Part 2 Question 3 Linear Regression

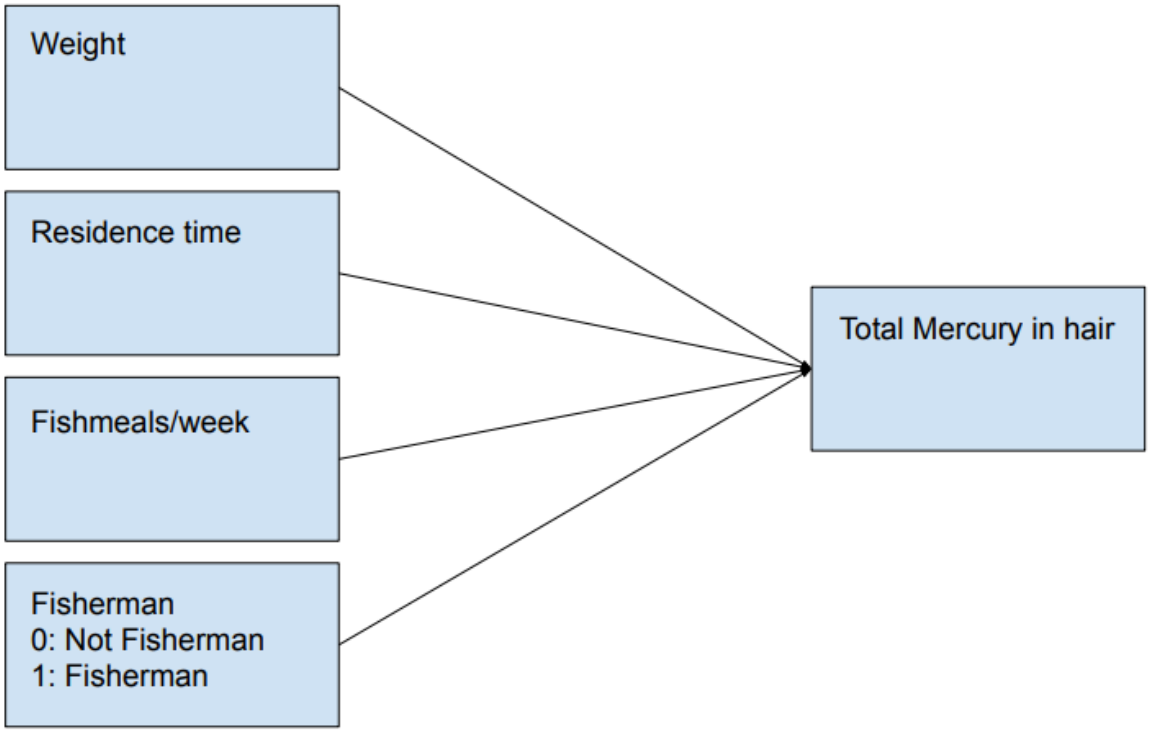
##Question 3 - Linear regression analysis

Conceptual model

Make a conceptual model underlying this research question

The model underlying the research question is that the amount of fish consumed by fishermen in Kuwait as well as factors such as their residence time, fishmeals consumed/week and weight have an impact on the amount of mercury in their hair.

Our independent variables are Fisherman, residence time, fishmeals/week and weight. The dependent variable is the Total mercury in the fishermans hair.



#

Visual inspection

Graphical analysis of the distribution of the dependent variable, e.g. histogram, density plot

```
# Question 2.3: Linear regression
library(foreign) #open various data files
library(car) #Package includes Levene's test
```

```
## Loading required package: carData
```

```
library(tidyr)  # for wide to long format transformation of the data
library(ggplot2)
library(QuantPsyc) #include lm.beta()
```

```
## Loading required package: boot

##
## Attaching package: 'boot'

## The following object is masked from 'package:car':
##
##     logit

## Loading required package: MASS

##
## Attaching package: 'QuantPsyc'

## The following object is masked from 'package:base':
##
##     norm
```

```
library(gmodels)
library(pander) #for rendering output
library(ez) #for ezANOVA
```

```
## Registered S3 methods overwritten by 'lme4':
##   method                          from
##   cooks.distance.influence.merMod car
##   influence.merMod                car
##   dfbeta.influence.merMod         car
##   dfbetas.influence.merMod        car
```

```
library(lmtest)
```

