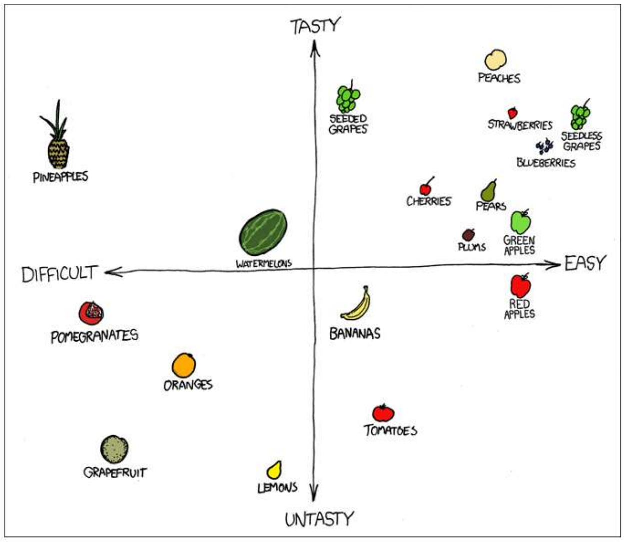
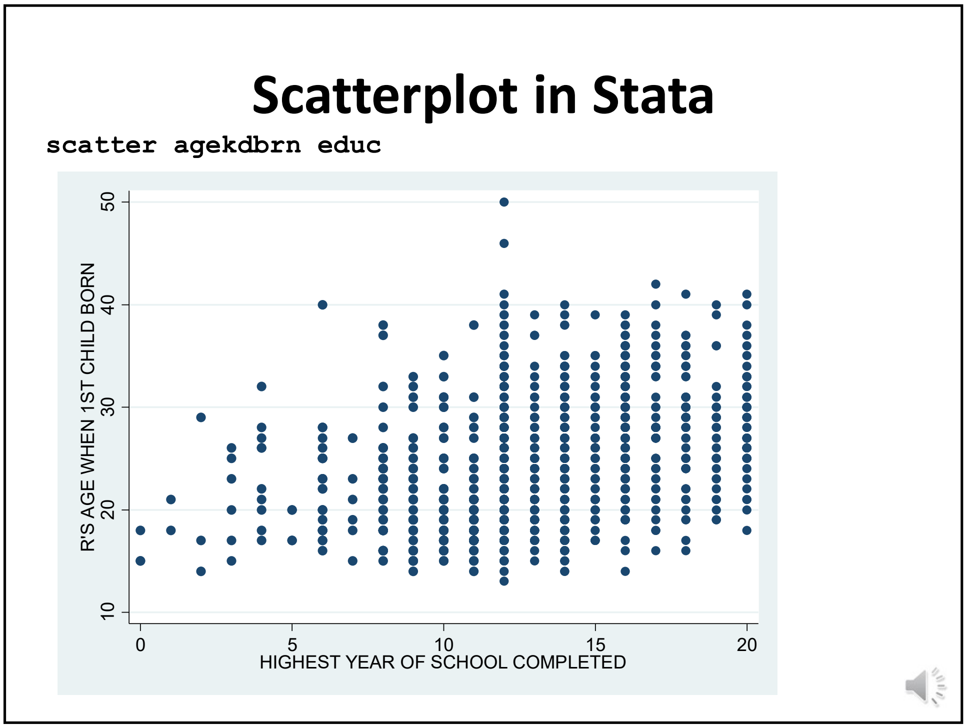
Correlation: Evaluating Relationships

" Why can't you tell me you love me without all the charts and graphs ?! "

