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

Alison Borenstein and Austin Zaccor
MIT Sloan School of Management

Motivation

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), our aim is to investigate the robustness and efficiency of this method alongside Mixed Integer Optimization (MIO) formulations of the Least Trimmed 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). Given that outliers can arise from different sources (*e.g.*, human error, system malfunctions, unintended data mutation), we evaluate the performance of these formulations under different outlier conditions utilizing synthetic data, which allows us to know the true signal a priori.

Background

In a Cartesian plane, an observation's *leverage*, h_{ii}, is a measure of how far the x value deviates from $\bar{x}$, while *influence*, δ_i, is a measure of how much the observation affects the slope of the regressor. Thus, high influence observations tend to have both high residuals (*i.e.*, outliers) and have high leverage. As shown by Figure 1, the OLS estimator is greatly affected by the presence of influential extreme points.

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

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

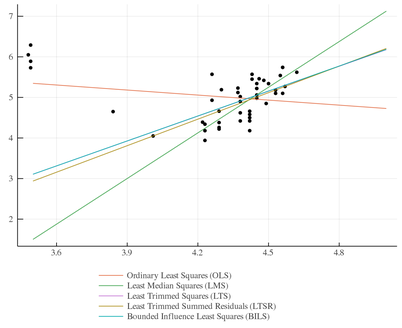

Fig. 1: Fitted Regressions on Hertzsprung-Russell Star Data

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 the univariate case (Figure 1), identifying outliers is relatively straightforward. However, most real world data sets are multivariate, and so the problem of detecting outliers becomes much more difficult. Several statistical methods have been developed to identify high leverage (Figure 2) or high influence (Figure 3) points for multivariate data sets.

Leveraging Cook's distance, a metric designed to quantify an observation's influence, we 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 (see MIO Formulations).

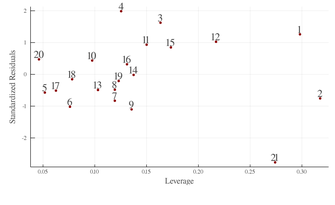

Fig. 2: Residuals vs. Leverage on Brownlee Stackloss Data

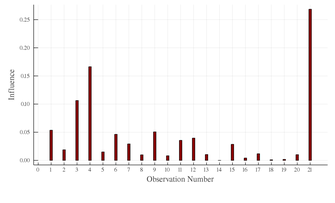

Fig. 3: Influence of Observations on Brownlee Stackloss Data

MIO Formulations

The Least Trimmed Squares (LTS) formulation below simultaneously minimizes over z and β and selects up to $n-k$ of the smallest residuals for fitting the model; the remaining observations are not considered when determining the values of the coefficients.

In the Bounded Influence Least Squares (BILS) formulation, we propose an objective function that incorporates the trade-off between including high influence observations and excluding potentially benign observations from the fitted model. The first term minimizes the sum of the squared residuals plus the squared influence for observations selected as outliers, while the second term minimizes the total number of outliers excluded. The objective is formulated such that if the trade-off parameter α is set to zero, we recover OLS, and if α is set to 1, we recover LTS with $k = (n-p-1)/2$.

Least Trimmed Squares

$$\begin{aligned} Min\ & \sum_{i=1}^{n} r_i^2 \\ s.t.\ & \beta_0 + \beta_p' X_i - y_i \le r_i + M z_i,\ \forall i \\ & -\beta_0 - \beta_p' X_i + y_i \le r_i + M z_i,\ \forall i \\ & \sum_{i=1}^{n} z_i \le k \\ & r_i \ge 0,\ \forall i \\ & z_i \in \{0,1\} \end{aligned}$$

Bounded Influence Least Squares

$$\begin{aligned} Min\ & \alpha \cdot \sum_{i=1}^{n} (r_i^2 + z_i \delta^2) + (1-\alpha) \cdot \sum_{i=1}^{n} z_i \\ s.t.\ & \beta_0 + \beta_p' X_i - y_i \le r_i + M z_i,\ \forall i \\ & -\beta_0 - \beta_p' X_i + y_i \le r_i + M z_i,\ \forall i \\ & \sum_{i=1}^{n} z_i \le (n-p-1)/2 \\ & r_i \ge 0,\ \forall i \\ & z_i \in \{0,1\} \end{aligned}$$

Computational Experiments

The estimators are evaluated on synthetic data with varying levels of outlier contamination. The covariates and error terms are drawn from normal distributions. Outliers for both covariates and response variables are generated from a uniform distribution. Reported below are the β estimates, number of outliers correctly detected, mean squared error (MSE), and runtime for each of the formuations solved to optimality, unless otherwise indicated by a (*).

Synthetic Data Set (SDS-1): n = 40, p = 5, 12% outlier density										
—	β_0	β_1	β_2	β_3	β_4	β_5	β_6	Outliers Detected	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	-	5/5	0.035	0.310
LTSR	0.06	0.17	-0.61	0.97	0.75	-0.44	-	4/5	0.038	0.303
BILS	0.05	0.21	-0.60	0.96	0.76	-0.50	-	5/5	0.035	0.552

In Synthetic Data Set 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 Data Set (SDS-2): n = 40, p = 5, 24% outlier density										
—	β_0	β_1	β_2	β_3	β_4	β_5	β_6	Outliers Detected	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	-	7/9	0.053	1.740
LTSR	0.03	0.18	-0.62	1.32	0.77	-0.64	-	6/9	0.053	2.398
BILS	0.04	0.19	-0.65	1.32	0.76	-0.67	-	7/9	0.055	1.182

Synthetic Data Set 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 Data Set (SDS-3): n = 80, p = 6, 36% outlier density										
—	β_0	β_1	β_2	β_3	β_4	β_5	β_6	Outliers Detected	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	17/24	0.040	180.000*
LTSR	0.05	0.30	-0.40	0.58	0.37	-0.61	0.61	17/24	0.040	180.000*
BILS	0.05	0.31	-0.40	0.58	0.37	-0.61	0.62	18/24	0.041	180.000*

Synthetic Data Set 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.

Interpretation of Results

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. Retaining these observations lowers the objective cost function more than dropping them.

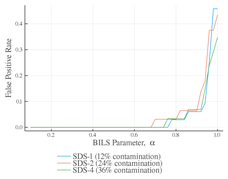

Fig. 4: False Positive Rate vs. α

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 data set 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 more observations than necessary being removed.

Application to Real World Data

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

Although we demonstrated that outlier-robust estimators outperform OLS in terms of MSE on our synthetic data sets, we wished to see how these results would extend to real data. We utilize the QSAR Aquatic Toxicity data set from the UCI Machine Learning Repository to predict the LC50 for a particular species of zooplankton in the presence of varying amounts of toxic contaminants. 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 the table above.

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. Additionally, BILS was not as efficient compared to LTS or LTSR as we hoped after observing the runtime in SDS-2. Overall, we are cautiously optimistic about our findings. Further testing on real world data is required to determine the true scalability of BILS, as well as the extent of its practical usability on outlier contaminated data sets. Still, the results recovered in both our synthetic and real world testing are encouraging for the future performance and relevance of this new model.

References

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

[2] 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.

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

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

[5] 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.

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