Body Fat Percentage Estimation

Tudor Liu, Samuel Tsui, Shirley Wang, William Wang

This version was compiled on November 4, 2023

In pursuit of an efficient method for estimating body fat percentage, this study endeavors to explore a model utilizing various body measurements. The research group seeks to establish a linear model that can calculate body fat percentage through the incorporation of pertinent body metrics.

Introduction. In this report, we delve into the vital subject of Body Fat Percentage Estimation. Our investigative journey is anchored by data sourced from the esteemed Brigham Young University's Human Performance Research Center. The focus is on male body fat percentage - a parameter of paramount importance in the realms of health and fitness.

Data Overview. Measuring body fat percentage accurately is difficult and often expensive. Our dataset includes precise body fat measurements and other accessible metrics for 250 cases across 16 categories. We've simplified the data for clear insights into body fat percentage essentials.

Data Refinement. We refined our dataset by calculating body fat percentage from underwater weighing, omitting duplicate density values. We resolved a redundancy by keeping abdominal over wrist measurements and standardized weight and height to kilograms and centimeters for analytical consistency.

Waist	33.54	32.68	34.61	34.02	39.37	37.17
Abdomen	85.20	83.00	87.90	86.40	100.00	94.40
Abdomen/2.54	33.54	32.68	34.61	34.02	39.37	37.17

Table 1. Redundancy between waist and abdomen

Stepwise variable selection. From the table, forward and backward search selected different variables. Yet, their AIC are very close to each other. Therefore, we decided to do 10-folds cross validation.

	Forward model		Backward model	
Predictors	Estimates	p	Estimates	p
(Intercept)	-32.57	<0.001	5.04	0.547
Abdopmen	0.88	<0.001	0.82	0.547
Weight	-0.24	<0.001		
Wrist	-1.76	<0.001	-1.73	0.001
Bicep	0.24	0.121		
Age	0.06	0.045	0.07	0.017
Thigh	0.18	0.1421	0.22	0.084
Height			-0.11	0.035
Neck			-0.45	0.039
Hip			-0.19	0.135
Forearm			0.30	0.124
Observations	250		250	
R^2/R^2 adjusted	0.742/0.736		0.747/0.739	
AIC	1443.187		1442.736	

Table 2. Forward and backward search with AIC

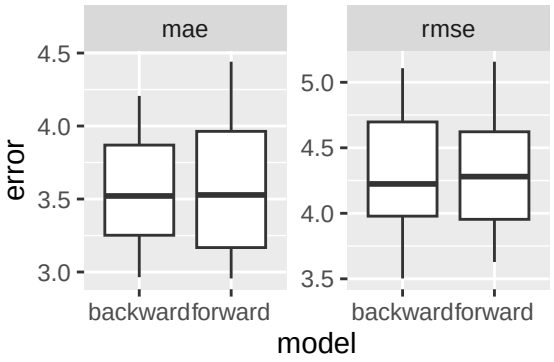

Fig. 1. Manual cross validation result

Boxplot of rmse and mae. The boxplots show the distribution of mean absolute error and root mean square error in the forward and backward model. In both comparisons of mae and rmse, both models give a similar distribution. Even when comparing the mean value of both errors, two models give approximately the same result.

Caret. Therefore, besides working the cross validation manually, we have done it by using Caret as well.
From the caret method, the forward model seems to have smaller errors than the backward model. Although Caret shows a slightly different distribution of mae and rmse, there is still an uncertainty on which model to choose.

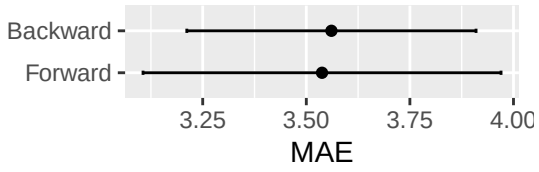

Fig. 2. Caret MAE figure

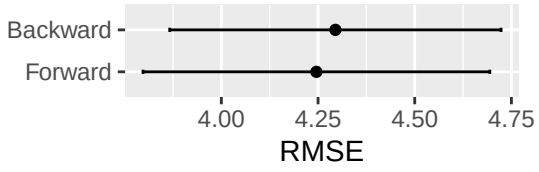

Fig. 3. Caret RMSE figure

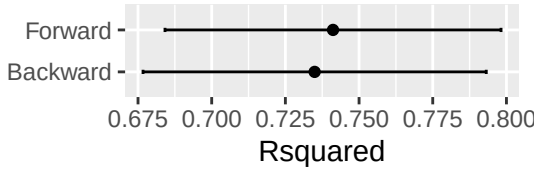

Fig. 4. Caret Rsquared figure

Choose forward mode. The less the variables we need for the model, the easier people can get their own percentage body fat. From table 2, forward model has selected less attributes. To sum up, with the evidence in caret and considering the amount of variables being used, we decided to use the forward model.

Assumption Checking. Before using the model to predict the average body fat percentage. It is pivotal to ensure the model follows Linear regression assumptions.

Independence: the assumption states that all errors are independent of each other. It is usually dealt within the experimental design phase - before data collection , so it is not assessed in the report.

Normality: the normality assumption assumes the errors follow aa normal distribution. In figure 5, the residuals follow normal distribution line, it suggests that the assumption of normality is met.

Linearity: the assumption states relationship between fitted values and percentage body fat are linear. Figure 6 shows a clear linear relationship between body fat percentage and fitted value. Similar trend is observed in figure 7, the residuals are roughly symmetric distributed around the 0 axis , However, the model seems to overestimate percentage body fat for fitted values above 30,it is trivial in this case as the model is not designed to only predict body fat percentage for fitted value above 30 and in overall ,only less than 10 percent of the data lines in that interval. Therefore, the assumption is not violated.

Homoscedasticity: the assumption suggests the error variance should remain constant across all levels of the independent variable, Upon observing figure 3.3, the residuals are roughly evenly spread out Except for residuals with a fitted value between 30 and 35 , suggesting potential heteroscedasticity. However, since we have 250 observations and only 9 of them lie in the fitted value interval between 30 and 35 ,The heteroskedasticity is trivial as it affects a very narrow range of the independent variable (between 30 and 35 less than 5% of the data).Thus, the model does not violate the assumption.

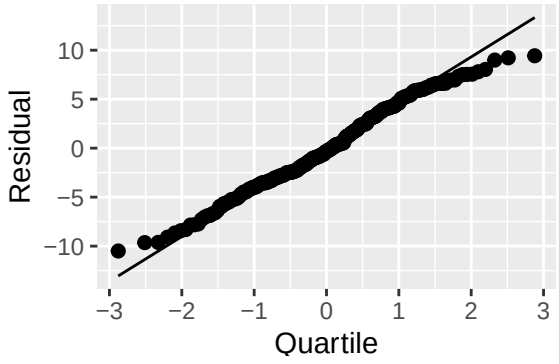

Fig. 5. qqplot of the residuals

Result. As heath becomes a prevailing issue, the report is interested in investigating the body fat percentage by predicting it indirectly through related variables. The regression model for estimating Body fat percentage is: Body fat percentage = -32.57 + 0.88 Abdomen -0.24 Weight -1.76 Wrist + 0.24 Bicep + 0.06 Age + 0.18 Thigh According to the regression found using Step-wise method, a percentage change in the predictors results the coefficient percentage change in the body fat.

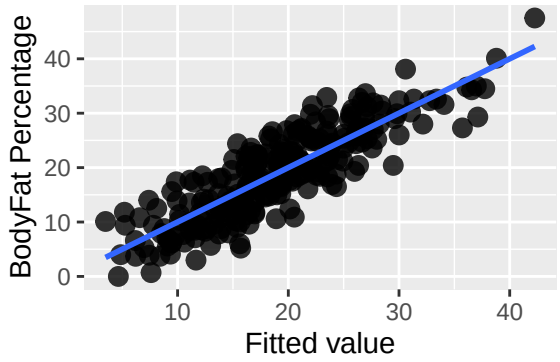

Fig. 6. relationship between bodyfate percentage and Fitted value

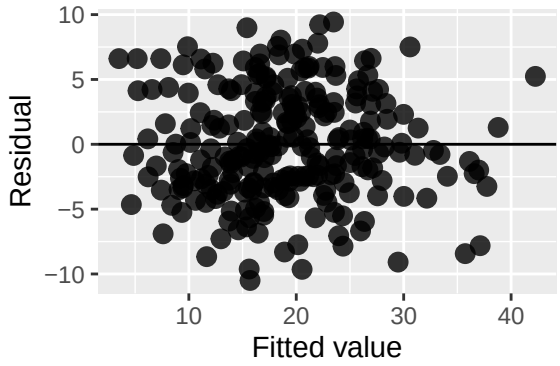

Fig. 7. relationship between residuals and Fitted value

Discussion and Conclusion. One notable limitation of this study pertains to the sample size, which may be considered inadequate for drawing comprehensive generalizations or statistically robust conclusions. The relatively small sample size (250 samples) utilized limits the extent to which the findings can be extrapolated. A larger and more diverse sample size could be considered to enhance the robustness for future researches, enabling a more representative analysis of body composition factors.

For the utilization of Siri's equation, it is imperative to underscore that Siri's equation operates based on a single variable (body density), and the outcomes it provides are heavily contingent on the accuracy of this singular metric. This equation may not fully account for the complexities inherent in specific body composition characteristics.

Another noteworthy limitation of this study revolves around the application of the AIC method. AIC operates under the assumption of certain statistical conditions, such as model linearity, independence of observations, and the absence of multicollinearity, among others. Failure to fully satisfy these assumptions could potentially compromise the accuracy and reliability of model selection, thus influencing the interpretability of the results.

In summary, our linear regression model, with distinct coefficients, shows promise for precise body fat estimation by considering variable impacts. To maximize its potential, addressing limitations and further refinement is crucial. This research contributes to evolving efficient body composition assessment, with broad implications for health, fitness, and science.