STATS 201 Experimental Class 2

runze liao, 222020321102007

Haemoglobin levels in athletes

Haeomoglobin concentration in the blood is a measure of how efficiently athletes can deliver oxygen to muscles during exercise. The aim of the study below was to generate reference values of haemoglobin concentration for athletes of different body sizes, against which blood samples from future athletes could be compared to inform their training programmes. The study measured haemoglobin concentration from 113 randomly selected Australian athletes, all of whom were performing at national level in their respective sports, and various body-size measurements for predictor variables.

Each row in athletes.csv corresponds to an athlete. The variables are • Hconc: Haemoglobin concentration in blood (grams per decalitre).

- Sex: Sex, either M for male or F for female.
- Height: Height (cm).
- Weight: Weight (kg).
- LBM: Body mass other than fat (kg).

Conduct a full analysis, and include Methods and Assumption Checks along with an Executive Summary. In particular, we are interested in addressing a few questions of interest in the Executive Summary:

- What are the predicted haemoglobin concentration levels for a male and a female athlete, both with height 170 cm, weight 70kg, and lean body mass 60 kg?
- Is the relationship between haemoglobin concentration and height different for males than it is for females?

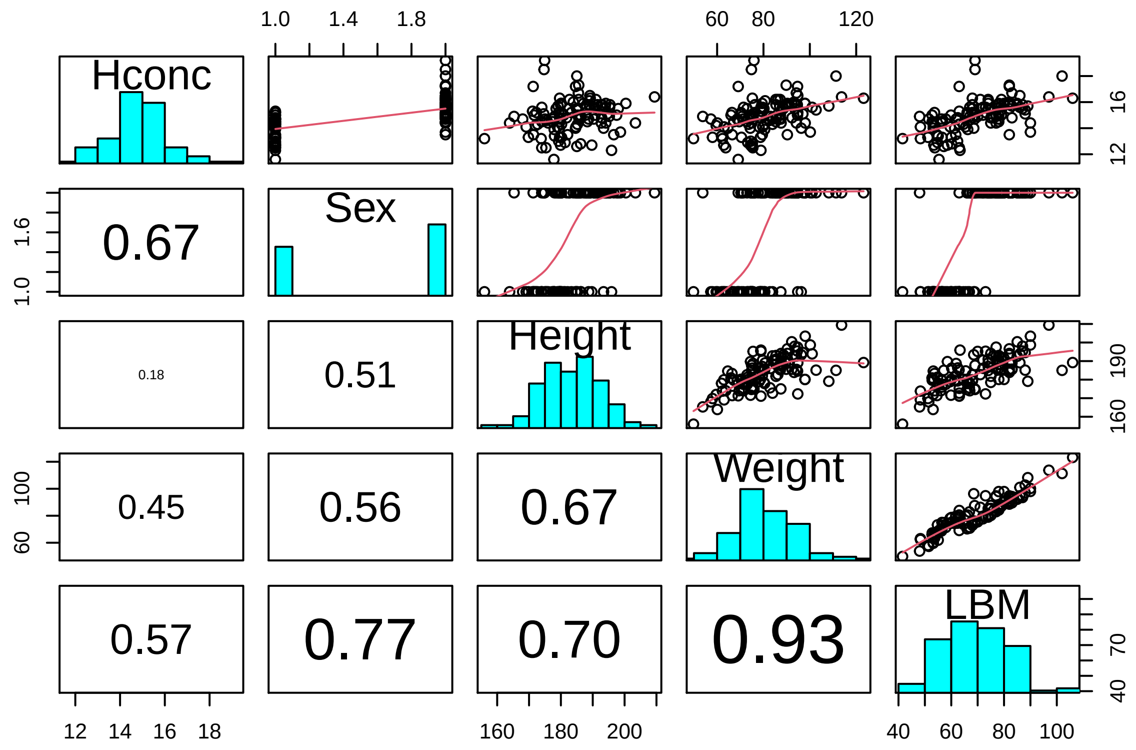
Code and output

```
library(s20x)
```

```
## Warning: package 's20x' was built under R version 4.0.5
```

