SI 618 Final Report (Part B)

Name: Omkar R Sunkersett ; Date: April 13, 2017

Motivation

My report tries to answer a research question about online recipes — "Is there really a relationship amongst the nutrients of these recipes?" We have millions of recipes available readily over the internet but do not understand the underlying relationships amongst their nutrients since such information is not instantly available with the recipes. My report focuses on such relationships to understand the "science" behind the nutrition of these recipes by answering four critical questions:

1. What is the probability value of a possible linear regression amongst the nutrients?
2. Can we accept/reject the null hypothesis of a one-sided test at some level of significance?
3. Is there any correlation observed while applying a linear model to the nutritional data?
4. What are the findings about the relationships from the estimated linear equations?

Data Sources

I used two sources of data for my analysis so that it could be double-checked with data from two disparate sources for the purpose of verification.

Dataset 1: Epicurious Website

"Epicurious" is an online food resource providing millions of users with over 33,000 recipes, menus, ingredients, food preparation tips and expert advice. I am focusing only on the highest rated recipes related to lunch for my analysis.

URL: http://www.epicurious.com/search/?meal=lunch&sort=highestRated&content=recipe

Data Format: HTML (parsed using BeautifulSoup in Python)

Important Variables and Their Types:

Recipe: This variable indicates the name of the recipe scraped from “Epicurious”. It is a string.

Source: This variable indicates the source of the recipe. It is a string.

Calories: This variable indicates the total calories present in the recipe. It is a floating-point value and is measured in the unit of “cal”.

Protein: This variable indicates the amount of protein present in the recipe. It is a floating-point value and is measured in the unit of “grams”.

Carbohydrates: This variable indicates the amount of carbohydrates present in the recipe. It is a floating-point value and is measured in the unit of “grams”.

Fat: This variable indicates the total fat present in the recipe. It is a floating-point value and is measured in the unit of “grams”.

Saturated Fat: This variable indicates the amount of saturated fat present in the recipe. It is a floating-point value and is measured in the unit of “grams”.

Cholesterol: This variable indicates the amount of cholesterol present in the recipe. It is a floating-point value and is measured in the unit of “milligrams”.

Fiber: This variable indicates the amount of fiber present in the recipe. It is a floating-point value and is measured in the unit of “grams”.

Sodium: This variable indicates the amount of sodium present in the recipe. It is a floating-point value and is measured in the unit of “milligrams”.

Number of recipes obtained from Epicurious = 1,827

These recipes have been published on the internet from over a decade ago.

Dataset 2: Spoonacular API

The Spoonacular API contains recipe related information for over 365,000 recipes and 86,000 food products, including price and nutrition related data for each recipe. I am extracting one recipe from Spoonacular per recipe of Epcurious based on the condition of the maximum commonality of ingredients between each pair of recipe from both the sources.

URL: https://market.mashape.com/spoonacular/recipe-food-nutrition

Data Format: JSON (using Python)

Important Variables and Their Types:

Recipe: This variable indicates the name of the recipe obtained from “Spoonacular API”. It is a string.

Source: This variable indicates the source of the recipe. It is a string.

Calories: This variable indicates the total calories present in the recipe. It is a floating-point value and is measured in the unit of “cal”.

Protein: This variable indicates the amount of protein present in the recipe. It is a floating-point value and is measured in the unit of “grams”.

Carbohydrates: This variable indicates the amount of carbohydrates present in the recipe. It is a floating-point value and is measured in the unit of “grams”.

Fat: This variable indicates the total fat present in the recipe. It is a floating-point value and is measured in the unit of “grams”.

Saturated Fat: This variable indicates the amount of saturated fat present in the recipe. It is a floating-point value and is measured in the unit of “grams”.

Cholesterol: This variable indicates the amount of cholesterol present in the recipe. It is a floating-point value and is measured in the unit of “milligrams”.

Fiber: This variable indicates the amount of fiber present in the recipe. It is a floating-point value and is measured in the unit of "grams".

Sodium: This variable indicates the amount of sodium present in the recipe. It is a floating-point value and is measured in the unit of "milligrams".

Number of recipes obtained from Spoonacular = 1,827

These recipes have been published on the internet from over a decade ago.

Method (used for each of the four questions)

I scraped the data from the Epicurious website using BeautifulSoup in Python, extracting the name of the recipe, ingredients and nutritional information from the webpage. While extracting the data, I removed the special characters (mainly # and ,) from the recipe names and ignored those recipes that had no nutritional information. There were around 2,400 recipes out of which 1,827 had their nutritional information available. The nutrients had their units associated with them and daily recommended values (%) that I had to split from the actual integer value (Protein 40 g (80% DV)). I stored these 1,827 recipes in a local cache called "epicurious-cache.txt" as a list of tuples of the form [(recipe-name-n, [(nutrient-1, value-1), …, (nutrient-n, value-n)], [ingredient-1, …, ingredient-n]),…]. I then used the Spoonacular API and passed these values into the API using HTTP GET requests with the help of the requests module in Python. The Spoonacular API took the following parameters as input —

X-Mashape-Key: This is the client token used for authentication with the web server.

Accept: This specifies the format in which to retrieve the data from the web server (i.e. application/json, etc).

fillIngredients: This is a Boolean parameter used to indicate that the list of ingredients will be provided by the user.

ingredients: This is a string of comma-separated values of ingredients given as the input.

limitLicense: This is a Boolean parameter used to indicate whether the user should be charged beyond the maximum usage limit of the API (number of requests) based on the plan of the user or the HTTP GET request be aborted.

number: This is an integer parameter used to indicate the number of recipe IDs to retrieve from the web server.

ranking: This is an integer parameter used to indicate the mechanism to be used to retrieve the results. An integer value of 1 indicates that the recipes should be selected on the basis of the maximum ingredient usage (i.e. the maximum number of common ingredients between the "ingredients" parameter and the actual ingredients present in the "Spoonacular" recipe) whereas an integer value of 2 indicates that the recipes should be selected on the basis of the minimum ingredient usage (i.e. the minimum number of common ingredients between the "ingredients" parameter and the actual ingredients present in the "Spoonacular" recipe).

I obtained the recipe ID of a "Spoonacular" recipe during each HTTP GET request as the feedback from the web server and passed this recipe ID as an input to another HTTP GET request to extract the nutritional information about the recipe. The additional parameter used during this HTTP GET request was "includeNutrition", which takes a Boolean value to indicate whether the user wants to retrieve the nutritional information about the recipe. Finally, I got a dictionary in JSON format as the return value from the web server for every recipe ID. I stored all these dictionaries in a local cache called "spoonacular-cache.txt" as a list of dictionaries. I also ensured that I ignored some recipes that had missing calorie values (i.e. 'NA' values) from the final dataset. These recipes were "Mini Chorizo Corn Dogs" and "Steak Rancheros". I also assigned 0.0 values to those nutrients that were missing and ignored recipes with greater than 2000 calories from my analysis. I also did not include "sugar (g)" values from "Spoonacular" because they were completely missing from the "Epicurious" data and I needed at least two datasets to corroborate my results. The final dataset was prepared by combining the "cleaned" data from both the sources in TSV (tab-separated value) format. The final dataset has 3,655 recipes in total.

I read the dataset into R using the read.table() function —

dt <- read.table("/Users/omkarsunkersett/Desktop/SI618/project/dataset-combined.tsv", header = TRUE, sep = "\t", quote = "")

I divided the data table "dt" into subsets based on two important conditions to improve the statistical significance of the data —

1. **Source:** I used the subset() function in R to split the dataset into smaller data tables called "epicurious" and "spoonacular" based on the "source" attribute of the recipe.

epicurious <- subset(dt,dt$Source=='Epicurious')

spoonacular <- subset(dt,dt$Source=='Spoonacular')

2. **Calorie:** I then applied the subset() function to split the "epicurious" and "spoonacular" data tables into smaller sub-tables based on the following conditions:

Epicurious:

A. **Subgroup 0 - 250 cal:** epi_g1 <- subset(epicurious, epicurious$Calories >=0 & epicurious$Calories <=250)

B. **Subgroup 250 - 500 cal:** epi_g2 <- subset(epicurious, epicurious$Calories >250 & epicurious$Calories <=500)

C. **Subgroup 500 - 750 cal:** epi_g3 <- subset(epicurious, epicurious$Calories >500 & epicurious$Calories <=750)

D. **Subgroup 750 - 2000 cal:** epi_g4 <- subset(epicurious, epicurious$Calories >750 & epicurious$Calories <=2000)

Spoonacular:

A. **Subgroup 0 - 250 cal:** spoon_g1 <- subset(spoonacular, spoonacular$Calories >=0 & spoonacular$Calories <=250)

B. **Subgroup 250 - 500 cal:** spoon_g2 <- subset(spoonacular, spoonacular$Calories >250 & spoonacular$Calories <=500)

C. **Subgroup 500 - 750 cal:** spoon_g3 <- subset(spoonacular, spoonacular$Calories >500 & spoonacular$Calories <=750)

D. **Subgroup 750 - 2000 cal:** spoon_g4 <- subset(spoonacular, spoonacular$Calories >750 & spoonacular$Calories <=2000)

Analysis and Results:

To answer the fundamental research question and four key questions, I derived the relationship amongst a given set of nutrients, namely the following pairs —

1. Saturated Fat and Cholesterol
2. Saturated Fat and Sodium
3. Total Fat and Carbohydrates
4. Protein and Fiber
5. Calories and Significant Nutrients (Total Fat, Carbohydrates, Protein and Fiber)

I performed a one-sided hypothesis test to determine the relationship between the above nutrients.

My hypothesis has been defined as follows —

Null Hypothesis, Ho = There is no known relationship amongst the given nutrients of a calorie subgroup.

Alternative Hypothesis, Ha = There is some relationship amongst the given nutrients of a calorie subgroup.

Using the linear model (lm() function) and level of significance () = 5% (i.e. 0.05), I performed my hypothesis test as follows —

<u>Studying the relationship between Saturated Fat and Cholesterol for Epicurious data</u>

A. Calorie Subgroup: 0 - 250 cal:

```
epi_g1_lm1 <- lm(Saturated.Fat..g. ~ Cholesterol..mg., data = epi_g1)

summary(epi_g1_lm1)

ggplot(data = epi_g1, mapping = aes(x=Cholesterol..mg., y=Saturated.Fat..g., col=factor(Saturated.Fat..g.))) +
  geom_jitter(col="darkorange",size=4,shape=15,alpha=0.6) +
  geom_quantile(color = "black",size=1,linetype=4,alpha=0.8) +
  ggtitle("Relationship Between Saturated Fat and Cholesterol (Calories: 0 - 250)") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Cholesterol (mg)") +
  ylab("Saturated Fat (g)”)

##
## Call:
## lm(formula = Saturated.Fat..g. ~ Cholesterol..mg., data = epi_g1)
```

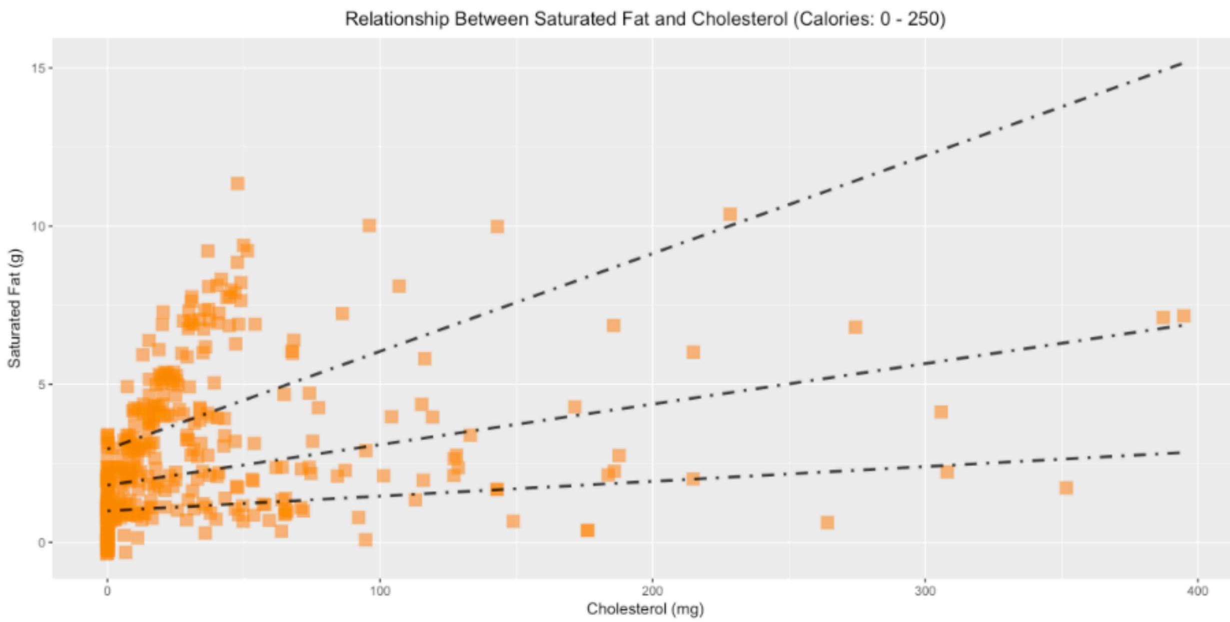
```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.605 -1.315 -0.423  1.250  8.145
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.263046   0.118735  19.060  < 2e-16 ***
## Cholesterol..mg. 0.012335   0.001892   6.519 1.95e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.161 on 438 degrees of freedom
## Multiple R-squared:  0.08844,    Adjusted R-squared:  0.08636
## F-statistic:  42.5 on 1 and 438 DF,  p-value: 1.952e-10
```

From the above output, p-value < 0.05 indicates that we have to reject the null hypothesis Ho.

We observe a strong positive (+) linear correlation between saturated fat and cholesterol.

The linear relationship for the above analysis is given by the following equation:

$Y = 2.263046 + 0.012335 * X + \epsilon$;   X: Cholesterol (mg), Y: Saturated Fat (g)



Finding: The saturated fat is estimated to increase by an average of 0.1234g for every 10mg average increase in cholesterol. Saturated fat contains a high proportion of low-density lipoprotein (LDL) cholesterol, which is a leading cause of heart disease whilst a person has high triglycerides (sugar). The average increase in saturated fat is very less for the average increase in cholesterol. This suggests that the proportion of the LDL cholesterol is comparatively lesser than the proportion of high-density lipoprotein (HDL) cholesterol and other cholesterol in the recipes. The HDL cholesterol is the good cholesterol and must be maximized in a person's lipid profile to prevent heart disease, whereas the LDL cholesterol is the culprit and must be minimized.

Note: The dotted lines indicate the first, second and third quantiles of the data points in the graph.

B. Calorie Subgroup: 250 - 500 cal:

epi_g2_lm1 <- lm(Saturated.Fat..g. ~ Cholesterol..mg., data = epi_g2)

summary(epi_g2_lm1)

ggplot(data = epi_g2, mapping = aes(x=Cholesterol..mg., y=Saturated.Fat..g., col=factor(Saturated.Fat..g.))) +

```
  geom_jitter(col="darkorange",size=4,shape=16,alpha=0.6) +
  geom_quantile(color = "black",size=1,linetype=4,alpha=0.8) +
  ggtitle("Relationship Between Saturated Fat and Cholesterol (Calories: 250 - 500)") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Cholesterol (mg)") +
  ylab("Saturated Fat (g)”)
##
## Call:
## lm(formula = Saturated.Fat..g. ~ Cholesterol..mg., data = epi_g2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4868 -2.7947 -0.8868  1.7128 20.6710
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       5.27997    0.19909  26.520  < 2e-16 ***
## Cholesterol..mg.  0.01225    0.00165   7.423 3.47e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

##

Residual standard error: 3.977 on 672 degrees of freedom

Multiple R-squared: 0.07579,   Adjusted R-squared: 0.07441

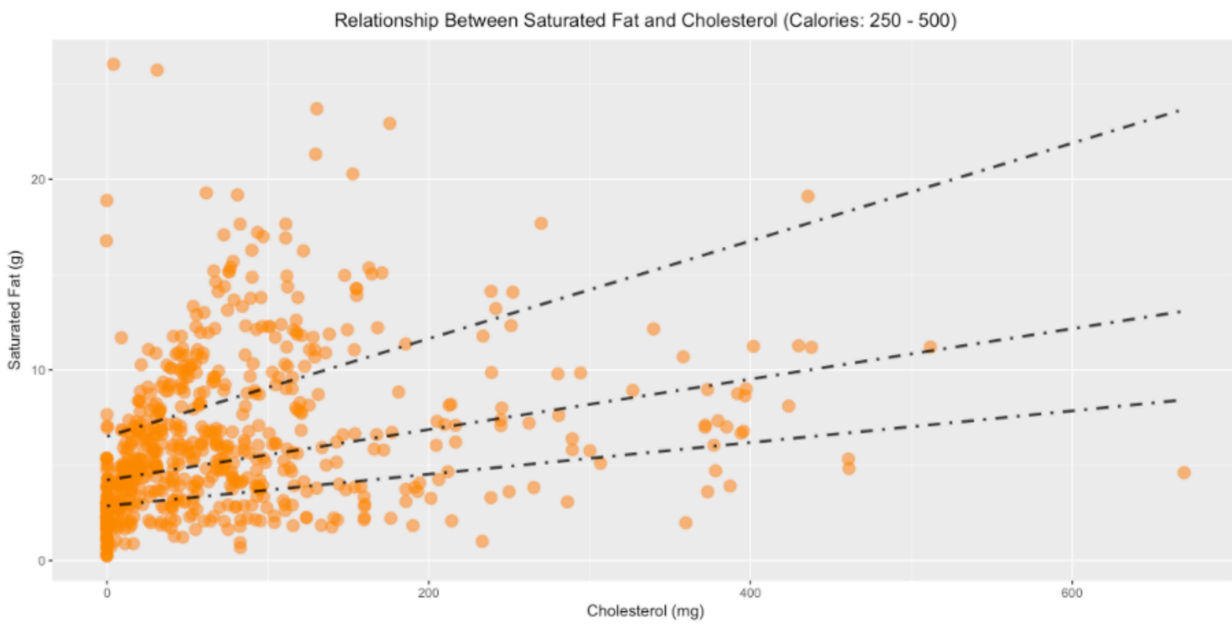
F-statistic: 55.11 on 1 and 672 DF, p-value: 3.466e-13

From the above output, p-value < 0.05 indicates that we have to reject the null hypothesis Ho.

We observe a strong positive (+) linear correlation between saturated fat and cholesterol.

The linear relationship for the above analysis is given by the following equation:

$Y = 5.27997 + 0.01225 * X + \epsilon$;   X: Cholesterol (mg), Y: Saturated Fat (g)



Finding: The saturated fat is estimated to increase by an average of 0.1225g for every 10mg average increase in cholesterol. Saturated fat contains a high proportion of low-density lipoprotein (LDL) cholesterol, which is a leading cause of heart disease whilst a person has high triglycerides (sugar). The average increase in saturated fat is very less for the average increase in cholesterol. This suggests that the proportion of the LDL cholesterol is comparatively lesser than the proportion of high-density lipoprotein (HDL) cholesterol and other cholesterol in the recipes. The HDL cholesterol is the good cholesterol and must be maximized in a person's lipid profile to prevent heart disease, whereas the LDL cholesterol is the culprit and must be minimized.

Note: The dotted lines indicate the first, second and third quantiles of the data points in the graph.

C. Calorie Subgroup: 500 - 750 cal:

