Gil Graybill – BAN 525
Assignment 5 – Predicting Body Fat
Professor Cetin Ciner
June 14, 2020

Introduction

Accurately measuring body fat is not easy to do. I had this performed on me once in high school. The method involved seeing how much I weighed underwater with all of the air expelled from my lungs…three times. The process seems based on scientific principle (https://www.dexafit.com/blog2/hydrostatic-weighing-is-it-accurate), but having gone through it, I've often wondered if there was an easier way, or if they really needed that accurate of a measurement of body fat for a basketball bench-warmer like me.

The dataset we are using contains various body measurements for 252 men and comes courtesy of "Generalized body composition prediction equation for men using simple measurement techniques", K.W. Penrose, A.G. Nelson, A.G. Fisher, FACSM, Human Performance Research Center, Brigham Young University, Provo, Utah 84602 as listed in _Medicine and Science in Sports and Exercise_, vol. 17, no. 2, April 1985, p. 189. The response variable is "Percentage of body fat", and the remaining variables are age, weight, height, and the circumference of 10 different parts of the body (see Appendix).

We will be running the following types of analysis on this dataset:

(1) Nominal Logistic Regression: This type of logistic regression model calculates the best fit through the data points that minimizes the squared differences between the observed values and the fitted values, and uses the calculated value to further determine the value of the nominal (or categorical) response variable.

(2,3) A Neural Network is a set of algorithms that are designed to recognize patterns, simulating the human brain. As these patterns are recognized they are grouped together and virtual connections are formed. As with a neuron in the brain, an input (or firing) determines what other nodes are signaled. Each Neural Network has an input layer, one or many "hidden layers" where these connected are formed and revised, and an output layer. Within the "hidden layer" is where the various algorithms are applied to the data across one or more nodes.

In our first Neural Network model we will have one layer with three nodes, and the algorithm applied will be the TanH activation function.

In our second Neural Network model we will have two layers with three nodes, and three algorithms will be applied: TanH, Linear, and Gaussian.

Analysis and Model Comparison

OLS regression will find any relationships that exist between response variable and the potential dependent variables. One disadvantage for OLS is that if there are many variables that are highly correlated, OLS may model random errors. All variables will be included in the model, for better or worse. This is our "simple" benchmark analysis.

Neural Networks are a "big hammer" that can be used to pound on any data. If there is a complex relationship (that is to say, not linear or represented by a polynomial expression), NN will find it. The disadvantage is that a Neural Network is a "black box" with its hidden layer, which makes it hard to troubleshoot or graphically represent the chosen model.

For cross-validation we employed a 60-20-20 split for training-validation-test split. This gave us 151 rows of training data, 50 rows of validation data, and 51 rows of test data, splitting up the data randomly.

"Percent Body Fat" was chosen as the response variable and the rest of the variables were tested as dependent variables using JMP Pro 15. The calculated value for the response variable, the calculated value for "Percent Body Fat" was determined and recorded and the resulting formulas determined using all 3 methods.

To compare results between the models, we calculate the difference between each observed value for the change in percent of body fat and the calculated value for the change in the percent of body fat. These differences can be compared in different ways. "RSquare" is a percentage that tells how much of the variability is explained by the model. "RASE" is "Root Average Standard Error", and because this describes values of error, smaller is better. "AAE" is "Absolute Average Error" is an error term that uses the absolute value of the error to compare the size of the error without regard for direction.

When comparing the models we will be using our cross-validation split. Because the test data is not used to create or refine the models, those are the results we are most interested in. The results are as follows:

Model Comparison

Measures of Fit for Percent body fat

Validation	Predictor	Creator	.2 .4 .6 .8	RSquare	RASE	AAE	Freq
Training	Pred Formula Percent body fat OLS	Fit Least Squares		0.7590	4.3207	3.5944	151
Training	Predicted Percent body fat NN	Neural		0.7293	4.5789	3.6968	151
Training	Predicted Percent body fat NN 3.2	Neural		0.7102	4.7375	3.9127	151
Validation	Pred Formula Percent body fat OLS	Fit Least Squares		0.7336	3.7612	3.0570	50
Validation	Predicted Percent body fat NN	Neural		0.7953	3.2975	2.7424	50
Validation	Predicted Percent body fat NN 3.2	Neural		0.7912	3.3302	2.6123	50
Test	Pred Formula Percent body fat OLS	Fit Least Squares		0.6864	4.2797	3.4800	51
Test	Predicted Percent body fat NN	Neural		0.5181	5.3051	4.4273	51
Test	Predicted Percent body fat NN 3.2	Neural		0.4868	5.4751	4.6246	51

All three models performed about the same when modeling with the Training and Validation data. But when applied to data that was never seen before, the “benchmark” of Ordinary Least Squares came out on top with an R-Square value of 0.6894, which means that 69% of the variance can be explained by the model, which is pretty good. Apparently the Neural Network models were prone to overfitting on this dataset.

Interpretation

The formula that was determined using OLS is:

Predicted Body Fat Percentage = (-27.4134972319472) +
0.057685904933055 * "Age (years)" +
-0.126419875968757 * "Weight (lbs)" +
-0.0462286757439306 * "Height (inches)" +
-0.379142133142742 * "Neck circumference (cm)" +
-0.0156515864893665 * "Chest circumference (cm)" +
0.96459385295931 *"Abdomen circumference (cm)" +
-0.203531590160979 *"Hip circumference (cm)" +
0.288954919019271 *"Thigh circumference (cm)" +
0.18870793873311 *"Knee circumference (cm)" +
0.200887229423258 *"Ankle circumference (cm)" +
0.33080240971537 *"Biceps (extended) circumference (cm)" +

0.412370366128199 *"Forearm circumference (cm)" +
-1.89904298070358 *"Wrist circumference (cm)"

When running variable importance with these variables against this model, we see the following:

Variable	Main Effect	Total Effect
Abdomen Circ.	0.763	0.787
Weight	0.115	0.139
Hip Circ.	0.011	0.022

Abdomen Circumference has a total effect of 79% on the model, followed by Weight at 14% and Hip Circumference at 2% . Both Weight and Hip have negative coefficients, indicating that even though Abdomen circ. must have an oversized effect, so much so that it must be "tamed" and brought back down by the other variables. "Main Effect" and "Total Effect" are very close to each other, indicating there is not a lot of interaction between the variables.

With OLS, these 3 variables account for 95% of the effect on the model. That indicates that this is a strong linear regression model. I believe that Neural Networks are overfitting the data because the linear relationship is so strong. Neural Networks work best when the relationships are non-linear (https://elitedatascience.com/machine-learning-algorithms). This effect may even be magnified with the 2nd Neural Network model where many more nodes and levels are being utilized. We can see that this model performs 3% worse on the test data than the "simpler" Neural Network model.

The other issue (also cited in https://elitedatascience.com/machine-learning-algorithms) is that with only 252 rows, this is a small dataset. Neural Networks require large amounts of training data to "learn" properly. To go back to the human brain analogy, our model, trained with only 252 observations, is merely a baby. To get the brain truly functioning as an "adult", it will need thousands more observations to train itself properly.

When using the following data as input, the OLS model determines that the Body Fat Percentage for this new observation is **17.70%:**

Age (years)=50 , Weight (lbs)= 167; Height (inches)= 67.75; Neck circumference (cm)= 38.8; Chest circumference (cm)=100.4; Abdomen circumference (cm)=89; Hip circumference (cm)= 93.2; Thigh circumference (cm)= 57.0; Knee circumference (cm)=34.8; Ankle circumference (cm)= 20.6; Biceps (extended) circumference (cm)= 33.9; Forearm circumference (cm)= 28.3;Wrist circumference (cm)= 18.0;

Appendix – Columns in this dataset

Percent Body Fat

Age (years)

Weight (lbs)

Height (inches)

Neck circumference (cm)

Chest circumference (cm)

Abdomen circumference (cm)

Hip circumference (cm)

Thigh circumference (cm)

Knee circumference (cm)

Ankle circumference (cm)

Biceps (extended) circumference (cm)

Forearm circumference (cm)

Wrist circumference (cm)