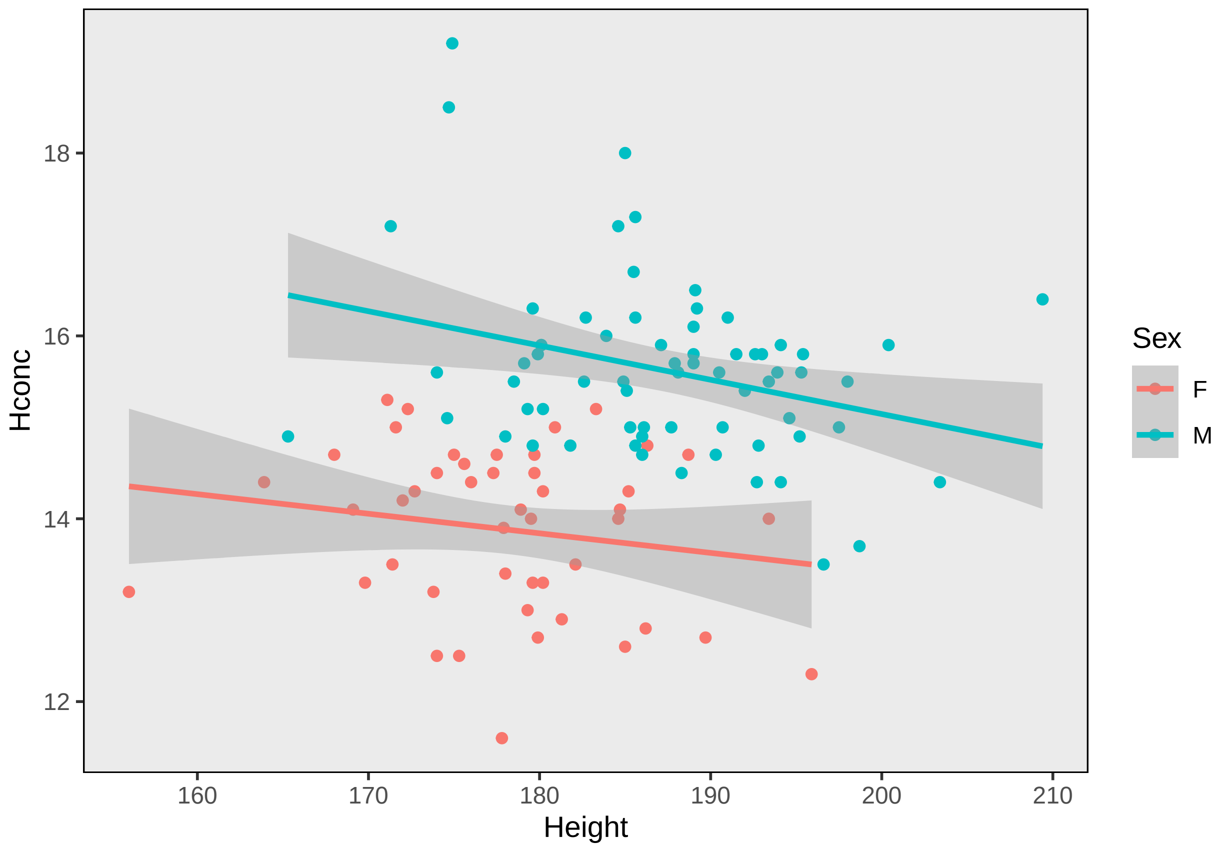
```
require(s20x)
## Loading in the data.
athletes.df = read.csv(file = "athletes.csv", header = TRUE)
athletes.df$Sex = factor(athletes.df$Sex)
## Plot the data.
## INSERT CODE HERE.
## Analyse the data.
## INSERT CODE HERE.
```

```
pairs20x(athletes.df[,c(1,2,3,4,5)])#the relation variable graph
```



```
#by using the ggplot to see the relationship
library(ggplot2)
ggplot(data=athletes.df, aes(x=Height, y=Hconc, group=Sex, color=Sex)) +
labs(x="Height", y="Hconc") +
geom_point() +
geom_smooth(method="lm") +
theme(panel.grid=element_blank(), panel.background=element_rect(color='black')) +
guides(color=guide_legend(title="Sex"))
```

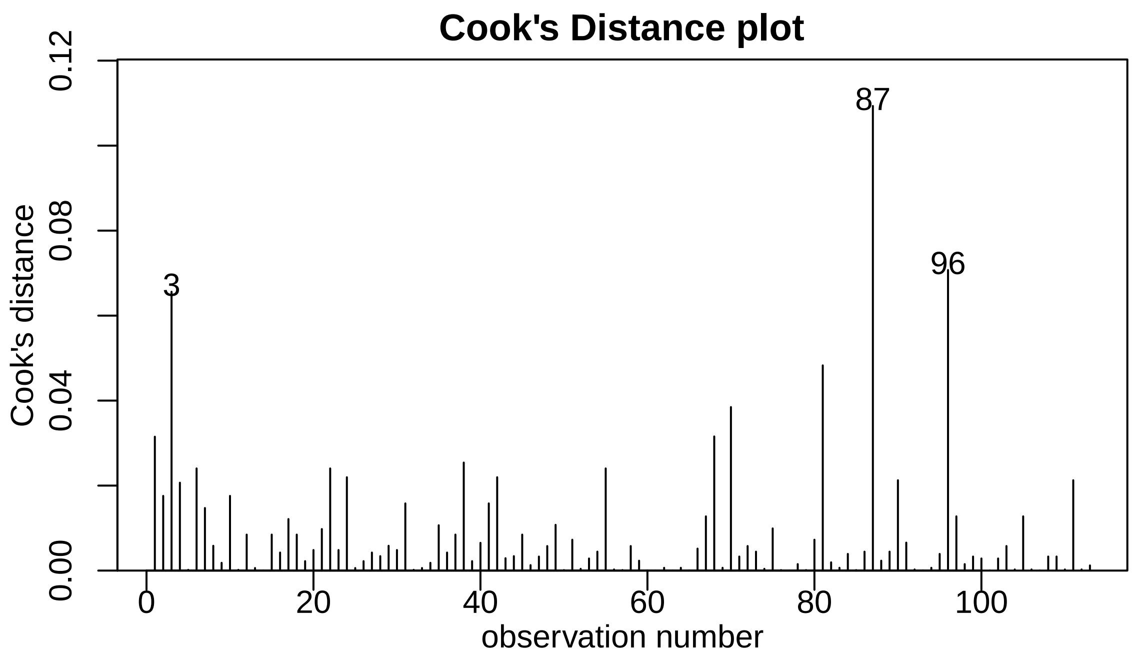
```
## `geom_smooth()` using formula 'y ~ x'
```



