**Biost 518, Winter 2003**
**Applied Biostatistics II**
**Final Examination**
**March 19, 2003**

**Name:** _____ **Mailbox:** _____

**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.**

All problems make use of the university salary dataset which we have used over the past two quarters. Recall that this dataset contains salary data over the years 1976 to 1995 for each faculty member still employed at the university in 1995. Hence the number of cases corresponding to a given faculty member varies according to when the faculty member was originally hired. The variables available in this dataset include:

- ***case*** = case number
- ***id*** = identification number for the faculty member
- ***sex*** = M (male) or F (female)
- ***deg*** = highest degree attained: PhD, Prof (professional degree, eg, medicine or law), or Other (Master's or Bachelor's degree)
- ***yrdeg*** = year highest degree attained
- ***field*** = Arts (Arts and Humanities), Prof (professional school, ie, Business, Law, Engineering or Public Affairs), or Other
- ***startyr*** = year in which the faculty member was hired (2 digits)
- ***year*** = year (2 digits)
- ***rank*** = rank of the faculty member in this year: Assist (Assistant), Assoc (Associate), or Full (Full)
- ***admin*** = indicator of whether the faculty member had administrative duties (eg, department chair) in this year: 1 (yes), or 0 (no)
- ***salary*** = monthly salary of the faculty member in this year in dollars

I also generated a variable containing log transformed salary data, as well as indicator (dummy) variables for the various fields, and terms modeling multiplicative interactions between administrative duties and field:

- ***logslry*** = log transformed salary data (natural log, i.e., base e)
- ***artsFld*** = indicator that the faculty member's field was Arts (0= nonArts, 1= Arts)
- ***profFld*** = indicator that the faculty member's field was Prof (0= nonProf, 1= Prof)
- ***otherfFld*** = indicator that the faculty member's field was Other (0= nonOther, 1= Other)
- ***adminArts = admin * artsFld***
- ***adminProf = admin * profFld***
- ***adminOther = admin * otherFld***

Lastly, in order to address the time between a given professor being promoted from associate professor, I computed a variable that measured the length of time a faculty member spent as associate professor, also recording whether that faculty member had been promoted to full professor by 1995. I note that I set these variables to missing if the faculty member was not hired as an Assistant Professor at this university between 1976 and 1995 or if the faculty member was not promoted to Associate Professor between 1976 and 1995.

- ***yrsAssoc*** = number of years between the time a professor was promoted from Assistant to Associate at this university and the time a professor was promoted to Full or 1995, whichever comes first (this variable is set to missing if the faculty member was not promoted from Assistant to Associate at this university between 1976 and 1995)
- ***promoted*** = indicator that the time recorded in *yrsAssoc* is time to promotion to full (0= faculty member is still an Associate Professor in 1995, 1= faculty member was promoted to Full professor)

1. Consider the following regression model based on 1995 data: *salary* (response) was regressed on predictors *female, admin, artsFld, profFld, adminArts, adminProf,* and *yrdeg* using classical linear regression. For this problem, assume that this model can be used to provide valid answers to all the questions posed.

```
. regress salary female admin artsFld profFld adminArts adminProf yrdeg if
year==95

      Source |       SS       df       MS              Number of obs =    1597
-------------+------------------------------           F(  7,  1589) =  152.52
       Model |  2.6608e+09     7   380112270           Prob > F      =  0.0000
    Residual |  3.9601e+09  1589  2492214.95           R-squared     =  0.4019
-------------+------------------------------           Adj R-squared =  0.3992
       Total |  6.6209e+09  1596  4148443.26           Root MSE      =  1578.7

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -465.9419   96.30577    -4.84   0.000    -654.8416   -277.0422
       admin |   1484.752   157.7927     9.41   0.000     1175.248    1794.256
     artsFld |  -877.0102   123.3168    -7.11   0.000    -1118.891   -635.1294
     profFld |   1180.539   109.0475    10.83   0.000      966.647    1394.431
   adminArts |  -625.2884   394.6466    -1.58   0.113    -1399.371    148.7944
   adminProf |   287.5401   327.2474     0.88   0.380     -354.342    929.4222
       yrdeg |  -88.22871   4.256975   -20.73   0.000    -96.57858   -79.87883
       _cons |    12959.5   322.5268    40.18   0.000     12326.87    13592.12
------------------------------------------------------------------------------

. test artsFld profFld
 ( 1)  artsFld = 0.0
 ( 2)  profFld = 0.0
           Prob > F =    0.0011

. test adminArts adminProf
 ( 1)  adminArts = 0.0
 ( 2)  adminProf = 0.0
           Prob > F =    0.1388

. test admin adminArts adminProf
 ( 1)  admin = 0.0
 ( 2)  adminArts = 0.0
 ( 3)  adminProf = 0.0
           Prob > F =    0.0003

. test artsFld profFld adminArts adminProf
 ( 1)  artsFld = 0.0
 ( 2)  profFld = 0.0
 ( 3)  adminArts = 0.0
 ( 4)  adminProf = 0.0
           Prob > F =    0.0000
```

a. (5 points) What is the interpretation of the intercept in the above model?

b. (5 points) What is your best estimate of the expected monthly salary in a male professor in the "Other" fields with no administrative duties and who received a degree in 1990?

c. (5 points) What is your best estimate of the expected monthly salary in a female professor in the "Other" fields with no administrative duties and who received a degree in 1990?

d. (5 points) What is your best estimate of the expected difference in monthly salary between a female professor in the "Other" fields with no administrative duties and who received a degree in 1990 and a male professor in the "Other" fields with no administrative duties and who received a degree in 1990?

e. (5 points) Is there statistical evidence that the difference you reported in part d would truly be different in the population? Explicitly specify the criterion you used to answer this question.

f. (5 points) Is there statistical evidence that there is a difference between male and female faculty members' expected salaries if they were in the Arts field with administrative duties and had received their degrees in 1975? Explicitly specify the criterion you used to answer this question.

g. (5 points) What is your best estimate of the expected monthly salary in a male professor in the "Other" fields who does have administrative duties and who received a degree in 1990?

h. (5 points) What is your best estimate of the expected difference in monthly salary between a female professor in the "Other" fields who does have administrative duties and who received a degree in 1990 and a female professor in the "Other" fields with no administrative duties and who received a degree in 1990?

i. (5 points) What is your best estimate of the expected difference in monthly salary between a female professor in the professional fields who does have administrative duties and who received a degree in 1990 and a female professor in the professional fields with no administrative duties and who received a degree in 1990?

j. (5 points) Is there statistical evidence that there is a difference in expected salaries between administrators and nonadministrators? Explicitly specify the criterion you used to answer this question.

k. (5 points) Is there statistical evidence that there the difference in expected salaries between administrators and nonadministrators varies by field? Explicitly specify the criterion you used to answer this question.

l. (10 points) Is there a statistically significant difference between your answers to part h and i? Explicitly specify the criterion you used to answer this question.

2. Now consider the possibility that the necessary statistical assumptions for classical linear regression might not be satisfied. For each of the following types of questions, specify the types of violated assumptions that might pose a problem.

   a. (5 points) Detection of an association between sex and salary.

b. (5 points) Detection of a difference in mean salary between the sexes.

c. (5 points) Estimation of the expected salary in males who got their degree in 1983, were faculty members in the professional fields, and had no administrative duties. (Be explicit about the problems which might be posed by each of the variables modeled.)

d. (5 points) Prediction of the central 95% of the range of salaries in males who got their degree in 1983, were faculty members in the professional fields, and had no administrative duties.

3. Consider the following analyses related to the time to promotion of associate professors to full professors across the sexes. For each model, provide a very brief interpretation for the slope parameter of the *female* predictor, pretending that the regression model was indeed valid. Then remark on the validity of the regression model to address the question.

   a. Linear regression of *yrsAssoc* on *female* using classical linear regression.

   b. Linear regression of *yrsAssoc* on *female* using robust standard error estimates.

   c. Linear regression of log(*yrsAssoc)* on *female* using robust standard error estimates.

d. Logistic regression of *promoted* on *female* using robust standard error estimates.

e. Proportional hazards regression of *yrsAssoc* and *promoted* on *female* using classical proportional hazards regression.

f. Proportional hazards regression of *yrsAssoc* and *promoted* on *female* using robust standard error estimates.

4. Consider the following regression analyses of *logslry* both adjusted and unadjusted for rank.

```
. regress logslry female admin artsFld profFld yrdeg if year==95
      Source |       SS       df       MS              Number of obs =    1597
-------------+------------------------------           F(  5,  1591) =  238.77
       Model |  62.8109237     5  12.5621847           Prob > F      =  0.0000
    Residual |  83.7064927  1591  .052612503           R-squared     =  0.4287
-------------+------------------------------           Adj R-squared =  0.4269
       Total |  146.517416  1596  .091802893           Root MSE      =  .22937
------------------------------------------------------------------------------
     logslry |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.0731779   .0139747    -5.24   0.000    -.1005887   -.0457671
       admin |   .2128639   .0188846    11.27   0.000     .1758226    .2499052
     artsFld |  -.1437737   .0170343    -8.44   0.000    -.1771857   -.1103617
     profFld |   .1840373   .0149357    12.32   0.000     .1547415     .213333
       yrdeg |  -.0138547   .0006183   -22.41   0.000    -.0150674    -.012642
       _cons |   9.749957   .0468388   208.16   0.000     9.658085     9.84183
------------------------------------------------------------------------------
```

```
. regress logslry female admin artsFld profFld yrdeg rank if year==95
      Source |       SS       df       MS              Number of obs =    1597
-------------+------------------------------           F(  6,  1590) =  333.77
       Model |  81.6727126     6  13.6121188           Prob > F      =  0.0000
    Residual |  64.8447038  1590  .040782833           R-squared     =  0.5574
-------------+------------------------------           Adj R-squared =  0.5558
       Total |  146.517416  1596  .091802893           Root MSE      =  .20195

------------------------------------------------------------------------------
     logslry |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.0491219   .0123545    -3.98   0.000    -.0733546   -.0248891
       admin |   .1659817   .0167688     9.90   0.000     .1330904    .1988731
     artsFld |  -.1269807   .0150178    -8.46   0.000    -.1564374   -.0975239
     profFld |   .1609472   .0131936    12.20   0.000     .1350686    .1868259
       yrdeg |  -.0034121   .0007294    -4.68   0.000    -.0048429   -.0019813
        rank |   .1972355   .0091713    21.51   0.000     .1792464    .2152247
       _cons |   8.496376   .0714031   118.99   0.000     8.356322     8.63643
------------------------------------------------------------------------------

. tabulate female rank, chi2
           |                rank
    female |    Assist      Assoc       Full |     Total
-----------+---------------------------------+----------
      Male |      2588       5064       8210 |     15862
    Female |      1460       1465       1001 |      3926
-----------+---------------------------------+----------
     Total |      4048       6529       9211 |     19788

          Pearson chi2(2) = 1164.2123   Pr = 0.000
```

a. (5 points) Provide an interpretation for the slope for the sex variable in the unadjusted (first) model.

b. (5 points) Provide an interpretation for the slope for the sex variable in the adjusted (second) model.

c. (10 points) How would you characterize the role of rank in answering the question regarding sex discrimination in salaries? Provide conclusions as you might report them in a scientific paper.