Predicting Body Composition Using Simple Measurement Techniques

Tim Johnson, Javier Navarro, Idongesit Idiong, and Mark Weeks
Seidenberg School of CSIS, Pace University, White Plains, New York

***Abstract*—The ability to take simple measurements such as neck or waist sizes, age, and possibly other factors and combine them to predict body fat content has been a long established practice for individuals in medical studies, athletes, and health specialists. This paper will assess an original study using modern techniques to validate this approach in estimating body composition.**

Keywords-Body fat, anthropometric equations, prediction

I. INTRODUCTION

An immense amount of study has been devoted to determining the percent of body fat from anthropometric equations since the late nineteen sixties when a wave of physical exercise swept the country. A mythology was established with the publishing in 1959 of *The Loneliness of the Long Distance Runner* by Alan Stillitoe that started people talking (movie in 1962). Followed by the promise of a "runner's high" in *The Complete Book of Running* (1978) by Jim Fixx, a health craze was in full swing. When Jim Fixx died during a training run in 1984 concerns about weight, body fat, exercise, and physical health in general and heart health in particular because he was initially obese and his family had a history of heart trouble.

A data file (bodyfat) posted in 2003 on StatLib at Carnegie Mellon University contains an article by Roger W. Johnson [1] of the South Dakota School of Mines written in 1995 that assess a formula for Lean Body Mass derived from data collected by K. W. Penrose [2] *et al* of Brigham Young University (1985) bring us back to that era. The formula was included in the short BYU paper along with four other formulas for comparison. These included studies were by Hodgdon and Beckett, Wright and Wilmore, Wilmore and Behnke, and McArdle et al. All these studies used circumferences as part of the anthropometric equations. Not included were later and overlapping studies by Jackson and Pollack (1985, 1978, and 1976) that followed a different approach by using skin-folds as measurement parameters [3]. Nothing was original in any of the previous studies except the formulas derived to predict body fat content based on easily obtainable measurements. All these studies including the hydrostatic weighing used as the gold standard to check the equations are now suspect because they used a two-component body model. The two-component model is based on a work by Siri in 1956 that divided the body into two types: lean body tissue and fat tissue. A four–component body model[3] consisting of fat mass, total body water, bone mineral mass and residual is considered more accurate using dual energy X-ray absorptiometry (DXA).

Since the body mass index (BMI) was created back in the mid-18th century by the Belgium mathematician Adolphe Quetelet[1] there have been hundreds of anthropometric equations. The focus for this study is to reassess the Johnson study using the KW Penrose equation and data. This equation:

$$LBW = 17.298 + .89946(\text{Weight in lbs}) - .2783(\text{age in years}) + .002617(\text{age})^2 + 17.819(\text{Height in inches}) - .6798(\text{Waist-Wrist in cm}) \quad (1)$$

was derived using step-wise linear regression analysis and its value, 92.4%, has not been exceeded in any study since using the same methods of measurements.

II OVERVIEW

The data set for 252 individual collected in 1985 included density, %body fat, age, weight, height, and

[1] http://en.wikipedia.org/wiki/Body_Mass_index accessed 4/5/2014.

10 circumferences: neck, chest, abdominal, hip, thigh, knee, ankle, biceps, forearm, and wrist. These 15 measurements become the attributes in Weka software. There are 3780 data points all together. The selection of individuals for this study was crucial to its eventual success. The first 143 individuals was based on a central composite rotatable design.

There are two calculated attributes in the data set: density and body fat (they are related by formulas). The density attribute was determined by hydrostatic measurement using the formula:

$$D = W_{air}/[(W_{air} - W_{water})/(c.f.) - LV] \quad (2)$$

W = weight in air/water (kg)
c.f. = water correction factor .997 at 76-78°F
LV = residue lung volume in liters

The % body fat attribute was derived from the formula:

$$D = 1/[(A/a) + (B/b)] \quad (3)$$

where,
D = Body Density in gm/cm^3
A = proportion of lean body tissue
a = density of lean body tissue = 1.1 gm/cm^3
B = proportion of fat body tissue
b = density of fat body tissue = 0.9 gm/cm^3

Once D is known B can be solved for:

$$B = (1/D)*[ab/(a-b)] - [b/(a-b)] \quad (4)$$

Or using Siri's equation:

$$B = 495/D - 450$$

III METHODOLOGY

The methodology for developing a predictive equation (Eq.1) was originally done by using a central composite rotatable design. This is a complex system of matrices that have parameters the original designers never revealed. In fact the authors speculate that the results are due to the sampling techniques used to gather the data.

