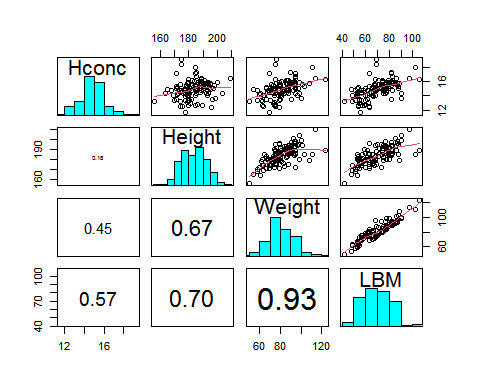
STATS 201 Lab Class 2

LiYuzhuo-2019210146

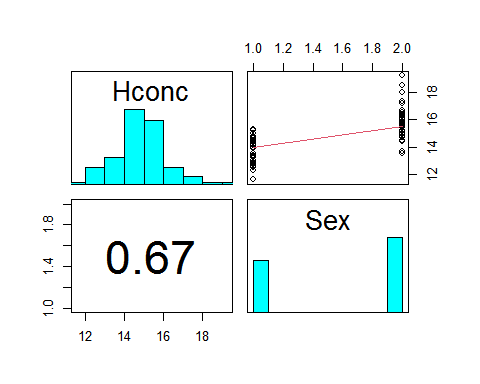
#Haemoglobin levels in athletes

# Code and output

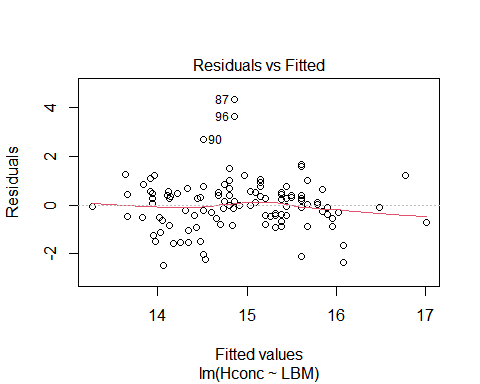
library(s20x)  
## Loading in the data.  
athletes.df = read.csv(file = "athletes.csv", header = TRUE)  
athletes.df$Sex = factor(athletes.df$Sex)  
## Plot the data.  
pairs20x(athletes.df[,c(1, 3, 4, 5)])



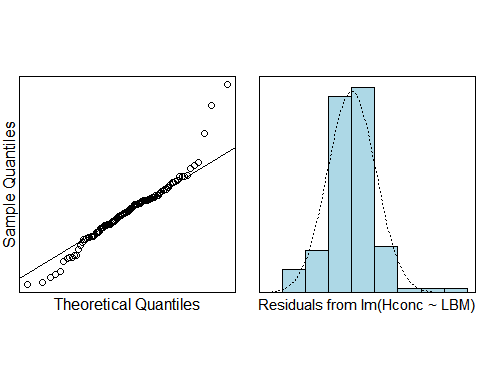
pairs20x(athletes.df[,c(1, 2)])



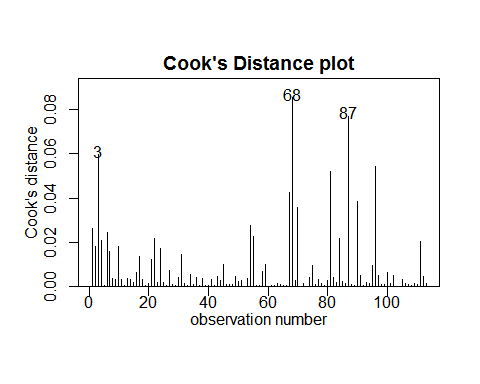
hconc.fit = lm(Hconc ~ LBM, data = athletes.df)  
plot(hconc.fit, which = 1)



normcheck(hconc.fit)



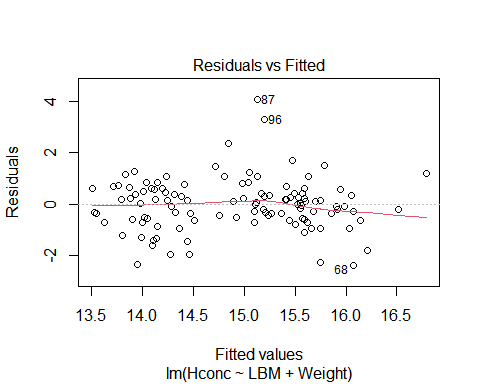
cooks20x(hconc.fit)



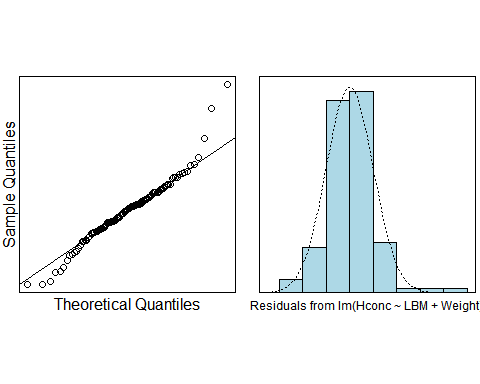
summary(hconc.fit)

##   
## Call:  
## lm(formula = Hconc ~ LBM, data = athletes.df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.4689 -0.5539 0.0127 0.5084 4.3403   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 10.859468 0.558884 19.43 < 2e-16 \*\*\*  
## LBM 0.057974 0.007867 7.37 3.25e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.051 on 111 degrees of freedom  
## Multiple R-squared: 0.3285, Adjusted R-squared: 0.3225   
## F-statistic: 54.31 on 1 and 111 DF, p-value: 3.251e-11

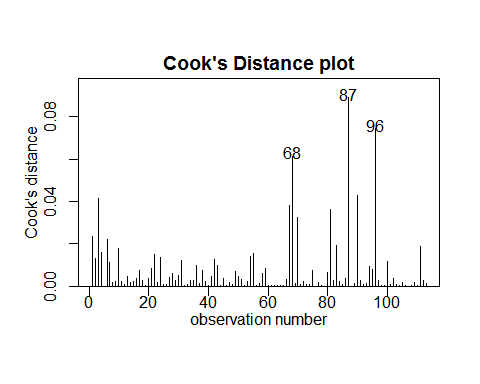
hconc.fit2 = lm(Hconc ~ LBM + Weight, data = athletes.df)  
plot(hconc.fit2, which = 1)



normcheck(hconc.fit2)



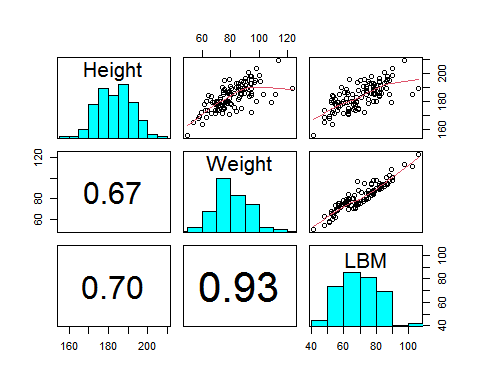
cooks20x(hconc.fit2)



summary(hconc.fit2)

##   
## Call:  
## lm(formula = Hconc ~ LBM + Weight, data = athletes.df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.3732 -0.5698 0.0462 0.5593 4.0749   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 11.76381 0.61154 19.236 < 2e-16 \*\*\*  
## LBM 0.11736 0.02052 5.719 9.31e-08 \*\*\*  
## Weight -0.06241 0.02004 -3.114 0.00235 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.012 on 110 degrees of freedom  
## Multiple R-squared: 0.3829, Adjusted R-squared: 0.3717   
## F-statistic: 34.13 on 2 and 110 DF, p-value: 2.936e-12

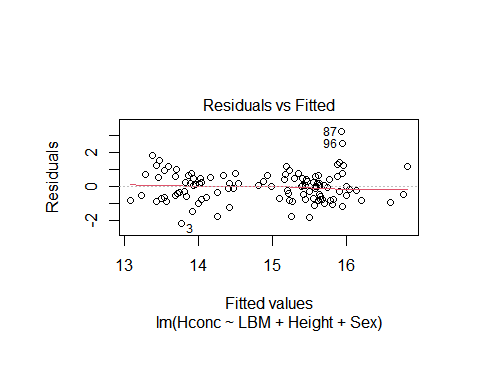
pairs20x(athletes.df[,c(3, 4, 5)])



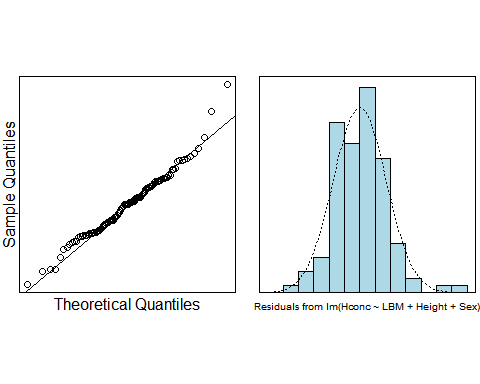
hconc.fit3 = lm(Hconc ~ LBM + Height, data = athletes.df)  
summary(hconc.fit3)

##   
## Call:  
## lm(formula = Hconc ~ LBM + Height, data = athletes.df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.3798 -0.6738 0.0488 0.5747 3.8270   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 20.19719 2.22356 9.083 4.91e-15 \*\*\*  
## LBM 0.08883 0.01022 8.692 3.80e-14 \*\*\*  
## Height -0.06263 0.01450 -4.319 3.45e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9765 on 110 degrees of freedom  
## Multiple R-squared: 0.4259, Adjusted R-squared: 0.4155   
## F-statistic: 40.8 on 2 and 110 DF, p-value: 5.555e-14

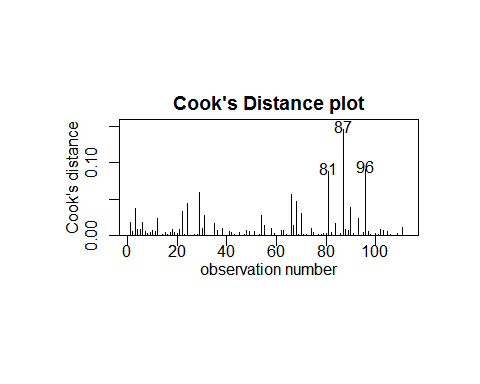
hconc.fit4 = lm(Hconc ~ LBM + Height + Sex, data = athletes.df)  
plot(hconc.fit4, which = 1)



normcheck(hconc.fit4)



cooks20x(hconc.fit4)



summary(hconc.fit4)

##   
## Call:  
## lm(formula = Hconc ~ LBM + Height + Sex, 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 \*\*\*  
## LBM 0.04479 0.01235 3.626 0.000439 \*\*\*  
## Height -0.05800 0.01301 -4.457 2.03e-05 \*\*\*  
## SexM 1.39123 0.26205 5.309 5.87e-07 \*\*\*  
## ---  
## 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

confint(hconc.fit4)

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

predict(hconc.fit4, data.frame("Sex"="M", "Height"=170, "Weight"=70, "LBM"=60), interval = "prediction")

## fit lwr upr  
## 1 15.82115 14.02687 17.61544

predict(hconc.fit4, data.frame("Sex"="F", "Height"=170, "Weight"=70, "LBM"=60), interval = "prediction")

## fit lwr upr  
## 1 14.42992 12.66232 16.19753

hconc.fit5 = lm(Hconc ~ Height \* Sex, data = athletes.df)  
summary(hconc.fit5)

##   
## Call:  
## lm(formula = Hconc ~ Height \* Sex, data = athletes.df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.28770 -0.69617 0.09872 0.53365 3.11403   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 17.69498 3.32119 5.328 5.4e-07 \*\*\*  
## Height -0.02141 0.01863 -1.150 0.253   
## SexM 4.94632 4.27494 1.157 0.250   
## Height:SexM -0.01607 0.02352 -0.683 0.496   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9236 on 109 degrees of freedom  
## Multiple R-squared: 0.491, Adjusted R-squared: 0.477   
## F-statistic: 35.05 on 3 and 109 DF, p-value: 6.159e-16

#predict(hconc.fit5, data.frame("Sex"="M", "Height"=170, "Weight"=70, "LBM"=60), interval = "prediction")  
#predict(hconc.fit5, data.frame("Sex"="F", "Height"=170, "Weight"=70, "LBM"=60), interval = "prediction")  
#preds.df = data.frame("Sex"="M", "Height"=170, "Weight"=70, "LBM"=60), interval = "prediction")  
#predict(hconc.fit4, preds, interval = "prediction")  
##plot(Hconc ~ Sex, data = athletes.df)  
##hconc.fit = lm(Hconc ~ Sex, data = athletes.df)  
##cooks20x(hconc.fit)  
  
## INSERT CODE HERE.  
## Analyse the data.  
## INSERT CODE HERE.

# Methods and Assumption Checks

From the pairs plot, we can concluded that Hconc was related to a number of our explanatory variables. So, we want to construct a multiple linear regression model with Hconc as the response variable.

We only use one parameter–LBM as explanatory variable at first, and all assumptions appear to be satisfied. Then we introduce Weight as another explanatory variable. But from the pairs plot, we saw that there is a strong correlation between Weight and LBM, so we shelved the case of introducing Weight as an explanatory variable. As we introduce Height and Sex, the Adjusted R-squared continuously increasing, so we keep them in our final model.

Our fimal model is:

where if the athlete is female and 1 if the athlete is male,

and

Our model explains about 54.38% of the variability in Haemoglobin concentration in blood.

# Executive Summary

We are interested in build a model to explain the Haemoglobin concentration in blood of athletes with their sex, height, weight, and LBM.

We can concluded from our analysis that:

The mean of the Haemoglobin concentration in blood is between 17.62 and 25.58 grams per decalitre.

We estimated that for each additional kg of athlete’s LBM, the Haemoglobin concentration in blood increases by 0.02 to 0.069 grams per decalitre.

We estimated that for each additional cm of athlete’s Height, the Haemoglobin concentration in blood decreases by 0.03 to 0.08 grams per decalitre.

And the predict value of haemoglobin concentration levels for a male with height 170 cm, weight 70kg, and lean body mass 60 kg is between 14.02 and 17.62 grams per decalitre. The predict value of haemoglobin concentration levels for a female with height 170 cm, weight 70kg, and lean body mass 60 kg is between 12.66 and 16.20 grams per decalitre.

When we fitted a model use Height and Sex as explanatory variable, as we saw, the p-value is 0.496, which means the relationship between haemoglobin concentration and height is the same for males and for females.