Biost 518, Winter 2003
Applied Biostatistics II
Final Examination Key
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
- ***female*** = indicator of female sex (0= male, 1= 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 (two digits)
- ***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?

Ans: The mean monthly salary for males in the "Other" fields who have no administrative duties and who received their degree in 1900 is estimated to be $12,959.50. *(Clearly extrapolating way outside the range of our data.)*

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?

Ans: 12959.5 + 90 * (-88.22871) = 5018.91 *(Values of the predictors for this problem: female=0; admin=0, artsFld=0, profFld=0, adminArts=0; adminProf=0; yrdeg=90)*

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?

Ans: 12959.5 + (- 465.9419) + 90 * (-88.22871) = 4552.974 *(Values of the predictors for this problem: female=1; admin=0, artsFld=0, profFld=0, adminArts=0; adminProf=0; yrdeg=90)*

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?

Ans: - 465.942 (so females are on average paid $465.94 less per month than comparable males) *(This is just the slope for the female predictor. You could also have taken the difference between your answers to parts c and b.)*

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.

Ans: The P value for the test that the *female* slope parameter is 0 yields P < 0.0005, hence we can with high confidence conclude that there is a difference between the average monthly salaries for men and women faculty in the same field who had similar administrative duties and who received their degrees in the same year. Thus we know that this is also true when we specify the exact level of the field, administrative duties, and year of degree predictors.

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.

Ans: Same answer as part e above. *(There were no modeled interactions with sex, so just as long as we make comparisons between similar fields, administrative duties, and year of degree, our answer about a sex effect will always be the same.)*

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?

Ans: 12959.5 + 1484.752 + 90 * (-88.22871) = 6503.668 *(Values of the predictors for this problem: female=0; admin=1, artsFld=0, profFld=0, adminArts=0; adminProf=0; yrdeg=90)*

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?

Ans: 1484.75 (so administrators in the "Other" fields are on average paid $1484.75 more per month than nonadministrator females in the same field) *(This is just the slope for the admin predictor, because the "Other" fields are our reference group.)*

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?

Ans: 1484.75 + 287.54 = 1772.29 (so administrators in the professional fields are on average paid $2665.29 more per month than nonadministrator females in the same field) *(This is the slope for the admin predictor plus the slope for the adminProf predictors, because we modeled an interaction..)*

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.

Ans: The P value for the test that the *admin, adminArts,* and *adminProf* slope parameters are simultaneously 0 yields P = 0.0003, hence we can with high confidence conclude that there is a difference between the average monthly salaries for administrators and nonadministrators of the same sex and who received their degrees in the same year. ***(Note that we had to test all three of the parameters that involved administrative duties in some way, because if there were a difference in at least one of the groups, there would have to be some sort of association).***

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.

Ans: The P value for the test that the *adminArts* and *adminProf* slope parameters are simultaneously 0 yields P = 0.1388, hence we can not with high confidence conclude that there is a difference across the fields in the effect of administrative duties on the average salaries, after adjusting for sex and year of degree. ***(This is a question about effect modification. Thus we had only to test the two parameters that involved the interaction of field and administrative duties.)***

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.

Ans: Here we are asking specifically about the difference between the administrative duties effect in the "Other" fields and the professional fields. Hence, we base our decision in part on the P value for the *adminProf* slope parameter. It is probably most appropriate to adjust for the multiple comparisons inherent in modeling field as dummy variables. There are three possible comparisons that could have been made, so I use the Bonferroni adjustment of 3 * 0.38, and report that P > 0.50. ***(People would sometimes differ on the adjustment for multiple comparisons according to whether this question was prespecified exactly or not. In any case, you should explicitly state whether adjustment for multiple comparisons was or was not made.)***

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.

Ans: We only need independence among the cases and a sufficiently large sample size so that the distribution of the parameter estimates is normal. Violations of homoscedasticity pose no problem with testing the null hypothesis of "no association whatsoever", because unequal

variances mean that there is an association between the variables. Similarly, nonlinearity poses no problem, because that too would mean that there is an association. Of course, with only two levels of sex, the linearity question is moot. If the reason that the other variables were included was to adjust for confounding, if the year of degree effect were not linear, then we may not have removed all confounding. All other variables were modeled in a flexible manner, so nonlinearity is impossible with them.

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

Ans: In addition to the assumptions needed above, we now do need equal variances between the groups. Unequal variances can cause inflation of the type I error above the desired level even when the means are in fact equal. Nonlinearity poses no problem, because if there is nonlinearity in the means, there must be a difference in the means. Again I note that with only two levels for sex, nonlinearity is impossible.

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.)

Ans: In addition to the assumptions needed above, we now do need the linearity assumption to hold for all variables. This is of concern only for the year of degree variable, because dummy variables were used to model all the other predictors.

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.

Ans: In addition to the assumptions needed above, we now need to know that the distribution within each group is normal.

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.

Ans: The slope parameter is the difference between the sexes in the average value of years as Associate Professor. As *yrsAssoc* is a right censored variable, this is not scientifically valid.

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

Ans: The slope parameter is the difference between the sexes in the average value of years as Associate Professor. As *yrsAssoc* is a right censored variable, this is not scientifically valid. *(Robust standard errors do nothing to help the censoring.)*

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

Ans: The exponentiated slope parameter is the ratio between the sexes' geometric mean of years as Associate Professor. As *yrsAssoc* is a right censored variable, this is not scientifically valid.

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

Ans: The exponentiated slope parameter is the odds ratio comparing the odds of a female being promoted and the odds of a male being promoted. As the length of time each faculty member was observed differs, this is not scientifically valid.

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

Ans: The exponentiated slope parameter is the hazard ratio comparing the instantaneous rate of a female being promoted and the instantaneous rate of a male being promoted. This analysis assumes proportional hazards between the sexes, a relatively strong assumption.

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

Ans: The exponentiated slope parameter is the hazard ratio comparing the instantaneous rate of a female being promoted and the instantaneous rate of a male being promoted. This analysis does not assume proportional hazards between the sexes, and thus seems the most appropriate to me.

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.

Ans: The geometric mean salary for females in 1995 is 7.06% lower than (only 92.94% as high as) that for a male in the same field who received his degree in the same year and who has the same level of administrative duties. (95% confidence interval 4.47% lower to 9.57% lower, P < .0005).

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

Ans: The geometric mean salary for females in 1995 is 4.79% lower than that for a male of the same rank in the same field who received his degree in the same year and who has the same level of administrative duties. (95% confidence interval 2.46% lower to 7.07% lower, P < .0005).

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.

Ans: Rank is potentially in the causal pathway in a policy that discriminates against women. Hence, my report would primarily focus on the unadjusted analysis, and I would use the rank adjusted model only to amplify the possible mechanisms of discrimination:

In 1995, women tended to be paid 7.06% less than otherwise comparable men. The geometric mean salary for females in 1995 is 7.06% lower than (only 92.94% as high as) that for a male in the same field who received his degree in the same year and who has the same level of administrative duties. (95% confidence interval 4.47% lower to 9.57% lower, P < .0005). This lower rate of pay reflects both a tendency for women to be less likely to be promoted to the higher ranks, as well as to receive lower pay than comparable men having the same professorial rank: The geometric mean salary for females in 1995 is 4.79% lower than that for a male of the same rank in the same field who received his degree in the same year and who has the same level of administrative duties. (95% confidence interval 2.46% lower to 7.07% lower, P < .0005).