International Journal of Mathematics and Statistics
ISSN (Online): 2208-2212
Volume 10 Issue 02 June 2024
DOI: https://doi.org/10.69980/evcf4246

ROBUST STATISTICAL ESTIMATION FOR MULTICOLLINEAR DATA UNDER RESPONSE CONTAMINATION: A COMPARATIVE ANALYSIS

Dr. Adrian J. Fletcher[1]*, Prof. Elena M. Rossi[2], Dr. Markus T. Vogel[3], Dr. Claire N. Beaumont[4]

[1]*Department of Statistical Modeling and Data Science, Westbridge Institute of Technology, Birmingham, United Kingdom*
[2]*School of Applied Mathematics and Statistics, Lombardy University of Science, Milan, Italy*
[3]*Institute for Computational Statistics and Econometrics, Rhine Academic University, Munich, Germany*
[4]*Department of Quantitative Research and Statistical Computing, École Européenne des Sciences, Lyon, France*

Abstract:
The reliability of the traditional estimators may be diminished due to regression data with highly correlated predictors and influential observations. This study employed and benchmarked ordinary least squares, Ridge, LASSO, Elastic Net, and Huber-loss SGD regression with a community-level dataset with high predictor dependence. The observations and predictors after pre-processing were 1,994 and 40, respectively. To check the multicollinearity, the pairwise correlations and the condition number were calculated. To evaluate the model performance, they used the 70:30 training-test split and the RMSE, MAE, R^2 metrics. Robustness was subsequently assessed with simulated responses contaminated at 0%, 5%, 10% and 20%. The predictor matrix was highly collinear (maximum absolute correlation 0.986, condition number 232.95). In the un-contaminated cases Huber-loss SGD regression and Ridge had the lowest RMSE values. However, as the contamination increased, the Huber loss SGD regression performed relatively well, while OLS, LASSO, Ridge and Elastic Net experienced significant loss in performance. The Huber-loss SGD regression was able to retain the same RMSE (366.50) and R^2 (0.6042) as the same contamination level as the OLS regression (479.08, 0.3232). With the estimators evaluated, Huber-loss SGD regression exhibited the strongest stability with respect to the investigated cases of multicollinearity and response contamination.

Keywords: *Multicollinearity; Robust regression; Ridge regression; Elastic Net; Huber-loss SGD regression*

1. Introduction:

Today, it is common to have high dimensional data sets, typically characterised by a large number of explanatory variables with complex dependence structures, in statistical modelling. While large sets of predictors can enrich a statistical model's descriptive and predictive power, they also mean significant methodological challenges. Of these, multicollinearity is especially significant since high correlation between the explanatory variables can lead to high variances of the regression coefficients, to instability in the regression coefficients, and to sensitivity of the regression results for small changes in the sample of observations. Regularized regression has thus become a significant tool for dealing with estimation instability due to redundant or tightly correlated predictor information (Sztepanacz & Houle, 2024). Variable selection and shrinkage methods are also required to get parsimonious and computationally stable models in high dimensional settings (Chamlal et al., 2024).

Ordinary least squares (OLS) is one of the most commonly used regression estimators due to the simplicity in computing the regression and its statistical properties under the standard assumptions. But it can be quite poor in cases of many correlated predictors or when there are influential observations and outliers. Penalized estimators have been shown to perform better under multicollinearity and contaminated observations via simulation evidence as compared to OLS (Nayem et al., 2024). Likewise, Herawati et al. (2024) indicated that Ridge, LASSO, and Elastic Net can cope well with the negative impact of the multicollinearity phenomenon, but the results of each of these methods vary depending on the data structure and the level of the correlation between predictors.

Ridge regression is used to deal with multicollinearity by minimizing the L2 norm of the regression coefficients, essentially achieving a trade-off between small bias and potentially large variance reduction. However, the Ridge regression performance is also sensitive to the proper specification of the Ridge regression penalty parameter, leading to the invention of other parameter-selection methods (Göktaş et al., 2021). The efforts to further investigate the optimal estimation of the Ridge penalty have also continued recently using simulation and empirical applications, with the emphasis on optimal parameter choice for stable estimation when predictors are correlated (Khan et al., 2024). LASSO adds an L1 penalty to the objective function and is appealing for simultaneous estimation and variable selection as it can shrink some coefficients exactly to zero. Elastic Net utilizes both L1 and L2 penalties and may be particularly beneficial in the presence of strongly correlated sets of relevant predictors (Su & Wang, 2021).

Regularization can alleviate multicollinearity-driven instability but can not guard against influential observations. This restriction has stimulated the creation of penalized procedures which include strong estimation. However, to mitigate the effect of atypical observations and maintain shrinkage properties desired for multicollinear models, weighted penalized M-estimators have been proposed, for instance (Wasim et al., 2024). Another solution for the simultaneous presence of multicollinearity and outliers is the quantile-based robust estimators which extend Kibria-Lukman and Ridge-type estimators (Wasim et al., 2025a). Other comparative studies on the robust Ridge M-estimators also argue that using robust loss functions with shrinkage can lead to significant improvement when both blemishes are present (Wasim et al., 2025b).

In recent years, the development of sound statistics has become more and more relevant in the field of modern high-dimensional analysis. Robust methods are aimed at making the results as insensitive as possible to violations of the idealized distributional assumptions, such as contamination, heavy-tailed errors, or other anomalous observations. Loh (2025) noted that in today's world of high dimensional data, robust statistics has changed significantly. The robust regression procedure has better estimation and prediction properties than procedures relying heavily on the Gaussian distribution under heavy-tailed distributions (Adomaityte et al., 2024). Wang and Karunamuni (2022) also showed that the high-dimensional regression setting with generally weaker error assumptions is important for loss functions.

