PPA-1.R

Yiming Zhang

2023-05-01

```
library(readxl)
library(dplyr)

## 
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
## 
##     filter, lag

## The following objects are masked from 'package:base':
## 
##     intersect, setdiff, setequal, union

library(tidyr)
library(ggplot2)
library(reshape2)

## 
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
## 
##     smiths

library(glmnet)

## Loading required package: Matrix

## 
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
## 
##     expand, pack, unpack

## Loaded glmnet 4.1-7

# Loading
setwd("C:/Users/Yiming Zhang/Desktop/R_workspace")
reader <- read_excel("Iowa_Housing_Data_Mod.xlsx")

## Warning: Expecting numeric in L2184 / R2184C12: got 'NA'
## Warning: Expecting numeric in M2184 / R2184C13: got 'NA'

## New names:
## • `` -> `...29`
## • `` -> `...30`

# Data process
columns <- c('Id', 'LotArea', 'OverallQual', 'OverallCond',
'YearBuilt', 'BsmtUnfSF',
             'TotalBsmtSF', 'CentralAir', '1stFlrSF', '2ndFlrSF',
'GrLivArea',
             'BsmtFullBath', 'BsmtHalfBath', 'FullBath', 'HalfBath',
'BedroomAbvGr',
             'KitchenAbvGr', 'TotRmsAbvGrd', 'Fireplaces',
'GarageCars',
             'GarageArea', 'WoodDeckSF', 'OpenPorchSF',
'EnclosedPorch',
             'ScreenPorch', 'PoolArea', 'YrSold', 'SalePrice')
features <- c('Id', 'LotArea', 'OverallQual', 'OverallCond',
'YearBuilt', 'BsmtUnfSF',
              'TotalBsmtSF', 'CentralAir', '1stFlrSF', '2ndFlrSF',
'GrLivArea',
              'BsmtFullBath', 'BsmtHalfBath', 'FullBath', 'HalfBath',
'BedroomAbvGr',
              'KitchenAbvGr', 'TotRmsAbvGrd', 'Fireplaces',
'GarageCars',
              'GarageArea', 'WoodDeckSF', 'OpenPorchSF',
'EnclosedPorch',
              'ScreenPorch', 'PoolArea', 'YrSold')
Target <- 'SalePrice'

data <- reader %>% select(all_of(columns))

data <- data %>%
  mutate(`Age of House` = YrSold - YearBuilt,
         `CentralAC Dummy` = ifelse(CentralAir == 'Y', 1, 0)) %>%
  na.omit() %>%
  select(-c(YrSold, YearBuilt, CentralAir))

# Scale features and target (SalePrice) to mean=0 and SD=1
features <- data %>% select(-SalePrice)
target <- data %>% select(SalePrice)

# Scale features and target
scaled_features <- scale(features)
scaled_target <- scale(target)

# Create a new data frame with the scaled data
data <- as.data.frame(cbind(scaled_features, SalePrice=scaled_target))
```

```r
# Linear regression
model <- lm(SalePrice ~ GrLivArea, data)
summary(model)
```

```
##
## Call:
## lm(formula = SalePrice ~ GrLivArea, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5816 -0.3851 -0.0193  0.2976  4.2245
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.838e-17  1.284e-02    0.00        1
## GrLivArea   7.219e-01  1.284e-02   56.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6921 on 2905 degrees of freedom
## Multiple R-squared:  0.5212, Adjusted R-squared:  0.521
## F-statistic:  3162 on 1 and 2905 DF,  p-value: < 2.2e-16
```

```r
# Multiple linear regression
features <- c('LotArea', 'OverallQual', 'OverallCond', 'Age of House',
'CentralAC Dummy', 'GrLivArea', 'GarageCars')
model2 <- lm(SalePrice ~ ., data = data %>% select(SalePrice,
all_of(features)))
summary(model2)
```

```
##
## Call:
## lm(formula = SalePrice ~ ., data = data %>% select(SalePrice,
##     all_of(features)))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5851 -0.2723 -0.0257  0.2150  3.5886
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.573e-16  8.212e-03   0.000   1.0000
## LotArea             1.231e-01  8.567e-03  14.369   <2e-16 ***
## OverallQual         4.085e-01  1.293e-02  31.593   <2e-16 ***
## OverallCond         8.173e-02  9.521e-03   8.584   <2e-16 ***
## `Age of House`     -2.071e-01  1.260e-02 -16.436   <2e-16 ***
## `CentralAC Dummy`  -1.976e-02  9.330e-03  -2.118   0.0343 *
## GrLivArea           3.607e-01  1.092e-02  33.037   <2e-16 ***
## GarageCars          1.185e-01  1.128e-02  10.507   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4428 on 2899 degrees of freedom
## Multiple R-squared:  0.8044, Adjusted R-squared:  0.804
## F-statistic:  1704 on 7 and 2899 DF,  p-value: < 2.2e-16
```

```
# Lasso regression
train <- data[1:1800,]
val <- data[1801:2400,]
test <- data[2401:2907,]

# Prepare the data
X_train <- model.matrix(SalePrice ~ ., train)[, -1]
y_train <- train$SalePrice

X_val <- model.matrix(SalePrice ~ ., val)[, -1]
y_val <- val$SalePrice

X_test <- model.matrix(SalePrice ~ ., test)[, -1]
y_test <- test$SalePrice

# Here we produce results for alpha=0.05 which corresponds to
lambda=0.1 in Hull's book
lambda <- 0.1/2
fit_lasso <- glmnet(X_train, y_train, alpha = 1, lambda = lambda,
standardize = FALSE)

# features and its respective coefficients
coef(fit_lasso)
```

```
## 27 x 1 sparse Matrix of class "dgCMatrix"
##                             s0
## (Intercept)       -0.009282444
## Id                 .
## LotArea            0.039306852
## OverallQual        0.336039810
## OverallCond        0.005920390
## BsmtUnfSF         -0.010263041
## TotalBsmtSF        0.158218912
## `1stFlrSF`         0.039731462
## `2ndFlrSF`         .
## GrLivArea          0.298517944
## BsmtFullBath       0.045738996
## BsmtHalfBath       .
## FullBath           .
## HalfBath           .
## BedroomAbvGr       .
## KitchenAbvGr      -0.026249322
## TotRmsAbvGrd       .
## Fireplaces         0.019987353
## GarageCars          0.021653665
## GarageArea          0.081679042
## WoodDeckSF          0.008250897
## OpenPorchSF         .
## EnclosedPorch       .
## ScreenPorch         .
## PoolArea            .
## `Age of House`     -0.095514433
## `CentralAC Dummy`   .
```

