BASIC PHD WRITTEN EXAMINATION

THEORY, SECTION 2

(9:00 AM–1:00 PM, July 28, 2017)

INSTRUCTIONS:

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

(b) The time limit for this examination is four hours.

(c) Answer both 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 is 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) Consider the linear model

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

where $E(\epsilon) = 0$ and $\text{Cov}(\epsilon) = V$, V is assumed known and positive definite, and (β, γ) are unknown. Further, let $A = X(X'V^{-1}X)^{-}X'V^{-1}$, where $-$ denotes generalized inverse, X is $n \times p$, Z is $n \times q$, and both X and Z may be less than full rank. Let $C(H)$ denote the usual label for the column space of an arbitrary matrix H.

(a) (2 points) Show that $(I - A)'V^{-1}(I - A) = (I - A)'V^{-1} = V^{-1}(I - A)$.

(b) (3 points) Show that A is the projection operator onto $C(X)$ along $C(V^{-1}X)^{\perp}$.

(c) (4 points) Let B denote the projection operator onto $C(X, Z)$ along $C(V^{-1}(X, Z))^{\perp}$. Assume that all matrix inverses exist. Show that

$$B = A + (I - A)Z\left[Z'(I - A)'V^{-1}(I - A)Z\right]^{-1} Z'(I - A)'V^{-1}.$$

(d) (5 points) Show that $(\hat{\gamma}, \hat{\beta})$ are generalized BLUE's for the linear model, where $(\hat{\gamma}, \hat{\beta})$ satisfy

$$\hat{\gamma} = \left[Z'(I - A)'V^{-1}(I - A)Z\right]^{-1} Z'(I - A)'V^{-1}(I - A)Y,$$

and

$$X\hat{\beta} = A(Y - Z\hat{\gamma}).$$

(e) (5 points) Suppose that $\epsilon \sim N_n(0, V)$ and V is known. Further, suppose that (β, γ) are both estimable. From first principles, derive the likelihood ratio test for the hypothesis $H_0 : \gamma = 0$, where (β, γ) are both unknown, and state the exact distribution of the test statistic under the null and alternative hypotheses.

(f) (6 points) Suppose that $\epsilon \sim N_n(0, \sigma^2 R)$, where R is known and positive definite, and $(\beta, \gamma, \sigma^2)$ are all unknown. Further, assume that (β, γ) are both estimable. Derive an exact joint 95% confidence region for $(\beta, \gamma, \sigma^2)$.

2. (25 points) Suppose that the pair (X, Y) is distributed such that $X \sim \text{normal}(0, 1)$ and $Y \sim \text{Bernoulli}(\theta), 0 < \theta < 1$. For example, X could be the log of the level of a certain biomarker and Y an indicator of some disease. Assume that the value of θ is known and given (e.g. we know that the disease prevalence is 0.001).

   Here we try to answer the following question: Given the above specifications, what is the largest possible correlation between X and Y?

   Notation: Define $\rho = \text{corr}(X, Y)$. Use $\phi(\cdot)$ and $\Phi(\cdot)$ to denote the standard normal pdf and cdf, respectively. Define any new notation you use.

   (a) (1 point) Is $\rho = 1$ possible? Explain.

   (b) (6 points) Find the joint distribution for (X,Y) that maximizes corr(X,Y) subject to the model stated above. Show that it (that distribution) has the property that $E[X|Y = 1] = \theta^{-1}\phi(\Phi^{-1}(1 - \theta))$.

   (c) (6 points) Obtain an explicit expression for $\rho^* = \text{corr}(X, Y)$ within the joint distribution found in the previous part. Compute the numerical value of ρ^* for the case $\theta = 0.001$.

   (d) (6 points) Now suppose we are interested in various diseases with different prevalences (θ) ranging in $(0, 1)$. Find the value of θ that leads to the largest possible value of ρ^*, and compute that largest value, to be denoted ρ^{**} (compute its numerical value).

   (e) (3 points) Note: This part is totally independent of the previous parts, even though the models have some similarity. You can reuse results obtained above if needed.

   Suppose that the iid pairs $(X_i, Y_i), i = 1, \ldots, n$, are distributed such that $X_i \sim \text{normal}(0, 1)$, $Y_i \sim \text{Bernoulli}(\theta), 0 < \theta < 1$, and $\text{corr}(X_i, Y_i) = \rho$, where both θ and ρ are unknown parameters. Develop an estimating equation for (ρ, θ) based on the vectors $Z_i := (T_i, Y_i)^\top, i = 1, \ldots, n$, where $T_i := X_i Y_i$. Obtain the estimates $(\hat{\rho}, \hat{\theta})$ in explicit form.

   (f) (3 points) Based on $\hat{\rho}$ from the previous part, develop a large-sample (as $n \to \infty$) 95% confidence interval for ρ. The interval should not depend on any unknown parameters. Describe and justify your procedure clearly.

   Note: $\Phi(-3.09) \approx 0.001$, $\Phi(-2.33) \approx 0.01$, $\Phi(-1.96) \approx 0.025$, $\Phi(-1.64) \approx 0.05$, $\Phi(-1.28) \approx 0.1$

2017 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