ValidationSetApproach.R

vijaykalmath

2022-01-04

```
# Validation Set Approach

library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ISLR)
library(boot)

set.seed(1)


# Get 196 random indexes
train = sample(392,196)

length(train)
```

```
## [1] 196
```

```
# using Auto data and target variable is mpg

colnames(Auto)
```

```
## [1] "mpg"          "cylinders"    "displacement" "horsepower"   "weight"
## [6] "acceleration" "year"         "origin"       "name"
```

```
lm.fit  = lm(mpg~horsepower,data = Auto,subset = train)


summary(lm.fit)
```

```
##
## Call:
## lm(formula = mpg ~ horsepower, data = Auto, subset = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3177 -3.5428 -0.5591  2.3910 14.6836
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 41.283548   1.044352   39.53   <2e-16 ***
## horsepower  -0.169659   0.009556  -17.75   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 5.032 on 194 degrees of freedom
## Multiple R-squared:  0.619,  Adjusted R-squared:  0.6171
## F-statistic: 315.2 on 1 and 194 DF,  p-value: < 2.2e-16
```

```
# Calculate MSE

mean((Auto$mpg - predict(lm.fit,Auto))[-train]^2)
```

```
## [1] 23.26601
```

```
lm.fit2 = lm(mpg~poly(horsepower,2),data = Auto,subset = train)
lm.fit3  = lm(mpg~poly(horsepower,3),data = Auto,subset = train)

summary(lm.fit2)
```

```
##
## Call:
## lm(formula = mpg ~ poly(horsepower, 2), data = Auto, subset = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.8711  -2.6655  -0.0096   2.0806  16.1063
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            23.5496     0.3175  74.182  < 2e-16 ***
## poly(horsepower, 2)1 -123.5881     6.4587 -19.135  < 2e-16 ***
## poly(horsepower, 2)2   47.7189     6.3613   7.501 2.25e-12 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.439 on 193 degrees of freedom
## Multiple R-squared:  0.705,  Adjusted R-squared:  0.702
## F-statistic: 230.6 on 2 and 193 DF,  p-value: < 2.2e-16
```

```
summary(lm.fit3)
```

```
##
## Call:
## lm(formula = mpg ~ poly(horsepower, 3), data = Auto, subset = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.6625  -2.7108   0.0805   2.0724  16.1378
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           23.5527     0.3185  73.946  < 2e-16 ***
## poly(horsepower, 3)1 -123.6143     6.4755 -19.089  < 2e-16 ***
## poly(horsepower, 3)2   47.8284     6.3935   7.481 2.58e-12 ***
## poly(horsepower, 3)3    1.3825     5.8107   0.238    0.812
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.45 on 192 degrees of freedom
## Multiple R-squared:  0.7051, Adjusted R-squared:  0.7005
## F-statistic:   153 on 3 and 192 DF,  p-value: < 2.2e-16
```

```
# Mean of Quadratic
mean((Auto$mpg - predict(lm.fit2,Auto))[-train]^2)
```

```
## [1] 18.71646
```

```
mean((Auto$mpg - predict(lm.fit3,Auto))[-train]^2)
```

```
## [1] 18.79401
```