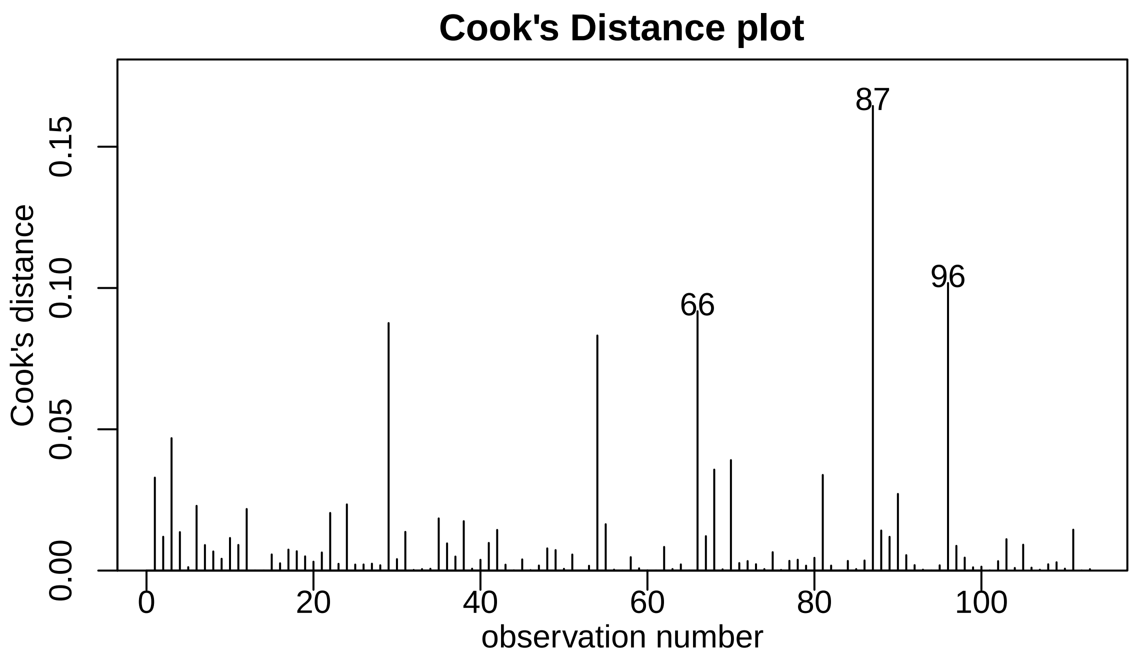
We can see that the the lines are parallel, so we consider there's no interaction between the Height and the Sex.

From the pairs graph it seems that the males have a higher mean Haemoglobin concentration in blood than female. The higher the height, the higher the Haemoglobin(a little bit fluctuation.), As Weight increases, the expected Hconc increases. There's a weak positive relationship between Weight and Hconc, howerver, not too much relationship between LBM and Hconc.

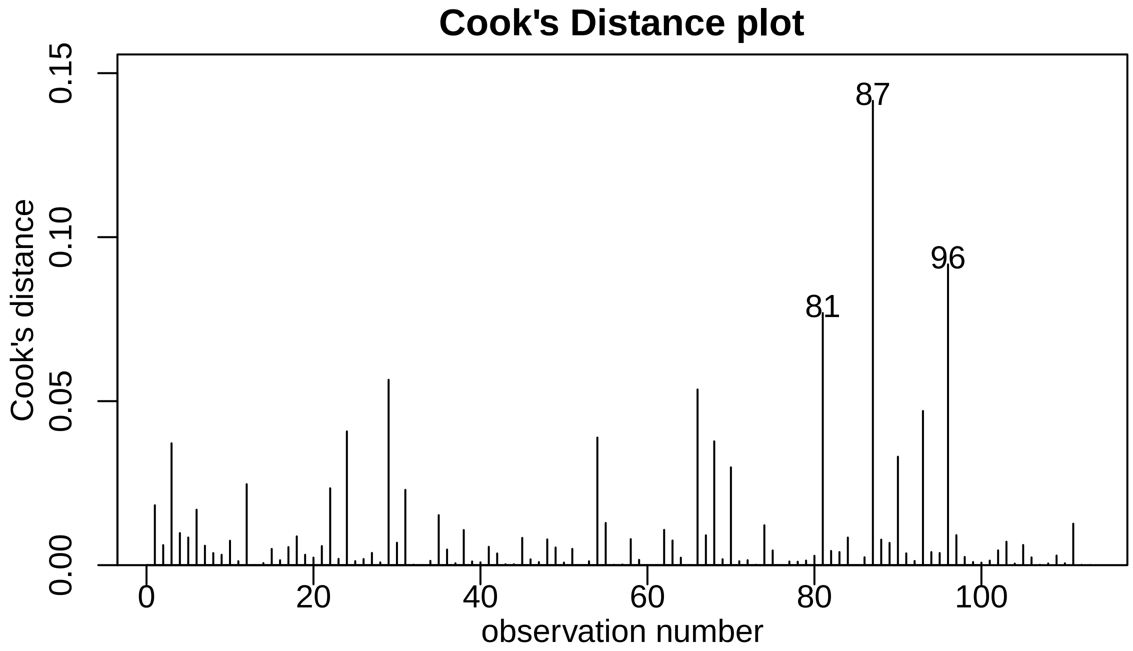
```
athletes.df $Sex = factor(athletes.df $ Sex)
athletes.fit = lm(Hconc~Sex,data = athletes.df)
cooks20x(athletes.fit)
```



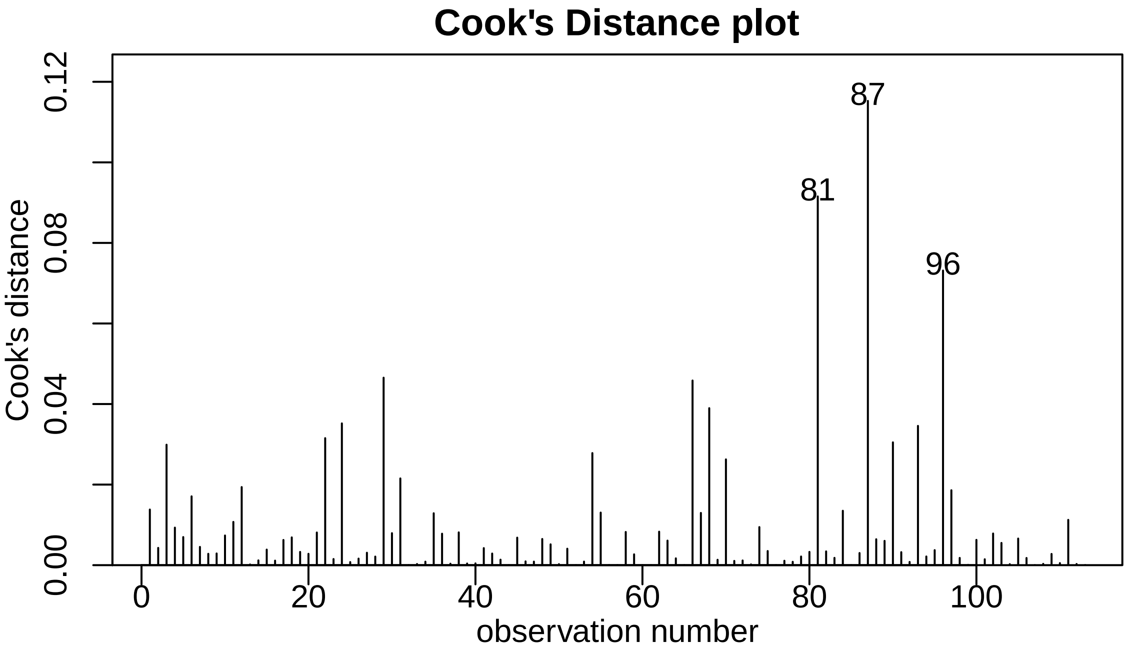
```
athletes.fit2 = lm(Hconc~Sex + Height,data = athletes.df)
cooks20x(athletes.fit2)
```



```
athletes.fit3 = lm(Hconc~Sex + Height + Weight,data = athletes.df)
cooks20x(athletes.fit3)
```

