Predicting Obesity Using DASL Dataset

460352996 470066919 480145820 480407614

GitHub code repository is here

This version was compiled on November 2, 2019

Excess weight, especially obesity, has become an epidemic in the 21st century. This study aims to investigate an alternative method to determine "overweight" individuals oppose to body fat percentage. Two alternative indicators are considered - BMI and body density. The results showed that BMI can be explained the best using simple body measurement and the measurement on abdomen is the most important predictor in estimating all three methods. A simpler predictive model for obesity has been developed using measurements on chest, abdomen and bicep. Limitations and implications are discussed.

Obesity | Regression | Correlation | Prediction

Introduction. Excess weight is the new epidemic of the 21st century and has resulted in many significant health and economic consequences for the global population (Stein and Colditz, 2004). In Australia, the obesity epidemic has spread drastically as 1 in 3 adults are classified as overweight or obese (Australian Institute of Health and Welfare, 2019). Researches have shown that this epidemic is more common in males than females and hence, BYU Human Performance Research Center has collected data from 250 men of various age and obtained estimates of the percentage of body fat through underwater weighing and various body circumference measurements (Rahman and Harding, 2013; DASL, n.d.). As body fat percentage is difficult to calculate in real life, the value for body fat percentage was derived from body density using the Siri's 1956 equation.

1.1 Data cleaning and processing. The DASL dataset contain 16 variables including body density, age, body fat percentage and body measurements. The dataset has already been cleaned and for the purpose of this analysis, an additional variable - BMI, has been added where

$$BMI = \frac{Weight(lbs) * 703}{Height(in)^2}$$

1.2 Sampling Method. Few details were provided with regards to the sampling method. However, from looking at the dataset, there is a gender bias as the epidemic question is one related to both genders, yet only male were involved in the sample. This suggests that any analysis based on this dataset cannot be applied to the whole population but only the male population.

Analysis. The analysis approach can be broken down into three steps.

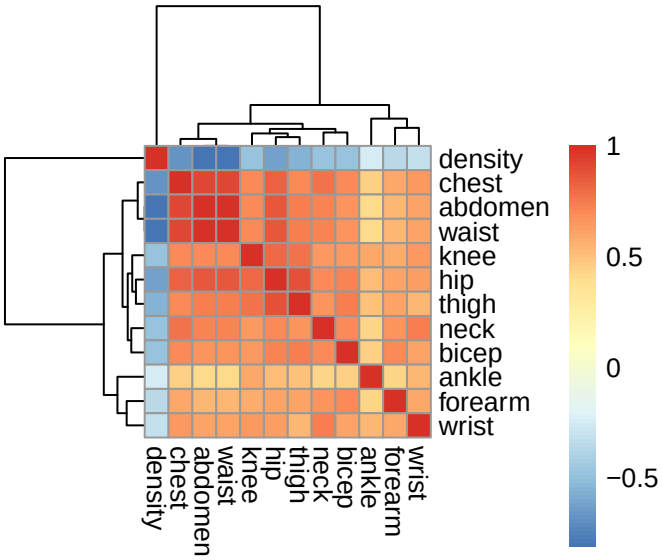
Step 1. Using multiple regression, firstly determine the number of body measurements that is significant in building an accurate prediction model for the three obesity indicators (Body Fat Percentage, BMI and Body Density) individually and how much variation can be explained using only body measurements to examine the ease of calculation.

Step 2. Compare the end results to determine the best indicator given that body measurements are the only available variables. In each sample, a relative importance test will also be run to determine which body measurement is relatively the most important.

Step 3. Using BMI as the obesity indicator, a binary indicator will be added to differentiate the sample into overweight individuals (1) and non-overweight individuals (0). A logistic regression is run on the binary indicator with significant variables that we have identified throughout research in order to build a simpler model to determine the odds of an individual being obese.

Multiple Regression.

Analysis. A correlation matrix is firstly drawn to show the general interactive correlation between variables.



Notably, waist, chest, abdomen are showing highly correlated relationships, and this may be due to the fact that they are from a similar body area. Hence, body measurements that are from similar areas are classified and linked together using the above graph.

All three obesity indicator follow a similar procedure for multiple regression analysis.

Backward stepwise model was used for body fat and BMI where a full model was selected at the start. The least informative variables were dropped using AIC until only the most relevant variables remain and a final fitted model is achieved. A forward variable selection method was used for body density and a null model was selected at the beginning with subsquent addition of the most statistically signicant variable. The final fitted model is formed when no further addition is required.

For the analysis of BMI, a transformation using log was required. However, as BMI is measured in unit increase, the interpretation of percentage changed is unreasonable. Hence, the non-transformed model was used as the final fitted model.

Assumptions for normality and homoskedasticity were checked via residuals plots and QQ plot. Then the predictor variable plots were drawn for illustrating the relative importance of the remaining variables.

Results. The final fitted models for the three obesity indicators were:

$$\hat{BodyFat} = 1.52 - 0.3965Neck - 0.128Chest + 1.01805Abdomen - 0.28758Hip + 0.26Bicep - 1.55084Wrist$$

$$\hat{BMI} = -10.94 + 0.161Chest + 0.127Abdomen + 0.050Hip + 0.150Thigh - 0.23Knee + 0.115Forearm$$

$$\hat{BodyDensity} = 1.1104052 + 0.0019085Neck - 0.0022064Abdomen + 0.0011314Hip - 0.0006094Thigh$$

All three QQ plots had shown a straight lines and this satisfies the normality assumption. However, the residual plots showed a slight variation for all three indicators, but given that the residual units were quite small, it is acceptable for the current analysis.

Overall, abdomen is relatively the most important measurement in predicting all three obesity indicator and this result was expected as it corresponds to the previous correlation matrix analysis.

BMI was identified as the best obesity indicator given only body measurements as 90.2% of its variation can be explained using solely body measurements compared to 73.5% for Body Fat Percentage and 70.4% for Body Density.

Logistic Regression.

Analysis. A binary logistic regression was used to calculate the probability of a person being overweight where overweight is indicated by a BMI greater or equal to 25.

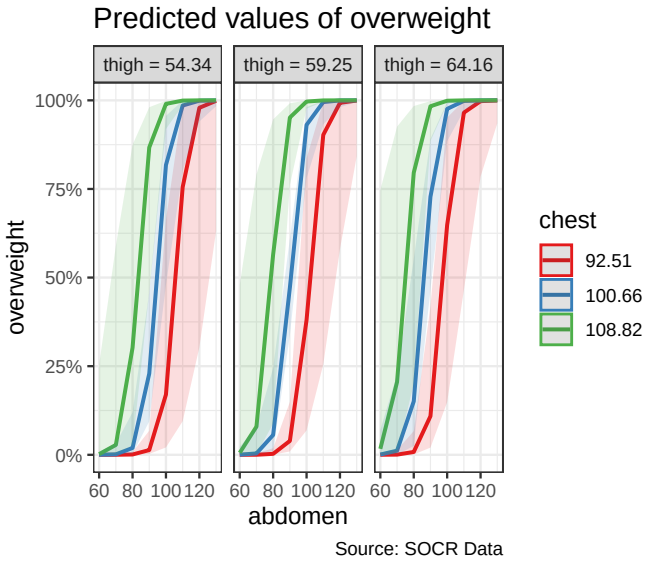
Similar to the multiple regression analysis, a backward stepwise selection method is also utilised to determine a final classification model.

Using the model, the probability of an individual being obese is then derived using a confusion matrix where its performance as a predictive model is evaluated.

Results. The final classification model is

$$logit(p) = log(\frac{p}{1-p}) = -75.89380 + 0.37854Chest + 0.27128Abdomen + 0.22381Thigh$$

The results of this model is visualised using the graph below through the sigmoid function. The predicted values to probabilities are mapped between 0 and 1. For predictions of 0.5 and above, these are classifed as people who are overweight. Whereas predictions of below 0.5 are classified as people who are not overweight.



A confusion matrix is created to derive the performance of our classification model.

```
#            Reference
#  Prediction   0   1
#           0 117  12
#           1   8 113
```

The accurate and inaccurate predictions are the diagonal and non-diagonal values respectively. For this model, there are 117 + 113 accurate predictions, and 12 + 8 are inaccurate predictions.

In fact, it has a sensitivity, ability to correctly identify those who are overweight, of 93.6%. This is calculated as the number of correct positive predictions (117) divided by the total number of positives (117 + 8).

The model also has a specificity of 90.4%, which is the ability to correctly identify those who are not overweight. This is calculated as the number of correct negative predictions (113) divided by the total number of negatives (113 + 12).

Along with an accuracy of 92%, we summarise that this new simplified model is, therefore, a good predictive model.

Limitations.

4.1 Gender bias. The data is taken from 250 males without any record of females. Therefore, the result of this analysis can only be applied to the male population rather than the entire population in general.

4.2 Age range. Majority of the participants are males between the age of 40-50. This is a potential bias in the sample that can compromise the prediction accuracy on younger or older males.

4.3 Multicollinearity. Several variables from the dataset are highly dependent, with the most significant correlation between waist and abdomen. Through plotting the points into a QQ-Plot, it is evident that all points are closely sitting on the line. Hence, during model selection, waist was dropped to prevent multicollinearity.

Conclusion. Through multiple regression and variable selection, a fitted model with solely body measurements was determined for each of the three obesity indicators. Using R^2, BMI was identified as the best indicator as it has the highest proportion of variance that can be explained using only body measurements.

Abdomen was the most important body measurement for determining obesity because for all three indicators, it ranked the highest in terms of relative importance in prediction.

By separating the dataset with the binary variable for overweight individuals, a simplified prediction model with 92% accuracy was built. The simplified model contains three body measurements - chest, abdomen and thigh, and should be relatively simpler to measure.

Reference List.

1. Australian Institute of Health and Welfare (AIHW). (2019). Overweight & obesity. Australian Government. Retrieved from https://www.aihw.gov.au/reports-data/behaviours-risk-factors/overweight-obesity/overview

2. DASL. (n.d.). Bodyfat. DASL. Retrieved from https://dasl.datadescription.com/datafile/bodyfat

3. Rahman, A., & Harding, A. (2013). Prevalence of overweight and obesity epidemic in Australia: some causes and consequences. JP Journal of Biostatistics, 10(1), 31-48.

4. Stein, C. J., & Colditz, G. A. (2004). The epidemic of obesity. The Journal of Clinical Endocrinology & Metabolism, 89(6), 2522-2525.