A^2(\tau)V^2\,(\mu_i)$. Denote $\nu(y,\mu,\tau) = \psi_c\,(r(y,\mu,\tau))\,/\,(V(\mu)A(\tau))$, with ψ_c an odd nondecreasing score function with tuning constant c, such as the Huber function, and

$$\rho(y,u,t) = -\left[\int_{s_0}^{H(u)} \nu(y,s,t)ds + G(H(u))\right], \tag{9}$$

where τ is such that $\nu(y,s,\tau) = 0$. To ensure Fisher–consistency, the correction term $G(s)$ satisfies $G'(s) = -\mathbb{E}_s\,(\nu(y,s,\tau))$. For the Binomial and Poisson families, explicit forms of the correction term $G(s)$ are given in Cantoni and Ronchetti (2001). The classical counterpart of this approach corresponds to the choice $\psi_c(u) = u$, $w_1 \equiv 1$.

Remark 2.1. The correction factor, denoted $G(s)$, is included to guarantee Fisher–consistency under the true model. Otherwise, one can only ensure that the estimators will be consistent

to the solution $\boldsymbol{\beta}(F)$ of the related functional equations, i.e., to $\boldsymbol{\beta}(F) = \operatorname{argmin}_{\mathbf{b}} S(\mathbf{b}, \tau)$ where $S(\mathbf{b}, t)$ is defined in (3). On the other hand, as it is well known, when $H(u) = u$, i.e., under the linear regression model $y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_i$, Fisher–consistency holds if, for instance, the errors ϵ_i have a symmetric distribution and the score function ψ_c is odd.

Under a logistic regression model, Fisher–consistency can easily be derived for the loss function given by (8), when ϕ satisfies the regularity conditions stated in Bianco and Yohai (1996) and

$$P\left(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} = \alpha\right) < 1, \qquad \forall(\boldsymbol{\beta}, \alpha) \neq 0 \quad . \tag{10}$$

Moreover, it is easy to verify that $\boldsymbol{\beta}$ is the unique minimizer of $S(\mathbf{b}, \tau)$ in this case. The same assertion can be verified for the robust quasi–likelihood proposal if ψ_c is bounded and increasing.

As shown below, under a generalized regression model with the response having a gamma distribution with a fixed shape parameter, Lemma 1 of Bianco *et al.* (2005) allows us to derive Fisher–consistency for the regression parameter by taking conditional expectation, if the score function ϕ is bounded and strictly increasing on the set where it is not constant and if (10) holds.

2.3 A particular case: The Poisson model

In the case of the Poisson distribution with parameter μ the density can be written as

$$f(y, \mu) = \begin{cases} \exp(-\mu)\mu^y/y! & y \in N \cup \{0\} \\ 0 & \text{in other case} \end{cases}$$

with $\mathbb{E}_\mu(y) = \mu$, $\textsc{Var}_\mu(y) = \mu$ and so $\tau = 1$. Hence, $\rho(y, u, \tau)$ given in (8) is $\rho(y, u, 1) = \phi(-y + y\ln y + H(u) - y\ln(H(u))) + G(H(u))$, where

$$G'(t) = -\varphi(t)\exp(-t) - \sum_{j=1}^{\infty} \varphi(j\ln j - j + t - j\ln t)\left(\frac{t-j}{t}\right)\exp(-t)t^j/j!\,.$$

In the particular case of the canonical link function, that is when $\log \mu = u$, i.e. $H(u) = \exp(u)$, we have that $\rho(y, u, 1) = \phi(-y + y \ln y + H(u) - yu + G(H(u))$.

2.4 A particular case: The log–Gamma model

An important application among generalized linear models is the gamma distribution with a $\log-$link. This model is called log–gamma regression and is introduced in Chapter 8 of McCullagh & Nelder (1983). We refer to Bianco *et al.* (2005) for a description on the robust estimators based on deviances for complete data sets and to Heritier *et al.* (2009) for a description on $M-$type estimators. For the sake of completeness, we will describe how to adapt the estimators based on deviances to the situation with missing responses since this will be the model used in our simulation study.

Denote by $d_i(\boldsymbol{\beta}, \tau)$ the deviance component of the i-th observation, i.e., $d_i(\boldsymbol{\beta}, \tau) = 2\tau\ d^*(y_i, \mathbf{x}_i, \boldsymbol{\beta})$ where

$$d^*(y, \mathbf{x}, \boldsymbol{\beta}) = -1 - (\log(y) - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}) + y \exp(-\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}).$$

Let us now assume that we are dealing with the situation in which some of the responses y_i, and so the transformed responses $z_i = \log(y_i)$, may be missing with $\delta_i = 1$ if z_i is observed, $\delta_i = 0$ if z_i is missing and $(z_i, \mathbf{x}_i) \in \mathbb{R}^{k+1}$ are such that $z_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + u_i$, where $u_i \sim \log(\Gamma(\tau, 1))$ and u_i and $\mathbf{x}_i$ are independent. Moreover, δ and z are conditionally independent given $\mathbf{x}$, i.e.,

$$P\left(\delta = 1|(y, \mathbf{x})\right) = P\left(\delta = 1|(u, \mathbf{x})\right) = P\left(\delta = 1|\mathbf{x}\right) = p\left(\mathbf{x}\right)\ ,$$

and so δ and u are independent. Besides, the density of u is $g(u, \tau)$, where

$$g(u, \tau) = \frac{\tau^\tau}{\Gamma(\tau)}\ \exp(\tau(u - \exp(u)))\ . \tag{11}$$

This density is asymmetric and unimodal with maximum at $u_0 = 0$. Note that

$$\begin{aligned} d^*(y, \mathbf{x}, \mathbf{b}) &= -1 - (z - \mathbf{x}^{\mathrm{T}}\mathbf{b}) + \exp(z - \mathbf{x}^{\mathrm{T}}\mathbf{b}) \\ &= -1 - (z - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{x}^{\mathrm{T}}(\boldsymbol{\beta} - \mathbf{b})) + \exp(z - \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{x}^{\mathrm{T}}(\boldsymbol{\beta} - \mathbf{b})) \\ &= -1 - u - \mathbf{x}^{\mathrm{T}}(\boldsymbol{\beta} - \mathbf{b}) + \exp(u)\exp(\mathbf{x}^{\mathrm{T}}(\boldsymbol{\beta} - \mathbf{b})) = \widetilde{d}(u, \mathbf{x}, \boldsymbol{\beta} - \mathbf{b}) \end{aligned}$$

The maximum likelihood estimator, MLE, of $\boldsymbol{\beta}$ is, thus, obtained as

$$\widehat{\boldsymbol{\beta}}_{\mathrm{ML}} = \underset{\mathbf{b}}{\operatorname{argmin}} \sum_{i=1}^{n} \delta_i d^*(y_i, \mathbf{x}_i, \mathbf{b}).$$

As described in Bianco *et al.* (2010) a three step procedure can be considered to compute the estimators when missing responses are present.

- **Step 1.** We first compute an initial $S-$estimate $\widetilde{\boldsymbol{\beta}}_n$ and the corresponding scale estimate $\widehat{\sigma}_n$ taking $b = \frac{1}{2}\sup\rho$ with the complete data set. To be more precise, for each value of $\mathbf{b}$ let $\sigma_n(\mathbf{b})$ be the $M-$scale estimate of $\sqrt{d^*(y_i, \mathbf{x}_i, \mathbf{b})}$ given by

$$\frac{1}{\sum_{i=1}^{n}\delta_i}\sum_{i=1}^{n}\delta_i\phi\left(\frac{\sqrt{d^*(y_i, \mathbf{x}_i, \mathbf{b})}}{\sigma_n(\mathbf{b})}\right) = b,$$

  where ϕ is Tukey's bisquare function.

  The $S-$estimate of $\boldsymbol{\beta}$ for the considered model is defined by

$$\widetilde{\boldsymbol{\beta}}_n = \underset{\mathbf{b}}{\operatorname{argmin}}\ \sigma_n(\mathbf{b}) \tag{12}$$

  and the corresponding scale estimate by

$$\widehat{\sigma}_n = \min_{\mathbf{b}}\ \sigma_n(\mathbf{b}).$$

  The functional related to this $S-$estimator is defined by $\boldsymbol{\beta}(F) = \operatorname{argmin}_{\mathbf{b}}\ \sigma(\mathbf{b})$. In Lemma 2.1 will show that the functional is Fisher–consistent.

Let u be a random variable with density (11) and write $\sigma^*(\tau)$ the solution of

$$\mathbb{E}_G\left[\rho\left(\frac{\sqrt{h(u)}}{\sigma^*(\tau)}\right)\right] = b,$$

where $h(u) = 1 - u - \exp(u)$. Note that since u and δ are independent, we have that $\sigma^*(\tau) = \sigma(\tau, \boldsymbol{\beta})$. Similar arguments to those considered in Theorem 5 in Bianco *et al.* (2005) allow to show that under mild conditions $\widetilde{\boldsymbol{\beta}}_n \xrightarrow{a.s.} \boldsymbol{\beta}$ and that $\widehat{\sigma}_n \xrightarrow{a.s.} \sigma^*(\tau)$. As mentioned above $\sigma^*(\tau)$ is a continuous and strictly decreasing function and so, an estimator of τ can be defined as $\widehat{\tau}_n = \sigma^{*-1}(\widehat{\sigma}_n)$ leading to a a strongly consistent estimator for τ.

- **Step 2.** In the second step, we compute $\widehat{\tau}_n = \sigma^{*-1}(\widehat{\sigma}_n)$ and

$$\widehat{c}_n = \max(\widehat{\sigma}_n, C_e(\widehat{\tau}_n)) = \max(\widehat{\sigma}_n, C_e(\sigma^{*-1}(\widehat{\sigma}_n))\,.$$

  We then have that $\widehat{c}_n \xrightarrow{p} c_0 = \max\{\sigma^*(\tau), C_e(\tau)\}$.

- **Step 3.** Let $\widehat{\boldsymbol{\beta}}_{\mathrm{GM},n}$ be the adaptive $GM-$estimator of $\boldsymbol{\beta}$ defined by

$$\widehat{\boldsymbol{\beta}}_{\mathrm{GM},n} = \underset{\mathbf{b}}{\operatorname{argmin}} \sum_{i=1}^{n} \delta_i \phi\left(\sqrt{d^*(y_i, \mathbf{x}_i, \mathbf{b})}/\widehat{c}_n\right) w_1(\mathbf{x}_i), \tag{13}$$

Lemma 2.1. *If score function ϕ is bounded and strictly increasing on the set where it is not constant and if (10) holds, we have that the functional defined as $\boldsymbol{\beta}(F) = \operatorname{argmin}_{\mathbf{b}}\ \sigma(\mathbf{b})$ is Fisher–consistent.*

Propensity score estimators are defined in an analogous way.

3 Consistency results

As mentioned in Heritier and Cantoni (1994), under Fisher–consistency and **N1** to **N5**, standard arguments allow to show that both the simplified and the propensity score estimators

introduced in Section 2.1 are consistent (see Huber, 1981, for instance). For the sake of completeness, we state here these results.

N1 The functions $w_1(\mathbf{x})$ and $w_1(\mathbf{x})\|\mathbf{x}\|$ are bounded.

N2 $\rho(y, u, v)$ is a continuous function.

N3 The class of functions $\mathcal{F} = \{f_{\mathbf{b},\tau}(y, \mathbf{x}, \delta) = \delta\rho\left(y, \mathbf{x}^{\mathrm{T}}\mathbf{b}, t\right) w_1(\mathbf{x})\mathbf{x}, t \in \mathbb{R}, \mathbf{b} \in \mathbb{R}^k\}$ has has finite bracketing number, $N_{[\,]}(\epsilon, \mathcal{F}, L^1(P)) < \infty$, for any $0 < \epsilon < 1$, where P is the distribution of $(y_1, \mathbf{x}_1)$ or that $\log N(\epsilon, \mathcal{F}, L^1(P_n)) = o_P(n)$ with P_n the empirical distribution.

N4 $\inf_{\mathbf{x}\in\mathcal{S}_{w_1}\cap\mathcal{S}_{\mathbf{x}}} p(\mathbf{x}) = A > 0$, where $\mathcal{S}_{w_1}$ and $\mathcal{S}_{\mathbf{x}}$ stand for the support of w_1 and $\mathbf{x}$, respectively.

N5 The estimators $\widehat{p}(\mathbf{x})$ of $p(\mathbf{x})$ satisfy either a) or b)

a) $\sup_{\mathbf{x}\in\mathcal{S}_{w_1}\cap\mathcal{S}_{\mathbf{x}}} |\widehat{p}(\mathbf{x}) - p(\mathbf{x})| \xrightarrow{a.s.} 0$ or

b) $p(\mathbf{x}) = p_{\boldsymbol{\lambda}}(\mathbf{x}) = G_{\mathrm{P}}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\lambda})$ for some continuous function $G_{\mathrm{P}} : \mathbb{R} \to (0, 1]$ with bounded variation and $\widehat{p}(x) = p_{\widehat{\boldsymbol{\lambda}}}(\mathbf{x})$ where $\widehat{\boldsymbol{\lambda}} \xrightarrow{a.s.} \boldsymbol{\lambda}$.

N6. The function $S_{\mathrm{P}}(\mathbf{b}, t, p)$ satisfies the following equicontinuity condition:

a) under **N5**a), for any $\epsilon > 0$ there exists $\delta > 0$ such that for any $t_1, t_2 \in \mathcal{K}$, a compact set in $\mathbb{R}$,

$$|t_1 - t_2| < \delta \Rightarrow \sup_{\mathbf{b}\in\mathbb{R}^k} |S_{\mathrm{P}}(\mathbf{b}, t_1, p) - S_{\mathrm{P}}(\mathbf{b}, t_2, p)| < \epsilon\,.$$

b) under **N5**b), for any $\epsilon > 0$ there exists $\delta > 0$ such that for any $t_1, t_2 \in \mathcal{K}$ and $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 \in \mathcal{K}$, compact sets in $\mathbb{R}$ and $\mathbb{R}^k$, respectively.

$$|t_1 - t_2| < \delta \text{ and } \|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\| < \delta \Rightarrow \sup_{\mathbf{b}\in\mathbb{R}^k} |S_{\mathrm{P}}(\mathbf{b}, t_1, p_{\boldsymbol{\lambda}_1}) - S_{\mathrm{P}}(\mathbf{b}, t_2, p_{\boldsymbol{\lambda}_2})| < \epsilon\,.$$

Proposition 3.1. *Assume that* ***N1*** *to* ***N6*** *hold, then* $\widehat{\boldsymbol{\beta}}_{\mathrm{P}} \xrightarrow{a.s.} \boldsymbol{\beta}$.

The proof of Proposition 3.1 can be found in the Appendix. Using similar arguments, we can obtain the following result.

Proposition 3.2. *Assume that* ***N1****,* ***N2*** *and* ***N3*** *hold, then* $\widehat{\boldsymbol{\beta}} \xrightarrow{a.s.} \boldsymbol{\beta}$.

