STAT 578 - Sample Final Exam

Instructor: Professor D. Wiens | Time: 3 hours

Instructions: Answer ANY SIX of these SEVEN equally weighted questions in the answer booklets supplied.

1. In a nonlinear regression model $\mathbf{Y} = \boldsymbol{\eta}(\boldsymbol{\theta}) + \boldsymbol{\varepsilon}$ with normally distributed errors, partition the $p \times 1$ vector $\boldsymbol{\theta}$ as $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)'$, where the subvectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ have p_1 and $p_2 = p - p_1$ elements respectively.

   (a) Write down expressions for

      i. the *conditional* likelihood region for $\boldsymbol{\theta}_1$;

      ii. the *profile* likelihood region for $\boldsymbol{\theta}_1$.

   (b) By using suitable approximations, obtain the normal approximation to the distribution of the LSE $\hat{\boldsymbol{\theta}}_1$.

2. Consider a nonlinear regression model $Y = \mathbf{x}'(\beta)\boldsymbol{\alpha} + \varepsilon$, where $\mathbf{x}$ and $\boldsymbol{\alpha}$ are p-dimensional vectors but each element of $\mathbf{x}$ is a function of the real variable β. Suppose that an initial estimate $\beta_{(0)}$ is available. Obtain an explicit expression for the first iterate $\beta_{(1)}$ obtained from the Golub-Pereyra algorithm.

3. Consider the problem of estimating an unspecified function $f(\mathbf{x})$ from data $\{\mathbf{x}_i, y_i\}_{i=1}^n$, with $y_i = f(\mathbf{x}_i) + \varepsilon_i$.

   (a) Write down the penalized regression problem whose solution is given by the *generalized additive model* fit. (You need not prove that this is the solution.)

   (b) Suppose that one were to approach this problem through *projection pursuit*, and thus modelled $f(\mathbf{x})$ as $\sum_{m=1}^{M} f_m\left(\boldsymbol{\alpha}_m^T \mathbf{x}\right)$ for suitable vectors $\boldsymbol{\alpha}_m$. If the functions $f_1, ..., f_M$ are given, how are the $\boldsymbol{\alpha}_m$ determined?

4. (a) Define the following terms:

      i. Generalized M-estimate of regression using Huber's ψ_c;

      ii. L1 estimate of regression;

      iii. Minimum Volume Ellipsoid estimates of multivariate location and scatter;

      iv. S-estimate of regression.

   (b) Define what is meant by the *breakdown point* of a regression estimate $\hat{\boldsymbol{\theta}}$.

   (c) Define what is meant by the *influence function* of a statistical functional $h\left(\hat{F}_n\right)$.

5. Describe explicitly the algorithm by which one computes a three-step Generalized M-estimate of regression.

6. (a) An estimate of the scale of the random errors in a regression can be defined, implicitly, as a statistical functional $S(\hat{F}_n)$ of the empirical distribution function through the equation

$$E_F\left[\chi\left(\frac{X}{S(F)}\right)\right] = 0,$$

for some function $\chi(\cdot)$. Suppose that the distribution function of these errors is F_0. Show that the influence function of S is then

$$IF\left[x; S, F_0\right] = \frac{S_0 \chi\left(\frac{x}{S_0}\right)}{E_{F_0}\left[\frac{X}{S_0}\chi'\left(\frac{X}{S_0}\right)\right]},$$

where $S_0 = S(F_0)$.

   (b) Apply relevant theory to determine the asymptotic distribution of $\sqrt{n}\left(S(\hat{F}_n) - S(F_0)\right)$.

7. Suppose that you are designing an experiment to estimate the parameters of a linear model $Y = \mathbf{x}'\boldsymbol{\theta} + \varepsilon$, with $\mathbf{x} = (x_1, x_2)'$. The independent variables x_1, x_2 must lie in the square $0 \leq x_1, x_2 \leq 1$, and observations are made sequentially. The first n observations $\{\mathbf{x}_i, y_i\}_{i=1}^n$ have resulted in a design matrix satisfying $\mathbf{X}'\mathbf{X} = \mathbf{I}$. If D-optimality is the goal, at what value of $\mathbf{x}$ should the next observation be taken?