Study of Least Absolute Deviation Methods

Lyle Paukner, Professor Jessica Kraker ❖ Mathematics ❖ University of Wisconsin-Eau Claire

Abstract

In the context of finding the “best” predictive model, this research project focuses on assessing and fitting models with least-absolute-deviation techniques. Such techniques are more robust to outliers, though there are computing considerations with the mechanics of fitting such models.

Linear multiple regression explores the relationship of a response variable to predictor variables through a linear combination of the terms, plus error (“noise”). Using predictors explored in a previous study, we simulated responses with a variety of types of errors. We assess the fits of a variety of models to the data via both least-squares and least-absolute-deviation measures. Additional measures of skewness and outliers were used in quantifying the impact of the types of errors on the assessment values.

Predictions were performed by fitting the model via ordinary and “penalized” regression methods. One of these methods, the L1-norm loss penalized regression, aims to minimize the least absolute deviation of the errors. While this model has been introduced in the literature, its preferential use has not been examined. Assessments for the various penalized methods will be computed and cross-compared over the different types of simulated errors, with the goal of better describing data situations in which this L1-norm loss penalized regression results in better predictions than with other regression-fitting methods.

Model Fitting Methods

Multiple regression

The general purpose of multiple regression is to learn more about the relationship between several independent (or predictor) variables $x_1, x_2, \ldots,$ and x_k and a dependent (or response) variable y via the model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$ where $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ are constants and ε is “noise.” Predictions, denoted $\hat{y}_i$, are computations of the above with coefficient *estimates*, for observation *i*.

Ordinary Least Squares Regression

The ordinary least squares (OLS) estimates are obtained by minimizing the sum of the squared errors (SSE) over all possible choices for β coefficients:

$$\min_{\beta}\{SSE\} = \min_{\beta}\left\{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right\}$$

Least Absolute Deviation Regression

Estimates for least absolute deviation (LAD) regression are obtained by minimizing the sum of absolute errors (SAE) over all possible choices for β coefficients. This method is known for being more robust to outlier but also for involving much more extensive computation techniques.

$$\min_{\beta}\{SAE\} = \min_{\beta}\left\{\sum_{i=1}^{n}|y_i - \hat{y}_i|\right\}$$

Penalized regression

Penalized regression modifies the method of least squares to allow biased estimators of the regression coefficients. This method is preferable when the bias is small, and the variance of the biased estimator is reduced compared to the unbiased estimator.

Ridge Regression :

Ridge regression is like OLS but shrinks the estimated coefficients towards zero. Ridge regression estimates have a closed form solution, and are more stable. Its solution can be stated in a matrix form similar to OLS and is determined by the following minimization (putting a penalty on the squared coefficients).

