BIOS 6612 Homework 1: Model Selection
Solutions

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) (5 points) 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.

   Answer: BMI is not associated with sodium levels in the univariate regression (with BMI as a linear term), but BMI has a significant quadratic relationship

with sodium levels. However the VIF is greater than 155 in the quadratic model. Therefore, BMI needs to be categorized in order to accommodate this quadratic relationship and avoid the issue of collinearity.

(b) (5 points) 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.

Answer: Dichotomizing the number of previous marathons run (howmany) improves the AIC. The decision to transform a variable should be made based on model fit and NOT p-values. Therefore, howmany should be transformed.

(c) (5 points) 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.

Answer: The diagnostic plots and AIC improve for a quadratic relationship with weight difference. It is possible that the quadratic relationship is driven by a few outliers, but without more information on these data points, it should be assumed that they are valid and therefore should not be ignored.

(d) (5 points) 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.

Answer: There is little difference in model fit if fluid frequency is treated as continuous or categorical. The AIC is similar for both models and the assumption of linearity looks justified. As a result, based on model fit considerations alone, fluid frequency could be treated as a continuous or categorical variable. However, treating this variable as categorical is to be preferred because of how discrete it is as well as the last category including 3 miles and up. A stronger argument for treating this as linear could be made if the exact frequencies were available.

(e) (5 points) 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.

Answer: The largest VIF of the model with the fluid variables is 1.09, so there is not an issue of collinearity here.

(f) (5 points) 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.

Answer: The correlation between runtime and trainpse is 0.79. The largest variance inflation factor is 2.7. Depending on the cutoff used for the VIF (some authors recommend a VIF cutoff of 2 or 10), it could be argued either way to include or exclude training pace. But it would be appropriate to include both covariates in the model.

(g) (5 points) **Should the outcome sodium levels be log transformed?** Justify your answer.

Answer: The diagnostic plots look similar both with and without the log transformation. Because it is easier to interpret a model with the outcome on its original scale, it is preferable not to log transform in this case since model fit is not affected.

2. Run the single variable analyses.

(a) (2 points) 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.)

Answer: gender (p=0.0004), fluid freq (p=0.0012), NSAID usage (p=0.0094), weight change (p¡0.0001), marathon time (p=0.0017), training pace (p=0.0013), BMI (p=0.0070)

(b) (5 points) **How do these univariate analyses compare to the original paper where sodium levels were dichotomous?**

Answer: Looking back at the original paper, they found (bolding variables that were significant also in our continuous analysis) **gender**, **BMI**, num. previous marathons, **training pace**, **race time**, **fluid frequency**, fluid volume, urination, and **weight change** to be significant. **NSAID usage** is not significant in the univariate analyses from the original paper although it is in our analysis.

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

Answer: The final model for sodium serum levels includes NSAID use (p-value= 0.09898), fluid consumption (p-value= 0.08642), low BMI (p-value= 0.02869), high BMI (p-value= 0.34469), weight gain (p-value $<$ 2e-16), and weight gain squared (p-value= 0.00177). In this multivariable model, weight gain has a significant quadratic relationship with sodium serum levels (p-value= 0.00177) and the low BMI category is significantly associated with sodium serum levels (p-value= 0.02869).

(b) (5 points) **What are some issues with this approach?**

Answer: The issues with this approach are that you are not comparing hierarchical models (wtdiff^2 can be included even if wtdiff is not), there are multiple comparisons issues, using a p-value < 0.05 is arbitrary, and a covariate with a p-value greater than 0.05 in the univariate analysis might be a strong predictor in the multivariable model.

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

Answer: The final model includes the low BMI category (p-value= 0.03399), the high BMI category (p-value= 0.25029), weight gain (p-value $<$ 2e-16) and weight gain squared (p-value= 0.00388).

(b) (2 points) **What are the results of the F test?**

Answer: The F statistic is 1.0797 on 9 numerator and 365 denominator degrees of freedom, for a p-value of 0.3769; therefore, we do not reject the null hypothesis and conclude that the simpler model is satisfactory.

5. (5 points) 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?

Answer: In the multivariable logistic regression analyses, hyponatremia was associated with weight gain, a racing time of more than 4 hours, and BMI extremes. This is similar to the continuous analysis except that we found a significant quadratic relationship with sodium serum levels and we did not find a significant relationship with racing time. The NEJM paper had 488 runners and we used a reduced dataset that had no missing values and only 370 subjects: because larger sample sizes are associated with greater precision (and therefore smaller standard errors), we would expect more significant coefficients to be found in the binary analysis.