Data Analytics I: Predictive Econometrics
Assignment III

Lauritz Storch
Sebastian Maser
Ivan Medvedev
Jan Gutjahr

2022-12-11

Data Preparation and Package Installation

Download the data sets browser_2006.csv and browser_new.csv[1] from Canvas. Load the data into R. Generate matrices for the outcome, control, as well as id variables for the 2006 and the new data. Install and load the packages grf, DiagrammeR, and glmnet.

```
# # Clear R
# rm(list=ls())
# cat("\014")

## Set working directory to load data
setwd("~/Downloads/")



## Packages and Library
# Packages function
install_if_missing <- function(p) {
  if (p %in% rownames(installed.packages()) == F) {
    try(install.packages(p, dependencies = T))
  } else {
    cat(paste("Skipping already installed package:", p, "\n"))
  }
}

# Define packages
packages = c("grf", "DiagrammeR", "glmnet")

# Install
invisible(sapply(packages, install_if_missing))
```

```
## Skipping already installed package: grf
## Skipping already installed package: DiagrammeR
## Skipping already installed package: glmnet
```

```
# read library
for(pkg in packages){
  library(pkg, character.only = TRUE)
```

[1] Please make sure your device is connected to the internet and you are logged into Canvas to download the dataset.

```
}
```

```
## Lade nötiges Paket: Matrix
```

```
## Loaded glmnet 4.1-6
```

```
## Read in data
# Load data
data_2006 <-read.csv("browser_2006.csv")
data_new <-read.csv("browser_new.csv")

# Data preparation
# 2006 data
y_2006 <- as.matrix(data_2006[, 2])
x_2006 <- as.matrix(data_2006[, c(3:ncol(data_2006))])
id_2006 <- as.matrix(data_2006[, 1])
# new data
x_new <- as.matrix(data_new[, c(2:ncol(data_new))])
id_new <- as.matrix(data_new[, 1])
```

Task 1

Q: How much is the average online spending in 2006? (0.5 points)

```
cat("The average online spending in 2006 is: ", mean(y_2006))
```

```
## The average online spending in 2006 is:  1959.921
```

Task 2

Q: On which webpage is the household with id = 1297 (first row of the 2006 sample) most of the time? (0.5 points)

```
# id <- which(id_2006 == 1297)
max_time <- max(x_2006[1,])
website_max <- names(x_2006[1,which(x_2006[1,] == max_time)])

cat("The household with id = 1297 spents most of its time (", max_time,"%)
on the website ", website_max)
```

```
## The household with id = 1297 spents most of its time ( 16.28118 %)
## on the website  weather.com
```

Task 3

Q: Which two webpages are together the best linear predictors for online spendings in 2006? You can use a Lasso with only two active control variables to answer the question. (1 point)

```
set.seed(2122020)
options(warn = -1) # turn warnings off
lasso.cv <- cv.glmnet(x_2006, y_2006, type.measure = "mse", family = "gaussian",
                      alpha = 1, pmax = 2) # Use Lasso for variable selection
options(warn = 0) # turn warnings on
coef_lasso1 <- coef(lasso.cv, s = "lambda.min")
index_lasso <- c(which(coef_lasso1 > 0)) # search for the index of selected variables
coef_lasso1[index_lasso,1] # print the variables and regarding coefficients
```

```
##     (Intercept) officedepot.com     staples.com
##       1944.3716        416.4678        206.3726
```

Task 4

Q: Estimate a post-Lasso model on your solution to the previous exercise. Use this model and a second OLS model that includes all 1000 websites as variables to predict the spendings for the households in the file browser new.csv. Plot the predictions against each other and calculate the correlation of your predictions. Do you think Lasso has selected a reasonable set of variables? (1 point)

