Division of Biostatistics Q2 Exam

Part I

August 27, 2010

1:00pm-5:00pm

Closed book

1. Write the question number in the upper left-hand corner and your 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. A study was designed to identify factors related to violent and non-violent infractions among inmates in Swiss prisons. The factors examined include demographic variables, information about the index offense that led to imprisonment, criminal history as well as in-prison behavior.

The sample consists of 318 male prisoners in Switzerland.

The code sheet for these data is:

	Variable	Label	Coding
Demographic	id	Identification number	
	age	Age in years (entry in prison)	Continuous number
	swiss	Swiss national	0 = no; 1 = yes
	civilstatus	Marital status	1 = single 2 = married 3 = divorced 4 = widowed
	illegal	Illegal residance status in Switzerland	0 = no; 1 = yes
Criminal history	nr_con	Number of convictions prior to the index offense	Continuous number
	prior_con	At least one prior conviction prior to the index offense	0 = no; 1 = yes
	viol_rec	Conviction for a violent offense prior to the index offense	0 = no; 1 = yes
Index offense	index_viol	Index offense that let to imprisonment was a violent offense (assault, murder, robbery)	0 = no; 1 = yes
	index_abuse	Index offense was child abuse	0 = no; 1 = yes
	index_sex	Index offense was a sex offense (adult victim) (e.g. rape)	0 = no; 1 = yes
	index_prop	Index offense was property offense	0 = no; 1 = yes
	index_drug	Index offense was drug offense	0 = no; 1 = yes
Imprisonment	time	Time spent in prison at the time of the investigation (month)	Continuous number
	infrac	Number of nonviolent infractions	Continuous number
	violence	Violent infractions during imprisonment	0 = no; 1 = yes

Of interest is to determine the factors that are predictive of which prisoners commit violent infractions during their imprisonment.

The 2×2 table of the outcome variable "violence" vs the categorical variable "swiss" is as follows:

```
. tab violence swiss

   violent |        swiss
 infracton |         0          1 |     Total
-----------+----------------------+----------
         0 |       144         91 |       235 
         1 |        60         23 |        83 
-----------+----------------------+----------
     Total |       204        114 |       318 
```

(a) Is there an association between being a Swiss national and committing a violent infraction while incarcerated? What would the coefficient be for the variable "swiss" in a logistic regression model? What would the standard error of the coefficient be? Using your point estimate and standard error, compute a 95% confidence interval for the log odds ratio of the association between "swiss" and "violence".

(b) Assuming violence is the response, carefully write the likelihood function that underlies the logistic regression model corresponding to the 2×2 table shown earlier.

A model was fit that included the variables: time, swiss, infrac and viol_rec. (Note: In this model, the variable "newtime" divides the variable "time" by 12.)

```
. gen newtime= time/12
. logit violence newtime swiss infrac viol_rec

Iteration 0:   log likelihood = -182.56597
Iteration 1:   log likelihood = -144.96926
Iteration 2:   log likelihood = -144.07517
Iteration 3:   log likelihood =  -144.0726
Iteration 4:   log likelihood =  -144.0726

Logistic regression                               Number of obs   =        318
                                                  LR chi2(4)      =      76.99
                                                  Prob > chi2     =     0.0000
Log likelihood =  -144.0726                       Pseudo R2       =     0.2108

------------------------------------------------------------------------------
    violence |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     newtime |   .1139039   .0486755     2.34   0.019     .0185016    .2093061
       swiss |  -1.029187   .3772512    -2.73   0.006    -1.768586   -.2897883
      infrac |   .2768106    .051555     5.37   0.000     .1757647    .3778564
    viol_rec |   .6653906   .3684903     1.81   0.071    -.0568372    1.387618
       _cons |  -1.948656   .2329258    -8.37   0.000    -2.405182    -1.49213
------------------------------------------------------------------------------
```

(c) This model assumes that the variable "time" is *linear in the logit.* Carefully explain the implications of that assumption in this model by computing the odds ratio corresponding to a 12 month increase at two different starting points in time (e.g., 12 months and 36 months).

(d) A number of methods can be used to assess the validity of the assumption that "time" is linear in the logit. Please summarize which methods you might use for this purpose.

Suppose you believe that the variable "time" is not linear in the logit. In fact, you believe that the relationship is quadratic rather than linear. As a result, the following model was fit. (Note: In this model, the variable "newtime" divides the variable "time" by 12 and "newtimesq" = "newtime"^{2}.

```
. gen newtimesq= newtime* newtime
. logit violence newtime newtimesq swiss infrac viol_rec

Iteration 0:   log likelihood = -182.56597
Iteration 1:   log likelihood = -140.48048
Iteration 2:   log likelihood = -139.14154
Iteration 3:   log likelihood = -139.13951
Iteration 4:   log likelihood = -139.13951

Logistic regression                               Number of obs   =        318
                                                  LR chi2(5)      =      86.85
                                                  Prob > chi2     =     0.0000
Log likelihood = -139.13951                       Pseudo R2       =     0.2379

------------------------------------------------------------------------------
    violence |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     newtime |   .5424386   .1505048     3.60   0.000     .2474547    .8374226
   newtimesq |  -.0340503   .0115048    -2.96   0.003    -.0565994   -.0115013
       swiss |  -1.197425   .3885879    -3.08   0.002    -1.959043   -.4358066
      infrac |   .2515902   .0522881     4.81   0.000     .1491074    .3540731
    viol_rec |   .6185321   .3776603     1.64   0.101    -.1216685    1.358733
       _cons |  -2.413786   .2937536    -8.22   0.000    -2.989533    -1.83804
------------------------------------------------------------------------------
```

(e) Provide a table estimating the odds ratio corresponding to a 12-month increase in time starting at 12 and at 60 months. Interpret these results, commenting on the appropriateness or inappropriateness of the assumption of linearity in the logit for time in this model.

After the model on page 2 was run, the variance-covariance matrix was generated as follows:

```
. vce
Covariance matrix of coefficients of logit model

             | violence
        e(V) |   newtime       swiss      infrac    viol_rec       _cons
-------------+------------------------------------------------------------
violence     |
     newtime |  .00236931
       swiss | -.00919989   .14231844
      infrac | -.00071935    .0001171   .00265791
    viol_rec |   .0009161  -.01637681   .00058747    .1357851
       _cons | -.00311284  -.01062054  -.00501075  -.02734419   .05425445
```

(f) Use this information to show how you would compute a 95% confidence interval for the probability that a non-swiss prisoner, in prison for 12 months, who had 3 previous nonviolent infractions and who was convicted for a violent offense prior to the index offense. Without actually carrying out the calculations, carefully show how you would set up the formulas you would need to solve.

2. Cure rate models are a class of survival models commonly used to model data from cancer clinical trials. Such models allow a certain percentage of the population to be cured; these subjects will never experience the event of interest (e.g., recurrence of cancer following chemotherapy) even if followed for an infinite amount of time. In this question, assume the following cure rate model:

$$S_{pop}(t) = \exp\{-\theta[1 - \exp(-\lambda t)]\}, \tag{1}$$

where $S_{pop}(t)$ is the survival function for a certain population evaluated at time t.

(a) Derive the hazard function of the population $h_{pop}(t)$.

(b) Provide the limit of $S_{pop}(t)$ as $t \to \infty$; this is the cure rate.

(c) Suppose that for an individual in the population, there are N metastatic-competent tumor cells left active following treatment where $N \sim$ Poisson(θ). Given $N = n \geq 1$, let $Y_1, \ldots, Y_n \overset{iid}{\sim} Exp(\lambda)$ denote the promotion times of these cells. Under this set-up, time to relapse of cancer is $T = \min\{Y_k, 0 \leq k \leq N\}$, where $P(Y_0 = \infty) = 1$, and hence the survival function of the population may be written as

$$S_{pop}(t) \quad = \quad P(N = 0) + P(Y_1 > t, \ldots, Y_N > t, N \geq 1). \tag{2}$$

Show that Equation 2 is equivalent to Equation 1.

(d) Derive the survival function for the portion of the population that is not cured; i.e.,

$$S^*(t) = P(T > t | N \geq 1).$$

3. Suppose $Y_1, \ldots, Y_T$ are mutually independent random variables with $Y_i \sim \text{Binomial}(m_i, p_i)$, where m_i is a *known* number of trials and p_i is an unknown success probability with $0 < p_i < 1$, for $i = 1, \ldots, T$. Additionally let $x_1, \ldots, x_T$ be a given set of T real numbers.

   (a) Find the variance of
   $$W = \sum_{i=1}^{T} Y_i x_i - \widehat{p} \sum_{i=1}^{T} m_i x_i.$$
   where $\widehat{p} = \sum_{i=1}^{T} Y_i / \sum_{i=1}^{T} m_i$.

   (b) When $p_1 = \cdots = p_T = p$ show that the variance of W reduces to
   $$p(1-p) \sum_{i=1}^{T} m_i \left(x_i - \overline{x}_w\right)^2$$
   where $\overline{x}_w := \sum_{i=1}^{T} m_i x_i / \sum_{i=1}^{T} m_i$.

   Now assume that the logistic model,
   $$\text{logit}(p_i) = \beta_0 + \beta_1 x_i,$$
   holds for $i = 1, \ldots, T$.

   (c) Show that the score statistic for testing H_0: $\beta_1 = 0$ against H_1: $\beta_1 \neq 0$ is given by
   $$X^2 = \frac{\left(\sum_{i=1}^{T} Y_i x_i - \widehat{p} \sum_{i=1}^{T} m_i x_i\right)^2}{\widehat{p}(1-\widehat{p}) \sum_{i=1}^{T} m_i \left(x_i - \overline{x}_w\right)^2}. \tag{1}$$
   This test statistic is the (well-known) Cochran-Armitage trend test.

   (d) Describe intuitively, by examining the quantity being squared in the numerator of the of (1), why this test statistic will be large if $\text{logit}(p_i)$ is positively associated with x_i.

4. Assume that an outcome variable y may be linearly related to predictor variables x, z, and w.

   (a) State whether or not each of the following conditions can occur, and justify each answer.

      i. y is significantly correlated with each of x and z, but when y is simultaneously regressed on both x and z, neither the regression coefficient for x nor the regression coefficient for z is significant.

      ii. y is not significantly correlated with either x or z, but when y is simultaneously regressed on both x and z, both the regression coefficient for x and the regression coefficient for z are highly significant.

      iii. The correlation of y with x is 0.3 and y with z is 0.2, but the correlation of y with a linear combination of x and z is 0.7.

   (b) Suppose that we regress y on x, z, and w.

      i. Define R^2 (that is, the multiple coefficient of determination) in terms of observed and predicted values and/or residuals.

      ii. Explain how you would use the usual tests of the model and/or its coefficients to test whether R^2 is significantly different from zero.

      iii. Define the partial correlation of y and x given z and w in terms of observed and predicted values and/or residuals from appropriate regression analyses.

      iv. Suppose that you have only the usual F-test for the regression of y on x, z, and w and the tests of the regression coefficients in this analysis. Explain how you would use these results to test whether or not the partial correlation of y and x given z and w is significantly different from zero.

   (c) Assume that there is only available a program to compute simple linear regressions. Explain how to compute the coefficients of the multiple linear regression of y on x and z using only simple linear regression(s). Be explicit.

Division of Biostatistics Q2 Exam

Part II

August 28, 2010

8:00am-5:00pm

Open book

1. There are two data analysis projects for this part. Submit a separate report for each project with your ID code on the title page of the reports. Do NOT put your name on any page of your reports.

2. Time allocation:

8 am – 12 pm	Project #1
12 pm – 1pm	Lunch Break
1 pm – 5 pm	Project #2

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

4. Part II is open book and you are allowed to bring up to 10 books and unlimited class notes as references.

5. Each report should be a self-contained scientific article, including the following sections (DO NOT merely provide answers to each question):

   a. Executive summary of scientific conclusions (one page).

      The executive summary should be addressed to research scientists in general, not necessarily statisticians. Therefore, be cautious with using jargon unfamiliar to those who have statistical knowledge not beyond PH-BIO 703.

b. Introduction (up to two pages)

It should include a description of the problem; discussions on the strength and limitations of the data, etc.

c. Overview of the analyses (up to two pages)

It may include brief description of the analyses attempted and the rationales.

d. Exploratory data analyses (up to three pages)

Summarize the analyses performed for the purpose of formulating the hypotheses, identifying the appropriate statistical test/model, etc.

e. Final analyses (up to five pages)

Provide the details for your final analyses, which may include model building, model checking, interpretation of the results, etc.

f. Appendix (optional; unlimited)

It may include computer code, important additional tables or graphs to justify your analyses, etc.

6. Each project worth 50 points with the following point distribution:

   a. Executive summary of scientific conclusions (10 pts)

   b. Introduction (5 pts)

   c. Overview of the analyses (5 pts)

   d. Exploratory data analyses (10 pts)

   e. Final analyses (20 pts)

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

8. Log-in Information

   For the exam in the lab, all students need to login with the following account:

   Username: bioexam
   Password: bioexam

   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 their machine before starting the exam.

Biostatistics Qualifier Exam

Day Two

Data Analysis Project #1

Open book

Four hours

The attached data set contains information on 709 premenopausal women with breast cancer that were recruited for a multi-site randomized clinical trial of hormonal therapy. Patients were randomized to receive either mastectomy plus 5 years of tamoxifen (treatment group) or mastectomy alone (control group). The primary purpose of the study was to study disease –free and overall survival between the two groups. Patients were scheduled visits every 3 to 6 months over the first 5 years following treatment assignment, and annually thereafter. The discovery of estrogen receptor-positive (ER+) tumors in women led to the design and conduct of this study. There was an expectation that only ER+ would benefit from treatment, but because there was no method for determining ER status at time of the trial, they enrolled all patients. Now there is an accurate identifier. So, an additional study goal is to determine whether ER status does make a difference in the effect of treatment. The dataset contains the following variables:

rnumber: patient randomization number

enroll_date: Date of patient enrollment

adjuvant_trt: Indicator of trt group (1 = adjuvant trt, 0 = observation)

dead_date: Date of patient death or date of censoring for OS outcome

dead: Indicator of patient death (or censoring) at dead_date (1= death, 0 = censored)

dfs_date: Date of disease recurrence or censoring for DFS outcome

dfs_status: Indicator of recurrence (or censoring) at dfs_date (1 = recur, 0 = censored)

path_size: Pathologic size of largest dimension (in cm) for primary tumor

path_tot_nodes: Total number of lymph nodes examined

path_pos_nodes: Number of positive lymph nodes (out of total number examined)

path_hist_grd: Histologic grade of primary tumor

er_pos: Estrogen receptor status of primary tumor (1= ER+, 0= ER-)

age: Patient age

You are asked to import the data into a statistical software package and perform the appropriate analysis to address the following questions. Prepare a report using the format described in the cover sheet. Do Not merely answer the questions one by one.

Question 1:

A) Please compute overall survival (OS) and disease free survival times (DFS).

B) Provide descriptive statistics (median, means, standard deviation) for age, DFS, OS. Please do it overall and by group.

C) Provide median overall follow-up time, overall and for patients with and without events.

D) Specify how you would calculate overall survival and disease-free survival time for patients that died without known recurrence or demonstrable non-disease-related death.

E) Which of these variables (overall survival or DFS) is better in this trial to focus on as primary endpoint?

Question 2:

In planning this trial choose and justify a testing method among the several alternatives. (Hints: What are the differences in terms of the relative weights and power among those tests? Discuss how would the amount of censored data affect these tests?) Remember that the purpose of the study was to compare DFS and OS between the two groups [null hypothesis of no differences and the alternative is that the treatment group (hormone therapy) has better outcome].

Question 3:

Given your choice, test the hypothesis of no difference in overall survival and disease-free survival between the two groups. Find a point estimate and a 95% confidence interval for the relative risk of recurrence for patients at a fixed time point, say 5 years (five-year recurrence risk), in the treatment group as compared to those in the control group.

Question 4:

Address the research question on ER status above (treatment more beneficial in ER+ patients). Include in the model important prognostic variables (such as number of positive nodes, age, histologic grade and pathologic size).

Question 5:

What are the underlying model assumptions? Test the model assumptions.

Question 6:

Plot the smoothed hazard functions for the ER+ patients (hint: use sts graph function in Stata). What is the interpretation of the hazard function? What is your interpretation on what is going on with the two hazard functions in these plots?

Question 7:

How would you address the problem of non-proportional hazard functions? Please provide two different ways of addressing this issue and what would be the implications in terms of answering the research question (Do not implement your proposed suggestion).

Biostatistics Qualifier Exam
Day Two

Data Analysis Project #2
Open book
Four hours

Many dietary and biochemical epidemiologic studies have shown an inverse association between beta-carotene and the risk of cancer. These findings have increased interest in factors influencing beta-carotene levels in human plasma. As part of a large multi-center clinical trial of skin cancer prevention with beta-carotene, information on personal characteristics and plasma beta-carotene levels from 315 participants were collected.

The attached text file contains 315 observations on 13 variables. The variable names in the order from left to right:

```
AGE: Age (years)
SEX: Sex (1=Male, 2=Female).
SMOKSTAT: Smoking status (1=Never, 2=Former, 3=Current Smoker)
QUETELET: Quetelet (weight/height^2))
VITUSE: Vitamin Use (1=Yes, fairly often, 2=Yes, not often, 3=No)
CALORIES: Number of calories consumed per day.
FAT: Grams of fat consumed per day.
FIBER: Grams of fiber consumed per day.
ALCOHOL: Number of alcoholic drinks consumed per week.
CHOLESTEROL: Cholesterol consumed (mg per day).
BETADIET: Dietary beta-carotene consumed (mcg per day).
RETDIET: Dietary retinol consumed (mcg per day)
BETAPLASMA: Plasma beta-carotene (ng/ml)
```

You are asked to import the data into a statistical software package and use appropriate regression modeling approach to identify the important risk factors associated with plasma beta-carotene levels. Prepare a report using the format described in the cover sheet. In particular, provide an executive summary of your findings and detailed description of your modeling practice, i.e. exploratory data analysis, modeling building, model selection, model diagnostics, interpretation of the final model, etc. You may also include justifications of your modeling approach. With your final model, predict the plasma beta-carotene level for a 40-year-old woman who is a current smoker, using vitamin fairly often, consuming 50g fat, 30g fiber, 300mg cholesterol, 5000 mcg dietary beta-carotene, 3000 mcg dietary retinol per day and also provide a 95% confidence interval for your prediction.