Assignment 1

Sherida van den Bent, Chang Liu, Kai Zhang

2020/2/26

Exercise 1

a) In this experience, we have two factor environment and humidity. Environment has a fixed level of $I = 3$, and humidity has a fixed level of $J = 2$. Also, we have 18 experimental units, and because of balanced design, we choose $N = 18 \div I \div J = 3$.

```
I=3; J=2; N=3
environment = rep(c('cold', 'intermediate', 'warm'),each=N*J)
humidity = rep(c('dry', 'wet'),N*I)
rbind(environment,humidity,sample(1:(N*I*J)))
```

```
##             [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]
## environment "cold" "cold" "cold" "cold" "cold" "cold" "intermediate"
## humidity    "dry"  "wet"  "dry"  "wet"  "dry"  "wet"  "dry"
##             "16"   "5"    "12"   "15"   "9"    "17"   "6"
##             [,8]           [,9]           [,10]          [,11]
## environment "intermediate" "intermediate" "intermediate" "intermediate"
## humidity    "wet"          "dry"          "wet"          "dry"
##             "4"            "2"            "7"            "18"
##             [,12]          [,13]  [,14]  [,15]  [,16]  [,17]  [,18]
## environment "intermediate" "warm" "warm" "warm" "warm" "warm" "warm"
## humidity    "wet"          "dry"  "wet"  "dry"  "wet"  "dry"  "wet"
##             "10"           "8"    "11"   "14"   "13"   "3"    "1"
```

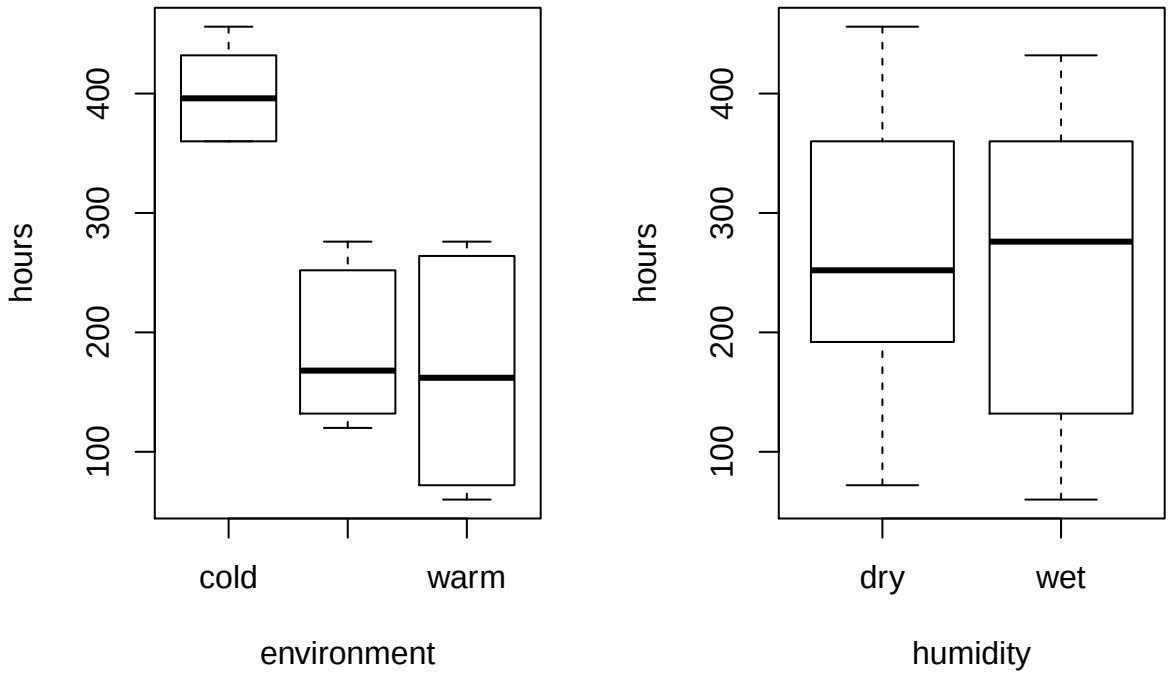
Result means for unit 15 use levels ('cold','dry'); for unit 14 use levels ('cold','wet'); . . .; for unit 5 use levels ('warm','wet').

b)

