**1.**

(a)

I would model *degree, field,* and *admin* as dummy variables since they are unordered categorical variables.

(b)

Robust standard errors do not assume equal variance across observations, an assumption which we would like to avoid making. Robust standard errors also account for within-cluster correlation. Classical linear regression could tend to be either conservative or anti-conservative, depending on the direction of the correlation in the data.

(i) In a real situation, I would make my choice of model before seeing the data. For this *a priori* choice, I would probably choose to adjust for year of degree and starting year with linear splines with knots that reflect my belief in where there might be discontinuities in an underlying time trend of salary at UW.

**2.**

In each case, we use a dummy variable for the degree and field variables.

(a)

Using linear regression with robust standard errors, we estimate an $89.87 difference in monthly salary between groups of observations similar in highest degree attained, sex, administrative duties, and field of study whose year of highest degree attainment differs by 1 year, with the group with later degree attainment having the lower estimated salary. The 95% CI suggests that this estimate is consistent with the group with more recent degree attainment having a monthly salary that is between $98.30 lower and $81.43 lower. The associated p-value is less than 10^-3 < 0.05, so we conclude that there is a statistically significant association between year of degree attainment and monthly salary, while accounting for highest degree attained, sex, administrative duties, and field of study.

(b)

Using linear regression with robust standard errors, we estimate a $56.88 difference in monthly salary between groups of observations similar in highest degree attained, sex, administrative duties, and field of study whose starting years differ by 1 year, with the group with later degree attainment having the higher estimated salary. The 95% CI suggests that this estimate is consistent with the group with more recent degree attainment having a monthly salary that is between $66.13 higher and $47.63 higher. The associated p-value is less than 10^-3 < 0.05, so we conclude that there is a statistically significant association between starting year and monthly salary, while accounting for highest degree attained, sex, administrative duties, and field of study.

(c)

Using linear regression with robust standard errors, we estimate a $111.96 difference in monthly salary between groups of observations similar in starting year, highest degree attained, sex, administrative duties, and field of study whose year of highest degree attainments differs by 1 year, with the group with later degree attainment having the lower estimated salary. The 95% CI suggests that this estimate is consistent with the group with more recent degree attainment having a monthly salary that is between $130.58 lower and $93.34 lower. The associated p-value is less than 10^-3 < 0.05, so we conclude that there is a statistically significant association between year of degree attainment and monthly salary, while accounting for starting year, highest degree attained, sex, administrative duties, and field of study.

(d)

Using linear regression with robust standard errors, we estimate a $27.15 difference in monthly salary between groups of observations similar in year of degree attainment, highest degree attained, sex, administrative duties, and field of study whose starting years differ by 1 year, with the group with later degree attainment having the higher estimated salary. The 95% CI suggests that this estimate is consistent with the group with more recent degree attainment having a monthly salary that is between $8.68 higher and $45.63 higher. The associated p-value is less than 0.004 < 0.05, so we conclude that there is a statistically significant association between starting year and monthly salary, while accounting for year of degree attainment, highest degree attained, sex, administrative duties, and field of study.

(e)

The model in part (d) estimates the association between salary and starting year, while accounting for level of degree attained, year of degree attainment, sex, administrative duties, and field of study. Models (a) – (c) are hierarchically nested in model (d), so they also estimate an association between salary and a potential explanatory variable, but hold fewer other covariates fixed than the model that comes next.

**3.**

In Table 1, we present a hierarchy of adjusted linear regression models (using robust standard errors) for monthly salary in the table below. The model fit in each row is nested in the one below it (e.g. the “year of degree” row includes an adjustment for year of degree as well as “degree,” which was in the previous row). The model fit in each row is nested in the one below it (e.g. the “year of degree” row includes an adjustment for year of degree as well as “degree,” which was in the previous row). In each case, we use a dummy variable for the degree, field, and rank variables as well as linear splines for year of degree and starting year as defined in problem 1(f).

**Table 1**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Adjustment** | **Estimate ($)** | **t** | **p-value** | **95% CI low** | **95% CI high** |
| None | -1335 | -14.04 | < .001 | -1521 | -1148 |
| Degree | -1266 | -13.40 | < .001 | -1452 | -1081 |
| Year of Degree (spline) | -614.1 | -7.17 | < .001 | -782.2 | -446.0 |
| Starting Year (spline) | -614.6 | -7.06 | < .001 | -785.3 | -443.8 |
| Field | -420.1 | -5.05 | < .001 | -583.1 | -257.0 |
| Administrative Duties | -419.7 | -5.17 | < .001 | -580.0 | -260.5 |
| Rank | -280.7 | -4.08 | < .001 | -415.5 | -145.8 |

**4.**

In Table 2, we present a hierarchy of adjusted linear regression models (using robust standard errors) for log-transformed monthly salary in the table below. The model fit in each row is nested in the one below it (e.g. the “year of degree” row includes an adjustment for year of degree as well as “degree,” which was in the previous row). In each case, we use a dummy variable for the degree, field, and rank variables as well as linear splines for year of degree and starting year as defined in problem 1(f).

