4 Workshop 3

4.1 Learning objectives

- To save commands as a script file so they can be reused.
- To fit a line to data
- To determine goodness of fit by calculating residuals
- To understand how much variance is explained by a model
- To understand the meaning of *significance* and *Explanatory Power*
- To assess the quality of a model fit

4.2 Recap

In the previous workshops we had loaded a dataset containing data from athletes. We had then used these data to calculate the Body Mass Index. We will reuse these data for this workshop. Set your working directory to the folder in which you saved the data files (using *Session > Set Working Directory > ...*). We had plotted the data and drawn a 'best fit' straight line through it using the `lm()` function. The mathematics behind regression fitting of straight lines is covered elsewhere [1], in this workshop we will be looking at how they are used and analysed.

Select the *File > New > R Script* menu option. A new pane will open at the top left.

Click on the pane and type in the following:

```
ais <- read.table("sport.txt", sep="\t", header=T)
attach(ais)
BMI <- Weight/((Height/100)**2)
plot(Height~Weight)
abline(lm(Height~Weight), col="red")
```

This set of commands will return us to the state at which we completed the last session. Click on *File > Save As ...* and save the file as `myscript.R`

Now click on the console pane (lower left) and enter

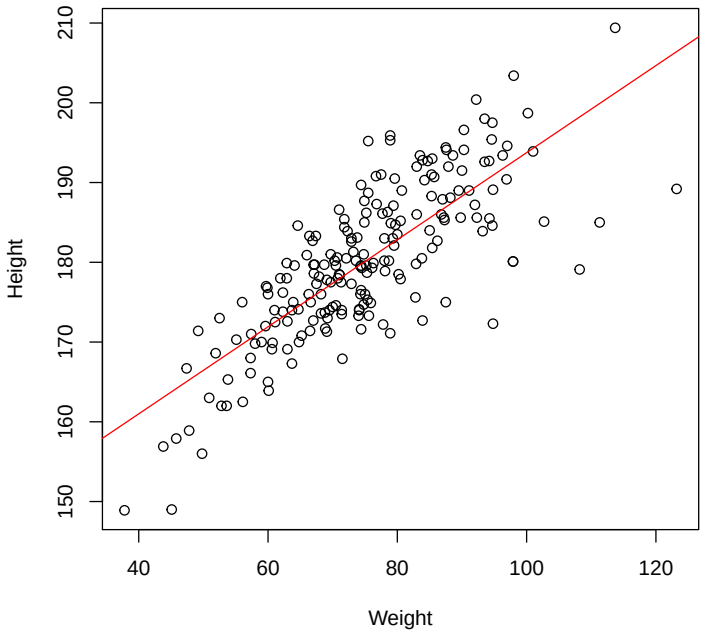
```
> source("myscript.R")
```

R will now run the commands you have saved in your script file.

When preparing figures for publicationit can be a good idea to write a script file to generate them. This lets you easily reproduce them, records where the data came from and allows changes to be made with very little effort.

```
> plot(Height~Weight)
> abline(lm(Height~Weight), col="red")
```

[1] See for example chapter 8 of *Crawley, M.J.* Statistics: An Introduction using R [ISBN 9780470022986]



The data look like they follow a pattern indicated by the straight line, but how well do they really follow it?

Every experimental observation can be described as a sum of the value expected from the model plus some error value. i.e for observation number i the values are:

$$Obs_i = Model_i + Error_i$$

The *error* or deviation of the observed value from the model is known as the *residual* and can be readily retrieved from the fitted model with the command `residuals()`

