STAA 553: HW5

Matthew Stoebe

See Canvas Calendar for due date.
48 points total, 2 points per problem unless otherwise noted.
Add or delete code chunks as needed.
Content for Q1-Q15 is from section 07.
Content for Q16-Q20 is from section 09.

Biomass (Q1 - Q15)

A greenhouse study was done to examine the effect of three herbicides (A, B or C) and two water regimes (Low or High) for two plant types (Grass or Forb). The response variable is biomass. There are three reps per treatment combination for a total of 36 observations. Each observation was a potted plant. The 36 pots were randomly assigned without restriction to locations in the greenhouse. The data is available from Canvas as "Biomass.csv".

Important notes:

- Remember to run str() and then define things as.factor where needed.
- Change contrasts options to get meaningful Type 3 tests (using Anova): options(contrasts=c("contr.sum", "contr.poly"))
- Diagnostic plots are considered for several questions. You do NOT need to include these plots in your assignment. But you do need to discuss your findings.

Q1

Fit the three-way model with all interactions and show the Type 3 ANOVA table. You should find evidence of a 3 way interaction.

```
## 'data.frame':    36 obs. of  4 variables:
##  $ Type   : chr  "Grass" "Grass" "Grass" "Grass" ...
##  $ Herb   : chr  "A" "A" "A" "A" ...
##  $ Water  : chr  "Low" "Low" "Low" "High" ...
##  $ Biomass: num  16.4 19.9 18.9 18.2 21.2 21.5 18 18.8 17.1 19.4 ...
```

