Bringing British Education to You

www.nccedu.com

Skills for Computing

Topic 11:

Regression Analysis

Learning Outcomes for this Topic

By the end of this topic, students should be able to:

- Understand and use simple linear regression
- Understand and use Pearson's (product moment) correlation coefficient
- Understand and use Spearman's (rank order) correlation coefficient

Learning Outcomes

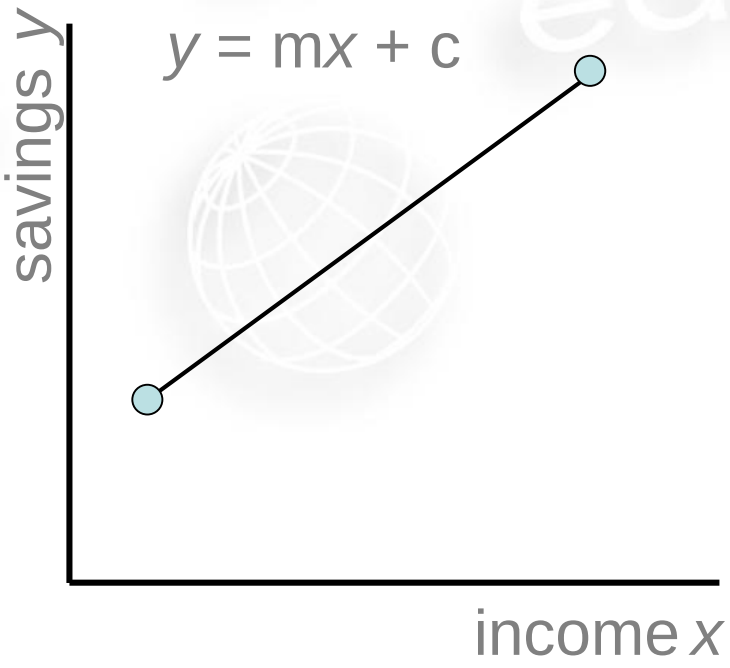
By the end of this topic students will be able to:

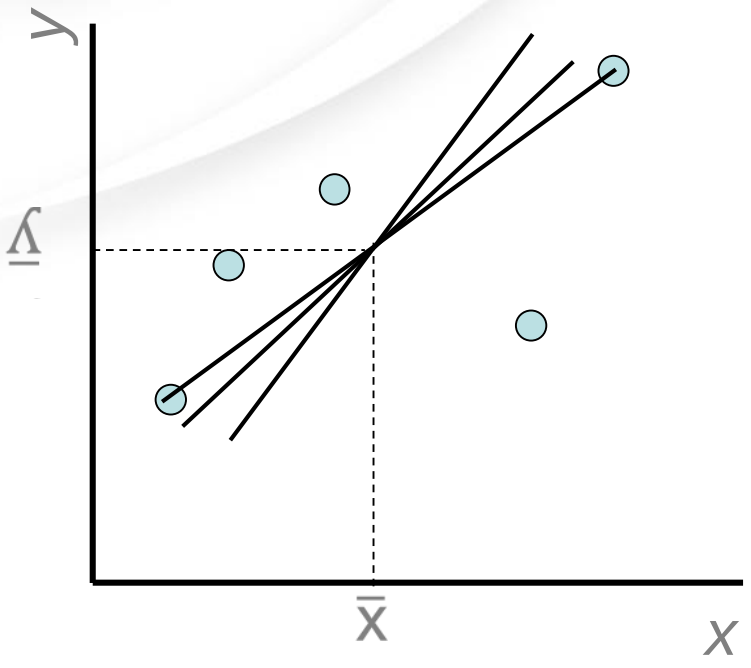
- Understand a straight line fit to bivariate data
- Calculate and interpret Pearson's correlation coefficient
- Calculate and interpret Spearman's correlation coefficient

Motivation

- The search for, and strength of, predictors
  - What is a good predictor of future job performance?
  - What is this product's price-demand curve?
  - Which process factors affect production yield?
  - How does a particular share price move with the market average?

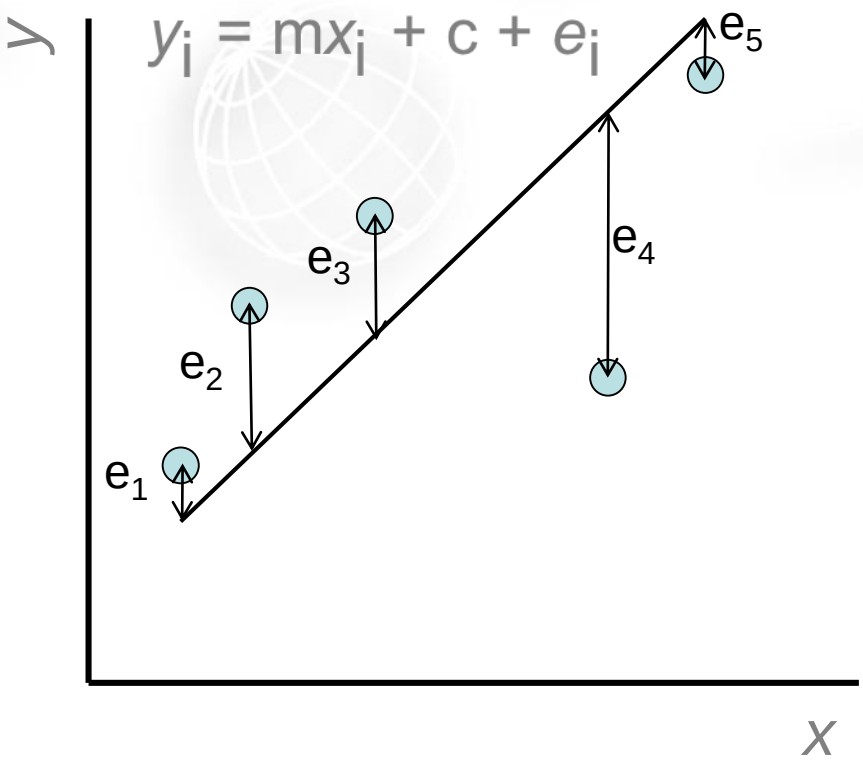
The Linear Relationship





- assuming interval or ratio data.

Least SSE Regression Criterion

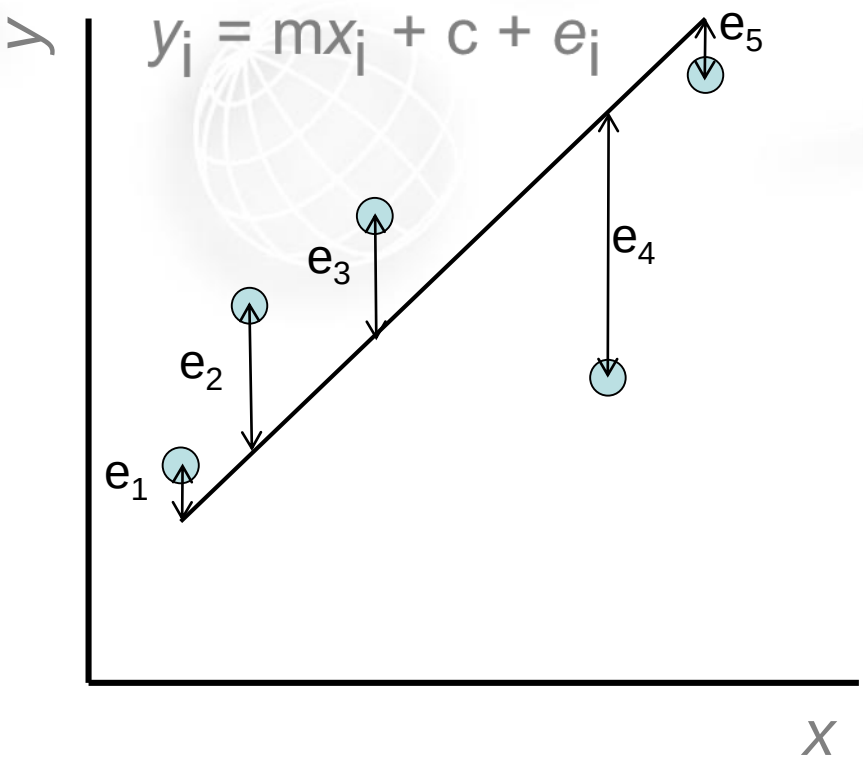


- $y_i = mx_i + c + e_i$
- The least squared line is the line that minimizes the sum of square errors

$$e_1^2 + e_2^2 + \cdots + e_n^2$$

- $\hat{y} = mx_i + c$

Least SSE Regression Criterion



- For the least SSE straight line, $\hat{y} = mx_i + c$

- $m = \dfrac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2}$

- $c = \bar{y} - m\bar{x}$

Least SSE Regression Criterion

- For the least SSE straight line, $\hat{y} = mx_i + c$
- $m = \dfrac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$
- $c = \bar{y} - m\bar{x}$

- For the least SSE straight line, $\hat{y} = mx_i + c$

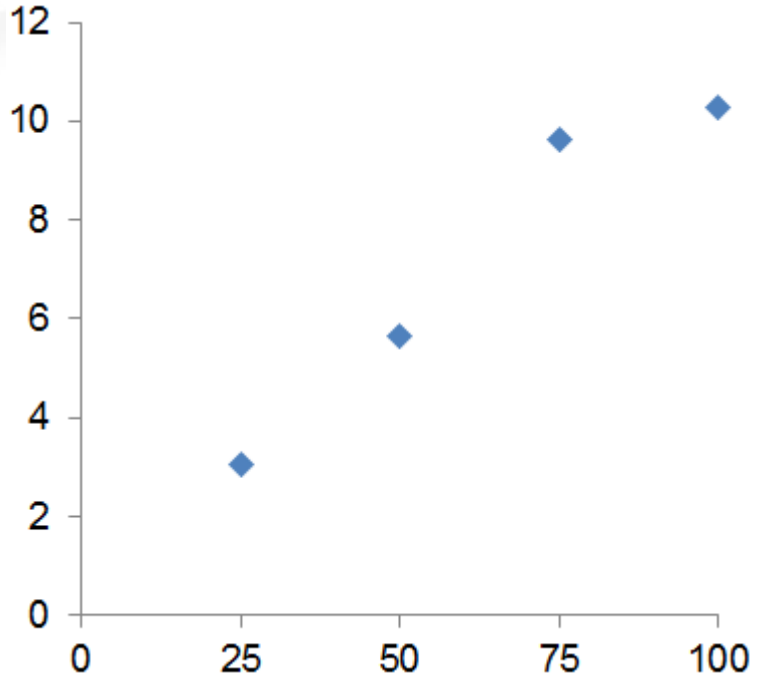
- $m = \dfrac{\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum\left(x_i - \bar{x}\right)^2}$

- $c = \bar{y} - m\bar{x}$

Example

- A company has been carrying out experiments with the position of a button on its website.

Page Position (% vertical)	Click Throughs (%)
25	3.07
50	5.64
75	9.63
100	10.26



Example

- A company has been carrying out experiments with the position of a button on its website.

	x	y	xy	x^2
	25	3.07	76.75	625
	50	5.64	282.00	2500
	75	9.63	722.25	5625
	100	10.26	1026.00	10000
total	**250**	**28.6**	**2107**	**18750**
mean	**62.5**	**7.15**		

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

$$m = \frac{\{(4 \times 2107) + (250 \times 28.6)\}}{\{(4 \times 18750) - (250 \times 250)\}}$$

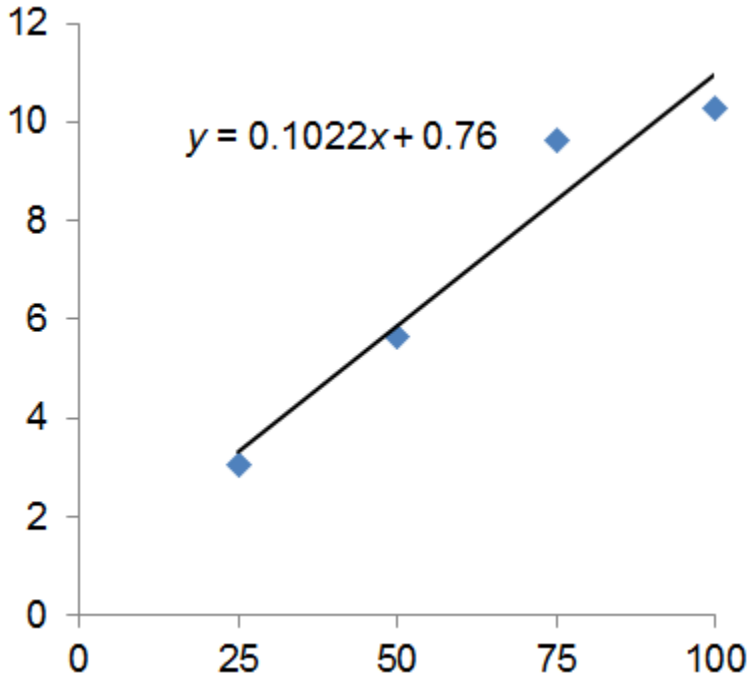
$$m = 0.10224, \quad c = \bar{y} - m\bar{x}$$

$$c = 7.15 - (0.10224 \times 62.5) = 0.76$$

Example

- A company has been carrying out experiments with the position of a button on its website.

Page Position (% vertical)	Click Throughs (%)
25	3.07
50	5.64
75	9.63
100	10.26



Exercise

- A company has been carrying out experiments with the position of a button on its website.

	x	y	xy	x^2
	25	1.44		5
	50	5.58		2500
	75	14.64		5625
	100	6.94		10000
total	**250**			**18750**
mean	**62.5**			

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

$$c = \bar{y} - m\bar{x}$$

Exercise

- A company has been carrying out experiments with the position of a button on its website.

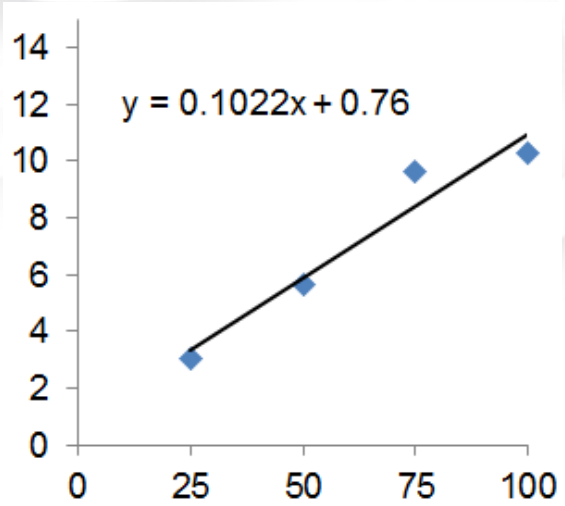
	x	y	xy	x^2
	25	1.44	36	5
	50	5.58	279	2500
	75	14.64	1098	5625
	100	6.94	694	10000
total	**250**	**28.6**	**2107**	**18750**
mean	**62.5**	**7.15**		

$$m = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = 0.10224$$

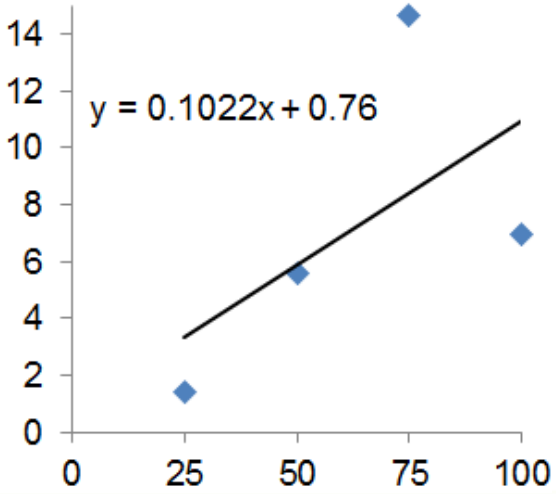
$$c = \bar{y} - m\bar{x} = 0.76$$

How Well Does the Line Fit?

x	*y*
25	3.07
50	5.64
75	9.63
100	10.26



x	*y*
25	1.44
50	5.58
75	14.64
100	6.94

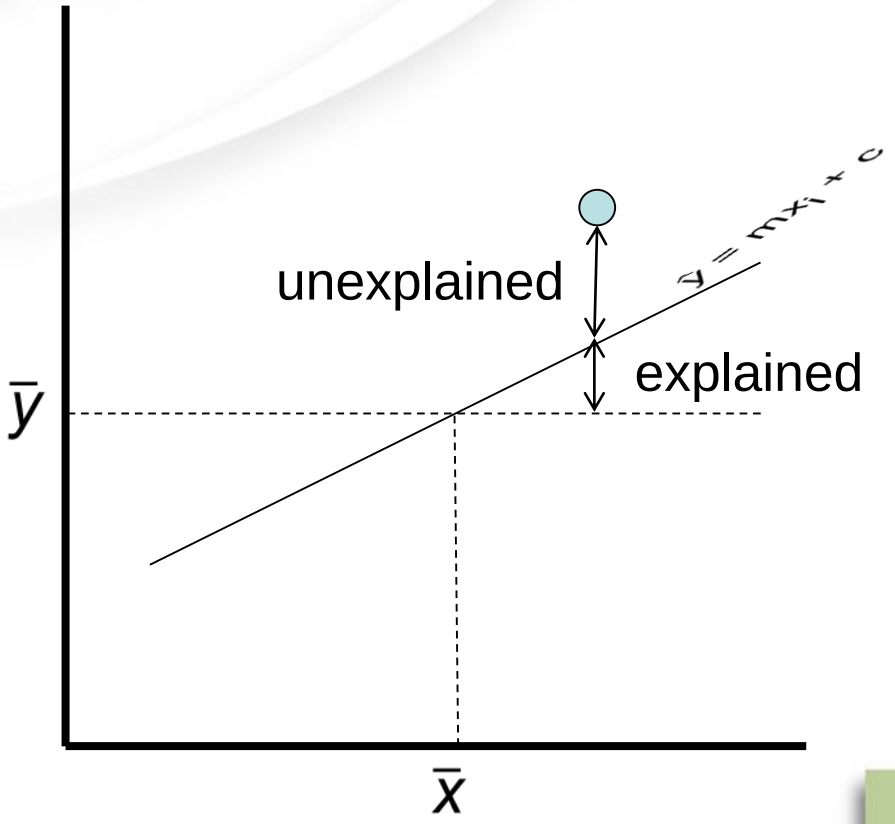


How Well Does the Line Fit?

- Total variation in two parts

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

- Total = unexplained + explained
- Fraction of the variation explained by the line
- $R^2 = r^2 = \dfrac{\Sigma(\hat{y} - \bar{y})^2}{\Sigma(y - \bar{y})^2}$



Pearson Correlation

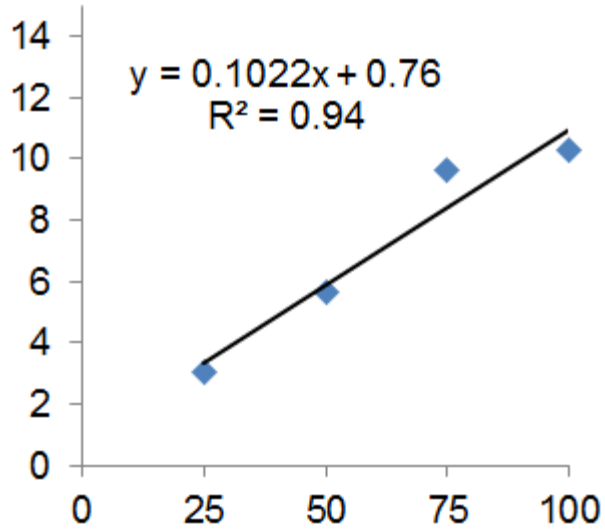
- R^2 or r^2 is called the coefficient of determination
- $0 \leq r^2 \leq 1$
- r is called the Pearson correlation coefficient
- $-1 \leq r \leq 1$
- Following rearrangement

$$R = r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{\left(n\sum x_i^2 - \left(\sum x_i\right)^2\right)\left(n\sum y_i^2 - \left(\sum y_i\right)^2\right)}}$$

How Well Does the Line Fit?

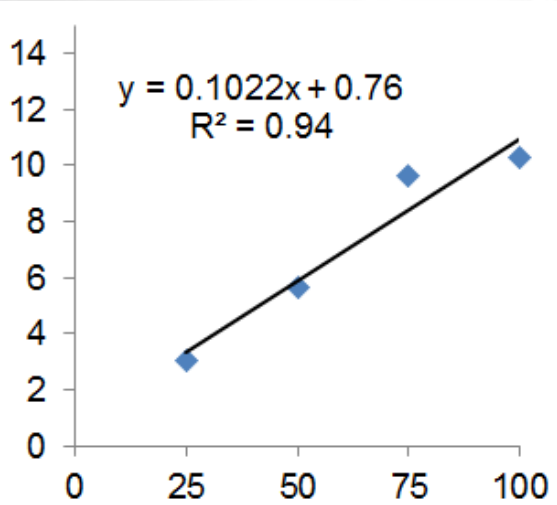
$$R = r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{\left(n\sum x_i^2 - \left(\sum x_i\right)^2\right)\left(n\sum y_i^2 - \left(\sum y_i\right)^2\right)}}$$

x	*y*	*xy*	x^2	y^2
25	3.07	76.75	625	9.425
50	5.64	282.00	2500	31.810
75	9.63	722.25	5625	92.737
100	10.26	1026.00	10000	105.268
250	**28.6**	**2107**	**18750**	**239.239**

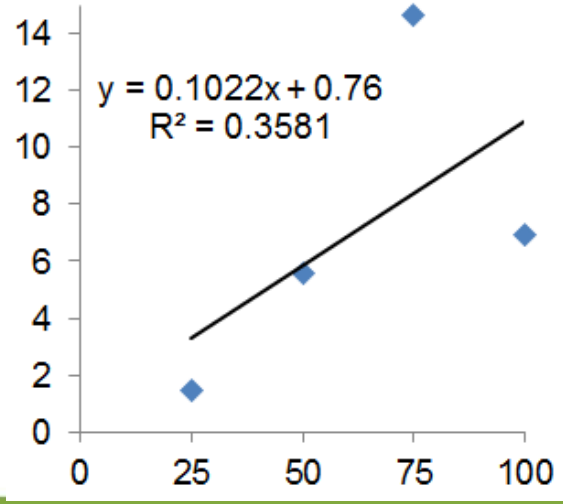


How Well Does the Line Fit?

x	*y*
25	3.07
50	5.64
75	9.63
100	10.26



x	*y*
25	1.44
50	5.58
75	14.64
100	6.94

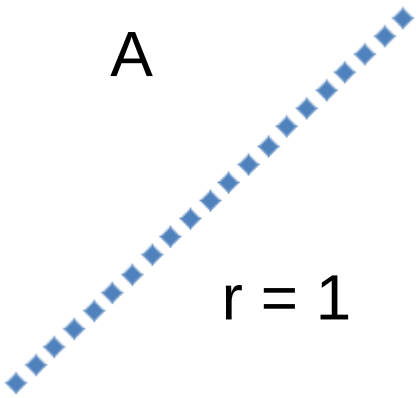


Language of Correlation

- Sign
  - $r > 0$ positive linear relationship
  - $r = 0$ no linear relationship
  - $r < 0$ negative linear relationship
- Strength
  - Physical sciences / engineering $R^2 > 0.6$ often found
  - Social sciences / policy $R^2 > 0.25$ sometimes useful
  - Business and management – includes science & social science!
  - But you will see language like $R^2 > 0.8$ strong, $R^2 > 0.5$ moderate, $R^2 > 0.25$ weak relationship
  - Be careful; context and numbers often more informative than descriptive word, but words help to communicate.

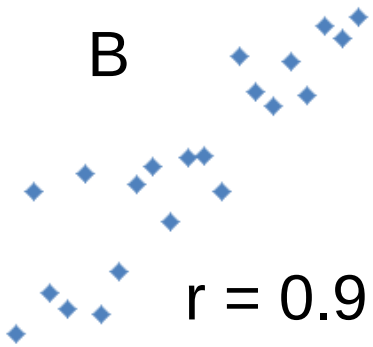
What is r?

A



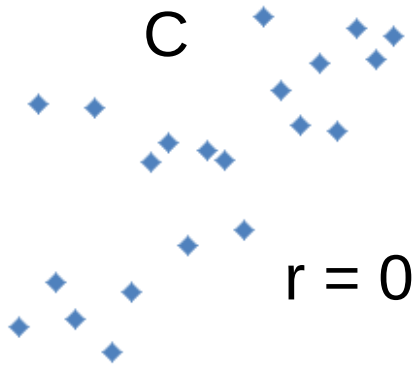
r = 1

B



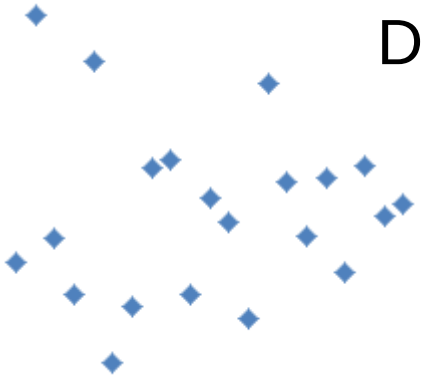
r = 0.9

C



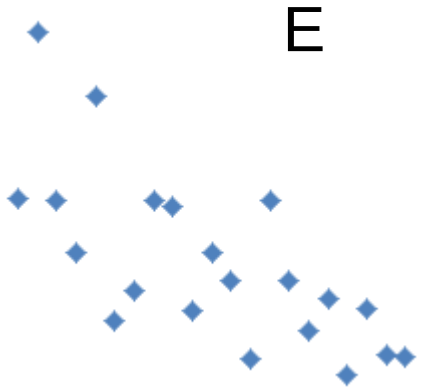
r = 0.7

D



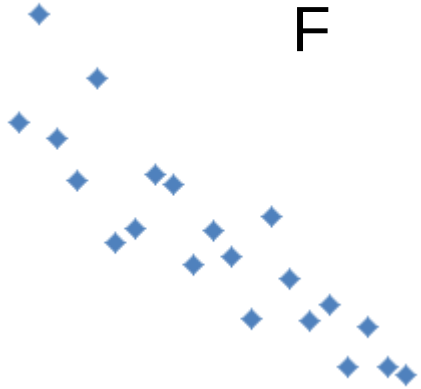
r = 0

E



r = -0.7

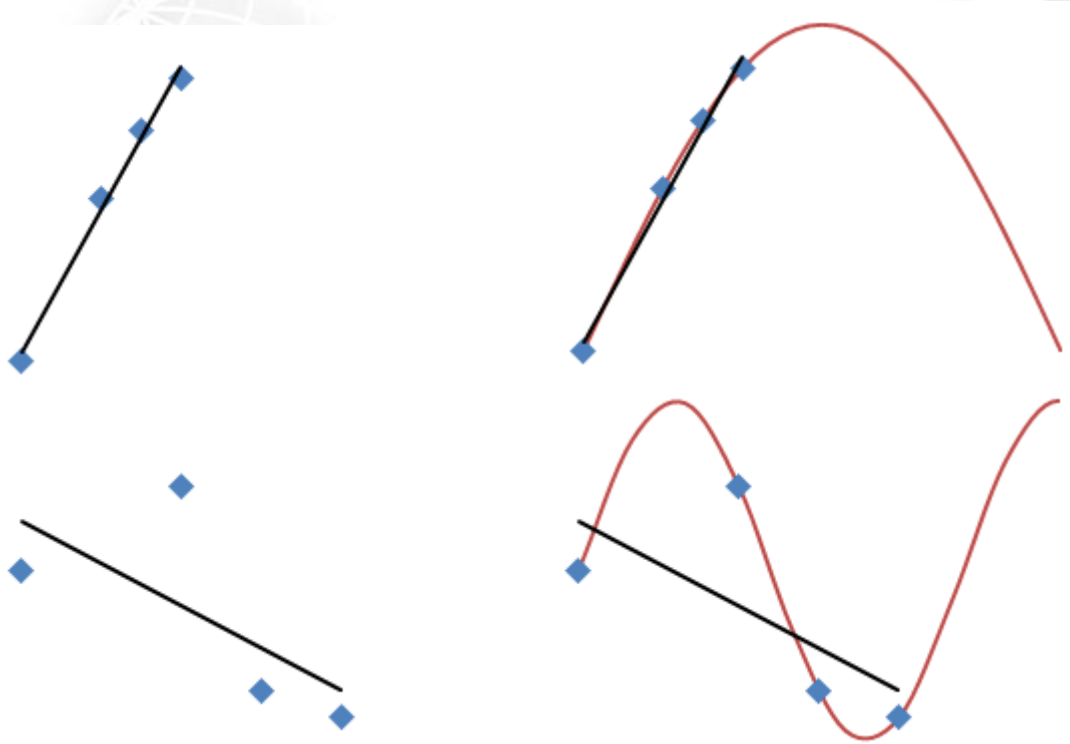
F



r = -0.9

Interpolation and Extrapolation

- Interpolation - Estimates between values already known
- Extrapolation - Estimates outside known values



Basics of Simple Linear Regression

- Plot scatter graph to intuit whether straight line is reasonable
- Look at $r^2 = R^2$ for strength of relationship
- Look at sign of r for direction
  - Check agrees with graph
- Look at m for gradient of relationship
- Use straight line equation to interpolate (with care)
- Use straight line equation to extrapolate (with caution)

Spearman's Rank Correlation

- Sometimes we only have **ordinal data**
  - two interviewers rank candidates
- Can we still define a correlation function? Yes
- $r_s = 1 - \frac{6 \sum d^2}{n(n^2-1)}$
  - where d is the difference between ranking.

Spearman's Correlation - Example

- Two interviewers individually rank prospective job candidates. What is the Spearman correlation coefficient?

Candidate	Interviewer 1	Interviewer 2		
Hidayat	3	E		
Elisa	2	A		
Nouman	1	B		
Bernie	4	C		
Li Ren	5	D		
Ahere	6	F		

Spearman's Correlation - Example

- Two interviewers individually rank prospective job candidates. What is the Spearman correlation coefficient?

Candidate	Interviewer 1	Interviewer 2	d	d^2
Hidayat	3	~~E~~ 5	-2	4
Elisa	2	~~A~~ 1	1	1
Nouman	1	~~B~~ 2	-1	1
Bernie	4	~~C~~ 3	1	1
Li Ren	5	~~D~~ 4	1	1
Ahere	6	~~F~~ 6	0	0

- $r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{8}{35} = 0.771$

Spearman Correlation - Ties?

- For tied ranks use mean rank, then
- Use formula for **Pearson** correlation

Candidate	Interviewer 1	Interviewer 2
Hidayat	3	OK
Elisa	2	Excellent
Nouman	1	Good
Bernie	4	OK
Li Ren	5	OK
Ahere	6	Poor

Recap

By the end of this topic students will be able to:

- Understand a straight line fit to bivariate data
- Calculate and interpret Pearson's correlation coefficient
- Calculate and interpret Spearman's correlation coefficient

Bibliography

- Burton, G., Carrol, G. and Wall, S. *Quantitative Methods for Business and Economics*. Longman.
- Buglear, J. *Quantitative Methods for Business*. Elsevier Butterworth Heinemann
- Hinton, PR. *Statistics Explained*. Routledge

Topic 11 – Regression Analysis

Any Questions?

Bringing British Education to You

www.nccedu.com