BASIC DOCTORAL WRITTEN EXAMINATION IN BIOSTATISTICS

DOCTORAL APPLICATIONS EXAM

(9:00 AM Friday, August 2 to 9:00 AM Wednesday, August 7, 2019)

INSTRUCTIONS:

a) This is an open book, take home examination. You may not communicate with anyone except Michael Hudgens (mhudgens@bios.unc.edu) and Fei Zou (feizou@email.unc.edu) about the content of this examination. Professors Hudgens and Zou will only answer questions for clarification purposes if it is deemed necessary.

b) Return the examination with a signed statement of the UNC Honor Pledge on a page separate from your answers. The pledge is attached at the end of the exam handout.

c) The time limit for this examination is five days. The time limit is strictly enforced and without exceptions, except by prior agreement. Any material turned in later than 9:00 am on the due date will be assigned a grade of 0.

d) Answer all four (4) of the questions that follow.

e) Put your answers to different questions on separate sets of paper and staple them separately by question. Please turn in two (2) copies of answers to each question.

f) Put your code letter, not your name, on each page. 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.

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. Be clear, precise and concise in presenting results and findings. Use only standard statistical language. Do not provide any computer code or output with your solution, unless otherwise directed. Pay attention to using precise notation and to providing clear interpretations.

h) Most questions should be answered in the equivalent of less than 5 typewritten pages (300 words per page with font no smaller than 12pt), and under no circumstances will more than the first 8 typewritten pages or the equivalent (including tables, figures, appendices, etc.) for each question be read by the graders.

i) All the answers should be turned in on paper (i.e., not electronically). Return the examination along with the signed UNC Honor Pledge to Melissa Hobgood in the Bios Conference Room by 9am on Wednesday, August 7th.

h) The computer files/data to which this examination refers can be obtained from the Department's website:

`https://www.bios.unc.edu/distrib/exam/DoctoralApplication%202009-present/2019aug/`

using your UNC onyen login information. Access to this site from off campus requires a VPN connection.

1. In many imaging genetics studies, out-of-sample prediction is of great interest. For example, we may use genetic variants to predict imaging biomarkers, which are associated with brain-related diseases. Different biomarkers may have different genetic architecture, and the number of causal variants may vary among traits.

   Imaging genetics studies are often expensive, thus in practice, the sample size n is often smaller than the number of genetic variants p. In this question, we will study 4966 genetic variants and two phenotypes (biomarkers) that are quantified from the brain via magnetic resonance imaging. For similarity, we have removed the effects of other covariates (such as age, sex) from the phenotypes and all variables are normalized.

   We have 500 subjects in the training data $(\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{X})$, where $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are the two phenotypes, and $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_p]$ are the 4966 genetic variants. Similarly, we have 100 subjects in the testing data $(\boldsymbol{y}_{1z}, \boldsymbol{y}_{2z}, \boldsymbol{Z})$, where $\boldsymbol{Z} = [\boldsymbol{z}_1, \cdots, \boldsymbol{z}_p]$. For both data sets, we have the following 4968 variables

Column	Variable
1	the phenotype generated from the first region of interest (ROI1)
2	the phenotype generated from the second region of interest (ROI2)
3-4968	4966 genetic variants, with names displayed in the datasets

   The training data set is available in the file `qual_2019_train_data_0.csv`, and the testing data set is contained in `qual_2019_test_data_0.csv`.

   A) Marginal screening is a common tool to assess associations between a large number of genetic variants and the primary variable. That is, for each genetic variant in $\boldsymbol{X}$, we build a linear regression model and test and estimate its association with the primary variable. Please run marginal screening for $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ separately in the training dataset. Display and compare the two sets of p-values (-log10(p-value)) and coefficient estimates. Comment on your findings. Are there significant genetic variants for the two variables? Suppose we know that one phenotype has dense genetic signals and another one has sparse signals; which phenotype do you think has sparse signals?

   B) Now we check the in-sample goodness of fit of the marginal screening procedure. Let $\widehat{\beta}_i$ and $\widehat{\eta}_i$ be the estimated effect size of the ith genetic variant for $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, respectively, $i = 1, \cdots, p$. Define $\widehat{\boldsymbol{y}}_1 = \sum_{i=1}^{p} \boldsymbol{x}_i \widehat{\beta}_i$ and $\widehat{\boldsymbol{y}}_2 = \sum_{i=1}^{p} \boldsymbol{x}_i \widehat{\eta}_i$. Report the variance in $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ that can be explained by $\widehat{\boldsymbol{y}}_1$ and $\widehat{\boldsymbol{y}}_2$, respectively. That is, calculate the in-sample R-squared (R^2) when fitting $\boldsymbol{y}_1 \sim \widehat{\boldsymbol{y}}_1$ and $\boldsymbol{y}_2 \sim \widehat{\boldsymbol{y}}_2$. Comment on your findings.

