Linear Statistical Analysis Project Proposal

NAME: SANNA TOURAY

PANTHER#: 002800278

DATE: October 2023

**Research Proposal**

Predicting Heart Disease Using Logistic Regression.

# Introduction

Heart disease is a major health concern globally, causing a significant number of deaths each year. Early detection and accurate prediction of heart disease risk can play a crucial role in preventing fatalities. This project aims to develop a predictive model using logistic regression, a fundamental statistical technique, to identify individuals at risk of heart disease on various clinical and lifestyle factors. We will use a heart disease dataset obtained from Kaggle, which includes patient information and heart disease outcomes.

# Research Objectives

The primary objectives of this study are as follows:

1. To build a predictive model for heart disease using logistic regression analysis.
2. To evaluate the model’s
3. accuracy, sensitivity, and specificity in predicting heart disease risk.
4. To create confidence intervals for the coefficient estimates and standard errors.
5. To identify the most Signiant factors contributing to heart disease

# Assumptions

We will ensure that the assumptions of logistic regression are met:

**1**.Binary Outcome: Logistic regression is designed for binary outcomes, meaning the dependent variable (in this case, the presence or absence of heart disease) is categorical and has only two possible outcomes.

**2**.Linearity of Log-Odds: Logistic regression assumes that the relationship between the independent variables and the log-odds of the dependent variable is linear, However, this linearity applies to the log-odds, not the raw probabilities.

**3**.Model Fitting: We will use a stepwise method for variable selection, assess goodness-of-fit using relevant statistical tests, and inspect deviance residuals and partial residuals. Independence of Error Terms.

**4.**Independence of errors: The observations used in logistics regression should be independent of each other. Each case should be independent of all other cases. In the context of heart disease prediction, this means that the presence or absence of heart disease in one individual should not affect the presence or absence of heart disease in another individual.

**5**.Large Sample Size: Logistic regression assumes a sufficiently large sample size for accurate estimation coefficients and valid statistical inference. A larger sample size is generally preferred to ensure stable estimates and reliable predictions. i.e n=303

# Conclusion

By successfully implementing logistic regression on a linear statistical model, this project aims to contribute valuable insights into predicting heart disease risks, potentially aiding in early intervention and prevention strategies.

Initially, I will present the data in a two-way table format for enhanced clarity. Subsequently, I will utilize R to develop a logistic regression model, interpreting the coefficients to understand their implications on predicted survival odds. Following this, I will generate confidence intervals for both the coefficient estimates and their standard errors. If there is ample space within the project, I intend to calculate the proportion of inaccurately predicted outcomes and showcase specific predicted values. Additionally, I plan to create a receiver operating characteristic curve, offering a graphical representation of the model's predictive performance.

Top of Form

2