Module 2 Report

Group 8

Lu Li
Jiawei Wu
Nan Yan
Youhui Ye

Contents

1. Background
2. Data Visualization and Pre-process
3. Modeling
4. Results
5. Strengths and Weaknesses

Background

- Body fat percentage is a good measure of one's fitness level.
- Due to the limitation of realistic conditions, it is hard to evaluate the body fat percentage directly.
- Try to nd a simple but also powerful linear model to predict body fat percentage accurately with a set of easily obtained body measurements, such as heights, weights and etc.

Contents

1. Background
2. Data Visualization and Pre-process
3. Modeling
4. Results
5. Strengths and Weaknesses

Target variable visualization

As Y is continuous, we fit linear regression models on data.

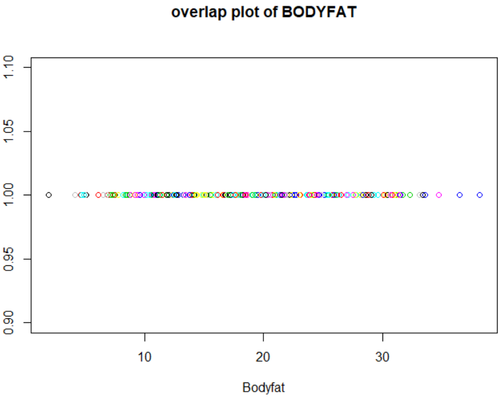


Figure: Bodyfat Variable

Data pre-process

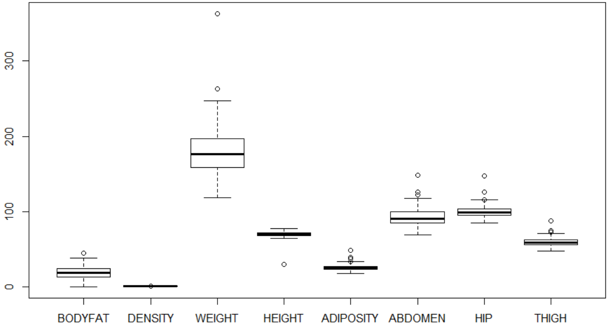


Figure: boxplot

Here we consider extreme values as potential outliers.

Potential outliers

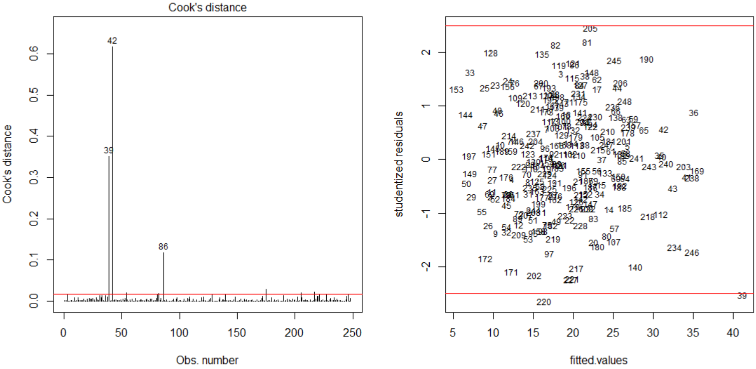


Figure: Cook's Distance

Basing on the above plots, we delete points [39, 42, 86, 220].

Potential outliers

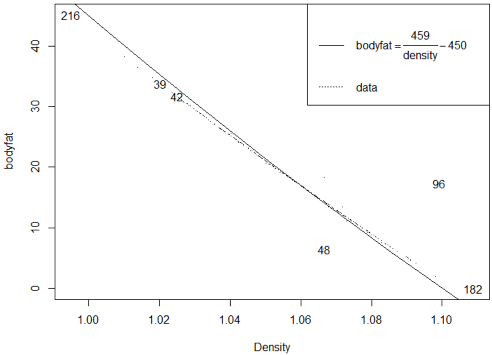


Figure: Bodyfat vs. Density

Delete 182, 216th datapoints. Modify the bodyfat value of 48th data(similar to 90th data) by function $bodyfat = \frac{459}{density} - 450$.

Contents

1. Background
2. Data Visualization and Pre-process
3. Modeling
4. Results
5. Strengths and Weaknesses

Original model

Multicolinearity

We check 2 measurements of multicolinearity:

- maximum VIF value : 128.532.
- condition number: 357615.1.

both of them show the design matrix has strong colinearity relationships.