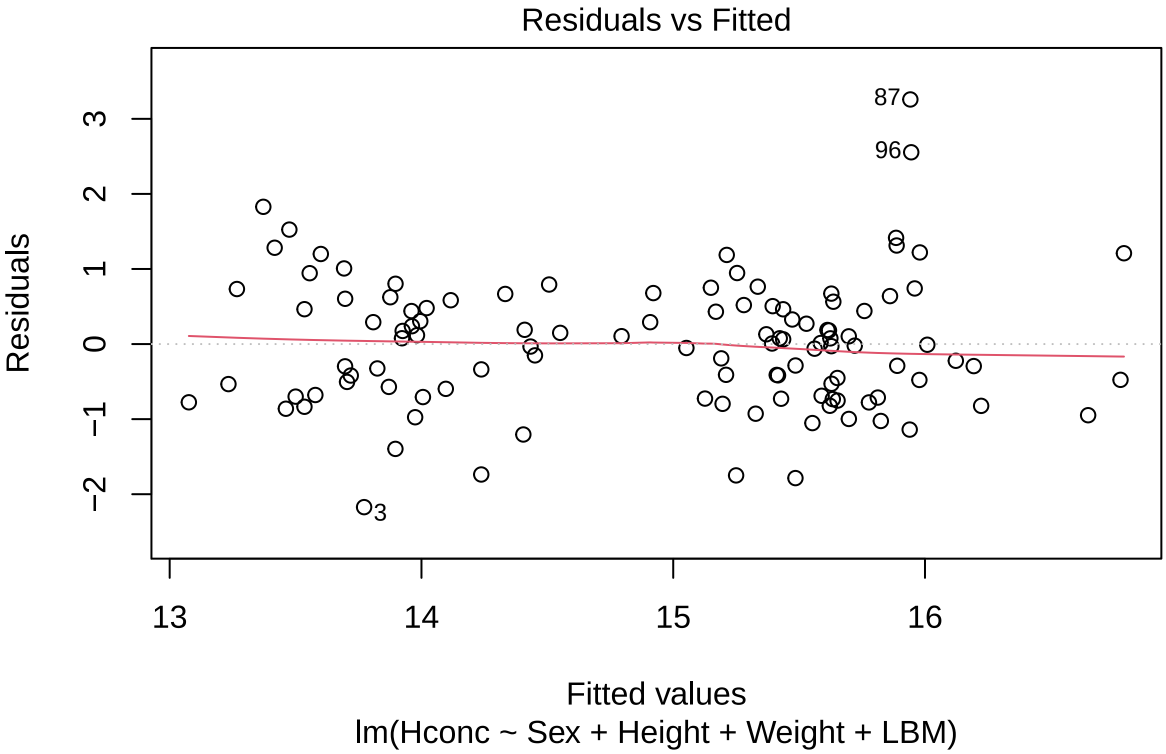


```
athletes.fit4 = lm(Hconc~Sex + Height + Weight + LBM,data = athletes.df)
cooks20x(athletes.fit4)
```



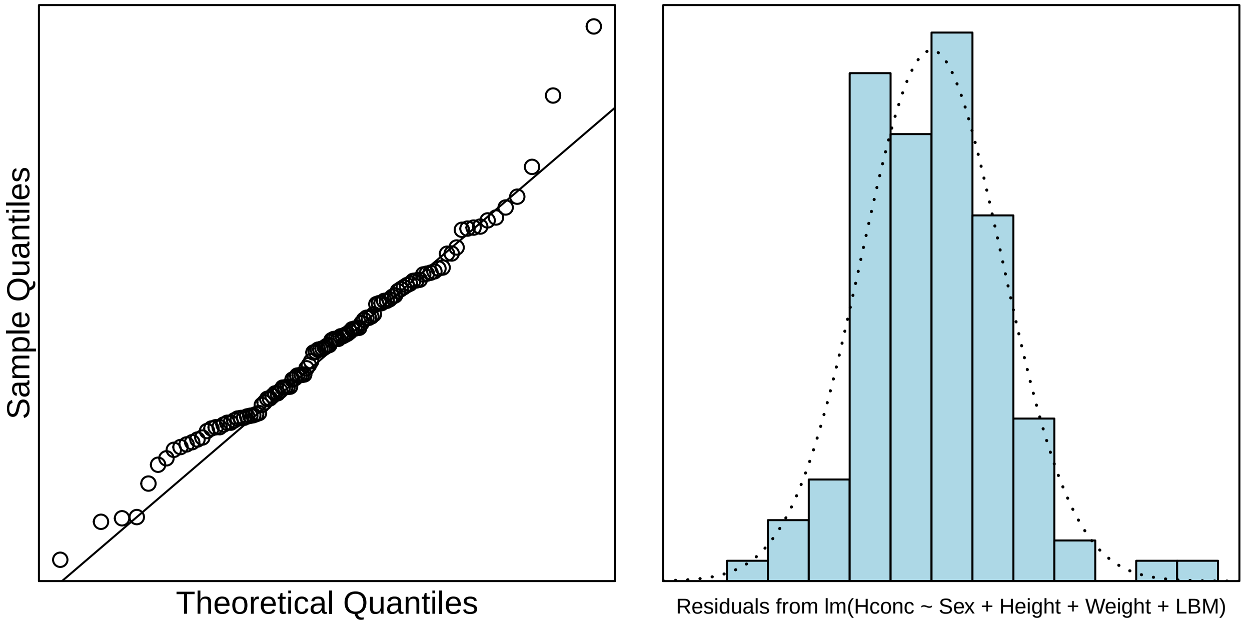
Next we will fit the model

```
athletes.fit = lm(Hconc~Sex+Height+Weight+LBM, data = athletes.df)
plot(athletes.fit,which =1)
```

