Problem Set 2, Winter 2021

Adam Ten Hoeve

1/13/2021

Let's say that, in my excitement about seeing flour in the grocery store again, I decided to conduct a factorial experiment inspired by the experiment conducted by Lowe (1935) to learn more about how much fat doughnuts absorb in different conditions. Like Lowe, I used four types of fats (fat_type). I also used three types of flour (flour_type): all-purpose flour, whole wheat flour, and gluten-free flour. Again like Lowe, I cooked six identical batches of doughnuts in each flour and fat combination. Each batch contained 24 doughnuts, and the total fat (in grams) absorbed by the doughnuts in each batch was recorded (total_fat).

Question 1 - 5 points

You may need to process your data before you begin your analysis. Specifically, you will need to make sure that the variable type is set to 'factor' for both of your grouping variables and 'num' for your outcome variable.

```
doughnuts.factorial <- read.csv("doughnutsfactorial.csv", header=TRUE, sep=",") # Loads the CSV file in

# Put any further code needed for processing below:
doughnuts.factorial$fat_type = as.factor(doughnuts.factorial$fat_type)
doughnuts.factorial$flour_type = as.factor(doughnuts.factorial$flour_type)
doughnuts.factorial$sim_tot_fat = as.numeric(doughnuts.factorial$sim_tot_fat)

head(doughnuts.factorial)
```

```
##   fat_type flour_type sim_tot_fat
## 1        1         ap          78
## 2        1         ap          71
## 3        1         ap          80
## 4        1         ap          88
## 5        1         ap          62
## 6        1         ap          72
```

Question 2 - 10 points

You will conduct a two-way ANOVA with an interaction to determine the effect of fat type and flour type on the absorption of fat into the doughnuts. Before you begin your analysis, do the following two things:

1) Fill in the blanks in this sentence with the appropriate numbers: "This experiment had two factors, fat type and flour type, making this a __ by __ factorial design with __ cell means."

Your answer here: "…making this a 4 by 3 factorial design with 12 cell means.""

2) Create a table showing the cell means. Label your table in such a way that someone unfamiliar with the data set would be able to identify the cell means. You may use any package/s you like to create the table, but please be sure to include any package/s you use in the set up code chunk at the top of this document.

```
# Code for your table here

# Calculate the mean sim_tot_fat for each group combination
doughnuts.summary = doughnuts.factorial %>% group_by(fat_type, flour_type) %>% summarise(mean=mean(sim_
```

```
## `summarise()` regrouping output by 'fat_type' (override with `.groups` argument)
```

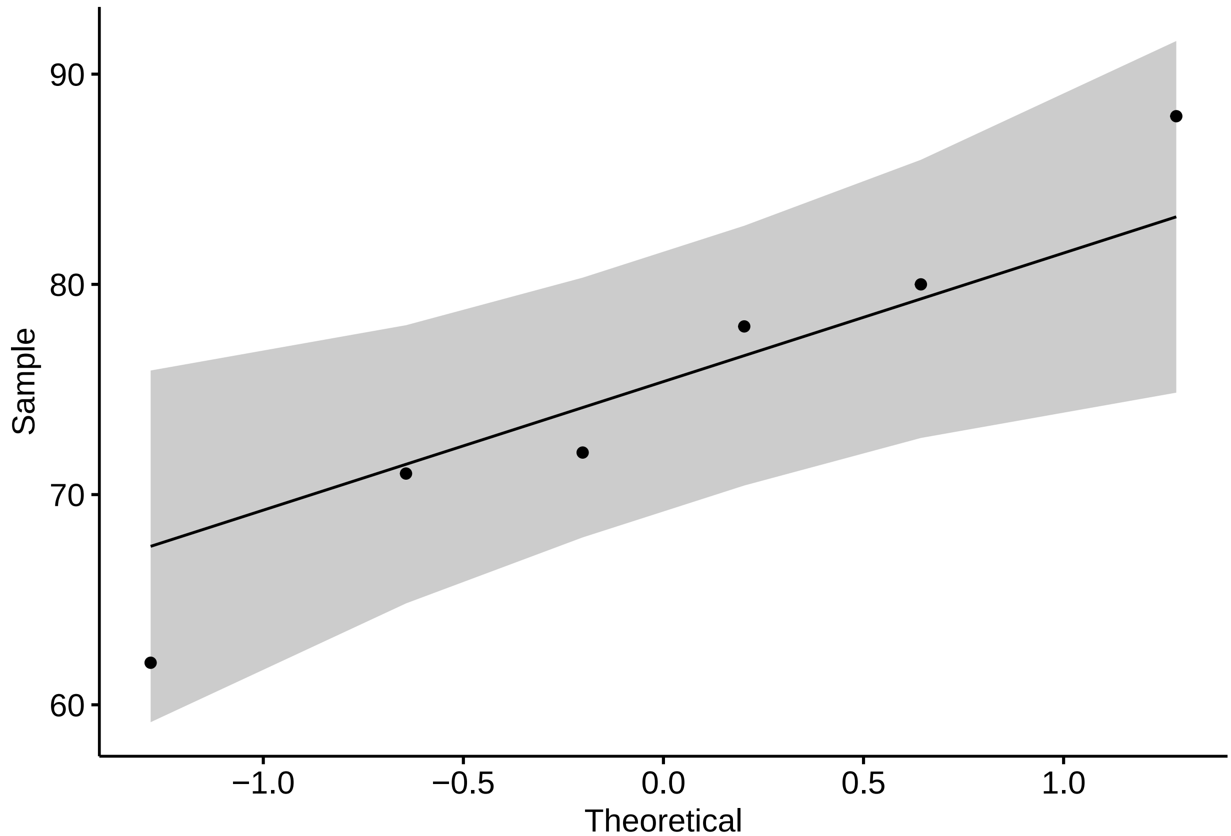
```
fat.types = c(1, 2, 3, 4)
flour.types = c("ap", "gf", "ww")
# Format the means into a matrix/table form.
doughnut.matrix = matrix(doughnuts.summary$mean, nrow=length(fat.types), ncol=length(flour.types),
                         byrow=TRUE, dimnames=list(fat.types, flour.types))
doughnut.matrix
```

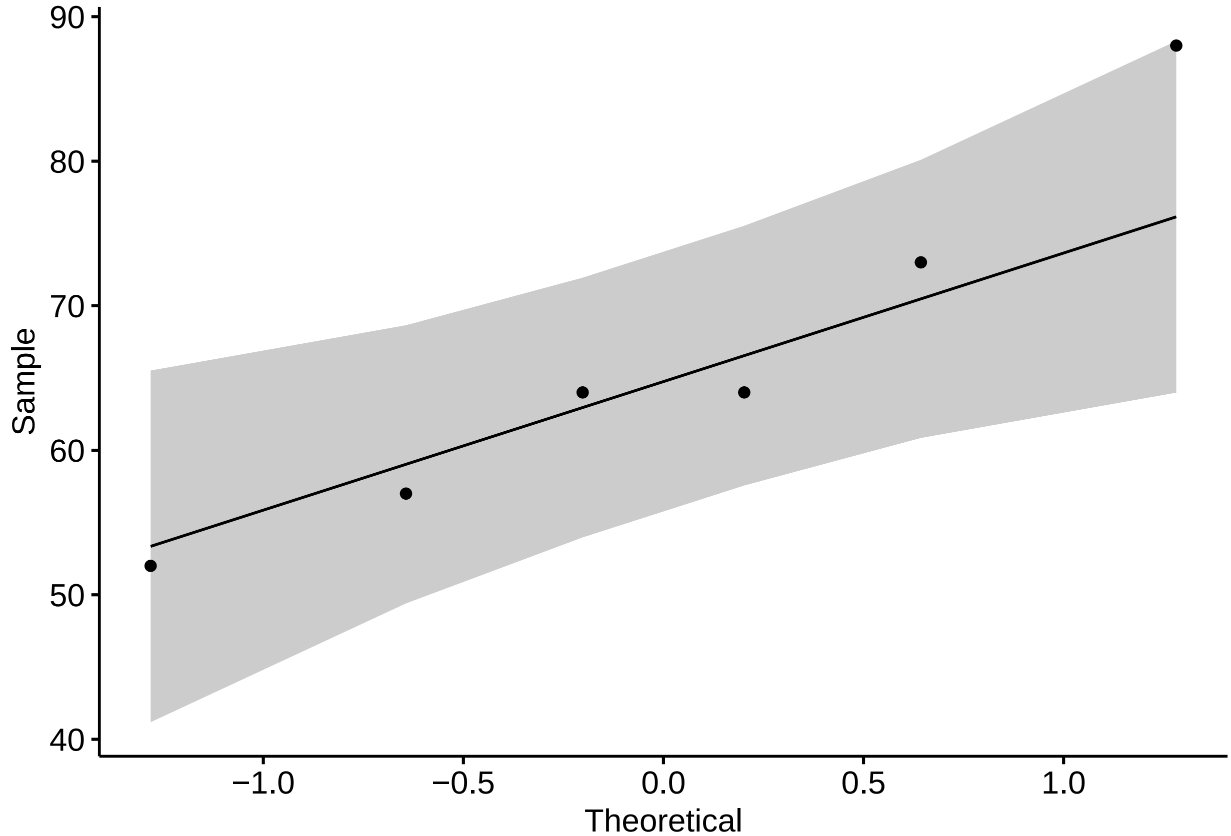
```
##         ap       gf       ww
## 1 75.16667 66.33333 59.33333
## 2 82.33333 77.16667 76.50000
## 3 78.83333 72.33333 75.83333
## 4 60.00000 47.33333 52.66667
```

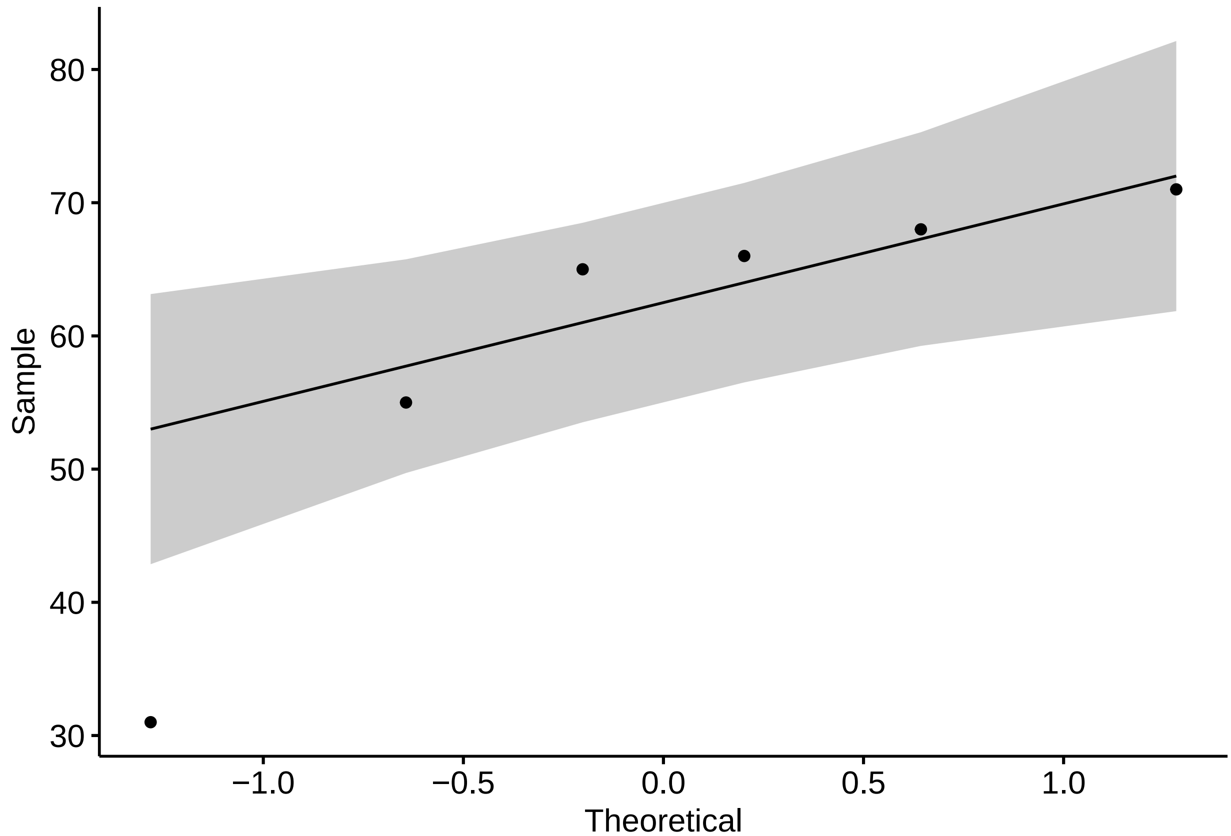
Question 3 - 10 points

Provide a visual assessment for the assumptions of normality and equality of variances for each cell.

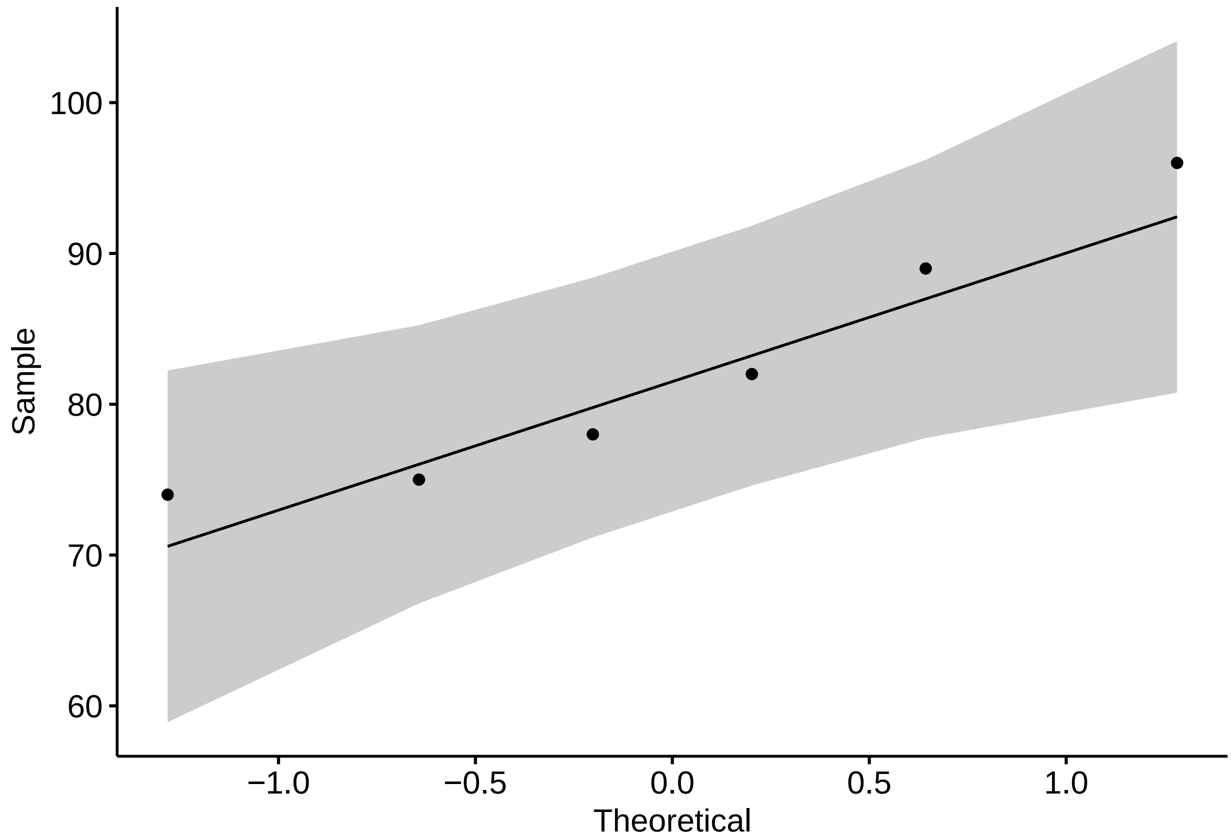
```
# Code for visual assessment of normality for each cell
par(mfrow=c(4, 4))
# Iterate through all groups of fat_type
for (fat in fat.types){
  # Iterate through all groups of flour_type
  for (flour in flour.types){
    # Select the points that are within these two categories
    cell.data = doughnuts.factorial[doughnuts.factorial$fat_type==fat & doughnuts.factorial$flour_type=
    # Plot a QQplot to visually assess normality of the group.
    print(ggqqplot(cell.data))
  }
}
```

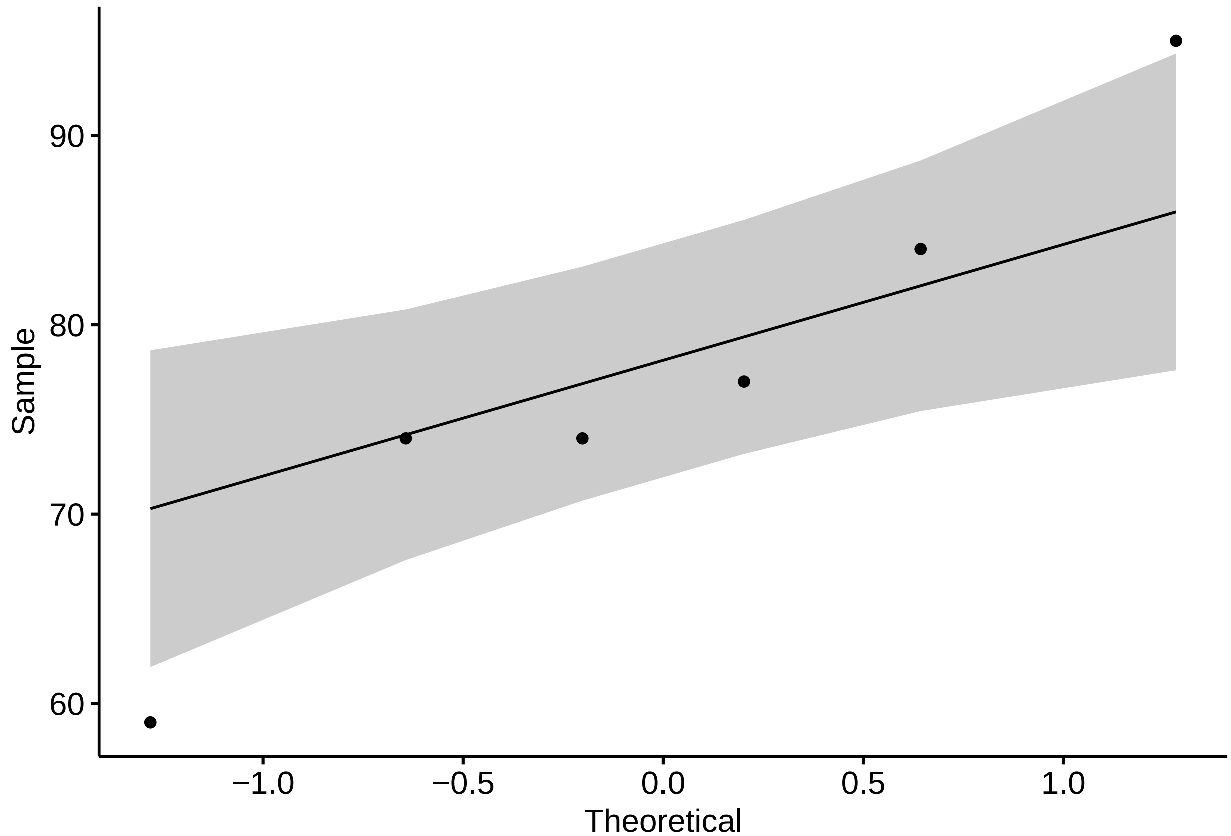

90
80
70
60
Sample
−1.0
−0.5
0.0
0.5
1.0
Theoretical

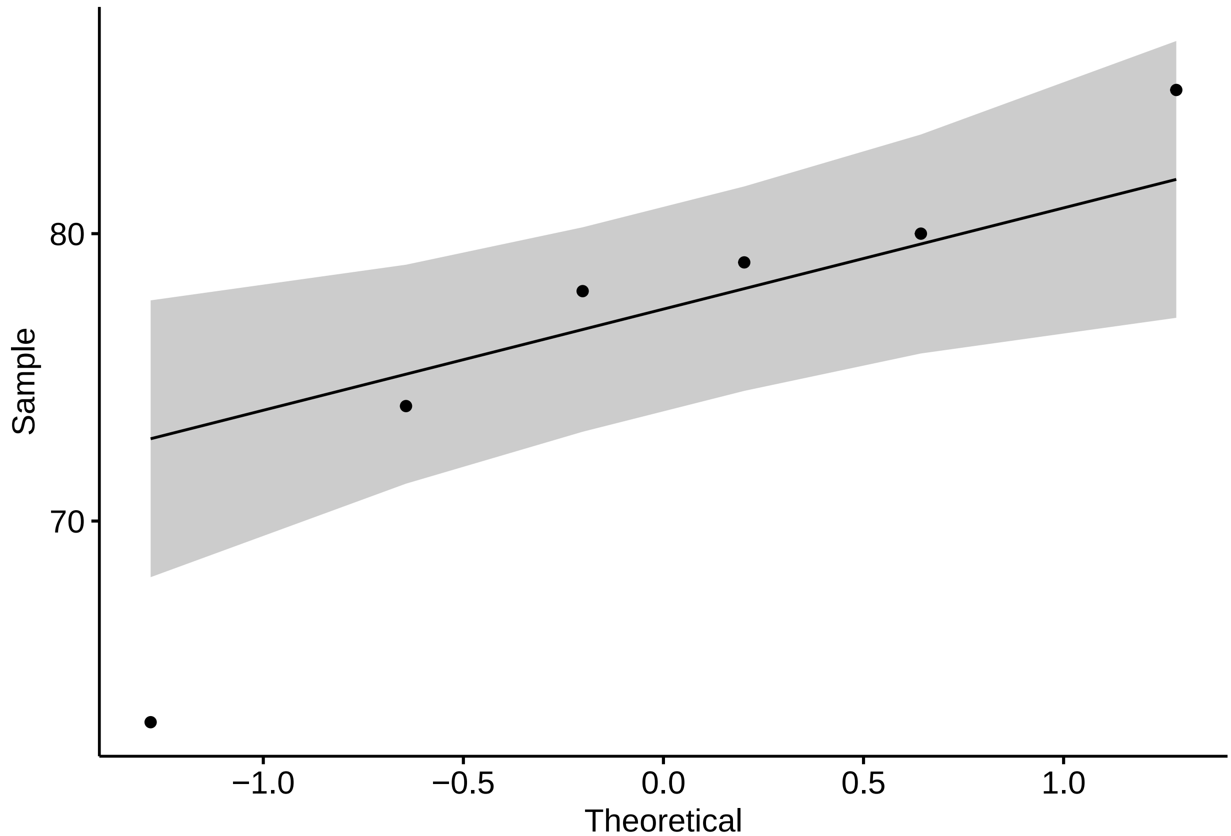

90
80
70
60
50
40
Sample
−1.0
−0.5
0.0
0.5
1.0
Theoretical

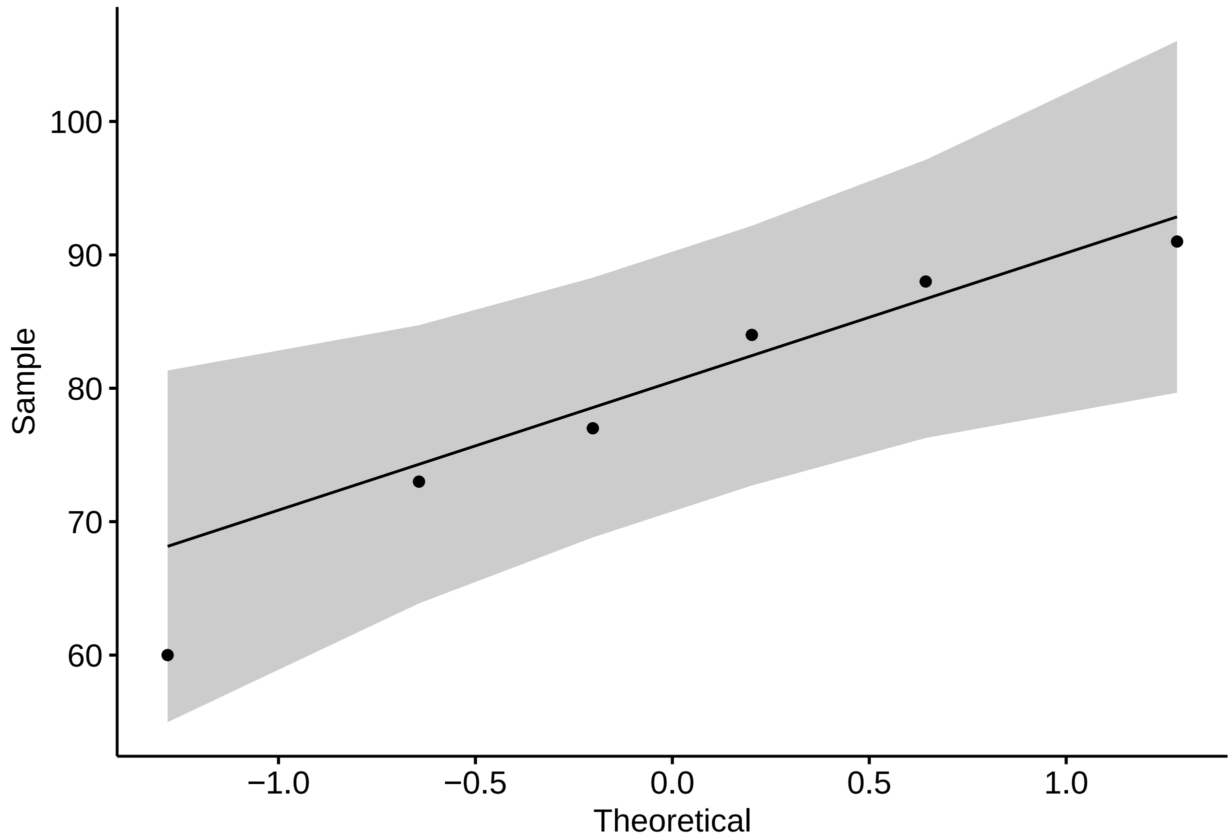


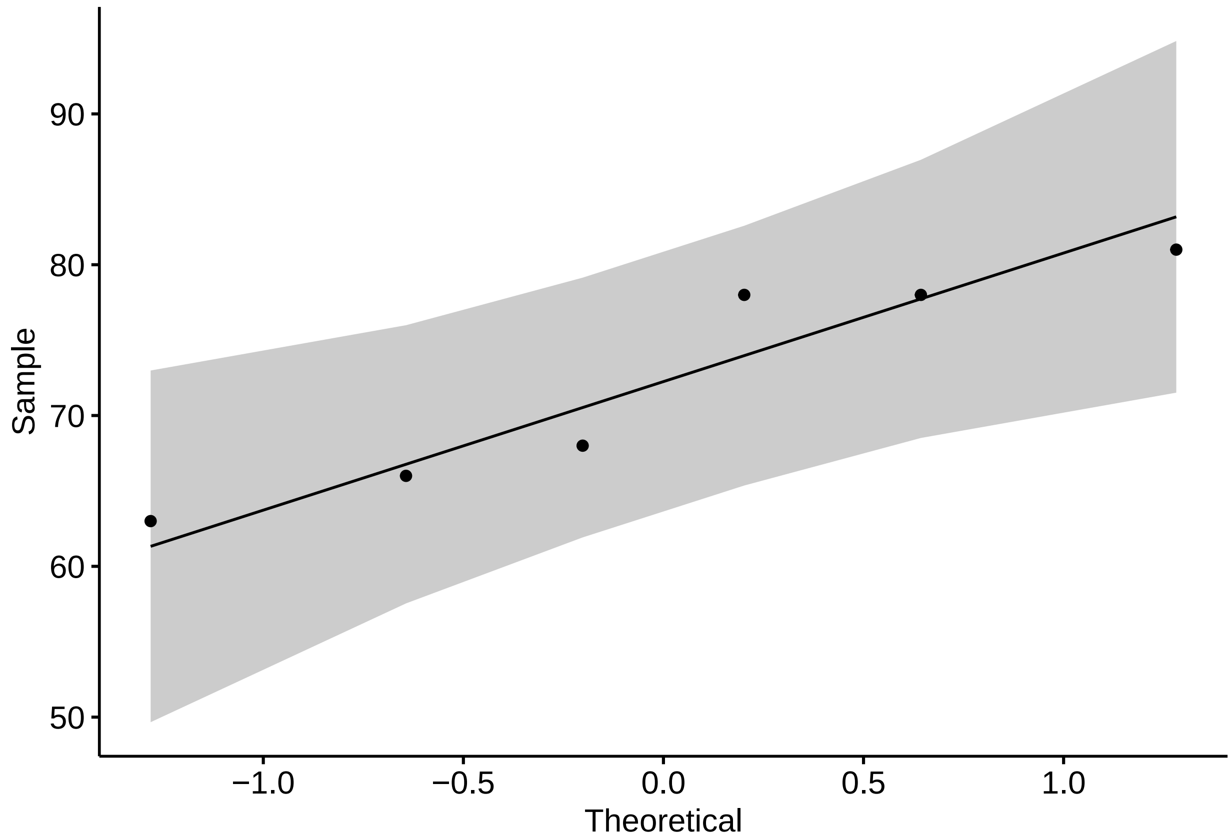
80
70
60
50
40
30
Sample
−1.0
−0.5
0.0
0.5
1.0
Theoretical

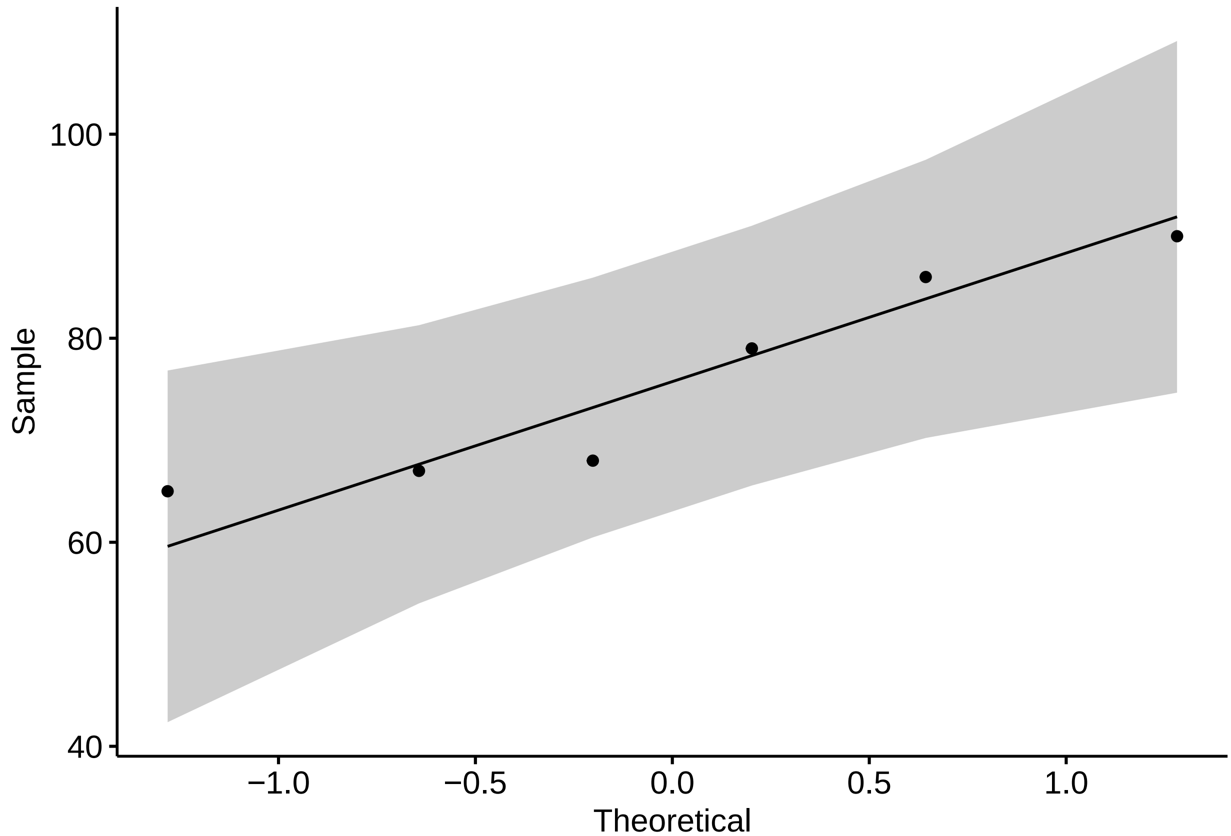

100
90
80
70
60
Sample
−1.0
−0.5
0.0
0.5
1.0
Theoretical


90
80
70
60
Sample
−1.0
−0.5
0.0
0.5
1.0
Theoretical

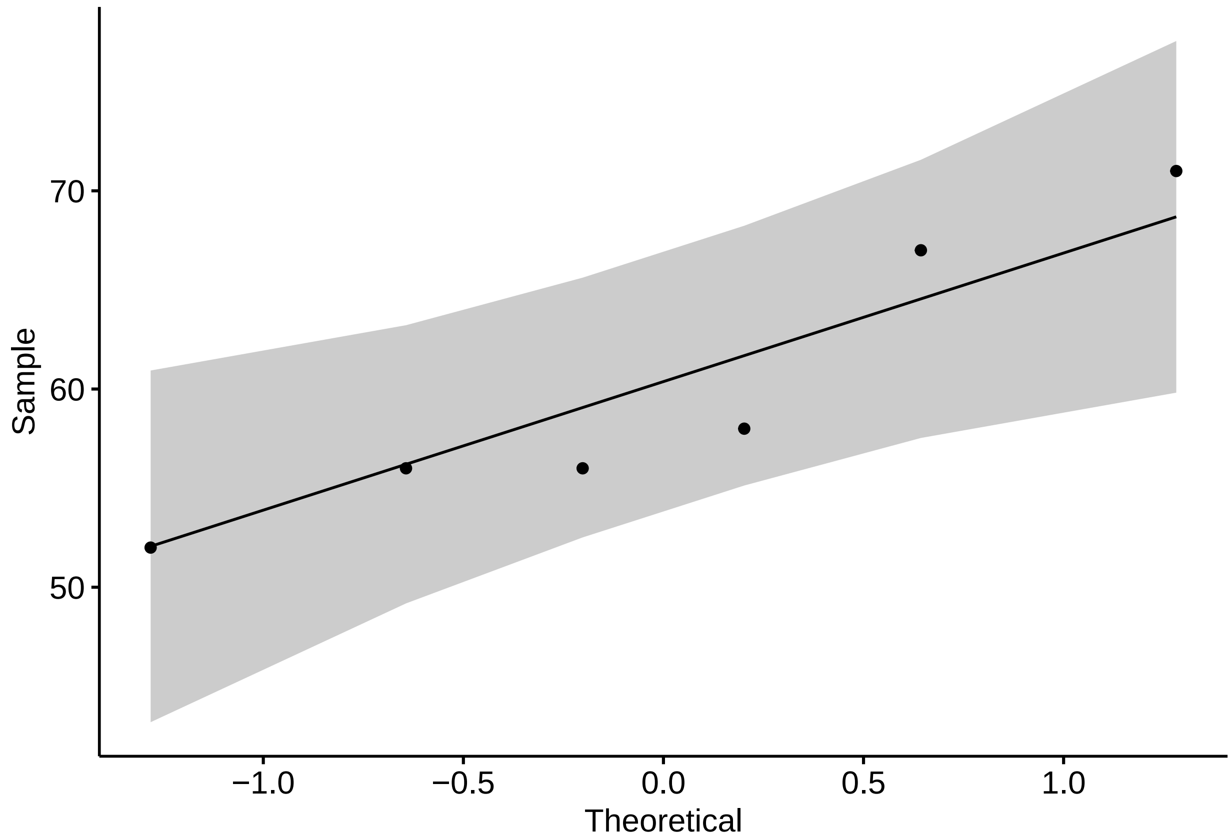

80
70
Sample
−1.0
−0.5
0.0
0.5
1.0
Theoretical


Sample
100
90
80
70
60
−1.0
−0.5
0.0
0.5
1.0
Theoretical

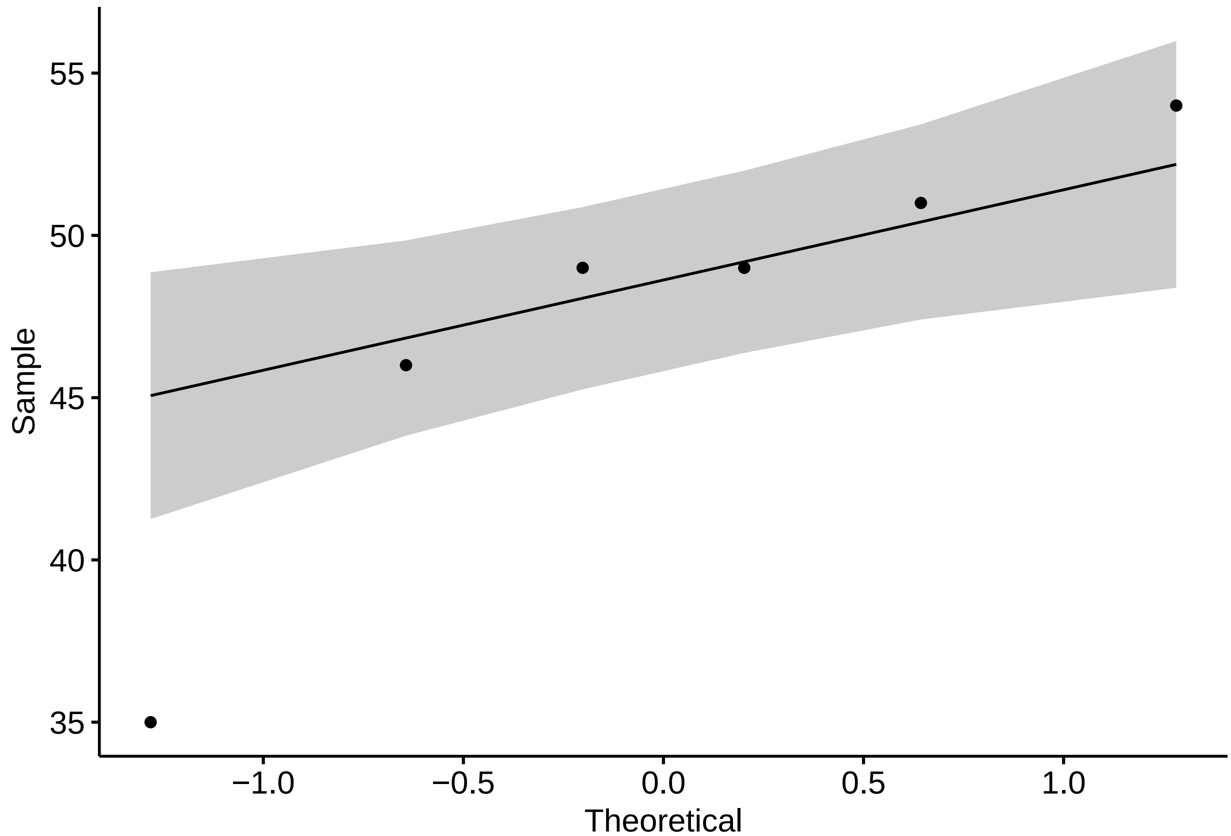

90
80
70
60
50
Sample
−1.0
−0.5
0.0
0.5
1.0
Theoretical

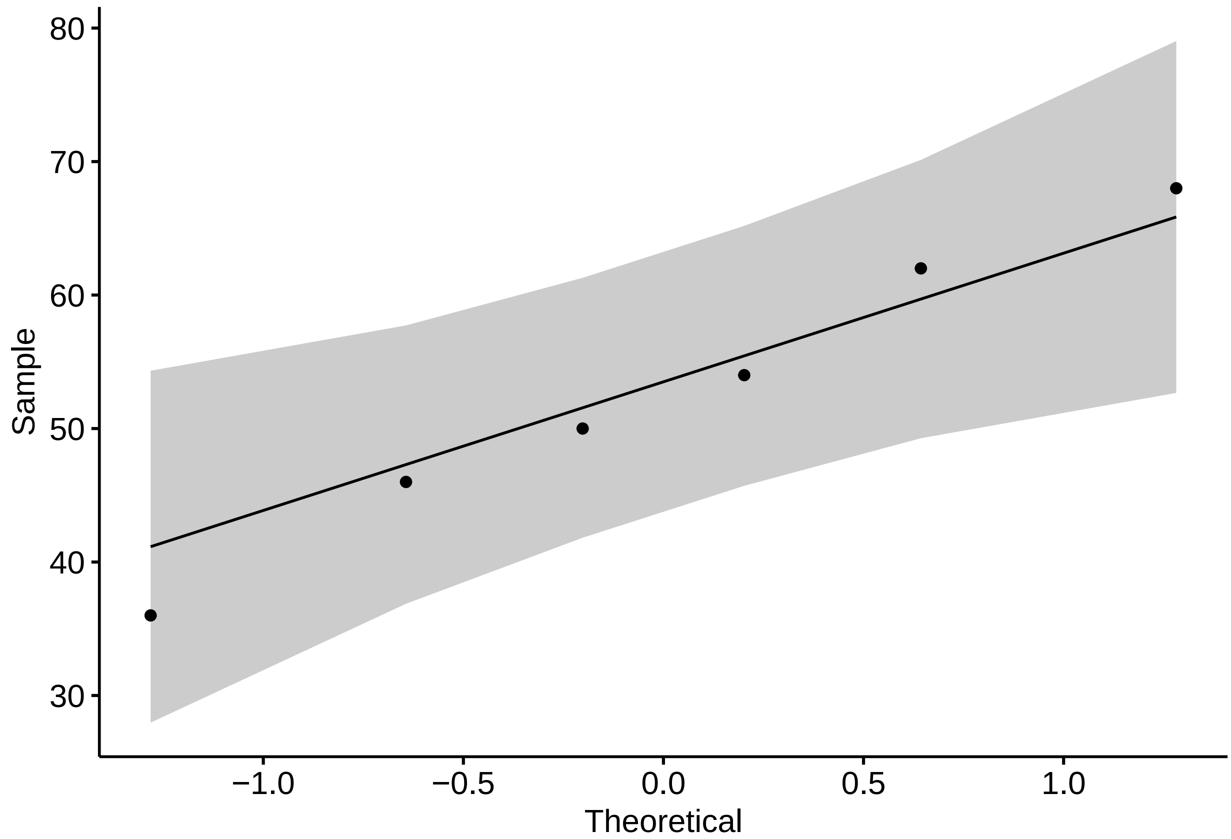


100
80
60
40
Sample
−1.0
−0.5
0.0
0.5
1.0
Theoretical



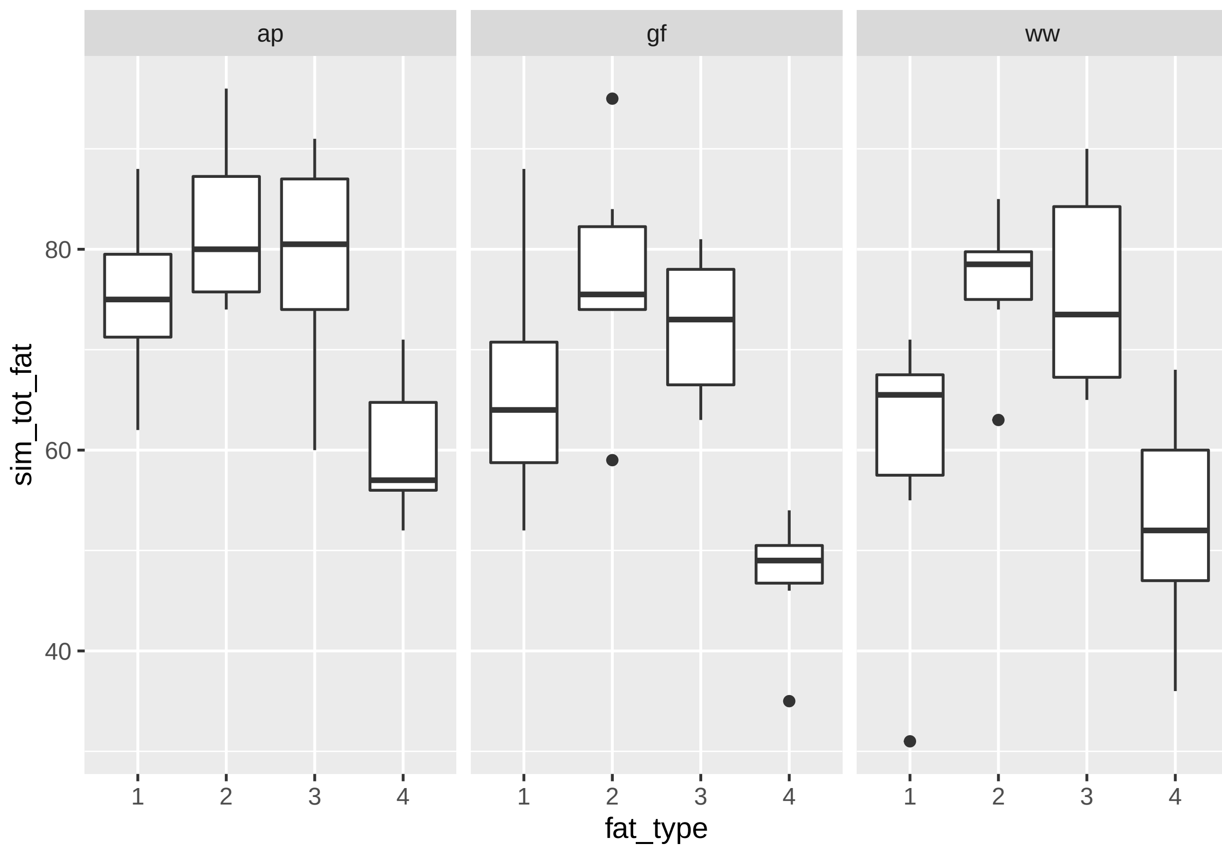
70
60
50
Sample
−1.0
−0.5
0.0
0.5
1.0
Theoretical


55
50
45
40
35
Sample
−1.0
−0.5
0.0
0.5
1.0
Theoretical




```r
# Code for visual assessment of equality of variances for each cell

# Seperates all the points into group combinations, then plots the points side-by-side to see if the sp
ggplot(doughnuts.factorial, aes(x=fat_type, y=sim_tot_fat)) +
    geom_boxplot() +
    facet_grid(. ~ flour_type)
```



Question 4 - 10 points

Conduct a two-way ANOVA with an interaction. Please be sure to display your ANOVA results using the summary() function.

```r
# Code for the two-way ANOVA
anova.model = aov(sim_tot_fat~fat_type*flour_type, data=doughnuts.factorial)
summary(anova.model)
```

```
##                     Df Sum Sq Mean Sq F value   Pr(>F)    
## fat_type             3   6967  2322.5  21.976 1.01e-09 ***
## flour_type           2   1063   531.3   5.028  0.00958 ** 
## fat_type:flour_type  6    427    71.2   0.674  0.67095    
## Residuals           60   6341   105.7                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question 5 - 10 points

Please state the three null hypotheses being tested in your analysis in Question 4 and interpret the result in the context of the research question. After interpreting the interaction, please note if the result of the interaction test changes how you view the results of the tests of main effects.

Fat type: The null hypothesis is that there is no difference between the means of different fat types, when flour type is held constant. Because the p-value is very small ($\sim 1 \times 10^{-9}$), we reject the null and assert that there is a difference between at least one pair of means of fat types. We can see that the interaction test was not significant, so it does not affect this result.

Flour type: The null hypothesis is that there is no difference between the means of different flour types, when fat type is held constant. Because the p-value is small (~ 0.01), we reject the null and can assert that there is a difference between the means of at least one pair of flour types. Similar to the fat type assessment, because the interaction is not significant, this does not affect our interpretation.

Interaction: The null hypothesis is that there is no interaction between fat types and flour types. From the ANOVA results, we see that the p-value of the interaction is large (~ 0.67), so we fail to reject the null and continue to assert that there is not an interaction between fat type and flour type.

Question 6 - 5 points

Much like how t-tests are special cases of ANOVA, ANOVA is a special case of regression. To demonstrate this, below is a regression model that is equivalently-specified to the one-way ANOVA on the Lowe data you conducted as part of Question 4 in Problem Set 1. Please run the code below and also have your results from Question 4 in Problem Set 1 in front of you before answering the two questions.

```
# Loading the data

doughnuts <- read.csv("doughnuts.csv",header=TRUE,sep=",") # You may need to change this line, but don’
doughnuts$fat_type_factor <- as.factor(doughnuts$fat_type)

doughnuts.reg = lm(total_fat ~ fat_type_factor, data=doughnuts) # Runs a simple linear regression.
summary(doughnuts.reg)
```

```
##
## Call:
## lm(formula = total_fat ~ fat_type_factor, data = doughnuts)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -16.00  -7.00   0.00   5.25  23.00
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        72.000      4.101  17.557 1.28e-13 ***
## fat_type_factor2   13.000      5.799   2.242   0.0365 *
## fat_type_factor3    4.000      5.799   0.690   0.4983
## fat_type_factor4  -10.000      5.799  -1.724   0.1001
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.04 on 20 degrees of freedom
## Multiple R-squared:  0.4478, Adjusted R-squared:  0.365
## F-statistic: 5.406 on 3 and 20 DF,  p-value: 0.006876
```

1) There were four types of fats used in the experiment, but why are there only three coefficients in the regression model associated with fat type?

The first fat level became the “control.” In other words, if a sample does not belong to one of the other three levels, then it is automatically placed within the first group without needing an extra coefficient to determine what that group would be. In this way, the other three groups coefficients are those groups' difference from the first group.

2) Identify two aspects of in the regression model output that demonstrate that this model is equivalent to the one-way ANOVA in Question 4 in Problem Set 1. Specifically, look at the Coefficients table and the model statistics shown under the Coefficients table. There are more than two aspects of this output that demonstrate the equivalent of these models, but just identify two. Please write your answer below:

Aspect 1: The first aspect is that the overall F-statistic, degrees of freedom and p-value of the regression model, which tests if the model is significantly different than the null model (which would be a constant), is that same as those values for the ANOVA results. In both cases, the F-statistic is ~ 5.406 on 3 and 20 degrees of freedom, resulting in the p-value of ~ 0.00688.

Aspect 2: The second equivalence is how the regression coefficients relate to the cell means of the ANOVA model. The regression's intercept is 72, which is the same value as the mean when fat type is 1. Then the remaining coefficients of the regression are the differences between the mean for that level of the ANOVA model and the first fat type. For example, the second fat type has mean 85, so the difference between fat type 2 and fat type 1 is $85 - 72 = 13$, which is the same as the coefficient for fat type 2 in the regression.