ADA Homework 6 Wenxin Liang wl2455

1. Consider the data set birthwt in R library MASS. Compare models selected using LASSO and a stepwise procedure to predict 'bwt' birth weight in grams using the following set of predictors:

'age' mother's age in years

'lwt' mother's weight in pounds at last menstrual period

'race' mother's race ('1' = white, '0' = other)

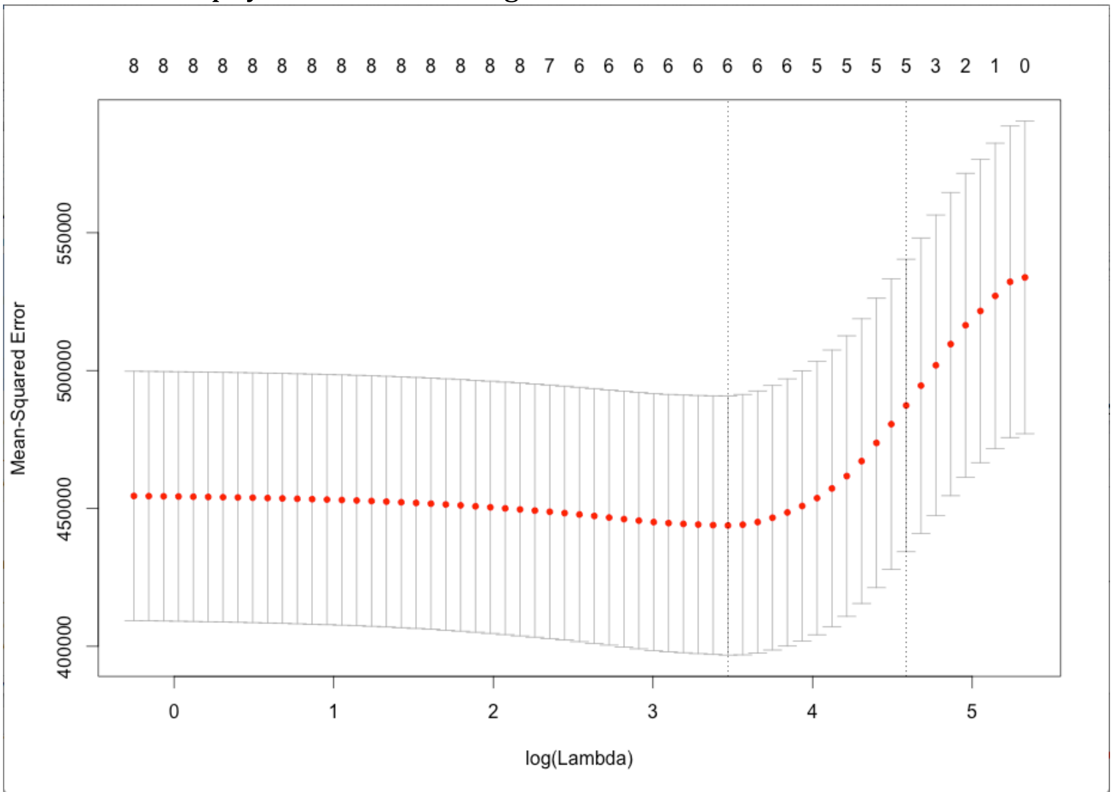
'smoke' smoking status during pregnancy

'ptl' number of previous premature labours

'ht' history of hypertension

'ui' presence of uterine irritability

'ftv' number of physician visits during the first trimester



Based on the Lasso procedure, first we get the graph above we know that the best model Lasso suggested for our problem is 6 predict variables then from the procedure we obtain the model to be selected is,

$$bwt = 2607.1680 + 3.0086 \times lwt + 328.9902 \times race - 302.2046 \times smoke - 26.8500 \times ptl - 457.5420 \times ht - 456.3827 \times ui$$

Then we know that the predict variables "age" and "ftv" are force to be zero.