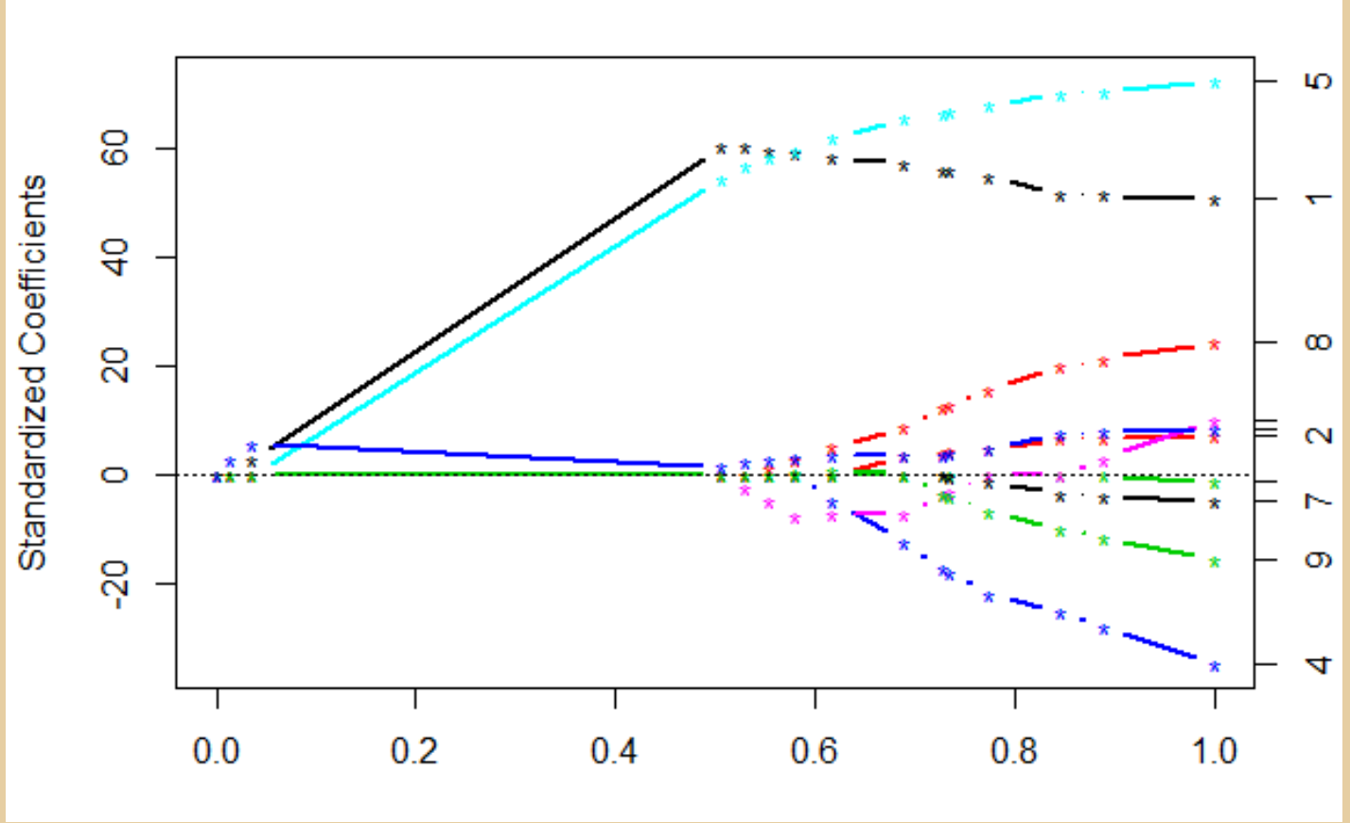
$$\min_{\beta,\lambda}\left\{SSE + \lambda\left(\sum_{j=1}^{k}{\beta_j}^2\right)\right\}$$

Lasso:

“The ‘lasso’ minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant.” (Tibshirani) This method often produces a more interpretable model, since some coefficients are reduced to zero (i.e. predictor selection).

$$\min_{\beta,\lambda}\left\{SSE + \lambda\left(\sum_{j=0}^{k}|\beta_j|\right)\right\}$$

All solutions are shown in the picture: The horizontal axis corresponds to penalties, with values decreasing left to right. A coefficient’s value for a particular penalty is the height of its curve at the point along the horizontal axis corresponding with the penalty.



Elasticnet:

This method uses a combination of Ridge Regression and Lasso penalties, along with a minimization of SSE. This helps to produce a more interpretable model, as some coefficients are reduced to 0, as well as enjoying the stability of the RR solution.

$$\min_{\beta,\lambda}\left\{SSE + \lambda\left[(1-\alpha)\left(\sum_{j=0}^{k}|\beta_j|\right) + \alpha\left(\sum_{j=0}^{k}{\beta_j}^2\right)\right]\right\}$$

Least-absolute deviation regression with ridge regression penalty (L1RR):

