Interdisciplinary PhD Program in Biostatistics

Qualifying Exam II

Day 1: Methods and Applications

Monday June 4, 2018, 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.
5. At the end of the exam, place your answers to each question in four different envelopes. Write the question number on the front of each envelope. Do **NOT** write your exam ID code on the envelopes.

1. Consider the general linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{Y}$ is an n-dimensional vector, $\mathbf{X} = \{\mathbf{X}_1 \ \ \mathbf{X}_2\}$ is an $n \times p$ design matrix of rank r, not necessarily of full rank, $\boldsymbol{\beta}$ is a coefficient vector of length p, and ε is an n-dimensional vector with $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $Var(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$, for $\sigma^2 > 0$. Let $\hat{\boldsymbol{\beta}}$ denote a solution of the normal equations for the full model, i.e., $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}$. Let $\boldsymbol{\beta} = \binom{\boldsymbol{\beta}_1}{\boldsymbol{\beta}_2}$ and $\hat{\boldsymbol{\beta}} = \binom{\hat{\boldsymbol{\beta}}_1}{\hat{\boldsymbol{\beta}}_2}$, where $\boldsymbol{\beta}_i$ and $\hat{\boldsymbol{\beta}}_i$ are coefficient vectors corresponding to $\mathbf{X}_i$.

   (a) (8 points) Show that every solution of the equation $\mathbf{X}_1'\mathbf{X}_1\boldsymbol{\beta}_1^* = \mathbf{X}_1'\mathbf{Y}$ (i.e., the least squares estimator of $\boldsymbol{\beta}_1$ under the hypothesis $\boldsymbol{\beta}_2 = \mathbf{0}$) is the $\hat{\boldsymbol{\beta}}_1$ part of a solution of the full normal equations if and only if $\mathbf{X}_1'\mathbf{X}_2\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$.

   (b) (8 points) Given that $\mathbf{X}$ is full rank and $\boldsymbol{\beta}_1^* = \hat{\boldsymbol{\beta}}_1$, prove or disprove that $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$.

   (c) (9 points) Consider a solution of the full normal equations $\hat{\boldsymbol{\beta}} = \binom{\hat{\boldsymbol{\beta}}_1}{\hat{\boldsymbol{\beta}}_2}$ that satisfies $\mathbf{X}_1'\mathbf{X}_1\hat{\boldsymbol{\beta}}_1 = \mathbf{X}_1'\mathbf{Y}$. Must it be the case that $\hat{\boldsymbol{\beta}}_2$ satisfies $\mathbf{X}_2'\mathbf{X}_2\hat{\boldsymbol{\beta}}_2 = \mathbf{X}_2'\mathbf{Y}$? Justify your answer.

2. Let $X \sim N(0, \sigma^2)$ and $Y \sim N(0, \sigma^2)$ denote two independent normal random variables. Also, let h denote a mapping from Cartesian coordinates to Polar coordinates such that $(R, \Theta) = h(X, Y)$. You may use the following facts throughout this question. The mapping from Polar coordinates to Cartesian coordinates is given by $(x, y) = h^{-1}(r, \theta) = (r\cos(\theta), r\sin(\theta))$, where $r > 0$ and $\theta \in [0, 2\pi)$. Also, $\frac{d}{dx}\sin(x) = \cos(x)$ and $\frac{d}{dx}\cos(x) = -\sin(x)$.

   (a) (4 points) Derive the joint probability density function of (R, Θ).

   (b) (3 points) Derive the cumulative distribution function of R and its inverse.

   (c) (3 points) Based on your solution to parts (a) and (b), propose an algorithm to sample two independent normal random variables with common mean μ and common variance σ^2 using two independent uniform random variables on $(0, 1)$.

   From now on, assume that we have a random sample $R_1, \ldots, R_n$ from a Rayleigh distribution. The probability density function of the Rayleigh distribution is given by:

$$f(r|\theta) = \frac{r}{\theta}\exp\left\{-\frac{r^2}{2\theta}\right\}, \quad r > 0,\ \theta > 0. \tag{1}$$

   (d) (4 points) Find the maximum likelihood estimator (MLE) of θ. Call it $\hat{\theta}$.