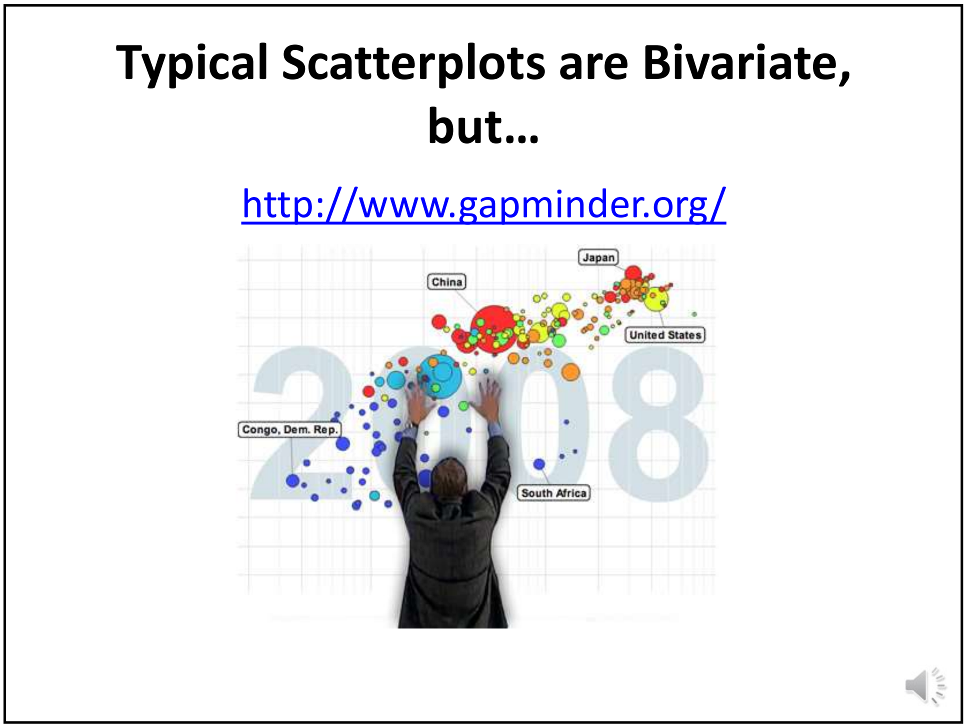
Scatterplot





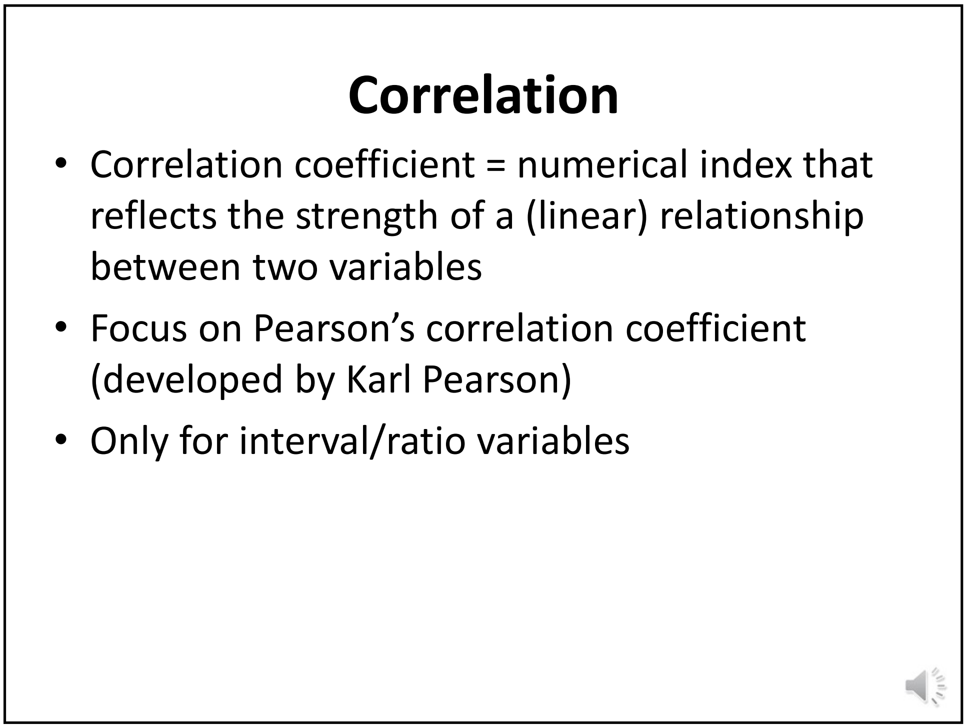
Scatterplot in Stata

```
scatter agekdbrn educ
```



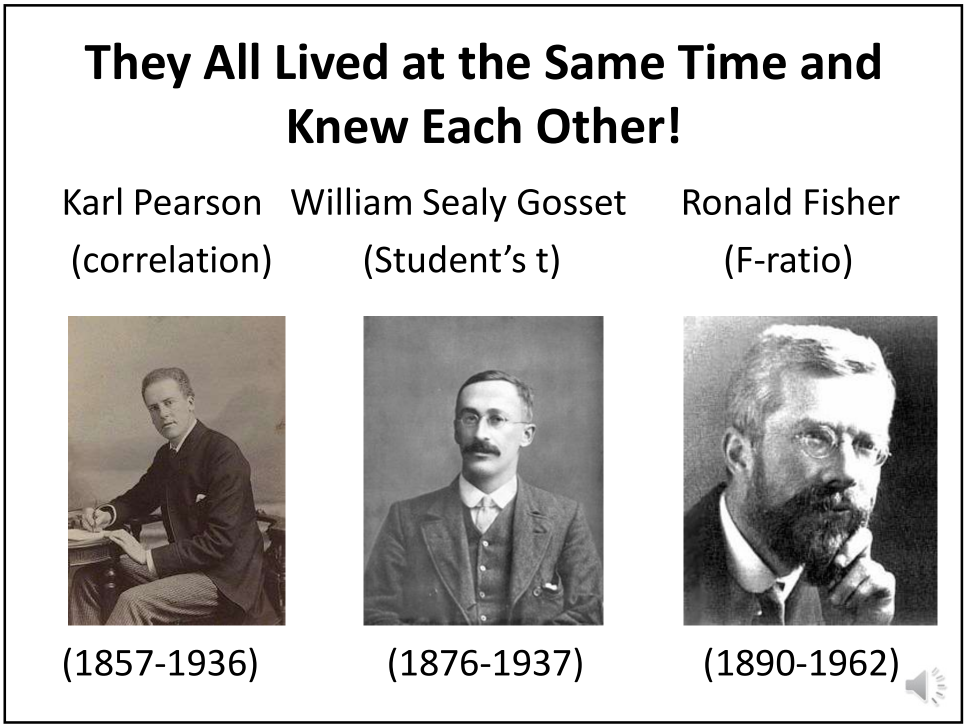
Typical Scatterplots are Bivariate, but…

http://www.gapminder.org/



Correlation

- Correlation coefficient = numerical index that reflects the strength of a (linear) relationship between two variables
- Focus on Pearson’s correlation coefficient (developed by Karl Pearson)
- Only for interval/ratio variables



They All Lived at the Same Time and Knew Each Other!

Karl Pearson	William Sealy Gosset	Ronald Fisher
(correlation)	(Student’s t)	(F-ratio)
(1857-1936)	(1876-1937)	(1890-1962)

Gosset, Pearson, Fisher

- Gosset studied under Karl Pearson at University College, London & published his article on t distribution (that came out of his work at Guinness!) in Pearson's journal Biometrika
- Gosset was for a long time the only figure on friendly terms with both Pearson and Fisher
- W. E. Deming, of the US Department of Agriculture: “Karl Pearson and R. A. Fisher disagree almost to the point of taking up arms on some questions in statistics.”

Interpreting Direction

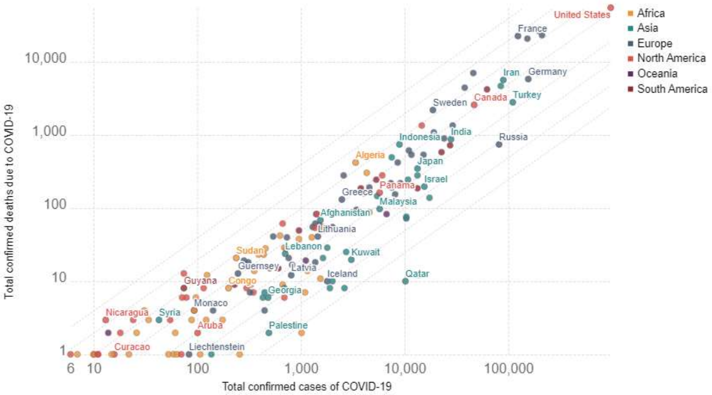
- Correlation coefficient: from –1 to +1 (0 = no relationship)
- Direction: positive (direct) and negative (indirect or inverse) correlation
- Positive = variables change in the same direction (both increase or both decrease)
- Negative = variables change in opposite directions (one increases, the other decreases)

Positive Correlation

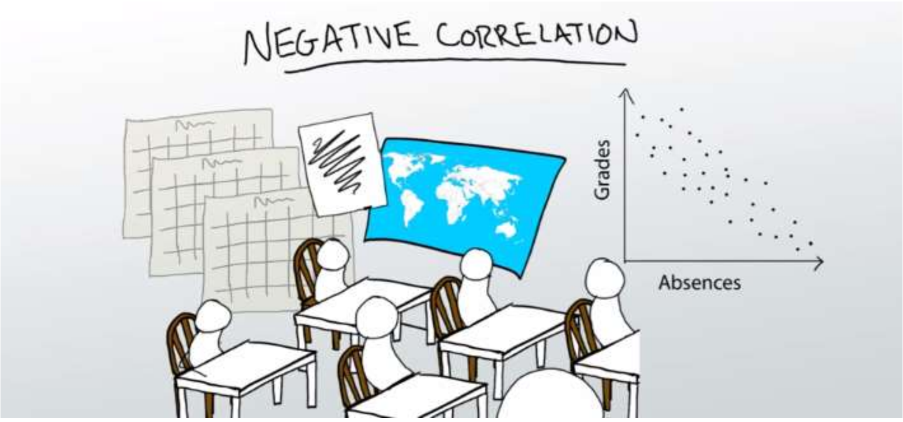
Total confirmed COVID-19 deaths vs. cases, Apr 27, 2020