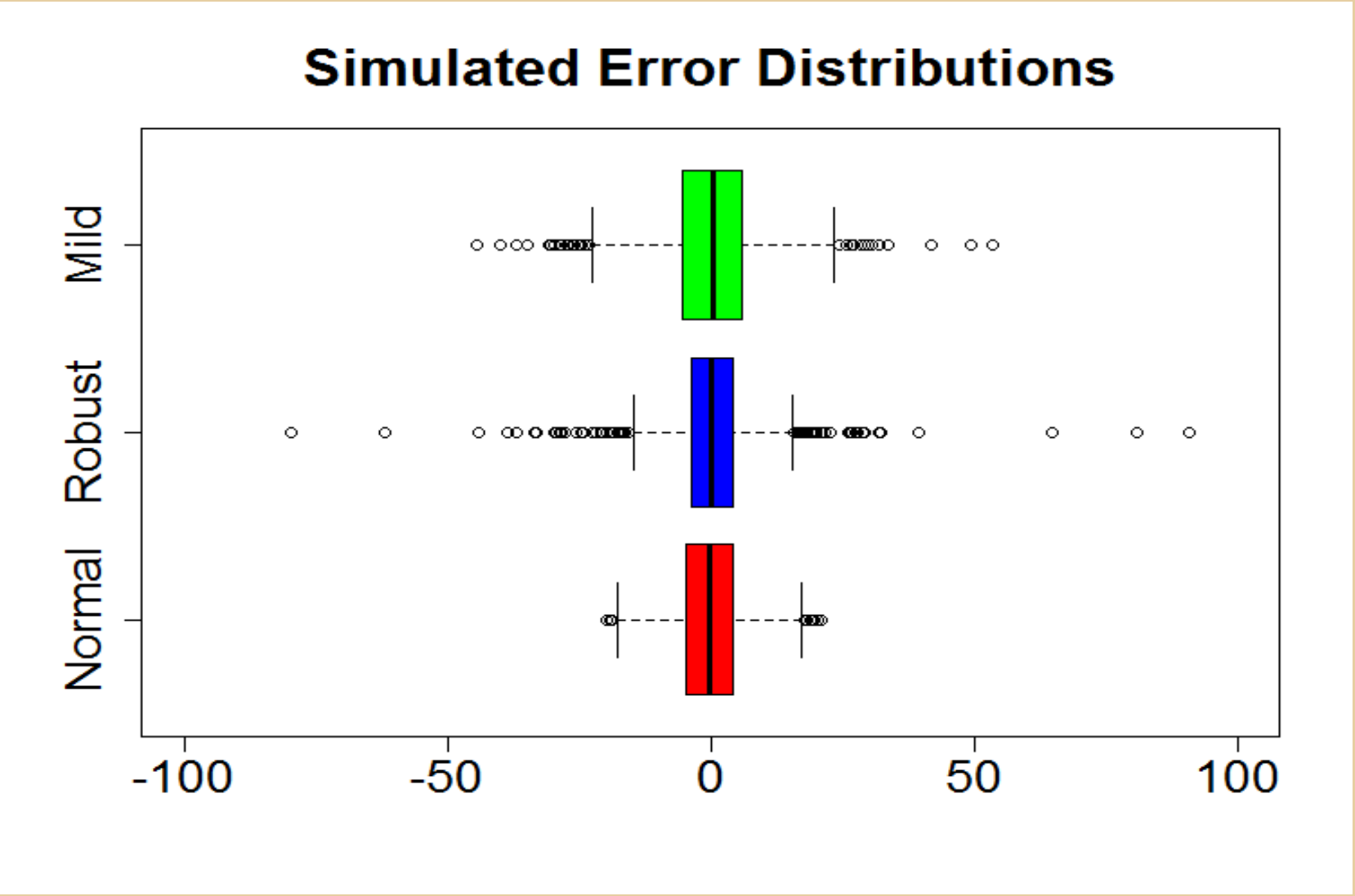
Another alternative to the LASSO uses both a squared component and an absolute-value component, but switches their application between the goal and penalty functions. This results in a method that shrinks the coefficients while fitting such that (k+1) of the residuals are 0.

$$\min_{\beta,\lambda}\left\{\sum_{i=1}^{n}|y_i - \hat{y}_i| + \lambda\left[\left(\sum_{j=1}^{k}{\beta_j}^2\right)\right]\right\}$$

Simulation Description

Design

- Each simulated data set consists of 100 observations, with 30 predictors each, simulated using R.
- The first five x values (predictors) were randomly simulated from a Gamma distribution to imitate skewed data.
- The next five x values were randomly simulated from a Normal distribution.
- The final 20 predictors were simulated sparse data. First, the predictor was set as zero or non-zero with a 50% probability (of either). The non-zero predictors were then randomly selected from a Poisson distribution.
- We used the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{30} x_{30} + \varepsilon$. The β_j ‘s are 0, 1, 2, or 3, with a large number of zero coefficients are included to test whether our cross-validation/penalized regression combos could correctly identify zero vs. non-zero coefficients.
- We used three different types of errors: The error ε is randomly selected from a either a normal distribution, a “mild” t-distribution, or a “robust” t-distribution.
- 100 datasets of each three different type were simulated, for 300 datasets total.



Model Selection

Cross-validation:

This is a model validation technique for assessing how the results of a particular analysis will generalize to an *independent* data set. One round of cross-validation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset (called the fitting or training set), and validating the analysis on the other subset (called the validation or testing set). To reduce variability, multiple rounds of cross-validation are performed using different partitions, and the validation results are averaged over the rounds.

