STAT 1291: Data Science

Lecture 17 - Cause and effect

Sungkyu Jung

Where are we?

- *data visualization*
- *data wrangling*
- *professional ethics*
- *statistical foundation*
- *Statistical modeling: Regression*
- **Cause and effect: Causality and confounding**
- *More statistical modeling*

Causality

- Does the death penalty have a deterrent effect?
- Is chocolate good for you?
- Is drinking coffee linked to pancreatic cancer?
- Isn't it odd that Florida has so many people living with Alzheimer's Disease?

Observation

Observation is a key to good science. An observational study is one in which scientists make conclusions based on data that they have observed but had no hand in generating.

Does drinking coffee cause cancer?

"A statistical link between the drinking of coffee and cancer of the pancreas . . . was reported by scientists of the Harvard School of Public Health. (NYT, 1981)"

The report continued:

"Data were obtained by interviews with *369 pancreatic cancer patients* at 11 hospitals in Boston. For comparison, the scientists . . . asked the same questions of *644 patients* comparable in age and sex who were hospitalized for a variety of reasons *unrelated to the pancreas*."

(http://www.nytimes.com/1981/03/12/us/study-links-coffee-use-to-pancreas-cancer.html)

Obvervational study

- *individuals*, study subjects, participants, units:
  - patients at 11 hospitals in the Boston metropolitan area
- *treatment*:
  - coffee consumption
- *outcome*:
  - pancreatic cancer

The fundamental question:

Does the treatment have an effect on the outcome?

First question

Has the coffee consumption *any relation* to the pancreatic?

> *Is the treatment associated with on the outcome? (Any relation = Association)*

The NYT report answers on this . . .

"When the results of all the interviews were analyzed, the scientists found only a weak association between cigarette smoking and pancreatic cancer, none with alcohol consumption, *an unexpected strong association with coffee consumption*, but none with tea."

Next question

Does the coffee consumption *lead to* the pancreatic?

Causation.

This question is often harder to answer.

From the report, "... Although the statistical association *does not prove that coffee causes cancer*..."

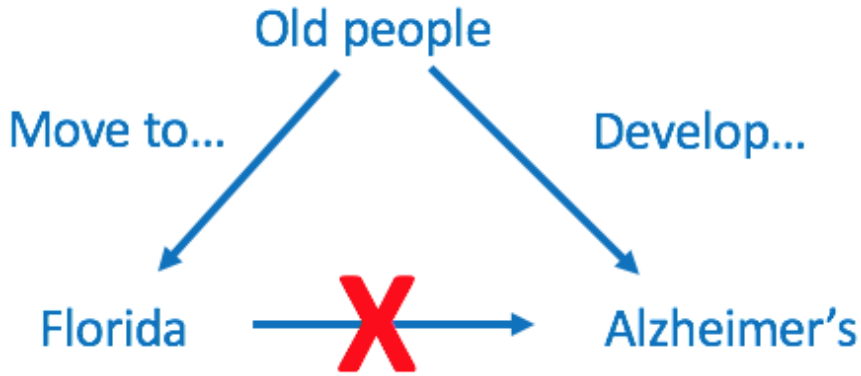
Why?

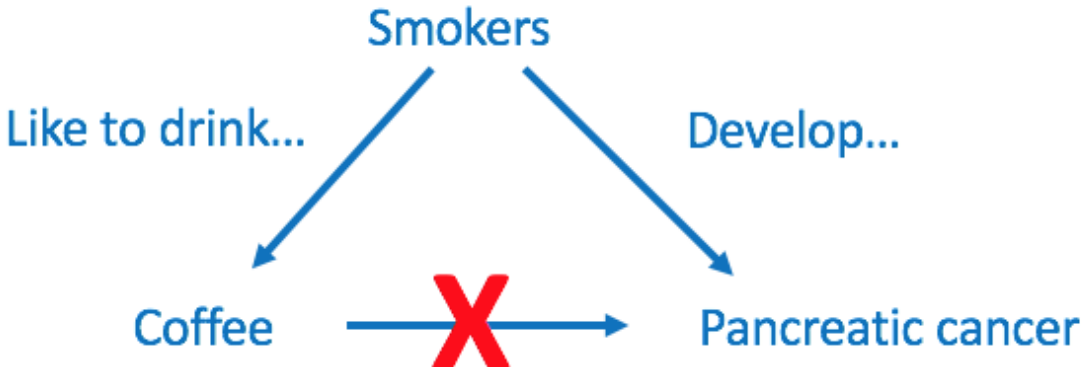
The problem is that the authors didn't *control for smoking*. A lot of people who drink coffee also smoke.

