The Accuracy of Body Fat Percentage Estimation with Body Composition Measurements

Brandon R. Russell

University of Maryland University College

Abstract

I would like to study the Fitting Body Fat % dataset for my research project. This dataset provides underwater weighing density, age, weight, height, and many different body measurements for over 252 men. Determining a person’s density with underwater weighing is an accurate method of determining body fat. For this project, I will examine the variations between different transactions to determine if there are any associations with measurements of the body and given body fat percentages and Adiposity index as an alternative method for accurately predicting these values. The measurement variables I will be examining are neck, chest, abdomen, hip, thigh, knee, ankle, biceps, forearm, and wrist circumference. I will start the project with the assumption that age is a completely independent variable and does not have any association with body fat percentage or the Adiposity index. Additionally, I will focus my initial analysis on the notion that the abdomen circumference will have the highest correlation with body fat percentage.

Keywords: Body, fat, percentage, composition, measurements

Contents

[Abstract 2](#_Toc516407929)

[The Accuracy of Body Fat Percentage 4](#_Toc516407930)

[Prior Research 4](#_Toc516407931)

[Research Methodology 4](#_Toc516407932)

[Data 4](#_Toc516407933)

[Variables 4](#_Toc516407934)

[Methods 4](#_Toc516407935)

[Regression something. 4](#_Toc516407936)

[Decision Tree something. 4](#_Toc516407937)

[Classification something. 4](#_Toc516407938)

[Experiments and Results Analysis 4](#_Toc516407939)

[Conclusion 4](#_Toc516407940)

[References 5](#_Toc516407941)

The Accuracy of Body Fat Percentage Estimation with Body Composition Measurements

This paper analyzes the Fitting Body Fat % dataset, comprised of 15 body composition variables of 252 men, to determine which data mining algorithms are effective in detemermining a coorleation or statistically probably pattern between body measurements and body fat percentage, as compared with preceise standard underwater weighing methods. Through the conduction of applicable mining methods, we will assert the accuracy of determining body fat percentage based upon measurements of various body components.

# Prior Research

The study of determining an individual’s body fat percentage has been a well-documented study over the past 70 years. A range of well-documented and precise methods have been created in this time to determine an individual’s accurate body fat percentage. For example, two-component body models have been created utilizing underwater hydrostatic weighing to determine density which is then added to a formula, either Siri or Brozek, to determine a percentage. Another accurate method is a four-component model utilizing a dual energy X-ray absorptiometry (DXA). These methods are highly accurate, but not practical for the average person to accomplish. For this reason, different studies have sought to determine if different algorithms can be utilized to predict a person’s body fat percentage, within a reasonable level of accuracy, based on measurements of various body components (Johnson, Navarro, Idiong, & Weeks).

One study, conducted by (Johnson, Navarro, Idiong, & Weeks), utilizes three different data mining algorithms to determine the accuracy of body measurements in predicting body fat percentage. The first algorithm they studied was a linear regression model. Based on their previous research, they decided to focus this method on waist and wrist circumference measurements. After applying linear regression on waist only, they discovered an 86% accuracy. When they applied both the waist and wrist coefficients accuracy improved by 1%. The second method they used was a M5R rule-based classifier. Utilizing this method, the algorithm determined 10 different parameters were needed in two separate rule sets: age, weight, height, neck, waist, hip, thigh, ankle, forearm, and wrist to achieve an 85% accuracy level. The last method they applied to the dataset was K-means clustering. With two clusters created, as the data was split in the previous rule-based method, the distinction of the weight variable between the two sets was highlighted, validating the discovery of two different rules in the previous method (Johnson, Navarro, Idiong, & Weeks).

Another similar study on predicting body fat percentage by body measurements was conducted by (Lean, Han, & Deurenberg, 1996). They applied a stepwise multiple regression model over a dataset consisting of 147 people, a mix of women and men. Measurements taken included: age, height, weight, BMI, waist, hip, thigh, MUAC, waist-hip ratio, lower leg length, arm span, triceps-skinfold thickness, density, ∑4Skinfolds, and body fat percentage. Applying a stepwise regression model to this dataset determined the best variables for predicting men’s density, and subsequently body fat percentage, was a combination of waist circumference with triceps-skinfold and age giving approximately 86% probability. Waist measurement alone for men provided an accuracy level of 77% (Lean, Han, & Deurenberg, 1996).

# Research Methodology

## Data

Description. The Fitting Body Fat % dataset used by (Johnson R. W., 1996) in his research contains 252 instances of 19 attributes of data points regarding body composition for men. The first attribute is a case number, some form of row identification. The next two attributes contain the Siri and Brozek body fat percentage calculated values. These attributes are determined by having the participants undergo underwater weighing and displacement which determines density. Density, the fourth attribute, is used with different formulas to calculate a percentage. The formula for Siri is (%BF = [4.950 / BD (kg/m3) – 4.500] x 100) and Brozek is (% BF = [4.570 / BD (kg/m3) – 4.142] x 100). Both the Siri and Brozek method BF% variables have shown a strong correlation with the highly accurate four-component DXA method of determining BF%. The Brozek method had a 1.7% closer relationship with DXA than Siri did (Guerra, Amaral, Marques, Mota, & Restivo, 2010). Attributes five through nine detail age, weight, height, Adiposity index, and fat-free weight. The adiposity index is calculated as ((hip circumference)/((height)1.5)–18) (Bergman, et al., 2011). Fat-free weight is calculated as the remaining percentage not covered by BF%. The remaining 10 attributes cover the circumference for neck, chest, abdomen, hip, thigh, knee, ankle, extended biceps, forearm, and wrist.

Preparation. Within this dataset there are no missing values we will need to rectify. The first attribute, the case number, serves our analysis no purpose and will be removed. Additionally, for the sake of simplicity we will only be comparing measurements to one of the BF% values. Since Brozek was noted earlier as being more closely accurate to DXA we will remove the Siri attribute and keep Brozek. Since the density attribute is a part of the Brozek formula it will be removed to avoid any false correlations. Additionally, because the Adiposity index and fat-free weight attributes are derived values from other attributes in the dataset they will also be removed, leaving us with 14 attributes overall (one independent and 13 dependent). All the values are continuous, meaning no categorial data. As will be noted in a later section, one of our selected algorithms will require categorical data so we will need to discretize some of the attributes for that specific method. Table 1 below shows the attributes of the 14 variables we will be using, and Figure 1 shows their distributions. Looking at Table 1, the only attribute with concerning numbers is the height variable with a minimum value of 29.5. This was likely a manual input error and we will rectify this anomaly by replacing this value with the mean value of that column. The Percent.body.fat.using.Brozek attribute is our independent variable. The minimum value is 0, and since it is unlikely any human has zero percent body fat we will change this value to the mean value of the column to avoid any skewing of our learning phases. Age and Weight have the greatest standard deviation, showing we have a wide variety of diverse data for an inclusive result. The Neck, Chest, Abdomen, Hip, Thigh, Knee, Ankle, Extended Bicep, Forearm, and Wrist circumference measurements will all be examined individually and combined in optimal pairs to determine their ability to accurately predict the Brozek body fat measurement.

Table 1. Numeric Attributes

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Attribute** | **Mean** | **Std Deviation** | **Median** | **Min** | **Max** |
| Percent.body.fat.using.Brozek | 18.93849 | 7.750855659 | 19 | 0 | 45.1 |
| Age | 44.88492 | 12.60203972 | 43 | 22 | 81 |
| Weight | 178.9244 | 29.38915989 | 176.5 | 118.5 | 363.15 |
| Height | 70.14881 | 3.662855788 | 70 | 29.5 | 77.75 |
| Neck.circumference | 37.99206 | 2.430913234 | 38 | 31.1 | 51.2 |
| Chest.circumference | 100.8242 | 8.430475532 | 99.65 | 79.3 | 136.2 |
| Abdomen.circumference | 92.55595 | 10.7830768 | 90.95 | 69.4 | 148.1 |
| Hip.circumference | 99.90476 | 7.164057667 | 99.3 | 85 | 147.7 |
| Thigh.circumference | 59.40595 | 5.249952028 | 59 | 47.2 | 87.3 |
| Knee.circumference | 38.59048 | 2.411804587 | 38.5 | 33 | 49.1 |
| Ankle.circumference | 23.10238 | 1.694893398 | 22.8 | 19.1 | 33.9 |
| Extended.biceps.circumference | 32.27341 | 3.021273751 | 32.05 | 24.8 | 45 |
| Forearm.circumference | 28.66389 | 2.020691165 | 28.7 | 21 | 34.9 |
| Wrist.circumference | 18.22976 | 0.933584929 | 18.3 | 15.8 | 21.4 |

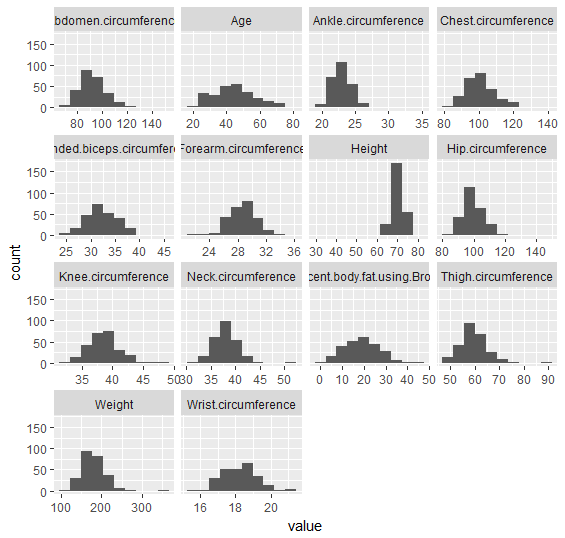


Figure 1. Body Fat Dataset attribute distribution

In the next section we will cover the different algorithm methods we will use to analyze the body fat data set, a multiple linear regression model, naïve bayes classification, and a recursive partitioning decision tree (rpart).

## Methods

### Model Selection. To determine the optimal algorithms to run on the body fat dataset, I did a literature review to see what algorithms other papers had implemented and what success they had, which can be noted in the prior research section above. I then implemented these models in R Studio and performed a visual analysis to determine which ones would provide the most interesting results. In total, I considered Apriori, Neural Networks, Support Vector Machines, Recursive Partitioning Decision Trees, Naïve Bayes Classification, and Multiple Linear Regression. I narrowed my focus to the latter three to remain within a reasonable scope of this paper, and because they provided the most interesting results when conducting my initial analysis.

### Multiple Linear Regression. With Simple Linear Regression we are attempting to apply a statistical model on two variables, with value Y as our dependent variable and value X as the independent variable. With this notion we are attempting to find a linear correlation between the two so that we can accurately predict a future value of Y based on the value of X. Based on this knowledge, we can define multiple linear regression as a higher level linear regression model that takes multiple independent attributes and combines them to further accurately predict Y. The assumptions with regression are that Y is independent, follows a normal distribution, the mean of that distribution is a linear function of each x, and Y has a constant variance. The process of conducting linear regression is to start by verifying a linear relation for each predictor, then estimate the model, assess if the model is an appropriate fit, draw inferences about the coefficients, remove insignificant predictors, then reassess the appropriateness of the model (Eberly, 2007).

### Recursive Partitioning Decision Tree (rpart). Decision trees are a greedy type of algorithm which takes a top-down approach of constructing a tree, making recursive decisions on how to best split and categorize data. The basic elements of the decision tree are internal nodes which holds a test, a branch which is the outcome of the test, and terminal nodes which have a class label. They are popular because of their accuracy, ability to handle multidimensional data, and they are generally easy for humans to understand (Han, Kamber, & Pei, 2011). Rpart is a form of decision tree, like the Classification and Regression Trees (CART). It uses a two-stage procedure and is represented as a binary tree. The splitting decision is a critical component of a decision tree and is responsible for measuring the impurity of data to determine the best pace to split to achieve a reduction in heterogeneity. Equation 1 represents how we can calculate impurity, where A is the node being measured and *f* is the chosen impurity function, for example the Gini index. Rpart also offers ways to prune the tree for unneeded data and cross-validation methods (Therneau, Atkinson, & others, 2018).

(1)

### Naïve Bayes Classification. These are statistical classifiers that predict whether data belongs to a class within a certain probability. They have class conditional independence, meaning the effect of one attribute value on a class is independent of values of other attributes. The formulas we are attempting to solve here would be the value X as our data, *H* as the hypothesis that X belongs to a class C, giving us P(*H*|X). Know this we can calculate the maximum posteriori hypothesis with Bayes’ theorem see in Equation 2. If data is missing, there is a process, called Laplacian correction, where 1 is added to the attribute to ensure this does not have a negative impact on the outcome of the algorithm results (Han, Kamber, & Pei, 2011).

(2)

# Experiments and Results Analysis

### Multiple Linear Regression.

### Recursive Partitioning Decision Tree (rpart).

### Naïve Bayes Classification.

# Conclusion

…

References

**There are no sources in the current document.**