STAT 210
Applied Statistics and Data Analysis:
Problem list 7
(Due on week 8)

Exercise 1

The dataset `anorexia` in the `MASS` package has information about 72 anorexia patients, 26 on the control (`Cont`) group, 29 in the `CBT` treatment group, and 17 in family treatment (`FT`). The data correspond to weight before and after treatment for each patient.

Do a complete analysis of variance for this set. Determine whether the treatments have an effect on the weight of the patient by means of a hypothesis test. Plot the diagnostic charts and comment on them. Use also Levene's test and Shapiro-Wilk. Use Tukey's HSD procedure to make pairwise comparisons and comment on the results.

Solution

We start by loading the `MASS` library and the data set

```
library(MASS)
str(anorexia)
```

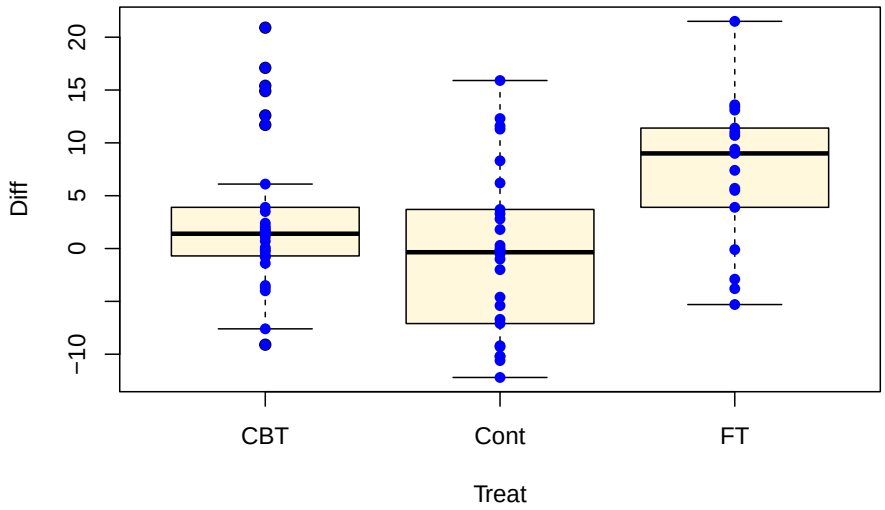
```
## 'data.frame':    72 obs. of  3 variables:
##  $ Treat : Factor w/ 3 levels "CBT","Cont","FT": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Prewt : num  80.7 89.4 91.8 74 78.1 88.3 87.3 75.1 80.6 78.4 ...
##  $ Postwt: num  80.2 80.1 86.4 86.3 76.1 78.1 75.1 86.7 73.5 84.6 ...
```

To determine whether the treatments have an effect on the weight of the patients we should look at the weight difference before and after treatment. We create a new variable `Diff` with the difference `Postwt - Prewt`.

```
anorexia$Diff <- anorexia$Postwt - anorexia$Prewt
```

We do a boxplot of the new variable as a function of treatment, to have an idea of the data.

```
plot(Diff ~ Treat, data = anorexia, col = 'cornsilk')
points(Diff ~ Treat, data = anorexia, col = 'blue', pch = 16)
```



We see that `CBT` does not seem to differ from the control group, but `FT` shows a marked difference. We need to carry out our analysis to confirm these results. Another thing that the graph shows is that the variability for the control group seems bigger than for the other two groups, particularly `CBT`.

We attach the data set and calculate means and standard deviations for the three groups.

```
attach(anorexia)
tapply(Diff, Treat, mean)
```

```
##       CBT      Cont        FT
##  3.006897 -0.450000  7.264706
```

```
tapply(Diff, Treat, sd)
```

```
##      CBT     Cont       FT
## 7.308504 7.988705 7.157421
```

The means look different but the standard deviations are similar.

We now carry out an analysis of variance for this data.