Based on R,

```
> library(glmnet)
> X <- model.matrix(bwt~.,data=birthwt[,-1])
> y <- birthwt$bwt
> fit <- glmnet(X,y)
> cvfit <- cv.glmnet(X,y)
> plot(cvfit)
```

```
> cv_out <- cv.glmnet(x,y,alpha=1)
> bestlammin <- cv_out$lambda.min
> result <- glmnet(X,y,alpha=1)
> lasso.coef <- predict(result,type="coefficients",s=bestlammin)
> lasso.coef
10 x 1 sparse Matrix of class "dgCMatrix"
                       1
(Intercept) 2607.167973
(Intercept)        .
age                  .
lwt                3.008559
race             328.990196
smoke           -302.204612
ptl              -26.849967
ht              -457.542001
ui              -456.382720
```

Based on the stepwise procedure, we obtain the model to be selected is

$$bwt = 2504.3049 + 3.8658 \times lwt + 389.6949 \times race - 370.2894 \times smoke - 584.4266 \times ht - 552.5400 \times ui$$

Then we conclude that five predict variables have significance on the response.

```
> fit <- lm(bwt~.,data = birthwt[,-1])
> step <- stepAIC(fit,direction = "both")
Start:   AIC=2456.95
bwt ~ age + lwt + race + smoke + ptl + ht + ui + ftv

          Df Sum of Sq      RSS    AIC
- age      1     36979 76035306 2455.0
- ftv      1     45750 76044077 2455.1
- ptl      1     91874 76090201 2455.2
<none>                 75998327 2456.9
- lwt      1   2373581 78371908 2460.8
- ht       1   3619607 79617933 2463.7
- smoke    1   5131191 81129518 2467.3
- ui       1   5772022 81770349 2468.8
- race     1   6282587 82280914 2470.0

Step:   AIC=2455.04
bwt ~ lwt + race + smoke + ptl + ht + ui + ftv

          Df Sum of Sq      RSS    AIC
- ftv      1     63545 76098851 2453.2
- ptl      1    110556 76145862 2453.3
```

```
<none>                    76035306 2455.0
+ age     1      36979 75998327 2456.9
- lwt     1    2338372 78373678 2458.8
- ht      1    3599309 79634615 2461.8
- smoke   1    5099798 81135104 2465.3
- ui      1    5736814 81772120 2466.8
- race    1    6353942 82389248 2468.2

Step:  AIC=2453.2
bwt ~ lwt + race + smoke + ptl + ht + ui

          Df Sum of Sq      RSS    AIC
- ptl     1     109225 76208075 2451.5
<none>                    76098851 2453.2
+ ftv     1      63545 76035306 2455.0
+ age     1      54774 76044077 2455.1
- lwt     1    2275785 78374636 2456.8
- ht      1    3538442 79637293 2459.8
- smoke   1    5062640 81161490 2463.4
- ui      1    5697773 81796624 2464.8
- race    1    6292956 82391807 2466.2

Step:  AIC=2451.47
bwt ~ lwt + race + smoke + ht + ui

          Df Sum of Sq      RSS    AIC
<none>                    76208075 2451.5
+ ptl     1     109225 76098851 2453.2
+ age     1      76147 76131928 2453.3
+ ftv     1      62214 76145862 2453.3
- lwt     1    2408206 78616282 2455.3
- ht      1    3575534 79783609 2458.1
- smoke   1    5501070 81709146 2462.6
- ui      1    6286035 82494110 2464.4
- race    1    6359552 82567628 2464.6
> step$anova #show the result we obtain
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
bwt ~ age + lwt + race + smoke + ptl + ht + ui + ftv

Final Model:
bwt ~ lwt + race + smoke + ht + ui
```

```
   Step Df    Deviance Resid. Df Resid. Dev          AIC
1                                  180      75998327 2456.946
2 - age    1    36979.00           181      76035306 2455.038
3 - ftv    1    63544.76           182      76098851 2453.196
4 - ptl    1 109224.64             183      76208075 2451.467
> step$coefficients
(Intercept)          lwt          race        smoke           ht           ui
2504.304943     3.865837    389.694924  -370.289368  -584.426550
-522.540014
```

