1 Motivation & Data Cleaning

Our group has developed a body fat prediction tool to simplify body fat estimation without using specialized body fat testing equipment. We achieve this by training a model using existing data. Upon pre-reviewing the raw data, we find that there are no duplicated or missing values. But 4 outliers are detected by interquartile range method[1]. In this step, we change the lower bound of 'BODYFAT' based on the data from the American Council on Exercise[2]. The outliers in 'BODYFAT' (0.0%, 1.9% and 45.1%) are deleted, this decision is based on the calculation of following formula $100 \times Bodyfat = \frac{495}{density} - 450$. For outlier in 'HEIGHT'(29.5inches), by the result of the calculation of $BMI(ADIPOSITY) = \frac{weight_{lbs} \times 703}{height_{inch}}$, 'HEIGHT' $= 29.5$ we deduce that the value may be a typo and we replace this data as 69.50.

2 Final Models

We provide 2 final models: AIC model and 3-features linear regression model. AIC model has the highest adjusted R^2 and smallest RMSE on the training set, but the complexity is high. 3-features model chooses ABDOMEN, WRIST, WEIGHT features for the model building, the coefficients are 0.915, -0.645, -0.132 respectively. AIC models chooses 11 features, including AGE, WEIGHT, ADIPOSITY(BMI), WRIST etc. Table 1 below shows 2 samples' body fat using final models to do the prediction. The prediction near the mean of body fat has smaller prediction error. Figure 1 shows the prediction error of AIC model and 3-features model on the test set.

Table 1: Model Performance

sample	True value	AIC	3-features
Sample A	21.7%	22.753%	21.381%
Sample B	9.6 %	16.274%	15.145%

3 Model Building

Training and the test dataset are split by 70% and 30%. We use the difference between the true body fat and the predicted value to evaluate model performance. A notable degree of collinearity is observed among features by using correlation matrix.

To deal with collinearity, we conducted Principal Component Analysis (PCA) on the explanatory variable data matrix. First three main components: PC1 explained 64% of the variance, revealing a substantial inverse relationship with 'WEIGHT'; PC2, explaining 10% of the variance, showed a positive association with 'HEIGHT.'; PC3 (explaining 7% of the variance) further underscored the positive relationship between 'HEIGHT' and this component.

Initially, we train a linear model using all the features. 5-level 5-fold cross-validation is used to increase the robustness of the model. We apply AIC and BIC in variable selection, and we select 3 features: "ABDOMEN", "WRIST" and "WEIGHT" to build a simple regression model. We also use regression tree and random forest in prediction. Regression tree predicts body fat by clustering and the random forest is much more complex and time-consuming.

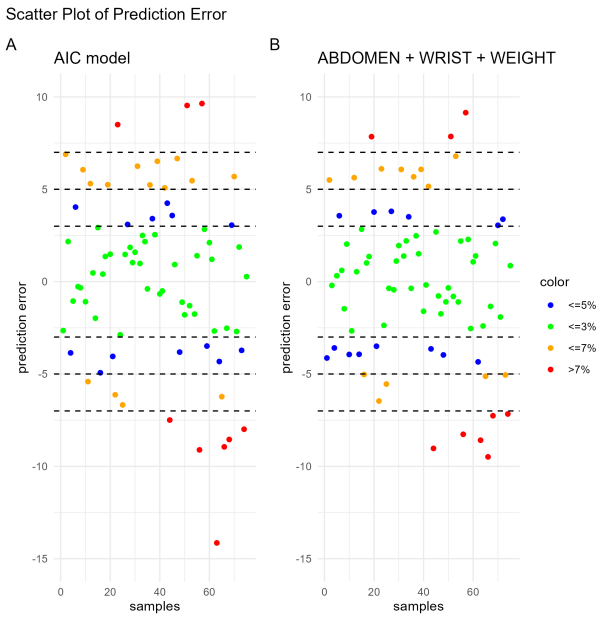

Figure 1: Prediction Error

By accessing the models' performance on the training set and making the trade-off between the complexity and the accuracy, we choose AIC and 3-features model as our final model. Table 2 below shows the performance of each model on training set using R^2 and RMSE and the ratio of samples whose absolute error is less than 3%, 5%, 7% on the test set.

Table 2: Model Performance

model	adjusted R^2	RMSE	$\pm 3\%$	$\pm 5\%$	$\pm 7\%$
All features	0.728	3.98	0.493	0.667	0.840
AIC	0.751	3.743	0.507	0.680	0.880
BIC	0.739	3.826	0.467	0.747	0.853
Abdomen+ Wrist+Weight	0.729	4.010	0.520	0.707	0.880
PCA	0.749	3.758	0.413	0.680	0.760
Regression Tree	0.633	4.708	0.400	0.693	0.787
Random Forest	0.684	4.236	0.507	0.693	0.840

4 Discussion and Improvement

Given the summary of the dataset and the background information, we know that those data were collected decades ago and the age group is bias[3] . Besides, gender has a strong impact on the body fat[4], but all the samples are male. To make our model more robust and accurate, increasing the sample size is important. We should consider collecting some variables about diet, exercise and gender that are highly related to the level of body fat.

Reference

[1] Hollander, M., & Wolfe, D. A. (1999). Nonparametric statistical methods (Vol. 119). John Wiley & Sons.
[2]Ace Fit: Percent body fat calculator. ACE Fit — Percent Body Fat Calculator. (n.d.). https://www.acefitness.org/resources/everyone/tools-calculators/percent-body-fat-calculator/
[3] Jiehua Lu, Yuchen Gu. Characteristics of population change in the United States since the 1950s and its implications for China [J]. Population Journal,2022,v.44; No.251(01):46-57
[4] Blaak, Ellen. Gender differences in fat metabolism. Current Opinion in Clinical Nutrition and Metabolic Care 4(6):p 499-502, November 2001.

Table 3: Member's Contribution

Contributions	**Wanxin Tu**	**Ziming Li**	**Baiheng Chen**
Presentation	Responsible for Slide 1, 6, 7, 9	Responsible for Slide 2, 3, 8, 9	Responsible for Slide 4, 5, 9
Summary	Responsible for model building and the discussion part	Responsible for introduction and the data cleaning part	Responsible for tables and the plot; Responsible for final models part
Code	Responsible for PCA part	Responsible for Data Cleaning and Regression Tree part	Responsible for all the other codes
Shiny app	Reviewed and provided feedback on Shiny app	Responsible for Shiny app	Reviewed and provided feedback on Shiny app

Chatgpt was used in the coding part.