BASIC DOCTORAL WRITTEN EXAMINATION IN BIOSTATISTICS

DOCTORAL APPLICATIONS EXAM

(9:00 AM Monday, August 2 to 9:00 AM Saturday, August 7th, 2021)

INSTRUCTIONS:

- This is an open book, take home examination. You may not communicate with anyone except Naim Rashid (naim@unc.edu) about the content of this examination. Professor Rashid 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 Tiffany Harris via email (tjharris@email.unc.edu) by 9am on Saturday, August 7th. 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 Tiffany Harris 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/2021aug/`

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

1. A study was conducted consisting of children at risk of injuries. Children were followed until they turned 5 years old. Each record (row) in the study data file "injdeath.csv" pertains to a distinct combination of covariate values, where each covariate is defined below. As a result, the contribution of each record is weighted by the total person years (childyrs) tabulated in that particular distinct covariate combination, as records may pertain to varying numbers of subjects who are followed for a varying amount of time. The variables are defined as the following in the file "injdeath.csv":

   - age: Age of the child in years at study entry
   - age_mom: Age of mother at birth of child, categorized as 19 (age $<$ 20), 24 ($20 \leq$ age ≤ 24), 29 ($25 \leq$ age ≤ 29), and 30 (age > 30).
   - lbw: low birth weight, categorized as 0 ($\geq$ 2500 g) and 1 ($<$ 2500 g)
   - educ_mom: Mother's years of education, categorized as 11(< 12) years ,12 (12 years) 15 (13-15 years), and 16 (> 16) years.
   - income: Quintile of average income in mother's neighborhood
   - prox: Proximity to playground equipment, categorized as 0 (greater than or equal to a half mile away) and 1 (less than a half mile away)
   - oth_chld: Number of other children, categorized as 0 (no siblings) , 1 (1 sibling) , 2 (2 siblings), 3 (3 siblings), and 4 (4 or more siblings).
   - race_mom: Mother's race, categorized as 0 (white) and 1 (black)
   - pnclate: Late or absent prenatal care, categorized as 0 (in first four months of pregnancy, care was received) and 1 (care was not received within the first four months of pregnancy)
   - childyrs: Total person-years recorded in the study pertaining to a particular distinct combination of values (record) with respect to the above covariates.
   - inj_dth: Total number of injuries or deaths recorded during the study period pertaining to a particular record in the study,

   Please use the lowest level of each categorical variable as the reference level, when relevant, in answering the following questions below.

   (a) Provide descriptive statistics of each variable above, with the exception of childyrs, in a table. Write a paragraph to describe the data displayed and give insights made from the descriptive study.

   (b) Estimate the risk of injury or deaths in children who live less than a half mile away from a playground, relative to those children who live greater than or equal to a half a mile away. Calculate and interpret a 95% confidence interval for this relative risk. What can be said about the likelihood of injury or death for children who live closer to playgrounds relative to those who live further away?

   (c) Using an appropriate model, estimate and interpret a 95% confidence interval for the relative risk of injury or death in children living close to playgrounds relative to children living further away. In this question, please adjust for all provided covariates, and assume that the logarithm of the relative risk is a linear function of the covariates. Write the likelihood for the model in mathematical notation. How does your estimate and confidence interval compare with your answer from part (b)? Interpret the regression coefficient associated with children living closer to playgrounds. Clearly state all assumptions of the model and whether the assumptions are likely to hold (provide justifications via plots/tables where possible).

(d) A subject matter expert suggests that you may need to account for the fact that the assumed conditional variance pertaining to the model from part (c) may be a linear function of the conditional mean. Utilize an appropriate statistical model to account for this proposed relationship. Based on this model, are there any covariates that suggest an increased risk of injury or deaths? Does the interpretation of the results from this new model differ from results from questions (a)-(c)? Why or why not?

(e) A second subject matter expert suggests that for various reasons a large proportion of the children in the study actually aren't at risk for injury or death. Furthermore, they suspect that among the children at risk for injury, the assumed form of the conditional variance will be a quadratic function of the conditional mean in the model. Suggest a new model to handle such a relationship and compute a confidence interval for the relative risk of injury or death for children living closer to playgrounds against those living further away. Write the likelihood for the model in mathematical notation, and clearly state the form of conditional mean (expected value of injury or death) each observation in terms of this model. Put all confidence intervals from parts (b) through (e) in a table. Given the evidence, argue whether proximity to a playground increases the risk for injury deaths compared to living further away from one.

