STAT 210
Applied Statistics and Data Analysis:
Homework 5 - Solution

Due on October 19/2025

You cannot use artificial intelligence tools to solve this homework.

Show complete solutions to get full credit. Writing code is not enough to answer a question. Your comments are more important than the code. Do not write comments in chunks. Label your graphs appropriately

Question 1

An experiment was conducted to evaluate the effect of four different diets on cattle weight gain. In addition to the experimental diets, a control group was included, which received the standard diet. The collected data is stored in the file 25Fhw5Q1. The diets are labeled as follows: `ctrl` for control and `dt1`, `dt2`, `dt3`, and `dt4` for the four experimental diets.

Perform a complete analysis of variance for this set. Visualize the data using appropriate plots to help understand the distribution and group differences. Determine whether the diets have an effect on the increase in weight through a hypothesis test and state explicitly the null and alternative hypotheses in this test. Estimate the cell means and calculate the effects of each diet. Write the equation for the model and state explicitly the assumptions on which the model is based. Generate diagnostic plots and comment on any patterns or concerns you observe. Use Levene's and Shapiro-Wilk's tests also. Use Tukey's HSD procedure to make pairwise comparisons between the diets and comment on the results. Use a non-parametric alternative to the analysis of variance and compare the results. Based on the analysis, identify which diet or diets you would recommend if the objective is to maximize weight gain, and explain your reasoning.

For all tests in this homework use a significance level of $\alpha = 0.02$.

Solution

We start by reading and exploring the data

```
Q1data <- read.table('25Fhw5Q1', header = T)
str(Q1data)
```

```
## 'data.frame':    30 obs. of  2 variables:
##  $ weight: num  3.6 3.3 2.3 0.4 2.7 3.5 14.2 11.1 18.1 14.3 ...
##  $ diet  : chr  "ctrl" "ctrl" "ctrl" "ctrl" ...
```

Transform `diet` into a `factor`:

```
Q1data$diet <- factor(Q1data$diet)
str(Q1data)
```

```
## 'data.frame':    30 obs. of  2 variables:
##  $ weight: num  3.6 3.3 2.3 0.4 2.7 3.5 14.2 11.1 18.1 14.3 ...
##  $ diet  : Factor w/ 5 levels "ctrl","dt1","dt2",..: 1 1 1 1 1 1 2 2 2 2 ...
```

Look at the levels for `diet`

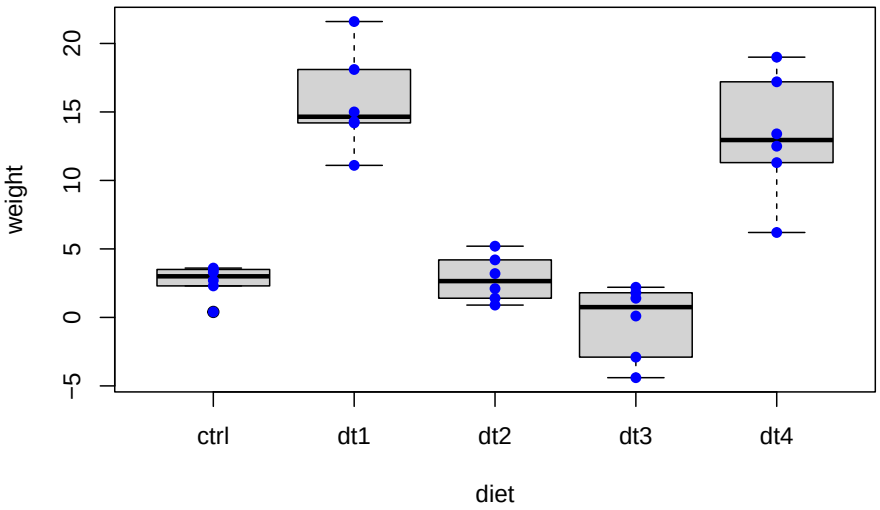
```
levels(Q1data$diet)
```

```
## [1] "ctrl" "dt1"  "dt2"  "dt3"  "dt4"
```

