**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.

**Degree, field, and admin are all potential effect modifiers on the variables startyr and yrdeg. Therefor, we should model them stratify our analyses of startyr and yrdeg when modeling these variables, such as modeling them as splines.**

* 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 is marked heterscedasticity in the sample, especially in terms of salary, where there are much more men who make a lot of money (and more overall men). Thus, inference based on classical linear regression will would be overly conservative when it trys to borrow information from smaller ranges (for females) to make inferences across the entire range of salaries.**

* 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?

**I would first plot the points to examine if we have heteroscedasticity at different distributions of these variables across salary and sex. I would use linear regression for obviously linear interactions, and higher order regression if the interaction calls for it. I would also model categorical variables as discrete variables if I think they would behave differently acorss different groups.**

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).

**All problems adjust for degree, field, administrative duties, and sex.**

**Without adjusting for starting year, for each year later an employee received their degree, they get $88.56 less per month.**

* 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).

**Without adjusting for year of degree, for each year later an employee starts, they get $56.93 less per month.**

* 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).

**For each year later an employee received their degree, after adjusting for start year, they get $106.94 less per month.**

* 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).

**For each year later an employee started, after adjusting year of degree, they get $25.17 more per month.**

* 1. 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.

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.

**The following table provides the difference in monthly salary between women and men in 1995, its T statistic, its two-sided P value, and its 95% CI for the alternative methods of modeling year of degree and starting year. Start year and year degree was earned were modeled as linear splines. Robust linear regression was performed on salary, adjusting for the variables listed below.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Adjusted for** | **$ Difference in monthly salary (Female – male)** | **T statistic** | **P Value** | **95% CI low** | **95% CI high** |
| *Difference in Means* |
| Unadjusted | -1334.731 | -14.04 | 0.000 | -1521.177 | -1148.286 |
| Degree | -1262.195 | -13.32 | 0.000 | -1448.02 | -1076.369 |
| Degree, Year of Degree | -629.7469 | -7.34 | 0.000 | -798.0564 | -461.4374 |
| Degree, Year of Degree, Start Year | -637.5013 | -7.33 | 0.000 | -808.0074 | -466.9953 |
| Degree, Year of Degree, Start Year, Field | -439.5629 | -5.28 | 0.000 | -602.9406 | -276.1852 |
| Degree, Year of Degree, Start Year, Field, Administrative Duties  | -441.7571 | -5.43 | 0.000 | -601.2291 | -282.2852 |
| Degree, Year of Degree, Start Year, Field, Administrative Duties, Rank | -284.9572 | -4.08 | 0.000 | -422.0893 | -147.8251 |

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.

**The following table provides the ratio of geometric mean salary between women and men in 1995, its T statistic, its two-sided P value, and its 95% CI for the alternative methods of modeling year of degree and starting year. Start year and year degree was earned were modeled as linear splines. Robust linear regression on the log transformed salary was performed, adjusting for the variables listed below.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Adjusted for | **Ratio of geometric mean monthly salary paid to women faculty vs men faculty in 1995** | **T statistic** | **P Value** | **95% CI low** | **95% CI high** |
| *Difference in Ratios of Geometric Mean* |
| Unadjusted | .8120 | -13.73 | 0.000 | .7882 | .8365 |
| Degree | .8209 | -13.02 | 0.000 | .7968 | .8456 |
| Degree, Year of Degree | .9068 | -7.15 | 0.00 | .8828 | .9315 |
| Degree, Year of Degree, Start Year | .9059 | -7.21 | 0.000 | .8819 | .9306 |
| Degree, Year of Degree, Start Year, Field | .9339 | -5.23 | 0.000 | .9103 | .9582 |
| Degree, Year of Degree, Start Year, Field, Administrative Duties  | .9337 | -5.38 | 0.000 | .9106 | .9573 |
| Degree, Year of Degree, Start Year, Field, Administrative Duties, Rank | .9572 | -4.04 | 0.000 | .9371 | .9778 |

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.

**The following table provides the ratio of mean salary between women and men in 1995, its T statistic, its two-sided P value, and its 95% CI for the alternative methods of modeling year of degree and starting year. Start year and year degree was earned were modeled as linear splines. Robust log linked generalized linear regression on the salary was performed, adjusting for the variables listed below.**

I am using the generalized linear model.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Adjusted** **for** | **Ratio of mean monthly salary for women faculty vs men faculty in 1995** | **T statistic** | **P Value** | **95% CI low** | **95% CI high** |
| *Ratio in Means* |
| Unadjusted | .8017 | -13.58 | 0.000 | .7765 | .8277 |
| Degree | .8099 | -12.96 | 0.000 | .7845 | .8361 |
| Degree, Year of Degree | .8951 | -7.32 | 0.000 | .8690 | .9221 |
| Degree, Year of Degree, Start Year | .8924 | -7.34 | 0.000 | .8656 | .9200 |
| Degree, Year of Degree, Start Year, Field | .9216 | -5.53 | 0.000 | .8953 | .9487 |
| Degree, Year of Degree, Start Year, Field, Administrative Duties  | .9203 | -5.81 | 0.000 | .8949 | .9464 |
| Degree, Year of Degree, Start Year, Field, Administrative Duties, Rank | .9493 | -4.27 | 0.000 | .9269 | .9723 |

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?

**Comparing questions 4 and 5, the ratio of mean monthly salary is smaller than the ratio of geometric mean monthly salary for each model done. In all models, the P value is highly statistically significant (P < .0005). The T statistic and regression parameters follow the same trend (usually decreasing, but increasing for some variables) when adding in more varaibles to the model. Overall, the inference is basically the same from each model.**

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.

**A priori, I would have selected the model that models the difference in mean monthly salary, as this is an easy to interpret model and reflects what a layperson would want to know.**

**The following table provides the difference in monthly salary between women and men in 1995, its T statistic, its two-sided P value, and its 95% CI for the alternative methods of modeling year of degree and starting year. Start year and year degree was earned were modeled as linear splines. Robust linear regression was performed on salary, adjusting for the variables listed below. Start year and year of degree were modeled as linear splines, and P values and confidence intervals were computing using Wald statistics. Missing data was omitted for missing vales only.**

**3926 of 19792 subjects were women. Before adjusting for any variables, women were found to make a mean of $1334.73 less per month than men. Based on a 95% confidence interval, this would be unsurprising if the true difference in mean monthly salary was between $1521.18 and $1148.29 less per month for females. This was statistically significant based on a two sided P value of <.0005. After adjusting for all of the variables in the table, with yrdeg and startyr modeled as linear splines, women were found to make a mean monthly salary of $284.96 less per month than men. Based on a 95% confidence interval, this would be unsurprising if the true difference in mean monthly salary was between $422.09 and $147.83 less per month for females. This was statistically significant based on a two sided P value of <.0005.**