BIOS 6612 Homework 1: Model Selection

The goal of the NEJM paper "Hyponatremia among Runners in the Boston Marathon" was to identify the principal risk factors of hyponatremia, a life-threatening illness among marathon runners. Hyponatremia in this data set is defined as a binary variable based on serum sodium concentration of 135 mmol per liter or less. For this homework, you will analyze serum sodium concentration as a continuous variable since you feel that some information may be lost by dichotomizing. You want to examine covariates that significantly predict decreases in serum sodium concentration.

The dataset includes the following variables:

- `sodium`: serum sodium concentration, **the outcome of interest**
- `bmi`: body mass index
- `howmany`: number of prior marathons run
- `fluidfr3`: fluid frequency through marathon (1=every one mile, 2=every two miles, 3=every third mile or more)
- `runtime`: time taken to run the marathon in minutes
- `trainpse`: training pace for a one-mile run in seconds
- `wtdiff`: weight change during the marathon
- `age`: age in years
- `female`: gender (1=female and 0=male)
- `lwobup01`: NSAID usage (1 if reported use of nonsteroidal anti-inflammatory medications and 0 otherwise)
- `wateld01`: water loading (1 if water loading prior to the race and 0 otherwise)
- `urinat3p`: urination (1 if urinated three or more times during the race and 0 otherwise)

Answer the following questions based on your analysis of this data set; raw output from R or SAS is not acceptable. **Turn in the code used for analysis with your answers.**

1. First consider transforming covariates and the outcome.

   (a) The original paper categorized BMI into 3 groups (`bmiC`=1 if BMI $>$ 20 *and* BMI $<$ 25, `bmiC`=2 if BMI $<$ 20, and `bmiC`=3 if BMI $>$ 25). This was done because BMI has a quadratic relationship with hyponatremia and the polynomials terms are collinear. **Is categorization necessary in this case?** Justify your answer.

   (b) The original paper dichotomized the number of previous marathons run (`howmany`) at the median due to model fit. **Should the number of previous marathons run be dichotomized?** Justify your answer.

(c) The original paper examined if there was a quadratic relationship between weight change (`wtdiff`) and hyponatremia. **Is there a quadratic relationship between weight change and sodium levels?** Justify your answer.

(d) Fluid frequency (`fluidfr3`) has 3 levels $(1, 2, 3)$. **Should fluid frequency be treated as a continuous variable or 2 indicator variables?** Justify your answer.

(e) The original paper was concerned that there was an issue of collinearity with the fluid variables (`fluidfr3`, `wtdiff`, `wateld01`, and `urinat3p`). **Therefore, they only used weight change and excluded the self-reported variables from the multivariable analysis. Is this an issue?** Justify your answer.

(f) The original paper was concerned that there was an issue of collinearity with the running variables (`runtime` and `trainpse`), **so only running time was used in the multivariable model and not training pace since it is self-reported. Is this an issue?** Justify your answer.

(g) **Should the outcome sodium levels be log transformed?** Justify your answer.

2. Run the single variable analyses.

   (a) Run the analysis of each variable with sodium levels separately. **Which variables are associated with sodium levels at the 0.05 level of significance?** (Give the description of the variable, not the variable name.)

   (b) **How do these univariate analyses compare to the original paper where sodium levels were dichotomous?**

3. Now consider multivariable analyses. You want to examine covariates that significantly predict serum sodium concentration. For approach 1, fit a multivariable regression with all the predictors that had a p-value less than 0.05 in question 2(a) and run stepwise regression based on AIC.

   (a) **What predictors are included in the final model?**

   (b) **What are some issues with this approach?**

4. For approach 2, fit the full model and then perform a partial F test with all covariates with a p-value less than 0.1, making sure to maintain the hierarchy principle.

   (a) **What predictors are included in the final model?**

   (b) **What are the results of the F test?**

5. Now think about how the final models in questions 3 and 4 compare to the final model chosen in the original paper. **Why do you think that there are more significant covariates in the final model for a binary outcome then there are for a continuous outcome?** Hint: how many subjects were used in the analyses for sodium levels binary and how many subjects were used in the analyses for continuous sodium levels?