Biost 518
Applied Biostatistics II

Final Examination
March 19, 2008

Name: _____ Box Nbr: _____

<u>Instructions:</u> Please provide concise answers to all questions. Rambling answers touching on topics not directly relevant to the question will tend to count against you. Nearly telegraphic writing style is permissible.

The examination is closed book and closed notes. If you come to a problem that you believe cannot be answered without making additional assumptions, <u>clearly</u> state the <u>reasonable</u> assumptions that you make, and proceed.

Problems 1-3 pertain to an analysis of data from 511 patients with primary biliary cirrhosis (PBC). Appendix A contains variable definitions and descriptive statistics for these data. The results of various regression analyses are given in the other Appendices.

1. Appendix B contains the results of six regression models (models B1, B2, B3, B4, B5, and B6) exploring the association between serum albumin and race.

    a. (5 points) Using model B1, what are the predicted values for mean serum albumin for whites, blacks, and Asians? What would be your conclusion about an association between albumin and race (provide a P value in support of your conclusion)?

    b. (5 points) Using model B2, what are the predicted values for mean serum albumin for whites, blacks, and Asians? What would be your conclusion about an association between albumin and race (provide a P value in support of your conclusion)?

    c. (5 points) Using model B3, what are the predicted values for mean serum albumin for whites, blacks, and Asians? What would be your conclusion about an association between albumin and race (provide a P value in support of your conclusion)?

d. (5 points) From the analyses provided in the Appendices, can you determine the sample means for serum albumin for whites, blacks, and Asians in this data? If so, do so. If not, explain why not.

e. (5 points) Suppose a researcher were to decide to use backward stepwise regression techniques in which all variables having a P value greater than 0.05 were dropped from the regression model. If the researcher **started with Model B1**, which of models B1 – B6 would the researcher use to report the results of the analysis? Provide a brief interpretation of the intercept and slope parameters for that model.

f. (5 points) Suppose a researcher were to decide to use backward stepwise regression techniques in which all variables having a P value greater than 0.05 were dropped from the regression model. If the researcher **started with Model B2**, which of models B1 – B6 would the researcher use to report the results of the analysis? Provide a brief interpretation of the intercept and slope parameters for that model.

g. (5 points) Suppose a researcher were to decide to use backward stepwise regression techniques in which all variables having a P value greater than 0.05 were dropped from the regression model. If the researcher **started with Model B3**, which of models B1 – B6 would the researcher use to report the results of the analysis? Provide a brief interpretation of the intercept and slope parameters for that model.

h. (10 points) In light of the above analyses, **very briefly** discuss the scientific difficulties posed by stepwise selection of models in this setting.

i. (10 points) Based on the model of your choice (specify which one), give a summary of the findings with respect to differences among the races with respect to serum albumin. What assumptions have to be satisfied for your statistical conclusions to be valid?

j. (10 points) Suppose the association between albumin and race were found to be not statistically significant (that is, assume nonsignificance no matter what you found above). Provide a "differential diagnosis" of the reasons that a such a result might be observed.

2. Appendix C contains the results of a logistic regression (using both the “logit” and “logistic” commands in Stata) performed on a variable indicating hepatomegaly as a function of albumin and nonwhite race..

    a. (5 points) Using this model, what is the estimated probability of hepatomegaly for a white subject having a serum albumin of 3?

    b. (5 points) Using this model, what is the estimated probability of hepatomegaly for a white subject having a serum albumin of 4?

    c. (5 points) Using this model, what is the estimated probability of hepatomegaly for a nonwhite subject having a serum albumin of 3?

    d. (10 points) What issues do you have to consider in deciding whether to rely on your answers to parts a-c?

    e. (5 points) What is the interpretation for the intercept in the “logit” model? What scientific use would you make of this estimate?

f. (5 points) What is the interpretation for the *albumin* slope parameter? What scientific use would you make of this estimate?

g. (5 points) What is the interpretation for the *nonwhite* slope parameter? What scientific use would you make of this estimate?

h. (10 pts) Suppose you discovered that the 511 measurements actually included up to 4 repeat measurements on just 150 individuals. How would you expect inference about race in a more appropriate model to differ from that provided in Appendix C?

i. (10 pts) Suppose you discovered that the 511 measurements actually included up to 4 repeat measurements on just 150 individuals. How would you expect inference about albumin in a more appropriate model to differ from that provided in Appendix C?

3. Appendix D contains the results of a proportional hazards regression analysis examining the association between survival and albumin and race. Use the results presented in Appendix D to answer the following questions. **For each question make certain that you identify which test and p value you use to answer the question.**

    a. (25 pts) Provide a brief interpretation of each of the regression parameter estimates.

    b. (5 pts) Is there evidence that there is a statistically significant association between survival and serum albumin?

    c. (5 pts) Is there evidence that there is a statistically significant association between survival and race?

    d. (5 pts) Is there evidence that any association between serum albumin and survival differs by race?

e. (5 pts) Is there evidence that any association between race and survival differs by albumin level?

4. A scientific colleague was examining how the relationship between C-reactive protein (CRP, a marker of inflammation) and age differed across the sexes. I would, of course, ideally wanted output from a linear regression of *crp* (C reactive protein) including terms for age (variable *age* measured in years), an indicator of male sex (variable *male*=0 for females, *male*=1 for males), and a variable *maleage*= *male* * *age*. He brought to me the following output from two linear regressions of CRP on age. From this output (he could not provide the data) he wanted to know the answer to a number of questions.

Linear regression model for females:

```
. regress crp age if male==0, robust
Linear regression                               Number of obs =    2861
                                                F(  1,  2859) =    8.10
                                                Prob > F      =  0.0045
                                                R-squared     =  0.0027
                                                Root MSE      =  5.4835
------------------------------------------------------------------------------
             |               Robust
         crp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0516989    .018163    -2.85   0.004    -.0873128   -.0160849
       _cons |   7.382912   1.346132     5.48   0.000     4.743424     10.0224
------------------------------------------------------------------------------
```

Linear regression model for males:

```
Linear regression                               Number of obs =    2072
                                                F(  1,  2070) =    1.18
                                                Prob > F      =  0.2770
                                                R-squared     =  0.0009
                                                Root MSE      =  6.9576
------------------------------------------------------------------------------
             |               Robust
         crp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0375293    .034512     1.09   0.277    -.0301526    .1052111
       _cons |  -1.354464   2.504879    -0.54   0.589    -6.266809    3.557881
------------------------------------------------------------------------------
```

a. (5 points) Supposing the researcher had fit the correct model including terms for *age, male,* and *maleage,* what would have been the estimated intercept?

b. (5 points) Supposing the researcher had fit the correct model including terms for *age, male,* and *maleage*, what would have been the estimated slope for *age*?

c. (5 points) Supposing the researcher had fit the correct model including terms for *age, male,* and *maleage,* what would have been the estimated slope for *male*?

d. (5 points) Supposing the researcher had fit the correct model including terms for *age, male,* and *maleage*, what would have been the estimated slope for male*age*?

e. (10 points) Is there a statistically significant difference between the age slope for females and the age slope for males?

f. (10 points) Suppose we had really wanted to know the association between CRP and age in the entire population, irrespective of sex. How might you approximate the slope of the age covariate if we had fit a regression model only including age to a sample that was 50% male and 50% female? Would that parameter likely indicate a statistically significant association between CRP and age in the population?

APPENDIX A: Descriptive Statistics for problems 1-3

Problems 1-4 pertain to analyses of data 511 subjects with primary biliary cirrhosis. Data is available on the following variables :

- ***age*** = age in years
- ***male*** = indicator of male sex (0= female, 1= male)
- ***race*** = coded variable for race (1= white, 2= black, 3= Asian). From this variable, four indicator variables were created:
  - ***white*** = indicator of white race (1= white, 0= black or Asian)
  - ***black***= indicator of black race (1= black, 0= white or Asian)
  - ***Asian*** = indicator of Asian race (1= Asian, 0= white or black)
  - ***nonwhite*** = indicator of nonwhite race (1= black of Asian, 0= white)
- ***hepmeg*** = indicator of hepatomegaly (1= enlarged liver, 0= not enlarged)
- ***bili*** = serum bilirubin (mg/dl)
- ***albumin*** = serum albumin (g/dl)
- ***obstime*** = time in years between start of study and the earlier of the subject's death or the time of data analysis
- ***death*** = an indicator that the subject died at the time recorded in *obstime* (0= subject still alive at time of data analysis, 1= subject observed to die)

Additional variable were constructed to model interactions between race and albumin:

- ***albWhite = albumin × white***
- ***albBlack = albumin × black***
- ***albAsian = albumin × asian***

Descriptive statistics on the entire dataset:

. tabstat age male race white black asian hepmeg bili albumin obstime death, stat(n mean sd min q max) col(s tat)

variable	N	mean	sd	min	p25	p50	p75	max
age	510	51.92	9.46	23.00	46.00	52.00	59.00	79.00
male	511	0.061	0.239	0.000	0.000	0.000	0.000	1.000
race	511	1.204	0.533	1.000	1.000	1.000	1.000	3.000
white	511	0.857	0.350	0.000	1.000	1.000	1.000	1.000
black	511	0.082	0.275	0.000	0.000	0.000	0.000	1.000
asian	511	0.061	0.239	0.000	0.000	0.000	0.000	1.000
hepmeg	487	0.283	0.451	0.000	0.000	0.000	1.000	1.000
bili	511	1.12	2.08	0.10	0.50	0.70	1.10	35.20
albumin	504	3.98	0.44	1.93	3.80	4.00	4.30	5.20
obstime	511	4.92	2.95	0.01	2.00	5.40	7.50	13.53
death	511	0.624	0.485	0.000	0.000	1.000	1.000	1.000

APPENDIX B: Linear regression analyses of serum albumin by race

MODEL B1:

```
. regress albumin asian black, robust

Linear regression                                      Number of obs =     504
                                                       F(  2,   501) =    7.41
                                                       Prob > F      =  0.0007
                                                       R-squared     =  0.0379
                                                       Root MSE      =  .43081

------------------------------------------------------------------------------
             |               Robust
     albumin |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       asian |   -.332767   .0925291    -3.60   0.000      -.51456   -.1509741
       black |  -.1317224   .0845676    -1.56   0.120    -.2978733    .0344286
       _cons |   4.006961   .0199291   201.06   0.000     3.967806    4.046115
------------------------------------------------------------------------------
```

MODEL B2:

```
. regress albumin white black, robust

Linear regression                                      Number of obs =     504
                                                       F(  2,   501) =    7.41
                                                       Prob > F      =  0.0007
                                                       R-squared     =  0.0379
                                                       Root MSE      =  .43081

------------------------------------------------------------------------------
             |               Robust
     albumin |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       white |    .332767   .0925291     3.60   0.000     .1509741      .51456
       black |   .2010447   .1221433     1.65   0.100    -.0389316    .4410209
       _cons |   3.674194   .0903575    40.66   0.000     3.496667     3.85172
------------------------------------------------------------------------------
```

MODEL B3:

```
. regress albumin white asian, robust

Linear regression                                      Number of obs =     504
                                                       F(  2,   501) =    7.41
                                                       Prob > F      =  0.0007
                                                       R-squared     =  0.0379
                                                       Root MSE      =  .43081

------------------------------------------------------------------------------
             |               Robust
     albumin |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       white |   .1317224   .0845676     1.56   0.120    -.0344286    .2978733
       asian |  -.2010447   .1221433    -1.65   0.100    -.4410209    .0389316
       _cons |   3.875238   .0821859    47.15   0.000     3.713767     4.03671
------------------------------------------------------------------------------
```

MODEL B4:

```
. regress albumin white, robust

Linear regression                                      Number of obs =     504
                                                       F(  1,   502) =   11.13
                                                       Prob > F      =  0.0009
                                                       R-squared     =  0.0304
                                                       Root MSE      =  .43204

------------------------------------------------------------------------------
             |               Robust
     albumin |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       white |   .2170975   .0650616     3.34   0.001     .0892709    .3449241
       _cons |   3.789863   .0619406    61.19   0.000     3.668168    3.911558
------------------------------------------------------------------------------
```

MODEL B5:

```
. regress albumin black, robust

Linear regression                                      Number of obs =     504
                                                       F(  1,   502) =    1.68
                                                       Prob > F      =  0.1960
                                                       R-squared     =  0.0048
                                                       Root MSE      =  .43773

------------------------------------------------------------------------------
             |               Robust
     albumin |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       black |  -.1093938   .0844854    -1.29   0.196    -.2753823    .0565947
       _cons |   3.984632   .0199178   200.05   0.000       3.9455    4.023764
------------------------------------------------------------------------------
```

MODEL B6:

```
. regress albumin asian, robust

Linear regression                                      Number of obs =     504
                                                       F(  1,   502) =   12.08
                                                       Prob > F      =  0.0006
                                                       R-squared     =  0.0310
                                                       Root MSE      =  .43191

------------------------------------------------------------------------------
             |               Robust
     albumin |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       asian |  -.3210707   .0923767    -3.48   0.001    -.5025632   -.1395782
       _cons |   3.995264   .0196275   203.55   0.000     3.956702    4.033826
------------------------------------------------------------------------------
```

APPENDIX C: Logistic regression analyses of hepatomegaly by serum albumin and race

```
. logit hepmeg albumin nonwhite, robust

Logistic regression                                     Number of obs   =        480
                                                        Wald chi2(2)    =      14.76
                                                        Prob > chi2     =     0.0006
Log pseudolikelihood = -279.83006                       Pseudo R2       =     0.0251

------------------------------------------------------------------------------
             |               Robust
      hepmeg |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     albumin |  -.2136686   .2364603    -0.90   0.366    -.6771222     .249785
    nonwhite |   .9563297   .2708027     3.53   0.000     .4255661    1.487093
       _cons |  -.2278416   .9507189    -0.24   0.811    -2.091216    1.635533
------------------------------------------------------------------------------

. logistic  hepmeg albumin nonwhite, robust

Logistic regression                                     Number of obs   =        480
                                                        Wald chi2(2)    =      14.76
                                                        Prob > chi2     =     0.0006
Log pseudolikelihood = -279.83006                       Pseudo R2       =     0.0251

------------------------------------------------------------------------------
             |               Robust
      hepmeg | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     albumin |    .807616   .1909691    -0.90   0.366     .5080771    1.283749
    nonwhite |   2.602128   .7046634     3.53   0.000     1.530457    4.424217
------------------------------------------------------------------------------
```

APPENDIX D: Proportional Hazards regression analysis on albumin, race

```
. stset obstime death
. stcox albumin asian black albAsian albBlack, robust

        failure _d:  death
  analysis time _t:  obstime

Cox regression -- no ties

No. of subjects      =          504                 Number of obs   =       504
No. of failures      =          315
Time at risk         =  2487.555601
                                                    Wald chi2(5)    =     16.38
Log pseudolikelihood =   -1793.8478                 Prob > chi2     =    0.0058

------------------------------------------------------------------------------
             |               Robust
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     albumin |   .6782549   .1102464    -2.39   0.017     .4932136     .932719
       asian |   8.959057   13.29865     1.48   0.140     .4883777    164.3497
       black |   1.745488   4.187453     0.23   0.816     .0158442    192.2927
    albAsian |   .4949193   .2017994    -1.73   0.085     .2225685    1.100538
    albBlack |   .8075404   .4980182    -0.35   0.729     .2411126    2.704635
------------------------------------------------------------------------------
```

HYPOTHESIS TEST D1:

```
. test asian black

 ( 1)  asian = 0
 ( 2)  black = 0

           chi2(  2) =    2.18
         Prob > chi2 =    0.3354
```

HYPOTHESIS TEST D2:

```
. test albAsian albBlack

 ( 1)  albAsian = 0
 ( 2)  albBlack = 0

           chi2(  2) =    3.00
         Prob > chi2 =    0.2229
```

HYPOTHESIS TEST D3:

```
. test asian black albAsian albBlack

( 1)  asian = 0
 ( 2)  black = 0
 ( 3)  albAsian = 0
 ( 4)  albBlack = 0

           chi2(  4) =    6.12
         Prob > chi2 =    0.1905
```

HYPOTHESIS TEST D4:

```
. test albumin

( 1)  albumin = 0

           chi2(  1) =    5.70
         Prob > chi2 =    0.0169
```

HYPOTHESIS TEST D5:

```
. test albumin albAsian albBlack

 ( 1)  albumin = 0
 ( 2)  albAsian = 0
 ( 3)  albBlack = 0

           chi2(  3) =   15.25
         Prob > chi2 =    0.0016
```