In this study, "smoking" is called a confounding variable.

Confounding

- Isn't it odd that Florida has so many people living with Alzheimer's Disease?
- Is drinking coffee linked to pancreatic cancer?



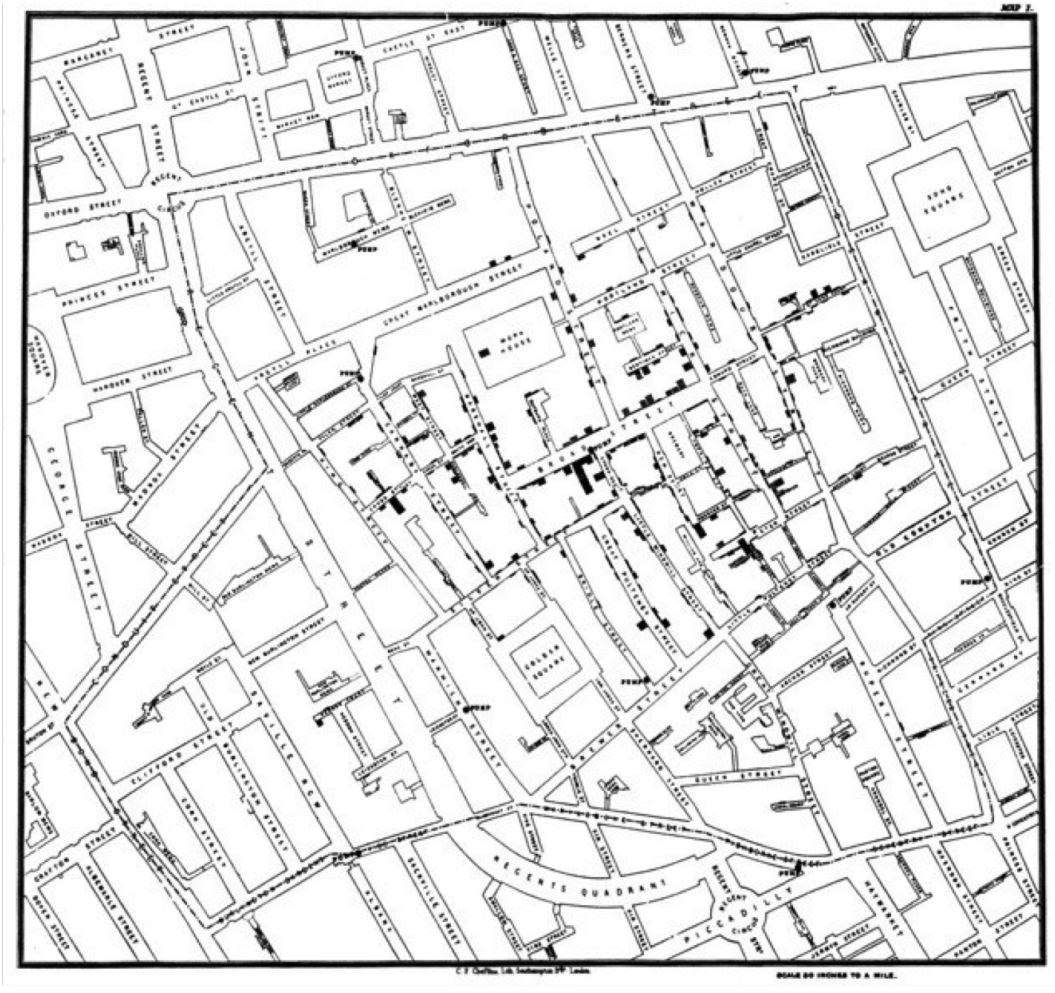


Strategies to reduce confounding

1. Randomization (aim is random distribution of confounders between study groups)
2. Matching (of individuals or groups, aim for equal distribution of confounders)
3. Stratification (confounders are distributed evenly within each stratum)
4. Adjustment (usually distorted by choice of standard)
5. Multivariate analysis (only works if you can identify and measure the confounders)

We will see a couple examples today.

