**STAT 462 – Applied Regression Analysis**

**Fall 2017, Lab 10**

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

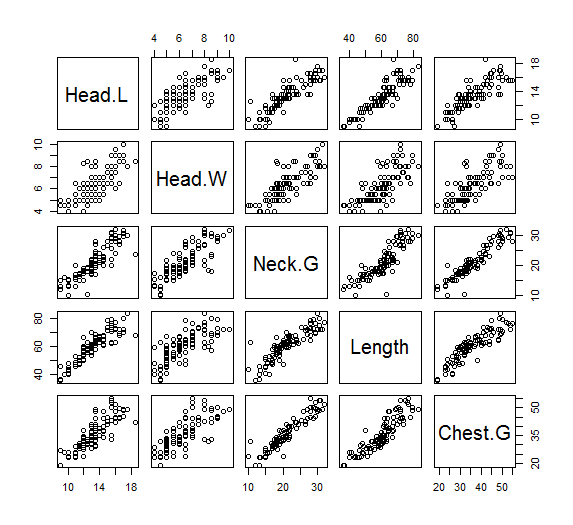
*Obs.No:* Observation number for this bear. For example, the bear with ID=41 (Bertha) was measured on four occasions. The value of Obs.No goes from 1 to 4 for these observations

*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 y=“Weight” and predictors x1=“Head.L”, x2=“Head.W”, x3=“Neck.G”, x4=“Length” and x5=“Chest.G”.

* Compute pairwise correlation among the 5 predictors and produce scatterplots for each pair of predictor. Do you see any strong pairwise correlation? Do you think you have a collinearity problem?



> cor(bears[,c(10,5:9)])

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

Weight 1.0000000 0.8222140 0.7580019 0.9462152 0.8755766 0.9631449

Head.L 0.8222140 1.0000000 0.7592926 0.8651442 0.8989541 0.8466939

Head.W 0.7580019 0.7592926 1.0000000 0.8063597 0.7261590 0.7561177

Neck.G 0.9462152 0.8651442 0.8063597 1.0000000 0.8714729 0.9390692

Length 0.8755766 0.8989541 0.7261590 0.8714729 1.0000000 0.8887218

Chest.G 0.9631449 0.8466939 0.7561177 0.9390692 0.8887218 1.0000000

**By looking at the plots, all of these predictors have strong correlation and pairwise between Length and Head.W has the least correlation while pairwise between Neck.G and Chest.G has the strongest correlation, which may indicate collinearity. However, we are not sure about this because strong linear interdependencies among the predictors can be present even when pairwise correlations are weak. Thus, we still may have collinearity problem.**

* For each of the 5 predictors xj in the model, consider it as a liner function of the other 4 predictors. Fit a linear regression model to predict xj using the other 4 predictors, and provide the coefficient of determination Rj2 and the variance inflation factor VIFj (you can check the VIF using the *vif* function in *car* library, but you need to compute it directly by yourself).

> X=model.matrix(Weight~Head.L+Head.W+Neck.G+Length+Chest.G)

> R2=vector("numeric",5)

> for(j in 1:5){

+ y\_tmp=X[,1+j]

+ x\_tmp=as.matrix(X[,-c(1,1+j)])

+ lm\_fit=lm(y\_tmp~x\_tmp)

+ R2[j]=summary(lm\_fit)$r.squared

+ }

> VIF=1/(1-R2)

> names(VIF)=c('Head.L', 'Head.W', 'Neck.G', 'Length', 'Chest.G')

> VIF

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

6.365595 2.999381 11.239028 7.451770 10.301203

> R2

[1] 0.8429055 0.6665979 0.9110243 0.8658037 0.9029240

* Looking at the VIF, do you have evidence of collinearity in your predictors?

**Since predictors except Head.W have VIFs close to 10, we have evidence to conclude that predictors have strong collinearity.**

Drop the predictor with the largest VIF and fit a new model to predict y=“Weight” without considering that predictor.

* Did the least square estimates of the betas for the 4 predictors changed a lot from the previous model?