Based on the R code of the Lasso procedure and the Stepwise procedure, we conclude that we obtain different final models selected from different procedure. Based Lasso procedure the final model includes six variables can have influence on our dependent variable "bwt", birth weight in grams. The six variables are 'lwt' mother's weight in pounds at last menstrual period , 'race' mother's race ('1' = white, '0' = other), 'smoke' smoking status during pregnancy, 'ptl' number of previous premature labours , 'ht' history of hypertension and 'ui' presence of uterine irritability. The coefficient of each of the predictor variables showed in the final model, $bwt = 2607.1680 + 3.0086\times lwt + 328.9902\times race - 302.2046\times smoke - 26.8500\times ptl - 457.5420\times ht - 456.3827\times ui$

Based on stepwise procedure the final model includes five variables can have influence on our dependent variable "bwt", birth weight in grams. The five variables are 'lwt' mother's weight in pounds at last menstrual period , 'race' mother's race ('1' = white, '0' = other), 'smoke' smoking status during pregnancy, 'ht' history of hypertension and 'ui' presence of uterine irritability. The coefficient of each of the predictor variables showed in the final model, $bwt = 2504.3049 + 3.8658\times lwt + 389.6949\times race - 370.2894\times smoke - 584.4266\times ht - 552.5400\times ui$

2. For the data set 'stackloss' in R, consider the multiple linear regression model of "stack loss" on the other explanatory variables

i) Investigate whether there is any multicollinearity, and suggest remedial measures if appropriate.

Based on R, we obtain,

```
> vif(mylm)
   Air.Flow Water.Temp Acid.Conc.
   2.906484    2.572632    1.333587
```

Since all the vif <10 then we conclude that there is no problem with collinearity within our linear model. Also we know that the $\overline{VIF}$ is not much larger than 1 we can verify that the there is no problem with collinearity within our linear model.

If we have the multicollineartiy problem, the remedial measure can be first we use Principal Component Analysis for the X matrix, then do regression on the eigen vetors. Second, we can do the ridge regression to force the both to zero. At last we drop a predictor variable from the model.

ii) Suppose the value of stack.loss[20] was changed from 14 to 1500, and those of Water.Temp[13] from 18 to 170, and Acid.Conc. [13] from 82 to 10.

a) Fit a multiple linear regression model on the new data.

Based on R we have the multiple linear regression model on the new data is

$$stack.loss = 1084.671 + 1.881 \times Air.Flow - 6.281 \times Water.Temp - 11.250 \times Acid.Conc.$$

```
> stackloss_new <- stackloss
> stackloss_new$stack.loss[20] <- 1500
> stackloss_new$Water.Temp[13] <- 170
> stackloss_new$Acid.Conc.[13] <- 10
> mylm1 <- lm(stack.loss~.,data=stackloss_new)
> mylm1 <- lm(stack.loss~Air.Flow+Water.Temp+Acid.Conc.,data=stackloss_new)
> summary(mylm1)

Call:
lm(formula = stack.loss ~ Air.Flow + Water.Temp + Acid.Conc.,
    data = stackloss_new)

Residuals:
    Min      1Q  Median      3Q     Max
-241.39  -84.00  -55.56  -19.23 1358.11

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1084.671   1304.149   0.832    0.417
Air.Flow       1.881     10.965   0.172    0.866
Water.Temp    -6.281      8.769  -0.716    0.484
Acid.Conc.   -11.250     16.696  -0.674    0.509

Residual standard error: 344.7 on 17 degrees of freedom
Multiple R-squared:   0.036, Adjusted R-squared:  -0.1341
F-statistic: 0.2116 on 3 and 17 DF,  p-value: 0.887
```

b) Identify influential points using DFFITS, DFBETAS, Studentized Deleted Residuals and Cook's D

Compare to p=3 since there are three variables then we conclude that 21 observations is a big sample.

```
> nrow(stackloss_new)
[1] 21
```

Based on R, we know there are 21 observations in our database then we can conclude we have a small database.