The birth of epidemiology



John Snow
(1813 – 1858)

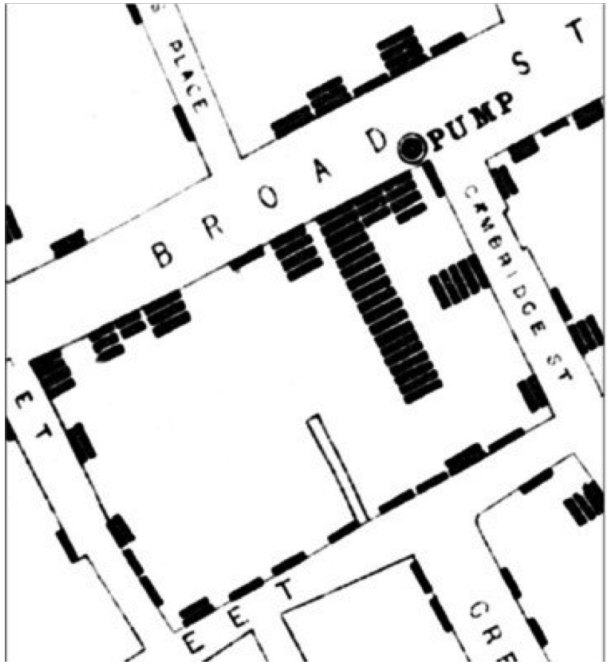


Figure 2: Remember Jon Snow?

- In 1854, cholera struck in London. As the deaths mounted, Snow recorded them diligently, using a method that went on to become standard in the study of how diseases spread: *he drew a map*.
- On a street map of the district, he recorded the location of each death.
- The pump's water was contaminated by sewage from the houses of cholera victims.
- Snow used his map to convince local authorities to *remove the handle of the Broad Street pump*.

Towards Causality

- The map gave Snow a strong indication that the cleanliness of the water supply was the key to controlling cholera
  - An asnwer to the *question of association*
- Still a long way from a convincing scientific argument that contaminated water was causing the spread of the disease.
  - He used *a method of comparison*

Snow's "Grand Experiment"

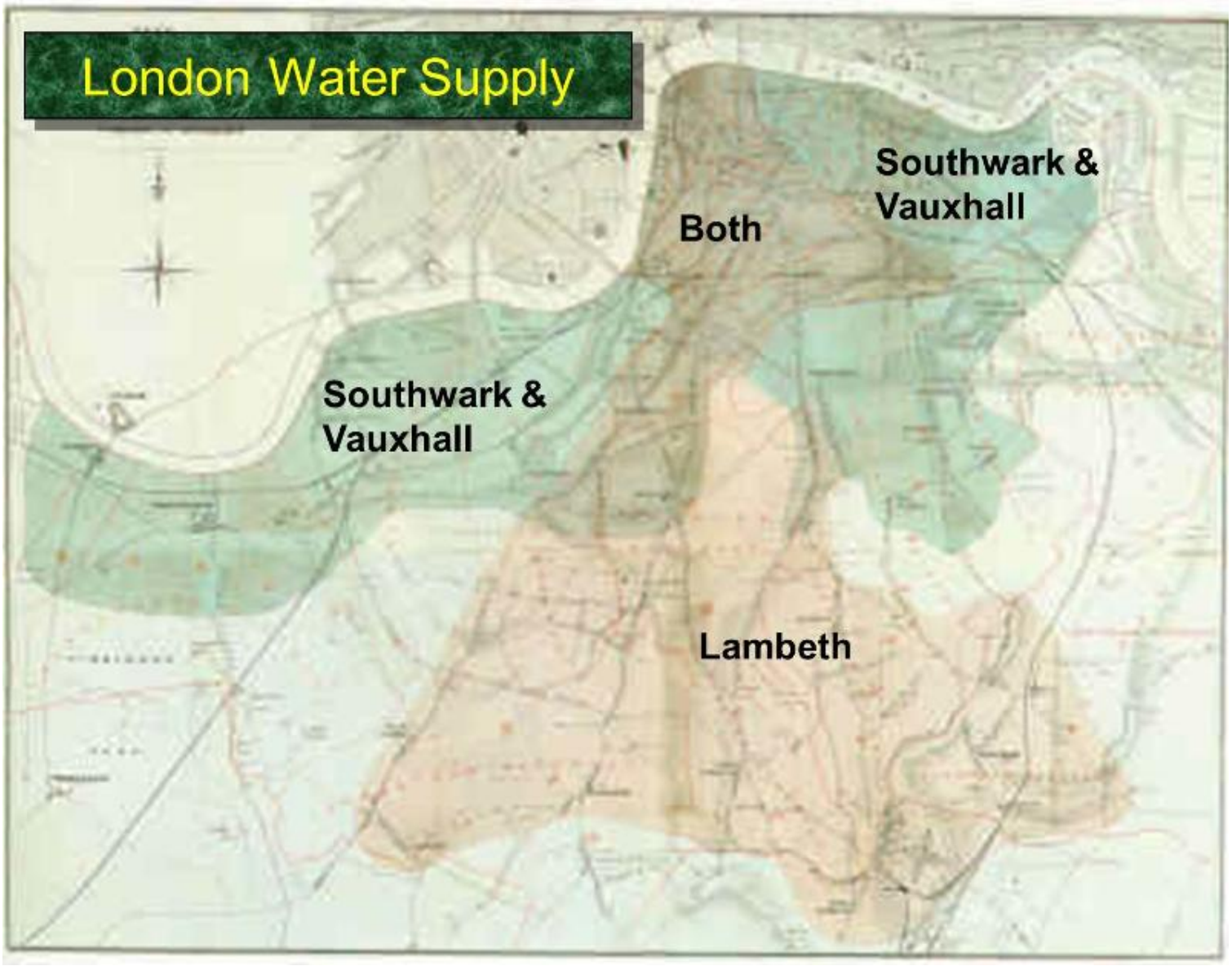


Figure 3:

- Lambeth water company : relatively clean.
- Southwark and Vauxhall (S&V) company : its supply was contaminated.

Snow wrote:

"In many cases a single house has a supply different from that on either side. Each company supplies both rich and poor, both large houses and small; there is no difference either in the condition or occupation of the persons receiving the water of the different Companies"

(Snow, On the Mode of Communication of Cholera, 1855)

Comparison

- Treatment group
  - Group of individuals who receive water from S&V
- Control group
  - does not receive the treatment; in this case, water supply from Lambeth

Snow's "Grand Experiment"

Snow's report continued:

"there is no difference whatever in the houses or the people receiving the supply of the two Water Companies, or in any of the physical conditions with which they are surrounded"

The two groups were **similar** except for the treatment.

Snow's result

The numbers pointed accusingly at S&V.

Supply Area	Number of houses	cholera deaths	rate
S&V	40046	1263	315
Lambeth	26107	98	37
Rest of London	256423	1422	59

(Rate = deaths per 10,000 houses)

Key in establishing causality

If the treatment and control groups are *similar apart from the treatment*, then differences between the outcomes in the two groups can be ascribed to the treatment.

Group homogeneity is often difficult (or impossible) to achieve in observational studies, where *other factors* may be the determinants of the observed relationship between two factors. Such other factors may confound the relationship being studied.

Randomize!

By controlling *at random* who receives a treatment and who receives a control, you ensure that, on average, all other factors are balanced between the two groups.

