The Accuracy of Body Fat Percentage Estimation with Body Composition Measurements

Brandon R. Russell

University of Maryland University College

Abstract

In this paper we will study the Fitting Body Fat % dataset. 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. In this project, we will examine the variations in data points to determine if there are any associations with measurements of the body and given body fat percentages as an alternative method for accurately predicting these values. The measurement variables of interest for examination are age, weight, and height with neck, chest, abdomen, hip, thigh, knee, ankle, biceps, forearm, and wrist circumferences. The different algorithms we will explore includes multiple linear regression, rpart decision trees, and Naïve Bayes classification. After a through analysis of these methods, along with a comparison of previous studies, we will attempt to make a confident decision in determining the best model for accurately predicting bodyfat percentage.

Keywords: Body, fat, percentage, composition, measurements

Contents

[Abstract 2](#_Toc518581088)

[The Accuracy of Body Fat Percentage 5](#_Toc518581089)

[Prior Research 5](#_Toc518581090)

[Research Methodology 6](#_Toc518581091)

[Data 6](#_Toc518581092)

[Description. 6](#_Toc518581093)

[Preparation. 7](#_Toc518581094)

[Methods 9](#_Toc518581095)

[Model Selection 9](#_Toc518581096)

[Multiple Linear Regression.. 10](#_Toc518581097)

[Recursive Partitioning Decision Tree (rpart). 10](#_Toc518581098)

[Naïve Bayes Classification.. 11](#_Toc518581099)

[Experiments and Results Analysis 11](#_Toc518581100)

[Multiple Linear Regression 12](#_Toc518581101)

[Recursive Partitioning Decision Tree (rpart) 19](#_Toc518581102)

[Naïve Bayes Classification 30](#_Toc518581103)

[Conclusion 35](#_Toc518581104)

[Model Comparison 35](#_Toc518581105)

[Prior Research Comparison 35](#_Toc518581106)

[Final Model Selection 35](#_Toc518581107)

[Lesson Learned 35](#_Toc518581108)

[Final Thoughts 35](#_Toc518581109)

[References 36](#_Toc518581110)

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. The goal is to determine which data mining algorithms are effective in detemermining a coorleation or statistically probable pattern between body measurements and body fat percentage, as compared with preceise standard underwater weighing methods. We will first look at previous research conducted in this field, and also what steps need to be taken to prepare the data for proper analysis. We will start with a multiple linear regression model, then rpart decision trees, and finally the Naïve Bayes algorithm. We will then use specific unique measurements to determine the best model from each section, followed by a comparision of cross-model measurement points to determine the best overall model for use. 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. After further examination of the row showing 0% BF, it was noted the fat free weight was equal to the weight value and the density value yielded a negative value. Because of the multiple erroneous entries in this row we will remove it to prevent 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 |

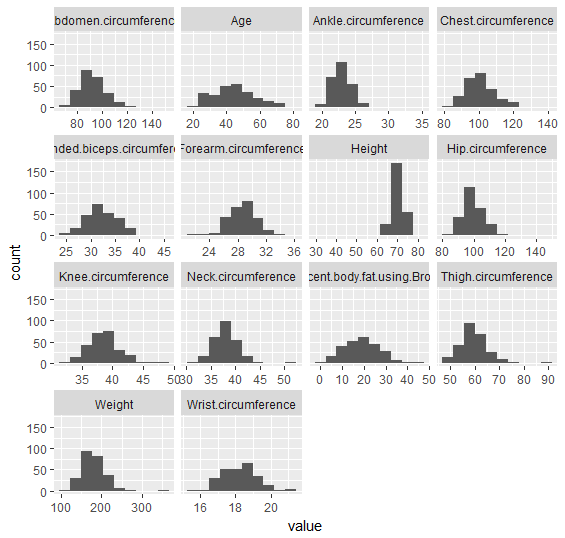


Figure 1. Body Fat Dataset attribute distribution

Before proceeding an examination of boxplots for each variable is applicable to determine if there are any other potential erroneous outlier data points, see figure 2.

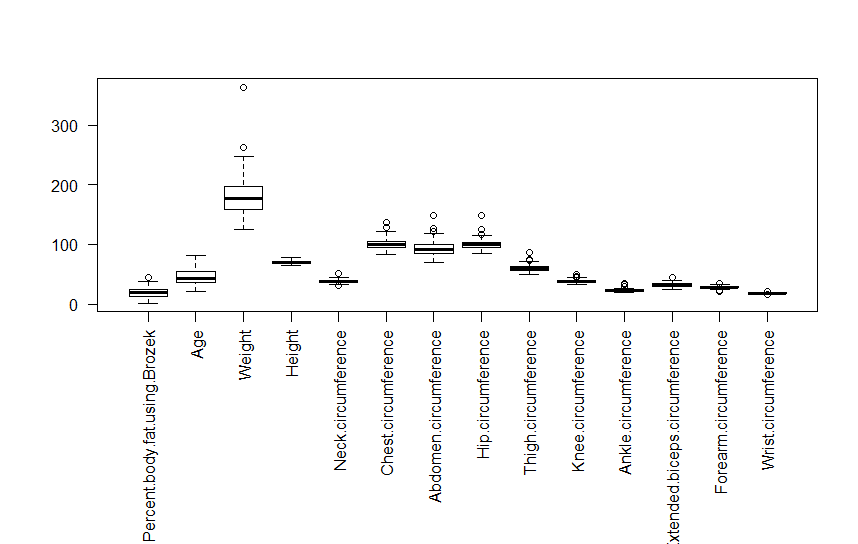


Figure 2. Boxplot analysis of bodyfat dataset.

After noting the most extreme potential outliers in the Weight, Chest, Abdomen, and Hip circumference columns, further analysis determined these all belonged to the same row of data indicating it is most likely normal data and will remain in the dataset for analysis.

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 focused on 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. The formula for linear regression, equation 1, gives us Y is equal to the intercept, a, plus the slope, b, times X.

(1)

Based on this knowledge, wedefine multiple linear regression as a higher level linear regression model that combines multiple independent attributesto 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 2 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).

(2)

### 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 3. 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).

(3)

# Experiments and Results Analysis

In the following section we will conduct several experiments within each previously described model to determine which provides the best predictive accuracy. In each section we will first assess the optimal variable selection and then apply this algorithm to cross validation procedures to proof our model accuracy and then apply the algorithm to our test data.

## Multiple Linear Regression

For linear regression we first need to look at all 13 independent variables compared to our dependent variable, the Brozek BF%, to determine which produces a linear model of interest. Before we start the analysis, the Bodyfat dataset will be split into a 70% train data set and 30% test data set. We will work with the train data set exclusively in the first portion of this experiment and will only use the test data set at the end to make predictions of the dependent variable. In figure 3, we first look at the Residuals vs Fitted, Normal Q-Q plot, Scale-Location, and Residuals vs Leverage plots. The first plot, Residuals vs Fitted, does not show any pattern, meaning it is suitable for linear regression. The Normal Q-Q plot follows a straight line, letting us know our residuals are normally distributed. Due to the horizontal line in the third plot we know the residuals are equally spread out across the independent variable range of values. The last plot show that rows 82 and 109 are above Cook’s line, however a manual examination of these rows does not indicate any odd values which might cause issues with the analysis (Kim, 2015).

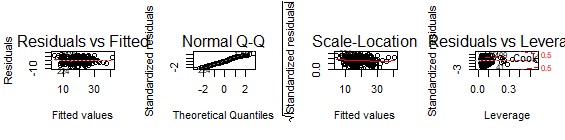


Figure 3. Different plots indicating ability to successfully produce a linear model.

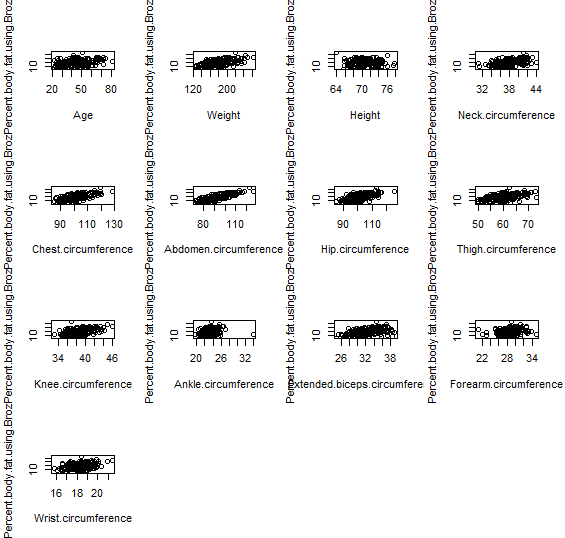


Figure 4. Linear model of Brozek BF% dependent variable compared to independent variables.

After examining the model plots in figure 3, we can move on to plotting relations between each independent attribute to the dependent variable to inspect for a linear relationship that justifies further analysis. After examining figure 4, we can make note that age and weight has a linear appearance. Additionally, neck, chest, abdomen, hip, and thigh circumferences have an interesting pattern worth investigating further. Upon producing the linear model in R and running the summary command we generate our initial coefficients and other significant statistics, see table 2. The first note of interest is the significance of the chest, abdomen, and wrist circumferences per their respective p values. Using this information, along with our earlier noted attributes of interest from the visual inspection will allow us to further refine our model to enhance to potentially enhance the overall accuracy.

Table 2. Summary command output of first linear model with all 13 independent attributes.

|  |
| --- |
| lm(formula = Percent.body.fat.using.Brozek ~ ., data = train.data) |
| Residuals: |
| Min 1Q Median 3Q Max |
| -9.6410 -2.7397 -0.2496 2.6990 8.6795 |
| Coefficients: |
| Estimate Std. Error t value Pr(>|t|) |
| (Intercept) 14.19019 27.41490 0.518 0.6054 |
| Age 0.05236 0.03534 1.481 0.1404 |
| Weight 0.01296 0.07954 0.163 0.8708 |
| Height -0.26239 0.21039 -1.247 0.2141 |
| Neck.circumference -0.38083 0.25377 -1.501 0.1353 |
| Chest.circumference -0.20290 0.12167 -1.668 0.0973 . |
| Abdomen.circumference 0.89585 0.10127 8.846 1.29e-15 \*\*\* |
| Hip.circumference -0.21019 0.16061 -1.309 0.1924 |
| Thigh.circumference 0.23160 0.16336 1.418 0.1581 |
| Knee.circumference -0.11579 0.25872 -0.448 0.6551 |
| Ankle.circumference -0.06281 0.26840 -0.234 0.8153 |
| Extended.biceps.circumference 0.09901 0.17636 0.561 0.5753 |
| Forearm.circumference 0.22254 0.20067 1.109 0.2690 |
| Wrist.circumference -1.42096 0.56559 -2.512 0.0129 \* |
| --- |
| Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |
| Residual standard error: 3.884 on 166 degrees of freedom |
| Multiple R-squared: 0.7586, Adjusted R-squared: 0.7397 |
| F-statistic: 40.13 on 13 and 166 DF, p-value: < 2.2e-16 |

In table 2, we should also make note of the other significant terms of measuring our model, to include the R-squared value of 0.7586, Adjusted R-squared value of 0.7397, Residual standard error of 3.884 and F-statistic of 40.13. After creating our second model with the reduced attributes, we find interesting results, seen in table 3. We see that the abdomen, thigh and wrist are the only attributes showing significance in this model.

Table 3. Summary command output of modified linear model with 8 independent attributes.

|  |
| --- |
| lm(formula = Percent.body.fat.using.Brozek ~ Age + Weight + Neck.circumference +  Chest.circumference + Abdomen.circumference + Hip.circumference +  Thigh.circumference + Wrist.circumference, data = train.data) |
| Residuals: |
| Min 1Q Median 3Q Max |
| -9.3740 -2.6337 -0.3196 2.6661 9.4745 |
| Coefficients: |
| Estimate Std. Error t value Pr(>|t|) |
| (Intercept) -14.09041 13.75351 -1.024 0.3070 |
| Age 0.04569 0.03298 1.385 0.1677 |
| Weight -0.05457 0.04934 -1.106 0.2703 |
| Neck.circumference -0.24954 0.24486 -1.019 0.3096 |
| Chest.circumference -0.10733 0.10777 -0.996 0.3207 |
| Abdomen.circumference 0.92709 0.09268 10.003 <2e-16 \*\*\* |
| Hip.circumference -0.20281 0.15447 -1.313 0.1910 |
| Thigh.circumference 0.32832 0.14118 2.326 0.0212 \* |
| Wrist.circumference -1.31987 0.54120 -2.439 0.0158 \* |
| --- |
| Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |
|  |
| Residual standard error: 3.884 on 171 degrees of freedom |
| Multiple R-squared: 0.7514, Adjusted R-squared: 0.7398 |
| F-statistic: 64.6 on 8 and 171 DF, p-value: < 2.2e-16 |

Still examining table 3, we see the residual standard error did not change from the first model, but the R-squared value decreased to 0.7514. Although R-squared decreased slightly, the Adjusted R-squared increased slightly. The F-statistic increased a measurable amount. Based on these results we will reduce our variable selection once more to only the significant attributes from the second model, namely the abdomen, thigh, and wrist circumferences. The results of the third iteration of our model can be seen in table 4. Although the F-statistic increased substantially in the third model, due to the significance of the included attributes, the residual standard error increased and both the R-squared and Adjusted R-squared values decreased.

Table 4. Summary command output of modified linear model with 3 independent attributes.

|  |
| --- |
| lm(formula = Percent.body.fat.using.Brozek ~ Abdomen.circumference +  Thigh.circumference + Wrist.circumference, data = train.data) |
| Residuals: |
| Min 1Q Median 3Q Max |
| -8.8622 -2.9158 -0.6771 3.0700 8.6055 |
| Coefficients: |
| Estimate Std. Error t value Pr(>|t|) |
| (Intercept) -7.54412 6.14819 -1.227 0.221 |
| Abdomen.circumference 0.77458 0.04673 16.574 < 2e-16 \*\*\* |
| Thigh.circumference -0.08625 0.09043 -0.954 0.342 |
| Wrist.circumference -2.20554 0.41504 -5.314 3.21e-07 \*\*\* |
| --- |
| Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |
| Residual standard error: 4.028 on 176 degrees of freedom |
| Multiple R-squared: 0.7248, Adjusted R-squared: 0.7201 |
| F-statistic: 154.5 on 3 and 176 DF, p-value: < 2.2e-16 |

After manually determining our variable selection we will not run a backwards step-wise regression to see if it can perform any better. We can see in table 5, our residual standard error is virtually the same as our second model, with a slightly higher Adjusted R-squared, and much higher F-statistic.

Table 5. Backwards step-wise regression model creation.

|  |
| --- |
| lm(formula = Percent.body.fat.using.Brozek ~ Height + Chest.circumference + |
| Abdomen.circumference + Wrist.circumference, data = train.data) |
| Residuals: |
| Min 1Q Median 3Q Max |
| -8.9772 -2.7684 -0.3005 2.7384 7.9712 |
| Coefficients: |
| Estimate Std. Error t value Pr(>|t|) |
| (Intercept) 12.94890 8.47188 1.528 0.12820 |
| Height -0.38730 0.12257 -3.160 0.00186 \*\* |
| Chest.circumference -0.21157 0.09315 -2.271 0.02435 \* |
| Abdomen.circumference 0.87462 0.06891 12.692 < 2e-16 \*\*\* |
| Wrist.circumference -1.44891 0.44669 -3.244 0.00141 \*\* |
| --- |
| Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |
| Residual standard error: 3.877 on 175 degrees of freedom |
| Multiple R-squared: 0.7465, Adjusted R-squared: 0.7407 |
| F-statistic: 128.8 on 4 and 175 DF, p-value: < 2.2e-16 |

After examining all four models we can deduce the step-wise model is the preferred model with the highest Adjusted R-square value of 0.7407, a Residual standard error of 3.887, and F-statistic of 128.8. The next logical step is to conduct cross validation of this model to ensure we have not overfit our model. Table 6 holds a summary of the results of a 10-fold cross validation on the second model. Table 6 shows our variables hold true with cross validation, indicating our model is ready to conduct predictions on our test data.

Table 6. 10-fold cross-validation results on the step-wise linear model.

|  |
| --- |
| lm(formula = .outcome ~ ., data = dat) |
| Residuals: |
| Min 1Q Median 3Q Max |
| -8.9772 -2.7684 -0.3005 2.7384 7.9712 |
| Coefficients: |
| Estimate Std. Error t value Pr(>|t|) |
| (Intercept) 12.94890 8.47188 1.528 0.12820 |
| Height -0.38730 0.12257 -3.160 0.00186 \*\* |
| Chest.circumference -0.21157 0.09315 -2.271 0.02435 \* |
| Abdomen.circumference 0.87462 0.06891 12.692 < 2e-16 \*\*\* |
| Wrist.circumference -1.44891 0.44669 -3.244 0.00141 \*\* |
| --- |
| Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |
| Residual standard error: 3.877 on 175 degrees of freedom |
| Multiple R-squared: 0.7465, |
| F-statistic: 128.8 on 4 and 175 DF, p-value: < 2.2e-16 |

After conducting a prediction on our test data set with the modified second model, we find our cross-model accuracy and error rates of our actual and predicted values in table 7, with table 8 showing the first ten rows of the predicted test data.

Table 7. Cross-model accuracy and error check for linear model.

|  |  |
| --- | --- |
| Test | Value |
| Correlation Accuracy | 0.81 |
| Min/Max Accuracy | 0.82 |
| Root Mean Square Error | 4.84 |
| Mean Absolute Error | 3.70 |
| Mean Absolute Percentage Error | 0.25 |

Table 8. Actual vs Predicted values with linear regression model

|  |  |  |
| --- | --- | --- |
| Row # | Actual | Predicted |
| 5 | 27.8 | 26.584 |
| 14 | 20.8 | 25.12584 |
| 16 | 20.5 | 23.09838 |
| 26 | 4.6 | 10.36149 |
| 28 | 22.4 | 19.14627 |
| 29 | 4.7 | 7.802323 |
| 36 | 38.2 | 36.9911 |
| 39 | 33.8 | 54.67484 |
| 40 | 31.3 | 30.74263 |
| 50 | 5.0 | 6.682976 |

In figure 5, we see a plot of our actual versus predicted values, visualizing their accuracy.

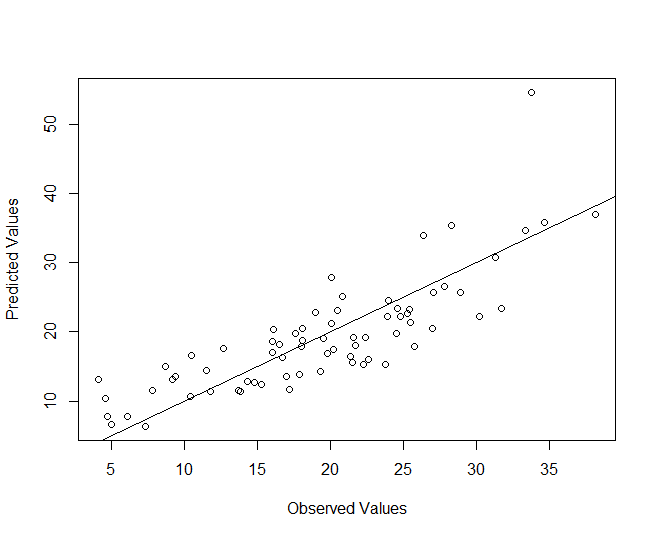


Figure 5. Observed/Predicted values of cross-validated step-wise linear model predictions on test data.

After an exhaustive study of applying multiple linear regression to the BodyFat data set, we can conclude that a modified linear model to predict the Brozek BF% with the independent variables: Height and Chest, Abdomen and Wrist circumference will provide the best accuracy. In the next section we will apply a decision tree algorithm on the Bodyfat data set to determine its effectiveness.

## Recursive Partitioning Decision Tree (rpart)

For this algorithm we will be employing the same data preprocessing we used in our linear regression model. The only exception, we will discretize our dependent variable, the Brozek BF%, into 6 equal frequencies for building a classification, vice regression, tree. Figure 6 shows the levels we created for classification. We will also split our data in the same 70%/30% training/test split. Just as we did with our linear regression model, we will first apply all 13 independent variables to our decision tree formula to see the effectiveness. For our first model we will employ the rpart function from the caret package. The resulting tree, see figure 7, is quite large and difficult to read. It is making use of the Abdomen, Thigh, Hip, Forearm, Ankle, and Chest circumferences and Height to make splitting decisions.

****

Figure 6. Frequency bins created for classification tree.

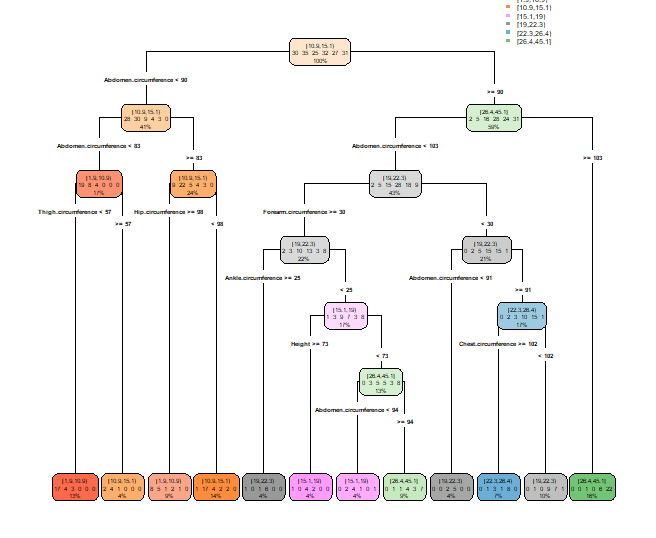


Figure 7. rpart decision tree for BF% training dataset.

Figure 8 and 9 show the confusion matrix for the model in figure 7 for training and test data. The accuracy measurements were 0.6167 and 0.3662 respectively, and the Kappa statistics were 0.5372 and 0.2416 respectively.

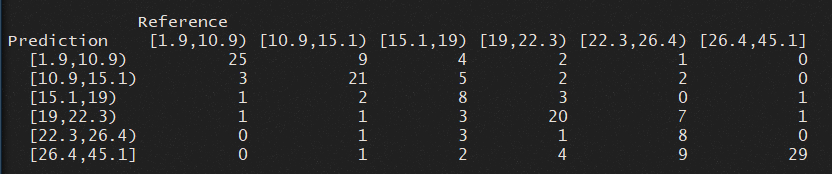


Figure 8. Confusion matrix for training set in first rpart model.

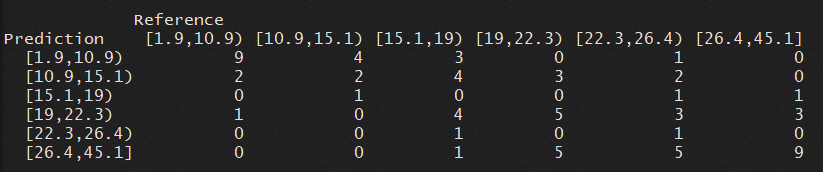


Figure 9. Confusion matrix for test set in first rpart model.

For the next test we decided to utilize the caret package version of rpart because it has cross-validation built-in, and autotunes to the best complexity parameter negating the need to prune the tree. For the initial formula we are using a 10-fold repeated cross-validation procedure with a tuned length of 50. Figure 10 highlights the iteration of our decision tree with all independent variables specified in the formula.

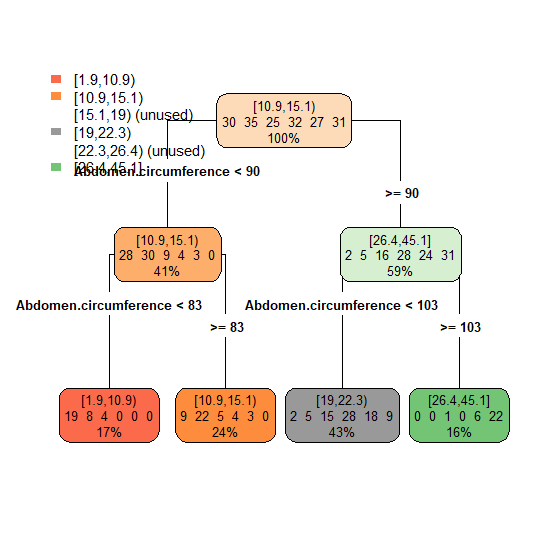


Figure 10. Cross-validated rpart decision tree with caret package and all variables in formula.

Here we can see the abdomen circumference is the only variable being used to decide on multiple levels. Figure 11 shows the corresponding confusion matrix with an overall accuracy of 0.5056 and kappa of 0.3978 on our training dataset .

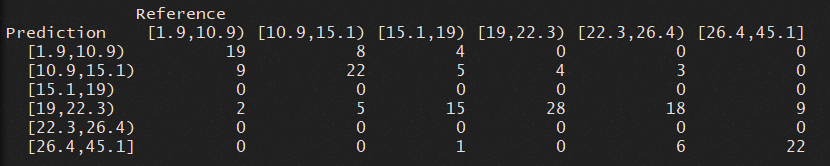


Figure 11. Confusion matrix for training set with all 13 independent variables specified.

When applied to our test dataset we get an accuracy of 0.4225 and kappa of 0.3115 with a confusion table seen in figure 12.

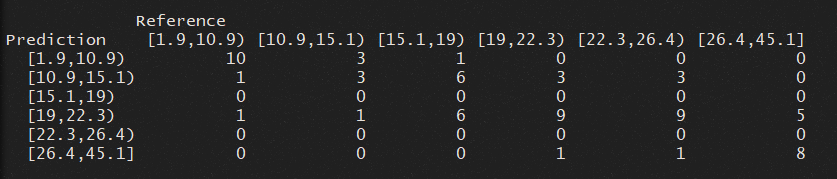


Figure 12. Confusion matrix for test set with all 13 independent variables specified.

To attempt to improve the accuracy of the model we will change the formula to specify the attributes discovered in the pervious linear regression section determined to be the best predictors: Height, along with Chest, Abdomen and Wrist circumference. We will also alter the numbers of the cross-validation rpart train function, changing the tune length to 100, remain at 10 folds, but add 5 repeats. We will also alter the datasets, changing the binning frequency from 6 to 8 to improve our classification granularity. Figure 13 shows the new frequency bins.



Figure 13. Adjusted frequency bins from 6 to 8.

Figure 14 shows our second progression of our rpart decision tree. In this model, even though we increased the bins from 6 to 8, the training set only classified data into 4 of 8, like the previous model which classified data into 4 of 6.

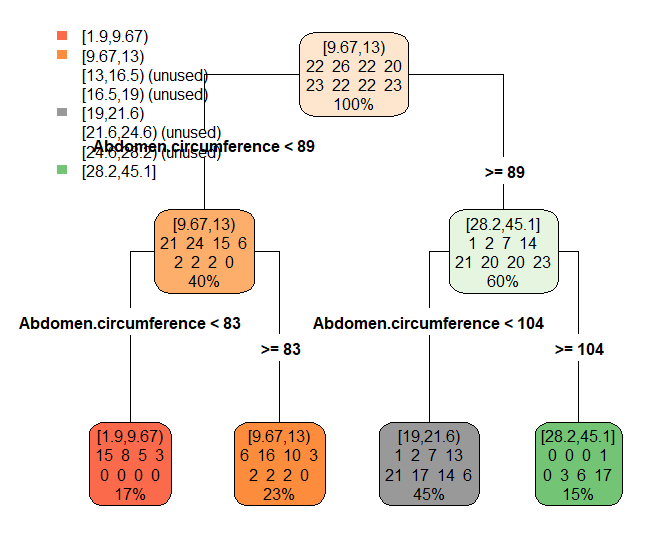


Figure 14. Second iteration of cross-validated rpart decision tree model.

Figure 15 shows the output of our training data set confusion matrix.

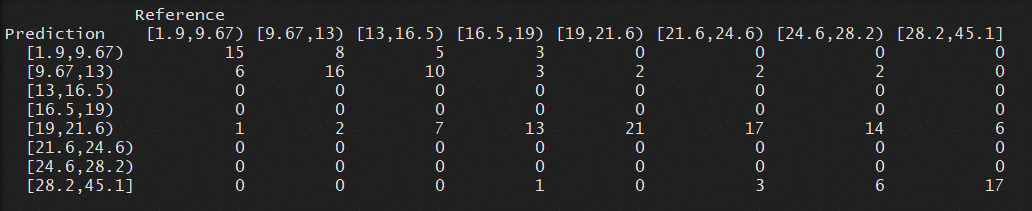


Figure 15. Adjusted model training set confusion matrix output.

The accuracy dropped to .3833 and the kappa to 0.2907. Figure 16 shows the confusion matrix for the modified model applied to the test data set.

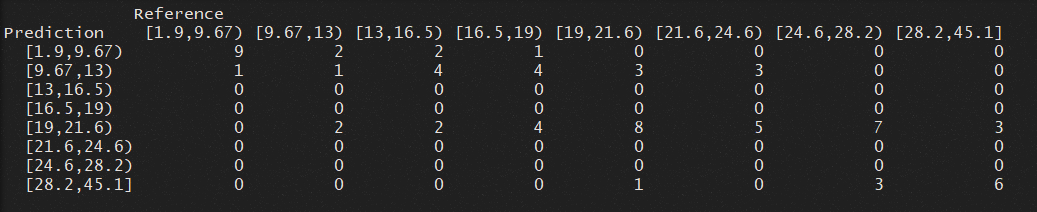


Figure 16. Adjusted model test set confusion matrix output.

The accuracy of the test set was .338 and kappa 0.2345. As we can see, our modified version performed worse across the board. In the next model we will reduce our frequency binning to 4 and reapply all 13 attributes to see if this improves accuracy.

Figure 17 shows the reduced frequency binning used in the next model.



Figure 17. Reduced frequency binning from 8 to 4.

Figure 18 shows the modified iteration of our rpart decision tree.

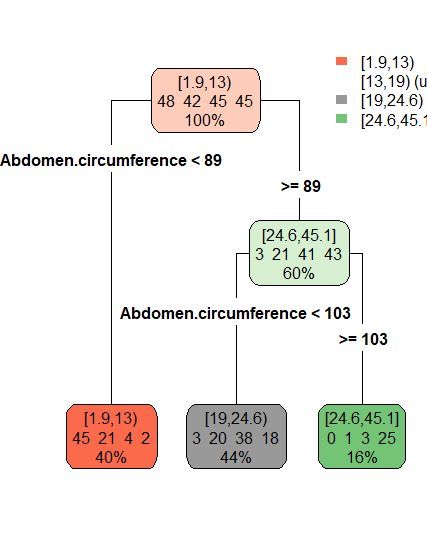


Figure 18. Modified iteration of rpart decision tree.

As can be seen, we are still only using the abdomen circumference measurement as the decision split criteria and only using 3 of the 4 frequency bins. Figure 19 shows the confusion matrix for the training set, with an accuracy of 0.6 and kappa of 0.4619.

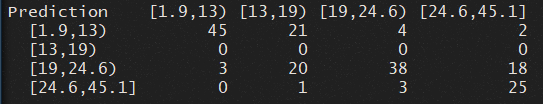


Figure 19. Third model training data confusion matrix.

Figure 20 shows the confusion matrix for the test data set with an accuracy of 0.493 and kappa of 0.324.

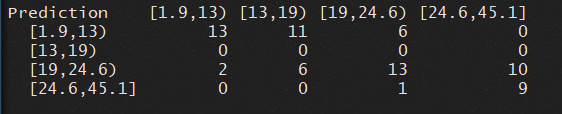


Figure 20. Third model test data confusion matrix.

Because the previous model is only using 3 of the bins we will make a final adjustment to this number of bins to see the effect on accuracy, see figure 21. Figure 22 shows the new decision tree with all 3 bins being used.



Figure 21. Adjusted dependent variable to 3 bins.

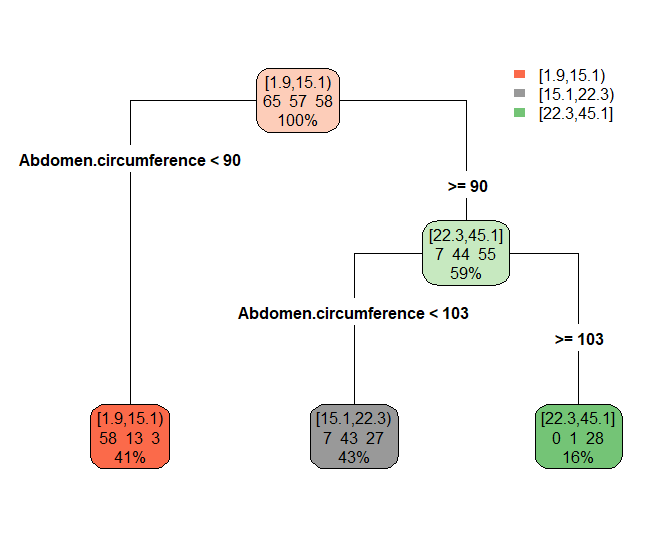


Figure 22. rpart cross-validated tree with 3 bins.

The resulting confusion matrix for the training set can be seen in figure 23 with an overall accuracy of 0.7167 and kappa of 0.5734.

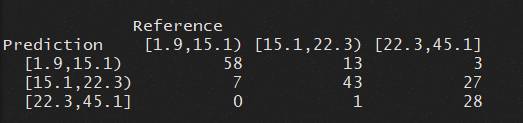


Figure 23. 3-bin model training set confusion matrix.

The confusion matrix for the test dataset can be seen in figure 24, with an overall accuracy of 0.5775 and kappa of 0.3744.

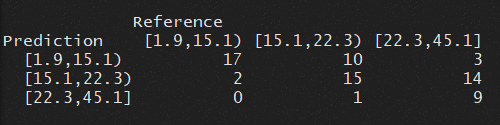


Figure 24. 3-bin model test set confusion matrix.

Although this last model has the lowest level of granularity, with 3 frequency bins, it does provide the best overall accuracy of 58% with the test data set out of all 5 models. Table 9 shows the number of true positive/negatives (TP/TN) and false positive/negatives (FP/FN) on the test set confusion matrix in figure 24, while table 10 shows the overall statistics.

Table 9. Number of true positive/negatives (TP/TN) and false positive/negatives (FP/FN) in test set.

|  |  |  |  |
| --- | --- | --- | --- |
| Type | [1.9, 15.1) | [15.1, 22.3) | [22.3, 45.1] |
| TP | 17 | 15 | 9 |
| Type 1 FP | 13 | 16 | 1 |
| TN | 39 | 29 | 44 |
| Type 2 FN | 2 | 11 | 17 |

Table 10. Overall statistics for final rpart cross-validated 3-bin model on test set.

|  |
| --- |
| Overall Statistics |
| Accuracy : 0.5775 |
| 95% CI : (0.4544, 0.6939) |
| No Information Rate : 0.3662 |
| P-Value [Acc > NIR] : 0.0002362 |
| Kappa : 0.3744 |
| Mcnemar's Test P-Value : 0.0002054 |
| Statistics by Class: |
| Class: [1.9,15.1) Class: [15.1,22.3) Class: [22.3,45.1] |
| Sensitivity 0.8947 0.5769 0.3462 |
| Specificity 0.7500 0.6444 0.9778 |
| Pos Pred Value 0.5667 0.4839 0.9000 |
| Neg Pred Value 0.9512 0.7250 0.7213 |
| Prevalence 0.2676 0.3662 0.3662 |
| Detection Rate 0.2394 0.2113 0.1268 |
| Detection Prevalence 0.4225 0.4366 0.1408 |
| Balanced Accuracy 0.8224 0.6107 0.6620 |

After conducting a prediction on our test data set with the modified second model, we find our cross-model accuracy and error rates of our actual and predicted values in table 11, with table 12 showing the first ten rows of the predicted test data.

Table 11. Cross-model accuracy and error check for linear model.

|  |  |
| --- | --- |
| Test | Value |
| Correlation Accuracy | 0.64 |
| Min/Max Accuracy | .82 |
| Root Mean Square Error | 0.74 |
| Mean Absolute Error | 0.46 |
| Mean Absolute Percentage Error | 0.20 |

Table 12. Actual vs Predicted test data set for rpart decision tree.

|  |  |  |
| --- | --- | --- |
| Row # | Actual | Predicted |
| 1 | 3 | 2 |
| 2 | 2 | 2 |
| 3 | 2 | 2 |
| 4 | 1 | 1 |
| 5 | 3 | 1 |
| 6 | 1 | 1 |
| 7 | 3 | 3 |
| 8 | 3 | 3 |
| 9 | 3 | 3 |
| 10 | 1 | 1 |

Figure 25 shows the ROC curve for each bin, with a mean Area Under Curve (AUC) of 0.76. After an exhaustive study of applying a recursive partitioning decision tree to the BodyFat data set, we can conclude that a cross-validated 10-fold 5-time repeat rpart algorithm with a tuned length of 100 will produce the best model using only the abdomen circumference attribute.

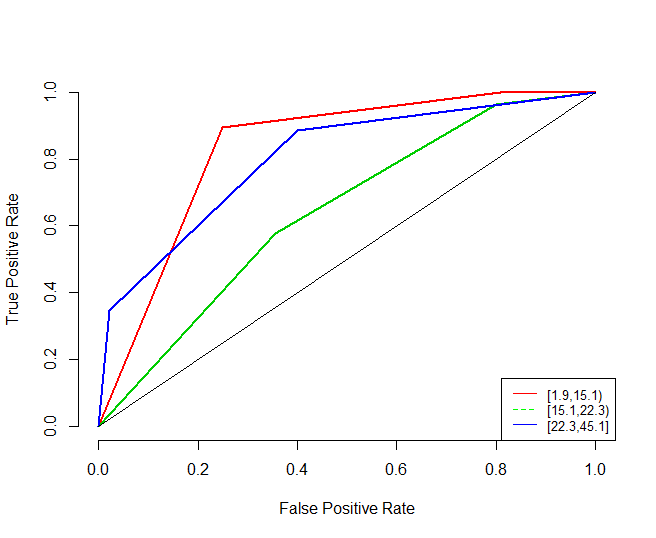


Figure 25. ROC curve plot for 3-bin cross-validated rpart decision tree model on test set data.

In the next section we will apply a Naïve Bayes algorithm on the Bodyfat data set to determine its effectiveness.

## Naïve Bayes Classification

For this algorithm we will be employing the same data preprocessing we used in our linear regression model. The only exception, we will discretize all the attributes, *to* convert our numerical data to categorical data. We will start by using all 13 independent variables in our formula and based on lessons learned from our previous section we will retain our binning at 3 for all attributes. Figure 26 shows the output of the model, a set of conditional probabilities for each independent attribute for the dependent attribute.

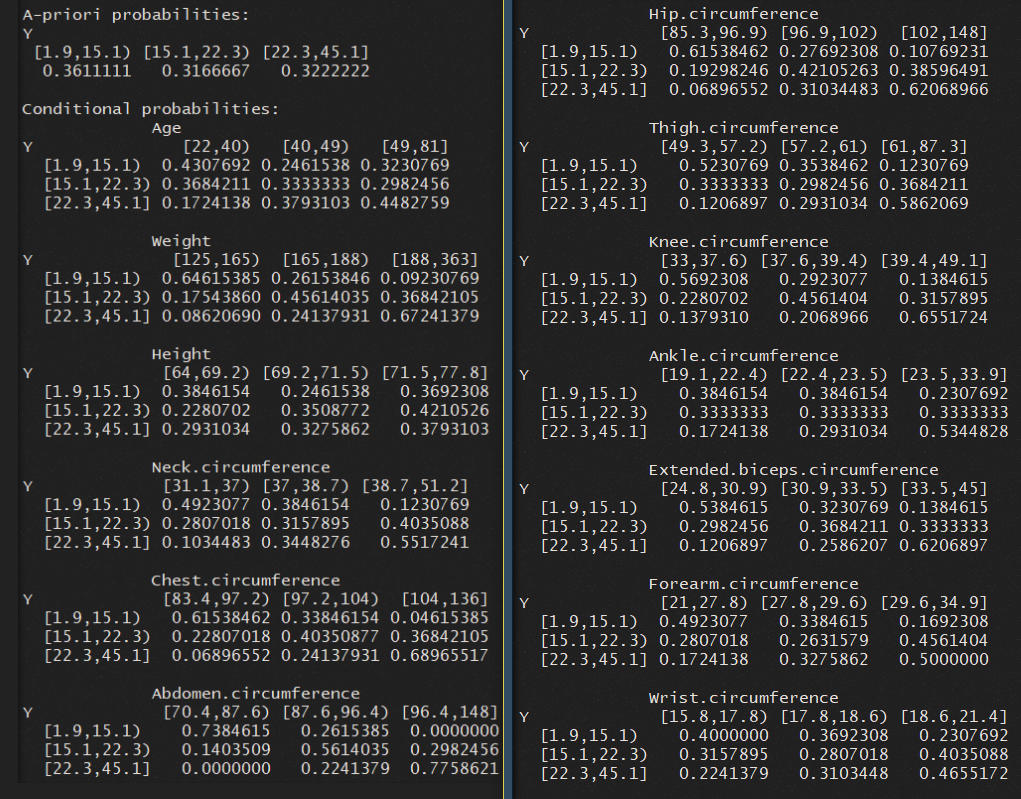


Figure 26. Conditional probabilities for each independent to dependent variable.

Figure 27 shows our prediction with this model on the training data with a prediction accuracy of 0.6222 and a kappa of 0.4312. This is followed by the prediction on the test data in figure 28 with a prediction accuracy of 0.5493 and kappa of 0.3395.

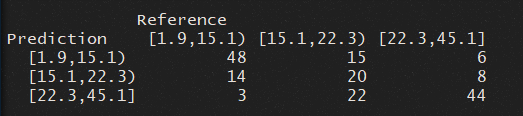


Figure 27. Naïve Bayes training set confusion matrix.

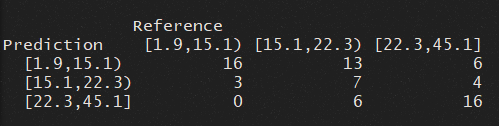


Figure 28. Naïve Bayes test set confusion matrix.

For the next iteration of our model we make use of the same train function we saw with our decision tree, which will provide us the ability to cross-validate our model in a single step. We will use the same parameters from before as well, a tune length of 100, repeated cross-validation, 10-fold, and with 5 repeats. Figure 29 shows the model output, still using all 13 attributes.

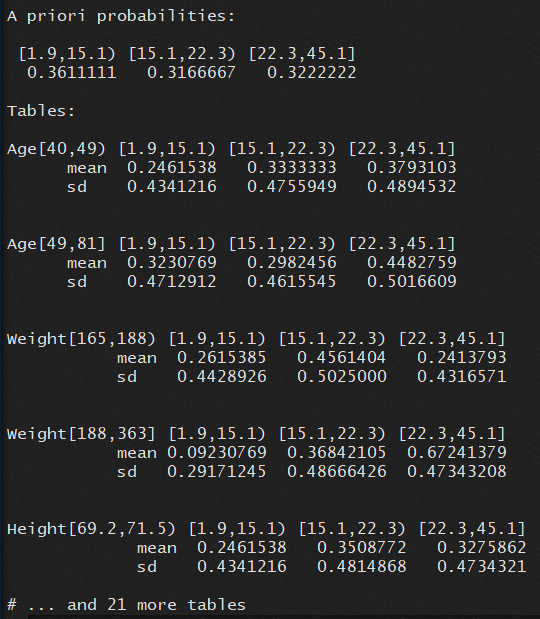


Figure 29. Cross-validated Naïve Bayes model output.

Figures 30 and 31 show the confusion matrix for this model with the training and test sets with a prediction accuracy of 0.5944 and 0.5211 respectively and kappa of 0.382 and 0.3109 respectively.

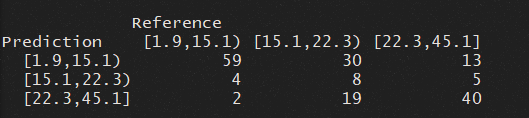


Figure 30. Cross-validated model training set confusion matrix.

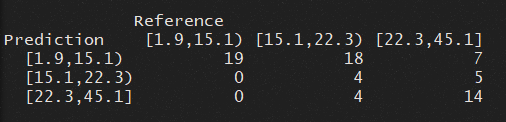


Figure 31. Cross-validated model training set confusion matrix.

We should make note that the overall accuracy did drop. Because this model has been cross-validated we can believe this to be a more accurate prediction. To try and improve the accuracy of this model we will reduce our attributes to those found in the previous section to have the most significance: Height, along with the Chest, Abdomen, and Wrist circumferences. Figure 32 shows the modified model output.

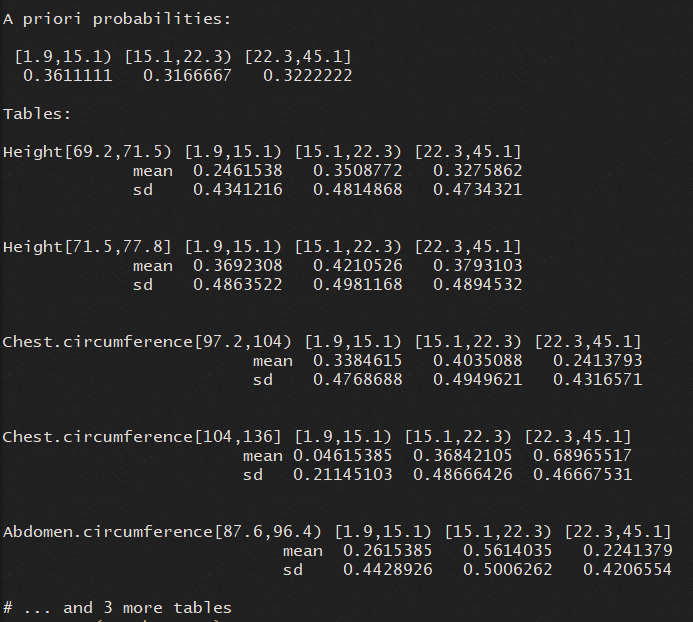


Figure 32. Modified Naïve Bayes model output.

Figures 33 and 34 show the confusion matrix for the training and test set. We see an accuracy of 0.6444 and kappa of 0.4576 for the training set and 0.5493 and 0.3527 for the test set.

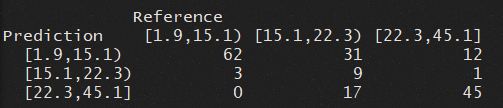


Figure 33. Modified Naïve Bayes training set confusion matrix.

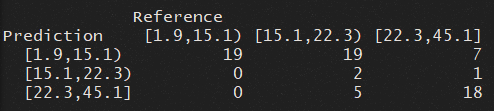


Figure 34. Modified Naïve Bayes test set confusion matrix.

This model does amount a small increase in accuracy. We will create one last model to see if we can increase the accuracy anymore by only discretizing the dependent variable, but not the other 13 independent variables. Figure 35 shows the model output.

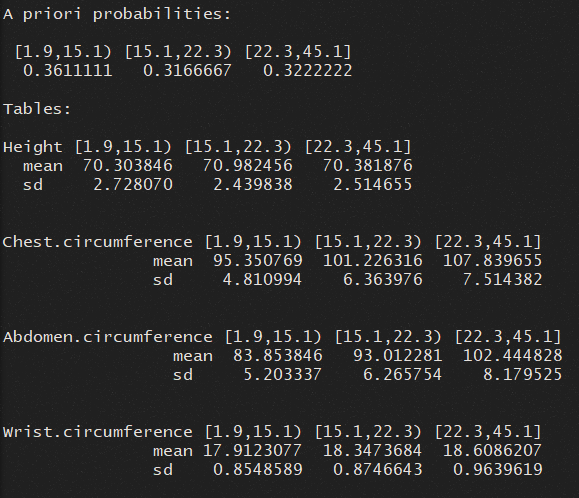


Figure 35. Modified Naïve Bayes output.

Figures 36 and 37 show the confusion matrix for both the training and test sets. The training set has an accuracy of 0.6611 and a kappa of 0.4904. The test set had an accuracy of 0.5634 and a kappa of 0.3602. This marks a clear increase in overall accuracy and is our model of choice.

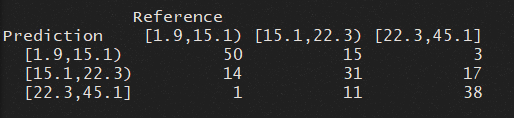


Figure 36. Modified model training set confusion matrix.

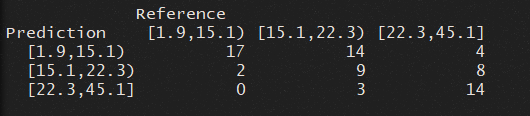


Figure 37. Modified model test set confusion matrix.

Table 13 shows our True Positive, False Positive, True Negative, False Negative counts, and table 14 has the overall statistics for our final model.

Table 13. Number of true positive/negatives (TP/TN) and false positive/negatives (FP/FN) in test set.

|  |  |  |  |
| --- | --- | --- | --- |
| Type | [1.9, 15.1) | [15.1, 22.3) | [22.3, 45.1] |
| TP | 17 | 9 | 14 |
| Type 1 FP | 18 | 10 | 3 |
| TN | 34 | 35 | 42 |
| Type 2 FN | 2 | 17 | 12 |

Table 14. Overall statistics for final Naïve Bayes cross-validated 3-bin model on test set.

|  |
| --- |
| Overall Statistics |
| Accuracy : 0.5634 |
| 95% CI : (0.4405, 0.6809) |
| No Information Rate : 0.3662 |
| P-Value [Acc > NIR] : 0.0005614 |
| Kappa : 0.3602 |
| Mcnemar's Test P-Value : 0.0015978 |
| Statistics by Class: |
| Class: [1.9,15.1) Class: [15.1,22.3) Class: [22.3,45.1] |
| Sensitivity 0.8947 0.3462 0.5385 |
| Specificity 0.6538 0.7778 0.9333 |
| Pos Pred Value 0.4857 0.4737 0.8235 |
| Neg Pred Value 0.9444 0.6731 0.7778 |
| Prevalence 0.2676 0.3662 0.3662 |
| Detection Rate 0.2394 0.1268 0.1972 |
| Detection Prevalence 0.4930 0.2676 0.2394 |
| Balanced Accuracy 0.7743 0.5620 0.7359 |

After conducting a prediction on our test data set with the final model, we find our cross-model accuracy and error rates of our actual and predicted values in table 15, with table 16 showing the first ten rows of the predicted test data.

Table 15. Cross-model accuracy and error check for linear model.

|  |  |
| --- | --- |
| Test | Value |
| Correlation Accuracy | 0.63 |
| Min/Max Accuracy | 0.80 |
| Root Mean Square Error | 0.78 |
| Mean Absolute Error | 0.49 |
| Mean Absolute Percentage Error | 0.22 |

Table 16. Actual vs Predicted test data set for rpart decision tree.

|  |  |  |
| --- | --- | --- |
| Row # | Actual | Predicted |
| 1 | 3 | 2 |
| 2 | 2 | 3 |
| 3 | 2 | 1 |
| 4 | 1 | 1 |
| 5 | 3 | 1 |
| 6 | 1 | 1 |
| 7 | 3 | 3 |
| 8 | 3 | 3 |
| 9 | 3 | 3 |
| 10 | 1 | 1 |

Figure 38 shows the ROC curve for each bin, with a mean Area Under Curve (AUC) of 0.83. After an exhaustive study of applying a Naïve Bayes method to the BodyFat data set, we can conclude that a cross-validated 10-fold 5-time repeat Naïve Bayes algorithm with a tuned length of 100 will produce the best model using the Height, along with the Chest, Abdomen, and Wrist circumferences when we discretize the dependent variables into 3 bins, but leave the independent variables in their original state.

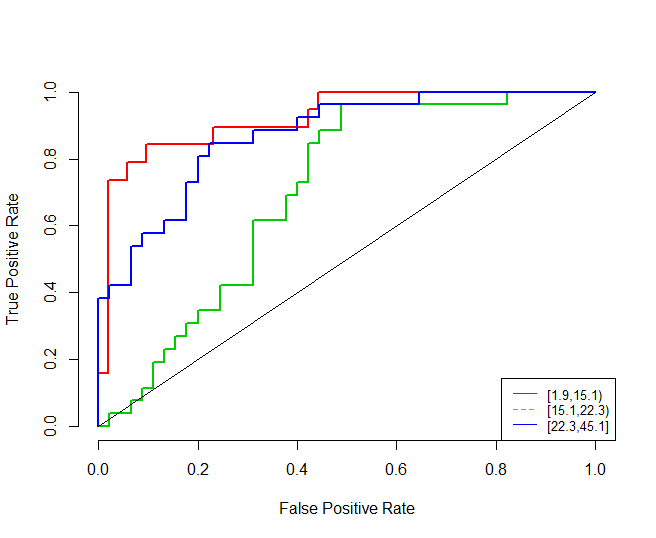


Figure 38. ROC curve plot for 3-bin cross-validated Naïve Bayes model on test set data.

In the next section we will do a comparison of the models presented here and of the models in our prior research, determine a final best model, and conclude with lessons learned and final thoughts.

# Conclusion

## Model Comparison

In this paper we conduced four different multiple linear regression model test, five rpart decision tree model test, and four Naïve Bayes model test. Through an extensive analysis involving variable selection, variable modification, and parameter tuning we determined the best overall model of each section based on specific criteria to each model and common criterion across all three models. In this section we will compare the best three models from each section to one another to determine the best model in terms of the chosen criterion we are judging on. Examining tables 17 through 19 we see the results of the best models from each section.

Table 17. Best linear model: Backwards step-wise regression model creation.

|  |
| --- |
| lm(formula = Percent.body.fat.using.Brozek ~ Height + Chest.circumference + |
| Abdomen.circumference + Wrist.circumference, data = train.data) |
| Residuals: |
| Min 1Q Median 3Q Max |
| -8.9772 -2.7684 -0.3005 2.7384 7.9712 |
| Coefficients: |
| Estimate Std. Error t value Pr(>|t|) |
| (Intercept) 12.94890 8.47188 1.528 0.12820 |
| Height -0.38730 0.12257 -3.160 0.00186 \*\* |
| Chest.circumference -0.21157 0.09315 -2.271 0.02435 \* |
| Abdomen.circumference 0.87462 0.06891 12.692 < 2e-16 \*\*\* |
| Wrist.circumference -1.44891 0.44669 -3.244 0.00141 \*\* |
| --- |
| Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |
| Residual standard error: 3.877 on 175 degrees of freedom |
| Multiple R-squared: 0.7465, Adjusted R-squared: 0.7407 |
| F-statistic: 128.8 on 4 and 175 DF, p-value: < 2.2e-16 |

Table 18. Best rpart model: cross-validated 3-bin model w/ abdomen only.

|  |
| --- |
| Overall Statistics |
| Accuracy : 0.5775 |
| 95% CI : (0.4544, 0.6939) |
| No Information Rate : 0.3662 |
| P-Value [Acc > NIR] : 0.0002362 |
| Kappa : 0.3744 |
| Mcnemar's Test P-Value : 0.0002054 |
| Statistics by Class: |
| Class: [1.9,15.1) Class: [15.1,22.3) Class: [22.3,45.1] |
| Sensitivity 0.8947 0.5769 0.3462 |
| Specificity 0.7500 0.6444 0.9778 |
| Pos Pred Value 0.5667 0.4839 0.9000 |
| Neg Pred Value 0.9512 0.7250 0.7213 |
| Prevalence 0.2676 0.3662 0.3662 |
| Detection Rate 0.2394 0.2113 0.1268 |
| Detection Prevalence 0.4225 0.4366 0.1408 |
| Balanced Accuracy 0.8224 0.6107 0.6620 |

Table 19. Best Naïve Bayes model: 3-bin selection with Height, Chest, Abdomen, and Wrist circumference.

|  |
| --- |
| Overall Statistics |
| Accuracy : 0.5634 |
| 95% CI : (0.4405, 0.6809) |
| No Information Rate : 0.3662 |
| P-Value [Acc > NIR] : 0.0005614 |
| Kappa : 0.3602 |
| Mcnemar's Test P-Value : 0.0015978 |
| Statistics by Class: |
| Class: [1.9,15.1) Class: [15.1,22.3) Class: [22.3,45.1] |
| Sensitivity 0.8947 0.3462 0.5385 |
| Specificity 0.6538 0.7778 0.9333 |
| Pos Pred Value 0.4857 0.4737 0.8235 |
| Neg Pred Value 0.9444 0.6731 0.7778 |
| Prevalence 0.2676 0.3662 0.3662 |
| Detection Rate 0.2394 0.1268 0.1972 |
| Detection Prevalence 0.4930 0.2676 0.2394 |
| Balanced Accuracy 0.7743 0.5620 0.7359 |

Table 20. Cross-model accuracy checks compared.

|  |  |  |  |
| --- | --- | --- | --- |
| Test | MLR | Rpart | Naïve Bayes |
| Correlation Accuracy | 0.81 | 0.64 | 0.63 |
| Min/Max Accuracy | 0.82 | .82 | 0.80 |
| Root Mean Square Error | 4.84 | 0.74 | 0.78 |
| Mean Absolute Error | 3.70 | 0.46 | 0.49 |
| Mean Absolute Percentage Error | 0.25 | 0.20 | 0.22 |

Our backwards step-wise linear regression model resulted in an adjusted R-squared of 0.7407 and a F-statistic of 128.8. The 3-bin rpart model has an accuracy of 0.5775 and kappa of 0.3744. The 3-bin Naïve Bayes model had an accuracy of 0.5634 and a kappa of 0.3602. Once we consider these numbers and the cross checks in table 20 we can conclude the linear regression model provides the best model fit for our data predictions. While the Root Mean Square Error and Mean Absolute Error do appear higher, we must also consider linear regression does not have the limitation of forcing predictions into only one of three bins. Linear regression is a more natural fit for this data and allows continuous predictions.

## Prior Research Comparison

If we recall from the prior research section, we examined two different studies which attempted to perform a similar analysis on bodyfat predictions using body measurements. In (Johnson, Navarro, Idiong, & Weeks), they utilized a linear regression model with abdomen and wrist measurements and were able to achieve a R value of 87%, notably higher than the 75% achieved in this paper. In the other study, by (Lean, Han, & Deurenberg, 1996), they also conducted a multiple step-wise linear regression and achieved 86% accuracy with the waist circumference, triceps-skinfold, and age variables.

Additional analysis would be needed to determine why the reason for the difference in accuracies, although we could infer it has something to do with the size of their training and test sets and the different software they used as contributing factors. From this comparison, however, we can positively deduce multiple step-wise linear regression does appear to be the optimal model selection.

## Final Model Selection

After an exhaustive examination of the Bodyfat dataset with multiple linear regression, rpart decision trees, and Naïve Bayes we can confidently decide of the best model to use in accurately predicting bodyfat percentage based on simple body measurements. Based on the findings of this paper, and cross-checked with similar research, we have determined a backwards multiple step-wise linear regression model is the best fit for accuracy in our predictions.

## Lesson Learned

This research project was challenging, but a definitively well-rounded learning experience of the depths of data mining algorithms. After researching previous works, it became apparent early on linear regression would be well suited due to the nature of the numerical continuous data. However, for the sake of exploring additional options I decided to pursue decision trees and Naïve Bayes classification to determine the similarities and differences in these algorithms and what their output might be.

The data preparation and outlier analysis were straightforward with no surprises. As predicted, linear regression was a good fit and provided solid numbers. The rpart decision tree section provided the first challenge. Knowing the dependent variable needed to be categorical, I discretized it into 6 bins. Making these decisions seemed arbitrary initially, impeding my confidence in the model selection. It became evident, after increasing the binning to 8 for increased granularity, this model would not be able to provide nearly the same level of accuracy has found in the linear regression model since I had to reduce binning to 3 just to make use of all bins. Naïve Bayes had similar challenges, requiring a binning level of 3 for proper performance. The final challenge was formulating functions that would allow a proper accuracy and error check across all three models. Ultimately, I decided to keep the unique measurements for each model, i.e. R-squared versus Kappa, and created functions which would provide numbers across all three, i.e. correlation accuracy and Root Mean Square Error. Overall, this was a demanding project, but well worth the investment with interesting results discovered in the process about both the data and processes utilized.

## Final Thoughts

The goal of this project was to examine a dataset of body measurements to determine if we could effectively predict body fat percentage. Backwards multiple step-wise linear regression gave us the best accuracy, both in this paper and other research papers which were compared. Although we can definitively say a best model fit for this dataset, this is not to say this is the gold standard. Our examination was limited, with only 251 observations. Future studies should include a much larger dataset. Additionally, other methods should be explored, such as Support Vector Machines (SVM) or Neural Networks. This paper focused on the Brozek formula, for reasons explained in detail in the data description section. However, another study might do a comparative analysis of the Siri measurements to see if there is a correlation or better prediction rates.

References

Bergman, R. N., Stefanovski, D., Buchanan, T. A., Sumner, A. E., Reynolds, J. C., Sebring, N. G., . . . Watanabe, R. M. (2011). A better index of body adiposity. *Obesity, 19*, 1083-1089.

Eberly, L. E. (2007). Multiple linear regression. In *Topics in Biostatistics* (pp. 165-187). Springer.

Guerra, R. S., Amaral, T. F., Marques, E., Mota, J., & Restivo, M. T. (2010). Accuracy of Siri and Brozek equations in the percent body fat estimation in older adults. *The journal of nutrition, health & aging, 14*, 744-748.

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: concepts and techniques.* Elsevier.

Johnson, R. W. (1996). Fitting percentage of body fat to simple body measurements. *Journal of Statistics Education, 4*.

Johnson, T., Navarro, J., Idiong, I., & Weeks, M. (n.d.). Predicting Body Composition Using Simple Measurement Techniques.

Kim, B. (2015, September 21). *Understanding Diagnostic Plots for Linear Regression Analysis*. Retrieved from Research Data Services + Sciences: https://data.library.virginia.edu/diagnostic-plots/

Lean, M. E., Han, T. S., & Deurenberg, P. (1996). Predicting body composition by densitometry from simple anthropometric measurements. *The American journal of clinical nutrition, 63*, 4-14.

Therneau, T. M., Atkinson, E. J., & others. (2018). An introduction to recursive partitioning using the RPART routines. Technical Report. URL https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf.