1.The mean of Cholesterol, grouped by PriorSmoke is 242.5 while the standard deviation is 132 rounded.

```
             Df  Sum Sq Mean Sq F value Pr(>F)
PriorSmoke    1   77258   77258   4.484  0.035 *
Residuals   313 5393183   17231
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

H0 = No difference in Cholesterol between PriorSmoke groups

HA = Difference in Cholesterol between PriorSmoke groups

The p-value in the ANOVA table (Pr(>F) = 0.035) provides the probability of observing the data or something more extreme if the null hypothesis were true. Since the p-value is less than the commonly used significance level of 0.05, we reject the null hypothesis in favor of the alternative hypothesis, concluding that there is a statistically significant difference in mean Cholesterol levels between the groups of "PriorSmoke".

2. Since the variables were labeled 1,2, and 3, and I do not have access to the data dictionary, I chose to leave out the dummy variable PriorSmoke_3. This is now the basis of interpretation.

Prediction equation:

Cholesterol = $\beta_0 + \beta_1 \times \text{PriorSmoke}_1 + \beta_2 \times \text{PriorSmoke}_2 + \varepsilon$

Where β_0 is the intercept, which is the mean cholesterol level for the reference group (the omitted PriorSmoke_3), and the other two β are the missing PriorSmoke_1 and PriorSmoke_2.

```
Call:
lm(formula = Cholesterol ~ PriorSmoke_1 + PriorSmoke_2, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-204.12  -90.02  -32.79   61.37  672.31

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    272.53      20.05  13.593   <2e-16 ***
PriorSmoke_1   -44.14      22.63  -1.951    0.052 .
PriorSmoke_2   -22.11      23.50  -0.941    0.348
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 131.5 on 312 degrees of freedom
Multiple R-squared:  0.01412,   Adjusted R-squared:  0.007803
F-statistic: 2.235 on 2 and 312 DF,  p-value: 0.1087
```

With the intercept at 0, Cholesterol will be at 272.53 given the other two are at 0.

PriorSmoke_1's coefficient of -44.14 means a decrease in average cholesterol level. This could mean this number represents people who do not have prior smoking habits.

PriorSmoke_2's coefficient is also -22.11, which could mean lightly smoking.

PriorSmoke_1 is barely outside of the hypothesis testing range with a p-value of 0.052, and PriorSmoke_2 is at 0.348 meaning we must accept the null-hypothesis that these are due to random chance instead of being statistically significant.

The F-statistic is also quite low at 2.235, with a p-value of 0.187, so we must also accept the null hypothesis.

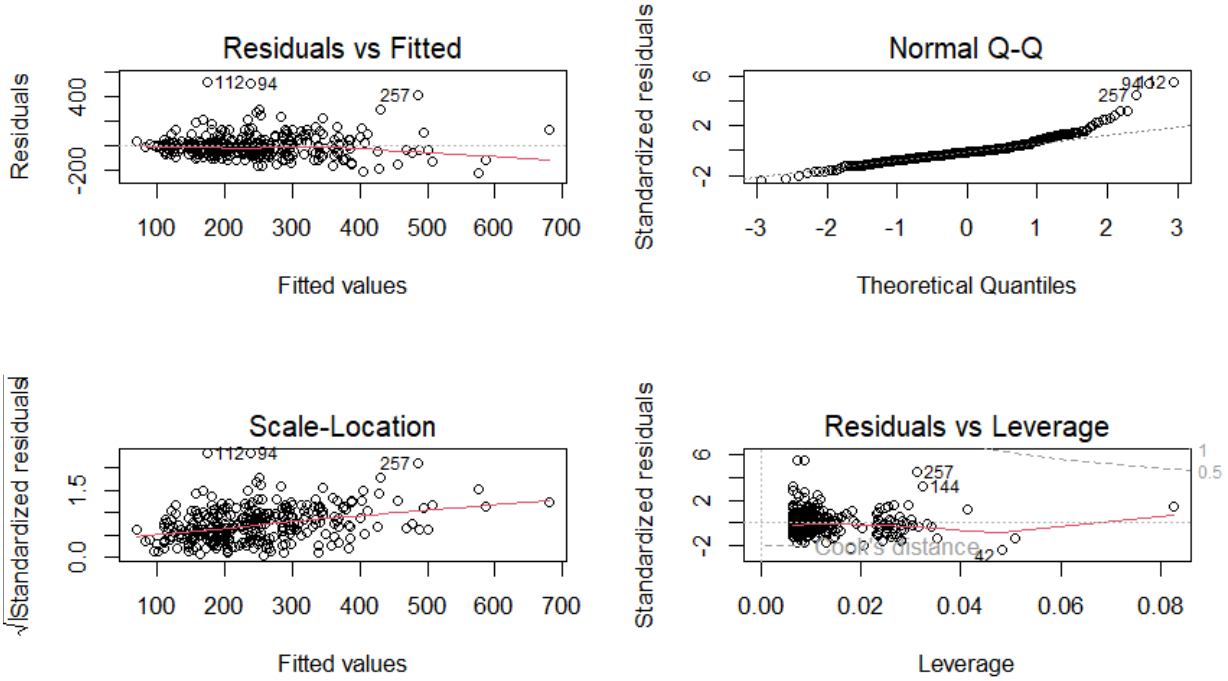
Comparing both models, it's evident that the second model is inferior to the first. We can see that the added coefficients are not statistically significant, albeit one is close to rejecting the null hypothesis. If we used a different coefficient as the basis of interpretation, it may be more impactful.

3.

Cholesterol=$\beta 0+\beta 1\times$PriorSmoke1+$\beta 2\times$PriorSmoke2+$\beta 3\times$Fat+ϵ

```
Call:
lm(formula = Cholesterol ~ PriorSmoke_1 + PriorSmoke_2 + Fat,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-214.06  -53.03  -12.01   33.24  514.58

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   39.5759    19.4585   2.034   0.0428 *
PriorSmoke_1 -10.6358    16.1763  -0.657   0.5113
PriorSmoke_2 -12.7500    16.6906  -0.764   0.4455
Fat            2.7630     0.1574  17.556   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 93.33 on 311 degrees of freedom
Multiple R-squared:  0.5048,    Adjusted R-squared:  0.5001
F-statistic: 105.7 on 3 and 311 DF,  p-value: < 2.2e-16
```

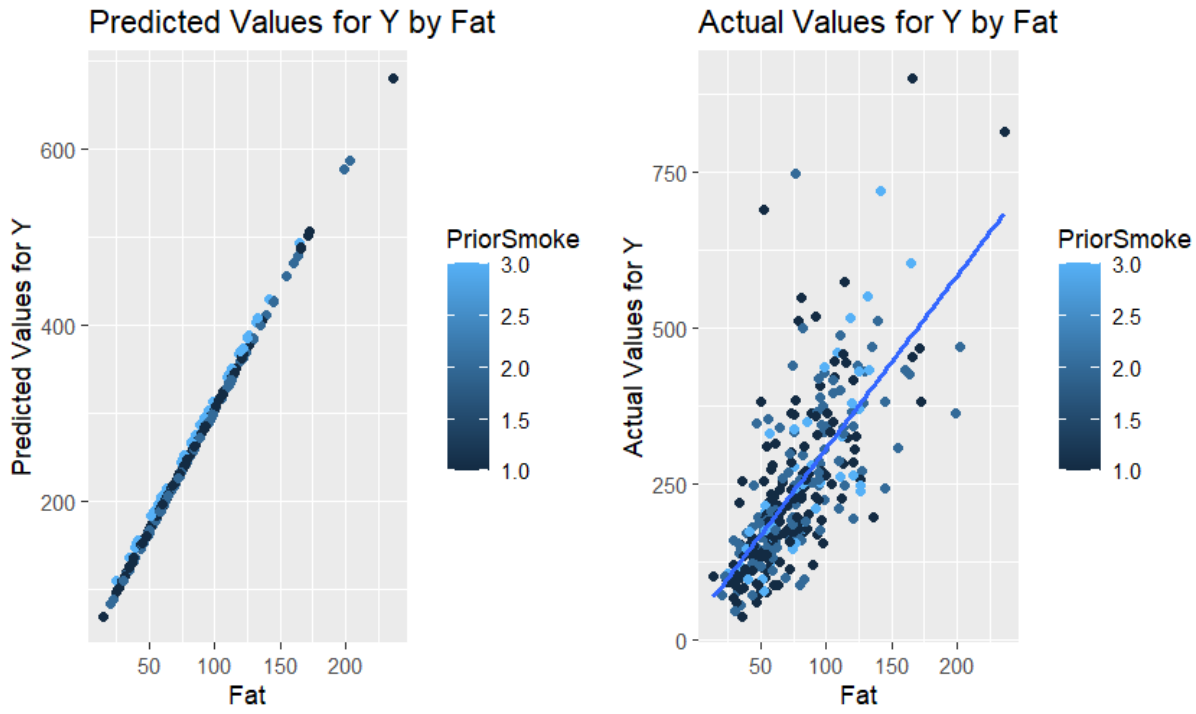
Much like our previous model, we must accept the null hypothesis for PriorSmoke 1 and PriorSmoke 2. Their p-values are well above the 0.05 criteria, in fact, they have gotten worse with the addition of Fat. However, Fat is incredibly significant at 2e-16 and therefor we must accept the alternative hypothesis. The F-statistic is much higher with the addition of Fat which adds credence to the addition of Fat being able to predict cholesterol levels.



Examining our ANCOVA

Our residuals vs fitted chart shows that the relationship may not be perfectly linear or perhaps there might be some non-constant variance. For the most part, our data points fit the Normal Q-Q graph, but taper upwards at the end, which could indicate some possible problems with normality. The spread in the Scale-Location is also a bit concerning as the data moves to the right. A few points also stick out as leverage in the Residuals vs Leverage chart which could be problematic such as point 257 and point 144.

4.



If I understand this correctly, the predicted values are plotted on the line that is shown on the right for actual values. I would not say that it plots the data all too well as the variance disperses the higher the fat increases. Although the line of fit is not terrible, I would imagine a model with more information would have a better fit.

5.

$$\text{Cholesterol}=\beta 0+\beta 1\times\text{PriorSmoke1}+\beta 2\times\text{PriorSmoke2}+\beta 3\times\text{Fat}+\beta 4\times\text{PriorSmoke1}\times\text{Fat}+\beta 5\times\text{PriorSmoke2}\times\text{Fat}+\epsilon$$

```
Call:
lm(formula = Cholesterol ~ PriorSmoke_1 + PriorSmoke_2 + Fat +
    PriorSmoke_1_Fat + PriorSmoke_2_Fat, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-222.37  -56.18   -9.74   35.48  518.67

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      -19.1791    38.0381  -0.504   0.6145
PriorSmoke_1      32.8823    42.2005   0.779   0.4365
PriorSmoke_2      84.2709    43.7383   1.927   0.0549 .
Fat                3.4598     0.4190   8.258 4.37e-15 ***
PriorSmoke_1_Fat  -0.4858     0.4787  -1.015   0.3110
PriorSmoke_2_Fat  -1.1697     0.4851  -2.411   0.0165 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 92.54 on 309 degrees of freedom
Multiple R-squared:  0.5163,    Adjusted R-squared:  0.5085
F-statistic: 65.97 on 5 and 309 DF,  p-value: < 2.2e-16
```

The F-statistic is lower at 65.97, the R^2 value is 0.5163 that means 51.63% of the variability in the dependent variable is explained by the model. The model has a decent fit.

Fat and PriorSmoke_2_Fat are the only statistically significant and reject the null hypothesis.

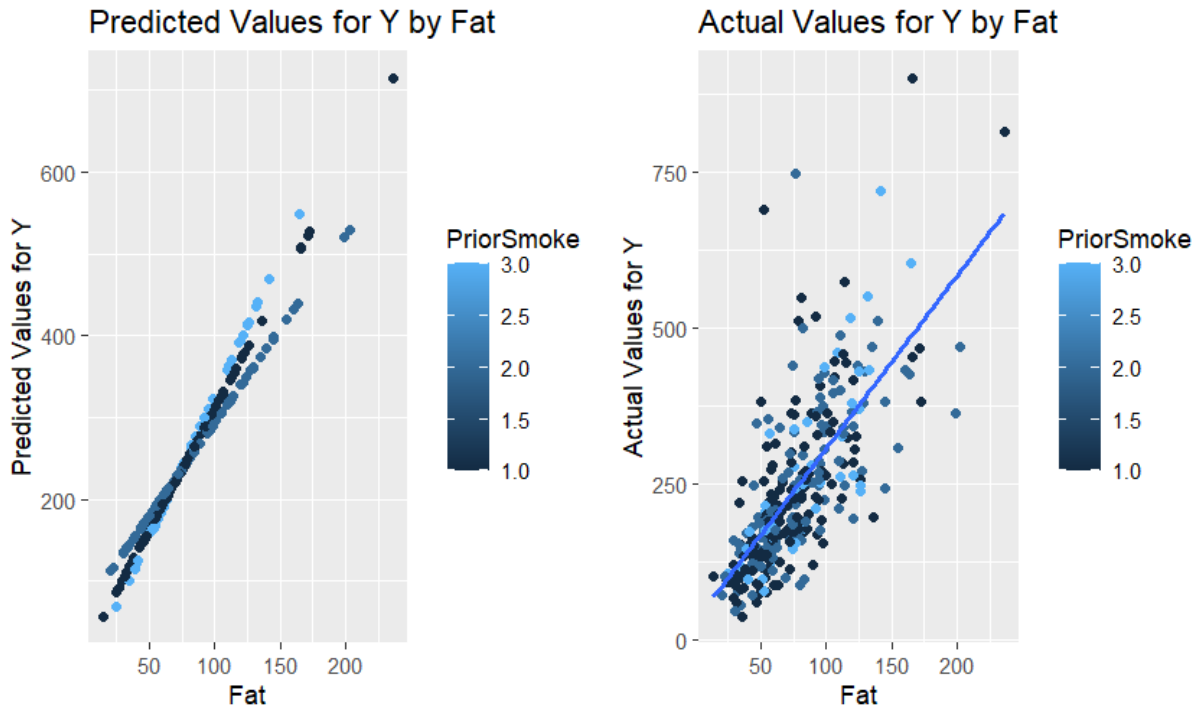
The residuals vs fitted shows a slight pattern like from before.

The normal Q-Q plot mimics the one earlier with the tail moving further along the x-axis, or large positive values.

The Scale-Location suggests increasing variability of residuals as the fitted values increase, which points towards heteroscedasticity.

Finally, we have points 257, 144, and 42 that could influence the model.

6.



Again, I'm not sure whether the predicted values are supposed to be in this format, but based on the left chart for predicted values, it appears that 3 has the steepest slope and the lowest slope is 2 with 1 splitting the middle. Compared to the other scatterplot we have outliers for 3, 2, and 1 with more outliers around the 750 to 850 range.

- Both plots show a positive relationship between Fat and Y.
- The predicted plot reflects the regression model's sfitted values, which show less variance compared to the actual values.
- The actual values plot exhibits more natural variability and potential outliers

7. Model 2 nests model 3 because model 3 is a more developed model and contains all the information from model 2 in model 3.

Model 2, the reduced model, predicts cholesterol using PriorSmoke_1 and PriorSmoke_2 and Fat without interaction terms. This model assumes the effect of Fat on cholesterol is the same regardless of the PriorSmoke status.

Model 3, the full model, includes the same predictors as model 2 plus interaction terms between PriorSmoke 1 and 2 and the multiplied coefficients with Fat. These interaction terms allow the slope of Fat to very by PriorSmoke, testing for unequal slopes.

Null Hypothesis

H0: The additional coefficients in the full model do not improve the fit of the model compared to the reduced model.

HA: The full model does improve the fit of the model compared to the reduced model.

```
Analysis of Variance Table

Model 1: Cholesterol ~ PriorSmoke_1 + PriorSmoke_2 + Fat
Model 2: Cholesterol ~ PriorSmoke_1 + PriorSmoke_2 + Fat + PriorSmoke_1_Fat +
    PriorSmoke_2_Fat
  Res.Df     RSS Df Sum of Sq     F  Pr(>F)
1    311 2708756
2    309 2645939  2     62817 3.668 0.02665 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

The F-statistic is 3.668 and the p-value is approximately 0.027 so we reject the null hypothesis. This means that the additional interaction terms in Model 3, or the full model, provide a statistically significant improvement in the model's fit over Model 2, the reduced model in predicting cholesterol levels.

Based on the results we can say that there are unequal slopes in this situation.

8. I grouped the information based off of alcohol, leaving out the None usage.

meanCholesterol	sdCholesterol
242.4606	131.9916

```
Call:
lm(formula = Cholesterol ~ alcohol_usage_Great + alcohol_usage_Moderate,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-246.72  -89.94  -33.42   69.33  662.55

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            238.150     12.515  19.029   <2e-16 ***
alcohol_usage_Great     46.266     29.683   1.559    0.120
alcohol_usage_Moderate   1.374     15.913   0.086    0.931
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 131.9 on 312 degrees of freedom
Multiple R-squared:  0.008383,  Adjusted R-squared:  0.002027
F-statistic: 1.319 on 2 and 312 DF,  p-value: 0.2689

>
```

The first iteration of the model shows that alcohol usage Great and Moderate both fail our p-value test therefore we must accept the null hypothesis which states that either one has a statistically significant effect on the model. The F-statistic is also quite low at only 1.319.

```
Call:
lm(formula = Cholesterol ~ alcohol_usage_Great + alcohol_usage_Moderate +
    Fat, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-217.38  -52.42  -10.67   32.88  516.68

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           33.3497    14.5909   2.286   0.0229 *
alcohol_usage_Great  -14.0311    21.2819  -0.659   0.5102
alcohol_usage_Moderate  -7.7098    11.2733  -0.684   0.4946
Fat                    2.7856     0.1577  17.663   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 93.31 on 311 degrees of freedom
Multiple R-squared:  0.505,     Adjusted R-squared:  0.5002
F-statistic: 105.7 on 3 and 311 DF,  p-value: < 2.2e-16

>
```

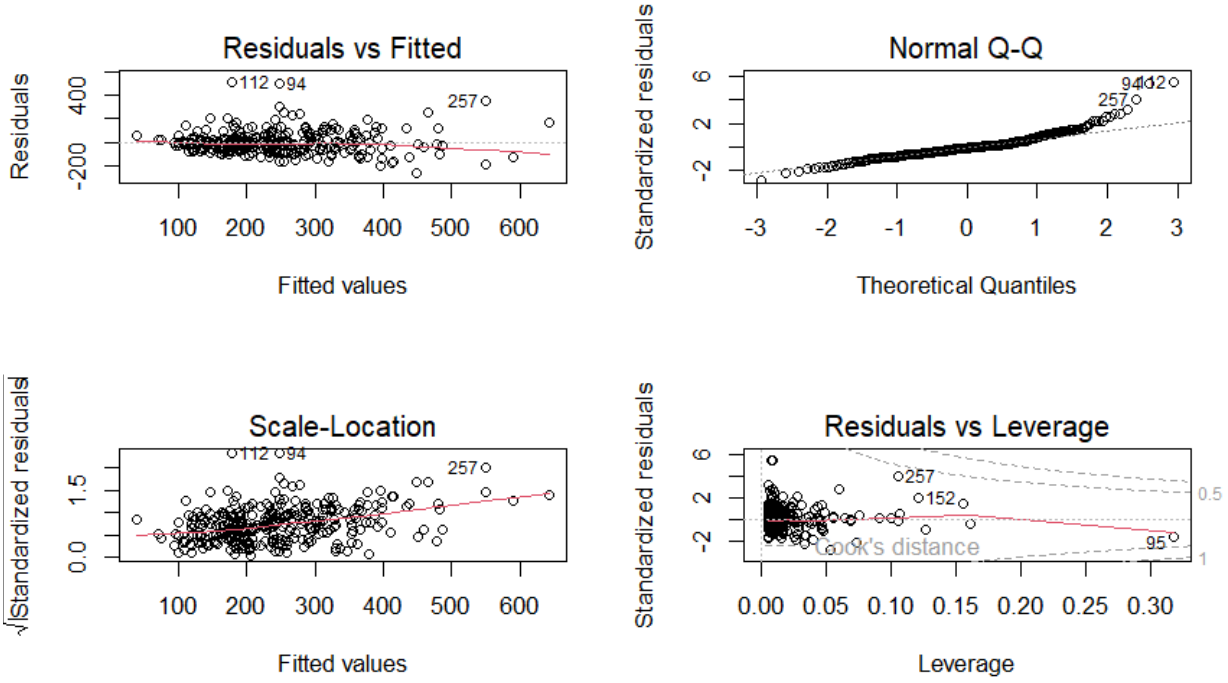
Again, like in our previous models, fat is statistically significant with a value of 2e-16 and a much higher F-statistic in our second model.

The first alcohol model is nested within this one, so although adding more coefficients raises the R^2 values, it would not account for shut a large jump up.

```
Call:
lm(formula = Cholesterol ~ alcohol_usage_Great + alcohol_usage_Moderate +
    Fat + alcohol_usage_Great_Fat + alcohol_usage_Moderate_Fat,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-253.33  -54.01   -8.59   34.12  510.73

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                -10.1124    24.6033  -0.411   0.6813
alcohol_usage_Great         24.1245    53.1620   0.454   0.6503
alcohol_usage_Moderate      54.7998    29.7383   1.843   0.0663 .
Fat                          3.3768     0.3124  10.808   <2e-16 ***
alcohol_usage_Great_Fat     -0.5354     0.5506  -0.972   0.3316
alcohol_usage_Moderate_Fat  -0.8392     0.3699  -2.269   0.0240 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 92.84 on 309 degrees of freedom
Multiple R-squared:  0.5131,    Adjusted R-squared:  0.5052
F-statistic: 65.13 on 5 and 309 DF,  p-value: < 2.2e-16
```

We reject the null hypothesis for Fat again, but also for the newly create interaction terms for Moderate Alcohol Usage created by multiplying alcohol usage with fat.

This is interesting because alone we had to accept the null hypothesis in the previous model, but when adding the new coefficients, it makes alcohol usage moderate much more relevant.

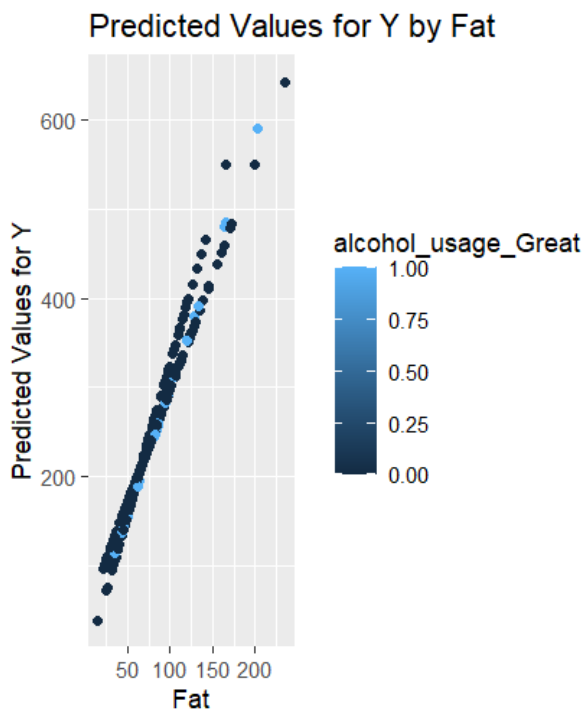


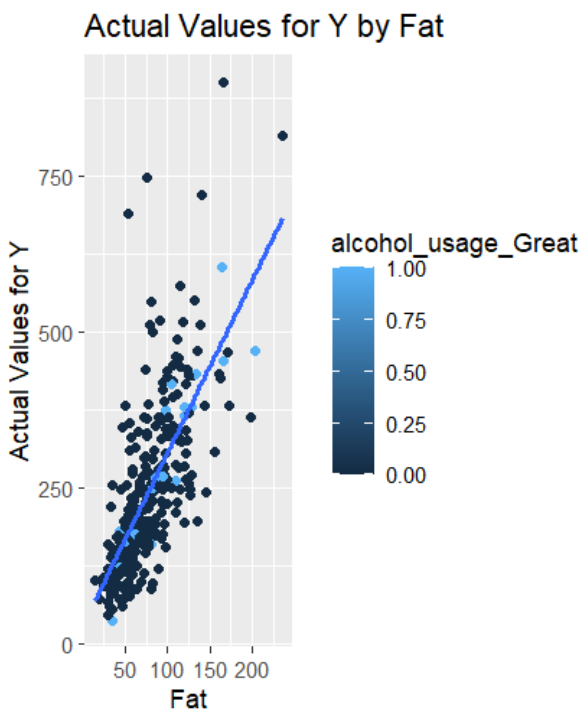
We see a slight tunnel shape in the residuals vs fitted plot which may indicate heteroscedasticity.

The normal q-q plot which means that the data may not be normally distributed.

The scale-location shows a slight increase in spread with higher fitted values, which could be heteroscedasticity.

Finally, we see the same points pop-up again in the previous models of 257, 152, and 95.





Compared to the previous models concerning cholesterol, we see that the line fits much better with the added alcohol with most data points belonging to moderate usage alcohol.

```
Analysis of Variance Table

Model 1: Cholesterol ~ alcohol_usage_Great + alcohol_usage_Moderate +
    Fat
Model 2: Cholesterol ~ alcohol_usage_Great + alcohol_usage_Moderate +
    Fat + alcohol_usage_Great_Fat + alcohol_usage_Moderate_Fat
  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1    311 2708082
2    309 2663564  2     44518 2.5822 0.07724 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Finally comparing the alcohol models together we can accept the null hypothesis, that the second model does not improve upon the first.

9.

Conclusion

I’ve learned that different combinations of coefficients can greatly change the underlying model. A few points that I still don’t fully understand is the need to plot the predicted points against the actual points for comparison since the predicted points fall on the slope each time, which is information we already know.

The initial models pointed towards a significant difference in cholesterol levels among different smoking groups, which indicated a slight effect. However, in delving deeper with more complex models, the direct impact of smoking alone was not statistically significant. However, adding fat to it increased the model, including when multiplying them together.

Over both models, fat played an important role in determining someone's cholesterol levels.

For future attempts, I would apply a few methods to smooth out the models. The outliers that kept popping up may have a strong effect on the underlying model, so creating two models. One with the outliers, and the other without could help as well as normalizing the data in advance since the charts showed signs of the data not being in a normal distribution.

```
#Load XLS
library(readxl)
library(ggplot2) # For data visualization
library(dplyr) # For data manipulation
library(GGally) # For pairwise plots
library(tidyr) # For handling missing values visualization
library(plyr)
library(fastDummies)
library(patchwork)
#install.packages('fastDummies')
#Load data.xls in current folder
data <- read_excel("NutritionStudy.xls")



#Recode Smoke to 1 if yes, 0 if no
data$Smoke <- ifelse(data$Smoke == "Yes", 1, 0)

#Gender male = 0 female = 1
data$Gender <- ifelse(data$Gender == "Male", 1, 0)

#VitamanUse get all unique variables
unique(data$VitaminUse)
#"Regular"    "Occasional" "No"

#Create Dummies for VitaminUse
data <- dummy_cols(data, select_columns = "VitaminUse")
#Create Dummies for PriorSmoke
data <- dummy_cols(data, select_columns = "PriorSmoke")
```

```
#Drop the original columns
#data <- data[, !(names(data) %in% c("VitaminUse", "PriorSmoke"))]

#If Alcohol = 0 None, if 0-10 moderate, if >= 10 great
data$alcohol_usage <- ifelse(data$Alcohol == 0, "None", ifelse(data$Alcohol <= 10, "Moderate", "Great"))

#Now dummies
data <- dummy_cols(data, select_columns = "alcohol_usage")

#Drop Alcohol
#data <- data[, !(names(data) %in% c("Alcohol", "alcohol_usage"))]

#Descriptive statistics by PriorSmoke

# Descriptive statistics by PRIORSMOKE, mean Cholesterol
priorSmoke_stats <- data %>%
  group_by(PriorSmoke)

#Summarizing the meand and SD of Cholesterol
priorSmoke_stats <- priorSmoke_stats %>%
  summarise(meanCholesterol = mean(Cholesterol, na.rm = TRUE),
            sdCholesterol = sd(Cholesterol, na.rm = TRUE))
#Anova test
anova_test <- aov(Cholesterol ~ PriorSmoke, data = data)
summary(anova_test)

#2
#Predicting Cholesterol as our Y variable using PriorSmoke_1, PriorSmoke_2, PriorSmoke_3
```

```
model1 <- lm(Cholesterol ~ PriorSmoke_1 + PriorSmoke_2, data = data)
summary(model1)

#3

model2 <- lm(Cholesterol ~ PriorSmoke_1 + PriorSmoke_2 + Fat, data = data)
summary(model2)

#Diagnostic plots
par(mfrow=c(2,2))
plot(model2)

#Influence measures
influence_measures <- influence.measures(model2)

#4



#Predicted values for Y
data$predicted_values <- predict(model2)



par(mfrow=c(1,2))
#Scatterplot
scatterplot_ <- ggplot(data, aes(x = Fat, y = predicted_values, color = PriorSmoke)) +
  geom_point() +
  labs(title = "Predicted Values for Y by Fat", x = "Fat", y = "Predicted Values for Y")
```

```
#Add line
lineplot_ <- ggplot(data, aes(x = Fat, y = Cholesterol, color = PriorSmoke)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Actual Values for Y by Fat", x = "Fat", y = "Actual Values for Y")

#Adding the two together
scatterplot_ + lineplot_

#5

#Create new product variables
data$PriorSmoke_1_Fat <- data$PriorSmoke_1 * data$Fat
data$PriorSmoke_2_Fat <- data$PriorSmoke_2 * data$Fat

#Unequal slopes model
model3 <- lm(Cholesterol ~ PriorSmoke_1 + PriorSmoke_2 + Fat + PriorSmoke_1_Fat +
PriorSmoke_2_Fat, data = data)

#Fit
summary(model3)

#diagnostic plots
par(mfrow=c(2,2))
plot(model3)

#6
#Use Model 3 to obtain predicted values.   Plot the predicted values
#for CHOLESTEROL (Y) by FAT(X).  Discuss what you see in this graph.
```

```
data_predicted_values <- predict(model3)
data$predicted_values_m3 <- data_predicted_values

#Plot
scatterplot_m3 <- ggplot(data, aes(x = Fat, y = predicted_values_m3, color = PriorSmoke)) +
  geom_point() +
  labs(title = "Predicted Values for Y by Fat", x = "Fat", y = "Predicted Values for Y")

#Add line
lineplot_m3 <- ggplot(data, aes(x = Fat, y = Cholesterol, color = PriorSmoke)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Actual Values for Y by Fat", x = "Fat", y = "Actual Values for Y")

scatterplot_m3
lineplot_m3

scatterplot_m3 + lineplot_m3

#7

#Nested F-test

nested_f_test <- anova(model2, model3)
print(nested_f_test)

#8
#Lets do the same thing with alcohol usage
```

```
alcohol_usage_stats <- data %>%
  group_by(alcohol_usage_Great, alcohol_usage_Moderate) %>%
  summarise(meanCholesterol = mean(Cholesterol, na.rm = TRUE),
            sdCholesterol = sd(Cholesterol, na.rm = TRUE))

# Basic Model - Predicting Cholesterol using alcohol usage levels
model_alcohol1 <- lm(Cholesterol ~ alcohol_usage_Great + alcohol_usage_Moderate, data = data)
summary(model_alcohol1)

# Extended Model - Adding another variable (e.g., Fat) to the model
model_alcohol2 <- lm(Cholesterol ~ alcohol_usage_Great + alcohol_usage_Moderate + Fat, data = data)
summary(model_alcohol2)



#Interaction terms
data$alcohol_usage_Great_Fat <- data$alcohol_usage_Great * data$Fat
data$alcohol_usage_Moderate_Fat <- data$alcohol_usage_Moderate * data$Fat

model_alcohol3 <- lm(Cholesterol ~ alcohol_usage_Great + alcohol_usage_Moderate + Fat +
alcohol_usage_Great_Fat + alcohol_usage_Moderate_Fat, data = data)
summary(model_alcohol3)

# Diagnostic plots for the extended model with interactions
par(mfrow=c(2,2))
plot(model_alcohol3)

#predicted values
```

```
data$predicted_values_alcohol <- predict(model_alcohol3)


#plot predicted vs actual
scatterplot_alcohol <- ggplot(data, aes(x = Fat, y = predicted_values_alcohol, color =
alcohol_usage_Great)) +
  geom_point() +
  labs(title = "Predicted Values for Y by Fat", x = "Fat", y = "Predicted Values for Y")


#actual
lineplot_alcohol <- ggplot(data, aes(x = Fat, y = Cholesterol, color = alcohol_usage_Great)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Actual Values for Y by Fat", x = "Fat", y = "Actual Values for Y")


scatterplot_alcohol + lineplot_alcohol




nested_f_test_alcohol <- anova(model_alcohol2, model_alcohol3)
print(nested_f_test_alcohol)
```