

# Week 6 Video Lecture Notes

## A. Correlation Coefficients and its Limitations

### Motivations

- how to determine the strength of a relationship? *Strong or Weak?*
  - how do we measure the relationship quantitatively

### The Correlation Coefficient

def: The **Correlation Coefficient**, denoted by  $r$ , is a measure of linear association between them

- summarizes the direction and strength of the association (ranging from  $-1$  to  $1$ )

### Interpreting the $r$ -value

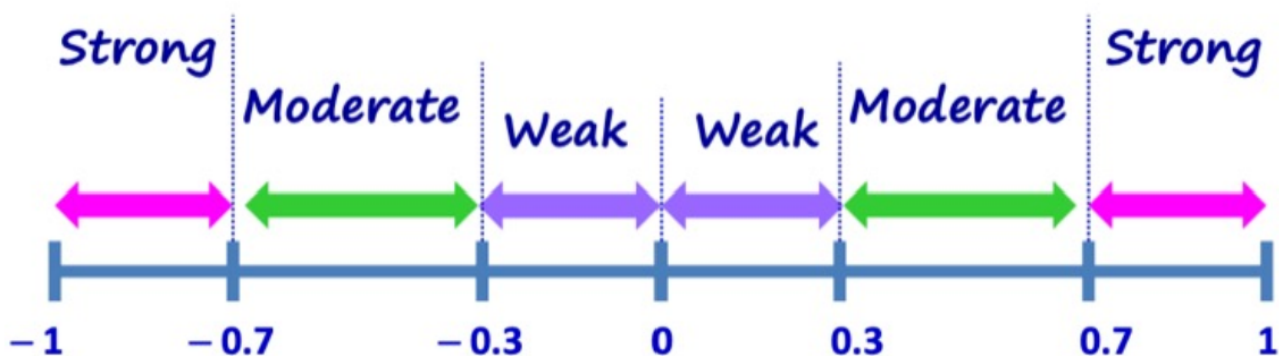
- the **sign** tells us the **direction** of the linear association,  $+$  for positive,  $-$  for negative and  $0$  for no linear association.
- the **magnitude of  $r$**  determines the **strength** of the linear association btwn two numerical variables
- **perfect linear association** occurs when all the data points line up on the line of best fit.

### Possible Values

- $r > 0 \implies$  positive linear association
- $r < 0 \implies$  negative linear association
- $r = 1 \implies$  perfect positive linear association
- $r = -1 \implies$  perfect negative linear association
- $r = 0 \implies$  no linear association

Magnitude of  $r$  tells us about the strength of the linear association

- closer the  $r$ -value to either  $-1$  or  $1$ , then the stronger the linear association
- closer the  $r$ -value to  $0$ , then the weaker the linear association



### Computation of Correlation Coefficient

Formulae:

$$\text{Standard Unit}_x = \frac{x - \bar{x}}{\sigma_x}$$

$$\text{Standard Unit}_y = \frac{y - \bar{y}}{\sigma_y}$$

Steps:

1. Compute mean and S.D. of  $x$  ( $\bar{x}$  and  $\sigma_x$ ) as well as mean and S.D. of  $y$  ( $\bar{y}$  and  $\sigma_y$ )

$x$	9	4	5	10	6	3	7	2	8	1
$y$	41	17	28	50	39	26	30	6	4	10

$$\bar{x} = \frac{9 + 4 + 5 + 10 + 6 + 3 + 7 + 2 + 8 + 1}{10}$$

$$= 5.5$$

$$\sigma_x = \sqrt{\frac{(9 - 5.5)^2 + (4 - 5.5)^2 + \dots + (1 - 5.5)^2}{10}}$$

$$= 2.87$$

$$\bar{y} = \frac{41 + 17 + \dots + 10}{10}$$

$$= 25.1$$

$$\sigma_y = \sqrt{\frac{(41 - 25.1)^2 + (17 - 25.1)^2 + \dots + (10 - 25.1)^2}{10}}$$

$$= 14.84$$

2. Compute  $x$  and  $y$  in their standard units for each data point, i.e.

$$SU_{x1} = \frac{9 - 5.5}{2.87} = 1.22(2 \text{ d.p.})$$

3. Compute  $SU_{x \cdot y}$  based on Step 2.

4. Compute  $r$  (sum the products obtained by  $SU_x$  and  $SU_y$ )

$$r = \frac{1}{N} \times \left( SU_{x1 \ y1} + SU_{x2 \ y2} + \dots + SU_{xN \ yN} \right)$$

## Important Properties of the Correlation Coefficient $r$

- $r$  is **not affected** by interchanging the  $x$  and  $y$  variables
- if we **add** (or **conversely subtract**) a constant  $k$  to all data points (supposing input error), it **does not affect** the value of  $r$
- if we **multiply** a constant  $m$  to all data points (supposing conversion or "shifting" by a certain factor  $m$ ), the value of  $r$  remains unaffected.

## Disadvantages of using Correlation Coefficient

1. Association is **not** causation.
  - there is often a strong association btwn variables (i.e.  $r \rightarrow 1$  or  $r \rightarrow -1$ ) which is misinterpreted as a cause-effect r/s
  - remember that statistical relationship  $\neq$  causal relationship
2. The correlation coefficient cannot tell us anything about non-linear relationships
  - $r$  does not apply for non-linear r/s since it does not give indications if variables are related non-linearly.
3. outliers may increase the **strength** ("closeness" to best fit) of the correlation coefficients
  - **removal of outliers** from the data set can have different effects on  $r$ !

## Ecological Correlations

- correlation based on aggregated "groups" at different points of the graph

### Disadvantage

- may overstate the strength of associations individuals **despite being in the same direction** as aggregates

def: **Ecological fallacy** occurs when one draws a false conclusion about correlation at the individual level (the big picture) based on what was observed at the aggregate level

def: **Atomistic fallacy** occurs when one mistakenly concludes that the same correlation observed at the individual level would *also exist* at the aggregate level.

Fallacy	Using	To conclude
Ecological	Ecological correlation (aggregate level)	Individual level correlation
Atomistic	Individual level correlation	Ecological correlation (aggregate level)

## B. Linear Regression

- used to "guess" some variable some time down the road through extrapolation.
- should not use regression line to predict values outside of the range of  $x$ .

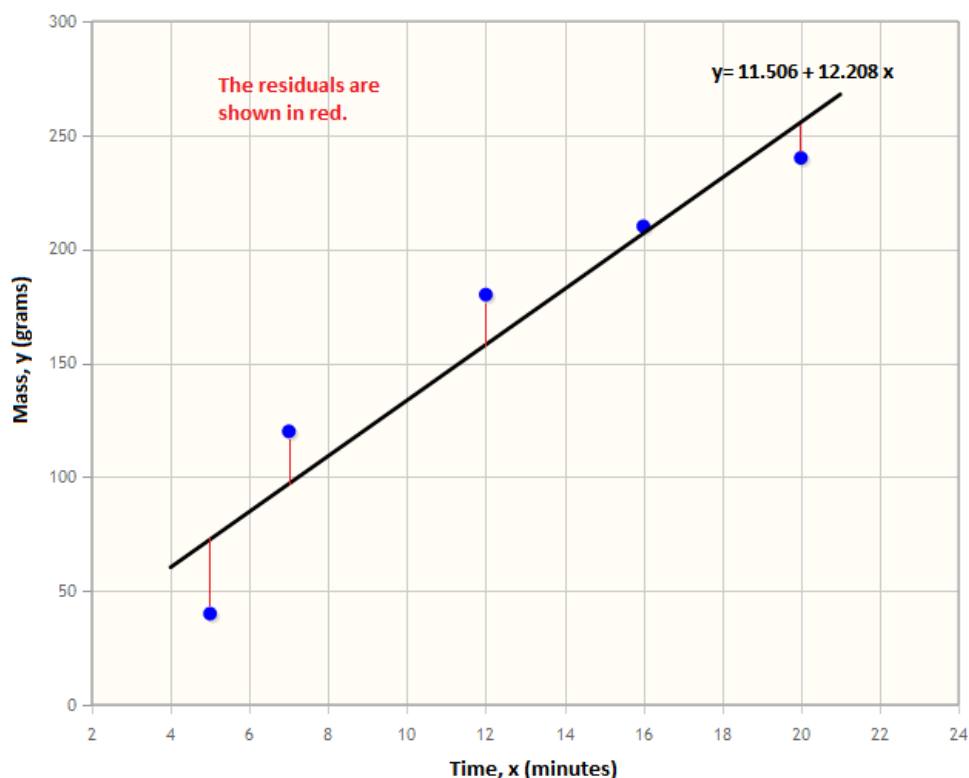
**Linear Association Formula:**

$$Y = mX + b$$

- where  $b \rightarrow y$  intercept;  $m \rightarrow$  gradient

### Residuals

- "error" of the  $i_{th}$  observation
  - formula:  $e_i = y_i - \hat{y}_i$

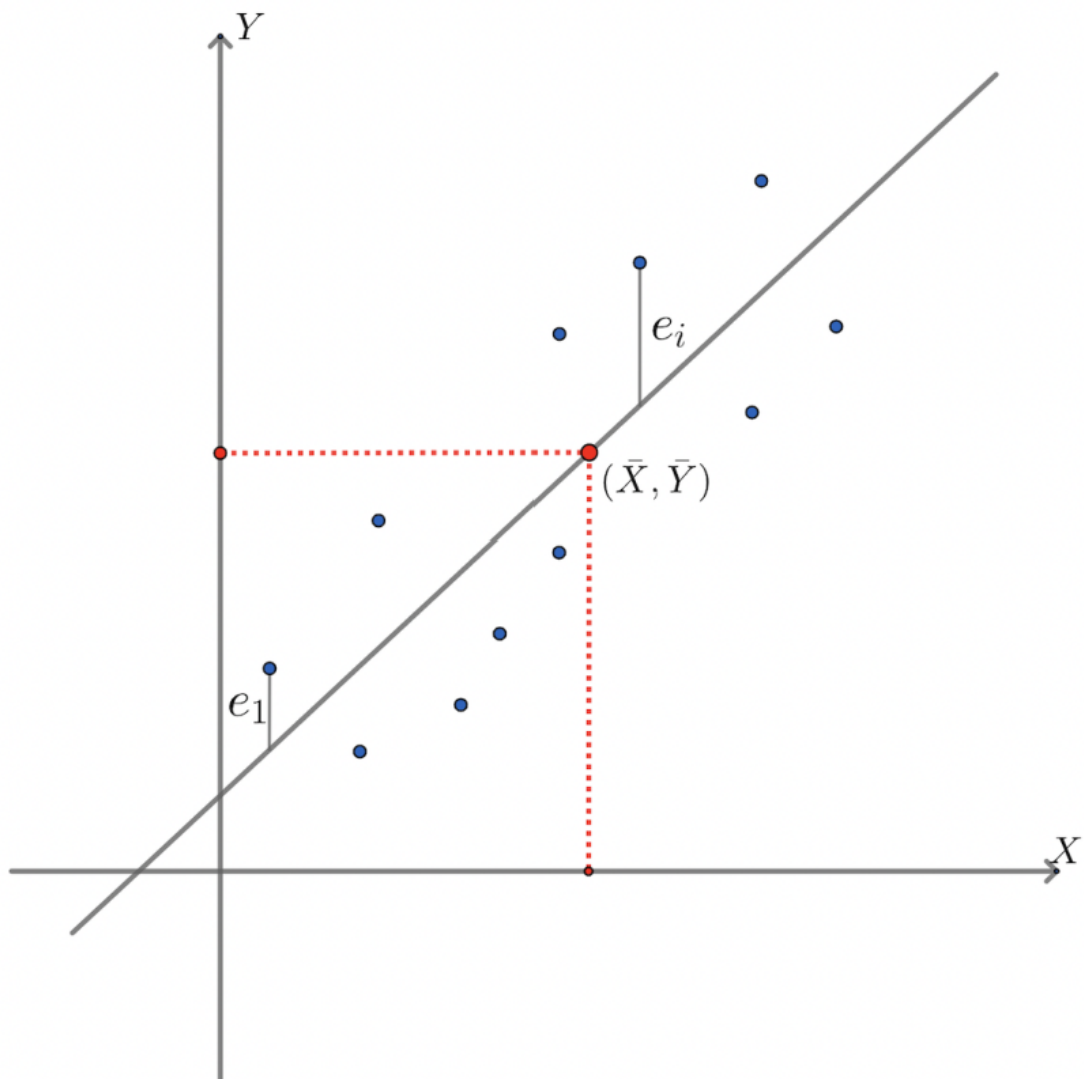


## Assessing Best Fit Lines

**Method of Least Squares** regression line - makes use of residuals

- need to square individual values  $\rightarrow$  same principle as S.D. computation

$$e_k = e_1^2 + e_2^2 + \dots + e_n^2 = \sum_{i=1}^k (e_i)^2$$



The line of best fit:

- will always pass through point of averages (i.e.  $\bar{x}$ ,  $\bar{y}$ )
- allows us to predict average resale price for a given age of a resale flat (but NOT vice versa)  $\implies$  two regression lines (price / age and age / price are not interchangeable!)

## Gradient of the regression line

Gradient is closely related to correlation coefficient  $r$ .

$$m = \frac{\sigma_y}{\sigma_x} \cdot r$$

- Gradient is the quotient of S.D. of  $y$  divided by S.D. of  $x$  multiply by  $r$ .
- $m$  may not equate to  $r$
- when  $r$  is positive, then  $m$  is also positive  $\rightarrow r_+ \implies m_+$  and  $r_- \implies m_-$  (vice versa)

---

## C. Studying non-linear relationships

- use a combination of correlation coefficient + scatter plot (for bivariate data) to test for non-linear relationships
  - show a **clear visual** of exponentiation or a quadratic / cubic function to disprove linearity and hence linear relationship between variables