STAT 210
Applied Statistics and Data Analysis:
Homework 5

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

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

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.

We begin by loading the required package and reading the data.

```
library(car)
```

```
## Loading required package: carData
```

```
data <- read.table("25Fhw5Q1", header = TRUE);
str(data)
```

```
## '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" ...
```

```
#change up the diet column into type factor.

data$diet <- as.factor(data$diet)

str(data)
```

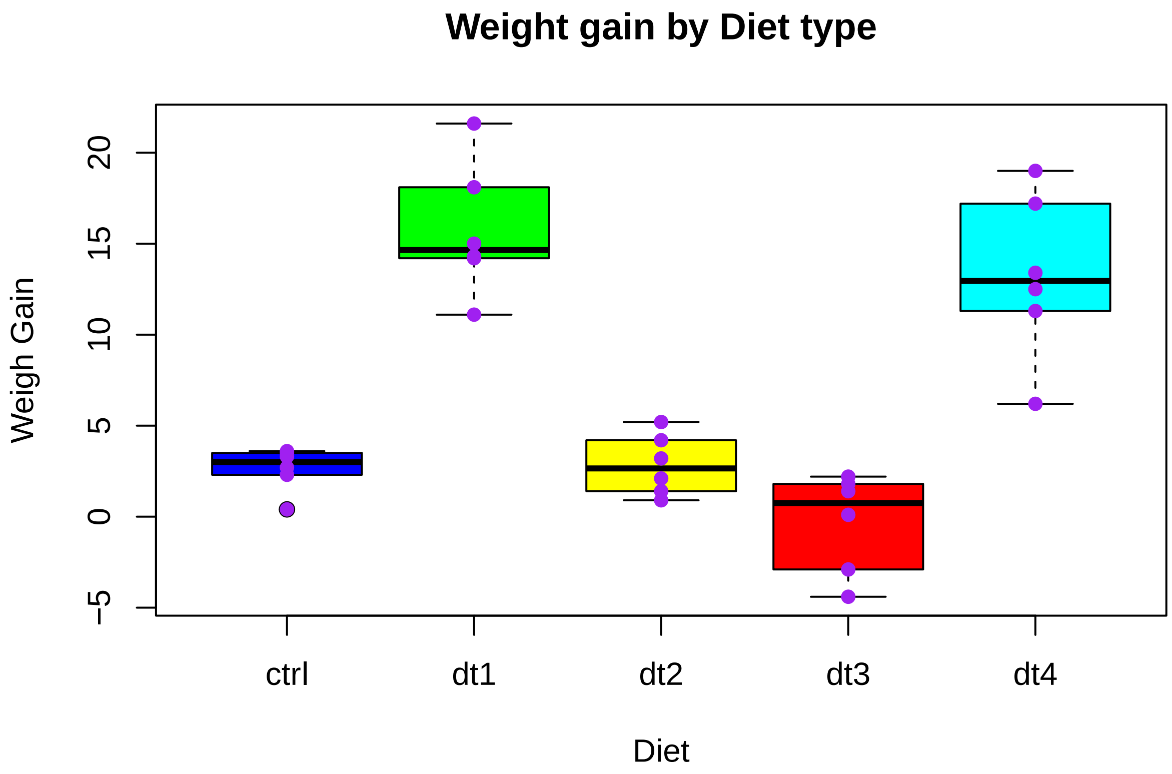
```
## 'data.frame':    30 obs. of  2 variables:
##  $ weight: num  3.6 3.3 2.3 0.4 2.7 3.5 14.2 11.1 18.1 14.3 ...
##  $ diet  : Factor w/ 5 levels "ctrl","dt1","dt2",..: 1 1 1 1 1 1 2 2 2 2 ...
```

```
levels(data$diet)
```

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

```
#here we show the data visually as a box plot and as points on top of the same plot.
boxplot(weight ~ diet, data = data,
        main = "Weight gain by Diet type",
        xlab = "Diet",
        ylab = "Weigh Gain",
        col = c("blue", "green", "yellow", "red", "cyan"))

points(weight ~ diet, data = data, pch = 16, col = 'purple')
```



For the hypothesis test, we choose the null hypothesis H0: mu_ctrl = mu_dt1 = mu_dt2 = mu_dt3 = mu_dt4. Which means that our hypothesis is that all diets have the same mean gain.

Thus the alternative hypothesis would be that at least one of the diets has a different mean, which means it probably had an effect on weight gain.