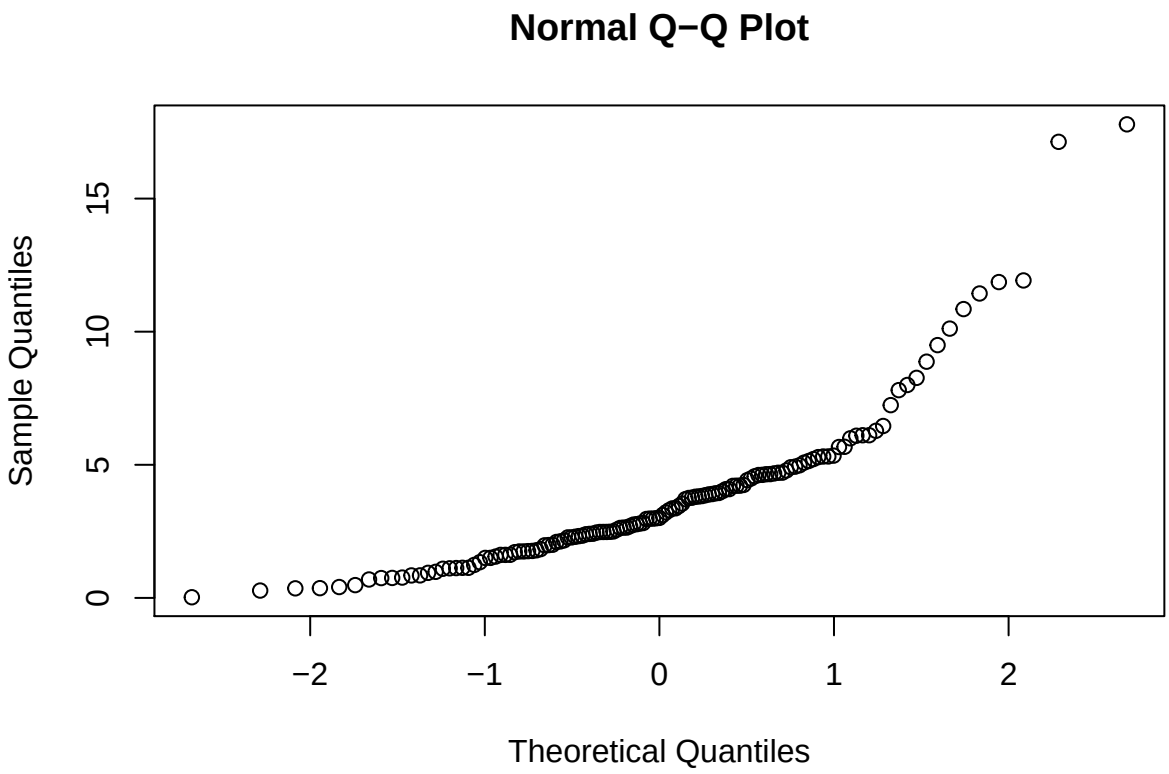
```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

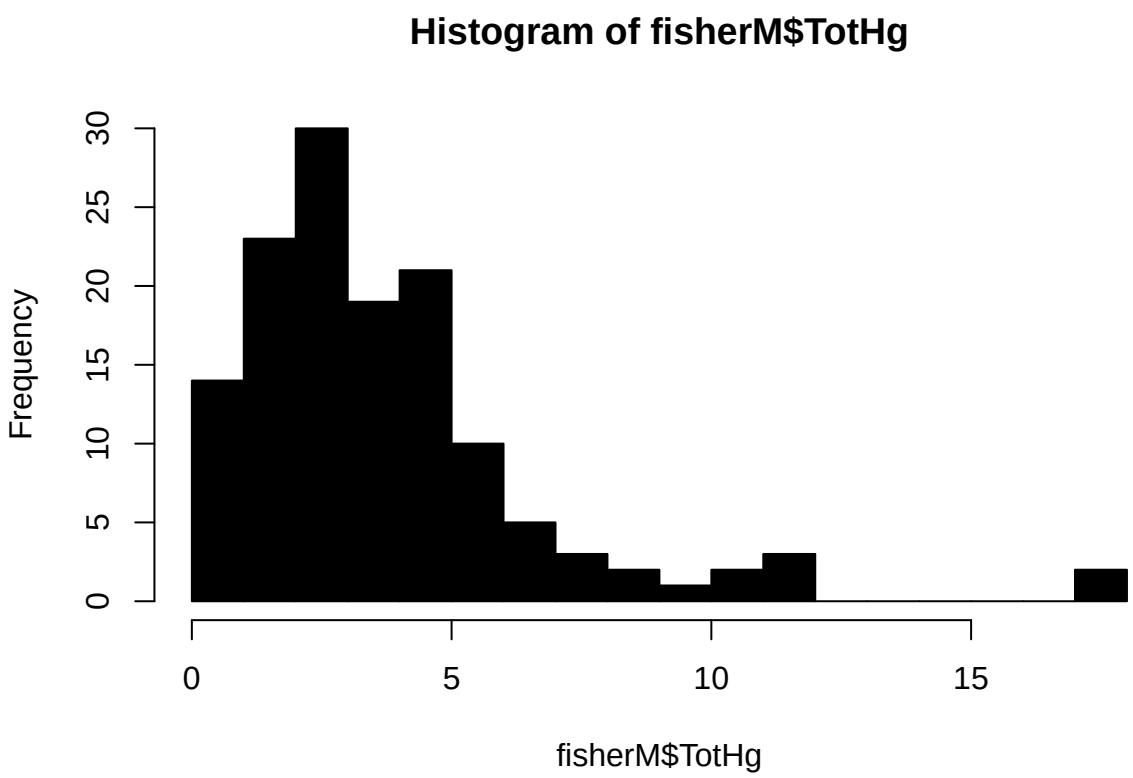
```
# Getting Data:
fisherM<-read.csv("fishermen_mercury.csv", header = TRUE)
fisherM$fisherman<-factor(fisherM$fisherman, levels =c(0:1), labels =c("NotFisherman","Fisherman"))
fisherM$fishpart<-factor(fisherM$fishpart, levels =c(0:3), labels =c("none","muscle tissue", "mt or whol

# Analysing Data:
qqnorm(fisherM$TotHg)
```

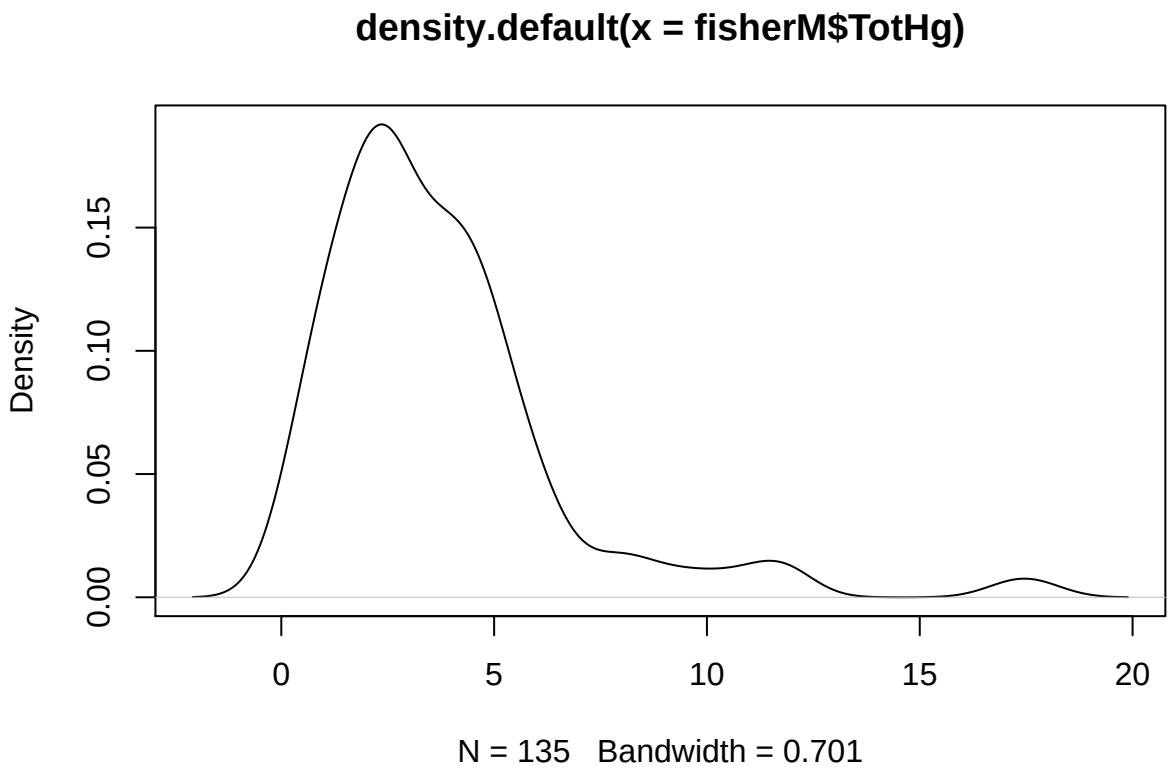


Performing a QQ plot analysis to check for normality shows a large deviation from a straight line at the ends of the plot, this shows that the distribution is non-normal

```
hist(fisherM$TotHg, 20, col="black")
```




```
d<-density(fisherM$TotHg)
plot(d)
```



The histogram and density plot do seem to resemble somewhat a normal distribution, but the long tail at the higher Mercury values shows that it is most likely not normally distributed.

