REPORT & ASSIGNMENTS

Statistical Data Analysis with R
(AGFO-301)

Spring 2020

Diana Crowe

Student nr.: 012056152

Analyses of Data Sets

I have divided the treatment of the problem(s) into the following:

- tab name and (short) description (all data sets are tabs from the file Rkurssi1.xlsx)
- Problem(s) to be solved in the assignment
- Conclusions – aka solution(s) to the problem(s)
- Analysis – with printouts or screenshots of data (charts, tables…) and comments

Assignment 1:

Data tab name and description: Smoking - Relationship between smoking and lung cancer
Problem: Smoking data: Examine with cross table analysis the association between variables smoking and cancer.
Conclusions: Based upon Fisher's test we can reject the Null Hypothesis (p = 7.82e-46), meaning that there is an association between the variables smoking and lungcancer. In our crosstable (see below) we can see the counts of cancer for people who don't smoke and people who smoke (observed frequencies). **The amount of non-smokers who got lung cancer** (30.357%) **is a lot lower than the number of smokers who got lung cancer** (75.000%). The test of proportions confirms that these percentages are statistically dissimilar.

Analysis:

In this data file we have only two variables (smoking and lungcancer). They are both dichotomous variables, which means that they can only have two values (in this case, yes and no).
When we think that there might be a cause-effect relationship between the variables, we should select the possible cause to be the column variable. In this instance, we then choose smoking as the column variable.

In crosstabulation we run several tests of independence. The most accurate test (and especially designed for 2x2 tables) is Fisher's test so that is the one we focused on. We also got results for Pearson's Chi Square test (which is problematic with a 2x2 table but good for very large amounts of data) and Pearson's Chi-square with Yates continuity correction (a bit better). They all agreed on rejecting the Null Hypothesis.

We also do a test of proportions with Yates continuity correction to check whether the percentual differences between people who go lung cancer who smokes and who didn't smoke were significant. We got a p-value < 2.2e-16, which rejects the Null Hypothesis (that the percentages would be statistically similar).

```
   Cell Contents
|-------------------------|
|                   Count |
|         Expected Values |
| Chi-square contribution |
|          Column Percent |
|-------------------------|

Total Observations in Table:  1000

                    | smoking$smoking
smoking$lungcancer |        no |       yes | Row Total |
-------------------|-----------|-----------|-----------|
                no |       390 |       110 |       500 |
                   |   280.000 |   220.000 |           |
                   |    43.214 |    55.000 |           |
                   |    69.643% |    25.000% |           |
-------------------|-----------|-----------|-----------|
               yes |       170 |       330 |       500 |
                   |   280.000 |   220.000 |           |
                   |    43.214 |    55.000 |           |
                   |    30.357% |    75.000% |           |
-------------------|-----------|-----------|-----------|
      Column Total |       560 |       440 |      1000 |
                   |    56.000% |    44.000% |           |
-------------------|-----------|-----------|-----------|


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  196.4286     d.f. =  1     p =  1.256701e-44

Pearson's Chi-squared test with Yates' continuity correction
------------------------------------------------------------
Chi^2 =  194.6469     d.f. =  1     p =  3.07661e-44


Fisher's Exact Test for Count Data
------------------------------------------------------------
Sample estimate odds ratio:  6.865775

Alternative hypothesis: true odds ratio is not equal to 1
p =  7.820319e-46
95% confidence interval:  5.146937 9.2105

Alternative hypothesis: true odds ratio is less than 1
p =  1
95% confidence interval:  0 8.796262

Alternative hypothesis: true odds ratio is greater than 1
p =  3.91016e-46
95% confidence interval:  5.380317 Inf



       Minimum expected frequency: 220



2-sample test for equality of proportions with continuity
        correction

data:  c(170, 330) out of c(560, 440)
X-squared = 194.65, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.5040208 -0.3888364
sample estimates:
   prop 1    prop 2
0.3035714 0.7500000
```

Assignment 2:

Data tab name and description:
Heart – heart data

Problem:
Heart data: calculate Pearson correlations for variables age, both blood pressures, height and weight. Make also graphical presentation with scatter plots in one picture (matrix plot).

Conclusions:
- The values for the different Pearson correlations can be seen in the table below.
- The matrix plot with all the scatter plots in one picture can be found after the table.

There is not a high correlation between most of the chosen variables. We do however see a positive correlation between weight and height (the taller you are, the heavier you are) and between weight and each of the blood pressures (heavier people have higher blood pressures). There is also a positive correlation between the two blood pressures.

Analysis:

Pearson's correlation coefficient calculation **r** is used for continuous variables (at least at the level of interval scales), ideally normally distributed.

