BIOST 518

Homework #8

Due Date: March 7, 2014

Question 1:

(a) Degree should be categorized as ordered with the lowest degree serving as the reference. There is a rank associated with degree such that a masters is higher than a bachelors; and a doctorate is higher than a masters.

Field should be categorized as unordered. There is no rank association with the field one has their degree or specialty in. However, there may be an association with salary, but it’s unlikely to determine which field is “better” or “higher” than another.

Admin should also be unordered since there is no rank associated with this variable.

(b) We cannot assume that this dataset is homoscedastic or the variances across groups are equal. As a result, it would be more appropriate to use robust standard error estimates over classical linear regression.

Classical linear regression would be anticonservative since the variances are presumed to be equal. Using unequal variances would result in a wider variance and result in wider confidence intervals.

(i) I would choose the model that would answer the a priori question. In this case, each model tells a different story. The linear model provide the most straight forward answer whether or not there is an association with sex and salary controlling for confounders. The quadratic model doesn’t provide much information unless you examine non-linearity, which is another step. Answers that are ratios of means provide an answer that is not easily interpretable by a decision maker. Therefore, it is also important to know who your audience is. Different in mean is easier to interpret and doesn’t require transformation or manipulation to get it into a context that can be understood.

Question 2:

(a)

Methods: Performed a linear regression model using monthly salary as a continuous response variable and year of degree as a continuous predictor variable adjusting for sex, field, degree, and administrative duties. 95% CI was determined using Huber-White sandwich estimator of the standard errors. We focused our analysis for the salary year 1995.

Inference:

The mean difference in salary was $89.97 lower in those faculty members who received their degree 1 year later compared to faculty members that received their degree earlier while having similar sex, degree, field, and administrative duties. This was highly significant with a two-tailed P<0.0001. From the 95% CI, the observed result would not be unusual if the true difference in mean monthly salary was anywhere between $98.30 and $81.42 lower in those faculty members who received their degree later.

(b)

Methods: Performed a linear regression model using monthly salary as a continuous response variable and start year as a continuous predictor variable adjusting for sex, field, degree, and administrative duties. 95% CI was determined using Huber-White sandwich estimator of the standard errors. We focused our analysis for the salary year 1995.

Inference:

The mean difference in monthly salary was $56.88 lower in faculty that started a year later compared to those who started earlier. The was highly significant at the two-tailed P<0.0001. From the 95% CI, the observed results were not unusual if the true different in mean monthly salary was anywhere from $47.63 and $66.13 lower in those faculty who started later.

(c)

Methods:

Methods: Performed a linear regression model using monthly salary as a continuous response variable and year of degree as a continuous predictor variable adjusting for sex, field, degree, administrative duties and start year. 95% CI was determined using Huber-White sandwich estimator of the standard errors. We focused our analysis for the salary year 1995.

Inference:

The mean difference in monthly salary was $111.96 lower in faculty who received their degree 1 year later compared to faculty who received their degree earlier. This was highly significant at the two-tailed P<0.0001. From the 95% CI, the observed resulted was not unusual if the true difference in mean salary was anywhere from $93.34 to $130.58 lower in faculty who received their degree later.

(d)

Methods: Performed a linear regression model using monthly salary as a continuous response variable and starting year as a continuous predictor variable adjusting for sex, field, degree, administrative duties and year of degree. 95% CI was determined using Huber-White sandwich estimator of the standard errors. We focused our analysis for the salary year 1995.

Inference:

The mean difference in month salary was $27.15 higher in faculty who started later compared to faculty who started earlier. This was highly significant with a two-tailed P=0.004. From the 95% CI, this observation was not unusual if the true difference in mean monthly salary was anywhere from $8.68 to $45.63 higher in faculty was started later.

(e)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  POI=yr of degree | Estimate | Z | P value | 95% CI low | 95% CI high |
| Linear (Model 1) | -89.86543 | -20.89 | <0.0001 | -98.3019 | -81.42895 |
| Linear 2 (Model 2) | -111.9608 | -11.79 | <0.0001 | -130.5796 | -93.34199 |
| POI = start year | Estimate | Z | P value | 95% CI low | 95% CI high |
| Linear (Model 3) | -56.88243 | -12.06 | <0.0001 | -66.13252 | -47.63234 |
| Linear 2 (Model 4) | 27.15361 | 2.88 | 0.004 | 8.680149 | 45.62707 |

When not adjusting for either the starting year or year of degree, there is an association for having a lower mean monthly salary if a faculty member started later or received their degree later. This relationship does not hold when both parameters are in the model. In Models 2 ad 4, the parameter estimates are the same for the variables start year and year of degree. Although Model 2 used year of degree as the POI adjusting for start year, the estimates are the same for Model 4 where start year was the POI and year of degree was adjusted. In Model 2, those who received their degree later had a lower monthly salary relative to those who started earlier. In Model 4, those who started later had a higher monthly salary. There appears to be some kind of interaction between the start year and mean monthly salary that should be further investigated.

Question 3:

For a-g, please see the table below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|   | Estimate | Z | P-value | 95% CI low | 95% CI high |
| Model 1 | -1334.731 | -14.04 | <0.0001 | -1521.177 | -1148.286 |
| Model 2 | -1266.152 | -13.4 | <0.0001 | -1451.555 | -1080.75 |
| Model 3 | -614.1284 | -7.17 | <0.0001 | -782.235 | -446.0218 |
| Model 4 | -614.5785 | -7.06 | <0.0001 | -785.3114 | -443.8456 |
| Model 5 | -420.0537 | -5.05 | <0.0001 | -583.1193 | -256.9881 |
| Model 6 | -419.7268 | -5.17 | <0.0001 | -578.9865 | -260.4672 |
| Model 7 | -280.6639 | -4.08 | <0.0001 | -415.5158 | -145.812 |

Question 4:

For a-g, please see table below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|   | Estimate | Z | P-value | 95% CI low | 95% CI high |
| Model 1 | 0.81201765 | -13.73 | <0.0001 | 0.78822019 | 0.83653359 |
| Model 2 | 0.82035484 | -13.09 | <0.0001 | 0.79637277 | 0.84505903 |
| Model 3 | 0.90901826 | -6.99 | <0.0001 | 0.88500693 | 0.93368104 |
| Model 4 | 0.90867517 | -6.98 | <0.0001 | 0.88453278 | 0.93347658 |
| Model 5 | 0.93624873 | -5.06 | <0.0001 | 0.91264737 | 0.96046043 |
| Model 6 | 0.9362918 | -5.17 | <0.0001 | 0.91320572 | 0.9599616 |
| Model 7 | 0.95742911 | -4.08 | <0.0001 | 0.93763109 | 0.97764516 |

Question 5:

Methods: Poisson regression was used to model the ratio of monthly mean salary for women and men in 1995.

Model 1 included salary as a continuous response variable and sex as a binary variable. Model 2, in addition to salary and sex, included degree as an ordered categorical variable. Model 3, in addition to salary, sex, and degree, included year of degree as linear splines. Model 4; in addition to salary, sex, degree, and year of degree; included start year at UW as linear splines. Model 5, in addition to the previous variables, included field as an unordered categorical variable. Model 6, in addition to the previous variables, included administrative duties as a binary categorical variable. Model 6, in addition to the previous variables, included rank as an ordered categorical variable.

Inference:

For a-g, please see table below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|   | Estimate | Z | P-value | 95% CI low | 95% CI high |
| Model 1 | 0.8017227 | -13.58 | <0.0001 | 0.7765495 | 0.8277119 |
| Model 2 | 0.8105014 | -12.98 | <0.0001 | 0.7851903 | 0.8366284 |
| Model 3 | 0.9008001 | -7.09 | <0.0001 | 0.8751422 | 0.9272103 |
| Model 4 | 0.9008235 | -7.01 | <0.0001 | 0.8748942 | 0.9275213 |
| Model 5 | 0.9286485 | -5.22 | <0.0001 | 0.9032124 | 0.9548009 |
| Model 6 | 0.9288995 | -5.33 | <0.0001 | 0.9041067 | 0.9543721 |
| Model 7 | 0.9512392 | -4.3 | <0.0001 | 0.9298091 | 0.9731633 |

Question 6:

Question 3 to 5 were similar in that the covariates used in each of the models were the same. The difference lay with the response variable. Question 3 focused on using a response variable that provided the difference in mean monthly salary. Question 4 focused on the ratio of geometric mean monthly salary; while Question 5 focused on the ratio of mean monthly salary (using Poisson regression).

Using the ratio of the geometric means and ratio of means provided very similar estimates for the comparison of monthly salary between females and males in 1995 for the different models. Although using the ratio of geometric means yielded slightly higher ratios than the Poisson regressions. Statistical significant was high for each model, even as more variables were introduced. Confidence intervals got narrower as more variables were introduced. Model 7 for both the ratio of geometric mean and Poisson models were lower compared to the other models.

For the difference in mean monthly salary as the response variable, similar trends in precision was observed as more covariates were introduced into the model.

In all cases, the Z-score decreased as each covariate were introduced. Significant drops in the Z-score occurred when we started to model for linear splines (e.g., start year and year of degree).

In Model 1, the difference in mean monthly salary was $1335 lower for women than men in 1995. This corresponds to 21% of the overall mean monthly salary ($6389), which is a ratio of approximately 0.80 and observed in the ratio of geometric mean and Poisson regression models. This correspondence occurred for all models as more covariates were introduced.

As more covariates an introduced the difference in mean monthly salary was reduced from $1335 lower in females than males for Model 1 to $281 lower in Model 7. The magnitude of difference was attenuated as more covariates were introduced. Similar results were seen when modeling the ratio of geometric means (ratio of 0.81 to 0.96); and ratio of means (0.80 to 0.95). In the cases where the ratios of monthly geometric and mean salary were modeled, the ratio was trending towards the null of 1.

Question 7:

Methods:

Linear regression was used to examine the association between mean monthly salary and gender in 1995. Potential confounders that were adjusted in the model included field, degree, rank, start year, year of degree, and administrative duties. Start year was modeled as linear splines with the following knots (less than 1960, 1960-65, 1965-70, 1970-75, 1975-80, 1880-85, 1985-90, and 1990 and greater). Year of degree was also modeled as linear spline using the same knots. 95% CI was determined using Huber-White sandwich estimator of the standard errors. Two-tailed P-value was calculated with a significance level set as 5%.

In order to covey the results, using a difference in mean monthly salary would make intuitive sense. However, selecting which model to use would be challenging. The unadjusted model does not take into consideration potential confounding variables. However, it is unclear whether or not all the variables should be included into the model. Regardless of an improvement in the precision of the CI, it should be necessary to determine whether or not effect modification is present. This should have been determined a priori. In our case, we decided to include all model terms in Model 7 as the representative inference we wished to convey.

Inference:

Mean difference in monthly salary for females was $280.66 lower compared to male while controlling for potential confounders in 1995. This was highly significant at the two-tailed P<0.0001. From the 95% CI, the observed results were not unusual if the true difference in mean monthly salary was $145.81 lower and $415.52 lower in females than males in 1995.