SML: Exercise 2

Thao Le, Finn-Ole Höner, Jason Wang, Ramon Snellen

2022-11-07

Introduction

This report aims to find the best prediction for past cumulative grocery sales (in dollars) for Dominick's Finer Foods based on an Elastic Net model by Zou and Hastie (2005).

Data

The data set contains seven years of store-level data collected at Dominick's Finer Foods by the University of Chicago Booth School of Business. The data can be found at https://www.chicagobooth.edu/research/kilts/datasets/dominicks. The data set contains 50 variables, which stem from:

1. customer count files, which contain information about in-store traffic;
2. a demographics file, which contains store-specific demographic data;
3. number identification files, which contain product information.

Of the fifty variables, `GROCERY_sum` is used as dependent variable. Furthermore, four categorical variables are dropped; `STORE`, `CITY`, `ZIP` and `SPHINDX`. The remaining variables are potential predictor variables.

Method

To find the optimal set of predictor variables, and there corresponding weights, we use a regression method that penalizes the size of coefficients. The penalty is useful when predictors are collinear, or the number of predictors destabilizes estimation. This data set only consists of 77 observations for 50 variables, hence the number of predictors would destabilize estimation if not penalized.

Let $P(\beta)$ denote a general penalty function. Then, the loss function of the regression equation becomes

$$L(\beta) = (\mathbf{y} - \mathbf{x}\beta)^T(\mathbf{y} - \mathbf{x}\beta) + \lambda P(\beta), \tag{1}$$

where λ is the hyperparameter that determines the strength of the penalty. When $P(\beta) = \beta^2$, the regression is called 'ridge' regression. Similarly, when $P(\beta) = |\beta|$, the regression is called 'LASSO' regression. Finally, any convex combination $P(\beta) = \alpha|\beta| + (1-\alpha)\beta^2$ of the 'ridge' and 'LASSO' penalty, where α denotes the weight on the 'LASSO' penalty, is called 'elastic net' (compare Zou and Hastie 2005). We use the elastic net to find the optimal set of predictor variables, since it exploits the LASSO property of variable selection, and the ridge ...

We use an MM-algorithm to estimate the elasting net. The MM-algorithm uses a majorizing function to find the coefficient vector β that minimizes the loss function specified in Equation 1. This is because the

minimum is not obtained analytically. Let ϵ denote the desired level of precision. The majorizing function is specified as

$$L_{MM}(\beta) = \frac{1}{2}\beta^{\mathrm{T}}(\frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda(1-\alpha)\mathbf{I} + \lambda\alpha\mathbf{D})\beta - \frac{1}{n}\beta^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} + c,$$

where $\mathbf{D}$ is a $p \times p$ diagonal matrix with elements

$$d_{jj} = 1/\max\left(|\beta_j^0|, \epsilon\right).$$

Furthermore,

$$c = \frac{1}{2n}\mathbf{y}^{\mathrm{T}}\mathbf{y} + \frac{1}{2}\lambda\alpha\sum_{i=1}^{p}|\beta_j^0|.$$

To find the optimal penalty strength λ and the weight of the elastic net α, K-fold cross-validation is used. K-fold cross-validation starts with splitting the data set into K folds. Subsequently, K iterations are ran as follows: for each iteration $k \in \{1, \ldots, K\}$, fold k is the test set, whereas the remaining $(K-1)$ folds together form the training set. On the training set, the elastic net is estimated for all $\lambda_i \in (10^{-2+i})_{i=0}^{12}$. Moreover, for each λ_i, all corresponding $\alpha \in \{0, 0.01, \ldots, 1\}$ are estimated, such that there are 2500 combinations of the hyperparameters α and λ. All estimated elastic nets (i.e., all combinations of the hyperparameters), are used to predict the dependent variable in the kth fold. When this procedure has been completed for all K folds, we have obtained fitted values for every observations in the data set, for all elastic nets. Consequently, prediction errors can be computed.

