Estimate Percentage of Body Fat Using Clinically Available Measurements: A MLR Method

Ruyan Zhou (rzhou84@wisc.edu) Peibin Rui (prui@wisc.edu) Tianrun Wang (twang494@wisc.edu)

Introduction:

Our main goal is to build a simple, robust model to predict bodyfat based on several body measurements. In this report, we showed the process of finding our best model.

Background Information:

The data we used is a real data set of 252 men with their percentage of body fat and various predictors including age, weight, height, bmi, and various body circumference measurements.

Data Cleaning:

The mean and sd of Y (bodyfat) are 18.94 and 7.75 respectively.

We remove individuals with body fat 0% and 45.1% because 0 body fat is clearly an error sample and 45.1% body fat is not a representative one for the whole population.

We fixed individual with height 29.50 inches by the value calculated from his weight and bmi because his height is too small comparing with his weight (205.00 lbs) and adiposity (29.9 bmi) (both of which are likely a normal person).

We fixed the adiposity of individuals with ID 163 and 221 by the value calculated from their weights and heights because we believe there were some miss calculation.

We removed individuals with ID 39, 41, 86, 31 because we think they are outliers based on our pairs plot and star plot.

Individuals with ID 48, 76, 96 showed violations with Siri's equation. We recognized No.96 as an anomaly because his body density is an outlier in the pairs plot. We imputed the body fat of No.48 and No.76 by the value calculated from their body density using Siri's equation.

Choosing Model:

We first considered a PCA model, but we drop this idea because it is not user-friendly (PCA requires the user to provide all 14 predictors). Finally, we chose to build our model by stepwise selection.

We got 3 candidate models and they are summarized in Table 1. From Table 1, we found all of these 3 models are significant by F-test but the third model is the best one because it has no insignificant coefficients, relatively high R-squared and the lowest residual standard error.

Source	#Predictors	#Insignificant Coefficients	R^2	RSE	p-value
Forward AIC	7	6	0.7342	3.878	<2.2e-16
Forward BIC	3	0	0.7206	3.942	<2.2e-16
Backward BIC	6	0	0.7335	3.875	<2.2e-16

Table 1: Summary of Candidate Models

We also compared the prediction performance of these 3 models by cross validation (divide the whole data set into training set and testing set, fit the model with training set and calculate the mean squared error when the fitted model is applied to the testing set). The results are summarized in Table 2.

Models	Mean squared error
Forward AIC	763.65
Forward BIC	777.93
Backward BIC	757.79

Table 2: Cross Validation

From Table 2, we found that the third model has the best performance of in-sample prediction. Thus, the third model is our final model. The summary of this model is shown in Table 3.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	48.3221	19.0800	2.53	0.0120
AGE	-1.0916	0.4016	-2.72	0.0071
ADIPOSITY	0.5213	0.2240	2.33	0.0208
CHEST	-0.2180	0.0868	-2.51	0.0127
ABDOMEN	0.7083	0.0699	10.13	0.0000
WRIST	-4.8784	1.0535	-4.63	0.0000
AGE:WRIST	0.0631	0.0218	2.90	0.0041

Table 3: Summary of the third model

From table 3, our final model has 6 predictors, all the coefficients are significant at a 5% type I error. The formula of our model is:

$$BODYFAT = 48.32 - 1.09AGE + 0.52ADIPOSITY - 0.22CHEST + 0.71ABDOMEN - 4.88WRIST + 0.06AGE \times WRIST$$

Usage: For example, a man with age 30, adiposity 25, chest 100, abdomen 110, wrist 15

is expected to have a body fat 38.52% based on our model). His 95% prediction interval is between 30.9% and 46.1%.

Model Interpretation:

Table 3 has shown that all the parameters are significant. (We talked about this above).

As age increases by 1 year, the bodyfat percent tends to decrease by (1.09-0.06*WRIST)% in average.

As BMI goes up by 1 unit, the bodyfat percent increases by 0.52% in average.

As CHEST goes up by 1 unit, the bodyfat percent increases by 0.22% in average.

As ABDOMEN goes up by 1 unit, the bodyfat percent increases by 0.71% in average.

As WRIST goes up by 1 unit, the bodyfat percent decreases by (4.88-0.06*AGE)%.

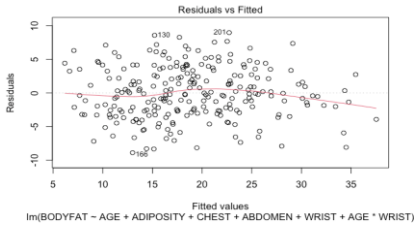
Statistical Analysis:

The overall F-statistics is 109.2 on 6 and 238 df. The p-value < 2.2e-16. And the t-test for each coefficient (shown in Table 3) are all significant, meaning every predictor is significant under 5% type I error.

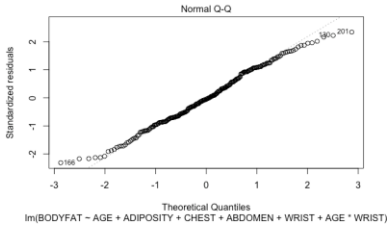
Adjusted R^2 is 0.7267, meaning 73% of the variation in bodyfat can be explained by these predictors.

Comparing with the model in which we delete the interaction, F-test gives a p-value of 0.004. Thus, the existence of interaction is significant.

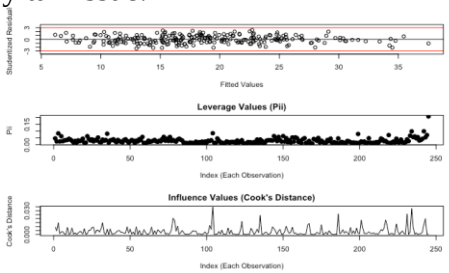
Model Diagnostics:



From the residuals vs Fitted value plot, linearity and homoscedasticity are reasonable since the residual plot shows a random pattern.



Normality also seems reasonable because the 45-degree line fits well. But there might be skinny tail issue.



The studentized residual plot shows there is no outlier. The leverage plot indicates the last observation might be a leverage point. And cook's Distance plot shows no significant influential points.

Model Strengths/Weaknesses:

Strengths:

We can get more accurate estimations of bodyfat compared to simple linear regression like the example of our professor. The model diagnostics indicate that the model satisfies the linear regression assumptions of linearity, homoscedasticity and normality. We take interactions into consideration and this will give us a more accurate and precise understanding of the association of each individual factor with the outcome.

Weakness:

1.The number of indicators is 5, and users are required to provide 6 measurements (ADIPOSITY requires WEIGHT and HEIGHT) which is a little bit complex.

2.Requires specific units (inches, cms)

3.This model is not suitable in calculating the body fat of females or teen males. Because we don't have those samples in our data.

Conclusion:

We found our simple, robust model through the steps shown in our report, which including data cleaning, stepwise selection, cross validation and diagnosing. And our final model works good in predicting body fat.

Contributions: Codes were written by RZ and PR. RZ designed the Shiny App. Summary was made by TW. Slides was made by PR.