In general we use the R code “influence.measure” to observe the influential points.

```
> IF <- influence.measures(mylm1)
> DFFITS1 <- IF$is.inf[,5]
> DFBETAS1 <- IF$is.inf[,1:4]
> HAT1 <- IF$is.inf[,8]
> COOK1 <- IF$is.inf[,7]
> which(DFFITS1 == TRUE)
13 20
13 20
> which(DFBETAS1 == TRUE)
[1] 20 41 62 83
> which(HAT1 == TRUE)
13
13
> which(COOK1 == TRUE)
13
13
```

Then we use the basic method to verify,

```
> # Identify influential observations using DFFITS
> DFFITS <- dffits(mylm1)
> which(abs(DFFITS) > 2*sqrt(4/21))
13 20
13 20
>
> # Index plot of DFFITS
> n <- nrow(data)
> plot(DFFITS)
> text(1:n,dffits(mylm),lab=1:n)
>
> # Identify influential observations using Cook's Distances
> D <- cooks.distance(mylm1)
> which(D >= qf(.5, 4, 21-4)) # none clearly identified as influential (though
13
13
>
>
> # Index plot of Cook's Distances
```

```
> plot(D, ylab = "Cook's Distance")
>
> # Identify influential observations using DFBETAS
> DFBETAS <- dfbetas(mylm1)
> max.DFBETAS <- apply(abs(DFBETAS), 1, max)
> which(max.DFBETAS > 2/sqrt(21))
13 17 20
13 17 20
> # Index plot of studentized residuals vs observation number
> plot(rstandard(mylm1), ylab = "studentized residuals", xlab = "observation")
> which(abs(rstudent(mylm1)) >= qf(1 - .05/(2 * nrow(stackloss)), df1 = 4, df2 =
17))
20
20
```

Based on R, we obtain the influential points using the DEFITS are the 13th and 20th observations, using the DFBETAS are the 13th observation, 17th observation and 20th observation, using the Cook’s Distance is the 20th observation, using the Studentized Deleted Residuals is the 20th observations. Therefore, the influential points are 13th, 17th and 20th observations.

c) Compare the estimates of the regression coefficients obtained before and after the above changes for each of the following:

• OLS

```
> mylm1_Bef <- lm(stack.loss~Air.Flow+Water.Temp+Acid.Conc.,data=stackloss)
> mylm1_Bef$coefficients
(Intercept)       Air.Flow    Water.Temp    Acid.Conc.
-39.9196744      0.7156402      1.2952861    -0.1521225
> mylm1_After <-
lm(stack.loss~Air.Flow+Water.Temp+Acid.Conc.,data=stackloss_new)
> mylm1_After$coefficients
(Intercept)       Air.Flow    Water.Temp    Acid.Conc.
1084.671247       1.880927      -6.280934    -11.249906
```

Compare the model from the data before and the model from the data after we conclude that there is a big changing for the intercept. For the coefficient of Air.Flow variable there is small increasing. For the coefficient of the Water.Temp variable there is a decreasing. For the coefficient of the Acid.Conc. variable there is a decreasing happened.

• Least median of squares regression

```
> set.seed(1)
> mylm2_Bef <-
lmsreg(stack.loss~Air.Flow+Water.Temp+Acid.Conc.,data=stackloss)
```

```
> mylm2_Bef$coefficients
  (Intercept)      Air.Flow    Water.Temp    Acid.Conc. 
-3.425000e+01  7.142857e-01  3.571429e-01 -3.489094e-17 
> # mylm2_After1 <- lmsreg(stack.loss~Air.Flow+Water.Temp+Acid.Conc.,data=stackloss_new)
> mylm2_After2 <- lmsreg(stack.loss~.,data=stackloss_new)
> # mylm2_After1$coefficients
> mylm2_After2$coefficients
  (Intercept)      Air.Flow    Water.Temp    Acid.Conc. 
-3.425000e+01  7.142857e-01  3.571429e-01 -1.046728e-16 
```

Compare the model from the data before and the model from the data after we conclude that there is a no change for the intercept. For the coefficient of Air.Flow variable there is no change. For the coefficient of the Water.Temp variable there is no change. For the coefficient of the Acid.Conc. variable there is a small change since the number changing is from -3.489094e-17 to -1.046728e-16 both of the numbers are really close to zero so we can conclude that the coefficient of the Acid.Conc. variable has nearly no change.

