Final Assignment

Andrea Di Dio, Chenghan Song, Jiacheng Lu — Group 22

Exercise *Trees*

a)

H_0 : The tree type does not affect volume.(without taking the diameter or height into account.)

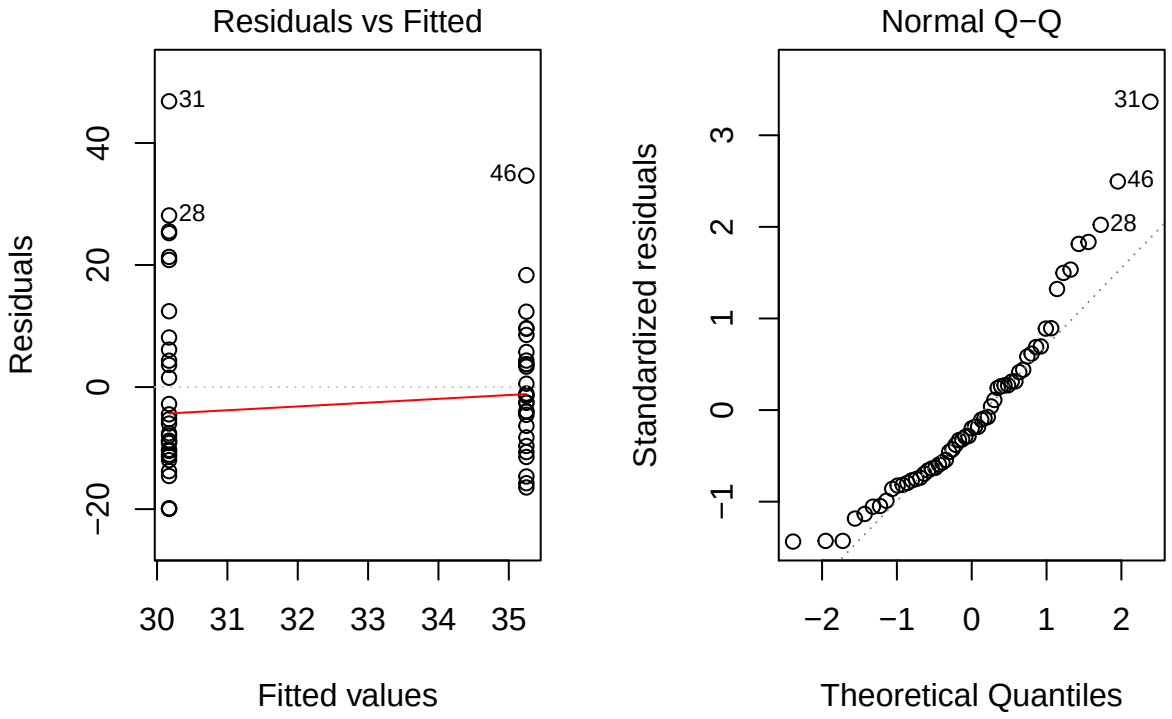
```
treeVolume=read.table(file="treeVolume.txt",header=TRUE)
treeVolumeaov=lm(volume~type, data = treeVolume)
anova(treeVolumeaov)
```

```
## 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
```

```
summary(treeVolumeaov)
```

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

```
par(mfrow=c(1,2))
plot(treeVolumeaov,1)
plot(treeVolumeaov,2)
```



From the test result, p-value is 0.1736, bigger than 0.05 (α), so we fail to reject H_0, concluding that, without taking the diameter ot height into account, tree type does not influence volume. For beech, the estimated volume is 30.171, for oak, the estimated volume is $30.171 + 5.079 = 35.25$. In the normal qq plot we can see that it shows a certain curve and not the straight qq line, so the model does not meet the normality assumption of anova.

b)