```
shapiro.test(fisherM$TotHg)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fisherM$TotHg
## W = 0.81642, p-value = 1.047e-11
```

```
shapiro.test(fisherM$fishmlwk)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fisherM$fishmlwk
## W = 0.76431, p-value = 1.898e-13
```

```
shapiro.test(fisherM$restime)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fisherM$restime
## W = 0.77266, p-value = 3.45e-13
```

```r
shapiro.test(fisherM$weight)
```

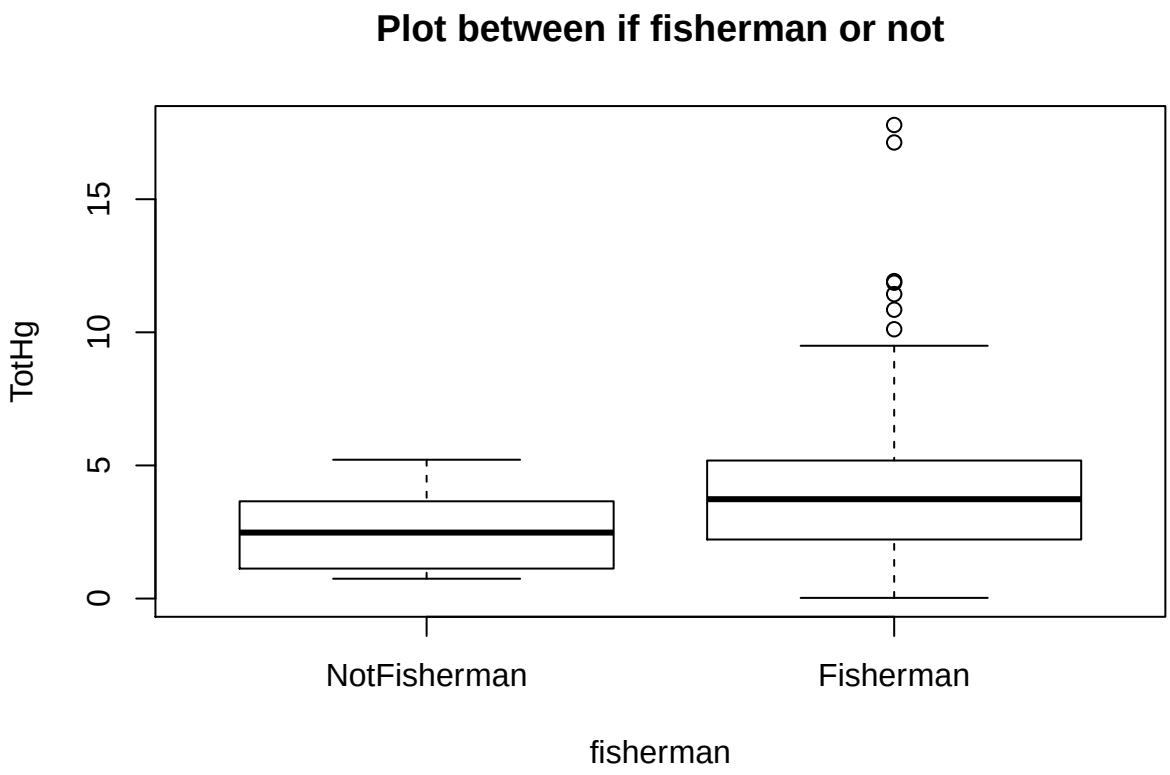
```
##
##  Shapiro-Wilk normality test
##
## data:  fisherM$weight
## W = 0.98449, p-value = 0.1295
```

The p values of most Independent Variables is less than 0.05 meaning that the null hypothesis of the data being normal can be rejected. Only the weight has a p value of >0.05 which shows that it is normally distributed, which intuitively makes sense.

Scatter plot

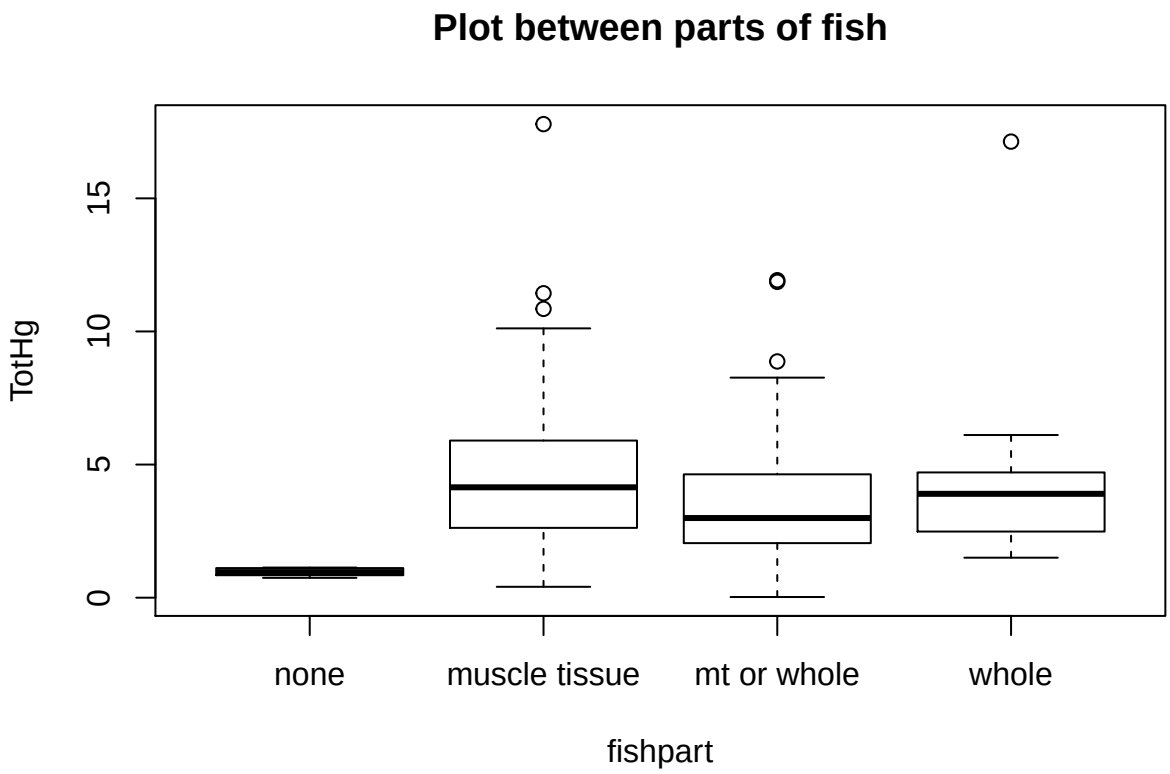
Scatter plots between dependent variable and the predictor variables

```r
# Scatterplots:
boxplot(TotHg~fisherman, data = fisherM,main="Plot between if fisherman or not")
```



The boxplot shows that pretty much all of the outliers of people with elevated mercury levels are indeed fishermen, leading to think that there is indeed some relation between having elevated levels and being a fisherman. Besides this it is worth noting that the standard deviation of the values for the fishermen is substantially larger than for the non fishermen. Important to take into account is that there is an imbalance of sample population as the amount of fishermen is 100 and non fishermen 35.

