ECON 256: Poverty, Growth & Inequality

Jack Rossbach

Correlation vs Causation

Oft repeated saying: “Correlation does not imply Causation”

- Just because two things move in similar directions, does not mean one is causing the other
- What are the dangers of inferring causation when there is none?
- How can we establish that a causal relationship exists?

What is Correlation?

Correlation means two variables (or two data series) move together

- For this class, we'll focus on linear correlation

Suppose we have two variables X and Y

- If X and Y are **positively correlated**, then X tends to be **higher** when Y is higher
- If X and Y are **negatively correlated**, then X tends to be **lower** when Y is higher

What is Correlation?

Correlation means two variables (or two data series) move together

- For this class, we'll focus on linear correlation

Suppose we have two variables X and Y

- If X and Y are **positively correlated**, then X tends to be **higher** when Y is higher

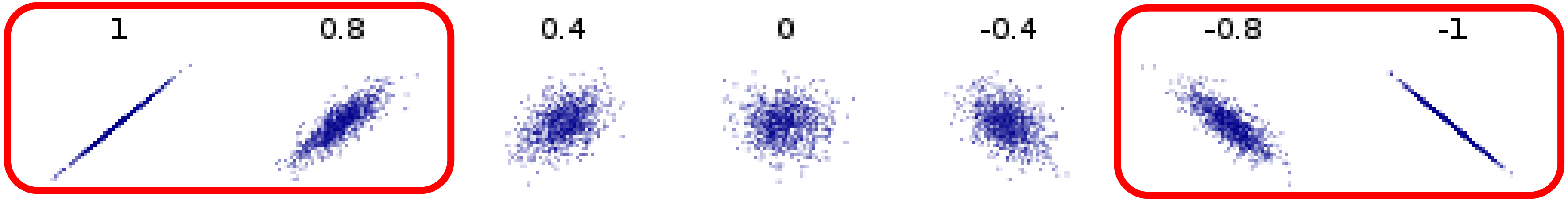
- If X and Y are **negatively correlated**, then X tends to be **lower** when Y is higher

Correlation coefficient will be between -1 and 1

- Closer to end points (-1 or 1) means stronger linear relationship

- Closer to zero means weaker linear relationship.

Examples of Linear Correlation Coefficients

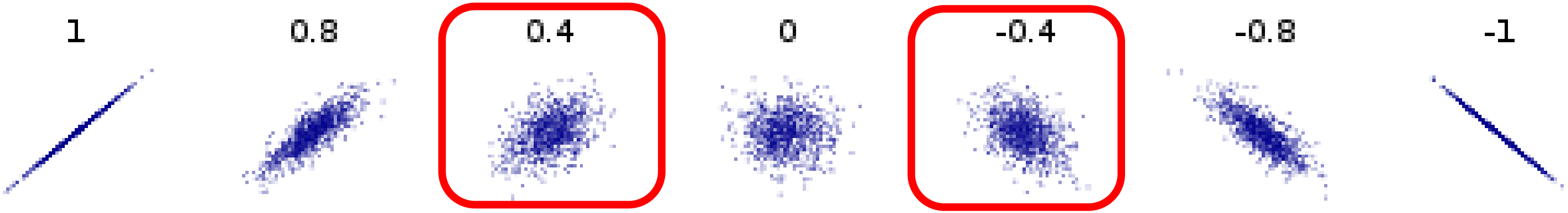


Strong Linear Relationship
(Positive)

Strong Linear Relationship
(Negative)

Source: Wikipedia

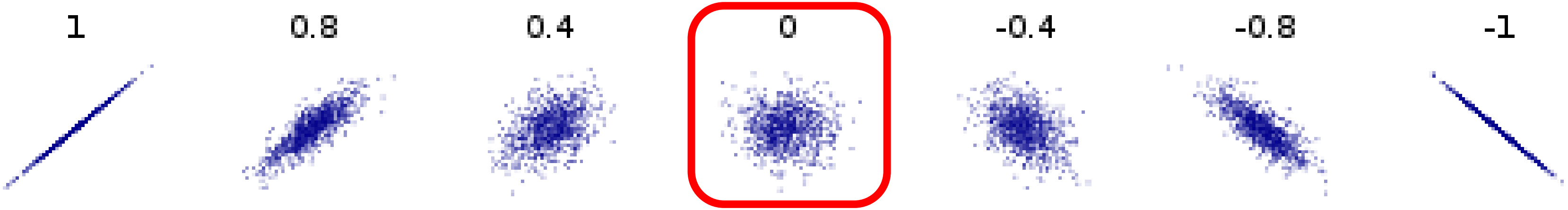
Examples of Linear Correlation Coefficients



