A question of some interest in public health is the relationship of sodium to health. While most researchers believe that sodium intake impacts poorly on health, particularly blood pressure level, a recent publication suggested that sodium intake might actually be protective for mortality (1):
``...dietary sodium intake is inversely associated with all-cause and CVD mortality, and … dietary sodium/calorie ratio is directly associated with both mortality rates." This statement is based on the results of a multivariate proportional hazards model estimated from the NHANES I followup study; the variables to be included in the model were selected using a stepwise backward variable selection algorithm.

We examined whether this finding would be inferred had sensitivity analyses of the variable selection methods been used. In particular we applied bootstrap sampling algorithms to the same dataset using the same candidate variables. Additionally, we applied the same algorithms to a second dataset, the NHANES II mortality follow-up study to ascertain whether the algorithms result in similar findings in this dataset. We also examined in the context of the example an algorithm suggested by Sauerbrei and Schumacher (2) for using bootstrap sampling as a procedure for variable selection. The procedure first selects a “usual” automatic variable selection technique and then applies the chosen elimination techniques to a large number of bootstrap samples. The frequency with each variable was selected in the model is tabulated and a model composed of the most frequently selected variables by the automatic elimination procedure is built. The correlation among the potential predicting factors is examined and some of the redundant variables are eliminated. The final model is selected based on the conditional relative frequency of the remaining variables.

The variables considered were age, race (white or black), gender, prevalence of cardiovascular disease (CVD), systolic blood pressure (SBP), sodium intake, fat intake, sodium/fat ratio, and use of salt at the table.

We examined these issues within two datasets that had similar information on the candidate set of variables: The NHANES I (n=10,861), and the NHANES II (n=9,141). These datasets contained similar information on the set of candidate variables as well as followup for mortality.

We use a proportional hazards model relating the candidate variables to mortality with a backward stepwise variable-selection procedure. We used a probability to remove of 0.10 and probability to enter of 0.05. 1,000 bootstrap samples were drawn and the results tabulated.

The best model based on the original data for NHANES I included age, gender, race, CVD, SBP, use of salt at the table, and total fat. This model was selected in only 274 of 1,000 bootstrap samples. The distribution of the number of variables selected in the bootstrap samples is given in Table 1.

| Table 1. Number of variables selected in 1,000 Bootstrap Samples. NHANES I. |
|------------------|---|---|---|---|---|---|---|---|
|                | <5 | 5  | 6  | 7  | 8  | 9  | 10 |
| Frequency       | 7  | 283| 436| 239| 35 | 0  |    |

Age, gender, race, CVD, and SBP were selected in all the samples, but no readily apparent pattern of variables selected could be discerned for the remaining variables. One or more of the dietary variables appeared in 993
of the models selected from the bootstrap samples. The most often selected set of dietary variables included both fat and always use salt in the model; but this model was selected in 444 bootstrap samples. Additionally, when a variable was selected for inclusion, its sign was not constant. Sodium appeared as a variable in 234 samples, in 54 (23%) of these samples the coefficient was positive; the remaining samples selected negative coefficients.

The process was repeated using NHANES II data. The model selected using the observed data in this case included age, gender, CVD, and SBP, but no dietary variables. This model was selected in only 182 out of 1,000 bootstrap samples. Age, gender, race, CVD, and SBP were selected in all samples. Again, no readily apparent pattern of variables selected could be discerned in the 1,000 bootstrap samples. The most often selected model included sodium/fat ratio, sodium, and use of salt at the table in the model, but this model was only selected in 135 bootstrap samples.

While making the problem of multicollinearity much easier to discern, the bootstrap sampling did not produce easily interpretable results. Additionally, algorithm that use bootstrap sampling to augment variable selection algorithms may need to include the sign of estimated effects in the procedure.

References: