Correlation and Regression

Dr. A M Yusufali; MD, FRCP(Glasg)
Head Cardiology and Consultant Physician/Cardiologist,
Cardiology and Cardiothoracic Surgery Centre,
Dubai Hospital,
DOHMS

Paired data

- Is there a relationship
- If so, in what way are they related or what is the equation?
- Use the equation for prediction

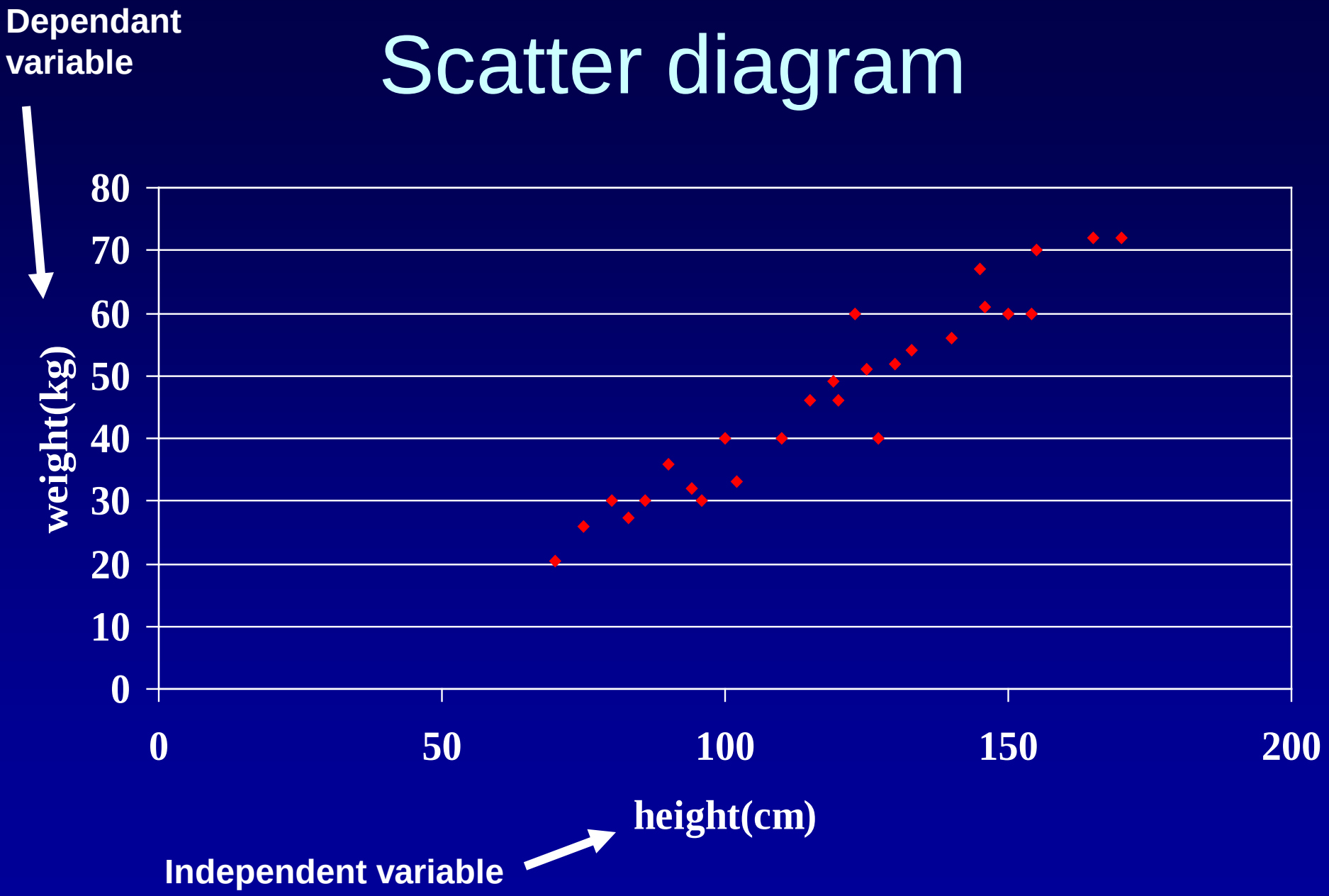
Definition

- Correlation : exits between two variables when one is related to the other in some way

