Relationship Between Interval and/or Ratio Variables: Correlation & Regression

Sorana D. BOLBOACĂ

OUTLINE

- Correlation
  - Definition
  - Deviation Score Formula, Z score formula
  - Hypothesis Test
- Regression
  - Intercept and Slope
  - Un-standardized Regression Line
  - Standardized Regression Line
  - Hypothesis Tests

Correlation: 3 Characteristics

1. Direction

- Positive(+)
- Negative (-)

2. Degree of association

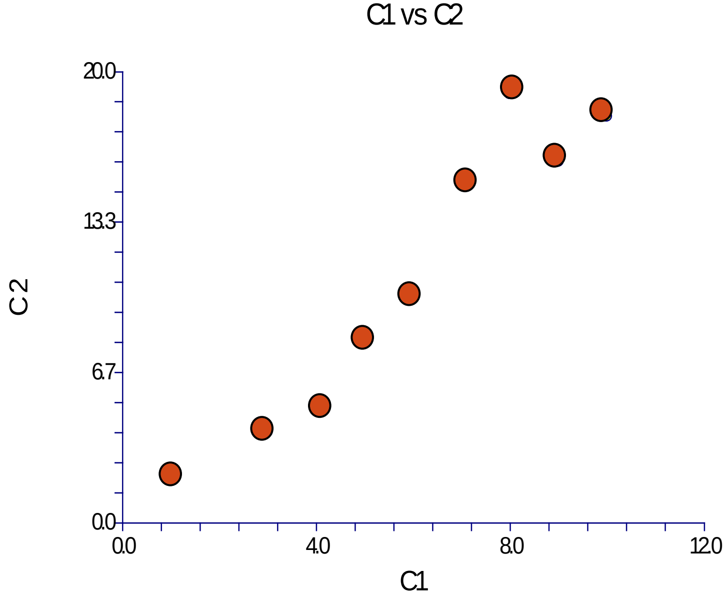
- Between –1 and 1
- Absolute values signify strength

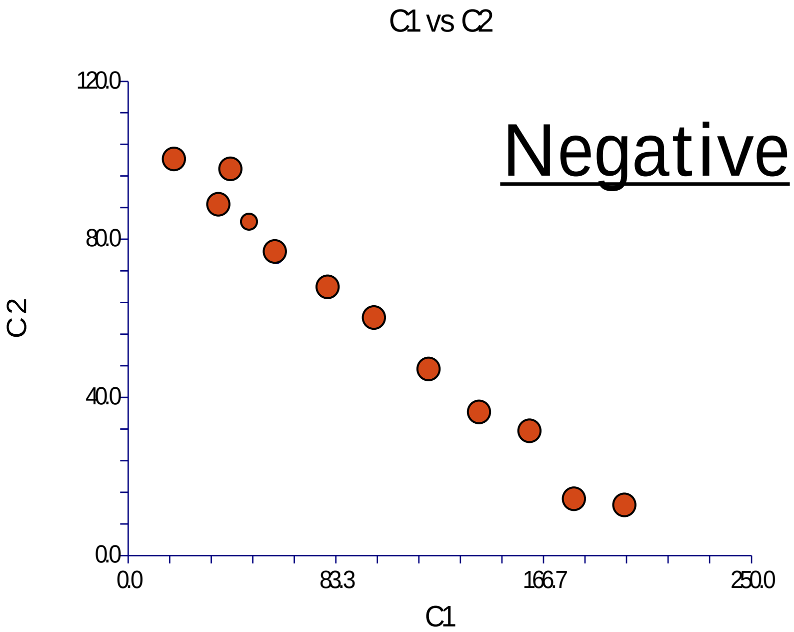
3. Form

- Linear
- Non-linear

Correlation: 1. Direction

<u>Positive</u>





<u>Negative</u>

Large values of X = large values of Y
Small values of X = small values of Y

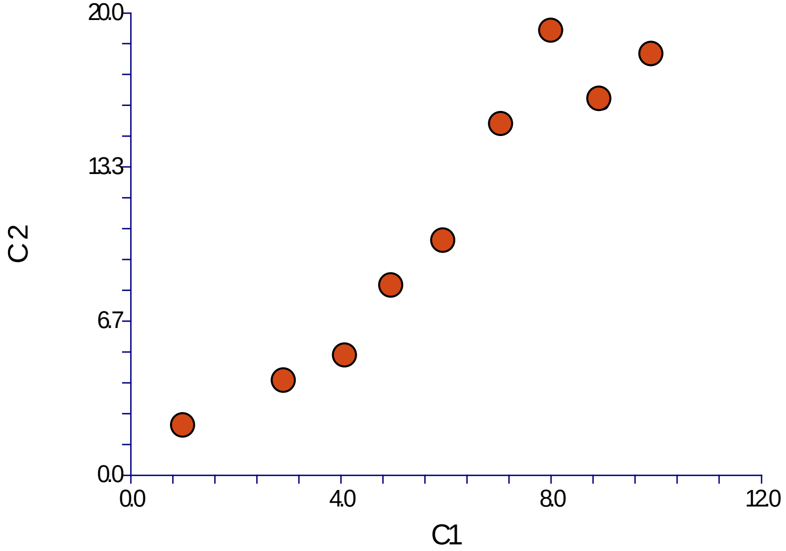
e.g. IQ and SAT

Large values of X = small values of Y
Small values of X = large values of Y

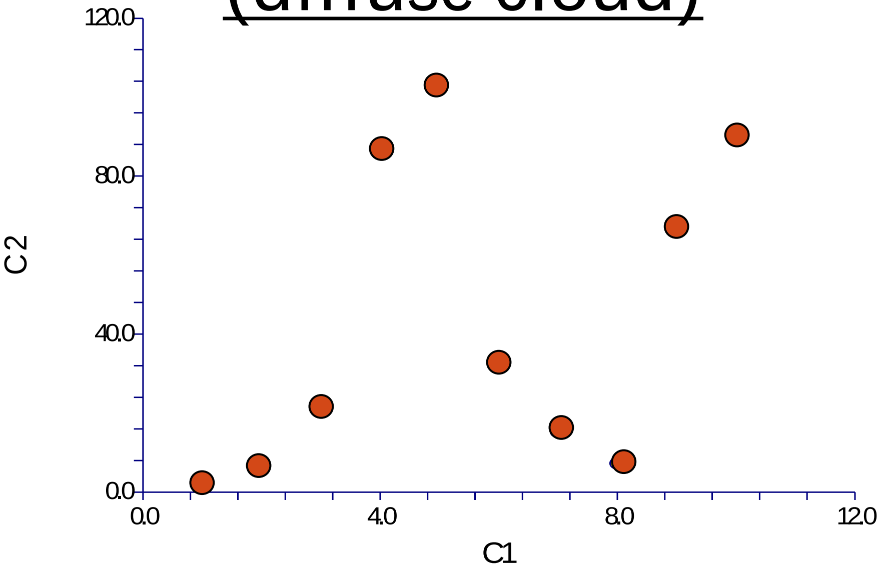
-e.g. SPEED and ACCURACY

Correlation: 2. Degree of association

<u>Strong</u>
<u>(tight cloud)</u>

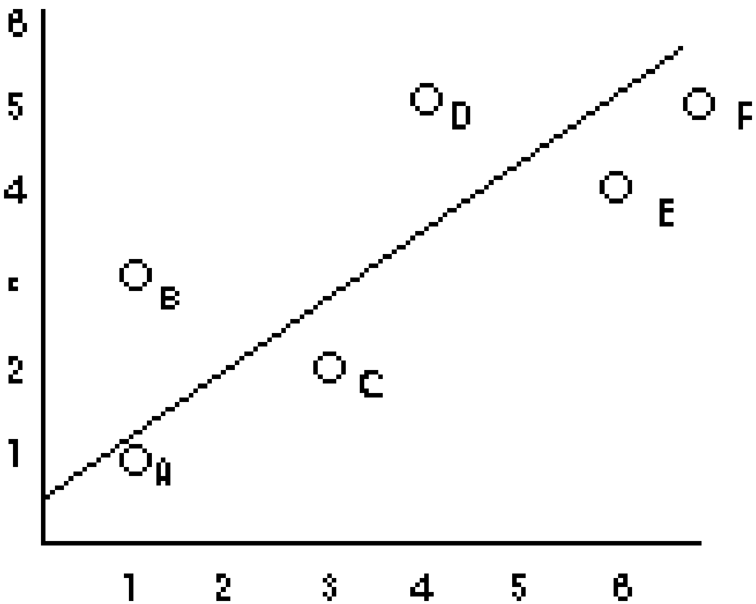


<u>Weak</u>
<u>(diffuse cloud)</u>

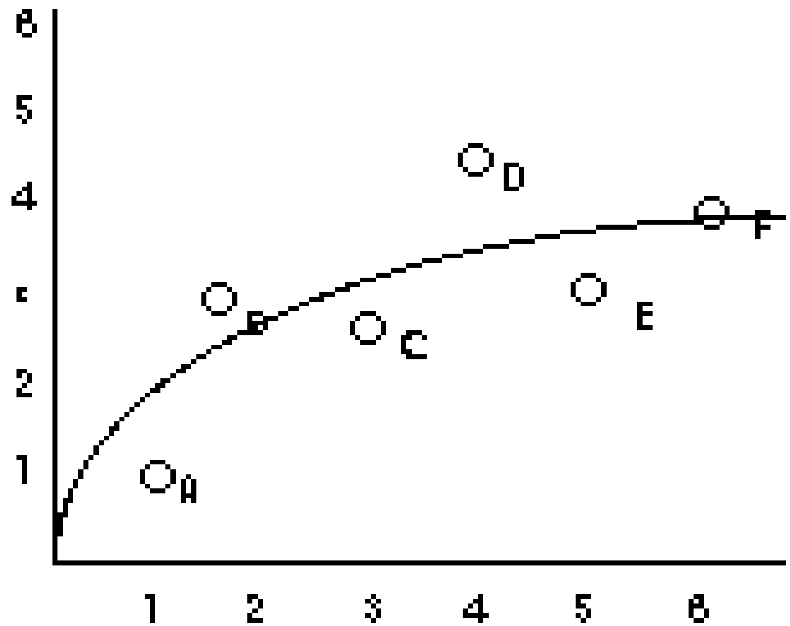


Correlation: 3. Form

Linear



Non- linear

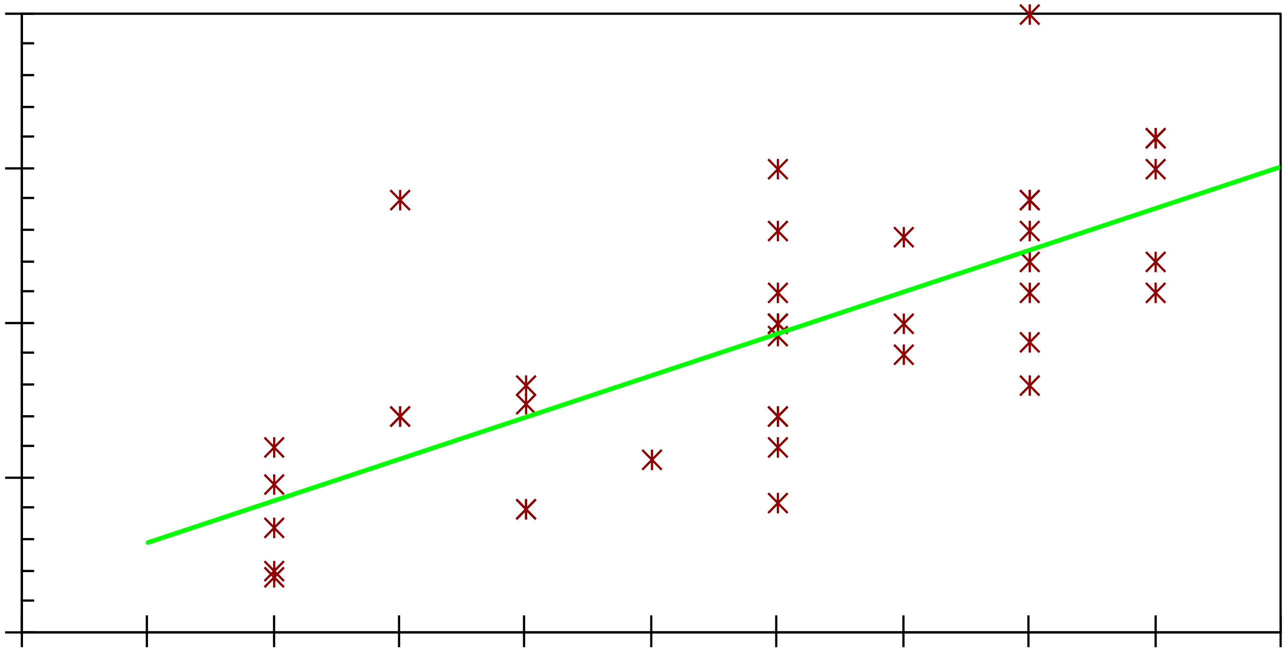


What is the best fitting straight line?

Regression Equation: $Y = a + bX$

How closely are the points clustered around the line?

Pearson's R



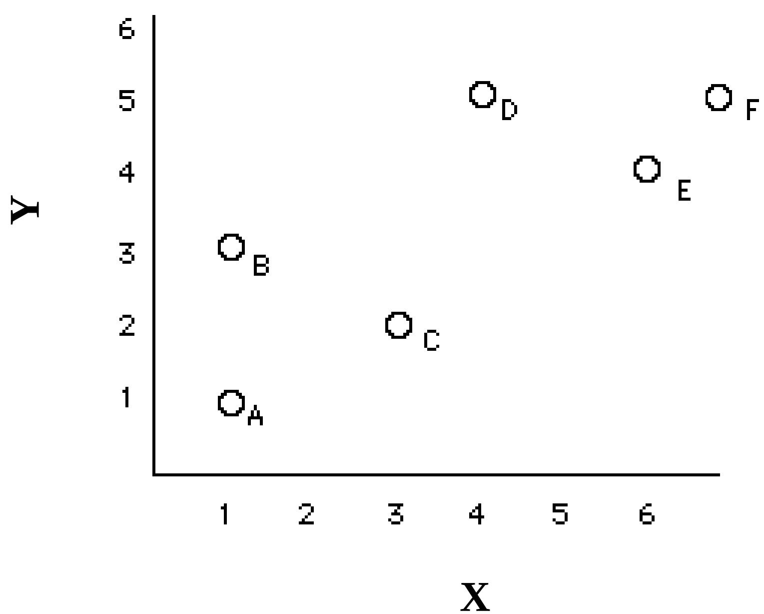
Correlation: Definition

Correlation: a statistical technique that measures and describes the degree of linear relationship between two variables

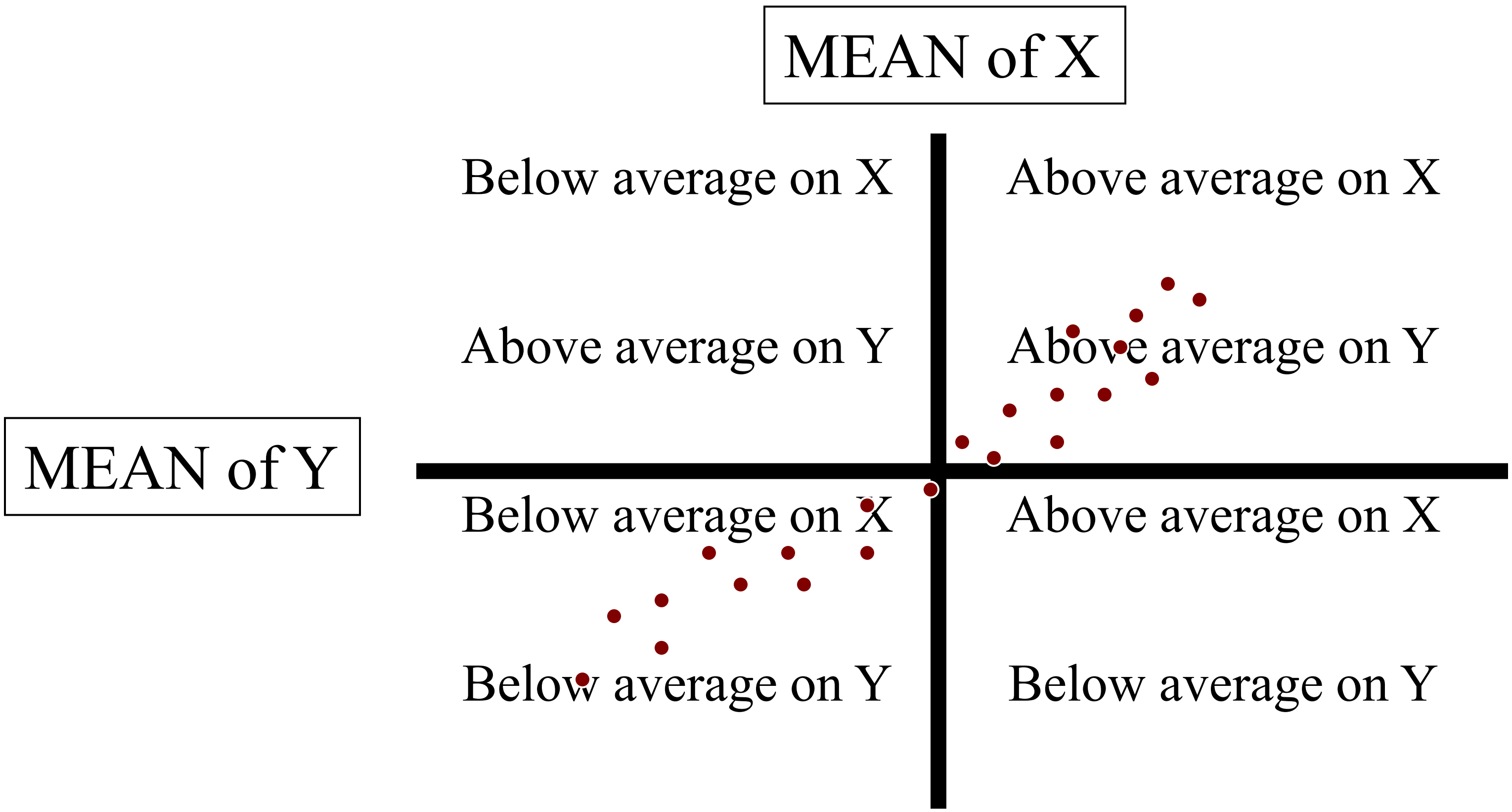
Dataset

Obs	X	Y
A	1	1
B	1	3
C	3	2
D	4	5
E	6	4
F	7	5

Scatterplot



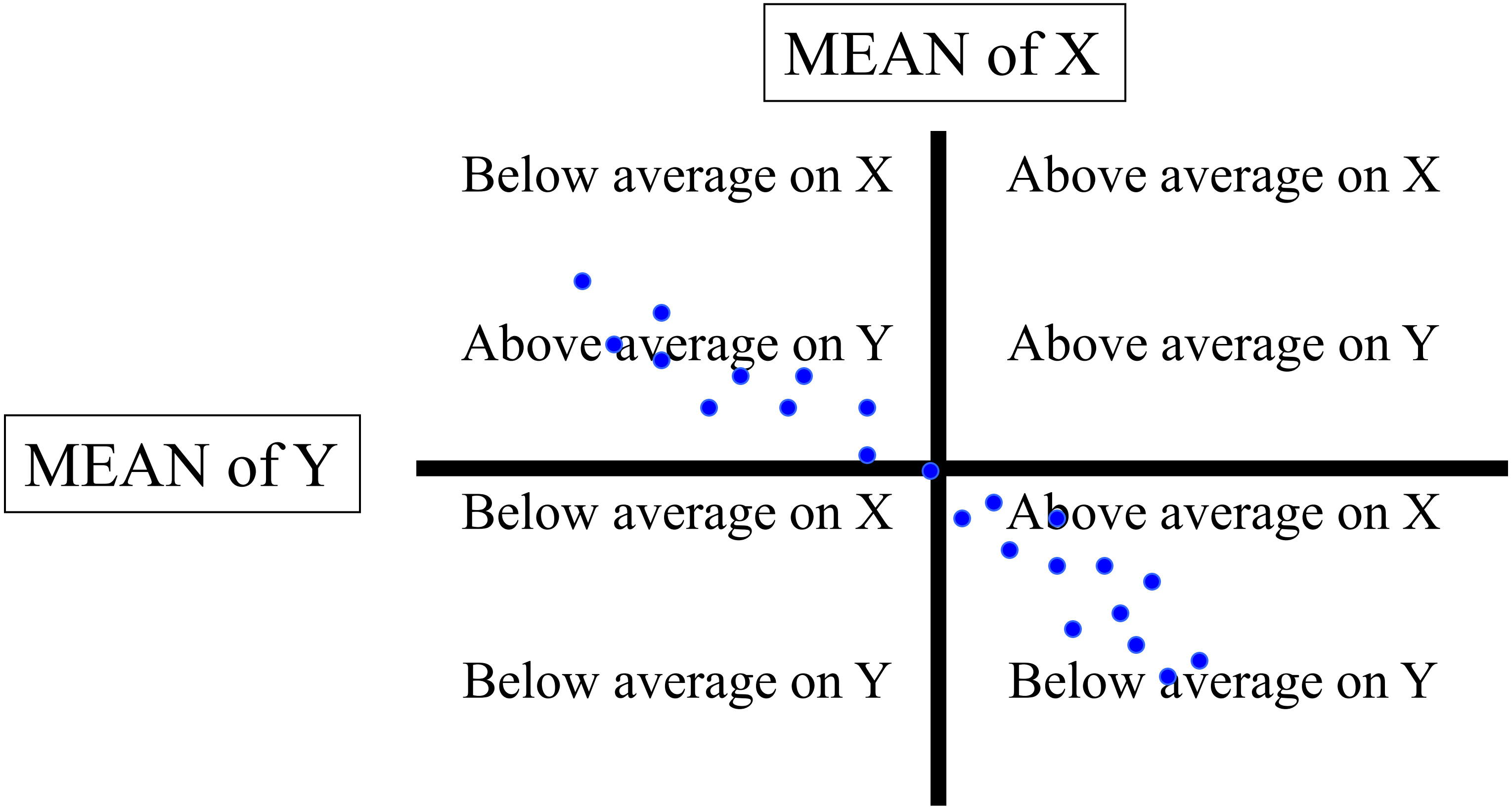
The logic of regression



Cross-Product = $(X - \overline{X})(Y - \overline{Y})$

For a strong positive association, the cross-products will mostly be positive

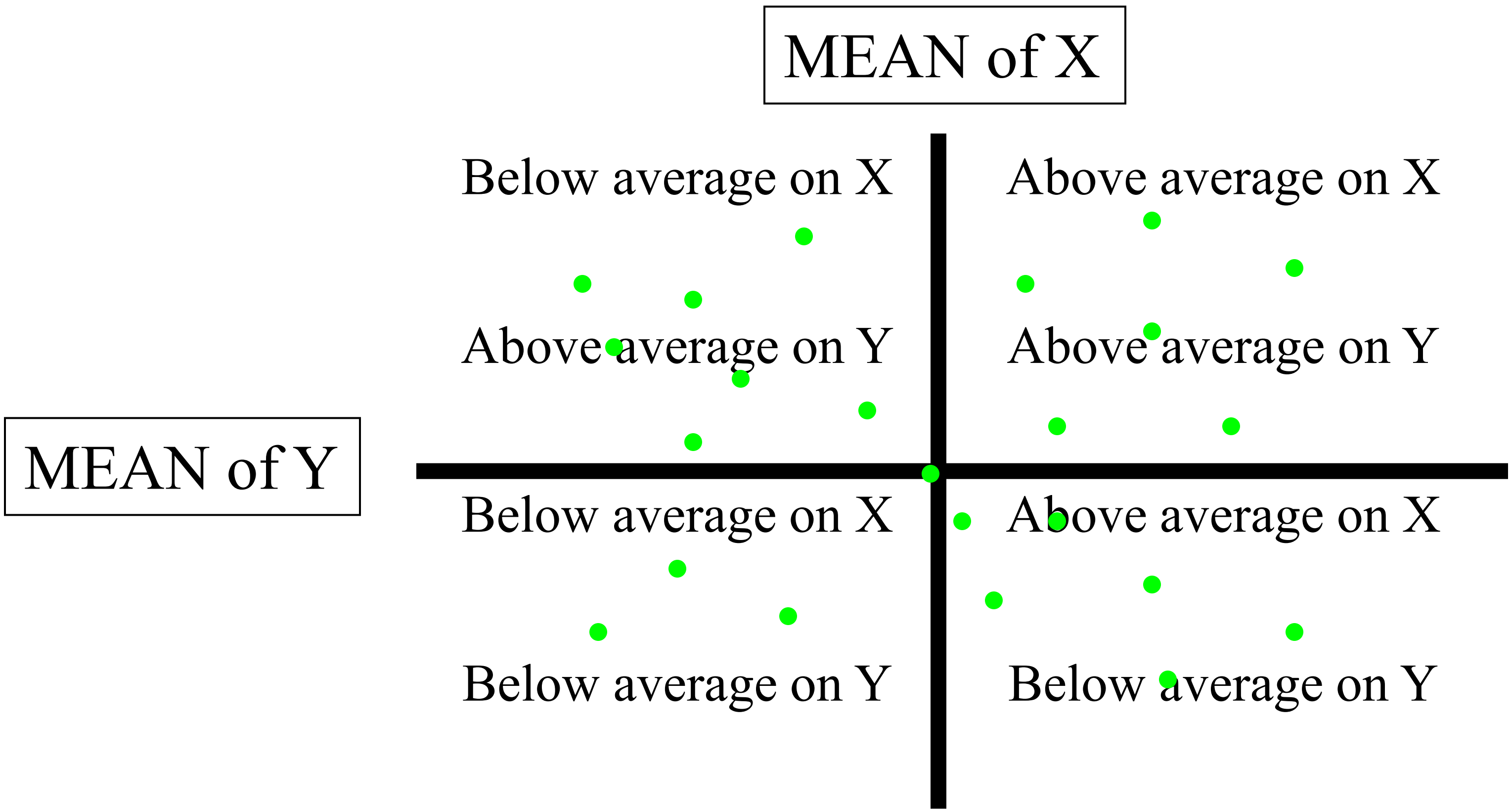
The logic of regression



Cross-Product = $(X - \bar{X})(Y - \bar{Y})$

For a strong negative association, the cross-products will mostly be negative

The logic of regression



Cross-Product = $(X - \bar{X})(Y - \bar{Y})$

For a weak association, the cross-products will be mixed

Pearson Correlation Coefficient

Symbol: r, R

A value ranging from -1.00 to 1.00 indicating the strength (look to the number of correlation coefficient) and direction (look to the sign of the correlation coefficient) of the linear relationship.

- Absolute value indicates strength
- +/- indicates direction

Sum of products

$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}}$$

Pearson Correlation Coefficient

- Assumptions:

1. The errors in data values are independent from one another
2. Correlation always requires the assumption of a straight-line relationship
3. The variables are assumed to follow a bivariate normal distribution

Bivariate Normal Distribution

- http://www.aos.wisc.edu/~dvimont/aos575/Handouts/bivariate_notes.pdf

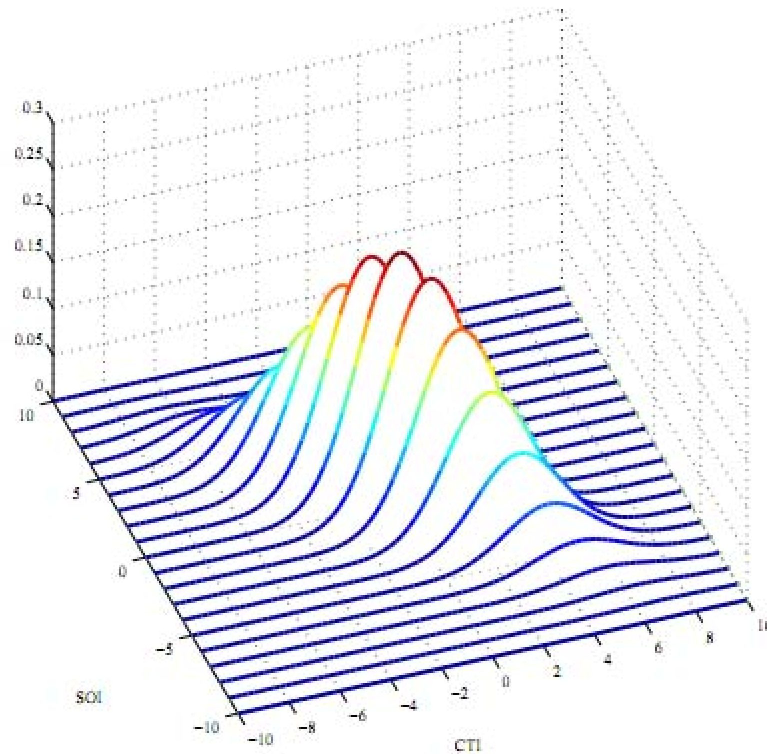


Figure 1: Bivariate Normal PDF calculated for parameters based on the Cold Tongue Index (x axis) and the Southern Oscillation Index (y-axis).

Pearson Correlation Coefficient

	Femur	Humerus	$(X-\bar{X})$	$(Y-\bar{Y})$	$(X-\bar{X})^2$	$(Y-\bar{Y})^2$	$(X-\bar{X})(Y-\bar{Y})$
A	**38**	**41**					
B	**56**	**63**					
C	**59**	**70**					
D	**64**	**72**					
E	**74**	**84**					
Mean	58.2	66.00					
					SS_X	SS_Y	SP

$$r = \frac{SP}{\sqrt{SS_X SS_Y}}$$

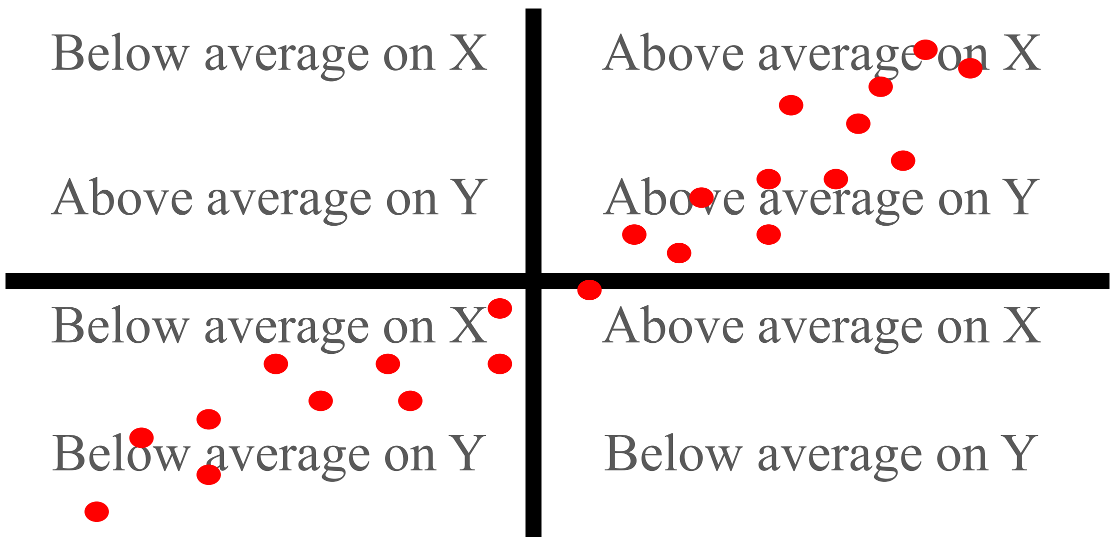
Pearson Correlation Coefficient

	Femur	Humerus	$(X-\bar{X})$	$(Y-\bar{Y})$	$(X-\bar{X})^2$	$(Y-\bar{Y})^2$	$(X-\bar{X})(Y-\bar{Y})$
A	**38**	**41**	-20.2	-25	408.04	625	505
B	**56**	**63**	-2.2	-3	4.84	9	6.6
C	**59**	**70**	0.8	4	.64	16	3.2
D	**64**	**72**	5.8	6	33.64	36	34.8
E	**74**	**84**	15.8	18	249.64	324	284.4
mean	58.2	66.00			**696.8**	**1010**	**834**
					SS_X	SS_Y	SP

r = 0.99

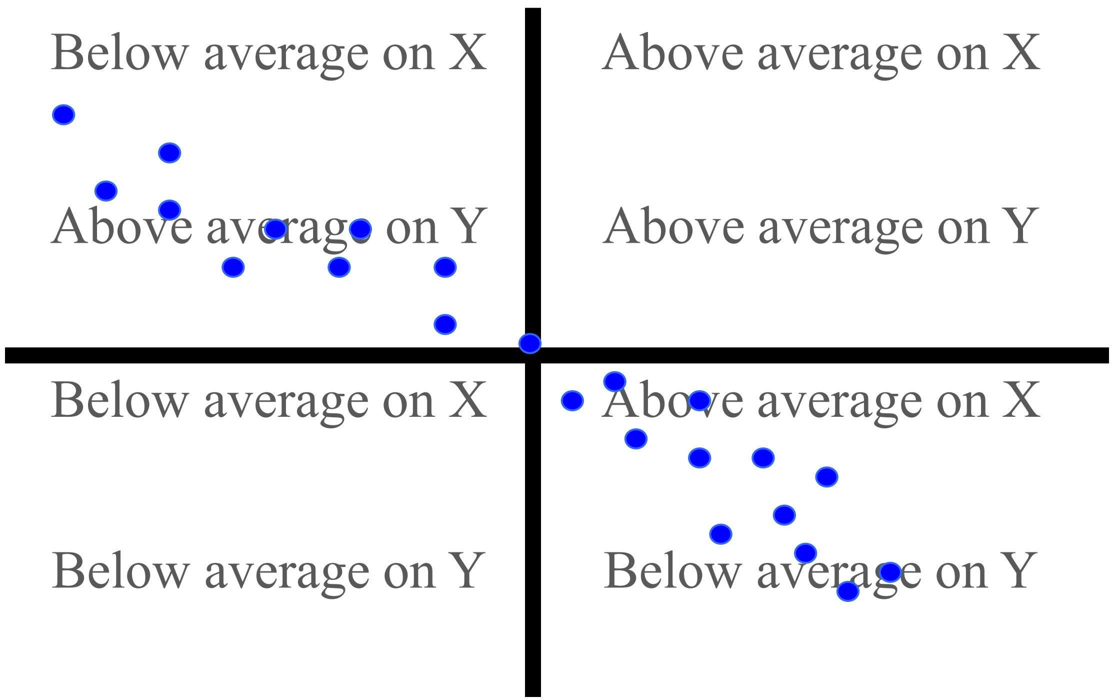
Pearson Correlation Coefficient

- For a strong positive association, the SP (sum of products) will be a big positive number



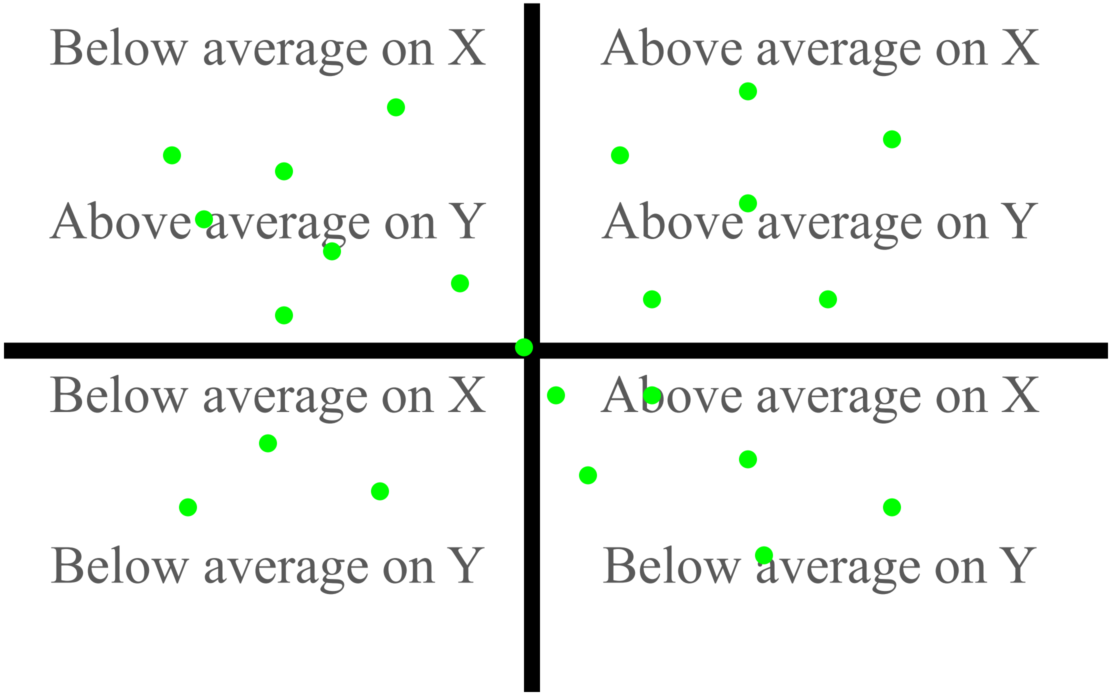
Pearson Correlation Coefficient

- For a strong <u>negative</u> association, the SP will be a big negative number



Pearson Correlation Coefficient

- For a <u>weak</u> association, the SP will be a small number (+ and – will cancel each other out)



Pearson Correlation Coefficient: Interpretation

- A measure of strength of association: how closely do the points cluster around a line?
- A measure of the direction of association: is it positive or negative?
- Colton [Colton T. Statistics in Medicine. Little Brown and Company, New York, NY 1974] rules:
  - $R \subset [-0.25 \text{ to } +0.25] \longrightarrow$ No relation
  - $R \subset (0.25 \text{ to } +0.50] \cup (-0.25 \text{ to } -0.50] \longrightarrow$ weak relation
  - $R \subset (0.50 \text{ to } +0.75] \cup (-0.50 \text{ to } -0.75] \longrightarrow$ moderate relation
  - $R \subset (0.75 \text{ to } +1) \cup (-0.75 \text{ to } -1) \longrightarrow$ strong relation

Pearson Correlation Coefficient: Interpretation

- The P-value is the probability that you would have found the current result if the correlation coefficient were in fact zero (null hypothesis).
- If this probability is lower than the conventional significance level (e.g. 5%) ($p < 0.05$) → the correlation coefficient is called statistically significant.

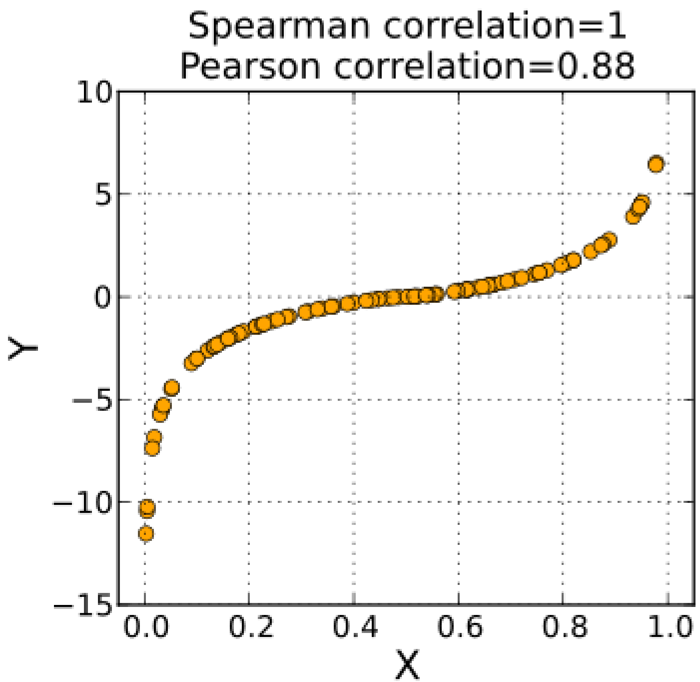
Spearman Rank Correlation Coefficient

- Not continuous measurements
- The assumption of bivariate normal distribution is violated
- Symbol: ρ (Rho Greek Letter)

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \times (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Spearman Rank Correlation Coefficient

- The sign of the Spearman correlation indicates the direction of association between X (the independent variable) and Y (the dependent variable).
- $\rho = 1 \rightarrow$ the two variables being compared are monotonically related. N.B. This does not give a perfect Pearson correlation.



Interpretation of r-squared (r^2)

- The amount of covariation compared to the amount of total variation
- The percent of total variance that is shared variance
- E.g. If $r = 0.80$, then X explains 64% of the variability in Y (and vice versa)

Properties of correlation coefficient

- A standardized statistic – will not change if you change the units of X or Y.
- The same whether X is correlated with Y or vice versa
- Fairly unstable with small n
- Vulnerable to outliers
- Has a skewed distribution

Correlation coefficient by example

- Hester KL, Macfarlane JG, Tedd H, Jary H, McAlinden P, Rostron L, Small T, Newton JL, De Soyza A. Fatigue in bronchiectasis. QJM. 2011 Oct 20. [Epub ahead of print]

- Results:
  - Fatigue correlated with MRCD score (Medical Research Council dyspnoea score) ($r = 0.57$, $P < 0.001$) and FEV(1)% predicted ($r = -0.30$, $P = 0.001$).

Correlation coefficient by example

www.gp-training.net/protocol/respiratory/copd/dyspnoea_scale.htm

Home | Protocols | Respiratory | COPD

MRC dyspnoea scale

Medical Research Council dyspnoea scale for grading the degree of a patient's breathlessness

1. Not troubled by breathlessness except on strenuous exercise
2. Short of breath when hurrying or walking up a slight hill
3. Walks slower than contemporaries on the level because of breathlessness, or has to stop for breath when walking at own pace
4. Stops for breath after about 100 m or after a few minutes on the level
5. Too breathless to leave the house, or breathless when dressing or undressing

Correlation coefficient by example

- Canan F, Ataoglu A, Ozcetin A, Icmeli C. The association between Internet addiction and dissociation among Turkish college students. Compr Psychiatry. 2011 Oct 13. [Epub ahead of print]
- **RESULTS:**

  According to the Internet Addiction Scale, 9.7% of the study sample was addicted to the Internet. The Pearson correlation analysis results revealed a significant positive correlation between dissociative experiences and Internet addiction ($r = 0.220$; $P < .001$) and weekly Internet use ($r = 0.227$; $P < .001$). Levels of Internet addiction were significantly higher among male students than female students ($P < .001$). The Internet use pattern also differed significantly between sexes.

Linear Regression

Simple Linear regression

Multiple linear regression

Linear Regression: Assumptions

- The errors in data values (e.g. the deviation from average) are independent from one another
- Regressions depends on the appropriateness of the model used in the fit
- The independent readings (X) are measured as exactly known values (measured without error)
- The variance of Y is the same for all values of X
- The distribution of Y is approximately normal for all values of X

Linear Regression

- But how do we describe the line?
- If two variables are linearly related it is possible to develop a simple equation to predict one variable from the other
- The outcome variable is designated the Y variable, and the predictor variable is designated the X variable
- E.g. centigrade to Fahrenheit:

$$F = 32 + 1.8°C$$

this formula gives a specific straight line

Linear Equation

- $F = 32 + 1.8(C)$
- General form is $Y = a + bX$
- The prediction equation: $\tilde{Y} = a + bX$
  - a = intercept, b = slope, X = the predictor, Y = the criterion
- *a and b are constants in a given line; X and Y change*

Slope and Intercept

- Equation of the line: $\tilde{Y} = a + bX$
- The slope b: the amount of change in Y with one unit change in X

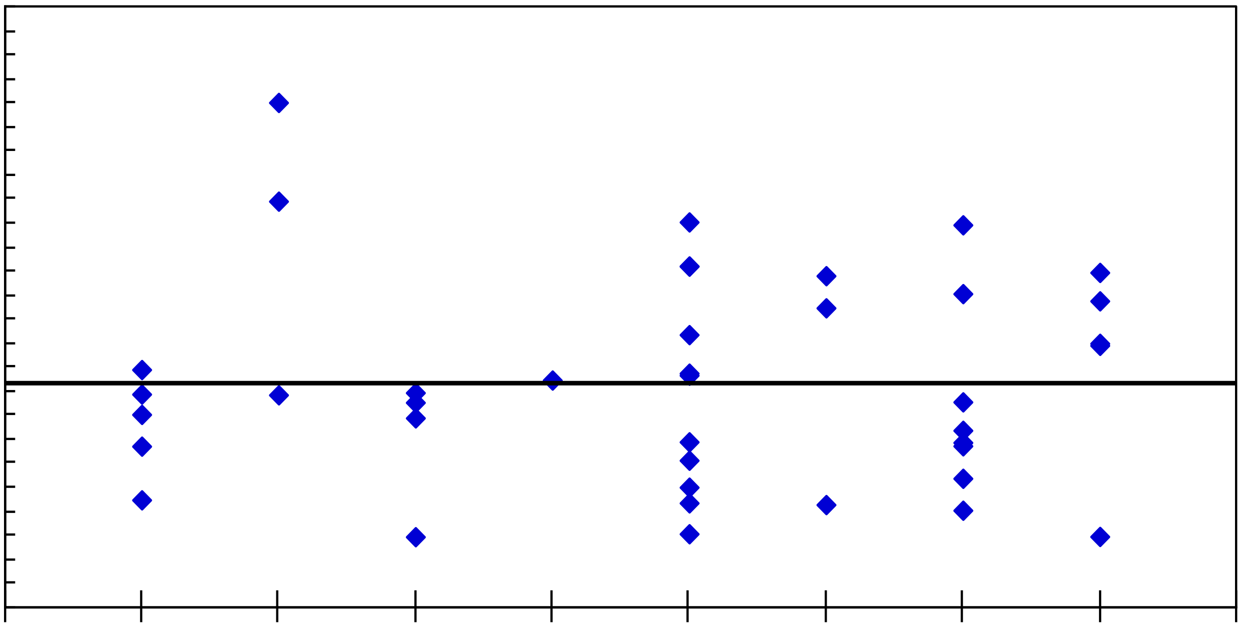
$$b = r\frac{s_y}{s_x} = \frac{SP}{SS_X}$$

- The intercept a: the value of Y when X is zero

$$a = \overline{Y} - b\overline{X}$$

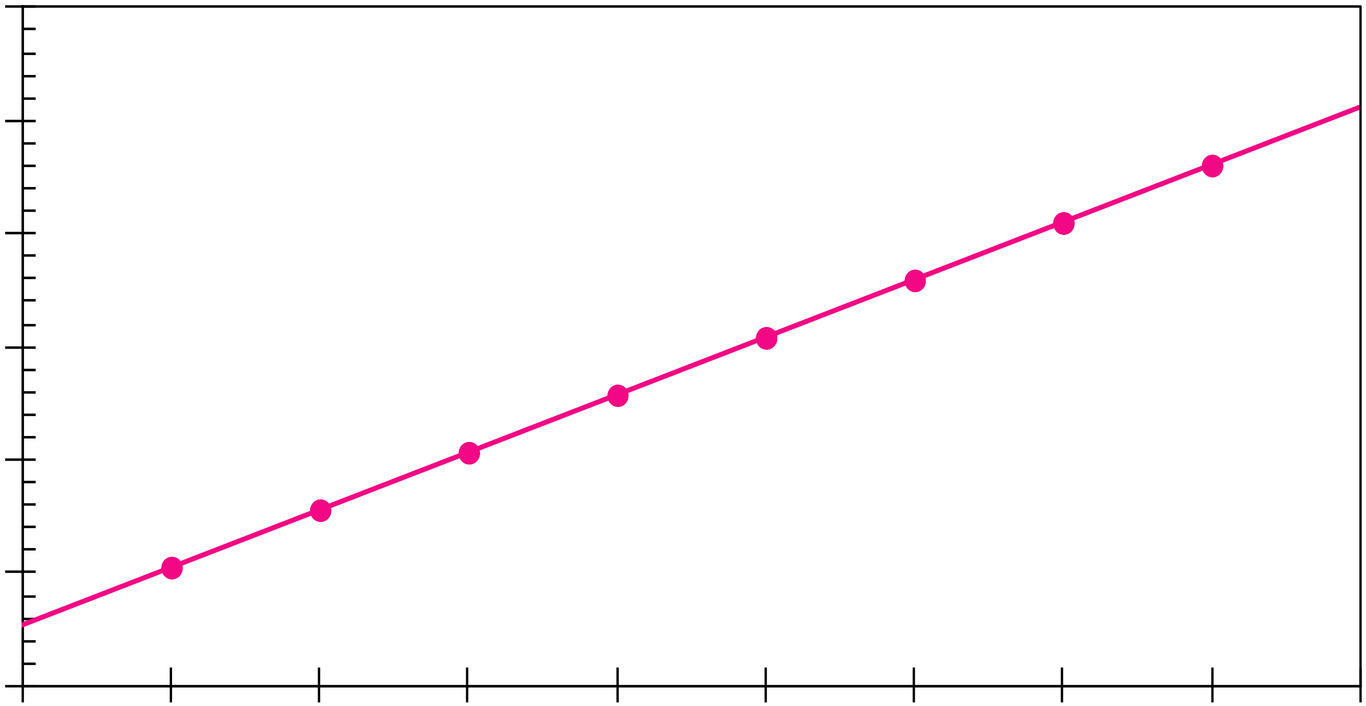
- The slope is influenced by r, but is not the same as r

When there is no linear association (r = 0), the regression line is horizontal, b=0

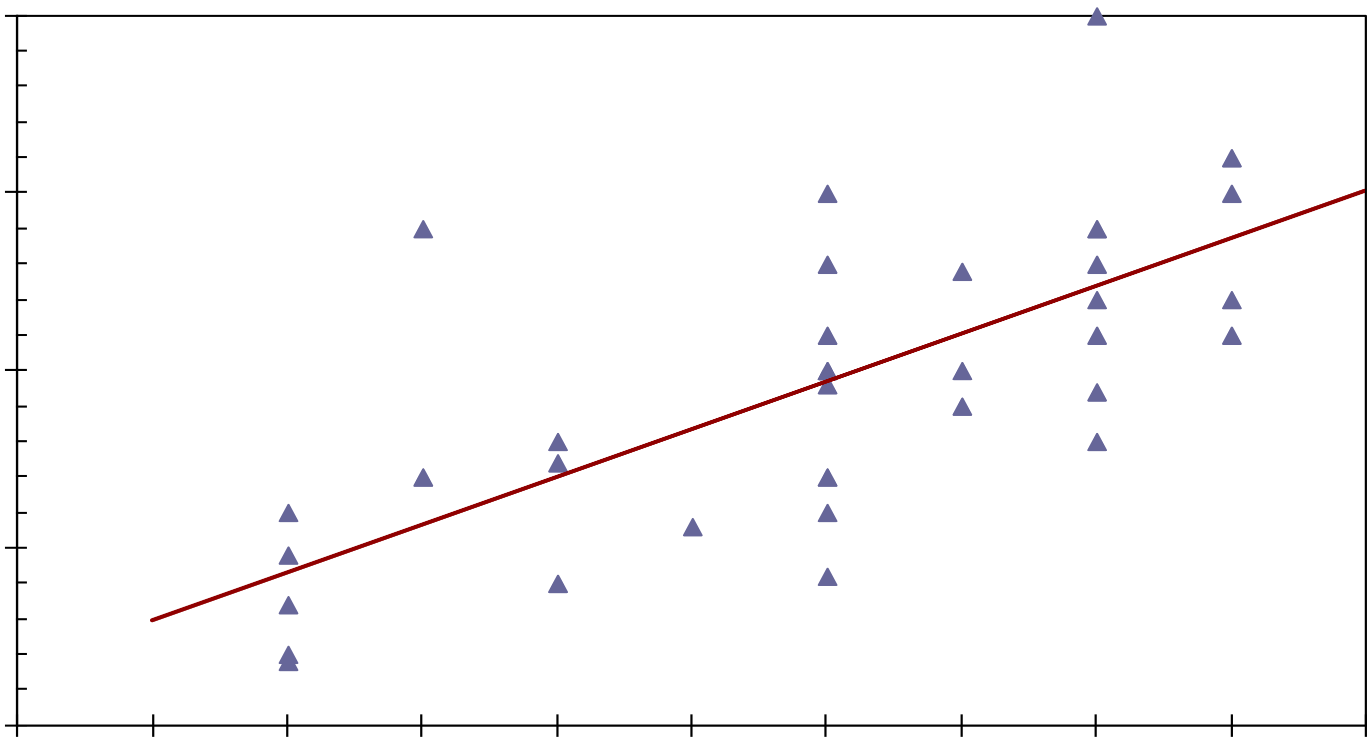


and our best estimate of age is 29.5 at all heights.

When the correlation is perfect ($r = \pm 1.00$),
all the points fall along a straight line with a slope



When there is some linear association (0<|r|<1), the regression line fits as close to the points as possible and has a slope

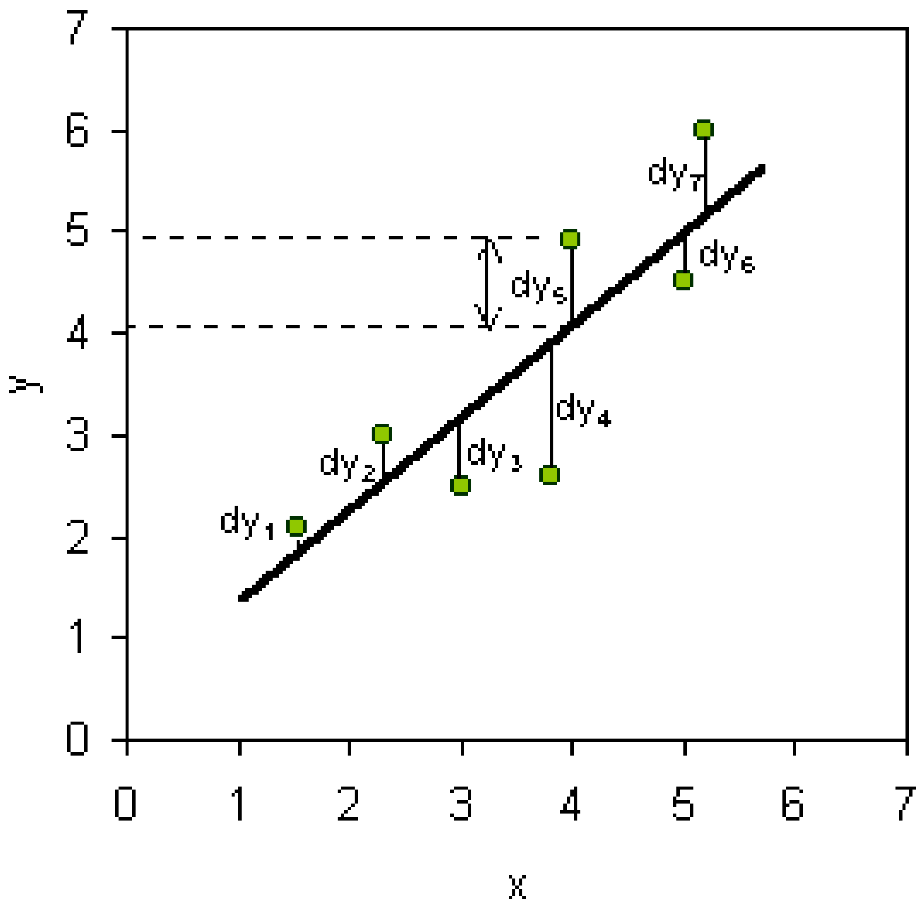


Where did this line come from?

- It is a straight line which is drawn through a scatterplot, to summarize the relationship between X and Y
- It is the line that minimizes the squared deviations $(\tilde{Y} - Y)^2$
- We call these vertical deviations "residuals"

Regression Line

- Minimizing the squared vertical distances, or "residuals"



Regression Coefficients Table

Predictor	Unstandardized Coefficient	Standard error	t	p
Intercept	a	SE_a	$t=a/SE_a$	
Variable X	b	SE_b	$t=b/SE_b$	

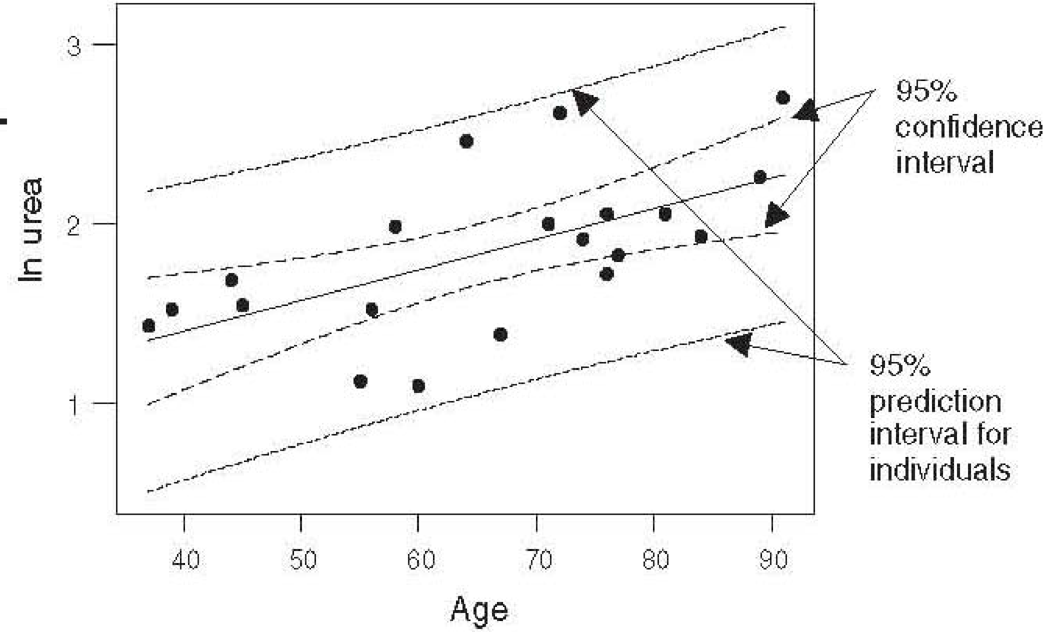
Regression parameter estimates, *P* values and confidence intervals for the accident and emergency unit data

	Coefficient	Standard error of coefficient	t	*P*	Confidence interval
Constant, or intercept	0.72	0.346	2.07	0.054	−0.01 to +1.45
ln urea	0.017	0.005	3.35	0.004	0.006 to 0.028

Linear Regression

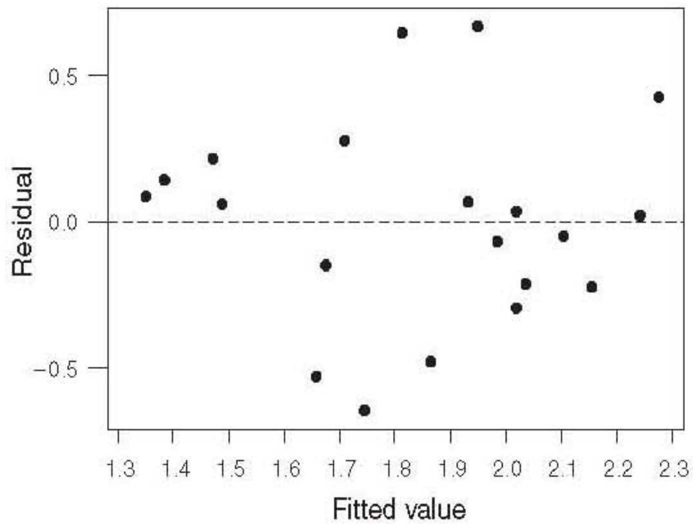
Analysis of variance for the accident and emergency unit data

Source of variation	Degrees of freedom	Sum of squares	Mean square	F	*P*
Regression	1	1.462	1.462	11.24	0.004
Residual	18	2.342	0.130		
Total	19	3.804			

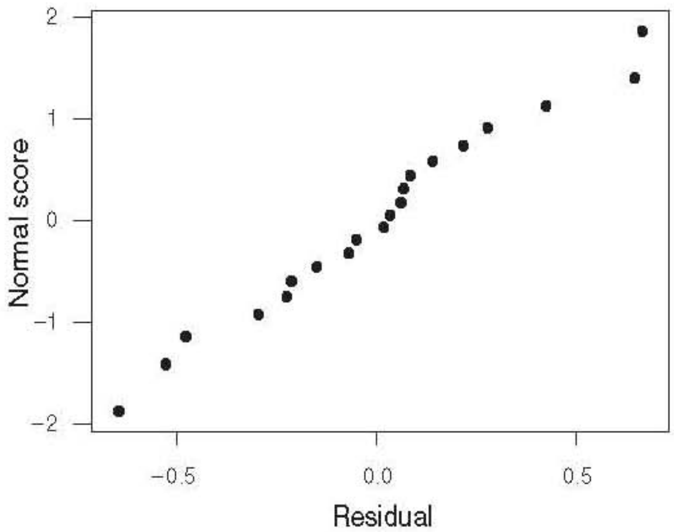


Regression line, its 95% confidence interval and the 95% prediction interval for individual patients.

Linear Regression



Plot of residuals against fitted values for the accident and emergency unit data.



Normal plot of residuals for the accident and emergency unit data.

Correlation & Regression: Summary

- Both correlation and simple linear regression can be used to examine the presence of a linear relationship between two variables providing certain assumptions about the data are satisfied.

- The results of the analysis need to be carefully interpreted, particularly when looking for a causal relationship or when using the regression equation for prediction.