```
boxplot(TotHg~fishpart, data = fisherM,main="Plot between parts of fish")
```

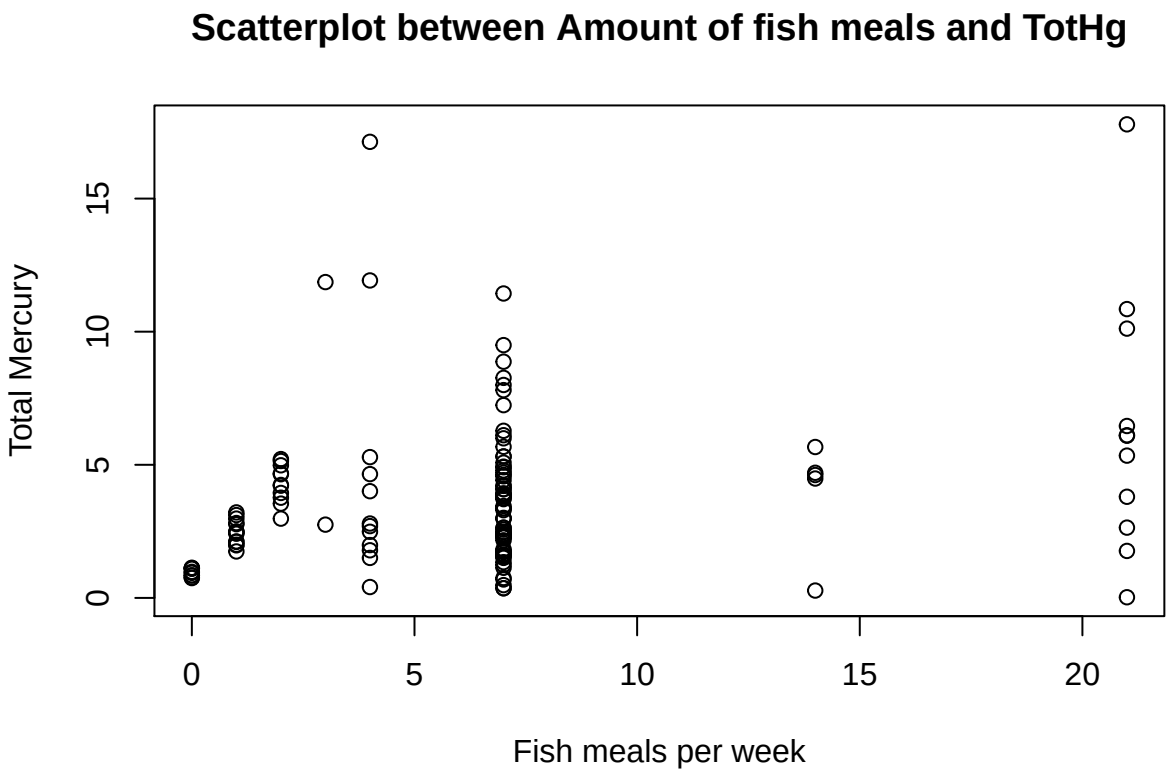


Additionally we checked whether eating a certain part of the fish would have an effect, even though we did not take this as an idependent variable in the model. Even though the values for muscle tissue look to have a larger standard deviation and mean, overall it does not seem to have much of a relation.

```
plot(TotHg~restime, xlab="Residence time in years", ylab = "Total Mercury", data = fisherM,main="Scatte
```

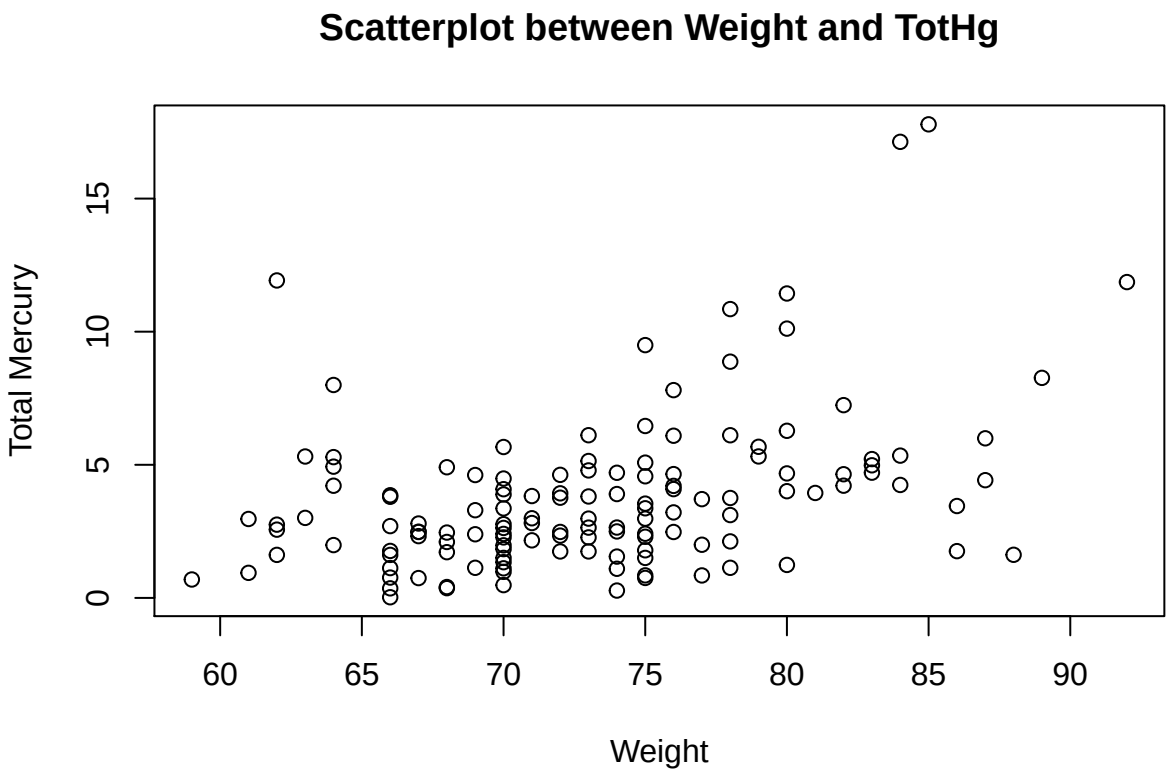
From the plot the residence time does not seem to have a relation on the mercury levels.

