605_HW12.Rmd

Kumudini Bhave

April 25, 2017

BIAS VARIANCE TRADE-OFF IN R

Problem Set 1 :

Using the stats and boot libraries in R perform a cross-validation experiment to observe the bias variance tradeoff. You'll use the auto data set from previous assignments. This dataset has 392 observations across 5 variables. We want to fit a polynomial model of various degrees using the glm function in R and then measure the cross validation error using cv.glm function. Fit various polynomial models to compute mpg as a function of the other four variables acceleration, weight, horsepower, and displacement using glm function.

For example: glm.fit=glm(mpg~poly(disp+hp+wt+acc,2), data=auto)

cv.err5[2]=cv.glm(auto,glm.fit,K=5)$delta[1]

will fit a 2nd degree polynomial function between mpg and the remaining 4 variables and perform 5 iterations of cross-validations. This result will be stored in a cv.err5 array. cv.glm returns the estimated cross validation error and its adjusted value in a variable called delta. Please see the help on cv.glm to see more information. Once you have fit the various polynomials from degree 1 to 8, you can plot the cross-validation error function as degree=1:8

plot(degree,cv.err5,type='b')

Solution :

Extracting Raw Data From GitHub Data File And Reading In CSV Format

```
knitr::opts_chunk$set(message = FALSE, echo = TRUE)

# Loading RCurl package to help scrape data from web (stored on GitHub).
library(RCurl)
```

```
## Loading required package: bitops
```

```
# Loading plyr package to help map abbreviated values to explained.
library(plyr)

# library for plotting
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.3
```

```
# library for regression and cross validation
library("stats")
library("boot")
```

```
## Warning: package 'boot' was built under R version 3.3.3
```

```
data.giturl <- "https://raw.githubusercontent.com/DataDriven-MSDA/DATA605/master/auto-mpg.data"
autompg.data <- read.table(data.giturl)
head(autompg.data)
```

```
##   V1 V2  V3    V4   V5   V6 V7 V8                        V9
## 1 18  8 307 130.0 3504 12.0 70  1 chevrolet chevelle malibu
## 2 15  8 350 165.0 3693 11.5 70  1         buick skylark 320
## 3 18  8 318 150.0 3436 11.0 70  1        plymouth satellite
## 4 16  8 304 150.0 3433 12.0 70  1             amc rebel sst
## 5 17  8 302 140.0 3449 10.5 70  1               ford torino
## 6 15  8 429 198.0 4341 10.0 70  1          ford galaxie 500
```

```
# Attempt extract the required data columns of displacement, horsepower, weight, acceleration, mpg from
autompg.datastudy <- subset(autompg.data, select = c(V3, V4, V5, V6, V1), V4 != "?")
autompg.datastudy <- na.omit(autompg.datastudy)


# Verifying the number of attributes
length(autompg.datastudy)
```

```
## [1] 5
```

```
# Verifying the number of observations selected
nrow(autompg.datastudy)
```

```
## [1] 392
```

```
# Renaming the attributes
colnames(autompg.datastudy) <- c("disp", "hp", "wt", "acc", "mpg")


# Viewing the data
head(autompg.datastudy)
```

```
##   disp    hp   wt  acc mpg
## 1  307 130.0 3504 12.0  18
## 2  350 165.0 3693 11.5  15
## 3  318 150.0 3436 11.0  18
## 4  304 150.0 3433 12.0  16
## 5  302 140.0 3449 10.5  17
## 6  429 198.0 4341 10.0  15
```

```
dim(autompg.datastudy)
```

```
## [1] 392   5
```

```
str(autompg.datastudy)
```

```
## 'data.frame':    392 obs. of  5 variables:
##  $ disp: num  307 350 318 304 302 429 454 440 455 390 ...
##  $ hp  : Factor w/ 94 levels "?","100.0","102.0",..: 17 35 29 29 24 42 47 46 48 40 ...
##  $ wt  : num  3504 3693 3436 3433 3449 ...
##  $ acc : num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
##  $ mpg : num  18 15 18 16 17 15 14 14 14 15 ...
```

