Stat 571B Appendix

Group 2

2025-04-29

Section 1: Importing Data

```
#-----------------------------------------
# Load data & Convert factors as categorical
#-----------------------------------------

# Data Import (may need to change based on where you saved the csv)
study <- read.csv("study.csv")

# Setting factors
study$Person <- as.factor(study$Person)
study$Day <- as.factor(study$Day)
study$Time <- as.factor(study$Time)
study$Replicate <- as.factor(study$Replicate)
study$Location <- as.factor(study$Location)

head(study)
```

```
##   X Person    Day Time Replicate      Location Score Rating   dB  Rating_n
## 1 1  Paige Monday   AM         1  Main Library  5.29      4 49.1 0.4444444
## 2 2  Paige Monday   AM         1   Common Room  3.66      2 40.7 0.2222222
## 3 3  Paige Monday   AM         1 Student Union  7.27      6 62.9 0.6666667
## 4 4  Paige Monday   PM         1 Student Union  8.04      7 66.4 0.7777778
## 5 5  Paige Monday   PM         1   Common Room  4.49      3 45.1 0.3333333
## 6 6  Paige Monday   PM         1  Main Library  6.18      5 54.3 0.5555556
##        dB_n
## 1 0.6145181
## 2 0.5093867
## 3 0.7872340
## 4 0.8310388
## 5 0.5644556
## 6 0.6795995
```

Section 2: Fit ANOVA model

```
#-----------------------------------------------------
# Fitting into ANOVA full factorial model
#-----------------------------------------------------
```

```r
library(lmerTest)
```

```
## Warning: package 'lmerTest' was built under R version 4.4.3
```

```
## Loading required package: lme4
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'lmerTest'
```

```
## The following object is masked from 'package:lme4':
##
##     lmer
```

```
## The following object is masked from 'package:stats':
##
##     step
```

```r
# Score as response
score_model <- lmer(Score ~ Day * Time * Location + (1 | Person), data = study)
anova(score_model)
```

```
## Type III Analysis of Variance Table with Satterthwaite's method
##                   Sum Sq Mean Sq NumDF DenDF  F value    Pr(>F)
## Day                 1.97   0.983     2   123   1.7601   0.17633
## Time               13.20  13.201     1   123  23.6353 3.479e-06 ***
## Location          384.19 192.096     2   123 343.9298 < 2.2e-16 ***
## Day:Time            0.37   0.184     2   123   0.3302   0.71940
## Day:Location        1.73   0.433     4   123   0.7756   0.54308
## Time:Location       3.93   1.966     2   123   3.5207   0.03259 *
## Day:Time:Location   0.94   0.235     4   123   0.4210   0.79327
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The three-way interaction Day:Time:Location has a P-value $0.79327 \geq 0.05$, therefore, this interaction has no significant effect on the response with 0.05 significance level. We can drop the three-way interaction term and fit a second-order model.

```r
#-------------------------------------------------------
# Fitting into a second order ANOVA model
#-------------------------------------------------------

library(lmerTest)

# Score as response
score_model_2 <- lmer(Score ~ (Day + Time + Location)^2 + (1 | Person), data = study)
anova(score_model_2)
```