Exploratory Analysis

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

```
boxplot(weight ~ diet, data = Q1data)
points(weight ~ diet, data = Q1data, pch = 16, col = 'blue')
```



Diets 1 and 4 seem to produce the largest increase of weight. Diet 2 has similar values to the control group, while diet 3 seems to produce a decrease in weight, at least in some cases. If we look at the boxes and the points, it seems that the variances may be different. Particularly, the variability for the control group seems smaller than for the other diets. We will need to check for homogeneous variances later in the analysis. Also, it looks as if the differences in weight between some diets 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 weight on the j-th replicate of the i-th diet, for $i = 1, \ldots, 5$ and $j = 1, \ldots, 6$, μ is the overall mean of the observations, τ_i is the effect of the i-th diet on the weight of the animals, 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

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

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

```
model1 <- aov(weight ~ diet, data = Q1data)
summary(model1)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## diet         4 1229.0  307.24   33.74 9.68e-10 ***
## Residuals   25  227.6    9.11
## ---
## 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 9.11, with standard deviation 3.02.

Description of the Model

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

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

```
##
## Call:
## aov(formula = weight ~ diet, data = Q1data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.067 -1.496  0.100  1.617  5.883
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.633      1.232   2.138   0.0425 *
## dietdt1       13.083      1.742   7.510 7.29e-08 ***
## dietdt2        0.200      1.742   0.115   0.9095
## dietdt3       -2.933      1.742  -1.684   0.1047
## dietdt4       10.633      1.742   6.104 2.22e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.017 on 25 degrees of freedom
## Multiple R-squared:  0.8437, Adjusted R-squared:  0.8187
## F-statistic: 33.74 on 4 and 25 DF,  p-value: 9.685e-10
```

The `Intercept` corresponds to the estimated cell mean for the control diet, i.e., to $\hat{\mu} + \hat{\tau}_1$. 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 diets we can use

```
model.tables(model1, 'mean', se = T)
```

```
## Tables of means
## Grand mean
##
## 6.83
##
##  diet
## diet
##   ctrl    dt1    dt2    dt3    dt4
##  2.633 15.717  2.833 -0.300 13.267
##
## Standard errors for differences of means
##          diet
##         1.742
```

