Lesson 4
Part 1

Relationships between two numerical variables

Correlation Coefficient

The correlation coefficient is a summary statistic that describes the linear relationship between two numerical variables

Relationship between two variables

- The summary statistics covered in the previous lessons are appropriate for describing a single variable.
- In many investigations, there is interest in evaluating the relationship between two variables.
- Depending on the type of data, different statistics are used to measure the relationship between two variables.

Measuring relationship between two variables

- Nominal binary data
  - Odds Ratio or Risk Ratio (covered in Lesson 4 part 2)
- Numerical data
  - Pearson Correlation Coefficient
- Ordinal data or Skewed Numerical data
  - Spearman Rank Correlation Coefficient

Correlation Coefficients

- Correlation coefficients measure the *linear* relationship between two numerical variables.
- Most often the ***Pearson Correlation Coefficient*** is used to describe the linear relationship
  - Note if just the term ‘correlation coefficient’ is used, it is referring to the Pearson Correlation Coefficient
- If one variable is numerical and the other variable is ordinal or if both variables are ordinal, use the ***Spearman rank correlation*** coefficient instead.

Correlation Analysis Examples

For correlation analysis, you have pairs of values.

The goal is to measure the linear relationship between these values.

For example, what is the linear relationship between:

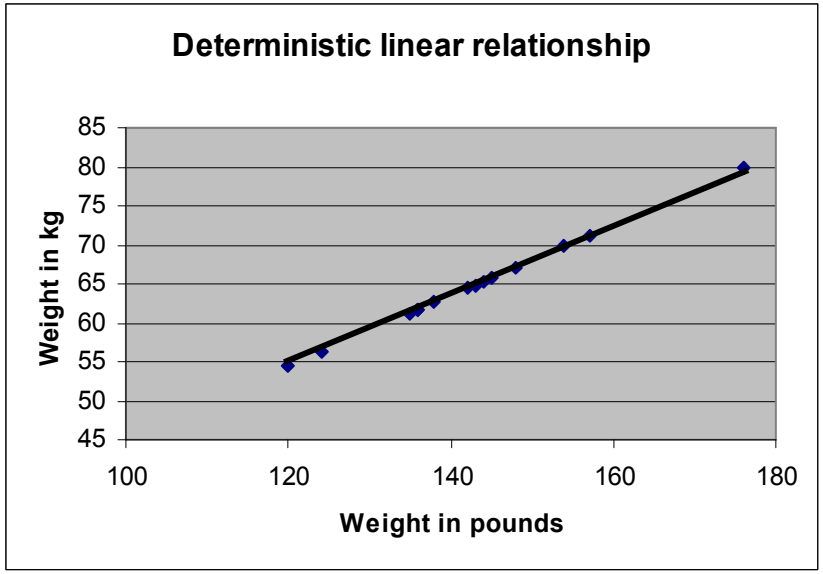
- Weight and height
- Blood pressure and age
- Daily fat intake and cholesterol
- Weight and cholesterol
- Age and bone density

The pairs of values are labeled (x,y) and can be plotted on a scatter plot

Types of Relationships between 2 variables

- There are 2 types of relationships between numerical variables
  - *Deterministic relationship* – the values of the 2 variables are related through an exact mathematical formula.
  - *Statistical relationship* – this is not a **perfectly linear** relationship. This is what we will explore through correlation analysis.

Example of Deterministic Linear relationship

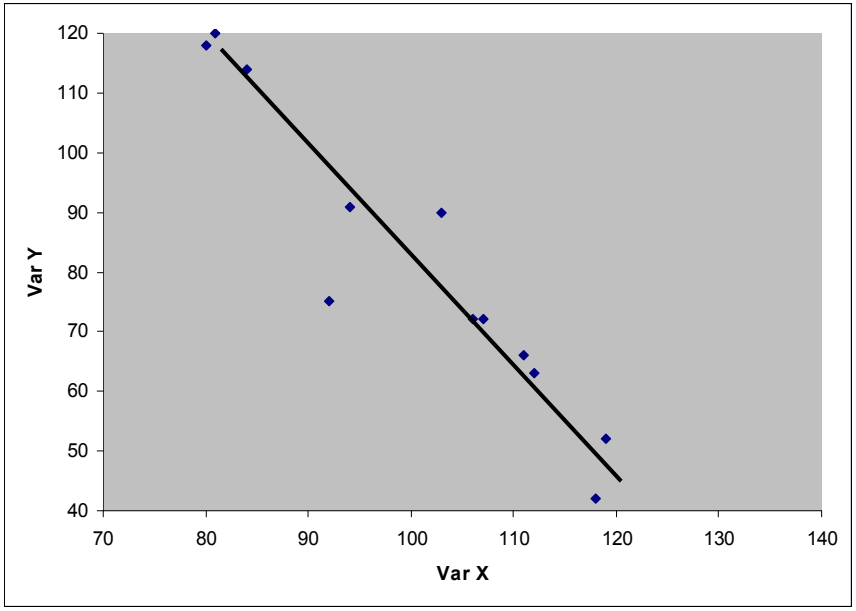


The relationship between weight measured in lbs. And weight measured in kg. is deterministic because it can be described with a Mathematical formula:

1 pound = 0.4536 *kg

All the points in the plot fall on a straight line in a deterministic relationship

Statistical Linear relationship

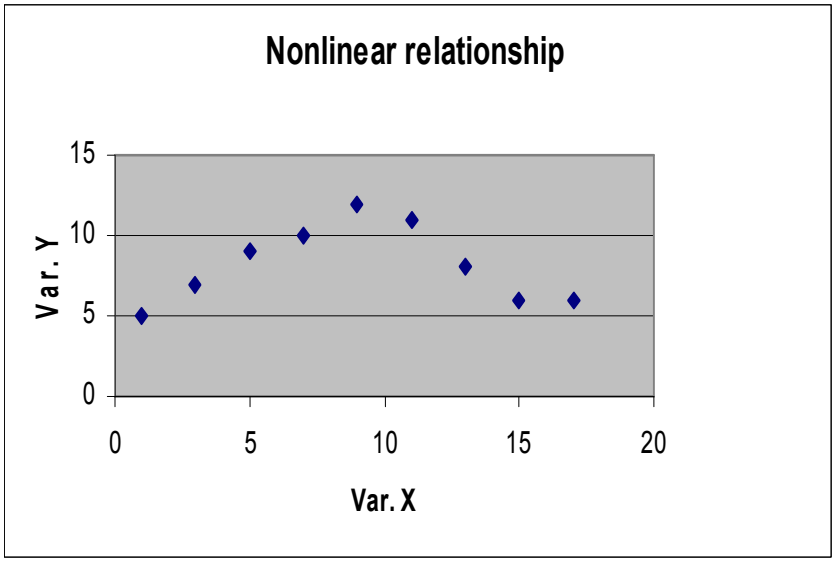


In a statistical linear relationship the plotted (x,y) values do not fall exactly on a straight line.

You can see, however, that there is a general linear trend in the plot of the (x,y) values.

The correlation coefficient provides information about the direction and strength of the linear relationship between two numerical values.

Nonlinear Relationship

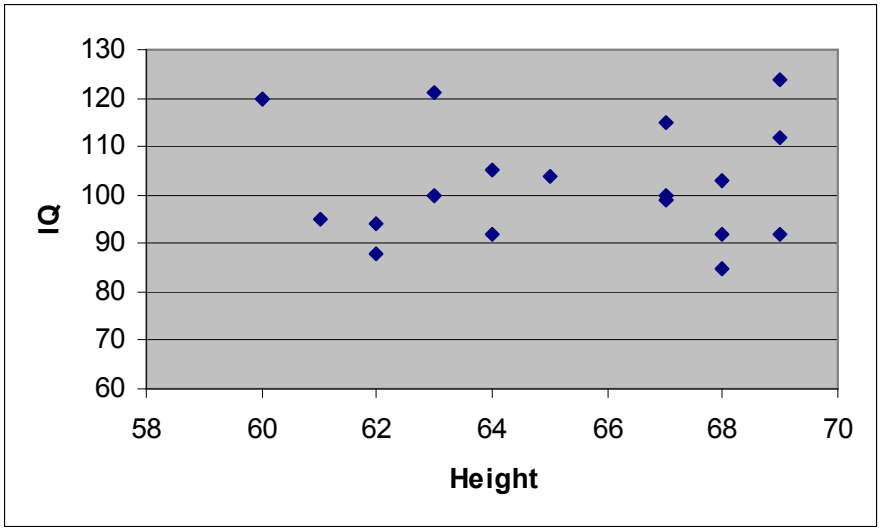


A nonlinear relationship between two variables results in a scatter plot of the (x,y) values that do not follow a straight line pattern.

The relationship between these two variables is a curvilinear relationship.

The correlation coefficient does not provide a description of the strength and direction of nonlinear relationships.

Random scatter: no pattern



This scatter plot of IQ and height for 18 individuals illustrates the situation where there is no relationship between the two variables.

The (x,y) values do not have either a linear or curvilinear pattern. These points have a random scatter, indicating no relationship.

