Assignment 2

Alexia Salomons, Nathan Maxwell Jones, Yauheniya Makarevich, group 71

15 March 2023

Exercise 1

a) To investigate whether tree type influences total wood volume, we can perform a one-way ANOVA.

```
tree_df$type <- as.factor(tree_df$type)
tree_type_lm <- lm(volume~type, data=tree_df)
anova(tree_type_lm)
```

```
## 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(tree_type_lm)
```

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

With $p > 0.05$, we can conclude that *type* does not have a significant effect on *volume*. Because the factor *type* has two levels, we can apply a two sample t-test.

```r
mask <- tree_df$type == "beech"
t.test(tree_df$volume[mask], tree_df$volume[!mask])
```

```
##
##  Welch Two Sample t-test
##
## data:  tree_df$volume[mask] and tree_df$volume[!mask]
## t = -1, df = 53, p-value = 0.2
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -12.33   2.17
## sample estimates:
## mean of x mean of y
##      30.2      35.2
```

This supports the result from the ANOVA test. The estimated volume is 30.2 for Beech trees and 35.2 for Oak trees.

b) To investigate this claim, we create two models, each including all three explanatory variables (*type*, *diameter* and *height*). In the first model, we also include the pairwise interaction between *type* and *diameter*.

```r
tree_type_d_lm <- lm(volume~height+type*diameter, data=tree_df)
anova(tree_type_d_lm)
```

```
## Analysis of Variance Table
##
## Response: volume
##               Df Sum Sq Mean Sq F value  Pr(>F)
## height         1   2188    2188  206.21 < 2e-16 ***
## type           1    431     431   40.65 4.2e-08 ***
## diameter       1   8577    8577  808.49 < 2e-16 ***
## type:diameter  1      6       6    0.52    0.47
## Residuals     54    573      11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(tree_type_d_lm)
```

```
##
## Call:
## lm(formula = volume ~ height + type * diameter, data = tree_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.350 -2.194 -0.141  1.701  8.176
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -63.873      5.539  -11.53  3.5e-16 ***
```

```
## height              0.434      0.079    5.49  1.1e-06 ***
## typeoak            -4.963      5.149   -0.96     0.34
## diameter            4.608      0.207   22.26  < 2e-16 ***
## typeoak:diameter    0.259      0.359    0.72     0.47
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.26 on 54 degrees of freedom
## Multiple R-squared:  0.951,  Adjusted R-squared:  0.948
## F-statistic:  264 on 4 and 54 DF,  p-value: <2e-16
```

```
tree_type_h_lm <- lm(volume~diameter+type*height, data=tree_df)
anova(tree_type_h_lm)
```

```
## Analysis of Variance Table
##
## Response: volume
##              Df Sum Sq Mean Sq F value  Pr(>F)
## diameter      1  10827   10827 1045.97 < 2e-16 ***
## type          1     45      45    4.37   0.041 *
## height        1    324     324   31.32 7.5e-07 ***
## type:height   1     19      19    1.88   0.176
## Residuals    54    559      10
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(tree_type_h_lm)
```

```
##
## Call:
## lm(formula = volume ~ diameter + type * height, data = tree_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.230 -2.113 -0.161  1.801  8.165
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -57.551      7.111   -8.09    7e-11 ***
## diameter            4.779      0.173   27.55   <2e-16 ***
## typeoak           -17.471     11.826   -1.48   0.1454
## height              0.321      0.102    3.14   0.0027 **
## typeoak:height      0.212      0.154    1.37   0.1761
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.22 on 54 degrees of freedom
## Multiple R-squared:  0.953,  Adjusted R-squared:  0.949
## F-statistic:  271 on 4 and 54 DF,  p-value: <2e-16
```

We see that both pairwise interactions are not significant. Therefore, we can conclude that both *height* and *diameter* have the same influence regardless of *type*. Both models suggest that all three explanatory variables have a significant effect individually.

c)

In (b), we saw that the interactions of *height* and *diameter* with *type* were not significant, and so we will investigate a purely additive model (assuming no interactions).

```
tree_add_all_lm <- lm(volume~diameter+height+type, data=tree_df)
anova(tree_add_all_lm)
```

```
## 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(tree_add_all_lm)
```

```
##
## Call:
## lm(formula = volume ~ diameter + height + type, data = tree_df)
##
## 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
```

We see that the effect of *type* is not significant in the additive model. Therefore we will investigate an additive model that excludes *type*.