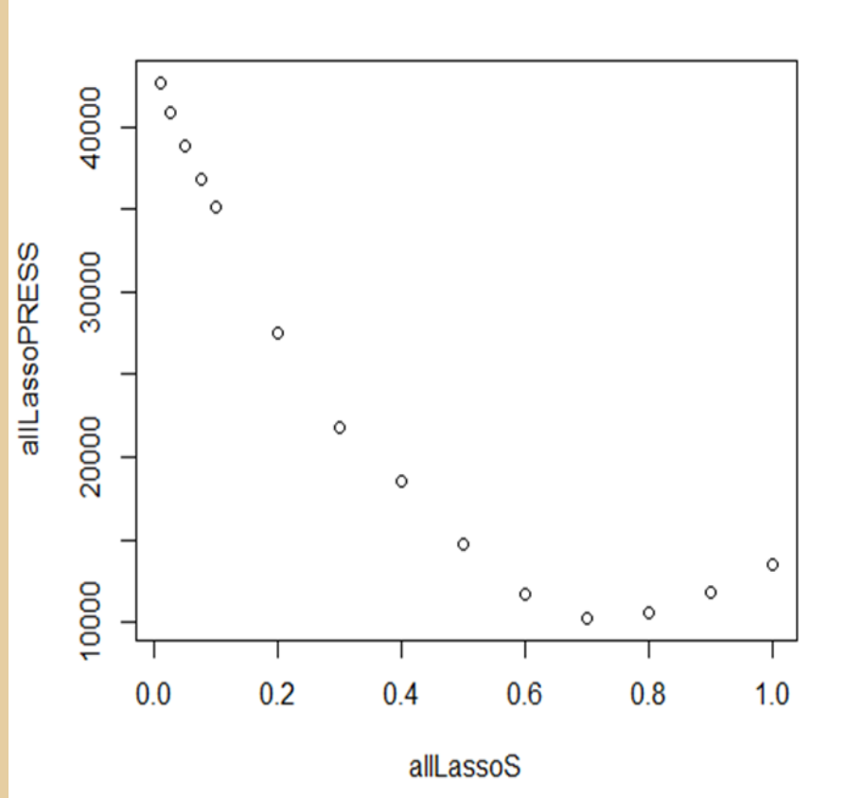
We performed Leave-one-out cross validation on the datasets using the different types of regression. One observation serves as the validation set, and the cross-validation is performed n times, where n is the size of the sample.

PRESS (Predicted residual sum of squares) Statistic:

We want to identify the minimum PRESS statistic to choose the best fitting penalty parameter(s) for the model. This is similar to minimizing SSE, except that the predictions are performed based on models fit without the validation set.

$$\min_{\beta,\lambda}\{PRESS\} = \min_{\beta,\lambda}\left\{\sum_{i=1}^{n}\left(y_i - \hat{y}_{(i)}\right)^2\right\}$$

The penalty parameter is selected by computing values of the PRESS statistic for models fit with different penalty values; the penalty value of the model resulting in the smallest PRESS statistic is then used. The image exemplifies selection of the penalty location for LASSO model.



Model Assessment & Results

External Datasets:

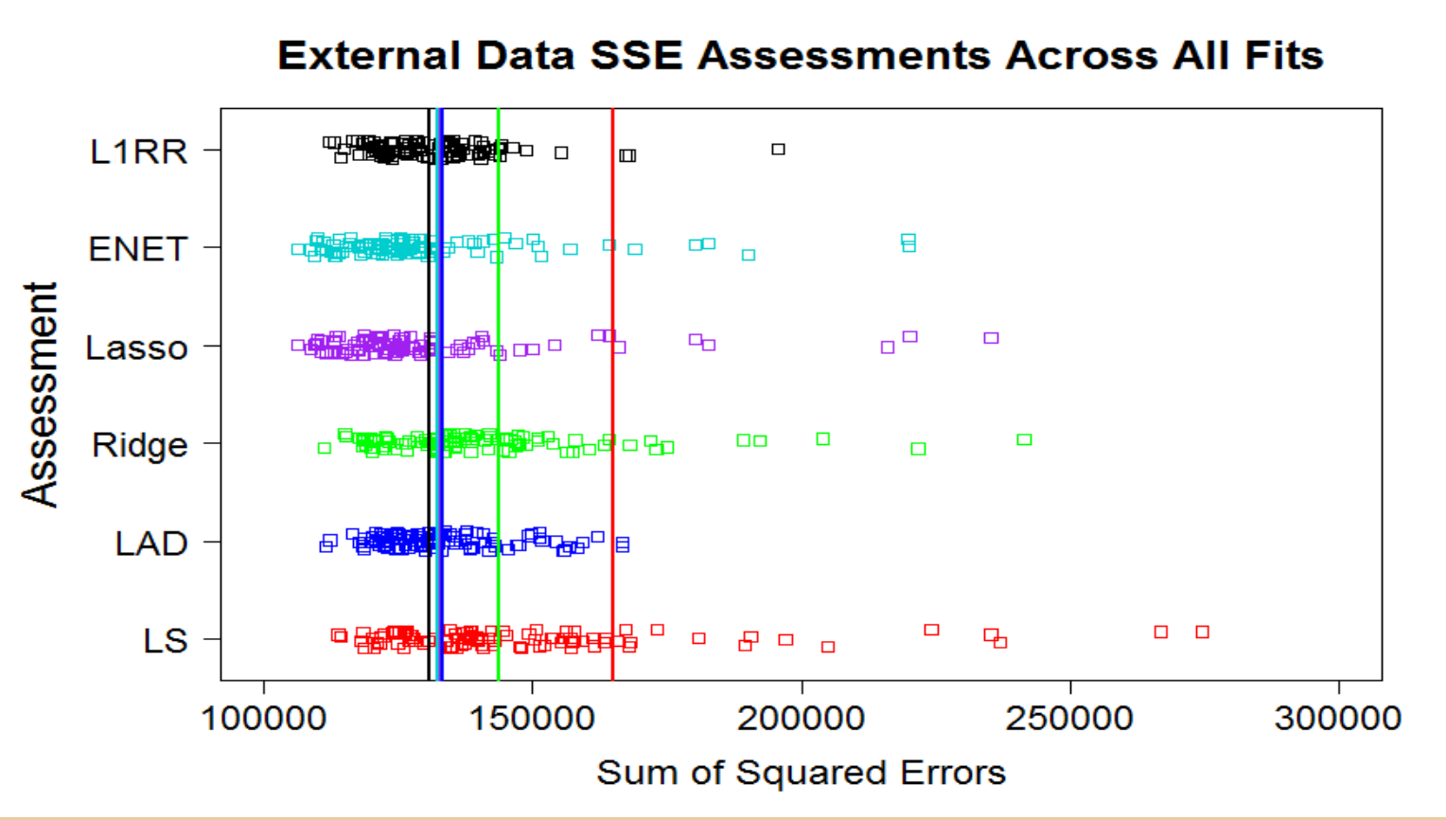
For each of the three different type of datasets, we created an “outside” dataset of 1000 observations. We then fit our predicted models to this new data in order to test how well each model performed. This results in a “true” assessment of the model predictive ability.