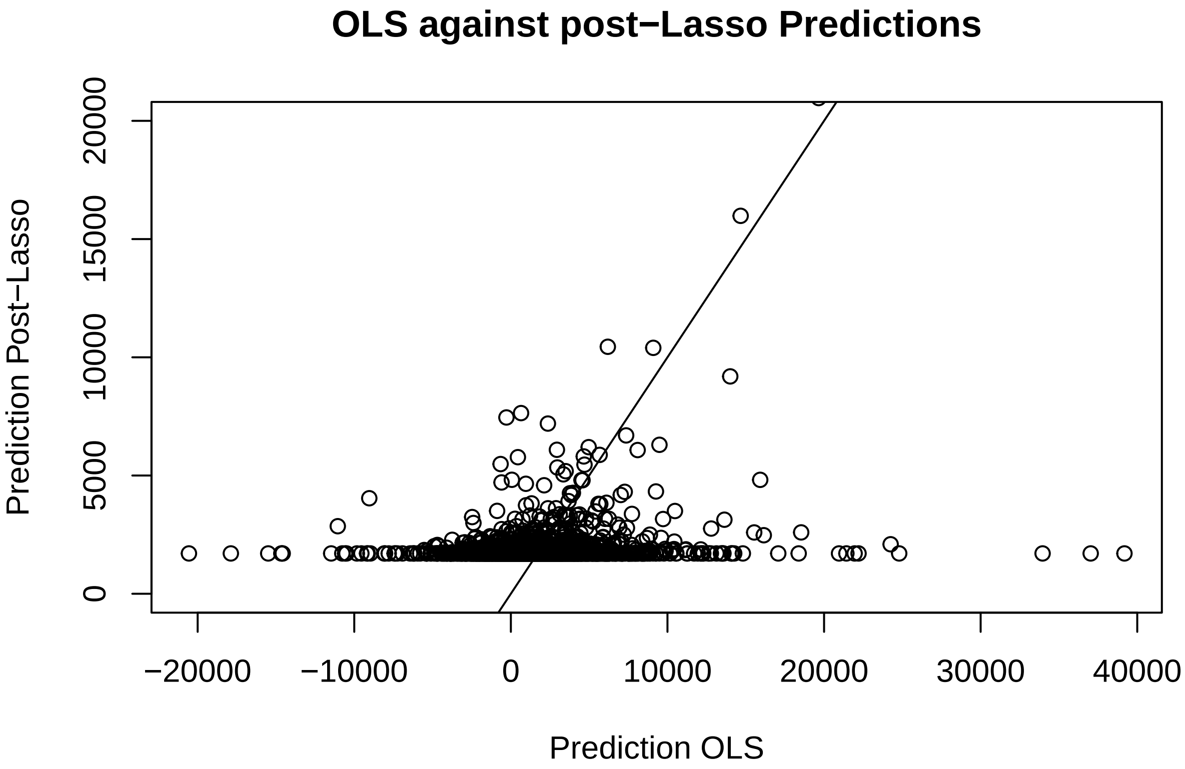
```
## Post-Lasso regression
data_no_id <- data_2006[,-1]
post_lasso <- lm(formula = spend ~ officedepot.com + staples.com, data = data_no_id)


## OLS
ols <- lm(formula = spend ~ .-vistaprint.com,
          data = data_no_id, na.action=na.exclude) # w.l.o.g can drop vistaprint.com
# lm(y_2006 ~ x_2006) # double check


## Comparison of coefficients (not asked)
# post_lasso$coefficients  # Post-Lasso
# ols$coefficients         # OLS


## Predict spendings in new data set
data_new$predlasso <- predict(post_lasso, newdata = data_new)
data_new$predols <- predict(ols, newdata = data_new)


## Plot
plot(data_new$predols, data_new$predlasso,ylim=c(0,20000),
     xlab = "Prediction OLS", ylab = "Prediction Post-Lasso",
     main = "OLS against post-Lasso Predictions")
par(new=TRUE)
abline(a=0,b=1)
```



```
## Correlation
cat("The correlation of OLS and post-Lasso Predictions is ",
round(cor(data_new$predols,data_new$predlasso),4))
```

```
## The correlation of OLS and post-Lasso Predictions is  0.185
```

Remark: We drop the variable "vistaprint.com" because it creates a NA in the OLS estimation. Dropping the variable does not influence the other variables estimation procedure, which can be checked by the out-commented line "# lm(y_2006 ~ x_2006) # double check".

Answer: The plot and the correlation both indicate that the post-Lasso prediction performs poorly compared to the OLS regression[2]. As the data set has 1000 control variables, Lasso selects only the two best predictive linear predictors and misses out on the information of 998 other variables. Therefore, the Lasso did not select a reasonable set of variables as it was restricted to only two active control variables. Breaking this restriction will most probably improve the post-Lasso performance, as the Lasso regression usually performs better than OLS.

[2] Points do not align on the plotted line, and the correlation coefficient is small