Weak Linear Relationship
(Positive)

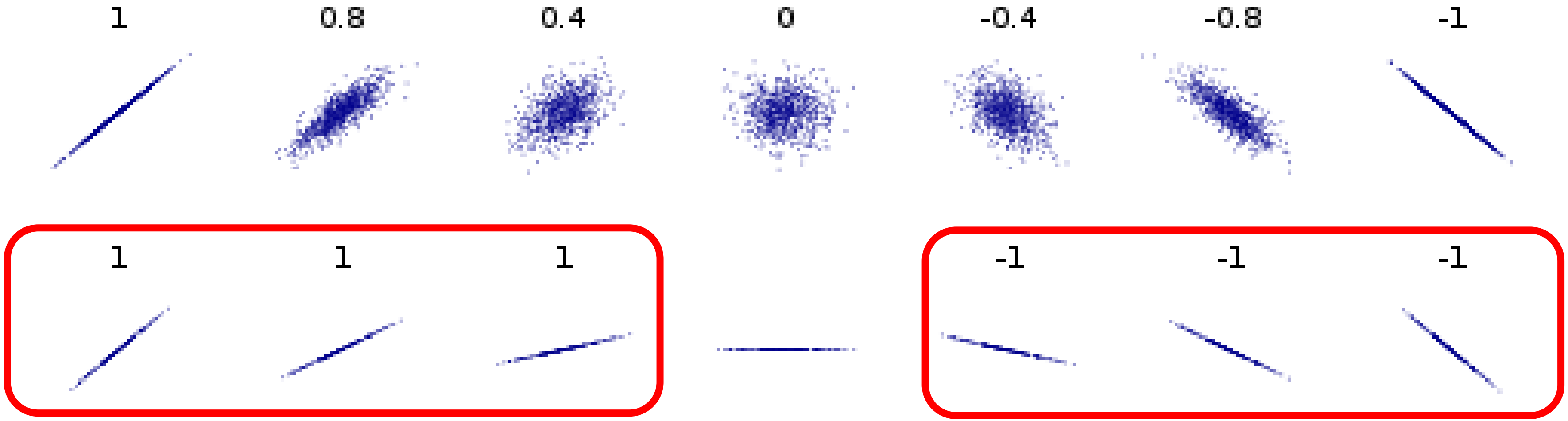
Weak Linear Relationship
(Negative)

Examples of Linear Correlation Coefficients



No Linear Relationship

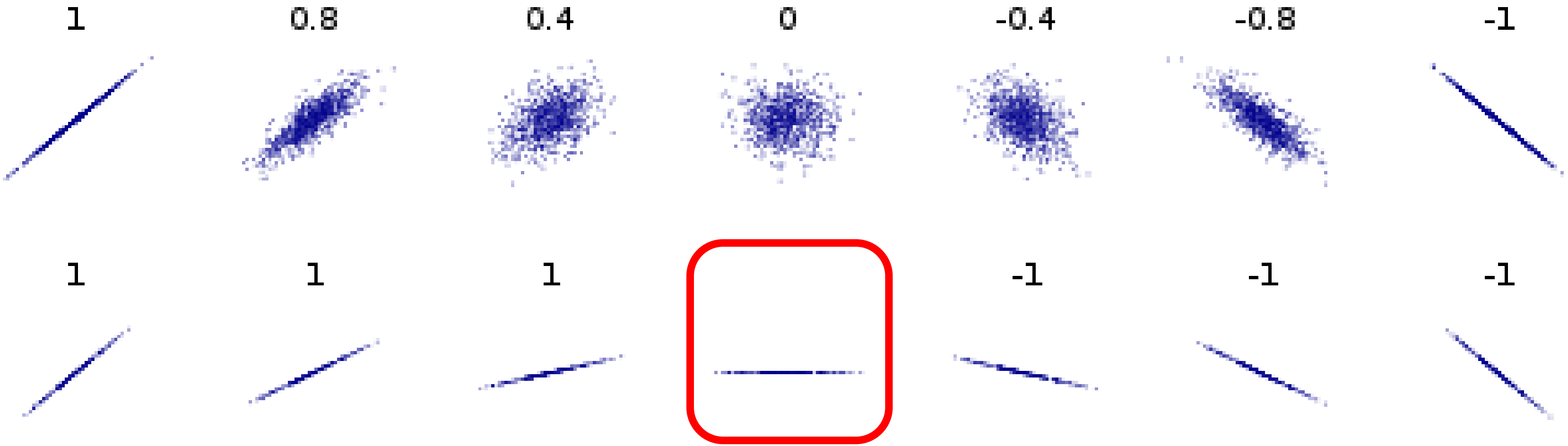
Examples of Linear Correlation Coefficients