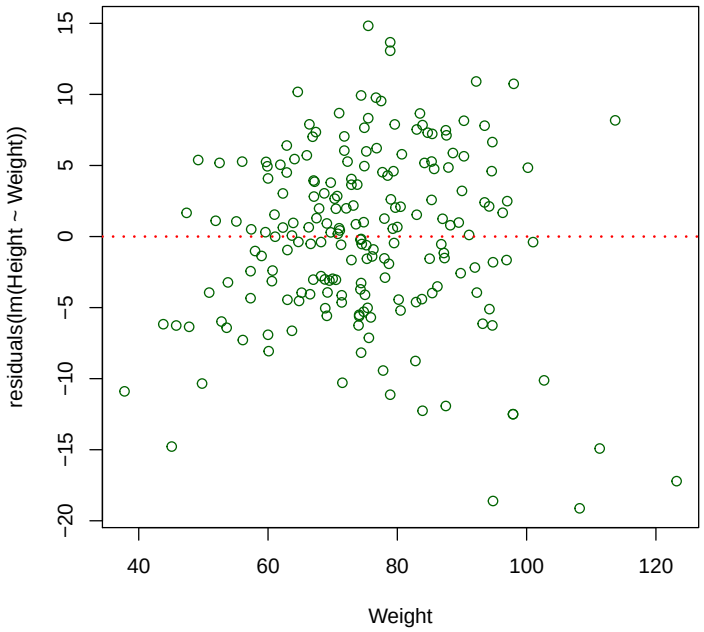
```
> residuals(lm(Height~Weight))
```

```
          1           2           3           4           5           6
13.67155658  9.92802808  0.92120562  4.95508680 10.17767713  0.06897143
          7           8           9          10          11          12
 5.99132203  0.63320700 -4.05949973  6.40567747  1.67320011  8.32755727
         13          14          15          16          17          18
-4.44891079 -5.20185551 -0.57973600  1.95697005  2.18308715  3.03955865
         19          20          21          22          23          24
 2.09814449 -1.65314809 -0.52656018 -5.01785448 -3.09714740  7.89508853
         25          26          27          28          29          30
 2.03485053  0.86473412 -1.91926691 -4.09950145 -10.34326105 3.85838248
         31          32          33          34          35          36
 5.71344155 -0.21738367 -2.89173738 -0.45597296  4.28990960  4.94332514
         37          38          39          40          41          42
-0.94891079  0.64967678 -2.39338091  4.04685191  1.97626469  1.29461771
         43          44          45          46          47          48
-5.50820715 -2.78750008 -3.01502961 -1.56326622  7.34920597 -2.97008868
         49          50          51          52          53          54
-6.25361890  1.11038558 -5.60820715 -3.71738367 -9.42797261  7.02214725
```

```
         55          56          57          58          59          60
-4.40326794 -4.61197364  5.45061840 -5.04232374 -4.53149939 -1.36538056
         61          62          63          64          65          66
 1.98355796 -7.12703099 -4.63432425  3.79367609  0.95979491  1.06356140
         67          68          69          70          71          72
-6.91126312 -1.01949801 -0.37691113 -11.92303340 -11.12844342 -12.25785620
         73          74          75          76          77          78
-8.75738539 -8.17197192 -18.60797606  5.38426848  5.05156003 -6.41761476
         79          80          81          82          83          84
-6.63102857 -5.98090872 -3.94985241 -3.94373186 -4.33738022  4.08873688
         85          86          87          88          89          90
-8.06585137  5.18285605  5.25250165 -2.43738022  0.30708990 -10.28891251
         91          92          93          94          95          96
 0.29367609 -7.28232115 -0.01173393  1.66685708  5.27226710 -6.25973083
         97          98          99         100         101         102
-6.35149594 -6.16796572 -10.89267038 -14.77761304 -3.03244101 -3.27197192
        103         104         105         106         107         108
 0.55320356  7.47696660  8.66049682 -1.53714912  1.26285088 -1.55832701
        109         110         111         112         113         114
 7.30543775 -2.17950490  5.27438145  7.53343810 -1.65432943  4.75955520
        115         116         117         118         119         120
-3.97668003  2.57790822  7.80167126 -0.54091561  4.85861358 -1.15926863
        121         122         123         124         125         126
-3.22679127 -2.57856328  0.11178940  5.87649579 -3.94326967  2.49108232
        127         128         129         130         131         132
 0.98520149  0.79484881 10.91131859 13.07155658  5.64849544  1.24990788
        133         134         135         136         137         138
 8.17484364 10.74519976  4.84425814  4.59861530  8.14849544  4.52661565
        139         140         141         142         143         144
 7.84214380 14.82755727 -3.13879265  8.68402877  6.04732272  6.21790995
        145         146         147         148         149         150
-10.12044825 -5.10774065  2.61696832 -0.51408799  7.04732272  1.00967506
        151         152         153         154         155         156
-0.38750008  3.03320700  1.54285433  9.53579216  0.50803153 -4.13432425
        157         158         159         160         161         162
 2.66614656 -4.43809074  5.17837903 -14.91503823  5.78896798 -12.50021198
        163         164         165         166         167         168
-17.21104064  3.64685191  1.53343810 -5.69079575  2.84779354  2.81297074
        169         170         171         172         173         174
-3.93338263  3.94026486 -3.04302995  0.19320528  0.58402877 -5.57879438
        175         176         177         178         179         180
 4.50567747 -1.80797606  4.60120045 -19.12280231 -12.50021198 -0.60867797
        181         182         183         184         185         186
-5.29032494  2.11955348 -1.39997227  6.64661220 -3.51338608  7.88943879
        187         188         189         190         191         192
 5.27790822 -0.17197192  2.40167126  7.12237834  7.22331997 -0.39244790
        193         194         195         196         197         198
 7.65508680 -1.51385689  3.21226021 -6.25338780 -0.90914878 -6.13456397
        199         200         201         202
 0.67108577  3.65555761  0.42944051  9.77249820
```

