Describing Relationship between Two Quantities

Covariance and Correlation

Covariance

- So far, we have been focusing analyzing properties describing aspects of a single list of numbers
- Frequently, however, we are interested in how multiple lists of numbers behave together
- The question is: Do the two lists of numbers change (or co-vary) together? To which extent they co-vary?

Covariance

- Remember that variance is defined as:

$$Var_X = \frac{\Sigma(X - \overline{X})^2}{N-1} = \frac{\Sigma(X - \overline{X})(X - \overline{X})}{N-1}$$

- Following intuition, the co-variance would be:

$$Cov_{XY} = \frac{\Sigma(X - \overline{X})(Y - \overline{Y})}{N - 1}$$

- How this works, and why?
- When would cov_{XY} be large and positive? Large and negative?

Covariance

$$\mathrm{cov}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

When X increases and Y increases: cov(x,y)= positive

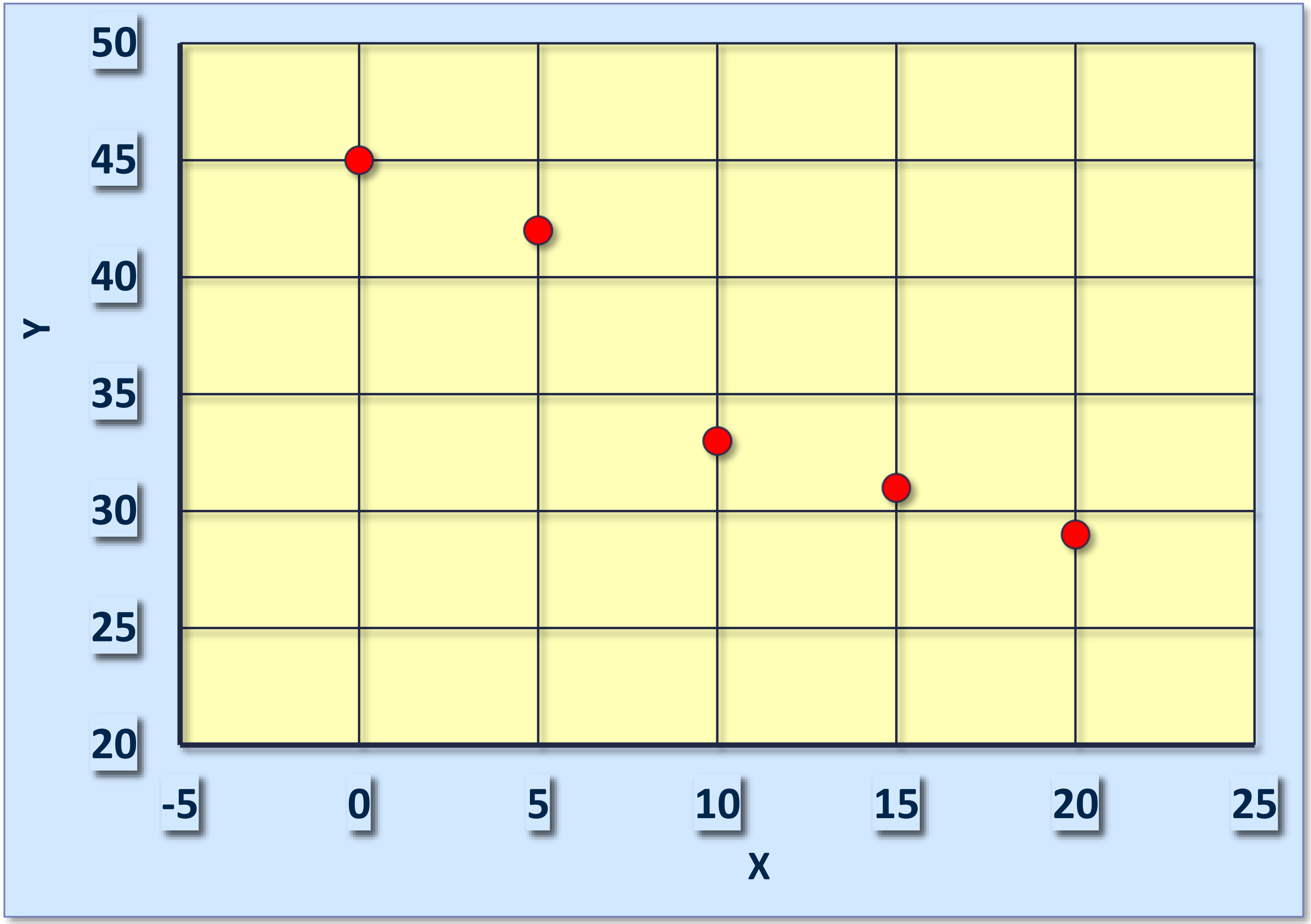
When X increases and Y decreases: cov (x,y)= negative

