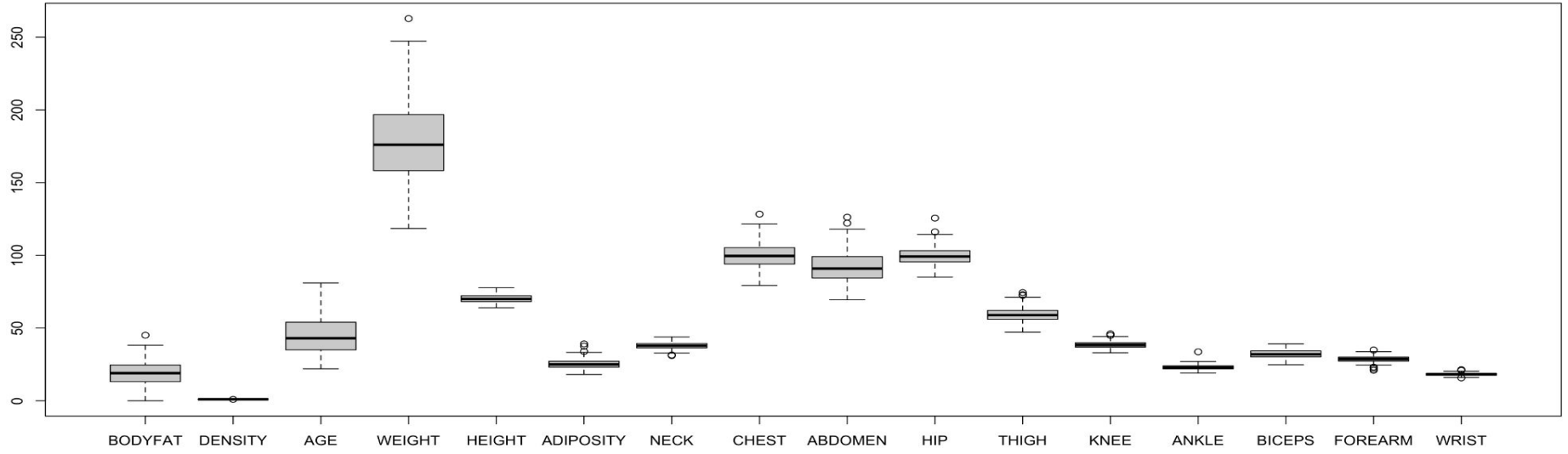
Stat 628 Module 2 Group 14

Body Fat Data Analysis

Data Description and Data Cleaning:

Look at the raw dataset, the boxplots show there are some outliers in the raw dataset.



Data Description and Data Cleaning:

- Use quantile to determine a range that detect the outliers and delete them.
- Range: (0.1qt - 1.5IQR, 0.9qt+1.5IQR)
- Find other outliers and delete them manually.

IDNO	BODYFAT
182	0

- Final Cleaned Data:
  row number: 245
  column number: 16 (remove the first column (IDNO) and the third column (DENSITY)).
- Standardize the data set.

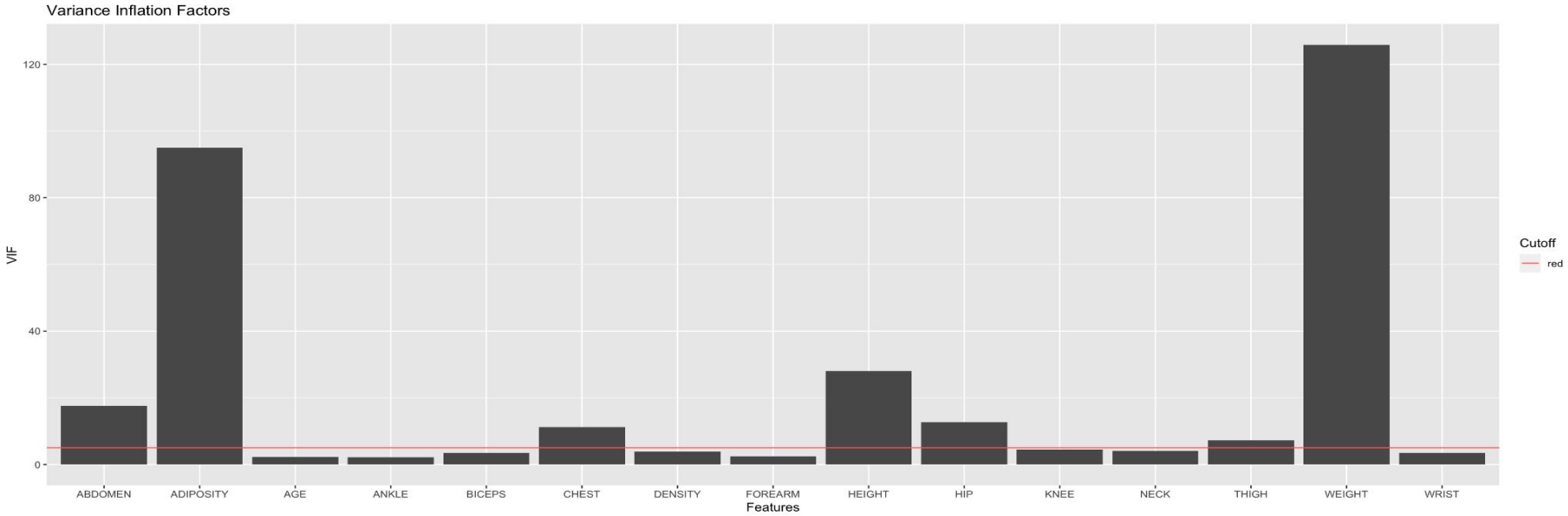
Three proposed Models:

1. OLS: $Y = \beta_0 + \beta_1 * \text{Weight} + \beta_2 * \text{Height} + \beta_3 * \text{Ankle} + \beta_4 * \text{Forearm}$
2. Lasso Regression
3. Ridge Regression

Consider lasso and ridge regression models with standardized inputs and penalty term, we tried to examine which model will outperform than the others.

Feature selection in OLS

Set variance inflation factor = 3 as the threshold, to select the features that potentially won't cause multi-collinearity issue.

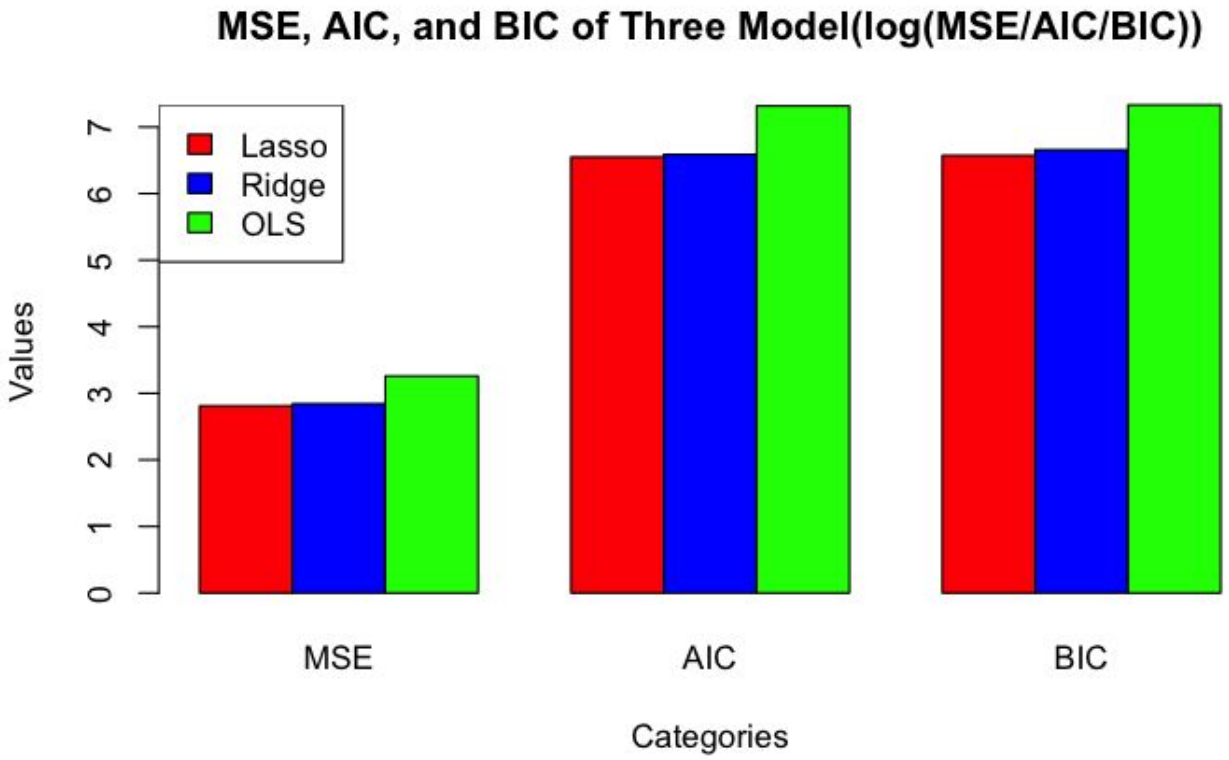


Rationale for the Final Model :

Lasso-R^2 = 0.7221

Ridge-R^2 = 0.7060

MLR-R^2 = 0.5594

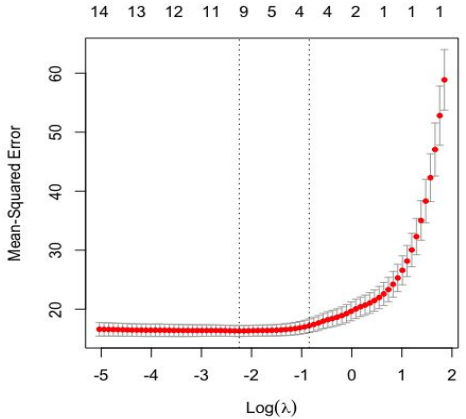


Final Model:

We finally choose **Lasso regression**. After doing the Lasso regression, there are four features to be kept.

- Body Fat (Y)
- Age (X1), Height (X2), Abdomen (X3), Wrist (X4).

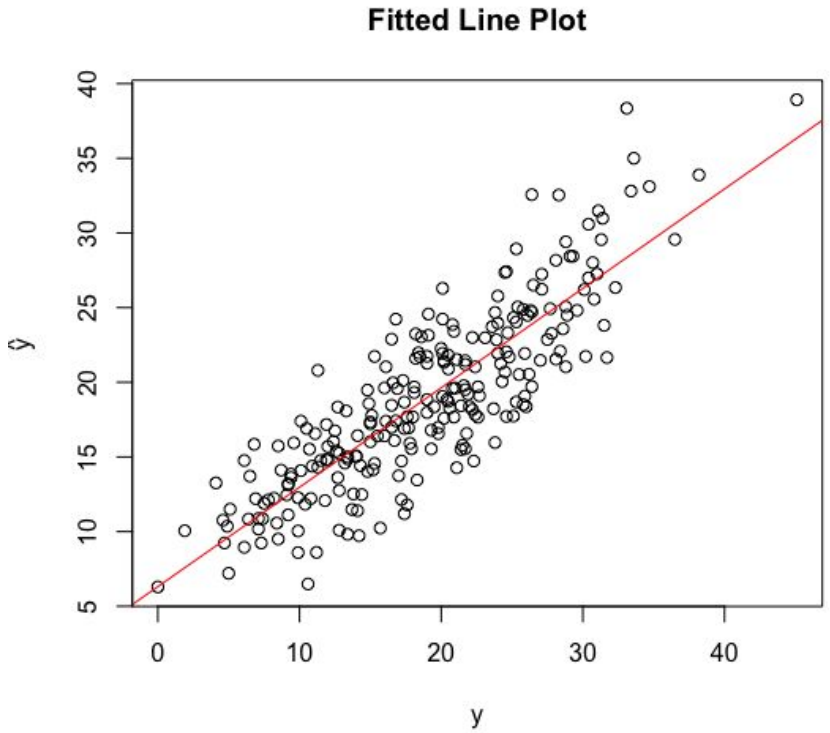
Choosing regularized parameter (λ):
Use cross-validation to determine the lse of λ to produce a sparser model, and keep the MSE at a lower level.



Y (BodyFat) = 18.8971429 + 0.2633502*X1 (Age)-0.6780748*X2 (Height)
+6.3043124*X3(Abdomen)-0.5269030*X4 (Wrist)

Final Model:

Visualized the final model

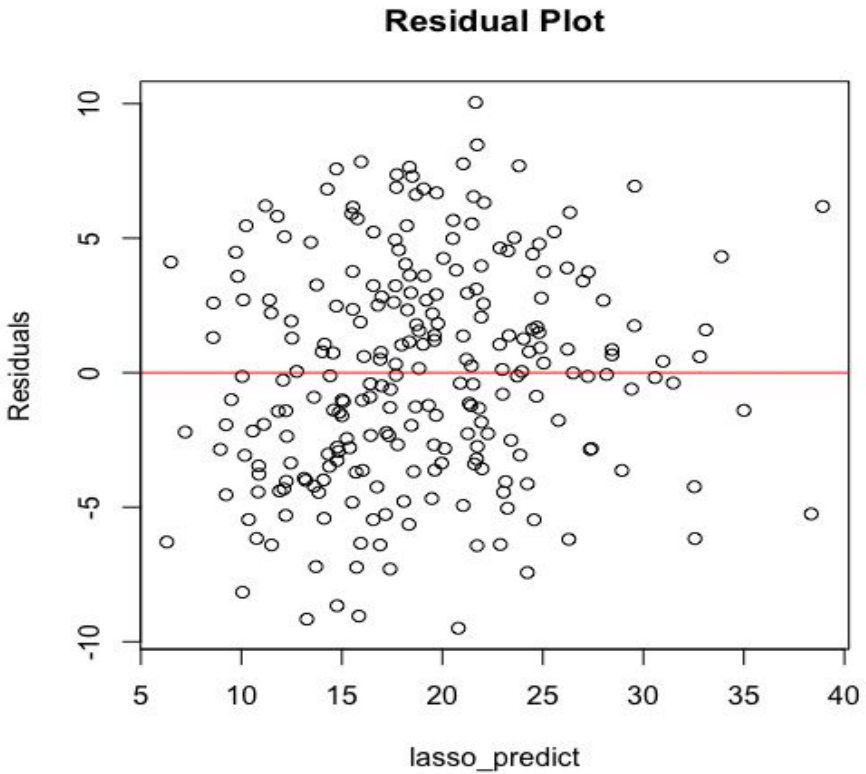


The x-axis represents the true value of y (Body Fat), the y-axis is the predicted value using our model. The red line is the linear regression model (lm) to fit a linear regression line of y_hat on y.

Model Diagnostics

We use residual plot to access our model.

1. No obvious pattern
2. Points are randomly distributed around the 0 line.
3. Points are evenly distributed around the 0 line.



Strength and Weakness

Y (BodyFat) = 18.8971429 + 0.2633502*X1 (Age)-0.6780748*X2 (Height)
+6.3043124*X3(Abdomen)-0.5269030*X4 (Wrist)

- Strength:
  - Our model is simple and could handle the collinearity problem, which improves model stability and interpretability.
  - The coefficient of our model is relatively accurate
  - Our independent variables of our model are easy to measure.
- Weakness:
  - Sensitive to small changes in the data and may lead to unstable feature selection results.
  - If the data set is small, estimated parameters may be imprecise.

Thanks for listening.