PSY 503: Foundations of Statistical Methods in Psychological Science

Correlation,
Regression (Group Models)

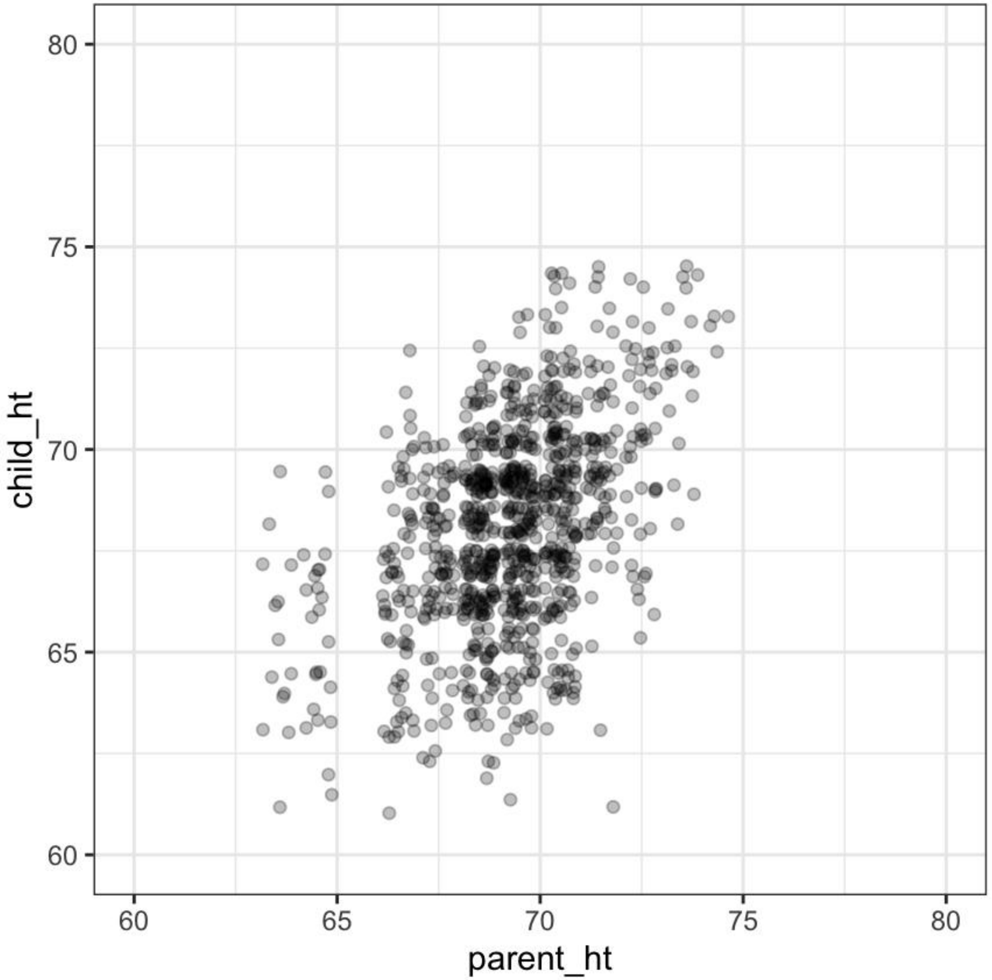
Suyog Chandramouli

311 PSH (Princeton University)

6th October, 2025

Last week..

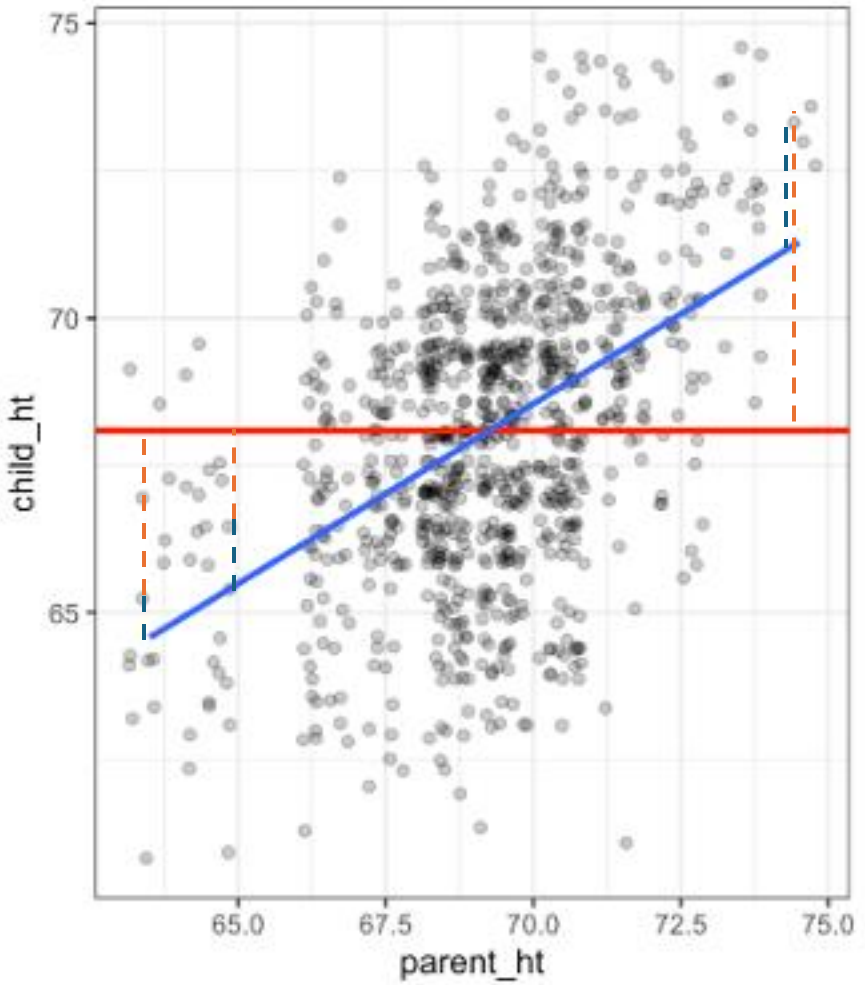
"There's a linear trend" – PSY 503 Student(s)



Last week..

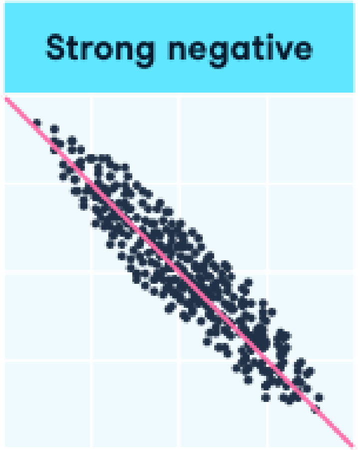
“There’s a linear trend” – PSY 503 Student (s)

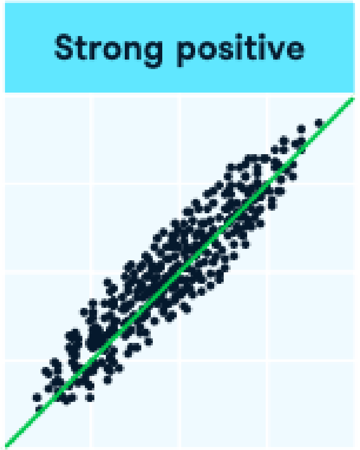
We indeed found that a line was a better fit to the data (in terms of error measures) than a line