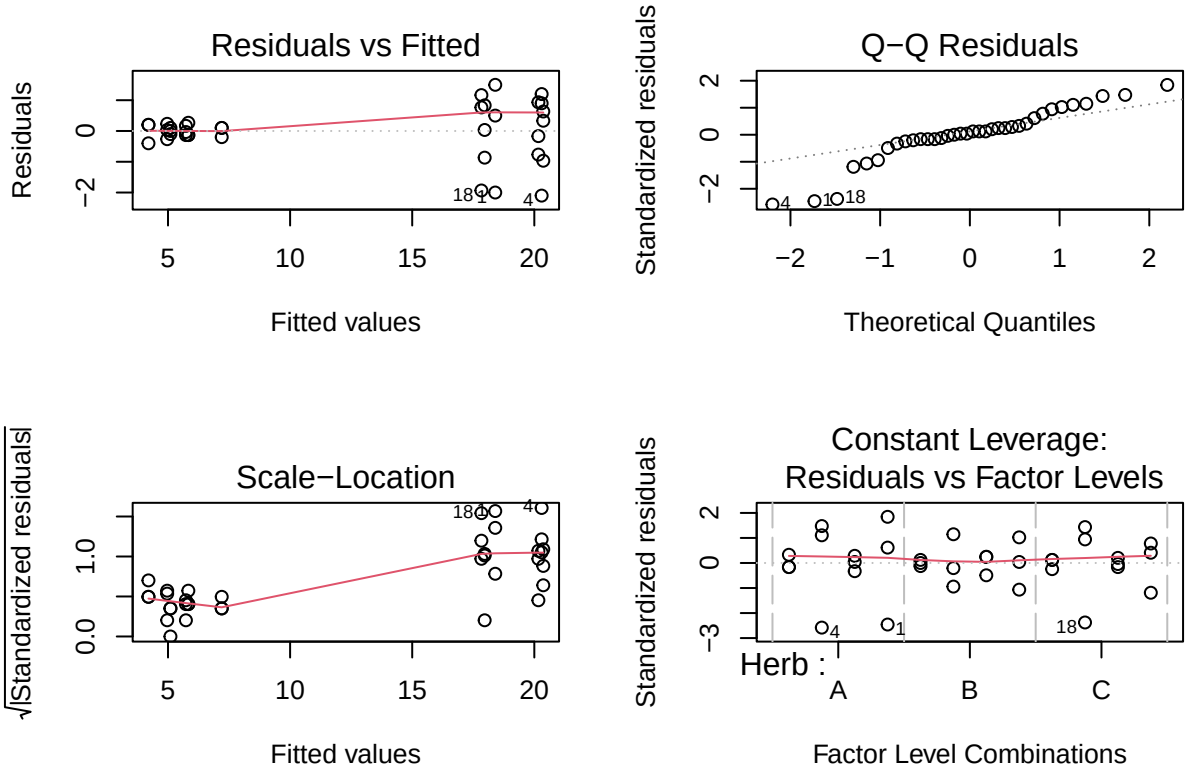
```
## Anova Table (Type III tests)
##
## Response: Biomass
##                Sum Sq Df   F value    Pr(>F)
## (Intercept)     5480.9  1 5520.8069 < 2.2e-16 ***
## Herb               5.2  2    2.5974  0.095253 .
```

```
## Water              5.8  1    5.8019  0.024046 *
## Type            1681.0  1 1693.2289 < 2.2e-16 ***
## Herb:Water         8.0  2    4.0501  0.030510 *
## Herb:Type          5.1  2    2.5845  0.096267 .
## Water:Type         0.7  1    0.6995  0.411203
## Herb:Water:Type   13.4  2    6.7356  0.004766 **
## Residuals         23.8 24
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Q2 (4 pts)

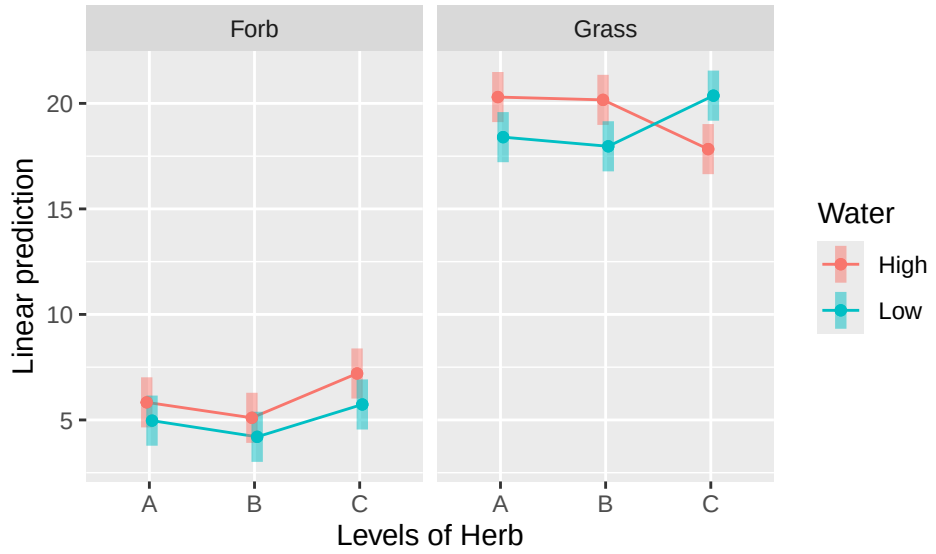
Use residual diagnostic plots to discuss whether model assumptions are satisfied. You do NOT need to include the plots in your assignment. But for full credit it should be clear which plot is being used to check which assumption.

Response By analyzing the "residuals vs fitted" graph I can determine that there is an issue of variance inequality as we see that the variance on the right side is farg reater than that on the left. We also see that on the !! plot it falls off the grey line heavily on either tail. Both of these cases indicate that there are some issues in our assumptions which could potentially be addressed by a log transform.



Q3

Create a summary graph (of emmeans) using code similar to what is provided.



Q4

Regardless of any concerns you may have about assumptions, use emmeans to calculate pairwise comparisons of Water (High vs Low) *for each level of Herb and Type*. Use code similar to what is provided.

