Multiple Regression Analysis of Body Mass Index Determinants

Students:

Ana Arabuli

Paula María Montoro Almagro

"A project/thesis submitted to the Faculty of the FABIZ master program Business Administration in fulfilment of the requirements for the subject of Research Methods for Business Administration"

Business Administration

Lecturers: Petre Caraiani

W.K. Härdle

Dan Traian Pele

Bucharest University of Economic Studies

Bucharest, 2025

Abstract

This research aims to examine how different demographic and lifestyle factors influence the Body Mass Index (BMI) of individuals. The dataset includes 1,134 participants, with variables such as sex, age, tobacco consumption, and smoking habits. A descriptive analysis revealed an average BMI of 44.34, showing a wide variation among individuals. Regression and correlation analyses were conducted to identify the direction and strength of each factor's impact on BMI.

The results indicate that sex has a positive influence on BMI, meaning that women generally have higher BMI values than men. In contrast, age, tobacco consumption, and smoking habits show negative effects. As age increases, BMI tends to decrease slightly, suggesting that older individuals have lower body mass compared to younger ones. Tobacco consumption has a minor negative effect, while smoking habits show the strongest negative relationship — daily and occasional smokers have significantly lower BMI compared to non-smokers.

Overall, the study demonstrates that both demographic characteristics and lifestyle behaviors play a significant role in shaping body composition. Understanding these relationships is important for promoting healthier habits and preventing obesity-related health problems in the population.

Keywords: Body Mass Index (BMI), Sex Differences, Age Groups, Tobacco Consumption, Smoking Habit, Regression Analysis, Correlation.

Table of Contents

List of Figures and Tables

List of Code

List of Table

List of Graphs

List of Abbreviations

BMI	Body Mass Index

Chapter 1. Introduction

The Body mass index: Is a measure very used in medicine to verify the adequacy of a person's body weight in relation to its height, and is used to identify weight categories that can lead to health problems. The WHO (World health organization) has created this concept to classify people in weight risk categories, which could be an indicator to prevent future health problems.

Relating population health to business administration, it has been shown that the worsening of population health has a direct impact on reduced productivity, absenteeism, and increased public spending on health (as we were able to verify in the recent COVID pandemic, in which a significant recession can be analyzed), directly affecting a country's economy.

Analyzing some of the individual factors that affect the BMI, there are big biological differences between men and women, also the relation between smoking and weight can be ambiguous, because the consume of nicotine is related to a low BMI, however a big amount of tobacco consumed is related to an unhealthy lifestyle, which we normally associate with greater weight.

Multiple studies have analyzed the effect of individual factors such as diet, exercise or genetics, however this study seeks to answer the following question: ¿How can factors such as age, sex or tobacco consume affect a person’s BMI ?

To analyze how these factors affect things, we will use a linear regression model, based on what we learned in the course.

The objective of this study is to validate a model that helps us understand the relationship between demographic factors (such as age, sex), tobacco consumption and people's body mass index, providing answers to potential problems in other studies.

Chapter 2. Methodology

We need to respond quickly to the challenges of sustainable development, so that public policy can change to fit the new reality. All of this needs changes to make sure there is enough money for environmentally friendly, steady, and fair economic growth. The financial system is very important for this process.

This chapter discusses sustainable development based on Green Finance (2.1). In subchapter 2.2, the Green Loans concept is discussed, also, Green Banking practices and banks' sources of Green Financing are discussed (2.3). This subsection presents customer pressure and Green Banking adoption (2.4). Also, this chapter discusses the relationship between Green Loans and consumer's loyalty (2.5). Sustainable Finance Taxonomy for Georgia is discussed is subchapter 2.6, Taxonomy adoption and implementation (2.7) and the practice of Green Loans in Georgia (2.8) are discussed in these subsections.

2.1 Data

For this project we have relied on the Eurostat, a free public database which contains information at the European level on economic, social and demographic aspects. All of this data is from 2019.

The data has been compiled in Excel and processed in Python after cleaning up missing data, obtaining 1,134 observations, which is a fairly large sample size and gives us reliability when analyzing the results.

2.2 Selected variables

1. Dependent variable Y: It is a continuous quantitative variable representing each individual's Body Mass Index, that has been measured in % . The objective of the model is

to explain this variable. The ranges used for this study and defined by the WHO (World Health Organization) are:

Underweight →Less than 18,5

Normal →18,5-24,9

Overweight→25-29,9

Obesity→30 or more

2. Independent variables Xi:

   a. Age: Is a qualitative continuous variable that has been measured in the following 7 intervals:

From 15 to 19 years

From 20 to 24 years

From 25 to 34 years

From 35 to 44 years

From 45 to 54 years

From 55 to 64 years

65 years or over

b. Sex: qualitative variable, male or female. To include it in the OLS model, it was transformed into a dummy variable where woman=1 and man=0.

c. Tobacco consumption: Is a quantitative variable measured as a percentage from 0 to 100%.

d. Smoking: describing the habit/type of smoking

Non-smoker = 0

Daily smoker = 1

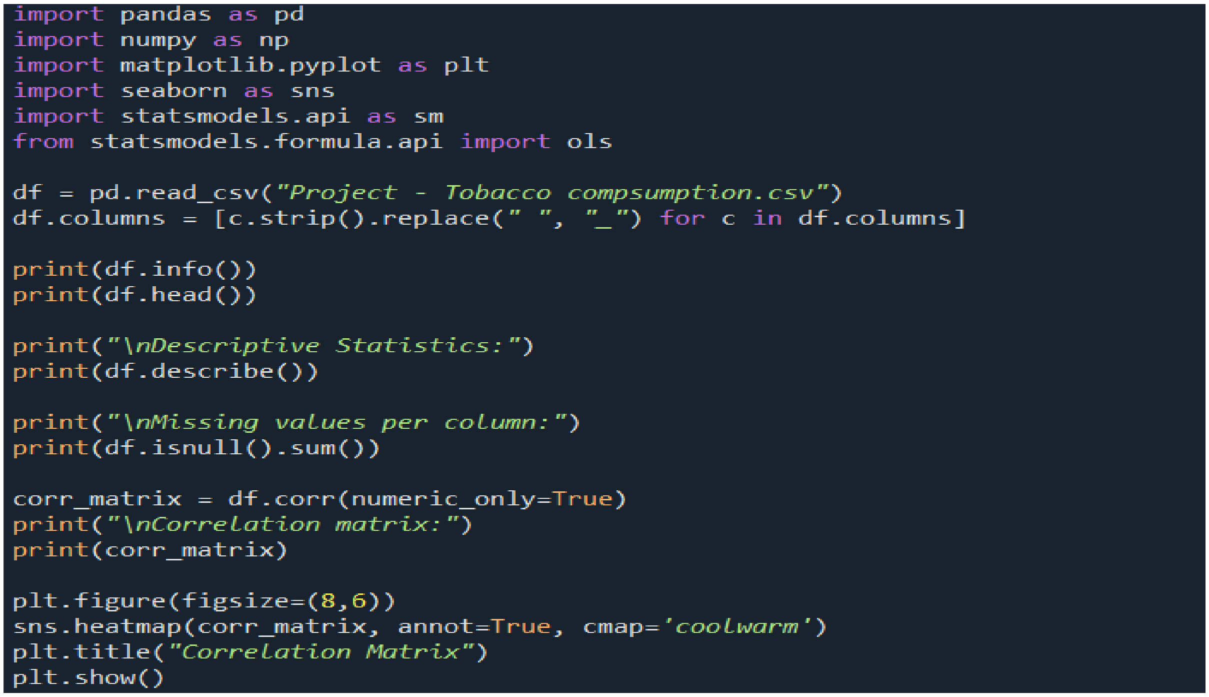
Occasional smoker = 2

Chapter 3. Methods

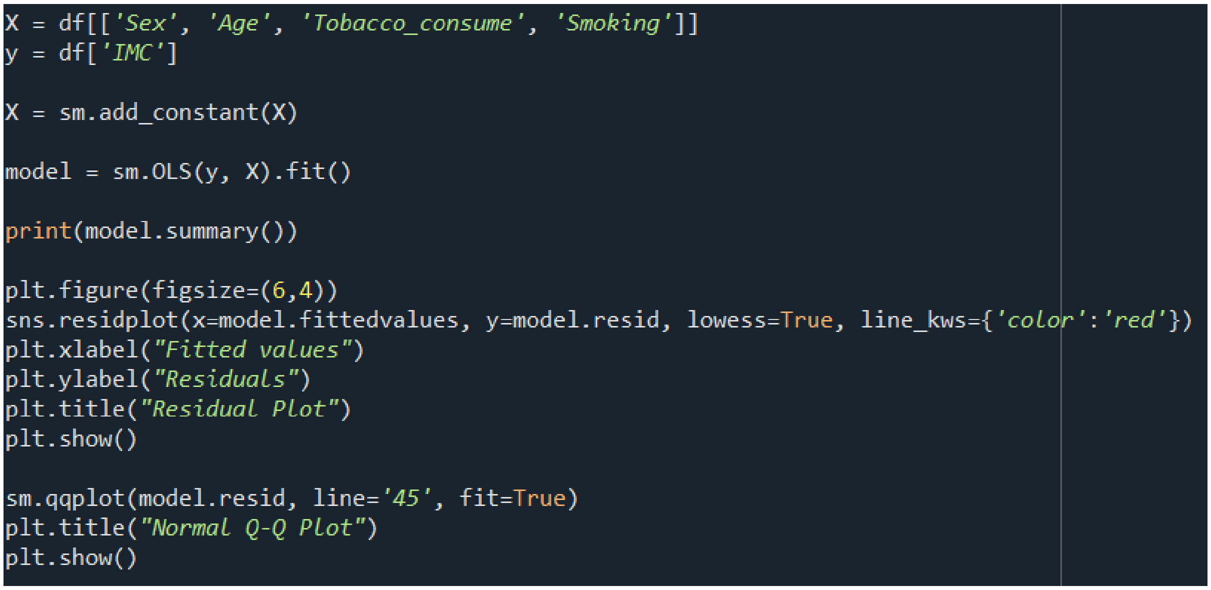
The analysis began by examining the dataset, which included 1,134 individuals with variables such as BMI, sex, age, tobacco consumption, and smoking habits. First, the data was checked for missing values and variable types to ensure completeness and correctness. Descriptive statistics, including mean, median, minimum, maximum, and standard deviation, were calculated to understand the general distribution and variability of each variable.

Next, as we can see down, a correlation analysis was conducted to explore the relationships between BMI and the independent variables. This provided a preliminary understanding of which factors might be associated with higher or lower BMI and the strength of these relationships. Based on these observations, a linear regression model using Ordinary Least Squares (OLS) was planned. BMI was set as the dependent variable, while sex, age, tobacco consumption, and smoking habits were included as independent variables. The purpose of this regression was to quantify the effect of each factor on BMI while controlling for the influence of the others. This approach allowed for a systematic evaluation of how demographic and lifestyle factors contribute to variations in BMI.

Code #3.1 Spyder codes



```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
from statsmodels.formula.api import ols

df = pd.read_csv("Project - Tobacco compsumption.csv")
df.columns = [c.strip().replace(" ", "_") for c in df.columns]

print(df.info())
print(df.head())

print("\nDescriptive Statistics:")
print(df.describe())

print("\nMissing values per column:")
print(df.isnull().sum())

corr_matrix = df.corr(numeric_only=True)
print("\nCorrelation matrix:")
print(corr_matrix)

plt.figure(figsize=(8,6))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm')
plt.title("Correlation Matrix")
plt.show()
```

