BASIC PHD WRITTEN EXAMINATION IN BIOSTATISTICS

THEORY, SECTION 2

(9:00 AM- 1:00 PM
Thursday, August 12, 2010)

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.

1. Consider independent observations $(X_1, Y_1), \cdots, (X_n, Y_n)$, where Y_i takes values 0 and 1. Suppose that $X_i | Y_i = m \sim N(\mu_m, \sigma^2)$ and $P(Y_i = m) = \pi_m$ for $m = 0, 1$, where $\pi_0 + \pi_1 = 1$ and $\pi_0 \in (0, 1)$.

   (a) Show that $P(Y_i = m | X_i), m = 0, 1$ satisfies a logistic model, that is

$$\text{logit}(P(Y_i = 1 | X_i, \alpha)) = \alpha_0 + \alpha_1 X_i,$$

   where $\text{logit}(u) = \log(u/(1-u))$, $\alpha = (\alpha_0, \alpha_1)$, and α_0 and α_1 are unknown parameters. Derive the explicit form of $\alpha = g(\theta)$ as a function of $\theta = (\pi_1, \mu_0, \mu_1, \sigma^2)$.

   (b) Based on the logistic model in (a), please give the explicit form of the Newton-Raphson algorithm for calculating the maximum likelihood estimate of α, denoted by $\hat{\alpha} = (\hat{\alpha}_0, \hat{\alpha}_1)$, and derive the asymptotic covariance matrix of $\hat{\alpha}$.

   (c) Please write down the joint distribution of $\{(X_i, Y_i) : i = 1, \cdots, n\}$ and calculate the maximum likelihood estimate of θ, denoted by $\hat{\theta}_F$, and its asymptotic covariance matrix.

   (d) Calculate the asymptotic covariance matrix of $g(\hat{\theta}_F)$.

   (e) In this part, suppose that $\mu_0 = \mu_1$. Show that $\text{Cov}(\hat{\alpha})^{-1}\text{Cov}(g(\hat{\theta}_F))$ converges to a matrix, which does not depend on θ. Please interpret the results.

   (f) Now, suppose that π_1 is known. Will the results in (b)-(e) be changed? Please explain. If so, then please derive the corresponding results and compare with those obtained above.

2. Consider the following model:

$$Y_i = X_{i1}\beta_1 + X_{i2}\beta_2 + \ldots + X_{ip}\beta_p + U + \epsilon_i, \tag{0.1}$$

$i = 1 \ldots n$, where $\beta_1, \ldots, \beta_p$ are unknown parameters, $\mathrm{Y} = (Y_1, \ldots, Y_n)$ is the vector of responses, and $X_{ij}, i = 1 \ldots n, j = 1 \ldots p$, are fixed covariates. Assume that $\epsilon_i \sim N(0, \sigma^2)$, $U \sim N(\alpha, k\sigma^2)$, where α and $\sigma^2 > 0$ are unknown, $k > 0$ is known, and ϵ_i are independent of each other and of U. Assume further that the $(n \times p)$ matrix with entries $X_{ij} - \overline{X}_{.j}$ has rank p, where $\overline{X}_{.j} = \frac{1}{n}\Sigma_{i=1}^{n} X_{ij}$.

(a) Find the distribution of Y, and show that the variance-covariance matrix for Y is positive-definite. (Hint: For a constant c the inverse of a matrix $I + cJ$ is in the form of $I + dJ$ for certain constant d, where I is the $(n \times n)$ identity matrix and J is the $(n \times n)$ matrix with all entries equal to 1.)

(b) Show that $\beta_1, \ldots, \beta_p$ and α are estimable.

(c) Let $\theta = (\alpha, \beta_1, \ldots, \beta_p)^T$. Derive the maximum likelihood estimator for θ, denoted by $\hat{\theta} = (\hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_p)^T$. What is the distribution of $\hat{\theta}$?

(d) Let $\tilde{\theta} = (\tilde{\alpha}, \tilde{\beta}_1, \ldots, \tilde{\beta}_p)$ be the value of the vector θ minimizing the sum of squares

$$\sum_{i=1}^{n} [Y_i - (\alpha + X_{i1}\beta_1 + \ldots + X_{ip}\beta_p)]^2.$$

Is it true that $\mathrm{Var}(\hat{\alpha}) < \mathrm{Var}(\tilde{\alpha})$? Carefully justify your answer.

3. To evaluate the diagnostic performance using two continuous biomarkers, we randomly select n diseased subjects and m non-diseased subjects. Let $X_1 = (X_{11}, X_{12})', ..., X_n = (X_{n1}, X_{n2})'$ be these two measured biomarkers for the diseased subjects and $Y_1 = (Y_{11}, Y_{12})'$, ..., $Y_m = (Y_{m1}, Y_{m2})'$ be the same two measured biomarkers for the non-diseased subjects. We aim to find an optimal linear combination of these two biomarkers to maximize some measure of the diagnostic performance. In particular, we need to find $\beta = (\beta_1, \beta_2)'$ such that the area under the receiver operating characteristics curve, defined by $AUC(\beta) \equiv P(\beta'\mathbf{X}_1 \geq \beta'\mathbf{Y}_1)$, is maximized.

   Assume $X_1, ..., X_n$ are i.i.d from $MN(\mu_1, \Sigma)$ and $Y_1, ..., Y_m$ are i.i.d from $MN(\mu_2, \Sigma)$, where $\mu_1 = (\mu_{11}, \mu_{21})', \mu_2 = (\mu_{12}, \mu_{22})', \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}$ are unknown parameters and Σ is a positive definite matrix. Moreover, assume $m = \tau n$ for a fixed constant $\tau > 0$.

   (a) Show $AUC(\beta) = \Phi\left(\beta'(\mu_1 - \mu_2)/\sqrt{2\beta'\mathbf{\Sigma}\beta}\right)$, where $\Phi(x)$ is the cumulative distribution function of $N(0, 1)$.

   (b) Show that the maximum of $AUC(\beta)$, denote as $A^{optimal}$, is

   $$\Phi\left(\left[(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2)/2\right]^{1/2}\right).$$

   Hint: the β maximizing $AUC(\beta)$ is unique up to some multiplicative scale.

   (c) Calculate the maximum likelihood estimator for $A^{optimal}$ and denote it by $\widehat{A}$.

   (d) Describe how you will obtain the asymptotic distribution of $\sqrt{n}(\widehat{A} - A^{optimal})$. You do not need to give the explicit expression of the asymptotic variance.

   (e) To test whether the combination of the two biomarkers is useful for diagnosis, we formulate the hypothesis $H_0 : A^{optimal} = 1/2$ vs $H_1 : A^{optimal} > 1/2$ and reject H_0 when $\widehat{A} > c_n$ for some threshold value c_n (depending on n).

      i. Determine c_n such that the type I error converges to a given level α, where c_n is a constant depending *only* on n and α; that is, $\lim_{n\to\infty} P(\widehat{A} > c_n | H_0) = \alpha$.

      ii. Calculate the asymptotic power of this test at a local alternative $H_1 : A^{optimal} = 1/2 + \delta/\sqrt{n}$ where δ is a fixed positive constant.

2010 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