**Biostat/Epi 536 2024**

**HW 2 (3 multi-part questions)**

1. A pharmaceutical company has run a randomized controlled trial of a drug intended for patients with symptomatic infection to prevent progression to severe disease. Assume that males with symptomatic infection are substantially more likely to progress to severe disease than females. In the trial, treatment vs. placebo was randomly assigned, with males and females each evenly divided between the treatment vs. placebo groups. You can also assume the treatment is effective – it does reduce the chances of severe disease.

a. Suppose the company chooses to summarize results with OR, and that the OR in males is the same as the OR in females. If the marketing department wants to make their drug look as impressive as possible, would they prefer the crude OR or the sex-adjusted OR? Why?

b. Which measure would you prefer for summarizing the effect of the treatment ? RR, RD, or OR? Explain briefly.

2. A major public research university has been accused of discriminating against women in admission to its graduate programs. A task force randomly selects 6 graduate programs (“majors”) from across the university to investigate the question. Use the dataset “sexbias” to investigate the following.

a. For the three variables in the dataset, draw a DAG representing the most appropriate scientific model to approach the question.

b. Use logistic regression to examine the unadjusted association between sex and acceptance to graduate school. Summarize the results in language suitable for the task force’s report.

c. Use logistic regression to examine the association between sex and acceptance to graduate school adjusted for “major”. Summarize the results in language suitable for the task force’s report.

d. Are the results from b and c very different? Why or why not? (Don’t answer in general terms, answer in terms of this dataset.)

e. Which analysis best addresses the question of whether the University discriminates against women in graduate school admissions?

f. Is there any other information you would have liked to have had for this analysis? E.g. any unmeasured potential confounders?

3. The course CANVAS site has a file of (fictitious) data from a case-control study of lung-cancer examining two exposures, smoking and asbestos. Asbestos is the exposure of interest. Fit a logistic regression model with a main effect for asbestos exposure, a main effect for smoking, and an interaction term for asbestos exposure and smoking. (This is “model A” in lecture 4.)

a. For each of the four regression parameters in the model: what population quantity does the parameter estimate? If the parameter does not estimate a population quantity, briefly explain why.

b. According to the fitted model, what is the OR for asbestos among non-smokers?

c. According to the fitted model, what is the OR for asbestos among smokers?

d. Summarize the evidence that smokers and non-smokers have different ORs for asbestos. Write your answer in a few sentences suitable for a scientific publication.

e. One could instead estimate the OR for asbestos among smokers by fitting a simple logistic regression model using the subset of the data on smokers. Do this. Compare your point estimates and confidence intervals here and part c and comment on whether any similarities or differences are to be expected.

f. Use an appropriate logistic regression model to estimate the smoking-adjusted OR for asbestos. Compare your results with b and c above.

g. For the model in part f, perform a test of the null hypothesis that the smoking-adjusted odds ratio is 1.