**Biost 518: Applied Biostatistics II**

**Biost 515: Biostatistics II**

Emerson, Winter 2014

**Homework #8**

February 28, 2014

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 variables *degree, field* and *admin* are naturally unordered categorical variables so it is most useful to model them with numerical values for each category. Specifically, the variable *degree* will be 1 for “other” degrees, 2 for “PhD” and 3 for “Professional” degrees. The variable *field* will be coded as 1 for “Arts,” 2 for “Other” fields, and 3 for “Professional” fields. The variable *admin* will be 0 for non-administrative duties, and 1 for administrative duties. This will allow us to easily adjust for all three variables in regression analyses.**

* 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.

**Using classical linear regression without robust standard error estimates does not adjust for unequal variances and also does not adjust for correlation within clusters that are identified in the data. In the salary dataset, it is expected that variables will be correlated at least minimally within clusters defined by the individual subjects. For example, it is reasonable to expect that rank is positively correlated with increasing years of employment for any given subject. Without using robust SE estimates, inference based on classical linear regression would be anti-conservative because this data is structured similarly to a study with repeated measures and would likely result in p-values that are too low and a 95% CI that is too narrow.**

* 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 real life, you must make decisions about the scientific question and what you expect to see in the data. If we expect a nonlinear association, then it would be most appropriate to use linear splines, especially if it was important to be able to detect the nonlinearity of the data.**

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.

**See Table 2 below for parts a-d of Problem 2**

* 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).
	2. Provide inference about the adjusted association between monthly salary and starting year (modeled as a linear continuous variable, not adjusted for year of degree).
	3. 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).
	4. 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).

**Table 2. Difference in mean monthly salary according to the year of degree receipt, and the starting year of employment at UW.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Adjustment** | **Estimate** | **t-statistic/****z-statistic** | **P Value** | **95% CI minimum** | **95% CI maximum** |
| Salary vs Year of degree | Degree, Field, Admin duties, Sex | -29.72 | -5.400 | <0.0001 | -40.51 | -18.93 |
| Salary vs Starting year | Degree, Field, Admin duties, Sex | 4.715 | 0.8400 | 0.3990 | -6.238 | 15.67 |
| Salary vs Year of degree | Starting year, Degree, Field, Admin duties, Sex | -57.40 | -6.470 | <0.0001 | -74.81 | -40.00 |
| Salary vs Starting year | Year of degree, Degree, Field, Admin duties, Sex | 37.16 | 4.460 | <0.0001 | 20.83 | 53.48 |

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

**From a scientific standpoint, it is reasonable to expect that mean salaries will vary based on the year the employee received their relevant degree, such that receiving a degree more recently implies shorter duration of employment. Similarly, starting year reflects length of employment, but if we also assume there is an annual increase in salaries overall (due to cost of living adjustment, for example), then it would be reasonable to expect an increased salary for employees who were hired later compared to the starting salary from earlier years. Parts a captures the decrease in salary based on the number of years ago the degree was received, but does not account for the fact that the date of hire does not necessarily coincide with date of degree receipt. Similarly, part b captures the increase in starting salaries for more recent years, but does not account for the fact that an employee may have earned their degree years ago and gained experience elsewhere before starting employment at UW. Parts c and d adjust for those discrepancies, and the potential correlation between year of degree and starting year from parts a and b.**

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.

**See Table 3 below for all parts of Problem 3**

* 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.
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.

**See Table 3 below for all parts of Problem 4**

* 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.
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.

**See Table 3 below for all parts of Problem 5. For this analysis, the ratio of the mean monthly salary between women and men was analyzed using Poisson regression.**

* 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.

**Table 3. Estimated comparison in monthly salary between women relative to men**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Adjustment** | **Estimate** | **t-statistic / z-statistic** | **P Value** | **95% CI minimum** | **95% CI maximum** |
| **Difference in Mean Monthly Salary** |
| Unadjusted | -1334.7 | -14.04 | <0.0001 | -1521.2 | -1148.3 |
| Degree | -1266.2 | -13.40 | <0.0001 | -1451.6 | -1080.8 |
| Degree, Year of degree | -614.1 | -7.170 | <0.0001 | -782.2 | -446.0 |
| Degree, Year of degree, Starting year | -614.6 | -7.060 | <0.0001 | -785.3 | -443.8 |
| Degree, Year of degree, Starting year, Field  | -420.1 | -5.050 | <0.0001 | -583.1 | -257.0 |
| Degree, Year of degree, Starting year, Field, Admin duties | -419.7 | -5.17 | <0.0001 | -579.0 | -260.5 |
| Degree, Year of degree, Starting year, Field, Admin duties, Rank | -280.7 | -4.08 | <0.0001 | -415.5 | -145.8 |
| **Ratio of Geometric Mean Monthly Salary** |
| Unadjusted | 0.8120 | -13.73 | <0.0001 | 0.7882 | 0.8365 |
| Degree | 0.8204 | -13.09 | <0.0001 | 0.7964 | 0.8451 |
| Degree, Year of degree | 0.9090 | -6.990 | <0.0001 | 0.8850 | 0.9337 |
| Degree, Year of degree, Starting year | 0.9087 | -6.980 | <0.0001 | 0.8845 | 0.9335 |
| Degree, Year of degree, Starting year, Field  | 0.9362 | -5.060 | <0.0001 | 0.9126 | 0.9605 |
| Degree, Year of degree, Starting year, Field, Admin duties | 0.9363 | -5.170 | <0.0001 | 0.9132 | 0.9600 |
| Degree, Year of degree, Starting year, Field, Admin duties, Rank | 0.9574 | -4.080 | <0.0001 | 0.9376 | 0.9776 |
| **Ratio of Mean Monthly Salary** |
| Unadjusted | 0.8017 | -13.58 | <0.0001 | 0.7765 | 0.8277 |
| Degree | 0.8105 | -12.98 | <0.0001 | 0.7852 | 0.8366 |
| Degree, Year of degree | 0.9008 | -7.090 | <0.0001 | 0.8751 | 0.9272 |
| Degree, Year of degree, Starting year | 0.9008 | -7.010 | <0.0001 | 0.8749 | 0.9275 |
| Degree, Year of degree, Starting year, Field  | 0.9286 | -5.220 | <0.0001 | 0.9032 | 0.9548 |
| Degree, Year of degree, Starting year, Field, Admin duties | 0.9289 | -5.340 | <0.0001 | 0.9041 | 0.9544 |
| Degree, Year of degree, Starting year, Field, Admin duties, Rank | 0.9512 | -4.300 | <0.0001 | 0.9298 | 0.9732 |

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 in problems 3-5 were conducted to evaluate and association between salaries paid to men and women, while adjusting for the same combinations of variables, but the questions each problem was addressing were slightly different. Problem 3 evaluated the difference in mean monthly salary between men and women, problem 4 evaluated the ratio of the geometric mean salary between men and women and therefore used a log-transformed salary variable, and problem 5 evaluated the ratio of the mean monthly salary between men and women using Poisson regression, in this case. For all three approaches, the precision increased with adjustment for more covariates. While the p-values were all highly statistically significant (<0.0001), the t- or z-statistic (presented in Table 3) decreased in magnitude as more covariates were included in the adjustments. Similarly, the 95% confidence intervals became narrower with adjustment for more covariates. While the point estimates are slightly different between the three analysis approaches, they are quite similar; however, the first analysis comparing the difference in means is more conservative than the second analysis using log-transformed salary data, which in turn is more conservative than the analysis using Poisson regression. Overall the inference would be essentially the same as all suggest a highly statistically significant difference in mean salaries between men and women.**

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.

**Methods: Mean salaries between male and female faculty were compared to identify an association between salary and sex. A linear regression model on log-transformed salary data was used to compare the ratio of geometric mean monthly salaries between men and women in the year 1995. By using log-transformed data we can evaluate the data in a multiplicative model as salary changes over time tend to behave in such a fashion. Also, we can attenuate the effect of outliers in the data, and make inference based upon the ratio of geometric mean salaries between females and males. Robust standard error estimation with the Huber-White sandwich estimator was used to account for potential heteroscedasticity and potential correlation among covariates. The model adjusted for several variables considered to be potential confounders, including highest degree received, the year that the degree was awarded, the year when employment began, what field of study the employee is in, whether or not the employee performs administrative duties, and faculty rank. No adjustment was made for the multiple comparisons that occur between strata of degree, field, and rank so p values and 95% confidence intervals between these strata should be interpreted as descriptive statistics only. The variables for the year when degree was awarded (*yrdeg*), and the year when employment began (*startyr*) were modeled using linear splines with knots defined at the years 1960, 1965, 1970, 1975, 1980, 1985, and 1990 to capture nonlinearities in the data. Two-sided p-values and 95% confidence intervals were Wald-based estimates. Statistical significance was defined by p<0.05.**

**Results: A total of 1597 faculty members who were employed by the University of Washington in 1995 provided data for this analysis. Overall, the mean monthly salary was $6,389.81 when not adjusting for any covariates. In 1995, approximately 25.6% of faculty were female. Linear regression on log-transformed salary data showed evidence of a statistically significant difference in mean monthly salaries between men and women, after adjusting for all covariates described in the Methods section above. The mean salary for female faculty was 4.257% lower (ratio of geometric mean monthly salary 0.9574) than that of men who were equivalent in field, degree, rank, year of degree, year starting employment at the University of Washington, and performance of administrative duties. This difference was statistically significant (p<0.0001), and was consistent with a true population difference falling within the 95% confidence interval with salaries paid to women from 2.236% to 6.237% lower than equivalent men (95% confidence interval of the ratio of geometric mean salaries 0.9376 to 0.9776). From this analysis, we reject the null hypothesis that there is no difference in mean monthly salaries based on sex, in favor of the alternative hypothesis that there is a difference in salary practices among male and female faculty.**