Estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico

DATA 690 Final Report

By

Sandeep Kemidi,

Sada Siva Madhav

Divya Sai Balaji

December 2023

**Estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico**

# Background

The dataset describes about the obesity levels and combination of life style and food

habits contributing to the obesity levels. The data is collected from the countries Mexico,

Peru and Columbia. Some of the factors like whether the individual is using any electronic

device to monitor calorie burns, amount of water drank, habit of having high caloric food,

habit of having vegetables, number of main meals taken per day etc are chosen of which key

factors can be decided that contributes to obesity levels. In this statistical analysis, initially

checks like sanity checks are performed, later on proceeded with the exploratory data analysis

to know what the data is, and to get insights from the data. Later on we headed on to know if

the number of meals taken per day actually effects and be a contributing factor to get obese.

We even did hypothesis testing on every variable to check its importance in the data against

obesity level.

**Null Hypothesis (H0):** The null hypothesis would state that there is no significant difference in obesity levels based on the number of main meals an individual consumes. In statistical terms, this means that the coefficient for NCP in your model is equal to zero, implying no effect.

**H0:** The coefficient for NCP is equal to 0 (NCP has no significant effect on obesity levels).

**Alternative Hypothesis (H1)**: The alternative hypothesis posits that there is a significant difference in obesity levels based on the number of main meals consumed. This means that the coefficient for NCP is not equal to zero, indicating an effect.

**H1:** The coefficient for NCP is not equal to 0 (NCP has a significant effect on obesity levels).

# Data

The data contains 17 columns and 2,111 rows, the records are labeled with the class

variable NObesity (Obesity Level), that allows classification of the data using the values of

Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type

I, Obesity Type II and Obesity Type III. These different levels are squeezed to four levels by

combining Overweight Level I, Overweight Level II to “Overweight”, Obesity Type I,

Obesity Type II and Obesity Type III to “Obesity”. Finally, the NObesity column has the

values Overweight, Obesity, Normal Weight, Insufficient Weight. The dataset includes

demographic details such as age, gender, and family history of overweight, providing a

foundation for understanding the distribution of obesity across different groups. Nutritional

Factors: Variables related to eating habits, including the consumption of food types (FAVC),

frequency of meals (FCVC), and adherence to specific diets (CAEC), offer insights into the

dietary patterns contributing to obesity. Lifestyle and Behavior: Information on smoking

habits (SMOKE), physical activity frequency (FAF), and daily water intake (CH2O) allows

for the exploration of lifestyle and behavioral factors associated with obesity. Medical

History: The presence of self-reported medical conditions, such as SCC (self-reported

cardiovascular diseases), adds a layer of understanding regarding the potential impact of

existing health issues on obesity. Physical Attributes: Data on height, weight, and Body Mass

Index (BMI) contribute to the core measurements of obesity, enabling a quantitative

assessment of participants&#39; physical conditions.

 Frequent consumption of high caloric food (FAVC)

 Frequency of consumption of vegetables (FCVC)

 Number of main meals (NCP)

 Consumption of food between meals (CAEC)

 Consumption of water daily (CH20)

 Consumption of alcohol (CALC)

 Calories consumption monitoring (SCC)

 Physical activity frequency (FAF)

 Time using technology devices (TUE)

 Transportation used (MTRANS)

Below is the overview of the data:

A screenshot of a computer

Description automatically generated

The summary of the data pointing the mean, min, max, standard deviation, quantiles of the data distribution.

A screenshot of a graph

Description automatically generated

**EDA**

A graph with a blue and orange rectangle

Description automatically generated

The insights from the above graph is that the number of individuals that have a overweight in their family history is about 1700 while those that does not have any background with the overweight is about 300.

A blue and orange rectangular graph

Description automatically generated

SCC defines the individual’s habit of using electronic gadgets through which they monitor the calories they burn, exercise etc. Our data show that out of the individuals, the data collected from, most of them, about 2000 have no habit of monitoring their health while only less than 100 has the habit of tracking their health metrics.

A graph of different colored bars

Description automatically generated

MTRANS refers to the mode of transportation used by the test individuals. These modes are categorised into usage of Public\_Transportation, Walking, Automobile, Motorbike, Bike. The data show that, most of the test individuals use public transportation as their mode, about 1600, followed by Automobile, used by more than 450. The key is to observe the physical activity amongst the individuals, and we found that lesser than 100 are found to be opting walking as their mode of transportation.

A graph of age and age

Description automatically generated

The graph shows the distribution of the age, most of the ages of the individuals are in the range of 20 - 30 while the youngest individual is around 8 years where as the oldest is above 60 years of age.

A graph of different colored columns

Description automatically generated with medium confidence

The graph shows that obesity levels, obesity 1, obesity type 2, obesity type 3 are interlinked with the individual’s family history with overweight, as you can see that there is hardly any evidence that an individual fell into a category of obesity while having no family history with overweight. While the other categories do have a dependency on whether or not the individual has family history with overweight.

# Methods

**Instantiate and Train the Model:**

Use the statsmodels package in Python to create logistic regression models for control vs AD and control vs LOBD diagnoses.

Fit the models using the statsmodels.formula.api object and the fit() function.

**Formula for Logistic Regression**:

Define the logistic regression formula specifying the dependent variable (diagnosis) and independent variables (biomarkers or predictors).

**Comparing Regression Slopes:**

Calculate the significance of the difference between the regression slopes for control vs AD and control vs LOBD.

Use statistical tests (e.g., t-value equation) from the scipy package, such as scipy.stats.t.sf(), to compare correlation coefficients and generate p-values for each biomarker..

Results

We made hypothesis testing on whether NCP i.e., number of main meals taken per day has significant effect over Obesity. Furthermore, we analyzed what are the variables or features that are important factors of obesity.

**NCP vs Obese**

Null Hypothesis:

Alternate Hypothesis:

**A screenshot of a computer

Description automatically generated**

**Obese vs All**

**A screenshot of a computer

Description automatically generated**

Based on the logistic regression results, we can interpret the findings in relation to the number of main meals (NCP) as a factor associated with the probability of being in a certain category of the NObeyesdad variable (which appears to be the obesity level).

Here is the interpretation:

* The coefficient for NCP is approximately − 0.0471.
* The standard error of this coefficient is 0.016.
* The z-score for this coefficient is − 3.017 .
* The p-value associated with this z-score is 0.003.
* The R-squared value is 0.001, indicating that approximately 0.1% of the variance in 'NObeyesdad' can be explained by 'NCP'. This is a very low value, suggesting that 'NCP' does not explain much of the variation in obesity levels.
* The F-statistic is 2.892 with a corresponding p-value of 0.0892, which is marginally above the conventional alpha level of 0.05, indicating that the model is not statistically significant at the 5% significance level.
* The coefficient for 'NCP' is 0.0237, with a standard error of 0.014. The t-value for 'NCP' is 1.701, and the p-value is 0.089, which, again, is not statistically significant at the 5% level (though it is close).
* The confidence interval for the 'NCP' coefficient ranges from -0.004 to 0.051, which includes zero, further suggesting that the effect of 'NCP' on 'NObeyesdad' may not be statistically significant.
* The Omnibus test has a p-value of 0.000, which suggests that the residuals of the model are not normally distributed. The Jarque-Bera test also indicates non-normality with a p-value close to zero.
* Both skewness and kurtosis values contribute to this assessment, with a skewness of 0.157 indicating a slight asymmetry in the data distribution and a kurtosis of 1.030 indicating a flatter peak compared to a normal distribution.

# Conclusion

* The negative coefficient for NCP indicates that as the number of main meals increases, the log odds of being in the reference category of NObeyesdad (likely 'Normal\_Weight' or another baseline category) decreases, suggesting a negative relationship between the number of meals and obesity level.
* The p-value is less than 0.05, which indicates that this finding is statistically significant. Therefore, we can reject the null hypothesis that there is no association between the number of main meals and the obesity level category. In other words, NCP is a statistically significant predictor of obesity level category.
* The Omnibus test has a p-value of 0.000, which suggests that the residuals of the model are not normally distributed. The Jarque-Bera test also indicates non-normality with a p-value close to zero. Both skewness and kurtosis values contribute to this assessment, with a skewness of 0.157 indicating a slight asymmetry in the data distribution and a kurtosis of 1.030 indicating a flatter peak compared to a normal distribution.
* The Durbin-Watson statistic is 0.201, which is far from the value of 2 that indicates no autocorrelation; this suggests that there may be positive serial correlation in the residuals.
* The p-value is used to determine the statistical significance of the results. A common threshold for significance is 0.05. Since the p-value in this case is 0.089, which is greater than 0.05, we do not have sufficient evidence to reject the null hypothesis at the 5% significance level. Therefore, based on this OLS regression analysis, we would conclude that we cannot affirm there is a significant effect of the number of main meals (NCP) on obesity levels. In other words, we cannot claim that NCP has a significant effect on obesity levels, and thus the null hypothesis cannot be rejected with high confidence.

