assignment2

nihat uzunalioglu - 2660298, emiel kempen - 2640580, saurabh jain - 2666959

2/26/2020

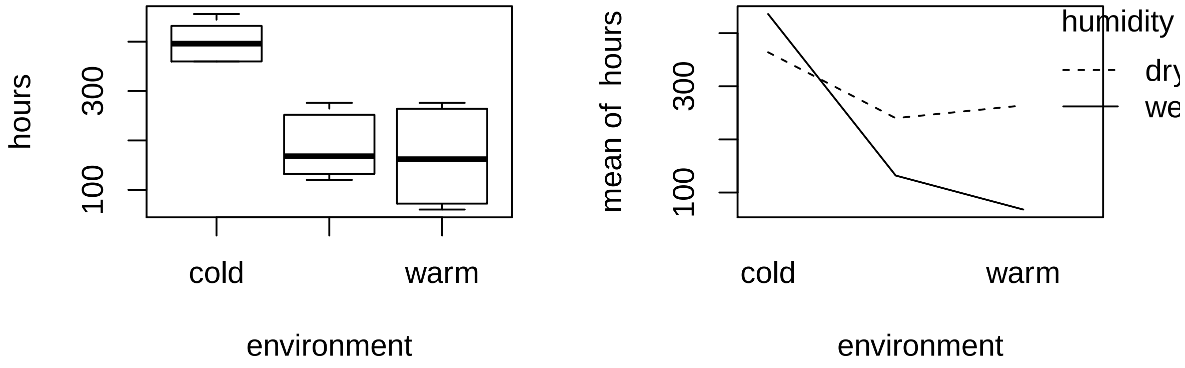
Exercise 1

a)

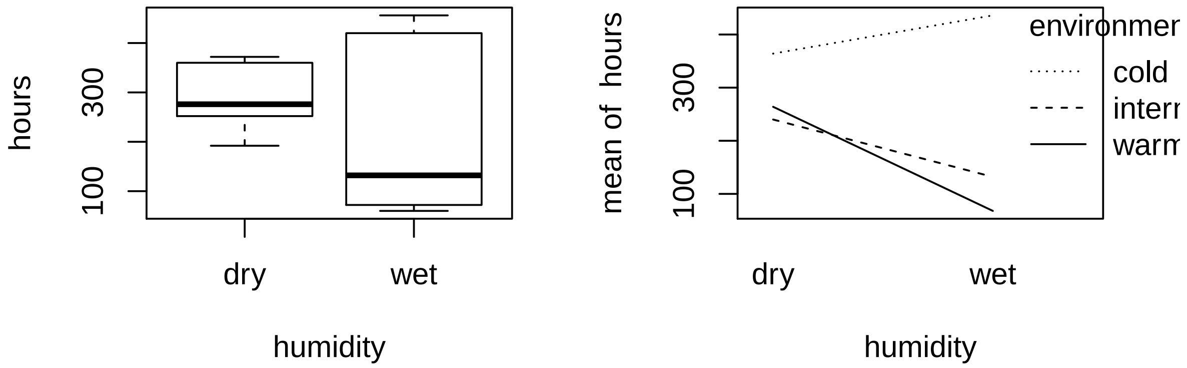
```
# The randomization process for 18 slices
N=3; I=2; J=3
rbind(rep(1:I, each = N*J), rep(1:J, N*I), sample(1:(N*I*J)))
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14]
## [1,]    1    1    1    1    1    1    1    1    1     2     2     2     2     2
## [2,]    1    2    3    1    2    3    1    2    3     1     2     3     1     2
## [3,]   13    6    7    1   16    9   18   17    5    15     4     3    11     8
##      [,15] [,16] [,17] [,18]
## [1,]     2     2     2     2
## [2,]     3     1     2     3
## [3,]    12     2    10    14
```

N: the number of units for each combination I: the levels of humidity J: the levels of environment - We performed randomization according to the levels of the dataset. This way we can randomly assign slices of bread to the different levels of the dataset.



b)



c)

```
# Creating linear model and ANOVA test
# Factorization
bread$humidity = as.factor(bread$humidity)
bread$environment = as.factor(bread$environment)
breadaov = lm(hours~environment*humidity, data = bread); anova(breadaov)
```

```
## Analysis of Variance Table
##
## Response: hours
##                      Df Sum Sq Mean Sq F value    Pr(>F)
## environment           2 201904  100952 233.685 2.461e-10
## humidity              1  26912   26912  62.296 4.316e-06
## environment:humidity  2  55984   27992  64.796 3.705e-07
## Residuals            12   5184     432
```

```
p_interaction = anova(breadaov)$Pr[3]
```

- The p-value for testing for H_0:$\gamma_{i,j} = 0$ for all i, j is 3.7054783×10^{-7}. Therefore, we reject the null hypothesis H_0 which means the interaction between environment and humidity is significant for this dataset.

d)

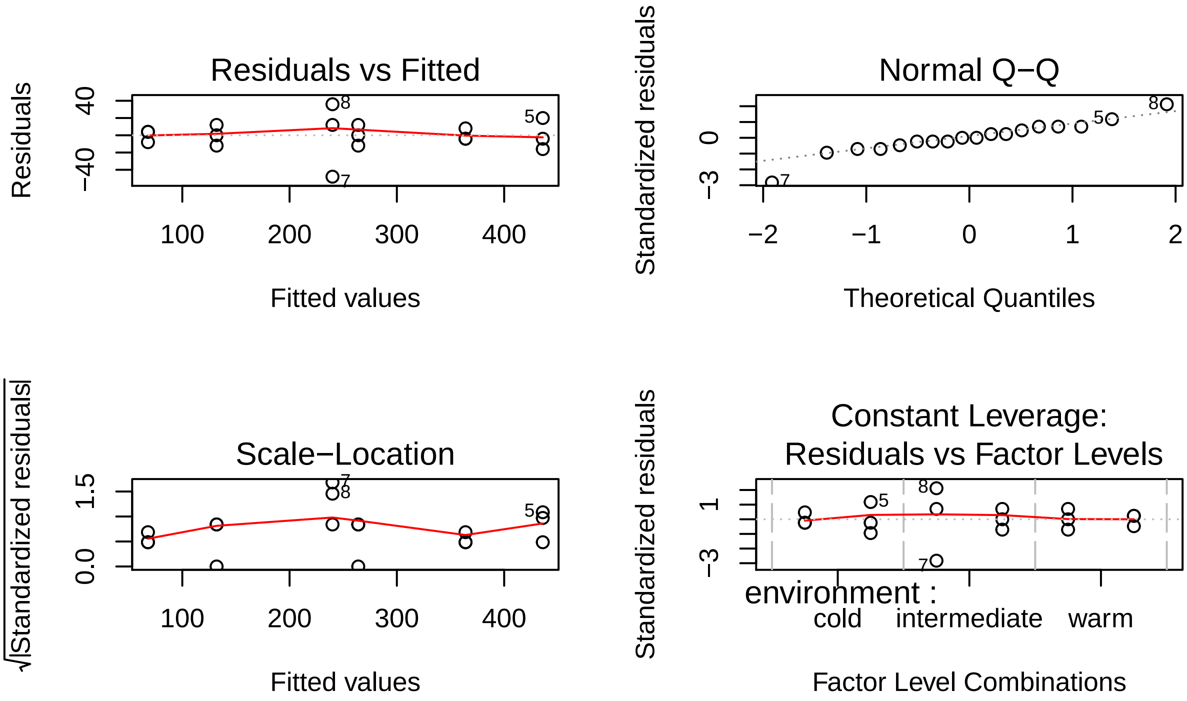
```
contrasts(bread$humidity)=contr.sum
contrasts(bread$environment)=contr.sum
breadaov2 = lm(hours~humidity*environment,data=bread)
summary(breadaov)[4]
```

```
## $coefficients
##                                     Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)                              364   12.00000 30.333333 1.032769e-12
## environmentintermediate                  -124   16.97056  -7.306770 9.389760e-06
## environmentwarm                          -100   16.97056  -5.892557 7.336887e-05
## humiditywet                                72   16.97056   4.242641 1.142103e-03
## environmentintermediate:humiditywet      -180   24.00000  -7.500000 7.233671e-06
## environmentwarm:humiditywet              -268   24.00000 -11.166667 1.073751e-07
```

- Here we can see that the environment factor affects the hours in a bigger way than humidity. We can see changes of 149.33 and −64.66 in comparison with 38.667.
- We don't think it's a good question though, because we agree that the root of impact lies in the relationship between these two variables rather than just one of them, even though it indicates bigger changes.

e)

```r
par(mfrow=c(2, 2))
# Plot the linear fitted model graphs
plot(breadaov)
```



- According to the tables we can say that 192, 2, 1 and 276, 2, 1 are the two that can be considered as outliers.

Exercise 2

```r
search = read.table("data/search.txt", header=TRUE)
```

a)

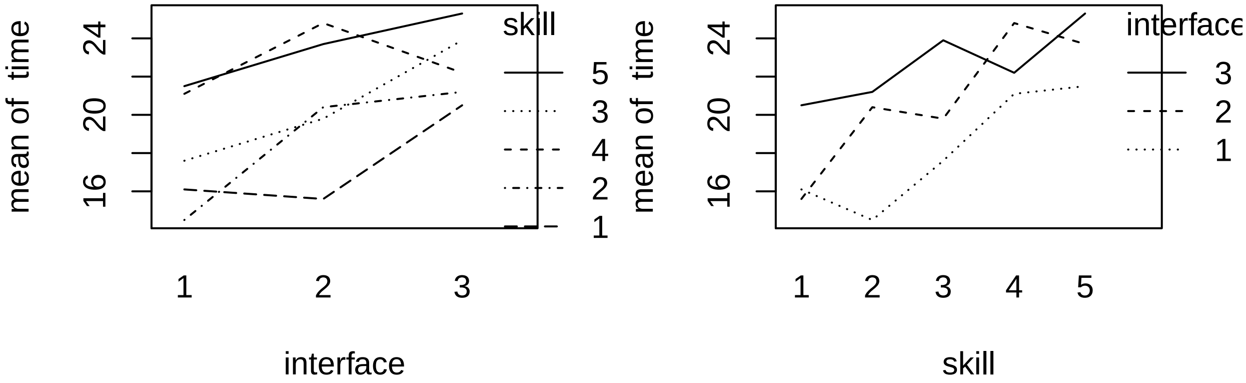
```
N = 1 #
I = 3
B = 5
for (i in 1:B){
  print(sample(1:(N*I)))
}
```

- The blocks created represent the students grouped per skill-level, so totaling to 5 blocks of 3 students each. For block 1 assign student 1 to interface 1, student 3 to interface 2, etc., for block 2 assign student 1 to interface 1, student 2 to interface 2, etc.

b)

```
attach(search)

par(mfrow=c(1,2))
interaction.plot(interface,skill,time)
interaction.plot(skill,interface,time)
```



- The pattern $(\alpha 1, \alpha 2, ..., \alpha_I)$ of treatment effects is assumed to be the same within every block. However, the lines in the seperate interaction plots do not seem to be parallel. Therefore, we can assume that there is an interaction between interface and skill.

c)

```
search$skill = as.factor(search$skill)
search$interface = as.factor(search$interface)

aovsearch=lm(time~interface+skill, data=search); anova(aovsearch)
```

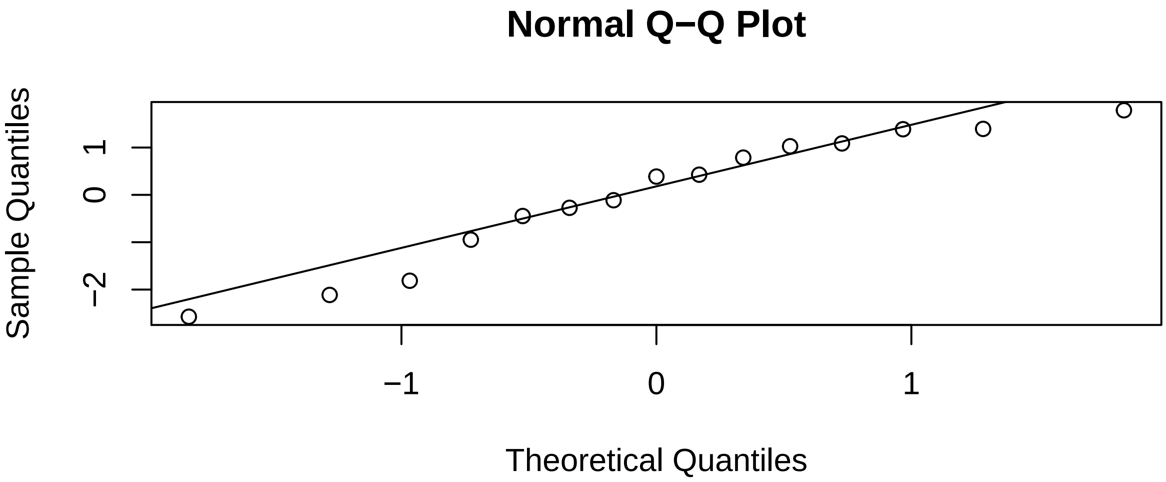
- From the ANOVA test follows a p-value for interface of 0.0130987. This indicates that we can reject that the null hypothesis H_0, that stated that the means of the search times for all interfaces is the same.

```
summary(aovsearch)[4]
```

- Data are assumed to follow the model $Y_{i,b,n} = \mu + \alpha_i + \beta_b + e_{i,b,n}$. Filling in for skill level 3 and interface 2: $Y_{2,3} = 15.013 + 2.700 + 3.033 = 20.746\ s$. This is the estimated time it takes a typical user of skill level 3 to find the product on the website if the website uses interface 2.

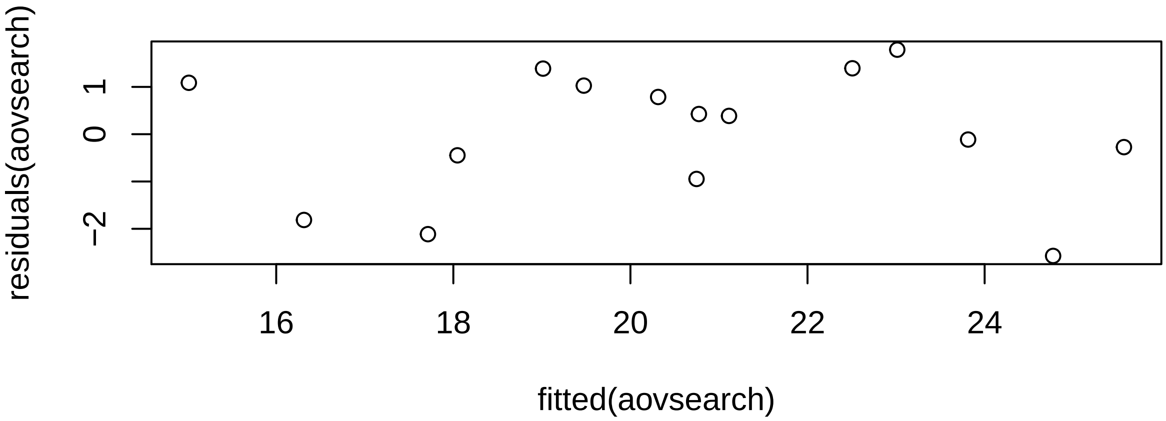
d)

```
qqnorm(residuals(aovsearch));qqline(residuals(aovsearch))
```



- The QQ-plot seems to deviate a bit from a straight line in the extremes, but the residuals can be assumed to be normally distributed.

```
plot(fitted(aovsearch),residuals(aovsearch))
```



- The scatter plot shows no clear pattern, so the residuals are (almost) symmetrically distributed.

