# 2020 - Exam

## Exercise Trees

### a)

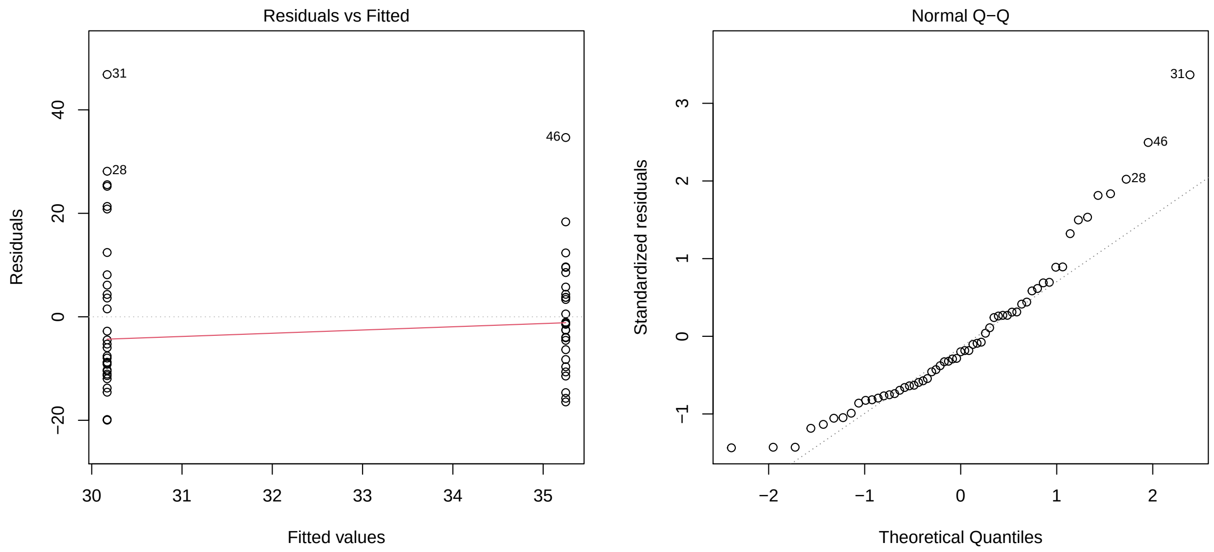
```
# read the data
data <- read.table(file="treeVolume.txt", header=TRUE)
data$type <- as.factor(data$type)
# perform one way-anova
model <- lm(volume~type, data = data)
anova(model)
```

```
## Analysis of Variance Table
##
## Response: volume
##           Df Sum Sq Mean Sq F value Pr(>F)
## type       1    380     380     1.9   0.17
## Residuals 57  11395     200
```

```
summary(model)
```

```
##
## Call:
## lm(formula = volume ~ type, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -19.97  -9.96  -2.77   5.94  46.83
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    30.17       2.54   11.88   <2e-16 ***
## typeoak         5.08       3.69    1.38     0.17
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 14.1 on 57 degrees of freedom
## Multiple R-squared:  0.0322, Adjusted R-squared:  0.0153
## F-statistic:  1.9 on 1 and 57 DF,  p-value: 0.174
```

```
par(mfrow=c(1,2));
plot(model, 1); plot(model, 2)
```



From the one-way ANOVA test above we can conclude that there is no significant effect of tree type on the volume. The estimate for *beech* type is 30.17, for *oak* type it is 30.17 + 5.08 = 35.25. Test diagnostics: Residuals vs Fitted plot look acceptable. QQ-plot poorly follows a straight line, therefore the normality here is questionable. It might be better to perfom a different test here too.

