18 April 2019

**Modeling Goal Refinement**

Previous goal: Predict the cadence at which a child will go from a walk to a run.

Refined goal: Assign a probability to which “gait state” an individual is in based upon their individual characteristic and their current cadence.

**Modeling Technique**

Logistic Regression. This logistic regression models the ‘log odds’ (can be converted into a probability) that a person has transitioned from walking to running given all their data.

**Data Update**

We used the same data as described in the previous iteration of this document with one adjustment. We have each individual in the dataset twice. Once at their last walking cadence (when transition = 0) and once at their first running cadence (when transition = 1).

We considered all the possible factors:

Dependent Variable: Transitioned (0/1): 1 if transitioned to run, 0 if not.

Considered Independent Variables: Age, HeightCMAvg, WeightKGAvg, Tanita.Avg, BMICont, BMIz, Sex, Race, Obesecat, Cadence

**Model Building Process**

We built the following model using the “purposeful selection” technique detailed in the textbook “Applied Logistic Regression” by Hosmer, Lemeshow, and Sturdivant. We can detail this method if you like.

Step1: Univariate Analysis. Note significate individual factors

Step2: Fit all possible predictors that pass step 1 triage. Remove variables one at a time whose P Values are greater than .05.

Step3: Monitor coefficients to ensure none vary more than 20% from step to step.

Step4: After oscillating between steps 2 and 3, add each variable not selected in step 1 to the model one at a time to determine if any of these happen to be significant in light of the other factors.

Step5: Check for linearity in the logit – one factor at a time.

Step6: Check for interactions if researchers determine there may be some that are clinically possible.

Step7: Check fit of model.

**Model**

Based on an individuals Age,Height,Weight, BMIz, and cadence, we can solve for the probability that an individual has transitioned to run.

Admittedly, the coefficients of each input are difficult to understand because they are in terms of the ‘log odds’, but essentially as age, height, bmiz, and cadence increase, the log odd of an individual transitioning goes up. The opposite is true for weight (due to the sign of the coefficient).

Purpose: Predict the cadence at which a child will go from a walk to a run.

Data Cleaning:

1. In order to conduct our analysis, we selected rows at which the child began to run. This was the fastest pace for all data as we understand that this is the point where kids went from walking to running.
2. There were two rows (one subject: ID# 68152) which contained missing information about sitting height.
   1. We did not include this subject in our analysis.
   2. However, after conclusion of our analysis, some of our models did not include Sitting Height as a predictor that impacted the cadence in which kids went from walking to running.
   3. Presently we have decided to keep this individual out of all models for consistency, but I do not feel there is a mathematical or statistical reason that this is necessary.
   4. The biggest factor for consideration of re-inserting this data point is: Does the fact we do not have data for this person reflect some sort of unique trait about this person?
3. Features used in model selection: 10 features and 1 response out of 21 total variables.
   1. Used Cadence\_stepsmin (1)
   2. Sex, Age\_years, Race, HeightCMAvg, WeightKGAvg, WaistCMAvg, leglengthCM, Tanita.Avg\_percentbodyfat, BMI\_rawscore, Obese\_status (10)
   3. Due to very high correlation with BMI\_rawscore, we removed BMI\_percentile and BMI\_zscore (2)
   4. Due to very high correlation, we only used HeightCMAvg and not SittingHeightCMAvg (1)
   5. Removed Agecat because it was a less granular version of age.
   6. Removed id as it is unique to each individual (1)
   7. Stage/TreadmillSpeed\_MPH because they are a different version of the Cadence\_stepsmin (2)
   8. Removed V02mlkgmin/METSYouth2/METSAdult because they require an invasive measurement and note practically useful for prediction (2)

Model 1: Linear Regression

1. We created a ‘best subsets’ regression of the above predictors using all the data. Best subsets were selected by lowest ‘bic’ which takes into consideration error, number of samples, and features
2. The best model is below. (5 predictors)
   1. Cadence\_stepsmin~Sex+Age\_years+WaistCMAvg+Racelimited+Tanita.Avg\_percentbodyfat
   2. BIC is -7.789.
   3. Leave one out cross validation RMSE is 18.95053.
3. Compare with second best model 6 predictors:
   1. Cadence\_stepsmin~Sex+Age\_years+WaistCMAvg+Racelimited+Tanita.Avg\_percentbodyfat +obeselimited(only ‘obese’)
   2. b. BIC is -7.6769
   3. c. Leave one out cross validation RMSE is 18.73
4. If we limit to only 3 predictors (for simplicity):
   1. Cadence\_stepsmin~Age\_years+WaistCMAvg+Racelimited
   2. BIC is -3.1176
   3. Leave one out cross validation RMSE is 19.967

Model 2: Survival Analysis

1) Best Subsets is not available for Survival Analysis so we used a ‘purposeful selection’ method of arriving at the best model. We tried all factors univariately. All factors that had a p-value below .25, we retained for the ‘saturated model.’ Also, compared with a ‘full model’ and factors that we would have removed from the univariate analysis – if significant in the full model, we retained to begin.

2) This may not be as relevant, because the model coefficients should be interpreted as ‘for every one unit increase in feature x is betax times more likely to ‘run’ than before.

3) The ‘best’ model according to the purposeful selection is:

Survival time ~ Sex+Age\_years+Race2+HeightCMAvg+WeightKGAvg+WaistCMAvg+ Tanita.Avg\_percentbodyfat+BMI\_rawscore+Obese\_status2

Model 3: Cluster Analysis

1. We wanted to see if there is any underlying structure in the static variables (those that don’t change in a subject over short duration). Clustering at high dimensions loses some of its benefit, so we looked at variables with high covariance and eliminated a few of them from our cluster analysis. The highest covariance was between waist and BMI (.958). Waist also had a high covariance with Weight (.919), so we eliminated it. BMI was strongly correlated with Tanita and weight as well (.83 and .889 respectively) so that was also eliminated. Finally, we took out Height as it correlated strongly with Age and leglength (.845 and .947, respectively).
2. We were then left with four primary factors to cluster: Age, leglength, Weight, and Tanita. We looked at a pair-wise depiction of these variables and only one pairing (Weight to Leglength) seemed to have some cluster aspects. However, after further analysis, this secondary cluster was only males as they had the larger height elements.
3. We used a GMM (Gaussian Mixture Model) in order to find the best possible model with the optimal number of clusters based on a BIC metric. The model assigns a vector of probabilities associated to each observation indicating the posterior probability of belonging to the respective clusters.
4. After doing this clustering, we then put the clusters into a linear regression model, using their assigned cluster as a factor. The results of this were inconclusive in that the cluster did not help in predicting run cadence time in children.