24 ① May

AM 5 5

this time: correlation, regression

next time:

read: FPP ch 10, 11

(hwk 5 due next (this thu) time)

Fri 27 May only: Ethan's office hr will be 11-noon

VIII. Correlation, regression, & prediction

big picture: p. 22 VIII A

we've talked about how to describe & do inference on 1 variable at a time; what about 2 at a time? ex. CS 17 (hts of F & S)

son	father
Y	X
+ 70 in	72 in +
- 67	66 -
⋮	⋮

n = 1078

1 row for each family

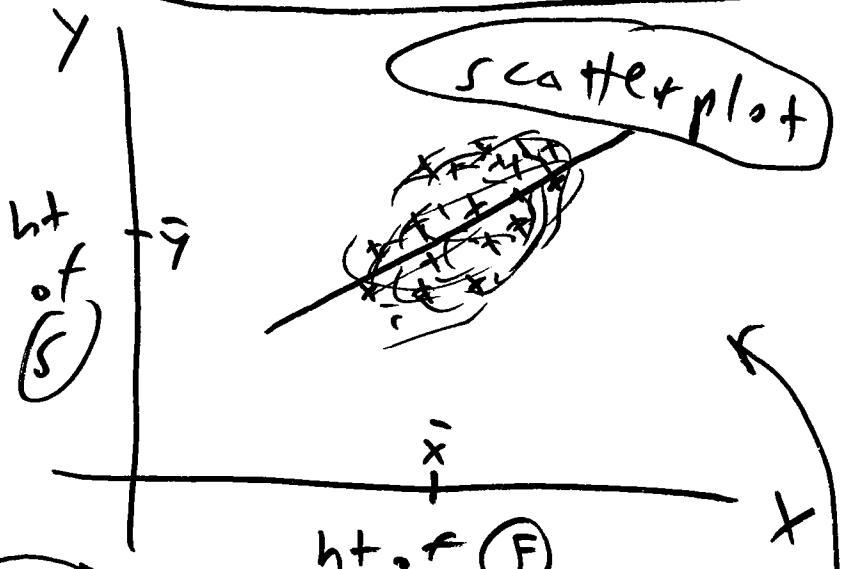
2 quant. cont.

ht of F, ht of S: both

mean	$\bar{y} = 69$ in	$\bar{x} = 68$ in
SD	$s_y = 2.7$ in	$s_x = 2.7$ in

hist of Y

hist of X



scatterplot

basic graphical descriptive summary of 2 var. at a time

Ams 5

secular trend in ht: (S) about 1 in taller than (F)

Case Study 17 (genetics): Human Inheritance of Height (nutrition)

How is your height related to that of your parents? Two eminent British scientists, Francis Galton and Karl Pearson, investigated this question in the late 1800s. They gathered data on the adult heights of 1,078 fathers and sons (one son per family). The picture below graphically summarizes this dataset; here are some numerical summaries:

variable	mean	SD	
father's height	68 inches	2.7 inches	n = 1078
son's height	69 inches	2.7 inches	

If you were going to draw a smooth curve through the data that captured the basic trend of how sons' heights depend on the heights of their fathers, a straight line would probably do pretty well. The plot below gives two natural lines you might try to fit to the data, one called the *SD line*, the other the *regression line*. We will talk next time about which is which, and which one is better. The question for today is: How strong is the linear relationship between these two variables?

(high) (bivariate normal)

(marginal hist of y)

(normal)

(dependent var)

(outcome var)

(low)

(oval / elliptical scatterplot)

(football-shaped)

