Exploring Risk Factors for Heart Disease: A Comprehensive Statistical Analysis

Elishel Sason, Julian Costa, Connor Horan

Introduction

Heart disease stands as one of the leading causes of mortality worldwide, impacting the lives of millions and placing extreme burden on healthcare systems. Understanding the various risk factors related to heart disease can help create more efficient prevention strategies and improve overall health outcomes. In this project, we seek to identify and study the significant predictors of heart disease in hopes to improve intervention strategies and encourage lifestyle changes that may aid at-risk populations.

Description of Data

This project utilizes the **Heart Disease Health Indicators** dataset from Kaggle, which contains health-related information on over 253,000 individuals. The data was originally collected from the Behavioral Risk Factor Surveillance System (BRFSS), a health-related telephone survey that is collected annually by the CDC.

The dataset contains 253680 observations of 22 variables.

Analysis

t-Test: Assessed differences in continuous variables (e.g., BMI, Mental Health) between heart disease groups to identify key predictors.

ANOVA: Compared group means for multiple levels of categorical variables (e.g., Age, General Health) to detect significant differences.

Chi-Squared: Evaluated associations between categorical variables (e.g., Smoking, HighBP) and heart disease status, highlighting relationships.

Logistic Regression: Modeled the combined effects of predictors and interactions on heart disease risk, utilizing odds ratios for interpretation.

Model Building: Starting with a base model that included all predictors, a stepwise selection process was ran to help select the best model, using AIC to identify the most significant factors.

Diagnostics

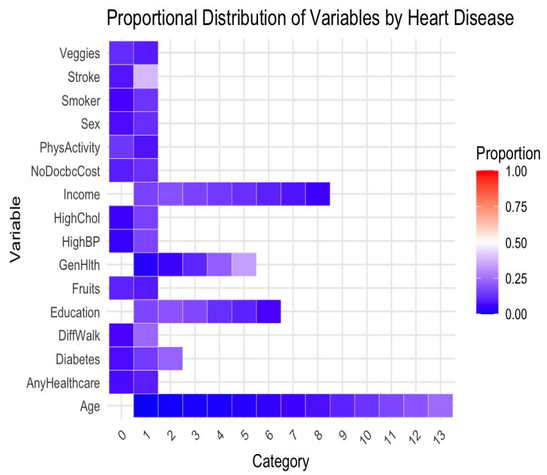
Key diagnostics included:

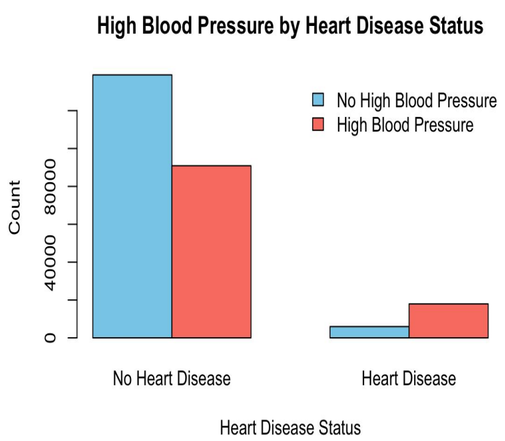
- **Multicollinearity**: Assessed using Variance Inflation Factors (VIF), confirming no severe multicollinearity issues.
- **Influential Observations**: Examined using Cook's Distance and leverage plots, with no observations requiring exclusion.
- **Linearity of Logit**: Verified using log-transformed continuous variables where applicable.

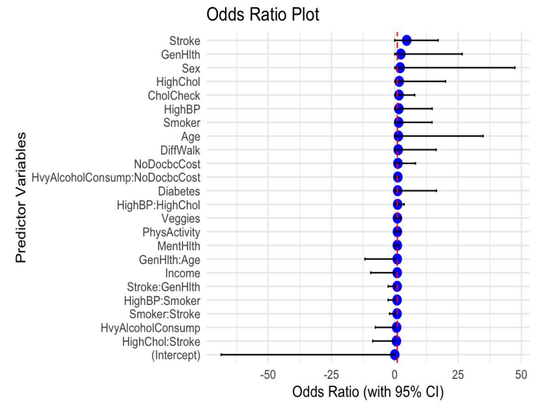
Hypothesis and Research Question

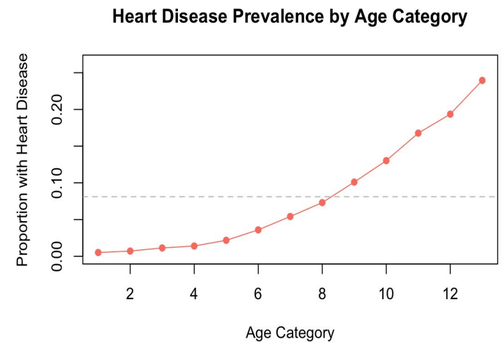
Research Question: What are the significant risk factors associated with heart disease?

