BASIC DOCTORAL WRITTEN EXAMINATION IN BIOSTATISTICS

DOCTORAL APPLICATIONS EXAM

(9:00 AM Friday, August 4 to 9:00 AM Wednesday, August 9, 2017)

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 9th.

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/2017aug/`

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

1. An investigator wishes to examine the relationship between the opium trade and terrorism in Afghanistan using data on terrorist attacks and casualties in the 34 Afghan provinces from 1996 to 2008. The analysis should consider various economic development, infrastructure, geographic, security, and cultural factors when examining causes of terrorism in the provinces. The data set contains two response variables: y_1 = terrorism incidents, and y_2 = terrorism casualties. Covariates include opium production, population, area, terrain, nutrition, literacy, drinking water, roads, infant mortality, Pashtun majority, and foreign troop presence for each province. The data are contained in the file terror.dat with the variables in the following order:

Variable	Column
Province	1
Total terrorism incidents (y_1)	2
Total terrorism casualties (y_2)	3
Average opium cultivation (hectares)	4
Population (1000s)	5
Area (1000s km^2)	6
Mountainous (%)	7
Literacy rate	8
Access to drinking water (%)	9
Below minimum calories (%)	10
All-season roads (%) 90-92	11
Under 5 mortality (per 1000)	12
Pashtun majority (1=Yes,0=No)	13
Average foreign troops (per year)	14

A) Rigorously show in detail that the negative binomial regression model can be obtained as an over-dispersed Poisson model, and explicitly give the form and the role of the over-dispersion parameter.

B) Analyze the data using a negative binomial regression model for the response y_1 treating province as the "subject" using all of the covariates in the model, assuming a canonical link. Use the logarithm of opium cultivation, province population, and province area. Carry out goodness of fit using other links such as the identity link, and inverse link.

C) Carry out a formal score test for the hypothesis that the dispersion parameter is zero using y_1 as the response variable and using all of the covariates. 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, the p-value, and your conclusions from the test.

D) Using the canonical link, carry out a residual analysis for the model in part B, using measures such as Cook's distance and the Andrews-Pregibon statistic to identify any outlying observations.

E) Instead of carrying out negative binomial regression, suppose we carry out a linear regression analysis on y_1 using all of the covariates in part B. Carry out a Box-

Cox transformation on the y_1 to find the best transformation on y_1 and provide an approximate 95% confidence interval for it. How do your results compare to parts B and C?

F) We would like to address the following three hypotheses using negative binomial regression and using both y_1 and y_2 in separate analyses:

   (i) Hypothesis 1: Afghan provinces characterized by higher levels of opium production are more likely to experience higher levels of terrorist activity

   (ii) Hypothesis 2: The positive relationship between opium production and terrorist activity in the Afghan provinces is independent of, and robust to the inclusion of social, economic, demographic, geographic, religious/cultural, and security factors that predict terrorism.

   (iii) Hypothesis 3: Additional troop deployment to economically impoverished provinces is correlated with a smaller increase in terrorist activity than in more economically stable regions. [Interpret the magnitude of any effect modification in your model and reflect on one's ability to make statements regarding the causality of these relationships given the available data.]

G) Discuss in detail how one might construct a joint regression model for (y_1, y_2).

H) Write a detailed two-paragraph summary of your findings.

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

2. Body mass index (BMI) is a measure that can be used to determine whether a child is overweight or obese. Overweight is defined as a BMI at or above the 85th percentile and below the 95th percentile for children of the same age and sex. Obesity is defined as a BMI at or above the 95th percentile for children of the same age and sex.

   A randomized intervention trial for body weight management was conducted in overweight and obese school-aged children. Children were randomized either to a group that would receive the weight management intervention or to a control group. Baseline data prior to randomization were collected on several variables for each child. Follow-up data on the children were collected six months later.

   The researchers would like to assess the effect of the intervention on managing weight, but they have some concern regarding attrition during the study. In particular, they are concerned that obese children may have been more likely to drop out of the trial prior to the follow-up visit and have missing values for that visit.

   CHILD.dat contains data on $n = 204$ children enrolled in this trial. Each observation in the dataset contains values for the following variables (in this order):

   - RX: the group to which the child was randomized (Intervention, Control)
   - Y0: indicator whether child was obese at baseline visit (1 obese, 0 not obese)
   - Y1: indicator whether child was obese at follow-up visit (1 obese, 0 not obese, NA missing)
   - AGE: child's age in years at baseline
   - SEX: child's sex (Male, Female)
   - RACE: child's race (African-American, Caucasian)
   - PARENT_WT: an indicator for whether either of the child's parents was classified as obese (i.e., body mass index ≥ 30 kg/m^2) at baseline.

A) Tabulate descriptive statistics for the collected baseline variables. Provide this information for the overall sample, as well as by intervention group. Also provide descriptive statistics for the outcome variable Y1 by intervention group. Provide brief commentary (e.g., one paragraph) on your findings.

B) Examine and describe the extent of missing data for the outcome variable. Assess whether missingness varies by intervention group, and describe how this could impact the analysis results.

C) Considering only the complete cases (i.e., excluding any observations with missing values), assess the effect of the intervention on the risk of being obese at follow-up using an appropriate statistical model, adjusting for child's baseline obesity status, age, sex, race, and parents' obesity status. Concisely summarize your findings, including an interpretation of the relevant parameter estimate(s) (or a meaningful transformation thereof) and corresponding 95% confidence interval(s).

D) Repeat part C using multiple imputation (MI) instead of complete case analysis. Specifically, perform multiple imputation for the outcome variable using regression imputation with the available baseline variables (including Y0) as predictors. Carefully describe each of the steps involved in the MI analysis, including an explicit (mathematical) form of the imputation model as well as how the final effect estimate(s), variance estimate(s), and confidence interval(s) are computed. Concisely summarize the results from the MI analysis as in part C.

E) Discuss the possible missing data mechanisms for which the analyses in parts C and D are valid. 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.

Point distribution: A 3, B 3, C 4, D 9, E 6

3. A multi-site study is collecting anthropometric data on adults over one year. Staff were trained according to standardized procedures for performing anthropometric measures on study participants. By protocol, 5% of the participants were randomly selected and re-measured by the clinic manager. As part of the quality assurance of the study, you have been asked to assess differences among staff and trends over time.

   The file ANTHRO.csv has data for 1057 participants measured over a period of one year by different staff.

   - ID: Unique participant identifier
   - AGE: Participant's age (years)
   - MALE: Male indicator (1 Yes, 0 No)
   - HEIGHT: Participant height (cm) measured by staff
   - WC: Waist circumference (cm) measured by staff
   - WC_GOLD: Waist circumference (cm) measured by clinic manager
   - STAFF: Unique staff ID
   - MONTH: Clinic visit month (1 to 12)

   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, and a brief interpretation of the result in terms of the subject matter.

A) Describe the demographic characteristics of participants by staff.

B) Provide descriptive statistics and figures displaying height over time by staff. Include a brief (e.g., one paragraph) commentary.

C) Specify and fit a linear model for height with independent variables staff and time to assess differences among staff and mean trends over time. Provide an explicit (mathematical) form for the model, with all variables and parameters carefully defined and all distributional assumptions clearly stated.

   i) Assess whether trends over time differ by staff.

   ii) Overall, is there a trend over time?

   iii) Are there differences among staff? If so, which?

   iv) Summarize the findings and results. Reference tables and/or figures to support your findings.

D) Specify and fit a linear mixed effects model for height including time as a fixed effect and random effects to account for between staff variability. Provide an explicit (mathematical) form for the model, with all variables and parameters carefully defined and all distributional assumptions clearly stated. Assess differences among staff and mean trends over time. Summarize the findings and results. Discuss whether this model or the model from part C is more appropriate for analysis of data from this study.

E) Provide descriptive statistics and a figure (or figures) for the two measurements of waist. Include a brief (e.g., one paragraph) commentary.

F) Provide an analysis assessing the extent to which there is agreement between waist measurements by the staff and the clinic manager. Interpret the results, assuming the clinic manager measurements are the gold standard (i.e., measured without error).

G) Write one or two paragraphs describing the results and recommendations, and reference graphs and/or tables to support your findings.

Point distribution: A 1, B 2, C 8, D 6, E 2, F 4, G 2

4. Low birth weight is an important public health problem because it is associated with perinatal mortality and morbidity, stunted growth and chronic diseases in adult life. Hence it is important to identify risk factors for low birth weight. A study was conducted to investigate the risk of low birth weight and how it relates to birth order and to mother's age. Data were collected (from hospital records) on all women with exactly five children (on the date the study started) in a certain geographic area in the United States. There were $K = 198$ women who qualified for the study and the data are available in files `babies1.dat` and `babies2.dat`. The two files contain the same data, but the first is formatted with five records per subject, while the second has one record per subject.

   Let the random variable Y_{ij} represent the birth weight in grams of the jth child for the ith woman, and let age_{ij} denote her age in years (truncated to an integer value) when she delivered her jth child. Define the vector Y_i to be $Y_i := (Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4}, Y_{i5})^\top$. In what follows, when fitting models for Y_i we will assume that Y_i is distributed as multivariate normal and specify various models for its mean and covariance. Subjects are assumed mutually independent throughout.

   Two common cutpoints used to define "low birth weight" are 2500 grams and 3000 grams. In this problem we will adopt the latter; a birth weight under 3000 grams will be considered a low birth weight. Hence, define Y^*_{ij} as follows: $Y^*_{ij} = 1$ if $Y_{ij} < 3000$ and $Y^*_{ij} = 0$ if $Y_{ij} \geq 3000$. Define the vector Y^*_i to be $Y^*_i := (Y^*_{i1}, \ldots, Y^*_{i5})^\top$.

   In what follows, for hypothesis testing, use likelihood ratio tests whenever 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.

   A) Present one or two useful summaries, numerical and/or graphical.

   B) Model M_1: Fit the model

   $$\mathrm{E}[Y_{ij}] = \alpha_j + \beta_j(age_{ij} - 20),$$

   $i = 1, \ldots, K, j = 1, \ldots, 5$, assuming that $\mathrm{Cov}(Y_i)$ is the same for all subjects, but otherwise has no special structure. Fit this model, present parameter and standard error estimates and comment on them. Interpret the parameters that appear in the mean structure.

   C) In the context of M_1, test the hypothesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_5$ against its complement. Report and interpret your findings.

   D) In the context of M_1, test the hypothesis $H_0 : \beta_j = 0, j = 1, \ldots, 5$ against its complement. Report and interpret your findings.

   E) In the context of M_1 and assuming that $\beta_1 = \cdots = \beta_5$, test the hypothesis $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$. Report and intepret your findings.

   F) In the context of M_1 and assuming that $\beta_1 = \cdots = \beta_5$, test the hypothesis $H_0 : \alpha_1 = ... = \alpha_5$ against its complement. Report and interpret your findings.

   G) Model M_2: Fit the model

   $$\mathrm{E}[Y_{ij}] = \alpha + \beta(age_{ij} - 20),$$

$i = 1, \ldots, K, j = 1, \ldots, 5$, assuming that $\mathrm{Cov}(Y_i)$ is the same for all subjects, but otherwise has no special structure. Fit this model, then calculate estimates of and 95% confidence intervals for the probability of a low-birth weight for the first birth if a woman's age at that birth was 15, 20, 25, 30, 35 or 40 years. Describe your methods and present the results in a table. Comment.

H) Fit the model M_3:

$$\Phi^{-1}(\mathrm{E}[Y^*_{ij}]) = \theta + \psi(age_{ij} - 20),$$

$i = 1, \ldots, K, j = 1, \ldots, 5$, assuming that $\mathrm{Cov}(Y^*_i)$ is the same for all subjects, but otherwise has no special structure. Fit this model, then explore possible theoretical connections between the mean structures in models M_2 and M_3, and how far these connections are borne out in the analysis results (the estimates).

I) Based on the estimates from M_3, calculate and present estimates and confidence intervals as in part G.

J) Fit the model M_4:

$$\Phi^{-1}(\mathrm{E}[Y^*_{ij}|U_i]) = \gamma + \delta(age_{ij} - 20) + U_i,$$

$i = 1, \ldots, K, j = 1, \ldots, 5$, where $U_1, \ldots, U_K$ are unobserved random variables distributed as iid normal with mean 0 and variance σ^2_u. Further, assume that $Y^*_{i1}, \ldots, Y^*_{i5}$ are conditionally independent given U_i. Explore possible theoretical connections between the mean structures in models M_3 and M_4, and how far these connections are borne out in the analysis results.

K) Based on the estimates from M_4, present a "between and within" decomposition of the variance of Y^*_{i3} assuming $age_{i3} = 25$, and compute the intra-class correlation based on this decomposition.

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

Points: A 2, B 2, C 1, D 1, E 1, F 1, G 4, H 3, I 2, J 3, K 2, L 3

2017 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