**STAT 462 – Applied Regression Analysis**

**Fall 2017, Lab 13**

Prepare a short report with relevant output, your comments, and answers to the questions (this does not need to be exhaustive or polished, but should contain enough to show that you completed all tasks and analyses).

Submit the report at the end of the lab session.

Consider again the dataset *bears.txt* used in previous labs.

This contains several variables measured on n=141 “bear capturing” occasions, with the following variables:

*ID:* Identification number

*Age:* Bear's age, in months

*Month:* Month when the measurement was made.

Sex. 1 = male 2 = female

*Head.L:* Length of the head, in inches

*Head.W:* Width of the head, in inches

*Neck.G:* Girth (distance around) the neck, in inches

*Length:* Body length, in inches

*Chest.G:* Girth (distance around) the chest, in inches

*Weight:* Weight of the bear, in pounds

*Obs.No:* Observation number for this bear. For example, the bear with ID=41 (Bertha) was measured on four occasions. The value of Obs.No goes from 1 to 4 for these observations

*Name:* The names of the bears given to them by the researchers.

As you did in previous labs, consider only the first observation for each bear (bears\_indep=bears[bears$Obs.No==1,]).

Consider a multiple linear regression model with response the square root of “Weight”. Consider as possible predictors x1=“Head.L”, x2=“Head.W”, x3=“Neck.G”, x4=“Length”, x5=“Chest.G” and d a dummy variable indicating if the bear is a female (note that “Sex” is codified as 1 = male 2 = female, so you cannot use it as it is, but you need to create the dummy variable 0 = male 1 = female). Do not consider interactions.

* Use forward selection with threshold alpha-to-enter 10% in order to select a simple model to predict the bear weight. Report which predictor you add at each step, and the final model that you obtain. Is this different from the model you obtained in Lab12 with backward selection? Do you think it is a golm(od model?

Step 1:

> pvalue\_for1

Head.L Head.W Neck.G Length Chest.G Sex

6.899848e-31 2.214437e-21 1.250115e-53 3.996562e-41 2.968438e-66 5.372814e-03

**We add “Chest.G” to the model since it has the smallest p-value which is smaller than alpha-to-enter = 0.1.**

Step 2:

> pvalue\_for2

Head.L Head.W Neck.G Length Sex

6.652309e-04 3.975049e-03 1.773151e-08 3.368083e-08 1.068002e-01

**We add “Neck.G” to the model since it has the smallest p-value which is smaller than alpha-to-enter = 0.1.**

Step 3:

> pvalue\_for3

Head.L Head.W Length Sex

9.255778e-02 4.392534e-01 5.222382e-07 9.943058e-01

**We add “Length” to the model since it has the smallest p-value which is smaller than alpha-to-enter = 0.1.**

Step 4:

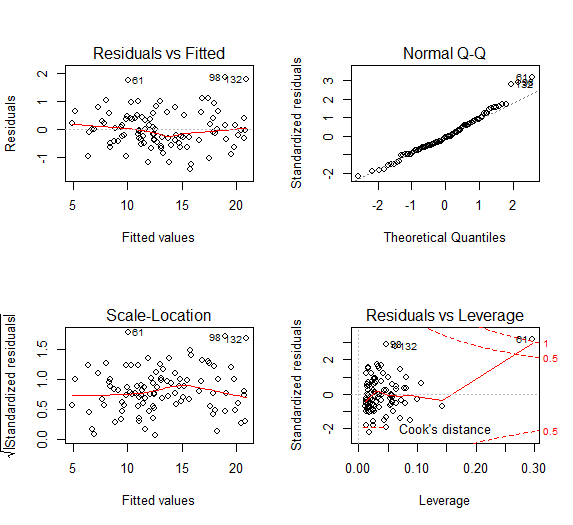
> pvalue\_for4

Head.L Head.W Sex

0.1629744 0.6958529 0.8359099

**Now, none of the p-values of the remaining predictors have p-value that is smaller than alpha-to-enter = 0.1, we stop adding predictors and the process stop.**

**From last lab, we obtained the model with predictor Head.L, Neck.G, Length, Chest.G by backward elimination. Thus, our model generated by forward selection is different from the model generated by backward elimination.**



**Since the scatter plot shows a non-linearity rend, the Q-Q plot shows a heavy tail, and the standardized residuals plot indicates an outlier, this model is not a good model.**

* Use best subset selection based on RSS (function regsubsets) in order to select the best model for each number of predictors (from 1 to 6 predictors). Is the model you selected in the previous point the best model according to RSS, among the ones with the same number of predictors?

