STAA 553: Final Exam

Spring 2025

Honor Pledge: I have worked independently on this exam. I have read the exam instructions. I have not given, received, or used any unauthorized assistance on this exam.
Matthew Stoebe

Instructions:

- This exam is due by Saturday 3/15 at midnight.
- **Students are required to work independently on the exam.** Do NOT discuss the exam with anyone else (including other students).
- You may use the textbook, class notes, examples, HW solutions posted in the current Canvas course. You may use any other publicly available (print or online) statistics references or resources that you find helpful. Use of homework "helper" websites (ex: Chegg, NoteHall, etc) is NOT allowed. Use of chatbots (ex: ChatGPT) is NOT allowed.
- Knit frequently to avoid last minute problems. You may add or delete code chunks as needed. **It is the student's responsibility to check the knitted document (for correctness and completeness) before submitting.**
- For any questions that require calculations, you should provide R code for full credit.
- For some questions, there may be more than one possible answer, analysis or graph that could be used for full credit. **Choose one approach**, making a reasonable choice and justifying if needed.
- Given this is the final exam, you should present your best work. I will deduct points for things like printing full data to knitted document, unreadable tables, unclear, excess or unnecessary output, etc.
- Use $\alpha = 0.05$ and/or 95% confidence where needed.
- All questions are worth 4 points except where noted. Maximum score is 100.
- **I believe all students can do well on this exam. Please don't cheat!!!**

Blueberries (Q1 - Q4)

A study is being **planned** to compare 3 fertilizer treatments (A, B, C) on blueberry bushes. The response variable is yield (in lbs) **per bush** at the end of the growing season. The study will be run at a single farm, using existing mature blueberry bushes. At this farm, there are a large number of fields of blueberry bushes representing different varieties and ages. A single field has multiple bushes of a single age and single variety. Initial soil nitrogen varies across locations (even within a field). It is known that yield will be associated with variety, age and initial soil nitrogen.

As a statistician, you are asked to make recommendations to design the study.

Q1 (2 pts)

Discuss how **replication** would be incorporated into the design. In other words, what is a replicate?

Response Replication is incorporated by measuring yield on multiple units for each treatment group. Multiple replicates helps us have higher confidence in the finidngs of our study, and allows us to get a better estimation of variability

---

Q2 (2 pts)

Discuss how **randomization** would be incorporated into the design.

Response Randomization will be handled in how the treatemnts are assigned to the bushes. In this case we would want each randomization unit (probably going to be bushes) to have an equal probability of being assigned each treatment.

Q3A

Blocking is one approach that can be used to **reduce noise**. Provide a brief but specific description of how **blocking** could be incorporated into this research scenario. How will you define/select blocks? How will treatments be applied?

Response In this case, I may consider blockign by field as growing conditions in each field are likely to be similar. We can randomly assign treatments to each bush in the block. this should help us reduce variability caused by cross -field differences.

Q3B

Propose one other approach that can be used to **reduce noise**. Since you already discussed blocking in the previous question, discuss something **different** here. (This can be an alternative to blocking or it may be used with blocking.) The proposed method does not have to be complicated. Provide a reasonable amount of detail specific to this research scenario and providing guidance about how to implement the method (1-3 sentences).

Response
We can control for Covariates in our model instead of just relying on blocking. For example, the question talks about soil nitrogen levels. This could be included as a predictor in our model and will account for some variability cause by it. This does not have to be in exclusion of blockign but could be done together with it.

Q4

Regardless of your answers to previous question, suppose the investigators plan to use a completely randomized design. They conjecture that $\mu_1 = 32$, $\mu_2 = 34$, $\mu_3 = 40$ and $\sigma = 8$. Find the sample size required (number of blueberry bushes per treatment) to achieve 80% power. Remember to state your answer as an integer value.

---

n = 19

```
##
##      Balanced one-way analysis of variance power calculation
##
##               k = 3
##               n = 18.82254
##               f = 0.4249183
##       sig.level = 0.05
##           power = 0.8
##
## NOTE: n is number in each group
```

Sheep (Q5 - Q8)

An investigator is interested in comparing 6 treatments (Trt 1-6) for bone injuries in sheep. To evaluate the treatments, he creates (and then treats) small bone defects in the femurs of sheep. The response variable (Y) is defect size at one month post treatment. They will use a total of 18 sheep (blocks) with 4 defects per sheep (block). They are considering a **balanced incomplete block design (BIBD)**.

