Assignment 3

Sara Salimzadeh,

Tung Nguyên Hũu Tùng,

Felipe Santos Batista de Souza

Group 18

March 12

library(multcomp)

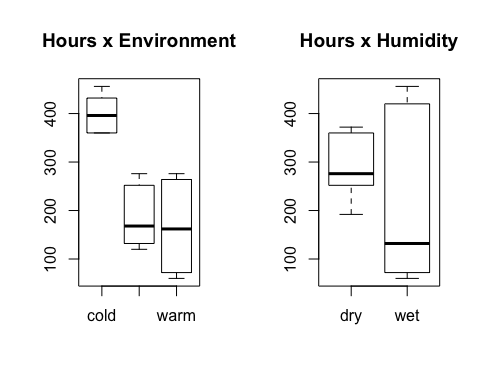
library(lme4)

### Exercise 1

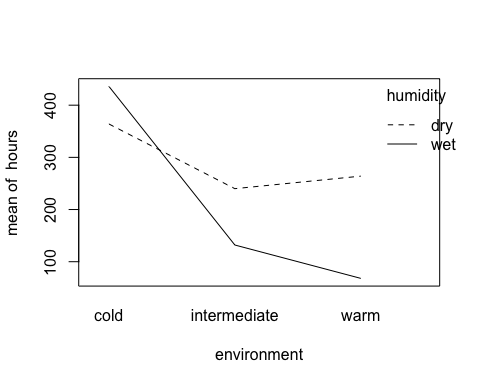
data\_bread <- read.table(file = "bread.txt", header = TRUE)  
  
# Question 1  
I = 3 #levels of temperature.  
J = 2 #levels of humidity.  
N = 3 #experimental units per combination of the two factors, given that the total of units is 18.  
rbind(rep(1:I,each=N\*J),rep(1:J,N\*I),sample(1:(N\*I\*J)))

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

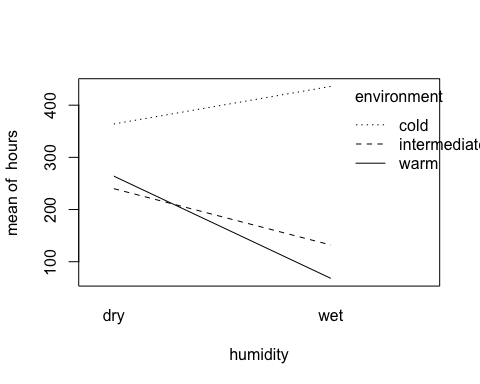
# Question 2  
par(mfrow=c(1,2))



par(mfrow=c(1,1))  
attach(data\_bread)  
interaction.plot(environment,humidity,hours)



interaction.plot(humidity,environment,hours)



# Question 3  
data\_bread$environment=as.factor(data\_bread$environment)  
data\_bread$humidity=as.factor(data\_bread$humidity)  
breadaov=lm(hours~environment\*humidity,data=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

summary(breadaov)

## lm(formula = hours ~ environment \* humidity, data = data\_bread)  
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 364.00 12.00 30.333 1.03e-12  
## environmentintermediate -124.00 16.97 -7.307 9.39e-06  
## environmentwarm -100.00 16.97 -5.893 7.34e-05  
## humiditywet 72.00 16.97 4.243 0.00114  
## environmentintermediate:humiditywet -180.00 24.00 -7.500 7.23e-06  
## environmentwarm:humiditywet -268.00 24.00 -11.167 1.07e-07  
##   
## (Intercept) \*\*\*  
## environmentintermediate \*\*\*  
## environmentwarm \*\*\*  
## humiditywet \*\*   
## environmentintermediate:humiditywet \*\*\*  
## environmentwarm:humiditywet \*\*\*

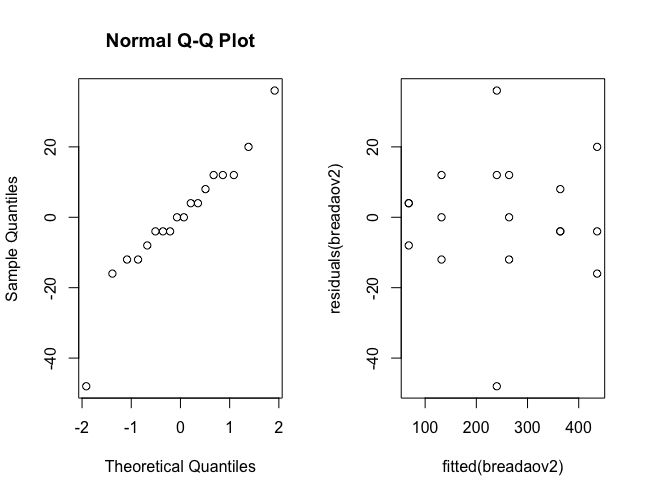
# Question 4  
contrasts(data\_bread$environment)=contr.sum  
contrasts(data\_bread$humidity)=contr.sum  
breadaov2=lm(hours~environment\*humidity,data=data\_bread)  
anova (breadaov2)

## 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

summary(breadaov2)

## lm(formula = hours ~ environment \* humidity, data = data\_bread)  
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 250.667 4.899 51.167 2.04e-15 \*\*\*  
## environment1 149.333 6.928 21.554 5.81e-11 \*\*\*  
## environment2 -64.667 6.928 -9.334 7.50e-07 \*\*\*  
## humidity1 38.667 4.899 7.893 4.32e-06 \*\*\*  
## environment1:humidity1 -74.667 6.928 -10.777 1.59e-07 \*\*\*  
## environment2:humidity1 15.333 6.928 2.213 0.047 \*

# Question 5



# An extra check is also the Shapiro test.  
shapiro.test(residuals(breadaov2))

##   
## Shapiro-Wilk normality test  
##   
## data: residuals(breadaov2)  
## W = 0.9296, p-value = 0.1911

**Exercise 1**

1. The resulting matrix shows a suggested random assignment for the different experiment units, for example: Elements of the matrix: (1,2) =1, (2,2) = 2, (3,2)=7 ==> the suggestion would be then to assign the unit 7 to temperature level 1 and humidity level 2.
2. It is interesting to see through the boxplots that under certain levels for the factors (environment and temperature) there is bigger dispersion on the sampled values for the hours variable, for example when the environment is warm or when the humidity is wet. Moreover, the two boxplots for intermediate and warm environment have overlap. Now looking at the interaction plots it’s possible to see that the lines are non-parallel both when humidity or environment are fixed. This can indicate interaction between these factors.
3. Conclusion: Considering the decomposition of the population mean as:

s per the results from ANOVA:

* The p-value testing : for all i is equal to 0 is 2.461e-10 (reject ).
* The p-value testing : for all j is equal to 0 is 4.316e-06 (reject )
* The p-value testing :for all i, j is equal to 0 is 3.705e-07 (reject )

Thus, it’s possible to say that both factors have a main effect given that the null hypothesis for both of them was rejected (so they are a factor <> 0) in the equation mentioned above. The fact that is also rejected for means that there might be interaction between both factors. As could be seen in the interaction plots from the item 1.2 there is an indication of interaction by the existence of non-parallel lines.

1. As can be seen on the estimations for the factors (environment1: 149.333 and environment2: -64.667), the environment has the biggest numerical influence over the time decay.

This **can’t** be a good question. Since, is also important to notice that, as shown on the results above, that there is an interaction factor between the two factors.

1. *Checking the normality of the population.*

The residuals seem to follow a normal distribution, there is are two extreme-value outliers) though.

An extra check is also the Shapiro test. In this case, we also fail to reject the null hypothesis (that the sample) follows a normal distribution (p-value = 0.1911).

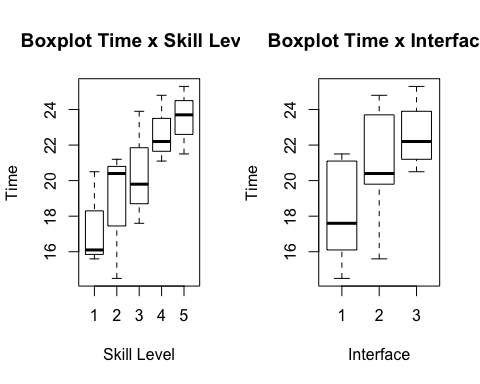
*Checking the assumption of equal population variances.*

As can be seen, there are two data-points that are extreme.

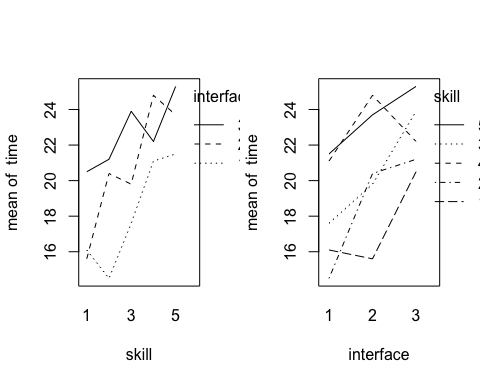
However, in general the residuals remain constant for different fitted values which is in line with the assumption of "of equal population variances".

### Exercise 2  
search = read.table("search.txt", header = TRUE)  
  
# Question 1  
B = 5  
I = 3  
N = 1  
  
rbind(rep(1:I,each=N\*B),rep(1:B,N\*I),sample(1:(N\*I\*B)))

## [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13]  
## [1,] 1 1 1 1 1 2 2 2 2 2 3 3 3  
## [2,] 1 2 3 4 5 1 2 3 4 5 1 2 3  
## [3,] 9 13 11 7 10 12 1 2 3 8 15 14 4  
## [,14] [,15]  
## [1,] 3 3  
## [2,] 4 5  
## [3,] 5 6  
  
# Question 2  
attach(search)  
par(mfrow=c(1,2))



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



# Question 3  
search$skill <- factor(search$skill)  
search$interface <- factor(search$interface)  
aovsearch = lm(time~interface+skill, data = search)  
anova(aovsearch)

## Analysis of Variance Table  
## Response: time  
## Df Sum Sq Mean Sq F value Pr(>F)   
## interface 2 50.465 25.2327 7.8237 0.01310 \*  
## skill 4 80.051 20.0127 6.2052 0.01421 \*  
## Residuals 8 25.801 3.2252

summary(aovsearch)

## lm(formula = time ~ interface + skill, data = search)  
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 15.013 1.227 12.238 1.85e-06 \*\*\*  
## interface2 2.700 1.136 2.377 0.04474 \*   
## interface3 4.460 1.136 3.927 0.00438 \*\*   
## skill2 1.300 1.466 0.887 0.40118   
## skill3 3.033 1.466 2.069 0.07238 .   
## skill4 5.300 1.466 3.614 0.00684 \*\*   
## skill5 6.100 1.466 4.160 0.00316 \*\*   
## F-statistic: 6.745 on 6 and 8 DF, p-value: 0.008395

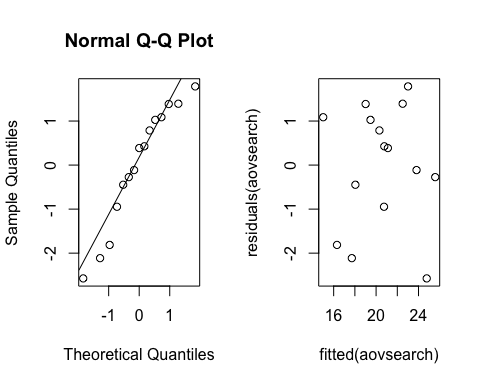
# Question 4  
contrasts(search$skill)=contr.sum  
contrasts(search$interface)=contr.sum  
aovsearch = lm(time~interface+skill, data = search)  
summary(aovsearch)

## lm(formula = time ~ interface + skill, data = search)  
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 20.5467 0.4637 44.311 7.43e-11 \*\*\*  
## interface1 -2.3867 0.6558 -3.640 0.00659 \*\*   
## interface2 0.3133 0.6558 0.478 0.64556   
## skill1 -3.1467 0.9274 -3.393 0.00946 \*\*   
## skill2 -1.8467 0.9274 -1.991 0.08161 .   
## skill3 -0.1133 0.9274 -0.122 0.90575   
## skill4 2.1533 0.9274 2.322 0.04877 \*   
## F-statistic: 6.745 on 6 and 8 DF, p-value: 0.008395

estimatedTime = 20.5467 + 2.1533 + (0 - (-2.3867) - 0.3133)   
estimatedTime

## [1] 24.7734

# Question 5:  
par(mfrow=c(1,2))



# Question 6  
friedman.test(time, interface, skill, data = search)

## Friedman rank sum test  
## data: time, interface and skill  
## Friedman chi-squared = 6.4, df = 2, p-value = 0.04076  
  
# Question 7  
oneaovsearch = lm(time~interface, data = search)  
anova(oneaovsearch)

## Analysis of Variance Table  
## Response: time  
## Df Sum Sq Mean Sq F value Pr(>F)   
## interface 2 50.465 25.233 2.8605 0.09642 .  
## Residuals 12 105.852 8.821

**Exercise 2**

1. As seen from randomized table, unit 3 will use levels (I=1, B=1), unit 7 will use levels (I=1, B=2), …, and unit 9 will use levels (I=3, B= 5).
2. Given the boxplot of time corresponding with skill time alone, we can observe that different levels of skill certainly affect the time. The lower the skill is (higher indicator), the more time is spent. Glancing at the boxplot of time with interface, it is fairly clear that time has large interval when using interface 3.The interaction plots show unparalleled lines so we can suspect the interaction between two factors: interface and skill.
3. By applying anova test to our data, we get the p-value for interface factor equals 0.013, which is less than 0.05. It means that we can reject the null hypothesis that the search time is the same for all interfaces.
4. Using the summary table, we can estimate that it takes 24.773 for a typical user of skill level 4 to find the product on the website if the website uses interface 3.
5. Hereby, the data might approximately follow the normal distribution even though there is a slight curve in the QQ-Plot. The plot nearby includes points which are all over the place. In other words, the plot can be considered diagnostic.
6. The Friedman test has the p-value for testing null hypothesis of “no interface effect” is 0.04076 and therefore we can reject the null hypothesis that there is significant effect of interface.
7. The p-value of one-way anova is 0.09642, which means we cannot reject the null hypothesis that there is no interface effect. This test is only valid if we ignore the skill variable as we assumed before. However, it isn’t too wise to apply this test because the data collected is always influenced by the skill factor. To be valid, the test should be taken by the way that skill factor is isolated, which means we should choose experiment units that have the same skill.

### Exercise 3  
  
# Question 1  
cream= read.table("cream.txt",header = TRUE)  
  
cream$batch= as.factor(cream$batch)  
cream$position= as.factor(cream$position)  
cream$starter= as.factor(cream$starter)  
  
creamaov= lm(acidity ~ starter + batch + position,data = cream)

anova(creamaov)

## Analysis of Variance Table  
##   
## Response: acidity  
## Df Sum Sq Mean Sq F value Pr(>F)   
## starter 4 44.136 11.0340 20.2080 2.904e-05 \*\*\*  
## batch 4 18.778 4.6944 8.5975 0.001632 \*\*   
## position 4 2.348 0.5870 1.0750 0.411191   
## Residuals 12 6.552 0.5460

summary(creamaov)

## lm(formula = acidity ~ starter + batch + position, data = cream)  
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.6616 0.5329 16.255 1.55e-09 \*\*\*  
## starter2 -0.1500 0.4673 -0.321 0.7538   
## starter3 -0.9800 0.4673 -2.097 0.0579 .   
## starter4 2.8100 0.4673 6.013 6.10e-05 \*\*\*  
## starter5 -0.4840 0.4673 -1.036 0.3208   
## batch2 -1.3480 0.4673 -2.884 0.0137 \*   
## batch3 0.2760 0.4673 0.591 0.5658   
## batch4 1.3680 0.4673 2.927 0.0127 \*   
## batch5 0.2000 0.4673 0.428 0.6763   
## position2 -0.6180 0.4673 -1.322 0.2107   
## position3 -0.0380 0.4673 -0.081 0.9365   
## position4 -0.7640 0.4673 -1.635 0.1280   
## position5 -0.2640 0.4673 -0.565 0.5825   
## F-statistic: 9.96 on 12 and 12 DF, p-value: 0.0001777

