MEASURING RELATIONSHIP BETWEEN CONTINUOUS VARIABLES

Correlation & Regression

Dr. Win Khaing

MBBS (Mdy),

MMedSc (PH)

Passed with Biostatistics Distinction

Cert. of Basic R for Epidemiology, Regression Analysis and Advanced Epidemiological Methods

WHO-TDR

Association

- Typically reserved for describing relationship between categorical variables

Correlation

- Typically reserved for describing relationship between continuous variables

Correlation

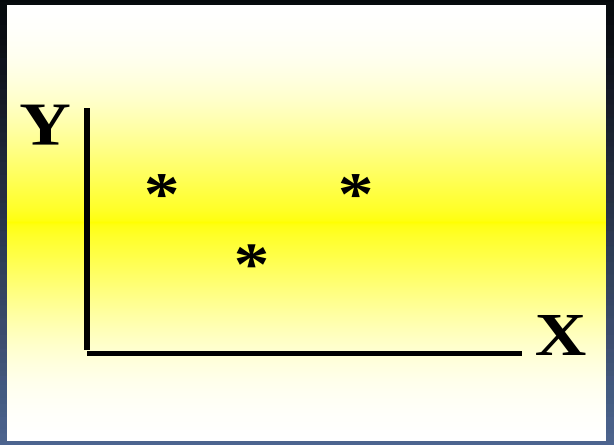
- When association between numerical data where a linear relationship is suspected, regression and correlation can be used
- Finding the relationship between two quantitative variables without being able to infer causal relationships
- Correlation is a statistical technique used to determine the degree to which two continuous variables are related

Scatter Diagram

First step in examining the relationship between two variables, measured on the same subjects, is always to draw a "Scatter diagram".

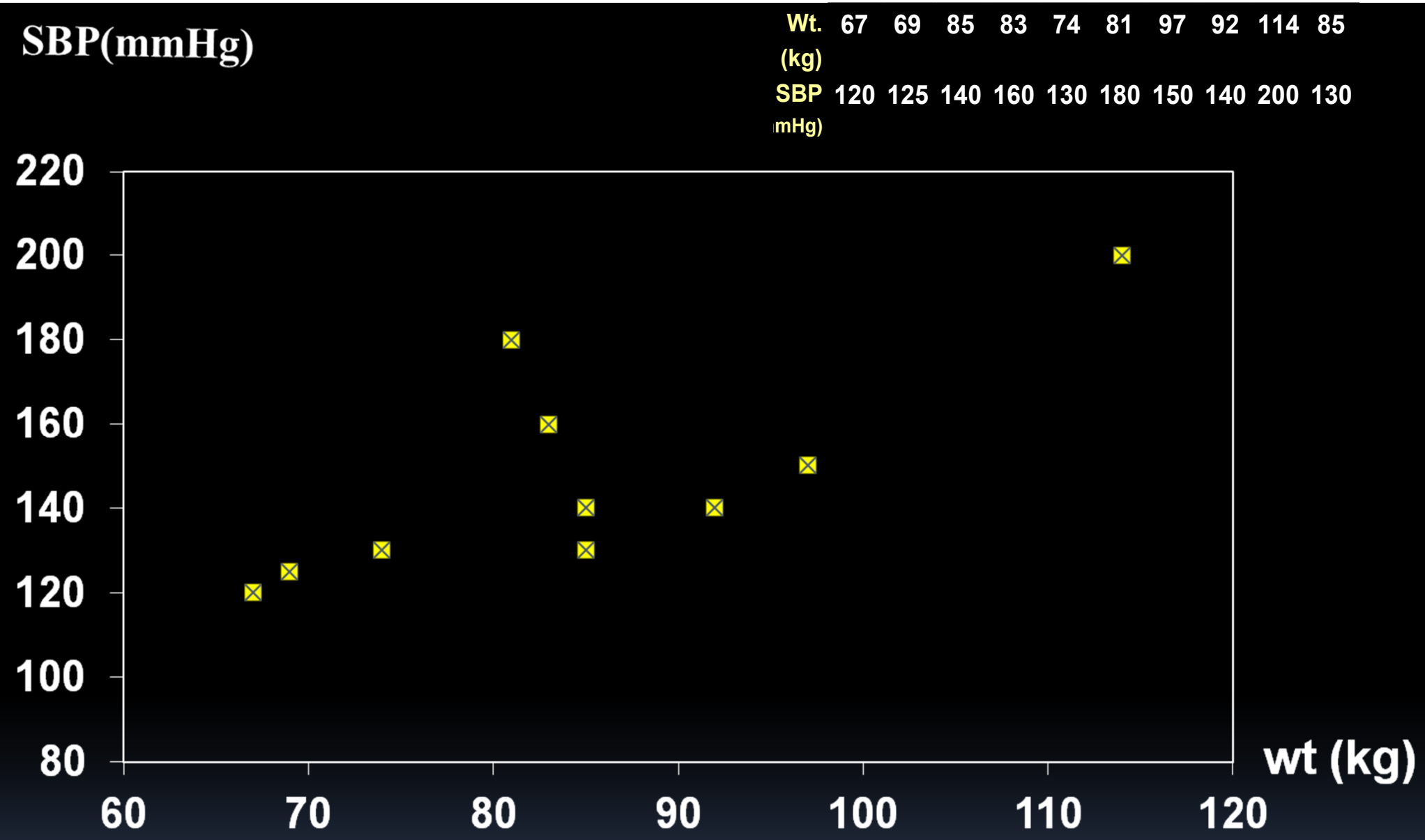
Scatter diagram

- Rectangular coordinate
- Two quantitative variables
- One variable is called independent (X) and the second is called dependent (Y)
- Points are not joined
- No frequency table

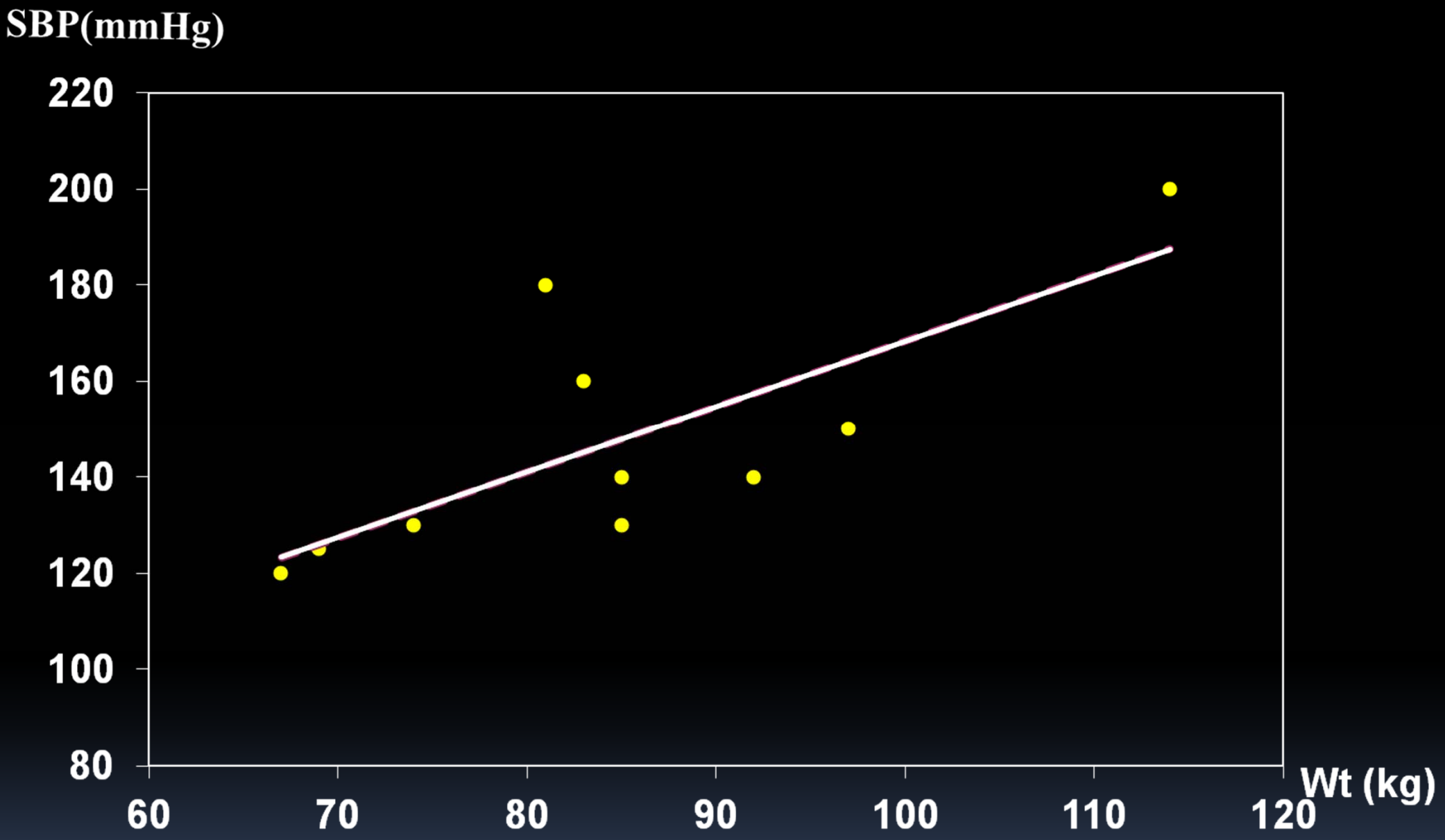


Example

Wt (kg)	67	69	85	83	74	81	97	92	114	85
SBP (mmHg)	120	125	140	160	130	180	150	140	200	130



Scatter diagram of weight and systolic blood pressure

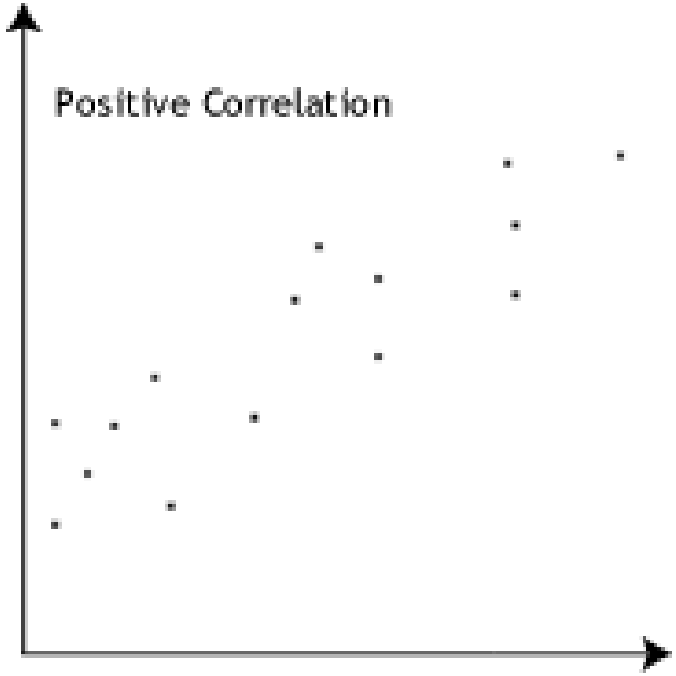


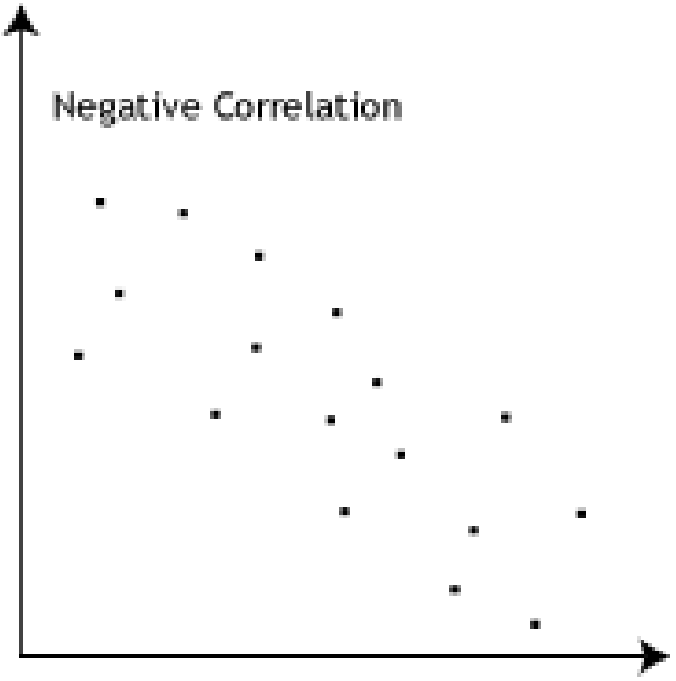
Scatter diagram of weight and systolic blood pressure

