

University of Illinois at Chicago
School of Public Health
Epidemiology-Biostatistics Division

PRELIMINARY EXAMINATION
Ph.D. in Biostatistics

Part II

Thursday, September 10, 2009, 1 pm
to
Thursday, September 17, 2009, 1 pm

This is a take-home exam. You are not to communicate about any aspect of this exam with anyone except Drs. Bhaumik (dbhaumik@psych.uic.edu), Chen (hychen@uic.edu), Demirtas (demirtas@uic.edu), Freels (sallyf@uic.edu), Hedeker (hedeker@uic.edu).

There are **six** questions in this exam. You need to answer only **three** of them. Only the chosen **three** will be counted towards your score. Datasets can be downloaded from the website: *www.uic.edu/~demirtas/prelim*. Please start each question on a separate page. Before you turn in your answers, number your pages consecutively in the upper right-hand corner, and put your code number (*not your name*) next to the page number. Turn in your completed exam to Dr. Demirtas (room 950) on September 17, between 10 am and 1 pm.

Code Number: _____

1. Linear Models

- (a) Consider Table 1 from a study of nonmetastatic osteosarcoma (A. M. Goorin, *J. Clin Oncol* 5: 1178-1184, 1987). The response is whether the subject achieved a three-year disease free interval.

Table 1:

Lymphocytic Infiltration	Gender	Osteoblastic Pathology	Disease Free	
			Yes	No
High	Female	No	3	0
		Yes	2	0
	Male	No	4	0
		Yes	1	0
Low	Female	No	5	0
		Yes	3	2
	Male	No	5	4
		Yes	6	11

- i. Show that each predictor has a significant effect when used individually without the others.
 - ii. Try to fit a main-effects logistic regression model containing all three predictors. Explain why the ML estimate for the effect of lymphocytic infiltration is infinite.
 - iii. Using conditional logistic regression (i) conduct an exact test for the effect of lymphocytic infiltration, controlling for the other variables, and (ii) find a 95% confidence interval for the effect. Interpret results.
- (b) Consider the following experimental data for copper from an interlaboratory study conducted by the Ford Motor company (see Table 2). These data were generated as part of a blind interlaboratory study of laboratories which hold Michigan State Drinking Water Certifications for the parameters tested. Samples were prepared by an independent source, randomized and submitted on a weekly basis over a five week period. Copper was analyzed by Inductively Coupled Plasma Atomic Emissions Spectroscopy (ICP/AES) using EPA method 200.7. The dataset consisted of five replicates at each of five concentrations (0, 2, 10, 50 and 200 $\mu\text{g/L}$) in each of seven laboratories.

Table 2: Interlaboratory data for Copper in $\mu\text{g/L}$

Lab	Rep	Concentration in $\mu\text{g/L}$				
		0	2	10	50	200
1	1	3.000	3.000	14.000	54.000	205.000
1	2	2.000	3.000	10.000	51.000	206.000
1	3	-1.000	5.000	11.000	52.000	208.000
1	4	1.000	2.000	12.000	54.000	211.000
1	5	-1.000	2.000	13.000	38.000	195.000
2	1	2.100	8.000	10.000	53.000	188.600
2	2	0.300	1.800	12.400	54.600	210.000
2	3	2.000	0.700	10.600	50.000	210.000
2	4	1.300	4.000	12.000	50.100	214.000
2	5	2.000	3.000	11.000	50.000	200.000
3	1	0.800	2.495	10.500	47.660	181.330
3	2	-0.185	2.695	10.335	45.390	173.205
3	3	0.990	2.410	9.735	44.270	180.560
3	4	0.905	1.840	10.245	46.910	183.650
3	5	0.365	2.840	10.325	47.240	181.585
4	1	1.661	3.243	12.250	48.140	205.400
4	2	1.996	3.432	13.510	54.450	200.400
4	3	0.000	9.246	11.160	51.010	199.700
4	4	2.993	3.390	13.440	52.860	189.600
4	5	2.042	4.109	10.470	48.720	187.700
5	1	0.090	0.860	10.030	50.060	193.400
5	2	-2.510	2.680	12.940	50.350	193.470
5	3	7.270	-0.400	8.970	49.320	203.160
5	4	7.140	4.730	9.610	49.930	190.020
5	5	0.280	5.200	9.120	48.080	191.050
6	1	7.226	4.964	4.713	48.242	191.020
6	2	-1.000	2.000	10.000	65.000	205.000
6	3	0.000	3.000	8.000	45.000	183.000
6	4	10.244	6.716	11.101	43.000	185.000
6	5	-2.177	8.844	8.249	47.000	182.000
7	1	0.018	1.323	6.000	45.500	162.000
7	2	-3.000	4.900	9.088	44.000	181.000
7	3	0.000	0.000	14.100	40.000	187.000
7	4	-2.000	0.000	6.000	43.000	178.300
7	5	-2.000	0.000	7.000	45.986	188.932

Fit an appropriate regression model and estimate the parameters. Write a report (at least of two pages) addressing the following points: (i) Why do you think that I should prefer to your model? (ii) How can I use your model for the prediction purpose?, (iii) How will you modify your model if you want to include one more lab in the study ?

2. Computational Statistics

Suppose $g'(x) = \frac{-3062(1-\epsilon)e^{-x}}{[\epsilon+(1-\epsilon)e^{-x}]} - 1013 + 1628/x$. When $g'(x) = 0$, it can be written that $G_1(x) = x = \frac{1628[\epsilon+(1-\epsilon)e^{-x}]}{3062(1-\epsilon)e^{-x}+1013[\epsilon+(1-\epsilon)e^{-x}]}$. Assume that $\epsilon = 0.61489$. The goal is to find a value of x that maximizes $g(x)$.

- a) Use fixed point iteration with $G_1(x)$.
- b) Demonstrate that $G_2(x) = x + g'(x)$ fails to converge.
- c) Try $G_3(x) = x + \alpha g'(x)$, where $\alpha = 1/1000$.
- d) Use Newton-Raphson with a starting value $x^{(0)} = 2$.
- e) Use Newton-Raphson with a starting value $x^{(0)} = 1.5$. What is the difference with part d?
- f) Use secant method. I suggest that you choose your second starting value as 1.49 since it is close to 1.5.
- g) Use Muller's method. Choose your third starting value as 1.48.
- h) Use bisection method.
- i) Use secant-bracket method.
- j) Use Illinois method.

3. Survival Analysis

A study was done comparing two treatment modalities to reduce the occurrence of muscle soreness among middle-aged men beginning weight training. The data file `recur.dat` includes data on episodes of soreness for 400 men. Each subject has a separate row of data reflecting up to 4 episodes. Subjects are randomized to receive either a new treatment or the old (standard) treatment. The number of days since treatment is then recorded for each episode. Variables appear in the following order:

ID Subject Identification (1 - 400)
AGE Age (years)
TREAT Treatment Assignment (0 = New, 1 = Old)
TIME0 Day of Previous Episode (days after treatment)
TIME1 Day of New Episode or Censoring (days after treatment)
CENSOR Indicator for Soreness Episode or Censoring
(1 = Episode Occurred at TIME1, 0 = Censored at TIME1)
EVENT Soreness Episode Number (0 - 4)

Analyze this data using models which do not make any parametric assumption about the underlying hazard distribution; which assume proportional hazards across time for effects of independent variables; and which account properly for the correlation amongst different episodes for the same patient.

a) Provide estimated hazard ratios for the effect of randomized treatment adjusted for age, as well as the effect of age (adjusted for treatment), on the hazard $h(t)$, where t is defined as

i) t = number of days from treatment assignment. ii) t = number of days since last episode.

b) Create a graph of the estimated hazard function across time in each treatment, adjusted for age, for each of the two definitions of t above.

c) How is the interpretation of results any different for approaches i) and ii)? Which one do you think is more appropriate, and why?

4. **Longitudinal Data Analysis** The data for this question are from the Riesby *et. al.*, article that is discussed in the Longitudinal Data Analysis class. This study examined the relationship in depressed inpatients between the drug plasma levels - the antidepressant imipramine (IMI) and its metabolite desimipramine (DMI) - and clinical response as measured by the Hamilton Depression Rating Scale (HDRS). In class, we noted that there was a significant relationship across time between the drug plasma levels (specifically, desimipramine) and depression change. Here, examine the degree to which this posited relationship is moderated by a subject's gender. The dataset is located at <http://tigger.uic.edu/~hedeker/RIESBYT4.DAT.txt> and contains the following variables:

field 1: Patient ID

field 2: HDRS change from baseline score

field 3: a field of ones

field 4: Week - from 0 (week 2) to 3 (week 5)

field 5: sex (0 = male 1 = female)

field 6: diagnostic group (0 = non-endogenous 1 = endogenous)

field 7: Imipramine (IMI) plasma levels (in ln units)

field 8: Desimipramine (DMI) plasma levels (in ln units)

Is there a different relationship between DMI and depression change for males and females? If so, describe how it varies. Are there any influential observations or outliers that might limit the robustness of your conclusions?

5. Missing Data

In a two-stage sampling design studying the effect of a certain biomedical measurement on the outcome, outcome and demographical variables are collected for all subjects in a random sample at the first stage. At the second stage, expensive bioassay is performed to obtain the biomedical measurement on a subsample of the subjects from the first stage. The selection of subjects to be included in the second stage depends on the outcome and demographical variables measured in the first stage.

- (a) Formulate this problem in terms of a statistical model.
- (b) Propose a method for statistical analysis of the data collected.
- (c) Discuss potential alternative statistical methods for this problem and compare their relative merits.

6. Bayesian Statistics

The characteristic function of a univariate random variable X is defined as $\phi_X(t) = E[e^{itx}]$, and it exists for any random variable. For a univariate random variable whose first moments are finite, and whose characteristic function ϕ is such that $\int |\phi(t)|dt$ and $\int |\phi''(t)|dt$ are finite, Devroye (1986) describes a method for generating random variates using the characteristic function. In the following algorithm, $p(\cdot)$ is the probability density function for a univariate continuous random variable.

- (a) Set $a = \sqrt{\frac{1}{2\pi} \int |\phi(t)|dt}$ and $b = \sqrt{\frac{1}{2\pi} \int |\phi''(t)|dt}$.
- (b) Generate u and v independently from a $U(-1, 1)$ distribution.
- (c) If $u < 0$, then set $y = bv/a$ and $t = a^2|u|$.
Otherwise, set $y = b/(va)$ and $t = a^2v^2|u|$.
- (d) If $t \leq p(y)$, then take y as the desired realization.
Otherwise, return to Step (b) above.

Consider three univariate posteriors of well-known distributions of your choice, and generate a large number of random variates using this algorithm. Compare your findings with the ones your favorite standard random number generation tools give. Write a report that summarizes your conclusions.

Reference: Devroye, L. (1986). Non-uniform random variate generation, Springer-Verlag, New York.