```
## Herb = A, Type = Forb:
##  contrast   estimate    SE df t.ratio p.value
##  High - Low    0.867 0.814 24   1.065  0.2973
##
## Herb = B, Type = Forb:
##  contrast   estimate    SE df t.ratio p.value
##  High - Low    0.900 0.814 24   1.106  0.2796
##
## Herb = C, Type = Forb:
##  contrast   estimate    SE df t.ratio p.value
##  High - Low    1.467 0.814 24   1.803  0.0840
##
## Herb = A, Type = Grass:
##  contrast   estimate    SE df t.ratio p.value
##  High - Low    1.900 0.814 24   2.335  0.0282
##
## Herb = B, Type = Grass:
##  contrast   estimate    SE df t.ratio p.value
##  High - Low    2.200 0.814 24   2.704  0.0124
##
## Herb = C, Type = Grass:
##  contrast   estimate    SE df t.ratio p.value
##  High - Low   -2.533 0.814 24  -3.114  0.0047
```

Biomass: Forb Only (Q4 - Q9)

Now fit a two-way model (including interaction) for **Forb only**.

Q5

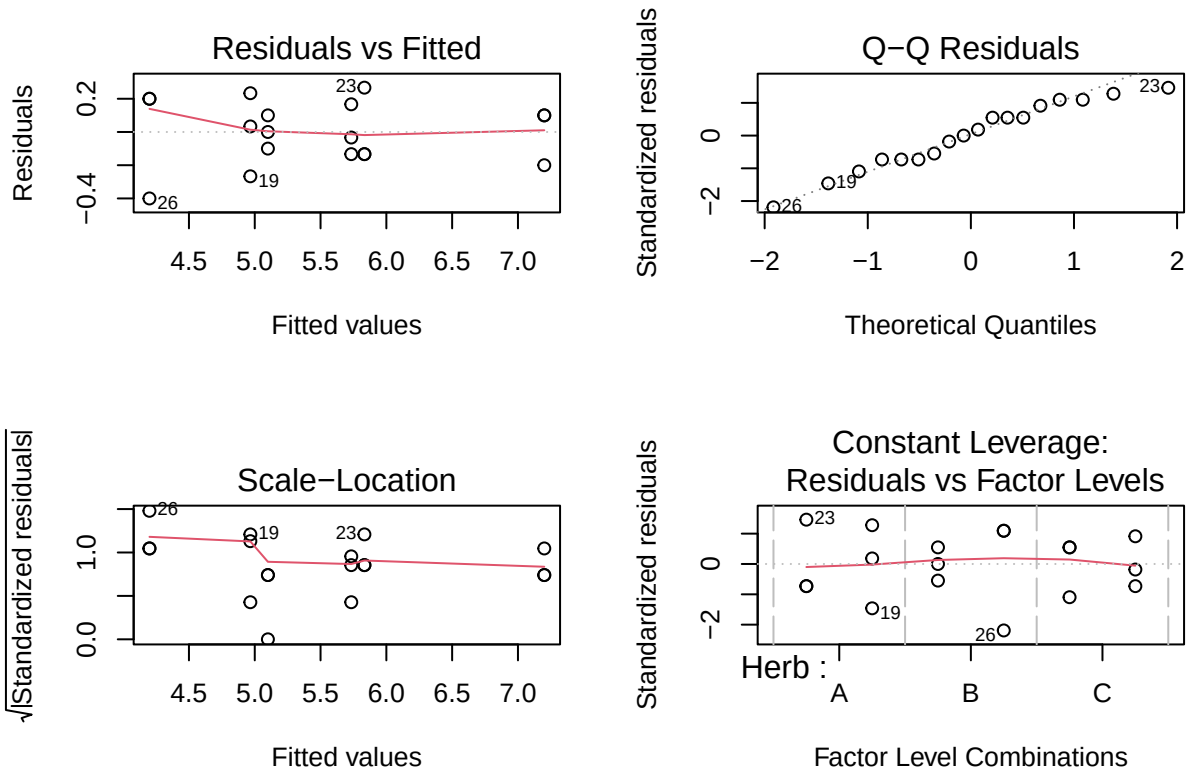
Show the Type 3 ANOVA table.

```
## Anova Table (Type III tests)
##
## Response: Biomass
##             Sum Sq Df    F value    Pr(>F)
## (Intercept) 545.60  1 10912.0111 < 2.2e-16 ***
## Herb         10.00  2   100.0111 3.287e-08 ***
## Water         5.23  1   104.5444 2.815e-07 ***
## Herb:Water    0.34  2     3.4111   0.06715 .
## Residuals     0.60 12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Q6

Consider the diagnostics plots and (briefly) discuss whether model assumptions are (better) satisfied.

