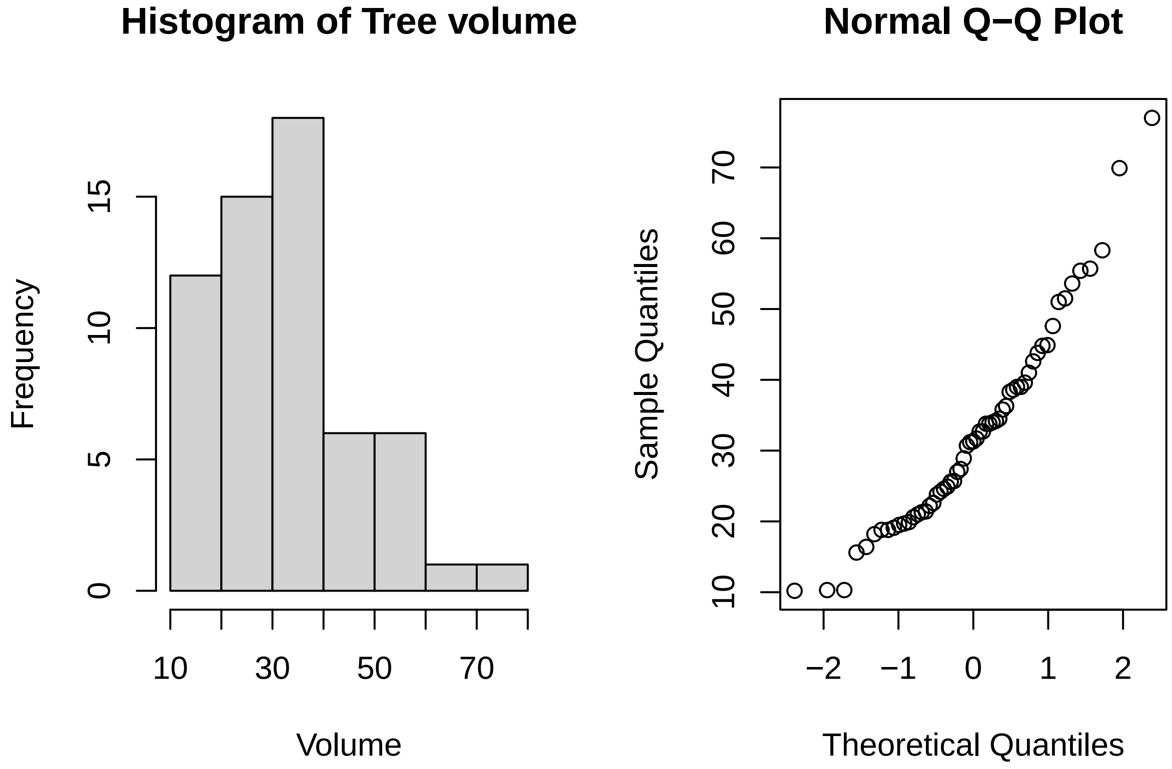
Assignment 2 - Group 27

Joost Driessen, Emma van Lipzig, Rohan Zonneveld

2023-03-15

Excercise 1. Trees

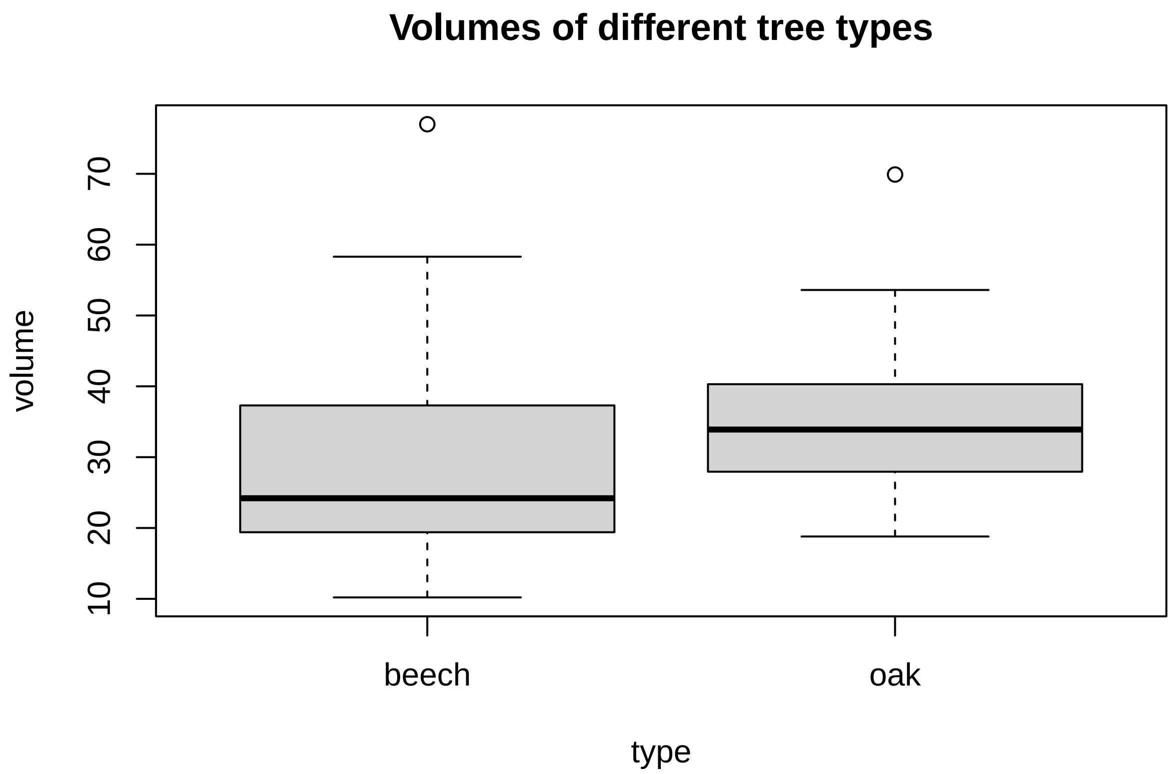
Intro



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

The histogram is slightly skewed to the right and the QQ-plot looks normal. So, the plots give no reason to suspect the data is not drawn from a normal distribution. However the Shapiro-Wilk test does reject normality ($p < 0.05$). However, since it is explicitly stated in the question that we should use anova we decided to use parametric tests throughout **Exercise 1**.

