Interdisciplinary PhD Program in Biostatistics
Public Health Specialization

Qualifying Exam II

Day 1: Methods and Applications

Monday June 6, 2016, 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. Let $X_1, X_2, \cdots, X_n \overset{iid}{\sim} Normal(\mu, 1)$. This question concerns inference for μ, which could be any real number.

   (a) (4 pts) Write down the likelihood function to make inference about μ and state the maximum likelihood estimator (MLE) for μ (no calculation needed).

   (b) (7 pts) Suppose we need to further restrict the parameter space to $\mu \in \{\cdots, -2, -1, 0, 1, 2, \cdots\}$, i.e., μ can only take on integer values. Find the MLE $\hat{\mu}$ for μ.

   (c) (7 pts) Identify the distribution of $\hat{\mu}$.

   (d) (7 pts) Show that $\hat{\mu}$ is unbiased.

2. Consider a variant of the susceptible-infective-susceptible (SIS) epidemic model in which any infected individual become again susceptible immediately after his/her infectious period is over (for instance, think about a common cold epidemic). Let us denote by $Y(t)$ the number of infectious individuals at t. Assume $Y(0) = 1$ and that the number of susceptibles at $t = 0$ is n. Under the usual contact homogeneity assumption (an infectious individual is equally likely to have contact with any susceptible one) $Y = \{Y(t); t \geq 0\}$ may be modeled as a continuous time random walk process with the state space $\{0, \ldots, n+1\}$ and the transition rates

from	to	at rate
i	$i+1$	$\lambda i(n-i)/n$
i	$i-1$	γi

   (a) (8 pts) For large n, what is the probability of no further infection (that is, the infected individual recovers before infecting anybody else)? How long on average does one need to wait for this recovery?

   (b) (8 pts) Calculate the extinction probability for Y using branching approximation [HINT: assume that early on $(n-i)/n \approx 1$].

   (c) (9 pts) Assume additionally that $\gamma = 0$. This corresponds to the so-called SI model or the "gossip spread model". In a population of n individuals a rumor is started at $t = 0$ by a single individual and circulates throughout at the above rate until everybody knows it. Assuming that n is large, show that the average amount of time ET_n it takes for the gossip to circulate is asymptotic to $c_n = \frac{2}{\lambda} \log n$, or more precisely that $\lim_{n\to\infty} ET_n/c_n = 1$.

3. A study of an experimental drug was conducted in a sample of n mice. Mice were treated with the drug or placebo at the start of the study and followed until J mice died. Of primary interest was the effect of the drug on the mortality hazard which was modeled using the following proportional hazards model:

$$\alpha(t|X_i) = \sum_{j=1}^{J} \alpha_{0j} e^{\beta X_i} I(t_{j-1} < t \leq t_j)$$

where $X_i = 1$ if mouse i was treated with the drug (0 otherwise), $\alpha_{01}, \ldots, \alpha_{0J}$ are positive constants, and $t_1 < t_2 < \cdots < t_J$ are the unique death times in the data set with $t_0 = 0$. The data for each mouse consist of $(X_i, \widetilde{T}_i, D_i)$ $(i = 1, \ldots, n)$, were $\widetilde{T}_i = \min(T_i, t_J)$, T_i = time of death (in years), and $D_i = I(\widetilde{T}_i = T_i)$. Assume no tied death times.

(a) (5 pts) Show that the log-likelihood equals

$$l(\boldsymbol{\alpha}, \beta) = \sum_{j=1}^{J} \left[\log(\alpha_{0j}) + \beta X_{i_j} - \alpha_{0j}(t_j - t_{j-1}) \sum_{l \in R_j} e^{\beta X_l} \right]$$

where X_{i_j} denotes the treatment indicator of the mouse that died at t_j and $R_j = \{l : \widetilde{T}_l > t_{j-1}\}$.

(b) (3 pts) Find $\widehat{\alpha}_{0j}$, the maximum likelihood estimate of α_{0j} given β $(j = 1, \ldots, J)$.

(c) (4 pts) Let $\widehat{\boldsymbol{\alpha}} = (\widehat{\alpha}_{01}, \widehat{\alpha}_{02}, \ldots, \widehat{\alpha}_{0J})'$. Provide the expression that would be used to find the value of β that maximizes the log-profile likelihood, $l(\widehat{\boldsymbol{\alpha}}, \beta)$. Simplify your expression as much as possible.

(d) (6 pts) Suppose the investigators of the study decided to not use a piece-wise constant baseline hazard function and instead chose to leave the baseline hazard unspecified and use Cox's partial likelihood function to estimate β. Derive the expression that would be solved to find the maximum partial likelihood estimate $\hat{\beta}$ and compare it to the expression from part c.

(e) (7 pts) The investigators decided to stick with the piece-wise constant hazard model, but added a mouse-specific frailty Z_i to account for heterogeneity in the hazard function not explained by treatment group. Their new model was as follows:

$$\alpha(t|X_i, Z_i) = \sum_{j=1}^{J} \alpha_{0j} Z_i e^{\beta X_i} I(t_{j-1} < t \leq t_j)$$

where Z_i is distributed gamma with pdf

$$g(z_i) = \frac{z_i^{1/\delta - 1}}{\delta^{1/\delta} \Gamma(1/\delta)} e^{-z_i/\delta}.$$

Derive a marginal likelihood, not conditional on $Z_1, \ldots, Z_n$, that could be used to perform inference on β.

4. Suppose that $X_1, \ldots, X_n$ are iid samples from a continuous distribution with pdf given by

$$f(x) = \frac{1}{b} \exp\left\{-\frac{x-a}{b}\right\} I\{x \geq a\},$$

where $-\infty \leq a \leq \infty$ and $b > 0$ are unknown constants, and $I(\cdot)$ is the indicator function.

(a) (5 pts) Let $Y_1 \leq \ldots \leq Y_n$ be the order statistics of $X_1, \ldots, X_n$. Find the joint distribution of $Y_1, \ldots, Y_n$.

(b) (6 pts) Let $Y_0 = 0$. $W_i = Y_i - Y_{i-1}$ for $1 \leq i \leq n$ are the *spacings*. Prove that W_i's are distributed independently, and show that the marginal pdfs are given by:

$$\frac{1}{b/n} \exp\left\{-\frac{w_1 - a}{b/n}\right\} \qquad \text{for } W_1$$

$$\frac{1}{b/(n-i+1)} \exp\left\{-\frac{w_i}{b/(n-i+1)}\right\} \qquad \text{for } W_i, 2 \leq i \leq n$$

(c) (7 pts) Find the MLEs of a and b.

(d) (7 pts) Denote $\hat{a}$ and $\hat{b}$ as the MLEs of a and b. Find the marginal distributions of $\hat{a}$ and $\hat{b}$. Are $\hat{a}$ and $\hat{b}$ independent? Why? If they are dependent, also find their joint distribution.

Interdisciplinary PhD Program in Biostatistics
Public Health Specialization

Qualifying Exam II

Day 2: Data Analysis #1

Tuesday June 7, 2016, 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.

A recent randomized controlled trial was conducted to compare two treatment regimens for behavior management for children with attention-deficit /hyperactivity disorder (ADHD). All of the children had severe aggression. The two treatments that were compared were:

Basic: parent training in behavior management + stimulant medication
Augmented: parent training in behavior management + stimulant medication + risperidone

The trial lasted 9 weeks, and the protocol was for children to come into the clinic for weekly visits. However, some children dropped out of the study and did not complete the full 9 weeks. The primary study goal was to determine whether adding the antipsychotic medication risperidone reduced aggressive behaviors. That analysis has already been completed. Instead, you will focus on evaluating the occurrence of adverse events in the two groups.

Risperidone is not without side effects. A major concern is the occurrence of adverse events (AEs) that are rated by a clinician as "severe" – some examples might be vomiting, headache, trouble falling asleep, etc. At each study visit, parents reported whether any AEs had occurred, and a clinician determined its severity (i.e., "severe" or "not severe"). You have been provided data on the severe AEs reported for each study participant during the course of the trial. To facilitate your data analyses, the data are provided both in "wide" form and in "long" form as described below:

"Wide" Data Set (q2_2016_day2_wide.csv)

Variable	**Description**
subnum	Unique child identifier
trt	Treatment group (0=Basic, 1=Augmented)
site	Clinical site (1,2,3,4)
age	Age of the child (at week 0)
gender	Sex of the child (1=male, 2=female)
vnumLastVisit	Last study visit: week number (e.g., 2=completed through week 2 of study)
daysLastVisit	Last study visit: days since treatment start *Note: a child's visits might not be exactly 7 days apart, hence a child who completed the entire study – 9 weeks – might not have his last visit exactly on day 9x7=63). Similarly a child who completed 4 visits might not have his last visit on exactly day 4x7=28.*
days1-days8	Days since treatment started when **1st through 8th** AE occurred *Note: No subject had more than 8 reported AEs*
ae1-ae8	Indicator for if an AE occurred – 1 if AE occurred, missing otherwise *Note: these variables take the value 1 if the corresponding "days" variable has a non-missing value, and are missing otherwise.*

"Long" Data Set (q2_2016_day2_AElong.csv)

Variable	**Description**
subnum	Unique child identifier
daysAE	Day on which an AE was reported (days since treatment start)

Note: If a child never had any AEs reported, they do not appear in this "long" data set.