Correlation Analysis Procedure

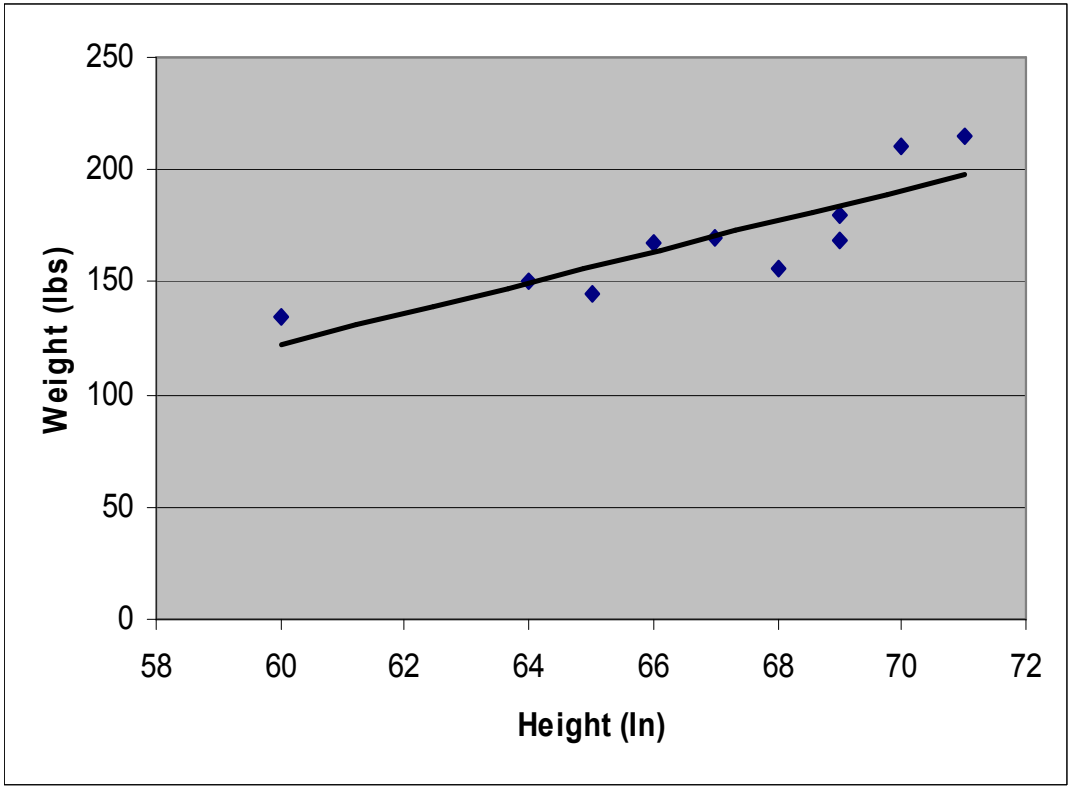
1. Plot the data using a scatter plot to get a visual idea of the relationship
    1. If there is a linear pattern, continue
    2. If the pattern is curved, stop
2. Calculate the correlation coefficient
3. Interpret the correlation coefficient

Scatter Plots

- Plot the 2 variables in a scatter plot (one of the types of charts in EXCEL).
- The pattern of the “dots” in the plot indicates the direction of the *statistical relationship* between the variables
  - Positive correlation
    - pattern goes from lower left to upper right
    - Increase in ‘X’ is associated with increase in ‘Y’
  - Negative correlation
    - pattern goes from upper left to lower right
    - Increase in ‘X’ is associated with decrease in ‘Y’

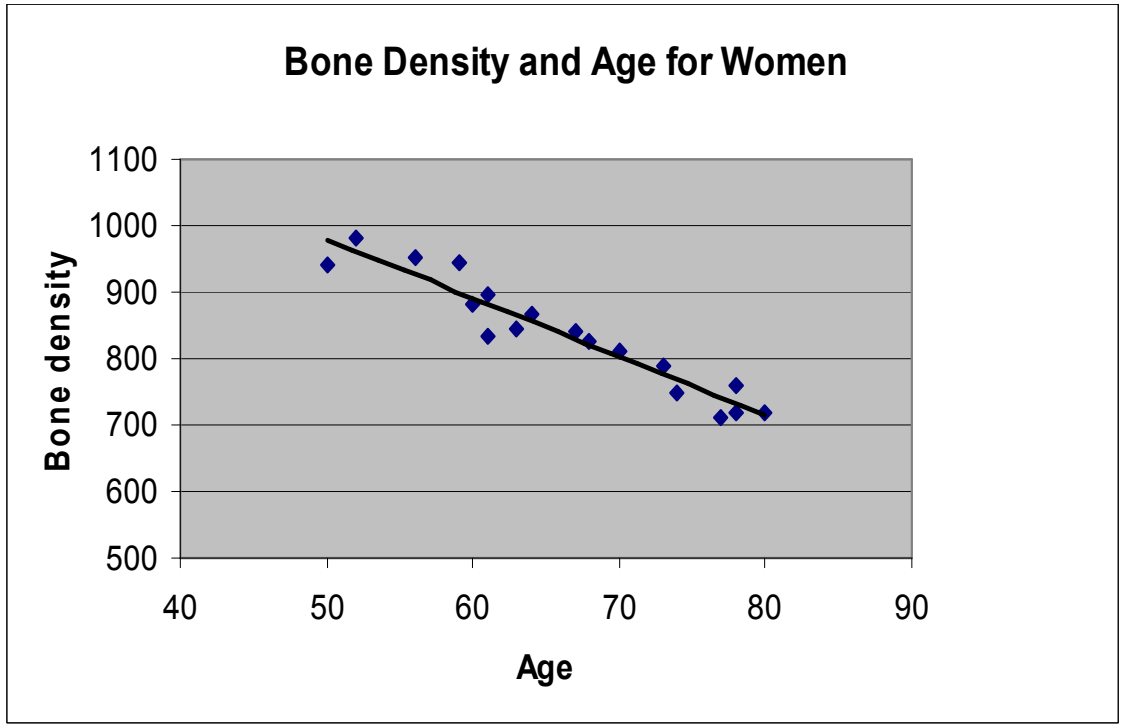
Height and weight: a positive linear relationship

Ht (in)	Wt (lbs)
64	150
67	170
69	180
60	135
68	156
69	168
70	210
71	215
65	145
66	167



Age and Bone Density (mg/cm^2): a negative linear relationship

Age	Bone Density
50	940
52	980
56	950
59	945
60	880
61	895
61	835
63	846
64	865
67	840
68	825
70	810
73	790
74	750
77	710
78	760
78	720
80	720



Strength of the Association

- The correlation coefficient provides a measure of the strength and direction of the linear relationship.

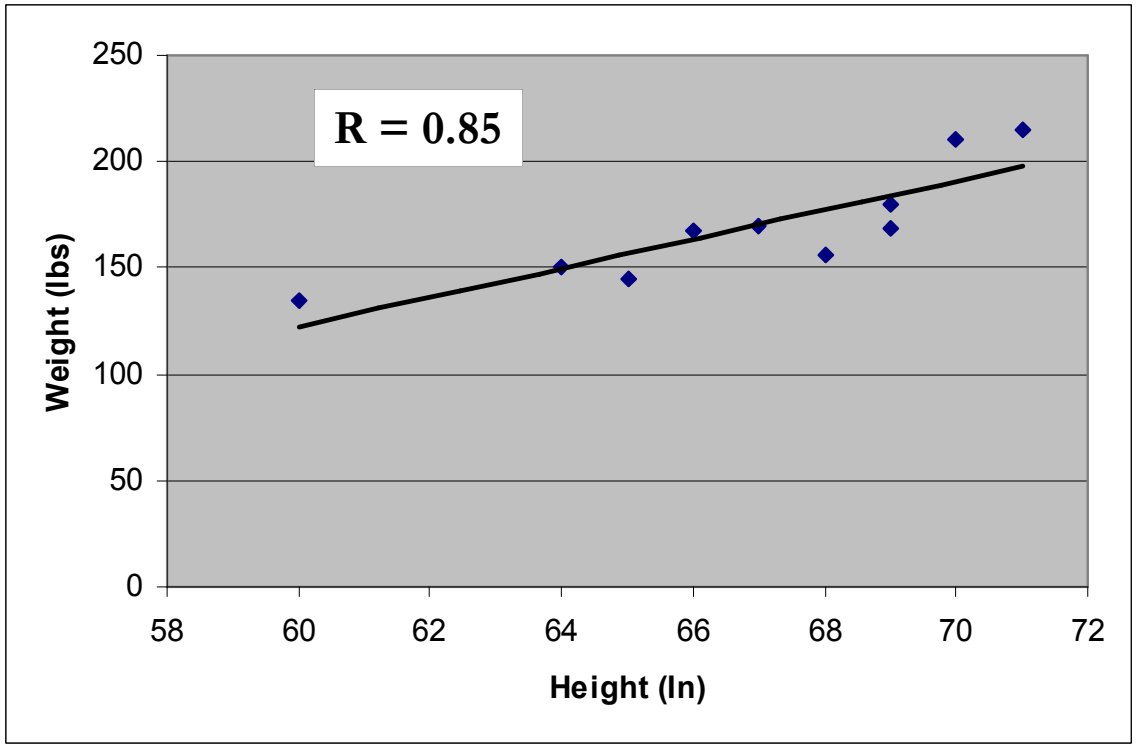
- The sign of the correlation coefficient indicates the direction
  - Positive coefficient = positive direction
  - Negative coefficient = negative direction
- Strength of association
  - Correlation coefficients closer to 1 or -1 indicates a stronger relationship
  - Correlation coefficients = 0 (or close to 0) indicate a non-existent or very weak linear relationship