> summary(lm.bear)

Call:

lm(formula = Weight ~ Head.L + Head.W + Neck.G + Length + Chest.G)

Residuals:

Min 1Q Median 3Q Max

-59.457 -17.969 -2.059 14.432 99.239

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -258.3771 20.8837 -12.372 < 2e-16 \*\*\*

Head.L -7.5230 3.3596 -2.239 0.0275 \*

Head.W 0.3087 3.3965 0.091 0.9278

Neck.G 8.5812 1.7639 4.865 4.65e-06 \*\*\*

Length 1.3305 0.7425 1.792 0.0764 .

Chest.G 7.8844 1.0190 7.738 1.19e-11 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27.24 on 93 degrees of freedom

Multiple R-squared: 0.9456, Adjusted R-squared: 0.9427

F-statistic: 323.5 on 5 and 93 DF, p-value: < 2.2e-16

> summary(lm.bear.drop)

Call:

lm(formula = Weight ~ Head.L + Head.W + Length + Chest.G)

Residuals:

Min 1Q Median 3Q Max

-82.817 -18.914 -0.591 17.441 101.371

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -286.9736 22.3251 -12.854 <2e-16 \*\*\*

Head.L -3.9334 3.6514 -1.077 0.2841

Head.W 5.9719 3.5548 1.680 0.0963 .

Length 1.3959 0.8270 1.688 0.0948 .

Chest.G 11.2612 0.8311 13.550 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 30.35 on 94 degrees of freedom

Multiple R-squared: 0.9318, Adjusted R-squared: 0.9289

F-statistic: 321 on 4 and 94 DF, p-value: < 2.2e-16

**Yes, the least square estimates of the betas for the 4 predictors changed a lot from the previous model.**

* For each of the 4 predictors xj in the new model, consider it as a liner function of the other 3 predictors. Fit a linear regression model to predict xj using the other 4 predictors, and provide the coefficient of determination Rj2 and the variance inflation factor VIFj (you can check the VIF using the *vif* function in *car* library, but you need to compute it directly by yourself).

> X2=model.matrix(Weight~Head.L+Head.W+Length+Chest.G)

> R2\_2=vector("numeric",4)

> for(j in 1:4){

+ y\_tmp=X2[,1+j]

+ x\_tmp=as.matrix(X2[,-c(1,1+j)])

+ lm\_fit=lm(y\_tmp~x\_tmp)

+ R2\_2[j]=summary(lm\_fit)$r.squared

+ }

> VIF2=1/(1-R2\_2)

> names(VIF2)=c('Head.L', 'Head.W', 'Length', 'Chest.G')

> VIF2

Head.L Head.W Length Chest.G

6.058524 2.647062 7.449330 5.521270

> R2\_2

[1] 0.8349433 0.6222227 0.8657597 0.8188823

* Looking at the VIF, do you still have evidence of strong collinearity in your predictors?

**Since predictors except Head.W still have VIFs close to 10, we still have evidence to conclude that predictors have strong collinearity.**

**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)

lm.bear=lm(Weight~Head.L+Head.W+Neck.G+Length+Chest.G)

plot(bears[5:9])

cor(bears[,c(10,5:9)])

X=model.matrix(Weight~Head.L+Head.W+Neck.G+Length+Chest.G)

R2=vector("numeric",5)

for(j in 1:5){

y\_tmp=X[,1+j]

x\_tmp=as.matrix(X[,-c(1,1+j)])

lm\_fit=lm(y\_tmp~x\_tmp)

R2[j]=summary(lm\_fit)$r.squared

}

VIF=1/(1-R2)

names(VIF)=c('Head.L', 'Head.W', 'Neck.G', 'Length', 'Chest.G')

VIF

R2

lm.bear.drop=lm(Weight~Head.L+Head.W+Length+Chest.G)

summary(lm.bear)

summary(lm.bear.drop)

X2=model.matrix(Weight~Head.L+Head.W+Length+Chest.G)

R2\_2=vector("numeric",4)

for(j in 1:4){

y\_tmp=X2[,1+j]

x\_tmp=as.matrix(X2[,-c(1,1+j)])

lm\_fit=lm(y\_tmp~x\_tmp)

R2\_2[j]=summary(lm\_fit)$r.squared

}

VIF2=1/(1-R2\_2)

names(VIF2)=c('Head.L', 'Head.W', 'Length', 'Chest.G')

VIF2

R2\_2