Scientific Data Analysis

Team 9: Rafael Muijsert , Shania Erica Barreto, Yuqiu Huang, Tommaso Bertonasco

The Dominance of Hereditary over Lifestyle Factors

Table of contents

01 **Introduction**

Project overview, research questions

02 **Methodology**

Sources, preparation, techniques used

03 **Analysis**

Data prep, Bivariate exploration, Multivariate modeling

04 **Results**

Key findings, model comparisons

05 **Conclusion**

Summarizing, implications, recommendations

01

Introduction

Is obesity more strongly predicted by hereditary or lifestyle factors?

Hypothesis

Null Hypothesis

There is no significant difference in the predictive power of hereditary and lifestyle factors in obesity outcomes

Alternative Hypothesis

There is a significant difference in the predictive power of hereditary factors and lifestyle factors in obesity outcomes

Dataset

Description

2111 records and 17 variables ranging from gender, to hereditary factors and lifestyle factors (e.g. daily water intake)

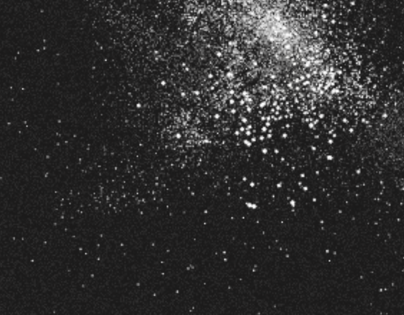
Survey

23% of the data was collected from 2111 participants from Peru, Mexico and Colombia using an online platform

Synthetic Data

77% of the data was generated synthetically utilising the weka tool and Synthetic Minority Oversampling TEchnique (SMOTE)

Methodology



Obesity Analysis: From Data to Insight

The study cleans the data, explores simple obesity relationships, and models hereditary versus lifestyle impacts to see which influences obesity more.

Stages

01

Data Preparation

Clean, encode, categorize, create outcomes

02

Bivariate Exploration

Test relationships, compare groups

03

Multivariate Modeling

Estimate effects, compare factor influence

Methodology

01 → 02 → 03 → 04

01 Remove Duplicates

Identify and delete repeated records to ensure each observation is unique and data integrity is preserved.

02 Measurable

Round variables like age, height, weight, and activity to reduce artificial precision and improve interpretability.

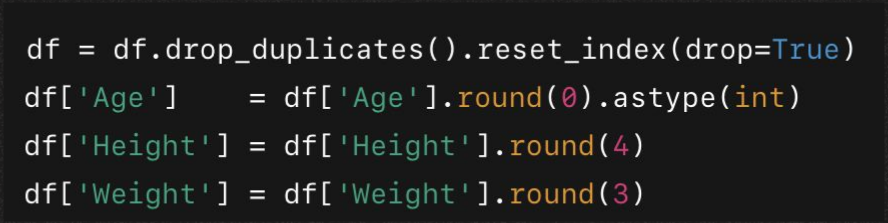
03 Data collection

Convert binary, ordinal, and nominal variables into numeric or one-hot formats suitable for modeling.

04 Specific sampling

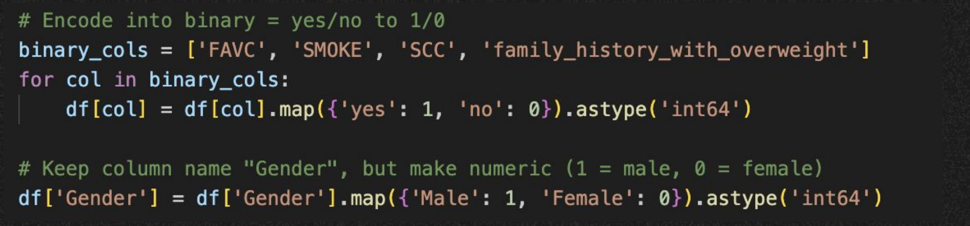
Construct ordered obesity levels and binary obese/non-obese variables to serve as clear analysis targets.

Code Samples



```
df = df.drop_duplicates().reset_index(drop=True)
df['Age']    = df['Age'].round(0).astype(int)
df['Height'] = df['Height'].round(4)
df['Weight'] = df['Weight'].round(3)
```

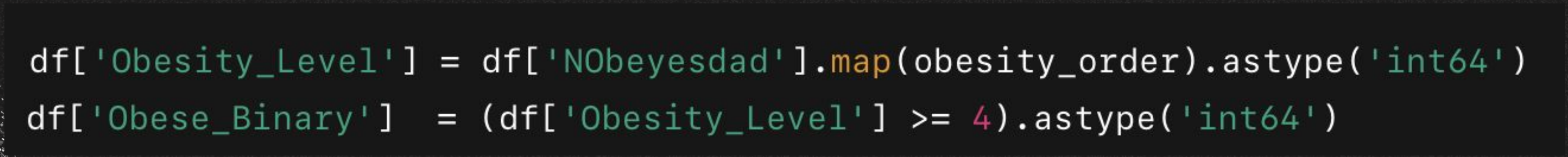
Ensures unique records and reduces unnecessary precision for better interpretability.



```
# Encode into binary = yes/no to 1/0
binary_cols = ['FAVC', 'SMOKE', 'SCC', 'family_history_with_overweight']
for col in binary_cols:
    df[col] = df[col].map({'yes': 1, 'no': 0}).astype('int64')

# Keep column name "Gender", but make numeric (1 = male, 0 = female)
df['Gender'] = df['Gender'].map({'Male': 1, 'Female': 0}).astype('int64')
```

Converts categorical responses into numeric formats suitable for analysis and modeling



```
df['Obesity_Level'] = df['NObeyesdad'].map(obesity_order).astype('int64')
df['Obese_Binary']  = (df['Obesity_Level'] >= 4).astype('int64')
```

Creates ordinal and binary obesity targets, encoding severity levels for regression models and simplifying classification for statistical and predictive analyses

Bivariate Analysis: Lifestyle Factors vs Obesity

We analyzed how key lifestyle behaviors relate to obesity status using appropriate statistical tests. Continuous and binary variables were examined separately to identify significant associations. This analysis provides an initial understanding of which lifestyle factors may influence obesity.

01

Variables Tested

Continuous variables include water intake, physical activity, and screen time; the binary variable is calorie monitoring.

02

Statistical Methods

Mann–Whitney U tests were applied for continuous variables, while Chi-Square tests assessed the binary variable

03

Objective

Identify lifestyle factors showing significant differences between obese and non-obese participants for further modeling.

Binary Variable: Chi-Square Test

Summary Results

- SCC (Calorie Monitoring): $\text{Chi}^2 = 74.527$, df = 1, p = 5.983e-18 (***), Min Expected = 44.71.

Calorie monitoring behavior differs strongly between obese and non-obese participants

This highly significant result indicates that individuals who monitor their calories are less likely to be obese, suggesting that calorie tracking is strongly associated with obesity prevention.

Mann–Whitney U Test

Summary Results

- FAF (Physical Activity): Non-Obese mean = 1.133, Obese mean = 0.875; Median = 1.0 vs 0.903; U = 623553, p = 2.303e-09 (***).
- CH2O (Water Intake): Non-Obese mean = 1.946, Obese mean = 2.073; Median = 2.0 vs 2.08; U = 462007, p = 4.923e-09 (***).
- TUE (Screen Time): Non-Obese mean = 0.715, Obese mean = 0.603; Median = 0.8 vs 0.553; U = 587964.5, p = 7.004e-04 (***).

All continuous lifestyle variables show statistically significant differences between obese and non-obese groups

Assessing Hereditary and Lifestyle Impacts on Obesity

The primary goal of this code is to answer a fundamental health question: Does our behavior matter more than our genetics?

From Raw Data to Ordered Insights

- **Standardization**: Uses StandardScaler to normalize variables, allowing a direct "apples-to-apples" comparison of effects across different units (e.g., liters vs. hours).
- **Ordinal Modeling**: Employs OrderedModel to respect the natural ranking of obesity levels (Underweight → Normal → Obese), which is more precise than standard classification.
- **Validation**: Calculates Odds Ratios to measure risk magnitude and uses Wald Tests to verify if lifestyle or genetic factors are statistically significant.

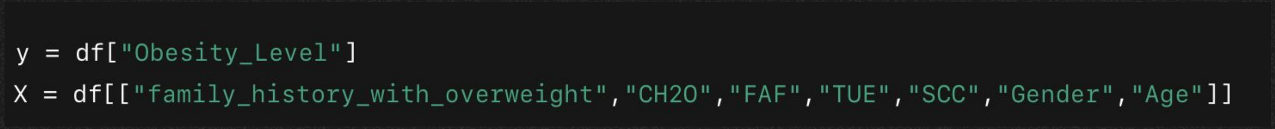
Assessing Hereditary and Lifestyle Impacts on Obesity

We used ordinal logistic regression to measure how genes versus lifestyle habits impact obesity progression. This model accounts for age and gender to isolate the exact probability of moving into higher-risk weight categories.

Key Predictors:

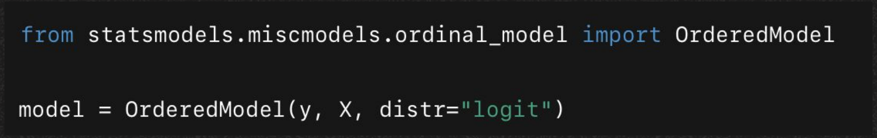
- Hereditary: Family history of overweight
- Lifestyle: Water intake (CH2O), Physical activity (FAF), Screen time (TUE), Calorie monitoring (SCC)
- Demographics: Age, Gender

Code Samples



```
y = df["Obesity_Level"]
X = df[["family_history_with_overweight","CH2O","FAF","TUE","SCC","Gender","Age"]]
```

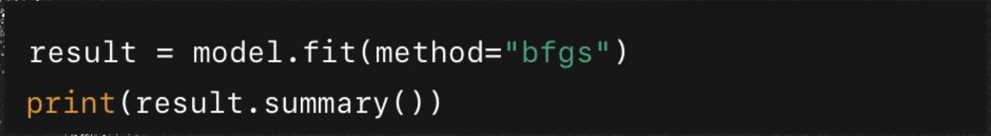
Specifies the dependent variable (ordered obesity levels) and independent predictors for the model.



```
from statsmodels.miscmodels.ordinal_model import OrderedModel

model = OrderedModel(y, X, distr="logit")
```

Initializes the ordinal logistic regression model using a logistic link for proportional odds.



```
result = model.fit(method="bfgs")
print(result.summary())
```

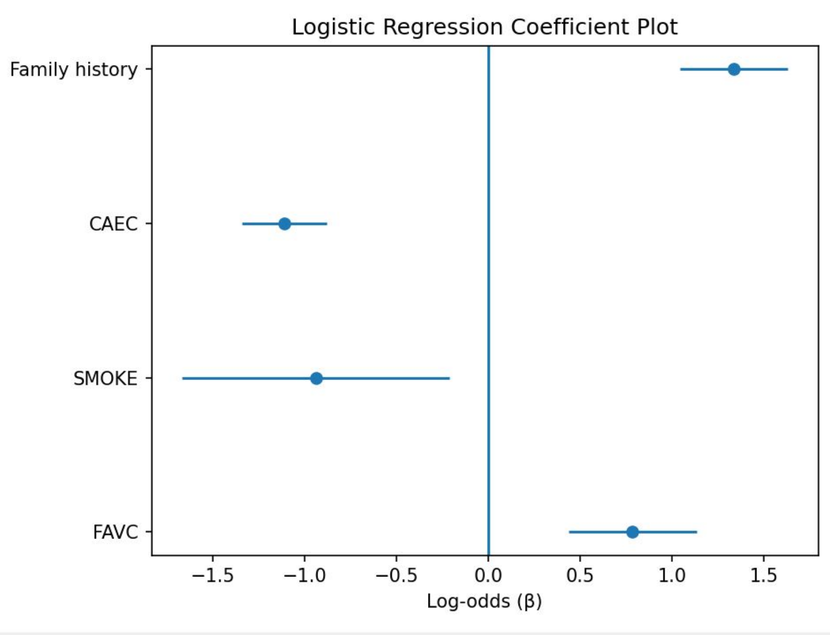
Fits the model using the BFGS optimization algorithm and outputs coefficients, standard errors, and significance

Results

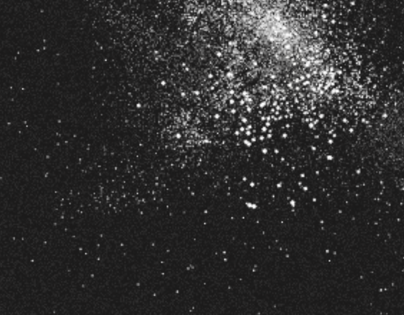
Main results:

- Family history of overweight shows a strong positive association with obesity
- Lifestyle variables (e.g. FAVC,CAEC) are significant predictors, but with comparing smaller effect sizes to genetic factors
- Smoking (SMOKE) shows a negative association with obesity. (p=0.012) This is not up to the common expectations about healthy behaviors.
- ▲ CAEC is also negatively associated with obesity—-> good way for modulating obesity risk.

- The model is **highly significant** (LLR $p < 0.001$)
- **Dominant predictor: Family history** shows the **largest positive effect** (β ≈ 1.34)---> higher odds of obesity when family history is present.
- **Negative associations: CAEC** (β ≈ −1.11) and **SMOKE** (β ≈ −0.94)
- **Positive: FAVC** shows a **positive association** (β ≈ 0.78)---> As this behavior increases, the likelihood of obesity also increases.
- All predictors are significant (all $p < 0.05$)---> **95% CIs not crossing 0** in the coefficient plot.



Conclusions



Key conclusions

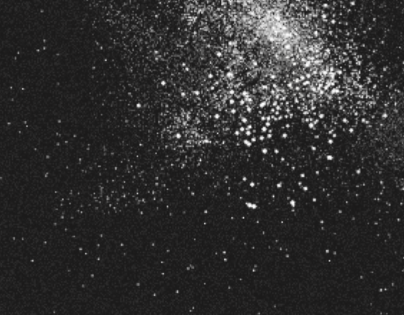
- Lifestyle as a whole is a stronger predictor of obesity than genetics alone
- However, family history was the most significant *single* predictive variable
- Genetics and lifestyle choices interact, creating a compounding effect on obesity risk

The Power of Predictors

- The single most powerful predictor was family history
- Incorporating all lifestyle factors (AIC: 7502) was significantly more accurate than a model containing only family history (AIC: 7614).
- Conclusion: genetic predisposition is highly significant but the cumulative effect of daily habits and environment is the dominant factor in determining obesity.

Implications & recommendations

- Primary Focus: Continue to prioritize broad public health interventions that target lifestyle and behavioral change.
- Secondary Focus: Develop targeted programs that provide personalized advice to individuals with a genetic risk.
- Future Work: Further research is needed to better understand the specific gene-lifestyle interactions to refine these personalized strategies.



Questions

Thank *you*

Do you have any questions?

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**