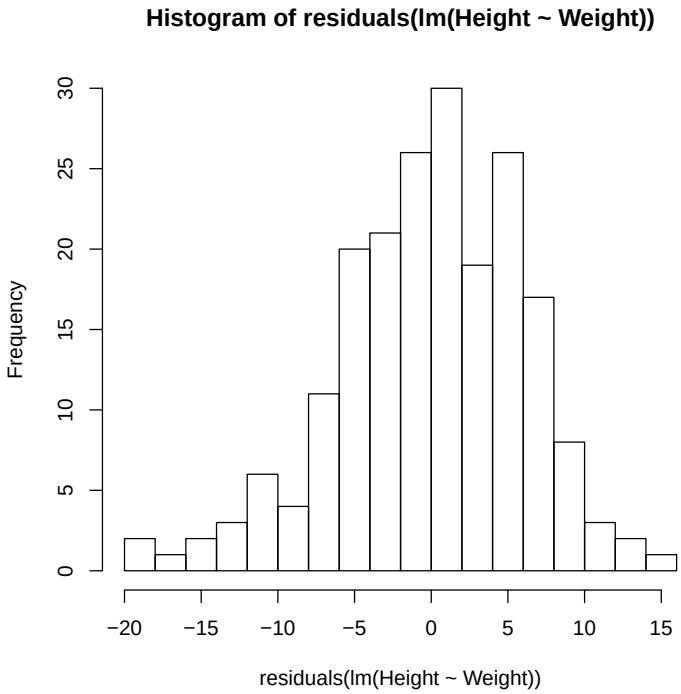
These values can be visualised in a plot.

```
> plot(residuals(lm(Height~Weight))~Weight, col='darkgreen', lty=2)
> abline(h=0, lty=3, lwd=2, col=2)
```



The mean residual is shown with a dashed red line. The sum of residuals is always zero (or as close to as a computer can calculate) so the mean will be zero. We can get an idea of how good or poor the fit is by plotting a histogram of the residuals with the `hist()` function. (The `breaks=` controls how many bars are plotted)

```
> hist(residuals(lm(Height~Weight)), breaks=20)
```

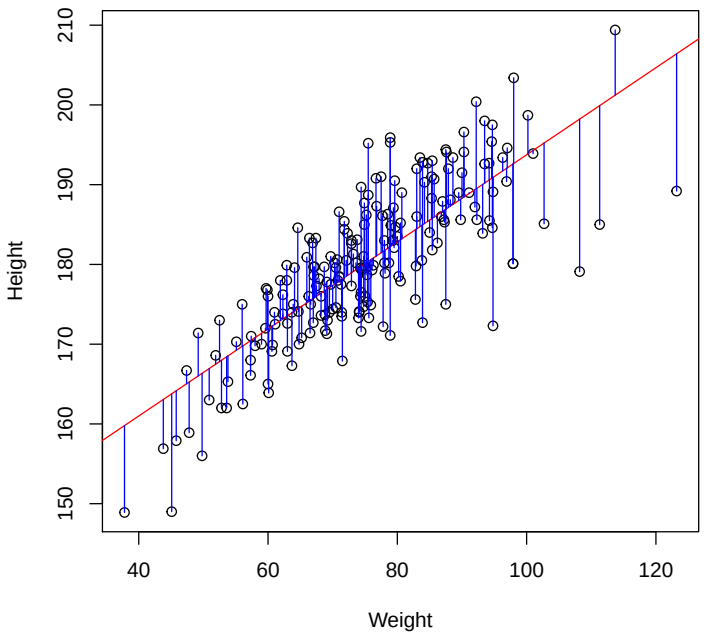


What is a residual? It can also be referred to as the error, or the deviation from prediction. The red line we have fitted to the data is the *Model*. Based on what we have already seen, this is our best estimate of where a new data point should appear. The residual describes how far from this predicted/fitted line the observed data actually lies. The better the model, the smaller the residuals. The analysis we will do later makes use of the residuals to calculate how good the model is.