e)

```
friedman.test(time,interface,skill)
```

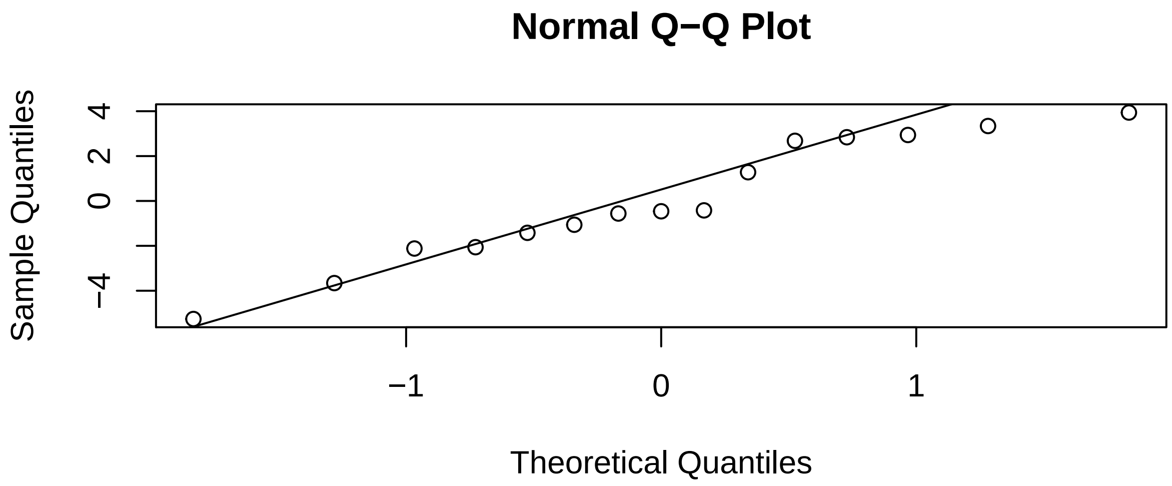
- We reject H_0 (= interface does not have an effect) as the p-value is 0.0407622039783662, which is lower than 5%.

f)

```
aovsearch2 = lm(time~interface, data=search); anova(aovsearch2)
```

- The one-wayANOVA, ignoring the variable `skill`, outputs a p-value of 0.0964165. Therefore, we cannot reject H_0, meaning that the means of the search times are the same for the different interfaces.
- As we the interaction plots in question 2b) showed that there is interaction between interface and skill, it is not right nor useful to ignore the variable `skill`.
- The assumption of a one-wayANOVA is that the data is normally distrubuted. However, the QQ-plot below shows that the data is not normal, so the assumption is not met, nor is it valid.

```
qqnorm(residuals(aovsearch2))
qqline(residuals(aovsearch2))
```



Exercise 3

```
## Loading required package: Matrix
```

```
cow = read.table("data/cow.txt", header=TRUE)
```

a)

```r
aovcow = lm(milk~id+per+treatment,data=cow)
anova(aovcow)
```

- The factor of interest here is type of feedingstuffs (treatment), which is therefore put in as the last factor of the ANOVA formula. The ANOVA outputs a p-value of 0.9346727, which means we accept the null hypothesis that treatment does not influence the milk production (H_0).

```r
cow$id = factor(cow$id); cow$per=factor(cow$per)
cowlm = lm(milk~treatment+per+id, data=cow)
summary(cowlm)
```

- The difference between treatment A (the Intercept) and treatment B is -0.51.

b)

```r
attach(cow)
cowlmer1 = lmer(milk~treatment+order+per+(1|id),REML=FALSE)
cowlmer2 = lmer(milk~order+per+(1|id),REML=FALSE)
anova(cowlmer2,cowlmer1)
```

- By performing a mixed effects analysis in the form of an ANOVA test, modelling the cow effect as a random effect using `lmer`, we find that the p-value equals 0.4460314. This leads us to accepting the hypothesis that treatment does not influence the milk production (H_0)

```r
summary(cowlmer1)
```

- From the `summary` function, it follows that - just as in question 3a) - the difference between treatment A (the Intercept) and treatment B is -0.51.

c)

```r
attach(cow)
```

```
## The following objects are masked from cow (pos = 3):
##
##     id, milk, order, per, treatment
```

```r
t.test(milk[treatment=="A"],milk[treatment=="B"],paired=TRUE)
```

- The t-test outputs the p-value 0.828095901847951, we therefore cannot reject H_0 that there is no difference in milk production given the two treatments. This is indeed compatible with 3a), where we concluded that the treatment did not influence the milk production. This t-test is thus a valid test.

Exercise 4

```
nauseatable = read.table(file = 'data/nauseatable.txt', header = TRUE)
```

a)

```
table_to_vector = unlist(nauseatable,  use.names = FALSE)
# Create nausea column, possible values 0 (No Nausea), 1(Nausea)
nausea = rep(c('0', '1'), each = 3, times = c(table_to_vector))
med = c('Chlorpromazine','Pentobarbital(100mg)','Pentobarbital(150mg)')
# Create medicine column, contains name of all the medicines
medicine = rep(c(med,med), each = 1, times = c(table_to_vector))
df = data.frame(cbind(nausea, medicine))
(df[c(1,101,133,181,233,268),])
```

```
##     nausea             medicine
## 1        0       Chlorpromazine
## 101      0 Pentobarbital(100mg)
## 133      0 Pentobarbital(150mg)
## 181      1       Chlorpromazine
## 233      1 Pentobarbital(100mg)
## 268      1 Pentobarbital(150mg)
```