```
epi_g3_lm1 <- lm(Saturated.Fat..g. ~ Cholesterol..mg., data = epi_g3)

summary(epi_g3_lm1)

ggplot(data = epi_g3, mapping = aes(x=Cholesterol..mg., y=Saturated.Fat..g., col=factor(Saturated.Fat..g.))) +
  geom_jitter(col="darkorange",size=4,shape=17,alpha=0.6) +
  geom_quantile(color = "black",size=1,linetype=4,alpha=0.8) +
  ggtitle("Relationship Between Saturated Fat and Cholesterol (Calories: 500 - 750)") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Cholesterol (mg)") +
  ylab("Saturated Fat (g)")

##
## Call:
## lm(formula = Saturated.Fat..g. ~ Cholesterol..mg., data = epi_g3)
##
```

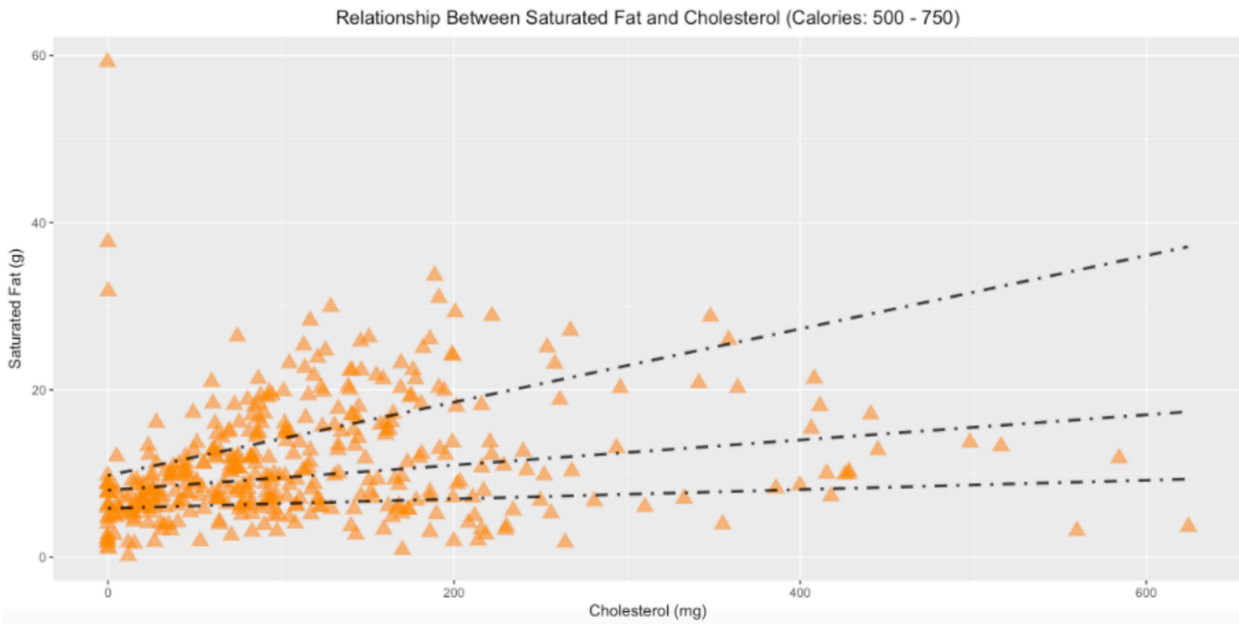
```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.965  -4.655  -1.582   3.494  49.351
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       9.64866    0.51941  18.576  < 2e-16 ***
## Cholesterol..mg.  0.01307    0.00329   3.971  8.5e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.864 on 392 degrees of freedom
## Multiple R-squared:  0.03868,    Adjusted R-squared:  0.03623
## F-statistic: 15.77 on 1 and 392 DF,  p-value: 8.501e-05
```

From the above output, p-value < 0.05 indicates that we have to reject the null hypothesis Ho.

We observe a strong positive (+) linear correlation between saturated fat and cholesterol.

The linear relationship for the above analysis is given by the following equation:

$Y = 9.64866 + 0.01307 * X + \epsilon$;   X: Cholesterol (mg), Y: Saturated Fat (g)



Finding: The saturated fat is estimated to increase by an average of 0.1307g for every 10mg average increase in cholesterol. Saturated fat contains a high proportion of low-density lipoprotein (LDL) cholesterol, which is a leading cause of heart disease whilst a person has high triglycerides (sugar). The average increase in saturated fat is very less for the average increase in cholesterol. This suggests that the proportion of the LDL cholesterol is comparatively lesser than the proportion of high-density lipoprotein (HDL) cholesterol and other cholesterol in the recipes. The HDL cholesterol is the good cholesterol and must be maximized in a person's lipid profile to prevent heart disease, whereas the LDL cholesterol is the culprit and must be minimized.

Note: The dotted lines indicate the first, second and third quantiles of the data points in the graph.

D. Calorie Subgroup: 750 - 2000 cal:

```
epi_g4_lm1 <- lm(Saturated.Fat..g. ~ Cholesterol..mg., data = epi_g4)

summary(epi_g4_lm1)

ggplot(data = epi_g4, mapping = aes(x=Cholesterol..mg., y=Saturated.Fat..g., col=factor(Saturated.Fat..g.))) +
```

```
  geom_jitter(col="darkorange",size=4,shape=18,alpha=0.6) +
  geom_quantile(color = "black",size=1,linetype=4,alpha=0.8) +
  ggtitle("Relationship Between Saturated Fat and Cholesterol (Calories: 750 - 2000)") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Cholesterol (mg)") +
  ylab("Saturated Fat (g)”)
##
## Call:
## lm(formula = Saturated.Fat..g. ~ Cholesterol..mg., data = epi_g4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.252  -6.515  -1.327   4.736  39.470
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      15.549555   0.912261  17.045  < 2e-16 ***
## Cholesterol..mg.  0.014694   0.003644   4.032  7.2e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

##

Residual standard error: 9.966 on 270 degrees of freedom

Multiple R-squared: 0.0568, Adjusted R-squared: 0.0533

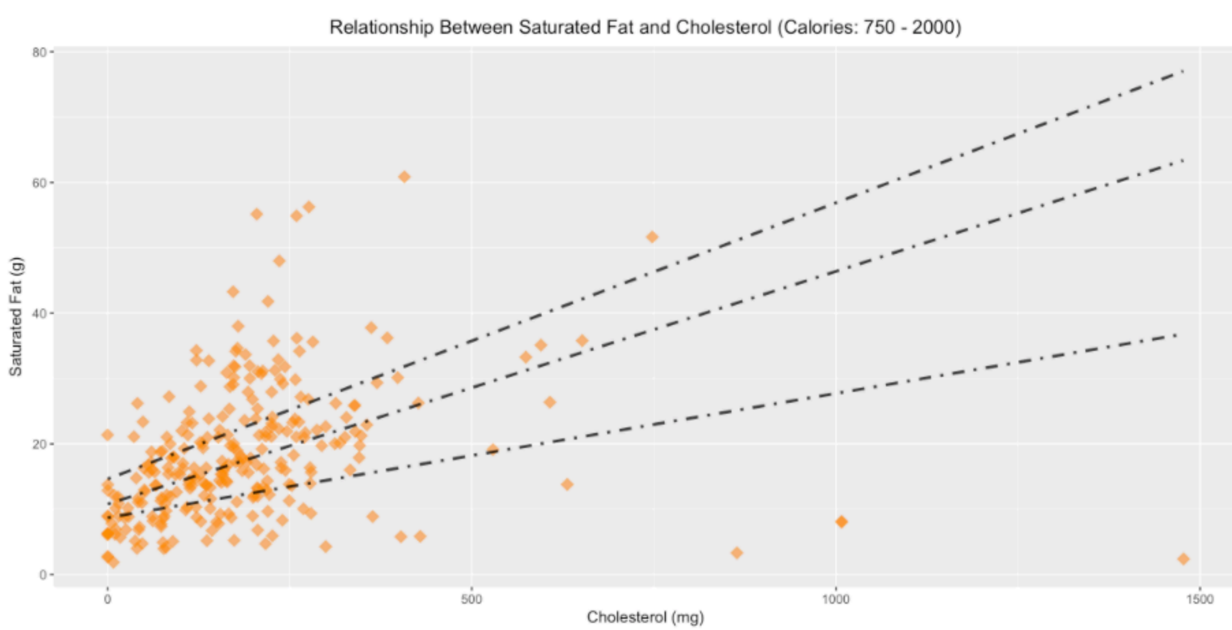
F-statistic: 16.26 on 1 and 270 DF, p-value: 7.196e-05

From the above output, p-value < 0.05 indicates that we have to reject the null hypothesis Ho.

We observe a strong positive (+) linear correlation between saturated fat and cholesterol.

The linear relationship for the above analysis is given by the following equation:

$Y = 15.549555 + 0.014694 * X + \epsilon$; X: Cholesterol (mg), Y: Saturated Fat (g)



Finding: The saturated fat is estimated to increase by an average of 0.1307g for every 10mg average increase in cholesterol. Saturated fat contains a high proportion of low-density lipoprotein (LDL) cholesterol, which is a leading cause of heart disease whilst a person has high triglycerides (sugar). The average increase in saturated fat is very less for the average increase in cholesterol. This suggests that the proportion of the LDL cholesterol is comparatively lesser than the proportion of high-density lipoprotein (HDL) cholesterol and other cholesterol in the recipes. The HDL cholesterol is the good cholesterol and must be maximized in a person's lipid profile to prevent heart disease, whereas the LDL cholesterol is the culprit and must be minimized.

Note: The dotted lines indicate the first, second and third quantiles of the data points in the graph.

Studying the relationship between Saturated Fat and Sodium for Epicurious data

A. Calories Subgroup: 0 - 250 cal:

```
epi_g1_lm2 <- lm(Saturated.Fat..g. ~ Sodium..mg., data = epi_g1)

summary(epi_g1_lm2)

ggplot(data = epi_g1, mapping = aes(x=Sodium..mg., y=Saturated.Fat..g., col=factor(Saturated.Fat..g.))) +
  geom_jitter(col="darkorange",size=4,shape=15,alpha=0.6) +
  geom_quantile(color = "black",size=1,linetype=4,alpha=0.8) +
  ggtitle("Relationship Between Saturated Fat and Sodium (Calories: 0 - 250)") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Sodium (mg)") +
  ylab("Saturated Fat (g)")

##
```

```
## Call:
## lm(formula = Saturated.Fat..g. ~ Sodium..mg., data = epi_g1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0116 -1.6097 -0.6511  1.3199  8.3988
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.5685216  0.1526463  16.827   <2e-16 ***
## Sodium..mg. 0.0002675  0.0003649   0.733    0.464
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.262 on 438 degrees of freedom
## Multiple R-squared:  0.001226,   Adjusted R-squared:  -0.001054
## F-statistic: 0.5376 on 1 and 438 DF,  p-value: 0.4638
```

From the above output, p-value > 0.05 indicates that we have to accept the null hypothesis Ho.

There is no correlation observed between saturated fat and sodium.

B. Calories Subgroup: 250 - 500 cal:

```
epi_g2_lm2 <- lm(Saturated.Fat..g. ~ Sodium..mg., data = epi_g2)

summary(epi_g2_lm2)

ggplot(data = epi_g2, mapping = aes(x=Sodium..mg., y=Saturated.Fat..g., col=factor(Saturated.Fat..g.))) +
  geom_jitter(col="darkorange",size=4,shape=16,alpha=0.6) +
  geom_quantile(color = "black",size=1,linetype=4,alpha=0.8) +
  ggtitle("Relationship Between Saturated Fat and Sodium (Calories: 250 - 500)") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Sodium (mg)") +
  ylab("Saturated Fat (g)”)

##
## Call:
## lm(formula = Saturated.Fat..g. ~ Sodium..mg., data = epi_g2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.894 -2.857 -1.000  1.906 20.186
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.6755810  0.2653128  21.392   <2e-16 ***
## Sodium..mg. 0.0010664  0.0004136   2.578   0.0101 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.116 on 672 degrees of freedom
## Multiple R-squared:  0.009795,   Adjusted R-squared:  0.008322
## F-statistic: 6.647 on 1 and 672 DF,  p-value: 0.01014
```

From the above output, p-value < 0.05 indicates that we have to reject the null hypothesis Ho.

There is a slight positive (+) correlation observed between saturated fat and sodium but this is not strong. Hence, we can ignore this finding.

C. Calorie Subgroup: 500 - 750 cal:

```
epi_g3_lm2 <- lm(Saturated.Fat..g. ~ Sodium..mg., data = epi_g3)

summary(epi_g3_lm2)

ggplot(data = epi_g3, mapping = aes(x=Sodium..mg., y=Saturated.Fat..g., col=factor(Saturated.Fat..g.))) +
  geom_jitter(col="darkorange",size=4,shape=17,alpha=0.6) +
  geom_quantile(color = "black",size=1,linetype=4,alpha=0.8) +
  ggtitle("Relationship Between Saturated Fat and Sodium (Calories: 500 - 750)") +
```

```
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Sodium (mg)") +
  ylab("Saturated Fat (g)”)
##
## Call:
## lm(formula = Saturated.Fat..g. ~ Sodium..mg., data = epi_g3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.123  -4.933  -1.682   3.602  47.990
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.2075023  0.6351831  19.219   <2e-16 ***
## Sodium..mg. -0.0012991  0.0006744  -1.926   0.0548 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.968 on 392 degrees of freedom
## Multiple R-squared:  0.009376,   Adjusted R-squared:  0.006849
```

F-statistic: 3.71 on 1 and 392 DF, p-value: 0.0548

From the above output, p-value > 0.05 indicates that we have to accept the null hypothesis Ho.

There is no correlation observed between saturated fat and sodium.

D. Calorie Subgroup: 750 - 2000 cal:

```
epi_g4_lm2 <- lm(Saturated.Fat..g. ~ Sodium..mg., data = epi_g4)

summary(epi_g4_lm2)

ggplot(data = epi_g4, mapping = aes(x=Sodium..mg., y=Saturated.Fat..g., col=factor(Saturated.Fat..g.))) +

  geom_jitter(col="darkorange",size=4,shape=18,alpha=0.6) +

  geom_quantile(color = "black",size=1,linetype=4,alpha=0.8) +

  ggtitle("Relationship Between Saturated Fat and Sodium (Calories: 750 - 2000)") +

  theme(plot.title = element_text(hjust = 0.5)) +

  xlab("Sodium (mg)") +

  ylab("Saturated Fat (g)")

##

## Call:

## lm(formula = Saturated.Fat..g. ~ Sodium..mg., data = epi_g4)

##

## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -19.245  -6.848  -1.890   5.094  43.133
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.690e+01  1.102e+00  15.328   <2e-16 ***
## Sodium..mg. 1.077e-03  6.968e-04   1.546    0.123
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.22 on 270 degrees of freedom
## Multiple R-squared:  0.008779,   Adjusted R-squared:  0.005108
## F-statistic: 2.391 on 1 and 270 DF,  p-value: 0.1232
```

From the above output, p-value > 0.05 indicates that we have to accept the null hypothesis Ho.

There is no correlation observed between saturated fat and sodium.

Studying the relationship between Total Fat and Carbohydrates for Epicurious data

A. Calorie Subgroup: 0 - 250 cal:

```
epi_g1_lm3 <- lm(Fat..g. ~ Carbohydrates..g., data = epi_g1)
```

```
summary(epi_g1_lm3)

ggplot(data = epi_g1, mapping = aes(x=Carbohydrates..g., y=Fat..g., col=factor(Fat..g.))) +
  geom_jitter(col="darkorange",size=4,shape=15,alpha=0.6) +
  geom_quantile(color = "black",size=1,linetype=4,alpha=0.8) +
  ggtitle("Relationship Between Total Fat and Carbohydrates (Calories: 0 - 250)") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Carbohydrates (g)") +
  ylab("Total Fat (g)")

##
## Call:
## lm(formula = Fat..g. ~ Carbohydrates..g., data = epi_g1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.2112  -3.5251   0.1945   3.6226  12.8608
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       11.21117    0.43532  25.754  < 2e-16 ***
## Carbohydrates..g. -0.11910    0.02445  -4.872 1.54e-06 ***
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##

Residual standard error: 5.059 on 438 degrees of freedom

Multiple R-squared: 0.05141,   Adjusted R-squared: 0.04925

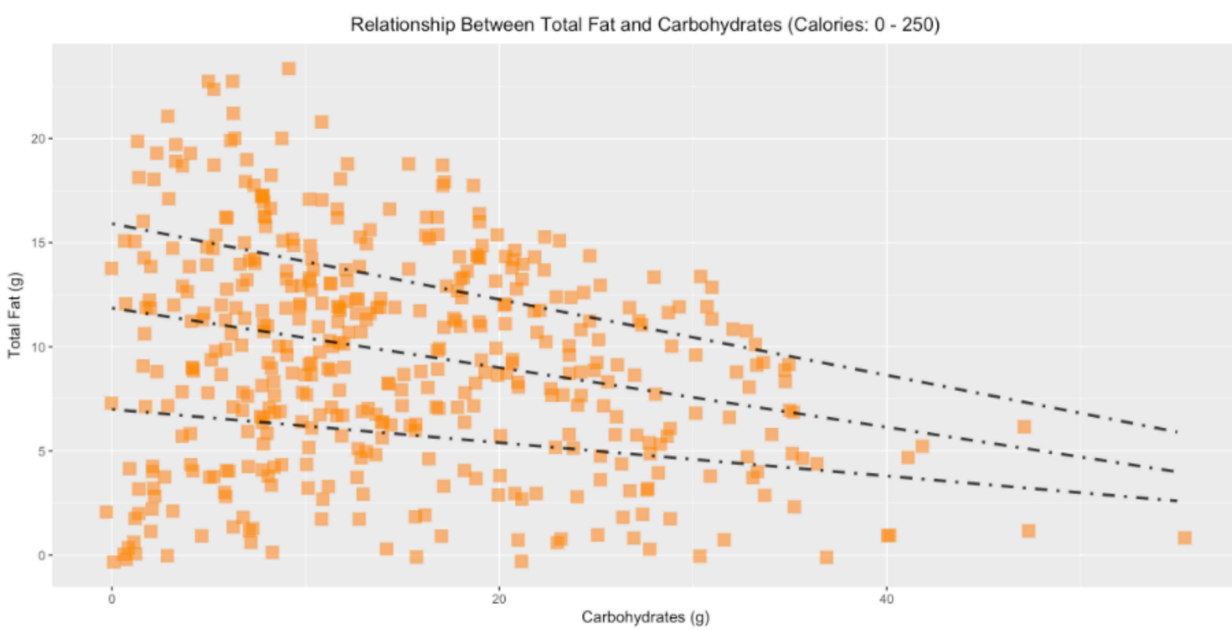
F-statistic: 23.74 on 1 and 438 DF, p-value: 1.545e-06

From the above output, p-value < 0.05 indicates that we have to reject the null hypothesis Ho.

We observe a strong negative (-) linear correlation between Total Fat and Carbohydrates.

The linear relationship for the above analysis is given by the following equation:

$Y = 11.21117 - 0.11910 * X + \epsilon$;   X: Fat (g), Y: Carbohydrates (g)



Finding: The total fat is estimated to decrease by an average of 1.1910g for every 10g average increase in carbohydrates. We see that the total fat to carbohydrate ratio of the recipes is well-balanced with the help of this linear relation. This indicates that a recipe having high total fat would likely have low carbohydrates and vice-versa.

Note: The dotted lines indicate the first, second and third quantiles of the data points in the graph.

B. Calorie Subgroup: 250 - 500 cal:

```
epi_g2_lm3 <- lm(Fat..g. ~ Carbohydrates..g., data = epi_g2)

summary(epi_g2_lm3)

ggplot(data = epi_g2, mapping = aes(x=Carbohydrates..g., y=Fat..g., col=factor(Fat..g.))) +
  geom_jitter(col="darkorange",size=4,shape=16,alpha=0.6) +
  geom_quantile(color = "black",size=1,linetype=4,alpha=0.8) +
  ggtitle("Relationship Between Total Fat and Carbohydrates (Calories: 250 - 500)") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Carbohydrates (g)") +
  ylab("Total Fat (g)”)

##
## Call:
## lm(formula = Fat..g. ~ Carbohydrates..g., data = epi_g2)
##
## Residuals:
```

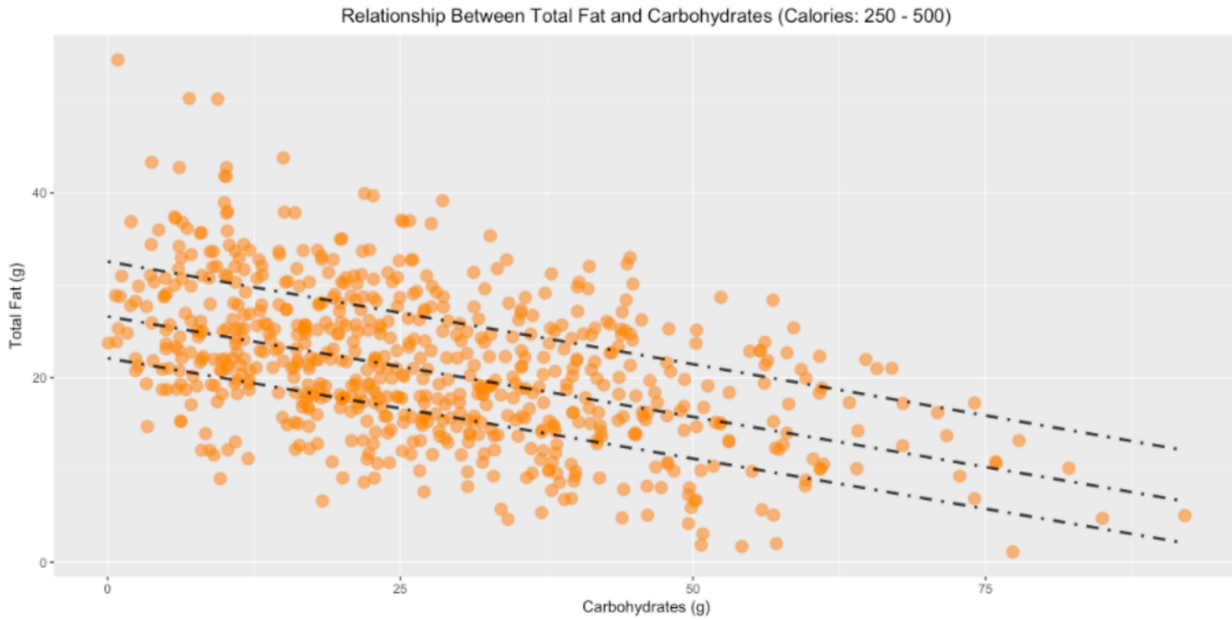
```
##      Min       1Q   Median       3Q      Max
## -16.5370  -5.1973  -0.6441   5.1547  26.5021
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       27.73094    0.52983   52.34   <2e-16 ***
## Carbohydrates..g. -0.23300    0.01645  -14.16   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.121 on 672 degrees of freedom
## Multiple R-squared:  0.2299, Adjusted R-squared:  0.2288
## F-statistic: 200.6 on 1 and 672 DF,  p-value: < 2.2e-16
```

From the above output, $p\text{-value} < 0.05$ indicates that we have to reject the null hypothesis Ho.

We observe a strong negative (-) linear correlation between Total Fat and Carbohydrates.

The linear relationship for the above analysis is given by the following equation:

$Y = 27.73094 - 0.23300 * X + \epsilon$;   X: Fat (g), Y: Carbohydrates (g)



Finding: The total fat is estimated to decrease by an average of 2.33g for every 10g average increase in carbohydrates. We see that the total fat to carbohydrate ratio of the recipes is well-balanced with the help of this linear relation. This indicates that a recipe having high total fat would likely have low carbohydrates and vice-versa.

Note: The dotted lines indicate the first, second and third quantiles of the data points in the graph.

C. Calorie Subgroup: 500 - 750 cal:

```
epi_g3_lm3 <- lm(Fat..g. ~ Carbohydrates..g., data = epi_g3)

summary(epi_g3_lm3)

ggplot(data = epi_g3, mapping = aes(x=Carbohydrates..g., y=Fat..g., col=factor(Fat..g.))) +

  geom_jitter(col="darkorange",size=4,shape=17,alpha=0.6) +

  geom_quantile(color = "black",size=1,linetype=4,alpha=0.8) +

  ggtitle("Relationship Between Total Fat and Carbohydrates (Calories: 500 - 750)") +

  theme(plot.title = element_text(hjust = 0.5)) +
```

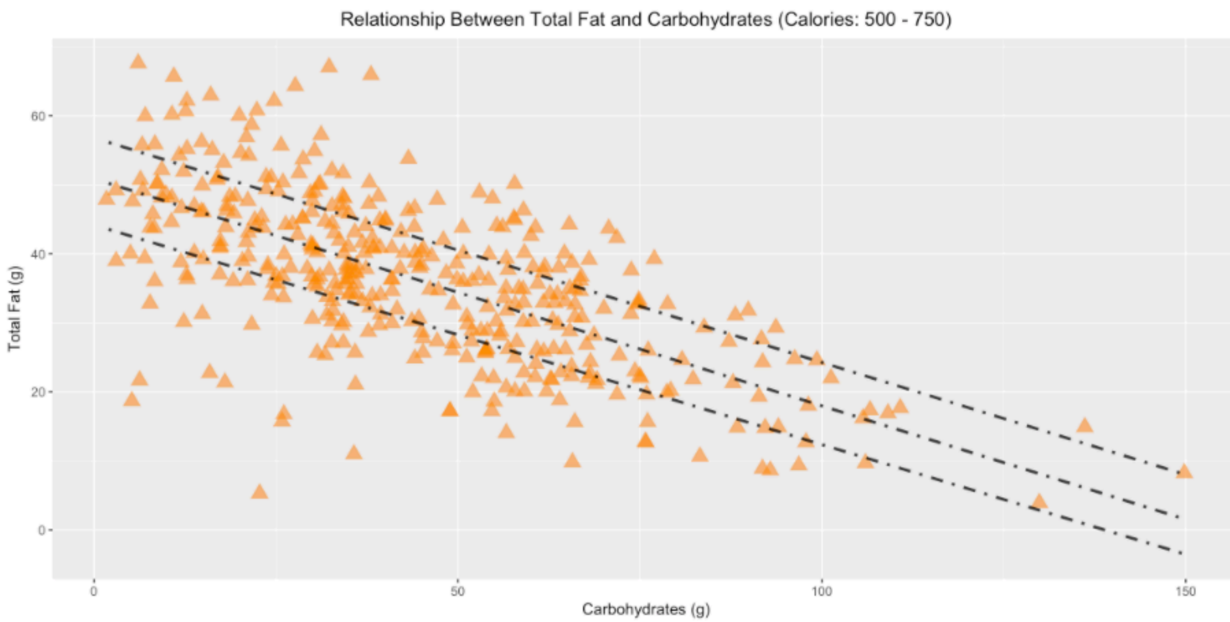
```
  xlab("Carbohydrates (g)") +
  ylab("Total Fat (g)")
##
## Call:
## lm(formula = Fat..g. ~ Carbohydrates..g., data = epi_g3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.841  -6.012   0.249   6.189  27.903
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        50.1153     0.9398   53.32   <2e-16 ***
## Carbohydrates..g.  -0.3163     0.0184  -17.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.086 on 392 degrees of freedom
## Multiple R-squared:  0.4298, Adjusted R-squared:  0.4284
## F-statistic: 295.5 on 1 and 392 DF,  p-value: < 2.2e-16
```

From the above output, p-value < 0.05 indicates that we have to reject the null hypothesis Ho.

We observe a strong negative (-) linear correlation between Total Fat and Carbohydrates.

The linear relationship for the above analysis is given by the following equation:

$Y = 50.1153 - 0.3163 * X + \epsilon$;   X: Fat (g), Y: Carbohydrates (g)



Finding: The total fat is estimated to decrease by an average of 3.163g for every 10g average increase in carbohydrates. We see that the total fat to carbohydrate ratio of the recipes is well-balanced with the help of this linear relation. This indicates that a recipe having high total fat would likely have low carbohydrates and vice-versa.

Note: The dotted lines indicate the first, second and third quantiles of the data points in the graph.

D. Calorie Subgroup: 750 - 2000 cal:

epi_g4_lm3 <- lm(Fat..g. ~ Carbohydrates..g., data = epi_g4)

```
summary(epi_g4_lm3)

ggplot(data = epi_g4, mapping = aes(x=Carbohydrates..g., y=Fat..g., col=factor(Fat..g.))) +
  geom_jitter(col="darkorange",size=4,shape=18,alpha=0.6) +
  geom_quantile(color = "black",size=1,linetype=4,alpha=0.8) +
  ggtitle("Relationship Between Total Fat and Carbohydrates (Calories: 750 - 2000)") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Carbohydrates (g)") +
  ylab("Total Fat (g)")

##
## Call:
## lm(formula = Fat..g. ~ Carbohydrates..g., data = epi_g4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.042 -18.758  -6.347  12.323 129.644
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        78.6701     2.7308  28.809  < 2e-16 ***
## Carbohydrates..g.  -0.1877     0.0327  -5.741 2.52e-08 ***
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##

Residual standard error: 26.94 on 270 degrees of freedom

Multiple R-squared: 0.1088, Adjusted R-squared: 0.1055

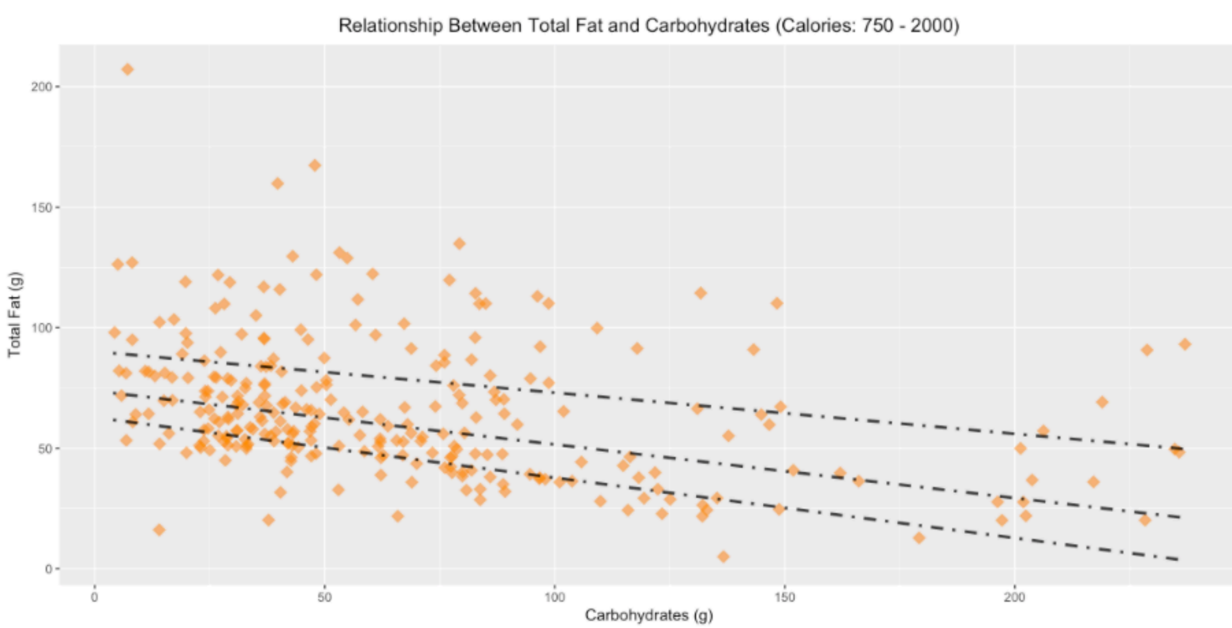
F-statistic: 32.96 on 1 and 270 DF, p-value: 2.522e-08

From the above output, p-value < 0.05 indicates that we have to reject the null hypothesis Ho.

We observe a strong negative (-) linear correlation between Total Fat and Carbohydrates.

The linear relationship for the above analysis is given by the following equation:

$Y = 78.6701 - 0.1877 * X + \epsilon$;   X: Fat (g), Y: Carbohydrates (g)



Finding: The total fat is estimated to decrease by an average of 1.877g for every 10g average increase in carbohydrates. We see that the total fat to carbohydrate ratio of the recipes is well-balanced with the help of this linear relation. This indicates that a recipe having high total fat would likely have low carbohydrates and vice-versa.

Note: The dotted lines indicate the first, second and third quantiles of the data points in the graph.

Studying the relationship between Protein and Fiber for Epicurious data

A. Calorie Subgroup: 0 - 250 cal:

```
epi_g1_lm4 <- lm(Protein..g. ~ Fiber..g., data = epi_g1)

summary(epi_g1_lm4)

ggplot(data = epi_g1, mapping = aes(x=Fiber..g., y=Protein..g., col=factor(Protein..g.))) +

  geom_jitter(col="darkorange",size=4,shape=15,alpha=0.6) +

  geom_quantile(color = "black",size=1,linetype=4,alpha=0.8) +

  ggtitle("Relationship Between Protein and Fiber (Calories: 0 - 250)") +

  theme(plot.title = element_text(hjust = 0.5)) +

  xlab("Fiber (g)") +

  ylab("Protein (g)”)

##

## Call:

## lm(formula = Protein..g. ~ Fiber..g., data = epi_g1)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.891 -3.955 -1.146  2.918 23.109
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.82747    0.41139  16.596   <2e-16 ***
## Fiber..g.    0.06374    0.11521   0.553     0.58
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.313 on 438 degrees of freedom
## Multiple R-squared:  0.0006983,  Adjusted R-squared:  -0.001583
## F-statistic: 0.3061 on 1 and 438 DF,  p-value: 0.5804
```

From the above output, p-value > 0.05 indicates that we have to accept the null hypothesis Ho.

There is no correlation observed between protein and fibre. This is bad because it indicates that the protein to fiber ratio is not well-balanced in most of the recipes. Ideally, protein and fibre should be positively correlated in a healthy meal.

B. Calorie Subgroup: 250 - 500 cal:

```
epi_g2_lm4 <- lm(Protein..g. ~ Fiber..g., data = epi_g2)

summary(epi_g2_lm4)

ggplot(data = epi_g2, mapping = aes(x=Fiber..g., y=Protein..g., col=factor(Protein..g.))) +
  geom_jitter(col="darkorange",size=4,shape=16,alpha=0.6) +
  geom_quantile(color = "black",size=1,linetype=4,alpha=0.8) +
  ggtitle("Relationship Between Protein and Fiber (Calories: 250 - 500)") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Fiber (g)") +
  ylab("Protein (g)")
```

```
##
## Call:
## lm(formula = Protein..g. ~ Fiber..g., data = epi_g2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.763  -7.463  -2.052   5.720  36.970
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  17.7628     0.6328  28.072   <2e-16 ***

## Fiber..g.    -0.2444     0.1184  -2.065   0.0393 *

## ---

## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 9.816 on 672 degrees of freedom

## Multiple R-squared:  0.006305,   Adjusted R-squared:  0.004826

## F-statistic: 4.264 on 1 and 672 DF,  p-value: 0.03931
```

From the above output, p-value < 0.05 indicates that we have to reject the null hypothesis Ho.

There is a slight negative (-) correlation observed between protein and fibre but it is not strong. Hence, we can ignore this finding. This finding is bad because it indicates that the protein to fiber ratio is not well-balanced in most of the recipes. Ideally, protein and fibre should be positively correlated in a healthy meal.

C. Calorie Subgroup: 500 - 750 cal:

```
epi_g3_lm4 <- lm(Protein..g. ~ Fiber..g., data = epi_g3)

summary(epi_g3_lm4)

ggplot(data = epi_g3, mapping = aes(x=Fiber..g., y=Protein..g., col=factor(Protein..g.))) +

  geom_jitter(col="darkorange",size=4,shape=17,alpha=0.6) +

  geom_quantile(color = "black",size=1,linetype=4,alpha=0.8) +

  ggtitle("Relationship Between Protein and Fiber (Calories: 500 - 750)") +
```

```
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Fiber (g)") +
  ylab("Protein (g)”)
## 
## Call:
## lm(formula = Protein..g. ~ Fiber..g., data = epi_g3)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -28.33  -9.29  -2.33   6.22  70.67 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  30.3707     1.0885  27.903  < 2e-16 ***
## Fiber..g.    -0.5201     0.1519  -3.425 0.000681 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 14.11 on 392 degrees of freedom
## Multiple R-squared:  0.02905,    Adjusted R-squared:  0.02657
```

F-statistic: 11.73 on 1 and 392 DF, p-value: 0.0006807

From the above output, p-value < 0.05 indicates that we have to reject the null hypothesis Ho.

