Project: Comprehensive Regression Analysis and Modeling

Yoonsung Jung

October 2024

Objective

The objective of this project is to perform comprehensive regression analysis using multiple regression models on a real-world dataset. This project includes Exploratory Data Analysis (EDA) to understand the data and then applying different regression techniques to build and evaluate models.

Dataset Suggestions

Data available from github.com or kaggle.com NOT allow on this project. You can consider to find your data from the following site:

- **OECD Data**
- **UNdata**
- **Knoema**
- **Harvard Dataverse**
- **European Union Open Data Portal**
- **Data.gov**
- **World Bank Open Data**
- **UCI Machine Learning Repository**

Instructions

1. Introduction and Objective

- **Purpose**: Provide a brief description of the project, its goals, and what you aim to achieve.

2. Dataset Selection and Description

- **Dataset Selection**: Choose a dataset relevant to your interests or major.
- **Dataset Description**: Describe the dataset, including the source, the number of observations, and the types of variables included.

3. Data Pre-Processing

- **Import Data**: Use Jupyter Notebook to import the dataset.
- **Data Cleaning**:
  - Handle missing values (e.g., imputation, removal).
  - Remove duplicates.
  - Correct any inconsistencies or errors in the data.

4. Exploratory Data Analysis (EDA)

Univariate Analysis

- Summary statistics (mean, median, mode, standard deviation, etc.).
- Visualize distributions using histograms, box plots, and density plots.

Bivariate Analysis

- Analyze relationships between two variables using scatter plots, correlation matrices, and pair plots.

Multivariate Analysis

- Explore interactions between multiple variables using heatmaps, pairwise relationships, and dimensionality reduction techniques like PCA (Principal Component Analysis).

5. Regression Analysis

Simple Linear Regression

- Perform a simple linear regression analysis using one predictor variable.
- Evaluate the model performance using metrics like R^2 and RMSE.
- Visualize the regression line and the residuals.

Multiple Linear Regression

- Perform a multiple linear regression analysis using multiple predictor variables.
- Evaluate the model performance using metrics like R^2 and RMSE.
- Analyze the significance of each predictor variable.

Polynomial Regression

- Apply polynomial regression to capture non-linear relationships.
- Compare the performance of polynomial regression with linear regression.

Logistic Regression

- Apply logistic regression for binary classification problems.
- Evaluate the model performance using metrics like accuracy, precision, recall, and ROC-AUC.

Regularization Techniques

- Apply LASSO and Ridge regression to handle multicollinearity and prevent overfitting.
- Compare the performance of LASSO, Ridge, and Elastic Net regression.

Advanced Regression Techniques

- Apply Quantile Regression to predict conditional quantiles.
- Use Poisson Regression for count data.
- Apply Negative Binomial Regression for overdispersed count data.
- Implement Zero Inflated and Hurdle Regression for zero-inflated data.
- Use Cox Regression for survival analysis.
- Apply Partial Least Squares Regression and Principal Component Regression for high-dimensional data.

6. Model Evaluation and Comparison

- Evaluate the models using appropriate metrics.
- Compare the performance of different models.
- Visualize and interpret the results.

7. Results and Interpretation

- **Summary of Findings**: Summarize the key findings from your regression analysis.
- **Insights**: Provide insights and potential implications of the results.
- **Limitations**: Discuss any limitations of your analysis.

8. Report and Presentation

- **Report**: Create a detailed report documenting all steps, findings, and interpretations.
- **Presentation**: Prepare a presentation to share your findings with the class. Include key visualizations and insights.

Submission Guidelines

- **Code**: Submit Jupyter Notebooks (.ipynb) including all Python scripts used for the project.
- **Data**: Submit your excel data used for the project.
- **Report**: Submit a report of your analysis (.pdf or .doc or .html).

Evaluation Criteria

- **Completeness**: Did the student complete all steps of the analysis?
- **Accuracy**: Are the calculations and results correct?
- **Clarity**: Is the code well-documented and easy to understand?
- **Insights**: Are the insights and interpretations meaningful and well-explained?
- **Presentation**: Is the presentation clear, engaging, and well-structured?