BMishra HW3

Bijesh Mishra

Due date: 2/15/2021

Machine Learning Homework 3

```
rm(list = ls()) #Clear environment.
setwd("~/Dropbox/OSU/PhD/SemVISp2021/STAT5063ML/Homeworks/hw3") #Mac.
# install.package("MASS")
library("MASS") # Load MASS package
data(Boston, package = "MASS") #Boston dataset from MASS Package.
attach(Boston) # Attach
# help(Boston) #information
```

Q1: Interpret the p-value for the overall F test for predicting crime rate with all the available predictors, and interpret the R-squared.

```
q1 = lm(crim ~ zn + indus + chas + nox + rm + age + dis +
          rad + tax + ptratio + black + lstat + medv)
summary(q1)
```

```
##
## Call:
## lm(formula = crim ~ zn + indus + chas + nox + rm + age + dis +
##     rad + tax + ptratio + black + lstat + medv)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.924 -2.120 -0.353  1.019 75.051
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.033228   7.234903   2.354 0.018949 *
## zn            0.044855   0.018734   2.394 0.017025 *
## indus        -0.063855   0.083407  -0.766 0.444294
## chas         -0.749134   1.180147  -0.635 0.525867
## nox         -10.313535   5.275536  -1.955 0.051152 .
## rm            0.430131   0.612830   0.702 0.483089
## age           0.001452   0.017925   0.081 0.935488
## dis          -0.987176   0.281817  -3.503 0.000502 ***
## rad           0.588209   0.088049   6.680 6.46e-11 ***
## tax          -0.003780   0.005156  -0.733 0.463793
## ptratio      -0.271081   0.186450  -1.454 0.146611
## black        -0.007538   0.003673  -2.052 0.040702 *
## lstat         0.126211   0.075725   1.667 0.096208 .
## medv         -0.198887   0.060516  -3.287 0.001087 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 6.439 on 492 degrees of freedom
## Multiple R-squared:  0.454,  Adjusted R-squared:  0.4396
## F-statistic: 31.47 on 13 and 492 DF,  p-value: < 2.2e-16
```

Q1 Answer:

- Ho: B_zn = B_indus = B_chas = B_nox = B_rm = B_age = B_dis = B_rad = B_tax =B_ptratio = B_black = B_lstat = B_mdev = 0; where B represent Beta-cofefficients of respective variable.
- Ha: Atleast one of the betas is different.
- Test Statistics: F(13, 492) = 31.47.
- Decision: Since p-value (< 2.2e-16) < 0.05, we reject Ho in favor of alternative hypothesis; atleast one of them is different. We need to include and test these variables in the model.
- R-squared = 0.454 means about 45.40% of variation in the crime rate is explained by the model.

Q2: Consider a model . . .

2a):

- Row design matrix X = [1 459 1 459]

2b)

```
q2 = lm(crim ~ tax*chas)
q2
```

```
##
## Call:
## lm(formula = crim ~ tax * chas)
##
## Coefficients:
## (Intercept)          tax         chas     tax:chas
##    -8.87163      0.03078      5.42284     -0.01706
```

- fhatt(x) = - 8.87163 + 0.03078*tax + 5.42284*chas - 0.01706(tax*chas)

2c) Verify using matrix multipication

```
# names(Boston) # Get names or Boston[1, ] to view variables
y = Boston[, 1]
x = as.matrix(cbind(1, Boston[,4], Boston[,10],
                    Boston[, 4]*Boston[, 10]))
matver = (solve(t(x)%*%x))%*%t(x)%*%y
matver
```

```
##              [,1]
## [1,] -8.87163081
## [2,]  5.42284390
## [3,]  0.03078064
## [4,] -0.01705803
```

The coefficients obtained from manual computation using formula is given on the table above and are same as Q2(b).

