Biost 312: Modern Regression Analysis

Extra Problems for Discussion

March 29, 2010 (and other dates)

All problems refer to the salary dataset as found on the class web pages. This is a very large file, so you need to make sure you have sufficient memory available when you start Stata. Also, it is probably most convenient if you code the variables as numbers, and use labels to make them more understandable.

In these problems, you are frequently asked to save the predicted values in order to examine them in a plot in a later question. The fitted values can be obtained following any regression. For instance, you might use Stata commands

```
regress salary female yrdeg startyr if year==95, robust
predict fit
```

You will then have a new variable *fit* that contains the mean values predicted by the model.

1. We are interested in making inference about the difference in the mean monthly salary paid to women faculty in 1995 and that paid to men faculty in 1995. In this problem, we focus on the modeling of the variables *yrdeg* and *startyr*.

    a. In all parts of this problem, in addition to the year of degree and starting year, you should adjust for the highest degree obtained, field, and administrative duties. What is the best way to model the variables *degree, field,* and *admin*? Briefly justify your answer.

    b. In all parts of this problem you should use robust standard error estimates. Briefly explain why inference based on classical linear regression (without robust SE estimates) would be incorrect. Do you think the classical linear regression inference would tend to be conservative or anti-conservative? Justify your answer.

    c. Model *yrdeg* and *startyr* as linear continuous variables. Report the inference you would make for the difference in mean salaries for men and women (a table of the results for parts c, d, e, f, and g will be sufficient). Save the predicted mean values from this regression model as variable *fitC*.

    d. Model *yrdeg* and *startyr* as quadratic continuous variables (so linear continuous plus a second order term). Report the inference you would make for the difference in mean salaries for men and women (a table of the results for parts c, d, e, f, and g will be sufficient). Save the predicted mean values from this regression model as variable *fitD*.

    e. Model *yrdeg* and *startyr* as dummy variables for groups defined by earlier than 1966, 1966 – 1975, 1976 – 1985, 1986 – 1995. Report the inference you would make for the difference in mean salaries for men and women (a table of the results for parts c, d, e, f, and g will be sufficient). Save the predicted mean values from this regression model as variable *fitE*.

f. Model *yrdeg* and *startyr* as linear splines with knots at years 1965, 1975, and 1985. Report the inference you would make for the difference in mean salaries for men and women (a table of the results for parts c, d, e, f, and g will be sufficient). Save the predicted mean values from this regression model as variable *fitF*.

g. Model *yrdeg* and *startyr* as dummy variables for groups defined by earlier than 1960, and then each year from 1960 to 1995. Report the inference you would make for the difference in mean salaries for men and women (a table of the results for parts c, d, e, f, and g will be sufficient). Save the predicted mean values from this regression model as variable *fitG*.

h. What is the difference in the assumed relationship between monthly salary and year of degree in models fit in parts c – g?

i. Examine the agreement between the fitted values derived from the various models in parts c – g. Do you find strong evidence that one or more of the models was superior to the others with respect to its ability to model the relationship between salary and either year of degree or year starting?

j. Examine the agreement between the inference about the adjusted association between monthly salary and sex. Did the inference vary substantially across the various models?

k. In a real situation, how would choose among the models you fit in parts c – g?

2. We are interested in making inference about the difference in the mean monthly salary paid to faculty according to the year in which faculty obtained their degree and the year in which they started. In all models in this problem, we will appropriately adjust for degree, field, administrative duties, and sex.

   a. Provide inference about the adjusted association between monthly salary and year of degree (modeled as a linear continuous variable, not adjusted for starting year).

   b. Provide inference about the adjusted association between monthly salary and starting year (modeled as a linear continuous variable, not adjusted for year of degree).

   c. Provide inference about the adjusted association between monthly salary and year of degree (modeled as a linear continuous variable, and adjusted for starting year as well as the other variables).

   d. Provide inference about the adjusted association between monthly salary and starting year (modeled as a linear continuous variable, and adjusted for year of degree as well as the other variables).

   e. Briefly discuss the scientific relevance between the results obtained in parts a,b and parts c,d of this problem.

Problems 3 – 5 ask you to fit a series of models in which you consider a hierarchy of adjusted analyses for each of three different summary measures. Your response to these problems might be best presented in a table of inference about the adjusted association between monthly salary and sex. Model *yrdeg* and *startyr* as linear splines as computed in problem 1f.

3. We are interested in making inference about the difference in the mean monthly salary paid to women faculty in 1995 and that paid to men faculty in 1995.

    a. Report inference regarding the unadjusted comparison of women's and men's salaries.

    b. Report inference regarding the comparison of women's and men's salaries after adjustment for degree.

    c. Report inference regarding the comparison of women's and men's salaries after adjustment for degree, year of degree.

    d. Report inference regarding the comparison of women's and men's salaries after adjustment for degree, year of degree, starting year.

    e. Report inference regarding the comparison of women's and men's salaries after adjustment for degree, year of degree, starting year, field.

    f. Report inference regarding the comparison of women's and men's salaries after adjustment for degree, year of degree, starting year, field, administrative duties. Save the predicted values of the mean salary for each individual as *fit3*.

    g. Report inference regarding the comparison of women's and men's salaries after adjustment for degree, year of degree, starting year, field, administrative duties, rank.

4. We are interested in making inference about the ratio of geometric mean monthly salary paid to women faculty in 1995 and that paid to men faculty in 1995.

    a. Report inference regarding the unadjusted comparison of women's and men's salaries.

    b. Report inference regarding the comparison of women's and men's salaries after adjustment for degree.

    c. Report inference regarding the comparison of women's and men's salaries after adjustment for degree, year of degree.

    d. Report inference regarding the comparison of women's and men's salaries after adjustment for degree, year of degree, starting year.

    e. Report inference regarding the comparison of women's and men's salaries after adjustment for degree, year of degree, starting year, field.

    f. Report inference regarding the comparison of women's and men's salaries after adjustment for degree, year of degree, starting year, field, administrative duties. Save the predicted values of the geometric mean salary for each individual as *fit4*.

    g. Report inference regarding the comparison of women's and men's salaries after adjustment for degree, year of degree, starting year, field, administrative duties, rank.

5. We are interested in making inference about the ratio of the mean monthly salary paid to women faculty in 1995 and that paid to men faculty in 1995. Stata has a regression function "glm" that

allows the specification of a log link function. Hence, you can fit the regression for part a using the command

```
glm salary female if year==95, link(log) robust
```

Parameter estimates will be interpretable as the log mean (intercept) and log mean ratio (slope). Note that this differs from log transformed salary as the outcome, in which case we would be modeling ratios of geometric means (that was problem 4).

a. Report inference regarding the unadjusted comparison of women's and men's salaries.

b. Report inference regarding the comparison of women's and men's salaries after adjustment for degree.

c. Report inference regarding the comparison of women's and men's salaries after adjustment for degree, year of degree.

d. Report inference regarding the comparison of women's and men's salaries after adjustment for degree, year of degree, starting year.

e. Report inference regarding the comparison of women's and men's salaries after adjustment for degree, year of degree, starting year, field.

f. Report inference regarding the comparison of women's and men's salaries after adjustment for degree, year of degree, starting year, field, administrative duties. Save the predicted values of the mean salary for each individual as *fit5*.

g. Report inference regarding the comparison of women's and men's salaries after adjustment for degree, year of degree, starting year, field, administrative duties, rank.

6. Briefly discuss the similarities and differences between the analyses performed in problems 3 – 5. How similar are the predicted values between the models? How different is the inference you would obtain. *A priori*, which set of analyses would you prefer when answering the question regarding sex discrimination?

7. For the analysis model of your choice, summarize the scientific relevance of the results from the 7 different models used in the comparison of salaries paid to women and men.