**Table 2**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Adjustment** | **Estimate** | **t** | **p-value** | **95% CI low** | **95% CI high** |
| None | -0.2082 | -13.73 | < .001 | -0.2380 | -0.1785 |
| Degree | -0.1980 | -13.09 | < .001 | -0.2277 | -0.1683 |
| Year of Degree (spline) | -0.0954 | -6.99 | < .001 | -0.1222 | -0.0687 |
| Starting Year (spline) | -0.0958 | -6.98 | < .001 | -0.1227 | -0.6884 |
| Field | -0.0659 | -5.06 | < .001 | -0.0914 | -0.0403 |
| Administrative Duties | -0.0658 | -5.17 | < .001 | -0.0908 | -0.0409 |
| Rank | -0.0435 | -4.08 | < .001 | -0.0644 | -0.0226 |

Here, exponentiating each estimate would give the estimated ratio of geometric mean monthly salaries scores of females relative to males, holding fixed each covariate in the hierarchical model.

**5.**

We use a generalized linear model with log link and report exponentiated parameter estimates in the table below. The model fit in each row is nested in the one below it (e.g. the “year of degree” row includes an adjustment for year of degree as well as “degree,” which was in the previous row). In each case, we use a dummy variable for the degree, field, and rank variables as well as linear splines for year of degree and starting year as defined in problem 1(f).

**Table 3**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Adjustment** | **Estimate** | **Z** | **p-value** | **95% CI low** | **95% CI high** |
| None | 0.8017 | -13.58 | < .001 | 0.7765 | 0.8277 |
| Degree | 0.8097 | -12.99 | < .001 | 0.7844 | 0.8359 |
| Year of Degree (spline) | 0.8981 | -7.12 | < .001 | 0.8719 | 0.9251 |
| Starting Year (spline) | 0.8964 | -7.04 | < .001 | 0.8695 | 0.9241 |
| Field | 0.9251 | -5.26 | < .001 | 0.8986 | 0.9523 |
| Administrative Duties | 0.9245 | -5.49 | < .001 | 0.8989 | 0.9501 |
| Rank | 0.9507 | -4.15 | < .001 | 0.9283 | 0.9736 |

**6.**

As shown in Figure 1 (where predicted salary is on the y-axis), the predicted salary values from the linear fit in problem 2(f) are noticeably higher for males than for females for all levels of actual salary. The predicted salary values from the geometric means and GLM models fit in problems 3(f) and 4(f) are also considerably higher for females than males. The difference between predicted salary values for males and females appears roughly the same for all of the models. This suggests that the discrepancy in salaries between the sexes cannot be fully explained by the degree type, year of degree, starting year, field, and administrative duties as they are specified in any of our regression models.

**Figure 1**

****

**7.**

**Methods:** *A priori,* I would have chosen to use linear regression over a single year. I will choose 1995 to be the fixed year and use the model problem 3. My outcome would be monthly salary at particular point in time and not its general time trend. My predictor interest would be a dummy variable indicating sex. I would have expected level of degree, year of degree, starting year, administrative duties, and rank to be effect modifiers. I would have expected field to be associated with both sex and salary, so I would treat it as a confounder. As a result, I would estimate the following linear regression model for reported monthly salaries in 1995:

$$Salary\_{i}= β\_{0}+ β\_{1}Female\_{i}+β\_{2}Degree\_{i}+β\_{3}YearDegree\_{i}+β\_{4}StartingYear\_{i}+β\_{5}Administrative\_{i}+β\_{6}Rank\_{i}$$

where, as in problem 3, female, degree, administrative duties, and rank are all treated as dummy variables, while year of degree and starting year are treated as linear splines with knots at 1960, 1965, 1970, 1975, 1980, 1985, and 1990. I will use robust standard errors that do not assume the presence of homoscedasticity. My coefficient of interest will be $β\_{1}$ and I will test for an association between sex and monthly salary, accounting for the included covariates, as my indicator of sex discrimination. We are particularly interested in whether $β\_{1} $is less than 0, which would indicate higher estimated salaries for males in the presence of the covariates. My null hypothesis is that $β\_{1}=0$ and the alternative is that $β\_{1}<0, $so I will report a one-sided p-value and Wald-based 95% CI associated with this coefficient estimate.

**Results:**

As shown in Table 1, we estimate that, for a group of male professors and a group of female professors at UW in 1995 who are similar in level of degree, year of degree, starting year, field, administrative duties, and rank, the female professors earn a monthly salary that is $280.70 lower than the males. The 95% CI associated with this estimate suggests that this estimate is consistent with females actually earning between $145.80 less and $415.50 less than a group of males who is similar with respect to this group of covariates. The corresponding one-sided p-value is < .001 < 0.05, so we have evidence that monthly salary for females is significantly lower than monthly salary for males in 1995, controlling for level of degree, year of degree, starting year, field, administrative duties, and rank. This may be interpreted as discrimination against women in awarding salaries at UW in 1995.