Exploratory Data Analysis and Hypothesis Testing on Heart Disease Factors

Surenther Selvaraj

Bellevue University

DSC530: Data Exploration and Analysis

Professor: Matthew Metzger

Date: Feb 28, 2025

Exploratory Data Analysis and Hypothesis Testing on Heart Disease Factors

Statistical/Hypothetical Question

The primary hypothesis explored in this analysis was: Do Age, Blood Pressure, Cholesterol Level, BMI, and Triglyceride Level significantly impact the likelihood of developing heart disease? This hypothesis was examined using exploratory data analysis (EDA) techniques such as histograms, probability mass functions (PMF), cumulative distribution functions (CDF), analytical distribution fitting, and scatter plots. Additionally, logistic regression analysis was conducted to assess the statistical relationship between the explanatory variables and the binary outcome of heart disease.

Outcome of EDA

The EDA revealed several insights into the dataset. Histograms and normal distributions suggested that Age, Blood Pressure, Cholesterol Level, BMI, and Triglyceride Level followed different distributions, with some showing skewness. Scatter plots indicated weak correlations between some variables, suggesting that no single variable had an overwhelming influence on heart disease. The logistic regression model revealed that BMI had a borderline significant effect on heart disease (p = 0.056), while other variables did not show strong statistical significance. The model's pseudo R-squared value was quite low, indicating that the selected variables did not fully explain the variance in heart disease occurrence.

Missed Aspects in Analysis

One key limitation was the potential exclusion of other influential variables such as genetic predisposition, lifestyle factors (smoking, exercise, and diet), and medical history. Additionally, potential interactions between the selected variables were not explored in depth. For example, the combined effect of Age and BMI might have provided additional insight into heart disease risk.

Additional Variables That Could Have Helped

Variables such as smoking history, alcohol consumption, stress levels, and physical activity levels could have significantly improved the model's explanatory power. Furthermore, socio-economic factors such as income level and education may have influenced lifestyle choices, indirectly affecting heart disease risk. Including categorical variables related to family history of heart disease could have also provided deeper insights.

Incorrect Assumptions

One possible incorrect assumption was treating all numerical variables as normally distributed in analytical distributions. Some variables exhibited skewness, which may have impacted statistical tests. Additionally, the assumption that all relevant predictors were included in the dataset was likely incorrect, leading to a potential omitted variable bias.

Challenges Faced and Areas of Uncertainty

One of the main challenges was handling categorical variables, particularly the conversion of “Yes/No” values into binary form for analysis. Initially, this led to NaN values in statistical tests due to incorrect formatting. Understanding and interpreting logistic regression outputs, such as odds ratios and pseudo R-squared values, required further study. Another challenge was determining the appropriate threshold for statistical significance, as some variables (e.g., BMI) were close to the conventional 0.05 significance level.

Conclusion

In conclusion, while the analysis provided valuable insights into the relationship between the selected variables and heart disease, the study’s limitations highlight the need for further research with additional explanatory variables. Future studies could incorporate machine learning techniques for improved prediction accuracy.