Lecture 2, Part IV: Simpson's Paradox and the Method of Adjusting for Confounders

Text references: Chapter 16.4 from *All of Statistics* and handouts

Simpson's Paradox

Simpson's paradox refers to the phenomenon that any statistical relationship between two variables can be reversed by including additional factors in the analysis. Here we use counterfactuals and the method of adjusting for confounding to make sense out of data sets that exhibit the paradox.

Let us start with the following example from "Probability and Statistics", 4th ed., by DeGroot and Schervish:

Example 1. Let X be a binary treatment variable and Y a binary outcome. The results of an experiment comparing two treatments are given by Table 10.29. If $X = 1$ denotes new treatment and $Y = 1$ improvement, then we have that

Now now that we consider the mean and the women separately; introduce

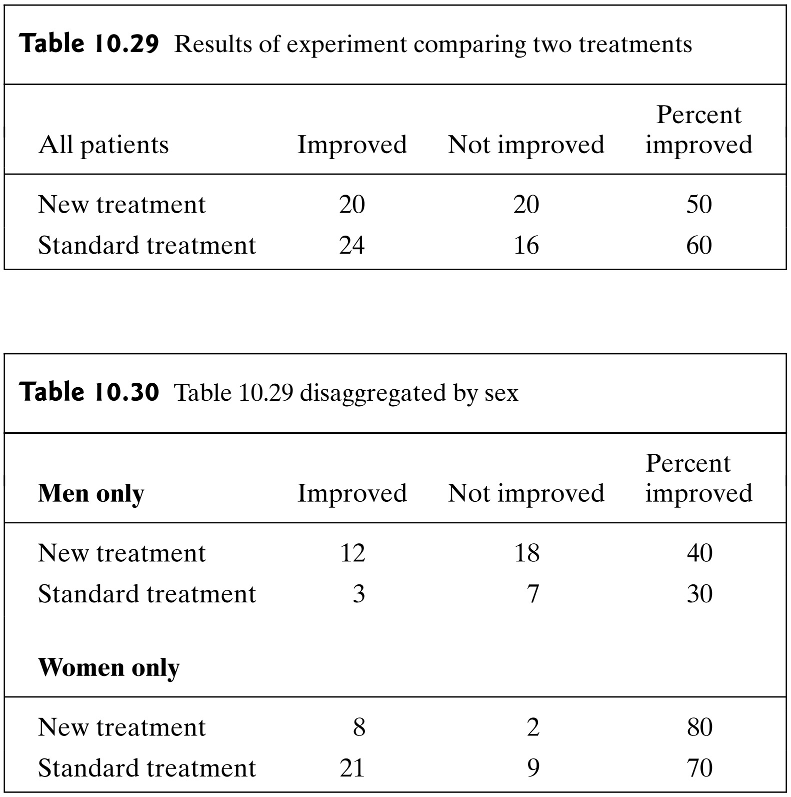

Table 10.29 Results of experiment comparing two treatments

All patients	Improved	Not improved	Percent improved
New treatment	20	20	50
Standard treatment	24	16	60

Table 10.30 Table 10.29 disaggregated by sex

Men only	Improved	Not improved	Percent improved
New treatment	12	18	40
Standard treatment	3	7	30
Women only			
New treatment	8	2	80
Standard treatment	21	9	70

Figure 1: From DeGroot and Schervish p. 654

the binary variable Z where $Z = 1$ if the subject is male. Table 10.30 shows the values disaggregated by considering the man and women separately. We have that

Explain the paradox.

Exercise 1: Explain and prove why the paradox above cannot occur if

$$\mathbb{P}(Z|X=1) = \mathbb{P}(Z|X=0).$$

Here's a hint (use proof by contradiction and the law of total probability)

Exercise 2: How would you adjust for gender? [Hint: review AOS Sec 16.3 on "adjusted treatment effect" and go through **Example 2** below]

Now let us look at a similar example from Agresti's book *An Introduction to Categorical Data Analysis* with real data from the *Florida Law Review 43:1-34 (1991)*.

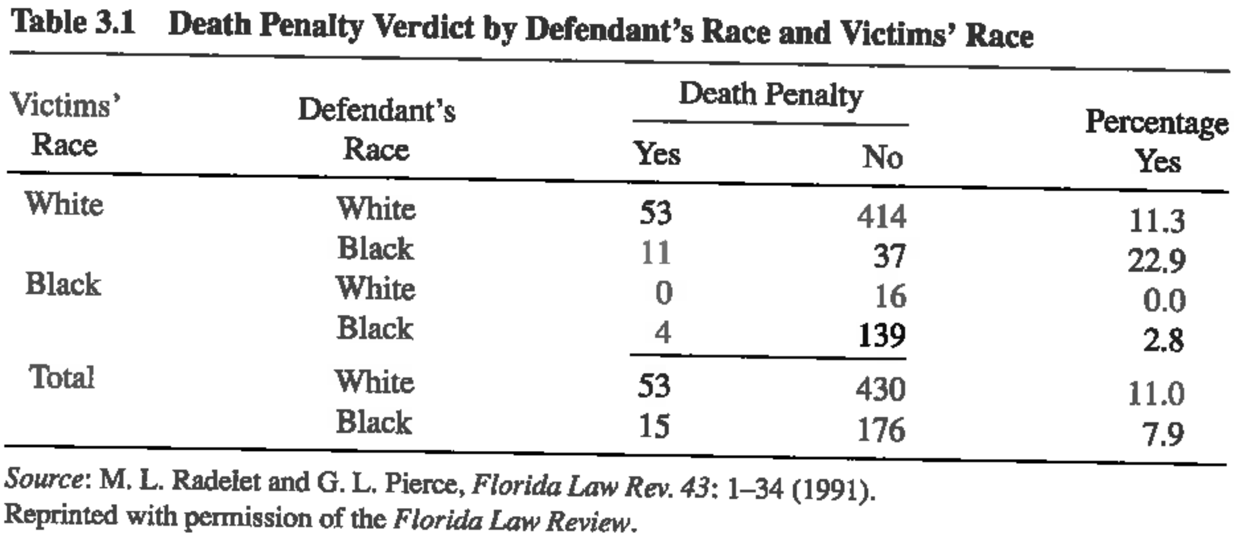

Table 3.1 Death Penalty Verdict by Defendant's Race and Victims' Race

Victims' Race	Defendant's Race	Death Penalty Yes	Death Penalty No	Percentage Yes
White	White	53	414	11.3
	Black	11	37	22.9
Black	White	0	16	0.0
	Black	4	139	2.8
Total	White	53	430	11.0
	Black	15	176	7.9

Source: M. L. Radelet and G. L. Pierce, *Florida Law Rev. 43*: 1–34 (1991). Reprinted with permission of the *Florida Law Review*.

Figure 2: From Agresti's "An Introduction to Categorical Analysis".

Example 2. (Death Penalty Example) The article studies the effects of racial characteristics on whether individuals convicted of multiple murders in Florida between 1976 and 1987 receive the death penalty. The variables are Y="death penalty verdict" (having categories no/yes), and X="race of defendant" and Z="race of victims" (each having categories white/black).

(a) First *regress Y (verdict) on X (defendant's race)*. Plot an estimate of the regression function $r(x)$ or the "marginal association". What's the relation to conditional probabilities and the numbers in the table?

(b) *Adjust for victim's race:* Note that Z (race of victims) depends on both X (race of defendant) as well as Y (verdict). Hence, now study the effect of defendant's race on the death penalty verdict, treating victim's race as a control variable. In other words, **regress** Y **on** X **and** Z. Plot an estimate of the regression function or the "conditional association" $r(x, z)$ for $z = B$ and $z = W$. Discuss your results.

