Body Fat Calculator

Hao Tong, Yuan Cao, Shushu Zhang

October 5, 2020

1 Overall Findings & Rule of Thumb

1.1 Rule of Thumb

$-24.42-0.12*WEIGHT-0.34*NECK+0.95*ABDOMEN+0.58*FOREARM-1.42*WRIST$

2 Statistical Analysis

2.1 Data Cleaning

In current context, it is natural to detect outliers by each variable seperately using techniques such as box plot. However, in data cleaning, we intend to impute the outliers that rise due to human errors instead of natural deviations in population distributions. In our case, the predictor variables have great redundancy (i.e. highly related) in this case. For example, "ADIPOSITY", "WEIGHT", and "HEIGHT" are strictly related with the following equation $ADIPOSITY(bmi) = BODYWEIGHT(kg) : BODYHEIGHT(m^2)$.. Anomalous observations that are supposed to be imputed or removed in advance, must have some inconsistency in different variables, while univariate outliers does not necessarily indicate anomalous observations. For example, observation No. 39 is way outlined in every variable, which will definitely be considered as an outlier if we consider variables seperately. However, it deviated in almost every variable at a similar degree, making it a reasonable observation of a highly overweight and obese person, thus the observation shouldn't been imputed.

In order to detect the anomalous observations to be imputed based on the aforementioned criteria, we introduce multivariate detections, namely, fitting linear regressions for every variable with respect to other related variables using Residuals or Cook's Distance as a anomalous measure. Then, we concentrate on the extreme (maximum/minimum) values of each variable to make sure they make sense with practical meaning of the certain variable. To be more specific, in step 1, we fit "DENSITY" with all other variables except "IDNO" and "BODYFAT" to make sure there is no obvious observation error in "DENSITY". Then, we fit "BODYFAT" with "DENSITY", resulting in two outliers *No. 48* and *No.*

96. In addition, we detect the relationship among "ADIPOSITY", "WEIGHT", and "HEIGHT" by the aforementioned equation, leading to an abnormal observation *No. 42*. Then, the rest of the variables are regressed similarly with all other variables except "IDNO", resulting in anomalous observations of *No. 31* and *No. 86* for "ANKLE", *No. 175* for "FOREARM". We impute all of the aforementioned anomalous observations with fitted values in the linear model or the defined equation. We can even infer the error mechanisms for some imputations. No. 42, which has anomalously low value only in "HEIGHT", can be imputed by the induced (or say redundent) variable "ADIPOSITY" defined as $ADIPOSITY(bmi) = BODYWEIGHT(kg) : BODYHEIGHT(m^2)$. As a result, "HEIGHT" OF No. 42, 29.50, is replaced by 69.45, indicating 6 is probably mistyped by 2. Note that No. 163 and No. 221 also have slight mismatches in the relation of "ADIPOSITY" with "WEIGHT" and "HEIGHT". But all three variables are in reasonable ranges, making it hard to determine the "bad" observations. Therefore, we determine to leave as they are.

In step 2, we use boxplot to detect the outliers in the target variable "BODYFAT", resulting in finding the higher-end outlier *No.216* and lower-end unreasonble point *No. 182* with zero body fat. Owing to the uncertainty of the anomalous mechanism, we simply remove these two observations.