mathematics

MDPI

Article

Robust Three-Step Regression Based on Comedian and Its Performance in Cell-Wise and Case-Wise Outliers

Henry Velasco [1], Henry Laniado [1], Mauricio Toro [2], Víctor Leiva [3,*] and Yuhlong Lio [4]

[1] Department of Mathematical Sciences, Universidad Eafit, Medellín 050022, Colombia; hgvelascov@eafit.edu.co (H.V.); hlaniado@eafit.edu.co (H.L.)
[2] Department of Informatics and Systems Engineering, Universidad Eafit, Medellín 050022, Colombia; mtorobe@eafit.edu.co
[3] School of Industrial Engineering, Pontificia Universidad Católica de Valparaíso, Valparaíso 2362807, Chile
[4] Department of Mathematical Sciences, University of South Dakota, Vermillion, SD 57069, USA; Yuhlong.Lio@usd.edu
* Correspondence: victor.leiva@pucv.cl or victorleivasanchez@gmail.com

Received: 19 June 2020; Accepted: 29 July 2020; Published: 1 August 2020

Abstract: Both cell-wise and case-wise outliers may appear in a real data set at the same time. Few methods have been developed in order to deal with both types of outliers when formulating a regression model. In this work, a robust estimator is proposed based on a three-step method named 3S-regression, which uses the comedian as a highly robust scatter estimate. An intensive simulation study is conducted in order to evaluate the performance of the proposed comedian 3S-regression estimator in the presence of cell-wise and case-wise outliers. In addition, a comparison of this estimator with recently developed robust methods is carried out. The proposed method is also extended to the model with continuous and dummy covariates. Finally, a real data set is analyzed for illustration in order to show potential applications.

Keywords: case-wise contamination; comedian; MAD; Monte Carlo simulation; R software; robustness; Rocke S-estimator; 3S-regression

1. Introduction

Regression models are one of most used statistical tools for diverse practitioners [1]. A well-known assumption to infer the parameters of these models is that the error term follows a normal distribution. However, as it was discussed in [2], the presence of outliers could affect the estimation of parameters under normality and lead to inaccurate results [3]. Because outliers frequently are present in real data sets, robust estimation methods are called to be considered in practice to avoid this inaccuracy [4–7].

A type of model introduced in [8] is known as three-step (3S) regression, which proceed as follows. In the first step, a univariate filter is applied to remove cell-wise outliers. In the second step, a generalized S-estimator (GSE) is used to down-weight the effect of case-wise outliers. In the third step, the regression coefficients are estimated. Nevertheless, some limitations of the 3S-regression method are mentioned in [9]. For instance, the GSE employed in the second step loses robustness against case-wise outliers, when the dimension is greater than ten. In addition, the extended minimum volume ellipsoid (EMVE) estimator, also utilized by the GSE as an initial value, is computationally slow and it does not scale well to higher dimensions. Hence, a new robust estimator, called generalized Rocke S-estimator (GRE), is proposed in [9] to replace the GSE in the second step. Furthermore, a cluster-based algorithm introduced in [9] for faster and more reliable sampling when computing

the EMVE estimator, named EMVE-C, is used as an initial value for the GRE algorithm. To our best knowledge, the comedian [10,11] has not been employed as an initial estimate in 3S-regression methods.

In robust linear regression, the case-wise contamination model is known as Tuckey-Huber contamination (THC). This model has been broadly studied in the literature, but its use in practice is not frequent. In the THC model, a small proportion of cases are contaminated. When the contamination is carried out using cell-wise, the independent contamination (IC) model arises. Note that the IC model has a small proportion of individual cells in the covariates that are independently contaminated [12]. However, the literature about the IC model is limited. To our best knowledge, there is few works that can deal with both types of outliers (case-wise and cell-wise) at the same time.

A classical robust regression for case-wise contamination is the least median of square (LMS) method. The LMS regression was proposed in order to optimize the median of the squares of residuals and it allows for a breakdown point of 50%, but it is computationally inefficient [13,14]. Two alternative methods that use iterative strategies are the regressions via the estimation of: (i) the minimal covariance determinant (MCD) [15,16]; and (ii) the iterative re-weighted least square (IRLS) [17,18]. The MCD regression minimizes the covariance matrix determinant of the central points. The IRLS regression includes additional information regarding error variance and covariance by incorporating a weight matrix into the model estimation, whose diagonal elements depend on a loss function. High breakdown affine equivariant estimators provide down-weighting to outlying cases, such as the least trimmed square (LTS) regression [14], S-regression [19], and MM-regression [20]; see also [21,22] for M-estimators in regression. All of these methods work well in practice under the THC model.

A number of authors [8,23–28] have proposed robust regression models that are resilient to case-wise and cell-wise outliers by robustifying the components of the covariance matrix in the solution of the least square (LS) optimization problem. Additionally, the multivariate S-estimator is incorporated instead of the empirical covariance and mean [24,25]. It has been also showed that, under mild assumptions (including symmetry and independence in the residuals), the 2S-regression estimator [8] is Fisher consistent and asymptotically normal, even if the multivariate S-estimators are not. Based on incomplete data, two kinds of estimators were constructed [26]: (i) the GSE and (ii) the extended S-estimator (ESE), which match with the multivariate S-estimator under complete data. Note that the GSE needs a robust initial estimate. Furthermore, the extended EMVE estimator is introduced in [26] as a particular case of the ESE. The EMVE estimator can be considered to be an initial value and a generalization of the minimum volume ellipsoid (MVE) estimator proposed in [15]. Moreover, the shooting S-estimator that is derived in [28] assigns individual weights to each cell in the data table, combining the shooting algorithm [29] and the simple S-regression [19]. Observe that the data may be snipped replacing cell-wise outliers by missing values NA [27]. Moreover, the Gervini-Yohai univariate filter [30] can be used followed by the GSE [26,31]. Notice that the 3S-regression [8] considers an estimator which is analogous to one defined in [26], but with the filter that is consistent for a broader range of distributions.

Based on this bibliographical review, the objective of this study is to propose a comedian-three-step (C3S) regression estimator that considers the comedian as the initial robust scatter value for the GRE algorithm. The proposed estimation method: (i) utilizes an adaptive consistent univariate filter to control the effect of extreme cell-wise outlier propagation; (ii) applies the GRE algorithm, but modified using the sample comedian matrix and wise-median as an initial robust scatter estimate for the filtered data; and (iii) estimates the regression coefficients using the GRE algorithm in the previous step.

This paper is organized, as follows. In Section 2, the general context and notations are provided. Then, the consistent factor in median absolute deviation (MAD), the comedian function, and the empirical comedian covariance matrix are defined. Section 3 introduces the models with continuous and dummy covariates and proposes an estimator based on the C3S-regression, as well as its asymptotic properties. In Section 4, an extensive simulation study is conducted in order to compare the performance of proposed estimator with recently developed robust methods. Additionally,

in this section, a real data example is used for illustration and for showing potential applications. Section 5 describes some conclusions and ideas for possible future works.

2. Comedian Covariance Matrix and Comedian Matrix

In this section, the general context and notations used in this work are presented. The consistent factor in MAD, the comedian function, and the empirical comedian covariance matrix are also defined here.

2.1. General Context and Notations

Let X and Y be two continuous random variables. Subsequently, the MAD of X, the comedian between X and Y, and the correlation median between X and Y (δ) are, respectively, given as [10]

$$\begin{aligned} \text{MAD}(X) &= \text{median}(|X - \text{median}(X)|), \\ \text{COM}(X,Y) &= \text{median}((X - \text{median}(X))(Y - \text{median}(Y))), \\ \delta &= \frac{\text{COM}(X,Y)}{\text{MAD}(X)\text{MAD}(Y)}. \end{aligned} \tag{1}$$

Note that the MAD(X) defined in Equation (1) is a robust measure of dispersion (or scatter) of X, COM(X, Y) is a robust measure of the covariance between X and Y, while δ is a robust measure of the correlation between X and Y. When $X = Y$, $(\text{COM}(X,X))^{1/2}$ can be used as a robust measure of variability, such as it occurs with the standard deviation (SD) and the covariance of X, that is, $\text{SD}(X) = (\text{COV}(X,X))^{1/2}$. Then, a robust measure of the covariance and correlation matrices for any random vector may be obtained by utilizing the robust measures defined in (1). Notice that the usual covariance between X and Y can be obtained as

$$\varsigma = \text{COV}(X,Y) = \frac{\text{MAD}(X)\text{MAD}(Y)}{g(1)} g(\varrho), \tag{2}$$

where g is the comedian function stated as $g(\varrho) = \text{COM}(X,Y)$ (see Lemma 2.1 in [10]) and ϱ is the correlation coefficient between X and Y. Observe that ϱ may be represented in terms of the correlation median as

$$\varrho = \text{COR}(X,Y) = g^{-1}(g(1)\delta). \tag{3}$$

The comedian function under a non-degenerate bivariate normal distribution was studied in [10], obtaining $g(1) = (\Phi^{-1}(0.75))^2$, where Φ is the standard normal cumulative distribution function and Φ^{-1} is the inverse of Φ or normal quantile function. In this work, we also extend $g(1)$ to other non-normal distributions. Note that Equation (2) can be written as $\varsigma = b_X \text{MAD}(X) b_Y \text{MAD}(Y) g(\varrho)$, where

$$g(1) = (b_X b_Y)^{-1} \tag{4}$$

and the consistent factors b_X and b_Y depend upon the marginal distributions of X and Y, respectively. The consistent factors for some distributions have been obtained and are presented in Table 1. Detailed calculations of these factors are available upon request from the authors.

Notice that the comedian can be considered as a robust initial scatter estimate for the GRE algorithm. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent random vectors following a bivariate distribution. Subsequently, the empirical comedian is established by

$$\widehat{\text{COM}}_n(X,Y) = \text{median}_{i \in \{1,\ldots,n\}} (X_i - \widehat{\text{median}}_n(X))(Y_i - \widehat{\text{median}}_n(Y)), \tag{5}$$

where $\widehat{\text{median}}_n(X)$ and $\widehat{\text{median}}_n(Y)$ denote the sample medians of $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$, respectively. Thus, the covariance matrix that is defined in Equation (2) is a sophisticated scatter estimate based on the comedian matrix.

Table 1. Consistent factor b_X of the indicated distribution.

Distribution of X	Notation	b_X
Exponential	$X \sim \text{Exp}(\lambda)$	$1/\log\left((1+\sqrt{5})/2\right)$
Logistic	$X \sim \text{Logistic}(\mu, s)$	$\sqrt{3}\pi/(3\log(3))$
Normal	$X \sim \text{N}(\mu, \sigma)$	$1/\Phi^{-1}(3/4)$
Student-t	$X \sim t(\nu)$ *	$\approx \sqrt{\frac{\nu}{\nu-2}}/m_{\text{median}}(\nu), \quad \nu > 2$
Uniform	$X \sim \text{U}(a, b)$	$2/\sqrt{3}$
Weibull	$X \sim \text{Wei}(\alpha, \beta)$ *	$\approx \left(\alpha\sqrt{\Gamma\left(1+2/\beta\right) - \left(\Gamma\left(1+1/\beta\right)\right)^2}\right)/m_{\text{median}}(\alpha, \beta)$

* m_{median} is the solution of m to the non-linear equations for $t(\nu)$ and $\text{Wei}(\alpha, \beta)$ given, respectively, by $\frac{2m\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu}\Gamma(\nu/2)\,{}_2F_1\left(\frac{1}{2}, \frac{\nu+1}{2}; \frac{3}{2} - \frac{m^2}{\nu}\right)} - \frac{1}{2} = 0, \quad \exp\left(-\left((\log(2))^{1/\beta} - \frac{m}{\alpha}\right)^{\beta}\right) - \exp\left(-\left((\log(2))^{1/\beta} + \frac{m}{\alpha}\right)^{\beta}\right) - \frac{1}{2} = 0.$

2.2. The Consistent Factor in MAD

The MAD is a very robust scatter estimate, which has 50% breakdown point (the best possible). According to [32], if we want to estimate the SD consistently, the MAD must be multiplied by a correction factor. Thereby, an alternative robust estimate of the SD of X is given by $\widehat{S}_X = b_X\widehat{\text{MAD}}_n(X)$, where $\widehat{\text{MAD}}_n(X) = \text{median}\{|X_i - \widehat{\text{median}}_n(X)|, i \in \{1, \ldots, n\}\}$. The consistent factor b_X depends exclusively on the distribution of the random variable X. If the marginal distribution of X is unknown, the consistent factor can be estimated via the non-parametric bootstrapping method [33]. Let $\widehat{F}_n$ be the empirical cumulative distribution function of the random variable X. Subsequently, the bootstrapping process used to obtain the consistent factor b_X is summarized in Algorithm 1.

Algorithm 1 Bootstrapping process used in order to obtain the consistent factor b_X.

1: Generate $X_1^*, \ldots, X_n^* \sim \widehat{F}_n(x)$ randomly.

2: Compute $T_n^* = g(X_1^*, \ldots, X_n^*) = \widehat{\text{MAD}}_n(X_1^*, \ldots, X_n^*)/\widehat{S}_n(X_1^*, \ldots, X_n^*)$.

3: Repeat steps 1–2 B times to get $T_{n,1}^*, \ldots, T_{n,B}^*$.

4: Evaluate $\widehat{b}_X = (1/B)\sum_{b=1}^{B} T_{n,b}^*$.

2.3. The Comedian and Empirical Comedian Covariance

The comedian function $g(\varrho)$ is needed in order to estimate the correlation median defined in Equation (3). The empirical correlation median $\widehat{\delta}_n = \widehat{\text{COM}}_n(X, Y)(\widehat{\text{MAD}}_n(X)\widehat{\text{MAD}}_n(Y))^{-1}$ stated in Equation (5) can be seen as a robust estimate of the correlation coefficient by

$$\widehat{\varrho}_n = g^{-1}(g(1)\widehat{\delta}_n). \tag{6}$$

The function g was analyzed in [10] when (X, Y) has a bivariate normal distribution, but an explicit form was not obtained. However, it may be approximated through Monte Carlo simulations. We conduct an extensive Monte Carlo simulation study for g via the `R` software [34]. This simulation was carried out by using an `R` package named `MASS` and its `mvrnorm` function. The empirical medians of 10 000 000 random numbers from a bivariate normal distribution were also used, with ϱ varying from -1 to 1 by 0.01 when $\varrho \notin [-0.1, 0.1]$, and by 0.001 when $\varrho \in [-0.1, 0.1]$. The number of replicates is $N = 10$ and a visualization of the approximation for g is shown in Figure 1.

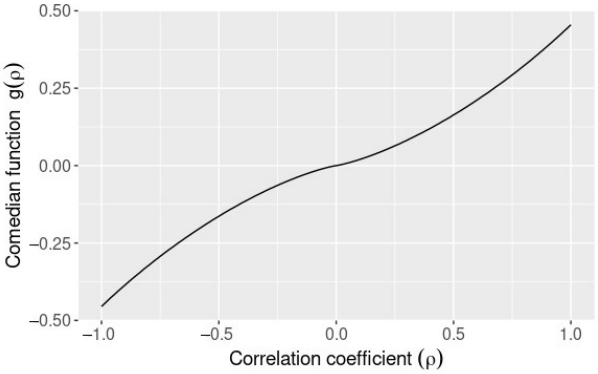

Figure 1. Approximation obtained by simulations of the comedian $g(\varrho) = \mathrm{COM}(X, Y)$ as a function of the correlation coefficient ϱ, where $g(1) = (\Phi^{-1}(0.75))^2$. Source: the authors.

In general, the exact image value of the inverse comedian function g^{-1} cannot be obtained, but it may be estimated through an approximation. A discrete approximation of the comedian function is obtained at the aforementioned simulation study. The expression that is given in Equation (6) may be approximated as $\widehat{\varrho}_n = \widehat{g}^{-1}(g(1)\widehat{\delta}_n)$, where $\widehat{g}^{-1}$ is an estimate of g^{-1} by interpolating the approximated points while using a cubic spline. To carry this method, it is necessary to know all of the values corresponding to the support of the inverse comedian function. If the consistent factors of the marginal distributions defined in Equation (4) are estimated properly, then $\widehat{\varrho}_n$ for other bivariate distributions can be obtained.

By using the empirical marginal distributions, the consistent factors may be estimated via bootstrapping, as described in Section 2. Thus, from Equation (6), we have $\widehat{\varrho}_n = \widehat{g}^{-1}((\widehat{b}_X\widehat{b}_Y)^{-1}\widehat{\delta}_n)$.

Let $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ be a set of p covariates. Then, a robust version of the empirical covariance between any pair of covariates (X_i, X_j), for $i, j \in \{1, \ldots, p\}$, can be stated as

$$\widehat{S}^{\mathrm{c}}_{X_iX_j} = \widehat{b}_{X_i}\widehat{\mathrm{MAD}}_n(X_i)\,\widehat{b}_{X_j}\widehat{\mathrm{MAD}}_n(X_j)\widehat{\varrho}_n, \tag{7}$$

where $\widehat{S}^{\mathrm{c}}_{X_i,X_j}$ is an element of the robust version of the empirical covariance matrix $\widehat{\boldsymbol{S}}^{\mathrm{c}}_{XX}$. We also propose to use the expression given in Equation (7) as an initial scatter estimate for the GRE algorithm instead of the EMVE estimator. This proposal is called here the full version of the C3S-regression estimator.

3. Comedian Three-Step Regression

In this section, the model with continuous and dummy covariates is introduced. Moreover, the proposed estimator that is based on the 3S-regression, as well as its asymptotic properties, are developed.

3.1. The Proposed Estimator

A multiple regression is used to model the linear relationship between a dependent (response) variable Y and p independent (covariates) variables $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ with observed values for the case i denoted by $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$. Subsequently, the multiple regression model can be written as

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad i \in \{1, \ldots, n\}, \tag{8}$$

where the error terms ε_i, for $i \in \{1, \ldots, n\}$, are independent and identically distributed random variables, which are also independent of the values of the covariates $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$.

The LS estimates of the parameters $(\beta_0, \boldsymbol{\beta})$ are defined as the solution to an optimization problem in order to minimize the sum squares of residuals as

$$(\widehat{\beta_0}_{\mathrm{LS}}, \widehat{\boldsymbol{\beta}}_{\mathrm{LS}}^{\top}) = \mathrm{argmin}_{(\beta_0, \boldsymbol{\beta}^{\top}) \in \mathbb{R}^{(p+1)}} \sum_{i=1}^{n} (Y_i - \beta_0 - \boldsymbol{x}_i^{\top} \boldsymbol{\beta})^2. \tag{9}$$

The solution of Equation (9) can be explicitly given by

$$\widehat{\beta_0}_{\mathrm{LS}} = \widehat{\mu}_Y - \widehat{\boldsymbol{\mu}}_X^{\top} \widehat{\boldsymbol{\beta}}_{\mathrm{LS}}, \quad \widehat{\boldsymbol{\beta}}_{\mathrm{LS}} = \widehat{\boldsymbol{\Sigma}}_{XX}^{-1} \widehat{\boldsymbol{\Sigma}}_{XY}, \tag{10}$$

where $\widehat{\boldsymbol{\Sigma}}_{XX}$ and $\widehat{\boldsymbol{\Sigma}}_{XY}$ are the components of the empirical covariance matrix, and $\widehat{\mu}_Y$ and $\widehat{\boldsymbol{\mu}}_X$ are empirical means of Y and X, respectively.