   (e) (6 points) What is the asymptotic distribution of $\hat{\theta}$?

   (f) (5 points) Derive an approximate $100(1-\alpha)\%$ confidence interval for $g(\theta) = E(R) = \sqrt{\frac{\pi}{2}\theta}$ using estimator $T = \sqrt{\frac{\pi}{2}\hat{\theta}}$ of the function $g(\theta)$.

3. A study was conducted comparing survival of melanoma patients who had a Sentinal Lymph Node (SNL) biopsy to melanoma patients who did not have an SNL. Patients were followed from diagnosis until death or study completion. Survival times were only censored if the subject was alive at the end of the study. The investigators wanted to control for thickness of the melanoma in their analysis, so they created four groups based on the quartiles of thickness and considered the following stratified test statistic:

$$Z_1(\tau) = \sum_{s=1}^{4} \left[\int_0^{\tau} dN_{1s}(t) - \int_0^{\tau} \frac{Y_{1s}(t)}{Y_{\cdot s}(t)} dN_{\cdot s}(t) \right]$$

where

- τ is the maximum follow-up time in the study
- $Y_{js}(t)$ is the number of patients in biopsy group j ($j = 0$ if no SNL biopsy, $j = 1$ if SNL biopsy) and thickness group s at risk of death at time t (i.e., number event-free and censoring-free just prior to t)
- $N_{js}(t)$ is number of patients in biopsy group j and thickness group s who died at or before time t
- $Y_{\cdot s}(t) = Y_{0s}(t) + Y_{1s}(t)$
- $N_{\cdot s}(t) = N_{0s}(t) + N_{1s}(t)$

The above test statistic can be used to test the following null hypothesis:

$$H_0 : \alpha_{0s}(t) = \alpha_{1s}(t) \text{ for } t \in [0, \tau], s = 1, 2, 3, 4$$

where $\alpha_{js}(t)$ is the hazard function for biopsy group j and thickness group s.

(a) (7 points) Show that $Z_1(\tau)$ is a sum of mean zero martingales under H_0.

(b) (7 points) Derive the asymptotic variance of $\frac{1}{\sqrt{n}} Z_1(\tau)$ under H_0, where n is the total number of patients.

(c) (4 points) Consider the following stratified Cox model:

$$\alpha(t|x, \text{stratum } s) = \alpha_{0s}(t) e^{\beta x}$$

where $x = 1$ if a patient had an SNL biopsy, 0 otherwise. The partial likelihood for this model is

$$L_{\text{strat}}(\beta) = \prod_{s=1}^{4} \prod_{T_{sj}} \frac{\exp\left\{\beta x_{si_j}\right\}}{\sum_{sl \in \Re_{sj}} \exp\left\{\beta x_{sl}\right\}}$$

where

- $T_{s1}, T_{s2}, \ldots$ denote the unique death times in thickness group s
- x_{sl} is the x value of the l^{th} patient in thickness group s
- x_{si_j} denotes the x value of the patient in thickness group s who died at T_{sj} (assuming no tied death times)
- R_{sj} denotes the set of patients in thickness group s at risk of death at T_{sj}

Derive the score function $U(\beta)$.

(d) (3 points) Show that $U(0)$ is equivalent to $Z_1(\tau)$.

(e) (4 points) What are the primary advantages and disadvantages of stratification compared to regression adjustment for melanoma thickness?

4. When estimating the association between two binary variables, i.e. X (exposed/unexposed) and Y (success/failure), we could consider a paired design or an independent two group design. The inference based on the paired design is usually referred to as conditional and the inference based on the independent design is referred to as marginal. Consider the following 2×2 table from a paired design, where n exposed subjects are each matched with a single unexposed subject.

		Unexposed		
		Success	Failure	Total
Exposed	Success	a	b	a+b
	Failure	c	d	c+d
	Total	a+c	b+d	n

   (a) (3 points) Express the above table in the format of an independent two-group design, where the exposed and unexposed subjects are not matched, i.e. a table of exposed/unexposed by success/failure.

   (b) (6 points) Relative risk (RR) is one popular measure of the association between two binary variables. Based on the independent two group design table in subquestion (a), express the estimate of the RR in terms of (a, b, c, d). Assuming Binomial distributions for the counts of observed successful events in each exposure group, find the large sample variance of the RR estimate in the logarithmic scale.

   (c) (6 points) Based on the literature (Chen, 1996), the maximum likelihood estimate (MLE) of $\log(RR)$ from the paired design has the same form as the independent design, with estimated asymptotic variance $\widehat{var}(\log(\widehat{RR})) = \frac{(b+c)}{(a+b)(a+c)}$. Suppose the paired design takes advantage of a positive relationship between subjects being matched, i.e. $\widehat{cov}(p_1, p_2) > 0$, where p_1 is the proportion of successful events in exposed group and p_2 is the proportion of successful events in unexposed group. Compare the variance estimates under both designs and identify the one with the smaller variance estimate (show detailed math work for reasoning).

   Hint: First work out what $\widehat{cov}(p_1, p_2) > 0$ implies under the paired design.

   (d) (6 points) The odds ratio (OR) is another popular measure for the association between two binary variables. Based on the independent design, write down an appropriate generalized linear model to estimate the marginal OR. Also derive the MLE of the OR based on your model (represent the answer using (a, b, c, d) from your table in subquestion (a)).

   (e) (4 pts) Based on the literature (Agresti, 2002), the MLE of the conditional OR from the paired design is $\frac{b}{c}$. Aside from the estimated variances, discuss the advantages and disadvantages of the paired (conditional) and independent (marginal) designs.

Interdisciplinary PhD Program in Biostatistics

Qualifying Exam II

Day 2: Data Analysis #1

Tuesday June 5, 2018, 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.
3. This part is open book and you are allowed to bring up to 6 bounded books and unlimited class notes as references. You may also bring a dictionary to the exam which will not count toward your 6 total books. Photocopied chapters of books or articles are not allowed.
4. You may not access the Internet during the exam except if needed to download software add-ons (e.g., R packages, Stata files) and to access the exam's Carmen site.
5. The data sets needed to complete the exam are available under the "Day 2 (June 5)" module on the "Biostatistics QII Exam" Carmen page.
6. Submit your report by 1:00 pm using the "Day 2 Submission" drop box on Carmen.

Monoclonal gammopathy of undetermined significance (MGUS) occurs in about 2% of persons over age 50 and 3% of those over 70. It can be a precursor of multiple myeloma or other plasma cell malignancies (PCM). The `R` data set `mgus2` from the `survival` package covers 1384 patients in southeastern Minnesota diagnosed with MGUS between 1960 and 1994. For each patient, it has the following 10 variables:

id	subject identifier
age	age in years at MGUS diagnosis
sex	factor with levels F and M
hgb	hemoglobin at MGUS diagnosis
creat	creatinine at MGUS diagnosis
mspike	size of the monoclonal serum spike at MGUS diagnosis
ptime	months from diagnosis to PCM or censoring
pstat	occurrence of PCM (0 = no, 1 = yes)
futime	months from diagnosis until death or censoring
death	occurrence of death (0 = no, 1 = yes)

It can be loaded in `R` using `data(mgus2)`, and it is available as `mgus2.csv` for other statistical packages.

1. Summarize the distribution of age at MGUS diagnosis and investigate its relationship with the sex of the patient:

   (a) (5 points) Give a graphical comparison of the distributions of age at MGUS diagnosis for males and females, and generate a table showing the mean, standard deviation, and quartiles separately for males and females. Does there appear to be any difference?

   (b) (5 points) Test the null hypothesis that the distribution of age at diagnosis is the same for males and females. Briefly explain the test you used and why you chose it. State your conclusion at the 5% significance level in language that would be understandable to a clinician.

2. We would like to investigate the effect of sex on the risk of death after MGUS diagnosis.

   (a) (6 points) Provide a graphical display of the time from MGUS diagnosis to death for males and females, clearly indicating which curve belongs to which sex. Does there appear to be a difference? Briefly explain the methods you used, including any underlying assumptions.

   (b) (6 points) Conduct a nonparametric test of the null hypothesis that time from MGUS diagnosis to death is the same in males and females. Briefly explain the test you used and why you chose it. State your conclusion at the 5% significance level in language that would be understandable to a clinician. Is this conclusion consistent with the plot in your answer to Question 2(a)?

3. Finally, we will look at the joint effects of age, sex, and other baseline covariates on the risk of death after MGUS diagnosis.

   (a) (14 points) Develop a model using sex and age that predicts the risk of death in a patient newly diagnosed with MGUS. Explain your model building process, assess model fit, and give a brief summary of the final model. Using the model, give point estimates and 95% confidence intervals for the 10-year survival probabilities of newly-diagnosed 70-year-old male and female MGUS patients.

   (b) (10 points) Starting with your model from 3(a), determine whether hemoglobin, creatinine, or monoclonal serum spike improve predictions of survival following MGUS diagnosis. Briefly explain the methods you used, including assessment of model assumptions. State your conclusions at the 5% significance level, interpreting the effects of significant predictors in language that would be understandable to a clinician.

   (c) (4 points) Now fit two simple Cox models: One with sex and age (as a linear main effect) and one with sex only. What is the coefficient on `sex` in each model? Using results of Question 1 and the coefficient estimates from the model with sex and age, explain the direction and approximate magnitude of the change in the coefficient on sex when age is removed from the model.

References

[1] Robert A. Kyle *et al.* (2002). A long-term study of prognosis in monoclonal gammopathy of undetermined significance. *New England Journal of Medicine* 346(8): 564–569.

Interdisciplinary PhD Program in Biostatistics

Qualifying Exam II

Day 3: Data Analysis #2

Wednesday June 6, 2018, 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.
3. This part is open book and you are allowed to bring up to 6 bounded books and unlimited class notes as references. You may also bring a dictionary to the exam which will not count toward your 6 total books. Photocopied chapters of books or articles are not allowed.
4. You may not access the Internet during the exam except if needed to download software add-ons (e.g., R packages, Stata files) and to access the exam's Carmen site.
5. The data sets needed to complete the exam are available under the "Day 3 (June 6)" module on the "Biostatistics QII Exam" Carmen page.
6. Submit your report by 1:00 pm using the "Day 3 Submission" drop box on Carmen.

Data Source and Context

The data for this problem are a slightly altered version of the data that are used in this paper:

> Farouk Musa A, Dillion J, Mohd Taib ME et al. A study on the effect of Haruan fish extract (Channa striatus) on wound healing and quality of life of coronary artery bypass grafting (CABG) patients: A prospective, double-blind, randomized, controlled trial [version 1; referees: awaiting peer review]. F1000Research 2018, 7:469 (doi: 10.12688/f1000research.13372.1)

Reading the complete paper is unlikely to help you answer the questions in this exam. (Note that this paper has not yet been peer reviewed, and is likely to contain some analytic errors.)

Extracts from the abstract and body of this paper summarize the background and methods for the study:

> **Background:** Wound healing remains a primary problem in all surgical cases especially … when the length of incision is very significant as with cardiac bypass patients. The main objective of this study is therefore to assess the effect of Haruan fish extract (Channa striatus) on chest and leg wounds post-coronary artery bypass grafting (CABG) surgery with the optimum and standard patient care in two groups of randomized patients.
>
> **Methods:** This is a randomized, double blind clinical trial being conducted at the National Heart Institute, Kuala Lumpur. Two randomized groups of similar demographic and co-morbid histories planned for CABG were enrolled into the study. Both groups were blinded to the capsules being given to them … post-operatively [one type of capsule contained fish extract; the other type was a placebo]. Assessments were … [made] on the health-related quality of life (HRQOL) of patients using the Nottingham Health Profile (NHP).
>
> **Measurement:** … **Health related quality of life.** In this study, we decided to assess the health-related quality of life (HRQOL) using the Nottingham Health Profile (NHP) - Part 1. … There were 38 subjective statements which were divided into six sections on NHP Part 1 namely physical mobility, … , pain, and sleep. Each section will have a score range from between 0 - 100 [where 0 means good quality of life and 100 means poor quality of life].
>
> The questionnaire was distributed to both groups of patients … at six weeks [after hospital discharge].

Extra notes:

- Three different surgeons performed the CABG surgery for this study.

- We will focus on the pain HRQOL index measurement made 6 weeks after the participants were discharged from the hospital.

Data Description

The same data are stored in two file formats:

- “HaruanData.csv” (a comma-separated text file)
- “HaruanData.Rdat” (an R data object)

These data include the variables described in Table 1. The first 10 rows of the data are displayed in Table 2. Note that this data set does have a small number of missing values.

Table 1: Descriptions of the variables

id	Unique patient identifier
capsule	Treatment (E= fish extract, O= placebo)
surgeon	Initial of the surgeon’s last name
surgery.time	Length of the surgery (in minutes)
w6.p	Pain quality of life (HRQOL) index measured 6 weeks after discharge (scale of 0-100)

Table 2: The first 10 rows of the dataset

id	capsule	surgeon	surgery.time	w6.p
1	E	J	220	13
2	O	A	170	0
3	E	A	145	13
4	O	A	175	13
5	O	A	150	NA
6	O	J	120	29
7	E	E	210	6
8	O	J	195	0
9	E	J	225	24
10	O	A	NA	26

Advice

If you are using an MCMC sampler, the code may take some time to run. Use this time to work on your write-up of your answers. If you do not have sufficient time to run your sampler as you would like, run it for fewer iterations and note in your write-up how you think this might affect your answers.

Software

For your convenience, the JAGS documentation is available on the exam's Carmen page under the module "Day 3 (June 6)." However, you are not required to use JAGS and may use a different software package for Bayesian computation if available in the computer lab.

Questions

In general, the researchers are interested in the effect of fish extract on the presence of pain (i.e., a pain HRQOL index greater than zero) and the magnitude of that pain 6 weeks after a patient is discharged from the hospital after surgery.

1. [10 points] Read through parts 2-5. Use exploratory data analysis, possibly including appropriate plots and summary statistics, to describe the distribution of the 6-week pain HRQOL index and its relationship with other variables.

2. [15 points] Focus on the presence of pain. The pain HRQOL index has been used in many different studies in the past. Typically, approximately 30 to 60% of patients report a pain HRQOL index greater than zero in these other studies. Use this information to form a reasonable prior distribution. Use this prior distribution and the data to estimate the probability that a randomly selected patient from the **control group** of this study reports a positive pain HRQOL index, for a 'typical' surgeon. **Write down the complete model you used (including all prior distributions) and report a 95% credible interval for this probability.** Be sure your model accounts for correlation due to patients being treated by the same surgeon. If you use a numerical method to calculate your answer, **justify the accuracy of your answer** (e.g., via plots or other summaries).

3. [10 points] The researchers' goal is to estimate any difference in pain HRQOL index for the fish extract group versus the control group. With this goal in mind, use your exploratory data analysis from part 1 to extend your model from part 2 to include the group that took fish extract.

   - **Write down your new model, including all prior distributions.**
   - **Write down the effect of treatment (fish extract versus placebo) as a parameter or function(s) of a parameter(s) in your model.**
   - **Report a 95% credible interval for the treatment effect.**

4. [5 points] Reflect on your answers from part 1 to 3. A physician in the medical center is planning to visit Malaysia. Based on your results, would you recommend he treat CABG patients there with fish extract? Write two paragraphs summarizing your recommendation and the evidence you used to support it. Make sure your language can be understood by someone without statistical expertise.

5. [10 points] Now consider the magnitude of the pain HRQOL index only among those whose HRQOL pain index is greater than zero. Consider a new patient who we are sure will have pain (i.e., their pain HRQOL index will surely be >0), will be treated by Doctor A, and whose surgery will last 100 minutes. [Hint: In your calculations, this new patient would look as if they were an existing participant that has a missing value for their pain HRQOL index.] **Report two 50% posterior prediction intervals for this person's pain HRQOL index - one if they have placebo and the second if they have fish extract.** (Your model for prediction should include both surgeon and duration of surgery.) **Write down the complete model (including prior distributions) you used to find these intervals.**