Final exam EDDA 2

Roel

5/12/2021

Trees

The 'Amsterdamsche Bos' Forestry wishes to estimate the total wood volume of the trees on itsdomain. To this end the Fore

```
trees <- read.table("treeVolume.txt", header=TRUE)
attach(trees)
```

A)

(2.0) Investigate whether the tree type influences volume by performing ANOVA, withouttakingdiameterandheightinto accour

```
trees$type <- as.factor(trees$type)
is.numeric(trees$volume); is.factor(trees$type)
```

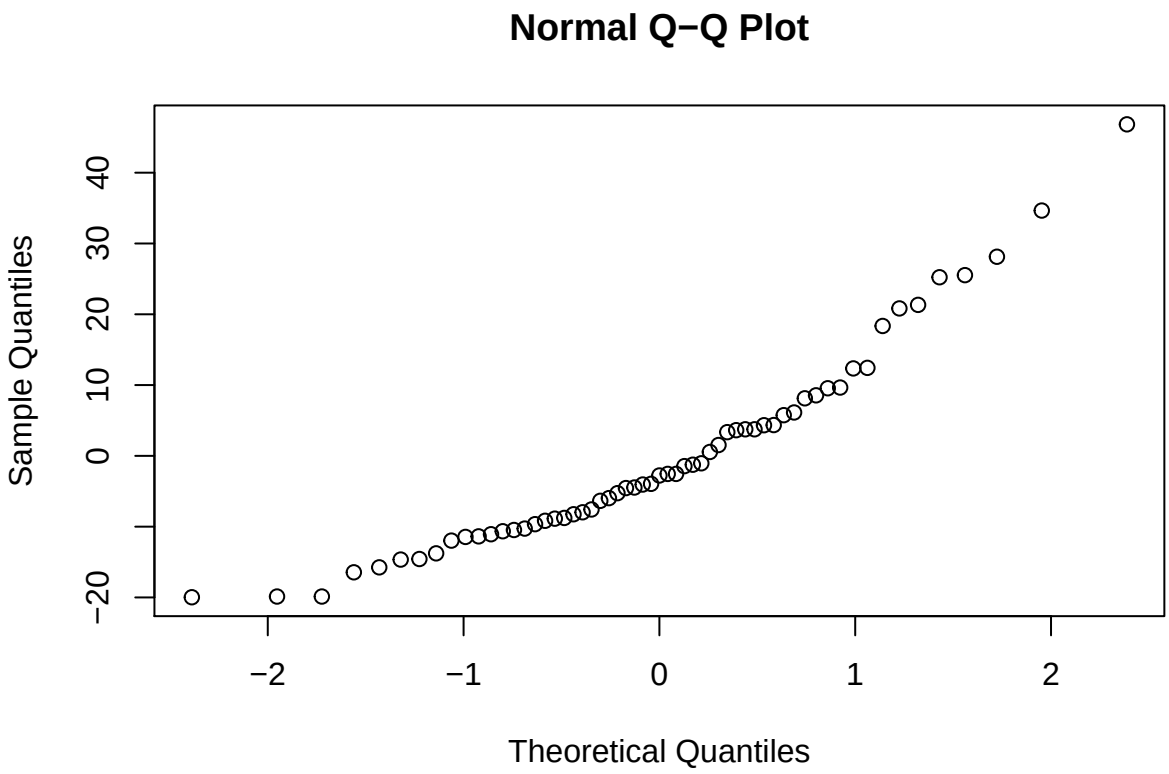
```
## [1] TRUE
```

```
## [1] TRUE
```

```
type_trees_aov = lm (volume~type, data=trees)
anova(type_trees_aov)
```

```
## Analysis of Variance Table
##
## Response: volume
##           Df  Sum Sq Mean Sq F value Pr(>F)
## type       1   379.5  379.52  1.8984 0.1736
## Residuals 57 11394.8  199.91
```

```
qqnorm(residuals(type_trees_aov))
```




```
summary(type_trees_aov)
```

```
##
## Call:
## lm(formula = volume ~ type, data = trees)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.971  -9.960  -2.771   5.940  46.829
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   30.171      2.539  11.881   <2e-16 ***
## typeoak        5.079      3.686   1.378    0.174
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.14 on 57 degrees of freedom
## Multiple R-squared:  0.03223,    Adjusted R-squared:  0.01525
## F-statistic: 1.898 on 1 and 57 DF,  p-value: 0.1736
```

