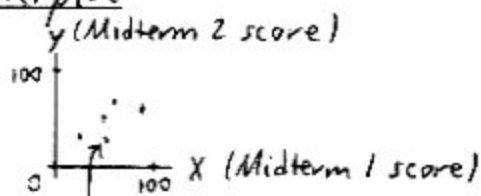


10-2 CORRELATIONPaired or bivariate dataScatterplot

$(x_i, y_i)$  for student #i  
 n pairs      n students

Karl Pearson developed:

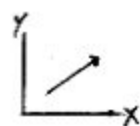
$\rho$  ("rho") = the linear correlation coefficient  
for a population

$r =$  ' sample

Properties of  $r$  (also apply to  $\rho$ ):

①  $-1 \leq r \leq 1$

② (a) If  $r > 0$ , then  $y$  tends to increase  
as  $x$  increases.



(b) If  $r < 0$ , then ' decrease



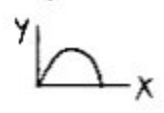
③ If  $r = 0$ , then there is no linear  
relationship between  $x$  and  $y$ .

③  $|r|$  measures the strength of the linear relationship between  $x$  and  $y$ .

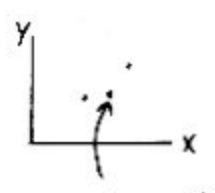
Handout

Warnings

①  $r$  might not "pick up" a strong nonlinear relationship.



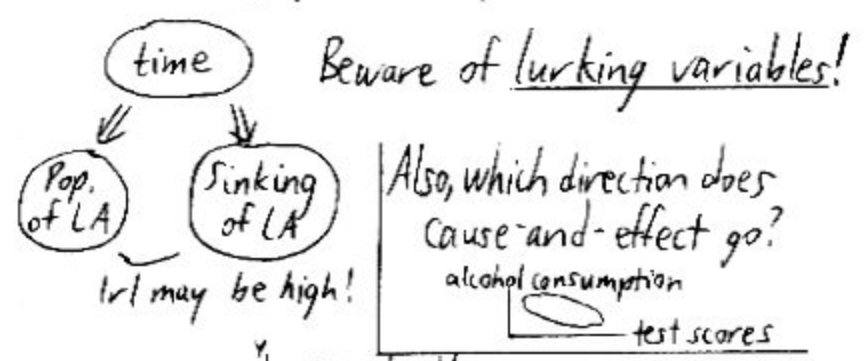
② Averaging tends to  $\nearrow |r|$ .



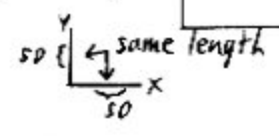
(avg. Mid 1, avg. Mid 2) for dorm # i

③ Correlation does not imply causality.

Against All odds  
 Bluman SOS-6  
 sed  
 alc  $\nearrow$  ?  
 test

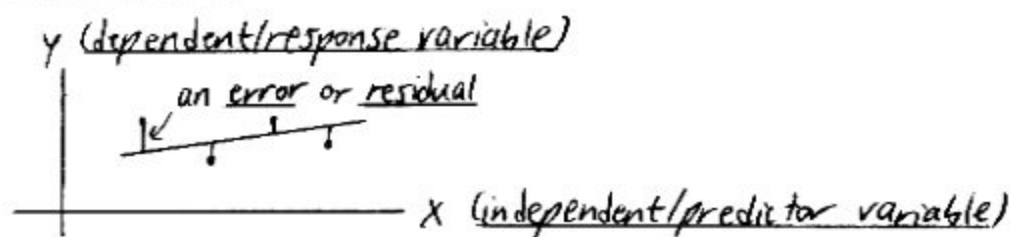


④ To avoid distortions,



Test  $H_0: \rho = 0$

(10-3) REGRESSION



What is the best linear model for a paired data set?

The least squares regression line is the line that minimizes the sum of the squares of the errors.

The higher  $|r|$  is, the more useful this line is.

"True" best line for pop. data: ~~\_\_\_\_\_~~

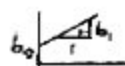
$$y = \beta_0 + \beta_1 x$$

pop. parameters

Best line for sample data:

$$\hat{y} = b_0 + b_1 x$$

y-int. slope



sample stats computed from data, formulas

CAUTION: Sensitive to outliers!

Freedman 127  
 Pts. cluster  
 around  $SO$  line  
  
 $(x, y)$   
 $SO$   
 "vs." human  $SO$

Freedman 188  
 slope =  $\frac{(r)(SO_y)}{(SO_x)}$   
 $\nearrow$   $SO$  in  $x$   
 $\nearrow$   $SO$  in  $y$   
 along line

Ex (Interpreting Slope as Marginal Change)

x = Midterm 1 scores,  
y = Mid 2 scores

$\hat{y} = 10 + 1.2x$   
↑ Predicted Mid 2 score      ↑ Mid 1 score

According to the reg. line, for every 1 point increase in the Mid 1 score, there is a 1.2 point increase in the Mid 2 score.

Ex	Mid 1 score	Predicted Mid 2 score
John	20 ↓+1	34
Jane	21 ↓+1	35.2 ↓+1.2

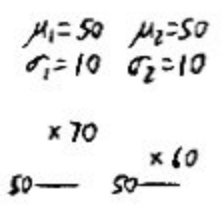
Regression to the Mean

In terms of SDs, you should expect that a person's Mid 2 score will be closer to the Mid 2 mean than his/her Mid 1 score was to the Mid 1 mean.

Usually, z-score for Mid 2 will be closer to 0.

Plous

Idea: If someone, aces Mid 1, it's probable that "errors" such as "luck" worked in his/her favor. It's probable that such errors will not be as favorable on Mid 2. ("Not as lucky.")



People who don't understand this may overestimate the value of punishment in "correcting" poor performance and unduly avoid praise for superior performance that declines. "Sports Illustrated Jinx" just reg. to the mean.

(10-4)

$r^2 =$  the coefficient of determination

$r^2$  in % form = the % of the variation in  $y$  that is explained by  $x$  and the reg. line.



$$r = 0.8 \text{ (actually, } \rho = 0.8)$$

$$\Rightarrow r^2 = 0.64$$

$\Rightarrow$  64% of the variation in the Mid 2 scores can be explained by the Mid 1 scores via the reg. line.