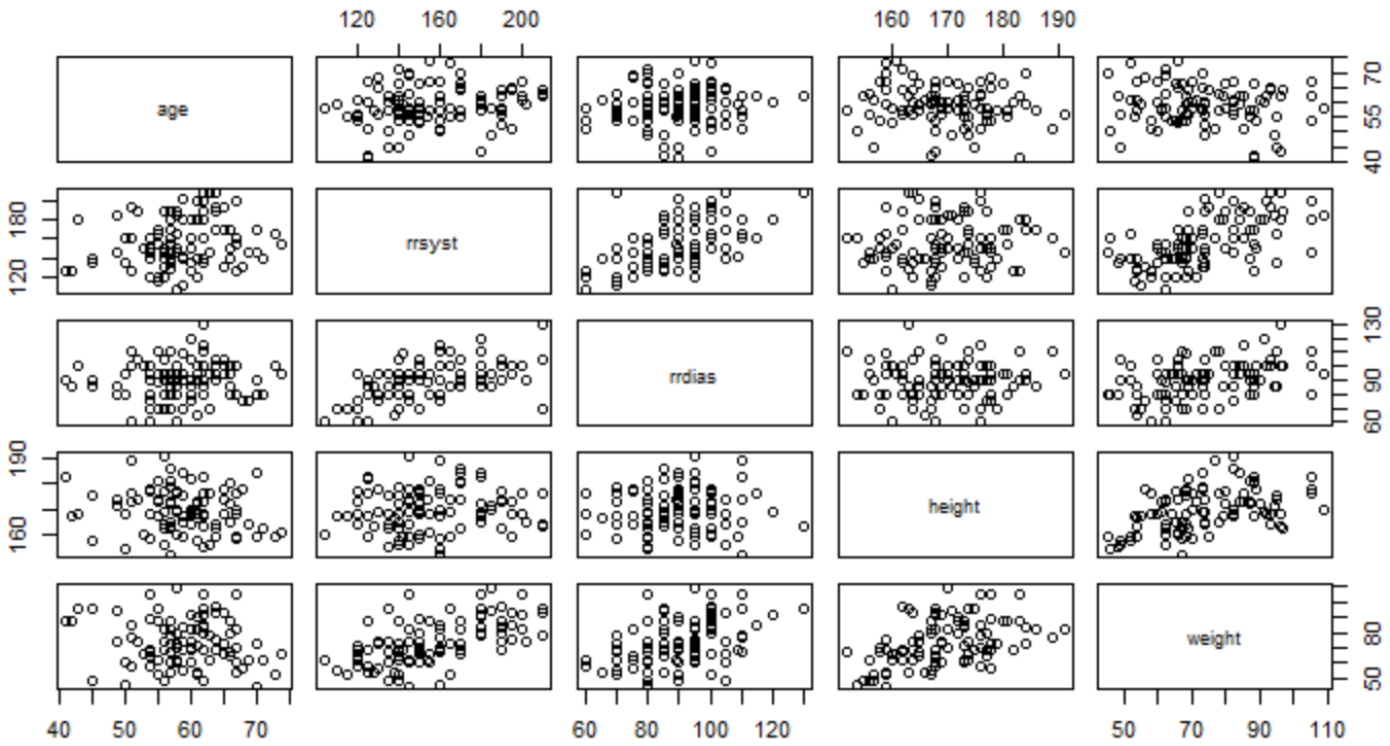
When studying correlation coefficients we should also look at the graphical output (scatter plots in this case, with 2 variables at a time). The shape of the output gives us an idea of what kind of correlation we have.

Correlations table obtained in R listing the Pearson's correlation coefficient for the variables age, systolic blood pressure, diastolic blood pressure, height and weight:

	age	rrsyst	rrdias	height	weight
age	1.00000000	0.18211602	0.07120694	-0.18180513	-0.1283974
rrsyst	0.18211602	1.00000000	0.55078257	0.07049758	0.5750153
rrdias	0.07120694	0.55078257	1.00000000	0.05215625	0.4264306
height	-0.18180513	0.07049758	0.05215625	1.00000000	0.4572624
weight	-0.12839743	0.57501528	0.42643061	0.45726244	1.0000000

R scatterplot matrix for the variables age, systolic blood pressure, diastolic blood pressure, height and weight:

Scatterplot Matrix



Assignment 3:

Data tab and description: Table 7.2 - body weight of ewes in two groups: Treatment and control group
Problem: Table 7.2: 2 independent samples, body weights of ewes, flushed and non-flushed: Examine with t-test whether the body weights of ewes are different in flushed (treatment) and non-flushed (control) groups.
Conclusions: The treatment group (flushed ewes) has a higher mean weight (67.37 Kg) than the control group (65.77 Kg).

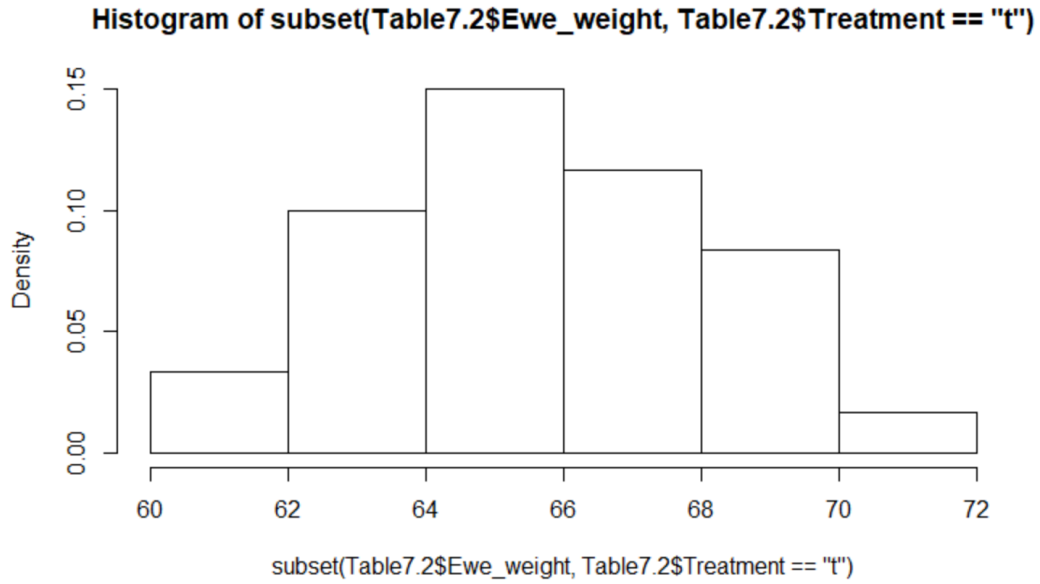
Analysis:

Since there is a small amount of data (fewer than 30 samples), we should check the normality of the distribution of ewe body weights first in order to make sure that we can use a t-test. I checked separately for each of the two groups (control and treatment):

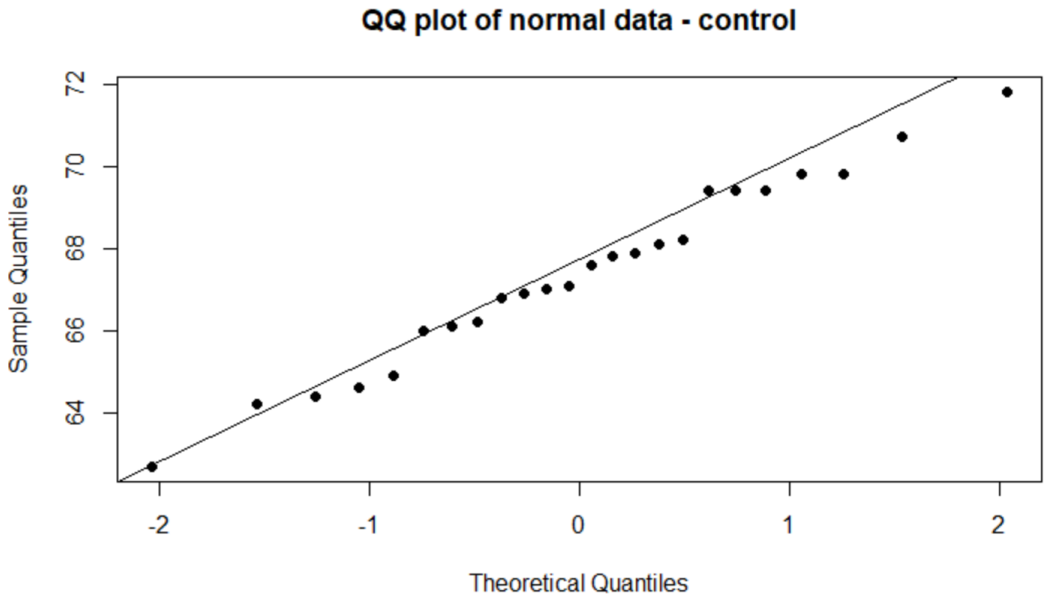
```
        Shapiro-Wilk normality test

data:  subset(Table7.2$Ewe_weight, Table7.2$Treatment == "c")
W = 0.98471, p-value = 0.9645
```

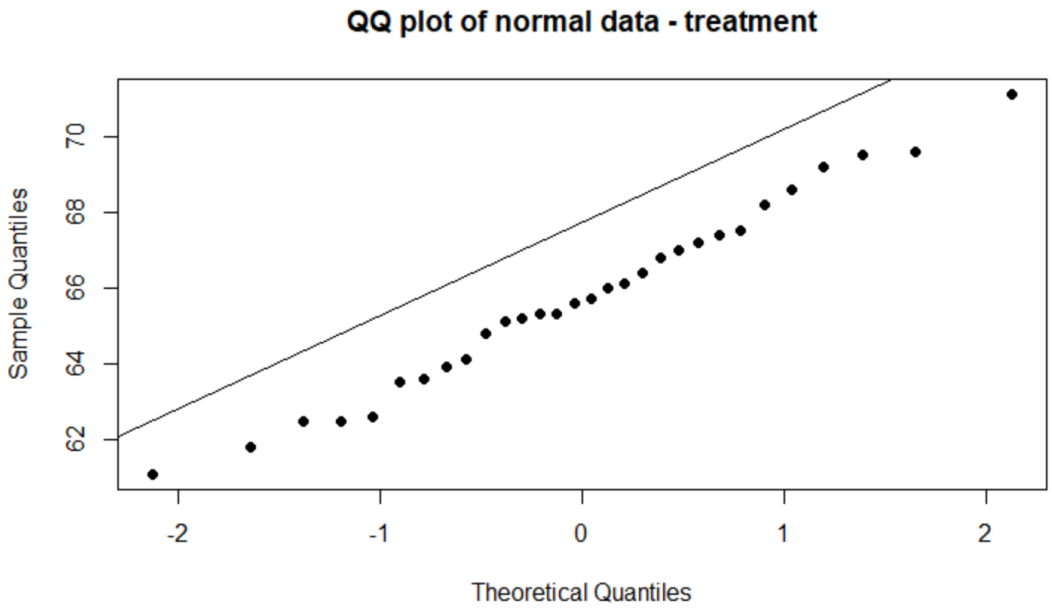
```
        Shapiro-Wilk normality test

data:  subset(Table7.2$Ewe_weight, Table7.2$Treatment == "t")
W = 0.98593, p-value = 0.952
```

From the p values in the Shapiro-Wilks test we see that both groups are normally distributed so we can use the t-test. The normality is also evident qualitatively when having a look at the corresponding histograms:



We can also do a QQ-plot to qualitatively judge normality (the closer the points are to the line, the more normally distributed they are).