H_0 : The tree type does not affect volume.(including diameter and height as explanatory variables)

```
treeVolumeaov2=lm(volume~diameter+height+type,data=treeVolume)
drop1(treeVolumeaov2,test="F")
```

```
## Single term deletions
##
## Model:
## volume ~ diameter + height + type
##          Df Sum of Sq    RSS    AIC  F value    Pr(>F)
## <none>                  578.4 142.68
## 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 ***
## type      1      23.2  601.6 143.00   2.2083     0.143
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(treeVolumeaov2)
```

```
##
## Call:
## lm(formula = volume ~ diameter + height + type, data = treeVolume)
##
## 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 ***
## diameter      4.69806    0.16450  28.559  < 2e-16 ***
## height        0.41725    0.07515   5.552 8.42e-07 ***
## typeoak      -1.30460    0.87791  -1.486    0.143
## ---
## 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
```

```
mean(treeVolume$diameter)
```

```
## [1] 13.90678
```

```
mean(treeVolume$height)
```

```
## [1] 75.84746
```

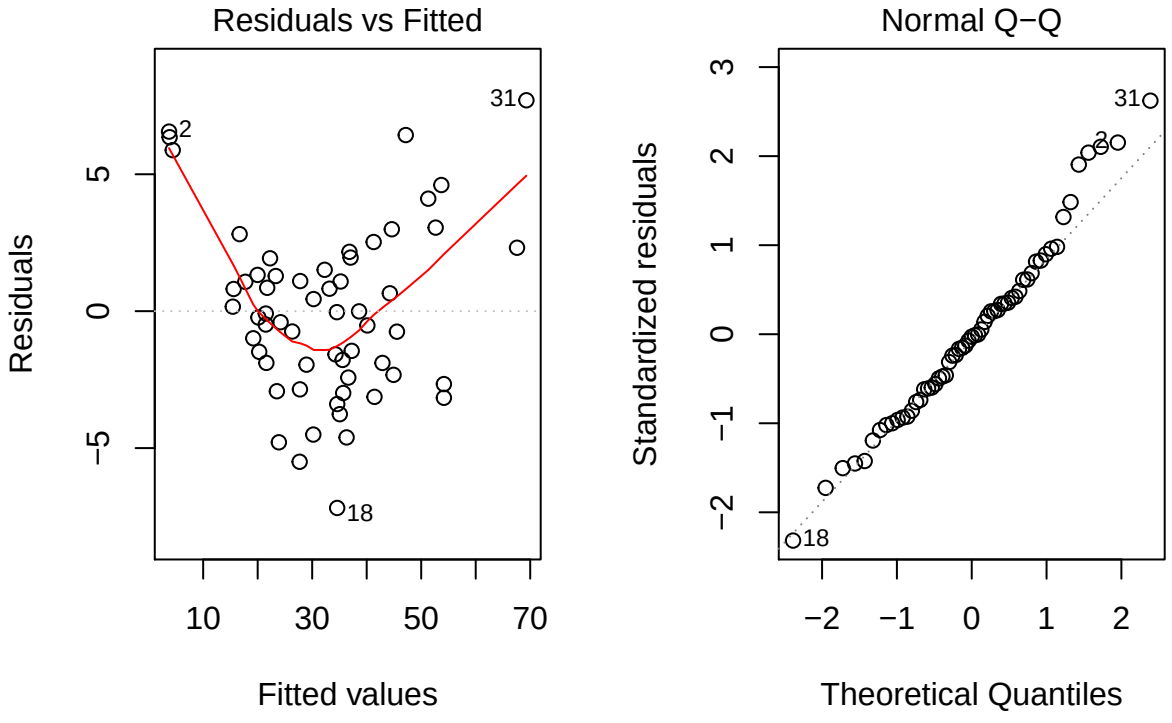
```
-63.78138-1.30460+4.69806*13.90678+75.84746*0.41725
```

```
## [1] 31.89626
```

```
-63.78138+4.69806*13.90678+75.84746*0.41725
```

```
## [1] 33.20086
```

```
par(mfrow=c(1,2))
plot(treeVolumeaov2,1)
plot(treeVolumeaov2,2)
```

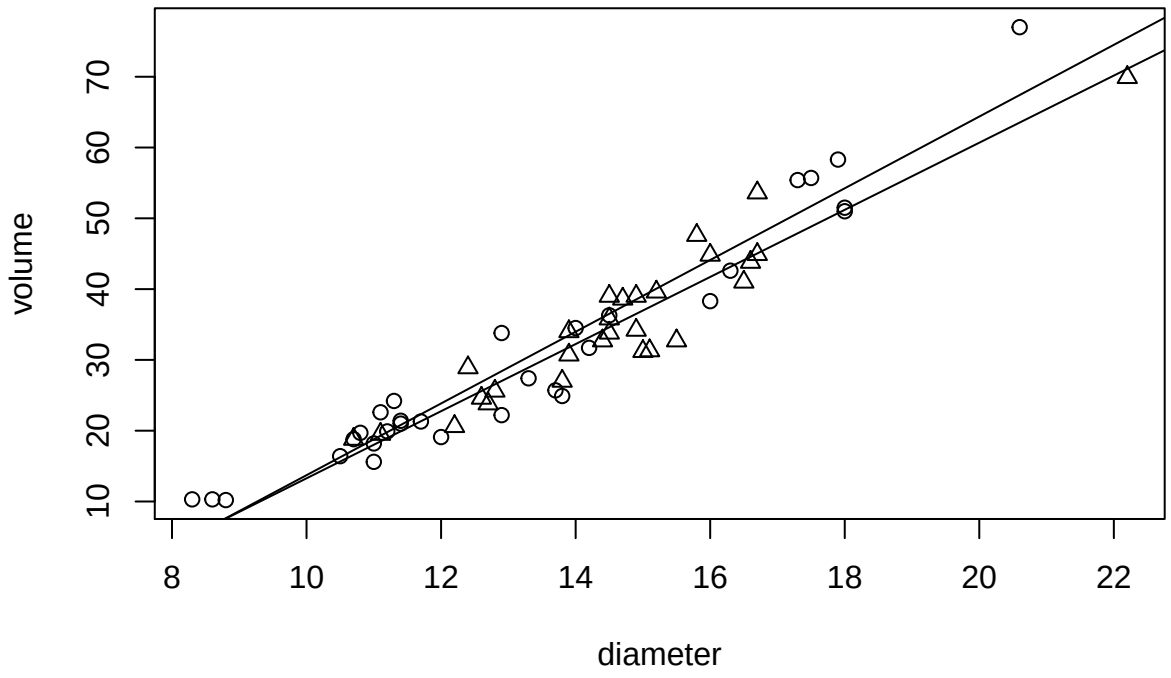


From the test result, we can see the p-value is 0.143, so we fail to reject H_0. $\mu = -63.78138, \beta_{diameter} = 4.69806, \beta_{height} = 0.41725, \alpha_{beech} = 0, \alpha_{oak} = -1.30460$ for beech type, the estimated value is 33.20086; for oak type, the estimated volume is 31.89626. In the plot of residual against fitted, we can see the spread shows certain pattern (i.e., there is a systematic change in the residuals based on the fitted values), so the model is not good, suggesting that the assumption of linear relationship is not reasonable.

c)

H_0 :The dependence between diameter and volume is the same for different tree types. (There is no dependence between diameter and volume for the tree types).

```
plot(volume~diameter,pch=unclass(type),data=treeVolume)
for (i in c("oak",'beech')) abline(lm(volume~diameter,data=treeVolume[treeVolume$type==i,]))
```




```
aovtreeVolumeinter=lm(volume~height+diameter*type,data=treeVolume)
summary(aovtreeVolumeinter)
```

```
##
## Call:
## lm(formula = volume ~ height + diameter * type, data = treeVolume)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3500 -2.1940 -0.1413  1.7012  8.1765
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -63.87254    5.53859 -11.532 3.47e-16 ***
## height             0.43412    0.07903   5.493 1.09e-06 ***
## diameter           4.60813    0.20701  22.261  < 2e-16 ***
## typeoak           -4.96300    5.14936  -0.964    0.339
## diameter:typeoak   0.25886    0.35897   0.721    0.474
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.257 on 54 degrees of freedom
## Multiple R-squared:  0.9513, Adjusted R-squared:  0.9477
## F-statistic:   264 on 4 and 54 DF,  p-value: < 2.2e-16
```

