MATH3823 - Solutions to Chapter 2 Exercises

Please note that the below calculations use the `lm` command in **R**, with a combination of `summary`, `anova` and `coefficients` to see the required numerical output. If it, however, the case that difference commands could have been used and, in particular, the more general `glm` command could have been used instead. The actual steps required would be a little different but the results would be the same.

Exercise 2.1 There is no fully right or wrong answers to these questions, but we rarely start an explanatory analysis without any information. In fact, if you ever start knowing *nothing* then make finding the background and context your first task. We will often come across unexpected features and our prior thoughts will sometimes be demonstrated incorrect, but be a detective and question every step.

a. It is well-known that a high calorie intake can lead to high weight and hence high BMI. The correlation is moderate and the scatter plot shows a general increase in BMI with weight. A linear model would be suitable, but might not be very reliable.
b. It is natural to expect that both consumption of electricity and gas will raise and fall together with both depending on number of daylight hours and general weather considerations, especially temperature. It is possible, however, that changes in lifestyles will increase the use of air conditioning, for example, which might be used more in summer than the winter use of gas for heating. The correlation is moderate and the scattering shows a general relationship, low values with low and high with high. A linear model would be suitable. It is important to note, however, that there are two suspected outliers. Before a regression is applied these should be investigated and corrected or removed before further modelling.
c. It is possible to imagine that a longer journey could lead to the accumulation of a bigger delay, but equally that if a delay occurs part-way through a short journey then they is little opportunity to *make up* the lost time. It is not clear which of these will be dominant. The correlation is negligible and there is no clear pattern in the data. There is no compelling argument that a linear model is suitable.
d. Having a large house is expensive to buy or rent and to run and hence large houses are likely to correspond to higher household income. Of course, it could be that households with high income may prefer to live in smaller houses. The correlation is moderate/high and the scatter plot shows the expected increasing trend, perhaps with a slight curve. A linear model is suitable, but goodness of fit should be carefully checked as a non-linear model may be necessary if an accurate model is required.

Exercise 2.2 We did not attempt to fit the model `birthweight` $\sim$ `sex`, that is `birthweight` $= \alpha + \beta$ `sex` where `sex` is coded 0 for girls and 1 for boys. This would be represented as two horizontal lines on the data plot, which does not look like a sensible explanation for the data and so we should not expect a good fit.

Let us fit the model and plot the resulting equations and the data:

```
birthweight = read.table("https://rgaykroyd.github.io/MATH3823/Datasets/birthwt-numeric.txt", header=T)

summary(birthweight)
```

```
##       sex           age            weight
##  Min.   :0.0   Min.   :35.00   Min.   :2412
##  1st Qu.:0.0   1st Qu.:37.00   1st Qu.:2785
##  Median :0.5   Median :38.50   Median :2952
##  Mean   :0.5   Mean   :38.54   Mean   :2968
```

