**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.

**From the homework of Biostat 514/517 we did last quarter(HW 7) , we know that degree, field and admin are all potential confounds. Here we might as well model these confounding variables as dummy variables in our model because they are categorical variables.**

* 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.

**We’ve known that we should use robust standard error estimates if we assume unequal variance across the data. Also, using robust standard error helps adjusting the correlated data among identified clusters. By using robust standard, we would have a wider confidence interval because we’ve been more 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?

**Using the spline model is an alternative methods for adjusting for year of degree and starting year. First of all, it’s the best fit according to 1(h). This makes sense since linear splines provide several slopes which helps to fit the model more precisely. Also, using linear spline method helps us to reduce the confounding among the estimates.**

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).

**Method: Linear regression model on monthly salary paid with robust standard error being used. Our predict of interest is the year in which faculty obtained their degree and we adjust for degree, field, administrative duties and sex. We will evaluate the estimate sex and the associated 95% confidence interval.**

|  |
| --- |
| **Linear regression model on monthly salary paid with respect to year of degree w/o adjust starting year.** |
|  | **Estimate** | **p-value** | **95% C.I** |
| **Year of degree** | **-89.87** | **< .0001** | **-98.302** | **-81.429** |

* 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).

|  |
| --- |
| **Linear regression model on monthly salary paid with respect to starting salary w/o adjust year of degree.** |
|  | **Estimate** | **p-value** | **95% C.I** |
| **Starting year** | **-56.88** | **< .0001** | **-66.133**  | **-47.632** |

* 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).

|  |
| --- |
| **Linear regression model on monthly salary paid with respect to year of degree after adjusting for starting year.** |
|  | **Estimate** | **p-value** | **95% C.I** |
| **Year of degree** | **-111.96** | **< .0001** | **-130.580** | **-93.342** |

* 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).

|  |
| --- |
| **Linear regression model on monthly salary paid with respect to starting year after adjusting for year of degree.** |
|  | **Estimate** | **p-value** | **95% C.I** |
| **Starting year** | **27.15361** | **0 .0004** | **8.680**  | **45.627** |

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

**If we compare the estimates from a, b and c, d, we can tell that the effect of year of degree on monthly salary is stronger after adjusting for starting year. And the effect of starting year on monthly salary is reversed after adjusting for year of degree. One reason of the effect of year of degree on monthly salary being stronger after adjusting might because people are paid less back the days. There’s inflation in the economy. One reason of the effect of starting year on monthly salary is reversed after adjusting is probably because people who having early starting year might have degree for a long time.**

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.

**for a-g in question 3, please see the table below.**

|  |
| --- |
| **Linear models of mean monthly salary by modeling yrdeg and startyr as linear splines** |
|  | **Estimate (Female)** | **p-value** | **95% C.I** |
| **a.** | **-1334.731** | **< .0001** | **-1521.177**  | **-1148.286** |
| **b.** | **-1266.152** | **< .0001** | **-1451.555**  | **-1080.75** |
| **c.** | **-614.1284** | **< .0001** | **-782.235** | **-446.0218** |
| **d.** | **-614.5785** | **< .0001** | **-785.3114**  | **-443.8456** |
| **e.** | **-420.0537** | **< .0001** | **-583.1193**  | **-256.9881** |
| **f.** | **-419.7268** | **< .0001** | **-578.9865**  | **-260.4672** |
| **g.** | **-280.6639** | **< .0001** | **-415.5158**  | **-145.812** |

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

**for a-g in question 4, please see the table below.**

|  |
| --- |
| **Linear models of geometric mean of monthly salary (log of the ratio of geometric mean is reported)** |
|  | **Estimate (Female)** | **p-value** | **95% C.I** |
| **a.** | **-.2082332** | **< .0001** | **-.2379778**  | **-.1784886** |
| **b.** | **-.198018** | **< .0001** | **-.2276879**  | **-.1683488** |
| **c.** | **-.0953901** | **< .0001** | **-.1221598**  | **-.0686204** |
| **d.** | **-.0957676** | **< .0001** | **-.1226957**  | **-.0688394** |
| **e.** | **-.0658741** | **< .0001** |  **-.0914057**  | **-.0403425** |
| **f.** | **-.0658281** | **< .0001** |  **-.0907941**  | **-.040862** |
| **g.** | **-.0435036** | **< .0001** | **-.0643987**  | **-.0226085** |

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

**for a-g in question 5, please see the table below.**

|  |
| --- |
| **poisson regression models of monthly salary (glm and eform command are used, ratio of probability is reported)** |
|  | **Estimate (Female)** | **p-value** | **95% C.I** |
| **a.** | **.8017227** | **< .0001** | **.7765495**  | **.8277119** |
| **b.** | **.809741** | **< .0001** | **.7843638**  | **.8359393** |
| **c.** | **.8981333** | **< .0001** |  **.8719496**  | **.9251032** |
| **d.** | **.8963554** | **< .0001** | **.8694502**  | **.9240932** |
| **e.** | **.9251028** | **< .0001** |  **.8986308 .** | **9523546** |
| **f.** | **.9244878** | **< .0001** |  **.8989153**  | **.9507877** |
| **g.** | **.9506864** | **< .0001** |  **.9282705**  | **.9736436** |

* 1. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree.
	2. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree.
	3. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree, starting year at UW.
	4. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree, starting year at UW, field.
	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, administrative duties. Save the predicted values of the mean salary for each individual as *fit5.*
	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, rank.
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 three models gave similar result according the fitted graph. For the male group, linear regression model of mean monthly salary and poisson regression model have really closed fitted value before salary hits 6000. Poisson ended up having a steeper trend. For both male and female group, geometric mean model gave relatively lower fitted values but still fairly closed to the other two models. However, there exists a pretty big discrepancy between female fitted salary and male fitted salary. Thus, there might be sexual discrimination on the salary.**

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.

**My priori is the linear regression model with geometric mean. Likely, we would have a male director of the department and his salary will be way higher than the average salary. This outlier will cause the average of male group salary higher than female group salary and make people think that there is sexual discrimination. But it might actually not exist. Similarly, if the director is female, it could cause the problem too. (It would look like no discrimination when there actually is.) Therefore, using log salary will reduce the effect of the outlier in the salary.**

**Method: We’d like to use the linear regression model on geometric mean of monthly salary with respect to the sex (female) and adjusting degree, year of degree, starting year at UW, field, administrative duties, rank. We’d like to use robust robust standard error. And we will report the estimate of sex, it’s p-value and 95% confidence interval.**