SML: Exercise 2

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

2022-11-05

Introduction

This report aims to find the best set of predictors for past cumulative grocery sales (in dollars) for Dominick's Finer Foods.

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 penalized regression equation becomes

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

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).

Results

In an estimation on simulated data, we get the same results for `glmnet` and our implementation based on the MM algorithm. We suspect that the algorithm used in the glmnet package (generalized linear model via penalized maximum likelihood), converges faster and hence delivers more precise estimates than our implementation of the elastic net with the MM algorithm.

```
dfCompareBetaTable %>% kableExtra::kable(align = "c")
```

GLMNET	MM	Predictor
45651.02	181882.881	ZIP
3473391.22	3080597.964	AGE9
6207250.85	6292294.324	AGE60
539235.73	592085.365	ETHNIC
74298.62	171182.276	EDUC
32420.41	3716067.874	NOCAR
-9346094.33	-10142944.188	INCOME
184405.07	-89534.759	INCSIGMA
341842.63	7121180.491	HSIZEAVG
-1618969.09	-200849.012	HSIZE1
-619077.05	0.000	

HSIZE2

190750.46

695678.203

HSIZE34

-150153.70

1979.621

HSIZE567

60390.68

1081328.520

HH3PLUS

-3498767.78

-6765715.045

HH4PLUS

-3074027.47

-64324.030

HHSINGLE

1038161.61

15530.931

HHLARGE

3723154.58

4415433.070

WORKWOM

4249458.82

254446.089

SINHOUSE

-674780.01

-551089.811

DENSITY

1369579.19

1567724.755

HVAL150

935586.99

898285.657

HVAL200

-1305677.32

-1198245.907

HVALMEAN

722826.68

563787.498

SINGLE

-1462241.66

-1534568.263

RETIRED

1175176.33

1364885.362

UNEMP

-4050903.63

-30887120.561

WRKCH5

-2312473.29

-2408.200

WRKCH17

1073906.26

504853.926

NWRKCH5

471486.42

14314.649

NWRKCH17

-500349.92

-4433927.985

WRKCH

207570.41

1221819.811

NWRKCH

1235861.39

29254690.291

WRKWCH

-4572046.01

-5148452.773

WRKWNCH

2932486.58

3134977.095

TELEPHN

-3209346.13

-3279539.537

MORTGAGE

-1539077.38

-1728503.399

NWHITE

-1944956.51

-181256.248

POVERTY

-4746032.48

-4882342.278

SHPCONS

3228436.05

5769056.960

SHPHURR

-4929322.53

-3603455.088

SHPAVID

567226.94

2784641.244

SHPKSTR

-1790732.83

0.000

SHPUNFT

-3703003.61

-2841869.636

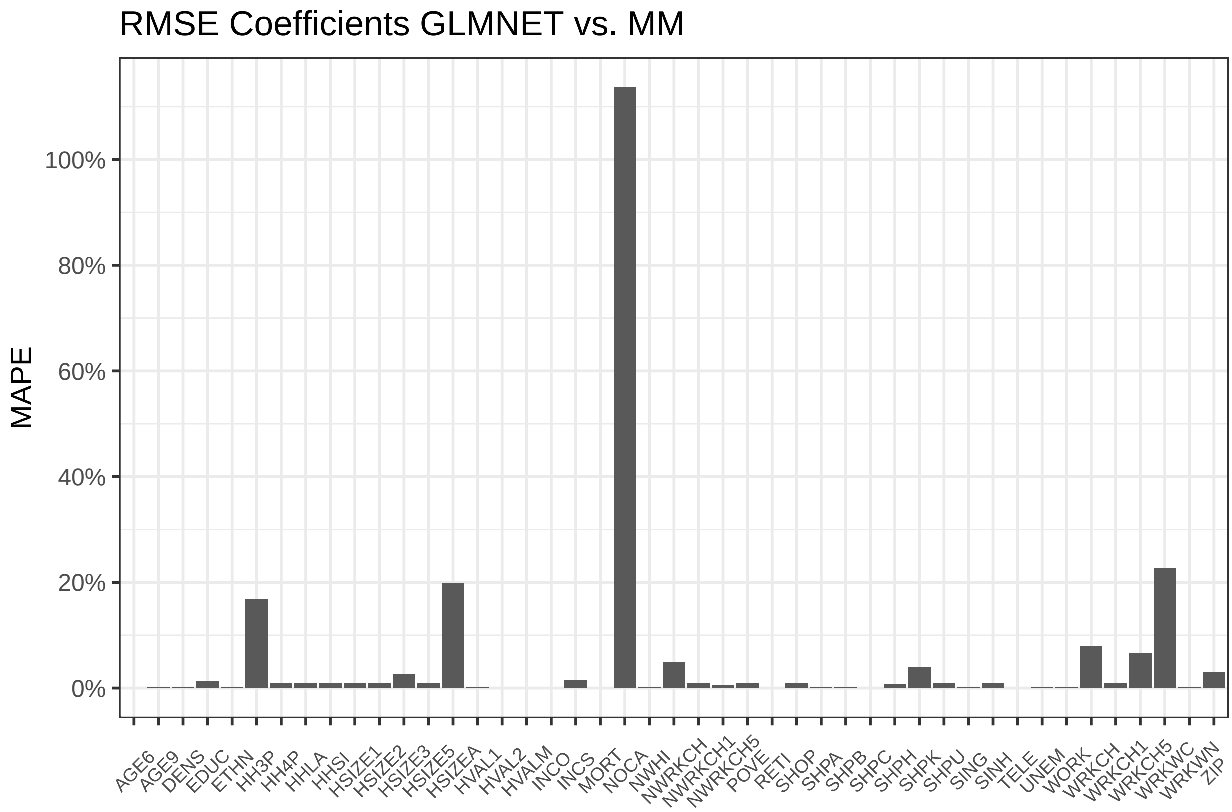
SHPBIRD

-7569247.80

0.000

SHOPINDX

```
plot_coef_rmse + labs(title = "RMSE Coefficients GLMNET vs. MM")
```



Conclusion and Discussion

Code

References

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.