The modern software used to process the data was developed at the University of Waikato in New Zealand. The *Waikato Environment for Knowledge Analysis* or Weka software is written in Java and distributed under the terms of the GNU General Public License. It is available at www.cs.waikato.ac.nz/ml/weka free of charge.

The decision to use the first 143 instances as a training set followed in the path of the original designers. The remaining 108 instances were used as the test set.

IV ANALYSIS

We first examined the test set using the training set and discovered a remarkably accuracy prediction of R= 99.8% for % Body fat:

$$-428.9*\text{Density} + 471 \quad (5)$$

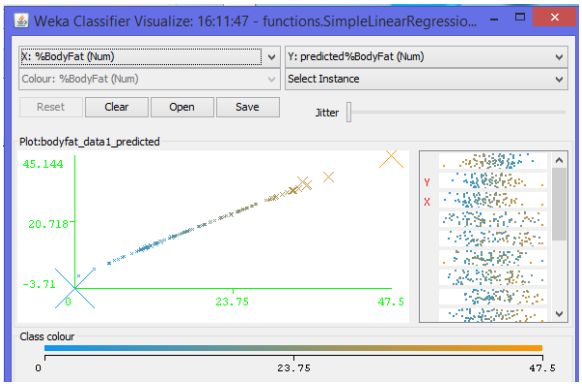

Figure 1 Here R = 99.8% for Eq.5 above that the % Body Fat in the test set was predicted by the equation using Density as its basis.

It's odd that some numbers didn't fit into this linear equation because body fat is derived from Density in the first place (see equation 4). We dropped Density, the gold standard against which all results are tested, from all future training/test data sets so there were now 14 categories of interest.

We next went on a systematic search for a better equation using linear regression but not based on density. When we examined the original equation, equation 1, two body circumference measurement were involved: Waist and Wrist. Their coefficient was the smallest of all the coefficients in the original equation. We built a new database that eliminated all the categories save two: % Body Fat and Waist and built a new training/test set combo. We were looking for the magic measurement that we could use to predict Body Fat. Testing for linear regression on this one measurement test set, R = 86%. The formula was

$$.6*\text{Waist} - 35.65. \quad (6)$$

Is this a good number? It is if we can't get one better.

We next tried for the magic 2 measurements to predict % Body Fat and built another dataset that included only %BodyFat, Waist, and Wrist. The

evaluation of this two measurement dataset was R = 87%. By including the wrist circumference we improved the accuracy 1%. The formula is

.7237*Waist – 2.6087* Wrist + .1211. (7)

We next tested out an obvious question: what would happen if we used all the categories but Density. Surely if wrist could improve the accuracy by 1% then all the measurements would be better. When we performed the regression analysis on all 14 categories we found Weka used nine of them for an R = 84.6%. Here is the formula for % BodyFat using the AllBut dataset:

.1069*Age -.1195*Weight -.4905*Neck +.9547*Waist -.2354 * Hip +.3219*Thigh +.3743*Ankle +1.0747*Forearm -2.4122*Wrist - 24.6889. (8)

In equation 8 we found 5 categories weren't contributing and of the nine that were used another 5 were throwing in less than .3 %. This is a reading of the coefficients that revealed the highest contributors to this equation were the Wrist, Forearm, Waist and Neck in descending order.

Suppose we made up a dataset of the measureable categories contained in the original equation: Age, Height, Weight, Waist, and Wrist? Could we regenerate the original equation? The results follow.

```
Linear Regression Model

%BodyFat =

      0.1087 * Age +
     -0.1821 * Height +
      0.6899 * Waist +
     -2.402  * Wrist +
      7.3358

Time taken to build model: 0.05 seconds

=== Evaluation on test set ===
=== Summary ===

Correlation coefficient                  0.8642
Mean absolute error                      3.683
Root mean squared error                  4.584
Relative absolute error                 50.7103 %
Root relative squared error             51.9442 %
Total Number of Instances              108
```

Figure 2 Weka attempt to match up with the original equation leaves out Weight.

