Lista 09

Matheus Cougias

28/02/2021

Leitura dos dados: Baseado nos resultados da lista anterior, a regressão será aplicada diretamente no modelo onde foi aplicado o logaritmo aos dados, devido sua maior capacidade tanto de compreensão dos dados (R^2 múltiplo) quanto de sua capacidade preditiva (R^2 preditivo).

```
dados <- read.delim("boston_corrected.txt")
dados <- subset(dados, select = -c(OBS., TOWN, TOWN., TRACT, LON, LAT, CMEDV))
dt <- data.frame(MEDV=log(dados$MEDV),
                 CRIM=log(dados$CRIM),
                 ZN=log(dados$ZN+1),
                 INDUS=log(dados$INDUS),
                 CHAS=log(dados$CHAS+1),
                 NOX=log(dados$NOX),
                 RM=log(dados$RM),
                 AGE=log(dados$AGE),
                 DIS=log(dados$DIS),
                 RAD=log(dados$RAD),
                 TAX=log(dados$TAX),
                 PTRATIO=log(dados$PTRATIO),
                 B=log(dados$B),
                 LSTAT=log(dados$LSTAT))
```

Modelo de regressão linear múltipla (com logaritmo): Através do modelo de regressão linear múltipla onde o logaritmo foi aplicado, alguns dados podem ser levantados: - Possui um R^2 múltiplo de 76,96% e R^2 ajustado de 76,35%; - Possu um R^2 preditivo de 75,42197%; - As variáveis CRIM e NOX se destacam com os maiores valores de VIF, 86,61% e 83,33% respectivamente.

```
require(car)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
modelo <- lm(MEDV ~ CRIM + ZN + INDUS + CHAS + NOX + RM + AGE + DIS + RAD + TAX + PTRATIO + B + LSTAT, d
summary(modelo)
```

```
##
## Call:
## lm(formula = MEDV ~ CRIM + ZN + INDUS + CHAS + NOX + RM + AGE +
##     DIS + RAD + TAX + PTRATIO + B + LSTAT, data = dt)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.95032 -0.10303 -0.00257  0.10887  0.82889
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.451574   0.437232  12.468  < 2e-16 ***
## CRIM        -0.031427   0.011183  -2.810 0.005148 **
## ZN          -0.016429   0.008551  -1.921 0.055274 .
## INDUS        0.007949   0.021823   0.364 0.715834
## CHAS         0.142429   0.052235   2.727 0.006625 **
## NOX         -0.367113   0.107536  -3.414 0.000693 ***
## RM           0.389532   0.109522   3.557 0.000412 ***
## AGE          0.042937   0.022258   1.929 0.054296 .
## DIS         -0.138655   0.036043  -3.847 0.000135 ***
## RAD          0.100878   0.022197   4.545 6.93e-06 ***
## TAX         -0.192292   0.045509  -4.225 2.84e-05 ***
## PTRATIO     -0.608403   0.094695  -6.425 3.12e-10 ***
## B            0.053979   0.012647   4.268 2.37e-05 ***
## LSTAT       -0.416449   0.025819 -16.130  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1988 on 492 degrees of freedom
## Multiple R-squared:  0.7696, Adjusted R-squared:  0.7635
## F-statistic: 126.4 on 13 and 492 DF,  p-value: < 2.2e-16
```

```
1 - 1/vif(modelo)
```

```
##       CRIM         ZN      INDUS       CHAS        NOX         RM        AGE
## 0.86613805 0.59263610 0.72781198 0.07468206 0.83330488 0.48292797 0.57651559
##        DIS        RAD        TAX    PTRATIO          B      LSTAT
## 0.79307218 0.79246685 0.75949611 0.43987992 0.18226256 0.67487689
```

```
Y <- as.matrix(dt$MEDV)
SQT <- sum((Y - mean(Y))^2)

X <- model.matrix(modelo)

H <- X%*%solve(t(X)%*%X)%*%t(X)
res <- Y - H%*%Y
aux <- 1 - diag(H)
res_del <- res/(1-diag(H))
SQRes <- sum(res_del^2)
(R2pred <- 1 - SQRes/SQT)
```

```
## [1] 0.7542197
```

Melhorando O R^2 preditivo: Na busca por um melhor resultado de R^2 preditivo, aplica-se um ajuste do Estimador de Ridge Regression, gerando um R^2 preditivo de 75,45716%.

```r
beta <- solve(t(X)%*%X)%*%t(X)%*%Y
Identidade <- diag(c(0,rep(1, nrow(beta)-1)))

R2predFun <- function(lmbd)
{
  Identidade <- diag(c(0,rep(1, nrow(beta)-1)))
  beta <- solve(t(X)%*%X + lmbd*Identidade)%*%t(X)%*%Y
  H <- X%*%solve(t(X)%*%X + lmbd*Identidade)%*%t(X)
  res <- Y - H%*%Y
  aux <- 1 - diag(H)
  res_del <- res/(1-diag(H))
  SQRes <- sum(res_del^2)
  R2pred <- 1 - SQRes/SQT
  return(R2pred)
}

saida <- optimize(R2predFun, lower=0.001, upper=1.2, maximum = TRUE)
lambda <- saida$maximum
(saida$objective)
```

```
## [1] 0.7545716
```

```r
beta <- solve(t(X)%*%X + lambda*Identidade)%*%t(X)%*%Y
```

Análise dos resultados: Ao compararmos os resutados, houve uma leve melhora entre o modelo inicial e o modelo de regressão onde o Estimador de Ridge foi apicado, cerca de 0,03519%. Uma ideia de que talvez o estimador não tenha afetado tão positivamente o modelo de regressão pode ser dada por, talvez, pela sua sensibilidade em relação aos outliers (ou pontos discrepantes) da base de dados. Para melhorar esse resultado, seria interessante realizar um tratamento prévio dos dados, de modo que o modelo se adeque melhor às informações.