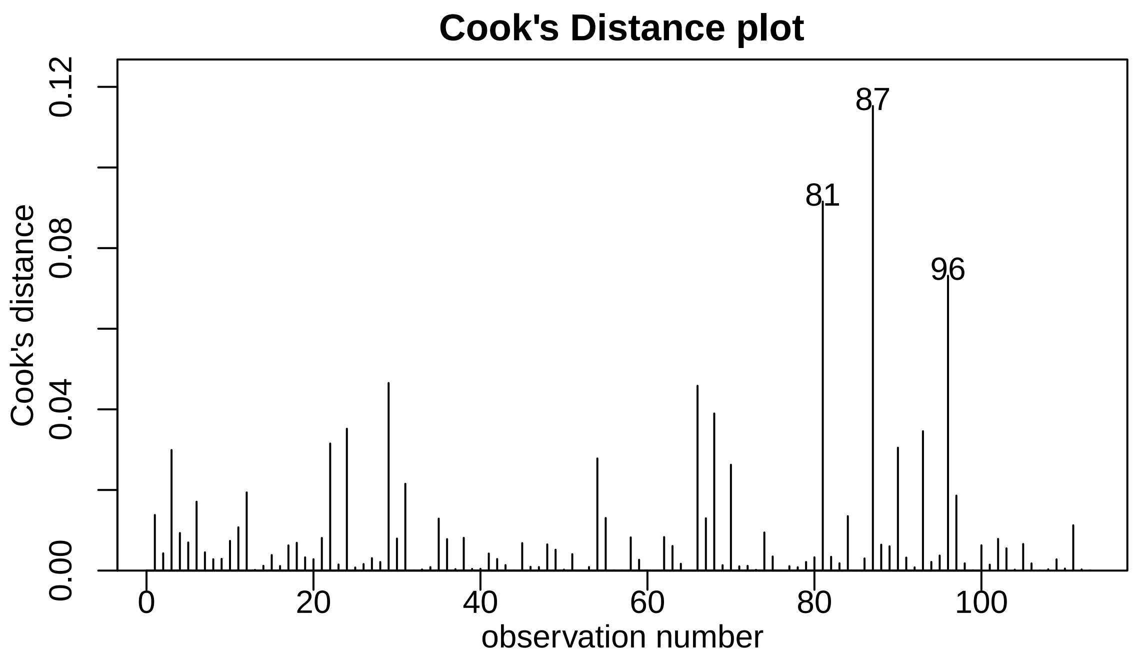


The EOV and no trend assumption seem to be okay.

```
normcheck(athletes.fit)
```



```
cooks20x(athletes.fit)
```



Seems Normal Distribution, and no strong influential points. We can trust our linear model.

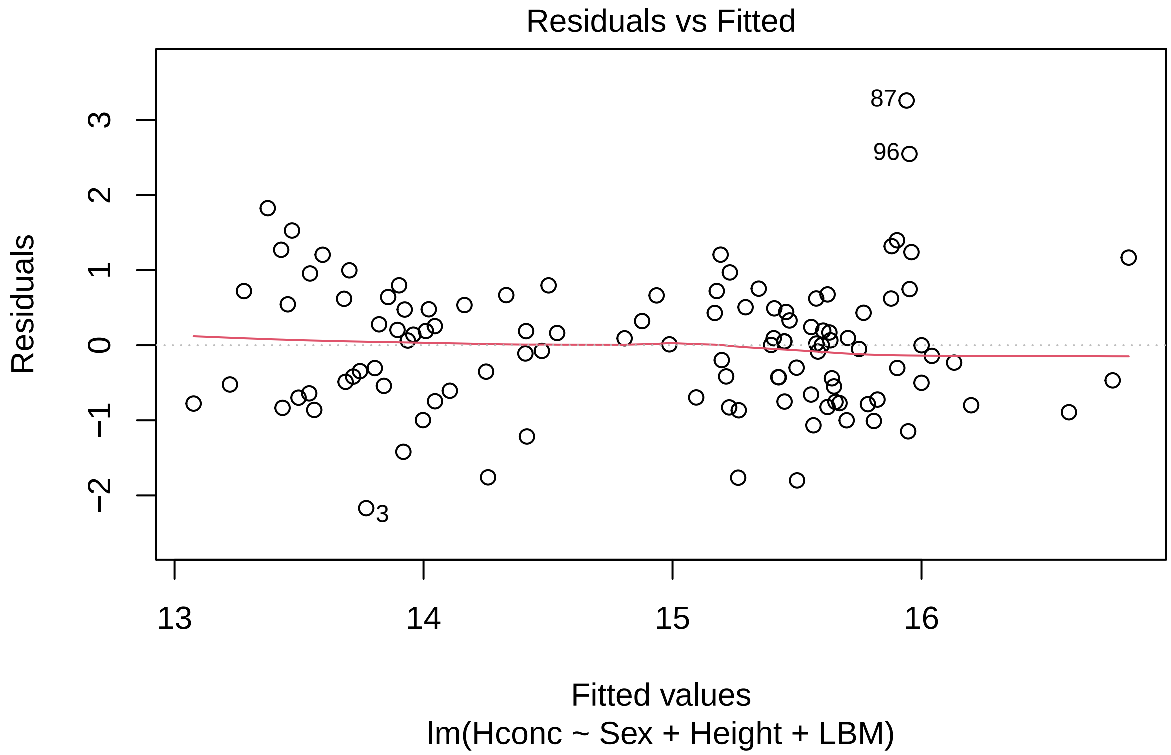
```
summary(athletes.fit)
```

```
##
## Call:
## lm(formula = Hconc ~ Sex + Height + Weight + LBM, data = athletes.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1722 -0.6786  0.0074  0.5195  3.2581
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.623636   2.018052  10.715  < 2e-16 ***
## SexM         1.465868   0.353174   4.151 6.64e-05 ***
## Height      -0.058106   0.013072  -4.445 2.14e-05 ***
## Weight       0.007407   0.023375   0.317    0.752
## LBM          0.035549   0.031677   1.122    0.264
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.878 on 108 degrees of freedom
## Multiple R-squared:  0.5443, Adjusted R-squared:  0.5274
## F-statistic: 32.25 on 4 and 108 DF,  p-value: < 2.2e-16
```

