**Advanced Statistical Modeling of NHANES Data**

The areas to be investigated include:

1. **Categorization of continuous factors** – categorization of continuous factors (e.g. age) can be done to introduce a nonlinear response without having to specify the form of the nonlinearity (e.g. quadratic, cubic, or other functional form). There are many ways in which a variable could be categorized, including:
   1. historical comparison,
   2. “science” based (e.g. the categorization of blood pressure readings into “normal” and “high”),
   3. interval based (e.g. categorization of age into each decade),
   4. percentile based (e.g. categorization of age by quantiles),
   5. k-means clustering based, and
   6. principle component analysis based.

Each of these techniques has its advantages and its drawbacks. However, it seems that how such categorization should be done and how the categorization method effects the outcomes of the modeling have not been studied.

1. **Multi-model modeling** – Most modeling using the NHANES dataset studies the correlation of a factor such as urinary arsenic concentration on a response such as hypertension by first “correcting” the data for covariates such as gender, age, ethnicity, etc. Frequently the dataset for the covariates is much larger than it is for the specific factor of interest; for the covariates mentioned it is about twice as large as the urinary arsenic dataset. If the covariate dataset could be used to create a separate model, more complex models could be considered and then incorporated into the final model where the only additional term(s) would be those arising from the factor(s) being studied. Of particular interest here is the inclusion of two- and perhaps three-factor interactions in the covariate model since there is insufficient data in the smaller datasets to include these terms.
2. **Dataset subsetting** – Studies using the NHANES dataset frequently look at responses where only a few percentage of the population exhibits the response of interest. For example, in studies of chronic kidney disease (CKD) about 5% of the population has CKD. Because the of the imbalance in the responses between having and not having CKD, any model built using the entire dataset will more heavily weight the not having CKD observations. Subsetting of the dataset so that approximately equal numbers of observations from each are used should result in models that better capture the effects of the factors being considered on having CKD. Furthermore, having subsetted the data into a training data fraction, the remainder of the data can be used to estimate the predictive quality of the model created.
3. **Model reduction** – Statistical models of the NHANES dataset that include interaction terms or even multiple levels of categorical factors can result in models with 100 or more model terms, many of which are not statistically significant. There are techniques for model reduction that have been used within the response surface modeling community for determining which terms in the model should be kept based on, for example, minimizing the AICc (corrected Akaike Information Criteria) or the BIC (Bayesian Information Criteria). Applying this procedure results in models with fewer terms and with most, if not all, of the remaining terms being statistically significant. It also results in nonsignificant terms in the reduced models to be eliminated. Thus, the final model is typically more robust in being able to identify if the variate factor(s) is(are) significantly correlated with the response being studied.
4. **Tukey-type comparisons** – Currently, the CDC uses pair-wise comparisons and t-tests to establish whether or not the responses for different levels (or combinations of levels for different factors) are statistically the same or not.
5. **Model validation and assessment** – Functional form misspecification tests(White’s test), AUC, Confusion matrix, precision, recall and f1 score.