Python Project Report

Multiple Linear Regression

Authors:
Bhashkar Paudyal [1]
Supreme khatiwada[2]
Bipasana Shrestha[3]
Shiksha bhattarai[4]

Institution:
Department of Mathematics, Kathmandu University

Course:
DSMA-113 - Introduction to Programming in Python

Instructor:
Er. Narayan Sapkota

Submission Date:
March 1, 2025

---

[1]Contributed to Python Implementation of Linear Regression
[2]Contributed to Data Cleaning
[3]Contributed to Project Report
[4]Contributed to Project Report

Abstract

This report investigates using multiple linear regression and potentially predicting diabetes progression based on many parameters of patient health attributes. The data will be built up with preprocessing, exploratory data analysis (EDA), including descriptive statistics and visualization, and model building using Python with all those pieces incorporated. The analysis lays out the key risk factors in case, importance feature correlation, model predictiveness assessment providing insight for health decisions in consideration of more future research on effective diabetes management.

This report is a perspective on the regression analysis for the prediction of diabetes progression by multiple linear regression. The exploratory study takes the approach of preprocessing data then EDA using descriptive statistics and visualization, and building a model in Python. It explores the significant risk factors, checks for important feature correlations, to be considered in decision-making in health and further broadening the research scope of diabetes diagnosis.

Contents

Introduction

1.1 Introduction to Diabetes

Diabetes is a disease that poses a major threat in the global health scenario today. The disease is mostly characterized by increased levels of glucose in the bloodstream, which can lead to a variety of complications if not monitored and managed properly. All around the world, millions of people have been with diagnosed with this disease which is primarily of two types:

- **TYPE I**, which is an autoimmune condition where the body attacks insulin producing cells.
- **TYPE II**, which is often brought upon by lifestyle factors and genetics

1.2 Diabetes Data Description

The data set consists of ten attributes on various medical factors for each of 442 patient record instances, namely:

Feature Name	**Description**
`age`	Age in years
`sex`	Gender (0 = female, 1 = male)
`bmi`	Body mass index
`bp`	Average blood pressure
`s1`	Total serum cholesterol (TC)
`s2`	Low-density lipoproteins (LDL)
`s3`	High-density lipoproteins (HDL)
`s4`	Total cholesterol / HDL ratio (TCH)
`s5`	Possibly log of serum triglycerides level (LTG)
`s6`	Blood sugar level (GLU)

Table 1.1: Data Description

1.3 Objectives

1. **To analyze the relationship between various medical factors and the risk of diabetes.** This involves examining how factors such as age, sex, BMI, blood pressure, and cholesterol levels impact the likelihood of developing diabetes.
2. **To develop a predictive model that can accurately identify individuals at high risk of diabetes.** This model will utilize machine learning techniques to analyze the dataset and predict the risk based on the input features.
3. **To evaluate the effectiveness of the predictive model in real-world scenarios.** This includes assessing the model's performance using metrics such as accuracy, precision, and recall, and comparing it with existing models.
4. **To provide insights for healthcare practitioners and policymakers.** The findings from this study aim to inform strategies for early detection, prevention, and management of diabetes.

1.4 Motivation

The goal of this project is to use Python to analyze the diabetes dataset "diabetes.csv" to determine patterns, using specific variables such as age, BMI, blood pressure, and blood glucose levels. Through this, we aim to establish foundations for larger future research initiatives that might help doctors develop more effective patient treatment and care methods. For the project, we have mainly focused on getting the data ready, analyzing it, and creating some visualizations. This project is a stepping stone for creating a model that we hope might actually forecast things which occur in the future.

1.5 Background/Theory

The project mainly focuses on examining the 'diabetes.csv' dataset, which consists of the medical and life-style traits utilized in diabetes studies, a chronic disease with excessive blood sugar because of the incapability of the body to synthesize or regulate insulin.
Starting off with the 'diabetes.csv' dataset, it was important for us to preprocess the data before working on it.
Our main task within the project was to perform Exploratory Data Analysis (EDA) using Python in order to investigate and analyze trends within the dataset.

1.5.1 Understanding Exploratory Data Analysis (EDA)

This includes summarizing the dataset, providing suitable pattern recognition, detection of outliers, and visualizing the data for better insights. EDA in this case is performed on the 'diabetes.csv' dataset, applying Python.

