# Predicting calories from the quantity of nutrients in the food

## Using Linear Regression

Mohsin Asif (M12467820) Smarak Das(M12435888) Faisal Ahmad(M12434762)

# Contents

## Introduction

In this paper we are analyzing nutritional data to predict amount of calories in food. In our data set we have nutritional variables as covariates such as fat, sugar, carbohydrates, protein etc. we use these variables to predict response variable calories.

In the first part, we do data cleaning by changing variable names and removing null values from our dataset. We then do linear regression to build a model and do adequacy checking on that. Once we established that our regression is valid, we looked for multicollinearity in our dataset. We checked multicollinearity in our model and it was very high and caused large variance inflation factors. We did anova test to check which variables are significant to get a clue on what will reduce multicollinearity in our model. We dropped some variables to reduce vif to an acceptable level. Based on that we selected our final model.

## Data Exploration and Data Cleaning

Firstly we will read the data file and load the CAR library

```
install.packages("car")
library(car)
calories<-read.csv("C:/Users/warfaisal/Documents/calories.csv", header = TRUE)
```

Let’s have a look at various variables in our data set

```
#get names of variables
names(calories)
```

```
> calories<-read.csv("C:/Users/warfaisal/Documents/calories.csv")
> names(calories)
 [1] "Fast.Food.Restaurant" "Type"                 "Serving.Size..g."
 [4] "Calories"             "Total.Fat..g."        "Saturated.Fat..g."
 [7] "Trans.Fat..g."        "Sodium..mg."          "Carbs..g."
[10] "Sugars..g."           "Protein..g."
> #rename variables as
> names(calories)=c( "FastFoodRest",    "Type",  "ServingSize",  "Calories",  "TotalFat",
+   "SaturatedFat",  "TransFat",  "Sodium_mg",  "Carbs",  "Sugars",  "Protein")
```

We can see the different variables are

1. FastFoodRest: which has the name of restaurant e.g. McDonalds, Wendy, Sonic etc.
2. Type: has the types of restaurants e.g. burger, MilkShake, Grilled Chicken etc.
3. ServingSize: contains the serving size in grams
4. Calories: has the number of calories per Serving Size
5. TotalFat: sum of saturated, monounsaturated and polyunsaturated fats in grams
6. SaturatedFat: saturated fat content in grams
7. TransFat: Trans fatty acids in grams which is unhealthy
8. Sodium_mg: Sodium content in milligrams
9. Protein: Protein content in grams

Let's Take a look at the structure of our data set, we have a total of 11 variables with 126 observations. It looks like FastFoodRest and Type are categorical variables.

```
> # check the structure of data
> str(calories)
'data.frame':   126 obs. of  11 variables:
 $ FastFoodRest: Factor w/ 12 levels "Burger King",..: 8 8 8 8 8 8 8 8 8 8 ...
 $ Type        : Factor w/ 6 levels "Breaded Chicken Sandwich",..: 2 2 2 2 2 2 6 1 5 3 ...
 $ ServingSize : int  98 113 211 202 270 283 257 213 200 65 ...
 $ Calories    : int  240 290 530 520 720 750 530 510 350 190 ...
 $ TotalFat    : num  8 11 27 26 40 43 15 22 9 12 ...
 $ SaturatedFat: num  3 5 10 12 15 19 10 3.5 2 2 ...
 $ TransFat    : num  0 0.5 1 1.5 1.5 2.5 1 0 0 0 ...
 $ Sodium_mg   : int  480 680 960 1100 1470 1280 160 990 820 360 ...
 $ Carbs       : num  32 33 47 41 51 42 86 55 42 12 ...
 $ Sugars      : num  6 7 9 10 14 10 63 10 8 0 ...
 $ Protein     : num  12 15 24 30 39 48 11 24 28 9 ...
```

Lets do some cleaning in data by checking and removing the Missing values.

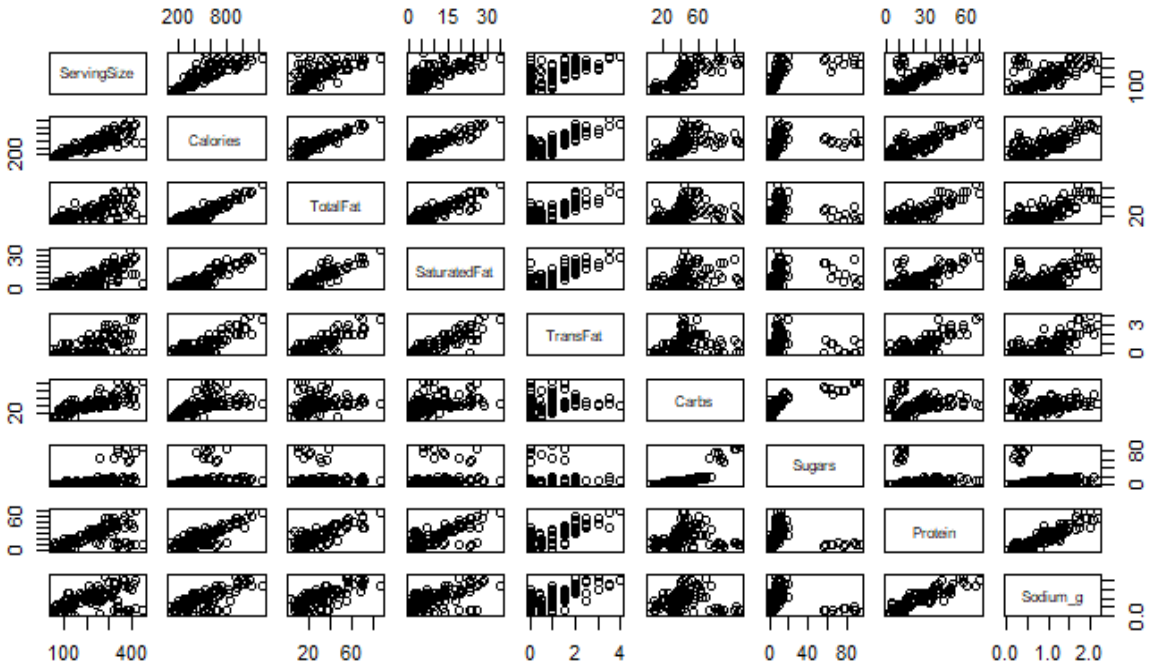
TransFat has missing values, after removing the NA values , we are left with 114 observations which is not a bad number

```
> #check for Missing values column wise
> apply(calories, 2, function(x) any(is.na(x)))
FastFoodRest         Type  ServingSize     Calories     TotalFat SaturatedFat
       FALSE        FALSE        FALSE        FALSE        FALSE        FALSE
    TransFat    Sodium_mg        Carbs       Sugars      Protein
        TRUE        FALSE        FALSE        FALSE        FALSE
> #remove missing values
> calories1<-na.omit(calories)
> apply(calories1, 2, function(x) any(is.na(x)))
FastFoodRest         Type  ServingSize     Calories     TotalFat SaturatedFat
       FALSE        FALSE        FALSE        FALSE        FALSE        FALSE
    TransFat    Sodium_mg        Carbs       Sugars      Protein
       FALSE        FALSE        FALSE        FALSE        FALSE
>
> #view number of observations
> nrow(calories1)
[1] 114
```

Lets convert Sodium from milligrams to grams as well as remove the categorical variables viz, FastFoodRest and type. Afterwards we will have a look at data and the summary statistics of our data.

```
> #converting Sodium to grams and removing unwanted variables sodim_mg, type, fastfoodresta
urant
> calories1$Sodium_g=calories1$Sodium_mg/1000
> calories2<-subset(calories1,select=-c(Sodium_mg,FastFoodRest,Type))
> #peek of data
> head(calories2, 5)
  ServingSize Calories TotalFat SaturatedFat TransFat Carbs Sugars Protein Sodium_g
1          98      240        8            3      0.0    32      6      12     0.48
2         113      290       11            5      0.5    33      7      15     0.68
3         211      530       27           10      1.0    47      9      24     0.96
4         202      520       26           12      1.5    41     10      30     1.10
5         270      720       40           15      1.5    51     14      39     1.47
>
> #descriptive stats of variables
> summary(calories2)
  ServingSize       Calories         TotalFat      SaturatedFat      TransFat
 Min.   : 44.0   Min.   : 130.0   Min.   : 3.50   Min.   : 1.00   Min.   :0.0000
 1st Qu.:115.8   1st Qu.: 322.5   1st Qu.:14.00   1st Qu.: 3.50   1st Qu.:0.0000
 Median :212.0   Median : 491.5   Median :22.00   Median : 7.00   Median :0.5000
 Mean   :217.2   Mean   : 515.7   Mean   :27.37   Mean   : 9.85   Mean   :0.8211
 3rd Qu.:292.5   3rd Qu.: 657.5   3rd Qu.:37.00   3rd Qu.:14.07   3rd Qu.:1.5000
 Max.   :450.0   Max.   :1240.0   Max.   :87.00   Max.   :35.00   Max.   :4.0000
     Carbs            Sugars          Protein         Sodium_g
 Min.   :  6.00   Min.   : 0.00   Min.   : 2.00   Min.   :0.0500
 1st Qu.: 32.25   1st Qu.: 3.00   1st Qu.:12.25   1st Qu.:0.5487
 Median : 42.00   Median : 7.00   Median :22.00   Median :0.9050
 Mean   : 43.61   Mean   :13.11   Mean   :24.33   Mean   :0.9290
 3rd Qu.: 50.75   3rd Qu.:10.00   3rd Qu.:33.75   3rd Qu.:1.2400
 Max.   :106.00   Max.   :93.00   Max.   :69.00   Max.   :2.1900
> #correlation between variables
> pairs(calories2)
```

Afterwards we will check the correlation between various variables. As we can see that Calories have positive correlation with almost all the variables except sugars. On the other hand, some covariates are highly correlated to other covariates which suggests that there might be multicollinearity in our model.