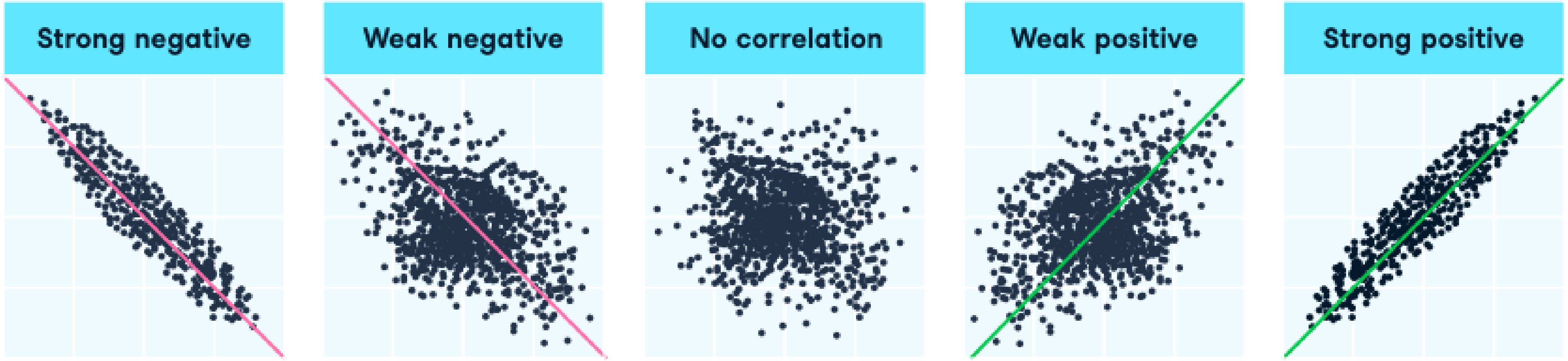
Discuss

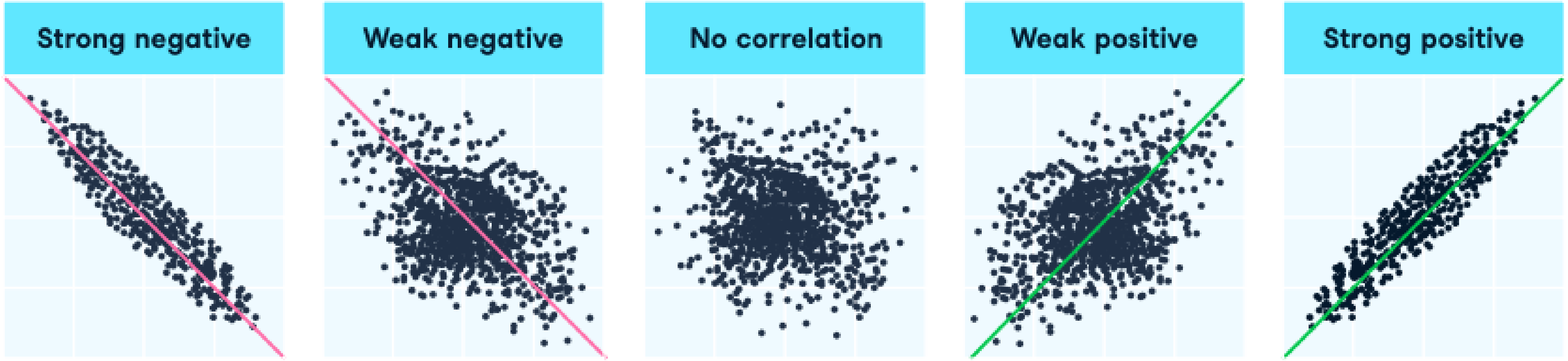
- What is a linear trend ?
- What would a strong linear trend look like ?
- What would a weak linear trend look ?


Strong negative


Strong positive


Strong negative
Weak negative
Weak positive
Strong positive


Strong negative
Weak negative
No correlation
Weak positive
Strong positive



r closer to -1

r = 0

r closer to 1

r closer to -1 r = 0 r closer to 1

Rule of thumb (varies by field):

- $|r| < 0.3$: Weak correlation
- $0.3 \leq |r| < 0.7$: Moderate correlation
- $|r| \geq 0.7$: Strong correlation

Formula

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2} * \sqrt{\sum(Y_i - \bar{Y})^2}}$$

Formula

$$r = \frac{\sum(X_i - \bar{X})\,(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2} * \sqrt{\sum(Y_i - \bar{Y})^2}}$$

- For each data point,
  - Numerator
    - Individual variation (from mean) is
      - $(X_i - \bar{X})$
      - $(Y_i - \bar{Y})$
    - Both above mean → positive product
    - Both below mean → positive product
    - One above, one below → negative product
    - Positive => X and Y deviate the same way
      Negative => X and Y deviate in the opposite ways
    - $(X_i - \bar{X})\,(Y_i - \bar{Y})$ is the joint deviation measure

Formula

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2} * \sqrt{\sum(Y_i - \bar{Y})^2}}$$

- For each data point,
  - Numerator
    - Individual variation (from mean) is
      - $(X_i - \bar{X})$
      - $(Y_i - \bar{Y})$
    - Both above mean → positive product
    - Both below mean → positive product
    - One above, one below → negative product
    - Positive => X and Y deviate the same way
      Negative => X and Y deviate in the opposite ways
    - $(X_i - \bar{X})(Y_i - \bar{Y})$ is the joint deviation measure
  - Denominator
    - How much does each variable vary alone?
    - Big spread in X? Increases denominator.
      Big spread in Y? Increases denominator
    - Larger denominator → smaller r.
    - Denominator = SD(X) × SD(Y)

Formula

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2} * \sqrt{\sum(Y_i - \bar{Y})^2}}$$

$$r = \frac{\text{coordinated movement}}{total\ (coordinated + independent) movement}$$

- For each data point,
  - Numerator
    - Individual variation (from mean) is
      - $(X_i - \bar{X})$
      - $(Y_i - \bar{Y})$
    - Both above mean → positive product
    - Both below mean → positive product
    - One above, one below → negative product
    - Positive => X and Y deviate the same way
      Negative => X and Y deviate in the opposite ways
    - $(X_i - \bar{X})\ (Y_i - \bar{Y})$ is the joint deviation measure
  - Denominator
    - How much does each variable vary alone?
    - Big spread in X? Increases denominator.
      Big spread in Y? Increases denominator
    - Larger denominator → smaller r.
    - Denominator = SD(X) × SD(Y)

Correlation Coefficient

[ˌkȯr-ə-ˈlā-shən ˌkō-ə-ˈfi-shənt]

A statistical measure of the strength of the relationship between the relative movements of two variables.

Investopedia

Correlation in R

family_id	child_ht	parent_ht
<chr>	<dbl>	<dbl>
F1	72.2	74.5
F2	73.2	74.5
F3	73.2	74.5
F4	73.2	74.5
F5	68.2	73.5
F6	69.2	73.5
F7	69.2	73.5
F8	70.2	73.5
F9	71.2	73.5
F10	71.2	73.5

Correlation in R

family_id <chr>	child_ht <dbl>	parent_ht <dbl>
F1	72.2	74.5
F2	73.2	74.5
F3	73.2	74.5
F4	73.2	74.5
F5	68.2	73.5
F6	69.2	73.5
F7	69.2	73.5
F8	70.2	73.5
F9	71.2	73.5
F10	71.2	73.5

````
```{r}
cor(height_data$child_ht,height_data$parent_ht )
```
````

```
[1] 0.4587332
```

Correlation in R

family_id <chr>	child_ht <dbl>	parent_ht <dbl>	gparent_ht <dbl>
F1	72.2	74.5	73.18186
F3	73.2	74.5	74.31761
F4	73.2	74.5	73.87115
F6	69.2	73.5	71.86452
F8	70.2	73.5	71.47048
F12	72.2	73.5	72.95794
F14	72.2	73.5	72.88320
F16	72.2	73.5	73.38607
F20	74.2	73.5	73.70816
F23	74.2	73.5	73.54220

Correlation in R

family_id <chr>	child_ht <dbl>	parent_ht <dbl>	gparent_ht <dbl>
F1	72.2	74.5	73.18186
F3	73.2	74.5	74.31761
F4	73.2	74.5	73.87115
F6	69.2	73.5	71.86452
F8	70.2	73.5	71.47048
F12	72.2	73.5	72.95794
F14	72.2	73.5	72.88320
F16	72.2	73.5	73.38607
F20	74.2	73.5	73.70816
F23	74.2	73.5	73.54220

````
```{r}
temp_data<- full_data %>%
  filter(!is.na(gparent_ht))

cor(temp_data[, c("child_ht", "parent_ht", "gparent_ht")])
```
````

```
            child_ht parent_ht gparent_ht
child_ht   1.0000000 0.4585804  0.8899514
parent_ht  0.4585804 1.0000000  0.7888931
gparent_ht 0.8899514 0.7888931  1.0000000
```

Correlation in R

family_id <chr>	child_ht <dbl>	parent_ht <dbl>	gparent_ht <dbl>
F1	72.2	74.5	73.18186
F3	73.2	74.5	74.31761
F4	73.2	74.5	73.87115
F6	69.2	73.5	71.86452
F8	70.2	73.5	71.47048
F12	72.2	73.5	72.95794
F14	72.2	73.5	72.88320
F16	72.2	73.5	73.38607
F20	74.2	73.5	73.70816
F23	74.2	73.5	73.54220

````
```{r}
temp_data<- full_data %>%
  filter(!is.na(gparent_ht))

cor(temp_data[, c("child_ht", "parent_ht", "gparent_ht")])
```
````

```
             child_ht parent_ht gparent_ht
child_ht   1.0000000 0.4585804  0.8899514
parent_ht  0.4585804 1.0000000  0.7888931
gparent_ht 0.8899514 0.7888931  1.0000000
```

````
```{r}
full_data %>%
  filter(!is.na(gparent_ht)) %>%
  corrr::correlate()
```
````

R Console

cor_df
3 x 4

A tibble: 3 × 4

term <chr>	child_ht <dbl>	parent_ht <dbl>	gparent_ht <dbl>
child_ht	NA	0.4585804	0.8899514
parent_ht	0.4585804	NA	0.7888931
gparent_ht	0.8899514	0.7888931	NA

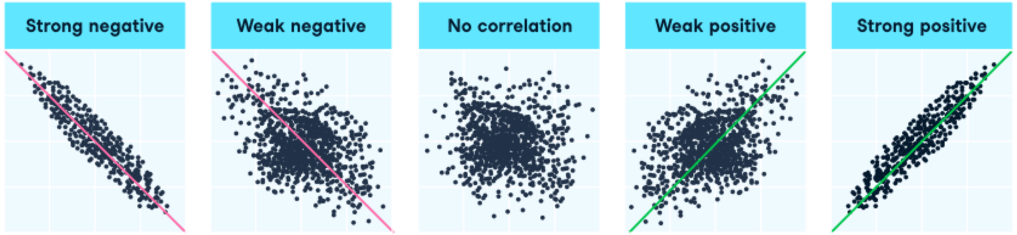
3 rows

Correlation Formula (Pearson's r)

- Summary statistic about a relationship between two variables.
- Intuition
  - How much do x and y vary together, compared to how much they vary individually?

