Interdisciplinary PhD Program in Biostatistics

Qualifying Exam II

Day 1: Methods and Applications

4 hours

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.

1. Suppose that the time to death X has an exponential distribution with hazard rate λ (i.e., $f_X(x) = \lambda e^{-\lambda x}$) and that the right-censoring time C is exponential with hazard rate θ. Let $T = \min(X, C)$ and $\delta = 1$ if $X \leq C$; $\delta = 0$ if $X > C$. Assume that X and C are independent.

   (a) (5 pts) Find $P(\delta = 1)$.

   (b) (5 pts) Find the distribution of T.
   Hint: this should be a "named" distribution.

   (c) (5 pts) Show that δ and T are independent.

   (d) (5 pts) Let $(T_1, \delta_1), \cdots, (T_n, \delta_n)$ be a random sample from this model. Show that the maximum likelihood estimator of λ is $\hat{\lambda} = \sum_{i=1}^{n} \delta_i / \sum_{i=1}^{n} T_i$.

   (e) (5 pts) Use parts (a)-(c) to find the mean and variance of $\hat{\lambda}$.
   Hint: If $X \sim Gamma(\alpha, \beta)$, then $1/X \sim InverseGamma(\alpha, 1/\beta)$.

Some (possibly) useful distributions:

Distribution	PDF	$E(X)$	$V(X)$
$X \sim Exponential(\lambda)$	$\lambda e^{-\lambda x}$	$1/\lambda$	$1/\lambda^2$
$X \sim Gamma(\alpha, \beta)$	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$	α/β	α/β^2
$X \sim InverseGamma(\alpha, \beta)$	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\beta/x}$	$\beta/(\alpha-1)$	$\beta^2/[(\alpha-1)^2(\alpha-2)]$

2. Consider the model $y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad i = 1, 2, 3, \quad j = 1, 2, 3.$

   (a) (5 pts) Write $\mathbf{X}$, $\mathbf{X}'\mathbf{X}$, $\mathbf{X}'\mathbf{y}$, and the normal equations.

   (b) (2 pts) What is the rank of $\mathbf{X}$ (or $\mathbf{X}'\mathbf{X}$)?

   (c) (5 pts) Find a set of linearly independent estimable functions.

   (d) (3 pts) Is μ estimable? Explain why or why not.

   (e) (5 pts) A *side condition* is a linear constraint that, when added to the normal equations, makes parameters unique and individually estimable; side conditions must be nonestimable functions of the parameters and of rank that is the same as the deficiency of rank in $\mathbf{X}$. For the model given above, define an appropriate side condition and find the resulting solution to the normal equations.

   (f) (5 pts) Show that $H_0 : \tau_1 = \tau_2 = \tau_3$ is testable.

3. Consider the regression model,

$$Y(x_i) = \beta_0 + \beta_1|x_i| + \epsilon_i, \quad i = 1, 2, \ldots, n\ , \tag{1}$$

where the $Y(\cdot)$ are observable random variables, the x_i are known constants satisfying $-1 \leq x_i \leq 1$, β_0 and β_1 are unknown parameters, and the ϵ_i are assumed to be independent and identically distributed Normal random variables with mean 0 and variance 1.

(a) (5 points) Suppose $n = 4$ and $x_1 = -\frac{1}{2}, x_2 = \frac{1}{2}, x_3 = -1, x_4 = 1$. Find the best linear unbiased estimate (BLUE) of β_1.

(b) (5 points) Suppose $n = 4$ and $x_1 = 0, x_2 = 0, x_3 = 1, x_4 = 1$. Find the BLUE of β_1.

(c) (5 points) For purposes of estimating β_1, would you prefer observing $Y(\cdot)$ at the 4 points in (a) or the 4 points in (b)? Justify your answer.

(d) (10 points) Notice in equation (1) that $\beta_0 + \beta_1|x_i| = \beta_0 + \beta_1|-x_i|$. So in part (a), couldn't you observe $Y(\cdot)$ only at $x_2 = 1/2$ and $x_4 = 1$, and use these same two values of $Y(\cdot)$ for the values at $x_1 = -1/2$ and $x_3 = -1$? You would observe $Y(\cdot)$ only twice but you could pretend you have 4 observations. In other words, couldn't you pretend that $n = 4$ with $x_1 = -1/2,\ x_2 = 1/2,\ x_3 = -1,\ x_4 = 1$, take $Y(1/2)$ as the observation at both $x_1 = -1/2$, $x_2 = 1/2$, and take $Y(1)$ as the observation at both $x_3 = -1$ and $x_3 = 1$? How would you model this? Find the BLUE for β_1 and estimate its variance for this model.

4. For a square matrix $\boldsymbol{A}$, let $\boldsymbol{A}^-$ denote the *generalized inverse* satisfying $\boldsymbol{A}\boldsymbol{A}^-\boldsymbol{A} = \boldsymbol{A}$.

   Now suppose that the random n-vector $\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n)$, where $\boldsymbol{X}$ is a $n \times p$ design matrix of rank r not necessarily of full rank, $\boldsymbol{\beta}$ is a coefficient vector of length p, and $\sigma^2 > 0$.

   (a) (5 points) Prove that ordinary least squares (OLS) solutions of $\boldsymbol{\beta}$, $\widetilde{\boldsymbol{\beta}}$, satisfy

   $$\widetilde{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{Y}.$$

   (b) (2 points) Is the estimator $\widetilde{\boldsymbol{\beta}}$ unbiased? Explain.

   (c) (5 points) Assuming that σ^2 is fixed and known, derive the sampling distribution of $\widetilde{\boldsymbol{\beta}}$.

   Now assume that σ^2 is unknown and also needs to be estimated from the data. Also assume that $n > r$ for the remainder of the question.

   (d) (3 points) Write down a $100(1-\alpha)\%$ joint confidence region for $\boldsymbol{\beta}$. You do not need to prove this result.

   (e) (8 points) Prove that the $100(1-\alpha)\%$ Scheffé simultaneous confidence interval for estimable functions $\boldsymbol{a}^T\boldsymbol{\beta}$ of $\boldsymbol{\beta}$ where $\boldsymbol{a}$ is any vector constructed from the linear independent basis $\{\boldsymbol{l}_1, \ldots, \boldsymbol{l}_s\}$ is given by

   $$\boldsymbol{a}^T\widetilde{\boldsymbol{\beta}} \pm \sqrt{sF_{s,n-r;1-\alpha}}\,\widehat{\sigma}\sqrt{\boldsymbol{a}^T(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{a}}.$$

   You may use the following result in your proof: Let $\boldsymbol{V}$ be a positive definite matrix. Then for any vector $\boldsymbol{b}$,

   $$\sup_{\boldsymbol{h}\neq\boldsymbol{0}} \frac{(\boldsymbol{h}^T\boldsymbol{b})^2}{\boldsymbol{h}^T\boldsymbol{V}\boldsymbol{h}} = \boldsymbol{b}^T\boldsymbol{V}^{-1}\boldsymbol{b}.$$

   (f) (2 points) Suppose that the constant n-vector $\boldsymbol{x} = (x_1, \ldots, x_n)^T$ is not proportional to the n-vector of ones, $\boldsymbol{1}_n$. For a design matrix $\boldsymbol{X} = [\boldsymbol{1}_n\ \boldsymbol{x}]$, provide the Scheffé multiplier for joint inference of

   $$a_1\beta_1 + a_2\beta_2$$

   for all $a_1, a_2 \in \mathbb{R}$. Explain why your answer is correct.

Interdisciplinary PhD Program in Biostatistics

Qualifying Exam II

Day 2: Data Analysis #1

4 hours

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 instructions of each question 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**.
4. Computer Login Information:

   Username: bioexam
   Password: xxx
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 on the desktop of the computer with the file name: Day2_*examID*, where *examID* is your assigned ID code. Also save a copy on the Flash drive provided by the proctor with the same file name.

Oropharyngeal cancer is a disease in which malignant cells form in the tissue of oropharynx (a middle part of the throat which includes the base of the tongue, the tonsils, the soft palate, and the walls of the pharynx). The Radiation Therapy Oncology Group conducted a large clinical trial in the treatment of carcinoma of the oropharynx in the United States. The full study included patients with squamous carcinoma at any one of 15 sites in the mouth and throat, with 16 participating institutions. A subset of the data, considering only three sites in the oropharynx at six largest institutions, are in the data set "pharynx.csv". Patients entering the study were randomly assigned to one of two treatment groups, radiation therapy alone or radiation therapy together with a chemotherapeutic agent.

The following variables are available in the dataset pharynx.csv:

Variable	Description
CASE	Case Number
INST	Participating Institution
SEX	1=male, 2=female
TX	Treatment: 1=radiation therapy alone, 2=combined treatment
GRADE	1=well differentiated, 2=moderately differentiated, 3=poorly differentiated, 9=missing
AGE	In years at time of diagnosis
COND	Condition: 1=no disability, 2=restricted work, 3=requires assistance with self care, 4=bed confined, 9=missing
SITE	1=faucial arch, 2=tonsillar fossa, 4=pharyngeal tongue
T_STAGE	1=primary tumor measuring 2 cm or less in largest diameter, 2=primary tumor measuring 2 cm to 4 cm in largest diameter with minimal infiltration in depth, 3=primary tumor measuring more than 4 cm, 4=massive invasive tumor
N_STAGE	0=no clinical evidence of node metastases, 1=single positive node 3 cm or less in diameter, not fixed, 2=single positive node more than 3 cm in diameter, not fixed, 3=multiple positive nodes or fixed positive nodes
ENTRY_DT	Date of study entry: Day of year and year, dddyy
STATUS	0=censored, 1=dead
TIME	Survival time in days from day of diagnosis

The T and N staging classifications (T_STAGE and N_STAGE) give a measure of the extent of the tumor at the primary site and at regional lymph nodes. T=1 refers to a small primary tumor 2 centimeters or less in largest diameter, whereas T=4 is a massive tumor with extension to adjoining tissue. T=2 and T=3 refer to intermediate cases. N=0 refers to there being no clinical evidence of a lymph node metastasis and N=1, N=2, N=3 indicate, in increasing magnitude, the extent of existing lymph node involvement. Patients with classifications T=1,N=0; T=1,N=1; T=2,N=0; or T=2,N=1, or with distant metastases were excluded from study.

The condition variable (COND) gives a measure of the functional capacity of the patient at the time of diagnosis (1 refers to no disability whereas 4 denotes bed confinement; 2 and 3 measure intermediate levels). The variable GRADE is a measure of the degree of differentiation of the tumor (the degree to which the tumor cell resembles the host cell) from 1 (well differentiated) to 3 (poorly differentiated).

Perform an analysis of this data set to answer the key questions of the researchers below. Provide a single report (no more than 5 pages) of the results, but clearly label the sections corresponding to each question. Computer output such as graphs may be added as an attachment (does not count in 5 page limit).

1. One objective of the study was to compare the two treatment policies with respect to patient survival. Without considering other covariates, describe and compare the survival experiences of the two groups using non-parametric methods only.
2. In addition to the primary question whether the combined treatment mode is preferable to the conventional radiation therapy, it is of considerable interest to determine the extent to which other covariates also relate to subsequent survival. Covariates available in the data are SEX, T-STAGE, N-STAGE, AGE, COND, SITE and GRADE. The possible differences between participating institutions require consideration as well. Create a model to determine which covariates relate to survival, and report on any resulting adjustments to the estimates of survival experience in the treatment groups due to these covariates. Be sure to include in your report a discussion of appropriateness of the assumptions for the model. Be sure to fully interpret the results for the primary question of interest, the survival experience in the two treatment arms, using metrics that practitioners will understand.
3. The full data set includes 15 sites in the mouth and 16 institutions. Describe how you would handle analysis of these two variables in your model (note: this data is not provided – no need to actually run this analysis, just describe your approach). Include a brief discussion of the advantages and disadvantages of your approach, and how to interpret the results of the resulting model.

Interdisciplinary PhD Program in Biostatistics

Qualifying Exam II

Day 3: Data Analysis #2

4 hours

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 instructions of each question 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**.
4. Computer Login Information:

   Username: bioexam
   Password: xxx
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 on the desktop of the computer with the file name: Day2_*examID*, where *examID* is your assigned ID code. Also save a copy on the Flash drive provided by the proctor with the same file name.

To investigate the treatment effect of lowering cholesterol, two statin class drugs are being studied, namely, A-statin and P-statin. (All datasets in this question are simulated)

Part I:

A single center, randomized controlled double blinded clinical trial of 160 patients was conducted. Patients were taking the drug daily at a fixed dose level and followed up for one year. Treatment indicator (treatment=1 for A-statin; treatment=0 for P-statin), baseline characteristic (age in years) and LDL cholesterol level at 12 month after treatment (LDL in mg/dl) were included in the dataset random.csv.

1. Estimate the treatment effect of taking A-statin over P-statin without using covariate information. Interpret your results.

2. Estimate the treatment effect of taking A-statin over P-statin with using age as a covariate. Interpret your results.

3. Compare the results in 1 and 2, and discuss whether covariate information should be included in the analysis of randomized clinical trials like the one presented here (with continuous outcomes).

Part II:

An epidemiologist used an observational dataset to investigate the treatment effect of the above two drugs in a much larger population (n=2000). The dataset observation.csv includes the following variables:

Treatment:	Treatment indicator, 1 for A-statin, 0 for P-statin
ldl:	LDL cholesterol level at 12 month after taking the drug
Age:	In years
Gender:	1 for male, 0 for female
CIMT:	Carotid intima-medial thickness in mm
Edu:	Education lelvel, 0 for low, 1 for medium, 2 for high

4. One way to analyze observational studies is to use a stratification based adjustment, as described below:

    a. Estimate the probability of receiving A-statin based on all observed covariates (you may call this estimated probability “p-score”).
    b. Use quintiles of p-score to stratify the dataset into five strata.
    c. Estimate the treatment effect separately within each stratum.
    d. Combine the treatment effect estimates across strata to obtain the overall treatment effect estimate.

    Implement this method and present both point and variance estimates.

5. Compare the results in part I and II. Do you think they should be the same or different? Explain why.

Write a report of no more than 4 pages to summarize your findings, which should incorporate your answers to the above questions. Clearly label the sections corresponding to each question. Only present key output/table/graph in the report. You may include additional details in an appendix.