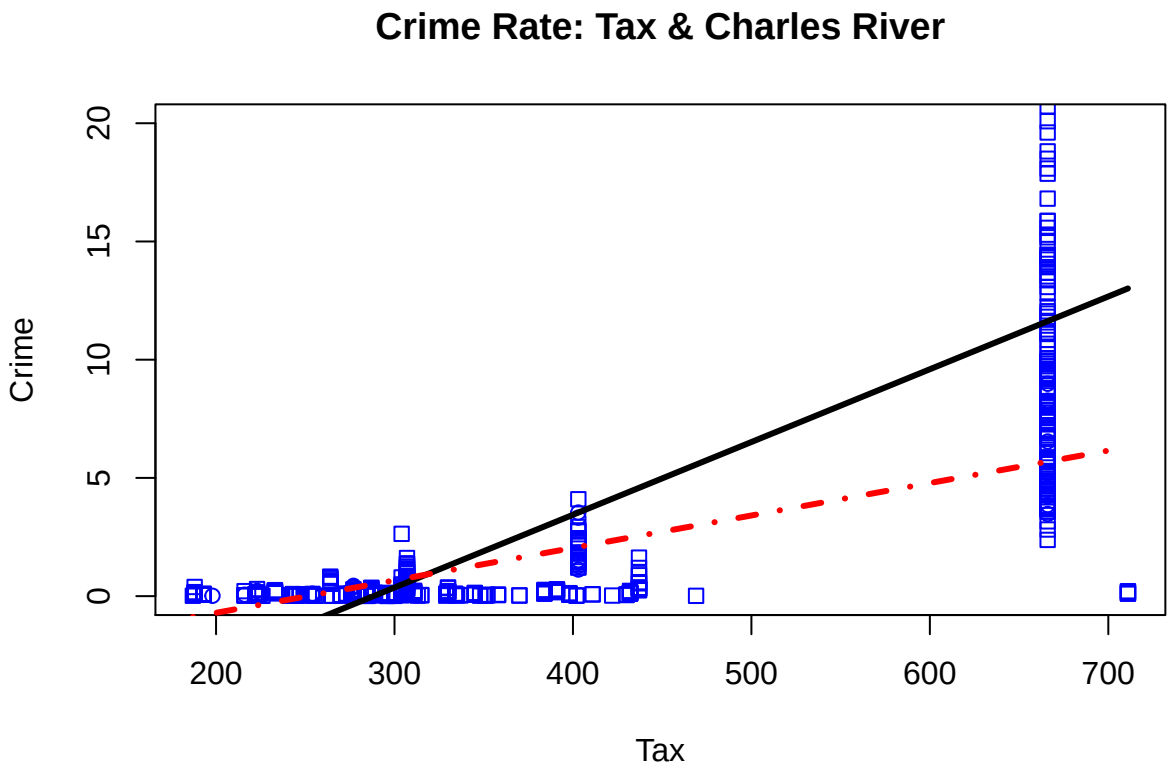
2d)

q2

```
##
## Call:
## lm(formula = crim ~ tax * chas)
##
## Coefficients:
## (Intercept)          tax         chas     tax:chas
##    -8.87163      0.03078      5.42284     -0.01706
```

- eq2: fhatt0(x) = -8.87163 + 0.03078*tax # when chas = 0
- eq1: fhatt1(x) = -8.87163 + 0.03078*tax* + *5.42284*chas - 0.01706*(tax*chas) # when chas = 1

```
# plot
plot(x = tax, y = crim, xlab = "Tax", ylab = "Crime", pch = chas,
     col = 500, ylim = c(0,20),
     main = "Crime Rate: Tax & Charles River")
# different plotting characteristics for "chas" variable.
curve(-8.87163 + 0.03078*x, add = TRUE, lwd = 3, col = 9, lty = 1) #chas = 0
curve(-8.87163 + 0.03078*x + 5.42284 - 0.01706*x,
      add = TRUE, lwd = 3, lty = 10, col = 2) #chas = 1
```



2e) Confidence interval:

```
#ci = [Bhatt_interaction + 1.96*st. er., Bhatt_interaction - 1.96*st. er]
# confint(q2)
```

```
ci = confint(q2)[4, ]
ci
```

```
##        2.5 %       97.5 %
## -0.031770196 -0.002345867
```

If we draw random samples with same number of observations from the same population repeatedly, the true B-coefficient of the interaction term between charles river and tax rate lies between the this interval 95 out of 100 times to predict per capita crime by town.

