DAB501 Final Project

## Student Information

– Group Number: Group 3 – Student Name and Student ID of Group members:

* Vishant Bhatia (0798567)
* Tulaib Bin Ayyub (0789141)
* Alisha Mahajan (0802631)

##Academic Integrity

*Replace the underscores below with your names acknowledging that you have read and understood the statement in the context of St. Clair College’s Academic Integrity policies.*

We, Vishant Bhatia, Tulaib Bin Ayyub and Alisha Mahajan, hereby state that we have not communicated with or gained information in any way from any other group or resource that would violate the College’s academic integrity policies, and that all work is my own.

## Instructions

Follow the instructions on the handout provided.

## Packages and Data

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.3.6 ✔ purrr 0.3.5   
## ✔ tibble 3.1.8 ✔ dplyr 1.0.10  
## ✔ tidyr 1.2.1 ✔ stringr 1.4.1   
## ✔ readr 2.1.3 ✔ forcats 0.5.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(here)

## here() starts at C:/F Drive/BAsic stats DAB 501/project 3

setwd("C:\\F Drive\\BAsic stats DAB 501\\project 3")

# Replace the \_\_\_\_\_ with the correct file name  
calorie\_df <- read\_csv(here('Group3.csv'))

## Rows: 77 Columns: 3  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## dbl (3): calories, fat, carb  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

### Function which we use in our Project

best\_line\_general <- function(slope, intercept, predictor,predicted, data, residuals = TRUE) {  
 mlb <- data %>% mutate(y\_predicted = intercept + slope \* predictor,   
 square\_residuals = (predicted - y\_predicted)^2)  
   
 ssr <- mlb %>% summarize(ssr = round(sum(square\_residuals), 0))  
   
 if (residuals) {  
 p <- ggplot(mlb, aes(x = predictor, y = predicted)) +  
 geom\_point(shape = 21, fill = 'skyblue', size = 2) +   
 geom\_line(aes(x = predictor, y = y\_predicted), colour = 'blue') +   
 geom\_segment(aes(x = predictor, y = predicted,   
 xend = predictor, yend = y\_predicted), linetype = 2) +  
 ggtitle(paste0("Sum of Squared Resdiduals: ", ssr[[1]]))  
 } else {   
 p <- ggplot(mlb, aes(x = predictor, y = predicted)) +  
 geom\_point(shape = 21, fill = 'skyblue', size = 2) +   
 geom\_line(aes(x = predictor, y = y\_predicted), colour = 'blue')  
 }  
 p  
}

# MODELING: First pair of variables

### Question 1 Identify the explanatory variable.

Ans1. Fat.

### Question 2. Identify the response variable.

Ans2. calories.

### Question 3.Create a linear regression model and display the full output of the model.

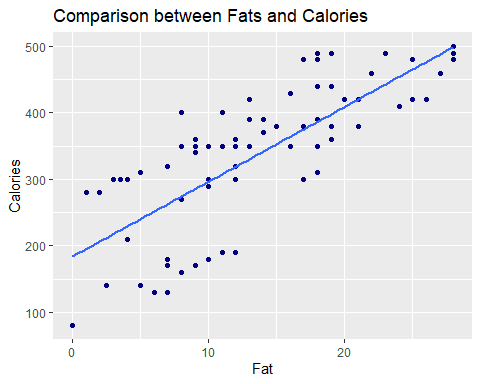
With First Explanatory Variable

cor(calorie\_df$fat, calorie\_df$calories)

## [1] 0.7586825

It has Strong positive relation

calorie\_df %>%   
 ggplot(aes(x = fat, y = calories))+  
 geom\_point(color = "navy")+  
 geom\_smooth(formula = 'y~x', method = 'lm', se = FALSE)+  
 labs(title = "Comparison between Fats and Calories",   
 x = "Fat",  
 y = "Calories")



### Question 4. Using the variables noted in #1 and #2 above and the results of #3, write the equation for your model.

For First Explanatory Variable

lin\_fat <- lm(calories ~ fat, data = calorie\_df)  
summary(lin\_fat)

##   
## Call:  
## lm(formula = calories ~ fat, data = calorie\_df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -132.599 -44.130 3.469 54.868 126.134   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 183.734 17.277 10.63 < 2e-16 \*\*\*  
## fat 11.267 1.117 10.09 1.32e-15 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 69.1 on 75 degrees of freedom  
## Multiple R-squared: 0.5756, Adjusted R-squared: 0.5699   
## F-statistic: 101.7 on 1 and 75 DF, p-value: 1.32e-15

In above model we have Slope = 11.267 and intercept = 183.734

Equation for this model is :-

y = mx+c Where y = response variable here it is Calories m = is slope of the line c = is intercept

intercept\_fat <- 183.734  
slope\_fat <- 11.267  
  
calorie\_df$calories\_pred\_fat =   
slope\_fat \* calorie\_df$fat + intercept\_fat

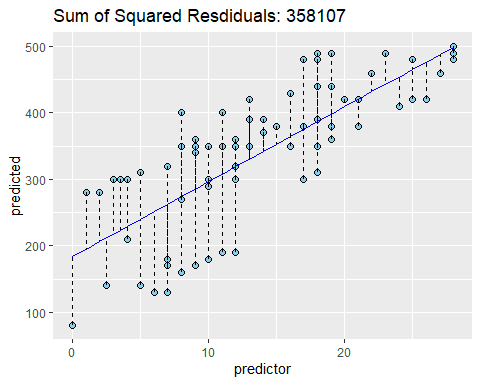
### Question 5. Explain what the intercept means in the context of the data

Ans5. The y-intercept is the value when x = 0. So, The Intercept from above model with respect to fat is 183.734 it means value of Calories is 183.734 when value of fat is zero.

### Question 6. Is the intercept a useful/meaningful value in the context of our data? If yes, explain. If not, explain what purpose it serves.

Ans6. Intercept is a useful value. It is a value where function crosses y - axis in y = mx + c. In our context to data intercept is useful as it tell us the value of calories present in particular item when value of fat iz zero. In our case that value is 183.734. In our data set there are some cases when value of fat is zero (at 67th index) at that time value of calories are 80. So in this case it should be near about 183.734, but it means it have error and we have to find that with slope and intercept. Now we will find this sum of squared residuals by best line fitted.

best\_line\_general(slope\_fat, intercept\_fat, predictor = calorie\_df$fat, predicted = calorie\_df$calories, data = calorie\_df, residuals = TRUE )

 In our case best fitted line is at sum squared residuals of 358107

### Question 7. Explain what the slope means in the context of the data.

Ans7. In equation y = mx + c, where m is a slope. Slope means for every unit increase of x value the value of y is increased by value slope unit. In our case value of slope is 11.267. It means every unit increase in fat our calories is increased by 11.267. So with increase of 1 gram in fat the value of calories is increased by 11.267.

## MODELING: Second pair of variables

### Question 1 Identify the explanatory variable.

Ans1. Carb.

### Question 2. Identify the response variable.

Ans2. calories.

### Question 3.Create a linear regression model and display the full output of the model.

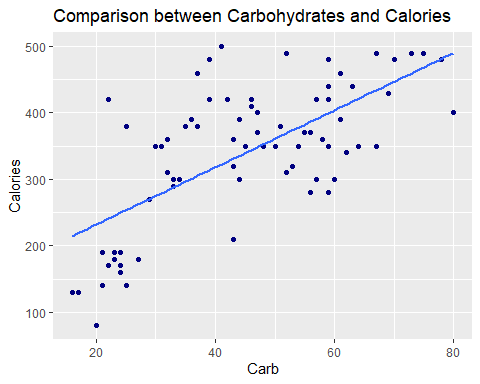
With Second Explanatory Variable

cor(calorie\_df$carb, calorie\_df$calories)

## [1] 0.674999

It is a Positive strong relation

calorie\_df %>%   
 ggplot(aes(x = carb, y = calories))+  
 geom\_point(color = "navy")+  
 geom\_smooth(formula = 'y~x', method = 'lm', se = FALSE)+  
 labs(title = "Comparison between Carbohydrates and Calories",   
 x = "Carb", y = "Calories")



### Question 4. Using the variables noted in #1 and #2 above and the results of #3, write the equation for your model.

For Second Explanatory Variable

lin\_carb <- lm(calories ~ carb, data = calorie\_df)  
summary(lin\_carb)

##   
## Call:  
## lm(formula = calories ~ carb, data = calorie\_df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -151.962 -70.556 -0.636 54.908 179.444   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 146.0204 25.9186 5.634 2.93e-07 \*\*\*  
## carb 4.2971 0.5424 7.923 1.67e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 78.26 on 75 degrees of freedom  
## Multiple R-squared: 0.4556, Adjusted R-squared: 0.4484   
## F-statistic: 62.77 on 1 and 75 DF, p-value: 1.673e-11

