COMPREHENSIVE WRITTEN EXAMINATION, PAPER III

FRIDAY AUGUST 15, 2008, 9:00 A.M.–1:00 P.M.

STOR 665 QUESTIONS

1. **[10 points]**

   Suppose that $Y|P \sim \text{Binomial}(m, p)$, and P has the beta distribution

   $$f_P(p) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}.$$

   Prove that

   $$\text{E}(Y) = m\pi,$$

   $$\text{var}(Y) = m\pi(1-\pi)\{1+(m-1)\tau^2\},$$

   where π and τ^2 are expressed in terms of α and β. Derive the expressions for π and τ^2.

   Hints: The following facts are useful:

   - The above beta random variable P has mean $\frac{\alpha}{\alpha+\beta}$ and variance $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.
   - Consider two random variables X and Y. Then,
     - $\text{E}(X) = \text{E}(\text{E}(X|Y))$;
     - $\text{var}(X) = \text{E}(\text{var}(X|Y)) + \text{var}(\text{E}(X|Y))$.

2. **[10 points]**

   Consider multinomial random variables Y_i with c categories with

   $$P(Y_i = j) = p_{ij}, \quad \text{for} \quad j = 1, \ldots, c,$$

   where $\sum_j p_{ij} = 1$ for each $i = 1, \ldots, n$.

   Let $\{(y_i, x_i), i = 1, \ldots, n\}$ denote the observations where y_i represents the observed values of Y_i and x_i denotes a vector of covariates.

   Nominal logistic models can be used in this case to model the responses, which choose one of the response categories as the reference category.

   Prove that the estimated probabilities $\{\hat{p}_{ij}\}$ do not depend on which category is chosen as the reference category.

3. **[80 points]**

Let the discrete random variables Y_i denote the ith binomial response associated with m_i number of trials. Let $\{(y_i, m_i, x_i),\ i = 1, \ldots, n\}$ denote observations where y_i represents the observed value of Y_i, and x_i is the covariate vector.

Suppose the success probability of Y_i depends on an *unobserved* random variable Z_i, representing the underlying component that generates Y_i. In addition, Z_i follows a multinomial distribution with c categories and

$$P(Z_i = j) = p_{ij}, \quad j = 1, \ldots, c.$$

Conditional on $Z_i = j$, we have

$$Y_i | Z_i = j \sim \text{Binoimal}(m_i, \pi_{ij}).$$

Furthermore, the set of "observations" (Y_i, Z_i) are sequentially independent.

(a) **[10 points]** Under the above model, prove that

$$\mathrm{E}(Y_i | x_i) = m_i \tilde{\pi}_i,$$

$$\mathrm{var}(Y_i | x_i) = m_i \tilde{\pi}_i (1 - \tilde{\pi}_i)(1 + (m_i - 1)\phi_i),$$

where $\tilde{\pi}_i$ and ϕ_i depend on $\{p_{ij}, \pi_{ij}, j = 1, \ldots, c\}$.

(b) **[10 points]** Derive the exact expressions for $\tilde{\pi}_i$ and ϕ_i in Question 3(a).

(c) **[5 points]** Compare the results in Questions 3(a) and 3(b) with the results under the binomial-beta model of Question 1. Explain how the current model generalizes the binomial-beta model in terms of modelling over-dispersion.

(d) **[5 points]** Suppose π_{ij} depends on the covariate vector $x_i^{(r)}$ that is part of the covariate vector x_i. Given $Z_i = j$, write down the logistic regression model for Y_i with α_i being the coefficient vector, and the corresponding likelihood function.

(e) **[5 points]** Write down the nominal logistic regression model for Z_i with category c as the reference category, $x_i^{(m)}$ as the covariate vector and β_j as the coefficient vector, as well as the corresponding likelihood function.

(f) **[10 points]** Write down the likelihood function of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ under the full model, where $\boldsymbol{\alpha} = (\alpha_1^T, \ldots, \alpha_c^T)^T$, $\boldsymbol{\beta} = (\beta_1^T, \ldots, \beta_{c-1}^T)^T$.

(g) As it turns out, the above likelihood is very complicated, which makes a direct maximization difficult. A nice alternative is to formulate the problem as a missing-value problem since the underlying component Z_i is indeed missing.

Suppose

$$z_{ij} = \begin{cases} 1, & \text{if } Z_i = j, \\ 0, & \text{otherwise,} \end{cases}$$

and denote $z_i = (z_{i1}, \ldots, z_{ic})^T$.

i. **[10 points]** Write down the likelihood function of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ given the full data $\{y_i, z_i\}$.

ii. **[15 points]** Suppose we have initial estimates as $\boldsymbol{\alpha}^{(0)}$ and $\boldsymbol{\beta}^{(0)}$. Derive the conditional expectation of z_i given the estimates and the data $\{y_i, m_i, x_i = (x_i^{(m)T}, x_i^{(r)T})^T, \ i = 1, \ldots, n\}$.

iii. **[10 points]** The maximum likelihood estimation can now be done via the EM algorithm. Describe the algorithm in as much detail as possible.