Assignment 2 - Group 27

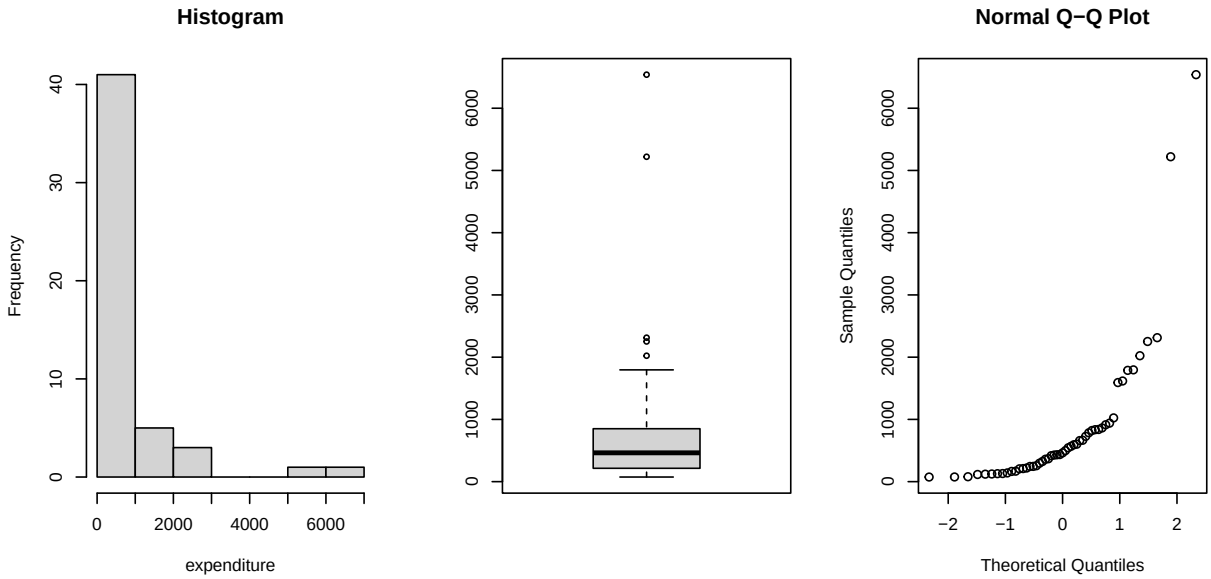
Joost Driessen, Emma van Lipzig, Rohan Zonneveld

2023-03-04

Excercise 1. Trees

Excercise 2. Expenditure on criminal activities

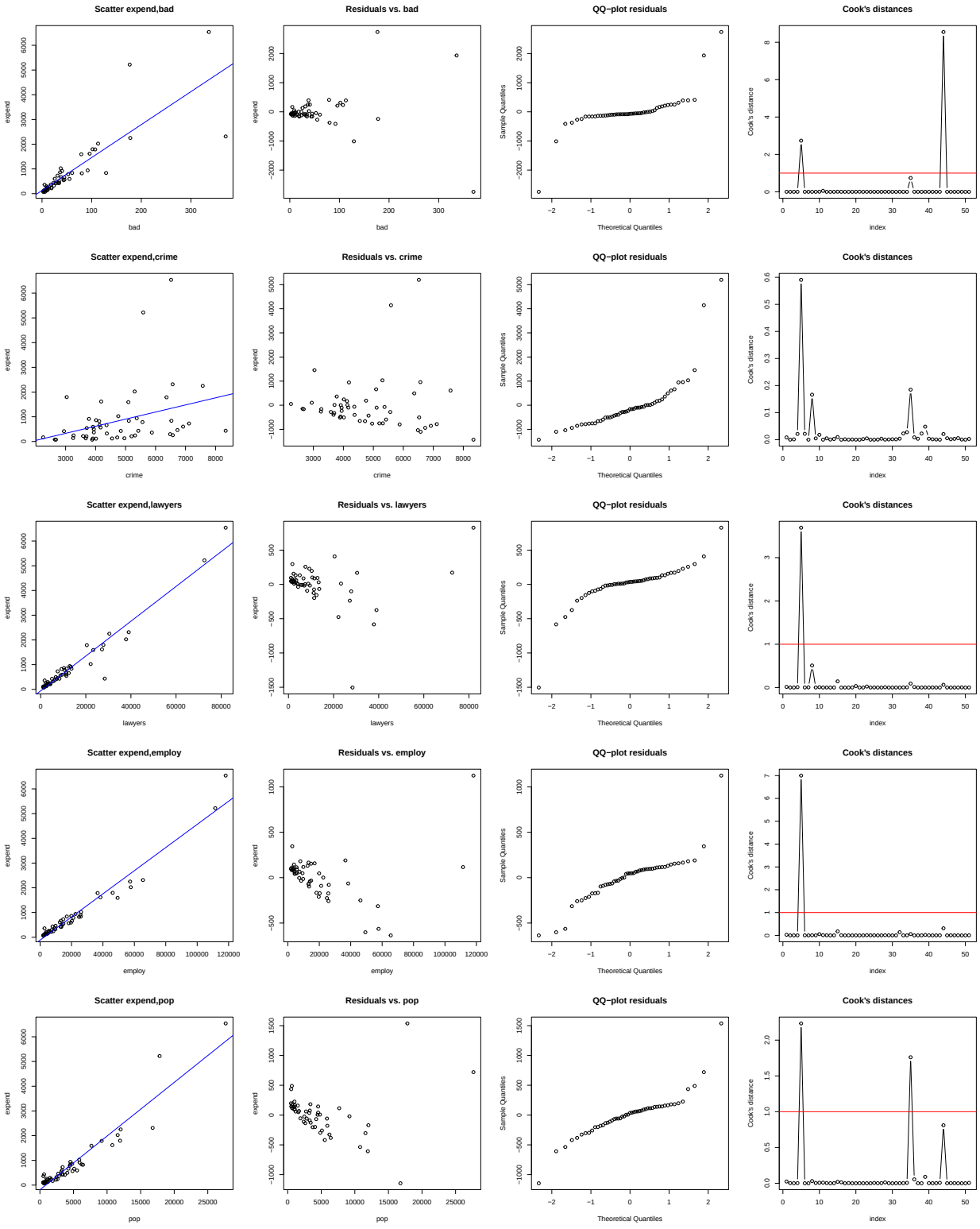
a



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.



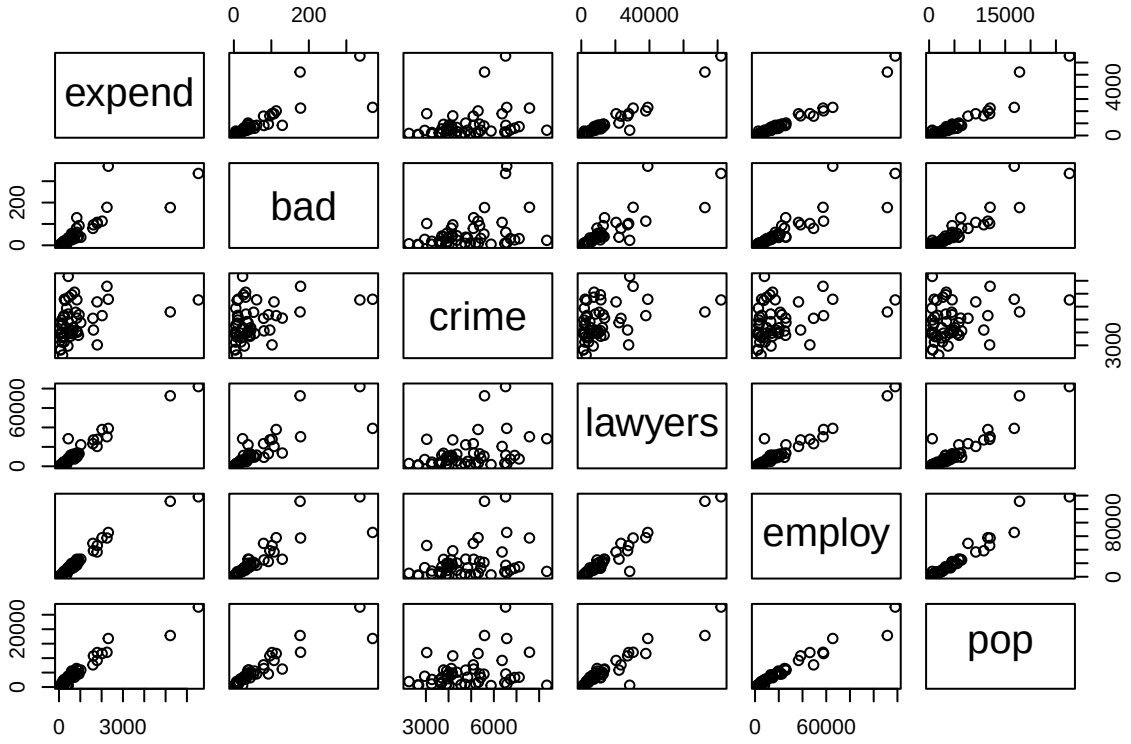
The outliers are apparent in all plots. In the residuals plot most points concentrate around zero, the points with high values appear on the right side of the residuals plot, which means they are likely due to a leverage point. To test if these points are influence points the Cook's distance is calculated:

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

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.

```
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.

```
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)`.

```
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)`.

```
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 this variables is up for debate. We decided not to add this 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

Excercise 3. Titanic

Excercise 4. Military coups