University of St.Gallen

Forests

University of St. Gallen
School of Management, Economics, Law,
Social Sciences, International Affairs
and Computer Science

Assignment 3

Data Analytics I: Predictive Econometrics
Prof. Jana Mareckova

submitted by

Cyril Janak, 16-611-287
Jonas Husmann, 16-610-917
Niklas Kampe, 16-611-618
Robin Scherrer, 18-617-969

15.12.2021

Contents

Requirements

To solve the following tasks, the required libraries and the data sets are loaded first.

```
library(ggplot2)
library(dplyr)
library(grf)
library(DiagrammeR)
library(glmnet)

browser_2006 <- read.csv(file = "GHA/browser_2006.csv")
browser_new <- read.csv(file = "GHA/browser_new.csv")
```

Exercise 1

The average online spending is $1959.921

```
mean(browser_2006$spend)
```

```
## [1] 1959.921
```

Exercise 2

The household with id = 1297 (first row of the 2006 sample) spends most of the time on *weather.com*

```
row_id_1297 <- browser_2006[browser_2006$id==1297,3:ncol(browser_2006)]
which.max(row_id_1297)
```

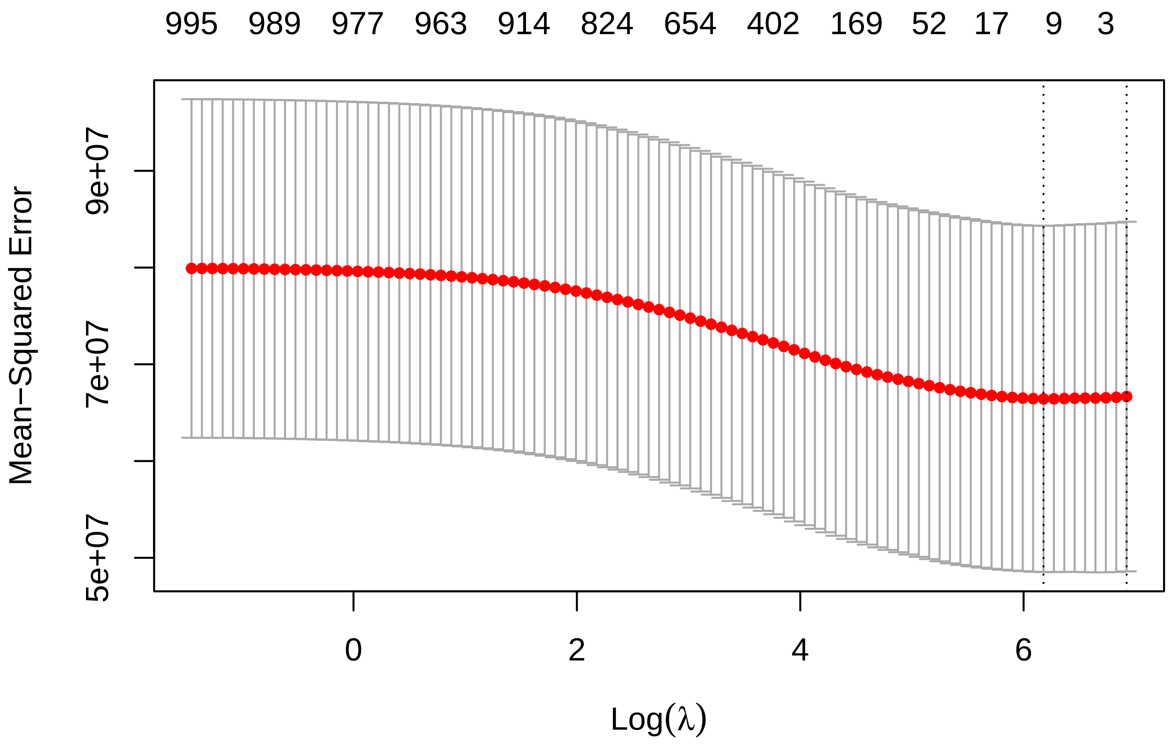
```
## weather.com
##          52
```

Exercise 3

To find the best two linear predictors, lasso was used. First, a 5-fold cross validation was performed to determine the optimal lambda parameter. In the plot below the result of the cross validation is shown. Where the minimum of the MSE is reached is the optimal value of the lambda parameter.

```
lasso.cv <- cv.glmnet(x = as.matrix(browser_2006[!names(browser_2006) %in%
                                                     c("id", "spend")]),
                      y = browser_2006$spend,
                      type.measure = "mse",
                      family = "gaussian",
                      nfolds = 5,
                      alpha = 1)

plot(lasso.cv)
```



Then, the largest coefficients in terms of their absolute value of the lasso model with the optimal lambda from before were determined, since they have the greatest influence on the prediction. The best two linear predictors are therefore the two websites *staples.com* and *officedepot.com*.

```
coef.lasso.cv <- coef(lasso.cv, s = "lambda.min")
(best.lin.pred <- data.frame(name = coef.lasso.cv@Dimnames[[1]][coef.lasso.cv@i + 1],
                             coefficient = coef.lasso.cv@x) %>%
      .[order(abs(.$coefficient), decreasing = T),] %>%
      .[1:2,])
```

```
##              name coefficient
## 9     staples.com    3113.788
## 8 officedepot.com    2053.332
```