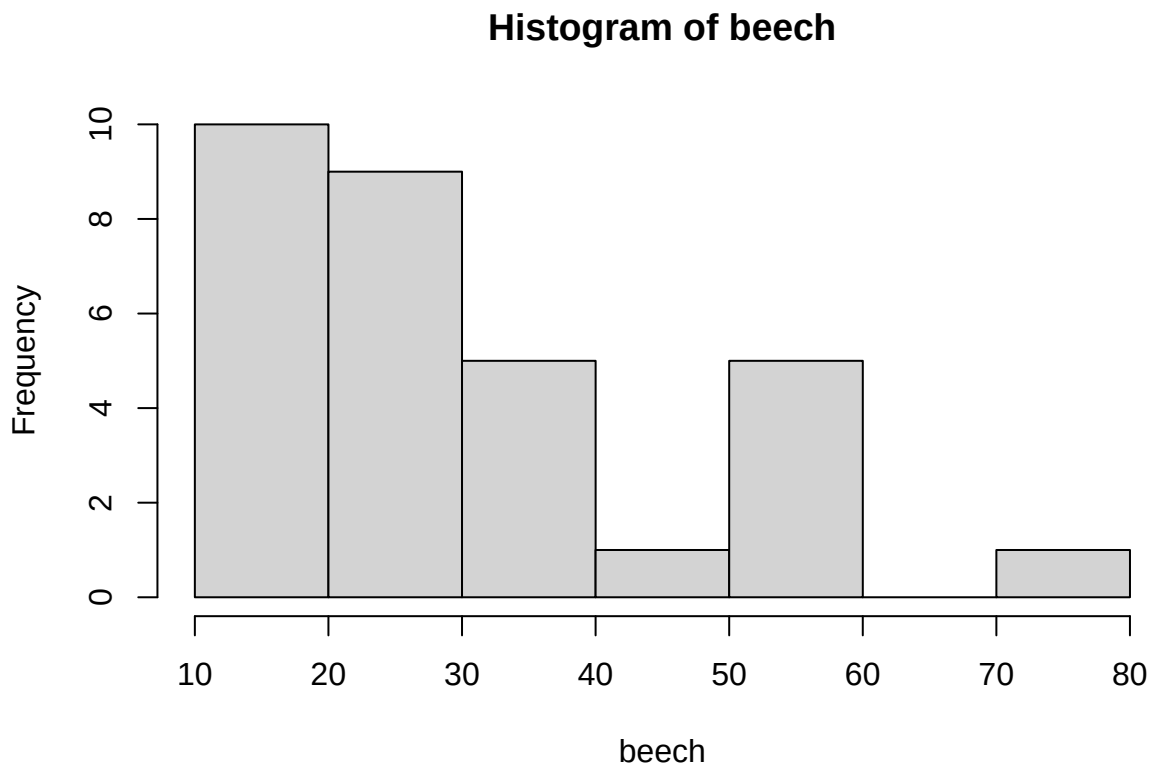
When performing the anova we see that the p-value is insignificant, as it is not smaller than alpha (0.05) but 0.1736. This means the H0 stating the the means for all factor levels are the same (no factor effect) cannot be rejected. Looking at the qqplot of the residuals you could say that the residuals are doubtfully normal. They are certainly not a straight line but also do not deviate enormously form the middle diagonal.