```
plot(TotHg~fishmlwk, xlab="Fish meals per week", ylab = "Total Mercury", data = fisherM,main="Scatterpl
```



The fishmeals consumed seems to have a linear relation, especially at the lower values, however the relation still looks fairly weak.

```
plot(TotHg~weight, xlab="Weight", ylab = "Total Mercury", data = fisherM,main="Scatterplot between Weig
```



The weight does look to have a relation with the total mercury in the persons hair, which makes sense as the amount of mercury would be more if the person has more mass overall.

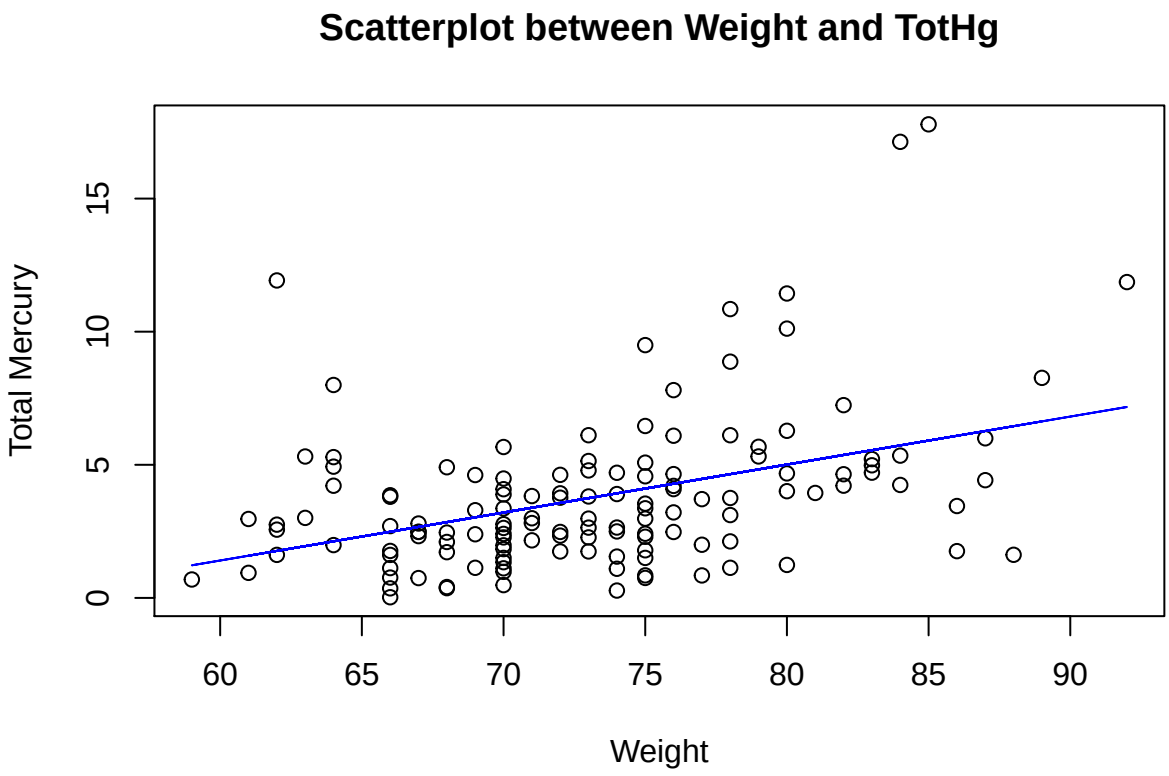
Linear regression

Conduct a multiple linear regression (including confidence intervals, and beta-values)

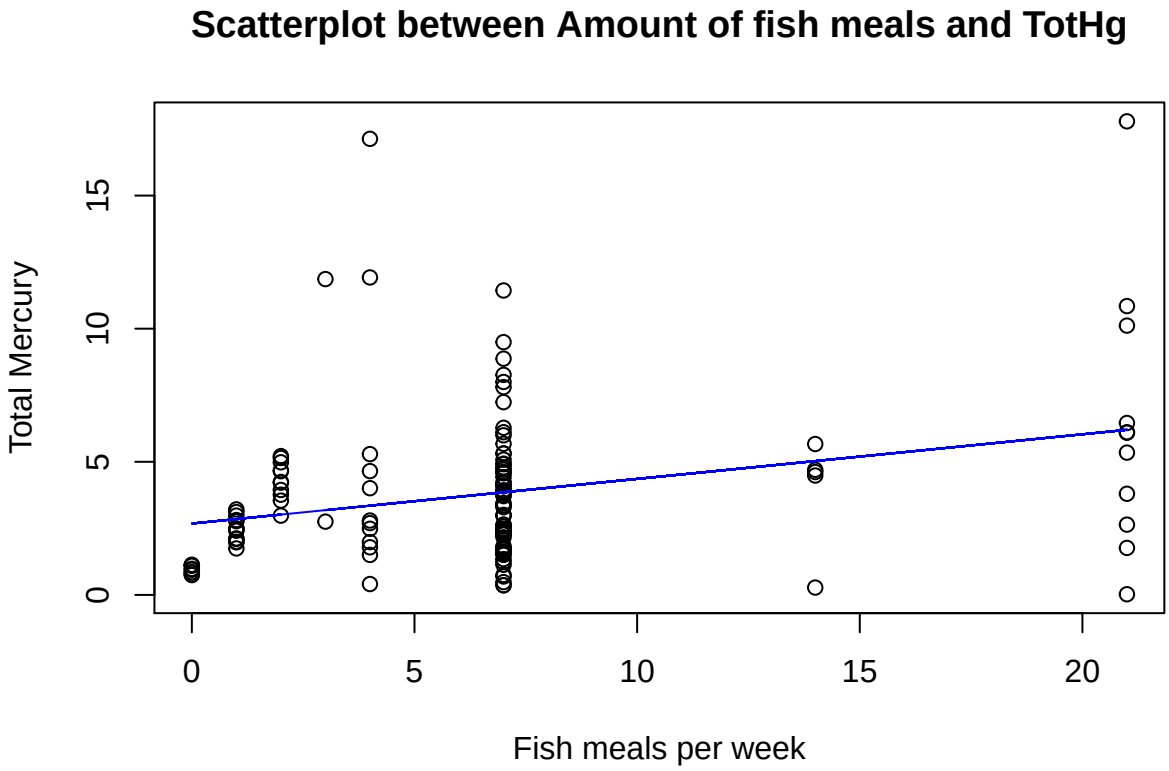
```
# Linear models:

model0 <-lm(TotHg~1 , data = fisherM, na.action = na.exclude)
model1 <-lm(TotHg~weight  , data = fisherM, na.action = na.exclude)
model2 <-lm(TotHg~fishmlwk , data = fisherM, na.action = na.exclude)
model3 <-lm(TotHg~restime , data = fisherM, na.action = na.exclude)
model4 <-lm(TotHg~weight+fishmlwk , data = fisherM, na.action = na.exclude)
model5 <-lm(TotHg~weight+restime , data = fisherM, na.action = na.exclude)
model6 <-lm(TotHg~fishmlwk+restime , data = fisherM, na.action = na.exclude)
model7 <-lm(TotHg~weight+fishmlwk+restime , data = fisherM, na.action = na.exclude)
modelfisherman <-lm(TotHg~fisherman+weight+fishmlwk+restime  , data = fisherM, na.action = na.exclude)

