Level M Regression Models Simple Linear Regression

In this session you will learn how to load data, produce plots and fit linear regression models in R

Power and Weight

In an experiment conducted by he Physiology department, a sample of volunteers had their power output measured in watts while they ran up stair as fast as possible under different test conditions. Their gender, weight and leg length were also recorded.

The data are available on Moodle under **phys1.csv** and contain

`Gender`
`Weight (kg)`
`LegLen (m)`
`Power1`: Power output in stair test
`Power2`: Power output with a ramp on the stairs
`Power3`: Power output with a ramp on the stairs and a fixed stride length

Download these data and save the csv file to your working directory. You can load the data in R by typing

```
phys <- read.csv("phys1.csv")
```

You can check the column names

```
names(phys)
```

```
## [1] "Gender" "Weight" "LegLen" "Power1" "Power2" "Power3"
```

The full dataset can be viewed

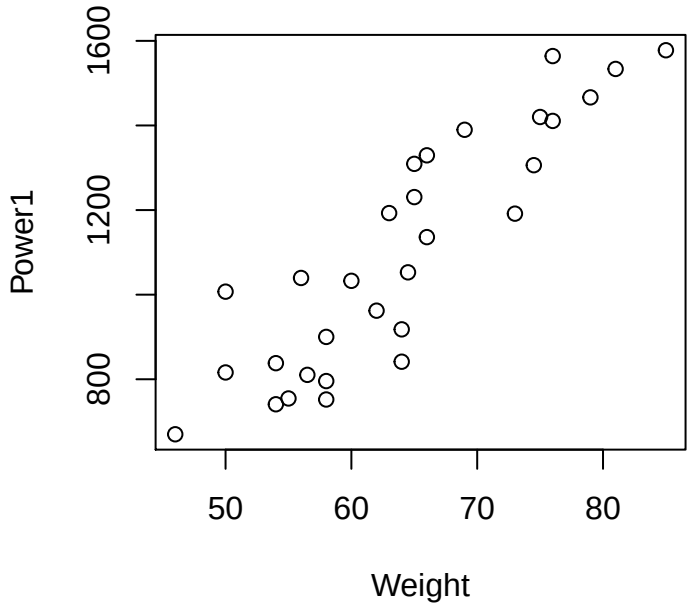
```
View(phys)
```

Note that this will open a new window and you must return back to your working R script.

Plotting the data

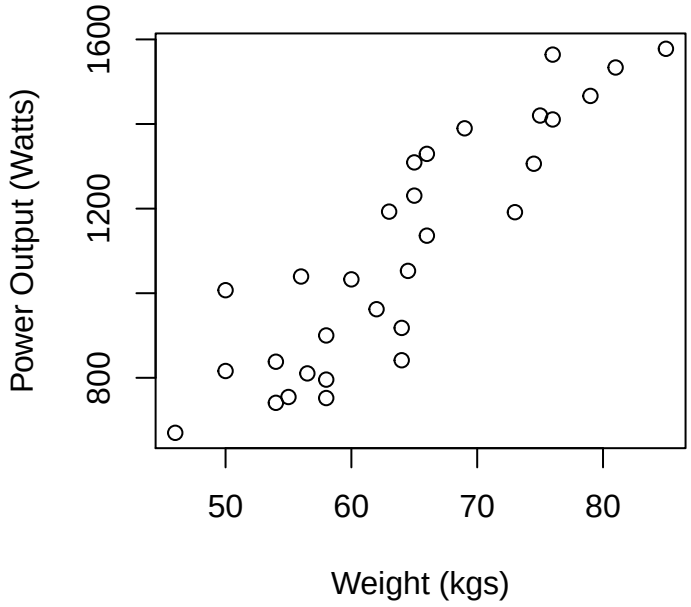
If we were interested in the relationship between weight and power 1, we can examine this relationship using a scatterplot

```
plot(Power1 ~ Weight, data=phys)
```



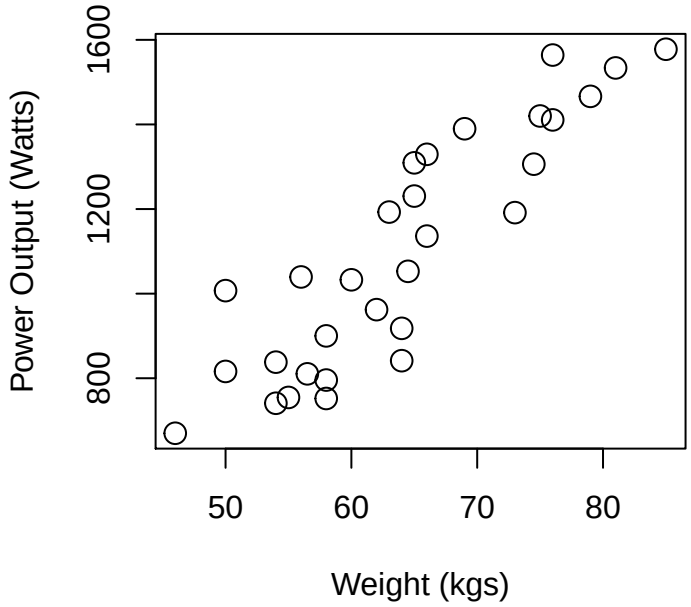
We can add axes labels to this plot

```
plot(Power1 ~ Weight, data=phys, xlab="Weight (kgs)", ylab="Power Output (Watts)")
```



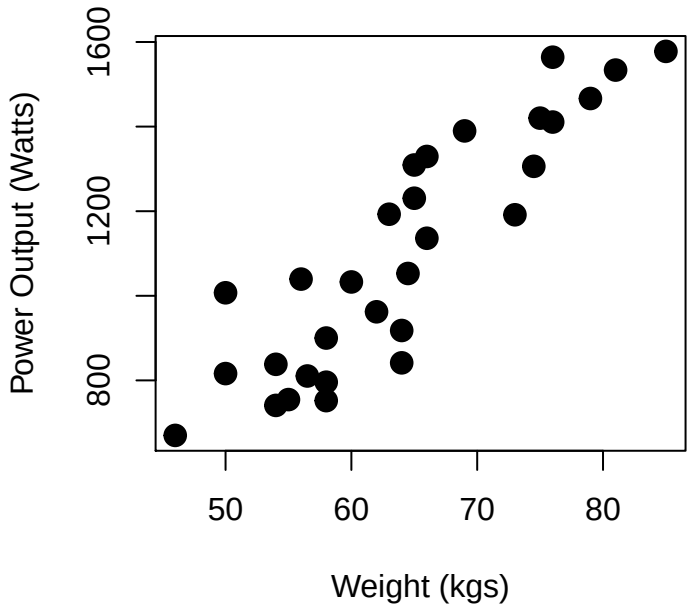
change the size of axes labels

```
plot(Power1 ~ Weight, data=phys, xlab="Weight (kgs)", ylab="Power Output (Watts)", cex=1.5)
```



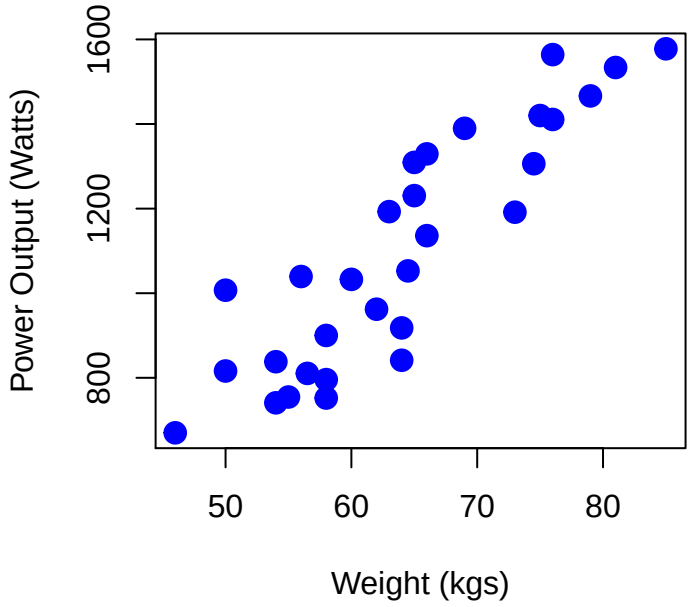
change the symbol

```
plot(Power1 ~ Weight, data=phys, xlab="Weight (kgs)",
     ylab="Power Output (Watts)", cex=1.5, pch=19)
```



