Division of Biostatistics
College of Public Health
Qualifying Exam II
Methodology

1-5 pm, June 4, 2014
Closed Book

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 being sure to show your work and justify your answers.

1. Consider a model for longitudinal outcome data Y_{ij} where i indicates the subject ($i = 1, ..., N$) and j denotes the time point ($j = 1, ..., m$ for all individuals):

$$Y_{ij} = \gamma_1(X_{ij}) + (Y_{i(j-1)}) + \epsilon_{ij} \tag{1}$$

where $Y_{i0} = 0$, the error terms (ϵ_{ij}) are iid Normal with mean 0 and variance σ_e^2 and the covariates x_{ij} are also iid Normal with mean 0 and variance τ^2. x_{ij} and ϵ_{ij} are independent of each other and $y_{i(j-1)}$.

(a) (2 pts) Show that in equation (1), y_{ij} can be written only in terms of prior covariate values (i.e., write an expression for y_{ij} that does not include any prior y values.)

(b) (2 pts) Find both the fully conditional expectation ($E(Y_{ij}|x_{i1}, ..., x_{ij})$) and the marginal expectation ($E(Y_{ij}|x_{ij})$) for the above (equation (1)).

(c) (3 pts) Give an example in which the marginal model $E(Y_{ij}|x_{ij}))$ might be of interest, despite the fact that it is not the "correct" model for the data.

Investigators are planning to use GEE to fit the marginal model $E(Y_i|X_i) = X_i\beta$, where $Y_i = (y_{i1}, ...y_{im})^T$, $X_i = (x_{i1}, ...x_{im})^T$, and β is a scalar.

(d) (6 pts) Show that for a linear model (identity link) and assumed known working covariance matrix $W^{-1} = A$ with elements a_{rc}, the estimate $\hat{\beta}$ is given by:

$$\hat{\beta} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{m}(\sum_{t=1}^{m} a_{tj}x_{it})y_{ij}}{\sum_{i=1}^{n} X_i^T A X_i}$$

(e) (8 pts) Using a working independence covariance matrix (so that $a_{rc} = 0$ when $r \neq c$ and $a_{rr} = 1$, find the (unconditional) expectation of $\hat{\beta}$ and show that the bias is negligible when N is large.

When the working covariance is not independence, the (unconditional) expectation of $\hat{\beta}$ is given by:

$$E(\hat{\beta}) = E(E(\hat{\beta}|X)) \approx \left[1 + \frac{\sum_{t=1}^{m-1}\sum_{j=t+1}^{m} a_{tj}}{\sum_{j=1}^{m}\lambda_j}\right]\beta, \tag{2}$$

where λ_j are the eigenvalues of A and must all be positive because A must be positive definite.

(f) (4 pts) Discuss implications of the result shown in equation (2) in terms of (a) potential bias and (b) implications for interpretation of results from the marginal model fit using GEE.

2. A study was conducted by investigators at a university in Australia to identify predictors of number of compensable injuries reported by cleaning services employees at a large hospital. A total of 137 employees were enrolled in the study with an average of 1.20 injuries reported per employee.

   a.) (3 pts) Injury counts are usually modeled using Poisson regression. However, the investigators in this study were worried that the Poisson assumption may be violated since the percentage of employees who reported zero injuries (53%) exceeded the level expected for a Poisson distribution. Do the data support this claim? Provide evidence supporting your answer. Hint: consider a Poisson distribution with mean equal to the average number of injuries per employee.

   b.) (5 pts) Since the investigators were concerned about a violation of the Poisson assumption, they considered the following Zero-Inflated Poisson (ZIP) model for number compensable injuries reported by an employee (Y):

$$\begin{aligned} P(Y=0) &= \phi + (1-\phi)e^{-\theta} \\ P(Y=y) &= (1-\phi)\frac{e^{-\theta}\theta^y}{y!} \quad y = 1, 2, \ldots, \end{aligned}$$

   where $\theta > 0$ is a Poisson mean and $0 < \phi < 1$ is parameter that accounts for more zeros than those allowed by the Poisson distribution. The parameter ϕ can also be thought of as the proportional of subjects who are immune to injury. Provide explicit expressions for the expected value and variance of Y based on the ZIP model.

   c.) (3 pts) Assuming the first 72 employees reported 0 injuries and the last 65 reported one or more, provide an expression for the likelihood based on the ZIP model from part b.

   d.) (7 pts) To simplify the estimation, define $p = (1-\phi)(1-e^{-\theta})$. Revise the log-likelihood from part c to be in terms of the parameters p and θ and show it can be factored into the product of two parts: 1.) The product of 137 Bernoulli(p) probability mass functions (pmfs) and 2.) the product of 65 truncated Poisson ($Y > 0$) pmfs.

   e.) (3 pts) The investigators considered the following regression models for p and θ:

$$\begin{aligned} \text{logit}(p_i) &= \log t_i + \alpha_0 + \alpha_1 \text{age}_i \\ \log(\theta_i) &= \log t_i + \beta_0 + \beta_1 \text{age}_i, \end{aligned}$$

   where $\log(t_i)$ is an offset term accounting for differences in number of hours worked by each employee (t_i) and age_i is the employee's age in years. The following maximum likelihood estimates were obtained:

$\hat{\alpha}_0$:	-0.409
$\hat{\alpha}_1$:	0.008
$\hat{\beta}_0$:	-1.072
$\hat{\beta}_1$:	0.015

Provide an interpretation of the value of $e^{\hat{\alpha}_1}$ that would be meaningful to someone who isn't a statistician.

f.) (4 pts) Suppose the investigators provided the following interpretation for $e^{\hat{\beta}_1}$: "Among employees reporting one or more injuries, the mean number of injuries increases approximately 1.5% per year increase in age controlling for number of hours worked." Is this a correct interpretation? Why or why not?

3. A hemoglobin molecule can carry one oxygen or one carbon monoxide molecule or none. Suppose that the molecules for these two types of gases arrive at rates μ_{ox} and μ_{cm} and attach to hemoglobin for an exponential amount of time with rates λ_{ox} and λ_{cm}, respectively.

   a.) (9 pts) Formulate a continuous time Markov chain model with 3-state space describing the state of a hemoglobin-gas molecule complex and find the long-run fraction of time the hemoglobin molecule is in each of its three states. Cite the appropriate theoretical result as necessary to justify your answer.

   b.) (8 pts) It is desired to perform a computer simulation from the above Markov chain. Describe the steps of an appropriate algorithm. Be specific with the distribution from which you are simulating the data.

   c.) (8 pts) Hypoxia (also known as Hypoxiation or Anoxemia) is a condition in which the body or a region of the body is deprived of adequate oxygen supply due to insufficient rate of oxygen transport. Describe briefly how could you model hypoxia with the above Markov chain model, that is, what conditions on the model parameters could be imposed to mimic the disease symptoms.

4. (a) (6 pts.) Let f_1, f_2 be two pdfs with known means μ_1, μ_2, and variances σ_1^2, σ_2^2. Consider the mixture pdf

$$f(x) = \alpha f_1(x) + (1-\alpha) f_2(x), \quad 0 < \alpha < 1, \tag{1}$$

where α is known. Find the mean and variance of X in terms of these known quantities in as simple form as possible.

(b) (3 pts.) How do you simulate a random variable X with pdf given in (1)?

For parts (c) - (e), consider the following set up. We have a system with two working components - labeled 1 and 2 whose lifetimes X and Y have exponential pdfs with rates $2\lambda_1$ and $2\lambda_2$ when both are working. When component 1 fails, the burden on component 2 is reduced and its remaining life-length will have exponential distribution with rate λ_2; and similarly when component 2 fails, the remaining life-length of component 1 will have exponential distribution with rate λ_1. (For example, the components could be two competing species.)

(c) (6 pts.) Show that the joint pdf of X and Y can be expressed as

$$f(x,y) = \begin{cases} 2\lambda_1\lambda_2 \exp\{-2(\lambda_1+\lambda_2)x - \lambda_2(y-x)\}, 0 < x \le y \\ 2\lambda_1\lambda_2 \exp\{-2(\lambda_1+\lambda_2)y - \lambda_1(x-y)\}, 0 < y \le x. \end{cases} \tag{2}$$

(d) (8 pts.) Using the above expression for the joint pdf, show that the marginal pdf of X can be expressed as

$$f(x) = \frac{2\lambda_1(\lambda_1+\lambda_2)}{\lambda_1+2\lambda_2} \exp\{-2(\lambda_1+\lambda_2)x\} + \frac{2\lambda_1\lambda_2}{\lambda_1+2\lambda_2} \exp\{-\lambda_1 x\}, \tag{3}$$

for $x > 0$. Identify (3) as a special case of (1) and determine the associated α, the component distributions and their parameters.

(e) (2 pts.) Using part (a), or otherwise, find the mean of X with the pdf given in (3).

Division of Biostatistics
College of Public Health
Qualifying Exam II
Data Analysis I

9 am—1 pm, June 5, 2014

1. This part contains one data analysis project. Submit a final report for the project with your exam ID code on the title page. Do **NOT** put your name on any page of your report.

2. The dataset is saved as read-only on the qualifier exam drive. You need to first download it to the desktop of your computer before you start working on it. At the end of exam period, print a copy of your final report and also save an electronic copy on the desktop with your exam ID as the file name.

3. This part is open book and you are allowed to bring up to 10 books and unlimited class notes as references.

4. The report should be self-contained. Follow the instruction of each question to prepare your answer.

5. The project is worth 50 points.

6. No access to internet is allowed during the exam.

7. Login information

   For the exam in the lab, all students need to login with the following account:
   Username: bioexam
   Password: Buckeyes2014! (case-sensitive)
   The dataset for the test can be found under My Computer in the T: Drive. They will be the only files available. Each student should copy the file to the desktop of his/her machine before starting the exam.

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.

Division of Biostatistics
College of Public Health
Qualifying Exam II
Data Analysis II

9 am—1 pm, June 6, 2014

1. This part contains one data analysis project. Submit a final report for the project with your exam ID code on the title page. Do **NOT** put your name on any page of your report.

2. The dataset is saved as read-only on the qualifier exam drive. You need to first download it to the desktop of your computer before you start working on it. At the end of exam period, print a copy of your final report and also save an electronic copy on the desktop with your exam ID as the file name.

3. This part is open book and you are allowed to bring up to 10 books and unlimited class notes as references.

4. The report should be self-contained. Follow the instruction of each question to prepare your answer.

5. The project is worth 50 points.

6. No access to internet is allowed during the exam.

7. Login information

   For the exam in the lab, all students need to login with the following account:
   Username: bioexam
   Password: Buckeyes2014! (case-sensitive)
   The dataset for the test can be found under My Computer in the T: Drive. They will be the only files available. Each student should copy the file to the desktop of his/her machine before starting the exam.

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.