Interdisciplinary PhD Program in Biostatistics

Qualifying Exam II

Day 1: Methods and Applications

Monday June 5, 2017, 1-5pm

1. Write the question number in the upper left-hand corner and your exam ID code in the right-hand corner of each page you turn in.
2. Do **NOT** put your name on any of your answer sheets.
3. Start each problem on a separate sheet of paper.
4. There are 4 questions, each worth 25 points, for a total of 100 points. Answer each question as completely as you can. Be sure to show your work and justify your answers.

1. (25 points) Let $X_1, X_2, \ldots, X_n$ be $i.i.d.$ with pdf

$$f(x|\theta) = \begin{cases} 1/\theta, & k\theta \leq x \leq (k+1)\theta \\ \\ 0, & \text{otherwise} \end{cases}$$

where $\theta > 0$ and k is a known, positive constant.

(a) (3 points) Let $\overline{X}_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$. Show that

$$T_n = \frac{2}{2k+1}\overline{X}_n$$

is the method of moments estimator of θ.

(b) (4 points) Show that T_n is a consistent estimator of θ.

(c) (4 points) Show that the maximum likelihood estimator (MLE) of θ is $\hat{\theta}_n = \frac{1}{k+1}X_{(n)}$, where $X_{(n)}$ is the n^{th} order statistic of the random sample $X_1, \ldots, X_n$.

(d) (9 points) Show that $\hat{\theta}_n$ is a biased and consistent estimator of θ.

(e) (5 points) Obtain the limiting distribution of $Y_n = n(\theta - \hat{\theta}_n)$ as $n \to \infty$.

2. (25 points) Metal pipes used by gas, water, power and communication companies are buried underground, and it is desirable to apply coating to these pipes to retard corrosion. In a study, effects of **four different coatings** for pipes that will be buried in **three types of soil** were investigated. The experiment was carried out by first selecting 12 pipe segments and applying each coating to three segments. The segments were then buried in soil for a specified time period in such a way that each soil type received one piece with each coating. The depth of corrosion is measured to see how fast pipes become rusty. Researchers decided to use the following two-way fixed effects ANOVA model to understand these effects:

$$Y_{ij} = \mu + \alpha_i + \tau_j + \epsilon_{ij}, \tag{1}$$

for $i = 1, 2, 3$ (soil type), $j = 1, 2, 3, 4$ (coating type) where ϵ_{ij} is a set of independent random variables following a $N(0, \sigma^2)$ distribution. Answer questions (a)−(e) using this model.

a) (4 points) Write down the statistical model (1) in matrix-vector form, defining all your quantities including design matrix $\boldsymbol{X}$, $\boldsymbol{Y} = \{Y_{ij}\}$, and $\boldsymbol{\beta} = [\mu, \alpha_1, \alpha_2, \alpha_3, \tau_1, \tau_2, \tau_3, \tau_4]'$. What is the rank of the design matrix? Do the normal equations have a unique solution?

b) (4 points) If G_1 and G_2 are two g-inverses (generalized inverses) for $\boldsymbol{X'X}$, show that fitted values corresponding to the two estimators $\tilde{\boldsymbol{\beta}}_{\mathbf{1}} = \boldsymbol{G_1X'Y}$ and $\tilde{\boldsymbol{\beta}}_{\mathbf{2}} = \boldsymbol{G_2X'Y}$ are equal.

c) (5 points) Show that if $\boldsymbol{c}'\boldsymbol{\beta}$ is estimable then $\boldsymbol{c}$ is orthogonal to the null space of $\boldsymbol{X}$, that is $c_1 = c_2 + c_3 + c_4 = c_5 + c_6 + c_7 + c_8$ where $\boldsymbol{c} = [c_1, c_2, \ldots, c_8]'$.

Hint: $\{\boldsymbol{\lambda_1}, \boldsymbol{\lambda_2}\}$ *where* $\boldsymbol{\lambda_1} = [1, -1, -1, -1, 0, 0, 0, 0]'$ *and* $\boldsymbol{\lambda_2} = [1, 0, 0, 0, -1, -1, -1, -1]'$ *is a basis for the null space of* $\boldsymbol{X}$ *denoted by* $N(X)$ *where* $N(X) = \{\boldsymbol{z} \in \mathbb{R}^8 : \boldsymbol{Xz} = \mathbf{0}\}$.

d) (6 points) Consider the null hypothesis $H_0 : \alpha_1 = \alpha_2 = \alpha_3$. Show that this hypothesis is testable. Assume that σ is known and clearly state the distribution of the test statistic under H_0.

Hint: The covariance matrix of $\boldsymbol{A}'\tilde{\boldsymbol{\beta}}$ *is* $\sigma^2 \begin{bmatrix} 1/2 & -1/4 \\ -1/4 & 1/2 \end{bmatrix}$ *where* $\tilde{\beta}$ *is the least squares estimator of* $\boldsymbol{\beta}$ *and* $H_0 : \boldsymbol{A}'\boldsymbol{\beta} = \mathbf{0}$. *You need to show all steps and use this calculated matrix at the end of the solution.*

e) (6 points) Although model (1) is the true model to generate depth corruptions, researchers chose to disregard the coating effect, and use the incorrect model

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}. \tag{2}$$

Is the least squares estimator under model (2) unbiased for estimating σ^2? If yes, prove it is unbiased; otherwise provide an expression for the bias.

* The pipe corruption data description is from “Biostatistical Analysis (4th Edition)” by Jerrold H. Zar.

3. (25 points) In this question, we will consider inference of the survival distribution for *current status data*, which are data subject to an extreme form of censoring. Consider a random sample of n individuals, who have unobserved event times (assumed to be $i.i.d.$) denoted by $T_i, i = 1, 2, \ldots n$. For each individual, the data consist of an observation time C_i, assumed to be independent of T_i, and knowledge of whether or not the event has occurred by time C_i, which we denote by $\delta_i = I(T_i < C_i)$. Thus, note that the observed data for individual i are $(C_i, \delta_i), i = 1, 2, \ldots, n$.

   (a) (3 points) Assume that T_i has density f_α and cdf F_α for some parameter α, and that C_i has density g and cdf G, independent of α. Write an expression for the likelihood of α.

   (b) Suppose that the event time distribution is exponential, i.e., $f_\alpha(t) = \alpha e^{-\alpha t}$, and that the same observation time C (assumed to be known) is used for all subjects.

      (i) (8 points) What is the relative efficiency (i.e., the ratio of the expected Fisher information) for estimation of α using the current status data compared to using the unavailable data consisting of complete observation of T_i?

      (ii) (2 points) Comment on how the observation time C could be selected in consideration of your answer to part (i).

   (c) Now consider a competing risks framework in which there are two possible independent causes of failure. Let α_j be the cause-specific hazard for cause j, $j = 1, 2$, and let $F_j(t) = P(T < t, J = j)$, where T is the time to failure and J is the cause of failure. Note that the overall survival function is $S(t) = 1 - F_1(t) - F_2(t)$. The data are a random sample of n individuals for which we observe $Y_i = (C_i, \Delta_i, \Phi_i)$, where C_i is the observation time, $\Delta_i = I(T_i < C_i, J = 1)$, and $\Phi_i = I(T_i < C_i, J = 2)$. In other words, at the observation time, we know whether or not each individual has failed as well as the cause of failure for individuals that have failed, but we do not know the specific failure time. We assume that C_i is independent of T_i and is uninformative about the cause of failure.

      (i) (4 points) Write an expression for the likelihood of α_1 and α_2.

   Now suppose that the distributions of the time to event due to cause 1 and cause 2 are exponential with parameters α_1 and α_2, respectively.

      (ii) (4 points) Find $F_j(t), j = 1, 2$.

      (iii) (4 points) Reparametrize your likelihood from (i) of part (c) under this exponential model by defining $p = \frac{\alpha_1}{\alpha_1 + \alpha_2}$ and $\alpha = \alpha_1 + \alpha_2$. Show that the MLE of α is the same as the MLE of an exponential hazard based on current status data given by the monitoring times C_i and the observation of whether any failure, of either type, has occurred by time C_i, e.g., the MLE for the likelihood in part (a).

4. (25 points) A study was conducted by researchers at the UCLA Fielding School of Public Health to examine "self pollution" in school buses resulting from a bus' own diesel emissions penetrating its cabin through cracks, doors, and windows. As part of the study, researchers measured counts of ultrafine particles (UFP), small airborne particles with diameters less than 100 nm that are known to be toxic to humans, inside a school bus for 20 minutes after the bus' engine had been turned on. These UFP counts (particles/cm^3) were collected using a Scanning Mobility Particle Sizer at a fixed location inside a single school bus parked in a remote location at $K = 6$ time steps during $M = 10$ separate "runs."

   Each run, denoted by i, began with the researcher turning the engine on and recording the UFP count, y_{ijk}, in size category j at each time step, k.

   Measurements were taken in two size categories: S25 corresponds to particles with diameters between 12.5 and 37.5nm and S50 corresponds to particles with diameters between 37.5 and 62.5nm. Let s_j be an indicator variable such that

$$s_j = \begin{cases} 0 & \text{if measurement } y_{ijk} \text{ is for size category S25} \\ 1 & \text{if measurement } y_{ijk} \text{ is for size category S50.} \end{cases}$$

   In addition, let t_k denote the time (in minutes) since the engine was turned on when observation y_{ijk} was collected, where $t_1 = 1$, $t_2 = 2$, $t_3 = 5$, $t_4 = 10$, $t_5 = 15$, and $t_6 = 20$.

   Half of the ten runs were collected with all windows closed and half were collected with the back four windows on each side open 20cm. The window position was recorded by the variable x_i, where

$$x_i = \begin{cases} 0 & \text{if the windows were closed on the } i\text{th run} \\ 1 & \text{if the windows were open on the } i\text{th run.} \end{cases}$$

   Let $\{\mathbf{y}_{ijk} : i = 1, \ldots, M; j = 1, \ldots, J; k = 1, \ldots, K\}$. Consider the following Bayesian hierarchical model for the data:

$$p(\mathbf{y}|\beta_0, \beta_1, \beta_2, \gamma, \boldsymbol{\alpha}, \sigma^2) = \prod_{i=1}^{M} \prod_{j=1}^{J} \prod_{k=1}^{K} p(y_{ijk}|\beta_0, \beta_1, \beta_2, \gamma, \alpha_i, \sigma^2),$$

   where $y_{ijk}|\beta_0, \beta_1, \beta_2, \gamma, \alpha_i, \sigma^2 \sim \mathrm{N}(\beta_0 + \beta_1 t_k + \beta_2 t_k^2 + \gamma s_j + \alpha_i, \sigma^2)$;

$$p(\beta_0, \beta_1, \beta_2, \gamma, \boldsymbol{\alpha}, \sigma^2|\delta, \sigma_\alpha^2) = p(\beta_0)p(\beta_1)p(\beta_2) \left[\prod_{i=1}^{M} p(\alpha_i|\delta, \sigma_\alpha^2)\right] p(\sigma^2),$$

   where $\beta_0 \sim \mathrm{N}(0, c^2)$, $\beta_1 \sim \mathrm{N}(0, c^2)$, $\beta_2 \sim \mathrm{N}(0, c^2)$,

$$\alpha_i|\delta, \sigma_\alpha^2 \sim \mathrm{N}(\delta x_i, \sigma_\alpha^2) \text{ for } i = 1, \ldots, M,$$

   and $\sigma^2 \sim IG(a, b)$; and

$$p(\delta, \sigma_\alpha^2) = p(\delta)p(\sigma_\alpha^2),$$

where $\delta \sim \mathrm{N}(0, c^2)$ and $\sigma_\alpha^2 \sim IG(a, b)$. $a > 0$, $b > 0$, and $c^2 > 0$ are known constants, $\mathrm{N}(m, c^2)$ denotes the normal distribution with mean m and variance c^2, and $\mathrm{IG}(a, b)$ denotes the inverse gamma distribution with density function

$$p(z) = \frac{b^a}{\gamma(a)} z^{-(a+1)} \exp(-b/z).$$

a) **[6 points]** Provide interpretations for the following model parameters in the context of the problem: β_0, α_3, δ, and σ_α^2.

b) **[8 points]** Derive the full conditional posterior distributions of γ and σ^2, $p(\gamma|\beta_0, \beta_1, \beta_2, \boldsymbol{\alpha}, \sigma^2, \delta, \sigma_\alpha^2, \mathbf{y})$ and $p(\sigma^2|\beta_0, \beta_1, \beta_2, \boldsymbol{\alpha}, \gamma, \delta, \sigma_\alpha^2, \mathbf{y})$, respectively. If possible, identify these distributions by name and provide expressions for their parameters.

c) **[4 points]** The table below lists the posterior means of several model parameters. Using these values, sketch the posterior expected UFP count as a *continuous* function of time since the engine was turned on (from 1 to 20 minutes) for all combinations of size category (S25 and S50) and window position (closed and open). (Do not simply approximate these functions by evaluating it at t_1, $\ldots$, t_6.) Be sure that your plot is detailed enough so that the relative differences in the expected counts across size category and window position are clear.

parameter	posterior mean
β_0	100
β_1	20
β_2	-0.5
γ	75
δ	500

d) **[3 points]** A reviewer for a manuscript describing the results of the study offers the following criticism of the Bayesian hierarchical model:

> *The model assumes that UFP counts are independent across the runs, which is unreasonable. Since runs were performed using the same school bus, which was parked in the same location, it does not make sense to assume the UFP counts are independent across runs.*

Provide a response to this critique indicating why you agree or disagree with it.

e) **[4 points]** Assume that the model was fitted using an MCMC algorithm and you have samples from the joint posterior distribution of all model parameters. Using these samples, describe how to approximate the posterior probability that the observed UFP count in size category S25 exceeded 800 particles/cm^3 25 minutes after the engine started during the 7th run. Note that $x_7 = 1$ (windows were open). Provide a reason why extrapolating outside the observation window for the study (1 to 20 minutes), as you are doing here, may be particularly inappropriate with this model.

Interdisciplinary PhD Program in Biostatistics

Qualifying Exam II

Day 2: Data Analysis #1

Tuesday June 6, 2017, 9am-1pm

1. This part contains one data analysis project, worth a total of 50 points. Submit a final report for the project with your exam ID code on the title page. Your final report should be one, self-contained document. Follow the project instructions to prepare your answers.
2. Do **NOT** put your name on any page of your report. Please only put your exam ID code on the report.
3. This part is open book and you are allowed to bring up to 10 books and unlimited class notes as references. However, **you may not access the Internet during the exam except if needed to download software add-ons (e.g., R packages, Stata files)**.
4. Computer Login Information:

   Username: bioexam
   Password: testtaker1! (case-sensitive)
5. The dataset is saved as read-only on the qualifier exam drive, located under My Computer in the T: Drive. You should copy it to the desktop of your computer before you start working on it.
6. At the end of exam period, save an electronic copy of your report (in **one** file) on the desktop of the computer with the file name: **Day2_*examID***, where *examID* is your assigned ID code.

Kramerica Pharmaceuticals conducted a randomized clinical trial to evaluate a drug treatment for patients with congestive heart failure in sinus rhythm. The statistician who helped design the study left Kramerica for a government position and is no longer able to analyze the study data. Thus, the investigators have contacted you for assistance. The primary goal of the trial is to determine if the drug improves overall survival (i.e., time to death due to any cause). The secondary goal of their study is to identify factors that impact the effect of the drug on overall survival. The data for your analysis are provided in the Excel file "Kramerica.xls." The variable definitions are as follows:

Variable	**Definition**
ID	Patient ID
TRTMT	0=Placebo, 1=Drug
AGE	Age at randomization to treatment (yrs)
RACE	1=White, 2=Nonwhite
SEX	1=Male, 2=Female
CHESTX	Chest X-ray (CT-Ratio) at randomization*
BMI	Body Mass Index (kg/m^2) at randomization
HYPERTEN	History of Hypertension
SYSBP	Systolic Blood Pressure (mmHg) at randomization
DEATH	Vital status at end of follow-up (1=Dead, 0=Alive)
DEATHDAY	Days since randomization to either last follow-up or death
REASON	Cause of Death (0=Alive, 1=Heart Failure, 2=Other Cause)

*A ratio < 0.5 is considered normal.

Perform the following analyses for Kramerica:

1. (8 points) Generate a table containing key descriptive statistics of study variables by treatment group. Label this table "Table 1." Don't include survival time in this table. Briefly describe any differences you observe across treatment groups.

2. (10 points) Provide a graphical display of overall survival by treatment group (call it Figure 1) and a table containing key descriptive statistics for overall survival (call it Table 2). Briefly describe any differences you observe across treatment groups.

3. (10 points) Perform a hypothesis test to determine if overall survival differs by treatment group (i.e., the primary analysis for the study). Briefly explain the method you used and why, using the results from part 1 to support your decision. If applicable, assess the assumptions of the method you used and make any necessary modifications to the data or model (if you used a model) to account for assumption violations. Then, clearly state your conclusion at the 0.05 significance level using language that would be understandable to a clinician (i.e., a non-statistician).

4. (12 points) Perform statistical tests to determine if the effect of the drug varies with patient demographics (age, race, sex) or any of the clinical measurements obtained at the time of randomization (chest x-ray, BMI, hypertension). Explain the methods you used

to perform these analyses including proper assessments of model assumptions. If you noticed an assumption violation, mention it and explain how you accounted for the violation. If you want to include any plots to support your assessment, include them in an appendix at the end of your exam. After explaining your analysis methods, provide a table containing p-values from each of your tests. Apply a correction for multiple testing.

5. (5 points) Identify the smallest corrected p-value from part 4. Explain, using language that a clinician would understand, how the effect of the drug differs according to the levels of that factor. Perform the interpretation even if the corrected p-value is insignificant (i.e., your interpretation shouldn’t be that the effect of the drug doesn’t differ by level of that factor since $p > 0.05$).

6. (5 points) The investigators would also like to perform a descriptive analysis comparing cause-specific mortality rates across treatment groups. Provide one or two figures for this descriptive analysis. Explain, in practical terms understandable to a clinician, what exactly it is that you are plotting and comment on the differences across treatment groups.

Interdisciplinary PhD Program in Biostatistics

Qualifying Exam II

Day 3: Data Analysis #2

Wednesday June 7, 2017, 9am-1pm

1. This part contains one data analysis project, worth a total of 50 points. Submit a final report for the project with your exam ID code on the title page. Your final report should be one, self-contained document. Follow the project instructions to prepare your answers.
2. Do **NOT** put your name on any page of your report. Please only put your exam ID code on the report.
3. This part is open book and you are allowed to bring up to 10 books and unlimited class notes as references. However, **you may not access the Internet during the exam except if needed to download software add-ons (e.g., R packages, Stata files)**.
4. Computer Login Information:

   Username: bioexam
   Password: testtaker1! (case-sensitive)
5. The dataset is saved as read-only on the qualifier exam drive, located under My Computer in the T: Drive. You should copy it to the desktop of your computer before you start working on it.
6. At the end of exam period, save an electronic copy of your report (in **one file**) on the desktop of the computer with the file name: **Day3_*examID***, where *examID* is your assigned ID code.

Question Part 1 (40 points total)

Researchers have developed new questionnaire-based instruments to measure sleep disturbance and sleep-related impairment. The first ("Disturbance") aims to measure the quality of one's sleep. The second ("Impairment") aims to measure one's sleepiness during the day. Both instruments result in a score between 8 and 40.

To examine the properties of these tests, the researchers asked a polling company to select a sample of individuals that can be thought of as a simple random sample (SRS) of 1897 individuals from the United States population. [Note the dataset contains additional individuals as described in Part 2 and indicated by the variable `SRS`=0.] The researchers administered a series of questionnaires to each of these individuals, including questions about demographics and health history, and the two new instruments designed to measure sleep-related health. The data for each participant contain the summary score for each of these instruments. Note that among the simple random sample, 701 self-identified as having been diagnosed with a sleep disorder.

Your task is to use the accompanying data to answer the following questions for the researchers. Present your answer to each as one to three paragraphs intended for a non-statistical audience. (Although you may use statistical language and mathematical equations, you should also explain your answers in plain language that the researchers could understand.) If you make figures or tables, be sure that you directly refer to them from the text of your response. Do not include figures, tables, or results that you do not reference in your answer.

1. (6 points) Describe the distribution of the new instrument scores (Disturbance and Impairment) in the United States population.
2. (16 points) The researchers would like for their new instrument scores to have the following two properties:

   - The score should be associated with diagnosis.
   - The magnitude of the score's association with diagnosis should not depend on demographic variables.

   For example, the score distributions should differ for those who do and do not have a sleep disorder diagnosis, but this difference should not depend on gender. Does the Impairment score have these two properties?

3. The researchers would like to use the new instrument scores (Disturbance and Impairment) to predict sleep disorder diagnoses, and thus would like you to create a predictive model. Because scores are often adjusted for age and gender, consider predictive models that include these two variables in addition to a single instrument score. Describe your predictive model, including the following considerations:

   (a) (6 points) Which single instrument score would be best to use in the predictive model?

(b) (6 points) Consider a person with a ‘typical’ demographic profile, based on the SRS data. Based on the predictive model you chose in part (a), what value of your chosen instrument score results in the largest probability of diagnosis? Calculate a confidence interval for the probability of diagnosis for a ‘typical’ person with this score.

(c) (6 points) Would you recommend that doctors consider gender and/or age when making diagnoses using the instrument score? If so, how?

Question Part 2 (10 points)

Do a random sample of people who self-report a sleep disorder diagnosis have Disturbance instrument scores that are substantially different from those of patients treated in sleep disorder clinics? To answer this question, the researchers measured the same variables on 234 patients being treated at one of 8 sleep disorder clinics across the country. Note that each clinic may specialize in different types of sleep disorders, and patients with different disorders may tend to have different instrument scores.

Do these data suggest any significant differences in Disturbance instrument score between the clinic population and the general population who report a sleep disorder diagnosis?

Present your answer as one or two paragraphs intended for a statistical audience (i.e., you may use statistical language and/or mathematical equations). Within your paragraph, refer to any separate figures, tables, or analysis results that you find helpful, but do not include figures, tables, or results that you do not reference in your answer.

Data

Data for this question is stored in two file formats:

- `sdat.Rdata`, which is an R object
- `sdat.csv`, which is a .csv file

The data files each contain the variables described in the table below.

Variable	Description
Age	Age, in years
Female	Indicator of female sex
Hispanic	Indicator of Hispanic ethnicity
Race	Race category, including White, Black, and Other
Income	Household income category, in thousands of dollars: <20, 20-49, 50-99, >100
BMI	Body Mass Index, in kg/m^2
Disturbance	Disturbance instrument score
Impairment	Impairment instrument score
SRS	Indicator of inclusion in the SRS data collection
Diagnosis	Indicator of sleep disorder diagnosis - either self-reported, or due to treatment at a sleep disorder clinic
Clinic	Identifier of the clinic (1-8) at which the patient is being treated; NA for SRS participants.