> sum\_subset$which

(Intercept) Head.L Head.W Neck.G Length Chest.G Sex

1 TRUE FALSE FALSE FALSE FALSE TRUE FALSE

2 TRUE FALSE FALSE TRUE FALSE TRUE FALSE

3 TRUE FALSE FALSE TRUE TRUE TRUE FALSE

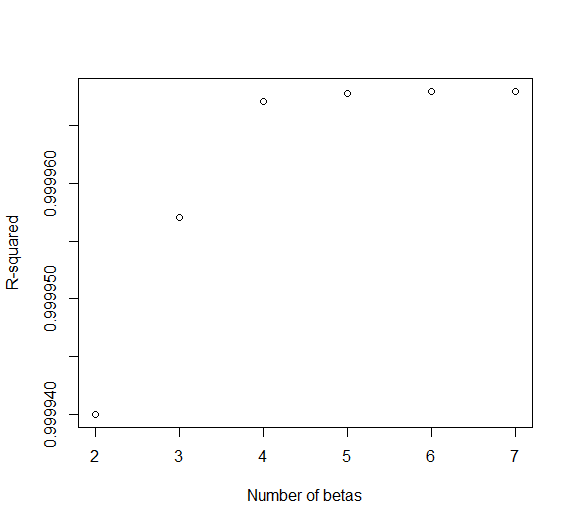
4 TRUE TRUE FALSE TRUE TRUE TRUE FALSE

5 TRUE TRUE TRUE TRUE TRUE TRUE FALSE

6 TRUE TRUE TRUE TRUE TRUE TRUE TRUE

**Among the ones with the same number of predictors, the model I selected in the previous point the best model according to RSS.**

* Compute the adjusted R-squared for each of the 6 models selected by best subset selection and plot it against the number of beta. Which model is the best?

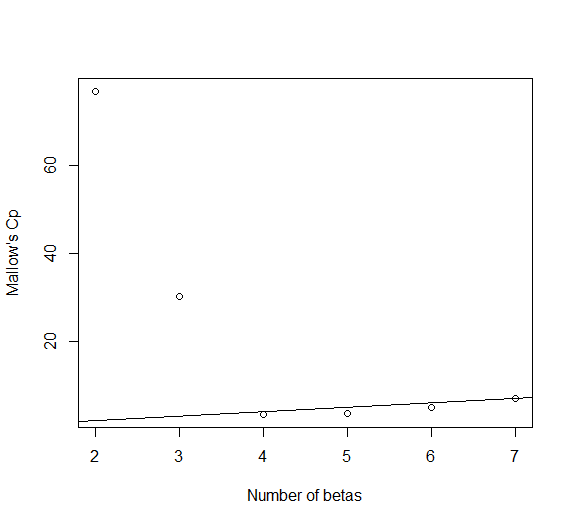


> R2\_adj

[1] 0.9999394 0.9999561 0.9999660 0.9999664 0.9999662 0.9999658

**Adjusted R square indicates that the model with 5 betas with 4 predictors (Head.L, Neck.G, Length, Chest.G) is the best model.**

* Compute the Mallow’s C for each of the 6 models selected by best subset selection and plot it against the number of beta. Add the line y=p to the plot. Which model is the best?



**Based on Mallow’s C, we should choose 4 betas with 3 predictors.**

* Compute the AIC for each of the 6 models selected by best subset selection and plot it against the number of beta. Which model is the best?

> aic\_p

[1] -22.05290 -52.93886 -77.31433 -77.37518 -75.87237 -73.91979

**Based on AIC, we should choose 5 betas with 4 predictors as the best model.**

* Compute the BIC for each of the 6 models selected by best subset selection and plot it against the number of beta. Which model is the best?

> bic\_p

[1] -16.86266 -45.15350 -66.93385 -64.39958 -60.30165 -55.75395

**Based on BIC, we should choose 4 betas with 3 predictors as the best model.**

* In conclusion, which model would you choose? Why?

