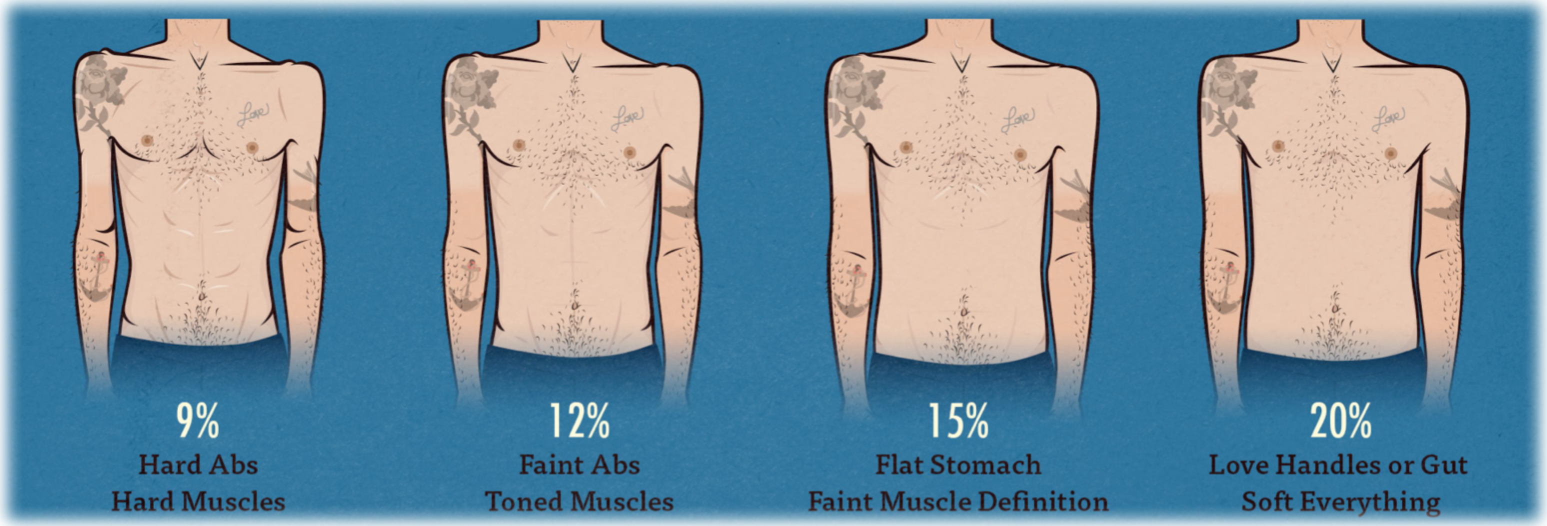
Goup4 Presentation

Body Fat Project



Summary of Data Cleaning

- We transformed the units of **NECK**, **CHEST**, **ABDOMEN**, **HIP**, **THIGH**, **KNEE**, **ANKLE**, **BICEPS**, **FOREARM** and **WRIST** from centimeter to inch to unify with the unit of **HEIGHT** and the custom of the US.

- We imputed **individual IDNO 42's HEIGHT**

Reasons:

The individual IDNO 42 has a height of 29.5 inches (about 75 cm) which is abnormal for an adult.

IDNO	Original HEIGHT	Imputed HEIGHT	Imputation Method
42	29.5 inches	69.5 inches	Using ABDOMEN(BMI) and WEIGHT to calculate to true value for HEIGHT

Summary of Data Cleaning

- We deleted **individual IDNO 172** and **individual IDNO 182**

Reasons:

The individual IDNO 172 has a **BODYFAT** of 0 and individual IDNO 182 has a **BODYFAT** of 1.9, which is below the normal range of human being. We tried using the formula $bodyfat = \frac{495}{\text{Density}} - 450$ to calculate the imputed value of **BODYFAT** for these two individuals but got smaller even negative values. So we just deleted them.

IDNO	**Original BODYFAT**	**Imputed BODYFAT**
172	1.9	0.697
182	0	-3.61

Summary of Data Cleaning

- Final Cleaned Data: **n=250** (from n=252) with p = 11 predictors

Predictor	Unit
WEIGHT	pound
HEIGHT	inch
ADIPOSITY	kg/m^2
NECK	inch
CHEST	inch
ABDOMEN	inch
HIP	inch
THIGH	inch
KNEE	inch
BICEPS	inch
WRIST	inch

Metric for Model Performance

- We define the desired model based on the following criteria:
  1. Prefer easy to measure predictors for users
  2. Prefer not confusing/clearly specified predictors for users
  3. Prefer using less predictors than more predictors
  4. Using *R-Squared* and *adjusted R-Squared* for evaluating models

Finding Desired Model for Bodyfat

- Decision making:

1. Try linear models with different dependent variables:

   BODYFAT vs DENSITY

2. After decided using BODYFAT as the dependent variable, try different simple linear regression models:

   BODYFAT ~ ABDOMEN

   BODYFAT ~ ADIPOSITY

   BODYFAT ~ CHEST

   BODYFAT ~ HIP

   BODYFAT ~ WEIGHT

   … …

Finding Desired Model for Bodyfat

- Decision making:

3. Starting from BODYFAT ~ ABDOMEN we select a new variable to be added to the SLR model:

BODYFAT ~ ABDOMEN + ADIPOSITY + ABDOMEN ADIPOSITY*

*BODYFAT ~ ABDOMEN + CHEST + ABDOMEN*CHEST*

*BODYFAT ~ ABDOMEN + HIP + ABDOMEN*HIP*

*BODYFAT ~ ABDOMEN + WEIGHT + ABDOMEN*WEIGHT*

*BODYFAT ~ ABDOMEN + HEIGHT + ABDOMEN*HEIGHT*

… …

4. Ended using **BODYFAT** ~ **ABDOMEN** + **WEIGHT** + **ABDOMEN*WEIGHT**

Results

Model	R-Squared	Adjusted R-Squared
BODYFAT ~ ABDOMEN	0.6522	0.6508
BODYFAT ~ ADIPOSITY	0.5193	0.5174
BODYFAT ~ CHEST	0.4808	0.4787
BODYFAT ~ HIP	0.3765	0.3739
BODYFAT ~ WEIGHT	0.3588	0.3563
… …		

Results

Model	R-Squared	Adjusted R-Squared
BODYFAT ~ ABDOMEN + ADIPOSITY + ABDOMEN ADIPOSITY*	0.671	0.667
*BODYFAT ~ ABDOMEN + CHEST + ABDOMEN*CHEST*	0.6795	0.6756
*BODYFAT ~ ABDOMEN + HIP + ABDOMEN*HIP*	0.6948	0.6911
*BODYFAT ~ ABDOMEN + WEIGHT + ABDOMEN*WEIGHT*	0.7235	0.7201
*BODYFAT ~ ABDOMEN + HEIGHT + ABDOMEN*HEIGHT*	0.6949	0.6912
… …		

Discussion of Results

- Reasons for using interaction terms:
  1. ABDOMEN and WEIGHT are two highly correlated variables.
  2. This interaction term is significantly important at significance level 0.01.
- Reasons for building model starting from ABDOMEN:
  1. In univariate models, BODAYFAT ~ ABDOMEN model gives the highest R-Squared and Adjusted R-Squared.
  2. For models using two predictors(with or without the interaction term), the formulas using ABDOMEN always perform better under the criteria of R-Squared and Adjusted R-Squared.

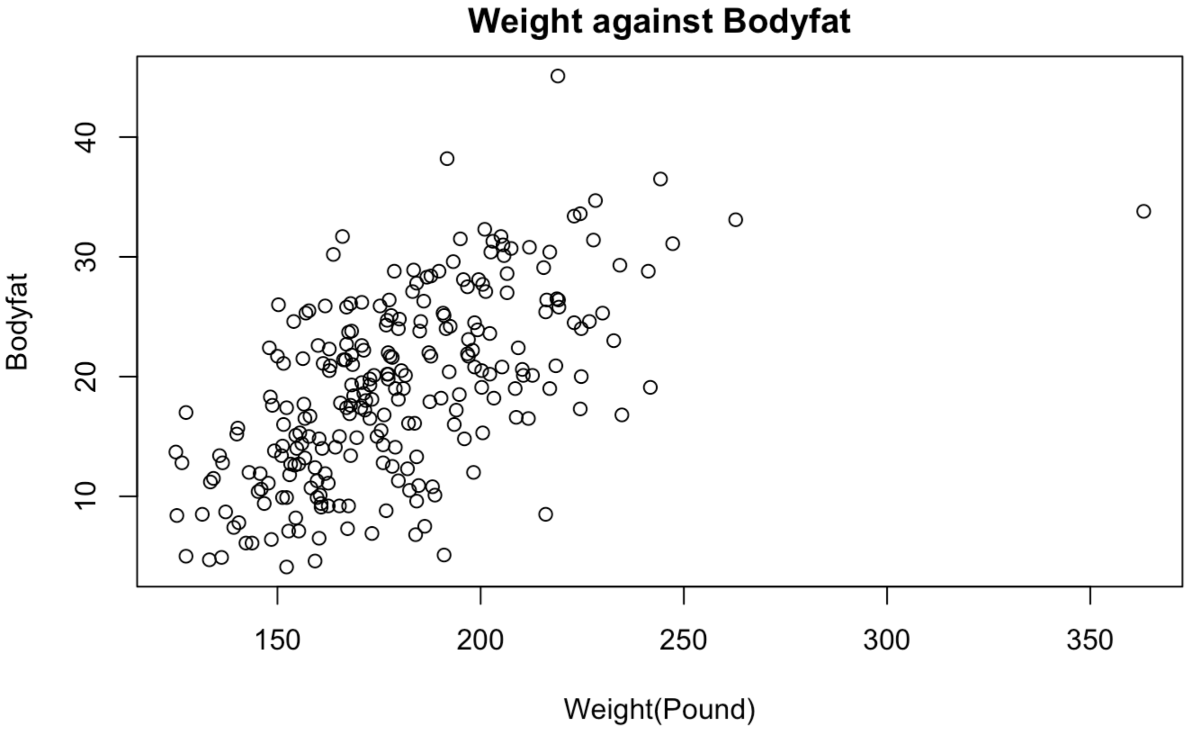
Final Model:

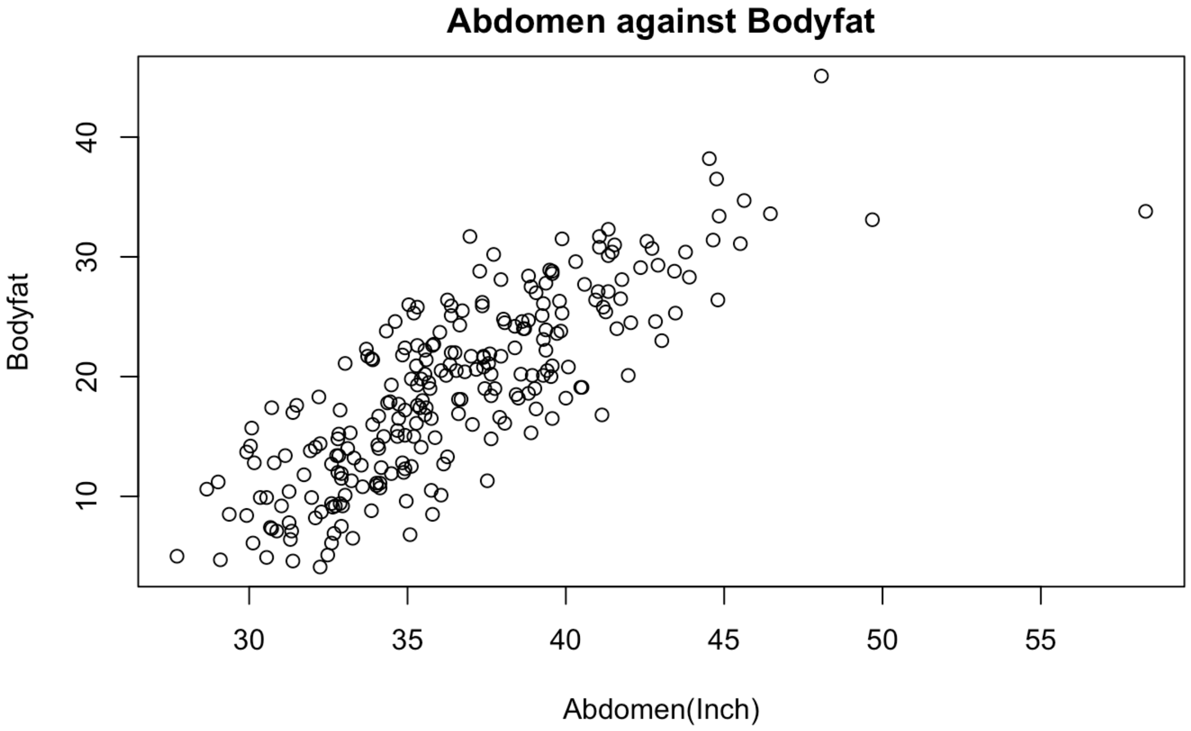
BODYFAT = -62.51 + 2.86*ABDOMEN - 0.014*WEIGHT - 0.003*ABDOMEN*WEIGHT

Explanation:

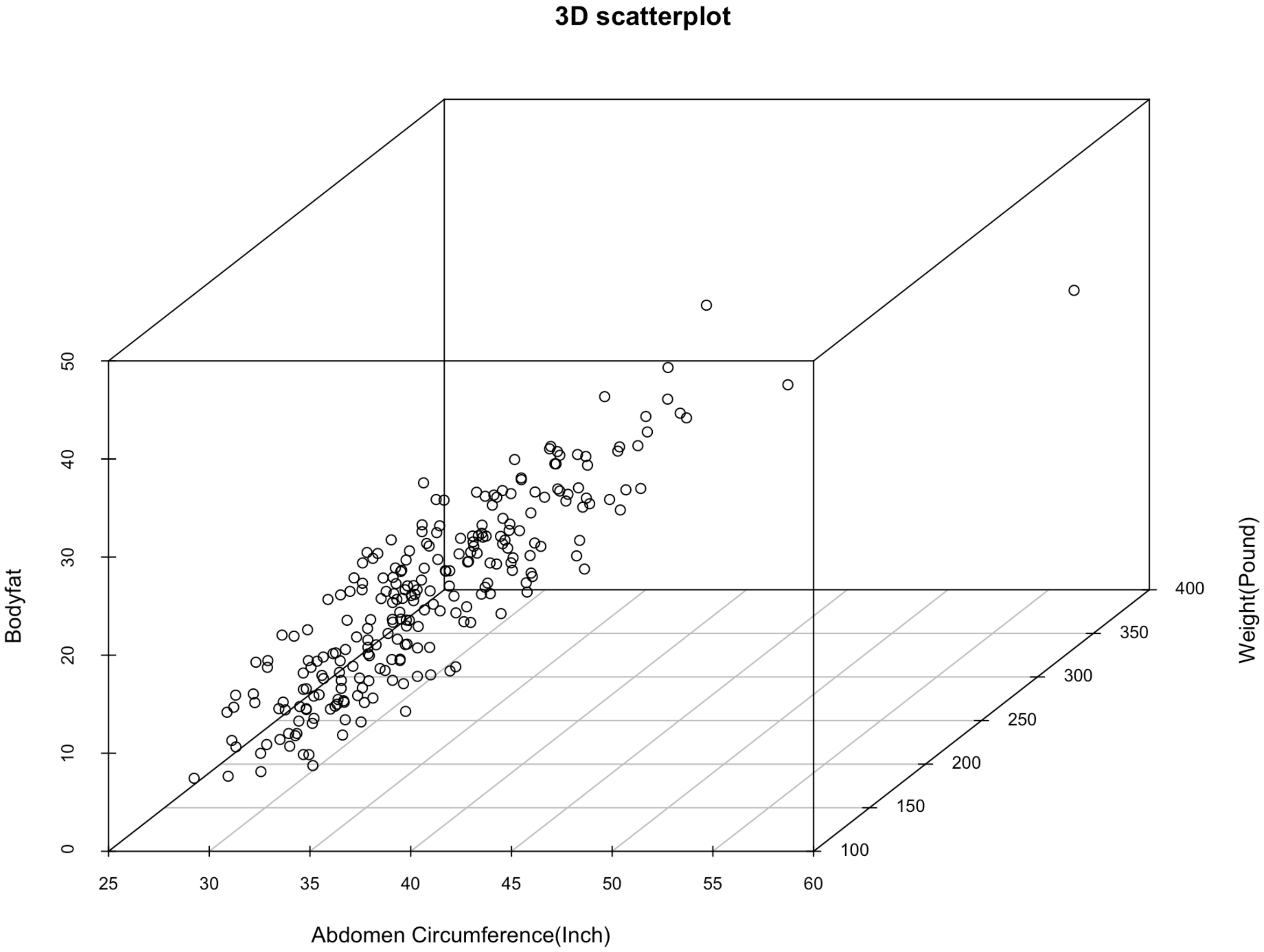
- As men's abdomen circumference get bigger by 1 inch, he is expected to gain about 2.9% in body fat while the actual increase depends on his weight.
- The increase of a man's weight does not tell us whether he is gaining muscle or fat. This explains why the coefficient for weight is negative and not significant.
- If a man gets heavier with the unchanged abdomen circumference, he is more likely to gaining muscle because muscle have higher density than fat. This is why the coefficient for the interaction term is negative.

Visualization of Final Model





Visualization of Final Model



Statistical Properties of Final Model

1. Coefficient for ABDOMEN is significant at significance level 0.001 based on two-sided t-test with p-value less than 2×10^{-16}.
2. Coefficient for the interaction term of ABDOMEN and WEIGHT is significant at significance level 0.001 based on two-sided t-test with p-value equals to 0.00313.
3. Coefficients of WEIGHT is negative and NOT significant at 0.1.
4. Overall model is significant at 0.05 based on F-test with p-value less than 2.2×10^{-16}.
5. The final model has an adjusted R-Squared equals to 0.72, making this model reliable and trustworthy.

Strengths and Weaknesses

Final Model:

BODYFAT = -62.51 + 2.86*ABDOMEN - 0.014*WEIGHT - 0.003*ABDOMEN*WEIGHT

- **Strengths**
  1. Simple: Easy to use for users, easy to explain for statisticians.
  2. Explains 72% of variation in body fat.

- **Weaknesses**
  1. Prediction is not accurate enough due to the limited volume of dataset.
  2. The coefficient for WEIGHT is not significant at level 0.1.

Thank you!