EDDA - Final Assignment 2

Ignas Krikštaponis

Forest

```
# read the data
data <- read.table(file="data/treeVolume.txt", header=TRUE)
# make the variables as factors
data$type <- as.factor(data$type)
```

a)

```
model <- lm(volume~type, data = data)
anova(model)
```

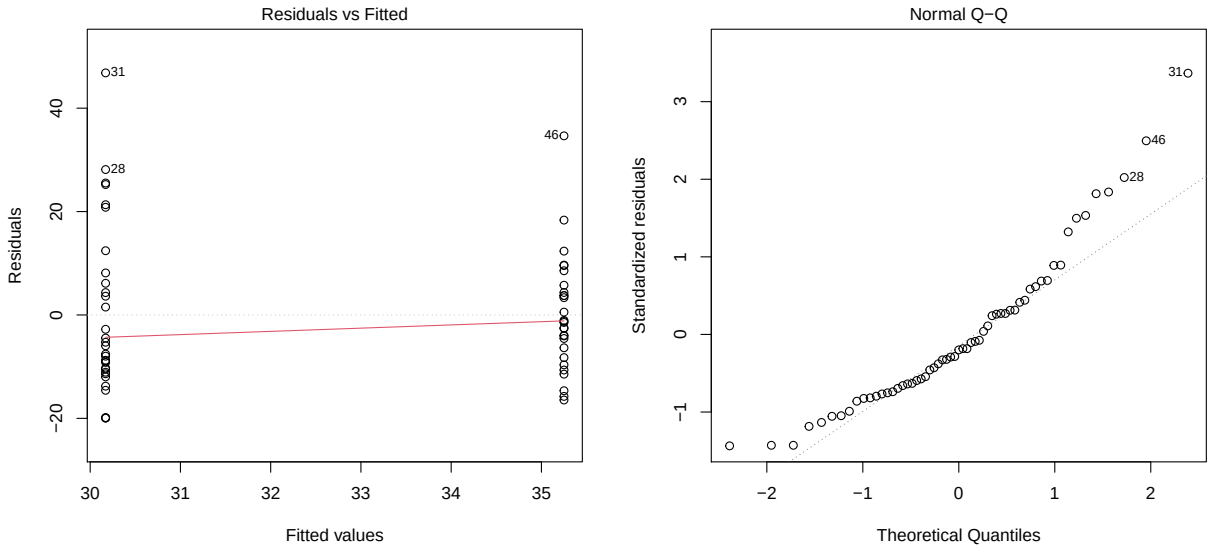
```
## Analysis of Variance Table
##
## Response: volume
##           Df Sum Sq Mean Sq F value Pr(>F)
## type       1    380     380     1.9   0.17
## Residuals 57  11395     200
```

```
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
```

From the ANOVA results (p-values > 0.05) we can conclude that there is no significant influence of tree type on the volume. For type = beech the estimate is 30.17, for type = oak it is 30.17 + 5.08 = 35.2. Let's do diagnostics.

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



From the diagnostics we can see that the qq plot does not follow a straight line well, therefore the ANOVA assumptions are invalidated and the analysis with ANOVA is not correct.

b)

```
## t-test check normality
# filter data for types

oak <- data %>% filter(type == "oak")
beech <- data  %>% filter(type == "beech")

shapiro.test(oak$volume); shapiro.test(beech$volume)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  oak$volume
## W = 0.9, p-value = 0.08

##
##  Shapiro-Wilk normality test
##
## data:  beech$volume
## W = 0.9, p-value = 0.004
```

