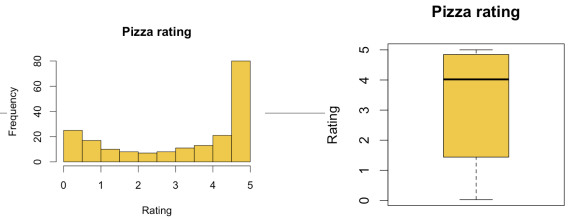
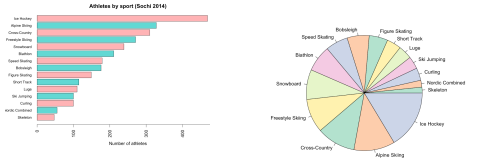
Linear models

Steven Moran, Marco Maiolini &
Alena Witzlack-Makarevich

Overview of visualization and summary techniques

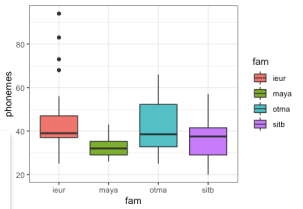


- **one quantitative variable**
  histogram, box plot
  five-number-summary (min., max., median, mean, IQR), SD

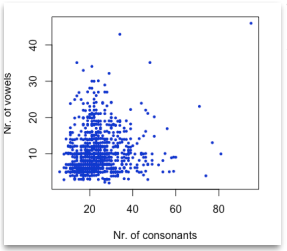


- **one categorical variable**
  bar plot, pie chart + frequency distribution

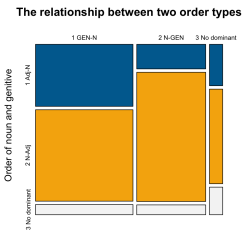
- **one categorical and one quantitative variable**
  parallel box plots



- **two quantitative variables**
  scatter plot, correlation coefficients *r*,
  regression line



- **two categorical variables**
  mosaic plot, grouped/stacked bar plots



Modeling: The big picture

- Scientists are interested in discovering/understanding something about a real-world phenomena
- Three types of goals:
  - **describe**
  - **explain**
  - **predict**: not necessarily some *future* events, rather predicting *new* observations which are not part of the sample

- Different disciplines prioritize different goals
- One generic tool for all three goals: **statistical modeling**

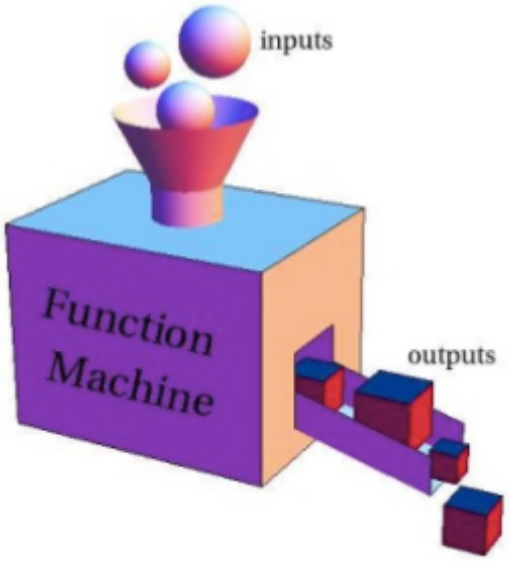
The big picture

- Statistical modeling is an attempt to describe some part of the real world **in mathematical terms**

- The relevant mathematical concept is the one of ***function***
  - Functions can show (describe) the relationship between two or more variables
  - These variables can represent e.g.
    - weight and height
    - number of vowels and the number of consonants
    - the frequency of a word and the speed of its pronunciation

Functions

- A function is any rule that assigns (or corresponds) to each element in one set precisely one element in another set
  - Aka *mapping* or *transformation*

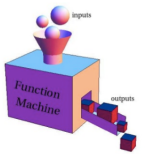


Input x	Function	Output y
1		2
2		4
3		6
4		8
5		10

Functions

In this example the function has the form:

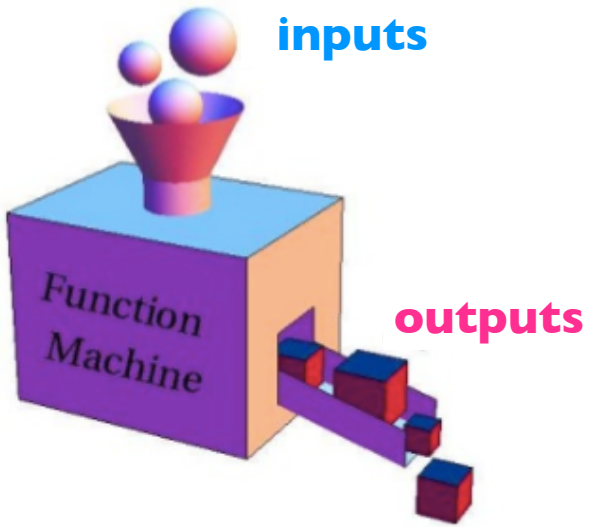
y = 2 * x



Input x	Function *y = 2 * x*	Output y
1	y = 2 * 1	2
2	y = 2 * 2	4
3	y = 2 * 3	6
4	y = 2 * 4	8
5	y = 2 * 5	10

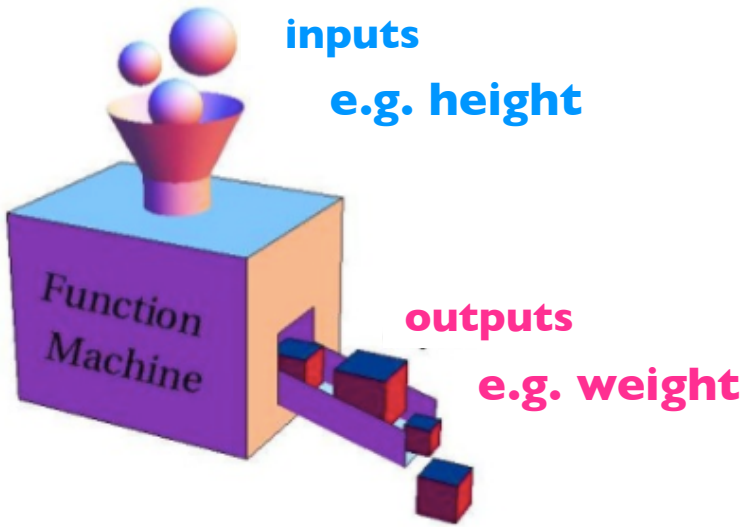
The big picture

- Statistical modeling is an attempt to describe some part of the real world **in mathematical terms**
- The relevant mathematical concept is the one of ***function***



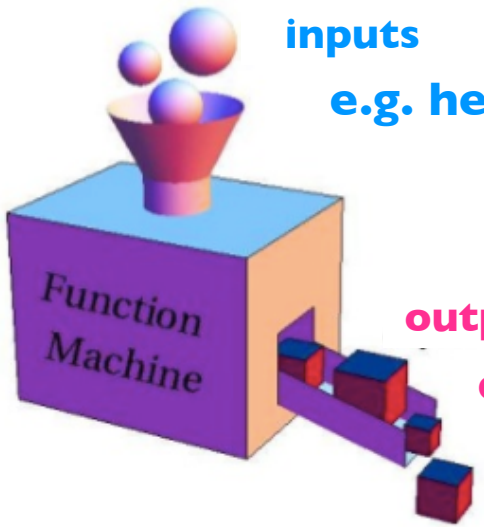
The big picture

- Statistical modeling is an attempt to describe some part of the real world **in mathematical terms**
- The relevant mathematical concept is the one of ***function***



The big picture

- Statistical modeling is an attempt to describe some part of the real world **in mathematical terms**
- The relevant mathematical concept is the one of ***function***



inputs

e.g. height

independent/ predictor/ explanatory *variable*

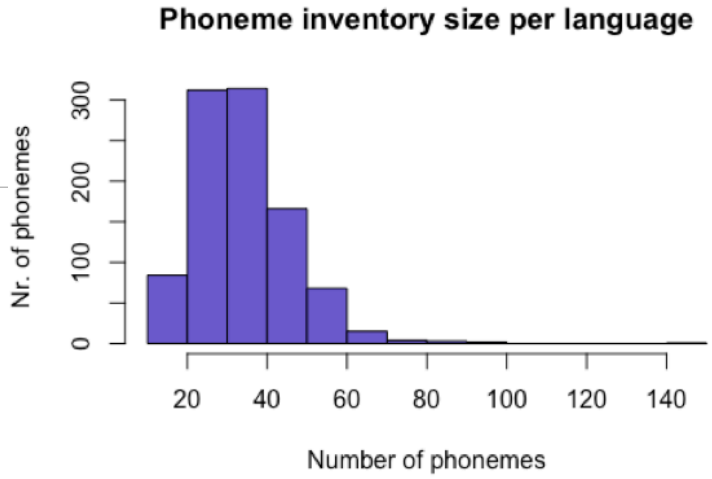
outputs

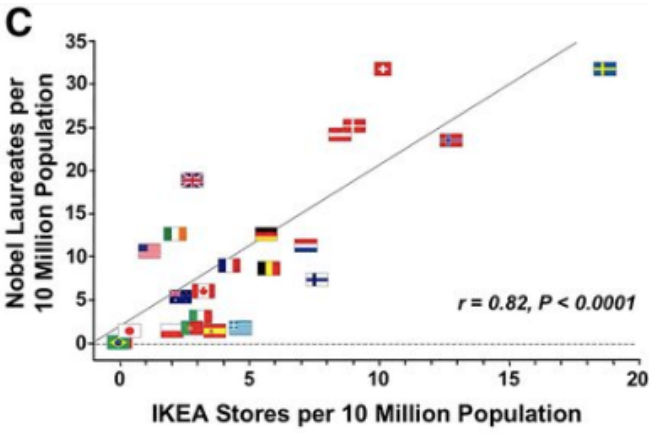
e.g. weight

dependent/ response *variable*

Bivariate distributions

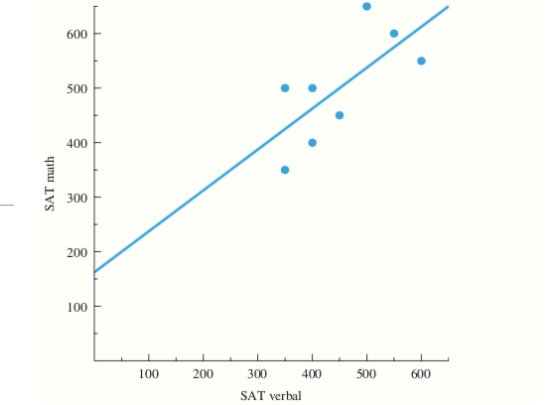
- ***Univariate*** distributions involve one variable
- In addition to considering individual variable, we can look at the association (or relation) between two variables
  ➙ the question of ***correlation***
- A correlation is exactly what its name suggests:
  a **co-relation** between two variables
- A bivariate distribution may show positive correlation, negative correlation, or zero correlation
- Statistically, the strength of the relation between two variables is indicated with the **correlation coefficient r**.
- Visually the relation between two variables is represented as a **scatter plot**