Our World in Data

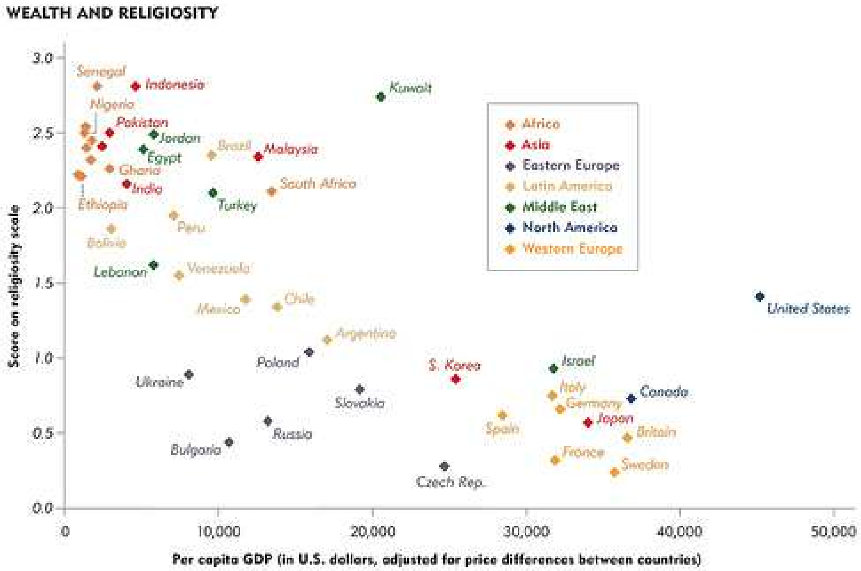
The number of confirmed cases is lower than the number of total cases. The main reason for this is limited testing. The grey lines show the corresponding case fatality rates, CFR (the ratio between confirmed deaths and confirmed cases).



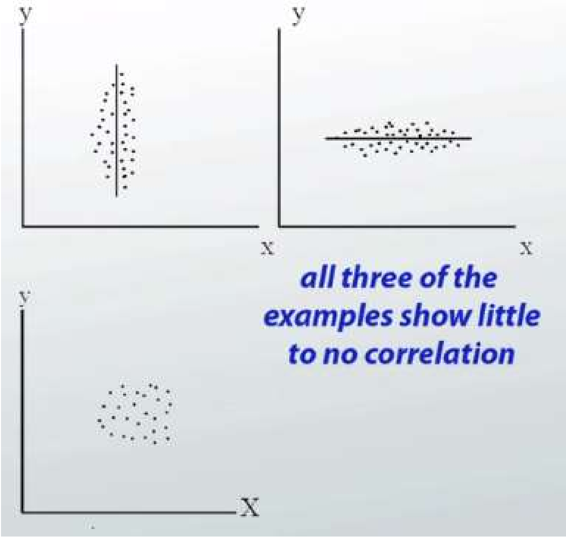
Source: European CDC – Situation Update Worldwide – Last updated 27th April, 11:15 (London time)   OurWorldInData.org/coronavirus • CC BY



Negative Correlation



No Correlation



all three of the examples show little to no correlation

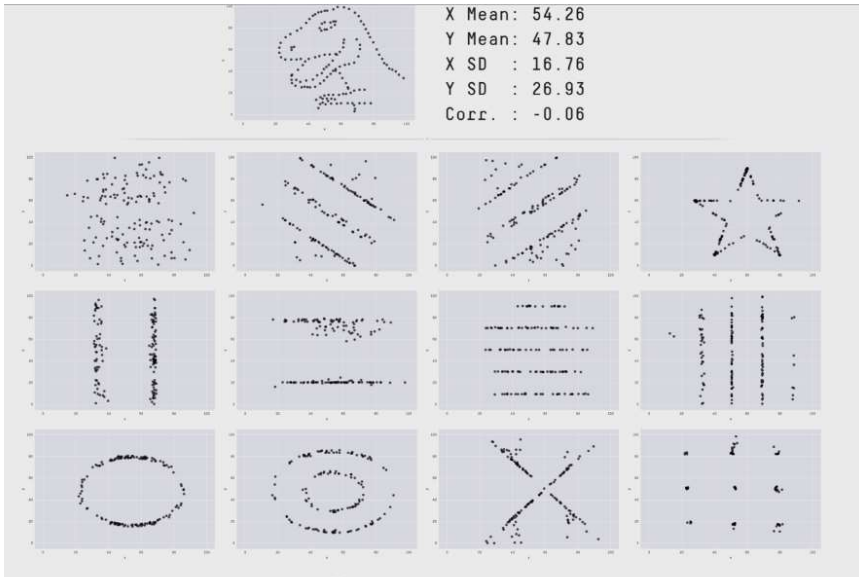
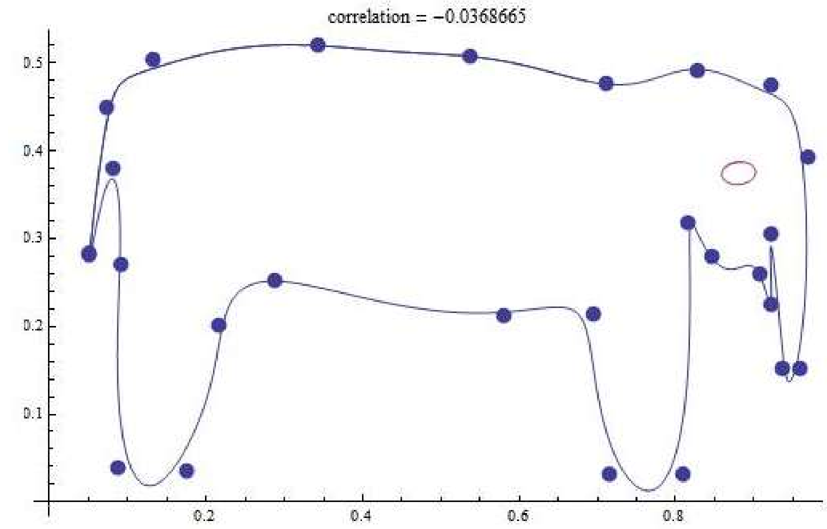


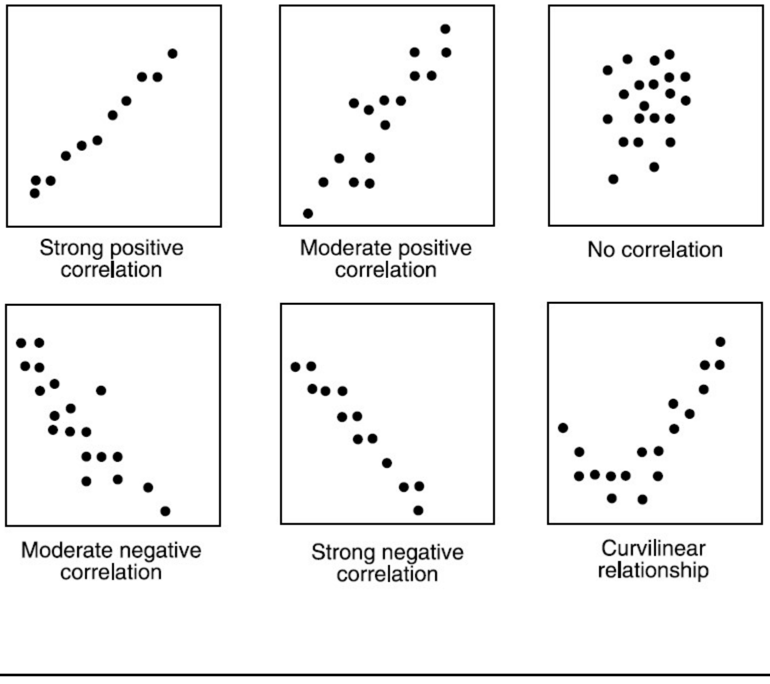
Fig 2. The ***Datasaurus Dozen****. While different in appearance, each dataset has the same summary statistics (mean, standard deviation, and Pearson's correlation) to two decimal places.*

- https://www.autodeskresearch.com/publications/samestats



Interpreting Strength

- Strength of the relationship = absolute value:
  - 0.8 to 1 = very strong
  - 0.6 to 0.8 = strong
  - 0.4 to 0.6 = moderate
  - 0.2 to 0.4 = weak
  - 0 to 0.2 = very weak or no relationship



Strong positive correlation

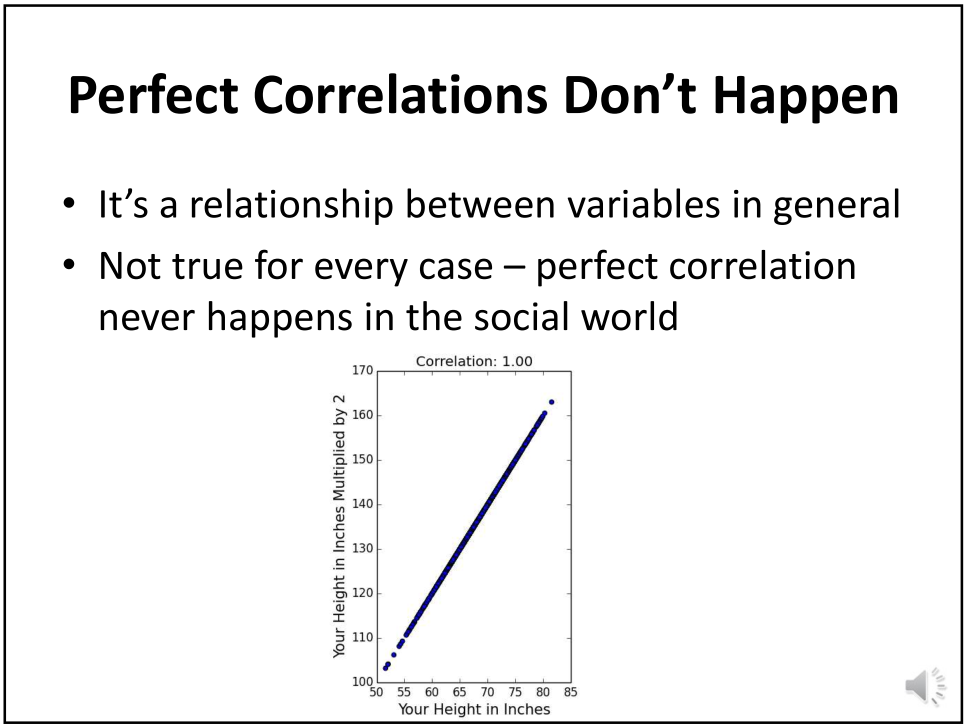
Moderate positive correlation

No correlation

Moderate negative correlation

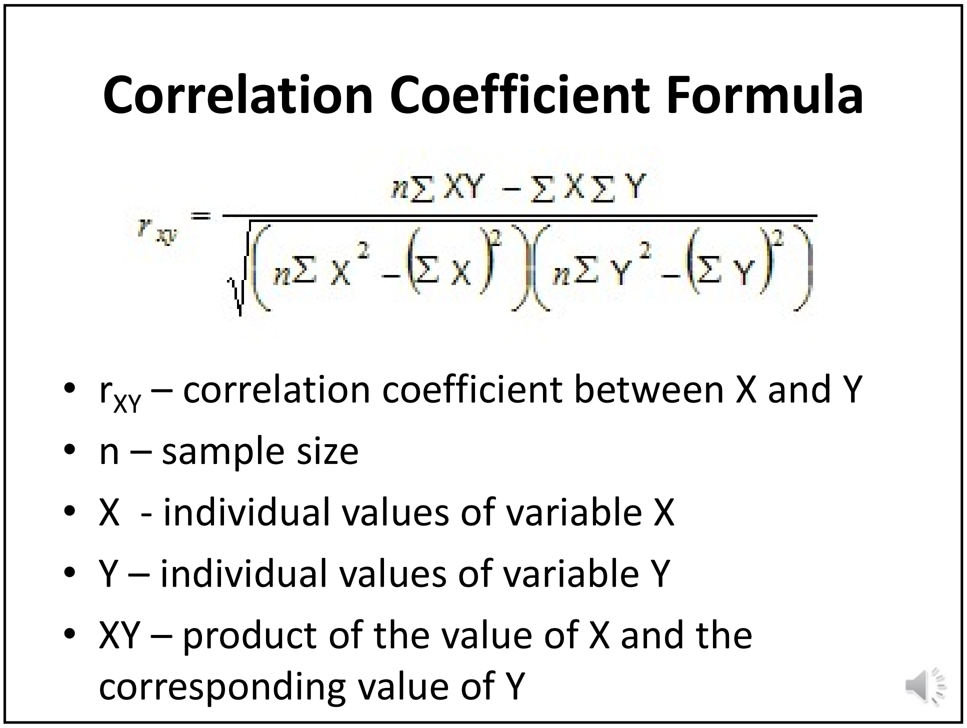
Strong negative correlation

Curvilinear relationship



Perfect Correlations Don't Happen

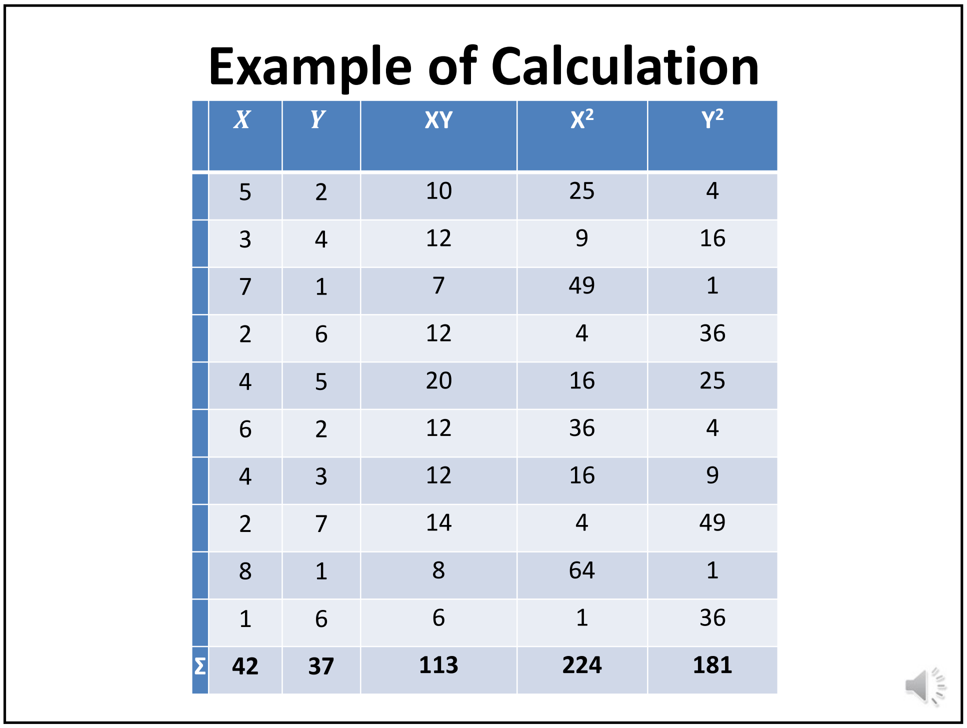
- It's a relationship between variables in general
- Not true for every case – perfect correlation never happens in the social world



Correlation Coefficient Formula

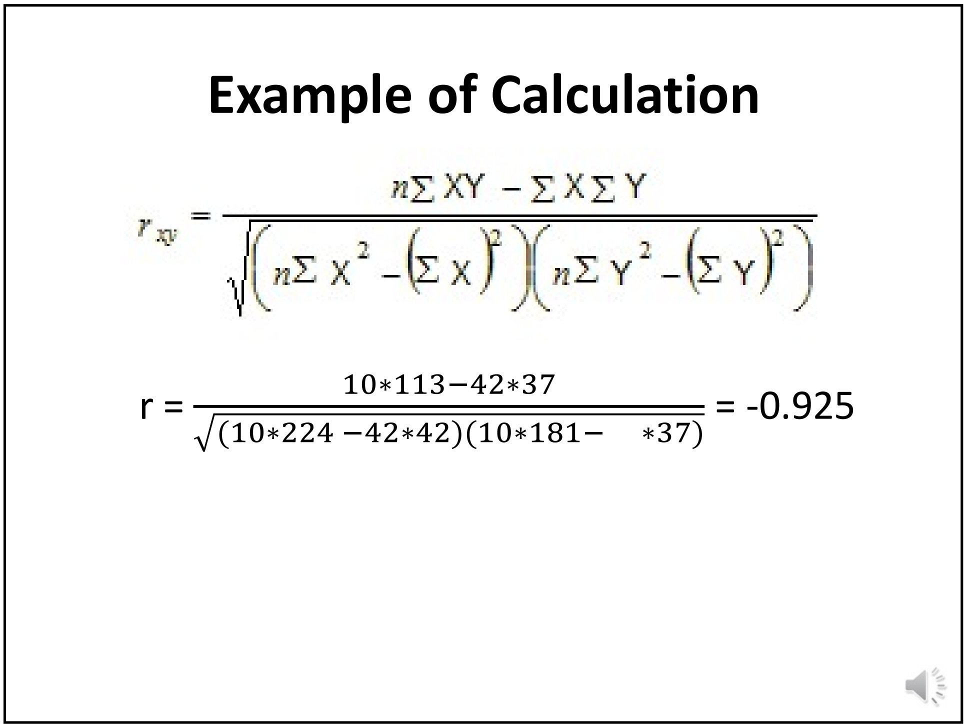
$$r_{xy} = \frac{n\Sigma XY - \Sigma X \Sigma Y}{\sqrt{\left(n\Sigma X^2 - \left(\Sigma X\right)^2\right)\left(n\Sigma Y^2 - \left(\Sigma Y\right)^2\right)}}$$

- r_{XY} – correlation coefficient between X and Y
- n – sample size
- X - individual values of variable X
- Y – individual values of variable Y
- XY – product of the value of X and the corresponding value of Y



Example of Calculation

	X	Y	XY	X^2	Y^2
	5	2	10	25	4
	3	4	12	9	16
	7	1	7	49	1
	2	6	12	4	36
	4	5	20	16	25
	6	2	12	36	4
	4	3	12	16	9
	2	7	14	4	49
	8	1	8	64	1
	1	6	6	1	36
Σ	**42**	**37**	**113**	**224**	**181**



Example of Calculation

$$r_{xy} = \frac{n\sum XY - \sum X \sum Y}{\sqrt{\left(n\sum X^2 - \left(\sum X\right)^2\right)\left(n\sum Y^2 - \left(\sum Y\right)^2\right)}}$$

$$r = \frac{10*113-42*37}{\sqrt{(10*224-42*42)(10*181- \quad *37)}} = -0.925$$

Testing Hypotheses About Correlation

1. State hypotheses:

- H0: ρ = 0
- H1: ρ > 0
- H1: ρ < 0
- H1: ρ ≠ 0

directional

non-directional

2. Select alpha: 0.05, 0.01, .001, .10
3. Test statistic: Correlation coefficient itself
4. Compute the correlation coefficient
5. Use the table to find critical value: Table B4 (df=n-2, alpha, one-tailed vs two-tailed)
6. Compare computed value and critical value
7. State your decision about H0
8. Make a substantive conclusion

Example

We want to find out if the length of marriage is related to marital satisfaction among U.S. adults. For a random sample of 100, we calculate a sample r = -.225. Can we conclude (with 95% confidence) that there is a relationship in the population?

THANK YOU FOR YOUR PATIENCE. THE DOCTOR WILL SEE YOU SHORTLY.
GLASBERGEN

"Remember how crazy we were about each other? Put that down under 'History of Mental Illness'."

"The secret to a long marriage? Never get divorced!"

Example: Step by Step

1. State hypotheses:

- H0: ρ = 0 No relationship between the length of marriage and marital satisfaction in the U.S.
- H1: ρ ≠ 0 → two-tailed test   There is a relationship between the length of marriage and marital satisfaction in the U.S.

2. Select alpha: 0.05

3. Test statistic: Correlation coefficient itself (although the test is based on t-distribution)

4. The correlation coefficient is r = -.225

5. Use the table to find critical value: Table B4 (df=n-2=100-2=98, alpha=.05, two-tailed) → 0.1946

Example: Step by Step

6. Compare computed and critical value: .225 > 0.1946.

7. State your decision about H0: Reject H0 in favor of H1

We report: r = -.225, p < .05

8. Conclusion: Based on the sample of 100 adults, we are 95% sure that the length of marriage has a weak negative relationship to marital satisfaction in the U.S. population (this relationship is statistically significant at .05 level)

Correlation Coefficient in Stata

- Problem: We would like to determine whether, for the U.S. population, there is a relationship between one's level of education and the age when they have their first child.
- H0: There is no relationship between one's level of education and the age when they have their first child.
- H1: There is a relationship between one's level of education and the age when they have their first child. [non-directional → two-tailed]

Correlation Coefficient in Stata

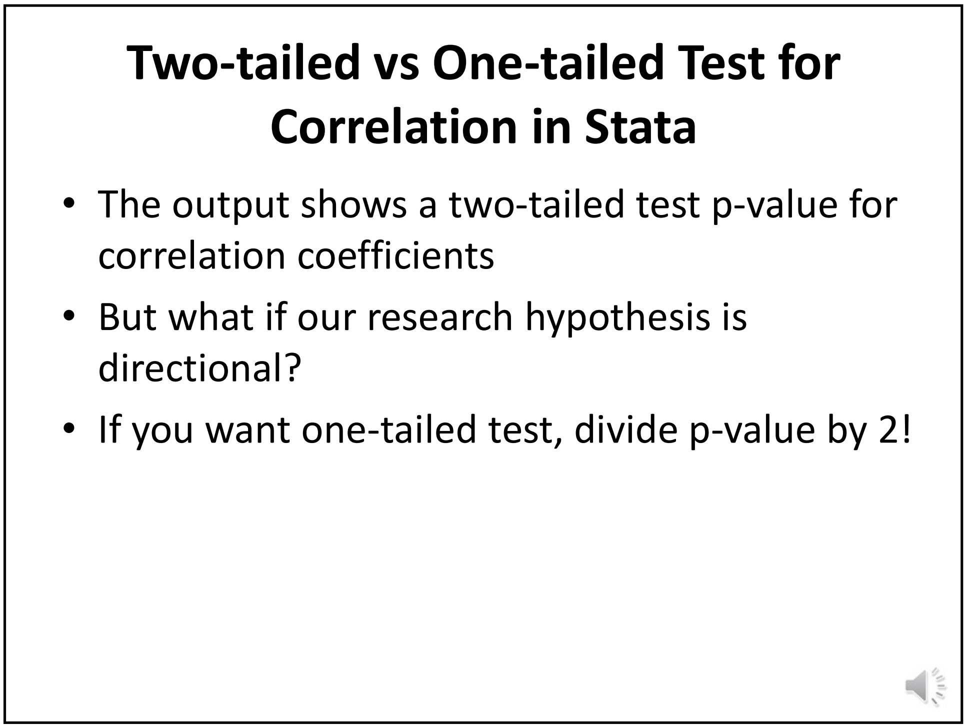
`pwcorr agekdbrn educ, sig`

```
             | agekdbrn     educ
-------------+------------------
    agekdbrn |   1.0000
             |
             |
        educ |   0.3596   1.0000
             |   0.0000
```

Correlation coefficient

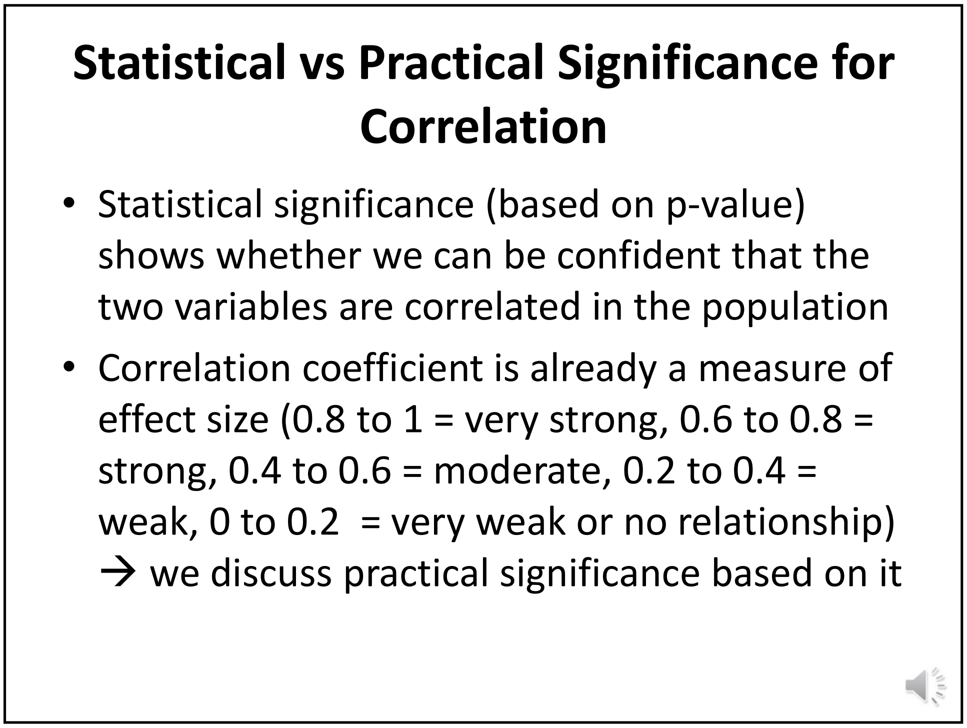
P-value

- Can report as: r = .360, p < .0001 (two-tailed) → reject null
- The positive correlation between education and age at first childbirth is statistically significant at .001 level → we are 99.9% confident this relationship exists in the population



Two-tailed vs One-tailed Test for Correlation in Stata

- The output shows a two-tailed test p-value for correlation coefficients
- But what if our research hypothesis is directional?
- If you want one-tailed test, divide p-value by 2!



Statistical vs Practical Significance for Correlation

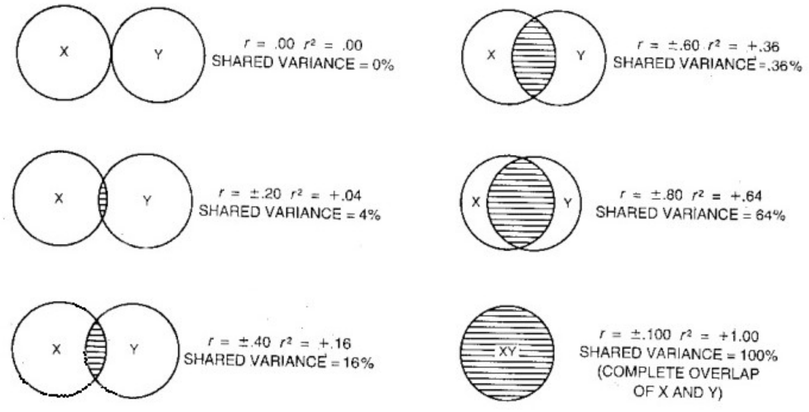
- Statistical significance (based on p-value) shows whether we can be confident that the two variables are correlated in the population
- Correlation coefficient is already a measure of effect size (0.8 to 1 = very strong, 0.6 to 0.8 = strong, 0.4 to 0.6 = moderate, 0.2 to 0.4 = weak, 0 to 0.2 = very weak or no relationship) → we discuss practical significance based on it

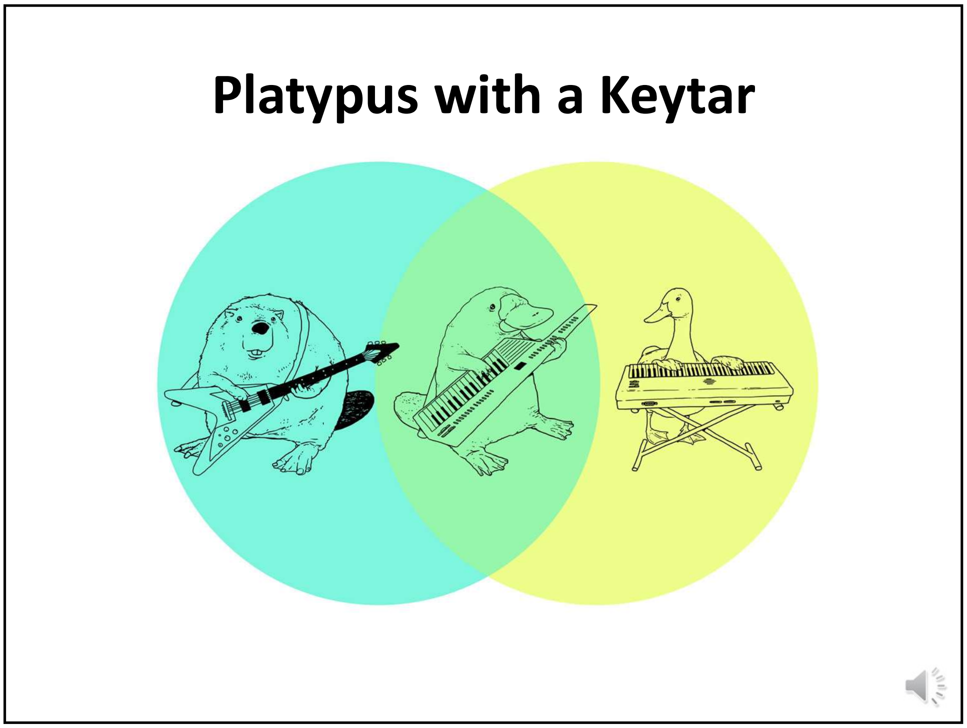
Coefficients of Determination and Alienation

- Coefficient of determination = r^2
  - – proportion of variance shared between X and Y
  - – to find, square the correlation coefficient
- Coefficient of alienation = $1 – r^2$
  - – proportion of variance that is unique, not shared, not explained by the other variable

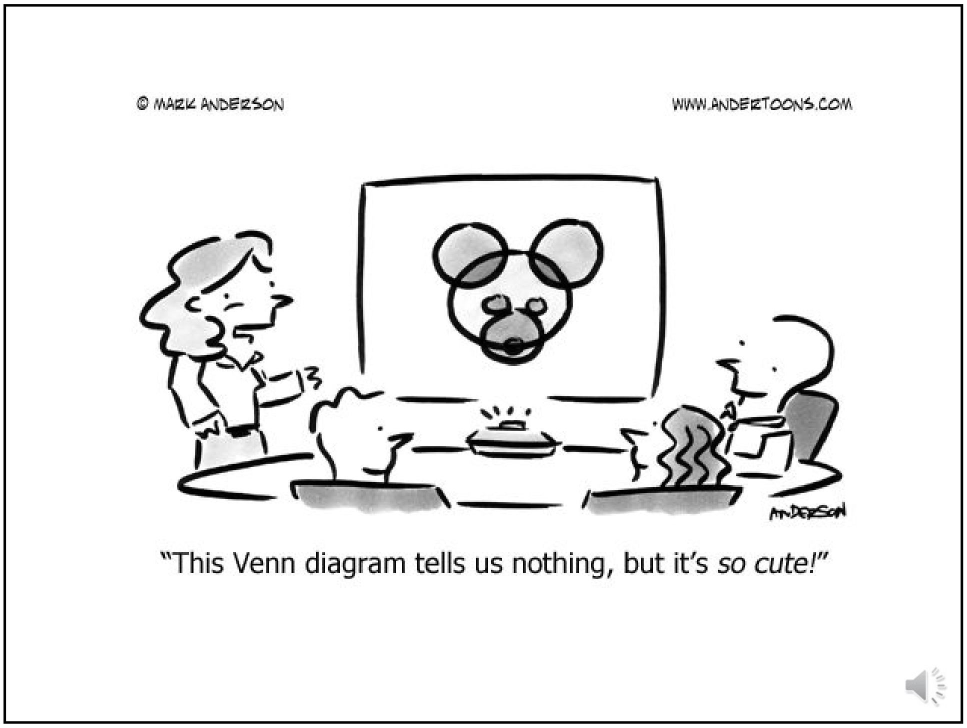
"Tell me about your feelings of alienation."

Correlation, R^2, and Shared Variance





Platypus with a Keytar



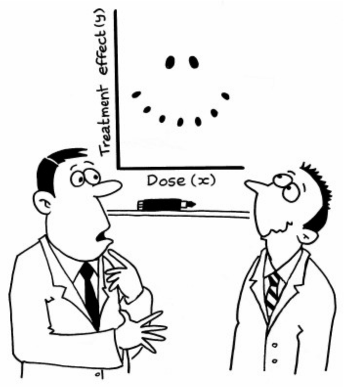
"This Venn diagram tells us nothing, but it's *so cute!*"

Example

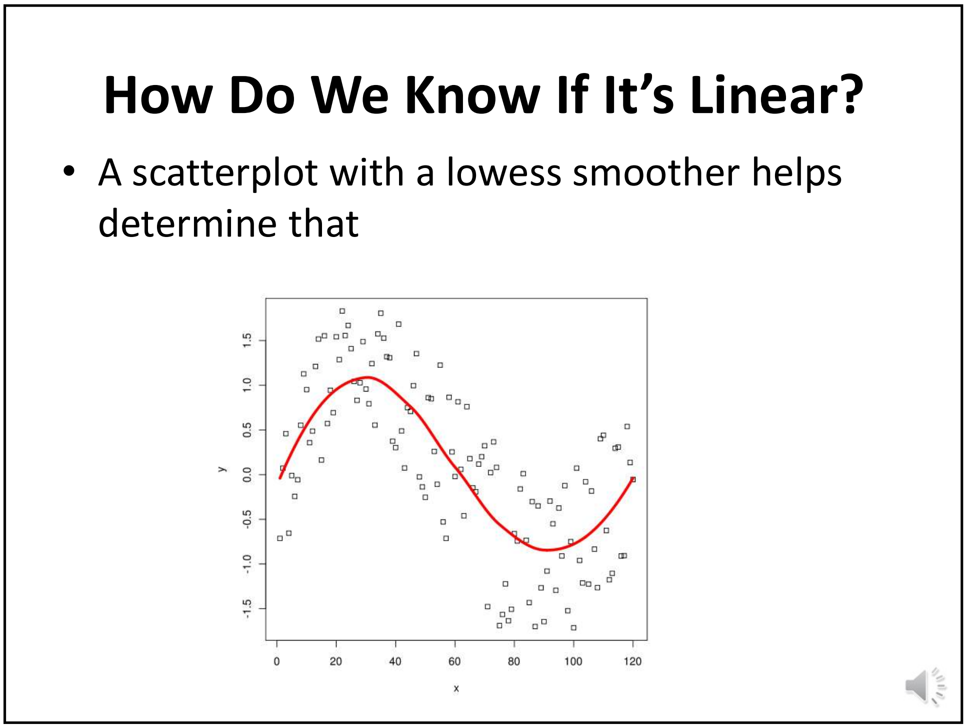
- For our example from hypothesis testing, coefficient of determination = $(-.225)^2$ = .051
- Coefficient of alienation = 1-.051 = .949
- 5.1% of variance is shared, 94.9% is unique

