Bodyfat Project

Group 2: Bowen Tian, Ouyang Xu, Tianhang Li

Summary of Data Cleaning

- **Imputation**

INDO	Original Obs.	Imputed Obs
42	29.5 inchs	69.5 inchs

$$Bodyfat = \frac{495}{Density} - 450$$

$$Adiposity = \frac{0.454 \times Weight}{(0.0254 \times Height)^2}$$

- **Deletion**

Reason	INDO
BODYFAT < 3 or BODYFAT > 40	172 182 216
BODYFAT and DENSITY not match	48 76 96
ADIPOSITY not match with HEIGHT and WEIGHT	63 221
Outliers	39 41

- **Final cleaned data**

- n = 242 with p = 14
- Predictors: AGE, WEIGHT, HEIGHT, ADIPOSITY, NECK, CHEST, ABDOMEN, HIP, THIGH, KNEE, ANKLE, BICEPS, FOREARM, WRIST

Model Fitting and Selection

- Candidate Models

Model	Predictors	Adjusted R^2	Rank of R^2	RMSE	Rank of RMSE	Method
1	AGE NECK ABDOMEN THIGH FOREARM WRIST	0.7125	1	3.894257	1	AIC
2	ABDOMEN ADIPOSITY CHEST	0.6736	4	4.175837	4	Correlation
3	ABDOMEN WEIGHT THIGH	0.7119	2	3.92342	2	AIC + Cor
4	HEIGHT ABDOMEN NECK	0.7106	3	3.932186	3	Searching

- Model Selection
  - **Model 1 - highest R-square and the lowest RMSE, but requires 6 predictors.**
  - **Model 3 - similar R-square and RMSE as Model 1, and only requires 3 predictors.**
  - **Select Model 3 to be the final model.**

Final Model

Bodyfat = -49.10679 + 0.90497×Abdomen - 0.15878×Weight + 0.21646×Thigh

- **Interpretation about final model**

increase 1 cm in abdomen circumference → 0.905% increase of body fat

increase 1 cm in thigh → 0.216% increase of body fat

Body fat percentage is negatively related to body weight within this model

- **Example of model using**
- **Weight = 180 lbs, Abdomen = 100 cm, Thigh = 60 cm**
- **Estimated Bodyfat: 25.79%**
- **95% confidence interval: [24.91%, 26.69%]**