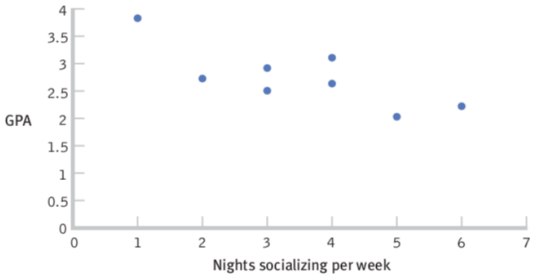




Bivariate distributions

- **Positive** correlation between two variables:
  - high measurements on one variable tend to be associated with high measurements on the other variable, and low measurements on one variable with low measurements on the other
    - e.g. tall fathers tend to have sons who grow up to be tall men; short fathers tend to have sons who grow up to be short men
- **Negative correlation**: increases in one variable are accompanied by decreases in the other variable (an inverse relationship)
- **Zero correlation:** no *linear* relationship between two variables
  High and low scores on the two variables are not associated in any predictable manner





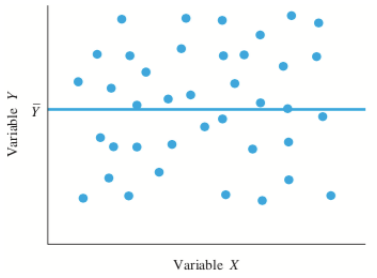
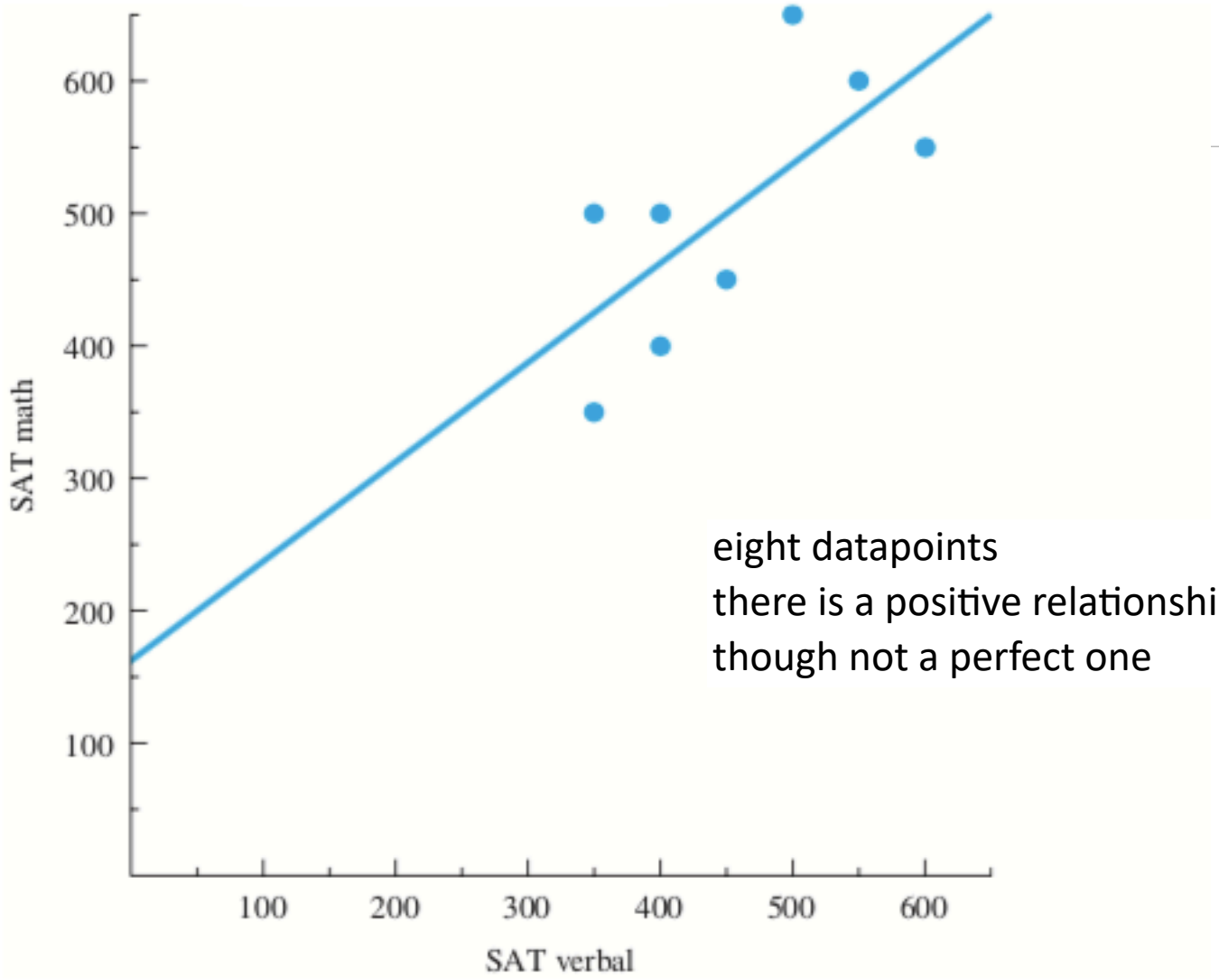
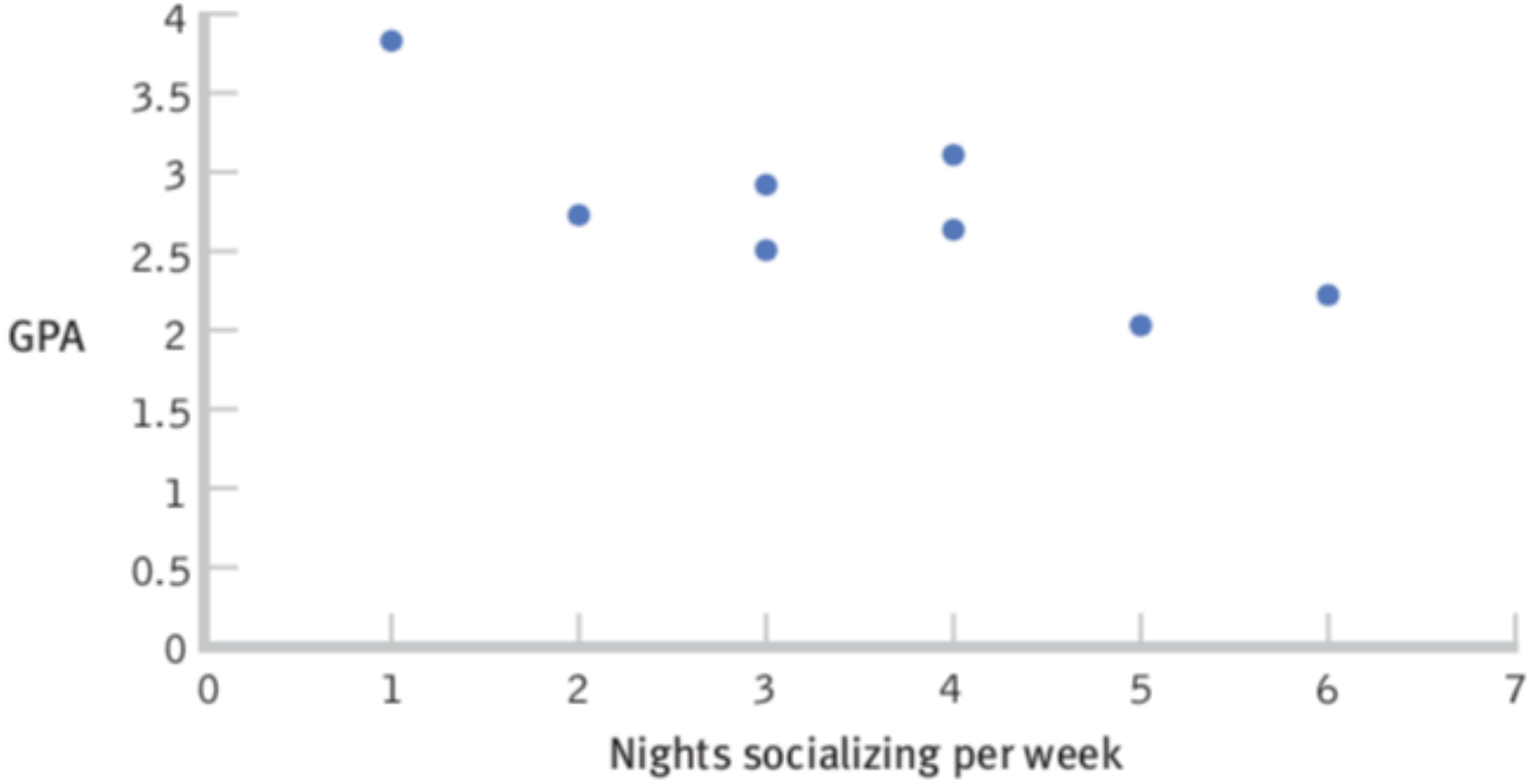


FIGURE 5.5 Scatterplot and regression line for a zero correlation



eight datapoints
there is a positive relationship,
though not a perfect one

These data points depict a negative correlation between nights socializing per week and GPA. Those who go out more tend to have lower GPAs, whereas those who go out less tend to have higher GPAs.



