EDDA_assign2_Group33

Wenbo Sun, Yuhao Qian, Meifang Li

3/7/2021

Exercise 1

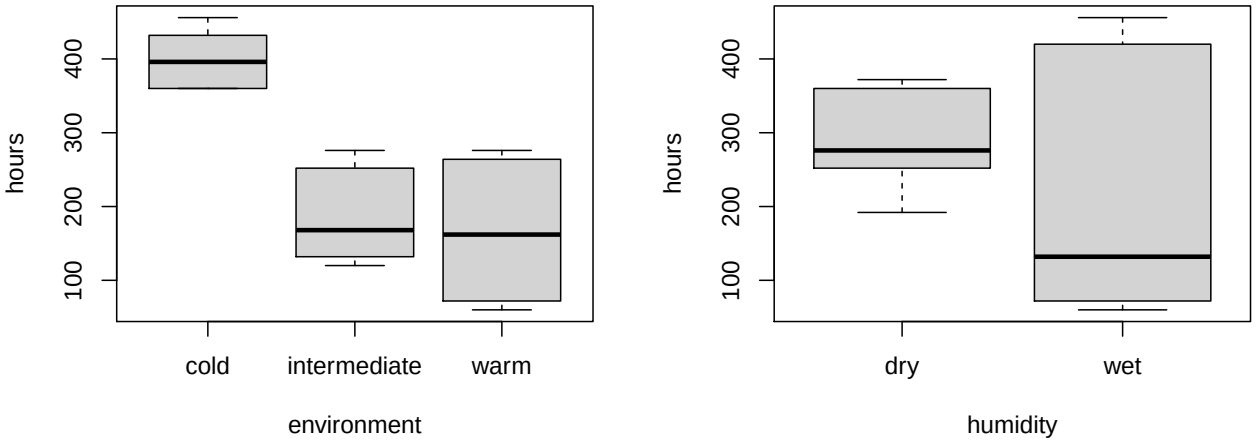
a) The randomization process can be implemented as follows: the 18 slices were randomized to the 6 combinations of conditions, cold&dry/cold&wet/intermediate&dry/intermediate&wet/warm&dry/warm&wet, each combinations sampled 3 times.

```
I=list("cold","intermediate","warm");J=list("dry","wet");N=3
rbind(rep(I,each=N*length(J)),rep(J,N*length(I)),sample(1:(N*length(I)*length(J))))
```

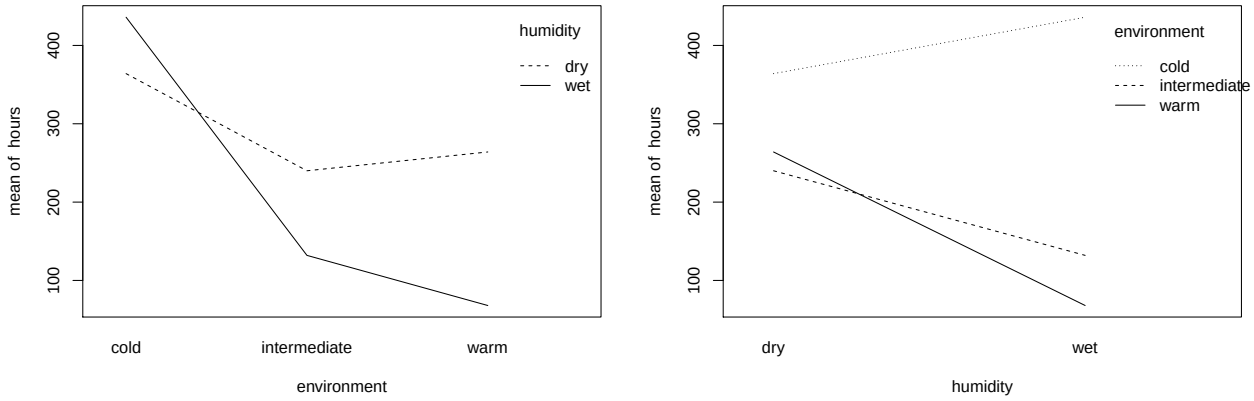
```
##      [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]           [,8]
## [1,] "cold" "cold" "cold" "cold" "cold" "cold" "intermediate" "intermediate"
## [2,] "dry"  "wet"  "dry"  "wet"  "dry"  "wet"  "dry"          "wet"
## [3,] 17     10     4      3      16     2      6              8
##      [,9]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

b) The boxplots show that the bread can keep longer in the cold environment than in the intermediate and warm environment while wet environment show more influence on hours than dry environment. In the interaction plots, we could also tell that the bread keep longer in wet environment when it is cold but shorter when it is intermediate or warm.

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



```
par(mfrow=c(1,2))
interaction.plot(environment,humidity,hours);
interaction.plot(humidity,environment,hours)
```



c) The p-value for $H_0 : \gamma_{i,j} = 0$ for all (i, j) is $3.705 * 10^{-7}$ which is smaller than 0.05, thus could be rejected. Hence, there is a significant evidence for interaction between the factors temperature and humidity.

```
bread$environment=as.factor(bread$environment);bread$humidity=as.factor(bread$humidity)
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   233.7 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
```

d) According to the coefficients of estimate in the result, the temperature has the greatest influence on the decay with cold=364/intermediate=364-124=240/warm=364-100=264, all keep shorter than in the humidity with dry=364/wet=364+72=436. However, this is not a good question because we have proved there is significant interaction effect between these factors and we cannot tell the exclusive effect of each factor without considering the interaction effect.

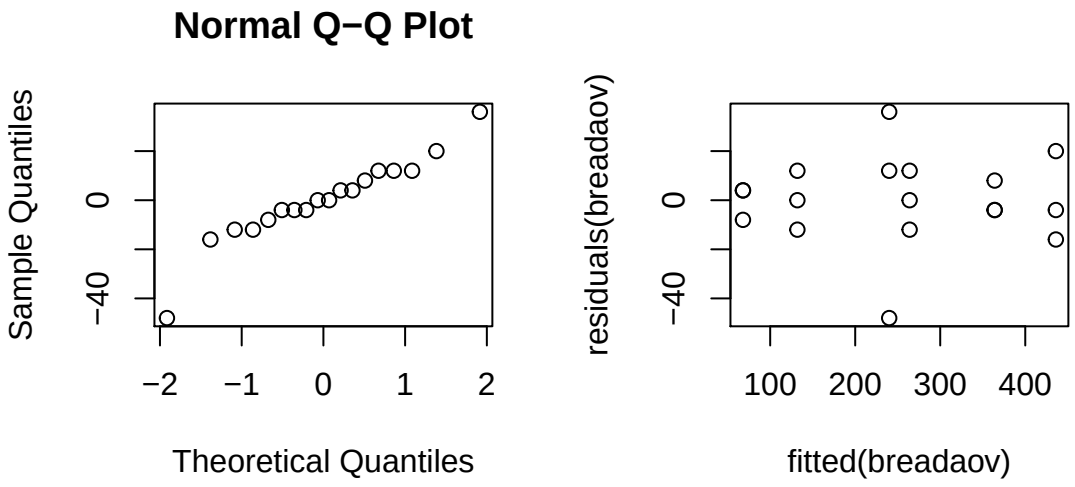
```
summary(breadaov)
```

```
##
## 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         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
```

e) The QQ-plot looks a bit deviated in the extremes but it could be normal. Some data-points seem extreme in the fitted plot. There are two outliers, one is too high and one is too low for the residuals.

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



Exercise 2

a) The Block represent the five different types of student and the Inter represent three different types of interfaces. We have 15 students so each combination we sample one time as follows: the 5 rows are the 5 blocks in which 1/2/3 represent 3 types of interfaces.

```
Block=5;Inter=3;N=1
for (i in 1:Block) print(sample(1:(N*Inter)))
```

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

b) We use two-way anova to test the null hypothesis. As shown in the result, the p-value of interface is 0.0131<0.05, thus the $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$ is rejected, the interface effects are siginificantly different from 0 and the search time is not the same for all interfaces. We could also tell that the interface 3 requires the longest search time and the combination of skill level 1 and interface 1 takes the shortest search time. The time it takes a skill level 3 user to find the product uses interface 3 is 15.013+3.033+4.46=22.506.

```
attach(search)
search$skill=as.factor(search$skill);search$interface=as.factor(search$interface)
searchanov=lm(time~skill+interface,data=search)
anova(searchanov)
```

```
## Analysis of Variance Table
```

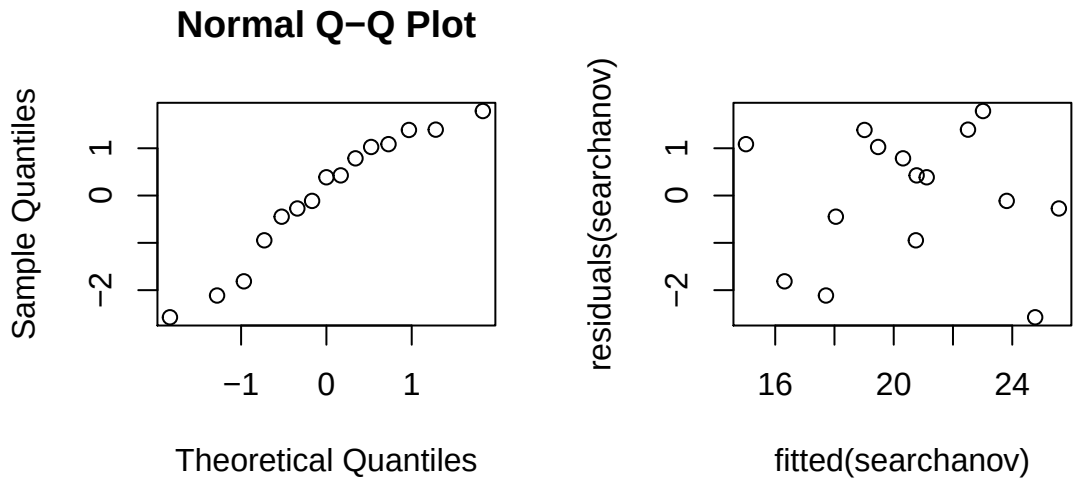
```
##
## Response: time
##           Df Sum Sq Mean Sq F value Pr(>F)
## skill      4   80.1   20.01    6.21  0.014 *
## interface  2   50.5   25.23    7.82  0.013 *
## Residuals  8   25.8    3.23
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(searchanov)[["coefficients"]]
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    15.01       1.23  12.238 1.85e-06
## skill2          1.30       1.47   0.887 4.01e-01
## skill3          3.03       1.47   2.069 7.24e-02
## skill4          5.30       1.47   3.614 6.84e-03
## skill5          6.10       1.47   4.160 3.16e-03
## interface2      2.70       1.14   2.377 4.47e-02
## interface3      4.46       1.14   3.927 4.38e-03
```

c) The QQ-plot look like normal and the residuals don't change systematically with the fitted values. Thus we can assume the populations have equal variances.