Q5

Identify the values for t (# treatments), r (# replicates per treatment), b (# blocks) and k (# "units" per block).

---

t = 6 treatments

b = 18 blocks

k = 4 defects

4*18 = 72 total treatment blocks. divided by 6 treatments = 12 replicates for each treatment r = 12:

Q6 (6 pts)

State and check the 3 conditions for a BIBD. Using your values from the description and responses to the previous question, is a BIBD possible?
Note: Even if there are mistakes in your responses to Q5, I am very open to giving partial credit as long as you show your work here.

Condition1: each treatment should appear 12 times in total

Condition2: each block has 4 distinct trreatments

Condition3: each pair of distinct treatments occur together the right number of blocks: 7.2

Is a BIBD possible?
7.2 is not an integer so it is not entirely possible.

```
## [1] 7.2
```

Q7

Regardless of your answer to the previous question, suppose the study will be run using a blocked design, even if we cannot achieve a balanced design.

Q7A (4 pts)

If the design is imbalanced, how does this affect the **analysis**? In other words, are changes required to the model or to generate the ANOVA table depending on whether we have an IBD (incomplete block design) versus a BIBD?

Response Yes we would have to adjust our analysis probalby using some mixed-modeling or a glm that can handle unbalanced data. The form should be similar, but not as straight forward as if the design were balanced.

Q7B (4 pts)

If the design is imbalanced, how does this affect the results? In other words, what benefit of the BIBD is lost if we have an IBD? Your discussion should be brief but still be specific.

Response The main difference if BIBD is lost is that the variances between treatments are not equal. This means we have varying confidence in our analysis for each parameter. *****

Q8

The investigators plan to use Tukey's method to run pairwise comparisons between the 6 treatments.

Q8A (4 pts)

Discuss the primary **benefit** of using Tukey's method (as compared to unadjusted pairwise comparisons). Your discussion should be brief but still be specific.

Response The tukey method controlls overall error rate instead of individual measurement error rate. This is important because as you do more experiments, it is likely that the .05 level of confidencw ends up failing eventualy.

Q8B (4 pts)

Discuss the primary **limitation** of using Tukey's method (as compared to unadjusted pairwise comparisons).

Response Tukey is super conservative and as such makes it harder ti find small incremental improvements without larger sample sizes (effectively reducing power)

Drug Study (Q9 - Q11)

A study was done to compare three drug Formulations (A = 5mg tablet, B = 100mg tablet, C = sustained release capsule). A total of n = 12 healthy adults volunteered to participate in the study. The study includes three periods, during which each subject received one of the formulations. There was a week long washout between each period. Subjects received the formulations in one of three sequences: ABC, BCA, CAB with 4 subjects randomly assigned to each sequence.

The data includes 36 rows and these variables:

- Subject: 1-12

- Formulation: A,B,C
- Period: 1, 2, 3
- Sequence: ABC, BCA, CAB
- PostBP: response, blood pressure after treatment

Notes:

- No data is provided! You have to think about the analysis based on the description above.
- Not all variables may be needed for analysis.

Q9 (2 pts)

Identify the ("blocking") design. Just provide the name of the design, no need to justify.

Response
This seems like a fun time to use a latin-square design

Q10

Considering your answer to the previous question, provide R code to fit an appropriate model. Note: I am primarily looking for a single line of code to set up the model. If you feel additional lines of code are needed, that is fine.

Response

lm(PostBP ~ Subject + Period + Formulation, data=drugData)

Q11

It is well known that blood pressure is variable even within a single person. Just for this question, suppose we also recorded the PreBP (blood pressure before treatment) for each subject and period.

Q11A (2 pts)

Suggest one way to incorporate this information into the analysis.

Response This could be included as a covariate and be used to adjust for each individuals baseline

---

Q11B (2 pts)

Modify your code from Q10 to incorporate your suggestion.

Response

lm(PostBP ~ Subject + Period + Formulation + PreBP, data=drugData)

Pine Beetles (Q12 - Q16)

Treatments for pine beetle infestation are suspected of affecting the "hardness" of the wood. A researcher is interested in comparing the wood hardness (Y) for three treatments (Trt: A, B, C). The researcher begins the experiment by identifying 5 infested trees (Tree: 1-5). For each tree, six branches are randomly assigned to treatments with each Trt represented exactly twice for each tree. At the end of the study, the hardness is measured for each branch. Hence, we have 5 trees x 3 treatments x 2 branches = 30 observations. The ANOVA table (from Anova with type = 3 and appropriate contrasts) is given here.

