BASIC PHD WRITTEN EXAMINATION IN BIOSTATISTICS

THEORY, SECTION 2

(9:00 AM- 1:00 PM
Thursday, August 8, 2013)

INSTRUCTIONS:

a) This is a **CLOSED-BOOK** examination.

b) The time limit for this Examination is four hours.

c) Answer any TWO (2) (BUT ONLY TWO) of the THREE (3) questions that follow.

d) Put the answers to different questions on separate sets of paper.

e) Put your code letter, **NOT YOUR NAME**, on each page. The same code will be used for Section 1 and Section 2 of the PhD Theory Exam. Please keep the code confidential and do not share this information with any students or faculty. Sharing your code with either students or faculty is viewed as a violation of the UNC honor code.

f) Return the examination with a signed statement of the UNC honor pledge, separately from your answers. The pledge statement is given on the last page of the exam handout.

g) In the questions to follow, you are required to answer only what is asked, and not to tell all you know about the topics involved.

1. Suppose that y_{ij} for $i = 1, \ldots, m$ and $j = 1, 2$ follow a Poisson mixed effects model:

$$y_{ij}|u_1, \ldots, u_m \sim \text{Poisson}(\lambda_{ij}), \quad \log(\lambda_{ij}) = x_{ij}^T\beta + u_i,$$

where x_{ij} is a $p \times 1$ covariate vector, β is an unknown $p \times 1$ parameter vector, and $u_1, \cdots, u_m$ are independent and identically distributed. Let $z_i = \exp(u_i)$ for all i and define γ to be the coefficient of variation of z_i, that is,

$$\gamma = \frac{\sqrt{\text{Var}(z_i)}}{E(z_i)}.$$

(a) Show that $\text{Var}(y_{ij}|x_{ij}) = \mu_{ij}(1 + \gamma^2\mu_{ij})$ and $\text{Cov}(y_{ij}, y_{ik}|x_{ij}, x_{ik}) = \gamma^2\mu_{ij}\mu_{ik}$ for $j \neq k$, where $\mu_{ij} = E(y_{ij}|x_{ij})$.

(b) It is assumed that the $z_i \sim \text{Gamma}(\alpha, 1/\alpha)$, for some unknown scalar $\alpha > 0$. Calculate μ_{ij} and γ^2. Write down the likelihood for (β, α) and show that it can be expressed in closed form using the gamma function.

(c) Suggest an algorithm to calculate the maximum likelihood estimator of $\theta = (\beta, \alpha)$, denoted by $\hat{\theta}_M = (\hat{\beta}_M, \hat{\alpha}_M)$. Derive the asymptotic distribution of $\hat{\beta}_M$. Please give the explicit form of the asymptotic covariance of $\hat{\beta}_M$.

(d) Suppose that $\hat{\beta}_E$ is the solution of a set of estimating equations for β given by

$$\sum_{i=1}^{m} \frac{\partial \mu_i}{\partial \beta}^T (y_i - \mu_i) = 0_p,$$

where $\mu_i = (\mu_{i1}, \mu_{i2})^T$, $y_i = (y_{i1}, y_{i2})^T$, and 0_p is a $p \times 1$ vector of zeros. Derive the asymptotic distribution of $\hat{\beta}_E$.

(e) Rigorously compare the asymptotic covariances of $\hat{\beta}_E$ and $\hat{\beta}_M$. Which estimator is more efficient? Are there scenarios where the asymptotic covariances are equal?

2. Consider the linear model

$$Y = X\beta + Z\gamma + \epsilon,$$

where Y is $n \times 1$, X is $n \times p$ of rank p, Z is $n \times q$ of rank q, β is an unknown $p \times 1$ parameter vector, γ is $q \times 1$, $\epsilon \sim N_n(0, R)$, $\gamma \sim N_q(0, D)$, R and D are positive definite matrices, ϵ and γ are independent, and $N_n(a, b)$ is an n variate normal random variable with mean vector a and covariance matrix b.

(a) For known R and D, the distribution of $Y|X, \gamma \sim N(X\beta + Z\gamma, R)$. Derive the marginal distribution of $Y|X$.

(b) In the following, continue to assume that R and D are known and treat γ as an unknown parameter in $Y|X, \gamma$.

(i) Show that the predictor of γ given by $\hat{\gamma} = DZ'V^{-1}(Y - X\hat{\beta})$ satisfies the conditional likelihood equations for (β, γ), where $\hat{\beta}$ is the MLE of β and $V = ZDZ' + R$.

(ii) Derive the exact distribution of $\hat{\gamma}$.

(iii) Show that $\hat{\gamma}$ is the best linear unbiased predictor of γ.