The one of the key pillars of robust regression is m-estimation. Squared-error estimation allows to give undue weight to outlying residuals, while robust M-estimators modify the loss function to give undue weight to atypical residuals. For the broad applicability of M-estimators to get stable regression estimates for contaminated data, see de Menezes et al. (2021). The Huber-loss SGD regression generalizes this idea by introducing a penalty term on the quadratic loss of the moderate residuals while keeping the sensitivity to extreme residuals low and is proved to have good theoretical properties in high-dimensional and heavy-tailed settings (Sun et al., 2020).

When there is high dimensionality and contamination both, robustness becomes even more important. In this paper, Wang et al. (2020) have proposed a robust method for high-dimensional regression without requiring any tuning procedure, which shows that reliable estimation can be obtained without heavily restricting the distributional assumptions. Similarly, Liu et al. (2020) considered robust sparse regression in high-dimensional contaminated setting, where the high-dimensional framework is needed when the classical estimators are sensitive toward corrupted data. In addition, methods that are able to deal with both influential observations and complex predictor structures are essential for robust high dimensional regression, as Filzmoser and Nordhausen (2021) noted.

The study addresses multicollinearity and growing contamination of response variables for OLS, Ridge, LASSO, Elastic Net and Huber-loss SGD regression. The most stable and accurate estimator is determined by the performance metrics: RMSE, MAE and R^2.

2. Methodology

2.1 Data Source and Analytical Sample

The analysis was conducted using the Communities and Crime Dataset (Unnormalized Data) obtained from Kaggle (Redmond, 2009). The original data comprised 2,215 observations at the community level and many socioeconomic, demographic, housing and law-enforcement variables. The dependent variable chosen to be analyzed was ViolentCrimesPerPop, which is the number of violent crimes per 100,000 people.

Observations with missing data for the dependent variable were removed, leaving 1,994 communities in the analysis. Community identifiers and crime related variables that may cause direct outcome leakage were dropped prior to the estimation of the model. This resulted in a first round of candidate predictors for non-crime events of 124.
Several predictors had a significant amount of missing data, and the variables with too much missingness were not included in the comparative modeling. Out of all the predictors, 40 were selected after pre-processing and screening for statistical analysis.

2.2 Preliminary Data Assessment

Empirical distribution of the dependent variable was described using descriptive statistics. The assessment involved the mean, standard deviation, median, first quartile, third quartile, minimum and maximum.
The distribution of ViolentCrimesPerPop was also graphed to check for skewness and for the presence of outliers. The preliminary examination was significant because the comparative analysis was to be used to examine estimator behavior in the setting of correlated predictors and possibly influential observations.

2.3 Assessment of Multicollinearity

wo complementary measures were used to check for multicollinearity among the predictors retained. To begin with, for every 40 predictors, pairwise Pearson correlation coefficients were computed. Highest absolute correlation values were found in order to see how much overlap exists between explanatory variables.
Second, the condition number of the standardized predictor matrix was determined as a general measure of multicollinearity. High values of the condition number and high pairwise correlations were believed to indicate high predictor dependence.
Next, the 10 predictor pairs with highest absolute correlation were selected and summarized to provide the strongest patterns of multicollinearity in the data.

2.4 Comparative Statistical Estimators

Five regression estimators were considered: Simple linear regression; Ridge regression; LASSO regression; Elastic Net regression; and Huber-loss SGD regression.
The classical benchmark estimator Ordinary Least Squares was added. Ridge regression was considered since coefficient shrinkage may help in estimation when predictors are highly correlated. LASSO was used as a regularized estimator that integrates shrinkage and variable selection, and Elestic Net was explored due to its integration of both L1 and L2 regularization. Huber-loss SGD regression was used as the robust estimator. The Huber loss is optimized using stochastic gradient descent approach in this model, which results in decreased influence of observations with large residuals and maintains the roughly squared-error behavior for observations with smaller residuals.

2.5 Training and Testing Procedure

The last analytical sample was split into training and test samples. The observations were divided into 70% for model estimation and 30% out of sample for model evaluation.
Each of the five estimators was trained on the same training set and tested on the same test set. The common data partition allowed the predictive performance of the methods to be compared directly in the same conditions.

2.6 Model Performance Measures

Three measures were used to assess the out-of-sample performance of each estimator:

- Mean Absolute Error (MAE); and
- Mean Absolute Error (MAE); and

Coefficient of Determination (R2).
The RMSE was employed to quantify the size of the prediction error, but with larger residuals given more weight. MAE was also used for an additional measure of average prediction error that was less sensitive to extreme errors. The percentage of the dependent variable's variance accounted for by the fitted model was determined using R^2.
The values of RMSE and MAE showed good predictive performance as the values got lower, while the higher the values of R^2, the better the out-of-sample predictive fit.

2.7 Simulated Response Contamination

For the purpose of evaluating the robustness of estimators, the training data were corrupted with control responses. Four contamination conditions were tested: 0%, 5%, 10% and 20% contamination.
Randomly selected training observations were modified by large random additions or subtractions (equal to six standard deviations of the training outcome) to the outcome of the training observations, for the contaminated conditions. The maintenance of the test sample was left unchanged for the contamination analysis. Thus, the changes in out-of-sample performance depended on the sensitivity of each estimator to the contamination added when estimating the model.
For each contamination level, the same five estimators were refitted and then compared with RMSE, MAE and R^2 values. This procedure allowed direct comparison of the performance of the estimators as the percentage of contaminated training observations was increased.