We then fit the ANOVA model to the data.

```
mod1 <- aov(weight ~ diet, data = data)
summary(mod1)
```

```
##             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
```

Since we have a very small p-value (9.68e-10), then we reject the null hypothesis which was that all diets are the same, or in other words, the diets had no effect.

We next try to find the cell means and analyze further. We use model.tables with argument “mean” to find the cell mean of the different types of diets.

```
model.tables(mod1, '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
```

To find the effects of each diet we use the same function without the “mean” argument.

```
model.tables(mod1, 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
```

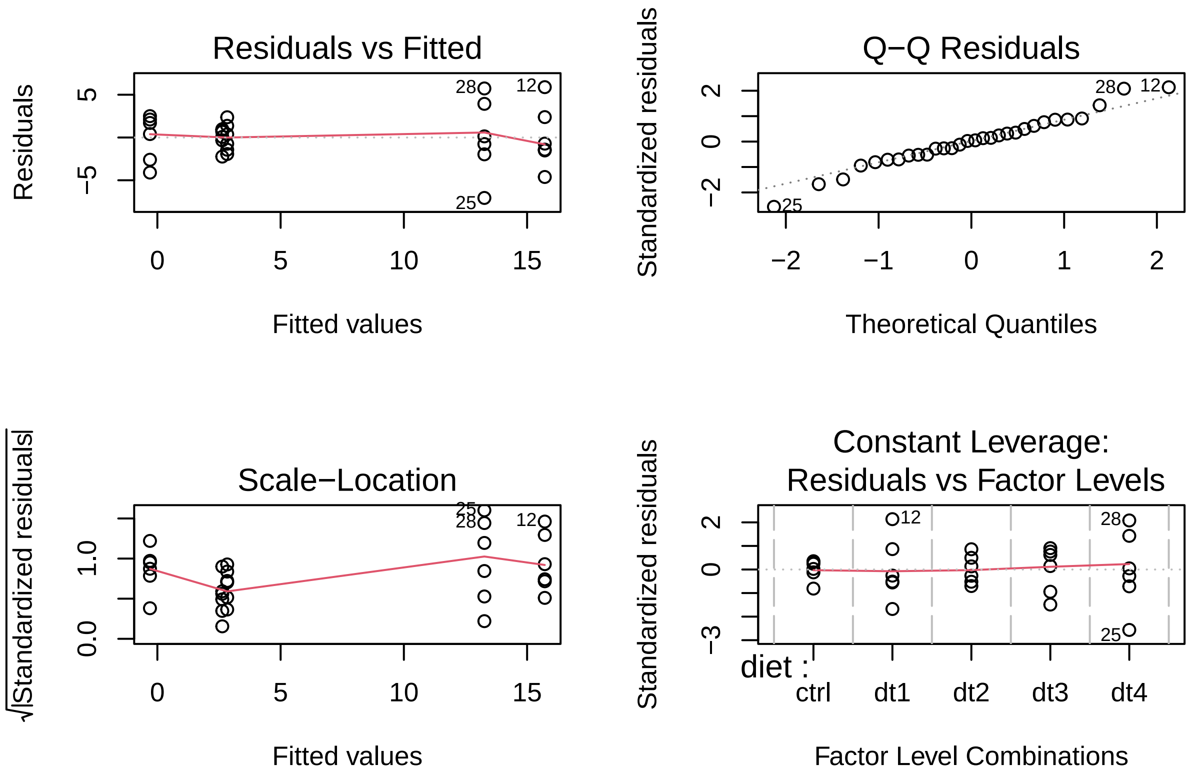
Equation of the ANOVA model: Y_ij = mu + tau_i + epsilon_ij

Where: Y_ij = weight gain for the j-th observation in the i-th diet group mu = overall mean weight gain tau_i = effect of the i-th diet (deviation from overall mean) epsilon_ij = random error term i = 1, 2, 3, 4, 5 (diet groups: ctrl, dt1, dt2, dt3, dt4) j = 1, 2, …, n_i (observations within each group)

Assumptions for the ANOVA model: Independence, Normality, Homogeneity of Variance.

Diagnostic Plots:

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



Residual vs fitted shows the trend line being close to 0 which indicates linearity, although there is some variation in the spread of residuals which violates homogeneity of variance; further testing may be needed. Q-Q plot shows the points lying roughly on the line, which suggests normality. The scale-location plot shows somewhat of an issue with our equal variance assumption, requiring further testing. Constant leverage plot shows all points within the dotted lines which shows no influential outliers.