Source	df	SS	F value	p-value
Tree	1	70	0.7303	0.4006
Trt	2	38	0.1984	0.8213
Residuals	26	2493		

Q12

Briefly identify the blocking and treatment design for this study. Include identifying details about the blocks (if applicable) and treatments.

Blocking Design: We can use each tree as a block and then use multiple branches as the replicates for each tree

Treatment Design: Randomly assign treatment to each of two branches on each tree

---

Q13 (2 pts)

The Type 3 ANOVA table is shown above. Would the Type 1 ANOVA SS and F-tests have been the same or different? Just answer same or different. No need to justify.

Response different

Q14 (2 pts)

The analyst made a mistake when setting up the model. Identify the mistake. Note: This question is NOT looking for a discussion of whether Tree should be fixed or random.

Response df equals 1 for tree but we know there are 5 trees so it should equal 4.

Q15

After addressing the mistake from Q14, a colleague expresses concern that the assumption of independence is "not satisfied because you have repeat observations on trees". Briefly discuss whether you agree with the colleague's concern. If so, propose a modelling approach that can be used to address this concern.

Response
this is true. the views are non independant. this is addressed however with the blocking effect of tree. If there are further concerns about correlated errors we could consdier some sort of mixed modeling

---

Q16

After addressing the mistake from Q14, could an interaction (Tree:Trt) have been included in the model? Briefly discuss.

Response Sure, but we would only want to do this if we really belive the treatment beahves differently for each tree. We should not just include interactions willy nilly to see what they do.

Feed Trial (Q17 - Q23)

A study was done to investigate a new vitamin supplement for cattle. A successful supplement would increase **ADG** (average daily gain) in cattle. Two levels of **Suppl** (no or yes) and four **Feed** formulations (A, B, C, D) and were considered. The primary research question is about the effect of Suppl. But the researchers are concerned that the effect of supplement could vary across feed types. A total of 24 **Pen**s of cattle were randomly assigned to treatment combinations such that there are exactly 3 pens per treatment combination. ADG was recorded (for each pen) at the end of the study. The data is available from Canvas as FeedTrial.csv. Use alpha = 0.05.

Q17

Briefly identify the blocking and treatment design for this study. Include identifying details about the blocks (if applicable) and treatments.

Blocking Design: there is no explicit blocking as treatments are assigned at the penn level.

Treatment Design: treatments are randomly assigned at the penn level. with 8 treatments and 24 pens each treatment will get 3 replciates.

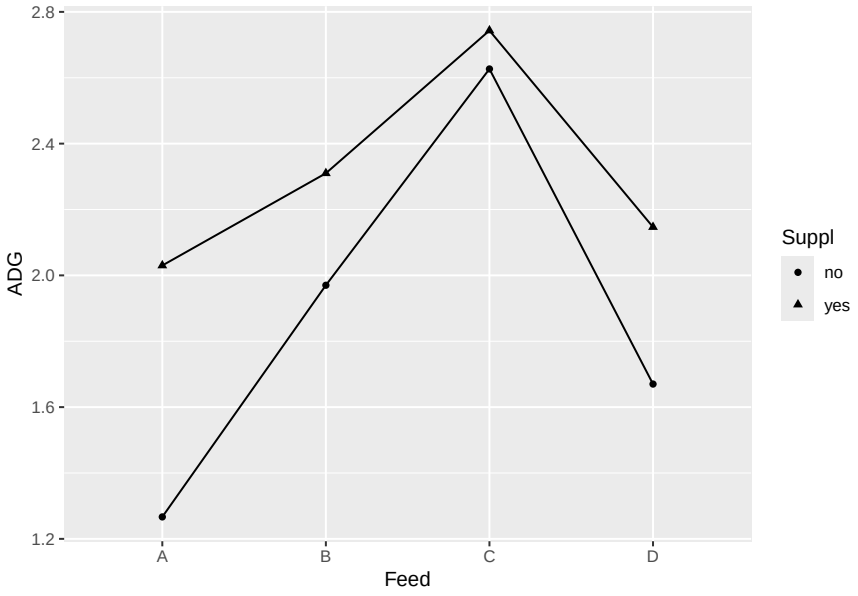
Q18 (6 pts)