```
## replic.     6
```

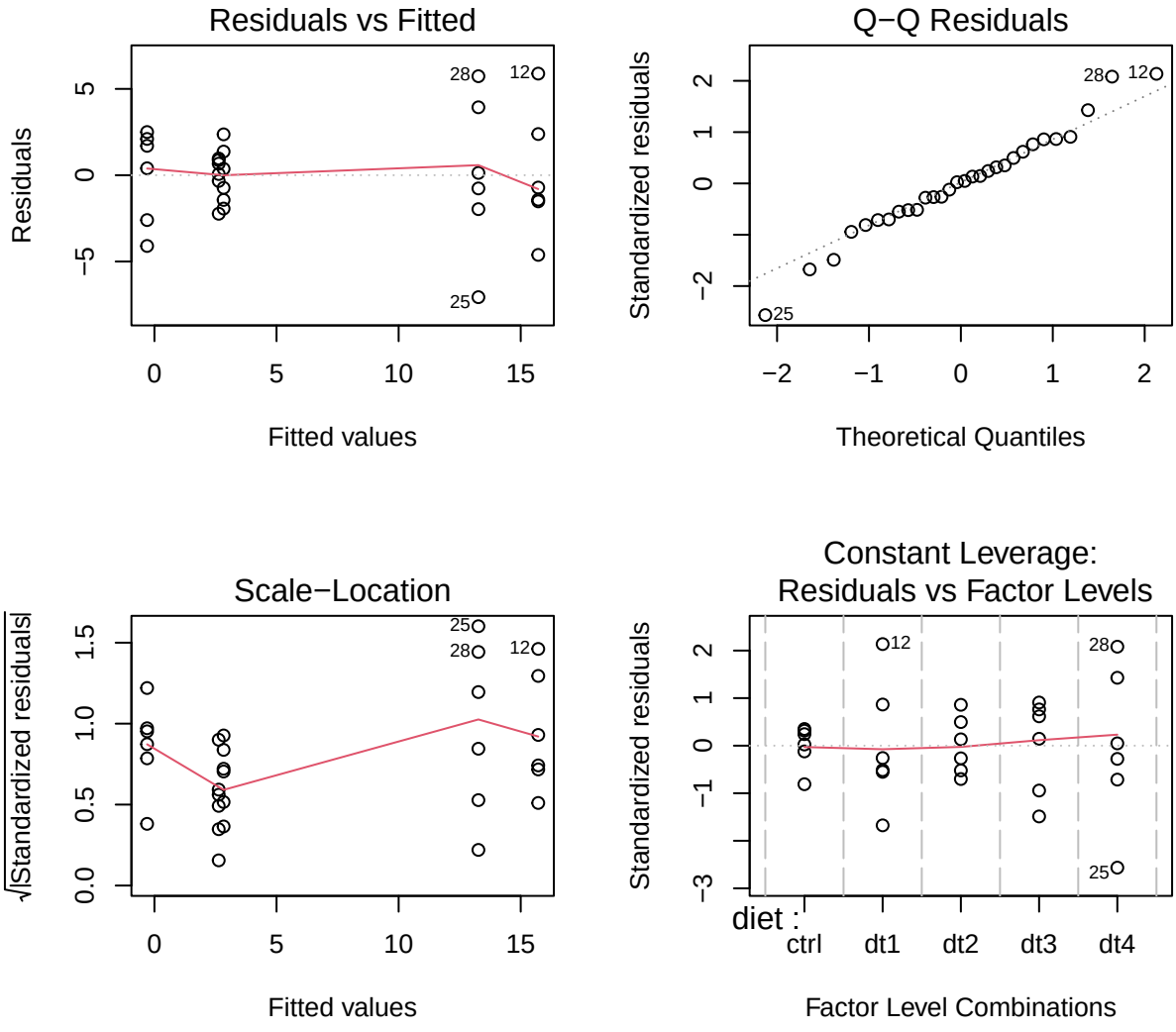
Similarly, we can obtain the estimated effects $\hat{\tau}_i, i = 1, \ldots, 5$:

```
model.tables(model1, se = T)
```

```
## Tables of effects
##
##  diet
## diet
##   ctrl    dt1    dt2    dt3    dt4
## -4.197  8.887 -3.997 -7.130  6.437
##
## Standard errors of effects
##          diet
##         1.232
## replic.     6
```

Diagnostic Plots

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



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

The first plot shows an almost horizontal red line, close to 0, except for a dip at the end. The residuals are approximately centered and symmetric. The variability of the residuals seems to be increase for higher fitted values. 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(model1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rstandard(model1)
## W = 0.97871, p-value = 0.7903
```

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

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

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  4  1.3822 0.2686
##       25
```

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

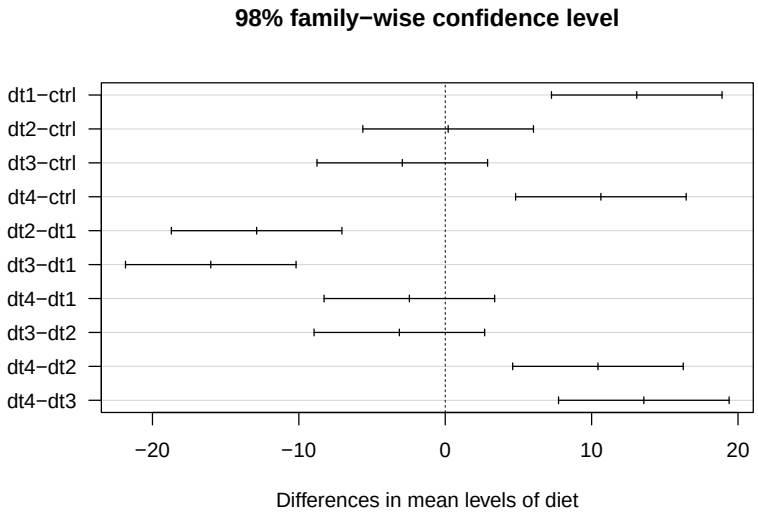
Pairwise comparisons

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

```
(model1.tky <- TukeyHSD(model1, conf.level = 0.98))
```

```
##   Tukey multiple comparisons of means
##     98% family-wise confidence level
##
## Fit: aov(formula = weight ~ diet, data = Q1data)
##
## $diet
##                 diff        lwr        upr     p adj
## dt1-ctrl  13.083333   7.257211  18.909456 0.0000007
## dt2-ctrl   0.200000  -5.626123   6.026123 0.9999578
## dt3-ctrl  -2.933333  -8.759456   2.892789 0.4613964
## dt4-ctrl  10.633333   4.807211  16.459456 0.0000205
## dt2-dt1  -12.883333 -18.709456  -7.057211 0.0000009
## dt3-dt1  -16.016667 -21.842789 -10.190544 0.0000000
## dt4-dt1   -2.450000  -8.276123   3.376123 0.6294387
## dt3-dt2   -3.133333  -8.959456   2.692789 0.3965289
## dt4-dt2   10.433333   4.607211  16.259456 0.0000274
## dt4-dt3   13.566667   7.740544  19.392789 0.0000004
```

```
plot(model1.tky, las = 1 )
```



We see that `dt1` and `dt4` are significantly different from the rest, but they are not distinguishable. On the other hand, `ctrl`, `dt2`, and `dt3` are not distinguishable.

Non-parametric test

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

```
kruskal.test(weight ~ diet, data = Q1data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  weight by diet
## Kruskal-Wallis chi-squared = 23.342, df = 4, p-value = 0.0001082
```

which leads to the same conclusion as the anova model.

Conclusion

Since we want more weight, we should go for either of `dt1` or `dt4`. These two have significantly higher cell means (15.717 and 13.267) than the other three diets. The pairwise comparison between `dt1` and `dt4` shows that they are not significantly different at the 98% confidence level, and therefore we cannot distinguish them. It is up to the manufacturer to decide which one to pick, or a further experiment could be designed to compare these two diets.

Question 2

A researcher wants to test whether different types of fertilizers have different effects on plant growth. Four types of fertilizers (A, B, C, D) plus a control group (no fertilizer) were tested. For each fertilizer, five plants are treated and their growth (height in cm) after 30 days is recorded. The results are stored in the file `25Fhw5Q2`.

Perform a complete analysis of variance for this set. Visualize the data using appropriate plots to help understand the distribution and group differences. Determine whether the fertilizers have an effect on the increase in height through a hypothesis test and state explicitly the null and alternative hypotheses in this test. Estimate the cell means and calculate the effects of each fertilizer. Write the equation for the model and state explicitly the assumptions on which the model is based. Generate diagnostic plots and comment on any patterns or concerns you observe. Use Levene's and Shapiro-Wilk's tests also. Use Tukey's HSD procedure to make pairwise comparisons between the fertilizers and comment on the results. Use a non-parametric alternative to the analysis of variance and compare the results. Based on the analysis, identify which fertilizer or fertilizers you would recommend if the objective is to maximize height, and explain your reasoning.

For all tests in this homework use a significance level of $\alpha = 0.05$.

Solution

We start by reading and exploring the data

```
Q2data <- read.table('25Fhw5Q2', header = T)
str(Q2data)
```

```
## 'data.frame':    25 obs. of  2 variables:
##  $ height    : num  24.7 23.6 24.7 25.2 26.5 27.7 29.8 29.8 28 26.4 ...
##  $ fertilizer: chr  "ctrl" "ctrl" "ctrl" "ctrl" ...
```

Transform `diet` into a `factor`:

```
Q2data$fertilizer <- factor(Q2data$fertilizer)
str(Q2data)
```

```
## 'data.frame':    25 obs. of  2 variables:
##  $ height    : num  24.7 23.6 24.7 25.2 26.5 27.7 29.8 29.8 28 26.4 ...
##  $ fertilizer: Factor w/ 5 levels "ctrl","ft1","ft2",..: 1 1 1 1 1 2 2 2 2 2 ...
```

Look at the levels for `fertilizer`

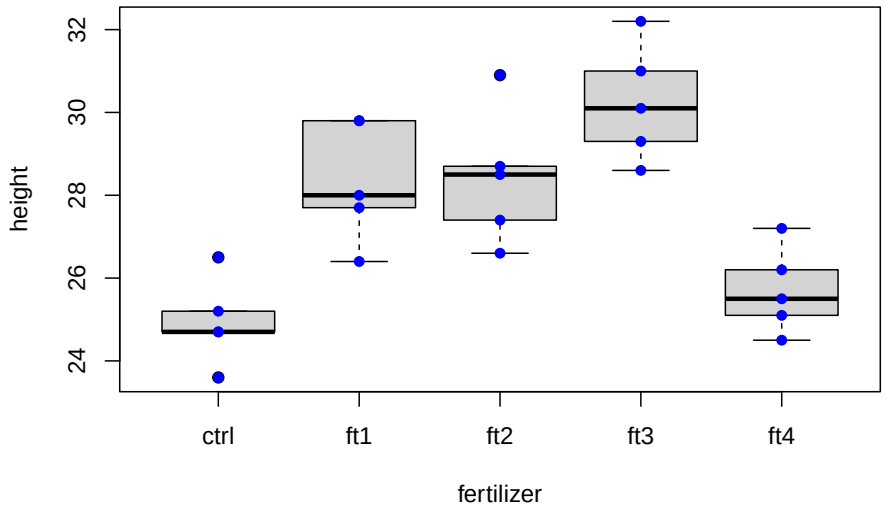
```
levels(Q2data$fertilizer)
```

```
## [1] "ctrl" "ft1"  "ft2"  "ft3"  "ft4"
```

Exploratory Analysis

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

```
boxplot(height ~ fertilizer, data = Q2data)
points(height ~ fertilizer, data = Q2data, pch = 16, col = 'blue')
```



Fertilizers 1, 2, and 3 seem to produce the largest increase of height, while fertilizer4 has similar values to the control group. If we look at the boxes and the points, it seems that the variances are similar. Also, it looks as if the differences in weight between some diets 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 height on the j-th replicate of the i-th diet, for $i = 1, \ldots, 5$ and $j = 1, \ldots, 5$, μ is the overall mean of the observations, τ_i is the effect of the i-th fertilizer on the height of the plants, 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

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

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

```
model2 <- aov(height ~ fertilizer, data = Q2data)
summary(model2)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## fertilizer   4  94.25  23.562   13.12 2.1e-05 ***
## Residuals   20  35.90   1.795
## ---
## 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 1.795, with standard deviation 1.34.

Description of the Model

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

```
summary.lm(model2)
```

```
##
## Call:
## aov(formula = height ~ fertilizer, data = Q2data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -1.94  -0.94  -0.20   0.76   2.48
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    24.9400     0.5992  41.622  < 2e-16 ***
## fertilizerft1   3.4000     0.8474   4.012 0.000684 ***
## fertilizerft2   3.4800     0.8474   4.107 0.000548 ***
## fertilizerft3   5.3000     0.8474   6.254 4.16e-06 ***
## fertilizerft4   0.7600     0.8474   0.897 0.380460
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.34 on 20 degrees of freedom
## Multiple R-squared:  0.7241, Adjusted R-squared:  0.669
## F-statistic: 13.12 on 4 and 20 DF,  p-value: 2.104e-05
```

The `Intercept` corresponds to the estimated cell mean for the control fertilizer, i.e., to $\hat{\mu}+\hat{\tau}_1$. 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 fertilizers we can use

```
model.tables(model2, 'mean', se = T)
```

```
## Tables of means
## Grand mean
##
## 27.528
##
##  fertilizer
## fertilizer
##  ctrl   ft1   ft2   ft3   ft4
## 24.94 28.34 28.42 30.24 25.70
##
## Standard errors for differences of means
##         fertilizer
##             0.8474
## replic.          5
```

Similarly, we can obtain the estimated effects $\hat{\tau}_i, i = 1, \ldots, 5$:

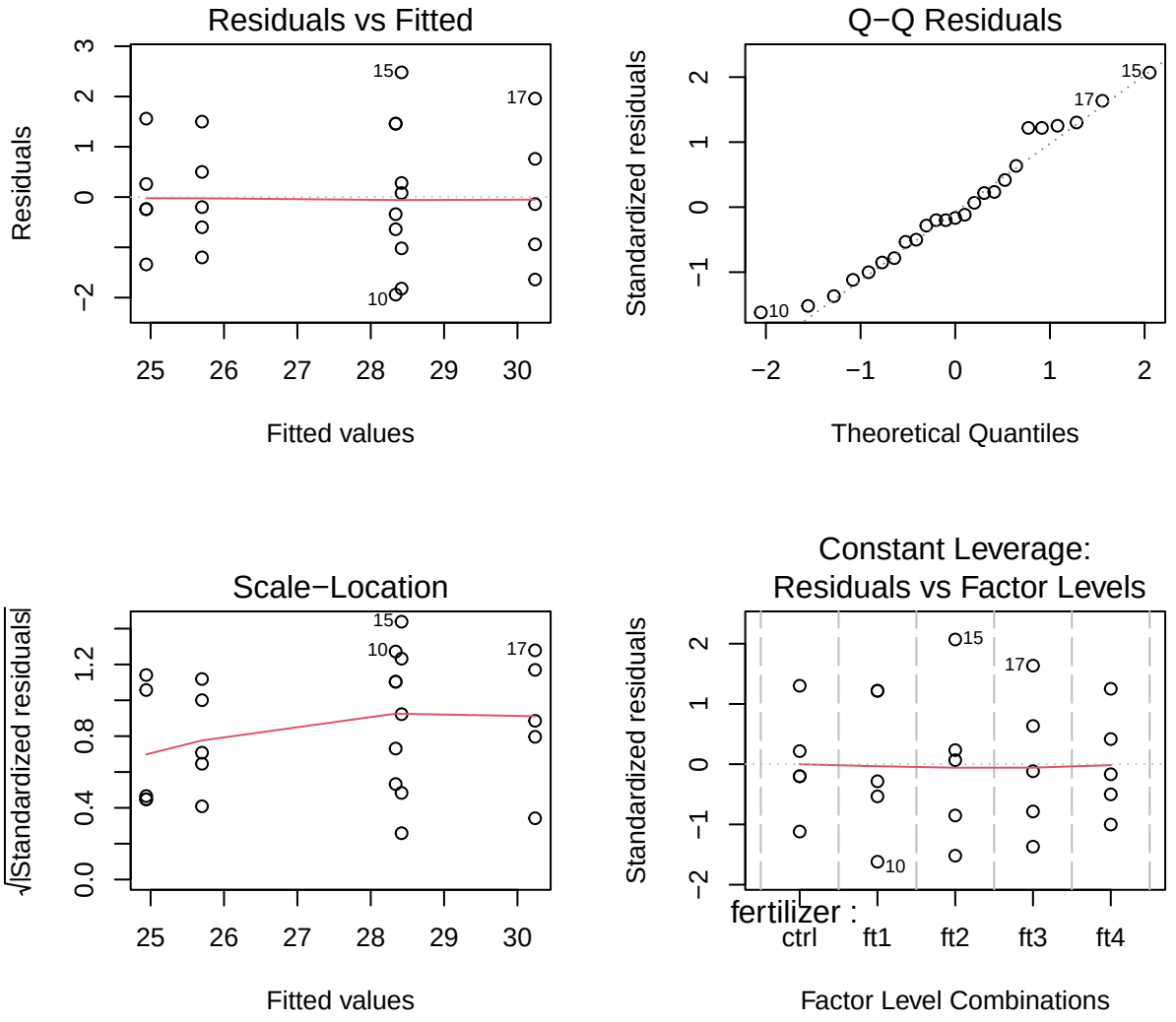
```
model.tables(model2, se = T)
```

```
## Tables of effects
##
##  fertilizer
## fertilizer
##   ctrl    ft1    ft2    ft3    ft4
## -2.588  0.812  0.892  2.712 -1.828
##
```

```
## Standard errors of effects
##         fertilizer
##             0.5992
## replic.          5
```