Yes, a t-test could be related to the test in a) - we could filter the data into two samples for two tree types and then perform a two-sample non-paired t-test. However, as we can see from Shapiro test that the normality

assumption is invalidated, therefore a t-test should not be applied here. All the other three mentioned test can be applied since they do not require the data to be normal, however the Mann-Whitney test would check for median not the mean.

```
# perform Mann-Whitney
wilcox.test(oak$volume, beech$volume)
```

```
## Warning in wilcox.test.default(oak$volume, beech$volume): cannot compute exact
## p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  oak$volume and beech$volume
## W = 568, p-value = 0.04
## alternative hypothesis: true location shift is not equal to 0
```

```
# perform Kolmogorov-Smirnoff
ks.test(oak$volume, beech$volume)
```

```
## Warning in ks.test(oak$volume, beech$volume): cannot compute exact p-value with
## ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  oak$volume and beech$volume
## D = 0.4, p-value = 0.03
## alternative hypothesis: two-sided
```

From the Wilcoxon test results we can see that the p-value < 0.05, therefore the medians among the two samples are significantly different. From Kolmogorov-Smirnov we have p-value < 0.05 and we can say that the means are significantly different. This contradict the results in a). Let's perform a permutation test:

```
mystat <- function(x) sum(residuals(x)^2)
B <- 1000
tstar <- numeric(B)
for (i in 1:B) {treatstar <- sample(data$type)
tstar[i] <- mystat(lm(data$volume~treatstar)) }
myt <- mystat(lm(data$volume~data$type))
```

```
pl <- sum(tstar<myt)/B
pr <- sum(tstar>myt)/B
2*min(pl,pr)
```

```
## [1] 0.358
```

From the results of the permutation test we see that p-value > 0.05, therefore the effect of type is insignificant - this aggress with a).

c)

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

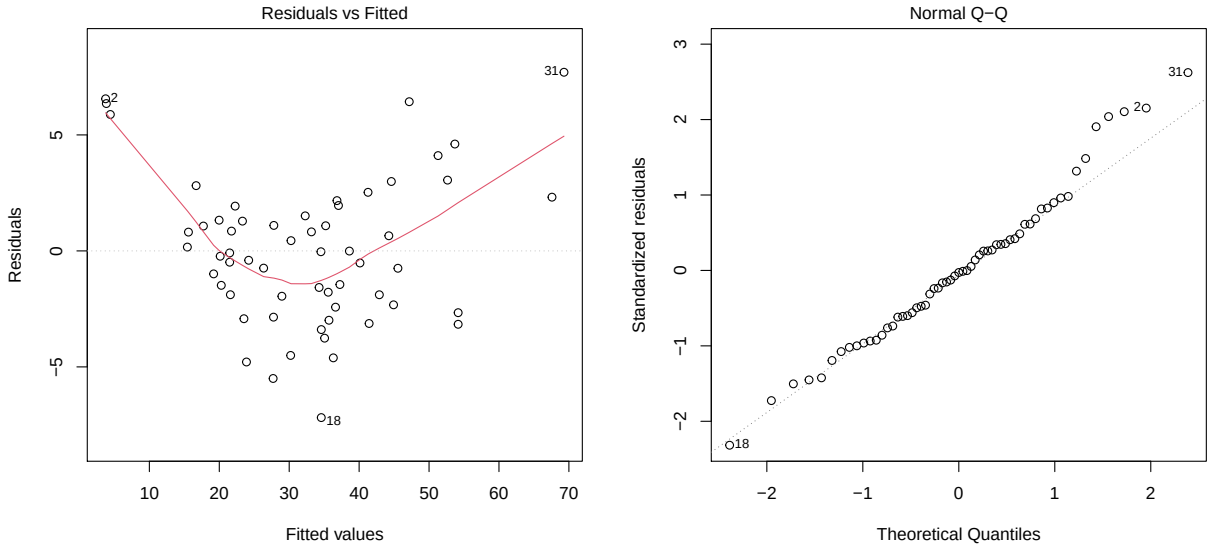
```
## Analysis of Variance Table
##
## Response: volume
##           Df Sum Sq Mean Sq F value  Pr(>F)
## diameter   1  10827   10827 1029.51 < 2e-16 ***
## height     1    346     346   32.92 4.3e-07 ***
## type       1     23      23    2.21    0.14
## Residuals 55    578      11
## ---
## Signif.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

```r
summary(model)$r.squared
```

```
## [1] 0.951
```

Here we performed ANCOVA. The outcome is the same as in a) - the effect of type is insignificant (p-value > 0.05). However, we can see that diameter and height do have a significant influence on volume. Let's do diagnostics.

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



From the diagnostics we can see that there are some outliers in the qqplot. Also there is obvious relationship in the residuals vs fitted plot, therefore the ANCOVA analysis is not valid.

```r
# perform predictions
new_data <- data.frame(type = c("oak", "beech"),
```

```
                       diameter = mean(data$diameter),
                       height = mean(data$height))

predict(model, new_data, type = "response")
```

```
##    1    2
## 31.9 33.2
```

From the results above - estimate for oak = 31.9, estimate for beech = 33.2.

d)

Start analysis with diameter:

```
# check normality
shapiro.test(data$volume)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$volume
## W = 0.9, p-value = 0.01
```

```
shapiro.test(data$diameter)
```

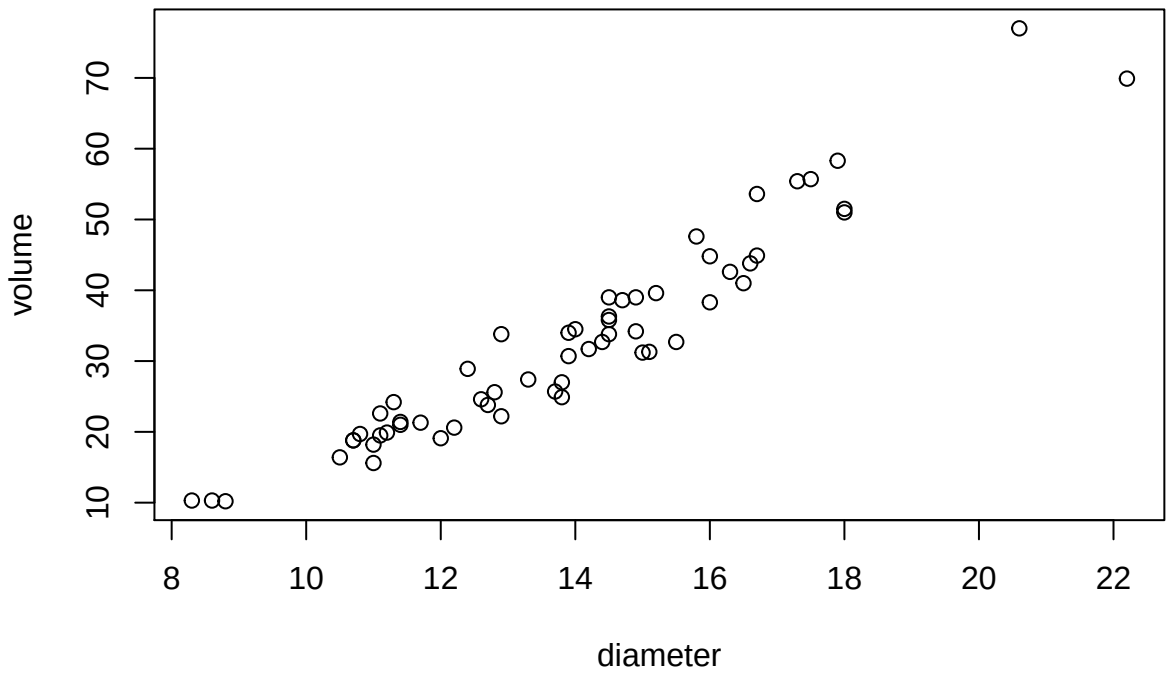
```
##
##  Shapiro-Wilk normality test
##
## data:  data$diameter
## W = 1, p-value = 0.4
```

