FACTORS INFLUENCING TOTAL MERCURY IN THE HAIR OF KUWAITI

➢ INTRODUCTION

As we all know, mercury is a poisonous pollutant to environment. For human being, mercury is a powerful neurotoxin. Nowadays, lots of human activities increase mercury content of environment, for instance, metal combustion, ore milling and etc. Various pollutants are getting involved in food chain and potentially doing harm to our health. Since mercury can be remaining in the hair and our hair almost has no process of metabolism for any elements in our body. Therefore, hair is a good way to measure the level of mercury containing in body. Because environment pollution and food safety have been hot issues these days. I decided to learn this topic for the final project.

The data I used for analysis is "Factors Influencing the Total Mercury and Methyl Mercury in the Hair of Fisherman in Kuwait" (N.B. Al-Majed and M.R. Preston, 2000). There are n = 135 observations collected from a group of Kuwaiti fisherman and a control group of non-fisherman in the data set. Since Methyl Mercury(mg/g) is included in the (TotHg)Total Mercury(mg/g), there will be some functional relationship between these two variables. Therefore, Total Mercury containing in the hair of Kuwaiti will be the response variable(Yi) of the regression model. Also, I guess the following variables are the explanatory variables.

X1(fisherman): Fisherman which indicates whether he's a fisherman or not;

X2(age): Age(years) ranging from 16 - 58 years old;

X3(restime): Residence Time(years) ranging form 0 – 25 years;

X4(weight): Weight(kg);

X5(fishmlwk): Fish meals consumed per week ranging from 0 - 21

X6(fishpart): Parts of fish consumed, indicate variable with 4 levels, 0 indicates not eating fish, 1 indicates muscle tissue only, 2 indicates muscle tissue and sometimes whole fish and 3 indicates whole fish;

X7(height): Height(cm).

Before analysis, I think fishmlwk and fishpart will directly affect TotHg in the hair. Also, fisherman will show more TotHg in the hair than non-fisherman. Weight, height, age and restime might be the factors of TotHg.

The main analysis for the regression model is hypothesis test and ANOVA. I will conduct the inference test from SLR to MLR to improve the model to a better reduced model.

- ANALYSIS

Step1: Simple Linear Regression

I conducted 7 single regression model for each explanatory variable. Using hypothesis test for each SLR where H_0: β_i (i ≠ 0) = 0 and H_a: β_i (i ≠ 0) ≠ 0, I want to check if we can directly conclude the existence of linear relationship between the explanatory variable and TotHg. Here is the p-value table for each SLR.

Xi	fisherman	age	restime	height	weight	fishmlwk	fishpart
P-value	0.006272	0.06384	0.474	0.02555	8.281e-07	0.0003557	three parts < 0.0016
evidence	strong	weak	no	moderate	strong	strong	strong-moderate

From each p-value, I found there is strong evidence to show linear relationship between TotHg and weight as well as fishmlwk. At the meantime, there is strong evidence to show difference of TotHg between fisherman and non-fisherman as well as which parts of the fish being consumed. I also found there is moderate evidence to show linear relationship between TotHg and height and weak evidence for TotHg and age. More importantly, there is no evidence to show linear relationship between TotHg and restime which is contrary to my guess. Therefore, I will take fisherman, age, height, weight, fishmlwk and fishpart as my independent variables in my MLR model.

Step2: Multiple Linear Regression

	fisherman	age	height	weight	fishmlwk	fishpart1	fishpart2	fishpart3
p-value	0.29571	0.69371	0.32005	4.6e-05	0.10419	0.09554	0.28232	0.15685
slope	0.76235	0.01165	0.03409	0.15245	0.08942	1.79684	1.06110	1.89060

Full_model: $\hat{Y}$ = -16.05524 + 0.76235 * fisherman + 0.01165*age + 0.03409*height + 0.15245*weight + 0.08942*fishmlwk + 1.79684*fishpart1 + 1.06110*fishpart2 + 1.89060 * fishpart3

Here I noticed that the p-value for each independent variable became larger than SLR. So there maybe exists some interaction terms in the final model. Therefore, I checked the correlation between these variables to see if there exits the interaction terms.