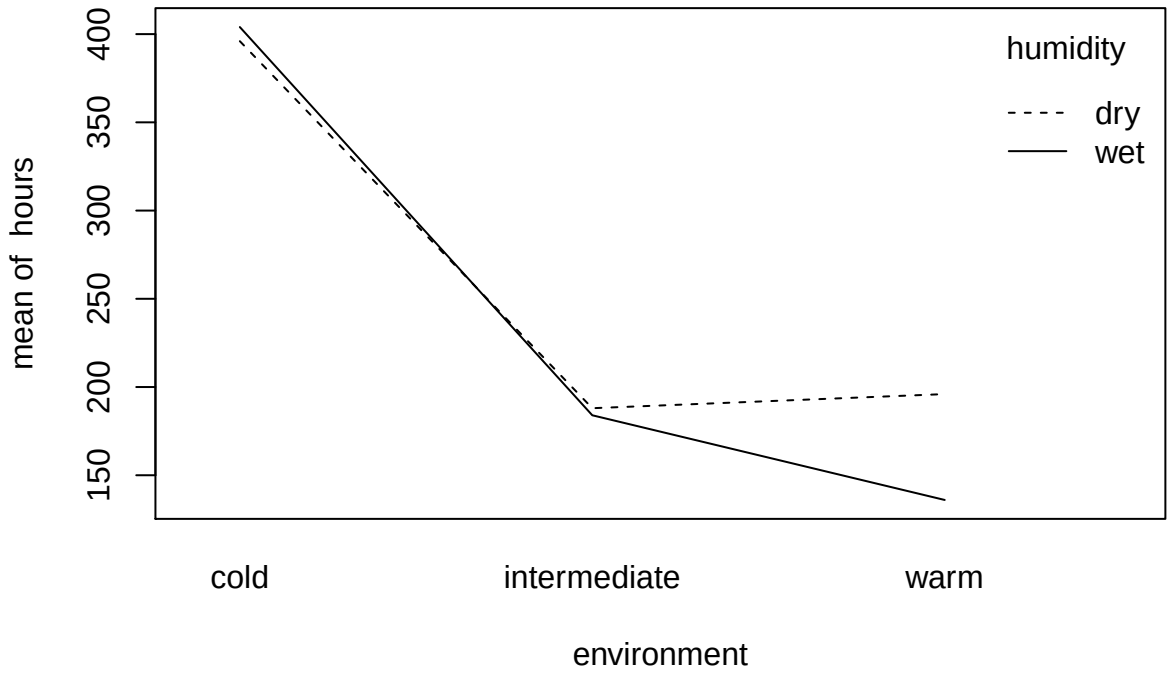
```
bread = read.table('bread.txt')
attach(bread)
```

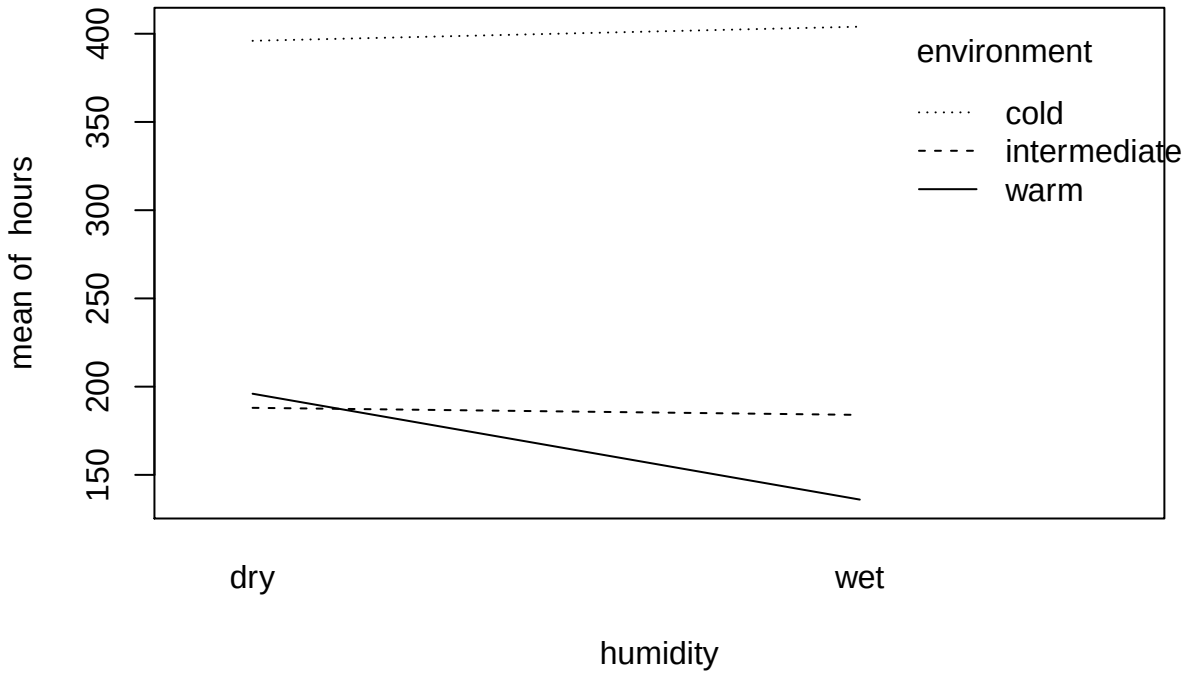
```
## The following objects are masked _by_ .GlobalEnv:
##
##     environment, humidity
```

```
par(mfrow=c(1,2))
boxplot(hours~environment); boxplot(hours~humidity);
```




```
par(mfrow=c(1,1))
interaction.plot(environment,humidity,hours); interaction.plot(humidity,environment,hours)
```






```
detach(bread)
```

c)

```
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
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value of testing: $H_0 : \gamma_{i,j} = 0$for all (i, j) is 3.705e-07, means interaction effect is highly significant, in other word, the relationship between humidity and the time to decay differs by the level of environment.

c)