and change the colour

```
plot(Power1 ~ Weight, data=phys, xlab="Weight (kgs)",
     ylab="Power Output (Watts)", cex=1.5, pch=19,col="blue")
```

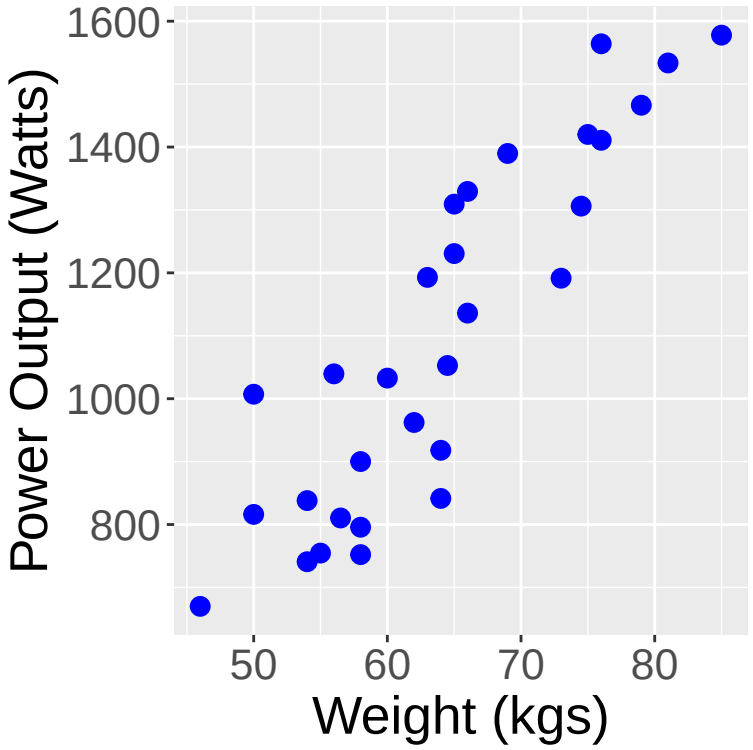


We can load the ggplot2 library

```
library(ggplot2)
```

which contains several functions that can be used to create nice plots. For example

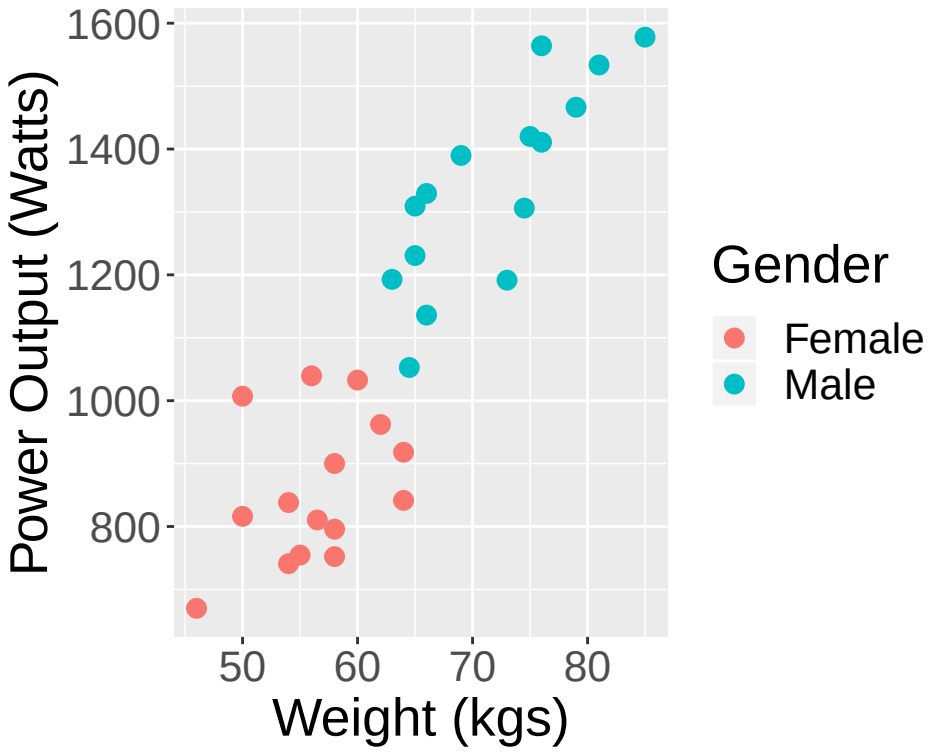
```
ggplot(data=phys, aes(y=Power1,x= Weight))+ #Specify x and y axes
geom_point(size = 3, color="blue") + #Specify to plot points size 2 coloured blue
labs(x="Weight (kgs)", y="Power Output (Watts)")+ #Specify axes labels
  theme(text = element_text(size=20)) #Change size of axes labels
```



