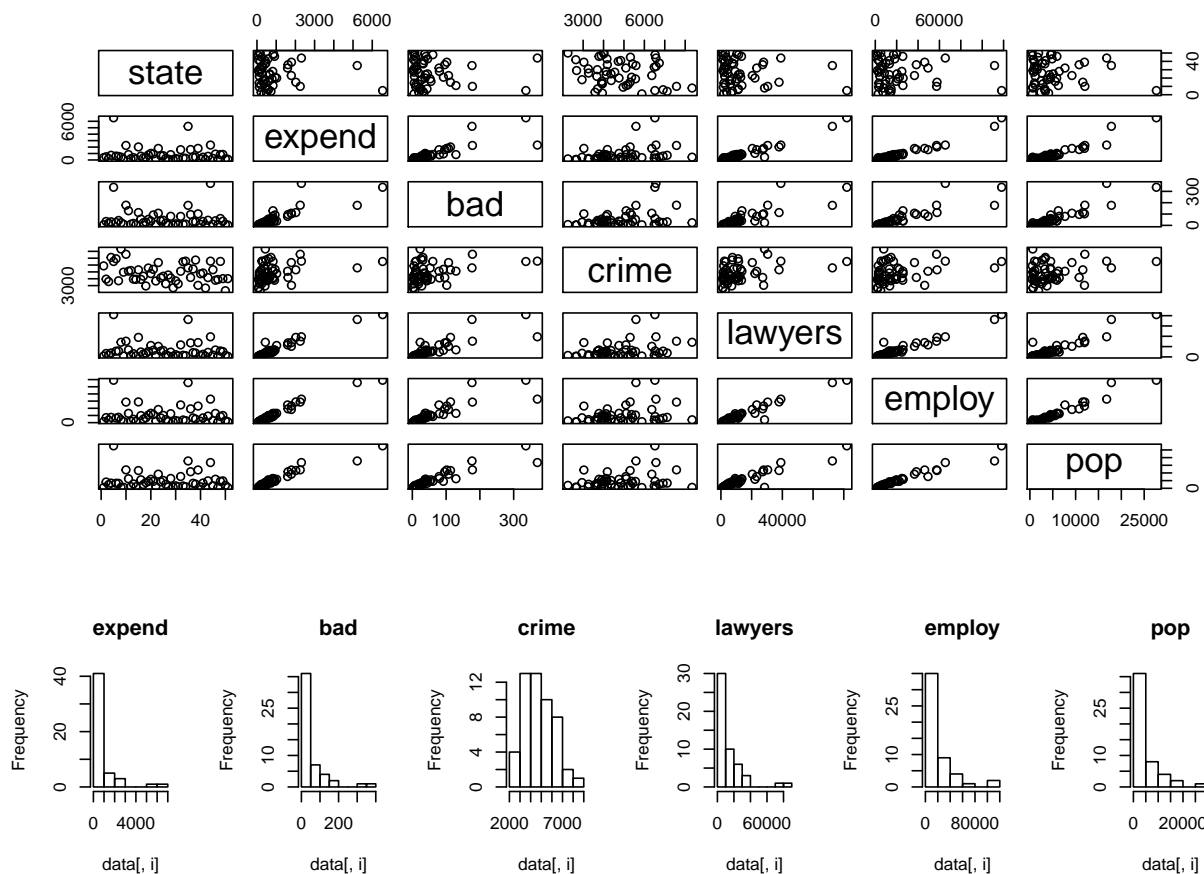


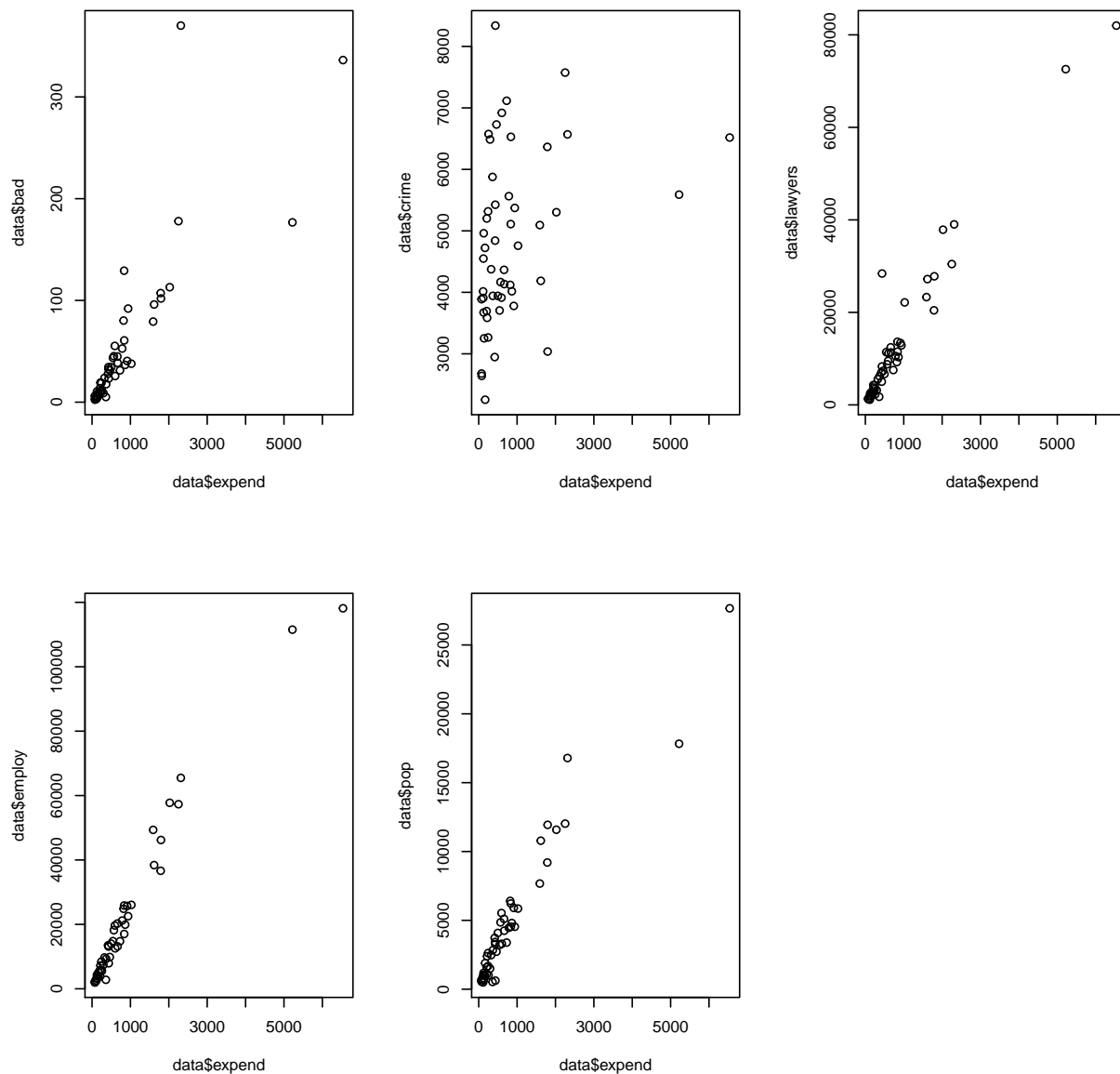
Exercise 5

In our regression analysis, the response variable is “expend” and the explanatory variables are: “bad, crime, lawyers, employ and pop”. The purpose is to explain expend by a numerical function of the explanatory variables.

a) First, we make a graphical summary of the data by plotting each variable against the others. Furthermore, we construct a histogram of all the numerical variables. Looking at the plots, we observe that expend, lawyers, employ and pop all approximate a linear relationship with each other. Furthermore, state and crime have nonlinear relationships with all the other variables. Lastly, the variable bad can be argued to have a weak linear relationship with the variables expend, lawyers, employ and pop. Looking at the histograms, it is interesting to see that almost all variables (expend, bad, lawyers, employ and pop) follow a similar pattern, namely, the lowest value appears frequently and as the value increases, the frequency decreases steeply. Except for a few outliers of frequently occurring high values. In contrast, the variable crime shows a different pattern. Namely, the values in the middle occur also relatively frequently. But the rule: as the value increases, the frequency decreases, applies as well.



To build an intuition of the linear relationship between the dependable variable (expend) and its explanatory variable, we have plotted the explanatory variables against the dependable variable below. These are the same plots as from the first plot with all factors plotted against each other, but now we can take a closer look. From the plots we can see a strong linear relationship between the dependable expend and 4 of the five factors: bad, lawyers, employ and pop. However we see very big outliers, which will have to be adjusted for, otherwise they will have a very big weight in the factor coefficient from the linear regression.



A potential point is an outlier in an explanatory variable. The effect can be studied by fitting the model with and without the potential point. If the estimated parameters change drastically when removing the potential point, the observation is called an influence point. Using the Cook's formula, the distance of an observation on the predictions can be calculated. Whenever the Cook's distance for an observation approximates or is larger than 1, the observation can be considered to be an influence point. As we have not constructed a model yet, we analyse the potential and influence points of our chosen model in c). Another relevant concept is collinearity. This is the problem of linear relations between explanatory variables. Collinearity can be detected by a straight line in a scatter plot or by calculating the correlation coefficient. Looking at the scatter plots of the data, we suspect collinearity between the variables `expend`, `lawyers`, `employ` and `pop`. We confirm this by calculating the correlation coefficients of all possible variable combinations. Looking at the output below, we observe that all the combinations of the variables `expend`, `lawyers`, `employ` and `pop` have a correlation coefficient above 0.93. Thus, we conclude that these variables have a collinear relation. The variable `bad` has a weaker collinear relation with the variables `expend`, `lawyers`, `employ` and `pop`, namely, ranging from 0.83 to 0.93. Lastly, the variable `crime` has no collinear relation with

any of the other variables. When collinearity is detected among variables, we should avoid having both explanatory variables in the model.

```
##      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
```

b) To fit a linear regression model to the data, first, we start with the step-up method. Using this method, we start by fitting all possible simple linear regression models and calculate the determination coefficient (R^2). The results are shown in the table below. Looking at this table, we observe that employ has the largest value of R^2 (0.954) and is thus selected. Next, we combine this variable with all the variables that do not have a collinear relation with employ. These are the explanatory variables bad and crime. Note that the variable bad still can be considered to be linearly correlated to employ. Adding the variables bad and crime to the models yields in $R^2 = 0.9551$. This is an improvement compared to the previous model. Therefore, we continue to add the other variables to the model. For both models, there is just one variable to add. This results in the last possible option: a model of employ, bad and crime combined. This result in $R^2 = 0.9568$, the highest value so far. As there are no more variables to add, the method stops here. The resulting model is: $\text{expend} = -2.857e^{+02} + 4.979e^{-02} * \text{employ} - 1.391e^{+00} * \text{bad} + 3.810e^{-02} * \text{crime} + \text{error}$. We have to be careful as it could be argued that the variables employ and bad are collinear. Therefore, the model that is constructed with the variables employ and crime ($\text{expend} = -2.484e^{+02} + 4.630e^{-02} * \text{employ} + 2.962e^{-02} * \text{crime} + \text{error}$) might be a better model as it contains fewer variables and the value of R^2 is similar.

| Explanatory Variable(s) | bad | crime | lawyers | employ | pop | bad, employ | crime, employ | bad, crime, employ |
|-------------------------|--------|--------|---------|--------|--------|-------------|---------------|--------------------|
| Multiple R-squared | 0.6964 | 0.1119 | 0.9373 | 0.954 | 0.9073 | 0.9551 | 0.9551 | 0.9568 |

Second, we use the step-down method. This method start with fitting all explanatory variables in the so-called full model. In each iteration, one explanatory variable is removed. This time, we try the model with all variables, regardless of collinearity. In round 1, we observe that the variable crime has the highest p-value, $0.25534 > 0.05$, therefore, the variable crime will be removed. In round 2, pop has the highest p-value, $0.06012 > 0.05$, therefore, the variable pop will be removed. In round 3, bad has the highest p-value, $0.34496 > 0.05$, therefore, the variable bad will be removed. In round 4, lawyers has the highest p-value, $0.00113 < 0.05$, therefore, the variable will not be removed and the method stops. This results in the model $\text{expend} = -1.107e^{+02} + 2.686e^{-02} * \text{lawyers} + 2.971e^{-02} * \text{employ} + \text{error}$.

Round 1: $\text{expend} \sim \text{bad, crime, lawyers, employ, pop}$

| Explanatory Variables | bad | crime | lawyers | employ | pop |
|-----------------------|---------|---------|---------|---------|---------|
| p-value | 0.02719 | 0.25534 | 0.00592 | 0.00354 | 0.03184 |

Round 2: $\text{expend} \sim \text{bad, lawyers, employ, pop}$

| Explanatory Variables | bad | lawyers | employ | pop |
|-----------------------|---------|---------|---------|---------|
| p-value | 0.05402 | 0.00106 | 0.00380 | 0.06012 |

Round 3: expend ~ bad, lawyers, employ

| Explantory Variables | bad | lawyers | employ |
|----------------------|---------|---------|--------------|
| p-value | 0.34496 | 0.00147 | $1.2e^{-06}$ |

Round 4: expend ~ lawyers, employ

| Explantory Variables | lawyers | employ |
|----------------------|---------|---------------|
| p-value | 0.00113 | $4.89e^{-07}$ |

Using the step-up and step-down method resulted in two different models. The advantage of the model constructed using the step-up method (expend ~ employ bad crime) is that the variables are not collinear. The advantages of the model constructed using the step-down method (expend ~ lawyers employ) are that the value of R^2 (0.9632) is higher compared to the R^2 value of the step-up model (0.9568) and the model contains fewer explanatory variables. However, as the collinearity of variables weight higher compared to number of variables and the difference of R^2 is relatively small, we prefer the model constructed using the step-up model. We even consider removing the variable bad of the step-up model as it can be argued to have a collinear relation with the variable employ.

c) We check the model (expend ~ employ crime) assumptions (linearity of the relation and normality of the errors) using both graphical and numerical tools. First, we look at the scatter plot of the response variable against each explanatory variable separately. This is visualized in a). For each combination with expend, we see two outliers at the right of the scatter plots. Moreover, the relation with the variables bad, lawyers, employ and pop is linear and nonlinear combined with state and crime. Second, we construct the scatter plot of the residuals against each explanatory variable that is in the model (crime and employ) seperately. The plot with crime contains a cluster around the middle line and the plot with employ has a cluster in the bottom left diagonal with two outliers in the upper right corner. This means... Third, we construct the added variable plot. In this plot, the residuals of the explanatory variables are plotted against the residuals of the model without that specific variable. This shows the effect of adding an explanatory variable to the model. Looking at the figure, we observe that the plots for bad, lawyers and pop contain compact clusters. The plots of crime and employ are more spread. This means... Next, we construct the scatter plots of the residuals against each explanatory variable that is not in the model (bad, lawyers and pop) seperatly. The outputs are similar to each other. When looking at the pattern we do not observe a linear relation and therefore should not include more variables. Afterwards, we construct the scatter plot of the reduals against the response variable and the fitted model. We observe little spread as in both plots there is a cluster around zero with only few outliers. This means... Lastly, we check the normality assumption by constructing a qq-plot and conduction Shapiro-Wilk's test. We cannot assume normality as the qq-plot does not approximate a straight line, this means that the model is invalid. Unfortunately, none of the other models resulted in a normal distribution of the residuals.

