Correlation

Lecture 44
Section 13.7

Robb T. Koether

Hampden-Sydney College

Fri, Apr 13, 2012

Outline

Outline

- So far our only way to judge the strength and direction of a linear relationship has been to look at a scatterplot and go with our feelings.

Introduction

- So far our only way to judge the strength and direction of a linear relationship has been to look at a scatterplot and go with our feelings.
- Statisticians don't like to go with their feelings.

Introduction

- So far our only way to judge the strength and direction of a linear relationship has been to look at a scatterplot and go with our feelings.
- Statisticians don't like to go with their feelings.
- Statisticians like to quantify everything.

Introduction

- So far our only way to judge the strength and direction of a linear relationship has been to look at a scatterplot and go with our feelings.
- Statisticians don't like to go with their feelings.
- Statisticians like to quantify everything.
- So we need to quantify the strength of the relationship.

Outline

The Correlation Coefficient

- The correlation coefficient r measures both the *direction* and *strength* of the linear relationship.
- Its value is always between -1 and $+1$.

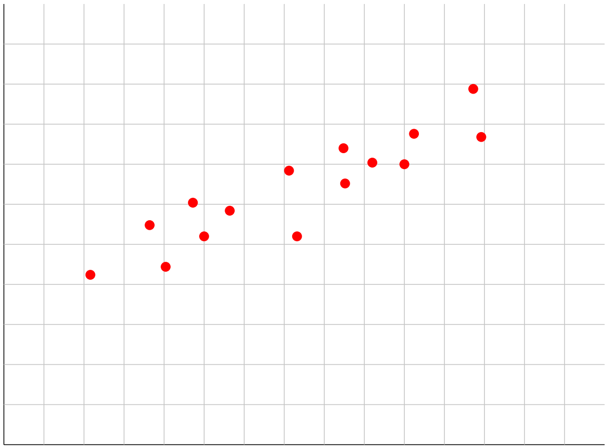
The Correlation Coefficient

- If $r > 0$, then the relationship is positive.
- If $r < 0$, then the relationship is negative.

The Correlation Coefficient

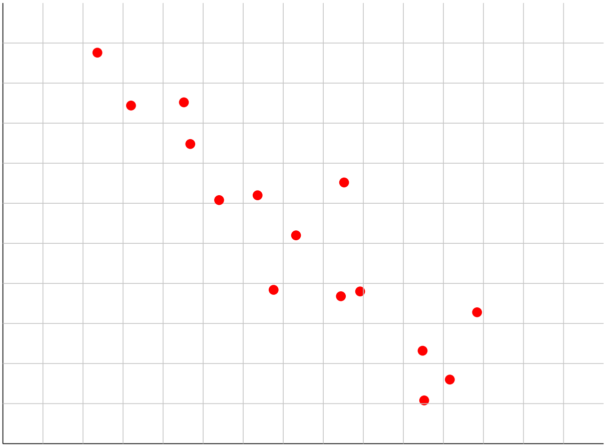
- If r is close to $+1$ or -1, the relationship is strong.
- If r is close to 0, the relationship is weak.
- In other words, $|r|$ measures the strength of the relationship.