## Model Building

Let's build our first linear model with Calories vs Serving size, total fat, saturated fat, transfat, sodium, carbs, sugars and proteins

```
> #multiple linear model 1 with
> model_cal_1<-lm(Calories~ServingSize+TotalFat+SaturatedFat+TransFat+Sodium_g+Carbs+Sugars
+Protein, data=calories1)
```

Let's look at the summary statistics of our model

```
> summary(model_cal_1)

Call:
lm(formula = Calories ~ ServingSize + TotalFat + SaturatedFat +
    TransFat + Sodium_g + Carbs + Sugars + Protein, data = calories1)

Residuals:
     Min       1Q   Median       3Q      Max
-125.975   -6.441    1.374    8.968   91.021

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.38129    7.56853   0.579 0.563908
ServingSize   0.06681    0.07459   0.896 0.372418
TotalFat      8.45673    0.44403  19.045  < 2e-16 ***
SaturatedFat -0.79955    1.09014  -0.733 0.464927
TransFat     20.17457    5.37673   3.752 0.000288 ***
Sodium_g     14.82005   11.43923   1.296 0.197973
Carbs         3.71000    0.27728  13.380  < 2e-16 ***
Sugars        0.10271    0.29215   0.352 0.725856
Protein       3.27742    0.47751   6.864 4.85e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24.3 on 105 degrees of freedom
Multiple R-squared:  0.9909,    Adjusted R-squared:  0.9902
F-statistic:  1423 on 8 and 105 DF,  p-value: < 2.2e-16
```

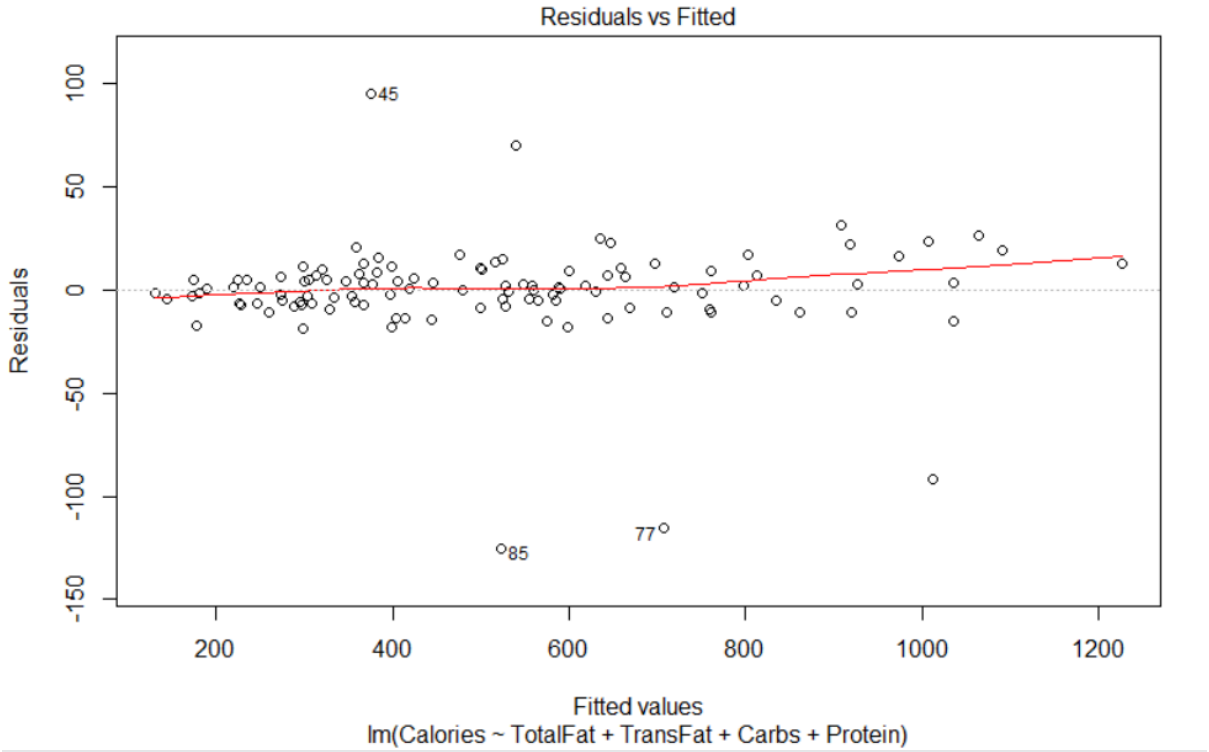
ServingSize, SaturatedFat, Sodium_g and Sugars are not good covariates as there p values are > 0.05 thus we cannot reject the null hypothesis. Which implies in this model there is not a linear relationship between Calories and ServingSize, SaturatedFat, Sodium_g and Sugars.