a)



```
model1 = lm(volume~type, data = mydata)
anova(model1)
```

```
## 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
```

```
beechvolume = mydata$volume[1:31]
oakvolume = mydata$volume[32:59]
t.test(beechvolume, oakvolume)
```

```
##
##  Welch Two Sample t-test
##
## data:  beechvolume and oakvolume
## 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
```

The p-value from the ANOVA-test is 0.174, which is bigger than 0.05, so we can not reject H0, thus there are equal variances between the volumes of the two tree types. A t-test can also be related to this result. The outcome of this test is that there is no difference between the mean volumes of the two types of trees. The mean volume for a beech is 30.17097 and for an oak it is 35.25.

b)

```
model2 = lm(volume~diameter*type, data = mydata)

model3 = lm(volume~height*type, data = mydata)

full_model = lm(volume~diameter * height * type, data = mydata)

anova(model2, full_model)
```

```
## Analysis of Variance Table
##
## Model 1: volume ~ diameter * type
## Model 2: volume ~ diameter * height * type
##   Res.Df RSS Df Sum of Sq    F  Pr(>F)
## 1     55 893
## 2     51 332  4       561 21.5 1.9e-10 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
anova(model3, full_model)
```

```
## Analysis of Variance Table
##
## Model 1: volume ~ height * type
## Model 2: volume ~ diameter * height * type
##   Res.Df  RSS Df Sum of Sq   F Pr(>F)
## 1     55 8413
## 2     51  332  4      8081 310 <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

The new ANOVA-tests show that both height and diameter significantly influence volume ($p < 0.01$). They also demonstrate that there is no interaction effect between diameter and type on volume, but also that there is an interaction effect between height and type on volume ($p < 0.01$).

c)

```
mydata$type = as.factor(mydata$type)
model4 = lm(volume~height + diameter + type, data = mydata)
drop1(model4, test = 'F')
```

```
## Single term deletions
##
## Model:
## volume ~ height + diameter + type
##          Df Sum of Sq  RSS AIC F value  Pr(>F)
## <none>                 578 143
## height    1       324  903 167   30.82 8.4e-07 ***
## diameter  1      8577 9155 304  815.61 < 2e-16 ***
## type      1        23  602 143    2.21    0.14
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model5 = lm(volume~height + diameter, data = mydata)
summary(model5)
```

```
##
## Call:
## lm(formula = volume ~ height + diameter, data = mydata)
##
## 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 ***
## height        0.4289     0.0755    5.68  5.1e-07 ***
## diameter      4.6325     0.1602   28.92  < 2e-16 ***
## ---
## 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
```

```
df = data.frame(diameter=mean(mydata$diameter), height = mean(mydata$height))
predict(model5, df, interval = 'prediction')
```

```
##    fit lwr  upr
## 1 32.6  26 39.2
```

All of the variables height, diameter and type have an influence on the volume of the tree, which means it would be wrong to exclude one of those variables, per the previous models. Model4 was constructed and drop1 was used to investigate whether all the used variables have an impact on the model. It shows that type is not significant, so type is excluded from the new model. The model to predict the average tree volume now only includes diameter and volume.

d)

```
mydata$newvolume = pi * (mydata$diameter/2)^2 * mydata$height
model6 = lm(volume ~ newvolume, data = mydata)
summary(model6)
```

```
##
## Call:
## lm(formula = volume ~ newvolume, data = mydata)
##
## 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
## newvolume    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
```

```
full_model = lm(volume~height + diameter + newvolume, data =mydata)
omega1 = lm(volume~height + diameter, data = mydata)
omega2 = lm(volume~newvolume, data =mydata)

anova(omega1, full_model)
```

```
## Analysis of Variance Table
##
## Model 1: volume ~ height + diameter
## Model 2: volume ~ height + diameter + newvolume
##   Res.Df RSS Df Sum of Sq    F  Pr(>F)
## 1     56 602
## 2     55 292  1       310 58.5 3.3e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

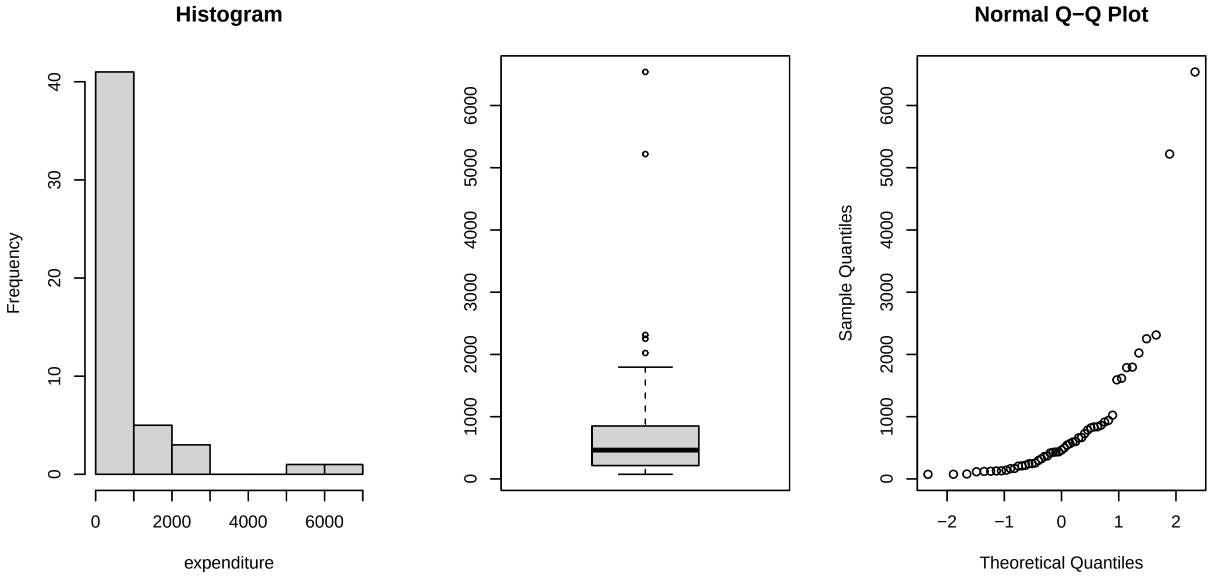
```
anova(omega2, full_model)
```

```
## Analysis of Variance Table
##
## Model 1: volume ~ newvolume
## Model 2: volume ~ height + diameter + newvolume
##   Res.Df RSS Df Sum of Sq    F Pr(>F)
## 1     57 297
## 2     55 292  2      5.61 0.53   0.59
```

This linear model might not be the best one to explain volume in terms of diameter and height. Since the volume is the product of the surface area and it's height, it would make more sense to construct as follows: $volume = \pi * (diameter/2)^2 * height$. In this model the surface area of a circle (tree) is calculated using πr^2 and it's multiplied by it's height to calculate the volume. The adjusted R-squared of this new model is 0.9743 and the ANOVA between the two submodels and the full model shows that this new model is a better predictor than the old model from *c*.

Excercise 2. Expenditure on criminal activities

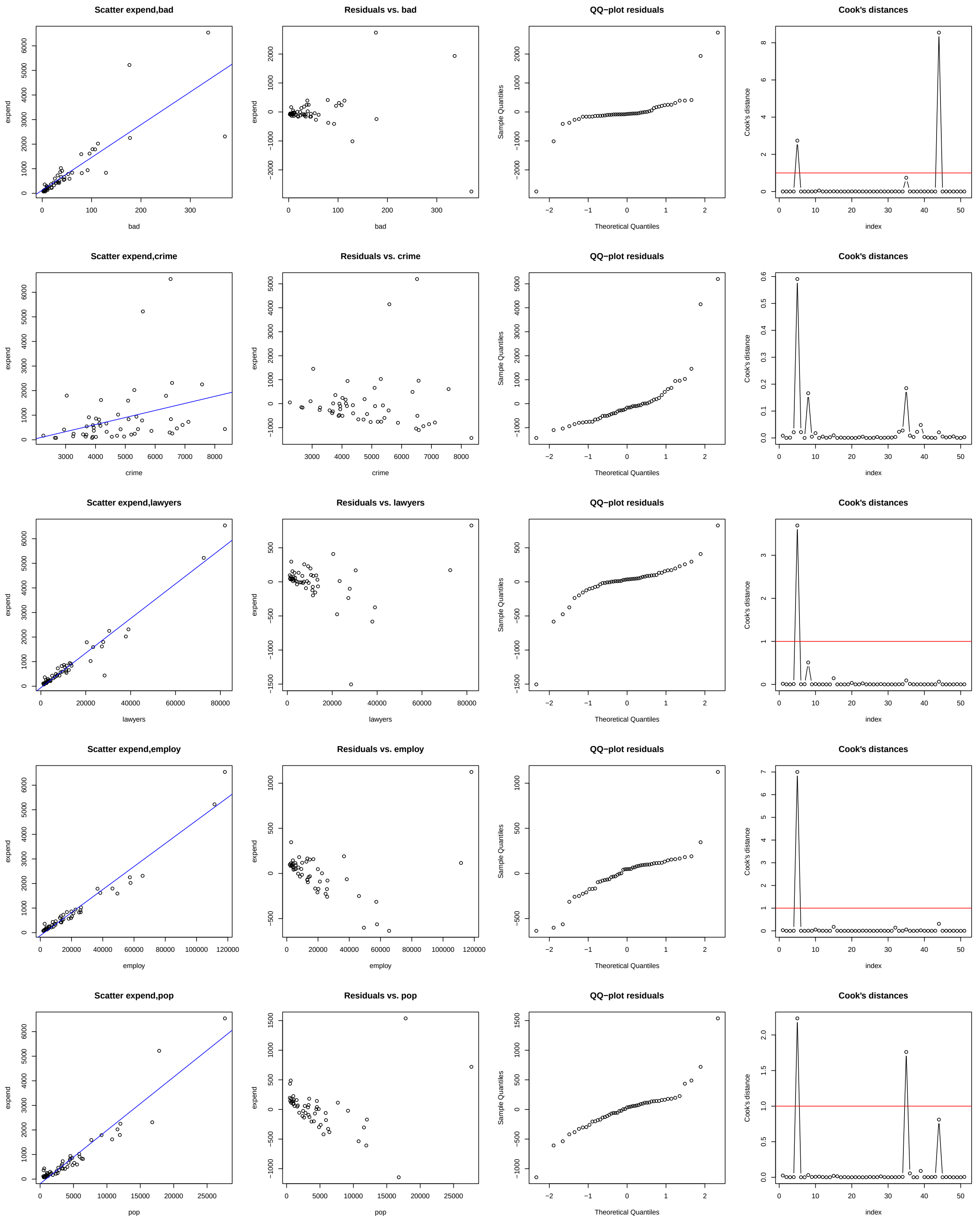
a)



This dataset contains a number outliers.

Influence Points

To investigate if these outliers are due to leverage points a scatter plot is made of the response variable expenditure against all individual explanatory variables. This plot also includes the linear regression model depicted as a blue line. Next, the residuals of a linear regression model containing the corresponding explanatory variable is plotted. Third, a QQ-plot of the residuals is plotted. Lastly, the Cook's distance is calculated and plotted with a line at cut-off value 1 depicted in red.



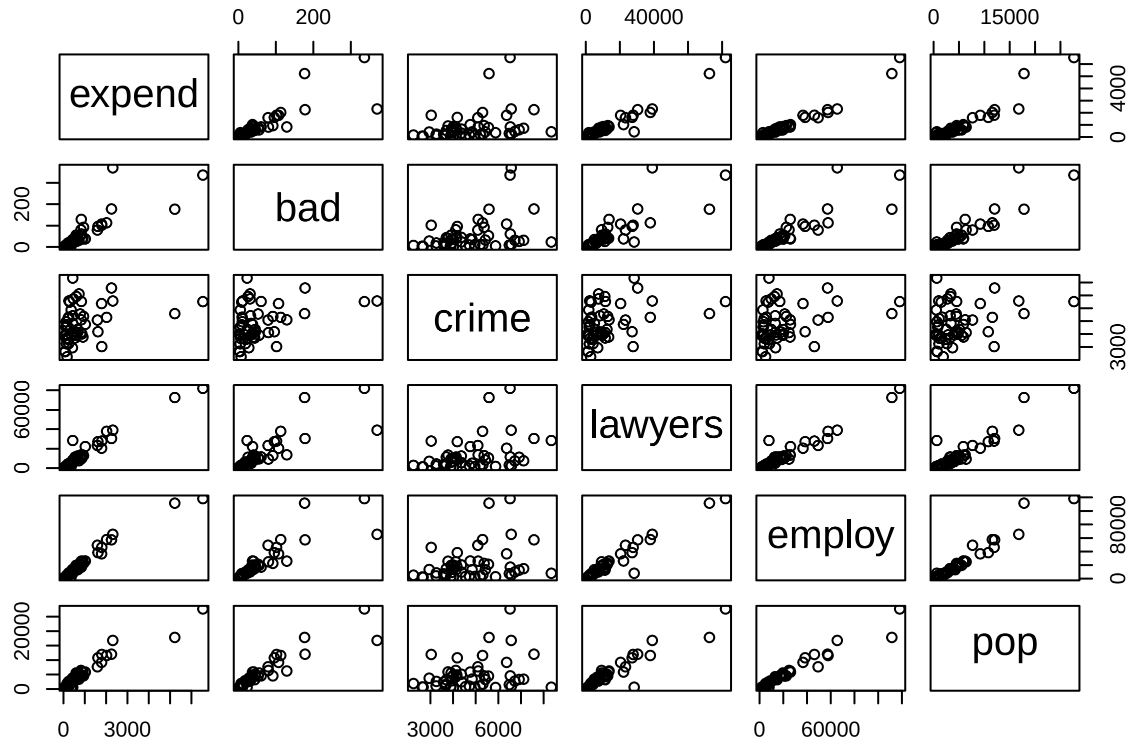
The outliers are apparent in all plots. In the residuals plot most points concentrate around zero, the points with high values appear on the right side of the residuals plot, which means they are likely due to a leverage point. To test if these points are influence points the Cook's distance is calculated:

$D_i = \frac{1}{(p+1)\hat{\sigma}^2} \sum_{j=1}^{n} (\hat{Y}_{(i),j} - \hat{Y}_j)^2.$

In words: the Cook's distance quantifies the influence of observation i on the prediction by calculating the sum of squared differences between the predicted values of the model with and without the i-th data point. In the last plot of every row the Cook's distance was plotted for all data points. From these plots we can conclude that all explanatory variables contain influence points except crime.

Collinearity

To investigate the problem of collinearity pairwise plots are made and a correlation table is produced.



```
round(cor(data),2)
```

```
##          expend  bad crime lawyers employ  pop
## expend     1.00 0.83  0.33    0.97   0.98 0.95
## bad        0.83 1.00  0.37    0.83   0.87 0.92
## crime      0.33 0.37  1.00    0.38   0.31 0.28
## lawyers    0.97 0.83  0.38    1.00   0.97 0.93
## employ     0.98 0.87  0.31    0.97   1.00 0.97
## pop        0.95 0.92  0.28    0.93   0.97 1.00
```

A lot of pairwise scatter plots show a linear pattern, which means these variables are probably collinear. Specifically, all explanatory variables are collinear except crime, which is clear from both the plot (data points scattered) and the correlation table ($R^2 < 0.80$).

This approach was only capable of finding pairwise collinearities. To find multi-collinearity the Variance Inflation Factor (VIF_j) is calculated according to the following formula:

$VIF_j = \frac{1}{1-R_j^2}$,

where R_j^2 the determination coefficient R^2 from the regression of the j-th explanatory X_j (as response) variable on the remaining explanatory variables.

```r
model <- lm(expend~bad+crime+lawyers+employ+pop,data=data)
vif(model)
```

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

All values except crime are bigger than 5 so there is a collinearity problem, which was already clear from the plots and the table.

b)

We apply the step-up method to fit a linear regression model to the data. The first step is to find the variable that yields the maximum increase in R^2.

```r
model <- lm(expend~employ,data=data)
summary(model)[[4]];summary(model)[[8]]
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -116.7052   47.06076   -2.48 1.66e-02
## employ         0.0468    0.00147   31.87 2.03e-34
```

```
## [1] 0.954
```

In this case the variable that yields the highest increase in R^2 is employ. Employ is significant so we add it to the model and repeat the process with the starting model `lm(expend~employ)`.

```r
model <- lm(expend~employ+lawyers,data=data)
summary(model)[[4]];summary(model)[[8]]
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -110.6588   42.56735   -2.60 1.24e-02
## employ         0.0297    0.00511    5.81 4.89e-07
## lawyers        0.0269    0.00776    3.46 1.13e-03
```

```
## [1] 0.963
```

Now the variable which leads to the highest increase in R^2 is lawyers. This variable is also significant so we add it to the model. Again the process is repeated, this time with starting model `lm(expend~employ+lawyers)`.

```r
model <- lm(expend~employ+lawyers+bad,data=data)
summary(model)[[4]];summary(model)[[8]]
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -110.5809   42.60733  -2.595 1.26e-02
## employ         0.0323    0.00580   5.569 1.20e-06
## lawyers        0.0263    0.00779   3.379 1.47e-03
## bad           -0.8627    0.90425  -0.954 3.45e-01
```

```
## [1] 0.964
```

```
model <- lm(expend~employ+lawyers+pop,data=data)
summary(model)[[4]];summary(model)[[8]]
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -123.3473   45.22265   -2.73  0.00894
## employ         0.0249    0.00764    3.26  0.00209
## lawyers        0.0272    0.00779    3.49  0.00105
## pop            0.0225    0.02642    0.85  0.39939
```

```
## [1] 0.964
```

Both bad and pop lead to an R^2 of 0.964, however pop is not significant while bad is. Because the change in R^2 is really small and a model with fewer explanatory variables is commonly better looked upon the choice whether or not to add these variables is up for debate. We decided not to add one of these variable as it is not essential in explaining the expenditure. So, we stop the process and define `lm(expend~employ+lawyers)` as the model. Looking back to **a** it is clear that this model will have a problem with collinearity. Also the change in R^2 after the first iteration of the algorithm is negligible. We conclude that this is not a good model because it contains redundant information and it contains variables that do not contribute significantly in added R^2.

c)

To find the 95% prediction interval for the expend of the hypothetical state, we perform the following r code:

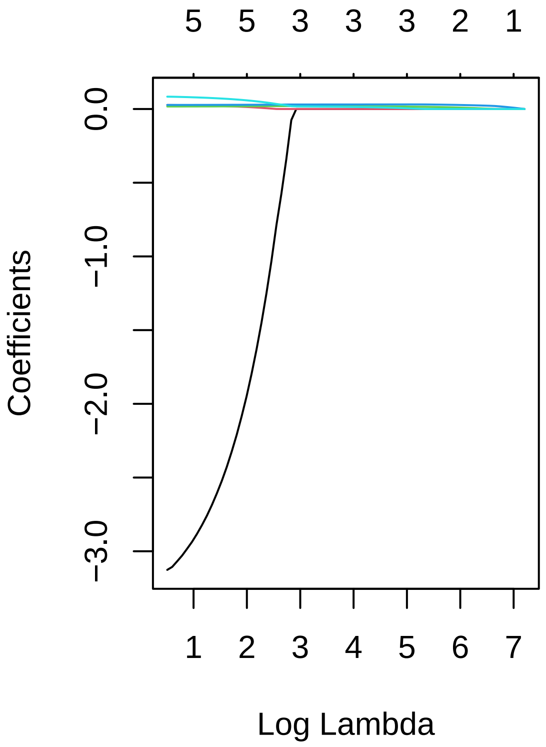
```
model <- lm(expend~employ+lawyers, data=data)
newxdata=data.frame(bad=50,crime=5000,lawyers=5000,employ=5000,pop=5000)
predict(model,newxdata,interval='prediction')
```

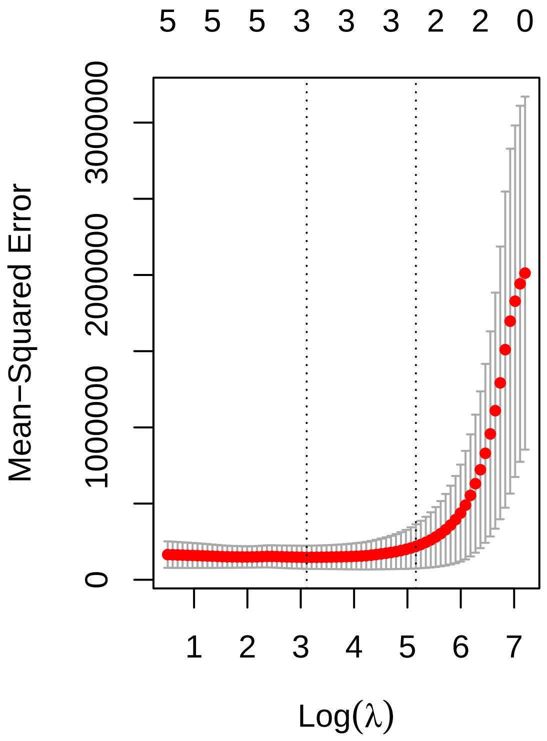
```
##   fit  lwr upr
## 1 172 -303 647
```

The interval is so big the lower confidence bound is negative, which is hard to interpret as expenditure is always a non-negative number. To improve this prediction interval we could lower the level or gather more data. However lowering the level would lead to a less weighty conclusion and gathering more data is not possible because there are a limited number of states. It is possible to find a better model with another method. If this model would explain more of the variance and have less errors the prediction interval would be smaller. Also, if a model is found with less regressors the degrees of freedom in the t-statistic would go up, that would lead to a smaller t-distribution and by result to a smaller prediction interval (more data points would have the same effect).

d)

We apply the lasso method to find a model. First the data is divided in a train and test set (2:1). Then the model is found by minimizing an equation with the squared error between the train set and the fitted model subject to β. Also a penalty term is added $(\sum_{k=0}^{p} \beta_k)$, which increases linearly with a factor of λ. A plot of the β-weights against the parameter λ is generated. Also, a plot is generated with the mean standard error of the model for every λ.





Contrasting to what one might expect the λ is chosen with the lowest mean standard error plus one standard deviation to penalize the β-weights even more and create an even simpler model. The lasso method with this λ yield the following β-weights:

```
## 6 x 1 sparse Matrix of class "dgCMatrix"
##                   s1
## (Intercept) 23.25933
## bad          .
## crime        .
## lawyers      0.01565
## employ       0.03122
## pop          0.00242
```

If the model with these β-weights is applied to predict the test set of the data the mean standard error equals:

```
## [1] 18362
```

Calculating the same statistic for the model from **b** yields:

```
## [1] 1458096
```

The most important variables are included in both models (lawyers and employ) and the β-weights do not substantially differ from one another (≈ 0.01). However, the mean squared error is significantly lower for the lasso model (up to tenfold). In conclusion, the lasso model is better able to find a good model although it does not substantially differ from the step-up model.

Exercise 3. Titanic

a)

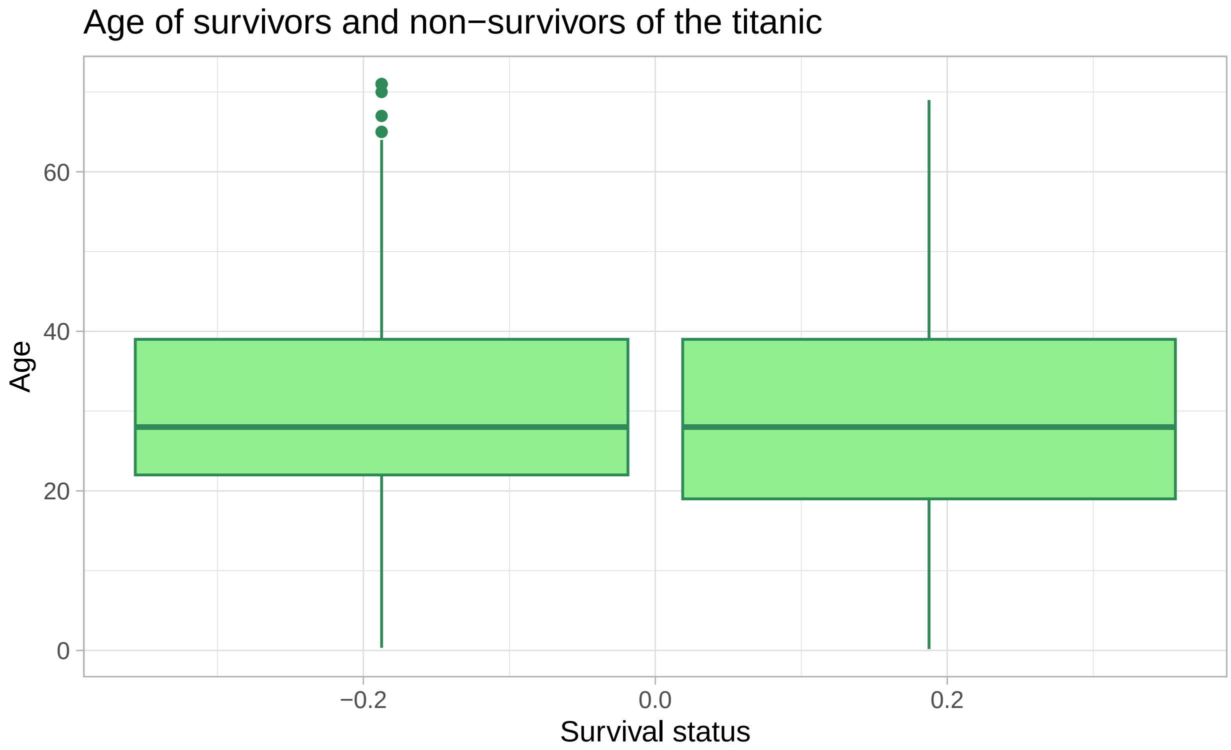
```
data_titanic <- read.table('titanic.txt', header=TRUE)
data_titanic$Sex <- as.factor(data_titanic$Sex)
data_titanic$PClass <-as.factor(data_titanic$PClass)
```

The tables below show the percentage of passengers who survived per class and sex, per sex and per class. These tables show that both females, and first class passengers have a higher survival rate than males and second or third class passengers respectively. Both first and second class females have a particularly high survival rate. Because of these differences, Sex and PClass are expected to be important variables in a model that predicts survival. The boxplot shows the distribution of ages among survivors and non-survivors. These distributions do not appear very different, therefore Age is not not expected to be an important variable for a model that predicts survival.

```
##         PClass
## Sex       1st  2nd  3rd
##   female 0.94 0.88 0.38
##   male   0.33 0.14 0.12
```

```
## Sex
## female   male
##   0.67   0.17
```

```
## PClass
##  1st  2nd  3rd
## 0.60 0.42 0.19
```



The linear regression model below shows that the odds for survival can be predicted approximately as exp{3.76 - 1.29SecondClass - 2.52ThirdClass - 2.63Sexmale - 0.04Age}. This means that the odds of survival are lower for second and third class passengers, male passengers. This is in line with the expectations from the tables. It also means the odds of survival decrease with age, which is an unexpected result. An example of odds predicted by this model would be the survival odds for a for a 30 year old second class male passenger, which are approximately 0.26.

```
lg_titanic <- glm(Survived ~ PClass + Sex + Age,
                    data=data_titanic,family=binomial)
summary(lg_titanic)$coefficients[,c("Estimate","Pr(>|z|)")]
```

```
##              Estimate Pr(>|z|)
## (Intercept)    3.7597 3.18e-21
## PClass2nd     -1.2920 6.78e-07
## PClass3rd     -2.5214 7.95e-20
## Sexmale       -2.6314 5.68e-39
## Age           -0.0392 2.69e-07
```

```
exponent <- 3.759662 -1.291962*1 -2.521419*0  -2.631357*1 -0.039177*30
exp(exponent) #survival odds 30yo 2nd class male
```

```
## [1] 0.262
```

b)

The models below show that there is a significant interaction effect between Age and Sex, but not between Age and PClass. Previous models showed that Age, PClass and Sex all have significant main effects. Therefore the resulting model includes PClass and an interaction between age and sex.

```
##                Estimate Pr(>|z|)
## (Intercept)     1.92298 1.04e-05
## Age            -0.03584 3.18e-04
## PClass2nd      -0.74428 1.93e-01
## PClass3rd      -2.29007 2.27e-05
## Age:PClass2nd -0.01321 4.05e-01
## Age:PClass3rd  0.00464 7.71e-01
```

```
##              Estimate Pr(>|z|)
## (Intercept)    0.3011 3.14e-01
## Age            0.0294 3.58e-03
## Sexmale       -0.5999 1.42e-01
## Age:Sexmale   -0.0657 1.57e-06
```

```
## Single term deletions
##
## Model:
## Survived ~ Age * PClass
##            Df Deviance AIC  LRT Pr(>Chi)
## <none>             909 921
## Age:PClass  2      910 918 1.17     0.56
```

```
## Single term deletions
##
## Model:
## Survived ~ Age * Sex
##         Df Deviance AIC LRT Pr(>Chi)
## <none>          771 779
## Age:Sex  1      796 802  25  5.6e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##              Estimate Pr(>|z|)
## (Intercept)   2.75656 3.00e-10
## PClass2nd    -1.54337 7.83e-08
## PClass3rd    -2.65398 8.47e-20
## Age           0.00244 8.30e-01
## Sexmale      -0.50819 2.51e-01
## Age:Sexmale -0.07559 4.74e-07
```

The table below shows the estimates of the survival probability for a 55 year passenger of each sex in every class.

```
##      Sex PClass Age Predictions
## 1 female    1st  55       2.891
## 2 female    2nd  55       1.348
## 3 female    3rd  55       0.237
## 4   male    1st  55      -1.775
## 5   male    2nd  55      -3.318
## 6   male    3rd  55      -4.429
```

c)

To predict the survival status of passengers, the logistic regression model from b) can be used to predict the estimated chance of survival for passengers with certain characteristics (age, sex and class). These estimates can then be turned into survival status predictions, by taking a survival chance of <0.5 to be 'not survived' and a survival chance of >0.5 as 'survived'. To test the quality of these predictions, k-cross validation can be used to test the accuracy of the predictions.

d)

The contingency tables are based on values from the tables created in a). The number of people who did not survive are calculated with `taba-tabb` and `tabc-tabd`. To test the effect of the factor class on survival, a chi-square test is performed on a contingency table of all the first, second and third class passengers who did and did not survive the titanic. This test is significant, which indicates that class has a significant effect on how the survivors/non-survivors are distributed over the contingency table. the table below shows the difference between the observed and expected values of the class contingency table. In first and second class more people survived than expected if class did not have an effect, and in third class more people died than expected if class did not have an effect. Although this was expected for third class, the results for second class are more surprising because in the logistic regression model, second class was associated with a decrease in survival chance. This difference has likely arisen because in the logistic regression model, the factor class is treatment parametrized with first class as baseline. Compared to first class, second class passengers do have a lower survival chance, which is apparent in the logistic regression model. However, second class passengers still have a higher survival rate than the average survival rate, which is why more second class passengers survived than would be expected if class did not matter to survival. This becomes apparent in the contingency table chisquare test.

To test the effect of the factor sex, a fisher test is used instead, because this is a 2 by 2 contingency table for which an exact p-value can be computed. The test is significant, meaning that sex matters to survival rate. The odds ratio result can be interpeted as for every surviving male there are approximately ten surviving women.

```
##
##  Pearson's Chi-squared test
##
## data:  class_matrix
## X-squared = 172, df = 2, p-value <2e-16
```

```
##          1st 2nd  3rd
## Survived  83  23 -106
## Died     -83 -23  106
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  sex_matrix
## p-value <2e-16
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##   7.6 13.1
## sample estimates:
## odds ratio
##       9.97
```

e)

The approach in d) is not wrong. The results a contingency table approach more easily interpretable. However, this approach does not take into account the influence of different variables and their potential interaction effects on survival rate. If only the variables survival and sex or survival and class were available in the data, the contingency table approach would be suitable, but because more data is available, it is better to use this for more accurate picture of what did and did not affect the survival of titanic passengers.

