Biost 518: Applied Biostatistics II
Biost 515: Biostatistics II
Emerson, Winter 2015

Homework #6 Key
March 4, 2015

Written problems: To be submitted as a MS-Word compatible file to the class Catalyst dropbox by 9:30 am on Wednesday, March 11, 2014. See the instructions for peer grading of the homework that are posted on the web pages.

*On this (as all homeworks) Stata / R code and unedited Stata / R output is **TOTALLY** unacceptable. Instead, prepare a table of statistics gleaned from the Stata output. The table should be appropriate for inclusion in a scientific report, with all statistics rounded to a reasonable number of significant digits. (I am interested in how statistics are used to answer the scientific question.)*

Unless explicitly told otherwise in the statement of the problem, in all problems requesting "statistical analyses" (either descriptive or inferential), you should present both

1. ***Methods:** A brief sentence or paragraph describing the statistical methods you used. This should be using wording suitable for a scientific journal, though it might be a little more detailed. A reader should be able to reproduce your analysis. DO NOT PROVIDE Stata OR R CODE.*
2. ***Inference:** A paragraph providing full statistical inference in answer to the question. Please see the supplementary document relating to "Reporting Associations" for details.*

Problems 1-3 of the homework relate to the dataset regarding MRI measurements of cerebral atrophy in elderly Americans (mri.doc and mri.txt). In this homework we will focus primarily on associations between mortality and serum LDL as possibly modified by race.

1. Suppose we are interested in exploring whether any association between time to death and serum LDL is adequately modeled by a relationship in which the log hazard function is linear in LDL. I ask you to compare several different alternative models that allow nonlinearity. In part f, I ask you to plot fitted HR estimates from each of these models on the same axis. In order to have comparability across models, we need to use the same reference group:

    - In all parts of this problem where you need to divide the LDL values into intervals, use 70, 100, 130, and 160 mg/dL as breakpoints for the LDL measurements. Stata commands that might be used are:

```
egen ldlctg= cut(ldl), at(0,70,100,130,160,400)
mkspline sldlA 70 sldlB 100 sldlC 130 sldlD 160 sldlE = ldl
```

    - In all parts of this problem where you model LDL continuously, we will use 1 mg/dL as the reference group (this will accommodate the log transformation). Thus you might create variables in Stata:

```
g logldl= log(ldl)
g cldl= ldl – 1
g cldlsqr= cldl^2
g cldlcub= cldl^3
```

Comments on this key: I included Stata output in order that you could explicitly see the analyses I ran to answer the questions. I did not expect you to include such output. More importantly, I expected you not to include such output.

a. Fit a regression model in which you test for a linear relationship using a step function as an alternative model. Briefly describe the model you fit and the parameters you evaluated to test the hypothesis that there were no departures from linearity. Provide a two-sided p value of the test. (Save fitted values for use in part f).

Instructions for grading: This problem is worth 5 points.

Answer: We fit a proportional hazards model relating the potentially censored observations of time to death to a linear predictor involving LDL modeled both continuously and as dummy variables. We use the Huber-White sandwich estimator to compute the standard errors. We test for departures from a linear relationship between the log hazard and LDL by simultaneously testing that all regression parameters for the dummy variables are 0 in a multiple partial chi squared test. The p value from such a test is 0.3609. We cannot detect departures a linear association.

```
. stcox cldl i.ldlctg, robust
Cox regression -- Breslow method for ties
No. of subjects      =          725                 Number of obs    =       725
No. of failures      =          131
Time at risk         =      1309564
                                                    Wald chi2(5)     =     15.84
Log pseudolikelihood =   -834.66339                 Prob > chi2      =    0.0073

             |               Robust
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cldl |   .9956353   .0087984    -0.49   0.621     .9785394     1.01303
             |
      ldlctg |
         70  |   .4564555   .2098229    -1.71   0.088     .1854041    1.123771
        100  |   .5078224   .3164298    -1.09   0.277     .1497335    1.722285
        130  |   .4294845   .3606346    -1.01   0.314     .0828324    2.226868
        160  |   .4649654   .5437737    -0.65   0.513     .0469836     4.60145

. testparm i.ldlctg*
 ( 1)  70.ldlctg = 0
 ( 2)  100.ldlctg = 0
 ( 3)  130.ldlctg = 0
 ( 4)  160.ldlctg = 0

           chi2(  4) =    4.35
         Prob > chi2 =    0.3609
```

b. Fit a regression model in which you test for a linear relationship using a quadratic polynomial as an alternative model. Briefly describe the model you fit and the parameters you evaluated to test the hypothesis that there were no departures from linearity. Provide a two-sided p value of the test. (Save fitted values for use in part f).

Instructions for grading: This problem is worth 5 points.

Answer: We fit a proportional hazards model relating the potentially censored observations of time to death to a linear predictor involving LDL modeled both as a linear continuous term and as a squared term. We use the Huber-White sandwich estimator to compute the standard errors. We test for departures from a linear relationship between

the log hazard and LDL by testing that the regression parameter for the squared term is 0 (we can use the partial Z test in the coefficient table). The p value from such a test is 0.055. We cannot detect departures a linear association using a 0.05 level of significance.

```
. stcox cldl cldlsqr, robust
Cox regression -- Breslow method for ties

No. of subjects      =          725                 Number of obs   =       725
No. of failures      =          131
Time at risk         =      1309564
                                                    Wald chi2(2)    =     15.28
Log pseudolikelihood =   -835.23455                 Prob > chi2     =    0.0005
                                          Robust
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cldl |   .9743794   .0094555    -2.67   0.007      .956022    .9930892
      cldlsqr |   1.000076   .0000398     1.92   0.055     .9999984    1.000154
```

c. Fit a regression model in which you test for a linear relationship using a cubic polynomial as an alternative model. Briefly describe the model you fit and the parameters you evaluated to test the hypothesis that there were no departures from linearity. Provide a two-sided p value of the test. (Save fitted values for use in part f).

***Instructions for grading*: This problem is worth 5 points.**

Answer: We fit a proportional hazards model relating the potentially censored observations of time to death to a linear predictor involving LDL modeled as all three of a continuous linear term, a squared term, and a cubic term. We use the Huber-White sandwich estimator to compute the standard errors. We test for departures from a linear relationship between the log hazard and LDL by simultaneously testing that the regression parameters for the squared and cubic terms are 0 in a multiple partial chi squared test. The p value from such a test is 0.0164. Using a 0.05 level of significance, we can with high confidence state that the association between the log hazard for death and LDL is nonlinear.

```
. stcox cldl cldlsqr cldlcub, robust
Cox regression -- Breslow method for ties

No. of subjects      =          725                 Number of obs   =       725
No. of failures      =          131
Time at risk         =      1309564
                                                    Wald chi2(3)    =     28.89
Log pseudolikelihood =   -835.04443                 Prob > chi2     =    0.0000
                                          Robust
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cldl |   .9590917   .0177986    -2.25   0.024     .9248339    .9946184
      cldlsqr |   1.000223   .0001933     1.15   0.248     .9998443    1.000602
      cldlcub |   .9999996   6.01e-07    -0.68   0.497     .9999984    1.000001

. test cldlsqr cldlcub
 ( 1)  cldlsqr = 0
 ( 2)  cldlcub = 0

           chi2(  2) =    8.22
         Prob > chi2 =    0.0164
```

d. Fit a regression model in which you test for a linear relationship using linear splines as an alternative model. Briefly describe the model you fit and the parameters you

evaluated to test the hypothesis that there were no departures from linearity. Provide a two-sided p value of the test. (Save fitted values for use in part f).

***Instructions for grading*: This problem is worth 5 points.**

Answer: We fit a proportional hazards model relating the potentially censored observations of time to death to a linear predictor involving LDL modeled both continuously and as linear splines. We use the Huber-White sandwich estimator to compute the standard errors. We test for departures from a linear relationship between the log hazard and LDL by simultaneously testing that all regression parameters for the linear splines are 0 in a multiple partial chi squared test. The p value from such a test is 0.1191. We cannot detect departures a linear association.

```
. stcox cldl sldl*
Cox regression -- Breslow method for ties
No. of subjects      =          725                 Number of obs    =       725
No. of failures      =          131
Time at risk         =      1309564
                                                    Wald chi2(5)     =     30.15
Log pseudolikelihood =   -834.63381                 Prob > chi2      =    0.0000

                |               Robust
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cldl |   .9938934   .0146726    -0.41   0.678     .9655476    1.023071
       sldlA |   .9842211   .0171546    -0.91   0.361     .9511666    1.018424
       sldlB |   .9853129   .0202419    -0.72   0.471     .9464277    1.025796
       sldlC |   1.005231   .0172634     0.30   0.761     .9719582    1.039642
       sldlD |   1.004187   .0240484     0.17   0.861     .9581421    1.052445
       sldlE |  (omitted)

. testparm sldl*
 ( 1)  sldlA = 0
 ( 2)  sldlB = 0
 ( 3)  sldlC = 0
 ( 4)  sldlD = 0

           chi2(  4) =    7.34
         Prob > chi2 =    0.1191
```

e. Fit a regression model in which you test for a linear relationship using a logarithmic transformation as an alternative model. Briefly describe the model you fit and the parameters you evaluated to test the hypothesis that there were no departures from linearity. Provide a two-sided p value of the test. (Save fitted values for use in part f).

***Instructions for grading*: This problem is worth 5 points.**

Answer: We fit a proportional hazards model relating the potentially censored observations of time to death to a linear predictor involving LDL modeled both as an untransformed linear continuous term and as a logarithmically transformed term. We use the Huber-White sandwich estimator to compute the standard errors. We test for departures from a linear relationship between the log hazard and LDL by testing that the regression parameter for the log transformed term is 0 (we can use the partial Z test in the coefficient table). The p value from such a test is 0.004. We thus can with high confidence reject the hypothesis that the association between the log hazard of death and LDL is linear..

```
. stcox cldl logldl
```

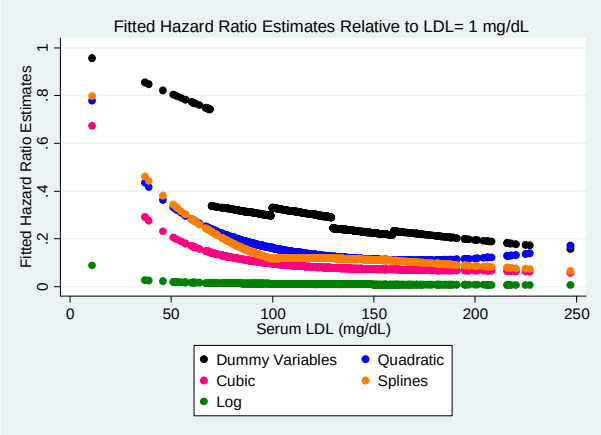
```
Cox regression -- Breslow method for ties

No. of subjects      =          725                 Number of obs    =       725
No. of failures      =          131
Time at risk         =      1309564
                                                    Wald chi2(2)     =     33.15
Log pseudolikelihood =   -835.33247                 Prob > chi2      =    0.0000
                                 Robust
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cldl |   1.002289   .005049     0.45   0.650      .992442    1.012234
      logldl |   .3595974   .1264391    -2.91   0.004     .1805191    .7163249
```

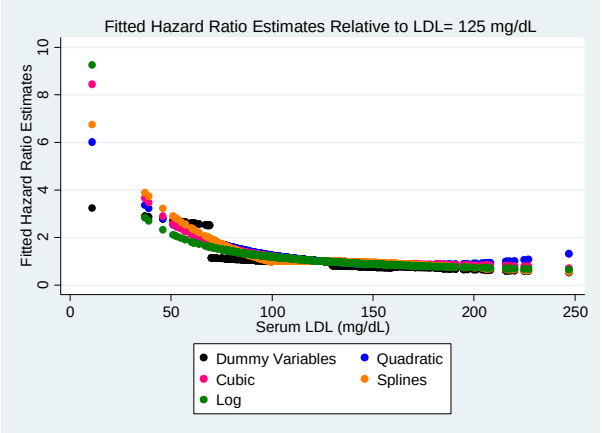
f. On the same set of axes, plot the fitted values from each of the above models, as well as a model that includes only the (centered) serum LDL values. Comment on the similarity and/or differences among these models. How might these results guide your choice of a particular model when investigating whether associations are not well described by a linear relationship?

***Instructions for grading*: This problem is worth 5 points.**

Answer: The following left hand graph displays the fitted values relative to the reference group of LDL of 1mg/dL. Except for the dummy variable alternative, it can be seen that the general shape of the curves are similar. This would argue that it is not too important which alternative curves you consider when exploring the possibility that associations are nonlinear.

To further explore the differences in the curve, we just need to consider where they disagree the most: Each of the curves extrapolates to the lowest values of LDL in a different manner. Because we were using an impossibly low value of 1 mg/dL as our reference HR of 1, this means that we cannot rely on the descriptive nature of these curves. We would really want to compare to some value that is in the middle of our data. The mean and median LDL is approximately 125 mg/dL, so I will re-calibrate our HR estimates to refer to that group. This can be done with the following code in which I first find the HR estimates for subjects with LDL = 125 mg/dL, and then I divide each fitted value by the corresponding value. When I plot the fitted values after such re-calibration, I get the right hand graph, which shows very similar fitted values in the center of the data. There are still marked differences at the lowest values of LDL, which highlights the difficulty of extrapolating with nonlinear functions: Our estimates of linear trends will tend to capture the relationships in the center of our data, but will sometimes diverge in the presence of nonlinearity. But different nonlinear curves can be in great disagreement outside the middle of our data. *(Hence my emphasis in first describing general first order trends, then trying to detect nonlinearities, then restricting further studies to areas where relationships are more linear.)*





2. Consider again a model exploring the associations between time to death and serum LDL when using linear splines.

    a. Explain the interpretation of the regression parameters in such a model.

***Instructions for grading*: This problem is worth 5 points.**

Answer: We fit a proportional hazards model relating the potentially censored observations of time to death to a linear predictor involving LDL modeled as linear splines with knots at 70, 100, 130, and 160 mg/dL and obtain the following output. The interpretation of the regression parameters is thus:

- **When comparing two groups having different levels of LDL with both levels below 70 mg/dL, the hazard ratio comparing the two groups is 0.9791 per 1 mg/dL difference in the LDL values, with the higher LDL group expected to have the lower hazard (given by the slope for sldlA).**

- **When comparing two groups having different levels of LDL with both levels between 70 and 100 mg/dL, the hazard ratio comparing the two groups is 0.9773 per 1 mg/dL difference in the LDL values, with the higher LDL group expected to have the lower hazard (given by the slope for sldlB).**

- **When comparing two groups having different levels of LDL with both levels between 100 and 130 mg/dL, the hazard ratio comparing the two groups is 1.000 per 1 mg/dL difference in the LDL values, with the higher LDL group expected to have the very slightly higher hazard (given by the slope for sldlC).**

- **When comparing two groups having different levels of LDL with both levels between 130 and 160 mg/dL, the hazard ratio comparing the two groups is 0.9771 per 1 mg/dL difference in the LDL values, with the higher LDL group expected to have the lower hazard (given by the slope for sldlD).**

- **When comparing two groups having different levels of LDL with both levels above 160 mg/dL, the hazard ratio comparing the two groups is 0.9943 per 1 mg/dL difference in the LDL values, with the higher LDL group expected to have the lower hazard (given by the slope for sldlE).**

```
. stcox sldl*
Cox regression -- Breslow method for ties
No. of subjects      =          725                 Number of obs    =       725
No. of failures      =          131
Time at risk         =      1309564
                                                    Wald chi2(5)     =     30.09
Log pseudolikelihood =    -834.5659                 Prob > chi2      =    0.0000
                                 Robust
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       sldlA |   .9790987   .0095063    -2.18   0.030      .960643    .9979091
       sldlB |   .9772977   .0137172    -1.64   0.102      .950779    1.004556
       sldlC |   1.000197   .0110792     0.02   0.986      .978716    1.022149
       sldlD |   .9971346   .0122882    -0.23   0.816     .9733387    1.021512
       sldlE |   .9942911   .0141451    -0.40   0.687     .9669502    1.022405
```

b. Is there evidence that the association between time to death and serum LDL is truly U-shaped? Explain your evidence.

***Instructions for grading*: This problem is worth 5 points.**

Answer: To have a U-shaped association, the slopes at the extreme values would have to be opposite in magnitude (i.e., one HR greater than 1 and the other HR lower than 1). Then to be confident that such an observation is not due to random variation, each of those HR estimates would have to be statistically significant. In this case, the HR estimates at the extremes are each less than 1, so there is not strong evidence of a U-shaped association.

3. Suppose we are interested in exploring the associations between time to death and serum LDL as possibly modified by race. In this problem you do not need to provide formal description of the methods or inference, though I do ask at times for specific inferential quantities.

    a. Fit a model of time to death regressed on a log transformation of serum LDL, race, and their interaction. Provide an explicit interpretation of each parameter in your model (be sure to include the actual numeric value in your interpretation, but you do not have to provide CI or p values for this part).

***Instructions for grading*: This problem is worth 15 points.**

Answer: We fit a proportional hazards model relating the potentially censored observations of time to death to a linear predictor involving log transformed LDL, dummy variables for race, and their multiplicative interactions. Interpretation of the various regression parameters includes:

- **The *baseline hazard function* (not shown) relates to whites having an LDL of 1 mg/dL (so log LDL= 0).**
- **The *2.race* parameter estimates that among subjects having an LDL of 1 mg/dL, blacks have a risk of death that is 0.154 times that of whites.**
- **The *3.race* parameter estimates that among subjects having an LDL of 1 mg/dL, Asians have a risk of death that is 305 times that of whites.**
- **The *4.race* parameter estimates that among subjects having an LDL of 1 mg/dL, Asians have a risk of death that is 333 million times that of whites.** *(We clearly do*

not have much precison in an estimate in such a small racial group having LDL measurements that are largely incompatible with life.)

- **The *logldl* parameter estimates that among white subjects every 2.718-fold higher LDL measurement is associated with a 53.9% lower risk of death.**
- **The *2.race#c.logldl* parameter is a ratio of HRs that estimates that the HR associated with a 2.718-fold higher LDL measurement in blacks is 1.55 times the HR associated with a 2.718-fold higher LDL measurement in whites.** *(Alternatively, the HR comparing blacks to whites at some specified LDL level is 1.55 times the HR comparing blacks to whites at an LDL level that is 2.718-fold higher.)*
- **The *3.race#c.logldl* parameter is a ratio of HRs that estimates that the HR associated with a 2.718-fold higher LDL measurement in Asians is 0.310 times the HR associated with a 2.718-fold higher LDL measurement in whites.** *(Alternatively, the HR comparing Asians to whites at some specified LDL level is 0.310 times the HR comparing Asians to whites at an LDL level that is 2.718-fold higher.)*
- **The *4.race#c.logldl* parameter is a ratio of HRs that estimates that the HR associated with a 2.718-fold higher LDL measurement in other races is 0.0179 times the HR associated with a 2.718-fold higher LDL measurement in whites.** *(Alternatively, the HR comparing other races to whites at some specified LDL level is 0.0179 times the HR comparing other races to whites at an LDL level that is 2.718-fold higher.)*

```
. stcox i.race##c.logldl, robust
Cox regression -- Breslow method for ties
No. of subjects      =          725                 Number of obs   =       725
No. of failures      =          131
Time at risk         =      1309564
                                                    Wald chi2(7)    =     70.63
Log pseudolikelihood =   -831.99421                 Prob > chi2     =    0.0000

------------------------------------------------------------------------------
             |               Robust
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        race |
          2  |   .1544611   .6242693    -0.46   0.644     .0000561    425.6295
          3  |   304.9796    1087.84     1.60   0.109     .2806063      331470
          4  |   3.33e+08   2.49e+09     2.62   0.009     142.8546    7.77e+14
             |
      logldl |   .4609821   .1013526    -3.52   0.000     .2995965    .7093022
             |
        race#|
    c.logldl |
          2  |   1.552511   1.309397     0.52   0.602     .2972533    8.108544
          3  |   .3101995   .2440707    -1.49   0.137      .066359    1.450048
          4  |   .0179322   .0304455    -2.37   0.018     .0006434    .4998192
```

b. Use the regression analysis in part a to perform a statistical test of the hypothesis that race does not modify the association between time to death and serum LDL. Make clear which parameters you test and provide a two-sided p value.

***Instructions for grading*: This problem is worth 5 points.**

Answer: **We test that all parameters that involve some aspect of effect modification between race and LDL are simultaneously 0. Hence we test the "interactions" between race and log ldl. The p value from such a test is P = .0452. We can thus state with some confidence (marginally significant at .05 level of significance) that there is some difference between the races in the association between LDL and survival.** *(Equivalently, I could have said that the association between race and survival differs across the various levels of LDL.)*

```
. testparm i.race#c.logldl
 ( 1)  2.race#c.logldl = 0
 ( 2)  3.race#c.logldl = 0
 ( 3)  4.race#c.logldl = 0

           chi2(  3) =    8.04
         Prob > chi2 =    0.0452
```

c. Use the regression analysis in part a to perform a statistical test of the hypothesis that there is no association between time to death and serum LDL. Make clear which parameters you test and provide a two-sided p value.

***Instructions for grading*: This problem is worth 5 points.**

Answer: **We test that all parameters that involve some aspect of LDL are simultaneously 0. Hence we test the "main effects" for log LDL and the "interactions" between race and log LDL. The p value from such a test is P < 0.0001. We can thus state with high confidence that there is some difference in survival by LDL value.**

```
. testparm c.logldl i.race#c.logldl
 ( 1)  logldl = 0
 ( 2)  2.race#c.logldl = 0
 ( 3)  3.race#c.logldl = 0
 ( 4)  4.race#c.logldl = 0

           chi2(  4) =   26.84
         Prob > chi2 =    0.0000
```

d. Use the regression analysis in part a to perform a statistical test of the hypothesis that there is no association between time to death and race. Make clear which parameters you test and provide a two-sided p value.

***Instructions for grading*: This problem is worth 5 points.**

Answer: **We test that all parameters that involve some aspect of race are simultaneously 0. Hence we test the "main effects" for all races and the "interactions" between race and log LDL. The p value from such a test is P < 0.0001. We can thus state with high confidence that there is some difference in survival by race.**

```
. testparm i.race i.race#c.logldl
 ( 1)  2.race = 0
 ( 2)  3.race = 0
 ( 3)  4.race = 0
 ( 4)  2.race#c.logldl = 0
 ( 5)  3.race#c.logldl = 0
 ( 6)  4.race#c.logldl = 0

           chi2(  6) =   42.40
         Prob > chi2 =    0.0000
```

e. Use the regression analysis in part a to perform a statistical test of the hypothesis that there is no difference in the distribution of time to death between whites and blacks. Make clear which parameters you test and provide a two-sided p value.

***Instructions for grading*: This problem is worth 5 points.**

Answer: We test that the two parameters that compare whites to blacks are both simultaneously 0. Hence we test the "main effect" for race 2 and the "interaction" between race 2 and log ldl. The p value from such a test is 0.5416. We cannot detect a statistically significant difference in the distribution of time to death between whites and blacks.

```
. testparm 2.race 2.race#c.logldl
 ( 1)  2.race = 0
 ( 2)  2.race#c.logldl = 0

           chi2(  2) =    1.23
         Prob > chi2 =    0.5416
```

Problem 4 of the homework relates to the university salary dataset.

4. We are interested in raises given to faculty hired in recent years. For this problem, restrict attention to faculty hired in 1990 or later and who started at the university within one year of the year in which they received their highest degree. In order to (at least in part) examine the patterns of raises given to faculty, we will model salaries by sex, calendar year, and an interaction between sex and calendar year. Use such a model to answer the following questions.

```
. drop if year < 90
. drop if startyr < 90
. drop if yrdeg < startyr-1
```

a. Is there evidence of sex discrimination in the mean salary given in recent years? You do not have to provide full inference, but you should make clear the basis for your answer.

***Instructions for grading*: This problem is worth 5 points.**

Answer: We fit a linear regression model relating the mean monthly salary to a linear predictor involving an indicator of female sex, a continuously modeled calendar year, and the multiplicative interaction. We model the correlation among measurements made on the same faculty member through use of estimated correlation within clusters and the Huber-White sandwich estimator for the standard errors. We use a multiple partial F test to investigate sex discrimination in salaries by simultaneously testing that all regression parameters involving female sex are 0. Hence we test the female "main effect" and the female-calendar year interaction. The p value from such a test is 0.0493. Using a .05 level of significance, the difference in salaries is thus marginally significant.

```
. regress salary i.female##c.year, cluster(id)
Linear regression                               Number of obs =     485
                                                F(  3,   122) =   21.81
                                                Prob > F      =  0.0000
                                                R-squared     =  0.1011
                                                Root MSE      =  1163.5
```

```
                                    (Std. Err. adjusted for 123 clusters in id)
             |               Robust
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.female |  -625.0284   4706.158    -0.13   0.895    -9941.337    8691.281
        year |   176.1796    41.3186     4.26   0.000     94.38529    257.9739
             |
      female#|
      c.year |
          1  |   .8382353   50.40342     0.02   0.987    -98.94036    100.6168
             |
       _cons |  -11488.38   3867.427    -2.97   0.004    -19144.34    -3832.42
------------------------------------------------------------------------------

. testparm 1.female 1.female#c.year
 ( 1)  1.female = 0
 ( 2)  1.female#c.year = 0

       F(  2,   122) =    3.09
            Prob > F =    0.0493

(The following code is for comparison in part d)
. testparm year 1.female#c.year
 ( 1)  year = 0
 ( 2)  1.female#c.year = 0

       F(  2,   122) =   27.89
            Prob > F =    0.0000
```

b. Is there evidence of sex discrimination in the geometric mean salary given in recent years? You do not have to provide full inference, but you should make clear the basis for your answer.

***Instructions for grading*: This problem is worth 5 points.**

Answer: We fit a linear regression model on logarithmically transformed salaries, thereby relating the geometric mean monthly salary to a linear predictor involving an indicator of female sex, a continuously modeled calendar year, and the multiplicative interaction. We model the correlation among measurements made on the same faculty member through use of estimated correlation within clusters and the Huber-White sandwich estimator for the standard errors. We use a multiple partial F test to investigate sex discrimination in salaries by simultaneously testing that all regression parameters involving female sex are 0. Hence we test the female "main effect" and the female-calendar year interaction. The p value from such a test is 0.0216. Using a .05 level of significance, the difference in salaries is thus statistically significant.

```
. g logslry= log(salary)
. regress logslry i.female##c.year, cluster(id) eform("GeomMn")

Linear regression                               Number of obs =     485
                                                F(  3,   122) =   23.25
                                                Prob > F      =  0.0000
                                                R-squared     =  0.1302
                                                Root MSE      =   .2238

                                    (Std. Err. adjusted for 123 clusters in id)
             |               Robust
     logslry |     GeomMn   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.female |   .5878337   .5978626    -0.52   0.602     .0784978    4.402014
        year |   1.037861   .0089541     4.31   0.000     1.020286    1.055739
             |
```

```
      female#|
      c.year |
          1  |   1.004424    .010856     0.41   0.684     .9831616    1.026146

. testparm 1.female 1.female#c.year
 ( 1)  1.female = 0
 ( 2)  1.female#c.year = 0

       F(  2,   122) =    3.96
            Prob > F =    0.0216

(The following code is for use in part d.)
. testparm year 1.female#c.year
 ( 1)  year = 0
 ( 2)  1.female#c.year = 0

       F(  2,   122) =   29.67
            Prob > F =    0.0000
```

c. What are the relative merits of the two models used in parts a and b?

***Instructions for grading*: This problem is worth 5 points.**

Answer: Because both models use the same linear predictors, this question must relate to the relative merits of inference based on means and inference based on geometric means. In most financial and economic settings, raises, interest, depreciation, etc. are calculated on a multiplicative scale. Variation in the rates would thus tend to lead to probability distributions for total values in which the geometric mean is likely more precisely measured than the mean. However, the mean allows us to use a sample to estimate the total cost of discrimination in the population. Because the question is more about existence of discrimination, I would tend to regard that the greater precision would be the deciding point.

d. If you answered parts a and b correctly, you accounted for the correlated observations used in the analysis. Compare that inference to what you would have obtained had you incorrectly treated the data as independent. In particular, consider whether these incorrect models would have tended to be conservative or anti-conservative when making inference about associations with sex. How would your answer differ when considering associations by year?

***Instructions for grading*: This problem is worth 5 points.**

Answer: Salaries received year to year are likely highly positively correlated: a faculty member who is highly paid relative to other faculty members in one year is likely to be similarly better paid in other years. When comparing across sexes, these positively correlated observations are always contributing to the same mean. Hence, failure to account for the correlated data within clusters will overstate the statistical information and lead to anti-conservative inference. Consistent with such a pattern, we find a much lower p value in an analysis that allows for heteroscedasticity, but not correlation: In part a, we found a p value of .0493 for an association between salary and sex, while in the linear regression performed below, the "p value" for the test of association is $P < 0.0001$.

On the other hand, when comparing salary patterns across years, the positively correlated observations are contributing to different means for different years. In this case, failure to

account for the correlated data will understate the statistical information and lead to conservative inference. Consistent with such a pattern, we find a much lower test statistic (and higher p value) when ignoring the correlation. In the additional analysis that was included in part a, we found an F statistic of 27.89 for an association between salary and sex, while in the linear regression performed below, the "F statistic" for the test of association is 14.98. *(There are some technical difficulties when trying to compare F statistics that assume different degrees of freedom, but in this case the difference is so striking that it does indicate a lower P value when the correlated observations are properly modeled.)*

```
. stcox cldl i.ldlctg
. regress salary i.female##c.year, robust
Linear regression                               Number of obs =     485
                                                F(  3,   481) =   19.69
                                                Prob > F      =  0.0000
                                                R-squared     =  0.1011
                                                Root MSE      =  1163.5
------------------------------------------------------------------------------
             |               Robust
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.female |  -625.0284   6105.549    -0.10   0.919    -12621.87    11371.81
        year |   176.1796   50.34239     3.50   0.001     77.26141    275.0978
             |
      female#|
      c.year |
          1  |   .8382353    65.6018     0.01   0.990    -128.0633    129.7397
             |
       _cons |  -11488.38    4688.26    -2.45   0.015    -20700.38   -2276.376
------------------------------------------------------------------------------

. testparm 1.female 1.female#c.year
 ( 1)  1.female = 0
 ( 2)  1.female#c.year = 0

       F(  2,   481) =   13.28
            Prob > F =    0.0000

. testparm year 1.female#c.year
 ( 1)  year = 0
 ( 2)  1.female#c.year = 0

       F(  2,   481) =   14.98
            Prob > F =    0.0000
```