**Executive Summary**

Jack Bressett, Yizhe Ge, Haoyang Yan

**Introduction & Results**

In our report, we attempt to demonstrate a simple, robust, and accurate model for predicting body fat percent from a collection of measurements from 252 men. Previous scientific work has demonstrated how underwater weighing can precisely measure density, then density can be used with certain physiological constants to calculate a body fat percent to a high degree of certainty (Siri. 1961) and (Katch and McArdle. 1977). This work is known as Siri’s Equation and despite its precision we cannot efficiently measure density, so a simpler model is necessary.

Our final model relies on the Height in inches and abdomen circumference in inches:

Body Fat % = 1.626\*Abdomen (Inches) - 0.572\*Height (Inches)

Our model states that for a one-inch increase in Abdomen Circumference, we expect an increase of 1.613 percent in body fat percentage. Conversely, a one-inch increase in height results in a decrease in body fat percentage by .566 percent. This decrease reveals the notion of becoming leaner as one grows taller. Also, an Abdomen Circumference and Height of 0 inches will result in a prediction of 0 percent body fat.

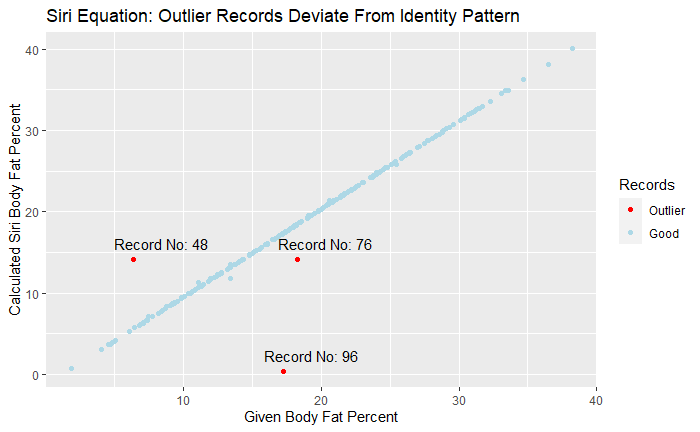
To simplify further, the variable Abdomen (Inches) can be represented by pant size and an easy rule of thumb for approximating body fat percent is expressed as:

Body Fat % 1.5\*(Pant Size)- 0.5\*Height (Inches)

A useful table of body fat percentages from the rule of thumb is as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Heights | Pant Sizes | | |
|  | 27 | 32 | 37 |
| 5 Ft. 3 In. | 9 % | 16.5 % | 24 % |
| 5 Ft. 6 In. | 7.5 % | 15 % | 22.5 % |
| 5 Ft. 9 In. | 6 % | 13.5 % | 21 % |
| 6 Ft. | 4.5 % | 12 % | 19.5 % |

**Data Exploration**

We want to make a model that could help most people estimate their body fat percentage. In exploratory analysis, three records were extreme cases in height, weight and body fat, respectively. Despite recovery efforts, these records negatively impact the model too much to be included, they have ID’s 39, 42,182. Next, we used the Siri Equation to calculate a body fat percentage then plot against our given body fat percentages. 

An identity line appears; points furthest away from the line are outliers and violate the assumption that the data are reliable. Removing these three points brings our final data count to 246 records. A note, centimeters has been changed to inches.

**Model & Feature Selection**

Exploratory analysis showed some features had stronger linear relationships than others, and after reverse stepwise regression, a full model with thirteen features was narrowed to the seven most significant features. Then, features easiest to measure and those which captured the most variance in body fat percentage were selected. Eventually, Abdomen only was a powerful model, but the combination of Abdomen and Height relieved error trends that appeared in the only Abdomen model. Thus, the final model consists of Abdomen and Height and has an r^2 value of .708 meaning nearly 71% of the variance in the body fat percentage data is captured by the model. The intercept was then forced through zero because it is natural for the scenario, and our features are shown to have strong linear relationships with body fat percentage. Weaknesses of the model include underpredicting low inputs of height and abdomen and overpredicting high inputs. Strengths of the model include simplicity and accuracy relative to the constraints.

**Conclusion**

The final model is not perfect and would benefit from a square term for height. This would complicate the model, however, as its main strength is interpretability and ease of calculation. As it stands, the model accounts for a relatively high amount of variance in body fat percentage and tells an interesting story about thinning out as one grows tall and what it may mean when you “bump up a pant-size”. Ultimately, the model can be efficiently used and does a fine job at predicting a wide range of body shapes.

Citations:

1. Siri, W. E. (1961). *Body composition from fluid space and density.* In J. Brozek & A. Hanschel (Eds.), **Techniques for measuring body composition** (pp. 223-244). Washington, DC: National Academy of Science.
2. Katch, Frank and McArdle, William (1977). \_Nutrition, Weight Control, and Exercise\_,   
   Houghton Mifflin Co., Boston.