Linear vs. Curvilinear

- Correlation coefficient measures the strength of a <u>linear</u> relationship; if curvilinear, not appropriate

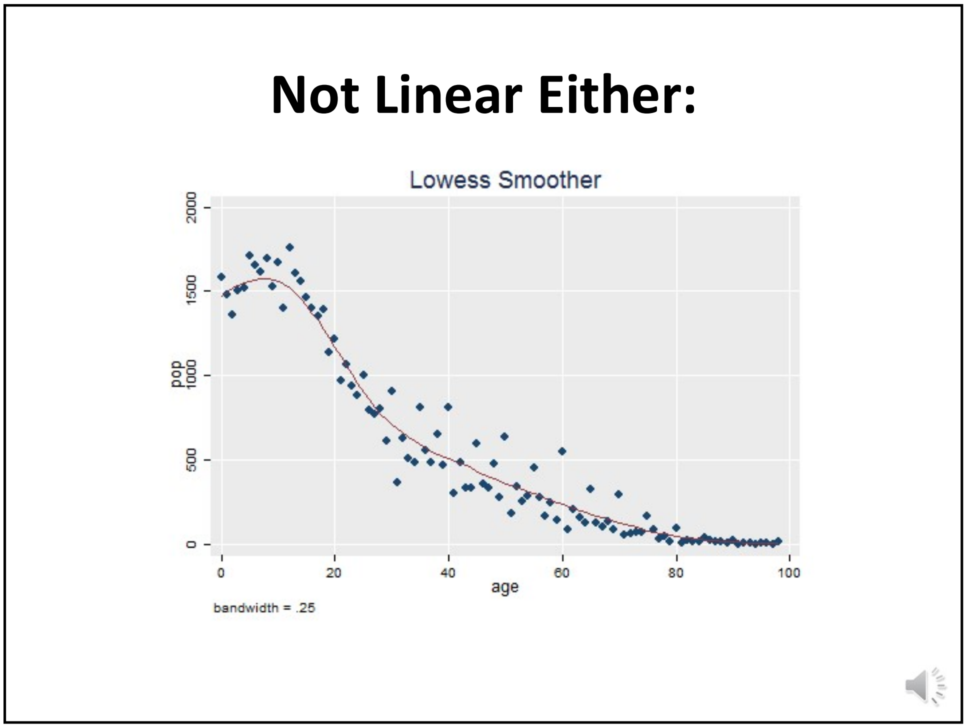


"It's a non-linear pattern with outliers.....but for some reason I'm very happy with the data."



How Do We Know If It’s Linear?

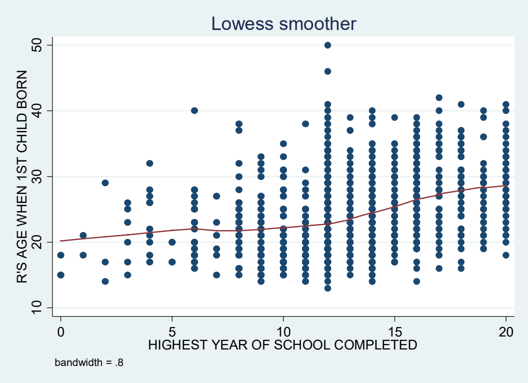
- A scatterplot with a lowess smoother helps determine that



Not Linear Either:

Scatterplot with a Lowess Line in Stata

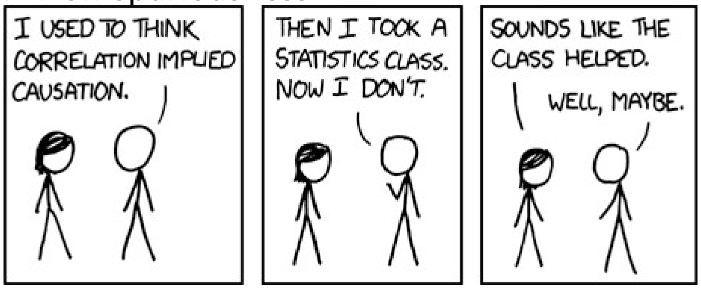
```
lowess agekdbrn educ
```



- LOWESS = locally estimated weighted scatterplot smoothing

Causation Is More Than Correlation

- "Correlation does not imply causation"
- Basic conditions for establishing causality:
  1. Association
  2. Time order
  3. Theory
  4. Non-spuriousness


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

1. Association

That’s where correlation comes in:

2 variables (X and Y) should be related

2. Time Order

One variable (X) should change before the other (Y)

THE WALL STREET JOURNAL.
Cable

"Do you think all these film crews brought on global warming or did global warming bring on all these film crews?"

THE FAMILY CIRCUS
8-5

"I wish they didn't turn on that seatbelt sign so much! Every time they do, it gets bumpy."

3. Theory

You should have a logical explanation of mechanisms

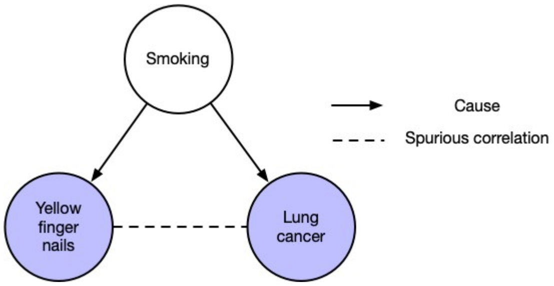
SALES
SHAVED HEADS
WE FOUND THIS CORRELATION IN THE DATA. EVERYONE TAKE A RAZOR.
TOM FISH BURNE

© marketoonist.com

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

4. Non-spuriousness

- You have to ensure that correlation is not due to variation in a third variable
- To rule out other explanations, we can control for such a third variable: remove its effect by holding it constant


Smoking
Cause
Spurious correlation
Yellow finger nails
Lung cancer