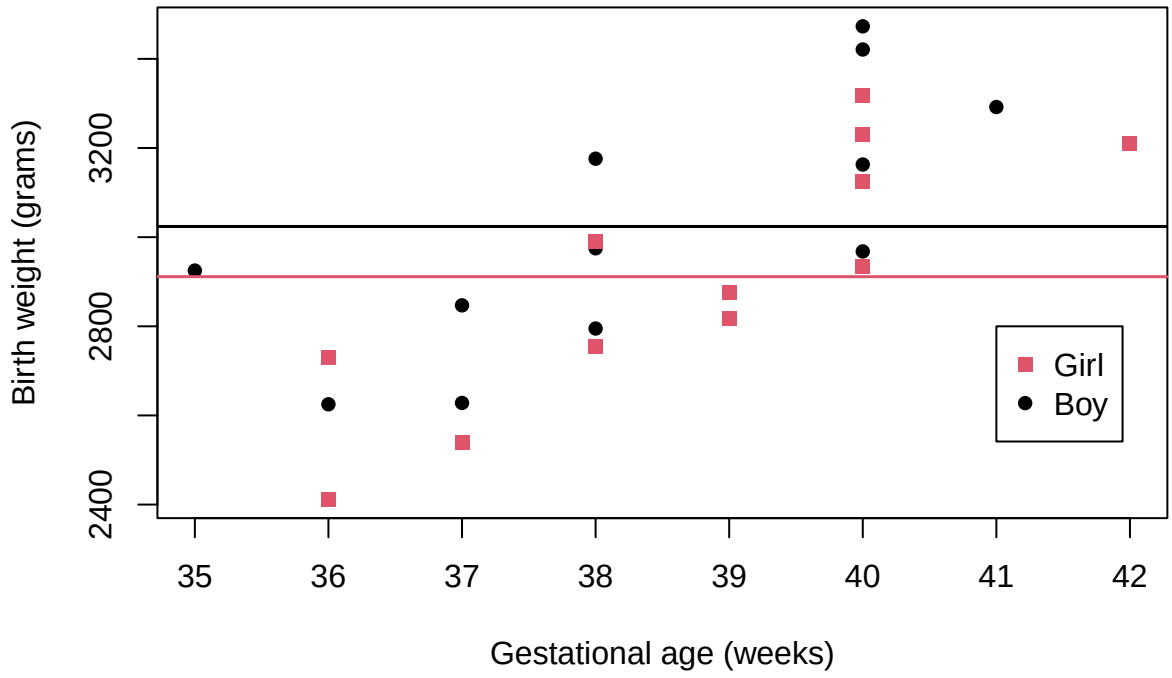
```
##  3rd Qu.:1.0   3rd Qu.:40.00   3rd Qu.:3184
##  Max.   :1.0   Max.   :42.00   Max.   :3473
```

```
attach(birthweight)

M1.fit = lm(weight~sex)

plot(age, weight, col=2-sex, pch=15+sex,
     xlab = "Gestational age (weeks)",
     ylab = "Birth weight (grams)")
legend(41,2800, c("Girl","Boy"), col = c(2,1), pch=c(15,16))

abline(h=M1.fit$coefficients[1] + c(0,1)*M1.fit$coefficients[2], col=c(2,1), lwd=1.5)
```



The model fitting summary output is:

```
summary(M1.fit)
```

```
##
## Call:
## lm(formula = weight ~ sex)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -499.33 -178.33  -42.67  228.00  449.00
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)   2911.3       81.5  35.720   <2e-16 ***
## sex            112.7      115.3   0.977    0.339
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 282.3 on 22 degrees of freedom
## Multiple R-squared:  0.04162,    Adjusted R-squared:  -0.001941
## F-statistic: 0.9554 on 1 and 22 DF,  p-value: 0.339
```

and we have fitted equations `birthweight` $= 2911.3$ when `sex` $= 0$ and `birthweight` $= 2911.3 + 112.7 = 3024$ when `sex` $= 1$.

Further, the overall birth weight mean and the mean birth weights of the girls and boys can be found as:

```
mean(weight)
```

```
## [1] 2967.667
```

```
weighted.mean(weight,sex==0)
```

```
## [1] 2911.333
```

```
weighted.mean(weight,sex==1)
```

```
## [1] 3024
```

We see that the mean of the girls is given by the intercept parameter and the mean of the boys by the sum of the two parameters. The overall mean birthweight is given by the simple average of these two group means (since there are equal numbers of girls and boys).

We did not expect a good fit and this is clear from the graph. To test the null hypothesis that including `sex` is important, or not, we consider the F-statistic, 0.9554, which follows an $F_{1,22}$ distribution. The p=value is 0.339, not significant, and hence we conclude that `sex` is not important in this case.

