IIIT SRI CITY
Institute Of National Importance

Regression Analysis for Heart Attack Prediction

Project report by:
Himanshu Kataria, Shubham Shandilya, Mayank Raj Gupta, Surya Naidu, Jagan Ram Jaswanth

This report is submitted to IIIT Sri City in fulfillment of the requirements for the Introduction to Data Analytics Project

November 14, 2023

1 Introduction

Heart disease remains a leading cause of mortality worldwide, emphasizing the critical need for effective predictive tools to identify individuals at a higher risk of cardiac events. In the pursuit of enhancing preventative measures and personalized healthcare, this project focuses on employing regression analysis to establish relationships between various risk factors, or regressors, and the likelihood of heart attacks. The importance of predicting heart attacks lies not only in the potential to save lives but also in optimizing resource allocation within healthcare systems and promoting proactive patient care.

Regression analysis serves as a powerful statistical method for discerning patterns and dependencies within datasets, making it an invaluable tool in the context of heart attack prediction. By investigating the relationship between specific attributes—such as resting blood pressure, cholesterol levels, chest pain type, and resting electrocardiographic results—we aim to unveil insights that contribute to a more nuanced understanding of cardiovascular risk factors.

Referencing heart attack prediction provides a solid foundation for this project, drawing upon existing research and datasets to inform our analysis. Leveraging the insights gained from regression analysis, we aspire to contribute to the ongoing efforts in refining risk assessment models and ultimately improving the accuracy of heart attack predictions. Through this exploration, we anticipate shedding light on the intricate interplay between various health indicators and the likelihood of a cardiac event, offering potential advancements in preventive healthcare strategies.

2 Theory

2.1 Regression Analysis

Regression analysis is a statistical technique used to model the relationship between a dependent variable (response variable) and one or more independent variables (regressors or predictors). It helps in understanding and quantifying the nature of the relationship between variables, making it a valuable tool in various fields, including medical research, economics, and social sciences.

2.1.1 Fundamental Concepts of Regression Analysis:

1. **Dependent Variable (Response Variable):** The variable that we want to predict or explain is called the dependent variable. In the context of heart attack prediction, this could be the likelihood or risk of experiencing a heart attack.

2. **Independent Variables (Regressor Variables):** These are the variables that are used to predict or explain the variation in the dependent variable. For instance, in the heart attack prediction scenario, independent variables might include resting blood pressure, cholesterol levels, chest pain type, and resting electrocardiographic results.

3. **Simple Linear Regression:** Simple linear regression is a type of regression analysis that involves only one independent variable. The relationship between the dependent variable (Y)

and the independent variable (X) is represented by a straight line equation: $Y = \beta_0 + \beta_1 X + \varepsilon$, where β_0 is the y-intercept, β_1 is the slope, and ε is the error term.

2.2 Attributes

The selected attributes for heart attack prediction provide crucial insights into various aspects of cardiovascular health:

- **Resting Blood Pressure (`trtbps`):** Resting blood pressure measures the force of blood against arterial walls when the body is at rest. Elevated blood pressure is a key indicator of hypertension, a significant risk factor for heart attacks.
- **Cholesterol (`chol`):** Cholesterol is a fatty substance in the blood. Abnormal levels, particularly high LDL ("bad" cholesterol) and low HDL ("good" cholesterol), are associated with an increased risk of atherosclerosis and heart attacks.
- **Chest Pain Type (`cp`):** Chest pain type characterizes the nature of chest discomfort. Specific patterns of chest pain, such as those associated with angina, can indicate compromised blood flow to the heart, aiding in risk assessment.
- **Resting Electrocardiographic Results (`restecg`):** Resting electrocardiography measures the heart's electrical activity at rest. Abnormalities in the electrocardiographic results, such as ST-segment depression, can indicate myocardial ischemia or other cardiac issues, contributing to risk assessment.

These attributes collectively contribute to a comprehensive understanding of cardiovascular risk, enabling informed predictions and interventions for heart attack prevention.

3 Data Pre-processing

3.1 Loading and Inspection

The initial step involved loading the dataset and inspecting the first few rows to gain a preliminary understanding of the data structure. The dataset, named `heart.csv`, was loaded into the R environment using the `read.csv()` function. A quick examination of the first few rows was performed using the `head()` function.

3.2 Handling Missing Values

Subsequent to the initial inspection, the dataset was checked for missing values. The number of rows with missing values was determined and reported using the `is.na()` function in combination with `sum()`. Rows with missing values were then removed from the dataset using the `na.omit()` function, ensuring a clean dataset for further analysis.

3.3 Feature Selection

The analysis focused on four key attributes: resting blood pressure (`trtbps`), cholesterol levels (`chol`), chest pain type (`cp`), and resting electrocardiographic results (`restecg`). These fields, along with the target variable (`output`), were selected to create a new dataset for in-depth analysis.

3.4 Exploratory Data Analysis

To gain deeper insights into the data, a summary statistics analysis was conducted using the `summary()` function. This provided measures of central tendency and dispersion for each selected attribute, aiding in the understanding of the dataset's distribution.

Additionally, a boxplot was generated to visually inspect the distribution of cholesterol levels (`chol`) concerning heart attack status (`output`). The boxplot provides a graphical representation of the central tendency, spread, and potential outliers in cholesterol levels for individuals with and without a heart attack.

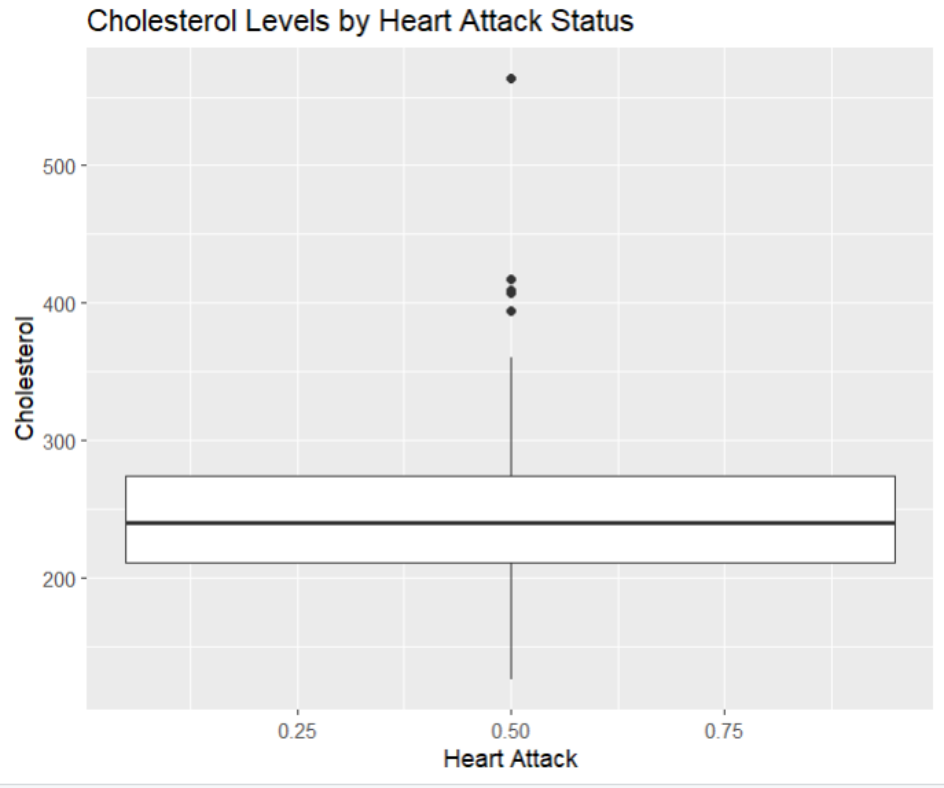

Figure 1: Cholesterol Levels by Heart Attack Status

The visualization enhances our understanding of how cholesterol levels vary between individuals

who experienced a heart attack and those who did not.

