# STP 531

# Applied Analysis of Variance

# Homework 4

Nathan A. Nguyen

31 March 2021

## Question 19.2

The student fails to understand that the number of treatments will increase in the number and complexity because you will have treatment combinations for multi-factor studies. Suppose you have one factor (A) with 3 levels and factor (B) with 3 levels. The number of treatments is $a = 3, b = 3 \implies 3 * 3 = 9$ combination of treatments. The interpretation will also become more complicated if you have interaction between the two factors as well. Imagine now that you had $2^+$ factor studies, you will have p-way interactions to analyze (if present) and many more combinations of treatments to look at.

## Question 19.5

### Part A

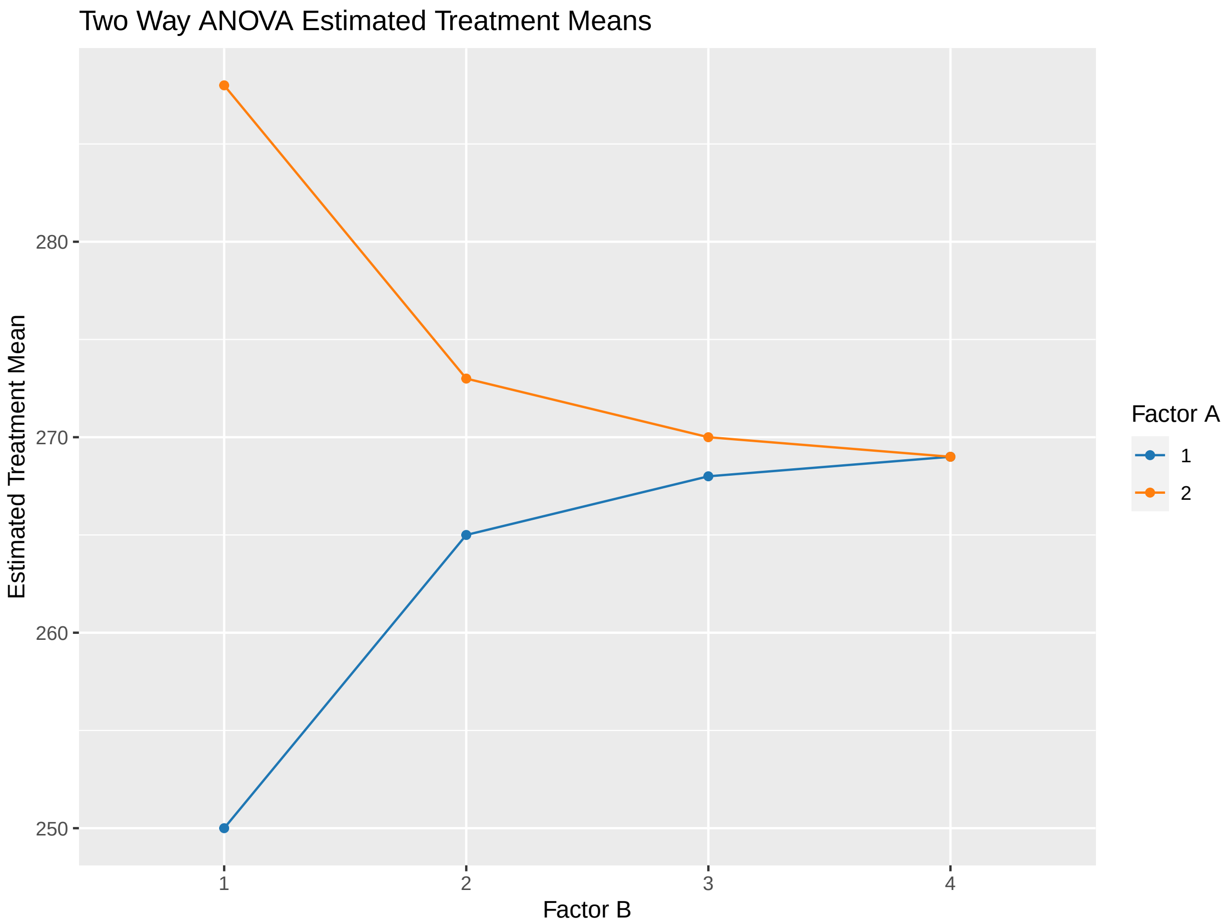
$$
\begin{aligned}
\beta_j &= \mu_{.j} - \mu_{..} \\
\mu_{..} &= 269, \mu_{.j} = 269, j = 1, 2, 3, 4 \\
\implies & \beta_j = 0
\end{aligned}
$$

In other words, there are no main effects for factor B.

```
## [1] "This is the grand mean: 269"
```

