STAT2402: Week 3 Computer Laboratory

Getting Started:
Log in at one of the PCs and start up the software package RStudio. If you have problems either logging in or starting RStudio, ask the lab demonstrator for help.

When typing in R commands in the console window you can use the arrow keys to speed things up. The 'up' arrow gives you the previous command that you typed. The usual prompt sign for R is `>`. If you get a `+` prompt sign instead, it means that R is awaiting the completion of the previous command that you typed in. This can happen because you have forgotten to close parentheses, for instance. Just type in the remainder of the command and press enter.

You can ask the lab demonstrator for help at any point when you have a problem.

This laboratory session covers the following topics:

1. Revision of linear models.
2. Probability.
3. Random variables.

Exercise 1: Linear models We will work on the weight loss clinical trial discussed in lectures this week. A randomised trial will be conducted to investigate the effect of exercise and diet on adults. Each healthy subject will be randomly allocated to an exercise regime and a diet level. The following variables will be collected.

- Exercise: Exercise regime, Yes or No
- Diet: Yes or No
- Sex: M or F
- Age: continuous
- Height: in metres
- Weight0: Initial Weight (kg), continuous
- Weight1: Weight (kg) after twelve weeks, continuous

1. An important consideration for trials is sample size. The calculation is based on *power* of the statistical test for a difference in weight loss due to the treatments (exercise and diet). We will use the `R` package `pwr` and the function `pwr.f2.test` to compute a sample size. Look at the document Sample Size Calculation on LMS in the folder Computer Labs.

   (a) You need to specify u, the numerator degrees of freedom. Note that u is simply the number of parameters in your model. You can determine this by writing your model equation.

   (b) Compute sample size for powers 0.85 to 0.95 in steps of 0.05.

   (c) You will also need to select an effect size f_2. Determine the effect of this on sample size be taking values between 0.3 and 0.8 in steps of 0.1

   The function gives you the value of v in the output. The required sample size is $u + v + 1$, rounded up to the next integer.

   Solution: Here the numerator degrees of freedom (df) are calculated as follows.

   - Exercise: categorical variable with two levels, so df = 1.
   - Diet: categorical variable with two levels, so df = 1.
   - Sex: categorical variable with two levels, so df = 1.

- Age: continuous variable, so df = 1.
- Height: continuous variable, so df = 1.
- Weight0: continuous variable, so df = 1.

This gives a total of 6. That is, $u = 6$. The following code calculates the sample size as a matrix.

```
library(pwr)
u <- 6
f2 <- seq(0.3, 0.8, 0.1)
power <- seq(0.85, 0.95, 0.05)
x <- matrix(nrow = 3, ncol = 6)
for (i in 1:3) for (j in 1:6) {
    a <- pwr.f2.test(u, f2 = f2[j], sig.level = 0.05, power = power[i])[2]
    x[i, j] <- a[[1]]
}
rownames(x) <- c("0.85", "0.90", "0.95")
colnames(x) <- c("0.3", "0.4", "0.5", "0.6", "0.7", "0.8")
(x <- ceiling(x + u + 1))
```

```
##      0.3 0.4 0.5 0.6 0.7 0.8
## 0.85  58  45  38  33  29  27
## 0.90  65  51  42  36  32  29
## 0.95  77  59  49  42  37  34
```

Observe that as power increases so does sample size, and as the effect size increases sample size decreases. The latter indicates simply that to detect smaller differences between treatments requires larger sample sizes, or equivalently, as the variation in the data decreases the required sample size decreases. As one expects, smaller variation in the data means that the data is well separated for the levels of the categories, or that the correlation between the response and explanatory variables is larger.

2. Take a sample size of 80, and produce a random allocation of subjects to the treatments.

   Solution: This is tricky if done properly. Ideally we would like to balance the design by Sex, so that equal numbers of males and females are in each combination of Exercise and Diet. So the first step is to select 40 males and 40 females. Then at random allocate 10 males and 10 females to each treatment combination. This requires generating two categorical variables with four levels, where each level represents a treatment combination. So we can take
   1 = Exercise Yes, Diet Yes, 2 = Exercise Yes, Diet No, 3 = Exercise No, Diet Yes, 4 = Exercise No, Diet No. Then use the first categorical variable to allocate males to a treatment combination, and use the second for the females. (We should not use the same variable for allocating both males and females, as this is no longer random allocation to the treatment combinations.)

```
# Function to simulate data from a discrete distribution.
sampleDist = function(n) {
    sample(x = c(1, 2, 3, 4), n, replace = T, prob = c(0.25, 0.25,
        0.25, 0.25))
}
# But this does not guarantee equal numbers from each category.
# So we need to check this. First create the allocations for
# Males.
check <- c(10, 10, 10, 10)
repeat {
    allocM <- sampleDist(40)
    tab <- table(allocM)
    if (sum(as.numeric(tab == check)) == 4)
        break
}
# Similarly for females.
repeat {
    allocF <- sampleDist(40)
    tab <- table(allocF)
    if (sum(as.numeric(tab == check)) == 4)
        break
}
```

```
allocM

##  [1] 1 3 4 4 2 3 1 1 3 3 1 1 2 2 4 3 1 1 4 4 2 4 2 2 3 1 2 1 3 1
## [31] 4 3 2 4 2 4 4 3 2 3

allocF

##  [1] 2 1 1 1 2 3 4 4 1 3 2 3 2 2 3 4 1 4 4 2 2 2 3 3 3 4 2 1 2 3
## [31] 3 1 4 1 4 1 4 4 3 1
```