```
## Type III Analysis of Variance Table with Satterthwaite's method
##               Sum Sq Mean Sq NumDF DenDF  F value    Pr(>F)
## Day             1.97   0.983     2   127   1.7928   0.17068
## Time           13.20  13.201     1   127  24.0744 2.784e-06 ***
## Location      384.19 192.096     2   127 350.3187 < 2.2e-16 ***
## Day:Time        0.37   0.184     2   127   0.3364   0.71501
## Day:Location    1.73   0.433     4   127   0.7900   0.53371
## Time:Location   3.93   1.966     2   127   3.5861   0.03055 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the output above, neither the factor 'Day' or its interaction terms are significant and therefore we can drop the factor 'Day'. Now we'll fit a second order but reduced model.

```
#-----------------------------------------------------
# Fitting into a second order reduced ANOVA model
#-----------------------------------------------------

score_model_red <- lmer(Score ~ Time*Location + (1 | Person), data = study)
anova(score_model_red)
```

```
## Type III Analysis of Variance Table with Satterthwaite's method
##               Sum Sq Mean Sq NumDF DenDF  F value    Pr(>F)
## Time           13.20  13.201     1   135  24.1785 2.508e-06 ***
## Location      384.19 192.096     2   135 351.8344 < 2.2e-16 ***
## Time:Location   3.93   1.966     2   135   3.6016   0.02993 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From this reduced model all factors, i.e., Time, Location and their interactions are statistically significant in terms of having effects on the response.

Section 3: Interaction Plot and Formal Test with Tukey's test for additivity

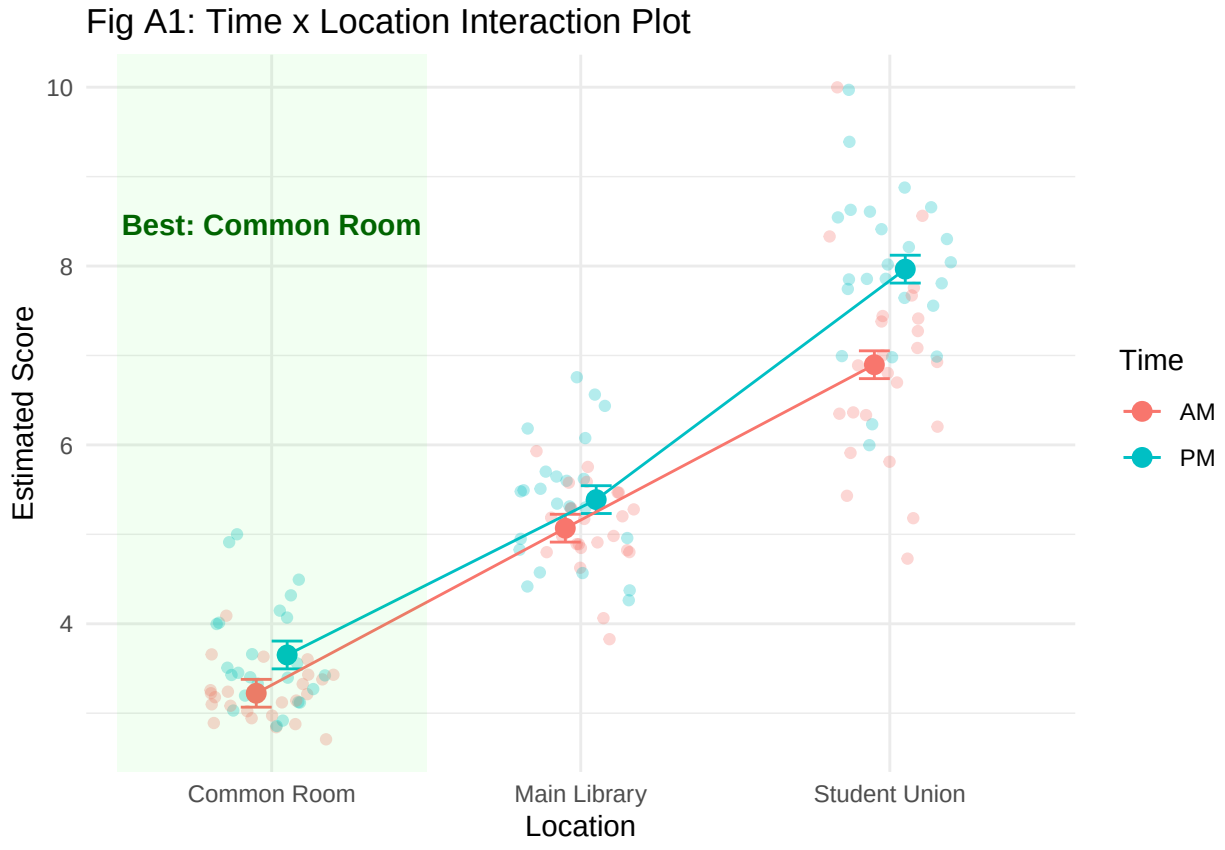
```
#-----------------------------------------------------
# Appendix Figure 1 (Fig. A1)
#-----------------------------------------------------