#scatterplots with fitted lines
plot(TotHg~weight, xlab="Weight", ylab = "Total Mercury", data = fisherM,main="Scatterplot between Weigh
lines(fisherM$weight, fitted(model1), col="blue")
```




```
plot(TotHg~fishmlwk, xlab="Fish meals per week", ylab = "Total Mercury", data = fisherM,main="Scatterpl
lines(fisherM$fishmlwk, fitted(model2), col="blue")
```




```
plot(TotHg~restime, xlab="Residence time in years", ylab = "Total Mercury", data = fisherM,main="Scatte
lines(fisherM$restime, fitted(model3), col="blue")
```

We plot the scatterplots with the corresponding trendline now. This further seems to confirm the earlier conclusions made about the relations of the data variables with the mercury levels. Only weight and fishmeals/week seem to have some linear relation.

```
pander(anova(model0, model1, model2, model3, model4, model5, model6, model7, modelfisherman))
```

Table 1: Analysis of Variance Table

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
134	1157	NA	NA	NA	NA
133	963.3	1	193.8	29.73	2.407e-07
133	1051	0	-87.65	NA	NA
133	1153	0	-101.7	NA	NA
132	868.1	1	284.5	43.65	9.15e-10
132	962.9	0	-94.78	NA	NA
132	1051	0	-88.04	NA	NA
131	866.7	1	184.2	28.26	4.475e-07
130	847.4	1	19.3	2.961	0.08768

Looking at the anova table only the models including the weight variable seem to have a significantly high F statistic value with the highest being model4 with weight and fishmlwk (43.65). For these models we can reject the null hypothesis meaning there is a relevant relation.

```
pander(summary(modelfisherman))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-11.23	2.52	-4.457	1.774e-05
fishermanFisherman	1.127	0.6548	1.721	0.08768
weight	0.1864	0.03371	5.529	1.699e-07
fishmlwk	0.1074	0.05306	2.023	0.04508
restime	-0.03508	0.04457	-0.7872	0.4326

Table 3: Fitting linear model: TotHg ~ fisherman + weight + fishmlwk + restime

Observations	Residual Std. Error	R^2	Adjusted R^2
135	2.553	0.2677	0.2451

From the table we can say that for the weight and fishmlwk variables the null hypothesis of there being no effect can be rejected, meaning that there is some relation between the amount of fish meals consumed and the mercury levels when considering the full model. Additionally as also assumed earlier, the weight has an effect. Being a fisherman looks to not have an effect, this is reasonable since most of the people with elevated levels were fishermen, but a majority of the fishermen also had no elevated levels, meaning that the two do not necessarily have any correlation.

```
pander(summary(model4))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-10.08	2.445	-4.122	6.604e-05
weight	0.1752	0.03322	5.273	5.337e-07
fishmlwk	0.1588	0.04175	3.805	0.0002161

Table 5: Fitting linear model: TotHg ~ weight + fishmlwk

Observations	Residual Std. Error	R^2	Adjusted R^2
135	2.564	0.2498	0.2384

```
pander(anova(model4))
```

Table 6: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
weight	1	193.8	193.8	29.47	2.633e-07
fishmlwk	1	95.21	95.21	14.48	0.0002161
Residuals	132	868.1	6.576	NA	NA

```
pander(lm.beta(model4))
```

weight	fishmlwk
0.3978	0.2871

Looking also at the impact of the individual predictors on the fit:

```
pander(anova(model0, model1))
```

Table 8: Analysis of Variance Table

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
134	1157	NA	NA	NA	NA
133	963.3	1	193.8	26.76	8.281e-07

```
pander(anova(model0, model2))
```

Table 9: Analysis of Variance Table

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
134	1157	NA	NA	NA	NA
133	1051	1	106.2	13.44	0.0003557

```
pander(summary(model7))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-10.12	2.454	-4.124	6.574e-05
weight	0.1767	0.03349	5.277	5.292e-07
fishmlwk	0.1625	0.04262	3.813	0.0002105
restime	-0.02028	0.04406	-0.4603	0.6461

Table 11: Fitting linear model: TotHg ~ weight + fishmlwk + restime

Observations	Residual Std. Error	R^2	Adjusted R^2
135	2.572	0.251	0.2338

```
pander(lm.beta(model7))
```

weight	fishmlwk	restime
0.4013	0.2937	-0.03562

```
pander(confint(model7))
```

	2.5 %	97.5 %
(Intercept)	-14.97	-5.265
weight	0.1105	0.243
fishmlwk	0.07819	0.2468
restime	-0.1074	0.06687

The confidence interval values can be interpreted that with a 2.5% chance the values will be lower than the given value for the variable and with 97.5% it will be lower than the value given in that column. meaning that with a 95% chance the value of the parameter will be between the given values.

The above results further strengthen the idea that mostly the weight and fishmeals has an impact. The residence time weights are negligible. The R and R squared values are fairly low, meaning that the data variance is spread quite wide around the fit line. This shows that even though there is a relation between the variables there is still a large variance for the individual cases around the fit.

Examine assumption

Examine assumptions underlying linear regression. E.g collinearity and analyses of the residuals, e.g. normal distributed, linearity assumption, homogeneity of variance assumption. Where possible support examination with visual inspection.