Step3: Multicollinearity

From the correlation plot, I found that there are no strong correlation between any two variables. Also, to make sure there will be no interaction term in the final regression model, I also checked VIF under the assumption of full model. There are no such a VIF which is greater than 3. As a result, there will be no interaction term in my final model.

Step4: Reduced model

From the fact that hypothesis test of H_0 = all $\beta_i = 0$ and the output of p-value from summary which is 3.425e-07, there is strong evidence to show that at least one independent variable has linear relationship with TotHg. So I first reduced variable age since it has the largest p-value. Following the same procedure, I found the final model which gave me a perfect p-value with H_0: $\beta_i = 0$. P-value for testing slope of weight and height are 5.34e-07 and 0.000216 respectively.

reduced_model: $\hat{Y}$ = -10.07582 + 0.17518*weight + 0.15884*fishmlwk

Step5: Check ANOVA

Checking anova(reduced_model, full_model) to make sure the final model is good enough to make prediction of TotHg. The p-value is based on the hypothesis test whose H_0: $\beta_{height} = \beta_{age} = \beta_{fisherman} = \beta_{fishpart1} = \beta_{fishpart2} = \beta_{fishpart3} = 0$. Since p-value is 0.2334, there is no evidence to show at least one of them has linear relationship with TotHg. Therefore, the final model is better than full model.

Step6: Check assumptions + Transformation

After getting the final model, I tried to check the assumptions for the MLR model. Firstly, the Residuals vs. Fitted plot shows that it almost a good fit linear model but the variance is not constant. It spreads out from small left to the right. As a result, I tried to find a transformation for my model according to box-cox plot which shows that $\lambda = ½$. After the transformation, the residuals are almost randomly distributed which shows a better constant variance.

Model after transformation: sqrt($\hat{Y}$) = -1.457879 + 0.041793*weight + 0.033532*fishmlwk

From Normal Q-Q plot, I found that the error terms are normally distributed with two tails.

From Residual vs. Leverage plot, since there is no points laying outside the cook's distance area, there is no influential cases for this model.

- Conclusion

Final model: $\widehat{\text{sqrt(TotHg)}}$ = -1.457879 + 0.041793*weight + 0.033532*fishmlwk

In conclusion, there is strong evidence to show that there is linear relationship between TotHg and weight as well as fishmlwk. Not surprisingly, the more fish we eat per week, the more mercury content will be shown in the hair since fish has been part of food chain of human beings. One person's weight also has positive linear relationship with mercury content in the hair. In this regression model, β_0 has no practical meaning since there's no one with weight 0. From the model, we can also know that when holding fishmlwk constant and increasing 1 kg in weight, the mean value of square root of total mercury will increase by 0.041793 mg/g. When holding weight constant and increasing 1 fish meal per week, the mean value of square root of total mercury will increase 0.033532 mg/g.

However, there's no statistically distinguishable relationship between total mercury in the hair and age as well as residence time in Kuwait. Meanwhile, following the procedure to the final model, variable fish part and fisherman indicator has been reduced although they show strong evidence of linear relationship towards total mercury. Consequently, according to SLR of fisherman indicator, I can conclude that group of fisherman shows 1.5642 mg/g higher mercury content in the hair than non-fisherman. Also, according to SLR of fish parts, the plot shows that people not consume fish has significantly low mercury content in the hair than those who consume fish while there is no evidence to show that fish parts we consume will affect total mercury.

As environment pollution becomes serious these days, the mercury coming from industrial pollutants and human being activities has access to the nature cycle. We can choose to consume less fish these days to prevent from intaking mercury and keep healthy.