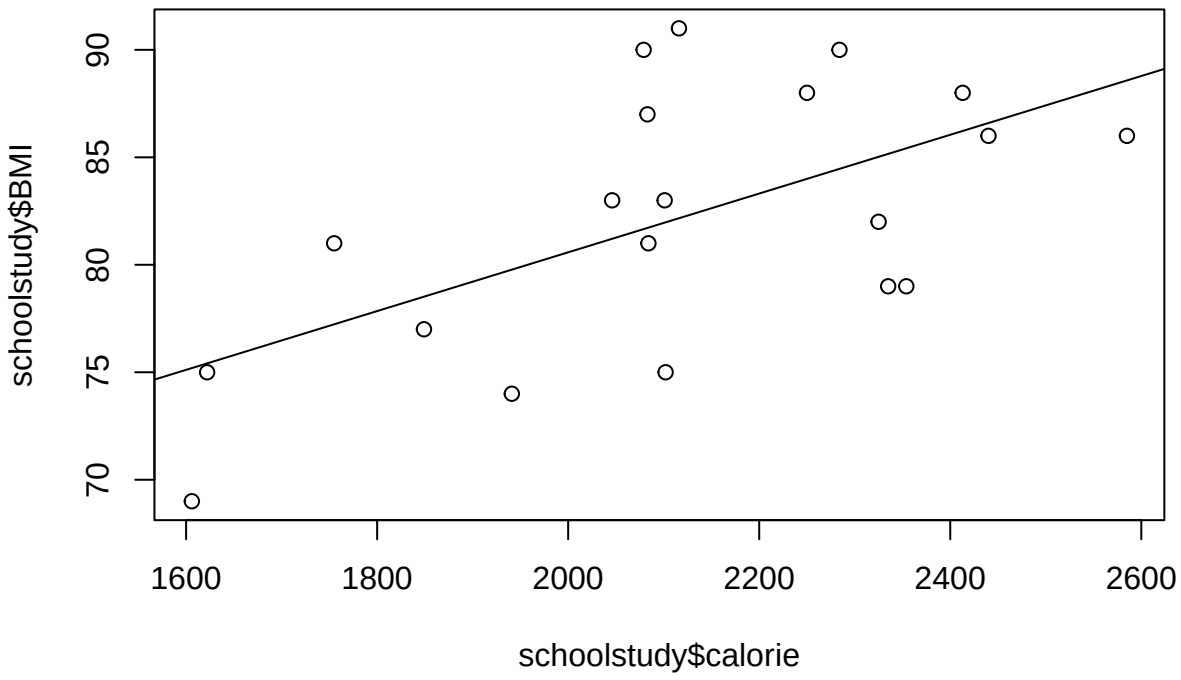
Exercise 2.3

a. Start by reading-in the data, then draw the scatter plot, perform the model fit and add the line to the scatter plot.

```
schoolstudy = read.csv("https://rgaykroyd.github.io/MATH3823/Datasets/schoolstudy.csv")

plot(schoolstudy$calorie, schoolstudy$BMI)

lin.model = lm(schoolstudy$BMI~schoolstudy$calorie)
abline(lin.model)
```




```
cor(schoolstudy$calorie, schoolstudy$BMI)
```

```
## [1] 0.5944361
```

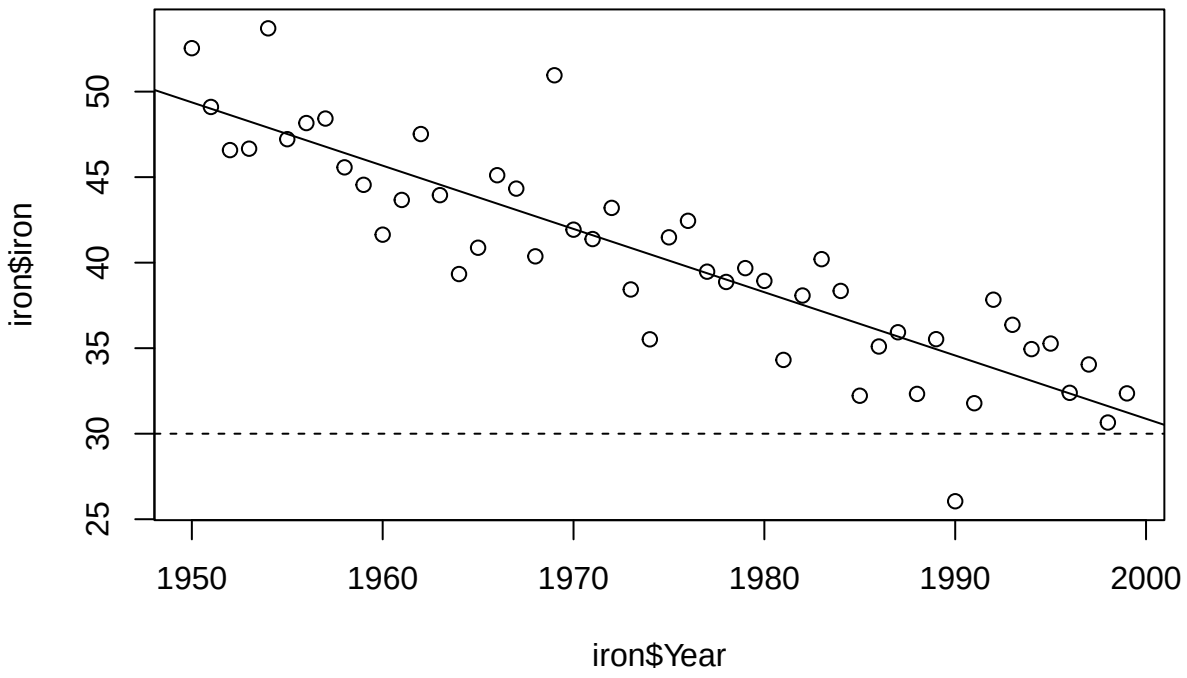
A linear model appears suitable, though points well scattered and so unlikely to be very reliable.

b. Start by reading-in the data, then draw the scatter plot, perform the model fit and add the line to the scatter plot.

```
iron = read.csv("https://rgaykroyd.github.io/MATH3823/Datasets/iron.csv")

plot(iron$Year, iron$iron)
abline(h=30, lty=2)

myfit = lm(iron~Year, data=iron)
abline(myfit)
```




```
cor(iron$Year, iron$iron)
```

```
## [1] -0.8795975
```

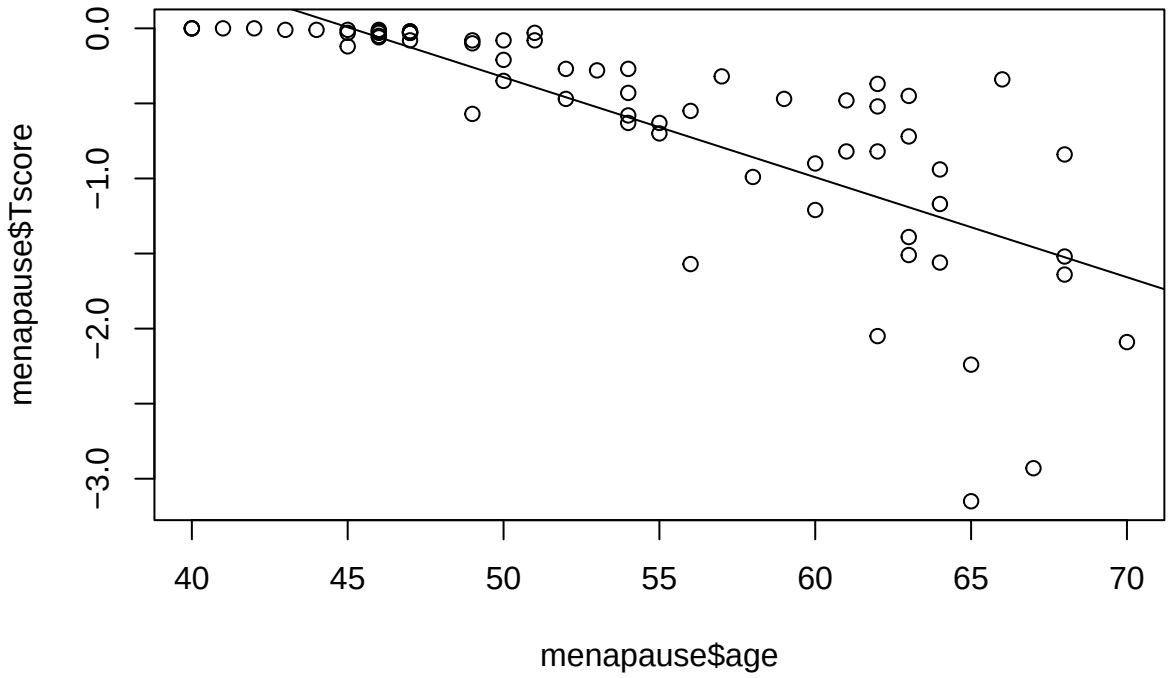
A clear linear relationship with a correlation of -0.88 showing a high correlation. A linear model is suitable and likely to be reasonably reliable. Here, recalling the early information, the downward trend in iron purity reached the critical 30% around 2000 – with an even lower figures in 1990. It appears as if the quarry will have become uneconomical.

c. Start by reading-in the data, then draw the scatter plot, perform the model fit and add the line to the scatter plot.

```
menapause = read.csv("https://rgaykroyd.github.io/MATH3823/Datasets/bmd.csv")

plot(menapause$age, menapause$Tscore)

themodel = lm(Tscore~age, data = menapause)
abline(themodel)
```




```
cor(menapause$age, menapause$Tscore)
```

```
## [1] -0.7713094
```