To evaluate which elastic net performs best (that is, the elastic net that maintains the optimal combination of hyperparameters), the Root Mean Square Error (RMSE) is computed for each elastic net; the elastic net that has the lowest RMSE is deemed to be the best model. Let $\hat{\mathbf{y}}_{\mathrm{test}(k)} = \mathbf{X}_{\mathrm{test}(k)}\hat{\beta}_{\mathrm{train(k)}}$ denote the fitted values, and n_k the number of observations in fold k. The RMSE is computed as

$$\mathrm{RMSE} = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\mathrm{MSE}_k},$$

where

$$\mathrm{MSE} = \frac{1}{n_k}(\mathbf{y}_{\mathrm{test}(k)} - \hat{\mathbf{y}}_{\mathrm{test}(k)})^{\mathrm{T}}(\mathbf{y}_{\mathrm{test}(k)} - \hat{\mathbf{y}}_{\mathrm{test}(k)}).$$

Results

In an estimation on simulated data, we get the same results for `glmnet` and our implementation based on the MM algorithm. However, for our estimation on the Dominick's Finer Food data, we get different estimates between these two implementations. Figure 1 shows the mean-absolute-percentage differences (MAPE) of the two implementations for the different predictors. For most predictors, the $MAPE$ is around to 1%.

One explanation for this difference could lie in the algorithms' speed of convergence. For a given data set the MM algorithm might not converge fast enough to provide the same estimates as the `glmnet` implementation. In fact, the cyclical coordinate decent used in `glmnet` is a very efficient algorithm according to Friedman, Hastie, and Tibshirani (2010).

We investigate this hypothesis by looking at the estimates' differences for lower convergence thresholds ϵ of the MM algorithm. Figure 2 shows the zero correlation between the Mean Absolute Error and the parameter ϵ: This precision parameter appears to play no role in the discrepancy between the coefficients of `glmnet` and the MM algorithm.

The uniform difference between the coefficients, and the result that this difference does not depend on the precision parameter ϵ leads us to believe that there are some structural differences between the `glmnet` and MM algorithm. However, as these discrepancies are low and similar across predictors, both algorithms seem to implement the same model.

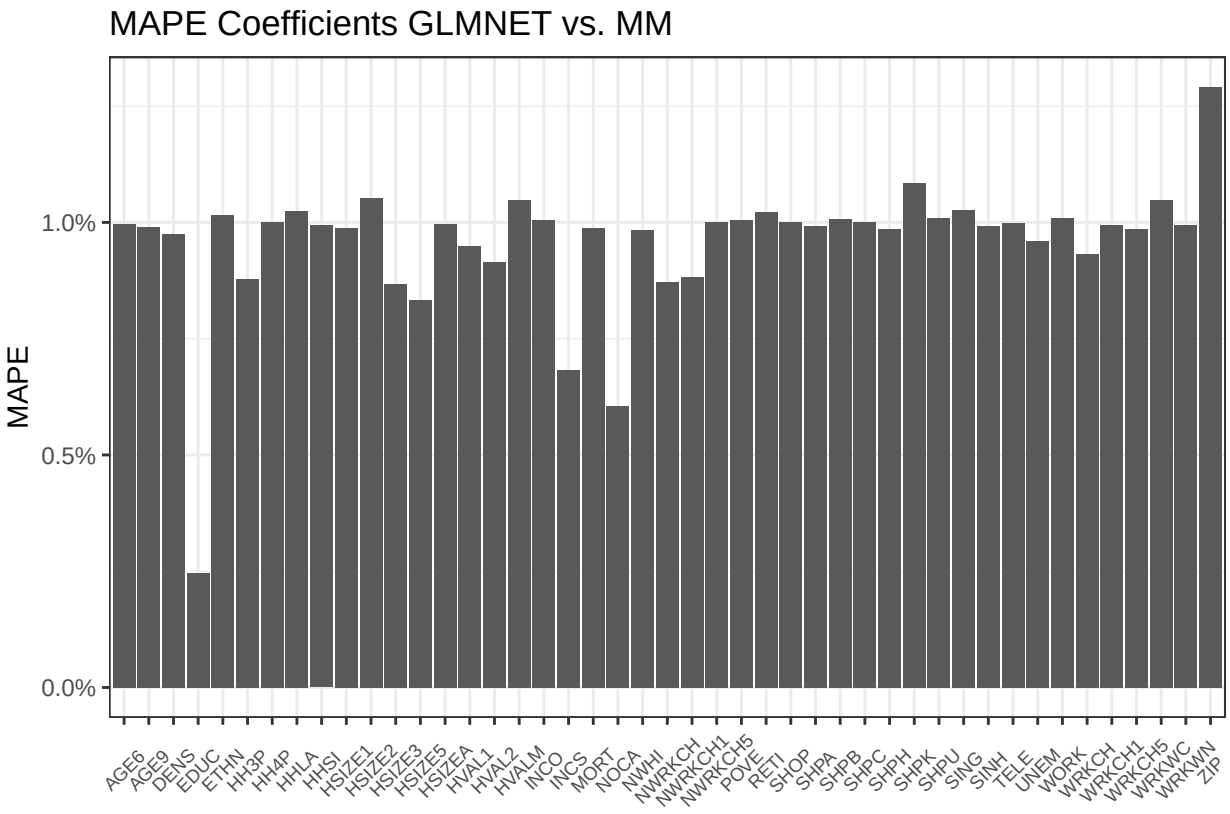


Figure 1: Mean absolute percentage difference between glmnet and MM coefficients.

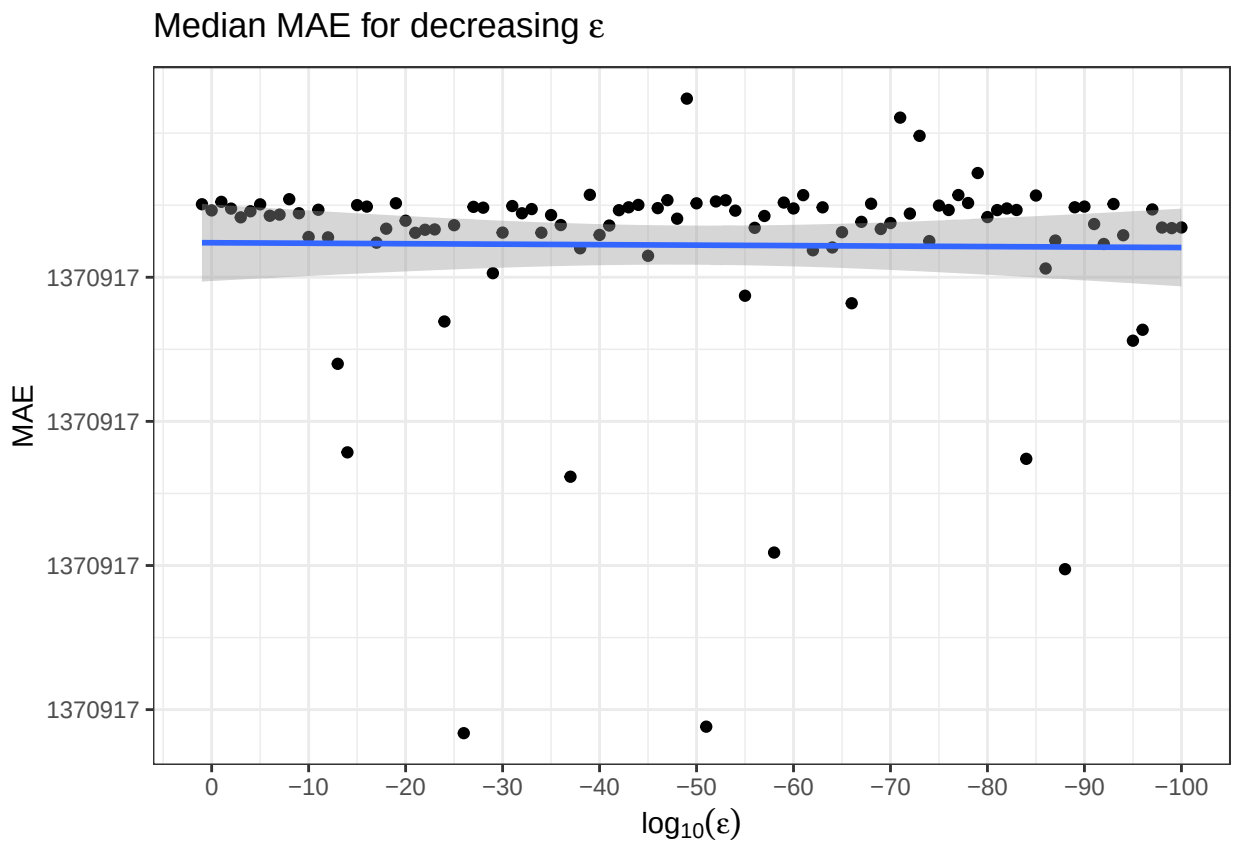
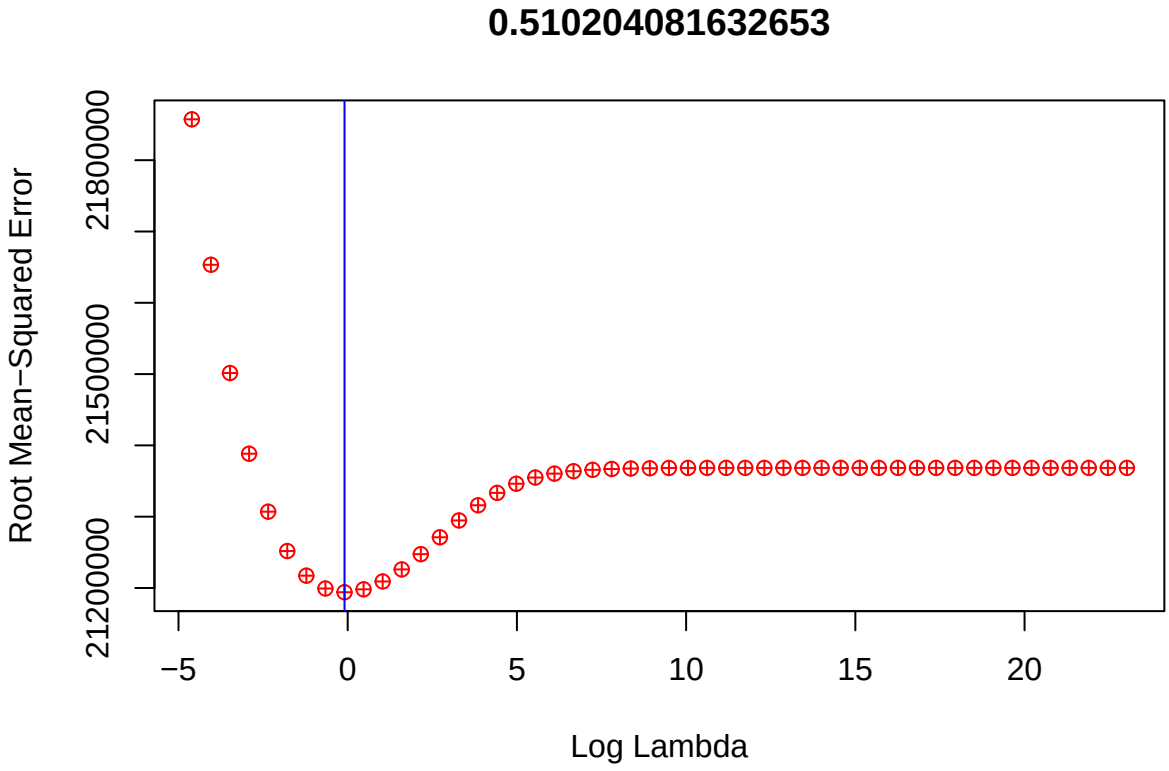


Figure 2: Development of MAE fo decreasing precision threshold ϵ.

