Multiple Regression Analysis Of Body Mass Index Determinants

Ana Arabuli
Paula María Montoro Almagro

Research methods for business administration

Agenda

01. Introduction

02. Methodology

03. Analysis

04. Results

05. Conclusion

Introduction

The Body Mass Index (BMI) is the WHO standard for identifying weight categories that can lead to health problems (overweight, obesity...)

Population health, reflected by BMI, has a direct economic impact. Poor public health is linked to:

- Reduced productivity and increased absenteeism.
- Increased public health spending, impacting the economy, as we saw during COVID pandemic.

Objetive: validate a model that helps us understand the relationship between demographic factors, tobacco consumption and people's body mass index, providing answers to potential problems in other studies

Methodology

To create this model we will use a linear regression model using Ordinary Least Squares (OLS)

- Data Source: Eurostat (European-level economic, social, and demographic data).
- Time Period: All data is from 2019.
- Processing: Data was compiled in Excel and processed using Python.
- Sample Size: After cleaning missing data, the final sample consists of 1,134 observations, providing a large and reliable basis for analysis.

Dependent Variable (Y):
BMI: A continuous quantitative measure explained in % in the regression model

The Independent Variables (X) used in the regression model are:

- Sex: A qualitative variable transformed into a dummy (woman=1, man=0)
- % of tobacco consume: A quantitative variable
- Age: Using 7 intervals of age between 15 and >65 years
- Smoking: Describing the habit of smoking (Non-smoker = 0

Daily smoker = 1; Occasional smoker = 2)

Analysis and results

The analysis began by examining the dataset of 1,134 individuals, checking for missing values and correct variable types. Descriptive statistics like mean, median, and standard deviation were calculated to understand the data's general distribution. A correlation analysis was then conducted to explore the preliminary relationships between BMI and the independent variables.

Based on these observations, an OLS linear regression model was planned, setting BMI as the dependent variable. Sex, age, tobacco consumption and smoking habits were included as independent variables to quantify the effect of each factor on BMI while controlling for the influence of the others.

01.

```
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
```

02.

```
df = pd.read_csv("Project - Tobacco
compsumption.csv")
#  Clean column names :
df.columns = [c.strip().replace(" ", "_") for c in df.columns]
# Check data info :
print(df.info())
print(df.head())
```

03.

```
# Check missing values:
print("\nMissing values per column:")
print(df.isnull().sum())
```

If we want to see the basic descriptive statistics:

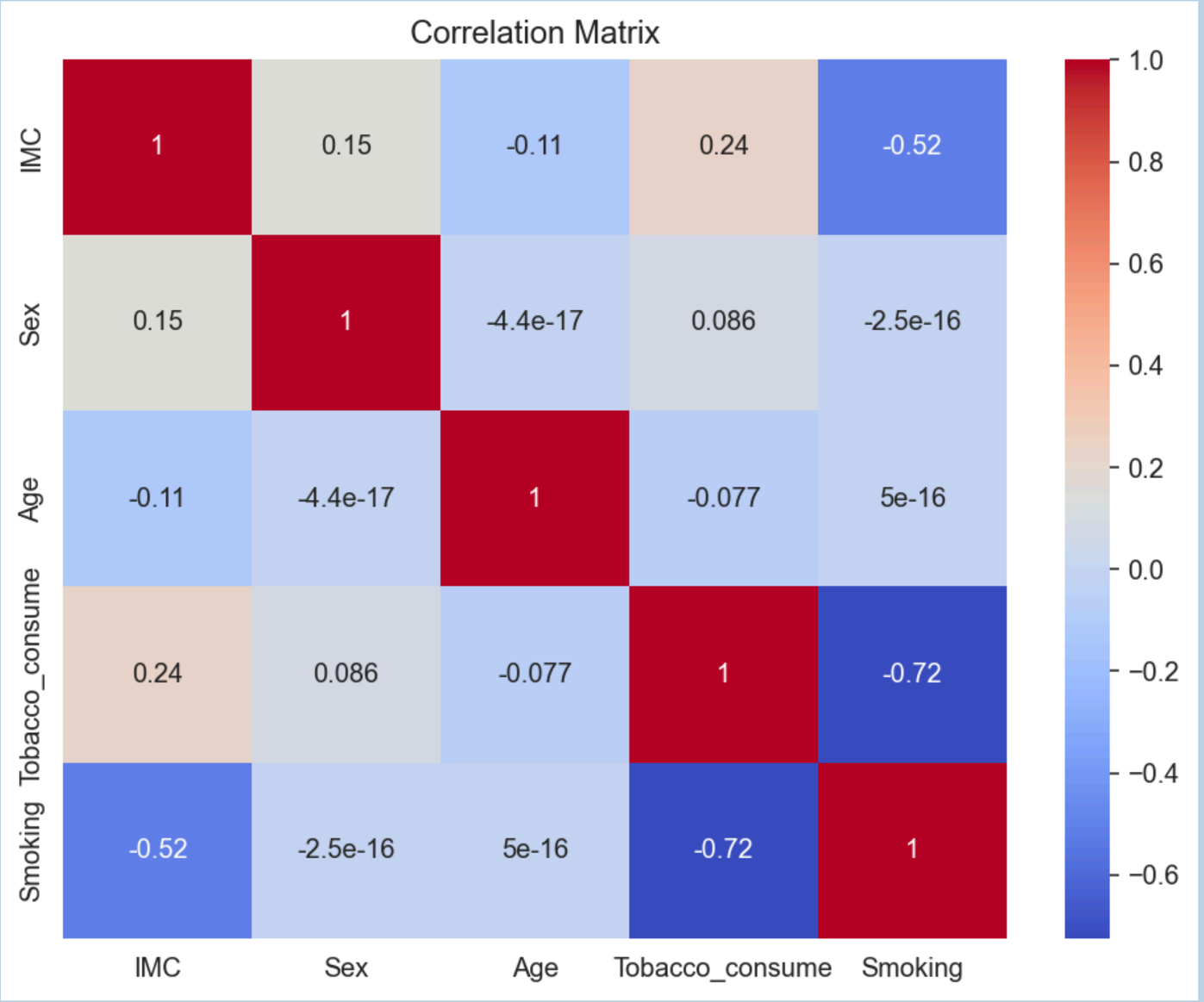
```
print("\nDescriptive Statistics:")
print(df.describe())
```

	MBI	Sex	Age	Tobacco_consume	Smoking Habit
Mean	44,341	0,500	3,000	52,7448	1
Standard error	21,401	0,500	2,001	18,7134	0,816857
Min	0,000	0,000	0,000	0	0
50%	45,809	0,500	3,000	45,845	1
Max	87,000	1,000	6,000	97,4	2

```
# Correlation matrix:
corr_matrix=df.corr(numeric_only=True)
print("\n Correlation matrix:")
print(corr_matrix)
```

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

	MBI	Sex	Age	Tobacco_consume	Smoking Habit
MBI	1,00	0,15	-0,11	0,24	-0,52
Sex	0,15	1,00	-4,386066 *10^17	0,09	-2,479667 *10^16
Age	-0,11	-4,38606 * 10^17	1,00	-0,08	5,004898 *10^16
Tobacco_consume	0,24	0,09	-0,08	1,00	-0,72
Smoking Habit	-0,52	-2,479667 * 10^16	5,004898 *10^16	-0,72	1,00


Correlation Matrix
IMC
Sex
Age
Tobacco_consume
Smoking
1
0.15
-0.11
0.24
-0.52
0.15
1
-4.4e-17
0.086
-2.5e-16
-0.11
-4.4e-17
1
-0.077
5e-16
0.24
0.086
-0.077
1
-0.72
-0.52
-2.5e-16
5e-16
-0.72
1
IMC
Sex
Age
Tobacco_consume
Smoking
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

Code to make the regression model

01.

```
# Regression Model with constant:

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

y = df [ 'IMC' ]

X = sm.add_constant (X)
```

02.

```
# Fit regression model:
model = sm.OLS (y , X). fit( )

# Show regression summary:
print (model.summary( ) )
```

03.

```
# Residual plot
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 ( )
```

The regression model:

R-squared = 0,361

$$y = 86,1841 + 7,6471 * Sex - 1,5154 * Age - 0,3968 * Tobacco\ consume - 20,1924 * Smoking\_habit$$

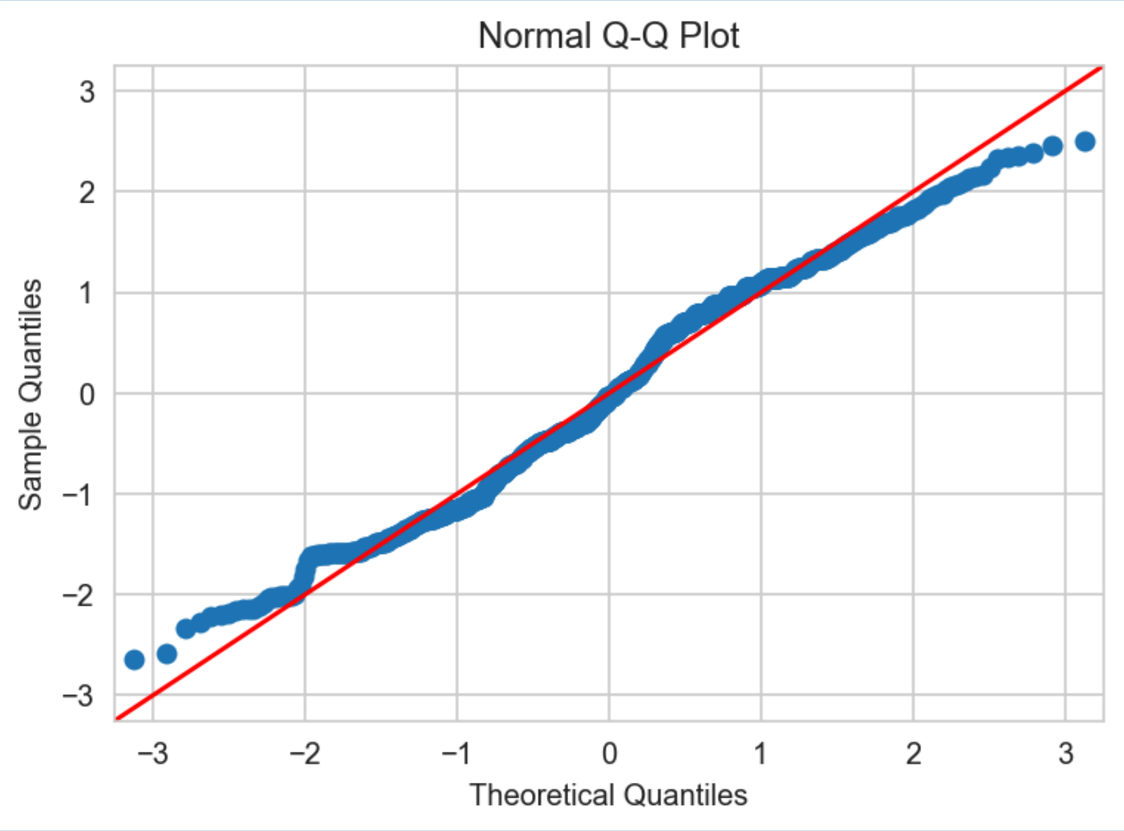
	coef	std err	t	P>|t|	[0.025	0.975]
Constant	86,1841	3,051	28,251	0,000	80,199	92,170
Sex	7,6471	1,026	7,455	0,000	5,634	9,660
Age	-1,5154	0,256	-5,917	0,000	-2,018	-1,013
Tobacco_consume	-0,3968	0,040	-9,920	0,000	-0,475	-0,318
Smoking Habit	-20,1924	0,910	-22,184	0,000	-21,978	-18,406

Residual plot vs. fitted values

In the graph from the right, 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.

```
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 ( )
```

Q-Q 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.

It can be seen that the majority of blue dots fall almost perfectly on the red line. This indicates that most of the residuals behave exactly as would be expected in a normal distribution. But at the end of the line we can observe that the blue dots move away slightly from the red line, however it doesn't make the model invalid.

```
# Normality check for residuals:
sm.qqplot ( model.resid , line=' 45 ', fit= True )
plt.title ( " Normal Q-Q Plot " )
plt.show ( )
```

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.

Thank you!