**Module 6 - Week 6: R Practice Assignment**

**(Heart Disease Data Analysis)**

**Sriram Kannepalli**

**Northeastern University: College of Professional Studies**

**ALY 6010.**[**71820**](https://northeastern.instructure.com/courses/196161)**: Probability Theory and Introductory Statistics**

**Professor Antonius Breur**

**November 1st, 2024**

**Abstract**

This report explores relationships between various clinical variables and the presence of heart disease. Through regression modeling, dummy variable creation, and visualization techniques, we investigate the impact of age, cholesterol levels, chest pain type, and other factors on the target variable (heart disease presence). Results indicate significant contributions from specific variables, particularly chest pain type and max heart rate, highlighting the differences across subsets.

**Introduction**

This analysis examines relationships between several clinical variables and the likelihood of heart disease in a dataset containing 1,190 records. We employ regression analysis, dummy variables, and visualization techniques to analyze the influence of factors such as age, cholesterol levels, and chest pain type. Additionally, we assess the impact of these factors within different subsets defined by chest pain type.

**Methods**

**Dataset Description**

The dataset consists of 1,190 observations of patients’ clinical attributes, including:

* **Age**: Age in years.
* **Sex**: 0 = Female, 1 = Male.
* **Chest Pain Type**: Categorical variable with four levels.
* **Clinical Measurements**: Resting blood pressure, cholesterol, fasting blood sugar, max heart rate, and oldpeak.
* **Target**: Binary outcome (0 = No heart disease, 1 = Presence of heart disease).

**Data Preparation**

1. **Missing Values Handling**: Missing values in numeric columns were replaced with column means to ensure a complete dataset for analysis.
2. **Factor Conversion**: Categorical variables (sex and chest\_pain\_type) were converted into factors to simplify the analysis.
3. **Descriptive Statistics**: Summary statistics were computed to explore the distribution and characteristics of each variable.

A screenshot of a computer screen

Description automatically generated**Insight**: The data summary provides an overview of each variable’s range, mean, and median values. Notably, the chest pain types vary in frequency, with a large number of asymptomatic cases, which may impact the interpretation of heart disease risk among different chest pain types.

**Results**

**Initial Regression Analysis**

The initial regression model used continuous predictors (age, cholesterol, max heart rate) to explore their impact on the presence of heart disease. The R-squared value for this model was 0.196, indicating that approximately 19.6% of the variation in heart disease presence is explained by these predictors.

A screenshot of a computer

Description automatically generated**Insight**: The initial regression analysis shows that age and max heart rate are significant predictors of heart disease, with max heart rate showing a negative association. This suggests that lower max heart rates are associated with higher likelihoods of heart disease presence, while age is positively associated with it.

**Enhanced Regression with Dummy Variables**

We introduced dummy variables for sex and chest\_pain\_type to improve the model's explanatory power. An interaction term (age:sex) was added to investigate gender-specific effects of age on heart disease. The model’s R-squared increased to 0.39, demonstrating improved predictive power.

A screenshot of a computer

Description automatically generated**Insight**: Including dummy variables, particularly for chest pain type, significantly improves the model. Asymptomatic chest pain type is strongly associated with higher heart disease risk, highlighting the importance of chest pain classification in predicting heart disease.

**Visualization of Variable Distributions**

To better understand the data distribution, we created visualizations, including:

* **Age Distribution**: A histogram showing the frequency of age values.
* **Cholesterol Levels by Heart Disease Presence**: A boxplot illustrating the cholesterol distribution among those with and without heart disease.

A graph of a number of blue bars

Description automatically generated

A graph with a chart and a diagram

Description automatically generated with medium confidence

**Insight**: The age distribution shows that most patients fall within the middle age range, with relatively fewer younger and older patients. The cholesterol boxplot reveals that patients with heart disease tend to have slightly higher cholesterol levels, though there is substantial overlap.

**Scatter Plot with Regression Lines for Chest Pain Type**

A scatter plot with regression lines was generated to observe the relationship between age and heart disease presence across different chest pain types. The plot illustrates how the impact of age on heart disease varies by chest pain type.