Correlation Formula (Pearson's r)

- Summary statistic about a relationship between two variables.
- Intuition
  - How much do x and y vary together, compared to how much they vary individually?
- Strengths
  - Standardized
  - Captures strength and direction
  - Scale-independent
  - Symmetric
  - Efficient
  - Used in other statistical calculations (PCA, factor analysis, etc.)
  - Intuitive
  - Computationally feasible, etc.



Correlation Coefficient & Regression

- Connections
  - r and slope share signs

Correlation Coefficient & Regression

- Connections
  - r and slope share signs
  - When variables are standardized, slope (β_1) = r

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2} * \sqrt{\sum(Y_i - \bar{Y})^2}}$$

Correlation Coefficient & Regression

- Connections
  - r and slope share signs
  - When variables are standardized, slope (β_1) = r
  - R-squared (is r^2) is the *coefficient of determination*
    - It represents the proportion of variance in the dependent variable explained by the independent variable(s)

Correlation Coefficient & Regression

- Connections
  - r and slope share signs
  - When variables are standardized, slope (β_1) = r
  - R-squared (is r^2) is the *coefficient of determination*
    - It represents the proportion of variance in the dependent variable explained by the independent variable(s)
      - $R^2 = 0$: The model explains none of the variability in the data
      - $R^2 = 1$: The model explains all the variability in the data
      - A measure of “goodness of fit” but it doesn’t account for “overfitting”

$$R^2 = \frac{Explained\ Variance}{Total\ Variance}$$

Aside: Adjusted-R-squared

- Modifies R^2 to account for the number of predictors in the model
  - Always <= R^2
    - *equal for simple linear regression*
    - *Higher the better*
    - *Can be negative*

Aside: Adjusted-R-squared

- Modifies R^2 to account for the number of predictors in the model
  - Always <= R^2
    - *equal for simple linear regression*
    - *Higher the better*
    - *Can be negative*
- Penalizes the addition of unnecessary predictors
- **Formula for Adjusted R-squared**

k = number or predictors

  - Where:
    - n is the number of observations
    - k is the number of predictors

$$Adjusted\ R^2 = 1 - \frac{(1-R^2)(n-1)}{n-k-1}$$

Aside: Adjusted-R-squared

- Modifies R^2 to account for the number of predictors in the model
  - Always <= R^2
    - *equal for simple linear regression*
    - *Higher the better*
    - *Can be negative*
- Penalizes the addition of unnecessary predictors
- **Formula for Adjusted R-squared**

k = number or predictors

  - Where:
    - n is the number of observations
    - k is the number of predictors

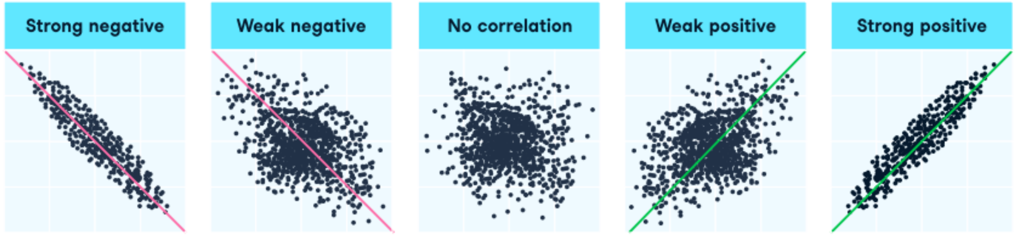
$$Adjusted\ R^2 = 1 - \frac{Unexplained\ Variance}{Total\ Variance} * \frac{n-1}{n-k-1}$$

Aside: When to use Adjusted-R-squared

- Comparing (nested) models with different numbers of predictors
- Assessing whether additional predictors improve the model
- Guard against overfitting in multiple regression

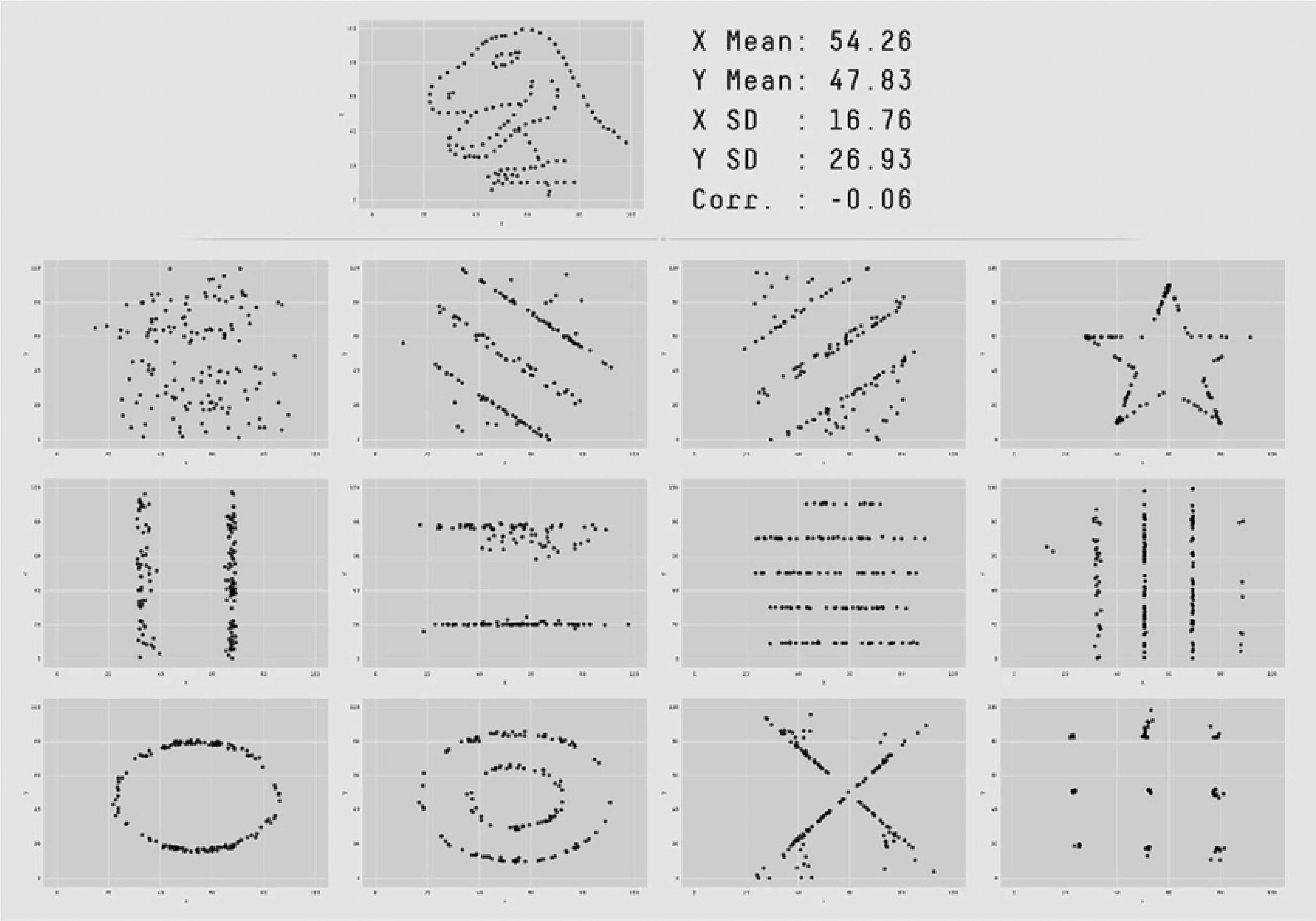
Correlation Formula (Pearson's r)

