

ANSS

This Time: Interpreting Correlation and Regression

5-31-05 (i)

Next Time: Final Exam Review

Dana Morton

Reading: FPP ch. 11, 12

* Overflow time for final exam review: Sat June 4, 12-2pm

* HW 6 due Thursday

* Final: 8-11am June 6 (same room)

On Thursday:

Exactly what's on final:

- how many problems of each type
- emphasis on new material (since Midterm 1)

Interpreting
Correlation and
Regression

Case Study 19:

- If you follow the typical person along in time, would the graph show what you observe?

- longitudinal question: following one person (or group) along in time.

- The data in the Health Examination Survey was not gathered longitudinally

• Cross sectional data: snapshot of a sample of the population at one moment in time.
not

• longitudinal data: following people along at multiple time points.

Key Point:

It is very hard to draw valid longitudinal conclusions from cross sectional data.

why?

- people born in different years (1880 vs 1930)
• this is the main PCF: secular trend: succeeding generations are taller.

(p119 in reader)

- with cross sectional data, people in snapshot differ from each other on x , y , and PCF's (z)

- observational study:

- can't assign people ages; the best you can do is a longitudinal study.

- age is confounded w/ the year of birth: "age-period cohort effect"

Q:

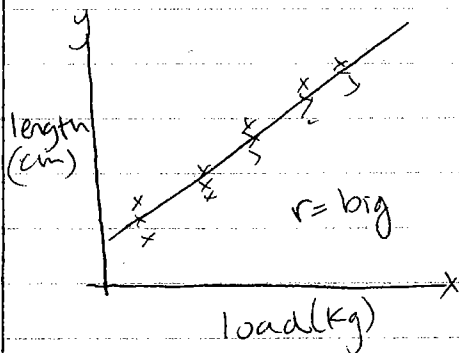
Then why do so many people try to draw longitudinal conclusions from cross sectional data?

A:

Cost: longitudinal data is very expensive and

P. 120 - In-
-terpreting
Correlation &
Regression
Results.

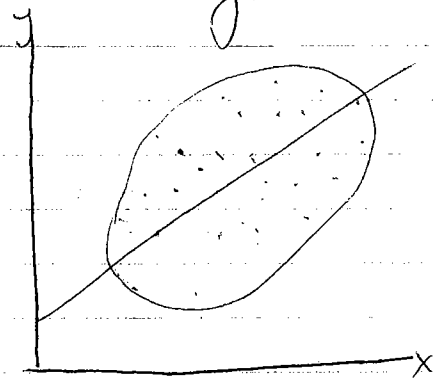
complicated to collect.



lobsters law

- Regression in controlled experiments:
 - when x is under experimental control and all other factors have been held constant, slope in regression equation has a valid cause & effect interpretation.

Regression in
Obs. Study:

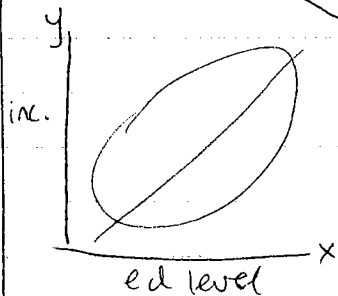


- cross sectional (snapshot) data on n individuals (measure x & y on each person)
- Slope does not necessarily have a valid causal interpretation. Often, the slope has no causal meaning at all.

Ex: #3 p122

variable	mean	SD
ed. level (x)	$\bar{x} = 14 \text{ yrs}$	$s_x = 3 \text{ yrs}$
income (y)	$\bar{y} = \$8000$	$s_y = \$3000$

$r = +0.4$



5-number summary

$$\text{Slope} = r \frac{s_y}{s_x} = (+.4) \left(\frac{\$3000}{3 \text{ yrs}} \right)$$

= \$400 income per year of education.

Q₂:

Does slope mean anything causally?

A₂:

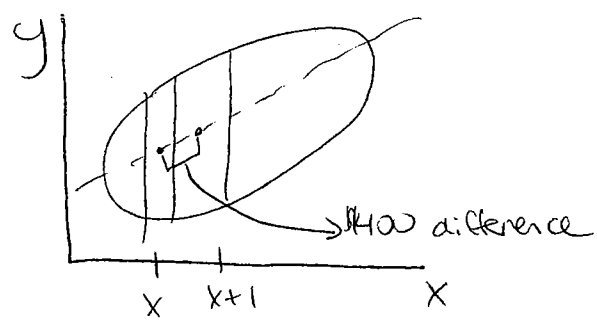
NO, The causal conclusion on pg 122 is an

attempt to draw a longitudinal conclusion from cross sectional data.

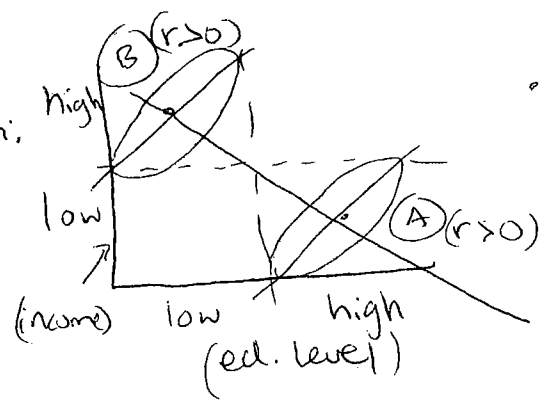
Q₃: Then what does the slope of the regression line mean?

A₃: "Associated with each extra year of education is an increase of \$400 in income on average."

- If I compare 2 groups of women whose ed. levels differ on avg by 1 yr, I expect to see their incomes differing on avg by \$400. This is a cross sectional conclusion from cross sectional data.



Note about economics problem:



• business or academic
is PCF

x - y
association ↔ correlation
correlation ✗ causation
causation → correlation

but x → y or y → x