- Appendix

R code:

```
f = read.csv("/Users/JeremyZhang/Desktop/2016fall/sta302/Assignment/A3/fishermen_mercury.csv",header=T)

par(mfrow=c(2,2))


# factor the dummy variables

f$fisherman <- as.factor(f$fisherman)

f$fisherman

f$fishpart <- as.factor(f$fishpart)

f$fishpart


# Simple Linear Regression


# X1 = fisherman indicator

fit1 <- lm(f$TotHg ~ f$fisherman)

summary(fit1)

plot(f$TotHg ~ f$fisherman)

# p-value: 0.006272

# We have strong evidence to show there is linear relationship between total mercury and whether the

# person is a fisherman or not.


# X2 = Age in years

fit2 <- lm(f$TotHg ~ f$age)

summary(fit2)

# p-value: 0.06384
```

```
# We have weak evidence to show there is linear relationship between
total mercury and age.


# X3 = Residence time in years

fit3 <- lm(f$TotHg ~ f$restime)

summary(fit3)

# p-value: 0.474

# We have no evidence to show there is linear relationship between total
mercury and residence time in Kuwait.


# X4 = Weight in kg

fit4 <- lm(f$TotHg ~ f$weight)

summary(fit4)

# p-value: 8.281e-07

# We have strong evidence to show there is linear relationship between
total mercury and weight.


# X5 = Fish meals per week

fit5 <- lm(f$TotHg ~ f$fishmlwk)

summary(fit5)

# p-value: 0.0003557

# There is strong evidence to show there is linear relationship between
total mercury and number of

# fish meal per week.


# X6 = Parts of fish consumed

fit6 <- lm(f$TotHg ~ f$fishpart)

summary(fit6)
```

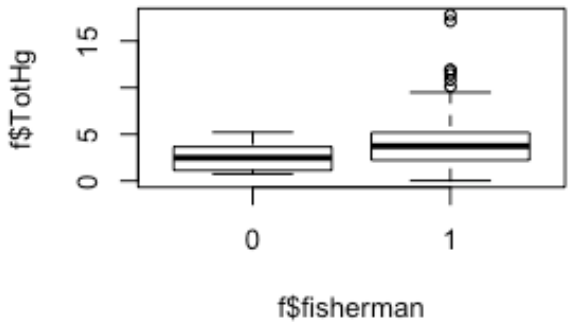
```
plot(f$TotHg ~ f$fishpart)

plot(fit6)

# p-value: 0.0004106

# There is strong evidence to show there is linear relationship between
total mercury and fish part consumed.


fit7 <- lm(f$TotHg ~ f$height)

summary(fit7)

# p-value: 0.0255

# There is moderate evidence to show there is linear relationship between
total mercury and height.


# Multiple Linear Regression


full_model = lm(f$TotHg ~ f$fisherman + f$age + f$height + f$weight +
f$fishmlwk + f$fishpart)

summary(full_model)

vif(full_model)


# check for correlation between all variables

mycor <- function ( data ){

  # ------------ put histograms on the diagonal ----------------

  panel.hist <- function (x , ...) {

    usr <- par ("usr") ; on.exit( par (usr ))

    par ( usr = c( usr [1:2] , 0, 1.5) )

    h <- hist (x , plot = FALSE )

    breaks <- h$ breaks ; nB <- length ( breaks )

    y <- h$ counts ; y <- y/ max (y)
```

