Lista 2 - Mineração

Victor Alves Dogo Martins, RA: 744878   Ana Beatriz Alves Monteiro, RA: 727838
Larissa Torres, RA: 631914

17-07-2022

Item 1

```r
# ITEM 1

set.seed(727838)

library(tidyverse)
library(rsample)
library(glmnet)
library(ggrepel)
library(forcats)
library(knitr)
library(kableExtra)

df <- ISLR::Carseats
df$US <-  as.factor(df$US)
df$Urban <- as.factor(df$Urban)
df$ShelveLoc <- as.factor(df$ShelveLoc)

# Divisao treino e teste

split <- initial_split(df, prop=0.6)

tre <- training(split)
tes <- testing(split)

x_tre <- model.matrix(Sales~., tre)
y_tre <- tre[,1]

x_tes <- model.matrix(Sales~., tes)
y_tes <- tes[,1]
```

Item 2

```r
# ITEM 2
```

```r
## Ajuste Minimos Quadrados

ajuste_mq <- glmnet(x_tre, y_tre, alpha=0, lambda=0)

## Ajuste Lasso

cv_lasso <- cv.glmnet(x_tre, y_tre, alpha=1)
ajuste_lasso <- glmnet(x_tre, y_tre, alpha=1, lambda = cv_lasso$lambda.1se)

## Erro x Lambda Lasso

tibble(
  lambda=cv_lasso$lambda,
  risco=cv_lasso$cvm
) |>
  ggplot()+
  aes(x=lambda, y=risco)+
  geom_line()+
  geom_vline(xintercept = cv_lasso$lambda.1se)+
  annotate(geom = 'text', y=5, x=0.25,
           label=paste0('lambda = ', round(ajuste_lasso$lambda,5)))+
  theme_minimal()+
  labs(title='Risco x Lambda estimado',
       subtitle = 'com Lambda escolhido')
```


Risco x Lambda estimado
com Lambda escolhido
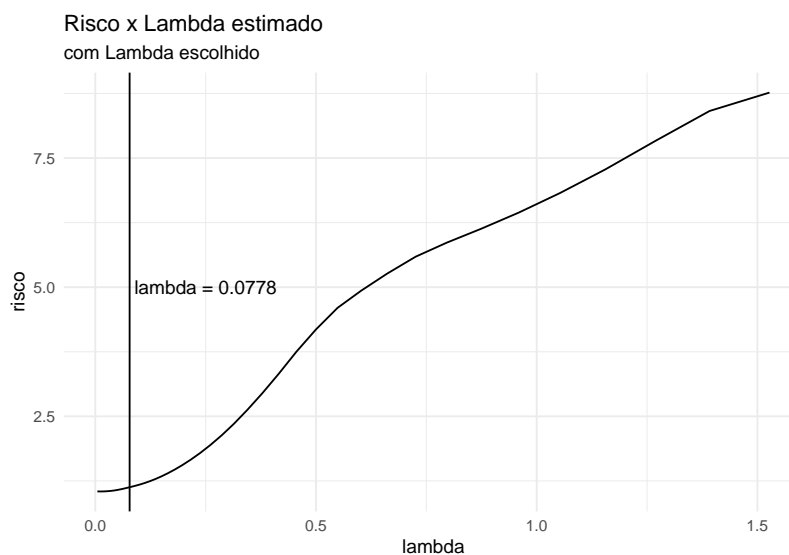
```r
## Apresentando melhor lambda (cv_lasso$lambda.1se)

lambdas <- cv_lasso$glmnet.fit$lambda

lam <- lambdas |>
  as.data.frame() |>
  mutate(penalty=names(cv_lasso$glmnet.fit$a0)) %>%
  rename(lambda=1)
```

```r
results <- cv_lasso$glmnet.fit$beta |>
  as.matrix() |>
  as.data.frame() |>
  rownames_to_column() |>
  gather(penalty,coefficients,-rowname) |>
  left_join(lam)

result_labels <- results |>
  group_by(rowname) |>
  filter(lambda==ajuste_lasso$lambda) |>
  ungroup()

ggplot()+
  geom_line(data=results, aes(lambda, coefficients,
                              group=rowname, color=rowname),
            show.legend = FALSE)+
  scale_x_log10()+
  geom_text(data=result_labels, aes(0.01, coefficients,
                                    label=rowname, color=rowname),
            nudge_x=-.06, show.legend = FALSE)+
  geom_vline(xintercept = ajuste_lasso$lambda)+
  annotate(geom = 'text', y=0.4, x=0.16,
           label=paste0('lambda = ', round(ajuste_lasso$lambda,5)))+
  theme_minimal()+
  labs(y='estimativa',
       title='Lambda x Estimativas dos Coeficientes')
```


Item 3

```r
# ITEM 3

coef_mq <- coefficients(ajuste_mq) |>
```

```r
    round(3)

coef_lasso <- coefficients(ajuste_lasso) |>
  round(3)

## Apresentando coeficientes

tibble(
  var=names(coef_mq[,1]),
  mq=coef_mq[,1],
  lasso=coef_lasso[,1]
) |> slice(-2) |>
  kable('latex', align='ccc',
        caption = 'Estimativas de Coeficientes para Mínimos Quadrados e Lasso',
        col.names=c('Variável', 'Mínimos Quadrados', 'Lasso')) |>
  kable_styling(position="center",
                latex_options="HOLD_position")
```

Table 1: Estimativas de Coeficientes para Mínimos Quadrados e Lasso

Variável	Mínimos Quadrados	Lasso
(Intercept)	6.018	6.672
CompPrice	0.093	0.080
Income	0.014	0.012
Advertising	0.119	0.110
Population	0.000	0.000
Price	-0.098	-0.089
ShelveLocGood	4.741	4.339
ShelveLocMedium	1.970	1.636
Age	-0.047	-0.042
Education	-0.032	-0.008
UrbanYes	0.230	0.033
USYes	0.089	0.021

```r
## Barras coeficientes minimos quadrados

tibble(
  var=names(coef_mq[,1]),
  val=coef_mq[,1]
) |>
  mutate(sinal=ifelse(val<0, 'negativo', 'positivo'),
         var=fct_reorder(var, val)) |>
  arrange(val) |>
  filter(var!='(Intercept)') |>
  ggplot()+
  aes(x=val, y=var, fill=sinal)+
  geom_bar(stat='identity', show.legend = FALSE)+
  theme_minimal()+
  labs(x='Estimativa', y='',
       title='Estimativas do Ajuste de Mínimos Quadrados')
```

Estimativas do Ajuste de Mínimos Quadrados
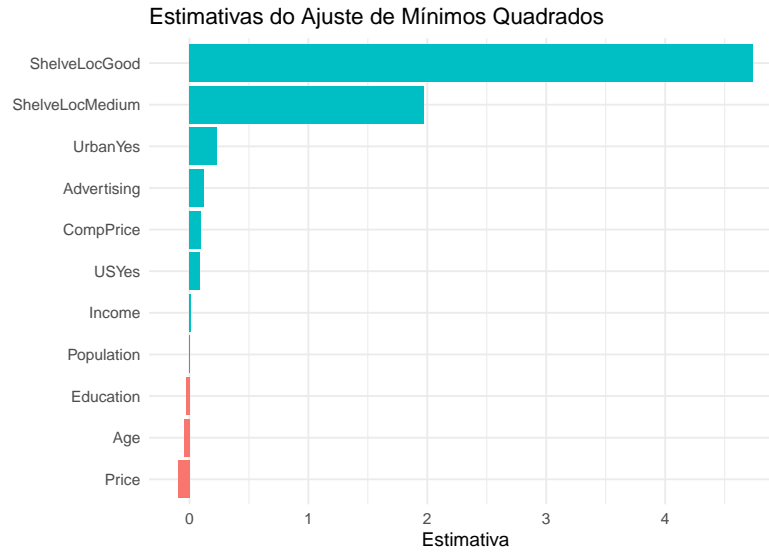

```
## Barras coeficientes lasso

tibble(
  var=names(coef_lasso[,1]),
  val=coef_lasso[,1]
) |>
  mutate(sinal=ifelse(val<0, 'negativo', 'positivo'),
         var=fct_reorder(var, val)) |>
  arrange(val) |>
  filter(var!='(Intercept)') |>
  ggplot()+
  aes(x=val, y=var, fill=sinal)+
  geom_bar(stat='identity', show.legend = FALSE)+
  theme_minimal()+
  labs(x='Estimativa', y='',
       title='Estimativas do Ajuste Lasso')
```

Estimativas do Ajuste Lasso
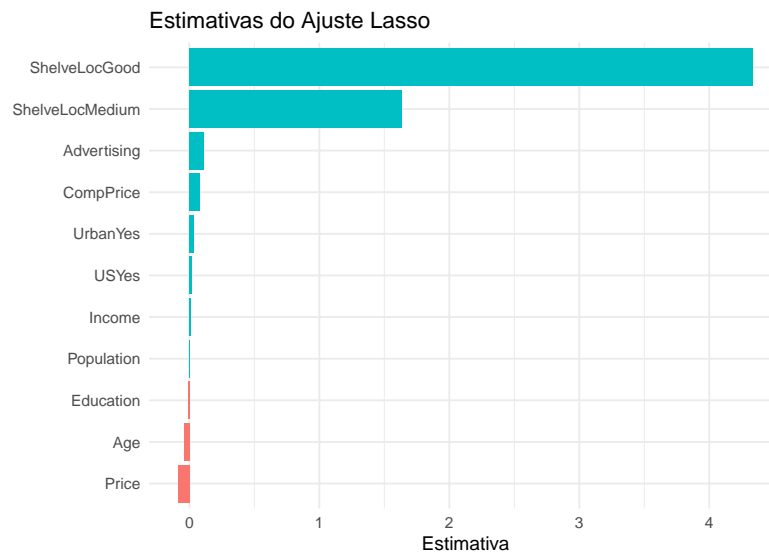
Item 4

```r
# ITEM 4

funcao_risco <- function(y_pred, y_obs){
  w <- (y_pred-y_obs)^2
  sigma <- var(w)
  risco <- mean(w)
  liminf <- risco - (2*sqrt((1/length(w))*sigma))
  limsup <- risco + (2*sqrt((1/length(w))*sigma))

  return(tibble(risco, liminf, limsup))
}

y_pred_mq <- predict(ajuste_mq, x_tes)
y_pred_lasso <- predict(ajuste_lasso, x_tes)

risco_mq <- funcao_risco(y_pred_mq, y_tes)
risco_lasso <- funcao_risco(y_pred_lasso, y_tes)

tibble(
  Estimativa=c('Risco', 'Limite Inferior', 'Limite Superior'),
  `Mínimos Quadrados`=unlist(c(risco_mq)),
  `Lasso`=unlist(c(risco_lasso))
) |>
  kable('latex', align='ccc',
        caption = 'Risco e Intervalos de Confiança para Lasso e Mínimos Quadrados',
        col.names=c('Variável', 'Mínimos Quadrados', 'Lasso')) |>
  kable_styling(position="center",
                latex_options="HOLD_position")
```

Table 2: Risco e Intervalos de Confiança para Lasso e Mínimos Quadrados

Variável	Mínimos Quadrados	Lasso
Risco	1.1815323	1.1869124
Limite Inferior	0.9231082	0.9441733
Limite Superior	1.4399564	1.4296515

Item 5