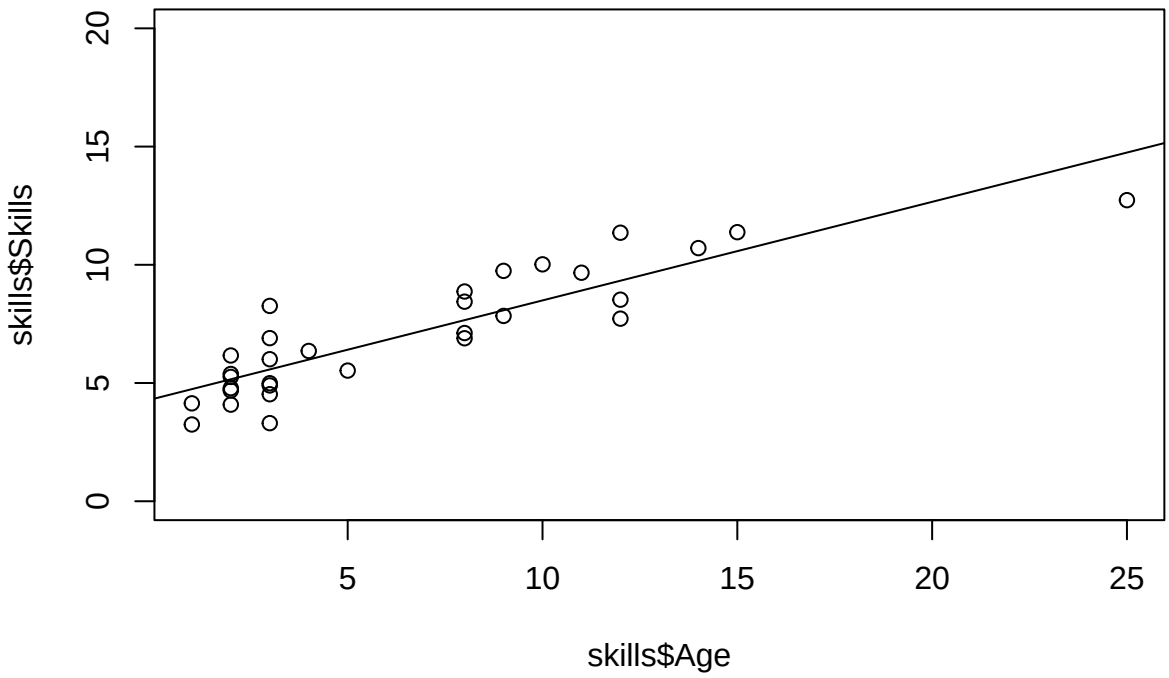
There is a definitely downward trend in Tscore with age, but it does not look as it a linear model fits very well. The relationship is non-linear with little change at lower ages and then a strong decline. Further there is very little variability in Tscore for younger ages but high variability for older ages. A non-linear model is required to get a good fit.

d. Start by reading-in the data, then draw the scatter plot, perform the model fit and add the line to the scatter plot.

```
skills = read.csv("https://rgaykroyd.github.io/MATH3823/Datasets/skills.csv")

plot(skills$Age, skills$Skills, ylim=c(0,20))

fitted.model = lm(Skills~Age, data=skills)
abline(fitted.model)
```




```
cor(skills$Age, skills$Skills, use="complete.obs")
```

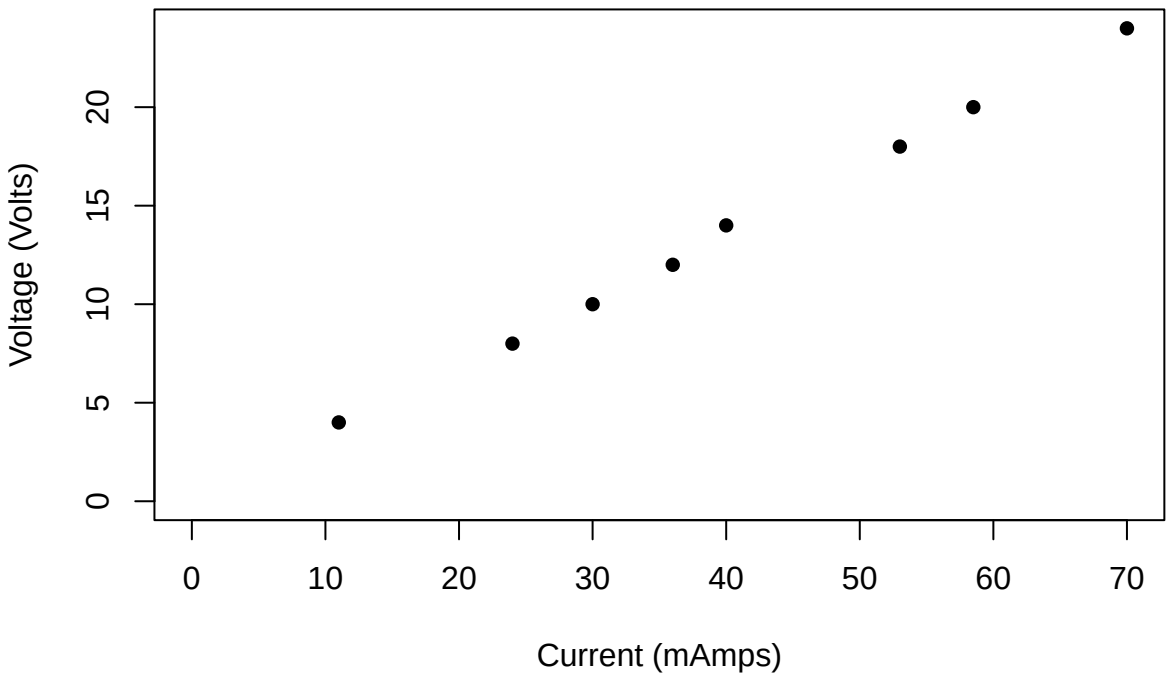
```
## [1] 0.8840017
```

There appears to be a clear link between the social skills of children and the age of their eldest sibling – the older the more skilled with a high correlation. The miss-recorded age, 25 instead of 15, should be corrected and the fitting process re-done. This shuld then provide a reliable linear model.

Exercise 2.4 Notice that **V=IR** is a linear equation with intercept fixed at zero, $y = \beta\, x$ say. To start the investigation, first plot the data:

```
Volts = c(4,8,10,12,14,18,20,24)
mAmps = c(11,24,30,36,40,53,58.5,70)

plot(mAmps, Volts,pch=16, ylim=c(0,24), xlim=c(0,70),
     ylab = "Voltage (Volts)",
     xlab = "Current (mAmps)")
```



From the graph, a linear model through the origin should fit well, but first let's fit a regular linear model:

```
M1.fit = lm(Volts ~ mAmps)

summary(M1.fit)
```

```
##
## Call:
## lm(formula = Volts ~ mAmps)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.27447 -0.18187 -0.03197  0.13910  0.35692
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.043001   0.215944  -0.199    0.849
## mAmps        0.342152   0.004886  70.022 5.71e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2503 on 6 degrees of freedom
## Multiple R-squared:  0.9988, Adjusted R-squared:  0.9986
## F-statistic:  4903 on 1 and 6 DF,  p-value: 5.708e-10
```

Note that this command fits a model which includes the intercept but we see that the result of the t-test on the intercept parameter has a p-value of 0.849 saying that it is not significantly different from zero.

Let's finish by fitting the suggested model

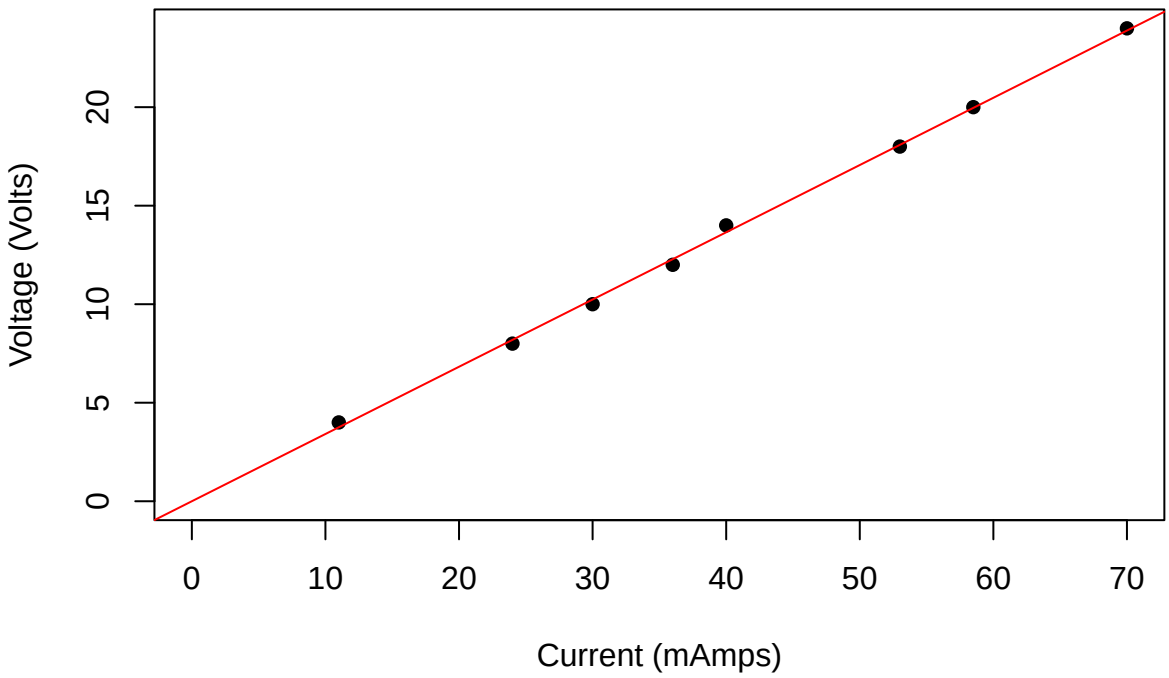
```
M2.fit = lm(Volts ~ mAmps-1)

summary(M2.fit)
```

```
##
## Call:
## lm(formula = Volts ~ mAmps - 1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.28552 -0.20224 -0.02549  0.14514  0.34942
##
## Coefficients:
##       Estimate Std. Error t value Pr(>|t|)
## mAmps  0.34126    0.00186   183.5 3.77e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2325 on 7 degrees of freedom
## Multiple R-squared:  0.9998, Adjusted R-squared:  0.9998
## F-statistic: 3.367e+04 on 1 and 7 DF,  p-value: 3.769e-14
```

Here, we see that the fitted model is $\texttt{Volts} = 0.34126\,\texttt{mAmps}$ with $\texttt{F-statistic} = 3.367e+04$, which follows an $F_{1,7}$ distribution, with a very small p-value and hence the slope parameter is highly significant. Seeing the fitted line along with the data reinforces that this is an almost perfect fit.

This does support Ohm's Law, with a resistance of 0.34Ω – given by the slope parameter estimate.



For information, this data set was collected by my father, Peter John Aykroyd, in 1956, while he was training as an electrical engineer.

Exercise 2.5 Let `pulse` denote the rest pulse rate, `sex` is recorded as 0 for Men and 1 for Women, and `time` as 0 for Before and 1 for After. Then, we suggest a model, $\texttt{pulse} = \alpha + \beta\,\texttt{sex} + \gamma\,\texttt{time} + \delta\,(\texttt{sex:time})$ to take into account that Men and Women might have different rest pulse rates, that there might be a change after a meal, and that Men and Women might be effected by a meal differently.

Suppose that we enter data row-by-row starting from the top-left. Then, $\texttt{pulse} = (105, 79, 79, 103, 87, 97, 109, 87, \ldots, 93, 81)$ and our design matrix, without selected columns removed to avoid aliasing (that is to remove identifiability

issues), is given by

$$X = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

where the first column represents the model intercept, the second indicates that the first 12 recorded values correspond to Men and the second 12 to Women; further the third column has blocks of 6 for the before and after for Men, then before and after for Women; finally the fourth column has the product of the second and third columns for the interaction term.

Defining the data vector and the design matrix in **R**,

```
pulse = c(105,79,79,103,87,97,109,87,86,109,100,101,74,73,82,78,86,77,81,80,90,90,93,81)

X=matrix(c(
1, 0, 0, 0,
1, 0, 0, 0,
1, 0, 0, 0,
1, 0, 0, 0,
1, 0, 0, 0,
1, 0, 0, 0,
1, 0, 1, 0,
1, 0, 1, 0,
1, 0, 1, 0,
1, 0, 1, 0,
1, 0, 1, 0,
1, 1, 0, 0,
1, 1, 0, 0,
1, 1, 0, 0,
1, 1, 0, 0,
1, 1, 0, 0,
1, 1, 1, 1,
1, 1, 1, 1,
```

```
1, 1, 1, 1,
1, 1, 1, 1,
1, 1, 1, 1,
1, 1, 1, 1
), ncol=4, byrow=T)
```

and then the matrix regression equation

```
beta = solve(t(X) %*% X) %*% t(X) %*% pulse

beta
```

```
##           [,1]
## [1,]  91.66667
## [2,] -13.33333
## [3,]   7.00000
## [4,]   0.50000
```

Using the `lm` command in **R** gives:

```
sex  = X[,2]   # this is easier than re-typing the 0 and 1's
time = X[,3]

my.fit.1 = lm(pulse ~ sex*time)

coefficients(my.fit.1)
```

```
## (Intercept)         sex        time    sex:time
##    91.66667   -13.33333     7.00000     0.50000
```

Notice that the coefficients, as expected, are identical to those obtained using the matrix regression equation above.

To check if the change in pulse rate due to a meal is the same for men and women we test the interaction term. From the anova table, we see that the `sex:time` interaction is not significant, with a p-value of 0.9440. Hence, there is no significant interaction and there is no evidence that the change is different.

```
anova(my.fit.1)
```

```
## Analysis of Variance Table
##
## Response: pulse
##           Df  Sum Sq Mean Sq F value   Pr(>F)
## sex        1 1027.04 1027.04 13.8524 0.001347 **
## time       1  315.37  315.37  4.2537 0.052396 .
## sex:time   1    0.38    0.38  0.0051 0.944010
## Residuals 20 1482.83   74.14
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Finally, we fit the reduced model without the interaction term:

```
my.fit.2 = lm(pulse ~ sex+time)

summary(my.fit.2)
```

```
##
## Call:
## lm(formula = pulse ~ sex + time)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.792  -4.896   0.375   5.917  13.458
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   91.542      2.971  30.809  < 2e-16 ***
## sex          -13.083      3.431  -3.813  0.00101 **
## time           7.250      3.431   2.113  0.04673 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.404 on 21 degrees of freedom
## Multiple R-squared:  0.4751, Adjusted R-squared:  0.4251
## F-statistic: 9.503 on 2 and 21 DF,  p-value: 0.001151
```

This gives final parameter estimates, all are significant, and shows that the model fits the data well, with an chi-squared statistic of 8.404 which follows a χ^2_{21} distribution and has a p-value of:

```
pchisq(8.404, 21, lower.tail = F)
```

```
## [1] 0.9931796
```

In conclusion, the best fitted model is: $\texttt{pulse} = 91.5 - 13.1\,(\texttt{sex}) + 7.25\,(\texttt{time})$ and hence the average pulse rate is about 91 beats/minute, the rate is 13 beats/minute lower for women than men, and there is a 7 beats/minute increase after a meal and this change in pulse rate is the same for men and women.

Exercise 2.5 A suitable model might say that seedling height depends on seed type, watering conditions and a possible interaction – that is some types of barley are more, or less, sensitive to water conditions. From the description, it seems that the incubator shelf level is not of interest.

```
barley = read.csv("https://rgaykroyd.github.io/MATH3823/Datasets/barley.csv")

attach(barley)
```

The response variable in the data set is `Height` with explanatory variables `Variety` with levels *G*, *M* and *Ig*, `Watering` with levels *N* and *W*, and `Position` with levels *Top*, *Second*, *Third* and *Bottom*.

Following the guidance in the question, the most complicated model that we might consider is

$$(\texttt{Height})_{ijk} = \texttt{mean} + (\texttt{Variety})_i + (\texttt{Watering})_j + (\texttt{Position})_k + (\texttt{VarietyWatering})_{ij}$$

where $i = 1, 2, 3$ for *G*, *M* and *Ig*, $j = 1, 2$ for *N* and *W*, and $k = 1, 2, 3, 4$ for *Top*, *Second*, *Third* and *Bottom*.

```
Variety  = as.factor(Variety)
Watering = as.factor(Watering)
Position = as.factor(Position)

M1.fit = lm(Height ~ Variety + Watering + Position + Variety:Watering)

anova(M1.fit)
```

```
## Analysis of Variance Table
##
## Response: Height
```

```
##                   Df   Sum Sq Mean Sq F value Pr(>F)
## Variety            2  3068488 1534244  0.9981 0.3918
## Watering           1  1708480 1708480  1.1114 0.3085
## Position           3  4628410 1542803  1.0036 0.4183
## Variety:Watering   2  3091330 1545665  1.0055 0.3892
## Residuals         15 23058076 1537205
```

The `Variety:Watering` interaction terms has an F-statistics of 1.0055 which follows an $F_{2,15}$ distribution and has a p-value of 0.3892 – which is non-significant. Hence there is no evidence that difference varieties respond differently to water levels.

```
M2.fit = lm(Height ~ Variety + Watering  )

anova(M2.fit)
```

```
## Analysis of Variance Table
##
## Response: Height
##           Df   Sum Sq Mean Sq F value Pr(>F)
## Variety    2  3068488 1534244  0.9970 0.3866
## Watering   1  1708480 1708480  1.1102 0.3046
## Residuals 20 30777816 1538891
```

From this anova table it appears that none of the variables are significant. This might suggest that the natural variability is large and that a bigger sample would be needed to identify any pattern.

End of Solutions to Chapter 2 Exercises