```
# perform correlation test
cor.test(data$volume, data$diameter, method = "spearman")
```

```
## Warning in cor.test.default(data$volume, data$diameter, method = "spearman"):
## Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  data$volume and data$diameter
## S = 1328, p-value <2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##   rho
## 0.961
```

```
plot(volume~diameter, data = data)
```



We perform a Spearman corelation since the data is not normal (from shapiro test we see that one p-value < 0.05). From the results we can see that there is a significant positive correlation between the variables - meaning as diameter increases volume increases too.

Now we will perform ANCOVA with interaction to see if the influence of diameter on volume is the same under all tree types.

```
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
```

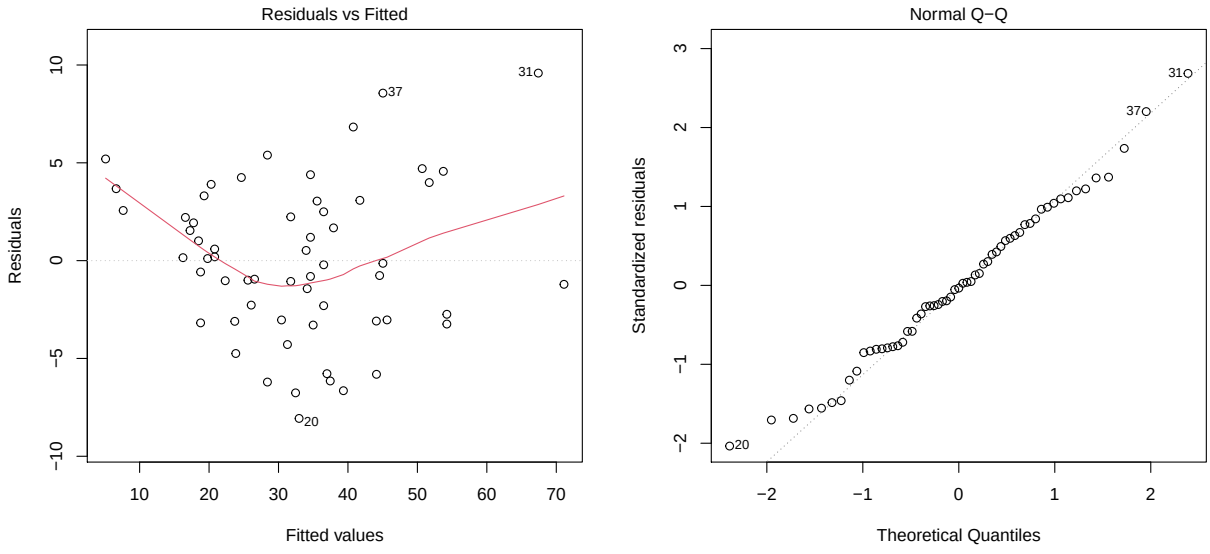
```
summary(model)
```

