BODY FAT CALCULATOR

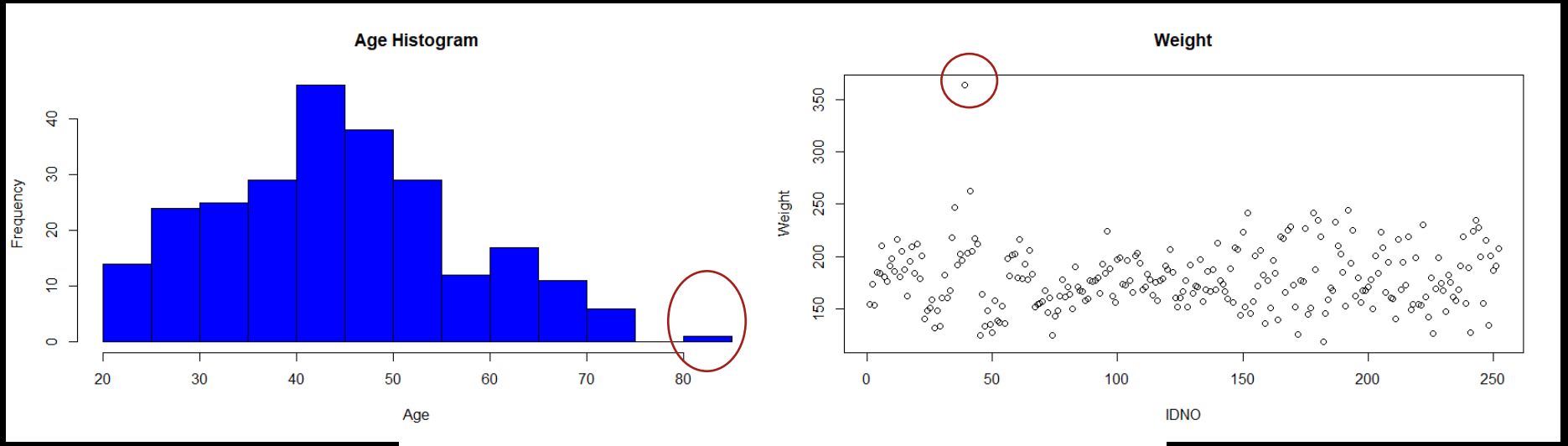
STAT 628 Module 2 Group 11

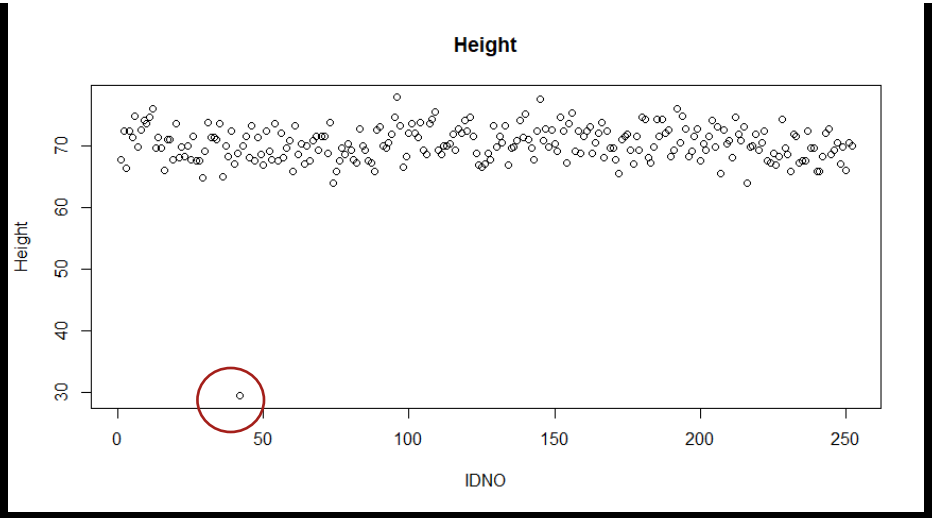
DATA EXPLORATION

- Searched for outliers by summary statistics and plots

- Imputed abnormal data by the existing relationship

PLOTS OF AGE, WEIGHT AND HEIGHT VARIABLES





OUTLIERS

IDNO	Reason	How to proceed
39	Has the max value for weight adiposity, neck, chest, abdomen, hip, thigh, and knee, clearly an outlier from the plot	Remove from data
42	Height unreasonably low	Attempt to recover
79	The sole individual with age above 80	Remove from data