(f) You have now been told that the maximum number of siblings observed in the data set is 4, so that the `oth_chld` variable could be measured as a numeric variable. Conduct the likelihood ratio, score, and Wald tests for test the null hypothesis of whether the number of siblings is linearly related to the logarithm of the relative risk. For the likelihood ratio and score tests, state the reduced model under H_0. For the Wald test, state the contrast matrix utilized for the test. Also, state the asymptotic distribution and report the test statistics and p-values of all three tests rounded to 4 decimal places. Use one or more plots to support your conclusion of whether the effect of siblings is linear or nonlinear.

Point distribution: a 2, b 3, c 4, d 4, e 5, f 7

2. A cross-sectional study was conducted to determine whether diabetes was associated with periodontal disease in a certain population of adults aged 20 and higher. There is one record per individual and 5000 individuals in the dataset, “dentalperio.csv”. The following variables were collected:

variable	Description:	Coding
periomodsev	Periodontal Disease:	1=yes, 0=no
agecat	Age:	1 = 20-34 yrs
		2 = 35-44 yrs
		3 = 45-54 yrs
		4 = 55 yrs and older
smoker	Smoker:	1=yes, 0 = no
hinsurance	Health insurance:	1=yes, 0=no
female	Gender:	1=female, 0=male
diab	Diabetes:	1=yes, 0 = no

In all questions, treat *agecat* as a nominal (categorical) variable. Also, use the lowest level of each categorical variable as the reference level in your models.

(a) Provide descriptive statistics for the data. Include the marginal observed counts and relative frequencies of periodontal disease for the levels of each of the following factors, separately: $agecat$, $smoker$, $hinsurance$, $female$, and $diab$. Summarize results in terms of the observed patterns of periodontal disease.

(b) Provide an estimate and a 95% confidence interval (CI) of the marginal odds ratio (OR) for periodontal disease with respect to diabetes status, unadjusted for the demographic and health factors. Additionally, provide estimates and 95% CIs for the ORs of periodontal disease and diabetes status for smokers and non-smokers, separately.

(c) Define a logistic regression model with periodontal disease as the outcome, main effects pertaining to the $agecat$, $smoker$, $hinsurance$, $female$ and $diab$ variables, and interaction term pertaining the $smoker * diab$ interaction. Using detailed notation, write out the model expression including coding used for all variables.

(d) Test the goodness-of-fit of the fitted model in part (c) where the model under the alternative hypothesis is the logistic regression model with all main effects and the four two-way interactions: $agecat * diab$, $smoker * diab$, $hinsurance * diab$, and $female * diab$. Specifically, report and interpret the goodness-of-fit based on the likelihood ratio test, the score test and the Wald test results, respectively.

(e) Report the maximum likelihood estimates and standard errors of regression coefficients for the model in part (c). Provide covariate-adjusted 95% confidence intervals of the odds ratio for periodontal disease and diabetes for smokers and non-smokers, respectively. Interpret results using language a non-statistician can understand.

(f) Based on weighted least squares (wls), define a linear model for the empirical logits, $f_i = \log[p_i/(1-p_i)]$ where $p_i = y_i/n_i$ are the observed proportions of individuals having periodontal disease (count $= y_i$) for each group with size n_i, $i = 1, \ldots, 64$ defined by the cross-classification of *agecat*, *smoker*, *hinsurance*, *female*, and *diab*. Specify the model to utilize the same main effects and interaction term as the model in part (c), and write it down using detailed notation. Additionally, provide an expression for the weighted least squares estimator of the regression coefficient vector in this model using inverse variance weights based on $\text{var}(f_i)$. Use matrix and vector notation as applicable.

(g) Provide the residual chi-squared goodness-of-fit test for the model in part (f) and interpret the result.

(h) Report the weighted least squares estimates of the model in part (f) and their associated standard errors. Provide covariate-adjusted 95% confidence intervals of the odds ratios relating to the association of periodontal disease and diabetes for smokers and non-smokers, respectively. Interpret the result using language a non-statistician can understand.

(i) Discuss at least one relative strength and at least one relative weakness for each of the maximum likelihood and weighted least squares estimation approaches for logistic regression.

Point distribution: a 2, b 2, c 3, d 4, e 3, f 3, g 2, h 4, i 2

3. High blood pressure is a major risk factor for recurrent stroke. Therefore, controlling blood pressure is an important strategy for secondary stroke prevention. You are collaborating with a team of physicians on the design of a clinical trial to evaluate the relative effectiveness of two treatment strategies for hypertension (high blood pressure) in stroke patients discharged home from the hospital. The trial will evaluate whether a novel strategy for blood pressure management, Intensive Personalized TeleHealth Management (IPTM), is superior to what is considered standard of care for this population, Intensive Clinic-Based Management (ICBM). Patients in the trial will be randomized to one of the two arms, each arm representing one of the two treatment strategies described. The treatments will be compared with respect to mean change from baseline in systolic blood pressure (SBP).

   In order to help design the trial, your collaborators have provided data from a previously conducted study in a similar stroke population that used only ICBM to treat hypertension. Therefore, only information relevant to ICBM is available in the pilot data, and not IPTM. The pilot data are available in the file “pilot.CSV”.

   The CSV file contains the following variables:

   (1) PATID = De-identified patient number

   (2) MONTHS = Number of months from baseline at measurement time

   (3) BASE = Baseline value of SBP (mmHg)

   (4) SBP = Value of SBP (mmHg)

   (5) CHANGE = Change from baseline in SBP (mmHg)

   Note that blood pressure measurements were not taken at specific intervals in the prior study, and so patients have measurements at different times and in different frequencies. Standard clinical practice is to treat hypertensive patients to reach a systolic blood pressure level between 120 mmHg and 130 mmHg. Thus patients who are comparatively less hypertensive at baseline would be expected to exhibit less change from baseline on average, regardless of treatment strategy.

   (a) Descriptively summarize the pilot data and comment briefly on any features of the data that you feel are important. Summaries should use professional quality data tabulations and visualizations to support your discussion.

   (b) Fit a linear mixed model (LMM) to the pilot data utilizing change from baseline in SBP as the outcome variable, assuming a linear trajectory over time (months), adjusting for baseline SBP, and including an interaction between baseline SBP and time. For analysis, center the baseline SBP covariate at its sample mean and convert it into 5 mmHg units.

      (i) Using the requested model for the mean, determine an appropriate structure for the random effects model. For this exercise, choose between a model with a random intercept and a model with a random intercept and slope. Formally specify and test whether the random slope is needed in the model using a likelihood ratio test conducted at significance level $\alpha = 0.05$.

(ii) Clearly and concisely state the final model you select (using rigorous mathematical notation). Present parameter estimates, standard errors, and 95% confidence intervals for the regression parameters as well as the estimated variance/covariance matrix for the random effect(s) and the estimated error variance. Describe the statistical methods used to estimate the model. Interpret the regression parameter estimates. Use professional quality data tabulations and visualizations to support your discussion.

(iii) Comment briefly on the model's fit and any concerns you may have regarding potential misspecification. Note that sensitivity analyses are NOT requested.

(c) The proposed trial will assess SBP at four in-person clinic visits (baseline, 2 months, 4 months, and 8 months) for all randomized patients (i.e., for both study arms). Because the trial proposed will compare two effective treatments, a decrease in SBP over time is expected in both arms on average. *Regardless of the baseline SBP value for the patient*, your collaborators hypothesize that patients receiving IPTM will have an additional SBP reduction of 4.0 mmHg on average at 8 months (compared to the average SBP reduction in the ICBM arm at 8 months) and that the additional SBP reduction afforded by IPTM will increase linearly up to 4.0 mmHg at 8 months.

Using the fitted model results from part (b), the baseline SBP distribution for the pilot data, and the information provided above about the hypothesized treatment effect of IPTM compared to ICBM, compute the sample size required in order to have 90% power to detect a difference in the SBP change from baseline trajectories for IPTM compared to ICBM over the period 2 to 8 months post-baseline. You may assume the information from the fitted model in part (b) provides accurate information on the ICBM mean change from baseline trajectory, that estimates of variances and/or covariance obtained from the model in (B) are applicable to both arms, that randomization will be 1:1 across the two arms, and that 100% follow-up occurs (i.e., no dropouts). Clearly state the required sample size and concisely describe the methods used for sample size determination with sufficient detail such that another qualified statistician could reproduce the results based on your written description of the methods used. You may identify the required sample size using simulation.

(d) Assume that in fact dropout *is* related to treatment (e.g., patients receiving IPTM are less likely to drop out compared to patients receiving ICBM) and that patients who fail to achieve a 2.0 mmHg or greater decrease in SBP by 4 months are more likely to drop out of the study before the ~~6~~ month assessment. For such a case, will analysis of the observed data using the LMM (without use of specialized methods to address missingness) be subject to bias due to the missing data? Rigorously justify your answer.

(e) Assume that in fact dropout *is* related to treatment (e.g., patients receiving IPTM are less likely to dropout compared to patients receiving ICBM) and that patient's who fail to achieve a 2.5 mmHg or greater decrease in SBP by 4 months are more likely to drop out of the study before the 8 month assessment. In such a case, will analysis of the observed data using the LMM (without use of specialized methods to address missingness) be subject to bias due to the missing data? Rigorously justify your answer.

(f) The primary analysis for the proposed trial (as is the case in most clinical trials) will be an *Intention-to-Treat* (ITT) analysis. An ITT analysis includes all randomized patients in the groups to which they were randomly assigned regardless of their adherence to the intervention. Your collaborators are concerned that as many as 10% of patients may be non-compliant with their medication (a key component of both interventions) and therefore have no reduction in SBP over the 8 month study period (regardless of whether they are in the IPTM or ICBM arm). Using the sample size identified from part (c) as providing 90% power, recompute power under the assumption that approximately 10% of participants will be non-compliant and thus experience no change from baseline in SBP on average over the 8 month study period. You may assume a 10% non-compliance rate in both arms and no dropout for this part of the question. Describe the implications of non-compliance on power for your collaborators.

Point distribution: a 2, b 5, c 5, d 3, e 5, f 5

4. You are the biostatistician on the research team that is conducting a randomized clinical trial (RCT) for a novel dietary and nutritional intervention that aims to lower hemoglobin A1c (A1C), which is a commonly used diabetes biomarker for hyperglycemia; it reflects the average blood sugar levels over the past over 3 months. Higher A1C values indicate a higher risk for developing diabetes. Specifically, the target population is parents at risk for diabetes and their at-risk children aged 13-18 years. For this trial, a parent and one of their at-risk teenaged children (considered together as a "dyad") were jointly randomized either to receive this novel intervention for 6 months or to a control intervention pertaining to standard of care (also lasting 6 months). Baseline data were collected on both the parents and their children, and post-intervention data were collected at single clinic visit 6 months later.

   The primary aim of the trial was to determine the efficacy of this intervention on lowering A1C at 6 months for these at-risk parents and for their at-risk teenaged children. Values below the clinical threshold of 5.7% are considered as 'normal', while values of at least 5.7% are indicative of 'prediabetes' or 'diabetes'.

   A1C.csv contains data on n=188 pairs of parents/children. Each observation has the following variables:

   - GROUP: randomized group, either the Novel intervention group ("Intervention"), or the Control intervention (standard of care) group ("Control").
   - ParentA1C_1: parent's A1C at baseline
   - ParentA1C_2: parent's A1C at follow-up
   - ChildA1C_1: child's A1C at baseline
   - ChildA1C_2: child's A1C at follow-up

(a) Tabulate descriptive statistics for all baseline variables, both for the overall sample as well as by intervention group. Also, provide descriptive statistics for all relevant follow-up variables by intervention group.

(b) Evaluate the effect of the novel intervention on mean change in A1C for the parents and for the children in a single statistical model at 6 months, adjusting for the respective baseline value of A1C. For this model, first provide a mathematical specification, interpret its parameters, state its assumptions, and provide a summary of the numerical results. Then, carry out a statistical test for whether these (baseline-adjusted) novel intervention effects (relative to the Control) are homogeneous across the parents and the children. Provide thorough diagnostics to evaluate the validity of the model assumptions. If any appear to be violated, please discuss your approach to remedy the violation(s).

(c) Evaluate the effect of the novel intervention on the proportion having A1C $\geq$ 5.7% for the parents and for the children in a single logistic model at 6 months , adjusting for the respective baseline value of this dichotomized variable. For this model, provide a mathematical specification, interpret its parameters, state its assumptions, and provide a summary of the numerical results. Expanding this model, provide a formal evaluation of whether these novel intervention effects (relative to the control) are homogeneous across the parents and the children. Provide a mathematical specification of the model that allows for heterogeneous novel intervention effects. Then, specify linear combinations of the regression parameters that will estimate and test these effects; provide a summary of these estimates and tests for the novel intervention effects.

(d) Report on your models from parts (b) and (c) in 2-3 paragraphs that would be suitable for the Methods section of a manuscript targeted to a clinical journal focusing on diabetes. In these paragraphs, you should include a clear description and rationale for your statistical modeling approach. Also, provide 2-3 paragraphs (plus related tables or figures) for the manuscript's Results section that comprehensively summarize your findings, including the estimates of the novel intervention effects (with appropriate measures of variability) for both the continuous and categorical versions of the A1C variable, separately for the parents and the children. Be sure to include in your discussion your findings regarding the presence of heterogeneity in these novel intervention effects across parents and children.

(e) Create a half-page summary of your findings regarding the novel intervention for easy reference by physicians. Use bullet points and figures/tables rather than complete sentences to summarize the most important findings of the RCT.

(f) Briefly describe two main advantages and two main disadvantages of this RCT design, considering why this dyadic approach to randomization may have been favored. Compare this to an alternative RCT design where the same at-risk parents would be randomized in one trial, and their (same) at-risk teenagers would be randomized in a separate trial: in 2-3 sentences, discuss two main advantages and two main disadvantages relative to the dyadic RCT design used for parts a-e of this problem.

Points: a 2, b 6, c 5, d 5, e 4, f 3