```
##
## Call:
## lm(formula = volume ~ type * diameter, data = data)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.065 -3.031 -0.136  2.805  9.587
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -36.943      3.189  -11.58   <2e-16 ***
## typeoak             2.809      6.125    0.46     0.65
## diameter            5.066      0.234   21.61   <2e-16 ***
## typeoak:diameter   -0.325      0.424   -0.77     0.45
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.03 on 55 degrees of freedom
## Multiple R-squared:  0.924,  Adjusted R-squared:  0.92
## F-statistic:  223 on 3 and 55 DF,  p-value: <2e-16
```

From the results above we can see that the interaction influence is not significant, therefore we can assume that the influence of diameter is similar under both tree types. Let's do diagnostics.

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



The qqplot follows a straight line pretty well. The residuals vs fitted does not seem to have any obvious relationship, there are some outliers that could be removed to improve the performace. The model assumptions are met.

Start analysis with height:

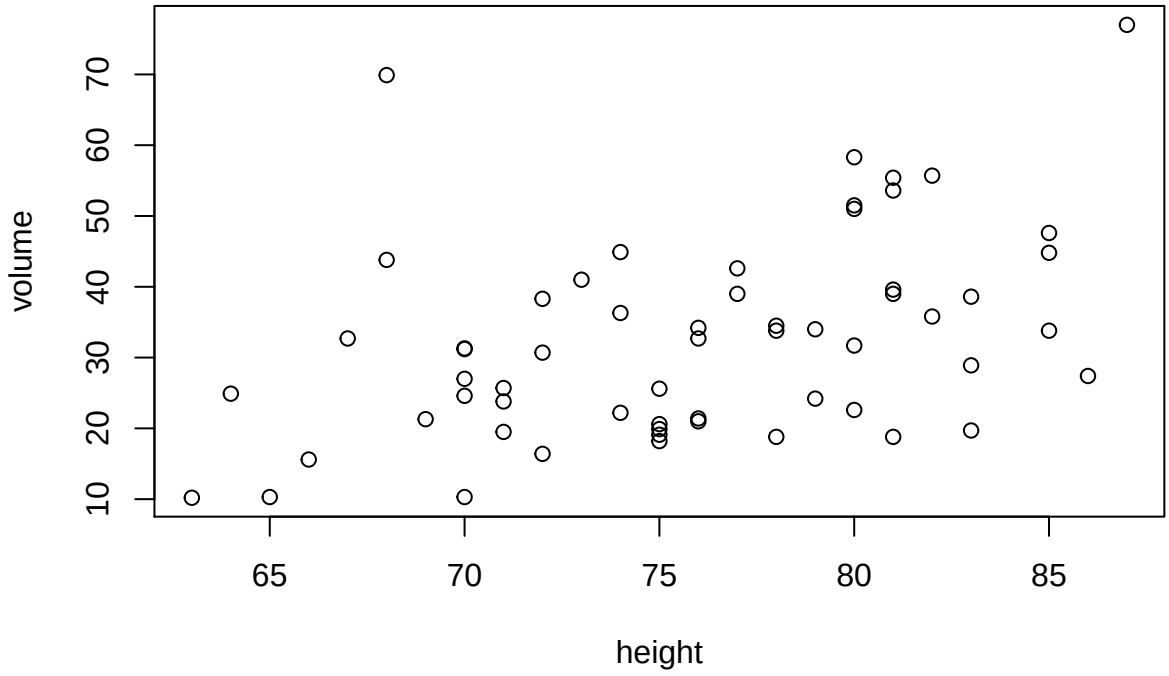
```
# perform correlation test
cor.test(data$volume, data$height, method = "spearman")
```

```
## Warning in cor.test.default(data$volume, data$height, method = "spearman"):
```

```
## Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  data$volume and data$height
## S = 19555, p-value = 7e-04
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##   rho
## 0.429
```

```
plot(volume~height, data = data)
```



From analysis before we know that Volume can not be assumed to be normally distributed, therefore we also perform a spearman correlation test. The correlation is positive (and significan) here also. Let's performe ANCOVA with interaction.