# Discussion

Several restrictions and ideas for further study should be highlighted while analysing the regression analysis's findings about the association between the number of major meals (NCP) and obesity levels (NObeyesdad):

Limitations:

1. **Sample Size and Power**: The study's possible sample size and data distribution are two of its limitations. Even if there are 2111 observations in the dataset, a bigger sample size would increase the analysis's power even further. This is particularly important if there are few observations for the dataset's subgroups (such as varying degrees of obesity). In order to fit data to a normal distribution, it is desirable to have at least 30 samples. This could lead to more reliable and broadly applicable findings.
2. **Exclusion of Confounding Variables**: Age, gender, and lifestyle characteristics were not taken into account in the analysis, which was limited to the NCP variable. These factors may complicate the association between NCP and obesity and have a major impact on obesity rates. The analysis might not adequately account for the underlying causes that drive obesity levels if such variables are excluded.
3. **Model Specification**: A linear relationship between the independent and dependent variables is assumed when using OLS regression. Given the intricacy of human obesity and the multitude of interconnected factors that drive it, this assumption might not be acceptable. Furthermore, the low R-squared value suggests that additional or other variables may be required to effectively capture the relationships at play because it shows that the model explains very little of the variance in the dependent variable.

Future directions:

1. Extending the Dataset: Research in the future may gain by employing a bigger and more varied dataset that provides sufficient representation across various degrees of obesity. This would allow for a more sophisticated understanding of how various factors interact to cause obesity in addition to improving the analysis's statistical strength.
2. Adding More Variables: A more complete model might be produced by adding a wider range of variables, such as age, gender, genetic characteristics, and extra lifestyle factors. This method could assist uncover more about the intricacies of the factors that lead to obesity and help control for confounding variables

**References**

1. A review article that provides an overview of obesity and its pathological mechanisms, presenting an evidence-based analysis of the effects of different dietary strategies on weight loss, metabolic responses, and diet adherence in obesity. This could provide context for understanding how meal frequency fits into broader dietary management practices​​.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7823549/#:~:text=,and%20diet%20adherence%20in%20obesity>

1. Research focusing on the influence of portion size and the timing of meals on weight balance and obesity in adults. This includes a discussion on meal patterns and suggests that regular eating habits might facilitate weight balance, which could be directly relevant to your hypothesis about meal frequency​​.

<https://pubmed.ncbi.nlm.nih.gov/26627086/#:~:text=Obesity%20%2F%20etiology,facilitate%20weight%20balance%2C%20while%20%E2%80%A6>

1. A study on the role of taste responsiveness and food preference in obesity, which reports that obese people have a lower taste sensitivity and a higher preference for fatty and sweet foods. This could offer insights into behavioral factors that influence meal frequency and dietary choices in the context of obesity​​.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8090462/#:~:text=Evidence%20of%20the%20role%20of,sweet%20foods%20in%20obese%20people>

1. Original research that defines overweight and obesity based on BMI scores and could provide foundational criteria for categorizing obesity levels in your dataset​​.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9486272/#:~:text=According%20to%20the%20WHO%2C%20Ow,30%20%28kg%2Fm%202>

# Appendix

import pandas as pd

data = pd.read\_csv('ObesityDataSet\_raw\_and\_data\_sinthetic.csv')

data.head()

A table with numbers and letters

Description automatically generated with medium confidence

data['NObeyesdad'].value\_counts()

A white background with black and white clouds

Description automatically generated

data['SMOKE'].value\_counts()

no 2067

yes 44

Name: SMOKE, dtype: int64

data.isna().sum()

A white background with black dots

Description automatically generated

data.describe()

A table of numbers with numbers

Description automatically generated

cleanup\_nums = {"Gender":{'Male':1,"Female":2},

"family\_history\_with\_overweight":{"yes":1,"no":2},

"FAVC":{"yes":1,"no":2},

"CAEC":{"Sometimes":1,"Frequently":2,"Always":3,"no":4},

"SMOKE":{"yes":1,"no":2},

"SCC":{"yes":1,"no":2},

"CALC":{"Sometimes":1,"Frequently":2,"Always":3,"no":4},

"MTRANS":{"Public\_Transportation":1,"Automobile":2,"Walking":3,"Motorbike":4,"Bike":5},

"NObeyesdad": {'Obesity\_Type\_I':1,

'Obesity\_Type\_III':1,

'Obesity\_Type\_II':1,

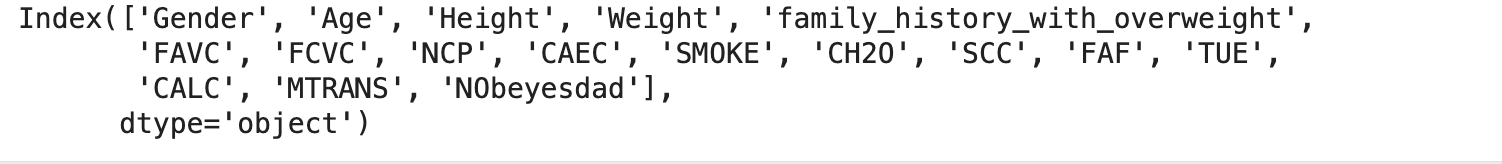
'Overweight\_Level\_I':0,

'Overweight\_Level\_II':0,

'Normal\_Weight':0,

'Insufficient\_Weight':0}}

data.columns



import matplotlib.pyplot as plt

import seaborn as sns

plt.title("family\_history\_with\_overweight")

sns.countplot(data = data,x='family\_history\_with\_overweight')

plt.show()

A blue and orange bar graph

Description automatically generated

data['FAVC'].value\_counts()

import seaborn as sns

plt.title("CAEC")

sns.countplot(data = data,x='CAEC')

plt.show()

A graph with a bar and a number of text

Description automatically generated with medium confidence

import seaborn as sns

plt.title("FAVC")

sns.countplot(data = data,x='FAVC')

A graph with a bar and a number of different colored squares

Description automatically generated with medium confidence

data['SMOKE'].value\_counts()

import seaborn as sns

sns.countplot(data = data,x='family\_history\_with\_overweight')

A graph with blue and orange bars

Description automatically generated

import seaborn as sns

plt.title("SCC")

sns.countplot(data = data,x='SCC')

A graph with blue and orange bars

Description automatically generated

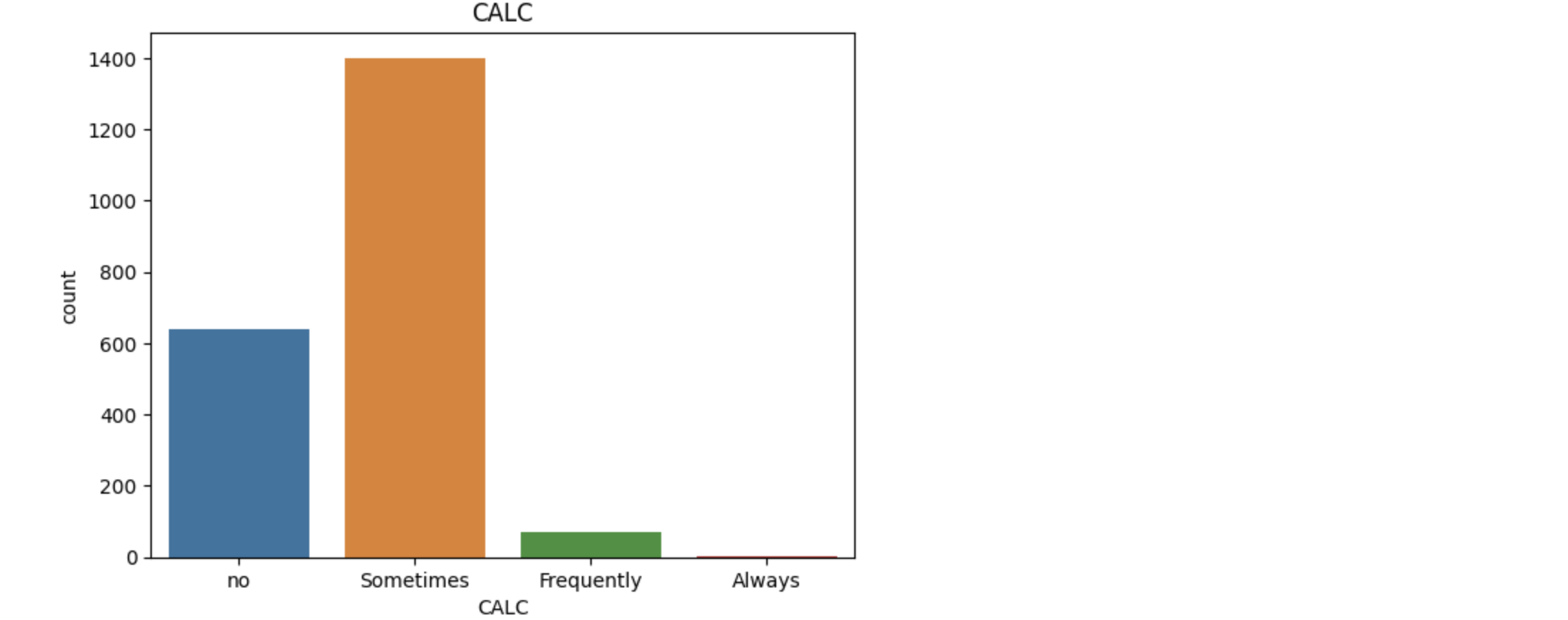
data['CALC'].value\_counts()

import seaborn as sns

plt.title("CALC")

sns.countplot(data = data,x='CALC')

plt.show()



data['MTRANS'].value\_counts()

import seaborn as sns

plt.title("MTRANS")

sns.countplot(data = data,x='MTRANS')

plt.xticks(rotation=90)

plt.show()

A graph with a green square and orange squares

Description automatically generated with medium confidence

data['Gender'].value\_counts()

import seaborn as sns

plt.title('Gender')

sns.countplot(data = data,x='Gender')

plt.show()

A blue and orange bar graph

Description automatically generated

plt.title("Age")

sns.histplot(data=data,x='Age')

plt.show()

A graph of age and age

Description automatically generated

plt.title('Height')

sns.histplot(data=data,x='Height')

plt.show()

A graph of a height

Description automatically generated with medium confidence

plt.title("Weight")

sns.histplot(data=data,x='Weight')

plt.show()

A graph of a weight

Description automatically generated with medium confidence

sns.histplot(data=data,x='Age')

A graph of age and age

Description automatically generated

plt.title("Age")

sns.boxplot(data=data, x='Age')

plt.show()

A graph with a number of dots

Description automatically generated with medium confidence

plt.title("Height")

sns.boxplot(data=data, x='Height')

plt.show()

A graph with a bar and a line

Description automatically generated with medium confidence

plt.title("Weight")

sns.boxplot(data=data, x='Weight')

plt.show()

A graph with a bar and a line

Description automatically generated with medium confidence

import matplotlib.pyplot as plt

plt.figure(figsize=(10,6))

plt.title("family\_history\_with\_overweight vs Obesity")

sns.countplot(x='NObeyesdad', hue='family\_history\_with\_overweight', data=data)

plt.xticks(rotation=90)

plt.show()

A graph with blue and orange bars

Description automatically generated

import seaborn as sns

import matplotlib.pyplot as plt

sns.scatterplot(data=data, x='Age', y='Weight')

plt.title('Scatter Plot: Age vs. Weight')

plt.show()

A graph with blue dots

Description automatically generated

import seaborn as sns

import matplotlib.pyplot as plt

sns.scatterplot(data=data, x='Height', y='Weight')

plt.title('Scatter Plot: Height vs. Weight')

plt.show()

A graph of a diagram

Description automatically generated with medium confidence

import seaborn as sns

import matplotlib.pyplot as plt

continuous\_vars = ['Age', 'Height', 'Weight']

sns.pairplot(data[continuous\_vars])

plt.suptitle('Pair Plot for Continuous Variables', y=1.02)

plt.show()

A group of blue and white graphs

Description automatically generated

data = data.replace(cleanup\_nums)

data['NObeyesdad'].value\_counts()

data

A table with numbers and letters

Description automatically generated

import statsmodels.api as sm

model = sm.Logit(endog=data['NObeyesdad'],exog=data['SMOKE']).fit()

print(model.summary())

A screenshot of a computer

Description automatically generated

data = data.replace({'family\_history\_with\_overweight':{'yes':1,'no':0}})

import statsmodels.api as sm

model = sm.Logit(endog=data['NObeyesdad'],exog=data['SMOKE']).fit()

print(model.summary())

A screenshot of a computer

Description automatically generated

object\_cols = data.drop('NObeyesdad', axis=1).select\_dtypes('object').columns

for col in object\_cols:

data[col] = le.fit\_transform(data[col])

A table with numbers and letters

Description automatically generated

X = data.drop('NObeyesdad', axis=1)

# Adding a constant to the independent variables

# Defining the dependent variable

y = data['NObeyesdad']

# Fitting the logistic regression model

model = sm.Logit(endog=y, exog=X).fit()

# Displaying the summary of the model

print(model.summary())

A screenshot of a computer

Description automatically generated

model = sm.Logit(endog=data['NObeyesdad'],exog=data['Age']).fit()

print(model.summary())

A screenshot of a computer

Description automatically generated

model = sm.Logit(endog=data['NObeyesdad'],exog=data['Weight']).fit()

print(model.summary())

A screenshot of a computer

Description automatically generated

model = sm.Logit(endog=data['NObeyesdad'],exog=data['NCP']).fit()

A screenshot of a computer

Description automatically generated

sns.regplot(data=data,x='NObeyesdad',y='NCP')

A graph with blue lines

Description automatically generated

import matplotlib.pyplot as plt

import pandas as pd

gender\_vs\_obesity = data[['Gender', 'NObeyesdad']]

# Group the data by 'Gender' and count the occurrences of each obesity level

gender\_obesity\_counts = gender\_vs\_obesity.groupby(['Gender', 'NObeyesdad']).size().unstack(fill\_value=0)

# Plotting the bar chart

ax = gender\_obesity\_counts.plot(kind='bar', figsize=(10, 6), width=0.8)

# Adding labels and title

plt.xlabel('Gender')

plt.ylabel('Count')

plt.title('Gender vs. Obesity')

# Adding a legend

plt.legend(title='Obesity Level', loc='upper right')

# Show the plot

plt.show()

A graph with blue and orange bars

Description automatically generated

X = data['NCP']

y = data['NObeyesdad']

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

print(model.summary())

A screenshot of a computer

Description automatically generated

import pandas as pd

import statsmodels.api as sm

# Load your dataset here

# data = pd.read\_csv('path\_to\_your\_dataset.csv')

# Assuming you have columns named 'NCP' and 'NObeyesdad'

# Replace these with the actual column names in your dataset

X = data['NCP'] # Independent variable (Number of main meals)

y = data['NObeyesdad'] # Dependent variable (Obesity levels)

# Add a constant (intercept) to the independent variable

X = sm.add\_constant(X)

# Create and fit the regression model

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

# Get the summary of the regression analysis

summary = model.summary()

# Extract the p-value for the coefficient of NCP

p\_value\_ncp = model.pvalues['NCP']

# Define your significance level (alpha)

alpha = 0.05

# Check if the p-value is less than alpha to determine significance

if p\_value\_ncp < alpha:

print("Null Hypothesis (H0) is rejected: There is a significant difference in obesity levels based on the number of main meals consumed.")

else:

print("Null Hypothesis (H0) is not rejected: There is no significant difference in obesity levels based on the number of main meals consumed.")

# Print the summary of the regression analysis

print(summary)

A screenshot of a computer

Description automatically generated