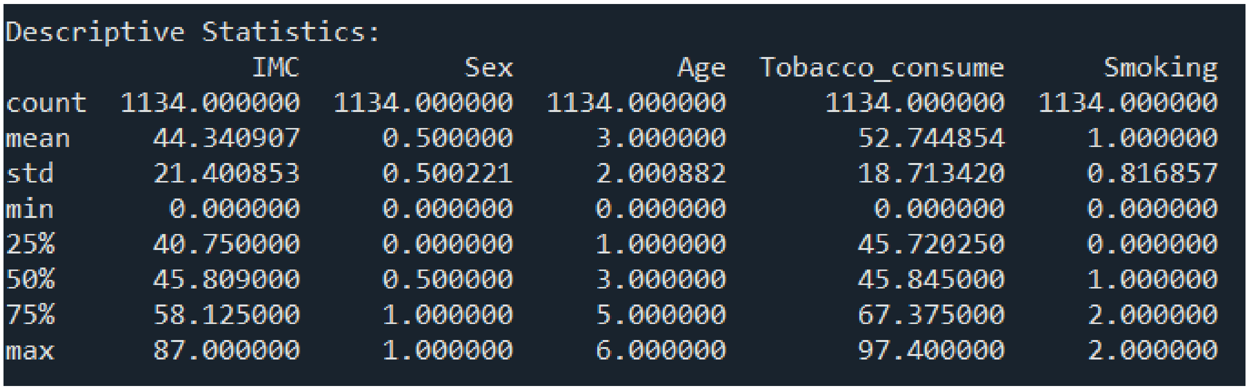


```
X = df[['Sex', 'Age', 'Tobacco_consume', 'Smoking']]
y = df['IMC']

X = sm.add_constant(X)

model = sm.OLS(y, X).fit()

print(model.summary())

plt.figure(figsize=(6,4))
sns.residplot(x=model.fittedvalues, y=model.resid, lowess=True, line_kws={'color':'red'})
plt.xlabel("Fitted values")
plt.ylabel("Residuals")
plt.title("Residual Plot")
plt.show()

sm.qqplot(model.resid, line='45', fit=True)
plt.title("Normal Q-Q Plot")
plt.show()
```

Chapter 4. Analysis of Results

4.1 Descriptive analysis

Table 4.1 Descriptive Statistics

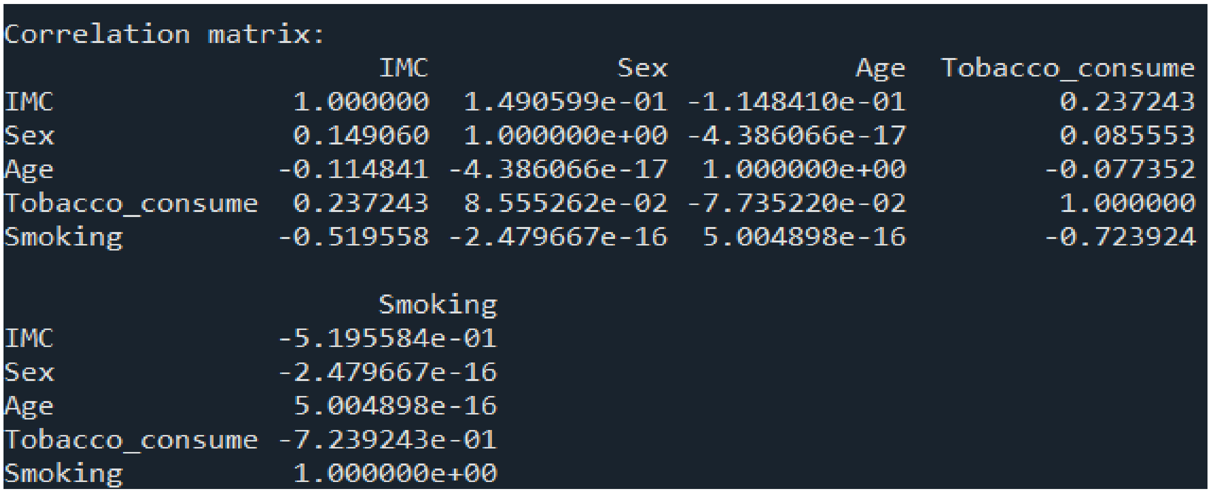


```
Descriptive Statistics:
               IMC          Sex          Age  Tobacco_consume      Smoking
count  1134.000000  1134.000000  1134.000000      1134.000000  1134.000000
mean     44.340907     0.500000     3.000000        52.744854     1.000000
std      21.400853     0.500221     2.000882        18.713420     0.816857
min       0.000000     0.000000     0.000000         0.000000     0.000000
25%      40.750000     0.000000     1.000000        45.720250     0.000000
50%      45.809000     0.500000     3.000000        45.845000     1.000000
75%      58.125000     1.000000     5.000000        67.375000     2.000000
max      87.000000     1.000000     6.000000        97.400000     2.000000
```

The dataset includes 1,134 individuals with variables such as BMI (IMC), sex, age, tobacco consumption, and smoking habits. A preliminary descriptive analysis showed that the average BMI was 44.34, with a standard deviation of 21.40, indicating considerable variation among individuals. Sex was evenly distributed, with a mean of 0.5, while age had a mean value of 3, reflecting the categorical coding used in the dataset. Tobacco consumption varied widely, with an average of 52.74 units and a standard deviation of 18.71. Smoking behavior showed clear differences among individuals, with values ranging from 0 to 2.

4.3 Correlation Analysis

Table # 4.2 correlation Matrix

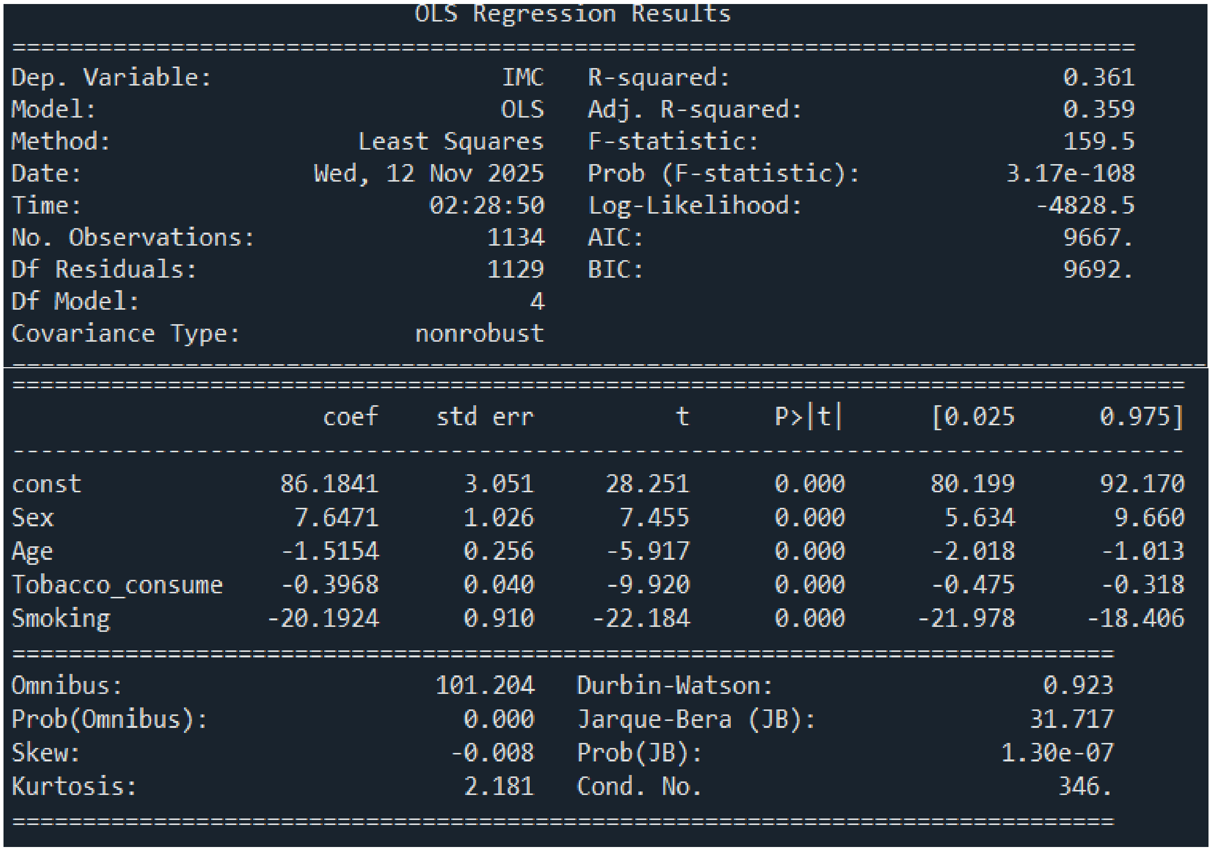


```
Correlation matrix:
                      IMC           Sex           Age  Tobacco_consume
IMC              1.000000  1.490599e-01 -1.148410e-01         0.237243
Sex              0.149060  1.000000e+00 -4.386066e-17         0.085553
Age             -0.114841 -4.386066e-17  1.000000e+00        -0.077352
Tobacco_consume  0.237243  8.555262e-02 -7.735220e-02         1.000000
Smoking         -0.519558 -2.479667e-16  5.004898e-16        -0.723924

                      Smoking
IMC             -5.195584e-01
Sex             -2.479667e-16
Age              5.004898e-16
Tobacco_consume -7.239243e-01
Smoking          1.000000e+00
```

A correlation analysis revealed that smoking had a strong negative relationship with BMI (-0.52), suggesting that individuals who smoke more tend to have lower BMI values. Tobacco consumption also showed a negative correlation (-0.24), while sex and age exhibited weaker associations (0.15 and -0.11, respectively).

4.4 Regression Analysis

Table # 4.3 Regression



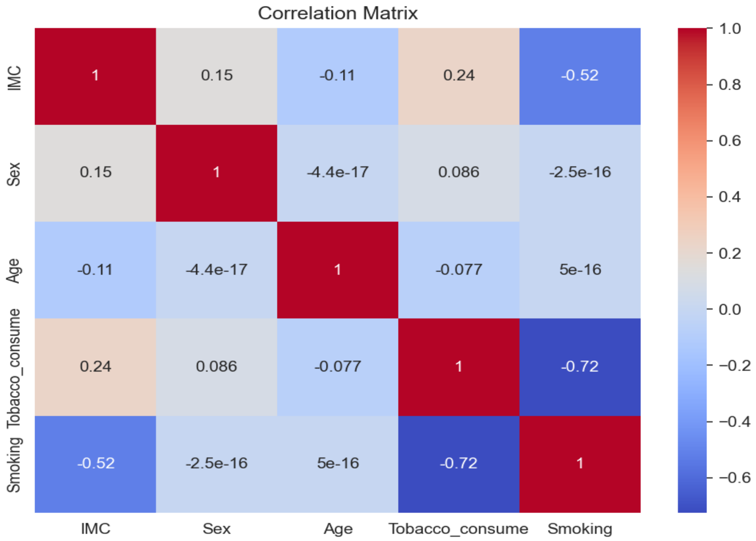
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    IMC   R-squared:                       0.361
Model:                            OLS   Adj. R-squared:                  0.359
Method:                 Least Squares   F-statistic:                     159.5
Date:                Wed, 12 Nov 2025   Prob (F-statistic):          3.17e-108
Time:                        02:28:50   Log-Likelihood:                -4828.5
No. Observations:                1134   AIC:                             9667.
Df Residuals:                    1129   BIC:                             9692.
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
===================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------
const              86.1841      3.051     28.251      0.000      80.199      92.170
Sex                 7.6471      1.026      7.455      0.000       5.634       9.660
Age                -1.5154      0.256     -5.917      0.000      -2.018      -1.013
Tobacco_consume    -0.3968      0.040     -9.920      0.000      -0.475      -0.318
Smoking           -20.1924      0.910    -22.184      0.000     -21.978     -18.406
==============================================================================
Omnibus:                      101.204   Durbin-Watson:                   0.923
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               31.717
Skew:                          -0.008   Prob(JB):                     1.30e-07
Kurtosis:                       2.181   Cond. No.                         346.
==============================================================================
```

Based on the patterns observed in the descriptive and correlation analyses, a linear regression model using Ordinary Least Squares (OLS) was conducted. In this model, BMI was set as the dependent variable, while sex, age, tobacco consumption, and smoking habits were included as independent variables.

The purpose of this model was to measure how each of these factors influences BMI while accounting for the effects of the others. The model produced an R-squared value of 0.361, which means that approximately 36% of the variation in BMI among individuals can be explained by differences in sex, age, tobacco consumption, and smoking. Also a significant F-statistic for the linear regression model was obtained with a p-value of 3’17*10-108 showing that the linear regression model in general is correct to analyze the relationship between the dependent and independent variables.

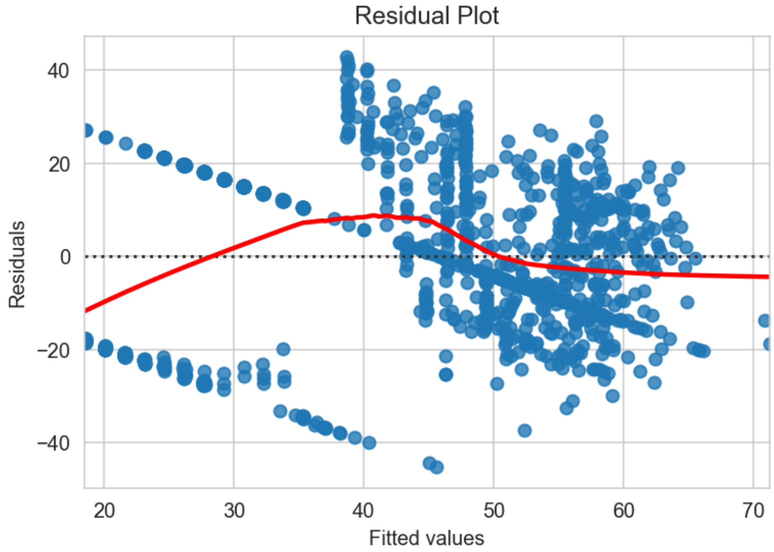