While studying the outcome of the table below, we see that with xtabs we get a contingency table from the medicine and nausea factors. There are more people suffering from nausea with the medicine Pentobarbital(100mg and 150mg combined) than with Chlorpromazine

```
xtabs(~medicine+nausea)
```

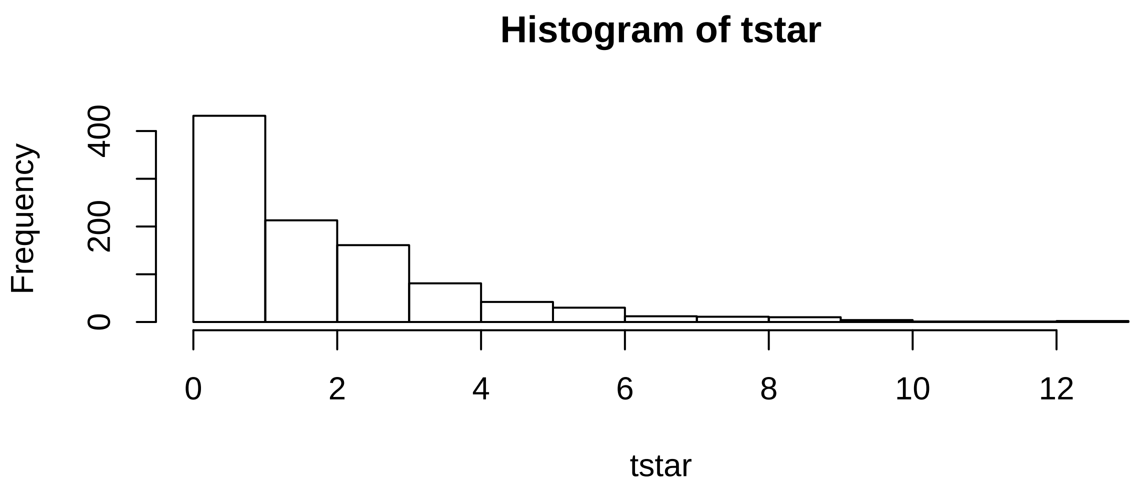
```
##                       nausea
## medicine                 0   1
##   Chlorpromazine       100  52
##   Pentobarbital(100mg)  32  35
##   Pentobarbital(150mg)  48  37
```

b)

```
#options(scipen = 999)
meds = factor(medicine)
mystat=function(x) sum(residuals(x)^2)
B=1000
tstar=numeric(B)
for (i in 1:B) {
  treatstar=sample(medicine)
  tstar[i]=chisq.test(xtabs(~treatstar+nausea, data = nauseatable))[[1]]
}
myt=chisq.test(xtabs(~medicine+nausea, data = nauseatable))[[1]]
myt
```

```
## X-squared
##  6.624765
```

```
hist(tstar)
```



```
pl = sum(tstar<myt)/B
pr = sum(tstar>myt)/B
pmin = min(pl,pr)
(pvalue = pmin)
```

```
## [1] 0.034
```

The obtained p-value from permutation test is over 5% thus we accept the null hypothesis (H_0) and conclude that the variables are not dependent.

c)

```
(pvalue_chisq = chisq.test(xtabs(~medicine+nausea, data = nauseatable))[[3]])
```

```
## [1] 0.03642928
```

```
(pvalue_tstar=pmin)
```

```
## [1] 0.034
```

We received very close p-values from both the permutation and chi-square tests. Even though we received 0.034 from the permutation test, since they both statiscally perform with regards to significance of the factors we accepted the null hypothesis (H_0). Since a single chi-squared test is one member from the permutation test, having close result is justifiable.

Exercise 5

```
expenses_crime = read.table(file = 'data/expensescrime.txt', header = TRUE)
```

b)

In the dataset there are 5 possible explantory `variables`, `bad`, `crime`, `lawyers`, `employ`, and `pop`.

```
# Step Up method
summary(lm( expend~bad ,data = expenses_crime))[[9]]
```

```
## [1] 0.6901876
```

```
summary(lm( expend~crime ,data = expenses_crime))[[9]]
```

```
## [1] 0.09373104
```

```
summary(lm( expend~lawyers ,data = expenses_crime))[[9]]
```

```
## [1] 0.9359988
```

```
summary(lm( expend~employ ,data = expenses_crime))[[9]]
```

```
## [1] 0.9530352
```

```
summary(lm( expend~pop ,data = expenses_crime))[[9]]
```

```
## [1] 0.9054348
```

Explanotory variable ‘employ’ delivers the highest R^2 value.

```
summary(lm( expend~employ+bad ,data = expenses_crime))[[9]]
```

```
## [1] 0.9532261
```

```
summary(lm( expend~employ+crime ,data = expenses_crime))[[9]]
```

```
## [1] 0.9531771
```

```
summary(lm( expend~employ+lawyers ,data = expenses_crime))[[9]]
```

```
## [1] 0.9616402
```

```
summary(lm( expend~employ+pop ,data = expenses_crime))[[9]]
```

```
## [1] 0.9524063
```