1

2

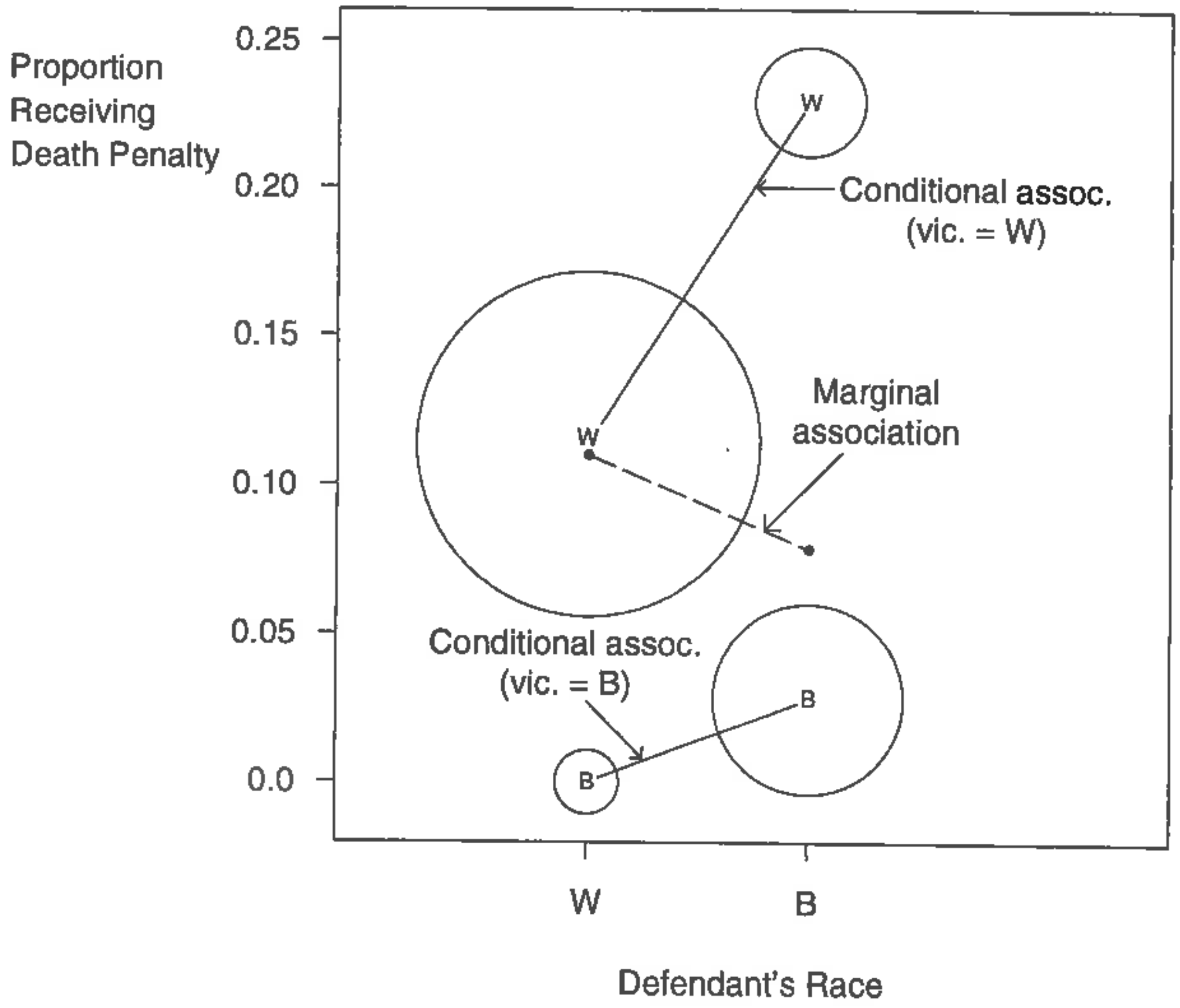

Figure 3.2 Proportion receiving death penalty by defendant's race, controlling and ignoring victim's race.

Figure 3: From Agresti's "An Introduction to Categorical Analysis".