Looking at the summary we see that type beech is embedded in the intercept. The estimated volume for beech is thus 30.171. To arrive at the estimated volume of oak we add typeoak (5.079) to the volume of beech and arrive at: 35.25

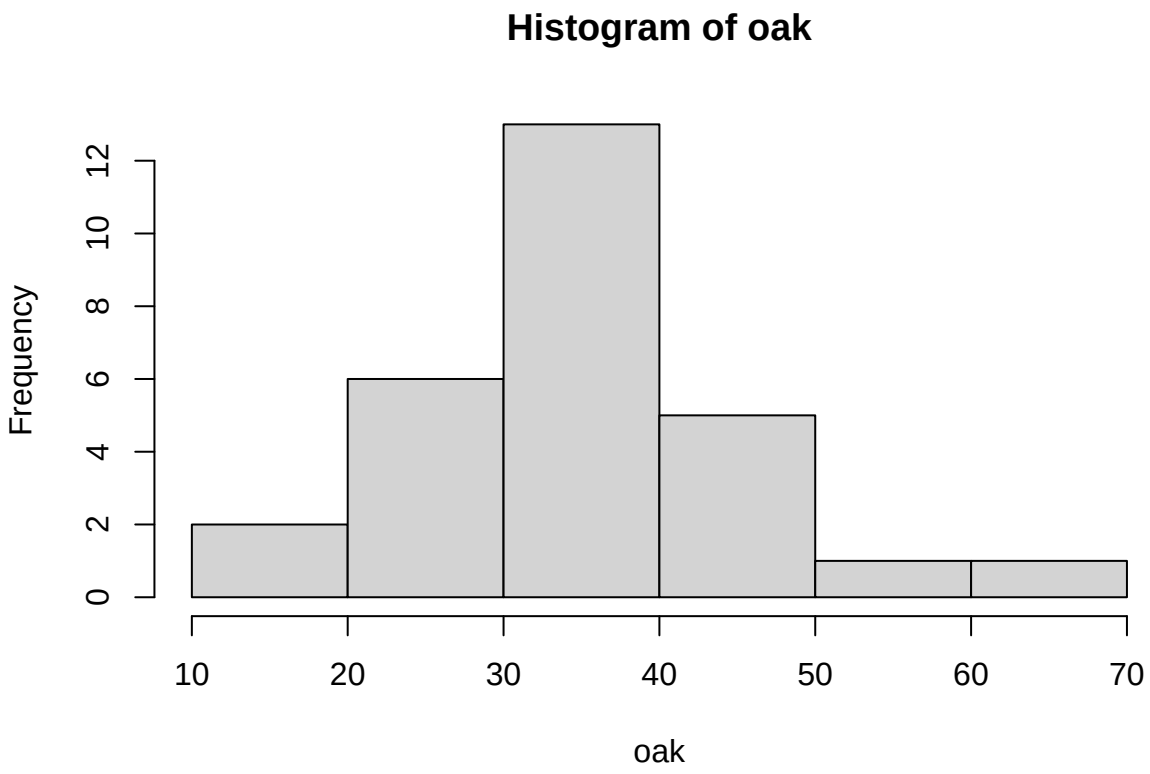
B)

(2.0) Can a t-test be related to the test in a)? Can Mann-Whitney, Kolmogorov-Smirnov andpermutation tests also be used

```
beech <- trees$volume[trees$type=="beech"] #31
oak <- trees$volume[trees$type=="oak"] # 28
hist(beech)
```




```
hist(oak)
```



As can be seen from the histograms the population of oaks looks pretty normally distributed whereas the population of beeches certainly is not. For that reason the t-test cannot be applied.

Since the data is paired, as the trees clearly grew in the same forest on the same ground, (it can be seen as the following example from the slides: "Comparing two car tire brands by putting both brands of tire on the same car and measuring the tires' wear.") we also cannot perform the Mann Whitney and Kolmogorov Smirnov test, since these two assume the data is unpaired.

The permutation test can be performed and can be done as follows (where it throws an error because they are not of the same length):

```
dataset = cbind(beech,oak)
```

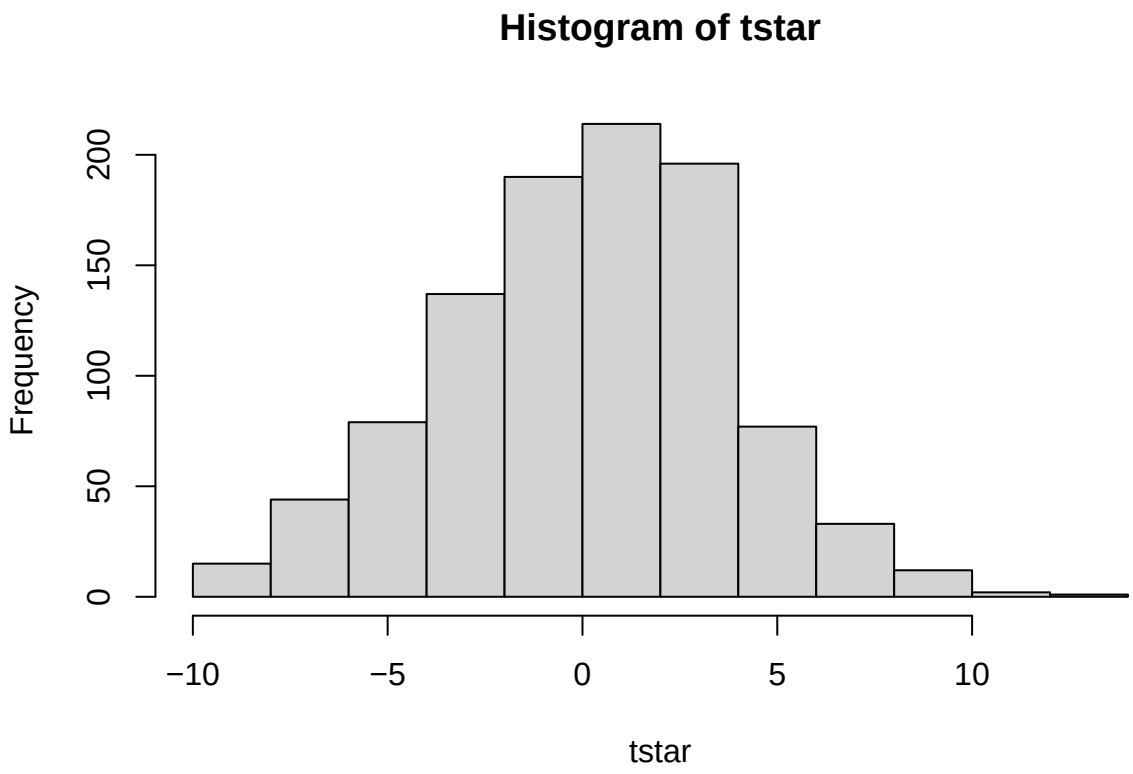
```
## Warning in cbind(beech, oak): number of rows of result is not a multiple of
## vector length (arg 2)
```

```
mystat=function(x,y) {mean(x-y)}
B=1000
tstar=numeric(B)
for (i in 1:B){
  datasetstar=t(apply(cbind(dataset[,1],dataset[,2]),1,sample))
  tstar[i]=mystat(datasetstar[,1], datasetstar[,2])
}

myt=mystat(dataset[,1],dataset[,2])
myt
```

```
## [1] -4.322581
```

```
hist(tstar)
```




```
pl=sum(tstar<myt)/B
pr=sum(tstar>myt)/B
p=2*min(pl,pr)
p # if the p-value is significant the two distributions have different means
```

```
## [1] 0.25
```

After performing the permutation test we get a p value of 0.25, which is insignificant with a as it is greater than our alpha of 0.05. Therefore we cannot reject the H0 for this permutation test that there is no difference between the distributions of X and Y within pairs. This corresponds with the conclusion from a. We see that tstar is normally distributed as expected.

C)

2.0) Investigate whether tree type influences volume, now includingdiameterandheightasexplanatory variables into the analys

When we include diameter and height as explanatory variables we arrive at a ancova.