3. The data for this trial is available in the file `Weight.txt`. Analyse the data and report on your findings. You should include in your analysis appropriate data exploration and model checking

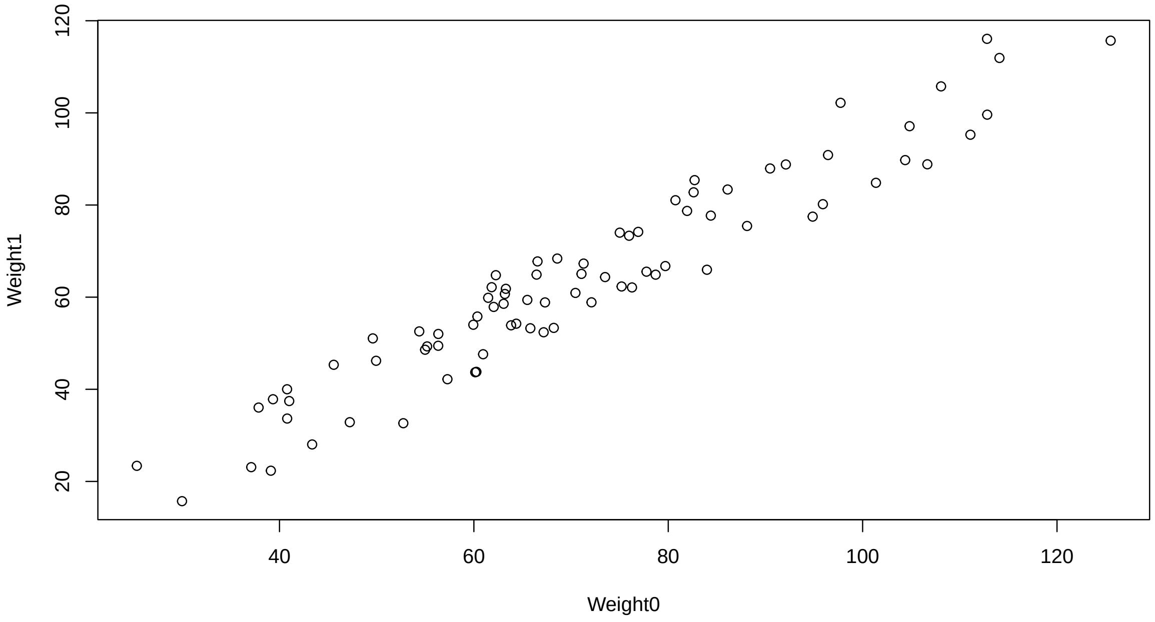
   Solution:

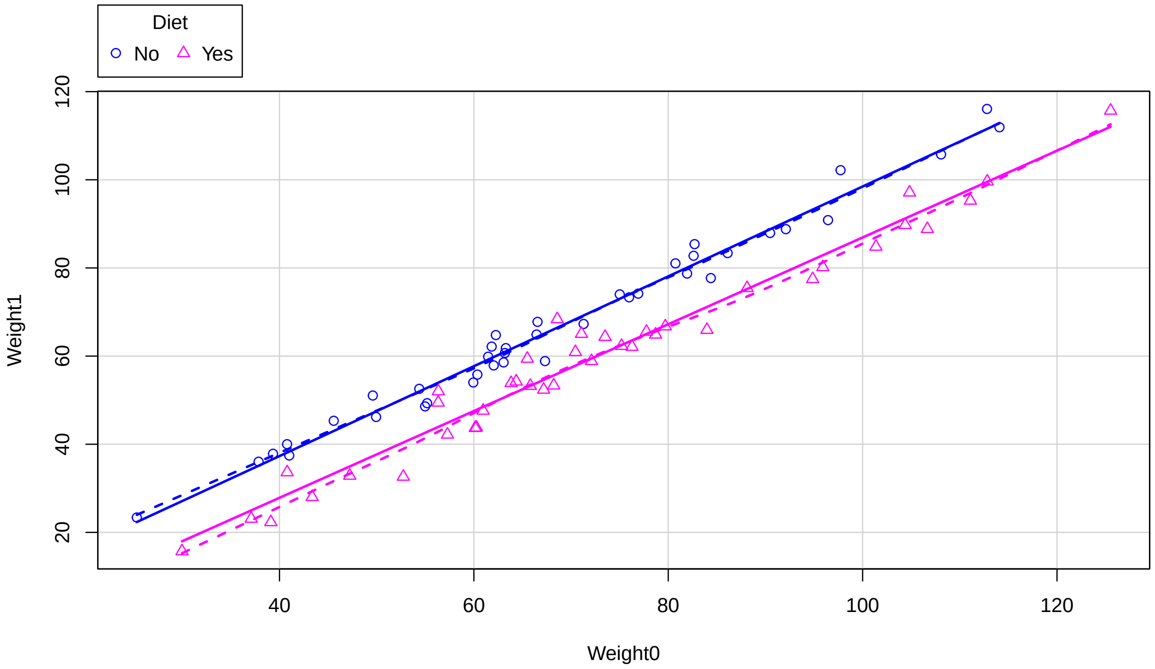
```
weight <- read.table("../Data/Weight.txt", sep = "\t", header = T,
    stringsAsFactors = T)
summary(weight)

##  Exercise  Diet    Sex         Age            Height     
##  No :40   No :40   F:40   Min.   :26.90   Min.   :155.3  
##  Yes:40   Yes:40   M:40   1st Qu.:35.48   1st Qu.:161.7  
##                           Median :39.95   Median :164.3  
##                           Mean   :42.58   Mean   :164.5  
##                           3rd Qu.:47.62   3rd Qu.:168.2  
##                           Max.   :78.60   Max.   :176.4  
##     Weight0          Weight1      
##  Min.   : 25.32   Min.   : 15.72  
##  1st Qu.: 57.05   1st Qu.: 49.41  
##  Median : 67.25   Median : 61.96  
##  Mean   : 70.88   Mean   : 63.49  
##  3rd Qu.: 83.03   3rd Qu.: 77.95  
##  Max.   :125.52   Max.   :116.07  
```