Statistical Properties of Final Model

- **All coefficients: significant at 0.05 level based on p-values:**

  p_Abdomen < 2e-16

  p_Weight < 2e-16

  p_Thigh = 0.0281

- **Overall model: significant at 0.001 level based on p-value**

  overall p-value < 2e-16

- **No multicolinearity issue based on VIF tests**

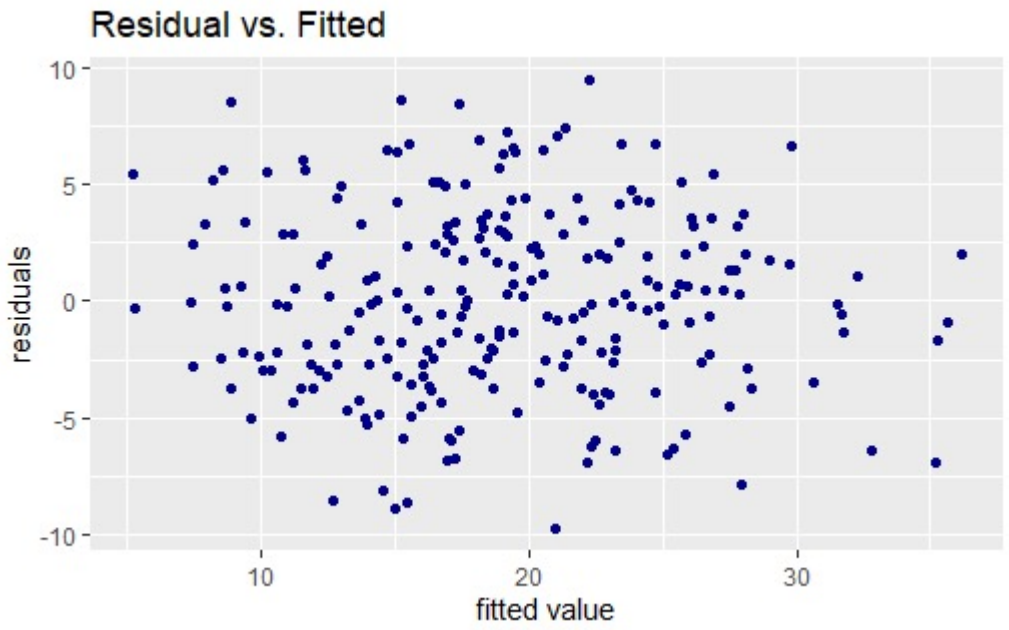
  vif_Abdomen = 4.0

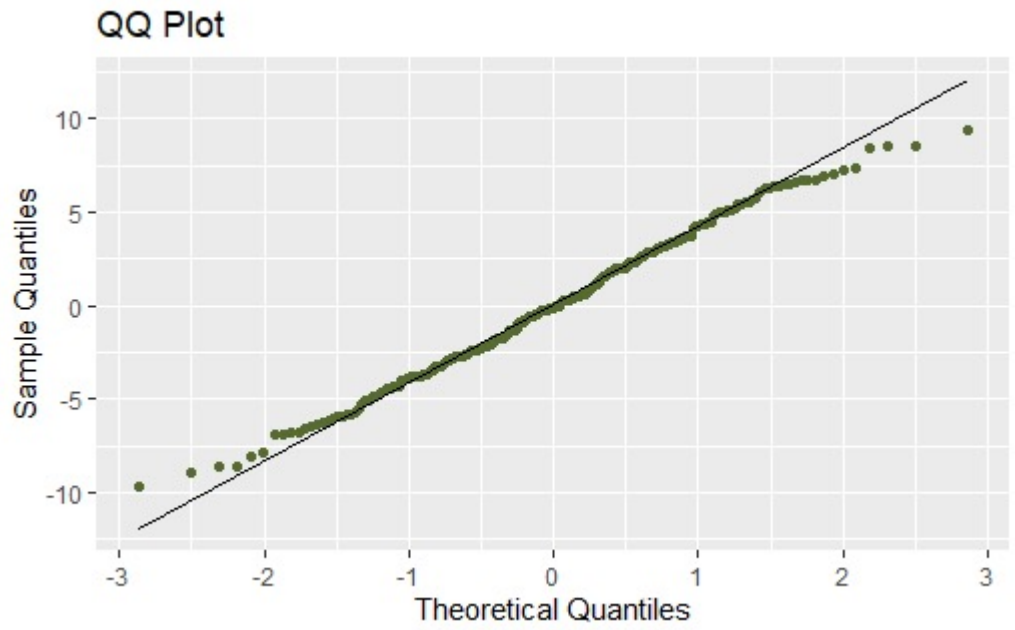
  vif_Weight = 6.6

  vif_Thigh = 3.4

- **Adjusted R^2 = 0.7119 RMSE = 3.92342**

Model Diagnostics





- **We do not find any pattern of the residual points. Random scatter indicates no serious departure from linearity.**
- **Residual points are distributed evenly and randomly, homoscedasticity checked.**
- **There is no outlier.**
- **In QQ plot, points at the tails are not close to the line, so the normality assumption maybe violated. However, for estimating and predicting of values of the response, the results will not be affected.**

Strengths and Weaknesses

$$\text{Bodyfat\%} = -49.10679 + 0.90497 \times \text{Abdomen} - 0.15878 \times \text{Weight} + 0.21646 \times \text{Thigh}$$

- **Strengths**
- All predictors in the final model are significant under alpha = 0.05.
- The model is simple but gives a fairly R-square and RMSE.
- The data of variables are easy to get.
- **Weakness:**
- The predictable range of the model is limited.

  eg. the estimated body fat for a male with 150 cm ofabdomen circumference, 60 cm thigh circumference, and 200 lbs body weight is 67 percent which is too high for aperson.

  Thus, the model is accurate only when data is within a certain range.
- Coefficient of Weight do not match with correlation.

  The correlation between BODYFAT and WEIGHT is 0.59. However, the coefficient for WEIGHT is negative which does not match with the correlation coefficient.

Thank you!