Remark 3.1. Assumptions **N1** and **N2** are standard requirements since they state that the weight function control large values of the covariates and that the score function bound large residuals, respectively. Note that **N6** holds if $\Psi\left(y, \mathbf{x}^{\mathrm{T}}\mathbf{b}, t\right)$ and $w_1(\mathbf{x})\|\mathbf{x}\|$ are bounded, which holds for the usual functions considered in robustness. Note that if w_1 has compact support, as it is the case for the Tukey weight function, **N4** holds for any continuous missingness probability such that $p(\mathbf{x}) > 0$. This includes, for instance, a logistic model for $p(\mathbf{x})$. On the other hand, if $\mathcal{S}_{\mathbf{x}} = \mathbb{R}^k$ and $w_1 \equiv 1$, i.e., if high leverage points are not downweighted, **N4** restricts the family of missing probabilities to be considered.

4 Monte Carlo Study

4.1 Gamma Regression Model

The gamma regression model considered was

$$y_i|\mathbf{x} \sim \Gamma(\tau, \mu(\mathbf{x})) \quad \text{with} \quad \mu(\mathbf{x}) = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3,\ i = 1, ..., n, \tag{14}$$

with $\tau = 3$, $\beta_1 = \beta_2 = \beta_3 = 0$ and $(x_{1i}, x_{2i}) \sim N(0, \mathbf{I})$. We considered two different values for the shape parameter: $\tau = 1$ and $\tau = 3$. The sample size was $n = 100$ and the number of Monte Carlo replications was $K = 1000$.

To compare the behaviour of the estimators, we consider samples that do not contain outliers and samples contaminated with $\epsilon = 5\%$ outliers. In the contaminated samples, the

outliers are all equal, say $(y_0, \mathbf{x}_0)$. Since the magnitude of the effect of these outliers depends on x_{10} and x_{20} only throughout $x_{10}^2 + x_{20}^2$, without loss of generality they were taken of the form $(y_0, \mathbf{x}_0)$ with $\mathbf{x}_0 = (x_0, 0, 1)$ and $y_0 = \exp(m\ x_0)$. The value m represents the slope of the outliers observations. We chose three values of x_0 corresponding to low leverage outliers with $x_0 = 1$, moderate leverage outliers with $x_0 = 3$ and high leverage outliers with $x_0 = 10$. As values for m we considered $m = 0.5$ and 2.5. These contaminations are denoted C_{m,x_0}. We have also considered an intermediate contamination C_1 by replacing 5 observation by $(y_0, \mathbf{x}_0^\star)$ where $\mathbf{x}_0^\star = (2.5, 2.6, 1)^{\mathrm{T}}$ and $y_0 = \exp(1)$ and an extreme contamination C_2 by replacing 5 observation by $(y_0, \mathbf{x}_0^\star)$ where $\mathbf{x}_0^\star = (4, 4, 1)^{\mathrm{T}}$ and $y_0 = \exp(1)$.

The robust estimators were computed as described in Section 2.4. For the weighted estimators, we used the Tukey's bisquare weight function with tuning constant $c = \chi^2_{k,0.95}$. The weights were computed over the robust Mahalanobis distances based on an $S-$estimator with breakdown point 0.25 using 500 subsamples. From now on, we denote by $\widehat{\boldsymbol{\beta}}_{\mathrm{ML}}$, $\widehat{\boldsymbol{\beta}}_{\mathrm{BGY}}$, $\widehat{\boldsymbol{\beta}}_{\mathrm{TUK}}$, the maximum likelihood estimators, the estimators related to those defined in Bianco *et al.* (2005), i.e., with $w_1 \equiv 1$, and their weighted version with Tukey's weights, respectively. The propensity score estimators will be denoted as $\widehat{\boldsymbol{\beta}}_{\mathrm{P,ML}}$, $\widehat{\boldsymbol{\beta}}_{\mathrm{P,BGY}}$, $\widehat{\boldsymbol{\beta}}_{\mathrm{P,TUK}}$, respectively.

We considered four models for the missing probability

- $p \equiv 1$
- $p \equiv 0.8$, missing completely at random
- $p(\mathbf{x}) = 0.4 + 0.5(\cos(\mathbf{x}^{\mathrm{T}}\boldsymbol{\lambda} + 0.4))^2$ with $\boldsymbol{\lambda} = (2, 2)^{\mathrm{T}}$.
- $p(\mathbf{x}) = 1/(1 + \exp(-\mathbf{x}^{\mathrm{T}}\boldsymbol{\lambda} - 2))$ with $\boldsymbol{\lambda} = (2, 2)^{\mathrm{T}}$, i.e., a logistic model for the missing probability.

Due to the skewness of the distribution of the norm, Figures 18 to **??** give the adjusted boxplots (see Hubert and Vandervieren, 2008) of $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2$. As expected, under C_0 the

classical or robust estimators perform quite similarly, under all the missing schemes. As expected, the simplified methods perform better than the propensity ones when considering a logistic missing probability. For large values of m and/or x_0, the classical estimator is meaningless. On the other hand, the estimators defined in Bianco *et al.* (2005) show their sensitivity to moderate outliers ($m = 0.5$ and $x_0 = 3$) and also for extreme outliers when considering a logistic missingness model. Their weighted version are stable with respect to all the contaminations considered.

4.2 Poisson Regression Model

We consider the Poisson regression model with the canonical link function, that is

$$y_i|\mathbf{x} \sim \mathcal{P}(\mu(\mathbf{x})) \quad \text{with} \quad \log(\mu(\mathbf{x})) = \beta_0 + \beta_1 x_{1i},\ i = 1, ..., n, \tag{15}$$

$x_{1i} \sim N(0, 1)$. The sample size was $n = 100$ and the number of Monte Carlo replications was $K = 500$.

We follow a similar scheme as in Bergesio and Yohai (2010). In order to compare the behaviour of the estimators, we consider samples that without outliers and samples contaminated with $\epsilon = 10\%$ outliers, where the outlying points, (y_0, x_0), are all equal and $x_0 = 2.5$ and $y_0 = 20$, which gives an expected value of marginal expectation of Y equal to 2.718.

We considered four models for the missing probability

- $p \equiv 1$
- $p \equiv 0.8$, missing completely at random
- $p(\mathbf{x}) = 1/(1 + \exp(-2x - 2))$, i.e., a logistic model for the missing probability, which gives a probability of missing equal to 0.999089 at x_0.
- $p(\mathbf{x}) = 0.7 + 0.2(\cos(2x + 0.4))^2$, which gives a probability of missing equal to 0.780567 at x_0.

For the weighted estimators, we used the Tukey's bisquare weight function with tuning constant $c = \chi^2_{1,0.975}$. The weights were computed over the robust Mahalanobis distances based on the median and the MAD. As before, we denote by $\widehat{\boldsymbol{\beta}}_{\mathrm{ML}}$, $\widehat{\boldsymbol{\beta}}_{\mathrm{BGY}}$, $\widehat{\boldsymbol{\beta}}_{\mathrm{TUK}}$, the maximum likelihood estimators, the estimators obtained with $w_1 \equiv 1$, and their weighted version with Tukey's weights, respectively. The propensity score estimators will be denoted as $\widehat{\boldsymbol{\beta}}_{\mathrm{P,ML}}$, $\widehat{\boldsymbol{\beta}}_{\mathrm{P,BGY}}$, $\widehat{\boldsymbol{\beta}}_{\mathrm{P,TUK}}$, respectively.

5 Examples

5.1 Leukemia Data

The data of Feigl and Zelen (1965) present the survivorship of 33 patients of acute myelogenous leukemia divided in two groups, that correspond to a factor variable AG which classifies the patients as positive or negative depending on the presence or absence of a morphological characteristic in the white cells. The original data are time at death and also the white blood cells count WBC, which is a useful tool for diagnosing the initial condition of the patient, indeed higher counts seem to be associated with more severe conditions. Bianco *et al.* (2005) fit, to the complete data set, the model

$$\log(y_i) = \beta_1 WBC_i + \beta_2 AG_i + \beta_3 + u_i,$$

where u_i has $\log \Gamma(\alpha_0, 1)$ distribution through their BGY–estimator. The QQ-plot of the residuals of the BGY-estimate computed by Bianco *et al.* (2005) reveals four clear outliers corresponding to patients with very high values of WBC who survived more than expected.

Denote by $\widehat{\boldsymbol{\beta}}_{\mathrm{ML}}$, $\widehat{\boldsymbol{\beta}}_{\mathrm{BGY}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{TUK}}$ the ML–estimates and the two robust estimates BGY and TUK–estimates computed with all the data, respectively. Besides, $\widehat{\boldsymbol{\beta}}_{\mathrm{ML}}^{-\{4\mathrm{OUT}\}}$ stands for the ML-estimator applied to the sample without the four outliers. Table 1 contains the values of these estimators. In this case, since AG is a factor variable, when computing the

weighted estimators with the Tukey's bisquare function, $\widehat{\boldsymbol{\beta}}_{\mathrm{TUK}}$, the weights $w_1(\mathbf{x})$ were based only on the variable WBC and the tuning constant was chosen as $c = \chi^2_{1,0.95}$. The robust Mahalanobis distance of WBC equals in this case $|WBC_i - \mathrm{median}_i(WBC_i)|/\mathrm{MAD}(WBC_i)$.

To evaluate the performance of the proposed estimators for incomplete data sets, we introduced artificially missing data to this example and we took the above analysis as a natural counterpart. Missing responses among the non–outlying points were introduced at random according to two missing schemes, a completely at random situation with $p(\mathbf{x}) = 0.9$ and a missing at random case with logistic probability of missing $p(\mathbf{x}) = 1/(1+\exp(0.2\,WBC-4))$. In this way, for the logistic case, 8 responses (almost 25% of the data) result in missing observations. The analysis was repeated for each of the obtained samples. In Table 2 we summarize the corresponding results. Different conclusions are derived depending on the missing scheme. As expected, when missing responses occur completely at random, analogous results to those obtained with the complete data set are obtained. On the other hand, for the incomplete sample obtained through a logistic missingness probability, the estimators $\widehat{\boldsymbol{\beta}}_{\mathrm{ML}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{BGY}}$ take similar values. Besides, if the 4 identified outlying observations were removed from this incomplete sample, the three estimators would lead to similar results than those obtained for the situation with no missing responses. These results show the advantage of introducing weights as a useful tool to prevent from outlying points even under different missing schemes.

5.2 Hospital Costs Data

Marazzi and Yohai (2004) introduced a data set that corresponds to the costs of 100 patients in a Swiss hospital in 1999 for *medical back problems*. They concerned on the relationship between the hospital cost of stay, y, (Cost, in Swiss francs) and some administrative explanatory variables:

	Estimated Coefficients			
	$\widehat{\boldsymbol{\beta}}_{\mathrm{ML}}$	$\widehat{\boldsymbol{\beta}}_{\mathrm{ML}}^{-\{4\mathrm{out}\}}$	$\widehat{\boldsymbol{\beta}}_{\mathrm{BGY}}$	$\widehat{\boldsymbol{\beta}}_{\mathrm{TUK}}$
$\frac{WBC}{1000}$	-0.007	-0.051	-0.051	-0.0895
AG	-1.101	-1.574	-1.802	-1.5096
$Intercept$	4.227	4.795	4.849	5.1007

Table 1: Analysis of Feigl & Zelen data. Complete data set.

	Estimated Coefficients		
	$\widehat{\boldsymbol{\beta}}_{\mathrm{ML}}$	$\widehat{\boldsymbol{\beta}}_{\mathrm{BGY}}$	$\widehat{\boldsymbol{\beta}}_{\mathrm{TUK}}$
	$p(\mathbf{x}) = 0.9$		
$\frac{WBC}{1000}$	-0.008	-0.050	-0.0842
AG	-0.974	-1.469	-1.3642
$Intercept$	4.333	4.841	5.0547
	$p(\mathbf{x}) = 1/(1+\exp(0.2\,WBC - 4))$		
$\frac{WBC}{1000}$	-0.0012	-0.0004	-0.1210
AG	-1.3718	-1.4371	-1.4315
$Intercept$	4.4432	4.5055	5.2617

Table 2: Analysis of Feigl & Zelen data with missing responses.

Variable	Description
LOS	length of stay in days
ADM	admission type (0 = planned; 1 = emergency)
INS	insurance type (0 = regular; 1 = private)
AGE	years
SEX	0 = female 1 = male
$DEST$	discharge destination (1 = home; 0 = other)

Table 3: Explanatory Variables for Hospital Costs Data.

Cantoni and Ronchetti (2006) fitted to the complete data set the model

$$\log(y_i) = \beta_1 \log LOS_i + \beta_2 ADM_i + \beta_3 INS + \beta_4 AGE + \beta_5 DEST + \beta_6 + u_i,$$

where u_i has $\log \Gamma(\alpha_0, 1)$. Using their robust proposal, they identified 5 outliers corresponding to observations labelled as $14, 21, 28, 44$ and 63, whose weights are less or equal than 0.5. They realized that the atypical points affected the classical estimates of the coefficient of variable INS and the shape parameter. In particular, the effect of the outliers on the shape parameter is remarkable since it achieved almost half the value obtained with the robust method.

As in the previous example, we compute $\widehat{\boldsymbol{\beta}}_{\mathrm{ML}}$, $\widehat{\boldsymbol{\beta}}_{\mathrm{BGY}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{TUK}}$ to the complete data set and the maximum likelihood estimator without the 5 outlying observations, $\widehat{\boldsymbol{\beta}}_{\mathrm{ML}}^{-\{5\mathrm{out}\}}$ and the corresponding estimators of the shape parameter τ. Table 4 summarized the obtained estimators. The obtained results are analogous to those obtained by Cantoni and Ronchetti (2006). Moreover, the value of $\widehat{\boldsymbol{\beta}}_{\mathrm{ML}}^{-\{5\mathrm{out}\}}$ and the related estimator of τ are very similar to those obtained with our robust proposal, showing its good performance in presence of outliers. and those obtained using $\widehat{\boldsymbol{\beta}}_{\mathrm{TUK}}$. In Figure 1 we show on the left the residuals corresponding to the fit obtained with the robust estimator in Cantoni and Ronchetti (2006) and on the right those based on $\widehat{\boldsymbol{\beta}}_{\mathrm{TUK}}$. On both plots we identify the outliers detected by

each of the methods, in the case of $\widehat{\boldsymbol{\beta}}_{\mathrm{TUK}}$ we label those residuals with absolute value greater than 0.40.

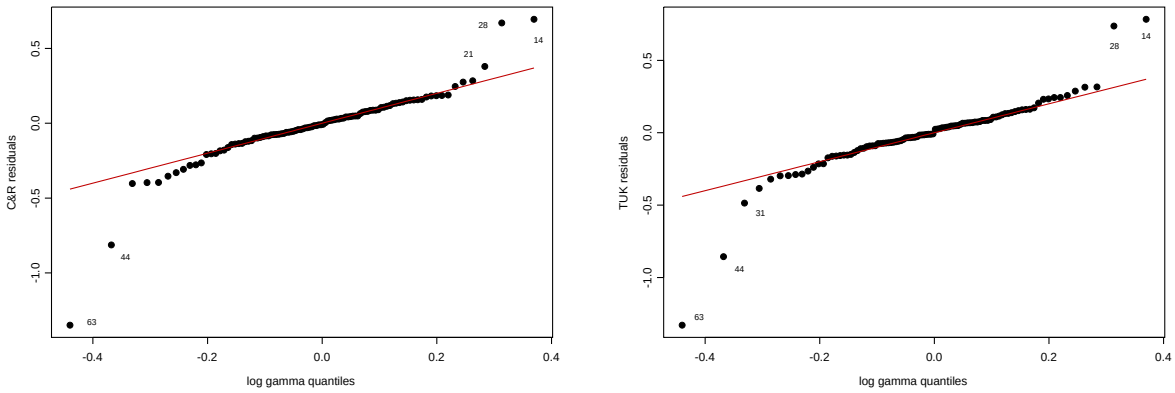

Figure 1: On the left Log–gamma residuals based on the robust estimator in Cantoni and Ronchetti (2006) and on the right those based on $\widehat{\boldsymbol{\beta}}_{\mathrm{TUK}}$

In a first stage, completely missing at random responses among the non atypical observations detected by Cantoni and Ronchetti (2006) were introduced with a probability $p(\mathbf{x}) = 0.85$ and then we repeat the study. In Table 5 we report the obtained results. It is worth noticing that both robust estimators remain very stable and very close to the values obtained with $\widehat{\boldsymbol{\beta}}_{\mathrm{ML}}^{-\{5\mathrm{out}\}}$, which we take as a natural counterpart. Besides, the estimator of τ obtained from $\widehat{\boldsymbol{\beta}}_{\mathrm{TUK}}$ is almost the same to the one obtained with the complete sample, while the estimate computed from $\widehat{\boldsymbol{\beta}}_{\mathrm{BGY}}$ changes and decreases to this value when missing observations are introduced.

In a second stage, a missing at random mechanism based on the logistic probability $p(\mathbf{x}) = 1/(1 + \exp(0.2 \log LOS - 2))$ was implemented in the same way and it results in about 20% of missing responses. Again, the study was repeated and we show in Table 5 the obtained results. Comparing with the counterpart estimates based on $\widehat{\boldsymbol{\beta}}_{\mathrm{ML}}^{-\{5\mathrm{out}\}}$, we observe

	Estimated Coefficients			
	$\widehat{\boldsymbol{\beta}}_{\mathrm{ML}}$	$\widehat{\boldsymbol{\beta}}_{\mathrm{ML}}^{-\{5\mathrm{out}\}}$	$\widehat{\boldsymbol{\beta}}_{\mathrm{BGY}}$	$\widehat{\boldsymbol{\beta}}_{\mathrm{TUK}}$
$\log LOS$	0.8218	0.8473	0.8640	0.8892
ADM	0.2132	0.2151	0.2576	0.2375
INS	0.0960	-0.0235	-0.0523	-0.0437
AGE	-0.0005	-0.0015	-0.0009	-0.0010
SEX	0.0954	0.0706	0.0489	0.0739
$DEST$	-0.1040	-0.1413	-0.1024	-0.1225
$Intercept$	7.2331	7.2764	7.1796	7.1268
$\widehat{\tau}$	20.1876	44.2838	48.9791	41.1086

Table 4: Analysis of Hospital Costs data.

that all the estimates based on $\widehat{\boldsymbol{\beta}}_{\mathrm{TUK}}$ remain very stable and close to these values, while the maximum likelihood estimates and the estimates based on $\widehat{\boldsymbol{\beta}}_{\mathrm{BGY}}$ are farther away; in particular, the estimators of the coefficient of INS and τ.

As in the previous example, we can see the benefits of introducing weights in order to avoid the effect of high leverage outlying data in the presence of missing responses.

6 Concluding Remarks

We have introduced resistant estimators for the regression parameter under a generalized regression model, when there are missing observations in the response variable and it can be suspected that anomalous observations are present in the sample. The estimators considered are Fisher–consistent and thus, lead to strongly consistent estimators.

The simulation study confirms the expected inadequate behaviour of the classical estimators and the sensitivity of the unweighted robust estimators in the presence of mild outliers.

	Estimated Coefficients		
	$\widehat{\boldsymbol{\beta}}_{\mathrm{ML}}$	$\widehat{\boldsymbol{\beta}}_{\mathrm{BGY}}$	$\widehat{\boldsymbol{\beta}}_{\mathrm{TUK}}$
	$p(\mathbf{x}) = 0.85$		
$\log LOS$	0.8168	0.8287	0.8528
ADM	0.2216	0.2473	0.2271
INS	0.1025	-0.0238	-0.0206
AGE	-0.0002	-0.0004	-0.0005
SEX	0.1078	0.0716	0.0862
$DEST$	-0.1004	-0.1247	-0.1374
$Intercept$	7.2089	7.2342	7.1915
$\widehat{\tau}$	18.9617	41.9101	41.1490
	$p(\mathbf{x}) = 1/(1+\exp(0.2\,\log LOS - 2))$		
$\log LOS$	0.8048	0.8194	0.8500
ADM	0.1891	0.2087	0.1852
INS	0.1487	-0.0021	-0.0106
AGE	-0.0010	-0.0014	-0.0012
SEX	0.0905	0.0711	0.0906
$DEST$	-0.0998	-0.1214	-0.1518
$Intercept$	7.3009	7.3259	7.2672
$\widehat{\tau}$	17.3636	35.2136	41.8678

Table 5: Analysis of Hospital Costs data with missing responses.

The proposed robust procedures for the regression parameter perform quite similarly under the central model or under the contaminations studied.

7 Appendix

PROOF OF LEMMA 2.1. Note that if we define $\sigma(\tau, \mathbf{b})$ as the $M-$scale functional solution of

$$\frac{\mathbb{E}_F\left(\delta\phi\left(\frac{\sqrt{d^*(y,\mathbf{x},\mathbf{b})}}{\sigma(\tau,\mathbf{b})}\right)\right)}{\mathbb{E}(\delta)} = b,$$

using the independence between δ and u, we have that

$$\frac{\mathbb{E}_F\left(\delta\phi\left(\frac{\sqrt{d^*(y,\mathbf{x},\mathbf{b})}}{\sigma(\tau,\mathbf{b})}\right)\right)}{\mathbb{E}(\delta)} = \frac{\mathbb{E}_F\left(p(\mathbf{x})\phi\left(\frac{\sqrt{\widetilde{d}(u,\mathbf{x},\boldsymbol{\beta}-\mathbf{b})}}{\sigma(\tau,\mathbf{b})}\right)\right)}{\mathbb{E}(p(\mathbf{x}))}.$$

Note that $\sigma(\tau, \mathbf{b})$is a function of $\boldsymbol{\beta} - \mathbf{b}$. Using Lemma 1 in Bianco *et al.* (2005), we get that for any fixed c

$$\begin{aligned}\mathbb{E}_F\left(\delta\phi\left(\frac{\sqrt{\widetilde{d}(u,\mathbf{x},\boldsymbol{\beta}-\mathbf{b}))}}{c}\right)\Big|\mathbf{x}\right) &= p(\mathbf{x})\mathbb{E}_F\left(\phi\left(\frac{\sqrt{\widetilde{d}(u,\mathbf{x},\boldsymbol{\beta}-\mathbf{b}))}}{c}\right)\Big|\mathbf{x}\right) \\ &\geq \mathbb{E}_F\left(\delta\phi\left(\frac{\sqrt{\widetilde{d}(u,\mathbf{x},\mathbf{0})}}{c}\right)\Big|\mathbf{x}\right) \qquad (16)\end{aligned}$$

Using (16), we get that for any $\mathbf{b} \neq \boldsymbol{\beta}$

$$\mathbb{E}_F\left(\delta\phi\left(\frac{\sqrt{\widetilde{d}(u,\mathbf{x},\mathbf{0})}}{\sigma(\tau,\mathbf{b})}\right)\Big|\mathbf{x}\right) < \mathbb{E}_F\left(\delta\phi\left(\frac{\sqrt{\widetilde{d}(u,\mathbf{x},\boldsymbol{\beta}-\mathbf{b}))}}{\sigma(\tau,\mathbf{b})}\right)\Big|\mathbf{x}\right) = b\,\mathbb{E}(p(\mathbf{x}))$$

Using that $\mathbb{E}_F\left(\delta\phi\left(\sqrt{\widetilde{d}(u,\mathbf{x},\mathbf{0})}/\sigma\right)\right)$ is decreasing in σ and that $\mathbb{E}_F\left(\delta\phi\left(\sqrt{\widetilde{d}(u,\mathbf{x},0)}/\sigma(\tau,\boldsymbol{\beta})\right)\right) = b\,\mathbb{E}(p(\mathbf{x}))$, we get that $\sigma(\tau,\boldsymbol{\beta}) < \sigma(\tau,\mathbf{b})$, which implies the Fisher–consistency of the functional. □

PROOF OF PROPOSITION 3.1. We will show that $S_{\mathrm{P},n}(\widehat{\boldsymbol{\beta}},\widehat{\tau},\widehat{p}) - S_{\mathrm{P},n}(\widehat{\boldsymbol{\beta}},\tau,p) \xrightarrow{a.s.} 0$. Note that $\mathbb{E}\left(S_{\mathrm{P},n}(\mathbf{b},t,q)\right) = S_{\mathrm{P}}(\mathbf{b},t,q)$

Let us first assume that **N5**b) holds then, using standard empirical process arguments, from **N3**, we have that

$$V_n = \sup_{\mathbf{b},t,\boldsymbol{\lambda}} \left| \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i}{G_{\mathrm{P}}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\lambda})}\rho(y_i,\mathbf{x}_i^{\mathrm{T}}\mathbf{b},t)w_1(\mathbf{x}_i) - \mathbb{E}\left(\frac{\delta_i}{G_{\mathrm{P}}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\lambda})}\rho(y_i,\mathbf{x}_i^{\mathrm{T}}\mathbf{b},t)\right) w_1(\mathbf{x}_i) \right| \xrightarrow{a.s.} 0 \ .$$

Therefore, since $\widehat{p}(\mathbf{x}) = p_{\boldsymbol{\lambda}}(\mathbf{x}) = G_{\mathrm{P}}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\lambda})$, we get

$$\begin{aligned}
\sup_{\mathbf{b}} |S_{\mathrm{P},n}(\mathbf{b},\widehat{\tau},\widehat{p}) - S_{\mathrm{P},n}(\mathbf{b},\tau,p)| &\leq \sup_{\mathbf{b}} |S_{\mathrm{P},n}(\mathbf{b},\widehat{\tau},\widehat{p}) - S_{\mathrm{P}}(\mathbf{b},\widehat{\tau},\widehat{p})| + \sup_{\mathbf{b}} |S_{\mathrm{P}}(\mathbf{b},\widehat{\tau},\widehat{p}) - S_{\mathrm{P}}(\mathbf{b},\tau,p)| \\
&+ \sup_{\mathbf{b}} |S_{\mathrm{P},n}(\mathbf{b},\tau,p) - S_{\mathrm{P}}(\mathbf{b},\tau,p)| \\
&\leq 2V_n + \sup_{\mathbf{b}} |S_{\mathrm{P}}(\mathbf{b},\widehat{\tau},\widehat{p}) - S_{\mathrm{P}}(\mathbf{b},\tau,p)|
\end{aligned}$$

Using the equicontinuity of $S_{\mathrm{P}}(\mathbf{b},\tau,p)$ and the consistency of $\widehat{\tau}$ and $\widehat{\boldsymbol{\lambda}}$, we get that, when **N5**b) holds, $\sup_{\mathbf{b}} |S_{\mathrm{P},n}(\mathbf{b},\widehat{\tau},\widehat{p}) - S_{\mathrm{P},n}(\mathbf{b},\tau,p)| \xrightarrow{a.s.} 0$.

Under **N5**a), we obtain easily from **N1** and **N2** that $S_{\mathrm{P},n}(\widehat{\boldsymbol{\beta}},\widehat{\tau},\widehat{p}) - S_{\mathrm{P},n}(\widehat{\boldsymbol{\beta}},\widehat{\tau},p) \xrightarrow{a.s.} 0$. Again, using standard empirical process arguments, from **N3**, we have that

$$V_n = \sup_{\mathbf{b},t} \left| \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i}{p(\mathbf{x}_i)}\rho(y_i,\mathbf{x}_i^{\mathrm{T}}\mathbf{b},t)w_1(\mathbf{x}_i) - \mathbb{E}\left(\frac{\delta_i}{p(\mathbf{x}_i)}\rho(y_i,\mathbf{x}_i^{\mathrm{T}}\mathbf{b},t)\right) w_1(\mathbf{x}_i) \right| \xrightarrow{a.s.} 0 \ ,$$

which implies that

$$\begin{aligned}
\sup_{\mathbf{b}} |S_{\mathrm{P},n}(\mathbf{b},\widehat{\tau},p) - S_{\mathrm{P},n}(\mathbf{b},\tau,p)| &\leq \sup_{\mathbf{b}} |S_{\mathrm{P},n}(\mathbf{b},\widehat{\tau},p) - S_{\mathrm{P}}(\mathbf{b},\widehat{\tau},p)| + \sup_{\mathbf{b}} |S_{\mathrm{P}}(\mathbf{b},\widehat{\tau},p) - S_{\mathrm{P}}(\mathbf{b},\tau,p)| \\
&+ \sup_{\mathbf{b}} |S_{\mathrm{P},n}(\mathbf{b},\tau,p) - S_{\mathrm{P}}(\mathbf{b},\tau,p)| \\
&\leq 2V_n + \sup_{\mathbf{b}} |S_{\mathrm{P}}(\mathbf{b},\widehat{\tau},p) - S_{\mathrm{P}}(\mathbf{b},\tau,p)|
\end{aligned}$$

and so, using the consistency of $\widehat{\tau}$ and the equicontinuity of $S_{\mathrm{P}}(\mathbf{b},\tau,p)$, we obtain that, when **N5**a) holds, $\sup_{\mathbf{b}} |S_{\mathrm{P},n}(\mathbf{b},\widehat{\tau},\widehat{p}) - S_{\mathrm{P},n}(\mathbf{b},\tau,p)| \xrightarrow{a.s.} 0$.

Therefore, $S_{\mathrm{P},n}(\widehat{\boldsymbol{\beta}},\widehat{\tau},\widehat{p}) - S_{\mathrm{P},n}(\widehat{\boldsymbol{\beta}},\tau,p) \xrightarrow{a.s.} 0$ and so the sequence of estimators $\widehat{\boldsymbol{\beta}}$ satisfies that $\inf_{\mathbf{b}} S_{\mathrm{P},n}(\mathbf{b},\tau,p) - S_{\mathrm{P},n}(\widehat{\boldsymbol{\beta}},\tau,p) \xrightarrow{a.s.} 0$ and so, the results from Huber (1967) can be applied. □

Acknowledgement. This work began while Isabel Rodrigues was visiting the Instituto de Cálculo from the Universidad de Buenos Aires partially supported by the *Fundação para a Ciência e a Tecnologia* (FCT), the *Center for Mathematics and its Applications* (CEMAT) and *Fundação Calouste Gulbenkian* (FCG). This research was also partially supported by Grants PIP 112-200801-00216 from CONICET, PICT 0821 from ANPCYT and X-018 from the Universidad de Buenos Aires at Buenos Aires, Argentina and by the joint cooperation program ANPCYT–GRICES PO09/05.

References

[1] Bianco, A., García Ben, M. and Yohai, V. (2005). Robust estimation for linear regression with asymmetric errors. *Canad. J. Statist.* **33**, 511-528.

[2] Bianco, A. and Martínez, E. (2009). Robust testing in the logistic regression model. *Comp. Statist. Data Anal.* **53**, 4095-4105.

[3] Bianco, A. and Yohai, V. (1996). Robust estimation in the logistic regression model. *Lecture Notes in Statistics*, **109**, 17-34. Springer-Verlag, New York.