- Least trimmed squares robust regression

```
> mylm3_Bef <- ltsreg(stack.loss~Air.Flow+Water.Temp+Acid.Conc.,data=stackloss)
> mylm3_Bef$coefficients
  (Intercept)      Air.Flow    Water.Temp    Acid.Conc. 
-3.429167e+01  7.142857e-01  3.571429e-01 -6.978189e-17 
> mylm3_After <- ltsreg(stack.loss~Air.Flow+Water.Temp+Acid.Conc.,stackloss_new)
> mylm3_After$coefficients
  (Intercept)      Air.Flow    Water.Temp    Acid.Conc. 
-3.580556e+01  7.500000e-01  3.333333e-01  2.355139e-16 
```

Compare the model from the data before and the model from the data after we conclude that there is a small decreasing for the intercept. For the coefficient of Air.Flow variable there is a small increasing. For the coefficient of the Water.Temp variable there is a small decreasing. For the coefficient of the Acid.Conc. variable there is a small change since the number changing is from -6.978189e-17 to 2.355139e-16 since both of the numbers are really close to zero so we can conclude that the coefficient of the Acid.Conc. variable has nearly no change.

- M-estimates of regression with Huber weights

```
> mylm4_Bef <-
```

```
rlm(stack.loss~Air.Flow+Water.Temp+Acid.Conc.,data=stackloss,scale.est="Hub
er")
> mylm4_Bef$coefficients
(Intercept)       Air.Flow    Water.Temp    Acid.Conc.
-41.1410914      0.8167062      0.9839117    -0.1314404
>                                          mylm4_After                                  <-
rlm(stack.loss~Air.Flow+Water.Temp+Acid.Conc.,stackloss_new,scale.est="Hube
r")
> mylm4_After$coefficients
  (Intercept)        Air.Flow      Water.Temp      Acid.Conc.
-36.85207014      1.11099732    -0.08852061    -0.11914091
```

Compare the model from the data before and the model from the data after we conclude that there is a small increasing for the intercept. For the coefficient of Air.Flow variable there is a small increasing. For the coefficient of the Water.Temp variable there is a decreasing. For the coefficient of the Acid.Conc. variable there is a small increasing.

Comparing the four methods for estimating of the regression coefficients obtained before and after the above changes, the OLS models have the big change for all of the coefficients, then the M-estimates of regression with Huber weights have some kind of changes for all the coefficients then the Least median of squares regression and Least trimmed squares robust regression have the least change for all the coefficients.