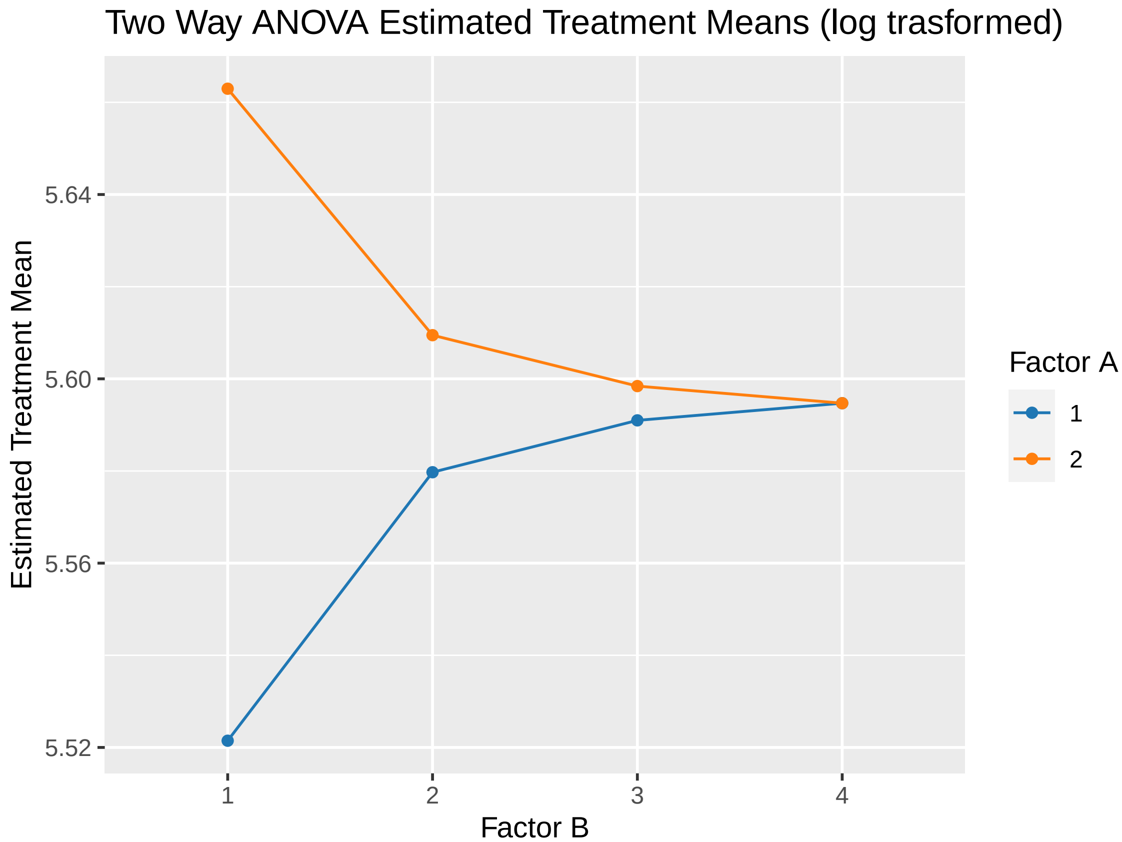
```
## [1] "These are the factor B means: 269" "These are the factor B means: 269"
## [3] "These are the factor B means: 269" "These are the factor B means: 269"
```

### Part B




The lack of parallelism between the two curves suggests that there are interactions present. The interactions look important because the two curves start off at very different "heights" (treatment means) and then converge towards a common treatment mean as the levels of factor B increase.

**Part C**




A logarithmic transformation of the treatment means does not make the interactions go away, so we still say that there are interactions present.

# Question 19.16

## Part A

Table 1: Fitted Values

| | Y11 | Y12 | Y13 | Y21 | Y22 | Y23 | Y31 | Y32 | Y33 |
|---|---|---|---|---|---|---|---|---|---|
| Est. Mean | 59.8 | 47.8 | 58.4 | 48.4 | 61.2 | 56.2 | 60.2 | 60.8 | 49.6 |

## Part B

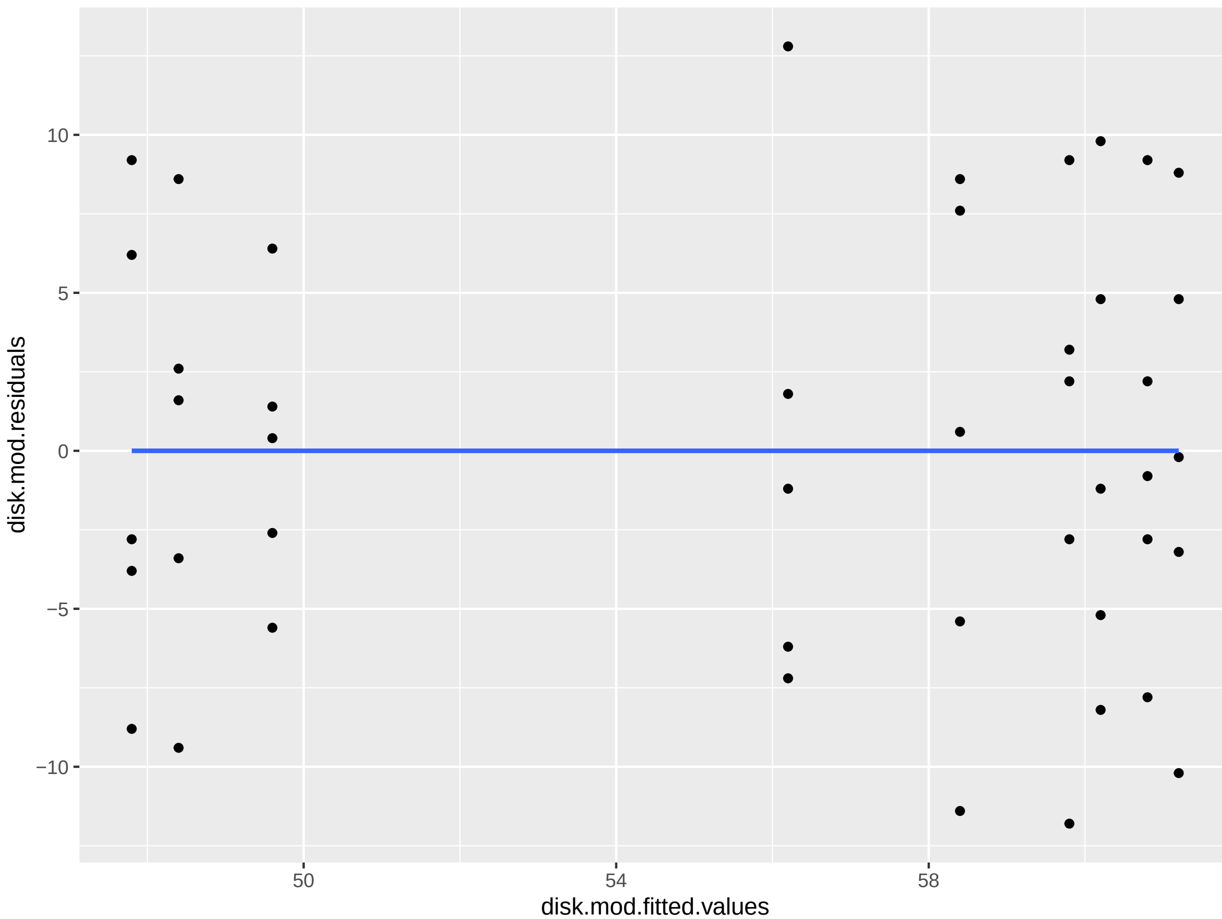
Table 2: Disk Residuals

| disk.mod.residuals |
|---|
| 2.2 |
| -11.8 |
| 3.2 |
| -2.8 |
| 9.2 |
| 9.2 |
| -2.8 |
| -8.8 |
| 6.2 |
| -3.8 |
| 0.6 |
| -5.4 |
| 8.6 |
| 7.6 |
| -11.4 |
| 2.6 |
| 8.6 |
| -3.4 |
| 1.6 |
| -9.4 |
| -0.2 |
| -3.2 |
| 8.8 |
| 4.8 |
| -10.2 |
| -1.2 |
| 1.8 |
| -6.2 |
| 12.8 |
| -7.2 |
| -1.2 |
| 4.8 |
| -5.2 |
| -8.2 |
| 9.8 |
| -2.8 |
| 2.2 |
| 9.2 |
| -7.8 |

| disk.mod.residuals |
| ---: |
| -0.8 |
| -2.6 |
| 6.4 |
| 1.4 |
| -5.6 |
| 0.4 |

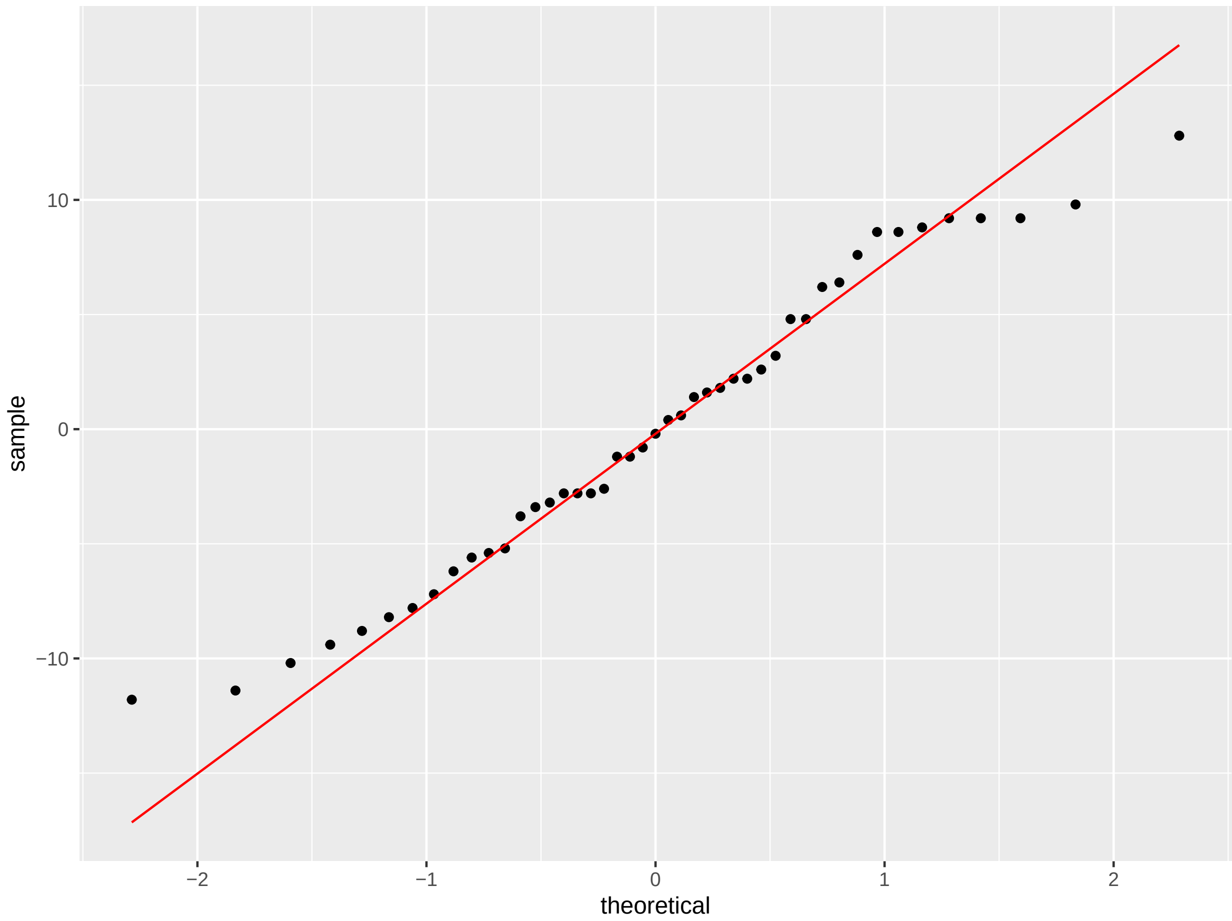
Not the prettiest format...

## Part C



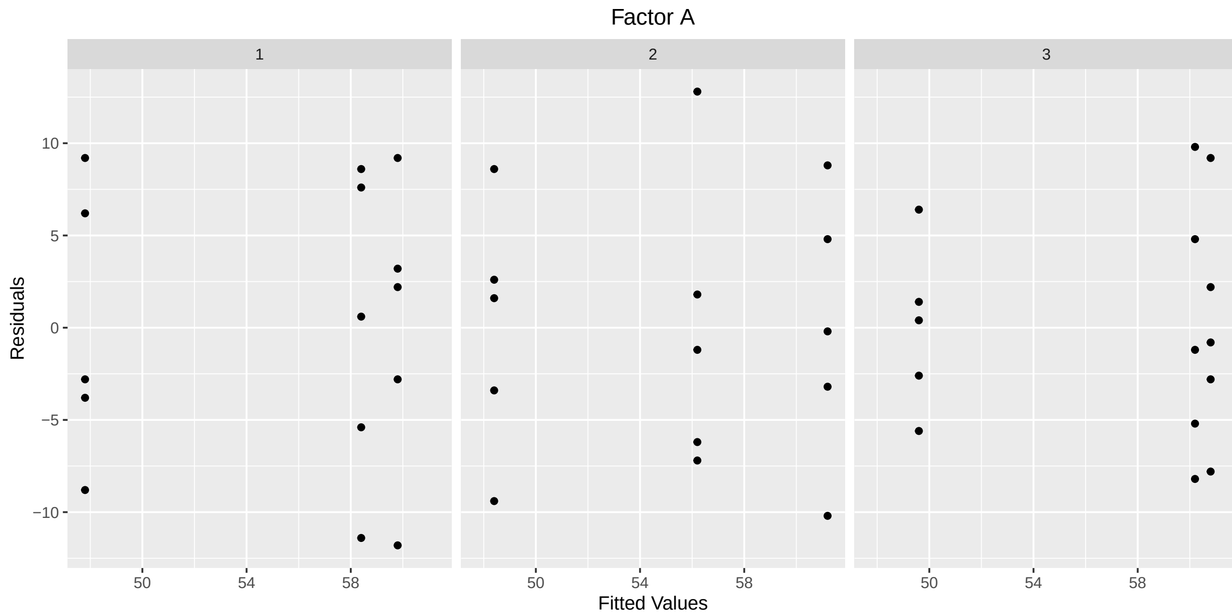

I don't see anything wrong with the residuals vs. fitted value plot. The assumptions are met. The QQ-plot (next) also don't suggest any serious departures as well.

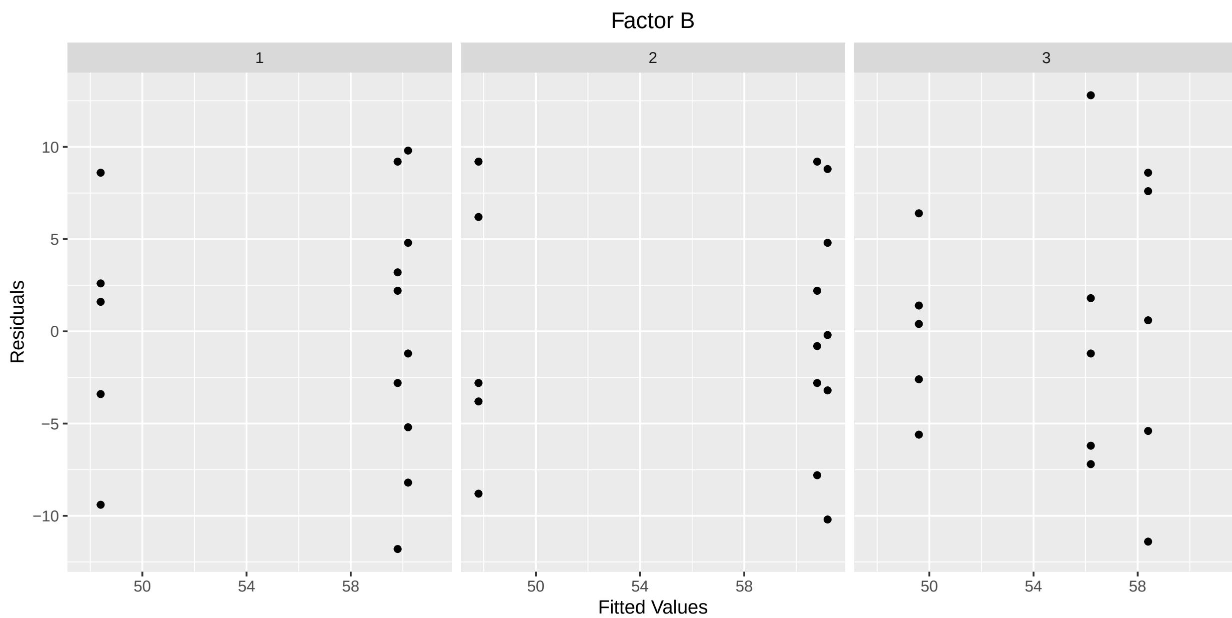
## Part D




The calculated residuals hug the line pretty well, so the normality assumption is reasonable. The coefficient of correlation is 0.9876149, more evidence that the normality assumption is appropriate.

## Part E

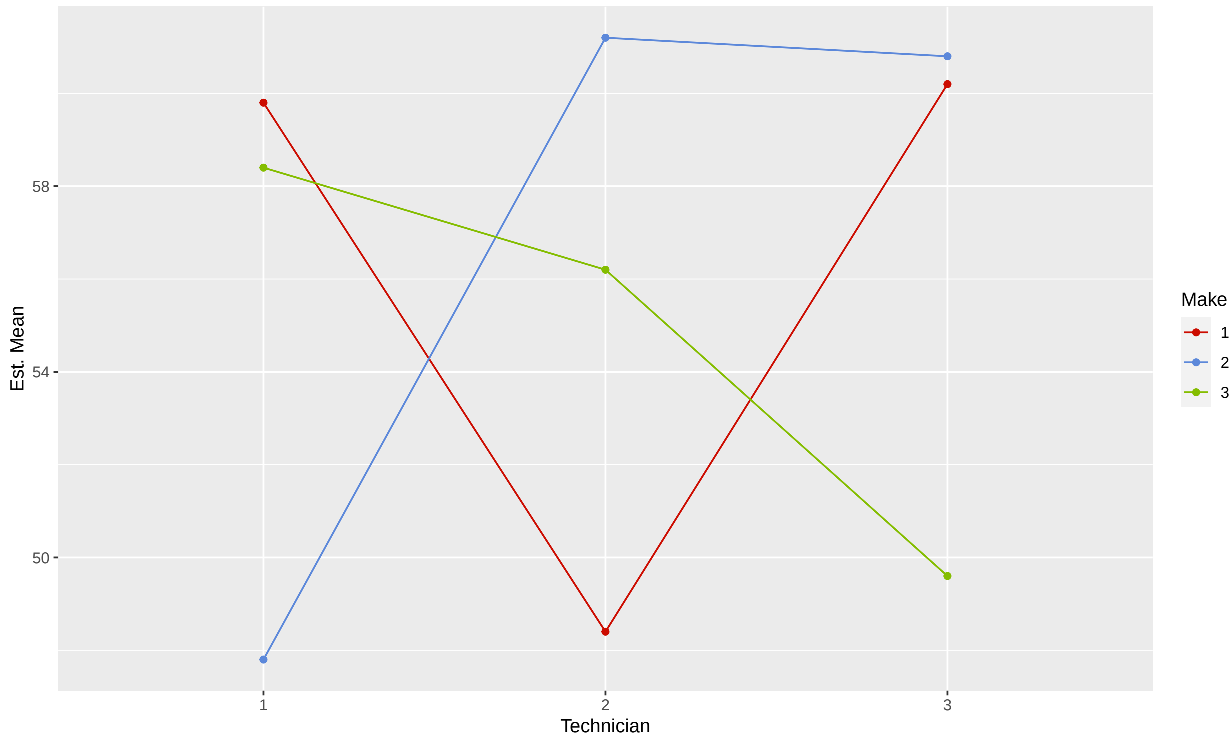



The homogeneity of variance assumption looks reasonable. I don't see any obvious patterns across the factor levels for both factors.

# Question 19.17

## Part A




There doesn't appear to be any main factor effects going on. It doesn't look like no one technician is outperforming the other and conversely, there doesn't appear to be one technician that is under-performing. The same can be said about the make as well. However, there appears to be significant interaction between the technician and type of disk drive.

## Part B

Table 3: Disk Drive ANOVA Table

| | Df | Sum.Sq | Mean.Sq | F.value | Pr..F. |
|---|---|---|---|---|---|
| A | 2 | 24.57778 | 12.28889 | 0.2362743 | 0.7907788 |
| B | 2 | 28.31111 | 14.15556 | 0.2721641 | 0.7632826 |
| A:B | 4 | 1215.28889 | 303.82222 | 5.8414869 | 0.0009941 |
| Residuals | 36 | 1872.40000 | 52.01111 | NA | NA |

## Part C

$$\begin{aligned}
H_0 &: (\alpha\beta)_{ij} = 0\ \forall_{ij} \\
H_a &: \text{ not all } (\alpha\beta)_{ij} = 0 \\
F^* &= \frac{MSAB}{MSE} = \frac{303.82}{52.01} \approx 5.841 \\
F_{critical} &= F(0.99; 4, 36) \approx 3.89 \\
F^* &> F_{critical},\ \text{reject } H_0\ (\alpha = 0.01) \\
p &\approx 0.000994
\end{aligned}$$

Since $F^* > F_{critical}$ and $p << 0.01$, there is sufficient evidence to reject the null hypothesis. There are interactions between the two factors present.

## Part D

$$\begin{aligned}
H_0 &: (\alpha)_i = 0\ \forall_i \\
H_a &: \text{ not all } \alpha_i = 0 \\
F^* &= \frac{MSA}{MSE} = \frac{12.29}{52.01} \approx 0.24 \\
F_{critical} &= F(0.99; 2, 36) \approx 5.25 \\
F^* &< F_{critical}, \text{fail to reject } H_0 \\
p &\approx 0.79 > \alpha = 0.01
\end{aligned}$$

$$\begin{aligned}
H_0 &: (\beta)_j = 0\ \forall_j \\
H_a &: \text{ not all } \beta_j = 0 \\
F^* &= \frac{MSB}{MSE} = \frac{14.16}{52.01} \approx 0.27 \\
F_{critical} &= F(0.99; 2, 36) \approx 5.25 \\
F^* &< F_{critical}, \text{fail to reject } H_0 \\
p &\approx 0.76 > \alpha = 0.01
\end{aligned}$$

We fail to reject both of the null hypothesis for testing for main effects for factor A and factor B. There is sufficient evidence to suggest that there are no meaningful main effects. Furthermore, it is not meaningful to test for the main effects since we have established that there are interactions present. Those are more interesting to look at.

## Part F

Yes. Part D does confirm my initial speculations in part A.

# Question 19.33

## Part A

$$
\begin{aligned}
\hat{\mu_{11}} &= Y\bar{1}1. = 59.8,\ \alpha = 0.01 \\
s^2\{\bar{Y_{11}}\} &= \frac{MSE}{n} = \frac{52.01}{5} \approx 10.402 \\
s\{\bar{Y_{11}}\} &\approx 3.23 \\
t^* &= t(0.995, 36) \approx 2.72 \\
\bar{Y_{11}} \pm t^* s\{\bar{Y_{11}}\} &= 59.8 \pm 2.792(3.23)
\end{aligned}
$$

```
##      Lower    Upper
## 1 51.02908 68.57092
```

The 99% confidence interval for $\mu_{11}$ is $51.03 \leq \mu_{11} \leq 68.57$. The mean time it takes for technician 1 working on disk drive type 1 is somewhere between 51.03 and 68.57 minutes required to complete the repair job. This is at the 99% confidence level.

## Part B

$$
\begin{aligned}
\hat{D} &= \bar{Y_{22}} - \bar{Y_{21}},\ \alpha = 0.01 \\
\bar{Y_{22}} &= 61.2,\ \bar{Y_{21}} = 48.4 \\
\implies \hat{D} &= 12.8 \\
s^2\{\hat{D}\} &= \frac{2MSE}{n} = \frac{2(52.01)}{5} =\approx 20.804 \\
s\{\hat{D}\} &= 4.56,\ t^* = 2.72 \\
\hat{D} \pm t^* s\{\hat{D}\} &= 12.8 \pm 2.72(4.56)
\end{aligned}
$$

```
##       Lower    Upper
## 1 0.3960493 25.20395
```

The 99% interval for this contrast is $0.396 \leq D \leq 25.204$ ($0.396 \leq \mu_{22} - \mu_{21} \leq 25.204$). So, this says that with 99% confidence that the mean time to complete a repair job for disk type 2 is longer than disk type 1 (FOR TECHNICIAN 2).

## Part C

$$\text{There are g comparisons, } g = 3$$
$$B = t\left(1 - \frac{0.05}{2(3)}; 36\right) \approx 2.51$$
$$D_i \pm Bs\{D_i\},\ i = 1, 2, ..., 9$$

Table 4: Comparisons

| D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 | D9 |
|---|---|---|---|---|---|---|---|---|
| 12 | 1.4 | -10.6 | -12.8 | -7.8 | 5 | -0.6 | 10.6 | 11.2 |

Table 5: Bonferroni 95% comparisons

| | Lower | Upper |
|---|---|---|
| D1 | 0.5467853 | 23.4532147 |
| D2 | -10.0532147 | 12.8532147 |
| D3 | -22.0532147 | 0.8532147 |
| D4 | -24.2532147 | -1.3467853 |
| D5 | -19.2532147 | 3.6532147 |
| D6 | -6.4532147 | 16.4532147 |
| D7 | -12.0532147 | 10.8532147 |
| D8 | -0.8532147 | 22.0532147 |
| D9 | -0.2532147 | 22.6532147 |

Using the Bonferroni procedure with family confidence 0.95:

- technician 1 works best with drive model 2 (time to repair is short than model 1)
- technician 3 works best with drive model 3; however not much different than the other models
- technician 2 works best with drive model 1

# Appendix

## Question 19.5

### Part A

```
Q19.5_data <- read.table("http://users.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/textdatasets/Kut
                         col.names = c("means", "FA", "FB"))
Q19.5_data$FA <- as.factor(Q19.5_data$FA)
Q19.5_data$FB <- as.factor(Q19.5_data$FB)

mod.1 <- lm(means ~ FA*FB, data = Q19.5_data)
blah1 <- aov(means ~ FA*FB, data = Q19.5_data)
mod.1.table <- model.tables(blah1, type = "means", se = T)
temp1 <- mod.1.table$tables
paste0("This is the grand mean: ", temp1$`Grand mean`)
paste0("These are the factor B means: ", temp1$FB)
```

### Part B

```
means <- Q19.5_data %>%
  group_by(FA, FB) %>%
  summarise(means = mean(means))

p2 <- means %>%
  ggplot(aes(x = FB, y = means, color = FA)) +
  geom_line(aes(group = FA)) + geom_point() +
  labs(x = "Factor B",
       color = "Factor A",
       y = "Estimated Treatment Mean",
       title = "Two Way ANOVA Estimated Treatment Means") +
  scale_color_d3()
p2
```

### Part C

```
# means$means <- log(means$means)
p2 <- means %>%
  ggplot(aes(x = FB, y = log(means), color = FA)) +
  geom_line(aes(group = FA)) + geom_point() +
    labs(x = "Factor B",
       color = "Factor A",
       y = "Estimated Treatment Mean",
       title = "Two Way ANOVA Estimated Treatment Means (log trasformed)") +
  scale_color_d3()
p2
```

## Question 19.16

### Part A

```r
getdata <- function(...){
  e = new.env()
  name = data(..., envir = e)[1]
  e[[name]]
}

data2 <- getdata("DiskDriveService")
data2$A <- as.factor(data2$A)
data2$B <- as.factor(data2$B)

disk.mod <- lm(y ~ A*B, data = data2)
temp1 <- data.frame(disk.mod$fitted.values)
disk.fit = data.frame(
  Y11 = 59.8,
  Y12 = 47.8,
  Y13 = 58.4,
  Y21 = 48.4,
  Y22 = 61.2,
  Y23 = 56.2,
  Y31 = 60.2,
  Y32 = 60.8,
  Y33 = 49.6
)
row.names(disk.fit) <- "Est. Mean"
kable(disk.fit[], caption = "Fitted Values", format = "markdown") %>%
  kable_styling(position = "center")
```

### Part B

```r
temp2 <- data.frame(disk.mod$residuals)
disk.res <- t(temp2)
# ugly fomat.
kable(temp2[], caption = "Disk Residuals", format = "markdown") %>%
  kable_styling(position = "center")
```

### Part C

```r
blah2 <- cbind(temp1,temp2)
p3 <- blah2 %>%
  ggplot(aes(x = disk.mod.fitted.values, y = disk.mod.residuals)) +
  geom_point() + geom_smooth(method = "loess", se = F)
p3
```

### Part D

```r
p4 <- blah2 %>%
  ggplot(aes(sample = disk.mod.residuals)) +
  stat_qq() + stat_qq_line(color = "red")
p4
```

```r
cat <- qqnorm(blah2$disk.mod.residuals, plot.it = FALSE)
res.cor <- cor(cat$x, cat$y)
```

**Part E**

```r
disk.residuals <- data2
disk.residuals$modFit <- fitted(disk.mod)
disk.residuals$modRes <- resid(disk.mod)

p5 <- disk.residuals %>%
  ggplot(aes(x = modFit, y = modRes)) + geom_point() +
  labs(x = "Fitted Values",
       y = "Residuals",
       title = "Factor A") +
  theme(plot.title = element_text(hjust = 0.5)) +
  facet_wrap(~ A)
p5

p6 <- disk.residuals %>%
  ggplot(aes(x = modFit, y = modRes)) + geom_point() +
  labs(x = "Fitted Values",
       y = "Residuals",
       title = "Factor B") +
  theme(plot.title = element_text(hjust = 0.5)) +
  facet_wrap(~ B)
p6
```

## Question 19.17

### Part A

```
disk.mod <- lm(y ~ A*B, data = data2)
anova.tab <- data.frame(anova(disk.mod))

means <- data2 %>%
  group_by(A,B) %>%
  summarise(means = mean(y))

p7 <- means %>%
  ggplot(aes(x = A, y = means, color = B)) +
  geom_line(aes(group = B)) + geom_point() +
  scale_colour_startrek() +
  labs(x = "Technician",
       y = "Est. Mean",
       color = "Make")
p7
```

### Part B

```
kable(anova.tab[], caption = "Disk Drive ANOVA Table", format = "markdown") %>% kable_styling(position =
```

No, one factor doesn't explain most of the variability over the other factor. Their sum of squares and mean squares are different, but only marginally. Furthermore both p-values $p > 0.05$ for the main effects.

### Part C

```
F.star <- 303.82/52.01
F.crit <- qf(0.99,4,36)
p.value <- pf(303.82,4,36,lower.tail = F)
```

### Part D

```
F.starA <- 12.29/52.01
F.crit <- qf(0.99,2,36)
F.starB <- 14.16/52.01
pA <- pf(0.24,2,36,lower.tail = F)
pB <- pf(0.27,2,36,lower.tail = F)
```

### Part F

## Question 19.33

### Part A

```
mu11 <- 59.8
s2 <- 52.01/(5)
s <- sqrt(s2)
t <- qt(0.995,36)
mu11.LL <- mu11 - t*s; mu11.UL <- mu11 + t*s
moo <- data.frame(
  Lower = mu11.LL,
  Upper = mu11.UL
)
moo
```

### Part B

```
D = 61.2-48.4
s2 = (2*52.01)/5
s = sqrt(s2)
D.LL = D - t*s; D.UL = D +t*s
meow = data.frame(
  Lower = D.LL,
  Upper = D.UL
)
meow
```

### Part C

```
# for the for loop
D1 = means$means[1]-means$means[2]
D2 = means$means[1]-means$means[3]
D3 = means$means[2]-means$means[3]
D4 = means$means[4]-means$means[5]
D5 = means$means[4]-means$means[6]
D6 = means$means[5]-means$means[6]
D7 = means$means[7]-means$means[8]
D8 = means$means[7]-means$means[9]
D9 = means$means[8]-means$means[9]
D = c(D1,D2,D3,D4,D5,D6,D7,D8,D9) # to iterate over
pig = data.frame(
  D1 = D1,
  D2 = D2,
  D3 = D3,
  D4 = D4,
  D5 = D5,
  D6 = D6,
  D7 = D7,
  D8 = D8,
  D9 = D9
)
sd <- sqrt((2*52.01)/5) # same as prior

#bonferroni correction
```

```r
B =qt(0.9916667, 36) # g = 3 => 9 comparisons
Lower = rep(0,9) # initialize
Upper = rep(0,9)
# for loop to construct the lower and upper bounds for each comparison
for(i in 1:9){
  Lower[i] = D[i]-B*sd
  Upper[i] = D[i]+B*sd
}
Bon_pairwise = data.frame(Lower, Upper)
rownames(Bon_pairwise) <- c("D1", "D2", "D3",
                            "D4", "D5", "D6",
                            "D7", "D8", "D9")

kable(pig[], caption = "Comparisons", format = "markdown") %>%
  kable_styling(position = "center")

kable(Bon_pairwise[], caption = "Bonferroni 95% comparisons", format = "markdown") %>% kable_styling(po
```