Fase 4: Modelación de datos

En la realización de este trabajo se tuvo que realizar varias veces todos los procesos durante el desarrollo hasta esta entrega final en la cual se tuvo que realizar los mismos procedimientos con las gráficas para marcar la regresión lineal considerando cada una de las variables con las que se decidió trabajar por lo que durante el trabajo general se realizó por lo menos unas 10 veces y en esta ultima parte se tuvo que repetir unas 5 veces.

A continuación, se mostrará el código en Python para realizar este trabajo:

import pandas as pd

import statsmodels.api as sm

import statsmodels.formula.api as smf

import matplotlib.pyplot as plt

from sklearn.linear\_model import LinearRegression

dt = pd.read\_csv("A01369947\_EtiquetasNutrimentales.csv")

calorias = dt["Calorias (kcal)"]

carbohidratos = dt["Carbohidratos (g)"]

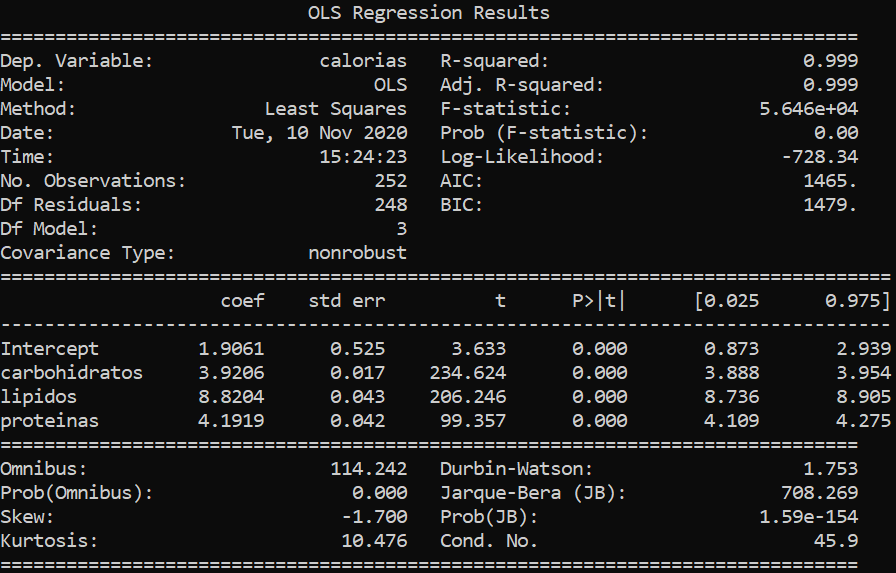
lipidos = dt["Lípidos (g)"]

proteinas = dt["Proteína (g)"]

#sodio = dt["Sodio (mg)"]

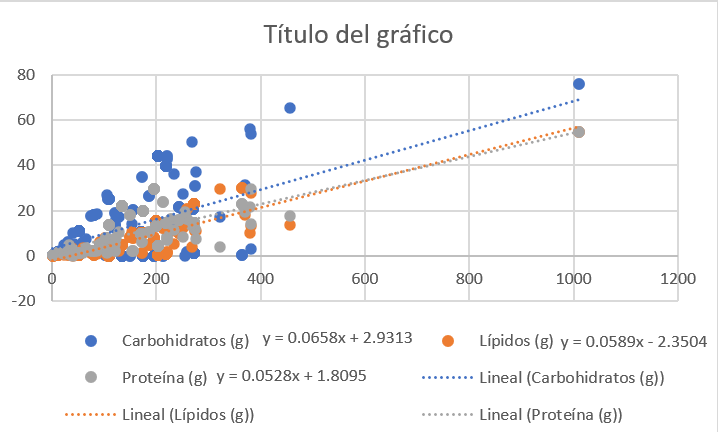
mod = smf.ols('calorias ~ carbohidratos + lipidos + proteinas', data=dt).fit()

print(mod.summary())



Podemos darnos cuenta de que los lípidos afectan en gran medida a la variable dependiente, seguido de las proteínas, los carbohidratos y al final la constante.

A continuación, se muestran las gráficas resultantes del análisis de cada variable:



Podemos percatarnos con la grafica que todos los datos tienden a irse a la izquierda a excepción de unos que se separan de la tendencia y que tienen relación con los alimentos chatarra, a diferencia del análisis inicial mi alimentación era muy variada pero en el último mes he estado repitiendo alimentos de forma sincronizada semanalmente y por días por lo que se ve muy reflejada la grafica actual con la inicial.

# Referencias

Seabold, Skipper, and Josef Perktold. “[statsmodels: Econometric and statistical modeling with python.](http://conference.scipy.org/proceedings/scipy2010/pdfs/seabold.pdf)” Proceedings of the 9th Python in Science Conference. 2010.