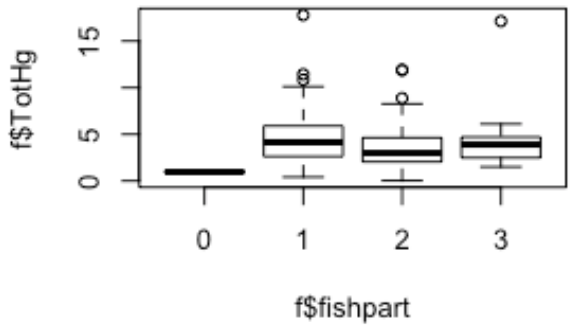
```
  rect ( breaks [ - nB ] , 0, breaks [ -1] , y , col ="lavender", ...)
 }

 panel.cor <- function (x , y , digits =4 , prefix ="", cex.cor , ...) {
  usr<- par ("usr") ;
  on.exit ( par ( usr ))
  par(usr = c(0 , 1 , 0 , 1) )
  txt1 <- format ( cor (x ,y) , digits = digits )
  txt2 <- format (cor.test (x ,y)$p.value , digits = digits )
  text (0.5 ,0.5 , paste ("r=",txt1 , "\n P.val =", txt2) , cex=0.8)
 }
 pairs (data , lower.panel = panel.cor , cex =0.7 , pch = 21 , bg
="steelblue",
      diag.panel = panel.hist , cex.labels = 1.1 ,
      font.labels =0.9 , upper.panel = panel.smooth )
}
# ----------- put correlations & P- value & 0.95 CIs on the lower panels -------
---------


a3cor = f[,c(2:7,9)]
str(a3cor)
mycor(a3cor)


# reducing model
reduced_model1 = lm(f$TotHg ~ f$fisherman + f$height + f$weight +
f$fishmlwk + f$fishpart)
summary(reduced_model1)
```

```
reduced_model2 = lm(f$TotHg ~ f$fisherman + f$weight + f$fishpart + f$fishmlwk)

summary(reduced_model2)


reduced_model3 = lm(f$TotHg ~ f$fisherman + f$weight + f$fishpart + f$fishmlwk)


reduced_model4 = lm(f$TotHg ~ f$fisherman + f$weight + f$fishmlwk)

summary(final_model)


reduced_model = lm(f$TotHg ~ f$weight + f$fishmlwk)

summary(reduced_model)

plot(reduced_model)


anova(reduced_model, full_model)

# box-cox to find a better model

library(MASS)

b <- boxcox(f$TotHg ~ f$fishmlwk + f$weight)


t_model = lm(sqrt(f$TotHg) ~ f$weight + f$fishmlwk)

summary(t_model)

plot(t_model)

# check if the reduced model is the best

nullmod = lm(f$TotHg ~ 1)

mboth = step(nullmod, scope=list(lower=formula(nullmod), upper=formula(full_model)), direction="both")
```

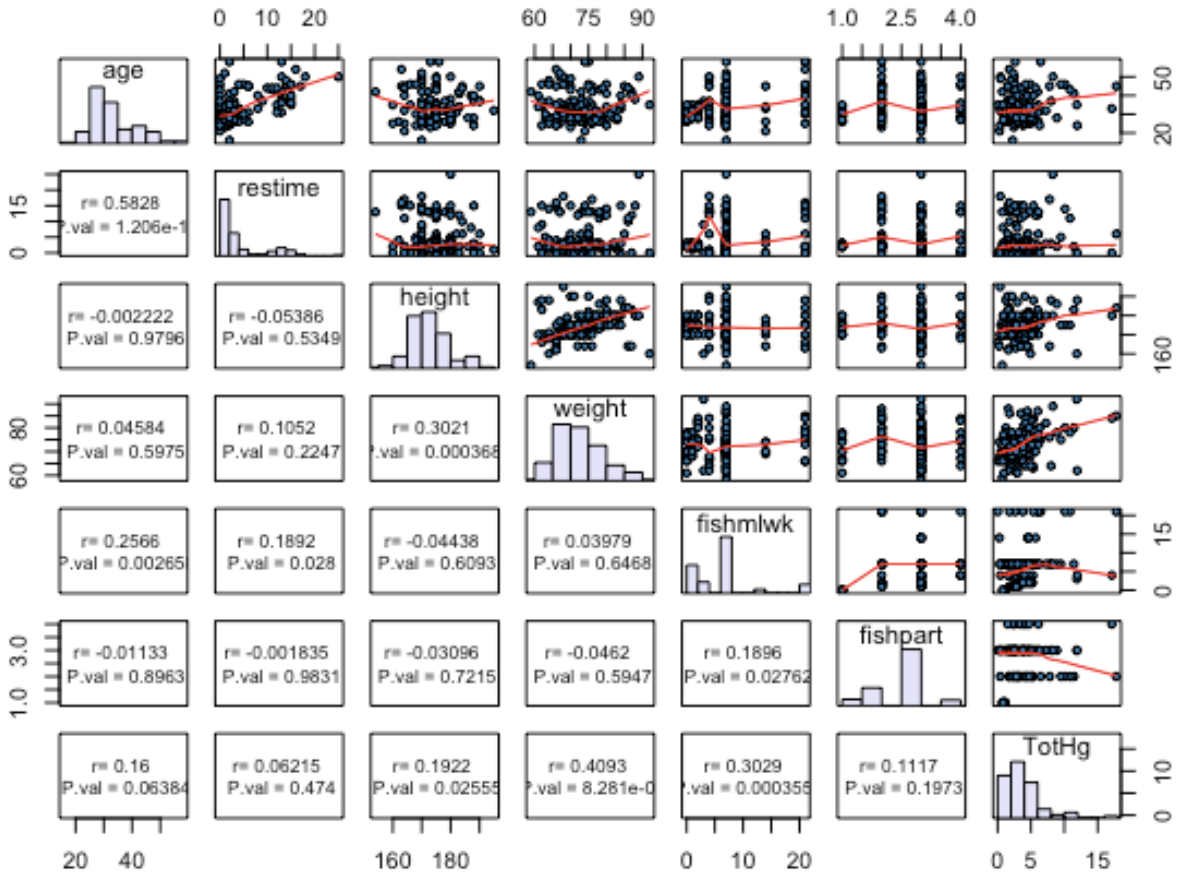
Graphs:

SLR: plot(f$TotHg ~ f$fisherman)

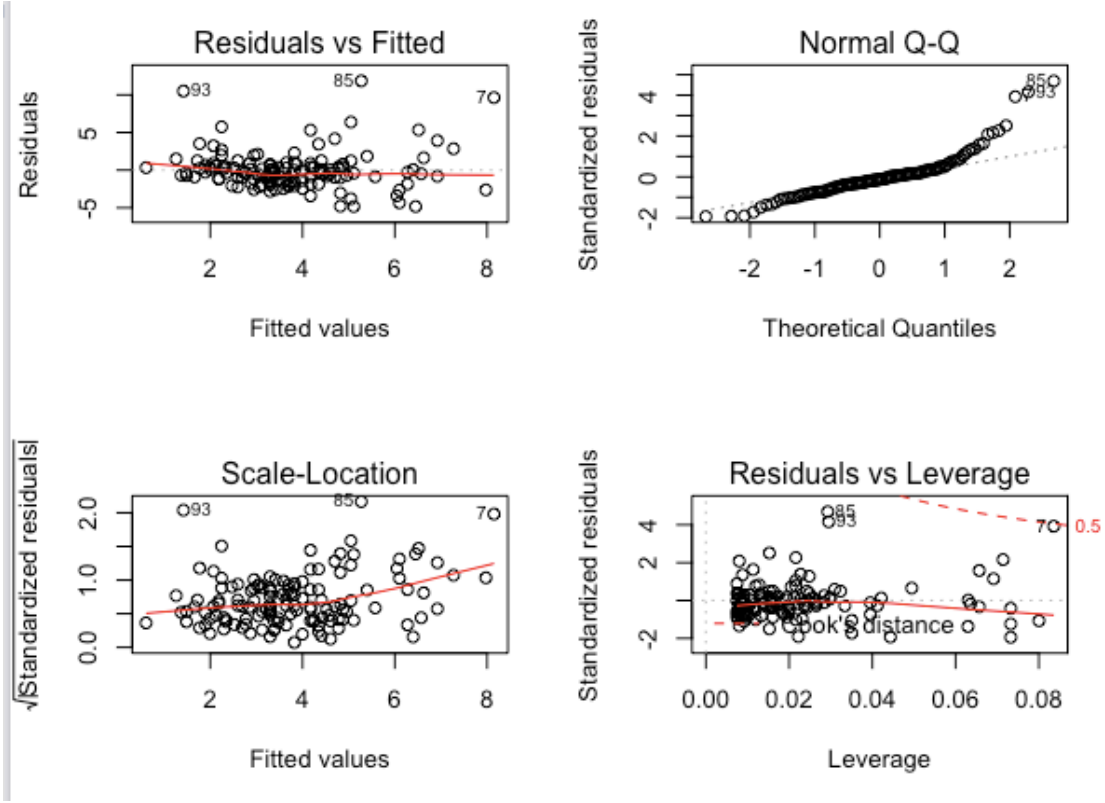


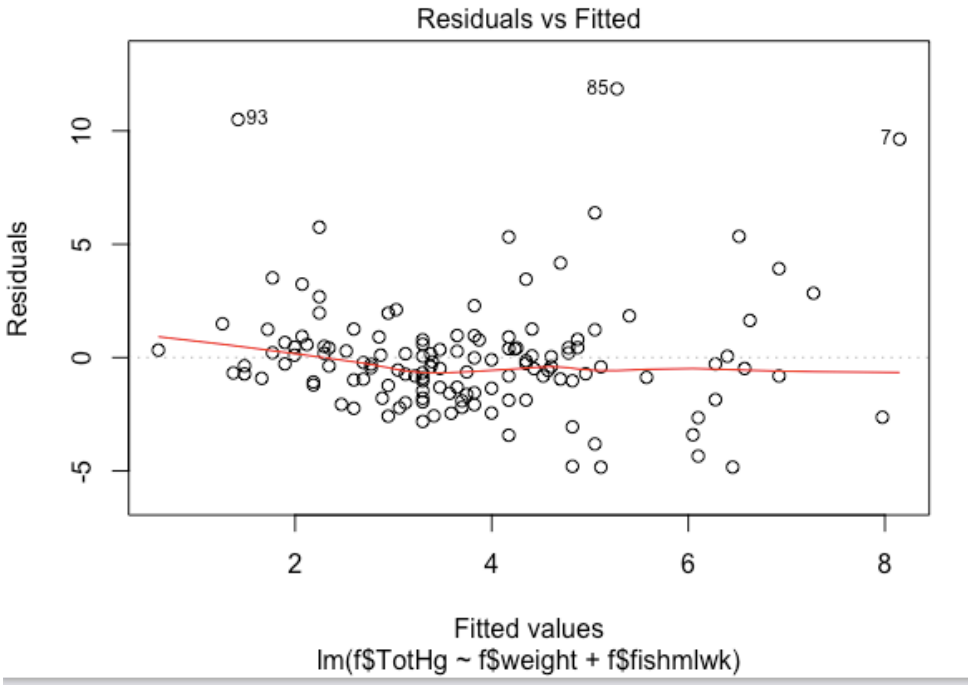