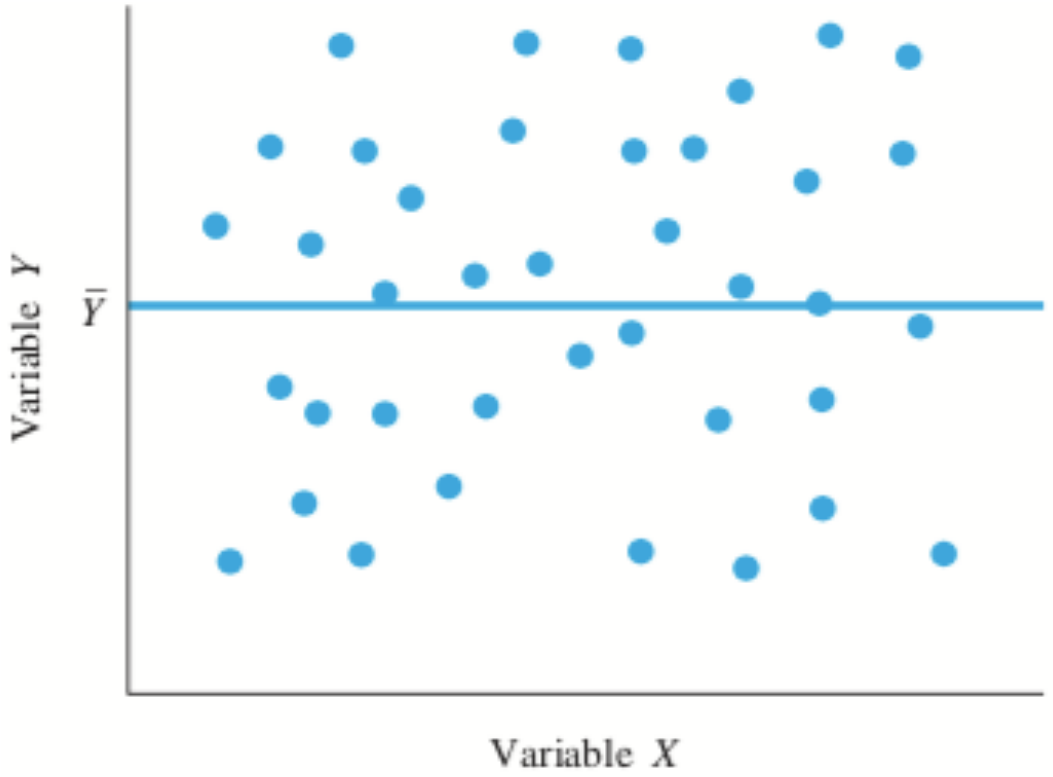


FIGURE 5.5 Scatterplot and regression line for a zero correlation

Bivariate relationship

When we have two numerical variables, we can distinguish:

Bivariate relationship

When we have two numerical variables, we can distinguish:

- *Response variable:* dependent variable, as known as Y

Bivariate relationship

When we have two numerical variables, we can distinguish:

- *Response variable:* dependent variable, as known as Y
- *Explanatory variable:* independent variable, as known as X

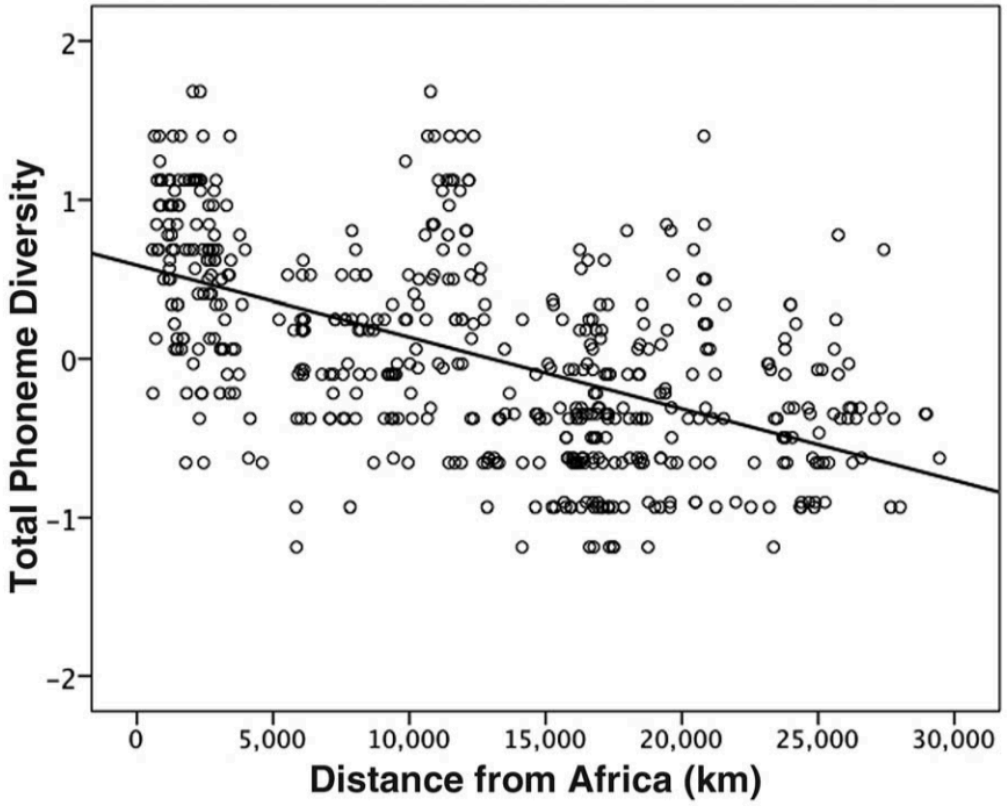
Bivariate relationship

When we have two numerical variables, we can distinguish:

- *Response variable:* dependent variable, as known as Y
- *Explanatory variable:* independent variable, as known as X

Scatter plot

- Visually the relation between two variables (X and Y) can be represented as a **scatter plot**



Response (dependent) and explanatory (independent) variable

- Warming Reducing Rice Yields (2004, *PNAS*) reported that
  - "an average daily temperature increase of 1°C resulted in a 10% reduction in the rice crop."
- In the US, corn and soybean yields were found to reduce in a manner similar to that of rice in the Philippines.
  - a. A positive or a negative correlation?
  - b. Which of the variables, rice yield or temperature, is the **explanatory** variable?

Example

Bentz C, Verkerk A, Kiela D, Hill F, Buttery P (2015) Adaptive Communication: Languages with More Non-Native Speakers Tend to Have Fewer Word Forms. PLoS ONE 10(6): e0128254. doi:10.1371 journal.pone.0128254

Abstract

Explaining the diversity of languages across the world is one of the central aims of typological, historical, and evolutionary linguistics. We consider the effect of language contact – the number of non-native speakers a language has – on the way languages change and evolve. By analysing hundreds of languages … we show that languages with greater levels of contact typically employ fewer word forms to encode the same information content (a property we refer to as *lexical diversity*). Based on three types of statistical analyses, we demonstrate that this variance can in part be explained by the impact of non-native speakers on information encoding strategies. …

What is the independent/predictor variable?

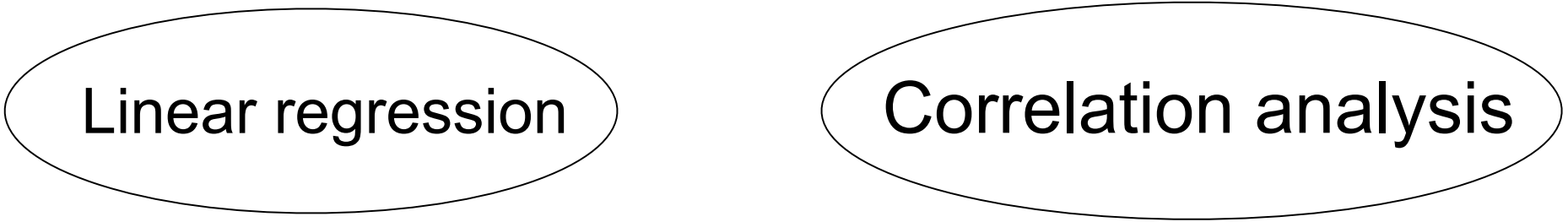
- What is the dependent/response variable?
- Which one goes where on a scatterplot?
- A positive or a negative correlation?

Relationship between X and Y

Techniques based on fitting a straight line to the data:

Relationship between X and Y

Techniques based on fitting a straight line to the data:



Linear regression

Example

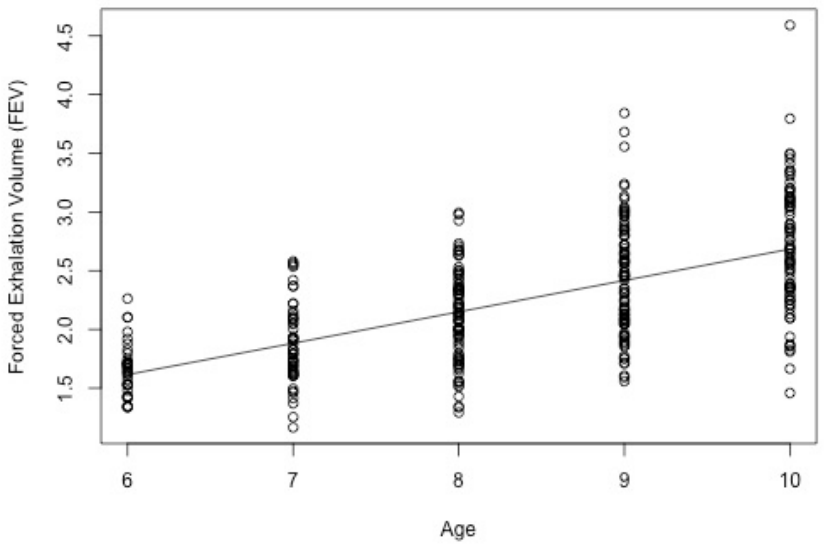
You want to test Lung Function in children. You measure the forced exhalation volume (FEV), the measure of how much air somebody can forcibly exhale from their lungs, from 6 to 10 year old children. You survey 345 children.

The data here are a part of dataset given in Kahn, Micheal (2005), “An Exhalent Problem for Teaching Statistics”, The Journal of Statistical Education, 13(2).

Linear regression

Example

You want to test Lung Function in children. You measure the forced exhalation volume (FEV), the measure of how much air somebody can forcibly exhale from their lungs, from 6 to 10 year old children. You survey 345 children.

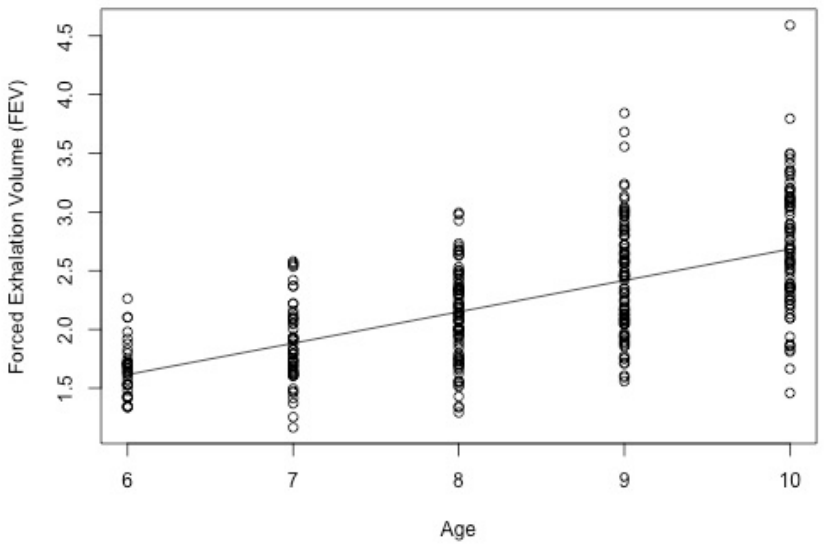


The data here are a part of dataset given in Kahn, Micheal (2005), “An Exhalent Problem for Teaching Statistics”, The Journal of Statistical Education, 13(2).

Linear regression

Example

You want to test Lung Function in children. You measure the forced exhalation volume (FEV), the measure of how much air somebody can forcibly exhale from their lungs, from 6 to 10 year old children. You survey 345 children.



The scatter plot suggests a definite age-relationship, with larger X tending to be associated with bigger values of Y

The data here are a part of dataset given in Kahn, Micheal (2005), “An Exhalent Problem for Teaching Statistics”, The Journal of Statistical Education, 13(2).

Correlation analysis

Example

You investigate whether standardized scores from high school (SAT) are related to academic grades in college (GPA). You predict that there's a positive correlation: higher SAT scores are associated with higher college GPAs while lower SAT scores are associated with lower college GPAs.

You gather a sample of 5000 college graduates and survey them on their high school SAT scores and college GPAs.

Correlation analysis

Example

You investigate whether standardized scores from high school (SAT) are related to academic grades in college (GPA). You predict that there's a positive correlation: higher SAT scores are associated with higher college GPAs while lower SAT scores are associated with lower college GPAs.

You gather a sample of 5000 college graduates and survey them on their high school SAT scores and college GPAs.

Correlation analysis

Example

You investigate whether standardized scores from high school (SAT) are related to academic grades in college (GPA). You predict that there's a positive correlation: higher SAT scores are associated with higher college GPAs while lower SAT scores are associated with lower college GPAs.

You gather a sample of 5000 college graduates and survey them on their high school SAT scores and college GPAs.

Correlation coefficient = 0.58

Correlation analysis

Example

You investigate whether standardized scores from high school (SAT) are related to academic grades in college (GPA). You predict that there's a positive correlation: higher SAT scores are associated with higher college GPAs while lower SAT scores are associated with lower college GPAs.

You gather a sample of 5000 college graduates and survey them on their high school SAT scores and college GPAs.

Correlation coefficient = 0.58

The scatter plot seems to confirm our prediction, with higher SAT scores associated with higher GPA values.

The Correlation Coefficient

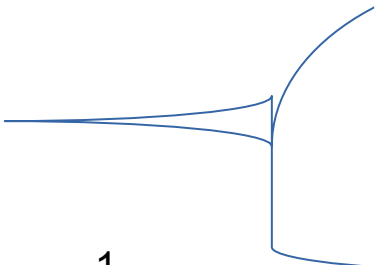
The Correlation Coefficient measures the strength of linear association between the two variables.

The Correlation Coefficient

The Correlation Coefficient measures the strength of linear association between the two variables.

+1

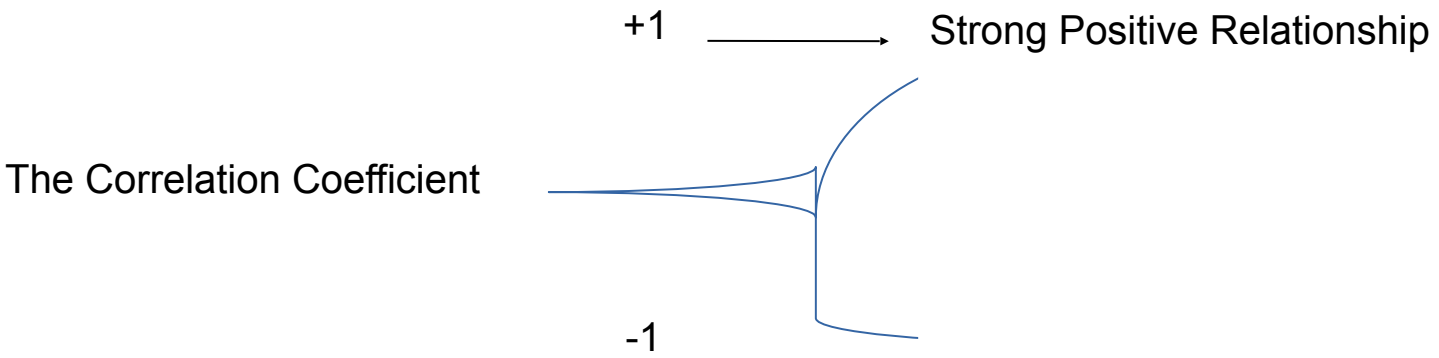
The Correlation Coefficient



-1

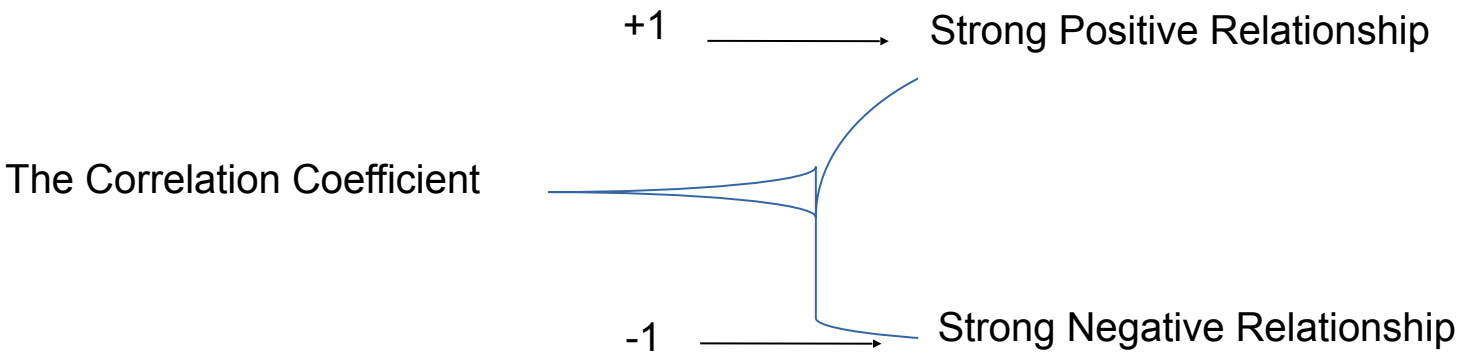
The Correlation Coefficient

The Correlation Coefficient measure the strength of linear association between the two variables.



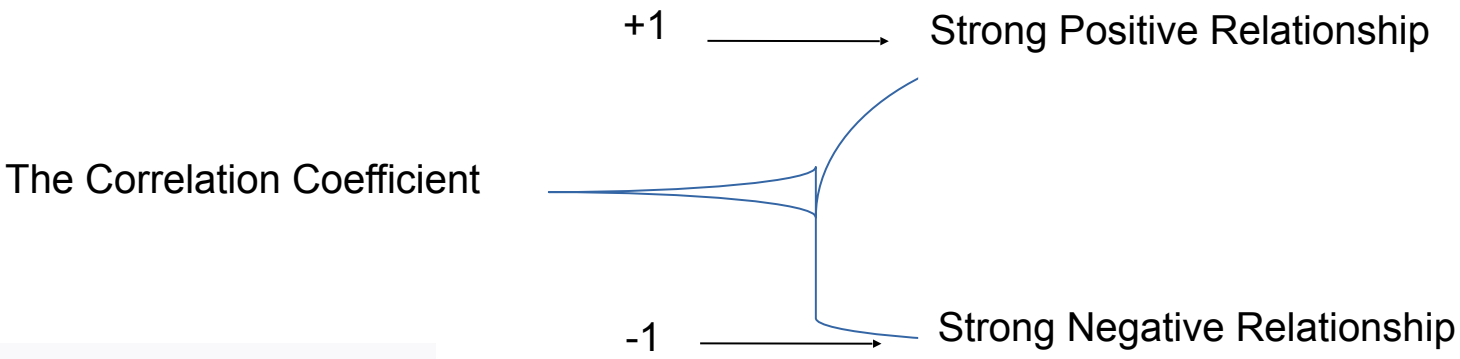
The Correlation Coefficient

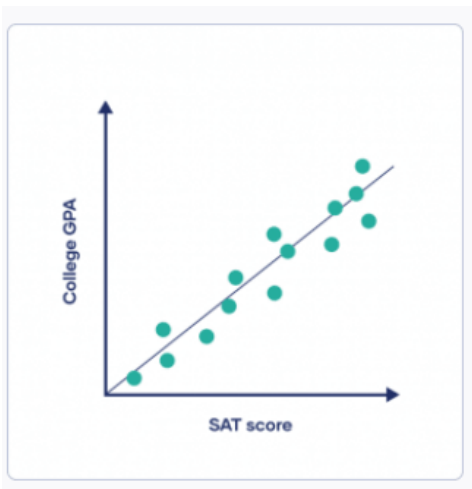
The Correlation Coefficient measure the strength of linear association between the two variables.



The Correlation Coefficient

The Correlation Coefficient measure the strength of linear association between the two variables.



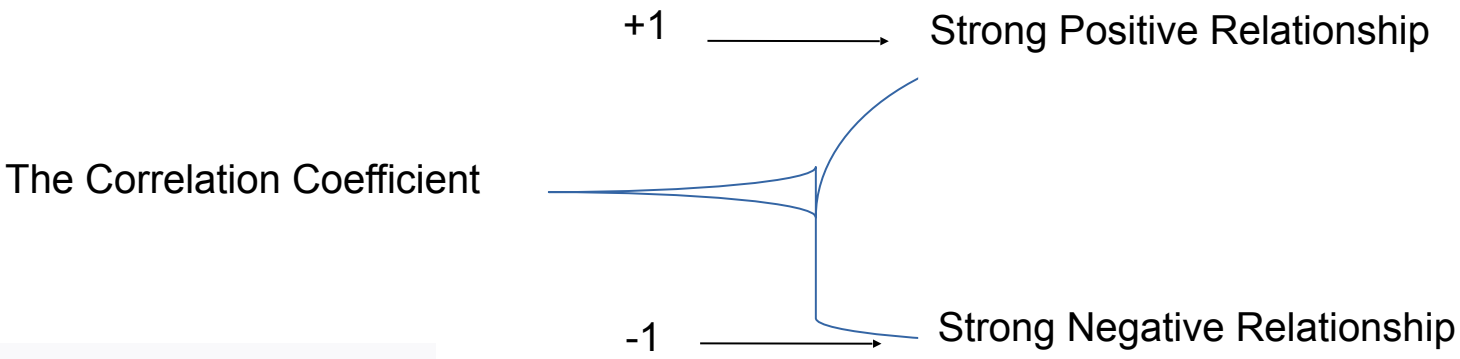


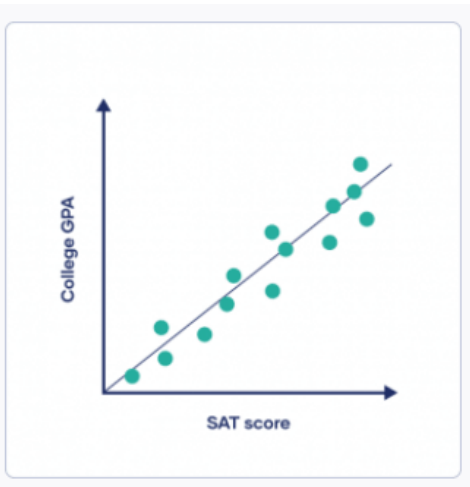
The Pearson's r correlation test:

- Variables are quantitative
- Variables normally distributed

The Correlation Coefficient

The Correlation Coefficient measure the strength of linear association between the two variables.





The Pearson's r correlation test:

- Variables are quantitative
- Variables normally distributed

In R you see if two variables are correlated by:

cor(x, y)

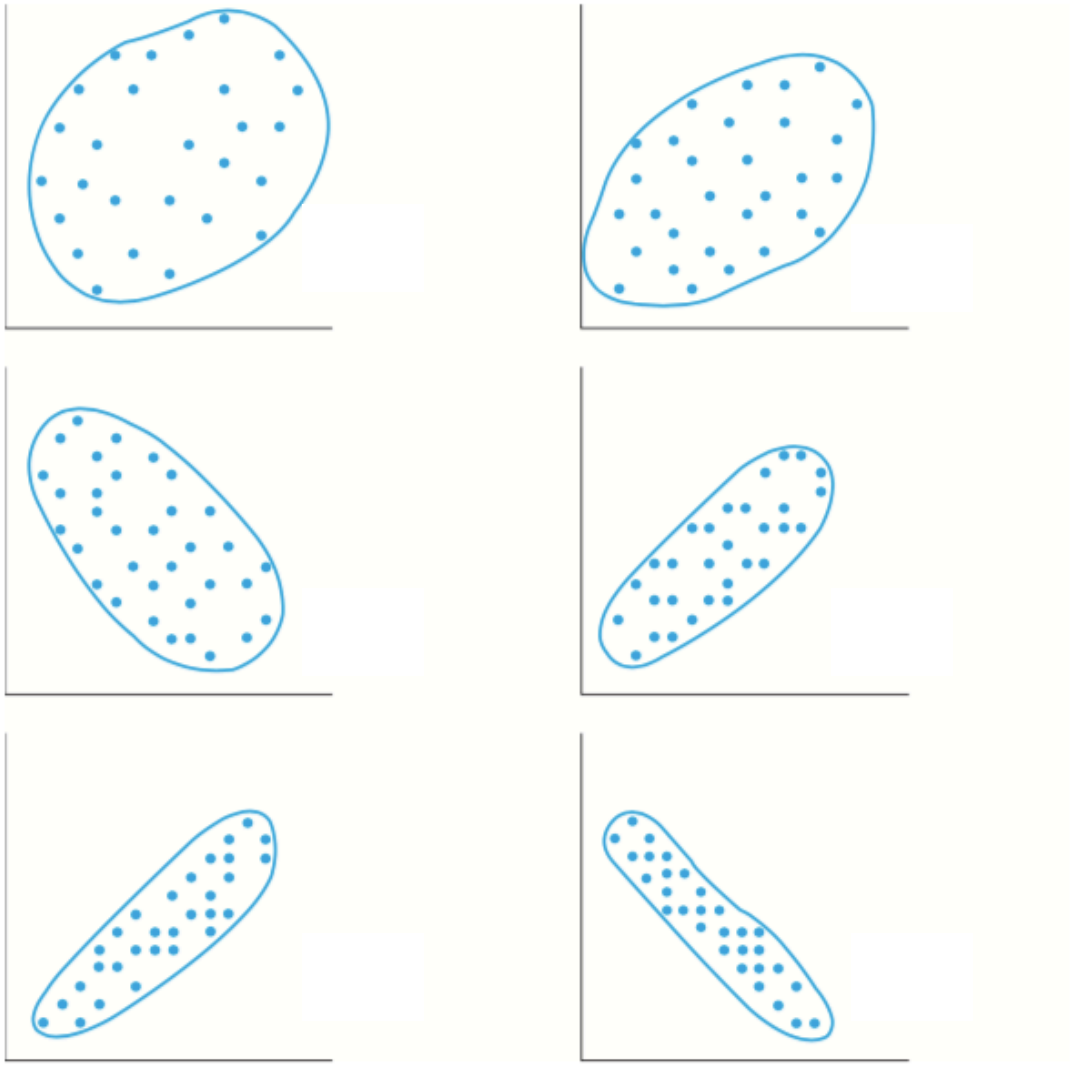


FIGURE 5.6 Scatterplots of data in which r = .20, .40, −.60, .80, .90, and −.95

$r = 0.4$

$r = -0.6$

$r = 0.8$

$r = -0.95$

$r = 0.2$

$r = 0.9$

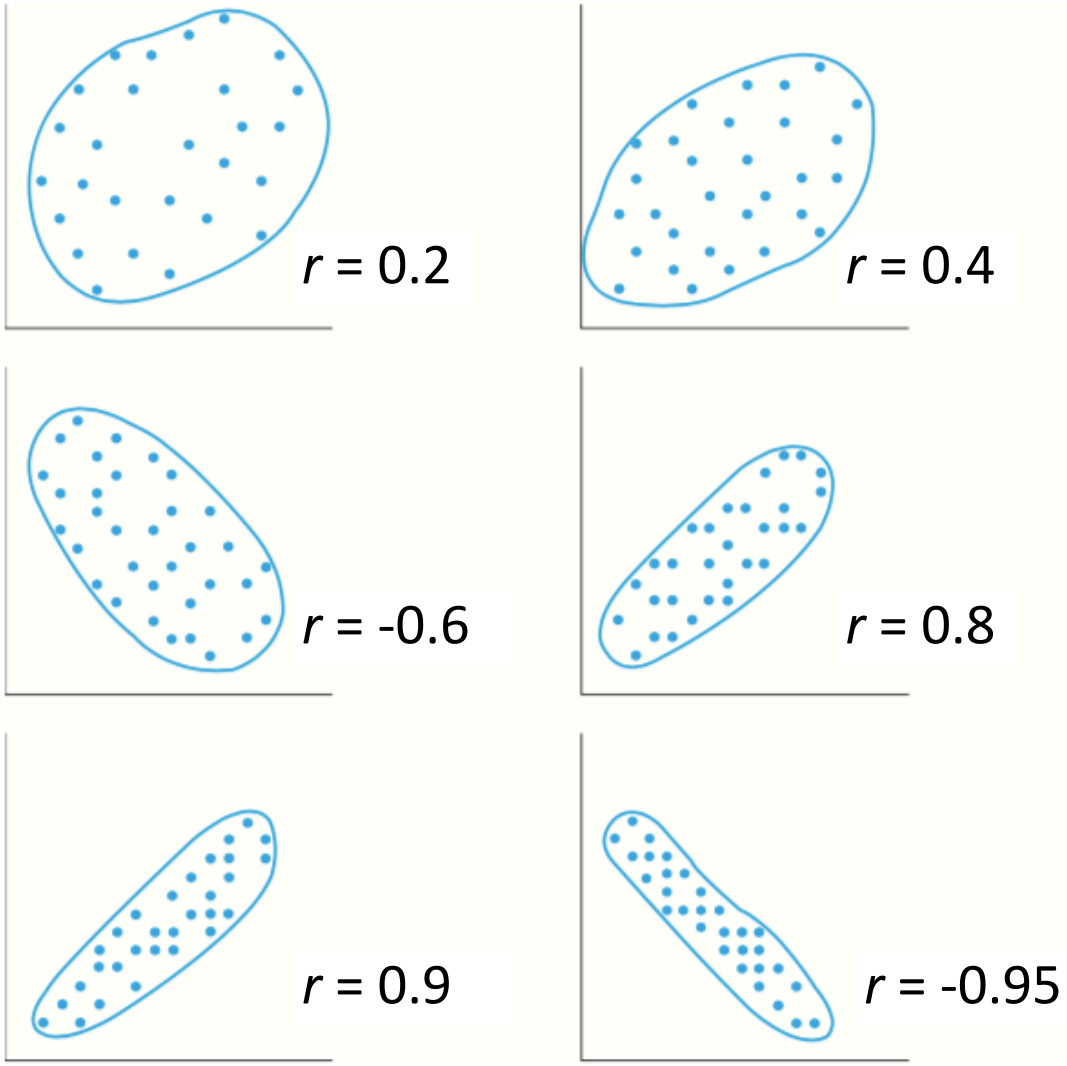


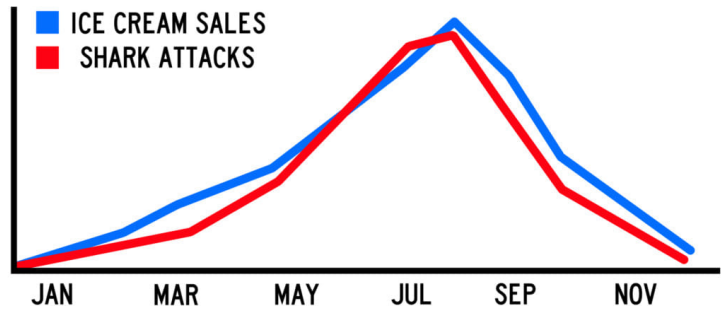
FIGURE 5.6 Scatterplots of data in which r = .20, .40, −.60, .80, .90, and −.95

The Correlation Coefficient

A strong correlation between two variables does not indicate any causal connection between them. It is important to remember this concept when interpreting correlation.

The Correlation Coefficient

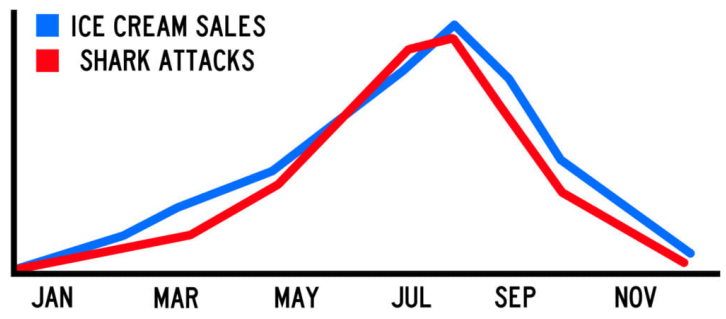
A strong correlation between two variables **does not indicate any causal connection** between them. It is important to remember this concept when interpreting correlation.



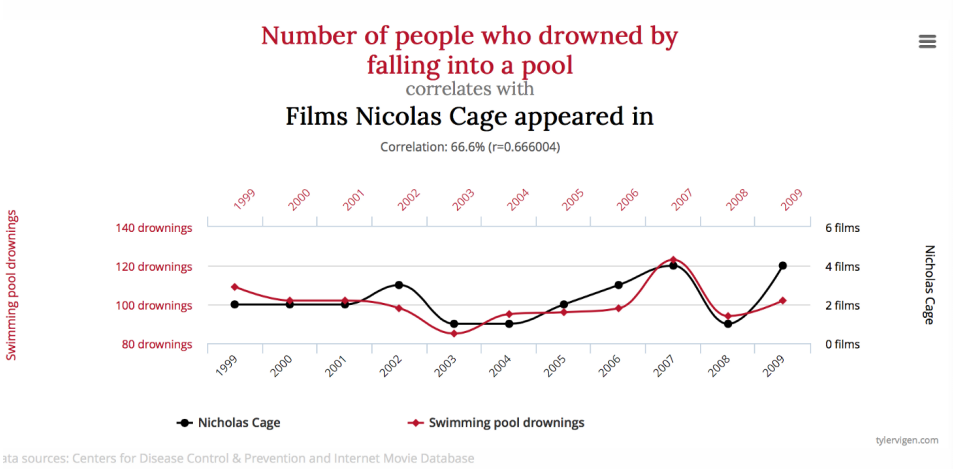
Both ice cream sales and shark attacks increase when the weather is hot and sunny, but they are not caused by each other (they are caused by good weather, with lots of people at the beach, both eating ice cream and having a swim in the sea)

The Correlation Coefficient

A strong correlation between two variables **does not indicate any causal connection** between them. It is important to remember this concept when interpreting correlation.



Both ice cream sales and shark attacks increase when the weather is hot and sunny, but they are not caused by each other (they are caused by good weather, with lots of people at the beach, both eating ice cream and having a swim in the sea)



The Correlation Coefficient

A strong correlation between two variables **does not indicate any causal connection** between them. It is important to remember this concept when interpreting correlation.

The Correlation Coefficient

A strong correlation between two variables **does not indicate any causal connection** between them. It is important to remember this concept when interpreting correlation.

The cat didn’t crush the awning

Correlation is not causation

- Be aware! A high correlation coefficient does not give you the kind of evidence that allows you to make ***cause-and-effect statements***. Therefore, don't do it! Ever.
- Of course, if you have a sizable correlation coefficient, it may be the result of a cause-and-effect relationship between the two variables.
- What is required **to establish a cause- and-effect relationship** is data from e.g., **controlled experiments or some independently attested causal mechanisms, not correlational data**.
- A sizable correlation is a necessary but not a sufficient condition for establishing causality.

https://en.wikipedia.org/wiki/Spurious_relationship

Spurious correlation

- A spurious relationship or spurious correlation is a mathematical relationship in which two or more events or variables are ***not causally*** related to each other, yet it may be wrongly inferred that they are due to either ***coincidence*** or the presence of a certain third, ***unseen factor*** (referred to as a “confounding factor”, or “lurking variable”)

https://www.dailymail.co.uk/health/article-2262513/Drink-milk--win-Nobel-prize-Nations-consume-white-stuff-laureates.html

Mail Online health

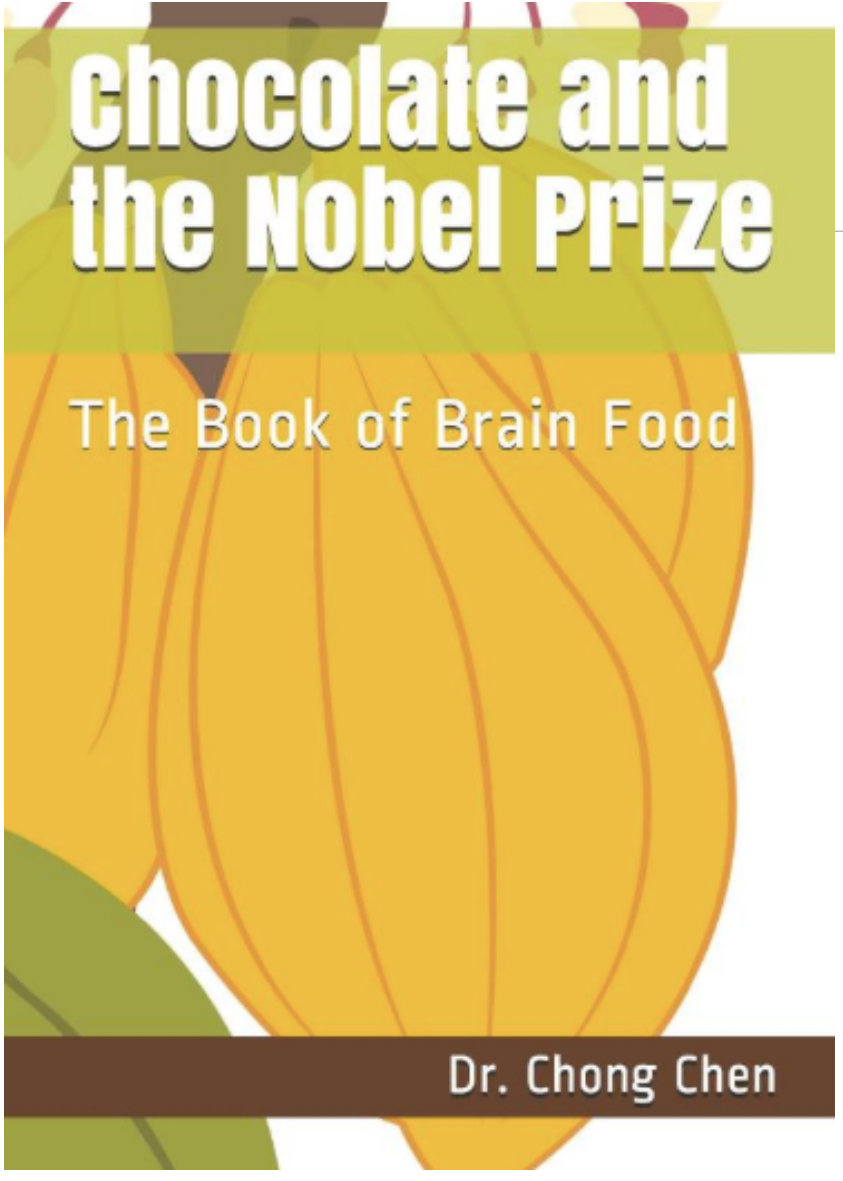
Home | News | U.S. | Sport | TV&Showbiz | Australia | Femail | Health | Science | Money | Video | Travel | DailyMailTV

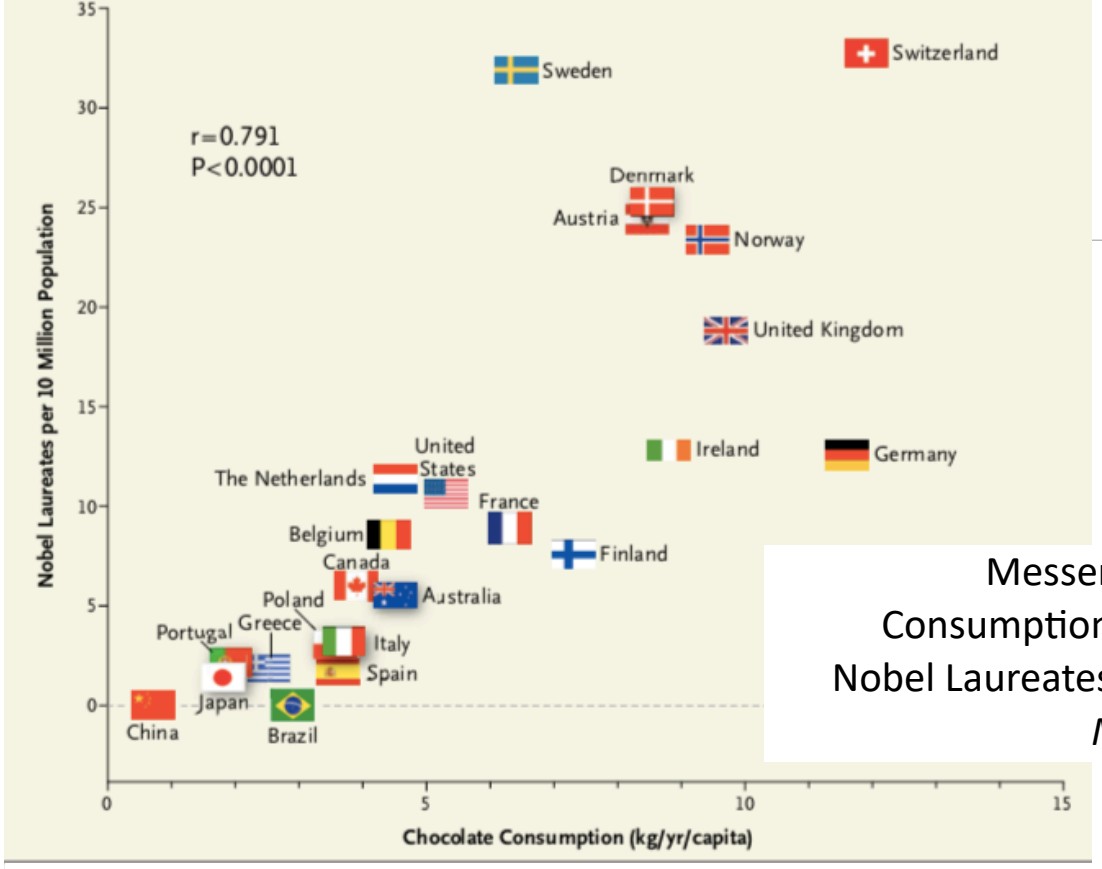
Latest Headlines | *Health* | Health Directory | Discounts Login

Drink more milk - you could win a Nobel prize! Nations that consume more of the white stuff have more laureates