Process

Considering the following reasons:

- Multicolinearity.
- Insignificant coefficients.
- Convenience for computation.

We select variables of the linear model by several criteria.

Approach 1: Stepwise regression

Direction

- Forward selection

  Starting with no variables in the model, repeat the process including the variable whose inclusion gives the most statistically significant improvement of the fit until no variables can be added.
- Backward elimination

  Starting with all candidate variables, repeat the process excluding variable whose loss gives the most statistically insignificant deterioration of the model fit, until no variables can be removed.
- Bidirectional elimination

  A combination of the above.

AIC & BIC

For model includes k predictors and estimated likelihood function $\hat{L}$:

$\text{AIC} = 2k - 2\ln\hat{L}$

Table: variables selected

Forward	*full model*
Backward	*age, weight, andomen, thigh, forearm, wrist*
Bidirectional	*age, weight, andomen, thigh, forearm, wrist*

$\text{BIC} = k\ln n - 2\ln\hat{L}$

Table: variables selected

Forward	*full model*
Backward	*weight, andomen, wrist*
Bidirectional	*weight, andomen, wrist*

Approach 2: Exhaustive Search

Mallows' Cp $= \frac{SSE_p}{MSE(full)} - (n - 2p)$

Weight, Abdomen, Thigh and Wrist

Contents

1 Background

2 Data Visualization and Pre-process

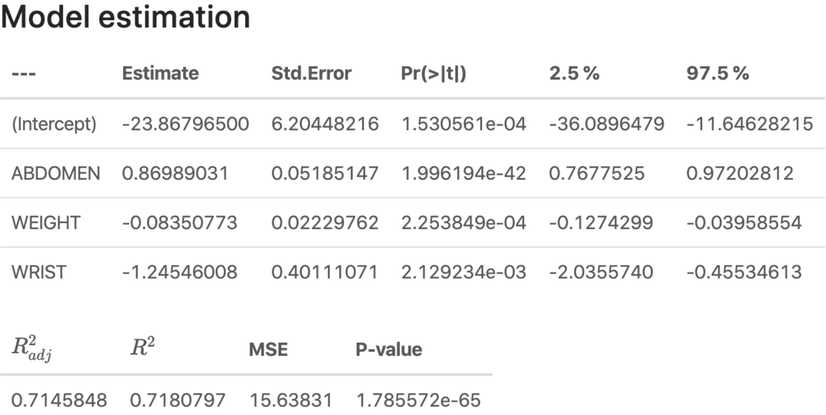
3 Modeling

4 Results

5 Strengths and Weaknesses

Final Model

According to the principle of simplicity, the second one is chosen to be our final model, and it is stated as follows. The Multiple R-squared is 0.718 and the Adjusted R-squared is 0.715.



Model estimation

---	Estimate	Std.Error	Pr(>|t|)	2.5 %	97.5 %
(Intercept)	-23.86796500	6.20448216	1.530561e-04	-36.0896479	-11.64628215
ABDOMEN	0.86989031	0.05185147	1.996194e-42	0.7677525	0.97202812
WEIGHT	-0.08350773	0.02229762	2.253849e-04	-0.1274299	-0.03958554
WRIST	-1.24546008	0.40111071	2.129234e-03	-2.0355740	-0.45534613

R^2_{adj}	R^2	MSE	P-value
0.7145848	0.7180797	15.63831	1.785572e-65

Diagnostic

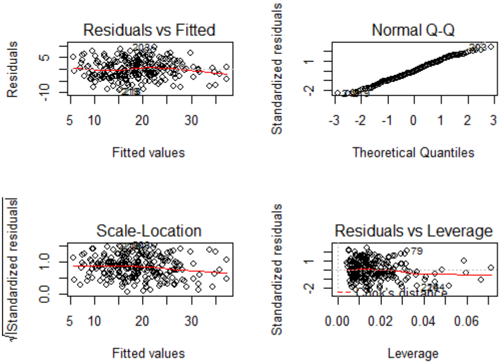


Figure: Diagnostic plot

Contents

1. Background

2. Data Visualization and Pre-process

3. Modeling

4. Results

5. Strengths and Weaknesses

Strengths and Weaknesses

Strengths

- The model has relatively high accuracy.
- The model is easy to be applied in real life.

Weaknesses

- For people who don not know their ABDOMEN and WRIST, it's hard for them to get a precise outcome.
- Due to the limitation of sample size, prediction range is relatively narrow.