```
par(mfrow=c(1,2))
qqnorm(residuals(searchanov));plot(fitted(searchanov),residuals(searchanov))
```



d) We use the Friedman test to test whether there is an effect of interface. The result show that the p-value for testing(H_0 : no interface effect)is 0.04076<0.05, thus H_0 is rejected and there is an effect of interface.

```
friedman.test(time,interface,skill,data=search)
```

```
##
##  Friedman rank sum test
##
## data:  time, interface and skill
## Friedman chi-squared = 6, df = 2, p-value = 0.04
```

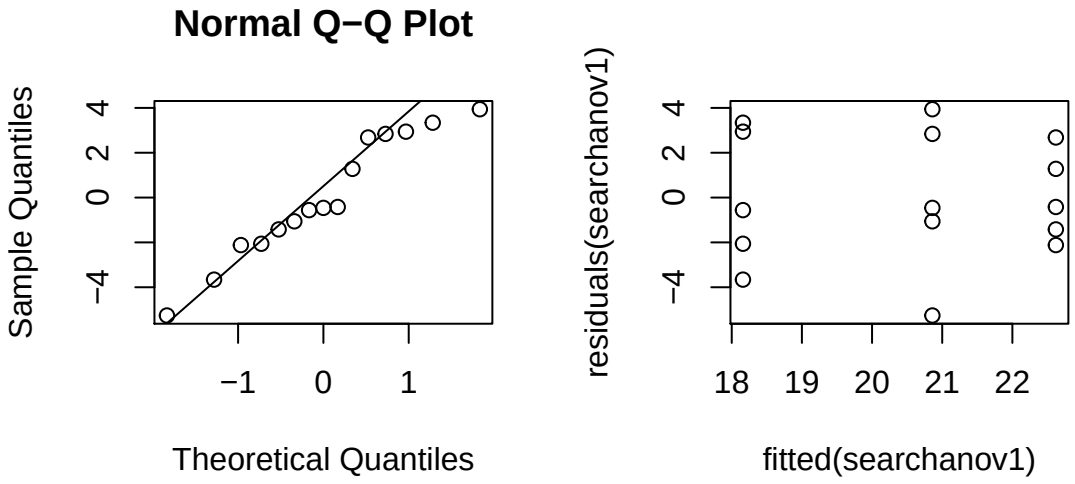
e) We use a one-way anova test to test whether the search time is the same for all interfaces, ignoring the variable sill. The p-value is 0.09642>0.05, which means we cannot reject the null hypothesis and the search time is the same for all interfaces. However, the QQ-plot does not show normality and the residuals spread in a systematical way in the fitted plot, thus the assumption of independent error fails in this experiment. Also we cannot tell the exclusive effect of interface with the exist of block factor because the units might be

dissimilar between the blocks so this one-way anova is both wrong and not useful on this dataset.

```
searchanov1=lm(time~interface,data=search)
anova(searchanov1)
```

```
## Analysis of Variance Table
##
## Response: time
##           Df Sum Sq Mean Sq F value Pr(>F)
## interface  2   50.5   25.23    2.86  0.096 .
## Residuals 12  105.9    8.82
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow=c(1,2))
qqnorm(residuals(searchanov1));qqline(residuals(searchanov1))
plot(fitted(searchanov1),residuals(searchanov1))
```



Exercise 3

a) In this question, we test the influence of treatment by the ordinary fixed effect model. We left out the order factor because the fixed effect model can not estimate the sequence influences. Regarding the order of factors, we put the 'treatment' at the end of the formula because the anova is a sequential analysing model in which order may affect the p-value of factors. The results are presented as follows:

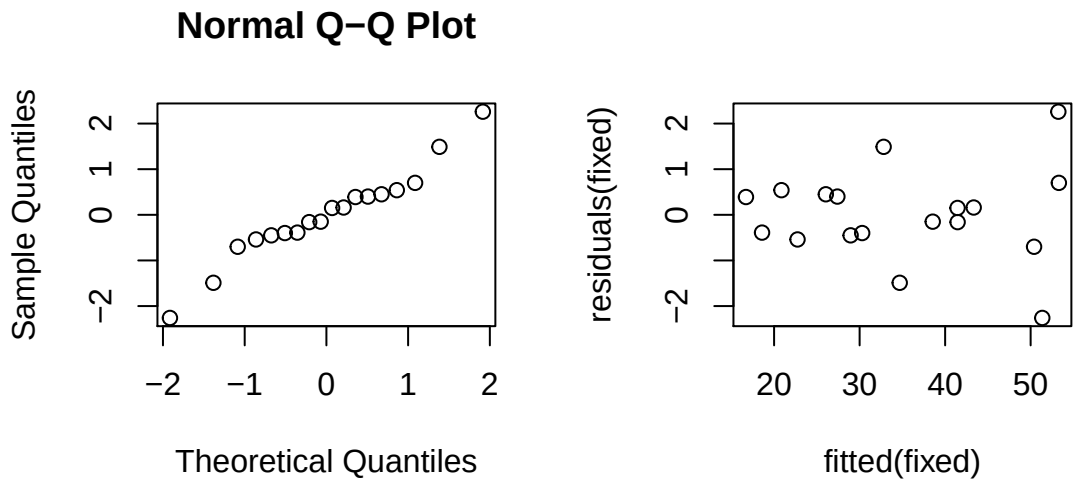
```
attach(data3)
fixed=lm(milk~id+per+treatment)
anv=anova(fixed);anv
```

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

```
summary(fixed)[["coefficients"]]
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    30.30      1.244  24.349 5.02e-08
## id2            23.00      1.574  14.612 1.68e-06
## id3            11.15      1.574   7.084 1.96e-04
## id4            -1.35      1.574  -0.858 4.19e-01
## id5            -7.05      1.574  -4.479 2.87e-03
## id6            23.45      1.574  14.898 1.47e-06
## id7            13.55      1.574   8.608 5.69e-05
## id8             4.90      1.574   3.113 1.70e-02
## id9           -11.20      1.574  -7.115 1.91e-04
## per2           -2.39      0.747  -3.201 1.50e-02
## treatmentB     -0.51      0.747  -0.683 5.17e-01
```

```
par(mfrow=c(1,2))
qqnorm(residuals(fixed))
plot(fitted(fixed),residuals(fixed))
```



The results show that 'treatment', also called feedingstuffs in this problem, has no significant influence on milk production because the p-value of 'treatment' is 0.516 larger than 0.05. We also diagnose the normality and spread of the residuals. The normality is doubtful because the qqnorm in the left plot is slightly curved. Also, the residuals do not distribute randomly: the spread shows bigger for larger fitted values. Hence, the test result is not sufficient, and we have to consider the influences of feed order.

b) For a comprehensive analysis, we conduct test by mixed effect model in which we consider the individual as 'random effect'. Because different individuals may have various reflection which is independent of the experiment setting, we regard the individuals as random samples from a population.

```
mixed=lmer(milk~(1|id)+order+per+treatment,REML=FALSE);
s_mixed=summary(mixed);s_mixed[["varcor"]];s_mixed[["coefficients"]]
```

```
##  Groups   Name        Std.Dev.
##  id       (Intercept) 11.54
##  Residual              1.39
```

```
##             Estimate Std. Error t value
## (Intercept)    38.50      5.811   6.625
## orderBA        -3.47      7.768  -0.447
## per2           -2.39      0.658  -3.630
## treatmentB     -0.51      0.658  -0.775
```

In the results above, we can figure out that the standard deviation of factor 'id' is 11.54, quite large. The effect of treatment and period are identical to results from fixed effect model. To futher determine the influence of treatment, we leave out the treatment and conduct to mixed effect model to the remaining factors. The difference between the two models indicates the treatment effect.

```
mixed1=lmer(milk~(1|id)+order+per,REML=FALSE)
anova(mixed1,mixed)
```

```
## Data: NULL
## Models:
## mixed1: milk ~ (1 | id) + order + per
## mixed: milk ~ (1 | id) + order + per + treatment
##        npar AIC BIC logLik deviance Chisq Df Pr(>Chisq)
## mixed1    5 118 122  -53.9      108
## mixed     6 119 125  -53.7      107  0.58  1       0.45
```

Likewise, the result shows p-value is 0.446 >0.05, indicating treatment does not significantly influence milk production, as the fixed effect model indicates. In the previous question, we can figure out the period has little effect. Hence, the difference between the fixed and mixed model results is minor, even considering the order effect.

c) Conducting t-test means ignoring the order effect. In b), we find that period and order's total effect contributes to -5.86, which is a large fraction, compared to the -0.51 contributed by treatment. Hence, ignoring the order effect is not a reasonable estimation method, indicating t-test on treatment is invalid to this experiment. But we still examine the treatment by t-test, regarding the order as exchangeable. For comparison, we refit the model by id and treatment, leaving out the order and period.

```
attach(data3)
```

```
## The following objects are masked from data3 (pos = 3):
##
##     id, milk, order, per, treatment
```

```
exchangable=lm(milk~id+treatment)
anv_ex=anova(exchangable);anv_ex
```

```
## Analysis of Variance Table
##
## Response: milk
##           Df Sum Sq Mean Sq F value  Pr(>F)
## id         8   2467   308.4   57.74 2.8e-06 ***
## treatment  1      0     0.3    0.05    0.83
## Residuals  8     43     5.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

```
##
##  Paired t-test
##
## data:  milk[treatment == "A"] and milk[treatment == "B"]
## t = 0.2, df = 8, p-value = 0.8
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.27  2.76
## sample estimates:
## mean of the differences
```

```
##                     0.244
```

We observe that p-value of both methods is the same, p-value=0.828>0.005, both insignificant, which is compatible with the conclusion from a).

