Stat631Final

Brandon Keck

1.

BK This submission is entirely my own work. This means you are not allowed to discuss the exam questions with anyone, including roommates or housemates who are enrolled in this course.

BK As a member of the academic community, I am expected to act with integrity and avoid plagerism and other forms of cheating. I agree to uphold the standards of academic integrity described at CSUEB Academic Dishonesty Policy.

BK If I am found to have participated in any form of academic dishonesty, I will fail the exam and an academic dishonesty incident report will be filed.

```
library(ggplot2)
library(car)
library(emmeans)
library(dplyr)
library(lme4)
library(lmerTest)
library(EMSaov)
```

2.

(2a)

```
# read in the dataset
fitness <- read.table("fitness.txt", header = T)
str(fitness)
```

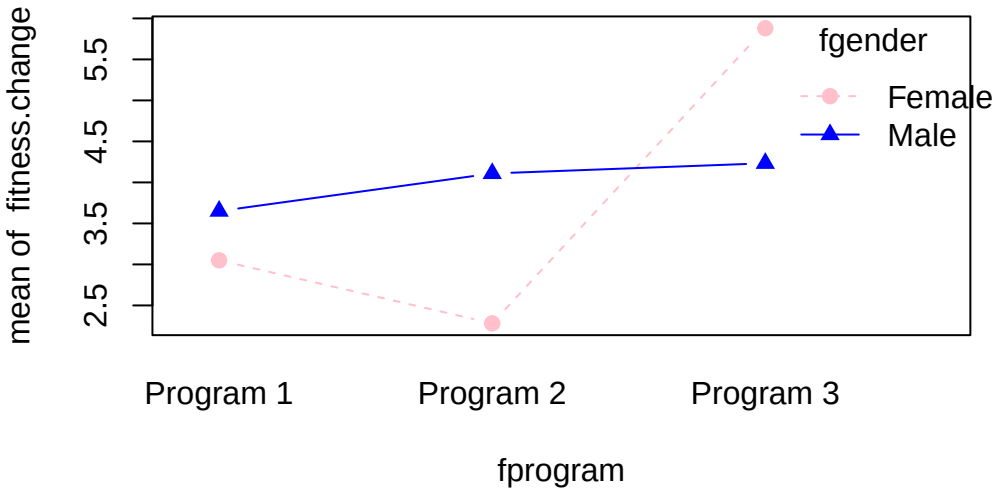
```
'data.frame':   78 obs. of  5 variables:
 $ person       : int  25 26 1 2 3 4 5 6 7 8 ...
 $ gender       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ program      : int  2 2 1 1 1 1 1 1 1 1 ...
 $ pre.fitness  : num  60 103 54.2 54 63.3 61.1 62.2 64 65 60.5 ...
 $ fitness6weeks: int  60 103 58 60 64 64 65 66 67 69 ...
```

```
# Change factor variables to factors
fitness <- within(fitness, {fitness.change <- fitness6weeks - pre.fitness
fgender <- factor(gender, labels = c("Female", "Male"))
fprogram <- factor(program, labels = c("Program 1", "Program 2", "Program 3"))})
# Check the structure
str(fitness)
```

```
'data.frame':   78 obs. of  8 variables:
 $ person        : int  25 26 1 2 3 4 5 6 7 8 ...
 $ gender        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ program       : int  2 2 1 1 1 1 1 1 1 1 ...
 $ pre.fitness   : num  60 103 54.2 54 63.3 61.1 62.2 64 65 60.5 ...
 $ fitness6weeks : int  60 103 58 60 64 64 65 66 67 69 ...
 $ fprogram      : Factor w/ 3 levels "Program 1","Program 2",..: 2 2 1 1 1 1 1 1 1 1 ...
 $ fgender       : Factor w/ 2 levels "Female","Male": 1 1 1 1 1 1 1 1 1 1 ...
 $ fitness.change: num  0 0 3.8 6 0.7 ...
```

