BASIC DOCTORAL WRITTEN EXAMINATION IN BIOSTATISTICS

DOCTORAL APPLICATIONS EXAM

(9:00 AM Tuesday, August 4 to 9:00 AM Sunday, August 9, 2020)

INSTRUCTIONS:

- This is an open book, take home examination. You may not communicate with anyone except Michael Hudgens (mhudgens@bios.unc.edu) about the content of this examination. Professor Hudgens will only answer questions for clarification purposes if it is deemed necessary.
- 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.
- Answer all four (4) of the questions that follow. For each question, 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.
- 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.
- Return your solution to the examination to Melissa Hobgood via email (mhobgood@bios.unc.edu) by 9am on Sunday, August 9th. Solutions should be submitted in a single pdf document. Use the file naming convention AppliedDocCodeXX.pdf with XX replaced by your exam code.
- Do not put your name anywhere on the exam. Keep your exam 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.
- When submitting your solution electronically to Melissa Hobgood via email, include the following statement in the body of the email: "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."
- 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/2020aug/`

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

1. Genetic correlation quantifies the shared genetic influences between two heritable traits. In this question, we evaluate the genetic correlation between cognition and the structure of two brain regions (hippocampus and thalamus). The data on cognition and brain regions are from two independent studies. For simplicity, we have removed the effects of other covariates (such as age, sex) from the traits and all variables are normalized.

   We have $n = 1,000$ subjects with brain regions data $(\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{X})$, where $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are measurements of the hippocampus and thalamus, respectively, and $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_p]$ are the $p = 901$ genetic variants. Similarly, we have $1,000$ subjects in cognition data $(\boldsymbol{y}_3, \boldsymbol{Z})$, where $\boldsymbol{Z} = [\boldsymbol{z}_1, \cdots, \boldsymbol{z}_p]$ are the same 901 genetic variants as in $(\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{X})$. Data can be found in the files `qual_2020_q1_data1.csv` and `qual_2020_q1_data2.csv`. We will evaluate the genetic similarity attributable to these genetic variants. The individuals in the two studies are independent. Suppose there are linear relationships between each trait and genetic variants

$$\boldsymbol{y}_1 = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{\epsilon}_\alpha, \quad \boldsymbol{y}_2 = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}_\beta, \quad \text{and} \quad \boldsymbol{y}_3 = \boldsymbol{Z}\boldsymbol{\eta} + \boldsymbol{\epsilon}_\eta, \tag{1}$$

   where $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\eta}$ are genetic effects, and $\boldsymbol{\epsilon}_\alpha$, $\boldsymbol{\epsilon}_\beta$, and $\boldsymbol{\epsilon}_\eta$ are random errors. The genetic correlations are defined by inner products of genetic effects

$$\varphi_{13} = \frac{\boldsymbol{\alpha}^T\boldsymbol{\eta}}{\|\boldsymbol{\alpha}\| \cdot \|\boldsymbol{\eta}\|} \quad \text{and} \quad \varphi_{23} = \frac{\boldsymbol{\beta}^T\boldsymbol{\eta}}{\|\boldsymbol{\beta}\| \cdot \|\boldsymbol{\eta}\|}. \tag{2}$$

   (a) For each of the 901 genetic variants, estimate its marginal genetic effects on hippocampus (y_1), thalamus (y_2), and cognition (y_3), respectively. Define them as $\widehat{\boldsymbol{\alpha}} = n^{-1}\boldsymbol{X}^T\boldsymbol{y}_1$, $\widehat{\boldsymbol{\beta}} = n^{-1}\boldsymbol{X}^T\boldsymbol{y}_2$, and $\widehat{\boldsymbol{\eta}} = n^{-1}\boldsymbol{Z}^T\boldsymbol{y}_3$. It is well known that linkage disequilibrium exists among genetic variants. Check the pair-wise correlation of the 901 genetic variants in $\boldsymbol{X}$ and $\boldsymbol{Z}$, What kind of pattern do you observe? Do you think the marginal estimators $\widehat{\boldsymbol{\alpha}}$, $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\eta}}$ are unbiased? Comment on your findings.