Eyeballing the two plots, the residuals look to be evenly distributed along the plot and their magnitude approximates a normal distribution.

We can also plot the residuals against the model. This requires learning the `arrows(`x_1`,`y_1`,`x_2`,`y_2`, ...)` command which plots an arrow from (x_1,y_1) to (x_2,y_2). In this case we will plot with x as the weight, y_1 value will be our observed value, and y_2 our predicted value for y given x. The predicted value is available with the function `predict()`.

```
> plot(Height~Weight)
> abline(lm(Height~Weight), col='red')
> arrows(Weight, Height, Weight, predict(lm(Height~Weight)), 0, 0, col='blue')
```



The two `0` parameters control the size and angle of the arrowheads. We have set this to 0 so no arrowhead is shown. More inforomation on the `arrows()` command can be found with `help(arrows)`

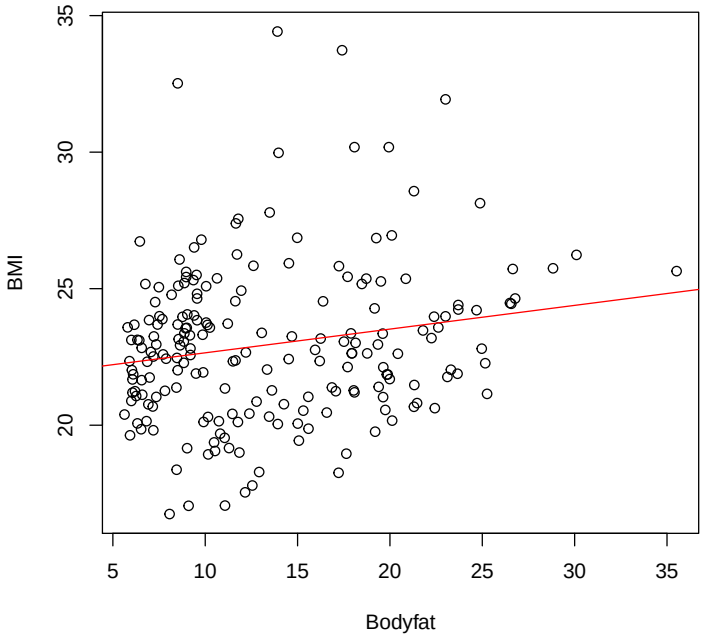
The `arrows()` command This takes the form `arrows(x1,y1,x2,y2,arrowheadlength, arrowheadangle, mode, ...)` draws a line from the point at coordinate x_1,y_1 to coordinate x_2,y_2. The arrowhead sides have a length of *arrowheadlength* (in inches - you have to play with this to get the desired results for your plot), and the sides of the arrow are at an angle of *arrowheadangle* from the straight line. An angle of 30-45 degrees gives a nice looking arrow. An angle of 90 degrees gives a flat bar like an error bar. The *mode* parameter describes whether arrowheads will be drawn at the start, end, both ends or neither end of the arrow.

4.2.1 Determining the goodness of fit

The plot of Height against Weight indicates a relationship between the two variables. How well does `lm` work when the data shows a less clear relationship?

Take a look at the plot of BMI against percent body fat. It might be thought a reasonable hypothesis that BMI (as a commonly used measure of obesity) is related to percent body fat.

```
> plot(BMI~Bodyfat)
> abline(lm(BMI~Bodyfat), col='red')
```



`lm` is happy to plot a best fit line throguh any dataset, whether it makes sens or not. We can explore the ability to which our x variable (the explanatory variable) explains the observations we get for the y value (the response variable).

