Biost 518
Applied Biostatistics II

Midterm Examination
February 11, 2008

Name: _____ Disc Sect: M W F

Instructions: 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.

NOTE: When you need to make calculations, always use at least four significant digits in your intermediate calculations, and report at least three significant digits. (Example: 1.045 and 0.0001234 and 1234000 each have four significant digits.)

If you come to a problem that you believe cannot be answered without making additional assumptions, clearly state the reasonable assumptions that you make, and proceed.

Please adhere to and sign the following pledge. Should you be unable to truthfully sign the pledge for any reason, turn in your paper unsigned and discuss the circumstances with the instructor.

PLEDGE:
On my honor, I have neither given nor received unauthorized aid on this examination:

Signed:_____

1. Suppose we are interested in the association between serum bilirubin, serum albumin, and the presence of edema (swelling) in a sample of patients with primary biliary cirrhosis. The following are the results of a linear regression analyses using the following variables
   - *bili:* serum bilirubin in mg/dl
   - *albumin:* serum albumin in g/dl
   - *edema:* indicator of edema (0= no, 1= yes)

```
. tabstat albumin bili, stat(n mean sd min q max) col(stat) by(edema)
Summary for variables: albumin bili
     by categories of: edema

 vaiable |  N    mean     sd    min    p25       p50       p75       max
-------------------------------------------------------------------------
→ edema == 0
albumin  |  77  3.623  .3852   2.54    3.4      3.65      3.87      4.64
   bili  |  77  2.439  3.078     .3     .8       1.2       3.2        20
-------------------------------------------------------------------------
→ edema == 1
albumin  |  23  3.098  .4204   2.27   2.74      3.13      3.41      4.06
   bili  |  23  8.643  7.528     .6    1.4       6.6      17.1      22.5
-------------------------------------------------------------------------
→ all patients
albumin  | 100  3.502  .4501   2.27  3.205     3.535      3.79      4.64
   bili  | 100  3.866  5.172     .3     .8       1.5       4.6      22.5
-------------------------------------------------------------------------
```

```
. regress bili albumin

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =   22.05
       Model |   486.49878     1   486.49878           Prob > F      =  0.0000
    Residual |  2162.02563    98   22.061486           R-squared     =  0.1837
-------------+------------------------------           Adj R-squared =  0.1754
       Total |  2648.52441    99  26.7527718           Root MSE      =   4.697

------------------------------------------------------------------------------
        bili |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     albumin |  -4.925514   1.048885    -4.70   0.000    -7.006992   -2.844035
       _cons |   21.11663   3.703417     5.70   0.000     13.76732    28.46594
------------------------------------------------------------------------------

. regress bili albumin edema

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  2,    97) =   20.72
       Model |  792.879897     2  396.439949           Prob > F      =  0.0000
    Residual |  1855.64451    97  19.1303558           R-squared     =  0.2994
-------------+------------------------------           Adj R-squared =  0.2849
       Total |  2648.52441    99  26.7527718           Root MSE      =  4.3738

------------------------------------------------------------------------------
        bili |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     albumin |  -2.706747   1.123111    -2.41   0.018     -4.93581    -.477683
       edema |   4.782688   1.195095     4.00   0.000     2.410755    7.154622
       _cons |   12.24582   4.099575     2.99   0.004     4.109298    20.38234
------------------------------------------------------------------------------
```

a. (5 points) Based on the regression model including both albumin and edema, what is the best estimate for the mean bilirubin in subjects with an albumin of 4.0 g/dl and no edema?

b. (5 points) Based on the regression model including both albumin and edema, what is the best estimate for the mean bilirubin in subjects with an albumin of 3.0 g/dl and no edema?

c. (5 points) Based on the regression model including both albumin and edema, what is the best estimate for the mean bilirubin in subjects with an albumin of 2.5 g/dl and no edema?

d. (5 points) Based on the regression model including both albumin and edema, what is the best estimate for the mean bilirubin in subjects with an albumin of 4.0 g/dl and edema present?

e. (5 points) Based on the regression model including both albumin and edema, what is the best estimate for the difference in mean bilirubin between subjects with an albumin of 4.0 g/dl and no edema and subjects with an albumin of 3.0 g/dl and no edema?

f. (5 points) Based on the regression model including both albumin and edema, what is the best estimate for the difference in mean bilirubin between subjects with an albumin of 4.0 g/dl and edema present and subjects with an albumin of 4.0 g/dl and no edema?

g. (5 points) Based on the regression model including both albumin and edema, what is the best estimate for the difference in mean bilirubin between two groups of subjects having the same edema status but who differ in serum albumin by 1.5 g/dl? Provide a confidence interval for this estimate.

h. (5 points) Provide an interpretation for the intercept in the regression model including both albumin and edema. What scientific use would you make of this estimate?

i. (5 points) Provide an interpretation for the slope for the albumin predictor in the regression model including both albumin and edema. What scientific use would you make of this estimate?

j. (5 points) Is there evidence that the slope for the albumin predictor is different from 0? State your evidence.

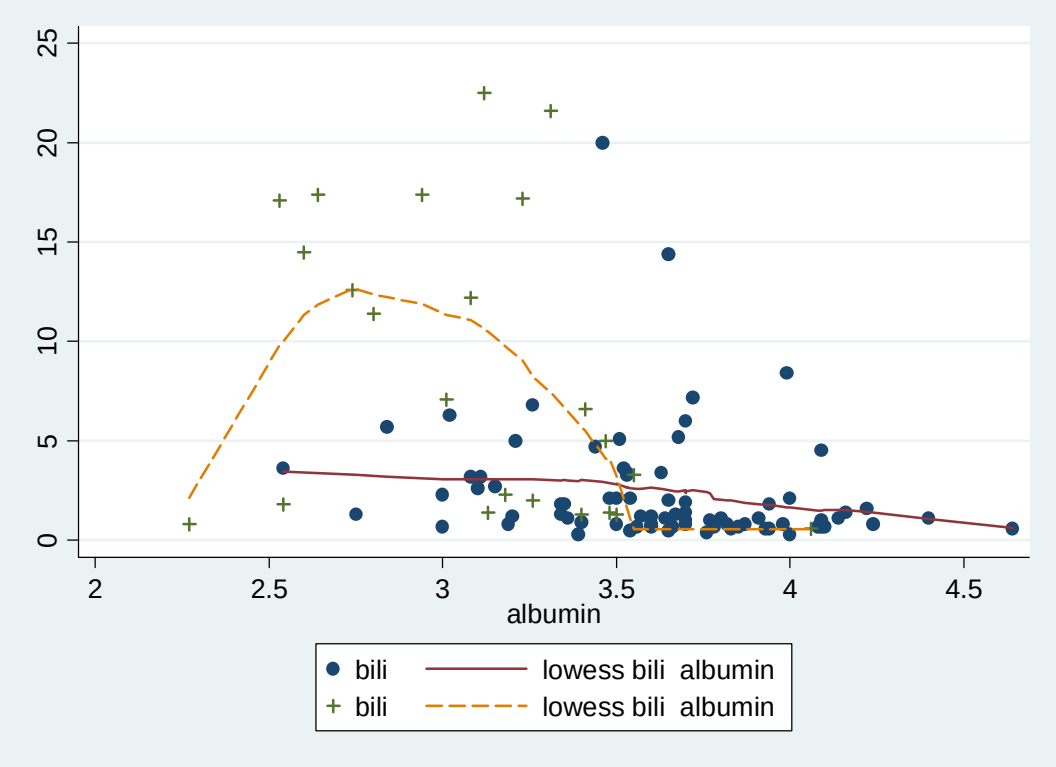
k. (5 points) Provide an interpretation for the slope for the edema predictor in the regression model including both albumin and edema. What scientific use would you make of this estimate?

l. (5 points) Is there evidence that the slope for the edema predictor is different from 0? State your evidence.

m. (5 points) Based on the regression model including both albumin and edema, what is the best estimate for the average standard deviation of bilirubin measurements within a group that is homogeneous with respect to albumin level and the presence of edema?

n. (5 points) Is there evidence that presence of edema would confound an analysis that merely considered the association between bilirubin and albumin? What would you have to consider?

2. The following scatterplot displays bilirubin (y axis) versus albumin (x axis) within strata defined by no edema (solid points and solid lowess curve) and presence of edema (points marked by + and dashed lowess curve).



a. (5 points) From this plot, comment on the reliability of your answers to parts a through d of problem 1.

b. (5 points) From this plot, comment on the reliability of your answers to parts f, g, h, i, and k of problem 1.

c. (5 points) From this plot, comment on the reliability of your answers to parts j and l of problem 1.

d. (5 points) In problem 1, I also presented a regression model of bilirubin regressed on albumin alone. How might this analysis have changed if I had used robust standard errors?

3. Now suppose we consider a log transformation of bilirubin: ***logbili= log(bili)***. Consider the following linear regression analysis.

```
. regress logbili albumin edema in 1/100

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  2,    97) =   20.65
       Model |  34.9186614     2  17.4593307            Prob > F      =  0.0000
    Residual |  81.9972366    97  .845332336           R-squared     =  0.2987
-------------+------------------------------           Adj R-squared =  0.2842
       Total |  116.915898    99  1.18096867           Root MSE      =  .91942

------------------------------------------------------------------------------
     logbili |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     albumin |  -.8377124   .2360884    -3.55   0.001    -1.306283   -.3691421
       edema |   .7307378   .2512203     2.91   0.004     .2321349    1.229341
       _cons |    3.47105   .8617694     4.03   0.000     1.760676    5.181423
------------------------------------------------------------------------------
```

a. (5 points) Provide an interpretation for the intercept in the above regression model. What scientific use would you make of this estimate?

b. (5 points) Provide an interpretation for the slope for the albumin predictor in the above regression model. What scientific use would you make of this estimate?

c. (5 points) Provide an interpretation for the slope for the edema predictor in the above regression model. What scientific use would you make of this estimate?

4. The following table presents the cross classification of a sample with respect to sex, prior cardiovascular disease, and death within 4 years (no subjects are censored).