There is a strong negative (-) correlation observed between protein and fibre. This finding is bad because it indicates that the protein to fiber ratio is not well-balanced in most of the recipes. Ideally, protein and fibre should be positively correlated in a healthy meal.

D. Calorie Subgroup: 750 - 2000 cal:

```
epi_g4_lm4 <- lm(Protein..g. ~ Fiber..g., data = epi_g4)

summary(epi_g4_lm4)

ggplot(data = epi_g4, mapping = aes(x=Fiber..g., y=Protein..g., col=factor(Protein..g.))) +
  geom_jitter(col="darkorange",size=4,shape=18,alpha=0.6) +
  geom_quantile(color = "black",size=1,linetype=4,alpha=0.8) +
  ggtitle("Relationship Between Protein and Fiber (Calories: 750 - 2000)") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Fiber (g)") +
  ylab("Protein (g)”)

##
## Call:
## lm(formula = Protein..g. ~ Fiber..g., data = epi_g4)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -46.688 -17.557  -5.314   9.662 167.011
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  51.0373     2.3877  21.375   <2e-16 ***
## Fiber..g.    -0.1747     0.2174  -0.804    0.422
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.21 on 270 degrees of freedom
## Multiple R-squared:  0.002386,   Adjusted R-squared:  -0.001308
## F-statistic: 0.6459 on 1 and 270 DF,  p-value: 0.4223
```

From the above output, p-value > 0.05 indicates that we have to accept the null hypothesis Ho.

There is no correlation observed between protein and fibre. This is bad because it indicates that the protein to fiber ratio is not well-balanced in most of the recipes. Ideally, protein and fibre should be positively correlated in a healthy meal.

<u>Studying the relationship between Calories and Significant Nutrients for Epicurious data</u>

A. Calorie Subgroup: 0 - 250 cal:

```
epi_g1_lm5 <- lm(Calories..cal. ~ (Fat..g.+Saturated.Fat..g.+Protein..g.+Carbohydrates..g.
+Fiber..g.+Sodium..mg.+Cholesterol..mg.), data = epi_g1)

summary(epi_g1_lm5)

ggplot(data = epi_g1, mapping = aes(x=(Fat..g.+Protein..g.+Carbohydrates..g.+Fiber..g.),
y=Calories..cal., col=Calories..cal.)) +
  geom_jitter(size=4,shape=15,alpha=0.6) +
  geom_quantile(color = "red",size=1,linetype=4,alpha=0.8) +
  ggtitle("Relationship Between Calories and Significant Nutrients (Calories: 0 - 250)") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Total Fat, Protein, Carbohydrates & Fiber (g)") +
  ylab("Calories (cal)”)

##
## Call:
## lm(formula = Calories..cal. ~ (Fat..g. + Saturated.Fat..g. +
##     Protein..g. + Carbohydrates..g. + Fiber..g. + Sodium..mg. +
##     Cholesterol..mg.), data = epi_g1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -20.856  -4.714  -0.996   3.133 122.052

##

## Coefficients:

##                      Estimate Std. Error t value Pr(>|t|)

## (Intercept)        6.0133849  1.6070984   3.742 0.000207 ***

## Fat..g.            8.4415615  0.1363637  61.905  < 2e-16 ***

## Saturated.Fat..g. -0.0289926  0.3096344  -0.094 0.925443

## Protein..g.        4.1123973  0.1238394  33.207  < 2e-16 ***

## Carbohydrates..g.  3.9396289  0.0741700  53.116  < 2e-16 ***

## Fiber..g.         -1.6656174  0.3167455  -5.259 2.29e-07 ***

## Sodium..mg.       -0.0008566  0.0019700  -0.435 0.663890

## Cholesterol..mg.   0.0176837  0.0121070   1.461 0.144849

## ---

## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 11.15 on 432 degrees of freedom

## Multiple R-squared:  0.9644, Adjusted R-squared:  0.9638

## F-statistic:  1671 on 7 and 432 DF,  p-value: < 2.2e-16
```

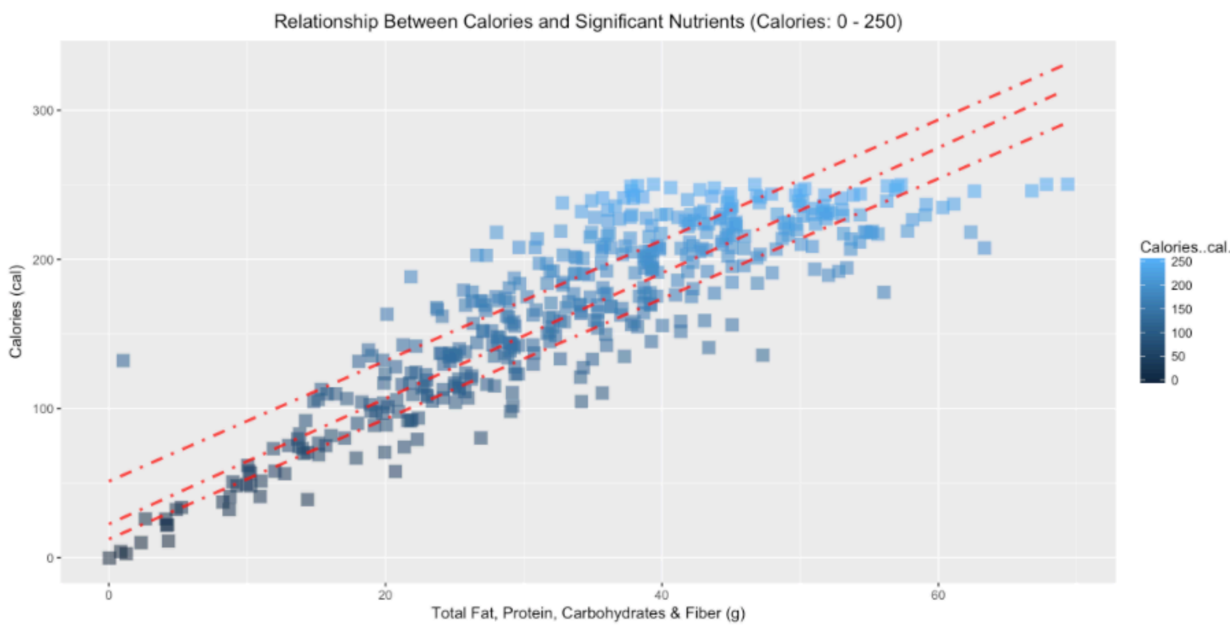
From the above output, p-value < 0.05 indicates that we have to reject the null hypothesis Ho.

There is a strong positive (+) linear relationship between the calories and above significant nutrients (fat, protein, fiber and carbohydrates).

The linear relationship for the above analysis is given by the following equation:

$Y = 6.0133849 + 8.4415615 * V + 4.1123973 * W + 3.9396289 * X - 1.6656174 * Z + \epsilon$

Y: Calories (cal), V: Fat (g), W: Protein (g), X: Carbohydrates (g), Z: Fiber (g)



Finding: An average increase of 10g in fat, protein, carbohydrates and fiber will likely result in an average increase in calories by 148.28 cal for the given recipes. This relationship can be used to optimize the total calories in the recipes by altering the proportions of their significant nutrients.

Note: The dotted lines indicate the first, second and third quantiles of the data points in the graph.

B. Calorie Subgroup: 250 - 500 cal:

```
epi_g2_lm5 <- lm(Calories..cal. ~ (Fat..g.+Saturated.Fat..g.+Protein..g.+Carbohydrates..g.
+Fiber..g.+Sodium..mg.+Cholesterol..mg.), data = epi_g2)

summary(epi_g2_lm5)

ggplot(data = epi_g2, mapping = aes(x=(Fat..g.+Protein..g.+Carbohydrates..g.+Fiber..g.),
y=Calories..cal., col=Calories..cal.)) +
  geom_jitter(size=4,shape=16,alpha=0.6) +
  geom_quantile(color = "red",size=1,linetype=4,alpha=0.8) +
  ggtitle("Relationship Between Calories and Significant Nutrients (Calories: 250 - 500)") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Total Fat, Protein, Carbohydrates & Fiber (g)") +
  ylab("Calories (cal)”)
##
## Call:
## lm(formula = Calories..cal. ~ (Fat..g. + Saturated.Fat..g. +
##     Protein..g. + Carbohydrates..g. + Fiber..g. + Sodium..mg. +
##     Cholesterol..mg.), data = epi_g2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.121  -4.308  -0.392   3.439  85.838
```

```
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.551069   1.812241   1.959   0.0505 .
## Fat..g.            8.621328   0.058383 147.670   <2e-16 ***
## Saturated.Fat..g.  0.026141   0.107332   0.244   0.8077
## Protein..g.        4.267933   0.043255  98.668   <2e-16 ***
## Carbohydrates..g.  3.988671   0.029738 134.125   <2e-16 ***
## Fiber..g.         -1.576052   0.140624 -11.208   <2e-16 ***
## Sodium..mg.        0.001395   0.001033   1.351   0.1771
## Cholesterol..mg.   0.010456   0.004573   2.287   0.0225 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.329 on 666 degrees of freedom
## Multiple R-squared:  0.9844, Adjusted R-squared:  0.9842
## F-statistic:  6000 on 7 and 666 DF,  p-value: < 2.2e-16
```

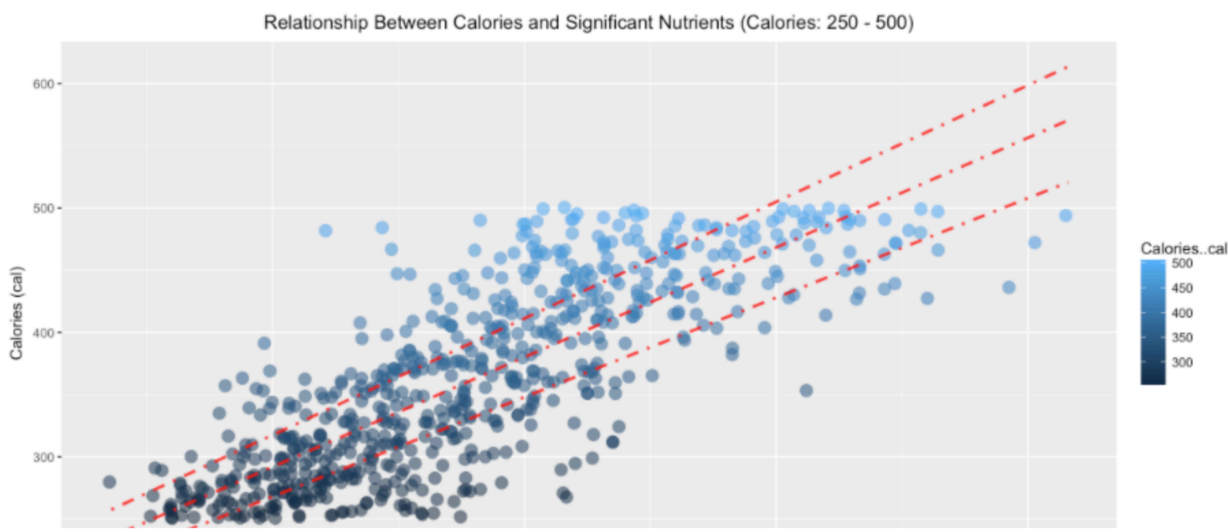
From the above output, p-value < 0.05 indicates that we have to reject the null hypothesis Ho.

There is a strong positive (+) linear relationship between the calories and above significant nutrients (fat, protein, fiber and carbohydrates).

The linear relationship for the above analysis is given by the following equation:

$Y = 3.551069 + 8.621328 * V + 4.267933 * W + 3.988671 * X - 1.576052 * Z + \epsilon$

Y: Calories (cal), V: Fat (g), W: Protein (g), X: Carbohydrates (g), Z: Fiber (g)



Finding: An average increase of 10g in fat, protein, carbohydrates and fiber will likely result in an average increase in calories by 153.02 cal for the given recipes. This relationship can be used to optimize the total calories in the recipes by altering the proportions of their significant nutrients.

Note: The dotted lines indicate the first, second and third quantiles of the data points in the graph.

C. Calorie Subgroup: 500 - 750 cal:

epi_g3_lm5 <- lm(Calories..cal. ~ (Fat..g.+Saturated.Fat..g.+Protein..g.+Carbohydrates..g.+Fiber..g.+Sodium..mg.+Cholesterol..mg.), data = epi_g3)

```
summary(epi_g3_lm5)

ggplot(data = epi_g3, mapping = aes(x=(Fat..g.+Protein..g.+Carbohydrates..g.+Fiber..g.),
y=Calories..cal., col=Calories..cal.)) +
  geom_jitter(size=4,shape=17,alpha=0.6) +
  geom_quantile(color = "red",size=1,linetype=4,alpha=0.8) +
  ggtitle("Relationship Between Calories and Significant Nutrients (Calories: 500 - 750)") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Total Fat, Protein, Carbohydrates & Fiber (g)") +
  ylab("Calories (cal)")

##
## Call:
## lm(formula = Calories..cal. ~ (Fat..g. + Saturated.Fat..g. +
##     Protein..g. + Carbohydrates..g. + Fiber..g. + Sodium..mg. +
##     Cholesterol..mg.), data = epi_g3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.249  -5.679  -0.564   4.819  75.036
##
## Coefficients:
```

```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       15.8377087  6.0430432   2.621  0.00912 **
## Fat..g.           8.5623806  0.0978442  87.510  < 2e-16 ***
## Saturated.Fat..g. -0.0917089  0.1143044  -0.802  0.42286
## Protein..g.        4.2231536  0.0643796  65.598  < 2e-16 ***
## Carbohydrates..g.  3.9694182  0.0492100  80.663  < 2e-16 ***
## Fiber..g.         -1.8394123  0.1741747 -10.561  < 2e-16 ***
## Sodium..mg.        0.0026466  0.0014045   1.884  0.06026 .
## Cholesterol..mg.  -0.0004335  0.0078601  -0.055  0.95605
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.12 on 386 degrees of freedom
## Multiple R-squared:  0.9635, Adjusted R-squared:  0.9628
## F-statistic:  1455 on 7 and 386 DF,  p-value: < 2.2e-16
```

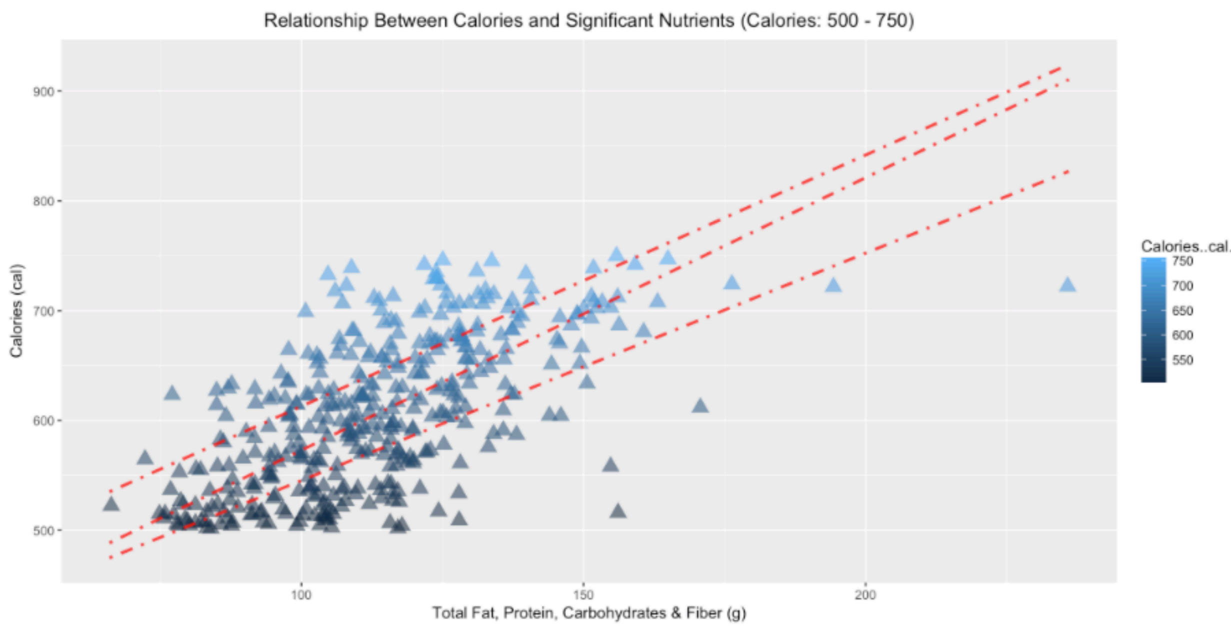
From the above output, p-value < 0.05 indicates that we have to reject the null hypothesis Ho.

There is a strong positive (+) linear relationship between the calories and above significant nutrients (fat, protein, fiber and carbohydrates).

The linear relationship for the above analysis is given by the following equation:

Y = 15.8377087 + 8.5623806 * V + 4.2231536 * W + 3.9694182 * X - 1.8394123 * Z + ϵ

Y: Calories (cal), V: Fat (g), W: Protein (g), X: Carbohydrates (g), Z: Fiber (g)



Finding: An average increase of 10g in fat, protein, carbohydrates and fiber will likely result in an average increase in calories by 149.16 cal for the given recipes. This relationship can be used to optimize the total calories in the recipes by altering the proportions of their significant nutrients.

Note: The dotted lines indicate the first, second and third quantiles of the data points in the graph.