We want to test the assumptions underlying the regression model. We test this on the model with the variables that were relevant in the previous analysis so weight and fishmlwk. We want to look at the assumptions of: Multicolinearity of predictors: we want to see if the predictors within the model have correlations with each other. If there exist correlations between predictors this can give trouble with knowing the impact of individual predictors. we can check this by looking at the variance inflation factor(vif) or tolerance (1/vif). Autocorrelation between the errors: we want to check whether the errors are independet meaning there is no correlation. This can be checked with the Durbin Watson Test with the null hypothesis that there is no correlation among residuals. Visual inspection of residuals: We want to check that the residuals are normally distributed as well as seeing if there would be a pattern between the residuals and the fitted values. For the regression assumptions to hold the residuals should be normally distributed and the residuals vs fitted values should show no pattern.

```
# Assumptions:
1/vif(model4)
```

```
##    weight  fishmlwk
## 0.9984171 0.9984171
```

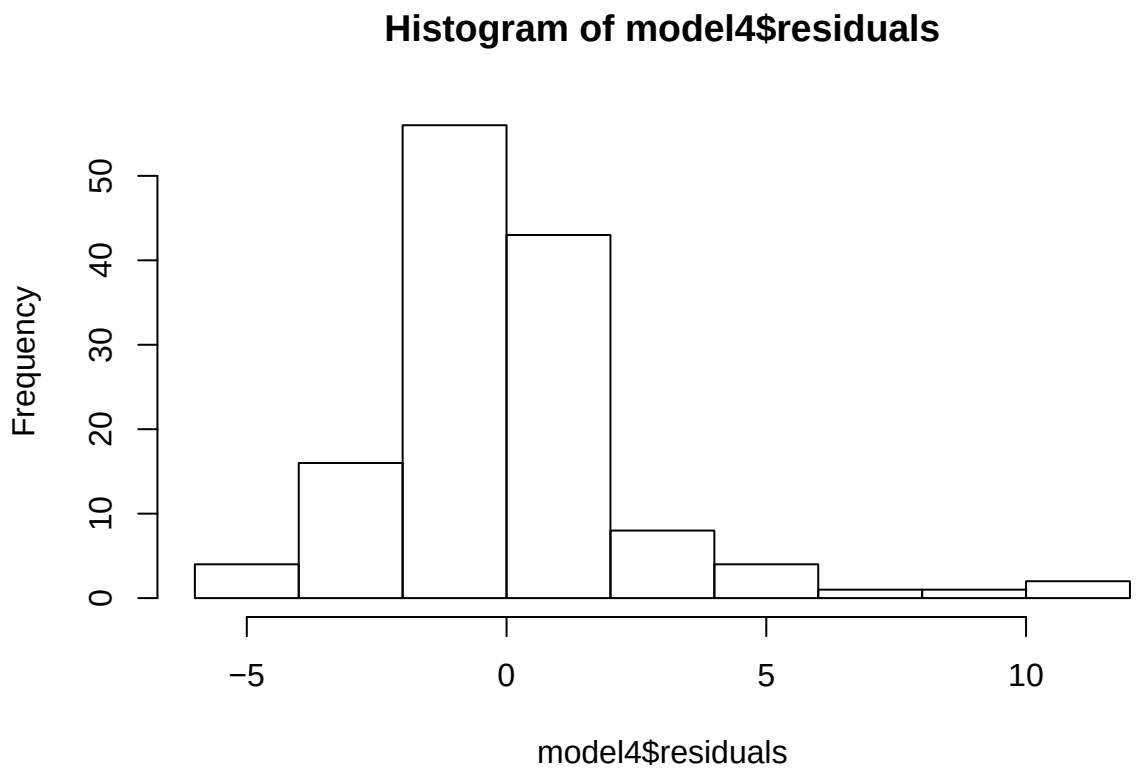
A tolerance value of > 0.2 like we see shows that there is no big multicolinearity problem.

```
durbinWatsonTest(model4)
```

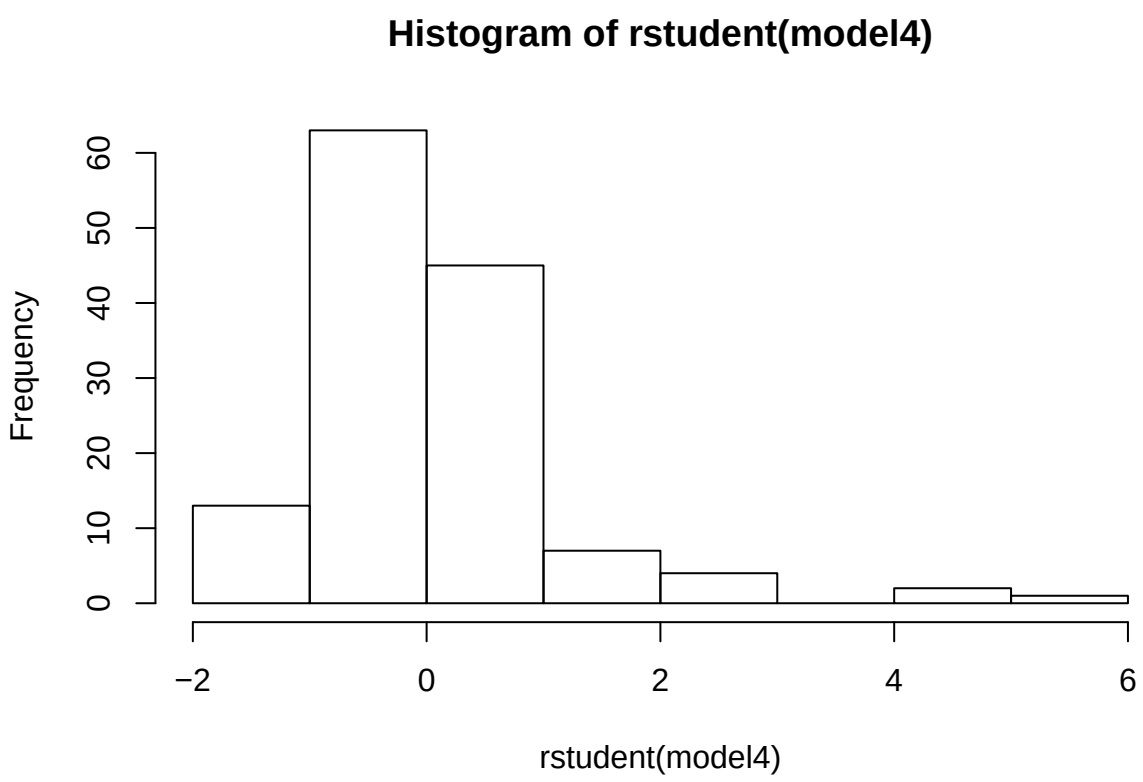
```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.2633075      1.472779   0.002
##  Alternative hypothesis: rho != 0
```

If we use a threshold of p value < 0.05 the DW test shows that the null hypothesis can be rejected, meaning there is some colinearity between predictors. The D-W statistic, however still shows not too big reason for concern. Good values would be around 1.5-2.5, in our case it is on the edge but still within a reasonable range.

```
hist(model4$residuals)
```




```
hist(rstudent(model4))
```

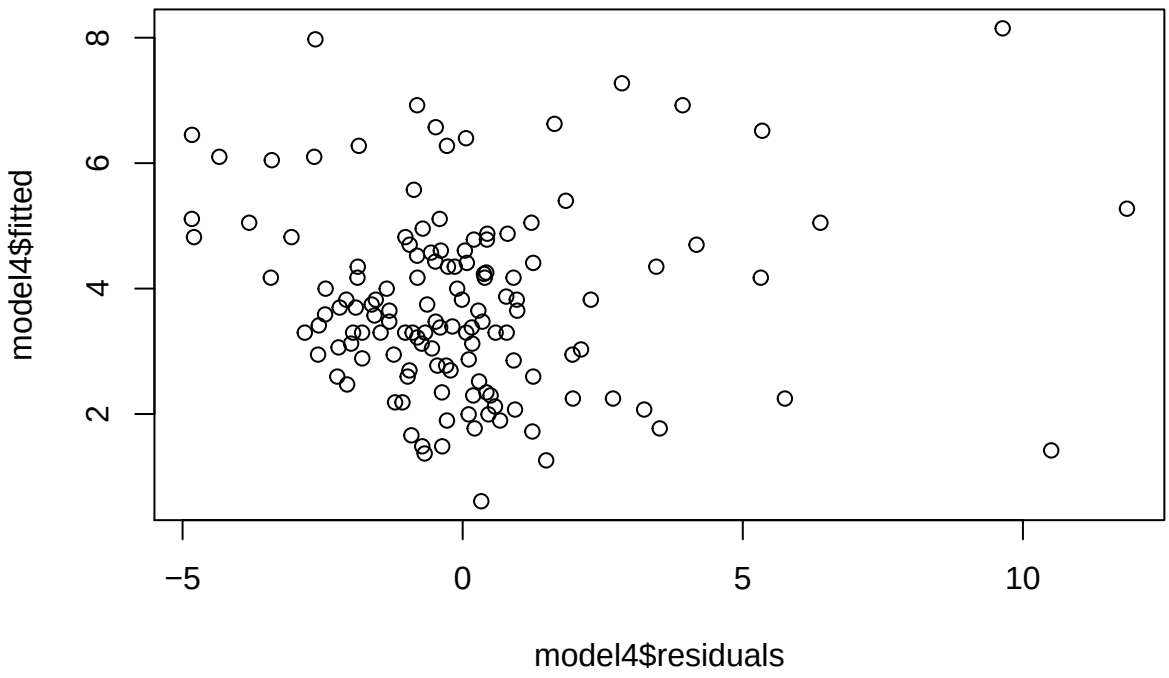



```
shapiro.test(model4$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model4$residuals
## W = 0.85395, p-value = 3.129e-10
```

Another assumption of the regression model is that the residuals are normally distributed. However, looking at the residual plots and the p value for the shapiro-wilk test we can say that the residuals are not normally distributed.

```
plot(model4$residuals, model4$fitted)
```



From the scatterplot it is clear that there is no pattern or relation between the residuals and fitted values.

```
fisherM$fitted.Hg <-fitted(model4)
pander(head(fisherM[, c("weight","fishmlwk","fitted.Hg" )], n=20L),
       caption = "Last 20 cases and their fitted values")
```

Table 14: Last 20 cases and their fitted values

weight	fishmlwk	fitted.Hg
70	14	4.41
73	7	3.823
66	7	2.597
80	7	5.05
78	21	6.923
75	21	6.397
85	21	8.149
68	7	2.947
80	21	7.273
75	7	4.174
76	21	6.573
66	21	4.821
66	21	4.821
87	7	6.276
68	7	2.947
69	14	4.234

weight	fishmlwk	fitted.Hg
70	7	3.298
72	7	3.648
70	7	3.298
70	14	4.41

Impact analysis of individual cases

Examine effect of single cases on the predicted values (e.g. DFBeta, Cook's distance)

We want to look at the impact of extreme values on the fitted values. For this we want to look at the hatvalues and see if the values have a high leverage. We take the average leverage as having to be around (#predictors + 1)/#observations. which in our case would be 2+1/135= 0.022. we want to investigate those cases that are 3 times the average leverage.

```
# Individual Cases:
#residuals
fisherM$stud.res<-rstudent(model4)
fisherM$dfbeta<-dfbeta(model4)
fisherM$leverage<-hatvalues(model4)
fisherM$cooks<-cooks.distance(model4)


troublingdata<-subset(fisherM, (leverage > 3*.022) | (abs(stud.res) > 2),
                      select = c("leverage", "stud.res", "dfbeta"))

head(troublingdata) # i did clean for the data with high leverage and it increased the adjusted R squar
```

```
##       leverage   stud.res dfbeta.(Intercept) dfbeta.weight dfbeta.fishmlwk
## 4  0.015274911  2.5615251      -0.4978893282  0.0074230331    0.0004422281
## 7  0.083591877  4.1619796      -1.6146691874  0.0196312178    0.0392966820
## 9  0.069123212  1.1503119      -0.2821631419  0.0031370379    0.0115373002
## 10 0.008023256  2.1104906      -0.0840397915  0.0016392389    0.0005908845
## 12 0.073254654 -0.4125989      -0.0868003314  0.0014580307   -0.0042951669
## 13 0.073254654 -1.9635298      -0.4073935989  0.0068432039   -0.0201591798
```

Report section for a scientific publication

Checking the means and standard deviations:

```
mean(fisherM$weight)
```

```
## [1] 73.15556
```

```
sd(fisherM$weight)
```

```
## [1] 6.673479
```

```r
mean(fisherM$fishmlwk)
```

```
## [1] 6.525926
```

```r
sd(fisherM$fishmlwk)
```

```
## [1] 5.310971
```

```r
mean(fisherM$TotHg)
```

```
## [1] 3.775304
```

```r
sd(fisherM$TotHg)
```

```
## [1] 2.938538
```

Taking the results of the model comparison and the analyses performed on these models we can report the results as follows. A linear model was fitted based on the amount of Total mercury in a hair of a Kuwaiti person. Taking being a fisherman, residence time, fishmeals in a week and weight as independent variables. The analysis found a significant main effect for the model including the weight and the amount of fishmeals per week ($F(1,132) = 43.65$, $p = 9.15e\text{-}10$). The relevant predictors that have shown to have an impact are weight($F(1,133)=26.76$, $p= 8.281e\text{-}07$) and fishmeals/week($F(1,133)=13.44$, $p= 0.0003557$). where the mean and standard deviation for the relevant independent variables and the dependent variable are weight(mean:73.15556, std:6.673479) fishmeals/week(mean:6.525926, std:5.310971) Total mercury(mean:3.775304, std:2.938538). For the model including the fisherman variable it found no significant main effect ($F(1,132) = 2.961$, $p = 0.08768$). An analysis of the assumptions made for the linear regression model does show that the model fails to uphold some assumptions such as normality of data and normality of residuals.