Now we can get some basic statistics (we want the means):

```
> describe(subset(Table7.2$Ewe_weight, Table7.2$Treatment=="c"))
   vars  n  mean   sd median trimmed  mad  min  max range  skew  kurtosis   se
X1    1 24 67.37 2.25  67.35   67.37 2.52 62.7 71.8   9.1 -0.08     -0.75 0.46

> describe(subset(Table7.2$Ewe_weight, Table7.2$Treatment=="t"))
   vars  n  mean  sd median trimmed  mad  min  max range skew  kurtosis   se
X1    1 30 65.77 2.5  65.65   65.73 2.59 61.1 71.1    10 0.11     -0.77 0.46
```

The next step is to do Levene's test to study the variances since the way we do the t-test is different depending on whether the two groups' variances are equal or not equal. This is important when studying both independent and related samples.

```
> leveneTest(Table7.2$Ewe_weight,factor(Table7.2$Treatment),location="mean")
Levene's Test for Homogeneity of Variance (center = median: "mean")
      Df F value Pr(>F)
group  1   0.238 0.6277
      52
```

It uses Fisher's test and the value to look at in our result is the p value of 0.6277. The Null Hypothesis is that the variances are equal, which is accepted in this case. The variances are similar. In practise, this test studies the ratios of the two variances and, if the ratio is around 1, the variances are similar.

And we can now (finally!) do the t-test for the case of two independent samples (paired=False) with similar variances (var.equal=True):

```
> t.test(formula= Table7.2$Ewe_weight ~ Table7.2$Treatment,data=Table7.2,paired =F,var
.equal=T,conf.level=0.95)

        Two Sample t-test

data:  Table7.2$Ewe_weight by Table7.2$Treatment
t = 2.4323, df = 52, p-value = 0.01848
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.2788147 2.9078519
sample estimates:
mean in group c mean in group t
       67.36667        65.77333
```

We get as a result that the Alternative hypothesis is accepted which means that the two means (control and treatment) are statistically different.

We therefore must conclude that the treatment group (flushed ewes) has a higher mean weight (67.37 Kg) than the control group (65.77 Kg).

(and since R so helpfully provides the groups means here, the descriptive basic statistics analysis step was actually redundant – but good practise)

Assignment 4:

Data tab name and description: table 8.1 – teeth calculus in 3 groups of dogs (1 control group and 2 treatment groups)
Problem: Table 8.1: one-way ANOVA, 3 diets of dogs for affecting on the build-up of calculus of teeth data: Examine with parametric ANOVA whether there are differences in calculus among the diets. Perform also post hoc tests if necessary.
Conclusions: The ANOVA test tells us that there are differences in the means of at least two groups. The Post-Hoc tests tell us that those differences are between the t3 group and the t1 group (control group). Looking at the means, the control group t1 has the most amount of calculus and group t3 has the least amount of calculus. Performing non-parametric tests confirms those results.

Analysis:

This dataset is a case of many independent samples (t1 = control group; t1and t2 are treatments being tested) and one dependent variable (the calculus in the dogs' teeth). As we have 3 groups, we cannot use a t-test as our parametric test but need to use instead an ANOVA table (**AN**alysis **O**f **VA**riance) in order to avoid a type 1 error.

As a starting point we look at the **group basic statistics**:

```
> describeBy(table81$Calculus, table81$Group)

 Descriptive statistics by group
group: t1
   vars n mean   sd median trimmed  mad  min  max range  skew kurtosis   se
X1    1 9 1.09 0.42   1.05    1.09 0.46 0.49 1.66  1.17 -0.11    -1.67 0.14
------------------------------------------------------------
group: t2
   vars n mean   sd median trimmed  mad  min  max range skew kurtosis   se
X1    1 9 0.75 0.37   0.76    0.75 0.46 0.22 1.38  1.16 0.13     -1.3 0.12
------------------------------------------------------------
group: t3
   vars n mean   sd median trimmed  mad  min  max range skew kurtosis  se
X1    1 8 0.44 0.29    0.4    0.44 0.25 0.05 0.95   0.9 0.39    -1.22 0.1
```

Our goal is to compare the three group means together.

As it is a small data set (just 26 dogs in total), we need to **check whether the data is normally distributed** in order to see whether we can apply a parametric test. It's such a small dataset we may want to also do the non-parametric test and compare the two results.

```
> tapply(table81$Calculus, table81$Group,shapiro.test)
$t1

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.94091, p-value = 0.5915


$t2

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.97834, p-value = 0.9552


$t3

        Shapiro-Wilk normality test

data:  X[[i]]
W = 0.97424, p-value = 0.929
```

Looking at the p-values obtained, we accept the Null Hypothesis in all three groups, so we can conclude that the data is normally distributed for all three groups and the parametric ANOVA test should be reliable.

We then do **Levene's test to study the groupwise variances**:

```
> leveneTest(table81$Calculus,factor(table81$Group),location="mean")
Levene's Test for Homogeneity of Variance (center = median: "mean")
      Df F value Pr(>F)
group  2  0.7437 0.4864
      23
```

According to the p-value obtained, the group variances are similar because we accept the Null Hypothesis.

