Lab Exercise 1 - Problem 2

Spencer Moon

10/15/2017

Before building the regression model, the data was loaded and normalized with the following code:

```
library(tidyverse)

bostonhousing <- read_tsv("bostonhousing.txt")
bostonhousing$CHAS <- factor(bostonhousing$CHAS)
```

Part A

Below is the linear model for the Boston housing data:

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

```
##
## Call:
## lm(formula = MEDV ~ CRIM + ZN + INDUS + factor(CHAS) + NOX +
##     RM + AGE + DIS + RAD + TAX + PTRATIO + B + LSTAT, data = bostonhousing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.595  -2.730  -0.518   1.777  26.199
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.646e+01  5.103e+00   7.144 3.28e-12 ***
## CRIM          -1.080e-01  3.286e-02  -3.287 0.001087 **
## ZN             4.642e-02  1.373e-02   3.382 0.000778 ***
## INDUS          2.056e-02  6.150e-02   0.334 0.738288
## factor(CHAS)1  2.687e+00  8.616e-01   3.118 0.001925 **
## NOX           -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
## RM             3.810e+00  4.179e-01   9.116  < 2e-16 ***
## AGE            6.922e-04  1.321e-02   0.052 0.958229
## DIS           -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
## RAD            3.060e-01  6.635e-02   4.613 5.07e-06 ***
## TAX           -1.233e-02  3.760e-03  -3.280 0.001112 **
## PTRATIO       -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
## B              9.312e-03  2.686e-03   3.467 0.000573 ***
## LSTAT         -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.745 on 492 degrees of freedom
## Multiple R-squared:  0.7406, Adjusted R-squared:  0.7338
## F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

From the result above, we can remove variables INDUS and AGE as the P-values associated with these predictors are too large and indicate the coeffcients are not significant.

Part B

Below is the adjusted linear model without variables INDUS and AGE:

```
reg.picked <- lm(MEDV ~ CRIM + ZN + factor(CHAS) + NOX +
                 RM + DIS + RAD + TAX + PTRATIO + B + LSTAT, bostonhousing)
summary(reg.picked)
```

```
##
## Call:
## lm(formula = MEDV ~ CRIM + ZN + factor(CHAS) + NOX + RM + DIS +
##     RAD + TAX + PTRATIO + B + LSTAT, data = bostonhousing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.5984  -2.7386  -0.5046   1.7273  26.2373
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    36.341145   5.067492   7.171 2.73e-12 ***
## CRIM           -0.108413   0.032779  -3.307 0.001010 **
## ZN              0.045845   0.013523   3.390 0.000754 ***
## factor(CHAS)1   2.718716   0.854240   3.183 0.001551 **
## NOX           -17.376023   3.535243  -4.915 1.21e-06 ***
## RM              3.801579   0.406316   9.356  < 2e-16 ***
## DIS            -1.492711   0.185731  -8.037 6.84e-15 ***
## RAD             0.299608   0.063402   4.726 3.00e-06 ***
## TAX            -0.011778   0.003372  -3.493 0.000521 ***
## PTRATIO        -0.946525   0.129066  -7.334 9.24e-13 ***
## B               0.009291   0.002674   3.475 0.000557 ***
## LSTAT          -0.522553   0.047424 -11.019  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.736 on 494 degrees of freedom
## Multiple R-squared:  0.7406, Adjusted R-squared:  0.7348
## F-statistic: 128.2 on 11 and 494 DF,  p-value: < 2.2e-16
```

Part C

Below is the MSE and MSA values associated with the linear models:

```
n=506
p1=13
p2=11

MSE1 = sum((reg$residuals)^2)/(n-1-p1)
MAE1 = sum(abs(reg$residuals))/(n-1-p1)

MSE2 = sum((reg.picked$residuals)^2)/(n-1-p2)
MAE2 = sum(abs(reg.picked$residuals))/(n-1-p2)
```