[4] Boente G., González–Manteiga W. and Pérez–González A. (2009). Robust nonparametric estimation with missing data. *J. Statist. Plann. Inference*, **139**: 571-592.

[5] Cantoni, E. and Ronchetti, E. (2001). Robust inference for generalized linear models. *J. Amer. Statist. Assoc.* **96**, 1022-1030.

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

[7] Croux, C. and Haesbroeck, G. (2003). Implementing the Bianco and Yohai estimator for logistic regression. *Comp. Statist. Data Anal.*, **44**, 273-295.

[8] Feigl, P. and Zelen, M. (1965). Estimation of exponential survival probabilities with concomitant information. *Biometrics* **21**, 826838.

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

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

[11] Hubert, M. and Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Comp. Statist. Data Anal.*, **52**, 51865201.

[12] McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. (2nd ed.) London: Champman and Hall.

[13] Marazzi A. and Yohai V. J. (2004). Adaptively truncated maximum likelihood regression with asymmetric errors. *Journal of Statistical Planning and Inference*, **122**, 271-291.

[14] Maronna R., Martin D. and Yohai V. (2006). *Robust statistics: Theory and methods* , Wiley, New York.

[15] McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. (2nd ed.) London: Champman and Hall.

[16] Nelder, J. and Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A*, **135**, 370384.

[17] Neyman, J. (1938). Contribution to the theory of sampling human populations. *J. Amer. Statist. Assoc.* **33** 101-116.

[18] Heritier, S. and Ronchetti, E. (1994). Robust bounded-influence tests in general parametric models. *J. Amer. Statist. Assoc.* **89**, 897-904.

[19] Heritier, S., Cantoni, E., Copt, S. and Victoria–Feser, M.P. (2009). *Robust Methods in Biostatistics*. Wiley Series in Probability and Statistics. Wiley.

[20] Stefanski, L., Carroll,R. and Ruppert, D. (1986). Bounded score functions for generalized linear models. *Biometrika* **73**, 413-424.

[21] Van der Vaart, A. (2000). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press

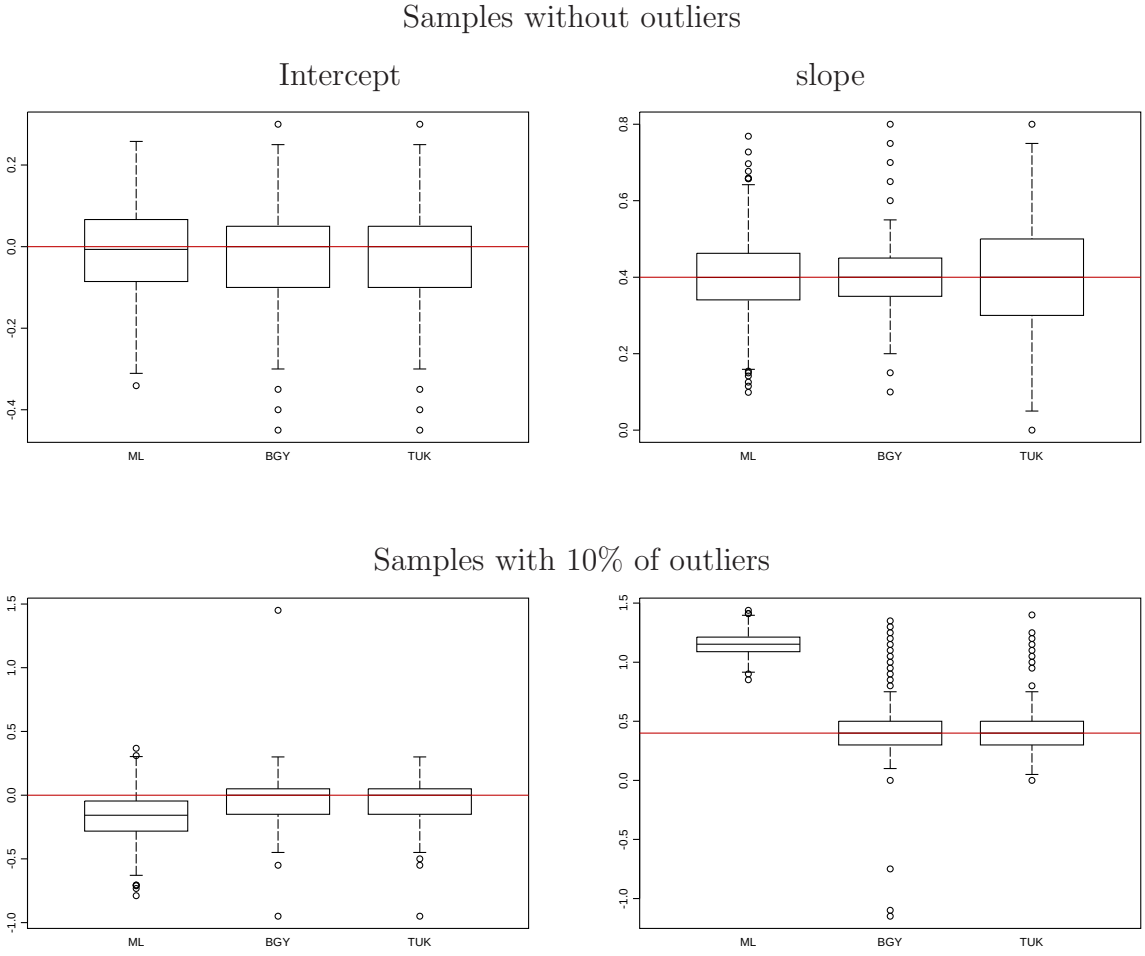

Figure 2: Boxplots for complete samples: $p \equiv 1$

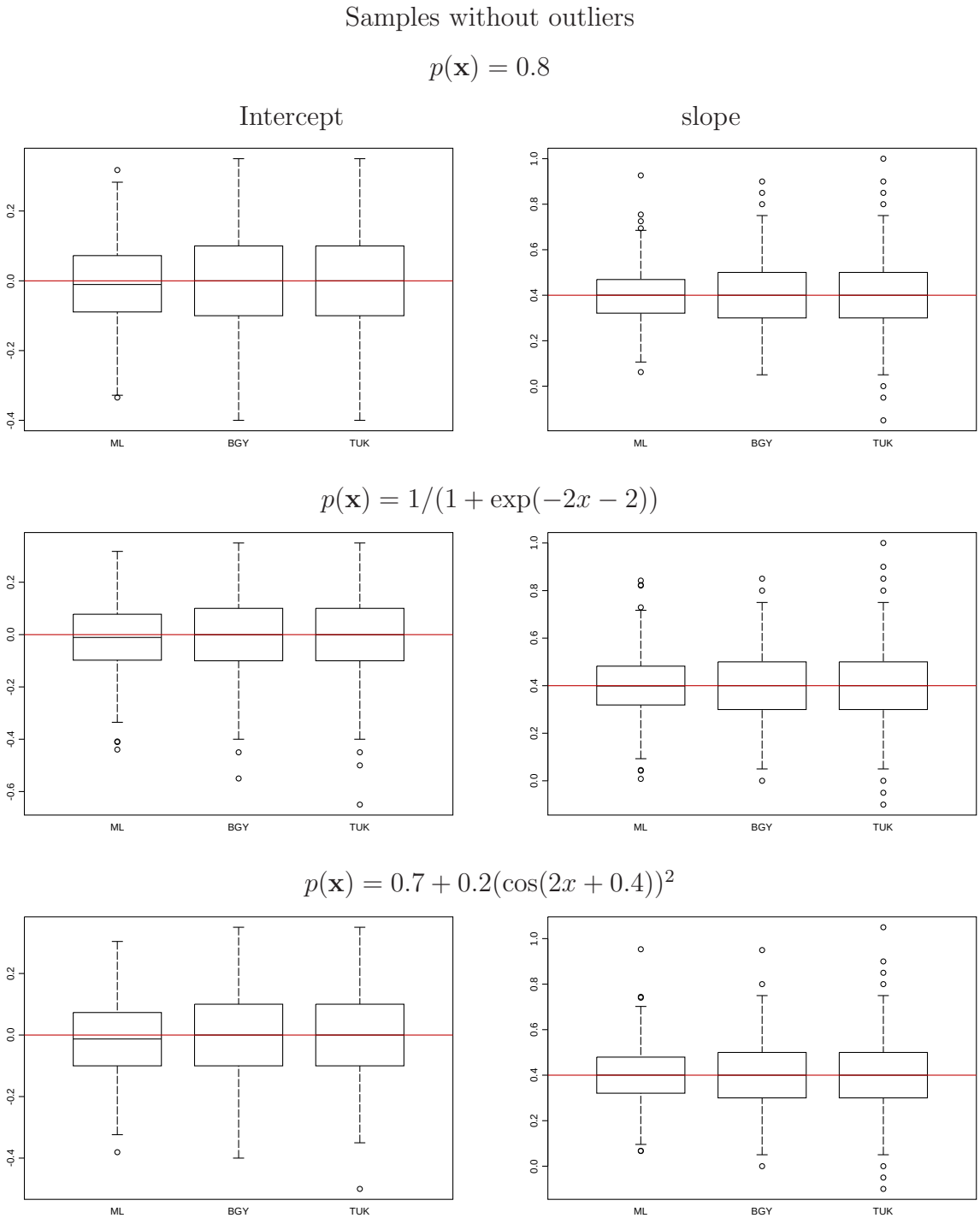

Figure 3: Boxplots for uncontaminated samples for different missing patterns: simplified estimators

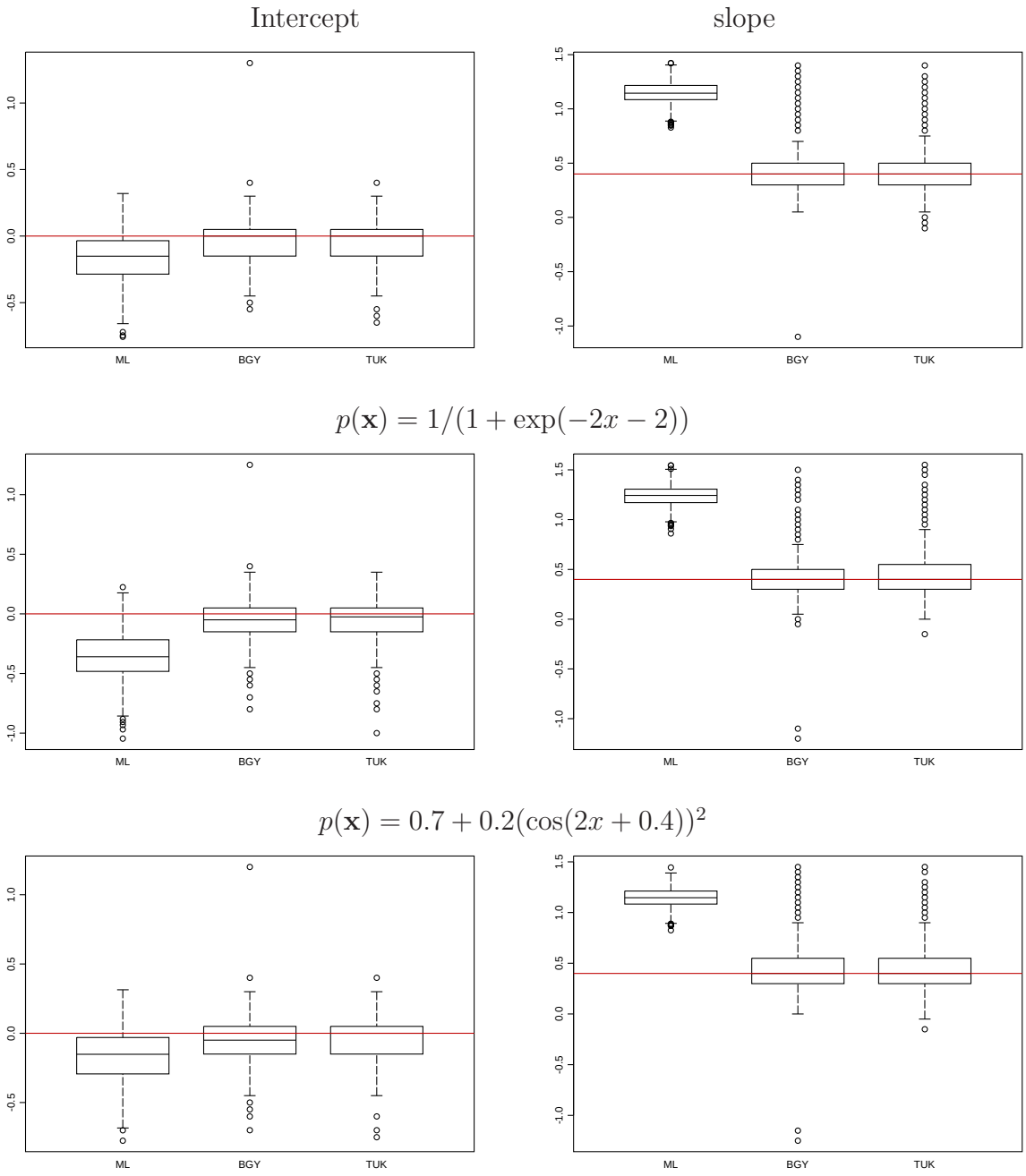

Figure 4: Boxplots for contaminated samples for different missing patterns: simplified estimators

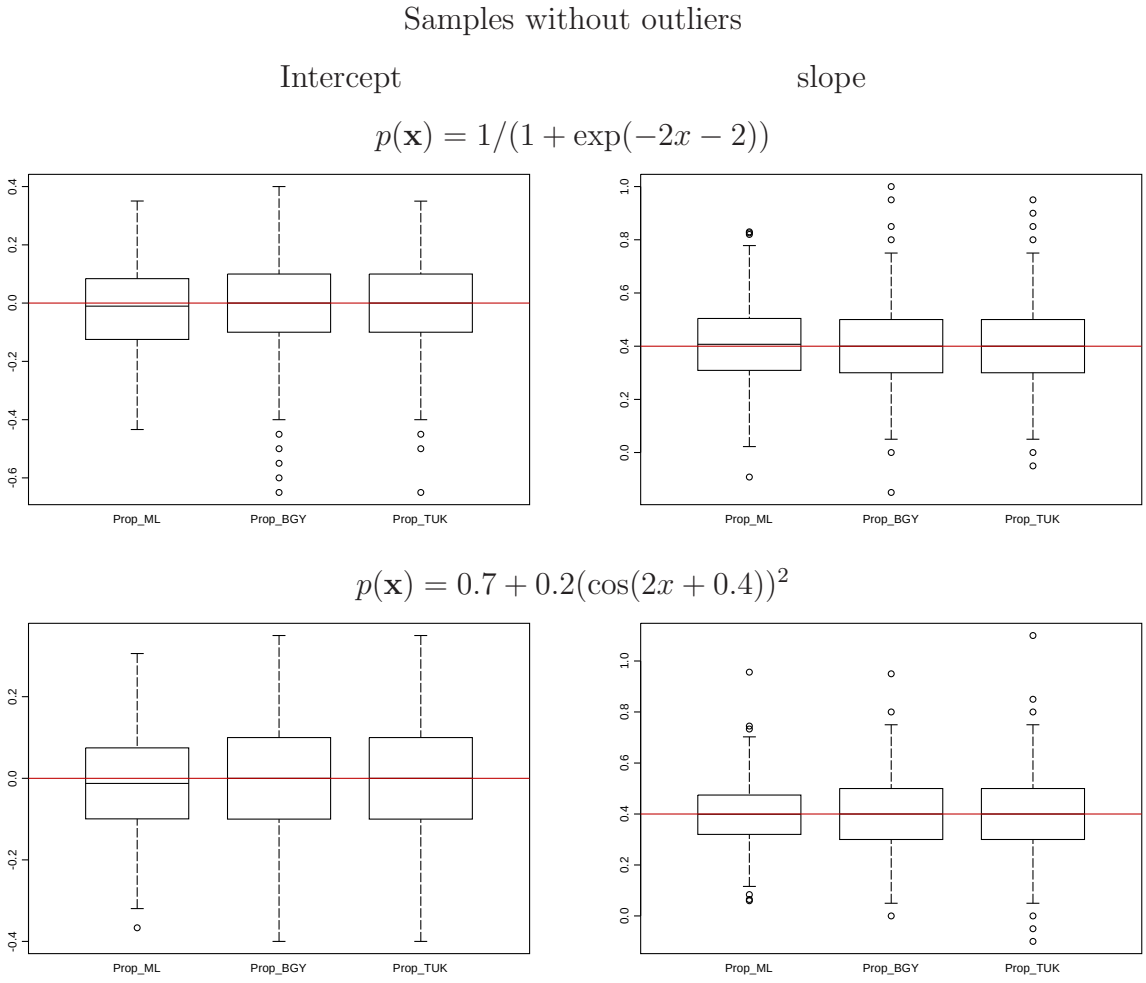

Figure 5: Boxplots for uncontaminated samples for different missing patterns: propensity estimators

