Biost 518
Applied Biostatistics II

Midterm Examination
February 12, 2003

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.

The examination is closed book and closed notes. 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.

Problems 1 - 5 refer to a clinical trial of methotrexate in the treatment of Primary Biliary Cirrhosis, a progressive disease of the liver which often leads to liver transplantation or death. Patients were accrued to the study over a nine year period from March 1, 1989 to March 1, 1998. They were randomized in a double blind fashion to receive either methotrexate or placebo, and then followed until they received a liver transplant, they died, or until the study ended on November 1, 2002. The variables available in this data set include the following. All variables except Obstime and Failure are measured at the time of randomization.

- **Age** = the patient's age in years
- **Male** = an indicator of the patient's sex (0 = female, 1= male)
- **Race** = a code indicating the patient's race/ethnicity (1= Caucasian, 2= Black, 3= Native American, 4= Hispanic, 5= Oriental/Pacific, 6= Mideast/Arabian, 7= Indian subcontinent, 8= Other)
- **Weight** = the patient's weight in kilograms
- **QoL** = a code indicating the patient's self-reported quality of life (1= normal health, 2= regular activity but not completely well, 3= not able to carry out regular activity, 4= confined to bed most the time, 5= in hospital most the time)
- **Bili** = patient's bilirubin in mg/dl (tends to be high in liver disease)
- **Albumin =** patient's albumin in mg/dl (tends to be low in liver disease)
- **Hepmeg** = an indicator of an enlarged liver (0= no, 1= yes)
- **Tx =** an indicator of treatment received by the patient (0= placebo, 1= methotrexate)
- **Obstime** = time of follow-up in years from start of study until death, liver transplant, or the time of data analysis, whichever comes first
- **Failure** = type of failure observed (0= none, 1= liver transplant, 2= death)

Table 1 presents selected descriptive statistics for these data.

Table 1: Descriptive statistics for 511 subjects in the study.

	n	msng	mean	std dev	min	25%ile	median	75%ile	maximum
Age	511	1	51.918	9.459	23.000	46.000	52.000	59.000	79.000
Male	511	0	0.061	0.239	0.000	0.000	0.000	0.000	1.000
Race	511	0	1.360	0.988	1.000	1.000	1.000	1.000	8.000
Weight	511	33	70.724	15.707	42.600	59.450	67.600	79.400	150.000
QoL	511	15	1.692	0.690	1.000	1.000	2.000	2.000	4.000
Bili	511	0	1.119	2.082	0.100	0.500	0.700	1.100	35.200
Albumin	511	7	3.965	0.442	1.800	3.800	4.000	4.300	5.200
Hepmeg	511	24	0.283	0.451	0.000	0.000	0.000	1.000	1.000
Tx	511	0	0.509	0.500	0.000	0.000	1.000	1.000	1.000
Obstime	511	0	4.923	2.945	0.009	2.006	5.396	7.504	13.530
Failure	511	0	0.775	0.690	0.000	0.000	1.000	1.000	2.000

1. Suppose we are interested in the association between serum bilirubin and serum albumin. The following is the Stata output from a classical linear regression (without robust standard error estimates) of serum bilirubin (as response) on serum albumin (as predictor). For this problem, assume that this analysis is entirely appropriate for all forms of inference with this data.

```
. regress bili albumin

      Source |       SS       df       MS              Number of obs =     504
-------------+------------------------------           F(  1,   502) =   69.23
       Model |  267.575754     1  267.575754           Prob > F      =  0.0000
    Residual |  1940.21185   502  3.86496385           R-squared     =  0.1212
-------------+------------------------------           Adj R-squared =  0.1194
       Total |  2207.78761   503  4.38923978           Root MSE      =   1.966

------------------------------------------------------------------------------
        bili |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     albumin |  -1.649092   .1981957    -8.32   0.000    -2.038487   -1.259697
       _cons |   7.664059   .7906476     9.69   0.000     6.110673    9.217445
------------------------------------------------------------------------------
```

   a. Based on the above regression model, what is the best estimate for the mean bilirubin in subjects having serum albumin of 3 mg/dl?

   b. Based on the above regression model, what is the best estimate for the mean bilirubin in subjects having serum albumin of 5 mg/dl?

c. Based on the above regression model, what is the best estimate for the standard deviation of bilirubin in subjects having serum albumin of 4 mg/dl?

d. Based on the above regression model, what is the best estimate for the difference in mean bilirubin between subjects having serum albumin of 3 mg/dl and subjects having serum albumin of 2 mg/dl?

e. Based on the above regression model, what is the best estimate for the difference in mean bilirubin between subjects having serum albumin of 4 mg/dl and subjects having serum albumin of 1 mg/dl?

f. Provide an interpretation for the intercept in the above regression model. What scientific use would you make of this estimate?

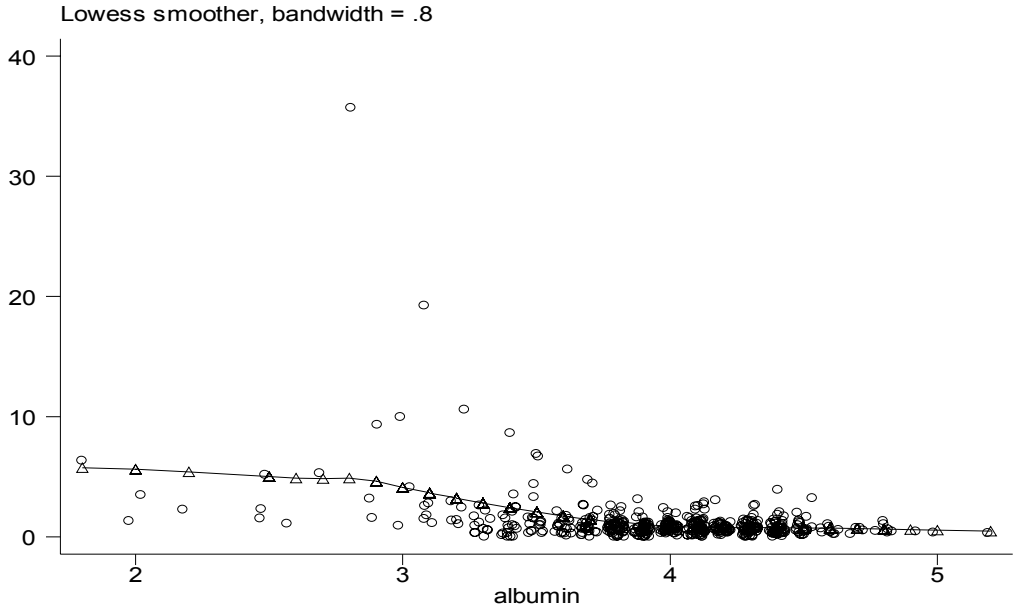
g. Provide an interpretation for the slope in the above regression model. What scientific use would you make of this estimate?

h. Is there evidence that the slope is different from 0? State your evidence.

i. Is there evidence of an association between bilirubin and albumin? Provide text suitable for inclusion in a scientific manuscript.

j. Is there a statistically significant correlation between bilirubin and albumin? State your evidence. Can you provide the correlation estimate?

2. Below is a scatterplot of serum bilirubin measurements and albumin levels with superimposed lowess curve.



Based on the appearance of this graph, do you have concerns about the validity of any of your answers to problem 1? If so, which answers and why?

3. The following is the Stata output from a regression of bilirubin (response) on sex using robust standard error estimates.

```
. regress bili male, robust

Regression with robust standard errors                Number of obs =     511
                                                      F(  1,   509) =    0.82
                                                      Prob > F      =  0.3657
                                                      R-squared     =  0.0014
                                                      Root MSE      =  2.0822

------------------------------------------------------------------------------
             |               Robust
        bili |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   .3305531    .365092     0.91   0.366    -.3867197    1.047826
       _cons |   1.098479   .0953814    11.52   0.000     .9110895    1.285869
------------------------------------------------------------------------------
```

What would be the inference derived from a t test comparing males to females? Provide the means within each group, an approximate P value, and a confidence interval for the difference between the sexes in mean bilirubin. To which version of the t test does this P value best correspond?

4. The following is the Stata output from a regression of log transformed bilirubin on age using robust standard error estimates. Use this analysis to answer the following questions.

```
. regress logbili age, robust

Regression with robust standard errors                Number of obs =     510
                                                      F(  1,   508) =    0.30
                                                      Prob > F      =  0.5815
                                                      R-squared     =  0.0006
                                                      Root MSE      =  .77991

------------------------------------------------------------------------------
             |               Robust
     logbili |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0020015   .0036293    -0.55   0.582    -.0091319    .0051289
       _cons |  -.1959544   .1886899    -1.04   0.300    -.5666631    .1747542
------------------------------------------------------------------------------
```

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

   b. Provide an interpretation for the slope in the above regression model. What scientific use would you make of this estimate?

   c. Based on the above regression model, what is the best estimate for the geometric mean bilirubin in 40 year old subjects?

   d. Is there evidence of an association between bilirubin and age? State your evidence.

e. Suppose there is no statistical evidence of an association between bilirubin and age in the above analysis. Provide four distinct reasons that such a result might be obtained. State your reasons specific to the model considered here. Please be brief.

5. Below is Stata output from a logistic regression of hepatomegaly (response) on race.

```
. logit hepmeg race

Logit estimates                                   Number of obs   =        487
                                                  LR chi2(1)      =      14.99
                                                  Prob > chi2     =     0.0001
Log likelihood = -282.80657                       Pseudo R2       =     0.0258

------------------------------------------------------------------------------
      hepmeg |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        race |   .3674674   .0958289     3.83   0.000     .1796463    .5552885
       _cons |  -1.448052   .1718669    -8.43   0.000    -1.784905    -1.1112
------------------------------------------------------------------------------
```

a. Based on the above regression model, what is the best estimate for the odds of hepatomegaly in Hispanic subjects?

b. Based on the above regression model, what is the best estimate for the probability of hepatomegaly in Hispanic subjects?

c. Why is this regression model inappropriate *a priori* to answer the scientific questions posed in parts a and b?

d. Provide an interpretation for the intercept in the above regression model. What scientific use would you make of this estimate?

e. Provide an interpretation for the slope in the above regression model. What scientific use would you make of this estimate?

f. What does this model say about an association between hepatomegaly and race? Do the problems you identified in part c above materially affect your answer to this question? That is, how can we interpret your answer regarding the existence of an association?

6. The following is Stata output from a proportional hazards regression of time until any failure (transplant or death) on bilirubin. (Note that I present two versions of the same analysis.)

```
. g Anyfail = 0
. replace Anyfail = 1 if failure > 0
(319 real changes made)

. cox obstime bili, dead(Anyfail)
Entry time 0                                    Number of obs   =        511
                                                LR chi2(1)      =       6.21
                                                Prob > chi2     =     0.0127
Log likelihood = -1826.5344                     Pseudo R2       =     0.0017

------------------------------------------------------------------------------
     obstime |
     Anyfail |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        bili |   .0538206   .0175379     3.07   0.002     .0194469    .0881943
------------------------------------------------------------------------------
```

```
. stset obstime, fail(Anyfail)
. stcox bili
No. of subjects =          511                     Number of obs   =       511
No. of failures =          319
Time at risk    =  2515.479945
                                                   LR chi2(1)      =      6.21
Log likelihood  =   -1826.5344                     Prob > chi2     =    0.0127

------------------------------------------------------------------------------
          _t |
          _d | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        bili |   1.055295   .0185077     3.07   0.002     1.019637      1.0922
------------------------------------------------------------------------------
```

Provide text suitable for inclusion in a scientific manuscript regarding the presence of an association between transplant-free survival and serum bilirubin.