(c) Now suppose that R is of the form $R = \sigma^2 I_n$, where I_n is the $n \times n$ identity matrix, and β, σ^2, and D are unknown. Devise a detailed EM algorithm for jointly estimating (β, σ^2, D).

(d) Next, consider the case that D, R, and β are unknown and that R has a general structure. Define $A = I_n - M$, where M is the orthogonal projection operator on the column space of X, and write $W = B'Y$ where $A = BB'$ and $B'B = I_n$. Consider estimation of the unknown parameters using the marginal distribution of $Y|X$ in (a).

(i) Let $\hat{\beta}$ denote the MLE of β when (D, R) are fixed. Show that $\mathrm{Cov}(W, \hat{\beta}) = 0$.

(ii) Use the result in (i) to derive the density of W.

(iii) Devise a joint estimation scheme for (D, R, β) using (i) and (ii).

(e) For $i = 1, \ldots, n$, consider the linear model $y_i = x_i^T\beta + \gamma_i + \epsilon_i$, where y_i is a scalar random variable, x_i is a $p \times 1$ vector of covariates, β is the $p \times 1$ regression parameter, γ_i are i.i.d. $N_1(0, \tau^2)$ and ϵ_i are independent $N_1(0, \sigma^2 \exp(\lambda^T x_i^*))$, where λ is a $q \times 1$ unknown parameter vector, $q < p$, and x_i^* is a subset of the covariate vector which does *not* contain a constant (intercept) term, and σ^2 and τ^2 are scalar parameters. Assume further that γ_i is independent of ϵ_i, for $i = 1, \ldots, n$. The parameters in $(\beta, \tau^2, \sigma^2, \lambda)$ are unknown. Using the marginal distribution of $y_i|x_i$, derive the score test for the hypothesis $H_0 : \lambda = 0$ and state its asymptotic distribution under H_0.

3. We consider N independent random variables, denoted by $Y_1, , ..., Y_N$, from a population of N subjects. We assume $Y_i = \beta x_i + N(0, \sigma^2)$, where $x_1, ..., x_N$ are known positive constants and β and σ^2 are unknown scalar parameters. To estimate β and σ^2, we take a random sample of $Y_i, i = 1, \ldots, N$, from these N subjects. For $k = 1, \ldots, N$, we observe x_k and R_k, the indicator variable for whether or not Y_k for the kth subject is selected. We assume $R_1, ..., R_N$ are mutually independent and independent of (Y_i, x_i), $i = 1, ..., N$. For $k = 1, \ldots, N$, assume $P(R_k = 1) = \pi_k$ for some known constant $\pi_k \in (0, 1)$. Thus, only Y_i from selected subjects with $R_i = 1$ are observable.

   (a) Write the likelihood function for the observed data.

   (b) Compute the maximum likelihood estimator for β and σ^2, denoted by $\widehat{\beta}$ and $\widehat{\sigma}^2$ respectively. If the sample size is zero, define $\widehat{\beta} = 0$ and $\widehat{\sigma}^2 = 0$.

   (c) Derive the mean and variance of $\widehat{\beta}$.

   (d) Calculate the distribution of $\widehat{\beta}$.

   (e) Construct a confidence interval of level $(1 - \alpha)$ for β based on the conditional distribution of $\widehat{\beta}$ given $R_1, ..., R_N$.

   (f) Define $\widetilde{\beta} = \{\sum_{i=1}^N \frac{R_i}{\pi_i} Y_i\}\{\sum_{i=1}^N x_i\}^{-1}$. Show $\tilde{\beta}$ is an unbiased estimator for β and derive the variance of $\widetilde{\beta}$.

   (g) Find the optimal π_i to minimize $var(\widetilde{\beta})$ under the condition that the expected sample size is fixed at n, i.e., $\sum_{i=1}^N \pi_i = n$.

   (h) For any given finite function $g(\cdot)$ and $\pi_i, i = 1, \ldots, N$, show

$$\widetilde{\beta}(g) \equiv \frac{\sum_{i=1}^N g(x_i) + \sum_{i=1}^N \frac{R_i}{\pi_i} \{Y_i - g(x_i)\}}{\sum_{i=1}^N x_i}$$

   is unbiased for β and calculate its variance.

   (i) For given $\pi_i, i = 1, \ldots, N$, determine the optimal $g(\cdot)$ minimizing the variance of $\widetilde{\beta}(g)$. Suggest a method for estimating the optimal g using the observed data.

2013 PhD Theory Exam, Section 2

Statement of the UNC honor pledge:

"In recognition of and in the spirit of the honor code, I certify that I have neither given nor received aid on this examination and that I will report all Honor Code violations observed by me."

(Signed) _____

NAME

(Printed) _____

NAME