As suggested by a number of authors [8,23–28], the components of the solution stated in Equation (10) can be robustified to immunize the estimator against case-wise and cell-wise outliers. Inspired by [9], we use a modified version of the GRE algorithm to obtain robust estimates of means and covariances needed in the solution presented in Equation (10). The modification proposed by us is that the GRE algorithm considers the comedian as an initial value instead of the EMVE-C estimate. The robust method basically utilizes the empirical median and the robust version of the covariance, introduced in Section 2, as the initial location and scatter estimates for the GRE algorithm. The proposed estimator uses the univariate filter given in [8] and the GRE algorithm for incomplete data developed in [9]. Our proposal works similarly to that used in the 3S-regression, but employing in the second step a different initial robust estimate for the GRE algorithm. In the present work, the initial estimates of location and scatter, the empirical median, and a robust estimate of the covariance, are computed after snipping the data. Therefore, the proposed robust regression estimator (C3S-regression) is established as

$$\widehat{\beta_0}_{\mathrm{C3S}} = \widehat{m}_Y - \widehat{\boldsymbol{m}}_X \widehat{\boldsymbol{\beta}}_{\mathrm{C3S}}, \quad \widehat{\boldsymbol{\beta}}_{\mathrm{C3S}} = \widehat{\boldsymbol{S}}_{XX}^{-1} \widehat{\boldsymbol{S}}_{XY}, \tag{11}$$

where both $\widehat{\boldsymbol{m}}$ and $\widehat{\boldsymbol{S}}$ come from the modified GRE algorithm proposed in this work, and they are computed as in Algorithm 2.

Algorithm 2 Computation of $\widehat{\boldsymbol{m}}$ and $\widehat{\boldsymbol{S}}$ from the modified GRE algorithm.

1: Filter extreme cell-wise outliers using a univariate filter to prevent cell-wise contaminated cases.
2: Compute the wise-median and the robust version of the covariance matrix (or comedian matrix) as initial robust location and scatter estimates.
3: Down-weight the effect of case-wise outliers by applying the GRE algorithm for computing robust location and scatter estimates with the filtered data from Step 1.

Now, consider a set of n data with observed covariates $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ and the corresponding response variable $\{Y_1, \ldots, Y_n\}$. Let $\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n\}$ be the joint data with $\boldsymbol{z}_i = (Y_i, \boldsymbol{x}_i^{\top})^{\top}$. In the first step, a univariate filter, as described in [8], is applied to each observed covariate x_j, for $j \in \{1, \ldots, p\}$. Let $\mathbb{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)^{\top}$ and $\mathbb{U}$ denote the resulting auxiliary matrices of zeros and ones, with zeros indicating the filtered (missing) entries. Subsequently, based on the GRE algorithm, we obtain

$$\widehat{\boldsymbol{m}} = \widehat{\boldsymbol{m}}_{\mathrm{GRE}}(\mathbb{Z}, \mathbb{U}), \quad \widehat{\boldsymbol{S}} = \widehat{\boldsymbol{S}}_{\mathrm{GRE}}(\mathbb{Z}, \mathbb{U}), \tag{12}$$

where $\widehat{\boldsymbol{m}}_{\mathrm{GRE}}$ and $\widehat{\boldsymbol{S}}_{\mathrm{GRE}}$ are robust location and scatter based on the GRE algorithm for the incomplete data $(\mathbb{Z}, \mathbb{U})$. Computation of the 3S-regression and C3S-regression estimates is summarized Algorithm 3.

Algorithm 3 Computation of 3S-regression and C3S-regression estimates.

1: Snip data.
2: Apply the GRE algorithm with robust initial location and scatter estimates.
3: Estimate the regression coefficients as in Equation (10).

Note that the C3S-regression uses a robust estimated covariance matrix as an initial scatter value instead of the EMVE estimate. In addition, the biflat ρ function [35] is employed instead of the Tukey bisquare ρ function for the GSE. More details of the definitions and algorithms of the EMVE estimate and GSE can be found in [26]. Furthermore, the GRE algorithm was studied in [9], showing that, in large dimension, the Rocke biflat function is more robust than the Tuckey bisquare function [35,36].

3.2. Models with Continuous and Dummy Covariates

Notice that the M-regression and 3S-regression were used in [8] in order to deal with continuous and dummy covariates. There, a 3S-regression was employed to estimate the coefficients of the continuous covariates, whereas an M-regression with the Huber ρ function given by $\rho_{\mathrm{H}}(t) = \min(1, t^2/2)$ [37] was considered to estimate the coefficients of the dummy covariates. This is a modification of the M-regression and 3S-regression proposed in [38]. We act similarly by using the C3S-regression to estimate the coefficients of the continuous covariates and the M-regression for the dummy coefficients. Consider the model with continuous and dummy covariates defined as

$$Y_i = \beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta}_1 + \boldsymbol{d}_i^\top \boldsymbol{\beta}_2 + \varepsilon_i, \quad i \in \{1, \ldots, n\}, \tag{13}$$

where $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip_1})^\top$ and $\boldsymbol{d}_i = (d_{i1}, \ldots, d_{ip_2})^\top$ are a $p_1 \times 1$ vector of continuous covariates and a $p_2 \times 1$ vector of dummy covariates, respectively. Let $\mathbb{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$, $\mathbb{D} = (\boldsymbol{d}_1, \ldots, \boldsymbol{d}_n)^\top$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$, where the columns in $\mathbb{X}$ and $\mathbb{D}$ are linearly independent. More precisely, our method that is based on the M-regression and C3S-regression works as

$$(\widehat{\beta}_0^{(r)}, \widehat{\boldsymbol{\beta}}_1^{(r)}) = h(\mathbb{X}, \boldsymbol{Y} - \mathbb{D}\widehat{\beta}_2^{(r-1)}), \quad \widehat{\boldsymbol{\beta}}_2^{(r)} = M(\mathbb{D}, \boldsymbol{Y} - \widehat{\beta}_0^{(r)} - \mathbb{X}\widehat{\boldsymbol{\beta}}_1^{(r)}), \quad r \in \{1, \ldots, R\}, \tag{14}$$

where h denotes the operator of a C3S-regression in each iteration for $(\mathbb{X}, \boldsymbol{Y})$; while M denotes the operator of an M-regression with no intercept for $(\mathbb{D}, \boldsymbol{Y})$, as stated in (11). To control the effect of propagation of cell-wise outliers, let $\widehat{\mathbb{X}}$ be the imputed $\mathbb{X}$ with the filtered entries by the linear predictor using $(\widehat{\boldsymbol{m}}^{(r)}, \widehat{\boldsymbol{S}}^{(r)})$ as defined in Equation (12) at the rth iteration of the GRE algorithm. The method that is presented in Equation (14) needs initial estimates $(\widehat{\beta}_0^{(0)}, \widehat{\boldsymbol{\beta}}_1^{(0)}, \widehat{\boldsymbol{\beta}}_2^{(0)})$ to start the algoritm until a maximum of $R = 20$ iterations [8]. Then, we first remove the effect of $\boldsymbol{d}_i$ from the continuous covariates and response. Let $\overline{\boldsymbol{Y}} = \boldsymbol{Y} - \mathbb{D}\,\boldsymbol{t}$ and $\overline{\mathbb{X}} = \mathbb{X} - \mathbb{D}\mathbb{T}$, where $\boldsymbol{t} = M(\mathbb{D}, \boldsymbol{Y})$ and $\mathbb{T}$ is a $p_1 \times p_2$ matrix with the jth column as $\boldsymbol{T}_j = M(\mathbb{D}, (x_{ij}, \ldots, x_{nj})^\top)$. Subsequently, the initial estimates are defined by $(\widehat{\beta}_0^{(0)}, \widehat{\boldsymbol{\beta}}_1^{(0)}) = h(\overline{\mathbb{X}}, \overline{\boldsymbol{Y}})$ and $\widehat{\boldsymbol{\beta}}_2^{(0)} = M(\mathbb{D}, \boldsymbol{Y} - \widehat{\beta}_0^{(0)} - \widehat{\mathbb{X}}\widehat{\boldsymbol{\beta}}_1^{(0)})$.

3.3. Asymptotic Properties of the Comedian Three-Step Regression

The strong consistency of the empirical comedian is proved in [10], as well as its asymptotic normality. The strong consistency, asymptotic normality, and regularity assumptions for the GSE were established in [8]. Because the respective estimates under the 3S-regression and C3S-regression are based on the same GSE, independent of the differences in the initial estimates and weight functions, asymptotic properties of the corresponding estimators from the C3S-regression and 3S-regression are guaranteed. Note that the 3S-regression and C3S-regression become a 2S-regression for an enough large n. Therefore, the estimators obtained from the C3S-regression inherit the asymptotic properties of the estimators obtained from the 2S-regression such as the 3S-regression does. The properties of the corresponding asymptotic covariance matrix are also found in [8].