2.8 Comparative Evaluation

This comparative analysis was split into two phases. The first step was to assess estimator performance with the uncontaminated data to see the behavior of classical, regularized, and robust regression methods when the multicollinearity is severe. Second, there was a progressive increase in response contamination across the estimators.
Overall, the comparison was directed to two statistical challenges that were manifested in the data: predictors were highly correlated and the responses were sensitive to extreme values. The performance of the estimators was compared in terms of prediction accuracy and consistency under four contamination settings.

2.9 Ethical Considerations

There was no direct contact with humans or collection of personally identifiable information and the study incorporated an existing secondary data set with community-level information. Only anonymised community-level data was available for all analyses, which were undertaken for research purposes only. Thus, there was no further recruitment of participants or informed consent or direct intervention in the present study.

3. Results

3.1 Data Characteristics and Preliminary Assessment

The first set of data consisted of 2215 community-level observations. Of the 2,034 observations that were available for statistical analysis, 1,994 had a valid value for ViolentCrimesPerPop. After excluding community variables and crime variables that might lead to direct outcome leakage, 124 non-crime candidate predictors were considered. Some variables in the law enforcement data were found to have a significant number of missing observations, so variables with a high proportion of missing values were eliminated and 40 informative ones were used in the comparative estimation analysis. The dependent variable was highly skewed and had a high amount of variance. The average rate of violent crime was 589.08 per 100,000 people, while the median rate was much lower at 374.06. The standard deviation was 614.78 and the range of observed values was from 0 to 4,877.06. The first and third quartiles were 161.70 and 794.40 respectively (Table 1). These characteristics suggest that there are high levels of variability in community crime rates and perhaps influential extreme observations.

Table 1. Dataset characteristics and descriptive statistics

Statistic	Value
Raw observations	2,215
Observations included in analysis	1,994
Initial non-crime candidate predictors	124
Predictors retained for comparative analysis	40
Overall missingness among candidate predictors	14.99%
Condition number	232.95
Maximum absolute predictor correlation	0.986
Mean violent-crime rate	589.08
Standard deviation	614.78
Median	374.06
First quartile	161.70
Third quartile	794.40
Minimum	0.00
Maximum	4,877.06

The distribution of the outcomes was highly skewed to the right as illustrated in Figure 1, with many communities having relatively low-to-moderate violent-crime rates and a smaller number of communities with extremely high rates of violent crime. This distribution is suitable as an empirical example to evaluate estimation procedures which are less sensitive to extreme observations.

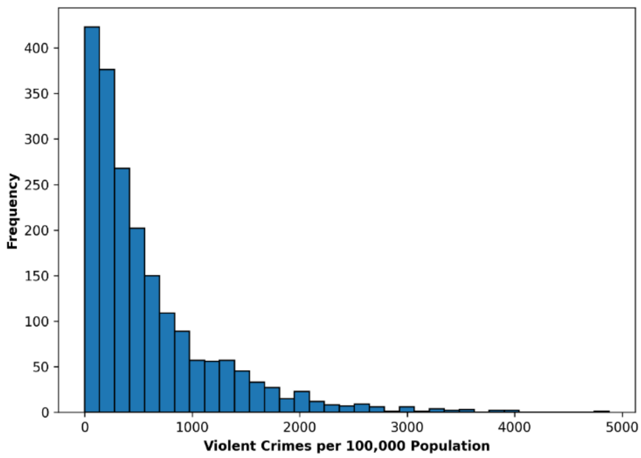

Figure 1. Distribution of violent crimes per 100,000 population across the analyzed communities

3.2 Evidence of Multicollinearity

There was a significant amount of multicollinearity in the predictor structure. The condition number of the standardized predictor matrix was extremely high, at 232.95, much greater than what would typically be considered in the absence of predictor dependence. In addition, there were strong (>0.90) correlations between several predictor pairs.

The highest correlation was found between the proportions of LargHouseFam and LargHouseOccup (Pearson correlation = 0.9862, Table 2). Likewise, the correlations between PctKids2Par and PctFam2Par were 0.9858 and between FemalePctDiv and TotalPctDiv were 0.9834. The redundancy of housing ownership measures was also significant in terms of the correlation between PctPersOwnOccup and PctHousOwnOcc (0.9822).

Table 2. Strongest correlations among retained predictors

Predictor 1	Predictor 2	Pearson r
PctLargHouseFam	PctLargHouseOccup	0.9862
PctKids2Par	PctFam2Par	0.9858
FemalePctDiv	TotalPctDiv	0.9834
PctPersOwnOccup	PctHousOwnOcc	0.9822
medFamInc	medIncome	0.9790
TotalPctDiv	MalePctDivorce	0.9753
PctOccupMgmtProf	PctBSorMore	0.9512
PctFam2Par	PctYoungKids2Par	0.9366
medFamInc	perCapInc	0.9336
PctNotHSGrad	PctLess9thGrade	0.9317

The 10 strongest absolute correlations are shown in Figure 2. The correlation of the coefficients, which is greater than 0.93 for several variables, indicates that there is redundancy in the information contained in these variables. Hence, the coefficients estimated via OLS are more likely to be sensitive to noise than those estimated using regularised or robust methods.

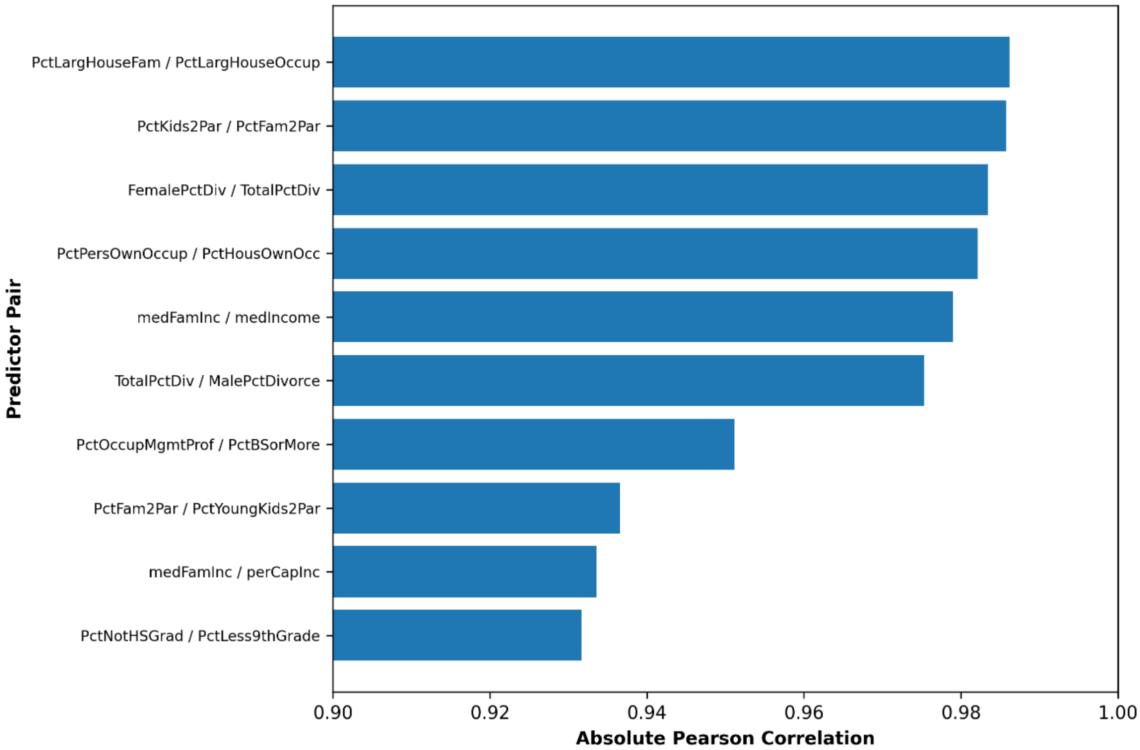

Figure 2. Ten strongest absolute correlations among the retained predictors, demonstrating severe multicollinearity

3.3 Comparative Predictive Performance of Statistical Estimators

The observations were split into two sets of 70% for model estimations and 30% for out-of-sample evaluation. Root mean squared error, mean absolute error and coefficient of determination were used to compare ordinary least squares, Ridge regression, LASSO regression, Elastic Net regression and Huber-loss SGD regression.

The Huber-loss SGD regression model had the lowest test RMSE of 365.09 and the lowest MAE of 235.47 (see Tab. 3). It also gave an R2 value of 0.6072, which was the most among all the other regression models. The ridge regression showed similar results in terms of RMSE and R^2, with an RMSE value of 365.17 and an R^2 value of 0.6071, but with a higher MAE which was 252.94.

Elastic Net also had an acceptable performance with the RMSE of 365.75 and R^2 of 0.6058. The ordinary least squares, however, yielded the highest RMSE value of 369.11 and R^2 of 0.5985, respectively, for the models that used clean data. LASSO was as effective as ordinary least squares.

Table 3. Out-of-sample performance of the comparative estimators

Estimator	RMSE	MAE	R²
OLS	369.11	256.25	0.5985
Ridge	365.17	252.94	0.6071
LASSO	368.56	256.14	0.5997
Elastic Net	365.75	253.58	0.6058
Huber-loss SGD regression	365.09	235.47	0.6072

The RMSE values are compared between the five estimators in Figure 3. The difference between the two models was not so significant in the absence of contamination, but the two models that yielded the lowest prediction errors were the robust estimator and Ridge regression. This implies that when there is a high level of multicollinearity, regularization is helpful and when there is good loss estimation, there is another benefit in minimizing the effect of extreme residuals.