Newly added variable ‘lawyers’ yields better R^2 compared to others.

```
summary(lm( expend~employ+lawyers+bad ,data = expenses_crime))[[9]]
```

```
## [1] 0.9615682
```

```
summary(lm( expend~employ+lawyers+crime ,data = expenses_crime))[[9]]
```

```
## [1] 0.9608384
```

```
summary(lm( expend~employ+lawyers+pop ,data = expenses_crime))[[9]]
```

```
## [1] 0.9614177
```

Adding additional variables leads to insignificant explanatory variables. Thus, 'step up' process need to be stopped at previous step.

```
#Final model for the step up approach
summary(lm( expend~employ+lawyers ,data = expenses_crime))[8]
```

```
## $r.squared
## [1] 0.9631745
```

Step Down method

```
summary(lm( expend~bad+crime+lawyers+employ+pop ,data = expenses_crime))[[4]]
```

```
##                  Estimate   Std. Error   t value    Pr(>|t|)
## (Intercept) -299.13408620 1.400527e+02 -2.135868 0.038166095
## bad           -2.83192107 1.240335e+00 -2.283190 0.027193547
## crime          0.03241186 2.813117e-02  1.152169 0.255336038
## lawyers        0.02324356 8.044089e-03  2.889521 0.005916572
## employ         0.02297074 7.461822e-03  3.078435 0.003538739
## pop            0.07786665 3.514981e-02  2.215279 0.031844579
```

Explanatory variable 'crime' has p-value is larger than 0.05. Thus removing it from the model.

```
summary(lm( expend~bad+lawyers+employ+pop ,data = expenses_crime))[[4]]
```

```
##                  Estimate   Std. Error   t value    Pr(>|t|)
## (Intercept) -146.42386127 45.410089800 -3.224479 0.002323987
## bad           -2.24065336  1.133206751 -1.977268 0.054022124
## lawyers        0.02646062  0.007570753  3.495111 0.001060463
## employ         0.02283249  0.007487368  3.049467 0.003795217
## pop            0.06368200  0.033040269  1.927406 0.060117598
```

Explanatory variable 'bad' has p-value is larger than 0.05. Thus removing them from the model.

```r
#Final model using Step down method
summary(lm( expend~lawyers+employ+pop ,data = expenses_crime))[[4]]
```

```
##                   Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept) -123.34725376 45.222648097 -2.7275549 0.008943060
## lawyers        0.02722640  0.007791357  3.4944365 0.001047007
## employ         0.02489534  0.007640376  3.2583924 0.002085147
## pop            0.02246543  0.026416233  0.8504405 0.399392044
```

Explanatory variable ‘pop’ has p-value is larger than 0.05. Thus removing them from the model.

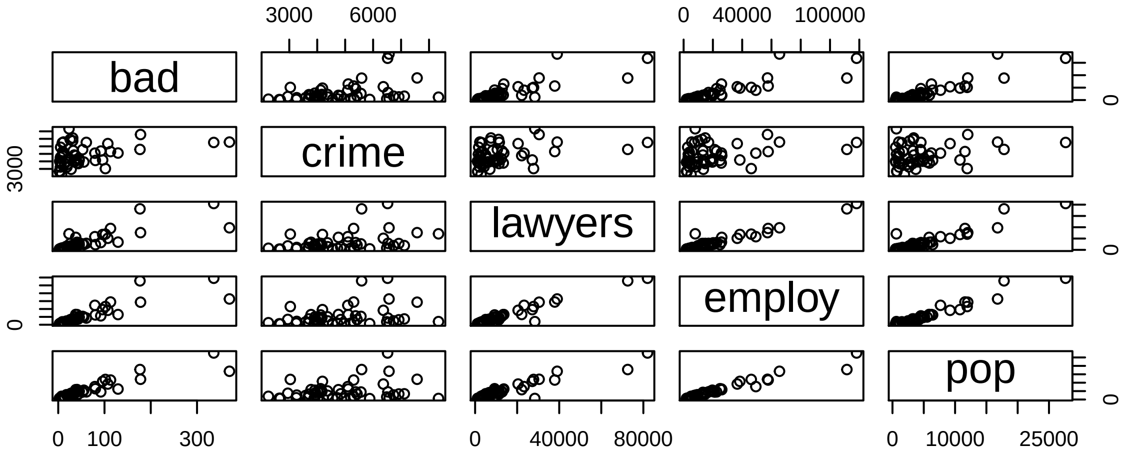
```r
summary(lm( expend~lawyers+employ ,data = expenses_crime))[8]
```

```
## $r.squared
## [1] 0.9631745
```

No need to remove further variables as all remaining explanatory variables in the model are significant. Conclusion: "Step up’ and ‘Step down’ methods results into same model with R-squared: 0.963174545600467.