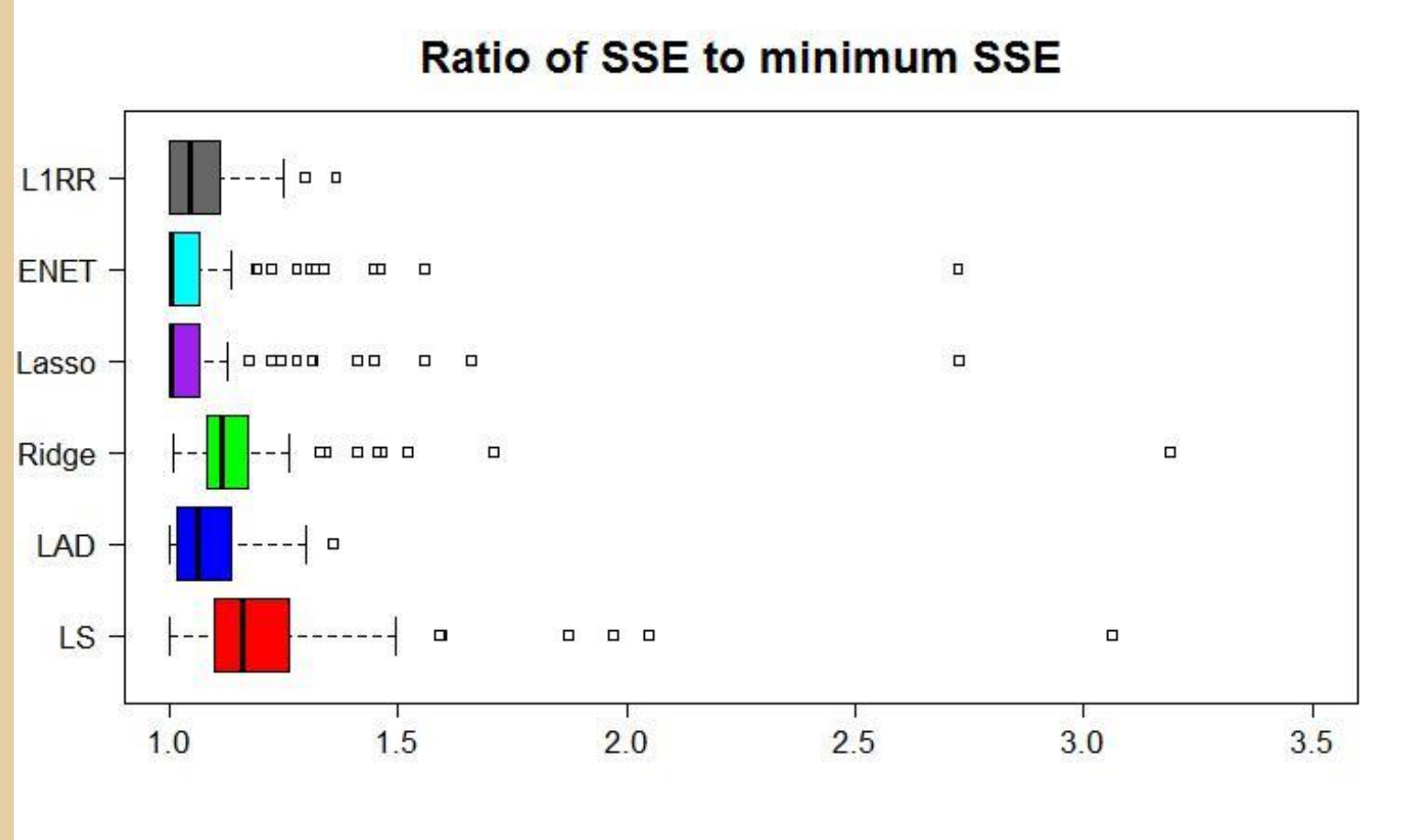
Assessment measure (SSE):

When we examine sparse data with “robust” errors, we can see that very often, Elasticnet and Lasso fits result in the lowest SSE. However, LAD and L1RR provide more consistent results, while Elasticnet and Lasso can still occasionally experience very high SSE.

The top plot shows the raw SSE values across all 100 datasets, for each of the six methods.

The boxplot at the bottom shows the ratio of the SSE measurement of each method to the minimum SSE measurement across the six methods, calculated for each of the 100 datasets.



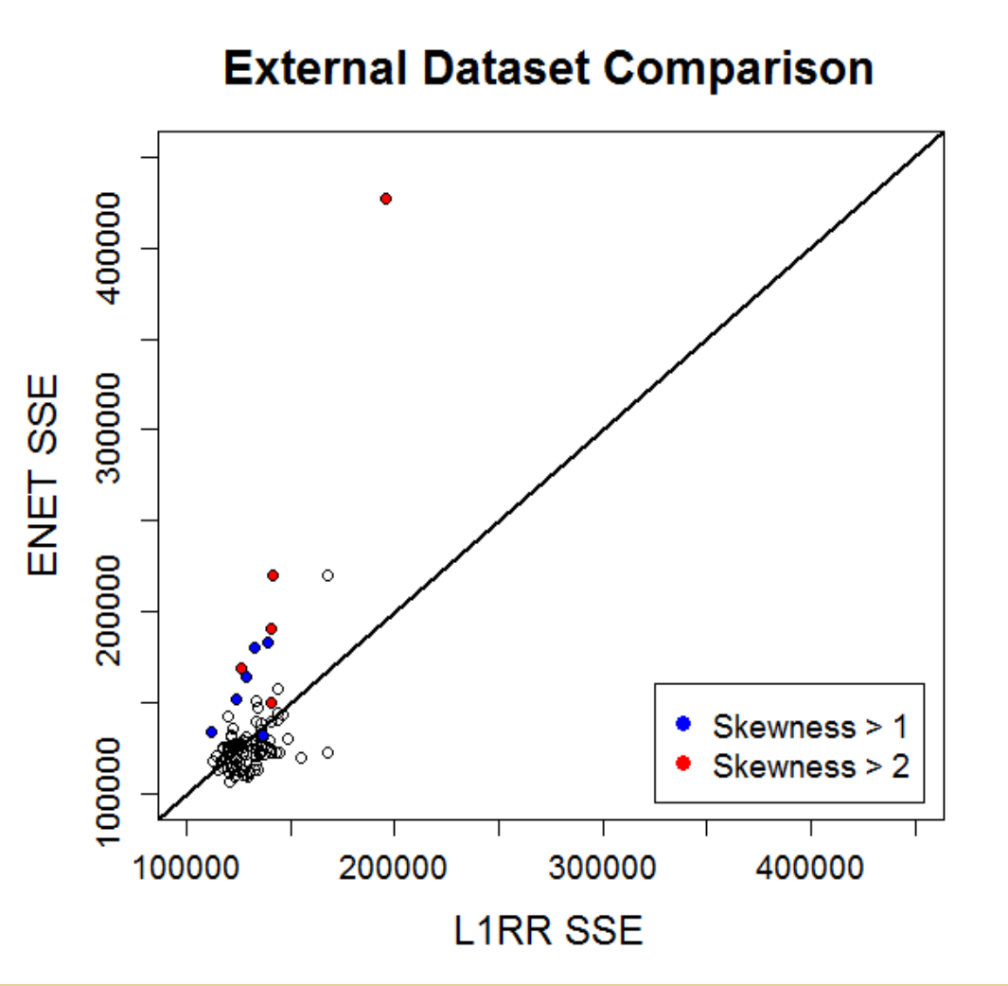


Observations regarding skewness:

Comparing the SSE’s for L1RR and Elasticnet from fits on each of 100 robust datasets, we see that L1RR outperforms Elasticnet when there is a high amount of skewness in the y-data which was used to fit the model. Quantifying this allows us to better identify types of responses for which L1RR may be the preferred model.

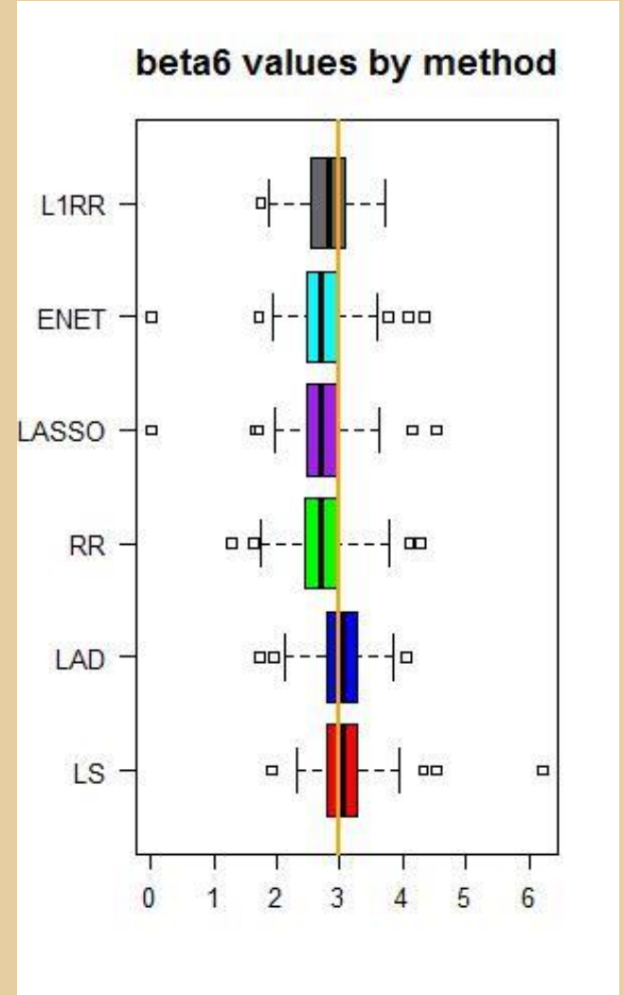
The skewness of the response values for each dataset is computed by the statistic:

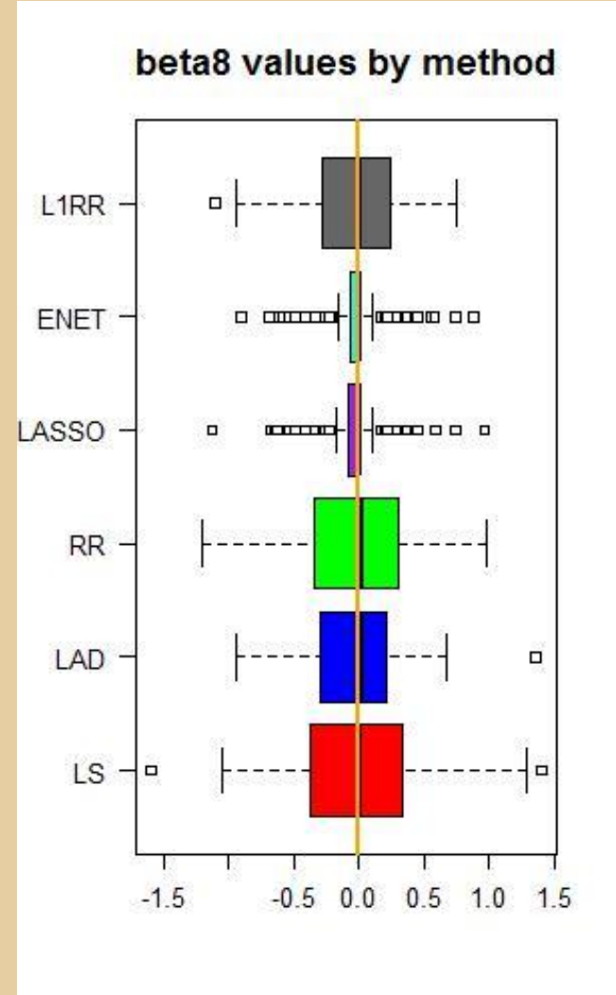
$$b_1 = \left(\frac{(n-1)^{3/2}}{n}\right)\frac{\sum_{i=1}^{n}(y_i - \bar{y})^3}{\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)^{3/2}}$$



Coefficient Estimation:

To the right, the boxplots show how two of the estimated coefficients from each fitting method are distributed over the 100 simulations. For the non-zero coefficients, with β_6 as the example, we can see that L1RR has about the same variance in the estimates, without outliers, plus it is not skewed by shrinkage of the coefficients as much as the other three penalized methods. For a zero coefficient, like β_8, Elasticnet and Lasso are capable of shrinking the coefficients to zero, while L1RR does not have the capability due to the quadratic penalty (which may advise adding an absolute-value penalty for “L1 Elasticnet”); however, L1RR still estimates better than RR.





Future Considerations

Explore other forms of penalized regression:

- Fused Lasso: Makes nearby coefficients more similar to one another
- Group Lasso: Dealing with grouped covariates, which are believed to have sparse effects on a group and within group level
- General Lasso: includes both fused and group lasso

These alternatives are attractive, as different types of predictors (e.g. skewed, sparse) can be penalized differently.

- L1 ENET: Extending the idea of the L1RR to include penalties on both the sum of squares and the absolute values of the coefficients, while implementing an absolute-error loss function, has been previously proposed by Kraker:

$$\min_{\beta,\lambda}\left\{\sum_{i=1}^{n}|y_i - \hat{y}_i| + \lambda\left[(1-\alpha)\left(\sum_{j=0}^{k}|\beta_j|\right) + \alpha\left(\sum_{j=0}^{k}{\beta_j}^2\right)\right]\right\}$$

Further computational work for the algorithm is necessary prior to application on similar simulated data.

Works Cited

Abramovich, F., and V. Grinshtein. “Estimation of a Sparse Group of Sparse Vectors.” *Biometrika* 100.2 (2013): 355-70. Print.

Joanes, D. N., and C. A. Gill, “Comparing measures of sample skewness and kurtosis”. *The Statistician* 47 (1998): 183–189. Print.

Koenker, Roger. *Quantile Regression (Econometric Society Monographs).* Cambridge University Press, 2005. Print.

Kraker, Jessica. “Penalized Regression Methods and Validation, with Particular Focus on Chemometric Data.” Dissertation paper, University of Minnesota (2008). Print.

Mendenhall, William, and Terry Sincich. *A Second Course in Statistics: Regression Analysis*. 6th ed. Upper Saddle River, NJ: Prentice Hall, 2003. Print.

Tibshirani, Robert. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1 (1996): 267-88. Print.

Zou, Hui, and Trevor Hastie. "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2 (2005): 301-20. Print.

Thanks To:

ORSP: for funding this project
Mathematics Department: for encouraging undergraduate research
Hans Kischel: for instruction on research tools