All of the independent variables were statistically significant with p-values less than 0.001, indicating that each factor has a meaningful and measurable impact on BMI. This regression analysis provides a clear and systematic framework for understanding how demographic characteristics and lifestyle behaviors contribute to differences in BMI within the study population, allowing us to identify which factors are most strongly associated with higher or lower BMI values.

Graphs # 4.1 Correlation Matrix



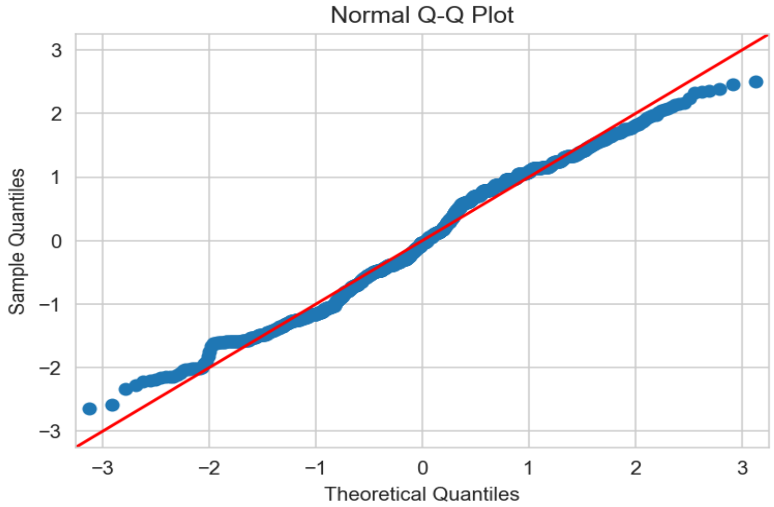
In the graph below, we can see our model fitted to the sample taken. Each blue dot is an observation, while the red line is the model fitted to the observations.

Graphs # 4.2 Residual Plot



The Q-Q plot that has been obtained shows if the residuals of the model follow a Normal distribution, which would be desirable. The red line shows where the points should be if the data were a perfect Gaussian bell curve, and the blue dots are the 1134 observations we have studied.

Graphs # 4.3 Normal Q-Q Plot



Chapter 5. Discussion

The linear regression model was conducted with BMI (IMC) as the dependent variable and sex, age, tobacco consumption, and smoking habits as independent variables. The model achieved an R-squared value of 0.361, indicating that approximately 36% of the variation in BMI can be explained by these factors. This represents a moderate explanatory power, showing that while the selected variables have a noticeable impact on BMI, other factors not included in the model may also influence body weight and could be added to the model to improve analysis and prediction. It might also be more appropriate to use a non-linear regression model, with interactions between variables and polynomials that better fit the data observed in the upper graph.

The estimated coefficients provide further insight into the effect of each variable. The intercept of 86.18 suggests that if all independent variables were zero, the predicted BMI would be approximately 86.18. Sex has a positive coefficient of 7.65, meaning that, on average, males have a BMI about 7.6 points higher than females. Age has a negative coefficient of -1.52, indicating that each increase in the age category is associated with a slight decrease in BMI. Tobacco consumption is associated with a small negative effect (-0.40 per unit), showing that higher tobacco use slightly lowers BMI. The strongest effect was observed for smoking, with a coefficient of -20.19, meaning that smokers, particularly daily smokers, have much lower BMI values on average.In plain terms, these results suggest that smoking habits have a strong influence on BMI, while sex and age have moderate effects. Tobacco consumption also contributes to a slight reduction in BMI. All variables were statistically significant with p-values less than 0.001, providing strong evidence that each factor has a measurable impact on BMI.

Model diagnostics were also assessed. The Durbin-Watson statistic of 0.923 indicates some positive autocorrelation in the residuals, suggesting that the residuals are not fully independent. The Omnibus and Jarque-Bera tests ($p < 0.001$) show that residuals deviate slightly from normality, which is common in large datasets. Overall, the diagnostics confirm that the model is generally valid, and the results can be interpreted with confidence.

Chapter 6 Conclusion

This study examined how demographic factors and lifestyle behaviors, including sex, age, tobacco consumption, and smoking habits, affect BMI (IMC) in a sample of 1,134 individuals. The results from the linear regression analysis showed that all factors have a statistically significant impact on BMI. Smoking habits exhibited the strongest negative effect, indicating that smokers tend to have much lower BMI values. Sex and age also influenced BMI, with males generally having higher BMI and older age groups showing a slight decline. Tobacco consumption had a small but consistent negative effect on BMI.

Overall, the model explained approximately 36% of the variation in BMI, highlighting that while these factors are important, other unobserved variables may also contribute to differences in body weight. These findings provide valuable insight into how lifestyle and demographic characteristics influence BMI and can help inform public health strategies aimed at promoting healthy body weight and preventing obesity-related health issues.

Appendix

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
from statsmodels.formula.api import ols

df = pd.read_csv("Project - Tobacco compsumption.csv")
df.columns = [c.strip().replace(" ", "_") for c in df.columns]

print(df.info())
print(df.head())

print("\nDescriptive Statistics:")
print(df.describe())

print("\nMissing values per column:")
print(df.isnull().sum())

corr_matrix = df.corr(numeric_only=True)
print("\nCorrelation matrix:")
print(corr_matrix)

plt.figure(figsize=(8,6))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm')
plt.title("Correlation Matrix")
plt.show()

X = df[['Sex', 'Age', 'Tobacco_consume', 'Smoking']]
y = df['IMC']

X = sm.add_constant(X)
```

```
model = sm.OLS(y, X).fit()

print(model.summary())

plt.figure(figsize=(6,4))
sns.residplot(x=model.fittedvalues, y=model.resid, lowess=True, line_kws={'color':'red'})
plt.xlabel("Fitted values")
plt.ylabel("Residuals")
plt.title("Residual Plot")
plt.show()

sm.qqplot(model.resid, line='45', fit=True)
plt.title("Normal Q-Q Plot")
plt.show()
```

Bibliography

Eurostat. 2019*. Body mass index (BMI) by sex, age and income quintile.* https://ec.europa.eu/eurostat/databrowser/view/hlth_ehis_bm1i__custom_18830102/default/table

Clínica Universidad de Navarra. 2025. *Índice de masa corporal.* https://www.cun.es/escuela-salud/indice-masa-corporal

https://docs.google.com/spreadsheets/d/2051877316/edit?gid=2051877316#gid=2051877316