D. Calorie Subgroup: 750 - 2000 cal:

epi_g4_lm5 <- lm(Calories..cal. ~ (Fat..g.+Saturated.Fat..g.+Protein..g.+Carbohydrates..g.+Fiber..g.+Sodium..mg.+Cholesterol..mg.), data = epi_g4)

summary(epi_g4_lm5)

```
ggplot(data = epi_g4, mapping = aes(x=(Fat..g.+Protein..g.+Carbohydrates..g.+Fiber..g.),
y=Calories..cal., col=Calories..cal.)) +

  geom_jitter(size=4,shape=18,alpha=0.6) +

  geom_quantile(color = "red",size=1,linetype=4,alpha=0.8) +

  ggtitle("Relationship Between Calories and Significant Nutrients (Calories: 750 - 2000)") +

  theme(plot.title = element_text(hjust = 0.5)) +

  xlab("Total Fat, Protein, Carbohydrates & Fiber (g)") +

  ylab("Calories (cal)")

##

## Call:

## lm(formula = Calories..cal. ~ (Fat..g. + Saturated.Fat..g. +

##     Protein..g. + Carbohydrates..g. + Fiber..g. + Sodium..mg. +

##     Cholesterol..mg.), data = epi_g4)

##

## Residuals:

##     Min      1Q  Median      3Q     Max

## -62.699  -7.508  -1.968   5.689  91.779

##

## Coefficients:

##                   Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)        0.414449   4.428075   0.094    0.926

## Fat..g.            8.852269   0.049903 177.389   <2e-16 ***

## Saturated.Fat..g.  0.131256   0.135946   0.965    0.335

## Protein..g.        4.406942   0.061130  72.091   <2e-16 ***

## Carbohydrates..g.  4.128899   0.031019 133.109   <2e-16 ***

## Fiber..g.         -2.536950   0.177095 -14.325   <2e-16 ***

## Sodium..mg.       -0.002018   0.001482  -1.362    0.174

## Cholesterol..mg.  -0.016485   0.010481  -1.573    0.117

## ---

## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 18.49 on 264 degrees of freedom

## Multiple R-squared:  0.9958, Adjusted R-squared:  0.9957

## F-statistic:  9016 on 7 and 264 DF,  p-value: < 2.2e-16
```

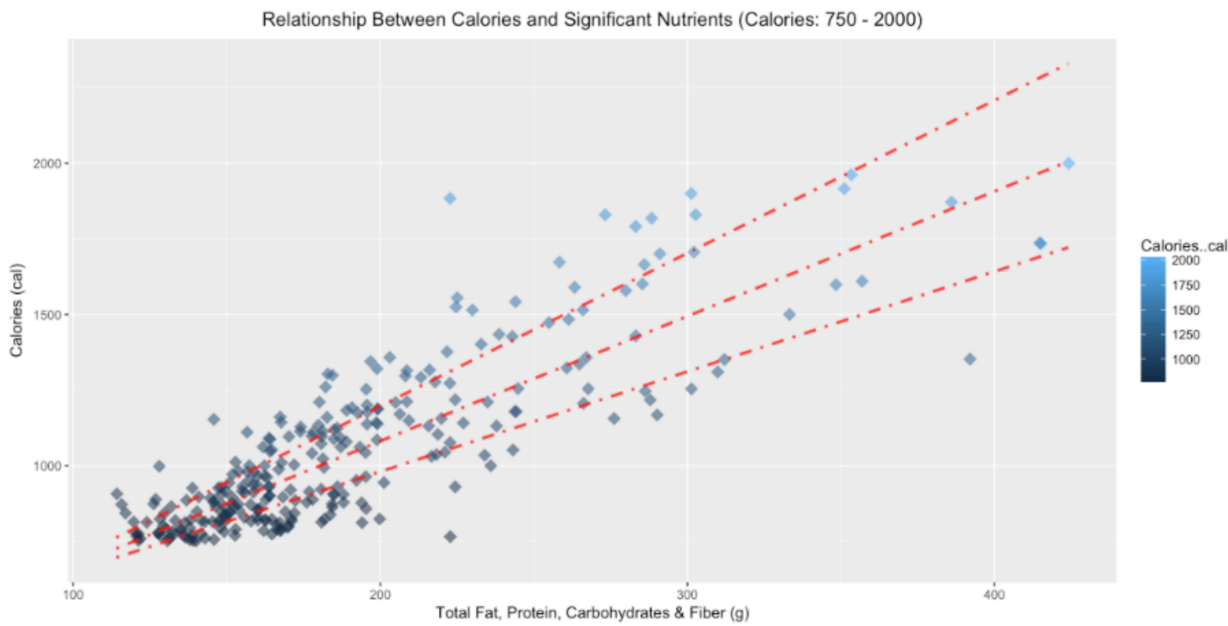
From the above output, p-value < 0.05 indicates that we have to reject the null hypothesis Ho.

There is a strong positive (+) linear relationship between the calories and above significant nutrients (fat, protein, fiber and carbohydrates).

The linear relationship for the above analysis is given by the following equation:

$$Y = 0.414449 + 8.852269 * V + 4.406942 * W + 4.128899 * X - 2.536950 * Z + \epsilon$$

Y: Calories (cal), V: Fat (g), W: Protein (g), X: Carbohydrates (g), Z: Fiber (g)



Finding: An average increase of 10g in fat, protein, carbohydrates and fiber will likely result in an average increase in calories by 149.16 cal for the given recipes. This relationship can be used to optimize the total calories in the recipes by altering the proportions of their significant nutrients.

Note: The dotted lines indicate the first, second and third quantiles of the data points in the graph.

<u>**Studying the relationship between Saturated Fat and Cholesterol for Spoonacular data**</u>

A. Calorie Subgroup: 0 - 250 cal:

spoon_g1_lm1 <- lm(Saturated.Fat..g. ~ Cholesterol..mg., data = spoon_g1)

summary(spoon_g1_lm1)

```
ggplot(data = spoon_g1, mapping = aes(x=Cholesterol..mg., y=Saturated.Fat..g., col=Saturated.Fat..g.)) +
  geom_jitter(col="darkgreen",size=4,shape=15,alpha=0.6) +
  geom_quantile(color = "black",size=1,linetype=4,alpha=0.8) +
  ggtitle("Relationship Between Saturated Fat and Cholesterol (Calories: 0 - 250)") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Cholesterol (mg)") +
  ylab("Saturated Fat (g)")
##
## Call:
## lm(formula = Saturated.Fat..g. ~ Cholesterol..mg., data = spoon_g1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7420 -1.5932 -0.6282  0.8771 10.9872
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.568228   0.126999  20.222  < 2e-16 ***
## Cholesterol..mg. 0.009637   0.001893   5.089  5.3e-07 ***
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##

Residual standard error: 2.393 on 448 degrees of freedom

Multiple R-squared: 0.05466,   Adjusted R-squared: 0.05255

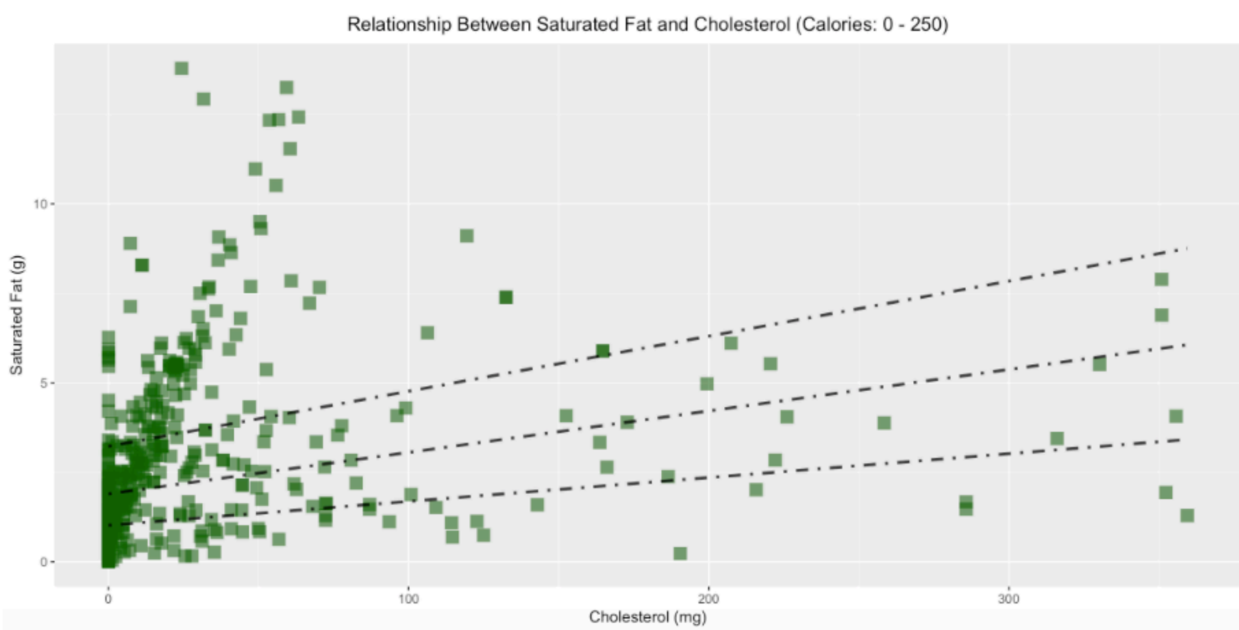
F-statistic: 25.9 on 1 and 448 DF, p-value: 5.295e-07

From the above output, p-value < 0.05 indicates that we have to reject the null hypothesis Ho.

We observe a strong positive (+) linear correlation between saturated fat and cholesterol.

The linear relationship for the above analysis is given by the following equation:

$Y = 2.568228 + 0.009637 * X + \epsilon$;   X: Cholesterol (mg), Y: Saturated Fat (g)



Finding: The saturated fat is estimated to increase by an average of 0.09637g for every 10mg average increase in cholesterol. Saturated fat contains a high proportion of low-density lipoprotein (LDL) cholesterol, which is a leading cause of heart disease whilst a person has high triglycerides (sugar). The average increase in saturated fat is very less for the average increase in cholesterol. This suggests that the proportion of the LDL cholesterol is comparatively lesser than the proportion of high-density lipoprotein (HDL) cholesterol and other cholesterol in the recipes. The HDL cholesterol is the good cholesterol and must be maximized in a person's lipid profile to prevent heart disease, whereas the LDL cholesterol is the culprit and must be minimized.

Note: The dotted lines indicate the first, second and third quantiles of the data points in the graph.

B. Calorie Subgroup: 250 - 500 cal:

```
spoon_g2_lm1 <- lm(Saturated.Fat..g. ~ Cholesterol..mg., data = spoon_g2)

summary(spoon_g2_lm1)

ggplot(data = spoon_g2, mapping = aes(x=Cholesterol..mg., y=Saturated.Fat..g., col=Saturated.Fat..g.)) +

  geom_jitter(col="darkgreen",size=4,shape=16,alpha=0.6) +

  geom_quantile(color = "black",size=1,linetype=4,alpha=0.8) +

  ggtitle("Relationship Between Saturated Fat and Cholesterol (Calories: 250 - 500)") +

  theme(plot.title = element_text(hjust = 0.5)) +

  xlab("Cholesterol (mg)") +

  ylab("Saturated Fat (g)”)

##

## Call:
```

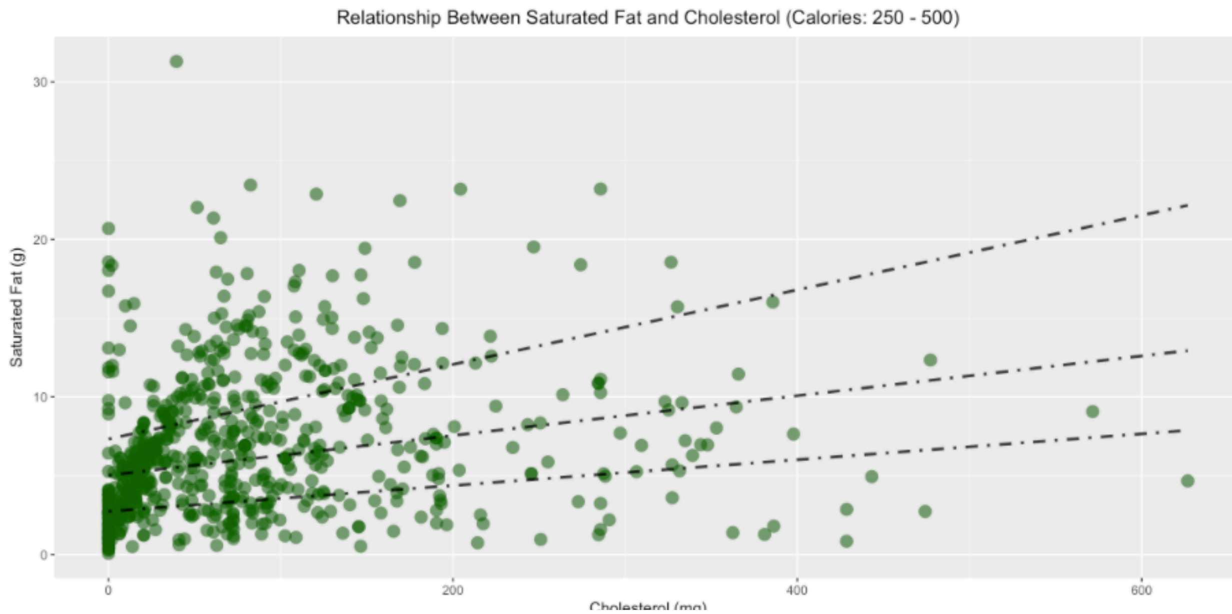
```
## lm(formula = Saturated.Fat..g. ~ Cholesterol..mg., data = spoon_g2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.4927 -3.3130 -0.7575  2.2460 24.9597
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.935490   0.227342  26.108  < 2e-16 ***
## Cholesterol..mg. 0.010259   0.001859   5.518 4.81e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.515 on 704 degrees of freedom
## Multiple R-squared:  0.04146,    Adjusted R-squared:  0.0401
## F-statistic: 30.45 on 1 and 704 DF,  p-value: 4.815e-08
```

From the above output, p-value < 0.05 indicates that we have to reject the null hypothesis Ho.

We observe a strong positive (+) linear correlation between saturated fat and cholesterol.

The linear relationship for the above analysis is given by the following equation:

$Y = 5.935490 + 0.010259 * X + \epsilon$;   X: Cholesterol (mg), Y: Saturated Fat (g)



Finding: The saturated fat is estimated to increase by an average of 0.10259g for every 10mg average increase in cholesterol. Saturated fat contains a high proportion of low-density lipoprotein (LDL) cholesterol, which is a leading cause of heart disease whilst a person has high triglycerides (sugar). The average increase in saturated fat is very less for the average increase in cholesterol. This suggests that the proportion of the LDL cholesterol is comparatively lesser than the proportion of high-density lipoprotein (HDL) cholesterol and other cholesterol in the recipes. The HDL cholesterol is the good cholesterol and must be maximized in a person's lipid profile to prevent heart disease, whereas the LDL cholesterol is the culprit and must be minimized.

Note: The dotted lines indicate the first, second and third quantiles of the data points in the graph.

C. Calorie Subgroup: 500 - 750 cal:

spoon_g3_lm1 <- lm(Saturated.Fat..g. ~ Cholesterol..mg., data = spoon_g3)

```
summary(spoon_g3_lm1)

ggplot(data = spoon_g3, mapping = aes(x=Cholesterol..mg., y=Saturated.Fat..g., col=Saturated.Fat..g.)) +
  geom_jitter(col="darkgreen",size=4,shape=17,alpha=0.6) +
  geom_quantile(color = "black",size=1,linetype=4,alpha=0.8) +
  ggtitle("Relationship Between Saturated Fat and Cholesterol (Calories: 500 - 750)") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Cholesterol (mg)") +
  ylab("Saturated Fat (g)”)
##
## Call:
## lm(formula = Saturated.Fat..g. ~ Cholesterol..mg., data = spoon_g3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.447  -5.154  -1.380   4.294  45.426
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      10.994214   0.480103  22.900   <2e-16 ***
```

Cholesterol..mg. 0.006188   0.002605   2.375   0.018 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##

Residual standard error: 6.95 on 421 degrees of freedom

Multiple R-squared: 0.01322,   Adjusted R-squared: 0.01088

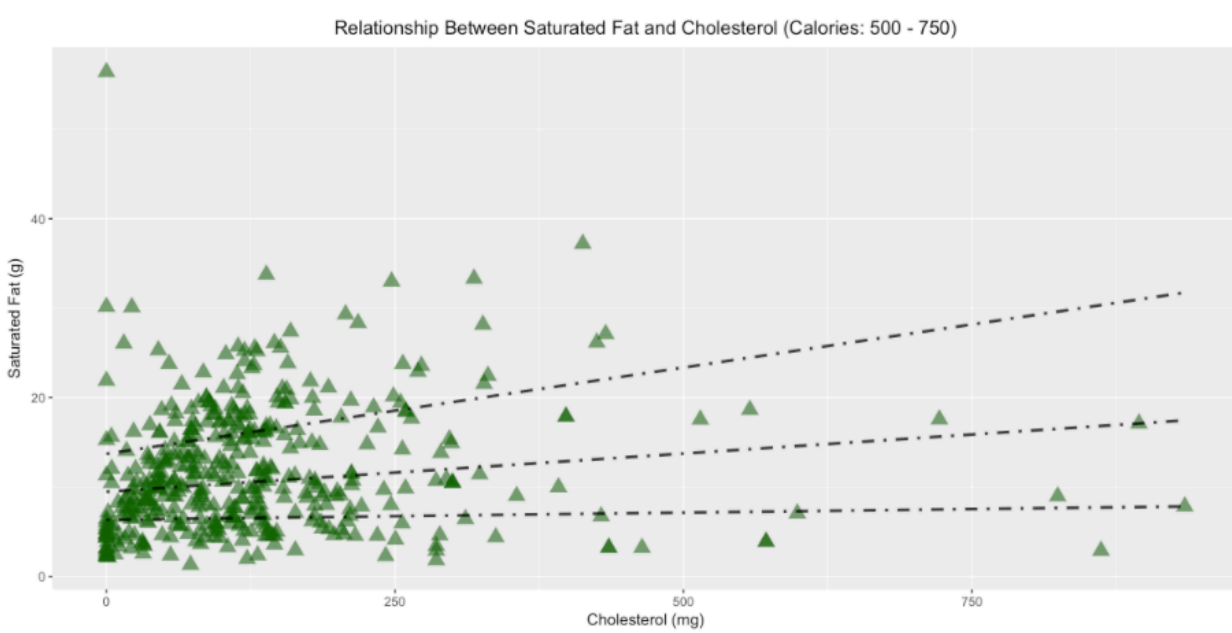
F-statistic: 5.642 on 1 and 421 DF, p-value: 0.01798

From the above output, p-value < 0.05 indicates that we have to reject the null hypothesis Ho.

We observe a strong positive (+) linear correlation between saturated fat and cholesterol.

The linear relationship for the above analysis is given by the following equation:

$Y = 10.994214 + 0.006188 * X + \epsilon$; X: Cholesterol (mg), Y: Saturated Fat (g)



Finding: The saturated fat is estimated to increase by an average of 0.06188g for every 10mg average increase in cholesterol. Saturated fat contains a high proportion of low-density lipoprotein (LDL) cholesterol, which is a leading cause of heart disease whilst a person has high triglycerides (sugar). The average increase in saturated fat is very less for the average increase in cholesterol. This suggests that the proportion of the LDL cholesterol is comparatively lesser than the proportion of high-density lipoprotein (HDL) cholesterol and other cholesterol in the recipes. The HDL cholesterol is the good cholesterol and must be maximized in a person's lipid profile to prevent heart disease, whereas the LDL cholesterol is the culprit and must be minimized.

Note: The dotted lines indicate the first, second and third quantiles of the data points in the graph.

D. Calorie Subgroup: 750 - 2000 cal:

```
spoon_g4_lm1 <- lm(Saturated.Fat..g. ~ Cholesterol..mg., data = spoon_g4)

summary(spoon_g4_lm1)

ggplot(data = spoon_g4, mapping = aes(x=Cholesterol..mg., y=Saturated.Fat..g., col=Saturated.Fat..g.)) +

  geom_jitter(col="darkgreen",size=4,shape=18,alpha=0.6) +

  geom_quantile(color = "black",size=1,linetype=4,alpha=0.8) +

  ggtitle("Relationship Between Saturated Fat and Cholesterol (Calories: 750 - 2000)") +

  theme(plot.title = element_text(hjust = 0.5)) +

  xlab("Cholesterol (mg)") +

  ylab("Saturated Fat (g)")

##

## Call:
```

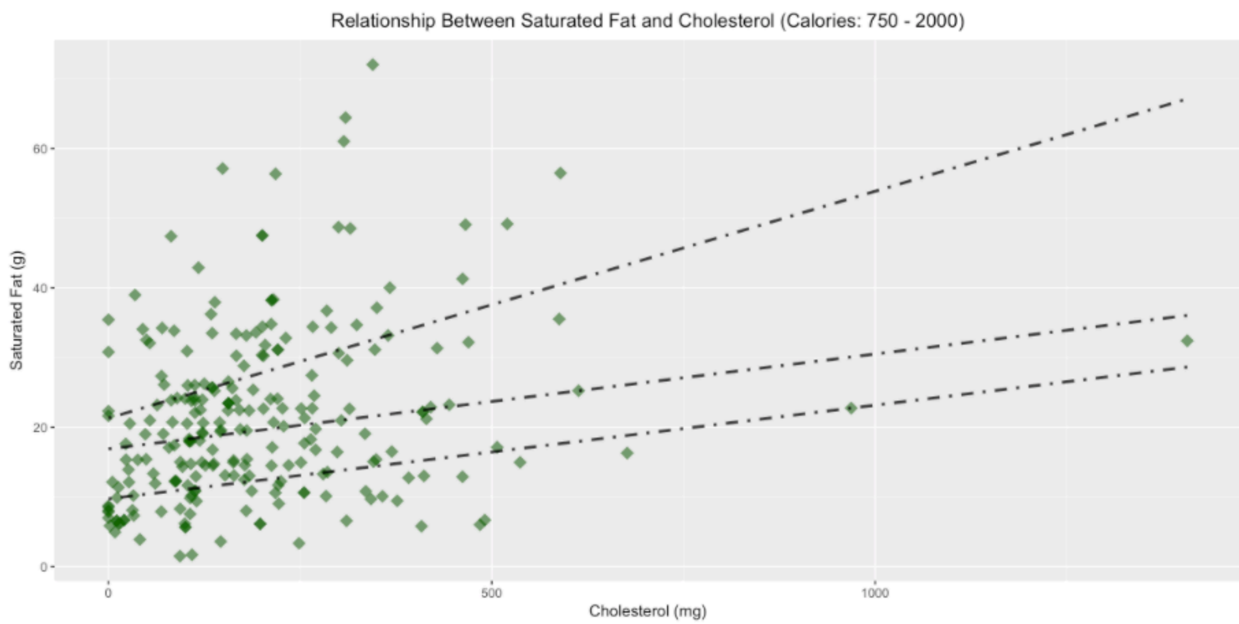
```
## lm(formula = Saturated.Fat..g. ~ Cholesterol..mg., data = spoon_g4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.922  -8.982  -1.727   6.082  47.824
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      17.592968   1.219576  14.425  < 2e-16 ***
## Cholesterol..mg.  0.019270   0.004765   4.044 7.18e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.99 on 230 degrees of freedom
## Multiple R-squared:  0.06637,    Adjusted R-squared:  0.06231
## F-statistic: 16.35 on 1 and 230 DF,  p-value: 7.184e-05
```

From the above output, p-value < 0.05 indicates that we have to reject the null hypothesis Ho.

We observe a strong positive (+) linear correlation between saturated fat and cholesterol.

The linear relationship for the above analysis is given by the following equation:

Y = 17.592968 + 0.019270 * X + ϵ ; X: Cholesterol (mg), Y: Saturated Fat (g)



Finding: The saturated fat is estimated to increase by an average of 0.06188g for every 10mg average increase in cholesterol. Saturated fat contains a high proportion of low-density lipoprotein (LDL) cholesterol, which is a leading cause of heart disease whilst a person has high triglycerides (sugar). The average increase in saturated fat is very less for the average increase in cholesterol. This suggests that the proportion of the LDL cholesterol is comparatively lesser than the proportion of high-density lipoprotein (HDL) cholesterol and other cholesterol in the recipes. The HDL cholesterol is the good cholesterol and must be maximized in a person's lipid profile to prevent heart disease, whereas the LDL cholesterol is the culprit and must be minimized.

Note: The dotted lines indicate the first, second and third quantiles of the data points in the graph.

<u>Studying the relationship between Saturated Fat and Sodium for Spoonacular data</u>

A. Calorie Subgroup: 0 - 250 cal:

```
spoon_g1_lm2 <- lm(Saturated.Fat..g. ~ Sodium..mg., data = spoon_g1)

summary(spoon_g1_lm2)

ggplot(data = spoon_g1, mapping = aes(x=Sodium..mg., y=Saturated.Fat..g., col=Saturated.Fat..g.)) +
  geom_jitter(col="darkgreen",size=4,shape=15,alpha=0.6) +
  geom_quantile(color = "black",size=1,linetype=4,alpha=0.8) +
  ggtitle("Relationship Between Saturated Fat and Sodium (Calories: 0 - 250)") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Sodium (mg)") +
  ylab("Saturated Fat (g)")

##
## Call:
## lm(formula = Saturated.Fat..g. ~ Sodium..mg., data = spoon_g1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8614 -1.7362 -0.7632  1.0592 10.9260
##
```

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.861e+00 1.415e-01 20.224   <2e-16 ***

Sodium..mg. 8.993e-06 1.745e-04   0.052   0.959

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##

Residual standard error: 2.461 on 448 degrees of freedom

Multiple R-squared: 5.929e-06, Adjusted R-squared: -0.002226

F-statistic: 0.002656 on 1 and 448 DF, p-value: 0.9589

From the above output, p-value < 0.05 indicates that we have to reject the null hypothesis Ho.

There is a slight positive (+) correlation observed between saturated fat and sodium but this is not strong. Hence, we can ignore this finding.

B. Calorie Subgroup: 250 - 500 cal:

spoon_g2_lm2 <- lm(Saturated.Fat..g. ~ Sodium..mg., data = spoon_g2)

summary(spoon_g2_lm2)

ggplot(data = spoon_g2, mapping = aes(x=Sodium..mg., y=Saturated.Fat..g., col=Saturated.Fat..g.)) +

  geom_jitter(col="darkgreen",size=4,shape=16,alpha=0.6) +

  geom_quantile(color = "black",size=1,linetype=4,alpha=0.8) +

```
  ggtitle("Relationship Between Saturated Fat and Sodium (Calories: 250 - 500)") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Sodium (mg)") +
  ylab("Saturated Fat (g)”)
##
## Call:
## lm(formula = Saturated.Fat..g. ~ Sodium..mg., data = spoon_g2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7024 -3.5001 -0.8445  2.5141 24.5358
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.763e+00  1.743e-01  38.803   <2e-16 ***
## Sodium..mg. 4.319e-06  1.163e-05   0.371     0.71
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.611 on 704 degrees of freedom
```

Multiple R-squared: 0.000196,   Adjusted R-squared: -0.001224

F-statistic: 0.138 on 1 and 704 DF, p-value: 0.7104

From the above output, p-value < 0.05 indicates that we have to reject the null hypothesis Ho.

There is a slight positive (+) correlation observed between saturated fat and sodium but this is not strong. Hence, we can ignore this finding.

C. Calorie Subgroup: 500 - 750 cal:

```
spoon_g3_lm2 <- lm(Saturated.Fat..g. ~ Sodium..mg., data = spoon_g3)

summary(spoon_g3_lm2)

ggplot(data = spoon_g3, mapping = aes(x=Sodium..mg., y=Saturated.Fat..g., col=Saturated.Fat..g.)) +

  geom_jitter(col="darkgreen",size=4,shape=17,alpha=0.6) +

  geom_quantile(color = "black",size=1,linetype=4,alpha=0.8) +

  ggtitle("Relationship Between Saturated Fat and Sodium (Calories: 500 - 750)") +

  theme(plot.title = element_text(hjust = 0.5)) +

  xlab("Sodium (mg)") +

  ylab("Saturated Fat (g)”)

##

## Call:

## lm(formula = Saturated.Fat..g. ~ Sodium..mg., data = spoon_g3)

##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.385  -5.212  -1.364   4.387  44.613
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.197e+01  3.594e-01  33.298   <2e-16 ***
## Sodium..mg. -1.072e-04  7.761e-05  -1.381    0.168
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.98 on 421 degrees of freedom
## Multiple R-squared:  0.004512,   Adjusted R-squared:  0.002147
## F-statistic: 1.908 on 1 and 421 DF,  p-value: 0.1679
```

From the above output, p-value < 0.05 indicates that we have to reject the null hypothesis Ho.

There is a slight negative (-) correlation observed between saturated fat and sodium but this is not strong. Hence, we can ignore this finding.

D. Calorie Subgroup: 750 - 2000 cal:

```
spoon_g4_lm2 <- lm(Saturated.Fat..g. ~ Sodium..mg., data = spoon_g4)
```

```
summary(spoon_g4_lm2)

ggplot(data = spoon_g4, mapping = aes(x=Sodium..mg., y=Saturated.Fat..g., col=Saturated.Fat..g.)) +
  geom_jitter(col="darkgreen",size=4,shape=18,alpha=0.6) +
  geom_quantile(color = "black",size=1,linetype=4,alpha=0.8) +
  ggtitle("Relationship Between Saturated Fat and Sodium (Calories: 750 - 2000)") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Sodium (mg)") +
  ylab("Saturated Fat (g)”)

##
## Call:
## lm(formula = Saturated.Fat..g. ~ Sodium..mg., data = spoon_g4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.066  -8.928  -1.838   6.092  50.567
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.7043859  0.9964516  21.782   <2e-16 ***
```

```
## Sodium..mg. -0.0001560  0.0002591  -0.602    0.548

## ---

## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 12.4 on 230 degrees of freedom

## Multiple R-squared:  0.001572,   Adjusted R-squared:  -0.002769

## F-statistic: 0.3622 on 1 and 230 DF,  p-value: 0.5479
```

From the above output, p-value < 0.05 indicates that we have to reject the null hypothesis Ho.

There is a slight negative (-) correlation observed between saturated fat and sodium but this is not strong. Hence, we can ignore this finding.

<u>Studying the relationship between Total Fat and Carbohydrates for Spoonacular data</u>

A. Calorie Subgroup: 0 - 250 cal:

```
spoon_g1_lm3 <- lm(Fat..g. ~ Carbohydrates..g., data = spoon_g1)

summary(spoon_g1_lm3)

ggplot(data = spoon_g1, mapping = aes(x=Carbohydrates..g., y=Fat..g., col=Fat..g.)) +

  geom_jitter(col="darkgreen",size=4,shape=15,alpha=0.6) +

  geom_quantile(color = "black",size=1,linetype=4,alpha=0.8) +

  ggtitle("Relationship Between Total Fat and Carbohydrates (Calories: 0 - 250)") +
```

```
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Carbohydrates (g)") +
  ylab("Total Fat (g)”)
##
## Call:
## lm(formula = Fat..g. ~ Carbohydrates..g., data = spoon_g1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3451  -3.7702  -0.3502   3.5271  14.8098
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       10.75364    0.44625   24.10  < 2e-16 ***
## Carbohydrates..g. -0.09867    0.02543   -3.88  0.00012 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.077 on 448 degrees of freedom
## Multiple R-squared:  0.03251,    Adjusted R-squared:  0.03035
```

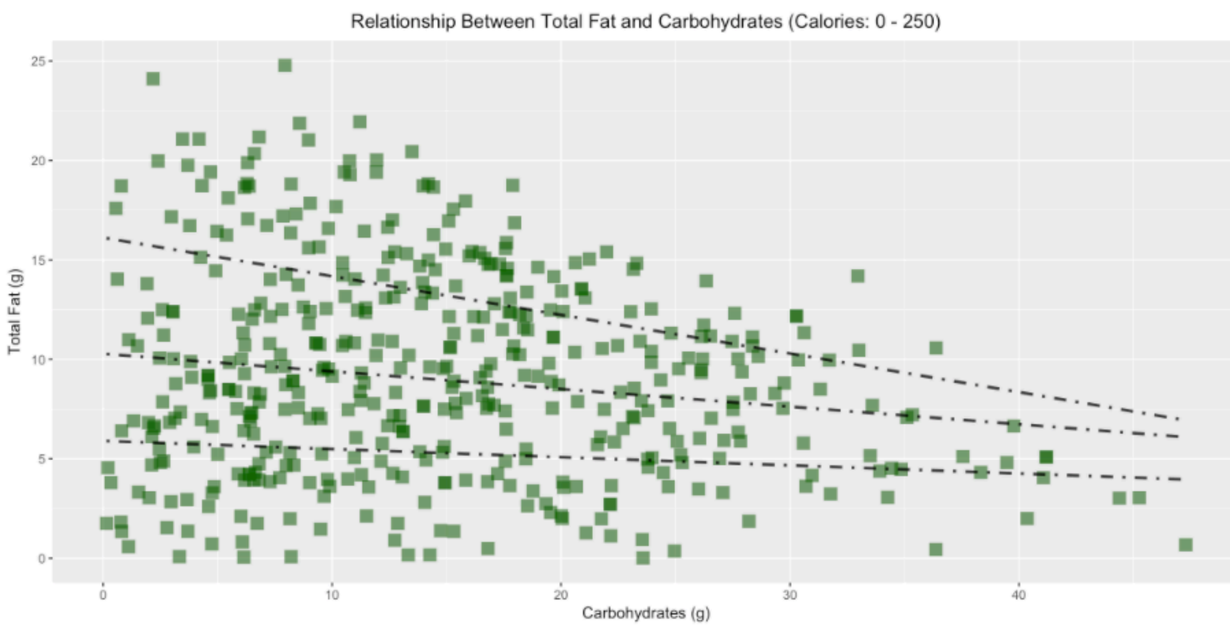
F-statistic: 15.05 on 1 and 448 DF, p-value: 0.0001201

From the above output, p-value < 0.05 indicates that we have to reject the null hypothesis Ho.

We observe a strong negative (-) linear correlation between Total Fat and Carbohydrates.

The linear relationship for the above analysis is given by the following equation:

$Y = 10.75364 - 0.09867 * X + \epsilon$;   X: Fat (g), Y: Carbohydrates (g)



Finding: The total fat is estimated to decrease by an average of 0.9867g for every 10g average increase in carbohydrates. We see that the total fat to carbohydrate ratio of the recipes is well-balanced with the help of this linear relation. This indicates that a recipe having high total fat would likely have low carbohydrates and vice-versa.

Note: The dotted lines indicate the first, second and third quantiles of the data points in the graph.

B. Calorie Subgroup: 250 - 500 cal:

```
spoon_g2_lm3 <- lm(Fat..g. ~ Carbohydrates..g., data = spoon_g2)

summary(spoon_g2_lm3)

ggplot(data = spoon_g2, mapping = aes(x=Carbohydrates..g., y=Fat..g., col=Fat..g.)) +

  geom_jitter(col="darkgreen",size=4,shape=16,alpha=0.6) +

  geom_quantile(color = "black",size=1,linetype=4,alpha=0.8) +

  ggtitle("Relationship Between Total Fat and Carbohydrates (Calories: 250 - 500)") +

  theme(plot.title = element_text(hjust = 0.5)) +

  xlab("Carbohydrates (g)") +

  ylab("Total Fat (g)”)

##

## Call:

## lm(formula = Fat..g. ~ Carbohydrates..g., data = spoon_g2)

##

## Residuals:

##      Min       1Q   Median       3Q      Max

## -17.8839  -5.4495  -0.5054   5.1286  19.6028

##

## Coefficients:

##                   Estimate Std. Error t value Pr(>|t|)
```

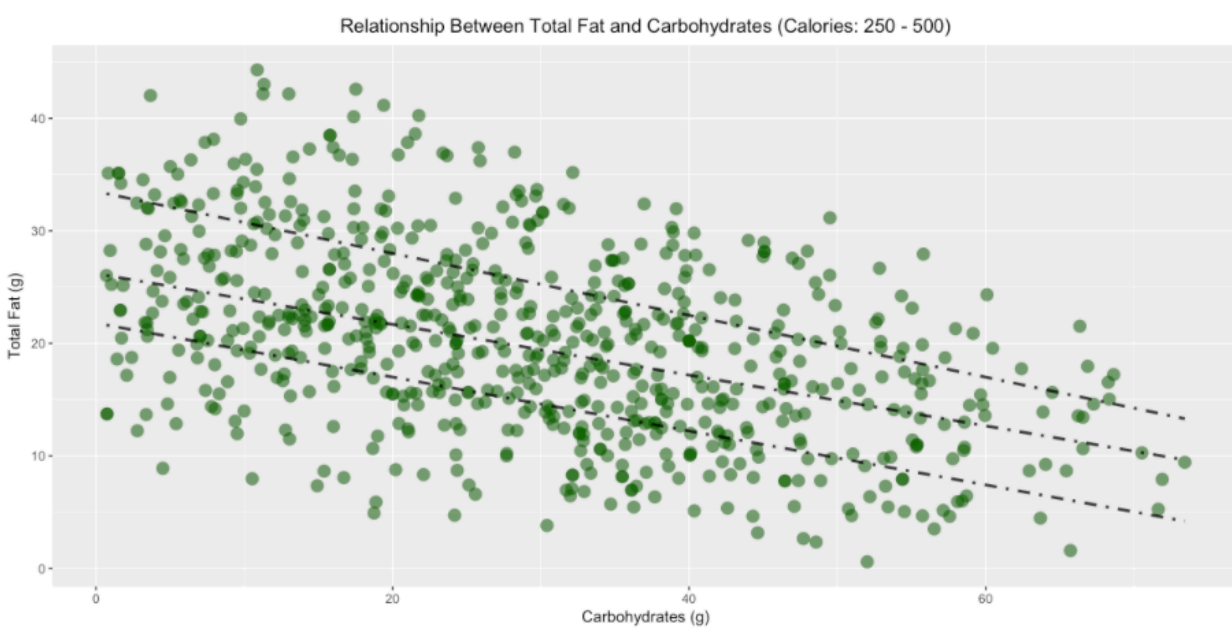
```
## (Intercept)       27.34660    0.57535   47.53   <2e-16 ***
## Carbohydrates..g. -0.24281    0.01711  -14.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.41 on 704 degrees of freedom
## Multiple R-squared:  0.2224, Adjusted R-squared:  0.2213
## F-statistic: 201.4 on 1 and 704 DF,  p-value: < 2.2e-16
```

From the above output, p-value < 0.05 indicates that we have to reject the null hypothesis Ho.

We observe a strong negative (-) linear correlation between Total Fat and Carbohydrates.

The linear relationship for the above analysis is given by the following equation:

$Y = 27.34660 - 0.24281 * X + \epsilon$;   X: Fat (g), Y: Carbohydrates (g)



Finding: The total fat is estimated to decrease by an average of 2.4281g for every 10g average increase in carbohydrates. We see that the total fat to carbohydrate ratio of the recipes is well-balanced with the help of this linear relation. This indicates that a recipe having high total fat would likely have low carbohydrates and vice-versa.

Note: The dotted lines indicate the first, second and third quantiles of the data points in the graph.

C. Calorie Subgroup: 500 - 750 cal:

