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.

**Admin is a binary variable so it should be modeled as a binary variable, it is already dichotomous and a dummy variable would fit it exactly so there is nothing to be gained or lost by modeling it as anything else. Field and degree are both unordered categorical variables. Because they are unordered it would not make sense to model them as continuous variables because the interpretation would depend on the order and there is no clear agreed upon order for either of these variables. The best way to model unordered categorical variables is with dummy variables. This will fit a step function across each category and order will not matter.**

* 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.

**Classical linear regression assumes homoscedasticity. Therefore, if data is heteroscedastic you could reject because the means are different or because the variances are different. If your question is about the means then you only want to reject if the means are truly different, not the variance. Using robust standard errors allows for heteroscedasticity so inference is only based on the means. In this dataset there are fewer females (n=409, sd = $1481.22) than males (n=1188, sd = $2089.76) and the standard deviation in monthly salaries is smaller for females than for males. Because the group with the smaller sample size has smaller variance I would expect any inference made using classical linear regression to be 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?

**I would model starting year and year of degree as continuous variables. Even though there is some evidence of nonlinearity from the splines the overall trend is the same. Additionally, because both of these variables are confounders I would not be overly concerned with fitting them exactly. The linear, continuous variables are also much easier to interpret than splines.**

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

**Year of degree and starting year are highly correlated (r=0.79), for most observations in this data set the starting year is within a few years of the year that highest degree was obtained. We would therefore expect both variables to measure similar latent variables which is the case in parts a and b. In part a the year that highest degree was obtained is used as a measure of experience and in part b starting year is used primarily as a measure of experience. However, when both variables are included in the model, as in parts c and d, the latent variable that is dominant for each measure changes. The interpretation for each parameter is the difference per unit change, with all other covariates held constant. This means that while year of degree continues to be a measure of experience, start year adjusted for year of degree becomes a measure of performance and productivity as well as experience. Individuals who received their degree in the same year but were hired 10 years apart represent very different hiring decisions and priorities and therefore different latent variables.**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **2.** | **POI** | **Estimate** | **Z** | **P Value** | **95% CI low** | **95% CI high** | **Interpretation** |
| **a** | Year of degree | -89.87 | -20.89 | <0.001 | -98.30 | -81.43 | Holding sex, field, degree, and administrative duties constant, for every year later that highest degree was obtained monthly there is an estimated $89.87 decrease in monthly salary. |
| **b** | Starting year | -56.88 | -12.06 | <0.001 | -66.13 | -47.63 | Holding sex, field, degree, and administrative duties constant, for every year later that an individual was hired there is an estimated $56.88 decrease in monthly salary. |
| **c** | Year of degree | -111.96 | -11.79 | <0.001 | -130.58 | -93.34 | Holding sex, field, degree, and administrative duties constant, for individuals who were hired in the same year there is an estimated $111.96 decrease in monthly salary for every year later their highest degree was obtained. |
| **d** | Starting year | 27.15 | 2.88 | 0.004 | 8.68 | 45.63 | Holding sex, field, degree, and administrative duties constant, for individuals who got their highest degree in the same year there is an estimated $27.15 increase in monthly salary for every year later they were hired. |

Problems 3 – 5 ask you to fit a series of models in which you consider a hierarchy of adjusted analyses for each of three different summary measures. Your response to these problems might be best presented in a table of inference about the adjusted association between monthly salary and sex.

For the benefit of the graders, we will agree on modeling *yrdeg* and *startyr* as linear splines as computed in problem 1f.

1. We are interested in making inference about the difference in the mean monthly salary paid to women faculty in 1995 and that paid to men faculty in 1995.
	1. Report inference regarding the unadjusted comparison of women’s and men’s salaries.
	2. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree.
	3. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree.
	4. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree, starting year at UW.
	5. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree, starting year at UW, field.
	6. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree, starting year at UW, field, administrative duties. Save the predicted values of the mean salary for each individual as *fit3.*
	7. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree, starting year at UW, field, administrative duties, rank.

**The unadjusted estimate suggests a higher disparity in salary than the adjusted estimates. Compared to only adjusting for degree and year of degree a slight increase in the difference was seen when also adjusting for starting year. This is likely due to the fact that starting year and year of degree are such similar measures.**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **3.** | **Adjusting For** | **Estimate** | **Z** | **P Value** | **95% CI low** | **95% CI high** | **Interpretation** |
| **a** | None | -1334.73 | -14.04 | <0.001 | -1521.18 | -1148.29 | The estimated mean monthly salary is $1,334.73 less for females than for males. |
| **b** | Degree | -1266.15 | -13.40 | <0.001 | -1451.56 | -1080.75 | Holding degree constant, there is an estimated $1,266.15 decrease in mean monthly salary for females than for males. |
| **c** | Degree, year of degree | -614.13 | -7.17 | <0.001 | -782.24 | -446.02 | Holding degree and year of degree constant, there is an estimated $614.13 decrease in mean monthly salary for females than for males. |
| **d** | Degree, year of degree, starting year | -614.58 | -7.06 | <0.001 | -785.31 | -443.85 | Holding degree, year of degree, and starting year at UW constant there is an estimated $614.58 decrease in mean monthly salary for females than for males. |
| **e** | Degree, year of degree, starting year, field | -420.05 | -5.05 | <0.001 | -583.12 | -256.99 | Holding degree, year of degree, starting year at UW, and field constant there is an estimated $420.05 decrease in mean monthly salary for females than for males. |
| **f** | Degree, year of degree, starting year, field, administrative duties | -419.73 | -5.17 | <0.001 | -578.99 | -260.47 | Holding degree, year of degree, starting year at UW, field, and administrative duties constant there is an estimated $419.73 decrease in mean monthly salary for females than for males. |
| **g** | Degree, year of degree, starting year, field, administrative duties, rank | -280.66 | -4.08 | <0.001 | -415.52 | -145.81 | Holding degree, year of degree, starting year at UW, field, administrative duties, and rank constant there is an estimated $280.66 decrease in mean monthly salary for females than for males. |

1. We are interested in making inference about the ratio of geometric mean monthly salary paid to women faculty in 1995 and that paid to men faculty in 1995.
	1. Report inference regarding the unadjusted comparison of women’s and men’s salaries.
	2. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree.
	3. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree.
	4. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree, starting year at UW.
	5. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree, starting year at UW, field.
	6. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree, starting year at UW, field, administrative duties. Save the predicted values of the geometric mean salary for each individual as *fit4.*
	7. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree, starting year at UW, field, administrative duties, rank.

**The unadjusted estimate suggests a higher disparity in salary than the adjusted estimates. Compared to only adjusting for degree and year of degree a slight increase in the difference was seen when also adjusting for starting year. This is likely due to the fact that starting year and year of degree are such similar measures.**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **4.** | **Adjusting For** | **Estimate** | **Z** | **P Value** | **95% CI low** | **95% CI high** | **Interpretation** |
| **a** | None | 0.8120 | -13.73 | <0.001 | 0.7882 | 0.8365 | The estimated geometric mean monthly salary is 18.80% less for females than for males. |
| **b** | Degree | 0.8204 | -13.09 | <0.001 | 0.7964 | 0.8451 | Holding degree constant, there is an estimated 17.96% decrease in the geometric mean monthly salary for females than for males. |
| **c** | Degree, year of degree | 0.9090 | -6.99 | <0.001 | 0.8850 | 0.9337 | Holding degree and year of degree constant, there is an estimated 9.10% decrease in geometric mean monthly salary for females than for males. |
| **d** | Degree, year of degree, starting year | 0.9087 | -6.98 | <0.001 | 0.8845 | 0.9335 | Holding degree, year of degree, and starting year at UW constant there is an estimated 9.13% decrease in geometric mean monthly salary for females than for males. |
| **e** | Degree, year of degree, starting year, field | 0.9362 | -5.06 | <0.001 | 0.9126 | 0.9605 | Holding degree, year of degree, starting year at UW, and field constant there is an estimated 6.38% decrease in geometric mean monthly salary for females than for males. |
| **f** | Degree, year of degree, starting year, field, administrative duties | 0.9363 | -5.17 | <0.001 | 0.9132 | 0.9600 | Holding degree, year of degree, starting year at UW, field, and administrative duties constant there is an estimated 6.37% decrease in geometric mean monthly salary for females than for males. |
| **g** | Degree, year of degree, starting year, field, administrative duties, rank | 0.9574 | -4.08 | <0.001 | 0.9376 | 0.9776 | Holding degree, year of degree, starting year at UW, field, administrative duties, and rank constant there is an estimated 4.26% decrease in geometric mean monthly salary for females than for males. |

1. We are interested in making inference about the ratio of the mean monthly salary paid to women faculty in 1995 and that paid to men faculty in 1995. You can use Poisson regression (with the irr option to get exponentiated parameter estimates), or you can use a generalized linear model with a log link. Stata has a regression function “glm” that allows the specification of a log link function. Hence, you can fit the regression for part a using the command

glm salary female if year==95, link(log) robust

Parameter estimates will be interpretable as the log mean (intercept) and log mean ratio (slope). (glm stands for “generalized linear model” and it includes as special cases linear regression, logistic regression, and Poisson regression. By default, it presumes the data are continuous and models the mean according to the value of the link function.) By specifying the “eform” option, it will return the exponentiated parameter estimates. In either case, make clear which analysis method you used.

* 1. Report inference regarding the unadjusted comparison of women’s and men’s salaries.
	2. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree.
	3. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree.
	4. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree, starting year at UW.
	5. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree, starting year at UW, field.
	6. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree, starting year at UW, field, administrative duties. Save the predicted values of the mean salary for each individual as *fit5.*
	7. Report inference regarding the comparison of women’s and men’s salaries after adjustment for degree, year of degree, starting year at UW, field, administrative duties, rank.