The amount of the variation in y which is explained by the model SSR (the line fitted by `lm` and known as the regression sum of squares) is the variation in y as SSY minus the variation remaining after the model has been fitted (the mean sum of squares of the error, SSE)

If:

$$SSY = SSR + SSE$$

then:

$$SSR = SSY - SSE$$

$$SSY = \Sigma(y - \bar{y})^2$$

where $\bar{y}$ is the mean value of y

$$SSE = \Sigma(y - \hat{y})^2$$

where $\hat{y}$ is the predicted value of y from the model. $y - \hat{y}$ is the residual which we have plotted before.

You need to look closely at the equations. $\bar{y}$ is a y with a - over it. $\hat{y}$ is a y with a hat (ˆ) over it.

The degree of fit (r^2) is the proportion of *SSY* which is explained by the model, ie $\frac{SSR}{SSY}$, and this is the square of the correlation coefficient r for the two datasets.

Let's calculate these values. We'll calculate SSY for the Height/Weight comparison as SSYhw and for the BMI/Bodyfat as SSYbb, and likewise SSEhw, SSEbb and then SSRhw SSRbb. From the SSR and SSY values we can determine r^2 as RSQhw and RSQbb.

First let's calculate the results for Height/Weight

```
> SSYhw <- sum((Height - mean(Height))^2)
> SSEhw <- sum(residuals(lm(Height~Weight))^2)
> SSRhw <- SSYhw - SSEhw
> RSQhw <- SSRhw/SSYhw
> sqrt(RSQhw)
```

```
[1] 0.7809063
```

Now calculate the results for BMI/Bodyfat

```
> SSYbb <- sum((BMI-mean(BMI))^2)
> SSEbb <- sum(residuals(lm(BMI~Bodyfat))^2)
> SSRbb <- SSYbb - SSEbb
> RSQbb <- SSRbb/SSYbb
> sqrt(RSQbb)
```

```
[1] 0.1876716
```

An r^2 value of near 1 indicates that most of the variation in our observed response variable is explained by the explanatory variable. If r^2 is close to 0 then very little of the variation can be explained by the explanatory variable.

In the two analyses we have performed, r^2 for the Height *vs* Weight comparison is substantial, indicating that over half of the variation in Height can be explained by the value of Weight. In the BMI *vs* Bodyfat analysis the value of r^2 is very close to 0, indicating very little explanatory power for Bodyfat with respect to BMI.

A point to ponder: *Is this a genuine result? What aspects of the experimental design might influence our confidence in the result? Do all sports act the same?*

Challenge: *Is the explanatory power the same for all sports or genders? Can you find the sport or set of sports with the strongest and weakest correlation for Height Weight and BMI Bodyfat, and does Gender make a difference?*

4.2.2 Explanatory power and significance

How reliable are these results? Should we always expect the same lines (gradient and intercept) to be fitted if we sample more athletes? In order to establish this we need to determine the uncertainty associated with the intercept and the uncertainty associated with the gradient. We cannot calculate these uncertainties until we know the error variance s^2.

Variance is calculated as the **sum of squares** divided by the **number of degrees of freedom**. Degrees of freedom is the number of observations minus the number of estimated parameters.

For SSY, the number of estimated parameters is 1, $\bar{y}$ in the formula $\Sigma(y-\bar{y})^2$ so the number of degrees of freedom is $n-1$.

For SSE the number of estimated parameters is 2, a and b in the formula $\Sigma(y-a-bx)^2$ ($\hat{y}=a+bx$) so the number of degrees of freedom is $n-2$.

For SSR the number of extra parameters in addition to the mean of y is 1, the gradient of the line giving the regression degrees of freedom as 1.

What are *Degrees of Freedom*? Degrees of Freedom refers to the number of parameters, or values, that can be allowed to change whilst still getting the same result. For example, for a certain mean value of a set of data we can allow all bar one of the data points to vary as they wish. That last data point *must* then take a specific value if the set is to have that mean. The degrees of freedom are therefore *N -1* where N is the number of datapoints. For the regression fit, the data must have a specific mean, so the number of parameters that can be allowed to vary at will is all bar one. That last parameter (whether it be the intercept or the gradient) is constrained by the requirement to fit the mean. So the degrees of freedom for our fit is *Q -1* where *Q* is the number of parameters in the fitted equation.

The *F-ratio* is given by the variance in the regression, divided by the variance in the Error. We can then see how likely it is to get an F-ratio that high by chance (ie establish it's significance.)