Exercise 4

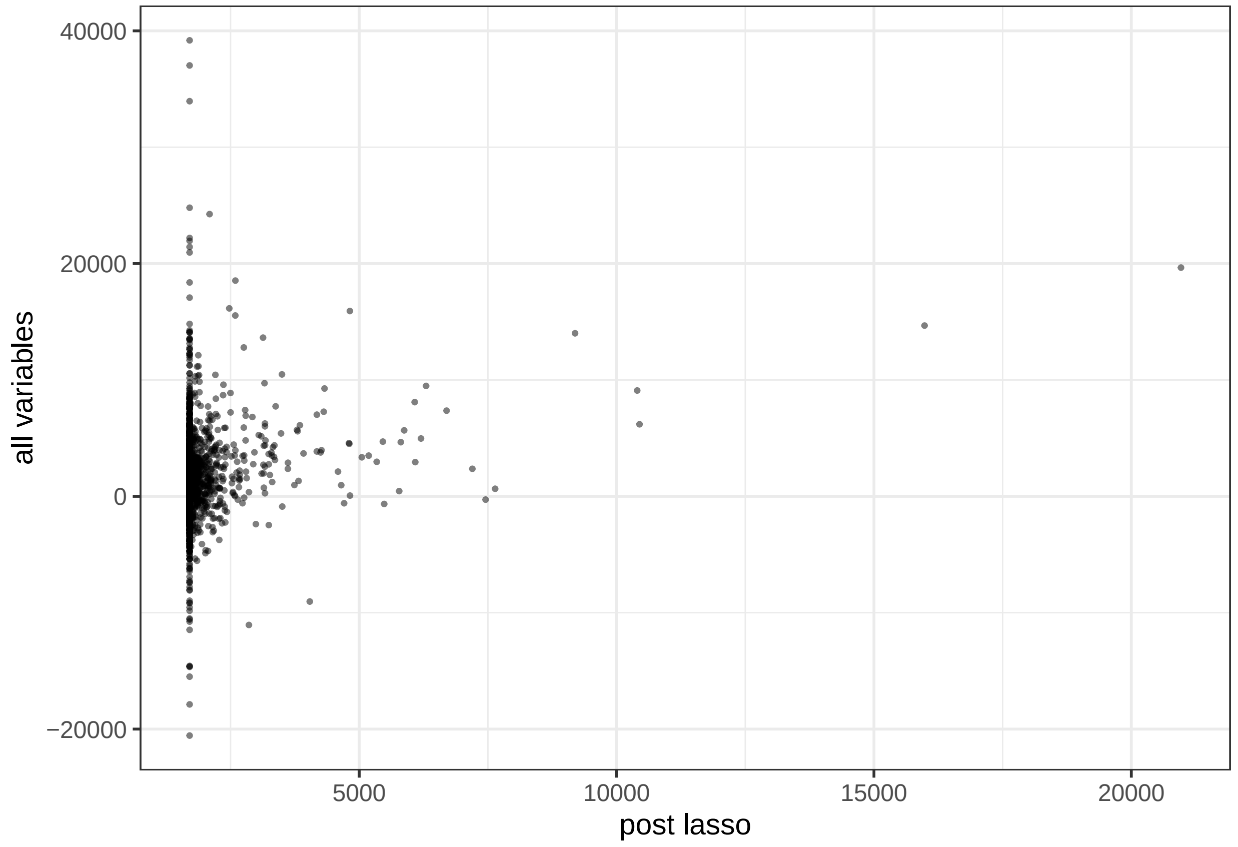
```
post_lasso = lm(data = browser_2006, "spend ~ officedepot.com + staples.com")
all_variables = lm(data = browser_2006, "spend ~ . - id")

predict_post_lasso = predict(post_lasso, newdata = browser_new)
predict_all_variables = predict(all_variables, newdata = browser_new)

predictions = data.frame(cbind(predict_post_lasso,predict_all_variables))
(summary(predictions))
```

```
##  predict_post_lasso predict_all_variables
##  Min.   : 1705      Min.   :-20557.57    
##  1st Qu.: 1705      1st Qu.:   -91.79    
##  Median : 1705      Median :  1446.64    
##  Mean   : 1898      Mean   :  1785.24    
##  3rd Qu.: 1787      3rd Qu.:  3198.47    
##  Max.   :20965      Max.   : 39182.27
```

```
(ggplot(data = predictions) +
  geom_point(aes(x = predict_post_lasso, y = predict_all_variables),
             size = 0.5, alpha = 0.5) +
  labs(x = "post lasso", y = "all variables") +
  theme_bw())
```



```
(correlations = cor(predictions))
```

```
##                       predict_post_lasso predict_all_variables
## predict_post_lasso             1.0000000             0.1849663
## predict_all_variables          0.1849663             1.0000000
```

What we can see in our plot is that our "all variables" model predicts a significant amount of values < 0, which is not reasonable. Additionally, the outliers are much higher (almost 40k versus 20k). For the Lasso model, however, the smallest prediction for money spending is 1'705, which is also not reasonable. In comparison the Lasso model is still more reasonable.