**The unadjusted estimate suggests a higher disparity in salary than the adjusted estimates. Compared to only adjusting for degree and year of degree a slight increase in the difference was seen when also adjusting for starting year. This is likely due to the fact that starting year and year of degree are such similar measures.**

Using Poisson regression with the irr option

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **5.** | **Adjusting For** | **Estimate** | **Z** | **P Value** | **95% CI low** | **95% CI high** | **Interpretation** |
| **a** | None | 0.8017 | -13.58 | <0.001 | 0.7765 | 0.8277 | The estimated mean monthly salary is 19.83% less for females than for males. |
| **b** | Degree | 0.8105 | -12.98 | <0.001 | 0.7852 | 0.8366 | Holding degree constant, there is an estimated 18.95% decrease in mean monthly salary for females than for males. |
| **c** | Degree, year of degree | 0.9008 | -7.09 | <0.001 | 0.8751 | 0.9272 | Holding degree and year of degree constant, there is an estimated 9.92% decrease in mean monthly salary for females than for males. |
| **d** | Degree, year of degree, starting year | 0.9008 | -7.01 | <0.001 | 0.8749 | 0.9275 | Holding degree, year of degree, and starting year at UW constant there is an estimated 9.92% decrease in mean monthly salary for females than for males. |
| **e** | Degree, year of degree, starting year, field | 0.9286 | -5.22 | <0.001 | 0.9032 | 0.9548 | Holding degree, year of degree, starting year at UW, and field constant there is an estimated 7.14% decrease in mean monthly salary for females than for males. |
| **f** | Degree, year of degree, starting year, field, administrative duties | 0.9289 | -5.34 | <0.001 | 0.9041 | 0.9544 | Holding degree, year of degree, starting year at UW, field, and administrative duties constant there is an estimated 7.11% decrease in mean monthly salary for females than for males. |
| **g** | Degree, year of degree, starting year, field, administrative duties, rank | 0.9512 | -4.30 | <0.001 | 0.9298 | 0.9732 | Holding degree, year of degree, starting year at UW, field, administrative duties, and rank constant there is an estimated 4.88% decrease in mean monthly salary for females than for males. |

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 results obtained for each model show a high degree of overlap indicating agreement between all the predicted fits. The model for difference in means gives slightly lower estimates than either the ratio of geometric means or the ratio of means. However, the inference obtained from each model would be the same: adjusting for degree, year of degree, start year at UW, field, and administrative duties there is a trend for lower monthly salaries for women compared to men.**

1. For the analysis model that you would have chosen *a priori*, summarize the scientific relevance of the single model that you think would best reflect any discrimination against women in awarding salaries. Give a formal report of your methods and results.

***A priori* I would have chosen to use linear regression on the log transformed salary adjusted for degree, year of degree, start year at UW, field, and administrative duties. I would not adjust for rank as that could be in the causal pathway if women are discriminated against in promotion decisions leading to lower salaries. However, adjustment for the other covariates should allow for distinction between historical practices regarding any discrimination or disparity in women’s pay and current practices.**

**Methods: Regression analysis was used to address the question of whether there is evidence of current sex based discrimination in the monthly salaries of male and female faculty at the University of Washington. The ratio of geometric mean monthly salaries across sex groups was compared using linear regression on log transformed salary data. The model was adjusted for degree, year of degree, start year at UW, field, and administrative duties. Year of degree and start year at UW were modeled as continuous variables. Degree, field, and administrative duties were modeled as dummy variables. The analysis was restricted to the year 1995 to prevent issues due to faculty with multiple measurements. P-values and 95% confidence intervals were calculated using the Huber-White sandwich estimator of the standard errors.**

**Inference: Linear regression analysis was performed on 1,597 subjects having mean monthly salary of $6,389.81 (SD $2,036.77; range $3,042.00-$14,464.00) in the year 1995. Of these subjects 1,188 were male with a mean monthly salary of $6,731.64 (SD $2,089.76; range $3,130.59-$14,464.00) and 409 were female with a mean monthly salary of $5,396.91 (SD $1,481.22; range $3,042.00-$11,036.00). Holding degree, year of degree, starting year at UW, field, and administrative duties constant the estimated ratio of geometric mean monthly salaries is 0.9347, with women tending to have a 6.53% lower monthly salary than men. A 95% confidence interval suggests that this observation is not unusual if the true relative difference in geometric mean monthly salaries was anywhere from 4.13% to 8.87% lower for women than for men. Using a two-sided p-vale this observation is highly statistically significant at a 0.05 level of significance (p < 0.001) and we reject the null hypothesis that the geometric mean monthly salary is the same for men and women in favor of a tendency for lower monthly salary for women.**