Experiment Design and Analysis: Final assignment

Lu Dai (qx8329)

Instructions:

1. Carefully write down detailed procedures you used in each problem; Due date is May 13-14 at blackboard, but you can turn it in as soon as you finish it; Attach the R code for the required problems.

2. You may discuss with your peers but should finish it independently; Recall the definitions when you have difficulties. Good luck!

Problem 1: Exercise 12.4. Do a) and d) at p. 308

(a) One of the expenses in animal experiments is feeding the animals. A company salesperson has made the claim that their new rat chow (35% less expensive) is equivalent to the two standard chows on the market. You wish to test this claim by measuring weight gain of rat pups on the three chows. You have a population of 30 inbred, basically exchangeable female rat pups to work with, each with her own cage.

We can use completely randomized design to compare these three chows. The chow can be as fixed factor with three levels (brands). Each 10 rat pups that are randomly chosen from a population of 30 inbred are assigned to each of three chows. Weight gains are as response.

N=30units; g=3 treatments; n=10 (10 units of each group).

$\mathbf{Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}}$

(d) The disposable diaper business is very competitive, with all manufacturers trying to get a leg up, as it were. You are a consumer testing agency comparing the absorbency of two brands of “newborn” size diapers. The test is to put a diaper on a female doll and pump body temperature water through the doll into the diaper at a fixed rate until the diaper leaks. The response is the amount of liquid pumped before leakage. We are primarily interested in brand differences, but we are also interested in variability between individual diapers and between batches of diapers (which we can only measure as between boxes of diapers, since we do not know the actual manufacturing time or place of the diapers). We can afford to buy 32 boxes of diapers and test 64 diapers.

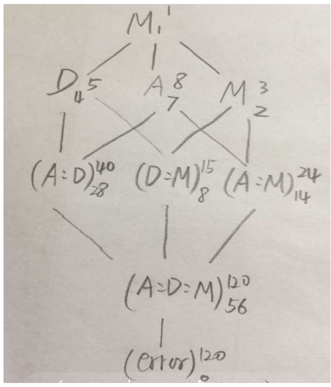
16 boxes are randomly selected from brand1, and 2 diapers are randomly selected from each box of brand1; meanwhile, 16 boxes are randomly selected from brand2, and 2 diapers are randomly selected from each box of brand2.

The brand is a fixed factor with 2 levels. The box is a random factor with 16 levels. Since boxes are selected from brands, the box is nested in the brands. 2 diapers are randomly selected from a box, so diapers are nested in boxes. So, the design should be mixed model with fixed effects and random effects.

$\mathbf{Y_{ijk} = \mu + \alpha_i + \beta j(i) + \varepsilon()k}$

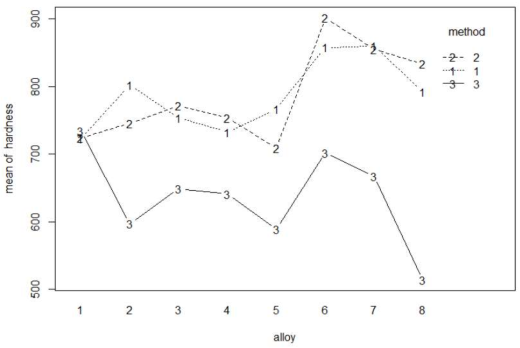
Problem 2: Problem 12.2 at p. 310. (data can be found at bb)

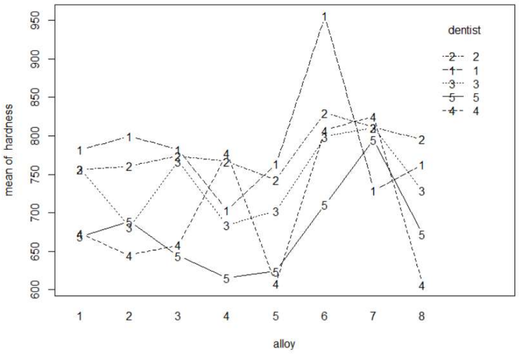
This case has three factors, dentist, alloy and method. I assume that dentist is the random factor, the method and alloy are fixed factors, and random interactions are independent. I assume the interaction of these 3 factors is as error due to a replication. This is mixed and unrestricted model.

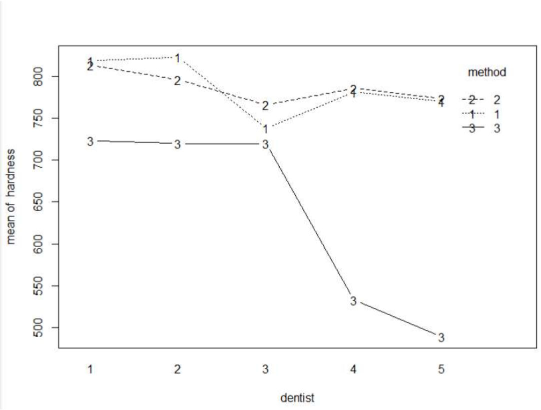


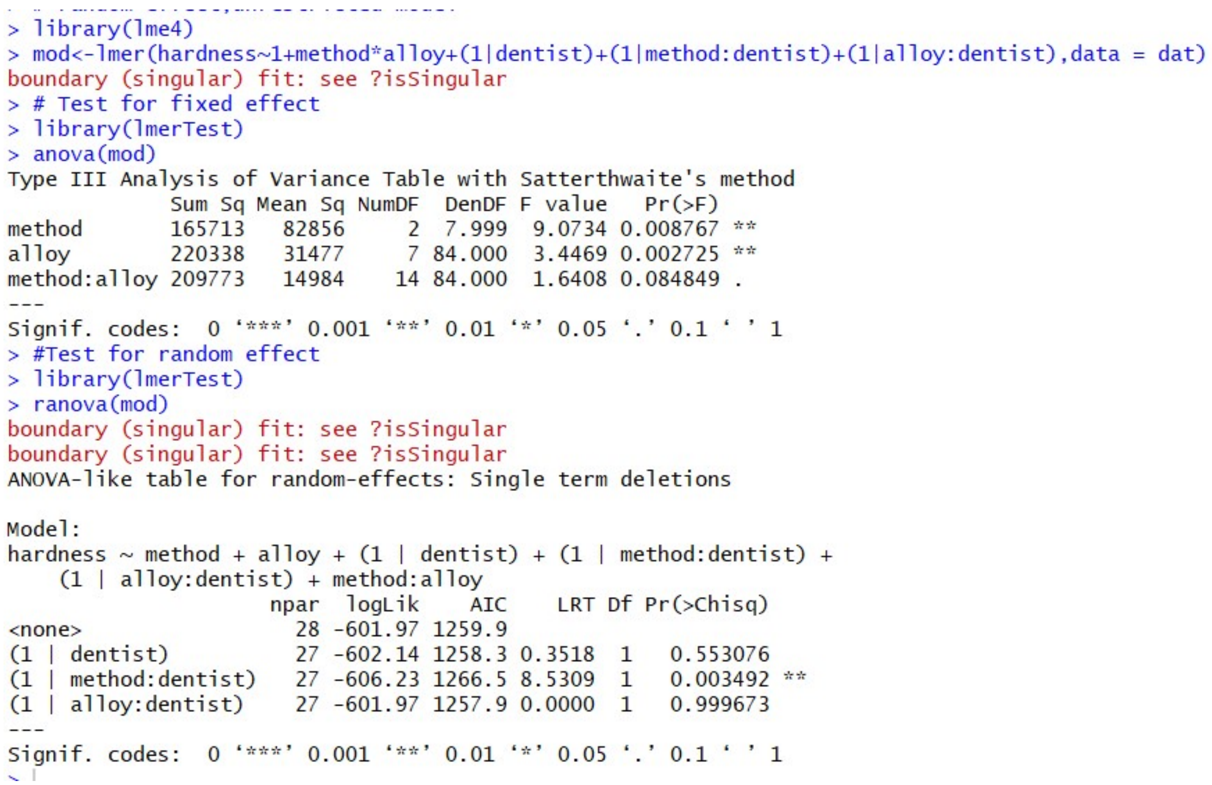
# check three factors interaction effects

I will use lmer to test fixed and random effects.