When no constant relationship: cov (x,y) = 0

A Simple Example

N	X	Y
1	0	45
2	5	42
3	10	33
4	15	31
5	20	29

Scatter Plot



Calculating Covariance

(X)	(Y)
0	45
5	42
10	33
15	31
20	29
$\bar{X} = 10$	$\bar{Y} = 36$

Covariance

- Variables that ***covary*** inversely, tend to appear on opposite sides of the group means
- Variables that ***covary*** simultaneously, tend to appear on the samesides of the group means

- Average product of deviation measures the degree to which the variables covary, i.e. the degree of linkage between them or covariance

Calculating Covariance

(X)	$(X-\bar{X})$	$(X-\bar{X})(Y-\bar{Y})$	$(Y-\bar{Y})$	(Y)
0	-10	-90	9	45
5	-5	-30	6	42
10	0	0	-3	33
15	5	-25	-5	31
20	10	-70	-7	29
		$\sum$= -215		

$$S_{xy} = \frac{1}{4}(-215) = -53.75$$

Computational Formula: Covariance

- The computational formula for covariance is similar to the one for variance. Indeed, the latter is a special case of the former, since variance of a variable is "its covariance with itself."

$$s_{xy} = \frac{1}{N-1}\left(\sum_{i=1}^{N} X_i Y_i - \frac{\sum_{i=1}^{N} X_i \sum_{i=1}^{N} Y_i}{N} \right)$$

Problems with Covariance

- The value obtained by covariance is dependent on the magnitude of the data's standard deviations
- If standard deviation is large, the value will be greater than if small…

even if the relationship between x and y is exactly the same in the large versus small standard deviation datasets

Developing a measure of “co-relation”:

“Problems” to address:

(1) Neither variable is an “outcome” or a “predictor”

(2) The measure of correlation should be dimensionless, (eg., applicable for inches or feet, water level or velocities)

Pearson’s solution: Re-express (transform) both variables on new “standard” scales that essentially eliminate the particular metrics of the original scales

Assortative Mating. Based on 1000 *to* 1050 *Cases of Husband and Wife.*

	Husband's Character	Wife's Character	Correlation and Probable Error	Symbol
Direct	Stature	Stature	·2804 ± ·0189	r_{12}
	Span	Span	·1989 ± ·0204	r_{34}
	Forearm	Forearm	·1977 ± ·0205	r_{56}
Cross	Stature	Span	·1820 ± ·0201	r_{14}
	Stature	Forearm	·1403 ± ·0204	r_{16}
	Span	Stature	·2023 ± ·0199	r_{32}
	Span	Forearm	·1533 ± ·0203	r_{36}
	Forearm	Stature	·1784 ± ·0201	r_{52}
	Forearm	Span	·1545 ± ·0203	r_{54}

Heredity: relationships between siblings and spouses

(Pearson & Lee, 1903, On the laws of inheritance in man, *Biometrika*)

Solution: Pearson's r

- Covariance alone does not really tell us much

» *Solution: standardise this measure*

- Pearson's r: standardises the covariance value.
- Divides the covariance by the multiplied standard deviations of X and Y:

$$r_{xy} = \frac{\operatorname{cov}(x,y)}{s_x s_y}$$

Computational Formulae

$$s_{xy} = \frac{1}{N-1}\left(\sum_{i=1}^{N} X_i Y_i - \frac{\sum_{i=1}^{N} X_i \sum_{i=1}^{N} Y_i}{N} \right)$$

$$r_{xy} = \frac{N\sum XY - \sum X \sum Y}{\sqrt{\left(N\sum X^2 - \left(\sum X\right)^2\right)\left(N\sum Y^2 - \left(\sum Y\right)^2\right)}}$$

Example Calculation of r_{xy}

	(X)	X^2	XY	Y^2	(Y)
	0	0	0	2025	45
	5	25	210	1764	42
	10	100	330	1089	33
	15	225	465	961	31
	20	400	580	841	29
$\sum=$	50	750	1585	6680	180

Computing r_{xy} from Table

$$r_{xy} = \frac{5(1585) \;\; - 50(180)}{\sqrt{\left(5(750 - 50^2)\right)\left(5(6680) - 180^2\right)}}$$

$$= \frac{7925 - 9000}{\sqrt{\left(3750 - 2500\right)\left(33400 - 32400\right)}}$$

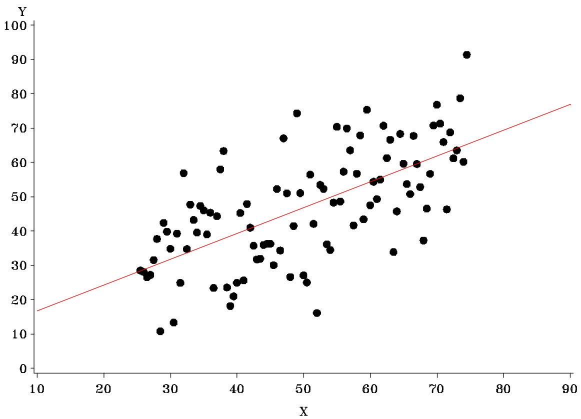
Computing Correlation

$$r_{xy} = \frac{-1075}{\sqrt{(1250)(1000)}}$$

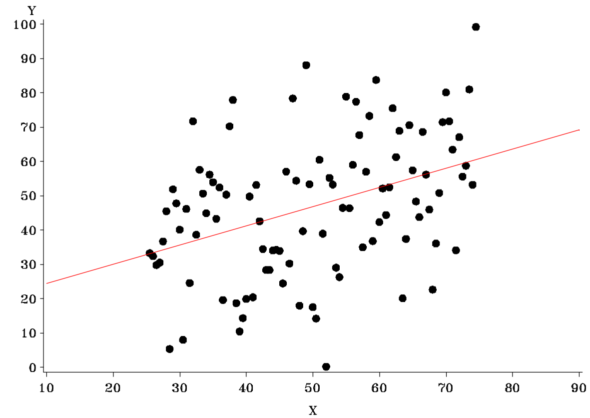
$$\boxed{r_{xy} = -0.9615}$$

Plots to help develop your intuition for interpreting r

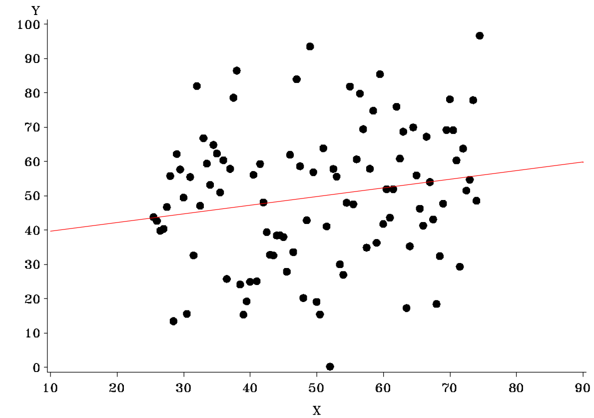
r=.65



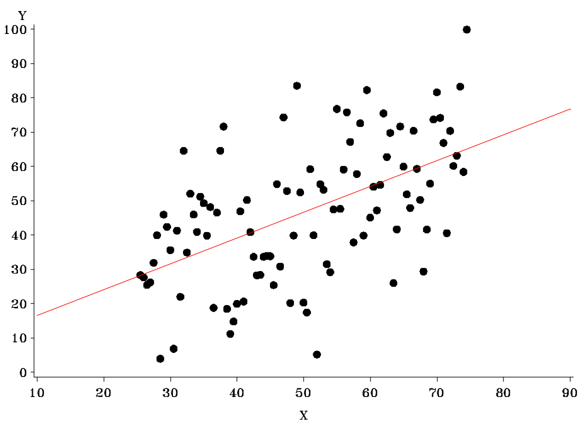
r=.39



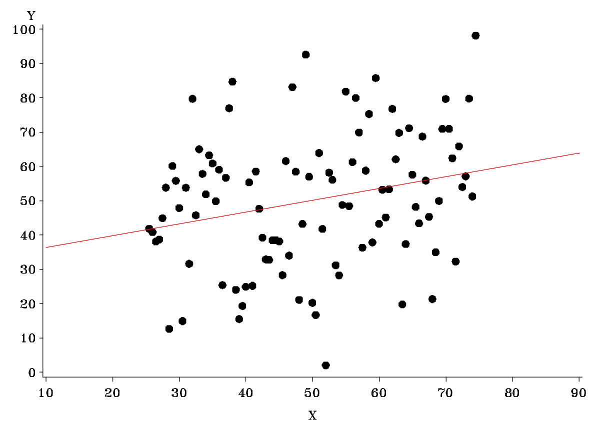
r=.18



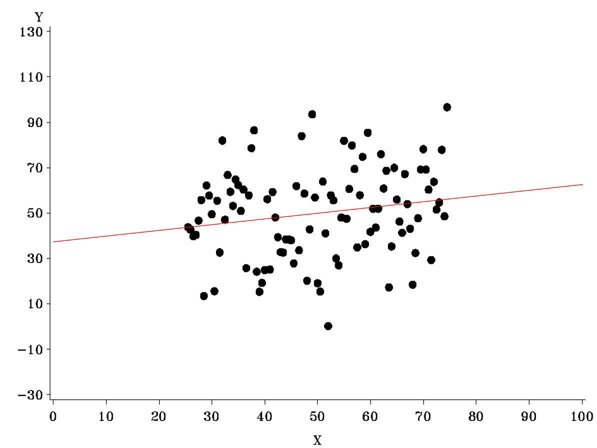
r=.54



r=.25



r=.18−−longer axes



The Correlation Coefficient

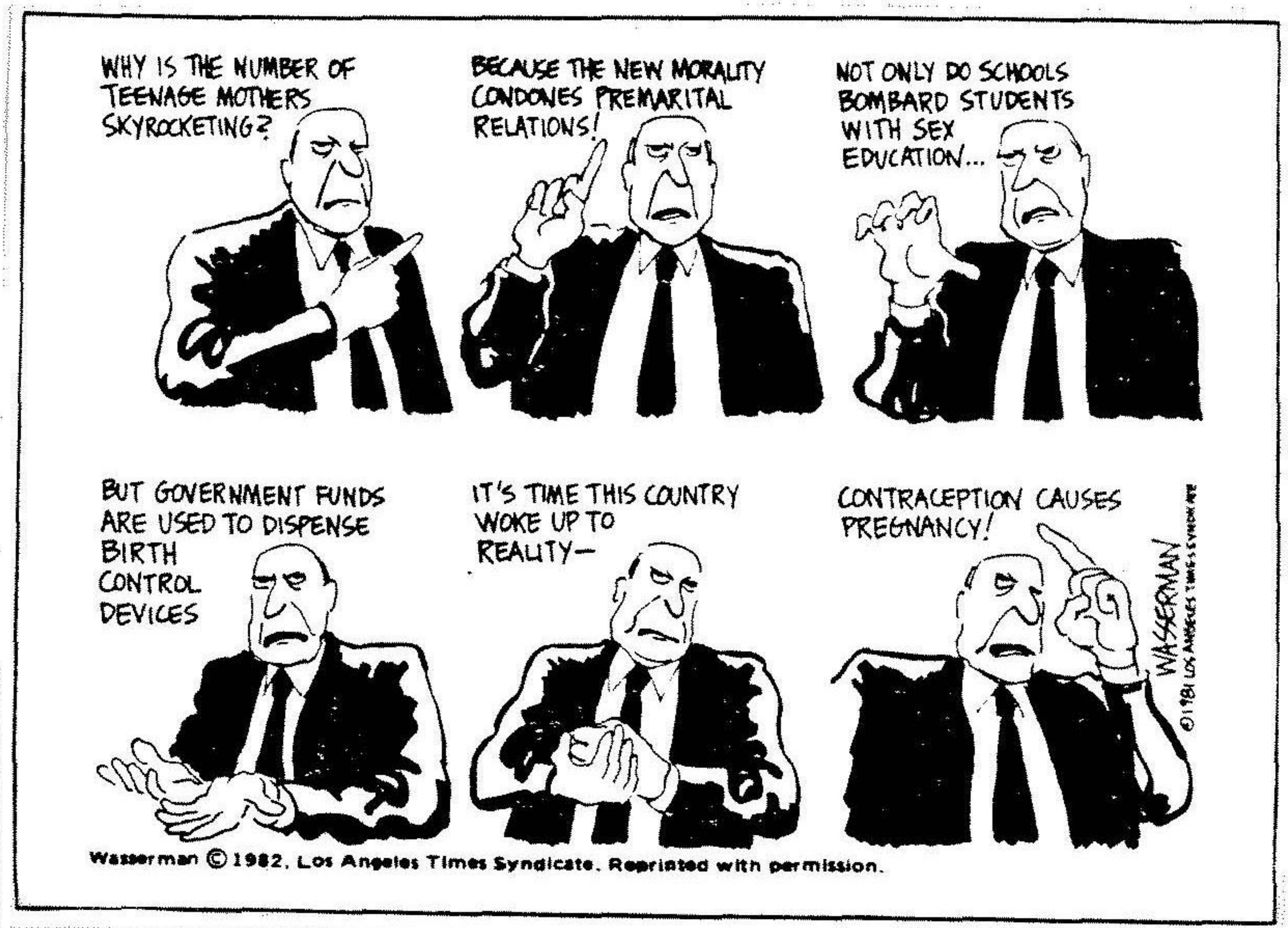
- An association between two variable can be *stronger* or *weaker*.
- Remember: a strong association means that knowing one variable helps to predict the other variable to a large extend.
- The *correlation coefficient* is a numerical value expressing the strength of the association.

Big issues to be aware of:

1. **Correlation does not imply causation.** For example, there is a strong correlation between golf scores and salaries for CEOs. This does not imply that one can improve their salary by getting better at golf. Often times there are **lurking variables**, which is something that affects both variables being studied, but is not included in the study.

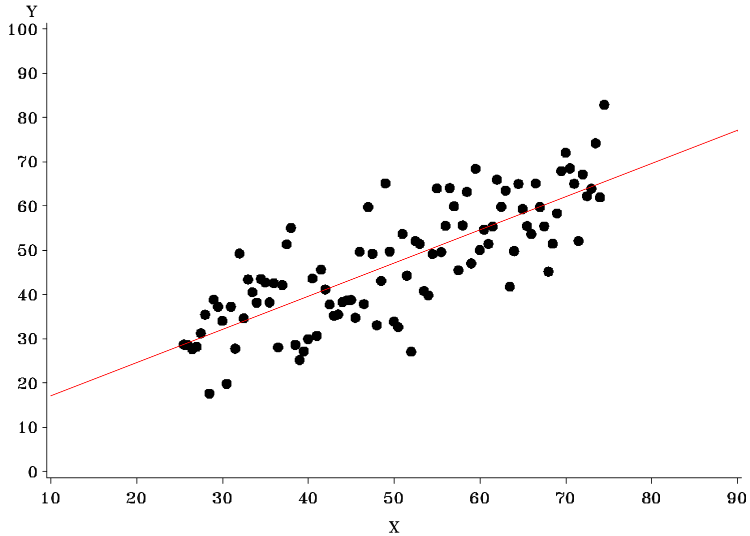
2. **Beware data based on averages**. Averages suppress individual variation, and can artificially inflate the correlation coefficient.

3. **Look out for non-linear relationships**. Just because there is no linear correlation does not mean that the variables might not be related in another way.