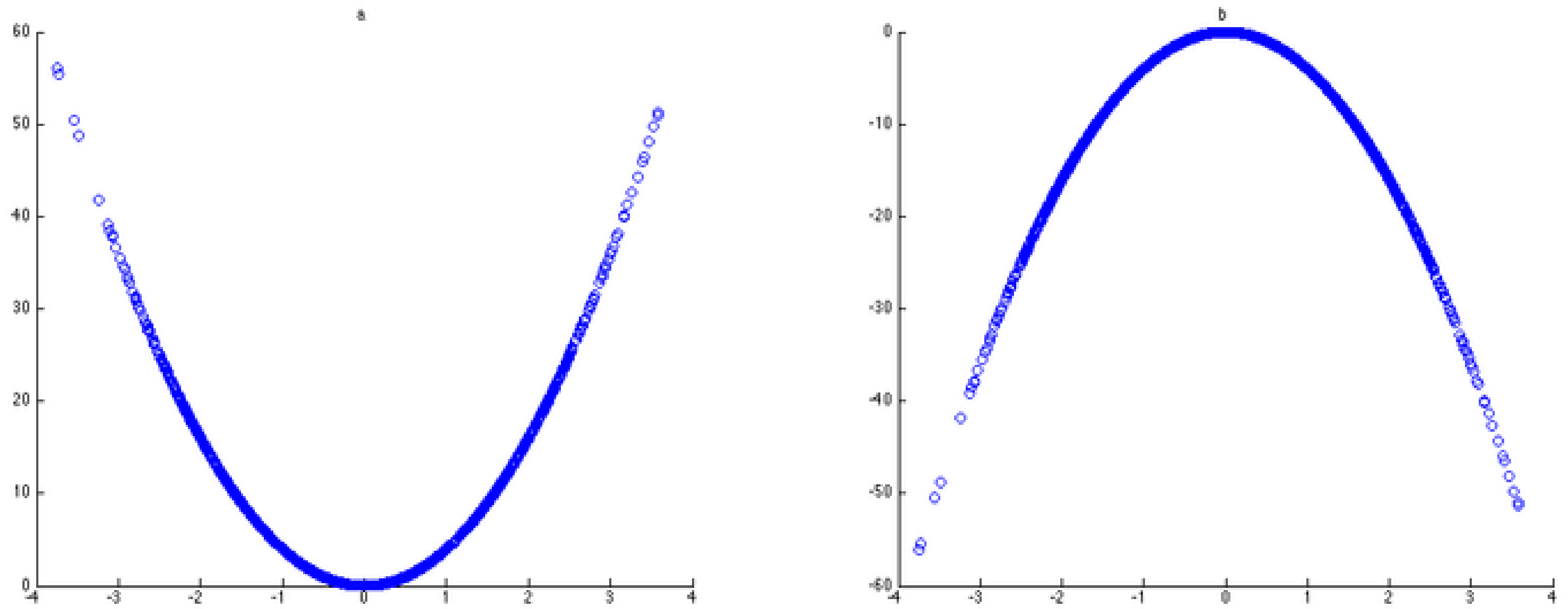
- Summary statistic about a relationship between two variables.
- Intuition
  - How much do x and y vary together, compared to how much they vary individually?
- Strengths
  - Standardized
  - Captures strength and direction
  - Scale-independent
  - Symmetric
  - Efficient
  - Used in other statistical calculations (PCA, factor analysis, etc.)
  - Intuitive
  - Computationally feasible, etc.



Discuss

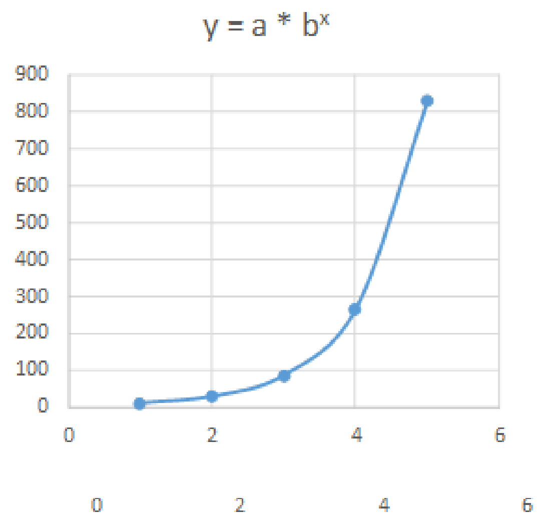
- Given that r captures strength & direction of linear relationship, is it of any use to visualize the data?

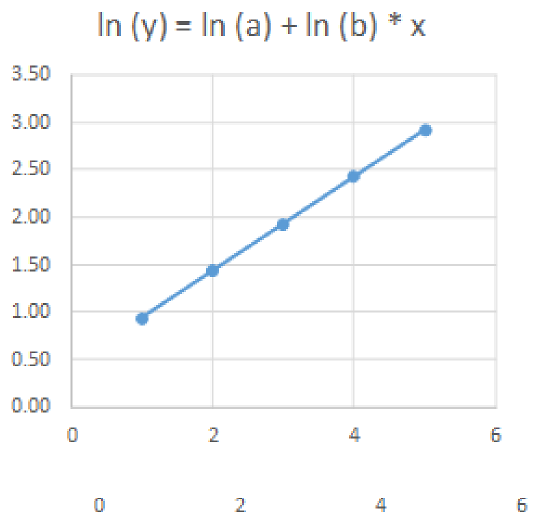

X Mean: 54.26
Y Mean: 47.83
X SD : 16.76
Y SD : 26.93
Corr. : -0.06


a
60
50
40
30
20
10
0
-4
-3
-2
-1
0
1
2
3
4
b
0
-10
-20
-30
-40
-50
-60
-4
-3
-2
-1
0
1
2
3
4

Aside: Handling non-linear relationships

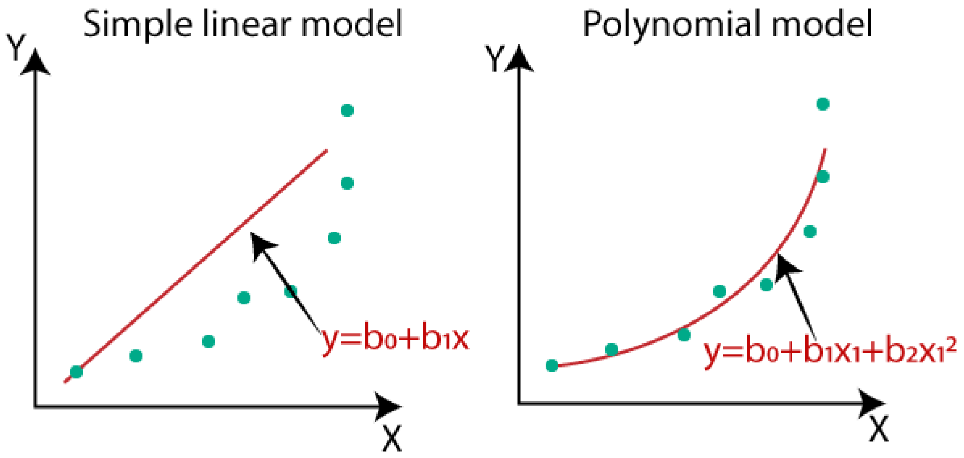
- For non-linear relationships, consider: Transforming variables (e.g., log, square root)





Aside: Handling non-linear relationships

- For non-linear relationships, consider: Transforming variables (e.g., log, square root)
- Using non-linear regression techniques



Aside: Handling non-linear relationships

- For non-linear relationships, consider: Transforming variables (e.g., log, square root)

- Using non-linear regression techniques

- Employing non-parametric correlation measures (e.g., Spearman's rho, mutual information, etc.)

