Zixiao Wu

Individual Assignment 4

2022-09-30

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com (http://rmarkdown.rstudio.com).

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

#4.7 Exercise, Problem 10 #(f)

```
library(ISLR)
library(MASS)
attach(Weekly)

train = (Year<2009)
Test = Weekly[!train ,]
Test_Direction= Direction[!train]

qda = qda(Direction ~ Lag2, data=Weekly[train,])
qda_pred = predict(qda,Test)
table(qda_pred$class, Test_Direction)
```

```
##       Test_Direction
##        Down Up
##   Down    0  0
##   Up     43 61
```

```
mean(qda_pred$class==Test_Direction)
```

```
## [1] 0.5865385
```

#5.4 Exercise, Problem 8 #(a)

```
library(boot)
set.seed (1)
x=rnorm(100)
y=x-2*x^2+rnorm (100)
head(x)
```

```
## [1] -0.6264538  0.1836433 -0.8356286  1.5952808  0.3295078 -0.8204684
```

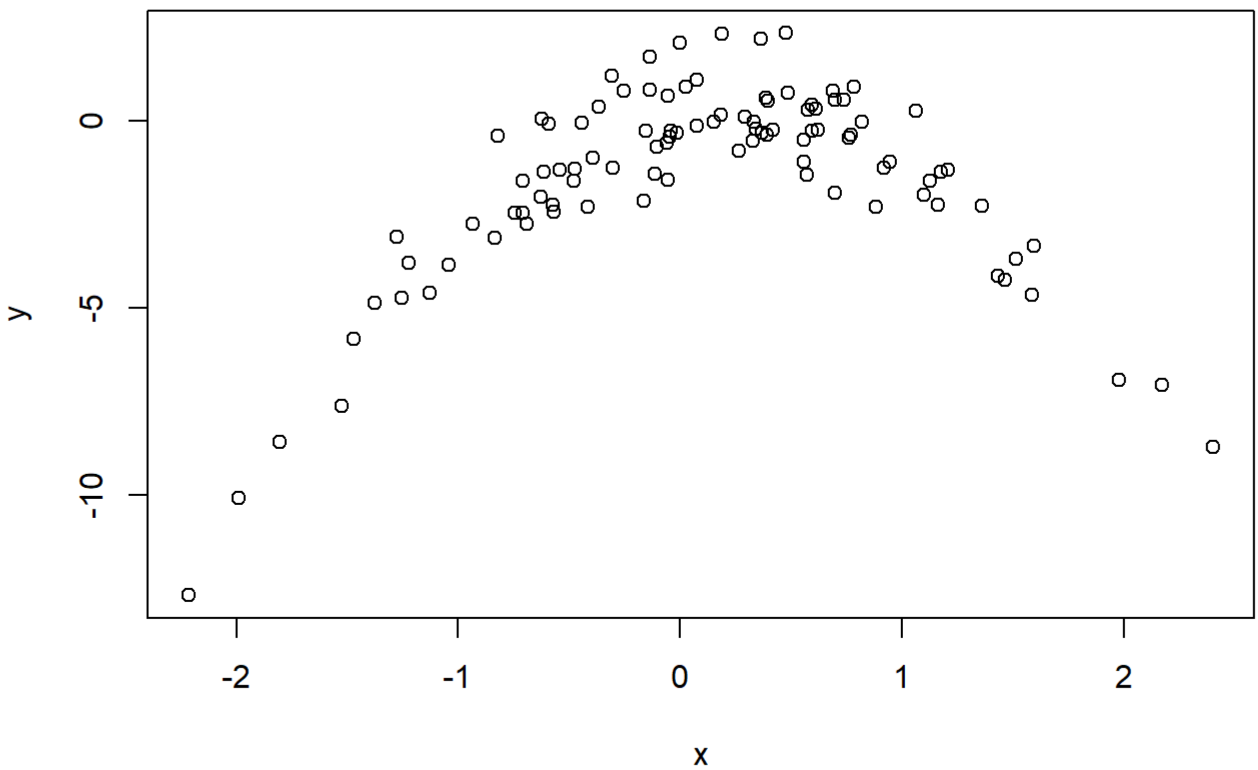
```
head(y)
```

```
## [1] -2.0317092  0.1583095 -3.1431006 -3.3365321 -0.5422276 -0.3995179
```

```
p = 100
n = 2
y = x-2*x^2+η
 η  is the error term
```

#(b)

```
plot(x,y)
```




```
We can know that the plot is non-linear, and this is a typical quadratic function plot.
```