   C) Rank the genetic variants according to the strength of their marginal association with the phenotype (e.g., by p-value or absolute value of coefficient estimate). Instead of using all p genetic variants, now only select the *top* variants to construct $\widehat{\boldsymbol{y}}_1$ and $\widehat{\boldsymbol{y}}_2$. Try a series of thresholds and display the corresponding in-sample R^2 values. Comment on the pattern.

   D) Use the coefficients estimated from the training dataset to predict the the phenotypes in the testing data. Use all p phenotypes first and construct the predictor as $\widehat{\boldsymbol{y}}_{1z} =$

$\sum_{i=1}^{p} \boldsymbol{z}_i \widehat{\beta}_i$ and $\widehat{\boldsymbol{y}}_{2z} = \sum_{i=1}^{p} \boldsymbol{z}_i \widehat{\eta}_i$. Report the out-of-sample R^2. Then only select the *top* variants for prediction. Try a series of cutoffs and display the corresponding out-of-sample R^2 values. Comment on the difference between the in-sample and out-of-sample R^2 values. For out-of-sample R^2, is the pattern of $\boldsymbol{y}_1$ similar to that of $\boldsymbol{y}_2$? If not, try to give a reasonable explanation about the difference.

E) Now put all the p variables in one regression model and add penalty to the coefficients. For example, the ridge estimator for $\boldsymbol{y}_1$ is $\widehat{\boldsymbol{\beta}}_R = (\boldsymbol{X}^T\boldsymbol{X} + \lambda n \boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{y}_1$, and for $\boldsymbol{y}_2$ is $\widehat{\boldsymbol{\eta}}_R = (\boldsymbol{X}^T\boldsymbol{X} + \lambda n \boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{y}_2$. Redo parts C and D with ridge estimators $\widehat{\boldsymbol{\beta}}_R$ and $\widehat{\boldsymbol{\eta}}_R$. Compare the results of ridge estimators to those of marginal estimators. Comment on your findings.

F) Try one or two *other* methods for out-of-sample prediction of the two phenotypes, such as XGBoost, LASSO, support vector machines, tree-based methods, principal component regression, etc. Report the best performance you can obtain from these methods. Does the new method have better performance than marginal screening? Do you find any difference when predicting $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$?

G) Write a half-page discussion of all your findings and thoughts when solving this question. You may use simulations to support your arguments if needed.

Point distribution: A 3, B 3, C 3, D 3, E 4, F 5, G 4

2. As the biostatistician for a research team submitting a grant proposal, you are responsible for determining an appropriate sample size to support a research project's ability to address relevant research questions with adequate power and to guide the budget. The research team is proposing a randomized trial to compare a new drug therapy versus usual care with regard to lowering cholesterol in hopes of preventing cardiovascular adverse events. Patients with elevated total cholesterol levels ($\geq$ 200 mg/dL) will be enrolled into the trial under a 1:1 randomization scheme, and they will be followed every two weeks for two months following their randomization.

   The trial's primary objectives are:

   I. Test the efficacy of the new drug therapy with respect to lowering the mean level of total cholesterol (relative to usual care) over an 8 week follow-up period, and

   II. Test the efficacy of the new drug therapy with respect to decreasing the proportion of patients with elevated ($\geq$ 200mg/dL) total cholesterol levels (relative to usual care) over an 8 week follow-up period.