Assumptions

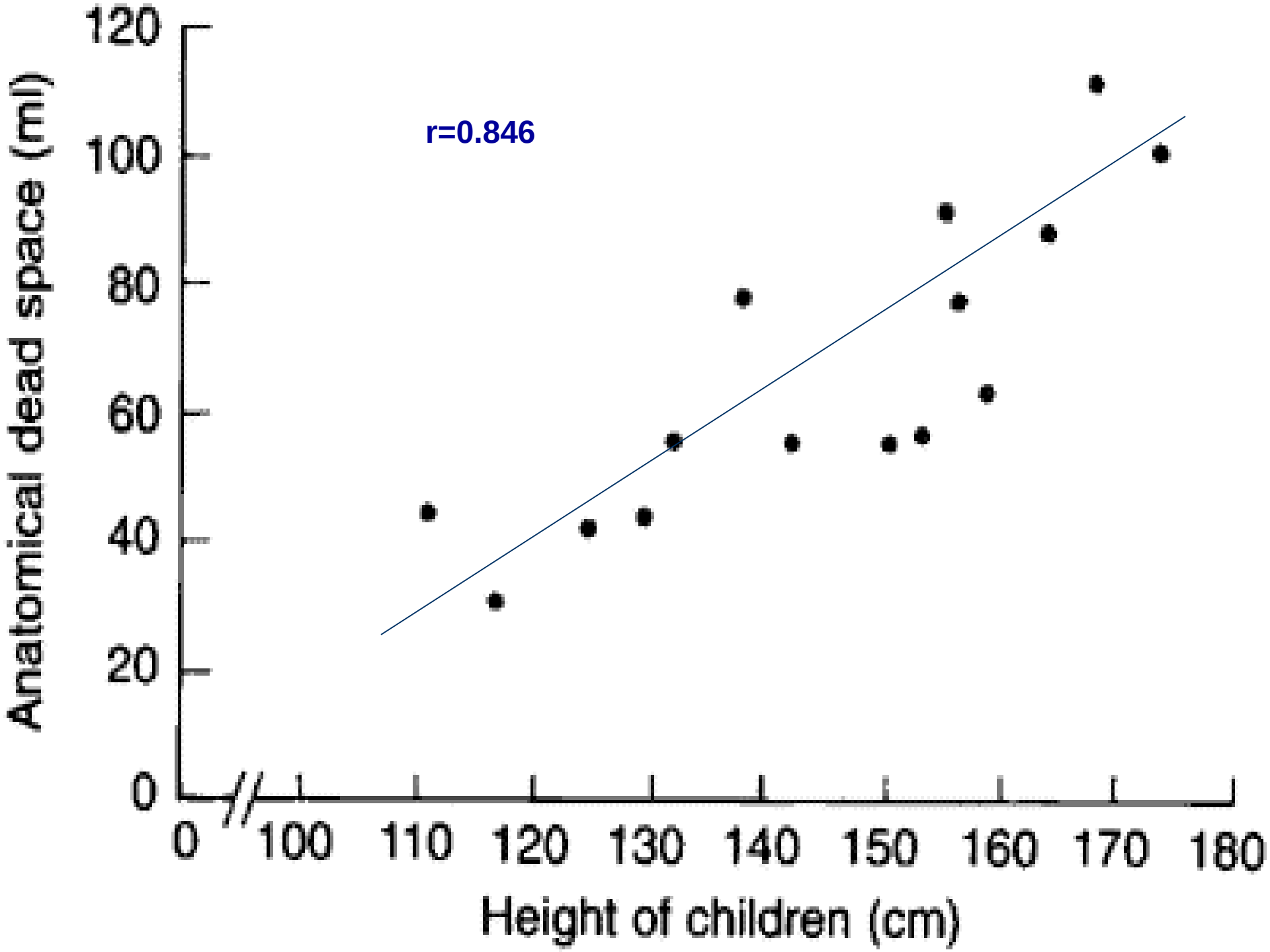
- The paired samples (x, y) are randomly selected.
- Both data are normally distributed