V MODIFIED DATA SET

Since this effort failed to generate anything close to the original formula a new dataset was created that contains two categories that were mathematical derivatives of the original data. The categories are seen in Figure 3.

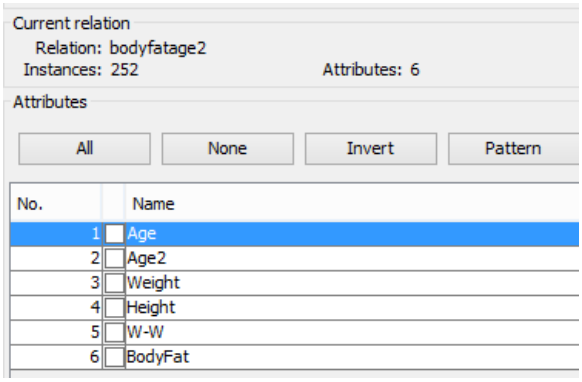

Figure 3 Recreated dataset from 1985.

The six attributes are 4 original attributes: Age, Height, Weight, and %BodyFat and two new attributes created from the original data: Age^2 and Waist - Wrist (W-W). All the other attributes were deleted.

The following images are the distributions of all 252 instances for the six attributes:

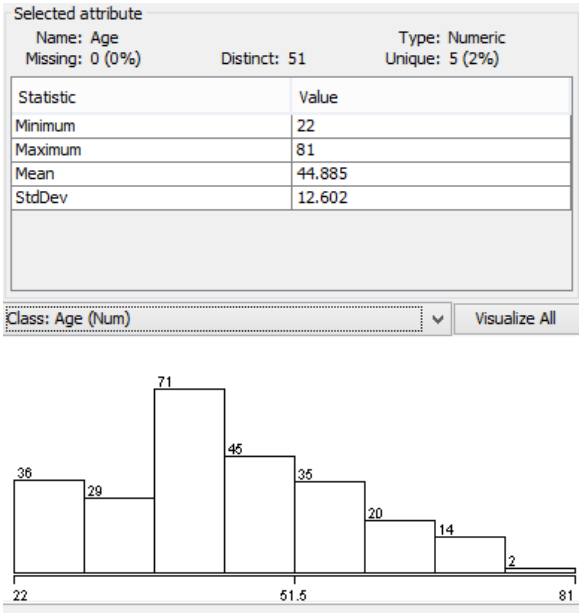

Figure 4 Distribution for Age

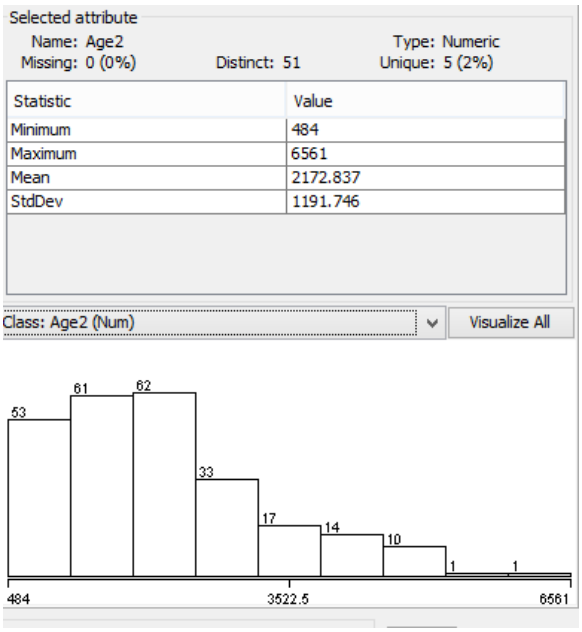

Figure 5 Distribution for Age squared

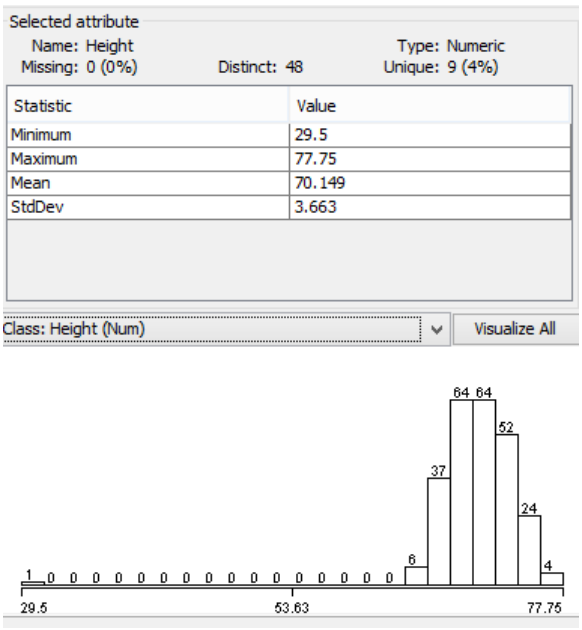

Figure 7 Distribution of Height

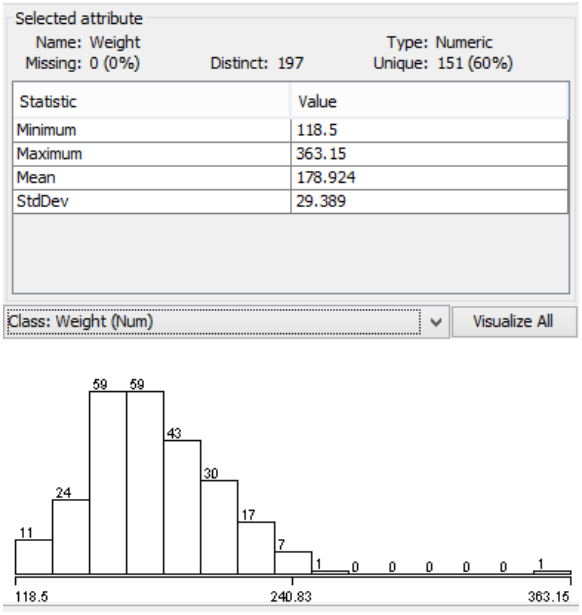

Figure 6 Distribution of Weight

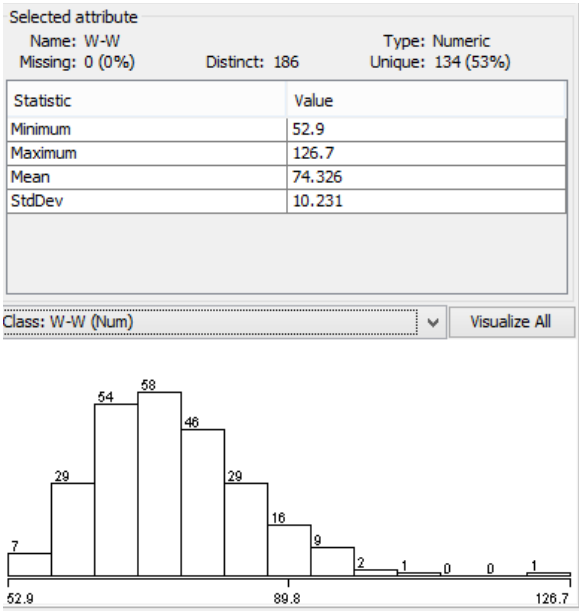

Figure 8 Distribution of Waist minus Wrist.

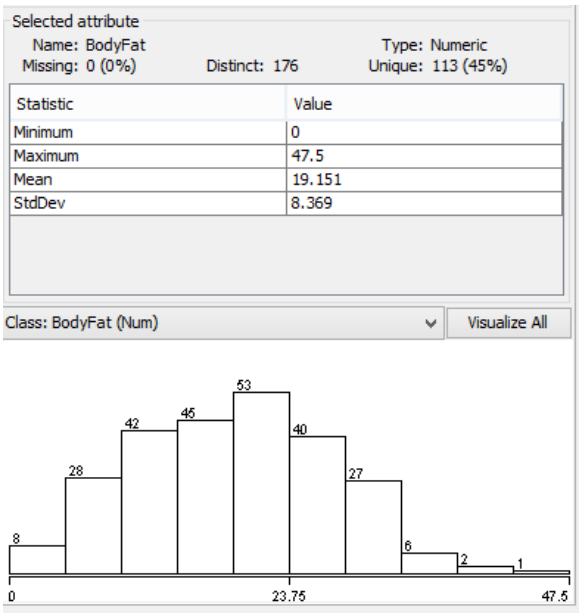

Figure 9 Distribution of % Body Fat

After dividing the data set up into a training set and a test set similar to previous trials the following output was observed for the linear regression analysis.

Results:

```
BodyFat =

      0.1592 * Age +
     -0.0013 * Age2 +
     -0.122  * Weight +
     -0.1354 * Height +
      0.9286 * W-W +
    -22.564

Time taken to build model: 0.06 seconds

=== Evaluation on test set ===
=== Summary ===

Correlation coefficient                  0.8622
Mean absolute error                      3.6604
Root mean squared error                  4.5885
Relative absolute error                 50.2409 %
Root relative squared error             51.9477 %
Total Number of Instances              109
```

Figure 10 Results of the trial failed to match 1985 results.

VI RULE BASE DATA SET

Next we examine Rules as another means to better fit the predicted value to the actual data for % Body Fat. We choose M5R as classifier with a minimum of 4 members of any particular rule.

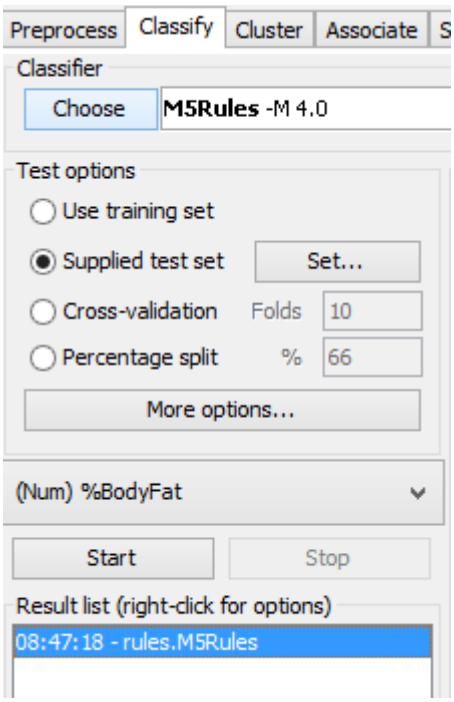

Figure 11 Setup for examination by rules

In this exercise two rules were discovered based upon a measurement of the waist. One third of the test set individuals fit their predicted % Body Fat if their waist measures less than 87.1 centimeters using formula LM1 and the other two-thirds fit their predicted % Body Fat when using formula LM2. The formulas were extensive using 10 of the available variables with an offset. The predicted fit has a regression value of 85.33% to the actual data.

```
LM num: 1
%BodyFat =
        0.0263 * Age
        - 0.0294 * Weight
        - 0.4452 * Height
        - 0.75 * Neck
        + 0.7771 * Waist
        - 0.0579 * Hip
        + 0.0792 * Thigh
        + 0.092 * Ankle
        + 0.2643 * Forearm
        - 0.5932 * Wrist
        + 11.8767
```

Figure 12 formula LM 1

```
LM num: 2
%BodyFat =
        0.0143 * Age
        - 0.1521 * Weight
        - 0.1582 * Height
        - 0.0657 * Neck
        + 0.8397 * Waist
        - 0.0316 * Hip
        + 0.0431 * Thigh
        + 0.0501 * Ankle
        + 0.1439 * Forearm
        - 0.3231 * Wrist
        - 15.5328
```

Figure 13 formula LM 2

Correlation coefficient	0.8533
Mean absolute error	3.6699
Root mean squared error	4.6852
Relative absolute error	50.5292 %
Root relative squared error	53.0907 %
Total Number of Instances	108

Figure 14 Details of the fit

A graphic for the rules was created in Weka

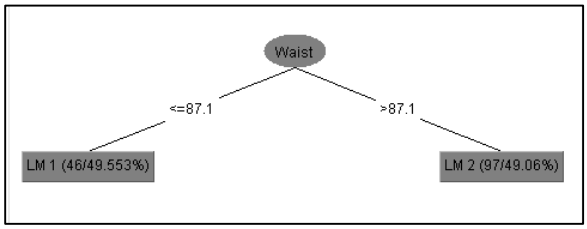

Figure 15 Rules graphic classifier

VII CLUSTERS

Next we tried to determine the effect of cluster on this data set using SimpleKMeans cluster filter. In Weka, all the categories are examined by default for clusters and we discovered some are more separated than others.