```
# Original regression model
MSE1
```

```
## [1] 22.51785
```

```
MAE1
```

```
## [1] 3.363936
```

```
# Adjusted regression model
MSE2
```

```
## [1] 22.43191
```

```
MAE2
```

```
## [1] 3.351519
```

From the values above, ***reg.picked*** is preferred because it has slightly lower MSE and MAE.

Part D

```
step(reg)
```

```
## Start:  AIC=1589.64
## MEDV ~ CRIM + ZN + INDUS + factor(CHAS) + NOX + RM + AGE + DIS +
##     RAD + TAX + PTRATIO + B + LSTAT
##
##                Df Sum of Sq   RSS    AIC
## - AGE           1      0.06 11079 1587.7
## - INDUS         1      2.52 11081 1587.8
## <none>                      11079 1589.6
## - factor(CHAS)  1    218.97 11298 1597.5
## - TAX           1    242.26 11321 1598.6
## - CRIM          1    243.22 11322 1598.6
## - ZN            1    257.49 11336 1599.3
## - B             1    270.63 11349 1599.8
## - RAD           1    479.15 11558 1609.1
## - NOX           1    487.16 11566 1609.4
## - PTRATIO       1   1194.23 12273 1639.4
## - DIS           1   1232.41 12311 1641.0
## - RM            1   1871.32 12950 1666.6
## - LSTAT         1   2410.84 13490 1687.3
##
## Step:  AIC=1587.65
## MEDV ~ CRIM + ZN + INDUS + factor(CHAS) + NOX + RM + DIS + RAD +
##     TAX + PTRATIO + B + LSTAT
##
##                Df Sum of Sq   RSS    AIC
## - INDUS         1      2.52 11081 1585.8
## <none>                      11079 1587.7
## - factor(CHAS)  1    219.91 11299 1595.6
## - TAX           1    242.24 11321 1596.6
## - CRIM          1    243.20 11322 1596.6
## - ZN            1    260.32 11339 1597.4
## - B             1    272.26 11351 1597.9
```

```
## - RAD           1    481.09 11560 1607.2
## - NOX           1    520.87 11600 1608.9
## - PTRATIO       1   1200.23 12279 1637.7
## - DIS           1   1352.26 12431 1643.9
## - RM            1   1959.55 13038 1668.0
## - LSTAT         1   2718.88 13798 1696.7
##
## Step:  AIC=1585.76
## MEDV ~ CRIM + ZN + factor(CHAS) + NOX + RM + DIS + RAD + TAX +
##     PTRATIO + B + LSTAT
##
##                Df Sum of Sq   RSS    AIC
## <none>                      11081 1585.8
## - factor(CHAS)  1    227.21 11309 1594.0
## - CRIM          1    245.37 11327 1594.8
## - ZN            1    257.82 11339 1595.4
## - B             1    270.82 11352 1596.0
## - TAX           1    273.62 11355 1596.1
## - RAD           1    500.92 11582 1606.1
## - NOX           1    541.91 11623 1607.9
## - PTRATIO       1   1206.45 12288 1636.0
## - DIS           1   1448.94 12530 1645.9
## - RM            1   1963.66 13045 1666.3
## - LSTAT         1   2723.48 13805 1695.0
##
## Call:
## lm(formula = MEDV ~ CRIM + ZN + factor(CHAS) + NOX + RM + DIS +
##     RAD + TAX + PTRATIO + B + LSTAT, data = bostonhousing)
##
## Coefficients:
##   (Intercept)           CRIM             ZN  factor(CHAS)1            NOX
##     36.341145      -0.108413       0.045845       2.718716     -17.376023
##            RM            DIS            RAD            TAX        PTRATIO
##      3.801579      -1.492711       0.299608      -0.011778      -0.946525
##             B          LSTAT
##      0.009291      -0.522553
```

Running the stepwise regression on the Boston housing dataset shows that AIC is the lowest in the model that excludes variables AGE and INDUS. This is equavalent to ***reg.picked*** in Part B, and all of the coefficients above match to those of ***reg.picked***.