Scatter diagram



Correlation between height and pulmonary dead space in 15 children

Child number	Height (cm)	Dead space (ml), y
1	110	44
2	116	31
3	124	43
4	129	45
5	131	56
6	138	79
7	142	57
8	150	56
9	153	58
10	155	92
11	156	78
12	159	64
13	164	88
14	168	112
15	174	101
Total	**2169**	**1004**
Mean	**144.6**	**66.933**


r=0.846
Anatomical dead space (ml)
120
100
80
60
40
20
0
0
100
110
120
130
140
150
160
170
180
Height of children (cm)

Linear coefficient correlation r

- Measures the strength of linear relationship between the paired x and y values in the sample.
- $-1 \geq r \leq +1$

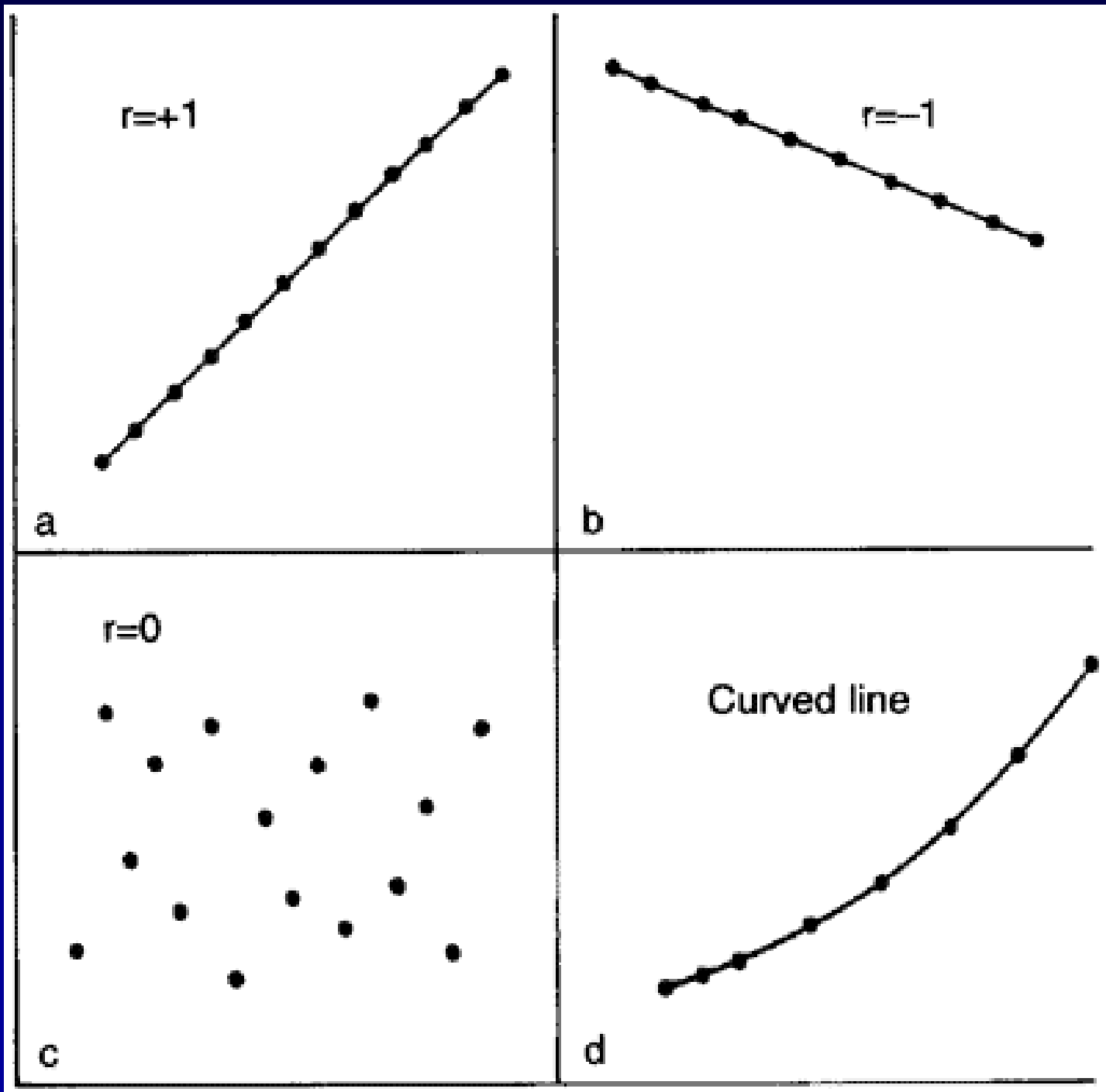
r value

- for absolute values of r,
- 0-0.19 is regarded as very weak,
- 0.2-0.39 as weak,
- 0.40-0.59 as moderate,
- 0.6-0.79 as strong and
- 0.8-1 as very strong correlation,
- but these are rather arbitrary limits, and the context of the results should be considered

Significance test

- To test whether the association is merely apparent, and might have arisen by chance a significance test is used.
- $P<0.001$: so the correlation coefficient may be regarded as highly significant.
- Thus we have a very strong correlation between dead space and height which is most unlikely to have arisen by chance.

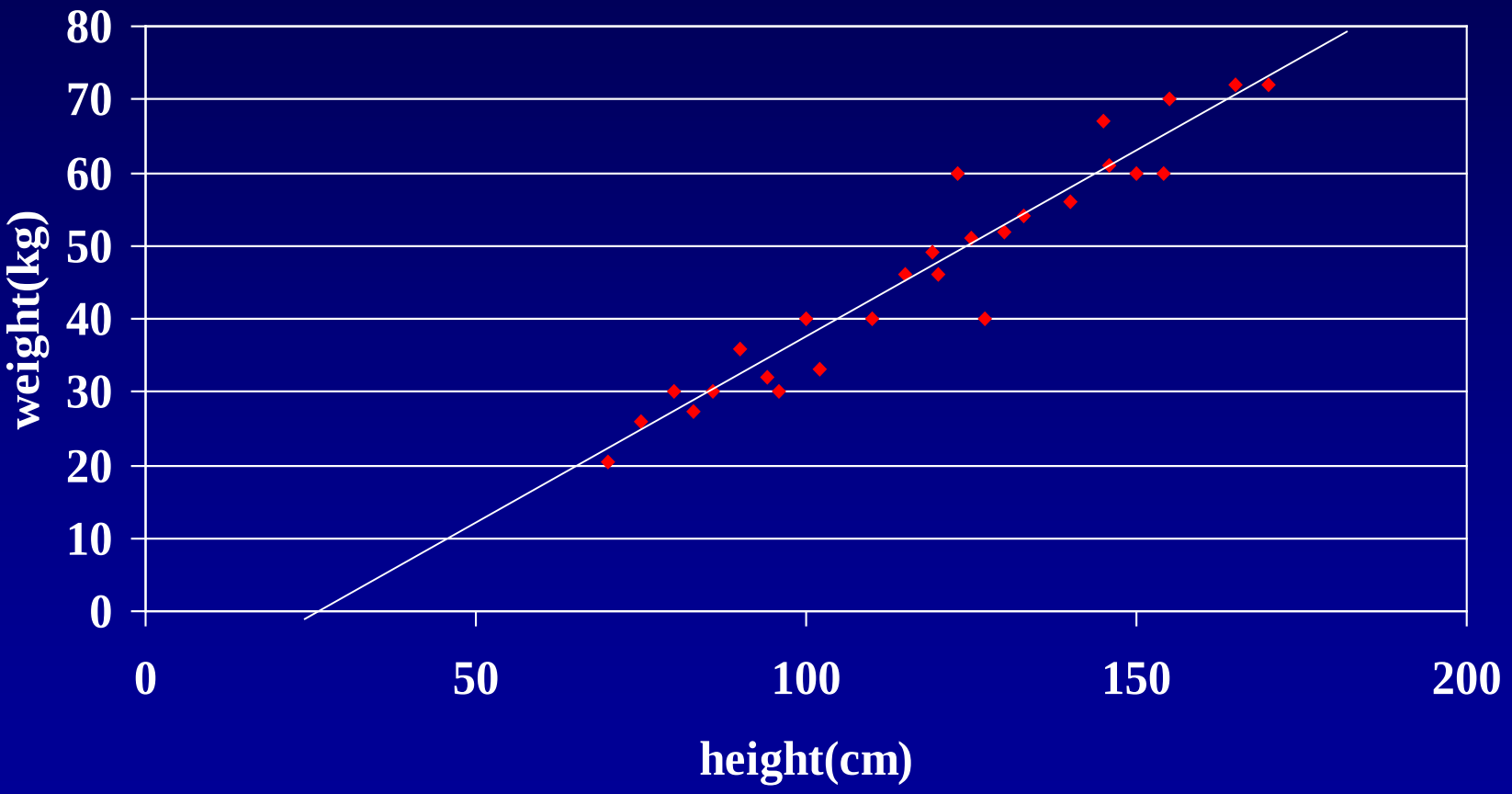
Correlation coefficient



Regression : Definitions

- **Regression Equation:**
  Given a collection of paired data, the regression equation
  $y = mx + c$
  describes the relationship between two variables.
- **Regression Line:**
  the graph regression equation.

Regression



Y= mx + c

Common errors involving correlations

- Significant linear correlation = causality
- Prediction beyond the original data
- No linear relationship, does not imply no relationship at all. There is a possibility of non-linear relationship e.g. curved

Tests for causation

- Is there evidence from true experiments in humans?
- Is the association strong?
- Is the association consistent from study to study?
- Is the temporal relation appropriate ?
- Is there a dose-response gradient ?
- Does the association make epidemiological sense?
- Does the association make biological sense?
- Is the association specific?

Multiple Regression

- "regression" refers to a mathematical equation- relationship of two variables
- In case of multiple regression, a far more complex mathematical equation.
- Allows several other variables to be factored into the model.
- Example: height and weight
  family hx, calorie intake, sex, physical activity

Multiple regression

- More than one variable is used to predict the criterion.
- For example, a college admissions officer wishing to predict the future grades of college applicants might use three variables (High School Grade, entrance exam, and Quality of letters of recommendation) to predict college Grade.
- The applicants with the highest predicted college Grade would be admitted.

- The prediction method would be developed based on students already attending college and then used on subsequent classes.
- Predicted scores from multiple regression are linear combinations of the predictor variables.
- Therefore, the general form of a prediction equation from multiple regression is:
  $Y' = b1X1 + b2X2 + ... + bkXk + A$

the predicted score

the score on the first predictor variable

The regression coefficients (b1, b2, etc.) are analogous to the slope in simple regression.

Y intercept

Strategies for Reducing the Door-to-Balloon Time in Acute Myocardial Infarction

Elizabeth H. Bradley. et al

- surveyed 365 hospitals to determine whether each of 28 specific strategies was in use.
- to determine the association between hospital strategies and the door-to-balloon time.
- To evaluate the association between specific hospital strategies and the door-to-balloon time, we used hierarchical generalized linear models.
- In multivariate analysis, six strategies were significantly associated with a faster door-to-balloon time.

The new england journal *o f* medicine November 13, 2006.

Strategies for Reducing the Door-to-Balloon Time in Acute Myocardial Infarction

- These strategies included having emergency medicine physicians activate the catheterization laboratory (mean reduction in door-to-balloon time,8.2 minutes),
- having a single call to a central page operator activate the laboratory (13.8 minutes),
- having the emergency department activate the catheterization laboratory while the patient is en route to the hospital (15.4 minutes),
- expecting staff to arrive in the catheterization laboratory within 20 minutes after being paged (vs. >30 minutes) (19.3 minutes),
- having an attending cardiologist always on site (14.6 minutes),
- and having staff in the emergency department and the catheterization laboratory use real-time data feedback (8.6 minutes).

For a group of young men with height 160-190cm the PEFR was measured and the results plotted on a scatter diagram.

- A. the correlation coefficient can't be higher than 1
- B. if the correlation coefficient is 0 there is no relationship between height and PEFR
- C. if the correlation coefficient is positive, the curve would have an upward slope
- D. this data can be used to calculate the PEFR for a height of 150cm
- E. height should be on the vertical axis on the graph

Statistics at Square One
Ninth Edition
Copyright BMJ Publishing Group

Contents

THANK YOU