PREDICTION OF BODYFAT USING A LINEAR REGRESSION MODEL AND BODY MEASUREMENTS

Apoorv Saraogee
Northwestern University, Practical Machine Learning
May 15, 2022

Abstract

Bodyfat percentage is an important estimator for health. The most accurate method for measuring bodyfat is by underwater weighing which is time-intensive. Predictive analytics (linear regression) using indirect measurements offer a faster method to compute bodyfat percentage. This study analyzes bodyfat data determined by underwater weighing with their corresponding indirect measurements. To predict the bodyfat using these measurements, we compare a traditional linear model, regularized model, subset models (indicator, dichotomous, piecewise, polynomial), and feature engineering models (principal components analysis). The correlation coefficients (R^2) are analyzed for each ten-fold cross-validated model. The best R^2 value (0.67) is found with a traditional linear model, keeping as much information as possible.

1. Introduction

Obesity, having excess body fat percentage (BFP), is a public health problem and increases risk of diseases such as diabetes and depression. The problem in diagnosis is that the standard method to assess BFP using underwater weighing is costly and requires specialized equipment (Fan et al. 2022). Prediction of bodyfat based on more cheaply obtained anthropometric measurements of different body features or indirect measurements offer an alternative to assessing BFP (Uçar et al. 2021). Accurate prediction of bodyfat can help in diagnosis of obesity and related health problems at a lower cost and prevent serious health problems.

This study's central research topic is to analyze a bodyfat dataset of 252 men measured with underwater weighing using a linear regression model of 13 explanatory variables - age, weight, height, neck, chest, abdomen, hip, thigh, knee, ankle, biceps, forearm, wrist (Penrose, Nelson, and

Fisher 1985). Research questions include whether subset models, regularized models work better than the traditional model (scored using the correlation coefficient R^2). Other questions will explore models with engineered features such as with a principal components analysis. The best model will be chosen based on the correlation coefficients R^2 between the predictions and the bodyfat measurements in the dataset.

2. Literature Review

Predictive equations using anthropometric data has been employed by others on the bodyfat data. The literature includes methods that use non-linear predictions and machine learning methods such as neural networks, support vector machines in a hybrid model with feature selection for the first stage and machine learning for the second stage (Fan et al. 2022), (Shao 2014), (Hussain, Cavus, and Sekeroglu 2021), (Uçar et al. 2021). These researchers all use feature selection to minimize the number of explanatory variables to the most information-rich ones. This investigation also uses 6 principal components as was the best found in (Fan et al. 2022). However, this investigation uses a more simple linear regression model for predictions compared to the more complicated models used in the literature. Linear methods are explored in the textbook (Izenman 2008).

3. Methods

This research will be conducted on the bodyfat data in Jupyter notebook using kernels for both R as well as Python3. The data will be visualized in R and then analyzed with various models using the sklearn package in Python3. Real bodyfat data is used to train the linear regression model and each model is evaluated using their correlation coefficients R^2. The linear assumption of the

model is tested by plotting the predicted values from the model against the real values for a straight line. The key objectives include evaluating and comparing a traditional regression model (linear regression with ten-fold cross validation), a regularized regression model (ridge regression), subset models (polynomial models to the second degree, indicator models for variables Chest and Abdomen with the 25^{th}, 50^{th} and 75^{th} percentile as cutoffs, dichotomous models for Chest and Abdomen with the 50^{th} percentile as a cutoff, piecewise transformation model with the Chest variable with the 50^{th} percentile cutoff) and a principal components model. The model with the best score (correlation coefficient R^2) will be chosen as the most optimal.

3. Results

It can be seen that the 'All variables' model or traditional model had the highest correlation coefficient of 0.67 in Table 1. This score is also shared with the Polynomial variables model, but since it includes 105 features (degree 2), the traditional model is still the best model for this dataset with a fewer 13 features. Regularization using ridge regression showed a slight reduction in the model correlation coefficients. Subset models showed significant reductions in the model correlation coefficients as considerable information is lost when categories are made. This is seen with the dichotomous model having a lower R^2 score of 0.52 compared to the more information rich indicator variable model having a R^2 score of 0.61. This large reduction could be due to the variables chosen for the subset models (Abdomen, Chest) being highly correlated with the bodyfat. This is seen in Figure 4 under the Body_Fat_Percent column with 0.81 and 0.70 for Abdomen and Chest respectively. The piecewise subset model had the worst test score of 0.47 likely because the cutoff used (50^{th} percentile of the Chest variable) didn't generalize well. The loss of generalization is supported by the big difference seen between the train and test scores in the piecewise model.

Feature engineering using principal components with 6 components worked quite well and close to the best model with a correlation coefficient of 0.65. This is quite good as there is a minimal reduction in the model score with half the number of features as the traditional model (6 instead of 13). However, because there are fewer explanatory variables, there is a reduction in the performance. The best model is still the one with most information used for the prediction. From these results, I learned that having more information in the model is better for prediction. This was clearly seen in Figure 1 with the All Possible Regression done in R, where the highest R^2 was found only using a combination of all 13 variables.

4. Conclusions

Within this study, the traditional linear model using all 13 explanatory variables performed the best with a decent correlation coefficient of 0.67. This is quite poor in obtaining accurate predictions for the body fat percentage as you would like to be above 0.8. The feature engineering is promising as it reduces the number of features with minimal predictive power loss, but the model parameters are not quite as interpretable. In the traditional model, the correlation coefficients can tell you about the predictive power of each feature. Subset models while offering simpler solutions often lead to a loss of information as a continuous variable is turned into a categorical variable. The best prediction of body fat using anthropometric or indirect body measurements has a correlation coefficient of 0.67. More information is needed in order to obtain a more accurate model of the bodyfat measurement and diagnose obesity. This should include taking more body measurements to use as explanatory variables for a model.

5. Appendices

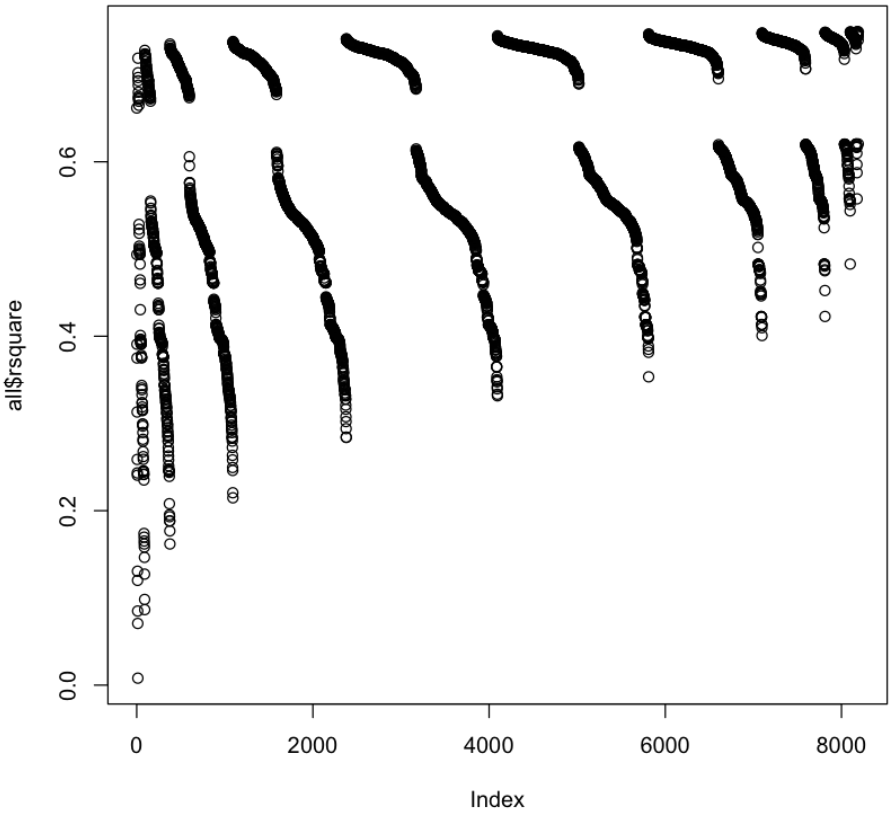

Figure 1. All Possible Regression in R

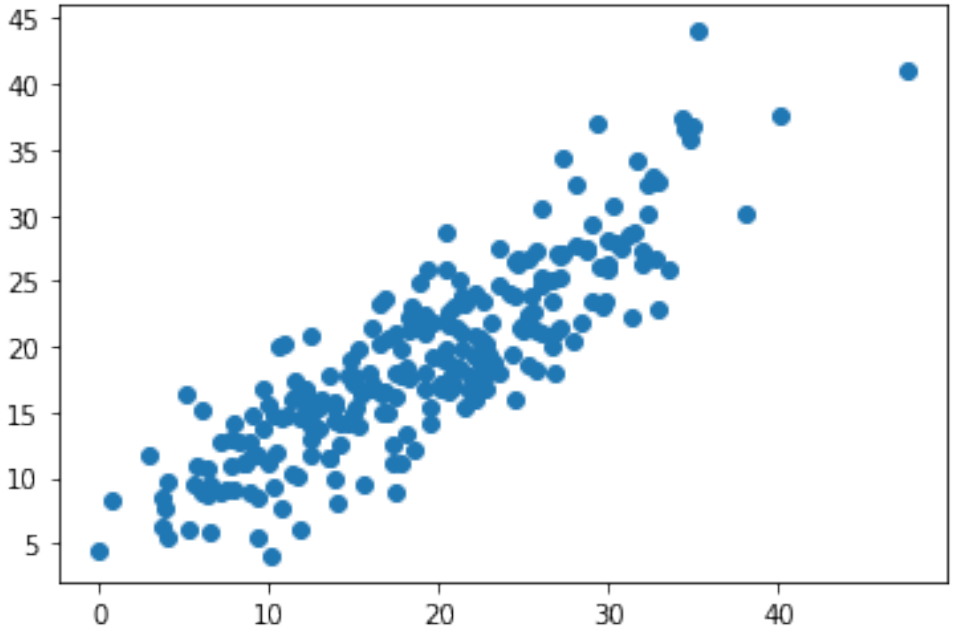

Figure 2. Linear Assumption of Model with real data on x-axis and predictions on y-axis

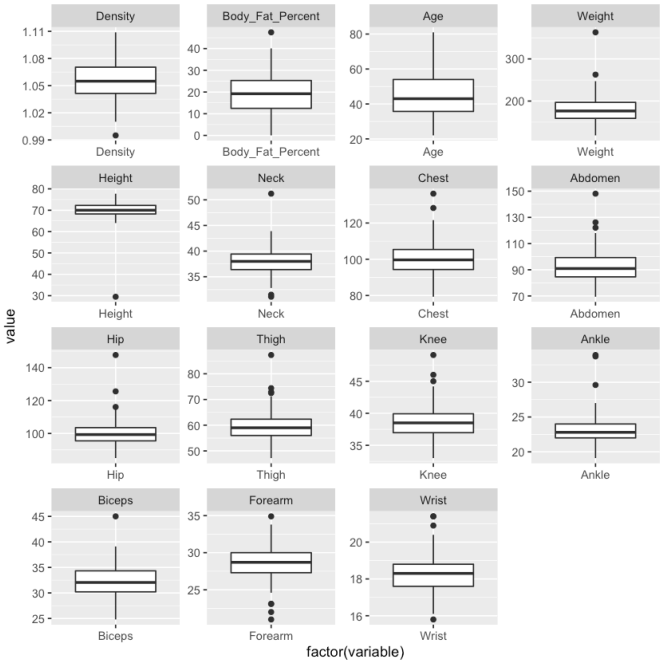

Figure 3. Box plots of explanatory variables of bodyfat data, using R

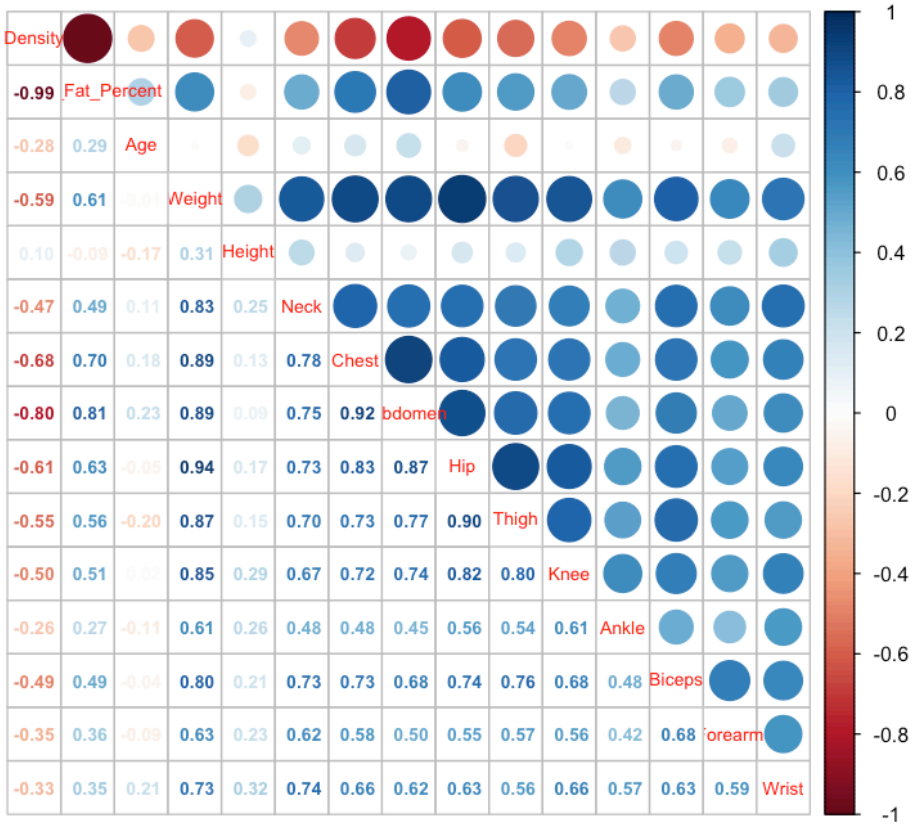

Figure 4. Correlation plot of explanatory variables for bodyfat data, using R

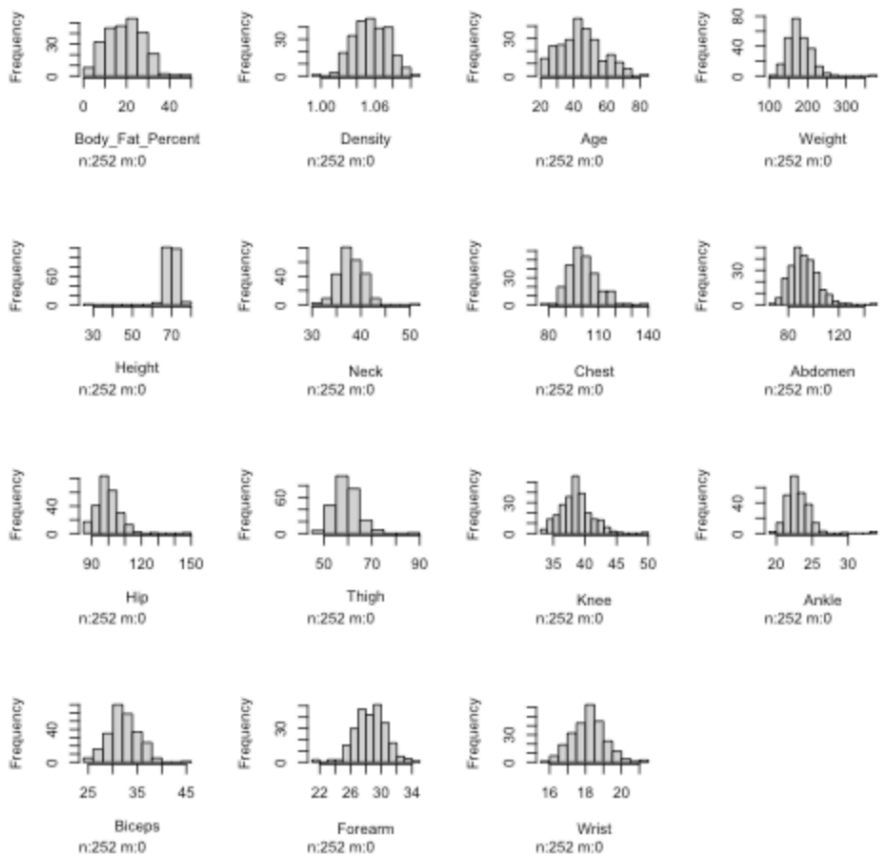

Figure 5. Histograms of explanatory variables, using R

Table 1: Comparison of different cross-validated models with training and testing score averages

model	train r2	test r2
All variables	0.750279	0.673734
Regularized all variables	0.750142	0.673657
Indicator variables	0.710344	0.606238
Dichotomous variables	0.655773	0.523038
Polynomial variables	0.750279	0.673734
Piecewise variables	0.669124	0.475196
Principal Components	0.71999	0.654166

Supporting Files

- Assignment1.html
- Assignment1.py
- Assignment1.ipynb
- assignment1bodyfat.csv

References

Fan, Zongwen, Raymond Chiong, Zhongyi Hu, Farshid Keivanian, and Fabian Chiong. 2022. "Body Fat Prediction through Feature Extraction Based on Anthropometric and Laboratory Measurements." Edited by Maciej Huk. *PLOS ONE* 17 (2): e0263333. https://doi.org/10.1371/journal.pone.0263333.

Hussain, Solaf A., Nadire Cavus, and Boran Sekeroglu. 2021. "Hybrid Machine Learning Model for Body Fat Percentage Prediction Based on Support Vector Regression and Emotional Artificial Neural Networks." *Applied Sciences 2021, Vol. 11, Page 9797* 11 (21): 9797. https://doi.org/10.3390/APP11219797.

Izenman, Alan J. 2008. "Modern Multivariate Statistical Techniques," Springer Texts in Statistics, . https://doi.org/10.1007/978-0-387-78189-1.

Penrose, Keith W, Arnold G Nelson, and Arnold Garth Fisher. 1985. "Generalized Body Composition Prediction Equations for Men Using Simple Measurement Techniques." *Medicine and Science in Sports and Exercise* 17: 189. https://doi.org/https://doi.org/10.1249/00005768-198504000-00037.

Shao, Yuehjen E. 2014. "Body Fat Percentage Prediction Using Intelligent Hybrid Approaches." *The Scientific World Journal* 2014. https://doi.org/10.1155/2014/383910.

Uçar, Muhammed Kürşad, Zeliha Uçar, Fatih Köksal, and Nihat Daldal. 2021. "Estimation of Body Fat Percentage Using Hybrid Machine Learning Algorithms." *Measurement* 167 (January): 108173. https://doi.org/10.1016/J.MEASUREMENT.2020.108173.