From the plot we can see that with the growth of diameter, volume grows bigger. The p-value for H_0 is 0.474, so we fail to reject H_0 concluding that there might not be a dependence between the diameter and the volume for the two tree types. The true lines in the plot are similar too.

d)

For volume, $v = \frac{(\pi * d^2 * h)}{4}$, so we transfrom diameter to $diameter^2$.

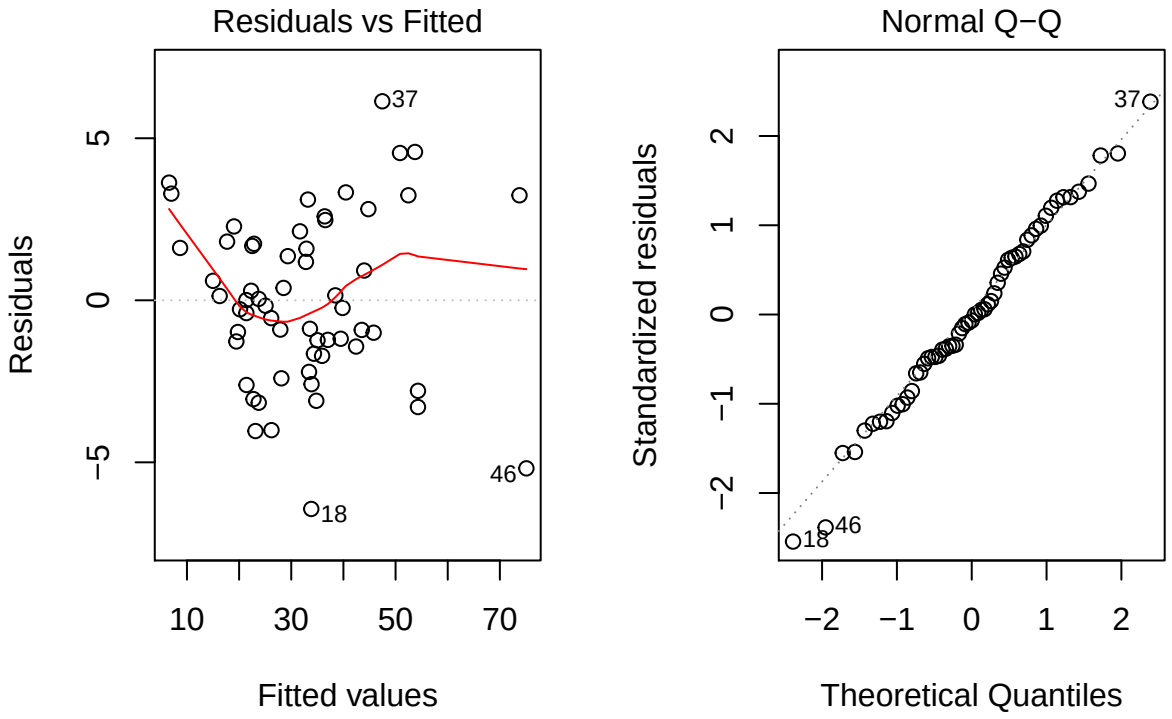
```
treeVolume$square_diameter=treeVolume$diameter^2
treeVolumeaov3=lm(volume~square_diameter+height+type,data=treeVolume)
drop1(treeVolumeaov3,test="F")
```

```
## Single term deletions
##
## Model:
## volume ~ square_diameter + height + type
##                 Df Sum of Sq    RSS    AIC   F value    Pr(>F)
## <none>                        385.7 118.78
## square_diameter  1    8769.7 9155.5 303.63 1250.3906 < 2.2e-16 ***
## height           1     469.8  855.5 163.78   66.9841 4.371e-11 ***
## type             1       0.3  386.0 116.83    0.0419    0.8385
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(treeVolumeaov3)
```

```
##
## Call:
## lm(formula = volume ~ square_diameter + height + type, data = treeVolume)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4389 -1.5381 -0.1682  1.7764  6.1405
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -37.065117   4.497367  -8.242 3.53e-11 ***
## square_diameter   0.159310   0.004505  35.361  < 2e-16 ***
## height            0.496786   0.060699   8.184 4.37e-11 ***
## typeoak          -0.145058   0.708391  -0.205    0.839
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.648 on 55 degrees of freedom
## Multiple R-squared:  0.9672, Adjusted R-squared:  0.9655
## F-statistic: 541.3 on 3 and 55 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(1,2))
plot(treeVolumeaov3,1)
plot(treeVolumeaov3,2)
```



We found out that the plot of residuals versus fitted, the spread does not show a certain pattern, the normal qq-plot shows a straight line as well, this suggests that using $diameter^2$ is better for our model as the assumption of a linear relationship is shown to be more likely by the spread of the residuals versus fitted values and the normality can also be assumed looking at the normal QQ plot given that the points lie around the normal QQ line.

Exercise *Jane Austen*

a)

The interest of contingency table test is in finding a possible dependency between the two factors. Thus, it is suitable and appropriate for studying possible dependency between Austen's novels and an Austen's admirer's novel (two factors) by counting the number of different words (categories). The table has some observed counts which can be used in a χ^2 analysis to test the independence by comparing those to the expected counts.

b)

H_0: There is no difference between the word counts/literary styles in Austen's three books.

```
book =read.table(file="austen.txt",header=TRUE);
ausmatrix = as.matrix(book[,1:3]);
auscq=chisq.test(ausmatrix);
auscq
```

```
##
##  Pearson's Chi-squared test
##
## data:  ausmatrix
## X-squared = 12.271, df = 10, p-value = 0.2673
```

```
(auscq$observed-auscq$expected)/sqrt(auscq$expected);
```

```
##               Sense       Emma      Sand1
## a       -1.02997736 -0.1290203  1.5937736
## an       0.44728806 -0.1590968 -0.3746273
## this     0.05133600  0.2938669 -0.5036577
## that     0.74817619  0.2865778 -1.4423521
## with    -0.04747379  0.5205063 -0.7035205
## without  1.06544255 -1.5884103  0.8926239
```

The p-value is $0.2673 > \alpha$ meaning that we fail to reject H_0 concluding that there seems to be no difference between the words counts/literary styles with different books. Austen herself was consistent in her different novels.

c)

H_0: There is no difference between the word counts/literary styles in Austen's three books.

```
book =read.table(file="austen.txt",header=TRUE);
bookmatrix = as.matrix(book);
z=chisq.test(bookmatrix);
print(z)
```

```
##
##  Pearson's Chi-squared test
##
## data:  bookmatrix
## X-squared = 45.578, df = 15, p-value = 6.205e-05
```

```
(z$observed-z$expected)/sqrt(z$expected);
```

```
##               Sense          Emma      Sand1       Sand2
## a       -1.0149156 -0.1120927868  1.6062866 -0.05889921
## an      -0.5906319 -1.2199545912 -1.0671306  3.72816398
## this     0.1388299  0.3904903154 -0.4436450 -0.32671736
## that     1.5943613  1.1798488360 -0.9099606 -3.04931581
## with    -0.5120944  0.0001916718 -1.0246069  1.74821745
## without  1.3919336 -1.3411962838  1.1365432 -1.06963011
```

The p-value is $6.205e-05 < \alpha$ meaning that we can safely reject H_0 concluding that there is significant difference between the words counts/literary styles with different books. The admiror was not successful in imitating Austen's style. From the table we can see the admirer used relatively more 'an','with' and less 'that','without' in Sanditon while Austen used relatively more 'a','without' and less 'an','with' in Sanditon, which made the whole novel inconsistent.