Stats 539 Homework 2: Due Thursday, Feb. 2 by 10:50am

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 identication 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?

   (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).

   (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.

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

(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).

(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.

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 $c(\cdot)$.

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.

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

(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 $\hat{\beta}_3$?

(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 $\hat{\beta}_2$?

(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 $\hat{\beta}_1$?

4. Agresti Exercise 2.42 (p. 77)