**STAT 462 – Applied Regression Analysis**

**Fall 2017, Lab 12**

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

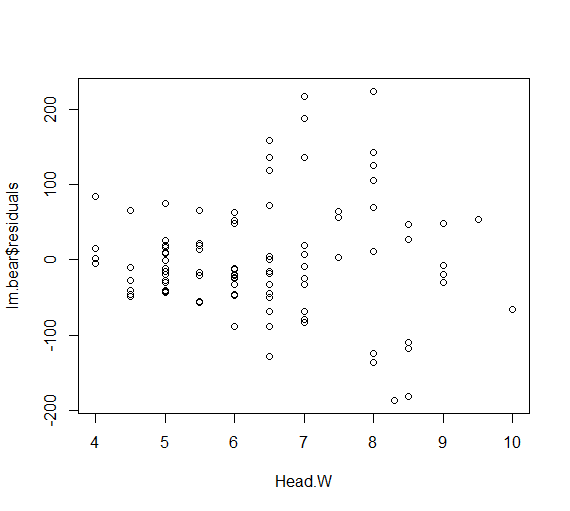
*Weight:* Weight of the bear, in pounds

*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,]).

* Fit a single linear regression model with response y=“Weight” and predictor x=“Head.W”. Draw a scatterplot of the residuals versus the predictor x. How does the variance of errors vary with x (constant, increase, decrease…)?



**This scatter plot indicates that variance of errors increases as the predictor increases.**

* Using the matrix formula, compute the weighted least square estimates beta\_w of a model with response y=“Weight” and predictor x=“Head.W”. Decide how to define the weight matrix **W** based on previous question.
  + Draw a scatterplot of the weighted residuals versus the predictor x. Is the variance approximately constant? If not, you probably didn’t use the “right” weight matrix **W**.

> w=1/(Head.W^2)

> W=diag(w)

> X=model.matrix(Weight~Head.W)

> y=Weight

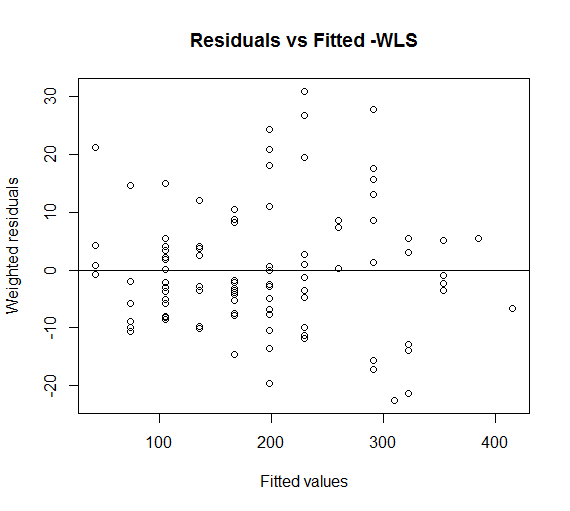
> beta\_hat\_weight=solve(t(X)%\*%W%\*%X,t(X)%\*%W%\*%y)

> y\_hat=X%\*%beta\_hat\_weight

> res\_weight=sqrt(w)\*(y-y\_hat)

> plot(y\_hat,res\_weight,xlab='Fitted values',ylab='Weighted residuals',main='Residuals vs Fitted -WLS')

> abline(h=0)



**In this plot, we can see that we still do not have the constant variance, which means our weight is not enough.**

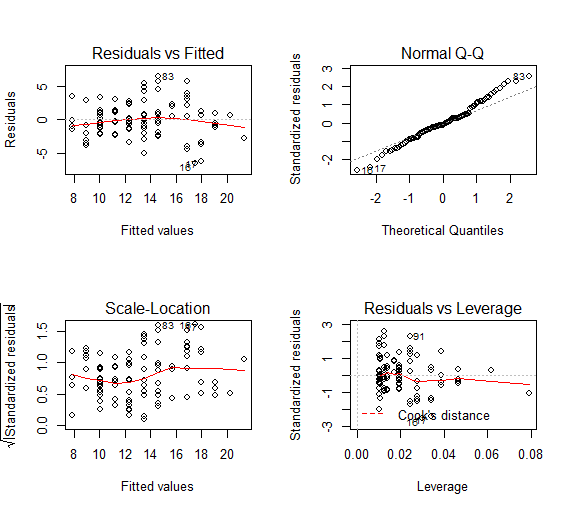
* An alternative to the use of weighted least squares is represented by variable transformation. Which variable do you need to transform? What kind of transformation can you use in order to fix the non-constant variance issue?
  + Fit a single linear regression model with response y=“Weight” and predictor x=“Head.W”, after performing the transformation. Produce diagnostic plots to make sure that the transformation worked.

I need to transform the response y = “Weight”. I use square root of y to transform the model to fix the non-constant variance issue.

> lm.bear.transf=lm(sqrt(Weight)~Head.W)

> par(mfrow=c(2,2))

> plot(lm.bear.transf)



**The scatter plot indicates approximately constant variance and linearity. The Q-Q plot also indicates normality. Thus, this transformation works.**

Consider now a multiple linear regression model with response the square root of “Weight” and 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 backward elimination with threshold alpha-to-remove 10% in order to select a simple model to predict the bear weight. Report which predictor you remove at each step, and the final model that you obtain. Do you think it is a good model?

