

## 4.2 Cautions about Correlation and Regression

### REMEMBER:

- correlation & regression are for linear relationships only
- correlation  $r$  and LSRL are NOT resistant
  - one influential observation can change  $r$  and slope
- extrapolation (be careful predicting beyond domain)

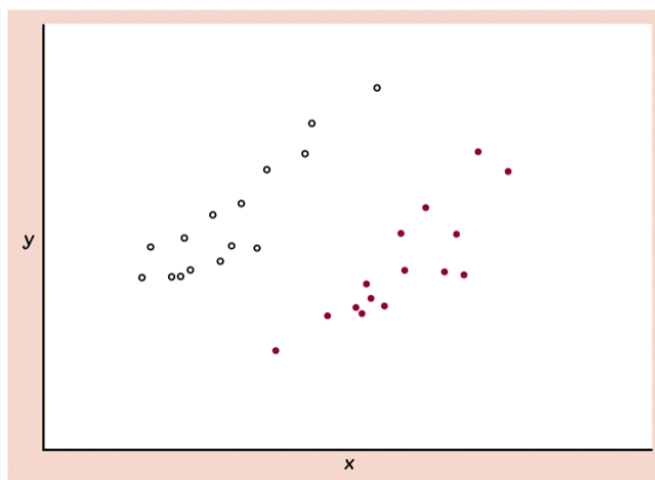
### LURKING VARIABLES:

- with correlation & regression, only measuring 2 variables ( $x$  &  $y$ )
- when a third variable (or even more) go unnoticed and/or unmeasured, but have an effect on the relationship between  $x$  and  $y$ .
- suggest a relationship that doesn't exist between  $x$  &  $y$  or can hide one
- always plot  $y$  and residuals over time to uncover lurking variables

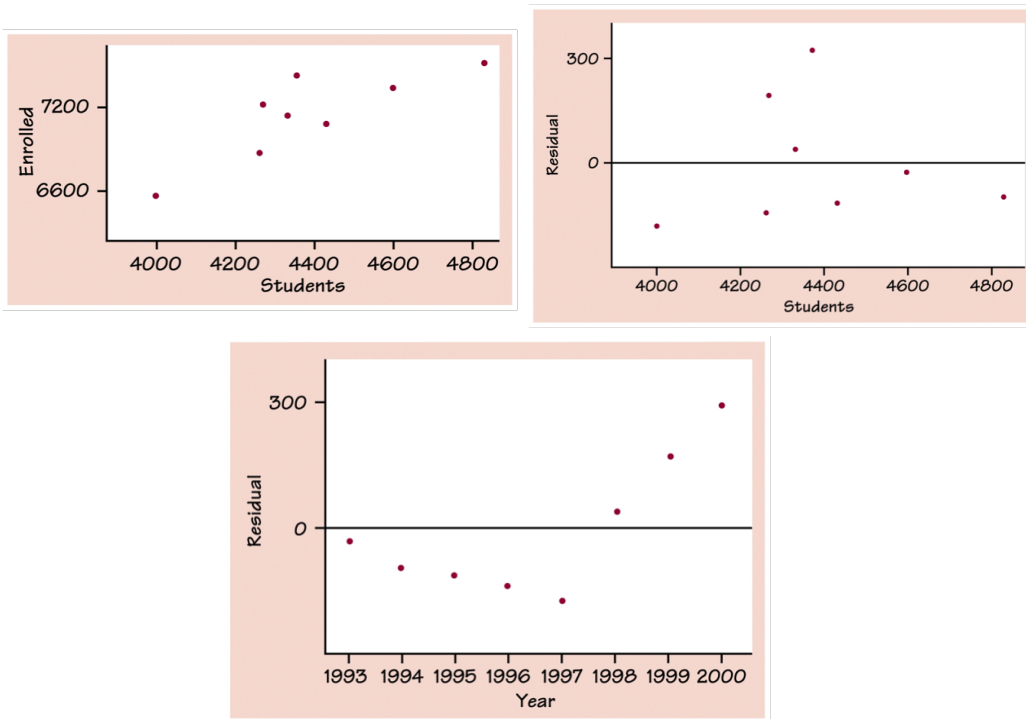
p.227/8 ex 4.10, 4.11, 4.12

### 4.11 Study of housing conditions

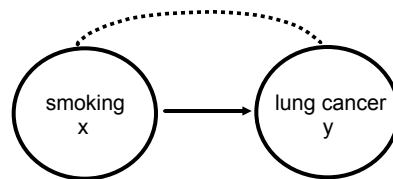
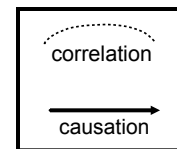
( $x$  = overcrowding,  $y$  = lack of indoor toilets,  $r = 0.08$ )



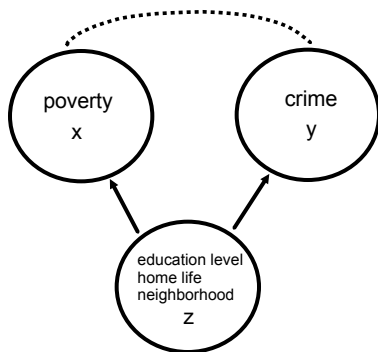
Ex 4.12 Predicting Enrollment (usefulness of plotting against time)



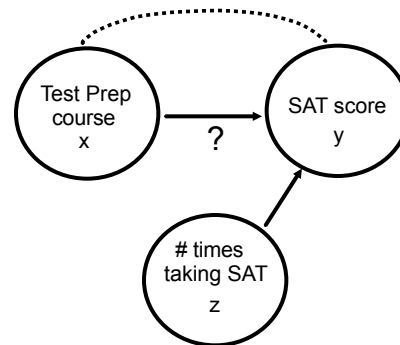
Correlation vs. Causation



**causation**  
changes in x cause changes in y



**common response**  
changes in z cause changes in both x and y, which change together



**confounding**  
changes in x and z both cause changes in y, can't tell the effects apart x and y change together

in both cases "z" is considered to be a lurking variable

RULE #1: High correlation does NOT imply causation

The best way to determine causation is by conducting an experiment with controls placed on lurking variables, but experiments are not always possible.

If an experiment is not possible, causation can be cautiously established if:

- strong and consistent association across populations and geographic locations
- the rate of change of y against x goes from positive to negative together (in place of a treatment showing effectiveness)
- the cause was in place before the effects of the response
- causation is plausible