**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.

The highest degree obtained variable, “degree”, is an unordered categorical variable with three values: “PhD”, “Prof”, or “Other”. The best way to model this variable is by creating three dummy variables representing the presence or absence of any of these three degree states. The first dummy variable would be equal to 1 if the highest degree obtained is a PhD, and 0 for either of the other two options. The second dummy variable would be equal to 1 for a value of “Prof” and 0 for either of the other two options, and the third would be similar with “Other” equal to 1. Field is also an unordered categorical variable with three values, and should be treated exactly the same way, with three dummy variables. Field is already a dummy variable with values of either “1” or “0” so it can be added into a regression directly.

* 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.

Inference based on classical linear regression assumes homogenaeity between groups, and as the current analysis is assessing the possibility of variance due to effect modification or confounding of other variables, ignoring this possible effect is unhelpful in determining the effects of adjusting for various variables.

* 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?

In a real situation, I would assess whether it makes scientific sense to consider either of them a confounder, and if so, I would determine whether either is likely to exhibit linear behavior. If either seems to be non-linearly associated with salary, I would choose linear splines to model them as covariates instead of modeling them as continuous linear variables.

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).

A Multiple Linear Regression was performed using a robust SE computed using the Huber-White Sandwich Estimator. Degree, field, administrative duties, and sex were used as dummy variable covariates, and the year of degree was the predictor of interest treated as a continuous linear variable. 95% confidence intervals and a 2-sided p-value were computed using Wald statistics. For every 1-year increase in the year a subject’s highest degree was earned, his or her mean salary is 89.865 dollars/month lower, with a robust standard error of 4.301, when adjusted for type of degree, field, administrative duties, and sex. According to a 95% confidence interval, finding a mean difference in salary for a 1-year increase in the year the subject’s highest degree was earned anywhere between -98.302 and -81.429 dollars/month would not be surprising. A 2-sided p-value tested the null hypothesis of a mean difference in salary of zero across a 1-year difference in the year the subject’s degree was earned when corrected for degree, field, administrative duties and sex. With a computed p>0.005, these results are statistically significant and this null hypothesis can be rejected.

* 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).

A Multiple Linear Regression was performed using a robust SE computed using the Huber-White Sandwich Estimator. Degree, field, administrative duties, and sex were used as dummy variable covariates, and the starting year of employment was the predictor of interest treated as a continuous linear variable. 95% confidence intervals and a 2-sided p-value were computed using Wald statistics. For every 1-year increase in the starting year of employment his or her mean salary is 56.882 dollars/month lower, with a robust standard error of4.414, when adjusted for type of degree, field, administrative duties, and sex. According to a 95% confidence interval, finding a mean difference in salary for a 1-year increase in the starting year of employment anywhere between -47.632 and -66.133 dollars/month would not be surprising. A 2-sided p-value tested the null hypothesis of a mean difference in salary of zero across a 1-year difference in the starting year of employment corrected for degree, field, administrative duties and sex. With a computed p>0.005, these results are statistically significant and this null hypothesis can be rejected.

* 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).

A Multiple Linear Regression was performed using a robust SE computed using the Huber-White Sandwich Estimator. Degree, field, administrative duties, and sex were used as dummy variable covariates, as well as the starting year of employment as a linear continuous covariate, and the year of degree was the predictor of interest treated as a continuous linear variable. 95% confidence intervals and a 2-sided p-value were computed using Wald statistics. For every 1-year increase in the year a subject’s highest degree was earned, his or her mean salary is 111.96 dollars/month lower, with a robust standard error of 9.492, when adjusted for type of degree, field, administrative duties, and sex. According to a 95% confidence interval, finding a mean difference in salary for a 1-year increase in the year the subject’s highest degree was earned anywhere between -130.6 and -93.34 dollars/month would not be surprising. A 2-sided p-value tested the null hypothesis of a mean difference in salary of zero across a 1-year difference in the year the subject’s degree was earned when corrected for starting year of employment, degree, field, administrative duties and sex. With a computed p>0.005, these results are statistically significant and this null hypothesis can be rejected.

* 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).

A Multiple Linear Regression was performed using a robust SE computed using the Huber-White Sandwich Estimator. Degree, field, administrative duties, and sex were used as dummy variable covariates, as well as the year the subject’s highest degree was earned used as a linear continuous covariate, and the starting year of employment was the predictor of interest treated as a continuous linear variable. 95% confidence intervals and a 2-sided p-value were computed using Wald statistics. For every 1-year increase in the starting year of employment his or her mean salary is 27.15 dollars/month higher, with a robust standard error of 9.418, when adjusted for year of degre, type of degree, field, administrative duties, and sex. According to a 95% confidence interval, finding a mean difference in salary for a 1-year increase in the starting year of employment anywhere between 8.680 and 45.63 dollars/month would not be surprising. A 2-sided p-value tested the null hypothesis of a mean difference in salary of zero across a 1-year difference in the starting year of employment corrected for year of degree, degree, field, administrative duties and sex. With a computed p=0.00399 (<0.005), these results are statistically significant and this null hypothesis can be rejected.

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

In parts a and b, either starting year or year of degree were modeled as continuous linear predictors of interest of 1995 monthly salary, when adjusted for degree, field, administrative duties, and sex. Both had similar associations with salary (the earlier the year the faculty member either earned his/her degree or were hired at the institution the study sampled from, the higher the monthly salary). Parts c and d involve modeling of either starting year or year of degree while adjusting for the other (alongside degree, field, administrative duties, and sex). From part c, it appears that by adjusting for starting year, the association between year of degree and salary for faculty members that started around the same time is still a negative association (lower salary as years increase), but from part d it can be seen that for faculty members that have similar years of degree, the association between starting year and salary is a positive one—those that obtained their degree at the same time that were hired at a later year tended to have higher salaries.

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.



* 1. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree.



* 1. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree.



* 1. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree, starting year at UW.



* 1. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree, starting year at UW, field.



* 1. 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.*



* 1. 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.



* 1. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree.



* 1. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree.



* 1. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree, starting year at UW.



* 1. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree, starting year at UW, field.



* 1. 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.*



* 1. 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.



* 1. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree.



* 1. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree.



* 1. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree, starting year at UW.



* 1. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree, starting year at UW, field.



* 1. 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.*



* 1. 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?

The analyses performed in problems 3-5 are similar in that they assess the disparity between women and men’s salaries in 1995, and all indicate a trend of slightly higher salaries for men after adjustment for degree, year of degree, starting year, field, administrative duties, and rank, though these differences all approach the null value (0 for the mean difference, 1 for the ratio of means or geometric means) as more variables are adjusted for. The predicted values between the models are very similar, and my inference would not be different between the models.

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.

I have little knowledge about the nature of salary data in terms of whether a transformation, such as a logarithmic transformation or modeling it quadratically, would be appropriate. Therefore, my *a priori* analysis model would be an untransformed linear regression, adjusting for the potential confounders of degree (different types of degrees might be associated with different levels of income, and with sex), year of degree (those having earned their degree longer ago may be paid more due to having more experience in the field), starting year (those that started their current job earlier may be paid more), field (different fields may attract different sexes, and may be associated with different levels of income), administrative duties (those with administrative duties may be paid more/less than their colleagues), and rank (those of a higher rank may be paid more).

A Multiple Linear Regression was performed using a robust SE computed using the Huber-White Sandwich Estimator. Degree, field, administrative duties, and rank were used as dummy variable covariates, and year of degree and start year were modeled as linear splines with knots at the years 1960, 1965, 1970, 1975, 1980, 1985, and 1990. Sex was the predictor of interest, a binary variable. 95% confidence intervals and a 2-sided p-value were computed using Wald statistics. The estimated mean salary for women is 280.7 dollars/month lower, with a robust standard error of 68.75 dollars/month, when adjusted for type of degree, field, administrative duties, rank, start year, and year of degree. According to a 95% confidence interval, finding a mean difference in salary between men and women anywhere between -415.5 and -145.8 dollars/month (women mean salary-men mean salary) would not be surprising. A 2-sided p-value tested the null hypothesis of a mean difference in salary of zero between mean and women corrected for degree, field, administrative duties, start year, and year of degree. With a computed p>0.005, these results are statistically significant and this null hypothesis can be rejected.