```
spoon_g3_lm3 <- lm(Fat..g. ~ Carbohydrates..g., data = spoon_g3)

summary(spoon_g3_lm3)

ggplot(data = spoon_g3, mapping = aes(x=Carbohydrates..g., y=Fat..g., col=Fat..g.)) +
  geom_jitter(col="darkgreen",size=4,shape=17,alpha=0.6) +
  geom_quantile(color = "black",size=1,linetype=4,alpha=0.8) +
  ggtitle("Relationship Between Total Fat and Carbohydrates (Calories: 500 - 750)") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Carbohydrates (g)") +
  ylab("Total Fat (g)”)

##
## Call:
## lm(formula = Fat..g. ~ Carbohydrates..g., data = spoon_g3)
##
## Residuals:
```

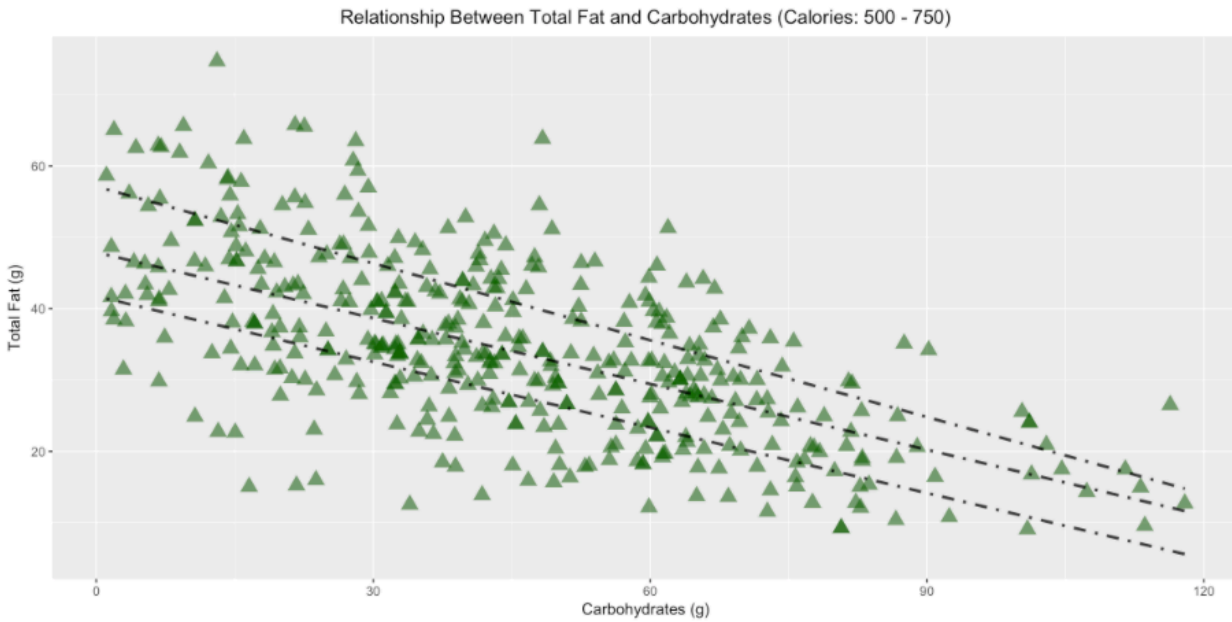
```
##      Min       1Q   Median       3Q      Max
## -28.4685  -6.5604  -0.0777   6.9043  30.5426
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       48.82247    0.98462   49.59   <2e-16 ***
## Carbohydrates..g. -0.32089    0.01927  -16.65   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.589 on 421 degrees of freedom
## Multiple R-squared:  0.3972, Adjusted R-squared:  0.3957
## F-statistic: 277.4 on 1 and 421 DF,  p-value: < 2.2e-16
```

From the above output, p-value < 0.05 indicates that we have to reject the null hypothesis Ho.

We observe a strong negative (-) linear correlation between Total Fat and Carbohydrates.

The linear relationship for the above analysis is given by the following equation:

$Y = 48.82247 - 0.32089 * X + \epsilon$;   X: Fat (g), Y: Carbohydrates (g)



Finding: The total fat is estimated to decrease by an average of 3.2089g for every 10g average increase in carbohydrates. We see that the total fat to carbohydrate ratio of the recipes is well-balanced with the help of this linear relation. This indicates that a recipe having high total fat would likely have low carbohydrates and vice-versa.

Note: The dotted lines indicate the first, second and third quantiles of the data points in the graph.

D. Calorie Subgroup: 750 - 2000 cal:

```
spoon_g4_lm3 <- lm(Fat..g. ~ Carbohydrates..g., data = spoon_g4)

summary(spoon_g4_lm3)

ggplot(data = spoon_g4, mapping = aes(x=Carbohydrates..g., y=Fat..g., col=Fat..g.)) +
  geom_jitter(col="darkgreen",size=4,shape=18,alpha=0.6) +
  geom_quantile(color = "black",size=1,linetype=4,alpha=0.8) +
  ggtitle("Relationship Between Total Fat and Carbohydrates (Calories: 750 - 2000)") +
  theme(plot.title = element_text(hjust = 0.5)) +
```

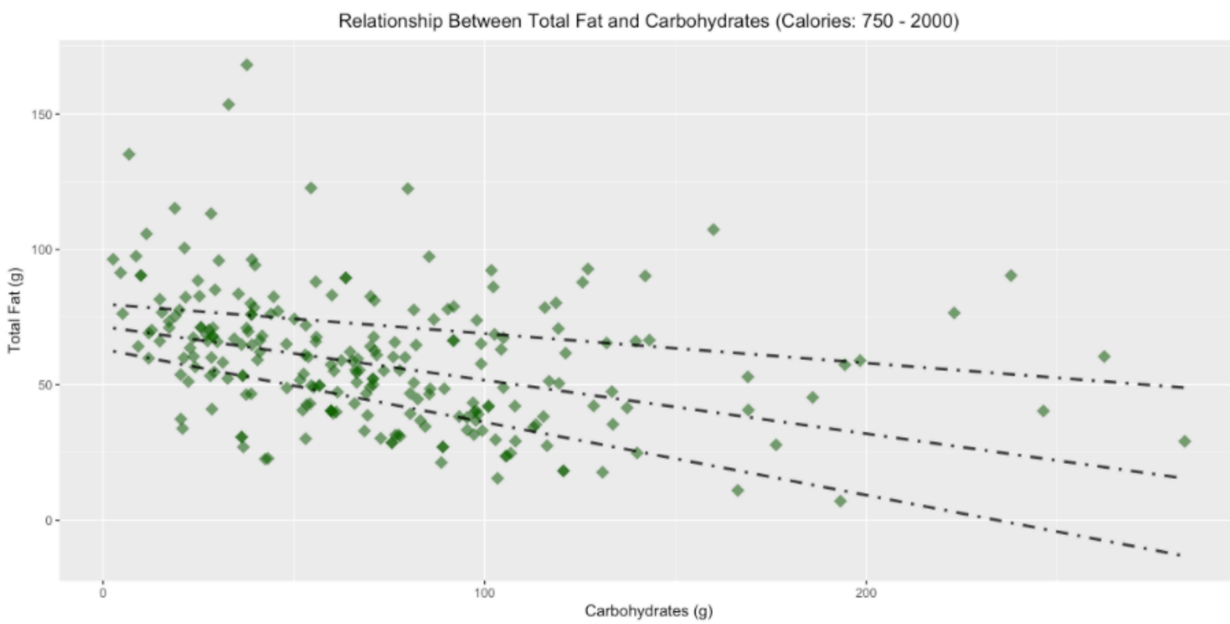
```
  xlab("Carbohydrates (g)") +
  ylab("Total Fat (g)")
##
## Call:
## lm(formula = Fat..g. ~ Carbohydrates..g., data = spoon_g4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.548 -15.039  -2.005  12.436 103.250
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       70.96911    2.73371  25.961  < 2e-16 ***
## Carbohydrates..g. -0.16019    0.03132  -5.114 6.63e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.17 on 230 degrees of freedom
## Multiple R-squared:  0.1021, Adjusted R-squared:  0.09821
## F-statistic: 26.16 on 1 and 230 DF,  p-value: 6.63e-07
```

From the above output, p-value < 0.05 indicates that we have to reject the null hypothesis Ho.

We observe a strong negative (-) linear correlation between Total Fat and Carbohydrates.

The linear relationship for the above analysis is given by the following equation:

$Y = 70.96911 - 0.16019 * X + \epsilon$;   X: Fat (g), Y: Carbohydrates (g)



Finding: The total fat is estimated to decrease by an average of 1.6019g for every 10g average increase in carbohydrates. We see that the total fat to carbohydrate ratio of the recipes is well-balanced with the help of this linear relation. This indicates that a recipe having high total fat would likely have low carbohydrates and vice-versa.

Note: The dotted lines indicate the first, second and third quantiles of the data points in the graph.

<u>Studying the relationship between Protein and Fiber for Spoonacular data</u>

A. Calorie Subgroup: 0 - 250 cal:

```
spoon_g1_lm4 <- lm(Protein..g. ~ Fiber..g., data = spoon_g1)

summary(spoon_g1_lm4)

ggplot(data = spoon_g1, mapping = aes(x=Fiber..g., y=Protein..g., col=Protein..g.)) +
  geom_jitter(col="darkgreen",size=4,shape=15,alpha=0.6) +
  geom_quantile(color = "black",size=1,linetype=4,alpha=0.8) +
  ggtitle("Relationship Between Protein and Fiber (Calories: 0 - 250)") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Fiber (g)") +
  ylab("Protein (g)")

##
## Call:
## lm(formula = Protein..g. ~ Fiber..g., data = spoon_g1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.502 -3.817 -1.309  2.130 32.856
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.6120     0.4561  16.688   <2e-16 ***
## Fiber..g.    -0.2307     0.1309  -1.762   0.0787 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.731 on 448 degrees of freedom
## Multiple R-squared:  0.006884,   Adjusted R-squared:  0.004667
## F-statistic: 3.106 on 1 and 448 DF,  p-value: 0.07871
```

From the above output, p-value > 0.05 indicates that we have to accept the null hypothesis Ho.

There is no correlation observed between protein and fibre. This is bad because it indicates that the protein to fiber ratio is not well-balanced in most of the recipes. Ideally, protein and fibre should be positively correlated in a healthy meal.

B. Calorie Subgroup: 250 - 500 cal:

```
spoon_g2_lm4 <- lm(Protein..g. ~ Fiber..g., data = spoon_g2)
summary(spoon_g2_lm4)
ggplot(data = spoon_g2, mapping = aes(x=Fiber..g., y=Protein..g., col=factor(Protein..g.))) +
  geom_jitter(col="darkgreen",size=4,shape=16,alpha=0.6) +
  geom_quantile(color = "black",size=1,linetype=4,alpha=0.8) +
  ggtitle("Relationship Between Protein and Fiber (Calories: 250 - 500)") +
```

```
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Fiber (g)") +
  ylab("Protein (g)")
##
## Call:
## lm(formula = Protein..g. ~ Fiber..g., data = spoon_g2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.308  -7.575  -2.163   6.527  47.566
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  19.6635     0.6096  32.256  < 2e-16 ***
## Fiber..g.    -0.2871     0.1022  -2.808  0.00512 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.1 on 704 degrees of freedom
## Multiple R-squared:  0.01108,    Adjusted R-squared:  0.009674
```

F-statistic: 7.887 on 1 and 704 DF, p-value: 0.005118

From the above output, p-value < 0.05 indicates that we have to reject the null hypothesis Ho.

There is a slight negative (-) correlation observed between protein and fibre but it is not strong. Hence, we can ignore this finding. This finding is bad because it indicates that the protein to fiber ratio is not well-balanced in most of the recipes. Ideally, protein and fibre should be positively correlated in a healthy meal.

C. Calorie Subgroup: 500 - 750 cal:

```
spoon_g3_lm4 <- lm(Protein..g. ~ Fiber..g., data = spoon_g3)

summary(spoon_g3_lm4)

ggplot(data = spoon_g3, mapping = aes(x=Fiber..g., y=Protein..g., col=Protein..g.)) +
  geom_jitter(col="darkgreen",size=4,shape=17,alpha=0.6) +
  geom_quantile(color = "black",size=1,linetype=4,alpha=0.8) +
  ggtitle("Relationship Between Protein and Fiber (Calories: 500 - 750)") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Fiber (g)") +
  ylab("Protein (g)”)
```

```
##
## Call:
## lm(formula = Protein..g. ~ Fiber..g., data = spoon_g3)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max

## -26.446 -10.545  -1.460   8.141  53.009

##

## Coefficients:

##             Estimate Std. Error t value Pr(>|t|)

## (Intercept)  31.6236     1.1130  28.414   <2e-16 ***

## Fiber..g.    -0.2976     0.1498  -1.987   0.0476 *

## ---

## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 14.34 on 421 degrees of freedom

## Multiple R-squared:  0.009287,   Adjusted R-squared:  0.006934

## F-statistic: 3.946 on 1 and 421 DF,  p-value: 0.04762
```

From the above output, p-value < 0.05 indicates that we have to reject the null hypothesis Ho.

There is a slight negative (-) correlation observed between protein and fibre but it is not strong. Hence, we can ignore this finding. This finding is bad because it indicates that the protein to fiber ratio is not well-balanced in most of the recipes. Ideally, protein and fibre should be positively correlated in a healthy meal.

D. Calorie Subgroup: 750 - 2000 cal

```
spoon_g4_lm4 <- lm(Protein..g. ~ Fiber..g., data = spoon_g4)
```

```
summary(spoon_g4_lm4)

ggplot(data = spoon_g4, mapping = aes(x=Fiber..g., y=Protein..g., col=Protein..g.)) +

  geom_jitter(col="darkgreen",size=4,shape=18,alpha=0.6) +

  geom_quantile(color = "black",size=1,linetype=4,alpha=0.8) +

  ggtitle("Relationship Between Protein and Fiber (Calories: 750 - 2000)") +

  theme(plot.title = element_text(hjust = 0.5)) +

  xlab("Fiber (g)") +

  ylab("Protein (g)”)

##

## Call:

## lm(formula = Protein..g. ~ Fiber..g., data = spoon_g4)

##

## Residuals:

##     Min      1Q  Median      3Q     Max

## -38.025 -17.359  -5.796  10.195 145.657

##

## Coefficients:

##             Estimate Std. Error t value Pr(>|t|)

## (Intercept)  51.7924     2.6407  19.613   <2e-16 ***

## Fiber..g.    -0.2318     0.2195  -1.056    0.292
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##

Residual standard error: 27.42 on 230 degrees of freedom

Multiple R-squared: 0.004826,   Adjusted R-squared: 0.0004989

F-statistic: 1.115 on 1 and 230 DF, p-value: 0.292

From the above output, p-value > 0.05 indicates that we have to accept the null hypothesis Ho.

There is no correlation observed between protein and fibre. This is bad because it indicates that the protein to fiber ratio is not well-balanced in most of the recipes. Ideally, protein and fibre should be positively correlated in a healthy meal.

<u>Studying the relationship between Calories and Significant Nutrients for Spoonacular data</u>

A. Calorie Subgroup: 0 - 250 cal:

spoon_g1_lm5 <- lm(Calories..cal. ~ (Fat..g.+Saturated.Fat..g.+Protein..g.+Carbohydrates..g.+Fiber..g.+Sodium..mg.+Cholesterol..mg.), data = spoon_g1)

summary(spoon_g1_lm5)

ggplot(data = spoon_g1, mapping = aes(x=(Fat..g.+Protein..g.+Carbohydrates..g.+Fiber..g.), y=Calories..cal., col=Calories..cal.)) +

  geom_jitter(size=4,shape=15,alpha=0.6) +

  geom_quantile(color = "red",size=1,linetype=4,alpha=0.8) +

```
  ggtitle("Relationship Between Calories and Significant Nutrients (Calories: 0 - 250)") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Total Fat, Protein, Carbohydrates & Fiber (g)") +
  ylab("Calories (cal)")
##
## Call:
## lm(formula = Calories..cal. ~ (Fat..g. + Saturated.Fat..g. +
##     Protein..g. + Carbohydrates..g. + Fiber..g. + Sodium..mg. +
##     Cholesterol..mg.), data = spoon_g1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.062  -3.337  -1.030   1.828 125.909
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        4.4750109  1.2746567   3.511 0.000493 ***
## Fat..g.            8.5103161  0.1108148  76.798  < 2e-16 ***
## Saturated.Fat..g.  0.0284420  0.2255688   0.126 0.899718
## Protein..g.        4.0520838  0.0923806  43.863  < 2e-16 ***
```

```
## Carbohydrates..g.  3.9330109  0.0557232  70.581  < 2e-16 ***
## Fiber..g.         -1.5325848  0.2604110  -5.885 7.86e-09 ***
## Sodium..mg.       -0.0003045  0.0006342  -0.480 0.631391
## Cholesterol..mg.   0.0210363  0.0095705   2.198 0.028464 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.757 on 442 degrees of freedom
## Multiple R-squared:  0.9773, Adjusted R-squared:  0.9769
## F-statistic:  2718 on 7 and 442 DF,  p-value: < 2.2e-16
```

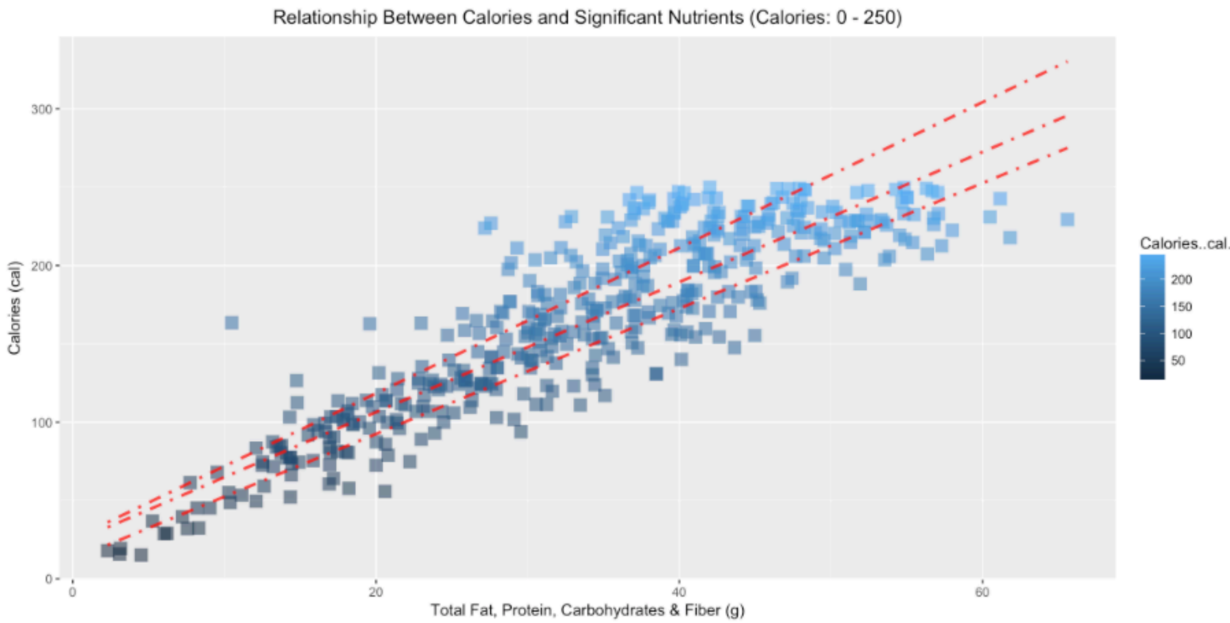
From the above output, p-value < 0.05 indicates that we have to reject the null hypothesis Ho.

