# Biost 518
# Applied Biostatistics II

## Midterm Examination Key
## February 28, 2003

Name: _____ Disc Sect: M W F

**<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 - 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. (5 points each) 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?

   **Ans: 7.664059 – 1.649092 x 3 = 2.717 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?

   **Ans: 7.664059 – 1.649092 x 5 = -0.581 mg/dl** (so the straight line model is pretty bad)

   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?

   **Ans: 1.966 mg/dl (from the root MSE)**

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?

**Ans: - 1.649092 mg/dl (from the slope)**

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?

**Ans: – 1.649092 x 3 = - 4.947 mg/dl ( a 3 unit difference in the predictor)**

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

**Ans: The mean bilirubin in a population having an albumin of 0mg/dl is estimated to be 7.664 mg/dl. There is no such population, so there is no scientific use of this estimate.**

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

**Ans: The average difference in mean bilirubin between two populations that differ by 1 mg/dl in their albumin is estimated to be 1.649 mg/dl, with the group having higher albumin tending to have a lower bilirubin. This describes a first order trend in bilirubin across groups defined by albumin, and suggests an association between bilirubin and albumin. (I would not necessarily believe the straight line relationship, particularly given the answer to part b.)**

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

**Ans: The P value for the t test of a slope equal to 0 is P < .0005, so I reject the null hypothesis that the slope is 0.**

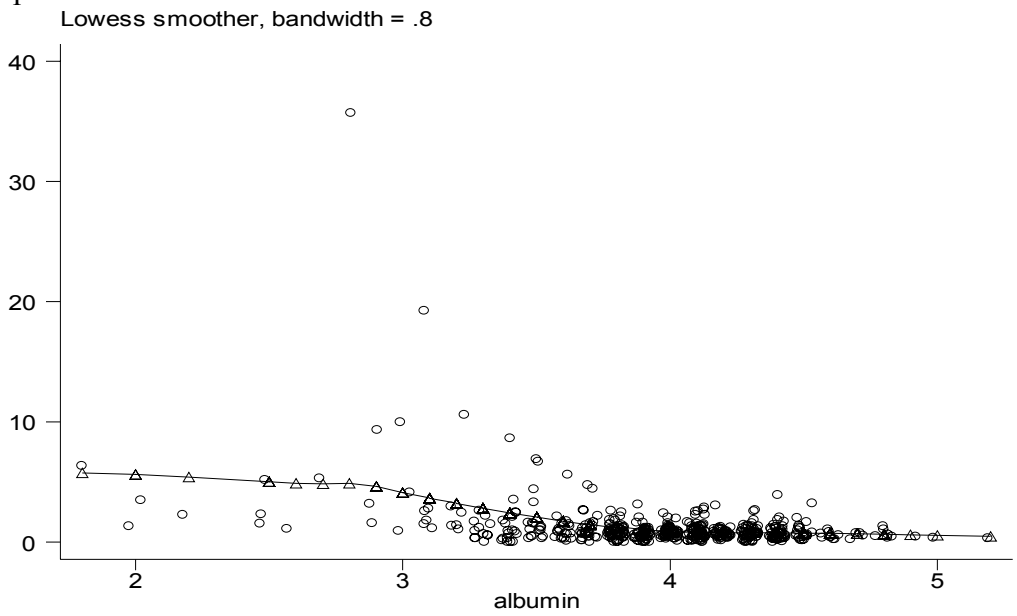
i. Is there evidence of an association between bilirubin and albumin? Provide text suitable for inclusion in a scientific manuscript.

**Ans: From linear regression, we estimate an average difference in mean bilirubin of -1.649 mg/dl (95% CI -2.038, -1.260) for each 1 mg/dl difference in albumin, an observation that is highly unusual in the absence of an association between bilirubin and albumin levels (P < .0005).**

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

**Ans: The t test of the slope is exactly the same as a test for correlation, so there is a statistically significant correlation. r = - sqrt(.1212) = - 0.348. (Note that the correlation is negative, because the slope is negative).**

2. (10 points) 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?

**Ans: The trend does not look particularly linear, so we cannot trust the estimated means for each group, though we can talk about the slope of the first order trend. There also appears to be heteroscedasticity, so I cannot trust the statistical inference (P values, CI). Given the increased variance in the groups where I have less data, I suspect the above inference is anti-conservative: Accurate CI would be wider; a proper P value would be higher.**

3. (15 points) 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?

**Ans: When using linear regression with robust standard errors with a single binary predictor, inference on the slope is approximately the same as a t-test which allows for unequal variances. Hence, given the P value of 0.366, we would not reject the hypothesis of no difference in mean bilirubin between the sexes. We estimate a mean bilirubin in females of 1.098 mg/dl and a mean bilirubin in males of 1.429 mg/dl. We estimate that the difference in mean bilirubin between males and females is 0.331 mg/dl (males higher than females), with a 95% confidence interval of -0.387 to 1.048 mg/dl.**

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

**Ans: The geometric mean of bilirubin in a population of newborns (age 0) with PBC is estimated to be exp(-.1959544) = 0.822 mg/dl. We are extrapolating way outside the range of our data, and thus I would not try to use this estimate scientifically at all.**

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

**Ans: The average ratio of geometric means of bilirubin when comparing two groups that differ in age by 1 year is estimated to be exp(-.0020015) = 0.998, hence we estimate a 0.2% lower geometric mean bilirubin for each year difference in age. I would use this estimate as indicative of a first order trend in order to assess the existence of an association between bilirubin and age.**

c. (5 points) Based on the above regression model, what is the best estimate for the geometric mean bilirubin in 40 year old subjects?

**Ans: exp( -.1959544 – 0.0020015 x 40) = 0.759 mg/dl**

d. (5 points) Is there evidence of an association between bilirubin and age? State your evidence.

**Ans: Based on the observed value of P= 0.582, I would conclude that the observed data is not atypical of what might occur when there is no true trend toward increasing or decreasing geometric mean bilirubin with age. Hence, I cannot reject the null hypothesis of no association between bilirubin and age.**

e. (10 points) 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.

**Ans: A "differential diagnosis" for a failure to reject the null includes:**

**1.) There is no association between bilirubin and age.**

**2.) There is an association between bilirubin and age, but it is not manifested in a difference in geometric means across groups.**

**3.) There is a difference in the geometric mean bilirubin across age groups, but on average the linear trend in log geometric mean is a flat line.**

**4.) There is a linear trend in the geometric mean bilirubin across age groups, but we lacked statistical precision to be able to rule out the absence of such a trend (a type II statistical error).**

***(Note: Many of you suggested confounding or effect modification as a reason for failure to detect an association between age and bilirubin. I note that the question of whether an association existed between age and bilirubin after adjustment for some other variable is inherently a different question than whether an association exists between age and bilirubin. It may well be that the adjusted question is of more scientific interest in some cases, but it is a different question and one that was not asked in this problem.)***

5. (5 points each) 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?

**Ans: exp ( - 1.448052 + 0.3674674 x 4 ) = 1.022 (Hispanics were coded as 4)**

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

**Ans: 1.022 / (1 + 1.022) = 0.505**

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

**Ans: Race is a nominal (unordered categorical) variable. There is no reason to order the races in any particular manner.**

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

**Ans: The odds of hepatomegaly in a race coded as 0 is estimated at exp (- 1.448052) = 0.235. There is no such group, hence this is scientifically meaningless.**

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

**Ans: On average, for each 1 unit difference in the coded race, the odds of hepatomegaly is exp ( 0.3674674 ) = 1.444 times higher. As there is no consistent scientific interpretation for a 1 unit difference in coded race, this estimate can only be interpreted according to whether it is different from 1.0 or not: Direction of effect makes no scientific sense.**

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?

**Ans: If there were no true differences among the races in the odds of hepatomegaly, the above analysis should tend to estimate a slope of 0. As we have a statistically significant nonzero slope (P < 0.0005), we can with high confidence state that the odds of hepatomegaly differs in some way across race, but not interpret the magnitude of any such difference. The estimates are completely meaningless. *(And, of course, this analysis would not be the most precise way in which any such associations could be detected.)***

6. (5 points) 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.

> **Ans: From the proportional hazards regression analysis, we estimate that the instantaneous risk of transplantation or death is on average 5.53% higher for each 1mg/dl difference in bilirubin level (95% confidence interval 1.96% to 9.22% higher). This observed trend is highly unusual in the absence of an association between bilirubin and transplant free survival (P = 0.002), and hence we reject the null hypothesis of no association.**

**Grade Distribution:**

Maximum Possible: 140
Highest Attained : 132

Mean (SD) : 109 (17.5)

Quantiles:

| 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% |
|---|---|---|---|---|---|---|---|---|
| 80 | 97 | 106 | 109 | 112 | 118 | 120 | 123 | 127 |