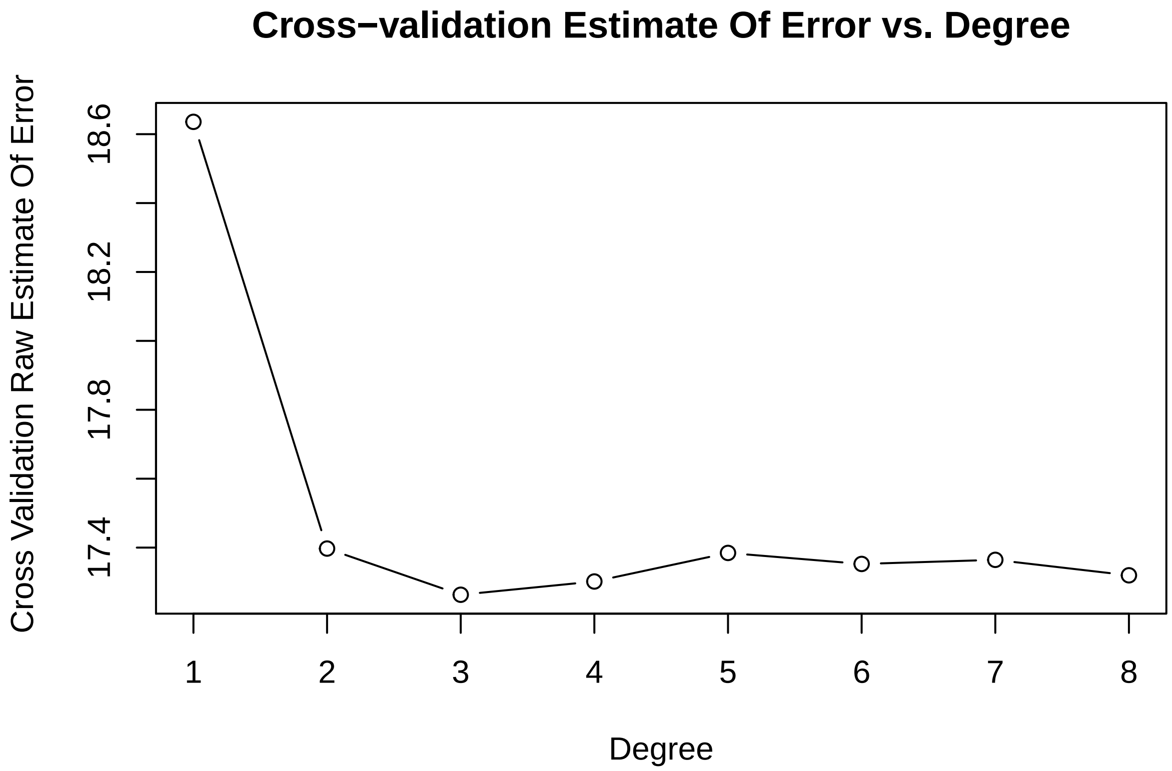
```
# Converting the horsepower as numeric as required , since it is quantitative variable
autompg.datastudy$hp <- as.numeric(autompg.datastudy$hp)
str(autompg.datastudy)
```

```
## 'data.frame':    392 obs. of  5 variables:
##  $ disp: num  307 350 318 304 302 429 454 440 455 390 ...
##  $ hp  : num  17 35 29 29 24 42 47 46 48 40 ...
##  $ wt  : num  3504 3693 3436 3433 3449 ...
##  $ acc : num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
##  $ mpg : num  18 15 18 16 17 15 14 14 14 15 ...
```

Auto Data Bias Variance Study

To understand the trade off between bias and variance, we try different fits between mpg and the remaining four variables (displacement, horsepower, weight, and acceleration) With model complexity increasing bias(overfitting) reduces upto a certain degree and variance (underfitting) increases. We need to find the optimal model where bias is minimal before it starts increasing