```
summary(breadaov)
```

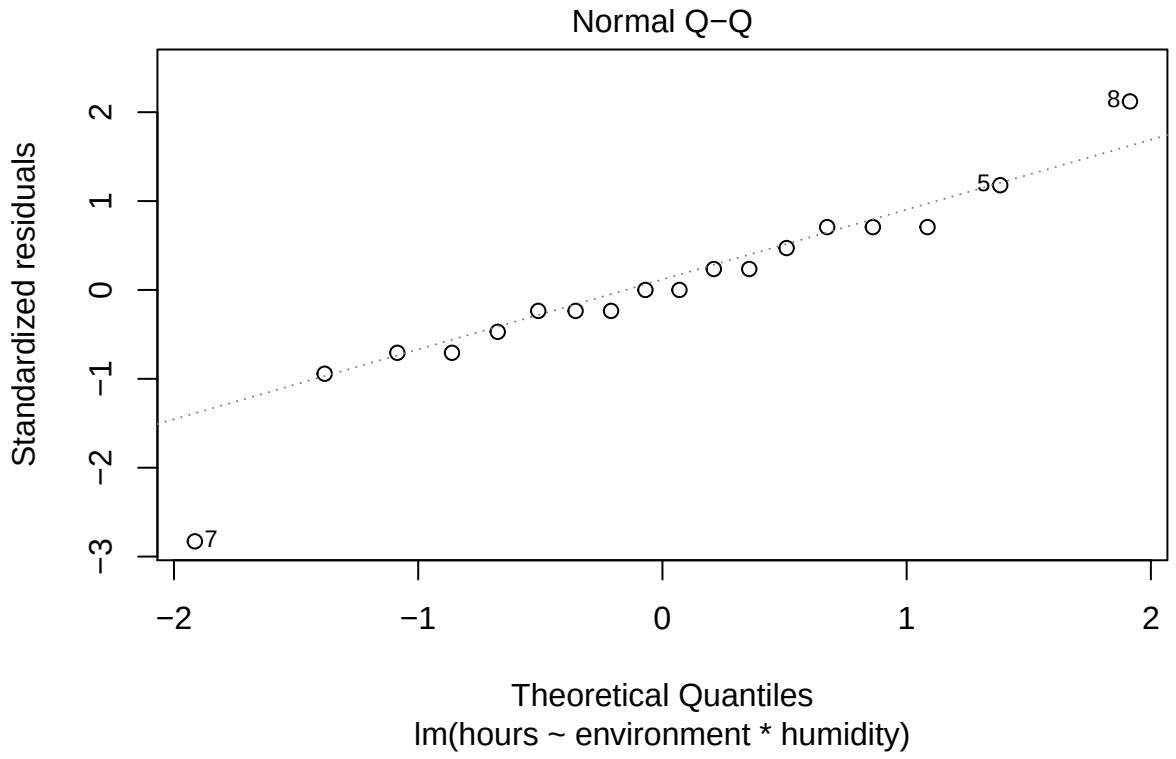
```
##
## Call:
## lm(formula = hours ~ environment * humidity, data = bread)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##    -48     -7      0     11     36
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        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          ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.78 on 12 degrees of freedom
## Multiple R-squared:  0.9821, Adjusted R-squared:  0.9747
## F-statistic: 131.9 on 5 and 12 DF,  p-value: 4.676e-10
```

This is not a good question, because since interaction effect is highly significan, we should NOT interpret the main effects without considering the interaction effect.

d)