Comment on the relationship between power and weight.

Now suppose we are interested in the relationship between power, weight and gender.

```
#Specify points to be coloured by gender.
ggplot(data=phys, aes(y=Power1,x= Weight,color=Gender)) +
geom_point(size = 3) +
labs(x="Weight (kgs)", y="Power Output (Watts)") +
  theme(text = element_text(size=20))
```



Comment on the relationship between power and weight and does this relationship differ for males and females?

Correlation between weight and power

What is the correlations between weight and power?

We can estimate the sample correlation

```
cor(phys$Power1, phys$Weight)
```

```
## [1] 0.889674
```

Therefore, there is a strong positive linear relationship between weight and power. Taking this one step further, we can test the null hypothesis that the population correlations is zero and provide a confidence interval for the correlation (i.e. a range of plausible values for the true population correlation). Remember we reject the null hypothesis, that the population correlation is zero, if the p-value < 0.05.

```
cor.test(phys$Power1, phys$Weight)
```

```
##
##  Pearson's product-moment correlation
##
## data:  phys$Power1 and phys$Weight
## t = 10.31, df = 28, p-value = 4.88e-11
## alternative hypothesis: true correlation is not equal to 0
```

```
## 95 percent confidence interval:
##  0.7791348 0.9465523
## sample estimates:
##      cor
## 0.889674
```

The p-value is 4.88×10^{-11} which is less than 0.05 and therefore we can reject the null hypothesis. In other words, these data give evidence on a significant, non-zero, correlation between power and weight. The output also provides a range of plausible values for the population correlation coefficient

$$(0.779, 0.947).$$

Notice this interval is centered around our estimate of 0.889. This interval is also entirely positive, and does not contain zero. If this interval did contain zero then that would suggest zero to be a plausible value. Since this interval does not contain zero, this suggests zero is not a plausible value. In summary, we have evidence of a significant strong positive correlation between power and weight.

Fitting a simple linear regression

We can use the `lm` function to fit a simple linear regression with weight as the explanatory variables and power as the response variable

```
model1<-lm(Power1 ~ Weight, data=phys)
summary(model1)
```

```
##
## Call:
## lm(formula = Power1 ~ Weight, data = phys)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -255.235  -93.830   -7.563   99.691  263.597
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -517.956    158.686  -3.264  0.00289 **
## Weight        25.229      2.447  10.310 4.88e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 130.3 on 28 degrees of freedom
## Multiple R-squared:  0.7915, Adjusted R-squared:  0.7841
## F-statistic: 106.3 on 1 and 28 DF,  p-value: 4.88e-11
```

Based on this output,

$$E(\texttt{Power1}|\texttt{Weight}) = -517.956 + 25.226 \times \texttt{Weight}$$

Comment on the value of R^2.