**b)**

```
# perform ancova
model <- lm(volume~diameter+height+type, data = data)
anova(model)
```

```
## Analysis of Variance Table
##
## Response: volume
##           Df Sum Sq Mean Sq F value  Pr(>F)
## diameter   1  10827   10827 1029.51 < 2e-16 ***
## height     1    346     346   32.92 4.3e-07 ***
## type       1     23      23    2.21    0.14
## Residuals 55    578      11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

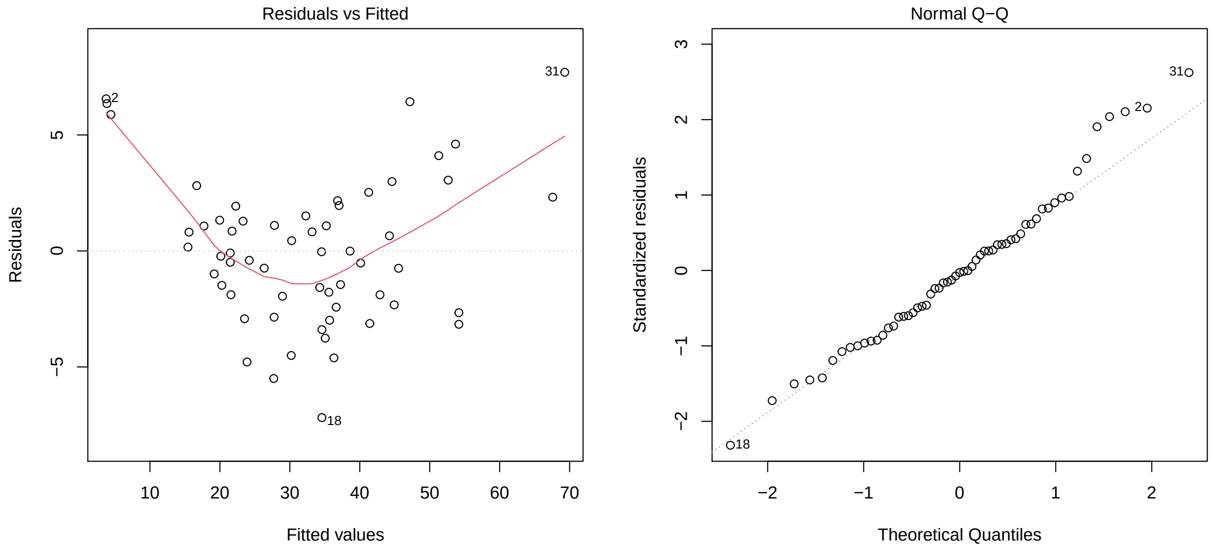
```
summary(model)
```

```
##
## Call:
## lm(formula = volume ~ diameter + height + type, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.186 -2.140 -0.087  1.721  7.701
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -63.7814     5.5129  -11.57  2.3e-16 ***
## diameter      4.6981     0.1645   28.56  < 2e-16 ***
## height        0.4172     0.0752    5.55  8.4e-07 ***
## typeoak      -1.3046     0.8779   -1.49     0.14
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.24 on 55 degrees of freedom
## Multiple R-squared:  0.951,  Adjusted R-squared:  0.948
## F-statistic:  355 on 3 and 55 DF,  p-value: <2e-16
```

ANCOVA analysis brings us to the same conclusion - there is no significant effect of the tree type on the volume. From the coefficients in the summary table it seems that oak type insignificantly results in a smaller volume.

```
# diagnostics
par(mfrow=c(1,2));
plot(model, 1); plot(model, 2)
```



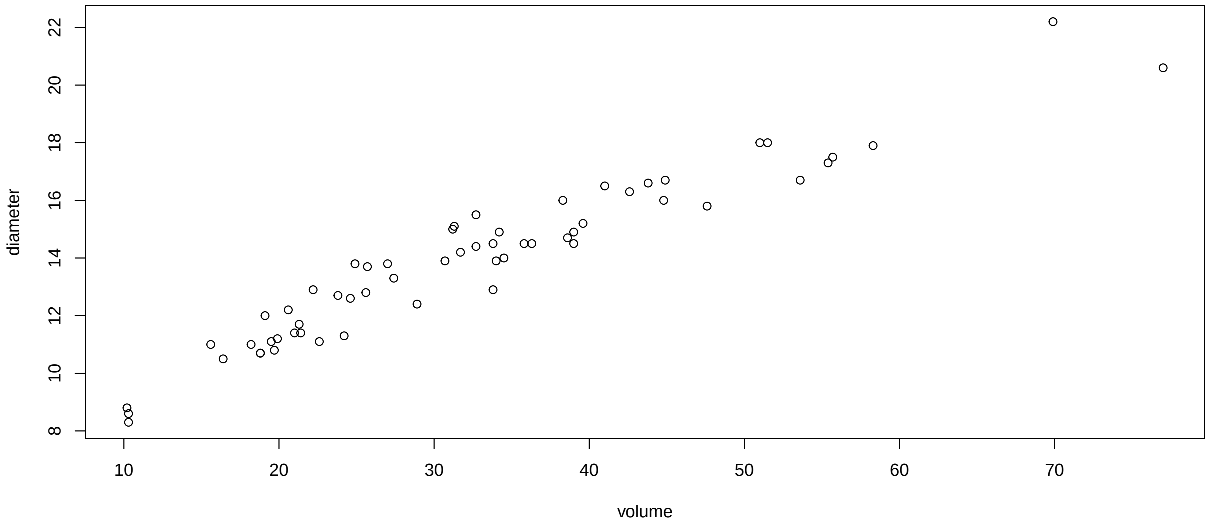
Diagnostics: from Residuals vs fitted we see some outliers that could raise doubts about normality of the data, however there does not seem to be any obvious relationship if these outliers would be removed. QQ-plot seems to follow a straight line pretty well (with some outliers).