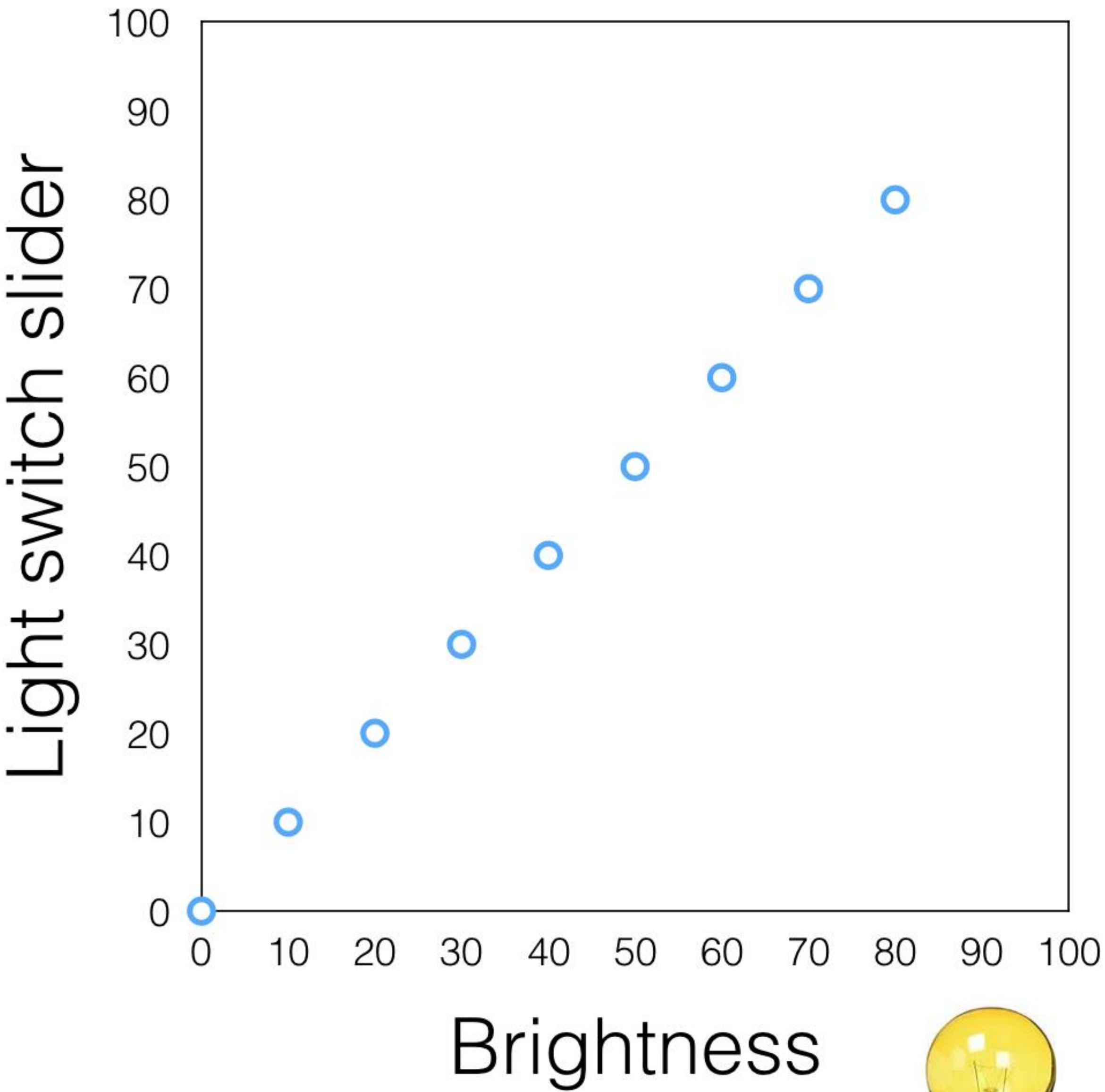
Correlation and causation

Causal influence diagram

- Light switch slider → Brightness

Causal influence diagram

- Light switch slider → Brightness
- Correlation is a common feature of causation


Light switch slider
100
90
80
70
60
50
40
30
20
10
0
0 10 20 30 40 50 60 70 80 90 100
Brightness

Correlation and causation

- One variable can cause changes in another variable and produce correlation

Correlation and causation

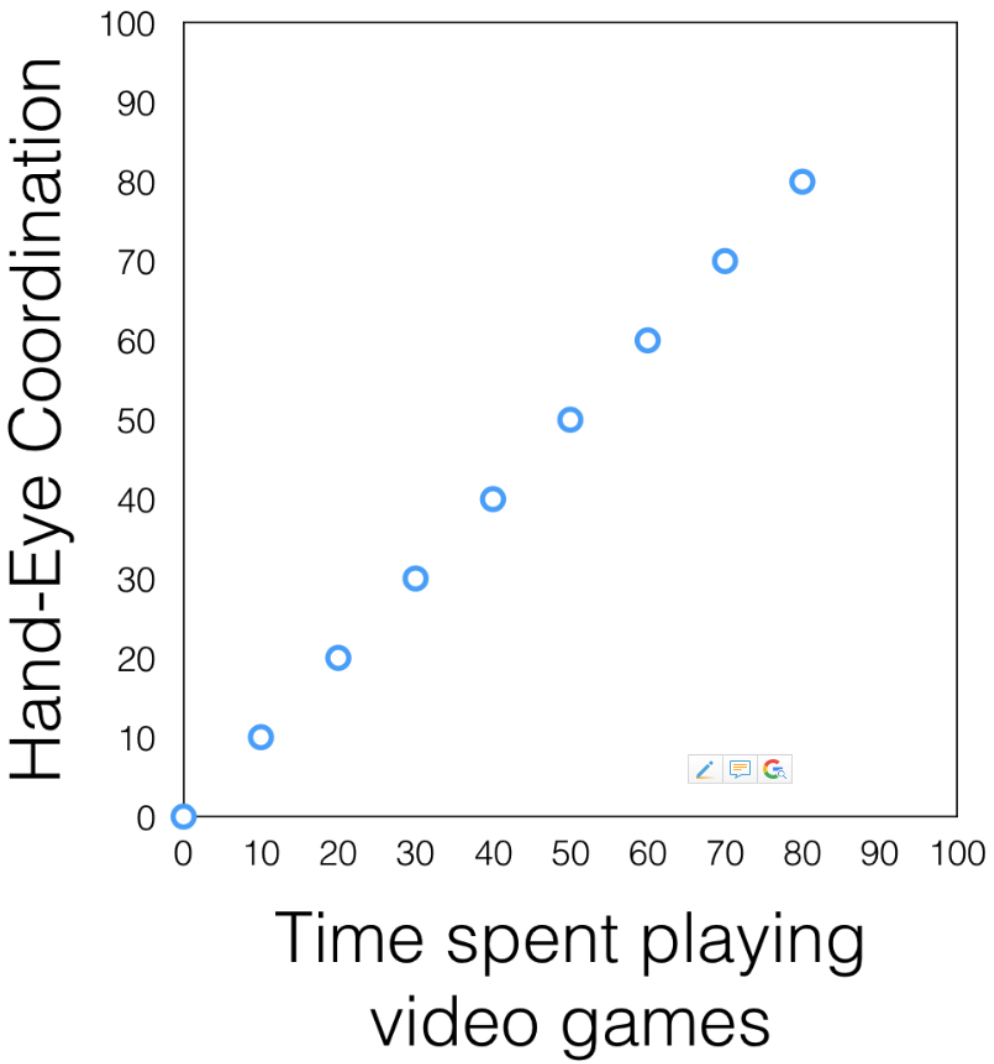
- One variable can cause changes in another variable and produce correlation

- **BUT**, correlation can also mean other things…

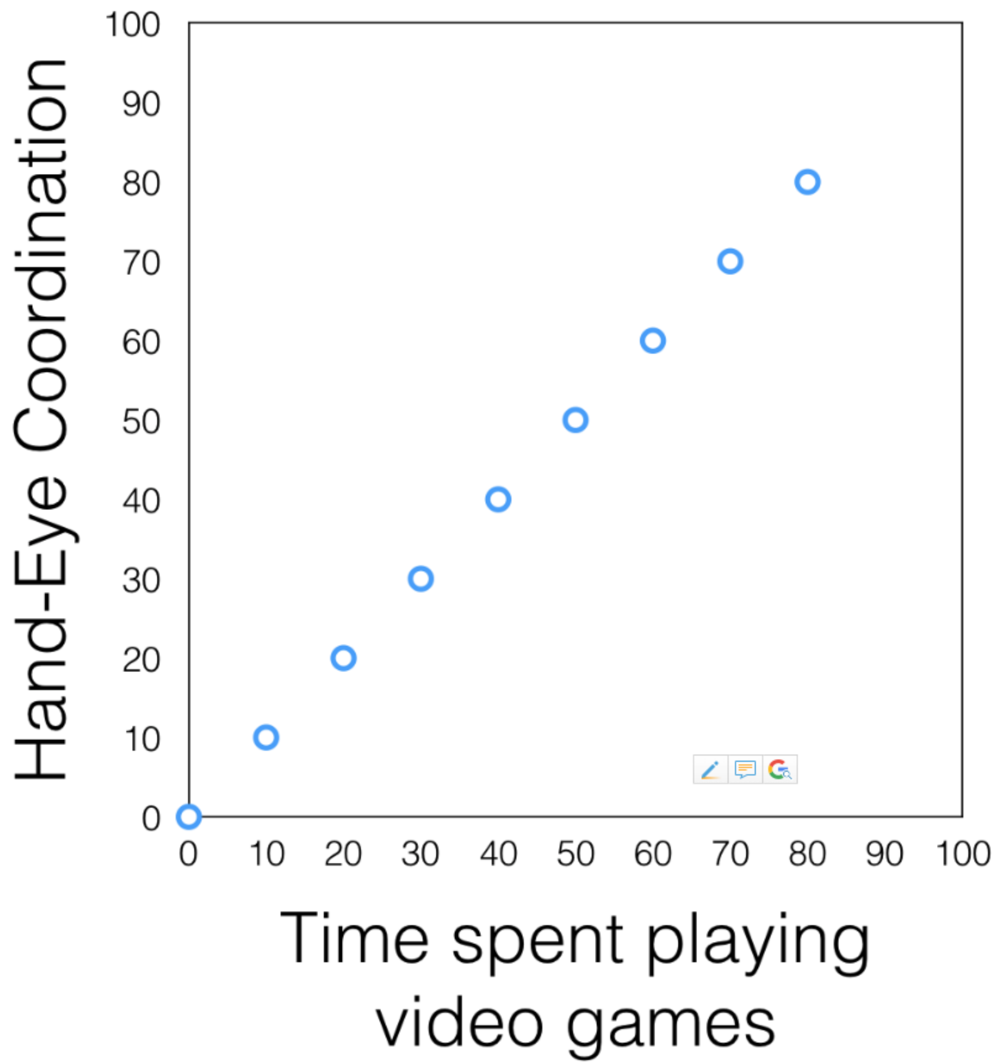
Correlation and causation

- One variable can cause changes in another variable and produce correlation
- **BUT**, correlation can also mean other things…