2f), & 2g):

```
# predict (q2, newdata = data.frame(tax = 666, chas = 1), interval = "confidence") #f
predict (q2, newdata = data.frame(tax = c(666, 666), chas = c(1, 0)),
         interval = "confidence") #g
```

```
##         fit      lwr      upr
## 1  5.690473  1.08623 10.29472
## 2 11.628278 10.48131 12.77525
```

- Interpretation (2f) (given by 1): If we draw random samples with same number of observations from the same population repeatedly, the true crime rate in the property paying $666 tax and bordering to the Charles River lies between 1.086 to 10.295, 95 out of 100 times.
- Interpreatation (2g) (given by 2): We can say with 95% confidence that the average crime rate falls between 1.086 to 10.295 in the property paying tax $666 and bordering to the Charles river.

2h):

```
predict (q2, newdata = data.frame(tax = 666, chas = 1),
         interval = "prediction") #h
```

```
##        fit       lwr      upr
## 1 5.690473 -8.753693 20.13464
```

- Interpreatation (h) (given by 1 on the bottom): We can say with 95% confidence that the average crime rate falls between -8.754 to 20.135 in the property paying tax $666 and bordering to the Charles river in Smithville.

2i) Use matrix multipication in R to verify points above:

```
newdata = c(1, 1, 666, 666)
newpred = newdata%*% matver
newpred1 = newdata%*% matver
newpred
```

```
##          [,1]
## [1,] 5.690473
```

```
newpred1
```

```
##          [,1]
## [1,] 5.690473
```

This result is same as above in q2(h) fit mean value. newpred and newpred1 are same; kept for better understanding of calculation.

3) Find a model that contains at least 3 predictors, statistically significant at 0.05:

```
# names(Boston)
q3 = lm(crim ~ zn + dis + rad + medv + black)
summary(q3)
```

```
##
## Call:
## lm(formula = crim ~ zn + dis + rad + medv + black)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.553  -1.869  -0.358   0.839  75.744
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.919933   1.778986   4.452 1.05e-05 ***
## zn           0.051799   0.017329   2.989 0.002935 **
## dis         -0.672189   0.202939  -3.312 0.000992 ***
## rad          0.472306   0.042102  11.218  < 2e-16 ***
## medv        -0.174219   0.036295  -4.800 2.10e-06 ***
## black       -0.008211   0.003615  -2.271 0.023562 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.473 on 500 degrees of freedom
## Multiple R-squared:  0.4393, Adjusted R-squared:  0.4337
## F-statistic: 78.34 on 5 and 500 DF,  p-value: < 2.2e-16
```

- model: fhatt(x) = 7.920 + 0.052*zn* - *0.672*dis + 0.472*rad* - *0.174*medv - 0.008*black + e
- All of the idepndent variables in the model above are statistically significant at 0.05. This can be confirmed by looking at p-value < 0.05 as indicated by at least *.

4) Compare your model above to full model using partial F-test and interpret p-value.:

```
anova(q3, q1) # Restricted model vs Full model F-test.
```

```
## Analysis of Variance Table
##
## Model 1: crim ~ zn + dis + rad + medv + black
## Model 2: crim ~ zn + indus + chas + nox + rm + age + dis + rad + tax +
##     ptratio + black + lstat + medv
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1    500 20950
## 2    492 20400  8    550.61 1.6599 0.1057
```

- Ho: B_indus = B_chas = B_nox = B_rm = B_age = B_tax = B_ptratio = B_lstat = 0; where B represent Beta-cofefficients of respective variable.
- Ha: Atleast one of the betas is different.
- Test Statistics: F(8, 492) = 1.6599 & p-value (0.1057) > 0.05.
- Decision: We fail to reject Ho because p-value > 0.05. So, this model is a complete model and no variables are missing in the model.

5) Consider a model . . .

5a)

```
taxsq = tax*tax
q5 = lm(crim ~ tax + taxsq)
summary(q5)
```

```
##
## Call:
## lm(formula = crim ~ tax + taxsq)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.810  -1.085  -0.011   0.256  76.982
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.934e+00  3.192e+00   1.859  0.06361 .
## tax         -4.318e-02  1.569e-02  -2.753  0.00612 **
## taxsq        7.850e-05  1.677e-05   4.680 3.69e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.856 on 503 degrees of freedom
## Multiple R-squared:  0.3672, Adjusted R-squared:  0.3647
## F-statistic: 145.9 on 2 and 503 DF,  p-value: < 2.2e-16
```

