4.2 Cautions about Correlation and Regression

Two statisticians were traveling in an airplane from Los Angeles to New York City. About an hour into the flight, the pilot announced that although they had lost an engine, there was no need for worry as the plane had three engines left. However, instead of 5 hours travel time it would now take them 7 hours to get to New York. A short while later, the pilot announced that a second engine failed. They still had two left, but it would take 10 hours to get to New York. Somewhat later, the pilot announced that a third engine had died. Never fear, he announced, because the plane could fly on a single engine. However, it would now take 18 hours to get to New York. At this point, one statistician turned to the other and said,

"Gee, I hope we don't lose that last engine, or we'll be up here forever!"
Pickles will kill you! Every pickle you eat brings you nearer to death. Pickles are associated with all major diseases of the body. They can be related to most airline tragedies. There exists a positive relationship between crime waves and consumption of this preserved fruit of the cucurbit family. For example:

* Nearly all sick people have eaten pickles. The effects are obviously cumulative.
* 99.7% of the people involved in air and auto accidents ate pickles within 14 days of preceding the accident.
* 93.1% of juvenile delinquents come from homes where pickles are served frequently.

Evidence points to the long-term effects of pickle-eating:

* Of the people born in 1869 who later ate pickles, 100% are now dead.
* All pickle eaters born between 1869 and 1879 have wrinkled skin, have lost most of their teeth, have brittle bones and failing eyesight, if the ills of eating pickles have not already caused their death.
* Even more convincing is the report of a noted team of medical specialists: Rats that were force-fed 2 pounds of pickles per day for 30 days developed bulging abdomens. Their appetites for wholesome food were destroyed.
extrapolating:
- using a model to predict outcomes outside the values used to create the model
- may not be accurate
- like driving a car blindfolded, getting directions from someone who's looking out the rear window

example:
Using world records for the 100 meter to make a LSRL, we predict that **at some point in the future**, runners will finish before they start.
lurking variable:
• may affect interpretation of relationships among variables
• not initially considered
• thought of as invisible or unseen

does use of sunblock cause a craving for ice cream?!
No, some lurking variable probably affects sales of both.

eample:
trend-- as sales of sunblock increase, so do ice cream sales
Averaging data reduces variation.

Models from averaged data are poor predictors of individual outcomes.

eample:
Hard to perfectly predict St. Louis temp. on a specific day in July based on average daily high for July.
Cause and Effect

- study relationship between two variables
- frequent goal: establish a cause-and-effect relationship
- hope to say changes in explanatory variable cause changes in response variable

examples
Does saccharin cause cancer?
Does living near power lines cause cancer?
Does education cause an increase in income?
Does praying cause you to live longer?
Does affirmative action cause less racism?
A detected association can be the result of:

- cause and effect
- common response to other variable(s)
- confounded relationship with other variable(s)
Causation

Strong evidence for causation comes from an experiment. (change x, hold all other variables constant and measure whether y changes with changes in x). If so, then we have established that x causes y.

This can't be done in observational studies (observe and record variables and changes in variables, without actively changing any of them).

Some experiments can't be conducted due to practical or ethical issues.

descriptions
Saccharin intake correlates highly with bladder cancer in lab rats. In such experiments, all other variables can be held constant. Through repeated experimentation, causation can be established.

When I add sugar to my tea, the calories increase.
**Common Response**

association between $x$ and $y$ is explained by a lurking variable, $z$. Both $x$ and $y$ change in response to $z$. This produces the association between $x$ and $y$.

**examples**

We see a high correlation between high school gpa and SAT scores.

higher intelligence results in a higher gpa and SAT score

When my heating gas bill goes up, the humidity levels in my house go down.

Cold air holds less water vapor, so the air is dry already and the cold temperatures cause me to crank up the heat, further lowering relative humidity.
Confounding

effects of explanatory and lurking variables on a response variable can't be distinguished

examples
We see a high correlation between education and income. What explains this?

Maybe causation, probably some confounding. Wealthy families can better afford the best colleges.

A runner wants to improve performance, so he runs an extra two miles a day and makes some diet changes. In two weeks his time has improved somewhat.

can't tell whether it was from the running or the diet.

Two years ago a new math teacher was hired at a particular school and last year the school adopted a new math textbook. State test scores came in and were much higher than in the past.

The increases could be due to the book, the teacher, or both.
lurking
*unaccounted for
*not considered initially
*plausibly explains the apparent x y relationship
*not measured but can still affect response variable

confounded
*identified but we can't determine which one(s) impact response variable
*associated with one another
*effect can't be distinguished or separated
Confounded variables: you can't tell which (or whether it's the combination) had an affect.

You might want to test a fertilizer on your lawn. Suppose you spread it on half the lawn to see if the grass will look better there. If you spread it on the sunny half, leaving the shady half unfertilized, you won't know whether the greener grass resulted from fertilizer or sunshine (or the two together).

Another possibility: Let's say that the sunny patch gets no fertilizer and the shady side gets the fertilizer. What if no differences are found? Can we say that NEITHER fertilizer nor sun condition has any impact? No, we just don't know.

In both cases above, we don't know which factor "caused" the difference.

Lurking (sometimes called common response) variables: one variable drives two others, creating the impression of an association between them.

For countries, pick any measure of technological modernity (# of TVs per capita) and life expectancy. You'll clearly see an association - countries with fewer TVs have lower life expectancy. Such lurking variables as general economic well-being and standard of living probably explain both. We don't think that having a TV increases your lifespan.

If you do a test of aspirin vs. nothing for a headache (i.e., no placebo), then your results will be confounded - is improvement due to the aspirin, or the placebo effect (just taking something helps)? There is nothing lurking, but confounding messes up the search for causation.

Similarly, other poor experimental designs can lead to the same problem. If you decide to try a new fertilizer, and the instructions say to use more water with it, and it leads to more growth than the old fertilizer, is it due to the new fertilizer or the extra water? Again, nothing lurking, but the variables are confounding.
Some experiments just can't be done:

Does smoking cause cancer?  
We can't ethically perform the required experiment; it would be evil to force people to smoke in order to test this.

Does living near power lines cause cancer?  
We can't ethically perform the experiment here either because we can't force people to live in a certain place.

Does drinking milk cause weight loss?  
We can't practically perform this because there is no way to track this while controlling for other variables-- people don't live in a vacuum.
establishing causation
How do you establish causation (statistically) if you can't perform the required experiment?

association must be strong
association must be consistent
higher doses (or equivalent) must be associated with stronger response
alleged cause must happen before effect
alleged cause must be plausible

remember!
When conducting observational studies, you can't infer causation.
1. Is the association positive, negative, or none?
2. Is this causation, common response, or confounding?
3. For common response or confounding, write down other plausible variables.

1. On a diet: daily calories eaten vs. the amount of weight lost
2. On a beach: amount of ice cream consumed vs. the number of people in the water
3. # of pets owned vs. amount spent on pet food
4. cost of someone's house vs. cost of someone's car
5. time studying vs. GPA
6. # of police visible on a road vs. the speed cars travel
7. algebra grade vs. geometry grade
8. person's height vs. $ that person has
9. # of wins a MLB team has in a season vs. total $ spent on concessions at their games
10. # of cigarettes sold per year vs. # of people who are diagnosed with lung cancer that year
11. # of people in a family vs. # of cars they own
12. # of problems on a math test vs. time for students to take it