**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2014

**Homework #8**

February 28, 2014

**Written problems:** To be submitted as a MS-Word compatible file to the class Catalyst dropbox by 9:30 am on Friday, March 7, 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***

* ***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.***
* ***Inference: A paragraph providing full statistical inference in answer to the question. Please see the supplementary document relating to “Reporting Associations” for details.***

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. The following file on the Datasets web pages contains commands you might find useful.

http://www.emersonstatistics.com/datasets/initsalary.doc

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 alternative modeling of the variables *yrdeg* and *startyr*. In all models in this problem, we will appropriately adjust for degree, field, administrative duties, and sex. ***(Note that I have provided answers to all parts of this problem except parts a, b and i, which you should answer.)***
   1. In all parts of this problem, in addition to the year of degree and year starting at the UW, 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.

**All of the variables: degree, field, and admin should be modeled as dummy variables as they are unordered categorical variables and there would be no sense in trying to fit a linear relationship between them.**

* 1. 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.

**There are fewer females in the dataset than males and we expect them to have a higher variance. Since the lower variance is assumed across groups in classical linear regression, this approach would be conservative.**

* 1. 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).

**Ans: (See table below)**

* 1. 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).

**Ans: (See table below)**

* 1. Model *yrdeg* and *startyr* as dummy variables for groups defined by earlier than 1960, 1960-64, 1965-69, 1970-74, 1975-79, 1980-84, 1985-89, and 1990 or later. 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).

**Ans: (See table below)**

* 1. Model *yrdeg* and *startyr* as linear splines with knots at years 1960, 1965, 1970, 1975, 1980, 1985, and 1990. 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).

**Ans: (See table below)**

* 1. Repeat parts c – f when modeling the ratio of mean salaries across sexes and when modeling the ratio of geometric mean salaries across sexes. These results can be included in the same table.)

**Ans: (See table below)**

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

**Ans: The following table provides the regression parameter estimates for the predictor indicating female sex, its Z statistic, its two-sided P value, and its 95% CI for the alternative methods of modeling year of degree and starting year. A few comments are in order**

* **In all cases, the linear splines provided the best fit to the data in the sense that adding the linear splines to each of the other models proved to be statistically significant. Adding the dummy variables to the model that included the linear splines did not improve the fit. I do not recommend doing this sort of testing unless your question was about the form of the relationship (e.g., linear vs nonlinear). My point here is that the linear splines did seem to model the true relationship with salary better when I was modeling sex, field, degree, and administrative duties.**
* **When modeling year of degree and start year as quadratic functions, I could not statistically establish nonlinearity in the linear regression model of the difference of means. When considering ratios of means or geometric means, I could detect the nonlinearity of either the year of degree or starting year when testing them combined, but because the terms are so correlated, I could not ensure that both were nonlinear when adjusting for the other.**
* **When modeling year of degree and start year as dummy variables or linear splines, there tended to be statistically significant departures from linearity for each variable separately and combined.**
* **Note that I included the Z statistic in this table only because the results were so strikingly statistically significant, that is only through looking at the Z statistic that we can assess whether there were any substantial differences (there were not).**
* **Note the similarity in ratios across all methods of modeling year of degree and start years and across the summary measures (means or geometric means).**
* **I provided inference about ratios of means using both Poisson regression and the generalized linear model when assuming Gaussian data with a log link. I prefer the Poisson regression, though this really only makes a big difference when looking at risk ratios with binary data. In that case, I *highly* recommend using Poisson regression rather than the generalized linear model with the binomial family and the log link. With means of positive continous random variables Poisson regression or the Gaussian GLM will both tend to behave okay.**
* **Lastly, the difference in means is of course a very different scale than the ratios of means or geometric means. But if you consider that the mean monthly salary for the entire sample was $6,389.81, the difference in means of about $420 is about 7% of the overall mean. So all models are giving quite similar answers.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Estimate** | **Z** | **P Value** | **95% CI low** | **95% CI high** |
| *Difference in Means* | | | | | |
| **Linear** | -428.3 | -5.23 | < .0001 | -588.9 | -267.8 |
| **Quadratic** | -428.1 | -5.25 | < .0001 | -588.1 | -268.0 |
| **Dummy** | -447.7 | -5.45 | < .0001 | -609.0 | -286.5 |
| **Splines** | -419.7 | -5.17 | < .0001 | -579.0 | -260.5 |
| *Ratio of Means (Poisson)* | | | | | |
| **Linear** | 0.9266 | -5.42 | < .0001 | 0.9014 | 0.9525 |
| **Quadratic** | 0.9280 | -5.36 | < .0001 | 0.9030 | 0.9537 |
| **Dummy** | 0.9244 | -5.63 | < .0001 | 0.8994 | 0.9500 |
| **Splines** | 0.9289 | -5.34 | < .0001 | 0.9041 | 0.9544 |
| *Ratio of Means (GLM)* | | | | | |
| **Linear** | 0.9227 | -5.55 | < .0001 | 0.8969 | 0.9493 |
| **Quadratic** | 0.9246 | -5.43 | < .0001 | 0.8988 | 0.9511 |
| **Dummy** | 0.9185 | -5.83 | < .0001 | 0.8926 | 0.9451 |
| **Splines** | 0.9245 | -5.49 | < .0001 | 0.8989 | 0.9508 |
| *Ratio of Geometric Means* | | | | | |
| **Linear** | 0.9347 | -5.22 | < .0001 | 0.9113 | 0.9587 |
| **Quadratic** | 0.9352 | -5.22 | < .0001 | 0.9119 | 0.9590 |
| **Dummy** | 0.9328 | -5.42 | < .0001 | 0.9096 | 0.9566 |
| **Splines** | 0.9363 | -5.17 | < .0001 | 0.9132 | 0.9600 |

* 1. In a real situation, how would choose among the alternative methods for adjusting for year of degree and starting year?

**We would rely on a priori assumptions about the nature of salary increase with respect to time in the workforce. Is it multiplicative, is it additive, is it evaluated in blocks of time with no adjustment in between or is it evaluated periodically with cost of living adjustments between as the spline model would suggest? Our analysis would rely more on substantive accuracy than statistical accuracy.**

1. 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 at UW. In all models in this problem, we will appropriately adjust for degree, field, administrative duties, and sex.
   1. Provide inference about the adjusted association between monthly salary and year of degree (modeled as a linear continuous variable, not adjusted for starting year).

**See Table**

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

**See Table**

* 1. 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)

**See Table**

* 1. 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).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Difference in Mean Monthly Salary | | | | |
|  | Estimate | Z | P Value | 95% CI low | 95% CI high |
| Degree Year Unadjusted for Starting Year | -89.87 | -20.89 | <0.0001 | -98.30 | -81.43 |
| Starting Year Unadjusted for Year of Degree | -56.88 | -12.06 | <0.0001 | -66.13 | -47.63 |
| Degree Year Adjusted for Starting Year | -111.96 | -11.79 | <0.0001 | -130.58 | -93.34 |
| Starting Year Adjusted for Year of Degree | 27.15 | 2.88 | 0.004 | 8.68 | 45.63 |
| \*All above models adjust for degree, field, administrative duties, and sex | | | | | |

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

**Adjusting for starting year when analyzing difference of means by year of degree and adjusting for degree year when analyzing difference of means by starting year at the UW add important information. Adjusting for starting year when analyzing degree year shows will take the people who started earlier because they were in a lower paid postdoctoral or graduate position from those who started earlier as full professors and magnify the positive relationship between time since obtaining degree and mean monthly salary (as demonstrated in a negative relation between the year of degree and mean monthly salary).**

**Similarly, when you adjust for year of degree, you separate the senior hires from new hires. People who started earlier but graduated the same year as somebody who started later is more likely to have been hired more quickly out of graduate school as lecturers, assistant and associate professors and may be on a slower or less upwardly mobile track. On the other hand, senior hires are morel likely to be full professors deans or other administrators.**

**Separating these groups gives a clearer idea of the relationship between starting year, year of degree, and salary.**

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.

**For problems 3-5 see table at the end of the questions**

For the benefit of the graders, we will agree on modeling *yrdeg* and *startyr* as linear splines as computed in problem 1f.

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.
   1. Report inference regarding the unadjusted comparison of women’s and men’s salaries.
   2. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree.
   3. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree.
   4. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree, starting year at UW.
   5. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree, starting year at UW, field.
   6. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree, starting year at UW, field, administrative duties. Save the predicted values of the mean salary for each individual as *fit3.*
   7. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree, starting year at UW, field, administrative duties, rank.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Difference in Mean Monthly Salary | | | | |
| Salary~ | Estimate | Z | P Value | 95% CI low | 95% CI high |
| Female | -1334.73 | -14.04 | <0.0001 | -1521.18 | -1148.29 |
| Female+Degree | -1266.15 | -13.40 | <0.0001 | -1451.56 | -1080.75 |
| Female+Degree+Degree Year | -614.43 | -7.18 | <0.0001 | -782.34 | -446.52 |
| Female+Degree+Degree Year+Starting Year | -621.46 | -7.14 | <0.0001 | -792.13 | -450.79 |
| Female+Degree+Degree Year+Starting Year+Field | -425.43 | -5.11 | <0.0001 | -588.75 | -262.11 |
| Female+Degree+Degree Year+Starting Year+Field+admin | -424.90 | -5.22 | <0.0001 | -584.41 | -265.38 |
| Female+Degree+Degree Year+Starting Year+Field+admin+rank | -286.77 | -4.17 | <0.0001 | -421.80 | -151.74 |

1. 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.
   1. Report inference regarding the unadjusted comparison of women’s and men’s salaries.
   2. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree.
   3. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree.
   4. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree, starting year at UW.
   5. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree, starting year at UW, field.
   6. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree, starting year at UW, field, administrative duties. Save the predicted values of the geometric mean salary for each individual as *fit4.*
   7. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree, starting year at UW, field, administrative duties, rank.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Ratio of Geometric Mean Monthly Salary | | | | |
| Salary~ | Estimate | Z | P Value | 95% CI low | 95% CI high |
| Female | 0.81 | -13.73 | <0.0001 | 0.79 | 0.84 |
| Female+Degree | 0.82 | -13.09 | <0.0001 | 0.80 | 0.85 |
| Female+Degree+Degree Year | 0.91 | -7.00 | <0.0001 | 0.89 | 0.93 |
| Female+Degree+Degree Year+Starting Year | 0.91 | -7.04 | <0.0001 | 0.88 | 0.93 |
| Female+Degree+Degree Year+Starting Year+Field | 0.94 | -5.10 | <0.0001 | 0.91 | 0.96 |
| Female+Degree+Degree Year+Starting Year+Field+admin | 0.94 | -5.21 | <0.0001 | 0.91 | 0.96 |
| Female+Degree+Degree Year+Starting Year+Field+admin+rank | 0.96 | -4.15 | <0.0001 | 0.94 | 0.98 |

1. 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. You can use Poisson regression (with the irr option to get exponentiated parameter estimates), or you can use a generalized linear model with a log link. 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). (glm stands for “generalized linear model” and it includes as special cases linear regression, logistic regression, and Poisson regression. By default, it presumes the data are continuous and models the mean according to the value of the link function.) By specifying the “eform” option, it will return the exponentiated parameter estimates.

In either case, make clear which analysis method you used.

* 1. Report inference regarding the unadjusted comparison of women’s and men’s salaries.
  2. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree.
  3. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree.
  4. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree, starting year at UW.
  5. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree, starting year at UW, field.
  6. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree, starting year at UW, field, administrative duties. Save the predicted values of the mean salary for each individual as *fit5.*
  7. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree, starting year at UW, field, administrative duties, rank.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Ratio of Mean Monthly Salary | | | | |
| Salary~ | Estimate | Z | P Value | 95% CI low | 95% CI high |
| Female | 0.80 | -13.58 | <0.0001 | 0.78 | 0.83 |
| Female+Degree | 0.81 | -12.98 | <0.0001 | 0.79 | 0.84 |
| Female+Degree+Degree Year | 0.90 | -7.10 | <0.0001 | 0.88 | 0.93 |
| Female+Degree+Degree Year+Starting Year | 0.90 | -7.08 | <0.0001 | 0.87 | 0.93 |
| Female+Degree+Degree Year+Starting Year+Field | 0.93 | -5.28 | <0.0001 | 0.90 | 0.95 |
| Female+Degree+Degree Year+Starting Year+Field+admin | 0.93 | -5.40 | <0.0001 | 0.90 | 0.95 |
| Female+Degree+Degree Year+Starting Year+Field+admin+rank | 0.95 | -4.37 | <0.0001 | 0.93 | 0.97 |

1. 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?

**All analyses when done adjusting for the same covariates are pretty similar. The linear regression on untransformed means and the poisson regression are more similar because they don’t downweight outliers in the way that the regression on geometric means does this is reflected in the predicted values where all are similar, but the poisson model and the linear regression are closest. The methods certainly make a lower impact than the choice of confounders and precision variables.**

1. For the analysis model that you would have chosen *a priori*, summarize the scientific relevance of the single model that you think would best reflect any discrimination against women in awarding salaries. Give a formal report of your methods and results.

**The model that includes degree, degree year, starting year, field and admin shows the ratio of mean salaries between male and female faculty members when licensure, seniority, experience, field and additional duties are held constant. This means that whatever disparity is present is there regardless of how long the faculty members have been in the school, how much experience they have accumulated, whether their field is lucrative and if they do additional work for the school beyond research/instruction. I don’t believe that field should be excluded because although I do believe that selection into fields is in part due to discrimination within the university system, it is a shared blame with society at large and appropriate response would be broader than at the university level. Additionally, I believe that the administrative duties variable will suppress some of the outliers that come from faculty members who are also deans or other sorts of administrators nullifying the necessity of the more difficult to explain geometric means regression.**

**Methods: To assess the percent difference between male faculty members’ average monthly salary and otherwise similar female faculty members’ average salary; we used a poisson regression with robust standard errors. We controlled for degree, degree year, starting year, field, and administrative duties.**

**Inference: On average, female faculty members at the UW in 1995 made 7% less per month than their male counterparts who hold the same type of degree, graduated in the same year, started at the same time at the UW, work in the same field and have the same amount of administrative duties. These results would not have been surprising if the true percent difference in mean salaries were anywhere between 5% and 10% less for female faculty members. With a two-sided p-value less than 0.0001, we can reject at the alpha=0.05 level the null hypothesis that there is no percent difference in average salaries between the sexes in favor of the alternative hypothesis that female faculty members make less on average than male faculty members.**