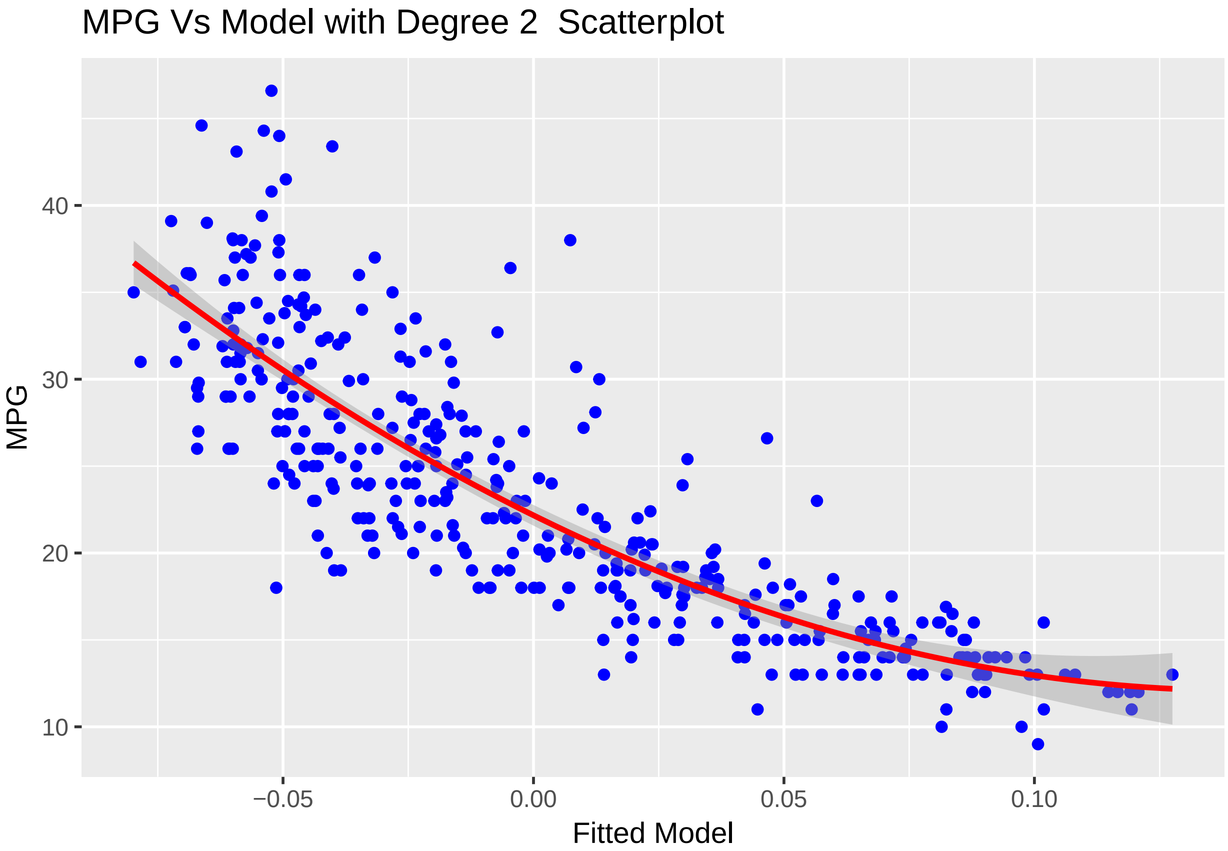
```
knitr::opts_chunk$set(message = FALSE, echo = TRUE)





# Function to randomly select samples from a given population dataset @param1 dataframe @param2 number
glm_crossv <- function(df, deg) {
    # df= autompg.datastudy deg <- 8
    cv.err5 <- c()

    cv.err5.adjcv <- c()
    # iterate for various degrees
    for (i in 1:deg) {

        degree <- i

        glm.fit = glm(mpg ~ poly(disp + hp + wt + acc, i), data = df)


        cv.err5[i] <- cv.glm(df, glm.fit, K = 5)$delta[1]  # raw cross validation estimate

        cv.err5.adjcv[i] <- cv.glm(df, glm.fit, K = 5)$delta[2]  # adjusted cross validation estimate



    }
    glmfitcv <- as.data.frame(cbind(seq(c(1:deg)), cv.err5, cv.err5.adjcv))
    colnames(glmfitcv) <- c("degree", "cvRaw", "cvAdj")
    return(glmfitcv)

}
glmcv <- glm_crossv(autompg.datastudy, 8)


# Mapping the crossvalidation raw estimate

plot(glmcv$degree, glmcv$cvRaw, type = "b", main = "Cross-validation Estimate Of Error vs. Degree", xlal
```



From the plot we observe that the mean cross-validation error is lowest at degree 2, and we get the characteristic U-shaped curve.

Plotting the model for the lowest error

```
knitr::opts_chunk$set(message = FALSE, echo = TRUE)

glm.fit.2 = glm(mpg ~ poly(disp + hp + wt + acc, 2), data = autompg.datastudy)


plotbiasvariancetradeoffmodel <- ggplot(data = autompg.datastudy, aes(y = mpg, x = poly(disp + hp + wt
    scale_y_continuous(name = "MPG") + stat_smooth(method = "glm", formula = y ~ poly(x, 2), size = 1,

plotbiasvariancetradeoffmodel
```


MPG Vs Model with Degree 2 Scatterplot
MPG
40
30
20
10
−0.05
0.00
0.05
0.10
Fitted Model