Predicting calories from the quantity of nutrients in the food

Using Linear Regression

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

Contents

1. [Introduction 2](#_Toc507082676)
2. [Data Exploration and Data Cleaning 2](#_Toc507082677)
3. [Model Building 5](#_Toc507082678)
4. [Model Adequacy Checking 7](#_Toc507082679)
5. [Final model: 13](#_Toc507082680)
6. [Conclusion and Interpretation 13](#_Toc507082681)

# 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")

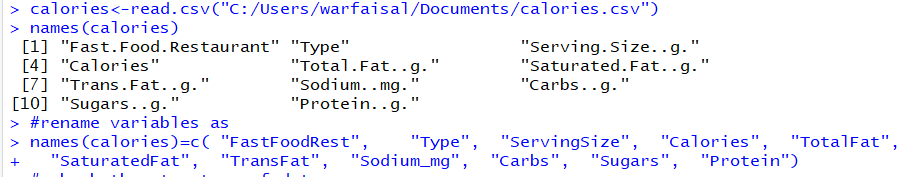
library(car)

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

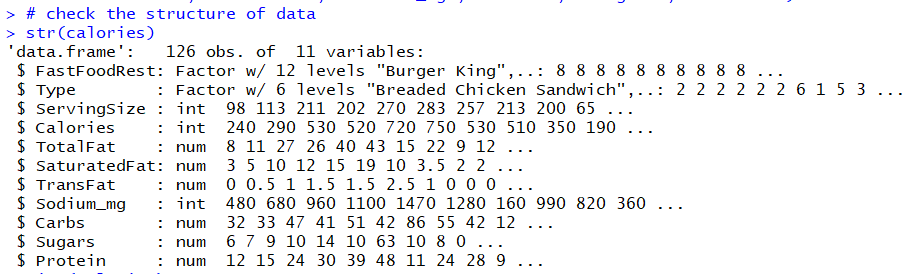
names(calories)



We can see the different variables are

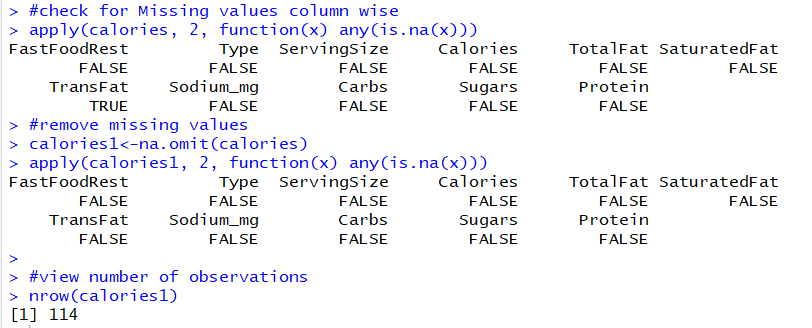
1. FastFoodRest: which has the name of restaurant e.g. McDonalds, Wendy, Sonic etc.
2. Type: has the types of restaurants e.g. burger, MilkShake, Grilled Chicken etc.
3. ServingSize: contains the serving size in grams
4. Calories: has the number of calories per Serving Size
5. TotalFat: sum of saturated, 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.