For this project, we have used:

- **Pandas**: For cleaning and accessing the data.
- **NumPy**: For numerical calculations.
- **Matplotlib & Seaborn**: For visualizations.
- **Scikit Learn**: For Machine learning tools.

1.5.2 Regression Model

> **Note** (Regression Formula). Here, X is the input matrix column stacked with a vector of ones and y is the output.
>
> $$B = (X^T X)^{-1} X^T y$$

For the python implementation, see the Appendix A.1.

Methodology

The systematic approach of the project focused on leveraging the understanding of the 'diabetes.csv' dataset in order to design a model to predict the progression of the disease. The methodology had several steps which included: data processing, exploratory data analysis (EDA), and formulating and testing a linear regression model. The model's precision and exactitude were highly regarded undertakings in each constituent of the project, which were cautiously conducted.

2.1 Data Preprocessing

2.1.1 Data Loading

The National Institute of Diabetes and Digestive and Kidney Diseases provided this dataset. The goal is to determine if it's possible to foresee whether a certain individual is diabetic based on target diagnostic measurements. The dataset comprises 442 instances with 10 attributes, some of which are demographic and clinical such as age, sex, BMI (Body Mass Index), blood pressure, and a collection of six serum measurements (s1–s6). On the one hand, the target variable y measures the progression of the disease one year post baseline.

> **Remark.** The dataset is particularly valuable for studying diabetes progression due to its inclusion of both demographic and clinical attributes, which provide a holistic view of the factors influencing the disease.

2.1.2 Treating the Missing Values

In data cleaning, the first step taken was assessing if there were any missing values present. Utilizing the `df.isnull().sum()` function, it was assured that there are no missing values (NaN) in the dataset. Inspection showed that zeros were present in certain features, which can be interpreted as missing or invalid values (e.g., blood pressure zeros or zero BMI). Exploring data analysis, these zeros were set aside to determine if they are real values or simply missing data.

> **Note.** The presence of zeros in features like blood pressure and BMI may indicate missing or erroneous data, which requires careful handling to avoid bias in the analysis.

2.2 Exploratory Data Analysis (EDA)

2.2.1 Summary Statistics

Summary statistics were constructed in order to prepare the database and understand its structure. Important findings were:

- Participants had a mean age of 48.5 years (SD = 13.1), which indicates that the sample is a middle-aged population.
- The average value of BMI was 26.4 (SD = 4.4), suggesting that the participants are on the upper side of average weight according to BMI ranges.
- Disease progression, the target variable, had a mean of 152.5 (SD = 77.1), indicating, with a high degree of variability, the severity of diseases among participants.

These statistics justify the further analysis as they reveal the need for normalization considering the different ranges of scales of the features.

Definition 2.2.1 (Normalization). Normalization is the process of scaling features to a common range (e.g., 0 to 1) to ensure that variables with larger magnitudes do not disproportionately influence the model.

2.2.2 Data Visualization

Visualization techniques were implemented in the dataset to determine the distributions of data, variable outliers, and variable dependencies.

Histograms

Remark. The skewed distributions observed in histograms indicate that certain features may require transformations (e.g., log or square root) to better align with the assumptions of statistical models.