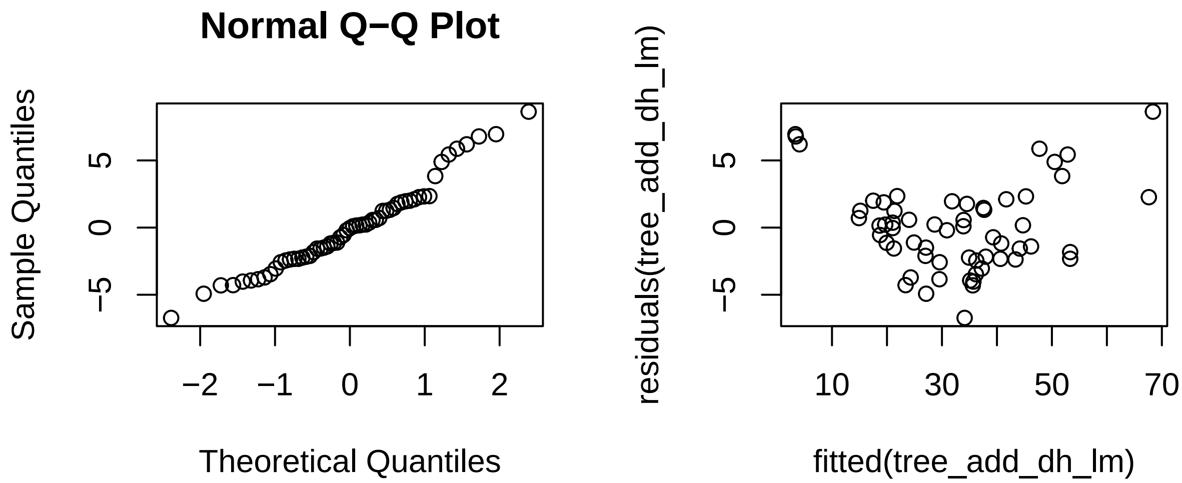
```
tree_add_dh_lm <- lm(volume~diameter+height, data=tree_df)
anova(tree_add_dh_lm)
```

```
## Analysis of Variance Table
##
## Response: volume
##           Df Sum Sq Mean Sq F value  Pr(>F)
## diameter   1  10827   10827  1007.8 < 2e-16 ***
## height     1    346     346    32.2 5.1e-07 ***
## Residuals 56    602      11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(tree_add_dh_lm)
```

```
##
## Call:
## lm(formula = volume ~ diameter + height, data = tree_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.724 -2.278 -0.034  1.820  8.629
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -64.3697     5.5577  -11.58  < 2e-16 ***
## diameter      4.6325     0.1602   28.92  < 2e-16 ***
## height        0.4289     0.0755    5.68  5.1e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.28 on 56 degrees of freedom
## Multiple R-squared:  0.949,  Adjusted R-squared:  0.947
## F-statistic:  520 on 2 and 56 DF,  p-value: <2e-16
```

This model has almost the same R-squared value as before, while using fewer variables. Since simpler models are generally preferred, this is our model of choice to make predictions. As a final test, we need to check this model's assumptions to ensure that the conclusions we draw from it are valid:



While these plots are not perfect, we believe the model assumptions to be valid.

Therefore, the effects of *type*, *diameter* and *height* can be summarized as follows:

- The tree *type* does not affect volume significantly.
- Looking at the coefficients, we see that increasing both height and diameter result in an increase in volume, with diameter having a bigger impact (with a gradient of 4.63 compared to *height's* 0.43). This makes sense given that we know volume is proportional to the the square of the diameter.

To predict the volume for a tree with the overall average diameter and height, we can use the following linear regression model:

$$volume = -64.37 + 4.63 * diameter + 0.43 * height$$

```
mean_d <- mean(tree_df$diameter)
mean_h <- mean(tree_df$height)
means <-  data.frame(diameter=c(mean_d), height=c(mean_h))

predict(tree_add_dh_lm, means, se.fit = TRUE)
```

```
## $fit
##    1
## 32.6
##
## $se.fit
## [1] 0.427
##
## $df
## [1] 56
##
## $residual.scale
## [1] 3.28
```

Therefore we expect the volume for such a tree to be 32.6.

d) Assuming that a tree is roughly cylindrical, we expect that *volume* would be proportional to the *height* multiplied by the square of *diameter*. We perform this transformation and add it as a new column in the data frame. We could apply the true transformation, $V = h \times \pi(d/2)^2$, but this would just add unnecessary constants which would already be captured in the regression coefficients.

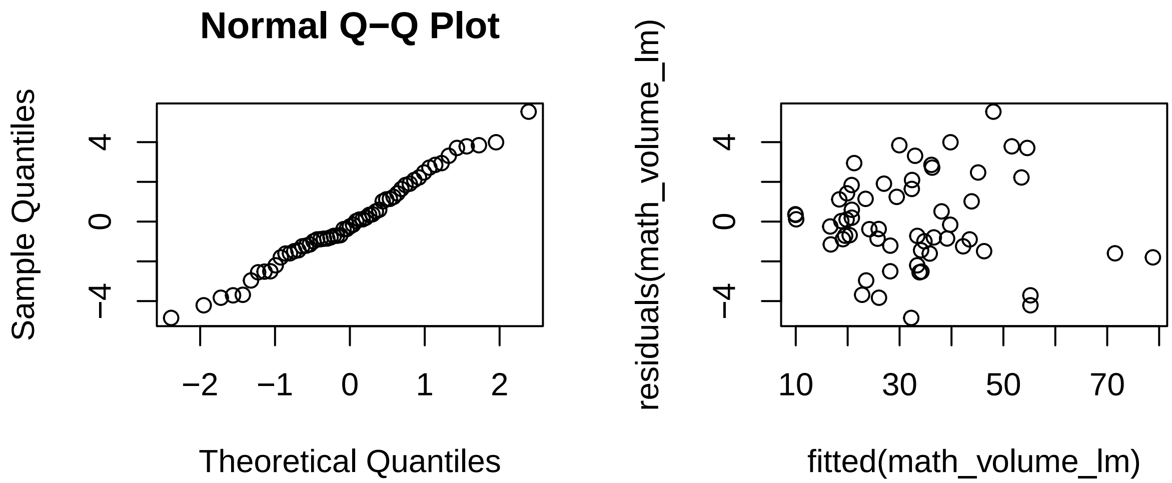
```
tree_df$math_volume <- tree_df$height * tree_df$diameter^2
math_volume_lm <- lm(volume~math_volume, data=tree_df)
anova(math_volume_lm)
```

```
## Analysis of Variance Table
##
## Response: volume
##             Df Sum Sq Mean Sq F value Pr(>F)
## math_volume  1  11477   11477    2201 <2e-16 ***
## Residuals   57    297       5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(math_volume_lm)
```

```
##
## Call:
## lm(formula = volume ~ math_volume, data = tree_df)
##
## 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
## math_volume  2.14e-03   4.57e-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
```

We see that this transformation does indeed produce an explanatory value with a significant effect. We also see that the R-squared value of 0.975 is higher than that of the previous models, indicating that it better explains the data. Finally, we check the assumptions of this model.

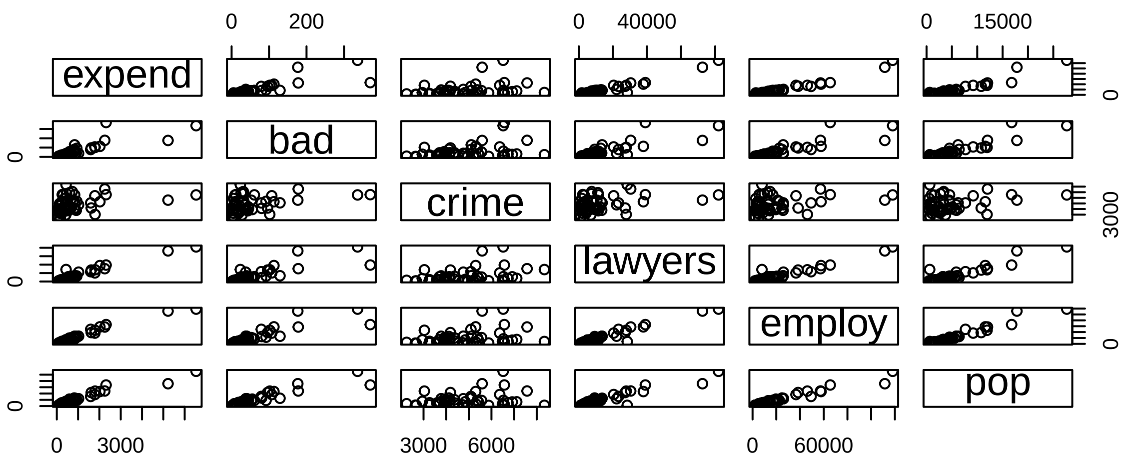


These plots are acceptable, meaning we can accept the model assumptions.

Exercise 2

a) «««« INCLUDE OTHER GRAPHICAL SUMMARIES??? »»»»

To investigate the interactions between all the variables of interest, we can plot the pairwise scatter plots for all their combinations:

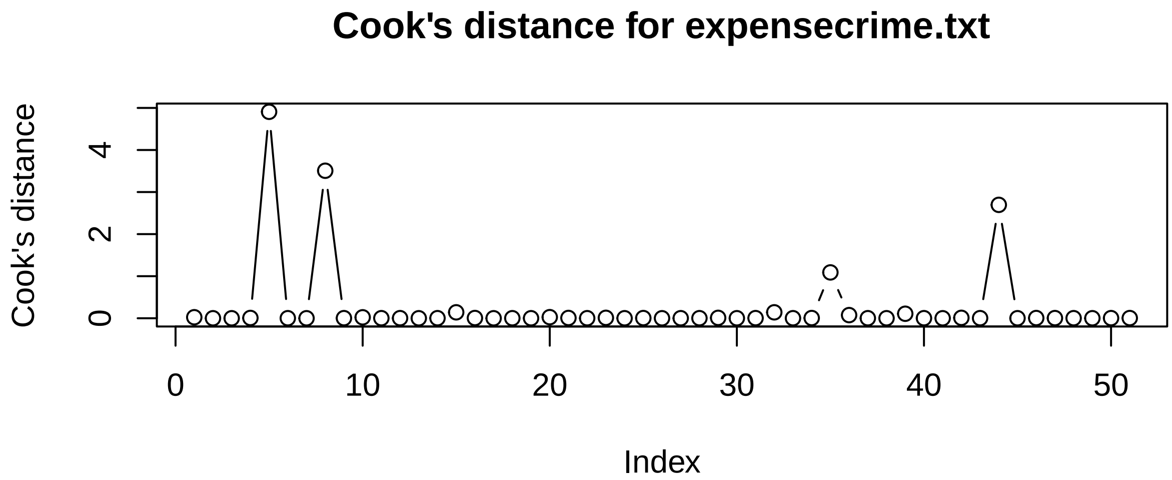


We see that *expend*, our response variable, appears to have a positive correlation with all the explanatory variables except for *crime*. There appear to be several outliers at the high end of the data which could skew the model. We can also see that collinearity exists between the explanatory variables *bad*, *lawyers*, *employ* and *pop*. This is a problem since the redundant information will make the regression coefficients difficult to estimate.

We can use Cook's distance to find the influence points (a distance greater than 1 indicates an outlier)

```
crime_lm <- lm(expend~bad+crime+lawyers+employ+pop, data=crime_df)
cooks.distance(crime_lm)[cooks.distance(crime_lm) > 1]
```

```
##    5    8   35   44
## 4.91 3.51 1.09 2.70
```



We can see that indices of 5, 8, 35 and 44 are outliers, which we can remove:

```
crime_df_upd <- crime_df[-c(5,8,35,44),]
```

To further investigate collinearity, we can examine the correlations between all the explanatory variables, which confirms strong correlations between *bad*, *lawyers*, *employ* and *pop*.

```
round(cor(crime_df[, c(columns)]), 2)
```

```
##          bad crime lawyers employ  pop
## bad     1.00  0.37    0.83   0.87 0.92
## crime   0.37  1.00    0.38   0.31 0.28
## lawyers 0.83  0.38    1.00   0.97 0.93
## employ  0.87  0.31    0.97   1.00 0.97
## pop     0.92  0.28    0.93   0.97 1.00
```

«««« IS VIF NECCESSARY? »»»»

To resolve the problem of collinearity, we can iteratively remove variables based on their VIF-values as follows:

Full model

```
vif(lm(expend~bad+crime+lawyers+employ+pop, data=crime_df))
```

```
##     bad   crime lawyers  employ     pop
##    8.36    1.49   16.97   33.59   32.94
```

Remove *employ*

```
vif(lm(expend~bad+crime+lawyers+pop, data=crime_df_upd))
```

```
##     bad   crime lawyers     pop
##    7.16    1.34   12.40   20.72
```

Remove *pop*

```
vif(lm(expend~bad+crime+lawyers, data=crime_df_upd))
```

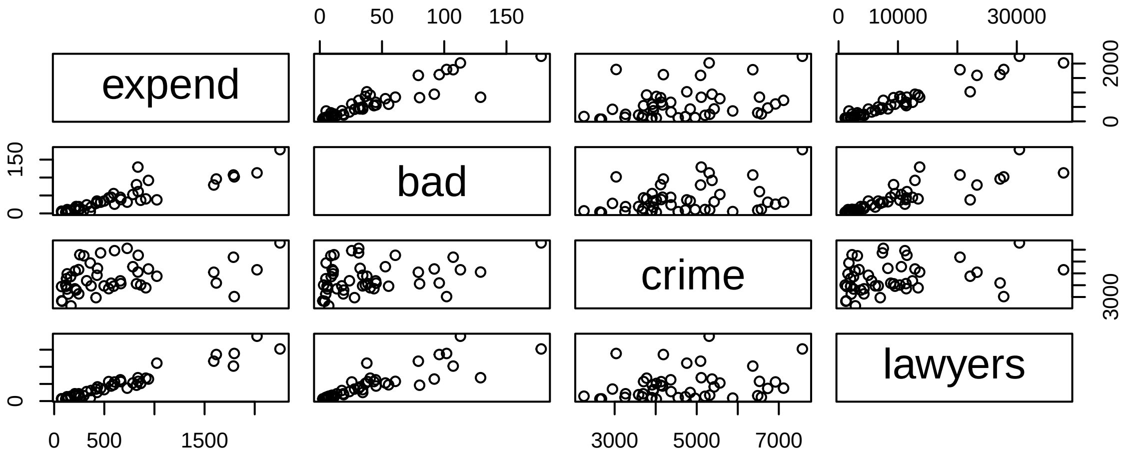
```
##     bad   crime lawyers
```

```
##     3.99     1.13     3.78
```

«««« SHOULD WE SHOW PLOT AGAIN?? »»»»

Therefore, after removing the influence points and collinear explanatory variables, the adjusted scatter plot appears as follows. We will work with this adjusted data for the remainder of this question.

```
pairs(crime_df_upd[, c(target, "bad", "crime", "lawyers")])
```



b)

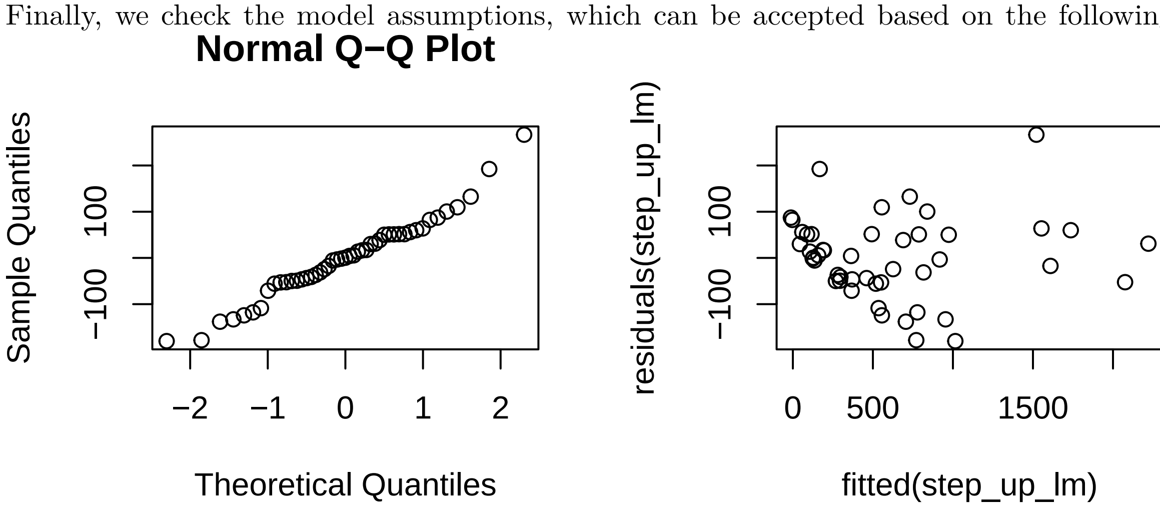
The step-up process was carried out. The variables added in order were *employ*, *crime* and *pop*, after which no further added variables had significant p-values. Hence the final model is as follows:

```
step_up_lm <- lm(expend~employ+crime+pop, data=crime_df_upd)
summary(step_up_lm)
```

```
##
## Call:
## lm(formula = expend ~ employ + crime + pop, data = crime_df_upd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -179.99  -49.64    0.48   51.19  266.63
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.47e+02   5.47e+01   -4.52  4.8e-05 ***
## employ       2.09e-02   3.95e-03    5.30  3.7e-06 ***
## crime        5.43e-02   1.13e-02    4.82  1.8e-05 ***
## pop          7.14e-02   1.79e-02    4.00  0.00025 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 91.4 on 43 degrees of freedom
## Multiple R-squared:  0.974,  Adjusted R-squared:  0.973
## F-statistic:  547 on 3 and 43 DF,  p-value: <2e-16
```

Final model: expend = -247 + 0.0209*employ + 0.0543*crime + 0.0714*pop ± error, with $R^2 = 0.974$. Step - up naturally removes collinearity and can be compared with VIF results.

Finally, we check the model assumptions, which can be accepted based on the following plots:



««« COMPARE TO MODEL FROM A??? »»»>

Worse R-squared! (0.957!)… expand… step up better way of avoiding non-linearity?

```
summary(lm(expend~bad+crime+lawyers, data=crime_df_upd))
```

```
##
## Call:
## lm(formula = expend ~ bad + crime + lawyers, data = crime_df_upd)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -328.4  -42.4  -14.2   33.8  355.3
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -113.5051    66.5587   -1.71  0.09535 .
## bad            3.7457     0.8845    4.23  0.00012 ***
## crime          0.0333     0.0145    2.30  0.02655 *
## lawyers        0.0456     0.0039   11.68  6.3e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 118 on 43 degrees of freedom
## Multiple R-squared:  0.957,  Adjusted R-squared:  0.954
## F-statistic:  322 on 3 and 43 DF,  p-value: <2e-16
```

c)

```
mean(crime_df_upd$expend)
```

```
## [1] 611
```

```r
new_data <- data.frame(bad=50, crime=5000, lawyers=5000, employ=5000, pop=5000)
predict(step_up_lm, new_data, interval="prediction", level=0.95)
```

```
##   fit lwr upr
## 1 486 258 713
```

If we think that improvement means making interval smaller, we can go for confidence interval.

```r
predict(step_up_lm, new_data, interval="confidence", level=0.95)
```

```
##   fit lwr upr
## 1 486 352 619
```

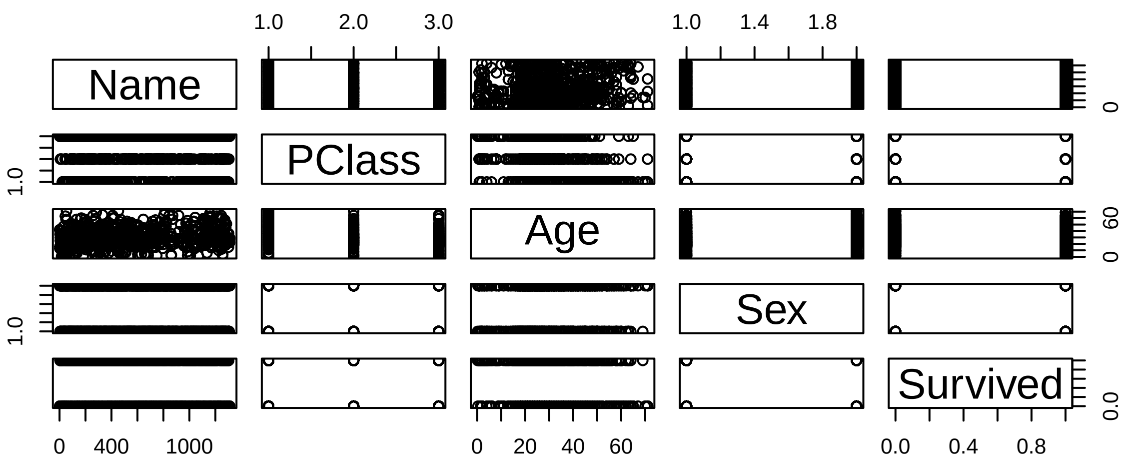
d)

Exercise 3

```r
head(titanic_df)
```

```
##                                            Name PClass   Age    Sex Survived
## 1                  Allen, Miss Elisabeth Walton    1st 29.00 female        1
## 2                   Allison, Miss Helen Loraine    1st  2.00 female        0
## 3           Allison, Mr Hudson Joshua Creighton    1st 30.00   male        0
## 4 Allison, Mrs Hudson JC (Bessie Waldo Daniels)    1st 25.00 female        0
## 5                 Allison, Master Hudson Trevor    1st  0.92   male        1
## 6                            Anderson, Mr Harry    1st 47.00   male        1
```

```r
plot(titanic_df)
```



a)

```r
titanic_df_upd <- na.omit(titanic_df)
titanic_df_upd$PClass <- as.factor(titanic_df_upd$PClass)
titanic_df_upd$Sex <- as.factor(titanic_df_upd$Sex)
head(titanic_df_upd)
```

```
##                                            Name PClass   Age    Sex Survived
## 1                  Allen, Miss Elisabeth Walton    1st 29.00 female        1
```

```
## 2                   Allison, Miss Helen Loraine    1st  2.00 female        0
## 3           Allison, Mr Hudson Joshua Creighton    1st 30.00   male        0
## 4 Allison, Mrs Hudson JC (Bessie Waldo Daniels)    1st 25.00 female        0
## 5                 Allison, Master Hudson Trevor    1st  0.92   male        1
## 6                            Anderson, Mr Harry    1st 47.00   male        1
```

b)

c)

d)

e)

Exercise 4

a)

b)

c)