STAT 461: Midterm 2 (Take-Home)

Xiangyu Ren

Oct 14, 2020

This is a take-home exam. You are allowed to use any non-human sources (internet, books, notes, etc), but you are NOT allowed to receive help from or work with any other person. If the instructor feels that cheating may have happened, an oral component will be added to the exam, with students each individually explaining their work. Any cheating will be addressed in accordance with Penn State's Academic Integrity policies.

For each experiment below, conduct a full analysis of the experimental data, and answer any additional questions as stated in each problem. **In all cases, you should check model assumptions and make transformations to the response variable as needed.** You must present your answers in a clear manner. That is, only show R code for your final model selection (that is you do not need to show all transformations you try). The answers to the question should be cleanly typed up and easy to read. Your answer should contain all R code used, and you should describe the results of all important hypothesis tests you conduct. You must provide your raw .Rmd code otherwise your exam will not be graded! You must submit the output as a HTML, PDF (preferred) or Word Document file.

This exam is due by 8:59 November 14th, 2020 EST.

Question 1 (25 points)

An experiment was run in order to compare the effects of different microbes on the production of healthy gut factors. Host lower intestine biopsies were studied in test subjects with inflammatory bowel disease. These cells were provided different bacterial supplements via capsule. One week after taking a supplement, the subjects were called back in order to measure host glucose levels. Subjects were given one of two different supplements "Sutterella" and "Akkermansia" at varying levels. The quantity of the respective supplement was varied at three levels: "low", "medium", and "heavy" dosages. Thus, there were two different treatments. "bact" indicates which bacterial supplement a patient was provivded. "dosage level" measures the amount of supplement they were given. The scientists are interested in determining the effects of bacterial supplement and dosage on glucose levels. The data are as follows:

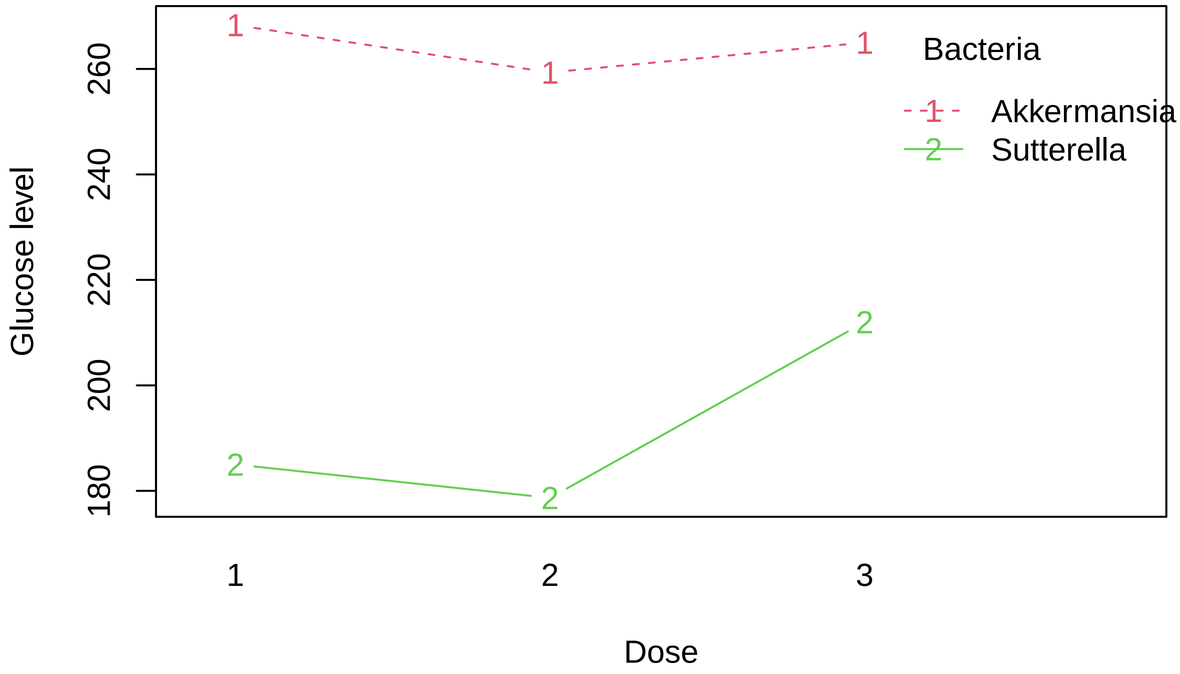
```r
bact=c(rep("Sutterella",9),rep("Akkermansia",9))
dose=rep(rep(1:3,each=3),2)
glucose.level = c(204,170,181,167,182,187,202,198,236,257,279,269,283,235,260,256,281,258)
```

1.1 Give a plot of either the response variable (glucose.level), or the mean response variable, versus the two treatment factors: bacteria and dose Your plot or plots should make it clear which treatments correspond to which response variables.

```r
df = data.frame(bact = as.factor(bact), dose = as.factor(dose), glucose.level)
df
```

```
##           bact dose glucose.level
## 1   Sutterella    1           204
## 2   Sutterella    1           170
## 3   Sutterella    1           181
## 4   Sutterella    2           167
## 5   Sutterella    2           182
## 6   Sutterella    2           187
## 7   Sutterella    3           202
## 8   Sutterella    3           198
## 9   Sutterella    3           236
## 10 Akkermansia    1           257
## 11 Akkermansia    1           279
## 12 Akkermansia    1           269
## 13 Akkermansia    2           283
## 14 Akkermansia    2           235
## 15 Akkermansia    2           260
## 16 Akkermansia    3           256
## 17 Akkermansia    3           281
## 18 Akkermansia    3           258
```

```r
interaction.plot(x.factor = df$dose, trace.factor = df$bact, response = df$glucose.level,
                 type = "b", col = 2:3, xlab = "Dose", ylab = "Glucose level",
                 trace.label = "Bacteria" )
```



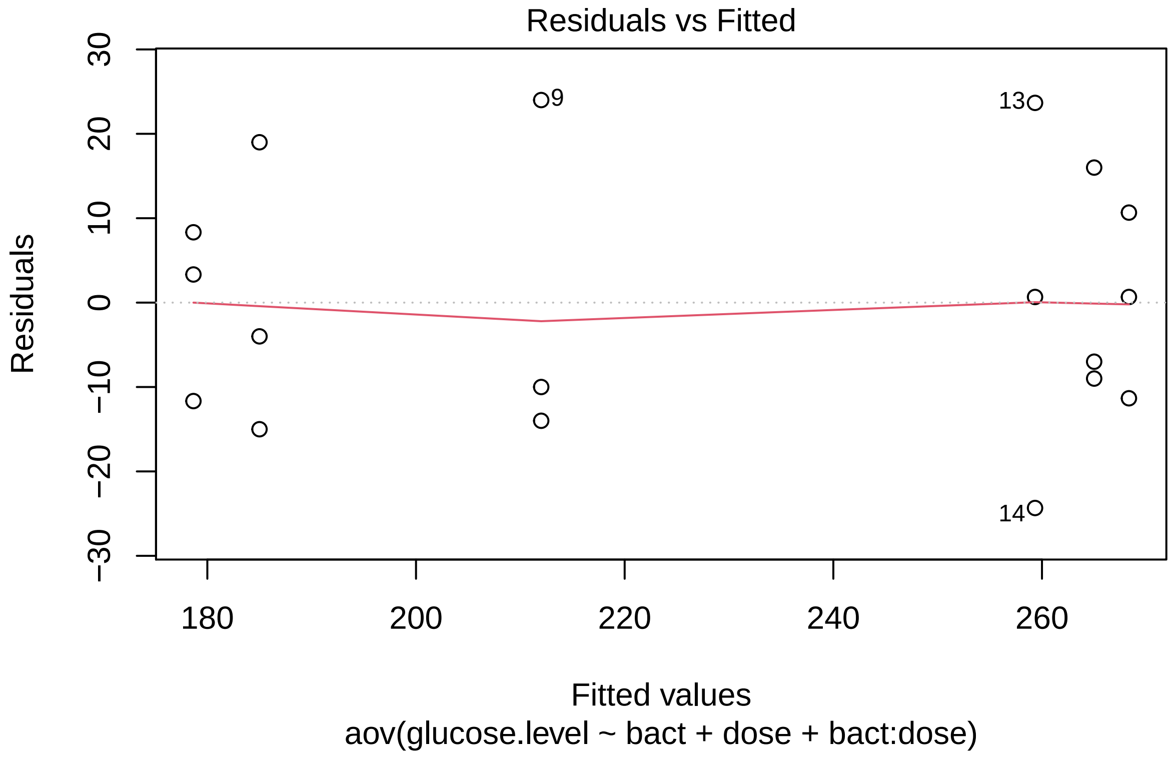
1.2 Give a complete analysis of this experimental data. You should show all R code used, write out the model, and explain all important choices and results in your analysis. Interpret the results in the context of the experiment, including pairwise differences if required.

