Report - Tuesday Group 4

Yuchen Xu, Mario Ma, Yiteng Tu, Yudi Wang

I. **Introduction/Motivation**: In this project, we used the dataset provided by Dr. A. Garth Fisher. With age, weight height and 10 body circumference measurements recorded from 252 men, we built up several models and found the best one to predict the value of body fat. We hope this model and our prediction can help to explain the true body fat accurately.

II. **Background Information/Data Cleaning**:

1.The data contains in total 17 columns, including IDNO, which we deleted in our experiment. The relationship between BODYFAT and DENSITY has been given by 'Siri's equation', so we used it to check if this two columns of data is correct and DENSITY is not included in our final model.

2.a. Clean Procedure: 1.We used histogram and LOF score to find outliers. 2.With KNN and the relationship between HEIGHT, WEIGHT and BMI, we found and replaced some of the wrong data points in these three columns. 3. We checked the points out of the normal range according to the internet. 4.We used KNN and the relationship between BODYFAT and DENSITY to fix these two columns. 5.we used KNN to replace other abnormal points in ANKLE column.

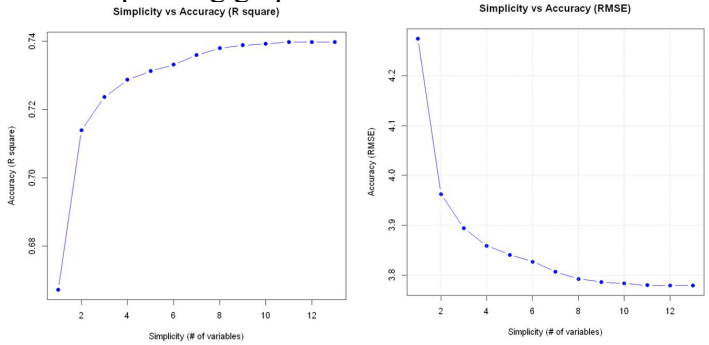
b. Fixed points:

Abnormal	Abnormal column	Original value	Replaced value	Imputation method
31	ANKLE	33.9(cm)	23.8(cm)	KNN
42	HEIGHT	29.5(inches)	69.43(inches)	BMI and WEIGHT calculate
48	BODYFAT	6.4(%)	14.14(%)	DENSITY calculate
76	BODYFAT	18.3(%)	14.09(%)	DENSITY calculate
86	ANKLE	33.7(cm)	22.5(cm)	KNN
96	DENSITY	1.0991(gm/cm^3)	1.059(gm/cm^3)	BODYFAT calculate
172	BODYFAT,DENSITY	1.9(%),1.0983(gm/cm^3)	6.99(%),1.083(gm/cm^3)	Regression by other columns
182	BODYFAT,DENSITY	0(%),1.1089(gm/cm^3)	4.53(%),1.089(gm/cm^3)	Regression by other columns
221	WEIGHT	153.25(lbs)	173.20(inches)	BMI and HEIGHT calculate

c. Leverage and Cook's distance check using full linear regression model: Deleted row 39,54,163,175,216.

III. **Final Model**

We applied the best subset selection using the *leaps* package in *R*, with the default criterion being adjusted R^2. This provided the optimal subsets corresponding to 1 to 13 predictors. To balance model simplicity and accuracy, we evaluated the accuracy of these linear regression models using the leave-one-out method, with root mean squared error (RMSE) and R^2 as the evaluation metrics. The corresponding graphs are listed as follows.



We found that using two variables to fit the model led to a significant improvement in both RMSE and R-squared compared to using just one variable. However, the improvement was not as pronounced as the model complexity increased from two to thirteen variables. Therefore, we selected the best subset with two variables as our final model, which is,

BODYFAT = - 41.547 - 0.124*WEIGHT (lbs) + 0.894*ABDOMEN (cm).

It means: for every centimeter increase in abdomen, the predicted body fat will increase 0.894%, and for every pound increase in weight, the predicted body fat will decrease 0.124%. The second interpretation may seem weird, but it's actually reasonable, for the increase of weight will lead to the increase of abdomen, contributing to the increase of body fat ultimately.

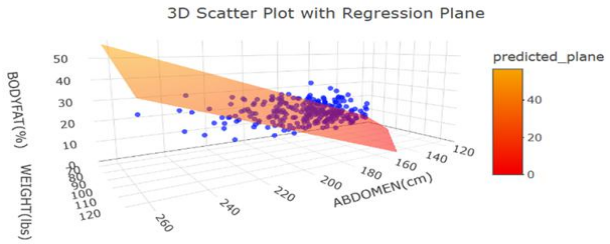
IV. Model Diagnostics

We used the Shapiro-Wilk test to check the normality of the residuals, the Durbin-Watson test to examine the independence of the residuals, the Breusch-Pagan test to assess homoscedasticity, and the Variance Inflation Factor (VIF) to detect multicollinearity. Our model passed all these tests, so we believe it satisfies the fundamental assumptions of linear regression.

V. Statistical Analysis

Firstly, we examined the overall significance of the model and the significance of each variable by checking the F-statistic and t-statistics. We found that both the overall model and each variable within it are significant at the 0.001 level, implying that our model can effectively explain the variation of dependent variables, and each variable is meaningful. Additionally, our residual standard error is 3.987, meaning that the average difference between the predicted value of our model and true value is less than 4% body fat. The R-squared of our model is 0.712, indicating that it can successfully explain over 70% of the variation.

Secondly, we focused on prediction effect. Based on the available data, our model successfully predicts 60% of the data within a 5% error margin. The related three-dimensional scatter is attached as follows for visualization. From the plot, we verified that our model does not consistently overestimate or underestimate body fat.



Finally, we assessed the robustness and sensitivity of our model. We introduced Gaussian noise with 2.5% standard error the corresponding mean to both independent and dependent variables, that is, allow for a 5% margin of error. Using our model to predict the data with added noise, we obtained a RMSE of 4.22, indicating that the discrepancy between the true values and predictions is 4.22% body fat. Compared to the data without noise, this difference only increased by 0.3% body fat, leading us to conclude that our model is not sensitive to errors and noise.

VI. Model Strengths & Weaknesses

As a linear model with only two variables, our model is simple, easy to interpret and visualize, and satisfies all the fundamental assumptions of linear regression. Additionally, it demonstrates a high level of predictive accuracy and robustness against noise. However, our model is trained on male data, which may need further transfer learning to expand this analysis to women.

VII. Conclusion

Through data cleaning and the selection of the best subset of variables, we established a linear regression model using weight and abdomen as predictors to forecast body fat. After testing the model and conducting statistical analysis on the prediction results, we believe that our model, "BODYFAT = - 41.547 - 0.124*WEIGHT (lbs) + 0.894*ABDOMEN (cm)", can accurately and robustly predict male body fat based on the data from weight and abdomen.

Contributions:

	Yuchen Xu	Mario Ma	Yiteng Tu	Yudi Wang
Code	Responsible for data cleaning code; Reviewed model selection, construction, diagnostic and plotting code	Responsible for data cleaning code; Reviewed model selection, construction, diagnostic and plotting code	Reviewed data cleaning code; Responsible for model selection, construction, diagnostic and plotting code	Reviewed data cleaning code; Responsible for model selection, construction and diagnostic code
Summary	Introduction, data cleaning sections	Introduction, data cleaning sections	Review	Final model, model diagnostics, statistical analysis, model strengths&weaknesses, conclusion sections
Shiny App	Frame building and basic function realization	Beautification	Review	Review
Presentation	Data cleaning	Conclusion	Statistical analysis, robustness and sensitivity, strengths and weaknesses	Final model, model diagnostics

References:

[1] Bailey, Covert (1994). Smart Exercise: Burning Fat, Getting Fit, Houghton-Mifflin Co., Boston, pp. 179-186.

[2] Behnke, A.R. and Wilmore, J.H. (1974). Evaluation and Regulation of Body Build and Composition, Prentice-Hall, Englewood Cliffs, N.J.

[3] Siri, W.E. (1956), "Gross composition of the body", in Advances in Biological and Medical Physics, vol. IV, edited by J.H. Lawrence and C.A. Tobias, Academic Press, Inc., New York.

[4] Katch, Frank and McArdle, William (1977). Nutrition, Weight Control, and Exercise, Houghton Mifflin Co., Boston.

[5] Wilmore, Jack (1976). Athletic Training and Physical Fitness: Physiological Principles of the Conditioning Process, Allyn and Bacon, Inc., Boston.