STAT 331 Lab 13

Instructions

Submit an HTML markdown document by the beginning of class. Answers to questions need to be included in text form in your HTML markdown document.

Simpson's Paradox

Simpson's paradox occurs when we observe a certain relationship between two variables, but when observations are divided into subgroups the opposite relationship is apparent. The paradox occurs because the grouping variable is a confounding variable which drives the observed relationship. Simpson's paradox may look like the below plots. Here, X and Y are negatively related (left). But when a third grouping variable is accounted for, we see that the relationship is positive within each group (right):

Simpson's paradox occurs frequently in real life data sets in the social sciences and public health. You are going to produce some plots like these using real data. The key point is that omitting confounding variables may produce an opposite relationship than the truth!

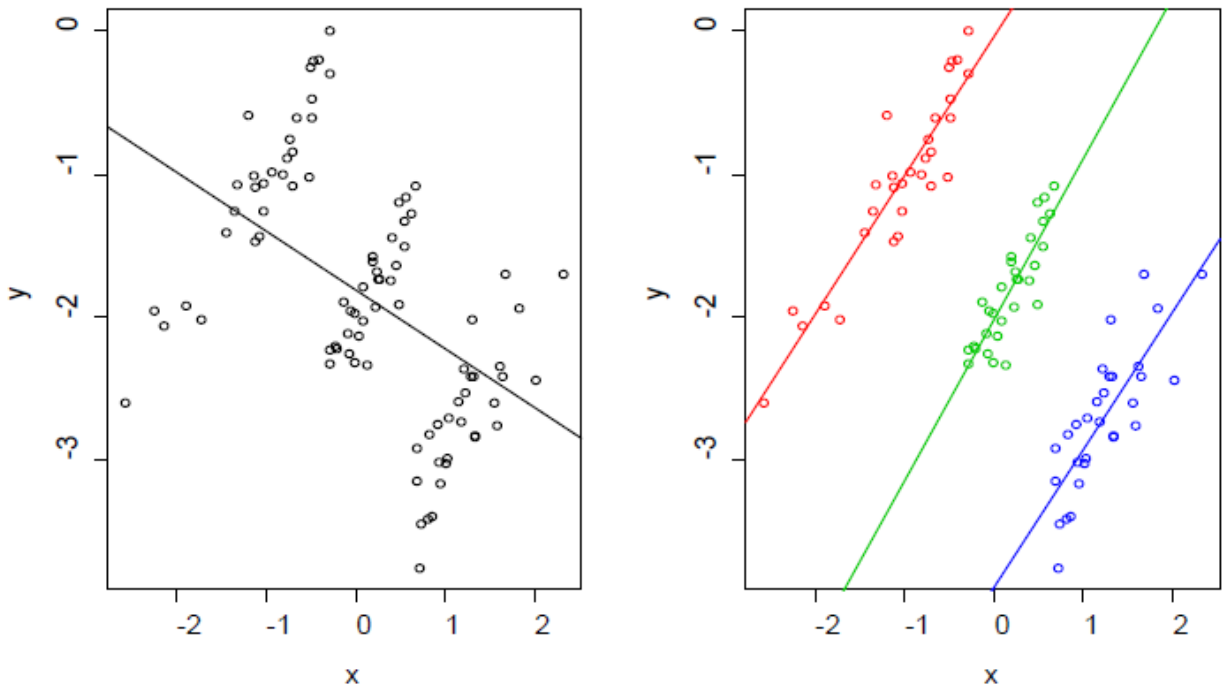
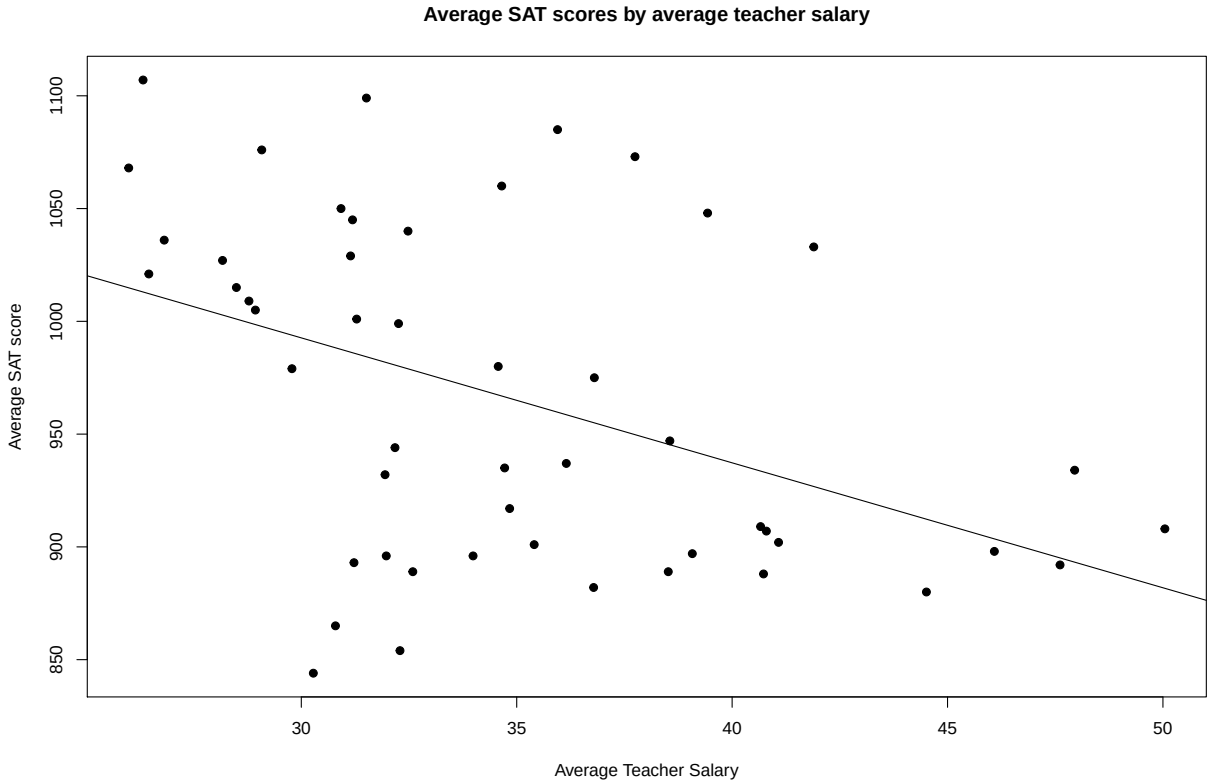


Figure 1:

Exercises

1. The data set we will work with is located in the mosaic library. Install and load the mosaic package, and load the mosaicData library. The data set is called SAT. Type data(SAT) to load the data frame into your workspace. This data set includes SAT scores, pupil/teacher ratio, average teacher salary, and other variables in all 50 states. Type ?SAT to see a description of the data set and variables
2. Replicate the following plot. This plot shows a negative relationship between average SAT score and teacher salary. Does this mean that increasing teacher salary decreases average SAT scores? Why or why not?

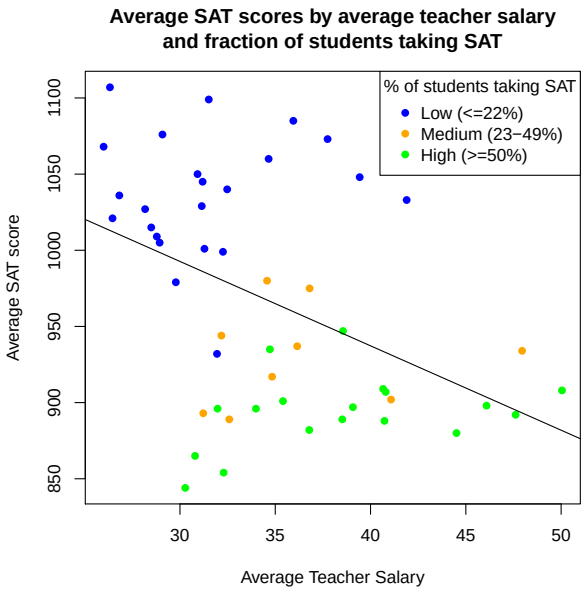


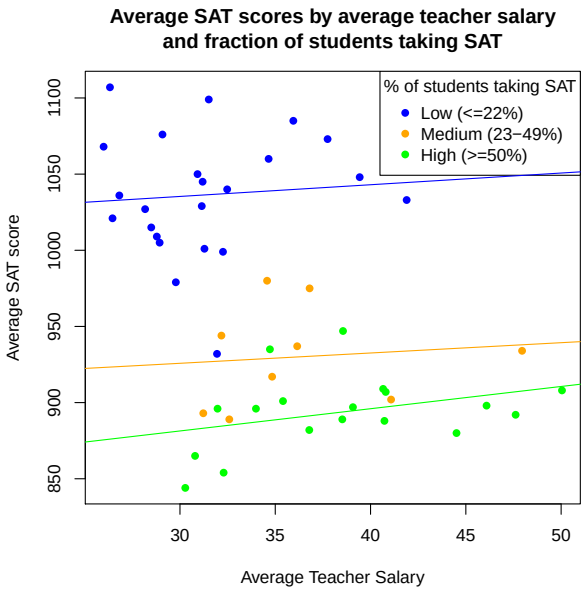
3. Now produce a scatterplot matrix showing the relationship between expend, ratio, salary, frac, and sat. Which variable appears to have the strongest relationship with SAT score? Describe this relationship.