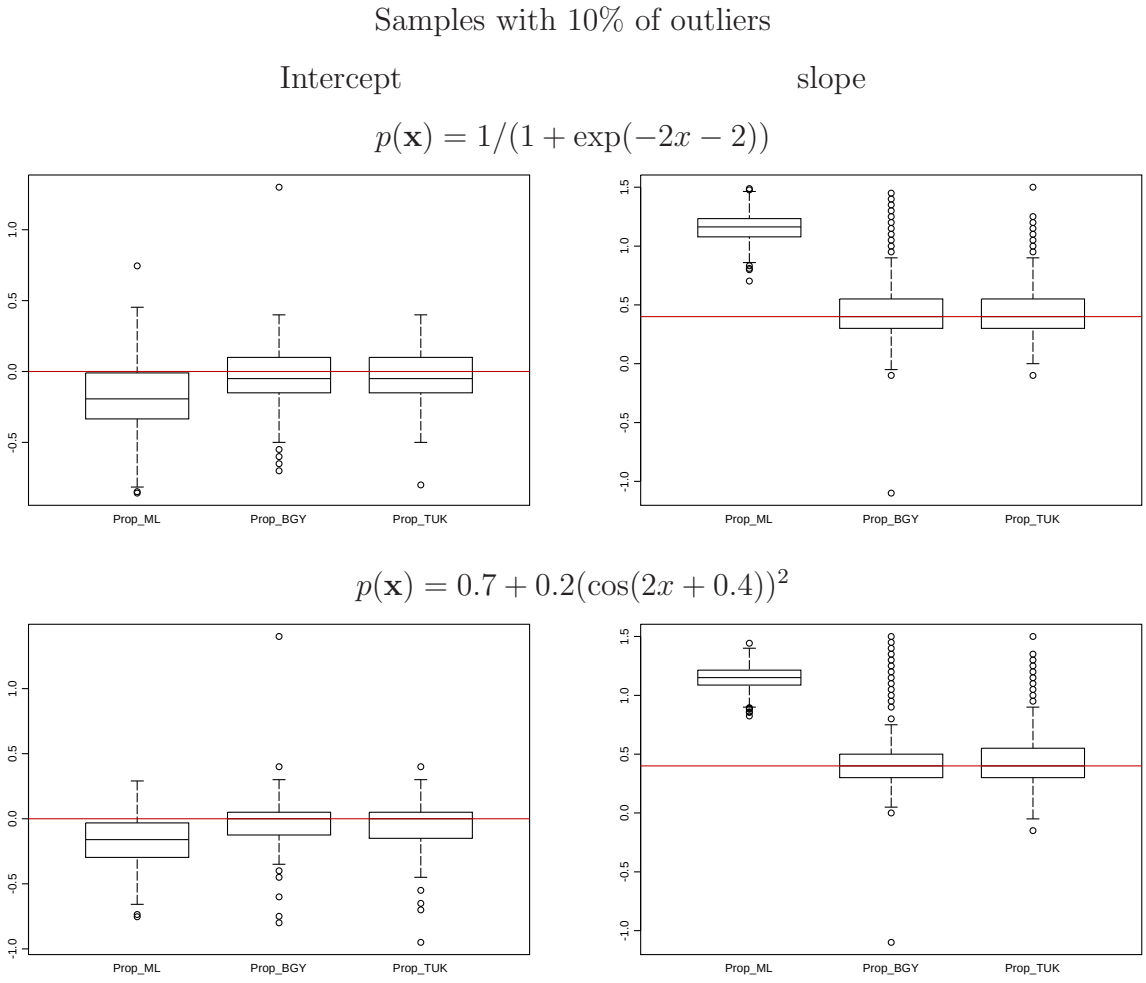

Figure 6: Boxplots for contaminated samples for different missing patterns: propensity estimators

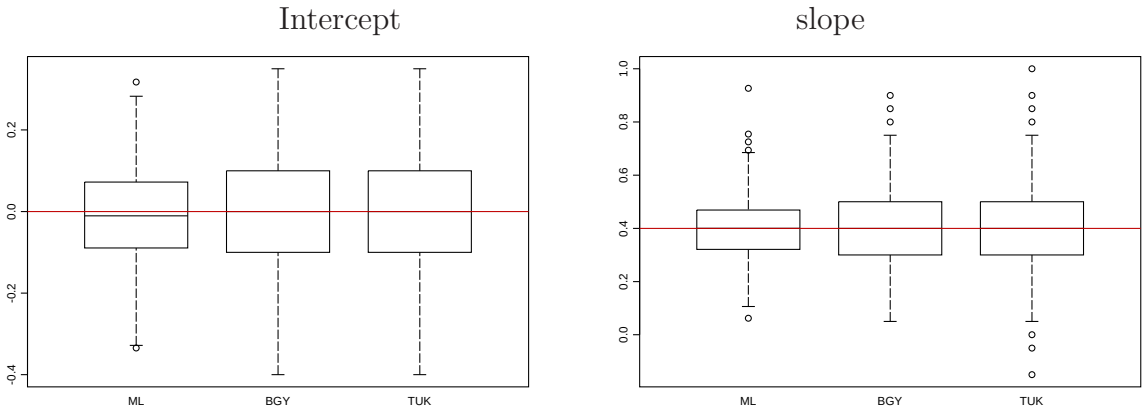

Figure 7: Boxplots $p \equiv 0.8$ without of outliers

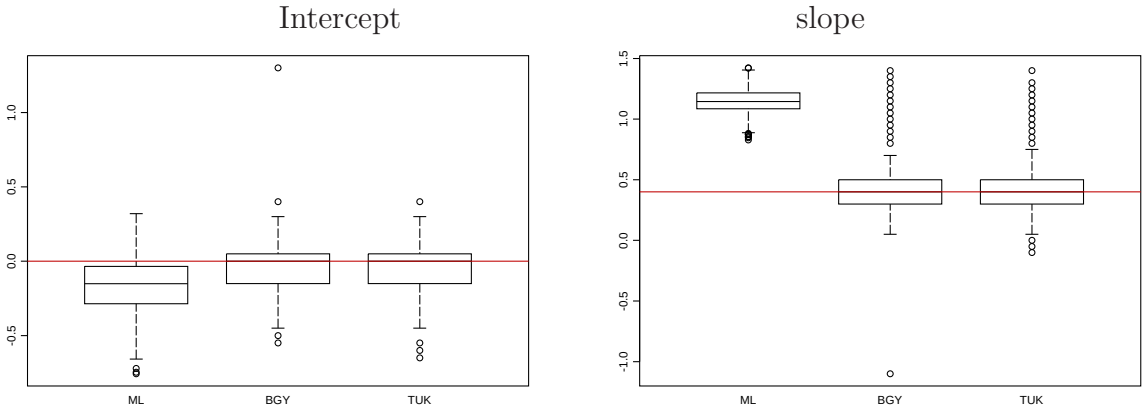

Figure 8: Boxplots $p \equiv 0.8$ with 10% of outliers

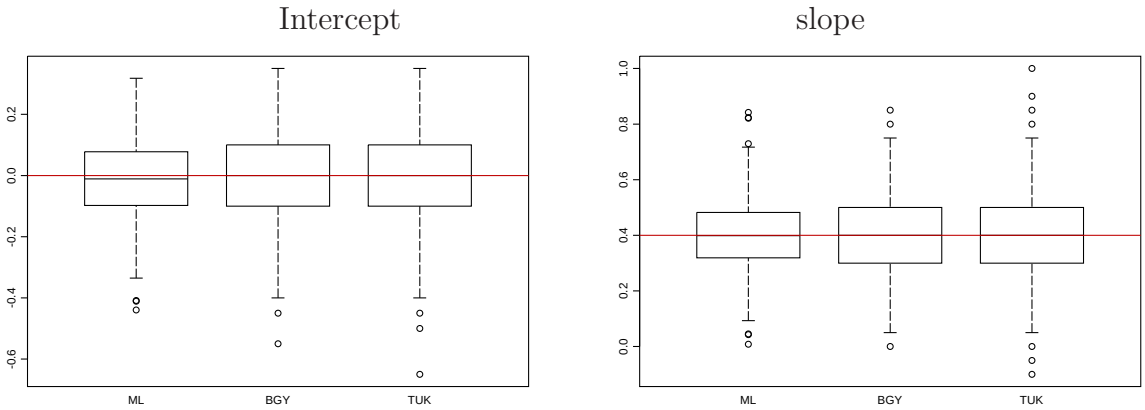

Figure 9: Boxplots $p(\mathbf{x}) = 1/(1 + \exp(-2x - 2))$ without of outliers

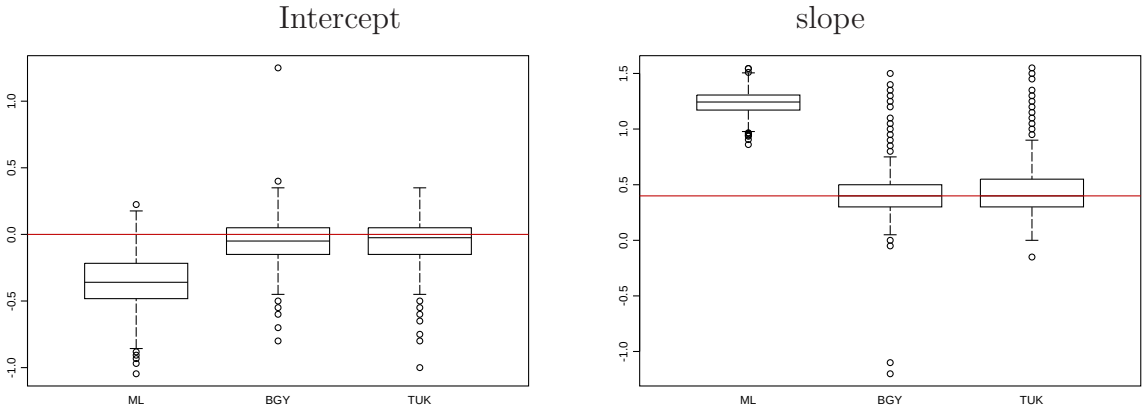

Figure 10: Boxplots $p(\mathbf{x}) = 1/(1 + \exp(-2x - 2))$ with 10% of outliers

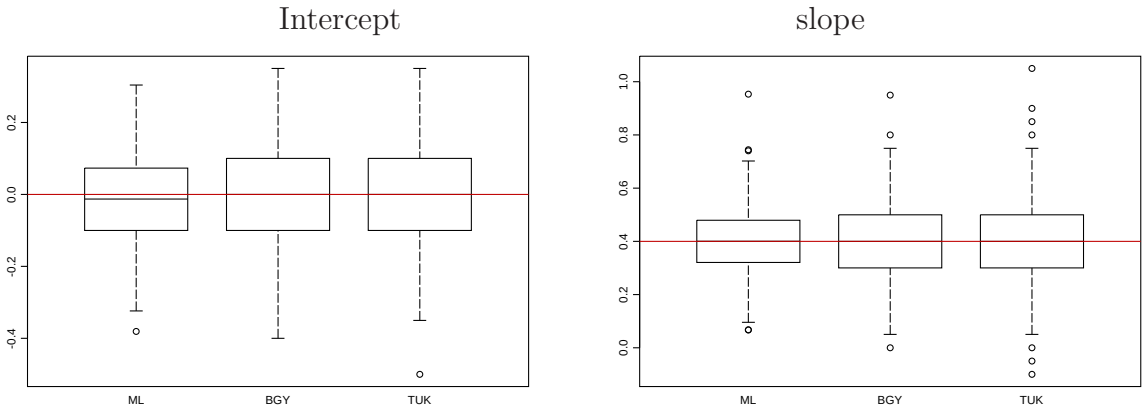

Figure 11: Boxplots $p(\mathbf{x}) = 0.7 + 0.2(\cos(2x + 0.4))^2$ without of outliers

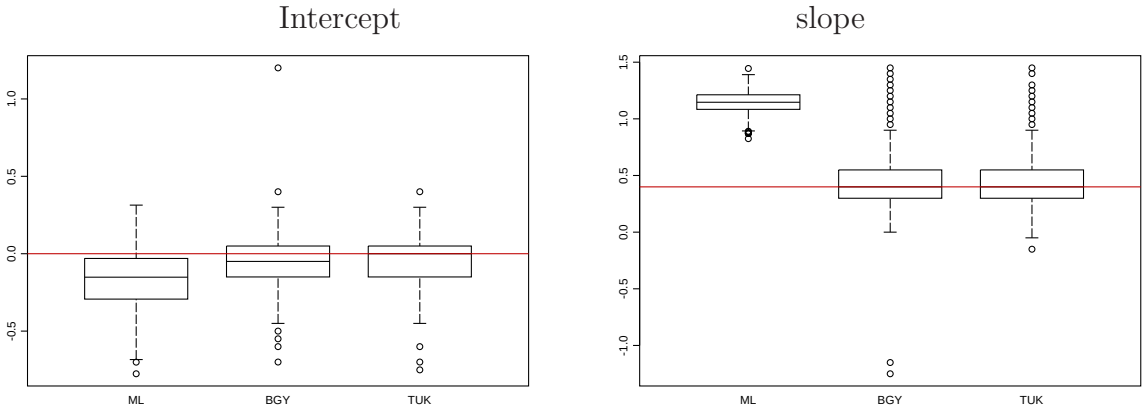

Figure 12: Boxplots $p(\mathbf{x}) = 0.7 + 0.2(\cos(2x + 0.4))^2$ with 10% of outliers