Levene's test for homogeneity of variance. H0 is that all groups have equal variance. H1 is at least one group has a different variance.

```r
leveneTest(mod1)
```

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

With p-value = 0.2686, we accept the null hypothesis and that the homogeneity of variance assumption is valid.

Shapiro-Wilk test for normality. H0: Residuals are normally distributed. H1: Residuals are not normally distributed.

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

```
##
##  Shapiro-Wilk normality test
##
## data:  rstandard(mod1)
## W = 0.97871, p-value = 0.7903
```

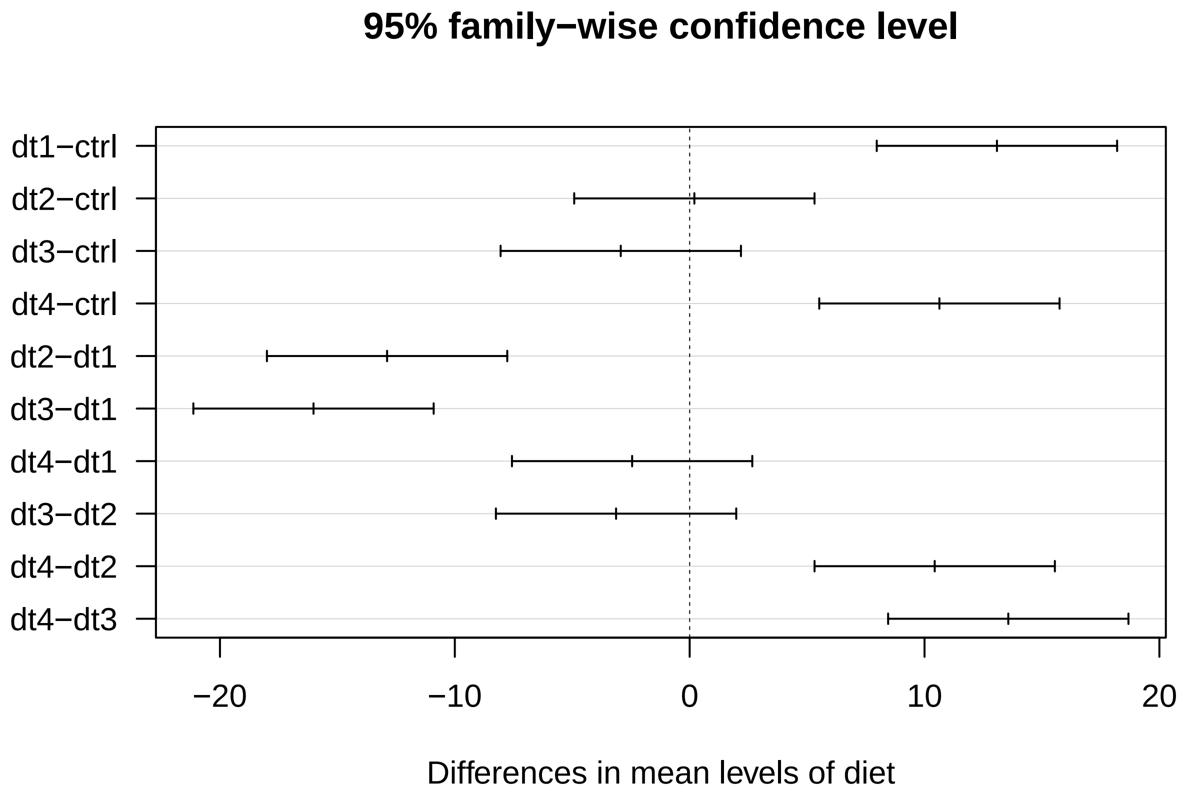
With p-value = 0.7903, we accept the null hypothesis and that the distribution of residuals is normal.

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

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = weight ~ diet, data = data)
##
## $diet
##                 diff        lwr        upr     p adj
## dt1-ctrl  13.083333   7.966899  18.199768 0.0000007
## dt2-ctrl   0.200000  -4.916435   5.316435 0.9999578
## dt3-ctrl  -2.933333  -8.049768   2.183101 0.4613964
## dt4-ctrl  10.633333   5.516899  15.749768 0.0000205
## dt2-dt1  -12.883333 -17.999768  -7.766899 0.0000009
## dt3-dt1  -16.016667 -21.133101 -10.900232 0.0000000
## dt4-dt1   -2.450000  -7.566435   2.666435 0.6294387
## dt3-dt2   -3.133333  -8.249768   1.983101 0.3965289
## dt4-dt2   10.433333   5.316899  15.549768 0.0000274
## dt4-dt3   13.566667   8.450232  18.683101 0.0000004
```

By analyzing Tukey's HSD procedure, we see that Diet 1 and 4 show significant difference compared to the control group. We also see that diets 2 and 3 have no significant difference to the control group and are not distinguishable. We can also see that diets 1 and 4 are also not distinguishable.