- Researchers found that nations that consume a lot of milk and milk products tend to have a lot of Nobel laureates
- Sweden had the most Nobel laureates and consumes most milk per head
- China had the lowest number of laureates and of milk consumption

"Researchers found that nations that consume a lot of milk and milk products tend to have a lot of Nobel laureates"


Chocolate and the Nobel Prize
The Book of Brain Food
Dr. Chong Chen

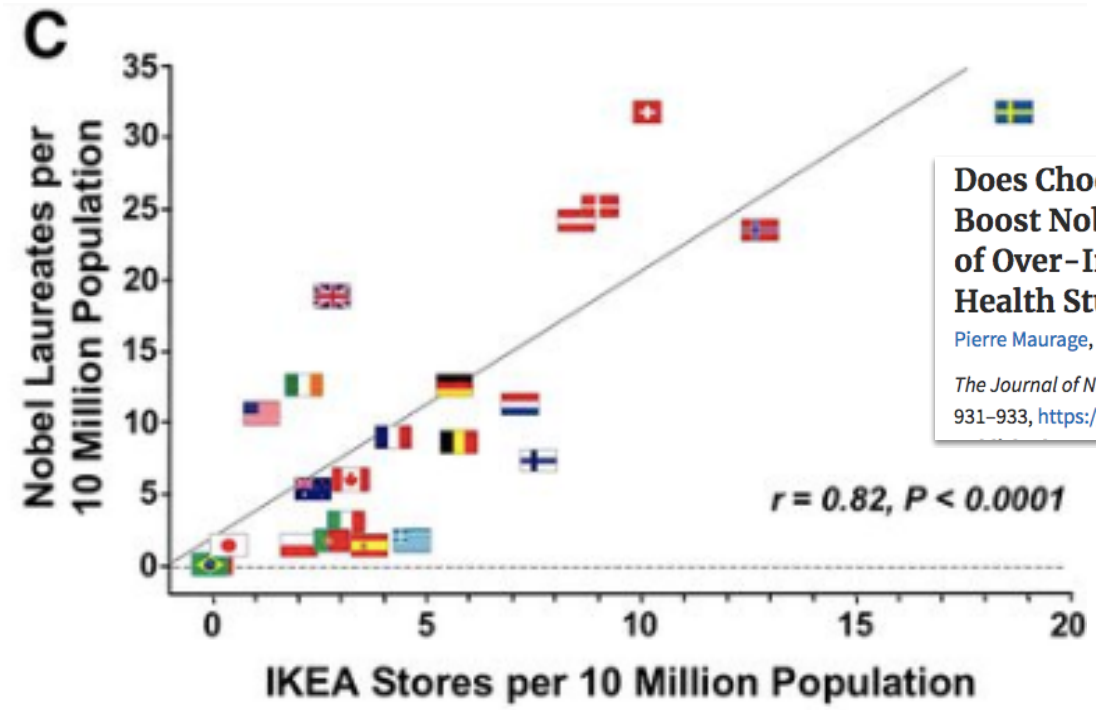


Messerli, F. H. (2012). Chocolate Consumption, Cognitive Function, and Nobel Laureates. *New England Journal of Medicine*, 367, 1562-1564

Dietary flavonoids … improve cognitive function. … Since chocolate consumption could hypothetically improve cognitive function … in whole populations, I wondered whether there would be **a correlation between a country's level of chocolate consumption and its population's cognitive function**.

Conclusions: It remains to be determined whether the consumption of chocolate is the underlying mechanism for the observed association with improved cognitive function.

We found an incredibly high correlation between the number of IKEA furniture stores and Nobel laureates (***r*= 0.82**; P< 0.0001), **although we could not come up with any mutual causal relationship** - and we doubt that … the need to understand and apply IKEA's furniture assembly instructions improves cognitive functioning at the population level.



Does Chocolate Consumption Really Boost Nobel Award Chances? The Peril of Over-Interpreting Correlations in Health Studies FREE

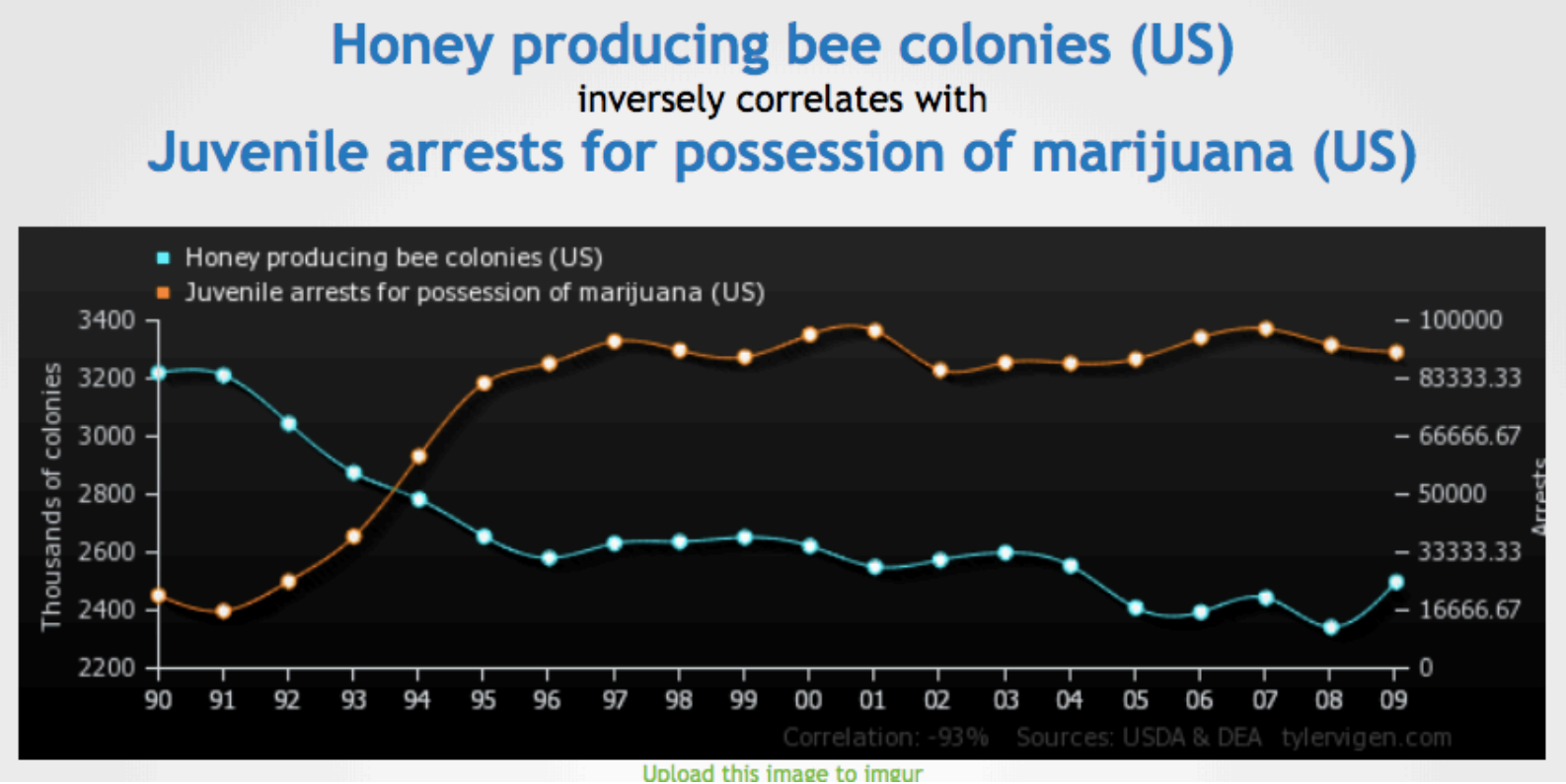
Pierre Maurage, Alexandre Heeren, Mauro Pesenti

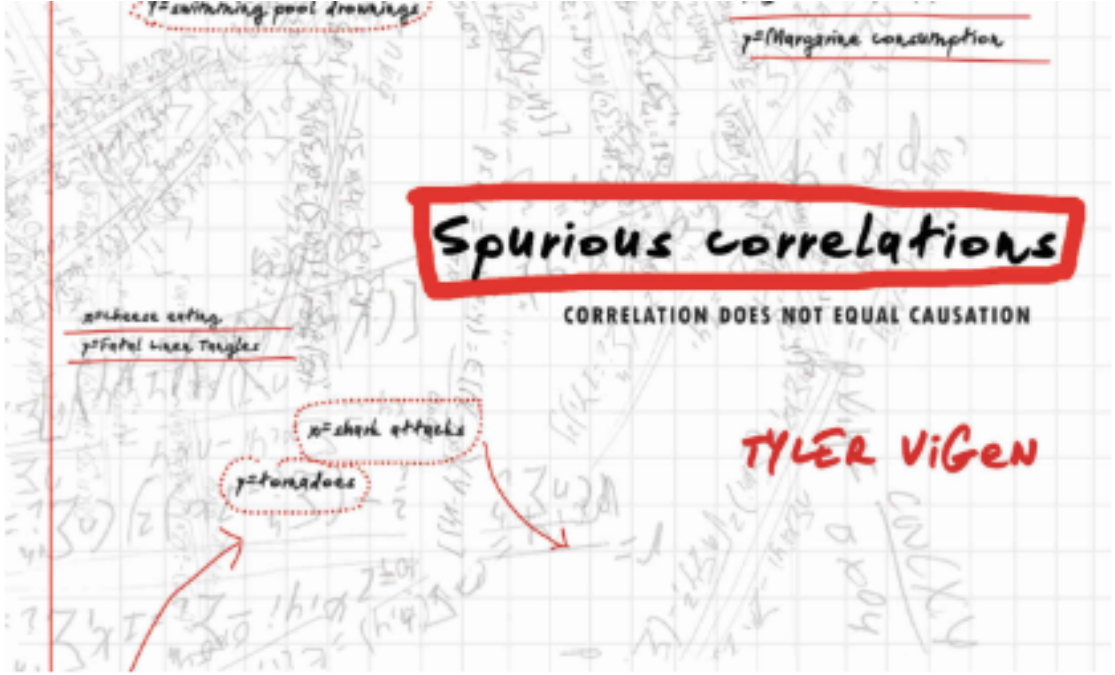
The Journal of Nutrition, Volume 143, Issue 6, 1 June 2013, Pages 931–933, https://doi.org/10.3945/jn.113.174813

http://tylervigen.com/view_correlation?id=1582

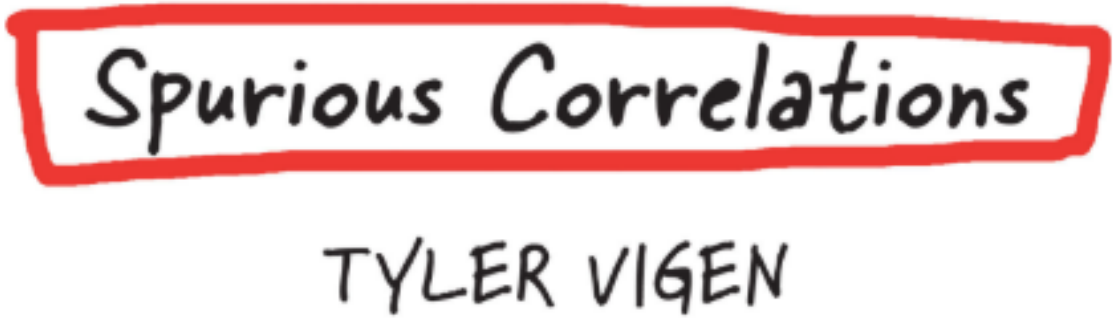
Spurious correlation: Examples

- Many examples of spurious relationships can be found in the ***time-series*** literature