```
plot(breadaov, 2)
```



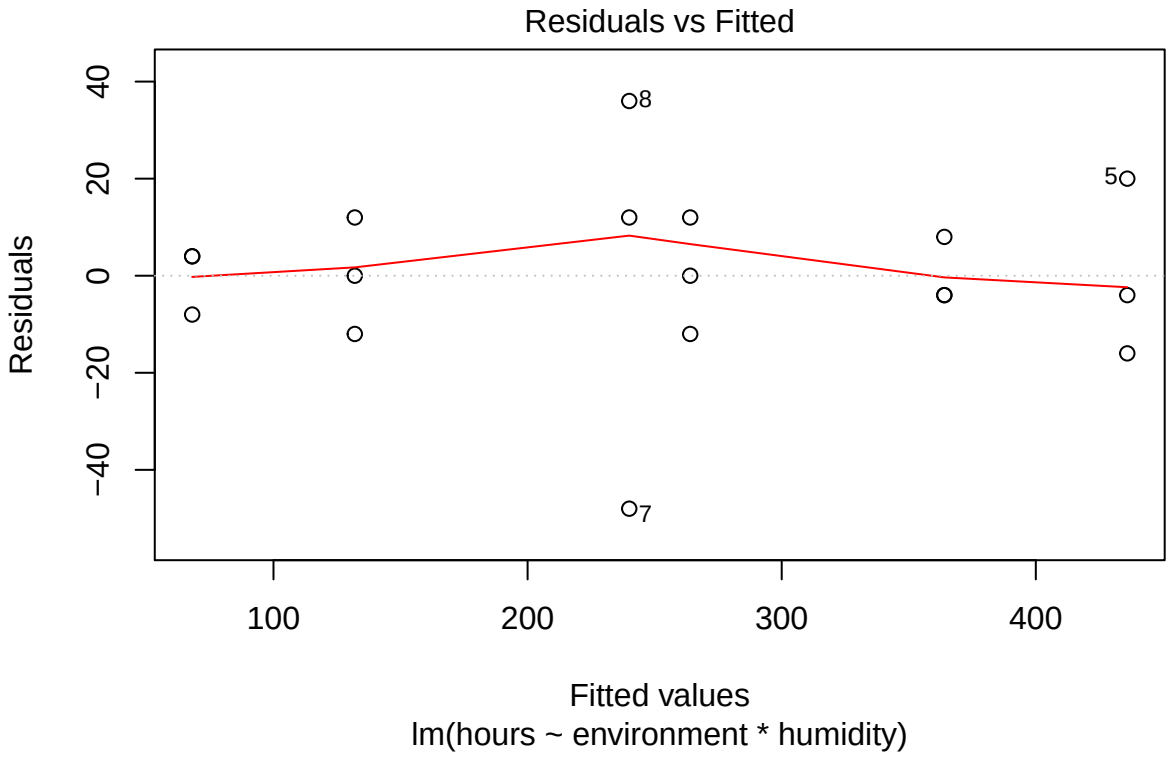

```
shapiro.test(residuals(object = breadaov))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(object = breadaov)
```

```
## W = 0.9296, p-value = 0.1911
```

First we need to check the assumption: normal distribution of the model residuals. After using QQ-plot, we can't tell whether it is normal distribution because of some outliers. So we also use Shapiro-Wilk normality test and based on the p-value($0.1911 > 0.05$), we could say it probably normal distributed. Conclusion: the first assumption of normal distribution of the model residuals has been met.

```
plot(breadaov, 1)
```



The plot seems to indicate that the residuals and the fitted values are uncorrelated. Conclusion: the second assumption of homogeneity of variance of the groups has been met.

Exercise 2

a) In this experience, we have two factor interface and skill. The treatment factor interface has a fixed level of $I = 3$, and the block variable skill has a fixed level of $B = 5$. Also, we have 15 experimental units, and because of balanced design, we choose $N = 15 \div I \div J = 1$.

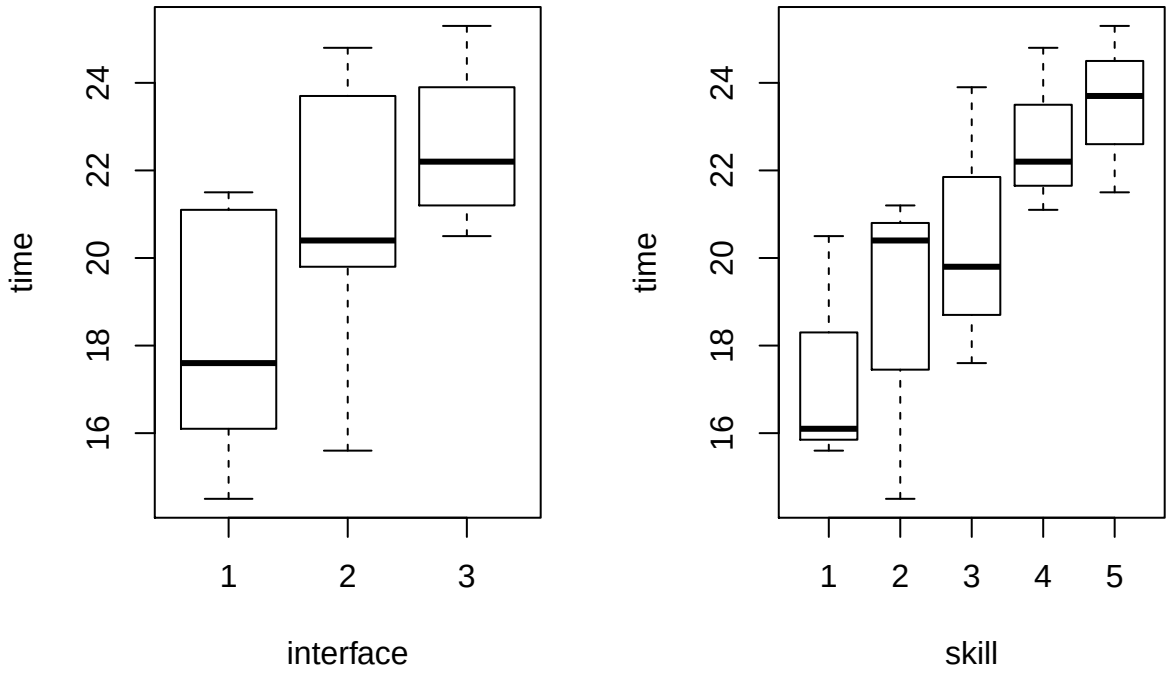
```
I=3; B=5; N=1
for (i in 1:B) print(sample(1:15)[(I*(i-1) + 1):(I*(i-1) + 3)])
```