```
par(mfrow=c(1,1))
plot(mod1.tky, las = 1)
```



For the non-parametric test we use the Kruskal-Wallis test which is the non-parametric alternative to the ANOVA model.

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

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

The test gives a p-value of 0.0001082, which leads to the same conclusion as the ANOVA model.

If the goal is to maximize weight gain, Diets 1 and 4 are recommended. Diets 1 and 4 showed significantly higher average weight gains compared to the other diets and control group, with no significant difference between them.

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.

```r
data2 <- read.table("25Fhw5Q2", header = TRUE)
str(data2)
```

```
## ’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" ...
```

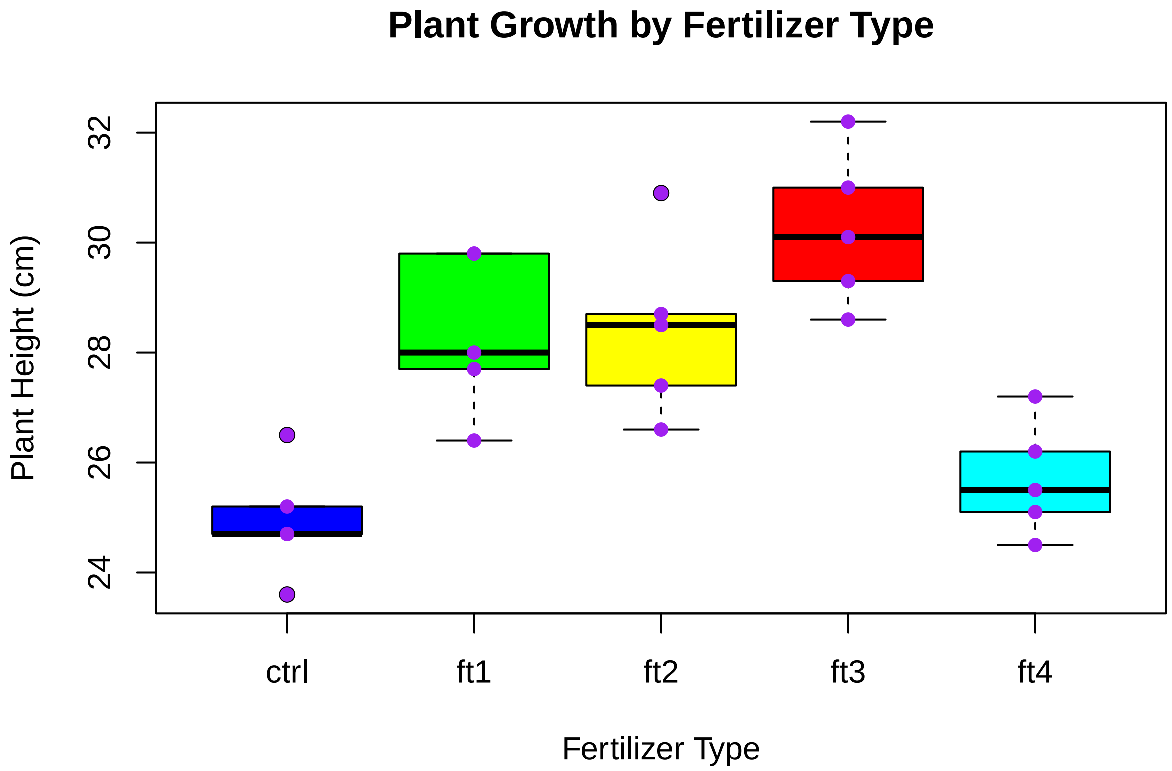
```r
data2$fertilizer <- as.factor(data2$fertilizer)
levels(data2$fertilizer)
```

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

```r
boxplot(height ~ fertilizer, data = data2,
main = "Plant Growth by Fertilizer Type",
xlab = "Fertilizer Type",
ylab = "Plant Height (cm)",
col = c("blue", "green", "yellow", "red", "cyan"))




points(height ~ fertilizer, data = data2, pch = 16, col = 'purple')
```



Hypothesis Test: H0: mu_ctrl = mu_dt1 = mu_dt2 = mu_dt3 = mu_dt4. (All fertilizers produce the same mean growth) H1: At least one fertilizer has a different mean effect on plant growth.

