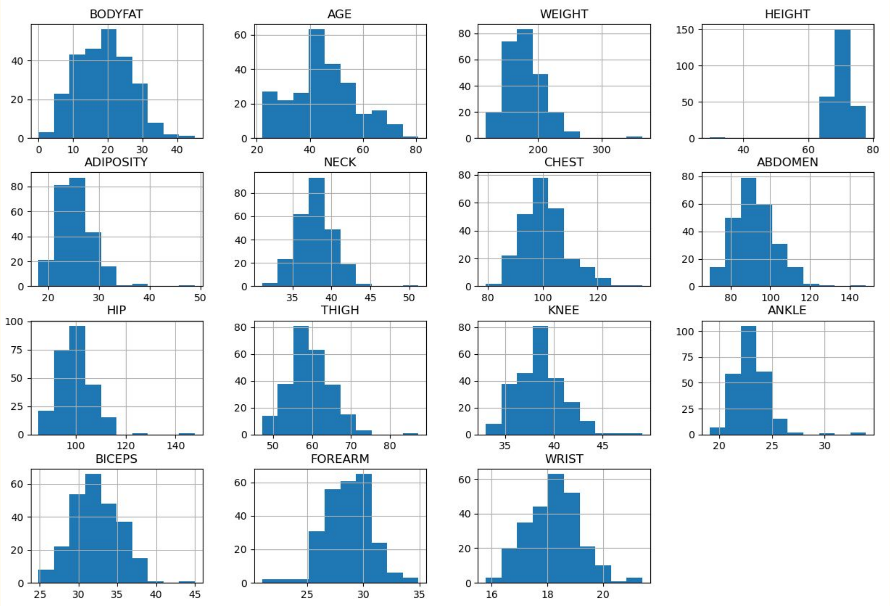
Body Fat Prediction Project

STAT628 - Module 2

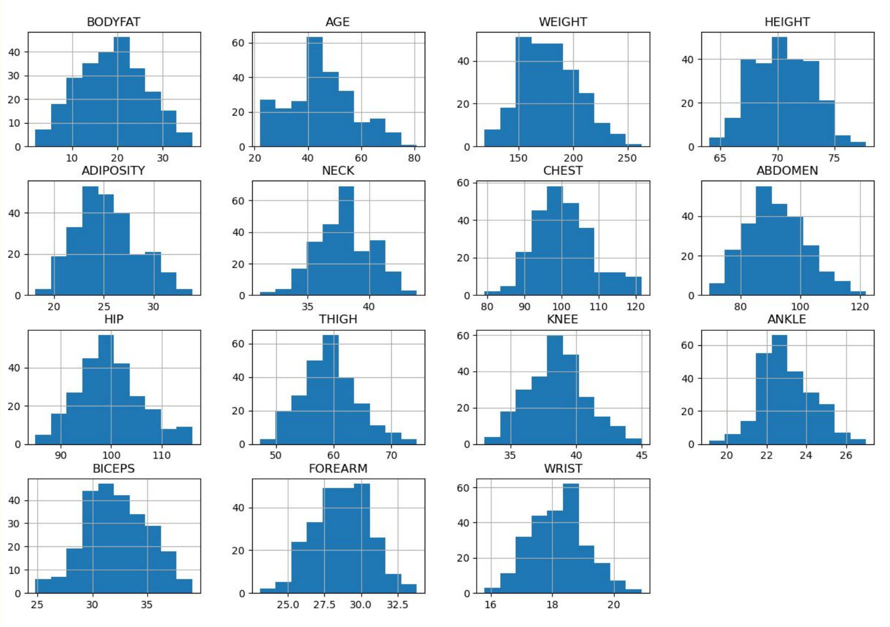
Group 13: Shan Leng, Siyan Wang, Ruotong Zhang

10/18/2023

Data Preview



Before cleaning



After cleaning

Data Cleaning

Body Fat:

Individual	Original body fat (%)	Imputed body fat (%)
182	0	19.62
216	45.1	17.97

Eg:

Mean of body fat within range:

[182’s age - 5, 182’s age + 5] excluding 182

Outlier Detection for Features:

- Z-score with a threshold value of 3

Imputation Method:

- Mean of feature within the age range of [outlier age - 5, outlier age + 5]

Final cleaned data: 252 samples with 14 features.

Dataset Segmentation

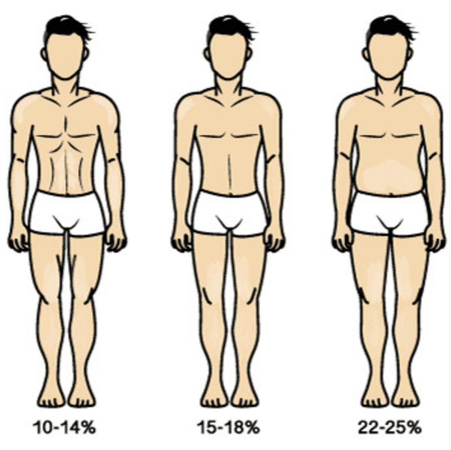
Train/Test Split

- Test: Randomly select same amount of samples from 3 age groups:
  [20-40), [40-60), [60, 81]
- Training: The remaining data automatically becomes the training set.
- Training : Test = 8:2.

Body Fat Grouping

In Shiny app, display prediction results by body fat group.

Arrange(%)	≤15	15<x≤20	20<x≤30	>30
Group	Athlete	Fitness	Normal	Obsess



Selection Criteria & Candidate Models

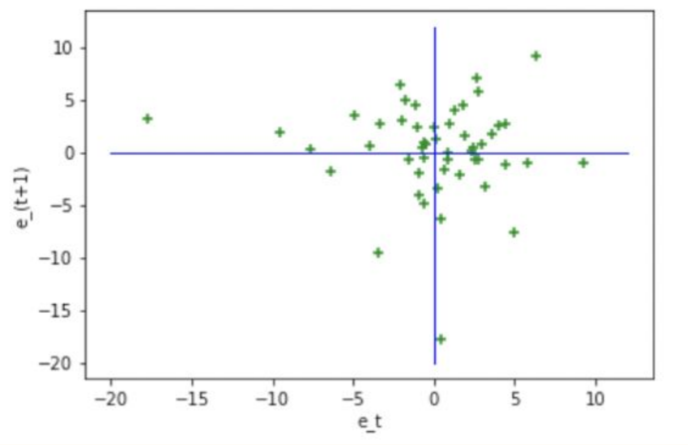
- Accuracy
  - Root mean square Error(RMSE) on training set
  - Coefficient of determination(R square) on training set
- Robustness
  - RMSE on test set
  - R square on test set
- Simplicity
  - Number of features included
  - Complexity of the methodology applied

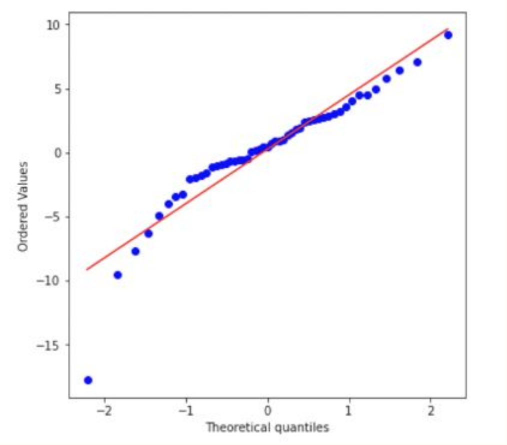
- Candidate models
  - Baseline: Linear regression
    Body Fat ~ Height + Weight
  - Ridge regression
    Body Fat ~ All features + | · |^2
  - LassoCV regression ***(final choice)***
    Body Fat ~ selected features + | · |
  - SVM regression
    Body Fat ~ features(same as lasso)

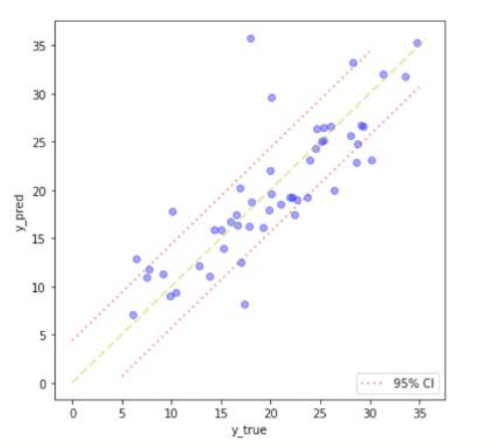
Comparison

Criteria Candidate	Accuracy(Training)		Simplicity		Robustness(Test)	
	RMSE	R square	Parameters	Methodology	RMSE	R square
Baseline:Linear	5.4799	0.4445	2	Linear regression	5.0462	0.5154
Ridge	4.2354	0.6682	14	Linear regression	4.3720	0.6363
LassoCV	**4.4722**	**0.6300**	**5**	**Linear regression**	**4.3568**	**0.6388**
SVM regression	4.5940	0.6096	5	SVM regression	4.4429	0.6244

Residual Analysis







Residual Plot

No obvious correlation.
Residuals are irrelevant.

Q-Q plot: Residuals

Except for a few outliers, most residuals are normally distributed.

Scatter Plot: y_pred v.s. y_true

Most true values are located in the 95% CI of predicted ones.

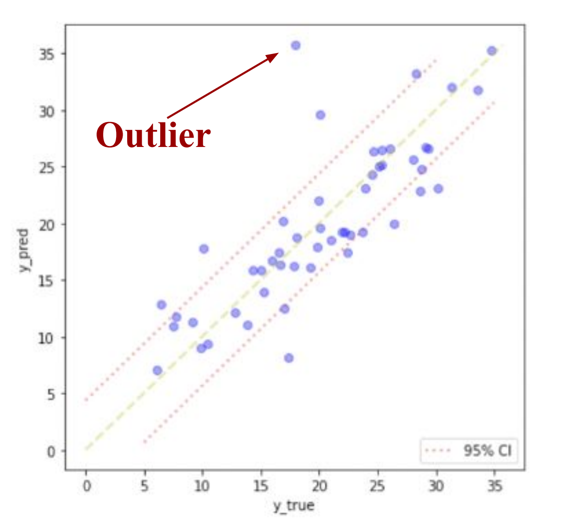
Outlier Analysis

Body Fat **Origin**	Body Fat **Imputed**	Body Fat **Predicted**	Age	Chest	Abdomen	Hip	Thigh
45.1	17.98	35.74	51	119.8	122.1	112.8	62.5

For the outlier with largest residual, we find that the prediction is closer to its origin value than the imputed one.

That means our model balances the system mistake.

We consider this as a confirmation of model robustness.


Outlier

Strengths & Weaknesses

BodyFat = 0.0345*AGE **+ 0.2623***CHEST **+ 0.7684***ABDOME **+ 0.3255***HIP **+ 0.3044***THIGH

- **Strengths**

  Simplicity(5 features instead of 14 and simple methodology).
  Acceptable accuracy.
  High Robustness.

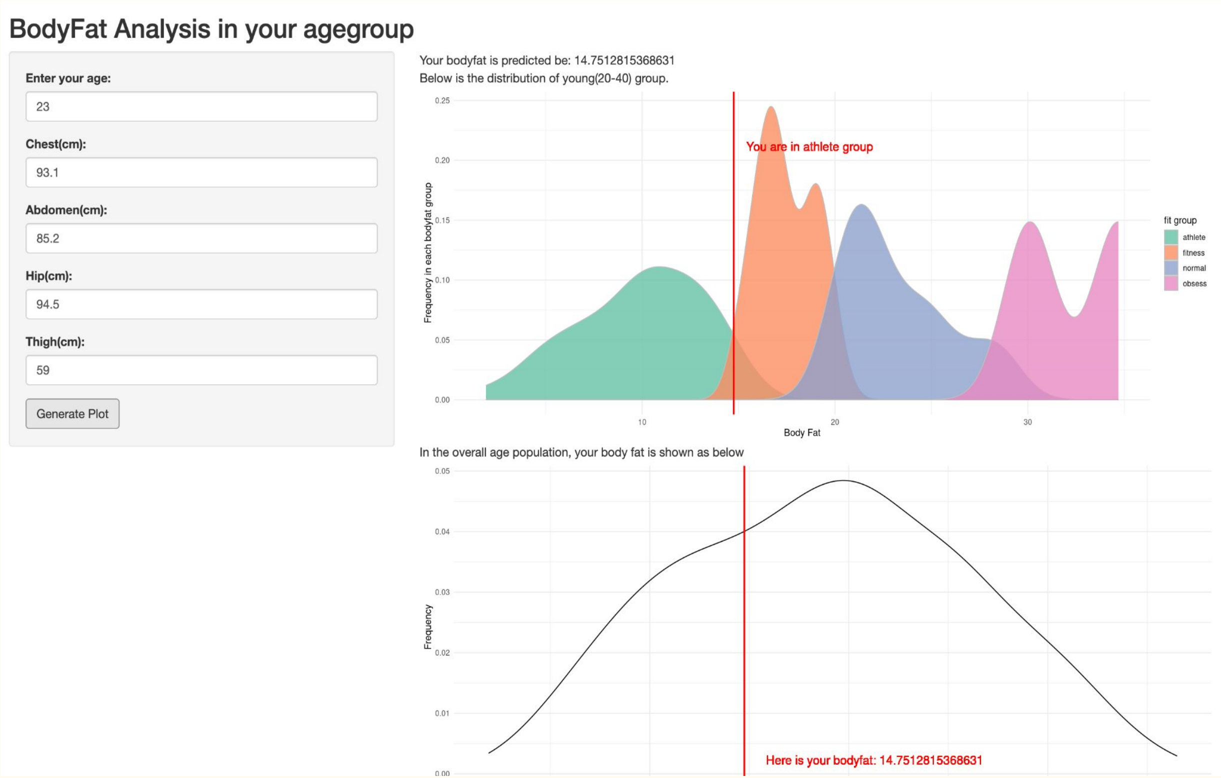
- **Weaknesses**

  Underfitting: R square is only 0.6388 on the test set:

  15.69% of predictions within +/- 3% of true value
  23.53% of predictions within +/- 5% of true value
  50.98% of predictions within +/- 10% of true value

Visualizable Product — Shiny App

- Input box
  - Five predictors
- Output
  - Body fat distribution for different age groups
  - User’s Body fat location in overall population.
- URL:https://stat628m2group13.shinyapps.io/stat628/



BodyFat Analysis in your agegroup

Enter your age: 23

Chest(cm): 93.1

Abdomen(cm): 85.2

Hip(cm): 94.5

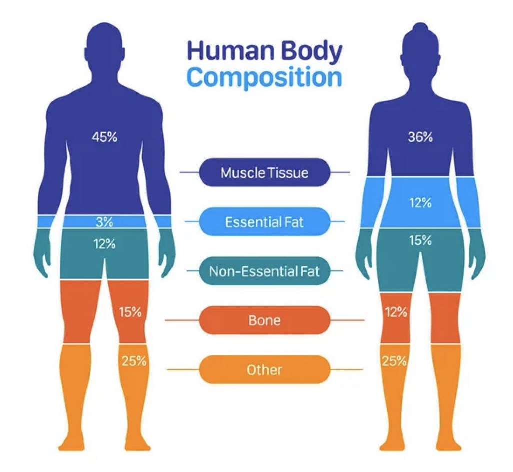
Thigh(cm): 59

Your bodyfat is predicted be: 14.7512815368631
Below is the distribution of young(20-40) group.

In the overall age population, your body fat is shown as below

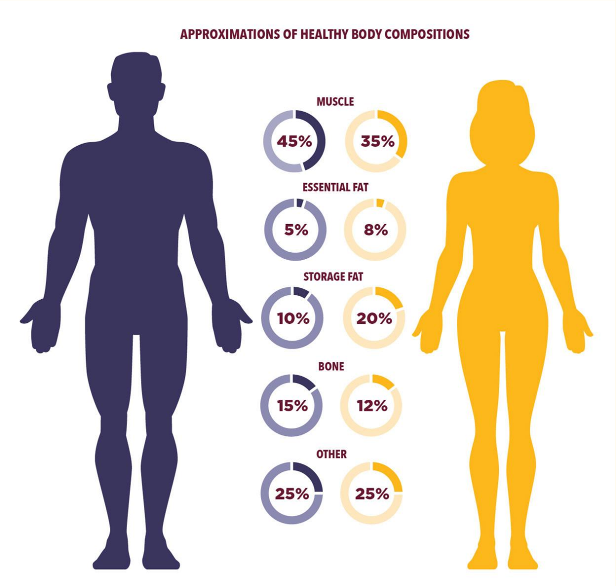
Discussion

- Dataset Improvement
  - The capacity of the current set is **not enough for more complicated model**(eg, MLP, random forest).
  - It **fails to distinguish the data between male and female**, considering that their body fat have very different distribution.
  - Possibles solutions: Further collection & Up-sampling.



Discussion

- Future Work
  - Fat is stored in different part of body.
  - Male and Female should have different models for calculation
  - Model accuracy can be enhanced.



Conclusion

This model is aimed to calculate body fat in an accurate, simple and robust way.

We choose **lasso regression with cross-validation** as our final model because it returns a lower RMSE and only requires five predictors. It shows good accuracy and robustness on both training and test sets, although limitations still exist.

If time permits, we believe it can be better polished with superior techniques.

Thank You for Listening.