Project Summary

Authors: Yuchen Zeng, Jiantong Wang, Ruixuan Zhao, Hao Pan

1. Introduction

Body fat percentage can be defined as mass of fat divided by total mass. Body fat percentage is a very important index for people to evaluate their health condition, health costs and even happiness. Though it's important, it's not easy to compute as it's impossible to get mass of fat. Many related works have been done. The only problem left is how to estimate body density, which is costly to get precise estimation. Thus, it would be great if we can come up with either a new way to estimate body fat percentage or an economical way to estimate body density.

Finally, our group built two models that can estimate body fat percentage fairly accurate with very limited body information. The first model we built estimates body fat percentage directly with abdomen circumference and weight. The second one estimates the reciprocal of body density to estimate body fat percentage with abdomen circumference and height.

2. Background Information

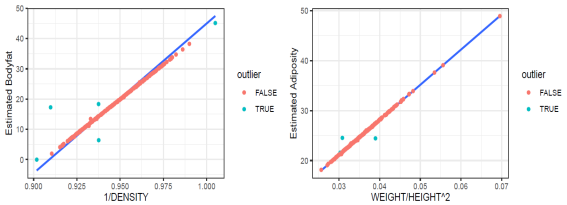
For a long time, people used various methods to estimate the percentage of body fat. Some people use predictive equation: "Siri's equation" to determine body fat using body density. However, it is difficult for doctors to determine body volume (body density) by underwater submersion. We should find a way to determine percentage of body fat by commonly available measurements. The dataset BodyFat.csv contains 252 records of available measurements: age, weight, height, adiposity, neck, chest, abdomen, hip, thigh, knee, ankle, biceps, forearm and wrist circumference. Also, we have density determined from underwater weighing and percentage of body fat from "Siri's equation". We should use the data to construct simple and robust statistical models to predict percentage of body fat.

3. Data Preprocessing

First, because we already know the relationship between density and body fat percentage which is

$$BodyFat = \frac{495}{Density} - 450.$$

We compared the predicted body fat which deduced from density and the input body fat and check which one is more reasonable. By this way, we could detect some potential outliers.



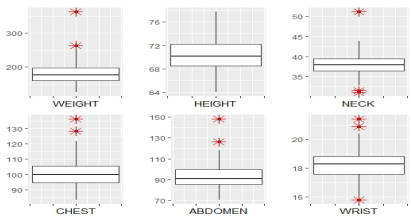
The potential outliers and shown in the figures above (left). Comparing the input body fat and estimated body fat, according to the 96th observation, the predicted body fat is impossible, we keep the original body fat percentage because the predicted body fat is too extreme to happen. As for 48th, 76th and 216th observation, we cannot tell which input body fat or density is correct, so we straightly delete them . As for 182nd observation, both predicted body fat and input body fat are impossible to happen, we delete it as an outlier.

Since we already know the formula to compute the adiposity by using weight and height, we use the weight and height to calculate the adiposity.

$$adiposity = \frac{0.4547 * weight}{(height * 0.025)^2}$$

The potential outlier detected in this way is shown as the figure (right) above. Looking into the estimated adiposity and the original data, we noticed the estimated adiposity of the 42rd sample is apparently problematic. So we took a further look at this sample, and found the extremely short height. We use known adiposity to estimate the height 69.45 and replace the original data. As for 163rd and 221st samples, we cannot tell which element is problematic, so we delete these two observations.

Now look at the boxplot to see other potential outliers.



These points all are potential outliers, but looking into the samples, we cannot delete them curtly.

4. Model Motivation and Selection

We came up with 3 idea to fit our model, the following part is how we fit and choose the models.

Model 1

We borrowed some idea from US Army, they have a formula for men to calculate body fat, it is like this:

$$Body\ Fat \sim \log(Waist - Neck) + \log(Height)$$

We borrowed this idea. But we don't have data of waist, but we thought abdomen is not much different from waist, so we just used abdomen instead of waist.