```
mod2 <- aov(height ~ fertilizer, data = data2)
summary(mod2)
```

```
##             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
```

With p-value = 2.1e-5, we reject the null hypothesis.

```
#Find cell means and standard errors
model.tables(mod2, "mean", se = TRUE)
```

```
## 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
```

```
#Find effects for each fertilizer
model.tables(mod2, se = TRUE)
```

```
## 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
```

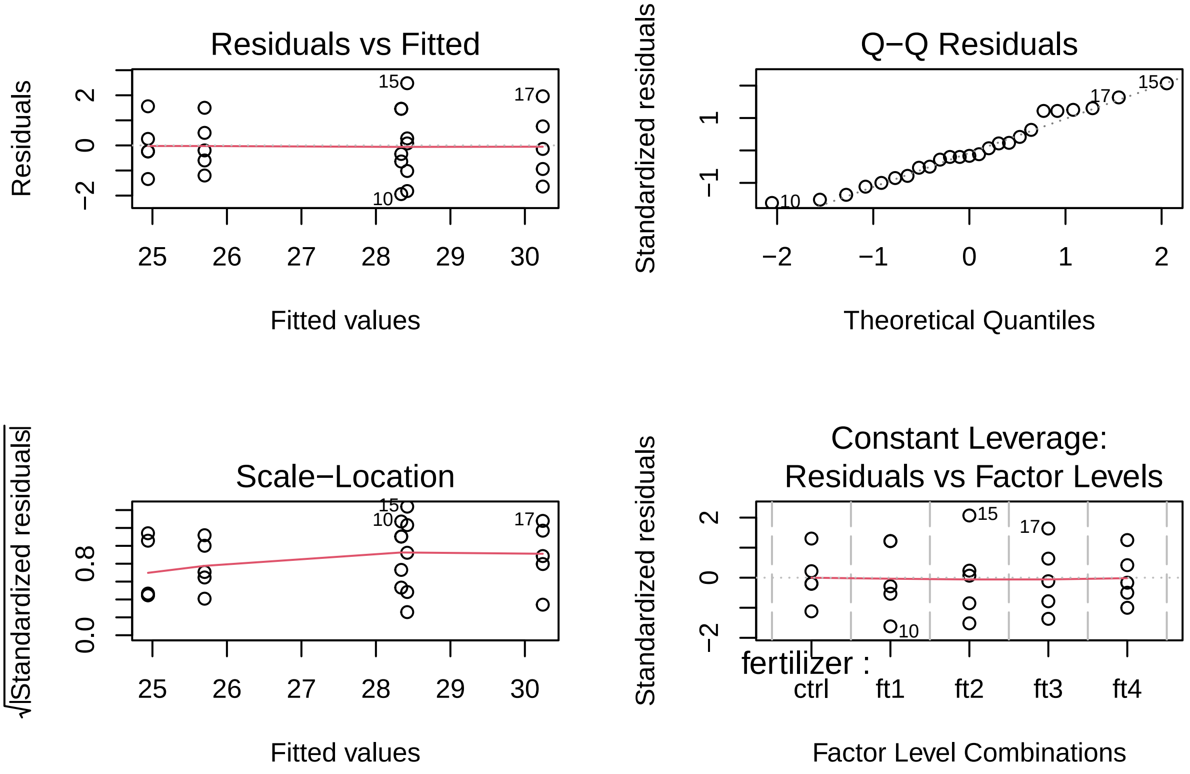
Equation of the ANOVA model: Y_ij = mu + tau_i + epsilon_ij

Where: Y_ij = weight gain for the j-th observation in the i-th diet group mu = overall mean weight gain tau_i = effect of the i-th diet (deviation from overall mean) epsilon_ij = random error term i = 1, 2, 3, 4, 5 (diet groups: ctrl, dt1, dt2, dt3, dt4) j = 1, 2, …, n_i (observations within each group)

Assumptions for the ANOVA model: Independence, Normality, Homogeneity of Variance.

Diagnostic Plots:

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



Residuals vs Fitted: shows random scatter around 0 (linearity, equal variance). Q-Q plot: points are roughly on the line, suggesting normality. Scale-location: variance is mostly constant across fitted values. Residual vs Leverage: shows no influential outliers.

Levene's Test for Homogeneity of Variance H0: all groups have equal variance H1: at least one group differs

