Betsabé Pérez Garrido

Exploring the performance of the GPSC method under several levels of outliers

Betsabé Pérez Garrido, Associate professor, Corvinus University of Budapest, Department of Computer Science, Hungary
Email: perez.betsabe@uni-corvinus.hu

The aim of this study is to evaluate the performance of the Groupwise Principal Sensitivity Components method which is a robust iterative procedure for fitting linear regression models with fixed group effects. A simulation study is carried out to assess its ability to detect multiple outliers located in the response variable (vertical outliers) or in the explanatory and response variable (high leverage outliers). Several levels of outliers are considered ranging from 5% to 45% within selected groups. The results suggest that the GPSC method is able to avoid the masking effect under low or moderate level of outliers -approximately below to 30%. Additionally, in almost all cases the GPSC method reports lower levels of false outlier detection under high leverage outliers.

Keywords: linear regression model with fixed effects, outlier detection, robust method, swamping effect, masking effect

Linear regression models have been widely used in many fields, such as the agricultural sector *(Dhulipala-Patil, 2020)* or marketing research *(Dumitrescu et al., 2012)*. The classical approach for fitting linear regression models is the least squares (LS) method, where the idea is minimizing the sum of squared residuals., It is well known, however, that estimators via LS can be highly affected by the presence of outliers *(Hadi–Simonoff, 1993; Molina et al., 2009; Pérez, 2011)*, which may appear in the response variable (vertical outliers) or in the explanatory and response variable (high leverage outliers).

Several proposals have been introduced in the literature to mitigate the effect of outliers, for instance, the *M* estimates *(Huber, 1981)*, the *S* estimates *(Rousseeuw–Yohai, 1984)* or the *MM* estimates *(Yohai, 1987)*. This study focuses on the Groupwise Principal Sensitivity Components (GPSC) method proposed by *Pérez et al. (2014),* which is a robust iterative procedure for linear regression models with fixed group effects. The GPSC method is based on an iterative

procedure for finding and removing potential outliers. Then, LS estimates are calculated based on a clean subset.

Detecting multiple outliers may be a challenging task due to the masking and swamping phenomena effects. The masking effect occurs when an outlier is not detected due to the presence of other outliers. The swamping effect occurs when a nonoutlier is erroneously considered an outlier due to the effect of some other hidden outliers *(Hadi–Simonoff, 1993).*

The goal of the present work is to explore the performance of the GPSC method in detecting multiple outliers located in the response variable (vertical outliers) or in the explanatory and response variable (high leverage outliers). In particular, it aims to analyse two main issues: (i) its capacity to detect all outliers – avoiding the masking effect – and (ii) its ability to minimize false outlier detection – the swamping effect. Considering the nature of the linear regression models with fixed group effects, several levels of outliers were artificially generated ranging from 5% to 45% within selected groups.

The work is structured as follows: Section 1 introduces the linear regression model with group effects. Section 2 presents the groupwise principal sensitivity components method. Section 3 presents the results of a simulation study. Section 4 concludes.

1. The model

Consider X a vector of continuous covariates, $X = (X_1, X_2, \dots, X_p)^T$ with $p \geq 1$, related to the variable of interest Y. Suppose there are n sample observations of X and Y coming from D different population groups of sizes $n_1, \dots, n_d$ with $n_d \geq 2$ for $d = 1, \dots, D$. Let n be the total number of observations where $n = \sum_{d=1}^{D} n_d$ D being a fixed value. The linear regression model with fixed group effects can be defined as

$$y_{dj} = \boldsymbol{x}_{dj}^T \boldsymbol{\beta} + \alpha_d + \varepsilon_{dj}, \text{ for } j = 1, \dots, n_d, \quad d = 1, \dots, D, \tag{1}$$

where y_{dj} is the variable of interest associated with the j-th sample unit within group d; $\boldsymbol{x}_{dj} = (x_{dj1}, \dots, x_{djp})^T$ is the vector of the p covariates for the same unit within group d; $\boldsymbol{\beta}$ is a p-vector of unknown parameters; α_d is the effect of the d-th group assumed to be fixed; and ε_{dj} is the model error with distribution $\varepsilon_{dj} \sim$ iid $N(0, \sigma^2)$, for $j = 1, \dots, n_d$, $d = 1, \dots, D$, where $\sigma^2 > 0$ is unknown.

Considering the group structure of the data, Model (1) can be rewritten as:

$$\boldsymbol{y}_d = \boldsymbol{X}_d \boldsymbol{\beta} + \alpha_d \mathbf{1}_{n_d} + \boldsymbol{\varepsilon}_d, \quad \text{for } d = 1, \dots, D, \tag{2}$$

where $\boldsymbol{y}_d = (y_{d1}, .., y_{n_d})^T$, $\boldsymbol{X}_d = (\boldsymbol{x}_{d1}, \dots, \boldsymbol{x}_{dn_d})^T$, $\boldsymbol{\varepsilon}_d = (\varepsilon_{d1}, \dots, \varepsilon_{dn_d})^T$ being $\boldsymbol{\varepsilon}_d \sim N(\boldsymbol{0}_{n_d}, \sigma^2 \boldsymbol{I}_{n_d})$. Expressions $\boldsymbol{0}_{n_d}$ and $\boldsymbol{1}_{n_d}$ represent the vectors of size n_d of zeros and ones; $\boldsymbol{I}_{n_d}$ is the $n_d \times n_d$ identity matrix. The LS estimators of $\boldsymbol{\beta}$ and α_d are given by

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{S}_X^{-1} \boldsymbol{s}_{XY}, \quad \hat{\alpha}_d = \bar{y}_d - \bar{\boldsymbol{x}}_d^T \widehat{\boldsymbol{\beta}}, \quad \text{for } d = 1, \dots, D,$$

where $\bar{\boldsymbol{x}}_d = (\bar{x}_{d1}, \dots, \bar{x}_{dp})^T$ is $\bar{x}_{dq}$ the mean of the q-th covariate X_q within group d, for $q = 1, \dots, p$; $\bar{y}_d = \frac{1}{n_d} \sum_{j=1}^{n_d} y_{dj}$ for $d = 1, \dots, D$,

$\boldsymbol{S}_X = \frac{1}{n} \sum_{d=1}^{D} \sum_{j=1}^{n_d} (\boldsymbol{x}_{dj} - \bar{\boldsymbol{x}}_d)(\boldsymbol{x}_{dj} - \bar{\boldsymbol{x}}_d)^T$, $\boldsymbol{s}_{XY} = \frac{1}{n} \sum_{d=1}^{D} \sum_{j=1}^{n_d} (\boldsymbol{x}_{dj} - \bar{\boldsymbol{x}}_d)(y_{dj} - \bar{y}_d)$.

The predicted value of the j-th unit within group d is given by $\hat{y}_{dj} = \boldsymbol{x}_{dj}^T \widehat{\boldsymbol{\beta}} + \hat{\alpha}_d$, for $j = 1, \dots, n_d$, $d = 1, \dots, D$, and the leverage effect can be written as $h_{jj}^d = \frac{1}{n_d} + (\boldsymbol{x}_{dj} - \bar{\boldsymbol{x}}_d)^T (n\boldsymbol{S}_X)^{-1} (\boldsymbol{x}_{dj} - \bar{\boldsymbol{x}}_d)$, for $d = 1, \dots, D$.

1.1 Sensitivity vectors

Peña–Yohai (*1999)* proposed sensitivity vectors for detecting high leverage outliers under linear regression models. The extension of the sensitivity vectors under linear regression models with fixed group effects was proposed by *Pérez et al. (2014)*. In the last case, considering the group structure of the data, the sensitivity vector $\boldsymbol{r}_{dj}$ associated with the j-th sample unit within group d can be defined as the vector of changes in the predicted value $\hat{y}_{dj}$ when each observation within group d is deleted, that is,

$$\boldsymbol{r}_{dj} = (\hat{y}_{dj} - \hat{y}_{dj(d1)}, \cdots, \hat{y}_{dj} - \hat{y}_{dj(dn_d)})^T.$$

The sensitivity matrix of group d can be written as

$$\boldsymbol{R}_d = \begin{pmatrix} \hat{y}_{d1} - \hat{y}_{d1(d1)} & \cdots & \hat{y}_{d1} - \hat{y}_{d1(dn_d)} \\ \vdots & \ddots & \vdots \\ \hat{y}_{dn_d} - \hat{y}_{dn_d(d1)} & \cdots & \hat{y}_{dn_d} - \hat{y}_{dn_d(dn_d)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{r}_{d1} \\ \vdots \\ \boldsymbol{r}_{dn_d} \end{pmatrix}. \tag{3}$$

The calculation of (3) can be computationally demanding as the number of observations increases within each group., This calculation, however, can be simplified using the fact that $\boldsymbol{R}_d = \boldsymbol{H}_{dd} \boldsymbol{W}_d$, where the hat matrix associated with *the* d-th group is given by $\boldsymbol{H}_{dd} = \left(h_{jk}^d\right)_{j,k=1,\dots,n_d} = \delta \widehat{\boldsymbol{y}}_d / \delta \boldsymbol{y}_d^T$ and $\boldsymbol{W}_d = \text{diag}_{1 \le j \le n_d} \left\{ \frac{e_{dj}}{1 - h_{jj}^d} \right\}$. Note that matrix $\boldsymbol{R}_d$ has rank $p+1$.

High leverage outliers are expected to appear as extreme values on at least one of the principal sensitivity components of matrix $\boldsymbol{M}_d = \boldsymbol{R}_d^T \boldsymbol{R}_d$, where λ_k^d and $\boldsymbol{v}_k^d$ denote the k-th eigenvalue and eigenvector of group d, and $\boldsymbol{z}_q^d = \boldsymbol{R}_d \boldsymbol{v}_q^d$ are the projections on the directions $\boldsymbol{v}_q^d$ for $q = 1, \dots, p+1$.

2. The GPSC method

The GPSC method is a robust procedure for fitting linear regression models with fixed group effects. The method is a trade-off between robustness and efficiency, computed in two stages. The first stage is iterative, it discards high- and low-leverage outliers within each group, providing an approximate *S* estimator of the regression parameters. In the second stage, the efficiency of the *S* estimator is improved by a reweighting procedure based on robust *t* tests. The GPSC method works as follows:

Stage 1 (iterative). In the first iteration, denoted as $r = 1$, it creates a set A_1 of candidate estimates of $\boldsymbol{\gamma} = (\beta^T, \alpha_1, \dots, \alpha_D)$ by calculating the sensitivity matrix $\boldsymbol{R}_d$ for each group $d = 1, \dots, D$ and computing the projections $\boldsymbol{z}_q^d$, $q = 1, \dots, p+1$. For each component *q* and group *d,* it considers two different datasets: (i) including all observations from the group and (ii) deleting 50% of the observations with the largest coordinates in the vector $\boldsymbol{d}_q^d = |\boldsymbol{z}_q^d - \text{median}(\boldsymbol{z}_q^d)|$. Considering *q=1,..,p+1* components and *d=1,..,D* groups, we obtain $2^D(p+1)$ potentially clean samples. Then, the LS estimators and their respective residuals of each potential sample are calculated. The selected estimate, $\boldsymbol{\gamma}^{(1)}$, is obtained by minimizing a robust scale of the residuals, that is, satisfying

$$\boldsymbol{\gamma}^{(1)} = \text{argmin}_{\boldsymbol{\gamma} \epsilon A_1} \; s\left(e_{11}(\boldsymbol{\gamma}), \dots, e_{Dn_d}(\boldsymbol{\gamma})\right),$$

where *s* is an *M*-scale estimator with a high breakdown point, such as the median absolute deviation (MAD). Let $\boldsymbol{\gamma}^{(\mathrm{r})} = \left((\beta^{(r)})^T, \alpha_1^{(r)}, \dots, \alpha_D^{(r)}\right)^{\boldsymbol{T}}$ be the selected estimate in iteration *r*. In the following iteration, *r+1,* the residuals associated with $\boldsymbol{\gamma}^{(\mathrm{r})}$, are obtained, that is,

$$e_{dj}^{(r+1)} = e_{dj}\left(\boldsymbol{\gamma}^{(\mathrm{r})}\right) = y_{dj} - \boldsymbol{x}_{dj}^T\boldsymbol{\beta}^{(r)} - \alpha_d^{(r)}, \text{ for} j = 1, \dots, n_d, \quad d = 1, \dots, D.$$

Let $s_d^{(r+1)} = s\left(e_{d1}^{(r+1)}, \dots, e_{dn_d}^{(r+1)}\right)^T$ be a robust scale for the *d*-th group, such as the normalized MAD. Then, in each group, all observations with $\left|e_{dj}^{(r+1)}\right| \geq C_1 \cdot s_d^{(r+1)}$ are eliminated being C_1 a constant, with the remaining observations from the *D* groups computing the LS estimators and again computing the principal sensitivity components. Construct the set of A_{r+1} with the new set of candidate estimates $\boldsymbol{\gamma}$ exactly as described before but including in the set the estimator obtained from the previous iteration$\boldsymbol{\gamma}^{(\mathrm{r})}$, too. The iterative process ends when $\boldsymbol{\gamma}^{(\mathrm{r+1})} \approx \boldsymbol{\gamma}^{(\mathrm{r})}$, and the resulting estimator is denoted as $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}^{(\mathrm{r+1})} = (\boldsymbol{\beta}^{*T}, \alpha_1^*, \dots, \alpha_D^*)^{\boldsymbol{T}}$, which is an approximate *S* estimator.

Stage 2 (improving efficiency). In the second stage, the efficiency of the robust preliminary estimator $\boldsymbol{\gamma}^*$ is improved. First, the residuals are calculated using the preliminary estimator $\boldsymbol{\gamma}^*$

$$e_{dj}^* = e_{dj}(\boldsymbol{\gamma}^*) = y_{dj} - \boldsymbol{x}_{dj}^T\boldsymbol{\beta}^* - \alpha_d^*, \text{ for } j = 1, \dots, n_d, \quad d = 1, \dots, D.$$

Let $s_d^* = s\left(e_{d1}^*, \ldots, e_{dn_d}^*\right)^T$ be a robust scale for the d-th group, such as the normalized MAD. For each group, eliminate all observations with $\left|e_{dj}^*\right| \geq C_2 \cdot s_d^*$, where C_2 is a constant. Let n^* be the total number of deleted observations, with the remaining $n - n^*$ observations compute the LS estimators and denote them as $\tilde{\beta}$ and $\tilde{\alpha}_d$, $d = 1, \ldots, D$. Additionally, compute the standard error $\tilde{\sigma}$ using the residuals of the remaining observations and the corresponding leverages $\tilde{h}_{jj}^{\,d}$.

Test the outlyingness of each of these n^* observations using the robust t test statistic

$$t_{dj} = \frac{y_{dj} - \boldsymbol{x}_{dj}^T \tilde{\boldsymbol{\beta}} - \tilde{\alpha}_d}{\tilde{\sigma}\sqrt{1 + \tilde{h}_{jj}^{\,d}}}$$

Each of the n^* observations will only be eliminated definitely if $\left|t_{dj}\right| \geq C_3$, where C_3 is a constant. Using the remaining observations, calculate the final LS estimator denoted as $\hat{\boldsymbol{\gamma}}^* = \left(\hat{\boldsymbol{\beta}}^{*T}, \hat{\alpha}_1^*, \ldots, \hat{\alpha}_D^*\right)^{\boldsymbol{T}}$. The recommended constants providing a good trade-off between robustness and efficiency are $C_1 = 2$ and $C_2 = C_3 = 3$.

3. Simulation study

A simulation study has been carried out to assess the performance of the GPSC method in identifying multiple outliers located in the response variable (vertical outliers) or in the explanatory and response variable (high leverage outliers). The goal is the analysis of two main issues: (i) its capacity to detect all true outliers (avoiding the masking effect) and (ii) its ability to minimize false outlier detection (swamping effect).

The simulated data are similar to those proposed by *Pérez et al. (2014).* The present study, however, considers more levels of outliers ranging from 5% to 45% within selected groups.

The data contain $D = 5$ groups, where the number of observations within each group is $(n_1, n_2, n_3, n_4, n_5) = (20,30,40,50,60)$. The total number of observations is $n = \sum_{d=1}^{D=5} n_d = 200$. The vector of continuous covariates is $X = (X_1, X_2, X_3, X_4)^T$, where $X_1 \sim N(3.31,0.82)$, $X_2 \sim N(1.74,1.10)$, $X_3 \sim N(1.70,1.28)$ and $X_4 \sim N(2.41,1.61)$. The true values of the regression parameters are $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)^T = (0.45,0.14,0.05,0.005)^T$. The fixed effects, α_d, were generated from $N(0,1)$ for $d = 1, \ldots, D = 5,$, and the errors ε_{dj} were produced independently from $N(0,0.1)$. The group effects and covariates were fixed during the 1000 Monte Carlo replicates, and in each replicate, the variable of interest y_{dj} was generated using Model (1).

Multiple outliers were artificially created in the response variable (vertical outliers) or in the explanatory and response variable (high leverage outliers). The mechanisms of generating the outliers are:

Type 1: Multiple outliers in the response variable (vertical outliers). Three groups (out of five) were selected for contamination, specifically groups d = 1, 3 and 5. In these groups, multiple outliers were generated by replacing a percentage of observations (5%, 10%, 15%, 20%, 25%, 30%, 35%, 40% or 45%) by $c_{d1} = \bar{y}_d + k \cdot s_{Y,d}$, where $\bar{y}_d$ and $s_{Y,d}$ are the mean and standard deviationof the generated clean observations within group d and considering k = 5, 10 or 15.

Type 2: Multiple outliers in the explanatory and response variable (high leverage outliers). Groups d = 1, 3 and 5 were again selected for contamination. In these groups, a percentage of observations (5%, 10%, 15%, 20%, 25%, 30%, 35%, 40% or 45%) in covariates X_3 and X_4 were replaced by $c_{d2} = \bar{x}_{dq} + k \cdot s_{X,d}$, where $\bar{x}_{dq}$ and $s_{X,d}$ are the mean and standard deviation of the generated clean observations within group d and considering q = 3 and 4 and considering k = 5, 10 or 15. Then, their corresponding variables of interest were contaminated as described in *Type 1*. In the case of their corresponding X_1 and X_2, unusual observations were introduced in the joint distribution of (X_1, X_2), but these were not outliers according to our definition.

The performance of the GPSC method was assessed using four criteria: (i) the percentage of simulations where all outliers were detected, that is, the ability of the GPSC method to avoid the masking effect; (ii) the average of false outliers detected over the simulations, which considered to be useful in evaluating the swamping effect; (iii) the empirical mean squared error (MSE) of the final estimator defined as $\text{MSE}(\widehat{\boldsymbol{\gamma}}) = \frac{1}{L}\sum_{l=1}^{L}\left\|\left(\widehat{\boldsymbol{\gamma}}^{(l)} - \boldsymbol{\gamma}\right)\right\|^2$ and (iv) the empirical median squared error (MNSE) defined as $\text{MNSE}(\widehat{\boldsymbol{\gamma}}) = \text{median}\left\{\left\|\left(\widehat{\boldsymbol{\gamma}}^{(l)} - \boldsymbol{\gamma}\right)\right\|^2\right\}$.

4. Conclusion

The performance of the GPSC method is reported in Tables 1 to 4. According to Table 1, the percentage of simulations where all outliers were detected is above 88% in the case of levels of contamination from 5% to 30%, considering both types of contamination (Type 1 and 2), and when the constant used for contamination is k = 5, 10 or 15. This percentage decreases from the level of contamination of 35%, and reports the lowest value at 45% contamination. The results suggest that the GPSC method is able to avoid the masking effect, particularly under low or moderate levels of contamination (approximately below 30% of outliers within groups).

Table 2 presents the average of false outliers detected over the simulations. This measure was introduced to assess the swamping effect. According to these results, the average of false outliers is almost similar in the case of Type 1 and Type 2 when the level of contamination ranges from 5% to 20% and the constant used for contamination is k = 5, 10 or 15. Moreover, in almost all cases, the GPSC method reports lower levels of false outlier detection in the case of contamination Type 2 (high leverage outliers).

Table 3 reports the empirical mean squared error of the final estimator. The results show that the mean squared error is almost similar in the case of levels of contamination from 5% to 20%. However, from the level of contamination 25%, the mean squared error is noticeably higher in the case of k = 10 and 15 (compared with k = 5). Additionally, from the level of contamination of 30%, the mean squared error is always higher in the case of Type 1 vertical outliers (compared with Type 2 high leverage outliers).

Table 4 presents the empirical median squared error of the final estimator. This value ranges from 0.45 to 0.65 when the level of contamination ranges from 5% to 30%. In particular, this value breaks at the 40% level of contamination in the case of Type 2 (high leverage outliers) and at 45% of contamination in the case of Type 1 (vertical outliers).

In future research, it would be interesting to test more configurations for the simulations, for instance, by increasing or decreasing the number of groups, increasing or decreasing the selected groups for contamination, etc. Another interesting issue could be the improvement of the recommended constants to obtain a better trade-off between robustness and efficiency, currently established as $C_1 = 2$ and $C_2 = C_3 = 3$.

Table 1

Percentage of simulations where all outliers were detected

(%)

Constant	Type 1	Type 2	Type 1	Type 2	Type 1	Type 2	Type 1	Type 2	Type 1	Type 2
	5%		10%		15%		20%		25%	
k = 5	100.0	98.2	100.0	98.8	99.9	99.5	99.1	99.1	97.1	99.0
k = 10	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.5	99.9
k = 15	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.7	99.9
	30%		35%		40%		45%			
k = 5	88.5	96.9	60.2	78.5	17.7	32.8	0.4	2.2		
k = 10	96.0	97.9	83.9	64.1	58.7	18.9	24.8	0.9		
k = 15	97.0	96.2	87.0	60.2	61.1	13.9	28.5	1.1		

Table 2

Average of false outliers detected over the simulations

Constant	Type 1	Type 2	Type 1	Type 2	Type 1	Type 2	Type 1	Type 2	Type 1	Type 2
	5%		10%		15%		20%		25%	
$k = 5$	0.438	0.451	0.390	0.357	0.294	0.285	0.219	0.218	0.206	0.206
$k = 10$	0.430	0.443	0.378	0.356	0.295	0.278	0.229	0.235	0.252	0.194
$k = 15$	0.435	0.431	0.376	0.372	0.294	0.283	0.212	0.226	0.229	0.18
	30%		35%		40%		45%			
$k = 5$	0.715	0.166	4.030	0.160	10.75	0.100	9.104	0.072		
$k = 10$	0.671	0.166	3.301	0.153	10.003	0.137	19.331	0.123		
$k = 15$	0.571	0.158	2.497	0.130	9.327	0.160	19.054	0.159		

Table 3

Empirical mean squared error of the final estimator

Constant	Type 1	Type 2	Type 1	Type 2	Type 1	Type 2	Type 1	Type 2	Type 1	Type 2
	5%		10%		15%		20%		25%	
$k = 5$	0.75	0.84	0.76	0.80	0.81	0.81	0.8	0.87	0.90	1.00
$k = 10$	0.75	0.82	0.77	0.79	0.81	0.81	0.77	0.83	1.88	1.45
$k = 15$	0.75	0.82	0.77	0.80	0.81	0.81	0.78	0.84	1.44	1.26
	30%		35%		40%		45%			
$k = 5$	8.29	2.36	62.26	15.21	215.97	41.19	449.63	51.28		
$k = 10$	27.19	10.16	215.99	165.06	757.33	351.36	1953.53	411.74		
$k = 15$	51.22	36.47	360.16	484.52	1581.68	1043.41	3929.68	1173.67		

Table 4

Empirical median squared error of the final estimator

Constant	Type 1	Type 2	Type 1	Type 2	Type 1	Type 2	Type 1	Type 2	Type 1	Type 2
	5%		10%		15%		20%		25%	
$k = 5$	0.46	0.50	0.46	0.46	0.49	0.48	0.45	0.50	0.50	0.52
$k = 10$	0.47	0.48	0.47	0.47	0.49	0.47	0.45	0.48	0.48	0.52
$k = 15$	0.45	0.49	0.46	0.47	0.50	0.47	0.45	0.48	0.49	0.52
	30%		35%		40%		45%			
$k = 5$	0.61	0.65	1.03	0.83	30.17	32.61	253.66	42.20		
$k = 10$	0.57	0.63	0.71	1.42	1.60	336.38	1661.94	379.36		
$k = 15$	0.56	0.63	0.65	1.62	1.45	1015.51	3577.29	1098.02		

References

Dhulipala, S. – Patil, G. R. (2020): Freight production of agricultural commodities in India using multiple linear regression and generalized additive modelling. *Transport Policy.* Vol. 97. pp. 245–258. https://doi.org/10.1016/j.tranpol.2020.06.012

Hadi, A. S. – Simonoff, J. S. (1993): Procedures for the identification of multiple outliers in linear models. *Journal of the American Statistical Association*. Vol. 88. No. 424. pp.1264–1272. https://doi.org/10.1080/01621459.1993.10476407

Huber, P. J. (1981): *Robust statistics*. Wiley, New York.

Molina, I. – Peña, D. – Pérez, B. (2009): Robust estimation in linear regression models with fixed effects. *Universidad Carlos III de Madrid. Working papers*. No. 09–88 (27). https://core.ac.uk/download/pdf/30041693.pdf

Peña, D. – Yohai, V. (1999): A fast procedure for outlier diagnostics in large regression problems. *Journal of the American Statistical Association.* Vol. 94. No. 446 pp. 434–445. https://doi.org/10.1080/01621459.1999.10474138

Pérez, B. (2011): *Robust estimation and outlier detection in linear models for grouped data.* Doctoral Thesis. Universidad Carlos III de Madrid. Madrid.

PÉREZ, B. – Molina, I. – Peña, D. (2014): Outlier detection and robust estimation in linear regression models with fixed group effects. *Journal of Statistical Computation and Simulation.* Vol. 84. No. 12. pp. 2652–2669. https://doi.org/10.1080/00949655.2013.811669

Rousseeuw, P. – Yohai, V. (1984): Robust regression by means of S-estimators. In: Franke, J. – Härdle, W. – Martin, D. (eds.): *Robust and Nonlinear Time Series Analysis.* Proceedings of a Workshop Organized by the Sonderforschungsbereich 123 "Stochastische Mathematische Modelle", Heidelberg. pp. 256–272.

Yohai, V. (1987): High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics.* Vol. 15. No. 2. pp. 642–656. https://doi.org/10.1214/aos/1176350366

Dumitrescu, L. – Stanciu, O. –Tichindelean, M. – Vinerean, S. (2012): The use of regression analysis in marketing research. *Studies in Business & Economics*. Vol. 7. No. 2.