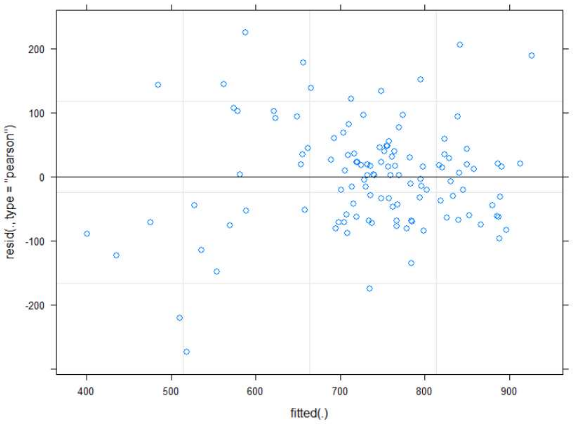


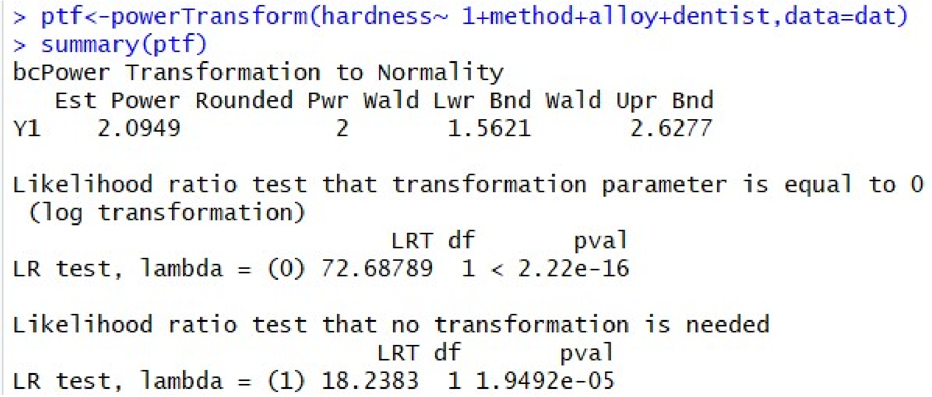
```
> library(lme4)
> mod<-lmer(hardness~1+method*alloy+(1|dentist)+(1|method:dentist)+(1|alloy:dentist),data = dat)
boundary (singular) fit: see ?isSingular
> # Test for fixed effect
> library(lmerTest)
> anova(mod)
Type III Analysis of Variance Table with Satterthwaite's method
             Sum Sq Mean Sq NumDF  DenDF F value   Pr(>F)   
method       165713   82856     2  7.999  9.0734 0.008767 **
alloy        220338   31477     7 84.000  3.4469 0.002725 **
method:alloy 209773   14984    14 84.000  1.6408 0.084849 . 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> #Test for random effect
> library(lmerTest)
> ranova(mod)
boundary (singular) fit: see ?isSingular
boundary (singular) fit: see ?isSingular
ANOVA-like table for random-effects: Single term deletions

Model:
hardness ~ method + alloy + (1 | dentist) + (1 | method:dentist) + 
    (1 | alloy:dentist) + method:alloy
                     npar  logLik    AIC    LRT Df Pr(>Chisq)   
<none>                 28 -601.97 1259.9                        
(1 | dentist)          27 -602.14 1258.3 0.3518  1   0.553076   
(1 | method:dentist)   27 -606.23 1266.5 8.5309  1   0.003492 **
(1 | alloy:dentist)    27 -601.97 1257.9 0.0000  1   0.999673   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

The p-values of method and alloy are less than 0.05, which means the method and alloy have significant effects (main effects). The interaction for method and alloy is insignificant (p-value is more than 0.05).

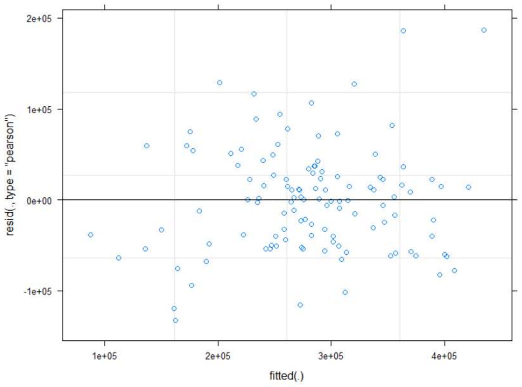
The p value of dentist is insignificant, which means dentist has no main effect for hardness. There are no interaction effects for alloy and dentist. But the interaction of method and dentist is significant (p-value=0.003493). The interaction plot indicates that dentist4,5 by method3 have effect most.

# check model assumptions: it looks a little non-constant variance. I try to use box-cox transformation to improve. Optimal power=2

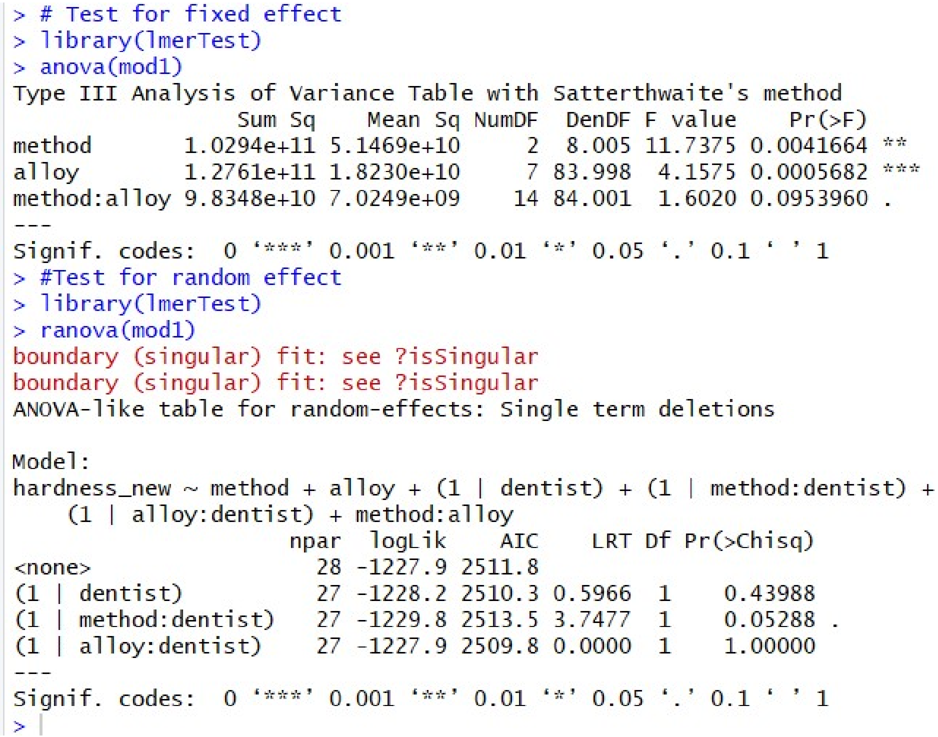




```
> ptf<-powerTransform(hardness~ 1+method+alloy+dentist,data=dat)
> summary(ptf)
bcPower Transformation to Normality
   Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Y1    2.0949           2       1.5621       2.6277

Likelihood ratio test that transformation parameter is equal to 0
 (log transformation)
                           LRT df       pval
LR test, lambda = (0) 72.68789  1 < 2.22e-16

Likelihood ratio test that no transformation is needed
                          LRT df       pval
LR test, lambda = (1) 18.2383  1 1.9492e-05
```



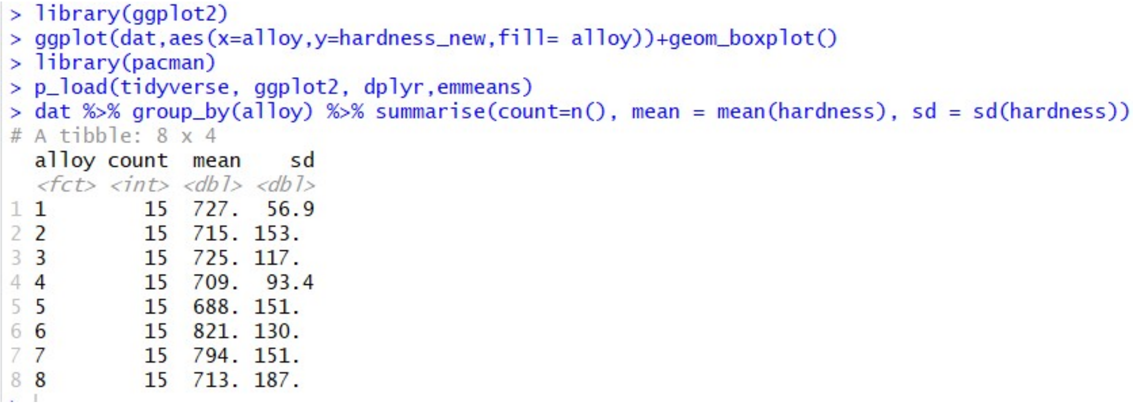
The plot is better.



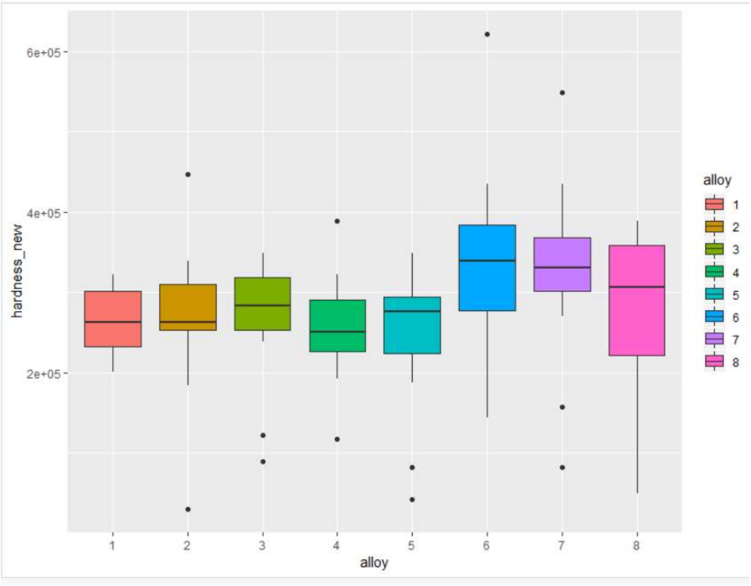
```
> # Test for fixed effect
> library(lmerTest)
> anova(mod1)
Type III Analysis of Variance Table with Satterthwaite's method
                 Sum Sq    Mean Sq NumDF  DenDF F value    Pr(>F)    
method       1.0294e+11 5.1469e+10     2  8.005 11.7375 0.0041664 ** 
alloy        1.2761e+11 1.8230e+10     7 83.998  4.1575 0.0005682 ***
method:alloy 9.8348e+10 7.0249e+09    14 84.001  1.6020 0.0953960 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> #Test for random effect
> library(lmerTest)
> ranova(mod1)
boundary (singular) fit: see ?isSingular
boundary (singular) fit: see ?isSingular
ANOVA-like table for random-effects: Single term deletions