   Consider the following mixed effects model to address trial objective I:

$$Y_{ij} = \beta_0 + \beta_1 X_i + \beta_2 t_j + \beta_3 X_i t_j + b_i + \epsilon_{ij}$$

   where Y_{ij} is the total cholesterol level for patient i at week t_j, and $X_i = 1$ if patient i is randomized to the new drug and 0 otherwise. Assume b_i iid $N(0, \sigma_b^2)$, ϵ_{ij} iid $N(0, \sigma_w^2)$, and $b_i \perp \epsilon_{ij}$. Likewise, consider the following marginal model to address objective II:

$$\text{logit}\{\Pr(Y_{ij} \geq 200)\} = \gamma_0 + \gamma_1 X_i + \gamma_2 t_j + \gamma_3 X_i t_j$$

   The research team would like to test the following null hypotheses, respectively: (i) the rate of change in total cholesterol level over time is the same in both trial arms, i.e., $H_0 : \beta_3 = 0$; and (ii) the rate of change in the log odds of elevated total cholesterol level over time is the same in both trial arms, i.e., $H_0 : \gamma_3 = 0$.

   Prior to this proposal, the research team conducted a pilot study in this same patient population with the same timing of follow-up visits as the proposed trial. Data from this pilot study may be used to obtain estimates useful for the sample size determination of the current project's planned randomized trial. The pilot study data can be found in the file `pilot.csv`, which contains the following variables:

   - ID: unique patient identification code
   - GROUP: (E = experimental therapy, C = usual care control)
   - WEEKS: weeks of follow-up since randomization (0, 2, 4, 6, 8)
   - TOTCHOL: Total cholesterol level (mg/dL) at the corresponding time of follow-up

   A) Conduct a preliminary analysis of the pilot data, including descriptive statistics and estimates useful for determining sample size for the randomized trial.

   B) Determine the sample size needed for the planned trial to test $H_0 : \beta_3 = 0$ versus an alternative of $H_1 : \beta_3 = -3$ with 0.85 power, assuming a one-sided 0.025 type I error rate. Provide justification of the sample size determination to be included in the grant proposal. Do not include software code in your solution, but do include enough detail that a grant reviewer could replicate your results.

C) Determine the sample size needed for the planned trial to test $H_0 : \gamma_3 = 0$ versus an alternative of $H_1 : \gamma_3 = -0.15$ with 0.85 power, assuming a one-sided 0.025 type I error rate. As in part B, provide justification of the sample size determination with sufficient detail that a reviewer could replicate the results.

D) Provide a brief (2 – 3 paragraphs) statistical analysis plan for the randomized trial. The study investigators do not anticipate patients will miss many visits and hence no adjustment for missing data is necessary in the sample size calculations above. Nonetheless, some patients may drop out of the study or miss some study visits. Include in the analysis plan a description of how such missing data will be handled. Describe the assumed missing data mechanism for the proposed approach and briefly discuss the plausibility of this assumption in the context of the trial.

Point distribution: A 3, B 7, C 7, D 8

3. A study was conducted to evaluate a new anti-psychotic drug for the treatment of schizophrenia. A random sample of $K = 400$ patients was obtained from a large medical center, and patients were randomly assigned in 1:1 ratio to receive either placebo or the drug. The study design called for patients to be evaluated at weeks 0, 1, 3 and 6. Week 0 was the baseline observation, before randomization. This analysis is restricted to Item 79 of the Inpatient Multidimensional Psychiatric Scale (Lorr and Klett (1966), *Inpatient Multidimensional Psychiatric Scale Manual*). The scale ranges from normal (0) to extremely ill (7). For the purpose of this analysis, the score is dichotomized into normal to mildly ill, score ≤ 3.5, coded as Illness=0, and moderately ill or severely ill, score > 3.5, coded as Illness=1.

   The data are available in files `illness1.dat` and `illness2.dat`. The two files contain the same data, but the first is formatted with four records per subject, while the second has one record per subject. The variable Week takes values 0, 1, 3, 6.

   Let the random variable Y_{ij} represent the illness outcome, Illness, of the ith subject at the jth time, $i = 1, \cdots, 400, j = 1, 2, 3, 4$. Define the vector Y_i to be $Y_i := (Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4})^\top$. Subjects are assumed mutually independent throughout. Further, let t_{ij} be the square root of week; $t_{i1} = 0, t_{i2} = 1, t_{i3} = \sqrt{3}, t_{i4} = \sqrt{6}$ for all i. For the ith patient, treatment is coded as $x_i = 0$ if the patient received placebo and $x_i = 1$ if the patient received the drug.

   In what follows, for hypothesis testing, use likelihood ratio tests when possible/feasible. Otherwise, use Wald-type tests. In either case, you should describe your methods, present test statistics, degrees of freedom and p-values. Use common statistical language with no reference to any computer code, statements or options. Do not assume that the reader knows anything about the software you used, and do not submit any computer code. Interpretations should be as clear and as practical as possible, and should be targeted at non-statisticians.

   To allow for uniform grading, if quadrature is needed, use 25 nodes (if your software allows it).

   (a) Present some relevant summaries, tabular and/or graphical, and provide some brief comments.

   (b) Without explicitly using regression methods, carry out an asymptotic 2-sample test of the null hypothesis of no effect of the experimental drug compared to the placebo. Do not assume a common covariance matrix in the two groups.

   (c) Fit the model M_1:

$$\operatorname{logit} \mathrm{E}[Y_{ij}|b_{1i}] = \beta_1 + b_{1i} + \beta_2 t_{ij} + \beta_3 t_{ij} x_i,$$

   $i = 1, \ldots, K, j = 1, \ldots, 4$, where $b_{1i}, \ldots, b_{1K}$ are unobserved random variables distributed as iid normal with mean 0 and variance σ^2. Further, assume that $Y_{i1}, \ldots, Y_{i4}$ are conditionally independent given b_{1i}. Present parameter and empirical (robust, sandwich) standard error estimates and 95% confidence intervals.

   (d) The linear predictor in model M_1 contains an interaction between time (t_{ij}) and treatment (x_i) but not a main effect of treatment. Explain why this is or is not justified.

(e) In the context of M_1, interpret the *estimates* of σ^2, β_2 and β_3.

(f) Fit the model M_2:

$$\operatorname{logit} \mathrm{E}[Y_{ij}|b_{1i}, b_{2i}] = \beta_1 + b_{1i} + \beta_2 t_{ij} + \beta_3 t_{ij} x_i + b_{2i} t_{ij},$$

$i = 1, \ldots, K, j = 1, \ldots, 4$, where $b_i = (b_{1i}, b_{2i})^\top, i = 1, \ldots, K$ are unobserved random vectors distributed as iid bivariate normal with mean vector 0 and covariance matrix G. Further, assume that $Y_{i1}, \ldots, Y_{i4}$ are conditionally independent given b_i. Present parameter and empirical (robust, sandwich) standard error estimates and 95% confidence intervals.

(g) One of the investigators suggested restricting G in model M_2 to be diagonal. Discuss whether this is or is not a good idea.

(h) In the context of M_2, test the null hypothesis $H_0 : g_{22} = 0$ against $H_1 : g_{22} \neq 0$, where g_{22} is the $(2, 2)$ element of G.

(i) Fit the model M_3:

$$\operatorname{logit} \mathrm{E}[Y_{ij}] = \gamma_1 + \gamma_2 t_{ij} + \gamma_3 t_{ij} x_i,$$

$i = 1, \ldots, K, j = 1, \ldots, 4$. Model the correlation among Y_{ij}'s within a subject via the pairwise log odds ratios, say $\alpha_1, \cdots, \alpha_6$. Assume that all six pairwise odds ratios are distinct but the same in both treatment groups. Present parameter and empirical standard error estimates and 95% confidence intervals. Further, test the null hypothesis of no effect of the experimental drug compared to the placebo.

(j) In the context of M_3, test the null hypothesis $H_0 : \alpha_1 = \cdots = \alpha_6$ against its complement.

(k) Write one or two paragraphs for a medical journal discussing the study design and your main findings.

Points: a 1, b 2, c 2, d 1, e 6, f 2, g 1, h 2, i 2, j 3, k 3

4. Unlike breast and colorectal cancers, gene expression-based subtyping for classifying pancreatic cancer tumors is in its infancy. In breast cancer, such subtyping methods have facilitated the application of precision medicine approaches in making treatment decisions, optimizing the selection of existing therapies to improve patient outcomes. For new breast cancer patients entering the clinic, a patient's determined subtype is used to recommend an optimal choice of therapy. Such recommendations are typically based upon prior clinical trial data, where treatments with the best response within a particular subtype category are typically recommended to new patients of the same subtype.

   Researchers are currently working towards developing a clinically relevant subtyping approach for pancreatic cancer. An ideal subtyping approach should consist of subtypes that differentiate treatment response in patients. Several tumor subtyping approaches for pancreatic cancer have been recently proposed. Subtyping approach A was initially proposed in 2017, consisting of three pancreatic tumor subtypes (A1, A2, and A3). A subsequent study of pancreatic cancer patients, based on a more diverse set of pancreatic cancer samples, proposed four tumor subtypes (subtyping approach B): B1, B2, B3, B4. A UNC lab proposed subtyping approach C, consisting of two subtypes C1 and C2. In their original publications, each proposed approach used different cohorts of patients to demonstrate clinical relevance. As a result, the generalizability, robustness, and relative clinical utility of each proposed subtyping schema remains unclear. This has presented a barrier to wider clinical adoption of such subtyping approaches in pancreatic cancer.

   To address this issue, a large clinical trial (n = 200 patients) was recently completed comparing two common first-line therapies in pancreatic cancer (Folfirinox and Gemcitabine + Abraxane). Patients were randomized to arms pertaining to each therapy. Prior to treatment, gene expression data was obtained from each patient's tumor via biopsy. Treatment response for each patient was assessed on an ordinal scale based on standardized criteria 8 weeks after treatment: Progressive Disease (PD, worst), Stable Disease (SD), Partial Response (PR), and Complete Response (CR, best).

   Based on this trial, we would like to make a recommendation to physician-scientists regarding the "best" subtyping approach in terms of discriminating responses to treatment. In `subtypeResponse.txt`, we have the following data obtained from this clinical trial:

   - Response category (responseCat): 1 Progressive Disease (PD), 2 Stable Disease (SD), 3 Partial Response (PR), 4 Complete Response (CR)
   - Arm: 1 Folfirinox, 2 Gemcitabine + Abraxane
   - Subtyping approach A (subA): 1 A1, 2 A2, 3 A3
   - Subtyping approach B (subB): 1 B1, 2 B2, 3 B3, 4 B4
   - Subtyping approach C (subC): 1 C1, 2 C2

   (a) Summarize the results of the trial as follows:

      i. Create a single, concise, and easy-to-read table that summarizes the proportion of subjects in each response category by subtype (pooling across arms). Repeat this for each subtyping approach. Summarize the results for researchers.

      ii. Create a single figure illustrating the proportion of participants with partial or complete response (PR+CR) by subtype for each subtyping approach (pooling

across arms). Label axes clearly. Summarize the results.

iii. Create a single figure with three panels that depicts overlap in the subtypes between approach A versus B, A versus C, and B versus C (pooling across arms). Summarize the result.

iv. Based on i-iii above, what can you say to the physician-scientists who advocate the use of the more complicated subtyping approach B (four subtype levels) instead of the simpler subtyping approach C (two subtype levels)? Justify your answer based on the figures and tables you created.

(b) One of the physician-scientists is interested in the relationship between response category and subtype within each subtyping approach.

i. Assuming response category is an ordinal outcome, fit a proportional odds model using subtype as the only predictor. Repeat this process for each subtyping approach. That is, fit separate models for subtyping approaches A, B, and C. For each fitted model, check the model assumptions using appropriate diagnostics.

ii. Perform a hypothesis test evaluating the null hypothesis that subtype is not associated with response category, one for each subtyping approach. Use the fitted models from (b)i for each hypothesis test. Report the test, the test statistic, its degrees of freedom, and p-value. Interpret the result for each test in plain language for the physician-scientists to understand.

(c) One of the goals for this trial is to determine which subtyping approach best explains clinical outcomes in patients. One of the investigators who advocates for subtyping approach C states that the approach with the smallest p-value from the tests performed in (b) is the one that best explains clinical outcomes in patients compared to the other models. Explain why this statement is (or is not) correct.

(d) Describe a statistical means to directly compare the ability of each subtyping approach to explain clinical outcomes in patients given the fitted models in (b). Implement this approach. Based on these results, make a recommendation to researchers regarding which approach best explains clinical outcome in patients.

(e) One of the investigators indicated that there is no prior evidence to believe that the subtype-specific response rates would be similar across arms in the study, suggesting that we cannot simply ignore arm or pool across arms when evaluating the relationship between subtype and response category in the study. Repeat your analysis in (b) and (d) addressing this question for the subtyping approach C.

i. Based on this fitted model, evaluate whether there is a difference in subtype specific response rates between arms given the model that you fit. Report the test statistic, its degrees of freedom and the p-value for this test.

ii. Based on these results, which drug would you recommend for each subtype? Why?

iii. Fit a similar model for subtyping approaches A and B. Does your conclusion about the best subtyping approach change after accounting for arm?

Point distribution: a 6, b 5, c 4, d 4, e 6

2019 DOCTORAL APPLICATIONS EXAM

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."

Print Name	Signature