	Females		**Males**		**All Subjects**	
	Alive	**Dead**	**Alive**	**Dead**	**Alive**	**Dead**
No CVD	2244	116	1317	174	3561	290
CVD	464	80	480	125	944	205
Total	**2708**	**196**	**1797**	**299**	**4505**	**495**

a. (10 points) Suppose we fit a logistic regression modeling the indicator of death within 4 years (response variable) as a function of a variable indicating male sex (so *male*=0 for females and *male*=1 for males). Can you find the intercept and slope for such a model? If so, do so. If not, explain the difficulty.

b. (10 points) Suppose we fit a logistic regression modeling the indicator of death within 4 years (response variable) as a function of a variable indicating male sex (so male=0 for females and male=1 for males) and an indicator of prior cardiovascular disease (so *cvd*=0 if none, *cvd=1* if so). Can you find the intercept and slopes for both the *male* and *cvd* variables for such a model? If so, do so. If not, explain the difficulty.

c. (20 points) Suppose we fit a logistic regression modeling the indicator of death within 4 years (response variable) as a function of a variable indicating male sex (so male=0 for females and male=1 for males), an indicator of prior cardiovascular disease (so *cvd*=0 if none, *cvd=1* if so) and their interaction *mcvd= male * cvd*.. Can you find the intercept and slopes for both the *male*, *cvd*, and *mcvd* variables for such a model? If so, do so. If not, explain the difficulty.

5. Suppose we are interested in the association between age and death in a population of patients recently admitted to a hospital for cardiovascular disease. Available data include

   - *age:* age of patient in years
   - *prevhosp:* an indicator that the patient had been previously hospitalized for cardiovascular disease (0= no, 1=yes)
   - *obs:* time in years that a patient was followed
   - *death:* an indicator that the patient was observed to die (0= patient was still alive at the time indicated by *obs*, 1= patient was observed to die at the time indicated by *obs*)

   The following proportional hazards analyses were performed on these data:

```
. tabstat age, by(prevhosp) stat(n mean sd min q max) col(stat)
Summary for variables: age
     by categories of: prevhosp
```

prevhosp	N	mean	sd	min	p25	p50	p75	max
0	50	72.56	4.558732	65	69	72	75	85
1	50	72.2	5.43233	65	68	72	74	89
Total	100	72.38	4.992479	65	68	72	75	89

Proportional hazards regression on age

```
. stcox age, robust
         failure _d:  death
   analysis time _t:  obs
Cox regression -- Breslow method for ties
No. of subjects      =          100                 Number of obs   =       100
No. of failures      =           70
Time at risk         =  123.0409999
                                                    Wald chi2(1)    =      2.76
Log pseudolikelihood =   -271.09993                 Prob > chi2     =    0.0966
------------------------------------------------------------------------------
             |               Robust
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.048262    .029735     1.66   0.097     .9915725    1.108192
------------------------------------------------------------------------------
```

Proportional hazards regression on age and previous hospitalization

```
. stcox age prevhosp, robust
         failure _d:  death
   analysis time _t:  obs
Cox regression -- Breslow method for ties
No. of subjects      =          100                 Number of obs   =       100
No. of failures      =           70
Time at risk         =  123.0409999
                                                    Wald chi2(2)    =     42.28
Log pseudolikelihood =   -252.31456                 Prob > chi2     =    0.0000
------------------------------------------------------------------------------
             |               Robust
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.074049   .0260873     2.94   0.003     1.024117    1.126416
    prevhosp |   5.003388   1.335928     6.03   0.000     2.964759    8.443819
------------------------------------------------------------------------------
```

a. (10 points) Based on the model that includes only age, provide the scientific conclusions you would reach about any association between time to death and age. Include estimates and inference.

b. (10 points) Based on the model that includes both age and previous hospitalization, provide the scientific conclusions you would reach about any association between time to death and age. Include estimates and inference.

c. (10 points) How would you explain any difference in your results? Is there evidence that previous hospitalization confounds the analysis of an association between age and time to death?

6. The following analysis also added a predictor ***agesqr= age ^ 2***.

```
. stcox age agesqr prevhosp, robust
         failure _d:  death
   analysis time _t:  obs
Cox regression -- Breslow method for ties
No. of subjects      =          100              Number of obs   =       100
No. of failures      =           70
Time at risk         =  123.0409999
                                                  Wald chi2(3)    =     52.77
Log pseudolikelihood =   -251.60793              Prob > chi2     =    0.0000
------------------------------------------------------------------------------
             |               Robust
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    .549638   .2950613    -1.11   0.265     .1919233    1.574077
      agesqr |   1.004503   .0035695     1.26   0.206     .9975314    1.011524
    prevhosp |   5.062474   1.341895     6.12   0.000     3.011183    8.511152
------------------------------------------------------------------------------
```

a. (10 points) Based on the above analysis, is there evidence that the effect of age on the log hazard rate is well approximated by a straight line? Explain your reasoning

b. (Bonus: 10 points) Using on the above analysis, how would you test for an association between age and survival?

7. 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 are in the population?