Diagnostic Plots

```
par(mfrow=c(2,2))
plot(model2)
```




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

The first plot shows a horizontal red line, close to 0, indicating that the residuals are approximately centered and symmetric. The variability of the residuals seems to be similar for all fitted values. 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(model2))
```

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

```
##
## data:  rstandard(model2)
## W = 0.96395, p-value = 0.4985
```

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

```
leveneTest(model2)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  4  0.3086 0.8688
##       20
```

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

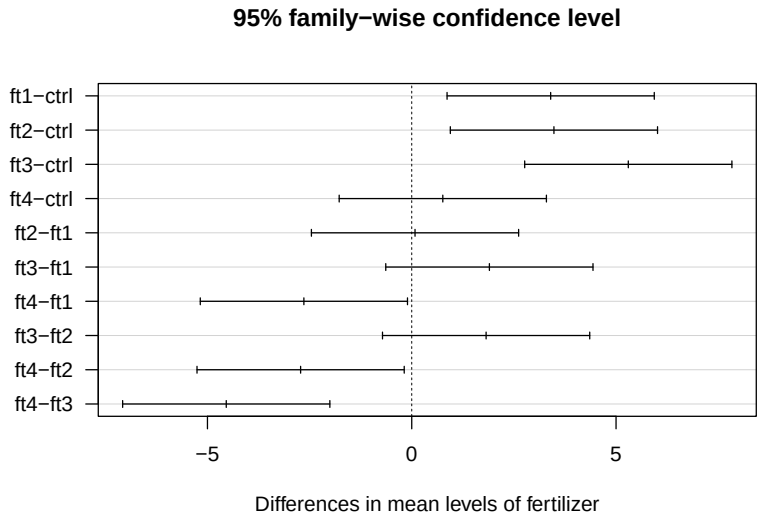
Pairwise comparisons

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

```
(model2.tky <- TukeyHSD(model2, conf.level = 0.95))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = height ~ fertilizer, data = Q2data)
##
## $fertilizer
##           diff        lwr        upr     p adj
## ft1-ctrl  3.40  0.8642737  5.9357263 0.0054814
## ft2-ctrl  3.48  0.9442737  6.0157263 0.0044333
## ft3-ctrl  5.30  2.7642737  7.8357263 0.0000373
## ft4-ctrl  0.76 -1.7757263  3.2957263 0.8947003
## ft2-ft1   0.08 -2.4557263  2.6157263 0.9999803
## ft3-ft1   1.90 -0.6357263  4.4357263 0.2050977
## ft4-ft1  -2.64 -5.1757263 -0.1042737 0.0387760
## ft3-ft2   1.82 -0.7157263  4.3557263 0.2395648
## ft4-ft2  -2.72 -5.2557263 -0.1842737 0.0318058
## ft4-ft3  -4.54 -7.0757263 -2.0042737 0.0002647
```

```
plot(model2.tky, las = 1 )
```



At the 5 % level, fertilizers 1, 2, and 3 are significantly better than the control or fertilizer 4, since they produce taller plants. We cannot distinguish among 1, 2, and 3 on the basis of this sample. Also, fertilizer 4 and the control are not significantly different.

Non-parametric test

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

```
kruskal.test(height ~ fertilizer, data = Q2data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  height by fertilizer
## Kruskal-Wallis chi-squared = 17.951, df = 4, p-value = 0.001262
```

which leads to the same conclusion as the anova model.

Conclusion

Since we want more taller plants, we should go for either of `ft1`, `ft2`, or `dt3`. These three have significantly higher cell means (28.34, 28.42, and 30.24) than the other two fertilizers. The pairwise comparison between `ft1`, `ft2`, and `ft3` shows that they are not significantly different at the 95% confidence level, and therefore we cannot distinguish them. It is up to the manufacturer to decide which one to pick, or a further experiment could be designed to compare these three fertilizers.