We can finally obtain the following **ANOVA table**:

```
> model<-aov(Calculus~factor(Group),data=table81)
> summary(model)
             Df Sum Sq Mean Sq F value Pr(>F)
factor(Group)  2  1.805  0.9023   6.668 0.0052 **
Residuals     23  3.112  0.1353
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We see that Fisher's test gives us a p-value of 0.0052 so we reject the Null Hypothesis (which was that there were no differences between the group means), which means that at least two group means are different.

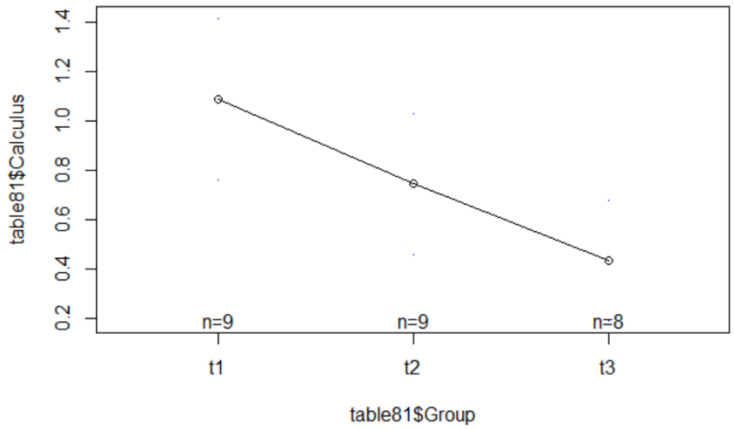
Since there are differences between at least two groups, we have to study things further to find out what kind of differences there are. We do this using **Post-Hoc tests**, which perform pairwise comparisons (comparing two group means at a time). There are many possible tests, but **we shall use Bonferroni** since it works well for a small number of groups (if 4 or more groups we might use Tukey instead).

```
> pairwise.t.test(table81$Calculus,table81$Group,p.adjust.method = "bonferroni")

        Pairwise comparisons using t tests with pooled SD

data:  table81$Calculus and table81$Group

   t1     t2
t2 0.1817 -
t3 0.0041 0.2912

P value adjustment method: bonferroni
```

We see in the pairwise comparison table that:

- for groups t1 and t2 the Null Hypothesis is accepted which means that the means for those two groups are similar.
- for groups t2 and t3 the Null Hypothesis is accepted which means that the means for those two groups are similar.
- We reject the Null Hypothesis for groups t1 and t3 – so the means for those two groups are different

We can also visualize our data for a qualitative approach and see how the means are distributed:



Conclusion: the control group has the most amount of calculus and group t3 has the least amount of calculus.

Checking what the **non-parametric** approach gives us (**Kruskal-Wallis test** because it is very good for studying 3 groups):

```
> kruskal.test(table81$Calculus ~ table81$Group,data=table81)

        Kruskal-Wallis rank sum test

data:  table81$Calculus by table81$Group
Kruskal-Wallis chi-squared = 9.1984, df = 2, p-value = 0.01006
```

This is analogous to our ANOVA table above and the p-value tells us to reject the Null Hypothesis since there are differences in at least two groups. We now do the post-hoc pairwise comparison – again using Bonferroni (paired = False because they are independent samples):

```
> pairwise.wilcox.test(table81$Calculus, table81$Group, p.adjust.method = "bonferroni"
,paired=F)

        Pairwise comparisons using Wilcoxon rank sum test

data:  table81$Calculus and table81$Group

   t1     t2
t2 0.3405 -
t3 0.0074 0.3357

P value adjustment method: bonferroni
```

When looking at the table of pairwise comparisons, we again get the conclusion that the only difference is between groups t1 and t3.

Conclusion: In our non-parametric analysis we reach the same conclusion as in the parametric analysis: the Control group t1 has the most amount of calculus and group t3 has the least amount of calculus.

Assignment 5:

Data tab name and description:

donkey – data on 386 donkeys: sex, age, body weight (in kg), heart girth, umbgirth (circumference around pelvis), length and height (in cm)

Problem:

Donkey data: Perform linear regression analysis in which bodyweight is the dependent, and the other continuous variables are independent variables. a) What is the value of Rsquare? b) Is there multicollinearity?

Conclusions:

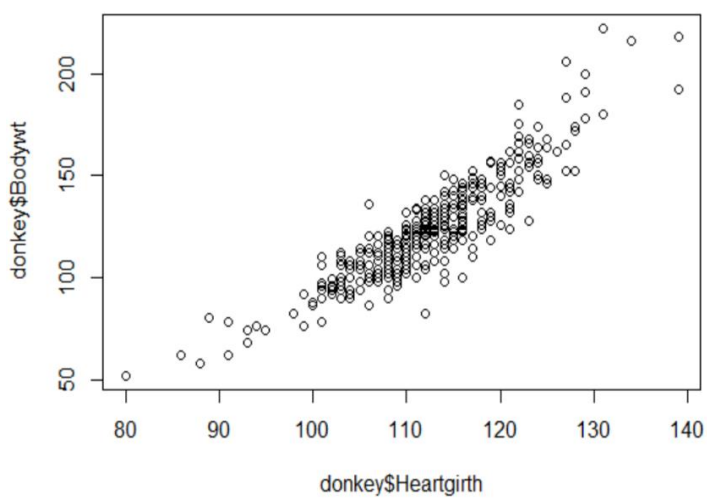
a) R square: Multiple R-squared: 0.8423, Adjusted R-squared: 0.8415

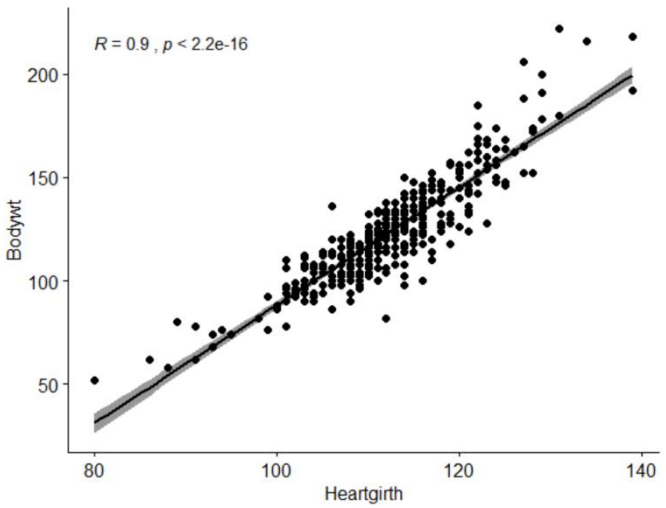
b) Multicollinearity: The values of Collinearity Tolerance for my chosen 2-variable model are half-way between zero and one (around 0.5) so there **is** some correlation – but not clear multicollinearity as that would be Tolerance values close to zero. The VIF values are well below the limit of 5 (just under 2), which tells us that the correlation is not too serious and we can choose to accept this model as it stands.

c) Regression equation: Weight = -213.04 + 2.23*Heartgirth + 1.04*Length

Analysis:

We want to predict the weight of the donkeys using other measurements since we may not always have specialized scales available to weight large animals. This is what our data looks like as a scatter plot (left plot) and as a scatter plot with a fitted regression line (right plot):





Starting point: If we do the full model using every possible independent variable, we get:

```
> model <-lm(Bodywt~Age+Heartgirth+Height+Length,data=donkey);summary(model)

Call:
lm(formula = Bodywt ~ Age + Heartgirth + Height + Length, data = donkey)

Residuals:
    Min      1Q  Median      3Q     Max
-29.731  -5.745  -0.047   5.446  44.955

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -212.7663     7.9442 -26.783  < 2e-16 ***
Age            0.4130     0.1872   2.206   0.0280 *
Heartgirth     2.0958     0.1020  20.544  < 2e-16 ***
Height         0.2449     0.1114   2.198   0.0285 *
Length         0.8769     0.1201   7.302 1.67e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.779 on 381 degrees of freedom
Multiple R-squared:  0.8465,   Adjusted R-squared:  0.8449
F-statistic: 525.2 on 4 and 381 DF,  p-value: < 2.2e-16
```

In the next-to-last line we have the goodness criteria (R square and adjusted R square).

R squared gives us multiple correlation (if many variables), or the correlation (if single). It varies from 0 to 1. Closer to 1 is better but sometimes even 0.3 is acceptable for a model (lower limit around 0.3 since below 0.3 we don't have a reasonable model). If we multiply the R squared by 100 we are then looking at the percentage of the variance of the dependent variable (weight of the donkey) that is explained by the current model. The rest of the variance is unexplained or error variance.

The model above explains almost 85% of the variance of the weight of the donkeys.

Adjusted R square is usually more reliable. The Adjusted R square decreases if irrelevant variables are added. The adjusted R squared should decrease if we add insignificant variables and increase of we add significant variables (in practise, it occasionally doesn't behave that way…).

In suitable models we should have a small p-value so that we can reject the Null Hypothesis in which the current model would be equivalent to a model with all regression coefficients equalling zero (in practise).

The t-test value studies the significance of the variable in the model (Null Hypothesis = variable is insignificant; ie. the coefficient is zero in practice). All of our variables seem to be significant in this instance.

****

We now study the **multicollinearity**.

We want to find independent variables with sufficiently high linear correlation with the dependent variable, and with no inter correlations (multi collinearity) between independent variables. We want also to use the smallest amount of independent variables possible to get a suitable but simple model.

```
> ols_vif_tol(model)# studying the collinearity
   Variables Tolerance      VIF
1        Age 0.8031579 1.245085
2 Heartgirth 0.3928135 2.545737
3     Height 0.3634864 2.751135
4     Length 0.3932648 2.542816
```

We have here two measures for studying collinearity: Collinearity Tolerance and Statistics VIF.

Tolerance ranges from 0 to 1 and if these values are very close to one (which is what we want), there is no multicollinearity (multicollinearity would be a value close to zero). In our case, we see that age has high Tolerance, but the other three variables have lower tolerances so there might be some multicollinearity.

The VIF value should be below 5, which is the case here for all our variables.

When there are only a few possible independent variables it is always best to study all the possible models:

- full model with all the variables (4 of them in our case, as shown above),
- remove one by one any insignificant variables (no insignificant variables in this case)
- all possible models with 3 variables
- all possible models with 2 variables

For the case of many variables, it is easier to use an automated procedure to compare all those different models.

In the automated procedures we can use for example the stepwise method. "Forwards" we add one variable at a time to the model; "backwards" we remove one variable at a time from the model; "bidirection" both removes and adds at the same time.

One should use at least two different methods – and if one gets similar outputs, then accept that result.

Method 1: Stepwise bidirection and Adjusted R square statistic