The combined data pre-processing steps ensure that the dataset is well-prepared for subsequent regression analysis. Further steps such as normalization or additional exploratory analyses can be incorporated based on the specific requirements of the analysis and chosen regression model.

4 Data Exploration

To gain insights into the dataset, we conducted exploratory data analysis, focusing on visualizing the distribution of selected attributes and examining the correlation between them.

4.1 Distribution of Selected Attributes

We visualized the distribution of key attributes to understand their frequency and variation.

4.1.1 Resting Blood Pressure

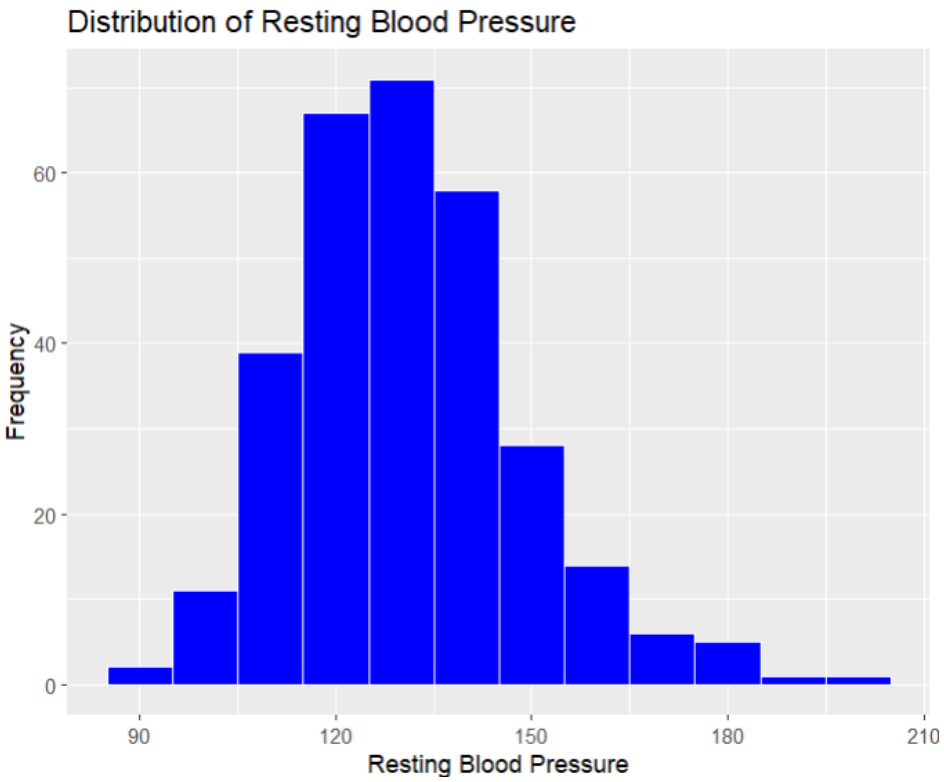

Figure 2: Distribution of Resting Blood Pressure

The histogram in Figure 2 illustrates the distribution of resting blood pressure in the dataset.

4.1.2 Cholesterol

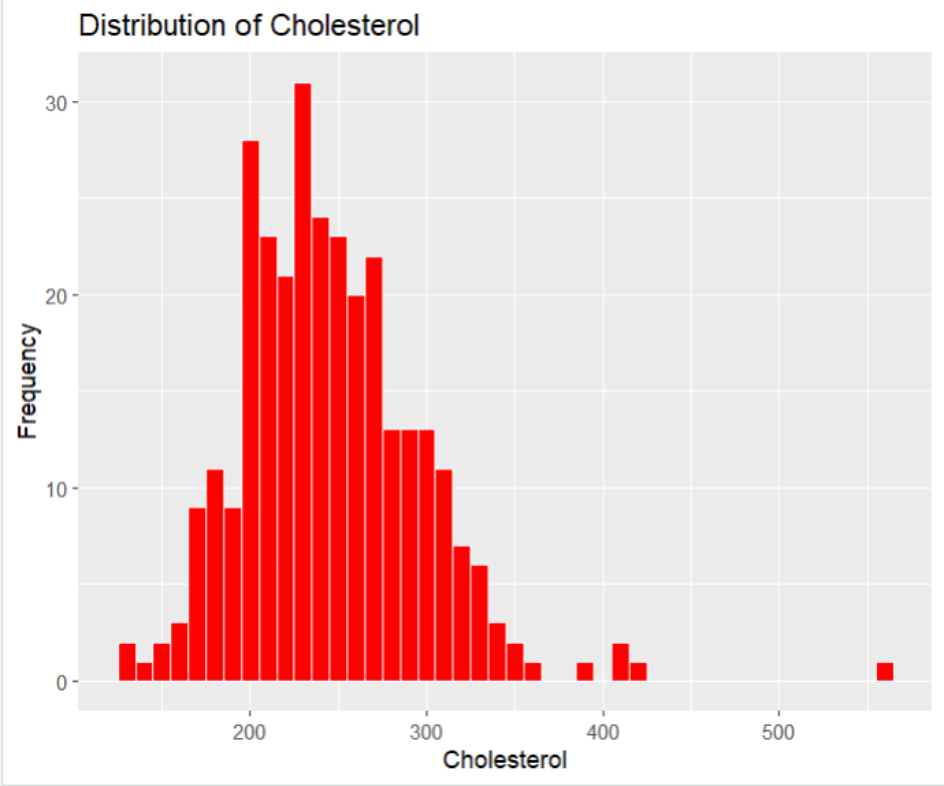

Figure 3: Distribution of Cholesterol

Figure 3 displays the distribution of cholesterol levels.

4.1.3 Chest Pain Type

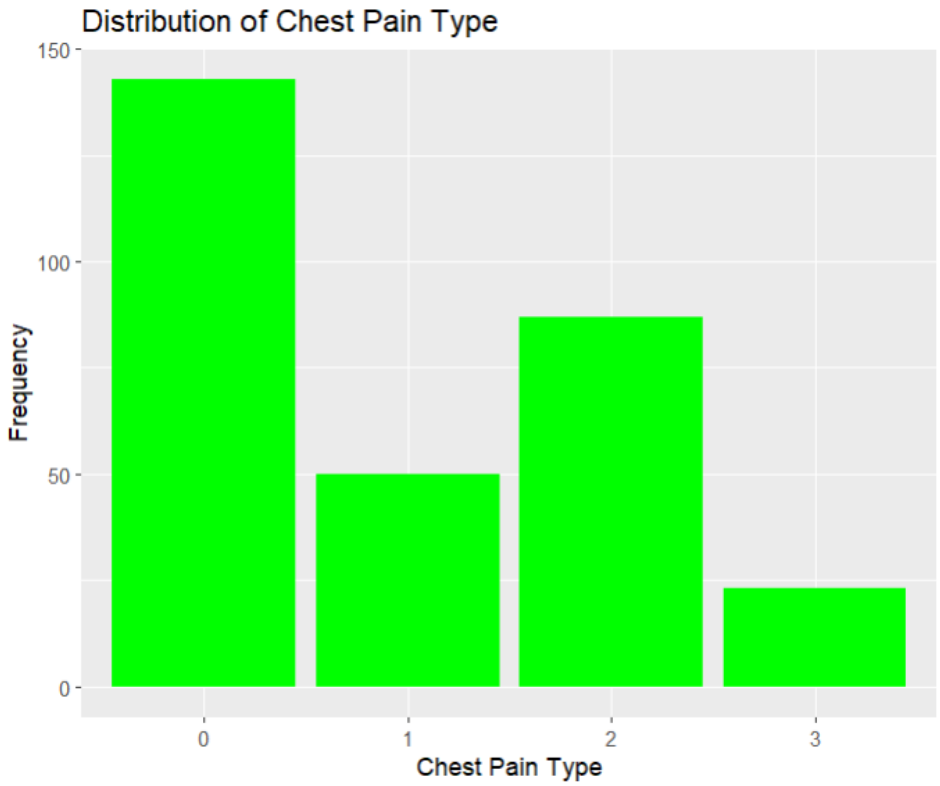

Figure 4: Distribution of Chest Pain Type

The bar chart in Figure 4 presents the distribution of chest pain types.

4.1.4 Resting Electrocardiographic Results

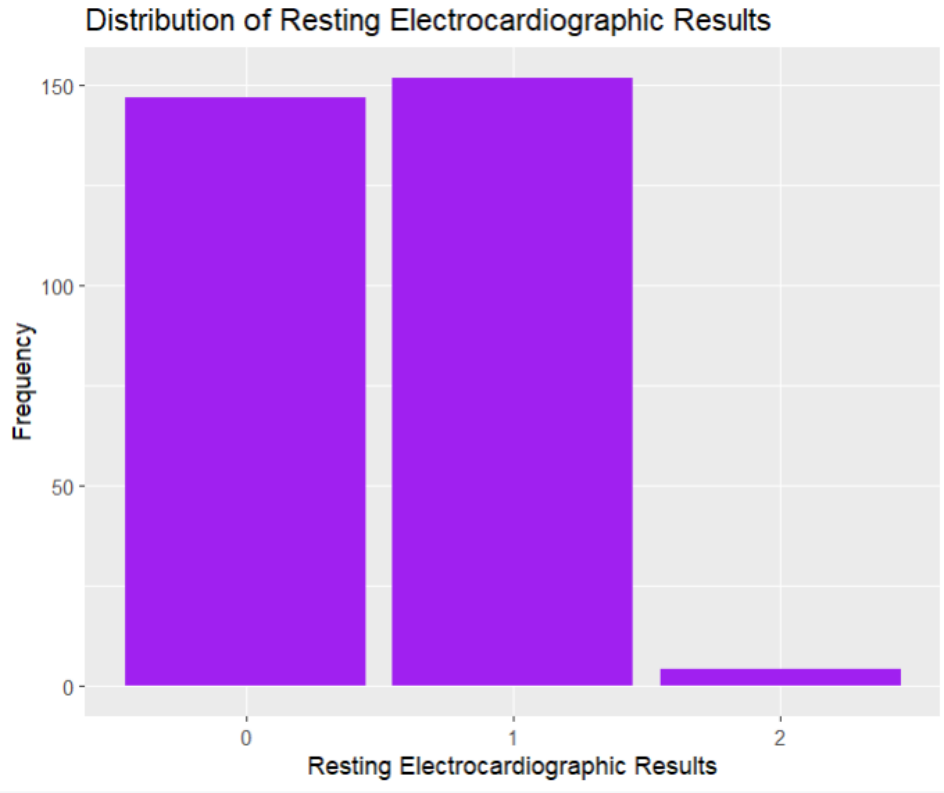

Figure 5: Distribution of Resting Electrocardiographic Results

Figure 5 visualizes the distribution of resting electrocardiographic results.

4.2 Correlation Matrix

We calculated the correlation matrix to quantify the relationships between different attributes.

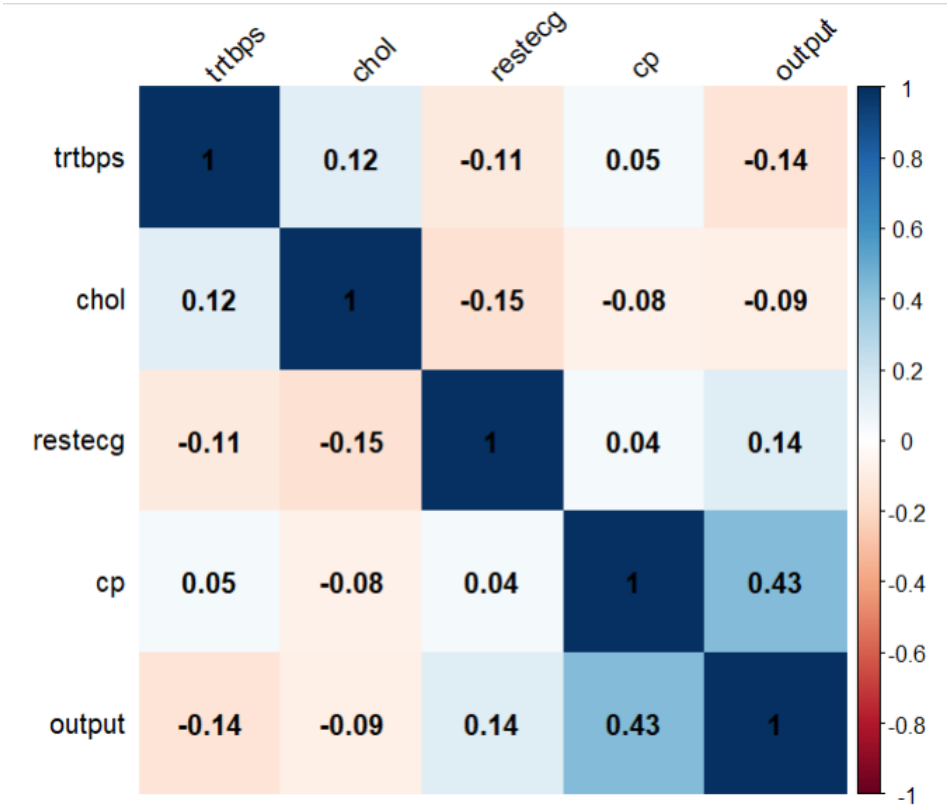

Figure 6: Correlation Matrix

Figure 6 visually represents the correlation matrix using the corrplot package. The values in the matrix provide insights into the strength and direction of relationships between the attributes.

These visualizations and analyses lay the foundation for our regression analysis, allowing us to understand the distribution of individual attributes and identify potential correlations between them.

5 Implementation

5.1 Simple Linear Regression for Resting Blood Pressure

The simple linear regression analysis for resting blood pressure is as follows:

```
# Simple Linear Regression for Resting Blood Pressure
model_trtbps <- lm(output ~ trtbps, data = heart_data)
summary(model_trtbps)
```

The results of the analysis indicate that the relationship between resting blood pressure and the likelihood of a heart attack is as follows:

- Coefficient for `trtbps`: -0.004122
- Residual standard error: 0.4944
- Multiple R-squared: 0.02101
- Adjusted R-squared: 0.01775
- F-statistic: 6.458 on 1 and 301 DF, p-value: 0.01155

5.2 Simple Linear Regression for Cholesterol

The simple linear regression analysis for cholesterol is as follows:

```
# Simple Linear Regression for Cholesterol
model_chol <- lm(output ~ chol, data = heart_data)
summary(model_chol)
```

The results of the analysis indicate that the relationship between cholesterol levels and the likelihood of a heart attack is as follows:

- Coefficient for `chol`: -0.0008204
- Residual standard error: 0.4978
- Multiple R-squared: 0.007266
- Adjusted R-squared: 0.003968
- F-statistic: 2.203 on 1 and 301 DF, p-value: 0.1388

5.3 Simple Linear Regression for Chest Pain Type

The simple linear regression analysis for chest pain type is as follows:

```
# Simple Linear Regression for Chest Pain Type
model_cp <- lm(output ~ as.factor(cp), data = heart_data)
summary(model_cp)
```

The results of the analysis indicate that the relationship between chest pain type and the likelihood of a heart attack is as follows:

- Coefficients for chest pain types (1, 2, 3)
- Residual standard error: 0.4285
- Multiple R-squared: 0.2696
- Adjusted R-squared: 0.2623
- F-statistic: 36.79 on 3 and 299 DF, p-value: ¡ 2.2e-16

5.4 Simple Linear Regression for Resting Electrocardiographic Results

The simple linear regression analysis for resting electrocardiographic results is as follows:

```
# Simple Linear Regression for Resting Electrocardiographic Results
model_restecg <- lm(output ~ as.factor(restecg), data = heart_data)
summary(model_restecg)
```

The results of the analysis indicate that the relationship between resting electrocardiographic results and the likelihood of a heart attack is as follows:

- Coefficients for electrocardiographic results (1, 2)
- Residual standard error: 0.4921
- Multiple R-squared: 0.03308
- Adjusted R-squared: 0.02663
- F-statistic: 5.132 on 2 and 300 DF, p-value: 0.006436

6 Experimental Results

6.1 Summary of Simple Linear Regression Analysis

The simple linear regression analyses for the selected attributes provide insights into the relationships with the likelihood of a heart attack. Here are the key findings:

6.1.1 Resting Blood Pressure (`trtbps`)

The simple linear regression model for resting blood pressure indicates a limited explanatory power with an R^2 value of 0.0210. The coefficient for resting blood pressure (`trtbps`) is statistically significant ($p < 0.05$), suggesting a modest negative association with the likelihood of a heart attack.

6.1.2 Cholesterol (`chol`)

The simple linear regression model for cholesterol levels shows a low R^2 value of 0.0073, indicating a weak relationship with the likelihood of a heart attack. The coefficient for cholesterol (`chol`) is not statistically significant ($p > 0.05$), suggesting that cholesterol alone might not be a strong predictor.

6.1.3 Chest Pain Type (`cp`)

The simple linear regression model for chest pain type exhibits a relatively higher R^2 value of 0.2696, signifying a stronger relationship with the likelihood of a heart attack. The coefficients for different chest pain types (1, 2, 3) are statistically significant ($p < 0.05$), highlighting their importance in predicting

7 Relation Analysis

7.1 Resting Blood Pressure

The regression analysis for resting blood pressure and the likelihood of a heart attack reveals a weak relationship. The coefficient estimate for resting blood pressure is negative ($\beta = -0.0041$), suggesting a slight decrease in the likelihood of a heart attack as resting blood pressure increases. However, the R^2 value is low (2.10%), indicating that resting blood pressure alone has limited explanatory power in predicting heart attacks.

7.2 Cholesterol

The regression analysis for cholesterol levels and the likelihood of a heart attack indicates a negligible relationship. The coefficient estimate for cholesterol is close to zero ($\beta = -0.0008$), and the R^2 value is only 0.73%. This suggests that cholesterol levels alone do not significantly contribute to the prediction of heart attacks in the given model.

7.3 Chest Pain Type

The analysis of chest pain type and the likelihood of a heart attack shows a more substantial relationship. The model indicates that different types of chest pain (`cp`) have a significant impact on the likelihood of a heart attack. The R^2 value for this model is 26.96%, suggesting a moderate level of explanation for the variability in heart attack likelihood based on chest pain type.

7.4 Resting Electrocardiographic Results

The regression analysis for resting electrocardiographic results and the likelihood of a heart attack yields mixed results. The coefficient estimates for the different levels of `restecg` indicate varying impacts on heart attack likelihood. However, the overall model has a relatively low R^2 value of 3.31%, suggesting that resting electrocardiographic results alone may not be highly predictive of heart attacks.

8 R^2 Values

The R^2 values for each attribute are as follows:

- Resting Blood Pressure: 2.10%
- Cholesterol: 0.73%
- Chest Pain Type: 26.96%
- Resting Electrocardiographic Results: 3.31%

These values indicate the proportion of variability in heart attack likelihood that is explained by each respective attribute.

9 Conclusion

In conclusion, the regression analysis suggests that while resting blood pressure and cholesterol levels have limited predictive power for heart attacks, chest pain type shows a more significant correlation. However, the overall explanatory power of the models remains modest. Further investigation, potentially with more attributes or advanced modeling techniques, may be required to enhance the accuracy of heart attack predictions.