```
## [1]  4 15  5
## [1] 12  3  7
## [1] 14  5  9
## [1] 15  5  1
## [1] 11  1 13
```

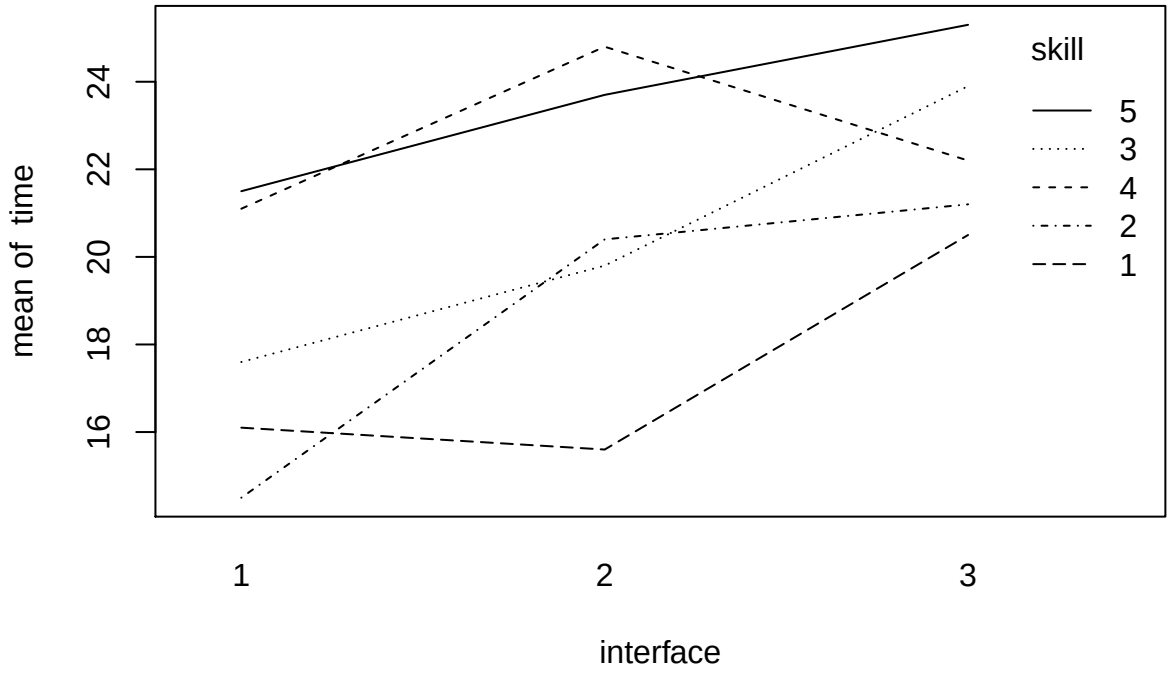
For block 1 assign unit 5 to treatment 1, unit 12 to treatment 2, etc., for block 2 assign unit 11 to treatment 1, unit 3 to treatment 2, etc. **b)**

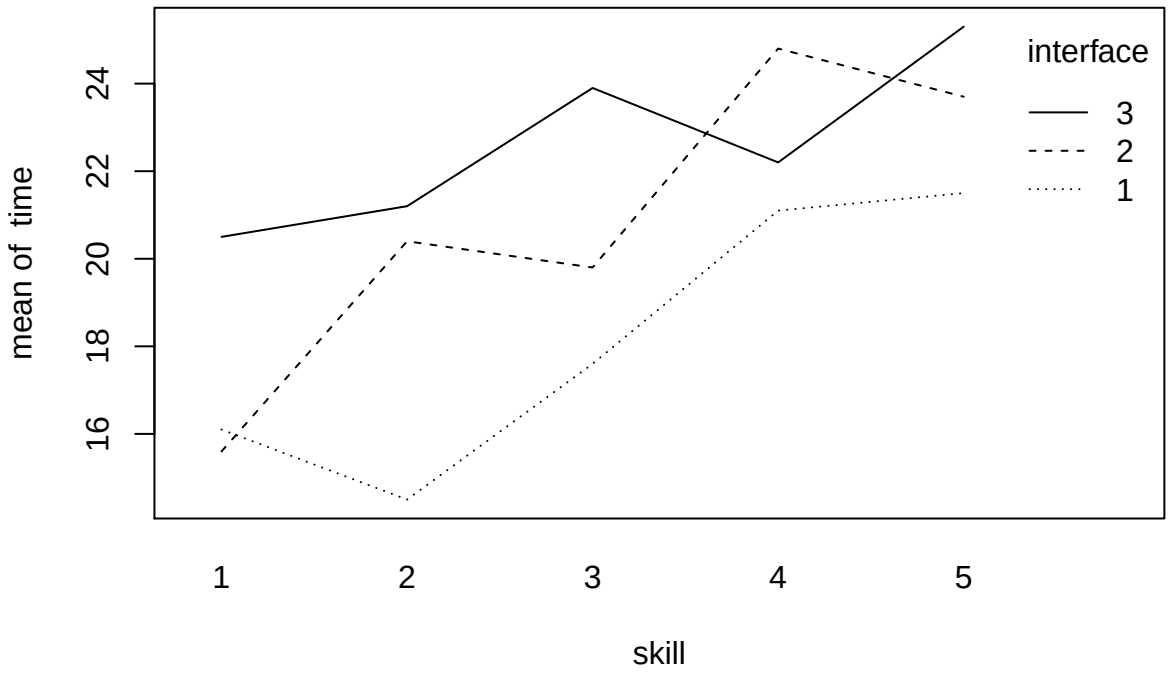
```
search = read.table('search.txt')
search$skill=as.factor(search$skill); search$interface=as.factor(search$interface);
attach(search)
```

```
par(mfrow=c(1,2))
boxplot(time~interface); boxplot(time~skill);
```




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






```
detach(search)
```

The lines in interaction plot are roughly parallel, so we can say there is no interactions between interface and skill. **c)**

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

```
## 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
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value for testing $H_0 : \alpha_i = 0$ for all (i) is 0.0007313. Conclusion: reject H_0 and search time is not same for all interfaces.

```
summary(searchaov)
```

```
##
## Call:
## lm(formula = time ~ interface + skill, data = search)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5733 -0.6967  0.3867  1.0567  1.7867
##
## 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 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.796 on 8 degrees of freedom
## Multiple R-squared:  0.8349, Adjusted R-squared:  0.7111
## F-statistic: 6.745 on 6 and 8 DF,  p-value: 0.008395
```

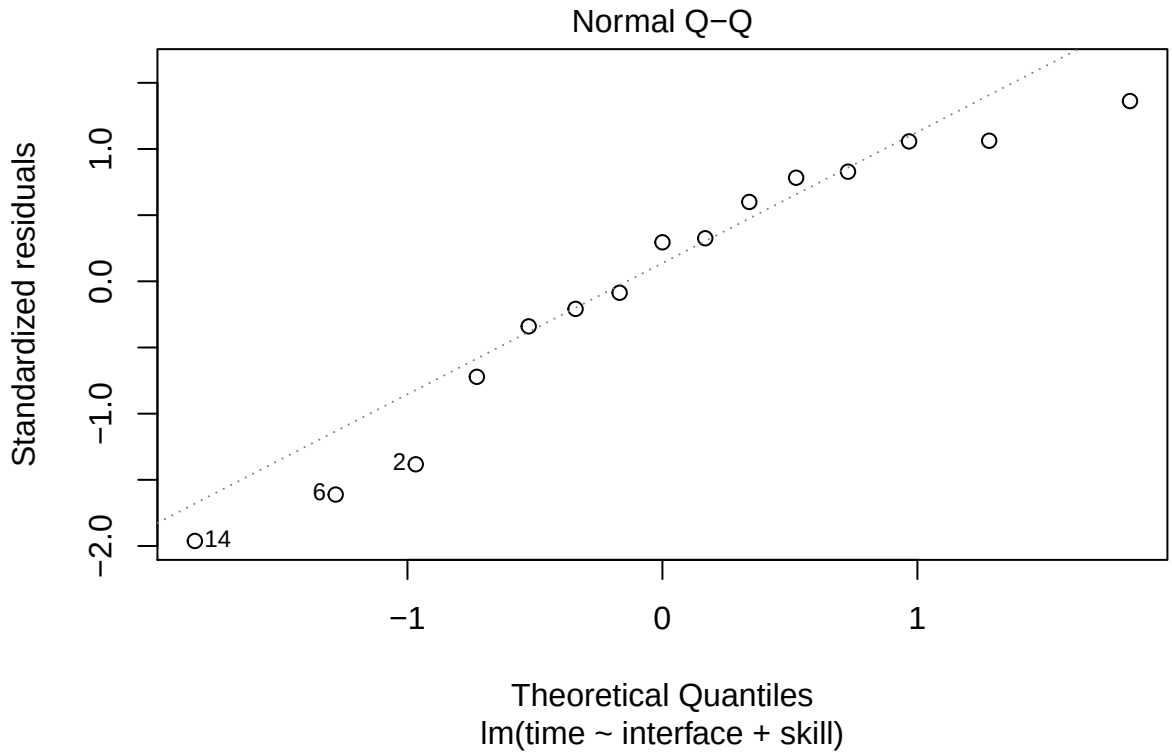
```
15.013+2.700+3.033
```

```
## [1] 20.746
```

The estimate value for $\hat{\mu}$=15.013, $\hat{\alpha_2}$=2.700, $\hat{\beta_3}$=3.033. So the additive model is $\hat{\mu_{23}} = \hat{\mu} + \hat{\alpha_2} + \hat{\beta_3}$=20.746.

