Comparative Analysis of Statistical Methods for Predicting Body Fat Percentage Using Brozek's Formula

Nazmus Sakib Sumon

May 2024

Abstract

This report explores the application of multiple linear regression methodologies to predict body fat percentage based on Brozek's formula, using seventeen potential predictors. The data set is divided into training and testing subsets to facilitate the evaluation of model performance. Various statistical models, including linear regression, ridge regression, LASSO, Principal Component Regression (PCR), and Partial Least Squares (PLS), are applied and assessed using mean squared error criteria. The effectiveness of each model is further examined through Monte Carlo Cross-Validation to ensure robustness and reliability of the findings.

Introduction

The prevalence of obesity is increasing globally, making the accurate measurement of body fat percentage critically important for medical diagnostics and health monitoring. The Brozek formula for estimating body fat percentage serves as the response variable in this analysis. The objective is to identify the best predictors among seventeen available physiological measurements and compare several statistical modeling approaches in terms of predictive accuracy and reliability.

Data Set Description Exploratory Data Analysis

The dataset comprises measurements from 252 individuals, each described by 18 attributes. The first attribute, 'brozek', represents the body fat percentage calculated using Brozek's formula and serves as the response variable. The remaining seventeen attributes are potential predictors, including physical measurements like density, age, weight, height, and various body circumferences.

Like brozek, the predictor variable 'siri' also displays percentage of body fat. Our scatterplot (figure 1) shows this relationship. We will remove this variable.

Table 1: Description of Dataset Variables

Variable Name	Description
brozek	Percent body fat using Brozek's equation, $\frac{457}{\text{Density}} - 414.2$
siri	Percent body fat using Siri's equation, $\frac{495}{\text{Density}} - 450$
density	Density (gm/cm^3)
age	Age (years)
weight	Weight (lbs)
height	Height (inches)
adipos	Adiposity index = Weight/Height^2 (kg/m^2)
free	Fat Free Weight = (1 - fraction of body fat) * Weight, using Brozek's formula (lbs)
neck	Neck circumference (cm)
chest	Chest circumference (cm)
abdom	Abdomen circumference (cm) at the umbilicus and level with the iliac crest
hip	Hip circumference (cm)
thigh	Thigh circumference (cm)
knee	Knee circumference (cm)
ankle	Ankle circumference (cm)
biceps	Extended biceps circumference (cm)
forearm	Forearm circumference (cm)
wrist	Wrist circumference (cm) distal to the styloid processes

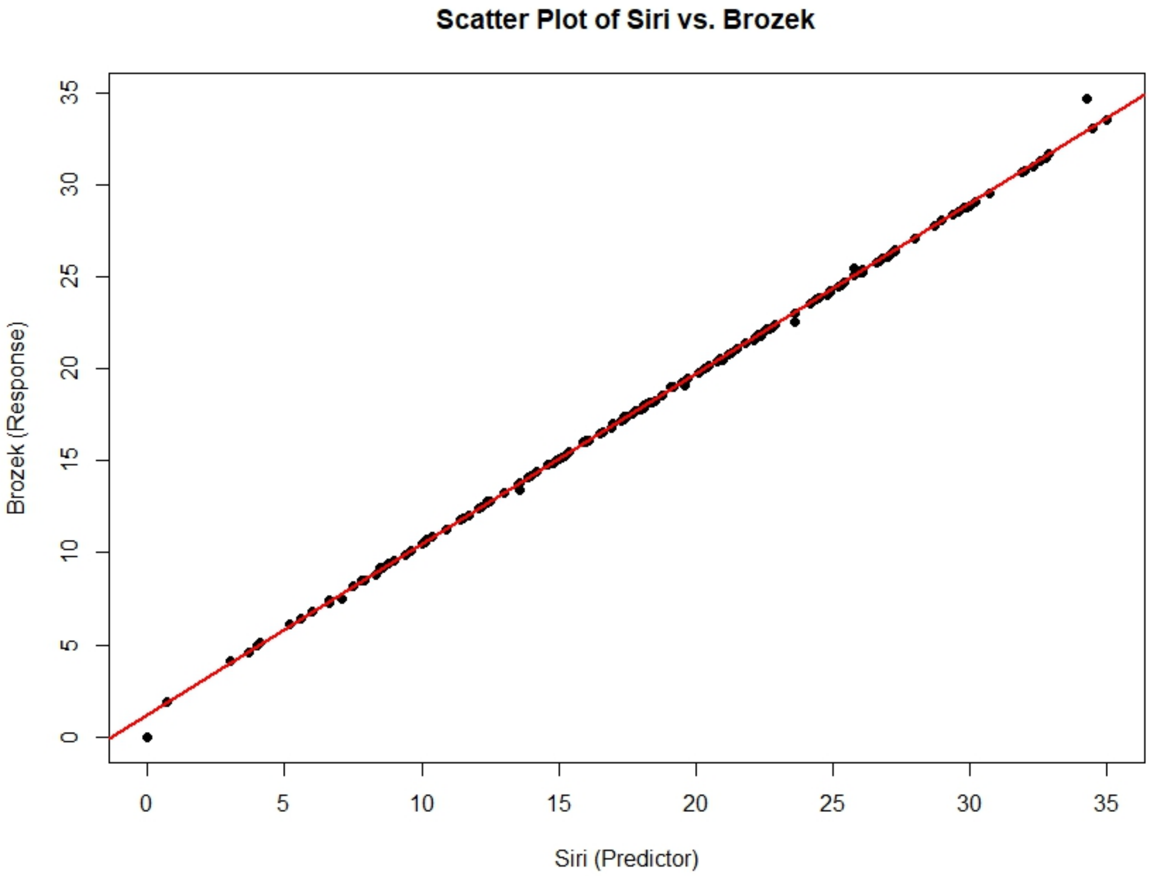

Figure 1: Scatter plot of Siri vs Brozek

We will also remove the “free” variable. Because it is derived from the weight variable, thus making this variable redundant. After removing these variables, the dimension of our training dataset: 227 rows and 16 columns. Now we will look into the box plot of training and testing dataset

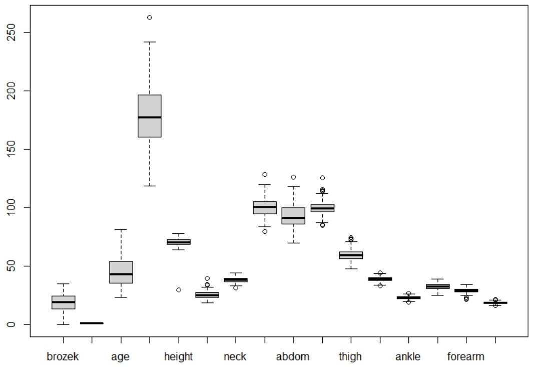

Figure 2: Boxplot of the training dataset

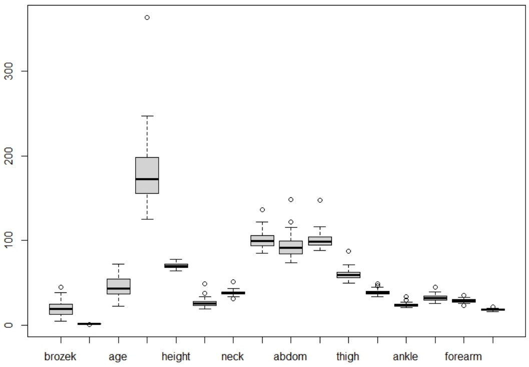

Figure 3: Boxplot of the test dataset

The datasets are overall similar, but in the testing dataset, some variable values have higher values compared to training dataset.

Methodology

The data was randomly split into a training set (70%) and a testing set (30%). The training data was used to develop several models:

1. **Linear regression with all predictors**: From the summary of this model, we see most of the variables’ p-values are not significant, except for density. The adjusted r-squared value is very high.
2. **Linear regression with the best subset of k = 5 predictors**: The best subset of k=5 predictors includes density, age, chest, abdom, and biceps.
3. **Linear regression with variables (stepwise) selected using AIC**: The stepwise selection gives us the same set of variables like the previous best subset of variables. This is an indication that we are getting the correct results.
4. **Ridge Regression**: We applied the optimal lambda value here.
5. **LASSO**: Similar process was followed for lasso regression.
6. **Principal Component Regression**: We chose the first four components to perform this PCR.
7. **Partial Least Squares**: The number of components here is 15. So, partial least squares method also reduces the number of usable variables.

Results

Each model's performance was evaluated based on the mean squared error (MSE) between the predicted and actual body fat percentages in the testing set. The initial findings indicated that while traditional linear regression models performed reasonably well, regularization and dimensionality reduction techniques like Lasso offered improvements in certain contexts, particularly in handling multicollinearity and model complexity. This is reflected also after monte carlo cross-validation.

Table 2: Initial Model Performance Table

Model Name	Testing Error
Linear regression	1.0079436
Linear Regression with k=5 variables	0.8679503
Linear Regression with step using AIC	0.8679503
Ridge Regression	2.3818815
Lasso Regression	0.4151663
PCR	35.1885895
PLS	3.5526131

Table 3: Model Performance Table After Cross-Validation

Model Name	Testing Error
Linear regression	0.7826105
Linear Regression with k=5 variables	0.2749667
Linear Regression with step using AIC	0.2694513
Ridge Regression	1.669678
Lasso Regression	0.3338994
PCR	40.25944
PLS	3.377984

Findings

The comparative analysis highlighted the strengths and weaknesses of each modeling approach. Regularization methods (Ridge, LASSO) and PLS provided the best balance between complexity and performance, indicating their suitability in scenarios with high-dimensional data. The Monte Carlo Cross-Validation reinforced these findings, demonstrating the robustness of the LASSO and PLS models across multiple training-test splits. Below is a summary of the important predictors identified by each model:

Table 4 showcases how each modeling approach selects and weighs different predictors, reflecting their individual methodologies and assumptions.

Future Research

Future studies could explore the integration of additional predictors, such as diet and lifestyle factors, into the regression models. Further, advanced machine learning techniques like ensemble methods and support vector machines could be evaluated to enhance predictive accuracy.

Table 4: Important Predictors Identified by Each Model

Model Name	**Significant Predictors**
Linear regression	Density (strong negative influence)
Best Subset Selection	Density, Age, Chest, Abdomen, Biceps
Stepwise Selection (AIC)	Density, Age, Chest, Abdomen, Biceps (same as Best Subset)
Ridge Regression	Uses all variables but with shrinkage
LASSO	Density, Age, Chest, Abdomen
Principal Component Regression	First four principal components
Partial Least Squares	First 15 components, emphasizing all measurements

Implementing a larger dataset might also provide deeper insights and a more stable performance evaluation.

Further Reading

For more detailed information on statistical methods used in this study, please refer to the faraway package documentation on CRAN. For general information on body fat percentage, see the Wikipedia article: Body Fat Percentage on Wikipedia.