To calculate the F-ratio for Height *vs* Weight, first calculate the variance in SSE. Note the -2 to get the correct number of degrees of freedom.

```
> varSSEhw <- SSEhw / (length(Height)-2)
```

Then calculate the variance in the regression:

```
> varSSRhw <- SSRhw / 1
```

Then calculate the F-ratio:

```
> fratio <- varSSRhw/varSSEhw
> fratio
```

```
[1] 312.5769
```

Is this a significant value? The F distribution can be accessed with several functions in R. The `pf()` function takes three arguments - the F-ratio, the degrees of freedom for the numerator, and the degrees of freedom for the denominator. It then returns the proportion fo the F-distribution for those degrees of freedom that has that F-ratio or lower.

```
> pf(fratio, 1, length(Height)-2)
```

```
[1] 1
```

We want the inverse, the proportion that would have that score or higher. This is our p-value for the fit.

```
> 1-pf(fratio, 1, length(Height)-2)
```

```
[1] 0
```

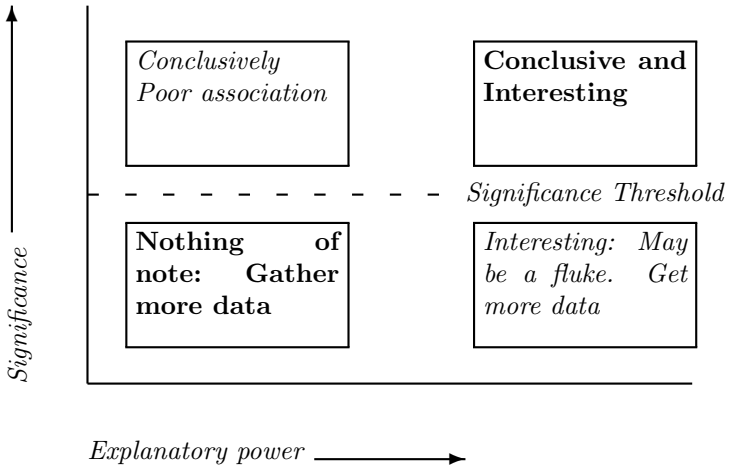
The score is very close to 0, so close it can't be measured. There is a highly significant relationship between Weight and Weight with good explanatory power.

BMI and percent body fat show a low correlation (<20%). Is there a significant relationship between the two variables?

Challenge: *Determine the p-value for BMI vs Bodyfat.*

4.2.3 The relationship between significance and explanatory power.

As has been seen above, explanatory power and significance are related but distinct concepts. Explanatory power describes the magnitude of any observed relationship between the x and y variables. Significance describes the likelihood that this observed association is genuine.



With low significance, no conclusions can be drawn from the data except that the number of data points is probably insufficient. High effect sizes at low significance are not conclusions - they are suggestive hints that may warrant further investigation to see if the experiment was just an artefact.

With high significance conclusions can be drawn. Even small relationships can be identified with high significance. However, a high explanatory power is needed for a predictive relationship between x and y.

Challenge: *Classify the sports into four categories for Weight vs body fat. Which of them demonstrate a significant relationship? Which demonstrate a high explanatory power? Which demonstrate both a high explanatory power and a significant relationship?*

4.2.4 Analysing the variance of the slope and the intercept

The standard error in the slope of the line $y = a + bx$ (se_b) is given by:

$$se_b = \sqrt{\frac{s^2}{SSX}}$$

where s^2 is the error variance (sum of squares of the error divided by the number of degrees of freedom in the error; $n-2$) and SSX is the corrected sum of sqaures of X^2 given by

$$SSX = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$$

For our Height *vs* Weight analysis, Weight is the x variable so this can be calculated as

```
> seb <- sqrt( (sum(residuals(lm(Height~Weight))**2) / (length(Weight)-2))
+             / (sum(Weight**2)-((sum(Weight)**2)/length(Weight)))  )
> seb
```

```
[1] 0.03087598
```

So for our Height *vs* Weight analysis the slope is 0.546 $\pm$ 0.031 (Standard Error).

Challenge: *What is the slope $\pm$ SE for the BMI* vs *Bodyfat relationship?*

The calculation of the Standard Error for the intercept se_a is a little more involved.

$$se_a = \sqrt{\frac{s^2 \Sigma x^2}{n \times SSX}}$$

It can be seen that this is equivalent to:

$$se_a = \sqrt{se_b^2 \times \frac{\Sigma x^2}{n}}$$

```
> sea = sqrt( seb**2 * sum(Weight**2)/length(Weight))
> sea
```

```
[1] 2.355331
```

So the intercept for the Height *vs* Weight plot is 139.2 *pm* 2.4 (Standard Error).