The R code for all the homework,

```
#Question 1
library(MASS)
birthwt$race[birthwt$race!=1]<- 0
#data <- birthwt[,2:9]

#Lasso
library(glmnet)

X <- model.matrix(bwt~.,data=birthwt[,-1])
y <- birthwt$bwt

fit <- glmnet(X,y)
cvfit <- cv.glmnet(X,y)
plot(cvfit)

cv_out <- cv.glmnet(x,y,alpha=1)
bestlammin <- cv_out$lambda.min
```

```
result <- glmnet(X,y,alpha=1)
lasso.coef <- predict(result,type="coefficients",s=bestlammin)
lasso.coef

# Stepwise
fit <- lm(bwt~.,data = birthwt[,-1])
step <- stepAIC(fit,direction = "both")
step$anova #show the result we obtain
step$coefficients

#Question 2
mylm <- lm(stack.loss~.,data=stackloss)

# Subquestion 1
# Variance inflation factor
library(car) #needed for access to vif function
vif(mylm)

# Subquestion 2
# Part a
stackloss_new <- stackloss
stackloss_new$stack.loss[20] <- 1500
stackloss_new$Water.Temp[13] <- 170
stackloss_new$Acid.Conc.[13] <- 10
mylm1 <- lm(stack.loss~.,data=stackloss_new)
mylm1 <- lm(stack.loss~Air.Flow+Water.Temp+Acid.Conc.,data=stackloss_new)
summary(mylm1)

# Part b
nrow(stackloss_new)

IF <- influence.measures(mylm1)
DFFITS1 <- IF$is.inf[,5]
DFBETAS1 <- IF$is.inf[,1:4]
HAT1 <- IF$is.inf[,8]
COOK1 <- IF$is.inf[,7]
which(DFFITS1 == TRUE)
which(DFBETAS1 == TRUE)
which(HAT1 == TRUE)
which(COOK1 == TRUE)

# Identify influential observations

# Identify influential observations using DFFITS
```

```
DFFITS <- dffits(mylm1)
which(abs(DFFITS) > 2*sqrt(4/21))

# Index plot of DFFITS
n <- nrow(data)
plot(DFFITS)

# Identify influential observations using Cook's Distances
D <- cooks.distance(mylm1)
which(D >= qf(.5, 4, 21-4)) # none clearly identified as influential (though

# Index plot of Cook's Distances
plot(D, ylab = "Cook's Distance")

# Identify influential observations using DFBETAS
DFBETAS <- dfbetas(mylm1)
max.DFBETAS <- apply(abs(DFBETAS), 1, max)
which(max.DFBETAS > 2/sqrt(21))

# Index plot of studentized residuals vs observation number
plot(rstandard(mylm1), ylab = "studentized residuals", xlab = "observation")
# No unusually large studentized residuals

# Determine whether any deleted studentized residuals exceed
# what is expected (at a .95 confidence level) for an F distribution with p
# numerator degrees of freedom and n - p denominator degrees of freedom.
which(abs(rstudent(mylm1)) >= qf(1 - .05/(2 * nrow(stackloss)), df1 = 4, df2 =
17))
# named integer(0) means that non exceeded the treshhold

# Part c
# OLS
mylm1_Bef <- lm(stack.loss~Air.Flow+Water.Temp+Acid.Conc.,data=stackloss)
mylm1_Bef$coefficients
mylm1_After <-
lm(stack.loss~Air.Flow+Water.Temp+Acid.Conc.,data=stackloss_new)
mylm1_After$coefficients

# Least Median of Square Error
set.seed(1)
mylm2_Bef <-
lmsreg(stack.loss~Air.Flow+Water.Temp+Acid.Conc.,data=stackloss)
mylm2_Bef$coefficients
# mylm2_After1 <-
```

```
lmsreg(stack.loss~Air.Flow+Water.Temp+Acid.Conc.,data=stackloss_new)
mylm2_After2 <- lmsreg(stack.loss~.,data=stackloss_new)
# mylm2_After1$coefficients
mylm2_After2$coefficients

# Least trimmed squares robust regression
set.seed(1)
mylm3_Bef <-
ltsreg(stack.loss~Air.Flow+Water.Temp+Acid.Conc.,data=stackloss)
mylm3_Bef$coefficients
mylm3_After <-
ltsreg(stack.loss~Air.Flow+Water.Temp+Acid.Conc.,stackloss_new)
mylm3_After$coefficients

#M-estimates of regression with Huber weights
set.seed(1)
mylm4_Bef <-
rlm(stack.loss~Air.Flow+Water.Temp+Acid.Conc.,data=stackloss,scale.est="Hub
er")
mylm4_Bef$coefficients
mylm4_After <-
rlm(stack.loss~Air.Flow+Water.Temp+Acid.Conc.,stackloss_new,scale.est="Hube
r")
mylm4_After$coefficients
```