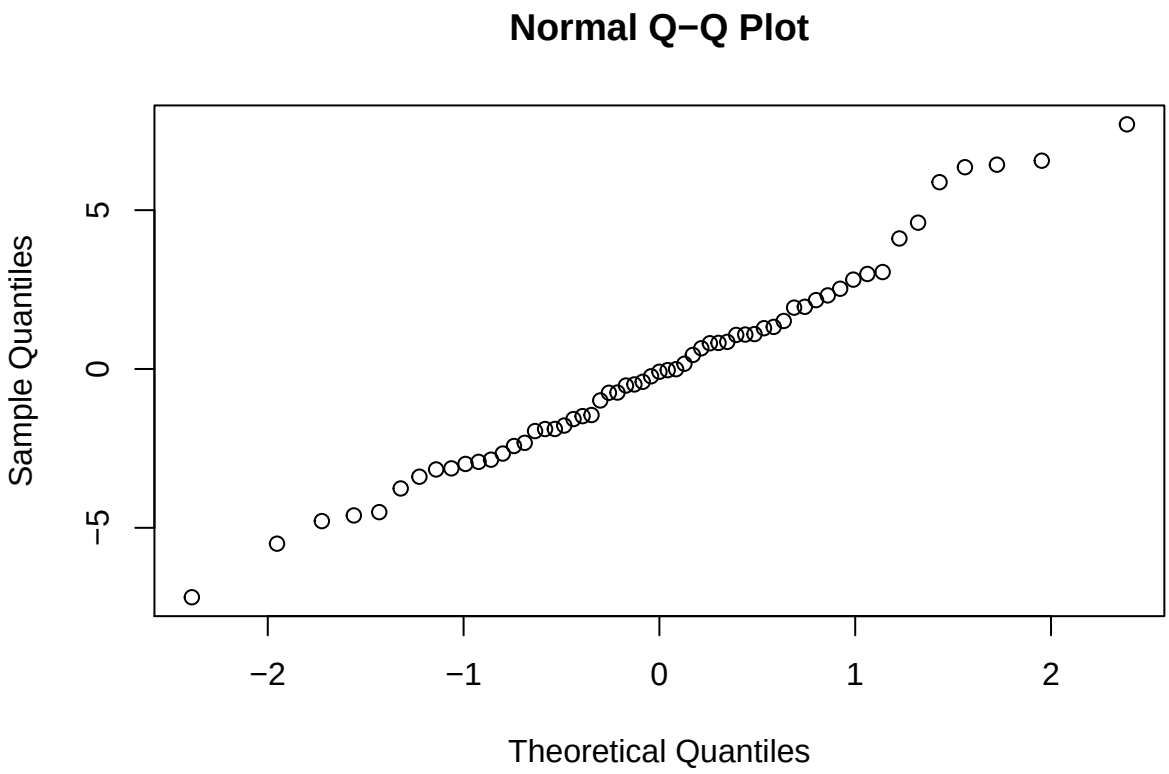
```r
model = lm(volume~type+diameter+height)
drop1(model, test = "F")
```

```
## Single term deletions
##
## Model:
## volume ~ type + diameter + height
##          Df Sum of Sq    RSS    AIC  F value    Pr(>F)
## <none>                 578.4 142.68
## type      1      23.2  601.6 143.00   2.2083     0.143
## diameter  1    8577.1 9155.5 303.63 815.6110 < 2.2e-16 ***
## height    1     324.2  902.5 166.93  30.8246 8.416e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

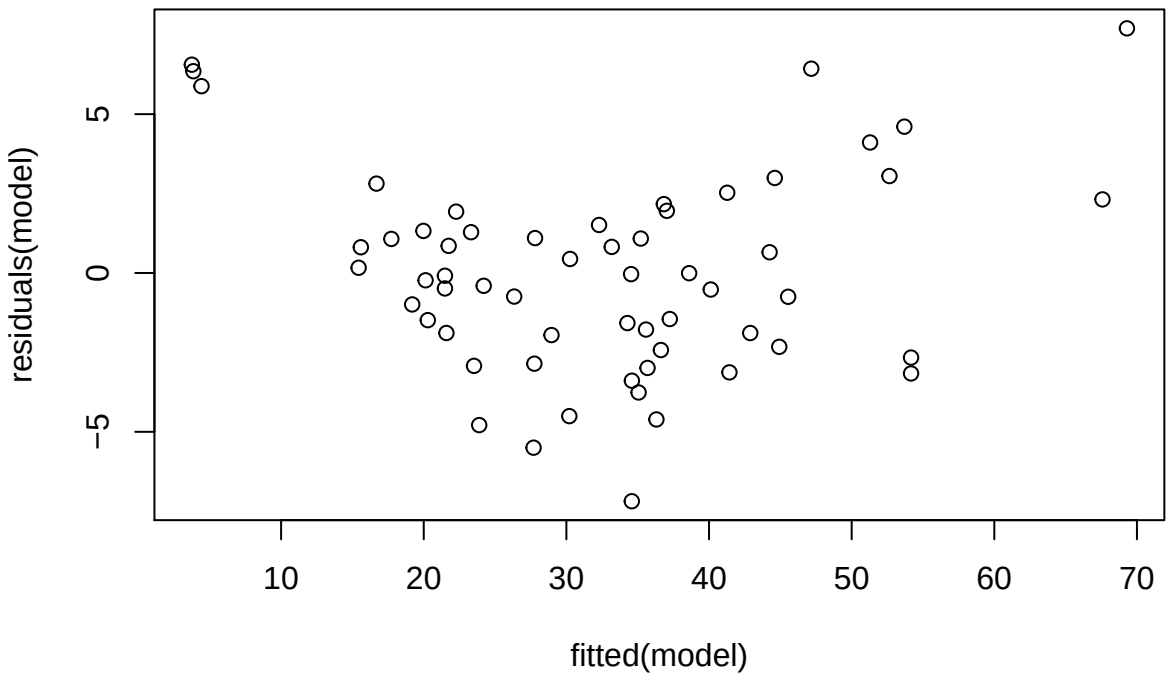
```r
summary(model)
```