Fit an appropriate model and show an appropriate ANOVA table in your exam.

```
## Analysis of Variance Table
##
## Response: ADG
##            Df Sum Sq Mean Sq F value    Pr(>F)
## Suppl       1 1.0795 1.07950 12.0913 0.0031093 **
## Feed        3 3.5069 1.16896 13.0933 0.0001417 ***
## Suppl:Feed  3 0.3291 0.10972  1.2289 0.3316943
## Residuals  16 1.4285 0.08928
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Q19

Create an appropriate summary plot of the data.

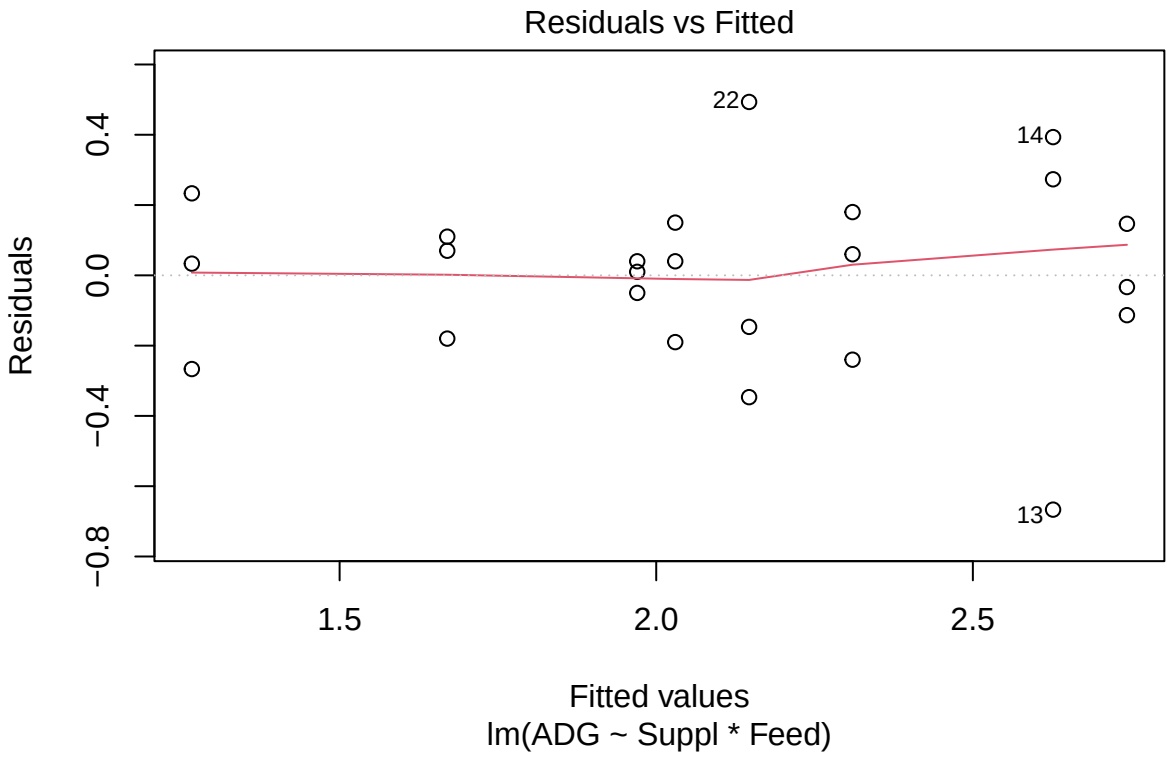


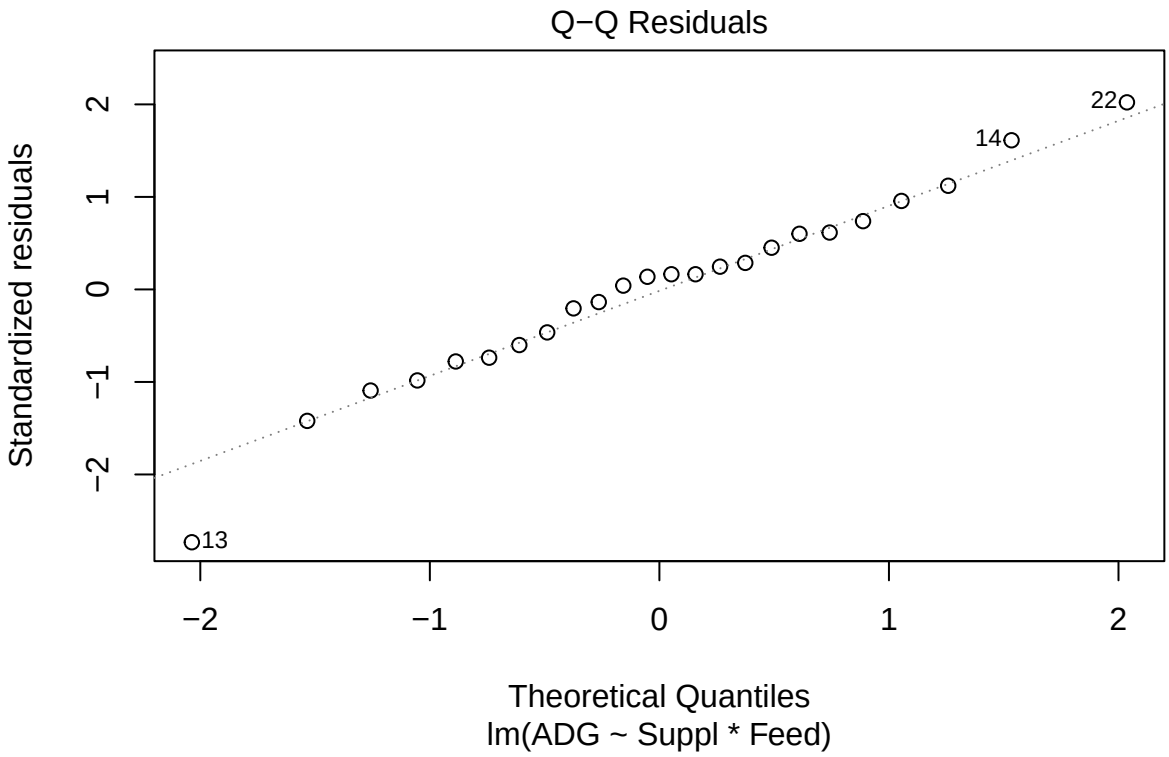
Q20 (6 pts)

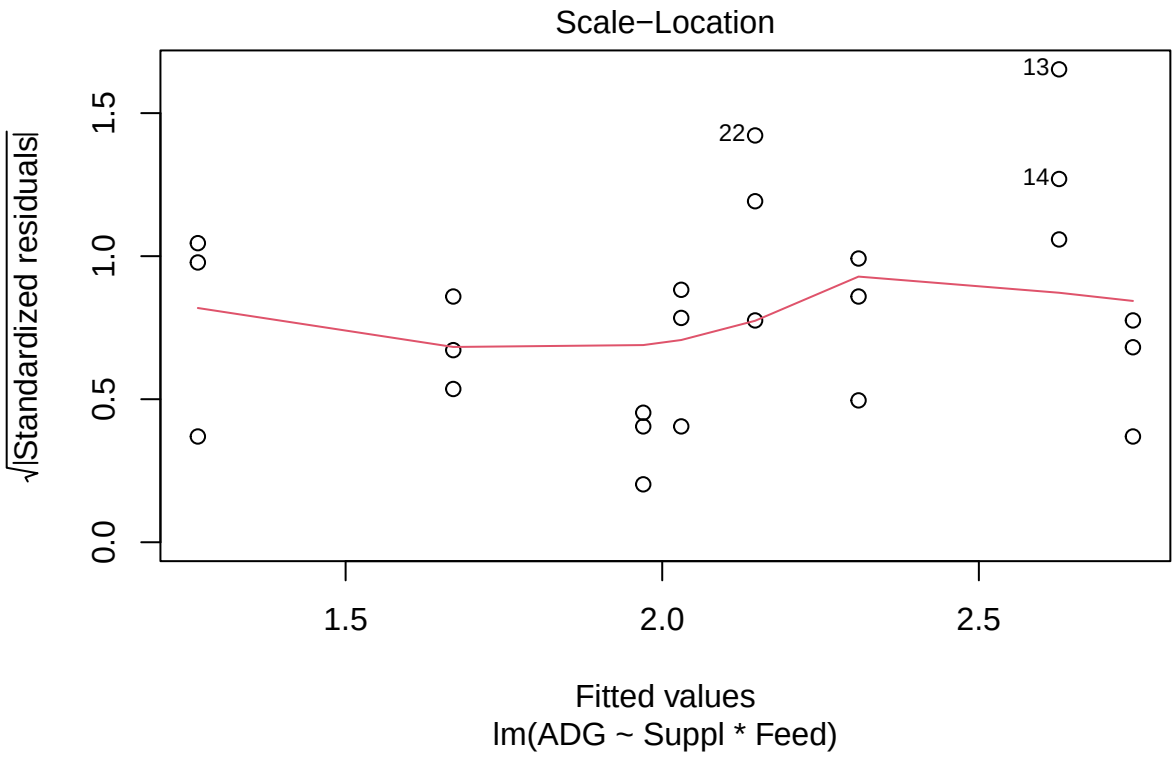
Provide two (common) diagnostic plots. Briefly discuss each plot and whether assumptions are satisfied. Notes: (1) Your discussion should be brief but still be specific. (2) Depending on how you set up your model, you may get a red note "hat values (leverages) are all = XXXX and there are no factor predictors". This message can safely be ignored.

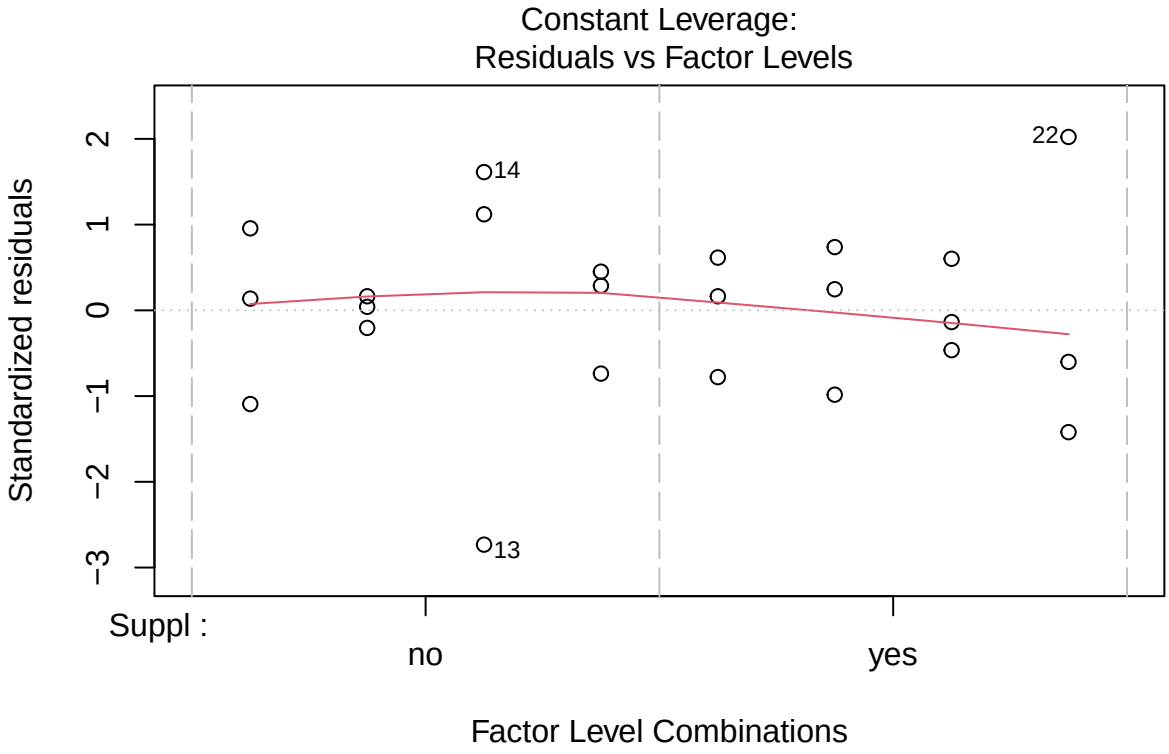
Response

For Plot 1. Residuals vs fitted i dont see any concerns of inequal or growing variance. For plot 2, Q-Q residulas, the data stays very close t line of normality. In both cases, requirements appear to be met.


Residuals vs Fitted
Residuals
0.4
0.0
−0.4
−0.8
22
14
13
1.5
2.0
2.5
Fitted values
lm(ADG ~ Suppl * Feed)


Q−Q Residuals
Standardized residuals
−2
−1
0
1
2
13
14
22
−2
−1
0
1
2
Theoretical Quantiles
lm(ADG ~ Suppl * Feed)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
1.5
2.0
2.5
Fitted values
lm(ADG ~ Suppl * Feed)
13
22
14



Q21

Considering the primary research question and your ANOVA table, provide appropriate pairwise comparisons. Use Tukey adjustment if appropriate.