```
> stepwise(donkey,y="Bodywt",exclude=c("Donkey","Sex"),selection="bidirection",select=
"adjRsq")
$process
  Step EffectEntered EffectRemoved EffectNumber    Select
1    0     intercept                          1 0.0000000
2    1    Heartgirth                          2 0.8029941
3    2        Length                          3 0.8414827
4    3      Umbgirth                          4 0.8534209
5    4        Height                          5 0.8546794
6    5           Age                          6 0.8556973

$variate
[1] "intercept"  "Heartgirth" "Length"     "Umbgirth"   "Height"
[6] "Age"
```

The model started with heartgirth and when it added length we see a significant increase in the R squared. When we add each of the other variables (umbgirth, height and age) even though the R Squared increases a bit, it is a very small increase and each variable makes the model more complicated. We should keep just the two first variables in our final model: heartgirth and length.

Method 2: Stepwise forward and Adjusted R square statistic

```
> stepwise(donkey,y="Bodywt",exclude=c("Donkey","Sex"),selection="forward",select="adj
Rsq")
$process
  Step EffectEntered EffectNumber    Select
1    0     intercept            1 0.0000000
2    1    Heartgirth            2 0.8029941
3    2        Length            3 0.8414827
4    3      Umbgirth            4 0.8534209
5    4        Height            5 0.8546794
6    5           Age            6 0.8556973

$variate
[1] "intercept"  "Heartgirth" "Length"     "Umbgirth"   "Height"     "Age"
```

We have the same results and can therefore reach the same conclusions using this second method. We can use the two-variable model, which is simpler and is already quite accurate.

Final solution with the two-variable model:

We first study the collinearity:

```
> ols_vif_tol(model2) # studying the collinearity
   Variables Tolerance      VIF
1 Heartgirth 0.5036443 1.985528
2     Length 0.5036443 1.985528
```

We see that there is a slight problem with the Tolerance (a bit of correlation). Collinearity Tolerance ranges from 0 to 1 and if these values are very close to one, there is no multicollinearity (multicollinearity would be a value close to zero). These values of Collinearity Tolerance are half-way (around 0.5) so there is some correlation but no clear multicollinearity.

The VIF values are well below the limit of 5, which tells us that the correlation is not too serious and we can choose to accept this model as it stands.

Then we calculate the model (since we chose to accept a bit of multicollinearity).

```
> model2 <-lm(Bodywt~Heartgirth+Length,data=donkey);summary(model2)

Call:
lm(formula = Bodywt ~ Heartgirth + Length, data = donkey)

Residuals:
    Min      1Q  Median      3Q     Max
-29.960  -5.939   0.169   5.583  43.674

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -213.03872    7.43787 -28.642   <2e-16 ***
Heartgirth     2.23231    0.09107  24.512   <2e-16 ***
Length         1.04140    0.10728   9.708   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.885 on 383 degrees of freedom
Multiple R-squared:  0.8423,   Adjusted R-squared:  0.8415
F-statistic:  1023 on 2 and 383 DF,  p-value: < 2.2e-16
```

According to the t-test values, both of the variables chosen (heartgirth and length) seem to be significant because we can reject the Null Hypothesis (in which the model would be similar to a model with null/zero coefficients).

The equation for our model becomes:

Weight = -213.04 + 2.23*Heartgirth + 1.04*Length

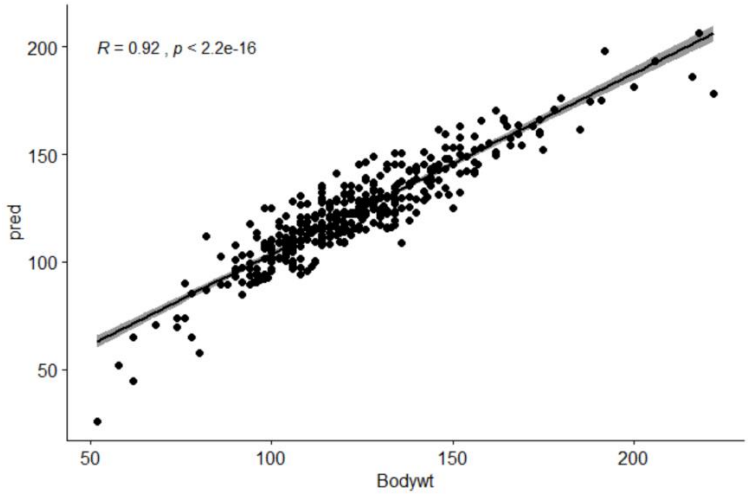
We need to still study our residuals, which should be normally distributed around zero.

We create a new variable to save the predicted values according to our model (there are some dependencies, so we need to run first the commands for predicting values according to the model where we use all of the variables):

```
> model <-lm(Bodywt~Age+Heartgirth+Height+Length,data=donkey);summary(model)
        (output already shown in page 13 above; I won't repeat it here)
> pred <- model$fitted.values;write_xlsx(data.frame(pred),"donkeypred.xlsx")
> pred2 <- model2$fitted.values;write_xlsx(data.frame(pred),"donkeypred2.xlsx")
```

We can now plot the estimated values versus the predicted values using the regression equation. We have a correlation coefficient of 0.92, which is fairly good but not the best possible (that would be a value of 1).

```
> ggscatter(cbind(donkey,pred2), x = "Bodywt", y = "pred2", add = "reg.line", conf.int
= TRUE, cor.coef = TRUE, cor.method = "pearson", xlab = "Bodywt", ylab ="pred" )
```

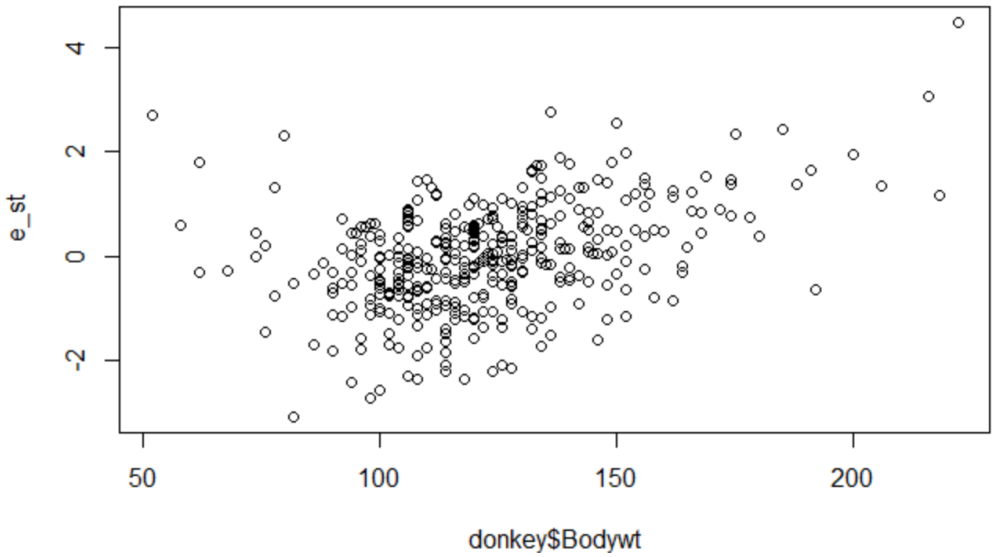


We can now create variables to hold the residuals ("e") and the corresponding standard residuals ("e_st"):

```
> e<-model2$residuals # residuals
> e_st<-rstandard(model2) #the corresponding standard residuals
```

We can plot the standard residuals:

```
> plot(donkey$Bodywt,e_st)
```



If we have any data points outside of the interval -3 to +3, then those are outliers (we can see at least one exceptionally fat donkey in the upper right corner).

When there are residuals located both above and below the zero line (as the current case), we have only random error, which is accepted to some extent.

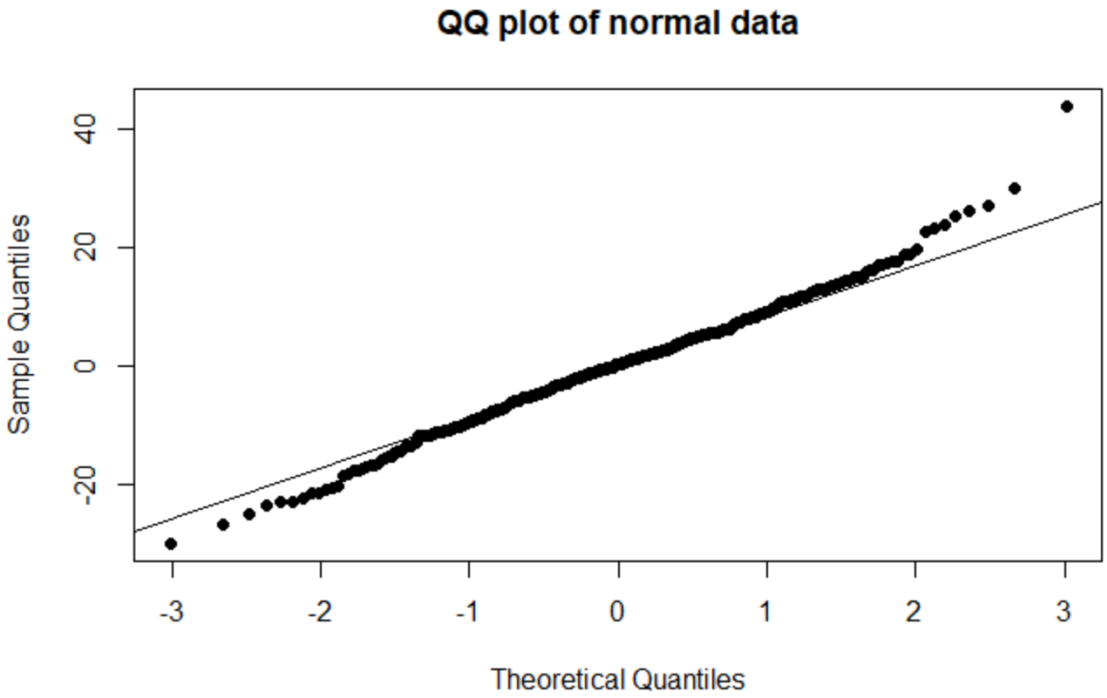
We now to the Shapiro-Willks test to see whether we have a normal distribution of residuals:

```
> shapiro.test(e)

        Shapiro-Wilk normality test

data:  e
W = 0.99088, p-value = 0.01743
```

The p-value means that we reject the Null Hypothesis (which is that there is a normal distribution), so our residuals are not normally distributed (which is bad and somewhat surprising).

We can follow the Shapiro-Wilks test with a QQ-plot. The closer the points are to the line, the more normally distributed they are.

```
> qqnorm(e,main="QQ plot of normal data",pch=19); qqline(e)
```



In the case of donkeys with smaller weights and with larger weights the points are not close to the line, which indicates a deviation from the normal distribution. For all the donkey weights in-between the situation is better.