> Sex=Sex-1

> lm.bear.full=lm(Weight~Head.L+Head.W+Neck.G+Length+Chest.G+Sex,data=bears)

> summary(lm.bear.full)

Call:

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

Sex, data = bears)

Residuals:

Min 1Q Median 3Q Max

-61.613 -19.255 -1.412 15.318 98.055

Coefficients:

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

(Intercept) -245.69864 24.18347 -10.160 < 2e-16 \*\*\*

Head.L -7.63277 3.35985 -2.272 0.0254 \*

Head.W -0.08356 3.41607 -0.024 0.9805

Neck.G 8.14327 1.81289 4.492 2.05e-05 \*\*\*

Length 1.37690 0.74349 1.852 0.0672 .

Chest.G 8.06466 1.03321 7.805 9.12e-12 \*\*\*

Sex -6.50900 6.26836 -1.038 0.3018

---

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

Residual standard error: 27.23 on 92 degrees of freedom

Multiple R-squared: 0.9463, Adjusted R-squared: 0.9428

F-statistic: 270 on 6 and 92 DF, p-value: < 2.2e-16

> lm.back1=update(lm.bear.full, .~.-Head.W)

> summary(lm.back1)

Call:

lm(formula = Weight ~ Head.L + Neck.G + Length + Chest.G + Sex,

data = bears)

Residuals:

Min 1Q Median 3Q Max

-61.762 -19.256 -1.403 15.315 98.072

Coefficients:

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

(Intercept) -245.8010 23.6906 -10.375 < 2e-16 \*\*\*

Head.L -7.6486 3.2795 -2.332 0.0218 \*

Neck.G 8.1297 1.7169 4.735 7.82e-06 \*\*\*

Length 1.3775 0.7391 1.864 0.0655 .

Chest.G 8.0650 1.0276 7.849 7.01e-12 \*\*\*

Sex -6.4920 6.1963 -1.048 0.2975

---

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

Residual standard error: 27.08 on 93 degrees of freedom

Multiple R-squared: 0.9463, Adjusted R-squared: 0.9434

F-statistic: 327.5 on 5 and 93 DF, p-value: < 2.2e-16

> lm.back2=update(lm.back1, .~.-Sex)

> summary(lm.back2)

Call:

lm(formula = Weight ~ Head.L + Neck.G + Length + Chest.G, data = bears)

Residuals:

Min 1Q Median 3Q Max

-58.881 -17.868 -2.097 14.646 99.186

Coefficients:

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

(Intercept) -258.118 20.579 -12.543 < 2e-16 \*\*\*

Head.L -7.463 3.276 -2.278 0.0250 \*

Neck.G 8.636 1.648 5.239 9.84e-07 \*\*\*

Length 1.328 0.738 1.799 0.0752 .

Chest.G 7.882 1.013 7.780 9.22e-12 \*\*\*

---

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

Residual standard error: 27.1 on 94 degrees of freedom

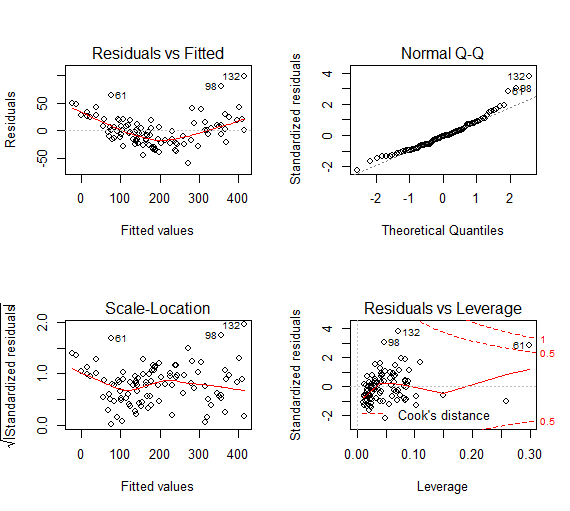
Multiple R-squared: 0.9456, Adjusted R-squared: 0.9433

F-statistic: 408.7 on 4 and 94 DF, p-value: < 2.2e-16

**First, we check the summary of the full model and we see that Head.L has the largest p-value that is larger than 0.1. So we remove this predictor and get a new model “lm.back1”.**

**Then, we continue to check the summary of our new model and see that the predictor Sex has a p-value that is larger than 0.1, so we remove this predictor and get a new model “lm.back2”.**

**Now, we check this model and see that no predictor has p-value that is larger than 0.1. So we obtain our final model with predictor Head.L, Neck.G, Length, Chest.G.**



**These diagnostic graph shows that data has approximately constant variance, but the linearity may not be satisfied. Q-Q plot shows us the normality. Thus, this model is a good model in general.**

**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.W)

plot(Head.W,lm.bear$residuals)

w=1/(Head.W^2)

W=diag(w)

X=model.matrix(Weight~Head.W)

y=Weight

beta\_hat\_weight=solve(t(X)%\*%W%\*%X,t(X)%\*%W%\*%y)

y\_hat=X%\*%beta\_hat\_weight

res\_weight=sqrt(w)\*(y-y\_hat)

plot(y\_hat,res\_weight,xlab='Fitted values',ylab='Weighted residuals',main='Residuals vs Fitted -WLS')

abline(h=0)

lm.bear.transf=lm(sqrt(Weight)~Head.W)

par(mfrow=c(2,2))

plot(lm.bear.transf)

Sex=Sex-1

lm.bear.full=lm(Weight~Head.L+Head.W+Neck.G+Length+Chest.G+Sex,data=bears)

summary(lm.bear.full)

lm.back1=update(lm.bear.full, .~.-Head.W)

summary(lm.back1)

lm.back2=update(lm.back1, .~.-Sex)

summary(lm.back2)

plot(lm.back2)