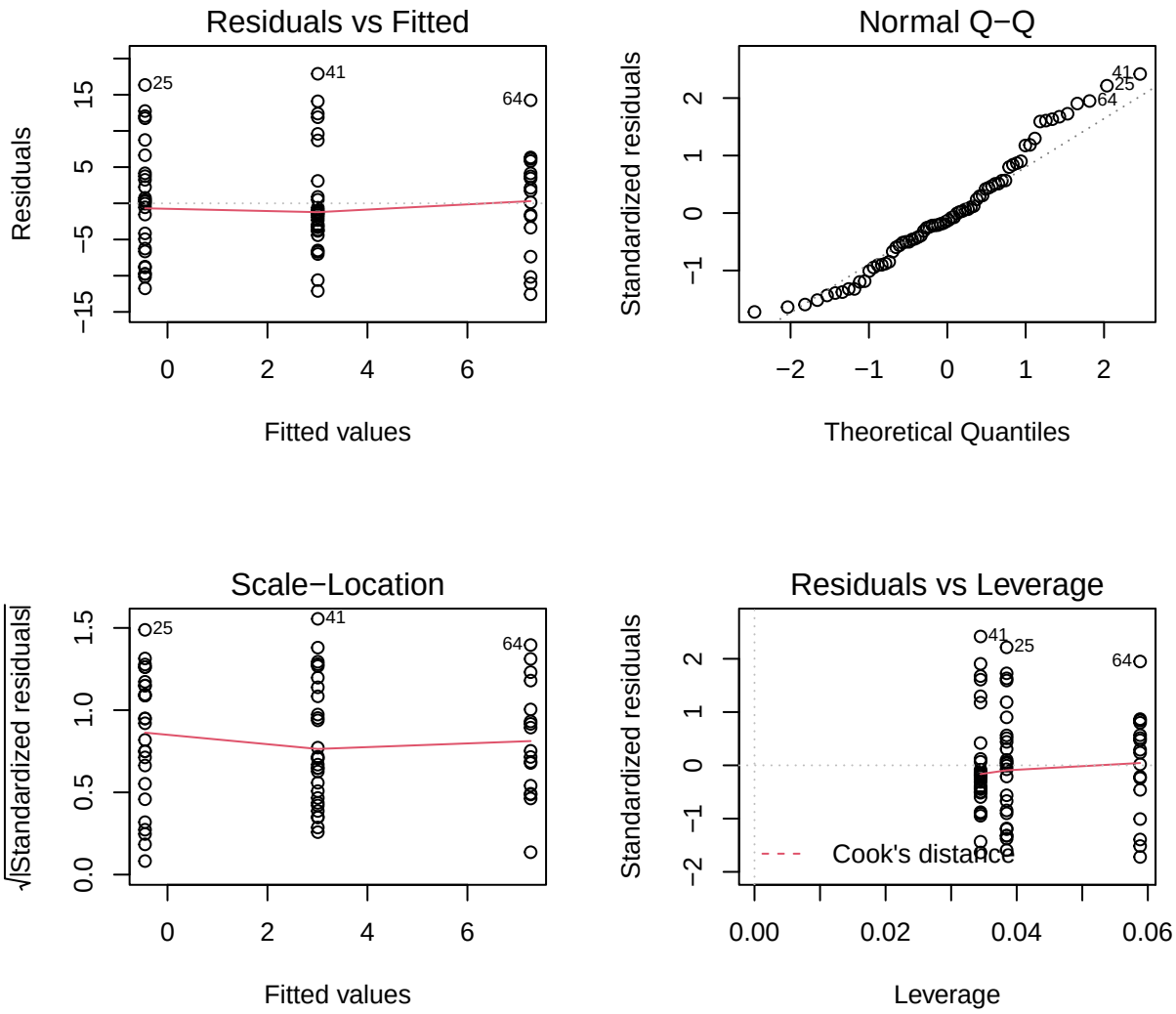
```
model1 <- aov(Diff ~ Treat, data = anorexia)
summary(model1)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Treat        2    615  307.32   5.422 0.0065 **
## Residuals   69   3911   56.68
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The effect of the treatments is significant at the usual levels since the p-value is 0.0065. This means that at least one treatment is different from another treatment.

Diagnostic plots:

```
par(mfrow=c(2,2))
plot(model1)
```




```
par(mfrow=c(1,1))
```

Perhaps the qq-plot causes some concern, since the upper tail departs from the straight line. The other plots look fine. Next, we do the Shapiro-Wilk and Levene tests for the residuals.

```
shapiro.test(resid(model1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model1)
## W = 0.96723, p-value = 0.05726
```

This shows that at the 5% level (or lower levels) we cannot reject the null hypothesis of Gaussianity.

```
library(car)
leveneTest(model1)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2  0.5207 0.5964
##       69
```

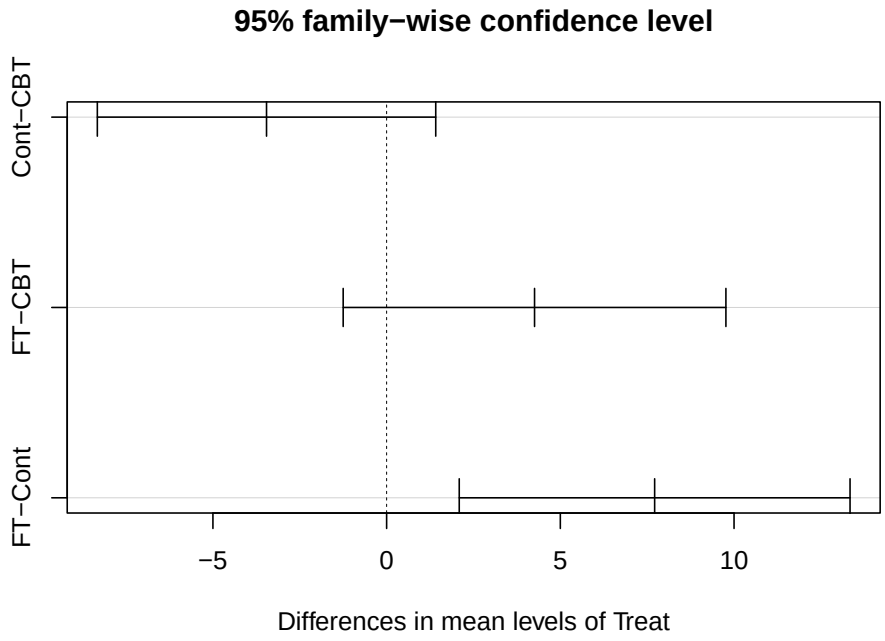
This test has a large p-value, saying that hypothesis of homoscedasticy is not rejected.

Finally, we use Tukey's honest significance difference to compare the different treatment levels.

```
(model1.tky <- TukeyHSD(model1))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Diff ~ Treat, data = anorexia)
##
## $Treat
##               diff       lwr       upr     p adj
## Cont-CBT -3.456897 -8.327276  1.413483 0.2124428
## FT-CBT    4.257809 -1.250554  9.766173 0.1607461
## FT-Cont   7.714706  2.090124 13.339288 0.0045127
```

```
plot(model1.tky)
```



We see that the only significant difference is between the control group and family treatment.

Exercise 2

In an experiment to study the effect of the amount of baking powder in a biscuit dough upon the rise heights of the biscuits, four levels of baking powder were tested and four replicate biscuits were made with each level in a random order. The results are shown in the table below.

.25 tsp	11.4	11.0	11.3	9.5
.50 tsp	27.8	29.2	26.8	26.0
.75 tsp	47.6	47.0	47.3	45.5
1.0 tsp	61.6	62.4	63.0	63.9

Do a complete analysis of variance for this set. Determine whether the treatments have an effect of the rise of the dough by means of a hypothesis test. Plot the diagnostic charts and comment on them. Use also Levene's test and Shapiro-Wilk. Use Tukey's HSD procedure to make pairwise comparisons and comment on the results.

Solution

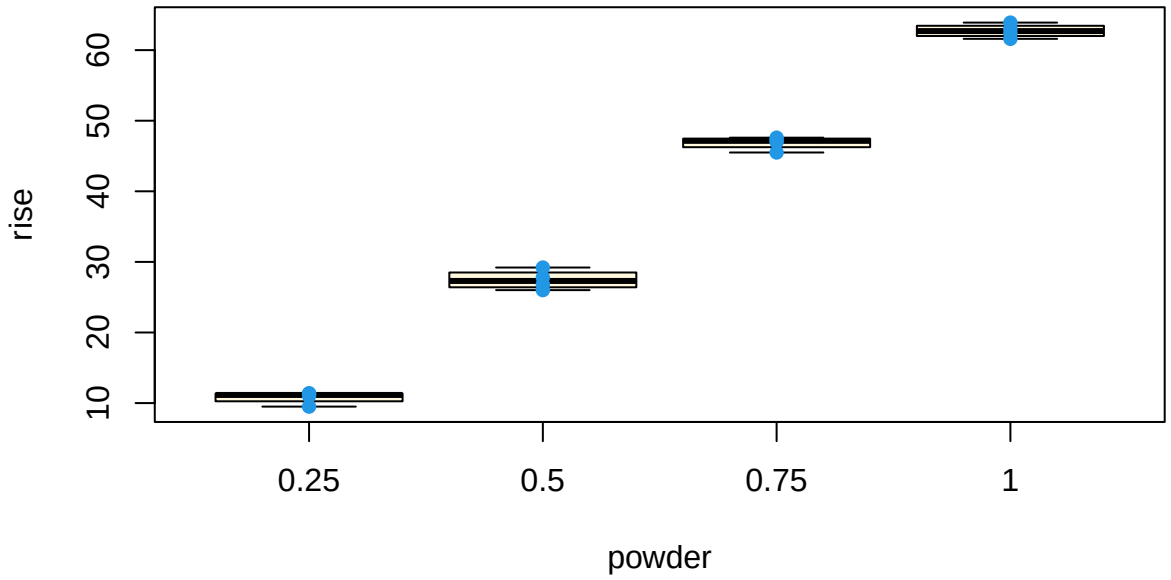
Start by loading the data:

```
rise <- c(11.4, 11.0, 11.3, 9.5, 27.8, 29.2, 26.8, 26.0, 47.6,
          47.0, 47.3, 45.5, 61.6, 62.4, 63.0, 63.9)
powder <- factor(rep(c(0.25,0.5,0.75,1), each = 4))
q2.df <- data.frame(powder,rise)
str(q2.df)
```

```
## 'data.frame':    16 obs. of  2 variables:
##  $ powder: Factor w/ 4 levels "0.25","0.5","0.75",..: 1 1 1 1 2 2 2 2 3 3 ...
##  $ rise  : num  11.4 11 11.3 9.5 27.8 29.2 26.8 26 47.6 47 ...
```

Boxplots for the data

```
boxplot(rise ~ powder, data = q2.df, col = 'cornsilk')
points(rise ~ powder, data = q2.df, pch = 16, col = 4)
```



We see that the height of the biscuits increases with the amount of baking powder. The boxes are similar in height, and there is little spread in the observed values for all treatment levels, so the homoscedasticity assumption seems to be valid. Also, there is no overlap on the rise value for different treatment levels, so the differences should all be significant.

We produce the Anova table with the following command

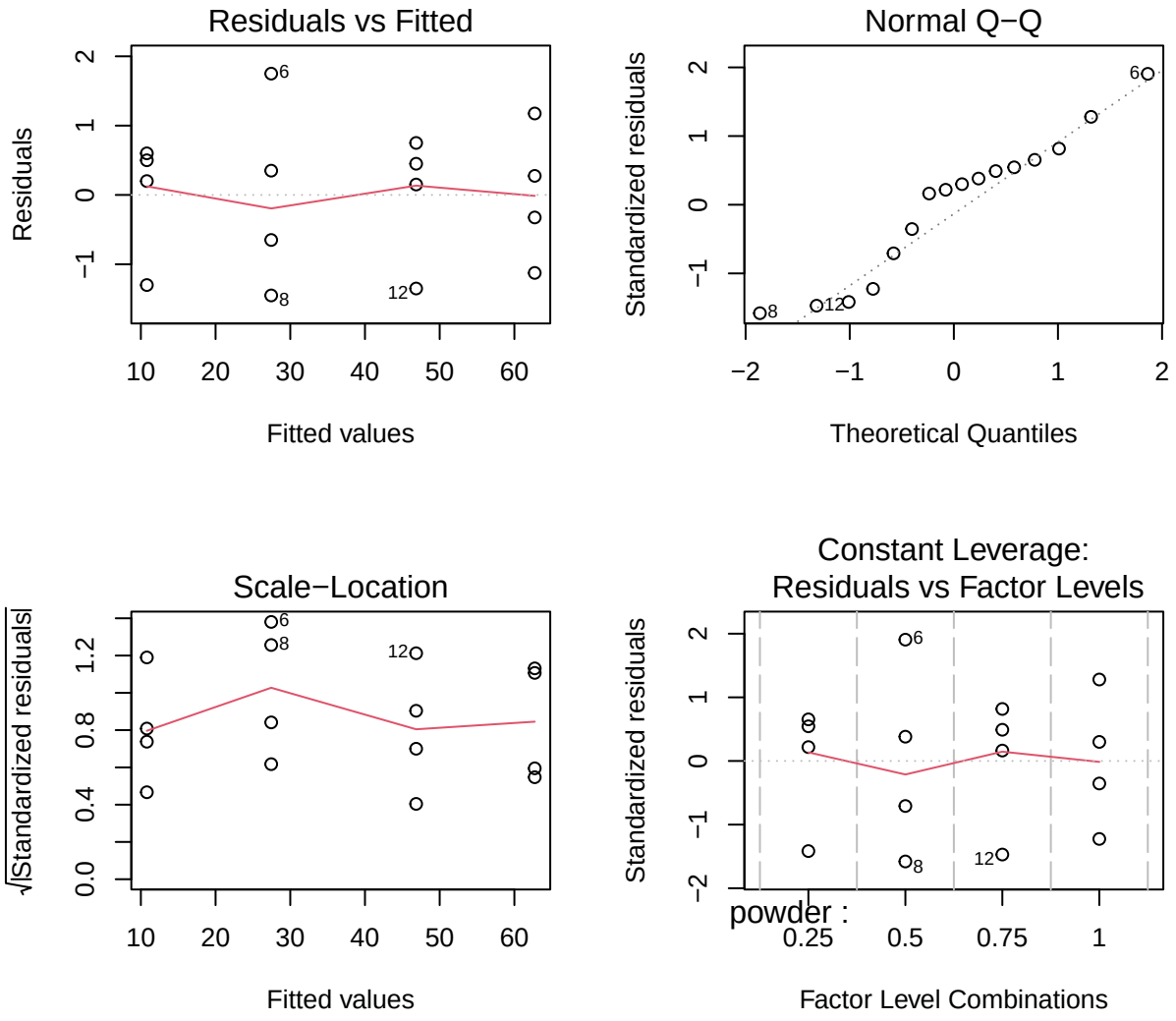
```
model1 <- aov(rise ~ powder, data = q2.df)
summary(model1)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## powder       3   6146  2048.6    1823 3.23e-16 ***
## Residuals   12     13     1.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p value is small, so we reject the hypothesis that all treatment levels have equal effects.

Diagnostic charts

```
par(mfrow=c(2,2))
plot(model1)
```




```
par(mfrow=c(1,1))
```

The quantile plot gives some cause for concern, as the points oscillate around the straight line, but departures are not large. The Shapiro-Wilk test will help deciding whether this assumption is satisfied. As for the other plots, there is some variability, but since there are only four meaasurements for each treatment level, they all seem reasonable.

We carry out the Shapiro-Wilk test on the residuals

```
shapiro.test(resid(model1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model1)
## W = 0.93838, p-value = 0.3297
```

This shows that at the 5% level we cannot reject the null hypothesis of Gaussianity.

```
library(car)
leveneTest(model1)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  3  0.4608 0.7148
##       12
```

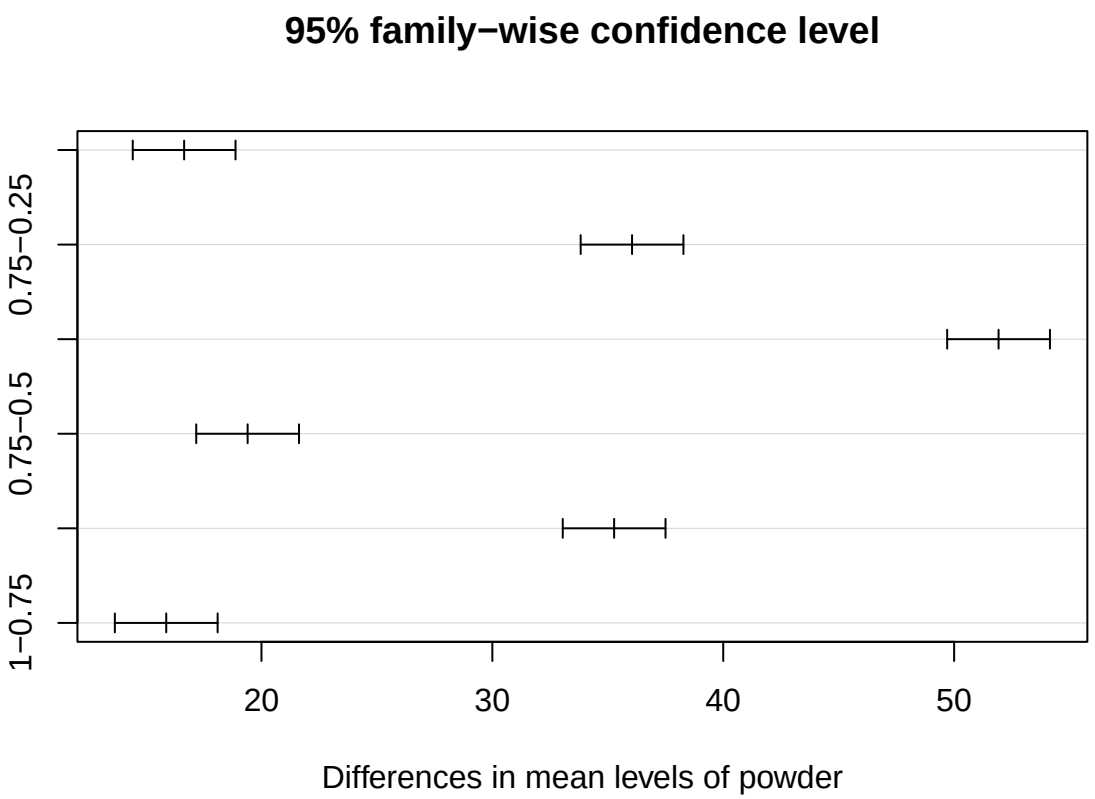
This test has a large p-value, saying that hypothesis of homoscedasticy is not rejected.

Finally, we use Tukey's honest significance difference to compare the different treatment levels.

```
(model1.tky <- TukeyHSD(model1))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = rise ~ powder, data = q2.df)
##
## $powder
##              diff      lwr      upr p adj
## 0.5-0.25   16.650 14.42436 18.87564     0
## 0.75-0.25  36.050 33.82436 38.27564     0
## 1-0.25     51.925 49.69936 54.15064     0
## 0.75-0.5   19.400 17.17436 21.62564     0
## 1-0.5      35.275 33.04936 37.50064     0
## 1-0.75     15.875 13.64936 18.10064     0
```

```
plot(model1.tky)
```



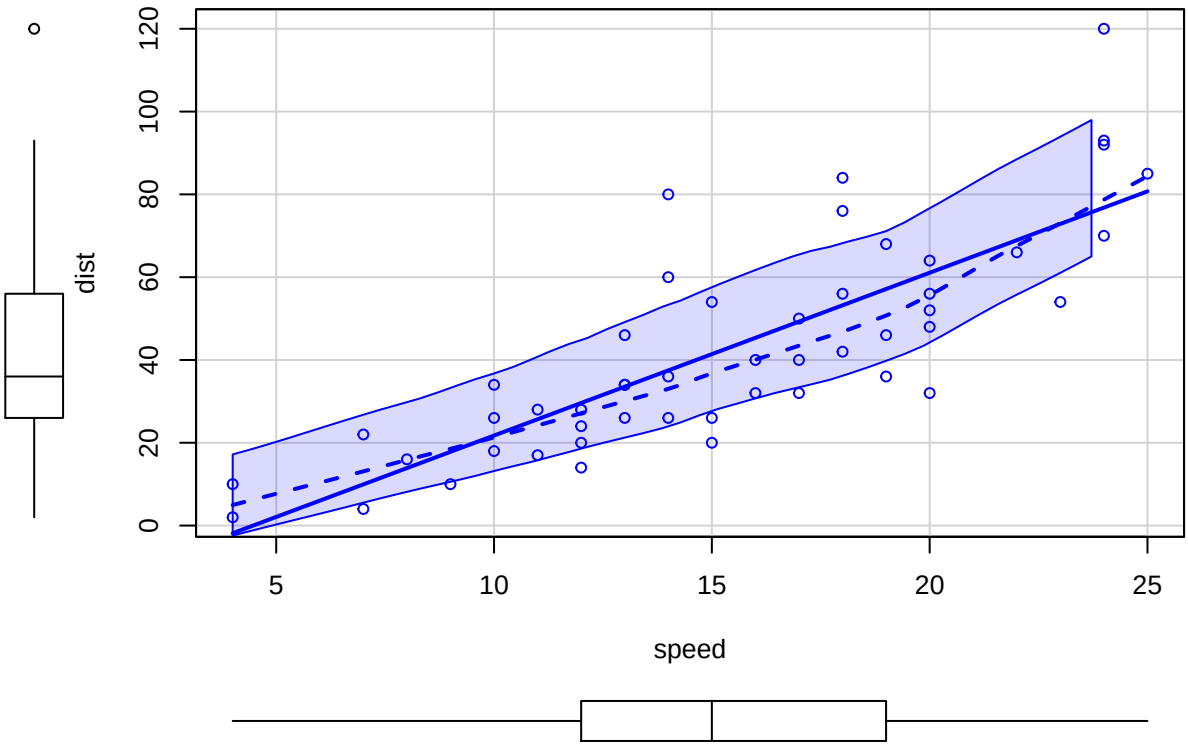
In this case we see that all differences are significant.