Exercise 4

a) The target of this question is testing the consistency of literary style using a contingency table. A word counting from one literature is a column of the contingency table. Comparing across pieces of literature is comparing word distribution between columns, which is a test for homogeneity.

b) We select three pieces of literature written by Austen to test the consistency of her word using. To find differences, we can compute the residuals, which reveals the deviation between expectation and observation.

```
z=chisq.test(data4[,c("Sense","Emma","Sand1")]);z
```

```
##
##  Pearson's Chi-squared test
##
## data:  data4[, c("Sense", "Emma", "Sand1")]
## X-squared = 12, df = 10, p-value = 0.3
```

```
residuals(z)
```

```
##           Sense   Emma  Sand1
## a       -1.0300 -0.129  1.594
## an       0.4473 -0.159 -0.375
## this     0.0513  0.294 -0.504
## that     0.7482  0.287 -1.442
## with    -0.0475  0.521 -0.704
## without  1.0654 -1.588  0.893
```

The p-value 0.267 indicates the difference between the three works is not statistically significant. But we also can find some inconsistency in detail. The word 'a' in Sand1 is apparently more frequent than in the other two works. On the contrary, the frequency of the word 'that' is lower.

c) To test the word consistency between the admirer's work and Austen's work, we add Sand2 to the contingency table, applying the chi-squared test again.

```
z1=chisq.test((data4));z1
```

```
##
##  Pearson's Chi-squared test
##
## data:  (data4)
## X-squared = 46, df = 15, p-value = 6e-05
```

```
residuals(z1)
```

```
##           Sense      Emma  Sand1   Sand2
## a       -1.015 -0.112093  1.606 -0.0589
## an      -0.591 -1.219955 -1.067  3.7282
## this     0.139  0.390490 -0.444 -0.3267
## that     1.594  1.179849 -0.910 -3.0493
## with    -0.512  0.000192 -1.025  1.7482
## without  1.392 -1.341196  1.137 -1.0696
```

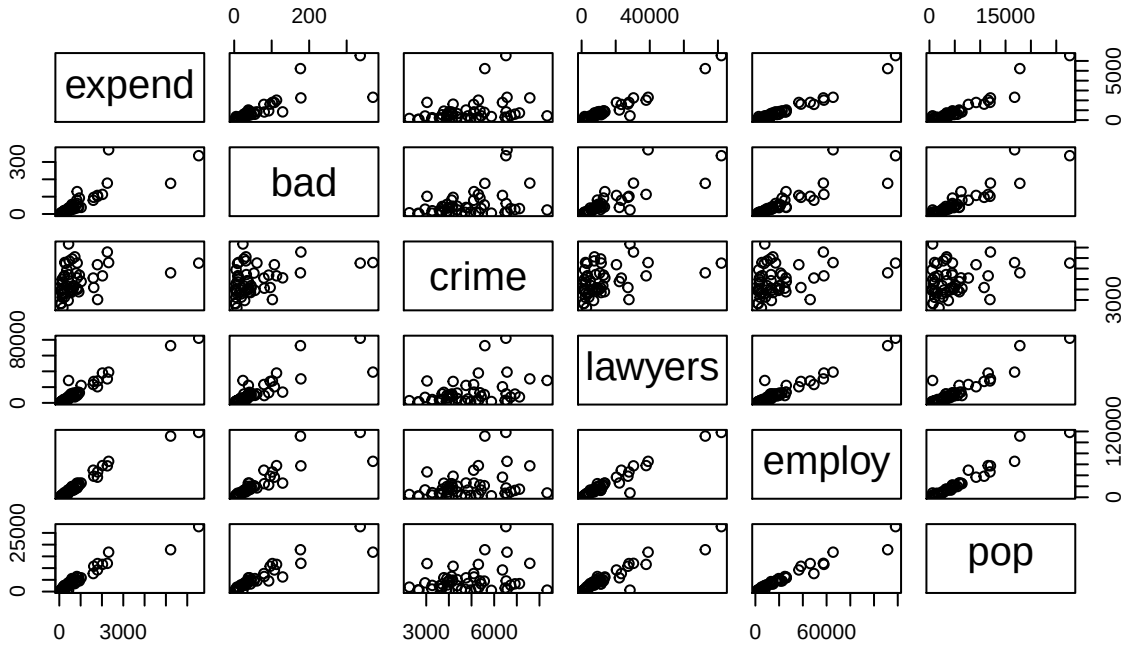
After adding Sand2, the p-value decreases dramatically to 6.2e-05, which means the distributions of these four works are significantly different. Incorporating the conclusion from b), we can say that the admirer's

imitation is not successful. Specifically, Sand2 has much more 'an' and 'with', but has significant less 'that' than Austen's three works.

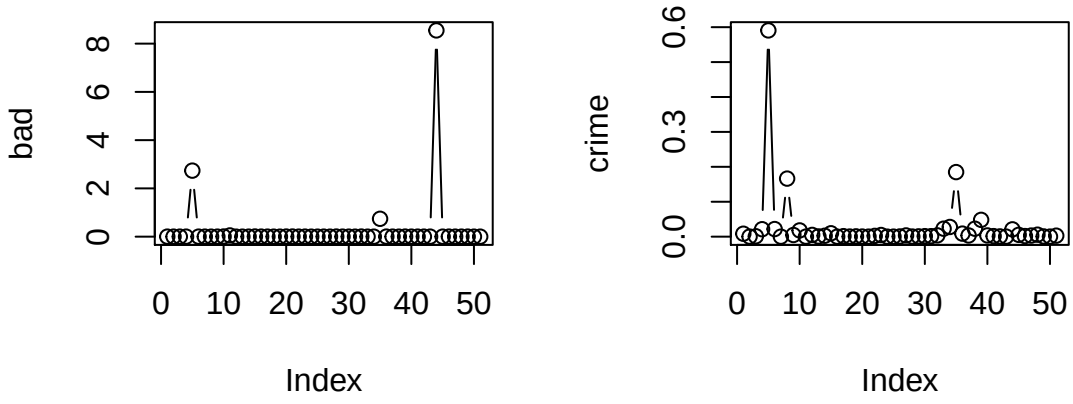
Exercise 5.

a) We first make some graphical summaries of the data.

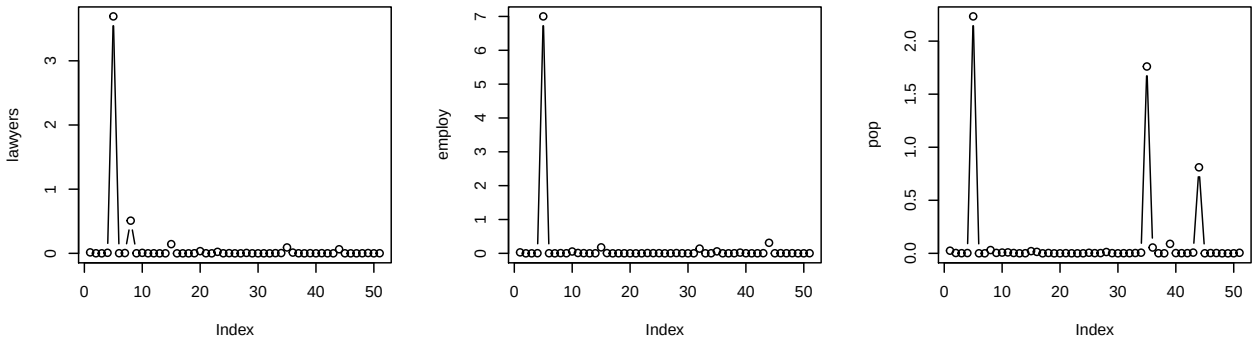
```
plot(crime[,2:7])
```




```
par(mfrow=c(1,2));plot(cooks.distance(lm(expend~bad,data=crime)),type ='b',ylab = 'bad') # 5, 35,44
plot(cooks.distance(lm(expend~crime,data=crime)),type ='b',ylab = 'crime')
```




```
par(mfrow=c(1,3))
plot(cooks.distance(lm(expend~lawyers,data=crime)),type ='b',ylab = 'lawyers') #5
plot(cooks.distance(lm(expend~employ,data=crime)),type ='b',ylab = 'employ') # 5
plot(cooks.distance(lm(expend~pop,data=crime)),type ='b',ylab = 'pop') # 5, 35,44
```



Since we consider a data point with close to or larger than 1 Cook's distance as an influence point. After computing and plotting the Cook's distance for each variable, we notice that the 5th, 35th, 44th points both in bad and pop, the 5th points both in lawyers and employ are all influence points. Hence, we would remove these data sets when we fitting a linear model in b) since all the VIF is larger than 5.

```
v1=vif(lm(expend~bad+lawyers,data=crime));v2=vif(lm(expend~bad+employ,data=crime))
v3=vif(lm(expend~bad+pop,data=crime));v4=vif(lm(expend~lawyers+employ,data=crime))
v5=vif(lm(expend~lawyers+pop,data=crime));v6=vif(lm(expend~employ+pop,data=crime))
v1;v2;v3;v4;v5;v6
```

```
##     bad lawyers
##    3.25    3.25
```

```
##    bad employ
##   4.15   4.15
```

```
##  bad  pop
## 6.53 6.53
```

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

```
## lawyers     pop
##    7.84    7.84
```

```
## employ    pop
##   17.3   17.3
```

As we can see in the first scatter plot in this section, there are four variables(bad/lawyers/employ/pop) seem to be collinears. After calculating the variance inflation factor(VIF) of each possible pair, we could tell that the following pair of variables have problem of collinearity:"bad+pop", "lawyers+employ", "lawyers+pop", "employ+pop" .

b) We also fit the raw data(with influence points) to another linear model using the same methods. Both results are $expend = -110.7 + 0.0297 * employ + 0.0269 * lawyers + e_n$, with $R^2 = 0.9632$. However, this model not only has collinearity problem but also fails the model assumption, as the qq-norm plot does not look well. Hence, fitting a linear model with cleaned data(without influence points) could improve the result.

Then we use the step-up and step-down method separately to fit a linear regression model to the cleaned data. Firstly we implemented the step-up method:

```
s=summary(lm(expend~bad,data=crime2))$coefficients;
s;summary(lm(expend~bad,data=crime2))$r.squared #0.82
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    109.9     48.270    2.28 2.75e-02
## bad             12.7      0.879   14.48 9.46e-19
```

```
## [1] 0.82
```

```
s=summary(lm(expend~crime,data=crime2))$coefficients;
s;summary(lm(expend~crime,data=crime2))$r.squared
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    41.23   275.0661    0.15   0.8815
## crime           0.12     0.0561    2.14   0.0374
```

```
## [1] 0.0908
```

```
s=summary(lm(expend~lawyers,data=crime2))$coefficients;
s;summary(lm(expend~lawyers,data=crime2))$r.squared # 0.81
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  78.1690   50.89134    1.54 1.31e-01
## lawyers       0.0548    0.00387   14.16 2.21e-18
```

```
## [1] 0.813
```

```
s=summary(lm(expend~employ,data=crime2))$coefficients;
s;summary(lm(expend~employ,data=crime2))$r.squared#0.95 best
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  23.2421   25.30964   0.918 3.63e-01
## employ        0.0371    0.00119  31.158 1.43e-32
```

```
## [1] 0.955
```

```
s=summary(lm(expend~pop,data=crime2))$coefficients;
s;summary(lm(expend~pop,data=crime2))$r.squared#0.92
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -19.998   35.00507  -0.571 5.71e-01
## pop            0.166    0.00714  23.289 4.23e-27
```

```
## [1] 0.922
```

We start with fitting all 5 possible simple linear regression models. All the variables are significant in the tests and we select "employ" with the highest $R^2 = 0.955$. The current model is $expend \sim employ$.

```
summary(lm(expend~employ+bad,data=crime2))$coefficients
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.4267   24.97085   0.898 3.74e-01
## employ        0.0333    0.00279  11.959 1.44e-15
## bad           1.5553    1.03032   1.510 1.38e-01
```

```
summary(lm(expend~employ+bad,data=crime2))$r.squared #0.9569
```

```
## [1] 0.957
```

```
summary(lm(expend~employ+crime,data=crime2))$coefficients
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -157.4420   54.96741   -2.86 6.33e-03
## employ         0.0363    0.00108   33.47 1.91e-33
## crime          0.0410    0.01139    3.60 7.80e-04
```

```
summary(lm(expend~employ+crime,data=crime2))$r.squared #0.9649
```

```
## [1] 0.965
```

```
summary(lm(expend~employ+lawyers,data=crime2))$coefficients
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.35063   24.56063   0.625 5.35e-01
## employ       0.03212    0.00253  12.684 1.84e-16
## lawyers      0.00896    0.00405   2.214 3.20e-02
```

```
summary(lm(expend~employ+lawyers,data=crime2))$r.squared #0.9592
```

```
## [1] 0.959
```

```
summary(lm(expend~employ+pop,data=crime2))$coefficients#
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.7459   26.18863   0.219 8.27e-01
## employ        0.0286    0.00456   6.266 1.26e-07
## pop           0.0403    0.02079   1.938 5.89e-02
```

```
summary(lm(expend~employ+pop,data=crime2))$r.squared #0.9582
```

```
## [1] 0.958
```

Then we extend the obtained model with other four possible variables. "crime" not only has highest $R^2 = 0.9649$, but also is significant in the second round. Thus the model turns to $expend \sim employ + crime$.

```
s=summary(lm(expend~employ+crime+bad,data=crime2))$coefficients #p 0.399
s;summary(lm(expend~employ+crime+bad,data=crime2))$r.squared#
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -147.5900   56.33943  -2.620 1.20e-02
## employ         0.0344    0.00254  13.512 2.91e-17
## crime          0.0387    0.01175   3.295 1.95e-03
## bad            0.8166    0.95974   0.851 3.99e-01
```

```
## [1] 0.965
```

```
s=summary(lm(expend~employ+crime+lawyers,data=crime2))$coefficients#p 0.237
s;summary(lm(expend~employ+crime+lawyers,data=crime2))$r.squared
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.39e+02   56.79585   -2.45 1.83e-02
## employ       3.37e-02    0.00240   14.05 7.09e-18
## crime        3.59e-02    0.01211    2.97 4.85e-03
## lawyers      4.79e-03    0.00399    1.20 2.37e-01
```

```
## [1] 0.966
```

```
s=summary(lm(expend~employ+crime+pop,data=crime2))$coefficients #1.953e-04
s;summary(lm(expend~employ+crime+pop,data=crime2))$r.squared #0.974
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -249.8556   52.56746   -4.75 2.17e-05
## employ         0.0210    0.00388    5.42 2.41e-06
## crime          0.0551    0.01041    5.29 3.68e-06
## pop            0.0708    0.01742    4.06 1.95e-04
```

```
## [1] 0.974
```

In the third round, only pop is significant, given that p-value=1.95e-04. Then we add pop to the linear model.

```
summary(lm(expend~employ+crime+pop+bad,data=crime2))$coefficients #0.521
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -263.2096   56.79091  -4.635 3.32e-05
## employ         0.0214    0.00394   5.419 2.54e-06
## crime          0.0577    0.01123   5.135 6.51e-06
## pop            0.0757    0.01906   3.970 2.69e-04
## bad           -0.5846    0.90241  -0.648 5.21e-01
```

```
summary(lm(expend~employ+crime+pop+lawyers,data=crime2))$coefficients #0.227
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.33e+02   54.13308   -4.30 9.74e-05
## employ       1.89e-02    0.00421    4.49 5.24e-05
## crime        5.04e-02    0.01103    4.57 4.13e-05
## pop          7.00e-02    0.01734    4.04 2.20e-04
## lawyers      4.22e-03    0.00344    1.23 2.27e-01
```

If continuously add the fourth possible variable bad or lawyers, we find that p-value is 0.521 and 0.227,so these two variables are not significant. Thus we terminate the step-up method here and the final model is $expend = -249.856 + 0.021 * employ + 0.055 * crime + 0.071 * pop + e_n$, with $R^2 = 0.974$.

Now we move to the step-down method and start with a full model, containing all explanatory variables. The first variable removed is 'bad' since it has the highest p-value 0.677 (>0.05).

```
summary(lm(expend~bad+crime+lawyers+employ+pop,data=crime2))$coefficients # remove crime
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.43e+02   59.62350  -4.068 2.04e-04
## bad         -3.85e-01    0.91791  -0.419 6.77e-01
## crime        5.24e-02    0.01217   4.309 9.69e-05
## lawyers      3.93e-03    0.00355   1.109 2.74e-01
## employ       1.93e-02    0.00435   4.439 6.43e-05
## pop          7.32e-02    0.01913   3.825 4.27e-04
```

```
summary(lm(expend~crime+lawyers+employ+pop,data=crime2))$coefficients # remove lawyers
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.33e+02   54.13308   -4.30 9.74e-05
## crime        5.04e-02    0.01103    4.57 4.13e-05
## lawyers      4.22e-03    0.00344    1.23 2.27e-01
## employ       1.89e-02    0.00421    4.49 5.24e-05
## pop          7.00e-02    0.01734    4.04 2.20e-04
```

"lawyers" has the highest p-value 0.227 in the second round and it should be deleted. Then the model becomes $expend \sim crime + employ + pop$.

```
summary(lm(expend~crime+employ+pop,data=crime2))$coefficients # stop
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -249.8556   52.56746   -4.75 2.17e-05
## crime          0.0551    0.01041    5.29 3.68e-06
## employ         0.0210    0.00388    5.42 2.41e-06
## pop            0.0708    0.01742    4.06 1.95e-04
```

In this round, all the remaining variables are significant as p-value smaller than 0.05. Thus we stop the step-down method here and the final result is $expend = -249.856 + 0.021 * employ + 0.055 * crime + 0.071 * pop + e_n$, with $R^2 = 0.974$. These two methods eventually have the same result.

c)

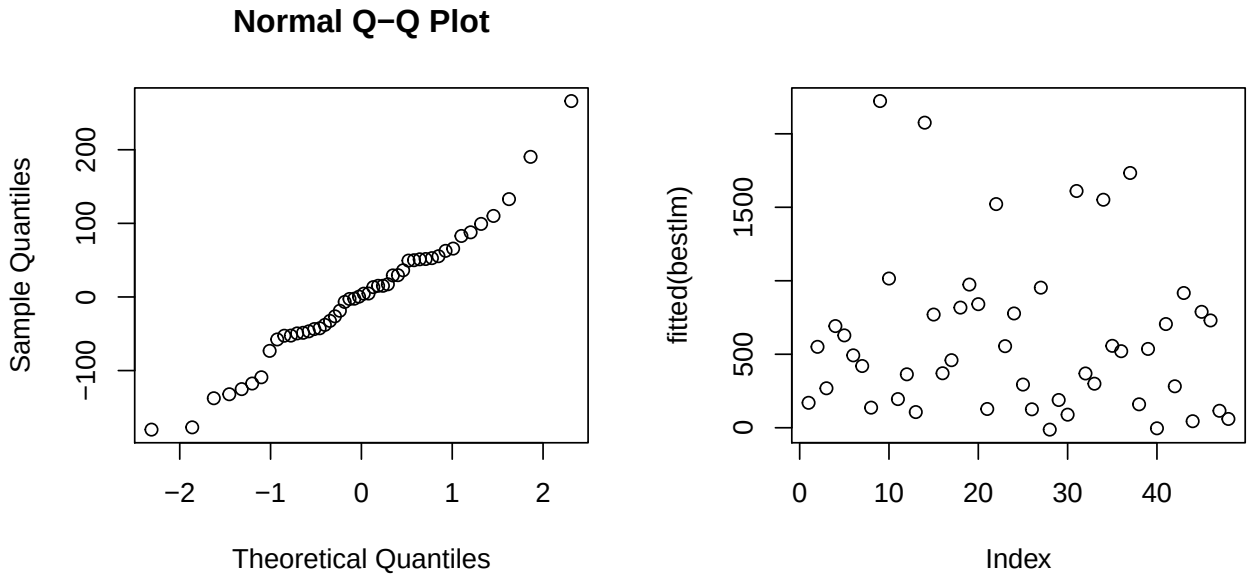
```
bestlm=lm(expend~crime+employ+pop,data=crime2);par(mfrow=c(1,2))
cor(crime2$crime,crime2$expend);cor(crime2$employ,crime2$expend);cor(crime2$pop,crime2$expend)
```

```
## [1] 0.301
```

```
## [1] 0.977
```

```
## [1] 0.96
```

```
qqnorm(residuals(bestlm));plot(fitted(bestlm));shapiro.test(residuals(bestlm))
```




```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(bestlm)
## W = 1, p-value = 0.3
```

As we already seen in the scatter plot, there are strong linearities between $expend \sim employ$ and $expend \sim pop$. We could confirm this assumption as the correlation is 0.977 and 0.960 respectively. But somehow $expend \sim employ$ does not have a clear linear relation. If we look at the qq-norm plot of the residual of this model, it looks good as almost all the points are on a straight line. Besides, the scatter plot of this fitted model does not have a systematical pattern. Thus, this model fits the data well.