Strong Linear Relationship (Positive)

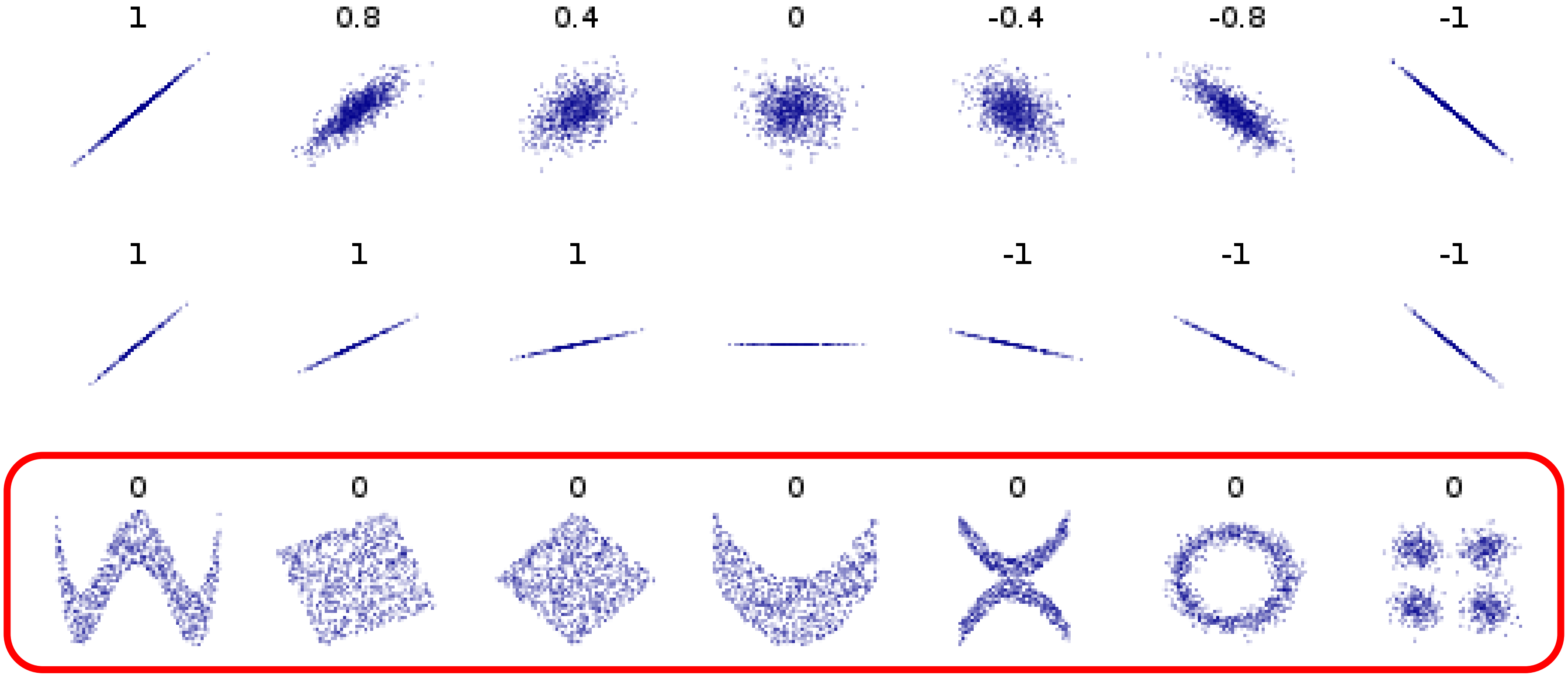
Strong Linear Relationship (Negative)

Examples of Linear Correlation Coefficients



Correlation Not Defined
(One of the Variables Doesn’t Change at All)

Examples of Linear Correlation Coefficients



No Linear Relationship
(Non-Linear Relationships aren’t Detected by Correlation)

Computing the Correlation Coefficient

Pearson's Correlation Coefficient, r, between two variables X and Y is computed as

$$r = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}[X]} \times \sqrt{\mathrm{Var}[Y]}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \times \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- Where x_i is the "i"th value of X, and y_i is the "i"th value of Y

- $\bar{x}$ is the mean (average value or expected value) of Y, and $\bar{y}$ is the mean of Y

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \, ; \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Note: There are several equivalent ways to rearrange the correlation formula, so you may see alternative expressions. Also, we are ignoring concerns of "sample" vs "population" statistics.