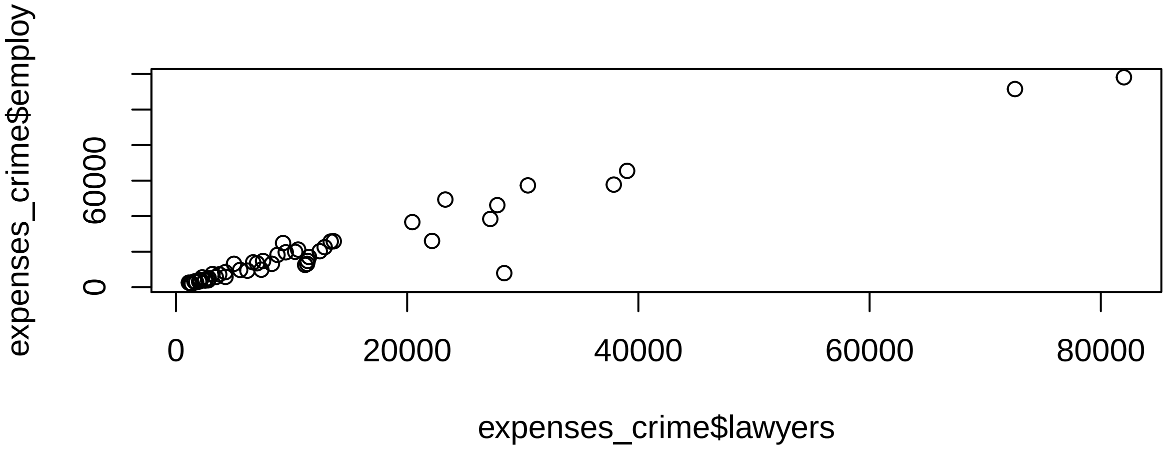
a)

```r
plot(expenses_crime[,c(3:7)])
```



In the plot below, there are ‘potential points’ on x axis between values 70000-85000.

```r
plot(expenses_crime$lawyers, expenses_crime$employ)
```

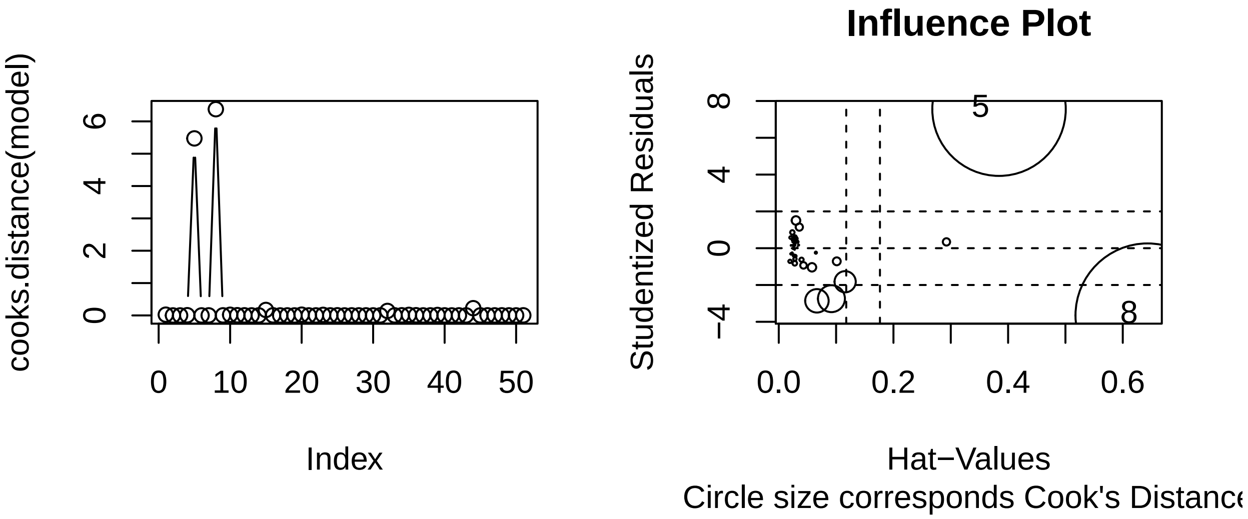


```
model = lm( expend~lawyers+employ ,data = expenses_crime)
library(car)
```

```
## Loading required package: carData

## Registered S3 methods overwritten by 'car':
##   method                          from
##   influence.merMod                lme4
##   cooks.distance.influence.merMod lme4
##   dfbeta.influence.merMod         lme4
##   dfbetas.influence.merMod        lme4
```

```
par(mfrow=c(1, 2))
plot(cooks.distance(model),type="b")
influencePlot(model, main="Influence Plot", sub="Circle size corresponds Cook's Distance")
```



```
##     StudRes       Hat    CookD
## 5  7.555051 0.3841429 5.473251
## 8 -3.652735 0.6431568 6.376359
```

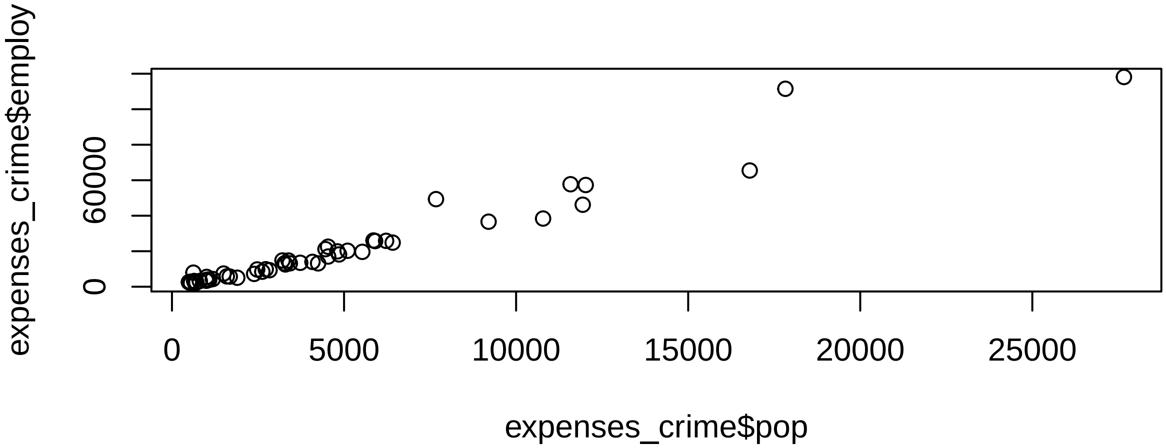
In the plot above, we clearly see influence points: the Cook's distance is 5.47 for the leverage point at index 5 and 6.38 for the leverage point at index 8.