We fit the model and check the results, we can get the adjusted R square is 0.7179, which is ideal. But the model requires 3 variables as input. It is kind or more complex than model 2. So we tried to dismiss NECK in our model, and see what would happen. Then, we find that although we exclude Neck from our model, the results are not getting so much worse. The adjusted R square is 0.6971, which is also fairly acceptable. So we just exclude Neck from our model. So our final model is:

$$Body\ Fat \sim \log(Waist) + \log(Height)$$

Model 2

For we already know that body fat is calculated by the body density, so our aim could be try to fit a model to predict density. We came up with two solutions:

Roughly, a guy could be regard as a cylinder, so a guy's density could roughly be described as:

$$\rho = \frac{Weight}{Volume} = \frac{Weight}{Height \times Area} = \frac{Weight}{(Height)^2} \times \frac{Height}{\pi R^2} = BMI \times \frac{Height}{\pi \times (\frac{Abdomen}{4\pi})^2}$$

And we could find that there is a linear relationship between BMI and Abdomen, so we could fit a regression for BMI on abdomen.

Then we could use abdomen to represent BMI, and further simplify our formula:

$$\rho = (k \times Abdomen + b)\frac{Height}{\frac{(Abdomen)^2}{16\pi}}$$

$$\frac{1}{\rho} = \frac{\frac{1}{16\pi} \times (Abdomen)^2}{(k \times Abdomen + b)Height} = \frac{1}{16\pi k^2}[\frac{(k \times Abdomen + b)}{Height} + \frac{b^2}{(k \times Abdomen + b)Height} - \frac{2b}{Height}]$$

$$Body\ fat \sim \frac{1}{\rho} \sim \frac{(k \times Abdomen + b)}{Height} + \frac{b^2}{(k \times Abdomen + b)Height} - \frac{2b}{Height}$$

For $(k \times Abdomen + b)$ is BMI based on the relationship with Abdomen. So we named it as BMI.Abdomen. I also tried just simply use adiposity here, but the adjusted R square is smaller than use abdomen here. I think it is kind of caused by that the Adiposity is relied on many other data, which may accumulate the variances. So we just use abdomen instead of adiposity.
To simplify our model, we could try to ignore some terms and simply run a regression like this:

$$BODYFAT \sim \frac{1}{HEIGHT} + \frac{1}{ABDOMEN \times HEIGHT} + \frac{ABDOMEN}{HEIGHT}$$

Model 3

	X	mindex	n	predictors	rsquare	adjr	predrsq	cp	aic	sbic	sbc
1	7	1	1	ABDOMEN	6.478017e-01	0.646358275	0.63726116	73.97764	1438.179	739.0771	1448.695
15	32	15	2	WEIGHT ABDOMEN	7.080663e-01	0.705663519	0.69825412	21.91079	1394.013	695.5229	1408.034
106	228	106	3	WEIGHT ABDOMEN WRIST	7.154836e-01	0.711956570	0.70221087	17.25624	1389.681	691.2710	1407.208

By using exhaustion method to list all possible linear models, we found the model using weight and abdomen as predictors is not only simple but also fit our data fairly well. Looking at the single-variable linear model, the one that performs best is using abdomen as predictor with R square equal to 0.65. And the one which performs best using two variables is taking weight and abdomen as predictors. The R square has been improved by about 0.06. However, as weight is a commonly used measurement that almost everyone knew their weight, adding weight as a predictor is good way to improve the accuracy of our model. According to the table above, if we still want to add a new predictor, wrist would be the best choice. However, the R square can only be improved by less than 0.01. So we select $bodyfat \sim abdomen + weight$ as the best linear model.

Model Performance

	Model 1	Model 2	Model 3
Adjusted R square	0.6971	0.703	0.7057
Residual standard error	4.132	4.091	4.073
F - statistics	282.9	194.3	294.7

Cross validation

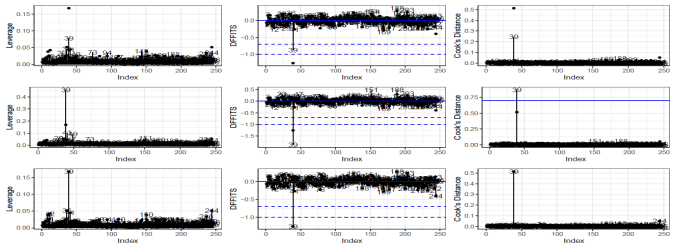
As the table of model performance shows, the three model are almost equally good. Further actions are necessary to determine the best model. A cross validation is thus conducted.

	Model 1	Model 2	Model 3
mean of MSE (standard error of MSE)	4.072 (0.510)	4.044 (0.406)	4.111(0.392)

From the results, we could find Model 2 and Model 3 are slightly better than Model1 due to slightly lower standard error of MSE. So we adopt Model 2 and Model 3. In this way, one can choose to input abdomen and height, or abdomen and weight to compute body fat. Our shiny will compute estimated body fat percentage based on what's input.

Outliers Detection

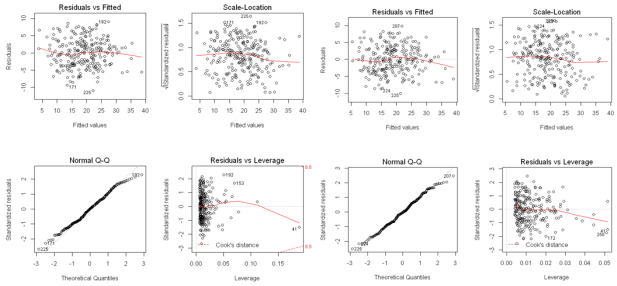
We use three criterions (leverage, dffits and cook's distance) to find the potential outliers.



For each of the three models, point 39 seems to be a leverage point. We look into the 39th observation and find this individual is an extreme sample. Deleting it would not affect the prediction of our model on common individuals. So we remove it from our dataset.

5. Model Interpretation

Model diagnostics



Diagnostics for Model 2 Diagnostics for Model 3

From the diagnostics plot, we could see the two models fit well. The residual fairly follows a normal distribution and the homoscedasticity assumption holds.

6. Conclusion

After deleting 39th observation, our final proposed models to predict bodyfat% are:

$$\begin{array}{lllll} Bodyfat = -21.62 + \dfrac{248324.17}{Abdomen \times Height} + 15.97\dfrac{Abdomen}{Height} + \dfrac{4095.9}{Height} & (1) \\ Std\ error \quad (7.336) \quad (11510) \quad (13.44) \quad (2623) \\ p-value \quad (0.004) \quad (0.032) \quad (0.236) \quad (0.120) \end{array}$$

And

$$\begin{array}{llll} Bodyfat = -41.96048 + 0.89851 Abdomen - 0.12384 Weight & (2) \\ Std\ error \quad (2.502) \quad (0.053) \quad (0.020) \\ p-value \quad (2e-16) \quad (2e-16) \quad (1.51e-09) \end{array}$$

Final model performance:

	Model (1)	Model (2)
Adjusted R square	0.7027	0.7103
Residual Standard error	4.069	4.016
F- statistic	193.2	300.1

Strength:
The model we use is quite simple, which only requires two variables as input but gives a fairly good R square. Moreover, it also provides us some flexibility that you can either input abdomen and height or abdomen and weight.

Weakness:
The confidence interval is kind of wide.

7. Contribution

- **Jiantong Wang:** Data preprocessing, Model 1, Model selection, Summary writing
- **Ruixuan Zhao :** Data preprocessing, Model 2, Slides design, Summary writing
- **Yuchen Zeng :** Data preprocessing, Model 2, Model selection, Summary writing, Shinny design, Github organization
- **Hao Pan :** Data preprocessing, Model 3, Slides design, Summary writing