STAT 210
Applied Statistics and Data Analysis:
Problem list 5 - Solution
(Due on week 6)

Exercise 1

An experiment was run to determine the effect of temperature on the density of certain type of bricks. The table below has the results for the experiment.

Table 1: Density of Bricks

Temperature	Density				
100	31.8	31.6	31.7	31.6	31.7
125	31.7	31.4	31.5	31.4	31.6
150	31.9	31.8	31.8	31.6	31.7
175	31.9	32.1	32.0	31.9	31.8

The data can be loaded by copying the commands below.

```
density = c(31.8, 31.6, 31.7, 31.6, 31.7,
            31.7, 31.4, 31.5, 31.4, 31.6,
            31.9, 31.8, 31.8, 31.6, 31.7,
            31.9, 32.1, 32.0, 31.9, 31.8)
temperature <- factor(rep(c(100,125,150,175), each = 5))
Q1data <- data.frame(density, temperature)
```

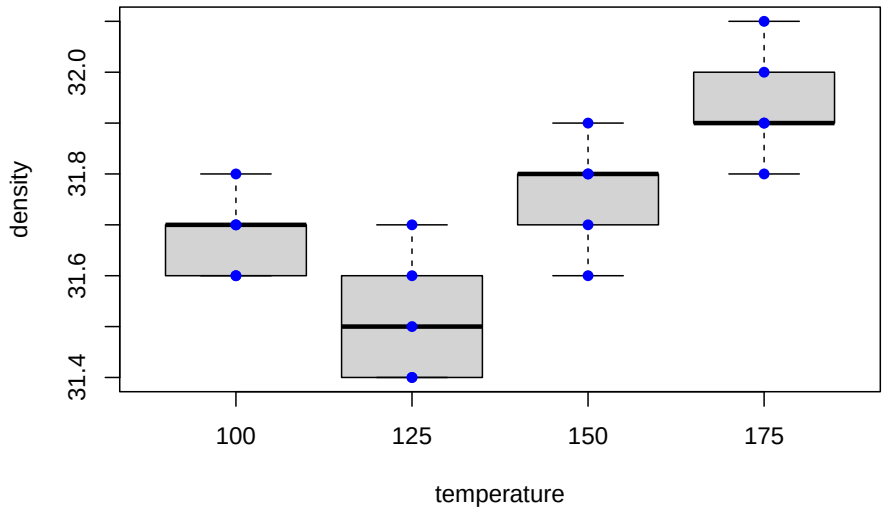
Do a complete analysis of variance for this set. Plot the data. Determine whether the temperatures have an effect on the brick density through a hypothesis test and state explicitly the null and alternative hypotheses in this test. Obtain the estimated values for the cell means and the effects. Write the equation for the model and state explicitly the assumptions on which the model is based. Plot the diagnostic charts and comment on them. Use Levene's and Shapiro-Wilk's tests also. Use Tukey's HSD procedure to make pairwise comparisons and comment on the results. Use a non-parametric alternative to the analysis of variance and compare the results. If the objective is more density, which temperature would you select and why?

Solution

Exploratory Analysis

We start by looking at the data. For this, we produce boxplots and include the data.

```
boxplot(density ~ temperature, data = Q1data)
points(density ~ temperature, data = Q1data, pch = 16, col = 'blue')
```



The density decreases as we go from 100 to 125 degrees, but then increases with temperature. If we look at the boxes, it seems that the variance for 125° may be different. However, looking at the points, we see that the density values for the different temperatures have similar ranges, supporting the assumption of equal variances. Also, it looks as if the differences in the density between some temperature values will be significant.

Equation for the model

The equation for the analysis of variance model is

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

where y_{ij} is the observed density on the j-th replicate of the i-th temperature level, for $i = 1, \ldots, 4$ and $j = 1, \ldots, 5$, μ is the overall mean of the observations, τ_i is the effect of the i-th temperature level on the density of the bricks, and ε_{ij} is the experimental error for the j-th replicate at treatment level i. The model is based on the assumptions that the errors are independent, centered random variables with common normal distribution with variance σ^2.

Model Fitting

The hypothesis tested in the anova table are

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_4 \qquad vs. \qquad H_1 : \tau_i \neq \tau_j \text{ for at least one pair } i, j$$

We fit the Anova model and print the anova table with the following commands

```
mod1 <- aov(density ~ temperature, data = Q1data)
summary(mod1)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## temperature  3 0.4575  0.1525    12.2 0.000209 ***
## Residuals   16 0.2000  0.0125
## ---
## 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. The estimated variance for this model, obtained from the table, is 0.125, with standard deviation 0.4.

Analysis of the Model

The estimated coefficients for the model can be obtained by printing the summary table with the function `summary.lm`:

```
summary.lm(mod1)
```

```
##
## Call:
## aov(formula = density ~ temperature, data = Q1data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -0.160 -0.080  0.000  0.065  0.180
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    31.68000    0.05000 633.600  < 2e-16 ***
## temperature125 -0.16000    0.07071  -2.263  0.03792 *
## temperature150  0.08000    0.07071   1.131  0.27457
## temperature175  0.26000    0.07071   3.677  0.00204 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1118 on 16 degrees of freedom
## Multiple R-squared:  0.6958, Adjusted R-squared:  0.6388
## F-statistic:  12.2 on 3 and 16 DF,  p-value: 0.000209
```

The `Intercept` corresponds to the estimated cell mean for temperature 100, i.e., to $\hat{\mu} + \hat{\tau}_1$. The other three values are differences of the cell means: $\hat{\tau}_i - \hat{\tau}_1$, for $i = 2, 3, 4$. To obtain the cell means for all circuits we can use

```
model.tables(mod1, 'mean', se = T)
```

```
## Tables of means
## Grand mean
##
## 31.725
##
##  temperature
## temperature
##   100   125   150   175
## 31.68 31.52 31.76 31.94
##
## Standard errors for differences of means
##         temperature
##             0.07071
## replic.           5
```

Similarly, we can obtain the estimated effects $\hat{\tau}_i, i = 1, \ldots, 4$:

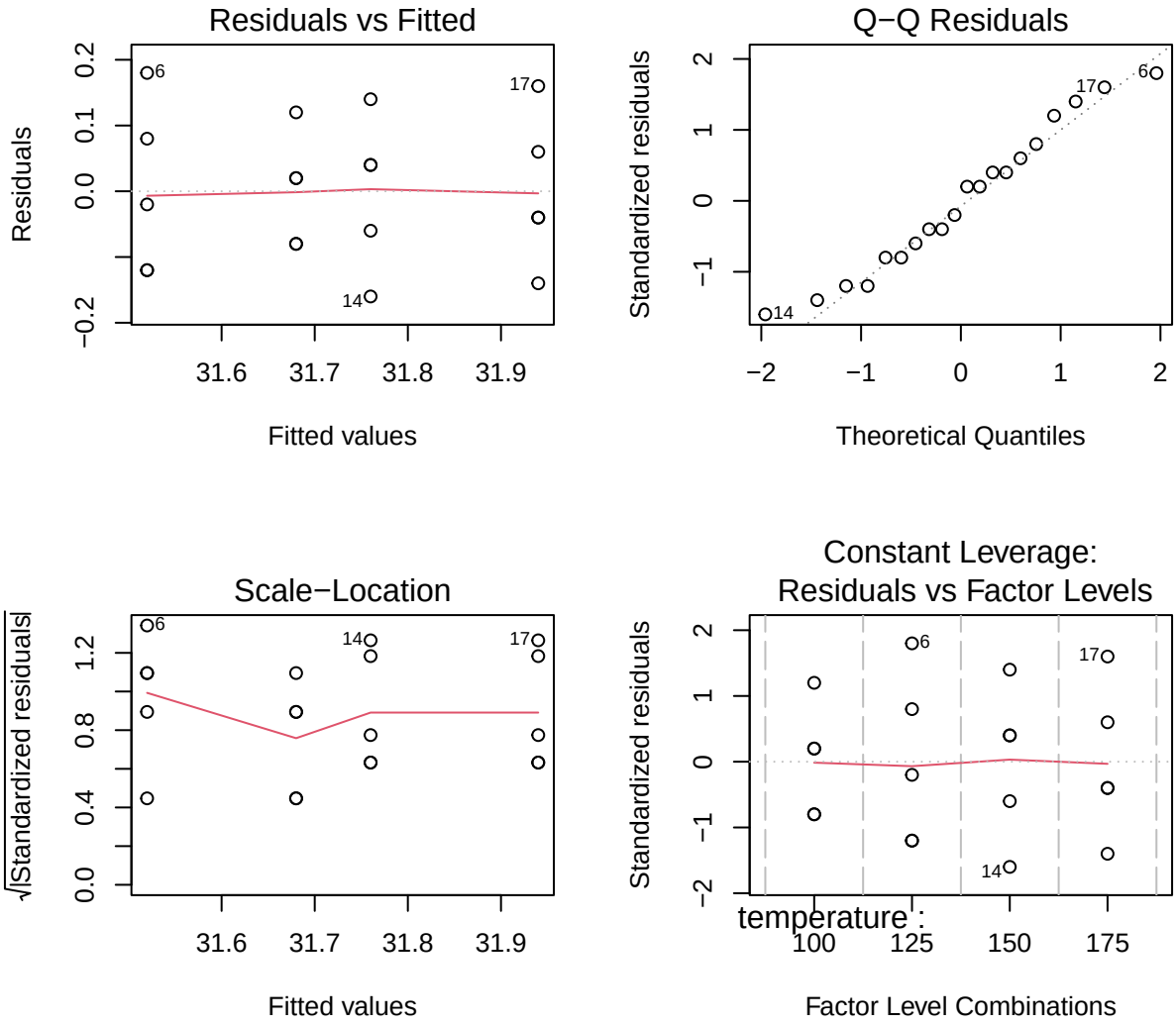
```
model.tables(mod1, se = T)
```

```
## Tables of effects
##
##  temperature
## temperature
##    100    125    150    175
## -0.045 -0.205  0.035  0.215
##
## Standard errors of effects
```

```
##         temperature
##                0.05
## replic.           5
```

Diagnostic Plots

```
par(mfrow=c(2,2))
plot(mod1)
```




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

The first plot shows an almost horizontal red line, close to 0, indicating that residuals are approximately centered and symmetric. The variability of the residuals seems to be stable. The normal quantile plot supports the assumption of normality. The red line in the scale-location plot shows some variability and suggests using the Levene test. The fourth plot is similar to the first and the same comments apply.

We now do the tests. We use the function `rstandard` to obtain the standardized residuals:

```
shapiro.test(rstandard(mod1))
```

```
##
##  Shapiro-Wilk normality test
##
```

```
## data:  rstandard(mod1)
## W = 0.96171, p-value = 0.5785
```

The p-values is large enough not to reject the null hypothesis of normality. For homoscedasticity we have

```
library(car)
leveneTest(mod1)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  3  0.2424 0.8655
##       16
```

And again the p-value is large, so we do not reject the null hypothesis of equal variances.

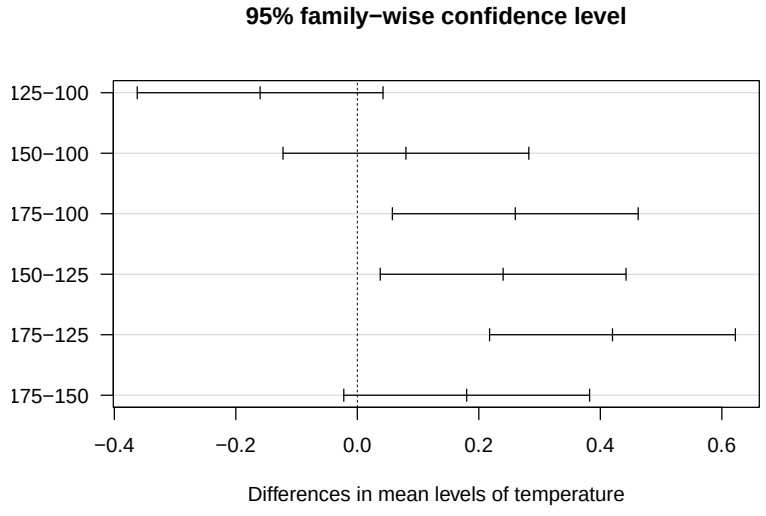
Pairwise comparisons

We use Tukey's honest significance difference method to compare the different treatment levels.

```
(mod1.tky <- TukeyHSD(mod1))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = density ~ temperature, data = Q1data)
##
## $temperature
##          diff         lwr        upr     p adj
## 125-100 -0.16 -0.36230465 0.04230465 0.1487864
## 150-100  0.08 -0.12230465 0.28230465 0.6762336
## 175-100  0.26  0.05769535 0.46230465 0.0098844
## 150-125  0.24  0.03769535 0.44230465 0.0175030
## 175-125  0.42  0.21769535 0.62230465 0.0001109
## 175-150  0.18 -0.02230465 0.38230465 0.0902649
```

```
plot(mod1.tky, las = 1 )
```



We see that only three comparisons are significant at the 5% level: 175-100, 175-125, and 150-125, but 175 and 150 are not distinguishable.

Non-parametric test

The non-parametric alternative to the anova model is the Kruskal-Wallis test:

```
kruskal.test(density ~ temperature, data = Q1data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  density by temperature
## Kruskal-Wallis chi-squared = 13.331, df = 3, p-value = 0.003973
```

which leads to the same conclusion as the anova model.

Conclusion

Since we want more density, we should go for either of the two higher temperatures. These two have significantly higher cell means (31.76 and 31.94) than the two lower temperatures (31.68 and 31.52). The pairwise comparison between 150 and 175 is not significant, and therefore qwe cannot conclude that they are different. It is up to the manufacturer to decide which one to pick, or a further experiment could be designed to compare these two temperatures.

Exercise 2

To determine the effect of polypropylene fibers on the compressive strength of concrete, and experiment was carried out in which five different levels of fibre content were considered: 0, 0.25, 0.50, 0.75, and 1%. For each level, four concrete cubes were produced and the compressive strength was measured. The table with the results is given below

Table 2: Compressive strength of concrete

Fibre content (%)	Strength (ksi)			
0	7.8	7.4	7.2	7.5
0.25	7.9	7.5	7.3	7.7
0.50	7.1	6.9	6.3	6.7
0.75	7.0	6.7	6.4	6.6
1.0	5.9	5.8	5.6	6.0

The data can be loaded by copying the commands below.

```
strength <- c(7.8, 7.4, 7.2, 7.5,
              7.9, 7.5, 7.3, 7.7,
              7.1, 6.9, 6.3, 6.7,
              7.0, 6.7, 6.4, 6.6,
              5.9, 5.8, 5.6, 6.0)
fib <- factor(rep(c(0, .25, .50, .75, 1), each = 4))
Q2data <- data.frame(strength, fib)
```

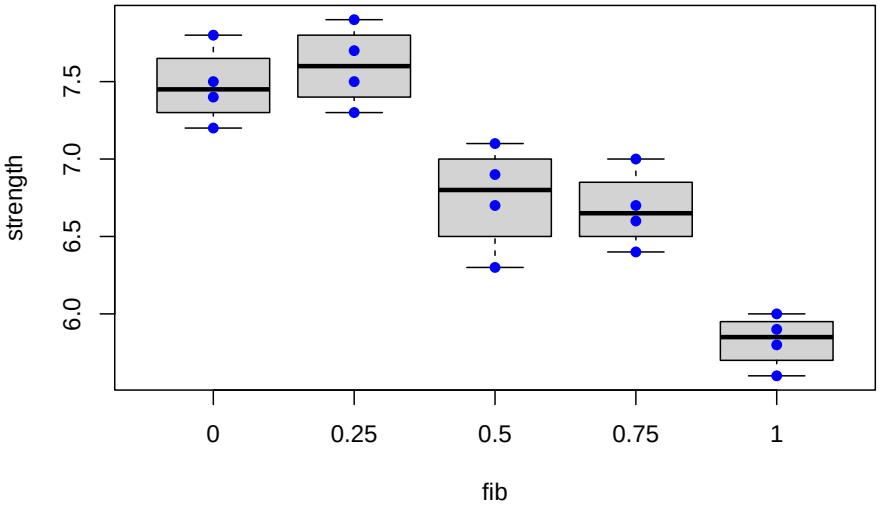
Do a complete analysis of variance for this set. Plot the data. Determine whether there is a difference in the lifetime of the fluids through a hypothesis test and state explicitly the null and alternative hypotheses in this test. Obtain the estimated values for the cell means and the effects. Write the equation for the model. Plot the diagnostic charts and comment on them. Use Levene's and Shapiro-Wilk's tests also. Use Tukey's HSD procedure to make pairwise comparisons and comment on the results. Use a non-parametric alternative to the analysis of variance and compare the results. If the objective is to have a higher strength, which fiber content percentage would you select and why?

Solution

Exploratory Analysis

We start by looking at the data. For this, we produce boxplots and include the data.

```
boxplot(strength ~ fib, data = Q2data)
points(strength ~ fib, data = Q2data, pch = 16, col = 'blue')
```



Roughly speaking, the strength decreases as the percentage of fiber increases, but not uniformly. In fact, 0 and 0.25 seem to produce the same strength, as do 0.5 and 0.75. The boxes are similar in size, supporting the assumption of equal variances.

Equation for the model

The equation for the analysis of variance model is

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

where y_{ij} is the observed compressive strength on the j-th replicate of the i-th fiber content level, for $i = 1, \ldots, 5$ and $j = 1, \ldots, 4$, μ is the overall mean of the observations, τ_i is the effect of the i-th fiber content level on the compressive strength of concrete, and ε_{ij} is the experimental error for the j-th replicate at treatment level i. The model is based on the assumption that the errors are independent, centered random variables with common normal distribution with variance σ^2.

Model Fitting

The hypothesis tested in the anova table are

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_5 \qquad vs. \qquad H_1 : \tau_i \neq \tau_j \text{ for at least one pair } i, j$$

We fit the Anova model and print the anova table with the following command

```
mod2 <- aov(strength ~ fib, data = Q2data)
summary(mod2)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## fib          4  8.173  2.0433   30.27 5.04e-07 ***
## Residuals   15  1.013  0.0675
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p value is small and at the 5% significance level we would reject the null hypothesis of equal effects. The estimated variance for this model, obtained from the table, is 0.0675, with standard deviation 0.26.

Analysis of the Model

The estimated coefficients for the model can be obtained by printing the summary table with the function `summary.lm`:

```
summary.lm(mod2)
```

```
##
## Call:
## aov(formula = strength ~ fib, data = Q2data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4500 -0.1313  0.0000  0.1562  0.3500
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.4750     0.1299  57.543  < 2e-16 ***
## fib0.25       0.1250     0.1837   0.680 0.506610
## fib0.5       -0.7250     0.1837  -3.946 0.001293 **
## fib0.75      -0.8000     0.1837  -4.355 0.000566 ***
## fib1         -1.6500     0.1837  -8.981 2.01e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2598 on 15 degrees of freedom
## Multiple R-squared:  0.8898, Adjusted R-squared:  0.8604
## F-statistic: 30.27 on 4 and 15 DF,  p-value: 5.037e-07
```

The `Intercept` corresponds to the estimated cell mean for 0% fibre content, i.e., $\hat{\mu} + \hat{\tau}_1$. The other four values are differences of the cell means: $\hat{\tau}_i - \hat{\tau}_1$, for $i = 2, 3, 4, 5$. To obtain the cell means for all fluids we can use

```
model.tables(mod2, 'mean', se = T)
```

```
## Tables of means
## Grand mean
##
## 6.865
##
##  fib
## fib
##     0  0.25   0.5  0.75     1
## 7.475 7.600 6.750 6.675 5.825
##
## Standard errors for differences of means
##            fib
##         0.1837
## replic.      4
```

Similarly, we can obtain the estimated effects $\hat{\tau}_i, i = 1, \ldots, 4$:

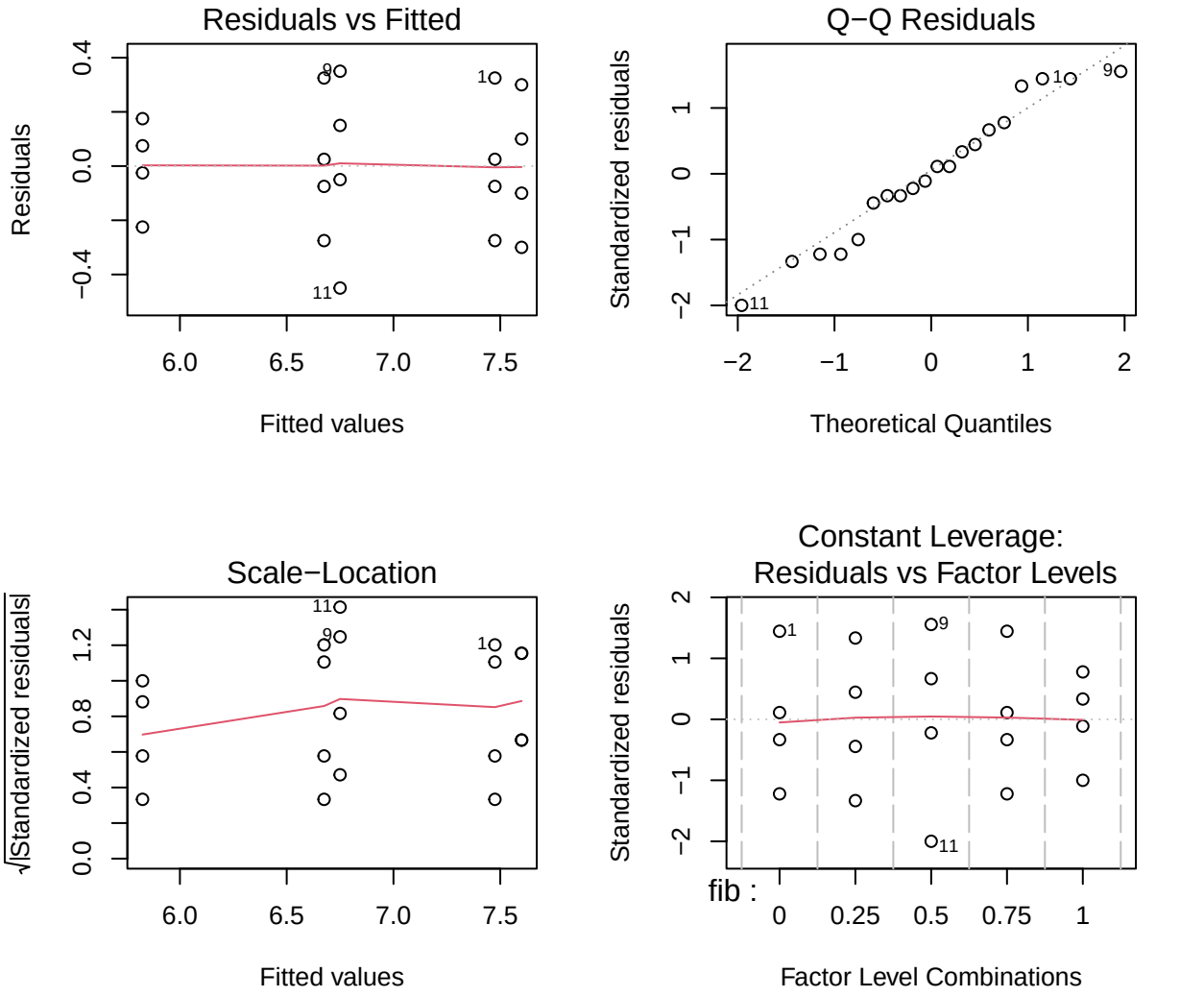
```
model.tables(mod2, se = T)
```

```
## Tables of effects
##
##  fib
## fib
```

```
##       0    0.25     0.5    0.75       1
##   0.610   0.735 -0.115 -0.190 -1.040
##
## Standard errors of effects
##             fib
##          0.1299
## replic.      4
```

Diagnostic Plots and Tests

```
par(mfrow=c(2,2))
plot(mod2)
```




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

The first plot shows an almost horizontal red line, close to 0, indicating that residuals are approximately centered and symmetric. The variability of the residuals seems bigger at the center than on the extremes. The normal quantile plot supports the assumption of normality. The red line in the scale-location plot is slightly increasing,and the assumption of homogeneous variances should be checked with Levene's test. The fourth plot is similar to the first and the same comments apply.

We do now the tests. We use the function `rstandard` to obtain the standardized residuals:

```r
shapiro.test(rstandard(mod2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rstandard(mod2)
## W = 0.96045, p-value = 0.5529
```

The p-value is large so we do not reject the null hypothesis of normality. For homoscedasticity we have

```r
library(car)
leveneTest(mod2)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  4   0.396 0.8084
##       15
```

Again, the p-value is large, so we do not reject the null hypothesis of equal variances.

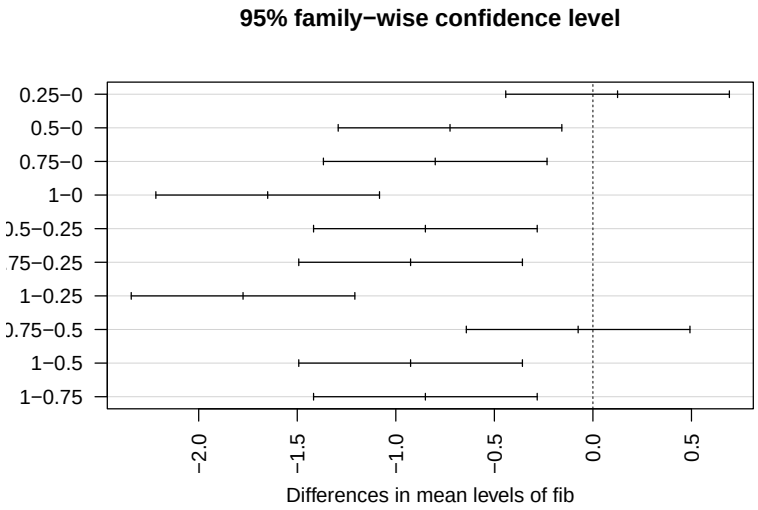
Pairwise comparisons

We use Tukey's honest significance difference method to compare the different treatment levels.

```r
(mod2.tky <- TukeyHSD(mod2))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = strength ~ fib, data = Q2data)
##
## $fib
##             diff       lwr       upr     p adj
## 0.25-0     0.125 -0.442288  0.692288 0.9577198
## 0.5-0     -0.725 -1.292288 -0.157712 0.0096590
## 0.75-0    -0.800 -1.367288 -0.232712 0.0043764
## 1-0       -1.650 -2.217288 -1.082712 0.0000018
## 0.5-0.25  -0.850 -1.417288 -0.282712 0.0025905
## 0.75-0.25 -0.925 -1.492288 -0.357712 0.0011913
## 1-0.25    -1.775 -2.342288 -1.207712 0.0000007
## 0.75-0.5  -0.075 -0.642288  0.492288 0.9935281
## 1-0.5     -0.925 -1.492288 -0.357712 0.0011913
## 1-0.75    -0.850 -1.417288 -0.282712 0.0025905
```

```r
plot(mod2.tky, las = 2)
```



We see that all pairwise comparisons are significant except two: 0 - 0.25 and 0.5 - 0.75.

Non-parametric test

The non-parametric alternative to the anova model is the Kruskal-Wallis test:

```
kruskal.test(strength ~ fib, data = Q2data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  strength by fib
## Kruskal-Wallis chi-squared = 16.604, df = 4, p-value = 0.002308
```

which leads to the same conclusion as the anova model at the 5% level of significance.

Conclusion

Since we want more strength, we should choose either 0 or 0.25% of fiber content, since these two levels cannot be distinguished and they produce the highest values for the compressive strength of concrete.

Exercise 3

The file `PL6-25_q1.csv` has the results of an experiment to test the strength of five different ferric alloys. The alloys are coded with the letters `A`, `B`, `C`, `D`, and `E`. **Use $\alpha = 0.05$ for the tests in this question.**

Read the data:

```
PL6_q1 <- read.csv('PL6-25_q1.csv')
str(PL6_q1)
```

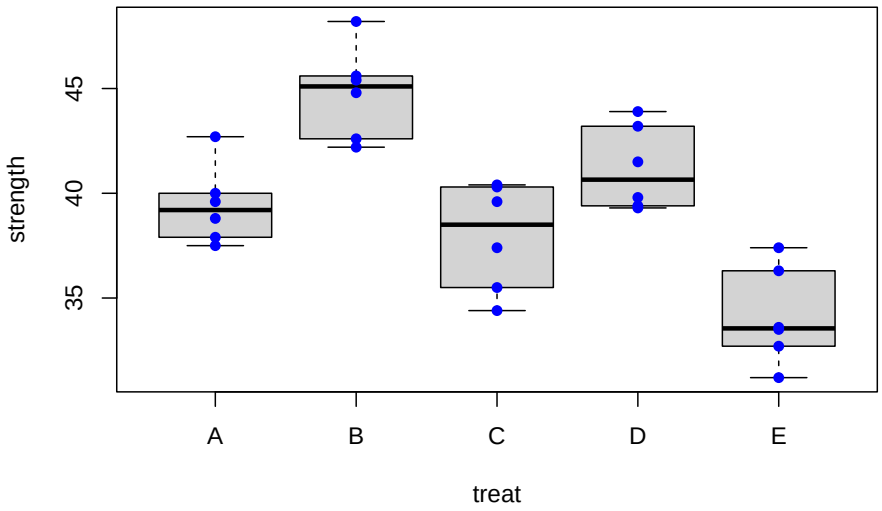
```
## 'data.frame':    30 obs. of  2 variables:
##  $ strength: num  42.7 38.8 37.5 40 37.9 39.6 48.2 42.6 45.4 44.8 ...
##  $ treat   : chr  "A" "A" "A" "A" ...
```

Transform `treat` into a factor

```
PL6_q1$treat <- factor(PL6_q1$treat)
```

(a) Plot the data and discuss the graphs.

```
plot(strength ~ treat, data = PL6_q1)
points(strength ~ treat, pch = 16, col = 'blue', data = PL6_q1)
```



The boxes have similar sizes, pointing to homogeneous variances for the different treatment levels. Alloy `B` has the highest median strength, while alloy `E` has the lowest. The range of values for `B` and `E` are disjoint, so we would expect to have at least one significant difference.

(b) Fit an analysis of variance model and print the anova table. What do you conclude from the table?

```
model1 <- aov(strength ~ treat, data = PL6_q1)
summary(model1)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## treat        4  374.2   93.55   19.21 2.47e-07 ***
## Residuals   25  121.7    4.87
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is small and we reject the null hypothesis of equal effects for all treatments.

(c) What are the estimated values for the variance and standard deviation of the errors in this experiment?

The estimated variance is the MSE, which is obtained from the anova table, and is 4.9. To get the standard deviation we need the square root

```
sqrt(4.87)
```

```
## [1] 2.206808
```

This can also be obtained as the residual standard error from the summary table.

(d) Obtain a summary table for the model fitted in (b) and discuss the information in the summary, including an interpretation of the coefficients for the model.

```
summary.lm(model1)
```

```
##
## Call:
## aov(formula = strength ~ treat, data = PL6_q1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5333 -1.7167 -0.2583  1.9292  3.4000
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  39.4167     0.9008  43.757  < 2e-16 ***
## treatB        5.3833     1.2739   4.226 0.000277 ***
## treatC       -1.4833     1.2739  -1.164 0.255265
## treatD        1.7667     1.2739   1.387 0.177752
## treatE       -5.3000     1.2739  -4.160 0.000328 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.207 on 25 degrees of freedom
## Multiple R-squared:  0.7546, Adjusted R-squared:  0.7153
## F-statistic: 19.21 on 4 and 25 DF,  p-value: 2.471e-07
```

The summary table starts by recalling the formula for the model. It then gives the quartiles for the residuals. We see from the values that the quartiles appear to be symmetric and that the median is reasonably close to zero.

The first coefficient, corresponding to the intercept, is the cell mean corresponding to the first level of treatment, which is alloy `A`. The value is 39.417. The following four coefficients correspond to differences between the other levels (`B` to `E`) and `A`. For instance, the third coefficient, `treatC` = -1.483, corresponds to the difference between the cell means for `A` minus `C`, and the cell mean for treatment level `C` is

$$39.417 - 1.483 = 37.934.$$

The values for the cell means for all levels will be obtained below in the table of means.

The residual standard error is the standard deviation for the residuals, with a value of 2.21. The adjusted R^2 has a value 0.715, and this is usually interpreted as the proportion of the uncertainty in the data that is explained by the model. Finally, the last line in the summary corresponds to the global test for the model, and is equivalent to the test we have in the anova table. It is an F test with test statistic 19.2 and p-value 2.47×10^{-7}, indicating that the null hypothesis of no treatment effect is rejected.

(e) Using the information in the summary table, calculate the effects and the marginal means.

Table of effects:

```
model.tables(model1)
```

```
## Tables of effects
##
##  treat
## treat
##      A      B      C      D      E
## -0.073  5.310 -1.557  1.693 -5.373
```

Table of cell means:

```
model.tables(model1, 'means')
```

```
## Tables of means
## Grand mean
##
## 39.49
##
##  treat
## treat
##     A     B     C     D     E
## 39.42 44.80 37.93 41.18 34.12
```

(f) Write down an equation for the anova model and explain the meaning of the parameters in this particular case.

There are five different alloys un the experiment, so we have five treatment levels, one for each alloy. There are 30 observations in the data set which correspond to six observation per treatment level. The equation is

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \quad i = A, B, C, D, E; 1 \leq j \leq 6,$$

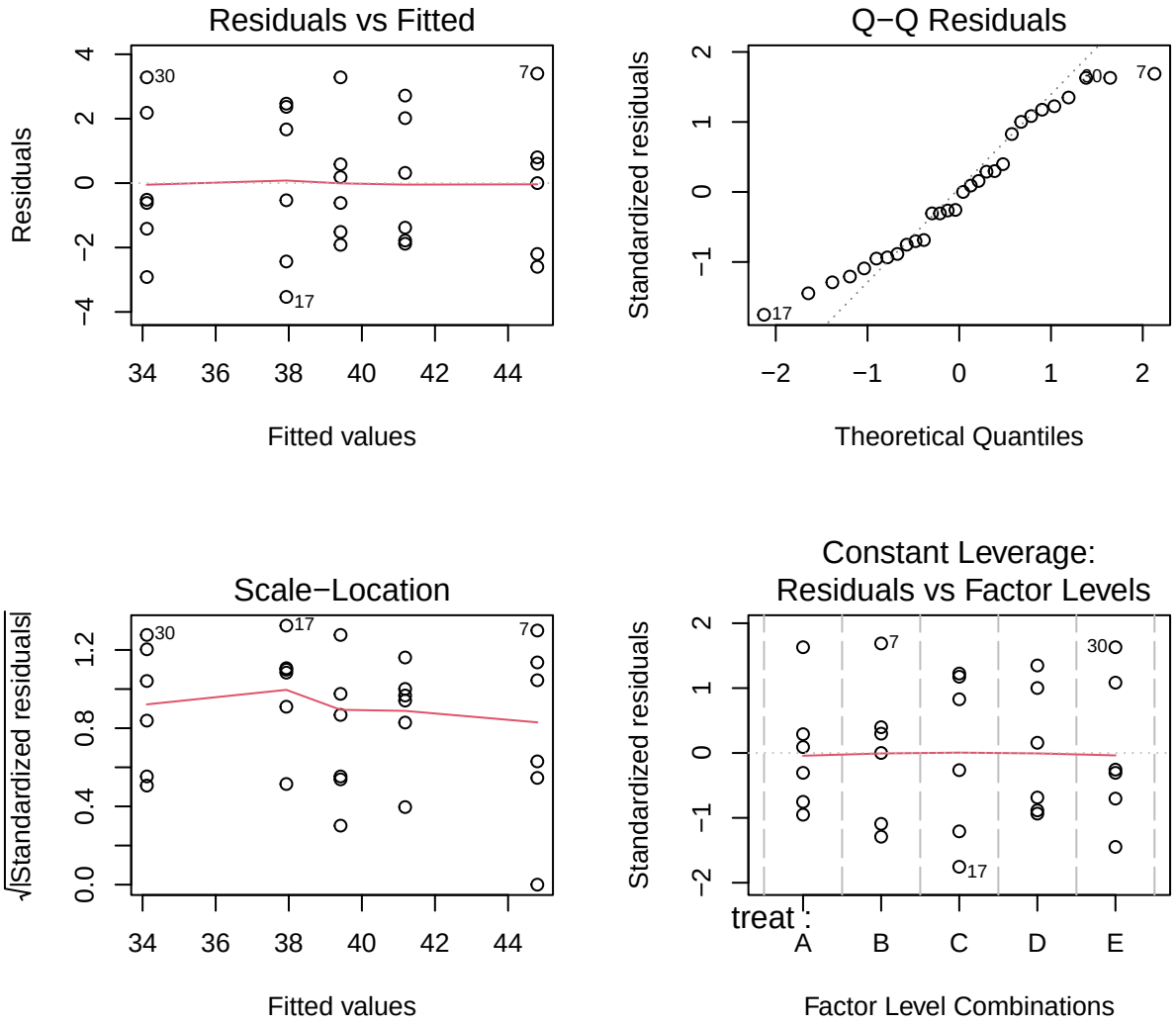
where y_{ij} is the measured strength for replicate j of alloy i, μ is the overall mean, τ_i is the effect of the i-th treatment level (alloy), and ε_{ij} is the error associated to the measurement of teh strength in the j-th replicate of level i. The values can be obtained from the tables in (e):

$$\mu = 39.49, \quad \tau_A = -0.73, \quad \tau_B = 5.31, \quad \tau_C = -1.557, \quad \tau_D = 1.693, \quad \tau_E = -5.373$$

(g) What are the assumptions on which the analysis of variance model is based? Using diagnostic plots and tests discuss whether these assumptions are satisfied.

The model assumes that the errors are independent and follow a normal distribution with mean zero and common unknown variance σ^2. We obtain the residual plots with the following commands

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



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

All the plots are good, and all the assumptions seem to be satisfied. The first plot shows that the residuals are randomly and symmetrically distributed with mean close to zero, and there are no unexpected patterns. The quantile plot shows a good fit to the reference line. The scale-location plot shows that all square roots of the normalized residuals have absolute values below 1.3, indicating that there are no large values, and the final plot, which is similar to the first, shows again the symmetry of the residuals.

We check normality and homogeneous variances with the corresponding tests

```r
shapiro.test(rstandard(model1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rstandard(model1)
## W = 0.95094, p-value = 0.1792
```

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

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  4   0.328 0.8565
##       25
```

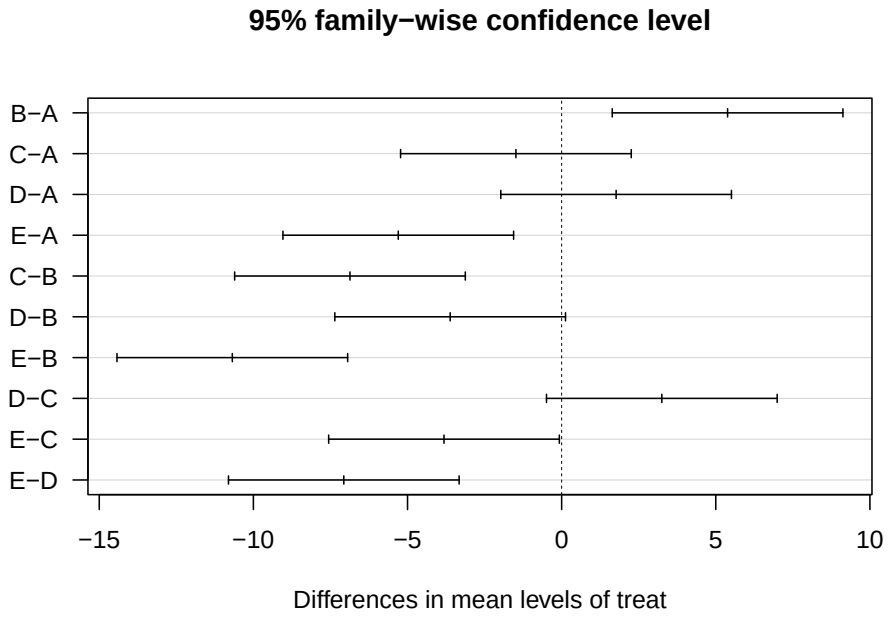
The p-values in both cases are large, and we do not reject the null hypotheses.

(h) Use Tukey's method for pairwise comparisons with a confidence level of 95% and comment on the results.

```r
(md1_tky <- TukeyHSD(model1))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = strength ~ treat, data = PL6_q1)
##
## $treat
##           diff         lwr        upr     p adj
## B-A   5.383333   1.6419495  9.1247172 0.0023519
## C-A  -1.483333  -5.2247172  2.2580505 0.7709589
## D-A   1.766667  -1.9747172  5.5080505 0.6413684
## E-A  -5.300000  -9.0413839 -1.5586161 0.0027696
## C-B  -6.866667 -10.6080505 -3.1252828 0.0001233
## D-B  -3.616667  -7.3580505  0.1247172 0.0616759
## E-B -10.683333 -14.4247172 -6.9419495 0.0000001
## D-C   3.250000  -0.4913839  6.9913839 0.1110806
## E-C  -3.816667  -7.5580505 -0.0752828 0.0439570
## E-D  -7.066667 -10.8080505 -3.3252828 0.0000829
```

```r
plot(md1_tky, las = 1)
```



We see that alloy `E`, which has the lowest median, is significantly different from the rest. On the other hand, alloy `B`, which has highest median, is different from `A`, `C`, and `E`, but not from `D`. The rest of comparisons are not significant at the 5% level.

(i) If stronger is better, which alloy would you recommend and why?

Either `B` or `D`, which have the highest medians and cell means, but are not distinguishable according to the tests above.

Exercise 4

A pharmaceutical company did an experiment to compare three different pain relievers for treating migraines. The data is stored in the file `migraine`. In the experiment, 27 volunteers participated, and nine were randomly selected for each pain reliever. The subjects were instructed to take the drug during their next migraine headache episode and to report their pain on a scale of 1 = no pain to 10 = extreme pain 30 minutes after taking the drug.

Read the data file into a data frame named `df2`. Make sure the data are read correctly. If `Drug` has character mode, transform it into a factor.

```
df2 <- read.csv('migraine.csv', header = T)
str(df2)
```

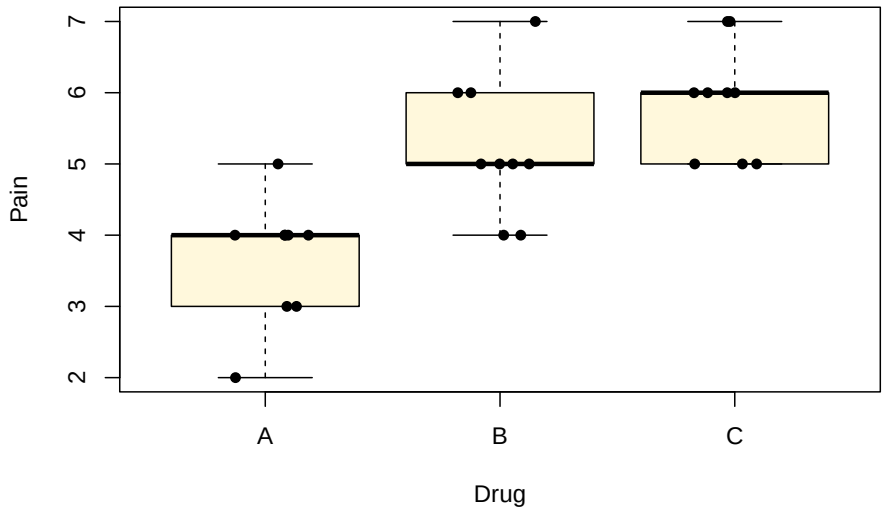
```
## 'data.frame':    27 obs. of  2 variables:
##  $ Pain: int  4 5 4 3 2 4 3 4 4 6 ...
##  $ Drug: chr  "A" "A" "A" "A" ...
```

```
df2$Drug <- factor(df2$Drug)
str(df2)
```

```
## 'data.frame':    27 obs. of  2 variables:
##  $ Pain: int  4 5 4 3 2 4 3 4 4 6 ...
##  $ Drug: Factor w/ 3 levels "A","B","C": 1 1 1 1 1 1 1 1 1 2 ...
```

(a) Do boxplots for `Pain` as a function of `Drug`. Add the points to this graph. Comment on what you observe.

```
boxplot(Pain ~ Drug, data = df2, col = 'cornsilk')
points(Pain ~ jitter(as.numeric(Drug)), data = df2, pch = 16)
```



We observe from the graph that drugs B and C have distributions with similar range of values, while drug A has lower values and is likely to have a different effect on `Pain` than the other two. The assumption of equal variances seems valid.

(b) Fit an analysis of variance model to this data using the function `lm` and print the anova table. Use $\alpha = 0.02$ for your test. What do you conclude from this analysis?

```
model2 <- lm(Pain ~ Drug, data = df2)
anova(model2)
```

```
## Analysis of Variance Table
##
## Response: Pain
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Drug       2 23.407 11.7037  15.229 5.368e-05 ***
## Residuals 24 18.444  0.7685
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p value is small and we reject the null hypothesis that all treatment levels have the same effect on `Pain`. At least two of the drugs have different effects.

(c) What are the estimated values for the variance and standard deviation of the errors in this experiment?

The estimated value for the variance of the errors is the mean square error, which can be read from the anova table: 0.7685. The standard deviation is the square root of this number:

```
sqrt(0.7685)
```

```
## [1] 0.8766413
```

(d) Using the function `summary`, obtain a summary table for the model fitted in (b). What is the meaning of the numbers in the `Estimate` column? Obtain the estimate for the mean response for each treatment from this table. Obtain also the effects.

```
summary(model2)
```

```
##
## Call:
## lm(formula = Pain ~ Drug, data = df2)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6667 -0.6667  0.1111  0.3333  1.7778
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.6667     0.2922  12.548 4.93e-12 ***
## DrugB         1.5556     0.4133   3.764 0.000954 ***
## DrugC         2.2222     0.4133   5.377 1.60e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8767 on 24 degrees of freedom
## Multiple R-squared:  0.5593, Adjusted R-squared:  0.5226
## F-statistic: 15.23 on 2 and 24 DF,  p-value: 5.368e-05
```

Let us use the notation $\hat{\mu}_A, \hat{\mu}_B$, and $\hat{\mu}_C$ for the mean observed responses for drugs A, B, and C. The first value in the `Estimate` column –named `(Intercept)`– corresponds to the average value for the observations corresponding to drug A, i.e., this is $\hat{\mu}_A$. The second and third rows in the `Estimate` column correspond to differences between the averages of observed values for drugs B and C and the average for drug A, i.e., they correspond to $\hat{\mu}_B - \hat{\mu}_A$ and $\hat{\mu}_C - \hat{\mu}_A$, respectively. Therefore, to obtain the mean response for drug B we need to add the value for drug A (`Intercept`) and the value for drug B (`DrugB`), and similarly for drug C:

$$\hat{\mu}_B = 3.667 + 1.556 = 5.223, \quad \text{and} \quad \hat{\mu}_C = 3.667 + 2.222 = 5.889$$

The overall mean is

```
(mn2 <- mean(df2$Pain))
```

```
## [1] 4.925926
```

The effects are

```
3.6667 - mn2; 5.223 - mn2; 5.889 - mn2
```

```
## [1] -1.259226
## [1] 0.2970741
## [1] 0.9630741
```

(e) Write down an equation for the anova model and explain the meaning of the parameters in this particular case.

There are three different pain relievers in the experiment, so we have three treatment levels. There are 9 observations (subjects) per treatment level. The equation is

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \quad i = A, B, C; 1 \leq j \leq 9,$$

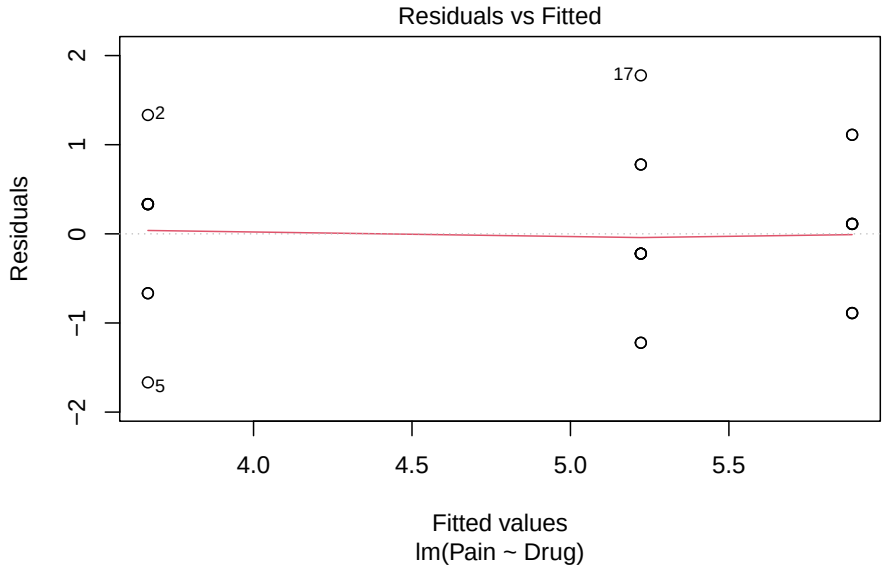
where y_{ij} is the measured pain for subject j using pain reliever i, μ is the overall mean, τ_i is the effect of the i-th treatment level (pain reliever), and ε_{ij} is the error associated to the measurement of the pain for the j-th subject of level i. The values can be obtained from the results in (d):

$$\mu = 4.926, \quad \tau_A = -1.259, \quad \tau_B = 0.297, \quad \tau_C = 0.963.$$

(f) What are the assumption on which the analysis of variance model is based? Draw diagnostic plots for checking these assumptions and discuss the results.

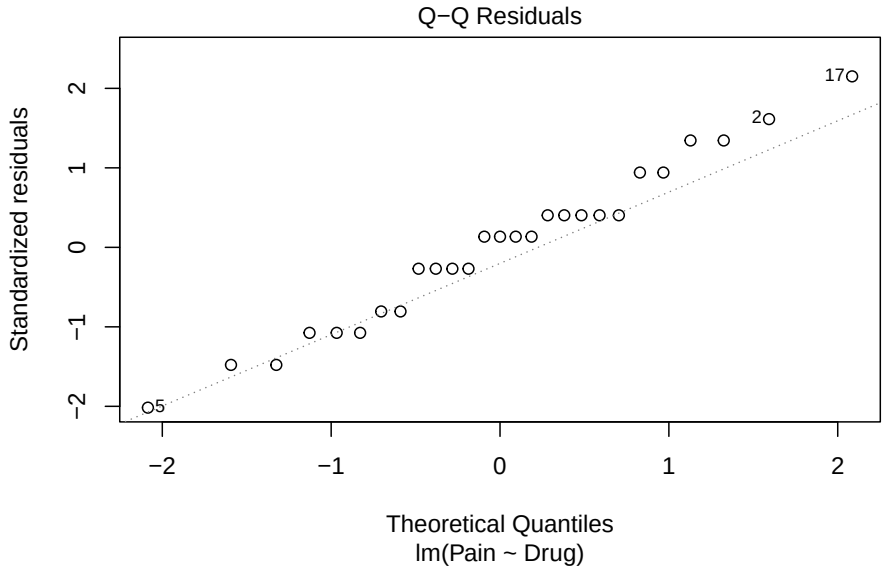
The model assumes that the errors are independent and have a common normal distribution with mean zero and variance σ^2.

```r
plot(model2, which = 1)
```



In this plot the red line is close to 0 and is almost horizontal. This shows that the residuals are centered and have a symmetric distribution. The dispersion of the three groups of points is similar,

```r
plot(model2, which = 2)
```



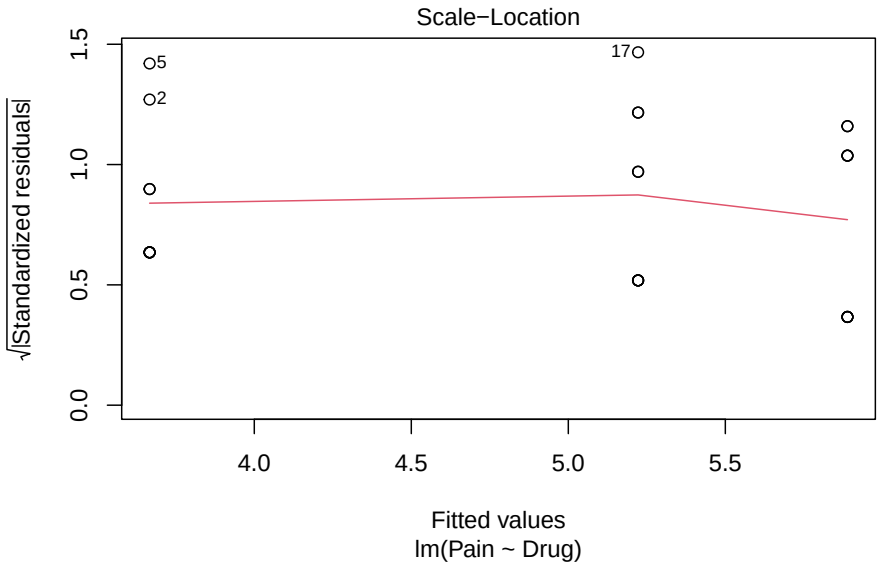
For the quantile plot, the points are close to the reference line and the graph supports the assumption of normality. To have a more solid base for this conclusion, we do a Shapiro-Wilk test on the standardized residuals

```r
shapiro.test(rstandard(model2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rstandard(model2)
## W = 0.9786, p-value = 0.8294
```

The p-value is large and we do not reject the null hypothesis of normality.

```
plot(model2, which = 3)
```



In the third plot the red line is almost horizontal, indicating that the variability within each group is similar and supporting the assumption of equal variances. There are no points beyond level 1.5, indicating that there are no outliers in the data.

We check this with the Levene test for hoscedasticity

```
library(car)
leveneTest(model2)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2  0.0851 0.9187
##       24
```

The p-value is large and we do not reject the null hypothesis of homogeneous variances.

In conclusion, the three plots and the tests support that the assumptions are valid for this experiment.

(g) We now want to use Tukey's method for pairwise comparisons, but since we fitted the model using `lm`, the command `TuketHSD` does not work. There are several alternatives, and we introduce one based on the `emmeans` package, a powerful tool for the analysis of statistically designed experiments. We load the package and calculate the marginal means for the model. The margina means are the means that correspond to each treatment level

```
library(emmeans)
(emDrug <- emmeans(model2, 'Drug'))
```

```
##  Drug emmean    SE df lower.CL upper.CL
##  A      3.67 0.292 24     3.06     4.27
##  B      5.22 0.292 24     4.62     5.83
##  C      5.89 0.292 24     5.29     6.49
##
## Confidence level used: 0.95
```

We get the marginal means and some additional information. To get pairwise comparisons using Tukey's methods, execute the following commands:

```
summary(contrast(emDrug, method = 'pairwise'))
```

and comment on the results

Solution:

```
summary(contrast(emDrug, method = 'pairwise'))
```

```
##  contrast estimate    SE df t.ratio p.value
##  A - B      -1.556 0.413 24  -3.764  0.0026
##  A - C      -2.222 0.413 24  -5.377  <.0001
##  B - C      -0.667 0.413 24  -1.613  0.2596
##
## P value adjustment: tukey method for comparing a family of 3 estimates
```

We see that Drug A is different from B and C, but B and C are not distinguishable. Since A gets lower pain scores, it is preferable.