Homework 8

1. Inference about the difference in the mean monthly salary paid to women faculty in 1995 and that paid to men faculty in 1995:
2. The best way to model the variables degree, field, and admin are as dummy variables since they are nominal variables. Dummy variables are the appropriate way to model nominal variables.
3. Inferences based on classical linear regression without robust SE estimates would be incorrect since we cannot assume equal variances across groups. Especially in the context of effect modification, there may be different associations shown across different strata. Classical linear regression inference with robust standard errors would tend to be conservative since it relaxes assumptions about the model fit.
4. In a real situation, I would choose to adjust for year of degree and starting year as a linear model since we would not be losing any information. We know from the answer to the previous question that the quadratic method is not the best method to adjust for year of degree and starting year, and we lose information from cutting intervals if we model year of degree and starting year as a dummy variable or linear splines.
5. Inference about the difference in mean monthly salary adjusting for degree, field, administrative duties, and sex:
6. Method: We perform a linear regression of salary on our predictor of interest year of degree and our covariates degree, field, admin, and sex modeled as dummy variables using robust standard errors which are based on Wald-based tests.

Inference: We find that the slope parameter for year of degree is -89.87, which means that with other covariates held constant, groups differing by one year in year of degree differ in mean salary by 89.87 dollars, with the group having earned their degree later tending towards lower mean salary. A 95% CI suggests that the observed difference is not unusual if the true difference in means between groups differing by one year in year of degree were anywhere between 81.43 dollars and 98.30 dollars, with the group having earned their degree later tending towards lower mean salary.

1. Method: We perform a linear regression of salary on our predictor of interest starting year and our covariates degree, field, admin, and sex modeled as dummy variables using robust standard errors which are based on Wald-based tests.

Inference: We find that the slope parameter for starting year is –56.88, which means that with other covariates held constant, groups differing by one year in starting year differ in mean salary by 56.88 dollars, with the group starting later tending towards lower mean salary. A 95% CI suggests that the observed difference is not unusual if the true difference in means between groups differing by one year in starting year were anywhere between 47.63 dollars and 66.13 dollars, with the group starting later tending towards lower salary.

1. Method: We perform a linear regression of salary on our predictor of interest year of degree and our covariates starting year, degree, field, admin, and sex (with the last three variables modeled as dummy variables) using robust standard errors which are based on Wald-based tests.

Inference: We find that the slope parameter for starting year is -111.96, which means that with other covariates held constant, groups differing by one year in year of degree differ in mean salary by 111.96 dollars, with the group having earned their degree later tending towards lower mean salary. A 95% CI suggests that the observed difference is not unusual if the true difference in means between groups differing by one year in starting year were anywhere between 93.34 dollars and 130.58 dollars, with the group having earned their degree later tending towards lower mean salary.

1. Method: We perform a linear regression of salary on our predictor of interest starting year and our covariates year of degree, degree, field, admin, and sex using robust standard errors which are based on Wald-based tests.

Inference: We find that the slope parameter for starting year is 27.15, which means that with other covariates held constant, groups differing by one year in starting year differ in mean salary by 27.15 dollars, with the group starting later tending towards higher mean salary. A 95% CI suggests that the observed difference is not unusual if the true difference in means between groups differing by one year in starting year were anywhere between 8.68 dollars and 45.63 dollars, with the group starting later tending towards higher salary.

1. The difference between the analyses performed in part a and part c is that part c adjusts for starting year. The difference between the analyses performed in part b and part d is that part d adjusts for year of degree. The scientific relevance is that the variables starting year and year of degree are relevant variables because adjusting for them causes our inferences about the association of salary and sex to change significantly.
2. Inference about the difference in the mean monthly salary paid to women faculty in 1995 and that paid to men faculty in 1995:

We perform a linear regression analysis where we model the variables degree, field, and rank as dummy variables and the variables year of degree and starting year as linear splines.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Adjusting for:** | **Estimate** | **P Value** | **95% CI Low** | **95% CI High** |
| **a** | **Unadjusted** | -1334.73 | <.001 | -1521.18 | -1148.29 |
| **b** | **deg** | -1266.15 | <.001 | -1451.56 | -1080.75 |
| **c** | **and yrdeg** | -614.12 | <.001 | -782.24 | -446.02 |
| **d** | **and startyr** | -614.58 | <.001 | -785.31 | -443.85 |
| **e** | **and field** | -420.05 | <.001 | -583.12 | -256.99 |
| **f** | **and admin** | -419.73 | <.001 | -578.99 | -260.47 |
| **g** | **and rank** | -280.66 | <.001 | -415.52 | -145.81 |

1. Inference about ratio of geometric mean monthly salary paid to women faculty in 1995 and that paid to men faculty in 1995:

We perform a linear regression analysis on geometric means where we model the variables degree, field, and rank as dummy variables and the variable year of degree and starting year as linear splines.

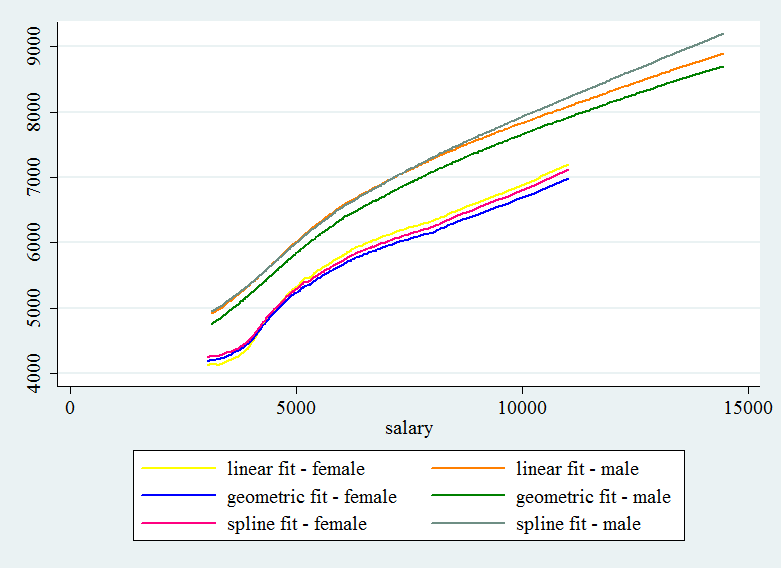
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Adjusting for:** | **Estimate** | **P Value** | **95% CI Low** | **95% CI High** |
| **a** | **Unadjusted** | 0.812 | <.001 | 0.7882 | 0.8365 |
| **b** | **deg** | 0.8204 | <.001 | 0.7964 | 0.8451 |
| **c** | **and yrdeg** | 0.9090 | <.001 | 0.8850 | 0.9337 |
| **d** | **and startyr** | 0.9087 | <.001 | 0.8845 | 0.9335 |
| **e** | **and field** | 0.9362 | <.001 | 0.9126 | 0.9605 |
| **f** | **and admin** | 0.9363 | <.001 | 0.9132 | 0.9600 |
| **g** | **and rank** | 0.9574 | <.001 | 0.9376 | 0.9776 |

1. Inference about ratio of mean monthly salary paid to women faculty in 1995 and that paid to men faculty in 1995:

We perform a poisson regression analysis where we model the variables degree, field, and rank as dummy variables and the variable year of degree and starting year as linear splines.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Adjusting for:** | **Estimate** | **P Value** | **95% CI Low** | **95% CI High** |
| **a** | **Unadjusted** | 0.8017 | <.001 | 0.7765 | 0.8277 |
| **b** | **deg** | 0.8097 | <.001 | 0.7844 | 0.8359 |
| **c** | **and yrdeg** | 0.8981 | <.001 | 0.8719 | 0.9251 |
| **d** | **and startyr** | 0.8964 | <.001 | 0.8695 | 0.9241 |
| **e** | **and field** | 0.9251 | <.001 | 0.8986 | 0.9524 |
| **f** | **and admin** | 0.9245 | <.001 | 0.8989 | 0.9508 |
| **g** | **and rank** | 0.9507 | <.001 | 0.9283 | 0.9736 |

1. The analyses performed in problems 3 to 5 are very similar in the predicted values that they yield. There is also no noticeable significant difference in the inferences that we’d obtain based on the different methods. Perhaps this is due to the fact that the sample is very large.



1. A priori, I would have chosen linear regression using robust standard errors because we lose the least amount of information. Modeling using dummy variables would cause loss of information since we have a continuous data, and using linear splines would not be as effective if we do not believe in different associations for each strata. Linear regression using robust standard errors would be effective because we have a large sample where we believe the relationship to be relatively linear. We do not have any reasons to believe in equal variances between groups, thus we can use robust standard errors based on Wald-based tests.