BASIC PHD WRITTEN EXAMINATION IN BIOSTATISTICS

THEORY, SECTION 2

(9:00 AM- 1:00 PM
Thursday, August 13, 2009)

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.

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.

Table 1: n pairs of binary observations

		Y_2		
		0	1	Total
Y_1	0	n_{00}	n_{01}	n_{0+}
	1	n_{10}	n_{11}	n_{1+}
Total		n_{+0}	n_{+1}	n

1. Table 1, a 2×2 contingency table, is based on n independent pairs of binary observations $(y_{i1}, y_{i2}), i = 1, \cdots, n$ from a cross-sectional study, where $Y_{ik} = 1$ denotes 'success' and 0 denotes 'failure' for $k = 1, 2$.

   (a) Assume $P(Y_{i1} = j, Y_{i2} = k) = \pi_{jk}$ for all i. Derive the maximum likelihood estimates of π_{jk}, denoted by $\hat{\pi}_{jk}$, based on Table 1 and show that

   $$\sqrt{n}(\hat{\pi}_{00} - \pi_{00}, \hat{\pi}_{01} - \pi_{01}, \hat{\pi}_{10} - \pi_{10}, \hat{\pi}_{11} - \pi_{11})^T$$

   converges in distribution to a multivariate normal random vector with zero-mean and covariance $\Sigma = \text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T$, where $\boldsymbol{\pi} = (\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})$.

   (b) Under the assumptions in (a), further assume that $\pi_{1+} = \pi_{11} + \pi_{10} = \exp(\alpha)/[1+\exp(\alpha)]$ and $\pi_{+1} = \pi_{11} + \pi_{01} = \exp(\alpha+\beta)/[1+\exp(\alpha+\beta)]$. Using the results from (a), construct an estimator for (α, β), denoted by $(\hat{\alpha}_M, \hat{\beta}_M)$. Use delta method to show that the asymptotic variance of $\sqrt{n}(\hat{\beta}_M - \beta)$ is

   $$(\pi_{1+}\pi_{0+})^{-1} + (\pi_{+1}\pi_{+0})^{-1} - 2(\pi_{11}\pi_{00} - \pi_{10}\pi_{01})/(\pi_{+1}\pi_{+0}\pi_{1+}\pi_{0+}).$$

   (c) Consider the subject-specific model

   $$P(Y_{i1} = 1) = \frac{\exp(\alpha_i)}{1 + \exp(\alpha_i)}, \quad P(Y_{i2} = 1) = \frac{\exp(\alpha_i + \beta)}{1 + \exp(\alpha_i + \beta)}.$$

   Assume independence of (Y_{i1}, Y_{i2}) across subjects, that is, across i. Show that $s_i = y_{i1} + y_{i2}$ is a sufficient statistic of α_i and that the conditional maximum likelihood estimate for β given $s_i, i = 1, \ldots, n$, is $\hat{\beta}_C = \log(n_{01}/(n_* - n_{01}))$, where $n_* = n_{01} + n_{10}$. Derive that the asymptotic variance of $\sqrt{n}(\hat{\beta}_C - \beta)$ is $1/\pi_{01} + 1/\pi_{10}$.

   (d) Next, consider the unconditional maximum likelihood estimation of $(\alpha_i, i = 1, \ldots, n, \beta)$ under the subject-specific model in (c). Show that the unconditional maximum likelihood estimator for β, denoted by $\hat{\beta}_{MLE}$, is inconsistent.

   (e) Show that $(\pi_{1+}\pi_{0+})^{-1} + (\pi_{+1}\pi_{+0})^{-1} \leq (\pi_{1+}\pi_{+0})^{-1} + (\pi_{+1}\pi_{0+})^{-1}$. Argue that $\text{var}[\sqrt{n}(\hat{\beta}_M)] \leq \text{var}[\sqrt{n}(\hat{\beta}_C)]$ when Y_{i1} and Y_{i2} are independent for each $i = 1, \ldots, n$, and $\alpha_i = \alpha$ are identical for $i = 1, \ldots, n$.

   Scoring: (a) 6 points; (b) 6 points; (c) 6 points; (d) 4 points; (e) 3 points.

2. We consider the model $Y_{ij} = \mu_i + (x_{ij} - \bar{x}_i)\gamma_i + \epsilon_{ij}$, $i = 1, 2$, $j = 1, \cdots, n_i$, where $n_i > 0$, $\bar{x}_i \equiv (1/n_i)\sum_j x_{ij}$, and μ_i and γ_i are scalar parameters. Suppose that x_{ij} are known scalars which are not all equal for each $i = 1, 2$. Further, suppose that $\{\epsilon_{ij}, i = 1, 2, j = 1, \cdots, n_i\}$ are assumed to be independent and identically distributed such that $\epsilon_{ij} \sim N(0, \sigma^2)$, where σ^2 is a scalar parameter.

   (a) Let $\beta = (\mu_1, \gamma_1, \mu_2, \gamma_2)^T$. We wish to write this model as $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, where $\mathbf{Y} = (Y_{11}, \ldots, Y_{1n_1}, Y_{21}, \ldots, Y_{2n_2})^T$, $\epsilon = (\epsilon_{11}, \ldots, \epsilon_{1n_1}, \epsilon_{21}, \ldots, \epsilon_{2n_2})^T$, and $\mathbf{X}$ is an appropriately defined matrix.

   (i) Specify $\mathbf{X}$ and the distribution of ϵ.

   (ii) Give the estimate (call it $\hat{\beta}$) for β which has minimum variance among the class of linear (in $\mathbf{Y}$) unbiased estimates.

   (b) (i) Specify a column vector $\mathbf{a}$ such that $\mathbf{a}^T\beta = (\gamma_1 - \gamma_2)$. Is $\mathbf{a}^T\beta$ estimable? Explain why or why not.
   (ii) Suppose σ^2 is known. Derive the distribution of $\mathbf{a}^T\hat{\beta}$ and give a $(1 - \alpha)$-level confidence interval for $\gamma_1 - \gamma_2$.
   (iii) Suppose σ^2 is unknown. Give a statistic to test $H_0 : \gamma_1 = \gamma_2$ and indicate its distribution under H_0.

   (c) (i) Do the results in (b)(ii) and (b)(iii) change if we fit the model under the restriction that $\mu_1 = \mu_2 = 0$? Justify your answer.
   (ii) Derive the least squares estimate of $\mu_i, i = 1, 2$ under the constraint that $\mathbf{a}^T\beta = 0$ [for $\mathbf{a}$ defined in (b)(i)]. Write the estimate in the scalar form explicitly, as opposed to the matrix representation.

   (d) Consider the linear model in the matrix form as in (a). Let Y_l denote the lth row of $\mathbf{Y}$ and $\mathbf{x}_l^T$ denote the lth row of $\mathbf{X}, l = 1, \ldots, n_1, n_1 + 1, \ldots, n_1 + n_2$. We are interested in deriving an F-test for the hypothesis that the observation in the kth row, Y_k, is not an outlier. Suppose we leave out the kth row from $\mathbf{Y}$ and $\mathbf{X}$ and we compute the least squares estimate, denoted by $\hat{\beta}_{(k)}$. Let $\mathbf{X}_{(k)}$ denote the design matrix obtained by deleting the kth row from $\mathbf{X}$ and let $\mathbf{Y}_{(k)}$ denote the vector obtained by deleting the kth row from $\mathbf{Y}$. Assume $\mathbf{X}_{(k)}$ is full rank. Define $D_k \equiv Y_k - \mathbf{x}_k^T\hat{\beta}_{(k)}$.
   (i) Give the matrix formulation for $\hat{\beta}_{(k)}$.
   (ii) Derive the distribution of D_k.
   (iii) Based on the distribution of D_k in (ii), provide an F-test for the hypothesis that Y_k is not an outlier (i.e. for the hypothesis $H_0 : E(Y_k) = \mathbf{x}_k^T\beta$).

Scoring: (a)(i)(2 points), (ii)(2 points); (b)(i)(2 points), (ii)(3 points), (iii) (3 points); (c)(i)(2 points); (ii)(2 points); (d)(i)(3 points), (ii)(3 points), (iii)(3 points).

3. To study the effect of a risk factor X on a count variable Y, data are collected from two clinical centers. For center $k = 1, 2$, the data, $(Y_{ik}, X_{ik}), i = 1, ..., n$, are i.i.d. from the following distribution: $X_{ik} \sim N(0, \sigma_k^2)$ and given X_{ik}, Y_{ik} follows a Poisson distribution having p.m.f. $\lambda_{ik}^y \exp(-\lambda_{ik})/y!, y = 0, 1, \ldots$, with $\lambda_{ik} = \exp(\alpha_k + \beta X_{ik})$, where $\sigma_k^2 > 0$. Both (σ_1^2, σ_2^2) and $(\alpha_1, \alpha_2, \beta)$ are unknown parameters.

   (a) Let $(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\beta})$ be the maximum likelihood estimator using all the data from the two centers. Find the asymptotic distribution of $\sqrt{2n}(\hat{\beta} - \beta)$ in terms of the true parameters. Hint:

$$\begin{pmatrix} A & B \\ B^T & C \end{pmatrix}^{-1} = \begin{pmatrix} (A - BC^{-1}B^T)^{-1} & -A^{-1}B(C - B^TA^{-1}B)^{-1} \\ -C^{-1}B^T(A - BC^{-1}B^T)^{-1} & (C - B^TA^{-1}B)^{-1} \end{pmatrix}.$$

   (b) In some practical situations, the individual level data may not be available and obtaining the maximum likelihood estimator in (a) is impossible. However, researchers from the two centers may report separate maximum likelihood estimators. Suppose $(\hat{\sigma}_k^2, \hat{\alpha}_k, \hat{\beta}_k)$ is the maximum likelihood estimator for $(\sigma_k^2, \alpha_k, \beta)$ using ONLY data from center $k = 1, 2$. In such situations, ONLY $(\hat{\sigma}_k^2, \hat{\alpha}_k, \hat{\beta}_k), k = 1, 2$, are available.

      (i) Find the asymptotic distribution of $\sqrt{n}(\hat{\beta}_k - \beta)$ in terms of the parameter $(\sigma_k^2, \alpha_k, \beta)$. Suggest a consistent estimator $\hat{V}_k$ of the asymptotic variance V_k using ONLY $(\hat{\sigma}_k^2, \hat{\alpha}_k, \hat{\beta}_k)$.

      (ii) To obtain a single estimator of β, one may consider $g(\hat{\beta}_1, \hat{\beta}_2)$, where $g(x, y)$ is a known, continuously differentiable, scalar function and $g(x, x) = x$ for any x. Find the asymptotic distribution of $\sqrt{2n}(g(\hat{\beta}_1, \hat{\beta}_2) - \beta)$ in terms of β, V_1, and V_2.

      (iii) Next, the goal is to find a function $g_{opt}(x, y)$ which satisfies the conditions in (ii) and which minimizes the asymptotic variance in (ii). Note that g_{opt} may not be unique. Write down the constraints which implicitly define $g_{opt}(x, y)$ and show that one such function is $g_{opt}^*(x, y) = (V_2 x + V_1 y)/(V_1 + V_2)$.

      (iv) Replacing V_1 and V_2 in $g_{opt}^*(x, y)$ in (iii) by $\hat{V}_1$ and $\hat{V}_2$ yields $\hat{g}_{opt}^*(x, y)$. Derive the asymptotic distribution of $\sqrt{2n}\left[\hat{g}_{opt}^*(\hat{\beta}_1, \hat{\beta}_2) - \beta\right]$. What is the asymptotic relative efficiency of $\hat{g}_{opt}^*(\hat{\beta}_1, \hat{\beta}_2)$ with respect to $\hat{\beta}$ given in (a)?

   (c) Now, suppose that we assume $\alpha_1 = \alpha_2 = \alpha$. Let $\hat{\beta}_r$ be the maximum likelihood estimator using the combined data from both centers. That is, we conduct maximum likelihood estimation using data from both centers, as in (a), under this additional restriction. What is the asymptotic distribution of $\sqrt{2n}(\hat{\beta}_r - \beta)$? Give a sufficient and necessary condition that $\hat{\beta}_r$ and $\hat{\beta}$ (in (a)) have the same asymptotic variances.

Scoring: (a) (5 points); (b) (i) (5 points), (ii) (3 points), (iii) (4 points), (iv) (4 points); (c) (4 points).