Correlation Coefficient

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

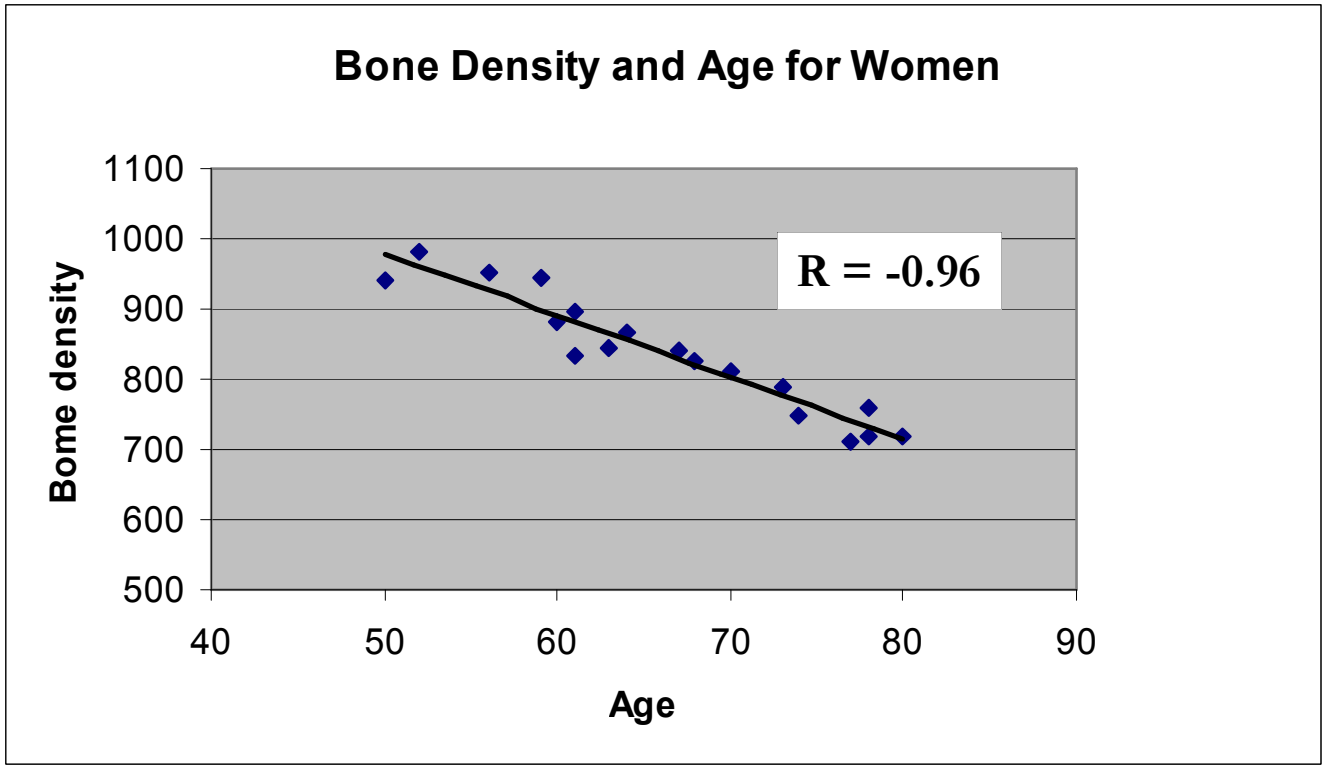
- The statistic r is called the Correlation Coefficient
- r is always between –1 and 1
- The closer r is to 1 or -1, the stronger the linear relationship

Strong positive relationship



The correlation coefficient for the (height, weight) data = 0.85 indicating a strong (close to 1.0), positive linear relationship between height and weight.

Strong negative relationship



The correlation coefficient for the (age, bone density) data = -0.96 indicating a strong (close to -1.0), negative linear relationship between age and bone density for women.

Correlation Coefficient Interpretation Guidelines*

These guidelines can be used to interpret the correlation coefficient

- 0 to 0.25 (or 0 to –0.25) – weak or no linear relationship
- 0.25 to 0.5 (or –0.25 to –0.5) – fair degree of linear relationship
- 0.50 to 0.75 (or –0.5 to –0.75) – moderately strong linear relationship
- > 0.75 (or -0.75) – very strong linear relationship
- 1 or –1 perfect linear relationship – deterministic relationship

* Colton (1974)

Calculating Correlation Coefficient in Excel

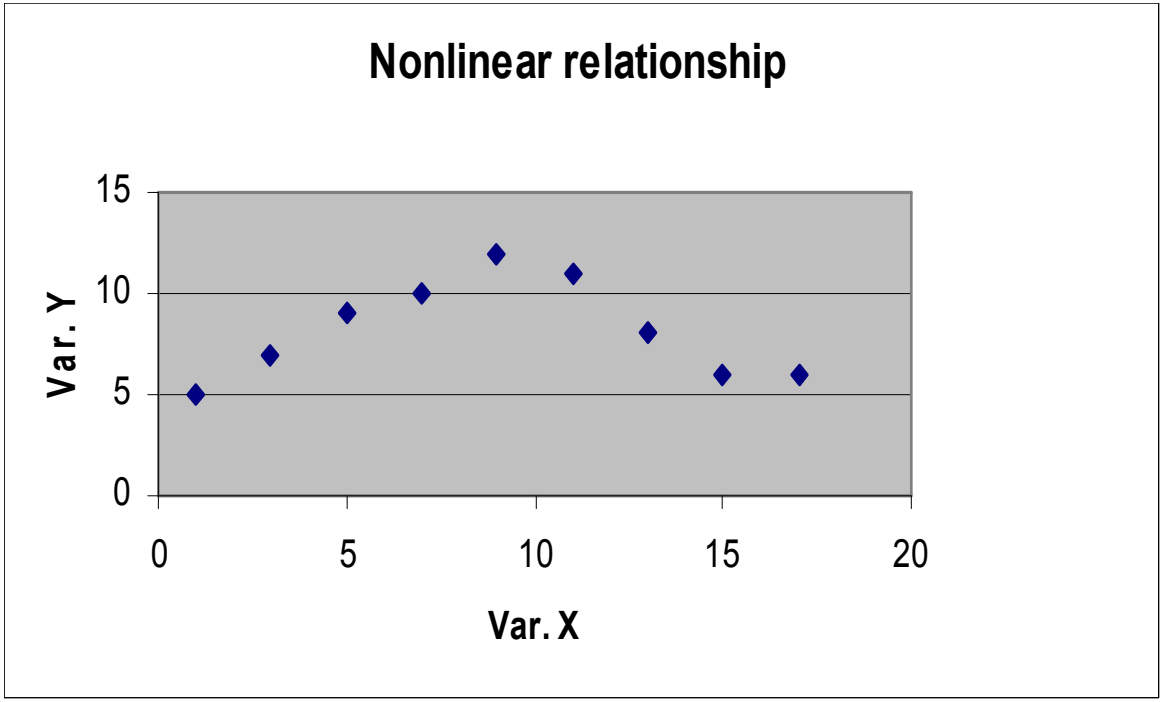
- If the original data are in B2:B11 and C2:C11
- Use the CORREL function and the data ranges

  =CORREL(B2:B11,C2:C11) to find the correlation coefficient : 0.8522
- Use the CORREL function for assignments and exams.

Interpretation of Correlation Coefficient = 0

- A correlation coefficient = 0 does not necessarily mean there is NO relationship between the variables.
- If r = 0 this means there is no **linear** relationship
- There may be a strong nonlinear relationship but r will not identify this.
- Plot the data first to determine if there is a nonlinear relationship

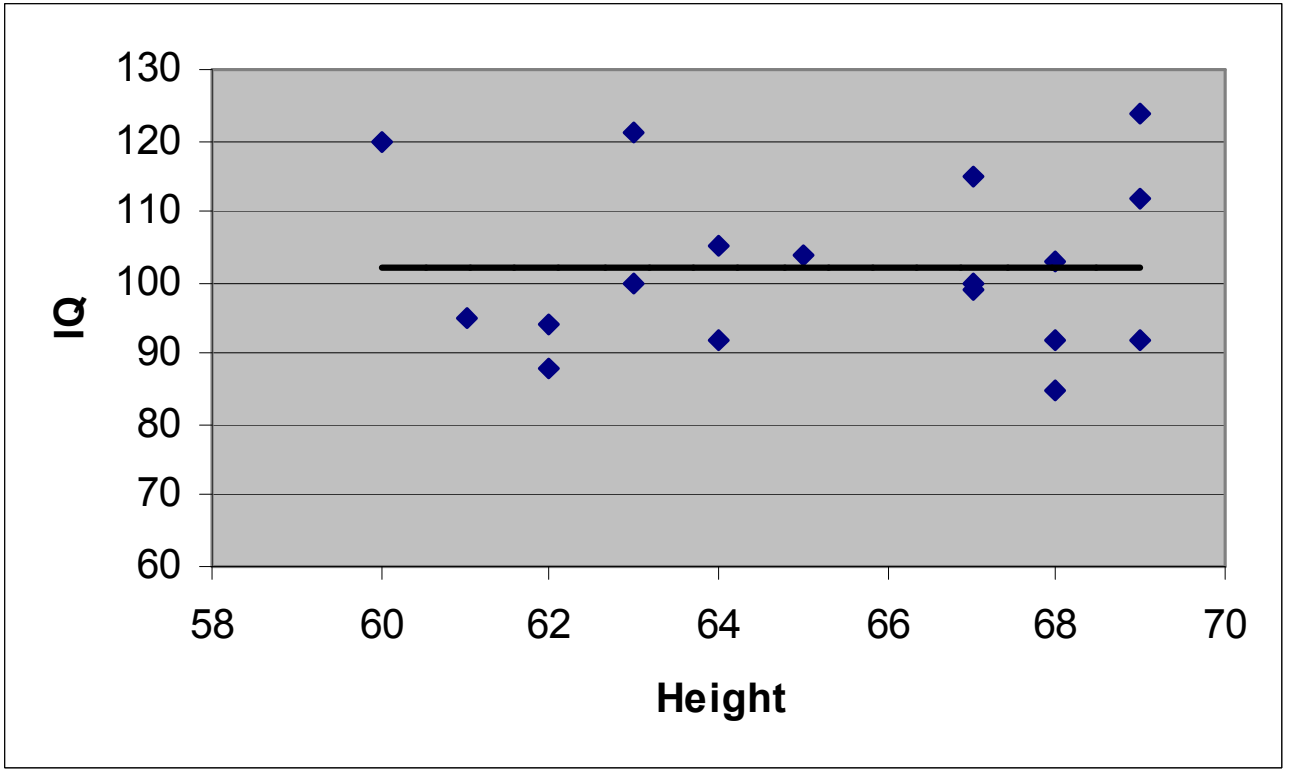
Nonlinear Relationship
r = 0



•This plot illustrates that the correlation coefficient only measures the strength of linear relationships.

•Here there is a nonlinear (curvilinear) relationship between X and Y but the correlation coefficient = 0.

Random scatter: no pattern
r = 0.002



When there is a random scatter of points, the correlation coefficient will be near 0 indicating no linear relationship between the variables

Notes about Correlation Coefficient

- The correlation coefficient is independent of the units used in measuring the variables

- The correlation coefficient is independent of which variable is designated the ‘X’ variable and which is designated the ‘Y’ variable

Correlation vs. Causation

- Strong correlation does not imply causation (i.e. that changes in one variable cause changes in the other variable).
- Example*: There is a strong positive correlation between the number of television sets /person and the average life expectancy for the world's nations. Does having more TV sets cause longer life expectancy?

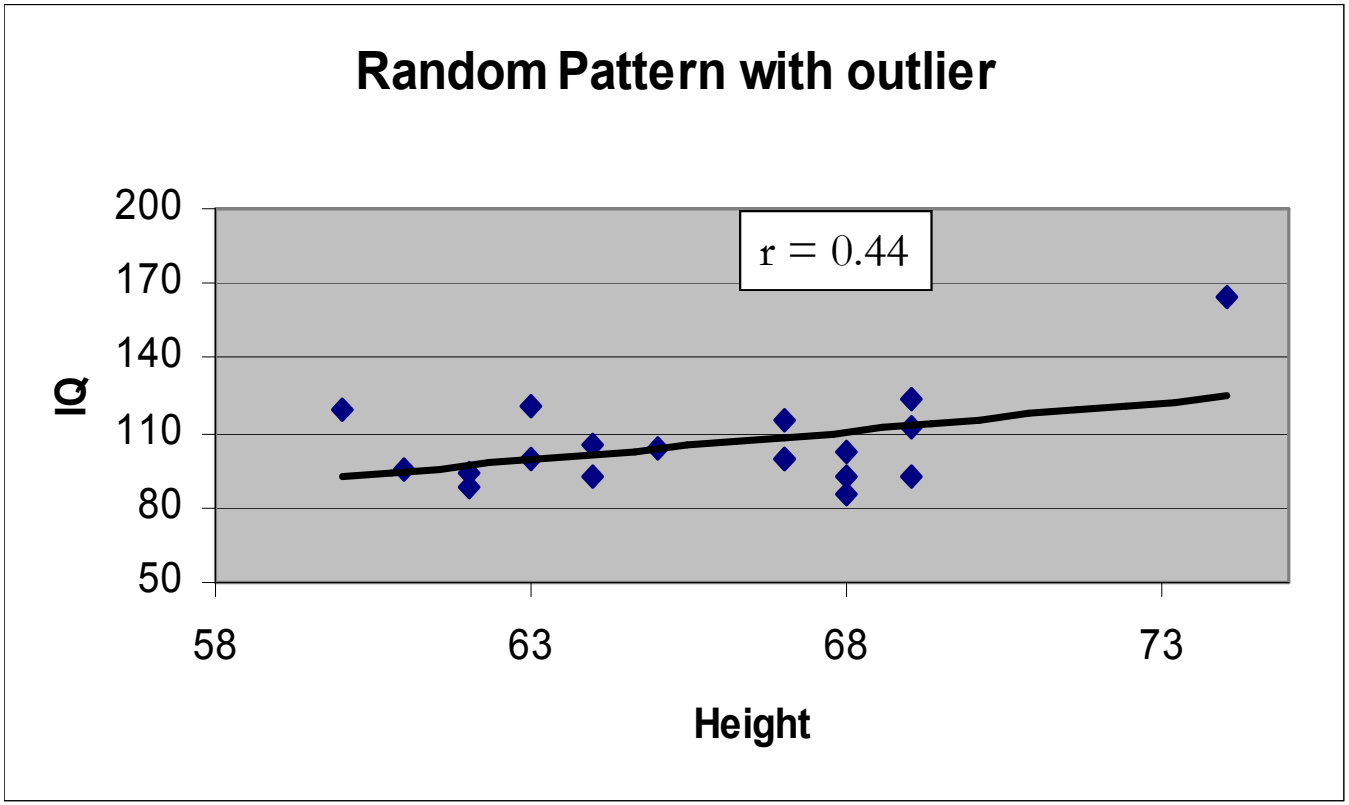
*Moore & McCabe

Coefficient of Determination

- The coefficient of Determination is the correlation coefficient squared (R^2)
- R^2 = the percent of variation in the Y variable that is explained by the linear association between the two variables
- From example of age / bone density:
  - Correlation coefficient for age and bone density = -0.96.
  - Coefficient of determination = $(-0.96)^2 = 0.92$
  - 92% of the variation in bone density is explained by the linear association between age and bone density

Effect of outlier on r

- One outlier has been added to the random pattern plot of height and IQ: a 74 inch tall person with IQ of 165
- Addition of this one point results in r changing from 0.002 to 0.44



Correlation Coefficient and Outliers

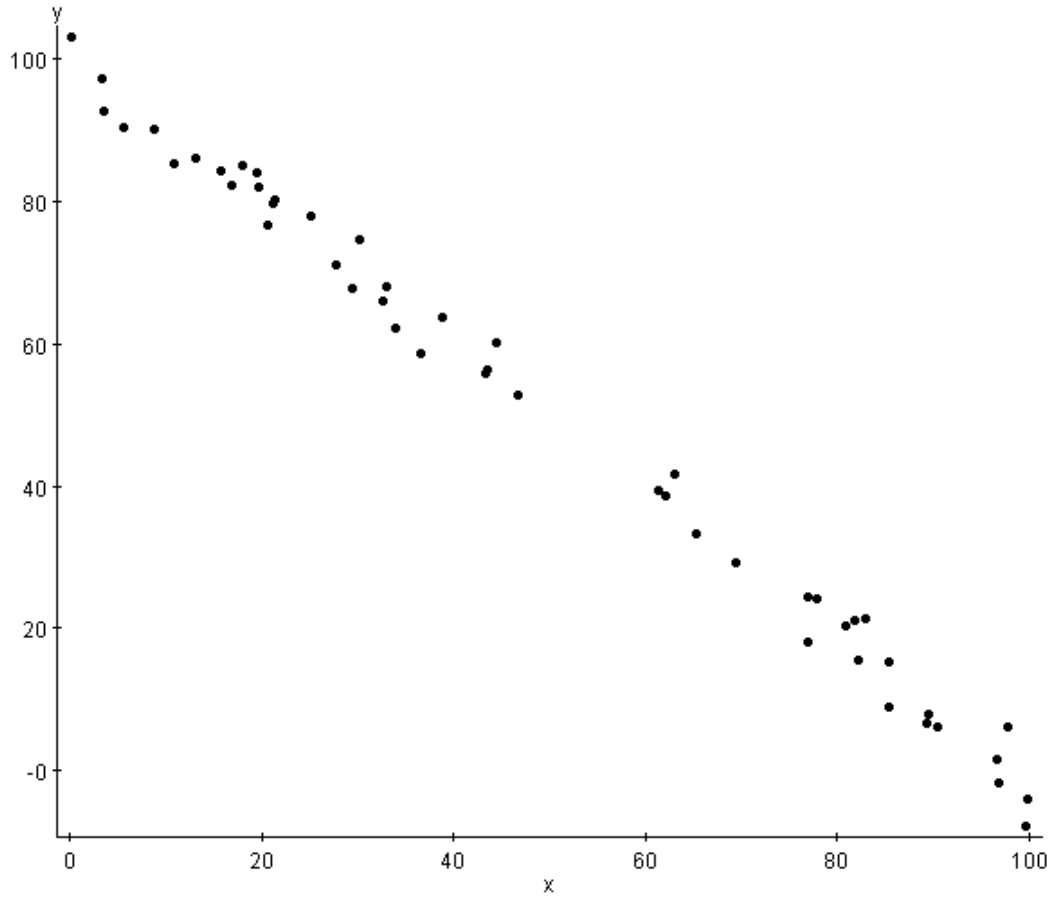
- The Pearson's correlation coefficient can be influenced by extreme values (outliers).
- Outliers are most easily identified from a scatter plot.
- If an outlier is suspected, calculate r with and without the outlier. If the results are very different, consider calculating the ***Spearman rank*** correlation coefficient instead.

Correlation measures the strength of any relationship between two continuous variables?

1. True
2. False

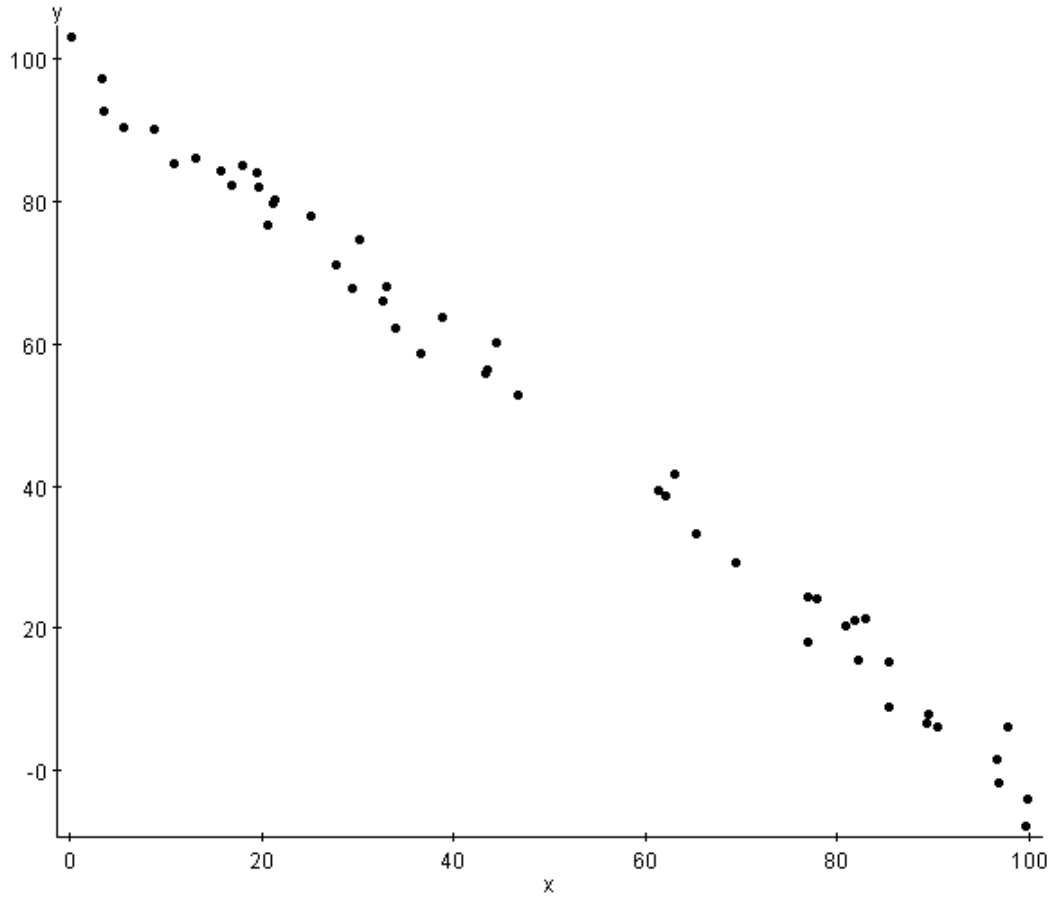
What is the direction of the association pictured in the scatter-plot below?

1. Positive
2. Negative
3. Cannot be determined.



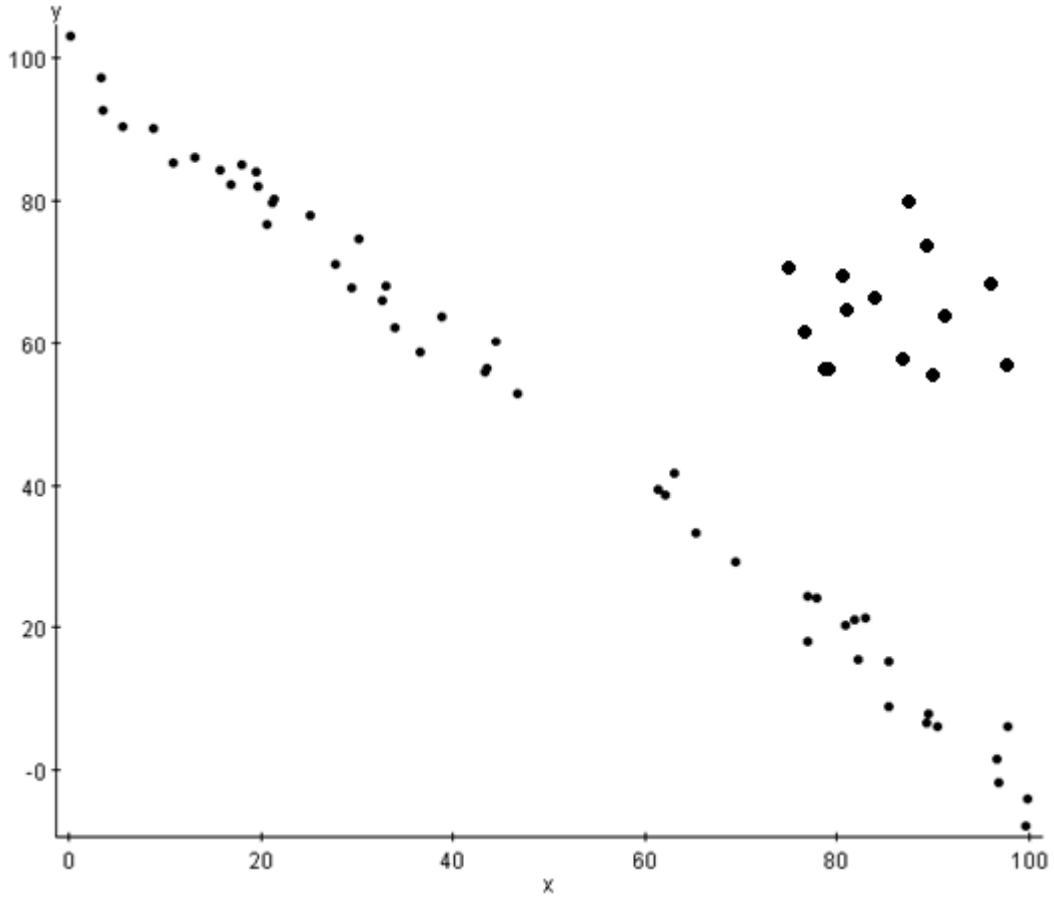
Which of the choices best describes the strength of the linear relationship pictured in the scatter-plot below?

1. Weak
2. Moderate
3. Strong



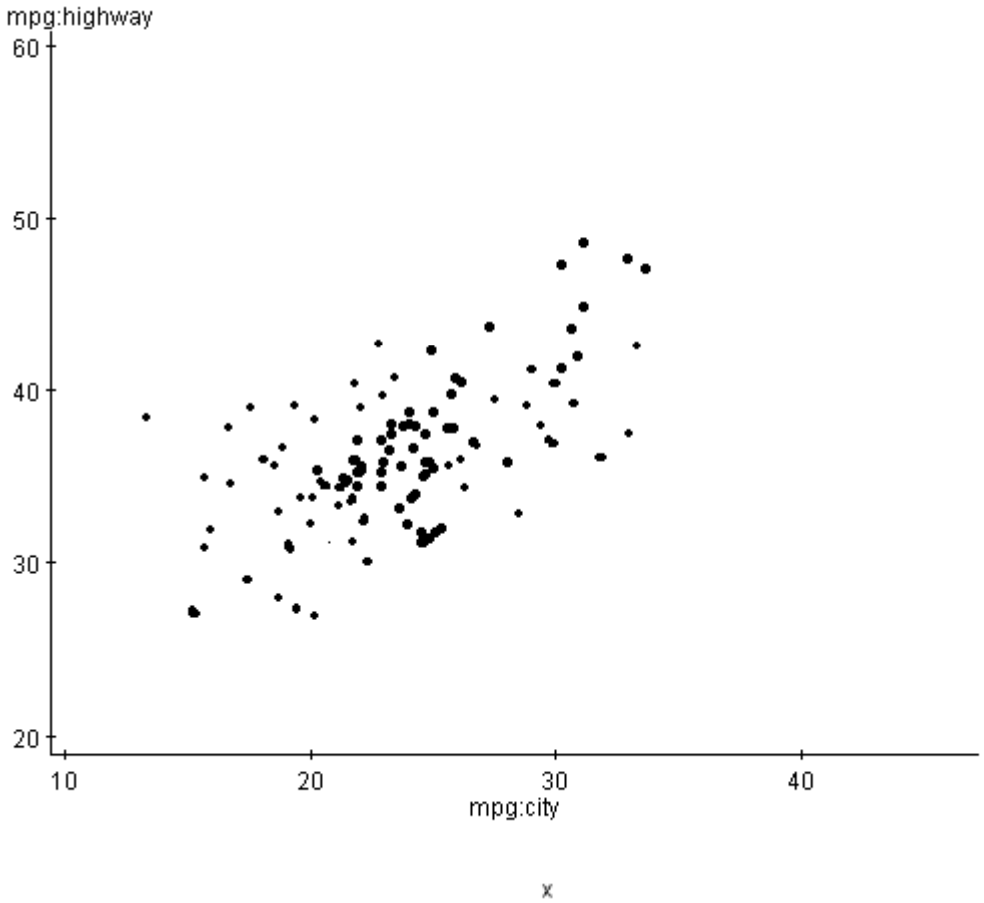
Suppose more data is added to the original plot, what happens to the strength of the linear relationship?

1. Weaker
2. Stronger
3. Cannot be determined.



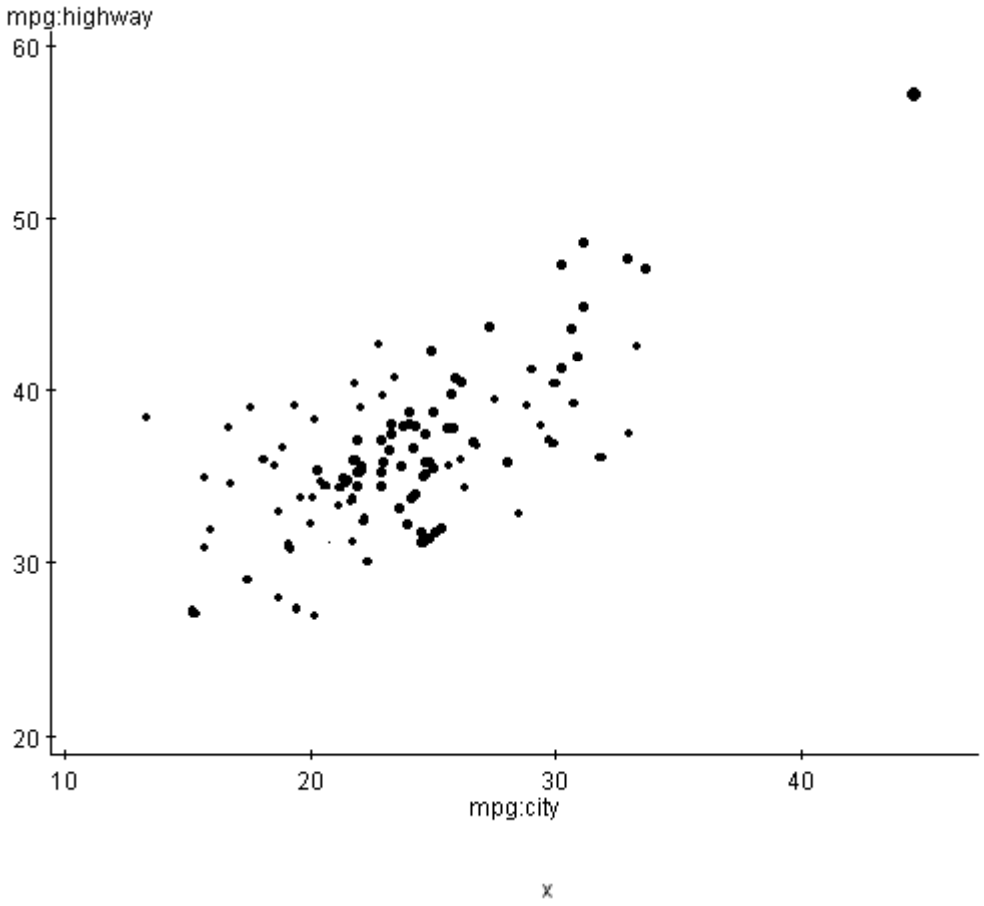
Which of the choices best describes the strength of the linear relationship pictured in the scatter-plot below?

1. Weak
2. Moderate
3. Strong



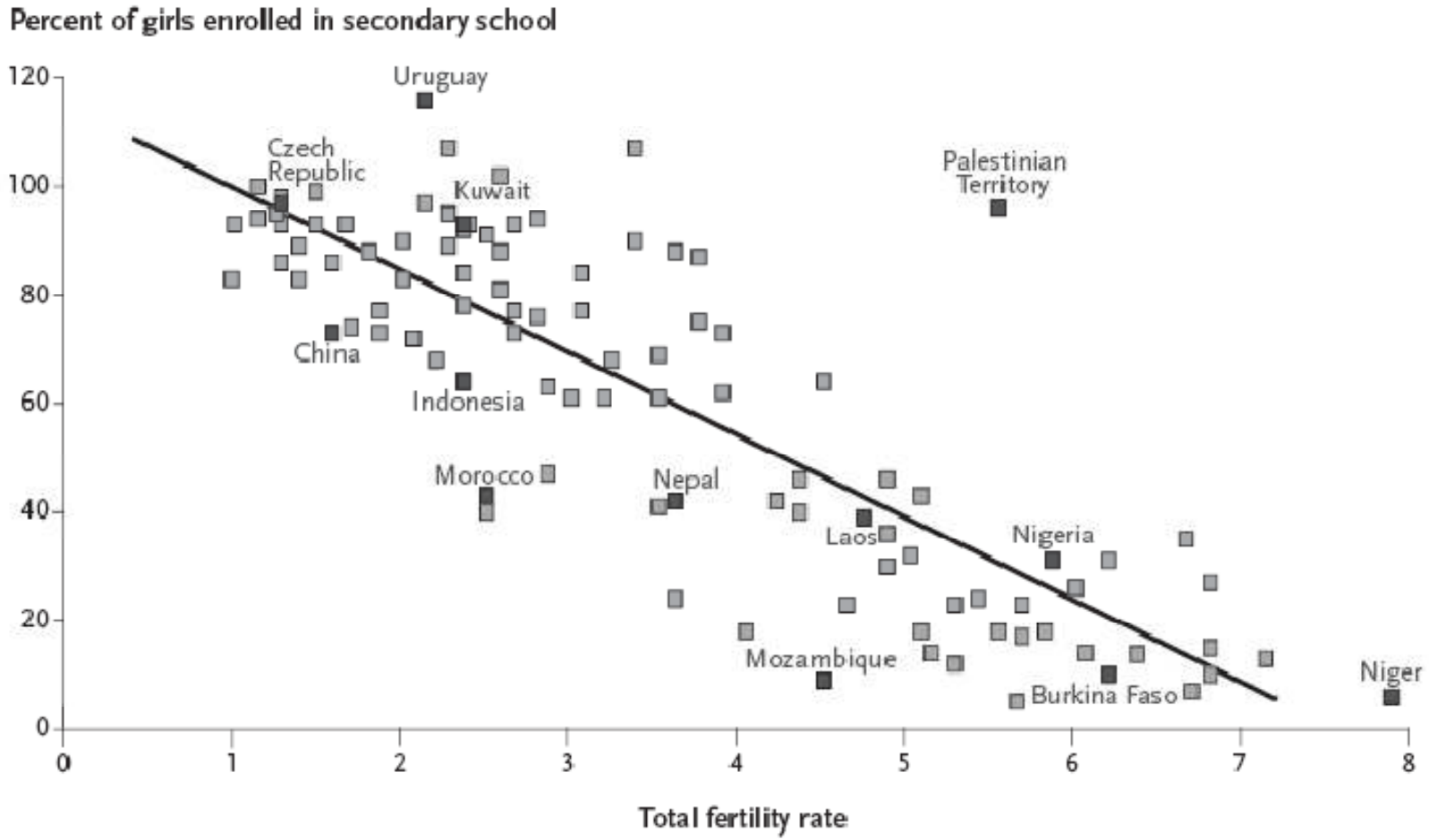
Suppose another piece data is added to the original plot, what happens to the strength of the linear relationship?

1. Weaker
2. Stronger
3. Cannot be determined



Scatter-plot girls education vs. fertility rate.

Source: ***Population Reference Bureau website:***
http://www.prb.org/Publications/Datasheets/2007/PopulationEconomicDevelopment2007.aspx



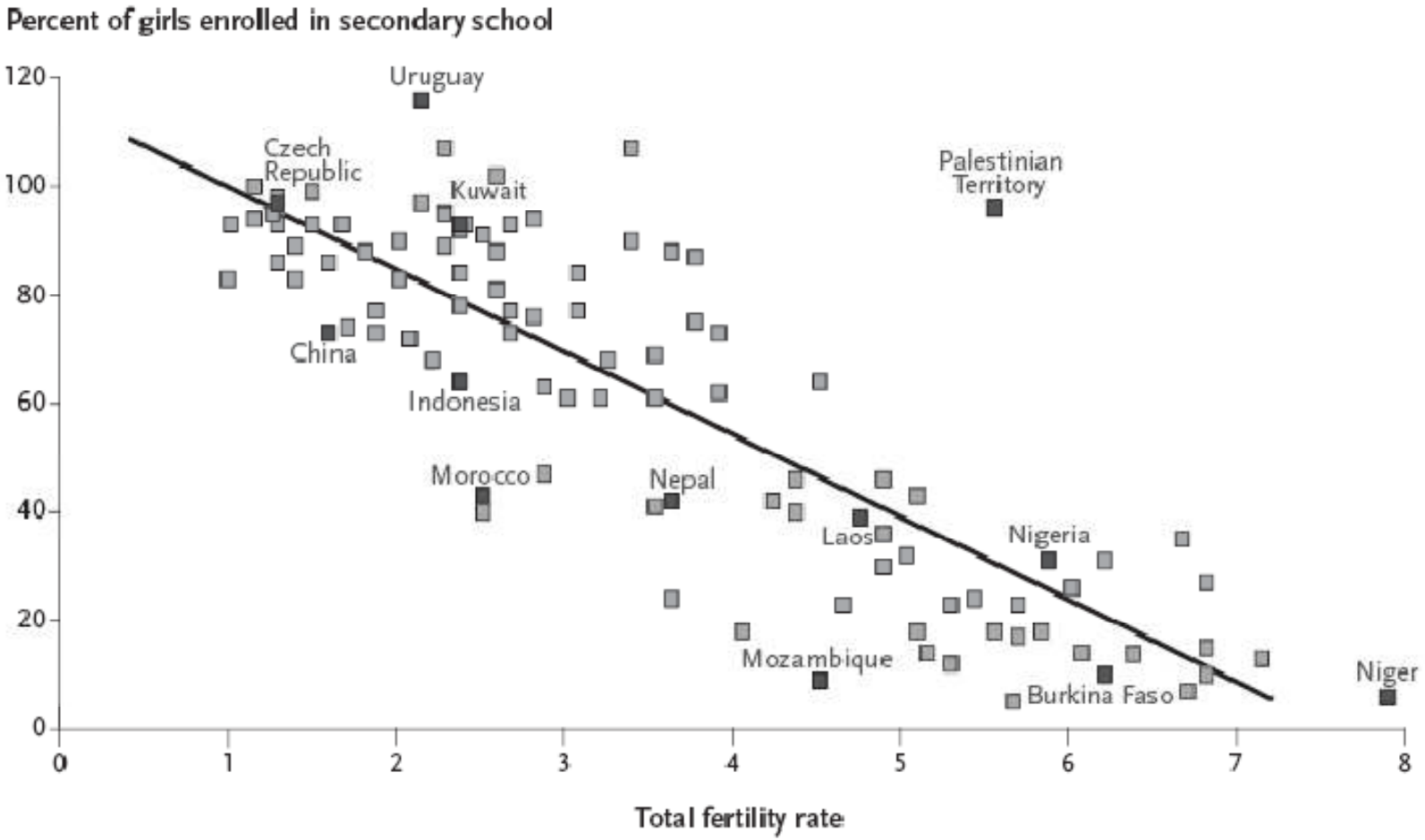
What is the general trend in this relationship?

What is the general trend in this relationship??

1. Positive
2. Negative

Scatter-plot girls education vs. fertility rate.

Source: ***Population Reference Bureau website:***
http://www.prb.org/Publications/Datasheets/2007/PopulationEconomicDevelopment2007.aspx

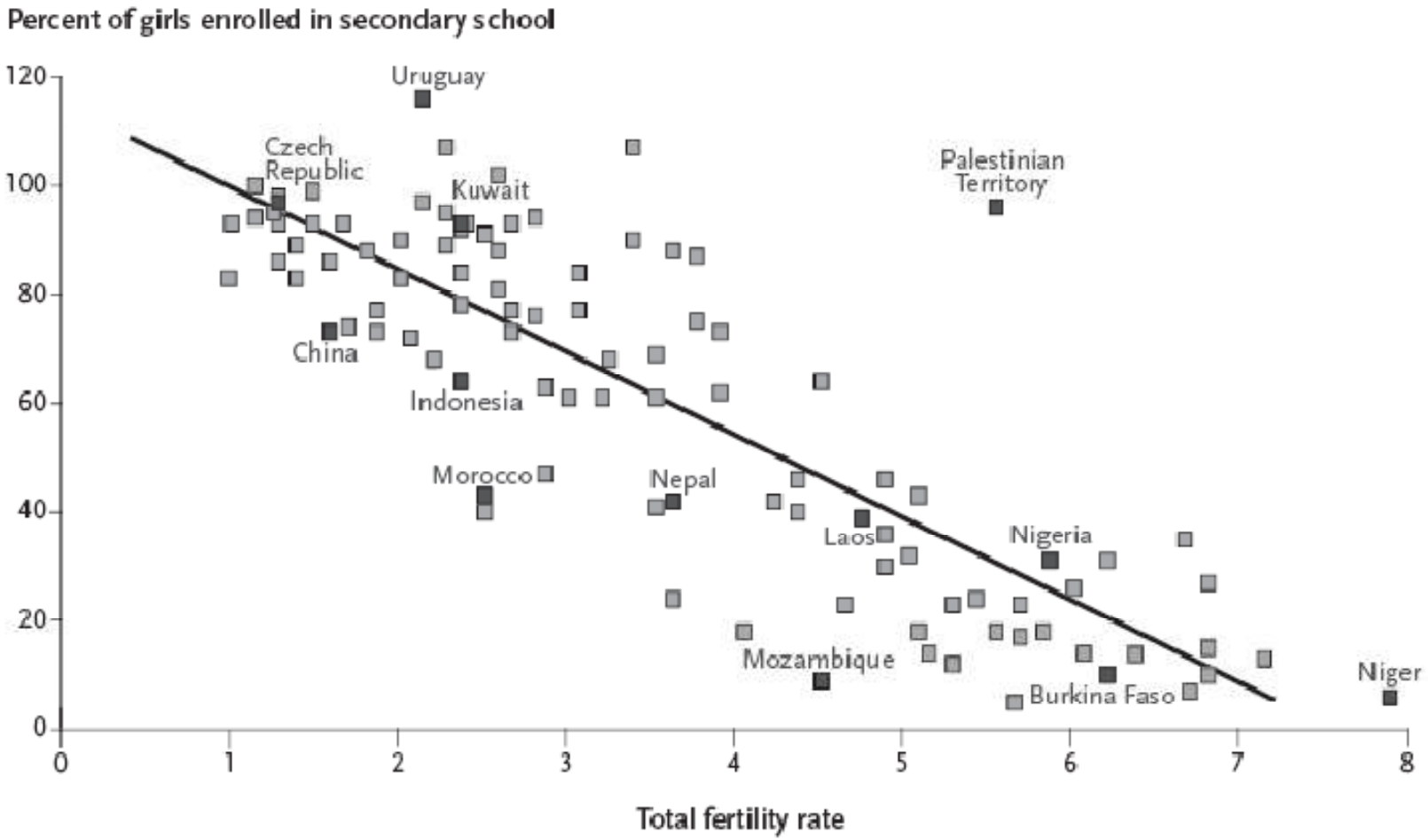


How strong is this linear relationship?

How strong is this linear relationship?

1. Weak
2. Moderate
3. Moderately Strong
4. Extremely Strong

Scatter-plot girls education vs. fertility rate.



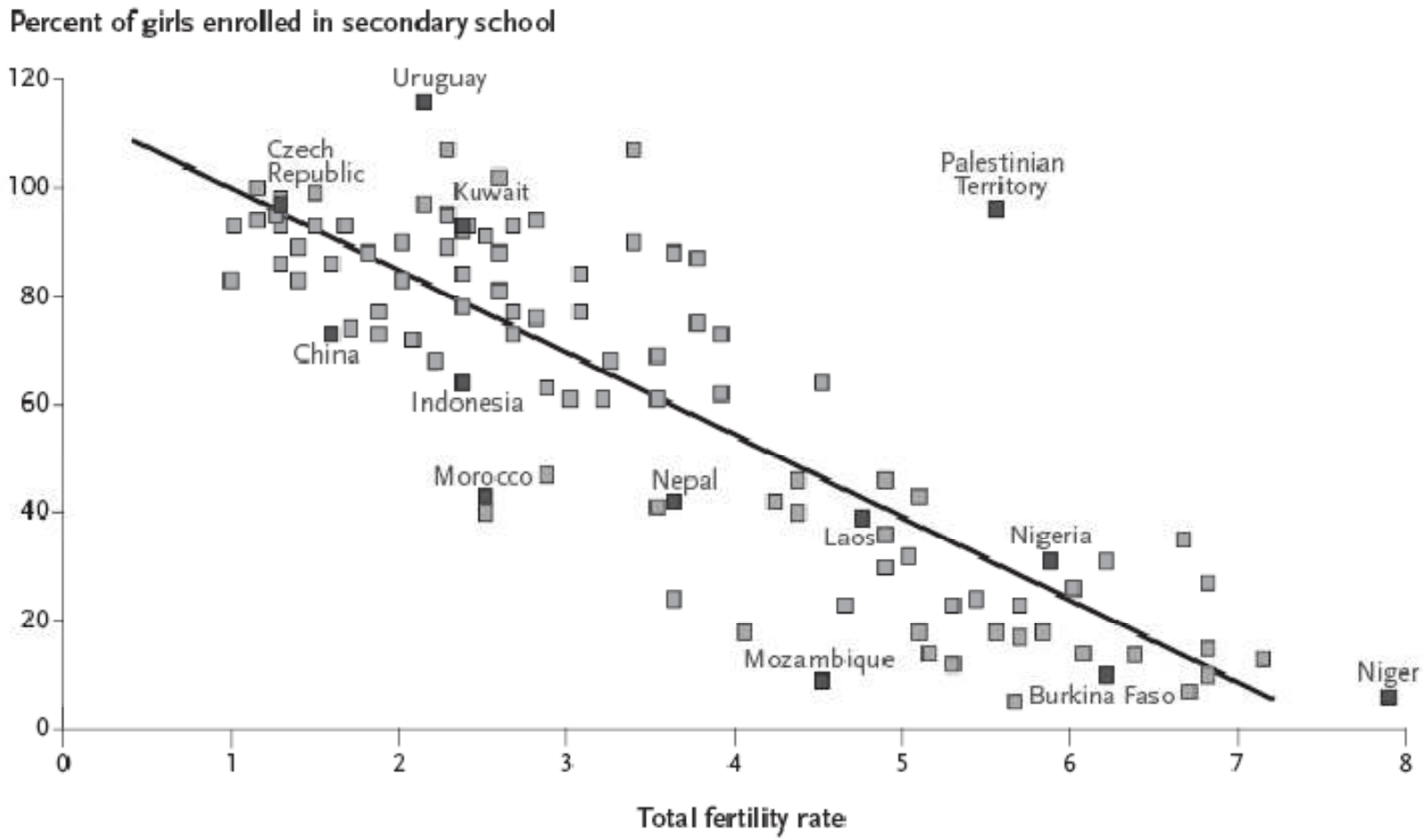
Which location appears to not follow the linear relationship?

Which location appears to not follow the linear relationship?

1. Palestinian Territory
2. Czech Republic
3. Burkina Faso
4. Laos

Scatter-plot girls education vs. fertility rate.

Source: ***Population Reference Bureau website:***
http://www.prb.org/Publications/Datasheets/2007/PopulationEconomicDevelopment2007.aspx



With what type of data can you create a scatter-plot?

With what type of data can you create a scatter-plot?

1. Two binary variables for each unit.
2. A continuous x and y for each unit.
3. An ordinal and a nominal variable for each unit.
4. Univariate Data

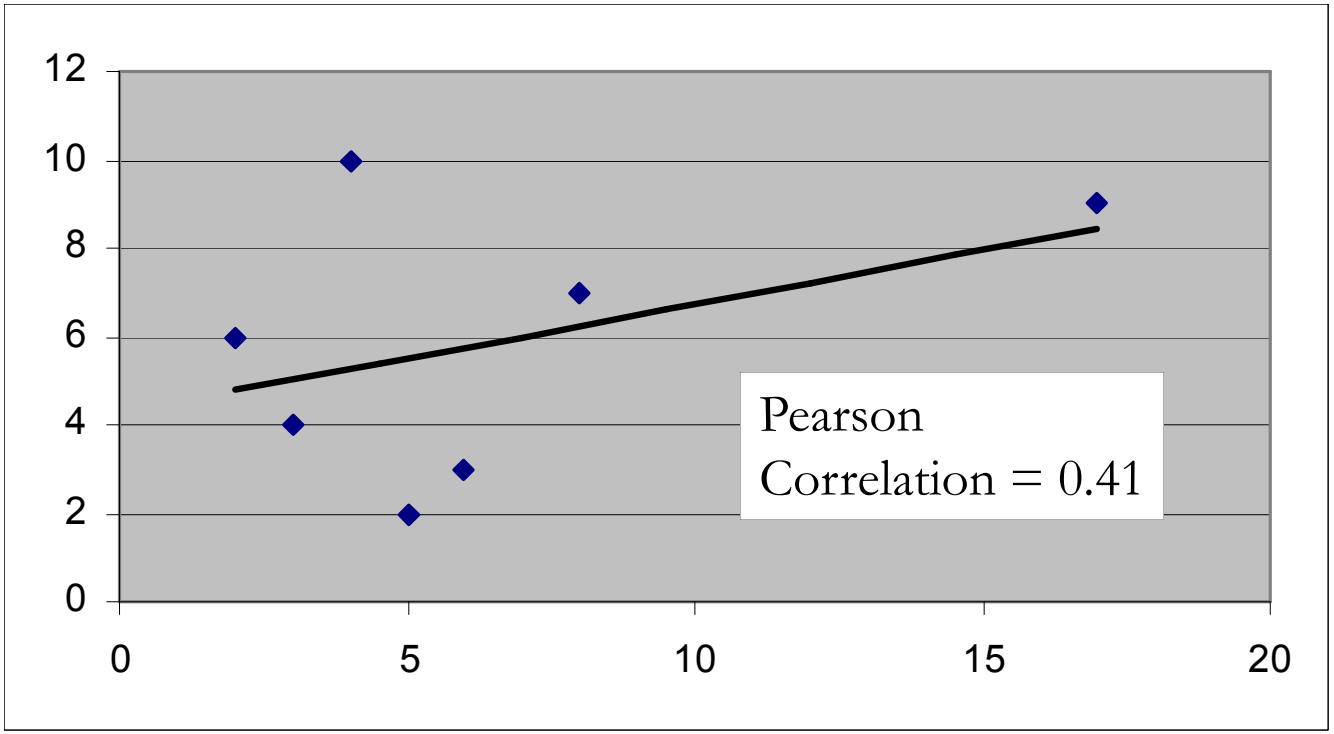
Spearman Rank Correlation (Spearman's Rho)

- Non-parametric: analysis methods that make no assumptions about the data distribution
- Spearman rank correlation is the appropriate correlation analysis to use when
  - Both variables are ordinal data
  - One variable is continuous and one is ordinal
  - Continuous data is skewed due to outlier(s)

Spearman rank correlation Procedure

- Order the data values for each variable from lowest to highest
- Assign ranks for each variable (1 up to n)
- Use the ranks instead of the original data values in the formula for the correlation coefficient

Skewed Data due to outlier



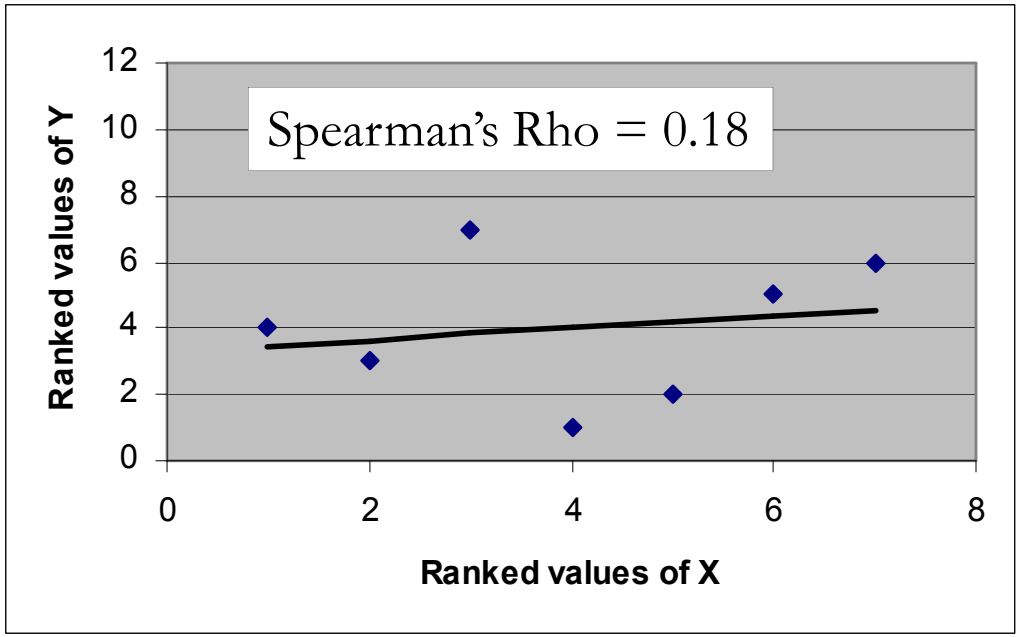
With the outlier point (17,9) r = 0.41. Without the outlier, r = -0.04.

Original values and Rank values

X	Y	X-rank	Y-rank
8	7	**6**	**5**
4	10	**3**	**7**
3	4	**2**	**3**
17	9	**7**	**6**
2	6	**1**	**4**
6	3	**5**	**2**
5	2	**4**	**1**

The smallest X value is given rank ‘1’, the largest is given rank ‘7’
Similarly, the smallest Y value is given rank ‘1’, the largest is given rank ‘7’
Other values are ranked according to their order.

Spearman's Rho: the Spearman rank correlation coefficient



Ranking the values of X and Y removes the influence of the outlier

Spearman's Rho is not as large as the Pearson's correlation coefficient for the original data (r = 0.44) and the outlier is not excluded from analysis.

Spearman rank correlation in Excel

- Excel does not have a function for Spearman rank correlation
- Use the RANK function to find the ordered rank for each value (separately for each variable)
- Use the CORREL function on the ranks.
- This will return the Spearman rank correlation coefficient (Spearman's Rho)

Reading and Assignments

- Reading: Chapter 3 pgs. 48 – 50
- Lesson 4 Part 1 Practice Exercises
- Lesson 4 Excel Module
- Begin Homework 2: Problems 1 and 2