Model:
hardness_new ~ method + alloy + (1 | dentist) + (1 | method:dentist) + 
    (1 | alloy:dentist) + method:alloy
                     npar  logLik    AIC    LRT Df Pr(>Chisq)  
<none>                 28 -1227.9 2511.8                       
(1 | dentist)          27 -1228.2 2510.3 0.5966  1    0.43988  
(1 | method:dentist)   27 -1229.8 2513.5 3.7477  1    0.05288 .
(1 | alloy:dentist)    27 -1227.9 2509.8 0.0000  1    1.00000  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

We assume null hypothesis is that three factors are not any effects for response (hardness), and all interaction effects are zeros. According to the test, the p-values of method and alloy are less more 0.05. We reject null hypothesis, which means the method and alloy have significant effects for the hardness. The dentist factor and interaction effects are not significant for hardness since the p-values are more than 0.05.

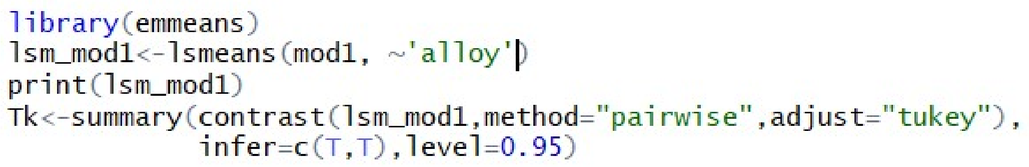
# check which method and alloy influence most the response(hardness):



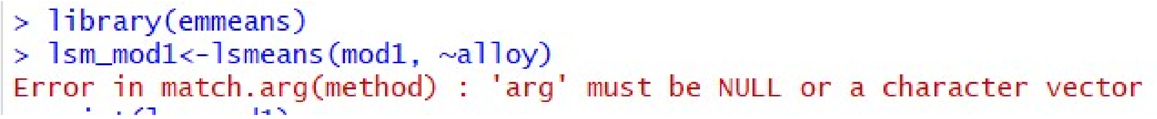
```
> library(ggplot2)
> ggplot(dat,aes(x=alloy,y=hardness_new,fill= alloy))+geom_boxplot()
> library(pacman)
> p_load(tidyverse, ggplot2, dplyr,emmeans)
> dat %>% group_by(alloy) %>% summarise(count=n(), mean = mean(hardness), sd = sd(hardness))
# A tibble: 8 x 4
  alloy count  mean    sd
  <fct> <int> <dbl> <dbl>
1 1        15  727.  56.9
2 2        15  715. 153.
3 3        15  725. 117.
4 4        15  709.  93.4
5 5        15  688. 151.
6 6        15  821. 130.
7 7        15  794. 151.
8 8        15  713. 187.
```



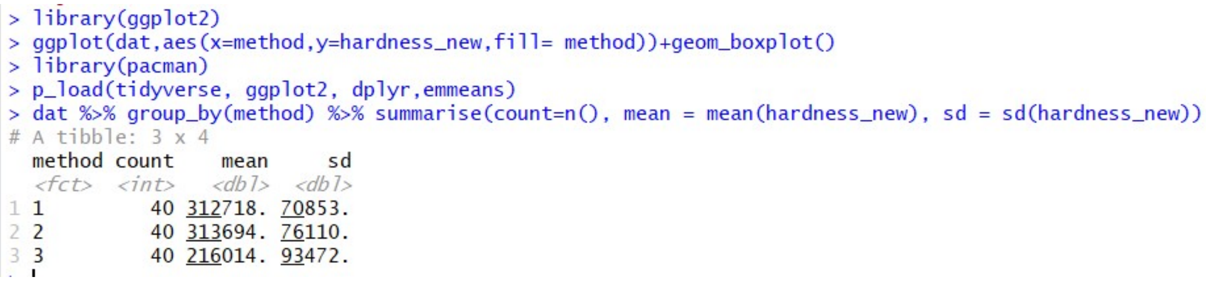
Based on the above methods, both results indicate that alloy6 is most influence the hardness. I used Tukey method to confirm the result. But the code has error. I also tried to use other contrast methods and different models on this data, still had same error. But same codes can work on the problem3. I have no more time to figure out this problem now. I will resolve it.



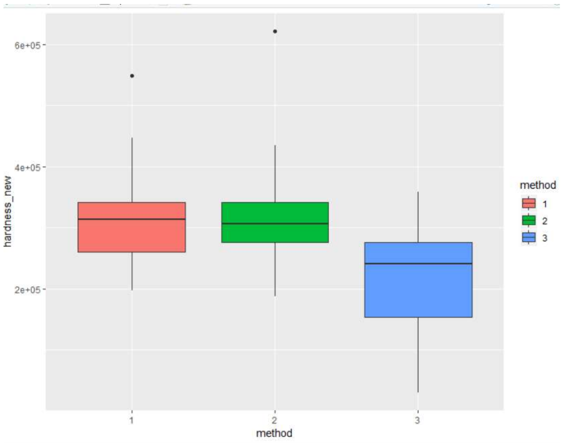
```
library(emmeans)
lsm_mod1<-lsmeans(mod1, ~'alloy')
print(lsm_mod1)
Tk<-summary(contrast(lsm_mod1,method="pairwise",adjust="tukey"),
             infer=c(T,T),level=0.95)
```



```
> library(emmeans)
> lsm_mod1<-lsmeans(mod1, ~alloy)
Error in match.arg(method) : 'arg' must be NULL or a character vector
```



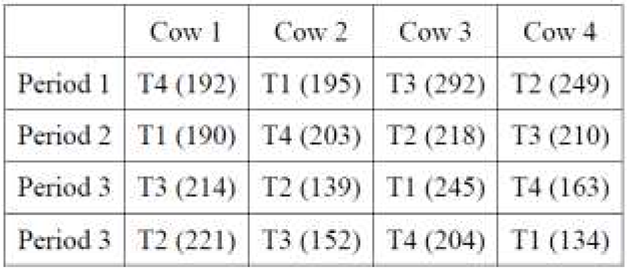
```
> library(ggplot2)
> ggplot(dat,aes(x=method,y=hardness_new,fill= method))+geom_boxplot()
> library(pacman)
> p_load(tidyverse, ggplot2, dplyr,emmeans)
> dat %>% group_by(method) %>% summarise(count=n(), mean = mean(hardness_new), sd = sd(hardness_new))
# A tibble: 3 x 4
  method count    mean     sd
  <fct>  <int>   <dbl>  <dbl>
1 1         40 312718. 70853.
2 2         40 313694. 76110.
3 3         40 216014. 93472.
```



The above results indicate the method1 and method2 look similar. These two methods are more influent to the response(hardness) than method3.

Problem 3:

An experiment to investigate the effects of various dietary starch levels on milk production was conducted on **four cows**. The **four diets, T1, T2, T3, and T4**, (in order of increasing starch equivalent), were fed for three weeks to each cow and the total yield of milk in the third week of each period was recorded (i.e. third week to minimize carry-over effects due to the use of treatments administered in a previous period). That is, the trial lasted 12 weeks since each cow received each treatment, and each treatment required three weeks. The investigator felt strongly that time period effects might be important (i.e earlier periods in the experiment might influence milk yields differently compared to later periods). Hence, the investigator wanted **to block on both cow and period**. However, each cow cannot possibly receive more than one treatment during the same time period; that is, all possible cow-period blocking combinations could not logically be considered. The randomization scheme and data are shown in the table below.



	Cow 1	Cow 2	Cow 3	Cow 4
Period 1	T4 (192)	T1 (195)	T3 (292)	T2 (249)
Period 2	T1 (190)	T4 (203)	T2 (218)	T3 (210)
Period 3	T3 (214)	T2 (139)	T1 (245)	T4 (163)
Period 3	T2 (221)	T3 (152)	T4 (204)	T1 (134)

1. Write down the model for this analysis.

Model: $Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \theta_k + \varepsilon_{ijk}$,

response ~ treatment + block + error

2. Which type of design is this?

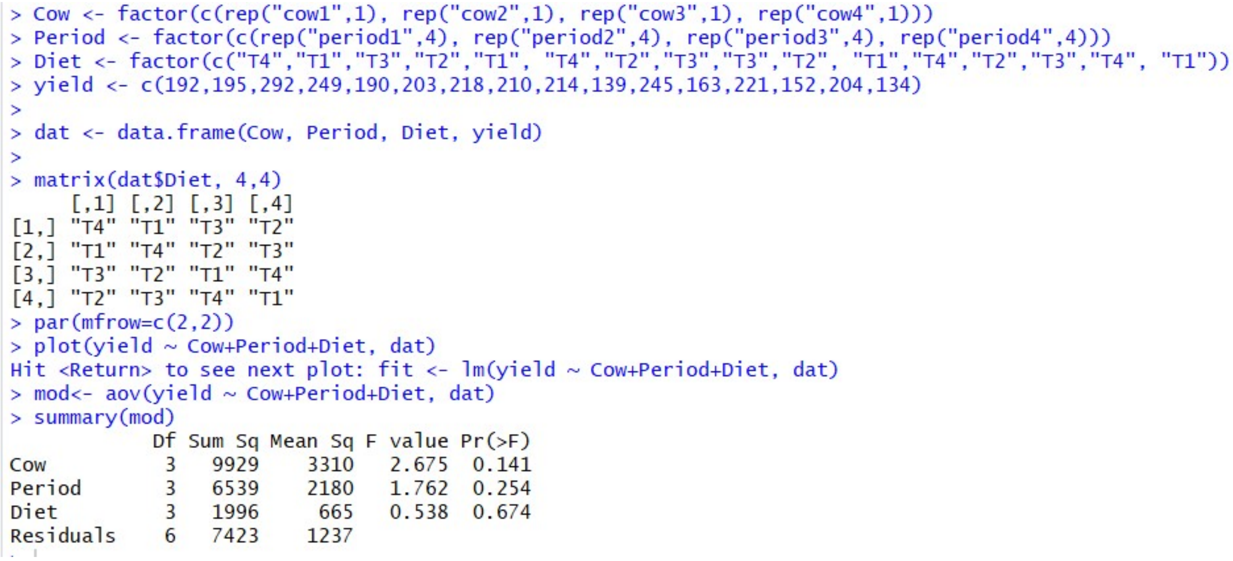
This is a completely randomized block design, I use Latin Square design for this experiment since there are two blocking factors (cow and period) which need to be controlled. The two blocking variables in a Latin square design are often generically labeled as row and column blocking variables. Cow is identified as the column variable and period as the row variable.

3. Use ANOVA to analyze the effects of four diets.

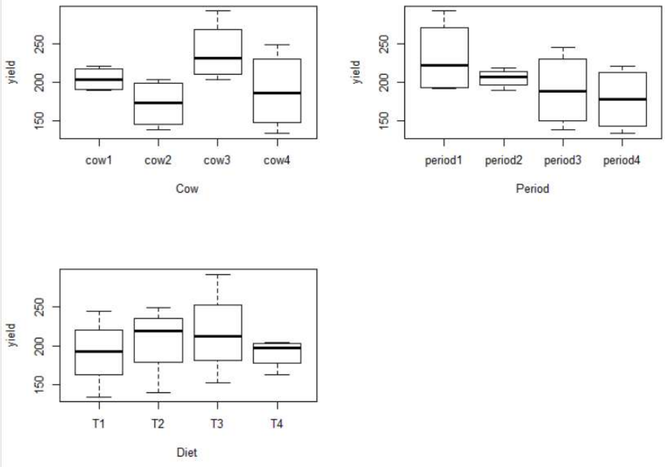
H_0: α_i=0

$H_{1:}$ not all α_i are zero

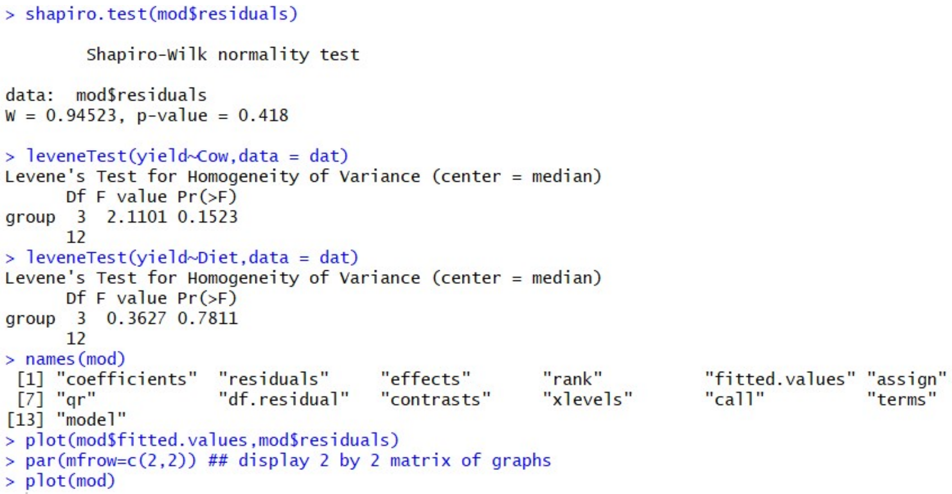
Since this is a completely block design, the factors are crossed, no interaction. I input data according to the randomization scheme table provided by the question.



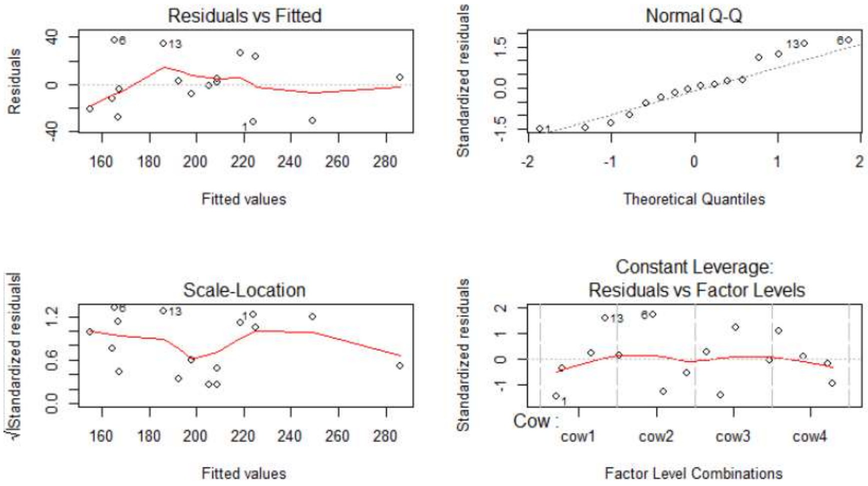
```
> Cow <- factor(c(rep("cow1",1), rep("cow2",1), rep("cow3",1), rep("cow4",1)))
> Period <- factor(c(rep("period1",4), rep("period2",4), rep("period3",4), rep("period4",4)))
> Diet <- factor(c("T4","T1","T3","T2","T1", "T4","T2","T3","T3","T2", "T1","T4","T2","T3","T4", "T1"))
> yield <- c(192,195,292,249,190,203,218,210,214,139,245,163,221,152,204,134)
>
> dat <- data.frame(Cow, Period, Diet, yield)
>
> matrix(dat$Diet, 4,4)
     [,1] [,2] [,3] [,4]
[1,] "T4" "T1" "T3" "T2"
[2,] "T1" "T4" "T2" "T3"
[3,] "T3" "T2" "T1" "T4"
[4,] "T2" "T3" "T4" "T1"
> par(mfrow=c(2,2))
> plot(yield ~ Cow+Period+Diet, dat)
Hit <Return> to see next plot: fit <- lm(yield ~ Cow+Period+Diet, dat)
> mod<- aov(yield ~ Cow+Period+Diet, dat)
> summary(mod)
            Df Sum Sq Mean Sq F value Pr(>F)
Cow          3   9929    3310   2.675  0.141
Period       3   6539    2180   1.762  0.254
Diet         3   1996     665   0.538  0.674
Residuals    6   7423    1237
```



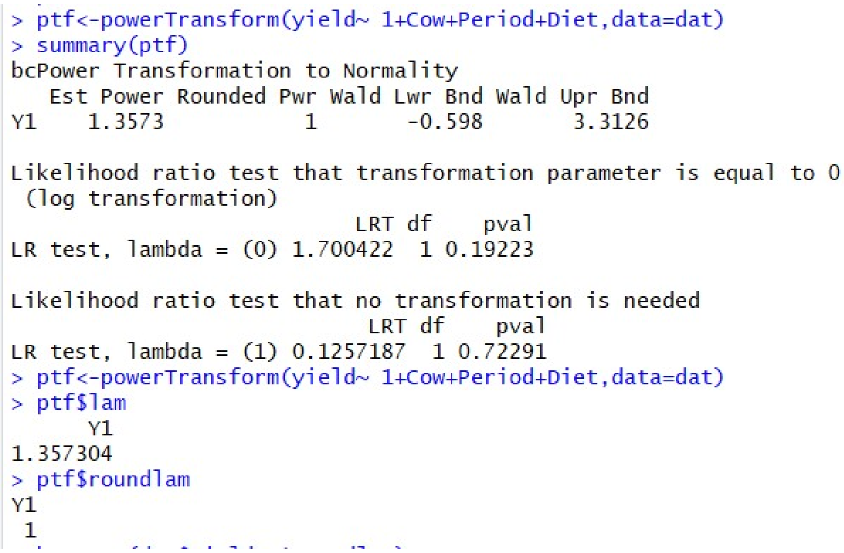
The p values of cow, period and diet are insignificant, which means there are no effects for yield of milk. From the boxplots, we also can observe the yields are not big different, but the periods plot indicate period1 has more effects for yield.



```
> shapiro.test(mod$residuals)

        Shapiro-Wilk normality test

data:  mod$residuals
W = 0.94523, p-value = 0.418

> leveneTest(yield~Cow,data = dat)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  3  2.1101 0.1523
      12
> leveneTest(yield~Diet,data = dat)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  3  0.3627 0.7811
      12
> names(mod)
 [1] "coefficients"  "residuals"     "effects"       "rank"          "fitted.values" "assign"
 [7] "qr"            "df.residual"   "contrasts"     "xlevels"       "call"          "terms"
[13] "model"
> plot(mod$fitted.values,mod$residuals)
> par(mfrow=c(2,2)) ## display 2 by 2 matrix of graphs
> plot(mod)
```

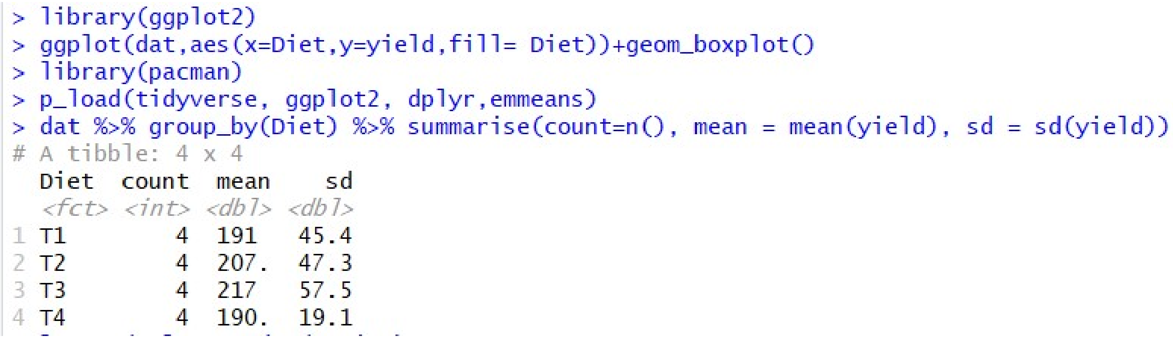


Shapiro test and Levene test indicate that data supports the assumptions. But the residual plots look like non-constant variance. Since this is blocking factors design, it is a little hard to detect the assumptions. I used the power transformations to test data and found there is no need to do transformations

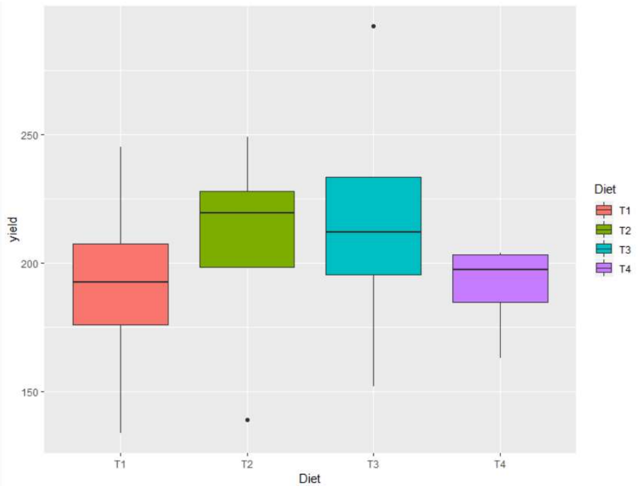


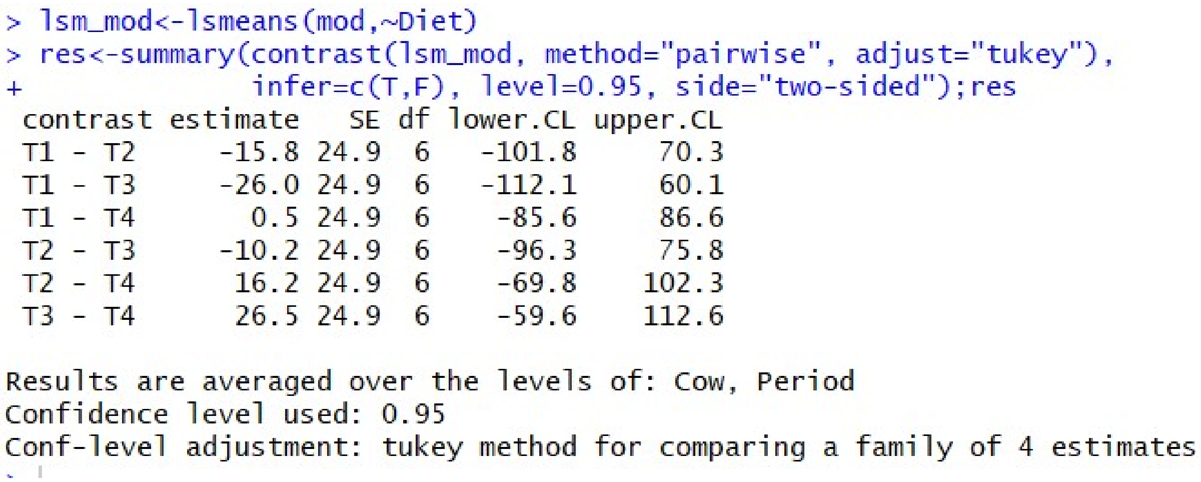
```
> ptf<-powerTransform(yield~ 1+Cow+Period+Diet,data=dat)
> summary(ptf)
bcPower Transformation to Normality
   Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Y1    1.3573           1       -0.598       3.3126

Likelihood ratio test that transformation parameter is equal to 0
 (log transformation)
                           LRT df     pval
LR test, lambda = (0) 1.700422  1 0.19223

Likelihood ratio test that no transformation is needed
                            LRT df     pval
LR test, lambda = (1) 0.1257187  1 0.72291
> ptf<-powerTransform(yield~ 1+Cow+Period+Diet,data=dat)
> ptf$lam
      Y1
1.357304
> ptf$roundlam
Y1
 1
```

4. Construct 95% Tukey's pairwise comparisons of the diets and report your findings.

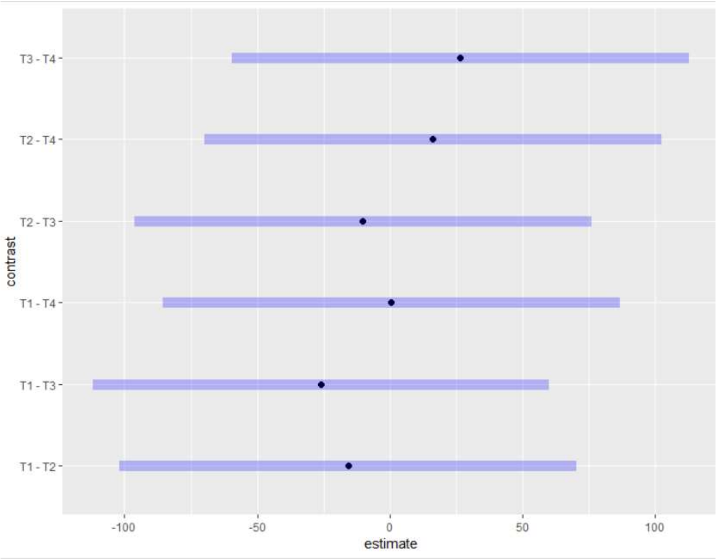


```
> library(ggplot2)
> ggplot(dat,aes(x=Diet,y=yield,fill= Diet))+geom_boxplot()
> library(pacman)
> p_load(tidyverse, ggplot2, dplyr,emmeans)
> dat %>% group_by(Diet) %>% summarise(count=n(), mean = mean(yield), sd = sd(yield))
# A tibble: 4 x 4
  Diet  count  mean    sd
  <fct> <int> <dbl> <dbl>
1 T1        4  191   45.4
2 T2        4  207.  47.3
3 T3        4  217   57.5
4 T4        4  190.  19.1
```





```
> lsm_mod<-lsmeans(mod,~Diet)
> res<-summary(contrast(lsm_mod, method="pairwise", adjust="tukey"),
+               infer=c(T,F), level=0.95, side="two-sided");res
 contrast estimate   SE df lower.CL upper.CL
 T1 - T2     -15.8 24.9  6   -101.8     70.3
 T1 - T3     -26.0 24.9  6   -112.1     60.1
 T1 - T4       0.5 24.9  6    -85.6     86.6
 T2 - T3     -10.2 24.9  6    -96.3     75.8
 T2 - T4      16.2 24.9  6    -69.8    102.3
 T3 - T4      26.5 24.9  6    -59.6    112.6

Results are averaged over the levels of: Cow, Period
Confidence level used: 0.95
Conf-level adjustment: tukey method for comparing a family of 4 estimates
```



Based on the results, the four diets have no significant effects for yield of milk. According Tukey's comparisons of the four diets, we can conclude that the four diets are similar.

Problem 4:

Write down 1 or 2 statistics/data science/machine learning-related concepts/methods/areas which may not be studied in your roadmap but you would like to learn before graduation, e.g., some practical useful stat skills. Your topics of interest may be discussed in a near future.

I would like to learn some skills about big data or data mining.

Thank you, Dr. Zou! I have learned a lot in your class. I appreciate you really care about your students. Thank you so much! Have a good summer!