```
##
## Call:
## lm(formula = volume ~ type + diameter + height)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.1859 -2.1396 -0.0871  1.7208  7.7010
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -63.78138    5.51293 -11.569 2.33e-16 ***
## typeoak      -1.30460    0.87791  -1.486    0.143
## diameter      4.69806    0.16450  28.559  < 2e-16 ***
## height        0.41725    0.07515   5.552 8.42e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.243 on 55 degrees of freedom
## Multiple R-squared:  0.9509, Adjusted R-squared:  0.9482
## F-statistic: 354.9 on 3 and 55 DF,  p-value: < 2.2e-16
```

```r
# Assumption check:
qqnorm(residuals(model)) # qqplot for residuals corrected for the different population means
```




```
plot(fitted(model),residuals(model))
```



When looking at the outcome of drop1 (similar to anova but takes into account that most important variable need not be at the end of the equation) we see that type is again not significant, since the p-value is above our alpha, while diameter and height are with p-values of respectively 0.143, 2.2e-16 and 8.416e-07. This means that for type we cannot reject the H0 the means for all factor levels are the same (no factor effect) while for the explanatory variables we can reject the H0 that explanatory variables (beta) are zero.

When checking the assumptions we see that the line in the qqplot looks fairly straight, suggesting the residuals are normally distributed. When looking at the plot of the residuals vs the fitted we see the points seem to be fairly homogenous, albeit without the two points in the upper left corner one could suggest there is some cone shape evident.

To estimate the volumes for the two tree types with average height and diameter we do the following:

```
beech_full = trees[type=="beech",]
oak_full = trees[type=="oak",]

avg_dia_beech = mean(beech_full$diameter); avg_dia_beech
```

```
## [1] 13.24839
```

```
avg_hei_beech = mean(beech_full$height); avg_hei_beech
```

```
## [1] 76
```

```r
avg_dia_oak = mean(oak_full$diameter); avg_dia_oak
```

```
## [1] 14.63571
```

```r
avg_hei_oak = mean(oak_full$height); avg_hei_oak
```

```
## [1] 75.67857
```

```r
estim_volume_beach = -63.7814 + 0.4172* avg_hei_beech + 4.6981 * avg_dia_beech; estim_volume_beach
```

