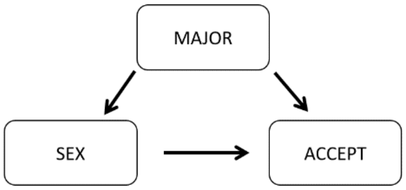
BIOST 536 HW2

John Schoof
10/17/2020

Question 1

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.

1a. For the three variables in the dataset, draw a DAG representing the most appropriate scientific model to approach the question.



The above DAG represents the scientific model proposed by question 1. Sex is the exposure variable, which is connected to Accept, the outcome variable. We are investigating this causal arc. Major is a third variable which is associated with both Sex and Accept as shown by the arcs above.

1b. 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.

$$\text{Logit}(Accept = 1|male) = \beta_0 + \beta_1 * male$$

I ran the above unadjusted logistic regression of the log odds of admission to a graduate program based on a sample of 4,526 individuals using robust standard errors. I estimate that, on average, the odds of a male being accepted are 1.841 (95% CI: 1.624, 2.087) times greater than the odds of a female being accepted. This finding is statistically significant at the alpha = 0.05 confidence level.

1c. 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.

$$\text{Logit}(Accept = 1|male, major) = \beta_0 + \beta_1 * male + \beta_2 * major$$

I ran the above logistic regression of the log odds of admission to a graduate program based on a sample of 4,526 individuals using robust standard errors. I estimate that, on average, the odds of a male being accepted are 0.905 **(95% CI:** 0.773, 1.059) times the odds of a female being accepted for individuals of the same major. This finding is not statistically significant at the alpha = 0.05 confidence level.

1d. 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.)

My findings from 1b and 1c. The findings from 1b suggest that males are much more likely than females to be accepted to graduate school. However, after we control for the major in 1c, we find no significant difference in the odds of admission for males and females. In fact, the findings from 1c suggest that odds of admission for males may actually be lower than that for females when accounting for major.

1e. Which analysis best addresses the question of whether the University discriminates against women in graduate school admissions?

The analysis in 1c, adjusting for major, best addresses this question. It is necessary to adjust for major because males and females apply to different programs in different amounts. It is possible that overall males appear to be accepted more often than females, but they may just apply to less competitive majors more often. It is necessary to account for this.

1f. Is there any other information you would have liked to have had for this analysis? E.g. any unmeasured potential confounders?

Other variables that could be potential confounders might include GRE scores, age, work experience, IQ, earnings in most recent year, whether they have children, home country.

Question 2

The course CANVAS site has a file of (fictitious) data from a case-control study of lung-cancer examining two exposures, smoking and asbestos. Fit the saturated logistic regression model discussed in class ("Model A") to these data. The model should have a "main effect" for asbestos exposure, and "main effect" for smoking, and an interaction term for asbestos exposure and smoking. Asbestos is the exposure of interest.

$$\text{Logit}(LungCancer = 1|asbestos, smoke) = \beta_0 + \beta_1 * asbestos + \beta_2 * smoke + \beta_3 * asbestos * smoke$$

2a. 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.

- β_0 estimates the log odds of developing lung cancer among individuals who do not smoke and were not exposed to asbestos.
- β_1 estimates the additional log odds of developing lung cancer among individuals who smoke.
- β_2 estimates the additional log odds of developing lung cancer among individuals who were exposed to asbestos.
- β_3 estimates the additional log odds of developing lung cancer among individuals who were exposed to asbestos and smoke.

2b. According to the fitted model, what is the OR for asbestos among non-smokers?

$e^{\beta_1} = e^{0.69} = 2.00$

The OR for asbestos among non-smokers is 2.00 (95% CI: 0.68, 6.58). This means that, among non-smokers, the odds of developing lung cancer for those exposed to asbestos are 2 times greater than those who were unexposed to asbestos on average. However, this point estimate is not statistically significant at the $\alpha = 0.05$ significance level.

2c. According to the fitted model, what is the OR for asbestos among smokers?

$e^{\beta_1+\beta_3} = e^{0.69+3.40} = 60$
The OR for asbestos among smokers is 60 (95% CI: 21.4, 168.0). This means that, among smokers, the odds of developing lung cancer for those exposed to asbestos are 60 times greater than those who were unexposed to asbestos on average.

2d. 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.

I ran the above logistic regression of the log odds of developing lung cancer based on a sample of 285 individuals using robust standard errors. The large difference between OR estimates for smokers and non-smokers is due to the effect modification of smoking on asbestos. The interaction term between smoking and asbestos is statistically significant at the $\alpha = 0.05$ significance level.

2e. 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.

I ran a simple logistic regression model using the subset of the data that smoked. My analysis found the exact same point estimate and confidence interval for the OR as in part c, 60 (95% CI: 21.4, 168.0). This is to be expected because this is exactly what happens when we run regression analysis and adjust for a confounding variable. We use the conditional method and only compare those that have the same value of the confounding variable.

2f. Use an appropriate logistic regression model to estimate the smoking-adjusted OR for asbestos. Compare your results with b and c above.

$$\text{Logit}(LungCancer = 1|asbestos, smoke) = \beta_0 + \beta_1 * asbestos + \beta_2 * smoke$$

I ran the above logistic regression of the log odds of developing lung cancer based on a sample of 285 individuals using robust standard errors. I estimate that, among those with the same smoking status, the odds of developing lung cancer if one is exposed to asbestos are 17.859 (95% CI: 9.295, 34.315) times higher than those who are not exposed to asbestos.

2g. For the model in part f, perform a Wald test and likelihood ratio test of the null hypothesis that the smoking-adjusted odds ratio is 1.

To test the null hypothesis that smoking-adjusted odds ratio is 1, we perform a chi-squared test of lung cancer on asbestos among non-smokers. This returns a p-value less than 0.001 which provides evidence against the null hypothesis.