> *Randomized trial is the ideal method of experiment for establishing causality.*

Randomized trial is not so practical

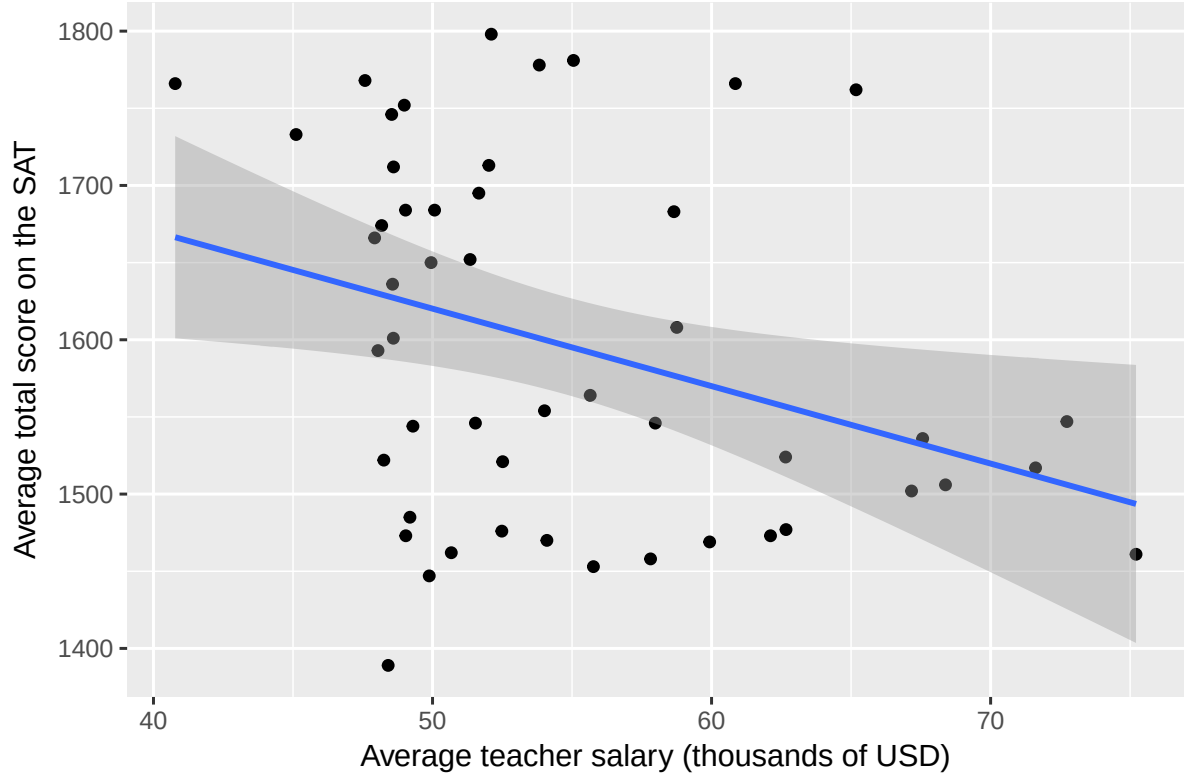
- Randomize some children to smoke and the others not to smoke in order to determine whether cigarettes cause lung cancer.
- Randomize adults to either drink coffee or abstain to determine whether it has long-term health impacts.
- Observational data may be the only feasible way to answer important questions.

- Stratification and multivariate analysis could identify and control confounding variables: See the next example.

Lower teacher salaries lead to high SAT scores?

A natural question is:

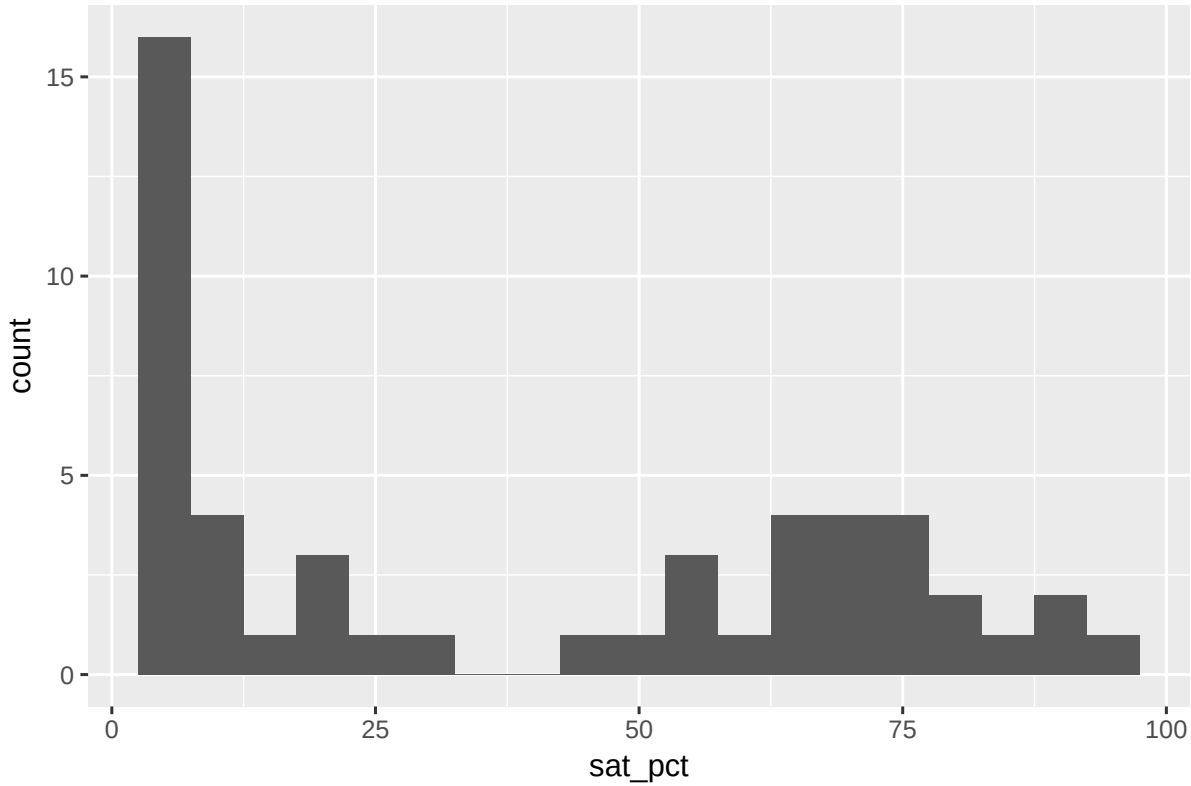
Are higher teacher salaries associated with better outcomes on the test at the state level?



```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1871.104    113.141  16.538   <2e-16 ***
## Salary        -5.019      2.048  -2.451   0.0179 *
##
## Residual standard error: 111.2 on 48 degrees of freedom
## Multiple R-squared:  0.1113, Adjusted R-squared:  0.092
## F-statistic: 6.008 on 1 and 48 DF,  p-value: 0.01793
```