Causal directionality



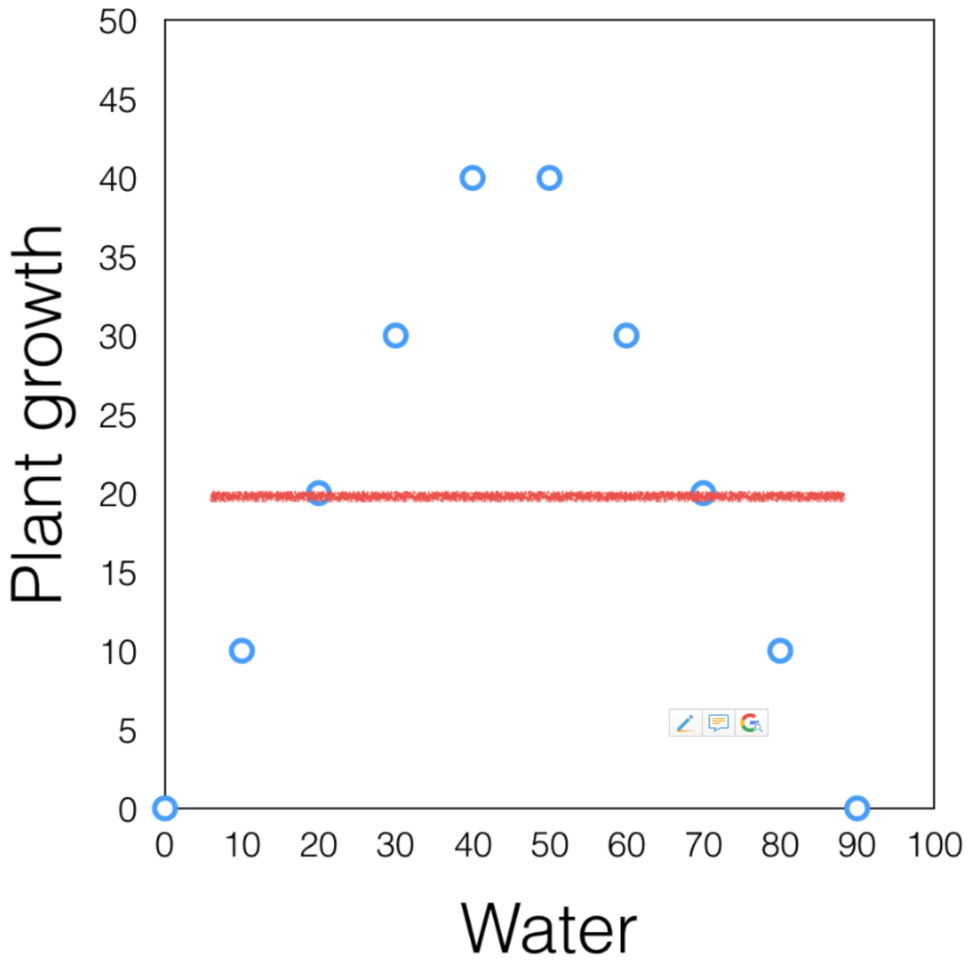
Causal directionality



A → B ?

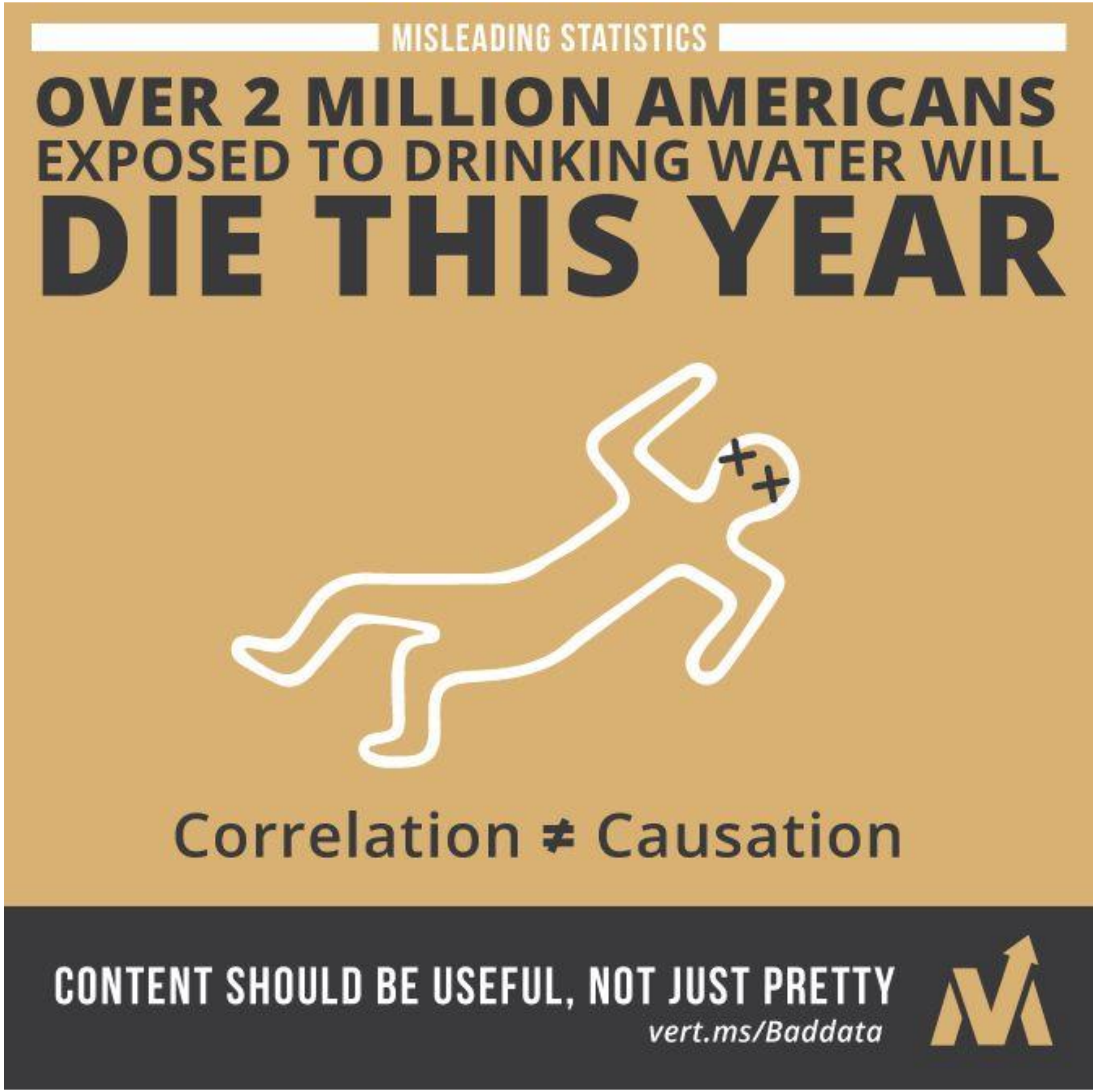
B → A ?

Nonlinearity problem

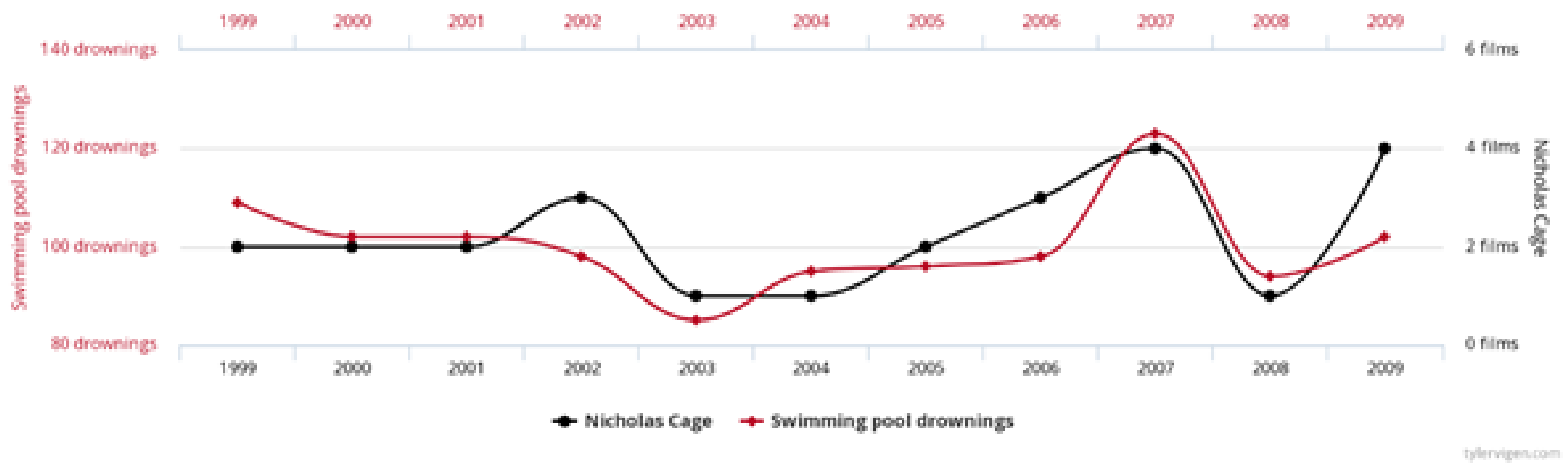


Chance problem

- Correlations between two variables can occur by chance, and be completely meaningless


