Stat 539 Homework 2

Kenny Flagg

February 2, 2017

1. *The overarching goal of this project is to explore whether there is any evidence suggestive of discrimination by sex in the employment of the faculty at a single university (University of Washington). To this end, salary data was obtained on all faculty members employed by the University during the 1995 academic year. You have been asked to provide an analysis of 1995 salaries with the primary goal of determining whether or not gender discrimination exists with respect to pay. Along with the 1995 salary the following additional variables were also collected:*

   ID = The anonymous identification number for the faculty member
   GENDER = Gender of the faculty member (coded as M or F)
   DEG = The highest degree obtained by the faculty member (PhD, Professional, Other)
   FIELD = Field of research during 1995 (Arts, Professional, Other)
   STARTYR = Year starting employment at the university
   YEAR = Year of data collection (1995 for all)
   RANK = Faculty rank as of 1995 (Assistant, Associate, Full)
   ADMIN = Does faculty member hold an administrative position as of 1995? (0=No, 1=Yes)
   SALARY = 1995 salary in US dollars

   Realizing that the strongest generalization of analysis results comes when the statistical question is decided before looking at the data, in this exercise we will simply think about the goal of the analysis and ways in which we should statistically approach the problem.

   (a) *First, consider the sampling scheme for the current project. What population will we be able to make inference about? Can you think of any way in which the sampling scheme could lead to misleading inference regarding gender discrimination at the university?*

   Data were collected on all faculty employed at one university during one year, so we an make inference only to the population of people who were University of Washington faculty during the 1995 academic year. Because we only have one year to work with, we do not observe people who left the university because they were discriminated against. Therefore, we may underestimate the level of discrimination at this university.

(b) *Recall that a confounder in the relationship between gender and salary must be causally related to salary and associated with gender. Thus we can begin looking for potential confounders by first considering those factors that may influence salary. List out those factors which you feel strongest influence salary in this setting and justify your choices (Note: You should not limit yourself to factors for which data has been collected).*

- $RANK$ — The largest pay differences would be due to differences in position within a department.
- Year of degree — The amount of time a person has been working since they graduated would be related to their salary (probably more so than the amount of time at one institution).
- Tenure status — This would be associated with year of degree and starting year, but people do not always get tenure at the same point in their careers so it would be an interesting variable on its own.
- DEG — PhDs earn a bit more than others in academia.
- $STARTYR$ — More time at the current institution should come with raises, but I would not expect this variable to have a strong association with salary after accounting for time since degree and tenure status.
- $FIELD$ — There would be some salary differences based on the productivity of the department and demand for its graduates in the labor force.

(c) *Among those factors listed above, decide which might reasonably be associated with gender and justify your choices. These are the potential confounders you would ideally like to adjust for in your analysis.*

DEG would be associated with gender because many disciplines have histories of gender imbalance. If gender discrimination was present, $RANK$ and tenure status would be associated with gender, as would $STARTYR$ because people who are harmed by discrimination might not stay at the university as long as those who benefit from it.

(d) *List any factors that you a priori feel would be effect modifiers in the relationship between gender and salary. Justify your choice(s).*

$FIELD$ would be an effect modifier because different fields have different attitudes about diversity and gender equity. $RANK$ and DEG could be effect modifiers because the positions and degrees with more prestige (and higher base salaries) would have more room for gender differences.

(e) *List any factors that you a priori feel would be precision variables you would ideally like to adjust for in your analysis. Justify your choice(s).*

Years since degree would be a precision variable because would be a good predictor of salary regardless of whatever other associations may be present. If $RANK$, DEG, and tenure status are not confounders or effect modifiers then they would be precision variables because they are associated with scheduled pay increases.

(f) *Using your answers to (b) through (e), describe what available adjustment variables you would include in your regression analysis to answer the question of interest. Classify your adjustment variables as potential confounders, effect modifiers, or precision variables.*

I do not think any of the variables I discussed in my previous answers are nuisance variables, so I would include all the available ones as follows:

- $RANK$ — potential confounder
- DEG — potential confounder
- $STARTYR$ — potential confounder
- $FIELD$ – effect modifier

2. *The gamma distribution has probability density function*

$$f(y|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1}\exp(-\beta y),$$

for $y > 0$ where $\Gamma(\alpha)$ is the Gamma function. Show that the gamma distribution is a member of the exponential dispersion family by putting its pdf into the form given in equation (4.1) on p. 121 of our textbook. Identify the natural parameter, the dispersion parameter, and the functions $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$.