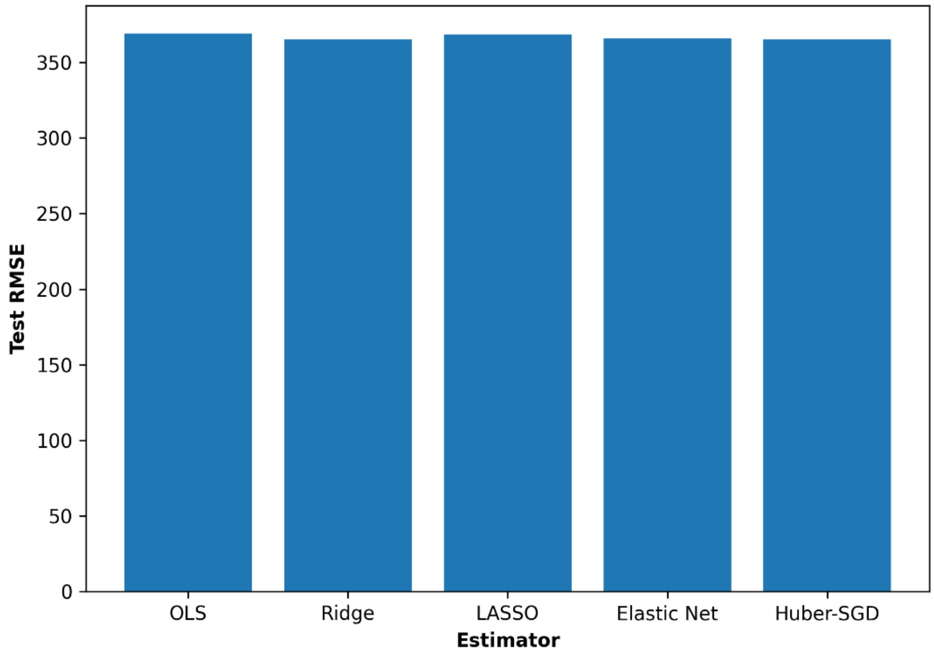

Figure 3. Comparison of out-of-sample RMSE across OLS, Ridge, LASSO, Elastic Net, and Huber-loss SGD regression

3.4 Performance Under Simulated Response Contamination

For evaluation of robustness, additional contamination at 5%, 10% and 20% was added to the training response. Random observations were subjected to large positive and negative perturbations of magnitude of six standard deviations from the training outcome. A comparison of the unchanged test sample with predictive performance was then made.

As seen in Table 4, there was significant difference in the performance of the estimators when the contamination level increased. The RMSE of ordinary least squares rose from 369.11 to 406.74 with contamination under 5%, but the Huber-loss SGD regression only slightly to 367.81 with contamination under 5%. The level of contamination was 10%, and OLS became even worse with an RMSE of 423.14 and an R^2 of 0.4724. The robust estimator, on the other hand, had an RMSE of 364.10 and R^2 of 0.6094.

The discrepancies grew really large at contamination levels under 20%. OLS produced an RMSE of 479.08, an MAE of 358.85, and an R^2 of only 0.3232. LASSO also had a large deterioration, with RMSE of 478.39. Ridge and Elastic Net were less unstable than OLS and LASSO, but still suffered significant loss of predictive accuracy. The RMSE value for Ridge was 459.64, and for Elastic Net was 463.91.

In contrast, the Huber-loss SGD regression resulted in an RMSE of 366.50, an MAE of 237.63 and an R^2 of 0.6042 when 20% of the training responses were contaminated.By comparison, the robust Huber-loss SGD regression retained an RMSE of 366.50, an MAE of 237.63, and an R^2 of 0.6042 even when 20% of the training responses were contaminated.

Table 4. Predictive performance under increasing simulated contamination

Contamination	Estimator	RMSE	MAE	R²
0%	OLS	369.11	256.25	0.5985
0%	Ridge	365.17	252.94	0.6071
0%	LASSO	368.56	256.14	0.5997
0%	Elastic Net	365.75	253.58	0.6058

Contamination	Estimator	RMSE	MAE	R^2
0%	Huber-loss SGD regression	365.09	235.47	0.6072
5%	OLS	406.74	289.37	0.5122
5%	Ridge	391.78	279.84	0.5477
5%	LASSO	404.77	288.47	0.5170
5%	Elastic Net	393.43	281.33	0.5439
5%	Huber-loss SGD regression	367.81	236.59	0.6014
10%	OLS	423.14	308.09	0.4724
10%	Ridge	402.03	288.86	0.5237
10%	LASSO	421.40	307.08	0.4767
10%	Elastic Net	404.62	291.44	0.5176
10%	Huber-loss SGD regression	364.10	235.20	0.6094
20%	OLS	479.08	358.85	0.3232
20%	Ridge	459.64	344.67	0.3773
20%	LASSO	478.39	358.51	0.3252
20%	Elastic Net	463.91	347.85	0.3657
20%	Huber-loss SGD regression	366.50	237.63	0.6042

The trends of the estimators are quite different with increasing contamination, as is evident in Figure 4. As the percentage of contamination increased, the prediction errors for OLS, LASSO, Ridge and Elastic Net regression grew, while Huber-loss SGD regression stayed more or less stable for all the contamination rates studied.

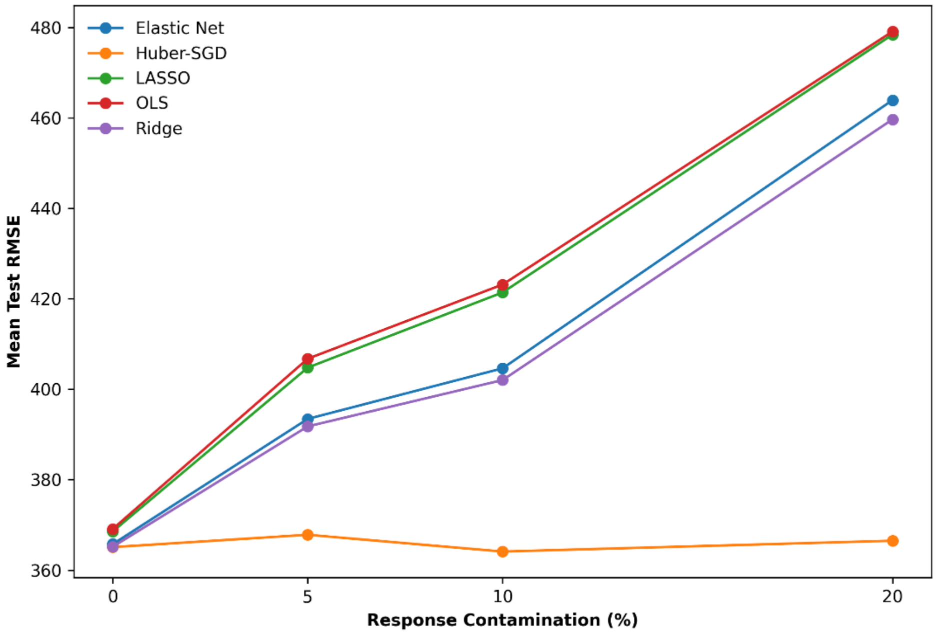

Figure 4. Changes in test RMSE across increasing levels of simulated response contamination

3.5 Overall Comparative Findings

Empirical analysis results combined with contamination analysis results show two significant statistical issues in the data set: the severe predicter multicollinearity and sensitivity to outlier observations. The advantage of coefficient shrinkage in ridge regression was demonstrated in the uncontaminated case, as expected because of the redundancy of many of the predictors. Elastic Net delivered comparable predictive performance and added shrinkage and variable selection.