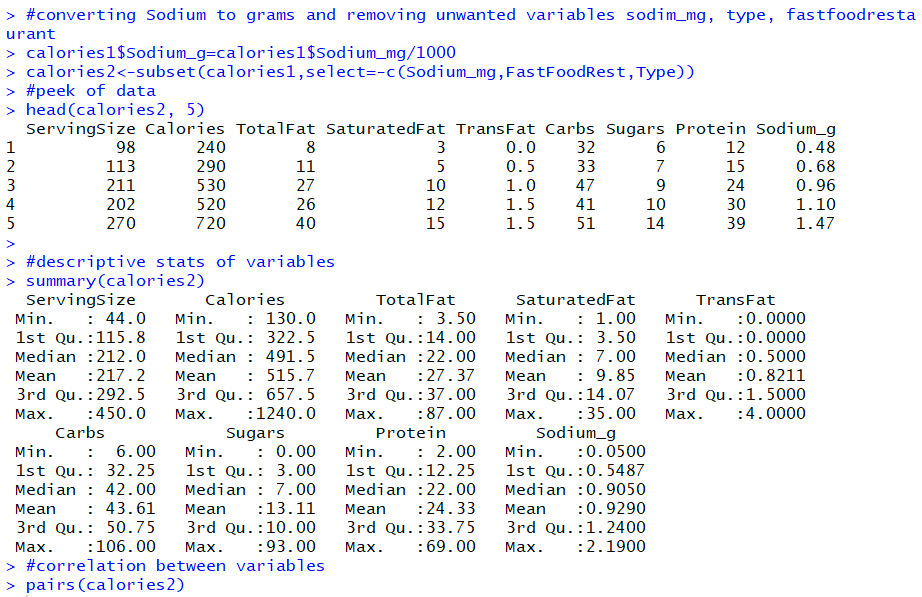


Lets do some cleaning in data by checking and removing the Missing values.

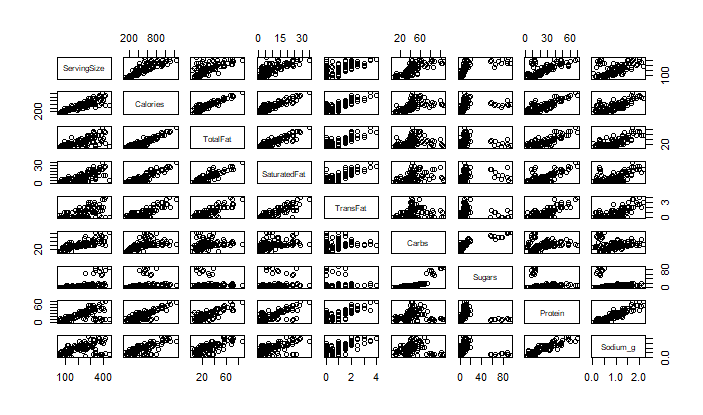
TransFat has missing values, after removing the NA values , we are left with 114 observations which is not a bad number



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.



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.