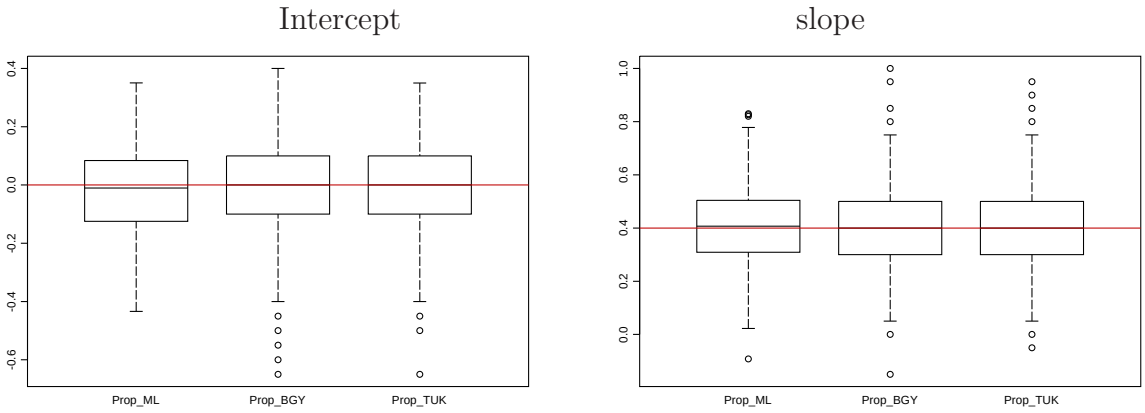

Figure 13: Boxplots $p(\mathbf{x}) = 1/(1 + \exp(-2x - 2))$ without of outliers

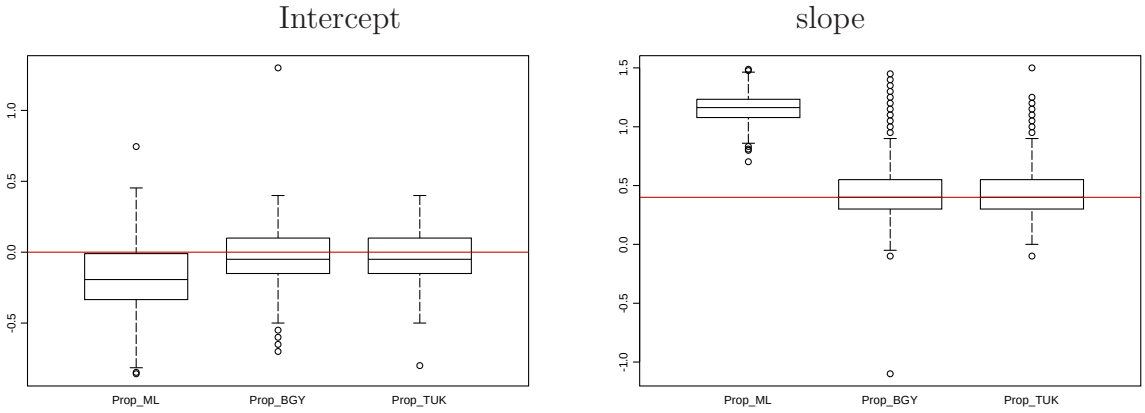

Figure 14: Boxplots $p(\mathbf{x}) = 1/(1 + \exp(-2x - 2))$ with 10% of outliers

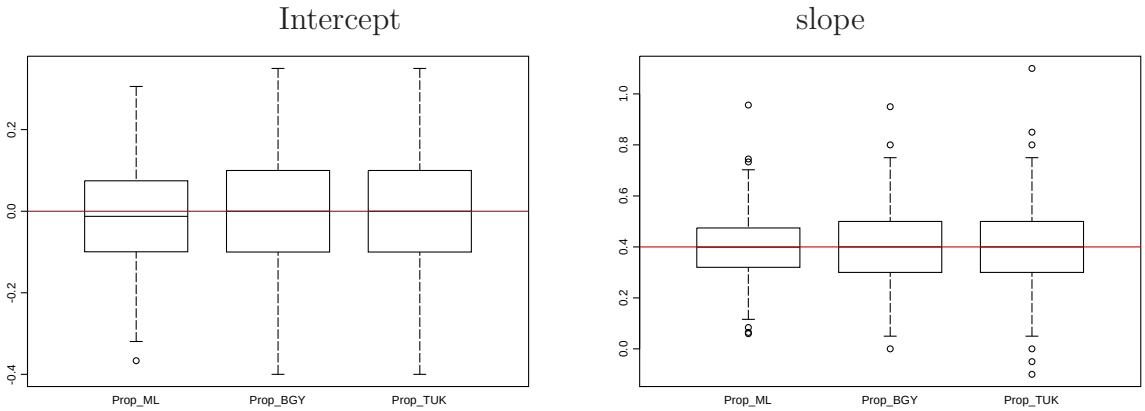

Figure 15: Boxplots $p(\mathbf{x}) = 0.7 + 0.2(\cos(2x + 0.4))^2$ without of outliers

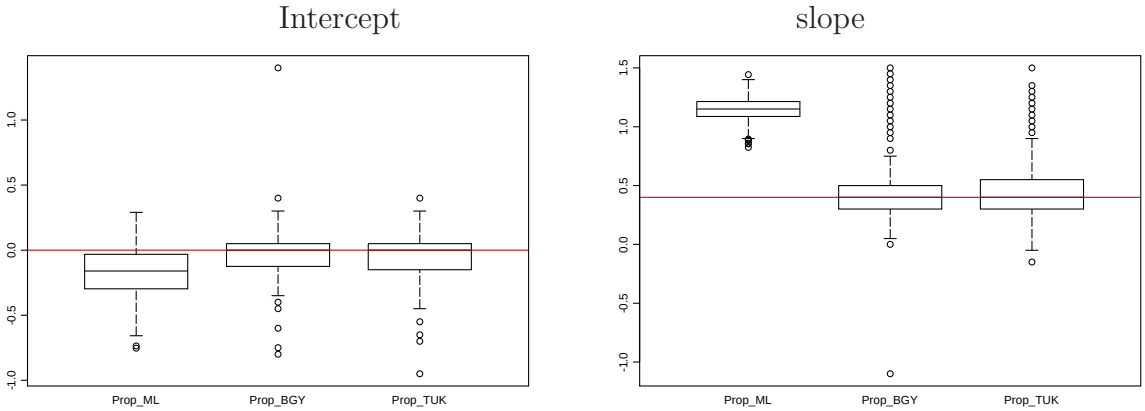

Figure 16: Boxplots $p(\mathbf{x}) = 0.7 + 0.2(\cos(2x + 0.4))^2$ with 10% of outliers

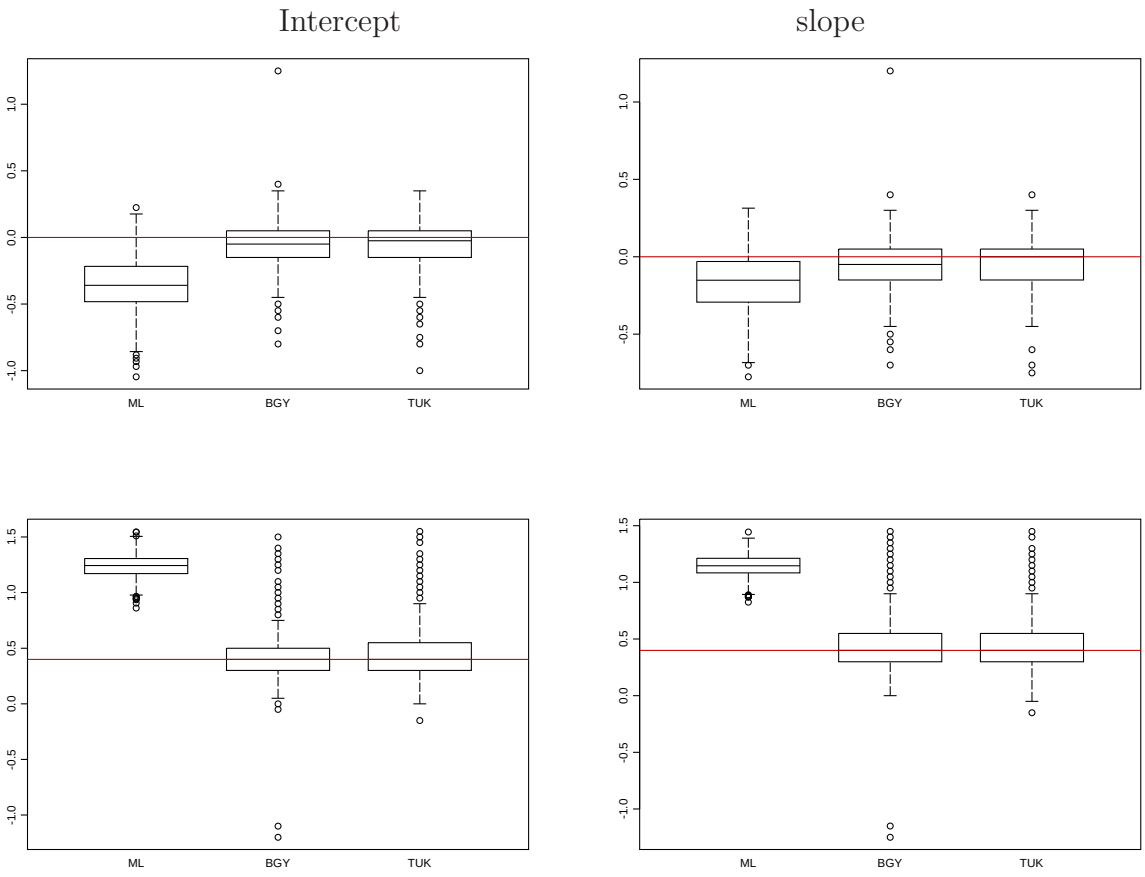

Figure 17: Boxplots $p(\mathbf{x}) = 1/(1 + \exp(-2x - 2))$ without of outliers

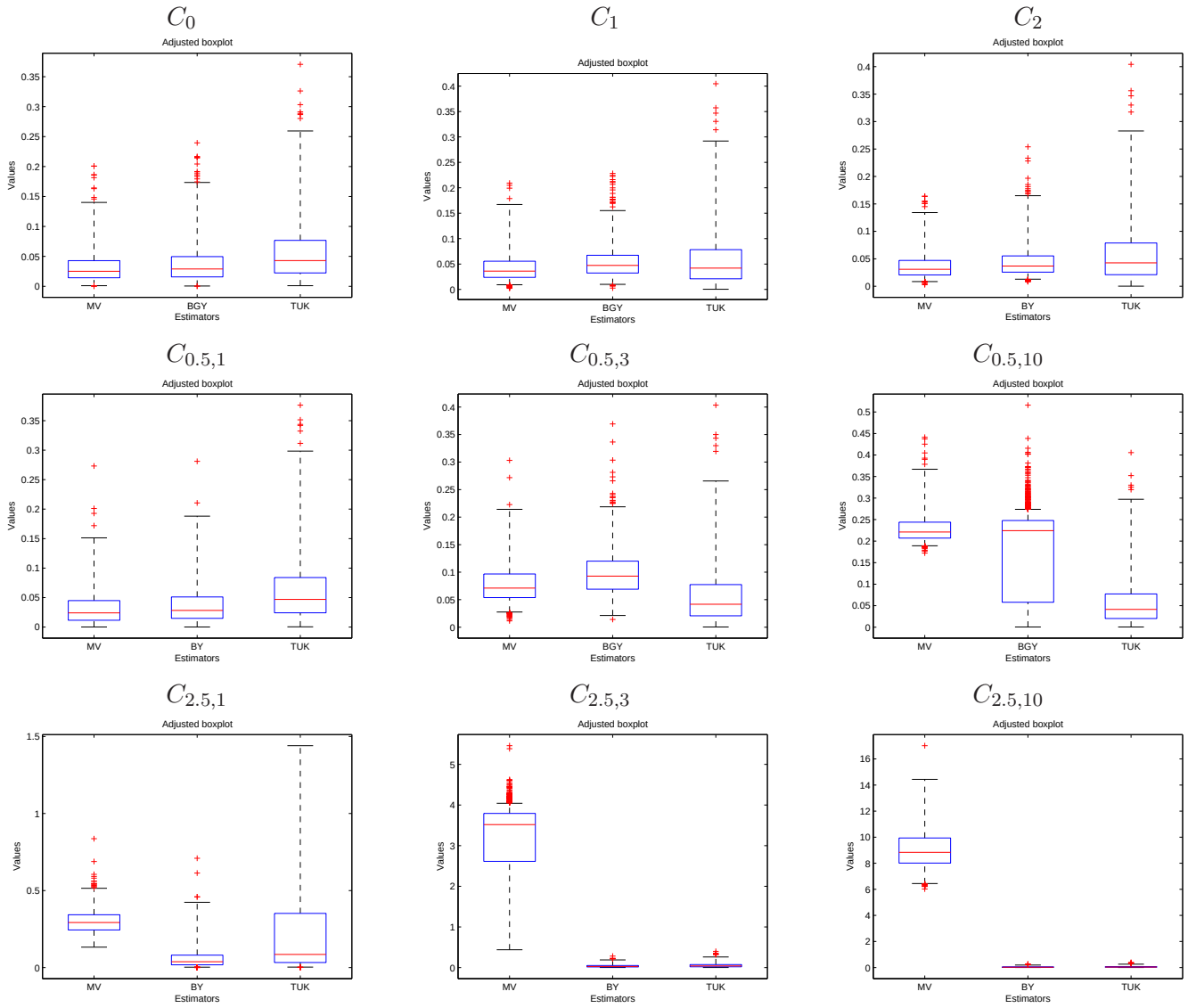

Figure 18: Adjusted boxplots for $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2$ under the Gamma model when $\tau = 1$, $c = \chi^2_{p,0.95}$, $p(\mathbf{x}) = 1$.

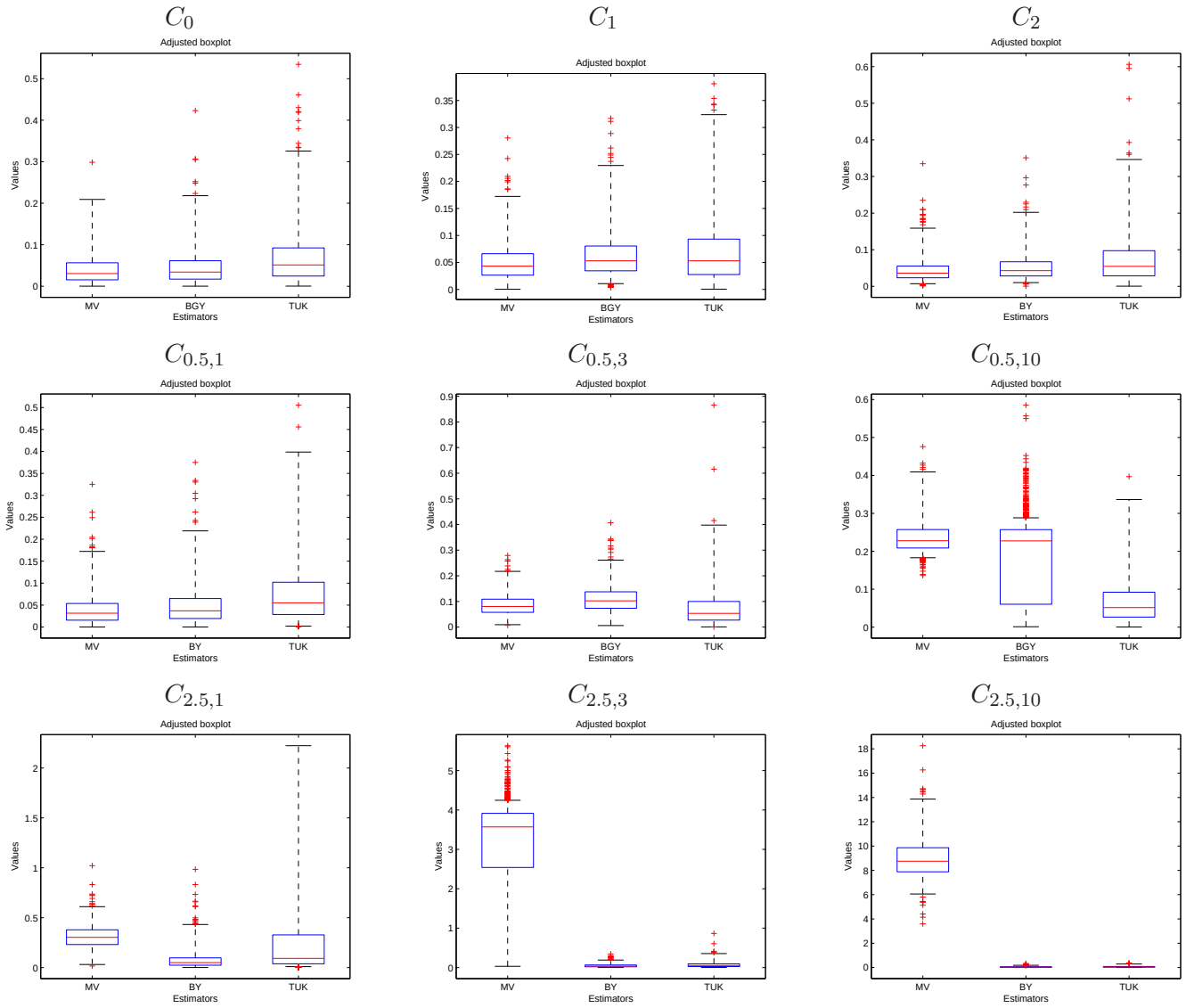

Figure 19: Adjusted boxplots for $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2$ under the Gamma model when $\tau = 1$, $c = \chi^2_{p,0.95}$, $p(\mathbf{x}) = 0.8$.

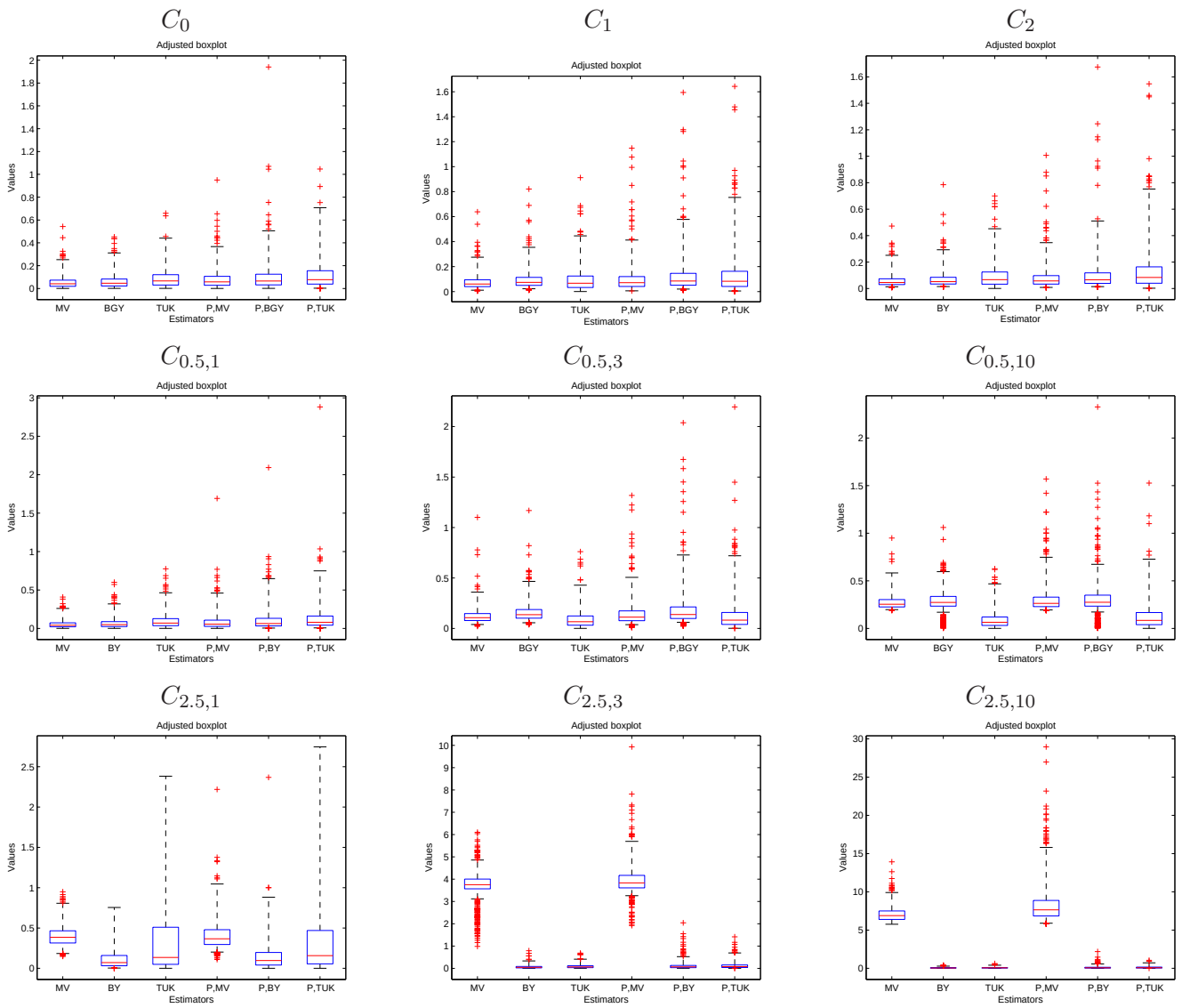

Figure 20: Adjusted boxplots for $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2$ under the Gamma model when $\tau = 1$, $c = \chi^2_{p,0.95}$, $p(\mathbf{x}) = 1/(1 + \exp(-\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{x} - 2))$ with $\boldsymbol{\lambda} = (2, 2)^{\mathrm{T}}$.

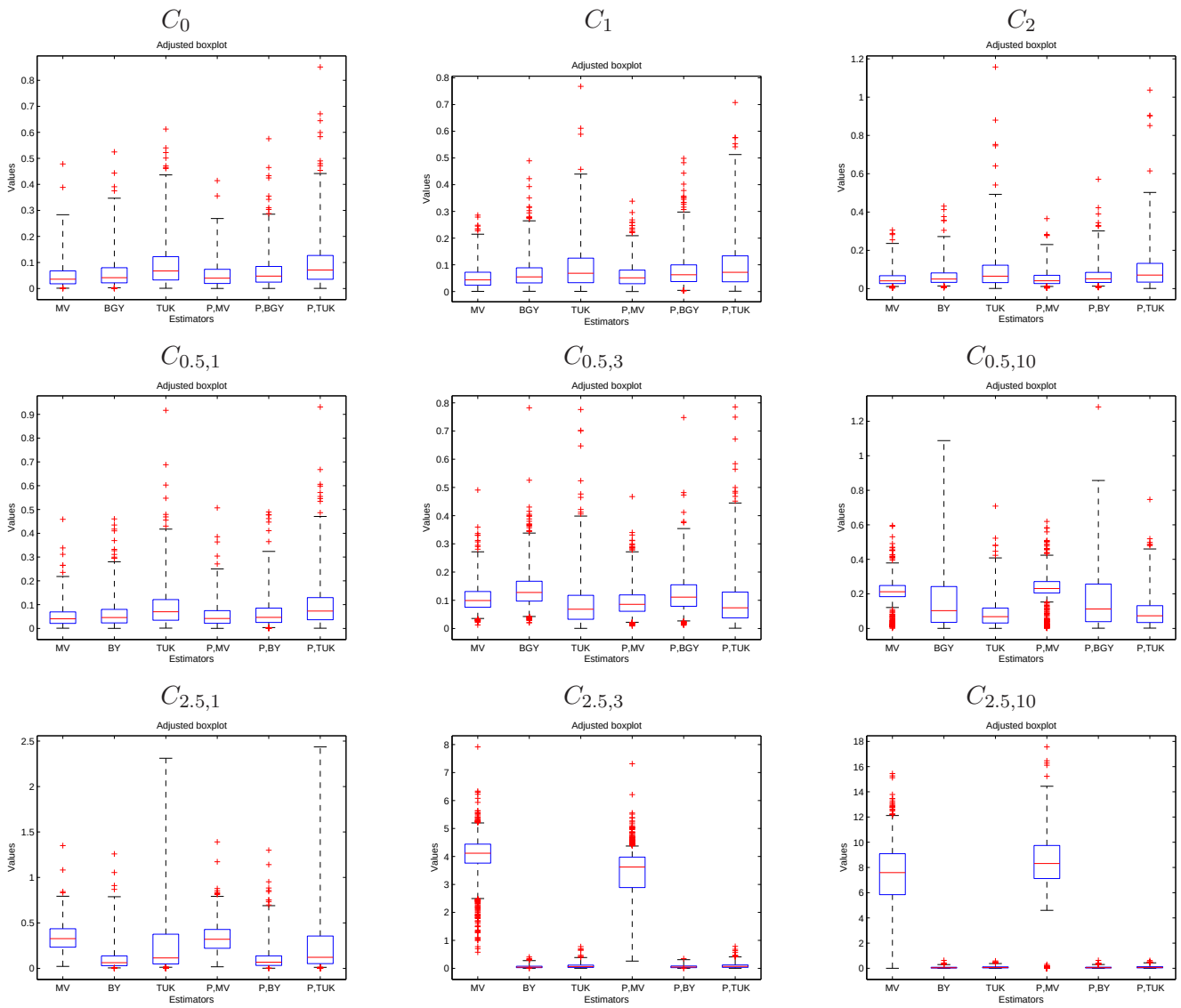

Figure 21: Adjusted boxplots for $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|^2$ under the Gamma model when $\tau=1$, $c=\chi^2_{p,0.95}$, $p(\mathbf{x})=0.4+0.5(\cos(\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{x}+0.4))^2$ with $\boldsymbol{\lambda}=(2,2)^{\mathrm{T}}$.

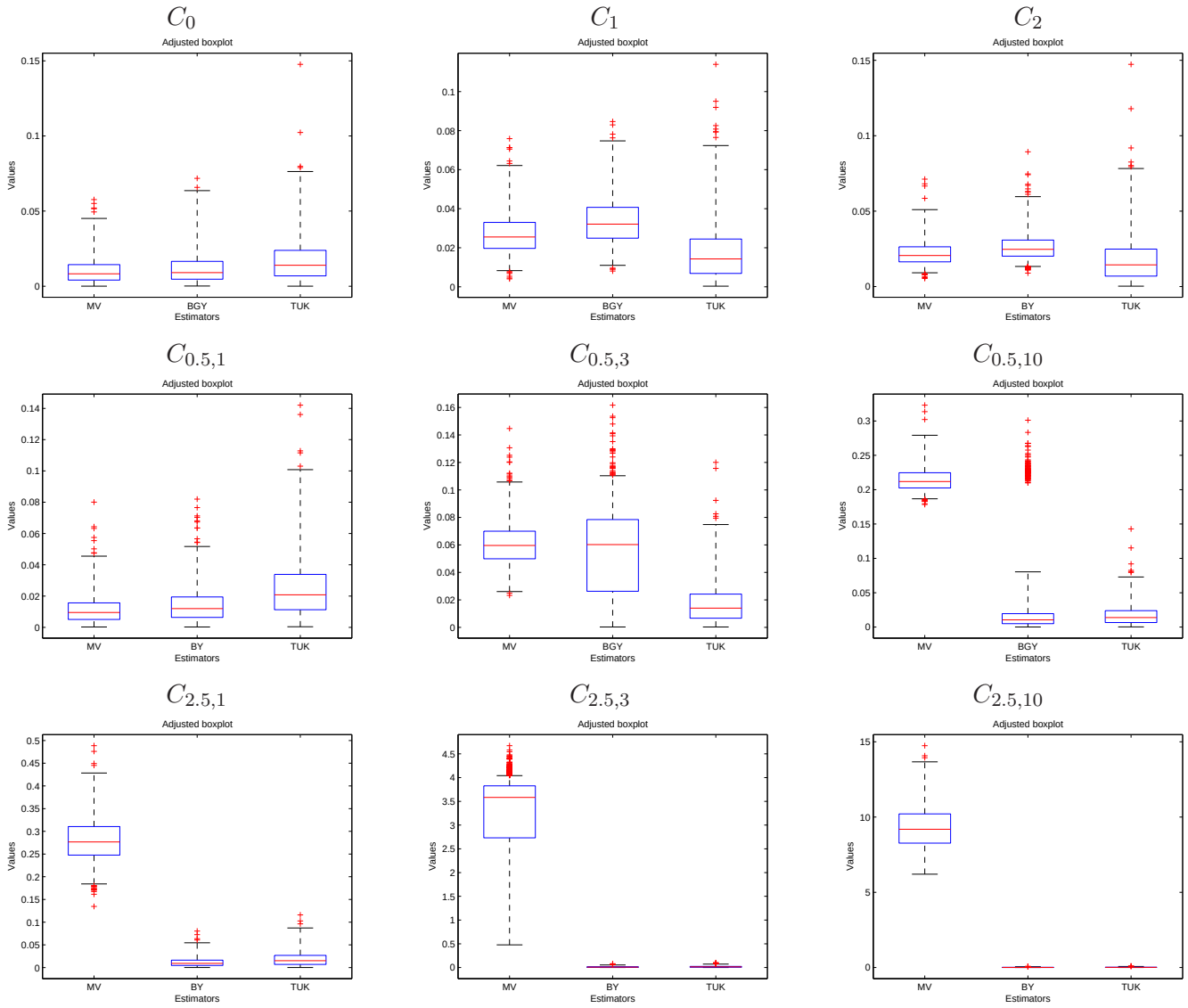

Figure 22: Adjusted boxplots for $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2$ under the Gamma model when $\tau = 3$, $c = \chi^2_{p,0.95}$, $p(\mathbf{x}) = 1$.