Computing the Correlation Coefficient

Variable	Observation 1	Observation 2	Observation 3
X	10	30	50
Y	20	0	10

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{3}(x_1 + x_2 + x_3) = \frac{1}{3}(10 + 30 + 50) = \frac{1}{3}(90) = 30$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{3}(y_1 + y_2 + y_3) = \frac{1}{3}(20 + 0 + 10) = \frac{1}{3}(30) = 10$$

Computing the Correlation Coefficient

Variable	Observation 1	Observation 2	Observation 3
X	10	30	50
Y	20	0	10

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = (x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + (x_3 - \bar{x})(y_3 - \bar{y})$$

$$= (10 - 30)(20 - 10) + (30 - 30)(0 - 10) + (50 - 30)(10 - 10)$$

$$= (-20)(10) + (0)(-10) + (20)(0)$$

$$= -200$$

Computing the Correlation Coefficient

Variable	Observation 1	Observation 2	Observation 3
X	10	30	50
Y	20	0	10

$$\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \sqrt{(x_1-\bar{x})^2+(x_2-\bar{x})^2+(x_3-\bar{x})^2} = \sqrt{(10-30)^2+(30-30)^2+(50-30)^2}$$

$$= \sqrt{(-20)^2+(0)^2+(20)^2} = \sqrt{400+0+400} = \sqrt{800}$$

$$\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2} = \sqrt{(y_1-10)^2+(y_2-10)^2+(y_3-10)^2} = \sqrt{(20-10)^2+(0-10)^2+(10-10)^2}$$

$$= \sqrt{(10)^2+(-10)^2+(0)^2} = \sqrt{100+100+0} = \sqrt{200}$$

Computing the Correlation Coefficient

Variable	Observation 1	Observation 2	Observation 3
X	10	30	50
Y	20	0	10

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \times \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$= \frac{-200}{\sqrt{800} \times \sqrt{200}} = \frac{-200}{\sqrt{800 \times 200}}$$

$$= \frac{-200}{\sqrt{1600}} = \frac{-200}{400} = \boxed{-0.5}$$

Correlation is −0.5

Computing the Correlation Coefficient

Variable	Observation 1	Observation 2	Observation 3
X	10	30	50
Y	20	0	10

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \times \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$= \frac{-200}{\sqrt{800} \times \sqrt{200}} = \frac{-200}{\sqrt{800 \times 200}}$$

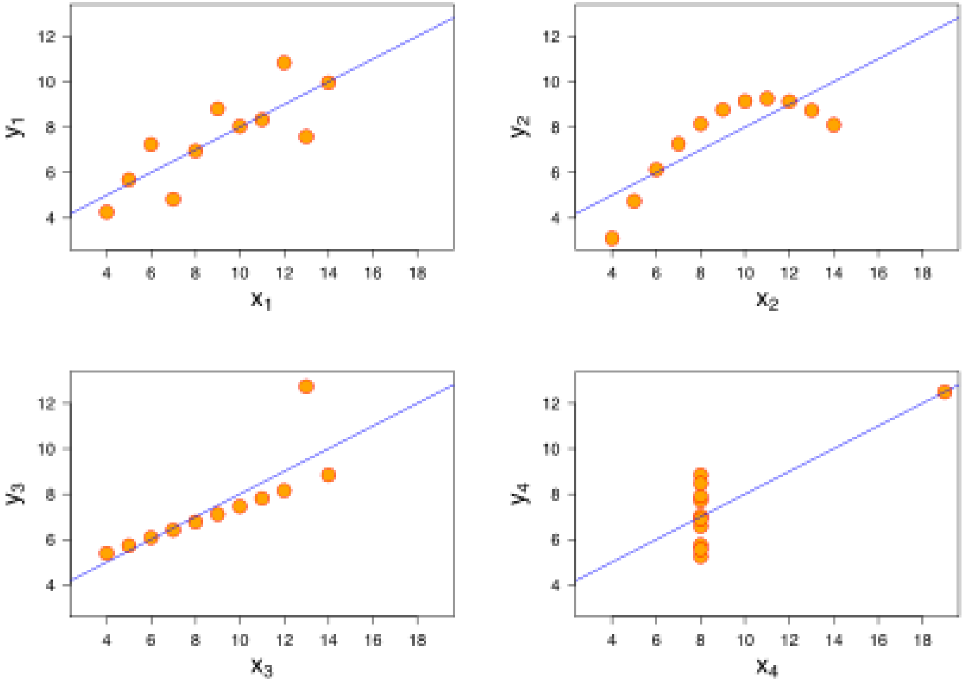
$$= \frac{-200}{\sqrt{1600}} = \frac{-200}{400} = \boxed{-0.5}$$

Correlation is –0.5

This means there is a negative correlation that is moderate strength.

Importance of Graphing the Data

Summary statistics such as the correlation coefficient are useful, however can be misleading



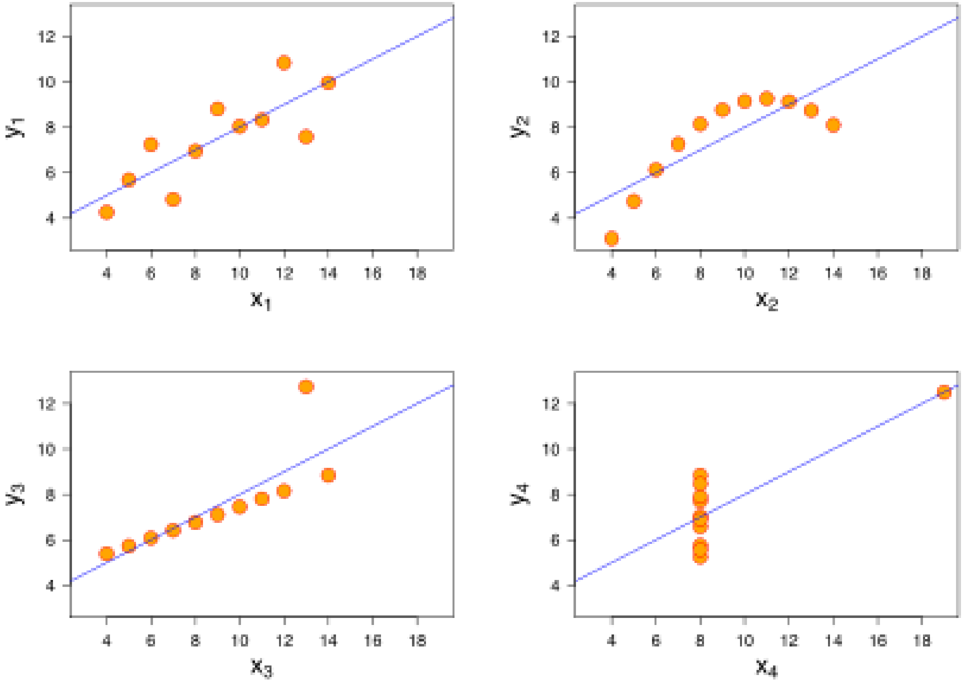
All four graphs have the same correlation coefficient! (r = 0.816)

Anscombe's quartet

Importance of Graphing the Data

Summary statistics such as the correlation coefficient are useful, however can be misleading

Graph 1 is a linear relationship, and ideal for summarizing with correlation coefficient



Graph 2 is a not a linear relationship (it's a quadratic relationship)

Graph 3 is linear, but with an outlier

In Graph 4, without the outlier, X would only take a single value

All four graphs have the same correlation coefficient! (r = 0.816)

Anscombe's quartet

What do We Use Correlation For?

Correlation hints that two variables might be related somehow

- Doesn’t tell us the nature of the relationship

Some Possible Relationships when X and Y are correlated

1. X causes Y, but Y does not cause X
2. Y causes X, but X does not cause Y
3. Both X causes Y and Y causes X (feedback loops or bidirectional causality)
4. No direct relationship between X and Y; both are related to an outside variable Z
5. No relationship or reason for the two to be related, correlation is just a statistical artifact.

Examples of Correlations and Causal Relationships

1. X causes Y, but Y does not cause X

$$X \rightarrow Y$$

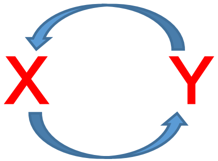
Rain and **Floods** are positively correlated

- There are more floods when it rains more
- Rain causes floods, but floods do not cause rain

$$\text{Rain} \rightarrow \text{Floods}$$

Examples of Correlations and Causal Relationships

3. Both X causes Y and Y causes X (feedback loops or bidirectional causality)



X

Y

Ideal gas law approximates relationship between pressure and temperature of a gas

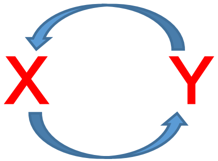
Ideal Gas Law: $PV = nRT$

Fix volume (V) and amount of substance (n). R is a constant.