MISLEADING STATISTICS
OVER 2 MILLION AMERICANS EXPOSED TO DRINKING WATER WILL DIE THIS YEAR
Correlation ≠ Causation
CONTENT SHOULD BE USEFUL, NOT JUST PRETTY
vert.ms/Baddata

Number of people who drowned by falling into a pool
correlates with
Films Nicolas Cage appeared in



More spurious correlations

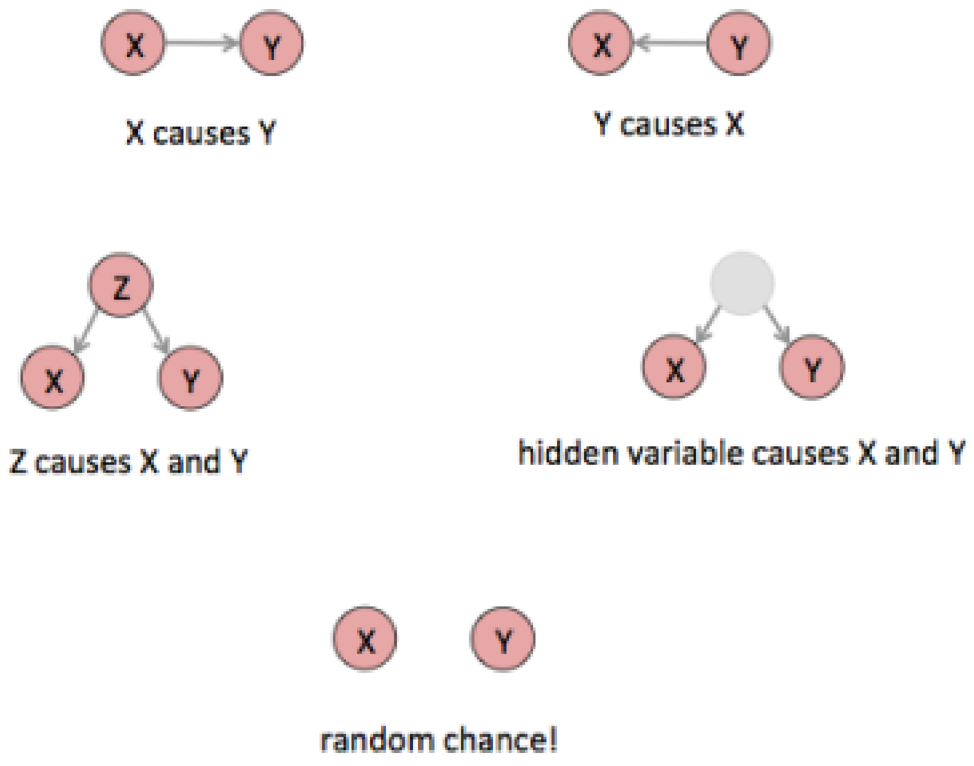
- https://www.tylervigen.com/spurious-correlations

Correlation and causation

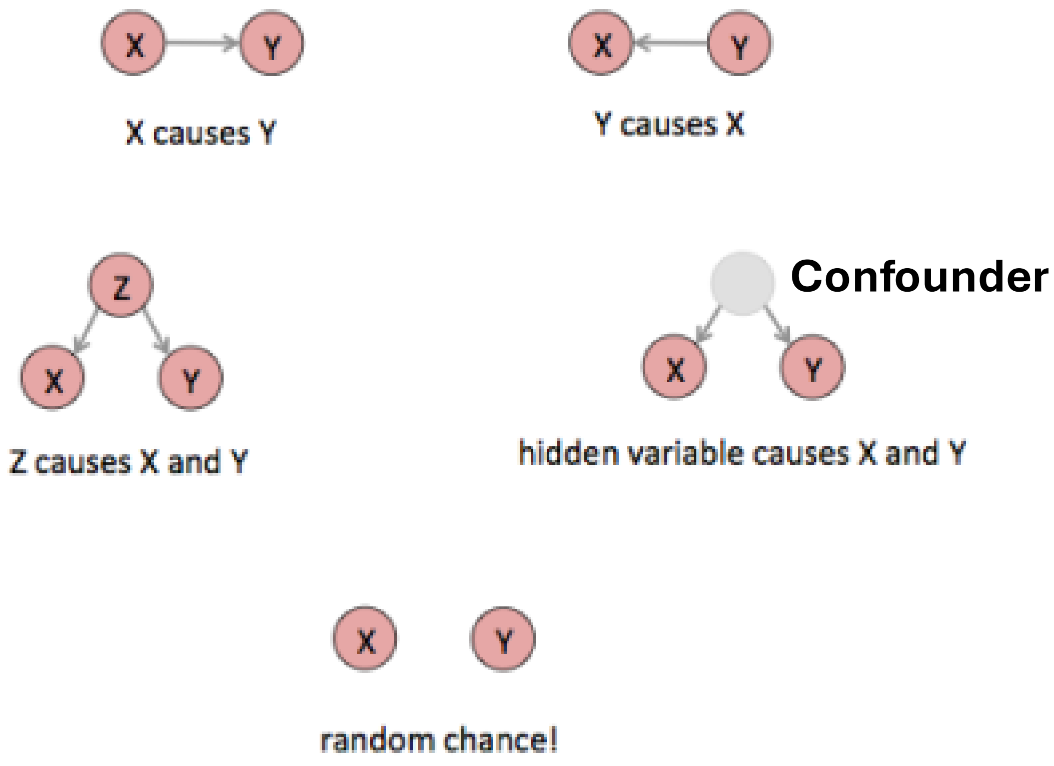
- One variable can cause changes in another variable and produce correlation

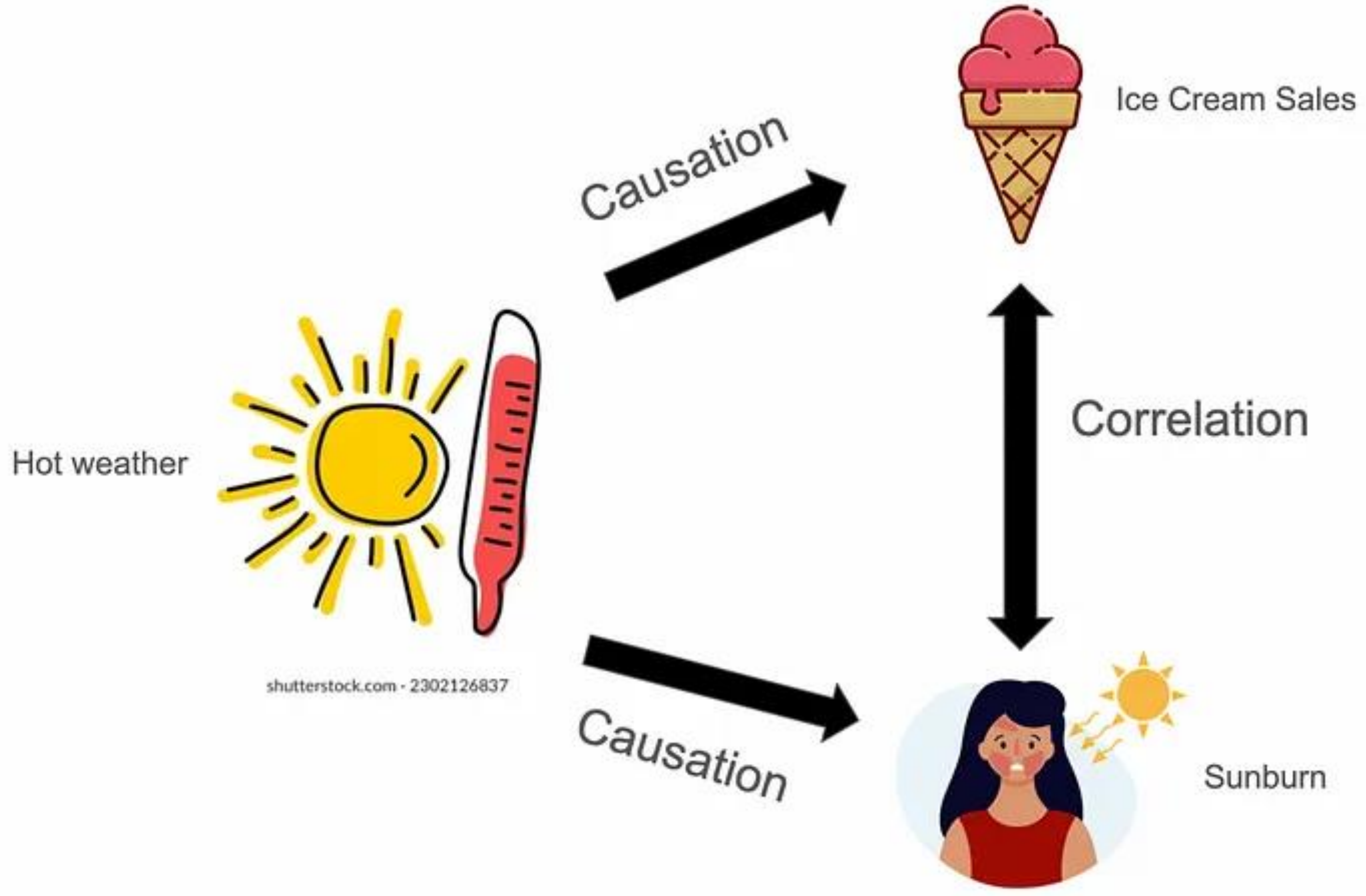
- **BUT**, correlation can also mean other things…
  1. Causal direction problem
  2. Common Causes
  3. Non-linear problem
  4. Spurious correlations
  5. Chance problem

How correlation happens



How correlation happens




Ice Cream Sales
Causation
Correlation
Hot weather
shutterstock.com · 2302126837
Causation
Sunburn

Regression (for categorical variables)

What Are Categorical Predictors?