A graph of a graph

Description automatically generated with medium confidence

**Insight**: The scatter plot indicates that patients with asymptomatic chest pain have a stronger positive relationship between age and heart disease. In contrast, typical angina appears to have a weaker association, highlighting potential differences in age-related risk based on chest pain type.

**Separate Regression Analyses by Chest Pain Type**

Separate regression models were created for each chest pain type to assess how age, cholesterol, and max heart rate influence heart disease within each subset.

* **Typical Angina**: This model showed no significant predictors, indicating limited explanatory power for this subset.

A screenshot of a computer

Description automatically generated

**Insight**: The lack of significant predictors for typical angina suggests that age, cholesterol, and max heart rate may not play as strong a role in predicting heart disease presence for this group.

* **Atypical Angina**: Age and max heart rate emerged as significant predictors, highlighting age-related risk within this subset.

A screenshot of a computer

Description automatically generated

**Insight**: For atypical angina, both age and max heart rate are significant predictors, indicating that older age and lower max heart rates increase heart disease risk for these patients.

* **Non-Anginal Pain**: Age, cholesterol, and max heart rate were significant predictors, suggesting these factors contribute more to heart disease presence in this subset.

A screenshot of a computer

Description automatically generated

**Insight**: Patients with non-anginal pain show strong associations with age, cholesterol, and max heart rate, suggesting that these variables are particularly important for assessing heart disease risk in this group.

* **Asymptomatic**: Age, cholesterol, and max heart rate were again significant predictors, demonstrating the relevance of these factors among asymptomatic patients.

A screenshot of a computer

Description automatically generated

**Insight**: The asymptomatic group has the highest significance for age, cholesterol, and max heart rate, indicating that these patients may carry a hidden risk profile detectable through these clinical metrics.

**Separate Panel Comparison of Regression Lines by Chest Pain Type (Step 9)**

We created a faceted scatter plot to visualize regression lines for each chest pain type separately. This plot highlights distinct trends for each category, emphasizing the differences in how age impacts heart disease presence among the subsets.

A graph of a normal pain type

Description automatically generated with medium confidence**Insight**: The faceted plot clearly shows different slopes and intercepts for each chest pain type. Asymptomatic and non-anginal pain groups display a steeper positive relationship between age and heart disease presence, whereas typical angina has a flat line, reflecting its lack of significant predictors.

**Discussion**

The results indicate that age, cholesterol, and max heart rate play significant roles in predicting heart disease presence, with some variability based on chest pain type. While the initial regression model provided a general overview, including dummy variables for categorical predictors and performing subset analyses allowed for a more nuanced understanding of the data. The faceted plot for chest pain type illustrates the importance of considering categorical distinctions in heart disease risk assessments.

**Key Findings:**

* **Enhanced Regression Model**: Including dummy variables improved the model's explanatory power, increasing the R-squared from 0.196 to 0.39.
* **Subset Analysis**: Separate analyses for each chest pain type reveal that risk factors influence each subset differently, particularly among those with non-anginal pain and asymptomatic presentations.

These findings underscore the necessity of incorporating categorical variables and performing subset analyses when examining complex relationships in clinical data.

**References**

**James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: with Applications in R* (2nd ed.). New York, NY: Springer. Available at:** [**https://www.statlearning.com/**](https://www.statlearning.com/) **(Accessed: October 30, 2024).**

**Wickham, H., & Grolemund, G. (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data* (1st ed.). Sebastopol, CA: O'Reilly Media. Available at:** [**https://r4ds.had.co.nz/**](https://r4ds.had.co.nz/) **(Accessed: October 29, 2024).**

**UCI Machine Learning Repository. (n.d.). *Heart Disease Data Set.* Irvine, CA: University of California, School of Information and Computer Science. Available at:** [**https://archive.ics.uci.edu/ml/datasets/Heart+Disease**](https://archive.ics.uci.edu/ml/datasets/Heart+Disease) **(Accessed: October 31, 2024).**