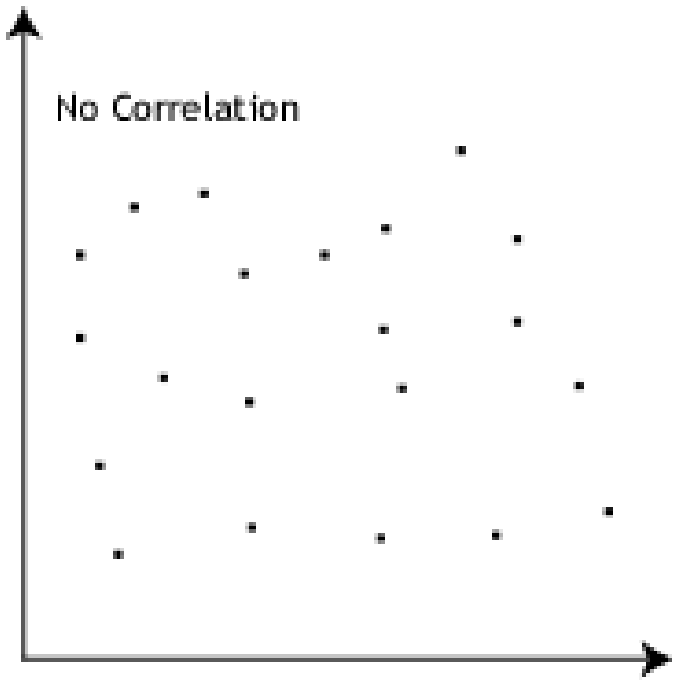
Scatter plots

The pattern of data is indicative of the type of relationship between your two variables:

- positive relationship
- negative relationship
- no relationship







Study Hours

Vs

Final Grade

Age of Car

Vs

Reliability

Body Weight

Vs

Pulse Rate

Correlation Coefficient

Commonly used

- Person's product-moment correlation coefficient (r)
- Others …

Simple Correlation coefficient (r)

- It is also called Pearson's correlation (or) product moment correlation coefficient.
- It measures the nature and strength between two variables of the quantitative type.

- The sign of r denotes the nature of association
- while the value of r denotes the strength of association.

- If the sign is +ve this means the relation is direct (an increase in one variable is associated with an increase in the other variable and a decrease in one variable is associated with a decrease in the other variable).

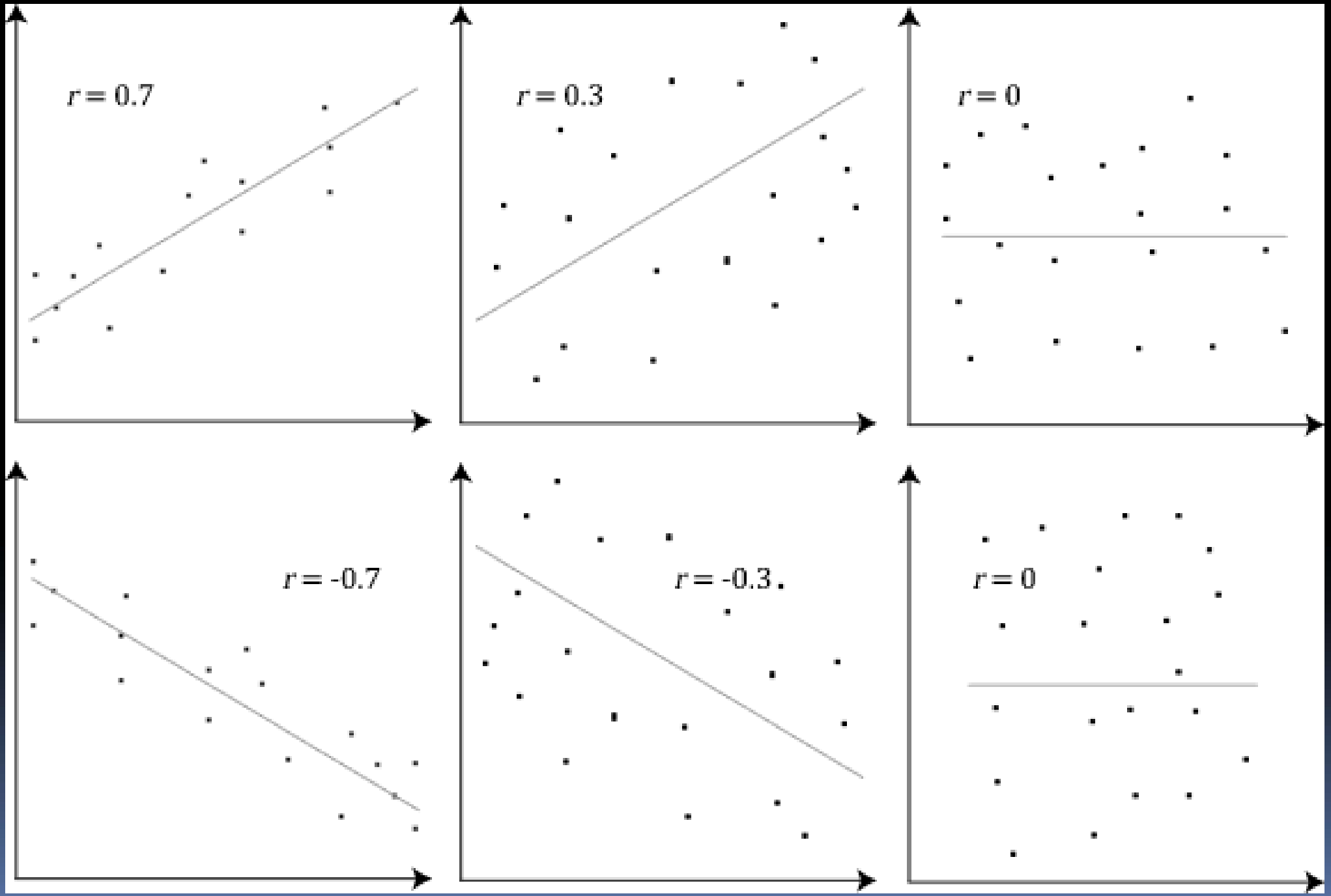
- While if the sign is -ve this means an inverse or indirect relationship (which means an increase in one variable is associated with a decrease in the other).

- The value of r ranges between (-1) and (+1)
- The value of r denotes the strength of the association as illustrated by the following diagram.

- If r = Zero this means no association or correlation between the two variables.
- If $0 < r < 0.25$ = weak correlation.
- If $0.25 \leq r < 0.75$ = intermediate correlation.
- If $0.75 \leq r < 1$ = strong correlation.
- If $r = 1$ = perfect correlation.

0 – 0.19	very weak
0.2 – 0.39	weak
0.40 – 0.59	moderate
0.6 – 0.79	strong
0.8 – 1	very strong

- Reference : Statistic Square One 11th Edition

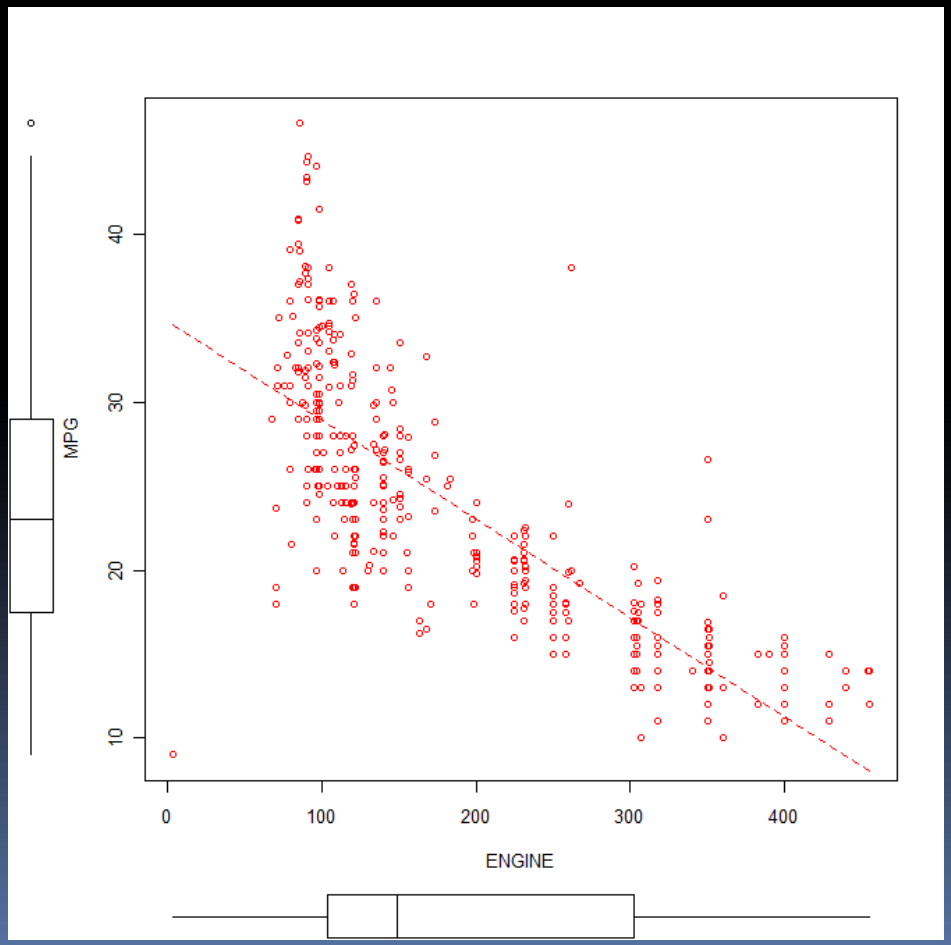

r = 0.7
r = 0.3
r = 0
r = -0.7
r = -0.3
r = 0

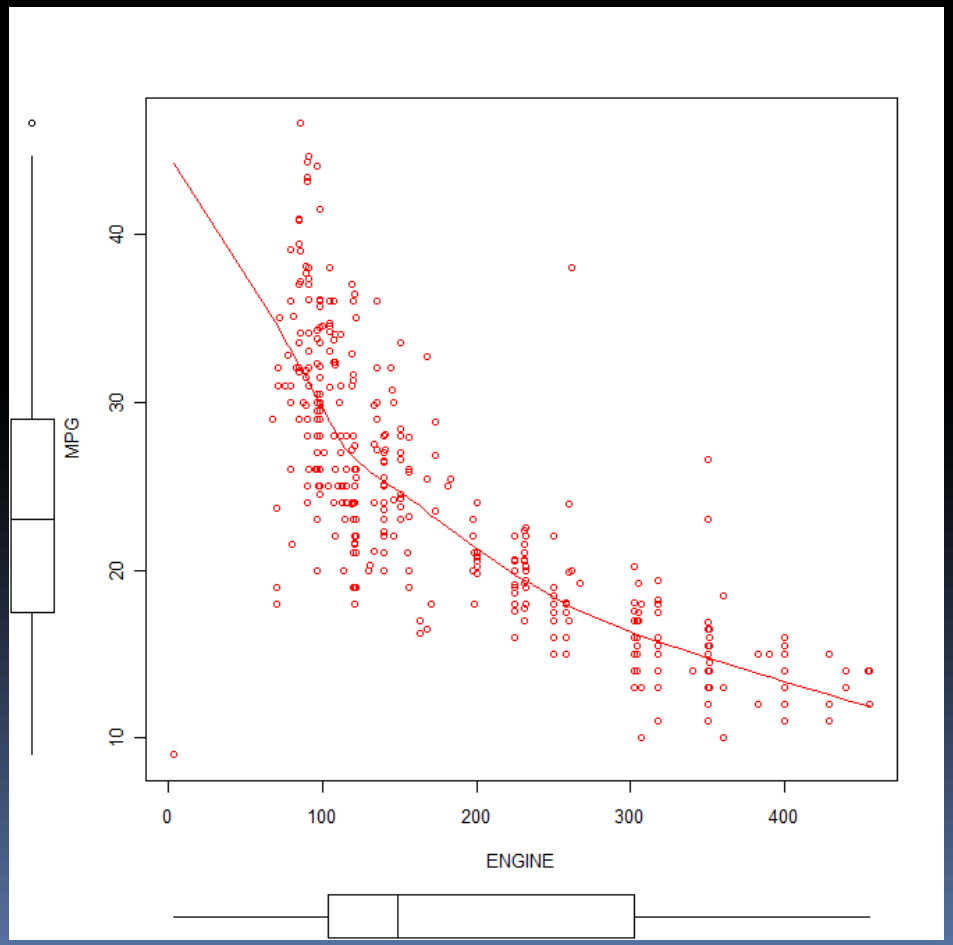
What assumptions does Pearson's correlation make?

