Problem Statement

Objective:

The goal of this assignment is to conduct an analysis using logistic regression to understand how different factors contribute to the risk of heart disease. This study aims to explore the relationship between various predictors and the likelihood of heart disease in a specific population.

The assignment should be done using either R or Python or SAS or Stata or SPSS programming languages/software packages.

Include in your Word document the question, its result, and a clear, precise interpretation. Marks will be deducted for inadequate responses. Ensure that screenshots are clear, and interpretations are well-articulated.

Dataset Description:

The dataset for this assignment is derived from the Coronary Risk-Factor Study (CORIS) baseline survey, conducted in 1983 in three rural areas of the Western Cape, South Africa. The focus of the study was to investigate the intensity of ischemic heart disease risk factors among a specific population. The dataset features a retrospective sample of white male individuals, aged between 15 and 64, from a high-risk region for heart disease in South Africa.

Response Variable: Presence(1) or absence(0) of coronary heart disease (CHD).

Data Cleaning and Manipulation

1. Explore the variables included in the dataset.
2. Load the data into your environment and perform any necessary cleaning steps and any data preprocessing steps needed for your analysis.
3. Conduct a descriptive analysis of the key variables. Note any major observations.
4. Use visualization techniques to explore the relationships between various predictors (such as age, ldl, blood pressure) and the binary outcome of heart disease.
5. Identify and discuss any observable trends or patterns that could influence your logistic regression analysis.

Formulate Questions

6. Based on your exploratory data analysis, formulate two statistical inference questions that logistic regression can address. Explain why these questions are relevant and how they relate to the dataset.
7. For each question, identify the predictor variable(s) and the binary outcome variable.

Analysis

8. For each question, perform logistic regression analysis to assess the relationship between the predictor(s) and the outcome.
9. Interpret the logistic regression coefficients and discuss their implications in the context of heart disease risk.
10. Evaluate the model's fit and discuss any limitations or assumptions in your analysis.