COMPREHENSIVE WRITTEN EXAMINATION, PAPER III

FRIDAY AUGUST 21, 2009, 9:00 A.M.–1:00 P.M.

STOR 665 QUESTIONS

Regression models analyze relationships between an outcome y and a predictor x. (Multiple predictors can also be incorporated.) The simplest example is a linear model that assumes the relationship between y and x as

$$E(y) = \beta_0 + \beta_1 x.$$

The above linear model is parametric with two parameters β_0 and β_1. As a comparison, a nonparametric model would simply specify $E(y)$ as some function of x,

$$E(y) = f(x).$$

More often, the function $f(\cdot)$ is assumed to be "smooth" in some sense. Intuitively speaking, "smooth" means when x and x' are close, the corresponding functional values $f(x)$ and $f(x')$ should be close as well. The problem of estimating the function $f(\cdot)$ is the subject of *nonparametric smoothing*.

The following problems concern an approach that performs nonparametric smoothing through linear mixed models.

We first break the range of the observed predictor x into n intervals and group the data points into these intervals. Then, we denote the center of the ith interval as x_i and the outcomes of the data points in the ith interval as y_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, n_i$. Note that some n_i's may be zero.

The statistical problem is as follows. Given the observations (x_i, y_{ij}), we assume that

$$y_{ij} = f_i + e_{ij} \equiv \beta + b_i + e_{ij},$$

where $f_i \equiv f(x_i)$ and e_{ij}'s are iid $N(0, \sigma^2)$. The goal is to estimate $f = (f_1, \ldots, f_n)'$ under the constraint that f satisfies some smoothness condition (to be defined later).

1. **[5 points]** Let y denote the response vector and $b = (b_1, \ldots, b_n)'$. Denote

   $$E(y) = X\beta + Zb.$$

   Find out the expressions of X and Z.

2. **[5 points]** Assume b is fixed and suppose $\sum_i n_i b_i = 0$. Derive the estimates of β, b and f.

3. **[60 points]** Now suppose b_1 is fixed and

   $$b_i = b_{i-1} + e_i \tag{1}$$

   where e_i's are iid $N(0, \sigma_b^2)$.

(a) **[4 points]** Find the $(n-1) \times n$ matrix Δ such that

$$\Delta b = (b_2 - b_1, b_3 - b_2, \ldots, b_n - b_{n-1})'.$$

(b) **[6 points]** Write down the conditional log-likelihood function of b given b_1, denoted as $\log p(b|b_1)$.

(c) **[6 points]** The function $\log p(b|b_1)$ involves a term of the form $b'Wb$. Find out the expression of W, showing the entries.

(d) **[6 points]** Write down the log-likelihood function based on y and b.

(e) **[4 points]** Given σ^2 and σ_b^2, find out the part of the log-likelihood that needs to be maximized to derive the MLE for β and b.

(f) **[6 points]** Given σ^2 and σ_b^2, the normal equation for β and b is of the following form,

$$U(\beta, b)' = V.$$

Find out the expression of U and V.

(g) **[6 points]** Show that the matrix U is singular.

(h) **[8 points]** The singularity of U implies that we need to specify a value for β to make the model identifiable. Under the natural choice that $\hat{\beta} = \bar{y}$, show that the estimate for f can be written as

$$\hat{f} = S\tilde{y},$$

where $\tilde{y} = (\bar{y}_1, \ldots, \bar{y}_n)'$. (Set $\bar{y}_i = 0$ if $n_i = 0$.) Identify the expression for S.

(i) **[7 points]** Show that the row sums of S are always one. What does this indicate regarding the interpretation of $\hat{f}$?

(j) **[7 points]** Write a *clear and concise* summary based on the above questions. The summary should state the differences between the fixed model and the mixed model, and offer explanations why the mixed model formulation helps to estimate a smooth f.

4. **[30 points]**

The above questions focus on the cases where y is normally distributed. The approach can be extended for nonparametric smoothing for non-normal data.

Consider the following problem. We group a sample of US voters into n equal length bins based on their age, and for the ith bin, let n_i denote the number of voters, y_i the number of voters supporting President Obama, x_i the medium age, and p_i the supporting rate for the president. We want to model $\text{logit}(p_i)$ as a smooth function of x_i, denoted as $f(x_i)$.

(a) **[8 points]**

Following the earlier framework and assuming Model (1) for the random effects b, write down the corresponding mixed model formulation for the current problem.

(b) **[8 points]**

Write down the log-likelihood based on y and b.

(c) **[14 points]**

Derive the iterative re-weighted least-squares procedure for estimating β, b and f. Clearly indicate the final updating formula.