# Question 2  
creammult= glht(creamaov,linfct = mcp(starter="Tukey"))  
summary(creammult)

## Simultaneous Tests for General Linear Hypotheses  
## Multiple Comparisons of Means: Tukey Contrasts  
## Fit: lm(formula = acidity ~ starter + batch + position, data = cream)  
##   
## Linear Hypotheses:  
## Estimate Std. Error t value Pr(>|t|)   
## 2 - 1 == 0 -0.1500 0.4673 -0.321 0.997367   
## 3 - 1 == 0 -0.9800 0.4673 -2.097 0.282005   
## 4 - 1 == 0 2.8100 0.4673 6.013 0.000497 \*\*\*  
## 5 - 1 == 0 -0.4840 0.4673 -1.036 0.834341   
## 3 - 2 == 0 -0.8300 0.4673 -1.776 0.428894   
## 4 - 2 == 0 2.9600 0.4673 6.334 0.000290 \*\*\*  
## 5 - 2 == 0 -0.3340 0.4673 -0.715 0.949081   
## 4 - 3 == 0 3.7900 0.4673 8.110 < 1e-04 \*\*\*  
## 5 - 3 == 0 0.4960 0.4673 1.061 0.822247   
## 5 - 4 == 0 -3.2940 0.4673 -7.048 0.000109 \*\*\*  
  
  
# Question 4  
confint(creammult)

## Simultaneous Confidence Intervals  
## Multiple Comparisons of Means: Tukey Contrasts  
## Fit: lm(formula = acidity ~ starter + batch + position, data = cream)  
## Linear Hypotheses:  
## Estimate lwr upr   
## 2 - 1 == 0 -0.1500 -1.6401 1.3401  
## 3 - 1 == 0 -0.9800 -2.4701 0.5101  
## 4 - 1 == 0 2.8100 1.3199 4.3001  
## 5 - 1 == 0 -0.4840 -1.9741 1.0061  
## 3 - 2 == 0 -0.8300 -2.3201 0.6601  
## 4 - 2 == 0 2.9600 1.4699 4.4501  
## 5 - 2 == 0 -0.3340 -1.8241 1.1561  
## 4 - 3 == 0 3.7900 2.2999 5.2801  
## 5 - 3 == 0 0.4960 -0.9941 1.9861  
## 5 - 4 == 0 -3.2940 -4.7841 -1.8039

**Exercise 3**

1. The p-values produced with anova and summary commands are not simultaneous. According to anova, starter with p-value 2.904e-05 and batch 0.001 affect the acidity. Considering the summary, the p-values in the lines starter2 to starter5 are for the hypothesis , and so on for the main effect of starter. This is the same for batch and position. According to the summary, starter4 has the least p-value equals to 6.10e-05. Therefore, we can conclude that the null hypothesis is strongly rejected and starter4 plays a significant role in the main model of the acidity. Similarly, bath2 and batch4 null hypothesis can be rejected as well with p-values equal to 0.0137 and 0.0127 respectively. The abstract model is: . To sum up, equals to for i = 2,3,5 (except ). For j = 2 and 4, is equal to (excluding . Finally, all in range (2,5) is the same as .
2. Here the table is based on simultaneous p-values for the null hypothesis. As it is clear from the table, starter4 has the most difference with others and significantly affects the acidity. P-values for cases, , , and, are less than 0.001 hence null hypothesis are rejected.
3. The p-value of second sector is 0.99 while it is equal to 0.75 in part 1. In simultaneous comparisons, the more inferences are made, the more likely erroneous inferences are to occur. Consequently, the p-value in part 2 is less trusted than in part 1 because of the error and its greater because of error propagation.
4. The intervals for , ,and, don’t contain 0. As we concluded in part 2, the starter4 lead to significantly different acidity. So, the confidence intervals shouldn’t include 0 to indicate the difference of starter4 from others.

### Exercise 4  
  
# Question 1  
cow = read.table("cow.txt", header = TRUE)  
cow$id = factor(cow$id)  
cow$per = factor(cow$per)  
cow$treatment = factor(cow$treatment)  
  
cowlm = lm(milk~treatment+id+per, data = cow)  
summary(cowlm)

## lm(formula = milk ~ treatment + id + per, data = cow)  
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 30.3000 1.2444 24.349 5.02e-08 \*\*\*  
## treatmentB -0.5100 0.7466 -0.683 0.516536   
## id2 23.0000 1.5741 14.612 1.68e-06 \*\*\*  
## id3 11.1500 1.5741 7.084 0.000196 \*\*\*  
## id4 -1.3500 1.5741 -0.858 0.419480   
## id5 -7.0500 1.5741 -4.479 0.002870 \*\*   
## id6 23.4500 1.5741 14.898 1.47e-06 \*\*\*  
## id7 13.5500 1.5741 8.608 5.69e-05 \*\*\*  
## id8 4.9000 1.5741 3.113 0.017011 \*   
## id9 -11.2000 1.5741 -7.115 0.000191 \*\*\*  
## per2 -2.3900 0.7466 -3.201 0.015046 \*   
## F-statistic: 100.6 on 10 and 7 DF, p-value: 1.349e-06

# Question 2 Result from question 1  
  
# Question 3  
cowlmer = lmer(milk~treatment+order+per+(1|id), data = cow, REML=FALSE)  
summary(cowlmer)

## Linear mixed model fit by maximum likelihood ['lmerMod']  
## Formula: milk ~ treatment + order + per + (1 | id)  
## Data: cow  
## Random effects:  
## Groups Name Variance Std.Dev.  
## id (Intercept) 133.145 11.539   
## Residual 1.927 1.388   
## Number of obs: 18, groups: id, 9  
##   
## Fixed effects:  
## Estimate Std. Error t value  
## (Intercept) 38.5000 5.8110 6.625  
## treatmentB -0.5100 0.6585 -0.775  
## orderBA -3.4700 7.7685 -0.447  
## per2 -2.3900 0.6585 -3.630  
##   
## Correlation of Fixed Effects:  
## (Intr) trtmnB ordrBA  
## treatmentB -0.063   
## orderBA -0.743 0.000   
## per2 -0.063 0.111 0.000

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

## 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 5 117.89 122.34 -53.946 107.89   
## cowlmer 6 119.31 124.65 -53.656 107.31 0.5807 1 0.446

# Question 4  
attach(cow)  
t.test(milk[treatment=="A"],milk[treatment=="B"],paired=TRUE)

## Paired t-test  
## data: milk[treatment == "A"] and milk[treatment == "B"]  
## t = 0.22437, df = 8, p-value = 0.8281  
## alternative hypothesis: true difference in means is not equal to 0  
## sample estimates:  
## mean of the differences   
## 0.2444444

**Exercise 4**

1. The effect of the type of feedingstuffs is not significant with the p-value = 0.516 > 0.05. In other words, we cannot reject the null hypothesis of “no effect of type of feeding stuffs”
2. TreamentA is estimated to produce 0.51 more milk than TreatmentB
3. With p-value = 0.446 > 0.05 from anova test, we can draw the same conclusion as the first section that we cannot reject the null hypothesis that there is no effect of type of feeding stuffs on milk production.
4. The pair t-test is invalid since it only tests whether there is difference between the true mean of the paired samples due to 2 different treatments. Nonetheless, we cannot ignore fixed period effect and fixed sequence effect in our current experiment, which in turn leads to the less accuracy of paired-test. The conclusion in t-test, which does not reject the null hypothesis of no effect of Treatment), is fairly compatible with the test in the question 1 (p-value equals 0.516536 and 0.8281 respectively). However, as mentioned before the pair t-test lacks of estimating other effects except “Treatment), we cannot trust the result from this test.