- Increase temperature (T) ⇒ increase pressure (P)
- Increase pressure (V) ⇒ increase temperature (T)
- Bidirectional causality (it works in either direction)

Examples of Correlations and Causal Relationships

3. Both X causes Y and Y causes X (feedback loops or bidirectional causality)



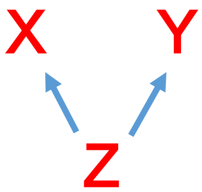
X Y

Some people argue **poverty** and **lack of education** is a feedback loop

- Poor, so can’t afford to go to school
- Don’t get an education, so stay poor

Examples of Correlations and Causal Relationships

4. No direct relationship between X and Y; both are caused by an outside variable Z



X = Number of Cows in the United States

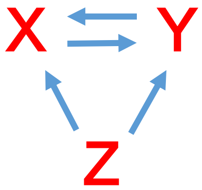
Y = Spending on Shoes in the U.S.

Z = Number of People in U.S.

- There’s more people in the United States now than before. More people means more spending on shoes and consume more milk and beef, even if rates stay relatively steady.

Examples of Correlations and Causal Relationships

Note: Can have multiple relationships between X, Y, and Z



X = Number of Firefighters dispatched to a Fire

Y = Total Damage caused by Fire; Z = Size of Fire when Firefighters are alerted to it

- Larger fires cause more damage, and more firefighters respond to larger fires
- Number of firefighters **positively correlated** with fire damage
- Actually firefighters **decrease** fire damage (correlation alone is misleading)

Examples of Correlations and Causal Relationships

5. No relationship or reason for the two to be related, correlation is just a statistical artifact

Sometimes there’s no causal relationship. Just dumb luck that two things are correlated.

- Example is wearing “lucky shirt” and your team wins (it’s only weird if it doesn’t work)
- Paul the Octopus correctly picked a large number of 2010 World Cup outcomes

Two features of spurious relationships that are statistical artifacts

- They tend to only last for a short time (it works…until it doesn’t)
- Will eventually show up if have enough trials: http://www.r-fiddle.org/#/fiddle?id=ZVISvZ4J

Dangers of Forgetting Statistical Artifacts Exist

Just because something is unlikely doesn't mean it's impossible

- Meadow's Law was a precept when prosecuting SIDS cases in 90's

> "One is a tragedy. Two is suspicious. Three is murder, unless proved otherwise."

- Sally Clark was convicted of murdering her sons in 1999. It was argued chance of two cases of SIDS occurring naturally to same parent was 73 million to 1 (99.999999% chance of guilt)

The Royal Statistical Society showed that reasoning was flawed, and got case overturned

- Prosecutor's fallacy: There are enough people that it is likely somebody has multiple children die from SIDS. Can't assume everybody guilty because it's unlikely for any one individual.
- Also, incorrectly calculated the 73 million to 1 number. Adjusting for prosecutors fallacy and incorrect odds suggested over 90% chance Sally Clark was innocent (based only on statistics)

Determining Causality

Have two variables X and Y with some sort of correlation

- How do we determine whether X causes Y?
- Need to rule out other cases

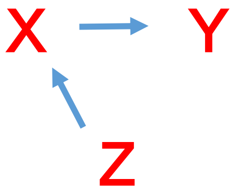
Methods for Determining Causality

Some Methods for Determining Causality

- **Controlled Lab Experiments**: make sure nothing changes except the dependent variable, and only when you change it

- **Randomized Assignment in Experiments**: Outside factors may change, but if random assignment, then they shouldn't be correlated with dependent variable or outcome

- **Instrumental Variable:** If we know Z can only affect Y indirectly through X, can use changes in Z to estimate causal impact of X on Y. We'll talk much more about this later.



Methods for Determining Causality

Some Methods for Determining Causality

- **Natural Experiments:** When you don't run the experiment (cause an event) yourself, but you can argue that it was unexpected and not related to dependent variable.
    - Political scientists like to use this to estimate effects of laws, by looking at places where laws pass by 50.1% or fail by 49.9%. They say it's essentially random whether it passed/failed.
- **Theory-based Methods:** Sometimes experiments and randomization are not possible. Need a strong reason for believing causality exists and for the data to match your theory.
    - An example is the effectiveness of new surgical procedures.
    - This is one of the reasons we rely on models in economics. It helps us tease out causality.

Dangers in Determining Causality

Some of the things that often fool people

- **Omitted Variable Bias:** We’re looking at X and Y, but ignore Z – which influences X and Y.
- **Selection Bias:** Our sample may not be representative of the general population
- **Regression to the Mean:** People attribute disappearance of extreme observations to causality, even though extreme observations should typically be expected to become less extreme
  - Suppose I flip a coin 10 times and somebody guesses it right each time. Regression to the mean says that for the next 10 flips, they’ll probably be closer to right half the time.

Dangers in Determining Causality

Some of the things that often fool people

- **Omitted Variable Bias:** We're looking at X and Y, but ignore Z – which influences X and Y.
- **Selection Bias:** Our sample may not be representative of the general population
- **Regression to the Mean:** People attribute disappearance of extreme observations to causality, even though extreme observations should typically be expected to become less extreme
  - Suppose I flip a coin 10 times and somebody guesses it right each time. Regression to the mean says that for the next 10 flips, they'll probably be closer to right half the time.
  - Important: Regression to the Mean does not mean they should get it wrong 10 times to average out their first 10 guesses.
  - It says moving forward, on average, things will be close to average.

Regression to the Mean: Example

Draw a card randomly from a standard deck

- Standard deck has 13 cards (1-10, J, Q, K) and 4 suits (♥,♦, ♠, ♣).
- Draw a card and put it back in deck, can you predict whether next card will be higher or lower?

Can do pretty well using regression to mean: http://www.r-fiddle.org/#/fiddle?id=dgOqceyR

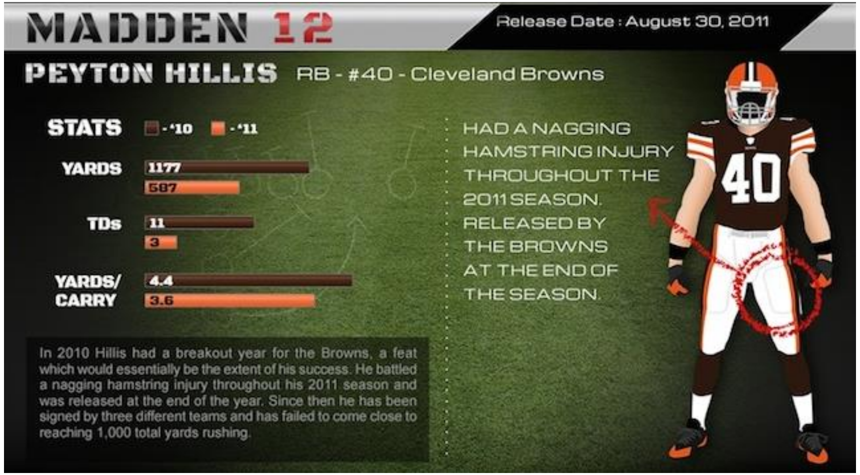
Draw a card over 7 ⇒ more often than not next card will be lower than the card you draw

- Draw a card under 7 ⇒ more often than not next card will be higher than the card you draw
- Correct around 75% of time with this method (only 45% of time with random guessing)

Examples of Why to Keep Regression to the Mean in Mind

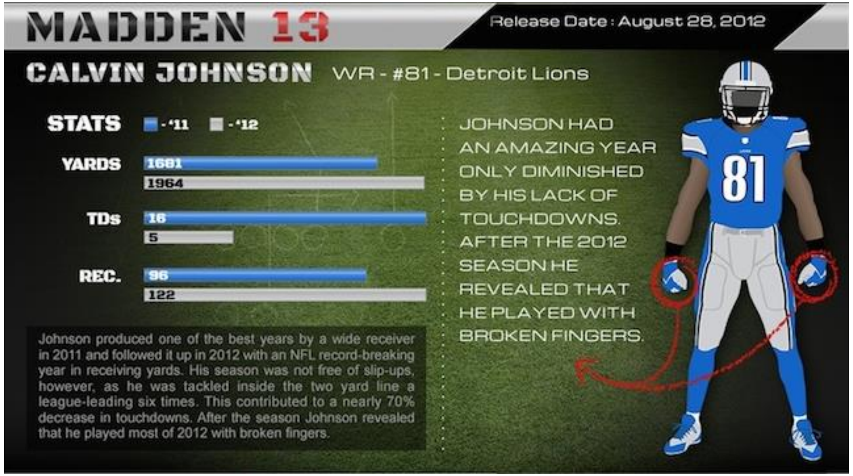
People expect extreme observations to continue when it’s due to natural fluctuations

- **Madden curse:** NFL Player with Outstanding Season gets on cover of Madden. Tend to not be as good the following year. This can mostly be explained by regression to the mean.



HAD A NAGGING HAMSTRING INJURY THROUGHOUT THE 2011 SEASON. RELEASED BY THE BROWNS AT THE END OF THE SEASON.

In 2010 Hillis had a breakout year for the Browns, a feat which would essentially be the extent of his success. He battled a nagging hamstring injury throughout his 2011 season and was released at the end of the year. Since then he has been signed by three different teams and has failed to come close to reaching 1,000 total yards rushing.



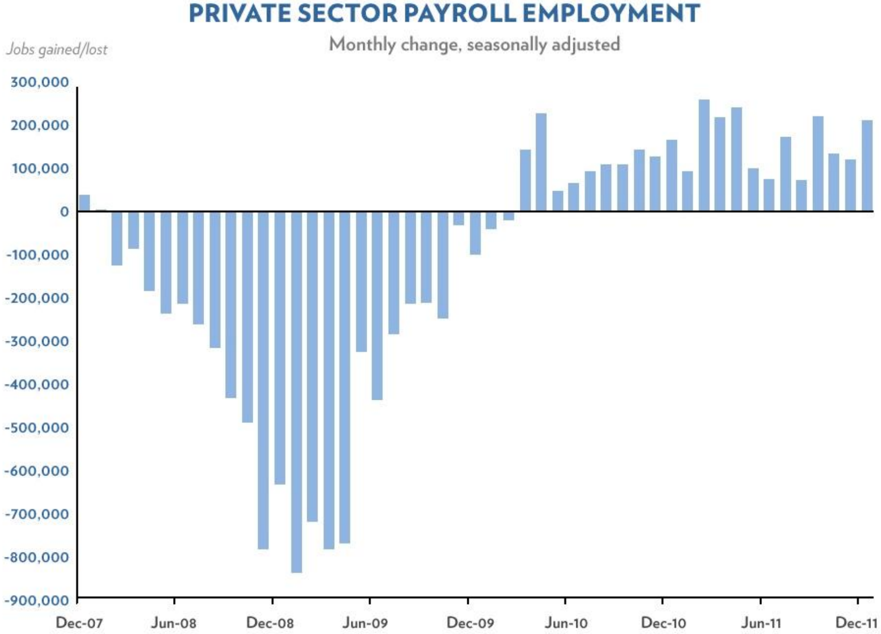
JOHNSON HAD AN AMAZING YEAR ONLY DIMINISHED BY HIS LACK OF TOUCHDOWNS. AFTER THE 2012 SEASON HE REVEALED THAT HE PLAYED WITH BROKEN FINGERS.

Johnson produced one of the best years by a wide receiver in 2011 and followed it up in 2012 with an NFL record-breaking year in receiving yards. His season was not free of slip-ups, however, as he was tackled inside the two yard line a league-leading six times. This contributed to a nearly 70% decrease in touchdowns. After the season Johnson revealed that he played most of 2012 with broken fingers.

Examples of Why to Keep Regression to the Mean in Mind

People expect extreme observations to continue when it’s due to natural fluctuations

- **Business Cycles:** Recessions don’t continue forever. They eventually get better. Politicians love to take credit for this fact, but it should be expected regardless of policy.



Examples of Why to Keep Regression to the Mean in Mind

Biggest problem is that when things are at their worst, people try lots of things.

When evaluating whether policies/treatments work, need to take into account reg. to mean

- Violent Crime Rates peaked in mid 90’s in US; local police departments across US took credit for reducing crime in their district, but was national phenomenon – not local

- People get sick, get desperate and try strange treatments, get better and think treatment worked. Need to keep in mind that the body naturally fights off most illnesses over time. Even though got better, treatment might have done nothing or even been detrimental.

Searching For Causality in Development

Want to know what things have a causal impact on economic growth and development

Some Broad Possibilities:

- Geography and Climate
- Culture
- Institutions

Searching For Causality in Development

Evaluating Effectiveness of Development and Poverty Reduction Policies

- Is microfinance effective in reducing Poverty?
- Does school quality matter for students? How much? How do we increase school quality?
- What policies are most effective for reducing deaths from Malaria?
- What determines the effectiveness of Foreign Aid?