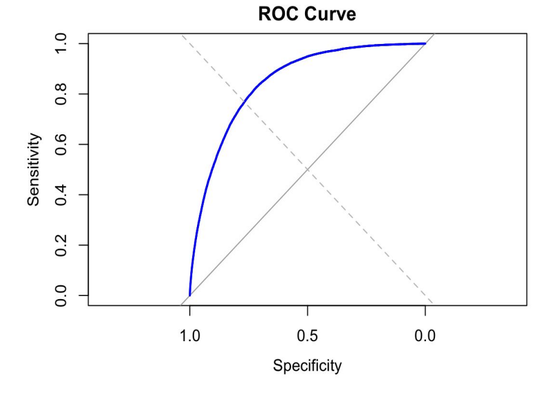
Hypothesis: Higher Age, Smoking, and High Blood Pressure are significantly associated with a greater likelihood of heart disease.

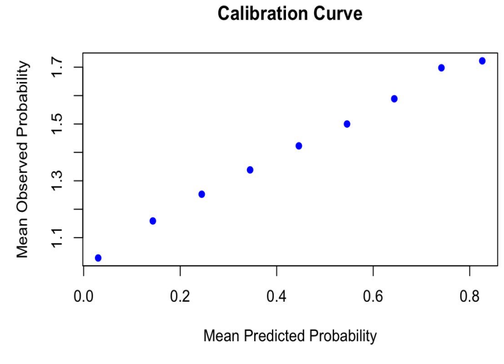












Conditions/Assumptions

Binary Outcome Variable: The dependent variable is HeartDiseaseorAttack, which has values: 0 = No, 1 = Yes. This variable is binary and confirms this assumption.
Independence of Observations: In our dataset, each observation represents a unique individual, and there are no repeated measures or clustering effects. Therefore, the independence assumption is confirmed.
Linearity of Logit: The age variable is categorized into groups, making it more complex to analyze in logistic regression. Our attempts to test its linearity caused technical issues with the model, so to keep the analysis straightforward, we left the variable as is. Accurate assessment of this assumption may require a more thorough refining of the age variable, be it through transformations or other means.
No Multicollinearity: Checking the model's Variation Inflation Factor (VIF) values, there seems to be no concerning multicollinearity issues present in our final model.
Large Sample Size: Given that our dataset contains over 250,000 observations, the assumption of large sample size is met.

Results

Main Effects:

Variables like high blood pressure, cholesterol, smoking, and stroke remain strong independent predictors of heart disease.
Lifestyle factors (e.g., physical activity, vegetable consumption) and socioeconomic factors (e.g., income, access to healthcare) also play key roles.

Interactions:

The interaction terms show how combinations of predictors influence heart disease differently compared to their independent effects. For example, the interaction between age and general health highlights how age can amplify the impact of poor health.

Variables Dropped in Interaction:

MentHlth (p=0.115): Number of bad mental health days is no longer significant in the interaction model.
HvyAlcoholConsump (p=0.113): This interaction term did not show a significant effect on heart disease risk.

Conclusion

The analysis sought to identify significant predictors of heart disease. Our hypothesis was supported: age, smoking, high blood pressure, and stroke are strongly associated with increased odds of heart disease. Factors like cholesterol, general health, and socioeconomic indicators like income also play critical roles. Further potential research could expand this investigation by exploring additional hypotheses such as if physical activity is inversely associated with heart disease, or if individuals with a history of stroke or diabetes are increased in risk for heart disease.

References

"CDC Diabetes Health Indicators." UCI Machine Learning Repository, 2015, archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators.

Teboul, Alex. "Heart Disease Health Indicators Dataset" Kaggle, 10 March. 2022, https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset/data