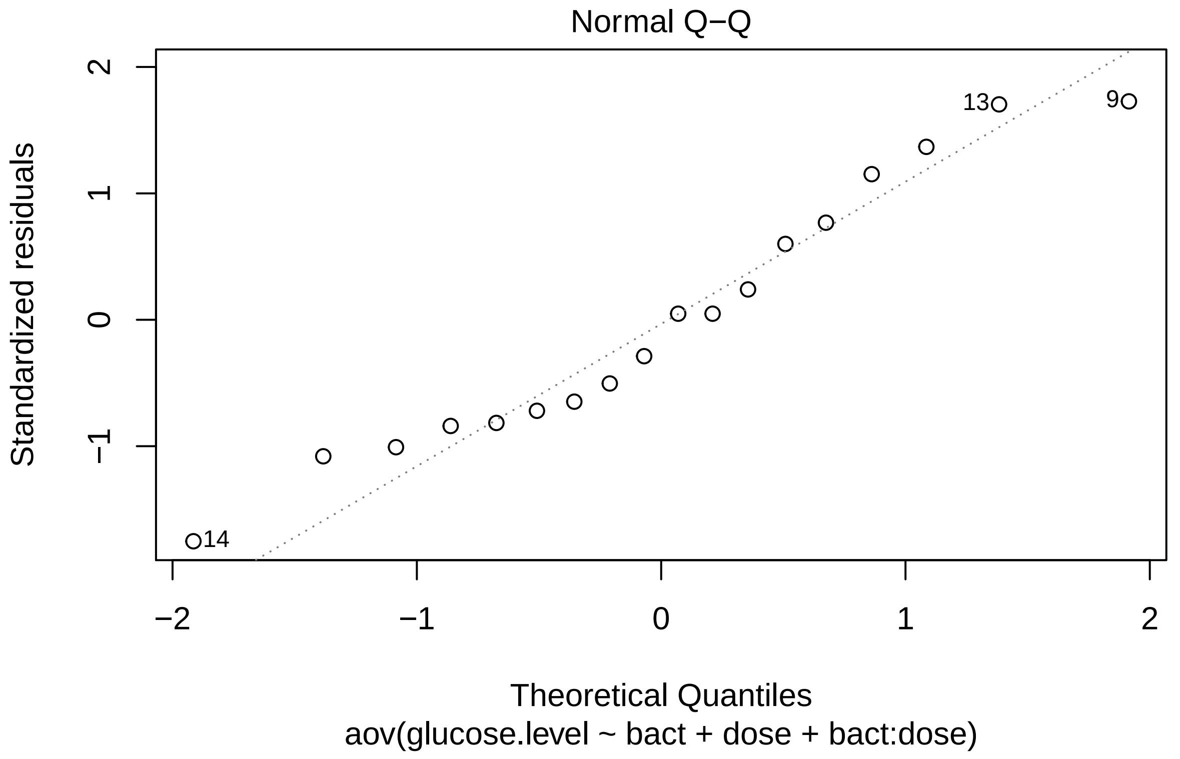
$$Y_{ijt} = \mu + \alpha_i + \beta_{j(i)} + \epsilon_{ijt}, \quad \epsilon_{ijt} \overset{iid}{\sim} N(0, \sigma^2)$$
$$i = Ak, Su \;\; j = 1, 2, 3 \;\; t = 1, 2, 3$$

```
library(car)
model1 = aov(glucose.level ~ bact + dose + bact:dose, data = df)
Anova(model1, type = "III")
```

```
## Anova Table (Type III tests)
##
## Response: glucose.level
##             Sum Sq Df  F value    Pr(>F)
## (Intercept) 216008  1 746.5726 3.559e-12 ***
## bact         10417  1  36.0023 6.215e-05 ***
## dose           124  2   0.2147    0.8098
## bact:dose      846  2   1.4626    0.2701
## Residuals     3472 12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(model1, which = c(1,2))
```





The residuals are approximately normal because the line of QQ-plot is nearly a straight line.

The assumption of constant error variance among treatments is close to justified, since the residuals are about separated in a square shape.

```
modelnint = aov(glucose.level ~ bact + dose, data = df)
anova(modelnint)
```

```
## Analysis of Variance Table
##
## Response: glucose.level
##           Df  Sum Sq Mean Sq F value   Pr(>F)
## bact       1 23544.5 23544.5 76.3311 4.84e-07 ***
## dose       2  1158.1   579.1  1.8773   0.1895
## Residuals 14  4318.3   308.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Transformation seems doesn't make a significant difference.

Give hypothesis test with following $\alpha = 0.05$

$$H_0 : (\alpha\beta)_{ij} = 0 \text{ for all } i, j \;\; vs \;\; H_a : \text{at least one treatment is different.}$$

```
library(lsmeans)
library(multcompView)
```

```r
library(multcomp)
library(knitr)
lsm.d = lsmeans(model1, ~ dose)
contrast(lsm.d, method = "pairwise")
```

```
##  contrast estimate   SE df t.ratio p.value
##  1 - 2        7.67 9.82 12   0.781  0.7215
##  1 - 3      -11.83 9.82 12  -1.205  0.4727
##  2 - 3      -19.50 9.82 12  -1.986  0.1582
##
## Results are averaged over the levels of: bact
## P value adjustment: tukey method for comparing a family of 3 estimates
```

```r
lsm.b = lsmeans(model1, ~ bact)
contrast(lsm.b, method = "pairwise")
```

```
##  contrast                 estimate   SE df t.ratio p.value
##  Akkermansia - Sutterella     72.3 8.02 12 9.021   <.0001
##
## Results are averaged over the levels of: dose
```

```r
lsm.inter = lsmeans(model1, ~ bact:dose)
cld(lsm.inter)
```

```
##  bact        dose lsmean   SE df lower.CL upper.CL .group
##  Sutterella  2       179 9.82 12      157      200  1
##  Sutterella  1       185 9.82 12      164      206  1
##  Sutterella  3       212 9.82 12      191      233  1
##  Akkermansia 2       259 9.82 12      238      281   2
##  Akkermansia 3       265 9.82 12      244      286   2
##  Akkermansia 1       268 9.82 12      247      290   2
##
## Confidence level used: 0.95
## P value adjustment: tukey method for comparing a family of 6 estimates
## significance level used: alpha = 0.05
```

Since bact:dose is larger than α, therefore we say we fail to reject the null and we can conclude that there are not significant interactions.

There is no significant difference between Elapsed Times since the p-value is greater than α, but there are significant differences between bacterial supplements since p-value is smaller than α.

Question 2 (25 pts)

A scientist wishes to study the boiling time of three polymers (coded P1–P3) and the industrial standard (coded P4). Thus, one can view the industrial standard as the control. These were boiled one by one with the system being reset each time before a new polymer was tested.