correlation chart



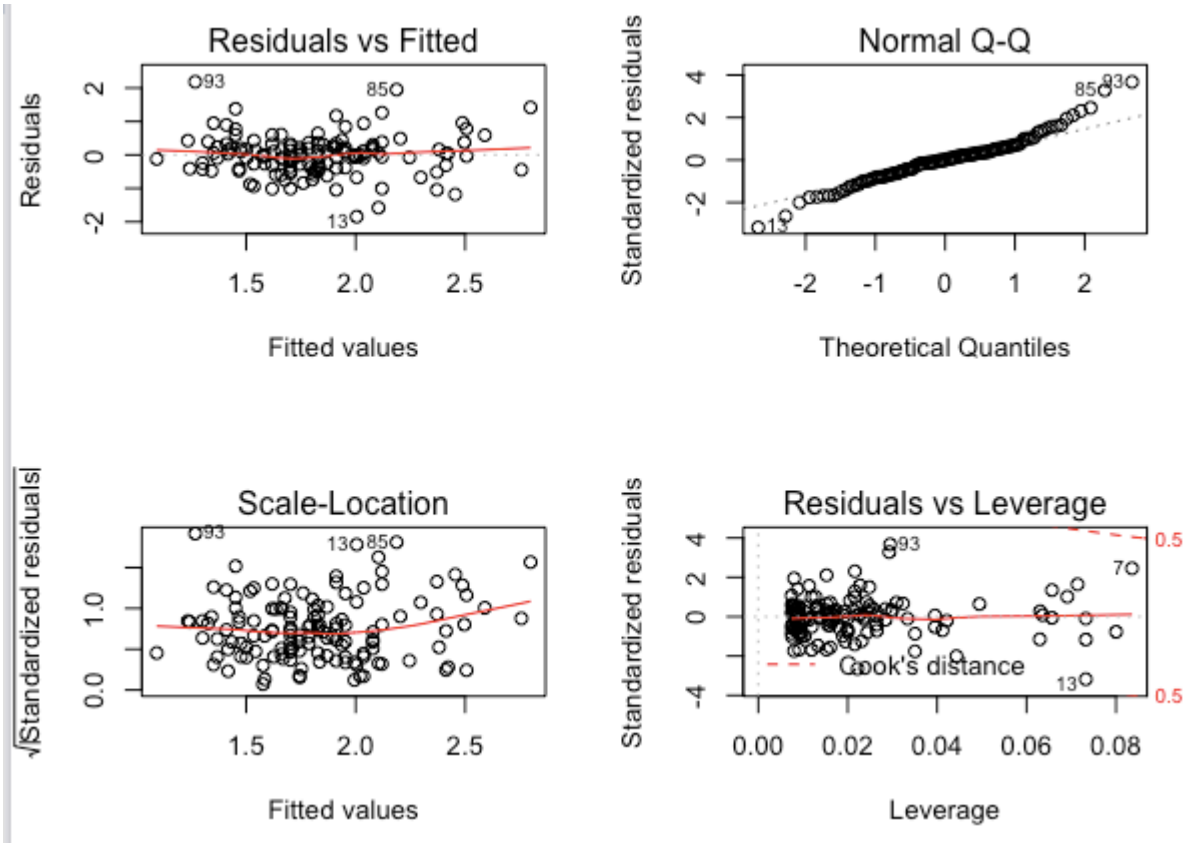
MLR: checking assumption for reduced_model

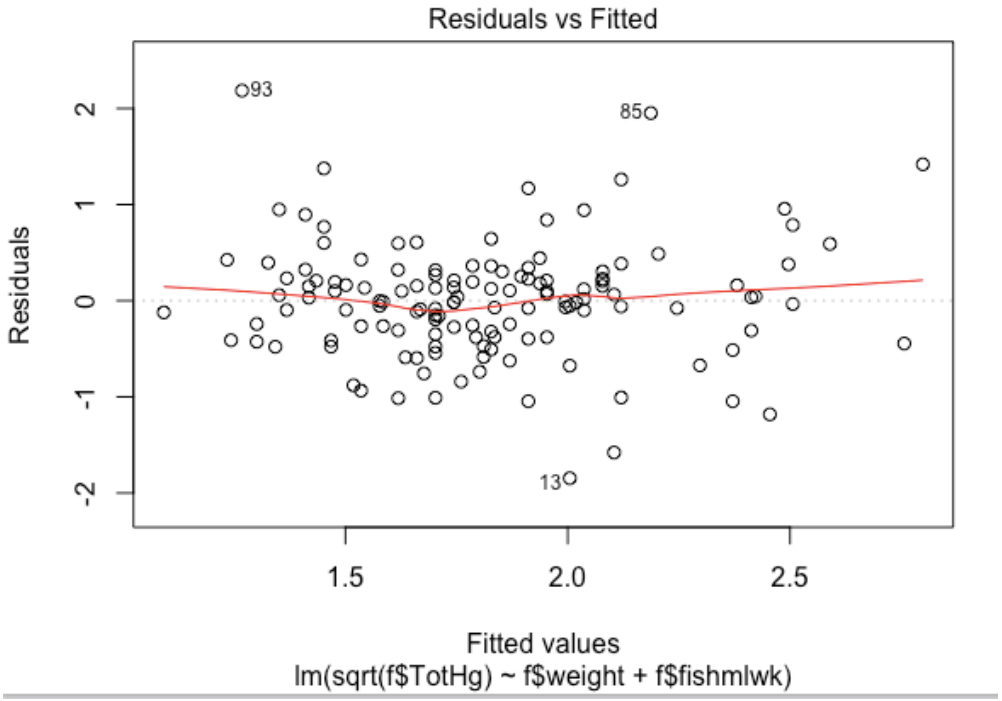




zoom in Residual vs. Fitted plot

MLR: checking assumption for model after transformation





zoom in Residuals vs. Fitted plot