Assignment 1

Group 69: Dmitrijs Voronovs (2779206), Nikita …, Alina …

24.02.2023

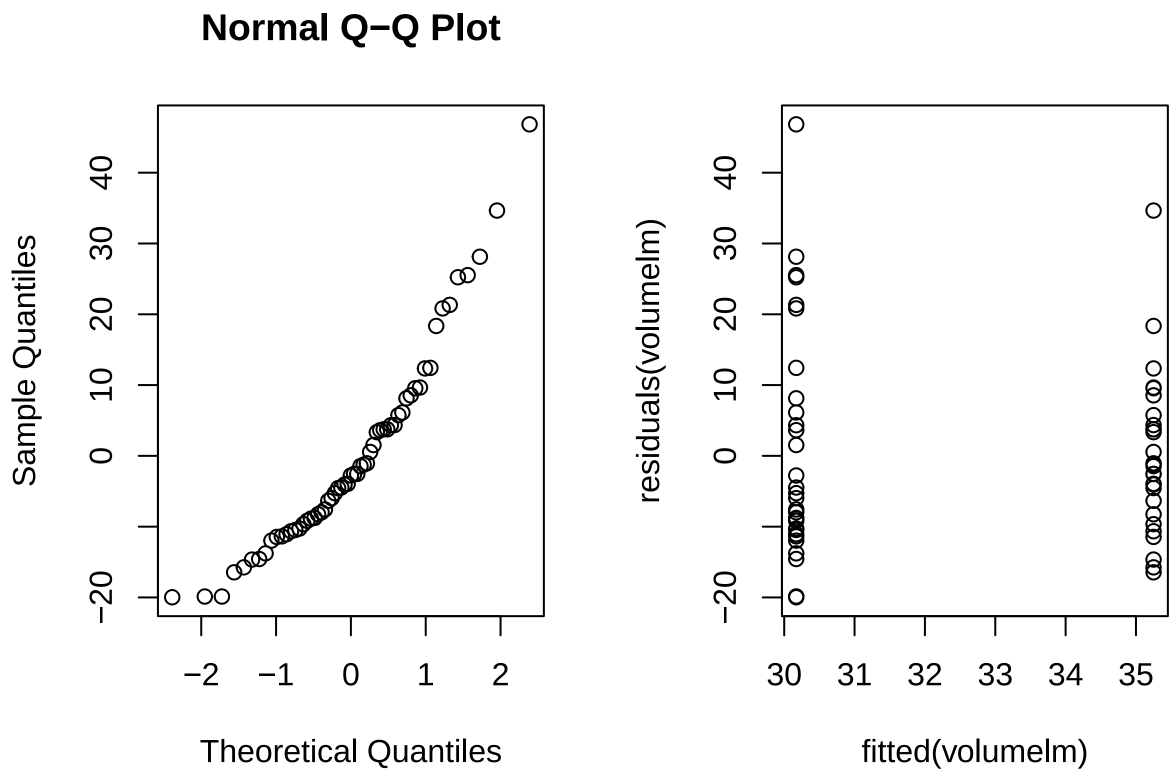
Exercise 1

a)

```r
treeVolume = read.table("treeVolume.txt", header = T)
head(treeVolume)
```

Since P value for type is 0.1736, it indicates that we fail to reject the null hypothesis and conclude that there is no significant difference in mean volume between the beech and oak trees.

```r
par(mfrow=c(1,2))
qqnorm(residuals(volumelm))
plot(fitted(volumelm), residuals(volumelm))
```




```
par(mfrow=c(1,1))
```

After testing the data for normality, it becomes obvious that there is a deviation of residuals from the normal distribution. As ANOVA assumptions have been violated, p-value may not be reliable.

```
t.test(volume ~ factor(type), data=treeVolume)
```

T-test shows the p-value 0.1659, indicating that we do not reject a null hypothesis. That indicated the difference in means between volumes of beech and oak is different.

T-test also displays estimates of mean for group beech (30.17097) and oak (35.25000)

b)

```
volumeFulllm1 = lm(volume ~ factor(height) + factor(type) * factor(diameter), data = treeVolume
anova(volumeFulllm1)
```

```
volumeFulllm2 = lm(volume ~ factor(diameter) + factor(type) * factor(height), data = treeVolume
anova(volumeFulllm2)
```

```
## Analysis of Variance Table
##
## Response: volume
##                              Df  Sum Sq Mean Sq  F value    Pr(>F)
## factor(diameter)             44 11606.8 263.790 2573.559 0.0003885 ***
## factor(type)                  1     2.6   2.602   25.386 0.0372075 *
## factor(height)               10   143.9  14.386  140.355 0.0070944 **
## factor(type):factor(height)   1    20.9  20.907  203.968 0.0048670 **
## Residuals                     2     0.2   0.102
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

c)

```r
volumeFulllm = lm(volume ~ factor(diameter) + factor(height) + factor(type), data = treeVolume)
anova(volumeFulllm)
```

```
## Analysis of Variance Table
##
## Response: volume
##                  Df  Sum Sq Mean Sq F value   Pr(>F)
## factor(diameter) 44 11606.8 263.790 37.4849 0.005969 **
## factor(height)   10   146.4  14.640  2.0804 0.297046
## factor(type)      1     0.1   0.064  0.0091 0.930038
## Residuals         3    21.1   7.037
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# TODO: how to use mean here? We were training on factors
# d = mean(treeVolume$diameter)
# h = mean(treeVolume$height)
d = median(treeVolume$diameter)
h = median(treeVolume$height)
avgTree = data.frame(diameter=d, height=h, type="oak")
avgTrees = data.frame(diameter=c(d,d), height=c(h,h), type=c("oak", "beech"))

predict(volumeFulllm, avgTree, interval = 'prediction')
```

```
## Warning in predict.lm(volumeFulllm, avgTree, interval = "prediction"):
## prediction from a rank-deficient fit may be misleading
```

```
##    fit      lwr      upr
## 1 32.8 8.186624 57.41338
```

d)

The least significant factor can be removed from the explanation, which is type, as when calculating the volume only height and diameter should be of a significance. However, while type by itself does nothing, it explains the height or the diameter, therefore its interaction with one of other factors should be included.

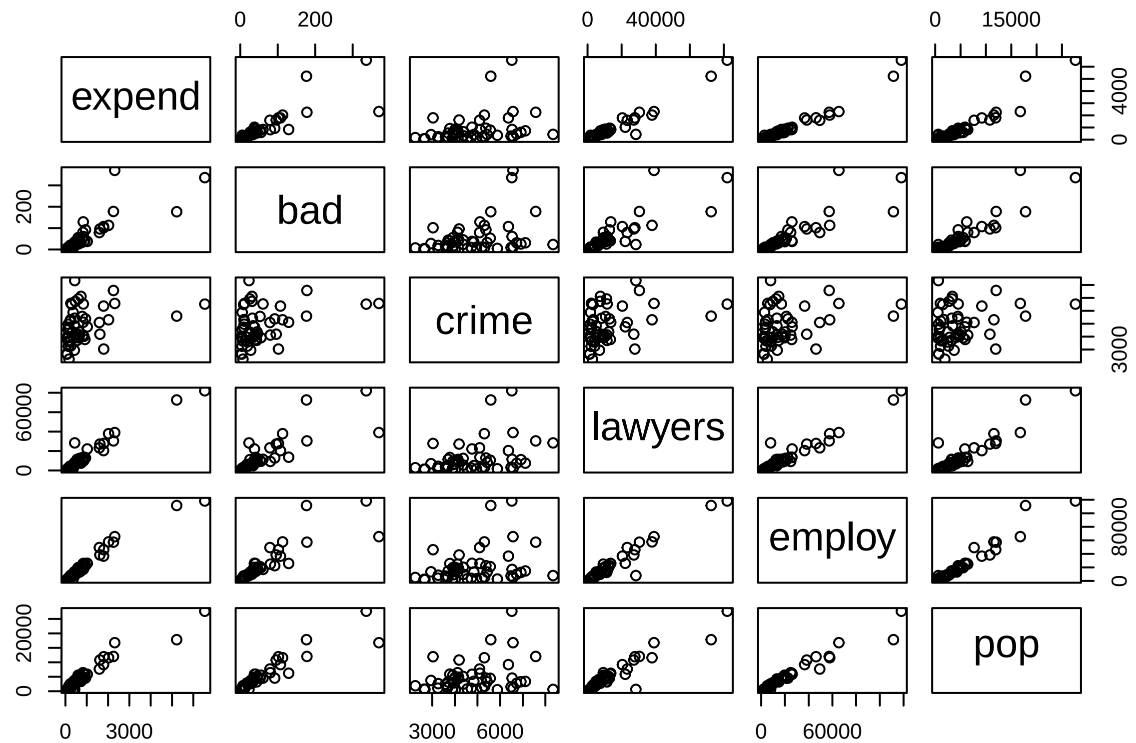
```
volumeOptimallm = lm(volume ~ factor(diameter) + factor(height) + factor(height):factor(type),
anova(volumeOptimallm)
```

```
## Analysis of Variance Table
##
## Response: volume
##                              Df  Sum Sq Mean Sq F value    Pr(>F)
## factor(diameter)             44 11606.8 263.790 2573.56 0.0003885 ***
## factor(height)               10   146.4  14.640  142.83 0.0069720 **
## factor(height):factor(type)   2    21.0  10.485  102.30 0.0096809 **
## Residuals                     2     0.2   0.102
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exercise 2

a)

```
expensescrime = read.table("expensescrime.txt", header = T)
pairs(expensescrime[,-1])
```



Influence points:

many paired scatter plots (i.e. expend vs bad, layers vs employ) have most of the data is skewed to the left and there are strong outliers on the right.

Collinearity:

Multiple variables are clearly collinear: bad & employ (`cor(expensescrime$bad, expensescrime$employ)` is 0.871), bad & lawyers (0.832), bad & pop (0.92), lawyers & employ (0.966), lawyers & pop (0.934)

b)

```
summary(lm(expend ~ factor(bad), data=expensescrime))$r.squared
```

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

```
summary(lm(expend ~ bad, data=expensescrime))$r.squared
```

```
## [1] 0.6963839
```

```r
summary(lm(expend ~ crime, data=expensescrime))$r.squared
```

```
## [1] 0.1118564
```

```r
summary(lm(expend ~ lawyers, data=expensescrime))$r.squared
```

```
## [1] 0.9372789
```

```r
summary(lm(expend ~ employ, data=expensescrime))$r.squared
```

```
## [1] 0.9539745
```

```r
summary(lm(expend ~ pop, data=expensescrime))$r.squared
```

```
## [1] 0.9073261
```

The first variable to add is **employ**, as it is significant and yields the best multiple R-squared value **0.9539745**

```r
summary(lm(expend ~ employ + bad, data=expensescrime))$r.squared
```

```
## [1] 0.955097
```

```r
summary(lm(expend ~ employ + crime, data=expensescrime))$r.squared
```

```
## [1] 0.9550501
```

```r
summary(lm(expend ~ employ + lawyers, data=expensescrime))$r.squared
```

```
## [1] 0.9631745
```

```r
summary(lm(expend ~ employ + pop, data=expensescrime))$r.squared
```

```
## [1] 0.95431
```

Next one to add is **lawyers** with the R-squared value **0.9631745**

```r
summary(lm(expend ~ employ + lawyers + bad, data=expensescrime))$r.squared
```

```
## [1] 0.9638741
```

```
summary(lm(expend ~ employ + lawyers + crime, data=expensescrime))$r.squared
```

```
## [1] 0.9631881
```

```
summary(lm(expend ~ employ + lawyers + pop, data=expensescrime))$r.squared
```

```
## [1] 0.9637326
```

Other variables upon testing showed no significance, therefore the final model is

```
model = lm(expend ~ employ + lawyers, data=expensescrime)
```

c)

```
newData = data.frame(bad=50, crime=5000, lawyers=5000, employ=5000, pop=5000)
predict(model, newData, interval = 'prediction')
```

```
##        fit       lwr      upr
## 1 172.2098 -302.9307 647.3504
```

Interval may be improved by adding interaction between variables. That could possible result in a better fit, thus precise interval and a predication.

d)

Exercise 3

a)

```
titanic = read.table("titanic.txt", header = T)
```

```
titlm = lm(Survived ~ factor(PClass) + Age + factor(Sex), data=titanic)
summary(titlm)
```

```
##
## Call:
## lm(formula = Survived ~ factor(PClass) + Age + factor(Sex), data = titanic)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.11851 -0.25363 -0.06171  0.22976  1.03436
```

```
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.130523   0.051941  21.766  < 2e-16 ***
## factor(PClass)2nd -0.207434   0.039240  -5.286 1.64e-07 ***
## factor(PClass)3rd -0.393344   0.037710 -10.431  < 2e-16 ***
## Age               -0.006005   0.001106  -5.430 7.63e-08 ***
## factor(Sex)male   -0.501326   0.029420 -17.040  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.388 on 751 degrees of freedom
##   (557 observations deleted due to missingness)
## Multiple R-squared:  0.3836, Adjusted R-squared:  0.3803
## F-statistic: 116.9 on 4 and 751 DF,  p-value: < 2.2e-16
```

All the predictor variables show significance. Female sex is not represented in the summary

b)

```
titIntlm1 = lm(Survived ~ factor(PClass) + Age + factor(Sex) + factor(PClass):Age, data=titani
summary(titIntlm1)
```

```
##
## Call:
## lm(formula = Survived ~ factor(PClass) + Age + factor(Sex) +
##     factor(PClass):Age, data = titanic)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.12349 -0.25341 -0.06777  0.22244  0.96212
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            1.135658   0.074103  15.325  < 2e-16 ***
## factor(PClass)2nd     -0.101386   0.097095  -1.044 0.296737
## factor(PClass)3rd     -0.489568   0.091035  -5.378 1.01e-07 ***
## Age                   -0.006086   0.001730  -3.518 0.000462 ***
## factor(Sex)male       -0.504762   0.029349 -17.198  < 2e-16 ***
## factor(PClass)2nd:Age -0.003774   0.002678  -1.409 0.159110
## factor(PClass)3rd:Age  0.003788   0.002617   1.447 0.148195
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3867 on 749 degrees of freedom
```

```
##   (557 observations deleted due to missingness)
## Multiple R-squared:  0.3894, Adjusted R-squared:  0.3845
## F-statistic: 79.62 on 6 and 749 DF,  p-value: < 2.2e-16
```

```
titIntlm2 = lm(Survived ~ factor(PClass) + Age + factor(Sex) + factor(Sex):Age, data=titanic)
summary(titIntlm2)
```

```
##
## Call:
## lm(formula = Survived ~ factor(PClass) + Age + factor(Sex) +
##     factor(Sex):Age, data = titanic)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.96406 -0.24565 -0.03889  0.25356  1.09610
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.9472923  0.0613408  15.443  < 2e-16 ***
## factor(PClass)2nd   -0.2091063  0.0385318  -5.427 7.75e-08 ***
## factor(PClass)3rd   -0.3950952  0.0370296 -10.670  < 2e-16 ***
## Age                  0.0002662  0.0015937   0.167  0.86739
## factor(Sex)male     -0.1810169  0.0662114  -2.734  0.00641 **
## Age:factor(Sex)male -0.0106501  0.0019809  -5.376 1.02e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.381 on 750 degrees of freedom
##   (557 observations deleted due to missingness)
## Multiple R-squared:  0.4065, Adjusted R-squared:  0.4025
## F-statistic: 102.7 on 5 and 750 DF,  p-value: < 2.2e-16
```

```
titIntlm3 = lm(Survived ~ factor(PClass) + Age + factor(Sex) + factor(PClass):Age + factor(Sex)
summary(titIntlm3)
```

```
##
## Call:
## lm(formula = Survived ~ factor(PClass) + Age + factor(Sex) +
##     factor(PClass):Age + factor(Sex):Age, data = titanic)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.95991 -0.23838 -0.05496  0.23793  1.02666
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)           9.600e-01  8.001e-02  11.998  < 2e-16 ***
## factor(PClass)2nd     -1.181e-01  9.544e-02  -1.237  0.21645
## factor(PClass)3rd     -4.907e-01  8.944e-02  -5.487 5.60e-08 ***
## Age                   -3.838e-05  2.048e-03  -0.019  0.98506
## factor(Sex)male       -1.898e-01  6.612e-02  -2.871  0.00421 **
## factor(PClass)2nd:Age -3.289e-03  2.632e-03  -1.249  0.21195
## factor(PClass)3rd:Age  3.700e-03  2.571e-03   1.439  0.15053
## Age:factor(Sex)male   -1.046e-02  1.977e-03  -5.292 1.59e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3799 on 748 degrees of freedom
##   (557 observations deleted due to missingness)
## Multiple R-squared:  0.4115, Adjusted R-squared:  0.4059
## F-statistic:  74.7 on 7 and 748 DF,  p-value: < 2.2e-16
```

PClass : age interaction is not significant, thus it should not be included in the final model, while *age : sex* interaction is shown as significant only for male. The proposal for the final model is as follows

```
tlm = lm(Survived ~ factor(PClass) + Age + factor(Sex) + factor(Sex):Age, data=titanic)
summary(tlm)
```

```
##
## Call:
## lm(formula = Survived ~ factor(PClass) + Age + factor(Sex) +
##     factor(Sex):Age, data = titanic)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.96406 -0.24565 -0.03889  0.25356  1.09610
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.9472923  0.0613408  15.443  < 2e-16 ***
## factor(PClass)2nd   -0.2091063  0.0385318  -5.427 7.75e-08 ***
## factor(PClass)3rd   -0.3950952  0.0370296 -10.670  < 2e-16 ***
## Age                  0.0002662  0.0015937   0.167  0.86739
## factor(Sex)male     -0.1810169  0.0662114  -2.734  0.00641 **
## Age:factor(Sex)male -0.0106501  0.0019809  -5.376 1.02e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.381 on 750 degrees of freedom
##   (557 observations deleted due to missingness)
## Multiple R-squared:  0.4065, Adjusted R-squared:  0.4025
## F-statistic: 102.7 on 5 and 750 DF,  p-value: < 2.2e-16
```

```r
tlm = lm(Survived ~ factor(PClass) + factor(Sex) + factor(Sex):Age, data=titanic)
summary(tlm)
```

```
##
## Call:
## lm(formula = Survived ~ factor(PClass) + factor(Sex) + factor(Sex):Age,
##     data = titanic)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.96406 -0.24565 -0.03889  0.25356  1.09610
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             0.9472923  0.0613408  15.443  < 2e-16 ***
## factor(PClass)2nd      -0.2091063  0.0385318  -5.427 7.75e-08 ***
## factor(PClass)3rd      -0.3950952  0.0370296 -10.670  < 2e-16 ***
## factor(Sex)male        -0.1810169  0.0662114  -2.734  0.00641 **
## factor(Sex)female:Age   0.0002662  0.0015937   0.167  0.86739
## factor(Sex)male:Age    -0.0103839  0.0013575  -7.649 6.20e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.381 on 750 degrees of freedom
##   (557 observations deleted due to missingness)
## Multiple R-squared:  0.4065, Adjusted R-squared:  0.4025
## F-statistic: 102.7 on 5 and 750 DF,  p-value: < 2.2e-16
```

Age is not significant by itself anymore, thus it can be removed.

c)

One method to predict the survival status is to use the final model tlm and apply it to new data containing the predictor variables for each passenger. The resulting model can be used to predict the probability of survival for each passenger.

To measure the quality of the prediction, we can use metrics such as accuracy, precision, recall, and F1-score. These metrics compare the predicted survival status to the actual survival status for each passenger. Accuracy measures the overall proportion of correct predictions, while precision measures the proportion of true positives among all positive predictions and recall measures the proportion of true positives among all actual positives. The F1-score is the harmonic mean of precision and recall and provides a balanced measure of the model's performance.

To implement this method, we would first split the titanic dataset into a training set and a test set. We would then fit the final model tlm to the training set and use it to predict the survival status for each passenger in the test set. We would calculate the quality measures for the predictions and use them to evaluate the performance of the model. If the performance is satisfactory, we can use the model to predict the survival status for new passengers in the future.

d)

e)

Exercise 4

a)

```r
coups = read.table("coups.txt", header = T)
```

```r
model <- glm(miltcoup ~ ., data = coups, family = "poisson")
summary(model)
```

```
##
## Call:
## glm(formula = miltcoup ~ ., family = "poisson", data = coups)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3443  -0.9542  -0.2587   0.3905   1.6953
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.5102693  0.9053301  -0.564  0.57301
## oligarchy    0.0730814  0.0345958   2.112  0.03465 *
## pollib      -0.7129779  0.2725635  -2.616  0.00890 **
## parties      0.0307739  0.0111873   2.751  0.00595 **
## pctvote      0.0138722  0.0097526   1.422  0.15491
## popn         0.0093429  0.0065950   1.417  0.15658
## size        -0.0001900  0.0002485  -0.765  0.44447
## numelec     -0.0160783  0.0654842  -0.246  0.80605
## numregim     0.1917349  0.2292890   0.836  0.40303
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 65.945  on 35  degrees of freedom
## Residual deviance: 28.668  on 27  degrees of freedom
## AIC: 111.48
##
## Number of Fisher Scoring iterations: 6
```

b)

c)