BASIC PHD WRITTEN EXAMINATION IN BIOSTATISTICS

THEORY, SECTION 2

(9:00 AM- 1:00 PM
Thursday, August 9, 2012)

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. After a certain surgical procedure, some patients develop a wound infection. Typically, the infection is treated and cleared. However, some patients develop another wound infection. The first infection is called the "primary infection", while the second is called a "secondary infection". An investigator is interested in the question whether the risk of a secondary infection in those who have had a primary infection is the same as the risk of a primary infection.

   Data are collected on a random sample of n patients. Assume that the n responses are independent and identically distributed. For the i-th patient, $1 \leq i \leq n$, let Y_{i1} denote a binary indicator of a primary infection and Y_{i2} a binary indicator of a secondary infection, both coded as 0 for "no", and 1 for "yes". Define $\alpha = P(Y_{i1} = 1)$ and $\beta = P(Y_{i2} = 1|Y_{i1} = 1)$. Both α and β take values in $(0, 1)$. Suppose there are X_1 patients with $Y_{i1} = 1, Y_{i2} = 1$; X_2 patients with $Y_{i1} = 1, Y_{i2} = 0$; X_3 patients with $Y_{i1} = 0, Y_{i2} = 0$. Note: $X_1 + X_2 + X_3 = n$. By definition, a secondary infection can occur only in patients who have had a primary infection.

   (a) Does the distribution of the data have the form of the exponential family? Give details.

   (b) Derive the maximum-likelihood estimators of α and β.

   (c) Derive the asymptotic covariance matrix of the estimators derived above.

   (d) Does there exist a UMP test for testing

   $$H_0 : \beta = 0.5 \quad \text{versus} \quad H_1 : \beta > 0.5?$$

   If so, then please find it. If not, then explain why such a test does not exist.

   (e) Derive the likelihood-ratio test statistic for testing

   $$H_0 : \alpha - \beta = 0 \quad \text{versus} \quad H_1 : \alpha - \beta \neq 0.$$

   (f) Derive the score test for the hypotheses in part (e).

   (g) Derive the Wald test statistic for the hypotheses in part (e).

   (h) Now, suppose we are interested in inference about β only, while considering α as a nuisance parameter. Derive a conditional likelihood for β which does not depend on α. Compute the maximum likelihood estimator for β and compare with the estimator for β in part (b). Is the result intuitive?

2. Consider the linear model

$$Y = X\beta + \epsilon, \tag{0.1}$$

where

$$E(\epsilon) = 0, \qquad \text{Cov}(\epsilon) = \Sigma, \tag{0.2}$$

X is $n \times p$ of rank $r \leq p$, β is $p \times 1$, Y is $n \times 1$, ϵ is $n \times 1$, (β, Σ) are both unknown and Σ is an unstructured positive semidefinite matrix. Let $\hat{\beta}$ be a least squares estimate (LSE) of β.

In the sequel, let $C(A)$ denote the column space of a matrix A, let $||a|| = \sqrt{a'a}$ for a column vector a, let $N_n(a, b)$ denote the n dimensional multivariate normal distribution with mean vector a and covariance matrix b, and let I_s be the s dimensional identity matrix.

(a) Let λ be a $p \times 1$ vector of scalars and let η be an $n \times 1$ vector of scalars. Let U be a conformable matrix such that $X'U = 0$ and $C(U) \cup C(X) = R^n$. Show that the following statements are all equivalent:

(i) $\lambda'\hat{\beta}$ is the best linear unbiased estimator (BLUE) of $\lambda'\beta$ for any $\lambda \in C(X')$.

(ii) $E(\lambda'\hat{\beta}\eta'Y) = 0$ for any $\lambda \in C(X')$ and any η such that $E(\eta'Y) = 0$.

(iii) $X'\Sigma U = 0$.

(iv) $\Sigma = XV_1X' + UV_2U'$ for some matrices V_1 and V_2.

(v) The matrix $X(X'X)^-X'\Sigma$ is symmetric, where $(X'X)^-$ denotes an arbitrary generalized inverse of $X'X$.

(b) Consider the model in (0.1) and (0.2) where $\epsilon \sim N_n(0, \sigma^2 I_n)$, X is of full rank, and β and σ^2 are unknown. We wish to estimate β under the loss function $L(\beta, a) = (\beta - a)'(\beta - a)$. Consider the estimator

$$\tilde{\beta} = \hat{\beta} - \frac{(p-2)\hat{\sigma}^2}{||(X'X)(\hat{\beta} - c)||^2}(X'X)(\hat{\beta} - c),$$

where $\hat{\sigma}^2 = SSE/(n - p + 2)$, SSE denotes the error sum of squares, and $c \in R^p$ is a column vector of fixed constants. Show that the frequentist risk of $\tilde{\beta}$ is smaller than the frequentist risk of $\hat{\beta}$.

(c) Consider the model in (0.1) and (0.2) with $\epsilon \sim N_n(0, \sigma^2 I_n)$, where β and σ^2 are unknown. We wish to conduct a hypothesis test of

$$H_0 : E(Y) \in C(X_0) \;\text{ vs. }\; H_1 : E(Y) \in C(X),$$

where $C(X_0) \subset C(X)$ for a known matrix X_0. In developing a test statistic, it is conjectured that a better estimator of σ^2 is $||P_2Y||^2/(n-q)$ since it is based on more degrees of freedom, where $q = \text{rank}(X_0) \leq r$ and P_2 is the orthogonal projection operator onto the orthogonal complement of $C(X_0)$. Consider the statistic

$$G = \left(\frac{n-q}{r-q}\right) \frac{||P_1Y||^2}{||P_2Y||^2},$$

where P_1 denotes the orthogonal projection operator onto $C(X) \cap C(X_0)^c$, and $C(X_0)^c$ denotes the complement of $C(X_0)$.

(i) Derive the distribution of G under H_0 and under H_1.

(ii) Derive the relationship of G to the usual F statistic for conducting the hypothesis test above.

(iii) Is G better than the usual F statistic in terms of statistical power? Justify your answer.

3. Consider independent observations $y_1, \ldots, y_n$, where $y_i = (y_{i1}, y_{i2})'$ is a bivariate binary random vector such that y_{ij} takes values 0 and 1 for $j = 1, 2$. Suppose that $y_i \sim QE(\theta, \lambda)$, where $QE(\theta, \lambda)$ is a bivariate binary distribution of quadratic exponential form

$$p(y_i|\theta, \lambda) = \Delta(\theta, \lambda)^{-1} \exp\{y_{i1}\theta_1 + y_{i2}\theta_2 + y_{i1}y_{i2}\lambda - C(y_{i1}, y_{i2})\},$$

where $\Delta(\theta, \lambda)$ is a normalizing constant and $C(y_{i1}, y_{i2})$ is a 'shape' function independent of $\theta = (\theta_1, \theta_2)'$ and λ.

(a) Derive both the marginal distribution of y_{i1} and the conditional distribution of y_{i2} given y_{i1}. Specify a sufficient and necessary condition such that y_{i1} and y_{i2} are independent.

(b) Calculate the marginal mean of y_i, denoted by $\mu = (\mu_1, \mu_2)' = E(y_i)$, the marginal product moment of $y_{i1}y_{i2}$, denoted by $\eta_{12} = E(y_{i1}y_{i2})$, and the marginal product centered moment of $(y_{i1} - \mu_1)(y_{i2} - \mu_2)$, denoted by $\sigma_{12} = E\{(y_{i1} - \mu_1)(y_{i2} - \mu_2)\}$.

(c) Calculate the Jacobian of the transformation from the canonical parameters θ and λ to the marginal parameters μ and η_{12}, denoted by $V = \partial(\theta, \lambda)/\partial(\mu, \eta_{12})$. Use V^{-1} to characterize the covariance matrix of $(y_i', y_{i1}y_{i2})'$ and specify a sufficient and necessary condition such that this transformation is one-to-one.

(d) Suppose that we also observe a $p \times 1$ column vector x_i for each i and that conditionally on x_i, $y_i \sim QE(\theta_i, \lambda_i)$, where $\theta_i = (\theta_{i1}, \theta_{i2})'$ and λ_i may depend on x_i, for $i = 1, \ldots, n$. Consider the model

$$E[y_i|x_i] = \mu_i = (\mu_{i1}, \mu_{i2})' = \mu(x_i, \beta), E[(y_{i1} - \mu_{i1})(y_{i2} - \mu_{i2})|x_i] = \sigma_{i12} = \sigma_{12}(x_i, \beta, \alpha),$$

where β is an unknown $p \times 1$ regression parameter and α is an unknown scalar parameter. Derive the likelihood score equations for $(\alpha, \beta')'$ and simplify them using the result obtained in part (c). Please clarify whether such estimating equations explicitly involve $C(y_{i1}, y_{i2})$.

(e) Consider generalized estimation equations for α and β given by

$$\sum_{i=1}^{n} \frac{\partial(\mu_i, \sigma_{i12})}{\partial(\alpha, \beta')} \frac{\partial \ell(y_i|\theta_i, \lambda_i)}{\partial(\theta_i, \lambda_i)} = 0$$

Compare the estimate of $(\alpha, \beta')'$ in part (d) with that in part (e) in terms of the statistical efficiency. To do so, provide an explicit comparison of the asymptotic variances of these estimators.

(f) Will the results in parts (a)-(e) be changed if y_{i1} and y_{i2} are continuous variables instead of binary variables? Please explain. If so, then derive the corresponding results and compare with those obtained above.

2012 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