Bios 312:
Modern Regression Analysis

Final Examination
April 27, 2010

Name: _____

Instructions: Please provide concise answers to all questions. Questions are of varying levels of difficult, so you may find it advantageous to skip questions you find especially difficult, and return to these questions at the end of the exam.

You are allowed three (3) pages of your own notes to assist you when taking the exam.

You may use a calculator to assist with arithmetic. When making intermediate calculations, always use at least four significant digits; report at least three significant digits.

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

Please adhere to the following pledge. If you are 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

_____

This exam consists of

- 12 total pages including the Appendix of Results
- There are 120 total points
  - Question 1: parts (a) – (n), 70 points
  - Question 2: parts (a) – (c), 15 points
  - Question 3: parts (a) – (b), 10 points
  - Question 4: parts (a) – (d), 25 points

All problems make use of the University salary dataset we have discussed this quarter. Recall that the dataset contains the average monthly salary for each faculty member over the years 1976 to 1995 for each faculty member still employed at the University in 1995. The variable names in the dataset include:

- id: Subject ID
- deg: Highest degree obtained (PhD, Professional, or Other)
- yrdeg: Year of degree (2 digits)
  - Note that this variables is coded using 2 digits, not 4 (e.g. 1980 is coded as 80)
- field: Arts, Professional (e.g. Business, Law, Engineering), and Other
- startyr: Starting year (also 2 digits, not 4 digits)
- year: Year of data (1976 to 1995)
- rank: Rank of faculty in given year (Assist, Assoc, or Full)
- admin: Indicator variable for administrative duties (0=none, 1=admin duties)
- salary: Monthly salary for the given year (dollars)
- male: Indicator for male gender (0=female, 1=male)

In addition to these variables, I created the following indicator variables

- rankassoc: Indicator variable for rank of Associate Professor (0=Asst/Full, 1=Assoc)
- rankfull: Indicator variable for rank of Full Professor (0=Asst/Assoc, 1=Full)
- degother: Indicator variable for degree of Other (0=PhD/Professional, 1=Other)
- degprof: Indicator variable for degree of Professional (0=PhD/Other, 1=Professional)

I also created the following variables to model interaction between administrative duties and degree

- adminother: admin*degother
- adminprof: admin*degprof

And the log transformed version of salary

- logsalary = log(salary)

Question 1 (70 points, 5 points each). Consider the following regression model based on the 1995 salary data. Salary (response) was regressed on the predictors *male*, *yrdeg, rankassoc, rankfull, degother, degprof, admin, adminother* and *adminprof* using classical linear regression. For all parts of question 1, assume that the model can be used to provide valid answers to all questions. Regression output is given in the Appendix. Note that you will not use all of the results in the appendix to answer the following questions.

1.a) What is the interpretation of the intercept in this model? Give a precise interpretation and indicate if it has any scientific meaning.

The intercept is the expected salary when all covariates in the model are set to 0. A female Assistant professor who received her PhD degree in 1900 and has no administrative duties will have an expected salary of $6452.95 per month. The year of degree is well outside the range of the data, so there is no scientific meaning to the intercept.

1.b) What is your best estimate of the expected monthly salary for a male, full professor with a PhD degree who received his degree in 1970 and has no administrative duties?

6453 + 461 + 2163 – 23.1*70 = $7460

1.c) What is your best estimate of the expected monthly salary for a female, full professor with a PhD degree who received her degree in 1970 and has no administrative duties?

6453 + 2163 -23.1*70 = $6999

1.d) What is your best estimate for the difference in expected monthly salary for a male, full professor with a PhD degree who received in 1970 and has no administrative duties compared to a female, full professor with a PhD degree received in 1970 and has no administrative duties?

Male professors make $461 more per month

1.e) Based on the output, are you able to give a 95% confidence interval for the difference in salary you report in part (d)? If so, give the 95% CI and provide an interpretation. If not, explain the difficulty with determining the CI.

Yes. The CI is the CI for the male term, [283.4, 639.2]. We are 95% confident that the true difference in salary comparing a male and female faculty member with the specified covariates is between $283 and $639 per month.

1.f) Is there any evidence to suggest that there is any difference in male and female faculty members' salaries if both received a "Professional" degree in 1985 and were now Associate Professors with administrative duties? Explicitly give the criterion you used to answer this question.

Yes. This is the same answer as above. Controlling for rank, year of degree, and administrative duties, males have higher monthly salaries than females ($p < 0.001$)

1.g) What is your best estimate of the expected monthly salary for a female, Assistant professor with a Professional degree received in 1980 who has administrative duties?

6453 – 23.1*80 + 1269 + 807 - 646= $6035

1.h) What is your best estimate of the expected monthly salary for a female, assistant professor with a Professional degree received in 1980 who has no administrative duties?

6453 – 23.1*80 = $5412

1.i) What is your best estimate of the difference in monthly salary comparing a female, Assistant professor with a Professional degree received in 1980 who has administrative duties to a female, Assistant professor with a Professional degree received in 1980 who has no administrative duties?

1269 – 646 = 623
The professor with admin duties makes $623 more per month

1.j) What is your best estimate of the difference in monthly salary comparing a female, Assistant professor with an Other degree received in 1980 who has administrative duties to a female, Assistant professor with an Other degree received in 1980 who has no administrative duties?

1269 – 502 = 767
The professor with admin duties makes $767 more per month

1.k) Is there any statistical evidence to suggest that there exists any difference in salary between faculty who have administrative duties and faculty who do not have administrative duties? Explicitly specify the criterion you used to answer this question.

Testing if administrative duties is a significant in the model, which will be the case if any of admin, adminother, or adminprof are significant. From the output, $p < 0.001$ ($F = 30.20$) indicating that there is evidence that admin duties are associated with salary.

1.l) Is there any statistical evidence to suggest that there is a difference in expected salary between administrators and non-administrators varies by degree obtained? Explicitly specify the criterion you used to answer this question.

Testing if adminother or adminprof are significant in the model. From the output, $p = 0.1672$ so there is no evidence to suggest that the difference in expected salary due to admin duties varies by degree obtained.

1.m) Is there any statistical evidence to suggest a difference in expected salary between a faculty member with a PhD degree and administrative duties compared to a faculty member with an Other degree and administrative duties? Explicitly specify the criterion you used to answer this question.

The answer I was expecting was: From the linear regression output, there is no evidence that the adminother coefficient is different from 0 ($p = 0.241$). One person answered this way. After reading the exams and re-reading how I worded the question, I believe I was unclear. Another correct answer, given by one student, is that the relevant output is not provided as you would have to look at a linear combination of adminother and degother. For grading purposes, everyone received credit on this question, and 5 bonus points were added if it was answered correctly.

1.n) Suppose that instead of comparing two different faculty members with and without administrative duties, we were interested in explaining the change in an individual faculty member's salary due to administrative duties. Explain why this question is not answered with the given analysis.

That is a longitudinal, or within-subject comparison. In this analysis we are focusing on data from 1995, so we only have salary information measured on each faculty at one time point. Any comparisons are thus between different faculty members, not comparisons within the same faculty member.

Question 2 (15 points, 5 points each). Now consider the possibility that the necessary statistical assumptions for classic linear regression might not hold. For each of the following questions, specify the types of violated assumptions that might pose a problem with the statistical analysis.

2.a) Detection of a statistically significant difference in mean salary between male and female faculty

Assuming that we have adequately addressed confounding, the estimated mean difference in salary for males and females should be correct. However, we did not use robust standard errors, so we are assuming the homoscedasticity holds. If constant variance does not apply, it is likely that the confidence intervals are too narrow (or p-values too small), giving us anti-conservative statistical inference.

2.b) Estimation of the expected salary for a female faculty member who received her PhD degree in 1980 and is now a full professor with administrative duties.

To estimate the expected salary for a a given set of covariates, we need to have correctly modeled the mean. Most of the covariates are indicator variables, so unless we missed important interactions, we have likely modeled the mean well for those covariates. The main concern would be year of degree, which we modeled just using a linear term. Using a linear term may be fine for precision or to control for confounding, but is less likely to be accurate when estimating means. I would prefer a spline model with at least 3 or 4 degrees of freedom instead.

2.c) Prediction of the central 95% range of salary for a female faculty member who received her PhD degree in 1980 and is now a full professor with administrative duties.

In addition to correctly modeling the mean (as outline above), additionally we need to have the residuals follow a Normal distribution with the same shape (e.g. variance) at each level of the grouping variables.

Question 3 (10 points, 5 points each). For the following questions, explain what will happen to your parameter estimate in terms of bias a precision due to the particular missing data mechanism.

3.a) A database storage error leads to a random 20% of the observations becoming unreadable, so they are treated as missing in your analysis.

The data should be missing completely at random, which will result in unbiased parameter estimates. Because we have lost a large percentage of the data, are estimate will be less precise than they would be with the full dataset.

3.b) The database error removes the fist 20% of the database, which primarily contained salary data for faculty who had been at the University the longest and thus predominately had the larger salaries.

The probability of missing is related to the outcome, so it will bias our parameter estimates. Furthermore, the loss of data will decrease precision, but the bias is of greater concern.

Question 4 (25 points; 5, 5, 10, 5). Consider the following regression model based on the 1995 salary data. Log salary (response) was regressed on the predictors *male*, *yrdeg, degother, degprof,* and *admin* in models that either did not adjust for rank or adjusted for rank using indicator variable (*rankassoc, rankfull*). Regression output is given in the Appendix along with the association of gender with rank.

4.a) Provide an interpretation for the slope parameter for male in the unadjusted model.

Holding year of degree, type of degree, and admin duties constant, male faculty are expected to make exp(0.0968) = 1.101 (or 10.1%) more than female faculty members. Salary is summarized using the geometric mean when we use this model.

4.b) Provide an interpretation for the slope parameter for male in the adjusted model.

Holding year of degree, type of degree, rank, and admin duties constant, male faculty are expected to make exp(0.0702) = 1.072 (or 7.2%) more than female faculty members. Salary is summarized using the geometric mean when we use this model.

4.c) Suppose your scientific hypothesis was that gender discrimination leads to lower promotion rates which in turn lead to lower salaries for women. In other words, rank lies within your causal pathway of interest. Based on all of these analyses, provide conclusions characterizing the role of rank in answering the question regarding the effect of sex discrimination on salary disparity at this University. Summarize your conclusions as you might for a scientific paper.

Rank is in the causal pathway of interest, so my report would primarily focus on the unadjusted analysis. I would use the rank adjusted model only to amplify the possible mechanisms of discrimination.

In 1995, men tended to be paid 10.1% more than otherwise comparable women. The geometric mean salary for a male in 1995 is 10.1% higher than that for a female in the same field who received her degree in the same year and who has the same level of administrative duties (95% confidence interval 7.3% higher to 13.1% higher, P < .001). 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 males in 1995 is 7.2% higher than that for a female 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 4.9% higher to 9.7% higher, P < .001).

4.d) Suppose the salary dataset included an additional variable indicating the reflux disease status (0=no disease, 1=diseased) for all faculty members. Further assume that having reflux disease (i.e. heartburn) is known to be strongly associated with gender, and we can safely presume is not associated with salary. If we adjusted for reflux disease in this analysis, how would that impact the statistical inference for gender?

We would be including a covariate that is correlated with our predictor, but is not associated with the outcome. Any time we do this in linear regression, it leads to variance inflation. Our standard error estimate for the gender effect effect would increase leading to a wider CI and larger p-values (anti-conservative inference).

Appendix for Question 1: Linear Regression Results

```
. regress salary male yrdeg rankassoc rankfull degother degprof admin adminother
adminpro
> f if year==95

      Source |       SS       df       MS              Number of obs =    1597
-------------+------------------------------           F(  9,  1587) =  155.16
       Model |  3.0990e+09     9   344335822           Prob > F      =  0.0000
    Residual |  3.5219e+09  1587  2219214.27           R-squared     =  0.4681
-------------+------------------------------           Adj R-squared =  0.4650
       Total |  6.6209e+09  1596  4148443.26           Root MSE      =  1489.7

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   461.2795   90.69516     5.09   0.000     283.3845    639.1744
       yrdeg |  -23.10927   5.479962    -4.22   0.000      -33.858   -12.36054
   rankassoc |   332.3686   121.4962     2.74   0.006     94.05859    570.6786
    rankfull |   2162.968   139.3667    15.52   0.000     1889.606     2436.33
    degother |  -634.0606   140.7909    -4.50   0.000    -910.2162   -357.9049
     degprof |   806.7996   169.1032     4.77   0.000     475.1105    1138.489
       admin |   1269.412   137.8633     9.21   0.000     998.9982    1539.825
  adminother |  -502.2977   428.5841    -1.17   0.241    -1342.948    338.3529
   adminprof |  -646.3103    403.546    -1.60   0.109     -1437.85    145.2289
       _cons |   6452.953   503.4613    12.82   0.000     5465.434    7440.472
------------------------------------------------------------------------------

.
.
. test rankassoc rankfull

 ( 1)  rankassoc = 0
 ( 2)  rankfull = 0

       F(  2,  1587) =  191.14
            Prob > F =    0.0000

.
. test male rankassoc rankfull

 ( 1)  male = 0
 ( 2)  rankassoc = 0
 ( 3)  rankfull = 0

       F(  3,  1587) =  143.15
            Prob > F =    0.0000

.
. test degother degprof

 ( 1)  degother = 0
 ( 2)  degprof = 0

       F(  2,  1587) =   23.65
            Prob > F =    0.0000
```

```
. test male degother degprof

 ( 1)  male = 0
 ( 2)  degother = 0
 ( 3)  degprof = 0

       F(  3,  1587) =   26.01
            Prob > F =    0.0000

.
. test rankassoc rankfull degother degprof

 ( 1)  rankassoc = 0
 ( 2)  rankfull = 0
 ( 3)  degother = 0
 ( 4)  degprof = 0

       F(  4,  1587) =  115.52
            Prob > F =    0.0000

.
. test male rankassoc rankfull degother degprof

 ( 1)  male = 0
 ( 2)  rankassoc = 0
 ( 3)  rankfull = 0
 ( 4)  degother = 0
 ( 5)  degprof = 0

       F(  5,  1587) =  104.31
            Prob > F =    0.0000

.
. test admin adminother adminprof

 ( 1)  admin = 0
 ( 2)  adminother = 0
 ( 3)  adminprof = 0

       F(  3,  1587) =   30.20
            Prob > F =    0.0000

.
. test admin rankassoc rankfull degother degprof

 ( 1)  admin = 0
 ( 2)  rankassoc = 0
 ( 3)  rankfull = 0
 ( 4)  degother = 0
 ( 5)  degprof = 0

       F(  5,  1587) =  123.79
            Prob > F =    0.0000

.
```

```
. test adminother adminprof

 ( 1)  adminother = 0
 ( 2)  adminprof = 0

       F(  2,  1587) =    1.79
            Prob > F =    0.1672

.
. test admin adminother adminprof rankassoc rankfull degother degprof

 ( 1)  admin = 0
 ( 2)  adminother = 0
 ( 3)  adminprof = 0
 ( 4)  rankassoc = 0
 ( 5)  rankfull = 0
 ( 6)  degother = 0
 ( 7)  degprof = 0

       F(  7,  1587) =   89.42
            Prob > F =    0.0000

.

. lincom admin + adminother

 ( 1)  admin + adminother = 0

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   767.1139    407.476     1.88   0.060     -32.1339    1566.362
------------------------------------------------------------------------------

.
. lincom admin + adminprof

 ( 1)  admin + adminprof = 0

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   623.1012   379.4114     1.64   0.101    -121.0991    1367.302
------------------------------------------------------------------------------
```

Appendix for Question 4: Linear regression with robust standard errors

```
. regress logsalary male yrdeg degother degprof admin if year==95, robust

Linear regression                                      Number of obs =    1597
                                                       F(  5,  1591) =  198.41
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.3603
                                                       Root MSE      =  .24272

------------------------------------------------------------------------------
             |               Robust
   logsalary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   .0968183   .0134115     7.22   0.000     .0705122    .1231245
       yrdeg |  -.0135732   .0006604   -20.55   0.000    -.0148685   -.0122779
    degother |  -.1627948   .0197673    -8.24   0.000    -.2015676   -.1240221
     degprof |   .0768916   .0254583     3.02   0.003     .0269562     .126827
       admin |   .2117278   .0189419    11.18   0.000     .1745742    .2488814
       _cons |   9.663533   .0556481   173.65   0.000     9.554381    9.772684
------------------------------------------------------------------------------

. regress logsalary male yrdeg degother degprof admin rankassoc rankfull  if
year==95, ro bust

Linear regression                                      Number of obs =    1597
                                                       F(  7,  1589) =  256.79
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.5212
                                                       Root MSE      =  .21012

------------------------------------------------------------------------------
             |               Robust
   logsalary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   .0702362   .0113602     6.18   0.000     .0479535    .0925188
       yrdeg |  -.0025019   .0007864    -3.18   0.001    -.0040444   -.0009595
    degother |  -.0967089   .0168383    -5.74   0.000    -.1297365   -.0636814
     degprof |   .1079014   .0218217     4.94   0.000     .0650991    .1507036
       admin |   .1604003   .0169345     9.47   0.000     .1271841    .1936166
   rankassoc |    .095727   .0161339     5.93   0.000     .0640809     .127373
    rankfull |   .3860021   .0191854    20.12   0.000     .3483708    .4236335
       _cons |   8.607883   .0711596   120.97   0.000     8.468306    8.747459
------------------------------------------------------------------------------

. tabulate rank male if year==95, chi2

           |         male
      rank |         0          1 |     Total
-----------+----------------------+----------
    Assist |       145        170 |       315
     Assoc |       138        299 |       437
      Full |       126        719 |       845
-----------+----------------------+----------
     Total |       409      1,188 |     1,597

          Pearson chi2(2) = 127.8958   Pr = 0.000
```