   (b) Use marginal estimators $\widehat{\boldsymbol{\alpha}}$, $\widehat{\boldsymbol{\beta}}$, and $\widehat{\boldsymbol{\eta}}$ to estimate φ_{13} and φ_{23} in Equation (2). Use re-sampling techniques (e.g., bootstrapping, permutation) to estimate the standard errors of your estimators and test whether φ_{13} and φ_{23} are zeros. Interpret your results.

   (c) Condition on thalamus, is there significant genetic correlation between hippocampus and cognition? Try to use the marginal estimators $\widehat{\boldsymbol{\alpha}}$, $\widehat{\boldsymbol{\beta}}$, and $\widehat{\boldsymbol{\eta}}$ to answer this question.

   (d) Put all 901 genetic variants in one linear regression model and calculate the OLS estimators of the genetic effects on y_1, y_2, and y_3, respectively. For example, the OLS estimator of α is $\widehat{\boldsymbol{\alpha}}_O = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}_1$, and similarly we can define $\widehat{\boldsymbol{\beta}}_O$, and $\widehat{\boldsymbol{\eta}}_O$. Are the OLS estimators similar to their corresponding marginal estimators in (a)? Redo part (b) with OLS estimators $\widehat{\boldsymbol{\alpha}}_O$, $\widehat{\boldsymbol{\beta}}_O$, and $\widehat{\boldsymbol{\eta}}_O$. Compare the new estimates of φ_{13} and φ_{23} to those using marginal estimators in part (b). Comment on your findings.

   (e) Now calculate the ridge estimators $\widehat{\boldsymbol{\alpha}}_R = (\boldsymbol{X}^T\boldsymbol{X} + n\lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{y}_1$, $\widehat{\boldsymbol{\beta}}_R = (\boldsymbol{X}^T\boldsymbol{X} + n\lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{y}_2$, and $\widehat{\boldsymbol{\eta}}_R = (\boldsymbol{Z}^T\boldsymbol{Z} + n\lambda\boldsymbol{I}_p)^{-1}\boldsymbol{Z}^T\boldsymbol{y}_3$. Use ridge estimators $\widehat{\boldsymbol{\alpha}}_R$, $\widehat{\boldsymbol{\beta}}_R$, and $\widehat{\boldsymbol{\eta}}_R$ to estimate φ_{13} and φ_{23}. Try a wide range of λ values; are these estimates of φ_{13}

and φ_{23} similar to the estimates in (b) or (d)? Comment on the relationships among the three sets of results.

(f) Use the model in Equation (1), simulate one trait using data $\boldsymbol{X}$. For example, $\boldsymbol{y}_{s1} = \boldsymbol{X}\boldsymbol{\alpha}_s + \boldsymbol{\epsilon}_{\alpha s}$. Then compare the performance of marginal and OLS estimators when estimating the genetic effects $\boldsymbol{\alpha}_s$. Which estimator has better performance in your simulation? Based on your simulation results, discuss the potential issues/limitations of marginal and OLS estimators when analyzing this dataset.

(g) Use the model in Equation (1), simulate another trait using data $\boldsymbol{Z}$. For example, $\boldsymbol{y}_{s3} = \boldsymbol{Z}\boldsymbol{\eta}_s + \boldsymbol{\epsilon}_{\eta s}$. Similar to φ_{13} in Equation (2), the genetic correlation between $\boldsymbol{y}_{s1}$ and $\boldsymbol{y}_{s3}$ is defined by the inner product of genetic effects $\boldsymbol{\alpha}_s$ and $\boldsymbol{\eta}_s$, denoted as φ_{s13}. Obtain the marginal estimators of $\boldsymbol{\alpha}_s$ and $\boldsymbol{\eta}_s$, and estimate φ_{s13} as you previously did in (b). Is this estimator a biased estimator of φ_{s13}? If biased, discuss some factors that may cause the bias in this data analysis. You may use additional simulations to support your arguments if needed.

Point distribution: a 3, b 3, c 3, d 4, e 4, f 4, g 4

2. A study was conducted to understand physicians' preferred therapeutic strategies for different scenarios involving patients with certain medical conditions. Two hundred twenty physicians were enrolled in the study, and they selected from among three therapeutic strategies for various scenarios. Data from this study can be found in the file `logit.csv`, which contains the following variables:

   - ID: unique physician identification code (1-220)
   - TherStrat: therapeutic strategy (0 treatment A, 1 treatment B, 2 treatment C)
   - Status: patient status (1 stable, 2 unstable)
   - Condition: medical condition (1 cardiovascular disease, 2 liver disease, 3 kidney disease)
   - Cost: monthly cost of treatment (1 less than \$1k, 2 \$1k-3k, 3 greater than \$3k)

   Your task is to fit an appropriate regression model with the preferred therapeutic strategy as the response variable, and patient status, medical condition, and monthly cost as explanatory variables.

   (a) Provide descriptive statistics for each study variable.

   (b) Consider a generalized logit model of the preferred therapeutic strategy using main effects (but no interactions) for the three explanatory variables. For this model, assume all observations are independent. Provide an explicit (mathematical) description of the model; carefully define all variables and parameters in the model, and clearly state all distributional assumptions. Give a description of how you fit the model and provide a table of all regression parameter estimates with their estimated standard errors. Summarize the results in a table containing the relevant estimated odds ratios and their corresponding 95% confidence intervals. Provide a brief interpretation of these results.

   (c) Repeat (b) using an approach that does not assume observations from the same physician to be independent.

   (d) Compare and contrast the findings from (b) and (c). Comment on how the interpretation of the results do or do not differ. Which model do you recommend to the investigator, and why?

   (e) Provide a succinct yet comprehensive summary of the findings of your selected model with regard to relationships between the response variable and the explanatory variables. Be sure to fully interpret the relevant relationships in a way the study investigators can understand.

   (f) Suppose, counter to fact, that some of the physician preferences (TherStrat) were missing and instead you were provided the data set `logit_miss.csv`. Discuss possible missing data mechanisms based on these data; provide an explicit mathematical statement of any missing data mechanism you discuss. Describe how you would fit the selected model from (d) for this data set (you need not actually fit the model).

   Point distribution: a 3, b 5, c 5, d 4, e 4, f 4

3. Researchers at the UNC Cancer Hospital are leading a multi-institutional study examining the effectiveness of several treatments in patients with HER2-positive breast cancer. Through a cooperative agreement, UNC researchers were able to obtain raw data from multiple independent clinical trials involving similar drugs that have been completed within the past 6 years. In each trial, patients were randomized to one of three arms: combination therapy (denoted AK), antibody only (A), or kinase inhibitor only (K). The file `her2.txt` contains data from nine clinical trials obtained under a data-sharing agreement between institutions. For each study the following variables are recorded:

   - Subject ID: Unique identifier for each subject
   - Trial ID: Identifier indicating the trial each patient participated in (1-9)
   - pCR: pathologic Complete Response. The primary outcome variable determining whether the patient responded to the assigned therapy (pCR=1) or not (pCR=0)
   - Arm: Treatment arm patient was randomized to (1=A, 2=K or 3=AK). Let A be the reference arm/category in the questions below, if relevant.
   - Genes 1-30: The expression measurement for 30 genes. Each gene expression measurement is a continuous variable

   Investigators are primarily interested in whether there is a consistent benefit of the combination therapy in each trial over the single agents. It is also of interest to assess which, if any, of the gene expression measurements are also related to patient response.

   (a) The investigators would like to summarize the pCR rate by arm within each trial. They would also like to evaluate the null hypothesis that there is no difference in response across arms within a given trial (unadjusted for gene expression). In a single easy to read table, summarize for each trial and arm the sample size and pCR response rate, along with 95% confidence intervals (CIs). Also report in the table the result from a hypothesis test comparing pCR between treatment arms for each study. Given this table, discuss the implications of these results in terms of (i) the effectiveness of individual therapies across trials, and (ii) the relative effectiveness of the combination therapy compared to the antibody only arm across trials.

   (b) The investigators would like to perform a test comparing pCR between treatment arms, now adjusting for the expression for gene 1. Report the results of a hypothesis test in each trial comparing pCR between arms adjusting for gene 1 (here and below do not include interaction terms between arm and gene). Also report any related quantities characterizing the estimated effect of gene 1 and adjusted effect for arm in a new table. Summarize the results for each trial in a single easy to read table. Comment on whether the estimated gene 1 effect is comparable across studies, and if not, how to mitigate this problem when comparing pCR between arms while adjusting for expression of gene 1. Use visual evidence to support your answer.