In order to overcome this, we build a new model model_cal_2 by dropping ServingSize, Sugars, SaturatedFat, Sodium_g. Then looking at summary statistics of our new model we can see all the covariates are having a significant P and t value to express Calories linearly.

```
> #Dropping ServingSize, Sugars, Sodium_g , SaturatedFat based on p value
> model_cal_2<-lm(Calories~TotalFat+TransFat+Carbs+Protein, data=calories1)
> summary(model_cal_2)

Call:
lm(formula = Calories ~ TotalFat + TransFat + Carbs + Protein,
    data = calories1)

Residuals:
     Min       1Q   Median       3Q      Max
-125.144   -6.245    0.987    7.886   95.208

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.9602     6.4227   1.084   0.2809
TotalFat      8.3500     0.2655  31.455   <2e-16 ***
TransFat     17.1957     4.2576   4.039   0.0001 ***
Carbs         3.9368     0.1143  34.433   <2e-16 ***
Protein       3.8796     0.2626  14.773   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24.19 on 109 degrees of freedom
Multiple R-squared:  0.9906,    Adjusted R-squared:  0.9903
F-statistic:  2872 on 4 and 109 DF,  p-value: < 2.2e-16
```

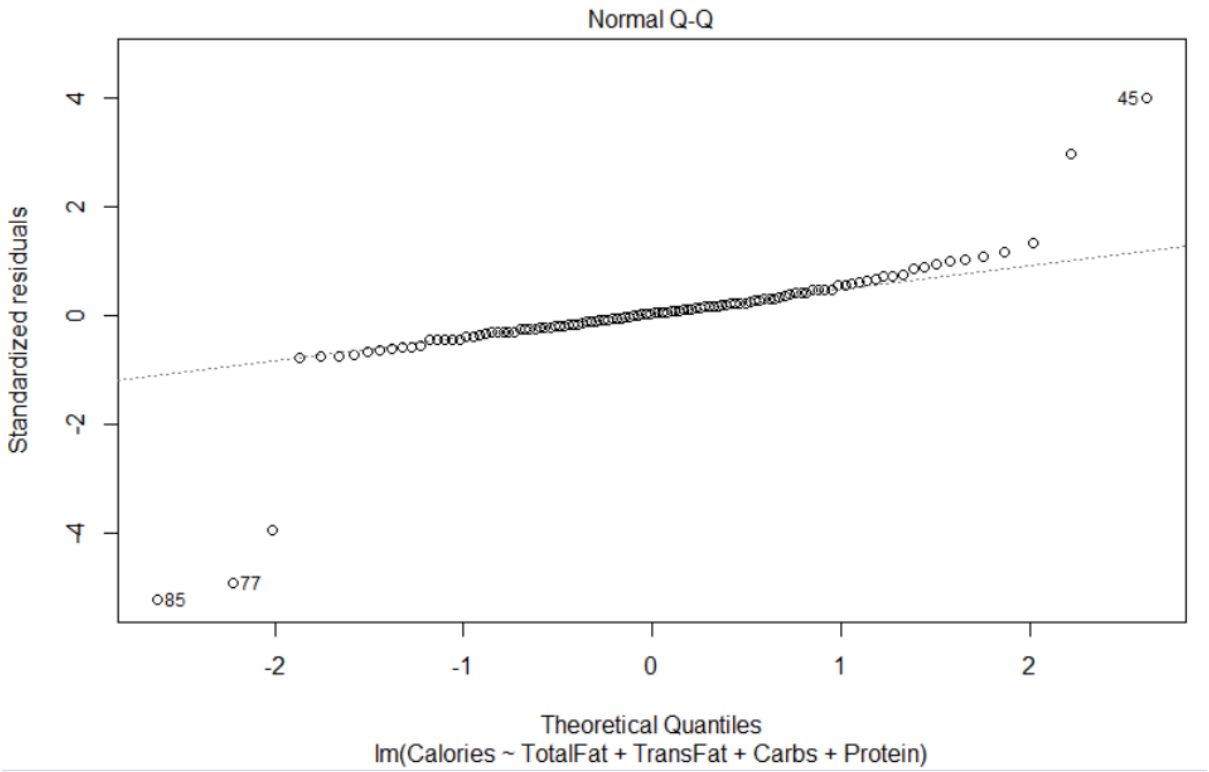
## Model Adequacy Checking

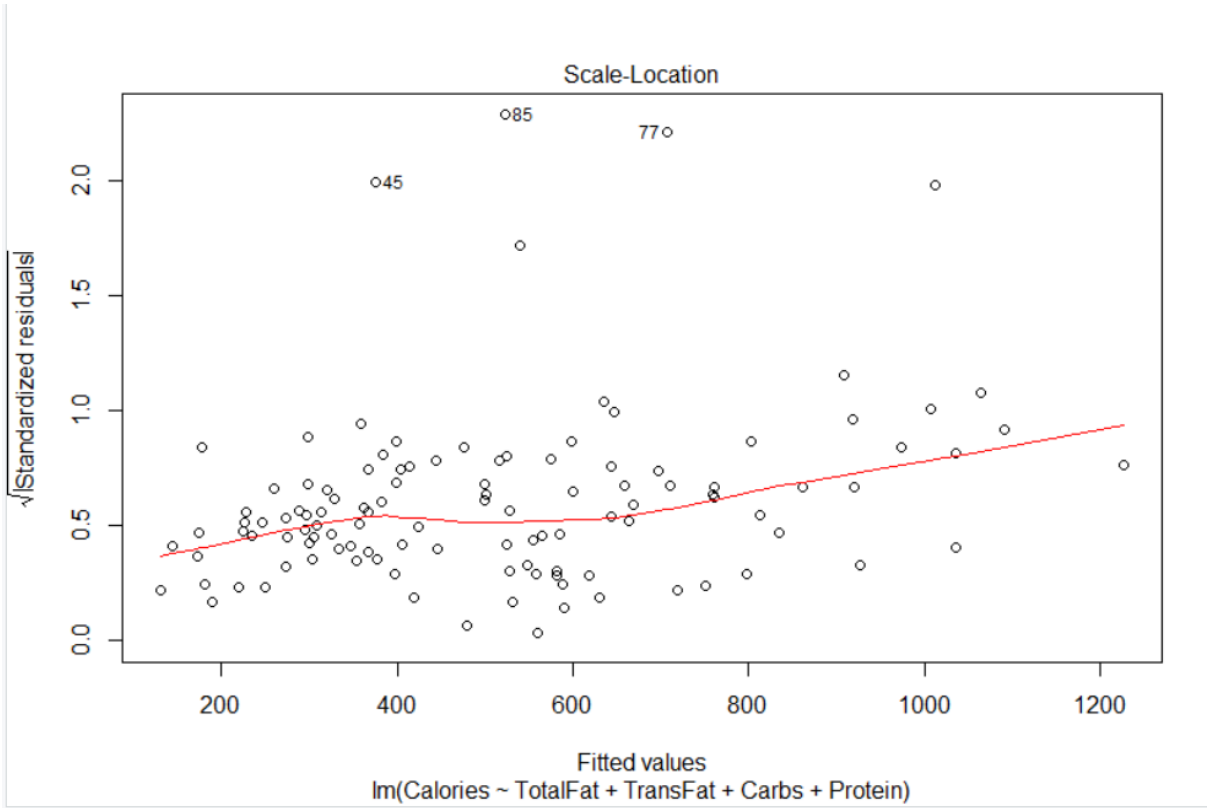
Now let's check the adequacy of our fitted model.

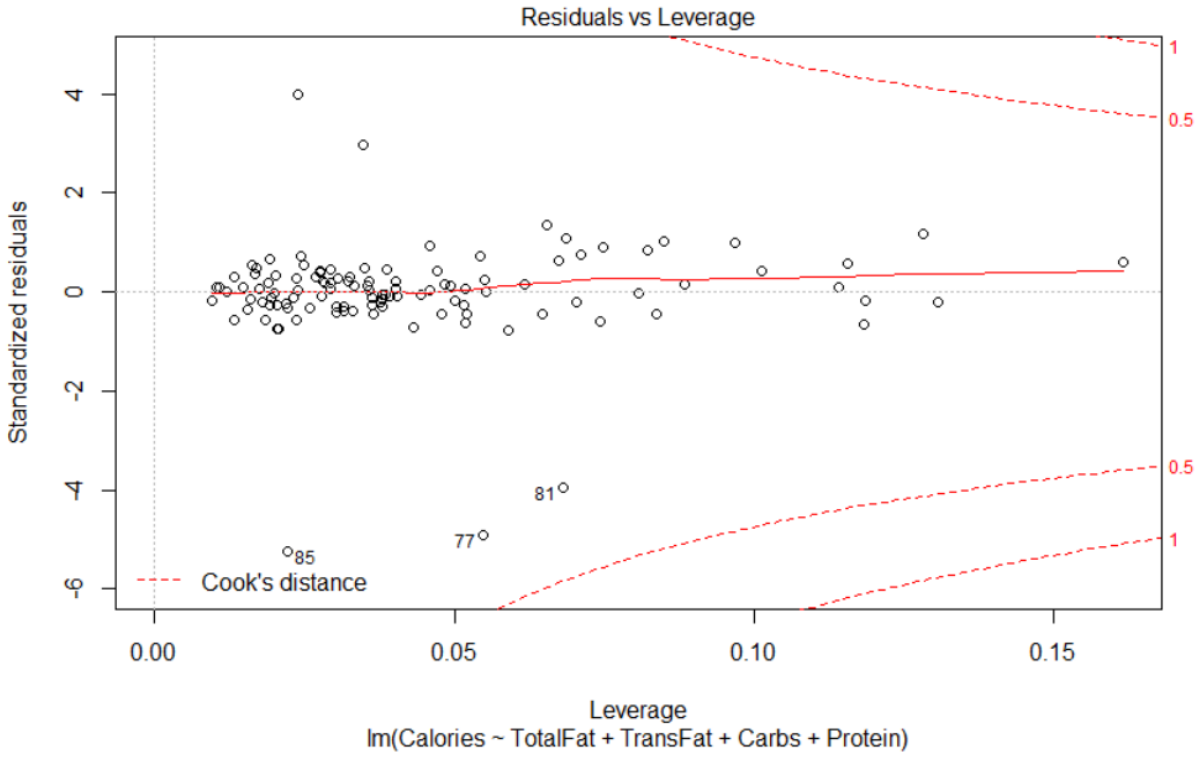



Above in the first plot we have Residual vs Fitted Values , we dont see any pattern on the red line Thus residuals are linearly distributed over fitted values and we can say approximately that variance is equal.

QQ plot is fairly linear except few outliers. Standardized residuals mostly follow the fitted model line. Thus, meeting our normality assumption


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Calories ~ TotalFat + TransFat + Carbs + Protein)
45
77
85

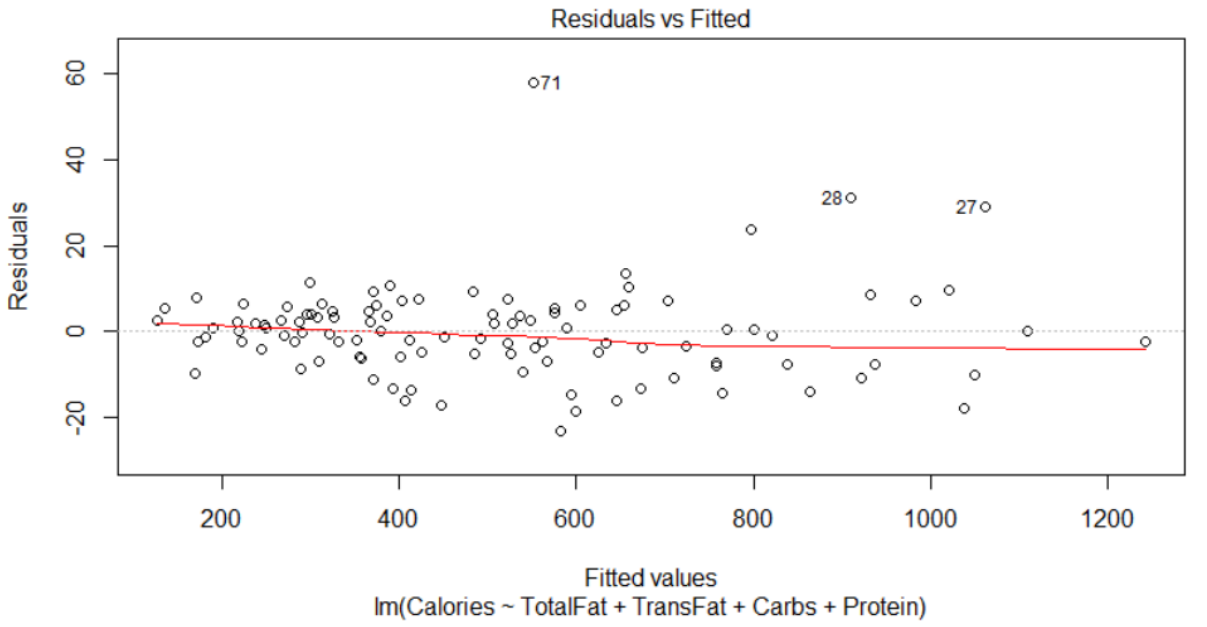

Scale-Location
√|Standardized residuals|
Fitted values
lm(Calories ~ TotalFat + TransFat + Carbs + Protein)
85
77
45

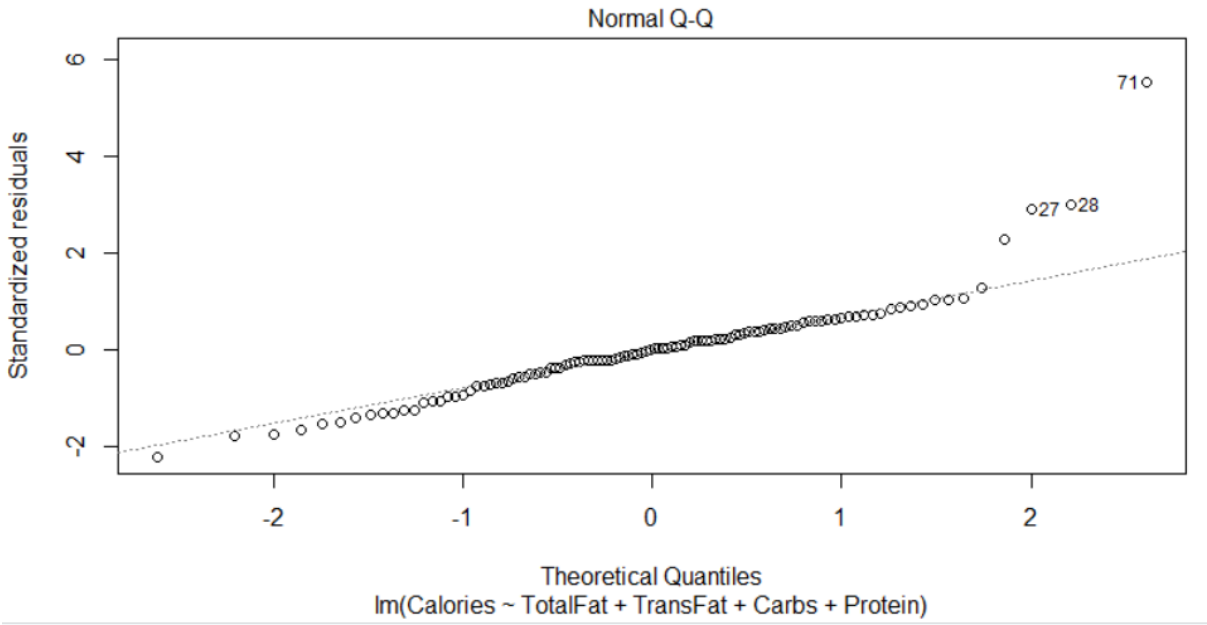


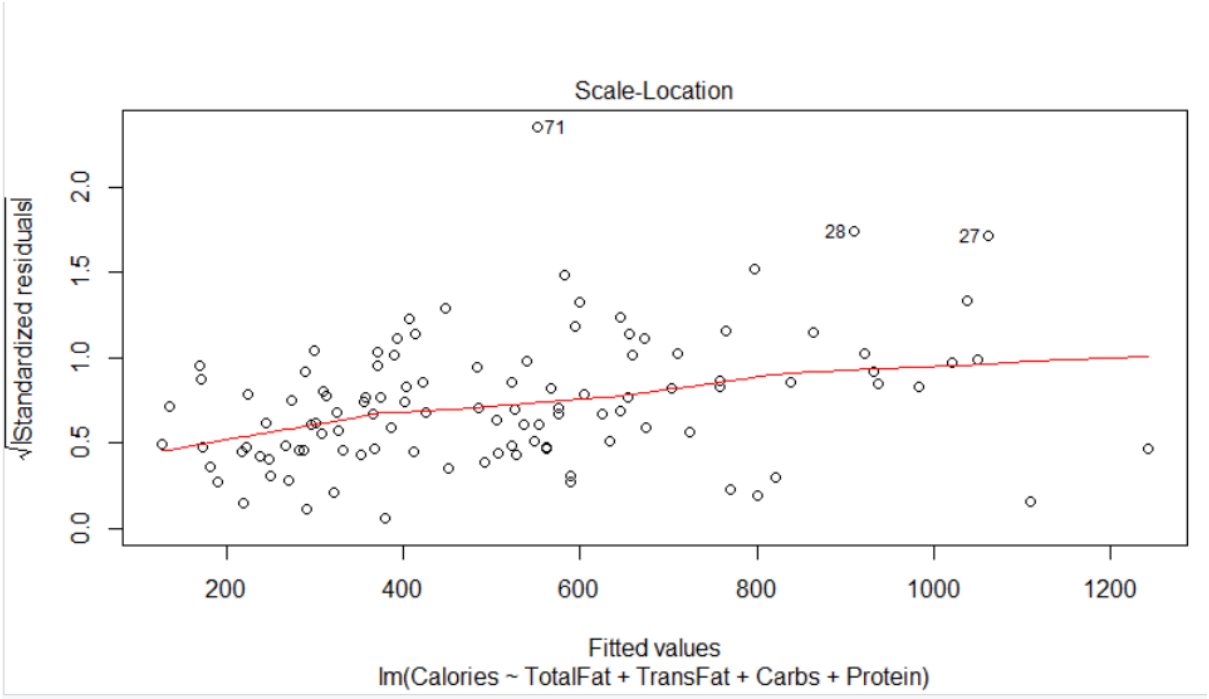

From the above graphs we can see there are few outliers in our model . To fix those let's remove them and build another model

