

Module 2 Report

Group 8

Lu Li
Jiawei Wu
Nan Yan
Youhui Ye

Contents

- 1 Data Visualization
- 2 Modeling
- 3 Results
- 4 Strengths and Weaknesses

Target variable visualization

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

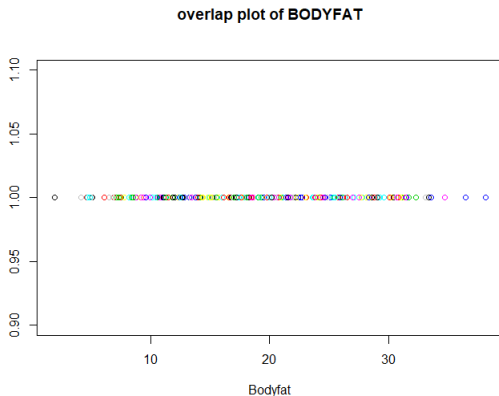


Figure: Bodyfat Variable

Data pre-process

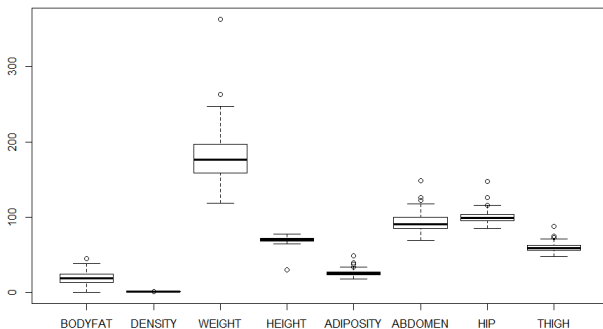


Figure: boxplot

Here we consider extreme values as potential outliers.

Potential outliers

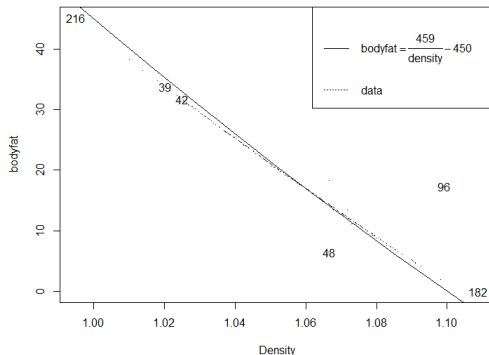


Figure: Bodyfat vs. Density

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

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.

Contents

- 1 Data Visualization
- 2 Modeling**
- 3 Results
- 4 Strengths and Weaknesses

Approaches– 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} :

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

Table: variables selected

Forward	<i>full model</i>
Backward	<i>age, weight, andomen, thigh, forearm, wrist</i>
Bidirectional	<i>age, weight, andomen, thigh, forearm, wrist</i>

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

Table: variables selected

Forward	<i>full model</i>
Backward	<i>weight, andomen, wrist</i>
Bidirectional	<i>weight, andomen, wrist</i>

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

Mallows' $cp =$

Contents

- 1 Data Visualization
- 2 Modeling
- 3 Results**
- 4 Strengths and Weaknesses

Contents

- 1 Data Visualization
- 2 Modeling
- 3 Results
- 4 Strengths and Weaknesses