4.2.5 Analysis of regression analysis the quick way

The quick way to analyse the fit of the linear model is to use the `summary` function which calculates the errors and variances we have laboriously performed earlier.

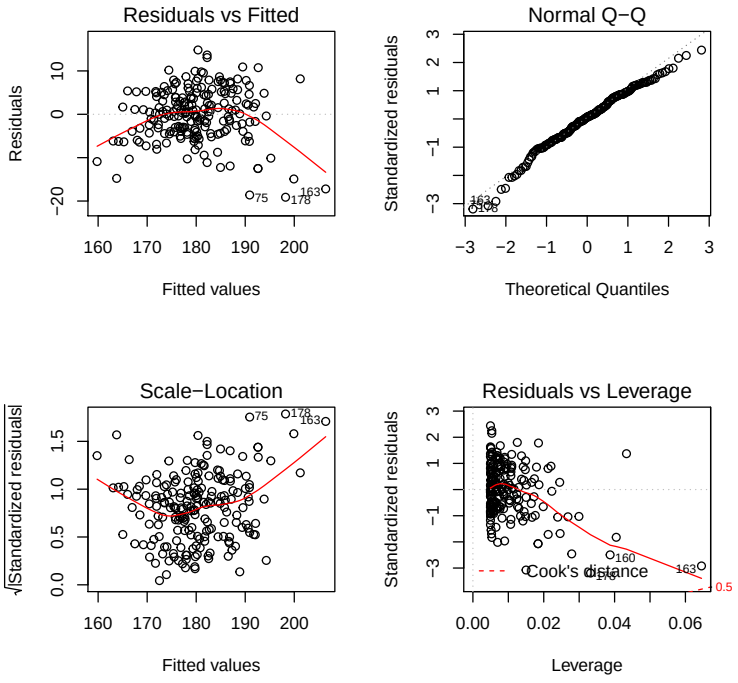
```
> summary(lm(Height~Weight))
```

[2] for a derivation of *SSX* see Crawley chapter 8.

```
Call:
lm(formula = Height ~ Weight)

Residuals:
     Min       1Q   Median       3Q      Max
-19.1228  -3.9483   0.3683   4.6006  14.8276

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 139.15831    2.35533   59.08   <2e-16 ***
Weight        0.54588    0.03088   17.68   <2e-16 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 6.096 on 200 degrees of freedom
Multiple R-squared:  0.6098,        Adjusted R-squared:  0.6079
F-statistic: 312.6 on 1 and 200 DF,  p-value: < 2.2e-16
```

Another useful feature is the plotting of various diagnostics when the model is passed to plot. This will print four plots. Press the <RETURN> key to move between plots.

```
> plot(lm(Height~Weight))
```



The first plot (top left) illustrates the residuals *vs* the predicted values. This should fit a straight line (shown in red). A very few outliers distort this set. The second plot (top right) shows the ranked qqnorm plot. An ideal fit would follow the diagonal. Significant deviations (a banana or S shape) sould indicate that the model being fitted is wrong. The third plot (lower left) shows the same as the upper left but with the absolute value of the residuals, scaled proportionately to the x-value. A good plot will be a straight line. A bad plot will show a triangular shape, indicating a bad fitting model. The fourth plot describes the influence of each point. The x-axis is the influence, the y axis the residual for that point. A contour line for Cook's distance is shown (just appearing on the lower right corner.) Strongly influential points will be to the right, points which

distort the model will be away from the 0 line. In this plot there are a small number of highly influential points that may distort the fit.

Challenge and discussion: *Which of male or female athletes fit best to Height* vs *Weight and BMI* vs *Bodyfat? Do certain sports distort these fits?*

4.3 Summary

A short summary of useful commands that you have used in this session.

```
> model <- lm(Ydata ~ Xdata)
```

Fit Ydata to Xdata and store the result in model.

```
> summary(model)
```

Provide all the analysis summary for the model.

```
> plot(Ydata~Xdata)
> abline(model)
```

Plot the graph and draw the best fit line through it.

```
> plot(model)
```

Plot the four graphs that analyse the fit of the model.