Assignment 2

Andrea Di Dio, Chenghan Song, Jiacheng Lu — Group 22

Exercise 1

a)

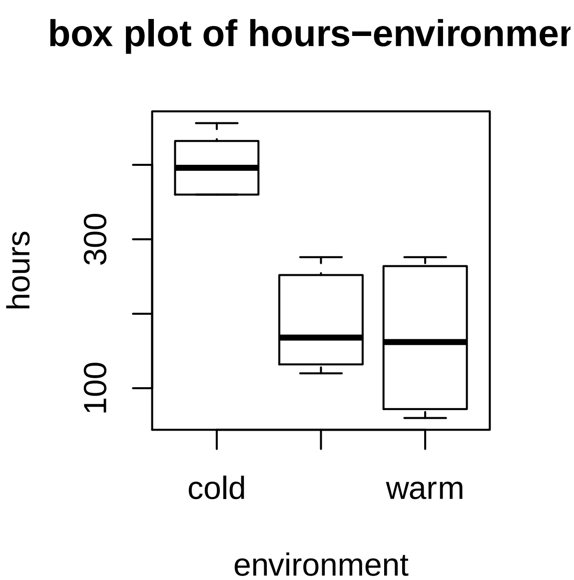
We assign combine levels (i,j) of the factors to a random set of N units

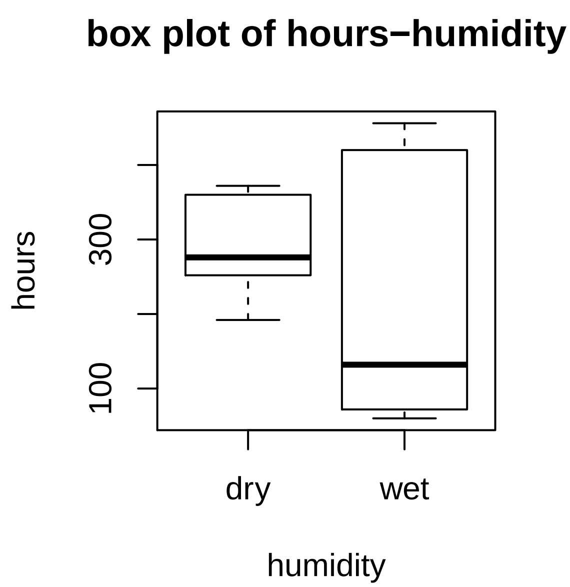
```
I=3; J=2; N=3
rbind(rep(1:I,each=N*J),rep(1:J,N*I),sample(1:(N*I*J)))
```

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

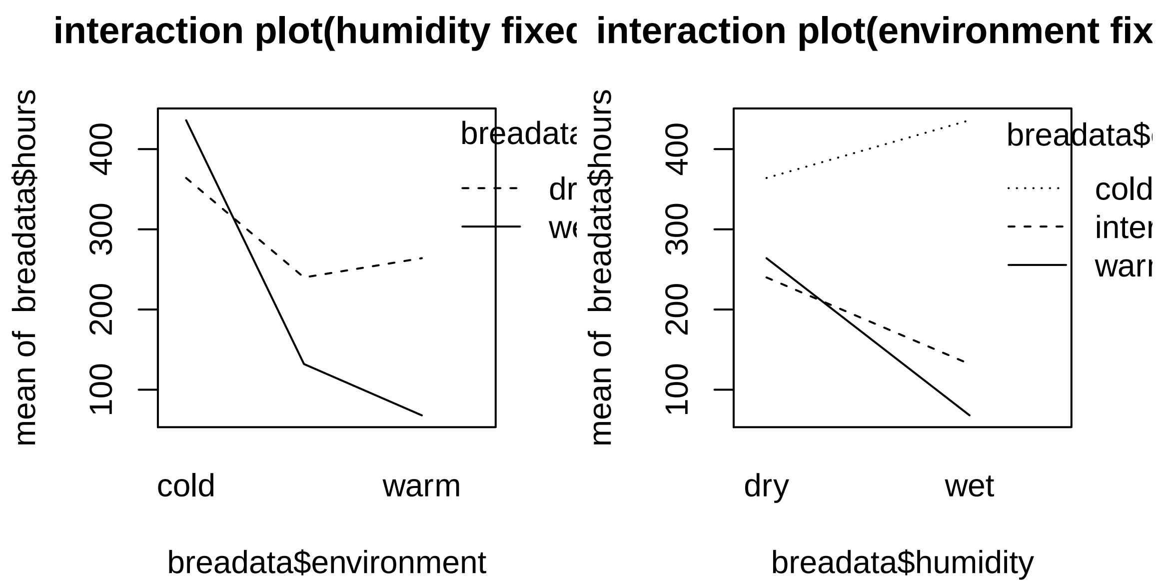
b)

```
breadata=read.table(file="bread.txt",header=TRUE);par(mfrow=c(1,2))
boxplot(hours~environment,data=breadata, main="box plot of hours-environment")
boxplot(hours~humidity, data=breadata,main="box plot of hours-humidity"); par(mfrow = c(1,2))
```






```
interaction.plot(breadata$environment,breadata$humidity,breadata$hours, main="interaction plot(humidity
interaction.plot(breadata$humidity,breadata$environment,breadata$hours, main="interaction plot(environm
```



The lines in the interaction plots are unparallel, thus we can assume that there are interactions between temperature and humidity.

c)

```
breadaov=lm(hours~environment*humidity, data = breadata); anova(breadaov)
```

```
## Analysis of Variance Table
##
## Response: hours
##                      Df Sum Sq Mean Sq F value  Pr(>F)
## environment           2 201904  100952   233.7 2.5e-10 ***
## humidity              1  26912   26912    62.3 4.3e-06 ***
## environment:humidity  2  55984   27992    64.8 3.7e-07 ***
## Residuals            12   5184     432
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

H_0:$\alpha_i = 0$ for all i: The temperature does not have a main effect on the time of decay

H_0: $\beta_j = 0$ for all j: The humidity does not have a main effect on the time of decay

H_0: $\gamma_{ij} = 0$ for all (i, j): Temperature and humidity do not have interaction effects on the time of decay

The p-value for testing H_0:$\alpha_i = 0$ for all i is 2.461e-10, hence H_0 is rejected; for H_0: $\beta_j = 0$ for all j is 4.316e-06, hence H_0 is rejected; for H_0: $\gamma_{ij} = 0$ for all (i, j) is 3.705e-07, hence H_0 is rejected. So both temperature and humidity have main effects and there are also interactions between two factors.

```
summary(breadaov)
```

```
##
## Call:
## lm(formula = hours ~ environment * humidity, data = breadata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##    -48     -7      0     11     36
##
```

```
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                             364         12   30.33  1.0e-12 ***
## environmentintermediate                -124         17   -7.31  9.4e-06 ***
## environmentwarm                        -100         17   -5.89  7.3e-05 ***
## humiditywet                              72         17    4.24   0.0011 **
## environmentintermediate:humiditywet    -180         24   -7.50  7.2e-06 ***
## environmentwarm:humiditywet            -268         24  -11.17  1.1e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.8 on 12 degrees of freedom
## Multiple R-squared:  0.982,  Adjusted R-squared:  0.975
## F-statistic:  132 on 5 and 12 DF,  p-value: 4.68e-10
```

H_0 for the interaction of intermediate environment and wet humidity: there is no interaction effect of intermediate environment and wet humidity. The p-value is 7.23e-06, therefore we reject H_0 and conclude that there is a significant interaction between intermediate temperature and wet humidity.

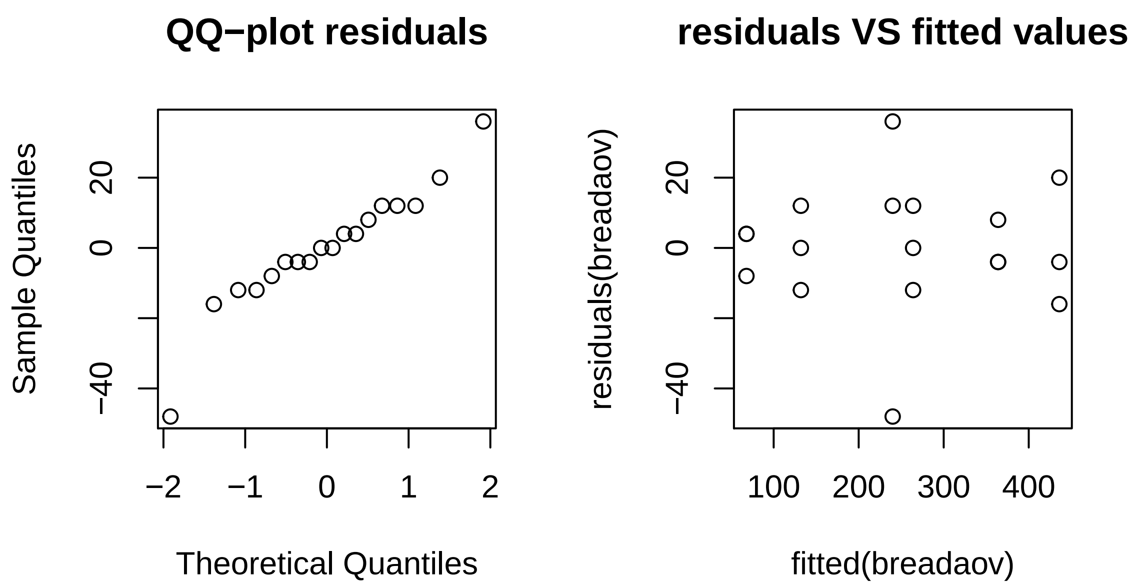
H_0 for the interaction of warm environment and wet humidity: there is no interaction effect of warm environment and wet humidity. The p-value is 1.07e-07, therefore we reject H_0 and conclude that there is a significant interaction between warm temperature and wet humidity.

d)

This is not a good question. Because the interaction effects between two factors are significant. We can not compare the influence of the first and second factor when there are interactions between them.

e)

```
par(mfrow=c(1,2));qqnorm(residuals(breadaov), main="QQ-plot residuals"); plot(fitted(breadaov),residuals
```



From the QQ-Plot we can see that the residuals are not normally distributed.

The spread in the residuals seems to be bigger for some certain fitted values. Due to the spread in the residuals should not change systematically with any variable, in particular not with the fitted values, there might be some outliers.

Exercise 2

a)

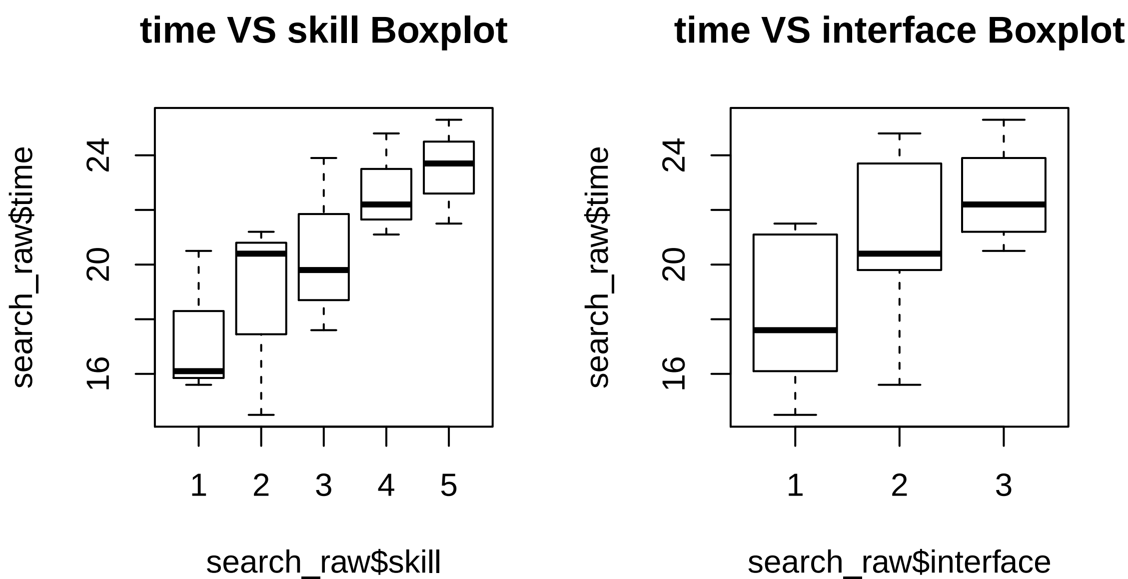
For the *randomized block design*, we perform the experiment with 5 blocks (B) using 3 interfaces (I) and have one replicate per treatment level per block (N).

```
search_raw <- read.delim("./search.txt", header = TRUE, sep = "")
I <- 3; B <- 5; N <- 1;
for(i in 1:B) print(sample(1:(N * I)))
```

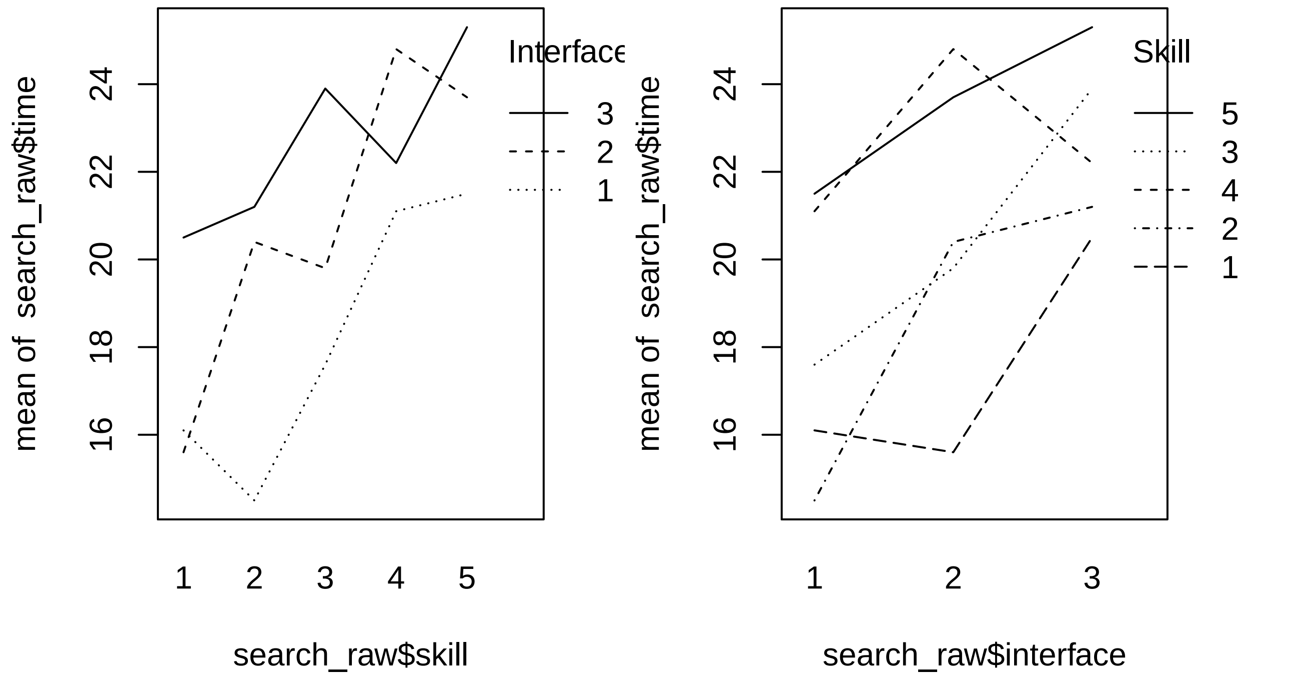
```
## [1] 2 3 1
## [1] 1 3 2
## [1] 3 2 1
## [1] 2 3 1
## [1] 1 2 3
```

b)

```
par(mfrow = c(1,2))
boxplot(search_raw$time ~ search_raw$skill, main = "time VS skill Boxplot")
boxplot(search_raw$time ~ search_raw$interface, main = "time VS interface Boxplot")
```




```
par(mfrow=c(1,2))
interaction.plot(search_raw$skill, search_raw$interface, search_raw$time, trace.label = "Interface")
interaction.plot(search_raw$interface, search_raw$skill, search_raw$time, trace.label = "Skill")
```



We have used two different graphical summaries in order to investigate the interaction between the interfaces and skill of the students on the time taken to complete the task. The boxplots show that both factors (*skill* and *interface*) affect the dependent variable (*time*). The lower the value of the skill factor (and hence the higher the competence of the student) result in lower times to complete the task. Also the interface factor seems to result in a difference of time taken to complete the task. This shows that there exists a dependence of the dependent variable on the factors of interest. The interaction plots show that the lines plotted are not parallel to one another, implying a significant interaction between the skill and the interface used to complete the task.

c)

H_0 : The search time is the same for all the interfaces (There is no interaction between the factors).
Given that our H_0 says that there is no interactions, we test using the additive model:

```
search_raw$skill <- as.factor(search_raw$skill)
search_raw$interface <- as.factor(search_raw$interface)
search_aov <- lm(search_raw$time ~ search_raw$skill + search_raw$interface, data = search_raw)
print(anova(search_aov), signif.stars = F)
```

```
## Analysis of Variance Table
##
## Response: search_raw$time
##                      Df Sum Sq Mean Sq F value Pr(>F)
## search_raw$skill      4   80.1   20.01    6.21  0.014
## search_raw$interface  2   50.5   25.23    7.82  0.013
## Residuals             8   25.8    3.23
```

The p-value $0.013 < \alpha$ meaning that we can safely reject H_0 concluding that the interfaces used affect the search time.

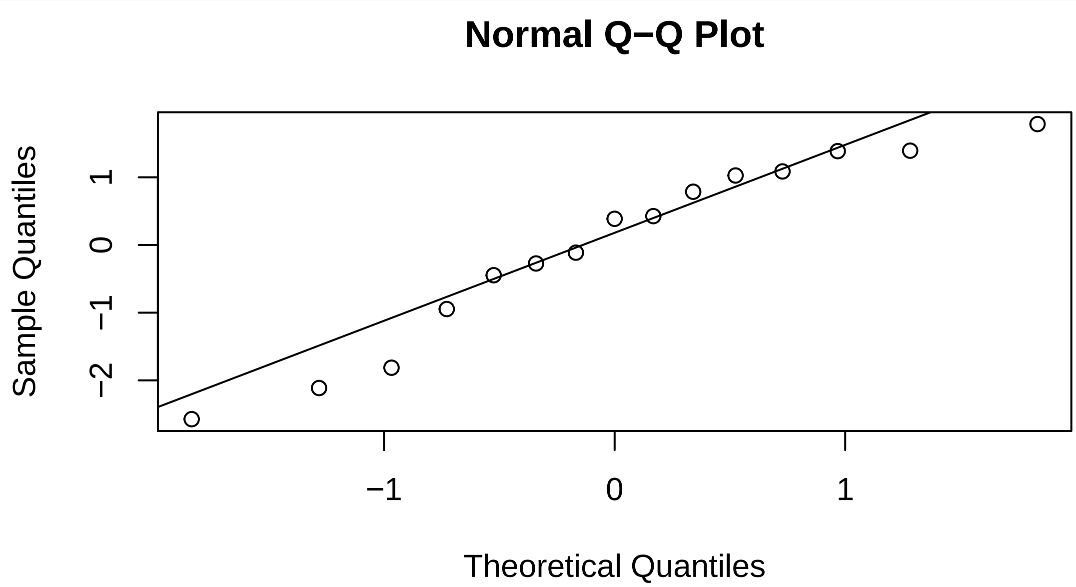
```
contrasts(search_raw$interface) <- contr.sum; contrasts(search_raw$skill) <- contr.sum
search_aov2 <- lm(search_raw$time  ~ search_raw$interface + search_raw$skill)
summary(search_aov2)
```

```
##
## Call:
## lm(formula = search_raw$time ~ search_raw$interface + search_raw$skill)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.573 -0.697  0.387  1.057  1.787
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             20.547      0.464   44.31  7.4e-11 ***
## search_raw$interface1   -2.387      0.656   -3.64   0.0066 **
## search_raw$interface2    0.313      0.656    0.48   0.6456
## search_raw$skill1       -3.147      0.927   -3.39   0.0095 **
## search_raw$skill2       -1.847      0.927   -1.99   0.0816 .
## search_raw$skill3       -0.113      0.927   -0.12   0.9057
## search_raw$skill4        2.153      0.927    2.32   0.0488 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.8 on 8 degrees of freedom
## Multiple R-squared:  0.835,  Adjusted R-squared:  0.711
## F-statistic: 6.74 on 6 and 8 DF,  p-value: 0.0084
```

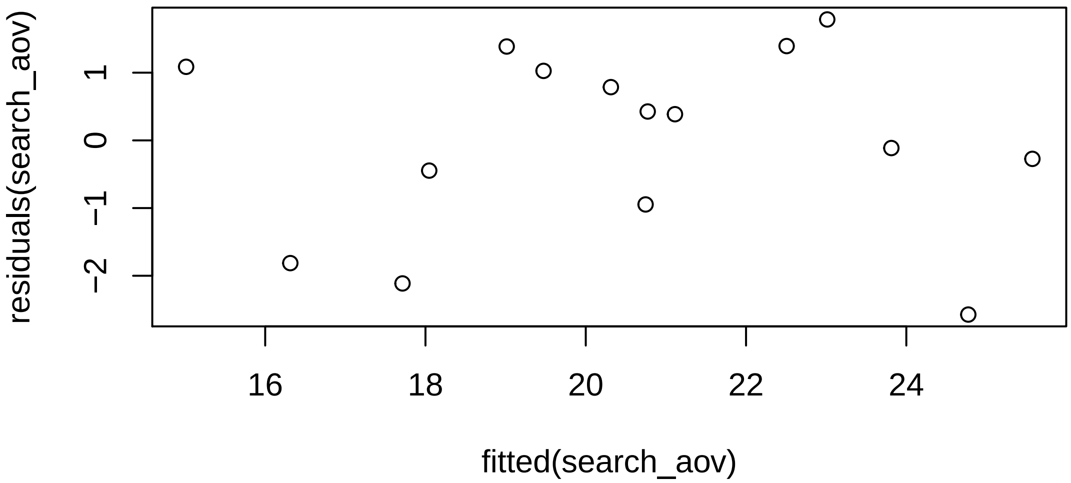
To estimate the time taken by a student with skill level 3 to complete the search task using interface 2, we should sum the intercept term with the coefficients of *skill3* and *interface2* $20.5467 + (-0.1133) + 0.3133 = 20.7467$.

d)

```r
qqnorm(residuals(search_aov));qqline(residuals(search_aov))
```



```r
plot(fitted(search_aov), residuals(search_aov))
```




```r
print(shapiro.test(residuals(search_aov)))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(search_aov)
## W = 0.9309, p-value = 0.282
```

The *QQ plot* of the residuals for the data shows that the data might seem to be normally distributed and this is also confirmed by the **Shapiro-Wilk test** which shows a p-value of 0.282. The plot which shows the fitted values VS the residuals show that there is no systematic change in the residuals based on the fitted values as the points are well-spread out, suggesting that the two populations have equal variances.

e)

H_0 : There is no effect in the interface used on the search time.

```r
friedman.test(search_raw$time, search_raw$interface, search_raw$skill)
```

```
##
##  Friedman rank sum test
##
## data:  search_raw$time, search_raw$interface and search_raw$skill
## Friedman chi-squared = 6.4, df = 2, p-value = 0.041
```

The non-parametric Friedman test gives a p-value of $0.041 < \alpha$ meaning that we can safely reject H_0 meaning that the choice in the interface used for the search engine has a statistically significant effect on the search time.

f)

H_0 : The interface used does not affect the search time.

```r
search_aov_one_way <- lm(search_raw$time ~ search_raw$interface, data = search_raw)
print(anova(search_aov_one_way), signif.stars = F)
```

```
## Analysis of Variance Table
##
## Response: search_raw$time
```

```
##                      Df Sum Sq Mean Sq F value Pr(>F)
## search_raw$interface  2   50.5   25.23    2.86  0.096
## Residuals            12  105.9    8.82
```

The resulting p-value is $0.096 > \alpha$ meaning that we fail to reject H_0 and hence could be possible that the interfaces have no effect on the search time. However, given that the testing in part `b)` suggest that there is interaction between the skill and the interface factors, this test is not useful and wrong. In order for this test to be valid, we must assume no interaction between the two factors on the dependent variable, but this is not the case meaning that this test is not statistically meaningful.

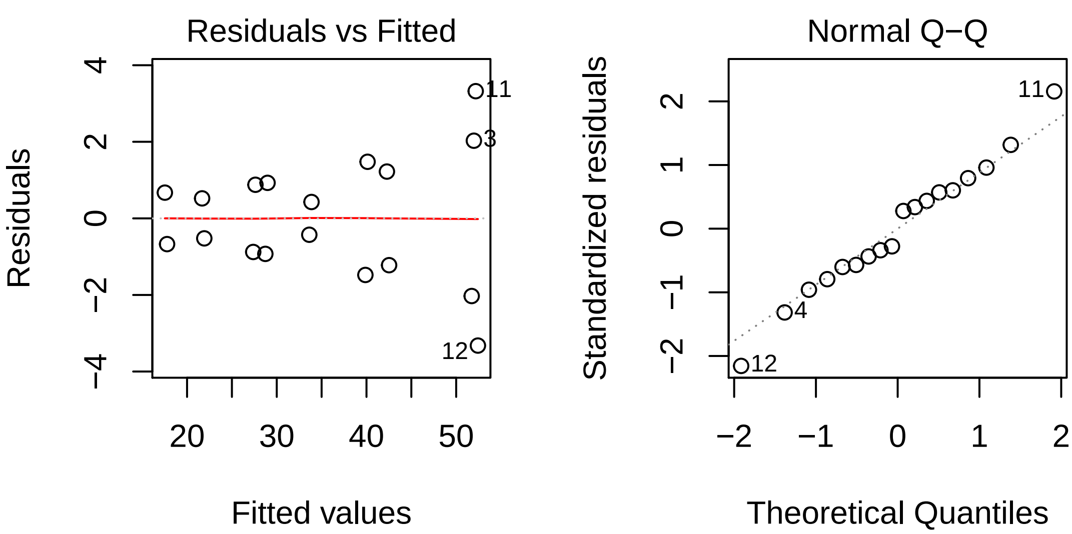
Exercise 3

a)

```
cow=read.table(file="cow.txt",header=TRUE)
cow$id=factor(cow$id); cow$per=factor(cow$per)
aovcow=lm(milk~treatment+id,data=cow);print(anova(aovcow), signif.stars = F)
```

```
## Analysis of Variance Table
##
## Response: milk
##           Df Sum Sq Mean Sq F value  Pr(>F)
## treatment  1      0     0.3    0.05    0.83
## id         8   2467   308.4   57.74 2.8e-06
## Residuals  8     43     5.3
```

```
par(mfrow=c(1,2));plot(aovcow,1);plot(aovcow,2)
```



H_0: the type of feedingstuffs does not influence milk production

The p-value for testing H_0 is 0.8281, thus we fail to reject H_0, the type of feedingstuffs does not have an effect on milk production. From the plot above, the residuals do not seem to deviate significantly from normal.

However, an ordinary "mixed effects" model is not suitable in this case where the assumption of "exchangebility" may fail. Cows may be happy with or bored at feedingstuff, then a block design is invalid.

b)

```r
cowlm=lm(milk~id+per+treatment,data=cow); print(anova(cowlm), signif.stars = F);summary(cowlm)[4]
```

```
## Analysis of Variance Table
##
## Response: milk
##           Df Sum Sq Mean Sq F value  Pr(>F)
## id         8   2467   308.4  124.48 7.5e-07
## per        1     25    24.5    9.89   0.016
## treatment  1      1     1.2    0.47   0.517
## Residuals  7     17     2.5
```

```
## $coefficients
##             Estimate Std. Error  t value   Pr(>|t|)
## (Intercept)    30.30    1.24442 24.34877 5.0185e-08
## id2            23.00    1.57408 14.61175 1.6798e-06
## id3            11.15    1.57408  7.08352 1.9649e-04
## id4            -1.35    1.57408 -0.85765 4.1948e-01
## id5            -7.05    1.57408 -4.47882 2.8703e-03
## id6            23.45    1.57408 14.89763 1.4723e-06
## id7            13.55    1.57408  8.60823 5.6925e-05
## id8             4.90    1.57408  3.11294 1.7011e-02
## id9           -11.20    1.57408 -7.11529 1.9108e-04
## per2           -2.39    0.74665 -3.20097 1.5046e-02
## treatmentB     -0.51    0.74665 -0.68305 5.1654e-01
```

```r
cowlmer=lmer(milk~treatment+order+per+(1|id),REML=FALSE, data = cow);summary(cowlmer)[10]
```

```
## $coefficients
##             Estimate Std. Error  t value
## (Intercept)    38.50    5.81104  6.62532
## treatmentB     -0.51    0.65848 -0.77451
## orderBA        -3.47    7.76847 -0.44668
## per2           -2.39    0.65848 -3.62956
```

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

```
## Data: cow
## Models:
## cowlmer1: milk ~ order + per + (1 | id)
## cowlmer: milk ~ treatment + order + per + (1 | id)
##          Df AIC BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## cowlmer1  5 118 122  -53.9      108
## cowlmer   6 119 125  -53.7      107  0.58      1       0.45
```

H_0: the type of feedingstuffs does not influence milk production

The difference between incorrect fixed effects analysis and correct mixed effects analysis is minor. The estimated treatment and period effects under fixed effects of mixed effects analysis are identical to those in fixed effects analysis. In fixed effects analysis, the p-value for testing H_0 is 0.517, thus we don't reject H_0, the type of feedingstuffs does not influence milk production. In mixed effects analysis, the p-value for testing H_0 is 0.45, thus we accept H_0, the type of feedingstuffs does not influence milk production.

c)

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

```
##
##  Paired t-test
##
## data:  milk[treatment == "A"] and milk[treatment == "B"]
## t = 0.224, df = 8, p-value = 0.83
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.2679  2.7568
## sample estimates:
## mean of the differences
##                 0.24444
```

The paired-sample t-test produces a invalid test for a difference in milk production, because repeated measures may not be exchangeable because of time effect: Cows may mature or get older and learning effect: Cows may be happy with or bored at feedingstuff.

The p-value for treatment is identical to the one of the previous "fixed effects" obtained in a) (the order of the treatments was ignored). They are compatible because in the case of two repeated measurements, t^2 value for paired t-test is identical to the F value of the repeated measures ANOVA.

Exercise 4

a)

```
nausea_raw <- read.delim("./nauseatable.txt", header = TRUE, sep = "")
med_vec <- c();nausea_vec <- c()

for(i in 1:3) { med_name <- rownames(nausea_raw)[i]
  for(j in 1:2) { nausea <- colnames(nausea_raw)[j]
    for(k in 1:nausea_raw[,j][i]) { med_vec <- append(med_vec, med_name)
      if(j == 1) { nausea_vec <- append(nausea_vec, "no")}
      else if(j == 2){ nausea_vec <- append(nausea_vec, "yes")}
    }
  }
}
nausea_frame <- data.frame(medicin = med_vec, naus = nausea_vec)
xtabs(~nausea_frame$medicin + nausea_frame$naus)
```

```
##                         nausea_frame$naus
## nausea_frame$medicin    no yes
##   Chlorpromazine       100  52
##   Pentobarbital(100mg)  32  35
##   Pentobarbital(150mg)  48  37
```

b)

H_0 : There is no difference between the treatment with different medicines (populations are the same).

```
permutation_test <- function() {
  B <- 1000; Tstar <- numeric(B)
  for(i in 1:B) { Xstar <- sample(nausea_frame$medicin)
    Tstar[i] <- chisq.test(xtabs(~Xstar + nausea_frame$naus))[[1]]}
```

```
  t <- chisq.test(xtabs(~nausea_frame$medicin + nausea_frame$naus))[[1]]
  p1 <- sum(Tstar < t) / B; pr <- sum(Tstar > t) / B
  p <- 2 * min(pr, p1)
  return(p)
}

p_val <- permutation_test()
```

The p-value resulting from the permutation test $0.078 > \alpha$ meaning that we fail to reject H_0 and thus could mean that the two medicines perform equally well to treat nausea. However, it must be said that in a permutation test, the p-value can change depending on the randomly generated samples meaning that a p-value which is just above the significance level, might be lower than the significance level on a different run.

c)

H_0 : There is no difference between the treatment with different medicines (populations are the same).

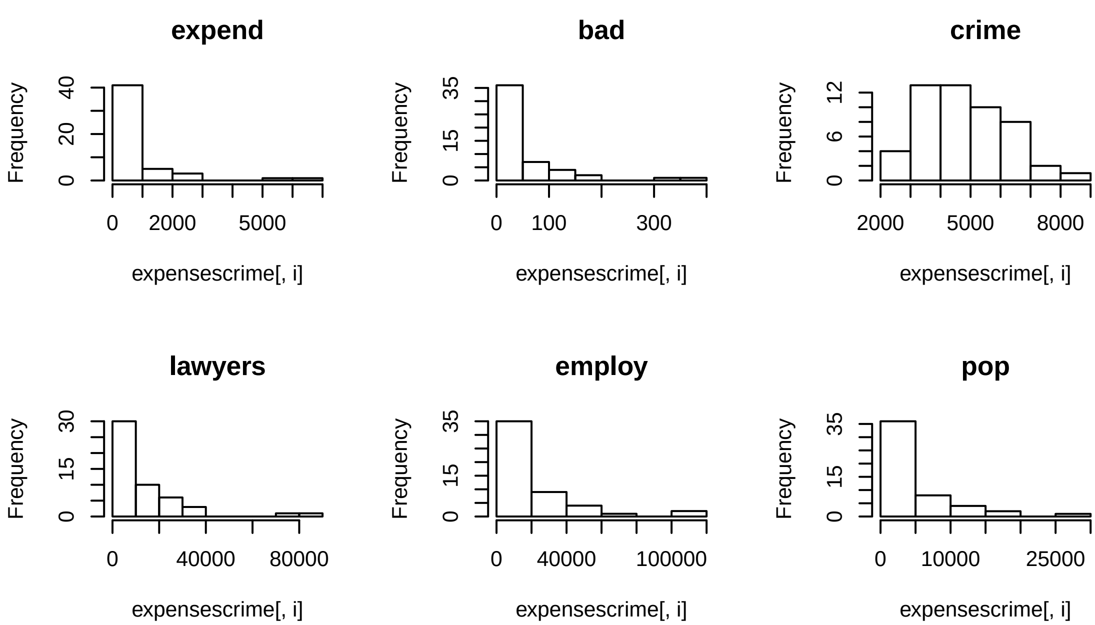
```
p_val_chi <- chisq.test(xtabs(~nausea_frame$medicin + nausea_frame$naus))[[3]]
```

The p-value for the χ^2-test for contingency tables is 0.036 which is lower than that computed in part b) and is, in fact less than trhe significance value α meaning that we can safely reject H_0 suggesting that the different medicines are not equally as good in treating nausea in patients. This could be due to the χ^2 T-values sampled in the permutation test being slightly different from the true χ-distribution.

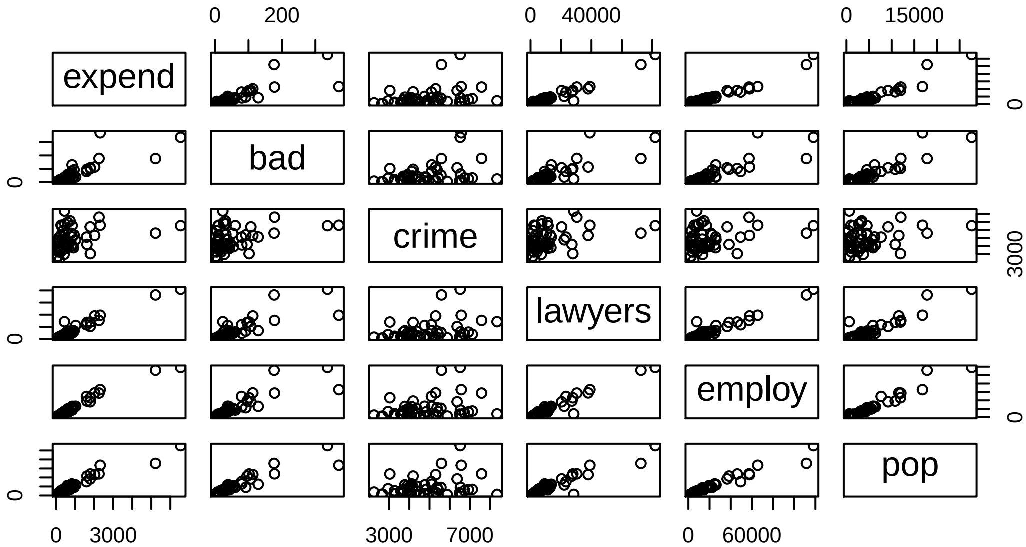
Exercise 5

a)

```
expensescrime=read.table(file="expensescrime.txt",header=TRUE)
par(mfrow=c(2,3))
for (i in c(2,3,4,5,6,7)) hist(expensescrime[,i],main=names(expensescrime[i]))
```




```
plot(expensescrime[,c(2,3,4,5,6,7)])
```



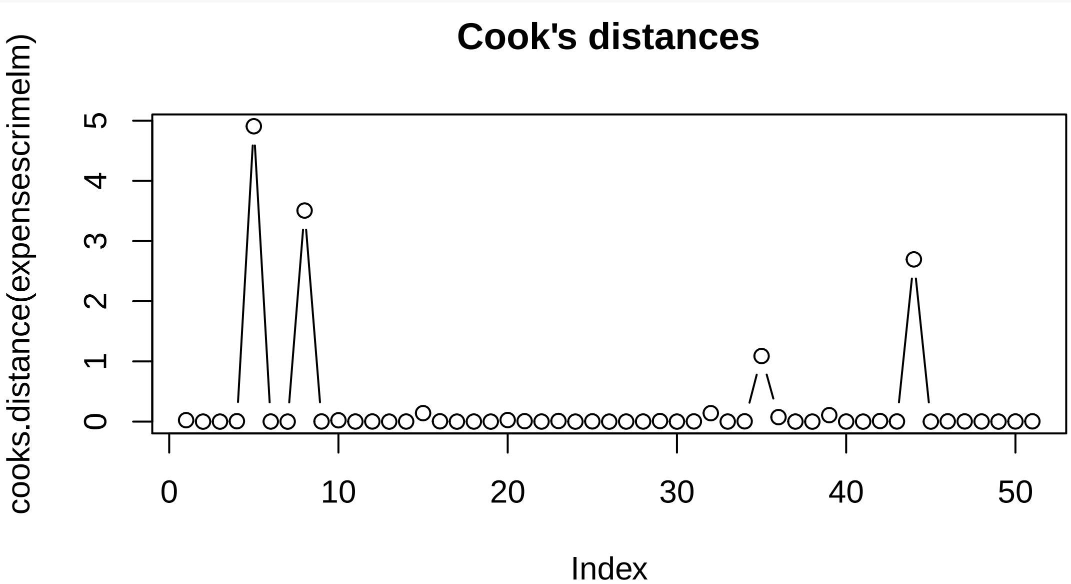

```
par(mfrow=c(2,3))
```

From the plot we can find potential points for each variable. For bad:336.2(CA) & 370.1(TX); For lawyers:82001(CA)&72575(NY);For employ: 118149(CA) & 111518(NY); For pop: 27663(CA).

```
expensescrimelm=lm(expend~bad+crime+lawyers+employ+pop,data=expensescrime)
round(cooks.distance(expensescrimelm),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
```

```
plot(cooks.distance(expensescrimelm),type="b",main="Cook's distances")
```



For influence points, we check the Cook's distance, from the plot we can see that there are four. No.5(CA) & No.8(DC) &

No.35(NY) & No.44(TX).

```r
vif(expensescrimelm)
```

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

```r
round(cor(expensescrime[,c(3,4,5,6,7)]),2)
```

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

From both the scatter plot and pairwise linear correlation and the vif, bad and pop (correlation value=0.92), lawyers and employ (correlation value=0.97), lawyers and pop(correlation value=0.93), employ and pop (correlation value=0.97), so crime and lawyers and employ and pop are collinear.

b)

First, in the step-up method, we starts with fitting all p possible simple linear regression models: $Y_n = \beta_0 + \beta 1 X_{nj} + e_n$ To save pages, Only parts of summary are shown.

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

```
## $r.squared
## [1] 0.696
```

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

```
## $r.squared
## [1] 0.112
```

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

```
## $r.squared
## [1] 0.937
```

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

```
## $r.squared
## [1] 0.954
```

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

```
## $r.squared
## [1] 0.907
```

The employ yields the highest R^2 increase.

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

```
## $r.squared
## [1] 0.955
```

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

```
## $r.squared
## [1] 0.955
```

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

```
## $r.squared
## [1] 0.963
```

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

```
## $r.squared
## [1] 0.954
```

Adding bad or crime or pop yields insignificant explanatory varibles. Therefore stop. The resulting model of the step-up method is $expend = -110.7 + 0.02971 * employ + 0.02686 * lawyers$.

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

```
## $r.squared
## [1] 0.968
```

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

```
## $r.squared
## [1] 0.967
```

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

```
## $r.squared
## [1] 0.964
```

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

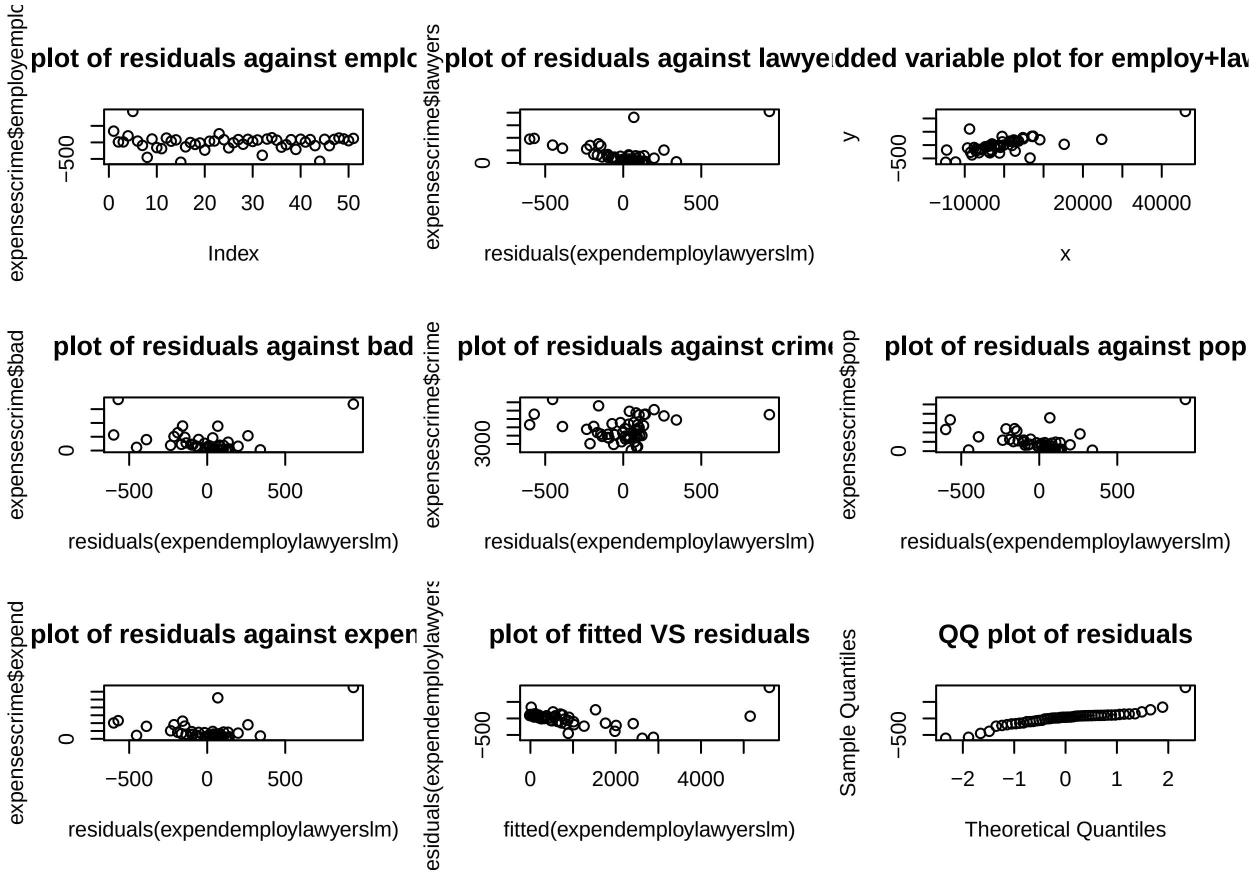
```
## $r.squared
## [1] 0.963
```

In the step-down method, we remove variables whose p-value is larger than 0.05, we stop after remove bad because remaining variables are significant. the model is the same as the model of step-up model.

c)

The model is $expend = -110.7 + 0.02971 * employ + 0.02686 * lawyers$.

```r
expendemploylawyerslm=lm(expend~employ+lawyers,data=expensescrime)
```



From the above plots we can see that spread of the residuals against employ and the residuals against lawyers are alike. The normal Q-Q plot of residuals doesn't show normal distribution. However, plot residuals against Y shows some pattern of decrease. From a), we already know that there is a problem of collinearity between lawyers and employ. Compare with other models, expend~employ has less variables and only a slightly lower value of R-squared.(Only parts of the summary are shown). The better model is $expend = -116.7 + 0.046811 * employ$.

```
vif(expendemploylawyerslm)
```

```
##   employ lawyers
##     14.8    14.8
```

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

```
## $r.squared
## [1] 0.937
```

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

```
## $r.squared
## [1] 0.954
```