d)

```
plot(searchaov, 2)
```

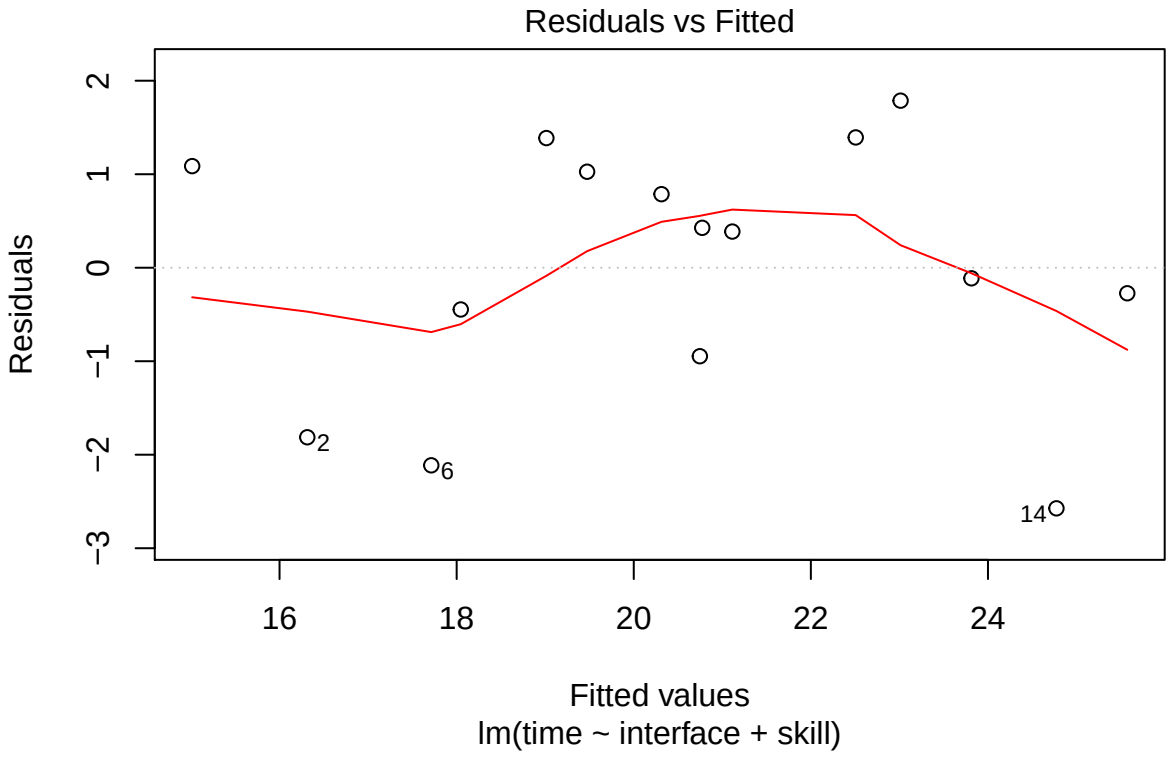



```
shapiro.test(residuals(object = searchaov))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(object = searchaov)
## W = 0.93092, p-value = 0.2817
```

First we need to check the assumption: normal distribution of the model residuals. After using QQ-plot, we can't tell whether it is normal distribution because of some outliers. So we also use Shapiro-Wilk normality test and based on the p-value($0.2817 > 0.05$), we could say it probably normal distributed. Conclusion: the first assumption of normal distribution of the model residuals has been met.

```r
plot(searchaov, 1)
```



The plot seems to indicate that the residuals and the fitted values are uncorrelated. Conclusion: the second assumption of homogeneity of variance of the groups has been met.

e)

```r
attach(search)
friedman.test(time, interface, skill)
```

```
##
##  Friedman rank sum test
##
## data:  time, interface and skill
## Friedman chi-squared = 6.4, df = 2, p-value = 0.04076
```

```r
detach(search)
```

p-value for testing H_0 : no treatment effect is 0.04076, so H_0 is rejected, there is an effect of interface.

f)

```r
searchoneaov = lm(time ~ interface, data = search)
anova(searchoneaov)
```

```
## 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
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It is not useful to perform this test on this dataset, because randomlized block design is to make the variability within blocks is less than the variability between blocks. and this design reduces variability within treatment conditions and potential confounding, producing a better estimate of treatment effects. Since we already gather enough data, we should always apply randomlized block design instead of one-way ANOVA.

Exercise 3

a)

```
cow = read.table('cow.txt')
cowlm=lm(milk~treatment+id,data=cow)
anova(cowlm)
```

```
## Analysis of Variance Table
##
## Response: milk
##           Df  Sum Sq Mean Sq F value Pr(>F)
## treatment  1    0.27   0.269  0.0017 0.9675
## id         1  161.01 161.008  1.0281 0.3267
## Residuals 15 2349.19 156.613
```

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 = 0.22437, df = 8, p-value = 0.8281
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.267910  2.756799
## sample estimates:
## mean of the differences
##               0.2444444
```

Exercise 4

a)

```
nausea = read.table('nauseatable.txt')
df = data.frame(matrix(0,304,2))
names(df) <- c("naus", "medicin")
df[1:180, 1] = 1; df[181: 304, 1] = 2
df[1:100, 2] = 1; df[101:132, 2] = 2; df[133:180, 2] = 3
df[181:232, 2] = 1; df[233:267, 2] = 2; df[268:304, 2] = 3
xtabs(~medicin+naus, data=df)
```

```
##        naus
## medicin   1   2
##       1 100  52
##       2  32  35
##       3  48  37
```

We use number 1 in column “naus” to indicate “Incidence of no nausea”, number 2 to indicate “Incidence of Nausea”. Number 1 in column to indicate “Chlorpromazine” and 2, 3 for “Pentobarbital(100mg)”,