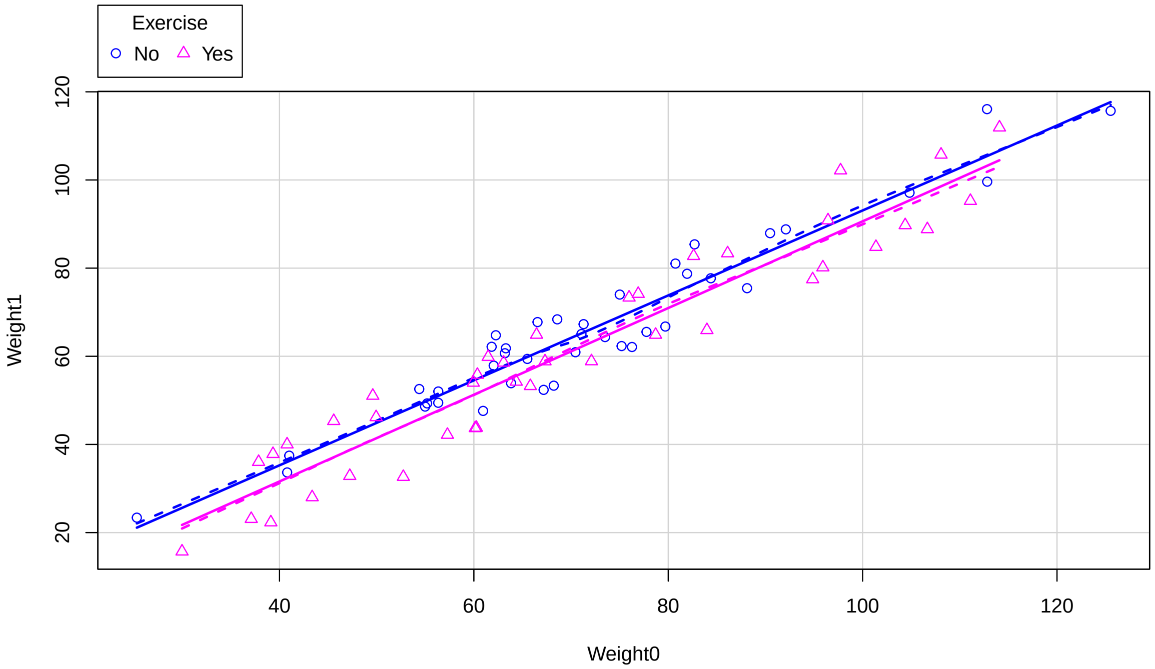
For brevity I will only produce a few plots and tables here, but you should produce all cross-tables and plots to examine the data for any anomalies and patterns.

```
## Create weight loss variable
weight$Wloss <- weight$Weight1 - weight$Weight0
attach(weight)

## The following objects are masked from weight (pos = 12):
## 
##     Age, Diet, Exercise, Height, Sex, Weight0, Weight1,
##     Wloss

table(Sex, Exercise)

##    Exercise
## Sex No Yes
##   F 18  22
##   M 22  18

table(Sex, Diet)

##    Diet
## Sex No Yes
##   F 21  19
##   M 19  21

tapply(Weight1, list(Exercise, Diet), mean)
```

```
##          No     Yes
## No  67.1665 65.2495
## Yes 66.6380 54.8945

# Define favourite statistics function.
favstats <- function(x) c(Mean = mean(x), N = length(x), StDev = sd(x),
    Min = min(x), Max = max(x))

tapply(Weight0, Sex, favstats)

## $F
##      Mean         N     StDev       Min       Max
##  71.16625  40.00000  21.36212  37.09000 125.52000
##
## $M
##      Mean         N     StDev       Min       Max
##  70.59450  40.00000  22.46951  25.32000 112.82000

tapply(Weight1, Sex, favstats)

## $F
##      Mean         N     StDev       Min       Max
##  63.90075  40.00000  21.96608  22.33000 115.68000
##
## $M
##     Mean        N    StDev      Min      Max
##  63.0735  40.0000  22.8487  15.7200 116.0700

tapply(Weight1, Exercise, favstats)

## $No
##      Mean         N     StDev       Min       Max
##  66.20800  40.00000  19.68125  23.39000 116.07000
##
## $Yes
##      Mean         N     StDev       Min       Max
##  60.76625  40.00000  24.54354  15.72000 111.92000

tapply(Weight1, Diet, favstats)

## $No
##      Mean         N     StDev       Min       Max
##  66.90225  40.00000  21.53528  23.39000 116.07000
##
## $Yes
##     Mean        N    StDev      Min      Max
##  60.0720  40.0000  22.7427  15.7200 115.6800

tapply(Wloss, list(Exercise, Diet), mean)

##          No      Yes
## No  -1.9075  -9.8980
## Yes -2.3420 -15.4255

plot(Weight1 ~ Weight0)
library(car)
scatterplot(Weight1 ~ Weight0 | Diet)
scatterplot(Weight1 ~ Weight0 | Exercise)
```


