Project 1 Requirements

Body fat percentage refers to the relative proportions of body weight in terms of lean body mass (muscle, bone, internal organs, and connective tissue) and body fat. The ideal body fat percentage varies considerably by gender and by age. Persons with higher than the recommended range of body fat are considered to be at increased risk for disease[1].

The most accurate means of estimating body fat percentage are cumbersome and require specialized equipment. For instance, the “gold standard, is completely submerging a person in water and calculating the volume of the displaced water. This is what this looks in practise[2]:

Image from: https://www.fau.edu/education/academicdepartments/eshp/images/underwater.jpg

Physicians estimate body fat percentage from anatomical measurements (e.g., abdomen circumference) that are much easier to obtain. In this project, you are given a dataset of 13 measurements from 225 subjects (all men) along with their bodyfat percentage. Your goal is to come up with a formula that can be used to estimate bodyfat percentage using only the 13 measurements. (The “density, column will not be used in this project.)

The challenge is to identify which combination of the 13 predictors will give the most accurate estimate and if transforming some of the variables will increase accuracy. You are encouraged to try a few different combinations of predictors. You will want to use some domain knowledge to pick the predictors.

[1] pennshape.upenn.edu/files/pennshape/Body-Composition-Fact-Sheet.pdf

[2] A full video is here: https://www.youtube.com/watch?v=kgIIcATPQWI

Evaluation

You should evaluate each combination of predictors using 10-fold cross-validation. Since you are estimating a continuous value, use mean squared error (MSE) as the evaluation metric.

An example of creating folds for cross-validation using the cut function in R is here: https://stats.stackexchange.com/questions/61090/how-to-split-a-data-set-to-do-10-fold-cross-validation

Submission:

1. Write a short report listing the different combinations of predictor variables you tried, and if you tried transforming any of the variables. The report should include a plot of the MSE after cross-validation for each of the combinations. [A PDF file]
2. A listing of your R code [.R file]
3. An R function of the following form that returns your best body fat percentage prediction [.R file]:

```
bodyfatpercentage <- function(Age, Weight, Height, Neck, Chest,
Abdomen, Hip, Thigh, Knee, Ankle, Biceps, Forearm, Wrist) {
# your code goes here
}
```

Due date:

Thursday 11/6, 5:30pm on Titanium. Submit three files: the PDF report, full code (.R), and bodyfatpercentage function (.R).

Group work:

You may work in groups of 1-3. Include all group member names in the PDF file. Only one person in the group needs to submit to Titanium.