4. We are going to examine a confounding variable - the fraction of students taking the SAT exam. In poorer states, where teacher salaries are lower, fewer students aspire to attend college, and a smaller fraction of students take the SAT, usually only the best and brightest. Type ?cut. This is a command to convert a numeric variable into a factor. Use the cut function to create a factor variable that divides states into the following groups:

- "low" = % of students taking SAT is ≤ 22
- "medium" = % of students taking SAT is > 22 and ≤ 49
- "high" = % of students taking SAT is > 49

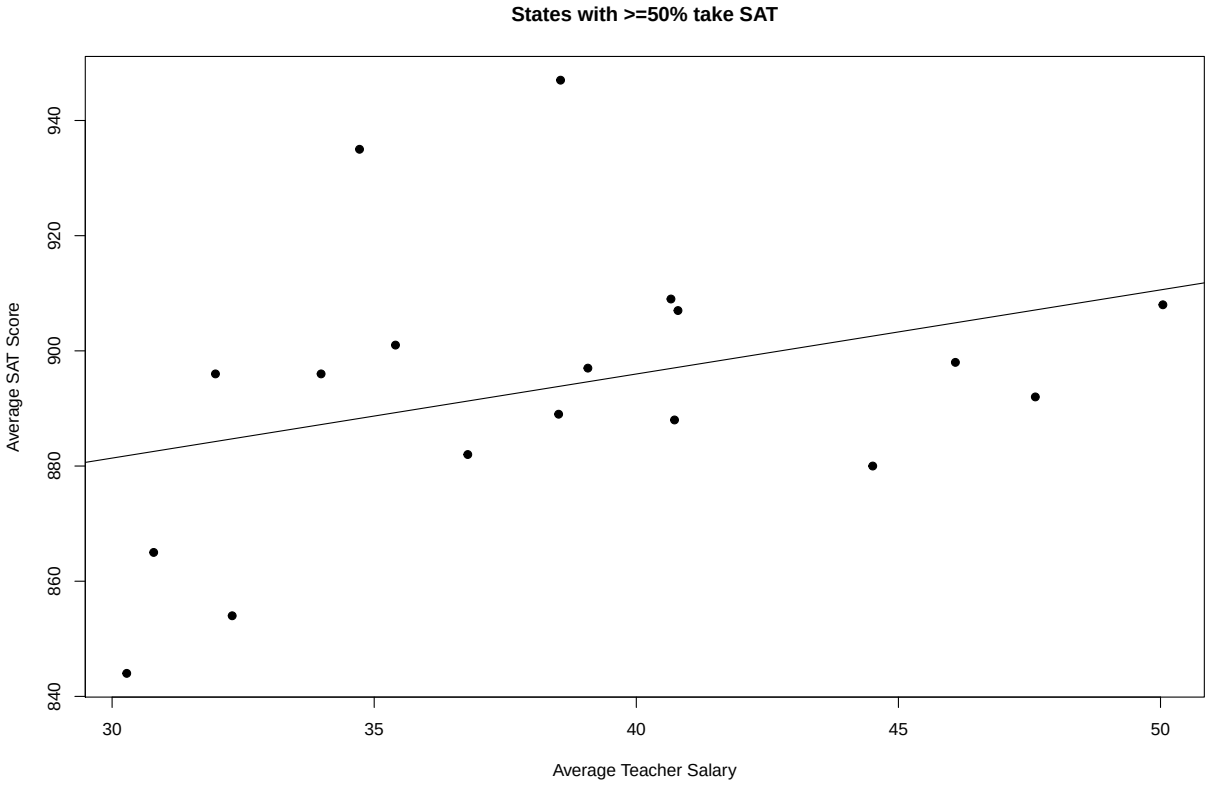
5. Produce the following two plots, showing that the relationship between SAT score and teacher salary is positive within each group:





6. In light of this new evidence, does increasing teacher salary cause SAT scores to decline? Why or why not?

7. For the remaining exercises, we'll focus only on states in which at least 50% of students take the SAT. Create a data set called fracGT50 which contains these observations. Produce the following scatterplot:



8. Compute the linear regression model of to predict average SAT score by average teacher salary among the states that have a high proportion of students taking the SAT. Display a summary table of your

model (coefficients, standard errors, t-statistics, p-values). Interpret the coefficients and describe the relationship between these variables.

9. Produce the following diagnostic plots: standardized residual vs fitted value, histogram of standardized residuals, and a normal quantile plot of standardized residuals. Do regression assumptions appear to be met? Why or why not? (Be sure to discuss which plot you are examining to assess which assumption.)