```
##  contrast estimate    SE df t.ratio p.value
##  no - yes   -0.424 0.122 16  -3.477  0.0031
##
## Results are averaged over the levels of: Feed
```

#We could also look at by level by feed *****

Q22

Briefly summarize your conclusions from the previous question, being sure to address the research question. Avoid statistical jargon (reject, hypotheses, etc).

Response Cattle receiving the supplement had, on average, higher ADG compared to those not receiving the supplement. The exact change had some variance across feed types, but was generally positive for subjects who recieved the supplement.

Q23

Just for this question, consider the output from code like the following:

```
## $emmeans
##  Suppl emmean     SE df lower.CL upper.CL
##  no      1.88 0.0863 16     1.70     2.07
##  yes     2.31 0.0863 16     2.12     2.49
##
## Results are averaged over the levels of: Feed
## Confidence level used: 0.95
##
## $contrasts
##  contrast estimate    SE df t.ratio p.value
##  no - yes   -0.424 0.122 16  -3.477  0.0031
##
## Results are averaged over the levels of: Feed

## $emmeans
##  Feed emmean    SE df lower.CL upper.CL
##  A      1.65 0.122 16     1.39     1.91
##  B      2.14 0.122 16     1.88     2.40
##  C      2.69 0.122 16     2.43     2.94
##  D      1.91 0.122 16     1.65     2.17
##
## Results are averaged over the levels of: Suppl
## Confidence level used: 0.95
##
## $contrasts
##  contrast estimate    SE df t.ratio p.value
##  A - B      -0.492 0.173 16  -2.850  0.0511
##  A - C      -1.037 0.173 16  -6.009  0.0001
##  A - D      -0.260 0.173 16  -1.507  0.4563
##  B - C      -0.545 0.173 16  -3.159  0.0280
##  B - D       0.232 0.173 16   1.343  0.5506
##  C - D       0.777 0.173 16   4.502  0.0018
##
## Results are averaged over the levels of: Suppl
## P value adjustment: tukey method for comparing a family of 4 estimates
```

Specifically, consider the contrasts "no – yes" vs "B – C". Surprisingly, the magnitude of the estimated difference is larger for "B-C" but the p-value is also larger for "B-C". This indicates lower power despite the larger magnitude difference. Give **two different** reasons why this happened. Note: You do not need show the output, just discuss the result

---

Reason 1: This could be because of different variability between groups

Reason 2: It could also be because of something we didnt consider like a correlation structure between the treatments.

Appendix

```
#Retain this code chunk!!!
library(knitr)
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(message = FALSE)
#Q4
library(pwr)
grandMean <- (32+34+40)/3
ssBetweenMeans <- (32 - grandMean)^2 + (34 - grandMean)^2 + (40 - grandMean)^2
f <- sqrt( ( ssBetweenMeans / 3 ) / (8^2) )

pwr.anova.test(k=3, f=f, sig.level=0.05, power=0.80)

#Q6
t <- 6
b <- 18
k <- 4
r <-12

cond3 <- r*(k-1)/(t-1)

cond3
#Q18
feedData <- read.csv("FeedTrial.csv")
FeedModel <- lm(ADG ~ Suppl * Feed, data=feedData)
anova(FeedModel)
#Q19
library(ggplot2)

ggplot(feedData, aes(x=Feed, y=ADG, group=Suppl, shape=Suppl)) +
  stat_summary(fun=mean, geom="point") +
  stat_summary(fun=mean, geom="line")
#Q20
plot(FeedModel)

#Q21
library(emmeans)


emm <- emmeans(FeedModel, ~ Suppl)
pairs(emm, adjust="tukey")

#Q23
a <- emmeans(FeedModel, pairwise ~ Suppl)
```

```r
b<- emmeans(FeedModel, pairwise ~ Feed)

a

b
```