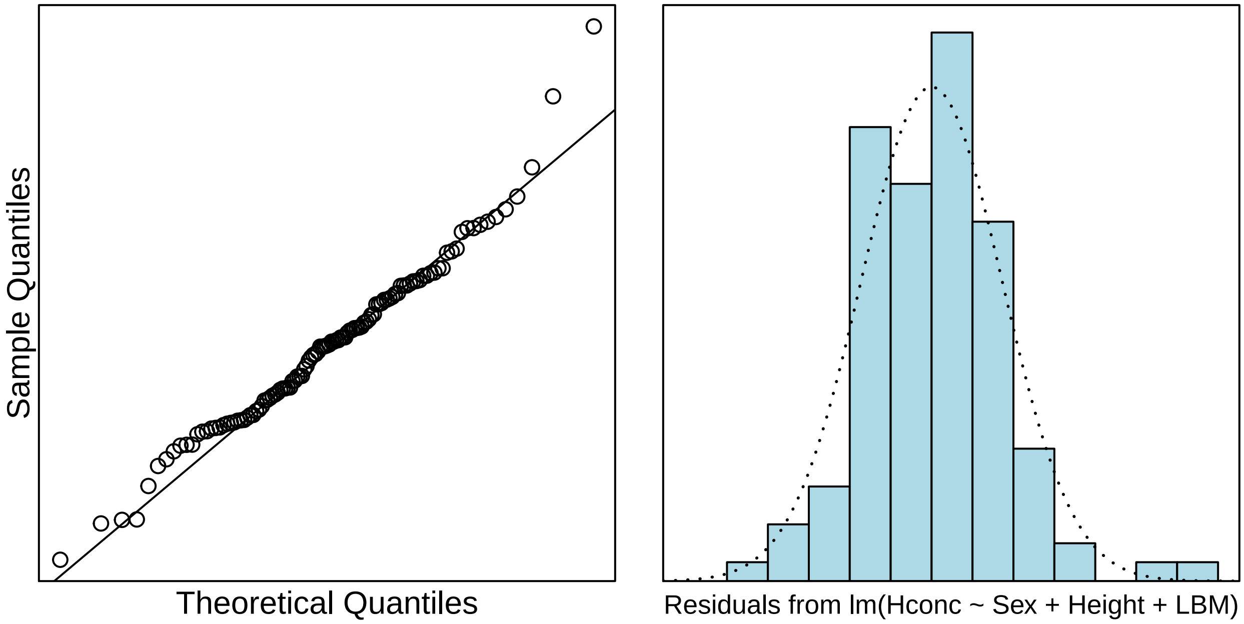
Worthy attention, the LBM and Height have already explained the Weight, so we remove the Weight, We can see that so many the P-value of Weight and LBM is > 0.05, According to the Occam Razor Principle, we only have the Sex, Height, LBM. So We fit the latest model.

```
athletes.fit5 = lm(Hconc~Sex+Height+LBM, data = athletes.df)
```

```
plot(athletes.fit5,which =1)
```

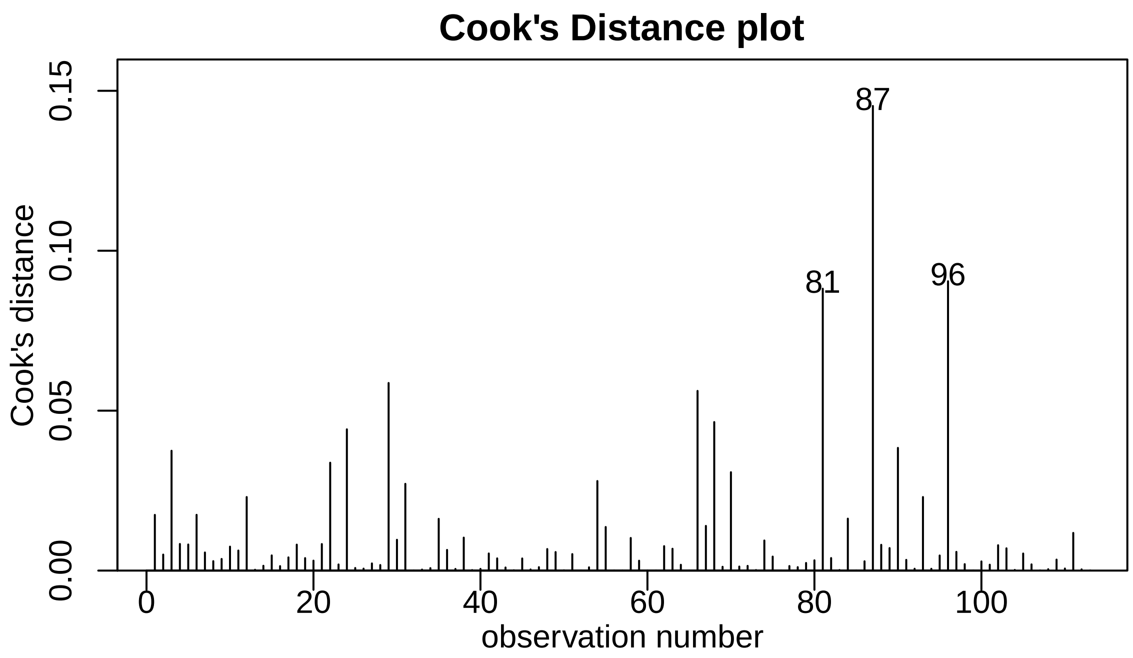


```
normcheck(athletes.fit5)
```



The EOV check is good.

```
cooks20x(athletes.fit5)
```



```
summary(athletes.fit5)
```

```
##
## Call:
## lm(formula = Hconc ~ Sex + Height + LBM, data = athletes.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1697 -0.6408  0.0125  0.5356  3.2600
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.60260    2.00862  10.755  < 2e-16 ***
## SexM         1.39123    0.26205   5.309 5.87e-07 ***
## Height      -0.05800    0.01301  -4.457 2.03e-05 ***
## LBM          0.04479    0.01235   3.626 0.000439 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8744 on 109 degrees of freedom
## Multiple R-squared:  0.5438, Adjusted R-squared:  0.5313
## F-statistic: 43.32 on 3 and 109 DF,  p-value: < 2.2e-16
```

The point 87 might seem stange, but we will keep it rather than delete it. No other strong influence point, everything is fine.

```
confint(athletes.fit5)
```

```
##                   2.5 %      97.5 %
## (Intercept) 17.62157912 25.58361470
## SexM         0.87184597  1.91061291
## Height      -0.08379183 -0.03220578
## LBM          0.02030639  0.06926445
```

Next we predict some situation:

```
pred.df = data.frame(Sex = c("F","M"), Height = c(170,170), LBM = c(60,60))
predict(athletes.fit5, pred.df, interval = "prediction")
```

```
##        fit      lwr      upr
## 1 14.42992 12.66232 16.19753
## 2 15.82115 14.02687 17.61544
```

Methods and Assumption Checks

Looking at the pairs plot, we saw that the : Haemoglobin concentration in blood was related to a number of our explanatory variables. So we will construct a multiple linear regression model with a suitable selection of the explanatory variables.

We decided to include the Sex and Height and LBM as the explanatory variable, but had to remove the weight as an explanatory due to multicollinearity. All model assumptions were satisfied by our final model.

Our final model is:

$$Hconc_i = \beta_0 + \beta_1 * Sex_i + \beta_2 * Height_i + \beta_3 * LBM_i + \epsilon_i$$

where $\epsilon_i \sim iid.N(0, \sigma^2)$. Here our indicator variable takes value 1 if the Sex is Male.

Our model explains about 31% of the variability in Haemoglobin concentration in blood.

Executive Summary

We wanted to have a model to explain the Haemoglobin concentration in blood.

Keeping all other variables constant:

- We estimate that for each additional centimetre in athletes' height, the Haemoglobin concentration in blood decreased by -0.032 to -0.084
- We estimate that for each one more LBM in athletes, the Haemoglobin concentration in blood increased by 0.02 to 0.07.
- We estimate that the Haemoglobin concentration in blood, the male is averagely more 1.39 than the female.

What are the predicted haemoglobin concentration levels for a male and a female athlete, both with height 170 cm, weight 70kg, and lean body mass 60 kg?

For a female and a male athlete, both with height 170 cm, weight 70kg, and lean body mass 60 kg, we predict the Haemoglobin concentration in blood(individual) between 12.7 to 16.2 and 14.0 to 17.6, respectively, these interval were very wide due to the high variability between athletes.

Is the relationship between haemoglobin concentration and height different for males than it is for females?
Nope, the graph above shows that the haemoglobin concentration and height for males and females, they have no interactions and For the same Height and LBM, the Male's Haemoglobin concentration is larger than 0.87 to 1.91 than the Female.