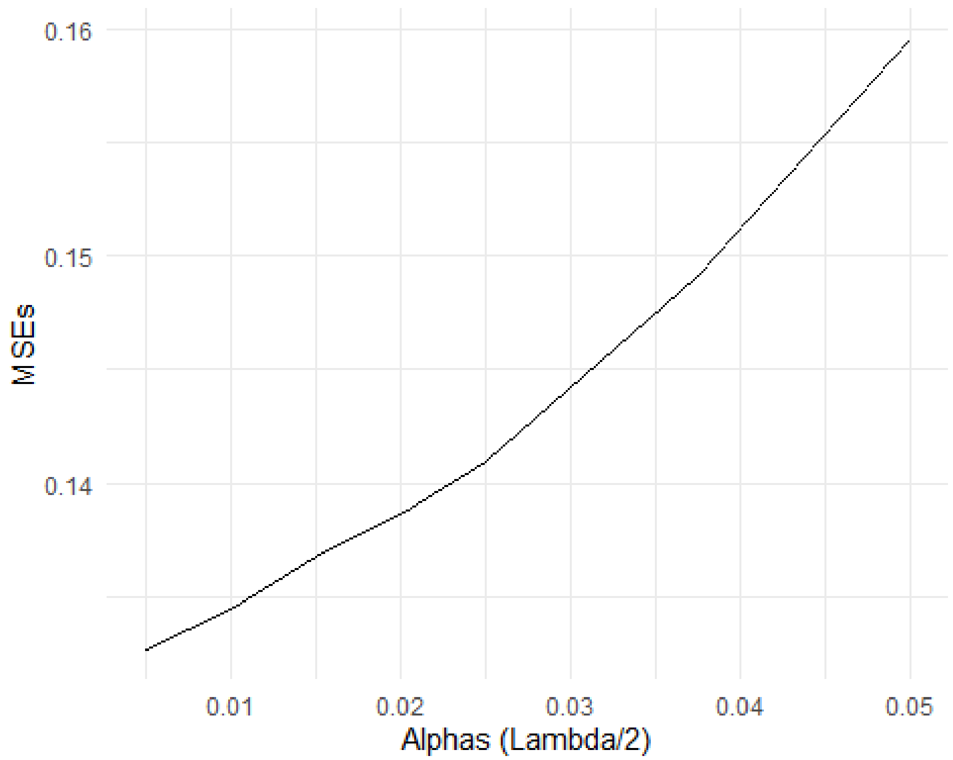
```
# We now consider different lambda values. The alphas are half the
lambdas
lambdas <- c(0.01, 0.02, 0.03, 0.04, 0.05, 0.075, 0.1) / 2
mses <- c()

for (lambda in lambdas) {
  lasso <- glmnet(X_train, y_train, alpha = 1, lambda = lambda,
standardize = FALSE)
  preds <- predict(lasso, X_val)
  mse_val <- mean((y_val - preds)^2)
  mses <- c(mses, mse_val)
  print(mse_val)
}
```

```
## [1] 0.132666
## [1] 0.1344193
## [1] 0.1367332
## [1] 0.1386108
## [1] 0.1409515
## [1] 0.1491544
## [1] 0.1595729
```

```
# Plot alphas (lambdas/2) vs MSEs
qplot(lambdas, mses, geom = 'line') +
  xlab("Alphas (Lambda/2)") +
  ylab("MSEs") +
  theme_minimal()
```

```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this
warning was
## generated.
```



```
# Calculate MSE for test set when Hull's lambda = 0.04
lambda <- 0.04 / 2
lasso <- glmnet(X_train, y_train, alpha = 1, lambda = lambda,
standardize = FALSE)
preds <- predict(lasso, X_test)
mse_test <- mean((y_test - preds)^2)
print(mse_test)

## [1] 0.1387251

# Calculate MSE for test set when Hull's lambda = 0.1
lambda <- 0.1 / 2
lasso <- glmnet(X_train, y_train, alpha = 1, lambda = lambda,
standardize = FALSE)
preds <- predict(lasso, X_test)
mse_test <- mean((y_test - preds)^2)
print(mse_test)

## [1] 0.1565691
```