### Exercise 5  
  
# Question 1  
nausea.frame=data.frame("nausea" = integer(),"medicine" = character(), stringsAsFactors = FALSE)  
index = 1  
for(i in 1:100){  
 nausea.frame[i,] <- rbind(0, "Chlorpromazine")  
 index = index + 1}  
  
for(i in 1:52){  
 nausea.frame[index,] <- rbind(1, "Chlorpromazine")  
 index = index + 1}  
  
for(i in 1:32){  
 nausea.frame[index,] <- rbind(0, "Pentobarbital(100mg)")  
 index = index + 1}  
  
for(i in 1:35){  
 nausea.frame[index,] <- rbind(1, "Pentobarbital(100mg)")  
 index = index + 1}  
  
for(i in 1:48){  
 nausea.frame[index,] <- rbind(0, "Pentobarbital(150mg)")  
 index = index + 1}  
  
for(i in 1:37){  
 nausea.frame[index,] <- rbind(1, "Pentobarbital(150mg)")  
 index = index + 1}  
attach(nausea.frame)

# Question 2  
nausea.frame$medicine=as.factor(nausea.frame$medicine)  
  
xtabs(~medicine+nausea)

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

# Question 3  
attach(nausea.frame)

t = chisq.test(xtabs(~medicine+nausea))[[1]]  
B=1000  
tstar=numeric(B)  
for (i in 1:B){  
 medicinestar=sample(medicine)  
 tstar[i]= chisq.test(xtabs(~medicinestar+nausea))[[1]]}  
t  
## X-squared   
## 6.624765

pl=sum(tstar<t)/B  
pr=sum(tstar>t)/B  
p=2\*min(pl,pr)  
pl;pr;p

## [1] 0.971

## [1] 0.027

## [1] 0.054

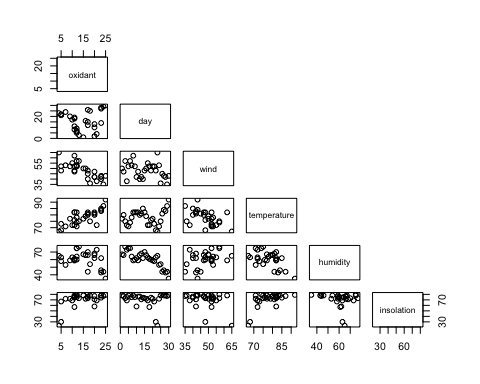
# Question 4  
chisq.test(xtabs(~medicine+nausea))

## Pearson's Chi-squared test  
## data: xtabs(~medicin + nausea)  
## X-squared = 6.6248, df = 2, p-value = 0.03643

**Exercise 5**

1. R Script presented above.
2. The outcome of xtabs (~ medicine + nausea) presents the data contained in the data frame in an aggregated view in which the different factors (in this case medicines) are presented in different rows and their respective outcomes (aggregated) are presented in the columns, in this case one column with the number of patients with or without nausea. It’s possible to see from xtabs that Chlorpromazine had the biggest number of samples (patients) compared to the other medicines. It also had the best performance (when considering the number of patients without nausea after the treatment).
3. Given that the value for p-value = 0.054 we can’t reject the null hypothesis that the populations are the same, in other words, there might not be a difference between the different treatments (bear in mind that the p-value can change for each computation which can lead to a different conclusion about rejecting or not the null hypothesis).
4. The p-value calculated from chi-square test is different from the value calculated by the permutation test. The p-value from chi-square would very unlikely be the same as the p-value from the permutation test, since the last is obtained from a random permutation (without replacement) of the labels which in turn leads to slightly different p-values at every computation, but the p-values could be the same if indeed the populations were exactly the same (which is the null hypothesis of this test that was rejected in the item 5.3 – above).

### Exercise 6  
  
# Question 1  
airpollution= read.table("airpollution.txt", header = TRUE)  
pairs(oxidant ~ day + wind + temperature + humidity + insolation , data= airpollution, upper.panel= NULL)



# Question 2  
oxidant\_day= lm(oxidant ~ day, data = airpollution)  
summary(oxidant\_day)

## lm(formula = oxidant ~ day, data = airpollution)  
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13.68966 2.28580 5.989 1.89e-06 \*\*\*  
## day 0.07164 0.12876 0.556 0.582   
## Residual standard error: 6.104 on 28 degrees of freedom  
## Multiple R-squared: 0.01093, Adjusted R-squared: -0.02439

#Multiple R-squared: 0.01093  
  
oxidant\_wind= lm(oxidant ~ wind, data = airpollution)  
summary(oxidant\_wind)

## lm(formula = oxidant ~ wind, data = airpollution)  
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 45.3171 4.8976 9.253 5.19e-10 \*\*\*  
## wind -0.6331 0.1005 -6.300 8.20e-07 \*\*\*  
## Multiple R-squared: 0.5863, Adjusted R-squared: 0.5715

#Multiple R-squared: 0.5863  
  
oxidant\_temperature=lm(oxidant ~ temperature, data = airpollution)  
summary(oxidant\_temperature)

## lm(formula = oxidant ~ temperature, data = airpollution)  
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -46.4292 9.9542 -4.664 6.94e-05 \*\*\*  
## temperature 0.7850 0.1273 6.168 1.17e-06 \*\*\*  
## Multiple R-squared: 0.576, Adjusted R-squared: 0.5609

#Multiple R-squared: 0.576  
  
oxidant\_humidity=lm(oxidant ~ humidity, data = airpollution)  
summary(oxidant\_humidity)

## lm(formula = oxidant ~ humidity, data = airpollution)  
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 27.4446 6.4368 4.264 0.000206 \*\*\*  
## humidity -0.2088 0.1049 -1.991 0.056317 .   
## Multiple R-squared: 0.124, Adjusted R-squared: 0.09273

#Multiple R-squared: 0.124  
  
oxidant\_insolation=lm(oxidant ~ insolation, data = airpollution)  
summary(oxidant\_insolation)

## lm(formula = oxidant ~ insolation, data = airpollution)  
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.43279 5.32967 -0.269 0.79003   
## insolation 0.22993 0.07424 3.097 0.00441 \*\*  
## Multiple R-squared: 0.2552, Adjusted R-squared: 0.2286

#Multiple R-squared: 0.2552  
  
## step 1 - select the highest R-squared (wind)  
oxidant1= lm(oxidant ~ wind, data = airpollution)  
summary(oxidant1)

## lm(formula = oxidant ~ wind, data = airpollution)  
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 45.3171 4.8976 9.253 5.19e-10 \*\*\*  
## wind -0.6331 0.1005 -6.300 8.20e-07 \*\*\*  
## Multiple R-squared: 0.5863, Adjusted R-squared: 0.5715

#Multiple R-squared: 0.5863

anova(oxidant1)

## Analysis of Variance Table  
## Response: oxidant  
## Df Sum Sq Mean Sq F value Pr(>F)   
## wind 1 618.45 618.45 39.684 8.205e-07 \*\*\*  
## Residuals 28 436.35 15.58   
  
## step 2 - select the highest R-squared (temperature as second variable) among below options  
oxidant2=lm(oxidant ~ wind + temperature, data = airpollution)  
summary(oxidant2)

## lm(formula = oxidant ~ wind + temperature, data = airpollution)  
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -5.20334 11.11810 -0.468 0.644   
## wind -0.42706 0.08645 -4.940 3.58e-05 \*\*\*  
## temperature 0.52035 0.10813 4.812 5.05e-05 \*\*\*  
## Multiple R-squared: 0.7773, Adjusted R-squared: 0.7608

#Multiple R-squared:0.7773   
  
oxidant2=lm(oxidant ~ wind + insolation, data = airpollution)  
summary(oxidant2)

## lm(formula = oxidant ~ wind + insolation, data = airpollution)  
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 32.32615 6.97098 4.637 8.07e-05 \*\*\*  
## wind -0.55639 0.09778 -5.690 4.81e-06 \*\*\*  
## insolation 0.13161 0.05383 2.445 0.0213 \*   
## Multiple R-squared: 0.6613, Adjusted R-squared: 0.6362

#Multiple R-squared:0.6613   
  
oxidant2=lm(oxidant ~ wind + humidity, data = airpollution)  
summary(oxidant2)

## lm(formula = oxidant ~ wind + humidity, data = airpollution)  
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 46.91570 5.68573 8.251 7.38e-09 \*\*\*  
## wind -0.60955 0.10971 -5.556 6.86e-06 \*\*\*  
## humidity -0.04516 0.07866 -0.574 0.571   
## Multiple R-squared: 0.5913, Adjusted R-squared: 0.561

#Multiple R-squared:0.5913 - one variable become insignificant  
  
oxidant2=lm(oxidant ~ wind + day, data = airpollution)  
summary(oxidant2)

## lm(formula = oxidant ~ wind + day, data = airpollution)  
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 47.84224 5.62785 8.501 4.10e-09 \*\*\*  
## wind -0.65984 0.10489 -6.291 9.87e-07 \*\*\*  
## day -0.07986 0.08691 -0.919 0.366   
## Multiple R-squared: 0.5989, Adjusted R-squared: 0.5691

#Multiple R-squared:0.5989 - one variable become insignificant

#selected  
oxidant2=lm(oxidant ~ wind + temperature, data = airpollution)  
anova(oxidant2)

## Analysis of Variance Table  
## Response: oxidant  
## Df Sum Sq Mean Sq F value Pr(>F)   
## wind 1 618.45 618.45 71.087 4.823e-09 \*\*\*  
## temperature 1 201.46 201.46 23.156 5.047e-05 \*\*\*  
## Residuals 27 234.90 8.70

## step 3 - The added variable makes one variable insignificant - stop stepping up

oxidant3=lm(oxidant ~ wind + temperature + insolation, data = airpollution)  
summary(oxidant3)

## lm(formula = oxidant ~ wind + temperature + insolation, data = airpollution)  
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4.45496 11.26714 -0.395 0.695778   
## wind -0.42353 0.08737 -4.848 5.02e-05 \*\*\*  
## temperature 0.47558 0.12564 3.785 0.000816 \*\*\*  
## insolation 0.03646 0.05071 0.719 0.478636   
## Multiple R-squared: 0.7816, Adjusted R-squared: 0.7565

#Multiple R-squared:0.7816 - one variable become insignificant

anova(oxidant3)

## Analysis of Variance Table  
## Response: oxidant  
## Df Sum Sq Mean Sq F value Pr(>F)   
## wind 1 618.45 618.45 69.8143 7.753e-09 \*\*\*  
## temperature 1 201.46 201.46 22.7418 6.186e-05 \*\*\*  
## insolation 1 4.58 4.58 0.5168 0.4786   
## Residuals 26 230.32 8.86   
  
oxidant3=lm(oxidant ~ wind + temperature + humidity, data = airpollution)  
summary(oxidant3)

## lm(formula = oxidant ~ wind + temperature + humidity, data = airpollution)  
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -16.60697 13.07154 -1.270 0.215   
## wind -0.44620 0.08513 -5.241 1.78e-05 \*\*\*  
## temperature 0.60190 0.11764 5.117 2.47e-05 \*\*\*  
## humidity 0.09850 0.06316 1.559 0.131   
## Multiple R-squared: 0.7964, Adjusted R-squared: 0.7729

#Multiple R-squared:0.7964 - one variable become insignificant

anova(oxidant3)

## Analysis of Variance Table  
## Response: oxidant  
## Df Sum Sq Mean Sq F value Pr(>F)   
## wind 1 618.45 618.45 74.8561 3.947e-09 \*\*\*  
## temperature 1 201.46 201.46 24.3842 3.957e-05 \*\*\*  
## humidity 1 20.09 20.09 2.4317 0.131   
## Residuals 26 214.81 8.26   
  
oxidant3=lm(oxidant ~ wind + temperature + day, data = airpollution)  
summary(oxidant3)

## lm(formula = oxidant ~ wind + temperature + day, data = airpollution)  
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.98987 10.94466 -0.273 0.787   
## wind -0.45604 0.08644 -5.276 1.63e-05 \*\*\*  
## temperature 0.52918 0.10568 5.008 3.29e-05 \*\*\*  
## day -0.09711 0.06328 -1.535 0.137   
## Multiple R-squared: 0.7958, Adjusted R-squared: 0.7722

#Multiple R-squared:0.7958 - one variable become insignificant

anova(oxidant3)

## Analysis of Variance Table  
## Response: oxidant  
## Df Sum Sq Mean Sq F value Pr(>F)   
## wind 1 618.45 618.45 74.654 4.053e-09 \*\*\*  
## temperature 1 201.46 201.46 24.318 4.028e-05 \*\*\*  
## day 1 19.51 19.51 2.355 0.137   
## Residuals 26 215.39 8.28   
  
## Best Model - step 2  
oxidant2

## lm(formula = oxidant ~ wind + temperature, data = airpollution)  
##   
## Coefficients:  
## (Intercept) wind temperature   
## -5.2033 -0.4271 0.5204

# Question 3  
## step 1 - full model  
oxidant5= lm(oxidant ~ wind + temperature + insolation + humidity + day, data = airpollution)  
summary(oxidant5)

## lm(formula = oxidant ~ wind + temperature + insolation + humidity +   
## day, data = airpollution)  
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -12.04010 21.20961 -0.568 0.57553   
## wind -0.44749 0.09103 -4.916 5.14e-05 \*\*\*  
## temperature 0.55714 0.15347 3.630 0.00133 \*\*   
## insolation 0.01822 0.05583 0.326 0.74694   
## humidity 0.06818 0.13336 0.511 0.61384   
## day -0.02997 0.13995 -0.214 0.83227   
## F-statistic: 19.01 on 5 and 24 DF, p-value: 1.203e-07

## step 2 - remove the highest p-value (day)  
oxidant4= lm(oxidant ~ wind + temperature + insolation + humidity , data = airpollution)  
summary(oxidant4)

## lm(formula = oxidant ~ wind + temperature + insolation + humidity,   
## data = airpollution)  
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -15.49370 13.50647 -1.147 0.26219   
## wind -0.44291 0.08678 -5.104 2.85e-05 \*\*\*  
## temperature 0.56933 0.13977 4.073 0.00041 \*\*\*  
## insolation 0.02275 0.05067 0.449 0.65728   
## humidity 0.09292 0.06535 1.422 0.16743   
## F-statistic: 24.69 on 4 and 25 DF, p-value: 2.279e-08

## step 3 - remove second highest p-value (insolation)  
oxidant3= lm(oxidant ~ wind + temperature + humidity , data = airpollution)  
summary(oxidant3)

## lm(formula = oxidant ~ wind + temperature + humidity, data = airpollution)  
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -16.60697 13.07154 -1.270 0.215   
## wind -0.44620 0.08513 -5.241 1.78e-05 \*\*\*  
## temperature 0.60190 0.11764 5.117 2.47e-05 \*\*\*  
## humidity 0.09850 0.06316 1.559 0.131   
## F-statistic: 33.89 on 3 and 26 DF, p-value: 3.904e-09

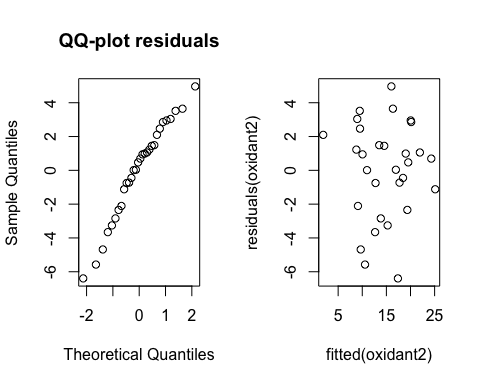
## step 4 - remove third highest p-value (humidity)  
oxidant2= lm(oxidant ~ wind + temperature , data = airpollution)  
summary(oxidant2)

## lm(formula = oxidant ~ wind + temperature, data = airpollution)  
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -5.20334 11.11810 -0.468 0.644   
## wind -0.42706 0.08645 -4.940 3.58e-05 \*\*\*  
## temperature 0.52035 0.10813 4.812 5.05e-05 \*\*\*  
## F-statistic: 47.12 on 2 and 27 DF, p-value: 1.563e-09

# Question 4  
## Best Model  
oxidant2

## lm(formula = oxidant ~ wind + temperature, data = airpollution)  
## Coefficients:  
## (Intercept) wind temperature   
## -5.2033 -0.4271 0.5204

# Question 5  
par(mfrow=c(1,2))



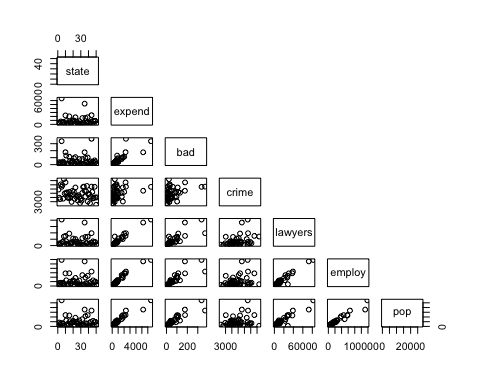
**Exercise 6**

1. According to scatterplots, wind and temperature have the most influence on oxidant. Their plots approximately show linear correlation. Afterward, insolation and humidity could affect oxidant linearly.
2. Among all explanatory variables, wind has the highest R-squared. (0.5863). Therefore, it should be selected as the basis to start to find linear regression model. In step 2, temperature is chosen to be added to the model with R-squared 0.7773 (highest among others). Adding third variable to the model in step 3 makes one variable of previous model insignificant. So, we should stop extending the model. Finally, the appropriate model consists of two explanatory variables which are wind and temperature. To investigate whether the extensions are useful, we apply ANOVA test after finding best model in each step (all p-values are less than 5%). In final step, we use this test for all possible options to prove that newly added variable has the p-value greater than 5% and we should stop extending model.
3. In the first and second and third step, day, insolation and humidity are removed from the model respectively due to their p-values. In this part, similarly we reach to the same model as in previous part. This model is: oxidant ~ wind + temperature
4. Wind and temperature equal to -0.427and 0.52 respectively.
5. To investigate the normality of residuals, we consider the QQ-plot. We can assume normality based on the graph. Moreover, the samples are scattered approximately in whole area according to the fitted-residuals plot.

### Exercise 7  
library(multcomp)

library(lme4)

crime\_expenses = read.table("expensescrime.txt", header = TRUE)  
  
#Identifying potential correlations  
pairs(crime\_expenses, upper.panel=NULL)



# Studying the collinearity between two variables:  
round(cor(crime\_expenses[,3:7]),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

# As can be seen, there is a strong correlation between employ and lawyers, so the model shouldn't contain both factors at the same time.  
  
attach(crime\_expenses)  
  
# step-up strategy  
## 1st Iteration  
expenseslm\_su = lm(expend~bad, data=crime\_expenses)  
summary(expenseslm\_su) # R-squared = 0.6902

## lm(formula = expend ~ bad, data = crime\_expenses)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2744.80 -129.97 -69.01 91.78 2738.99   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 126.721 114.865 1.103 0.275   
## bad 13.324 1.257 10.601 2.8e-14 \*\*\*  
## Multiple R-squared: 0.6964, Adjusted R-squared: 0.6902   
## F-statistic: 112.4 on 1 and 49 DF, p-value: 2.796e-14

expenseslm\_su = lm(expend~lawyers, data=crime\_expenses)  
summary(expenseslm\_su) # R-squared = 0.936

## lm(formula = expend ~ lawyers, data = crime\_expenses)  
## Residuals:  
## Min 1Q Median 3Q Max   
## -1504.24 -26.77 36.66 95.05 827.01   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -59.611979 53.799425 -1.108 0.273   
## lawyers 0.070385 0.002601 27.060 <2e-16 \*\*\*  
## Multiple R-squared: 0.9373, Adjusted R-squared: 0.936   
## F-statistic: 732.2 on 1 and 49 DF, p-value: < 2.2e-16

expenseslm\_su = lm(expend~employ, data=crime\_expenses)  
summary(expenseslm\_su) # R-squared = 0.953 ==> Highest factor

## lm(formula = expend ~ employ, data = crime\_expenses)  
## Residuals:  
## Min 1Q Median 3Q Max   
## -636.04 -84.35 47.60 107.99 1124.70   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.167e+02 4.706e+01 -2.48 0.0166 \*   
## employ 4.681e-02 1.469e-03 31.87 <2e-16 \*\*\*  
## Multiple R-squared: 0.954, Adjusted R-squared: 0.953   
## F-statistic: 1016 on 1 and 49 DF, p-value: < 2.2e-16

expenseslm\_su = lm(expend~pop, data=crime\_expenses)  
summary(expenseslm\_su) # R-squared = 0.9054

## lm(formula = expend ~ pop, data = crime\_expenses)  
## Residuals:  
## Min 1Q Median 3Q Max   
## -1144.59 -153.41 35.38 137.34 1537.41   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.888e+02 6.968e+01 -2.709 0.00927 \*\*   
## pop 2.172e-01 9.916e-03 21.903 < 2e-16 \*\*\*  
## Multiple R-squared: 0.9073, Adjusted R-squared: 0.9054   
## F-statistic: 479.7 on 1 and 49 DF, p-value: < 2.2e-16

## 2nd Iteration  
expenseslm\_su = lm(expend~employ+bad, data=crime\_expenses)  
summary(expenseslm\_su) # R-squared = 0.9532 !! this combination of factors doesn’t contribute to R squared, also high p-value

## lm(formula = expend ~ employ + bad, data = crime\_expenses)  
## Residuals:  
## Min 1Q Median 3Q Max   
## -655.96 -99.41 37.23 101.23 1154.08   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.164e+02 4.697e+01 -2.479 0.0167 \*   
## employ 4.966e-02 2.986e-03 16.630 <2e-16 \*\*\*  
## bad -1.090e+00 9.948e-01 -1.095 0.2788   
## Multiple R-squared: 0.9551, Adjusted R-squared: 0.9532   
## F-statistic: 510.5 on 2 and 48 DF, p-value: < 2.2e-16

expenseslm\_su = lm(expend~employ+lawyers, data=crime\_expenses)  
summary(expenseslm\_su) # R-squared = 0.9616 ==> Highest factor on the iteration

## lm(formula = expend ~ employ + lawyers, data = crime\_expenses)  
## Residuals:  
## Min 1Q Median 3Q Max   
## -599.47 -94.43 36.01 91.98 936.55   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.107e+02 4.257e+01 -2.600 0.01236 \*   
## employ 2.971e-02 5.114e-03 5.810 4.89e-07 \*\*\*  
## lawyers 2.686e-02 7.757e-03 3.463 0.00113 \*\*   
## Multiple R-squared: 0.9632, Adjusted R-squared: 0.9616   
## F-statistic: 627.7 on 2 and 48 DF, p-value: < 2.2e-16

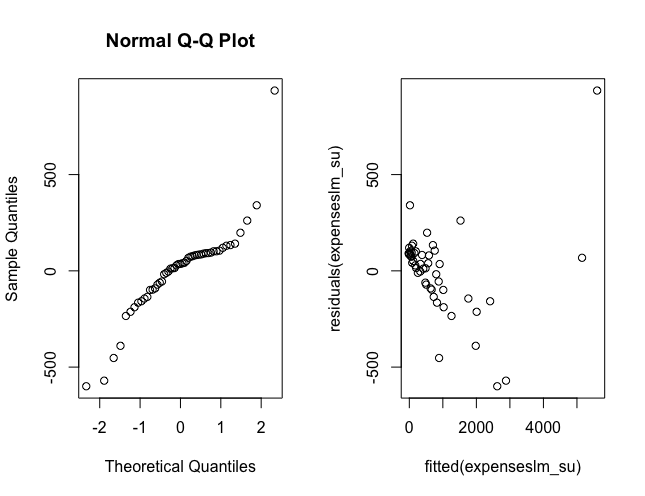
expenseslm\_su = lm(expend~employ+pop, data=crime\_expenses)  
summary(expenseslm\_su) # R-squared = 0.9524 !! this combination of factors doesn’t contribute to R squared, also high p-value

## lm(formula = expend ~ employ + pop, data = crime\_expenses)  
## Residuals:  
## Min 1Q Median 3Q Max   
## -685.73 -93.48 44.54 112.94 1072.78   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.266e+02 5.022e+01 -2.521 0.0151 \*   
## employ 4.326e-02 6.158e-03 7.026 6.72e-09 \*\*\*  
## pop 1.739e-02 2.930e-02 0.594 0.5555   
## Multiple R-squared: 0.9543, Adjusted R-squared: 0.9524   
## F-statistic: 501.3 on 2 and 48 DF, p-value: < 2.2e-16

# Re-executing the highest R-squared for analysis  
expenseslm\_su = lm(expend~employ+lawyers, data=crime\_expenses)  
summary(expenseslm\_su)

## lm(formula = expend ~ employ + lawyers, data = crime\_expenses)  
## Residuals:  
## Min 1Q Median 3Q Max   
## -599.47 -94.43 36.01 91.98 936.55   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.107e+02 4.257e+01 -2.600 0.01236 \*   
## employ 2.971e-02 5.114e-03 5.810 4.89e-07 \*\*\*  
## lawyers 2.686e-02 7.757e-03 3.463 0.00113 \*\*   
## Multiple R-squared: 0.9632, Adjusted R-squared: 0.9616   
## F-statistic: 627.7 on 2 and 48 DF, p-value: < 2.2e-16

# Graphical analysis



shapiro.test(residuals(expenseslm\_su)) # The p-value led us to reject the null hypothesis

## Shapiro-Wilk normality test  
## data: residuals(expenseslm\_su)  
## W = 0.8475, p-value = 1.118e-05

## 3rd Iteration (adding bad since this factor had the highest coeficient, removing lawyers since employ and lawyers are correlated)  
expenseslm\_su = lm(expend~employ+bad, data=crime\_expenses)  
summary(expenseslm\_su) #R squared = 0.9532

## lm(formula = expend ~ employ + bad, data = crime\_expenses)  
## Residuals:  
## Min 1Q Median 3Q Max   
## -655.96 -99.41 37.23 101.23 1154.08   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.164e+02 4.697e+01 -2.479 0.0167 \*   
## employ 4.966e-02 2.986e-03 16.630 <2e-16 \*\*\*  
## bad -1.090e+00 9.948e-01 -1.095 0.2788   
## Multiple R-squared: 0.9551, Adjusted R-squared: 0.9532   
## F-statistic: 510.5 on 2 and 48 DF, p-value: < 2.2e-16

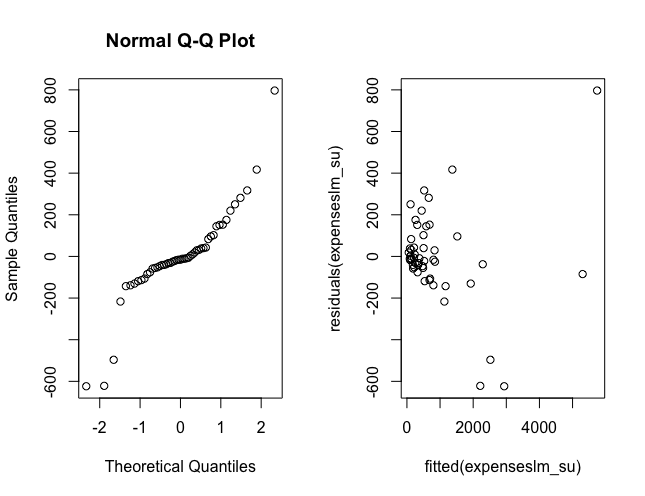
expenseslm\_su = lm(expend~employ+pop, data=crime\_expenses)  
summary(expenseslm\_su) #this combination of factors doesn’t contribute to R squared

## lm(formula = expend ~ employ + pop, data = crime\_expenses)  
## Residuals:  
## Min 1Q Median 3Q Max   
## -685.73 -93.48 44.54 112.94 1072.78   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.266e+02 5.022e+01 -2.521 0.0151 \*   
## employ 4.326e-02 6.158e-03 7.026 6.72e-09 \*\*\*  
## pop 1.739e-02 2.930e-02 0.594 0.5555   
## Multiple R-squared: 0.9543, Adjusted R-squared: 0.9524   
## F-statistic: 501.3 on 2 and 48 DF, p-value: < 2.2e-16

## 4th Iteration - After trying the previous options, bad^2 is added to the model - This will generate the best R-squared !  
crime\_expenses$bad2 = crime\_expenses$bad^2  
expenseslm\_su = lm(expend~employ+bad+bad2, data=crime\_expenses)  
summary(expenseslm\_su) #R squared 0.9634

## lm(formula = expend ~ employ + bad + bad2, data = crime\_expenses)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -623.59 -56.79 -15.77 62.87 796.72   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -5.937628 50.814988 -0.117 0.907478   
## employ 0.055576 0.003071 18.097 < 2e-16 \*\*\*  
## bad -7.885913 2.001834 -3.939 0.000270 \*\*\*  
## bad2 0.016219 0.004290 3.780 0.000442 \*\*\*  
## Multiple R-squared: 0.9656, Adjusted R-squared: 0.9634   
## F-statistic: 439.3 on 3 and 47 DF, p-value: < 2.2e-16

# Graphical analysis from the 4th Iteration (last improvement on R squared)



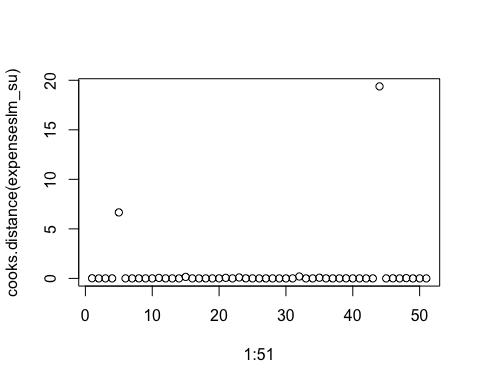
shapiro.test(residuals(expenseslm\_su)) # The p-value still led us to reject the null hypothesis (that residuals follow a normal dist.)

## Shapiro-Wilk normality test  
## data: residuals(expenseslm\_su)  
## W = 0.85613, p-value = 1.932e-05

#Checking for potential/influence points:  
round(cooks.distance(expenseslm\_su),2)

## 1 2 3 4 5 6 7 8 9 10 11 12   
## 0.01 0.00 0.00 0.00 6.66 0.00 0.00 0.00 0.00 0.00 0.05 0.00   
## 13 14 15 16 17 18 19 20 21 22 23 24   
## 0.00 0.00 0.15 0.00 0.00 0.00 0.00 0.00 0.06 0.00 0.09 0.01   
## 25 26 27 28 29 30 31 32 33 34 35 36   
## 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.19 0.00 0.00 0.07 0.00   
## 37 38 39 40 41 42 43 44 45 46 47 48   
## 0.00 0.00 0.01 0.00 0.00 0.00 0.00 19.38 0.00 0.01 0.00 0.03   
## 49 50 51   
## 0.00 0.00 0.00

plot(1:51,cooks.distance(expenseslm\_su))



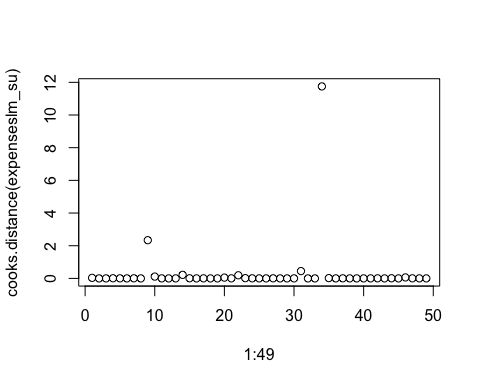
#As can be seen there are two potential points.  
new\_crime\_expenses <- crime\_expenses[ which(crime\_expenses$state != 'TX'),]  
new\_crime\_expenses <- new\_crime\_expenses[ which(new\_crime\_expenses$state != 'CA'),]  
detach(crime\_expenses)  
attach(new\_crime\_expenses)  
  
# 5th Iteration - after removing the potential point.  
new\_crime\_expenses$bad2 = new\_crime\_expenses$bad^2  
expenseslm\_su = lm(expend~employ+bad+bad2, data=new\_crime\_expenses)  
summary(expenseslm\_su) #Contributes ==> R squared moves from 0.9634 to 0.9704, it's possible to say that those points were influence points, given the

## lm(formula = expend ~ employ + bad + bad2, data = new\_crime\_expenses)  
## Residuals:  
## Min 1Q Median 3Q Max   
## -393.85 -61.21 -25.19 51.45 377.78   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 39.830469 39.459692 1.009 0.31818   
## employ 0.044580 0.002485 17.941 < 2e-16 \*\*\*  
## bad -5.339659 1.646717 -3.243 0.00223 \*\*   
## bad2 0.026952 0.009429 2.858 0.00643 \*\*   
## Multiple R-squared: 0.9722, Adjusted R-squared: 0.9704   
## F-statistic: 524.9 on 3 and 45 DF, p-value: < 2.2e-16

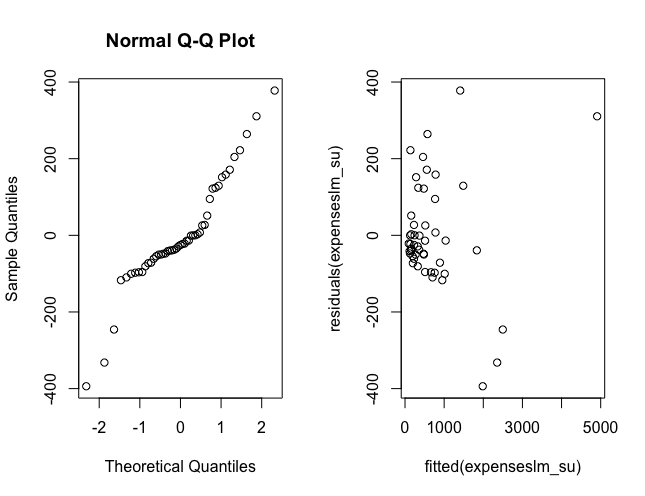
#changes in the estimated parameters.  
  
#Checking again for new potential/influence points:  
round(cooks.distance(expenseslm\_su),2)

## 1 2 3 4 6 7 8 9 10 11 12 13   
## 0.03 0.00 0.00 0.01 0.00 0.00 0.01 0.00 2.34 0.11 0.00 0.00   
## 14 15 16 17 18 19 20 21 22 23 24 25   
## 0.00 0.22 0.01 0.00 0.00 0.00 0.00 0.05 0.00 0.19 0.02 0.00   
## 26 27 28 29 30 31 32 33 34 35 36 37   
## 0.00 0.00 0.01 0.00 0.00 0.00 0.45 0.00 0.00 11.75 0.02 0.00   
## 38 39 40 41 42 43 45 46 47 48 49 50   
## 0.01 0.00 0.00 0.00 0.00 0.00 0.00 0.01 0.00 0.06 0.00 0.00   
## 51   
## 0.00

plot(1:49,cooks.distance(expenseslm\_su))



# Graphical analysis



shapiro.test(residuals(expenseslm\_su)) # The p-value calculated via Shapiro improves but still led us to reject the null hypothesis

## Shapiro-Wilk normality test  
## data: residuals(expenseslm\_su)  
## W = 0.9271, p-value = 0.004789

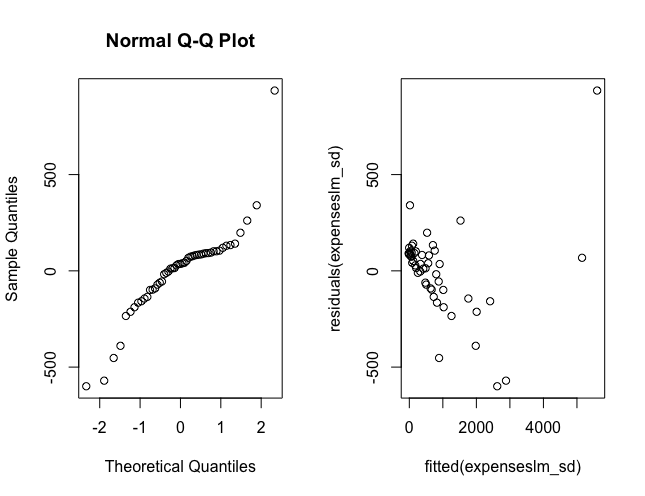
###### !!!!!!!! Step-down strategy !!!  
  
  
detach(new\_crime\_expenses)  
attach(crime\_expenses)  
  
#1st Iteration - all factors  
expenseslm\_sd = lm(expend~bad+lawyers+employ+pop, data=crime\_expenses)  
summary(expenseslm\_sd) #R squared = 0.9637

## lm(formula = expend ~ bad + lawyers + employ + pop, data = crime\_expenses)  
## Residuals:  
## Min 1Q Median 3Q Max   
## -635.62 -80.18 18.77 114.54 809.66   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.464e+02 4.541e+01 -3.224 0.00232 \*\*  
## bad -2.241e+00 1.133e+00 -1.977 0.05402 .   
## lawyers 2.646e-02 7.571e-03 3.495 0.00106 \*\*  
## employ 2.283e-02 7.487e-03 3.049 0.00380 \*\*  
## pop 6.368e-02 3.304e-02 1.927 0.06012 .   
## Multiple R-squared: 0.9666, Adjusted R-squared: 0.9637   
## F-statistic: 332.5 on 4 and 46 DF, p-value: < 2.2e-16

#2nd Iteration - factors bad and pop are removed given that their respective p-values are above 0.05.  
expenseslm\_sd = lm(expend~lawyers+employ, data=crime\_expenses)  
summary(expenseslm\_sd) #this combination of factors doesn’t contribute to R squared

## lm(formula = expend ~ lawyers + employ, data = crime\_expenses)  
## Residuals:  
## Min 1Q Median 3Q Max   
## -599.47 -94.43 36.01 91.98 936.55   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.107e+02 4.257e+01 -2.600 0.01236 \*   
## lawyers 2.686e-02 7.757e-03 3.463 0.00113 \*\*   
## employ 2.971e-02 5.114e-03 5.810 4.89e-07 \*\*\*  
## Multiple R-squared: 0.9632, Adjusted R-squared: 0.9616   
## F-statistic: 627.7 on 2 and 48 DF, p-value: < 2.2e-16

#Graphical analysis



shapiro.test(residuals(expenseslm\_sd)) # The p-value led us to reject the null hypothesis

## Shapiro-Wilk normality test  
## data: residuals(expenseslm\_sd)  
## W = 0.8475, p-value = 1.118e-05

# Final model (obtained via step up strategy – 4th iteration):  
#expenses =-5.937+0.055 x employ-7.885 x bad+0.0162 x bad^2 + error

**Exercise 7**

Below the approach for this analysis will be described, each of the steps, evidences and conclusions can be verified/seen in the above excerpts of script and results from R.

To execute this analysis, basically two linear regression strategies were used: initially **step-up** and later **step-down**.

The first step was to understand the existence of **correlation** between:

* the variable that should be explained ‘expend’ and the other variables/factors given in the data set (in order to have an impression of potential candidates for the model);
* the factors themselves in order to understand the possibility of **collinearity** (factors that have a strong correlation to each other, in other words, that are not linearly independent from each other).

For that, two mechanisms were used, the method pairs (which shows a graph where a potential correlation between all the factors in the data set can be depict) and also the function ‘cor’ (which presents the correlation between each pair of variables).

From this analysis it was possible to conclude that employ and lawyers have a strong correlation – for that reason is expected that in the final model these factors will not appear together. The other factors which seem to have a correlation with expend (⍴ next to 1).

Afterwards, using the step-up strategy the linear regression was calculated for each individual factor and the one with highest R squared was chosen to be added to the model, in this case employ.

In the second iteration other factors were added to employ and the lin. Reg. again calculated, this time the highest R squared was found when employ and lawyers were combined, but since these factors have a strong correlation to each other, this combination was discarded.

An iteration (4th) was done with bad2 since this factor (bad), despite of the p-value above the confidence level, had a big numerical coefficient. For the combination of employ and bad2 the R squared had a small improvement compared to the previous iterations.

At this point, **diagnostic techniques** were used to assess the current model (which had: employ, bad and bad2 as factors). For that, a **qqnorm** plot was done which in this case presented a curve not very similar to a normal distribution. Additionally, a **Shapiro test** was done on the residuals of the linear regression, the calculated p-value was small enough to reject the null hypothesis (that the distribution follows the normal dist.). A **fitted-residual plot** was also used to identify the spread of the points – in this case it was possible to observe a high concentration of points for fitted values < 1000 (so the points were not well scattered across the graph).

After this analysis the next step was to identify **potential and influence points**, for that the Cook distance was calculated. As a result, two points were identified as potentials, they were then removed from the data set and the linear regression was again calculated using the best combination of factors so far, as could be seen after checking the calculated R squared that those points indeed had influence on the parameter of the model.

With this we conclude the step-up strategy, the final model was then:

The step-down strategy was also executed. In this approach rather than R square, the decision on which factors should be in the model is based on the p-value of each factor.

After two iterations there were only two factors left lawyers and employ (which are correlated among each other) besides that the R squared for the linear regression was worse than the one achieved by the model identified on the step-up strategy. So at this point the analysis on this strategy was interrupted.

The final model is then the one calculated via the step-up strategy: