Section 1.4 – Scatterplots and Correlation vs. Causation

MDM4U
David Chen

Part 1: Scatterplots Video

Let's start by watching a video on scatterplots:

http://www.learner.org/courses/againstallodds/unitpages/unit10.html

Answer the following questions while watching the movie:

1. What does a scatterplot show about the relationship between the number of powerboats registered in Florida and the number of manatees killed by powerboats?

There is a positive association between manatees killed by powerboats and the number of powerboat registrations. In other words, as the number of powerboat registrations increases, the number of manatees killed also tends to increase.

2. Why is the number of boats plotted on the horizontal axis of this scatterplot?

The number of powerboat registrations is the explanatory variable.

3. What trend would you expect to see in a scatterplot of two variables that have a negative association?

As one variable increases, the other tends to decrease. For example, in factoring quadratics, the time it takes for you to factor a particular type of quadratic decreases with the number of times you have practiced factoring. (The more you practice, the faster you get)

Note: *a scatterplot of manatee deaths and the number of powerboat registrations shows a positive correlation between the two variables. However, the fact that there is a relationship between two variables is not sufficient evidence to prove cause-and-effect linkage. A well-designed randomized experiment in which the researcher imposes some treatment on its subjects to see how they respond is THE ONLY WAY to give good evidence for cause and effect as you will learn in next unit.*

Part 2: Scatterplot Basics

Remember that quantitative variables are measured using numerical values. When you have two quantitative variables (bivariate data), you can use a scatterplot to examine the correlation (association) between the two variables.

In many cases, changes in a variable *x* are thought to "e**x**plain" changes in a second variable *y*. In such examples, *x* is called the e**x**planatory (or independent) variable and *y* is called the response (or dependent) variable.

A scatterplot is a plot of observations of quantitative variables x and y as points in the plane. The explanatory variable, if any, is always plotted on the horizontal scale (x-axis) of the scatterplot.

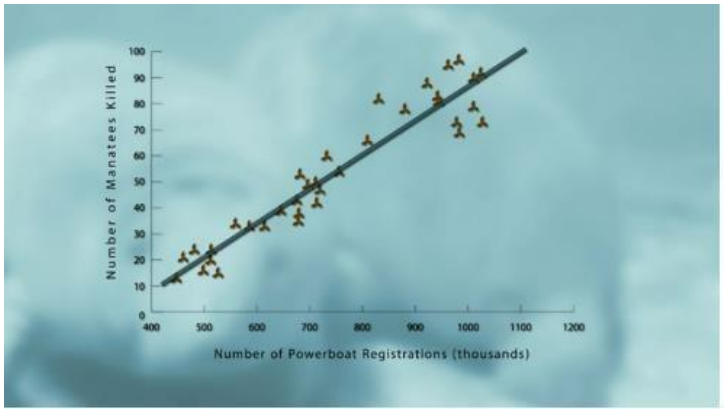
In the video on manatees,

The explanatory variable was:

# of powerboats sold

The response variable was:

# of manatees killed

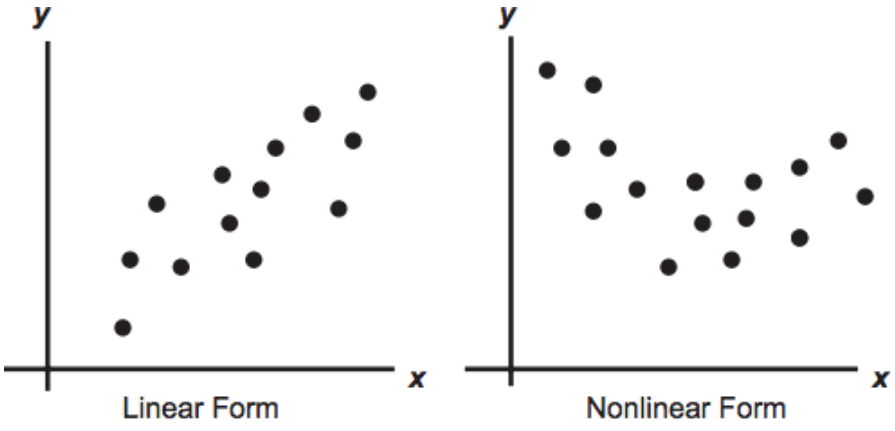


Part 3: Correlation

When analyzing a scatterplot of bivariate data, we look for:

- the overall pattern (linear, curved, random scatter)
- direction (positive, negative)
- strength of the relationship (strong, moderate, weak, no correlation)

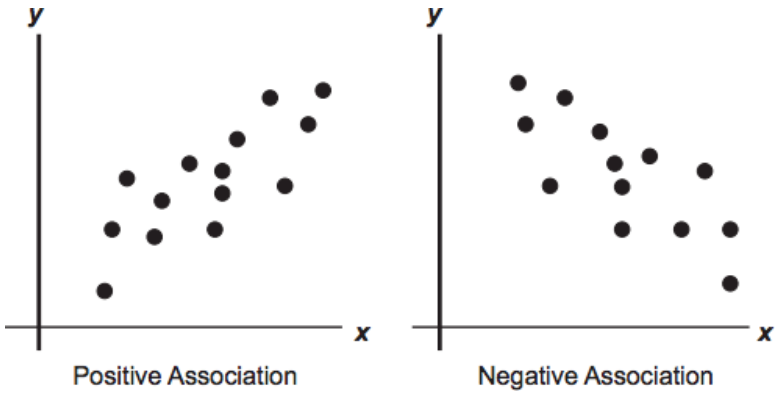
Pattern: A scatterplot has linear form when the dots appear to be randomly scattered on either side of a straight line. However, sometimes the data form a curved pattern. In that case, we say the scatterplot has non-linear form.



Direction:

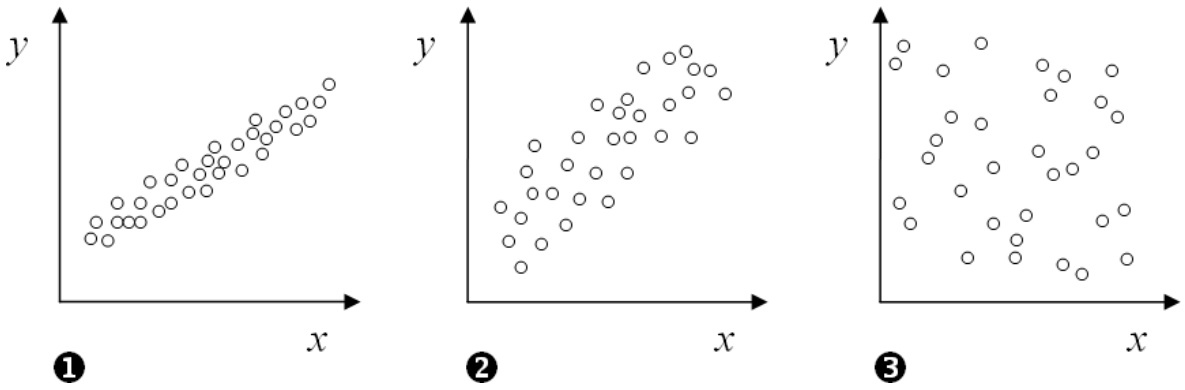
Two variables are positively associated (correlated) when above-average values of one tend to accompany above-average values of the other and below-average values of one tend to accompany below-average values of the other. In a scatterplot a positive association would appear as a pattern of dots in the lower left to the upper right.

Two variables are negatively associated (correlated) when above-average values of one accompany below-average values of the other, and vice versa. In a scatterplot a negative association would appear as a pattern of dots in the upper left to the lower right.



Strength:

Correlations can also be strong or weak depending on how close together or spread out the points on the graph are. If there seems to be no trend in the data, we say that there is no correlation.



Graph 1 shows a stronger correlation than graph 2 because the points are more closely clustered together.

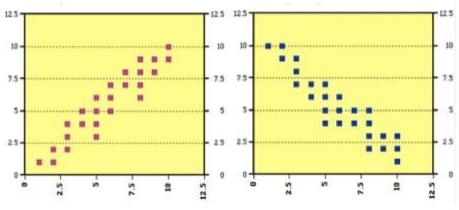
Graph 3 shows no correlation.

Time to check your understanding of what we have covered so far today.

Example 1: Circle the independent (explanatory) variable in each pair of variables

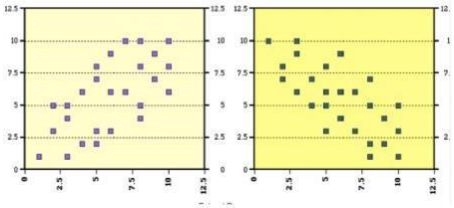
Height	**vs.**	Stride Length
Exam Score	**vs.**	Study Time
Smoking	**vs.**	Cancer Rates
Absences	**vs.**	Final Grade
Annual Income	**vs.**	Age

Example 2: Identify the type of correlation for each scatterplot. Identify the pattern, direction, and strength.



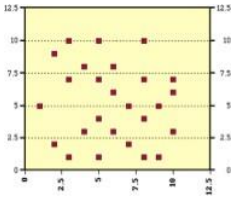
Strong Positive Linear

Strong Negative Linear



Weak Positive Linear

Weak Negative Linear

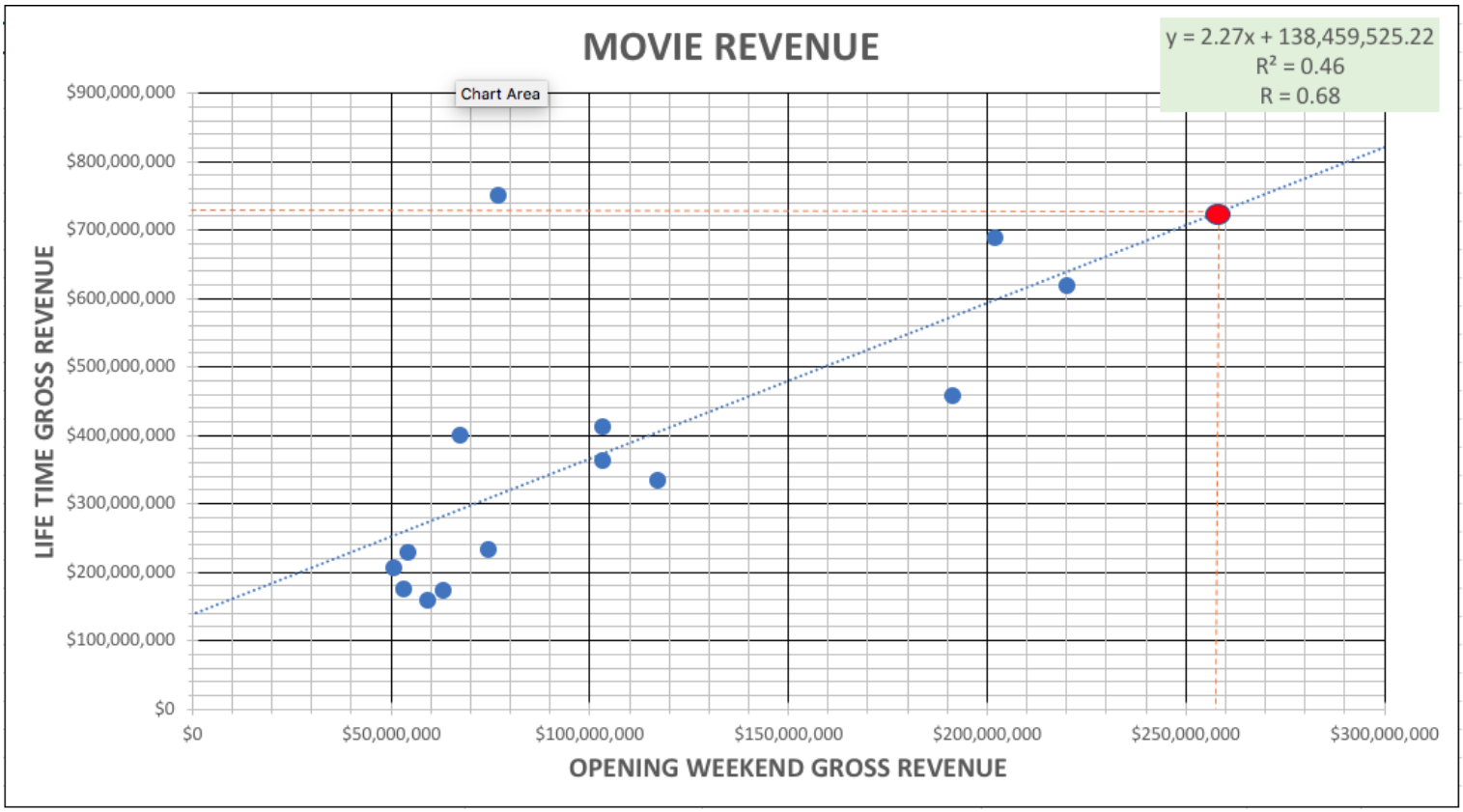


No correlation

Example 3: Avengers: Infinity War had a successful opening weekend by earning more than 250 million dollars. In opening weekends, a movie's opening gross income is a way of predicting the movies eventual success. Can you predict a movie's total gross income from its opening weekend gross income?

Movie	Opening	Total Gross	Release Date
The Martian	$54,308,575	$228,433,663	10/2/15
Star Trek Beyond	$59,253,211	$158,848,340	7/22/16
LEGO Batman	$53,003,468	$175,750,384	2/10/17
Spider Man: Homecoming	$117,027,503	$334,201,140	7/7/17
Pirates of the Caribbean: Dead Men Tell no Tales	$62,983,253	$172,558,876	5/26/17
Fantastic Beasts and Where to Find Them	$74,403,387	$234,037,575	11/18/16
Coco	$50,802,605	$207,389,121	2/19/18
The Jungle Book	$103,261,464	$364,001,123	4/15/16
Frozen	$67,391,326	$400,738,009	11/27/13
Avengers: Age of Ultron	$191,271,109	$459,005,868	5/1/15
Avatar	$77,025,481	$749,766,139	12/18/09
Star Wars: The Last Jedi	$220,009,584	$618,199,339	12/15/17
Wonder Woman	$103,251,471	$412,563,408	6/2/17
Black Panther	$202,003,951	$688,796,094	2/16/18
Avengers: Infinity War	$257,698,183		4/27/18

a) Make a scatter plot of the data.



b) Describe the trend in the data

There appears to be a moderate, positive, linear correlation between opening weekend revenue and total gross revenue. The more a movie makes in the opening weekend, the more money it will gross in total.

c) Use a trend line to make a prediction for the total gross revenue of Avengers: Infinity War.

Using the trend line on the graph, the predicted total gross revenue for Avengers is about $725 000 000

Note: Using the equation of the trend line we can more accurately make an estimate.

$$predicted\ total\ revenue = 2.27(opening\ weekend\ revenue) + 138\,459\,525.22$$
$$predicted\ total\ revenue = 2.27(257\,698\,183) + 138\,459\,525.22$$
$$predicted\ total\ revenue = \$723\,434\,400.60$$

The actual gross revenue so far for Avengers: Infinity War is **$678 630 680**

https://phet.colorado.edu/sims/html/least-squares-regression/latest/least-squares-regression_en.html