**Predicting Body Fat Applying Multivariable Linear Regression**

Jingchen Chai, Ruihan Zhang, Wenjia Zhu

1. **Abstract**

We conducted initial visualization of the body fat dataset with plots and tables, followed by model selection, model interaction, and cross-validation. Model diagnostics was also applied to check whether the model needs any transformation before omitting the outliers. We then select an optimal model that can best predict the body fat.

1. **Introduction**

Body fat is an important factor in many medical situations. Measuring body fat can help to assess whether a person is at a healthy weight. Additionally, obese or overweight can be identified by calculating the body fat. A normal body fat is between 25 and 30 percent in women and 18 and 23 percent in men (PennMedicine.). More than 68% of US adults are considered overweight, and 35% are obese(Flegal et al., 2010). While the number of people who are overweight or obese are gradually increasing, more people are becoming more concerned about this issue. The goal of this project is to predict the body fat in different patients based on their physical measurements.

1. **Methods**

***3.1 Data Exploratory Analysis***

We have chosen bodyfat\_siri as our final target variable. The goal of this part is to have a sense of the distribution of each variable. Correlation was also considered to see whether any variables are highly correlated with each other so that we can try to eliminate the multicollinearity effect. In order to visualize the above information, a descriptive statistics table was made, a boxplot for the target variable, histogram for the predictors, and a correlation plot. Multicollinearity will also be checked by calculating the VIF.

***3.2 Model Selection***

In order to select the most appropriate model, we used automatic procedures (Stepwise Regression) as followed for comparison between bigger and smaller models.

***3.3 Model interaction***

To see if there are interaction effects between any two independent variables, separate regression models were graphed. If the graph showed paralleled lines, then it means there are no interaction effect. Different slopes suggests that interaction might be present. The main effects of these two variables should be taken into consideration together in the model. After forming the interaction, we would use ANOVA to confirm whether the large model is superior to the original model without interaction.

***3.4 Model Validation***

A 5-fold cross-validation will be constructed to test the validity of the model. RMSE(root mean squared deviation), R square, and adjusted R square will be used to evaluate the model. The model with lower RMSE and higher R square will suggest a better fitness of the model.

***3.5 Model Diagnostics***

After obtaining the model that we have chosen, several plots will be made to check whether the model meets the regression assumptions. The QQ plot will be used to determine the normality of the data. Outliers in the data will also be identified by plotting the residuals vs leverage plot and also calculating the Cook's distance, which can show the influence of a specific case on all fitted values. Finally, a model without outliers will be shown and will be compared to the original model to examine the influence of the outliers. The influential outliers will be omitted in the final model.

1. **Results**

***4.1 Data Exploratory Analysis***

The table (Table 1 ) has shown the mean, standard deviation, maximum, minimum, and IQR for the 14 variables. After plotting a boxplot for our target variable, we noticed there exists an outlier, which we will be considering in our subsequent analysis. The 13 histograms (Figure 2) for the predictors, many of them performed a normal distribution. Only a few of them performed slightly right-skewed. Therefore, we made an inverse transformation for some of the variables. The improved transformation is shown in Figure 3. We also applied a correlation plot (Figure 4) to observe correlation between independent variables. The plot shows that many of them are highly correlated with each other. In order to verify this, we calculated the variance inflation factor for each of them(VIF), which a value a value between 1 and 5 indicates a moderate correlation, and value greater than 5 indicates a severe correlation. Figure 5 shows the VIF for each of the predictors, we remained i\_abdomen as it performed significant in the multiple linear regression, and removed variables i\_weight,i\_chest,i\_hip, and i\_thigh.

***4.2 Model Selection***

The model we obtained by Stepwise Regression is bodyfat\_siri = 1.257e+02 + 5.425e-02 \* age - 4.106e-01 \* height + 6.568e+02 \* i\_neck - 6.876e+03 \* i\_abdomen + 3.315e-01 \* forearm - 1.741e+00 \* wrist.

***4.3 Model interaction***

We made the interaction terms between forearm and height, age and i\_neck, wrist and age, which show significant interaction(Figure 6). After selecting the three models, we used ANOVA to compare the original model and the three adjusted models with the interaction terms, in which the original model is nested in the three adjusted models. Based on the pvalues given in each result, we reject H0 and conclude that model 2 and 3 are superior than the original model.

***4.4 Model Validation***

Cross-Validation was used by folding the data 5-fold and testing the model using 1/5 odf data. compare the two above models. The model with lower RMSE and higher R square will be selected as optimal due to small deviation and good fitness (Table 2). Our new updated model will be : bodyfat\_siri ~ age + height + i\_neck + i\_abdomen + forearm + wrist + age\*i\_neck

***4.5 Model Diagnostics***

Next, we operated diagnostics by the following diagnostic plots for checking the adequacy of the regression model (Figure 7). As shown in the Residuals vs Fitted plot, residual values bounce around 0 and residuals form a generally horizontal 'band' around zero, showing no unequal error variance. There are three observations #192, #224, #225 standing out from the random pattern, making them potential outliers. According to the QQ plot, which is almost straight, exhibiting a nice normality without heteroscedasticity. #192, #224, #225 still stand out again from the random pattern in the QQ plot. The Scale-location plot is almost a horizontal line with equally spread points, confirming the assumption of equal variance. In the Residuals vs Leverage plot, observations #153, #175, #216 appears to be outlying values with high statistics in Cook's distance D (Figure 8). Next, we compared the model with and without the three observations. It came out that though excluding three values slightly increased R^square by 0.002, but all variables in the model turned out to be statistically significant (Table3). Therefore, we decide to exclude outliers #153, #175, #216 from our analysis. Overall, these plots look like this model is fitting the data well.

1. **Conclusion/Discussion**

In conclusion, we established a multiple linear regression model to predict the body fat based on 6 variables. Interaction between age and the inverse of neck circumference, which means that the inverse of neck circumference in different ages may have different effects on the body fat. The R square is 0.7522, which means that it can explain 75.22% of the variance of the dependent variable, which is a quite good result.

**Figures and Tables**

Figure 1 : Distribution of bodyfat\_siri

Chart, box and whisker chart

Description automatically generated

Figure 2 :Distribution of Predictors Before Inverse Transformation

Chart

Description automatically generated

Figure 3: Distribution after Inverse Transformation

Diagram, timeline

Description automatically generated

Figure 4: Correlation Plot

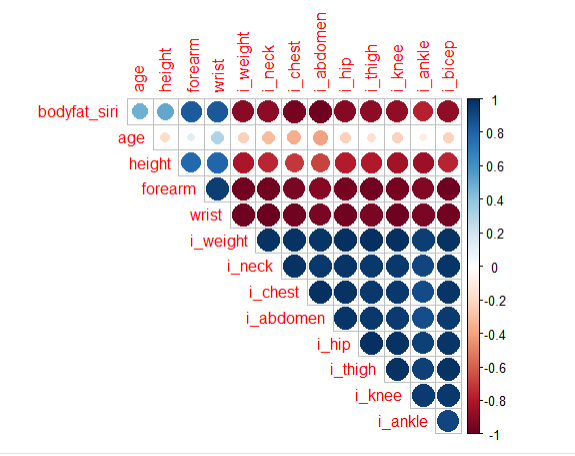


Figure 5: VIF for the Predictors

Chart

Description automatically generated

Figure 6: Interaction Plots

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

Figure 7: Model Diagnostics

Chart, diagram

Description automatically generated

Figure 8 : Cooks Distance

Chart

Description automatically generated

Table 1: Descriptive Statistics for all Variables

A screenshot of a computer

Description automatically generated with low confidence

Table 2:

Text, letter

Description automatically generated

Model 2

Text, letter

Description automatically generated

Model 3

Table 3: Summary Statistics for the Updated Model

Table

Description automatically generated

**Reference**

[1] https://www.pennmedicine.org/for-patients-and-visitors/find-a-program-or-service/bariatric-surgery/who-is-a-candidate/weight-loss-and-obesity-facts#:~:text=Obesity%20Definition%20and%20Criteria,body%20fat%20are%20considered%20obese.

[2] Flegal, K.M., Carroll, M.D., Ogden, C.L., and Curtin, L.R. (2010). Prevalence and Trends in Obesity Among US Adults, 1999-2008. JAMA 303, 235-241.