#(c)

```
library(MASS)
set.seed(1000)
x = c(rnorm(100))
y = c(x-2*x^2+rnorm (100))
data = data.frame(x,y)

#i
lm1 = glm(y~x, data = data)
err1 = cv.glm(data, lm1)$delta[1]

#ii
lm2 = glm(y~x+I(x^2), data = data)
err2 = cv.glm(data, lm2)$delta[1]

#iii
lm3 = glm(y~x+I(x^2)+I(x^3), data = data)
err3 = cv.glm(data, lm3)$delta[1]

#iv
lm4 = glm(y~x+I(x^2)+I(x^3)+I(x^4), data = data)
err4 = cv.glm(data, lm4)$delta[1]

err1
```

```
## [1] 9.507444
```

```
err2
```

```
## [1] 0.8513634
```

```
err3
```

```
## [1] 0.8698437
```

```
err4
```

```
## [1] 0.8868636
```

```
We can see that the model ii has the smallest estimate MSE, because it is a quadratic model as the real model.
```

#(d)

```
library(boot)
set.seed(1001)
x = c(rnorm(100))
y = c(x-2*x^2+rnorm (100))
data = data.frame(x,y)

#i
lm1 = glm(y~x, data = data)
err1 = cv.glm(data, lm1)$delta[1]

#ii
lm2 = glm(y~x+I(x^2), data = data)
err2 = cv.glm(data, lm2)$delta[1]

#iii
lm3 = glm(y~x+I(x^2)+I(x^3), data = data)
err3 = cv.glm(data, lm3)$delta[1]

#iv
lm4 = glm(y~x+I(x^2)+I(x^3)+I(x^4), data = data)
err4 = cv.glm(data, lm4)$delta[1]

err1
```

```
## [1] 13.55765
```

```
err2
```

```
## [1] 0.6757288
```

```
err3
```

```
## [1] 0.7303769
```

```
err4
```

```
## [1] 0.7271295
```

```
We can see that the result is the same as above, the lowest MSE is given by the model with a qu
adratic term.
```

#(e)

```
The model with the quadratic term had the lowest LOOCV error. This is as we expected, because t
he real model is a quadratic model
```

#(f)

```
summary(lm1)
```

```
##
## Call:
## glm(formula = y ~ x, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -18.253   -1.320    1.152    2.291    4.106
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.5880     0.3524  -7.344 6.19e-11 ***
## x             0.3915     0.3095   1.265    0.209
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 12.41851)
##
##     Null deviance: 1236.9  on 99  degrees of freedom
## Residual deviance: 1217.0  on 98  degrees of freedom
## AIC: 539.69
##
## Number of Fisher Scoring iterations: 2
```

```
summary(lm2)
```

```
##
## Call:
## glm(formula = y ~ x + I(x^2), data = data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.54899  -0.62732  -0.01035   0.44389   2.21924
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.01372    0.10167  -0.135    0.893
## x            0.95496    0.07241  13.188   <2e-16 ***
## I(x^2)      -1.98477    0.04735 -41.914   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.6565082)
##
##     Null deviance: 1236.894  on 99  degrees of freedom
## Residual deviance:   63.681  on 97  degrees of freedom
## AIC: 246.66
##
## Number of Fisher Scoring iterations: 2
```

```
summary(lm3)
```

```
##
## Call:
## glm(formula = y ~ x + I(x^2) + I(x^3), data = data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.48682  -0.60365  -0.09639   0.52858   2.14225
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.04162    0.10420  -0.399    0.690
## x            1.06546    0.11848   8.993 2.17e-14 ***
## I(x^2)      -1.95390    0.05405 -36.149  < 2e-16 ***
## I(x^3)      -0.03280    0.02787  -1.177    0.242
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.6539123)
##
##     Null deviance: 1236.894  on 99  degrees of freedom
## Residual deviance:   62.776  on 96  degrees of freedom
## AIC: 247.23
##
## Number of Fisher Scoring iterations: 2
```

```
summary(lm4)
```

```
##
## Call:
## glm(formula = y ~ x + I(x^2) + I(x^3) + I(x^4), data = data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.45615  -0.56562  -0.05031   0.56969   2.11563
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.02577    0.11659   0.221   0.8255
## x            1.14958    0.13534   8.494 2.71e-13 ***
## I(x^2)      -2.07369    0.10846 -19.120  < 2e-16 ***
## I(x^3)      -0.06731    0.03882  -1.734   0.0862 .
## I(x^4)       0.02136    0.01678   1.273   0.2063
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.6497195)
##
##     Null deviance: 1236.894  on 99  degrees of freedom
## Residual deviance:   61.723  on 95  degrees of freedom
## AIC: 247.54
##
## Number of Fisher Scoring iterations: 2
```

From model ii, iii and iv we can know that both x and x^2 term is statistical significant, but other term is not significant with a large p-value. Also the model ii has the smallest AIC, so we should pick it. These results agree with the conclusions drawn based on the cross-validation results.