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.

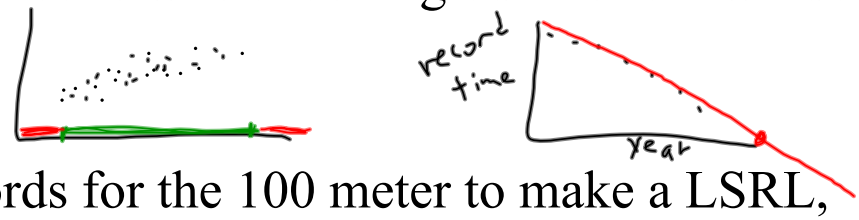
Evidence points to the long-term effects of pickle-eating:

* 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.

<u>extrapolating</u>:

- using a model to predict outcomes outside the values used to create the model
- may be inaccurate
- 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 **<u>at some point in the future</u>**, runners will finish before they start.

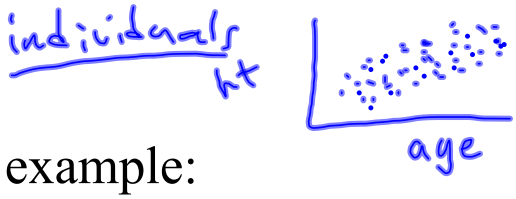
lurking variable:

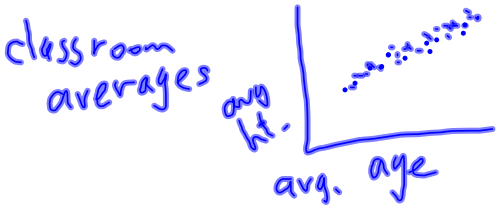
- may affect interpretation of relationships among variables
- not initially considered
- thought of as invisible or unseen

example:
trend-- as sales of sunblock increase, so do ice cream sales
Sunblock use causes cravings for ice cream?!
No, some lurking variable probably affects sales of both.

Averaging data reduces variation.

Models from averaged data are poor predictors of individual outcomes.





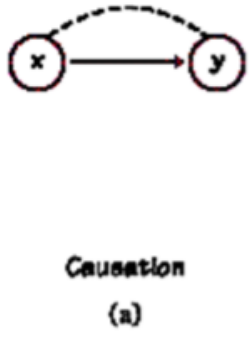
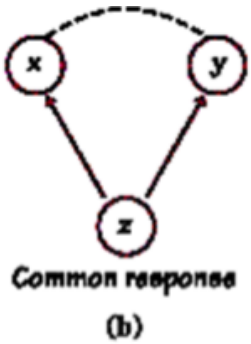
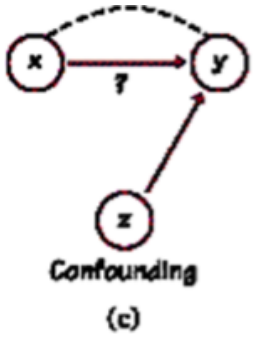
example:
Hard to perfectly predict St. Louis temp. on a specific day in July based on average daily high for July.

Now try exercises 27, 29, & 31
on page 230.

Cause and Effect
goal: verify that changes in explanatory variable **cause** changes in response variable
how: study relationship between two variables

examples
Does saccharin cause cancer?
Does living near power lines cause cancer?
Does drinking milk cause weight loss?
Does praying cause people to live longer?

A detected association can be the result of:

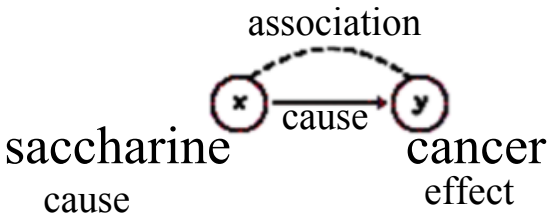
		
cause and effect	common response to other variable(s)	confounded relationship with other variable(s)

Strong evidence for **causation** comes from an experiment: keep other variables constant & measure changes in y as we change x.

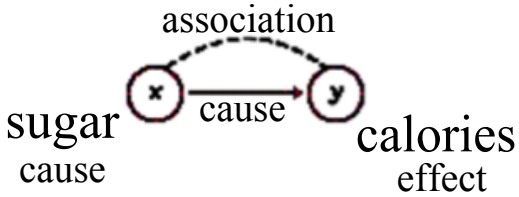
Some experiments can't be conducted due to practical or ethical issues.

examples
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.



Causation
(a)



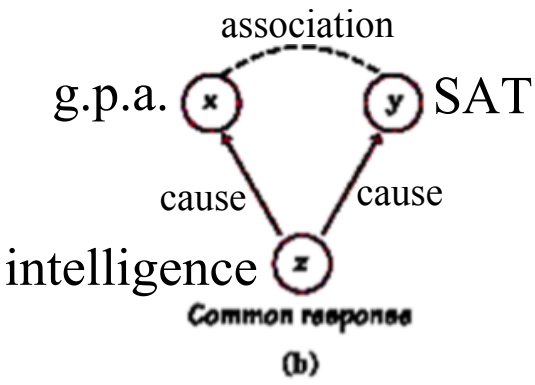
Causation
(a)

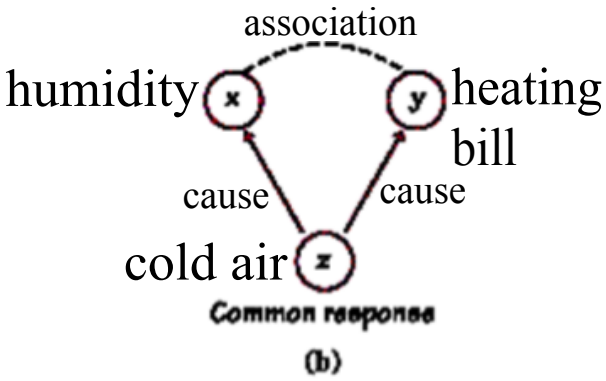
If an association between x and y is explained by a lurking variable, z, both x and y have a **common response** to z.

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 is responsible for both the drier air and the increased heating bill.)



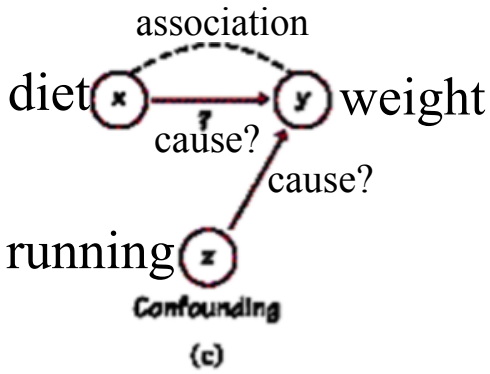


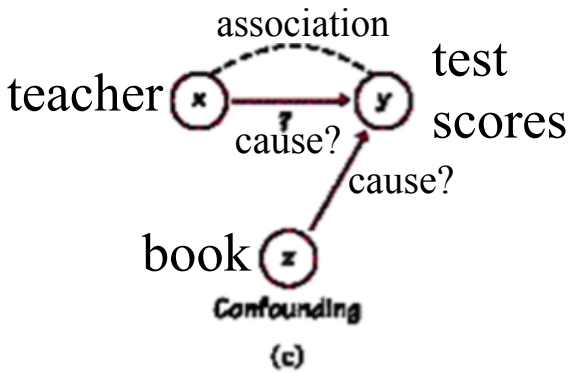
When the effects of explanatory and lurking variables on a response variable can't be distinguished, we say the variables are **confounded**.

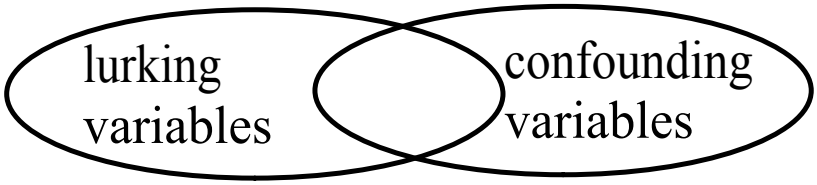
examples

To lose weight, Chris runs an extra 2 miles/day, makes some diet changes, & dropped 6 lbs in 2 weeks. (How much was due to running & how much from diet?)

A new teacher was hired, the school adopted a new math textbook, & state test scores increased. (The increases could be due to the book, the teacher, or both.)







lurking:
often unaccounted for or not considered and may affect the response variable.

confounded:
identified, but the effect can't be distinguished or separated.

Confounded variables: you can't tell which (or whether it's the combination) had an affect.

Why is a bunch of the grass in my backyard dead?

- my dog
- poor drainage
- too much shade
- the fact that I don't water it
- tree roots compete with the grass

Who knows?
Is it one of these
or a combination?

Lurking variables: we may not have considered that some other variable could affect.

Why did I have a hard time getting grass to grow in the far corner of my yard?

- I thought it needed more water.
- I thought the soil was washing away.
- I eventually dug around and discovered concrete and rocks just below the surface.

Who knew?
I never thought of that before!

Some experiments just can't be done:

Does smoking cause cancer? unethical
Does living near power lines cause cancer? unethical
Does drinking milk cause weight loss? impractical

1.Association: +, -, none?
2.Causation, common response, or confounding?
3.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 dogs owned vs. amount spent on dog 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

Now try exercises 33, 35, & 37
on page 237.