Project 1 Requirements

Due: Tuesday 4/16, 5:30pm on Titanium.
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.

The attached dataset was collected as part of a long running study conducted by researchers at CSU Fullerton's College of Health and Human Development. The study collects data relating to the physical health of college students.

This project has two objectives:

1. Perform an exploratory data analysis of the dataset.
2. One of the variables is Total Fitness Factor Score (column with header "FF") which is computed using a formula (hidden from you). Come up with an approximation of Total Fitness Factor Score for a subject using only the other available variables.

Exploratory data analysis

all 5 models should be linear regression models
1 model should have all the 35 variables

Exploratory data analysis is:

1. Generating summary statistics (mean, median, any outliers, any missing data points) for a variable
2. Visualizing the values of a variable
3. Visualizing the relationship between pairs of variables

The dataset contains approximately 35 variables, so it is not expected that every variable or every pair of variables will be explored. For this project, it is sufficient to consider any 5 variables and any 5 pairs of variables.

Approximating Total Fitness Factor Score

The challenge is to identify which combination of the nearly 40 potential predictors will give the most accurate estimate and if transforming some of the variables will increase accuracy. You should explore at least 5 different combinations of predictors and choose the best combination. You will want to use some domain knowledge to pick the predictors. The attached Data dictionary file gives information on the units and meaning of the different columns. Note this file will *not* be read into the R code.

Evaluation

You should evaluate each combination of predictors using 10-fold cross-validation. Since you are estimating a continuous value, use root mean squared error (RMSE) 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 that includes [a PDF file]:
    a. The exploratory data analysis (of your selected variables)
    b. The different combinations of predictor variables you tried to approximate Total Fitness Factor Score, and if you tried transforming any of the variables. The report should include a plot of the RMSE after cross-validation for each of the combinations.
2. A listing of your R code [.R file]
3. An R function of the following form that returns your best Total Fitness Factor Score approximation. Name the file totalfitnessfactorscore.R. Hint: this function will be based on the coefficients of your best linear model and will have the form

```
totalfitnessfactorscore <- function(Sex, Age, Weight, Height, Stress,
...) {
# formula for computing fitness factor goes here
}
```

Grading:

Grading will be based on how complete your report is and if the R code does the analysis correctly. It will NOT be based on how accurate is your approximation to total fitness factor score (though you can strive for higher accuracy).

Acknowledgment

The data was provided by Dr. Bill Beam (Department of Kinesiology), Dr. Archana McEligot (Department of Public Health), and Dr. Sinjini Mitra (Department of Information Systems and Decision Sciences).