"Pentobarbital(150mg)" respectively. xtabs are a convenient function to creat contingency table, so the result is same as the data in "nauseatable.txt". **b)**

```
B=1000
tstar=numeric(B)
for (i in 1:B) {
  treatstar=df
  treatstar[,2] = sample(df[,2])
  tstar[i] = chisq.test(xtabs(~medicin+naus, data = treatstar))[[1]]
}
myt = chisq.test(xtabs(~medicin+naus, data = df))[[1]]
pr=sum(tstar>myt)/B
pr
```

```
## [1] 0.032
```

The p-value for testing $H_{0}: $the different medicins work equally well against nausea is 0.031 Conclusion: we reject H_0 and we have confidence that there is difference between different medicins. **c)**

```
chisq.test(xtabs(~medicin+naus, data = df))[[3]]
```

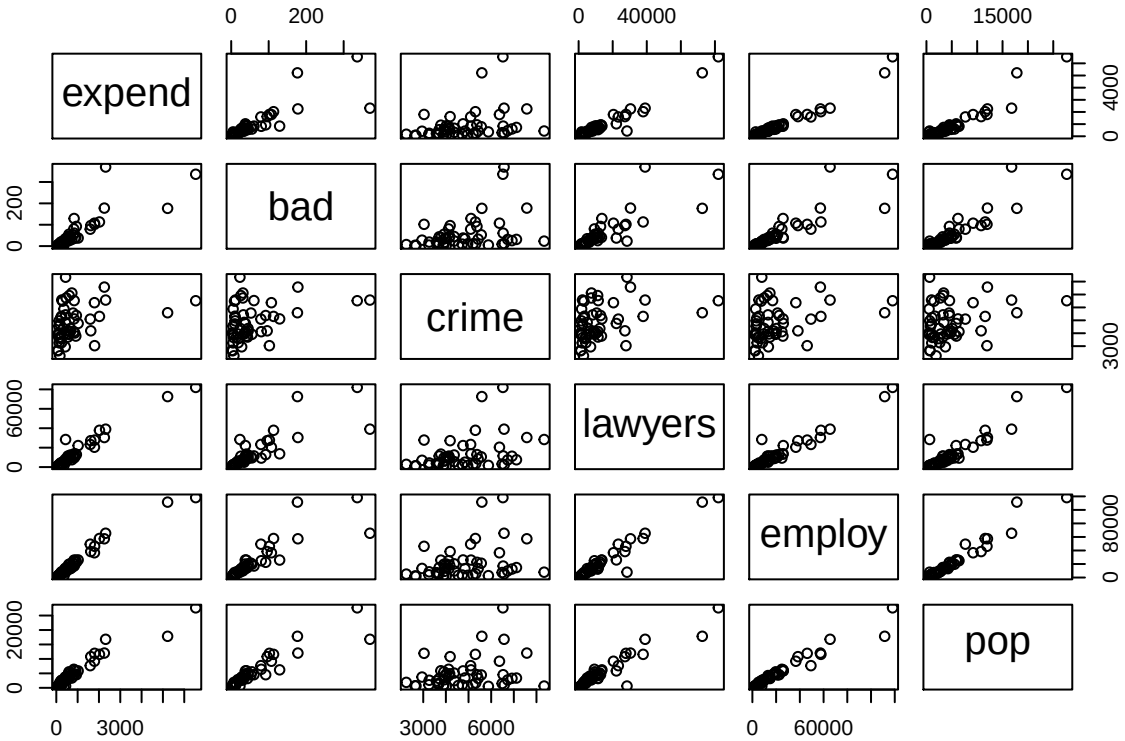
```
## [1] 0.03642928
```

p-value from permutation test and chisquare test for contingency tables are very close.

Exercise 5

a)

```
crime = read.table('expensescrime.txt', header = TRUE)[, 2:7]
pairs(crime)
```



b)

```
summary(lm(expend~bad,data=crime))
summary(lm(expend~crime,data=crime))
summary(lm(expend~lawyers,data=crime))
summary(lm(expend~employ,data=crime))
summary(lm(expend~pop,data=crime))
```

Model expend~employ has max determination coefficient: 0.954, so wo chose this model for next step.

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

In those four model only lawyers in expend~employ+lawyers is significant, and it has determination coefficient 0.9632 larger than 0.954, so wo chose this model for next step.

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

All of those newly added feature yields insignificant explanatory variables, so we can stop and take model expend~employ+lawyers as our final step-up model.

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

Feature crime has the largest p-vlaue 0.25534, and it is large than 0.05, so we remove crime from the model.

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

Feature pop has the largest p-vlaue 0.06012, and it is large than 0.05, so we remove pop from the model.

```
summary(lm(expend~bad+lawyers+employ,data=crime))
```

Feature bad has the largest p-vlaue 0.34496, and it is large than 0.05, so we remove bad from the model.

```
summary(lm(expend~lawyers+employ,data=crime))
```

All remaining explanatory variables in the model are significant, so we can stop and take model expend~employ+lawyers as our final step-up model.

Both method generate same model: expend~employ+lawyers.