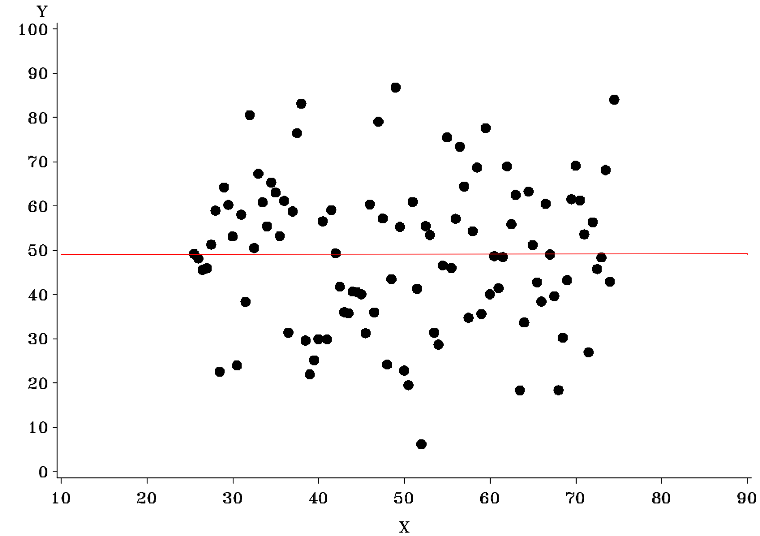

WHY IS THE NUMBER OF TEENAGE MOTHERS SKYROCKETING?
BECAUSE THE NEW MORALITY CONDONES PREMARITAL RELATIONS!
NOT ONLY DO SCHOOLS BOMBARD STUDENTS WITH SEX EDUCATION...
BUT GOVERNMENT FUNDS ARE USED TO DISPENSE BIRTH CONTROL DEVICES
IT'S TIME THIS COUNTRY WOKE UP TO REALITY—
CONTRACEPTION CAUSES PREGNANCY!
WASSERMAN

How do we interpret r?

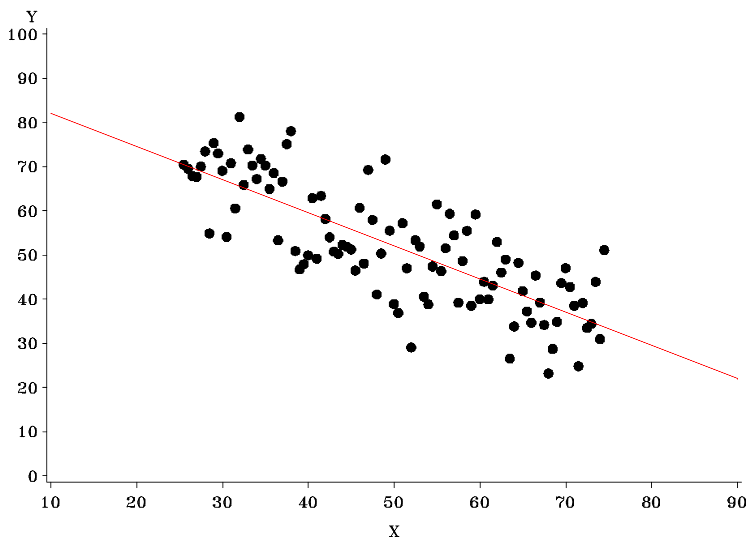
Positive correlation (r=+.79)



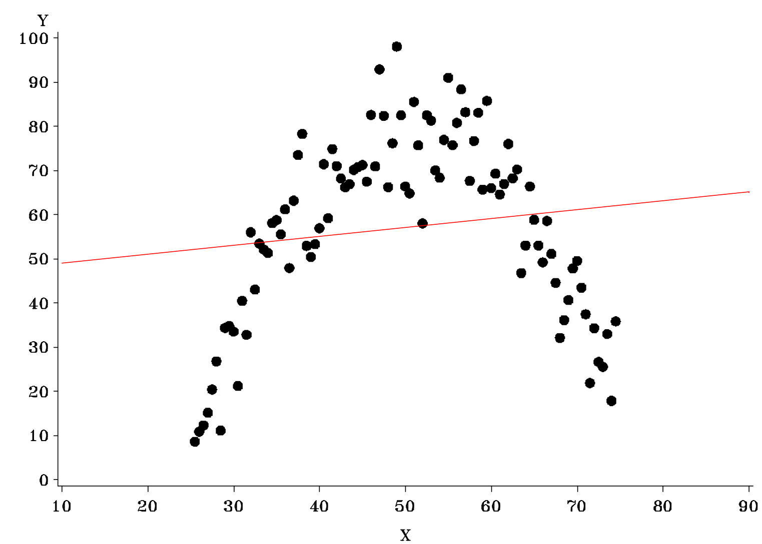
No correlation (r=.00)



Negative correlation (r=−.79)



(Almost) no correlation (r=.14)



Regression

- Correlation tells you if there is an association between x and y but it doesn’t describe the relationship or allow you to predict one variable from the other.

- To do this we need REGRESSION!