Body Fat Project Presentation

Group-20

October 20, 2022

Summary of Data Cleaning

- We drop the sample whose estimated body fat is negative. The estimated fat is infered from Siri's equation:

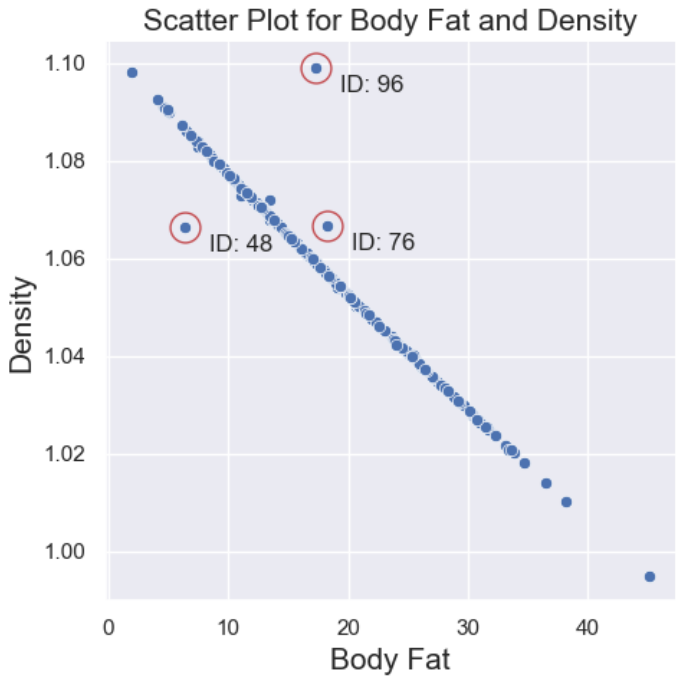
$$\text{Fat} = \frac{495}{\text{Density}} - 450$$

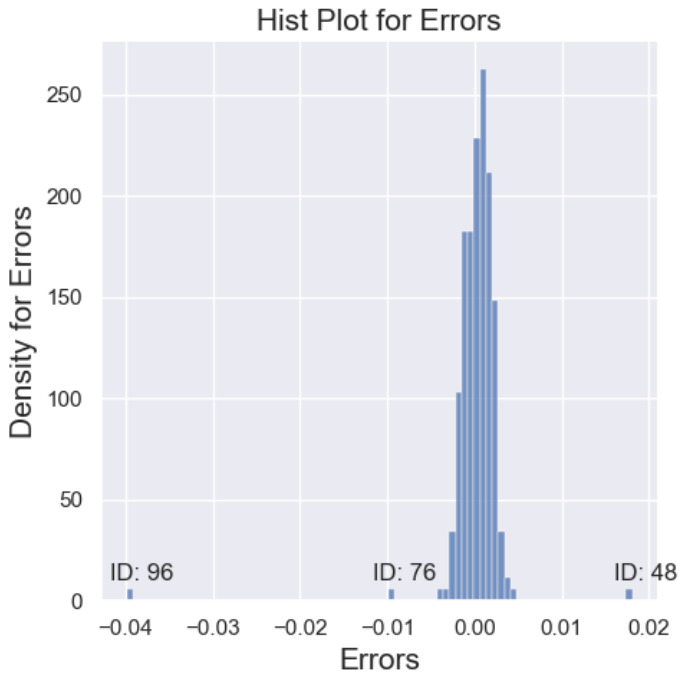
The sample's statistics are listed below:

IDNO	BODYFAT	DENSITY	Estimated Body Fat
182	0.0	1.1089	-3.6117

Summary of Data Cleaning

- From the scatter plot of body fat and density, we find there are 3 obvious outliers. We use the Siri's equation to compute the estimated body fat and substract it by corresponding the original BODYFAT to get errors.





Summary of Data Cleaning

- We use $3-\sigma$ criterion on the errors to find out these outliers. Finally, we use estimated body fat from Siri's equation to replace the original body fat for these samples.

IDNO	BODYFAT	Estimated Body Fat	Error
48	6.4	14.1350	7.7350
76	18.3	14.0915	-4.2085
96	17.3	0.3685	-16.9315

Feature Selection

Here we consider two useful methods for the feature selection step: LASSO and stepwise regression.

- LASSO is use L_1 norm to keep the sparsity of the estimated solution. The features with non-zero coefficient are selected to be the important variables.

$$L(\beta) = \|y - X\beta\|^2 + \lambda\|\beta\|_1$$

Finally, we select four features: Weight, Height, Abdomen and Thigh.

Feature Selection

- In general, the process of selecting variables by the stepwise regression method consists of two basic steps: one is to remove variables from the regression model that are not significant by t-test, and the other is to introduce new variables into the regression model that are significant by F-test.

  Here, we set both thresholds for introducing and removing variables as 0.05. Finally, we select four features: Weight, Wrist, Abdomen and Forearm.

Model Selection

- Candidate Models: Linear model with different predictors

$$\text{BodyFat} \sim \text{Abdomen} + \text{Weight} + \text{Wrist} + \text{Forearm}$$

(Based on stepwise regression)

$$\text{BodyFat} \sim \text{Weight} + \text{Height} + \text{Abdomen} + \text{Thigh}$$

(Based on Lasso method)

- Model performance result:

Model	Lasso Model	Stepwise Regression
R^2	0.7245	0.7351
MSE on test set (CV=6)	18.228	17.208

Model Statistical Analysis

Coefficients	Estimate	t value	p-value
Intercept	-31.30	-4.67	5.6×10^{-6}
Abdomen	0.92	17.75	2.0×10^{-16}
Wrist	-1.39	-3.40	8.0×10^{-4}
Forearm	0.45	2.65	8.5×10^{-3}
Weight	-0.13	-5.48	1.05×10^{-7}

F-statistics $= 162.4$, p-value $= 2.2 \times 10^{-16}$

- T-test reports the significant partial effect of adding each predictors to the model.
- F-test shows there exists relationship between response and predictors.

Model Interpretation

Final Model

$$100 \times \text{Bodyfat\%} = -31.30 + 0.92 \times \text{Abdomen} - 1.39 \times \text{Wrist} + 0.45 \times \text{Forearm} - 0.13 \times \text{Weight}$$

Some description of the final model in words:

- As men's abdomen increases by one centimeter, he is expected to gain 0.92% in body fat.
- If two men have almost the same abdomen, forearm and weight, then having the smaller wrist means having the higher percentage of body fat.

Model Diagnostics

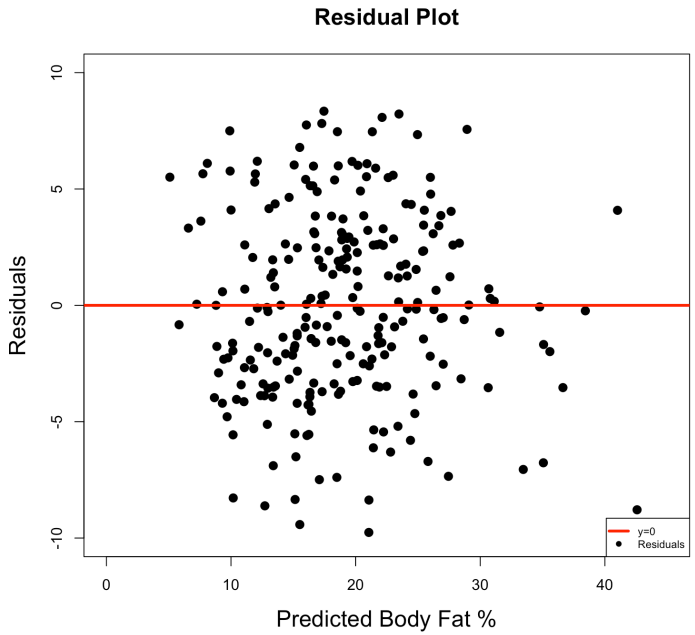
Assumptions for the SLR Model

- Linearity: The relationship between X and Y must be linear.
- Independence of errors: There is not a relationship between the residuals and the Y variable;
- Normality of errors: The residuals must be approximately normally distributed.
- Homoskedasticity (Constant variance): The variance of the residuals is the same for all values of X.

Model Diagnostics

Residual Plot

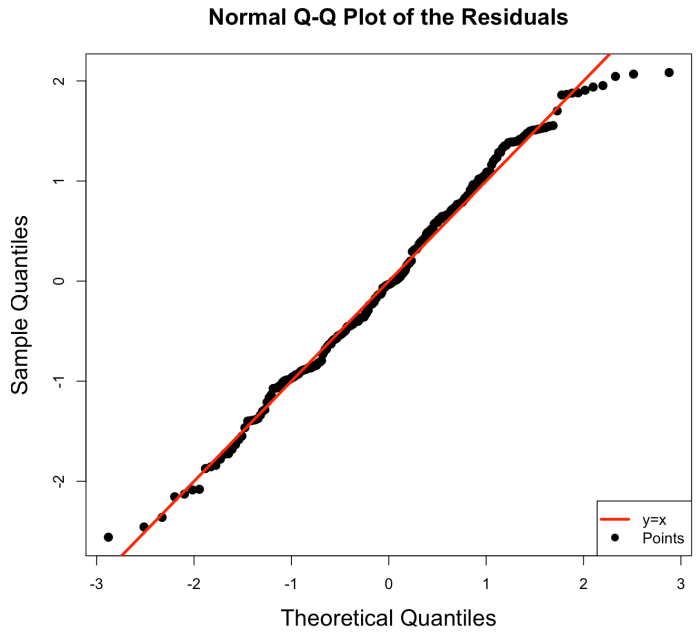
- We checked linearity, homoskedasticity and independence using Residual Plot.



Model Diagnostics

QQ Plot

- We checked normality using QQ Plot.



Model Diagnostics

Shapiro-Wilk test

- We used the Shapiro-Wilk test to test normality

$$W = 0.98926, p_{val} = 0.05935$$

- We failed to reject "the sample is normal" because p-value is $0.05935 > 0.05$.

Strength and Weekness

- Strength
  - Our model satisfies the linear regression assumption of homoskedasticity and linearity and it is explainable about its result.
  - The stepwise regression method overcomes the multiple linearity of predictors of the model.
- Weekness
  - The final model contains 4 predictors, which is kind of complicated.
  - If a man has extreme values, the model may produce negative body fat prediction.