Test on real data using k-fold

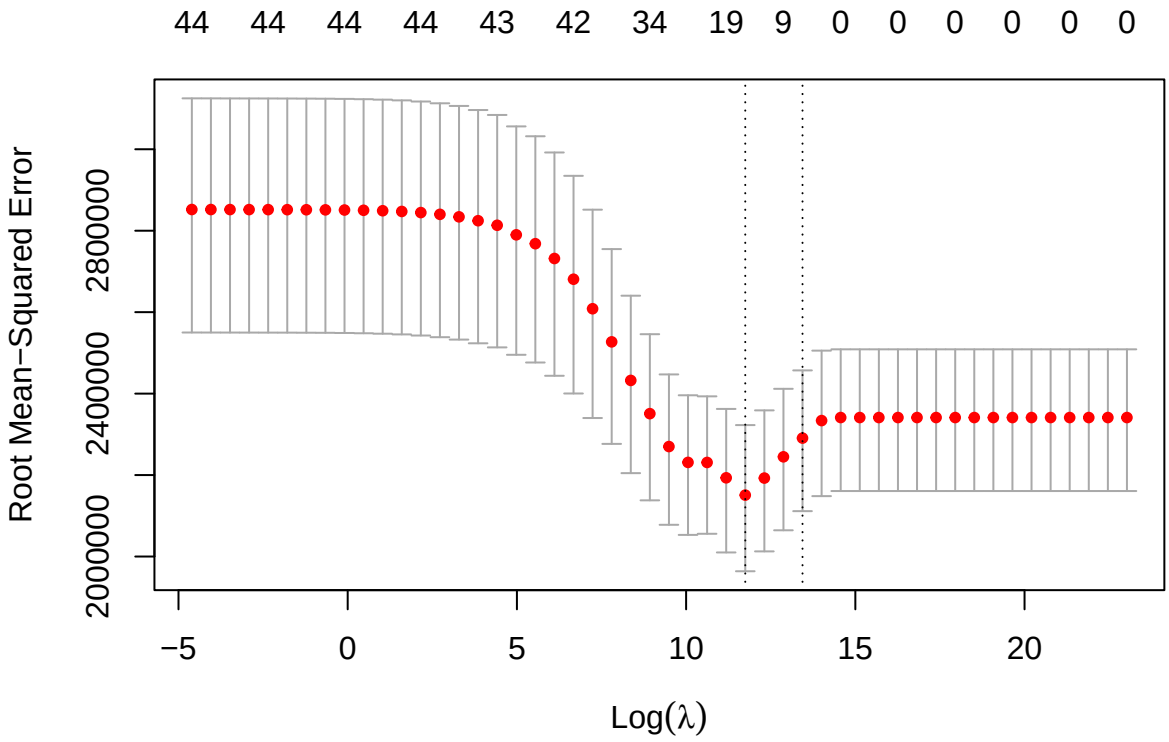
First we find the optimal Lambda and Alpha value for our MM-algorithm

```
## Alpha is:  0.5102041 . The minimum lambda is:  0.9102982
##  The minimum RMSE is:  21194047
```



Then we compare to the optimal lambda and alpha value using the GLMNET package.

```
## [1] 126485.5
## [1] 686648.8
```




```
## NULL
```

Conclusion and Discussion

References

Code

Friedman, Jerome, Trevor Hastie, and Rob Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33 (1): 1.

Zou, Hui, and Trevor Hastie. 2005. "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2): 301–20.