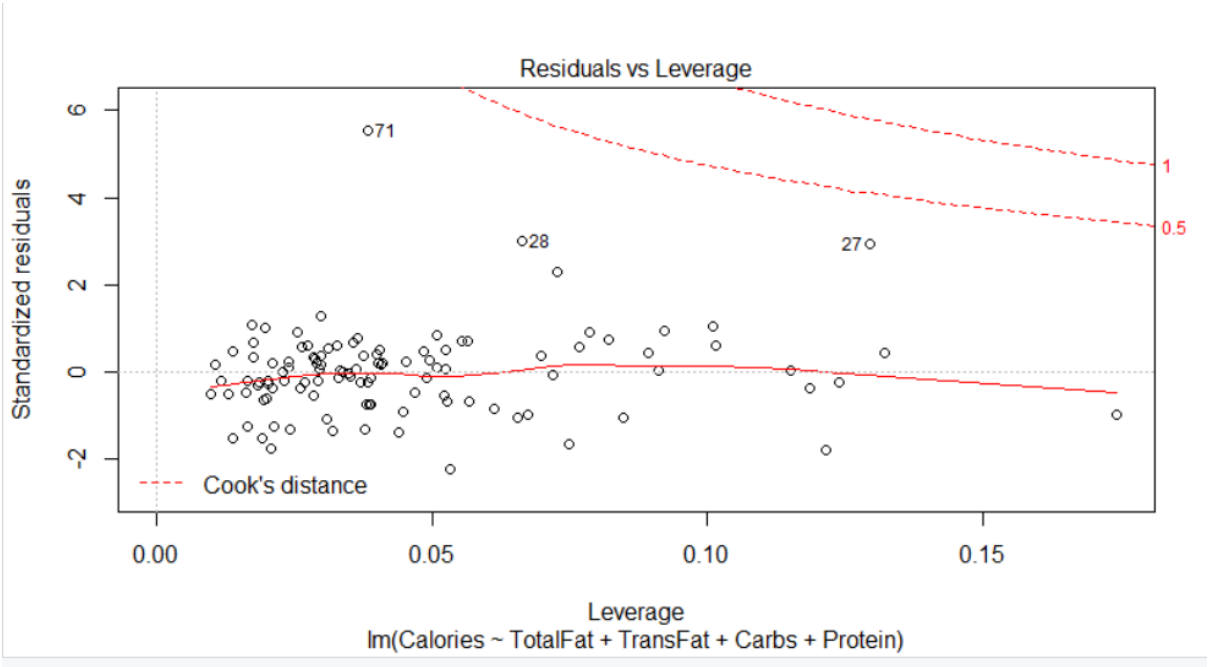
```
> #Removing outliers
> calories3=calories2[-c(77,45,81,85),]
> model_cal_3<-lm(Calories~TotalFat+TransFat+Carbs+Protein, data=calories3)
> summary(model_cal_3)

Call:
lm(formula = Calories ~ TotalFat + TransFat + Carbs + Protein,
    data = calories3)

Residuals:
    Min      1Q  Median      3Q     Max
-22.926  -5.596   0.141   4.736  57.794

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.50614    2.87663  -0.524  0.60168
TotalFat     8.98511    0.12191  73.701  < 2e-16 ***
TransFat     6.71535    1.94268   3.457  0.00079 ***
Carbs        3.94256    0.05046  78.136  < 2e-16 ***
Protein      3.95596    0.11609  34.077  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.66 on 105 degrees of freedom
Multiple R-squared:  0.9982,    Adjusted R-squared:  0.9981
F-statistic: 1.45e+04 on 4 and 105 DF,  p-value: < 2.2e-16

> plot(model_cal_3)
```

After removing building the new model we can see a fair increase in F statistics, Adjusted R-square value. Let's recheck the model adequacy with the help of Plot(model_cal_3) function.






The model looks fairly good meeting all the assumption of Linearity between response and regressor, Normality of error distribution, Independence of errors i.e. non-correlation, and equal variance of errors

Looking at the VIF values of our model we can say that we do not have multicollinearity problem. Looking at the correlations between the variables earlier, we could see that there

will be high multicollinearity but by dropping variables in model in the early phase, we got rid of multicollinearity.

All of the values are below 10 so we are good.

```
> vif(model_cal_3)
TotalFat TransFat    Carbs  Protein
4.353981 3.421443 1.069910 3.082291
```

## Final model:

Calories=6.96+8.35*TotalFat+17.19*TransFat+3.93*Carbs+3.87*Protein

## Conclusion and Interpretation

After building our final model, we can say that while determining calories in a product nutrients such as total fat, trans fat, carbohydrates, and protein are most significant variables that largely explain the variation in calories.

Keeping all variables fixed, a unit increase in total fat in a food, increases calories by 8.35 on average. Similarly, trans fat causes 17.19 unit increase on average for every one unit increase. Lastly, carbs and proteins, cause calories to increase by 3.93 and 3.87 on average for every one unit increase keeping all other variables fixed.