Cluster centroids:

Attribute	Full Data (143)	Cluster# 0 (83)	1 (60)
%BodyFat	19.093	15.1506	24.5467
Age	44.7692	43.6747	46.2833
Weight	177.1811	161.3464	199.0858
Height	69.9913	69.7289	70.3542
Neck	37.8224	36.5482	39.585
Chest	100.065	95.5867	106.26
Waist	91.7406	85.6181	100.21
Hip	99.7007	96.1867	104.5617
Thigh	59.3245	56.8795	62.7067
Knee	38.414	37.2843	39.9767
Ankle	23.1063	22.5337	23.8983
Biceps	32.2126	30.6687	34.3483
Forearm	28.5825	27.5843	29.9633
Wrist	18.1783	17.7578	18.76

Figure 16 Summary of clusters for the fourteen cateories

A visualization is also provided and seen is the two clusters of Weight in Figure 17. In the figure one can almost see the regression line for each cluster with different slopes verifying the Rule division under two rules. The Weight attribute clusters are quite distinguished from each other by nearly 38 pounds (circled in red) so the cluster is almost distinct with little cross over. Since the Rule divided into two parts based on the Waist dimension it was of interest to see this visualization of a cluster verified the Rule. This is seen in Figure 18.

Figure 17 The two clusters of Weight vs %Body Fat

Figure 18 The two clusters of Waist vs %Body Fat

The disjoint visible in the clusters of Weight are not seen in the cluster of Waist. Added to Figure 18 is a vertical line at approximately 87.1 which is the dividing line for the Rules. This shows the clusters interface is mostly to the right of the demarcation for the Rules. All of Cluster 1, in red, is valid for LM 2 but cluster 0 has included a portion of instances where Waist is > 87.1.

VIII CONCLUSIONS

In this paper we have examined a data set from the 1980s whose authors were able to determine to a greater degree a better fit using central composite rotational design methodology than anything we have been able to determine to date with Weka. Our best effort generated a regression value of 87%. This is not a bad number and we do not risk over fitting the training set to the test data. These numbers are if anything an indicator. There is a 4-component model [4] for predicting % Body Fat that is being offered as a substitute for the 2-component model; however, nothing beats the ease of the one measurement prediction formula. Since the goal is to produce a prediction equation for body fat with easily accessible measurements, the one or two-measurement formulas develop here are quite useful as indicators of body fat. The topic of obesity is richly studied and despite the multitude of studies there is still room for improvement. One aspect that was not tried in this analysis was removing outliers nor did we check for data errors. A second paper published a year later (1996) by RW Johnson [4] at a different institute covers this aspect but only mentions two instances of concern. The authors of this analysis elected not to consider the effect of these errors.

REFERENCES

[1] R. W. Johnson, "Bodyfat," Department of Mathematics & Computer Science, South Dakota School of Mines and Technology, 1995. Hosted by StatLIB, Department of Statistics, Carnegie Mellon, http://lib.stat.cmu.edu/modules.php, accessed 3/23/2014.

[2] K. W. Penrose, A.G. Nelson and A.G. Fisher, "Generalized body composition prediction equation for men using simple measurement techniques," FACSM, Human Performance Research Center, Brigham Young University, "Medicine and Science in Sports and Exercise, V17, No. 2, April 1985, p. 189.

[3] SD Ball, TS Altena and PD Swan, "Comparison of anthropometry to DXA: a new prediction equation for men," European Journal of Clinical Nutrition, published online 2004, http://www.readcube.com /articles/10.1038/sj.ejcn.1602003 , accessed 4/7/2014.

[4] R. W. Johnson, "Fitting percent of body fat to simple body measurements," Journal of Statistics Education [Online], 4(1), 1996, www.amstat.org/publications/jse/v4n1/datasets.johnson.html, accessed 4/9/2014.