There is a strong positive (+) linear relationship between the calories and above significant nutrients (fat, protein, fiber and carbohydrates).

The linear relationship for the above analysis is given by the following equation:

$$Y = 4.4750109 + 8.5103161 * V + 4.0520838 * W + 3.9330109 * X - 1.5325848 * Z + \epsilon$$

Y: Calories (cal), V: Fat (g), W: Protein (g), X: Carbohydrates (g), Z: Fiber (g)



Finding: An average increase of 10g in fat, protein, carbohydrates and fiber will likely result in an average increase in calories by 149.63 cal for the given recipes. This relationship can be used to optimize the total calories in the recipes by altering the proportions of their significant nutrients.

Note: The dotted lines indicate the first, second and third quantiles of the data points in the graph.

B. Calorie Subgroup: 250 - 500 cal:

spoon_g2_lm5 <- lm(Calories..cal. ~ (Fat..g.+Saturated.Fat..g.+Protein..g.+Carbohydrates..g.+Fiber..g.+Sodium..mg.+Cholesterol..mg.), data = spoon_g2)

summary(spoon_g2_lm5)

ggplot(data = spoon_g2, mapping = aes(x=(Fat..g.+Protein..g.+Carbohydrates..g.+Fiber..g.), y=Calories..cal., col=Calories..cal.)) +

geom_jitter(size=4,shape=16,alpha=0.6) +

geom_quantile(color = "red",size=1,linetype=4,alpha=0.8) +

```
  ggtitle("Relationship Between Calories and Significant Nutrients (Calories: 250 - 500)") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Total Fat, Protein, Carbohydrates & Fiber (g)") +
  ylab("Calories (cal)”)
##
## Call:
## lm(formula = Calories..cal. ~ (Fat..g. + Saturated.Fat..g. +
##     Protein..g. + Carbohydrates..g. + Fiber..g. + Sodium..mg. +
##     Cholesterol..mg.), data = spoon_g2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.713  -4.837  -0.394   3.388  52.526
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.785e+00  1.835e+00   2.063   0.0395 *
## Fat..g.           8.629e+00  6.057e-02 142.472   <2e-16 ***
## Saturated.Fat..g.  1.872e-01  1.048e-01   1.787   0.0744 .
## Protein..g.       4.247e+00  4.061e-02 104.590   <2e-16 ***
```

```
## Carbohydrates..g.  3.938e+00  2.911e-02 135.301   <2e-16 ***
## Fiber..g.        -1.409e+00  1.170e-01 -12.048   <2e-16 ***
## Sodium..mg.       6.479e-06  2.357e-05   0.275   0.7835
## Cholesterol..mg.   1.003e-02  4.662e-03   2.151   0.0319 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.308 on 698 degrees of freedom
## Multiple R-squared:  0.9837, Adjusted R-squared:  0.9835
## F-statistic:  6015 on 7 and 698 DF,  p-value: < 2.2e-16
```

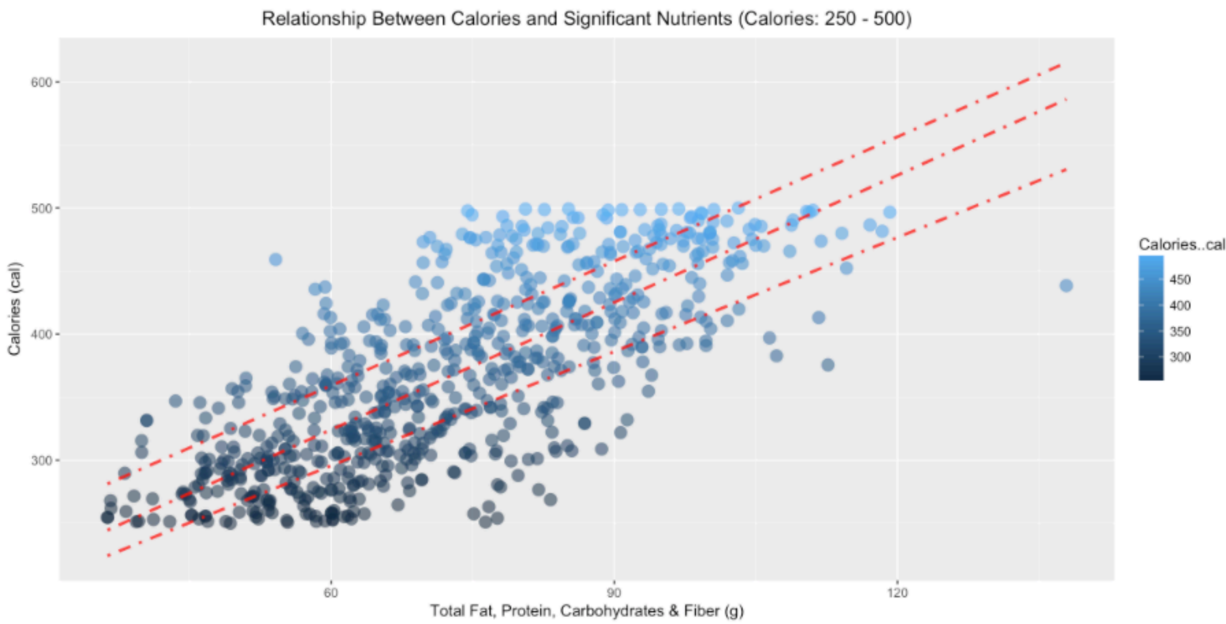
From the above output, p-value < 0.05 indicates that we have to reject the null hypothesis Ho.

There is a strong positive (+) linear relationship between the calories and above significant nutrients (fat, protein, fiber and carbohydrates).

The linear relationship for the above analysis is given by the following equation:

$Y = 3.785e{+}00 + 8.629e{+}00 * V + 4.247e{+}00 * W + 3.938e{+}00 * X - 1.409e{+}00 * Z + \epsilon$

Y: Calories (cal), V: Fat (g), W: Protein (g), X: Carbohydrates (g), Z: Fiber (g)



Finding: An average increase of 10g in fat, protein, carbohydrates and fiber will likely result in an average increase in calories by 154.05 cal for the given recipes. This relationship can be used to optimize the total calories in the recipes by altering the proportions of their significant nutrients.

Note: The dotted lines indicate the first, second and third quantiles of the data points in the graph.

C. Calorie Subgroup: 500 - 750 cal:

```
spoon_g3_lm5 <- lm(Calories..cal. ~ (Fat..g.+Saturated.Fat..g.+Protein..g.+Carbohydrates..g.
+Fiber..g.+Sodium..mg.+Cholesterol..mg.), data = spoon_g3)

summary(spoon_g3_lm5)

ggplot(data = spoon_g3, mapping = aes(x=(Fat..g.+Protein..g.+Carbohydrates..g.+Fiber..g.),
y=Calories..cal., col=Calories..cal.)) +

  geom_jitter(size=4,shape=17,alpha=0.6) +

  geom_quantile(color = "red",size=1,linetype=4,alpha=0.8) +
```

```
  ggtitle("Relationship Between Calories and Significant Nutrients (Calories: 500 - 750)") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Total Fat, Protein, Carbohydrates & Fiber (g)") +
  ylab("Calories (cal)”)
##
## Call:
## lm(formula = Calories..cal. ~ (Fat..g. + Saturated.Fat..g. +
##     Protein..g. + Carbohydrates..g. + Fiber..g. + Sodium..mg. +
##     Cholesterol..mg.), data = spoon_g3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.182  -8.486  -2.634   4.396 120.050
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       23.5903781  6.6296743   3.558 0.000416 ***
## Fat..g.            8.5151477  0.1145101  74.362  < 2e-16 ***
## Saturated.Fat..g. -0.2701340  0.1492601  -1.810 0.071047 .
## Protein..g.        4.1499186  0.0728799  56.942  < 2e-16 ***
```

```
## Carbohydrates..g.  3.8955723  0.0536885  72.559  < 2e-16 ***
## Fiber..g.        -1.4923789  0.1974398  -7.559 2.63e-13 ***
## Sodium..mg.      -0.0002173  0.0001812  -1.199 0.231086
## Cholesterol..mg.   0.0216187  0.0072207   2.994 0.002918 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.09 on 415 degrees of freedom
## Multiple R-squared:  0.9506, Adjusted R-squared:  0.9497
## F-statistic:  1140 on 7 and 415 DF,  p-value: < 2.2e-16
```

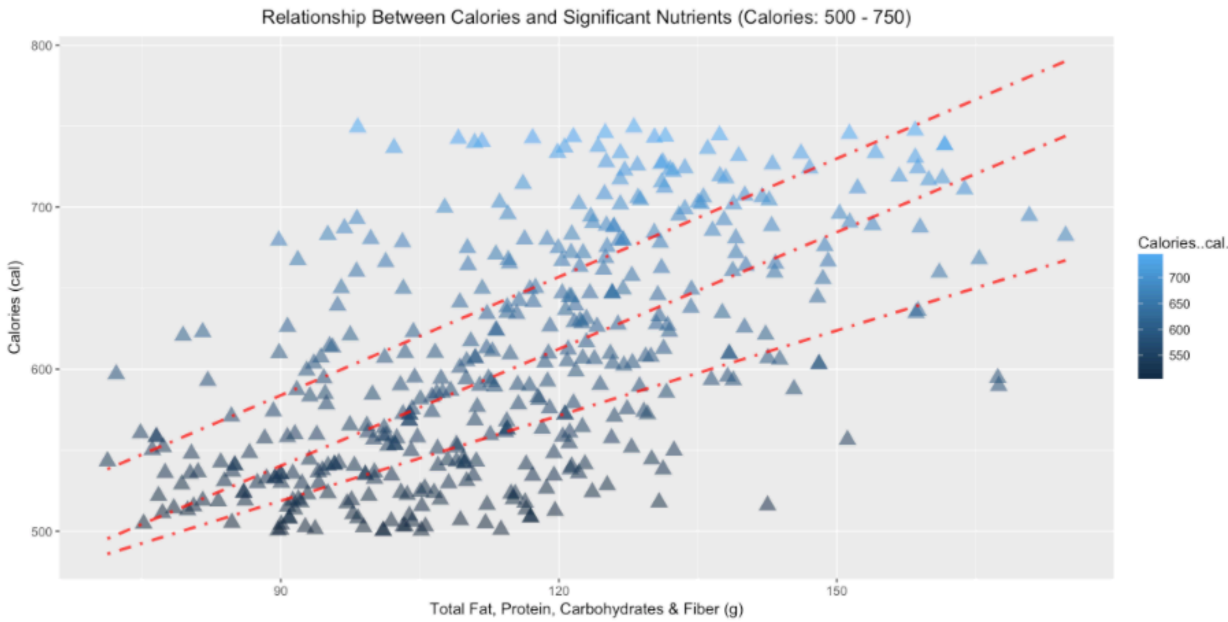
From the above output, p-value < 0.05 indicates that we have to reject the null hypothesis Ho.

There is a strong positive (+) linear relationship between the calories and above significant nutrients (fat, protein, fiber and carbohydrates).

The linear relationship for the above analysis is given by the following equation:

$$Y = 23.5903781 + 8.5151477 * V + 4.1499186 * W + 3.8955723 * X - 1.4923789 * Z + \epsilon$$

Y: Calories (cal), V: Fat (g), W: Protein (g), X: Carbohydrates (g), Z: Fiber (g)



Finding: An average increase of 10g in fat, protein, carbohydrates and fiber will likely result in an average increase in calories by 150.68 cal for the given recipes. This relationship can be used to optimize the total calories in the recipes by altering the proportions of their significant nutrients.

Note: The dotted lines indicate the first, second and third quantiles of the data points in the graph.

D. Calorie Subgroup: 750 - 2000 cal:

spoon_g4_lm5 <- lm(Calories..cal. ~ (Fat..g.+Saturated.Fat..g.+Protein..g.+Carbohydrates..g.+Fiber..g.+Sodium..mg.+Cholesterol..mg.), data = spoon_g4)

summary(spoon_g4_lm5)

ggplot(data = spoon_g4, mapping = aes(x=(Fat..g.+Protein..g.+Carbohydrates..g.+Fiber..g.), y=Calories..cal., col=Calories..cal.)) +

geom_jitter(size=4,shape=18,alpha=0.6) +

geom_quantile(color = "red",size=1,linetype=4,alpha=0.8) +

```
  ggtitle("Relationship Between Calories and Significant Nutrients (Calories: 750 - 2000)") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Total Fat, Protein, Carbohydrates & Fiber (g)") +
  ylab("Calories (cal)")
##
## Call:
## lm(formula = Calories..cal. ~ (Fat..g. + Saturated.Fat..g. +
##     Protein..g. + Carbohydrates..g. + Fiber..g. + Sodium..mg. +
##     Cholesterol..mg.), data = spoon_g4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.927 -11.899  -3.737   5.394 142.559
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       6.4889974  6.6633707   0.974    0.331
## Fat..g.           8.8638347  0.0990851  89.457  < 2e-16 ***
## Saturated.Fat..g.  0.0647521  0.1923898   0.337    0.737
## Protein..g.       4.1642540  0.0718631  57.947  < 2e-16 ***
```

```
## Carbohydrates..g.  4.0997946  0.0426726  96.076  < 2e-16 ***
## Fiber..g.         -2.2085363  0.2572068  -8.587 1.53e-15 ***
## Sodium..mg.       -0.0009410  0.0005747  -1.638    0.103
## Cholesterol..mg.   0.0191717  0.0125819   1.524    0.129
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.24 on 224 degrees of freedom
## Multiple R-squared:  0.9914, Adjusted R-squared:  0.9911
## F-statistic:  3682 on 7 and 224 DF,  p-value: < 2.2e-16
```

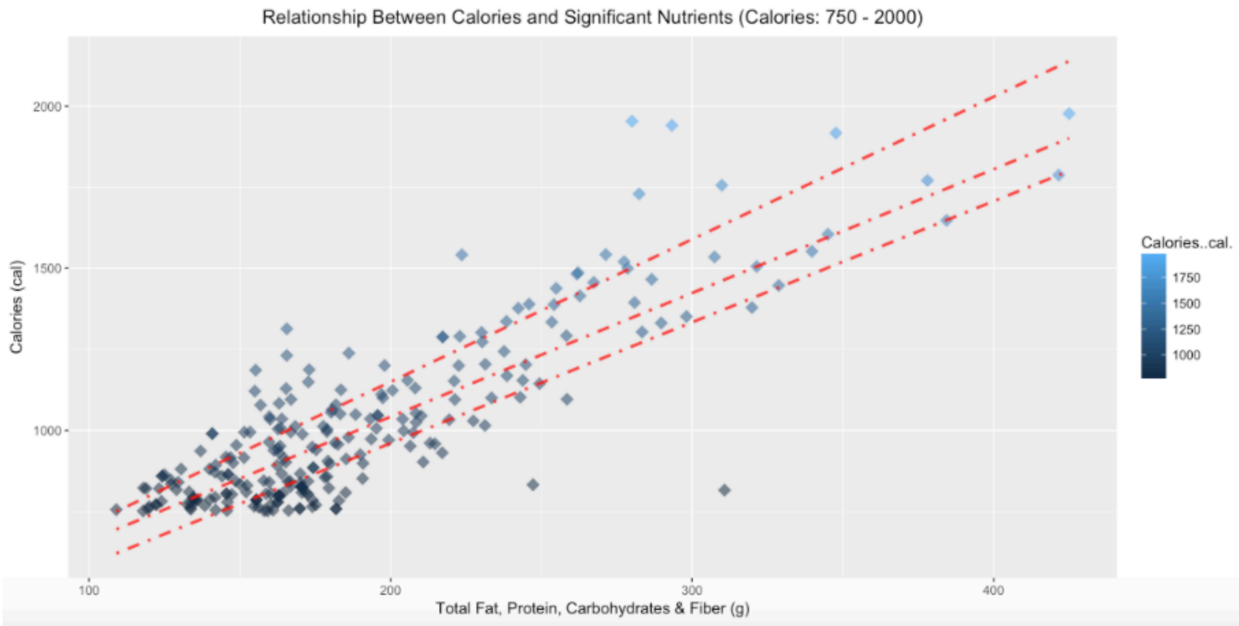
From the above output, p-value < 0.05 indicates that we have to reject the null hypothesis Ho.

There is a strong positive (+) linear relationship between the calories and above significant nutrients (fat, protein, fiber and carbohydrates).

The linear relationship for the above analysis is given by the following equation:

$$Y = 6.4889974 + 8.8638347 * V + 4.1642540 * W + 4.0997946 * X - 2.2085363 * Z + \epsilon$$

Y: Calories (cal), V: Fat (g), W: Protein (g), X: Carbohydrates (g), Z: Fiber (g)



Finding: An average increase of 10g in fat, protein, carbohydrates and fiber will likely result in an average increase in calories by 149.19 cal for the given recipes. This relationship can be used to optimize the total calories in the recipes by altering the proportions of their significant nutrients.

Note: The dotted lines indicate the first, second and third quantiles of the data points in the graph.