(c) Suppose the victims' race is the *only* confounding variable. Compute the *adjusted treatment effect* of X on Y and plot an estimate of the "causal" regression function $\theta(x)$. How is this computation different from regressing Y directly on X? [Hint: See AOS Remark 16.7 on p.259]

R Demo 2.3 (SAT Example)

Description of Data from the *Statistical Sleuth* by Ramsey and Schafer

In 1982, average SAT scores were published with breakdowns of state-by-state performance in the United States. The average SAT scores varied considerably by state, with mean scores falling between 790 (South Carolina) to 1088 (Iowa).

Two researchers examined compositional and demographic variables to examine to what extent these characteristics were tied to SAT scores. The variables in the data set were:

1. *state*: state name
2. *sat*: mean SAT score (verbal and quantitative combined)
3. *takers*: percentage of total eligible students (high school seniors) in the state who took the exam
4. *income*: median income of families of test takers, in hundreds of dollars
5. *years*: average number of years that test takers had in social sciences, natural sciences, and humanities (combined)
6. *public*: percentage of test takers who attended public schools
7. *expend*: state expenditure on secondary schools, in hundreds of dollars per student
8. *rank*: median percentile of ranking of test takers within their secondary

school classes. Possible values range from 0-99, with 99th percentile students being the highest achieving.

"Notice that the states with high average SATs had low percentages of takers. One reason is that these are mostly midwestern states that administer other tests to students bound for college in-state. Only their best students planning to attend college out of state take the SAT exams. As the percentage of takers increases for other states, so does the likelihood that the takers include lower-qualified students."

Research Question: "After accounting for the percentage of students who took the test and the median class rank of the test takers (to adjust, somewhat, for the selection bias in the samples from each state), which variables are associated with state SAT scores? After accounting for the percentage of takers and the median class rank of the takers, how do the states rank? Which states perform best for the amount of money they spend?"

(a) Exploratory Data Analysis. Read in the data and look at the data set before beginning any formal analysis. First examine the variables individually through *histograms* or box plots; look at the general range of the data, shape (skewed, gapped, symmetric, etc.), as well as any other trends.

Do you see any outliers? One state has almost double the amount of secondary schooling expenditures of all the other states. Which state is it?

Next we look the variables together. A *scatterplot matrix* shows the relationships between the variables at a glance (use the 'round' command to make the output more easily readible). Generally we are looking for *trends* here. Does the value of one variable tend to affect the value of another? If so, is that relationship linear? These types of questions help us think of what type of interaction and higher order terms we might want to include in the regression model. (Note however that *such trends do not imply causation, and that each trend may be nullified or even reversed when accounting for the other variables in the data set!*) We should also look for outliers (we may want to remove high influence points later from the analysis).

(b) Linear Regression. Fit a full regression line to all data. Check the residuals. What do you see?

(c) Using Residuals to Create Better Rankings for SAT Data. First rank the states based on raw SAT scores. This approach however doesn't seem reasonable: Some state universities require the SAT and some require a competing exam (the ACT). States with a high proportion of takers probably have "in state" requirements for the SAT. In states without this requirement, only the more elite students will take the SAT, causing a bias. We would like to rank the states by SAT scores, corrected for percent taking the exam and median class rank. Let's explore this thinking further.

To address the research question of how the states rank after accounting for the percentage of takers and median class rank, we define a reduced model that fits a regression line to 'takers' and 'rank'. Instead of ranking by actual SAT score, we can then rank the schools by how far they fall above or below their fitted regression line value. A residual is defined as the difference between the observed value and the predicted value.

Sort the states by residual value and display the old ranking. What do you see? Do the ranking shift once we control for the variables 'takers' and 'rank'? Analyze the ranks further by accounting for such things as expenditures to get a sense of which states appear to make 'efficient' use of their spendings.

(d) Check the Residuals of the Reduced Model. One of the assumptions of the basic regression model is that the magnitude of residuals is relatively constant at all levels of the response. It is important to check that this assumption is upheld here. Hence, plot the residuals of the reduced model versus 'sat', 'takers' and 'rank'. Do you see any patterns in the residual plots? Can you make some of the relationships in the reduced model l look more linear after a transformation of variables?