- (Independent) Variables that represent categories or groups, not continuous values
  - Examples:
    - Gender (M/F/Non-binary)
    - Treatment groups (Control/ Treatment)
    - Education level..

What Are Categorical Predictors?

- (Independent) Variables that represent categories or groups, not continuous values
    - Examples:
        - Gender (M/F/Non-binary)
        - Treatment groups (Control/ Treatment)
        - Education level..

- We are still assuming continuous outcomes

What Are Categorical Predictors?

- (Independent) Variables that represent categories or groups, not continuous values
  - Examples:
    - Gender (M/F/Non-binary)
    - Treatment groups (Control/ Treatment)
    - Education level..

- We are still assuming continuous outcomes

- Types:
  - Nominal: Categories without a natural order (e.g., color)
  - Ordinal: Categories with a natural order (e.g., education level)

What Are Categorical Predictors?

- (Independent) Variables that represent categories or groups, not continuous values
  - Examples:
    - Gender (M/F/Non-binary)
    - Treatment groups (Control/ Treatment)
    - Education level..
- We are still assuming continuous outcomes
- Types:
  - Nominal: Categories without a natural order (e.g., color)
  - ~~Ordinal: Categories with a natural order (e.g., education level)~~

Previously..

Model = lm(child_height ~ parent_height , data = galton_data)

$\hat{Y}_i$ = intercept + slope * parent_ht

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon$$

Previously..

Model = lm(child_height ~ parent_height, data = galton_data)

$\hat{Y}_i$ = intercept + slope * parent_ht

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon$$

Let's assume we don't have parent heights.
But, we know the gender of the children.

lm() for categorical predictors

*Model = lm(child_height ~ **child_gender** , data = galton_data)*

- Use of lm () doesn’t change. R automatically handles this.

lm() for categorical predictors

*Model = lm(child_height ~ **child_gender** , data = galton_data)*

- Use of lm () doesn't change. R automatically handles this.
- Turns out that even the equation does not change much

coefficients

$\hat{Y}_i$ = intercept + ~~slope~~ * **child_gender**

$$Y_i = \beta_0 + \beta_1 \boldsymbol{D}_i + \epsilon$$

lm() for categorical predictors

*Model = lm(child_height ~ **child_gender** , data = galton_data)*

- Use of lm () doesn't change. R automatically handles this.

- Turns out that even the equation does not change much

- **BUT**
  - Regression needs numerical data
    - We convert categorical data to ***dummy variables***
      - Gender
        - Female = 0, Male = 1..

coefficients

$\hat{Y}_i$ = intercept + ~~slope~~ * **child_gender**

$$Y_i = \beta_0 + \beta_1 \boldsymbol{D}_i + \epsilon$$

- D is the dummy variable
- Interpreting coefficients
  - β_0: Mean of the reference group
  - β_1: Difference between groups

Dummy coding : The 0-1 Representation

- Dummy variables are typically represented using 0 and 1
  - 0: Absence of the category
  - 1: Presence of the category
- Example with three categories:
  - Category A: (1, 0)
  - Category B: (0, 1)
  - Category C: (0, 0)
- This 0-1 coding simplifies interpretation and computation

coefficients

$\hat{Y}_i$ = intercept + ~~slope~~ * **categorical_predictor**

$$Y_i = \beta_0 + \beta_1 \boldsymbol{D_i} + \epsilon$$

Dummy coding: practical example

- Consider a categorical variable "Region" with levels: North, South, East, West
- Dummy coding using 0-1 representation:
  - North: (0, 0, 0) [Reference level]
  - South: (1, 0, 0)
  - East: (0, 1, 0)
  - West: (0, 0, 1)
- Resulting model: $Y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3 + \varepsilon$
  - Where D_1, D_2, D_3 are dummy variables for South, East, and West respectively
- R automatically handles this coding in lm() function

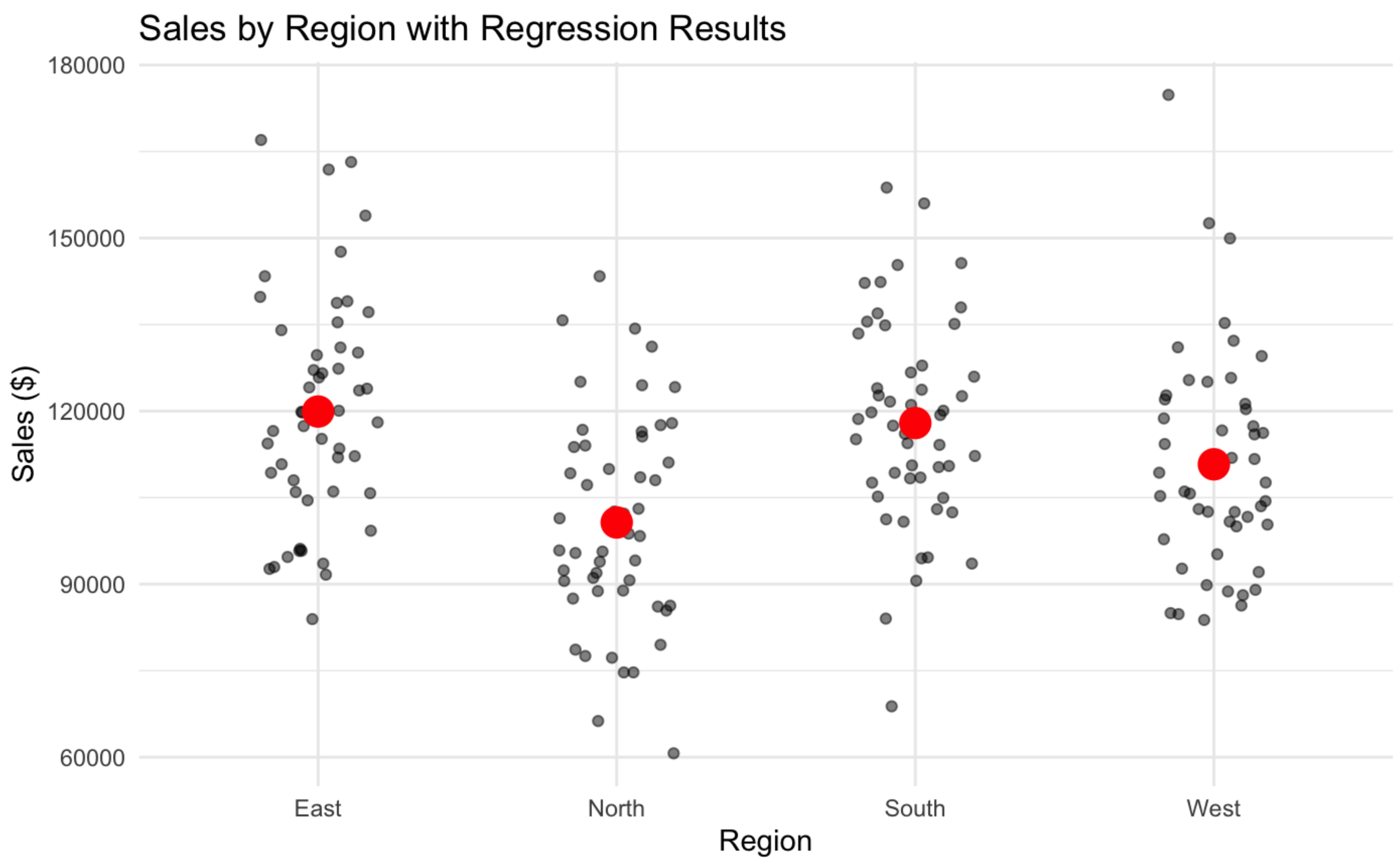
Dummy coding: practical example

- $y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3 + \varepsilon$
  - β_0 is still called the intercept
  - β_1, β_2, β_3 are often called "coefficients" rather than slopes
- Interpretation of coefficients:
  - β_0: Mean of the reference group (intercept)
  - β_1, β_2, β_3: Differences from the reference group

- Example with Region: $y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3$
  - β_0: Mean for North (reference level)
  - β_1: Difference between South and North
  - β_2: Difference between East and North
  - β_3: Difference between West and North

- The term "slope" is less commonly used with categorical predictors, but the coefficients represent the "change" associated with each category

Visualizing the data



- To be continued