BASIC DOCTORAL WRITTEN EXAMINATION IN BIOSTATISTICS

DOCTORAL APPLICATIONS EXAM

(9:00 AM Friday, August 3 to 9:00 AM Wednesday, August 8, 2018)

INSTRUCTIONS:

a) 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.

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 8th.

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/2018aug/`

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

1. In many recent complex disease studies, both genetic data (such as single nucleotide polymorphisms [SNPs]) and imaging markers (such as brain volumes on regions of interest [ROIs]) are collected to help identify subjects at risk and predict the disease progression. In this question, we consider data from a study on Alzheimer's Disease (AD) to explore the prediction capability of image and genetic biomarkers for this disease using different models. For subjects in this study, disease severity was categorized into three levels: Alzheimer's disease (AD), mild cognitive impairment (MCI), or normal (NL).

   The dataset ad.csv consists of the following 128 variables:

Column	Variable
1	ID: subject identifier
2	Disease: disease severity/status (AD, MCI or NL)
3-110	108 image ROI biomarkers
111-120	10 SNPs: genotype category 1 or 2
121	Age
122	Gender: 1 female, 2 male
123-127	gPC: genetic principal components
128	Phase: in which phase of the study that this subject is enrolled

   For the following questions, briefly comment on your results/findings if applicable.

   A) Provide a brief summary of the disease status variable. Report how many subjects are enrolled in each phase of the study, and their corresponding disease status. Image data are often collected across different time points by different lab members, creating *batch effects* due to such non-biological experimental factors. Assess graphically if there is evidence of strong batch effects in any image ROI biomarkers (Columns 3-110) that were caused by enrolling in different phases of the study.

   B) We would like to assess whether there is strong association between SNPs and disease status. For each SNP marker (Columns 111-120), fit a separate proportional odds model on the ordinal disease status $\{NL < MCI < AD\}$, while adjusting for the same set of covariates: Age, Gender, gPCs and Phase (Columns 121-128). Briefly explain the key assumption of the proportional odds model and summarize your results of *primary variables* in one or more table(s); note the primary variables in this question are the SNP markers.

   C) Carry out goodness of fit by repeating (B) using other links such as the probit link and/or complementary log-log link for ordinal regression models.

   D) Assess goodness of fit by calculating some variants of pseudo R^2 statistics of the proportional odds models in (B). Discuss whether these SNP markers (as well as these covariates) are good predictors of disease status. Provide evidence to support your answer.

   E) Instead of considering disease status as an ordinal variable $\{NL < MCI < AD\}$, now treat it as a polytomous variable $\{NL, MCI, AD\}$. Fit a multi-categorical logit model between the disease status and *all* of the other variables (Columns 3-128) *in one model*, and then perform variable selection using a proper method you prefer,

such as LASSO, Elastic Net, AIC/BIC, or SIS. Provide the list of variables the method selects.

F) We would like to use the Phase 1 data to predict the disease status of subjects in Phase 2/3 using the variables selected in (E). Here zero-one loss is used, that is, given one new subject with true disease status D_N, we want to find the optimal predicted disease status $\widehat{D_N}$ which minimizes the expectation of the following loss function:

$$E[I(D_N \neq \widehat{D_N})]$$

where $I(\cdot)$ is the indicator function.

i) Treat disease status as a polytomous variable $\{NL, MCI, AD\}$ and use a multi-categorical logistic regression model for prediction. Write down the objective function of your prediction model in detail.

ii) Remove the subjects with MCI disease status in the whole dataset (both Phase 1 and Phase 2/3), and then use a logistic regression model for prediction.

Compare the prediction accuracies and comment on your findings.

G) Treat disease status as a polytomous variable $\{NL, MCI, AD\}$, repeat question (F(i)) with using one of any *other* suitable method, such as random forest, tree-based methods, XGBoost, support vector machines, random effects model, ridge regression, principal component regression, etc. Compare the results with those obtained from the multi-categorical logit model.

H) Write a half-page discussion of all your findings and thoughts when solving this question.

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

2. Rational design of combination drug therapies is a critical component to developing new approaches for treating cancer. Such combinations consist of two or more FDA approved therapies, and when given together may show significantly greater response over what would be predicted from the individual effects of each drug. Many potential drug combinations exist, and therefore preclinical model systems, such as cell lines, are an important experimental tool for evaluating new combinations of existing therapies before moving to clinical trials.

   Researchers at UNC hospitals recently isolated tumors from 100 patients, and were able to culture them into cell lines for testing in the lab. The researchers wish to test the impact of two chemotherapy drugs (palbociclib and oxaliplatin), alone and in combination, on the abundance of a target protein of interest (CDK4) over time. Each patient cell line was randomly assigned to either no treatment, palbociclib only, oxaliplatin only, or to both drugs (four groups total). CDK4 protein abundance was then measured 0, 4, 8, and 16 hours after treatment (0 hours pertains to pre-treatment abundance). One of the goals of this study is to determine whether the decrease in protein abundance over time is greater in the combination group relative to what would be predicted by the effects of each drug alone.

   As the consulting statistician on this project, you receive the data from the researchers in file protein.txt. Due to technological limitations, the abundance of the protein itself cannot be measured directly, and instead the abundances of fragments of this protein (peptides, eight in total) are reported. The peptide abundances are measured in terms of log ionization intensity. The peptides, despite being from the same original protein, have their own baseline levels of ionization, and therefore may differ in their baseline intensities despite having the same abundance. Given the known variation in baseline CDK4 abundance between patients, as well the known variation in baseline ionization between peptides, we need to account for such sources of variation when evaluating protein abundance differences between groups. We assume such sources of variability are random and separate from each other (non-nested).

   In file protein.txt, the following variables are present

   - logInt: The log ionization intensity for a given peptide, cell line, and timepoint
   - palbo: indicator of whether logInt value came from sample receiving palbociclib treatment (0 no, 1 yes)
   - ox: indicator of whether logInt value came from sample receiving oxaliplatin treatment (0 no, 1 yes)
   - time: The timepoint that logInt was taken from (0, 4, 8, 16 hours). This variable can be considered as continuous.
   - cellID: Unique identifier of the cell line pertaining to logInt
   - pepIndex: The index of the particular peptide pertaining to logInt (8 detected peptides total)

   A) For the first patient, visually depict the log intensity of each peptide over time in a single informative plot. Do the same for the seventy-sixth patient in a separate plot. Given that individual peptides derive from the same original protein, what is the scientific relevance of the magnitude of logInt versus the relative change in logInt

over conditions? Based on these two figures, discuss the relative evidence behind true protein abundance change over time for patient 1 and patient 76.

B) Using a linear mixed model (with random intercepts), test whether there is a significant difference in trend over time in protein abundance for the combination therapy samples versus what would be predicted from the main effects of each drug alone. Account for peptide-level and sample-level random effects (do not have to be nested). Report the model you utilize, the parameter estimates of this model, the test and statistic utilized, the distribution of this test statistic under the null hypothesis, and the p-value. Interpret the result of this test.

C) Evaluate the model assumptions for these data given the fitted model from part (B).

D) Assume that there is no difference between groups in the change in protein abundance over time (time fixed effect is the same across treatment groups). For this reduced model, describe how one may evaluate whether a pepIndex-level random slope for time may be necessary, and carry out this procedure. Discuss your results. Even if you do not reject the null hypothesis, describe how one may interpret the model if a pepIndex-level random slope for time is included.

E) Suppose we discretize the log intensity values for each peptide and patient such that values > 12 represent high intensity (1), and values ≤ 12 represent low intensity (0). Then, fit a model using `glmer()` from the `lme4` package in R, including only an intercept, a fixed effect for time, and a peptide-level random intercept. Refit this model using the argument `nAGQ = 10`. Alternatively, use `proc glimmix` in SAS with the option `method=quad(qpoints=10)` and also with `method=LAPLACE`. Show the results differ numerically, and explain which approach may yield more accurate results.

F) In typical proteomic analyses logInt values are often missing. This is due to the fact that the machine utilized for the experiment can only measure a fixed number of peptides per run, and that patients with less abundant levels of protein expression will have higher probabilities their peptides are not detected. Discuss how the estimates of the fixed effect model coefficients in (B) may be biased in such a setting. Some scientists propose replacing the missing values in a given sample using a single imputation step, replacing missing values in a sample with those drawn from a Gaussian distribution with mean μ and standard deviation (SD) σ, where μ is a value three SDs below the minimum observed value in the sample, and σ is calculated from the observed data. Discuss the pros and cons of such an approach and the assumptions made to handling missing data in this setting.

Point distribution: A 3, B 7, C 3, D 6, E 3, F 3

3. Veterinary investigators interested in the risk of horses acquiring West Nile virus wish to examine the relationship between the frequency of cases of horses with this virus and the corresponding frequency of cases in birds. The investigators collected data on cases of West Nile in horses and birds at the county level within a particular state, as well as related information on the numbers of farms in the county, the area of the county, and the human population of the county.

   They have pre-specified interest in a standardized bird virus rate, defined as the count of bird West Nile cases divided by the respective human population count. Likewise, the standardized horse virus rate is defined as the count of horse West Nile cases divided by the respective number of farms.

   The investigators have asked you, as the project's biostatistician, to fit an appropriate statistical model to these count data, where the standardized horse virus rate would be considered as the response, and both the standardized bird virus rate and the human population density as explanatory.

   The dataset VIRUS.DAT contains the following variables:

   - CountyID: Unique county identifier
   - HorseCases: Number of West Nile cases in horses
   - NumFarms: Number of farms
   - PopDensity: Human population per square mile
   - BirdRate: Standardized bird virus rate

   A) Examine and thoroughly describe the distribution of the count of horse West Nile cases. Include quantitative measures, as well as appropriate visualizations.

   B) As an initial step, fit a Poisson model to the standardized horse virus rate, including main effects for both the standardized bird virus rate and the human population density, as well as their interaction. Use an appropriate offset to account for the number of farms in each county. Report the parameter estimates and their standard errors. Interpret the results relative to the investigators' primary interest in assessing to what extent there is an association between West Nile virus in horses and West Nile virus in birds.

   C) Based on prior experience, there is concern about possible overdispersion. Briefly, describe this phenomenon in statistical terms, and modify the model in (B) to account for potential overdispersion using the negative binomial distribution. Report the parameter estimates and their standard errors. Interpret the results as in part (B).

   D) There may be excess zeroes in the data. Briefly, describe this phenomenon in statistical terms, and make modifications to the model in (B) to allow for a separate statistical model for these excess zeroes. Discuss your approach to fitting this model (in addition to the Poisson model for the rate). Report both sets of parameter estimates and their standard errors. Interpret the results as in part (B).

   E) To be comprehensive, fit a model that accounts for both overdispersion and excess zeroes in the data. Discuss your approach to fitting such a model or models. Report

appropriate parameter estimates and their standard errors, and interpret the results as in part (B).

F) Determine your preferred model from among (B) through (E), and justify your choice. Compare and contrast the findings of this model to the other three candidate models.

G) Provide a concise description of the various statistical methods you have used in (A) through (E) that would be suitable for a scientific veterinary journal that is similar to scientific clinical journals (e.g., *JAMA*, *New England Journal of Medicine*). Likewise, provide a concise description of your findings for your preferred model from (B) through (E), including appropriate tables or figures.

Point distribution: A 2, B 3, C 4, D 4, E 4, F 4, G 4

4. A two arm, unblinded, randomized efficacy trial was conducted in adults living in the US to assess a Mediterranean-style diet weight loss intervention. The study was conducted in four primary care sites and included $N = 400$ participants. The intervention was delivered in three phases over three years: phase I (6 months) to adopt and maintain a Mediterranean-style diet, phase II (12 months) to focus on weight loss, and phase III (18 months) to maintain or continue weight loss. Participants were randomized to the intervention or control (usual standard care) in a 1:1 ratio, stratified by diabetes status. Average weight loss was hypothesized to be the same by diabetes status. Data were collected at 0, 6, 12, 18, 24 and 36 months. The primary objective of the trial was to assess the effect of the intervention on weight loss at 36 months. The secondary objective was to assess the intervention's effect at the other time points.

   The file WEIGHTLOSS.csv includes data for 400 participants measured over a period of three years, with the following variables:

   - ID: Unique participant identifier
   - SITE: Primary care identifier (1, 2, 3, 4)
   - GROUP: Intervention indicator (1 Intervention, 0 Control)
   - DIABETIC: Diabetes indicator (1 Yes, 0 No)
   - AGE: Participant's age (years) at baseline
   - MALE: Male indicator (1 Yes, 0 No)
   - SMOKING: Smoking status (1 never, 2 former, 3 current)
   - DEPRESSION: Depressive symptoms at baseline (1 Yes, 0 No)
   - MONTH: Clinic visit month (0, 6, 12, 18, 24, 36)
   - WEIGHT: Measured participant weight (kg)

   In answering the following questions, when reporting the results of a statistical test, state the null and alternative hypotheses, the test statistic, the distribution of the statistic under the null, the p-value, the significance level, your decision (reject/do not reject), and a brief interpretation of the result in terms of the subject matter.

   A) Provide descriptive statistics for baseline variables and for the outcome. Graph (i) the outcome over time for some (or all) participants and (ii) the average outcome over time by diabetes status. Summarize the findings briefly in one paragraph.

   B) Describe any missing data patterns. Discuss the possible missing data mechanisms. Provide an explicit (mathematical) statement of any missing data mechanism you discuss. Explain to what extent the observed data provide evidence about the missingness mechanism.

   C) Provide an explicit (mathematical) form for a linear mixed effects model which can be used to assess the primary and secondary objectives of the trial. Carefully define all variables and parameters in the model, and clearly state all distributional assumptions.

      i) Provide a justification for how baseline weight is handled in the modeling.

ii) Provide a justification for how other baseline measurements are (or are not) included in the model.

iii) Specify a hypothesis test corresponding to the primary aim of the trial.

D) Fit the model specified in (C), conduct the test in part (iii), and estimate the effect of the intervention on weight loss at 36 months. Test whether the effect is modified by diabetes status.

E) Discuss how missing data was handled in part (D) and under what assumption(s) about the missing data mechanism are the resulting inferences valid. Do the results in part (B) support the assumption(s)?

F) Evaluate the fit of the model specified in (C) for these data. Use graphical methods to look for isolated departures from the model. Also look for subjects and individual observations with too much influence on the estimated regression coefficients. Also explore the distribution of the residuals relative to the model assumptions. (Hint: The diagnostics may be obtained from a marginal model).

G) Repeat part (D) at other time points to address the secondary objective of the trial.

H) Write one or two paragraphs summarizing all the findings, and reference figures and/or tables to support your statements.

Point distribution: A 2, B 2, C 5, D 4, E 3, F 5, G 2, H 2

2018 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