Excercise 4. Military coups

a)

We perform a poisson regression on the full dataset to create a model for the number of military coups. To find out what variables are significantly contributing to this model we compare a model without this variable to the full model. The output of this analysis is given below.

```
data$pollib= as.factor(data$pollib)
 model=glm(miltcoup~oligarchy+pollib+parties+pctvote+popn+size+numelec+numregim,family=poisson,data=data
drop1(model, test = "Chisq")[5]
```

```
##           Pr(>Chi)
## <none>
## oligarchy    0.043 *
## pollib       0.026 *
## parties      0.008 **
## pctvote      0.128
## popn         0.125
## size         0.320
## numelec      0.670
## numregim     0.368
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Oligarchy, pollib and parties have a significant effect on the number of military coups. To what extent these variables influence the response variable is given below.

```
summary(model)$coefficients[,c("Estimate","Pr(>|z|)")]
```

```
##               Estimate Pr(>|z|)
## (Intercept) -0.233427  0.81500
## oligarchy    0.072566  0.04007
## pollib1     -1.103244  0.09252
## pollib2     -1.690306  0.01249
## parties      0.031221  0.00517
## pctvote      0.015441  0.12641
## popn         0.010959  0.12531
## size        -0.000265  0.32444
## numelec     -0.029619  0.67054
## numregim     0.210943  0.36720
```

From this table we can conclude that the number of years the country was ruled by a military oligarchy and the number of legal political parties is associated with an increase in the number of military coups. Also, the political liberalization is inversely related to the number of military coups.

b)

We use the step-down approach to reduce the number of explanatory variables in the model. We removed variables with the lowest p-value sequentially until all p-values were lower than $\alpha = 0.05$.

```
summary(model)$coefficients[, "Pr(>|z|)"]
```

```
## (Intercept)   oligarchy     pollib1     pollib2     parties     pctvote
##     0.81500     0.04007     0.09252     0.01249     0.00517     0.12641
##        popn        size     numelec    numregim
##     0.12531     0.32444     0.67054     0.36720
```

We remove numelec.

```
model <- glm(miltcoup~oligarchy+pollib+parties+pctvote+popn+size+numregim,family=poisson,data=data)
summary(model)$coefficients[, "Pr(>|z|)"]
```

```
## (Intercept)   oligarchy     pollib1     pollib2     parties     pctvote
##     0.59464     0.00483     0.08625     0.00404     0.00443     0.14164
##        popn        size    numregim
##     0.13862     0.31710     0.42075
```

We remove numregim.

```
model <- glm(miltcoup~oligarchy+pollib+parties+pctvote+popn+size,family=poisson,data=data)
summary(model)$coefficients[, "Pr(>|z|)"]
```

```
## (Intercept)   oligarchy     pollib1     pollib2     parties     pctvote
##    0.942048    0.000936    0.084412    0.003412    0.004796    0.111513
##        popn        size
##    0.207510    0.332621
```

We remove size.

```
model <- glm(miltcoup~oligarchy+pollib+parties+pctvote+popn,family=poisson,data=data)
summary(model)$coefficients[, "Pr(>|z|)"]
```

```
## (Intercept)   oligarchy     pollib1     pollib2     parties     pctvote
##     0.66168     0.00123     0.16799     0.00707     0.00388     0.13728
##        popn
##     0.30204
```

We remove popn.

```
model <- glm(miltcoup~oligarchy+pollib+parties+pctvote,family=poisson,data=data)
summary(model)$coefficients[, "Pr(>|z|)"]
```

```
## (Intercept)   oligarchy     pollib1     pollib2     parties     pctvote
##    0.820609    0.000044    0.202919    0.007371    0.007036    0.183834
```

The highest p-value is for pollib1, however this is one level of the entire factor pollib which is significant. We remove the variable with the next biggest p-value, pctvote.

```
model <- glm(miltcoup~oligarchy+pollib+parties,family=poisson,data=data)
summary(model)$coefficients[,c("Estimate","Pr(>|z|)")]
```

```
##             Estimate Pr(>|z|)
## (Intercept)   0.2080 6.41e-01
## oligarchy     0.0915 5.04e-05
## pollib1      -0.4954 2.98e-01
## pollib2      -1.1121 1.55e-02
## parties       0.0224 1.40e-02
```

We stop iterating because there are no more insignificant variables in the model. The remaining variables are the same variables which were significant in the original full model. However this model is better since it contains less variables. The p-value for oligarchy is smaller in comparison to the p-value in the full model, so there is even more certainty about the influence of oligarchy on the number of military coups.

c)

To calculate the predicted number of military coups of an hypothetical average country seperated for political liberalization we perform the following r code.

```
countries <- data.frame(oligarchy=mean(data$oligarchy),pollib=as.factor(c(0,1,2)),parties=mean(data$par
predict(model,countries,interval="prediction")
```

```
##       1       2       3
##  1.0676  0.5722 -0.0445
```

This result tells us that an average country with no civil rights for political expression is expected to have approximately one successful military coup between independence and 1989. From this prediction a trend is visible with the number of coups decreasing as the political liberties increase. For the country with full political liberties the number of succesful coups is negative. This is hard to interpret but in practice this would come down to no coups.