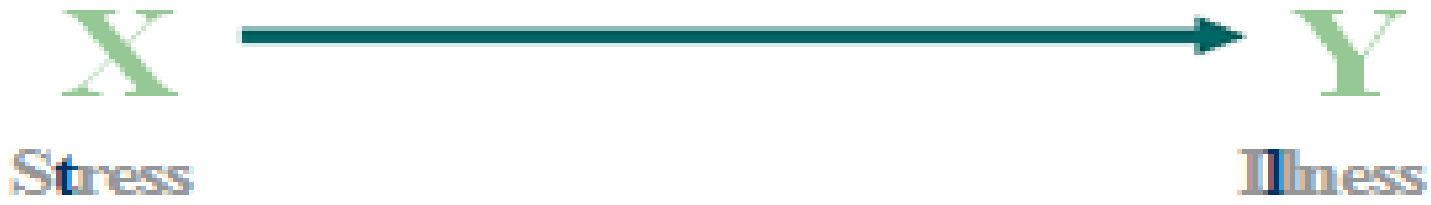
Correlation(Pearson & spearman) &Linear Regression

Correlation

* Semantically, Correlation means **Co**-together and **Relation.**
* Statistical correlation is a statistical technique which tells us if two variables are related.



PEARSON CORRELATION

* measures the degree of linear association between two interval scaled variables analysis of the relationship between two quantitative outcomes, e.g., height and weight

$$\rho_{X,Y} = \operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y},$$

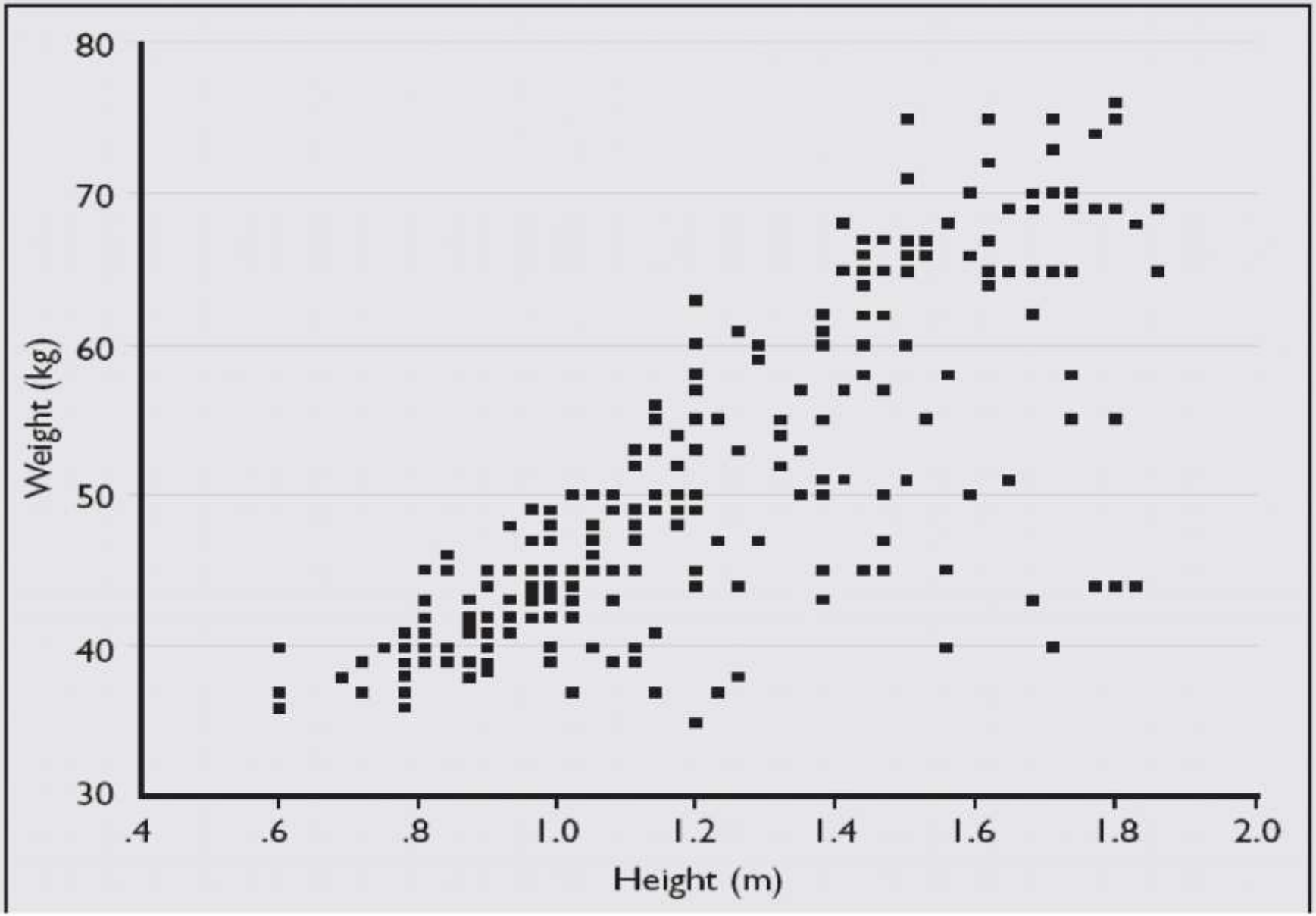
Assumption under Pearson's Correlation Coefficient

* **Assumption 1: The correlation coefficient *r* assumes that the two variables measured form a bivariate normal distribution population.**
* **Assumption 2: The correlation coefficient r measures only linear associations: how nearly the data falls on a straight line. It is not a good summary of the association if the scatterplot has a nonlinear (curved) pattern.**

Assumptions Contd.

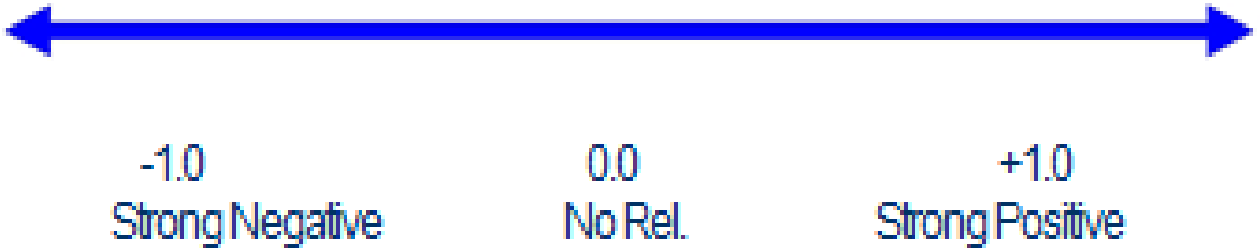
* **Assumption 3: The correlation coefficient *r* is not a good summary of association if the data are heteroscedastic.(when random variables have the same finite variance. It is also known as homogenity of variance)**
* **Assumption 4: The correlation coefficient *r* is not a good summary of association if the data have outliers.**

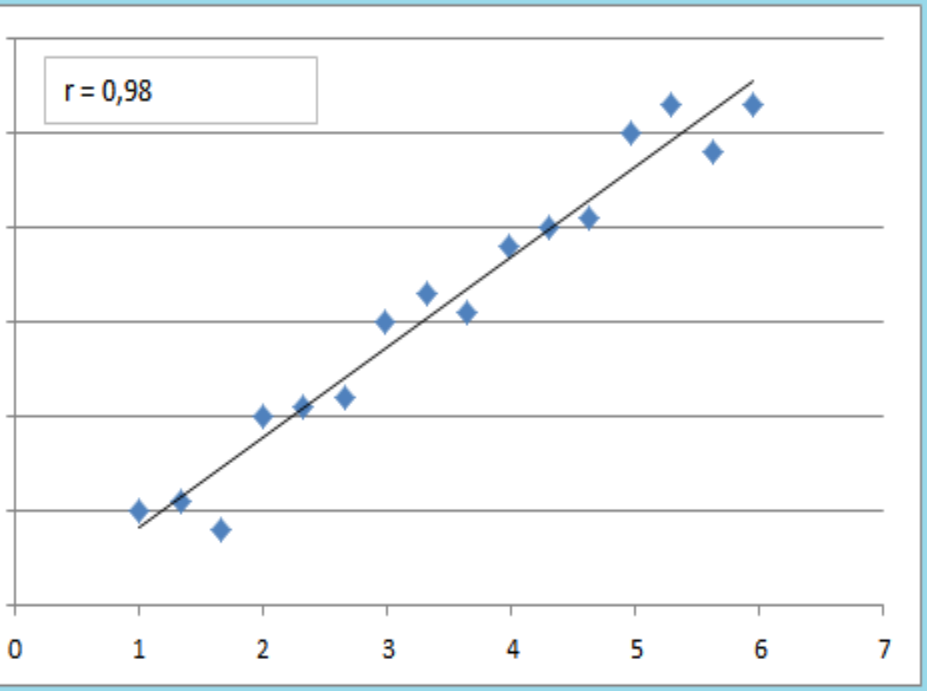
Fig. 1 Relationship between height and weight.

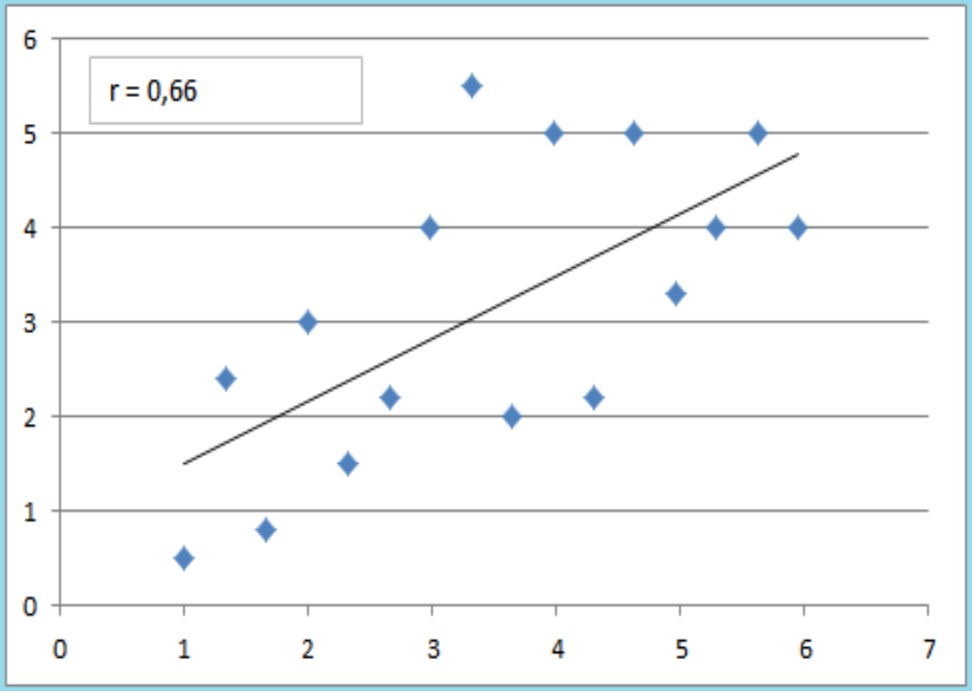


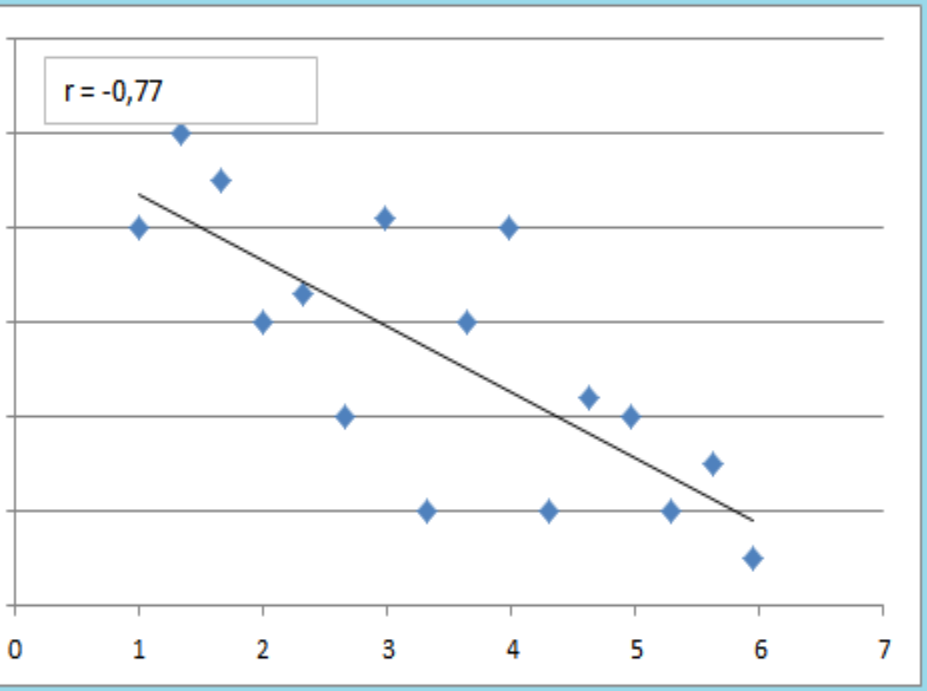
Strength of Relationship

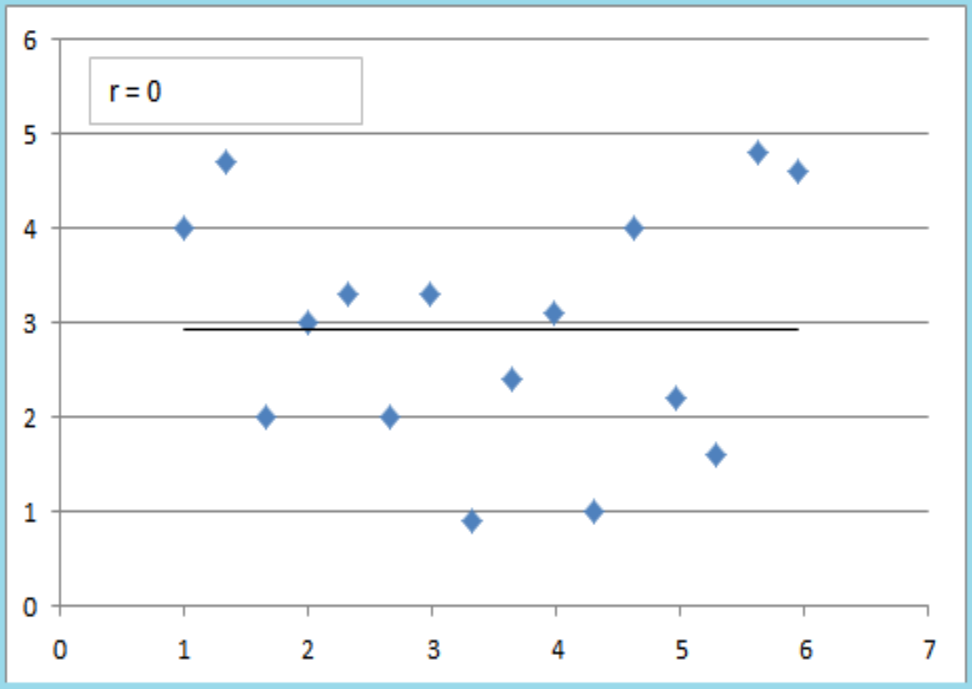
* r lies between -1 and 1. Values near 0 means no (linear) correlation and values near ± 1 means very strong correlation.
*




r = 0,98
0
1
2
3
4
5
6
7


r = 0,66
0
1
2
3
4
5
6
7


r = -0,77
0
1
2
3
4
5
6
7


r = 0
0
1
2
3
4
5
6
7

Interpretation of the value of r

Table II. Strength of linear relationship.

Correlation Coefficient value	Strength of linear relationship
At least 0.8	Very strong
0.6 up to 0.8	Moderately strong
0.3 to 0.5	Fair
Less than 0.3	Poor

Coefficient of Determination

* Pearson's *r* can be squared , *r* 2, to derive a coefficient of determination.
* Coefficient of determination - the portion of variability in one of the variables that can be accounted for by variability in the second variable

Example

* Pearson's *r* can be squared , *r* 2
* If *r*=0.5, then *r*2=0.25 If *r*=0.7 then *r*2=0.49
* Thus while *r*=0.5 versus 0.7 might not look so different in terms of strength, *r*2 tells us that *r*=0.7 accounts for about twice the variability relative to *r*=0.5

Spearman's Rank Correlation

* **Spearman's rank correlation** coefficient or **Spearman's** rho, is a measure of statistical dependence between two variables.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}.$$

Interpretation

* The sign of the Spearman correlation indicates the direction of association between *X* (the independent variable) and *Y* (the dependent variable). If *Y* tends to increase when *X* increases, the Spearman correlation coefficient is positive. If *Y* tends to decrease when *X* increases, the Spearman correlation coefficient is negative. A Spearman correlation of zero indicates that there is no tendency for *Y* to either increase or decrease when *X* increases

Repeated Ranks

* If there is more than one item with the same value , then they are given a common rank which is average of their respective ranks as shown in the table.

Variable X_i	Position in the ascending order	Rank x_i
0.8	1	1
1.2	2	$\frac{2+3}{2} = 2.5$
1.2	3	$\frac{2+3}{2} = 2.5$
2.3	4	4
18	5	5

Example

* The raw data in the table below is used to calculate the correlation between the IQ of an with the number of hours spent in front of TV per week.

IQ, X_i	Hours of TV per week, Y_i
106	7
86	0
100	27
101	50
99	28
103	29
97	20
113	12
112	6
110	17

Example Contd.

IQ, X_i	Hours of TV per week, Y_i	rank x_i	rank y_i	d_i	d_i^2
86	0	1	1	0	0
97	20	2	6	−4	16
99	28	3	8	−5	25
100	27	4	7	−3	9
101	50	5	10	−5	25
103	29	6	9	−3	9
106	7	7	3	4	16
110	17	8	5	3	9
112	6	9	2	7	49
113	12	10	4	6	36

$$\rho = 1 - \frac{6 \times 194}{10(10^2 - 1)}$$

$$\rho = -29/165 = -0.175757575.$$

Regression

* One variable is a direct cause of the other or if the value of one variable is changed, then as a direct consequence, the other variable also change or if the main purpose of the analysis is prediction of one variable from the other
*

Regression

* Regression: the dependence of dependent variable Y on the independent variable X.

* Relationship is summarized by a regression equation.

  $y = a + bx$

* A=intercept at y axis
* B=regression coefficient

The Least Squares Method

* The line of regression is the line which gives the best estimate to the value of one variable for any specific value of the other variable. Thus the line of regression is the line of "best fit" and is Obtained by the *principle of least squares.*
* This principle consists in minimizing the sum of the squares of the deviations of the actual values of y from their estimate values given by the line of best fit

Formulas to be used

$$
\begin{aligned}
\mathrm{ss}_{x\,x} &= \sum_{i=1}^{n} (x_i - \bar{x})^2 \\
&= \left(\sum_{i=1}^{n} x_i^2\right) - n\,\bar{x}^2 \\
\mathrm{ss}_{y\,y} &= \sum_{i=1}^{n} (y_i - \bar{y})^2 \\
&= \left(\sum_{i=1}^{n} y_i^2\right) - n\,\bar{y}^2 \\
\mathrm{ss}_{x\,y} &= \sum_{i=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y}) \\
&= \left(\sum_{i=1}^{n} x_i\,y_i\right) - n\,\bar{x}\,\bar{y},
\end{aligned}
$$

$$
b = \frac{\mathrm{cov}\,(x, y)}{\sigma_x^2} = \frac{\mathrm{ss}_{x\,y}}{\mathrm{ss}_{x\,x}},
$$

$$
a = \bar{y} - b\,\bar{x}.
$$

Example

* Fit a least square line to the following data

X	1	2	3	4	5
Y	2	5	3	8	7

Solution

X	Y	XY	X^2	$\hat{Y} = 1.1 + 1.3X$
1	2	2	1	2.4
2	5	10	4	3.7
3	3	9	9	5.0
4	8	32	16	6.3
5	7	35	25	7.6
$\sum X = 15$	$\sum Y = 25$	$\sum XY = 88$	$\sum X^2 = 55$	

$a = 1.1$ and $b = 1.3$.

THANK YOU.