- The variables must be approximately normally distributed
- There is a linear relationship between the two variables
- The variables must be either interval or ratio measurements
- Outliers are either kept to a minimum or are removed entirely
- There is homoscedasticity of the data

Linearity

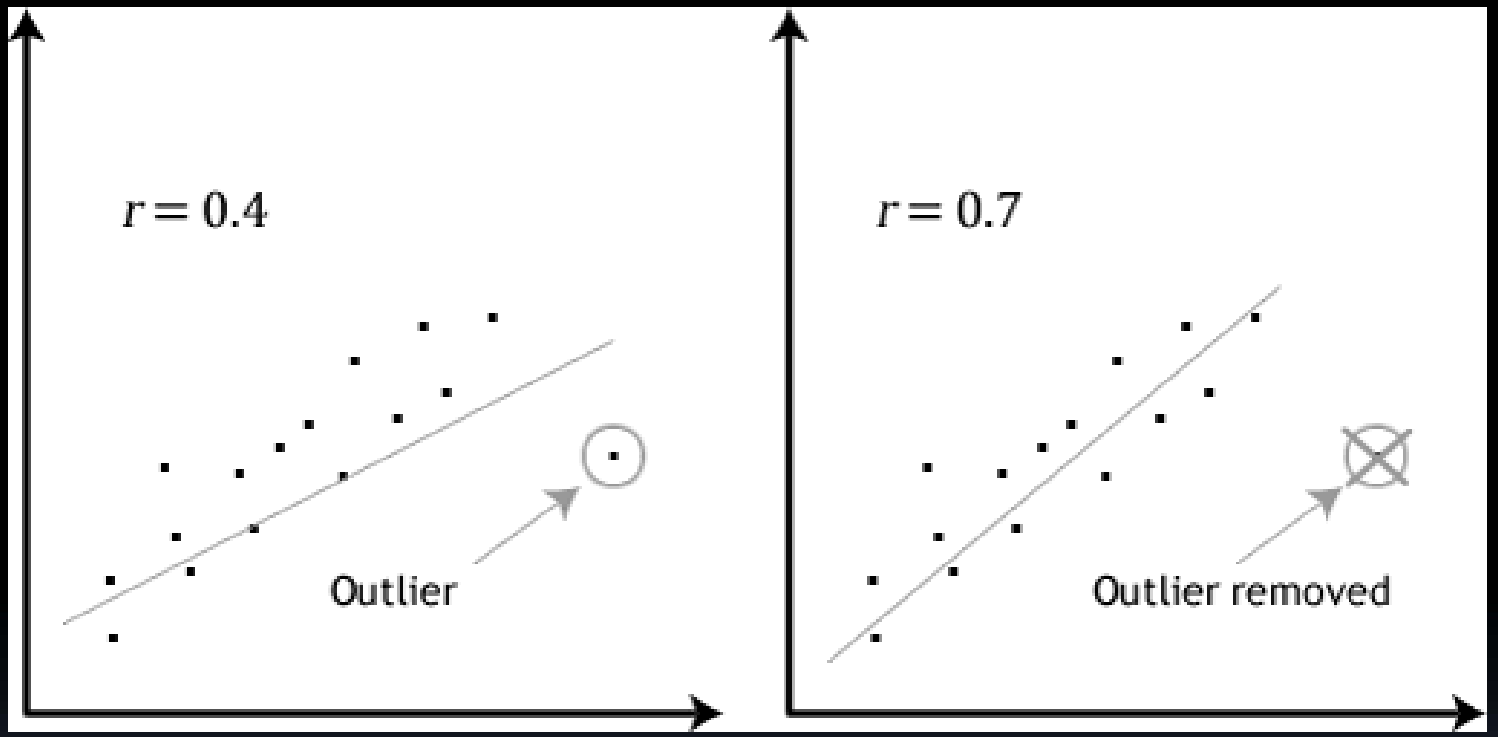
- Nonlinear relationships will have an adverse effect on a measure designed to find a linear relationship





Effect of Outliers

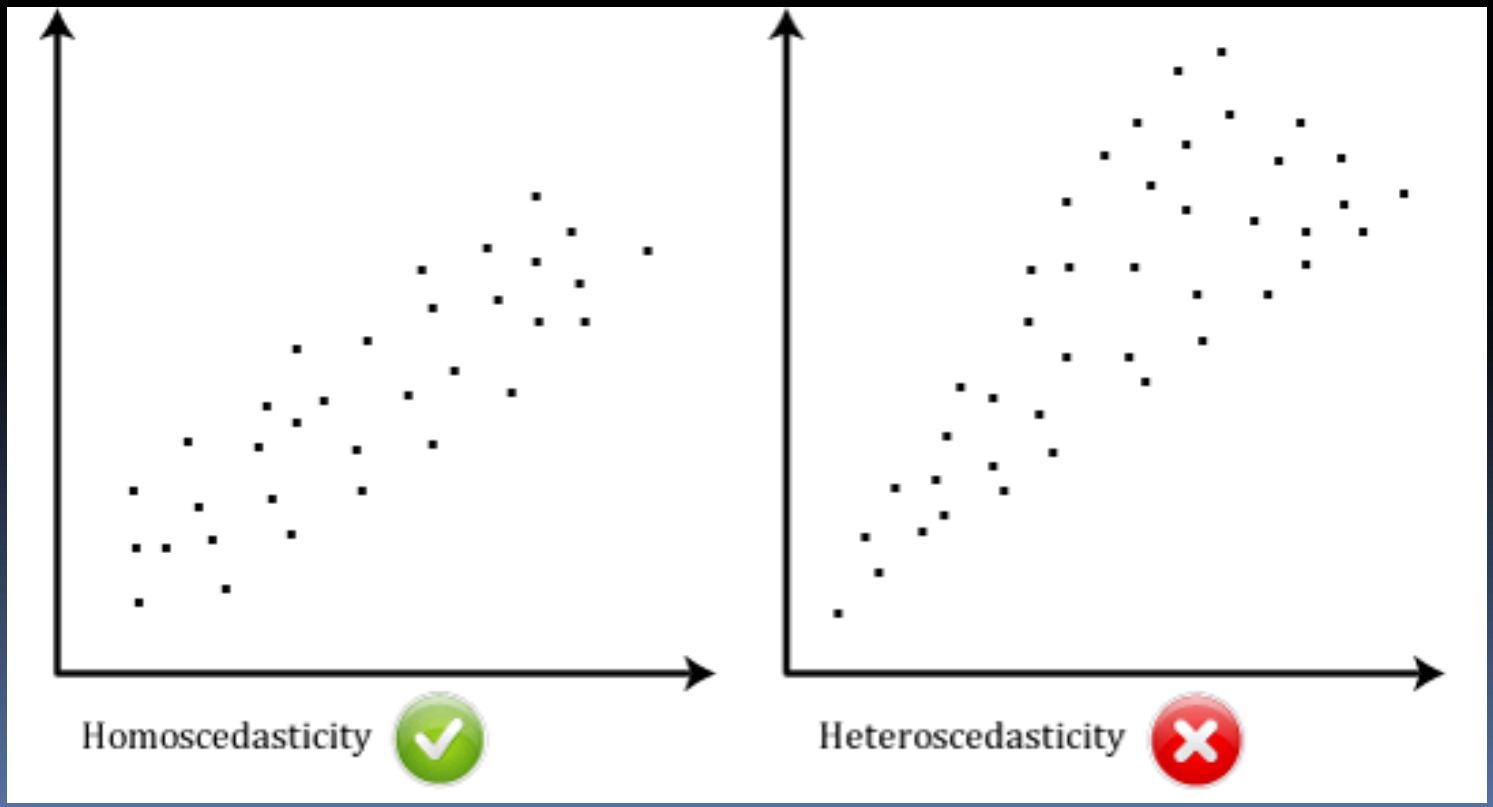
- Outliers can artificially increase or decrease *r*



- Options
  - Compute *r* with and without outliers
  - Transform variables (last resort)

What is homoscedasticity?

- variances along the line of best fit remain similar as you move along the line
- Homoscedasticity is most easily demonstrated diagrammatically as below:



Advantages of correlational studies

- Show the amount (strength) of relationship present
- Can be used to make predictions about the variables studied
- Often easier to collect correlational data, and interpretation is fairly straightforward.

Disadvantages of correlational studies

- Can't assume that a cause-effect relationship exists
- Little or no control (experimental manipulation) of the variables is usually seen
- Relationships may be accidental or due to a third variable, unmeasured factor

Common Questions on Correlation

Q1. Can you use any type of variable for Pearson's correlation coefficient?

A: **No**, the two variables have to be measured on either an interval or ratio scale. However, both variables do not need to be measured on the same scale (e.g., one variable can be ratio and one can be interval)

Common Questions on Correlation

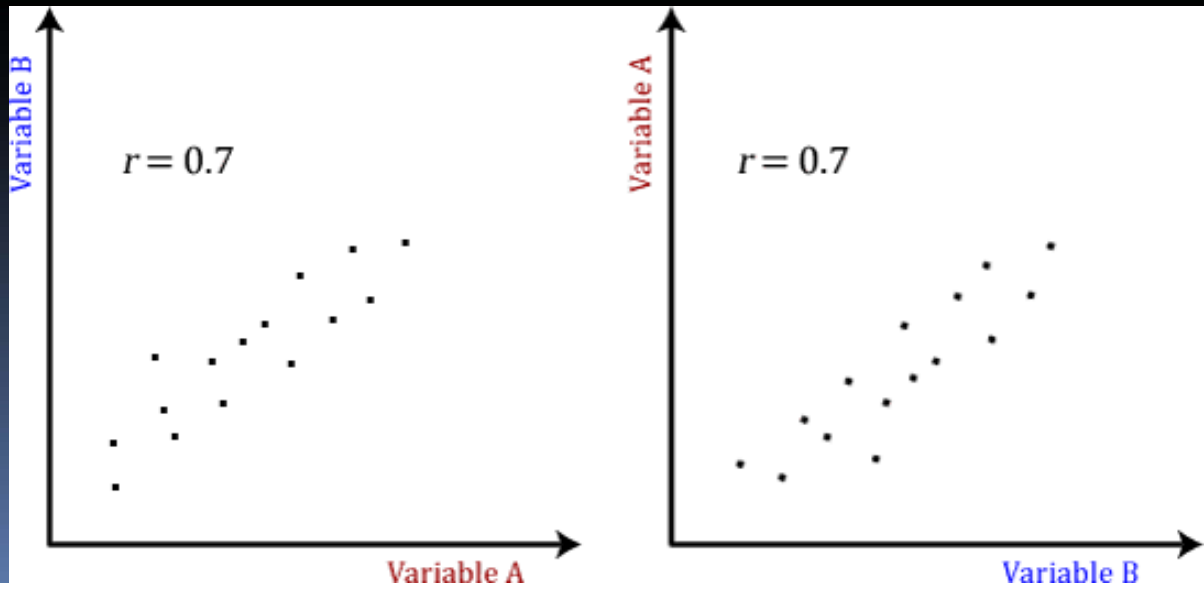
Q2. Do the two variables have to be measured in the same units?

A: **No**, the two variables can be measured in entirely different units. For example, you could correlate a person's age (measured in years) with their blood sugar levels (measured in mmol/L).

Common Questions on Correlation

Q3. What about dependent and independent variables?

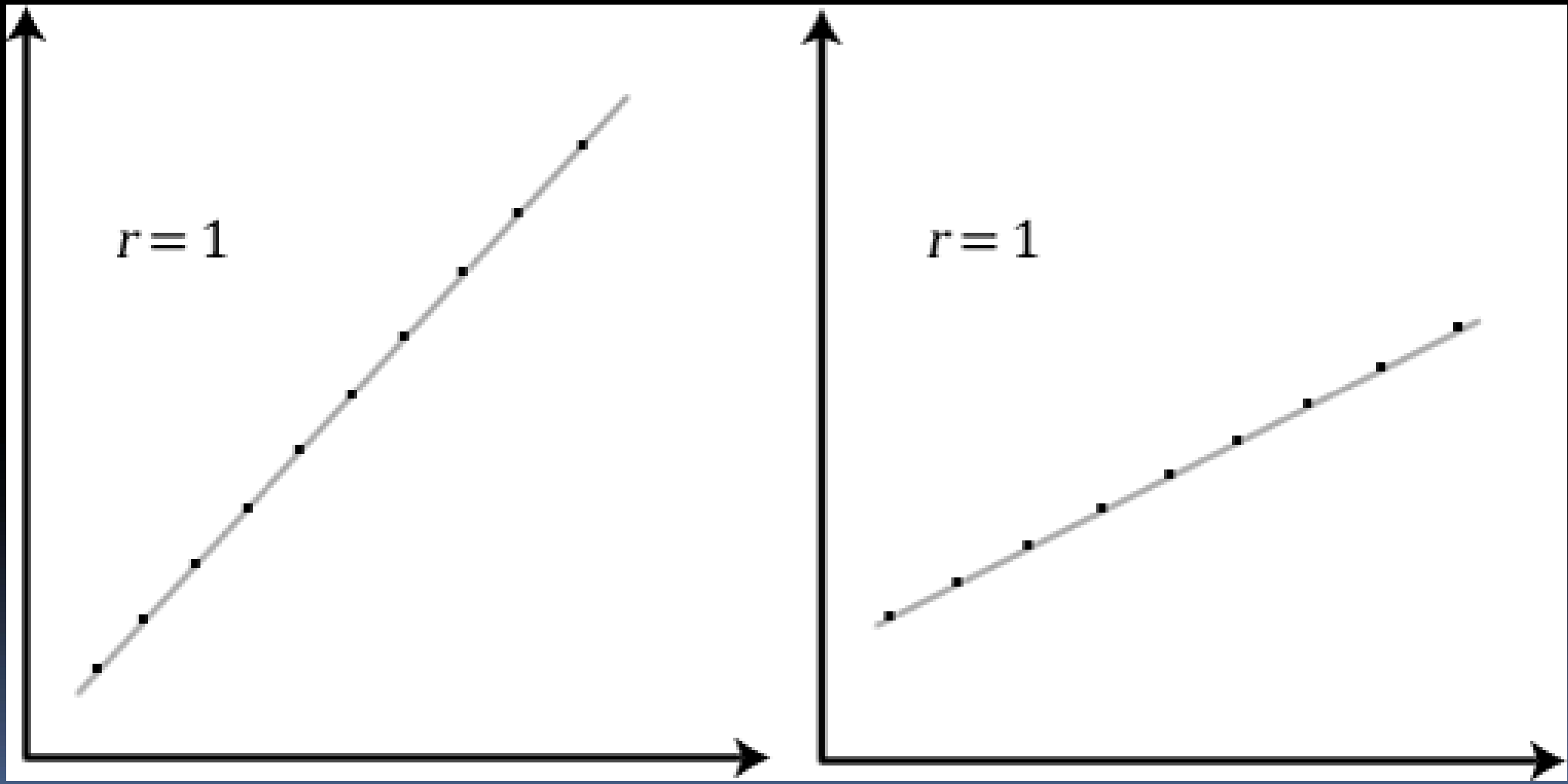
A: The Pearson product-moment correlation does not take into consideration whether a variable has been classified as a dependent or independent variable.



Common Questions on Correlation

Q4. Does the Pearson correlation coefficient indicate the slope of the line?

A: It is important to realise that the Pearson correlation coefficient, *r*, does not represent the slope of the line of best fit. Therefore, if you get a Pearson correlation coefficient of +1 this does not mean that for every unit increase in one variable there is a unit increase in another. It simply means that there is no variation between the data points and the line of best fit.


$r = 1$
$r = 1$

How to compute the simple correlation coefficient (r)

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)\left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}}$$

Example :

A sample of 6 children was selected, data about their age in years and weight in kilograms was recorded as shown in the following table . It is required to find the correlation between age and weight.

serial No	Age (years)	Weight (Kg)
1	7	12
2	6	8
3	8	12
4	5	10
5	6	11
6	9	13

These 2 variables are of the quantitative type, one variable (Age) is called the independent and denoted as (X) variable and the other (weight) is called the dependent and denoted as (Y) variables to find the relation between age and weight compute the simple correlation coefficient using the following formula:

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right).\left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}}$$

Serial n.	Age (years) (x)	Weight (Kg) (y)	xy	X^2	Y^2
1	7	12	84	49	144
2	6	8	48	36	64
3	8	12	96	64	144
4	5	10	50	25	100
5	6	11	66	36	121
6	9	13	117	81	169
Total	∑x= 41	∑y= 66	∑xy= 461	∑x2= 291	∑y2= 742

$$r = \frac{461 - \frac{41 \times 66}{6}}{\sqrt{\left[291 - \frac{(41)^2}{6}\right]\left[742 - \frac{(66)^2}{6}\right]}}$$

r = 0.759

strong direct correlation

EXAMPLE:

Relationship between Anxiety and Test Scores

Anxiety (X)	Test score (Y)	X^2	Y^2	XY
10	2	100	4	20
8	3	64	9	24
2	9	4	81	18
1	7	1	49	7
5	6	25	36	30
6	5	36	25	30
$\sum X = 32$	$\sum Y = 32$	$\sum X^2 = 230$	$\sum Y^2 = 204$	$\sum XY = 129$

Calculating Correlation Coefficient

$$r = \frac{(6)(129) - (32)(32)}{\sqrt{\left(6(230) - 32^2\right)\left(6(204) - 32^2\right)}} = \frac{774 - 1024}{\sqrt{(356)(200)}} = -.94$$

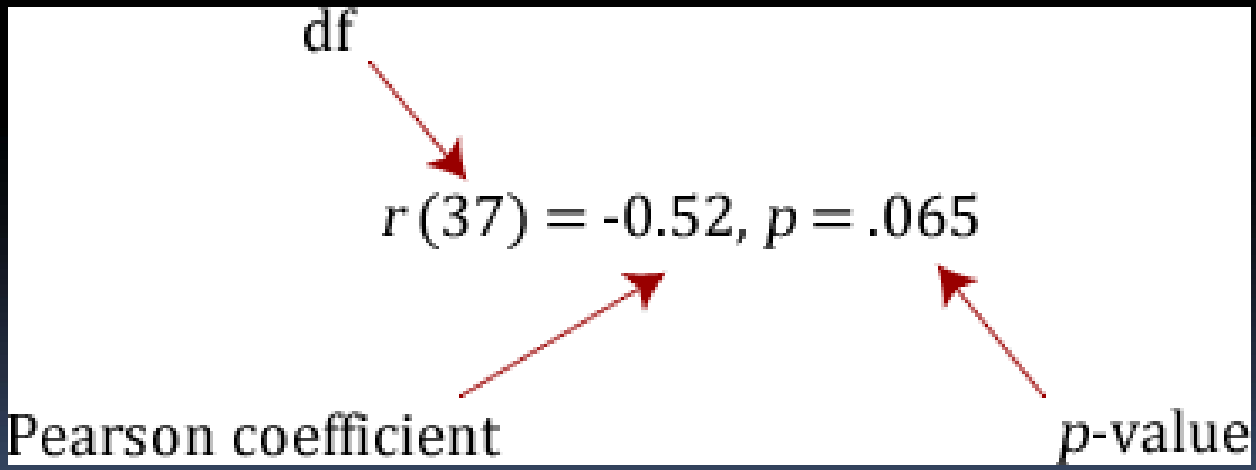
r = - 0.94

Indirect strong correlation

How do I Report the output of correlation in Dissertation/Thesis

Dentene lead levels correlated well and inversely with family income, indicating that poorer children have higher levels of lead in their systems

- n =
- r (df) =
- p =



degrees of freedom (df) is the number of data points minus 2 (*N* - 2)

Other Kinds of Correlation

- Spearman Rank Correlation Coefficient

  ρ [rho] (or) r_s

- Kendall's Rank Correlation Coefficient

  tau (τ)

- Point Biserial Correlation Coefficient (r_{pb})
- Point Multiserial Correlation Coefficient

Confusing Point … Beware …

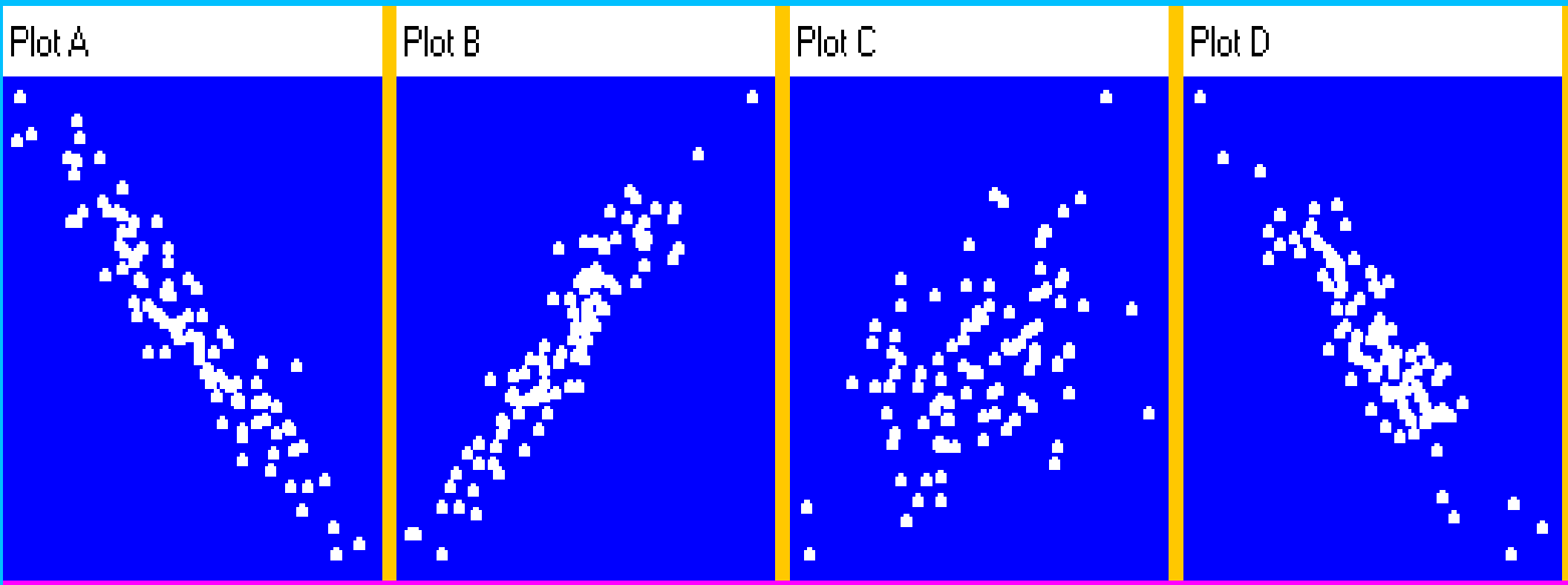
Related to, but different from Correlation is

Measurement of Agreement

- Kappa Statistics (κ) calculating inter-rater reliability.

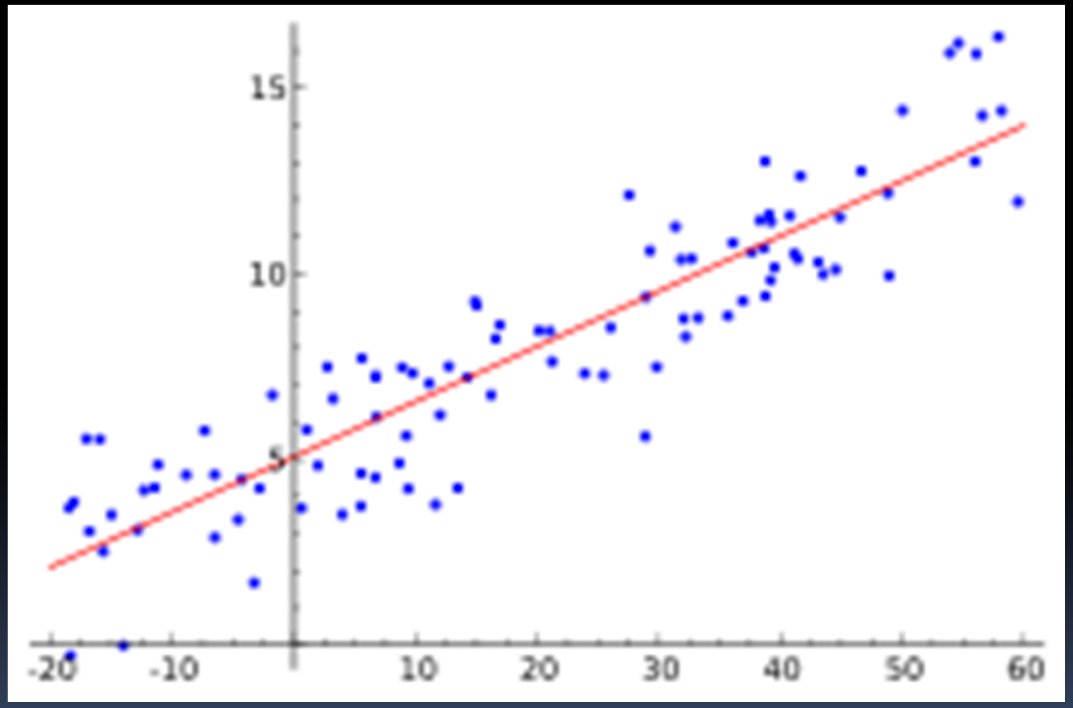
  - Agreement Between Diagnosis of Chest X-ray by Radiologist 1 and Radiologist 2
  - Agreement Between Diagnosis of mp slide by Microbiologist 1 and Microbiologist 2

Exercise



r = -0.95	✓ A	B	C	D
r = -0.9	A	B	C	✓ D
r = 0.49	A	B	✓ C	D
r = 0.93	A	✓ B	C	D

Regression Analyses



History

- The term "regression" was coined by Francis Galton in the nineteenth century to describe a biological phenomenon.
- The phenomenon was that the heights of descendants of tall ancestors tend to regress down towards a normal average (a phenomenon also known as regression toward the mean)

Regression Analyses

- Regression: technique concerned with predicting or estimating some variables by knowing others
- The process of predicting variable Y using variable X
- Tells you how values in y change as a function of changes in values of x

Regression Analysis

- An area of statistics that attempts to predict or estimate the value of a response (dependent) variable from the know values of one or more explanatory variable (independent) variables.

Correlation or Regression

- Correlation describes the strength of a **linear** relationship between two variables
- **Linear** means "**straight line**"
- **Regression** tells us how to draw the straight line described by the correlation

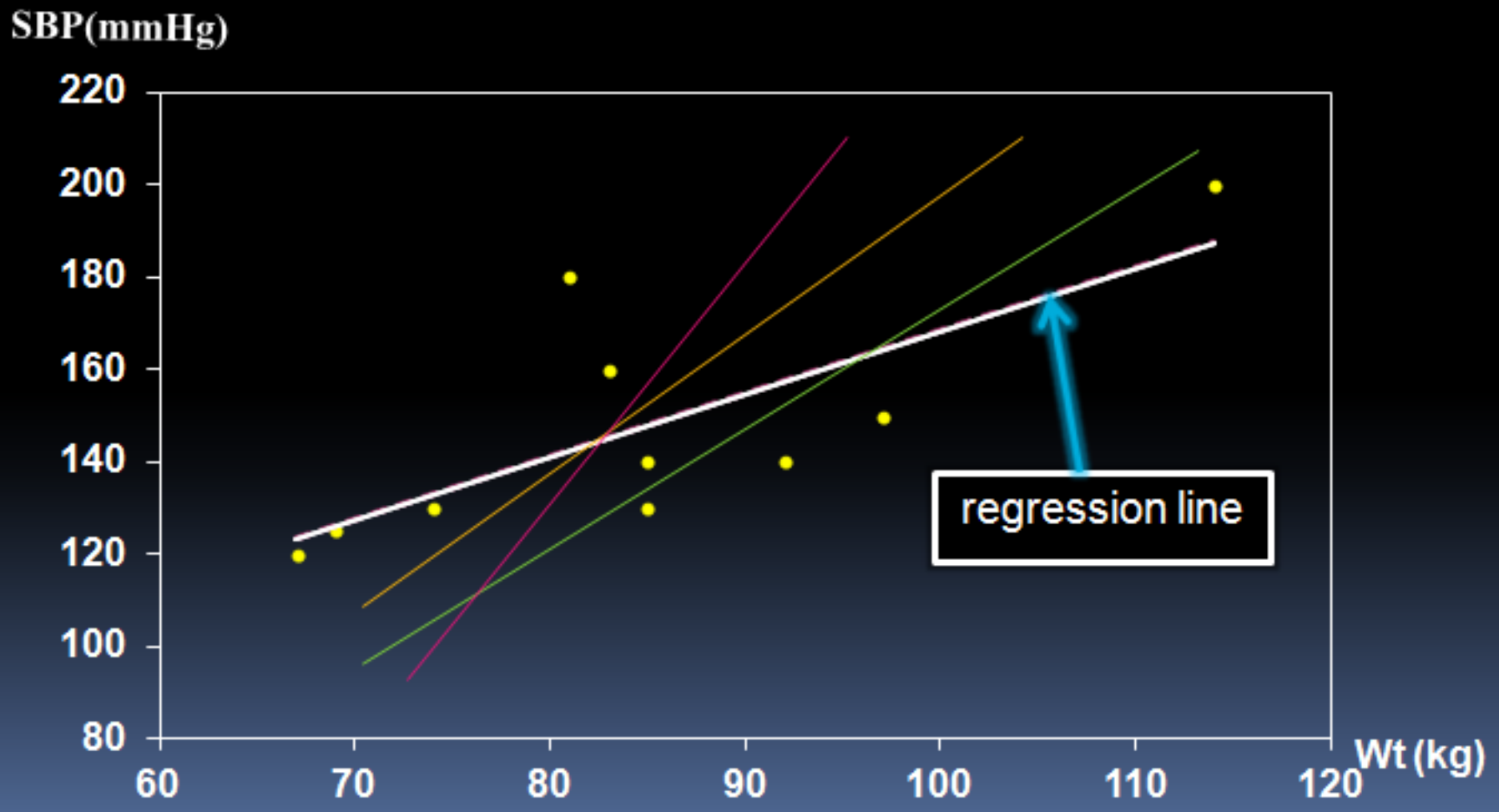
Correlation or Regression

In correlation, there is no question of a dependence relationship: issue of interest is, " Are the two variables related?"

In regression, there is a clear dependent / explanatory relationship between the two variables explaining or describing this relationship is key goal

Regression Analysis

The regression line, "best-fit" line, "least squared line"



By using the **least squares method** (a procedure that minimizes the vertical deviations of plotted points surrounding a straight line) we are able to construct a best fitting straight line to the scatter diagram points and then formulate a regression equation in the form of:

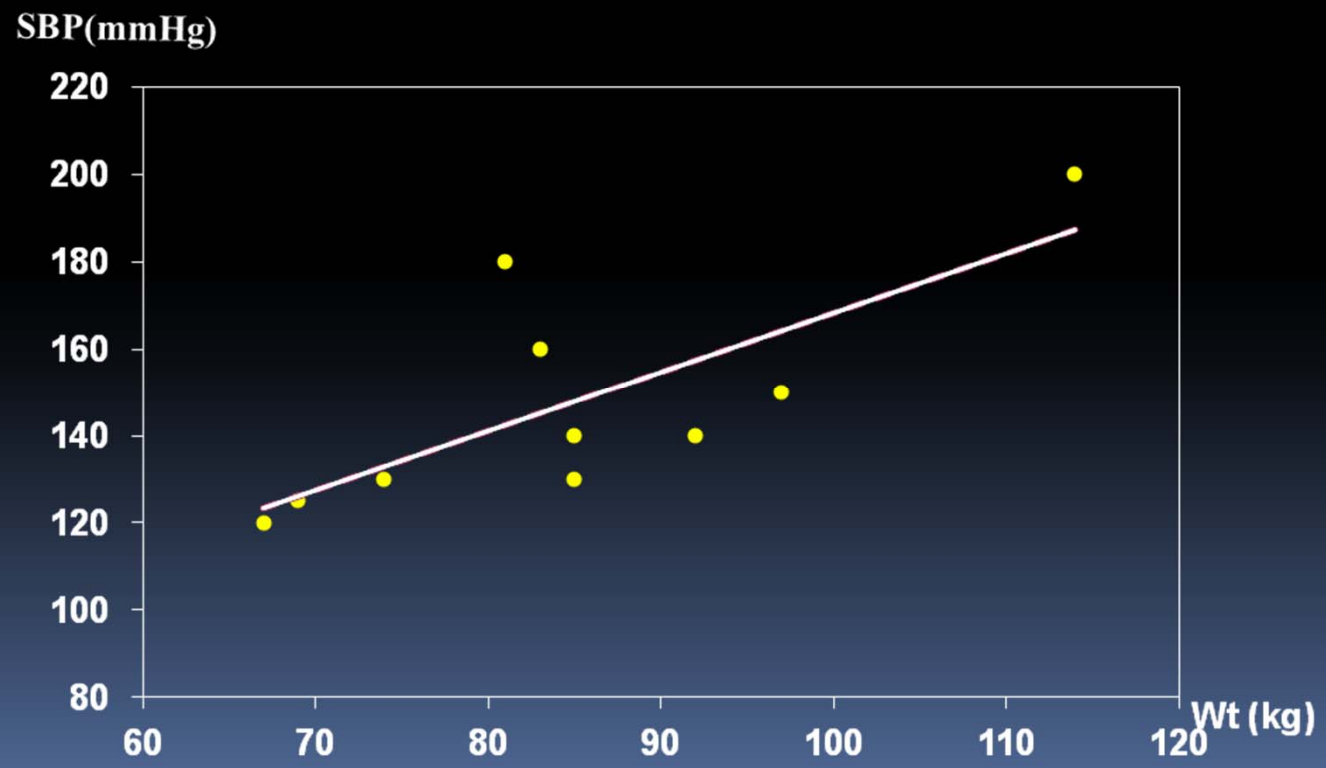
$$\hat{y} = \bar{y} + b(x - \bar{x})$$

$$\hat{y} = a + bX$$

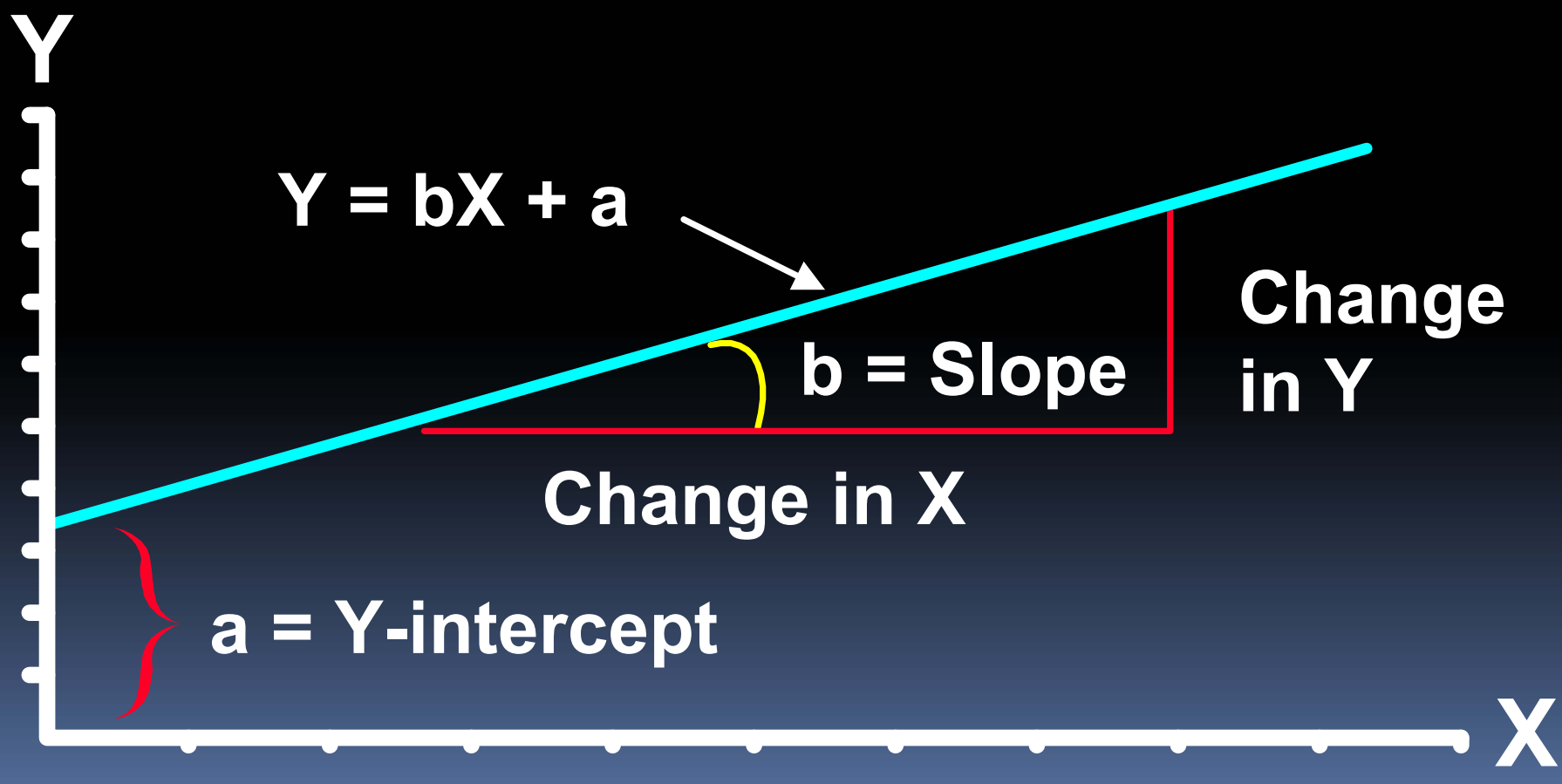
$$b = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

Regression Equation

- Regression equation describes the regression line mathematically
  - Intercept
  - Slope



Linear Equations



Regression Equation

$$y = a + bx$$

y = dependent variable

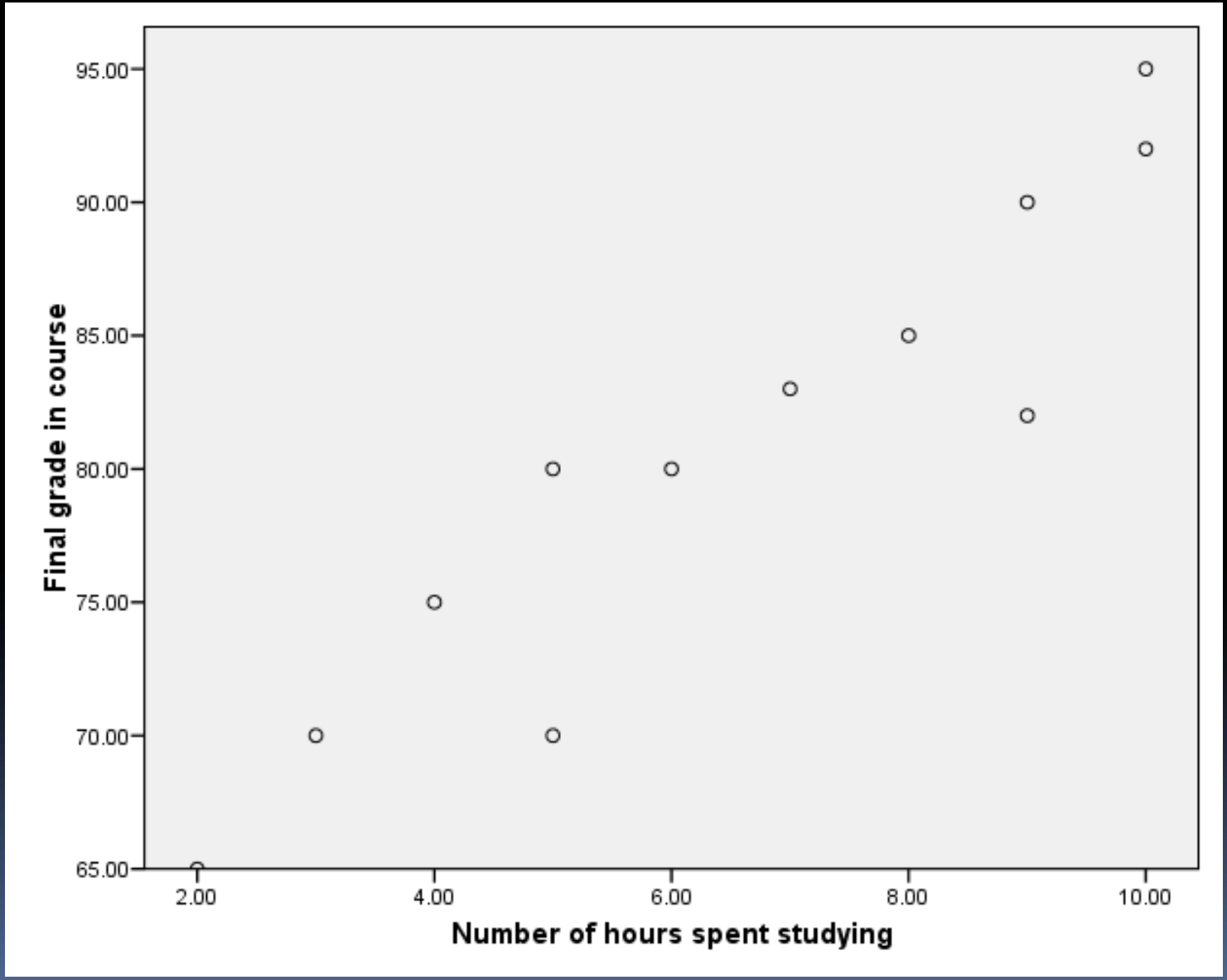
a = intercept of the regression line (value when y takes whenever x is zero)

b = slope of the regression line (regression coefficient)

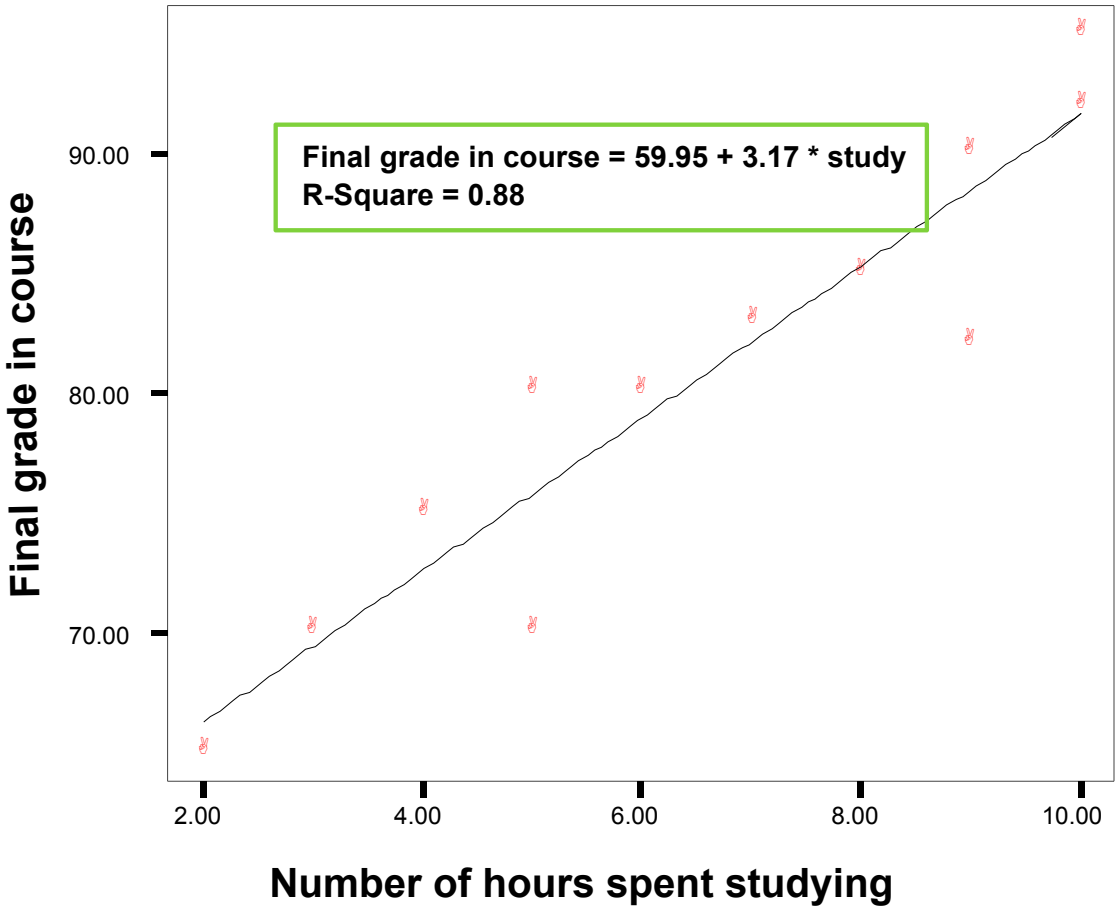
= change in y value for each unit change in x value.

x = independent variable

Hours studying and grades



Regressing grades on study hours



Predicted final grade in class =
59.95 + 3.17*(number of hours you study per week)

```
Predicted final grade in class =
59.95 + 3.17*(hours of study)
```

Predict the final grade of…

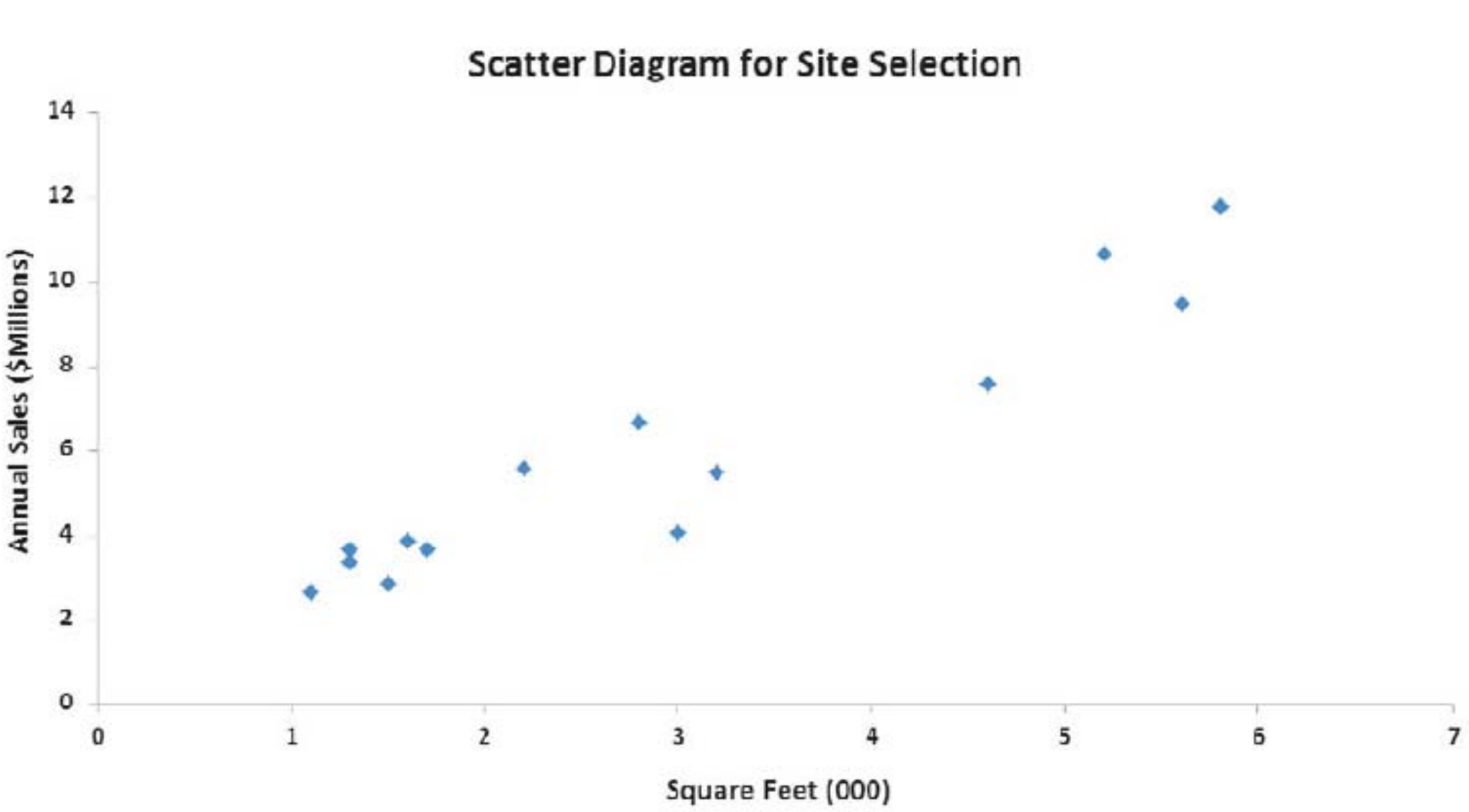
- Someone who studies for 12 hours
- Final grade = 59.95 + (3.17*12)
- Final grade = 97.99

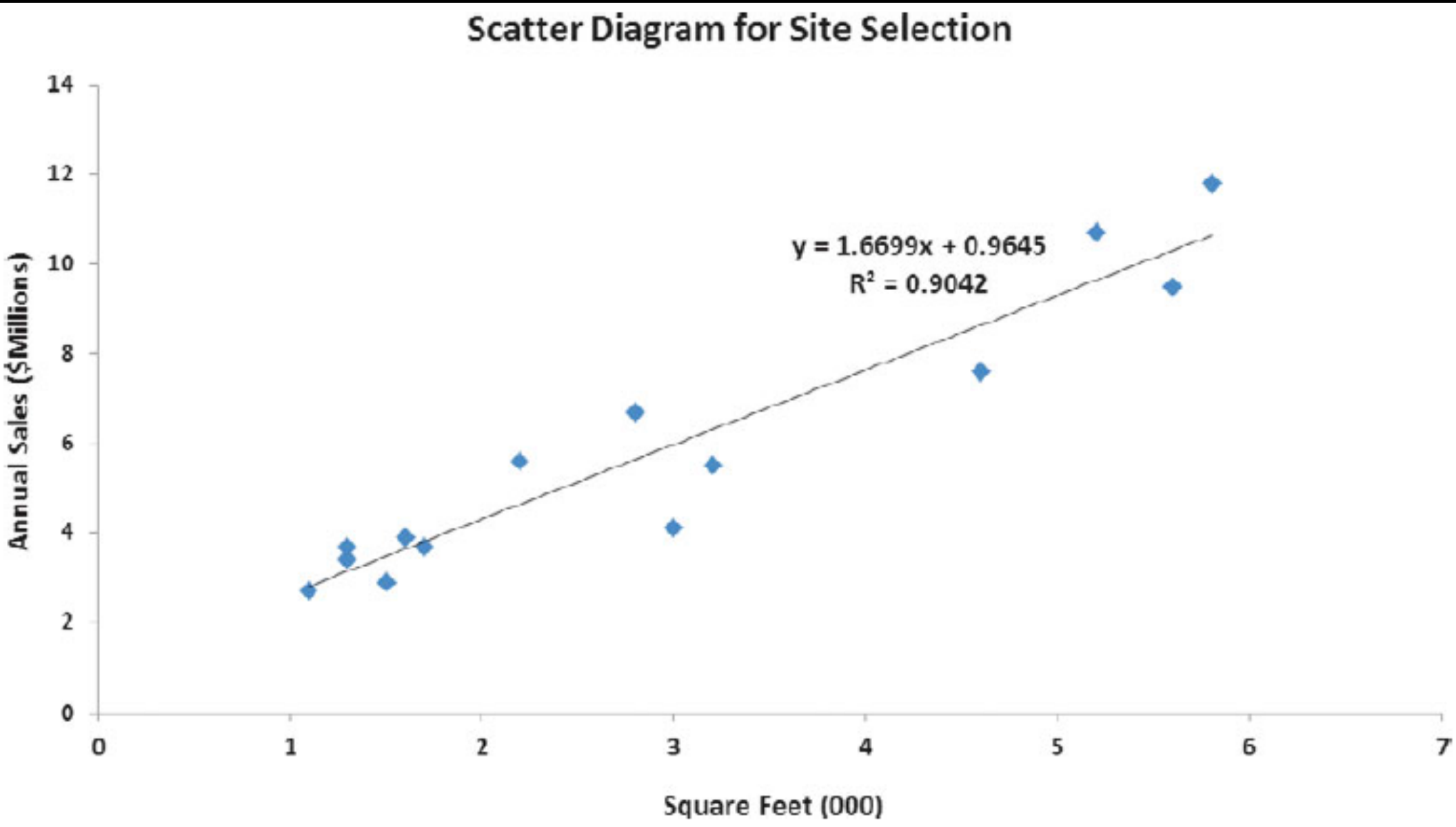
- Someone who studies for 1 hour:
- Final grade = 59.95 + (3.17*1)
- Final grade = 63.12

Example 2

One business director of planning is to forecast annual sales for all new stores, based on store size. To examine the relationship between the store size in square feet and its annual sales, data were collected from a sample of 14 stores.

Store	Square Feet (Thousands)	Annual Sales (in Millions of Dollars)	Store	Square Feet (Thousands)	Annual Sales (in Millions of Dollars)
1	1.7	3.7	8	1.1	2.7
2	1.6	3.9	9	3.2	5.5
3	2.8	6.7	10	1.5	2.9
4	5.6	9.5	11	5.2	10.7
5	1.3	3.4	12	4.6	7.6
6	2.2	5.6	13	5.8	11.8
7	1.3	3.7	14	3.0	4.1


Scatter Diagram for Site Selection
Annual Sales ($Millions)
14
12
10
8
6
4
2
0
0
1
2
3
4
5
6
7
Square Feet (000)


Scatter Diagram for Site Selection
y = 1.6699x + 0.9645
R² = 0.9042
Annual Sales ($Millions)
Square Feet (000)
0
2
4
6
8
10
12
14
0
1
2
3
4
5
6
7

Annual Sale =

0.9645 + 1.6699*(Square Feet)

- The slope, b1 is +1.6699. This means that for each increase of 1 unit in X, the predicted value of Y is estimated to increase by 1.6699 units.
- In other words, for each increase of 1.0 thousand square feet in the size of the store, the predicted annual sales are estimated to increase by 1.6699 millions of dollars.
- Thus, the slop represents the portion of the annual sales that are estimated to vary according to the size of the store.

Be aware …
Predictions in Regression Analysis:

Two words of caution:

- A straight line should be fitted only if the scatter diagram suggest that the relationship between the two variables is **roughly linear**.
- It is dangerous to **extrapolate** the regression line **out side the range of the data**. In out example, extrapolating the line to an income of $2000 per would yield an estimated mean weight of 34.8 Kg, which is of course absurd.

The Coefficient of Determination (r^2)

- The coefficient of determination measures the proportion of variation in Y that is explained by the variation in the independent variable X in the regression model.

- It is a measure of the "goodness-of-fit" of the model to the data

- Coefficient of determination measures the proportion (%) of valid correlation against the variability (i.e by chance) in variable y accounted for by the linear relationship with variable x.

From the example of Final Grade and Study Hour, the calculation gives;

$$r^2 = 0.88 = 88\ \%$$

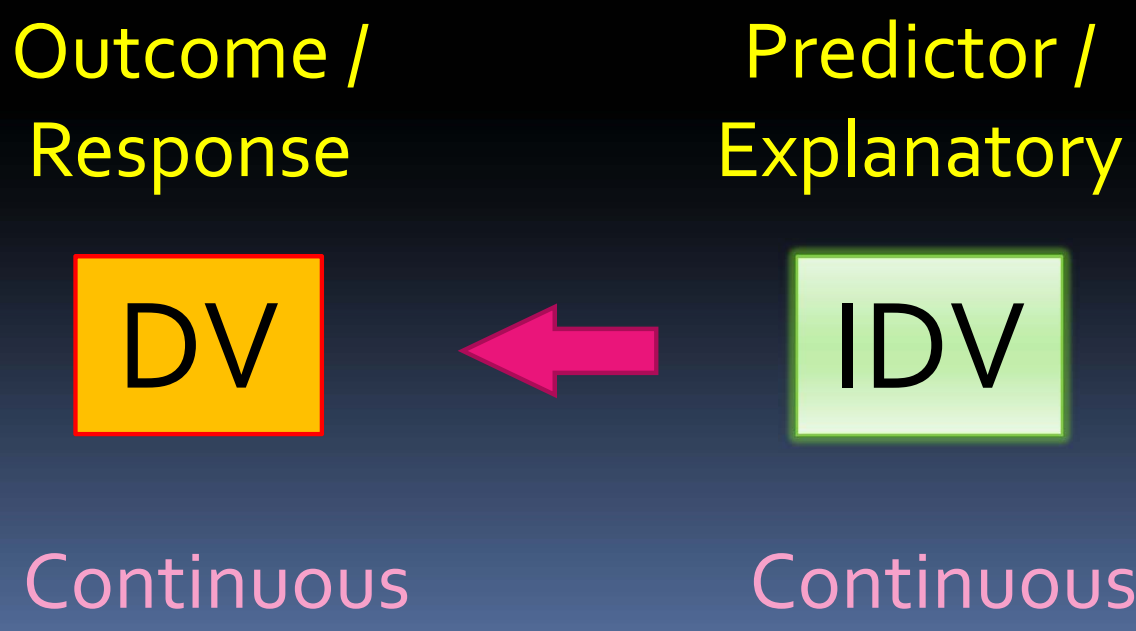
which indicates the valid correlation between Final Grade and Study Hour corresponds to 88% and variation by chance is 12%

Several Types of Regression Analyses

- Simple Linear Regression
- Multiple Liner Regression
- Simple Logistic Regression
- Multiple Logistic Regression
- Cox-proportional Hazards Regression

Simple Linear Regression

- Relationship between a single continuous explanatory variable and a single continuous response variable that varies linearly over a range of value



Multiple Linear Regression

- Linear relationship between two or more continuous or categorical explanatory variables and single continuous response variable

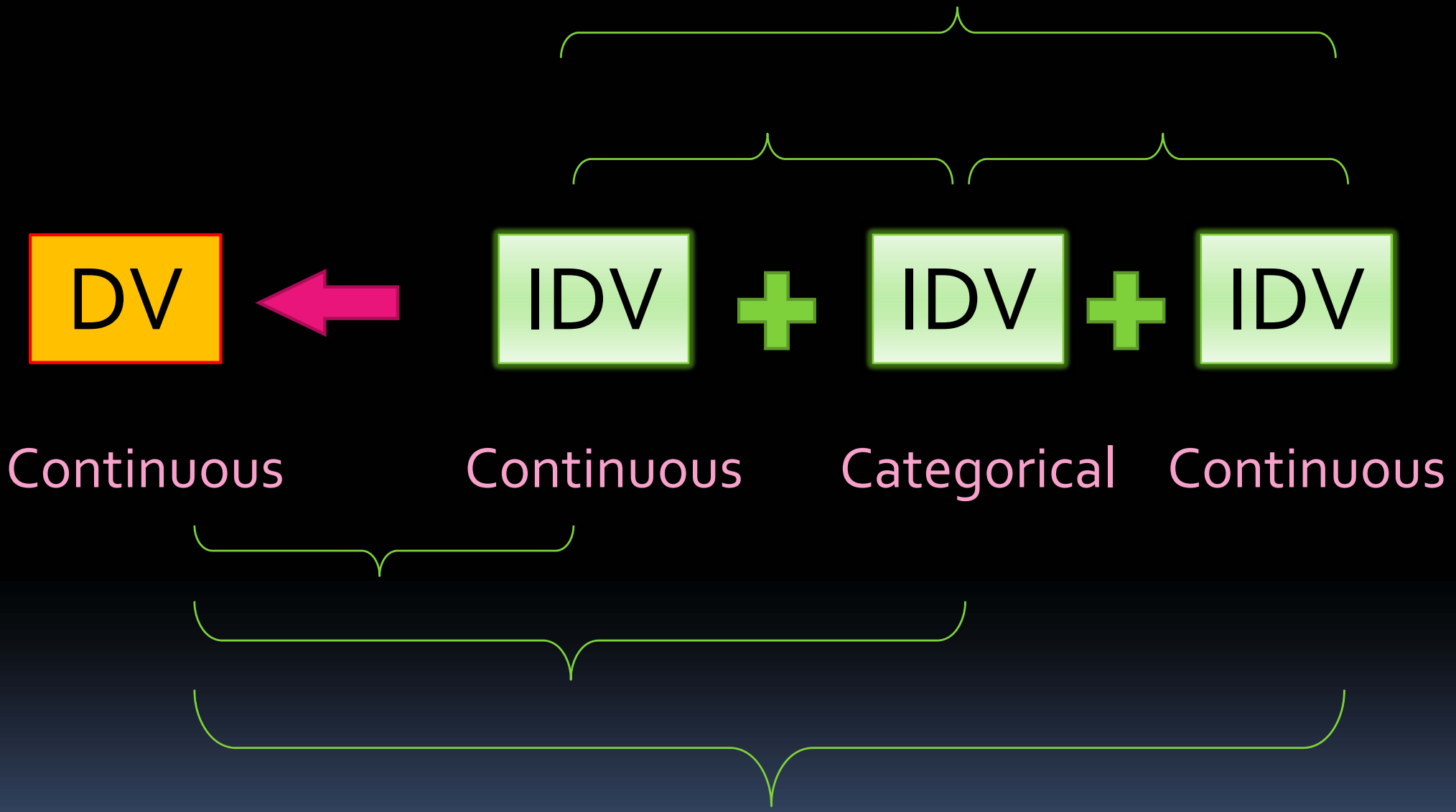
Multiple Linear Regression (Example)

SBP (mmHg) = 4 + (1.5 x **Age**) +
(0.333 x **Hip girth**) +
(10 x **sex**) +
(10 x **smoker**)

If the patient is a 50-year-old non-smoking male, with a hip girth of 240 cm, then you would estimate his systolic blood pressure to be:

SBP = 4 + (1.5 x 50) + (0.333 x 240)+ (10 x 1)+ (10 x 0)

= 149mmHg


DV
IDV
IDV
IDV
Continuous
Continuous
Categorical
Continuous

Simple Logistic Regression

- Relationship between a single continuous or categorical variable and a single categorical response variable, usually a binary variable, such as whether or not a heart attack has occurred

Binary

Continuous (or) Categorical

DV ← IDV

Yes / No
Alive / Death
High / Low
D^+ / D^-

Simple Logistic Regression (Example)

- Among 453 patients with either high serum levels (>220 mg/dL) or low serum levels (≤ 220 mg/dL), weight prove to be a significant predictor for serum levels.

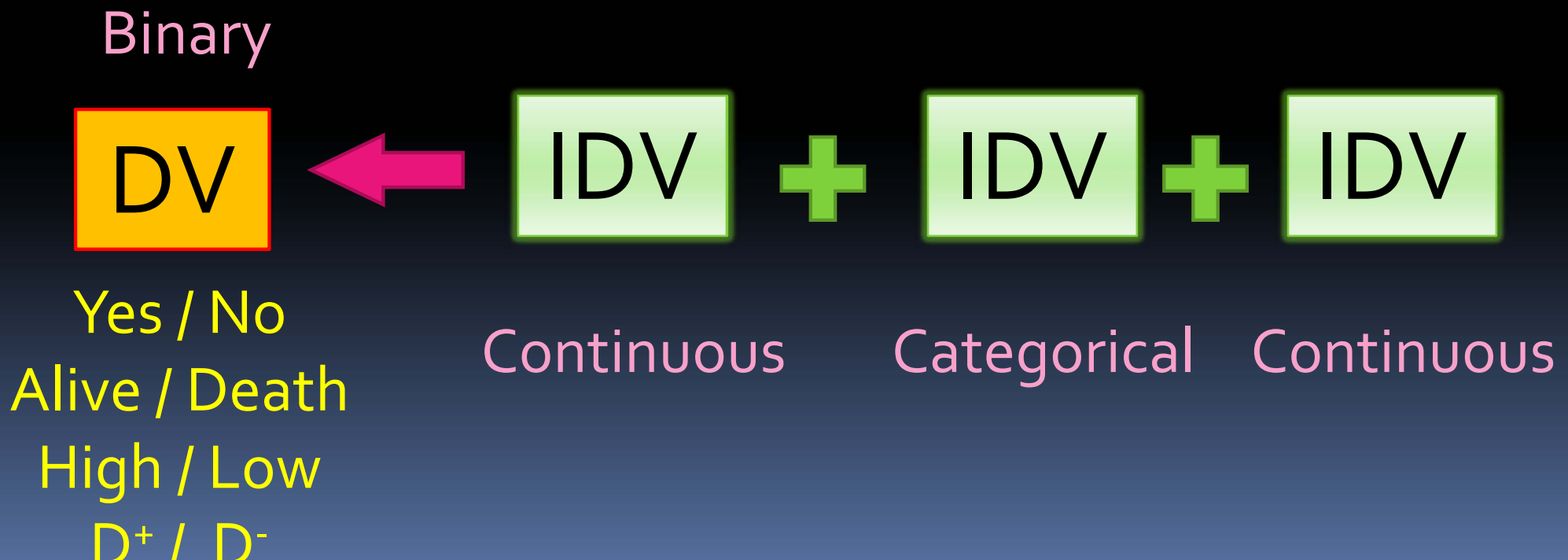
Variable	Coefficient (β)	Standard Error	Wald $\chi 2$	P value	Odds Ratio	95 % CI
Intercept	- 1.89	0.48	-	-	-	-
Weight	0.44	0.11	16.0	<0.001	1.55	1.25 to 1.93

Hint : $e^{0.44} = 1.55$

Serum (high/low) = -1.89 + (1.55 x weight)

Multiple Logistic Regression

- Relationship between two or more continuous or categorical explanatory variables and a single categorical response variable



Overweight Children and Adolescents : A Risk Group for Iron Deficiency

Nead Karen G et al.

	Odds Ratio	95% CI	P Value
Age			
2-5 y	1.2	0.7-2.2	.4
6-11 y	1.0		
12-16 y	2.6	1.4-4.6	.002
Gender			
Female	2.6	1.7-3.9	< .0001
Male	1.0		
Race/ethnicity			
White	1.0		
Black	1.3	0.7-2.2	.3
Mexican American	3.5	2.2-5.8	< 0.0001
Other	3.8	2.0-7.1	< 0.0001
Poverty status			
Below poverty line	2.0	0.9-4.2	.06
Above poverty line	1.0		
Care-taker education			
< 12th grade	0.7	0.3-1.5	.4
12th grade	1.0	0.5-2.0	.9
> 12th grade	1.0		

Prevalence and risk factors of diabetes mellitus among middle-aged school teachers in Mandalay City

Variables	Total n=522	No. of diabetes (%)	OR (95%CI)	p
Age (years)				**0.004**
<=45	127	10 (7.9)	1	
>45	395	73 (18.5)	2.65 (1.30-5.95)	
Sex				0.231
Male	67	14 (20.9)	1	
Female	455	69 (15.2)	0.68 (0.35-1.40)	
Marital status				**0.024**
Single	209	24 (11.5)	1	
Ever married	313	59 (18.8)	1.79 (1.05-3.12)	
History of delivering big baby				
No	212	33 (15.6)	1	0.112
Yes	50	13 (26.0)	1.9 (0.84 - 4.15)	
Other (male, un-married women)	260	37 (14.2)	0.9 (0.52 - 1.55)	
Hypertension				**0.031**
No	372	51 (13.7)	1	
Yes	150	32 (21.3)	1.7 (1.01-2.85)	
Family history in father				**0.009**
No	469	68(14.5)	1	
Yes	53	15(28.3)	2.32 (1.12-4.61)	
Family history in mother				0.231
No	466	71(15.2)	1	
Yes	56	12(21.4)	1.52 (0.69-3.1)	
Family history in siblings				**<0.001**
No	451	59(13.1)	1	
Yes	71	24(33.8)	3.38 (1.84-6.13)	
Family history in paternal side				**0.018**
No	478	70(14.6)	1	
Yes	44	13(29.5)	2.44 (1.11-5.08)	
Family history in maternal side				0.058
No	475	71(14.9)	1	
Yes	47	12(25.5)	1.95 (0.88-4.07)	
Smoking				0.989
No	497	79(15.9)	1	
Yes	25	4 (16)	1.01 (0.24-3.1)	
Alcohol				0.773
No	488	77 (15.8)	1	
Yes	34	6 (17.6)	1.14 (0.37-2.94)	
Exercise				0.405
No	372	56(15.1)	1	
Yes	150	27(18.0)	1.24 (0.72-2.1)	
BMI				0.013
Normal	335	42(12.5)	1	
Pre-obese	162	34 (21.0)	1.85 (1.09-3.13)	
Obese	25	7 (28)	2.7 (0.9-7.3)	

Khin Phyu Phyu, Maung Maung, Lei Lei Win, Khaing Khaing Mar, Zaw Win Tun, Mya Marlar, **Win Khaing**, **Khin Than Aye** & **Kyaw Zin Thant**

MHSR Journal Vol 22, Number 3, December 2010

Multiple Logistic Regression Model

Factors	Adj. OR (95%CI)	p (Wald's test)	p (LR test)
Age			
Above 45 vs. 40-45	2.69 (1.17, 6.15)	0.019	0.010
Gender marital: ref. = Male single			0.034
Female single	1.16 (0.14, 9.6)	0.888	
Male ever-married	4.39 (0.51, 37.96)	0.179	
Female ever-married without history of big baby	1.36 (0.17, 11.11)	0.774	
Female ever-married with history of big baby	2.56 (0.29, 22.86)	0.399	
Family history in siblings ref.=Yes vs. No	2.39 (1.16, 4.91)	0.018	0.024
BMI ref =Normal			0.017
Pre-obese	2.08 (1.15, 3.77)	0.015	
Obese	3.35 (1.1, 10.18)	0.033	

Cox-proportional Hazards Regression

- An aspect of time-to-event (survival) analysis, is used to assess the relationship between two or more continuous or categorical explanatory variable and a single continuous response variable (the time to the event)

$$\text{Hazard} = h_0 + e^{(\beta_1 X_1 + \beta_2 X_2 + \ldots)}$$

Laparoscopy-assisted colectomy versus open colectomy for treatment of non-metastatic colon cancer : a randomised trial

Antonio M Lacy et al.

	Hazard ratio (95% CI)	p
Probability of being free of recurrence		
Lymph-node metastasis (presence *vs* absence)	0·31 (0·16–0·60)	0·0006
Surgical procedure (OC *vs* LAC)	0·39 (0·19–0·82)	0·012
Preoperative serum CEA concentrations (≥4 ng/mL *vs* <4 ng/mL)	0·43 (0·22–0·87)	0·018
Overall survival		
Surgical procedure (OC *vs* LAC)	0·48 (0·23–1·01)	0·052
Lymph-node metastasis (presence *vs* absence)	0·49 (0·25–0·98)	0·044
Cancer-related survival		
Lymph-node metastasis (presence *vs* absence)	0·29 (0·12–0·67)	0·004
Surgical procedure (OC *vs* LAC)	0·38 (0·16–0·91)	0·029

OC=open colectomy; LAC=laparoscopy-assisted colectomy; CEA=carcinoembryonic antigen.

Q & A Section

Exercise (1)

The birth weights of 1,333 fifty-year-old Swedish men were traced through birth records. Adult height and birth weight were significantly correlated ($r = 0.22$, $P<0.001$) (Leon et al., 1996).

a) What is meant by ‘correlated’ and ‘r = 0.22’?

b) What assumptions are required for the calculation of the P value?

c) What can we conclude about the relationship between adult height and birth weight?

Solution to Exercise (1)

a) Two variables are correlated :

one high -> other high, one low -> other low

one high -> other low, one low -> other high

r = 0.22 – positive correlation, showing that adult height tend to be greater for subjects with high birth weight, but the relationship is weak

b) Two variable must follow a Normal Distribution for the P value to be valid.

c) Adult height is related to birth weight, but the relationship is weak. However, we cannot conclude from these data that the relationship is causal.

Exercise (2)

- A Statistics professor wants to use the number of hours a student studies for a statistics final exam (X) to predict the final exam score (Y). A regression model was fit based on data collected for a class during the previous semester, with the following results:

$$Y = 35.0 + 3X$$

- What is the interpretation of the Y intercept, β_0 and the slope β_1?

Solution to Exercise (2)

- The Y intercept $\beta 0=35.0$ indicates that when the student does not study for the final exam, the predicted final exam score is 35.0
- The slope $\beta 1=3$ indicates that for each increase of one hour in studying time, the mean change in the final exam score is predicted to be +3.0. In other words, the final exam score is predicted to increase by 3 points for each one hour increase in studying time.

References

- Martin Bland and Janet Peacok, *Statistical Questions in Evidence-based medicine*, Oxford, 2000.
- Thomas A. Lang and Michelle Secic, *How to Report Statistics in Medicine*, 2nd Edition, American Collage of Physicians, 2006
- David Bowers et al, *Understanding Clinical Papers*, 2nd Edition, Wiley 2006
- David M Levine et al, *Business Statistics, A first course*, 5th Edition, Person, 2011

Any Discussion ??

Thank You