(2b)

```
# Create the interaction plot
with(fitness, interaction.plot(x.factor = fprogram, trace.factor = fgender,
                               response = fitness.change, type = "b",
                               col = c("pink", "blue"), pch = c(19, 17)))
```




```
# Blue = Males
# Pink = Females
```

The interaction plot shows average fitness change by program and gender. For males (blue), the mean fitness change steadily increased across programs, but only slightly. For females (pink), there is a sharp increase in fitness change from Program 2 to Program 3. The lines intersect, suggesting a possible interaction between gender and program. Program 3 appears to be especially effective for females, indicating that the Program effect may differ by gender.

(2c)

Gender is the natural blocking factor since we could simply divide our subjects into gender classes. Genders have biological differences such as height, weight, etc., that can cause variation in the response variable. By using gender as a blocking factor we can group similar responses together and reduce the extra variability.

(2d)

```r
# Create the model
mod1 <- aov(fitness.change ~ fprogram + fgender, data = fitness)
summary(mod1)
```

```
            Df Sum Sq Mean Sq F value  Pr(>F)
fprogram     2   71.1   35.55   6.130 0.00344 **
fgender      1    1.0    1.04   0.178 0.67389
Residuals   74  429.1    5.80
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(2e)

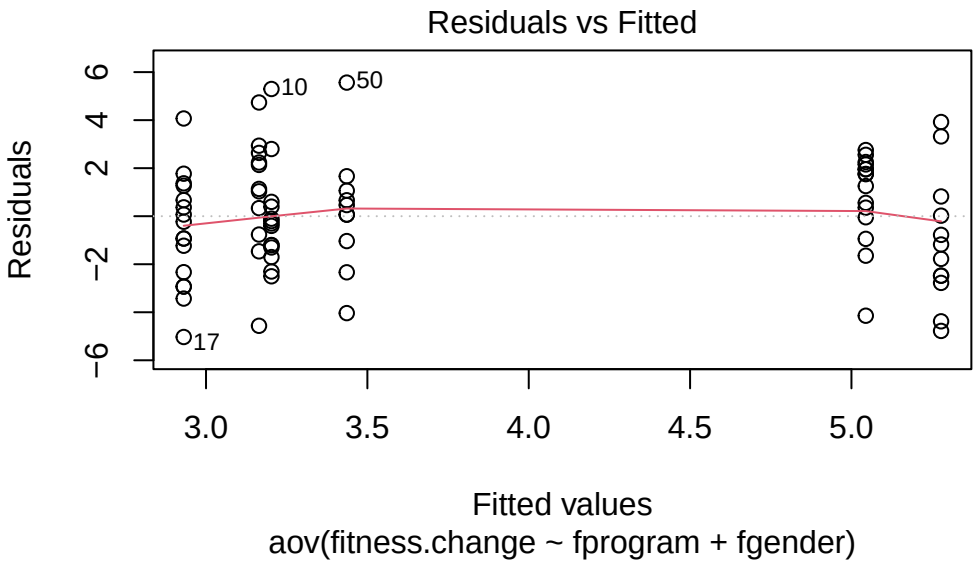
From the ANOVA table and using a significance level of $\alpha = 0.05$, the p-value for gender is 0.67389. Since this p-value is greater than our significance level of 0.05, we fail to reject the null hypothesis. This indicates that there is no statistically significant difference in fitness change between males and females.

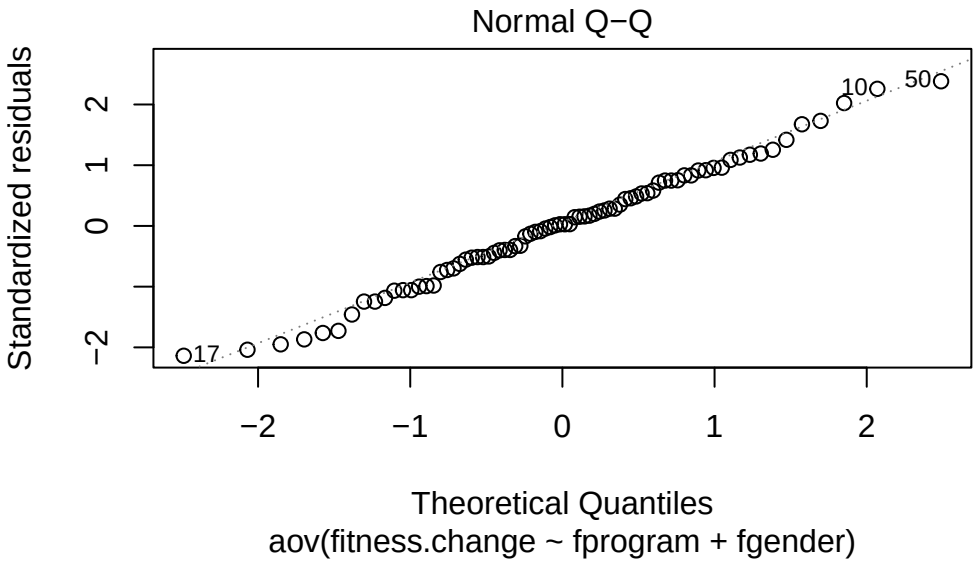
(2f)

From the ANOVA table and using a significance level of $\alpha = 0.05$ the p-value for program is 0.00344. Since this is less than our significance level of 0.05, we reject the null hypothesis and conclude that there is a statistically significant difference in mean fitness change among programs.

(2g)

```r
# Check ANOVA assumptions
plot(mod1, 1:2)
```






```
shapiro.test(resid(mod1))
```

```
    Shapiro-Wilk normality test

data:  resid(mod1)
W = 0.99068, p-value = 0.847
```

```r
leveneTest(fitness.change ~ fprogram * fgender, data = fitness)
```

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  5  0.4351 0.8227
      72
```

The Residuals vs Fitted plot does not show any discernible pattern or fan shape, so the assumptions of constant variance appears to be satisfied. The Levene's test for Homogeneity of variance returned a p-value of 0.8227, which is greater than the significance level $\alpha = 0.05$. The Normal Q-Q plot shows that the residuals are approximately normally distributed, with only minor deviations in the tails. The Shapiro-Wilk test produced a p-value of 0.847 providing no evidence against the normality assumption. Therefore, we can conclude that the assumptions for ANOVA are reasonably met.

(2h)

Based on the ANOVA table output there is strong evidence that the fitness programs have different effects on fitness level. With a p-value of $0.00344 < \alpha = 0.05$ this indicates that the program a person participated in has meaningful impact on their fitness.

3.

```r
# read in the dataset
drug <- read.table("drug.txt", header = T)
#Convert factor variables to a factor
drug <- within(drug, {Facility = as.factor(Facility);
Batch = as.factor(Batch);
Sample = as.factor(Sample)})
# Check the structure
str(drug)
```

```
'data.frame':   60 obs. of  4 variables:
 $ Facility: Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 1 ...
 $ Batch   : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 2 2 2 2 2 ...
 $ Sample  : Factor w/ 5 levels "1","2","3","4",..: 1 2 3 4 5 1 2 3 4 5 ...
 $ Potency : num  22.4 21.9 23.2 22.6 23 ...
```

(3a)

Three factors: Facility, Batch and Sample

(3b)

All three factors; Facility, Batch and Sample are random.

(3c)

```
# From Dr. Myung
knitr::include_graphics("(3c).png")
```

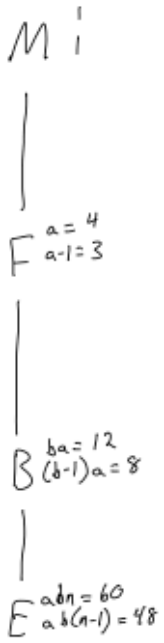
(3c) Hasse Diagram

3 factors

- Facility (4) $A = 4$
- Batch (3) $B = 3$
- Sample (5) $n = 5$

$N = 4 * 3 * 5 = 60$

$g = 4 * 3 = 12$



(3d)

Both Batch and Facility are nested factors in the model.

(3e)

$Y_{ijk} = \mu + F_i + B_{j(i)} + \epsilon_{k(ij)}$

- μ is the overall mean
- F_i is the effect on the i-th facility
- $B_{j(i)} \sim \mathrm{N}(0, \sigma^2)$ is the random effect on the j-th batch
- $\epsilon_{k(ij)} \sim \mathrm{N}(0,\sigma^2)$ is the residual error

(3f)

```r
#Create the model
mod2 <- lmer(Potency ~ 1 + (1|Facility) + (1|Facility:Batch), data = drug)
summary(mod2)
```

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
Formula: Potency ~ 1 + (1 | Facility) + (1 | Facility:Batch)
   Data: drug

REML criterion at convergence: 127.9

Scaled residuals:
     Min       1Q   Median       3Q      Max
-1.77808 -0.59968  0.02403  0.74865  1.49754

Random effects:
 Groups         Name        Variance Std.Dev.
 Facility:Batch (Intercept)  0.8195  0.9053
 Facility       (Intercept) 15.4697  3.9332
 Residual                    0.2245  0.4738
Number of obs: 60, groups:  Facility:Batch, 12; Facility, 4

Fixed effects:
            Estimate Std. Error     df t value Pr(>|t|)
```

```
(Intercept)   24.011      1.985  3.000    12.1  0.00122 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The largest source of variability in drug potency is Facility with a variance estimate of 15.4697. The smallest component of variance is residual variance, with 0.2245.

(3g)

```
ranova(mod2)
```

```
ANOVA-like table for random-effects: Single term deletions

Model:
Potency ~ (1 | Facility) + (1 | Facility:Batch)
                     npar  logLik    AIC    LRT Df Pr(>Chisq)
<none>                  4 -63.960 135.92
(1 | Facility)          3 -73.083 152.17 18.245  1  1.942e-05 ***
(1 | Facility:Batch)    3 -88.056 182.11 48.191  1  3.866e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Facility and Batch are both significant. There is significant variability in drug potency between the production facilities and between batches within the facilities.

(3h)

```
fit <- EMSanova(Potency ~ Facility + Batch, data = drug, type = c("R", "R"),
                nested = c(NA, "Facility"))
fit
```

```
               Df        SS          MS  Fvalue  Pvalue Sig
Facility        3 709.08922 236.3630728 54.6891 <0.0001 ***
Batch(Facility)  8  34.57556   4.3219450 19.2545 <0.0001 ***
Residuals       48  10.77428   0.2244642
                                          EMS
Facility        Error+5Batch(Facility)+15Facility
Batch(Facility)           Error+5Batch(Facility)
Residuals                                 Error
```

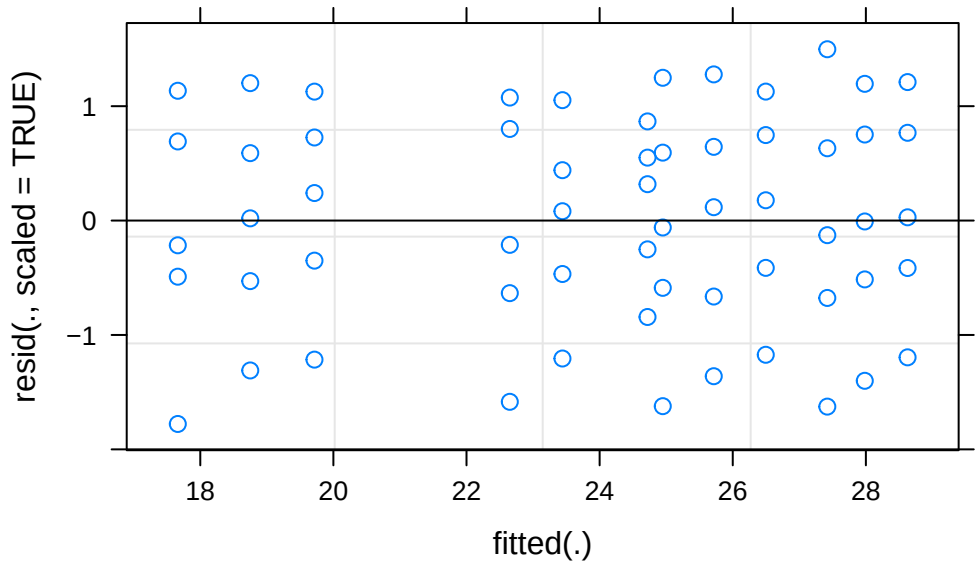
Facility has highly significant effects with a p-value of 0.0001. Batches also had a significant effects with a p-value of 0.001. So both are still significant.

(3i)

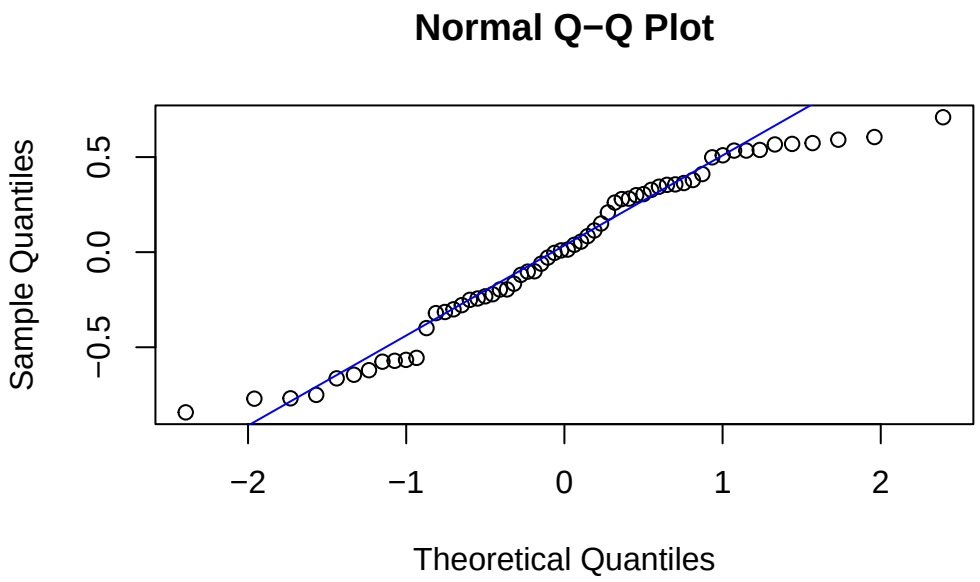
Both part h and part f are consistent with each other. In both models we see that Facility is significant. Also Batch was significant in part f. Both parts h and f lead to the same conclusions. Facility has the largest variance in both models and the residuals has the smallest variance in both parts h and f.

(3j)

```
# Checking Equal Variance
plot(mod2,resid(.,scaled =TRUE)~fitted(.),abline =c(-2,0,2))
```




```
# Check Normality
qqnorm(resid(mod2))
qqline(resid(mod2), col = "blue")
```




```
shapiro.test(resid(mod2))
```

```
    Shapiro-Wilk normality test

data:  resid(mod2)
W = 0.95288, p-value = 0.02133
```

The Shapiro-Wilk test gave a p-value of 0.02133, which is greater than the significance level of $\alpha = 0.01$. Therefore, we fail to reject the null hypothesis of normality, and the assumption of normality is satisfied. Therefore, the assumptions of ANOVA are satisfied.