```r
poly <- c(rep("P1",10), rep("P2", 10), rep("P3",10), rep("P4", 10))
boiling <- c(167, 171, 178, 175, 184, 176, 185, 172, 178, 178,
         231, 233, 236, 252, 233, 225, 241, 248, 239, 248,
         176, 168, 171, 172, 178, 176, 169, 164, 169, 171,
         201, 199, 196, 211, 209, 223, 209, 219, 212, 210)
```

Give a complete analysis of this experimental data and answer if the type of polymer affects boiling time. You should show all R code used, and explain all important choices and results in your analysis. Interpret the results in the context of the experiment, including pairwise differences if required.

$$Y_{it} = \mu + \tau_i + \epsilon_{it}, \quad i = P_1, P_2, P_3, P_4 \;\; t = 1, 2, ..., 10$$

$$\epsilon_{it} \overset{iid}{\sim} N(0, \sigma^2)$$

Give hypothesis test with following $\alpha = 0.05$

$$H_0 : \tau_1 = \tau_2 = \tau_3 = \tau_4 \;\; vs. \;\; H_a : \text{At least one type of polymer is different from others}$$

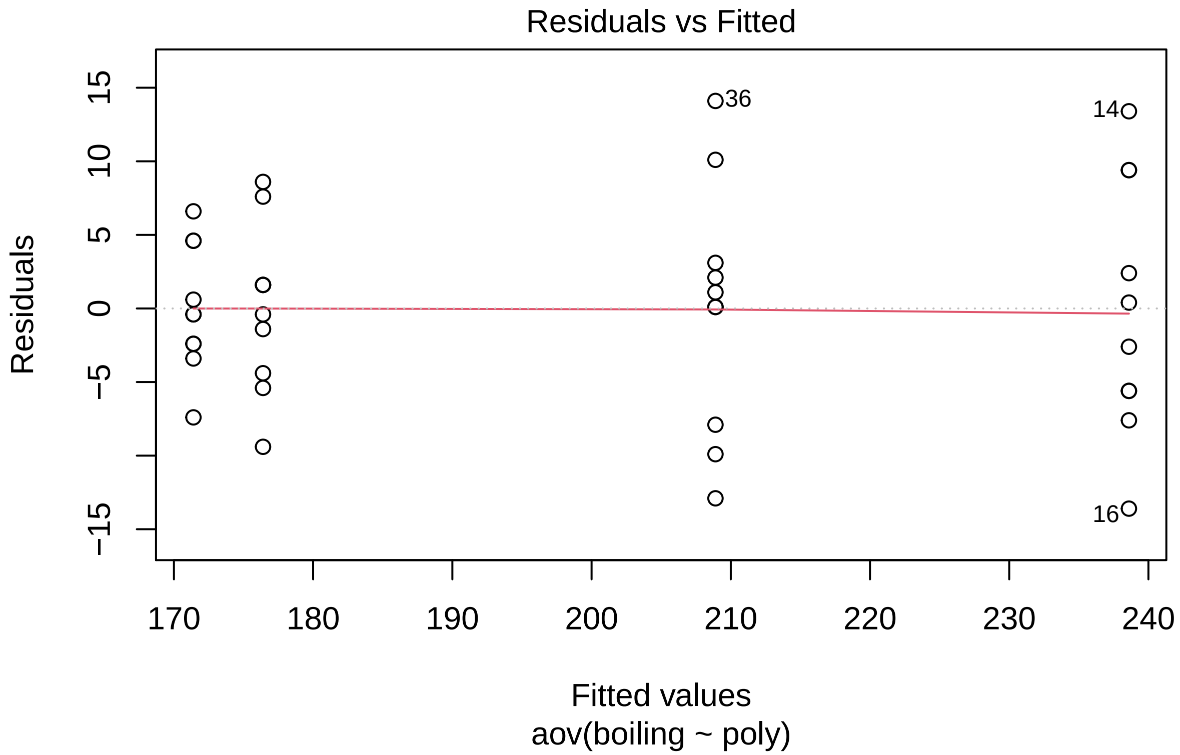
```r
df = data.frame(poly, boiling)
df
```

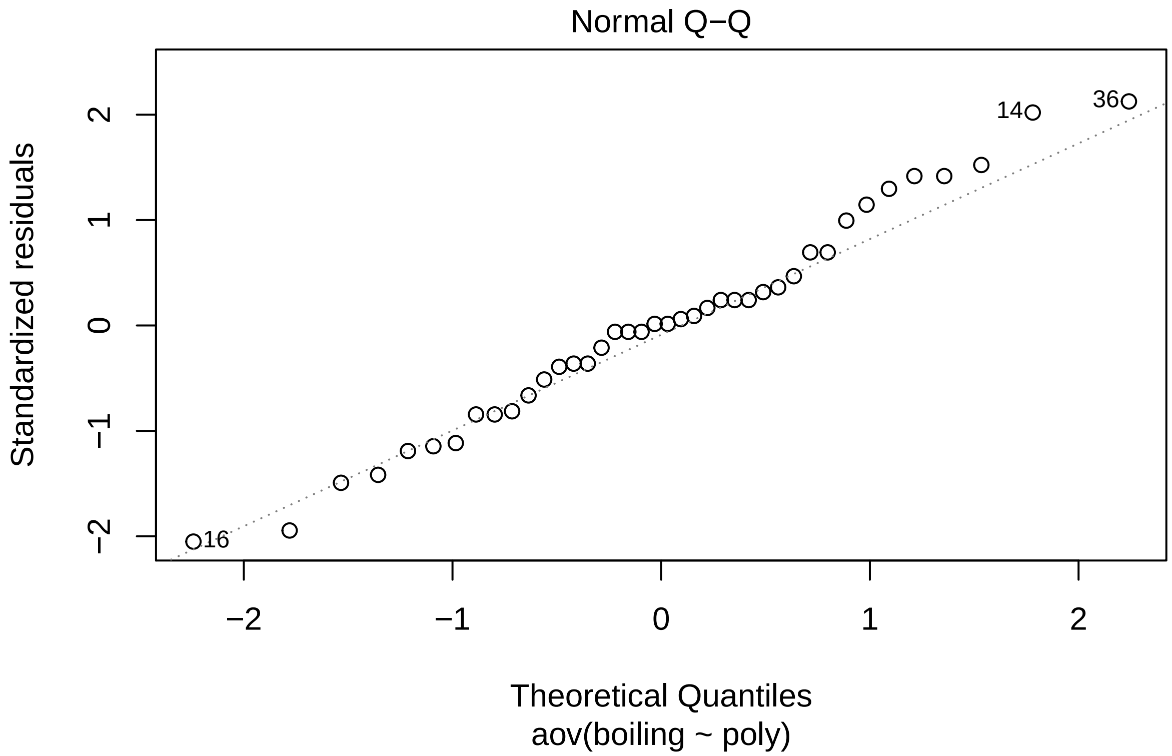
```
##    poly boiling
## 1    P1     167
## 2    P1     171
## 3    P1     178
## 4    P1     175
## 5    P1     184
## 6    P1     176
## 7    P1     185
## 8    P1     172
## 9    P1     178
## 10   P1     178
## 11   P2     231
## 12   P2     233
## 13   P2     236
## 14   P2     252
## 15   P2     233
## 16   P2     225
## 17   P2     241
## 18   P2     248
## 19   P2     239
## 20   P2     248
## 21   P3     176
## 22   P3     168
## 23   P3     171
## 24   P3     172
## 25   P3     178
## 26   P3     176
## 27   P3     169
## 28   P3     164
## 29   P3     169
## 30   P3     171
## 31   P4     201
## 32   P4     199
## 33   P4     196
## 34   P4     211
## 35   P4     209
## 36   P4     223
## 37   P4     209
## 38   P4     219
## 39   P4      212
## 40   P4      210
```

```
model2 = aov(boiling ~ poly, data = df)
anova(model2)
```

```
## Analysis of Variance Table
##
## Response: boiling
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## poly       3 29385.7  9795.2  200.35 < 2.2e-16 ***
## Residuals 36  1760.1    48.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(model2, which = c(1, 2))
```





The residuals are normal because the line in QQ-plot is nearly a straight line.

The assumption of constant error variance among treatments is justified, since the residuals are about separated in a square shape.

```
lsm.model2 = lsmeans(model2, "poly")
kable(summary(contrast(lsm.model2, method = "pairwise", adjust = "tukey"), infer = c(T, T),
              level = 0.95, side = "two-sided"))
```

contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
P1 - P2	-62.2	3.127033	36	-70.621809	-53.77819	-19.891062	0.0000000
P1 - P3	5.0	3.127033	36	-3.421809	13.42181	1.598960	0.3919934
P1 - P4	-32.5	3.127033	36	-40.921809	-24.07819	-10.393240	0.0000000
P2 - P3	67.2	3.127033	36	58.778191	75.62181	21.490022	0.0000000
P2 - P4	29.7	3.127033	36	21.278191	38.12181	9.497822	0.0000000
P3 - P4	-37.5	3.127033	36	-45.921809	-29.07819	-11.992200	0.0000000

```
lsm.inter = lsmeans(model2, ~ poly)
cld(lsm.inter)
```

```
##  poly lsmean   SE df lower.CL upper.CL .group
##  P3      171 2.21 36      167      176  1
##  P1      176 2.21 36      172      181  1
##  P4      209 2.21 36      204      213   2
##  P2      239 2.21 36      234      243    3
##
## Confidence level used: 0.95
## P value adjustment: tukey method for comparing a family of 4 estimates
## significance level used: alpha = 0.05
```

Since only the p-value of P1-P3 is great than α, we reject the null, therefore all contrasts except of P1-P3 are significant and there is at least one of the polymer that is different from others.

Question 3 (25 points)

An analysis is conducted to determine if different species and types of wood influence the nitrogen content in specific trees. Trees are divided into two kinds of wood: hard wood (oak, ash, and maple), and soft (pine, spruce, and fir). A random selection of 4 trees of each kind (24 trees total) was chosen from all trees in the State Game Lands, and the nitrogen content was measured.

```
wood=read.table("wood.csv",header=TRUE)
wood
```

```
##        Type Species Nconc
## 1  softwood    pine    12
## 2  softwood    pine    13
## 3  softwood    pine    11
## 4  softwood    pine    12
## 5  softwood  spruce    15
## 6  softwood  spruce    19
## 7  softwood  spruce    17
## 8  softwood  spruce    17
## 9  softwood     fir    10
## 10 softwood     fir    12
## 11 softwood     fir    11
## 12 softwood     fir    17
## 13 hardwood   maple    18
## 14 hardwood   maple    20
## 15 hardwood   maple    21
## 16 hardwood   maple    16
## 17 hardwood     oak    20
## 18 hardwood     oak    14
## 19 hardwood     oak    17
## 20 hardwood     oak    15
## 21 hardwood     ash    19
## 22 hardwood     ash    22
## 23 hardwood     ash    21
## 24 hardwood     ash    21
```

3.1 Is this experiment a completely randomized design? Why or why not?

No, this is not a completely randomized design, because all the trees are been selected from the same state, also the types and species are fixed.

3.2 Give a complete analysis of this data. Show all R code used, and explain all important choices and results in your analysis. Interpret the results in the context of the experiment, including pairwise differences if required.

$$Y_{ijt} = \mu + \alpha_i + \beta_{j(i)} + \epsilon_{ijt}, \;\; \epsilon_{ijt} \overset{iid}{\sim} N(0, \sigma^2)$$

$$i = SW, HW \;\; j = P, S, F, M, O, A \;\; t = 1, 2, 3, 4$$

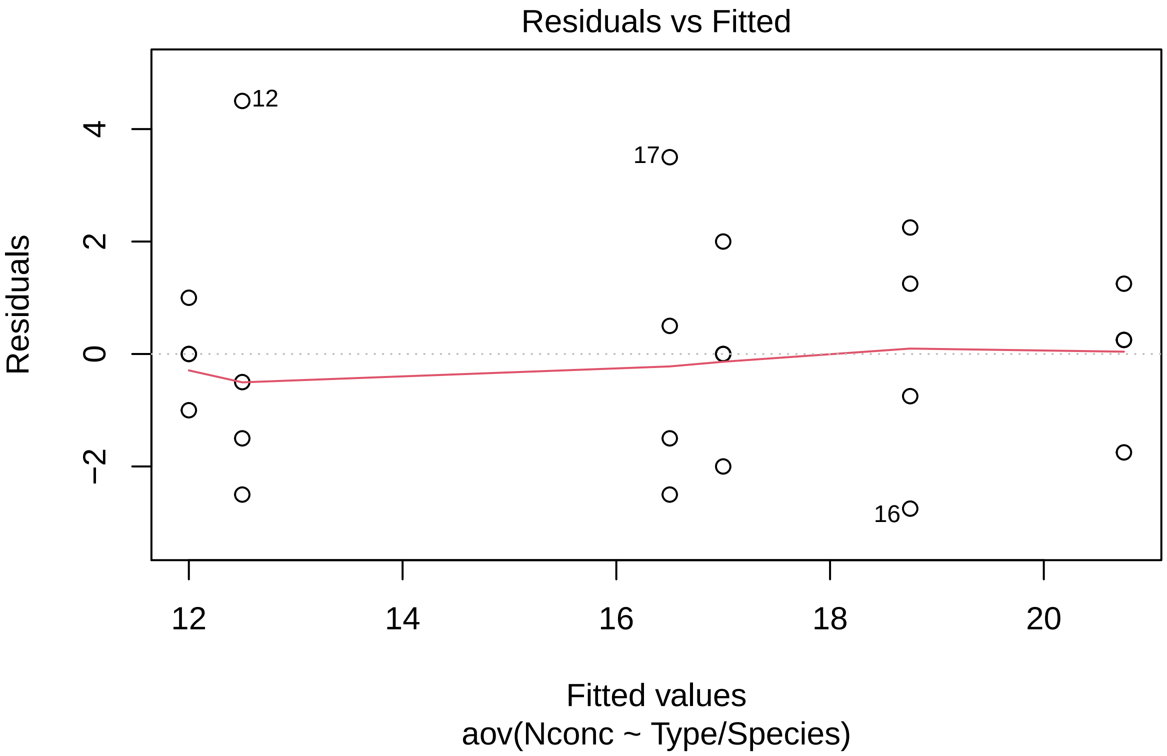
```
library(car)
df = data.frame(Type = wood$Type, Species = wood$Species, Nconc = wood$Nconc)
df
```

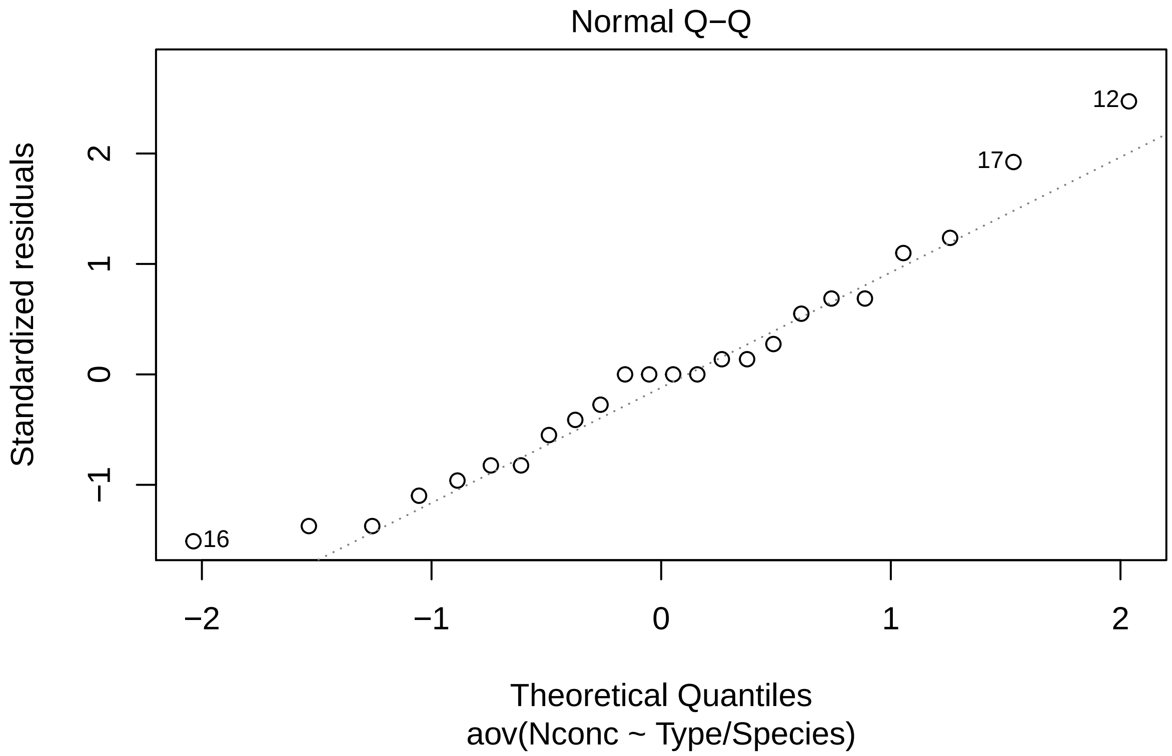
```
##        Type Species Nconc
## 1  softwood    pine    12
## 2  softwood    pine    13
## 3  softwood    pine    11
## 4  softwood    pine    12
## 5  softwood  spruce    15
## 6  softwood  spruce    19
## 7  softwood  spruce    17
## 8  softwood  spruce    17
## 9  softwood     fir    10
## 10 softwood     fir    12
## 11 softwood     fir    11
## 12 softwood     fir    17
## 13 hardwood   maple    18
## 14 hardwood   maple    20
## 15 hardwood   maple    21
## 16 hardwood   maple    16
## 17 hardwood     oak    20
## 18 hardwood     oak    14
## 19 hardwood     oak    17
## 20 hardwood     oak    15
## 21 hardwood     ash    19
## 22 hardwood     ash    22
## 23 hardwood     ash    21
## 24 hardwood     ash    21
```

```
model3 = aov(Nconc ~ Type / Species, data = df)
Anova(model3)
```

```
## Anova Table (Type II tests)
##
## Response: Nconc
##                Sum Sq Df F value    Pr(>F)
## Type          140.167  1 31.7358 2.408e-05 ***
## Type:Species   96.833  4  5.4811  0.004574 **
## Residuals      79.500 18
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
plot(model3, which = c(1,2))
```


Residuals vs Fitted
Residuals
4
2
0
−2
12
17
16
12
14
16
18
20
Fitted values
aov(Nconc ~ Type/Species)



The residuals are normal because the line in QQ-plot is nearly a straight line.

The assumption of constant error variance among treatments is close to justified, since the residuals are about closed in a square shape.

Transformation is not necessary since all transformations seems close to original model.

Give hypothesis test with following $\alpha = 0.05$

$$H_0 : \alpha_{Softwood} = \alpha_{Hardwood} \;\; vs. \;\; H_a : \alpha_{Softwood} \neq \alpha_{Hardwood}$$

```
lsm.t = lsmeans(model3, ~ Type)
cld(lsm.t, alpha = 0.05)
```

```
##  Type     lsmean    SE df lower.CL upper.CL .group
##  softwood   13.8 0.607 18     12.6     15.1  1
##  hardwood   18.7 0.607 18     17.4     19.9   2
##
## Results are averaged over the levels of: Species
## Confidence level used: 0.95
## significance level used: alpha = 0.05
```

looking at the output, we see softwood is significant different from hardwood.

Since, the p-value is less than 0.05, thus we reject the null and we can conclude that there are significant differences between different types.

$$H_0 : \left\{ \begin{array}{c} \beta_{1(softwood)} = \beta_{2(softwood)} = \beta_{3(softwood)} = \beta_{4(softwood)} = \beta_{5(softwood)} = \beta_{6(softwood)} \\ \vdots \end{array} \right\}$$

$$vs.\ H_a\text{: The null is not true}$$

```
lsm.s = lsmeans(model3, ~ Type:Species)
cld(lsm.s, alpha = 0.05)
```

```
##  Species Type     lsmean   SE df lower.CL upper.CL .group
##  pine    softwood   12.0 1.05 18     9.79     14.2  1
##  fir     softwood   12.5 1.05 18    10.29     14.7  12
##  oak     hardwood   16.5 1.05 18    14.29     18.7  123
##  spruce  softwood   17.0 1.05 18    14.79     19.2   23
##  maple   hardwood   18.8 1.05 18    16.54     21.0    3
##  ash     hardwood   20.8 1.05 18    18.54     23.0    3
##
## Confidence level used: 0.95
## P value adjustment: tukey method for comparing a family of 6 estimates
## significance level used: alpha = 0.05
```

looking at the output, we can see pine is significant different from spruce, maple and ash

Since, the p-value is less than 0.05, thus we reject the null and we can conclude that there are significant differences between different species from same type.

Question 4 (25 points)

An experiment was conducted to determine the best recipe for different kinds of canned beans. Beans are divided into four different crocks (i=1,2,3,4). The beans are soaked before cooking for either a long or a short time (j=short,long). Two of the crocks are randomly chosen to soak for a short time, and the other two crocks are allowed to soak for a long time. After soaking, the beans from each crock are divided into three jars, and are used to make baked beans using one of three recipes (k=Original, Barbecue, or Refried). Finally, beans from each jar are fed to people, and the average taste rating of for each jar is recorded.

```
beans=read.table("Beans.csv",header=TRUE)
beans
```

```
##    Crock SoakTim   Recipe Jar Rating
## 1      1    Long Original   1     45
## 2      1    Long Barbecue   2     50
## 3      1    Long  Refried   3     44
## 4      2   Short Original   4     33
## 5      2   Short Barbecue   5     40
## 6      2   Short  Refried   6     40
## 7      3    Long Original   7     46
## 8      3    Long Barbecue   8     49
## 9      3    Long  Refried   9     45
## 10     4   Short Original  10     32
## 11     4   Short Barbecue  11     41
## 12     4   Short  Refried  12     41
```

4.1 Explain why Jar is not treated as a factor in this experiment.

Because Jar is not either crossed or nested with any other factors, instead it is only a counting figure.

4.2 Nested models can be combined with complete models in order to yield more complex models. In these cases, one can add multiple treatments in order to build a more complex model. Thus, a two-way model can be extended further into a more general “k”-factor model.

Give a complete analysis of this data, under the following model.

$$Y_{ijk} = \mu + \alpha_j + \beta_{i(j)} + \gamma_k + (\alpha\gamma)_{jk} + \epsilon_{ijk}, \quad \epsilon_{ijk} \sim N(0, \sigma^2)$$

$$\beta_{i(j)} \sim N(0, \sigma^2_{crock})$$

Show all R code used, and explain all important choices and results in your analysis. Interpret the results in the context of the experiment.

Hint: You should be able to do this by extending the code for the different two-way models we have learned in class. As noted in class, reading the output of ANOVA models remain similar regardless of how many treatments you add. First, given the model shown above and the notation written in the question description, figure out what treatments correspond to α, β and γ.

```
library(lme4)
library(lmerTest)
library(car)
library(multcompView)

df = data.frame(crock = beans$Crock, soaktime = beans$SoakTim, recipe = beans$Recipe,
                rating = beans$Rating)
df
```

```
##    crock soaktime   recipe rating
## 1      1     Long Original     45
## 2      1     Long Barbecue     50
## 3      1     Long  Refried     44
## 4      2    Short Original     33
## 5      2    Short Barbecue     40
## 6      2    Short  Refried     40
## 7      3     Long Original     46
## 8      3     Long Barbecue     49
## 9      3     Long  Refried     45
## 10     4    Short Original     32
## 11     4    Short Barbecue     41
## 12     4    Short  Refried     41
```

```
model4 = lmer(log(rating) ~ + soaktime + (1 | crock:soaktime) + recipe + soaktime:recipe,
              data = df)
```

I picked log transformation because the residual seems more likely to be normal, and assumption of constant error variance among treatments looks like it’s closer to be justified.

```
anova(model4, which = "3")
```

```
## Type III Analysis of Variance Table with Satterthwaite's method
##                     Sum Sq  Mean Sq NumDF DenDF F value    Pr(>F)
## soaktime        0.132900 0.132900     1     6 447.650 7.266e-07 ***
## recipe          0.047713 0.023857     2     6  80.357 4.662e-05 ***
## soaktime:recipe 0.029505 0.014753     2     6  49.692 0.0001846 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
rand(model4)
```

```
## ANOVA-like table for random-effects: Single term deletions
##
## Model:
## log(rating) ~ soaktime + recipe + (1 | crock:soaktime) + soaktime:recipe
##                        npar logLik     AIC        LRT Df Pr(>Chisq)
## <none>                    8 13.773 -11.547
## (1 | crock:soaktime)    7 13.773 -13.547 2.4869e-14  1          1
```

```
lsm.r = lsmeans(model4, ~ soaktime:recipe)
cld(lsm.r, alpha = 0.05)
```

```
##  soaktime recipe   lsmean     SE df lower.CL upper.CL .group
##  Short    Original   3.48 0.0122  6     3.45     3.51  1
##  Short    Refried    3.70 0.0122  6     3.67     3.73   2
##  Short    Barbecue   3.70 0.0122  6     3.67     3.73   2
##  Long     Refried    3.80 0.0122  6     3.77     3.83    3
##  Long     Original   3.82 0.0122  6     3.79     3.85    3
##  Long     Barbecue   3.90 0.0122  6     3.87     3.93     4
##
## Degrees-of-freedom method: kenward-roger
## Results are given on the log (not the response) scale.
## Confidence level used: 0.95
## P value adjustment: tukey method for comparing a family of 6 estimates
## significance level used: alpha = 0.05
```

We give a hypothesis test with following $\alpha = 0.05$

$$H_0 : \sigma^2_{crock} = 0 \;\; vs. \;\; H_a : \sigma^2_{crock} \neq 0$$

Since the p-value for randomness is greater than α, we fail to reject the null, therefore we can conclude that beans are same with different crocks and the same soak time.

We then reject the null for everything else since the p-value for other variables with the same recipe is smaller than 0.05, therefore we can conclude that beans with different other variables with same recipes have different contributions.