Correlation

Correlation

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 variables are related

Properties of Correlation coefficient

- The correlation coefficient lies between -1 & +1 symbolically ($-1 \le r \le 1$)
- The correlation coefficient is independent of the change of origin & scale.
- The coefficient of correlation is the geometric mean of two regression coefficient.

$$\mathbf{r} = \sqrt{\mathbf{b}_{\mathbf{yx}}\mathbf{b}_{\mathbf{xy}}}$$

- The one regression coefficient is (+ve) other regression coefficient is also (+ve) correlation coefficient is (+ve)

(i.e. Same sign)

Methods of Studying Correlation

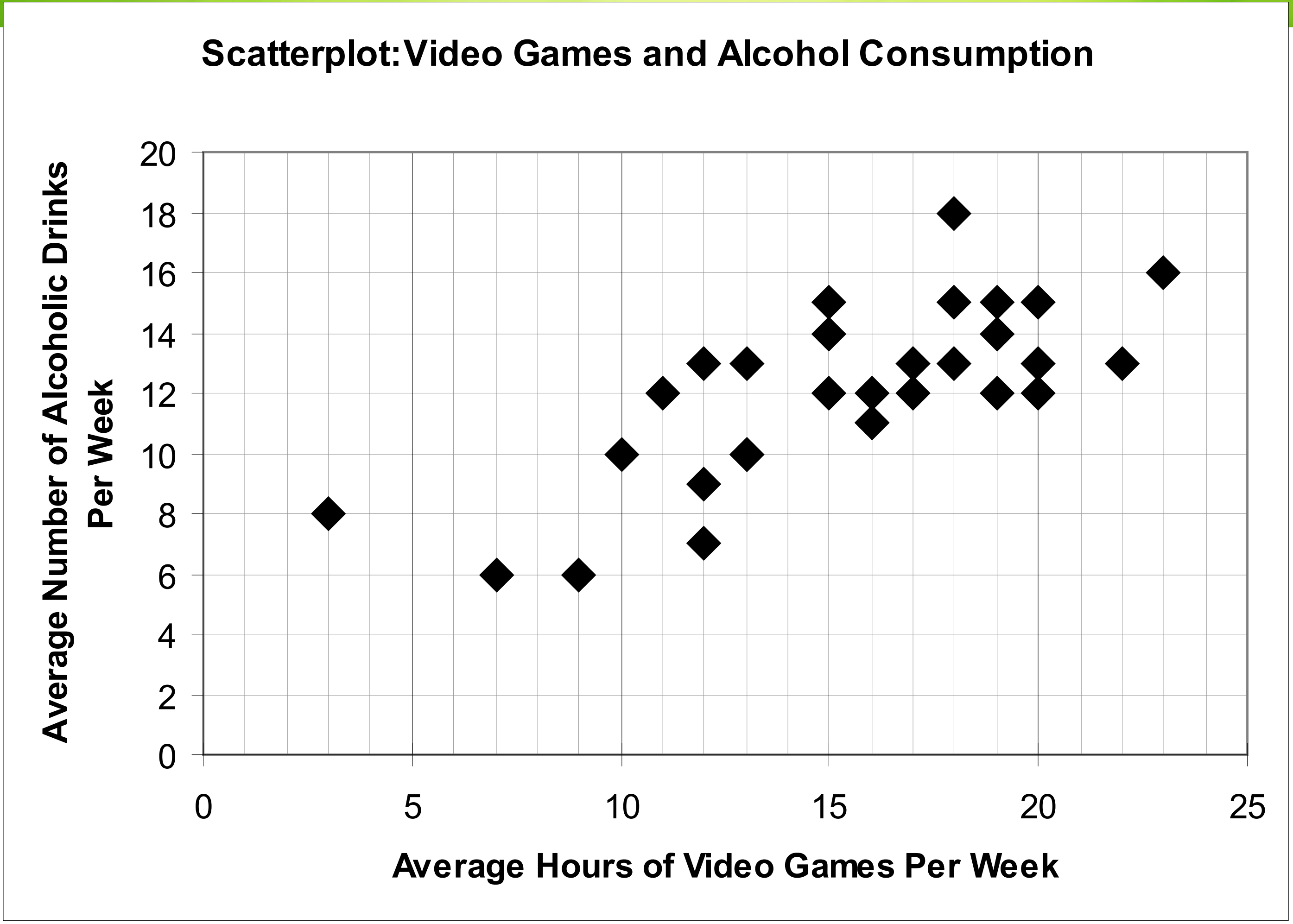
- Scatter Diagram Method
- Karl Pearson’s Coefficient of Correlation
- Spearman’s Rank Correlation

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

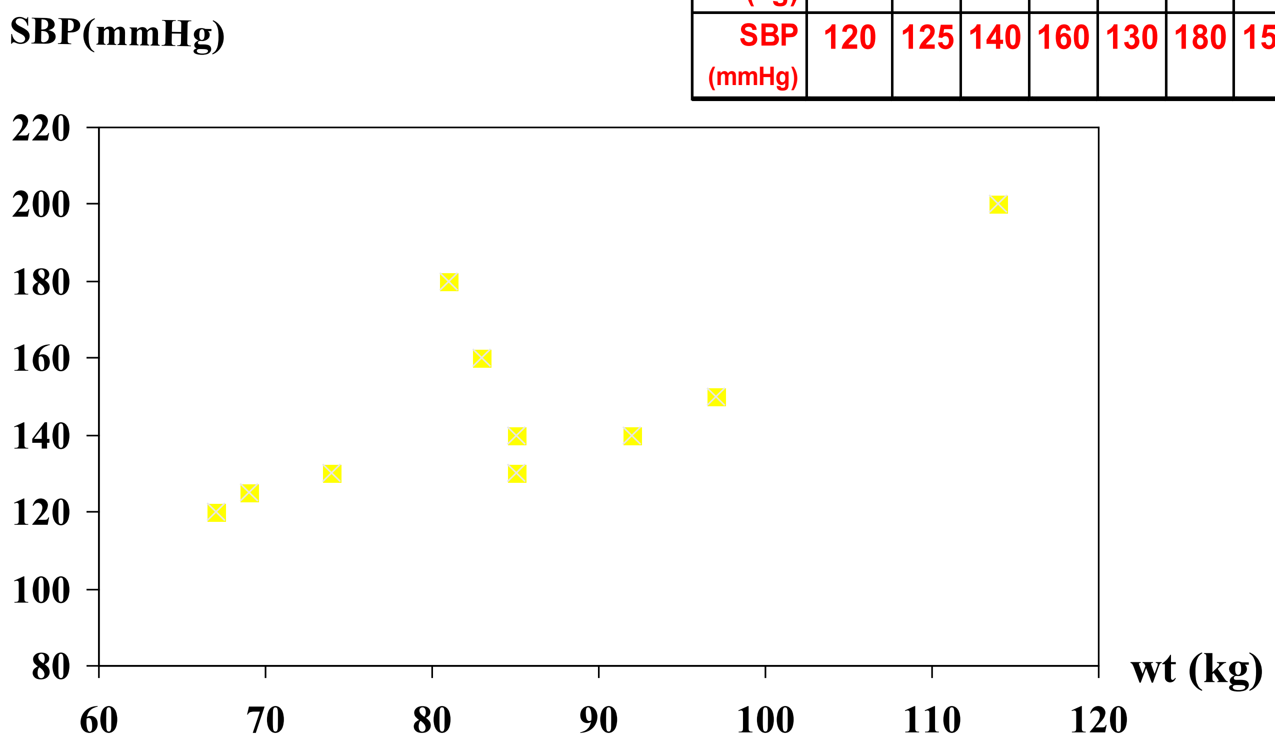
Example of Scatter Plot



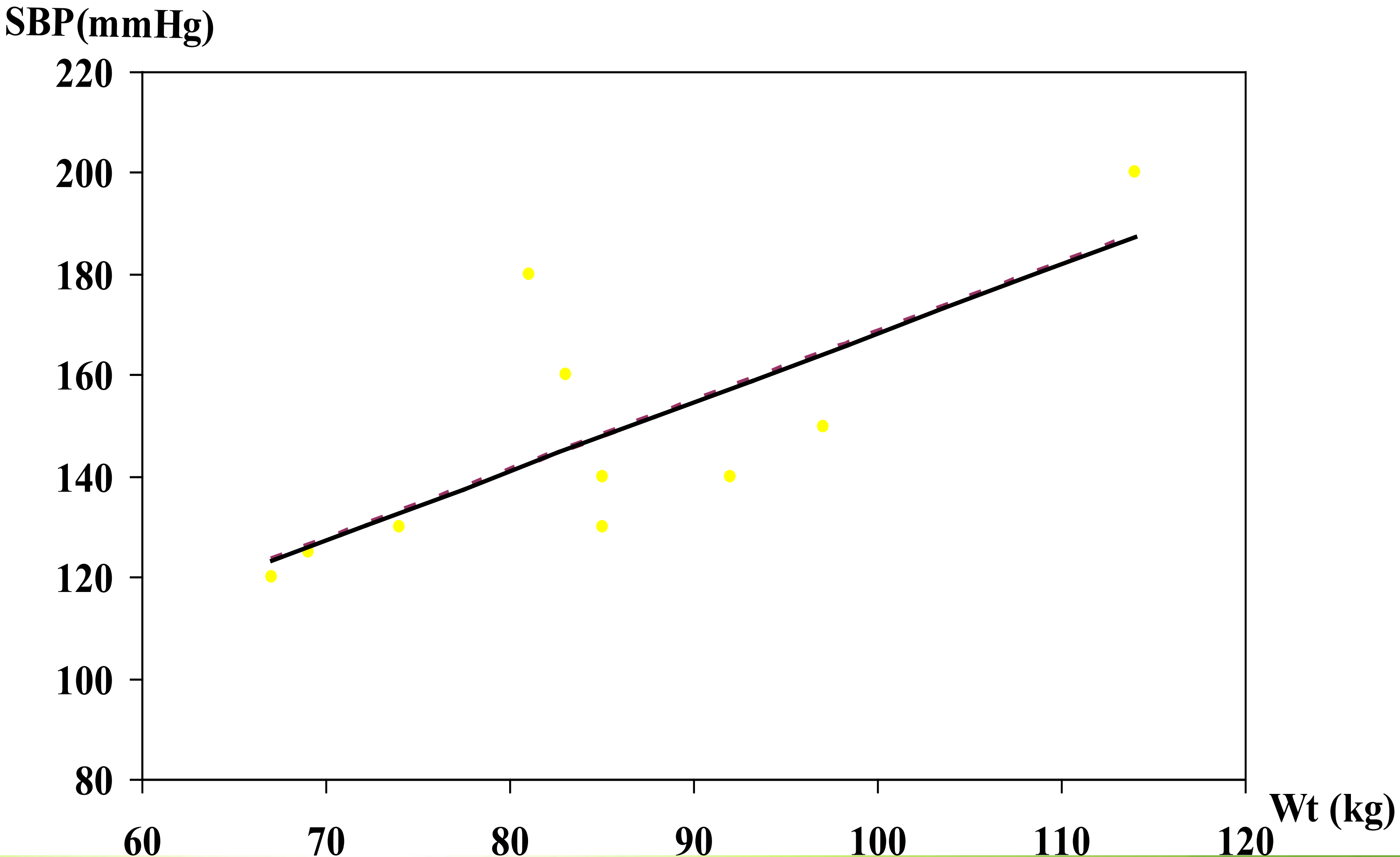
Example

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

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



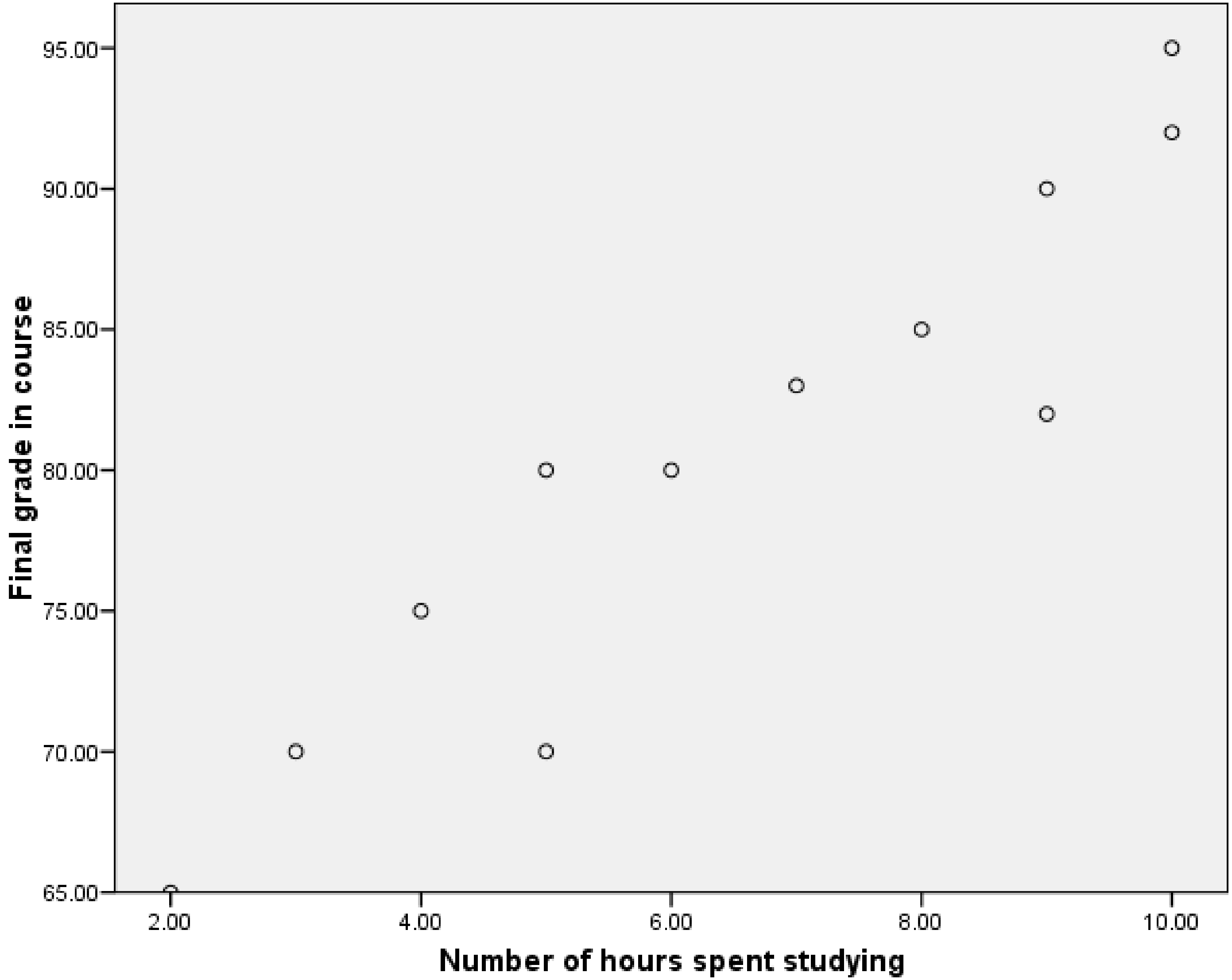
Scatter diagram of weight and systolic blood pressure

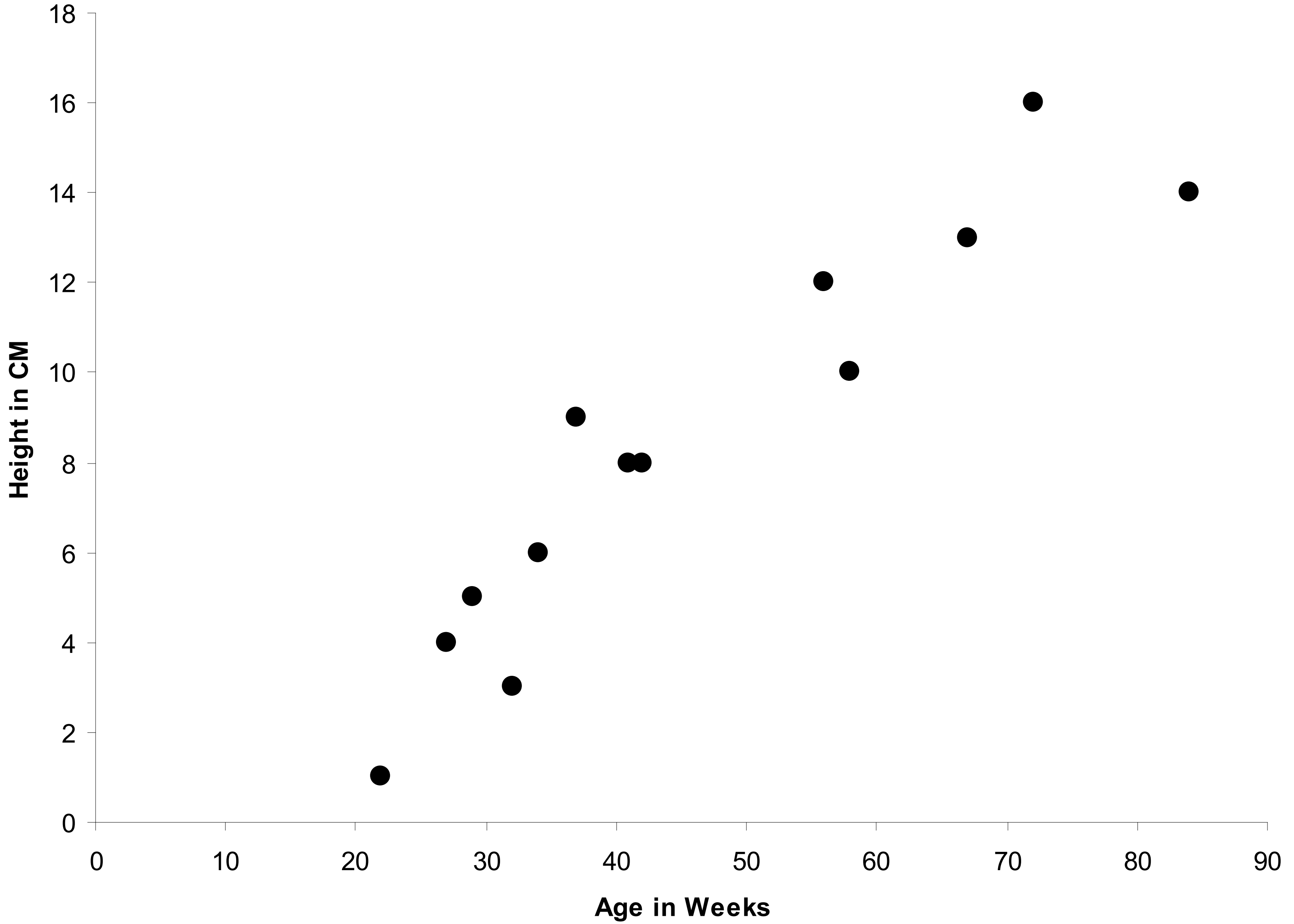
Scatter plots

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

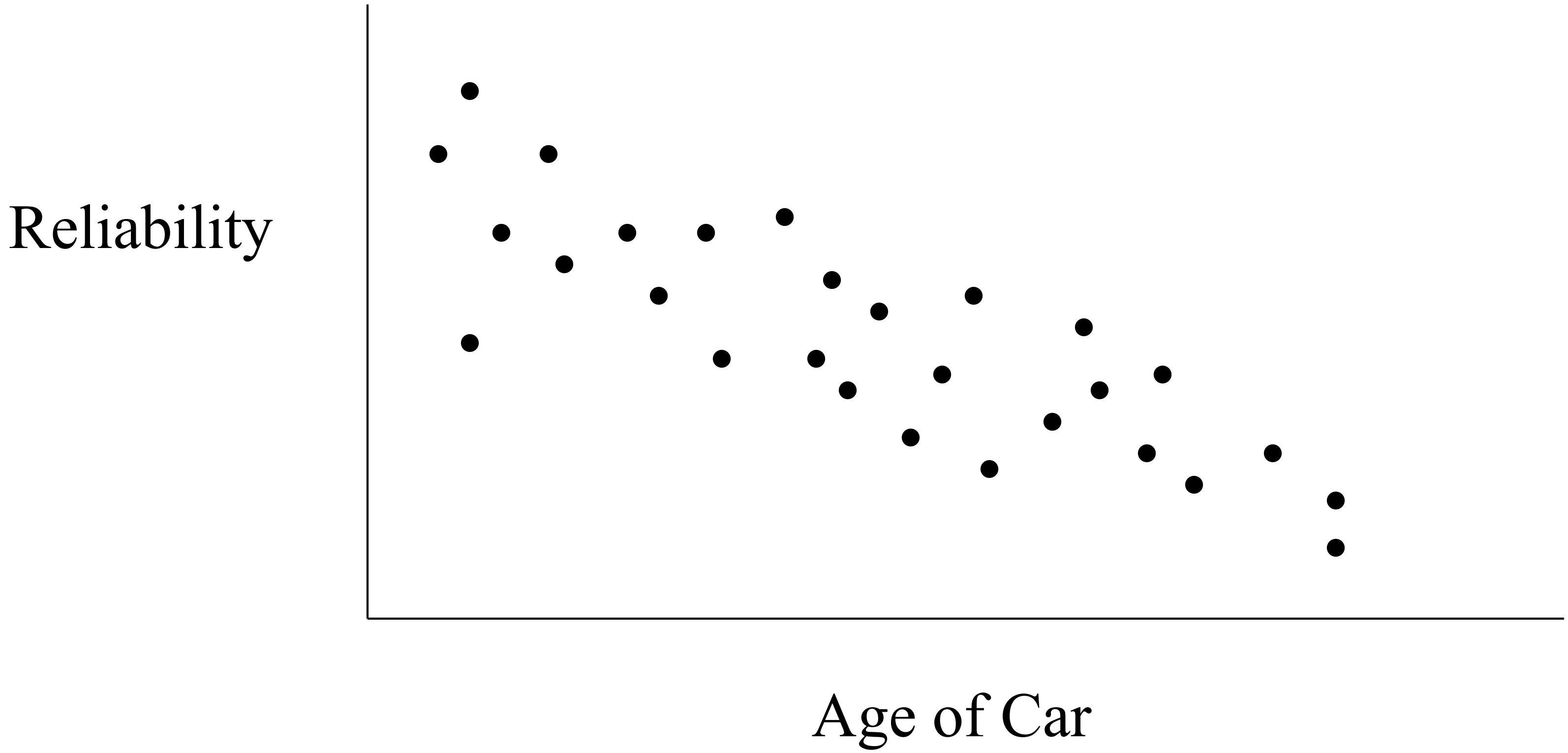
- positive relationship
- negative relationship
- no relationship

Positive relationship

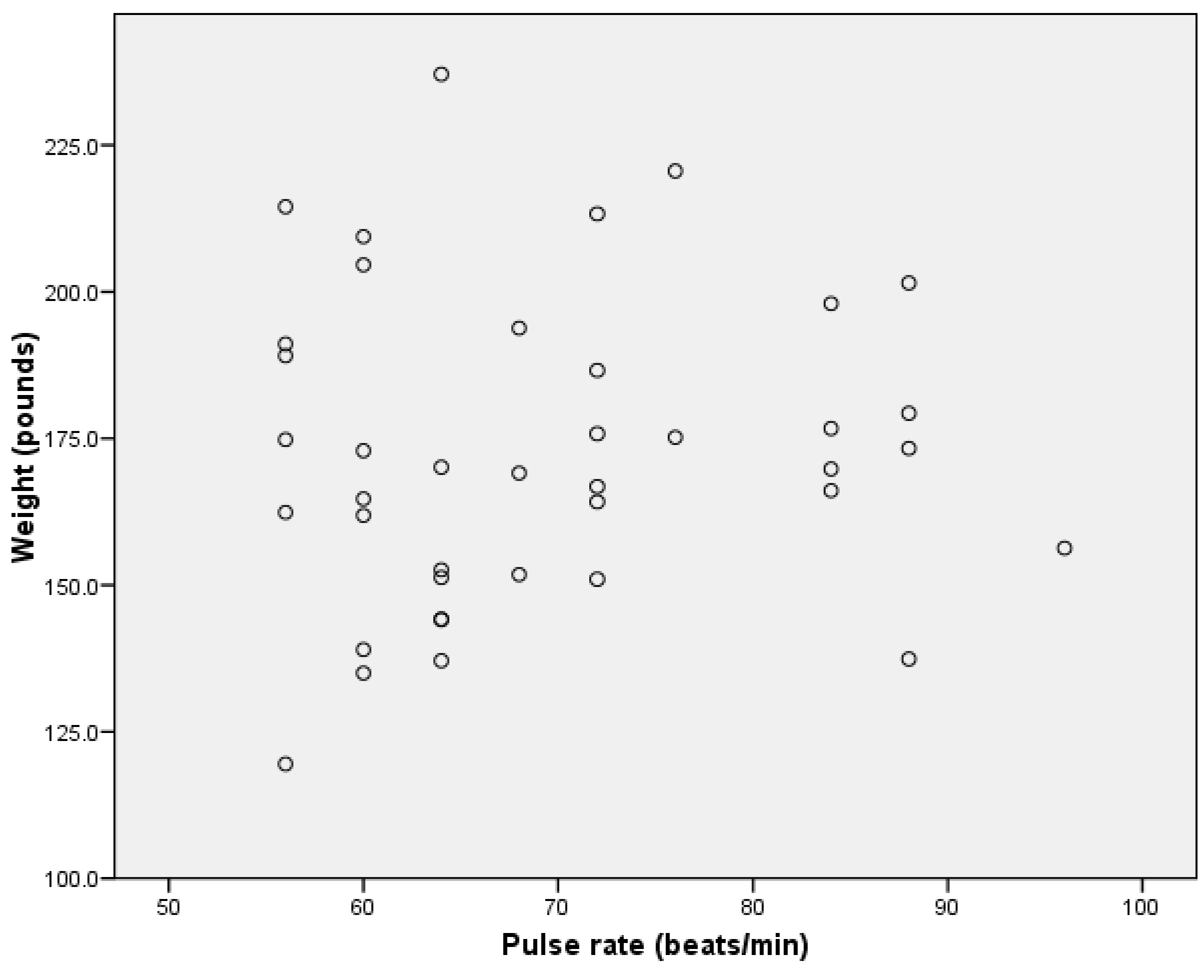



Height in CM
18
16
14
12
10
8
6
4
2
0
0
10
20
30
40
50
60
70
80
90
Age in Weeks

Negative relationship

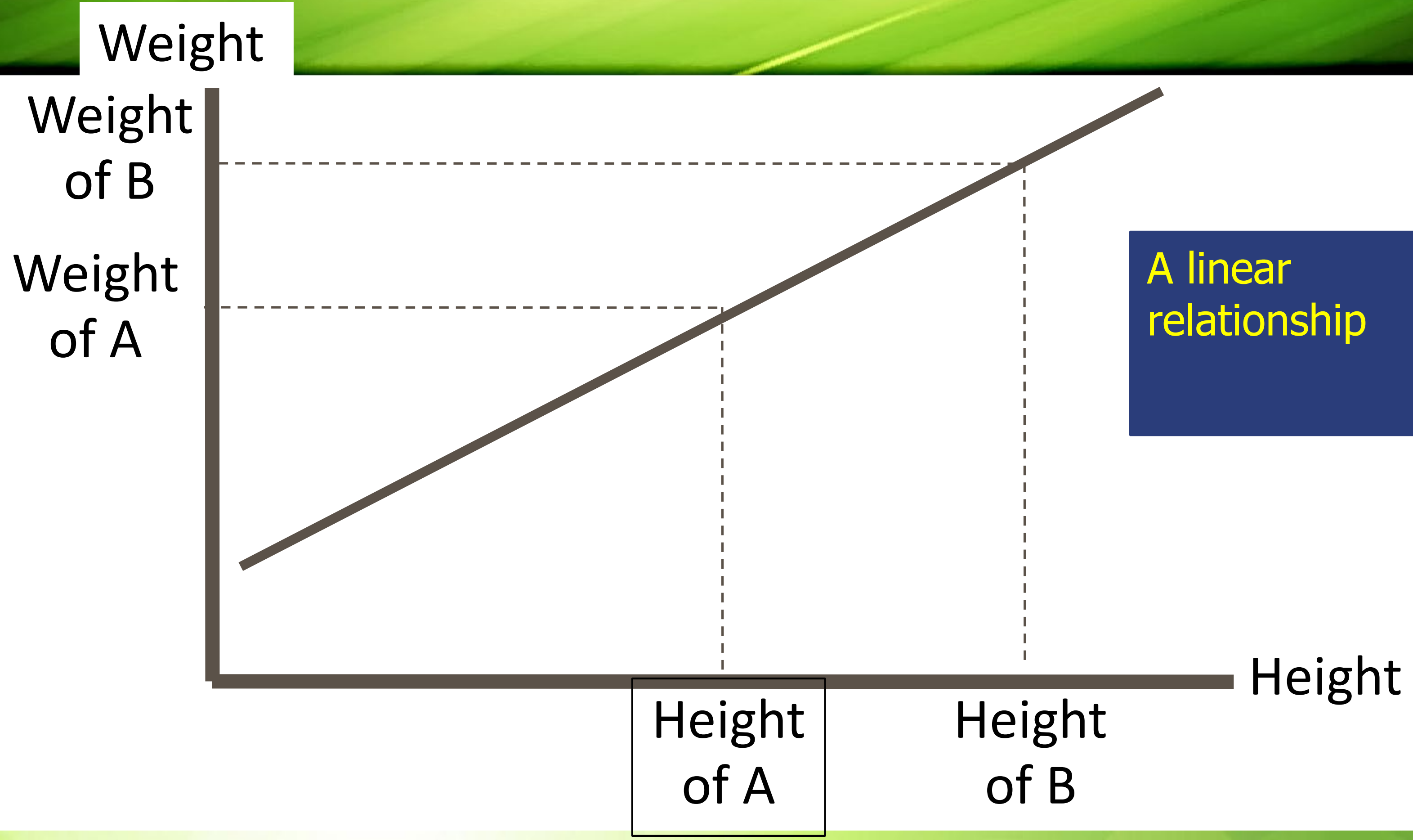


No relation



225.0
200.0
175.0
150.0
125.0
100.0
Weight (pounds)
50
60
70
80
90
100
Pulse rate (beats/min)

A perfect positive correlation



A linear relationship

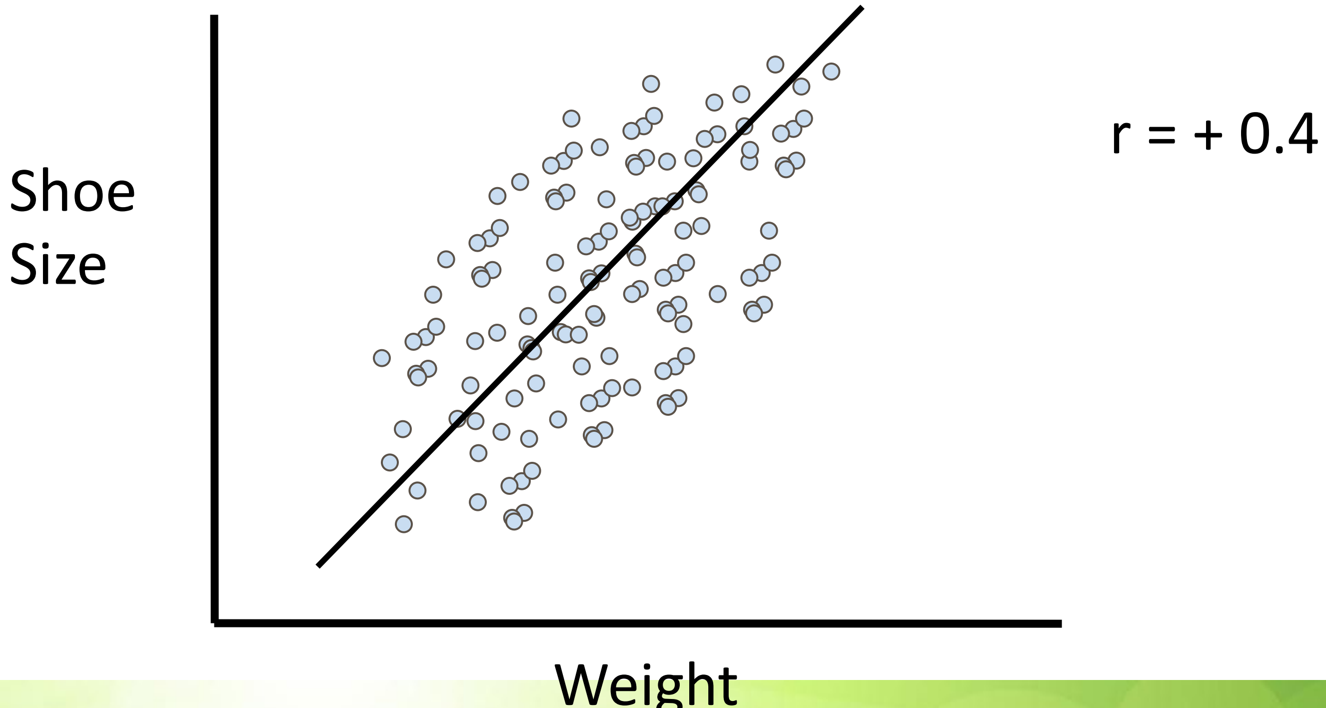
High Degree of positive correlation

- Positive relationship

r = +.80

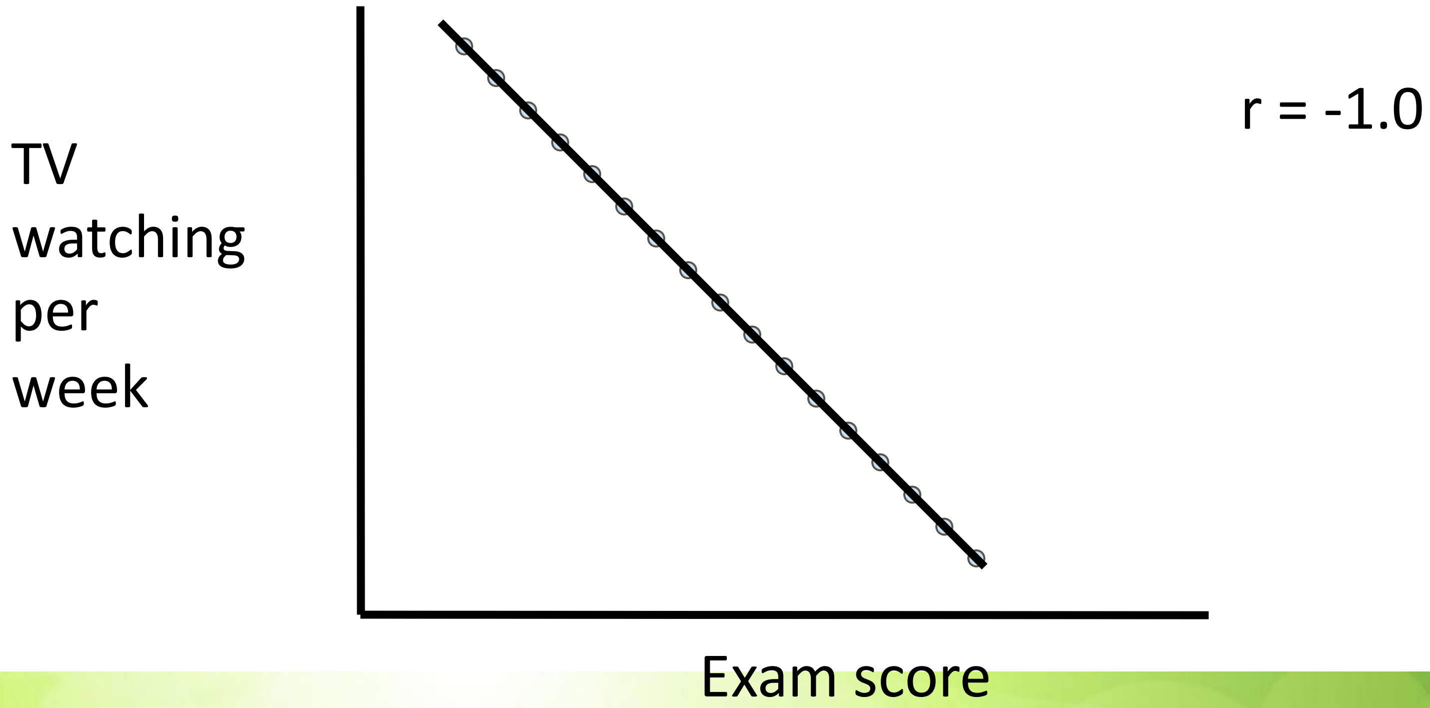
Degree of correlation

- **Moderate Positive Correlation**



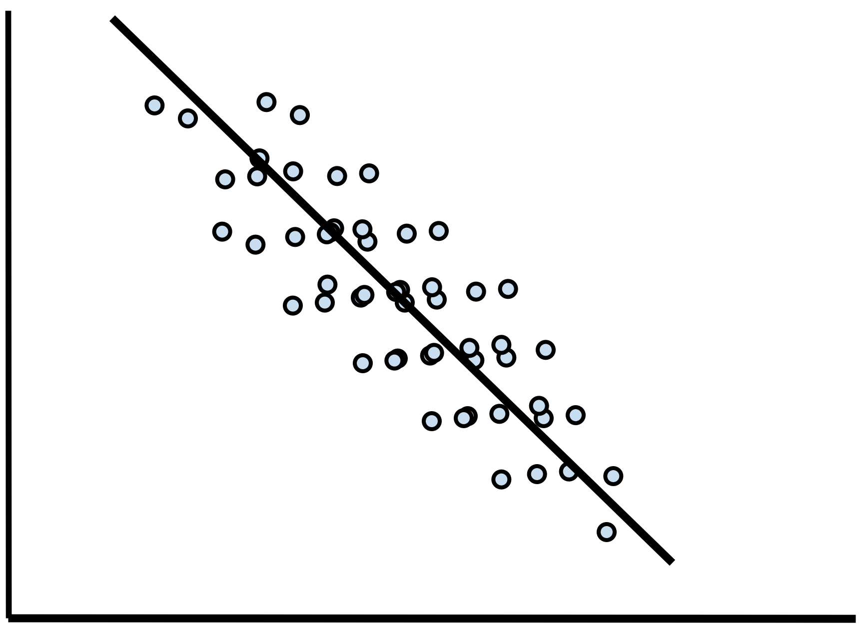
Degree of correlation

- **Perfect Negative Correlation**



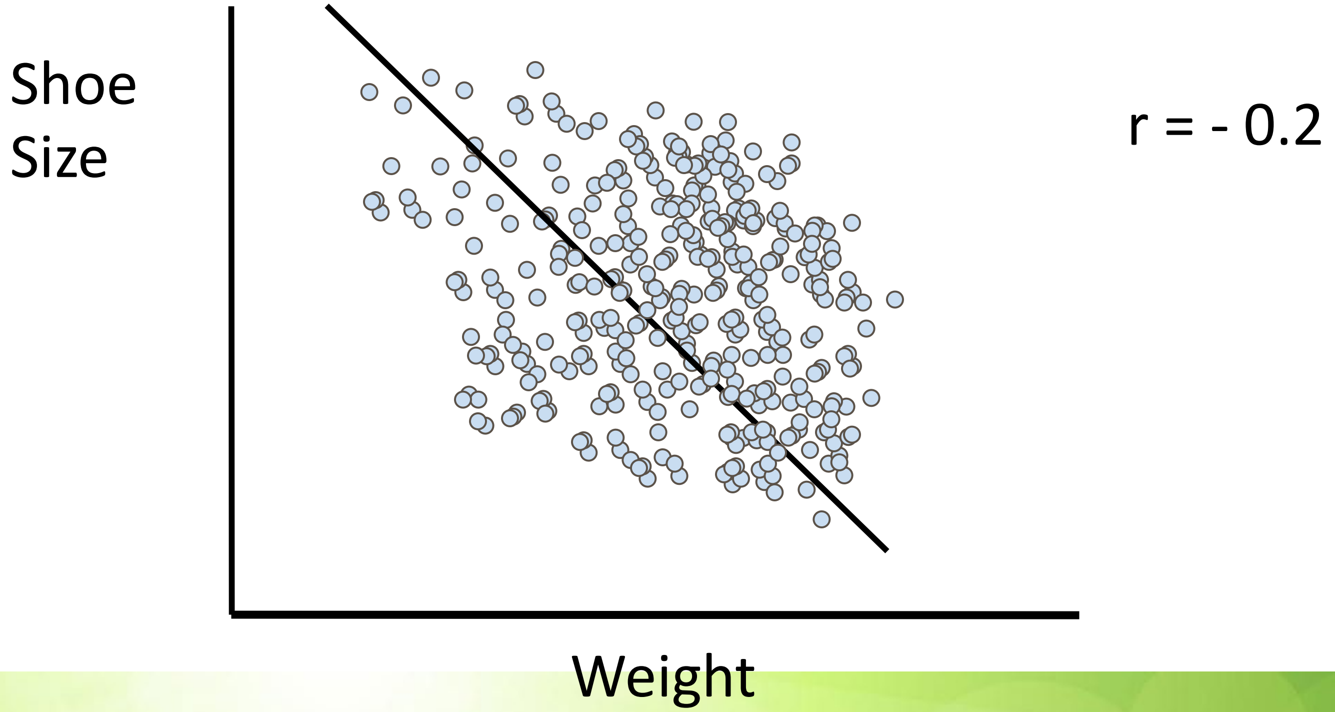
Degree of correlation

- **Moderate Negative Correlation**



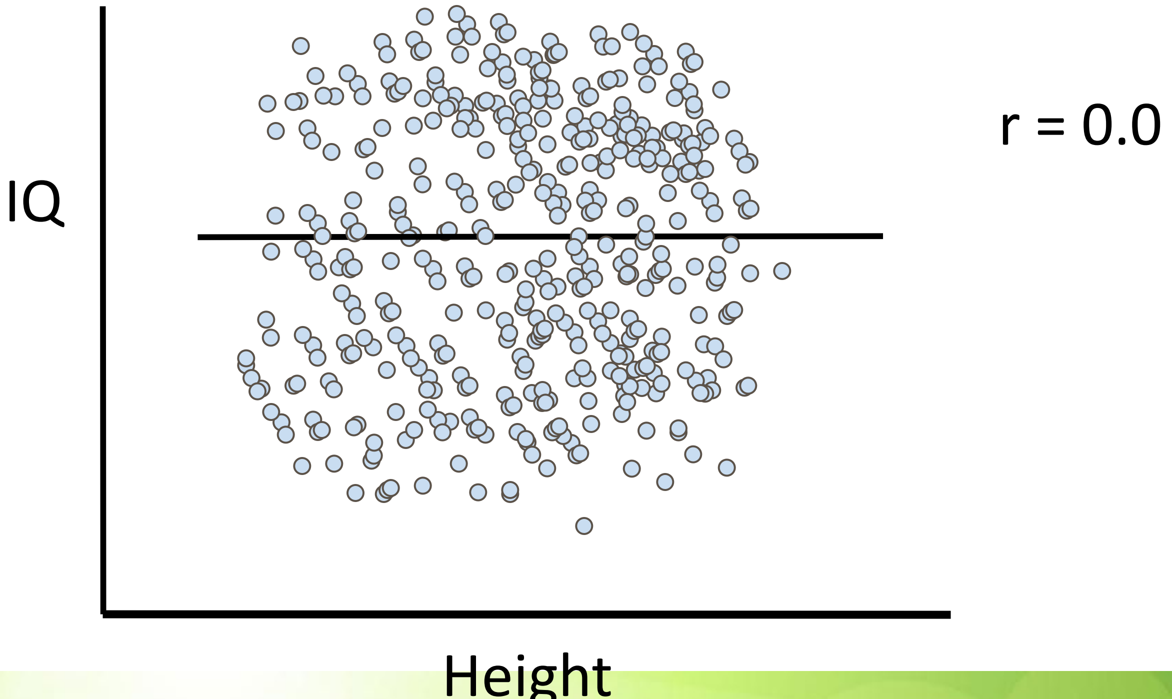
Degree of correlation

- **Weak negative Correlation**

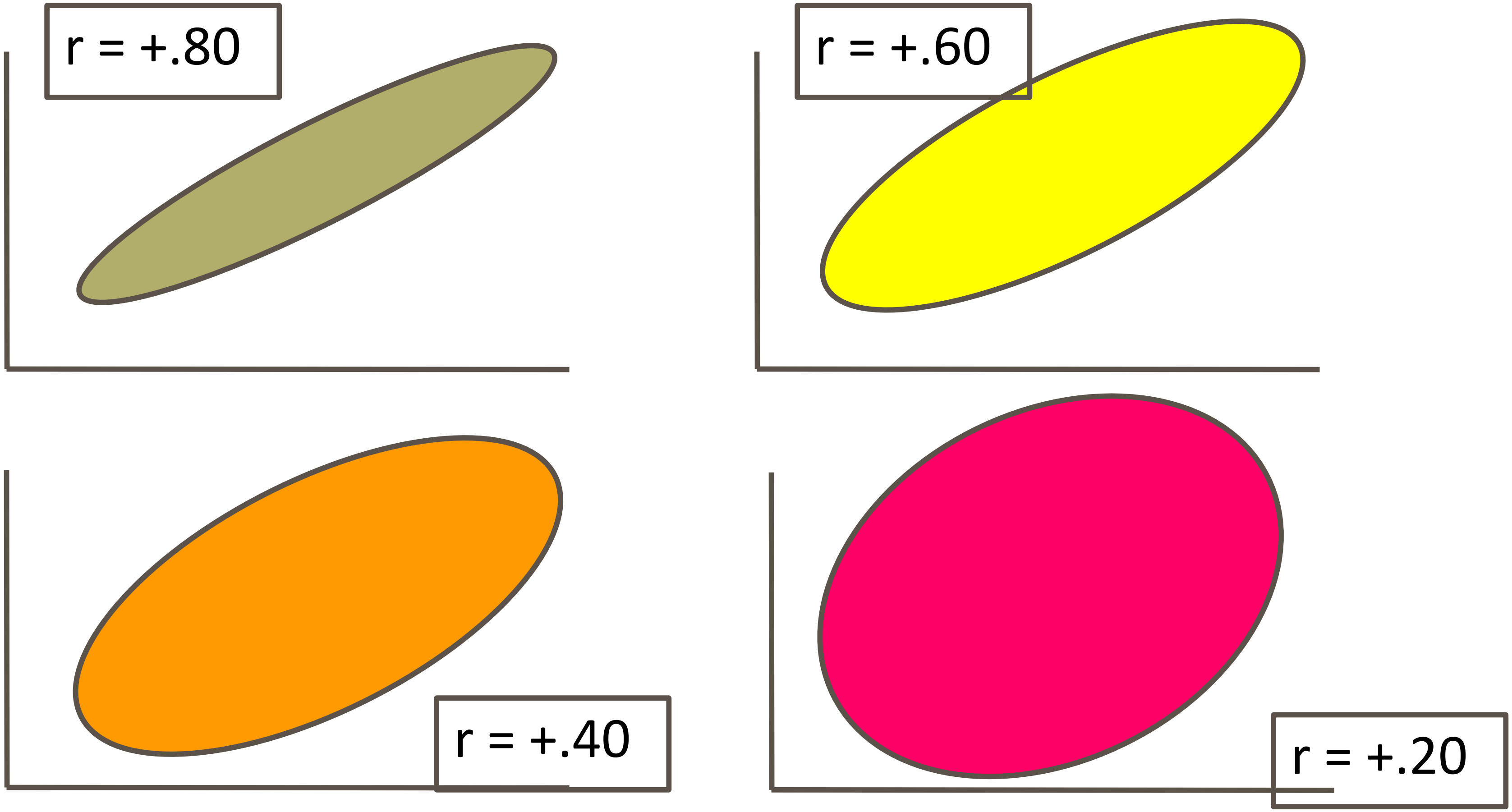


Degree of correlation

- **No Correlation (horizontal line)**



Degree of correlation (r)



Spurious/Non-sense Correlation:

- The correlation in absence of causation is called Spurious or Non-sense Correlation.

- Ex. Correlation between ***Marks of Student*** and ***Gold Prices***.

Advantages of Scatter Diagram

- Simple & Non Mathematical method
- Not influenced by the size of extreme item
- First step in investing the relationship between two variables

Disadvantage of scatter diagram

Can not adopt the an exact degree of correlation

Correlation

- **Correlation**: The degree of relationship between the variables under consideration is measure through the correlation analysis.
- The measure of correlation called the correlation coefficient .
- The degree of relationship is expressed by coefficient which range from correlation **($-1 \leq r \geq +1$)**

1st way of classification: Types of Correlation

- **Positive Correlation:** The correlation is said to be positive correlation if the values of two variables changing with same direction.

  Ex. Pub. Exp. & sales, Height & weight.

- **Negative Correlation:** The correlation is said to be negative correlation when the values of variables change with opposite direction.

  Ex. Price & qty. demanded.

More examples

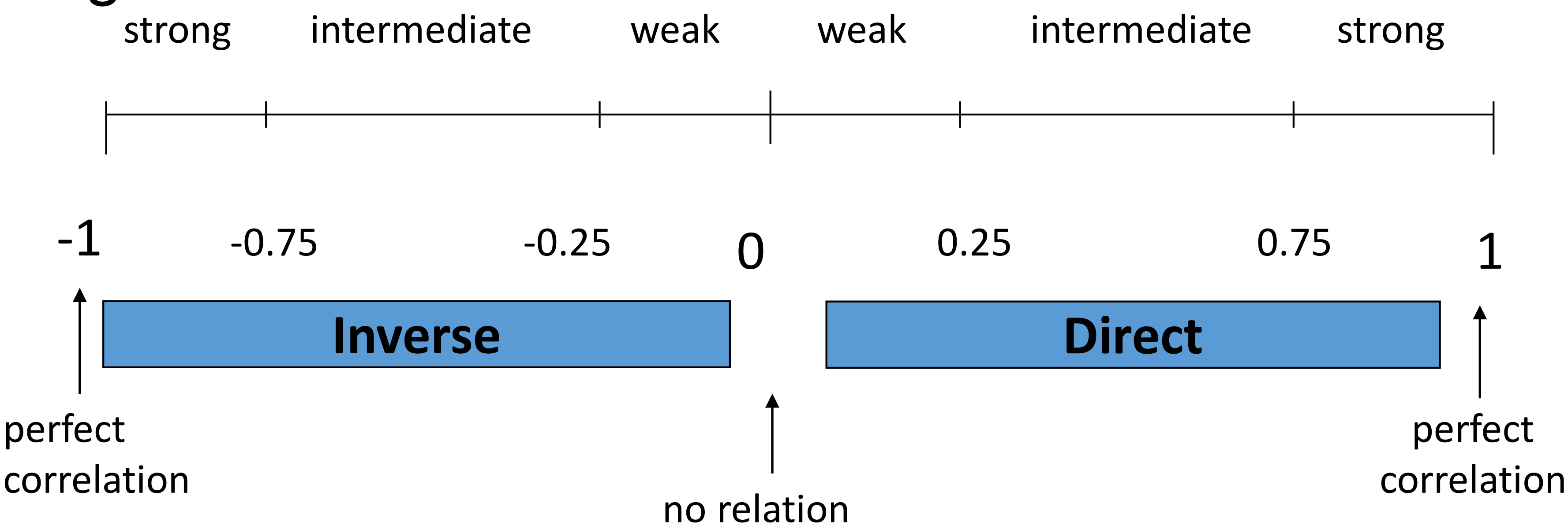
- **Positive relationships**
- water consumption and temperature.
- study time and grades.

- **Negative relationships**:
- alcohol consumption and driving ability.
- Price & quantity demanded

2nd way of classification: Types of Correlation

- **Simple correlation:** Under simple correlation problem there are only two variables are studied.
- **Multiple Correlation:** Under Multiple Correlation three or more than three variables are studied.
- **Partial correlation:** analysis recognizes more than two variables but considers only two variables keeping the other constant.

- 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.



Karl Pearson's Coefficient of Correlation

Karl Pearson's Coefficient of Correlation

- **Formula**

$$r_{xy} = \frac{cov(x,y)}{\sqrt{var(x) * var(y)}}$$

where, $cov(x,y) = \frac{\sum(x - \overline{x})\,(y - \overline{y})}{(n-1)}$

Advantages of Pearson’s Coefficient

- It summarizes in one value, the degree of correlation & direction of correlation also.

Limitation of Pearson's Coefficient

- Always assume linear relationship
- Interpreting the value of r is difficult.
- Value of Correlation Coefficient is affected by the extreme values.
- Time consuming method

3. Spearman's Rank Coefficient of Correlation

Spearman's Rank Coefficient of Correlation

- When statistical series arranged in serial order, in such situation Spearman Rank correlation can be used.

$$\rho_{xy} = 1 - \frac{6\sum d^2}{n^3 - n}$$

where $d_i = R_1 - R_2$

- R = Rank correlation coefficient
- D = Difference of rank between paired item in two series.
- N = Total number of observation.

Rank Correlation Coefficient (R)

a) Steps after finding ranks:

1) Calculate the difference 'D' of two Ranks i.e. ($R1 - R2$).
2) Square the difference & calculate the sum of the difference i.e. $\sum D^2$
3) Substitute the values obtained in the formula.

Rank Correlation Coefficient (R)

- **Equal Ranks or tie in Ranks:**

In such cases average ranks should be assigned to each individual.

and

$$\rho_{xy} = 1 - \frac{6\sum(d^2 + CF)}{n^3 - n}$$

m = The number of time an item is repeated

$$CF = \frac{1}{12\,({m_1}^3 - m_1)} + \frac{1}{12\,({m_2}^3 - m_2)} + \cdots$$

Merits Spearman’s Rank Correlation

- This method is simpler to understand and easier to apply compared to karl Pearson’s correlation method.
- This method is useful where we can give the ranks and not the actual data. (qualitative term)
- This method is to use where the initial data in the form of ranks.

Limitation Spearman's Correlation

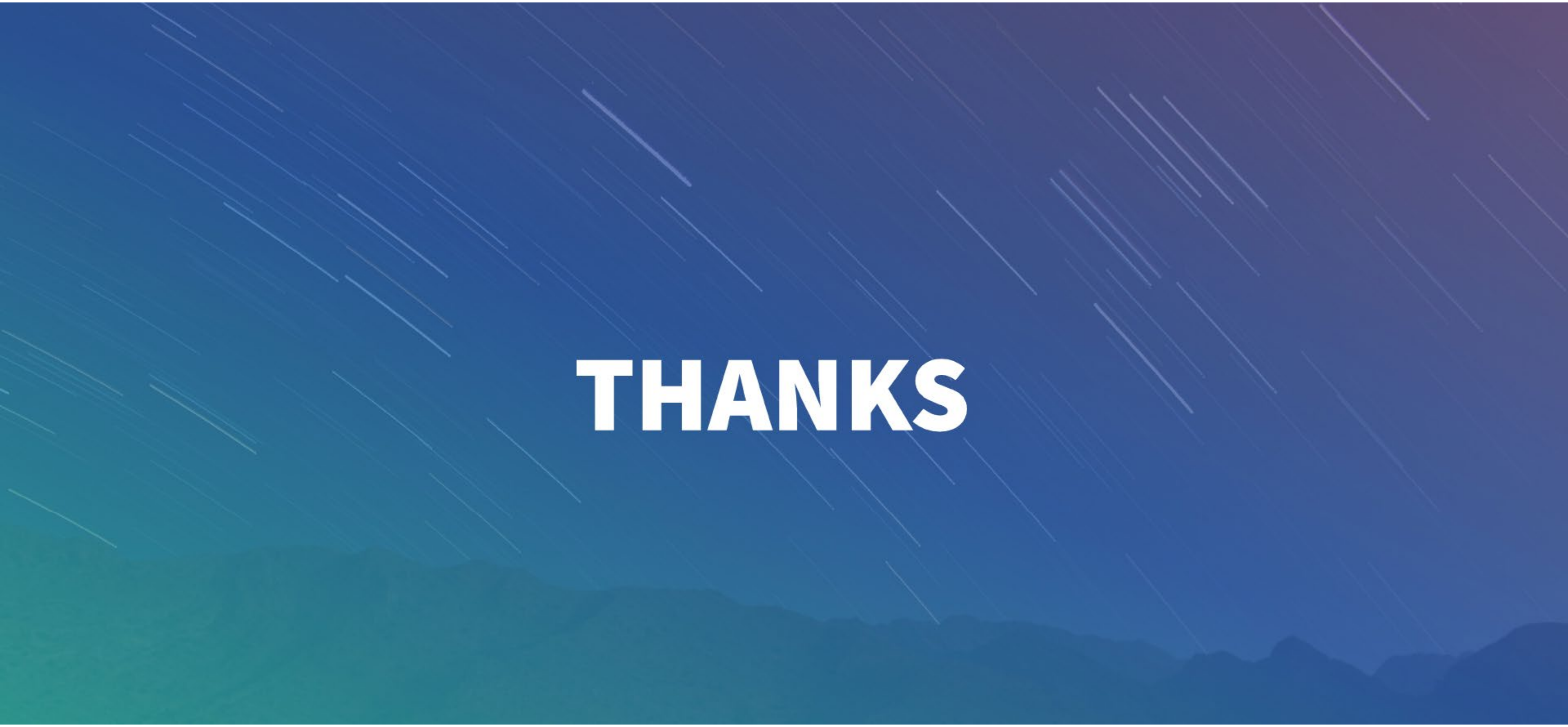
- Cannot be used for finding out correlation in a grouped frequency distribution.
- This method should be applied where N exceeds 30.

Advantages of Correlation studies

- Show the amount (strength) of relationship present
- Can be used to make predictions about the variables under study.
- Can be used in many places, including natural settings, libraries, etc.
- Easier to collect co relational data

Disadvantages of correlation studies

- Can’t assume that a cause-effect relationship exists
- Little or no control (experimental manipulation) of the variables is possible
- Relationships may be accidental or due to a third, unmeasured factor common to the 2 variables that are measured


THANKS

Simple Linear Regression

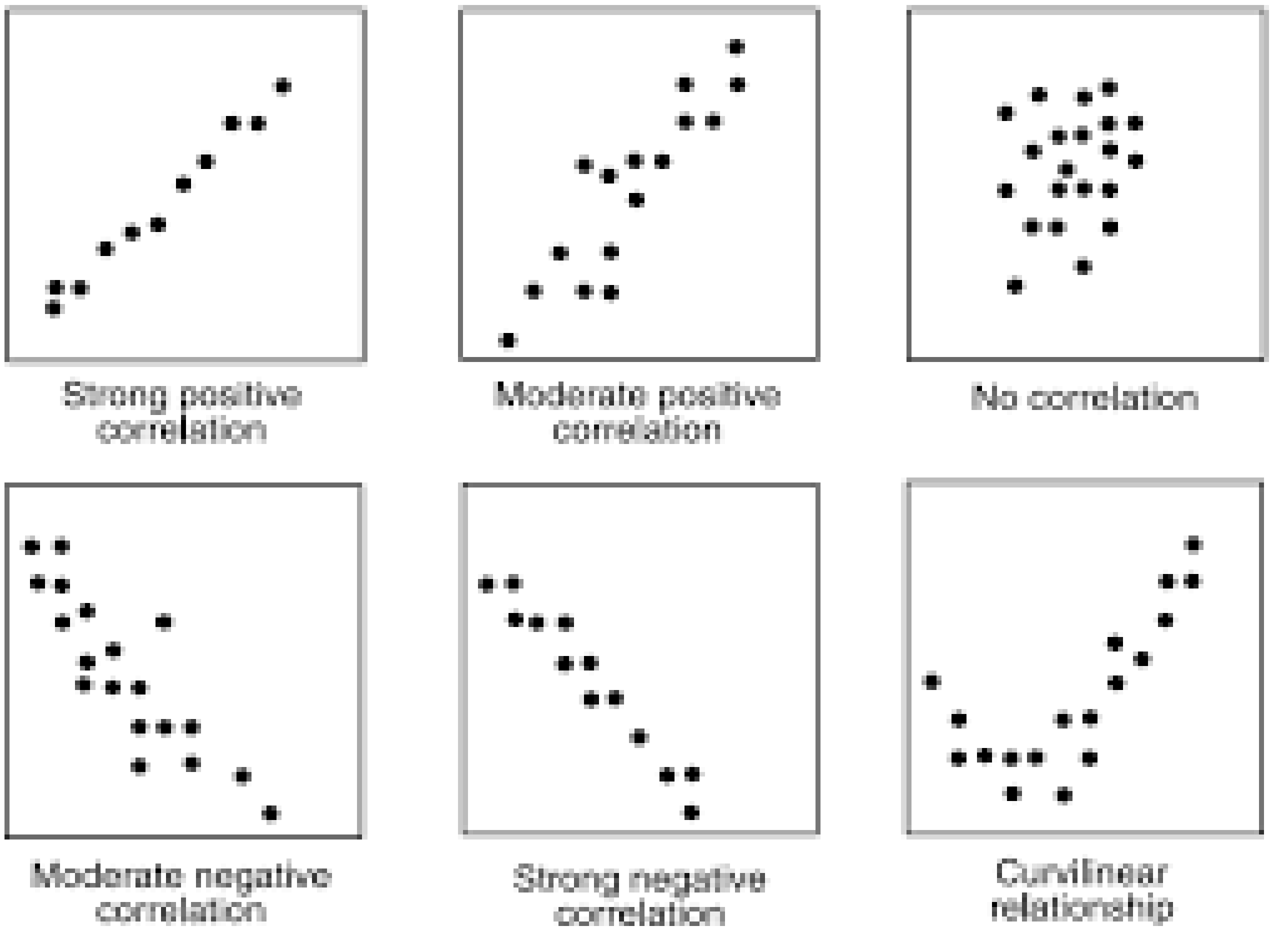
Introduction

- Correlation is the measure of linear relationship between two variables.
- Regression analysis is the tool to model the linear relationship between two variables.
- Regression analysis is useful when there is strong positive or negative correlation between two variables.
- Poor correlation between two variables does not implies independency of two variables. There may be non-linear relationship between given two variables.

Correlation

- When?
- When variables under consideration are continuous and you want to check whether there is any linear relationship between variables (>=2) under consideration.
- When you want to check whether the observed correlation is significant or not.

Correlation...



Correlation...

$$r_{xy} = \frac{\text{cov}(x, y)}{\sqrt{\sigma_x * \sigma_y}}$$

Moderate Negative

Moderate Positive

-1

-0.5

0

0.5

+1

Strong Negative

No correlation

Strong Positive

Regression

- When?
- If you get significant p-value for correlation coefficient between two variable and you want to model the linear relationship between these two variables.
- If you want a linear model (if exist) for prediction purpose.

Regression...

- The equation that describes how y is related to x and an error term is called the regression model.
- The simple linear regression model is:

$$y = b_0 + b_1 x + e$$

  - b_0 and b_1 are called parameters of the model.
  - e is a random variable called the error term.

Regression...

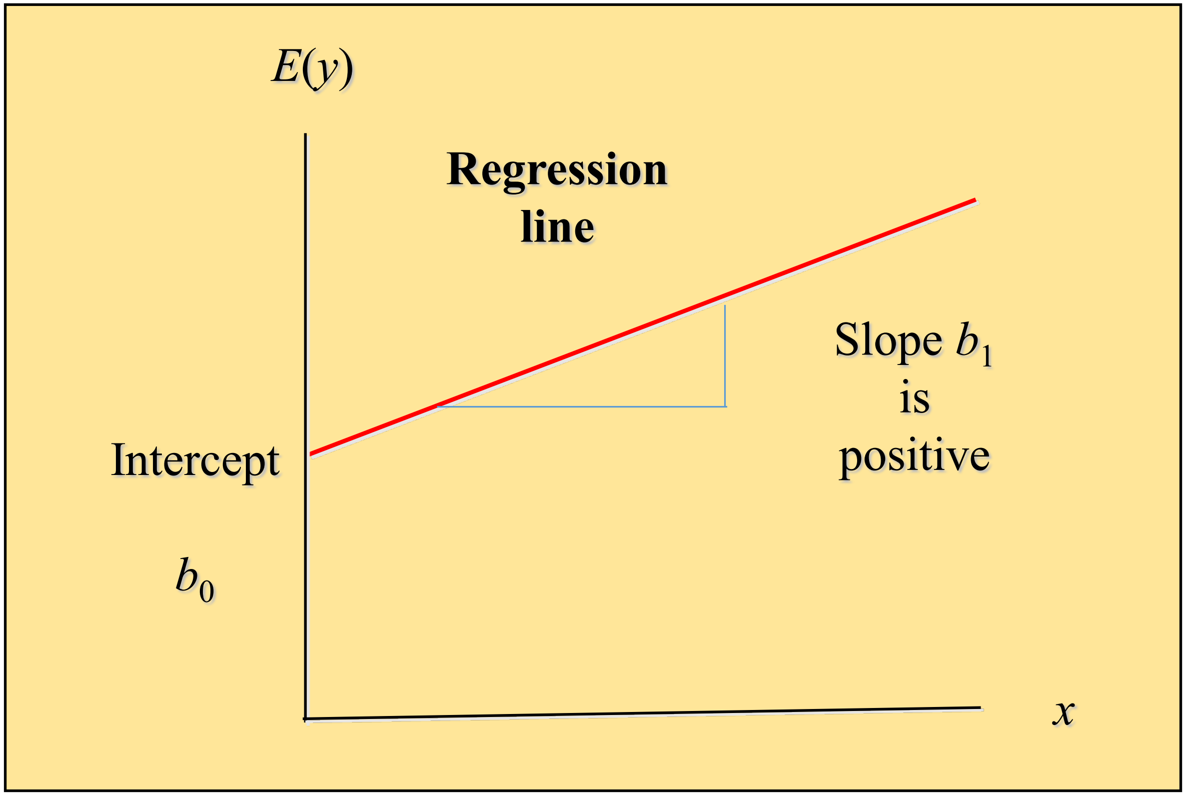
- The simple linear regression equation is:

$$E(y) = b_0 + b_1x$$

- Graph of the regression equation is a straight line.
- b_0 is the y intercept of the regression line.
- b_1 is the slope of the regression line.
- $E(y)$ is the expected value of y for a given x value.

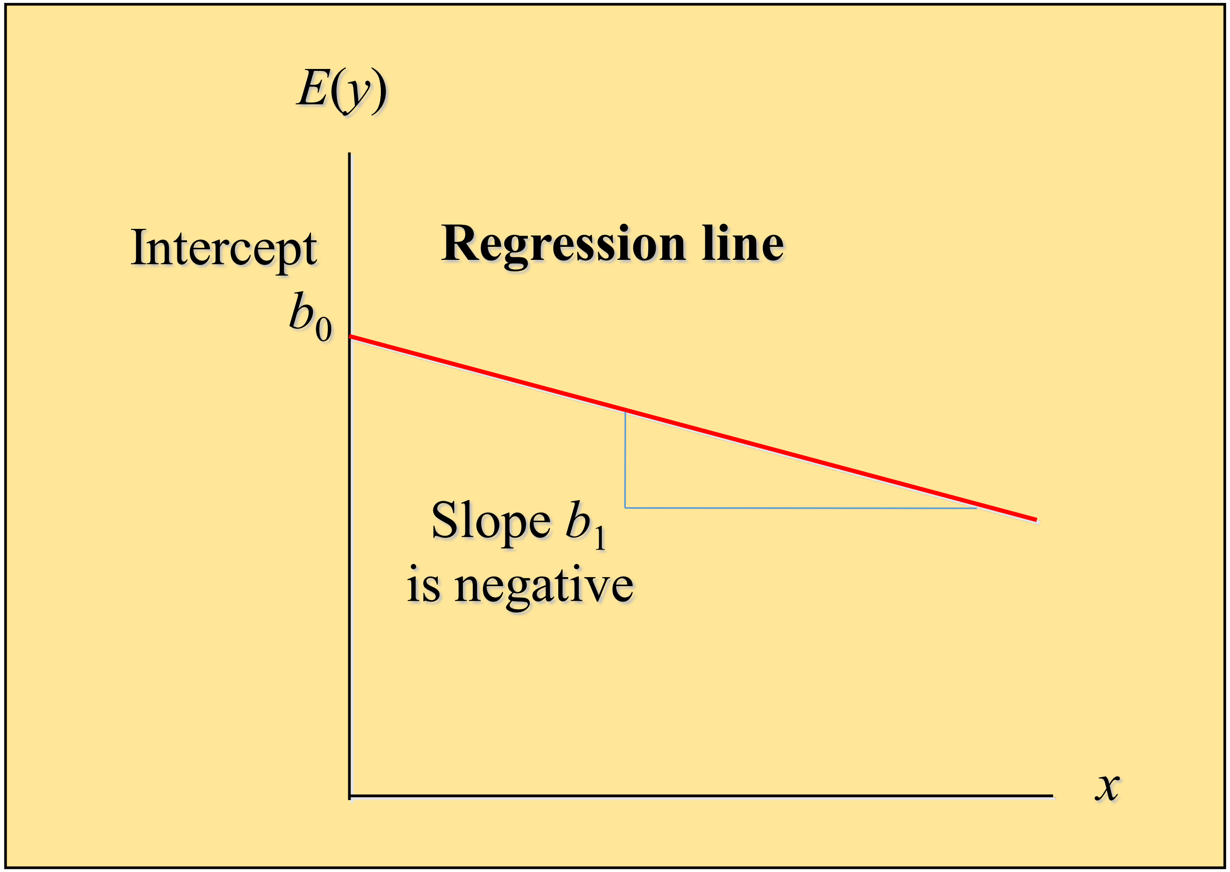
Regression...

- Positive Linear Relationship



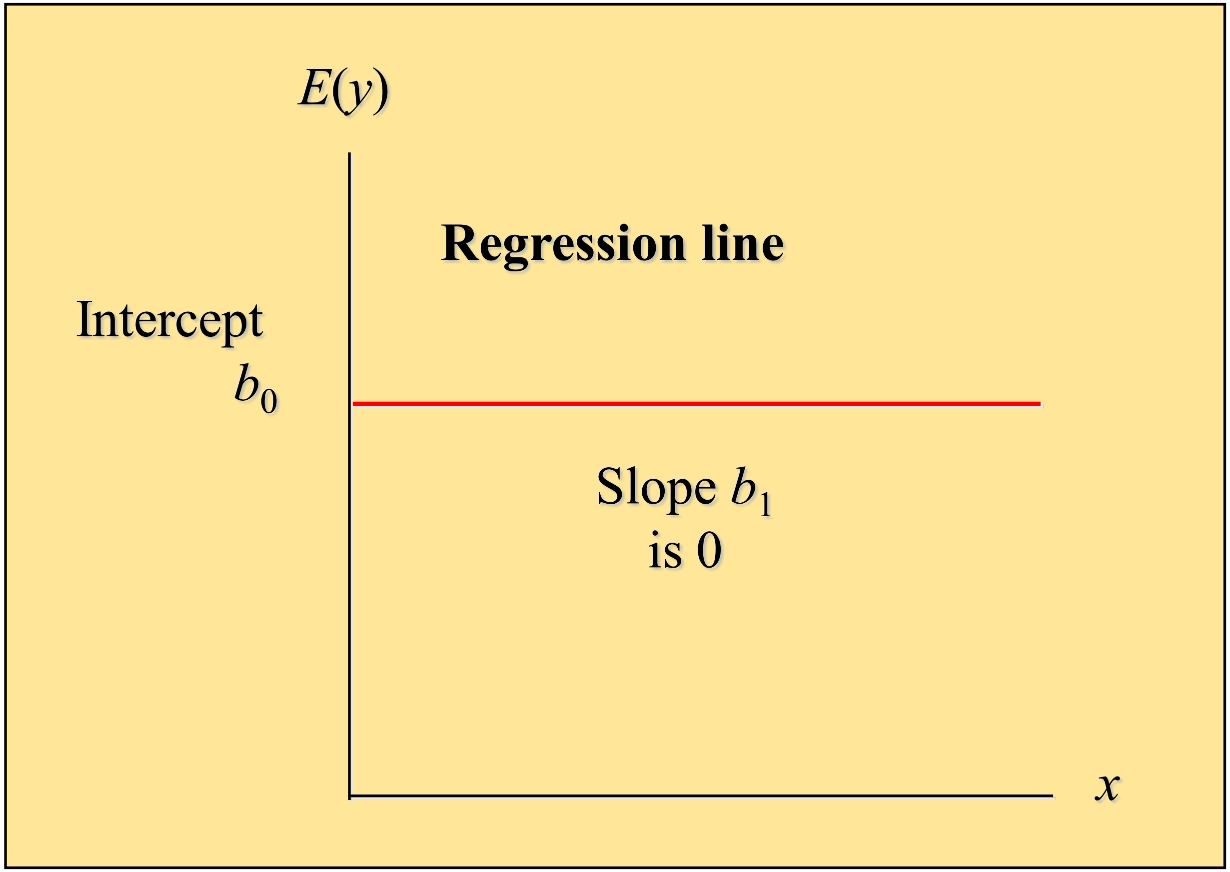
Regression...

- Negative Linear Relationship



Regression...

- No Relationship



Parameters

- In SLR equation discussed above, for any data (y, x) corresponding 'b_0' and 'b_1' are unknown constants and are called as parameters.
- These parameters cab be estimated using some estimation procedure.
- Least squares method is used to estimate these parameters.
- The estimated parameters are denoted as $\hat{b}_0$ and $\hat{b}_1$ respectively.

Least Squares Method

- Least Squares Criterion

$$\min \sum (y_i - \hat{y}_i)^2$$

where:

y_i = <u>observed</u> value of the dependent variable for the i^{th} observation

$\hat{y}_i$ = <u>estimated</u> value of the dependent variable for the i^{th} observation

Parameter Estimator

- Slope for the Estimated Regression Equation

$$b_1 = \frac{\sum x_i y_i - (\sum x_i \sum y_i)/n}{\sum x_i^2 - (\sum x_i)^2/n}$$

Parameter Estimator...

- y-Intercept for the Estimated Regression Equation

$$b_0 = \overline{y} - b_1\overline{x}$$

where:

x_i = value of independent variable for ith observation

y_i = value of dependent variable for ith observation

$\bar{x}$ = mean value for independent variable

$\bar{y}$ = mean value for dependent variable

n = total number of observations

Estimated SLR Equation

The estimated simple linear regression equation is:

$$\boxed{\hat{y} = \hat{b}_0 + \hat{b}_1 x}$$

- The graph is called the estimated regression line.
- $\hat{b}_0$ is the y intercept of the line.
- $\hat{b}_1$ is the slope of the line.
- $\hat{y}$ is the estimated value of y for a given x value.

Example: Reed Auto Sales

- Simple Linear Regression

  Reed Auto periodically has a special week-long sale. As part of the advertising campaign Reed runs one or more television commercials during the weekend preceding the sale. Data from a sample of 5 previous sales are shown on the next slide.

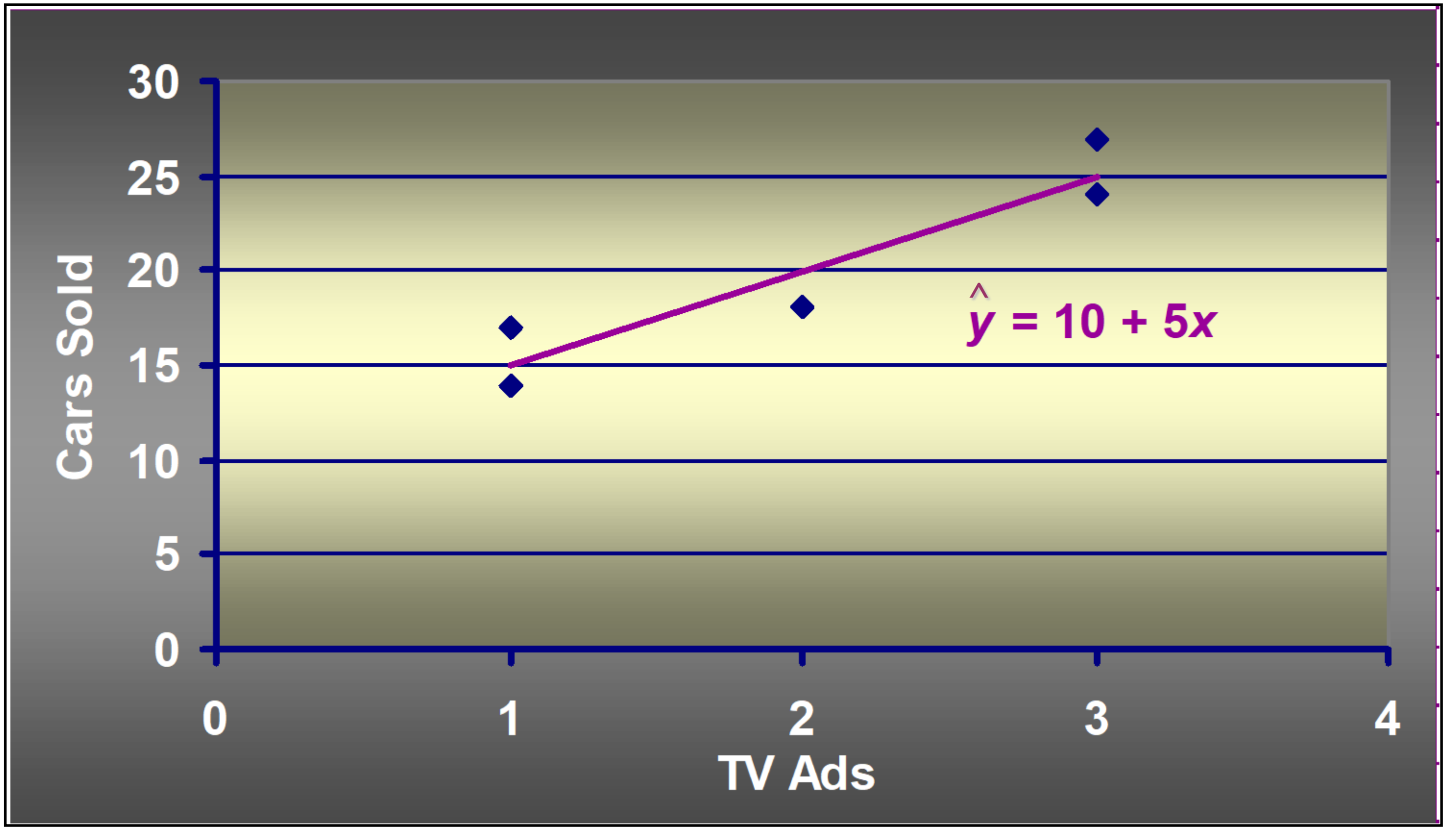
Example: Reed Auto Sales...

- Simple Linear Regression

Number of TV Ads	Number of Cars Sold
1	14
3	24
2	18
1	17
3	27

Example: Reed Auto Sales...

- Scatter Diagram



The Coefficient of Determination

- Relationship Among SST, SSR, SSE

$$\boxed{\text{SST} = \text{SSR} + \text{SSE}}$$

$$\boxed{\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2}$$

where:

SST = total sum of squares
SSR = sum of squares due to regression
SSE = sum of squares due to error

The Coefficient of Determination...

- The coefficient of determination is:

$$\boxed{R^2 = SSR/SST}$$

where:

SST = total sum of squares

SSR = sum of squares due to regression

Model Assumptions

- Assumptions About the Error Term ε

1. The error ε is a random variable with mean of zero.
2. The variance of ε, denoted by σ^2, is the same for all values of the independent variable.
3. The values of ε are independent.
4. The error ε is a normally distributed random variable.

Model Assumptions

- Assumptions About the Error Term ε
  1. The error ε is a random variable with mean of zero.
  2. The variance of ε, denoted by σ^2, is the same for all values of the independent variable.
  3. The values of ε are independent.
  4. The error ε is a normally distributed random variable.

Validating Model Assumptions

- Using qqplot for errors, we can check whether they follow Normal distribution or not.
- Using plot of residuals vs fitted values, we can check whether they are independent of each other or not. We can use Durbin-Watson's test for checking existence of autocorrelation in errors.
- Using the plot above, we can check assumption of constant variance or we can use 'non-constant error variance test' for the same.

Measures of goodness of Model

- S^2
- R^2
- Adj_R^2 …

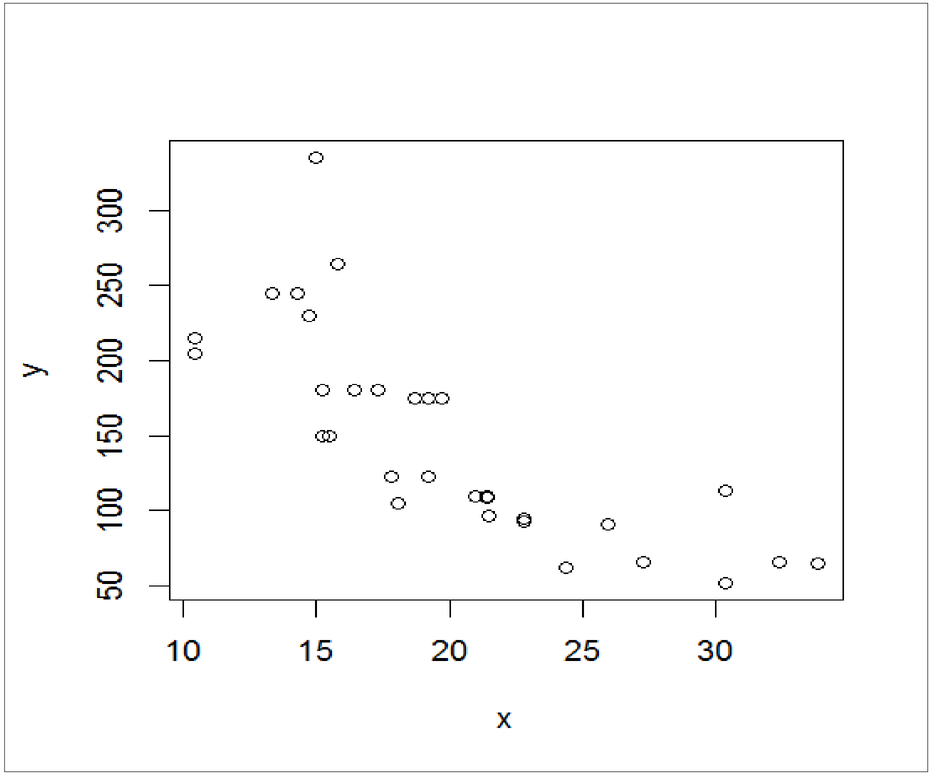
Issues in fitting Linear Model

- Outliers: Outliers leads to poor model. Using boxplot one can identify them and treat them differently. We can also use ‘Outlier test’ for their validation.
- Influential observations: Some cases could be very influential even if they look to be within a reasonable range of the values. They could be extreme cases against a regression line and can alter the results if we exclude them from analysis. We can use leverage plot or Cook’s distance to identify influential observations.

- **Correlation and Regression Using R**

SLR in R

```
> y <- mtcars$hp
> x<- mtcars$mpg
> plot(x,y)
```



SLR in R...

```
> cor(x,y)
[1] -0.7761684
> cor.test(x,y)

        Pearson's product-moment correlation

data:  x and y
t = -6.7424, df = 30, p-value = 1.788e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.8852686 -0.5860994
sample estimates:
       cor
-0.7761684
```

SLR in R...

```
> Z<-lm(y~x)
> Z

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)          x
    324.08       -8.83
```

$$y = 324.08 - 8.83 * x$$

SLR in R...

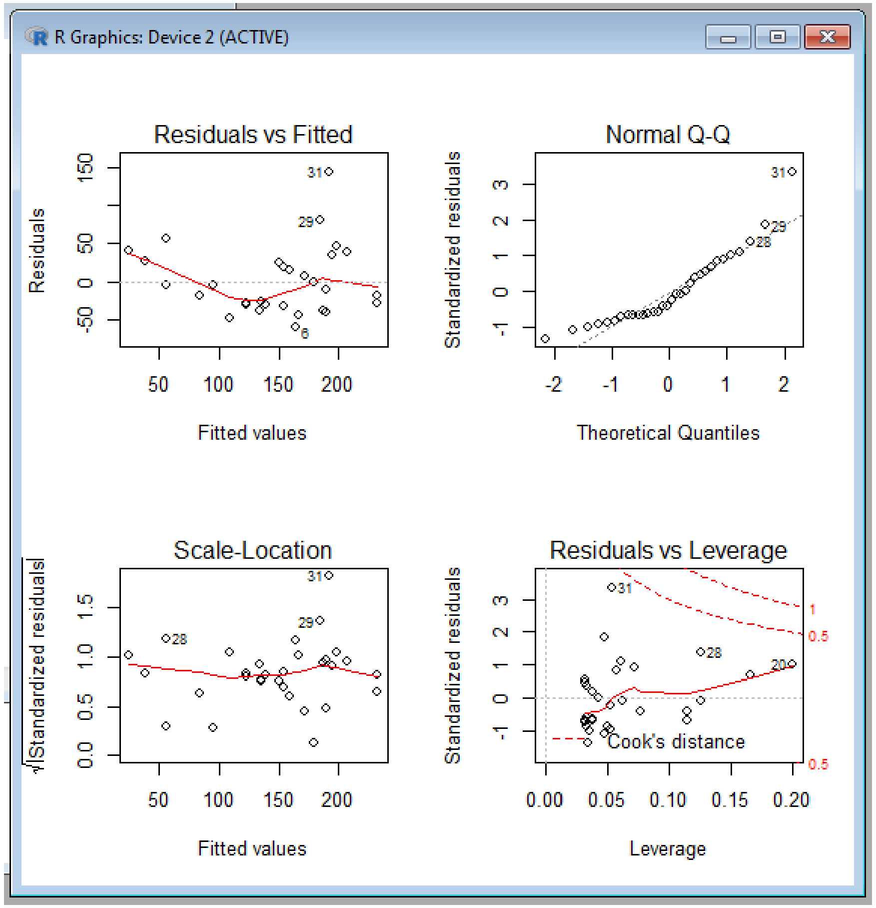
```
> summary(Z)
Call:
lm(formula = y ~ x)
Residuals:
   Min      1Q Median      3Q     Max
-59.26 -28.93 -13.45  25.65 143.36
Coefficients:
             Estimate  Std. Error  t value    Pr(>|t|)
(Intercept)   324.08      27.43   11.813   8.25e-13 ***
  x            -8.83       1.31   -6.742   1.79e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 43.95 on 30 degrees of freedom
Multiple R-squared:  0.6024,    Adjusted R-squared:  0.5892
F-statistic: 45.46 on 1 and 30 DF, p-value: 1.788e-07
```

SLR in R...

Model Diagnostics:

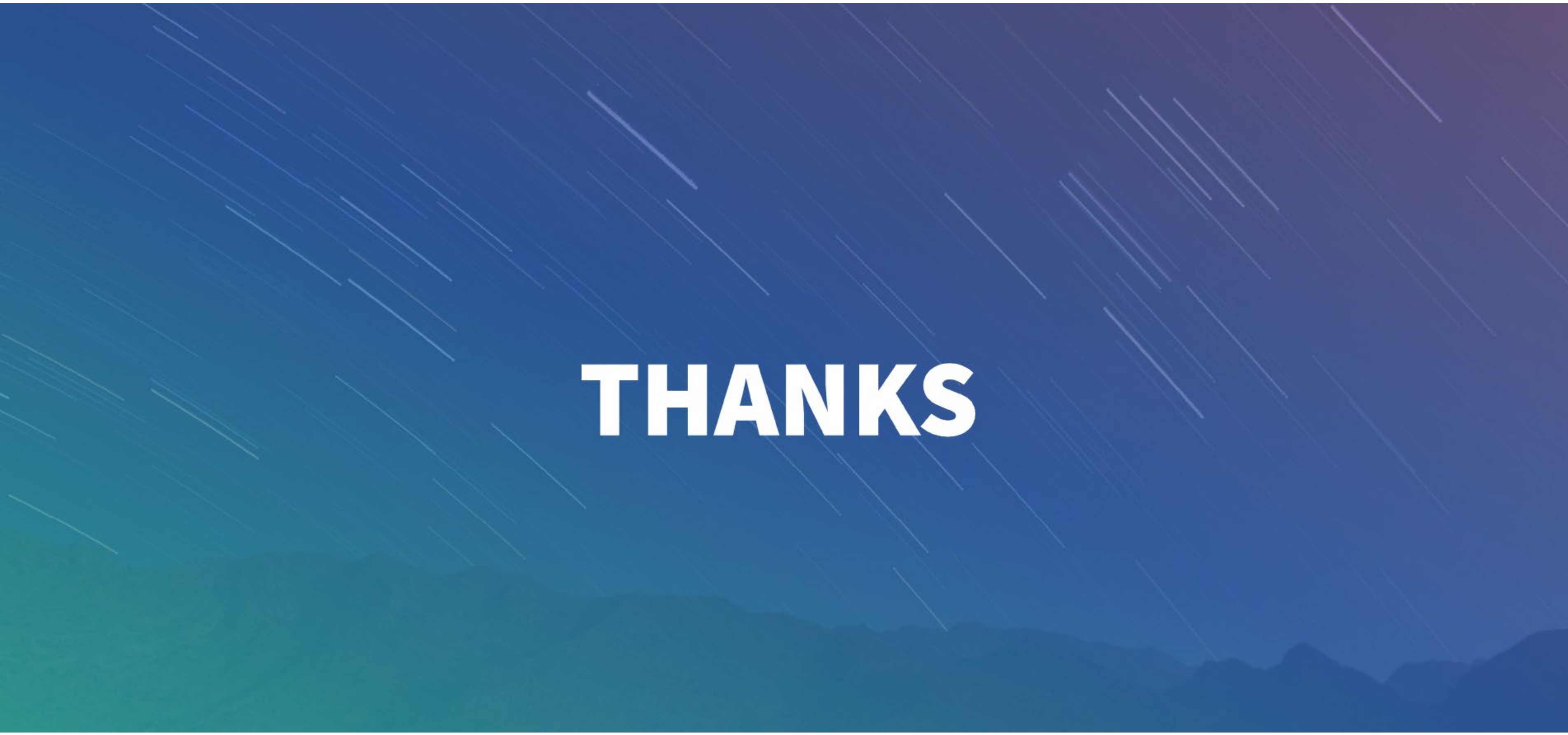
>par(mfrow=c(2,2))

>plot(Z)



SLR in R...

Refer the code given along with the data set mentioned in it for more detailed SLR in R.


THANKS