In above model we have Slope = 4.2971 and intercept = 146.0204

Equation for this model is :-

y = mx+c

Where y = response variable here it is Calories m = is slope of the line c = is intercept

intercept\_carb <- 146.0204  
slope\_carb <- 4.2971  
  
calorie\_df$calories\_pred\_carb = slope\_carb \* calorie\_df$carb + intercept\_carb

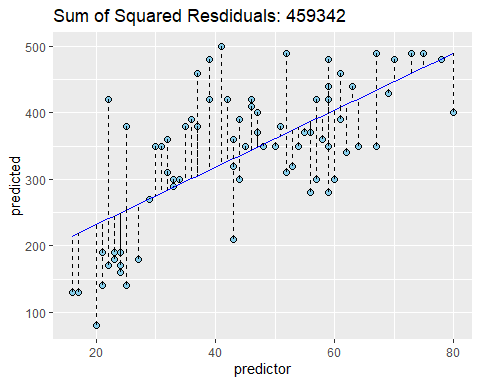
### Question 5. Explain what the intercept means in the context of the data

Ans5. The y-intercept is value when x = 0. So,the Intercept from above model with respect to carb is 146.0204 it means value of Calories is 146.0204 when value of carb is zero.

### Question 6. Is the intercept a useful/meaningful value in the context of our data? If yes, explain. If not, explain what purpose it serves.

Ans6. Value of calories is 146.0204 when value of carbs is zero. But, in our data set there is no record having carb = 0 (zero) or near about zero. So, in this case intercept is not very useful. We will calculate best fitted line.

best\_line\_general(slope\_carb, intercept\_carb, predictor = calorie\_df$carb, predicted = calorie\_df$calories, data = calorie\_df, residuals = TRUE )



In our case best fitted line is at sum squared residuals of 459342

### Question 7. Explain what the slope means in the context of the data.

Ans7. In equation y = mx + c, m is a slope. Slope means for every unit increase of x value the value of y is increased by value slope unit. In our case value of slope is 4.2971. It means every unit increase in carbohydrates our calories is increased by 4.2971. So with increase of 1 gram in carbs the value of calories is increased by 4.2971.

## MODEL ASSESSMENT

### Question 1. Which metric can you use to choose between the two models you just created?

Ans1. The coefficient of determination which is called R-squared is the metrics used to choose the model.

### Question 2. Explain what this metric means and why it is good for comparing models.

Ans2. Coefficient of determination (R-squared) is a metrics in a regression model that express or tell us proportion of variance in response variable that is explained by the explanatory variable. It means it shows the best fitted line in the regression model that tells how well the data is fitted. Value of R-squared lies between 0 and 1. It is always good for comparing models because it interpreted the percentage of variance in response variable. For example one of our model it have 0.57 of R-square, as it is proved below.

cor(calorie\_df$fat, calorie\_df$calories) ^ 2

## [1] 0.5755991

It might mean that the model explains about 57% of the variation in calories.

### Question 3. According to this metric, which model is the best of the two you created? Why?

cor(calorie\_df$fat, calorie\_df$calories) ^ 2

## [1] 0.5755991

cor(calorie\_df$carb, calorie\_df$calories) ^ 2

## [1] 0.4556237

Model R-squared

Model

R-Squared

fat and calories

0.5755

carbs and calories

0.4556

The R-squared is higher in case of explanatory variable fat as it has value of 0.5755. It means fat can better express the variance as compare to carb in calories.

## MODEL DIAGNOSTICS

### Question 1. Create two new data columns based on your best model: predicted values for your response variable and the corresponding residuals.

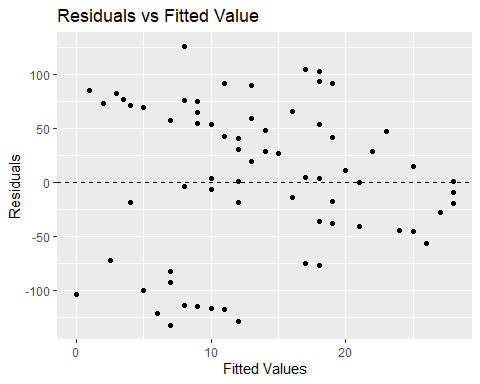
lin\_fat <- lm(calories ~ fat, data = calorie\_df)  
summary(lin\_fat)

##   
## Call:  
## lm(formula = calories ~ fat, data = calorie\_df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -132.599 -44.130 3.469 54.868 126.134   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 183.734 17.277 10.63 < 2e-16 \*\*\*  
## fat 11.267 1.117 10.09 1.32e-15 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 69.1 on 75 degrees of freedom  
## Multiple R-squared: 0.5756, Adjusted R-squared: 0.5699   
## F-statistic: 101.7 on 1 and 75 DF, p-value: 1.32e-15

intercept\_fat <- 183.734  
slope\_fat <- 11.267  
  
calorie\_df <- calorie\_df %>%   
 mutate(calories\_pred\_fat = intercept\_fat +   
 slope\_fat \* fat,   
 residuals\_pred\_fat = calories - calories\_pred\_fat)

### Question 2. Create a plot to check the assumption of linearity. State whether or not this condition is met and explain your reasoning.

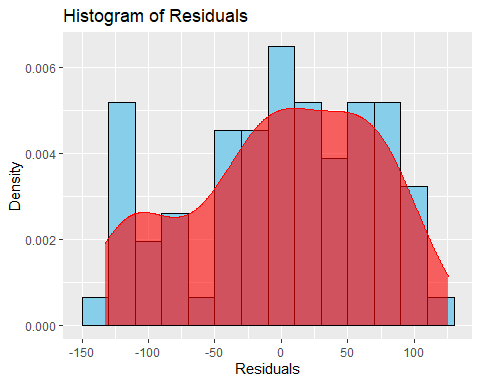
calorie\_df %>%   
 ggplot(aes(x = fat , y = residuals\_pred\_fat))+  
 geom\_point()+  
 geom\_hline(yintercept = 0, linetype = "dashed")+   
 labs(title = "Residuals vs Fitted Value",   
 y="Residuals",   
 x="Fitted Values")



Ans2. In Above plot there are some points which are above the line it means error in that case is almost zero. But in some cases points are far from that dashed line it means either error is positive or negative. In this case there is no pattern of linearity between residual and fitted value (fat). It means assumptions of linearity satisfied.

### Question 3. Create a plot to check the assumption of nearly normal residuals. State whether or not this condition is met and explain your reasoning.

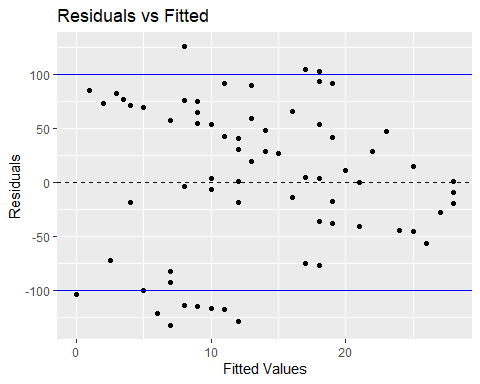
calorie\_df %>%   
ggplot(aes(x = residuals\_pred\_fat)) +  
 geom\_histogram(aes(y = ..density..), binwidth = 20,   
 color = "black", fill = "skyblue" )+  
 geom\_density(color = "red",fill = 'red', alpha = 0.6)+  
 labs(title = "Histogram of Residuals",   
 x = "Residuals",   
 y="Density")



Ans3. From above plot we can say that our plot is not “bell shaped” plot and it is not uniform plot. most of the data lies towards the right side. So, this is our left skewed plot. In this model, assumption of nearly normal residuals is not satisfied.

### Question 4. Create a plot to check the assumption of constant variability. State whether or not this condition is met and explain your reasoning

calorie\_df %>%   
 ggplot(aes(x = fat, y = residuals\_pred\_fat ))+  
 geom\_point()+  
 geom\_hline(yintercept = 0, linetype = "dashed")+  
 geom\_hline(yintercept = 100, color = "blue")+  
 geom\_hline(yintercept = -100, color = "blue")+  
 labs(title = "Residuals vs Fitted",   
 y="Residuals",   
 x="Fitted Values")



Ans4. In this plot we can see the trend between fat and residuals. The value of fat is increasing with decrease in value of residuals and vice-versa. The variability of points around the least squares line are constant here. This implies that the variability of residuals around the 0 line is constant as well. So, assumption of constant variability is satisfied.

## CONCLUSION

### Question 1. Based on the results of the “Model Diagnostics” section above, what can you conclude about your model?

Ans1. According to above “Model Diagnostics”, we can conclude that our model is not best fitted model but it is moderately fitted model. It is satisfying all the diagnostics above done except one assumptions of nearly normal residuals. In general model, most of the error will lie near zero, as we move far from line the number of error must be less. Which mean most of the error must be near zero and its mean must be near about zero. But, in our case it is not. In our model assumption of constant variability is satisfied which means variation in our model is constant and have no particular pattern. In our model assumption of linearity is also satisfying and without this accurate predictions cannot be done.