Exercise 3

For this exercise we will use the data set `cars`, that has information about the speed, in miles per hour, and braking distance, in feet, for 50 cars. The data was recorded in the 1920s.

(i) Draw a scatterplot of `speed` against `dist`. For this, use the function `scatterplot` in the `car` package. This function draws the points and also a simple regression line for the two variables. Moreover, it also plots a broken line that represents a local smoother function for the points as well as confidence bands for the smoother. The function also graphs boxplots for both variables on the corresponding axes. How would you interpret the differences between the regression line and the local smoother function that you see on the graph?

```
library(car)
scatterplot(dist ~ speed, data = cars)
```



The difference is a measure of the goodness-of-fit of the linear model to the data. In this case the local variability is reduced and the fit look good.

(ii) Use the function `lm` to fit a regression line to this data. Write down explicitly the model that you get and interpret the meaning of the coefficients.

```
model1 <- lm(dist ~ speed, data = cars)
summary(model1)
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
## speed         3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

The model is

$$dist = -17.58 + 3.93 \times speed$$

The intercept is the value of the breaking distance when the speed is 0. In this case it is a negative number, which does not make sense. The slope 3.93 indicates the increase in feet for the breaking distance for every increase in one mile per hour in the speed.

(iii) Use the function `summary` on the output of the regression. Interpret the t-tests in the table. Are the parameters different from zero?

The slope has a very small p-value, so we can reject the null hypothesis of zero slope. The intercept is marginally significant.

(iv) Describe the sampling distribution for the estimated parameters in this regression.

The estimated parameters are $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)$, which have a normal distribution:

$$\hat{\boldsymbol{\beta}} = N\big((\beta_0, \beta_1)', \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\big)$$

The matrix $(\mathbf{X}'\mathbf{X})^{-1}$ is obtained in `R` with

```
(invXtX <- summary(model1)$cov.unscaled)
```

```
##             (Intercept)        speed
## (Intercept)  0.19310949 -0.011240876
## speed       -0.01124088  0.000729927
```

The variance is unknown and is estimated by the mean square. The standard deviation is

```
summary(model1)$sigma
```

```
## [1] 15.37959
```

and the estimated variance is

```
summary(model1)$sigma^2
```

```
## [1] 236.5317
```

The estimated covariance matrix for $\hat{\boldsymbol{\beta}}$ can be obtained with

```
vcov(model1)
```

```
##             (Intercept)      speed
## (Intercept)   45.676514 -2.6588234
## speed         -2.658823  0.1726509
```

or multiplying $\hat{\sigma}^2$ times $(\mathbf{X}'\mathbf{X})^{-1}$

```
(summary(model1)$sigma^2)*invXtX
```

```
##             (Intercept)      speed
## (Intercept)   45.676514 -2.6588234
## speed         -2.658823  0.1726509
```

(v) Give confidence intervals at a confidence level of 98% for the parameters of the regression.

```
confint(model1, level = 0.98)
```

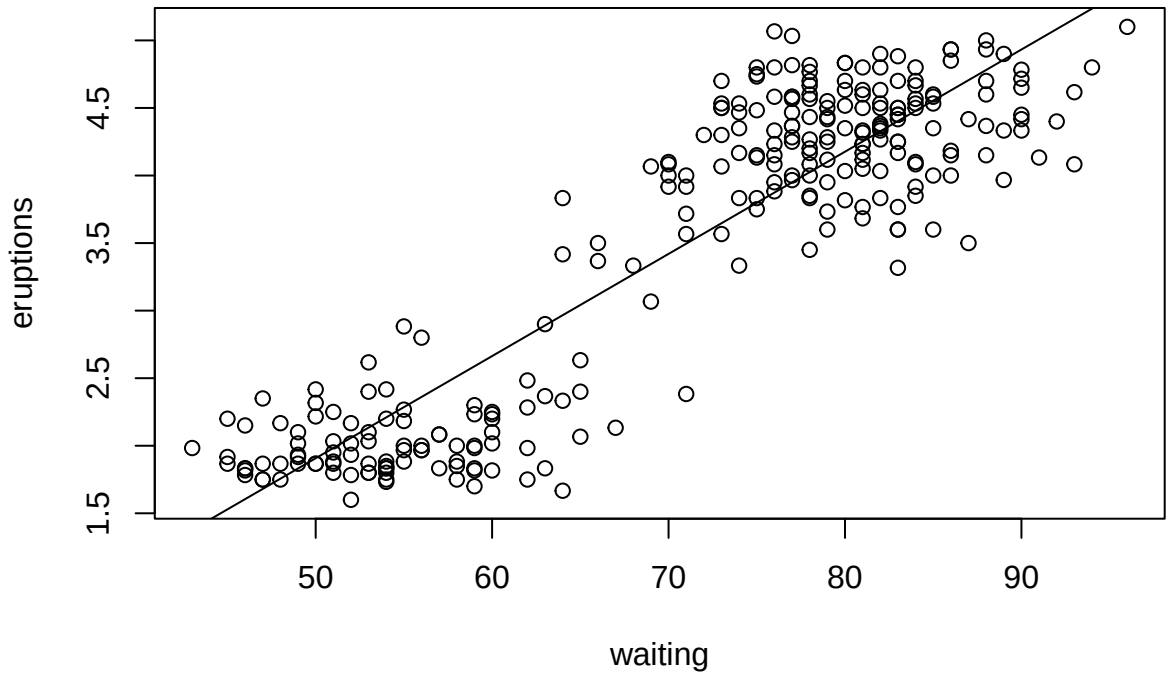
```
##                    1 %      99 %
## (Intercept) -33.843830 -1.314359
## speed         2.932443  4.932374
```

Exercise 4

For this exercise we will use the data set `faithful`. Look at the help for this set and familiarize yourself with the variables and their meaning.

(i) Plot a scatterplot of `eruptions` as a function of `waiting`. Add the corresponding regression line.