Weight1
20
40
60
80
100
120
40
60
80
100
120
Weight0


Diet
○ No
△ Yes
Weight1
20
40
60
80
100
120
40
60
80
100
120
Weight0



A linear model seems appropriate. There does seem to be a difference in mean weight loss by Exercise and Diet level. In particular, from the cross-table of mean weight loss, Diet seems to have a larger effect. Without Exercise, the diet group has an average weight loss of 9.90 kg, while with exercise the diet group has an average weight loss of 15.43 kg.

```
wt.lm <- lm(Weight1 ~ Weight0 + Sex + Age + Height + Exercise * Diet,
    data = weight)
summary(wt.lm)
```

```
##
## Call:
## lm(formula = Weight1 ~ Weight0 + Sex + Age + Height + Exercise *
##     Diet, data = weight)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.8453 -2.1442 -0.1162  1.8327  9.9400
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -19.53537   12.91657  -1.512 0.134803
## Weight0               0.99656    0.01582  62.981  < 2e-16 ***
## SexM                 -0.22245    0.69693  -0.319 0.750511
## Age                  -0.04949    0.03443  -1.438 0.154870
## Height                0.12143    0.07664   1.584 0.117496
## ExerciseYes          -0.37763    0.95983  -0.393 0.695165
## DietYes              -7.68150    0.98706  -7.782 3.91e-11 ***
## ExerciseYes:DietYes  -5.36851    1.42200  -3.775 0.000326 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.009 on 72 degrees of freedom
## Multiple R-squared:  0.9834,Adjusted R-squared:  0.9818
## F-statistic: 608.2 on 7 and 72 DF,  p-value: < 2.2e-16
```

We reduce the model. One guide for large models is to use stepwise function.

```
library(MASS)
stepAIC(wt.lm)
```

```
## Start:  AIC=183.81
## Weight1 ~ Weight0 + Sex + Age + Height + Exercise * Diet
##
##                 Df Sum of Sq   RSS    AIC
## - Sex            1         1   653 181.92
## <none>                         652 183.81
## - Age            1        19   670 184.07
## - Height         1        23   674 184.55
## - Exercise:Diet  1       129   781 196.26
## - Weight0        1     35906 36557 503.97
##
## Step:  AIC=181.92
## Weight1 ~ Weight0 + Age + Height + Exercise + Diet + Exercise:Diet
##
##                 Df Sum of Sq   RSS    AIC
## <none>                         653 181.92
## - Age            1        19   671 182.18
## - Height         1        25   678 182.96
## - Exercise:Diet  1       129   782 194.37
## - Weight0        1     35956 36609 502.08
##
## Call:
## lm(formula = Weight1 ~ Weight0 + Age + Height + Exercise + Diet +
##     Exercise:Diet, data = weight)
##
## Coefficients:
##         (Intercept)              Weight0                  Age
##           -20.38474              0.99672             -0.04941
##              Height          ExerciseYes              DietYes
##             0.12589             -0.39242             -7.72925
## ExerciseYes:DietYes
##            -5.28970
```

The procedure omits Sex from the model but then stops. We can go further and omit Age and Height.

```
# Omit Sex
wt.lm1 <- update(wt.lm, . ~ . - Sex)
summary(wt.lm1)

##
## Call:
## lm(formula = Weight1 ~ Weight0 + Age + Height + Exercise + Diet +
##     Exercise:Diet, data = weight)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7415 -2.1021 -0.1429  1.9340  9.8800
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -20.38474   12.56149  -1.623 0.108944
## Weight0               0.99672    0.01572  63.417  < 2e-16 ***
## Age                  -0.04941    0.03421  -1.444 0.152950
## Height                0.12589    0.07489   1.681 0.097027 .
## ExerciseYes          -0.39242    0.95280  -0.412 0.681650
## DietYes              -7.72925    0.96964  -7.971 1.59e-11 ***
## ExerciseYes:DietYes  -5.28970    1.39176  -3.801 0.000297 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.99 on 73 degrees of freedom
## Multiple R-squared:  0.9833,Adjusted R-squared:  0.982
## F-statistic: 718.4 on 6 and 73 DF,  p-value: < 2.2e-16
```

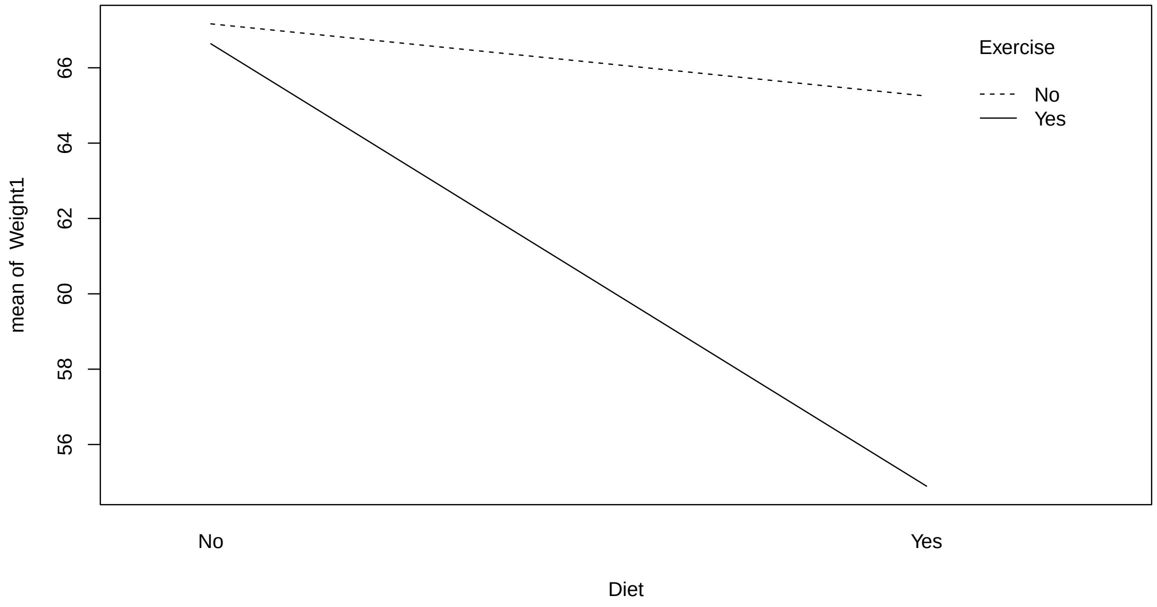
```r
# Omit Age
wt.lm2 <- update(wt.lm1, . ~ . - Age)
summary(wt.lm2)
```

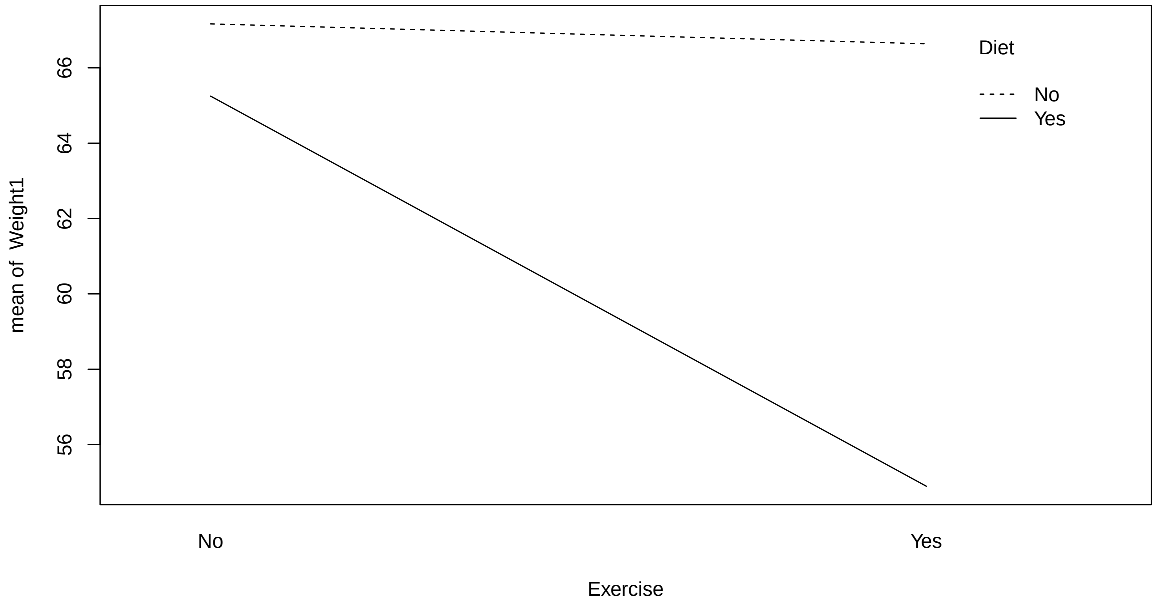
```
##
## Call:
## lm(formula = Weight1 ~ Weight0 + Height + Exercise + Diet + Exercise:Diet,
##     data = weight)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3365 -2.0158 -0.1406  1.7012 10.1518
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -23.72732   12.43669  -1.908 0.060291 .
## Weight0               0.99340    0.01566  63.429  < 2e-16 ***
## Height                0.13540    0.07515   1.802 0.075641 .
## ExerciseYes          -0.53992    0.95423  -0.566 0.573230
## DietYes              -8.00440    0.95769  -8.358 2.72e-12 ***
## ExerciseYes:DietYes  -4.80548    1.36065  -3.532 0.000715 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.012 on 74 degrees of freedom
## Multiple R-squared:  0.9829,Adjusted R-squared:  0.9817
## F-statistic: 849.2 on 5 and 74 DF,  p-value: < 2.2e-16
```

```r
# Omit Height.
wt.lm3 <- update(wt.lm2, . ~ . - Height)
summary(wt.lm3)
```

```
##
## Call:
## lm(formula = Weight1 ~ Weight0 + Exercise + Diet + Exercise:Diet,
##     data = weight)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1293 -2.1595 -0.1038  1.7276  9.6532
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -1.43618    1.29325  -1.111 0.270322
## Weight0              0.99318    0.01589  62.487  < 2e-16 ***
## ExerciseYes         -0.43514    0.96662  -0.450 0.653888
## DietYes             -7.94906    0.97143  -8.183 5.35e-12 ***
## ExerciseYes:DietYes -5.12530    1.36907  -3.744 0.000353 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.057 on 75 degrees of freedom
## Multiple R-squared:  0.9821,Adjusted R-squared:  0.9812
## F-statistic:  1030 on 4 and 75 DF,  p-value: < 2.2e-16
```

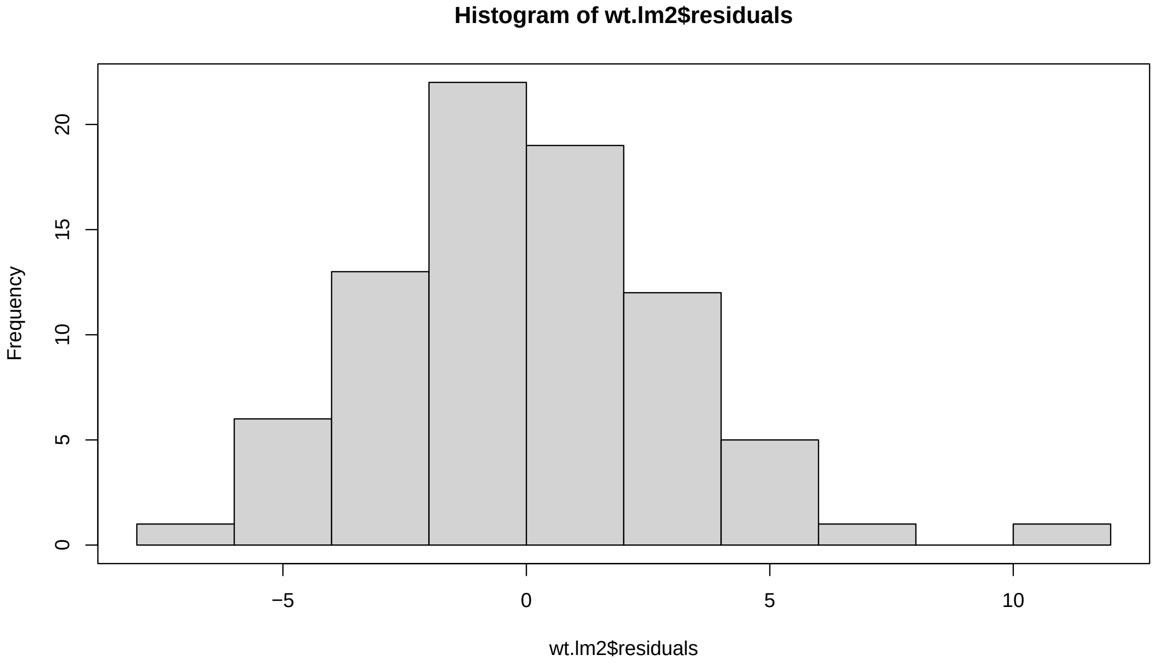
```r
interaction.plot(Diet, Exercise, Weight1)
interaction.plot(Exercise, Diet, Weight1)
```

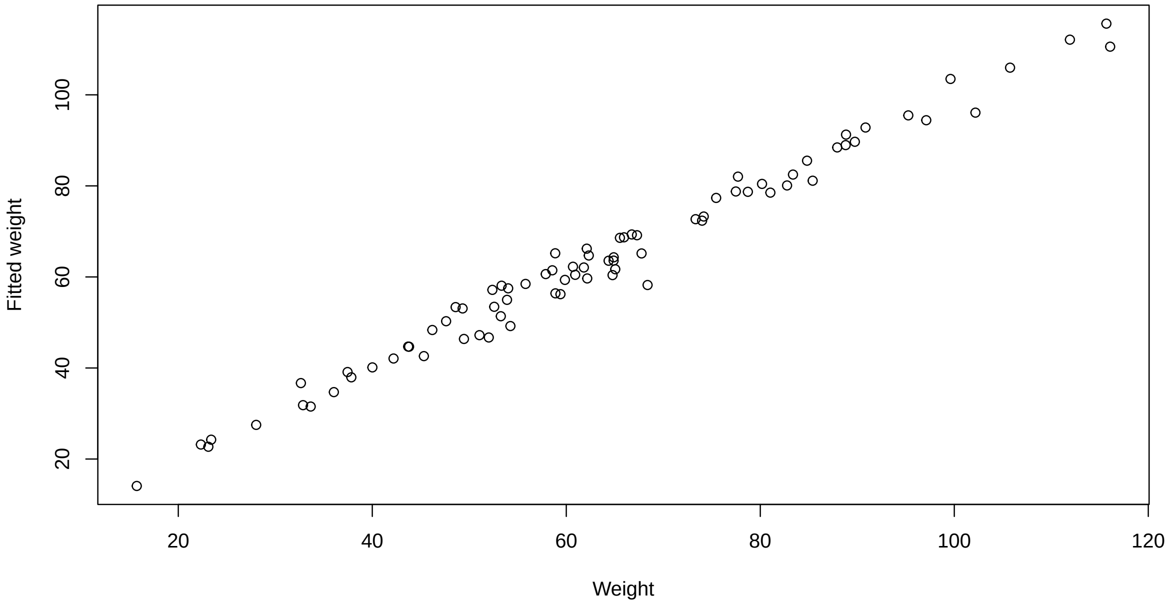


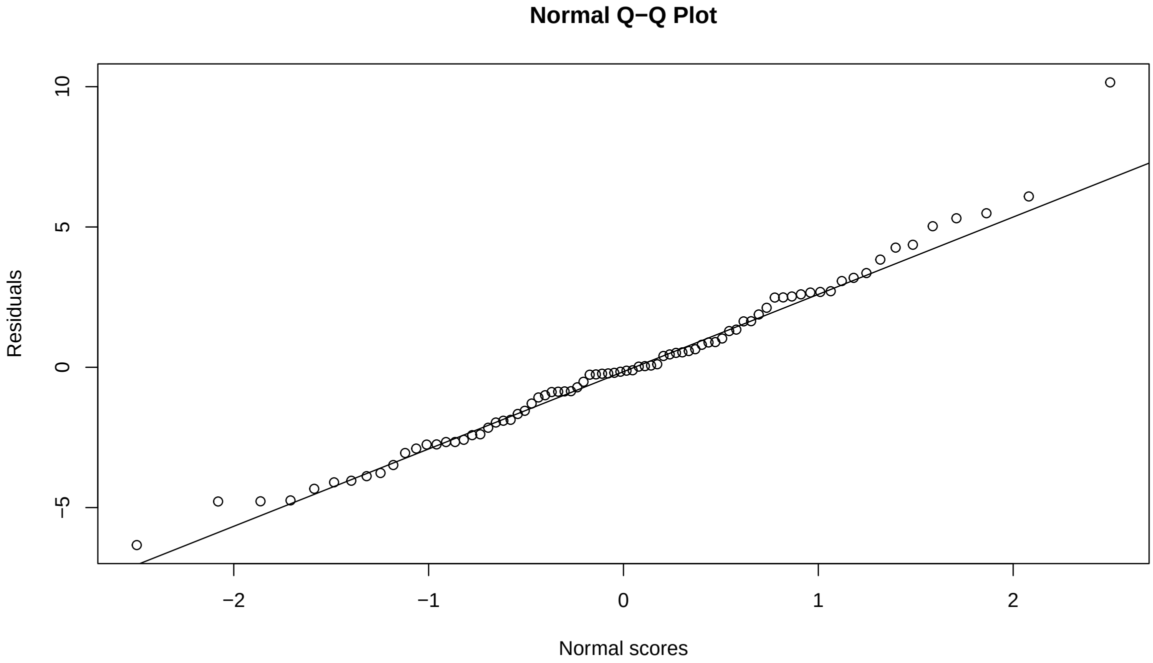


The final model indicates that: Exercise and Diet together result in a mean weight loss of $(7.95 + 5.13) = 13.08$ kg; for a person who diets but does not exercise the mean weight loss is 7.94 kg; for no exercise and no diet the final weight is almost the same as the initial weight (note the coefficient of Weight0).

```
hist(wt.lm2$residuals)
box()
plot(wt.lm2$fitted.values ~ Weight1, xlab = "Weight", ylab = "Fitted weight")
qqnorm(wt.lm2$residuals, xlab = "Normal scores", ylab = "Residuals")
qqline(wt.lm2$residuals)
```


Histogram of wt.lm2$residuals
Frequency
0
5
10
15
20
−5
0
5
10
wt.lm2$residuals


Fitted weight
20
40
60
80
100
20
40
60
80
100
120
Weight



Based on the model diagnostics we conclude there is no evidence against the assumptions of the linear model.

Extra
On further data exploration, it appears that the minimum weight is 25.32 kg for 30 year old male in the No diet and No Exercise group. This person has a weight loss of 1.93 kg. This data record is not credible. Other similar checks for data should be conducted, not just in this example but routinely.

Exercise 2: Binomial Problems

1. On Pingelap Island, 10% of the population is colour blind. A researcher selects 50 people at random from the island. Let the random variable X denote the number of people, out of the 50, who are colour blind.

   (a) State the distribution of the random variable X.
   Solution:
   $X \sim \text{Bin}(50, 0.10)$.

   (b) Determine the probability that of these 50 people, exactly 7 are colour blind.
   Solution: $\text{P}(X = 7) = 0.1076$ (4 d.p., from R).

   (c) Determine also the following:
      i. $\text{P}(X = 4)$; Solution: 0.1809 (4 d.p., from R)
      ii. $\text{P}(X \leq 6)$. Solution:0.7702 (4 d.p., from R)

   (d) What are the mean and variance of the random variable X?
   Solution: Since X is a Binomial random variable,

$$\begin{aligned}\mathbb{E}[X] &= np = 50 \times 0.1 = 5\\ \text{Var}(X) &= np(1-p) = 50 \times 0.1 \times 0.9 = 4.5.\end{aligned}$$

2. Let X be the number of heads from 10 tosses of a fair coin. Evaluate the following probabilities:

   (a) $\text{P}(X = 5)$; Solution: From Tables, $\text{P}(X = 5) = 0.2461$ (4 d.p.). Alternatively,

$$\text{P}(X = 5) = \binom{10}{5} 0.5^5 (1 - 0.5)^{10-5} = 0.2461 \text{ (4 d.p.)}.$$

(b) $\mathrm{P}(X > 7)$; Solution: $\mathrm{P}(X > 7) = 1 - \mathrm{P}(X \leq 7) = 1 - 0.9453 = 0.0547$ (4 d.p., Tables).

(c) $\mathrm{P}(3 \leq X \leq 8)$. Solution: $\mathrm{P}(3 \leq X \leq 8) = \mathrm{P}(X \leq 8) - \mathrm{P}(X < 3) = \mathrm{P}(X \leq 8) - \mathrm{P}(X \leq 2) = 0.9893 - 0.0547 = 0.9346$ (4 d.p., from Tables).

3. A botanist researching flower bulbs knows that 90% of its bulbs will flower. They are sold in packets of 12 randomly selected bulbs with a guarantee that the packet will be replaced if 100% flowering is not achieved.

(a) What is the probability that it will be necessary to replace a given packet under this guarantee? Interpret this probability.

(b) What would be the probability of replacing a packet if the guarantee covered only at least 10 out of 12 bulbs flowering? Comment on your findings in parts (a) and (b).

Solution: Let X denote the number of bulbs flowering out of 12. Then $X \sim \mathrm{Bin}(12, 0.9)$.

(a) $P(X \leq 11) = 1 - \binom{12}{12}(0.9)^{12}(0.1)^0 = 0.7175$ (by calculation or `R`).
In the long run, 71.75% of the packets are replaced (surely the business will not survive!).

(b) $P(X \leq 9) = 1 - \binom{12}{12}(0.9)^{12}(0.1)^0 - \binom{12}{11}(0.9)^{11}(0.1)^1 - \binom{12}{10}(0.9)^{10}(0.1)^2 = 0.1109$.
In the long run, 11.09% of the packets are replaced (still very high!).

Exercise 3: Poisson Problems

1. While John is in his office, he receives 4 phone calls per hour on average. Assume that the number of calls within any interval of time is Poisson distributed.

(a) What is the probability that the phone rings at least 4 times between 10am and 11am? Solution: Let X be the number of calls received during one hour. Then X is a Poisson random variable with parameter $4 \times 1 = 4$, and the required probability is

$$\begin{aligned} \mathrm{P}(X \geq 4) &= 1 - \mathrm{P}(X \leq 3) \\ &= 1 - 0.4335 \qquad \text{(from Tables)} \\ &= 0.5665. \end{aligned}$$

(b) If John takes a 30 minute lunch break, what is the probability that the phone does not ring during that time? Solution: 30 minutes = 0.5 hours; the number of calls received in any 30 minute interval is a Poisson random variable with parameter $4 \times 0.5 = 2$. Letting $X \sim \mathrm{Poisson}(2)$, the required probability is $\mathrm{P}(X = 0)$:

$$\mathrm{P}(X = 0) = e^{-2} \cdot \frac{2^0}{0!} = e^{-2} = 0.1353 \text{ (4 d.p.).}$$

(c) What is the expected number of times that the phone will ring during John's lunch break? What is the variance? Solution: $\mathbb{E}[X] = \mathrm{var}(X) = 2$.

(d) If John arrives at work at 9 am and leaves at 5 pm, what is the expected number of times that the phone will ring during the day? What is the variance? Solution: Since this is an 8 hour interval, the mean and variance are $\mathbb{E}[X] = \mathrm{var}(X) = 4 \times 8 = 32$.

2. Hummingbirds arrive at a flower at a rate λ per hour.

(a) How many visits are expected in x hours of observation?

(b) What is the variance of the number of visits in x hours?

(c) If significantly more variance is observed than expected, what might this tell you about hummingbird visits?

Solution: Let X = number of visits in x hours. Then $X \sim Poisson(\lambda x)$

(a) $\mathbb{E}(X) = \lambda x$

(b) $\text{Var}(X) = \lambda x$

(c) The mean and variance for a Poisson distribution are the same. If the variance is much larger than the mean then the distribution is unlikely to be Poisson.

3. Let $Y \sim \text{Poi}(6)$. Without using `R`, find:

(a) $\text{P}(Y \geq 3)$; Solution: $\text{P}(Y \geq 3) = 1 - \text{P}(Y \leq 2) = 1 - 0.0620 = 0.9380$ (4 d.p., Tables).

(b) $\text{P}(Y \leq 15)$; Solution: $\text{P}(Y \leq 15) = 0.9995$ (4 d.p., Tables).

(c) $\text{P}(3 \leq Y \leq 15)$. Solution:

$$\begin{aligned} \text{P}(3 \leq Y \leq 15) &= \text{P}(\{Y \geq 3\} \cap \{Y \leq 15\}) \\ &= \text{P}(\{Y \geq 3\}) + \text{P}(\{Y \leq 15\}) - \text{P}(\{Y \geq 3\} \cup \{Y \leq 15\}) \\ &= 0.9380 + 0.9995 - 1 = 0.9375. \end{aligned}$$

4. Bacteria are spread across a plate at an average density of 1000 per square cm.

(a) What is the chance of seeing no bacteria in the viewing field of a microscope if this viewing field is 4×10^{-4} square cm?

(b) What is therefore the probability of seeing at least one bacterium cell?

Solution: Let X = number of bacteria cells in a 4×10^{-4} square cm area. Then the mean number of bacteria in this area are $4 \times 10^{-4} \times 1000 = 4 \times 10^{-1} = 0.4$, and $X \sim \text{Pois}(0.4)$.

(a) $P(X = 0) = \frac{e^{-0.4}}{0.4^0 \times 0!} = 0.6703$.

(b) $P(X \geq 1) = 1 - P(X = 0) = 1 - 0.6703 = 0.3297$ (From `R`)

5. A firm collects large quantities of data. Occasionally, typing errors cause data to be incorrectly entered. The number of typing errors per 20 pages of data is a Poisson random variable with mean 3.

(a) What is the probability of there being 10 or more typing errors in 40 pages of data? Solution: Let X denote the number of typing errors in 40 pages of data. Then since $3 \times \frac{40}{20} = 6$, $X \sim \text{Poisson}(6)$. The required probability is

$$\text{P}(X \geq 10) \;=\; 1 - \text{P}(X \leq 9) = 1 - 0.9161 = 0.0839 \text{ (4 d.p.; Tables)}.$$

(b) What is the probability of there being between 5 and 9 (inclusive) typing errors in 40 pages of data? Solution: The required probability is

$$\begin{aligned} \text{P}(5 \leq X \leq 9) &= \text{P}(X \leq 9) - \text{P}(X \leq 4) \\ &= 0.9161 - 0.2851 = 0.6310 \text{ (4 d.p.; Tables)}. \end{aligned}$$

(c) What is the probability of there being less than 5 typing errors in 20 pages of data? Solution: Let Y be the number of typing errors in 20 pages of data; it is known that $Y \sim \text{Poisson}(3)$. The required probability is

$$\text{P}(Y < 5) \;=\; \text{P}(Y \leq 4) = 0.8153 \text{ (4 d.p.; Tables)}.$$

(d) What is the mean number of typing errors in 40 pages of data? What is the variance? Solution: The required mean and variance are $\mathbb{E}[X] = \text{var}(X) = 6$.

Reminder: Logging Off:
When you have finished, close down RStudio. Remember to log off from your computer before leaving.