Summary

Introduction: (Yuhan)

Nowadays, people are more concerned about health problems. One of the important factors is the body fat content. Excessive fat content will greatly increase the risk of disease. Therefore, accurately predicting body fat content is particularly important for human health. In this report, we try to build a simple model, which could make the precise prediction of body fat.

Data Cleaning: (Yuhan)

First, we look at the raw data there are some outliers in the raw data. We process the raw data by using quantiles to detect the outliers and delete them. We calculate the 10th percentile and the 90th percentile data of the data set. Then, using lower quantile minus all outliers 1.5 times the interquartile range and upper quantile plus all outliers 1.5 times the interquartile range as the range to determine the outlier. After doing so, we delete the row of outliers. We scan the first three columns and delete the outliers manually (e.g. the row of body fat contains 0). And then standardize the data.

Final Model Statement: (Yuhan)

We finally choose Lasso regression with regularization parameters to describe the relationship between the body fat (Y) and other features (AGE (X_1), HEIGHT (X_2), ABDOMEN (X_3), WRIST (X_4)). The model is:

$$Y = 18.8971429 + 0.2633502 \cdot X_1 - 0.6780748 \cdot X_2 + 6.3043124 \cdot X_3 - 0.5269030 \cdot X_4$$

When we use lasso regression, we use cross-validation to select proper λ, using lse of λ as the regularized parameter. We try to keep the MSE at a lower level and get a simple model simultaneously. (See **Figure 2.**) A 22-year-old man, whose height is 72.25 inches, abdomen is 83.0 cm, and wrist is 18.2 cm. Based on our model, his bodyfat is predicted to be 12.37(%), and 95% predicted interval is (11.69%,12.73%). Our estimated coefficients are 0.2237683,-0.6684841, 6.2077191, and -0.4217869, which are in the units of AGE, HEIGHT, ABDOMEN, and WRIST. This means that when the value of AGE increases by 1 unit, the predicted value of body fat % will increase by 0.2237683 units when the value of HEIGHT decreases by 1 unit, the predicted value of body fat % will decrease by 0.6684841 units when the value of ABDOMEN increases by 1 unit, the predicted value of body fat % will increase by 6.2077191 unit and when WRIST the value of decreases by 1 unit, the predicted value of body fat % will decrease by 0.4217869 unit.

Rationale for the Final Model and Relevant Statistical Analysis: (Ziyi)

We found Lasso's R^2 is closest to 1 compared with OLS and ridge regression. This implies that the Lasso model will be best in our prediction. Since AIC, BIC will penalize models for adding more parameters and they will ensure the model is not overfit or complex. And MSE can help us

to ensure the difference between the model and actual data. We also calculate the MSE, AIC, and BIC values of the three models and compare them together. From MSE, AIC, and BIC comparisons, we still get Lasso as best from them.

We use the residual plot (See **Figure 1.**) to access our model. Our residual plot shows that there is no obvious pattern, points are randomly distributed around the 0, and points are evenly distributed around the 0 line. Therefore, the model fitting results are good.

Model Strengths and Weaknesses: (Yuhan)

Strength: The model is simple and the R^2 is relatively high. Our model could handle the multicollinearity problem, which improves model stability and interpretability. Additionally, the coefficient of our model is relatively accurate, making the model more precise.
Weakness: Our model is sensitive to small changes in the data and may lead to unstable feature selection results. And if the data set is small, estimated parameters may be imprecise.

Conclusion: (Yifan)

Upon all the mentioned measurements of metrics, we agreed with each other that the Lasso regression retains the most robustness while reducing the complexity of the model itself. We considered 3 different metrics describing the goodness of fit, one focuses on the divergence of real and predicted values, and the others consider estimating the likelihood function with a penalty term of the number of parameters in the model. Based on all the evidence, we claim that the Lasso regression model satisfies both robustness and simplicity.

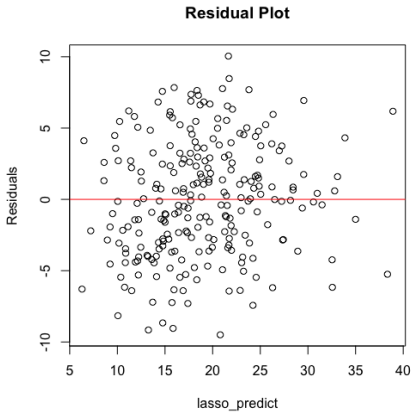

Figure 1. Residual Plot

Contribution:
Yifan Ren: Multiple Linear Regression, Shiny App & Github, Figure Variance Inflation Factor.
Yuhan Zheng: Data cleaning, Ridge regression, Lasso regression, and relevant statistical analysis.
Ziyi Yang: Rationale for final model and model diagnostics