For the purposes of this exam, you will focus on **three specific research questions**:

1. Is there a difference between the treatments regimens in the occurrence of at least one adverse event?
2. The study was conducted at four different sites (clinics). Since the AEs are rated by clinicians, there is concern that there may be a "site effect". Is there evidence of a "site effect" in the occurrence of at least one adverse event?
3. In addition to caring about whether a child experienced any AEs, we also care about the cumulative experience of AEs – i.e., did children in one treatment group experience more AEs across the whole study period than children in the other treatment group?

Your Assignment

You are asked to perform the appropriate data analysis to answer each of the three research questions above. Summarize your results in a 1-2 page report in lay terms that explains (a) the analysis methods you used, (b) the results that answer the questions, and (c) alternative strategies you considered and the strengths/weaknesses of your chosen approach. Key output from statistical software may be included in an appendix or as tables or figures that you refer to in your report.

Please clearly label the sections in your report corresponding to each research question.

Interdisciplinary PhD Program in Biostatistics
Public Health Specialization

Qualifying Exam II

Day 3: Data Analysis #2

Wednesday June 8, 2016, 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.

Background

Primary biliary cirrhosis is a serious disease of the liver in which the intrahepatic bile ducts become scarred and blocked. The damaged ducts impair the ability of the liver to excrete bile into the gastrointestinal tract. The disease affects women more than men, and is most often first diagnosed between the ages of 35 and 60.

A randomized trial was conducted to test whether the drug methotrexate might lead to better (less worse) liver function in affected patients. Toxicities affecting the lung function of patients taking methotrexate (at higher doses) have been previously reported, thus lung function was also a concern in this study. Participants were randomized to either receive the study drug or placebo, and then were followed for 5 years. They were evaluated at baseline (year 0) and then 1 year and 5 years post-randomization.

Liver Function Measures

- **Bilirubin** is a breakdown product of the body's natural clearance of old red blood cells. It is usually excreted in bile (and urine). Because the liver is responsible for making bile, high levels of bilirubin in the blood tend to indicate worse liver function.

- **Prothrombin time** is a standardized measure of how long it takes blood to clot. The liver makes proteins that are essential to normal blood clotting, therefore long prothrombin times indicate fewer of these proteins are present in the blood (due to poor liver function). Prothrombin time is reported as international normalized ratio (INR).

- **Splenomegaly** is an enlargement of the spleen. In diseased livers, scarring affects the flow of fluids into and around the liver, which can in turn result in fluid build-up and congestion in the spleen.

Lung Function Measures

- **DLCO** is a measure of how well carbon monoxide can diffuse from inhaled air into the blood, with low values of DLCO suggesting some degree of lung disease.

- **FVC** is a measure of how much air can be forcibly expelled after a full inhalation and is a measure of total lung volume. Low values of FVC indicate poor lung function.

For the purposes of this exam, you will focus on **five specific research questions**:

1. Does treatment with the study drug lead to better measures of liver function (bilirubin and/or prothrombin time) relative to placebo?
2. Is there any evidence that the effect of the study drug on bilirubin at 5 years post-randomization is different for older patients? Among those with more severe liver disease?
3. Of particular concern for potential lung function toxicities are DLCO values below 80% of that predicted for sex, age, and height. In the placebo group, is there a difference in the proportion of individuals with impaired DLCO (values below 80%) 5 years post-randomization compared to baseline?
4. Does treatment with the study drug lead to greater short-term (1 year) or long-term (5 year) risk of impaired DLCO post-randomization?
5. Is there evidence of a greater toxic effect of treatment among certain groups of patients?

The data you have are available are a subset of all the data available, but include measures at baseline (time=0), 1 year and 5 years after randomization.

Data Set (q2_2016_day3.csv)

Variable	Description
ptid	Patient identification number
tx	Treatment arm (0=Placebo, 1=Study drug)
time	Time of observation in years (0=baseline)
age	Age at randomization (years)
sex	Sex (0=Female, 1=Male)
height	Height (cm)
weight	Weight (kg)
durdis	Days between diagnosis with PBC and randomization
stage	Histologic stage of liver disease (1=low, 2, 3, 4=high, high is worse disease)
splen	Splenomegaly at baseline (0=Absent, 1=Present)
comply	Was patient was taking study drug at time of measurement? (0=No, 1=Yes)
bili	Total bilirubin (mg/dl)
alb	Albumin (g/dl)
ptinr	Prothrombin time (INR)
fvc	Forced vital capacity (liters)
fvcpred	Forced vital capacity (percent of that predicted for sex, age, height)
dlco	DLCO - diffused lung carbon monoxide- (ml/min/mmHg)
dlcopred	DLCO (percent of that predicted for sex, age, height)

Your Assignment

You are asked to perform the appropriate data analysis to answer each of the five research questions above. Summarize your results in a written report in lay terms that explains (a) the analysis methods you used, (b) the results that answer the questions, and (c) strengths and weaknesses of your chosen approach. Key output from statistical software may be included in an appendix or as tables or figures that you refer to in your report.

Please clearly label the sections in your report corresponding to each research question.