RECOVER HEIGHT USING WEIGHT+ADIPOSITY

- Notice that IDNO42 has height 29.5 inches/74.93cm, which is unreasonable
- Can recover using $Adiposity = \frac{703 \cdot Weight(lb)}{Height(inch)^2} = \frac{Weight(kg)}{Height(cm)^2}$
- We have:

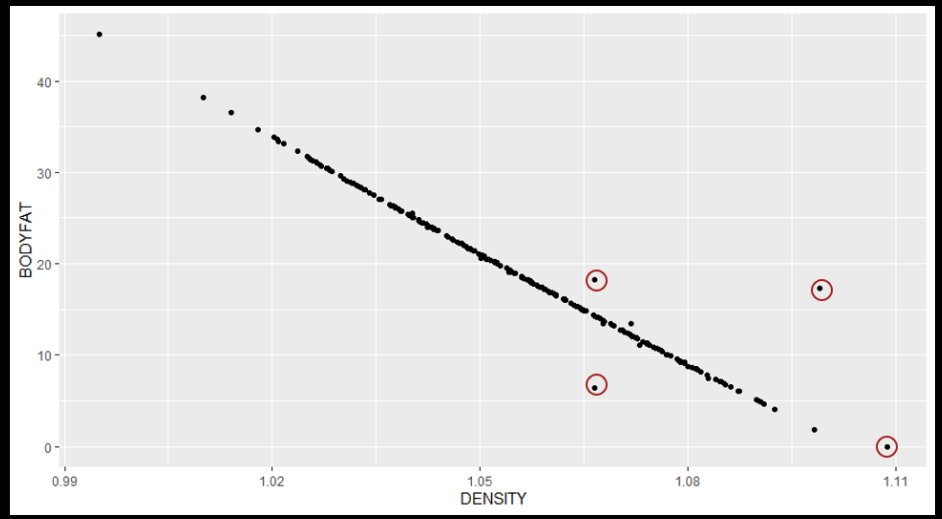
IDNO	Original	Recovered	Reasonable ?
42	29.5 inches/74.93cm	69.43 inches/176.34cm	Yes

BODYFAT VARIABLE

- There is a direct relationship between body fat and density that is shown by Siri's equation: $Bodyfat = \frac{495}{Density} - 450$

- Recovered abnormal data by Siri's equation

- Removed data of outliers after recovery



RECOVER DATA USING SIRI'S EQUATION

- Clearly, some values are wrong as they deviate from the line
- There are issues for IDNO48, 76, 96, 182
- Using Siri's Equation, we have the following:

IDNO	Original	Recovered	Reasonable and Keep?
48	6.4	14.1	Yes
76	18.3	14.1	Yes
96	17.3	0.4	No
182	0	-3.6	No

BODYFAT DATA

- From Google, we learn that bodyfat percentage <3% is unreasonable
- Using this, we can filter out the data for IDNO172, which has bodyfat value of 1.9

DATA CLEANING SUMMARY

- We deleted IDNO39, 79, 96, 172, 182 from data
- We recovered data for IDNO42, 48, 76
- Final cleaned data: n=247 individuals with all the original predictors

VARIABLE SELECTION

- Used LASSO for variable selection

(Intercept)	-7.930884769
AGE	0.004195674
WEIGHT	.
HEIGHT	-0.119194937
ADIPOSITY	.
NECK	.
CHEST	.
ABDOMEN	0.562906646
HIP	.
THIGH	.
KNEE	.
ANKLE	.
BICEPS	.
FOREARM	.
WRIST	-0.208526144

- Age, Height, Abdomen and Wrist are the most important variables

AGE VARIABLE

- By the article Effect of Aging on Body Fat, percentage body fat increased slightly between ages 20 to 39 and 40 to 84.