```
leveneTest(mod2)
```

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

p-value = 0.8688, therefore we accept H0, variances are homogeneous.

Shapiro–Wilk Test for Normality of Residuals H0: residuals are normally distributed H1: residuals are not normally distributed

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

```
##
##  Shapiro-Wilk normality test
##
## data:  rstandard(mod2)
## W = 0.96395, p-value = 0.4985
```

p-value = 0.4985, therefore we accept H0, residuals are normally distributed.

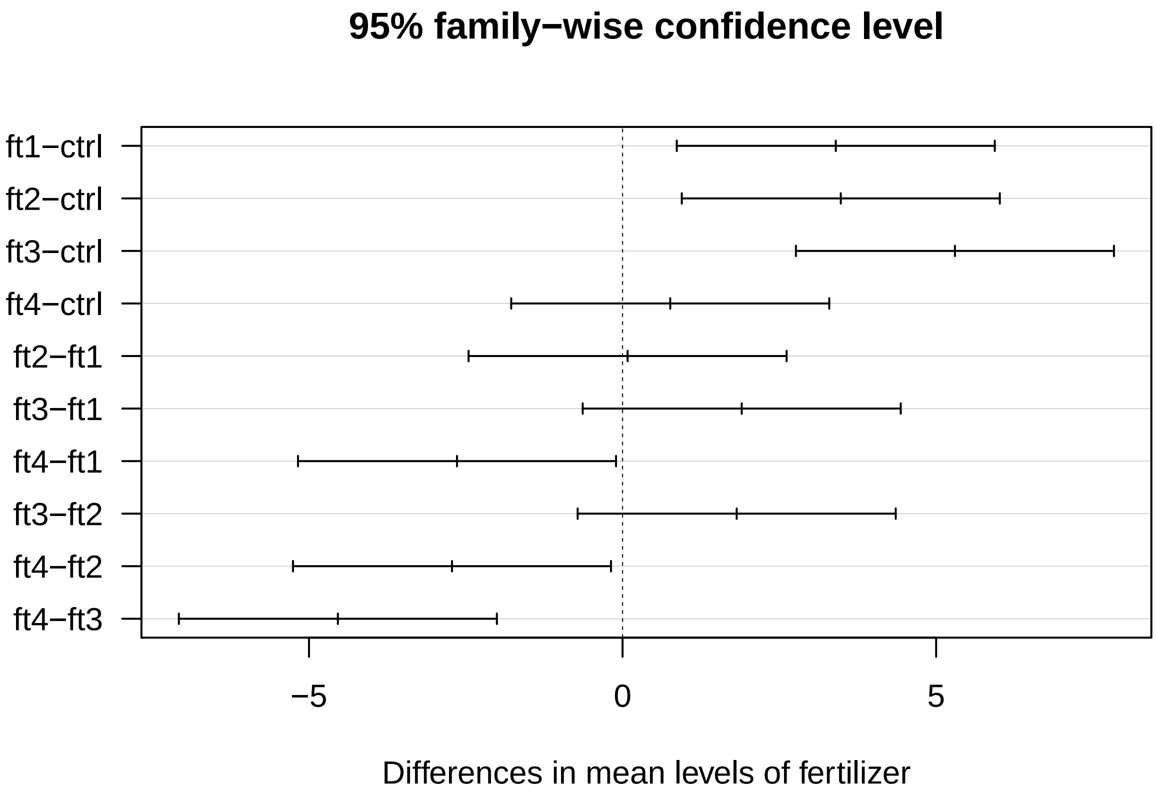
Tukey's HSD Comparisons

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

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = height ~ fertilizer, data = data2)
##
## $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
```

We see that F1 and F2 are indistinguishable and have significantly higher mean than the control group; they are also somewhat indistinguishable from F3 which had the strongest results. F4 is not significantly different from the control group. Therefore, for best performance we see F3 first, then F1 or F2, and lastly F4.

```
par(mfrow = c(1, 1))
plot(mod2.tky, las = 1)
```



Non-parametric Alternative: Kruskal–Wallis Test

```
kruskal.test(height ~ fertilizer, data = data2)
```

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

With p-value of 0.001262, it leads to the same conclusion as the ANOVA model.

For the goal of maximizing plant height, Fertilizer 3 is recommended as it has shown the most significant improvement. Fertilizers 1 and 2 also showed improvement, but not to the level of Fertilizer 3. Fertilizer 4 showed no significant improvement in plant height.