BIOST 515/518

Winter 2014

Homework 8

ID: 8658

1. (a) The best way to model the variables degree, field, and admin is to include them as dummy variables into a multiple regression model, as they are unordered categorical variables. Since admin is a binary variable we would model them as binary variables.

(b) Classical linear regression (without robust SE estimates) would be incorrect as it assumes equal variances of salaries between men and women, which may not be true. In 1995, there are 1188 observations for males and 409 observations for females. The standard deviation of salary is $2090 for males and $1481 for females. Since the variance is lower in the group with smaller sample size, I expect the inference for classical linear regression to be conservative.

(i) In a real situation, one would adjust for year of degree and starting year by modeling them as either continuous variables or linear splines, instead of dummy variables, because a lot of information would be loss and a step function would be forced. To decide between the options, one would think about the scientific question – perhaps the most appropriate way would be to model them using linear spline according to cost of living (i.e. inflation).

2. (a) From a linear regression analysis modeling year of degree as a linear continuous variable using standard errors calculated with the Huber-White sandwich estimator, we estimate that mean monthly salary differs by $89.87 between two individuals differing by one year of year of degree obtained, with the group with a later year of degree tending towards a lower average monthly salary. This result is significantly different from 0 (two-sided p<0.001), with a 95% confidence interval suggesting that such observed results would not be unusual if the true difference in mean monthly salary between year of degree were anywhere between $81.43 lower and $98.30 lower for each year difference in degree obtained, comparing a group of later year of degree to a group of earlier year of degree. We thus reject the null hypothesis that mean salary does not differ across year of degree, in favor of a hypothesis that mean monthly salary tends to be lower for later year of degrees.

(b) From a linear regression analysis modeling start year as a linear continuous variable using standard errors calculated with the Huber-White sandwich estimator, we estimate that mean monthly salary differs by $56.88 between two individuals differing by one year in starting year, with the group with a later year of degree tending towards lower average monthly salary. This result is significantly different from 0 (two-sided p<0.001), with a 95% confidence interval suggesting that such observed results would not be unusual if the true difference in mean monthly salary between year of degree were anywhere between $47.63 lower and $66.13 lower for each year difference in degree obtained, comparing a group of later start year to a group of earlier start year. We thus reject the null hypothesis that mean salary does not differ across starting year, in favor of a hypothesis that mean monthly salary tends to be lower for later starting years.

(c) From a linear regression analysis modeling year of degree as a linear continuous variable using standard errors calculated with the Huber-White sandwich estimator, we estimate that, after adjusting for start year modeled as a linear continuous variable, mean monthly salary differs by $111.96 between two individuals differing by one year of year of degree obtained, with the group with a later year of degree tending towards a lower average monthly salary. This result is significantly different from 0 (two-sided p<0.0001), with a 95% confidence interval suggesting that such observed results would not be unusual if the true difference in mean monthly salary between year of degree were anywhere between $93.34 lower and $130.58 lower for each year difference in degree obtained, comparing a group of later year of degree to a group of earlier year of degree. We thus reject the null hypothesis that mean salary does not differ across year of degree, in favor of a hypothesis that mean monthly salary tends to be lower for later year of degrees.

(d) From a linear regression analysis modeling start year as a linear continuous variable using standard errors calculated with the Huber-White sandwich estimator, we estimate that, after adjusting for year of degree modeled as a linear continuous variable, mean monthly salary differs by $27.15 between two individuals differing by one year in starting year, with the group with a later year of degree tending towards higher average monthly salary. This result is significantly different from 0 (two-sided p = 0.004), with a 95% confidence interval suggesting that such observed results would not be unusual if the true difference in mean monthly salary between year of degree were anywhere between $8.68 higher and $45.63 higher for each year difference in degree obtained, comparing a group of later start year to a group of earlier start year. We thus reject the null hypothesis that mean salary does not differ across starting year, in favor of a hypothesis that mean monthly salary tends to be higher for later starting years.

(e) The unadjusted analyses showed that year of degree and start year individually are associated with average monthly salary of faculty. After adjustment, the associations of year of degree and start year with average monthly salary become weaker (attenuates towards the null). Thus scientifically, these two variables could be confounders when we attempt to answer the scientific question of the association of the other variable and mean monthly salary. This is because start year and year of degree are positively correlated variables.

3. (a) – (g) will be presented in the following table with the models defined as:

Linear regression with robust standard errors are run to compute the difference in mean monthly salary (salary modeled as a linear continuous variable), comparing women to men (men as baseline).

A: Unadjusted comparison

B: Adjusted for degree modeled as a dummy variable

C: Adjusted for degree modeled as a dummy variable and year of degree modeled as linear splines

D: Adjusted for degree modeled as a dummy variable, year of degree modeled as linear splines, start year modeled as linear splines

E: Adjusted for degree modeled as a dummy variable, year of degree modeled as linear splines, start year modeled as linear splines, field modeled as a dummy variable

F: Adjusted for degree modeled as a dummy variable, year of degree modeled as linear splines, start year modeled as linear splines, field modeled as a dummy variable, administrative duties modeled as a binary variable. The predicted values are saved as fit3

G: Adjusted for degree modeled as a dummy variable, year of degree modeled as linear splines, start year modeled as linear splines, field modeled as a dummy variable, administrative duties modeled as a binary variable, rank modeled as a dummy variable

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Estimate | t-statistic | Two-sided p-value | 95% C.I. |
| A | - 1334.73 | -14.04 | < 0.001 | -1521.18, -1148.29 |
| B | -1266.15 | -13.40 | < 0.001 | -1451.56, -1080.75 |
| C | -614.13 | -7.17 | < 0.001 | -782.24, -446.02 |
| D | -614.58 | -7.06 | < 0.001 | -785.31, -443.85 |
| E | -420.05 | -5.05 | < 0.001 | -583.12, -256.99 |
| F | -419.73 | -5.17 | < 0.001 | -578.99, -260.47 |
| G | -280.66 | -4.08 | < 0.001 | -415.52, - 145.81 |

4. (a) – (g) will be presented in the following table with the models defined as:

Linear regression with robust standard errors are run to compute the ratios of geometric mean monthly salary (salary modeled as a log linear continuous variable), comparing women to men (men as baseline).

A: Unadjusted comparison

B: Adjusted for degree modeled as a dummy variable

C: Adjusted for degree modeled as a dummy variable and year of degree modeled as linear splines

D: Adjusted for degree modeled as a dummy variable, year of degree modeled as linear splines, start year modeled as linear splines

E: Adjusted for degree modeled as a dummy variable, year of degree modeled as linear splines, start year modeled as linear splines, field modeled as a dummy variable

F: Adjusted for degree modeled as a dummy variable, year of degree modeled as linear splines, start year modeled as linear splines, field modeled as a dummy variable, administrative duties modeled as a binary variable. The predicted values are saved as fit4

G: Adjusted for degree modeled as a dummy variable, year of degree modeled as linear splines, start year modeled as linear splines, field modeled as a dummy variable, administrative duties modeled as a binary variable, rank modeled as a dummy variable

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Point estimate | T-statistic | Two-sided p-valiue | 95% C.I. |
| A | 0.8120 | -13.73 | < 0.001 | 0.7882, 0.8365 |
| B | 0.8204 | -13.09 | < 0.001 | 0.7964, 0.8451 |
| C | 0.9090 | -6.32 | < 0.001 | 0.8850, 0.9337 |
| D | 0.9087 | -6.98 | < 0.001 | 0.8845, 0.9335 |
| E | 0.9362 | -5.06 | < 0.001 | 0.9126, 0.9605 |
| F | 0.9363 | -5.17 | < 0.001 | 0.9132, 0.9600 |
| G | 0.9574 | -4.08 | < 0.001 | 0.9376, 0.9776 |

5. (a) – (g) will be presented in the following table with the models defined as:

Poisson regression with robust standard errors (irr option) are run to compute the ratios of arithmetic mean monthly salary (salary modeled as a linear continuous variable), comparing women to men (men as baseline).

A: Unadjusted comparison

B: Adjusted for degree modeled as a dummy variable

C: Adjusted for degree modeled as a dummy variable and year of degree modeled as linear splines

D: Adjusted for degree modeled as a dummy variable, year of degree modeled as linear splines, start year modeled as linear splines

E: Adjusted for degree modeled as a dummy variable, year of degree modeled as linear splines, start year modeled as linear splines, field modeled as a dummy variable

F: Adjusted for degree modeled as a dummy variable, year of degree modeled as linear splines, start year modeled as linear splines, field modeled as a dummy variable, administrative duties modeled as a binary variable. The predicted values are saved as fit5

G: Adjusted for degree modeled as a dummy variable, year of degree modeled as linear splines, start year modeled as linear splines, field modeled as a dummy variable, administrative duties modeled as a binary variable, rank modeled as a dummy variable

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Point estimate | T-statistic | Two-sided p-value | 95% C.I. |
| A | 0.8017 | -13.58 | < 0.001 | 0.7765, 0.8277 |
| B | 0.8105 | -12.98 | < 0.001 | 0.7852, 0.8366 |
| C | 0.9008 | -7.09 | < 0.001 | 0.8751, 0.9272 |
| D | 0.9008 | -7.01 | < 0.001 | 0.8749, 0.9275 |
| E | 0.9286 | -5.22 | < 0.001 | 0.9032, 0.9548 |
| F | 0.9289 | -5.34 | < 0.001 | 0.9041, 0.9544 |
| G | 0.9512 | -4.30 | < 0.001 | 0.9298, 0.9732 |

6. The similarities between the analyses performed in problems 3-5 were that all of them included unadjusted analyses to estimate the difference of salaries between male and female faculty, in addition to adjusted analyses with variables modeled the exact same way. Problem 3 modeled the difference as a linear continuous variable (difference in arithmetic mean); problem 4 modeled the difference as a log-transformed continuous variable (ratio of geometric means); while problem 5 modeled the ratio of arithmetic means. To gauge how similar the predicted values are, I tabulated the values of fit3, fit4, and fit5 (see below), and concluded that the predicted values are rather similar. Besides that, the unadjusted ratios agree exactly, while the adjusted ones vary a little bit. The inference obtained is different in each problem – problem 4 views salary as a multiplicative model, while problems 3 and 5 view salary linearly, using differences and ratios respectively. In all cases, even after adjusting for all variables, we get statistically significant results that female faculty tend to receive a lower salary compared to otherwise similar male faculty.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Male | | | Female | | | Total | | |
| Model | N | Mean (S.D.) | Min - Max | N | Mean (S.D.) | Min - Max | N | Mean (S.D.) | Min - Max |
| fit3 (difference of means) | 1188 | 6732 (1235) | 3395 - 10693 | 409 | 5397 (1134) | 2836 - 9252 | 1597 | 63890 (1343) | 2836 – 10693 |
| fit4 (geometric means) | 1188 | 6543 (1227) | 3754 - 11284 | 409 | 5309 (989.1) | 3397 - 9169 | 1597 | 6227 (1289) | 3397 - 11284 |
| fit5 (ratio of arithmetic means) | 1188 | 6732 (1279) | 3771 - 11594 | 409 | 5397 (1021) | 3394 - 9390 | 1597 | 63890 (1350) | 3394 – 11594 |

7. A priori, the single model that would best reflect discrimination against women in awarding salary is to adjust for all variables except for rank, and model salary as a linear continuous variable, reporting the ratio of means (Problem 5 (f)). This is because it makes the most scientific sense to think about the arithmetic mean of salary. In addition, money is typically multiplicative, thus a multiplicative model with the ratio of means make sense. In addition, we should compare men and women who are “otherwise similar”, so we should adjust for all other variables. Nonetheless since rank could be a basis for discrimination (e.g. more women just never make past the assistant professor level) we do not want to adjust for that.

Methods: The difference in mean salary between male and female faculty in 1995, who are otherwise similar was compared using a Poisson regression model with robust standard errors. Salary was modeled as a linear continuous variable, while adjusted covariates include degree modeled as a dummy variable, year of degree modeled as linear splines, start year modeled as linear splines, field modeled as a dummy variable, and administrative duties modeled as a binary variable. Statistical inference was based on the Wald statistic computed from the regression slope parameter and its standard error as estimated using the Huber-White sandwich estimator, with two-sided p-values and 95% confidence intervals computed using the approximate normal distribution for linear regression parameter estimates.

Inference: Of the 1597 otherwise similar faculty members (1188 males and 409 females) whose monthly salary was reported in 1995, the average salary for male faculty is $25100.52, while the average salary for female faculty is $24680.79. This ratio of mean salary of 0.9289 (females have mean salaries 7.11% lower compared to males), with female tending to a lower salary, would not be deemed unusual if, based on a 95% confidence interval, the true ratio of mean salary was anywhere between 0.9041 lower and 0.9544 lower (females have mean salaries 4.56% to 9.59% lower), comparing female to male faculty. A two-sided p-value of <0.001 suggest that we can with high confidence reject the null hypothesis that salary is not associated with sex in favor of a hypothesis that female faculty tends towards a lower average monthly salary in 1995.