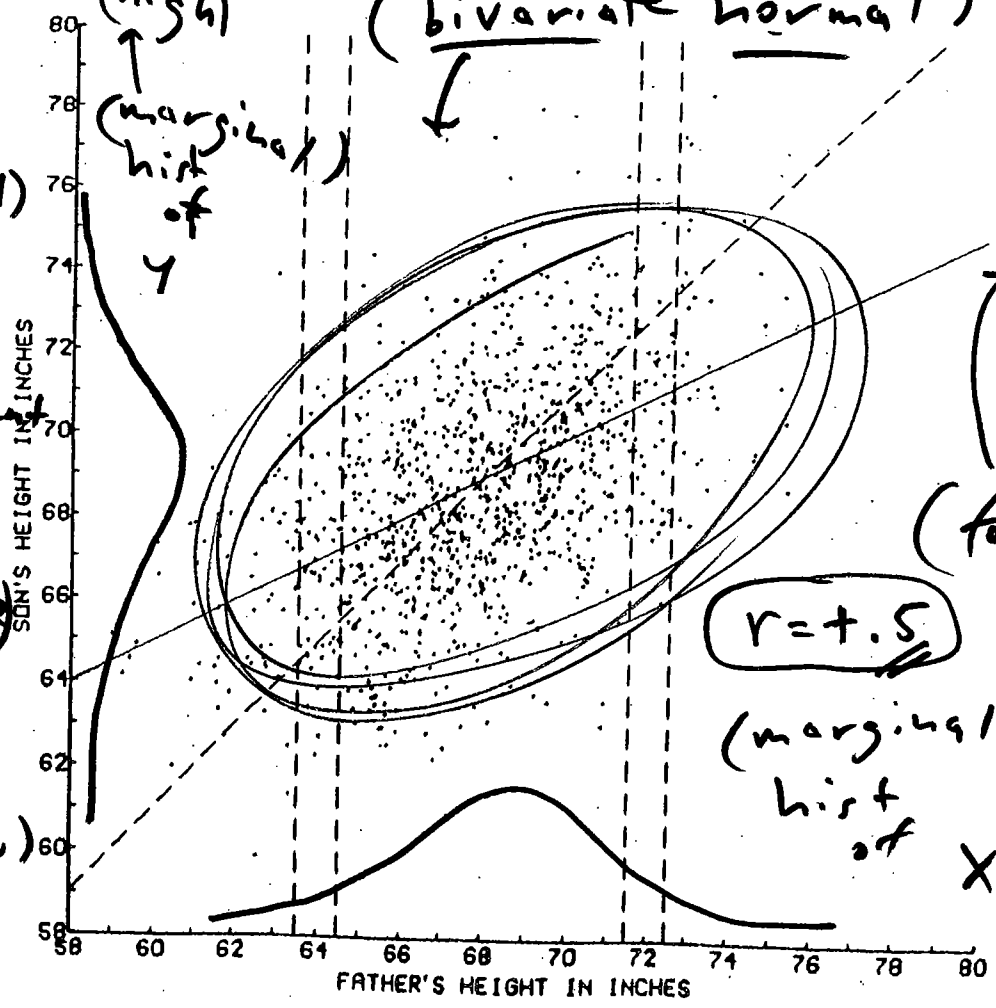
r=+.5

(marginal) hist of x

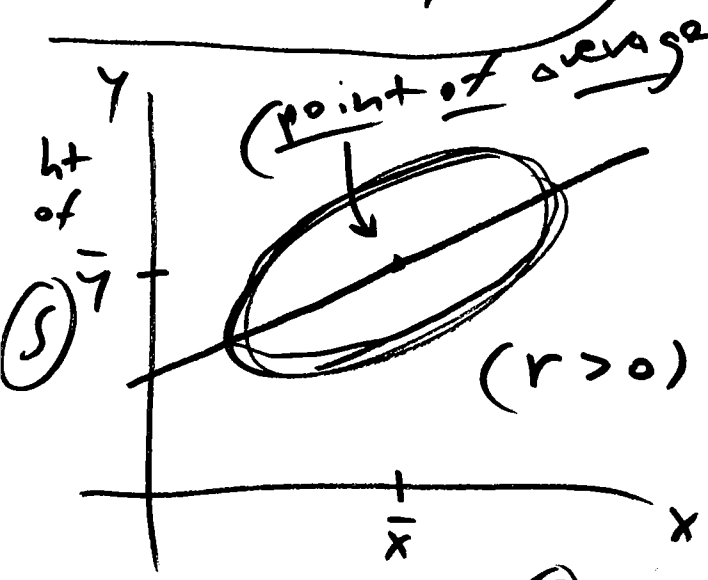
(predictor var)

(normal)

(independent var)

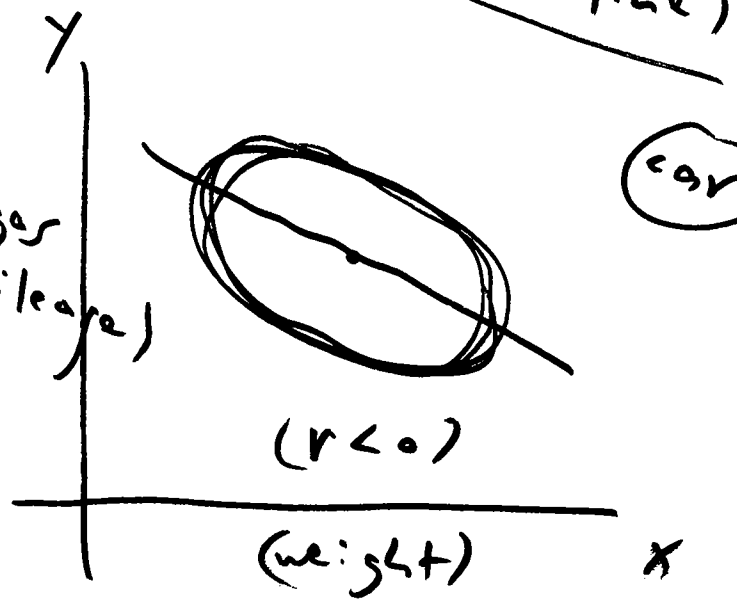


Q: How strongly related (in a linear way) are x & y? ← (straight line)



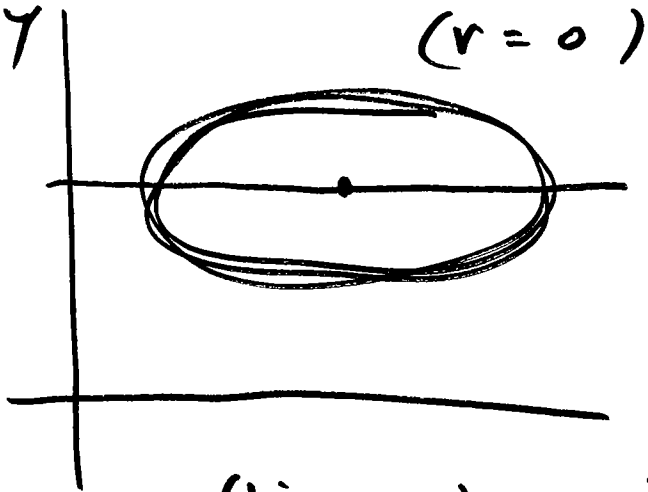
slope positive

x, y positively associated



slope negative

x, y negatively associated



no (linear) association

r = correlation (coefficient)