```
## [1] 30.16805
```

```r
estim_volume_oak = -63.7814 + 0.4172* avg_hei_oak + 4.6981 * avg_dia_oak; estim_volume_oak
```

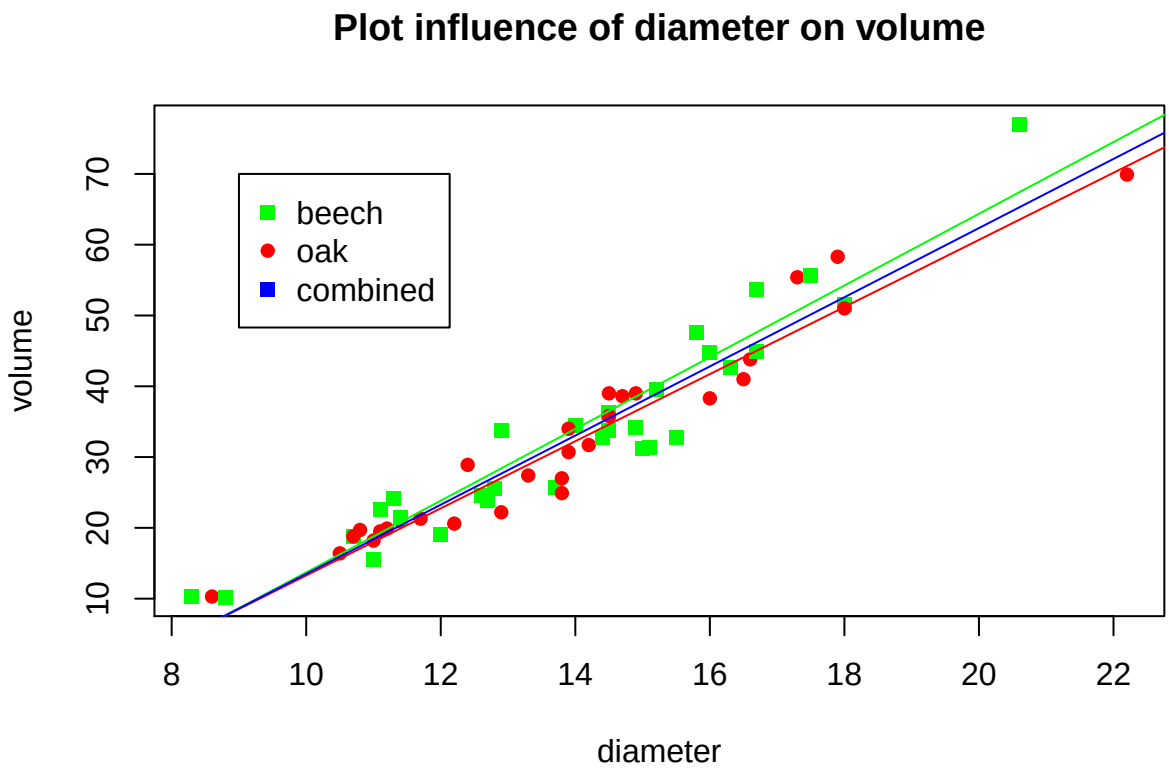
```
## [1] 36.55175
```

After which we can conclude that the estimated volume with the average diameter and height for beech is 30.168 and for oak 36.55, which is similar to the results obtained in a, but a little different since we now discriminate between tree types.

D)

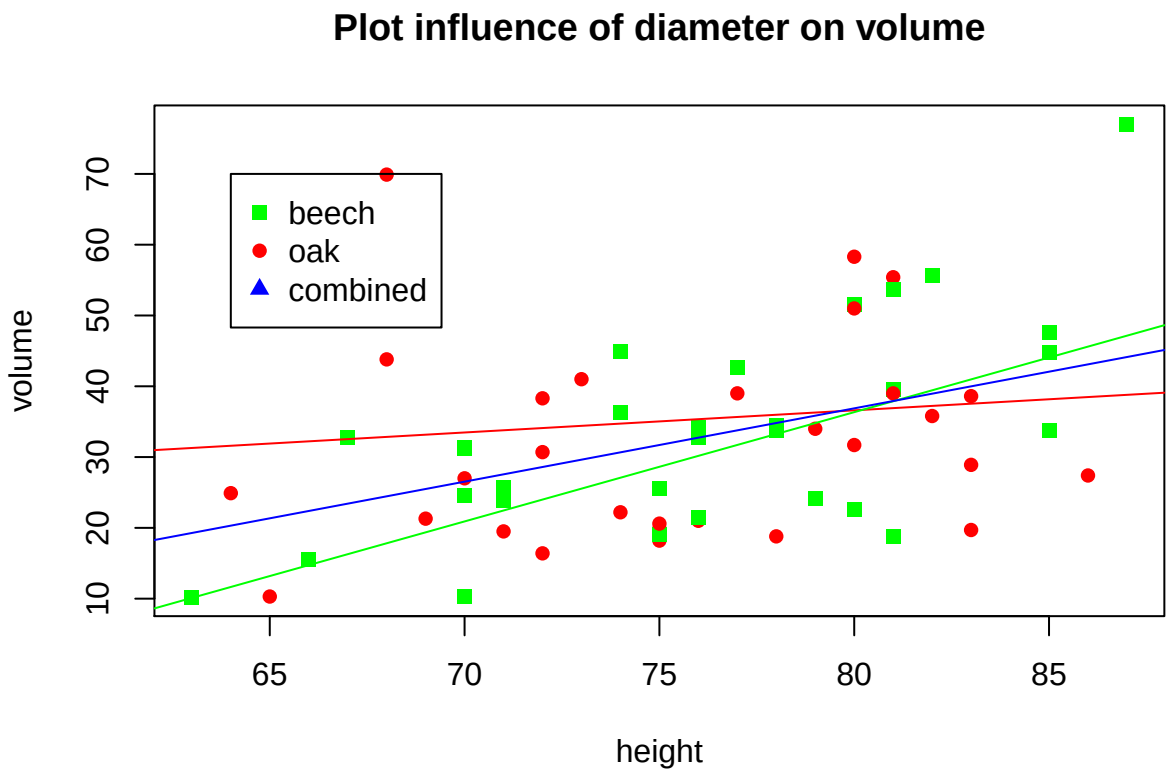
(2.0) How doesdiameterinfluencevolume? Investigate whether this influence is similar forthe both tree types. Do the same fo

```r
plot(volume~diameter, pch= c(15, 16), col=c("green", "red"))
legend(9, 70, legend=c("beech", "oak", "combined"), pch= c(15, 16), col=c("green", "red", "blue"))
title("Plot influence of diameter on volume")
abline(lm(volume~diameter,data=trees[type=='oak',]), col = "red")
abline(lm(volume~diameter,data=trees[type=="beech",]), col = "green")
abline(lm(volume~diameter,data=trees), col = "blue")
```



When looking at how diameter influences volume it is evident that as the diameter grows, the volume grows as well. Given that oak is red and beech is green it is also evident that this relation is the same for both type of trees

```
plot(volume~height, pch= c(15, 16), col=c("green", "red"))
legend(64, 70, legend=c("beech", "oak", "combined"), pch= c(15, 16, 17), col=c("green", "red", "blue"))
title("Plot influence of diameter on volume")
abline(lm(volume~height,data=trees[type=='oak',]), col = "red")
abline(lm(volume~height,data=trees[type=="beech",]), col = "green")
abline(lm(volume~height,data=trees), col = "blue")
```



When looking at the influence of height on volume we again see an increase in volume when height increased. This time the slope of the combined line is way smaller, and when we look at the different types of trees we see that especially for oak the slope is is really small. This suggests the volume of an oak does not quickly get more as it gets higher, whereas for beeches that is more the case. An example for this could be that an oak mostly sprouts branches with leafs after a certain height, while the massive trunk only gradually grows.

E)

(2.0) Propose a transformation of the explanatory variables that possibly yields a better model(verify this). (Hint: think of a

The volume of a round tube is calculated as it's height times the area of the tube (hA). To better capture this mathematical relation it is probably better to replace the diameter with the the area of the tube. I believe the calculation is ' 2*pi * r^2' , where r is half of the diameter, but despite the exact implementation the transformation I propose is to rewrite the diameter (as present in the model) to the area of the circle.