**In conclusion, we should choose 4 betas with 3 predictors because it is simple and useful.**

R code:

setwd("//udrive.win.psu.edu/Users/j/q/jql5883/Desktop/math462")

getwd()

bears=read.csv("bears.txt", header=T, sep="")

bears=bears[bears$Obs.No==1,]

head(bears)

attach(bears)

Sex=Sex-1

#1----------------------------------------------------------------------

lm.bear=lm(sqrt(Weight)~1,data=bears)

summary(lm.bear)

pvalue\_for1=numeric(6)

names(pvalue\_for1)=c("Head.L","Head.W","Neck.G","Length","Chest.G","Sex")

pvalue\_for1[1]=summary(update(lm.bear,.~.+Head.L))$coefficients[2,4]

pvalue\_for1[2]=summary(update(lm.bear,.~.+Head.W))$coefficients[2,4]

pvalue\_for1[3]=summary(update(lm.bear,.~.+Neck.G))$coefficients[2,4]

pvalue\_for1[4]=summary(update(lm.bear,.~.+Length))$coefficients[2,4]

pvalue\_for1[5]=summary(update(lm.bear,.~.+Chest.G))$coefficients[2,4]

pvalue\_for1[6]=summary(update(lm.bear,.~.+Sex))$coefficients[2,4]

pvalue\_for1

lm.for1=update(lm.bear,.~.+Chest.G)

summary(lm.for1)

pvalue\_for2=numeric(5)

names(pvalue\_for2)=c("Head.L","Head.W","Neck.G","Length","Sex")

pvalue\_for2[1]=summary(update(lm.for1,.~.+Head.L))$coefficients[3,4]

pvalue\_for2[2]=summary(update(lm.for1,.~.+Head.W))$coefficients[3,4]

pvalue\_for2[3]=summary(update(lm.for1,.~.+Neck.G))$coefficients[3,4]

pvalue\_for2[4]=summary(update(lm.for1,.~.+Length))$coefficients[3,4]

pvalue\_for2[5]=summary(update(lm.for1,.~.+Sex))$coefficients[3,4]

pvalue\_for2

lm.for2=update(lm.for1,.~.+Neck.G)

summary(lm.for2)

pvalue\_for3=numeric(4)

names(pvalue\_for3)=c("Head.L","Head.W","Length","Sex")

pvalue\_for3[1]=summary(update(lm.for2,.~.+Head.L))$coefficients[4,4]

pvalue\_for3[2]=summary(update(lm.for2,.~.+Head.W))$coefficients[4,4]

pvalue\_for3[3]=summary(update(lm.for2,.~.+Length))$coefficients[4,4]

pvalue\_for3[4]=summary(update(lm.for2,.~.+Sex))$coefficients[4,4]

pvalue\_for3

lm.for3=update(lm.for2,.~.+Neck.G)

summary(lm.for3)

pvalue\_for4=numeric(3)

names(pvalue\_for4)=c("Head.L","Head.W","Sex")

pvalue\_for4[1]=summary(update(lm.for3,.~.+Head.L))$coefficients[5,4]

pvalue\_for4[2]=summary(update(lm.for3,.~.+Head.W))$coefficients[5,4]

pvalue\_for4[3]=summary(update(lm.for3,.~.+Sex))$coefficients[5,4]

pvalue\_for4

par(mfrow=c(2,2))

plot(lm.for3)

#2----------------------------------------------------------------------

install.packages("leaps")

require(leaps)

subset=regsubsets(sqrt(Weight)~Head.L+Head.W+Neck.G+Length+Chest.G+Sex,data=bears)

sum\_subset=summary(subset)

sum\_subset$which

#3----------------------------------------------------------------------

p\_full=7

p=2:p\_full

RSS\_p=sum\_subset$rss

totalSS=sum((bears$Weight-mean(bears$Weight))^2)

R2\_p=1-RSS\_p/totalSS

n=nrow(bears)

R2\_adj=1-(RSS\_p/(n-p))/(totalSS/(n-1))

R2\_adj

plot(p,R2\_p,xlab="Number of betas",ylab="R-squared")

#4----------------------------------------------------------------------

lm\_full=lm(sqrt(Weight)~Head.L+Head.W+Neck.G+Length+Chest.G+Sex,data=bears)

sigma\_hat\_full=summary(lm\_full)$sigma

C\_p=RSS\_p/(sigma\_hat\_full^2)+2\*p-n

C\_p

plot(p,C\_p,xlab="Number of betas",ylab="Mallow's Cp")

abline(0,1)

#-----------------------------------------------------------------------

aic\_p=n\*log(RSS\_p/n)+2\*p

aic\_p

bic\_p=n\*log(RSS\_p/n)+p\*log(n)

bic\_p