library(ggplot2)
library(emmeans)
```

```
## Warning: package 'emmeans' was built under R version 4.4.3
```

```
## Welcome to emmeans.
## Caution: You lose important information if you filter this package's results.
## See '? untidy'
```

```
# Estimated marginal means
emm <- emmeans(score_model_red, ~ Time * Location)
emm_df <- as.data.frame(emm)
```

```r
# Plot
ggplot(emm_df, aes(x = Location, y = emmean, color = Time, group = Time)) +

# Raw data jittered (from full dataset)
geom_jitter(data = study, aes(x = Location, y = Score, color = Time),
            width = 0.2, alpha = 0.3, inherit.aes = FALSE) +

# Estimated means with error bars
geom_point(size = 3, position = position_dodge(0.2)) +
geom_errorbar(aes(ymin = emmean - SE, ymax = emmean + SE),
            width = 0.2, position = position_dodge(0.2)) +
geom_line(position = position_dodge(0.2)) +

# Annotate the best location
annotate("rect", xmin = 0.5, xmax = 1.5, ymin = -Inf, ymax = Inf,
          alpha = 0.05, fill = "green") +
annotate("text", x = 1, y = max(emm_df$emmean) + 0.5,
          label = "Best: Common Room", color = "darkgreen", fontface = "bold") +

labs(title = "Fig A1: Time x Location Interaction Plot",
       y = "Estimated Score", x = "Location") +
theme_minimal()
```



We investigated the interaction between the two factors through an interaction plot which can be shown in the Appendix Figure 1. In Fig A1, the colored lines (AM = red, PM = blue) represent the estimated

mean score for each location across times of day. Since these two lines are not perfectly parallel, there is an interaction present between factors 'Time' and 'Location' on the score. For example, the effect of particular location (i.e., Student Union) on the score varies from time (AM) to time (PM). Interaction is significantly pronounced through Student Union, i.e., slightly higher score in PM than AM.

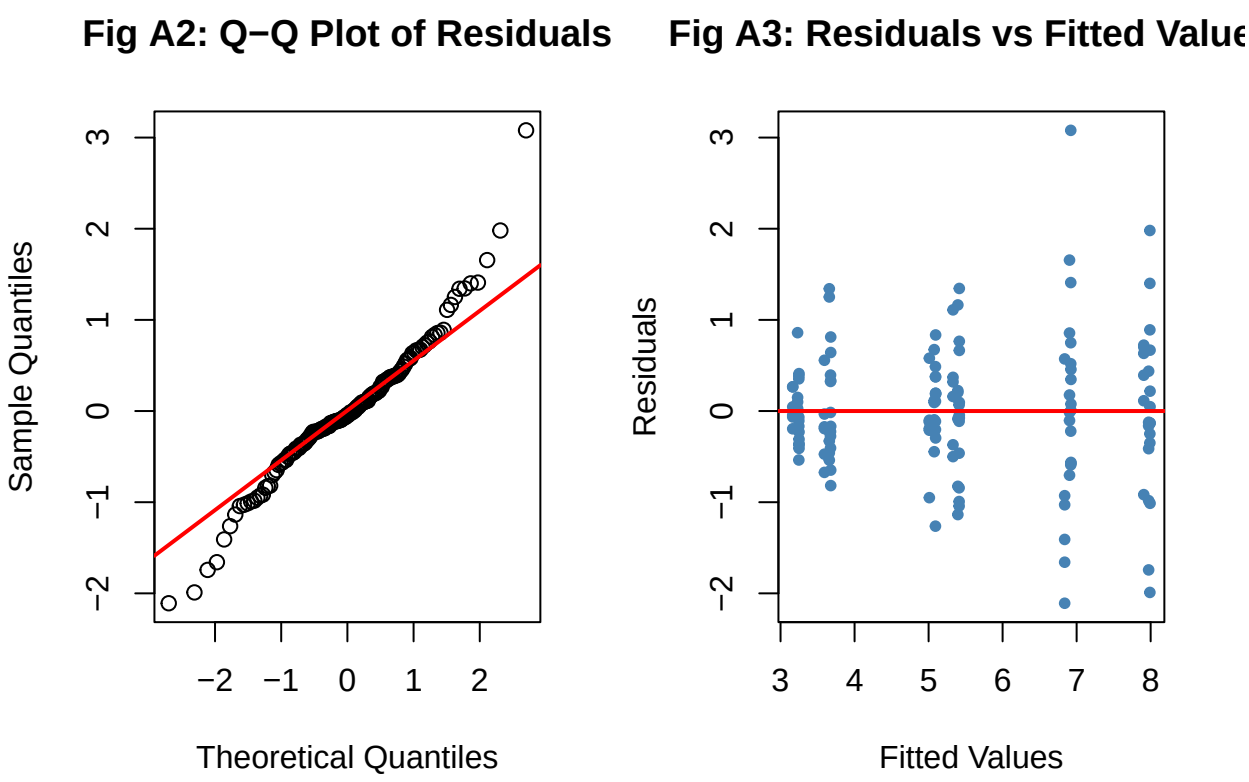
```
#-----------------------------------------------------
# Tukey's Test for NonAdditivity
#-----------------------------------------------------

score_add <- aov(Score ~ Time + Location, data = study)
study$q3 <- fitted(score_add)^2
score_tukey_add <- aov(Score ~ Time + Location + q3, data = study)
anova(score_tukey_add)
```

```
## Analysis of Variance Table
##
## Response: Score
##            Df Sum Sq Mean Sq  F value    Pr(>F)
## Time        1  13.20  13.201  23.7634 2.933e-06 ***
## Location    2 384.19 192.096 345.7939 < 2.2e-16 ***
## q3          1   2.68   2.683   4.8292   0.02964 *
## Residuals 139  77.22   0.556
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The above test confirms that the interaction effect is significant, as also suggested by the (diagnostic) interaction plot.

Step 4: Model Assumption Check

```
#-----------------------------------------------
# Normal plot (Appendix Figure 2) and Residuals vs. Fitted (Appendix Figure 3)
#-----------------------------------------------

resid_values <- resid(score_model_red)
fitted_values <- fitted(score_model_red)

par(mfrow = c(1, 2))

# Q-Q Plot for Normality
qqnorm(resid_values, main = "Fig A2: Q-Q Plot of Residuals")
qqline(resid_values, col = "red", lwd = 2)

# Residuals vs. Fitted
plot(fitted_values, resid_values,
     main = "Fig A3: Residuals vs Fitted Values",
     xlab = "Fitted Values", ylab = "Residuals",
     pch = 20, col = "steelblue")
abline(h = 0, col = "red", lwd = 2)
```



Q-Q Plot of Residuals:

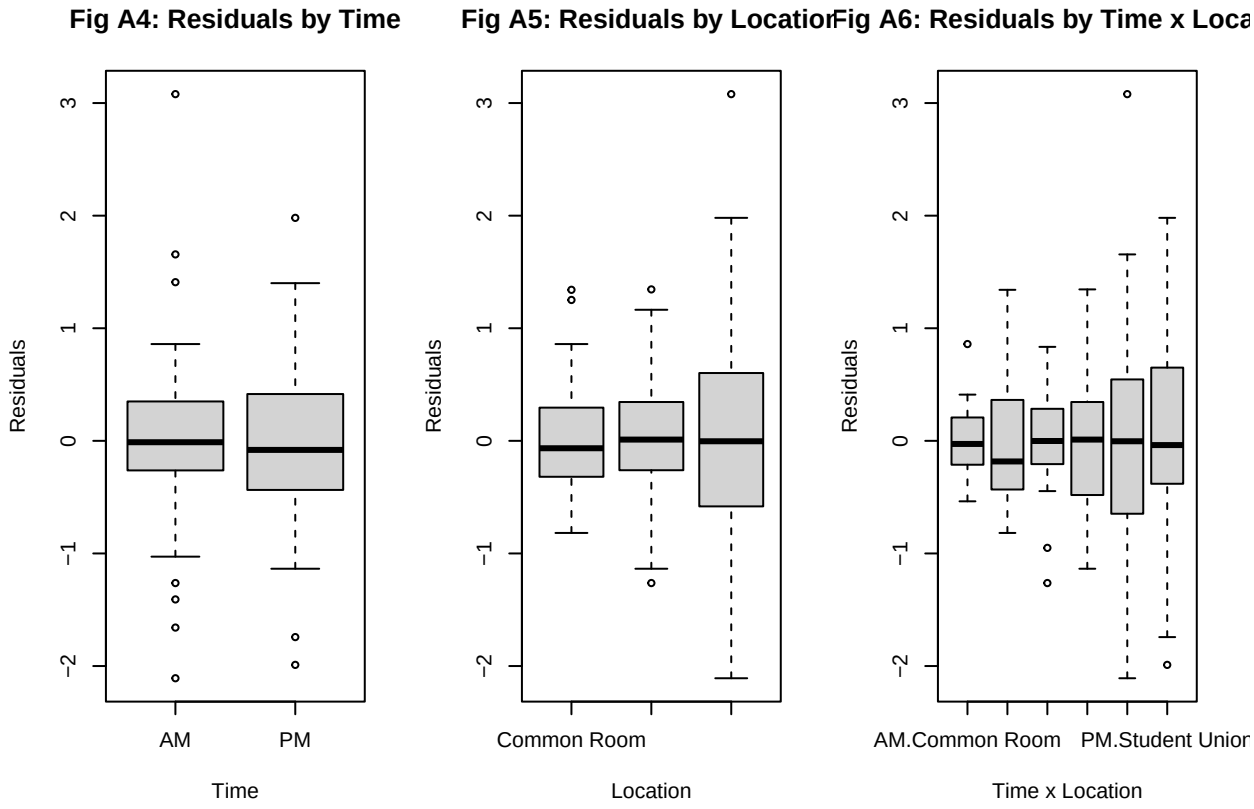
From Fig A2. Q-Q Plot of Residuals we can see that most of the residual points are wrapped well across the reference line. Some observations are deviated which is common for a large dataset. Only one observation on the top appears to be an outlier but not at a significant margin. Thus, we can conclude that the normality assumption is reasonably satisfied for the residuals.

Residuals vs Fitted Plot:

From Fig A3. Residuals vs Fitted Plot we can observe that the residuals appear randomly scattered around the horizontal line at zero. Also, there is no clear pattern or shape forming and no monotonous increasing or decreasing trend is observed. Therefore, the homoscedasticity (constant variance) assumption seems to be satisfied. We can also see that the higher predicted values near 7 and 8, there are some points showing high variance than the other fitted scores.

```
#-----------------------------------------------------
# Residual boxplots (Appendix Figures 4, 5, 6)
#-----------------------------------------------------

par(mfrow = c(1, 3), mar = c(5, 4, 4, 2) + 0.1)
resid_values <- resid(score_model_red)

boxplot(resid_values ~ study$Time,
        main = "Fig A4: Residuals by Time",
        ylab = "Residuals",
        xlab = "Time")

boxplot(resid_values ~ study$Location,
```

```
        main = "Fig A5: Residuals by Location",
        ylab = "Residuals",
        xlab = "Location")

interaction_group <- interaction(study$Time, study$Location)

boxplot(resid_values ~ interaction_group,
        main = "Fig A6: Residuals by Time x Location",
        ylab = "Residuals",
        xlab = "Time x Location")
```



From Appendix Figure 4 we can observe that both AM and PM data have similar spread, so the variance is constant across 'Time'. the second plot in Fig A5 shows that the Common Room and Main Library have similar variance. However, the Student Union has a larger spread indicating high variance. In Fig A6, both AM.Student Union and PM.Student Union show larger spread and therefore may have more outliers which can be a potential source for heteroscedasticity.

Section 5: Tukey-adjusted confidence intervals to make recommendations

```
#-------------------------------------------------
# Tukey-adjusted confidence intervals
#-------------------------------------------------

library(emmeans)
```

```r
emm <- emmeans(score_model_red, ~ Location)
```

```
## NOTE: Results may be misleading due to involvement in interactions
```

```r
summary(pairs(emm, adjust = "tukey"))
```

```
##  contrast                     estimate    SE  df t.ratio p.value
##  Common Room - Main Library      -1.79 0.151 135 -11.877  <.0001
##  Common Room - Student Union     -3.99 0.151 135 -26.480  <.0001
##  Main Library - Student Union    -2.20 0.151 135 -14.603  <.0001
##
## Results are averaged over the levels of: Time
## Degrees-of-freedom method: kenward-roger
## P value adjustment: tukey method for comparing a family of 3 estimates
```

```r
emm <- emmeans(score_model_red, ~ Time)
```

```
## NOTE: Results may be misleading due to involvement in interactions
```

```r
summary(pairs(emm, adjust = "tukey"))
```

```
##  contrast estimate    SE  df t.ratio p.value
##  AM - PM    -0.606 0.123 135  -4.917  <.0001
##
## Results are averaged over the levels of: Location
## Degrees-of-freedom method: kenward-roger
```

```r
emm <- emmeans(score_model_red, ~ Time:Location)
summary(pairs(emm, adjust = "tukey"))
```

```
##  contrast                           estimate    SE  df t.ratio p.value
##  AM Common Room - PM Common Room      -0.428 0.213 135  -2.008  0.3432
##  AM Common Room - AM Main Library     -1.846 0.213 135  -8.654  <.0001
##  AM Common Room - PM Main Library     -2.165 0.213 135 -10.152  <.0001
##  AM Common Room - AM Student Union    -3.674 0.213 135 -17.223  <.0001
##  AM Common Room - PM Student Union    -4.742 0.213 135 -22.233  <.0001
##  PM Common Room - AM Main Library     -1.417 0.213 135  -6.645  <.0001
##  PM Common Room - PM Main Library     -1.737 0.213 135  -8.144  <.0001
##  PM Common Room - AM Student Union    -3.245 0.213 135 -15.215  <.0001
##  PM Common Room - PM Student Union    -4.314 0.213 135 -20.225  <.0001
##  AM Main Library - PM Main Library    -0.320 0.213 135  -1.498  0.6660
##  AM Main Library - AM Student Union   -1.828 0.213 135  -8.570  <.0001
##  AM Main Library - PM Student Union   -2.897 0.213 135 -13.580  <.0001
##  PM Main Library - AM Student Union   -1.508 0.213 135  -7.071  <.0001
##  PM Main Library - PM Student Union   -2.577 0.213 135 -12.082  <.0001
##  AM Student Union - PM Student Union  -1.069 0.213 135  -5.010  <.0001
##
## Degrees-of-freedom method: kenward-roger
## P value adjustment: tukey method for comparing a family of 6 estimates
```

- Common Room is significantly quieter (lower distraction) than the others.
- Student Union is the loudest
- AM is a quieter/better time to go for studying in the campus.
- AM Common Room is the quietest interaction combination (all point estimates are negative). Not significantly quieter than PM Common Room.