- Changed Age into a categorical variable with age 40 as cutoff

MODELS FOR BODYFAT

- Tried the following models and compare their performance:

  1. Random Forest

  2. XGBoost

  3. Multiple Linear Regression using Leave-one-out Cross Validation

MODEL PERFORMANCE

Model	Adjusted R^2	MSE
Random Forest	0.62	21.34
XGBoost	0.63	20.67
Multiple Linear Regression	With Wrist:0.74 Without: 0.71	With Wrist:15.75 Without:17.06

- Based on the results above, we proceed with using Multiple Linear Regression Model
- Note that we tried one model with wrist and another without so that wrist can be an optional input in the Shiny App for user convenience

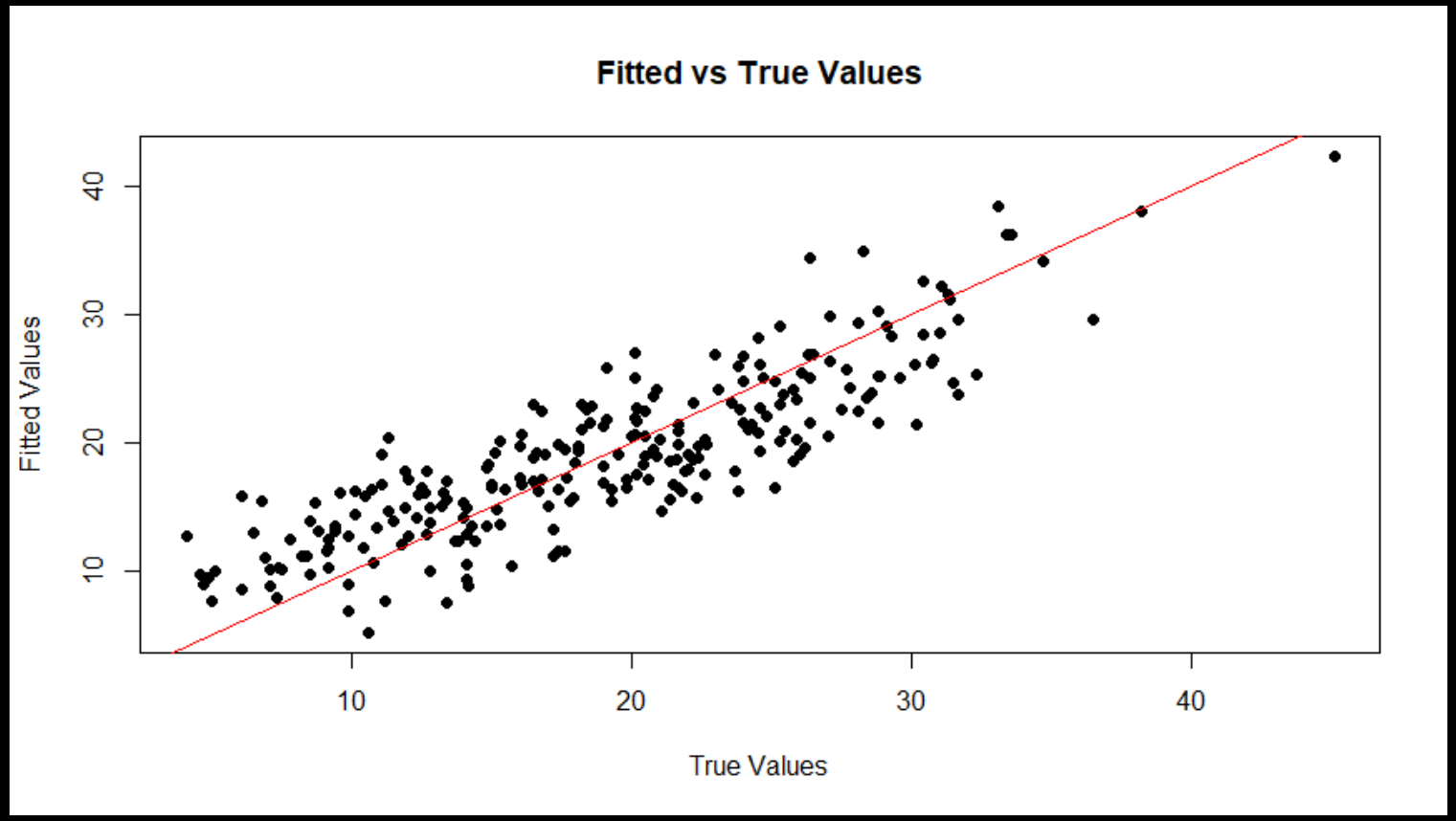
FINAL MODEL

- $Bodyfat\ \% = 9.39 + 0.71Abdomen - 0.14Height\ - 1.74Wrist + 1.14(Age\ 40\sim80)$

- Simpler and user-friendly model without wrist:

$$Bodyfat\ \% = -1.73 + 0.63Abdomen - 0.21Height\ + 0.75(Age\ 40\sim80)$$

MODEL PLOT


Fitted vs True Values
Fitted Values
True Values
10
20
30
40
10
20
30
40

EXAMPLES

- Average American male (175.3cm height, 101.25cm waist, 18.42cm wrist):

  ≤40: 25.76% 95% CI:[24.60%, 26.91%]

  >40: 26.91% 95% CI:[26.08%, 27.73%]

- Usain Bolt (35 years old, 195cm height, 83.8cm waist, 18.42cm wrist):
  Body fat 9.58% 95% CI:[8.06%, 11.09%]

MODEL DIAGNOSTICS

- Check Multiple Linear Regression assumptions:

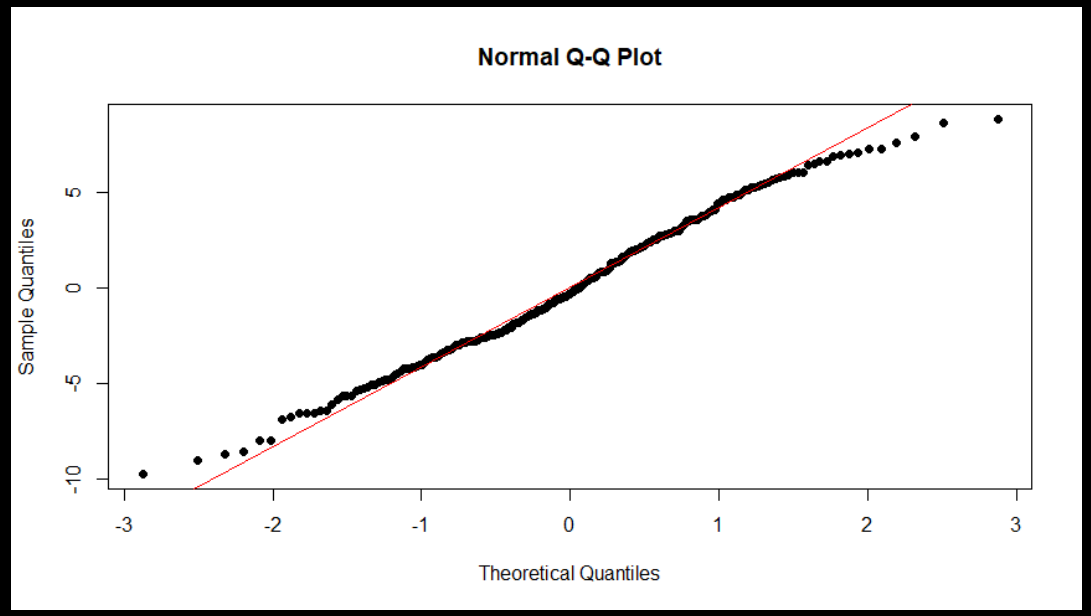
  - Normality

  - Homoscedasticity

  - Multicollinearity

NORMALITY

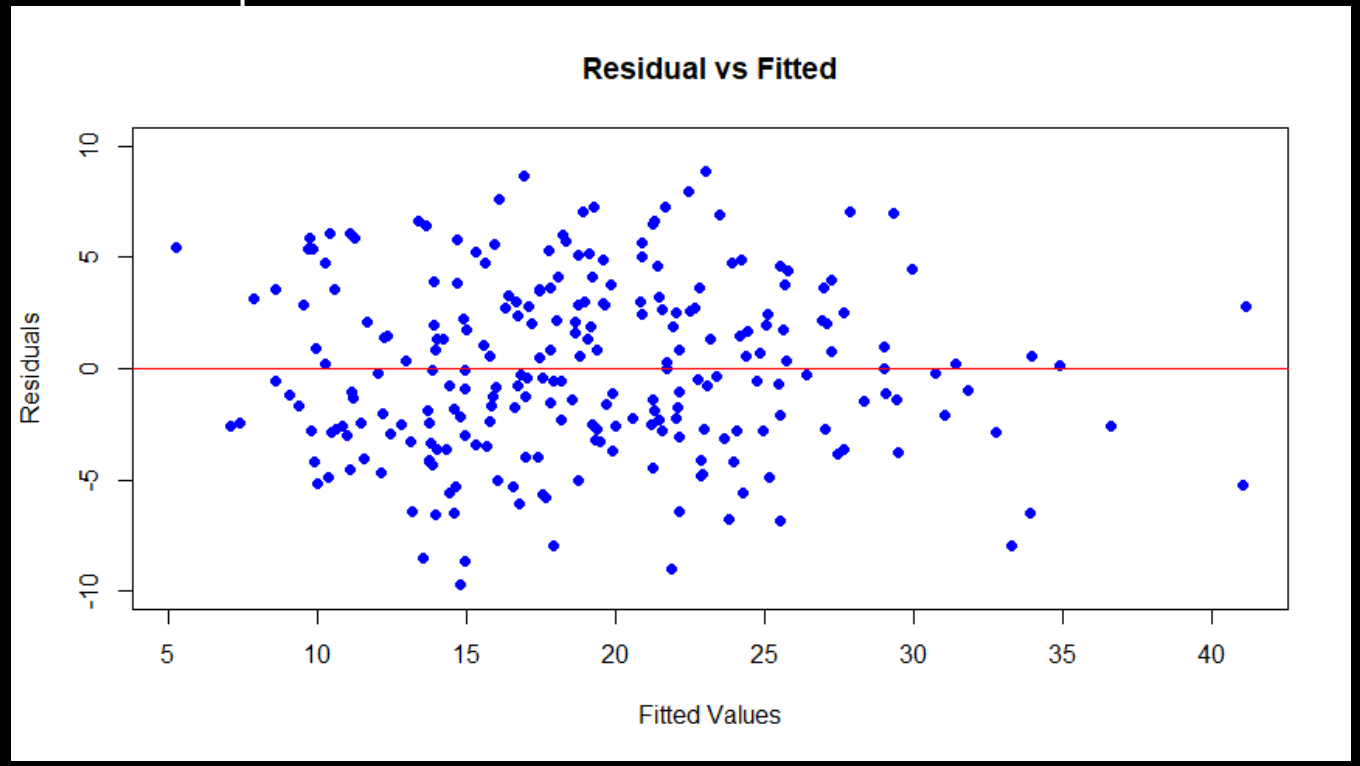
- We look at QQplot for residuals as well as perform the Shapiro test



- Shapiro test p-value: 0.076

HOMOSCEDASTICITY

- Residual plot:



MULTICOLLINEARITY

- Used the variance inflation factor (VIF) to check multicollinearity

- VIF value for each variable is lower than 2

`AGEgroup(40,80]`	ABDOMEN	HEIGHT	WRIST
1.110597	1.550051	1.227032	1.767819

- Conclusion: No multicollinearity issue

STRENGTHS/WEAKNESSES

- Strengths:
  - Decent R^2
  - Simple and easy to use and interpret

- Weaknesses:
  - It might not generalize well for individuals outside the range of our data, such as people <20 or > 80
  - Dataset population is low

SHINY APP DEMONSTRATION

- https://tinghuixu1114.shinyapps.io/bodyfatcalculator-group11/

Thank you!