BASIC PHD WRITTEN EXAMINATION IN BIOSTATISTICS

THEORY, SECTION 2

(9:00 AM–1:00 PM, July 29, 2015)

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 exam code, **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. (25 points) Suppose that $Y \sim N(\mu, \Sigma)$ where Σ is symmetric and full rank. Let A be a symmetric matrix.

   (a) (6 points) Show that the quadratic form $Y^{\mathrm{T}}AY$ can be represented as

   $$Y^{\mathrm{T}}AY = \sum_{i=1}^{k} \lambda_i W_i$$

   where the W_i's are independently distributed as noncentral chi-squared variables with d_i degrees of freedom and noncentrality parameter δ_i, that is, $W_i \sim \chi^2_{d_i}(\delta_i), i = 1, 2, ..., k$. Indicate what λ_i, d_i, δ_i are equal to.

   (b) (6 points) Use part (a) to derive the moment generating function of $Y^{\mathrm{T}}AY$. Let $m(t)$ denote the moment generating function. Show that $m(t)$ exits in a small neighborhood of $t = 0$, say, $|t| < t_0$ for some positive constant t_0. Find the maximal value of t_0.

   (c) (6 points) Use part (a) to show that $tr[(A\Sigma)^2] = tr(A\Sigma) = r$, where r is the rank of A, then $Y^{\mathrm{T}}AY$ has a chi-squared distribution. distribution. Determine its degrees of freedom and noncentrality parameter.

   (d) (7 points) Show that $Y^{\mathrm{T}}AY$ has a noncentral chi-squared distribution if and only if $A\Sigma$ is idempotent.

2. (25 points) Consider the linear model $Y = X\beta + \epsilon$, where $\epsilon \sim N_n(0, \sigma^2 I)$, X is $n \times p$ of rank p, β is $p \times 1$, and (β, σ^2) are unknown. Define $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ and let h_{ii} denote the ith diagonal element of H. Further let $\hat{\sigma}^2$ denote the usual unbiased estimator of σ^2 for the linear regression model and let $\hat{\epsilon}_i$ denote the ordinary residual. Let $A_i = \frac{\hat{\epsilon}_i^2}{\sigma^2(1-h_{ii})}$ and $B_i = \frac{(n-p)\hat{\sigma}^2}{\sigma^2} - \frac{\hat{\epsilon}_i^2}{\sigma^2(1-h_{ii})}$.

   (a) (3 points) Show that $B_i \sim \chi^2_{n-p-1}$.

   (b) (5 points) Show that A_i and B_i are independent.

   (c) (5 points) Let $r_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{(1-h_{ii})}}$. Using parts (a) and (b), derive the *exact* distribution of $r_i^2/(n-p)$.

   (d) Suppose we suspect that the ith case is an outlier and we consider the mean shift outlier model $Y = X\beta + d_i\phi + \epsilon$, where $\epsilon \sim N_n(0, \sigma^2 I)$, ϕ is an unknown scalar parameter, and d_i is an $n \times 1$ vector with a 1 in the ith position and zeroes elsewhere.

      (i) (3 points) Derive the maximum likelihood estimate of ϕ.

      (ii) (3 points) Suppose we wish to test $H_0 : \phi = 0$. Derive the test statistic for this hypothesis and derive its exact distribution under H_0.

   (e) (6 points) Let $I = \{1, \ldots, m\}$ be the subset of the first m cases in the dataset. Let D_I denote Cook's distance based on simultaneously deleting m cases from the dataset, which is given by

$$D_I = \frac{(\hat{\beta} - \hat{\beta}_I)^{\mathrm{T}}(X^{\mathrm{T}}X)(\hat{\beta} - \hat{\beta}_I)}{p\hat{\sigma}^2},$$

   where $\hat{\beta}_I$ denotes the least squares estimate of β with the cases deleted from set I and $\hat{\beta}$ denotes the estimate of β based on the full data. Show that D_I can be written as

$$D_I = \frac{1}{p}\sum_{i=1}^{m} h_i^2 \left(\frac{\lambda_i}{1-\lambda_i}\right),$$

   where the λ_i, $i = 1, \ldots, m$, are the eigenvalues of the matrix $P_I = X_I(X^{\mathrm{T}}X)^{-1}X_I^{\mathrm{T}}$ based on a spectral decomposition of P_I, P_I is $m \times m$, X_I denotes the $m \times p$ matrix with rows corresponding to the m cases in the set I, $h_i^2 = \frac{(\gamma_i^{\mathrm{T}}\hat{\epsilon}_I)^2}{\hat{\sigma}^2(1-\lambda_i)}$, γ_i is the eigenvector corresponding to λ_i for $i = 1, \ldots, m$, and $\hat{\epsilon}_I = y_I - X_I\hat{\beta}$, where y_I is the $m \times 1$ response vector corresponding to the m cases in the set I.

3. (25 points) Consider a sequence of i.i.d. random vectors (X_i, Y_i, D_i) in which D_i is a Bernoulli variable with $D_i = 1$ for cases and $D_i = 0$ for controls; X_i is a $k \times 1$ vector and Y_i is another response variable. The case-control design samples individuals based on their disease status as follows. Let S_i be a Bernoulli variable where $S_i = 1$ denotes that subject i was sampled (selected into the study) while $S_i = 0$ denotes that subject i was not sampled (and will not participate in the study). The distribution of S_i is such that $\Pr(S_i = 1|D_i, X_i, Y_i) = \Pr(S_i = 1|D_i)$. Define $\tilde{p} = P(D_i = 1)$ and $\tilde{\pi} = P(D_i = 1|S_i = 1)$, and assume that both $\tilde{p}$ and $\tilde{\pi}$ are known. This problem focuses on inference on the mean of Y_i given X_i defined as $E[Y_i|X_i] = \mu(X_i)$ and assumed to have the form $\mu(x) = (1, x^{\mathrm{T}})\beta_0$, where β_0 is a $(k+1) \times 1$ vector of regression coefficients.

   For convenience, the sampled subjects will be renumbered from 1 to n, so the sample will be denoted $(X_1, Y_1, D_1), \cdots, (X_n, Y_n, D_n)$. Of course, $S_i = 1$ for each subject in the case-control sample. Inference will be based on this sample. To simplify notation, the subscript i will be dropped when it is not essential.

   (a) (7 points) Consider the conditional mean of Y given (X, D), denoted by $\tilde{\mu}(X, D) = E[Y|X, D]$. Show that

$$\mu(X) \;=\; \tilde{\mu}(X, 1)\Pr(D = 1|X) + \tilde{\mu}(X, 0)\Pr(D = 0|X) \tag{1}$$

   and

$$\tilde{\mu}(X, D) \;=\; E[Y|X, D, S = 1] = \mu(X) + \{D - \Pr(D = 1|X)\}\gamma(X), \tag{2}$$

   where $\gamma(X)$ describes the association between Y and D. Derive the explicit form of $\gamma(X)$ in terms of $\tilde{\mu}(X, 1)$ and $\tilde{\mu}(X, 0)$.

   (b) (6 points) Suppose that the disease is rare in the population so that $p(X) = \Pr(D = 1|X) \approx 0$ and, further, that $\gamma(X) = (1, X^T)\gamma_0$. Show that $\tilde{\mu}(X, D)$ can be approximated by model that is linear in the parameters $\xi_0 = (\beta_0, \gamma_0)$. Calculate the least-square estimate of ξ_0, denoted as $\hat{\xi}_0$. Derive the covariance matrix of $\hat{\xi}_0$ and suggest a reasonable estimator of it. (The covariance here is conditional on both the X_i's and the D_i's). Is $\hat{\xi}_0$ an efficient estimator? Explain.

   (c) (6 points) Let $\pi(X) = \Pr(D = 1|X, S = 1)$ be the risk of disease at X in the case-control sample. Recall that the selection indicator S is conditionally independent of (Y, X) given D. Show that logit $p(X) =$ logit $\pi(X) + \log \frac{\tilde{p}(1-\tilde{\pi})}{\tilde{\pi}(1-\tilde{p})}$.

Suppose that $\pi(X)$ follows the model logit $\pi(X) = \eta_0 + X^T\psi_0$. Based on the observed data $\{(x_i, d_i)\}_{i=1}^n$, we can directly calculate the maximum likelihood estimate of (η_0, ψ_0), denoted by $(\hat{\eta}_0, \hat{\psi}_0)$. Write down the score function for (η_0, ψ_0) and the covariance matrix of $(\hat{\eta}_0, \hat{\psi}_0)$. (Again, the covariance here is conditional on both the X_i's and the D_i's).

(d) (6 points) Based on the results in (c), we can plug the estimated $p(X)$, denoted as $\hat{p}(X)$, into (1) and (2) . Therefore, $\tilde{\mu}(X, D)$ can be also approximated by a model that is linear in $\xi_0 = (\beta_0, \gamma_0)$. Calculate the least-square estimate of ξ_0, denoted as $\hat{\xi}_0$, and derive the covariance matrix of $\hat{\xi}_0$ (conditional on both the X_i's and the D_i's).

2015 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