Let H be the distribution of $(\boldsymbol{X}^\top, Y)$, $(\widehat{\boldsymbol{m}}, \widehat{\boldsymbol{S}})$ be the GSE, and $(\widehat{\beta_0}_{\text{C3S}}, \widehat{\boldsymbol{\beta}}_{\text{C3S}}^\top)$ be the estimated 3S-regression coefficients. Subsequently, $\boldsymbol{z}_i = (\boldsymbol{x}_i^\top, Y_i)$ is replaced by $\widehat{\boldsymbol{z}}_i = (\widehat{\boldsymbol{x}}_i^\top, Y_i)$ and $\widetilde{\boldsymbol{x}}_i = (1, \boldsymbol{x}_i^\top)^\top$ by $\widehat{\widetilde{\boldsymbol{x}}}_i = (1, \widehat{\boldsymbol{x}}_i^\top)^\top$, where $\widehat{\boldsymbol{x}}_i$ is the best linear prediction of $\boldsymbol{x}_i$. Thus, the asymptotic covariance matrix is estimated through the asymptotic S-estimator variance (ASV) matrix stated as $\widehat{\text{ASV}}(H) = \widehat{C}(H)^{-1}\widehat{D}(H)\widehat{C}(H)^{-1}$ in [8], where $\widehat{C}(H) = (1/n)\sum_{i=1}^n (w(d_n(\widehat{\boldsymbol{z}}_i)) + 2/(\widehat{\sigma}_{\varepsilon,n}^2 w(d_n(\widehat{\boldsymbol{z}}_i))\widehat{r}_i^2))\widehat{\widetilde{\boldsymbol{x}}}_i\widehat{\widetilde{\boldsymbol{x}}}_i^\top$, $\widehat{D}(H) = (1/n)\sum_{i=1}^n w^2(d_n(\widehat{\boldsymbol{z}}_i))\widehat{r}_i^2\widehat{\widetilde{\boldsymbol{x}}}_i\widehat{\widetilde{\boldsymbol{x}}}_i^\top$, $\widehat{\sigma}_{\varepsilon,n} = (\widehat{S}_{YY} - \widehat{\boldsymbol{\beta}}_{\text{C3S}}^\top \widehat{\boldsymbol{S}}_{XX}\widehat{\boldsymbol{\beta}}_{\text{C3S}})^{1/2}$, $d_n(\widehat{\boldsymbol{z}}_i) = (\widehat{\boldsymbol{z}}_i - \widehat{\boldsymbol{m}})^\top \widehat{\boldsymbol{S}}^{-1}(\widehat{\boldsymbol{z}}_i - \widehat{\boldsymbol{m}})$, $\widehat{r}_i = Y_i - \widehat{\widetilde{\boldsymbol{x}}}_i^\top \widehat{\boldsymbol{\beta}}_{\text{C3S}}$, and $w(d_n(\widehat{\boldsymbol{z}}_i)) = \rho_{\text{R}}(d_n(\widehat{\boldsymbol{z}}_i))$, with ρ_{R} being the Rocke biflat function and d_n defined in Equation (13).

4. Numerical Studies

In this section, the computational framework and simulation scenarios are described. Subsequently, we report the results of an intensive simulation study, which is conducted to evaluate the statistical performance of the C3S-regression coefficient estimators and to compare the proposed estimator and other existing estimators. In addition, the illustration with real data is provided.

4.1. Computational Framework and Simulation Scenarios

Our simulation study is performed by utilizing the `R` software with a Hewlett–Packard HP Compaq computer, Pro 6300 SFF with 8 cores processor GenuineIntel Intel(R) Core(TM) i7-3770 CPU @ 3.40 GHz. The simulation is similar to the one carried out in [8] using the same criteria for evaluating the performance of the C3S-regression. The method proposed in the present paper is also compared with the performance of the LS regression and following two robust alternatives:

- The first one is the 2S-regression [25] that uses an MVE estimate as an initial value. The MVE estimate is computed by means of an iterative subsampling with a concentration step. The MVE estimate is implemented in an `R` package named `rrcov`, using the function `CovSest` with the option `method = "bisquare"` [39].
- The second one is the 3S-regression [8] that reduces the high computational burden of uniform subsampling for the EMVE estimate. The GSE with bisquare ρ function is computed by an iterative algorithm that employs the EMVE-C estimate as an initial value. The 3S-regression without modifications is implemented in an `R` package, named `robreg3S`, using the function `robreg3S` as the default option [40]. Nevertheless, the GSE with the EMVE-C estimate as an initial value is implemented in the `GSE` package, using the function `GSE` with the option `init = "emve_c"` [41].

The univariate filter that is needed by the C3S-regression is implemented in the `robreg3S` package, while the GRE is computed using the `GSE` function with the option `method = "rocke"`. Two versions of the C3S-regression to be considered are: (i) the full version using the comedian covariance matrix; and, (ii) the light version using the raw comedian matrix as an initial scatter estimate. The last one is called light version, because it uses less operations to compute than the full version. From now on, the C3S-regression is referred to both versions, unless that an indication is done.

Next, the regression model presented in Equation (8) with $p = 15$ and $n \in \{150, 300, 500, 1000, 5000\}$ is considered. The values of covariates $\boldsymbol{x}_i$, for $i \in \{1, \ldots, n\}$, are generated from a multivariate normal distribution $\text{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We set $\boldsymbol{\mu} = \boldsymbol{0}$ and $\Sigma_{jj} = 1$ [8], for $j \in \{1, \ldots, p\}$, without loss of generality, because the GSE used by the C3S-regression is location and scale equivariant. (Note that from the location equivariant of the GSE, $\beta_0 = 0$ can be set.) To address the fact that the C3S-regression and 3S-regression are neither affine-equivariant nor regression equivariant, the correlation structure $\boldsymbol{\Sigma}$ may be used. Observe that this correlation structure is described in [27], with the condition number fixed at 100 and random generation of $\boldsymbol{\beta}$ as $\boldsymbol{\beta} = R\boldsymbol{b}$. Let $R = 10$ and $\boldsymbol{b}$ follow a uniform distribution on the unit spherical surface. The response variable Y_i is given by $Y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$, where $\varepsilon_i \sim \text{N}(0, \sigma = 0.5)$, for $i \in \{1, \ldots, n\}$, are independent. The scenarios assumed in the simulation study are:

S1 Clean data: the data generation is not altered.

S2 Cell-wise contamination: randomly replace a proportion q of the cells in the covariates by outliers $x_{ij}^{\text{cont}} = \text{E}(X_{ij}) + k\,\text{SD}(X_{ij})$ and of the responses by outliers $Y_{ij}^{\text{cont}} = \text{E}(Y_{ij}) + k\,\text{SD}(\varepsilon_i)$, where $k \in \{1, \ldots, 10\}$.

S3 Case-wise contamination: randomly replace a proportion q of the cases by leverage outliers $(\boldsymbol{x}_i^{\text{cont}\top}, Y_i^{\text{cont}})$, where $\boldsymbol{x}_i^{\text{cont}} = c\boldsymbol{v}$, and $Y_i^{\text{cont}} = \boldsymbol{x}_i^{\text{cont}}\boldsymbol{\beta} + \varepsilon_i^{\text{cont}}$, with $\varepsilon_i^{\text{cont}} \sim \text{N}(k, \sigma^2)$, for $k \in \{1, \ldots, n\}$. Here, $\boldsymbol{v}$ is the eigenvector corresponding to the smallest eigenvalue of $\boldsymbol{\Sigma}$ with length such that $(\boldsymbol{v} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{v} - \boldsymbol{\mu}) = 1$. To compute the value of $c \in \{1, \ldots, 100\}$, we follow the same process introduced in [8,27]; that is, a Monte Carlo study with the same number of replicates $N = 500$. We observe that $c = 22$ is the value that produces the worst performance of the scatter estimator.

Let $q \in \{0.01, 0.05, 0.09\}$ for the cell-wise contamination. From the fact that case-wise outliers are unusual in practice, we consider $q = 0.03$ for the case-wise contamination. The number of replicates for each setting is $N = 1000$. In addition, the simulation study is also carried out in order to consider the regression model presented by Equation (13) with $p_1 = 12$ continuous covariates, $p_2 = 3$ dummy covariates, and $n \in \{500, 1000\}$. Then, the performance of the M-regression and C3S-regression is evaluated. The values of covariates $(\boldsymbol{x}_i, \boldsymbol{d}_i)$, for $i \in \{1, \ldots, n\}$, are first generated from a multivariate normal distribution $\text{N}_{p_1+p_2}(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is the randomly generated correlation matrix with a fixed condition number of 100. Subsequently, d_{ij} is dichotomized at $\Phi^{-1}(\pi_j)$, with $\pi_j \in \{1/4, 1/3, 1/2\}$, for $j \in \{1, 2, 3\}$, respectively. The generation of the model with continuous and dummy covariates follows the scenarios S1-S2 and for the case-wise contamination follows the scenario S3.

Let $\boldsymbol{\Sigma}_1$ be a sub-matrix of $\boldsymbol{\Sigma}$, which quantifies the covariance of the continuous covariates. In this new scenario, randomly replace a proportion q of the cases in $\mathbb{X}$ by leverage outliers $(\boldsymbol{x}_i^{\text{cont}\top}, Y_i^{\text{cont}})$, where $\boldsymbol{x}_i^{\text{cont}} = c\boldsymbol{v}$, $Y_i^{\text{cont}} = \boldsymbol{x}_i^{\text{cont}}\boldsymbol{\beta}_1 + \boldsymbol{d}_i^\top\boldsymbol{\beta}_2 + \varepsilon_i^{\text{cont}}$, with $\varepsilon_i^{\text{cont}} \sim \text{N}(k, \sigma^2)$, and $k \in \{1, \ldots 10\}$. Here, $\boldsymbol{v}$ is now the eigenvector corresponding to the smallest eigenvalue of $\boldsymbol{\Sigma}_1$, with length such that $(\boldsymbol{v} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{v} - \boldsymbol{\mu}) = 1$, and the corresponding least favorable case-wise contamination size for the twelve continuous variables is $c = 18$.

Once again, we consider $q \in \{0.01, 0.05, 0.09\}$ for the cell-wise contamination and $q = 0.03$ for the case-wise contamination. The number of replicates for each setting is $N = 1000$. Furthermore, the simulation study is conducted for non-normal covariates to compare the performance of the C3S-regression, 3S-regression, 2S-regression and LS estimators. For the C3S-regression, the full and light versions of the proposed estimator are considered. The same regression model with $p = 15$ and $n = 500$ is used, but the covariates are generated from a non-normal distribution [8]. The covariates $\boldsymbol{X}_i$, for $i \in \{1, \ldots, n\}$, are first generated from a multivariate normal distribution with zero mean and covariance matrix $\boldsymbol{\Sigma}$, which is, $\boldsymbol{X}_i \sim \text{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$, where, again, $\boldsymbol{\Sigma}$ is a randomly generated correlation matrix with a fix condition number of 100. Subsequently, the covariates are transformed by means of $(X_{i1}, \ldots, X_{ip}) \leftarrow (G_1^{-1}(\Phi(X_{i1})), \ldots, G_p^{-1}(\Phi(X_{ip})))$. We consider a distribution for G_j as: $\text{N}(0, 1)$, with $j \in \{1, 2, 3\}$; $\chi^2(20)$, with $j \in \{4, 5, 6\}$; $F(90, 10)$, with $j \in \{7, 8, 9\}$; $\chi^2(1)$, with $j \in \{10, 11, 12\}$; and $\text{Pareto}(1, 3)$, with $j \in \{13, 14, 15\}$. The scenarios that are evaluated in this simulation study are as S1. For the cell-wise contamination, we replace $q = 0.05$ by the proportion of cells in the covariates with outliers $x_{ij}^{\text{cont}} = kG_j(0.999)$, and by the proportion of responses with outliers $Y_{ij}^{\text{cont}} = \text{E}(Y_{ij}) + k\,\text{SD}(\varepsilon_i)$.

4.2. Simulation Results

The statistical performance in the estimation of regression coefficients due to the effect of cell-wise and case-wise outliers can be evaluated using the empirical mean squared error (MSE), defined as

$$\overline{\text{MSE}} = \frac{1}{Np} \sum_{i=1}^{N} \sum_{j=1}^{p} (\widehat{\beta}_j^{(i)} - \beta_j^{(i)})^2, \tag{15}$$

where $\widehat{\beta}_j^{(i)}$ is the estimate of $\beta_j^{(i)}$ at the ith Monte Carlo replicate. Tables 2 and 3 report the $\overline{\text{MSE}}$ defined in Equation (15) for $k = 1$ in all the settings with $n \in \{500, 1000\}$. The results for $k = \{5, 10\}$ are omitted, because they are similar to $k = 1$. Figures 2 and 3 show curves of $\overline{\text{MSE}}$ for cell-wise and case-wise contamination in models with $p = 15$ continuous covariates and $n = 1000$.

Table 2. Maximum $\overline{\text{MSE}}$ in all of the considered scenarios for models with continuous covariates.

Estimator	Clean		1% Cell-Wise		5% Cell-Wise		9% Cell-Wise		Case-Wise	
	n = 500	1000	n = 500	1000	n = 500	1000	n = 500	1000	n = 500	1000
C3SFull	0.0037	0.0017	0.0054	0.0026	0.5017	0.4509	1.7417	1.7287	0.0042	0.0019
C3S	0.0037	0.0017	0.0055	0.0026	0.5182	0.4709	1.7999	1.7671	0.0042	0.0019
3S	0.0028	0.0014	0.0105	0.0064	0.8682	0.9009	2.1563	1.9819	0.0033	0.0015
2S	0.0027	0.0014	0.0092	0.0062	3.1863	3.0689	4.3996	4.3861	0.0031	0.0014
LS	0.0026	0.0013	2.3581	2.3459	4.7799	4.7558	5.4603	5.4615	1.3299	1.3141

Table 3. Maximum $\overline{\text{MSE}}$ in all scenarios for models with continuous and dummy covariates.

Estimator	Clean		1% Cell-Wise		5% Cell-Wise		9% Cell-Wise		Case-Wise	
	n = 500	1000	n = 500	1000	n = 500	1000	n = 500	1000	n = 500	1000
C3SFull	0.0029	0.0013	0.0039	0.0020	0.1653	0.1212	1.5194	1.4519	0.0031	0.0015
C3S	0.0028	0.0013	0.0039	0.0020	0.1700	0.1238	1.5572	1.4845	0.0031	0.0015
3S	0.0067	0.0041	0.0109	0.0076	0.5196	0.5874	1.8069	1.7518	0.0077	0.0051
2S	0.0020	0.0010	0.0042	0.0025	1.2884	1.2638	3.8409	3.8522	0.0023	0.0011
LS	0.0018	0.0009	2.4618	2.4173	4.9249	4.9155	5.6736	5.5816	0.5230	0.5037

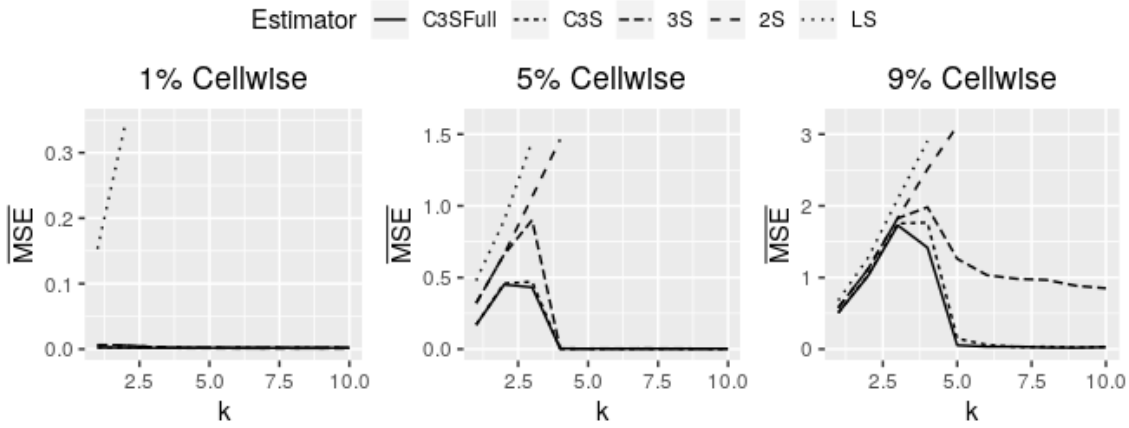

Figure 2. $\overline{\text{MSE}}$ for indicated cell-wise contamination values k in models with $p = 15$ continuous covariates and $n = 1000$. Source: the authors.

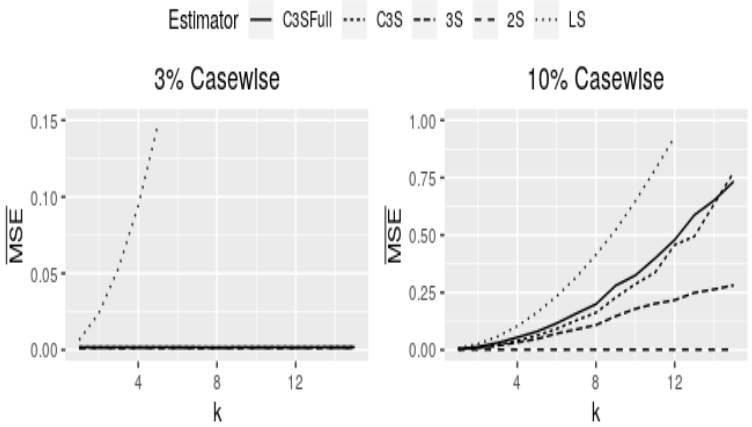

Figure 3. $\overline{\text{MSE}}$ for indicated case-wise contamination values k, in models with $p = 15$ continuous covariates and $n = 1000$. Source: the authors.

Figures 4 and 5 display curves of $\overline{\text{MSE}}$ for cell-wise and case-wise contamination in models with continuous and dummy covariates, and $n \in \{500, 1000\}$. Figure 6 shows curves of $\overline{\text{MSE}}$ for cell-wise and case-wise contamination in models with continuous covariates and $n = 500$. Note that models with continuous covariates in the M-regression and C3S-regression outperform in all of the assumed scenarios for both cell-wise and case-wise contaminations. In addition, in the four panels of Figure 6, both versions of the C3S-regression have almost the same behavior for all settings assumed. The full version of the C3S-regression is little more robust than its light version, but the estimates of both are almost equal for all contamination settings. The results for $n = 1000$ are similar to the cell-wise contamination settings. In the cell-wise contamination setting for small and moderate contamination proportions ($q \leq 0.05$), the C3S-regression is highly robust against moderate and large cell-wise outliers ($k \geq 3$), but less robust against inliers ($k \leq 2$). The 3S-regression and C3S-regression perform similarly for moderate and large outliers, but in the presence of inliers ($k \leq 3$), the 3S-regression is less robust; see first two panels of Figure 6.

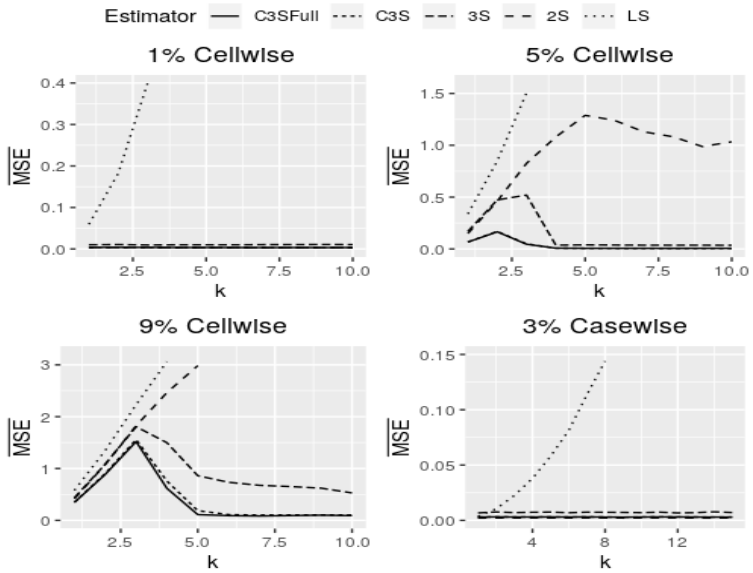

Figure 4. $\overline{\text{MSE}}$ for indicated cell-wise and case-wise contamination values k in models with continuous and dummy covariates, and $n = 500$. Source: the authors.

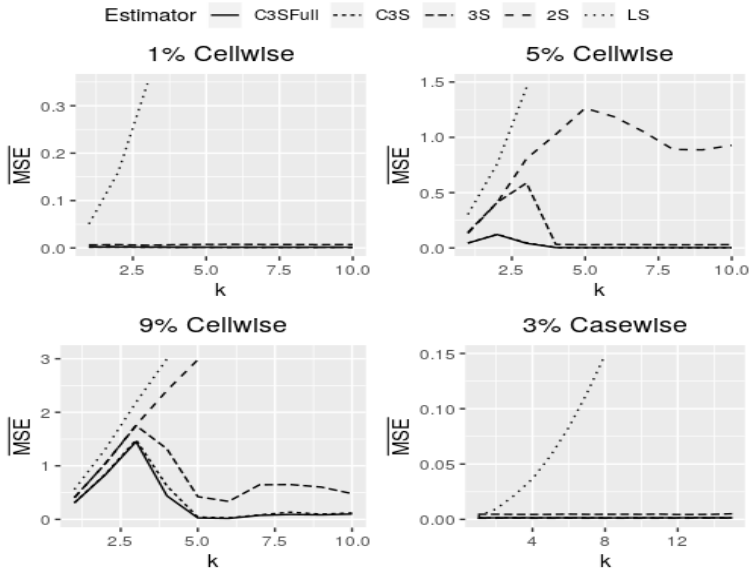

Figure 5. $\overline{\text{MSE}}$ for indicated cell-wise and case-wise contamination values k in models with continuous and dummy covariates and $n = 1000$. Source: the authors.

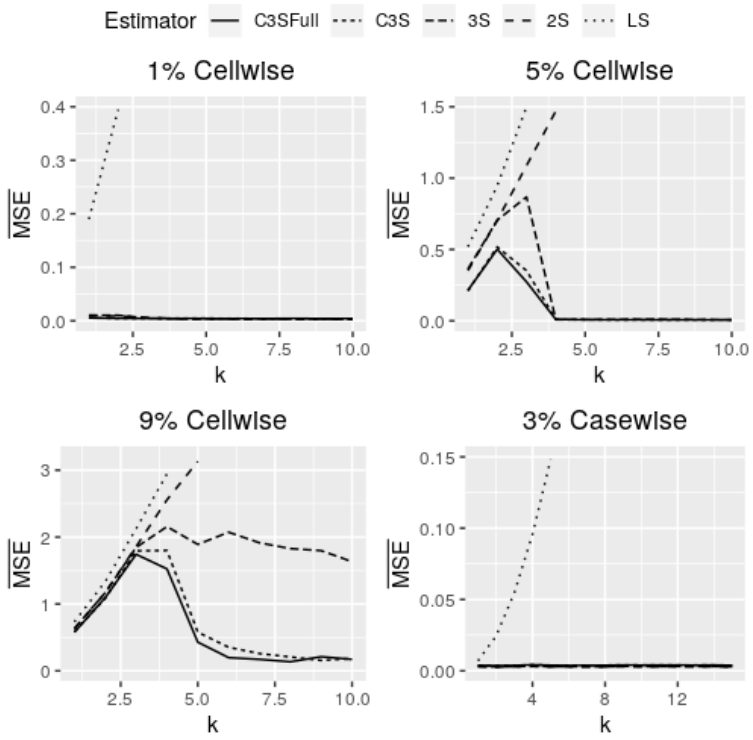

Figure 6. $\overline{\text{MSE}}$ for indicated cell-wise and case-wise contamination values k in models with continuous covariates and $n = 500$. Source: the authors.

The 2S-regression and 3S-regression perform similarly in the presence of inliers, as expected from the simulation studies carried out in [8]. However, the 2S-regression breaks down in cases when the proportion of contaminated cells is $q > 0.5$; that is, when the propagation of large cell-wise outliers is expected to affect more than 50% of the cases.

For a large contamination proportion ($q = 0.09$), the C3S-regression, 3S-regression, and 2S-regression perform similarly in the presence of inliers ($k \leq 3$), but the 3S-regression breaks down for moderate and large cell-wise outliers ($k \geq 4$). However, the C3S-regression is highly robust against large cell-wise outliers ($k \geq 5$) although less robust against moderate outliers. In the case-wise contamination setting, the C3S-regression, 3S-regression and 2S-regression perform fairly well and similarly. Nevertheless, the 2S-regression has the best performance, followed by the 3S-regression, which is followed in performance by the C3S-regression.

We also study the performance of the estimator with moderate and large case-wise contamination levels of 10% and 20%, in which at a size of leverage outliers of 22, the C3S-regression and 3S-regression break down as k increases. In this settings, the C3S-regression outperforms the 3S-regression, but, as expected, the 2S-regression maintains its robustness for any contamination level.

Note that, in practice, it is unusual to find case-wise outliers and even more at moderate or large levels. Thus, the loss of robustness for the C3S-regression and 3S-regression does not present a disadvantage. We detect that models with continuous and dummy covariates in the M-regression and C3S-regression outperform in all assumed scenarios. Table 4 reports a summary of the performance of the estimators evaluated by $\overline{\text{MSE}}$. The performance of the 3S-regression considering non-normal covariates is comparable to all the other estimators for clean data. However, both versions of the C3S-regression outperform all other estimators for any contamination size k in the cell-wise contamination setting. In the cases of non-normal covariates, the C3S-regression maintains its competitive performance, followed by the 3S-regression, while the 2S-regression, as expected, breaks down in the presence of moderate and large cell-wise outliers proportion.

Table 4. $\overline{\text{MSE}}$ for the indicated estimator with clean data and cell-wise contaminated data.

Estimator	Clean	Cell-Wise		
		$k = 1$	$k = 5$	$k = 10$
C3SFull	0.0050	0.0295	0.0180	0.0382
C3S	0.0040	0.0360	0.0173	0.0392
3S	0.0094	0.1712	0.0242	0.0494
2S	0.0011	6.4893	5.3185	5.8407
L2	0.0006	5.2217	6.4807	6.6118

Next, the statistical performance of confidence intervals (CIs) for the regression coefficients based on the asymptotic covariance matrix, as described in Subsection 3.3, is evaluated. The asymptotic $100(1-\tau)\%$ CIs for the coefficients of C3S-regression can be established as

$$\text{CI}(\widehat{\beta}_j) = \left(\widehat{\beta}_j - \Phi^{-1}(1-\tau/2)\sqrt{\widehat{\text{ASV}}(\widehat{\beta}_j)/n}; \widehat{\beta}_j + \Phi^{-1}(1-\tau/2)\sqrt{\widehat{\text{ASV}}(\widehat{\beta}_j)/n}\right), j = 0, 1, \ldots, p. \quad (16)$$

The performance of CIs defined in Equation (16) may be evaluated using the empirical mean coverage rate (CR) given by

$$\overline{\text{CR}} = \frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p} I(\beta_j^{(i)} \in \text{CI}(\widehat{\beta}_j^{(i)})) \quad (17)$$

and the empirical mean CI length (CIL) defined as

$$\overline{\text{CIL}} = \frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p} 2\Phi^{-1}(1-\tau/2)\sqrt{\widehat{\text{ASV}}(\widehat{\beta}_j)/n}. \quad (18)$$

Table 5 reports the average CIL defined in Equation (18) obtained from the C3S-regression and 3S-regression in the case of clean data and contaminated data with 1% cell-wise ($k = 9$), 5% cell-wise ($k = 6$), 9% cell-wise ($k = 3$) and 3% case-wise ($k = 3$), for $n \in \{150, 300, 500, 1000, 5000\}$. The results of the LS and 2S-regression estimates are not included here, because we are interested in comparing the CIL between the 3S-regression and C3S-regression. The CIL that is obtained from the C3S-regression is comparable to that of the 3S-regression for all considered scenarios. The CIL reached from the 3S-regression are shorter than that for the C3S-regression with clean data and data with small and moderate cell-wise contamination levels. For data with large cell-wise contamination levels or case-wise contamination, the CILs of the C3S-regression are shorter than the CILs of the 3S-regression. Moreover, for any assumed scenario, CILs of the 3S-regression and C3S-regression decrease as the sample size n increases.

Figure 7 shows the $\overline{\text{CR}}$ defined in Equation (17) in the case of clean data and contaminated data with 5% cell-wise contamination ($k = 5$), and 3% case-wise contamination ($k = 3$), and for different sample sizes $n \in \{150, 300, 500, 1000\}$. Although the results for the sample size $n = 5000$ are not shown here for visualization, it can be noticed that, for the C3S-regression and 3S-regression, the evaluations of $\overline{\text{CR}}$ under $n = 5000$ are better than those when $n = 1000$. For contamination settings, the 3S-regression yields the best CR, which is the closest to the nominal level. In general, the CR for the C3S-regression is similar to that of the 3S-regression, and it tends to be equal as the sample size n increases.

Table 5. Average CIL for clean data and for cell-wise and case-wise contamination.

Size (n)	Clean		1% Cells, k = 9		5% Cell, k =6		9% Cell, k = 3		3% Cases, k = 3	
	C3S	3S	C3S	3S	C3S	3S	C3S	3S	C3S	3S
150	0.3959	0.3507	0.3818	0.3401	0.4496	0.5196	1.7126	1.8352	0.3957	0.3537
300	0.2824	0.2467	0.2740	0.2414	0.2580	0.3336	1.3654	1.2992	0.2821	0.2498
500	0.2181	0.1915	0.2118	0.1869	0.1917	0.2533	1.1206	1.0047	0.2189	0.1944
1000	0.1543	0.1357	0.1493	0.1323	0.1354	0.1767	0.8204	0.7069	0.1546	0.1375
5000	0.0686	0.0605	0.0670	0.0596	0.0609	0.0785	0.3779	0.3150	0.0694	0.0618

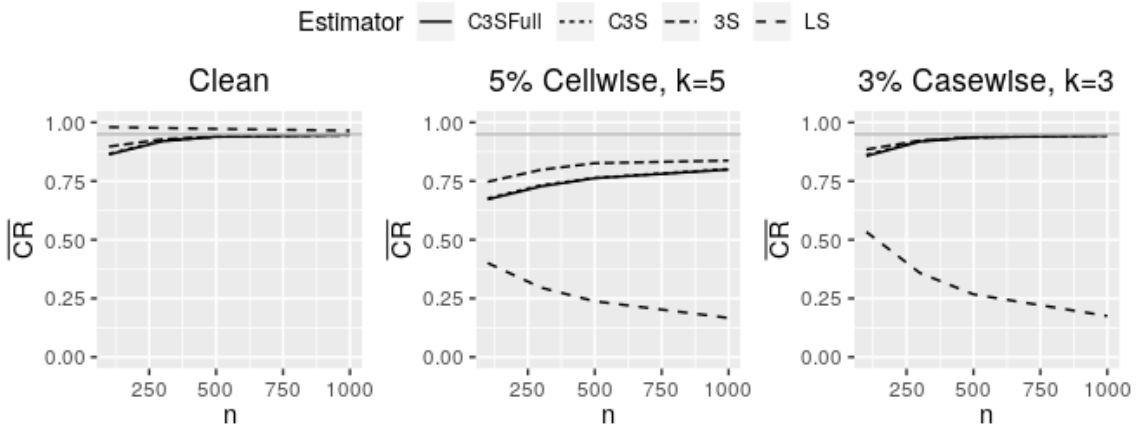

Figure 7. $\overline{\text{CR}}$ for clean data and for cell-wise and case-wise contaminated data with the indicated n. Source: the authors.

4.3. Analysis of Real Data

The airfoil self-noise data set is used for the illustration purpose. These data were obtained from a series of aerodynamic and acoustic tests of two and three-dimensional airfoil blade sections conducted in an anechoic wind tunnel by the NASA. The data set comprises airfoils of different sizes at various wind tunnel speeds and angles of attack with $n = 1503$ observations (cases). For this data set, Table 6 shows five covariates and one response variable along with their statistical summaries. This data set is available at the UCI repository [42]. The aim of this empirical study is to predict the noise generated by an airfoil, from dimensions, speed and angle of attack. Specifically, the objective is to explain the scaled sound pressure level.

Table 6. Description of the variables in the airfoil self-noise data set.

Variable	Label	Units	Type	Minimum	Mean	Maximum
X_1	Frequency	Hertz	Covariate	200	2886.38	20000
X_2	Angle of attack	Degrees	Covariate	0.0000	6.7823	22.2000
X_3	Chord length	Meters	Covariate	0.0254	0.1365	0.3048
X_4	Free stream velocity	Meters	Covariate	31.7000	50.8607	71.3000
X_5	Suction side displacement thickness	Meters	Covariate	0.0004	0.0111	0.0584
Y	Scaled sound pressure level	Decibels	Response	103.38	124.836	140.987

The data set is fitted with the model given by

$$\log(Y_i) = \beta_0 + \beta_1 \log(X_{1_i}) + \beta_2 X_{2_i} + \beta_3 X_{3_i} + \beta_4 X_{4_i} + \beta_5 X_{5_i} + \varepsilon_i, \quad i \in \{1, \ldots, 1503\},$$

where the log function is used for X_1 due to its wide range and high skewness, while the log function is employed for Y in order to improve the R2-adjusted. The corresponding parameters with

C3S-regression (in both versions and full version computed by bootstrap estimation), 2S-regression, 3S-regression, and LS estimates are obtained. The regression coefficient estimates and the corresponding p-values are reported in Table 7. Note that the regression coefficients are similar for all the estimates, except for the covariate X_5, (that is, the suction side displacement thickness). The coefficient of X_5 estimated by 3S-regression and 2S-regression are similar, but are very different from the C3S-regression and LS estimates. For the C3S-regression, X_5 is highly not significant, while for the 2S-regression and 3S-regression, it is only not significant. However, the LS method indicates that X_5 is significant.

Table 7. Estimates and p-values of the regression coefficients for the airfoil self-noise data set.

Variable	C3SFull		C3S		3S		2S		LS	
	Coeff.	*p*-Value	Coeff.	*p*-Value	Coeff.	*p*-Value	Coeff.	*p*-Value	Coeff.	*p*-Value
$\log(X_1)$	−0.0319	<0.0001	−0.0319	<0.0001	−0.0311	<0.0001	−0.0311	<0.0001	−0.0290	<0.0001
X_2	−0.0032	<0.0001	−0.0032	<0.0001	−0.0034	<0.0001	−0.0034	<0.0001	−0.0032	<0.0001
X_3	−0.3299	<0.0001	−0.3299	<0.0001	−0.3026	<0.0001	−0.3026	<0.0001	−0.2828	<0.0001
X_4	0.0006	<0.0001	0.0006	<0.0001	0.0006	<0.0001	0.0006	<0.0001	0.0007	<0.0001
X_5	−0.3008	0.7186	−0.3020	0.7165	−0.8505	0.2110	−0.8561	0.2690	−1.3347	<0.0001

The squared norm distance, defined as $\text{SND} = n\sum_{j=1}^{p}(\widehat{\beta}_{j,A} - \widehat{\beta}_{j,B})^2\text{MAD}(X_{ij},\ldots,X_{nj})^2$, is used to compare the four estimators. Table 8 reports the corresponding SND, which shows that these distances from each two pairs are not large. Therefore, it suggests that the data are not contaminated or the contamination level is very small (inliers).

Table 8. Pairwise squared norm distance between the estimates for the airfoil self-noise data set.

	C3SFull	C3S	3S	2S	LS
C3SFull	-	4.3055×10^{-08}	0.0107	0.0107	0.0389
C3S		-	0.0107	0.0107	0.0389
3S			-	3.0751×10^{-6}	0.0130
2S				-	0.0128
LS					-

5. Conclusions and Future Works

We have provided a new form for robustifying the estimation of parameters of a linear regression model in order to immunize these estimators against case-wise and cell-wise outliers. The main idea here was to modify the generalized Rocke S-estimator in order to obtain robust estimators of the corresponding means and covariances. The difference in our proposal was changing, in the generalized Rocke S-estimator, the initial scatter estimate from the extended minimum volume ellipsoid estimate by the empirical median. The proposed estimator used a univariate filter introduced in the literature and the generalized Rocke S-estimator modified for incomplete data. Our method worked well and similar to that used in the 3S-regression, but in the second step with a different initial robust estimate for the generalized Rocke S-estimator. The initial estimates of location and scatter, the empirical median, and the robust version of the covariance, were computed after snipping the data. Therefore, we have obtained the following findings:

- A new method, called comedian-three-step regression, was proposed, which showed an overall outperformance over the recent developed robust methods.
- An exact correction factor (b_X) was calculated in order to estimate consistently the standard deviation by using the median absolute deviation for the exponential, logistic and uniform distributions. In addition, a numerical solution for this correction factor was introduced in the Student-t and Weibull distributions.

- In continuous covariates, for small contamination proportion and large cell-wise outliers, the 3S-regression performed similarly to the C3S-regression. However, in general, the C3S-regression outperformed the 3S-regression when the cell-wise contamination proportion increase.
- In continuous and dummy covariates, the C3S-regression outperformed both the 3S-regression and 2S-regression, for different contamination proportions. However, for the case-wise outliers, the performance of the three estimators was quite similar.
- The performance of the full version of the C3S-regression estimator proposed in this work was better than its light version. However, the latter one is computationally faster and it can also be used without significant loss of robustness.

Therefore, we have contributed to the robust statistic literature modifying the original three-step regression model by introducing a new family of initial estimates based on the comedian. Our method and the original one are useful to deal with both cell-wise and case-wise outliers. Nevertheless, the numerical results reported that the method proposed in the present study showed an overall outperformance over the recent developed robust methods and a better performance for models with continuous and dummy covariates.

The following aspects derived of this paper may be considered for future work:

- The C3S-regression and 3S-regression estimators work well for cell-wise contamination. However, the performance of these estimators with moderate and large case-wise contamination levels (for example, between 10% and 20%) do not work well when the contamination level increases. Some new kind of shrinkage estimator for the initial scatter estimate should be investigated.
- A bivariate filter can be considered in the first step in order to snip deviation of cells, which could improve the performance of the estimator.
- A numerical procedure must be studied to calculate the correction factor for any distribution.

Author Contributions: Data curation, H.V., H.L. and M.T.; formal analysis, H.L., M.T., V.L. and Y.L; investigation, H.V., H.L., M.T., V.L. and Y.L.; methodology, H.V., H.L., M.T., V.L. and Y.L.; writing–original draft, H.V., H.L., M.T., V.L. and Y.L.; writing–review and editing, H.L., M.T., V.L. and Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: The research was partially supported by the Departamento Administrativo de Ciencia, Tecnología e Innovación (Colciencias), currently Ministerio de Ciencia y Tecnología de Colombia, (Project 7252015), by the Vicerectoría de Descubrimiento y Creación from the Universidad Eafit (H. Velasco, H. Laniado, and M. Toro), and by FONDECYT (grant 1200525) from the National Agency for Research and Development (ANID) of the Chilean government (V. Leiva).

Acknowledgments: The authors thank the Editors and Reviewers for their constructive comments on an earlier version of this manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Draper, N.R.; Smith, H. *Applied Regression Analysis*; Wiley: New York, NY, USA, 2014.
2. Andrews, D.F. A robust method for multiple linear regression. *Technometrics* **1974**, *16*, 523–531. [CrossRef]
3. Liu, Y.; Mao, G.; Leiva, V.; Liu, S.; Tapia, A. Diagnostic analytics for an autoregressive model under the skew-normal distribution. *Mathematics* **2020**, *8*, 693. [CrossRef]
4. Peña, D.; Prieto, F.J. Multivariate outlier detection and robust covariance matrix estimation. *Technometrics* **2001**, *43*, 286–310. [CrossRef]
5. Rousseeuw, P.J.; Leroy, A.M. *Robust Regression and Outlier Detection*; Wiley: New York, NY, USA, 2005.
6. Sánchez, L.; Leiva, V.; Galea, M.; Saulo, H. Birnbaum-Saunders quantile regression models with application to spatial data. *Mathematics* **2020**, *8*, 1000. [CrossRef]
7. Athayde, E.; Azevedo, A.; Barros, M.; Leiva, V. Failure rate of Birnbaum-Saunders distributions: Shape, change-point, estimation and robustness. *Braz. J. Probab. Stat.* **2019**, *33*, 301–328. [CrossRef]
8. Leung, A.; Zhang, H.; Zamar, R. Robust regression estimation and inference in the presence of cell-wise and case-wise contamination. *Comput. Stat. Data Anal.* **2016**, *99*, 1–11. [CrossRef]

9. Leung, A.; Yohai, V.; Zamar, R. Multivariate location and scatter matrix estimation under cell-wise and case-wise contamination. *Comput. Stat. Data Anal.* **2017**, *111*, 59–76. [CrossRef]
10. Falk, M. On MAD and comedians. *Ann. Inst. Stat. Math.* **1997**, *49*, 615–644. [CrossRef]
11. Di Palma, M.A.; Gallo, M. A co-median approach to detect compositional outliers. *J. Appl. Stat.* **2016**, *43*, 2348–2362. [CrossRef]
12. Alqallaf, F.; VanAelst, S.; Yohai, V.J.; Zamar, R.H. Propagation of outliers in multivariate data. *Ann. Stat.* **2009**, *37*, 311–331. [CrossRef]
13. Hampel, F.R. Beyond location parameters: Robust concepts and methods. *Bull. Int. Stat. Inst.* **1975**, *46*, 375–382.
14. Rousseeuw, P.J. Least median of squares regression. *J. Am. Stat. Assoc.* **1984**, *79*, 871–880. [CrossRef]
15. Rousseeuw, P.J. Multivariate estimation with high breakdown point. *Math. Stat. Appl.* 1985, 8, 283–297.
16. Rousseeuw, P.J.; Driessen, K.V. A fast algorithm for the minimum covariance determinant estimator. *Technometrics* **1999**, *41*, 212–223. [CrossRef]
17. Holl, P.W.; Welsch, R.E. Robust regression using iteratively reweighted least-squares. *Commun. Stat. Theory Methods* **1977**, *6*, 813–827.
18. Wager, T.D.; Keller, M.C.; Lacey, S.C.; Jonides, J. Increased sensitivity in neuroimaging analyses using robust regression. *Neuroimage* **2005**, *26*, 99–113. [CrossRef]
19. Rousseeuw, P.; Yohai, V. *Robust Regression by Means of S-Estimators*; Springer: New York, NY, USA, 1984.
20. Yohai, V.J. High breakdown-point and high efficiency robust estimates for regression. *Ann. Stat.* **1987**, *20*, 642–656. [CrossRef]
21. Leiva, V.; Sanhueza, A.; Sen, P.K.; Araneda, N. M-procedures in the general multivariate nonlinear regression model. *Pak. J. Stat.* **2010**, *26*, 1–13.
22. Sanhueza, A.; Sen, P.K.; Leiva, V. A robust procedure in nonlinear models for repeated measurements. *Commun. Stat. Theory Methods* **2009**, *38*, 138–155. [CrossRef]
23. Maronna, R.; Morgenthaler, S. Robust regression through robust covariances. *Commun. Stat. Theory Methods* **1986**, *15*, 1347–1365. [CrossRef]
24. Davies, P.L. Asymptotic behaviour of s-estimates of multivariate location parameters and dispersion matrices. *Ann. Stat.* **1987**, *15*, 1269–1292. [CrossRef]
25. Croux, C.; VanAelst, S.; Dehon, C. Bounded influence regression using high breakdown scatter matrices. *Ann. Inst. Stat. Math.* **2003**, *55*, 265–285. [CrossRef]
26. Danilov, M.; Yohai, V.J.; Zamar, R.H. Robust estimation of multivariate location and scatter in the presence of missing data. *J. Am. Stat. Assoc.* **2012**, *107*, 1178–1186. [CrossRef]
27. Agostinelli, C.; Leung, A.; Yohai, V.J.; Zamar, R.H. Robust estimation of multivariate location and scatter in the presence of cell-wise and case-wise contamination. *TEST* **2015**, *24*, 441–461. [CrossRef]
28. Öllerer, V.; Alfons, A.; Croux, C. The shooting s-estimator for robust regression. *Comput. Stat.* **2016**, *31*, 829–844. [CrossRef]
29. Fu, W.J. Penalized regressions: The bridge versus the lasso. *J. Comput. Graph. Stat.* **1998**, *7*, 397–416.
30. Gervini, D.; Yohai, V.J. A class of robust and fully efficient regression estimators. *Ann. Stat.* **2002**, *30*, 583–616. [CrossRef]
31. Farcomeni, A. Robust constrained clustering in presence of entry-wise outliers. *Technometrics* **2014**, *56*, 102–111. [CrossRef]
32. Rousseeuw, P.J.; Croux, C. Alternatives to the median absolute deviation. *J. Am. Stat. Assoc.* **1993**, *88*, 1273–1283. [CrossRef]
33. Efron, B. Bootstrap Methods: Another Look at the Jackknife. *Ann. Stat.* **1979**, *7*, 1–26. [CrossRef]
34. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2017.
35. Rocke, D.M. Robustness properties of s-estimators of multivariate location and shape in high dimension. *Ann. Stat.* **1996**, *24*, 1327–1345. [CrossRef]
36. Maronna, R.A.; Martin, D.R.; Yohai, V.J. *Robust Statistics: Theory and Methods*; Wiley: New York, NY, USA, 2006.
37. Huber, P.J.; Ronchetti, E.M. *Robust Statistics*; Wiley: Hoboken, NJ, USA, 2009.
38. Maronna, R.A.; Yohai, V.J. Robust regression with both continuous and categorical predictors. *J. Stat. Plan. Inference* **2000**, *89*, 197–214. [CrossRef]

39. Todorov, V.; Filzmoser, P. An object-oriented framework for robust multivariate analysis. *J. Stat. Softw.* **2009**, *32*, 1–47. [CrossRef]
40. Leung, A.; Zhang, H.; Zamar, R. *robreg3S: Three-Step Regression and Inference for Cellwise and Casewise Contamination*; R Package Version 0.3; R Foundation for Statistical Computing: Vienna, Austria, 2015.
41. Leung, A.; Danilov, M.; Yohai, V.J.; Zamar, R. *GSE: Robust Estimation in the Presence of Cellwise and Casewise Contamination and Missing Data*; R Package Version 4.1; R Foundation for Statistical Computing: Vienna, Austria, 2016.
42. Dheeru, D.; Karrataniskidou, E. *UCI Machine Learning Repository*; University of California: Irvine, CA, USA, 2017.

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).