```
# perform predictions
avg_diameter <- mean(data$diameter); avg_height <- mean(data$height)
types <- unique(as.character(data$type))
new_data <- expand.grid(diameter = avg_diameter, height = avg_height,
                        type = types)
results <- predict(model, new_data, type="response")
final <- new_data %>% bind_cols('Estimated volume' = results)
knitr::kable(final)
```

| diameter | height | type | Estimated volume |
|---:|---:|:---|---:|
| 13.9 | 75.8 | beech | 33.2 |
| 13.9 | 75.8 | oak | 31.9 |

**c)**

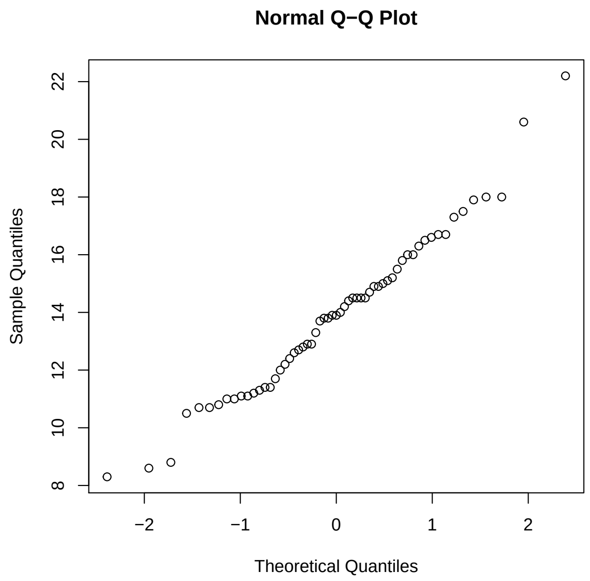
```
# plot to see relationship
plot(diameter~volume, data = data)
```

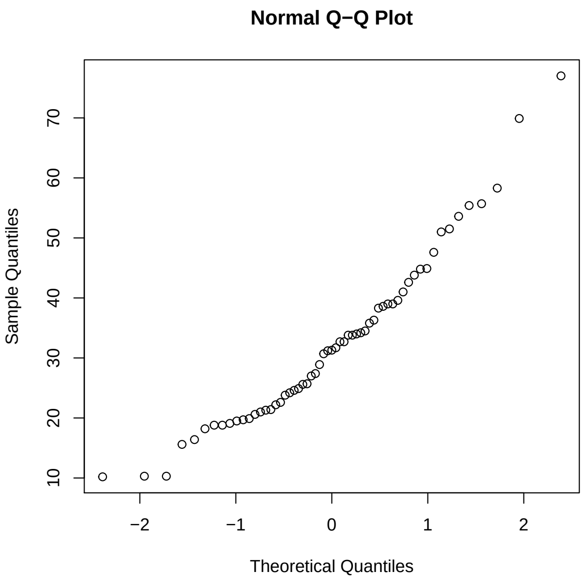


```
cor.test(data$diameter, data$volume)
```

```
##
##  Pearson's product-moment correlation
##
## data:  data$diameter and data$volume
## t = 26, df = 57, p-value <2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.932 0.975
## sample estimates:
##   cor
## 0.959
```

```
# diagnostics
par(mfrow=c(1,2)); qqnorm(data$diameter); qqnorm(data$volume)
```





From the plot above there seems to be an obvious positive linear relationship between volume and diameter. By performing Pearson correlation test we see that there is significant positive correlation between the two variables. Test diagnostics confirm normality of the data

```r
# perform ANCOVA with interaction
model <- lm(volume ~ type*diameter, data = data)
anova(model)
```

```
## Analysis of Variance Table
##
## Response: volume
##               Df Sum Sq Mean Sq F value   Pr(>F)
## type           1    380     380   23.37 1.1e-05 ***
## diameter       1  10492   10492  646.21 < 2e-16 ***
## type:diameter  1     10      10    0.59    0.45
## Residuals     55    893      16
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

From the p-value for interaction we see that it is >0.05, therefore there is no interaction between diameter and type i.e. the hypothesis that the diameter influences volume in the same way based on the tree type is not rejected.

### d)

```r
# create new variable
data_1 <- data %>% mutate(new_var = (diameter/2)**2*pi*height)

# perform ancova
model <- lm(volume~new_var+type, data = data_1)
anova(model)
```

```
## Analysis of Variance Table
```

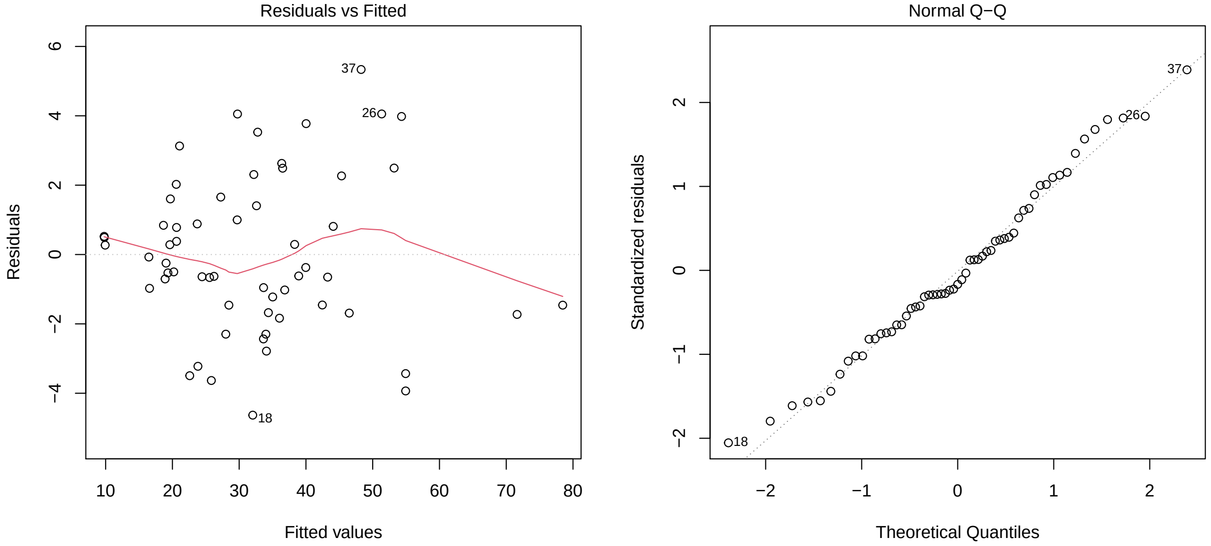
```
##
## Response: volume
##           Df Sum Sq Mean Sq F value Pr(>F)
## new_var    1  11477   11477 2183.80 <2e-16 ***
## type       1      3       3    0.56   0.46
## Residuals 56    294       5
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
summary(model)
```

```
##
## Call:
## lm(formula = volume ~ new_var + type, data = data_1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.632 -1.460 -0.375  1.504  5.335
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.06e-01   7.84e-01   -0.64     0.52
## new_var      2.72e-03   5.93e-05   45.96   <2e-16 ***
## typeoak      4.53e-01   6.06e-01    0.75     0.46
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.29 on 56 degrees of freedom
## Multiple R-squared:  0.975,  Adjusted R-squared:  0.974
## F-statistic: 1.09e+03 on 2 and 56 DF,  p-value: <2e-16
```

The new varialbe introduced is the calculated volume from the provided data. This results in a bettwe fit - from *b*) the r-squared is 0.948, here it is 0.974.

```
# diagnostics
par(mfrow=c(1,2));
plot(model, 1); plot(model, 2)
```



Diagnostics: Fitted vs Residuals seems to not produce any obvious relationship. qq-plot follows are straight line very well. The assumptions are met.