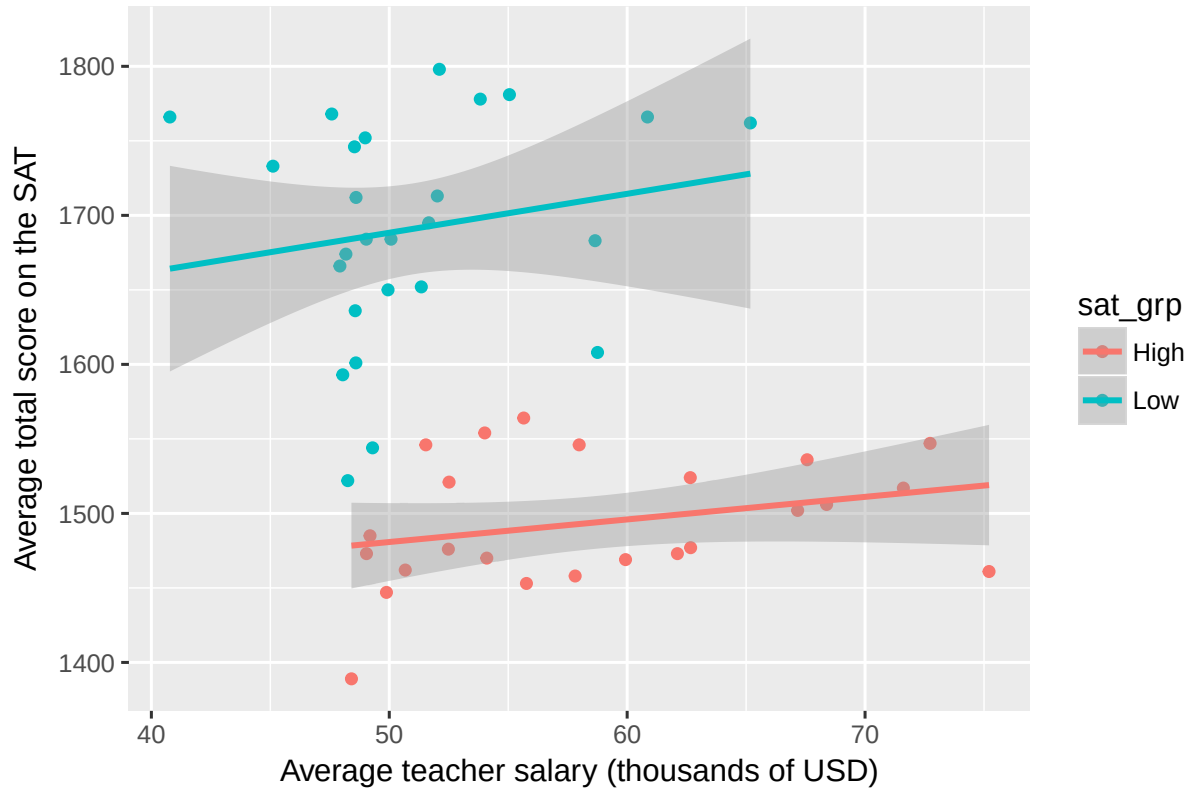
Lurking in the background

The percentage of students who take the SAT in each state varies dramatically (from 3% to 93% in 2010).



We can create a variable called `sat_grp` that divides the states into two groups.

```
SAT_2010n <-
  SAT_2010 %>%
  mutate(sat_grp = ifelse(sat_pct > 30, "High", "Low"))
```


Average total score on the SAT
1800
1700
1600
1500
1400
40
50
60
70
Average teacher salary (thousands of USD)
sat_grp
High
Low

```r
coef(lm(total ~ Salary,
        data = filter(SAT_2010n, sat_grp == "Low")))
```

```
## (Intercept)      Salary
## 1557.658858    2.613381
```

```r
coef(lm(total ~ Salary,
        data = filter(SAT_2010n, sat_grp == "High")))
```

```
## (Intercept)      Salary
## 1405.048718    1.515035
```

Stratification

Stratification (simply means grouping) can control the confounding variable `sat_grp`.

- For each group given by the values of `sat_grp`, average teacher salary is positively associated with average SAT score.
- When we collapse over this variable, average teacher salary is negatively associated with average SAT score. (`sat_grp` or `sat_pct` is confounding here.)
- This form of confounding is called *Simpson's paradox*.

Multiple regression

Multiple regression is another way of controling confounding variables.

```
SAT_mod2 <- lm(total ~ Salary + sat_pct, data = SAT_2010)
msummary(SAT_mod2)
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1589.0065    58.4707  27.176   <2e-16 ***
## Salary         2.6370     1.1488   2.295   0.0262 *
## sat_pct       -3.5526     0.2785 -12.756   <2e-16 ***
##
## Residual standard error: 53.18 on 47 degrees of freedom
## Multiple R-squared:  0.8008, Adjusted R-squared:  0.7924
## F-statistic: 94.49 on 2 and 47 DF,  p-value: < 2.2e-16
```

- the slope for `Salary` is positive and statistically significant when we control for `sat_pct`.

Conclusion

- Correlation does not imply causation
- Because of potentially lurking counfounding variables
- To infer causal relations, do randomized trials (as opposed to observational studies)
- Almost all examples of Big data are observational
- Consider other methods in controling confounding variables: conditional modeling (stratification / multiple regression), matching.