Individual Assignment 2: Logistic Regression Analysis

Objective

Analyze the **Heart Disease dataset** to predict the presence of heart disease using logistic regression. The dataset contains medical information about individuals, and the goal is to explore how various factors contribute to the likelihood of heart disease.

Instructions

Follow the steps outlined below to conduct a logistic regression analysis. Answer all questions and include R code and explanations for your results.

1. Data Preparation

a. Download and Load the Dataset

Download the Heart Disease dataset from Kaggle: `https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset`. Import the dataset into R and inspect the first few rows.

- Load the dataset and provide the R code you used.
- **Question:** What are the variables in the dataset? Provide a brief description of each.

b. Select Variables for Analysis

Use the following variables:

- **Target (Presence of Heart Disease):** `target` (1 = Presence of heart disease, 0 = No heart disease)
- **Age (Age of the person):** `age`
- **Sex (Gender):** `sex` (1 = Male, 0 = Female)
- **Chest Pain Type:** `cp` (Categorical: 1, 2, 3, 4)
- **Cholesterol Level:** `chol`

2. Data Cleaning

- Remove rows with missing values, if there are any such.
- Convert categorical variables (`cp` and `sex`) into factors.
- Convert the target variable (`target`) into a factor.
- Provide the R code you used to clean the dataset.
- **Question:** How many rows and columns are in the cleaned dataset?

3. Exploratory Data Analysis

- Summarize the variables using the `summary()` function.
- Create boxplots of `chol` across levels of `target`.
- Create bar plots for `sex` and `cp` with respect to `target`.
- **Question:** What patterns do you observe in the data? For example, does cholesterol seem to differ by heart disease status? Are there differences in heart disease prevalence by sex or chest pain type?

4. Fitting a Logistic Regression Model

- Fit a logistic regression model to predict `target` using `age`, `sex`, `cp`, and `chol` as predictors.
- **Question:** Interpret the coefficients of the model. Which variables significantly affect the likelihood of heart disease? Explain the direction of their effects.

5. Model Interpretation

- Convert the coefficients into odds ratios using `exp(coef())`.
- **Question:** Provide interpretations for the odds ratios of `sex`, `cp`, and `chol`. For example, how does an increase in cholesterol affect the odds of heart disease?

6. Model Comparison

- Fit a reduced model excluding the `chol` variable.
- Perform a likelihood ratio test between the full and reduced models using `anova()`.
- **Question:** Does including `chol` significantly improve the model? Explain your answer based on the p-value.

7. Model Predictions and Performance

- Predict probabilities of heart disease for all individuals.
- Convert these probabilities into binary predictions using a threshold of 0.5.
- Create a confusion matrix to evaluate the model's performance.
- **Question:** What is the accuracy of the model? How many false positives and false negatives are there?

8. ROC Curve and AUC

- Generate an ROC curve and calculate the AUC for the model using the `pROC` package.
- **Question:** What does the AUC tell you about the model's performance? Is the model good at distinguishing between individuals with and without heart disease?

Submission Guidelines

- Submit a well-organized R Markdown file or PDF report containing:
  - All R code used for the analysis.
  - Answers to each question, including visualizations and interpretations.
- Ensure your code is commented and outputs are clear.