Response These assumptions are better satisfied. residuals vs fitteded does not show the same cone paggern, and the QQ i closer to the line of normality



Q7

Use emmeans to calculate pairwise comparisons of Water (High vs Low) *for each level of Herb*.

```
## Herb = A:
##  contrast   estimate    SE df t.ratio p.value
##  High - Low    0.867 0.183 12   4.747  0.0005
##
## Herb = B:
##  contrast   estimate    SE df t.ratio p.value
##  High - Low    0.900 0.183 12   4.930  0.0003
##
## Herb = C:
##  contrast   estimate    SE df t.ratio p.value
##  High - Low    1.467 0.183 12   8.033  <.0001
```

Q8

Use emmeans to calculate the comparison of Water (High vs Low) *averaging over the levels of Herb.*

```
##  contrast   estimate    SE df t.ratio p.value
##  High - Low     1.08 0.105 12  10.225  <.0001
##
## Results are averaged over the levels of: Herb
```

Q9

Considering the SE for the comparisons from Q7 (interaction comparisons) and Q8 (main effect comparison), which has higher power? Briefly discuss.

Response The main effect has higher comparison as indicated by p value, standard error, and estimate. it may make sense to focus on this effect as opposed to the interactive effects *****

Biomass: Grass Only (Q10 - Q13)

Now fit a two-way model (including interaction) for **Grass only**.

Q10

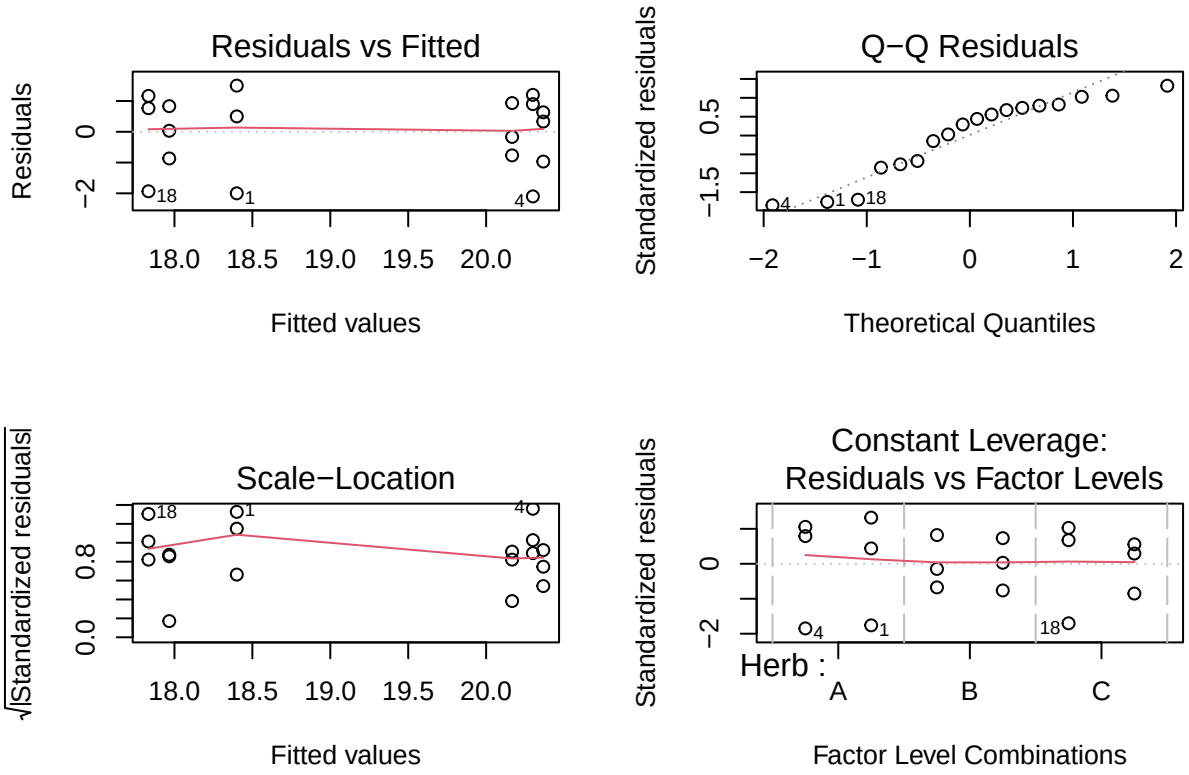
Show the Type 3 ANOVA table.

```
## Anova Table (Type III tests)
##
## Response: Biomass
##             Sum Sq Df   F value   Pr(>F)
## (Intercept) 6616.3  1 3418.3126 4.14e-16 ***
## Herb           0.3  2    0.0743  0.92878
## Water          1.2  1    0.6340  0.44135
## Herb:Water    21.1  2    5.4440  0.02077 *
## Residuals     23.2 12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Q11

Consider the diagnostics plots and (briefly) discuss whether model assumptions are (better) satisfied.

These plots indicate that the assumptions are reasonably satisfied, Errors do not have a clear pattern, and the QQ residuals adhere relatively closely to the line of normality



*****

Q12

Use emmeans to calculate pairwise comparisons of Water (High vs Low) *for each level of Herb.*

```
## Herb = A:
##  contrast   estimate   SE df t.ratio p.value
##  High - Low     1.90 1.14 12   1.673  0.1203
##
## Herb = B:
##  contrast   estimate   SE df t.ratio p.value
##  High - Low     2.20 1.14 12   1.937  0.0767
##
```

```
## Herb = C:
##  contrast   estimate   SE df t.ratio p.value
##  High - Low    -2.53 1.14 12  -2.230  0.0456
```

Q13

Would it be appropriate to calculate the comparison of Water (High vs Low) *averaging over the levels of Herb*? Briefly discuss.

Response This may not be wise as we saw that there is a signifficant interarctionb etween grass herb and water earlier in the assignment *****

Biomass: Compare Models (Q14 - Q15)

Now we compare the three-way model to the separate two-way models.

Q14

Give (at least) *one benefit* of splitting the analysis by Type (running separate 2way ANOVAs for Grass and Forb). Your answer should be *based on specific output.*

Response One benefit of running separate two-way ANOVAs is that each model is tailored to the specific variability of that plant type. For example, if the Forb model shows a particularly strong Herb $\times$ Water interaction that is diluted in the full three-way analysis, analyzing Forb separately can reveal these differences more clearly. Additionanlly, it allows us to seperate the fact that while forb has strong signal, Grass does not. Seperating the dataset makes this easier to see and to proceed with. *****

Q15

Give (at least) *one weakness* of splitting the analysis by Type as compared to the full 3way ANOVA model.

Response We loose tha ability to test differences between plant types, and are not able to analyze the 3-way interaction that we saw was signifficant *****

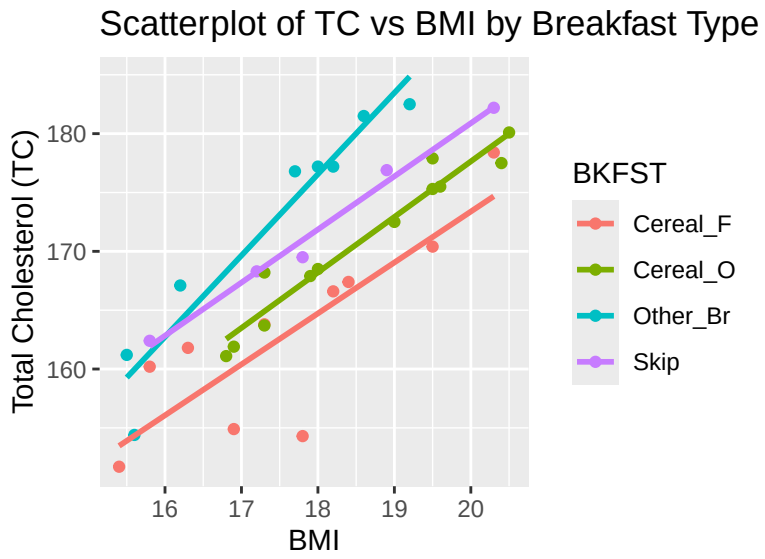
Breakfast (Q16 - Q20)

We return to the breakfast data from HW3. A study was done to examine whether breakfast choice was associated with cholesterol levels in children. A total of n=35 fourth and fifth graders were included in the study. Based on survey response, children were identified as one of (g = 4) four (BKFST) breakfast types: Cereal_F (cereal with fiber), Cereal_O (other cereal), Other_Br (other breakfast) or Skip (no breakfast). Note that the sample sizes are unequal. The height and weight of each child was used to determine their Body Mass Index (BMI). BMI is not of direct research interest, but will be considered as a covariate in some models. The response variable is plasma total cholesterol (TC). The data is available from Canvas as Breakfast.csv.

Q16

Construct a scatterplot of TC (Y) vs BMI (X) for all BKFST groups on the same plot. Overlay a separate regression line for each BKFST group.

```
## 'data.frame':    35 obs. of  3 variables:
##  $ BKFST: chr  "Other_Br" "Other_Br" "Other_Br" "Other_Br" ...
##  $ BMI  : num  18 18.2 19.2 18.6 16.2 15.6 17.7 15.5 17.2 20.3 ...
##  $ TC   : num  177 177 182 182 167 ...
```



Q17 (0 pts)

Calculate a table of summary statistics including sample size, mean, sd by BKFST group. (0 pts, because we already did this for HW3).

```
## # A tibble: 4 x 4
##   BKFST        n mean_TC sd_TC
##   <fct>    <int>   <dbl> <dbl>
## 1 Cereal_F    10    163.  8.19
```

```
## 2 Cereal_O    12    171.  6.54
## 3 Other_Br     8    172. 10.2
## 4 Skip         5    172.  7.75
```

Q18

Fit a one-way model (using BKFST as the predictor).

Q18A (0 pts)

Show the ANOVA table. (0 pts, because we already did this for HW3).

```
## Anova Table (Type III tests)
##
## Response: TC
##             Sum Sq Df    F value  Pr(>F)
## (Intercept) 904001  1 13724.3164 < 2e-16 ***
## BKFST          531  3     2.6891 0.06343 .
## Residuals     2042 31
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Q18B

Calculate Tukey adjusted pairwise comparisons for BKFST.

```
##  contrast            estimate   SE df t.ratio p.value
##  Cereal_F - Cereal_O   -7.892 3.48 31  -2.271  0.1270
##  Cereal_F - Other_Br   -9.287 3.85 31  -2.413  0.0956
##  Cereal_F - Skip       -8.910 4.45 31  -2.004  0.2082
##  Cereal_O - Other_Br   -1.396 3.70 31  -0.377  0.9814
##  Cereal_O - Skip       -1.018 4.32 31  -0.236  0.9953
##  Other_Br - Skip        0.378 4.63 31   0.082  0.9998
##
## P value adjustment: tukey method for comparing a family of 4 estimates
```

Q19

Now fit a model including both BKFST and BMI (but no interaction).

Q19A

Show the Type 3 ANOVA table.

```
## Anova Table (Type III tests)
##
## Response: TC
##              Sum Sq Df F value    Pr(>F)
## (Intercept) 1347.89  1 119.700 5.413e-12 ***
## BKFST        524.77  3  15.534 2.785e-06 ***
## BMI         1704.11  1 151.335 2.992e-13 ***
## Residuals    337.82 30
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Q19B

Show the emmeans for BKFST.

```
##  BKFST    emmean   SE df lower.CL upper.CL
##  Cereal_F    165 1.07 30      162      167
##  Cereal_O    168 1.00 30      166      170
##  Other_Br    175 1.21 30      173      177
##  Skip        172 1.50 30      168      175
##
## Confidence level used: 0.95
```

Q19C

Calculate Tukey adjusted pairwise comparisons for BKFST.

```
##  contrast            estimate   SE df t.ratio p.value
##  Cereal_F - Cereal_O    -3.04 1.49 30  -2.042  0.1957
##  Cereal_F - Other_Br   -10.36 1.59 30  -6.502  <.0001
##  Cereal_F - Skip        -6.86 1.85 30  -3.715  0.0044
##  Cereal_O - Other_Br    -7.32 1.61 30  -4.561  0.0004
##  Cereal_O - Skip        -3.81 1.80 30  -2.119  0.1703
##  Other_Br - Skip         3.51 1.93 30   1.818  0.2851
##
## P value adjustment: tukey method for comparing a family of 4 estimates
```

Q19D (4 pts)

Briefly summarize your findings from the previous question (using alpha = 0.05).

Response all relationships besides Cereal F - Cereal 0 and Cereal 0 - skip and Other_BR-Skip are significant *****

Q20

Compare the results from the one-way model (Q18) vs the ANCOVA model (Q19). Briefly explain why we were able to detect differences using the ANCOVA model, when we did not detect differences using the one-way model. Your answer should be based on *specific output*. Hint: You may want to calculate MSResid.

Response In the one way model, we do not account for variablility explained by BMI which goes into the residuals. This results in a lower power test, higher variability, and the inability to isolate the affect of the cereal eaten. we confirm this in the second analysis where after controlling for BMI, the relationship between Breakfasts becomes apparent and signifficant. *****

Appendix

```
#Retain this code chunk!!!
library(knitr)
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(message = FALSE)
knitr::opts_chunk$set(warning = FALSE)
#Q1
library(car)

options(contrasts = c("contr.sum", "contr.poly"))
biomass <- read.csv("Biomass.csv")
str(biomass)

biomass$Herb  <- as.factor(biomass$Herb)
biomass$Water <- as.factor(biomass$Water)
biomass$Type  <- as.factor(biomass$Type)

BM_3way <- lm(Biomass ~ Herb * Water * Type, data = biomass)

Anova(BM_3way, type = 3)

#Q2
par(mfrow = c(2,2))
plot(BM_3way)
#Q3
library(emmeans)
```

```r
emmip(BM_3way, Water ~ Herb | Type, CIs = TRUE)
#Q4
emout1 <- emmeans(BM_3way, ~ Water|Herb*Type)
pairs(emout1)
#Q5
forb <- subset(biomass, Type == "Forb")
BM_forb <- lm(Biomass ~ Herb * Water, data = forb)
Anova(BM_forb, type = 3)

#Q6
par(mfrow = c(2,2))
plot(BM_forb)
#Q7
emout_forb <- emmeans(BM_forb, ~ Water | Herb)
pairs(emout_forb)
#Q8
emout_forb_avg <- emmeans(BM_forb, ~ Water)
pairs(emout_forb_avg)
#Q10
grass <- subset(biomass, Type == "Grass")
BM_grass <- lm(Biomass ~ Herb * Water, data = grass)
Anova(BM_grass, type = 3)
#Q11
par(mfrow = c(2,2))
plot(BM_grass)
#Q12
emout_grass <- emmeans(BM_grass, ~ Water | Herb)
pairs(emout_grass)
#Q16
library(ggplot2)
breakfast <- read.csv("Breakfast.csv")
str(breakfast)

breakfast$BKFST <- as.factor(breakfast$BKFST)

ggplot(breakfast, aes(x = BMI, y = TC, color = BKFST)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Scatterplot of TC vs BMI by Breakfast Type",
       x = "BMI",
       y = "Total Cholesterol (TC)")
#Q17
library(dplyr)
breakfast_summary <- breakfast %>%
  group_by(BKFST) %>%
  summarize(n = n(), mean_TC = mean(TC, na.rm = TRUE), sd_TC = sd(TC, na.rm = TRUE))
breakfast_summary
#Q18A
lm_oneway <- lm(TC ~ BKFST, data = breakfast)
Anova(lm_oneway, type = 3)
#Q18B
library(emmeans)
emmeans_oneway <- emmeans(lm_oneway, ~ BKFST)
```

```r
pairs(emmeans_oneway, adjust = "tukey")
#Q19A
lm_ancova <- lm(TC ~ BKFST + BMI, data = breakfast)
Anova(lm_ancova, type = 3)
#Q19B
emmeans_ancova <- emmeans(lm_ancova, ~ BKFST)
emmeans_ancova
#Q19C
pairs(emmeans_ancova, adjust = "tukey")
```