Regularisation was not enough, however, to adequately mitigate the effects of severe response contamination. OLS and LASSO suffered the greatest loss as the level of contamination increased, and Ridge and Elastic Net provided only limited protection. The Huber-loss SGD regression showed the highest overall stability with an RMSE of around 364–368 for all contamination conditions and an R^2 of about 0.60.

The results suggest that if the data are multicollinear and influenced by one or more unusual observations, then it is possible to gain a significant benefit from using robust loss-based estimation. The regularized procedures are still useful to deal with predictor dependence; however, it seems that robust estimation is especially important if the contamination is an added source of instability in the estimation process.

4. Discussion

The results show that multicollinearity and contamination can have very different impacts on the regression estimators and that a regularization technique that works well at one end does not necessarily solve both issues. The predictor matrix was highly dependent and the condition number was 232.95, with the largest absolute pairwise correlation value being 0.986. In the uncontaminated case, Ridge, Elastic Net and the Huber-loss SGD regression performed slightly better than OLS and LASSO. But the more the differences became after adding the contamination of the responses. The Huber-loss SGD regression was also quite resistant to contamination, while the performance of OLS and LASSO was severely affected by contamination and that of Ridge and Elastic Net was partially.

The low relative stability of OLS to contamination is in line with the basic sensitivity of least-squares estimation to large residuals. At 20% contamination, OLS RMSE increased from 369.11 to 479.08, while R^2 declined from 0.5985 to 0.3232. The findings are consistent with earlier studies that show that conventional regression estimators break down when multicollinearity and outliers are present together. Under a combined multicollinearity-outlier problem, Affindi and Ahmad (2022) showed that the Ridge-type regularization in combination with robust procedures can offer advantages over the standard estimation. In a similar fashion, Arum and Ugwuowo (2022) showed that the RDMs can significantly improve estimation when there are highly correlated predictors and anomalous data simultaneously present.

Under uncontaminated conditions, one of the best performances was from ridge regression with an RMSE value of 365.17, as compared to 369.11 for OLS. This is in line with the anticipated variance-reduction effect of L2 regularization in the presence of strong multicollinearity. But, Ridge RMSE rose to 459.64 at 20% contamination. Therefore, while predictor dependence resulted in improved estimation, coefficient shrinkage did not completely shield the model from influential response observations. Majid et al. (2023) also proposed that the traditional Ridge estimators might still be susceptible to outliers, and presented the performance of robust Ridge alternatives. Similarly, Roozbeh et al. (2024) showed that strong mechanisms could be used to enhance Ridge estimation in the presence of both multicollinearity and outlier contamination.

The results with Lasso were also informative. In the clean conditions, RMSE values were 368.56 and R^2 of 0.5997 which are relatively close to RMSE value of OLS. However, when the contamination level was under 20%, its RMSE rose to 478.39 and R^2 decreased to 0.3252. LASSO works well for coefficient shrinkage and sparsity of variables, but the L1 penalty does not render the squared-error fitting criterion robust to extreme residuals. This finding aligns with the simulation results found by Altelbany (2021), who demonstrated that the multicollinearity effects on Ridge, LASSO and Elastic Net are different based on the predictor correlation and model structure. The present results also suggest that the benefits of penalization can become significantly diluted if the main source of instability is contamination as opposed to predictor dependence.

Elastic Net showed a good performance in resistance but also suffered with increasing contamination. Under uncontaminated conditions, its RMSE was 365.75 while under 20% contamination, the RMSE increased to 463.91. The relatively good clean data performance is consistent with the power of Elastic Net to select variables using an L1 penalty and stabilize the magnitude of the corresponding coefficients using an L2 penalty, when the variables are correlated. When extreme observations are available, Kim and Lee (2021) have shown that further robustification of Elastic Net may be required to enhance reliability in the case of high-dimensional sampling uncertainty. In a related line of work, Su and Wang (2021) incorporated Elastic Net penalization with quantile regression, demonstrating the possibility of regularization being used in conjunction with less outlier-sensitive principles of estimation.

The most interesting finding was that the Huber-loss SGD regression was stable. The RMSE for it were 365.09, 367.81, 364.10 and 366.50 under uncontaminated data, 5% contaminated data, 10% contaminated data, and 20% contaminated data, respectively. The same applied to the R^2 which was also around 0.60 for all conditions. The pattern is similar to what Huber-loss SGD regression is theoryed to do: moderate outliers are treated roughly quadratically, while extreme outliers are not allowed to dominate the objective function. The result is in agreement with the rich literature of M-estimation reviewed by De Menezes et al. (2021) and the properties of adaptive Huber regression claimed by Sun et al. (2020).

