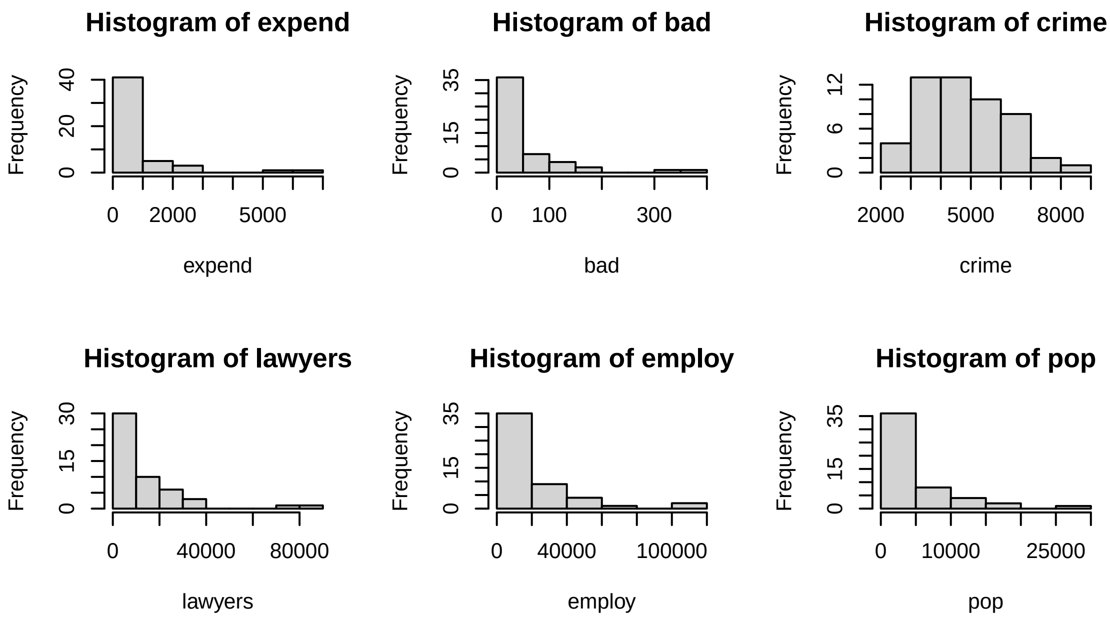
A2E5

Xinyu Hu

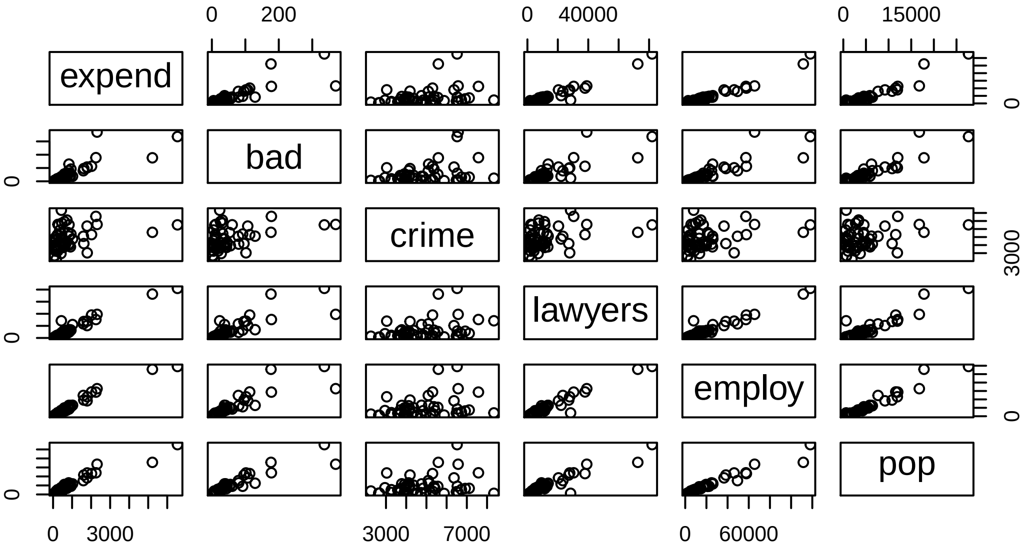
2021/3/5

Exercise 5

a)



