Math 33110: Applied Regression Analysis Final Project

Instructions: Choose one of the two problems to answer. Submit a hard copy of your work to the Instructor's mailbox by Friday, June 24, 2016.

Problem A (Heteroscedastic Regression). This problem explores methods for joint regression modeling of the mean and variance functions. Given the responses Y_i and vectors of predictors $\mathbf{x}_i$ and $\mathbf{z}_i$, $i = 1, \ldots, n$, consider the heteroscedastic regression model

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i,$$

where $\varepsilon_i \sim N(0, \sigma_i^2)$ are independent with

$$g(\sigma_i^2) = \mathbf{z}_i^T \boldsymbol{\gamma}.$$

Here $g(\cdot)$ is a known link function and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are vectors of regression coefficients. See Verbyla (1993).

(a) Derive iterative algorithms for computing the maximum likelihood and residual (or restricted) maximum likelihood estimators.

(b) Implement and test your algorithms on the Minitab cherry tree data, which is available in the R package `dispmod` (data file: `minitab`).

(c) Summarize the advantages and disadvantages of the log and identity link functions.

(d) Discuss briefly how the methodology can be extended to the cases of (i) count responses and (ii) multivariate responses. Be sure to cite appropriate references.

Problem B (Citation Networks). The aim of this problem is to develop two types of regression models for citation networks. See Varin, Cattelan and Firth (2016) for the problem background. In the network of citations between journals, data are collected on the cross-citation counts (data file: `cross-citation-matrix10.csv`). Also available are nodal attributes including the immediacy index, impact factor, impact factor without self-citations, 5-year impact factor, article influence score, and two export scores based on the Stigler model (data file: `journal-scores.csv`).

(a) Following the perspective of Zhu et al. (2016), consider the cross-citation matrix as fixed. Describe a regression model to study the network effect and nodal effect for a univariate nodal response. Since temporal structure is not present in our setting, there is no need to consider the momentum effect.

(b) More often in network modeling, the edges are considered random observations. Describe a regression model to study the network effect and nodal effect for the number of citations. To model the network effect, refer to Fosdick and Hoff (2015) for obtaining a low-rank representation of the network in terms of node-specific network factors.

(c) Implement and test your methods on the journal citation data.