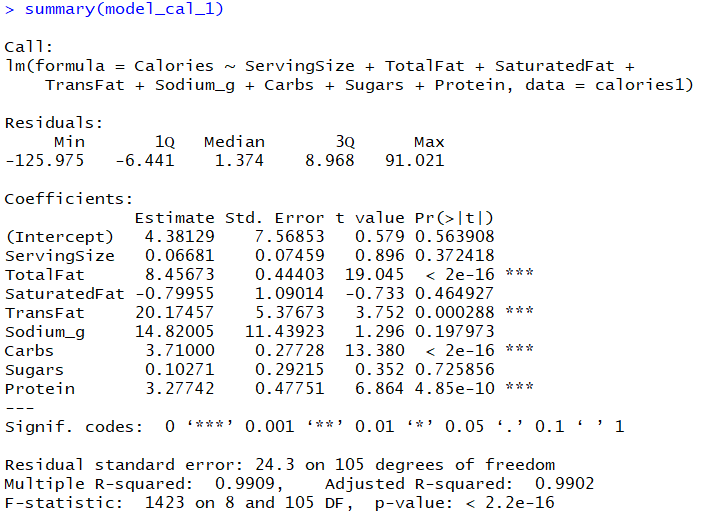


# Model Building

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

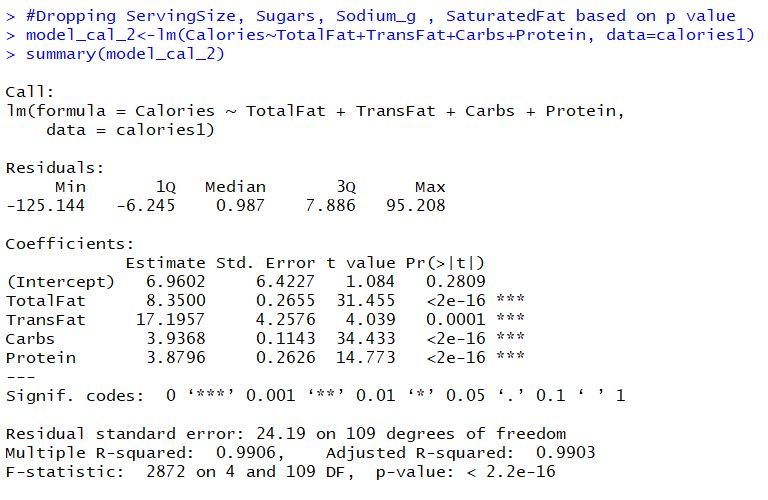


Let’s look at the summary statistics of our model



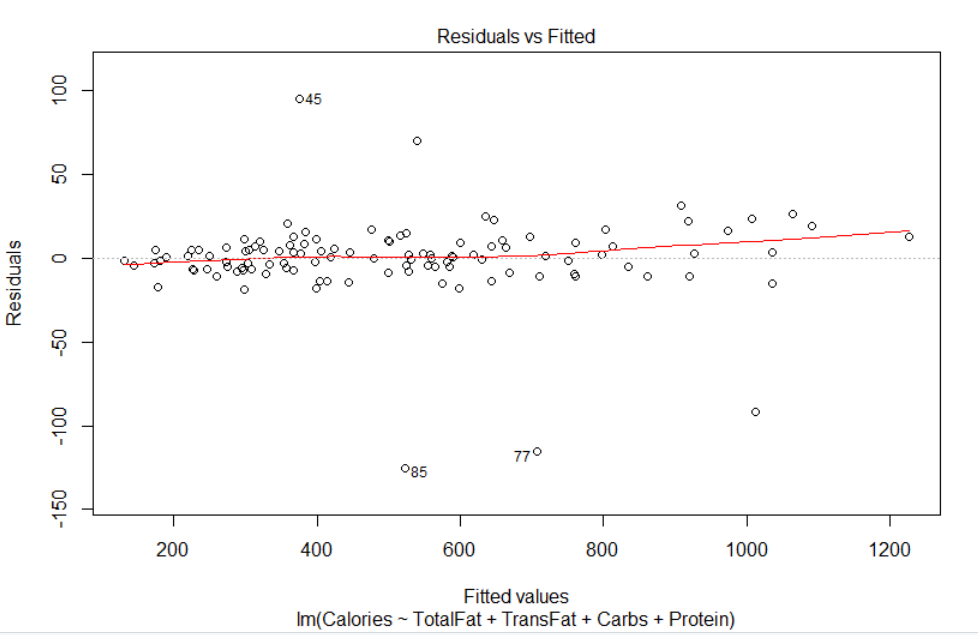
ServingSize, SaturatedFat, Sodium\_g and Sugars are not good covariates as there p values are > 0.05 thus we cannot reject the null hypothesis. Which implies in this model there is not a linear relationship between Calories and ServingSize, SaturatedFat, Sodium\_g and Sugars.

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.



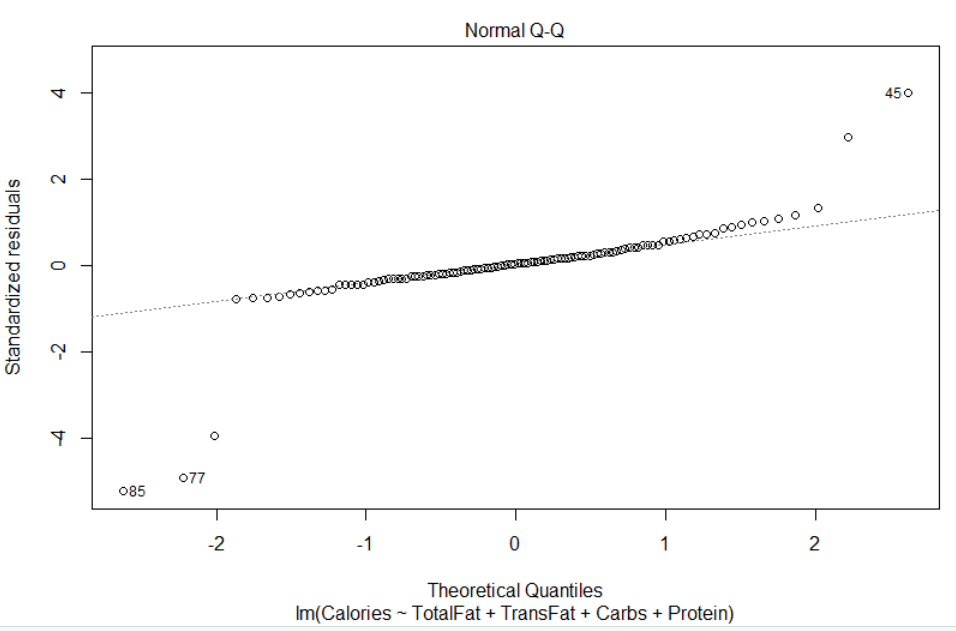
# Model Adequacy Checking

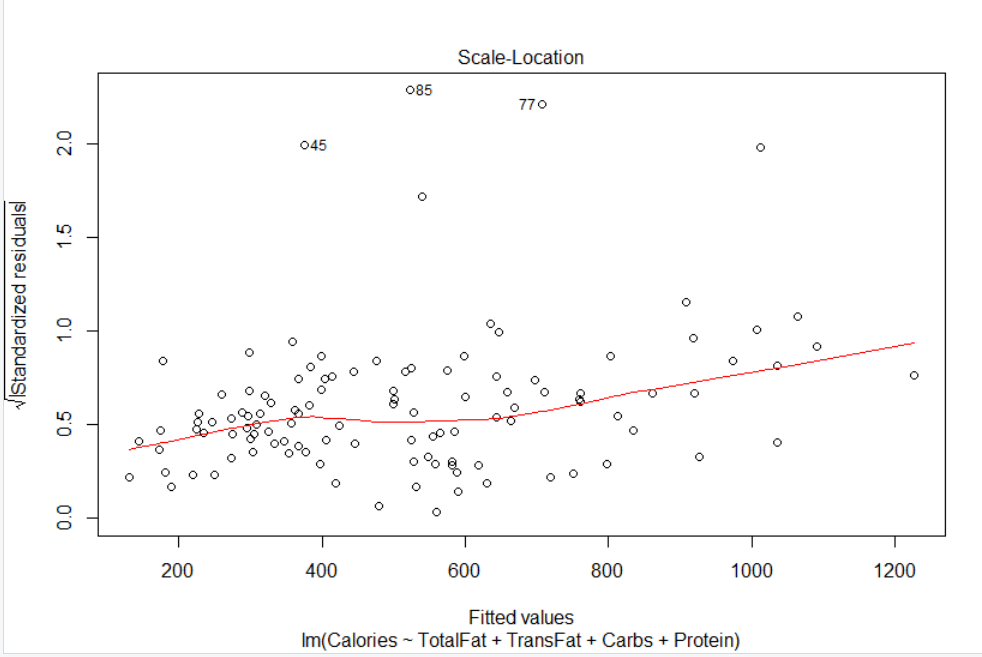
Now let's check the adequacy of our fitted model.

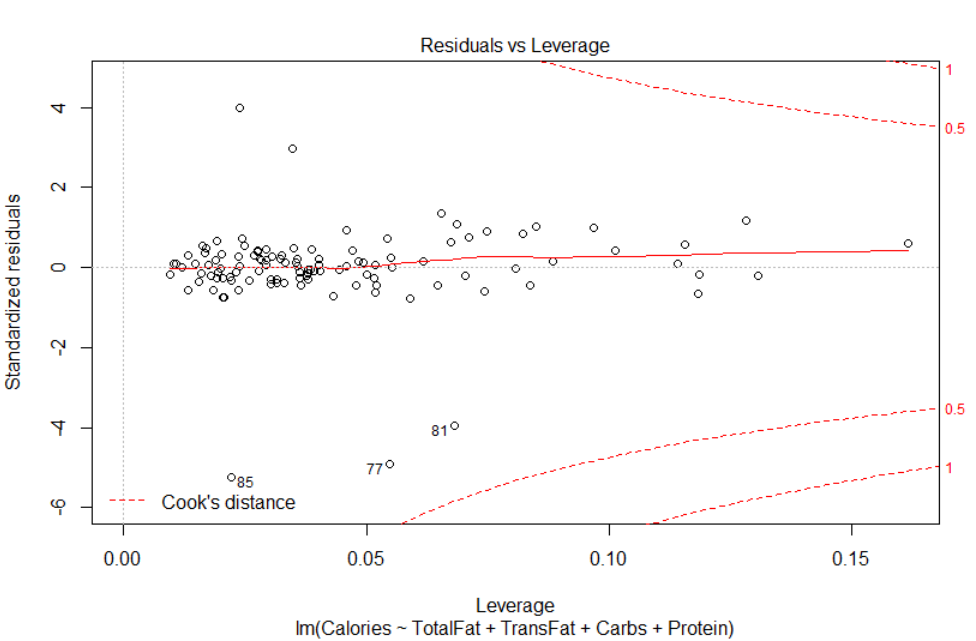


Above in the first plot we have Residual vs Fitted Values , we dont see any pattern on the red line Thus residuals are linearly distributed over fitted values and we can say approximately that variance is equal.

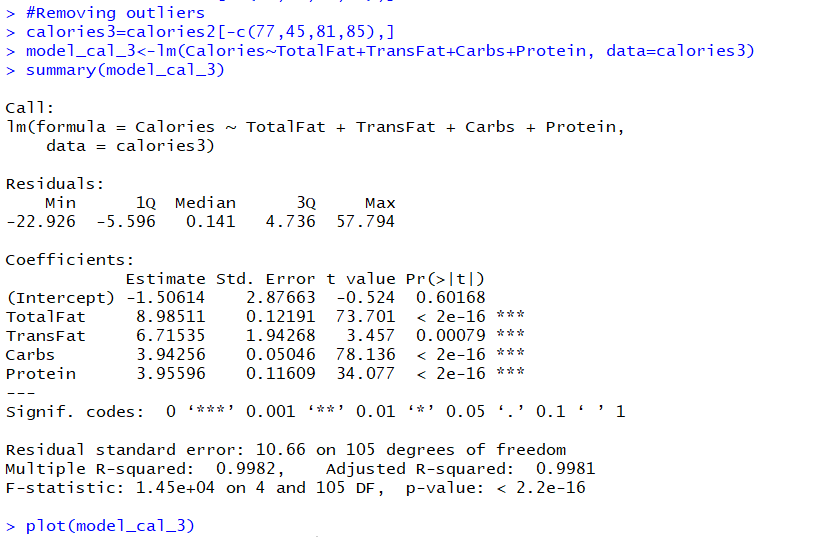
QQ plot is fairly linear except few outliers. Standardized residuals mostly follow the fitted model line. Thus, meeting our normality assumption



