Bodyfat Summary

Jingyu Ji, Yunwen Jiang, Luwei Liang

1.Introduction

1.1 Motivation and thesis statement

Body fat percentage as a measure of obesity plays an important role in deciding an individual's health condition. There already exists some useful but inconvenient and costly predictive equations for body fat using body circumference measurements. So the goal of this project is to build a simple, robust, accurate and precise equation to estimate body fat percentage.We build a linear function with 2 variables on this data.

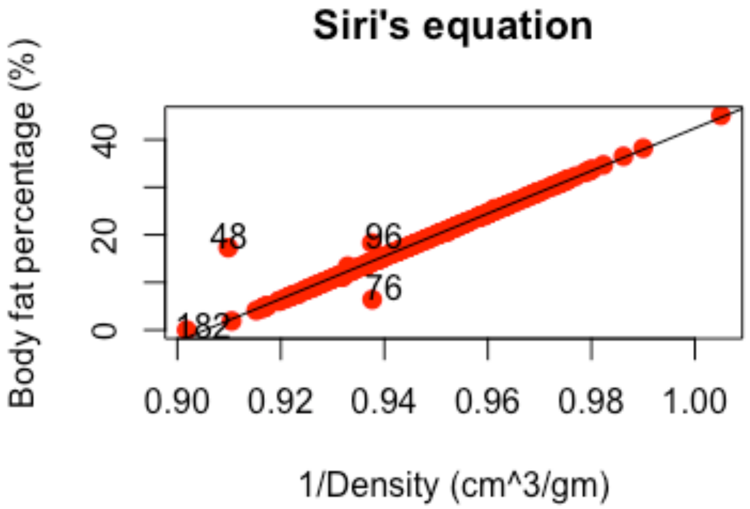
1.2 Background information about data

The data set is based on 252 men. It contains Bodyfat(%), Density(gm/cm^3) and 14 body circumference measurements. Gender is important, but for this data set is all men, no gender difference, The average bodyfat for men is 18-24%. Usually Abdomen is an important indicator.

2.Data cleaning

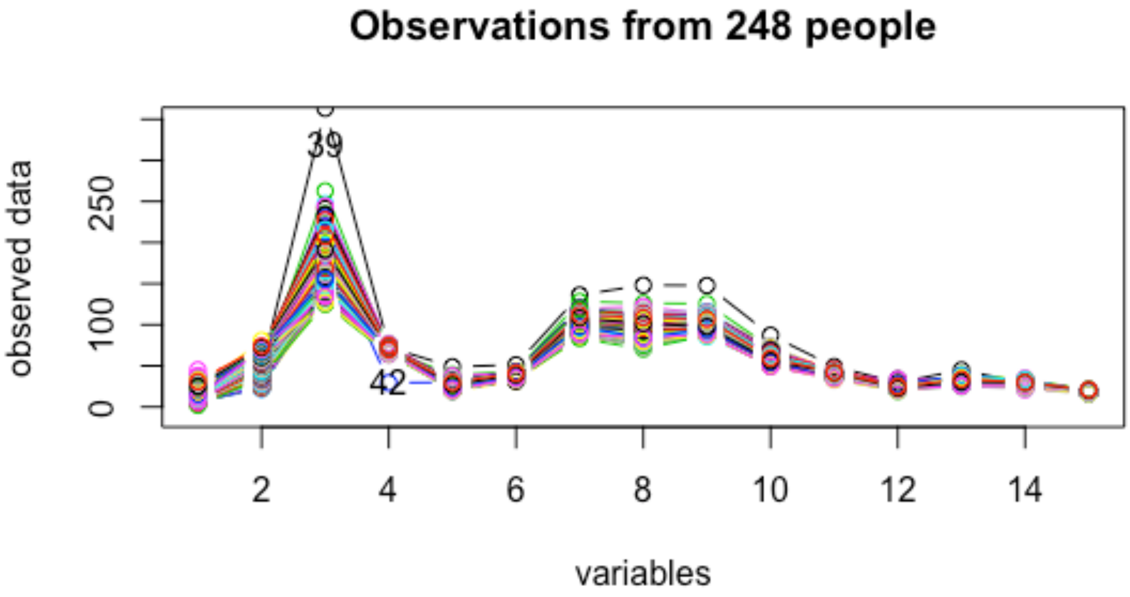
2.1 Check Bodyfat records by Siri's equation

We use the siri's eqaution provided in data description to check bodyfat records. The following plot shows that the **48**th, **76**th and **96**th, **182**nd observations are obviously not on the line(more obvious if image is larger). First we suspect that this mistake comes from computation, so we compute through the equation and results are 14.4, 14.09, 0.37, -3.61 respectively. It is impossible for a person to have a negative body fat. So it is possible that density is wrongly recorded. Therefore, we have no exact evidence to tell which one is wrong, bodyfat or density? In this case we choose to delete them.



2.2 'Stick-out' points

Then we check if there's any weird observations in each variable. We use the following line chart to show these 248 people's records.



From this plot, we can see the value of most variables like weight, height, chest and abdomen of the **39**th observation are extremely large. However, the body fat of 39th is just 33.8, quite small in comparison. It is quite unnormal and definitely causes huge influence on the model, so we remove it. For the **42**nd observation, it has extremely short height 29.5 while other variables are in normal range. We suspect that the value of height may be wrongly recorded, so we use the linear relationship between $HEIGHT^2$ with $WEIGHT/ADIPOSITY$ to check it. The 42nd point turns out to be the outlier. So we fit the linear model based on the dataset to recalculate its height.We also do a simple linear regression to see if there is any observation of strong influence, no one seems to be extremely weird.

3.Variable selection

Too many variables, we use Mallow's Cp, adjusted R^2 , AIC/BIC criteria to help us narrow the selection.

3.1 Selected variables

Method	Number of selected variables	Selected variables
Mallow's Cp	6	Age, Height, Chest, Abdomen, Biceps and Wrist
Adjusted R^2	9	Age, Adiposity, Neck, Chest, Abdomen, Hip, Thigh, Forearm and Wrist
Stepwise AIC	4	Abdomen, Weight, Wrist and Bliceps
Stepwise BIC	3	Abdomen, Weight and Wrist

3.2 Cross Validation

We execute 1000 times, for each time randomly select 200 observations as the training set, and the remaining 47 observations as the test set to test the trained model. We compute the RMSE as an estimate of the prediction error of the unknown data. The results are as follows:

Method	Mallow's Cp	Adjusted R^2	Stepwise AIC	Stepwise BIC
RMSE	4.0312	4.0319	4.0241	4.0431

3.3 Delete Wrist from the BIC model

It is clear that the four models have similar performances. Considered simplicity, we choose the model based on BIC as our basic model. Based on four reasons below we decide to retain Abdomen, Weight and delete Wrist:
1.It is hard to explain the cooperation relationship between Wrist and Abdomen:
(1)We do a single simple regression BODYFAT~ABDOMEN, the R^2 could be around 0.7, which is extremely high for a single variable.
(2)ABDOMEN appears in all four selected models, with large coefficient and broad range at the same time, which shows that Abdomen is absolutely of decisive importance of the whole data set. All the variables are just playing the role of cooperation, adjustment and optimization. However, it is hard to explain the cooperation relationship between Wrist and Abdomen. It is also hard to explain the relationship between Wrist and Bodyfat. Wrist is less informative and intuitively understandable to users.
2.The linear model with variables ABDOMEN+WEIGHT gets relatively larger R^2 than the one with ABDOMEN+WRIST.
3.The linear model with variables ABDOMEN+WEIGHT still performed well in CV based on RMSE: 4.09

4.Model fitting: inference and interpretation

The final model is

$$(BodyFat\%) = -42.45868 + 0.89817Abdomen - 0.12103Weight$$

All the coefficients of the model are significant.The associated $p-value$ is less than $2.2*e^{-16}$, the residual standard error is 4.058 and the adjusted R^2 is 0.7174.

1.For p-value: Since the p-value is small, we can declare that there is a linear relationship between **abdomen, weight** and **bodyfat%**. However, our conclusion carries a 5% error rate where we may have falsely declared that there is a relationship even though there truly isn't a relationship. Correspondingly, we are 95% confident that the interval (0.795,1.000) contains the true slope value of variable Abdomen, the interval (-0.160,-0.082) contains the true slope value of variable Weight.

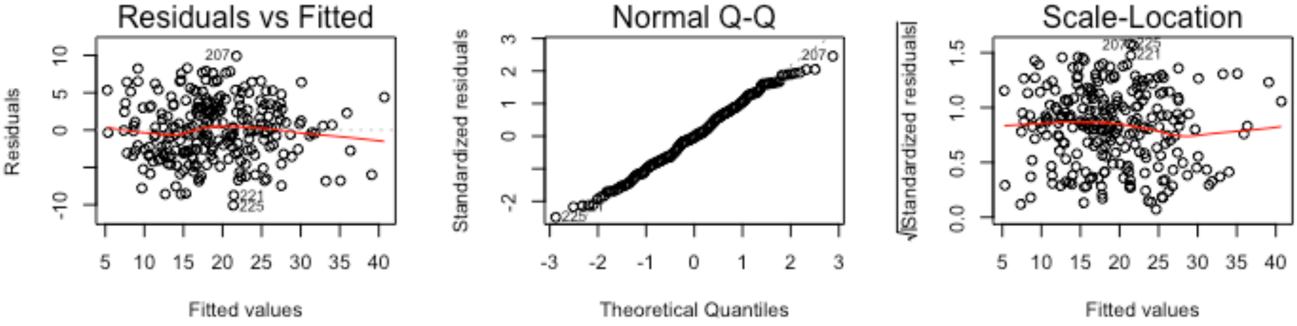
variable	(Intercept)	ABDOMEN	WEIGHT
2.5%	-47.3247658	0.7954474	-0.1600073
97.5%	-37.59259082	1.00088499	-0.08204662

2.The residual standard error is 4.058. It shows the precision of our prediction that our predicted bodyfat is on average 4.058 difference from the original bodyfat. The R^2 represents the two variables we selected could explain about 72% of all the variation in body fat.
3.Model explanation: Each centimeter of abdomen is associated with a 0.9 increase in bodyfat% among men. Each pound of weight is associated with a 0.12 decrease in bodyfat% among men. In general cases, one's percent body fat is mainly determined by the abdomen circumference. For people with the same abdomen circumference, the heavier people will get a lower percent of body fat for they may have more muscle tissue.

5.Model diagnostic

5.1 Diagnostic plots

- **For Residual Plot**: The scatters are randomly dispersed around the horizontal axis and the line is almost straight, so it is appropriate to use linear regression on this data.
- **For QQ plot**: the points generally fall on the 45-degree reference line which indicate residuals follows **normality**.
- **For Scale-location plot**: points are randomly distributed around the horizontal line, i.e., there is **no homoskedasticity violations**.



5.2 Formal Assumption Test

5.2.1 Normality Assumption

H_0: The population is not significantly different from the normal distribution. We do shapiro-wilk normality test and get p-value: 0.2092, which is larger than 0.05,we do not reject H_0, i.e., residual follows **normality**.

5.2.2 Multicollinearity Test

The variance inflation factor of Abdomen and Weight are 4.166731, which are below 5, i.e., there is **no multicollinearity violations**.

5.2.3 Homoscedasticity Test

H_0:The model has constant error variance. We do Non-constant Variance Score Test and get p-value:0.6135939, which is larger than 0.05, we do not reject H_0, i.e., model follows **homoskedasticity**.

6.Conclusion

6.1 Rule of thumb

"multiply your abdomen (cm) by 0.9, minus your weight (lbs) multiply by 0.1 and minus 42"

$$(BodyFat\%) = -42 + 0.9Abdomen(cm) - 0.1Weight(lbs)$$

Example usage: For a people, his abdomen is 90 cm, weight is 170 lbs, his predicted body fat percentage would be around 17.8%. There is a 95% probability that his body fat is between 9.79% and 25.81%.

6.2 Strength and weakness

Our model is simple to understand and implement, users can get useful information. All variables we used are easy to measure.It follows the linear regrssion assumptions, but there still exists the problem that the precision of the prediction is not so high.

6.3 Contributions

Jingyu Ji: Do variables selection, model diagnostic and jupyter notebook summary; **Yunwen Jiang**: Clean raw data, variables selection and jupyter notebook summary; **Luwei Liang**: Clean raw data, model evaluation and shiny app

6.4 References

Body fat percentage, Wikipedia, https://en.wikipedia.org/wiki/Body_fat_percentage (https://en.wikipedia.org/wiki/Body_fat_percentage)