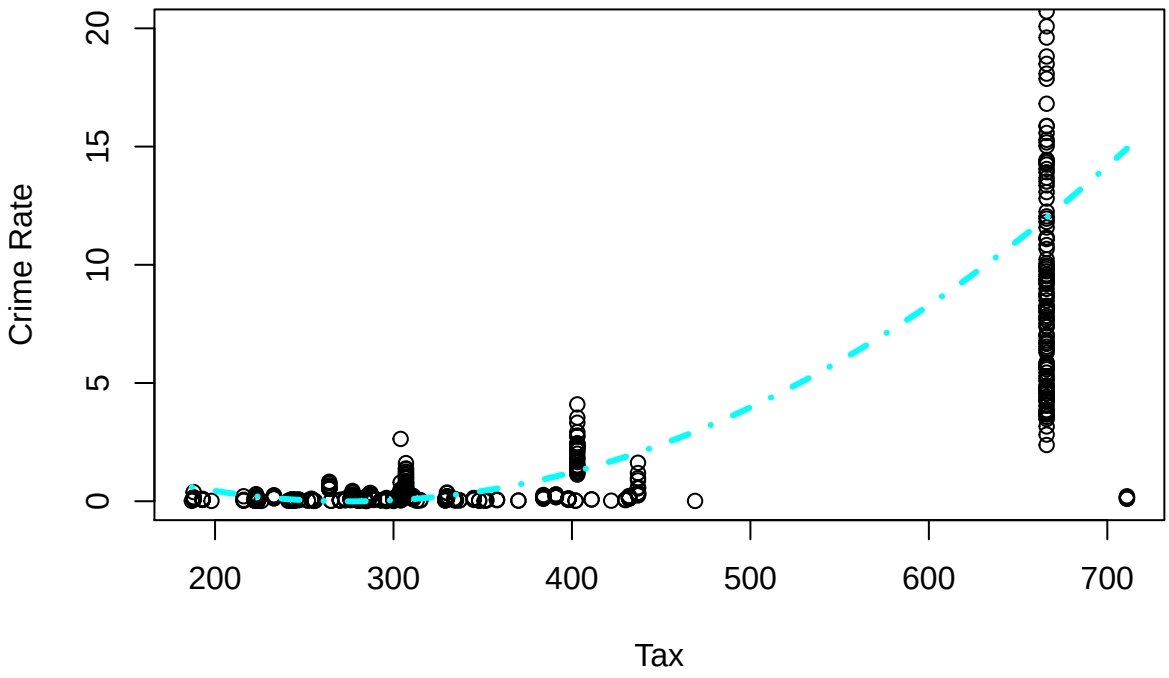
```
t.test(taxsq, alternative = "two.sided", conf.level = 0.95,
       paired = FALSE)
```

```
##
##  One Sample t-test
##
## data:  taxsq
## t = 27.831, df = 505, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  181240.1 208772.3
## sample estimates:
## mean of x
##  195006.2
```

- Ho: B_taxsq = 0; Ha: B_taxsq != 0 (!=: is not equal to)
- Test Statistics: t-test: T(505, a = 0.05/2) = 27.831, and p-value = < 0.000
- Decision: We reject null hypotheis (Ho) in favor of alternative hypothesis (Ha) and thus the quadratic term is not equal to zero. So, the crim and tax have non-linear relationship.

5b) Plot:

```
plot(x = tax, y = crim, xlab = "Tax", ylab = "Crime Rate",
     ylim = c(0,20))
curve(q5$coefficients[1] + q5$coefficients[2]*x + q5$coefficients[3]*x^2,
      add = TRUE, col = 5, lwd = 3, lty = 4)
```



5c) Get prediction interval for crime rate using tax = 666.

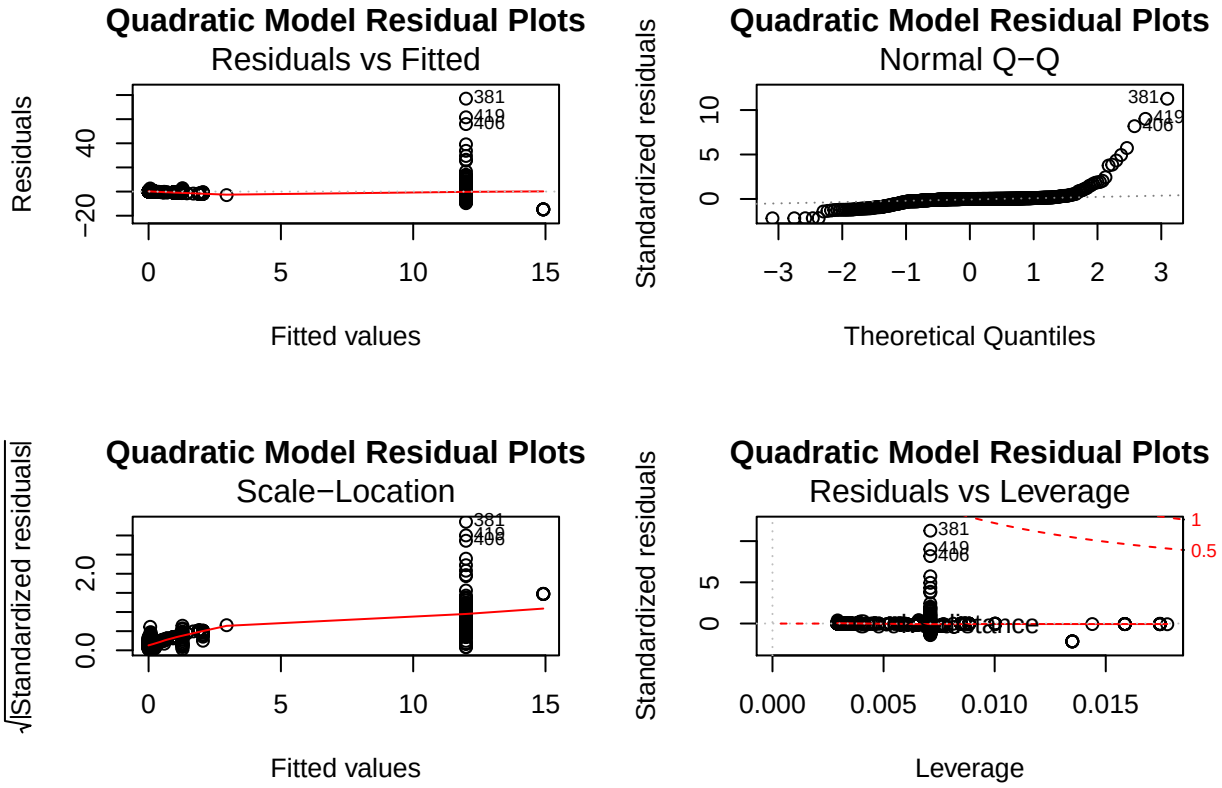
```
predict (q5, newdata = data.frame(tax = 666, taxsq = 666*666), interval = "prediction")
```

```
##        fit       lwr      upr
## 1 11.99463 -1.523389 25.51265
```

- The prediction interval of crime rate for tax = 666 for the quadratic model is between -1.523389 and 25.51265.

5d) Assume the mean 0, constant variance & normality assumption using plot.

```
par(mfrow = c(2,2))
# plot(q5,  main = "")
plot(q5,  main = "Quadratic Model Residual Plots")
```



- (Error) Mean = 0: May not be valid as the residual vs fitted plots shows that the value of residuals are more widely spread as we progress towards higher vaule of x and y. The dots in the plot is not randomly scattered indicating there is some kind of patter in the data. The residual plots might be showing heteroscadesticity.
- Constant Variance: May not be valid as the Scale-location plot does not have red line roughly horizantal. This like is increasing at the beinging and bending to become more or less horizantal showing some type of non-linear pattern. This might be a symbol of heteroscadesticity.
- The normal Q-Q plot ideally should have all points falling on straight line rising more or less diagonally from bottom left to top right. But the plot is curvy which shows residual is not normally distributed. This means either data is skewed or have more extreme values. But Residual vs Leverage plots (which should be showing extreme values beyong cook's distance) does not give much indication of extreme values in the data. So, data is more likely to be skewed.

```
# Breusch-Pagan Test:
# install.packages("lmtest")
# library(lmtest)
# bptest(q5)
# Ho: Homoscadestic; Ha: Heteroscadestic; p-value < 0.005; Decision: Reject Ho.
```

6)

Since inches and centeimers are convertible by multiplying one of them by a constant (1 inch = 2.54 cm), this creates a multicollinearity and the variance cannot be computed. We cannot compute XX‘ matrix and thus variance covariance matrix is not possible. Asymptotic VIF = 1/(1-R^2j1 - j) = infinity as R^2j1 = 1.