We can also fit a multiple linear regression model to assess the relationship between power, weight and gender

```
model2<-lm(Power1 ~ Weight + Gender, data=phys)
summary(model2)
```

```
##
## Call:
## lm(formula = Power1 ~ Weight + Gender, data = phys)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -177.50  -80.21   11.48   55.19  243.97
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   13.391    188.265   0.071 0.943818
## Weight        14.995      3.304   4.538 0.000105 ***
## GenderMale   249.708     64.259   3.886 0.000598 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 106.3 on 27 degrees of freedom
## Multiple R-squared:  0.8663, Adjusted R-squared:  0.8564
## F-statistic: 87.47 on 2 and 27 DF,  p-value: 1.595e-12
```

Based on this output,

$$\begin{aligned} E(\texttt{Power1}|\texttt{Weight, Male}) &= 13.391 + 249.708 + 14.995 \times \texttt{Weight} \\ E(\texttt{Power1}|\texttt{Weight, Female}) &= 13.391 + 14.995 \times \texttt{Weight} \end{aligned}$$

Lastly, we can fit two separate regression lines to assess the relationship between power and weight for males and females separately.

```
model3<-lm(Power1 ~ Weight * Gender, data=phys)
summary(model3)
```

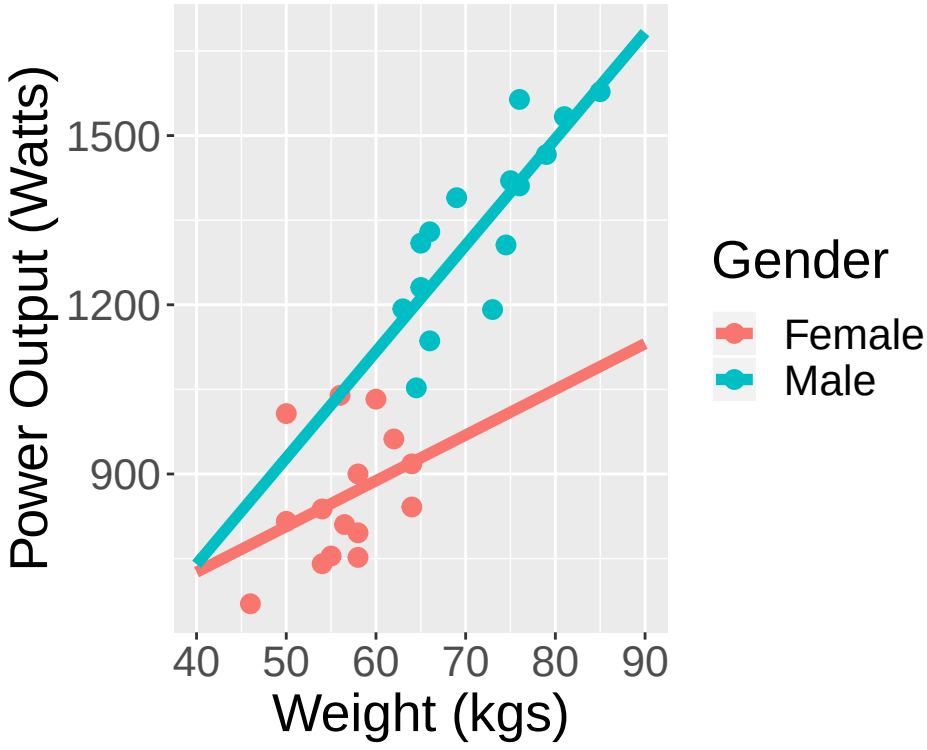
```
##
## Call:
## lm(formula = Power1 ~ Weight * Gender, data = phys)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -170.590  -82.972   -1.924   50.538  200.088
##
```

```
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        401.931    303.383   1.325    0.197
## Weight               8.102      5.361   1.511    0.143
## GenderMale        -416.113    419.349  -0.992    0.330
## Weight:GenderMale   10.751      6.696   1.606    0.120
##
## Residual standard error: 103.3 on 26 degrees of freedom
## Multiple R-squared:  0.8784, Adjusted R-squared:  0.8643
## F-statistic: 62.58 on 3 and 26 DF,  p-value: 5.031e-12
```

Based on this output,

$$\begin{aligned} E(\texttt{Power1}|\texttt{Weight, Male}) &= 401.931 - 416.113 + 14.995 \times \texttt{Weight} + 10.751 \times \texttt{Weight} \\ &= -14.182 + 25.746 \times \texttt{Weight} \\ E(\texttt{Power1}|\texttt{Weight, Female}) &= 401.931 + 8.102 \times \texttt{Weight} \end{aligned}$$

Let's add these two fitted lines to our plot of weight against power coloured depending on gender.



Do you think it is reasonable to have two separate regression lines, one for each gender?

Hubble's Constant

In 1929 Edwin Hubble investigated the relationship between distance and radial velocity of extra-galactic nebulae (celestial objects). It was hoped that some knowledge of this relationship might give clues as to the way the universe was formed and what may happen later. His findings revolutionized astronomy and are the source of much research today on the 'Big Bang'. Given here is the data that Hubble used for 24 nebulae. It is of interest to determine the effect of distance on velocity.

The data are available on Moodle under **hubble.csv** and contain

`Distance` in megaparsecs from Earth
`Velocity` the recession velocity in km/sec

Download these data and save the csv file to your working directory and load into R.

```
hubble <- read.csv("hubble.csv")
```

1. Produce a scatterplot of the data with velocity on the y-axis and distance on the x-axis.

```
plot(Velocity~Distance, data=hubble)
```

2. Describe the relationship between velocity and distance.
3. Fit a simple linear regression model with velocity as the reponse variable and distance as the explanatory variable.

```
model <- lm(Velocity ~ Distance, data=hubble)
```

4. Write down the fitted line.
5. Add this fitted line to your plot

```
plot(Velocity~Distance, data=hubble)
abline(model) #This function adds the fitted line from a simple linear regression.
```

6. Notice this intercept from this model looks quite close to zero. Try fitting a regression model with no intercept term.

```
model2 <- lm(Velocity ~ Distance - 1, data=hubble)

#Plot the data
plot(Velocity~Distance, data=hubble)
abline(model,col="blue") #This function adds the fitted line from a simple
                         #linear regression.
abline(model2,col="red")
```

7. Comparing the two fitted lines, is the model that passes through the origin plausible?

Publishing History

The Short Title Catalogue (STC) is a list of all the books that were published in Scotland, England and Ireland between 1475 and 1640. We are interested in finding out if there are any changes in the number of books published between 1500 and 1640.

The data are available on Moodle under **books.csv** and contain

```
Year
Number.of.Books
```

Download these data and save the csv file to your working directory and load into R.

```
books <- read.csv("books.csv")
```

1. Produce a plot of the data with number of books on the y-axis and year on the x-axis.
2. Describe the relationship between number of books published and year.
3. Using the command given below, fit a quadratic regression

```
model <- lm(Number.of.Books~poly(Year, 2), data=books)
```

4. Write down the fitted line of the form

$$\text{E(number of books)} = \quad + \quad \times Year + \quad \times Year^2$$

5. Plot the data with fitted line
6. Do you think this model provides a good fit to these data?

```
plot(Number.of.Books ~ Year, data=books)
lines(fitted(model))
```

The Taste of Cheese

Data are available from an experiment involving the chemical constituents of cheese and its taste. It contains concentrations of acetic acid, hydrogen sulphide (H2S) and lactic acid and a subjective taste score. It is of interest to investigate the effects of the different acids on the taste score.

The data are available on Moodle under **cheese.csv** and contain

`Case` Sample ID
`Taste` A taste score
`Acetic.Acid` Concentration
`H2S` Concentration
`Lactic.Acid` Concentration

Download these data and save the csv file to your working directory and load into R.

```
cheese <- read.csv("cheese.csv")
```

1. Produce plots of response variable Taste and each of the three chemicals Acetic Acid, H2S and Lactic Acid.
2. Which chemical do you think best describes Taste?
3. Fit a simple linear regression with your choosen chemical.
4. Produce a plot of your fitted line.
5. Is the population correlation between Taste and your choosen variable significantly differen from zero? Describe the correlations between Taste and your choosen variable.