Strong Positive Linear Association



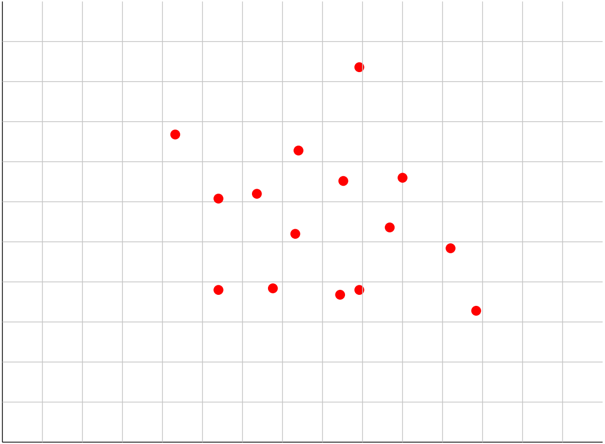
r is close to $+1$

Strong Positive Linear Association



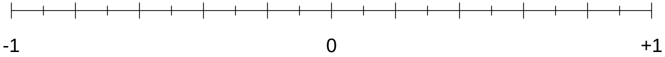
r is close to -1

Strong Positive Linear Association

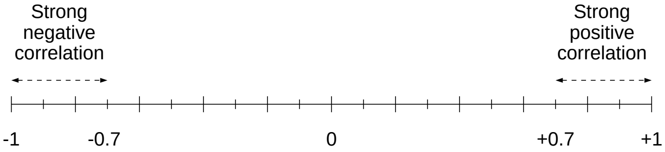


r is close to 0

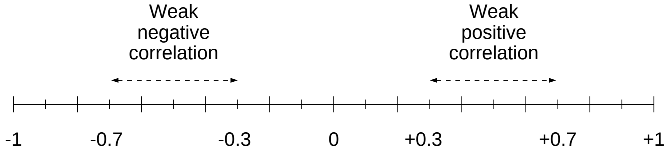
Strong Positive Linear Association



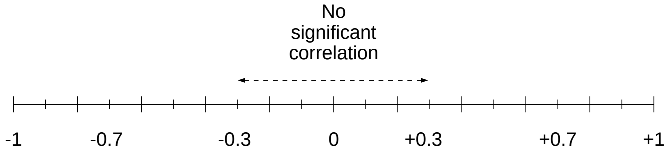
Strong Positive Linear Association



Strong Positive Linear Association



Strong Positive Linear Association



Outline

Calculating the Correlation Coefficient

- There are many formulas for r.
- The basic formula is

$$r = \frac{\sum (x - \overline{x})(y - \overline{y})}{\sqrt{\sum (x - \overline{x})^2 \sum (y - \overline{y})^2}}$$

which has the alternative form

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

Calculating the Correlation Coefficient

- If we define

$$\text{SSX} = \sum (x - \overline{x})^2,$$
$$\text{SSY} = \sum (y - \overline{y})^2,$$
$$\text{SSXY} = \sum (x - \overline{x})(y - \overline{y}),$$

then the formula simplifies to

$$r = \frac{\text{SSXY}}{\sqrt{\text{SSX} \cdot \text{SSY}}}.$$

Example

Example (Correlation Coefficient)

- Consider the following height and weight data.

Height (x)	Weight (y)
70	185
65	140
71	180
76	220
68	150
67	170
68	185
72	200
74	210
69	160

Example

Example (Correlation Coefficient)

- Enter x into list $\mathtt{L}_1$ and y into list $\mathtt{L}_2$.
- Then enter `2-Var Stats L1,L2`.
- We get $\overline{x} = 70$, $\overline{y} = 180$,

Example

Example (Correlation Coefficient)

- Compute the deviations, squared deviations, and their sums.

Height (x)	Weight (y)	$x - \overline{x}$	$(x - \overline{x})^2$	$y - \overline{y}$	$(y - \overline{y})^2$	$(x - \overline{x})(y - \overline{y})$
70	185					
65	140					
71	180					
76	220					
68	150					
67	170					
68	185					
72	200					
74	210					
69	160					

Example (Correlation Coefficient)

- Compute the deviations, squared deviations, and their sums.

Height (x)	Weight (y)	$x - \overline{x}$	$(x - \overline{x})^2$	$y - \overline{y}$	$(y - \overline{y})^2$	$(x - \overline{x})(y - \overline{y})$
70	185	0				
65	140	-5				
71	180	1				
76	220	6				
68	150	-2				
67	170	-3				
68	185	-2				
72	200	2				
74	210	4				
69	160	-1				

Example

Example (Correlation Coefficient)

- Compute the deviations, squared deviations, and their sums.

Height (x)	Weight (y)	$x - \overline{x}$	$(x - \overline{x})^2$	$y - \overline{y}$	$(y - \overline{y})^2$	$(x - \overline{x})(y - \overline{y})$
70	185	0	0			
65	140	−5	25			
71	180	1	1			
76	220	6	36			
68	150	−2	4			
67	170	−3	9			
68	185	−2	4			
72	200	2	4			
74	210	4	16			
69	160	−1	1			
			100			

Example

Example (Correlation Coefficient)

- Compute the deviations, squared deviations, and their sums.

Height (x)	Weight (y)	$x - \overline{x}$	$(x - \overline{x})^2$	$y - \overline{y}$	$(y - \overline{y})^2$	$(x - \overline{x})(y - \overline{y})$
70	185	0	0	5		
65	140	−5	25	−40		
71	180	1	1	0		
76	220	6	36	40		
68	150	−2	4	−30		
67	170	−3	9	−10		
68	185	−2	4	5		
72	200	2	4	20		
74	210	4	16	30		
69	160	−1	1	−20		
			100			

Example

Example (Correlation Coefficient)

- Compute the deviations, squared deviations, and their sums.

Height (x)	Weight (y)	$x - \overline{x}$	$(x - \overline{x})^2$	$y - \overline{y}$	$(y - \overline{y})^2$	$(x - \overline{x})(y - \overline{y})$
70	185	0	0	5	25	
65	140	−5	25	−40	1600	
71	180	1	1	0	0	
76	220	6	36	40	1600	
68	150	−2	4	−30	900	
67	170	−3	9	−10	100	
68	185	−2	4	5	25	
72	200	2	4	20	400	
74	210	4	16	30	900	
69	160	−1	1	−20	400	
			100		5950	

Example

Example (Correlation Coefficient)

- Compute the deviations, squared deviations, and their sums.

Height (x)	Weight (y)	$x-\overline{x}$	$(x-\overline{x})^2$	$y-\overline{y}$	$(y-\overline{y})^2$	$(x-\overline{x})(y-\overline{y})$
70	185	0	0	5	25	0
65	140	−5	25	−40	1600	200
71	180	1	1	0	0	0
76	220	6	36	40	1600	240
68	150	−2	4	−30	900	60
67	170	−3	9	−10	100	30
68	185	−2	4	5	25	−10
72	200	2	4	20	400	40
74	210	4	16	30	900	120
69	160	−1	1	−20	400	20
			100		5950	700

Example

Example (Correlation Coefficient)

- We now have

$$\begin{aligned} SSX &= 100, \\ SSY &= 5950, \\ SSXY &= 700. \end{aligned}$$

Example

Example (Correlation Coefficient)

- Then calculate,

$$r = \frac{700}{\sqrt{(100)(5950)}} = 0.9075.$$

- Thus, there is a *strong positive linear association* between height and weight.

Outline

TI-83 Calculating *r*

- Be sure that `Diagnostic` is turned on (once and for all).
- (To do so, press `CATALOG` and select `DiagnosticOn`, and press `ENTER` twice.)
- Then follow the procedure that produces the regression line.
- In the same window, the TI-83 reports the values of r^2 and r.

Calculating r on the IT-83

Practice

- Use the TI-83 to calculate r in the preceding example.

Outline

Correlation vs. Cause and Effect

- If the value of r is close to $+1$ or -1, that indicates that x is a good predictor of y.
- It does not indicate that x causes y.
- The correlation coefficient alone cannot be used to determine cause and effect.

Correlation vs. Cause and Effect

- For example, one's shoe size is a very good predictor of one's score on the SAT Math test.
- How so?

Correlation vs. Cause and Effect

- For example, one's shoe size is a very good predictor of one's score on the SAT Math test.
- How so?
- Well, assuming that the person is between 3 and 21 years old.

Correlation vs. Cause and Effect

- People who wear larger shoe sizes tend to have larger feet.

Correlation vs. Cause and Effect

- People who wear larger shoe sizes tend to have larger feet.
- People with larger feet tend to be larger people.

Correlation vs. Cause and Effect

- People who wear larger shoe sizes tend to have larger feet.
- People with larger feet tend to be larger people.
- Larger people tend to be older people, at least within the age range 3 to 21.

Correlation vs. Cause and Effect

- People who wear larger shoe sizes tend to have larger feet.
- People with larger feet tend to be larger people.
- Larger people tend to be older people, at least within the age range 3 to 21.
- Older people tend to have had more education, at least within the age range 3 to 21.

Correlation vs. Cause and Effect

- People who wear larger shoe sizes tend to have larger feet.
- People with larger feet tend to be larger people.
- Larger people tend to be older people, at least within the age range 3 to 21.
- Older people tend to have had more education, at least within the age range 3 to 21.
- People with more education tend to have been taught more math.

Correlation vs. Cause and Effect

- People who wear larger shoe sizes tend to have larger feet.
- People with larger feet tend to be larger people.
- Larger people tend to be older people, at least within the age range 3 to 21.
- Older people tend to have had more education, at least within the age range 3 to 21.
- People with more education tend to have been taught more math.
- People who have been taught more math, tend to know more math.

Correlation vs. Cause and Effect

- People who wear larger shoe sizes tend to have larger feet.
- People with larger feet tend to be larger people.
- Larger people tend to be older people, at least within the age range 3 to 21.
- Older people tend to have had more education, at least within the age range 3 to 21.
- People with more education tend to have been taught more math.
- People who have been taught more math, tend to know more math.
- People who know more math tend to do better on the SAT Math test.

Correlation vs. Cause and Effect

- However, do not go out and buy larger shoes, hoping to improve your score on the final exam.

Correlation vs. Cause and Effect

- However, do not go out and buy larger shoes, hoping to improve your score on the final exam.
- Neither should you avoid studying math for fear that your feet will get larger.

Outline

Assignment

Homework

- Read Section 13.7, pages 841 - 854.
- Let's Do It! 13.8, 13.9, 13.10, 13.12.
- Exercises 13 - 16, 22 - 24, 26 - 29, page 858.