   (c) In addition to the results from part (a), the investigators would like to combine data across trials and estimate an overall odds ratio (OR) for pCR comparing arm AK versus A, adjusted for each patient's gene 1 expression. Fit a logistic regression model for pCR with arm and gene 1 as predictors. Report the estimated OR and corresponding 95% CI for AK versus A based on this model. Provide an interpretation and discuss any potential concerns about this result. Now fit a random effects logistic regression model with a trial-level random intercept and trial-level random slope for arm, assuming fixed effects for arm and gene 1 expression. Comment on the interpretation of this model relative to the previous one, interpreting the model coefficients and estimated OR (and 95% CI) for AK vs A. Clearly explain the differences in the estimated adjusted OR for AK versus A between the two models to the investigators. Conduct a likelihood ratio tests comparing the two models to assess whether there is correlation between observations within the same trial. Include in your answer a formal (mathematical) statement of the null and alternative hypotheses, the test statistic, the distribution of the statistic under the null, and the p-value.

(d) The investigators are interested in whether gene 1 has any association with response in the random effects model in the previous question. Using a likelihood ratio test, carry out a hypothesis test evaluating the null hypothesis that the gene 1 is unrelated to pCR using the random effect model from (c). Then, repeat this test for each of the remaining 29 genes, replacing gene 1 in the previous model with gene 2, and so forth. The researchers are interested in further study of genes where the likelihood ratio test p-value is less than 0.05. Explain why all such genes may not be good candidates to follow up, given the large number of genes that were considered. Suggest an alternative approach to selecting a subset of genes for further study; provide a rationale for the proposed approach and the corresponding list of genes selected.

(e) There is concern that the assumptions for the random effects model from part (d) may not hold for the model including gene 3, and therefore the corresponding hypothesis test result may not be valid. Devise a permutation test to evaluate the null hypothesis that pCR is unrelated to gene 3 expression conditional on arm and trial. Discuss whether the two tests yield different results, and if so, which is preferred.

Points: a 5, b 5, c 7, d 5, e 3

4. Cardiovascular diseases (CVD) are leading causes of death worldwide. A study collected data on a random sample of adults from the United States. The data included sociodemographic and behavior characteristics and CVD risk factors at baseline (visit 1) and at two follow-up visits approximately 5 and 10 years later (visits 2 and 3, respectively). The primary aim of this analysis is to prepare for one or more manuscripts focusing on the relationship between baseline depression and five major cardiovascular risk factors (high cholesterol, high blood pressure, obesity, alcohol use and smoking) over time.

   The file `cvdriskfactors.dat` contains data, formatted as three records per subject. The variables are:

   - id: participant identifier
   - visit: 1, 2, 3
   - female: Sex (1 female, 0 male)
   - age: Age (years)
   - educ: Education at baseline (1 Less than high school, 2 High school, 3 More than high school)
   - hei2010: Healthy Eating Index score at baseline (0 to 100; higher score means healthier diet quality)
   - pag2008: Indicator for meeting physical activity guidelines at baseline (1 Yes, 0 No)
   - depr1: Indicator for depressive symptoms at baseline (1 Yes, 0 No)
   - bmi: Body mass index (kg/m^2)
   - highchol: High cholesterol indicator (1 Yes, 0 No)
   - highbp: High blood pressure indicator (1 Yes, 0 No)
   - obese: Indicator for BMI $>$ 30 (1 Yes, 0 No)
   - alcuse: Alcohol use indicator (1 Yes, 0 No)
   - smoker: Smoking indicator (1 Yes, 0 No)

   Please note that much of the notation will be recycled between parts. For example, Y and β in part (b) are completely different from the same symbols appearing in part (c). But within any one part, the notation will be consistent.

   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. Violating these requirements will result in loss of points. To allow for uniform grading, if quadrature is needed, use 25 nodes. Interpretations should be as clear and as practical as possible, and should be targeted at non-statisticians. For hypothesis testing, use likelihood ratio tests whenever possible/feasible. Otherwise, use Wald-type tests. If an alternative hypothesis is not given, it is understood to be the complement of the null hypothesis. You should describe your methods, present test statistics, degrees of freedom and p-values.

   (a) Given the study objectives, present descriptive table(s) that might be useful for a manuscript.

(b) The study investigator is interested in the relationship between age and BMI. The following model was considered:

$$\mathrm{E}[Y_{ij}] = \beta_1 + \beta_2 x_{ij} + \beta_3 x_{i1} \quad i = 1, \ldots, K, j = 1, 2, 3,$$

where the response Y_{ij} is BMI, i indexes subjects, j indexes visits, and x_{ij} is the age (years) of subject i at visit j. Assume that the 3×1 response vectors $Y_i = (Y_{i1}, Y_{i2}, Y_{i3})^\top$ are independent and follow a multivariate normal distribution with a covariance matrix that has no specific structure. In the context of this model, do the following:

i. Test the null hypothesis that at visit 1, there is no association between age and BMI.

ii. Test the null hypothesis that there is no association between changes in age and changes in BMI over time.

iii. Report parameter and standard error estimates of the two parameters that were tested in the last two sub-parts.

iv. Test the null hypothesis that those two parameters are equal.

(c) The investigator is interested in the association between depression and the five CVD risk factors at visit 1, adjusting for linear age effect, education (factor, 3 levels), sex, diet quality (HEI2010) and physical activity (PAG2008).

Consider the following model for data from visit 1 (only):

$$\mathrm{logit}(\mathrm{E}[Y_{ij}]) = x_i^\top \beta_j \quad i = 1, \ldots, K, j = 1, \ldots, 5,$$

where the response Y_{ij} is a CVD risk factor, i indexes subjects, j indexes the five CVD risk factors, and x_i is a vector that consists of 1 (for an intercept), depression status and other components that appropriately represent the adjustments desired (listed above). Note that each β_j is a vector.

i. Using the data from visit 1 (only), fit the model for each risk factor *separately.* Use Wald test statistics to test the null hypothesis that depression has no effect on any of the CVD risk factors, adjusted as described above. Test the hypothesis at the 0.05 level, using a simple correction for multiplicity (which should be reported).

ii. Again, using the data from visit 1 (only), fit the model for all risk factors *together.* Use the odds ratio as a measure of dependence between CVD risk factors, and, as a working model, assume that all 10 pairwise odds ratios are equal (exchangeable odds ratios). Report the estimated exchangeable odds ratio (point estimate and 95% confidence interval). Test the same null hypothesis given in the previous subpart (but now it is a single test based on a single model fit - do not use a multiplicity correction).

iii. Define $T_i = Y_{i1} + \cdots + Y_{i5}$, the number of CVD risk factors present. Assume that the regression model given above still holds. If, further, $Y_{i1}, \ldots, Y_{i5}$ were all mutually independent, would you expect T_i to have extra-binomial, sub-binomial or binomial variation? Justify your answer.

(d) Now, we focus on the relationship between depression at visit 1 and the CVD risk factors obesity, alcohol use and smoking at all three visits. To this end, consider the following model:

$$\text{logit}(\text{E}[Y_{ijk}]) = x_{ij}^\top \beta_k \quad i = 1, \ldots, K, j = 1, 2, 3, k = 1, 2, 3,$$

where the response Y_{ijk} is a CVD risk factor, i indexes subjects, j indexes the three visits while k indexes the CVD risk factors ($k = 1$ for obesity, $k = 2$ for alcohol use, $k = 3$ for smoking). The vector x_{ij} consists of 1, (age-50)/5, sex and baseline (visit 1) depression status.

i. Fit this model using an unstructured working correlation matrix. Present parameter estimates, standard error estimates and 95% confidence intervals.

ii. Provide a careful and detailed interpretation of the estimated effect of sex on alcohol use. Repeat for the estimated effect of age on alcohol use. The difference between the two interpretations, if any, must be made clear.

iii. Test the null hypothesis that the coefficients corresponding to baseline depression status are all zero. Interpret this hypothesis.

(e) Give a brief summary of the study and the main findings in the above analyses.

Points: a 2, b 7, c 7, d 7, e 2