```
plot(eruptions ~ waiting, data = faithful)
abline(lm(eruptions ~ waiting, data = faithful))
```



(ii) Fit a regression to this data. What are the estimated values for the intercept and slope? Write down the regression model in this case and interpret the meaning of the coefficients.

```
eruptions.lm <- lm(eruptions ~ waiting, data = faithful)
summary(eruptions.lm)
```

```
##
## Call:
## lm(formula = eruptions ~ waiting, data = faithful)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.29917 -0.37689  0.03508  0.34909  1.19329
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.874016   0.160143  -11.70   <2e-16 ***
## waiting      0.075628   0.002219   34.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4965 on 270 degrees of freedom
## Multiple R-squared:  0.8115, Adjusted R-squared:  0.8108
## F-statistic:  1162 on 1 and 270 DF,  p-value: < 2.2e-16
```

$$eruptions = -1.874 + 0.756 \times waiting$$

(iii) What are the results of the t-tests in this example?

Both p values are very small, so we reject the null hypotheses of parameters equal to zero.

(iv) What would be the predicted value according to this model for the eruption time corresponding to a waiting time of 72 minutes?

```
predict.lm(eruptions.lm, data.frame(waiting = 72), interval = 'p')
```

```
##        fit     lwr      upr
## 1 3.571196 2.59186 4.550533
```

(v) Describe the sampling distribution for the estimated parameters in the previous regression.

The estimated parameters are $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)$, which have a normal distribution:

$$\hat{\boldsymbol{\beta}} = N\big((\beta_0, \beta_1)', \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\big)$$

The matrix $(\mathbf{X}'\mathbf{X})^{-1}$ is obtained in R with

```
(invXtX <- summary(eruptions.lm)$cov.unscaled)
```

```
##              (Intercept)       waiting
## (Intercept)  0.104029479 -1.415475e-03
## waiting     -0.001415475  1.996521e-05
```

The variance is unknown and is estimated by the mean square. The standard deviation is

```
summary(eruptions.lm)$sigma
```

```
## [1] 0.4965129
```

and the estimated variance is

```
summary(eruptions.lm)$sigma^2
```

```
## [1] 0.2465251
```

The estimated covariance matrix for $\hat{\boldsymbol{\beta}}$ can be obtained with

```
vcov(eruptions.lm)
```

```
##               (Intercept)       waiting
## (Intercept)  0.0256458773 -3.489501e-04
## waiting     -0.0003489501  4.921926e-06
```

or multiplying $\hat{\sigma}^2$ times $(\mathbf{X}'\mathbf{X})^{-1}$

```
(summary(eruptions.lm)$sigma^2)*invXtX
```

```
##                (Intercept)       waiting
## (Intercept)  0.0256458773 -3.489501e-04
## waiting     -0.0003489501  4.921926e-06
```

(vi) Give a confidence interval at a confidence level of 98% for the parameters of the regression.

```
confint(eruptions.lm,level = 0.98)
```

```
##                      1 %        99 %
## (Intercept) -2.24878944 -1.49924253
## waiting      0.07043603  0.08081986
```