**Homework #8**

1. A. The best way to model degree, field, and admin is as dummy variables, since these are unordered categorical variables.B. Inference based on classical linear regression would be incorrect because classical linear regression would assume that the variance of salary is the same within each group that is homogeneous with respect to sex, degree, field, and administrative duties, which is most likely not true since some fields and degrees lead to a large range of potential salaries, while others lead to a more standard, smaller range of potential salaries. Classical linear regression inference about salary would tend to be conservative if the smaller groups defined by a given variable had lower variance, while it would tend to be anti-conservative if the smaller groups defined by a given variable had higher variance. We might expect that of those people who had an earlier year of degree or an earlier start year, fewer will still be in the dataset at 1995 because many will have retired by this point. Note that when two people first attain their degree (or when they first start at the university), their salaries are more likely to be close than when it has been many years since they attained their degree (or began working at the university). This is because early after graduation (or starting work at the university), all people are viewed as approximately equally valuable, and so they will be paid about the same, while years later, some people are more productive and will have received many raises, while others are less productive and will not have received many raises. This gap between those who are productive and those who are not will create greater variance in salary amongst those people who have an earlier year of degree or an earlier start year, and hence, we expect that classical linear regression would tend to give anti-conservative inference for this data.I. In a real situation, we would choose the method for adjusting for year of degree and starting year that made the most sense for addressing our scientific question. Modeling these as dummy variables does not make much sense, since we want our within 5-year group estimates to borrow information from other group estimates. However, if we want to create a more flexible model that gives more accurate within-group estimates if there is a nonlinear trend we would not want to model year of degree or starting year as linear continuous variables. Also, modeling year of degree or starting year as quadratic continuous variables gives inference that is difficult to interpret, and since we have no scientific reason to believe that the squared year of degree or the squared starting year have any relevance to salary, it makes no sense to include quadratic terms in our model. Hence, we should model year of degree and starting year as linear splines with knots 5 years apart, since this will allow us to detect some nonlinear trends, as well as easing interpretation.

2. A. Methods: We tested to see if there was an association between monthly salary and year of degree modeled as a continuous variable, after adjusting for degree, field, administrative duties, and sex, by using a linear regression model that treated salary as the response and year of degree as the predictor of interest. Salaries were compared across groups homogeneous with respect to degree, field, administrative duties, and sex, and differing one year in the year of degree. Inference was made using Huber-White robust standard errors, and a point estimate, 95% confidence interval, and two-sided p-value were derived for the adjusted association between salary and year of degree.Inference: We found that among groups homogeneous with respect to degree, field, administrative duties, and sex, and differing one year in the year of degree, the mean salary was $89.87 lower for those with the later year of degree. This point estimate would not be unusual if the true salary for those with a degree one year later was anywhere from $81.43 to $98.30 lower than for those with the earlier year of degree, with a two-sided p-value less than 0.0005 and a t-statistic of -20.89. Thus, we can with high confidence reject the null hypothesis that there is no adjusted association between salary and year of degree.B. Methods: We tested to see if there was an association between monthly salary and starting year modeled as a continuous variable, after adjusting for degree, field, administrative duties, and sex, by using a linear regression model that treated salary as the response and starting year as the predictor of interest. Salaries were compared across groups homogeneous with respect to degree, field, administrative duties, and sex, and differing one year in the starting year. Inference was made using Huber-White robust standard errors, and a point estimate, 95% confidence interval, and two-sided p-value were derived for the adjusted association between salary and starting year.Inference: We found that among groups homogeneous with respect to degree, field, administrative duties, and sex, and differing one year in the starting year, the mean salary was $56.88 lower for those with the later starting year. This point estimate would not be unusual if the true salary for those with a starting year one year later was anywhere from $47.63 to $66.13 lower than for those with the earlier starting year, with a two-sided p-value less than 0.0005 and a t-statistic of -12.06. Thus, we can with high confidence reject the null hypothesis that there is no adjusted association between salary and starting year.C. Methods: We tested to see if there was an association between monthly salary and year of degree modeled as a continuous variable, after adjusting for starting year, degree, field, administrative duties, and sex, by using a linear regression model that treated salary as the response and year of degree as the predictor of interest. Salaries were compared across groups homogeneous with respect to starting year, degree, field, administrative duties, and sex, and differing one year in the year of degree. Inference was made using Huber-White robust standard errors, and a point estimate, 95% confidence interval, and two-sided p-value were derived for the adjusted association between salary and year of degree.Inference: We found that among groups homogeneous with respect to starting year, degree, field, administrative duties, and sex, and differing one year in the year of degree, the mean salary was $111.96 lower for those with the later year of degree. This point estimate would not be unusual if the true salary for those with a degree one year later was anywhere from $93.34 to $130.58 lower than for those with the earlier year of degree, with a two-sided p-value less than 0.0005 and a t-statistic of -11.79. Thus, we can with high confidence reject the null hypothesis that there is no adjusted association between salary and year of degree.D. Methods: We tested to see if there was an association between monthly salary and starting year modeled as a continuous variable, after adjusting for year of degree, degree, field, administrative duties, and sex, by using a linear regression model that treated salary as the response and starting year as the predictor of interest. Salaries were compared across groups homogeneous with respect to year of degree, degree, field, administrative duties, and sex, and differing one year in the starting year. Inference was made using Huber-White robust standard errors, and a point estimate, 95% confidence interval, and two-sided p-value were derived for the adjusted association between salary and starting year.Inference: We found that among groups homogeneous with respect to year of degree, degree, field, administrative duties, and sex, and differing one year in the starting year, the mean salary was $27.15 lower for those with the later starting year. This point estimate would not be unusual if the true salary for those with a starting year one year later was anywhere from $8.68 to $45.63 lower than for those with the earlier starting year, with a two-sided p-value of 0.004 and a t-statistic of 2.88. Thus, we can with high confidence reject the null hypothesis that there is no adjusted association between salary and starting year.E. The results obtained for the association between monthly salary and year of degree in part A only compared salary between groups of people differing in year of degree by one year and having the same degree, field, administrative duties, and sex, while the results obtained for the association between monthly salary and year of degree in part C compared salary between groups of people differing in year of degree by one year and having the same degree, field, administrative duties, and sex, as well as the same starting year. We can see that both parts showed that salary would be lower when comparing a group with a year of degree one year later to a group with an earlier year of degree. However, in part A, the magnitude of the unadjusted difference in salaries was smaller than for the adjusted difference in salaries from part C. However, the confidence interval is larger in part C, to the point where it includes one of the bounds of the 95% confidence interval from part A. Hence, we cannot with high confidence say that the inference for the adjusted and the unadjusted regression models are different, and so comparing only those people with the same starting year may not change our results. The results obtained for the association between monthly salary and starting year in part A only compared salary between groups of people differing in starting year by one year and having the same degree, field, administrative duties, and sex, while the results obtained for the association between monthly salary and starting year in part C compared salary between groups of people differing in starting year by one year and having the same degree, field, administrative duties, and sex, as well as the same year of degree. We can see that in part B, inference shows that unadjusted salary would be lower when comparing a group with a starting year one year later to a group with an earlier starting year, but in part D, inference shows that adjusted salary would be higher when comparing a group with a starting year one year later to a group with an earlier starting year. The 95% confidence intervals from these two parts are non-overlapping, so we can with high confidence say that the inference for the adjusted and the unadjusted regression models are different, and comparing only those people with the same starting year changes our results from giving a negative association to giving a positive association.

3. Below is a table listing inference made on the difference of mean salaries between men and women, adjusted for the variables described in the homework assignment. The letter for the part corresponding to the adjustment is in the left-most column, and the point estimate for the difference in mean salaries between men and women (salary for women minus salary for men), t-statistic, two-sided p-value, and 95% confidence interval endpoints for inference with robust standard errors are given in each other column.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Part** | **Estimate** | **t** | **P Value** | **95% CI low** | **95% CI high** |
| **A** | -1334.73 | -14.04 | < .0001 | -1521.18 | -1148.29 |
| **B** | -1266.15 | -13.40 | < .0001 | -1451.56 | -1080.75 |
| **C** | -614.13 | -7.17 | < .0001 | -782.24 | -446.02 |
| **D** | -614.58 | -7.06 | < .0001 | -785.31 | -443.85 |
| **E** | -420.05 | -5.05 | < .0001 | -583.12 | -256.99 |
| **F** | -419.73 | -5.17 | < .0001 | -578.99 | -260.47 |
| **G** | -280.66 | -4.08 | < .0001 | -415.52 | -145.81 |

4. Below is a table listing inference made on the ratio of geometric mean salaries between men and women, adjusted for the variables described in the homework assignment. The letter for the part corresponding to the adjustment is in the left-most column, and the point estimate for the ratio of geometric mean salaries between men and women (salary for women divided by salary for men), t-statistic, two-sided p-value, and 95% confidence interval endpoints for inference with robust standard errors are given in each other column.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Part** | **Estimate** | **t** | **P Value** | **95% CI low** | **95% CI high** |
| **A** | 0.8120 | -13.73 | < .0001 | 0.7882 | 0.8365 |
| **B** | 0.8204 | -13.09 | < .0001 | 0.7964 | 0.8451 |
| **C** | 0.9090 | -6.99 | < .0001 | 0.8850 | 0.9337 |
| **D** | 0.9086 | -6.98 | < .0001 | 0.8845 | 0.9335 |
| **E** | 0.9362 | -5.06 | < .0001 | 0.9127 | 0.9605 |
| **F** | 0.9363 | -5.17 | < .0001 | 0.9132 | 0.9599 |
| **G** | 0.9574 | -4.08 | < .0001 | 0.9376 | 0.9777 |

5. Below is a table listing Poisson regression-based inference made on the ratio of mean salaries between men and women, adjusted for the variables described in the homework assignment. The letter for the part corresponding to the adjustment is in the left-most column, and the point estimate for the ratio of mean salaries between men and women (salary for women divided by salary for men), z-statistic, two-sided p-value, and 95% confidence interval endpoints for Poisson regression-based inference with robust standard errors are given in each other column.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Part** | **Estimate** | **z** | **P Value** | **95% CI low** | **95% CI high** |
| **A** | 0.8017 | -13.58 | < .0001 | 0.7765 | 0.8277 |
| **B** | 0.8105 | -12.98 | < .0001 | 0.7852 | 0.8366 |
| **C** | 0.9008 | -7.09 | < .0001 | 0.8751 | 0.9272 |
| **D** | 0.9008 | -7.01 | < .0001 | 0.8749 | 0.9275 |
| **E** | 0.9286 | -5.22 | < .0001 | 0.9032 | 0.9548 |
| **F** | 0.9289 | -5.34 | < .0001 | 0.9041 | 0.9544 |
| **G** | 0.9512 | -4.30 | < .0001 | 0.9298 | 0.9732 |

6. 

The graph above shows the actual salary of patients along the x-axis and the predicted salary from various models along the y-axis, where the blue dots for “Diff fit” are predicted salaries from question 3 part F, the pink dots for “Geom fit” are predicted salaries from question 4 part F, the green dots for “Ratio fit” are predicted salaries from question 5 part F, and the black dots labeled “salary” are a straight line plotting salary against salary, and therefore giving a visual representation of what how far the predicted models are from the truth. Based on this graph, it seems as though all three of these models give very similar predicted salaries since their scatterplots are very similar, with the predicted salaries from question 3 part F being slightly lower on average than those from question 4 part F, which were slightly lower than those from question 5 part F. The inference for each of these models should be very similar.

7. A priori, we would choose ratio of geometric means, since pay is on a multiplicative scale, and it makes more sense to compare someone differing in salary by a factor of 2 than differing in salary by a difference of $100 or some other arbitrary amount. We would adjust for sex since we want to see how salary differs by sex. We would choose a model that adjusted for field, administrative duties, degree, year of degree, and starting year, since these are associated with salary and could be potential confounders. However, we would not adjust for rank, because choosing not to promote someone to a higher rank of professor could be a major form of sex discrimination in itself, and so there might be a smaller sample of women within higher ranks, making comparisons between men and women in the higher ranks less powerful.Methods: Geometric mean monthly salaries were compared between groups of men and of women who were homogeneous with respect to field, administrative duties, degree, year of degree, and starting year. Adjusted differences in the mean of log transformed salaries were tested using a linear regression model which used Huber-White robust standard errors. 95% confidence intervals and two-sided p-values for the difference in population means for log salary were similarly based on that same handling of variances. Estimates and CI were then exponentiated in order to obtain inference on the geometric mean.Inference: We found that among people differing by sex but homogeneous with respect to field, administrative duties, degree, year of degree, and starting year, the geometric mean salary for women was 6.37% lower than the geometric mean salary for comparable men. This point estimate for the adjusted percent difference between salaries of men and women would not be unusual if the true geometric mean salary for women was anywhere from 4.00% to 8.68% lower than the geometric mean salary for comparable men, with a two-sided p-value for this percent difference less than 0.0005 and a t-statistic of -5.17. Thus, we can with high confidence reject the null hypothesis that there is no association between sex and geometric mean salary, after adjusting for field, administrative duties, degree, year of degree, and starting year, in favor of the alternative hypothesis that women have lower salaries than comparable men.