Outlier Robustification: A Comparative Study of Extreme Point Removal and Censorship

Alison Borenstein, Austin Zaccor

December 2019

1 Problem Statement

Outliers present a serious issue for parametric regression formulations such as Ordinary Least Squares (OLS). However, various formulations of the regression problem can reduce or even eliminate the impact that extreme points have on the estimators. Based on the presentation of the Least Median of Squares (LMS) estimator, our aim is to investigate the performance, in terms of both robustness and efficiency, of this method alongside Mixed Integer Optimization (MIO) formulations of the Trimmed Ordinary Least Squares (LTS) estimator and a modified version of LTS we call Least Trimmed Summed Residuals (LTSR), as well as a new formulation we call Bounded Influence Least Squares (BILS).

The common motivation behind each of these methods is to find a regression model that is immune to outliers. Outliers arise from various sources (*e.g.*, human error, system malfunctions, unintended data mutation, natural deviations) that are likely to present themselves in different ways in data. We would like to evaluate the performance of these formulations under various outlier conditions knowing the ground truth, so we leverage synthetic datasets. Given that the goal of our investigation is to better understand how each of these formulations reacts to outliers, we will vary the percentage of perturbed points (outlier density) and experiment with the presence of outliers in the covariates as well as in the response variable. By using an optimized version of one of these four alternative formulations of the OLS estimator, one can obtain a high-quality solution robust to datasets with extreme outliers. The question of which one tends to perform the best in the real-world we think is an important question for anyone encountering outlier contaminated data. Therefore, we also evaluate the performance of these methods on a real world dataset on the toxicity of various chemicals to *Daphnia Magna*, a planktonic crustacean. Robustness and tractability considerations will need to be made when evaluating final performance for each estimator.

2 Importance

The Ordinary Least Squares estimator is one of the most widely used statistical methods; however, it relies on assumptions that are often not met by the datasets to which it is applied. Specifically, the OLS estimator assumes the error terms to be independently and identically distributed, and the error terms in the population to be approximately normally distributed. When data does not satisfy these underlying assumptions, it is important to use a more robust modeling method.

Outlier robust alternatives to OLS include the LMS and LTS estimators. The LMS estimator finds regression coefficients that minimize the median residual. Alternatively, the LTS estimator minimizes the sum of squared residuals over an outlier-removed subset of all observations. Both methods seek to prevent the most influential points from dominating the coefficients. In order to further investigate methods of outlier detection, it is important to first distinguish between the various properties of an observation.

In a Cartesian plane, an observation's *leverage* (Equation 1), h_{ii}, is a measure of how far the x value deviates from $\bar{x}$, while *influence* (Equation 2), δ_i, is a measure of how much the observation affects the slope of the regressor. An outlying observation may appear as a result of an outlier in either the predictor variable, the response variable, or both. It is important to note, however, that a point of high leverage might not be an outlying point in the sense that its influence is not large enough to affect the slope of the regression equation. High influence observations, on the other hand, tend to have both high residuals (*i.e.*, outliers) and have high leverage. As shown in Figure 1, the OLS estimator is greatly affected by the presence of influential extreme points.

$$h_{ii} = x_i^T (X^T X)^{-1} x_i \tag{1}$$

$$\delta_i = \frac{r_{i,OLS}^2}{\sigma_{OLS}^2} \cdot \frac{1}{p} \cdot \frac{h_{ii}}{(1-h_{ii})^2} \quad , \tag{2}$$

$$\text{where } \sigma_{OLS}^2 = \frac{(y - X\beta_{OLS})^T (y - X\beta_{OLS})}{n - p - 1} \tag{3}$$

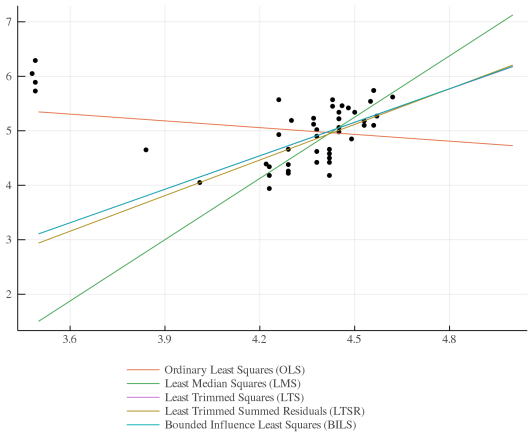

Figure 1: Fitted Regressions on Hertzsprung-Russell Star Data

In the univariate case (Figure 1), identifying outliers is relatively straightforward. However, most real world datasets are multivariate, so the problem of detecting outliers becomes much more difficult. Furthermore, as the number of outliers increases, the measures of central tendency are pulled toward the outliers, leading to a masking effect where it becomes even more difficult to detect outlying observations. Several statistical methods have been developed to identify high leverage (Figure 2) or high influence (Figure 3) points for multivariate datasets.

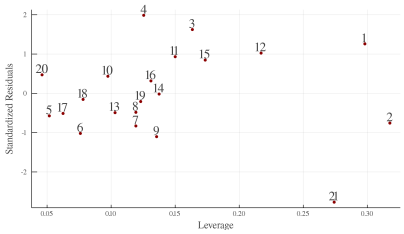

Figure 2: Standardized Residuals vs. Leverage on Brownlee Stackloss Data

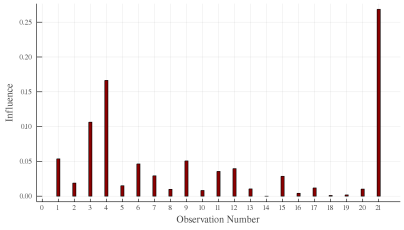

Figure 3: Influence of Observations on Brownlee Stackloss Data

In the next section we leverage Cook's distance (a metric designed to quantify an observation's influence) to present a quadratic MIO formulation that seeks to bound the influence of otherwise compromising outliers while parametrizing the trade-off between avoiding these high influence outliers and including as many of the inliers as possible.

3 Data

3.1 Benchmark Datasets

There are several widely studied outlier datasets in the field of robust regression. Prior investigators (*e.g.*, Andrews, (1974); Rupert and Carrol, (1980); Rousseeuw and Leroy, (1987)) have used these datasets to test the performance of numerous proposed robust regression methods. We will use two of these datasets to evaluate our formulations, namely, the "Hertzsprung-Russell Star Dataset" and the "Stackloss Dataset" [Brownlee, (1965)].

3.1.1 Hertzsprung-Russell Star Data

The Hertzsprung-Russell Star Dataset contains 47 observations, each of which corresponds to stars of the CYG OB1 cluster in the direction of the constellation Cygnus. The predictor variable is the logarithm of the effective surface temperature of the star (T_e), and the response variable y is the logarithm of the light intensity of the star (L/L_0). The dataset is interesting with regard to outliers because there are four high influence observations, namely 11, 20, 30, and 34, that correspond to giant stars.

3.1.2 Brownlee Stackloss Data

The Stackloss Dataset is multivariate, containing three predictor variables, each of which corresponds to measurements for a plant that oxidizes ammonia to nitric acid. The three covariate measurements are (i) air flow to the plant, (ii) cooling water inlet temperature, and (iii) acid concentration. The response variable is the stackloss, or amount of ammonia lost. Rousseeuw and Leroy (1987) stated that prior research has led to the conclusion that observations 1, 3, 4, and 21 are outliers.

3.2 Synthetic Data

The robust estimators are also evaluated on synthetic data with varying levels of outlier contamination, where outliers are randomly selected from among all x_{ij} and y_i. For all synthetic datasets, the x_{ij} values are drawn from the standard normal distribution $\mathcal{N}(0,1)$. Values for β are chosen such that they are approximately centered around 0 with a similar amount of spread to those used in the background literature on this topic. Response variables are then the sum of two components. The first component, $y = X\beta$, is completely deterministic, while the second component is 0 for inlying points and a draw from the uniform distribution $\mathcal{U}(6,10)$ for outliers. Once y is calculated, we create our outliers in X by adding a draw from the uniform distribution $\mathcal{U}(4,6)$.

4 Methods

4.1 Least Quantile Squares (LQS)

$$Min\ \gamma$$

$$s.t.\ r_i = y_i - \beta_0 - \beta_p' X_i,\ \forall i = 1...n$$

$$a_i \geq r_i,\ \forall i = 1...n$$

$$a_i \geq\ -r_i,\ \forall i = 1...n$$

$$\gamma \geq a_i - \mu_i,\ \forall i = 1...n$$

$$\mu_i \leq M(1 - z_i),\ \forall i = 1...n$$

$$\sum_{i=1}^{n} z_i = q$$

$$\mu_i \geq 0,\ \forall i = 1...n$$

$$z_i \in \{0,1\}$$

The first constraint establishes r_i as the i^{th} residual resulting from the standard OLS regression equation, and then sets a_i equal to the absolute value of this residual in the next two constraints. The fourth constraint states that the objective value γ must be greater than or equal to the absolute value of the residual less some value μ where μ_i is zero if the i^{th} observation is included and γ is unconstrained if it is excluded. We also ensure that exactly q observations are included, where $q = \frac{n}{2}$ if n is odd, and $\frac{n}{2} - 1$ if n is even, then we have recovered the Least Median Squares (LMS) regression. Thus, minimizing γ is equivalent to minimizing the subset q of residuals that are included in the regression.

4.2 Least Trimmed Squares (LTS)

$$Min \sum_{i=1}^{n} r_i^2$$

$$s.t.\ \beta_0 + \beta_p' X_i - y_i \leq r_i + M z_i,\ \forall i = 1...n$$
$$-\beta_0 - \beta_p' X_i + y_i \leq r_i + M z_i,\ \forall i = 1...n$$
$$\sum_{i=1}^{n} z_i \leq k$$
$$r_i \geq 0,\ \forall i = 1...n$$
$$z_i \in \{0, 1\}$$

This formulation simply minimizes the sum of $n - k$ residuals. A binary variable, z_i, is introduced to determine which points will be deleted from the set of observations. If z_i is 1, observation i will not be included, and the first two constraints combined with the non-negativity constraint will force the residual, r_i, to equal zero. If z_i is 0, the residual value is r_i. The third constraint allows for up to k points to be deleted. With this formulation, which simultaneously minimizes over z and β, up to $n - k$ of the smallest residuals will be selected for fitting the model and the remaining observations will not be considered when determining the values of the coefficients.

From a robustification standpoint, there is one main drawback to using the LTS estimator: the formulation is extremely sensitive to the k parameter. Drastically different fits may be obtained depending on the number of points that are to be excluded. Without *a priori* knowledge of the number of outliers in a dataset, which is typically not realistic, we need sufficient data for cross validation to obtain a satisfactory value for k. However, as we mentioned, the random variation involved in splitting the dataset is sufficient in some cases to perturb the estimator quite severely.

4.3 Least Trimmed Summed Residuals (LTSR)

$$Min \sum_{i=1}^{n} r_i$$

$$s.t.\ \beta_0 + \beta_p' X_i - y_i \leq r_i,\ \forall i = 1...n$$
$$-\beta_0 - \beta_p' X_i + y_i \leq r_i,\ \forall i = 1...n$$
$$\sum_{i=1}^{n} z_i \leq k$$
$$r_i \geq 0,\ \forall i = 1...n$$
$$z_i \in \{0, 1\}$$

The formulation above is similar in nature to the Least Trimmed Squares formulation, with a modified objective function. The new objective function minimizes the sum of a subset of residuals, determined using an equivalent process as that of the LTS estimator. We present this formulation with the intention to discover whether or not it is more computationally efficient than the other methods we evaluate.

4.4 Bounded Influence Least Squares (BILS)

$$Min\ \alpha \cdot \sum_{i=1}^{n}(r_i^2 + z_i\delta^2) + (1-\alpha)\cdot \sum_{i=1}^{n} z_i$$

$$s.t.\ \beta_0 + \beta_p' X_i - y_i \leq r_i + Mz_i,\ \forall i = 1...n$$

$$-\beta_0 - \beta_p' X_i + y_i \leq r_i + Mz_i,\ \forall i = 1...n$$

$$\sum_{i=1}^{n} z_i \leq (n-p-1)/2$$

$$r_i \geq 0,\ \forall i = 1...n$$

$$z_i \in \{0,1\}$$

The Bounded Influence formulation can be thought of as an extension to the LTS formulation. In LTS, k is a user-specified parameter meant to capture the actual number of potential outliers in the dataset. BILS does not require the number of possible outliers to be passed in. Instead, we present an objective function that captures the trade-off between including observations with high influence and excluding observations from the model. The first term in the new objective function minimizes the sum of the squared residuals plus the squared influence for observations selected as outliers. The second term minimizes the total number of outliers selected (and thus removed from the model). The α parameter represents the trade-off between these objectives. The objective is formulated such that with $\alpha = 0$, the optimal solution will be for the sum of z_i to equal zero, recovering OLS. If $\alpha = 1$, LTS with $k = (n-p-1)/2$ will be recovered by removing the maximum number of observations, before reaching the breakdown point of the estimator.

5 Results

5.1 Benchmark Dataset Performance

5.1.1 Hertzsprung-Russell Star Data

Returning to the Hertzsprung-Russell Star Data, we see in Table 1 the β recovery results and the outliers detected for each of the robust estimators formulated above, as well as the OLS estimator for comparison.

Outlying Points: 11, 20, 30, 34			
—	β_0	β_1	Outliers Detected
OLS	6.793	-0.413	—
LMS	-11.624	3.750	—
LTS	-4.057	2.046	11, 20, 30, 34
LTSR	-4.686	2.178	20, 30, 34
BILS	-4.070	2.051	11, 20, 30, 34

Table 1: Regression Coefficients and Outliers Detected for Hertzsprung-Russell Star Data

Interpreting the output of the OLS estimator would lead one to the inaccurate conclusion that light intensity of a star decreases as surface temperature increases. By contrast, all four of the robust methods return the opposite sign for β_1, and the relationship between light intensity and surface temperature can be correctly recovered. Each of the robust regression methods detects at least 3 of the 4 known outliers. The LTSR estimator fails to detect observation 11, which of the four outlying points has the least amount of influence.

5.1.2 Brownlee Stackloss Data

Outlying Points: 1, 3, 4, 21					
—	β_0	β_1	β_2	β_3	Outliers Detected
OLS	-39.920	0.716	1.295	-0.152	—
LMS	-34.468	0.709	0.335	0.012	—
LTS	-37.652	0.780	0.577	-0.067	1, 3, 4, 21
LTSR	-35.941	0.822	0.437	-0.070	1, 3, 4, 21
BILS	-37.652	0.780	0.577	-0.067	1, 3, 4, 21

Table 2: Regression Coefficients and Outliers Detected for Brownlee Stackloss Data

Recall from Figure 2 that observation 2 has the highest leverage of any point. However, because the residual is comparatively low, observation 2 has much lower influence than the four outliers correctly identified by each of the robust estimators above, as is illustrated in Figure 3.

5.2 Synthetic Dataset Performance

Reported below are the β estimates, mean squared error (MSE), and runtime for each of the formulations solved to optimality, unless otherwise indicated by an (*).

Synthetic dataset (SDS-1): n = 40, p = 5, 12% outlier density									
—	β_0	β_1	β_2	β_3	β_4	β_5	β_6	MSE	Runtime (s)
True Signal	0	0.3	-0.6	1	0.7	-0.5	-	-	-
OLS	0.06	0.19	-0.37	0.50	0.21	-0.24	-	0.560	0.001
LMS	0.04	0.17	-0.64	0.96	0.67	-0.49	-	0.043	86.481
LTS	0.05	0.21	-0.61	0.97	0.75	-0.50	-	0.035	0.310
LTSR	0.06	0.17	-0.61	0.97	0.75	-0.44	-	0.038	0.303
BILS	0.05	0.21	-0.60	0.96	0.76	-0.50	-	0.035	0.552

Table 3: Signal Recovery and Runtime for Various Models on SDS-1

In synthetic dataset 1 (SDS-1), we can see that in the presence of even 12% outlier contamination Gauss-Markov theorem has been sufficiently violated such that OLS is no longer the best linear unbiased estimator. By contrast, each of the outlier robust estimators are able to recover most of true signal with a high degree of accuracy, reducing MSE by over an order of magnitude.

Synthetic dataset (SDS-2): n = 40, p = 5, 24% outlier density									
—	β_0	β_1	β_2	β_3	β_4	β_5	β_6	MSE	Runtime (s)
True Signal	0	0.3	-0.6	1.3	0.7	-0.7	-	-	-
OLS	0.17	0.15	-0.45	0.45	0.16	-0.22	-	0.96	0.001
LMS	-0.04	0.05	-0.76	1.23	0.73	-0.62	-	0.109	54.153
LTS	0.04	0.19	-0.65	1.32	0.76	-0.66	-	0.053	1.740
LTSR	0.03	0.18	-0.62	1.32	0.77	-0.64	-	0.053	2.398
BILS	0.04	0.19	-0.65	1.32	0.76	-0.67	-	0.055	1.182

Table 4: Signal Recovery and Runtime for Various Models on SDS-2

Synthetic dataset 2 (SDS-2) further supports the conclusions we began drawing above. In the presence of 24% contamination, the extreme points that render OLS inaccurate have much less effect on the four outlier-robust estimators' ability to recover the true signal.

Synthetic dataset (SDS-3): n = 80, p = 6, 36% outlier density									
—	β_0	β_1	β_2	β_3	β_4	β_5	β_6	MSE	Runtime (s)
True Signal	0	0.3	-0.4	0.6	0.4	-0.6	0.6	-	-
OLS	0.19	0.86	-0.31	0.48	0.22	-0.34	0.32	0.678	0.002
LMS	0.04	0.28	-0.45	0.55	0.40	-0.61	0.60	0.041	180.000*
LTS	0.05	0.31	-0.41	0.58	0.37	-0.61	0.62	0.040	180.000*
LTSR	0.05	0.30	-0.40	0.58	0.37	-0.61	0.61	0.040	180.000*
BILS	0.05	0.31	-0.40	0.58	0.37	-0.61	0.62	0.041	180.000*

Table 5: Signal Recovery and Runtime for Various Models on SDS-3

Synthetic dataset 3 (SDS-3) could not be solved to optimality within a reasonable amount of time, so a cutoff of 180 seconds was imposed, as the BILS formulation had once solved on SDS-3 in that amount of time during trials. Although far more efficient, OLS continues to suffer from outlier contamination much more than the outlier-robust estimators.

Based on our computational experiments, we confirm that OLS is by far the most computationally efficient formulation and that LMS is many times more computationally expensive than the other methods, quickly becoming intractable as we scale n and p. Interestingly, BILS is faster than the other formulations in SDS-2 but not in SDS-1. Another interesting result is that LTSR often selects less than the upper bound of $(n-p-1)/2$ observations for removal, as some observations have high leverage but low influence, and thus do not materially affect the values of the coefficients. Therefore, retaining these observations lowers the objective cost function more than dropping them.

5.3 Outlier Detection and False Positive Rate

Method	SDS-1	SDS-2	SDS-3
LTS	5/5	7/9	17/24
LTSR	4/5	6/9	17/24
BILS	5/5	7/9	18/24

Table 6: Outliers Detected

In Table 6 we see that as the contamination percentage increases, the number of outliers detected decreases. SDS-1 contains outliers in covariates only, while SDS-2 and SDS-3 contain outliers in covariates as well as in the response variable. In SDS-2 and SDS-3, all three methods fail to detect the outliers in the response variables, which make up most of the false positive rate for these datasets.

We further investigate the BILS formulation with respect to the sensitivity of the trade-off parameter, α. We vary the level of contamination on a dataset of $n = 40$, $p = 5$, and record the number of non-outlying observations that are identified by the estimator as outliers to generate false positive rates for increasing values of α. We find that, in general, α values greater than 0.7 tend to result in non-outlying points being removed.

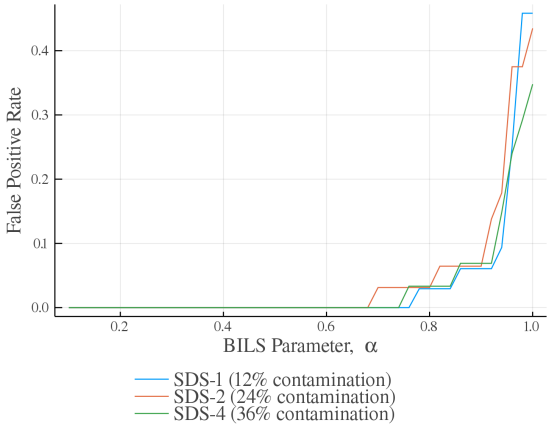

Figure 4: False Positive Rate of BILS Outlier Detection by α

5.4 Comparison of Methods on a Real World dataset

Although we demonstrated that outlier-robust estimators outperform OLS in terms of MSE on our synthetic datasets, we wished to see how these results would extend to real data. We utilize the QSAR Aquatic Toxicity dataset from the UCI Machine Learning Repository to predict the LC50 for a particular species of zooplankton in the presence of different chemicals. Basic exploratory analysis confirms the presence of outliers in both the dependent and explanatory variables; however, since the *true* amount of outlier contamination is unknown, we utilize a validation set to select the hyperparameter k in LTS and LTSR, as well as the hyperparameter α in BILS before obtaining the out of sample MSEs reported in Table 7.

Aquatic Toxicity: n = 80, p = 6		
Method	MSE	Runtime (s)
OLS	0.93	0.002
LMS	0.85	180.000*
LTS	0.84	2.522
LTSR	0.85	6.958
BILS	0.85	31.335

Table 7: MSE and Runtime by Model on QSAR Aquatic Toxicity Dataset

As expected, the outlier-robust estimators obtain consistently lower out of sample MSE values than OLS, though the relative improvement is much smaller than it was in the synthetic data experiments. In this dataset, it may be that there was not as much room to improve over OLS as in the syntheitc datasets. Additionally, BILS was not as efficient compared to LTS or LTSR as we hoped after observing the runtimes in SDS-2.

6 Conclusions

In the pursuit of outlier-robust regression, we compared four MIO formulations that either censor or ignore outliers when calculating fit. We have shown that these formulations are highly effective at recovering the true signal and reducing out of sample error on synthetic datasets.

When outlier density is low (12%), our experiments show that the trimmed estimators are able to correctly identify outliers with high accuracy. As the percentage of contamination increases, however, the trimmed estimators successfully identify a lower percentage of the outliers. The BILS estimator does the best, and is still able to identify 75% of the outliers in the dataset even with a large percentage of contamination (36%). For the BILS estimator, the fact that outliers in the response variables go undetected reflects the formulation's design to seek high influence observations. Since an outlier in the response variable cannot be high leverage, it makes it much more difficult to identify it as high influence. Alternatively, if the value has high leverage, but a low residual, it will not be identified as an outlier or an influential observation.

Another important evaluative criterion for these formulations is their efficiency. Despite the fact that no model comes close to the scalability of OLS, we show that LMS is much less efficient than the other three outlier-robust methods. We did not experiment with large-scale datasets in this project, but it is worthwhile to note that heuristic methods could be used to provide a warm start for our MIO formulations.

Although they require tuned hyperparameters, another advantage of LTS, LTSR, and BILS is that observations identified as outliers or high influence can be returned to the modeler for further inspection. We believe that this transparency has important implications and adds value when working with real world data. The ability to separate out the outlying observations could lead to discoveries about a significant sub-population in the data, or may reveal potential errors in how the data is generated or recorded. In both scenarios it is helpful to view outliers separately and decide how best to proceed with modeling (*e.g.*, through removal, imputation, etc.).

Overall, we are cautiously optimistic about our findings. In testing on real world data, we found significant reductions to out of sample MSE for all four of our outlier-robust models, though not to the same extent as on our synthetic datasets. Further testing on real data is required to determine the true scalability of BILS, as well as the extent of its practical usability on outlier contaminated datasets. Still, the results recovered in both our synthetic and real world testing are encouraging for the future performance and relevance of this new model.

Appendix

Code

`https://github.com/azaccor/Outlier_Robust_Regression_Models`

Least Trimmed Squares (Alternative Formulation)

In addition to our final formulation for Least Trimmed Squares, we also developed a formulation that sorts the residuals for each observation. Early results from computational experiments revealed the inefficiency of this method; thus, this formulation was not included in the estimator comparisons. However, we present it here to show yet another way to formulate the trimmed estimator.

$$Min \sum_{i=1}^{n}\sum_{j=1}^{s} u_{ij}$$

$$s.t. \ -\beta_0 - \beta_p' X_i + y_i \leq r_i + M(1-z_i), \ \forall i = 1...n$$

$$-\beta_0 - \beta_p' X_i + y_i \geq r_i - M(1-z_i), \ \forall i = 1...n$$

$$\beta_0 + \beta_p' X_i - y_i \leq r_i + Mz_i, \ \forall i = 1...n$$

$$\beta_0 + \beta_p' X_i - y_i \geq r_i - Mz_i, \ \forall i = 1...n$$

$$\sum_{i=1}^{n} a_{ik} = 1, \ \forall k = 1...n$$

$$\sum_{k=1}^{n} a_{ik} = 1, \ \forall i = 1...n$$

$$u_{ij} \geq -Ma_{ij}, \ \forall i = 1...n, \forall j = 1...n$$

$$u_{ij} \leq Ma_{ij}, \ \forall i = 1...n, \forall j = 1...n$$

$$r_i - M(1-a_{ij}) \leq u_{ij}, \forall i = 1...n, \forall j = 1...n$$

$$r_i \geq u_{ij}, \ \forall i = 1...n, \forall j = 1...n$$

$$r_i \geq 0, \ \forall i = 1...n$$

$$z_i \in \{0, 1\}$$

Hertzsprung-Russell Star Data

Table 8: Table of Observations: Outliers in Bold

Obs.	x	y
1	4.37	5.23
2	4.56	5.74
3	4.26	4.93
4	4.56	5.74
5	4.30	5.19
6	4.46	5.46
7	3.84	4.65
8	4.57	5.27
9	4.26	5.57
10	4.37	5.12
11	**3.49**	**5.73**
12	4.43	5.45
13	4.48	5.42
14	4.01	4.05
15	4.29	4.26
16	4.42	4.58
17	4.23	3.94
18	4.42	4.18
19	4.23	4.18
20	**3.49**	**5.89**
21	4.29	4.38
22	4.29	4.22
23	4.42	4.42
24	4.49	4.85
25	4.38	5.02
26	4.42	4.66
27	4.29	4.66
28	4.38	4.90
29	4.22	4.39
30	**3.48**	**6.05**
31	4.38	4.42
32	4.56	5.10
33	4.45	5.22
34	**3.49**	**6.29**
35	4.23	4.34
36	4.62	5.62
37	4.53	5.10
38	4.45	5.22
39	4.53	5.18
40	4.43	5.57
41	4.38	4.62
42	4.45	5.06
43	4.50	5.34
44	4.45	5.34
45	4.55	5.54
46	4.45	4.98
47	4.42	4.50

Stackloss Data

Table 9: Table of Observations: Outliers in Bold

Obs.	x_1	x_2	x_3	y
1	**80**	**27**	**89**	**42**
2	80	27	88	37
3	**75**	**25**	**90**	**37**
4	**62**	**24**	**87**	**28**
5	62	22	87	18
6	62	23	87	18
7	62	24	93	19
8	62	24	93	20
9	58	23	87	15
10	58	18	80	14
11	58	18	89	14
12	58	17	88	13
13	58	18	82	11
14	58	19	93	12
15	50	18	89	8
16	50	18	86	7
17	50	19	72	8
18	50	19	79	8
19	50	20	80	9
20	56	20	82	15
21	**70**	**20**	**91**	**15**

References

[1] D. Bertsimas and R. Mazumder. "Least Quantile Regression Via Modern Optimization". In: The Annals of Statistics, **42** (2014), pp. 2494–2525.

[2] Brownlee, K. A. (1960, 2nd ed. 1965) Statistical Theory and Methodology in Science and Engineering. New York: Wiley. pp. 491–500.

[3] A. Giloni and M. Padberg. "Least Trimmed Squares Regression, Least Median Squares Regression, and Mathematical Programming". In: Mathematical and Computer Modelling, **35** (2002), pp. 1043–1060.

[4] Cedric E. Ginestet. "Detecting Outliers and Influence Analysis". In: MA 575 Linear Models: Boston University(Jan. 2010), pp. 20–24.

[5] D. J. Hand, F. Daly, A. D. Lunn, K. J. McConway and E. Ostrowski (1994). A Handbook of Small Datasets, Chapman and Hall/CRC, London.

[6] P.J. Rousseeuw and A.M Leroy, (1987), Robust Regression and Outlier Detection, Wiley. New York.

[7] P.J. Rousseeuw, M. Hubert, and S. Van Aelst. "High-Breakdown Robust Multivariate Methods". In: Statistical Science, **23** (2008), pp. 92–119.

[8] R. Todeschini. "Prediction of acute aquatic toxicity towards daphnia magna using GA-kNN method". In: Alternatives to Laboratory Animals (ATLA), **42** (2014), pp. 31–41.

[9] G. Zioutas and A. Avramidis. "Deleting Outliers in Robust Regression with Mixed Integer Programming". In: Acta MathematicaeApplicatae Sinica, **21** (2005), pp. 323–334.