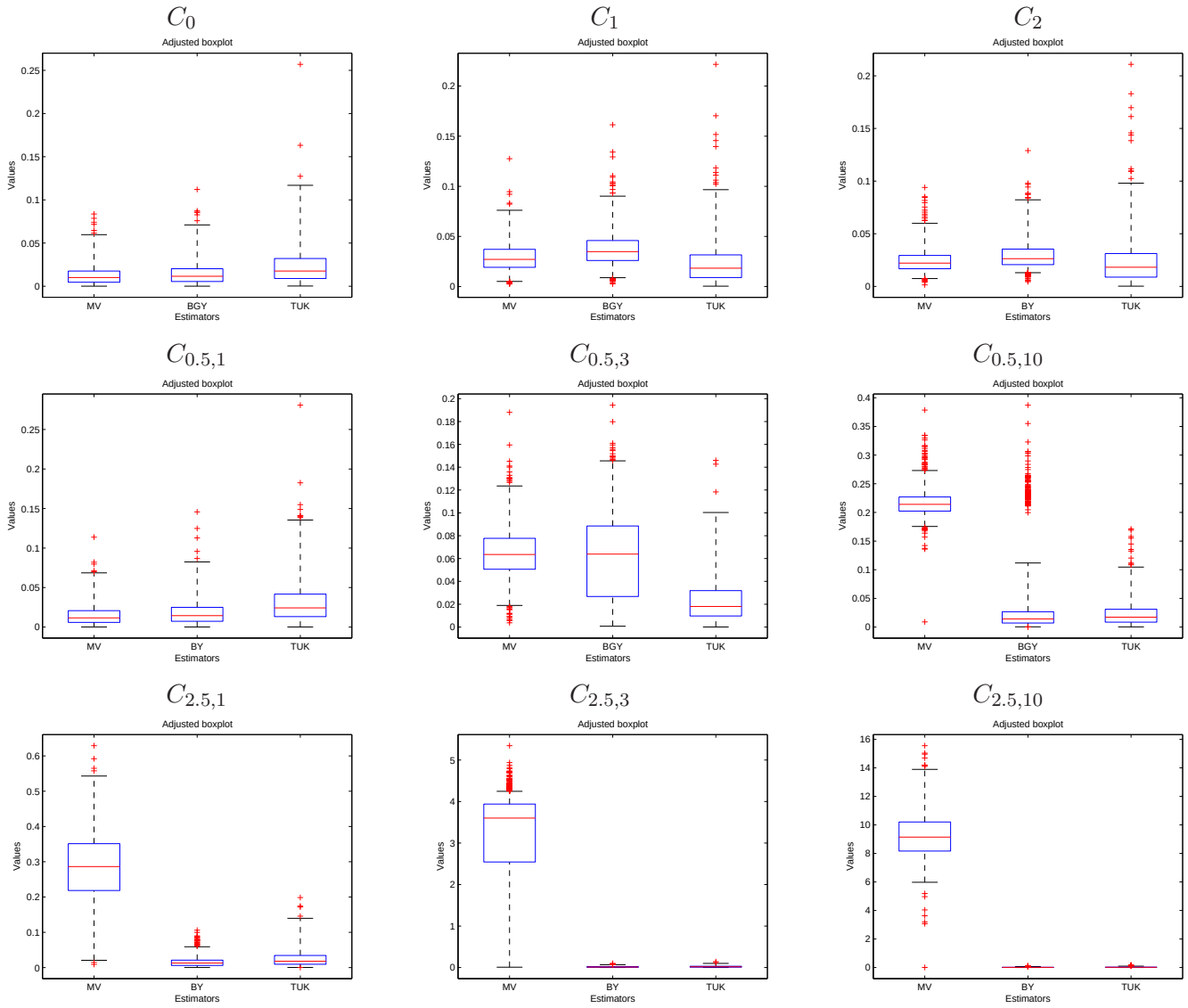

Figure 23: Adjusted boxplots for $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2$ under the Gamma model when $\tau = 3$, $c = \chi^2_{p,0.95}$, $p(\mathbf{x}) = 0.8$.

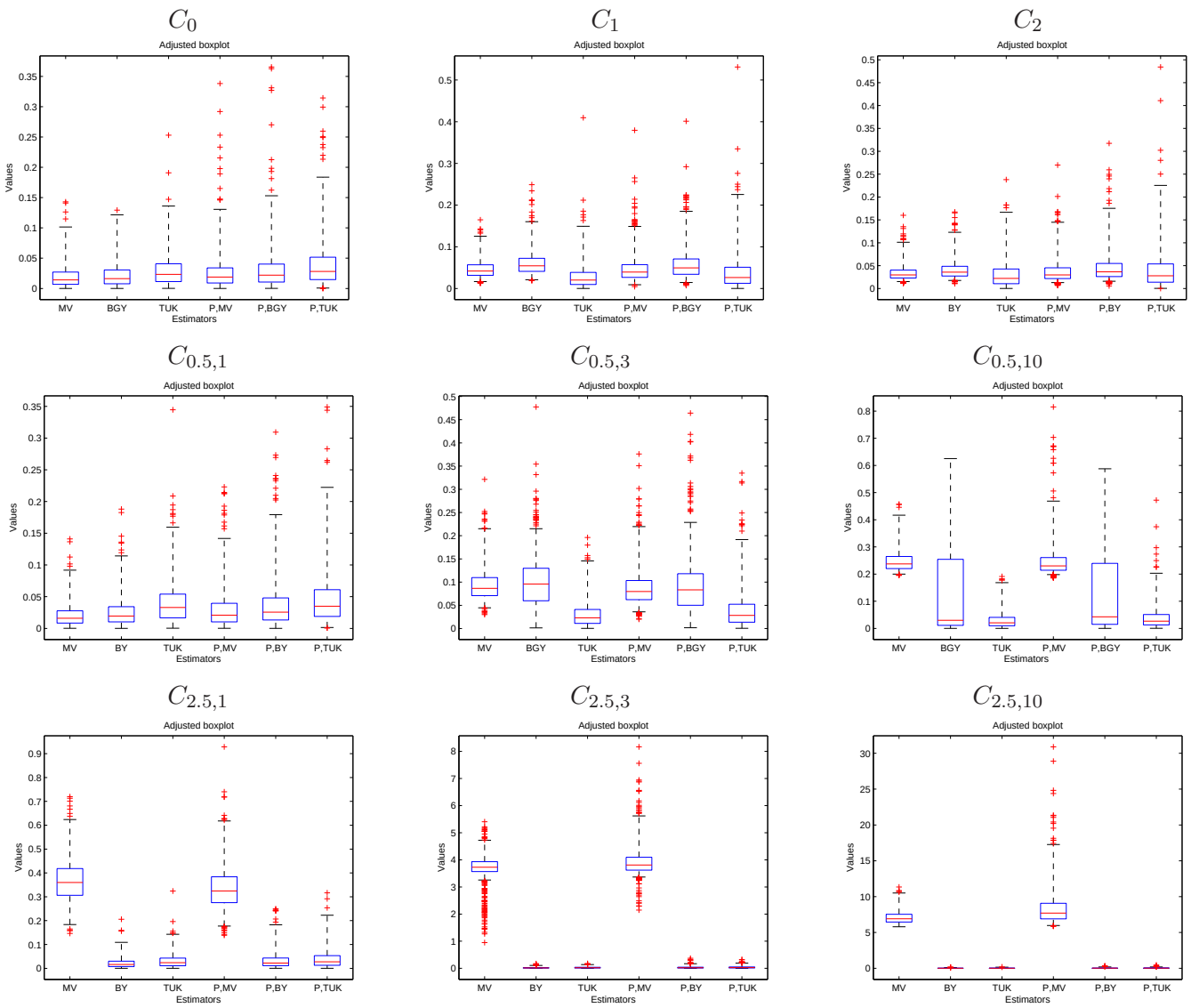

Figure 24: Adjusted boxplots for $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2$ under the Gamma model when $\tau = 3$, $c = \chi^2_{p,0.95}$, $p(\mathbf{x}) = 1/(1 + \exp(-\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{x} - 2))$ with $\boldsymbol{\lambda} = (2, 2)^{\mathrm{T}}$.

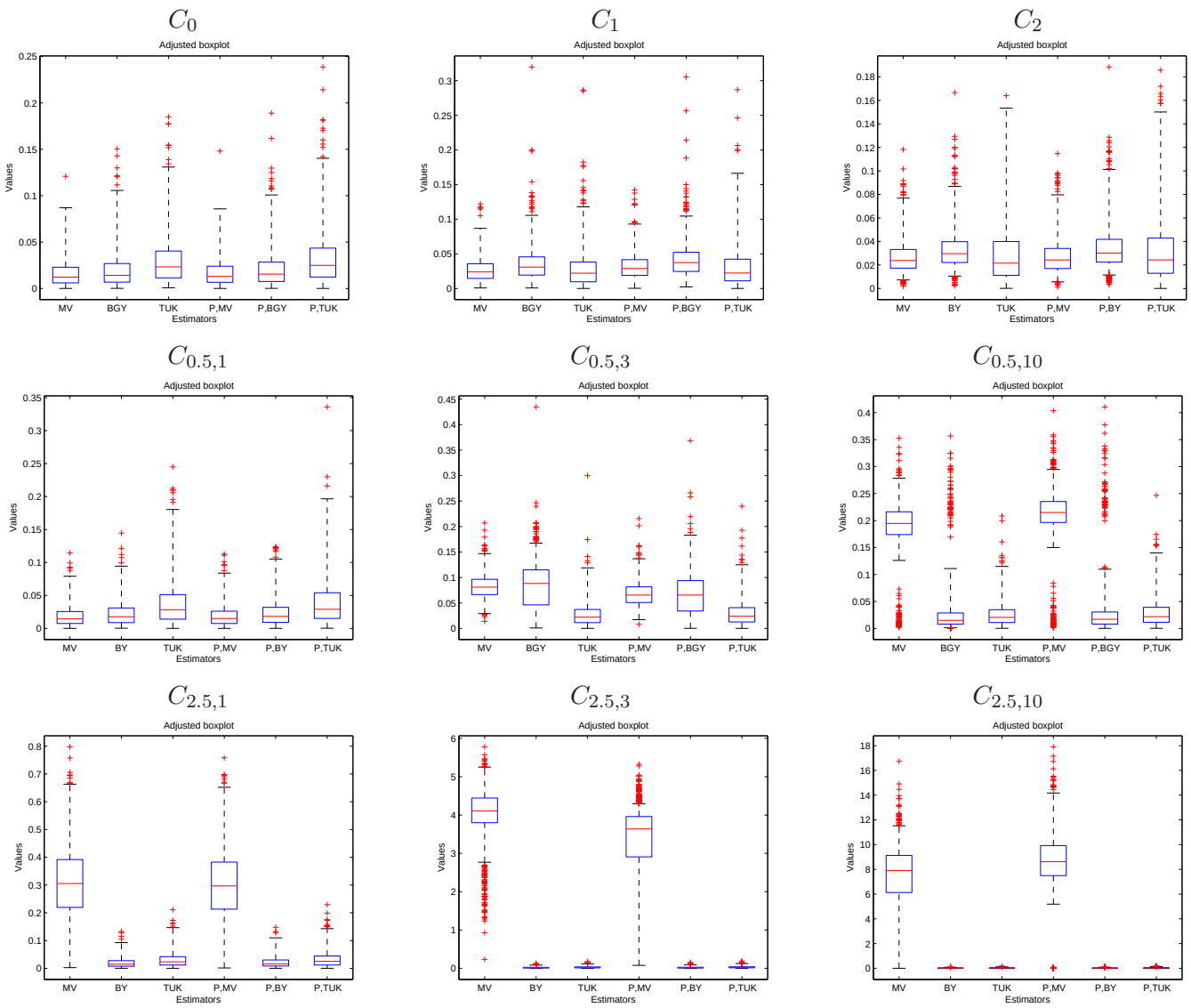

Figure 25: Adjusted boxplots for $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|^2$ under the Gamma model when $\tau=3$, $c=\chi^2_{p,0.95}$, $p(\mathbf{x})=0.4+0.5(\cos(\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{x}+0.4))^2$ with $\boldsymbol{\lambda}=(2,2)^{\mathrm{T}}$.

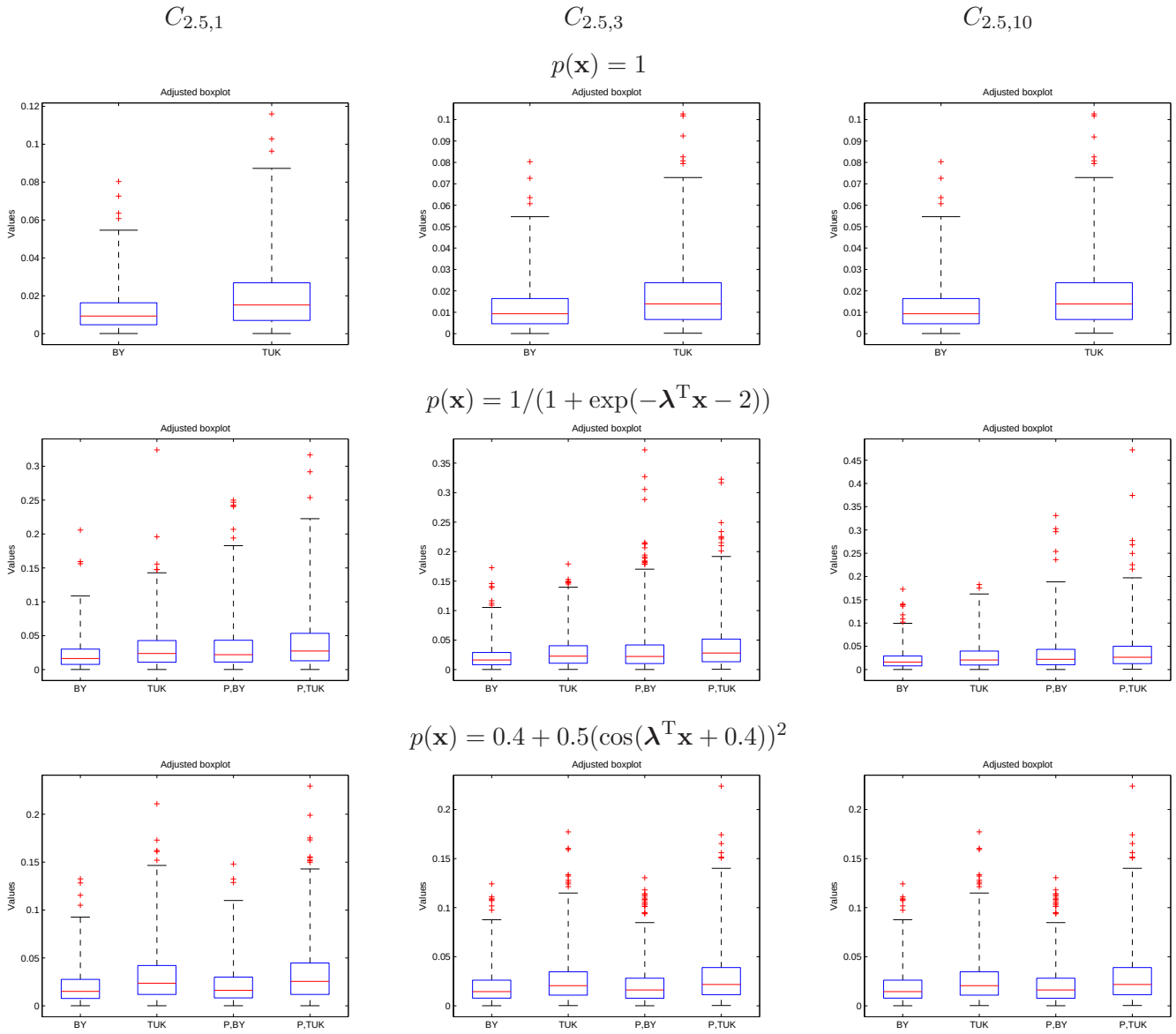

Figure 26: Adjusted boxplots for $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2$ when considering the robust estimators, under the Gamma model when $\tau = 3$, $c = \chi^2_{p,0.95}$, when $m = 2.5$.