STAT 689: Flexible Regression Using R

**Text and Topics:** This course will be based, somewhat loosely, on the soon to be published book *Semiparametric Regression in R*, by J. Harezlak, D. Ruppert and M. P. Wand. As the title of the book indicates, there will be much use of the R programming framework for the analysis of data examples, as is true in other courses in the department. It is a follow up of *Semiparametric Regression* by D. Ruppert, M. P. Wand and R. J. Carroll (2003).

**Grade Policy:** There is a midterm and a final, along with regular homework. The homework will be 40% of the course grade, and midterm 30% and the final 30%. The final course grade will be based on the standard scale where a total of 90 to 100 percent will be an A, 80 to less than 90 percent will be a B, etc.

**Tentative Course Outline:**
- Spline regression basics: weeks 1-2
- Penalized Spline regression: weeks 3-7
- Generalized additive models: weeks 8-11
- Longitudinal and Grouped Data: weeks 12-14

**Background About the Course:**
*Semiparametric Regression* refers to a field of statistics that produces flexible models that allow for general regressions to take on many shapes. From the book “Regression is used to understand the relationships between predictor variables and response variables and for predicting the latter using the former. In parametric regression, the effect of each predictor has a simple form, for example, is a linear or exponential function, so that its overall shape is dictated by the model, not the data. In contrast, with nonparametric regression the model is flexible enough to allow any smooth trend in the data. Semiparametric regression combines parametrically modeled effects for some predictors with nonparametric modeling of the effects of the other variables.”

The course will cover many aspects of semiparametric regression, especially involving generalized linear models such as logistic regression, and including completely nonparametric regression, partially linear models, additive models and grouped data (including longitudinal data). Other topics will be discussed if time permits, including shape constraints, spatial models, robustness and accounting for missing observations. Multiple R packages will be used, including gamlss, nlme, mgev, quantreg, refund, rstan, VGAM, scar and SCAM. The HRW package that the authors are creating will be used extensively.

**Data:** The book includes many data sets that can be accessed directly from R: the current version of the introduction of the book includes 22 such data sets.

**Programming:** All programming will be in R. This is a course in analyzing data in R, and not a course about learning how to create R packages or to do sophisticated R programming. Students will be expected to analyze the data sets, and to comment current R code to make the code more self-contained. Students will be expected to analyze new data and, where appropriate, write code to do such analyses.
**Requirements:** Students will be expected to have taken the regression course STAT 408/608. We particularly recommend the book *A Modern Approach to Regression with R*, authored by Simon J. Sheather, as good background.

**ACADEMIC INTEGRITY STATEMENT:** "An Aggie does not lie, cheat, or steal or tolerate those who do." The Aggie Honor Council Rules and Procedures are available at [http://www.tamu.edu/aggiehonor](http://www.tamu.edu/aggiehonor).

**STATEMENT ON PLAGIARISM:** As commonly defined, plagiarism consists of passing off as one's own ideas, words, writing, etc., which belong to another. In accordance with this definition, you are committing plagiarism if you copy the work of another person and turn it in as your own, even if you should have the permission of that person. Plagiarism is one of the worst academic sins, for the plagiarist destroys the trust among colleagues without which research cannot be safely communicated. If you have any questions regarding plagiarism, please consult the latest issue of the Texas A&M University Student Rules, under the section "Scholastic Dishonesty."

**STATEMENT ON DISABILITIES:** The Americans with Disabilities Act (ADA) is a federal anti-discrimination statute that provides comprehensive civil rights protection for persons with disabilities. Among other things, this legislation requires that all students with disabilities be guaranteed a learning environment that provides for reasonable accommodation for their disabilities. If you believe you have a disability requiring an accommodation, please contact the Office of Disability Services in Room B118 of Cain Hall. The phone number is 845-1637.