```
model <- lm(volume~type*height, data = data)
anova(model)
```

```
## Analysis of Variance Table
##
## Response: volume
##             Df Sum Sq Mean Sq F value  Pr(>F)
```

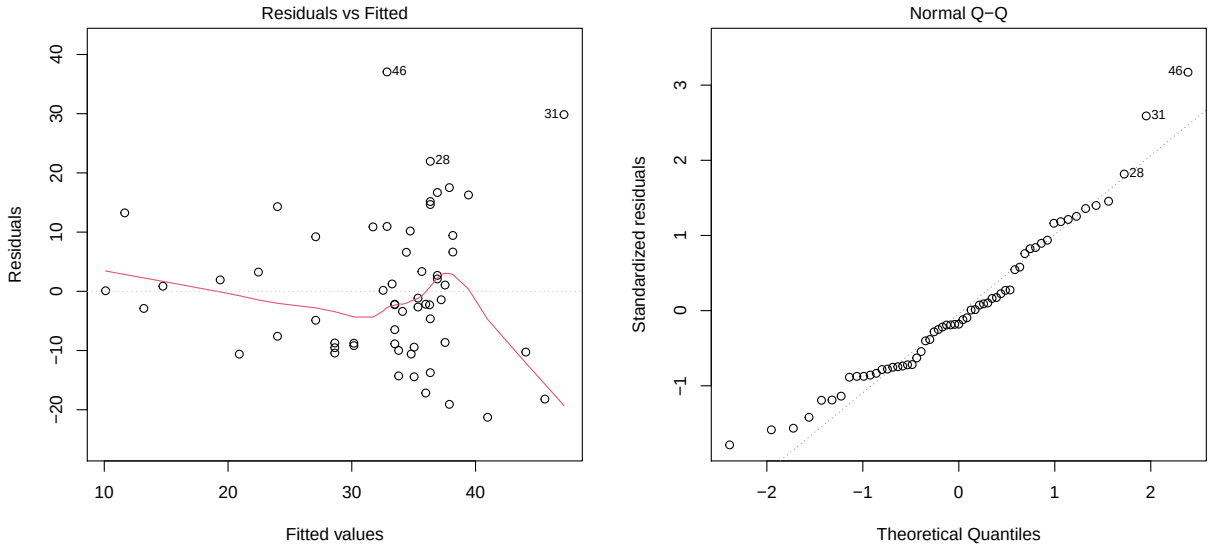
```
## type         1    380     380    2.48 0.12097
## height       1   2239    2239   14.64 0.00033 ***
## type:height  1    742     742    4.85 0.03183 *
## Residuals   55   8413     153
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(model)
```

```
##
## Call:
## lm(formula = volume ~ type * height, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -21.27  -9.02  -2.17   7.92  37.05
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -87.124     27.025   -3.22   0.0021 **
## typeoak          98.699     42.475    2.32   0.0239 *
## height            1.543      0.354    4.35  5.8e-05 ***
## typeoak:height   -1.231      0.559   -2.20   0.0318 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.4 on 55 degrees of freedom
## Multiple R-squared:  0.285,  Adjusted R-squared:  0.246
## F-statistic: 7.32 on 3 and 55 DF,  p-value: 0.000323
```

From the results above we can see that the interaction influence is now significant, therefore we can assume that the influence of diameter is different under two different tree types. Let's do diagnostics.

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



The qqplot follows a straight line pretty well. From residuals vs fitted, however, we can see that there is wider spread of values for higher values of height. The assumptions are not valid for this analysis.

e)

Let's create a new variable which is just mathematically calculated volume.

```
# let's create a new variable - calculated volume

data <- data %>%
  mutate(c_volume = height * (diameter/2)**2*pi)
```

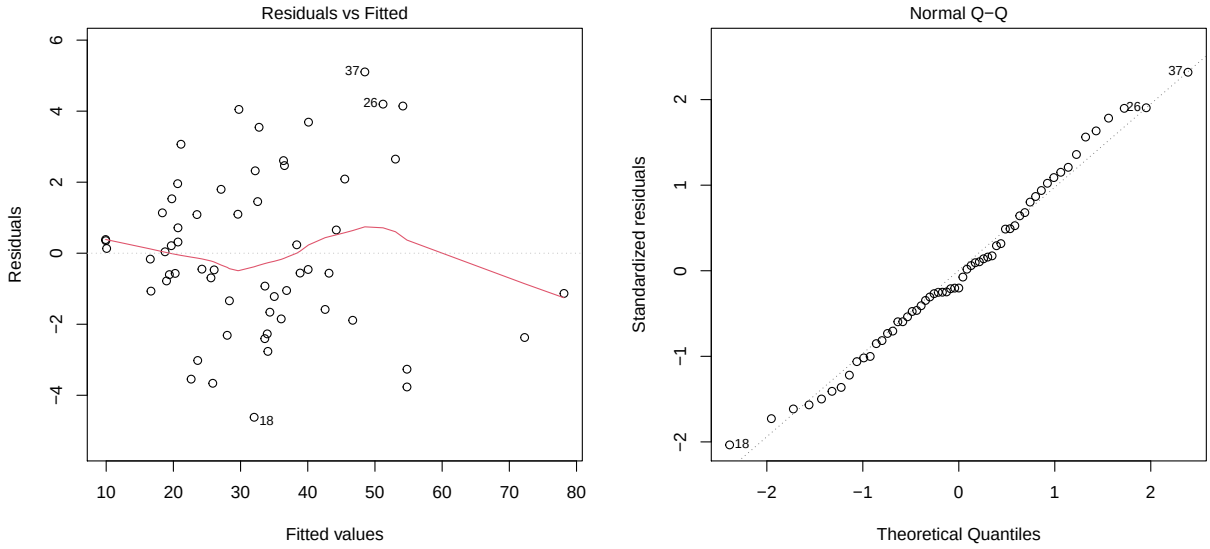
Let's fist perform ANCOVA with interaction.

```
model <- lm(volume~type*c_volume, data = data)
anova(model)
```

```
## Analysis of Variance Table
##
## Response: volume
##               Df Sum Sq Mean Sq F value  Pr(>F)
## type           1    380     380   71.25 1.7e-11 ***
## c_volume       1  11101   11101 2083.89 < 2e-16 ***
## type:c_volume  1      1       1    0.25    0.62
## Residuals     55    293       5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the results we can see that the interaction effect is insignificant. Let's do diagnostics.

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



From diagnostics we don't see any relationships in the residuals vs fitted plot. qqplot also follows the line very well, therefore the assumptions are met.

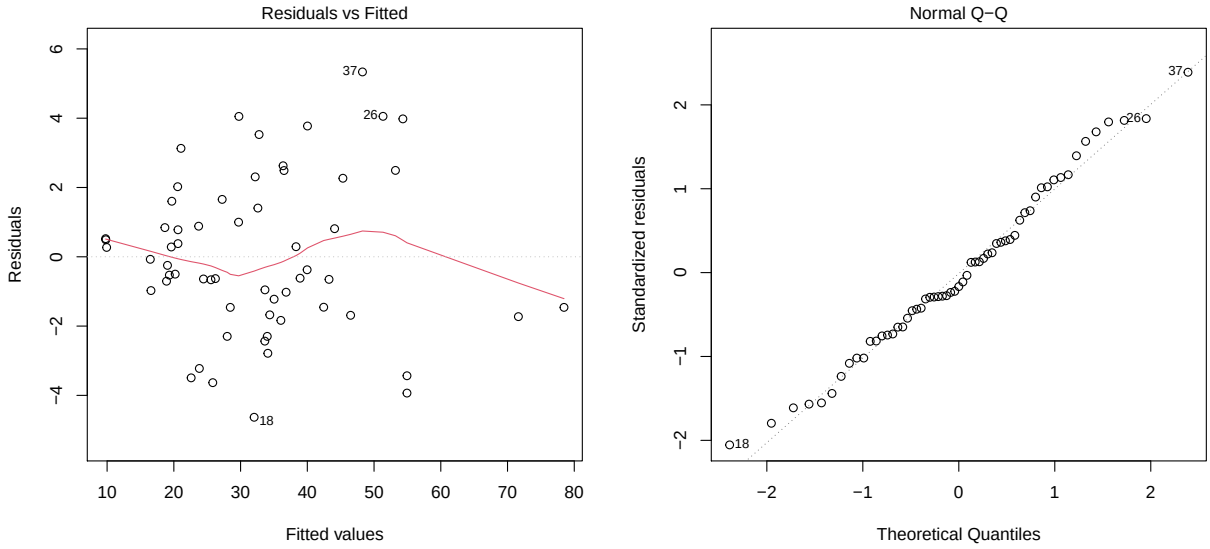
As there is no interaction, let's move on to the additive model.

```
model <- lm(volume~c_volume+type, data = data)
anova(model)
```

```
## Analysis of Variance Table
##
## Response: volume
##           Df Sum Sq Mean Sq F value Pr(>F)
## c_volume   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
```

As previously, the effect of type is insignificant. Let's do diagnostics:

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



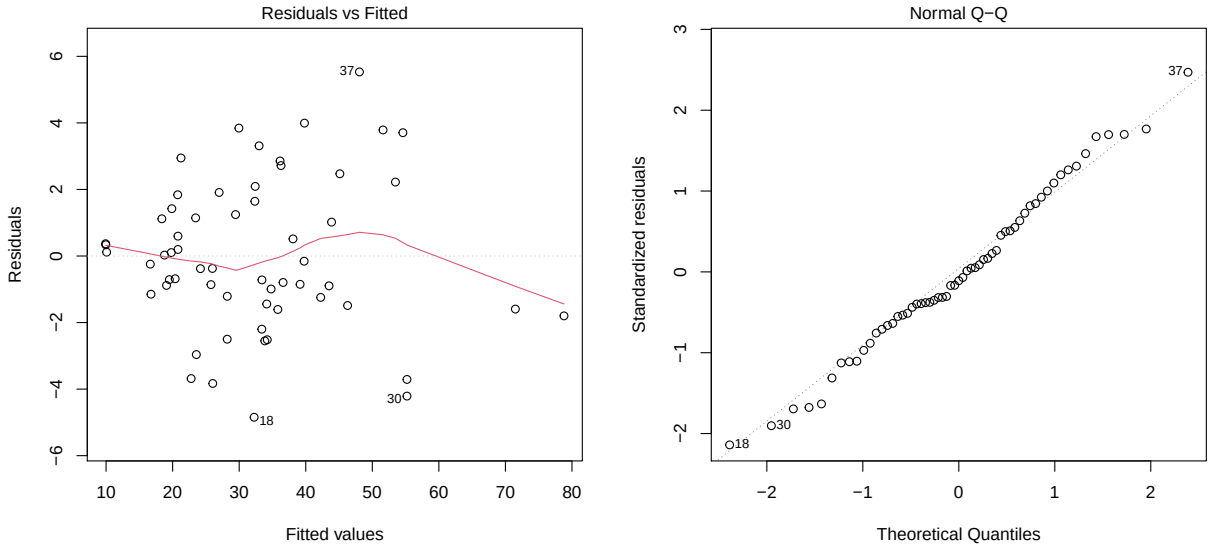
From diagnostics we don't see any relationships in the residuals vs fitted plot. qqplot also follows the line very well, therefore the assumptions are met. Let's now remove type from the model - this will create a linear model.

```
model <- lm(volume~c_volume, data = data)
summary(model)
```

```
##
## Call:
## lm(formula = volume ~ c_volume, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.846 -1.343 -0.245  1.533  5.532
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.79e-01   7.63e-01    -0.5     0.62
## c_volume     2.73e-03   5.82e-05    46.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.28 on 57 degrees of freedom
## Multiple R-squared:  0.975,  Adjusted R-squared:  0.974
## F-statistic: 2.2e+03 on 1 and 57 DF,  p-value: <2e-16
```

From the results we can see that the influence of c_volume is significant. Let's do diagnostics.

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



From diagnostics we don't see any relationships in the residuals vs fitted plot. qqplot also follows the line very well, therefore the assumptions are met.

From the r-squared value here (0.974) in comparison to what we had in d) with height (0.285) or d) with diameter (0.924) or c) (0.951), we can conclude that this model is more accurate (and therefore better).