In the following graph and correlation table, 'pop' and 'employ' are collinear with correlation value of 0.9707407.

```r
cor(expenses_crime[,c(3:7)])[24]
```

```
## [1] 0.9707407
```

```r
plot(expenses_crime$pop, expenses_crime$employ)
```



VIF values of both the variables in the model is higher than 5, which represents the collinearity problem.

```r
vif(model)
```

```
##  lawyers   employ
## 14.83915 14.83915
```

Since both the variables have same value, we need to remove one of the variables from the model as below:

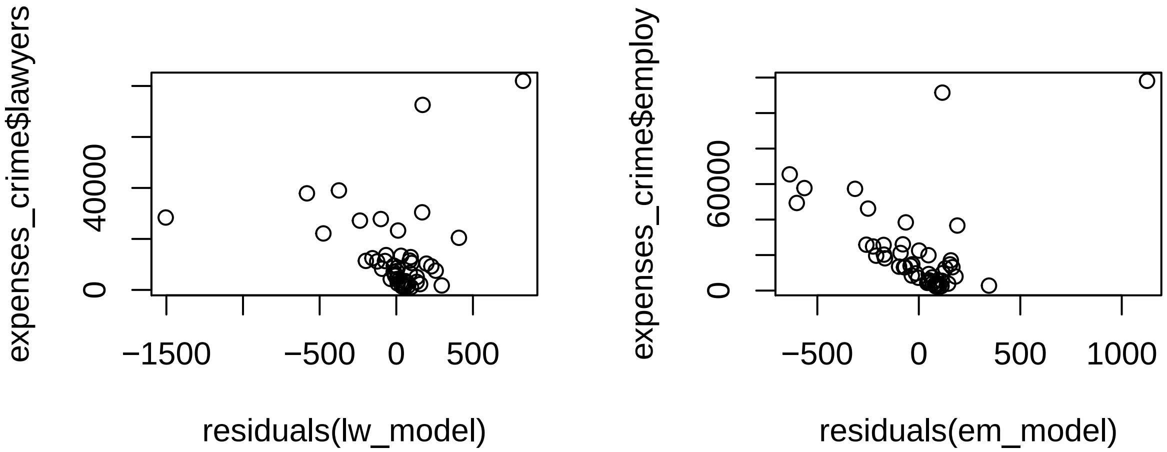
```r
model_new = lm( expend~lawyers ,data = expenses_crime)
```

c)

Scatter plot of residuals against each Xk in the model separately. There is no visible pattern in the plots.

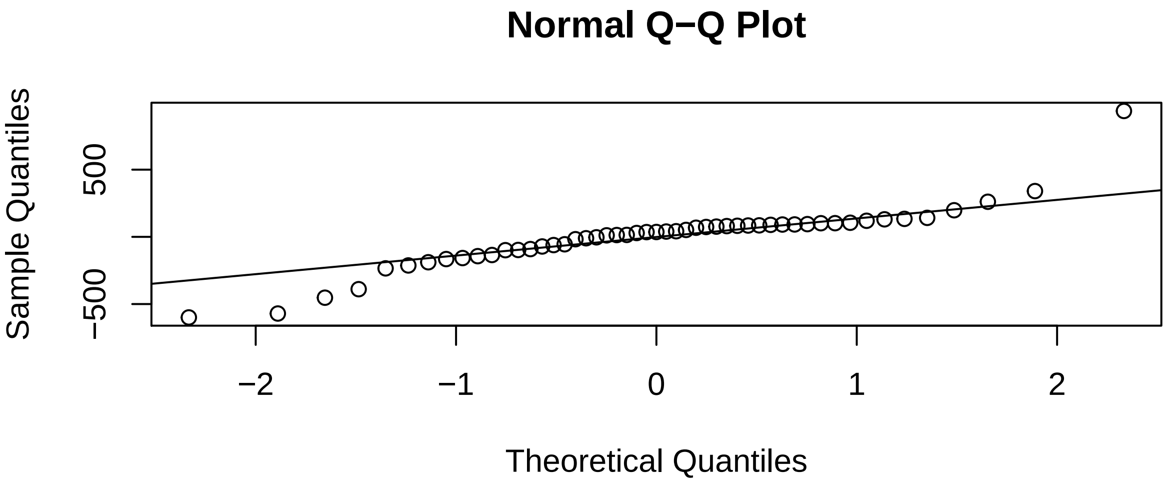
```
lw_model = lm( expend~lawyers ,data = expenses_crime)
em_model = lm( expend~employ ,data = expenses_crime)

par(mfrow=c(1, 2))
plot(residuals(lw_model),expenses_crime$lawyers)
plot(residuals(em_model),expenses_crime$employ)
```



From the normal QQ-plot of the residuals it is evident that error is not normally distributed.

```
qqnorm(residuals(model));qqline(residuals(model))
```


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles