BODY FAT STUDY SUMMARY

Introduction

Using a multiple linear regression model with a real dataset of 252 men to predict body fat.

Data Cleaning

1. View the characteristics of variables, including mean, max, min, median, and quartiles.
   a) *Bodyfat*(DV): has two values(ID 172, 182) less than 2%(Essential fat).
   b) Independent Variable: the minimum of *height* is abnormal(ID 42). *adiposity* can be calculated from *height* and *weight*.
   c) Measurements: *weight* and *height* are in non-metric units.
2. Convert units and impute outliers.
   a) Convert *weight* and *height* to metric units.
   b) Cross-validate *adiposity*, *height*, and *weight*. Replace abnormally low *height*(ID 42) with calculated results from *weight* and *adiposity*; substitute significantly differing *adiposity* results with those calculated from *height* and *weight*(ID 163, 211).
   c) Exclude two abnormal values, fit a linear model using *density* and *bodyfat* to identify outliers(ID 48, 76). Impute these two data points using the linear fit results.
   d) Impute the two abnormal low *bodyfat* values(ID 172, 182) using KNN.

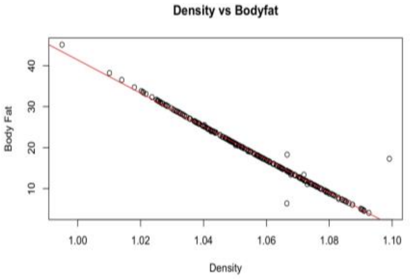

Figure 1 Before Imputation

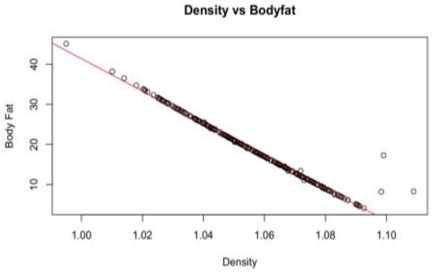

Figure 2 After Imputation

Motivation for Model

1. Final Model:

$$bodyfat = -25.02 + 0.92abdomen - 0.25weight_{inkg} - 1.16wrist + 0.47forearm - 0.36neck + \epsilon$$

2. Example Usage: A man with 88.6cm *abdomen*, 89.92kg, 19.2cm *wrist*, 30cm *forearm* and 42.1cm *neck* is expected to have a body fat % of 11.23%, while the real data is 12%. His 95% prediction interval is between 9.62% and 12.84%.
3. Interpret Model: Our estimated coefficients are -25.02, 0.92, -0.25, -1.16, 0.47, and -0.36, in the units of %, cm, kg, cm, cm, and cm. For instance, the coefficient for wrist is -1.16, meaning that for every 1 cm increase in wrist circumference, body fat percentage is predicted to decrease by 1.16 percentage points, holding other variables constant.
4. Choose Model:
   a) We use model selection(Forward selection, Backward selection, Stepwise selection).
   b) We use lasso regression.
   c) Due to the low rate in some age groups, we used weighted least squares, adjusting weights based on 2024 global male age distribution. We grouped the data in 5-year interval and calculated the proportion of each group within the 20-85 range for the sample and real data. The ratio of true to sample proportions was used as weights.
   d) Use leave-one-out cross-validation to calculate the average MSE, and select the model with the smallest MSE based on the checking model assumptions.

Table 1 Choose Model

Model	MSE(LOOCV)	AIC	BIC
LASSO	17.9561	717.5898	738.7663
OLS (Backward)	15.9935	1412.9470	1455.3002
OLS (Forward)	16.0438	1414.0327	1438.7387
OLS (Stepwise)	15.9935	1414.0327	1438.7387
Weighted(Backward)	16.2674	1459.9384	1484.6444
Weighted (Forward)	16.6847	1459.3374	1484.0434
Weighted (Stepwise)	16.2674	1459.3374	1484.0434

Hypothesis Testing

1. We use the t-test to assess the significance of each independent variable and the F-test to evaluate the significance of the model. Detailed values are showed in Table 2.
2. We found our R^2 is 0.7375, adjusted R^2 is 0.7321. The model explains approximately 73.75% of the variability in the dependent variable.
3. We calculate the CI, detailed values are showed in Table 2.

Table 2 Hypothesis Testing Results

	Estimate	Std.Error	t value	2.5%	97.5%	P-value	
(Intercept)	-25.01713	6.96761	-3.59	-38.7409	-11.2933	0.000398	***
ABDOMEN	0.92274	0.05129	17.991	0.8217	1.0238	< 2E-16	***
WEIGHT_inkg	-0.24749	0.05189	-4.77	-0.3497	-0.1453	3.16E-06	***
WRIST	-1.15995	0.4262	-2.722	-1.9994	-0.3205	0.006961	**
FOREARM	0.46668	0.16745	2.787	0.1369	0.7965	0.005737	**
NECK	-0.35585	0.20157	-1.765	-0.7529	0.0412	0.078747	.
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1							
F-statistic: 138.2 on 5 and 246 DF, p-value: < 2.2e-16							

Model Diagnostics

1. Residuals: Our model satisfies the assumptions of MLR as the residuals plot confirmed both linearity and homoscedasticity, the Q-Q plot validated the normality of residuals.

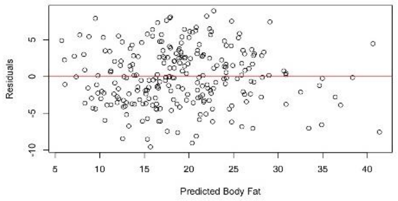

Figure 3 Residual plot

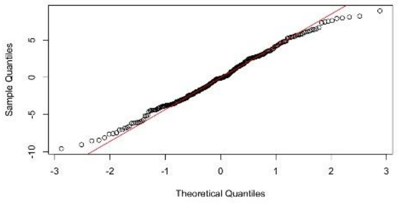

Figure 4 Residual QQ plot

2. Collinearity: The VIF test showed no multicollinearity problem among the variables. The biggest VIF is 7.7398, which is smaller than 10.

Model Strengths and Weakness

1. Strengths: Our model includes five easily measurable variables, making it simple and interpretable. It satisfies the assumptions of homoscedasticity and no multicollinearity, supporting the validity of our results in predicting body fat.
2. Weaknesses: Our model may overfit, relies on linear relationships, and struggles with extreme low body fat values.

Future Work

Handling Extreme Body Fat Values: We will explore local regression for better handling.

Conclusion

Based on data cleaning, we choose to perform multiple linear regression with forward selection to obtain a simple and effective model that includes five independent variables.

References

Neter, John, William Wasserman, and Michael H. Kutner. *Applied Linear Regression Models.* 4th ed., McGraw-Hill, 1996.

Johnson, Roger W. "Fitting Percentage of Body Fat to Simple Body Measurements." Journal of Statistics Education, Mar. 1996, https://doi.org/10.1080/10691898.1996.119105.

"Population Pyramids of the Nations of the World 2024 - Population Pyramids." *Population Pyramids of the Nations of the World 2024 - Population Pyramids,* population-pyramid.net/. Accessed 14 Oct. 2024.

Contributions

Contributions	Minyuan Zhao	Meiyi Yan	Siyu Wang
Presentation	1) Responsible for slide 10-15 presentation. 2) Responsible for feedback.	1) Responsible for slide 1-6 presentation. 2) Responsible for feedback.	1) Responsible for slide 7-9 presentation. 2) Responsible for design the slides.
Summary	1) Responsible for ideas of the whole summary. 2) Responsible for Checking errors.	1) Responsible for Data visualization and tables. 2) Responsible for Model Diagnostics, Model Strengths and Weakness, Future Work.	1) Responsible for ideas of the whole summary. 2) Responsible for page 1-3, except from Model to Future Work. 3) Responsible for typesetting.
Code	1) Responsible for model selection (forward-selection, backward-elimination, stepwise). 2) Responsible for lasso regression.	1) Responsible for data cleaning. 2) Responsible for code merge. 3) Responsible for model picking. 4) Responsible for LOOCV.	1) Responsible for weighted least squares. 2) Responsible for model selection (forward-selection, backward-elimination, stepwise).
Shiny App	1) Responsible for building a framework. 2) Responsible for Beautification and interface design. 3) Responsible for testing results.	1) Reviewed/edited and provided feedback on Shiny app. 2) Responsible for Checking errors.	1) Reviewed/edited and provided feedback on Shiny app. 2) Responsible for Checking errors.