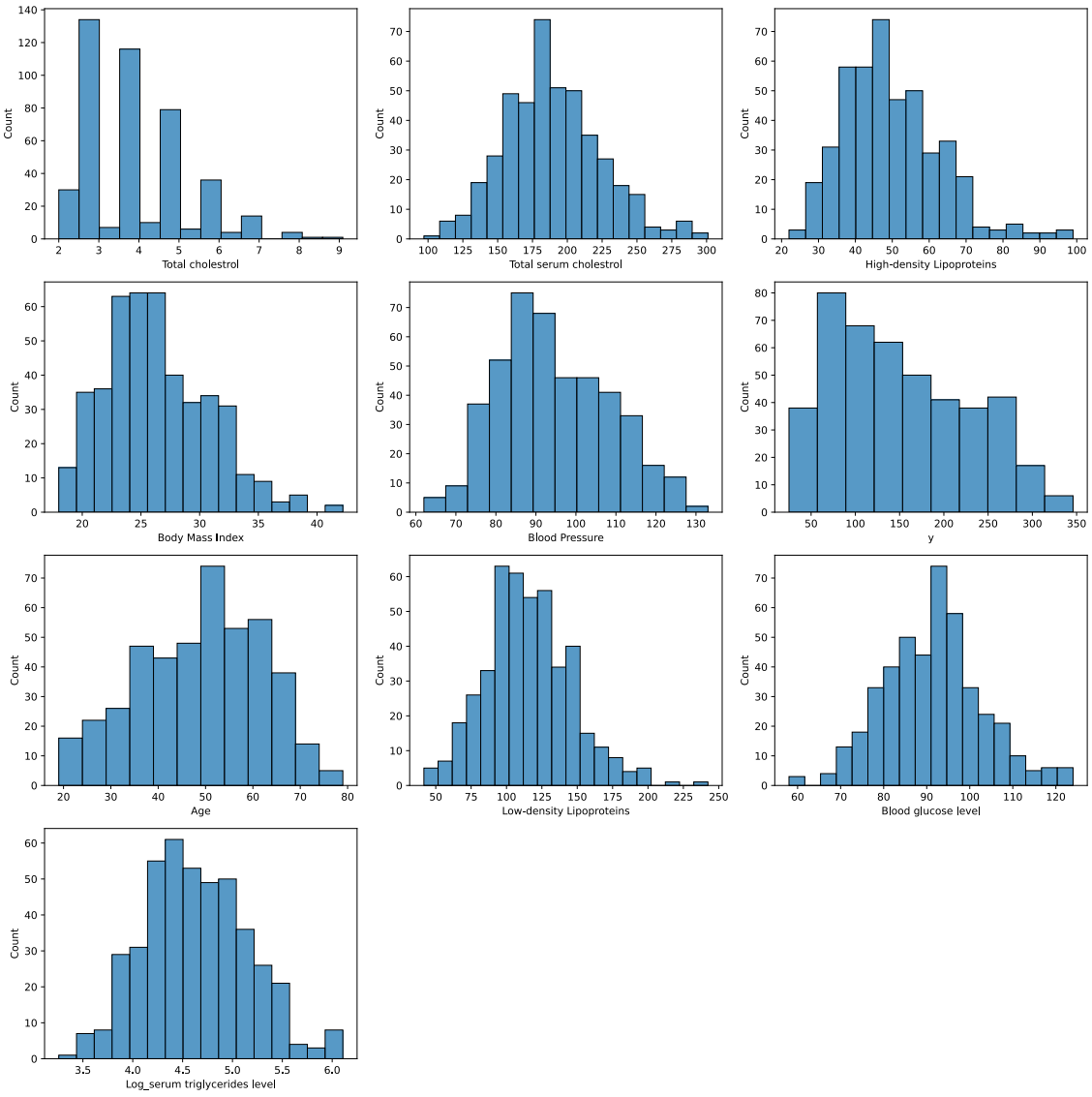

Figure 2.1: Histograms

Box Plots Feature box plots were captured to determine the outlier values if any existed.

> **Note.** Outliers in features like BMI and blood pressure may represent extreme cases or data entry errors, and their impact on the model should be carefully evaluated.

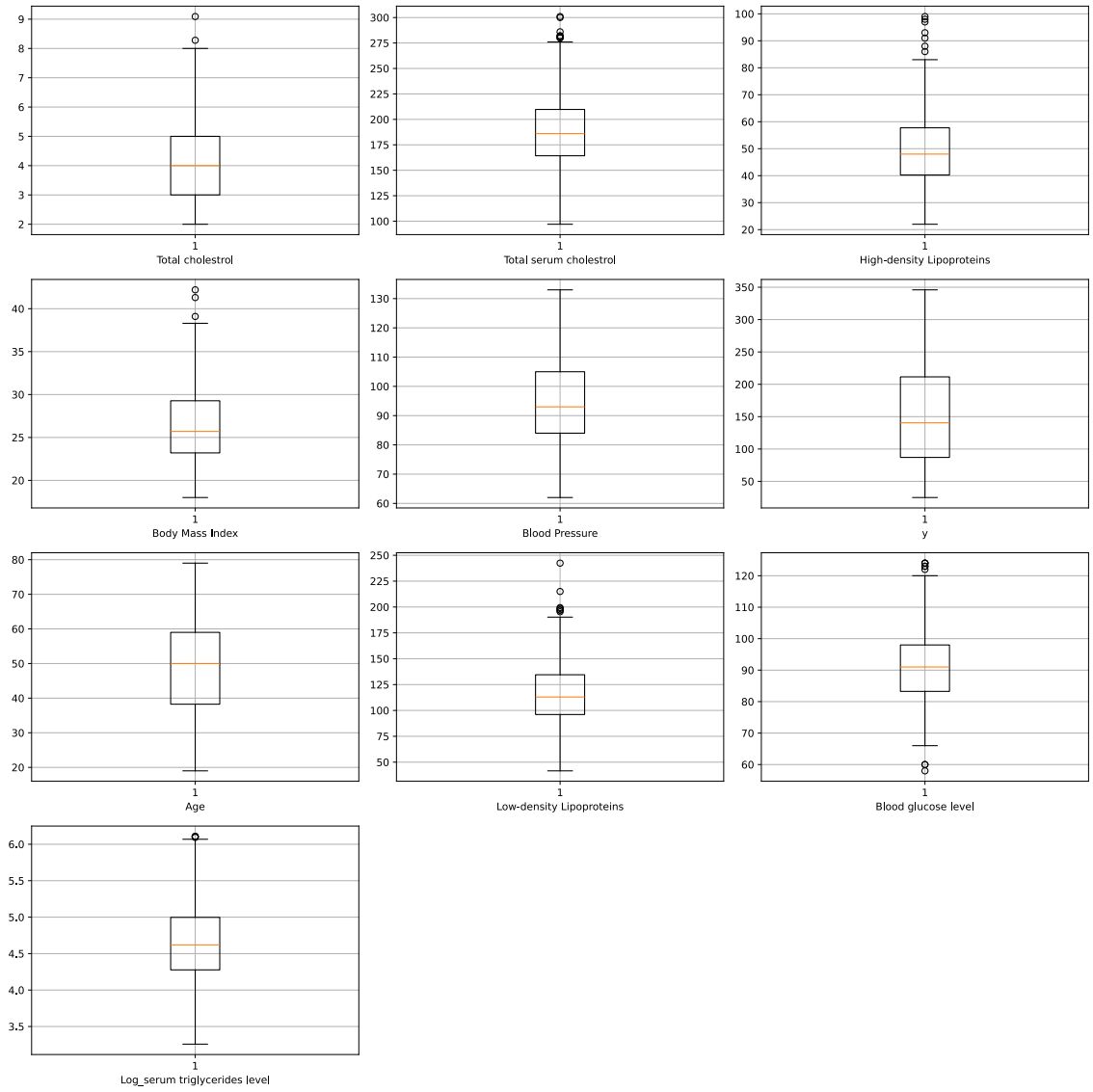

Figure 2.2: Boxplots

Q-Q Plots Quantile-Quantile (Q-Q) plots helped to establish the normality of distribution within datasets.

> **Lemma 2.2.1.** If the data points in a Q-Q plot deviate significantly from the reference line, the data distribution is non-normal, and transformations or robust statistical methods are recommended.

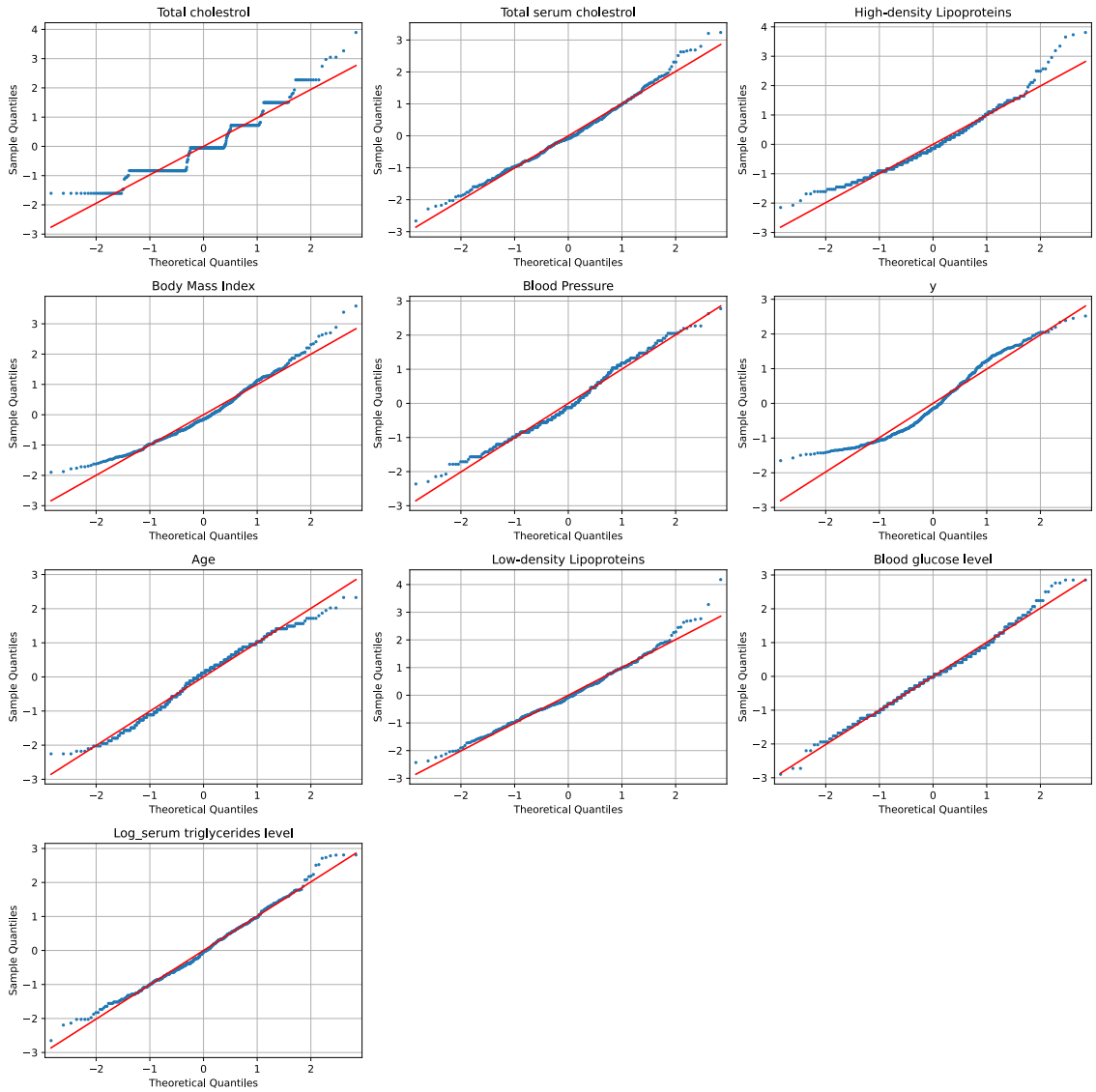

Figure 2.3: Q-Q Plots

Correlation Heatmap To analyze the interaction between the features and the target variable, a correlation heatmap was prepared.

> **Corollary 2.2.1.** Features with strong correlations (e.g., BMI and s5) are likely to be significant predictors of disease progression, while weakly correlated features (e.g., sex) may have limited predictive power.

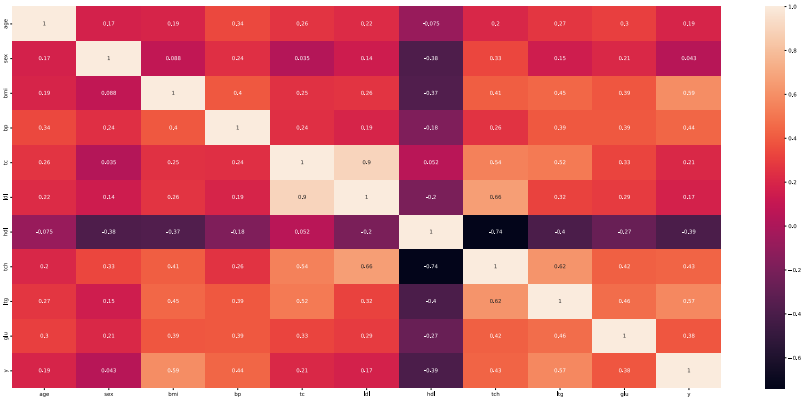

Figure 2.4: Correlation Heatmap

Bar Plots Bar plots were created to show the distribution of various categories of the population such as sex.

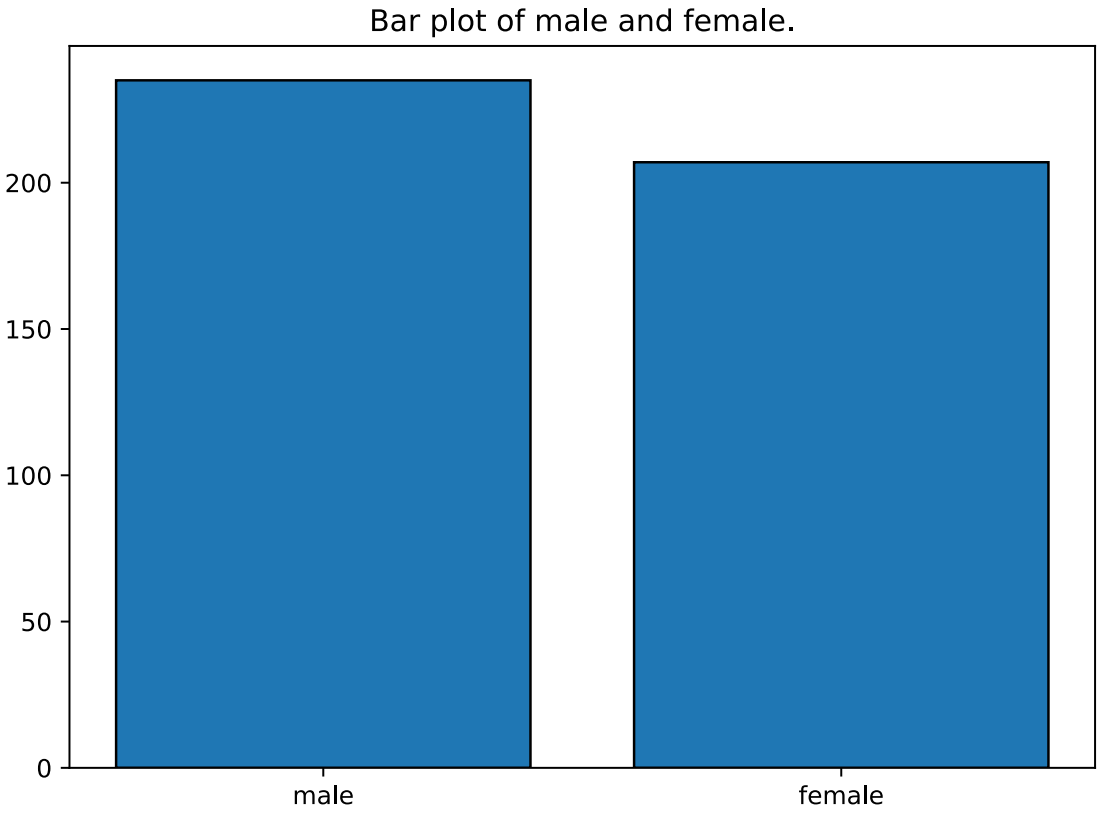

Figure 2.5: Bar Plot

2.3 Linear Regression

2.3.1 Normalization

In order to create the regression model, the defined elements had to be normalized so that each variable had an equal weight in the analysis. Normalization adjusted each feature's value to a common range, usually zero to one, so that features with higher magnitudes would not influence the model incompatibly. This adjustment was essential because it enhanced the efficacy and clarity of the model.

Remark. Normalization ensures that all features contribute equally to the model, preventing bias toward variables with larger scales.

2.3.2 Data Splitting

Data is split into training and testing sets at a ratio of 80:20. The training set is utilized for building the linear regression model selected, and the retained parts are stored for evaluation of the model's performance after the model has already been trained. This method enabled evaluation of the model's accuracy on predictive tasks in which the data was not initially available, therefore fostering better accuracy in generalization during the model's application.

Definition 2.3.1 (Training and Testing Sets)**.** The training set is used to build the model, while the testing set is used to evaluate its performance on unseen data, ensuring the model generalizes well.

2.3.3 Data Modeling

Linear Regression

In order to establish the relations that exist among the attributes and the target variable, a small-scale linear regression model is selected. Linear Regression is one of the pillars in the world of predictive modeling and works best when multiple independent variables need to be analyzed against a singular variable that is continuously measured. The model was fit using the normalized training set, while assuming that all essential features contribute to worsened advancement of the disease and estimating the coefficients to each of them.

Lemma 2.3.1. The coefficients in a linear regression model represent the change in the target variable for a unit change in the corresponding feature, holding all other features constant.

2.3.4 Model Evaluation

Evaluation of the linear regression model in this case was based on its performance on the test dataset using the Mean Squared Error (MSE) metric. The model's achieved MSE on the test set was 0.038, which indicates a reasonable model fit. The scatter plot of the predicted values against the actual ones showed moderate spread, especially for higher cases of disease progression. This showed that although the model had performed reasonably well, there was still room for improvement around more chronic cases.

Remark. A lower MSE indicates better model performance, but the scatter plot suggests that the model struggles with accurately predicting extreme cases of disease progression.

Results

Some of the major findings from the analysis of the ‘diabetes.csv‘ dataset in this multiple-regression project disclosed significant insights into the relationships among various health measures and disease extremity.

3.1 Descriptive Statistics

On average, the ages of participants were 48.5 ± 13.1 years, which indicates that the subjects were of a varied age group. The mean BMI was established as 26.4 ± 4.4, which denotes that most of the participants were in the overweight category but varied widely. Disease progression, being the dependent variable, spanned a huge range, with a mean of 152.5 and a standard deviation of 77.1, indicating a wide variation of degree in the progression across the dataset.

> **Remark.** The wide standard deviation in disease progression suggests significant variability in the severity of the condition among participants, which may require further investigation into contributing factors.

3.2 Data Visualization

Various visualization techniques were used to examine the distribution and relationship among the various attributes in the dataset.

3.2.1 Histograms

Histograms [2.1] showed skewed distributions for variables such as blood pressure and serum measurements, suggesting non-normality for these variables.

> **Note.** The skewness in the distributions indicates that transformations (e.g., logarithmic or square root) may be necessary to meet the assumptions of statistical models.

3.2.2 Box Plots

Outliers were identified by Box Plots [2.2] in BMI (5 outliers), blood pressure (8 outliers), and serum measurements, suggesting potential anomalies that could influence the analysis.

> **Remark.** Outliers in features like BMI and blood pressure may represent extreme cases or data entry errors, and their impact on the analysis should be carefully evaluated.

3.2.3 Q-Q Plots

The Q-Q Plots [2.3] validated the non-normal distribution deviations for the majority of the variables, reinforcing the observations from the histograms.

Lemma 3.2.1. If the data points in a Q-Q plot deviate significantly from the reference line, the data distribution is non-normal, and transformations or robust statistical methods are recommended.

3.2.4 Correlation Heatmap [2.4]

A heatmap of correlations revealed moderate to high correlations between the target variable (disease progression) and certain features, including BMI ($r = 0.59$), s5 ($r = 0.57$), and blood pressure ($r = 0.44$), suggesting that these are strong predictors. Conversely, the sex variable had a low correlation ($r = 0.04$) with disease progression.

Corollary 3.2.1. Features with strong correlations (e.g., BMI and s5) are likely to be significant predictors of disease progression, while weakly correlated features (e.g., sex) may have limited predictive power.

3.2.5 Bar Plots

A Bar Graph [2.5] presented an even gender split, with 52% male and 48% female respondents, thereby indicating that the dataset accurately reflects gender distribution.

3.3 Linear Regression Model

After normalizing the data and splitting it into training and test sets, a linear regression model was developed to predict disease progression. The model achieved a Mean Squared Error (MSE) of 0.038 on the test set, demonstrating reasonable predictive accuracy. A scatter plot of predicted versus actual values showed a moderate linear trend, with points grouping along the diagonal line but with considerable scatter at larger values. The scatter indicated that the model was less able to predict more extreme cases of disease progression accurately, possibly because the underlying causes were complex and not well represented in the data. The correlations already found, especially with BMI, s5, and blood pressure, would have helped the performance of the model, although variability with higher values of progression showed that there was scope for improvement in subsequent versions.

Remark. The moderate MSE and scatter plot trends suggest that the model performs reasonably well but struggles with extreme cases, indicating the need for more advanced modeling techniques or additional data.

Discussion

The analysis of the `diabetes.csv` data has provided valuable insights into the factors influencing diabetes progression and its potential implications for clinical decision-making. However, it also highlighted certain limitations and areas for further research.

4.1 Key Findings

The associations between blood pressure, BMI, and disease progression established in this analysis align with existing medical literature, reinforcing their role as major risk factors. These findings underscore the importance of monitoring these parameters in clinical settings. However, the weak correlation with sex ($r = 0.04$) suggests that gender may not be a strong predictor in this dataset. This raises questions about the presence of confounding variables, such as hormonal or lifestyle factors, which should be explored in future research.

> **Remark.** The weak correlation with sex highlights the need for additional variables, such as hormonal profiles or lifestyle factors, to better understand their potential influence on diabetes progression.

4.2 Model Performance

The linear regression model achieved a Root Mean Squared Error (RMSE) of 0.0387 on the test set and 0.0399 on the training set, indicating a good fit. However, the proximity of these errors suggests potential overfitting or that the model's simplicity may prevent it from capturing complex relationships within the data. The scatter plot revealed significant dispersion at higher disease progression values, highlighting the model's struggle to accurately predict severe cases. This suggests that linear assumptions may not sufficiently model the variability in advanced stages of the disease.

> **Note.** The model's difficulty in predicting extreme cases accurately suggests the need for more sophisticated modeling techniques, such as non-linear regression or ensemble methods.

4.3 Limitations

The analysis faced several limitations:

- **Missing Values**: The presence of zeros in variables like blood pressure and BMI may indicate implicit missing values, which were not explicitly handled, potentially leading to biased results.
- **Skewness**: While normalization addressed differences in scale, it did not correct for skewness in the data, which can distort regression coefficients and affect model predictions.
- **Dataset Constraints**: The absence of detailed clinical or lifestyle data limited the model's ability to account for all potential variables influencing diabetes progression.

Remark. Addressing these limitations, such as by imputing missing values or applying transformations to reduce skewness, could improve the model's accuracy and reliability.

4.4 Directions for Future Research

To enhance the analysis and model performance, the following directions are recommended:

- **Feature Engineering**: Incorporating interaction terms, polynomial features, or regularization techniques (e.g., Ridge or Lasso regression) could better capture non-linear relationships and improve generalization.
- **Domain Knowledge**: Collaborating with medical experts to interpret outliers and assess their biological plausibility would strengthen the model's clinical relevance.
- **Additional Data**: Collecting more detailed information on lifestyle factors, medication usage, or genetic markers would enrich the dataset and improve the understanding of diabetes progression.

Corollary 4.4.1. Future research should focus on integrating advanced modeling techniques and additional data sources to better predict diabetes progression and inform clinical decision-making.

Conclusion

The analysis of the ‘diabetes.csv‘ dataset provided valuable insights into the factors influencing the progression of diabetes. Key findings revealed that BMI, blood pressure, and serum measurements (e.g., s5) have significant predictive power, aligning with established medical knowledge. Our linear regression model performed well, achieving a Mean Squared Error (MSE) of 0.038 on the test dataset. The consistency of these findings with traditional medical experience underscores the potential of data-driven approaches in understanding and managing diabetes.

> **Remark.** The strong performance of the model, particularly in predicting disease progression based on BMI, blood pressure, and serum measurements, highlights the importance of these factors in clinical practice.

However, the project also encountered several challenges:

- The model struggled to accurately predict more complex cases of disease progression, suggesting potential overfitting or the influence of outliers.
- The dataset had limitations, such as skewed distributions and zero values in features like blood pressure, which may represent missing or erroneous data.
- These issues emphasize the need for improved data cleaning and refinement in future methodologies.

> **Note.** Addressing these challenges, such as by handling outliers more effectively or applying advanced modeling techniques, could further enhance the model's accuracy and reliability.

Despite these limitations, the project has established a strong foundation for understanding the etiology of diabetes and has opened new avenues for innovation and real-world application. The findings underscore the importance of maintaining key health indicators, such as BMI and blood pressure, in diabetes management. Furthermore, the project highlights the potential of data science in advancing healthcare outcomes and informing public health policies.

> **Corollary 5.0.1.** This study serves as a starting point for future research, paving the way for more sophisticated models and the integration of additional data sources to better predict and manage diabetes progression.

Appendix

Code

A.1 Jupyter Notebook

See the code at here