**8392**

**Biost 515 (Winter 2014)**

**Instructor: Scott Emerson**

**Homework 8**

*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.

**Answer:** The variables *degree* and *field* are unordered categorical variables and the variable *admin* is a binary variable. As such, the best way to model the variables *degree*, *field*, and *admin* is as indicator variables for each potential category within each variable.

* 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.

**Answer:** Inference based on classical linear regression (without robust standard error estimates) would be incorrect in our analysis concerning the difference in mean monthly salary paid to women faculty in 1995 versus that paid to men faculty in 1995. Since an individual’s salary in 1995 is based on a rate of pay increase since the individual’s start year and/or year of degree, we must account for these varying years in female faculty and in male faculty. That is, we must account for the mean-variance relationship by using robust standard errors. Classical linear regression would tend to provide conservative estimates because its’ corresponding standard error will be larger.

* 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).

 **Answer: (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).

 **Answer: (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).

 **Answer: (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).

 **Answer: (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.)

 **Answer: (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?

 **Answer:** 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 continuous 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?

**Answer:** In a real situation, I would have a priori chosen a model which treats year of degree and start year as linear splines variables. I do not believe there is ever an instance in which continuous variables should be transformed into dummy variables, and I also have no intuitive reason to believe that year of degree and start year may be quadratic continuous predictors of salary. Furthermore, although I could have treated the variables as continuous and linear, I believe that enough latent variables (e.g. changes in gender discrimination over the years) exist for me to justify knots in the linearity of my 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.
	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).

**Answer:** (See table below)

* 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).

**Answer:** (See table below)

* 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).

**Answer:** (See table below)

* 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).

**Answer:** (See table below)

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

**Answer:** The following table suggests confounding between degree year and starting year in each variable’s association with mean monthly salary. If we model year of degree without adjusting for start year, we are estimating the mean difference of salary within faculty with the same degree year but who may not have worked at UW for the same length of time. It is not scientifically relevant to draw inference from these results, since for example, we may be comparing a recent department chair hire to an associate professor with administrative duties, who will tend to earn a lower wage. If we adjust for starting year, our comparisons are a more accurate reflection of the differences in mean salary according to degree year.

 Similarly, if we model starting year without adjusting for year of degree, we are again making potentially erroneous comparisons. In this model, we are comparing a new hire who recently received her PhD and is entering as an assistant professor to a recent department chair hire. Clearly, the salary differences in these two faculty members will not be scientifically meaningful. Adjusting for year of degree will correct for these situations.

|  |
| --- |
| **Table 2: Difference in Mean Monthly Salary by Year of Degree & Starting Year, 19951** |
|  | **Estimate** | **T** | **P-Value** | **95% CI Low** | **95% CI High** |
| Year of degree |  -56.93 | -12.06 | <0.001 |  -66.19  | -47.68 |
|  Adjusted2 | -106.94 | -11.70 | <0.001 | -124.87 | -89.01 |
| Starting Year |  -88.56 | -20.90 | <0.001 |  -96.87 | -80.25 |
|  Adjusted3 |  23.17 |  2.53 |  0.012 |  5.18 |  41.15 |
| 1All models adjust for degree, field, administrative duties, and sex.2Model adjusts for starting year.3Model adjusts for year of degree. |

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.

**Answer:** (See table below)

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

**Answer:** (See table below)

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

**Answer:** (See table below)

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

**Answer:** (See table below)

* 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.

**Answer:** (See table below)

* 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.*

**Answer:** (See table below; predicted values saved)

* 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.

**Answer:** (See table below)

|  |
| --- |
| **Table 3: Difference in Mean Monthly Salary by Sex, 19951** |
|  | **Estimate** | **t** | **P Value** | **95% CI Low** | **95% CI High** |
| Unadjusted | -1334.73 | -14.04 | <0.001 | -1521.18 | -1148.29 |
| +Degree+Year of Degree+Start Year at UW | -1262.20 -629.82 -642.99 | -13.32  -7.35 -7.40 | <0.001<0.001<0.001 | -1448.02 -797.96  -813.54  | -1076.37 -461.68 -472.44 |
| +Field+Admin Duties |  -443.98 -445.96  |  -5.32 -5.47 | <0.001<0.001 |  -607.65 -605.74  |  -280.31 -286.19 |
| +Rank |  -290.50 |  -4.15 | <0.001 |  -427.86  |  -153.14 |
| 1Difference = women - men.+ Model adjusts for previous adjustment variables and current variable. |

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.

**Answer:** (See table below)

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

**Answer:** (See table below)

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

**Answer:** (See table below)

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

**Answer:** (See table below)

* 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.

**Answer:** (See table below)

* 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.*

**Answer:** (See table below; predicted values saved)

* 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.

**Answer:** (See table below)

|  |
| --- |
| **Table 4: Ratio of Geometric Mean Monthly Salary by Sex, 19951** |
|  | **Estimate** | **z** | **P Value** | **95% CI Low** | **95% CI High** |
| Unadjusted | 0.8120 | -13.74 | <0.001 | 0.7882 | 0.8365 |
| +Degree+Year of Degree+Start Year at UW | 0.82090.90680.9053 | -13.03 -7.18 -7.29 | <0.001<0.001<0.001 | 0.79680.88290.8814 | 0.84560.93130.9298 |
| +Field+Admin Duties | 0.93350.9332 |  -5.29 -5.44 | <0.001<0.001 | 0.91000.9102 | 0.95760.9567 |
| +Rank | 0.9566 |  -4.12 | <0.001 | 0.9366 | 0.9770 |
| 1Ratio = women/men.+ Model adjusts for all previous adjustment variables and current variable. |

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.

(I used the generalized linear model for this question)

* 1. Report inference regarding the unadjusted comparison of women’s and men’s salaries.

**Answer:** (See table below)

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

**Answer:** (See table below)

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

**Answer:** (See table below)

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

**Answer:** (See table below)

* 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.

**Answer:** (See table below)

* 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.*

**Answer:** (See table below; predicted values saved)

* 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.

**Answer:** (See table below)

|  |
| --- |
| **Table 5: Ratio of Mean Monthly Salary by Sex, 19951** |
|  | **Estimate** | **z** | **P Value** | **95% CI Low** | **95% CI High** |
| Unadjusted | 0.8017 | -13.58  | <0.001 | 0.7765 | 0.8277 |
| +Degree+Year of Degree+Start Year at UW | 0.8099 0.8951 0.8915  | -12.96 -7.33 -7.41 | <0.001<0.001<0.001 | 0.78450.86900.8648 | 0.83610.92200.9190 |
| +Field+Admin Duties | 0.9207 0.9195 |  -5.58 -5.85 | <0.001<0.001 | 0.89430.8941  | 0.94780.9457 |
| +Rank | 0.9483  |  -4.33 | <0.001 | 0.9258  | 0.9714 |
| 1Ratio = women/men.+ Model adjusts for all previous adjustment variables and current variable. |

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?

**Answer:**  The following graph (next page) displays loess curves derived from the predicted monthly salary for male and for female faculty using simple linear (LR), log transformed linear (Log LR), and generalized linear with a log link function (GLM) models. Each model adjusts for degree, year of degree, starting year at UW, field, and administrative duties. For comparison, a loess curve of faculties’ salaries within the sample is also plotted.

 As shown, the linear regression model with the log transformation of salary consistently estimated lower salaries for both genders than the other two models. The simple linear regression and the generalized linear regression models had very similar fitted values. These differences aside, it is very clear in all three models that a significant difference in mean salary exists between males and females. Therefore, any of the three models could potentially be a good fit for this dataset, and the same conclusions could be drawn from each one.



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.

**Answer:**  A priori, I would have estimated any discrimination against women with a simple linear regression model to test for a difference in mean monthly salary between men and women.

**Methods:** The difference in estimated mean salary of UW faculty members in 1995 was compared between males and females using a linear regression model and adjusting for degree, year of degree, starting year at UW, field, and administrative duties. Statistical inference on the difference in mean salary between sexes holding all else equal was based on the Wald statistic computed from the regression slope parameter and its standard error as estimated using the Huber-White sandwich estimator. A two-sided p-value and 95% confidence interval were computed using the approximate normal distribution for linear regression parameter estimates.

**Inference:** In 1995, the observed mean monthly salary for female faculty members at UW was $5397, while for male faculty members, the observed mean monthly salary was $6732. The estimated difference in mean salary between males and females with the same degree, year of degree, starting year at UW, field, and administrative duties was $290.5, with females being awarded the lower salary. Based on a 95% confidence interval, this estimated difference in salaries would not be judged unusual if males were truly awarded salaries anywhere between $427.9 and $153.1 higher than females with all else equal. These results are statistically sufficient enough (p-value < 0.001) to reject the null hypothesis of no difference in salaries between men and women in favor of a hypothesis of gender discrimination at UW.