

Correlation, Regression, and Prediction

Case Study 17:
Human Inheritance
of height:
Fathers (F) and
Sons (S)

• We've talked about how to describe & do inference on 1 variable at a time;
- What about 2 at a time?

(S)	(F)
Y	X
70"	72"
69"	66"
⋮	

← 1 row for each family
n = 1078

• The heights of F and S are both quantitative continuous variables
• The \bar{x} and \bar{y} show a secular trend:
- Sons are about 1" taller than fathers on average due to better nutrition.

mean of F = $\bar{x} = 68"$

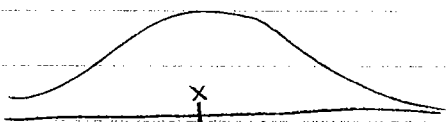
mean of S = $\bar{y} = 69"$

SD of F = $s_x = 2.7"$

SD of S = $s_y = 2.7"$

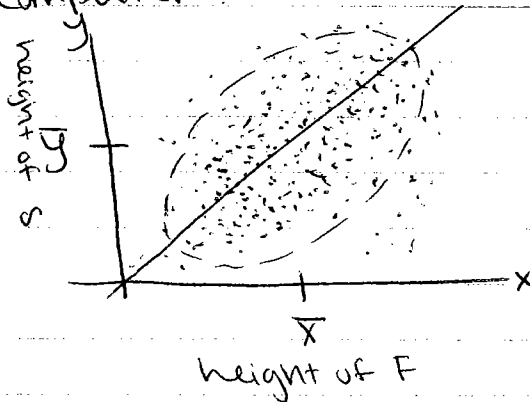
hist. of F: normal

hist. of S: normal



How can we see if there is another trend in the data?

- Scatter Plot: plot x vs y; usually done by computer



- See pg 103 for actual plot

A scatter plot is the basic graphical descriptive summary for 2 variables at a time.

• this is an elliptical shaped (aka: oval or 'football shaped') scatter plot.

- you get an elliptical scatter plot if the histogram of x is normal shaped and if the histogram of y is normal shaped.

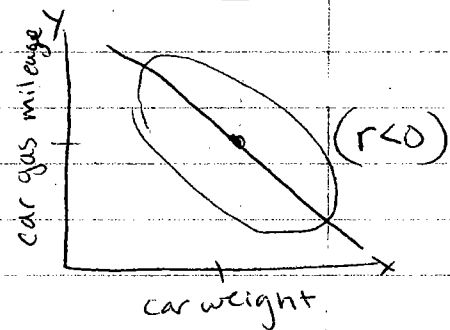
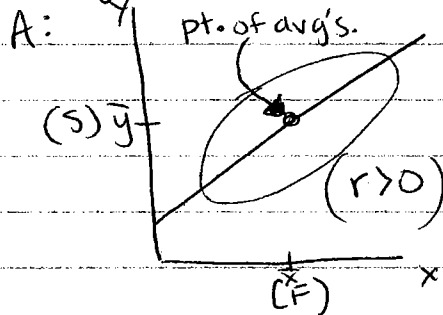
- describes a "bivalent normal" distribution.

Q: How strongly related are x and y ?
(in a linear way)

- the best summary for this plot is a straight line.

° x : normal independent variable: predictor variable

° y : normal dependent variable: outcome variable

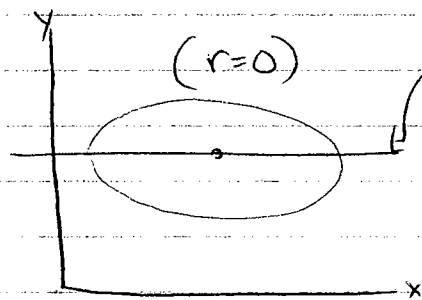


° slope is positive

° slope is negative

° x, y positively associated

° x, y negatively associated



No (linear) association between 2 variables

r is...

- the correlation coefficient

- not the slope of the line.

- 2 plots can have the same r value when the slopes are different.

Defining r :

° There are 4 quadrants in a scatter plot:



- when $r = 0$, all 4 quadrants are the same size

- the top right and bottom left are bigger in this plot (have more data)