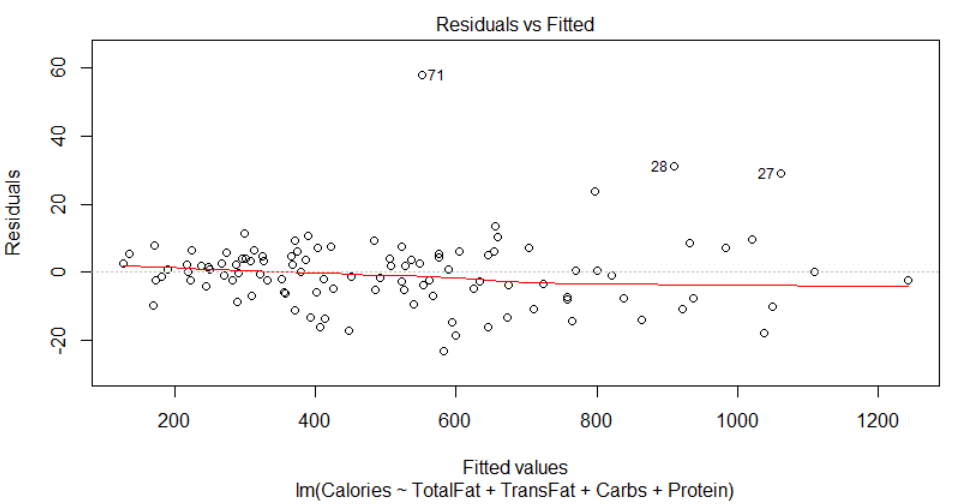


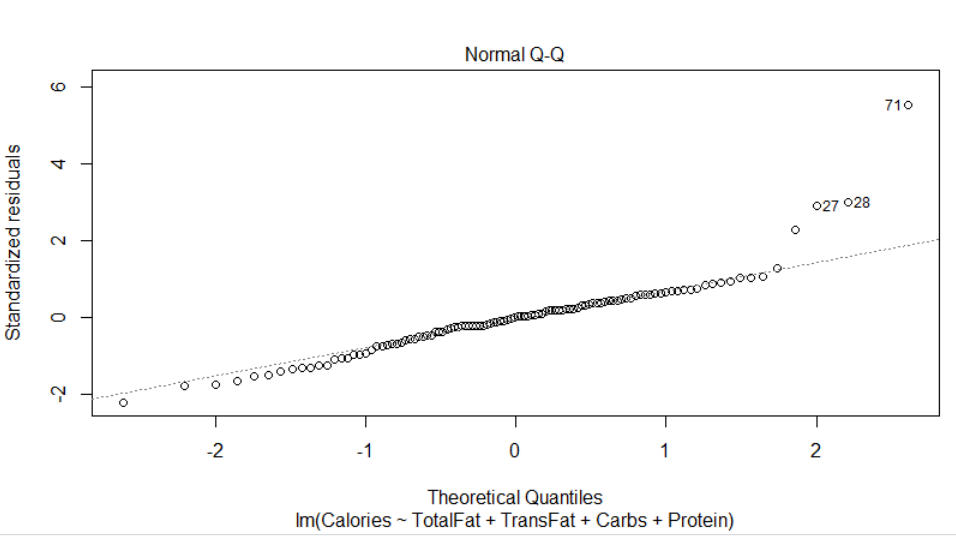


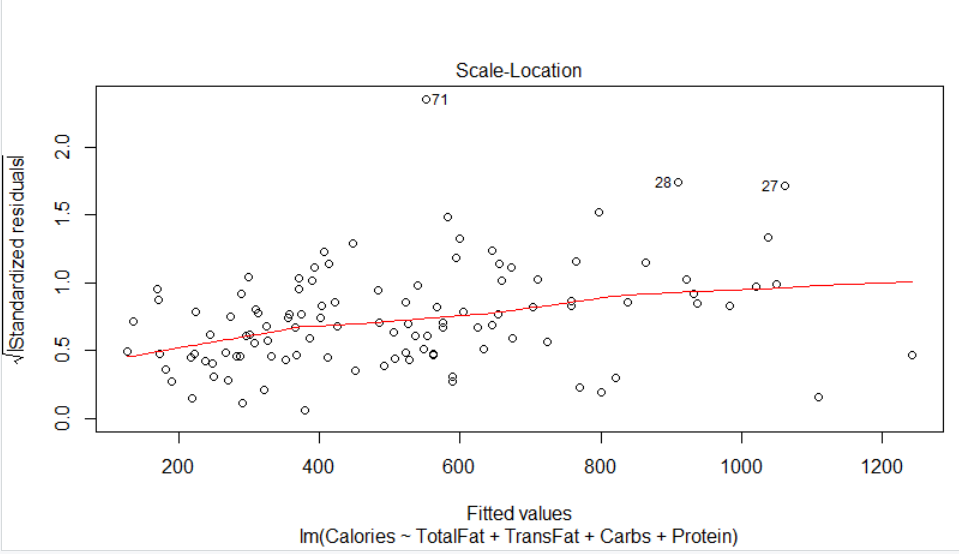
From the above graphs we can see there are few outliers in our model . To fix those let’s remove them and build another model

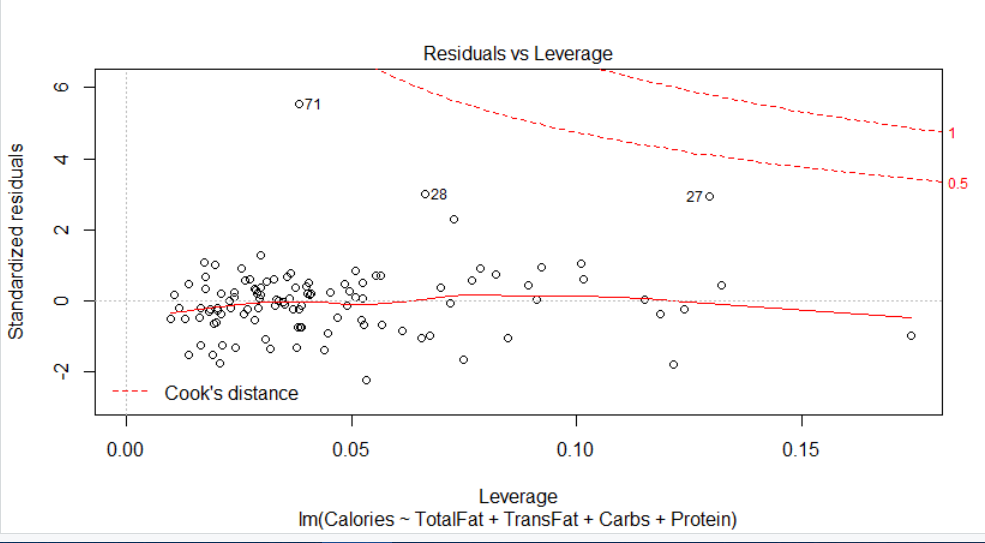


After removing building the new model we can see a fair increase in F statistics, Adjusted R-square value. Let's recheck the model adequacy with the help of Plot(model\_cal\_3) function.









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.



# 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.