What is seen here is also consistent with broader advances in robust estimation in high dimensions. Ghosh and Thoresen (2021) highlighted the importance of powerful screening techniques if the high dimensional data are observed to be able to bias the variable importance. Similarly, Yang et al. (2022) suggested strong variable-selection rules that would still be efficient and robust to the presence of influential observations. These contributions contribute to the interpretation that robustness is not just another fitting criterion, but a general criterion for reliable high dimensional modelling.

A combination of multicollinearity with the presence of outliers has led to the development of some special estimators. Suhail et al. (2021) presented that quantile based robust Ridge M-estimation can be used to tackle correlated regressors and contamination. A related development of modified robust Ridge M-estimators for two parameter models of the Ridge type was made by Yasin et al. (2021). When there is multicollinearity and outlier data, the performance of robust Stein-type estimators has been shown by Lukman et al. (2023) to outperform the performance of conventional estimators. Indeed, the present result that shrinkage is not sufficient if contaminated responses have a significant impact on coefficient estimation is corroborated by these findings.

The observed pattern is very well supported by recent strong Ridge research. In a recent study, Wasim et al. (2024) demonstrated that weighted penalized M-estimators have the ability to enhance the robustness of the Ridge-type regularization while maintaining its benefits. Wasim et al. (2025a) further built upon this work by implementing a robust Kibria-Lukman estimator based on the quantile approach and found its results to be favourable when both multicollinearity and outlier issues exist simultaneously. In a similar manner, Wasim et al., (2025b) showed that the robust Ridge M-estimators are superior to the classic ones in the presence of both estimation problems. The present results confirm this overall impression: In case of correlated predictors, an appropriate regularization method is helpful in overcoming the problem of coefficient instability, whereas in the presence of contaminated observations, an appropriate choice of the loss function is essential for controlling the influence.

More generally, the results align with the current robust high-dimensional theory. Adomaityte et al. (2024) demonstrated that robust regression can still inherit desirable properties when the data has heavy tails and are high-dimensional, and Loh (2025) noted that robust statistical theory has become more and more important as the dimensionality and the complexity of contamination grow. Similarly, Wang and Karunamuni (2022) showed that new robust loss functions for high dimensional regression are useful. The present analysis provides an applied comparison and demonstrates that these theoretical benefits hold up well in real-world scenarios of increasingly polluted responses.

The results also support the notion that sparsity, dependence of the predictors, and contamination should be treated as a combined problem and not as separate ones. Liu et al. (2020) proved that the robust sparse regression is an important problem when there are corrupted observations in high dimensional data, and Wang et al. (2020) proved that when data are high dimensional, reliable performance of the high dimensional robust estimation is possible without being too sensitive to the conventional tuning assumption. Similarly, Filzmoser and Nordhausen (2021) highlighted the need for techniques to handle complex predictor structures and influential observations in high-dimensional robust regression.

The overall results show that the methods Ridge and Elastic Net are helpful when the main statistical challenge is multicollinearity, however, they are not as robust when significant contamination is added to the data. The Huber-loss SGD regression model was the most stable of the tested models, highlighting the importance of stable loss-based methods in the presence of strongly correlated predictors and anomalous responses. It is also consistent with the overall comparative evidence presented in Göktaş et al. (2021), and in Sztepanacz and Houle (2024), which highlight the improvement of Ridge estimation in the case of problematic regression structures and limited information. Combined, the results justify a modelling approach where the type of estimator to use depends on the dependence structure of the predictors, and the expected level of contamination, and not exclusively by dimensions.

5. Conclusion

This study made a comparison between OLS, Ridge, LASSO, Elastic Net, and Huber-loss SGD regression when the data have high multicollinearity and when the response is contaminated progressively. Results indicated that the predictive performance of Ridge and Elastic Net were better with multicollinearity, while OLS and LASSO were sensitive to the increasing contamination. The Huber-loss SGD regression showed the highest overall stability with RMSE values around 365-368 and R^2 values of around 0.60 in all levels of contamination. OLS performance, however, worsened considerably with RMSE growing to 479.08 and R^2 falling to 0.3232 at 20% contamination level. The ridge and elastic net algorithms also had some protection and were heavily decimated as the contamination level rose.

These results suggest that regularization is effective when the predictors are correlated but it is inadequate when influential response observations also have an impact on the data. A solid loss-based estimation is an important extra layer of protection against contamination. Thus, when using estimator in the multicollinear regression, it is important to take into account the dependence among the predictors in addition to the possibility of existing abnormal values. In the data structure and contaminations investigated Huber-loss SGD regression showed the most robust predictive performance, whereas Ridge and Elastic Net were found to be helpful when the multicollinearity was the primary concern.

References

1. Adomaityte, U., Defilippis, L., Loureiro, B., & Sicuro, G. (2024). High-dimensional robust regression under heavy-tailed data: Asymptotics and universality. Journal of Statistical Mechanics: Theory and Experiment, 2024(11), 114002.
2. Affindi, A. N., & Ahmad, S. (2022). Robust ridge regression approach for combined multicollinearity-outlier problem. Applied Mathematics and Computational Intelligence, 11(1), 123–132.
3. Altelbany, S. I. (2021). Evaluation of Ridge, Elastic Net and Lasso Regression Methods in Precedence of Multicollinearity Problem: A Simulation Study. Journal of Applied Economics & Business Studies (JAEBS), 5(1), 131.
4. Arum, K. C., & Ugwuowo, F. I. (2022). Combining principal component and robust ridge estimators in linear regression model with multicollinearity and outlier. Concurrency and Computation: Practice and Experience, 34(10), e6803.
5. Chamlal, H., Benzmane, A., & Ouaderhman, T. (2024). Elastic net-based high dimensional data selection for regression. Expert Systems with Applications, 244, 122958.
6. De Menezes, D. Q., Prata, D. M., Secchi, A. R., & Pinto, J. C. (2021). A review on robust M-estimators for regression analysis. Computers & Chemical Engineering, 147, 107254.
7. Filzmoser, P., & Nordhausen, K. (2021). Robust linear regression for high-dimensional data: An overview. Wiley Interdisciplinary Reviews: Computational Statistics, 13(4), e1524.
8. Göktaş, A., Akkuş, Ö., & Kuvat, A. (2021). A new robust ridge parameter estimator based on search method for linear regression model. Journal of Applied Statistics, 48(13-15), 2457-2472.
9. Ghosh, A., & Thoresen, M. (2021). A robust variable screening procedure for ultra-high dimensional data. Statistical Methods in Medical Research, 30(8), 1816-1832.
10. Herawati, N., Wijayanti, A., Sutrisno, A., Nusyirwan, & Misgiyati. (2024). The performance of ridge regression, LASSO, and elastic-net in controlling multicollinearity: A simulation and application. Journal of Modern Applied Statistical Methods, 23(2), 1–13.
11. Khan, M. S., Ali, A., Suhail, M., Alotaibi, E. S., & Alsubaie, N. E. (2024). On the estimation of ridge penalty in linear regression: Simulation and application. Kuwait Journal of Science, 51(4), 100273.

12. Kim, H., & Lee, T. H. (2021). A robust elastic net via bootstrap method under sampling uncertainty for significance analysis of high-dimensional design problems. Knowledge-Based Systems, 225, 107117.
13. Liu, L., Shen, Y., Li, T., & Caramanis, C. (2020, June). High dimensional robust sparse regression. In International Conference on Artificial Intelligence and Statistics (pp. 411-421). PMLR.
14. Loh, P. L. (2025). A theoretical review of modern robust statistics. Annual Review of Statistics and Its Application, 12(1), 477-496.
15. Lukman, A. F., Farghali, R. A., Kibria, B. G., & Oluyemi, O. A. (2023). Robust-stein estimator for overcoming outliers and multicollinearity. Scientific Reports, 13(1), 9066.
16. Majid, A., Amin, M., Aslam, M., & Ahmad, S. (2023). New robust ridge estimators for the linear regression model with outliers. Communications in Statistics-Simulation and Computation, 52(10), 4717-4738.
17. Nayem, H. M., Aziz, S., & Kibria, B. G. (2024). Comparison among ordinary least squares, ridge, lasso, and elastic net estimators in the presence of outliers: Simulation and application. International Journal of Statistical Sciences, 24(2), 25–48.
18. Redmond, M. (2009). Communities and crime dataset (unnormalized data) [Data set]. Kaggle. https://www.kaggle.com/datasets/johnp47/communities-and-crime-dataset
19. Roozbeh, M., Maanavi, M., & Mohamed, N. A. (2024). A robust counterpart approach for the ridge estimator to tackle outlier effect in restricted multicollinear regression models. Journal of Statistical Computation and Simulation, 94(2), 279-296.
20. Su, M., & Wang, W. (2021). Elastic net penalized quantile regression model. Journal of Computational and Applied Mathematics, 392, 113462.
21. Suhail, M., Chand, S., & Kibria, B. G. (2021). Quantile-based robust ridge M-estimator for linear regression model in presence of multicollinearity and outliers. Communications in Statistics-Simulation and Computation, 50(11), 3194-3206.
22. Sun, Q., Zhou, W. X., & Fan, J. (2020). Adaptive huber regression. Journal of the American Statistical Association, 115(529), 254-265.
23. Sztepanacz, J. L., & Houle, D. (2024). Regularized regression can improve estimates of multivariate selection in the face of multicollinearity and limited data. Evolution Letters, 8(3), 361-373.
24. Wang, Y., & Karunamuni, R. J. (2022). High-dimensional robust regression with Lq-loss functions. Computational Statistics & Data Analysis, 176, 107567.
25. Wang, L., Peng, B., Bradic, J., Li, R., & Wu, Y. (2020). A tuning-free robust and efficient approach to high-dimensional regression. Journal of the American Statistical Association, 115(532), 1700-1714.
26. Wasim, D., Suhail, M., Albalawi, O., & Shabbir, M. (2024). Weighted penalized m-estimators in robust ridge regression: an application to gasoline consumption data. Journal of Statistical Computation and Simulation, 94(15), 3427-3456.
27. Wasim, D., Suhail, M., Khan, S. A., Shabbir, M., Awwad, F. A., Ismail, E. A. A., Ahmad, H., & Ali, A. (2025a). Quantile-based robust Kibria–Lukman estimator for linear regression model to combat multicollinearity and outliers: Real life applications using T20 cricket sports and anthropometric data. Kuwait Journal of Science, 52(1), 100336.
28. Wasim, D., Zaman, Q., Ahmad, M., & Kibria, B. G. (2025b). Mitigating multicollinearity and outliers in regression: comparison of some new and old robust ridge M-estimators. Journal of Statistical Computation and Simulation, 95(16), 3526-3547.
29. Yang, Z., Fu, L., Wang, Y. G., Dong, Z., & Jiang, Y. (2022). A robust and efficient variable selection method for linear regression. Journal of Applied Statistics, 49(14), 3677-3692.
30. Yasin, S., Salem, S., Ayed, H., Kamal, S., Suhail, M., & Khan, Y. A. (2021). Modified robust ridge M-estimators in two-parameter ridge regression model. Mathematical Problems in Engineering, 2021, 1845914.