The pdf can be written as

$$\begin{aligned}
f(y|\alpha,\beta) &= \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1}\exp(-\beta y) \\
&= \exp\left(\alpha\log(\beta) - \log(\Gamma(\alpha)) + (\alpha-1)\log(y) - \beta y\right) \\
&= \exp\left(\alpha\log\left(\alpha\frac{\beta}{\alpha}\right) - \log(\Gamma(\alpha)) + (\alpha-1)\log(y) - \alpha\frac{\beta}{\alpha}y\right) \\
&= \exp\left(\frac{y\frac{\beta}{\alpha} - \log\left(\frac{\beta}{\alpha}\right)}{-\frac{1}{\alpha}} + \alpha\log(\alpha) - \log(\Gamma(\alpha)) + (\alpha-1)\log(y)\right) \\
&= \exp\left(\frac{y\theta - \log(\theta)}{-\frac{1}{\phi}} + \phi\log(\phi) - \log(\Gamma(\phi)) + (\phi-1)\log(y)\right)
\end{aligned}$$

where

- $\theta = \dfrac{\beta}{\alpha}$ is the natural parameter,
- $\phi = \alpha$ is the dispersion parameter,
- $a(\phi) = -\dfrac{1}{\phi}$,
- $b(\theta) = \log(\theta)$, and
- $c(y,\phi) = \phi\log(\phi) - \log(\Gamma(\phi)) + (\phi-1)\log(y)$

so the gamma distribution is a member of the exponential dispersion family.

3. *Suppose we are analyzing data regarding the effect of smoking and/or sex on systolic blood pressure (SBP). Suppose that the variance of SBP is σ^2 within each category defined by sex and smoking habits, and further suppose that the true average SBP within each group is as follows:*

	Males	Females
Nonsmokers	125	120
Smokers	133	128

Suppose that the percentage of smokers among both males and females is 30%. Let $MALE$ be an indicator variable that is 1 for males and 0 for females and $SMOKE$ be an indicator variable that is 1 for smokers and 0 for nonsmokers.

(a) *Is $MALE$ an effect modifier for the relationship between $SMOKE$ and SBP? Justify your answer.*

No, $MALE$ is not an effect modifier because, for both sexes, the true mean SBP for smokers is 8 units higher than the true mean SBP for nonsmokers.

(b) *Is $MALE$ a confounding variable for the relationship between $SMOKE$ and SBP? Justify your answer.*

No, $MALE$ is not a confounder because it is not related to both SBP and smoking status. It is related to the response SBP because the true mean SBP varies by sex within each smoking status, but it is not related to smoking status because the proportion who smoke is the same for both sexes. ($MALE$ is a precision variable.)

(c) *Suppose we fit the model*

$$E(SBP) = \beta_0 + \beta_1 \times SMOKE + \beta_2 \times MALE + \beta_3 \times SMOKE \times MALE$$

What are the interpretations of the linear regression estimates $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$? What are the expected values of the linear regression estimates $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$?

- We estimate that the mean SBP for female nonsmokers is $\hat{\beta}_0$.

$$E\left(\hat{\beta}_0\right) = \beta_0 = 120$$

- We estimate that the mean SBP for female smokers is $\hat{\beta}_1$ higher than the mean SBP for female nonsmokers.

$$E\left(\hat{\beta}_1\right) = \beta_1 = 8$$

- We estimate that the mean SBP for male nonsmokers is $\hat{\beta}_2$ higher than the mean SBP for female nonsmokers.

$$E\left(\hat{\beta}_2\right) = \beta_2 = 5$$

- We estimate that the difference in mean SBP between male smokers and male nonsmokers is $\hat{\beta}_3$ higher than the difference in mean SBP between female smokers and female nonsmokers.

$$E\left(\hat{\beta}_3\right) = \beta_3 = 0$$

(d) *Suppose we fit the model*

$$E(SBP) = \beta_0 + \beta_1 \times SMOKE + \beta_2 \times MALE$$

What are the interpretations of the linear regression estimates $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$? What are the expected values of the linear regression estimates $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$?

- We estimate that the mean SBP for female nonsmokers is $\hat{\beta}_0$.

$$E\left(\hat{\beta}_0\right) = \beta_0 = 120$$

- We estimate that the mean SBP for smokers of a given sex is $\hat{\beta}_1$ higher than the mean SBP for nonsmokers of the same sex.

$$E\left(\hat{\beta}_1\right) = \beta_1 = 8$$

- We estimate that the mean SBP for males of a given smoking status is $\hat{\beta}_2$ higher than the mean SBP for females of the same smoking status.

$$E\left(\hat{\beta}_2\right) = \beta_2 = 5$$

(e) *Suppose we fit the model*

$$E(SBP) = \beta_0 + \beta_1 \times SMOKE$$

What are the interpretations of the linear regression estimates $\hat{\beta}_0$ and $\hat{\beta}_1$? What are the expected values of the linear regression estimates $\hat{\beta}_0$ and $\hat{\beta}_1$?

- We estimate that the mean SBP for nonsmokers is $\hat{\beta}_0$.

$$E\left(\hat{\beta}_0\right) = \beta_0 = 120$$

- We estimate that the mean SBP for smokers is $\hat{\beta}_1$ higher than the mean SBP for nonsmokers.

$$E\left(\hat{\beta}_1\right) = \beta_1 = 8$$

4. *Agresti Exercise 2.42 (p. 77) In some applications, such as regressing annual income on the number of years of education, the variance of y tends to be larger at higher values of x. Consider the model* $E(y_i) = \beta x_i$*, assuming* $var(y_i) = x_i\sigma^2$ *for unknown* σ^2*.*

   (a) *Show that the generalized least squares estimator minimizes* $\sum_i (y_i - \beta x_i)^2 / x_i$ *(i.e., giving more weight to observations with smaller* x_i*) and has* $\hat{\beta}_{GLS} = \bar{y}/\bar{x}$*, with* $var\left(\hat{\beta}_{GLS}\right) = \sigma^2 / \left(\sum_i x_i\right)$*.*

   The model matrix is $\mathbf{X} = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^T$. Assuming the y_i are independent, we have

$$var(\mathbf{y}) = \sigma^2 \mathbf{V} = \sigma^2 \begin{bmatrix} x_1 & & 0 \\ & \ddots & \\ 0 & & x_n \end{bmatrix}$$

   so the GLS estimator needs to minimize

$$\begin{aligned} (\mathbf{y} - \mathbf{X}\beta)^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\beta) &= \begin{bmatrix} y_1 - \beta x_1 & \cdots & y_n - \beta x_n \end{bmatrix} \begin{bmatrix} \frac{1}{x_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{x_n} \end{bmatrix} \begin{bmatrix} y_1 - \beta x_1 \\ \vdots \\ y_n - \beta x_n \end{bmatrix} \\ &= \begin{bmatrix} \frac{y_1 - \beta x_1}{x_1} & \cdots & \frac{y_n - \beta x_n}{x_n} \end{bmatrix} \begin{bmatrix} y_1 - \beta x_1 \\ \vdots \\ y_n - \beta x_n \end{bmatrix} \\ &= \sum_{i=1}^{n} \frac{(y - \beta x_i)^2}{x_i}. \end{aligned}$$

The GLS estimator is

$$\begin{aligned}
\hat{\beta}_{GLS} &= \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} \\
&= \left(\begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} \frac{1}{x_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{x_n} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}\right)^{-1} \mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} \\
&= \left(\begin{bmatrix} 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}\right)^{-1} \mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} \\
&= \frac{1}{\sum_{i=1}^n x_i}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} \\
&= \frac{1}{\sum_{i=1}^n x_i} \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} \frac{1}{x_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{x_n} \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \\
&= \frac{1}{\sum_{i=1}^n x_i} \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \\
&= \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i} \\
&= \frac{\bar{y}}{\bar{x}}
\end{aligned}$$

with variance

$$\begin{aligned}
var\left(\hat{\beta}_{GLS}\right) &= var\left(\frac{\bar{y}}{\bar{x}}\right) \\
&= var\left(\frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i}\right) \\
&= \frac{\sum_{i=1}^n var(y_i)}{\left(\sum_{i=1}^n x_i\right)^2} \\
&= \frac{\sum_{i=1}^n x_i\sigma^2}{\left(\sum_{i=1}^n x_i\right)^2} \\
&= \frac{\sigma^2}{\sum_{i=1}^n x_i}.
\end{aligned}$$

(b) *Show that the ordinary least squares estimator is* $\hat{\beta} = \left(\sum_i x_i y_i\right) / \left(\sum_i x_i^2\right)$ *and has* $var\left(\hat{\beta}\right) = \sigma^2 \left(\sum_i x_i^3 / \left(\sum_i x_i^2\right)^2\right)$.

The OLS estimator is

$$\begin{aligned}
\hat{\beta} &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \\
&= \left(\begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}\right)^{-1}\mathbf{X}^T\mathbf{y} \\
&= \frac{1}{\sum_{x_2^2}}\mathbf{X}^T\mathbf{y} \\
&= \frac{1}{\sum_{x_2^2}}\begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix}\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \\
&= \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}
\end{aligned}$$

with variance

$$\begin{aligned}
var\left(\hat{\beta}\right) &= var\left(\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}\right) \\
&= \frac{\sum_{i=1}^n x_i^2 var(y_i)}{\left(\sum_{i=1}^n x_i^2\right)^2} \\
&= \frac{\sum_{i=1}^n x_i^2 x_i \sigma^2}{\left(\sum_{i=1}^n x_i^2\right)^2} \\
&= \sigma^2 \frac{\sum_{i=1}^n x_i^3}{\left(\sum_{i=1}^n x_i^2\right)^2}.
\end{aligned}$$

(c) *Show that* $var\left(\hat{\beta}\right) \geq var\left(\hat{\beta}_{GLS}\right)$.

Note that

$$\begin{aligned}
var\left(\hat{\beta}\right) &= \frac{\sum_{i=1}^n x_i^3 \sum_{i=1}^n x_i}{\left(\sum_{i=1}^n x_i^2\right)^2} var\left(\hat{\beta}_{GLS}\right) \\
&= \frac{\sum_{i=1}^n \sum_{j=1}^n x_i^3 x_j}{\sum_{i=1}^n \sum_{j=1}^n x_i^2 x_j^2}\left(\hat{\beta}_{GLS}\right).
\end{aligned}$$

Since $var(y_i) = x_i\sigma^2$ implies $x_i > 0$ for all i,

$$\frac{\sum_{i=1}^n \sum_{j=1}^n x_i^3 x_j}{\sum_{i=1}^n \sum_{j=1}^n x_i^2 x_j^2} \geq 1$$

so $var\left(\hat{\beta}\right) \geq var\left(\hat{\beta}_{GLS}\right)$.