In case of finding the potential and influence points, the histograms are shown above. It is clear that the crime factor is normally distributed, and the rest of factors have the similar curve. It shows that there exists collinearity.




```
##          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
```

In the graph, (expend, crime), (bad, crime), (crime, lawyers), (crime, employ), (crime, pop) are not linear independently. And we need to see which predictor variables are involved in collinearity.

```
exlm1 = lm(expend~bad+crime+lawyers+employ+pop, data=ex);vif(exlm1)
```

```
##       bad     crime   lawyers    employ       pop
##  8.364321  1.487978 16.967470 33.591361 32.937517
```

```
exlm2 = lm(expend~crime+lawyers+employ+pop, data=ex); vif(exlm2)
```

```
##     crime   lawyers    employ       pop
##  1.233263 16.372292 33.106158 17.576977
```

```
exlm3 = lm(expend~crime+employ+pop, data=ex); vif(exlm3)
```

```
##     crime    employ       pop
##  1.121163 17.967808 17.568906
```

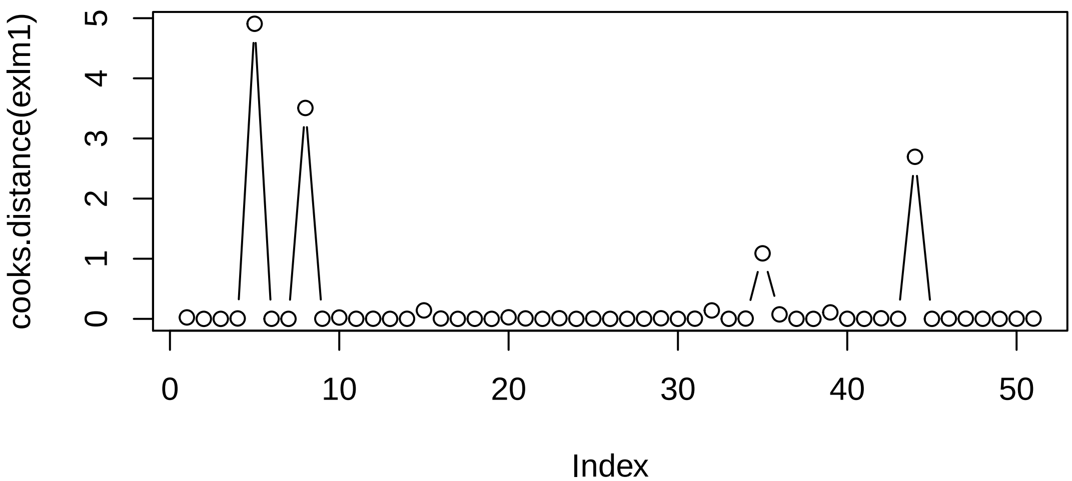
```
exlm4 = lm(expend~crime+pop, data=ex); vif(exlm4)
```

```
##   crime     pop
## 1.08213 1.08213
```

```
exlm5 = lm(expend~crime, data=ex);vif(exlm5)
# Error in vif.default(exlm5) : model contains fewer than 2 terms
```

In exlm1, exlm2 and exlm3, all VIF's are large, so there is a collinearity problem, but the exlm4 and exlm5 are OK.

```r
plot(cooks.distance(exlm1),type="b")
```




```r
round(cooks.distance(exlm1),2)
```

```
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16
## 0.02 0.00 0.00 0.01 4.91 0.00 0.00 3.51 0.00 0.02 0.00 0.00 0.00 0.00 0.14 0.01
##   17   18   19   20   21   22   23   24   25   26   27   28   29   30   31   32
## 0.00 0.00 0.00 0.03 0.01 0.00 0.01 0.00 0.00 0.00 0.00 0.00 0.01 0.00 0.00 0.14
##   33   34   35   36   37   38   39   40   41   42   43   44   45   46   47   48
## 0.00 0.00 1.09 0.07 0.00 0.00 0.11 0.00 0.00 0.01 0.00 2.70 0.00 0.00 0.00 0.00
##   49   50   51
## 0.00 0.00 0.00
```

Thus, the potential and influence points are Point(5), Point(8), Point(35) and Point(44).

b)
First, we start with step-up method.

```r
summary(lm(expend~bad,data=ex))[[8]]
```

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

```r
summary(lm(expend~crime,data=ex))[[8]]
```

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

```r
summary(lm(expend~lawyers,data=ex))[[8]]
```

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

```r
summary(lm(expend~employ,data=ex))[[8]]
```

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

```r
summary(lm(expend~pop,data=ex))[[8]]
```

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

The employ has highest value: 0.9539745.

```
summary(lm(expend~employ+bad,data=ex))[[8]]
```

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

```
summary(lm(expend~employ+crime,data=ex))[[8]]
```

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

```
summary(lm(expend~employ+pop,data=ex))[[8]]
```

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

```
summary(lm(expend~employ+lawyers,data=ex))[[8]]
```

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

The model of expend~employ+lawyers has highest value: 0.9631745.

```
summary(lm(expend~employ+lawyers+bad,data=ex))[[8]]
```

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

```
summary(lm(expend~employ+lawyers+crime,data=ex))[[8]]
```

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

```
summary(lm(expend~employ+lawyers+pop,data=ex))[[8]]
```

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

Since the models did not yield any significant results, the step-up method stopped.

```
summary(lm(expend~bad+crime+lawyers+employ+pop,data=ex))[[8]]
```

```
## [1] 0.9675314
```

```
summary(lm(expend~bad+lawyers+employ+pop,data=ex))[[8]]
```

```
## [1] 0.9665736
```

```
summary(lm(expend~bad+lawyers+employ,data=ex))[[8]]
```

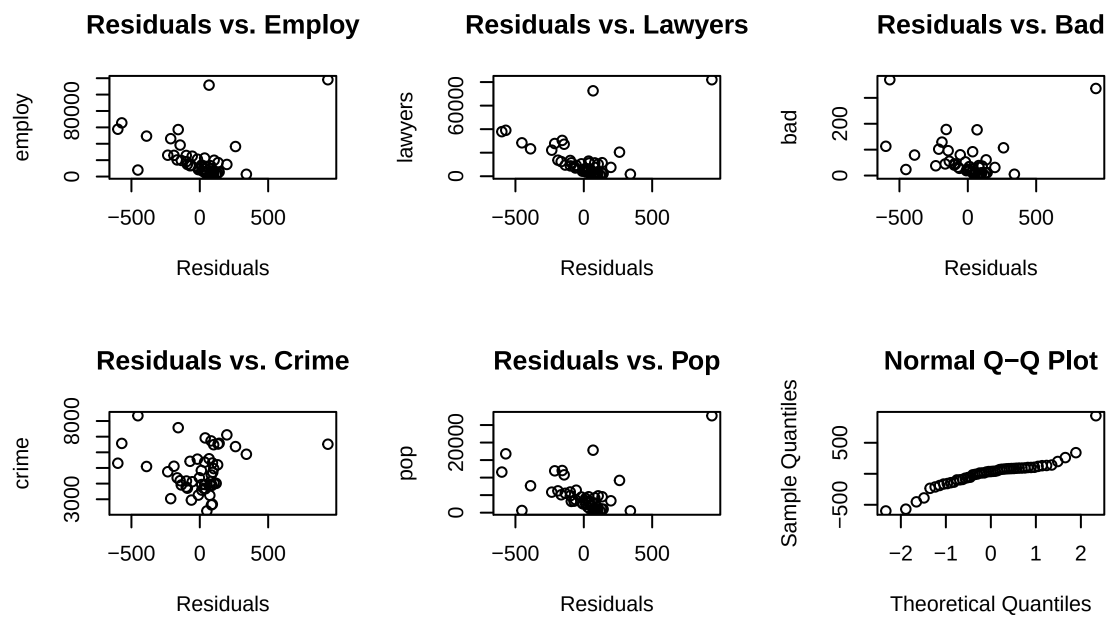
```
## [1] 0.9638741
```

```
summary(lm(expend~lawyers+employ,data=ex))[[8]]
```

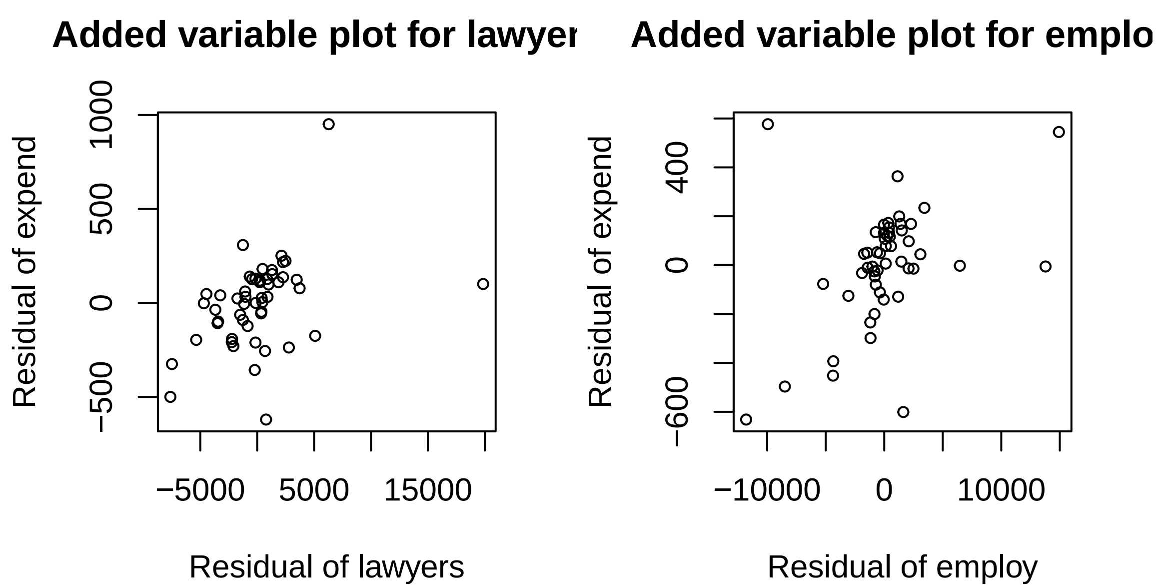
```
## [1] 0.9631745
```

All of these models did not yield any significant results, so the step-down method stopped. Hence, expend~lawyers+employ is the final model for both methods, which $expend = -110.7 + 0.002971 \times employ + 0.02686 \times lawyers$.

c)



From question(a), it already shows the collinearity of dependent and independent variables. The above graphs claims that the spread of residuals against variables did not show such a pattern existing. And the QQ-plot shows the residuals are normally distributed.



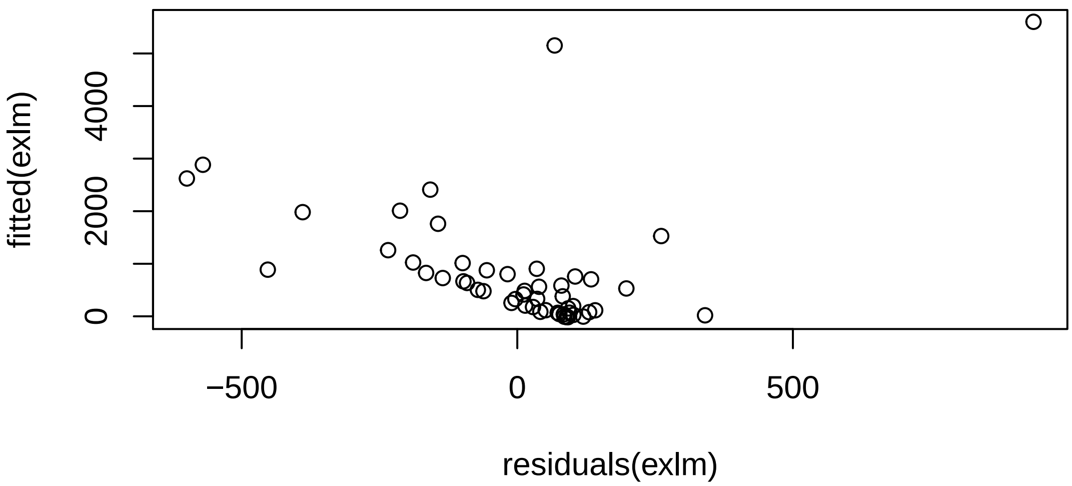
The added variable plots also show that there is no such specific curved pattern visible.

```
shapiro.test(residuals(exlm))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(exlm)
```

```
## W = 0.8475, p-value = 1.118e-05
```

The Shapiro-Wilk normality test shows the same as the QQ-plot, which means it is still normally distributed since p-value=1.118e-05 < 0.05.



Moreover, there is no patterns or errors are visible in the scatter plot of residuals against Y (and $\hat{Y}$).