Regression Analysis and EDA for BodyFat percentage Prediction from Body Measurements

Biswajit Rana
B2330026

RKMVERI, belur

1 Introduction

This report details a comprehensive analysis focused on understanding and predicting body density using regression techniques. Body density, a key indicator of body composition, is often measured through methods that can be time-consuming or inconvenient. This study explores the potential of using easily obtainable body measurements—including age, weight, height, and various circumferences—to predict density through a regression model. The primary goal is to establish a predictive relationship that allows for a more accessible and less invasive assessment of body composition. To achieve this, we will first conduct a thorough Exploratory Data Analysis (EDA). This initial phase will involve examining the distributions of each feature, identifying potential outliers, and uncovering correlations between the measurements and body density. The insights gained from EDA will not only inform our feature selection process but also assist in identifying any assumptions that need to be addressed during modeling. Finally, a regression model will be constructed and rigorously evaluated using standard validation techniques, ensuring the model's accuracy and generalizability for predicting body density. The results of this study may allow for a simplified and more efficient method of estimating body composition.

2 Problem Statement

The goal is to understand the relationship between various body measurements (like age, weight, height, circumferences) and body density using regression analysis. This analysis aims to predict body density based on these measurements, enabling a potentially less invasive way to assess body composition. EDA will be employed to explore data distributions, uncover potential correlations, and assess the suitability of features for regression modeling. Finally, a regression model will be built and evaluated on the data using techniques like train/test splitting. The insights from EDA will further guide the model development for improved prediction accuracy.

3 Data Description

The dataset utilized for this analysis comprises 252 records, each representing an individual, and encompasses 16 distinct features designed to capture various aspects of their physical characteristics. These features are a mix of continuous numerical measurements and a categorical class label, each with a specific purpose for the study.

- **Density:** This continuous variable represents the body density of the individual, typically measured using a complex laboratory procedure. It serves as the target variable for our regression model.
- **Age:** This numerical feature represents the age of the individual in years.
- **Weight:** The weight of the individual is recorded in kilograms (kg).
- **Height:** The height of the individual is provided in centimeters (cm).
- **Neck, Chest, Abdomen, Hip, Thigh, Knee, Ankle, Biceps, Forearm, Wrist:** These features represent circumferences of different body parts, all measured in centimeters (cm). They provide detailed information about the body's dimensions and can reveal relationships with body composition.
- **Class:** This is a neumerical feature that represents the bodyfat percentage of people.
- **Gender:** This column is encoded in the 'class' column and is not explicitly present in the raw data.

The Pandas DataFrame used in this analysis contains the following information:

Table 1: DataFrame Column Details

Column	Non-Null Count	Data Type
Age	249	int64
Weight	249	float64
Height	249	float64
Neck	249	float64
Chest	249	float64
Abdomen	249	float64
Hip	249	float64
Thigh	249	float64
Knee	249	float64
Ankle	249	float64
Biceps	249	float64
Forearm	249	float64
Wrist	249	float64
class	249	float64
Gender	249	int64

4 Exploratory Data Analysis

The analysis of numerical features reveals the following key observations:

- Most numerical features exhibit distributions that are approximately normal, resembling a bell-shaped curve. However, some features show evidence of skewness, indicating an imbalance in their distributions.
- Several features, such as 'Age', 'Weight', 'Height', 'Abdomen', 'Hip', and 'Thigh', possess a relatively wide range of values, suggesting considerable variability within the dataset.
- The distributions of certain features indicate the presence of outliers. These outliers warrant further investigation as they may impact the robustness of the model and require special consideration during pre-processing.

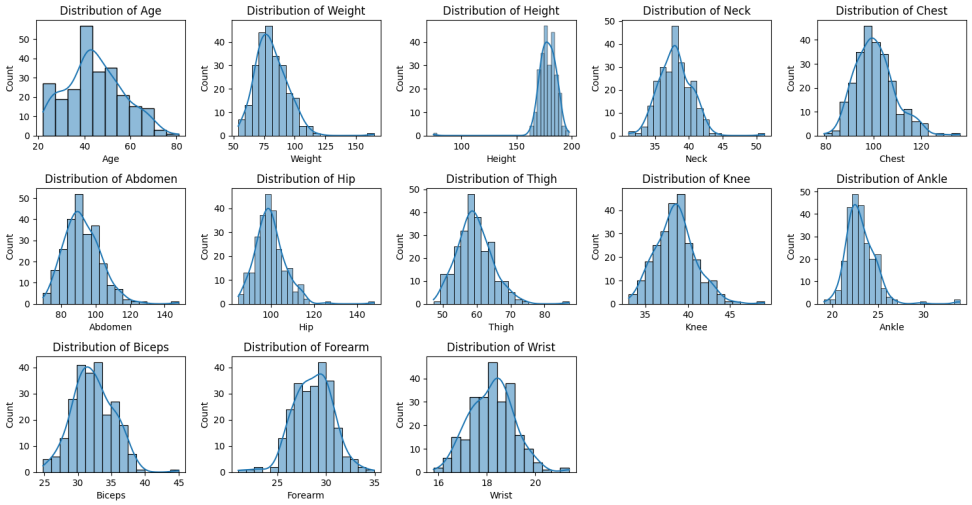

Figure 1: Distribution of all the Regressors except Gender.

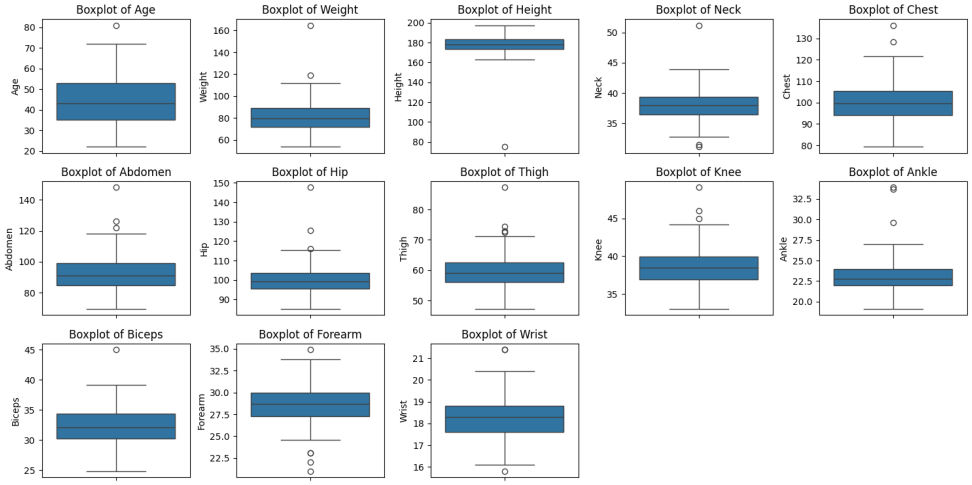

Figure 2: Boxplot of all the Regressors except Gender.

- Only one Catagorical column present in this data is Gender column .Below is the frequency plot of it.0 represents male and 1 represents female.

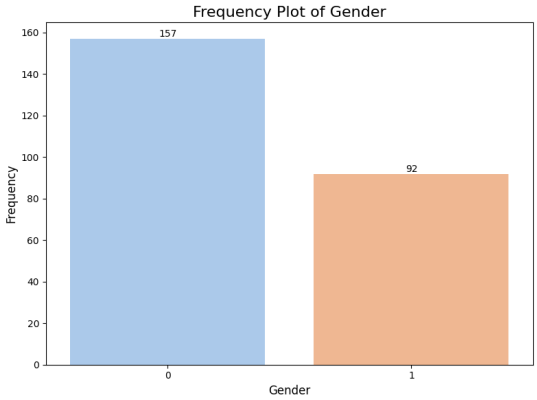

Figure 3: Distribution of Gender.

- Next is the correlation Heatmap between all the features.

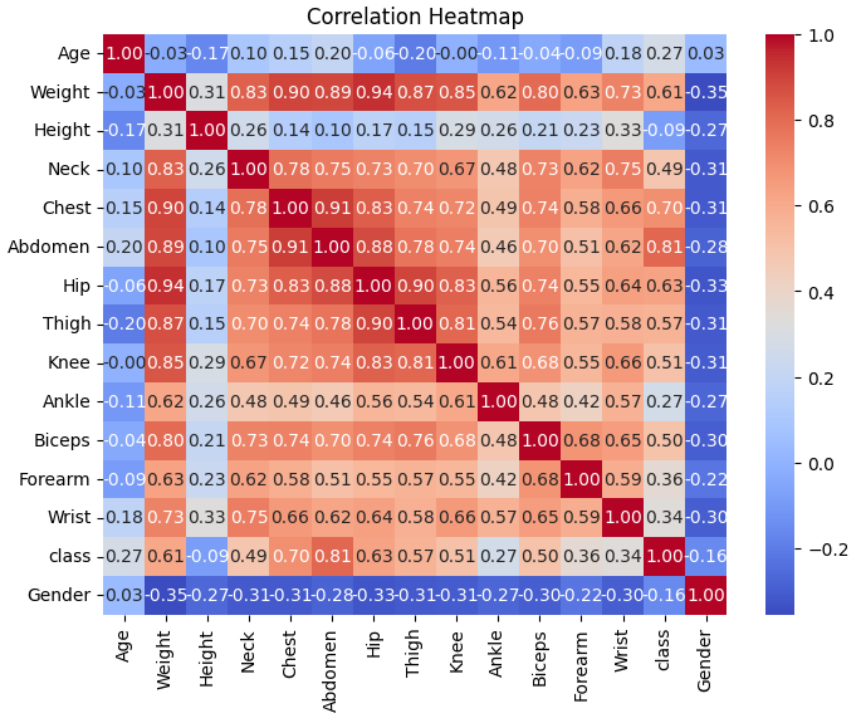

Figure 4: Correlation Heatmap.

- Many columns are correlated with other columns . Example - Weight is correlated with most of the columns.
- Height feature is not correlated to other features as such.

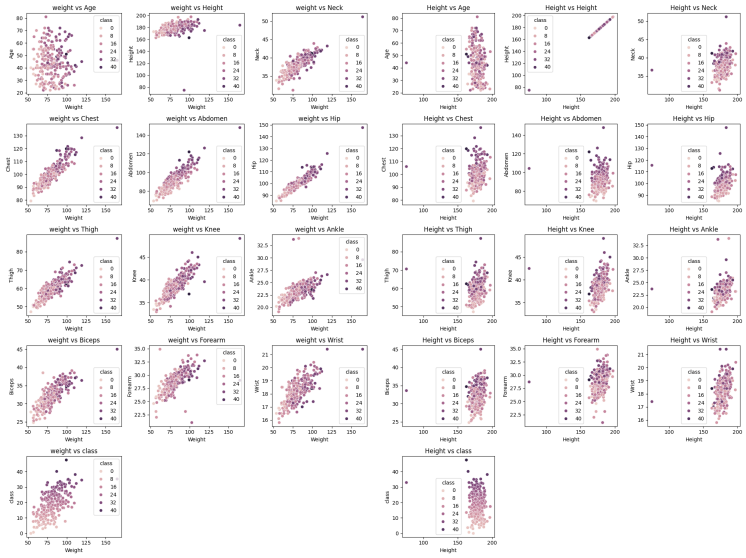

Figure 5: Scatter plots against Weight and height.

5 Regression Analysis

Following is the result of the OLS regression on the Target variable.

- **Dep. Variable:** class
- **R-squared (uncentered):** 0.960
- **Model:** OLS
- **Adj. R-squared (uncentered):** 0.957
- **Method:** Least Squares
- **F-statistic:** 398.9
- **Date:** Wed, 18 Dec 2024
- **Prob (F-statistic):** 5.74e-155
- **Time:** 13:01:04
- **Log-Likelihood:** -709.13
- **No. Observations:** 249
- **AIC:** 1446.
- **Df Residuals:** 235
- **BIC:** 1495.
- **Df Model:** 14
- **Covariance Type:** nonrobust

Table 2: Parameter Estimates

				95% Confidence Interval	
Coefficient	**coef**	**std err**	**t**	**P>|t|**	**Upper**
Age	0.0670	0.032	2.073	0.039	0.131
Weight	-0.0934	0.055	-1.706	0.089	0.014
Height	-0.0438	0.031	-1.417	0.158	0.017
Neck	-0.5266	0.222	-2.371	0.019	-0.089
Chest	-0.0611	0.090	-0.682	0.496	0.115
Abdomen	0.9658	0.087	11.104	0.000	1.137
Hip	-0.3063	0.117	-2.608	0.010	-0.075
Thigh	0.2649	0.145	1.827	0.069	0.551
Knee	-0.1055	0.239	-0.441	0.659	0.365
Ankle	0.1428	0.218	0.654	0.514	0.573
Biceps	0.1654	0.171	0.966	0.335	0.503
Forearm	0.4285	0.198	2.160	0.032	0.819
Wrist	-1.8492	0.535	-3.455	0.001	-0.795
Gender	0.0882	0.607	0.145	0.885	1.284

- **Omnibus:** 2.675
- **Durbin-Watson:** 1.746
- **Prob(Omnibus):** 0.262
- **Jarque-Bera (JB):** 2.003
- **Skew:** 0.013
- **Prob(JB):** 0.367
- **Kurtosis:** 2.561
- **Cond. No.:** 625.

Notes:

1 R^2 is computed without centering (uncentered) since the model does not contain a constant.

2 Standard Errors assume that the covariance matrix of the errors is correctly specified.

- Residual vs Y Predicted plot

5.1 Heteroskedasticity

- Goldfeld-Quandt Test Statistic: 1.1351833916521799
- Goldfeld-Quandt p-value: 0.25301248368804397
- There is no heteroskedasticity (fail to reject the null hypothesis of homoskedasticity).

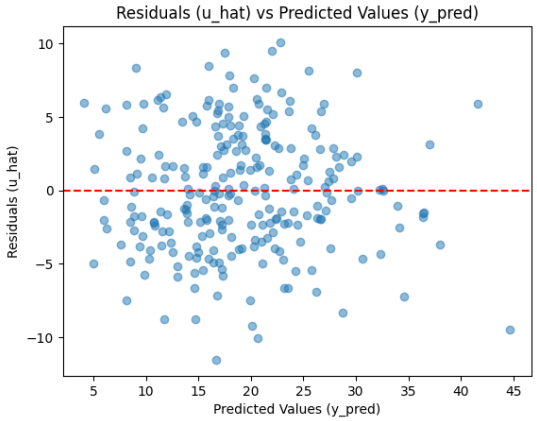

Figure 6: Distribution of Gender.

5.2 Endogeneity

- Correlation between y pred and u hat: 0.002384014315947323 There is no endogeneity (correlation is close to 0).

5.3 Test for normality of Residuals

5.3.1 Q-Q Plot

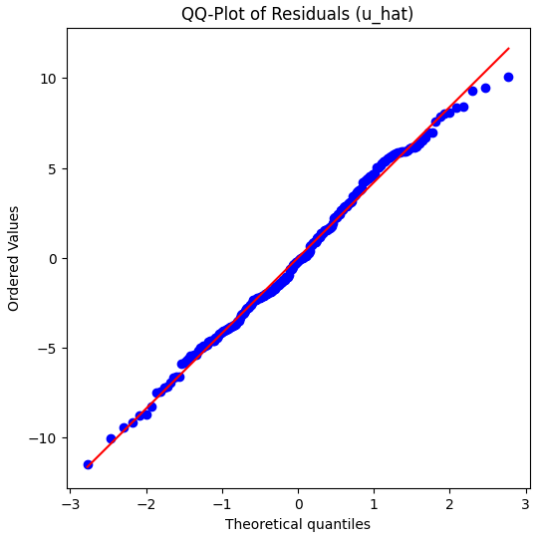

Figure 7: Q-Q plot of the Residuals.

5.3.2 Jarque-Bera Test

- Jarque-Bera test statistic: 2.0031582740579688
- Jarque-Bera p-value: 0.36729896756799313
- Fail to reject the null hypothesis: Residuals are normally distributed.