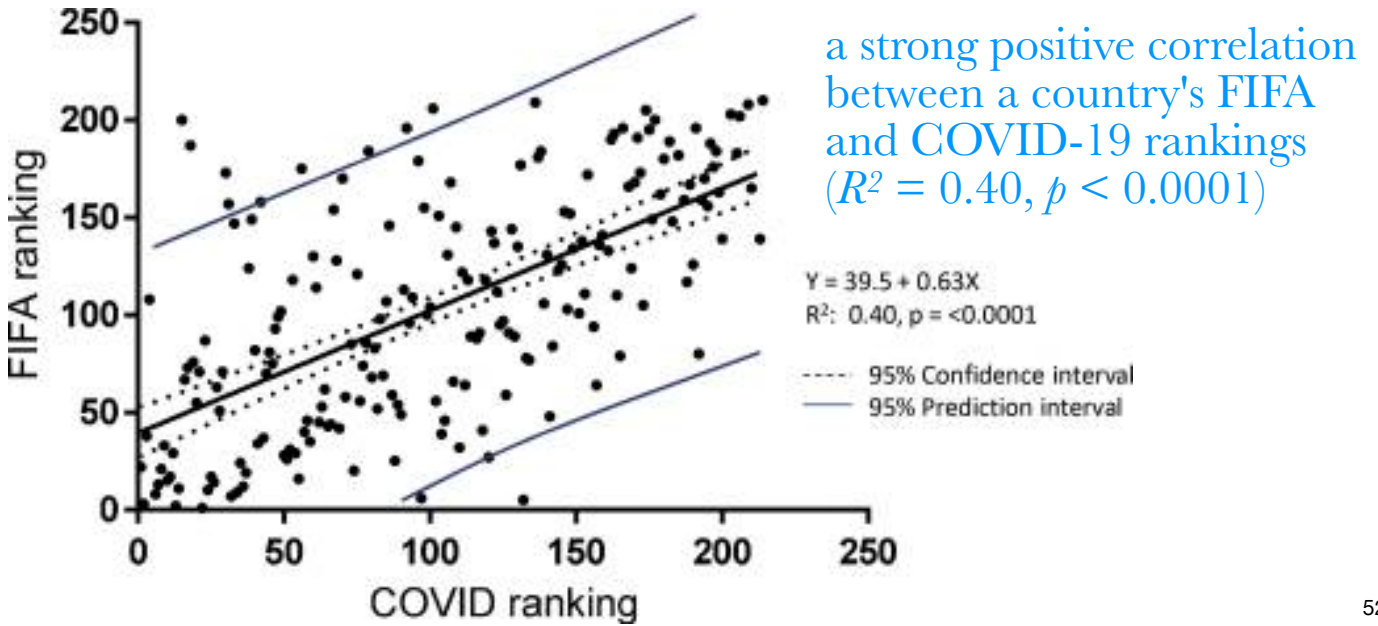
http://www.tylervigen.com/spurious-correlations



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7470813/
doi:

Correlation

- COVID-19 ranking of countries based on the highest total number COVID-19 cases (16.06.2020)
- FIFA ranking of national football teams on the basis of international matches over the period of the last 4 years



a strong positive correlation between a country's FIFA and COVID-19 rankings ($R^2 = 0.40$, $p < 0.0001$)

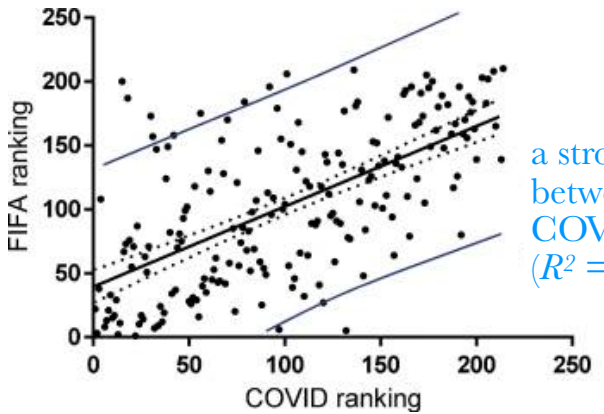
Clin Microbiol Infect. 2020 Sep 3 PMCID: PMC7470813
doi: 10.1016/j.cmi.2020.08.034 [Epub ahead of print] PMID: 32891763

Football and COVID-19 risk: correlation is not causation

Fares Ayoub,[1] Toshiro Sato,[2,3] and Atsushi Sakuraba[1,*]

Correlation is not causation

"Does this mean that people who are skilled in playing football are at increased risk of catching SARS-CoV-2 or of spreading it? Or does COVID-19 make you a better football player? This is unlikely to be the correct conclusion to draw from these findings. However, there might be **alternative explanations**.



a strong positive correlation between a country's FIFA and COVID-19 rankings ($R^2 = 0.40$, $p < 0.0001$)

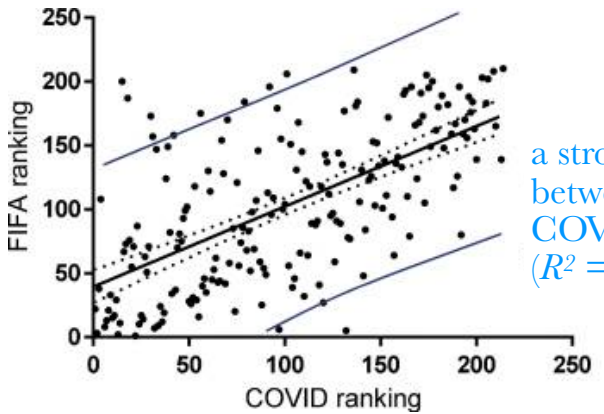
Clin Microbiol Infect. 2020 Sep 3 PMCID: PMC7470813
doi: 10.1016/j.cmi.2020.08.034 [Epub ahead of print] PMID: 32891763

Football and COVID-19 risk: correlation is not causation

Fares Ayoub,[1] Toshiro Sato,[2,3] and Atsushi Sakuraba[1,*]

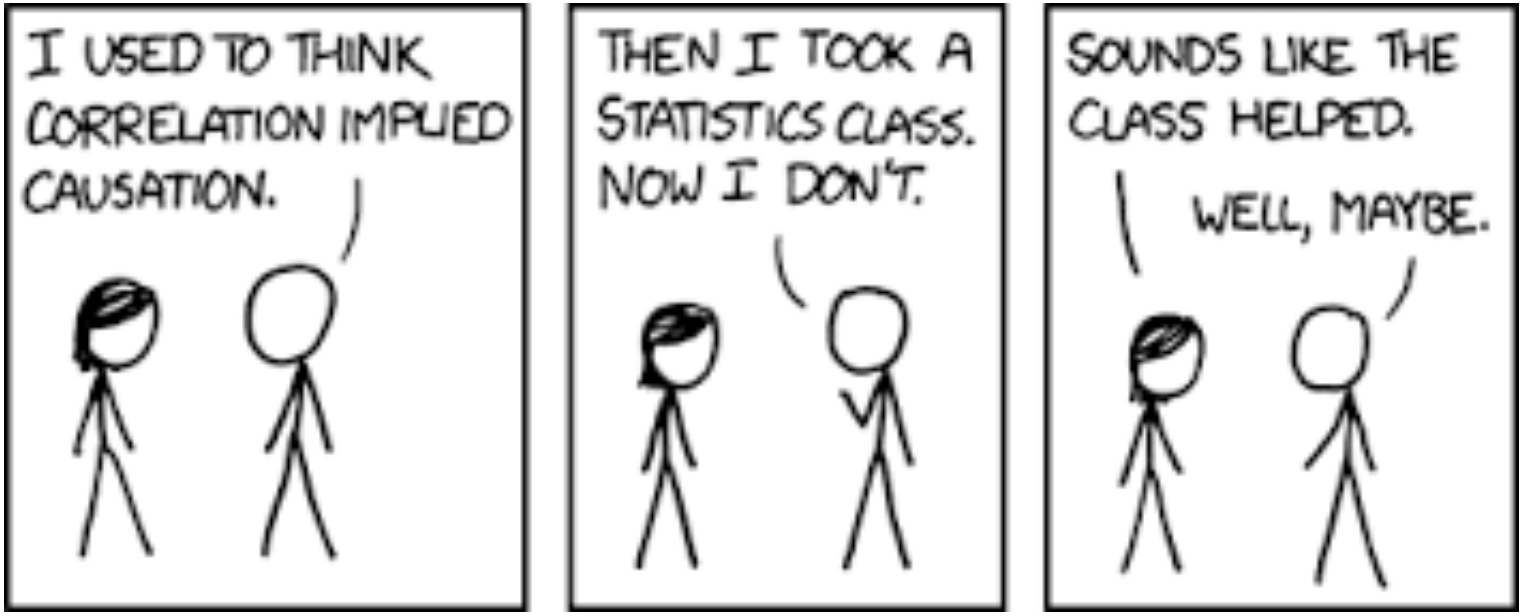
Correlation is not causation

"The mass gatherings at football stadiums early in the pandemic may have contributed to the spread of SARS-CoV-2 or the cross-border travel of the supporters during the UEFA Champions League may have spread it to multiple European countries. Or is it possible that SARS-CoV-2 spreads among the guests in the football pubs where the supporters got drunk without social distancing and face coverings? These latter explanations appear more reasonable than the former, but no matter how strong **correlation does not equal causation**."



a strong positive correlation between a country's FIFA and COVID-19 rankings ($R^2 = 0.40$, $p < 0.0001$)

Correlation is not causation



The Regression Line

In perfect linear relationships the line that fits exactly the data have slope Sy/Sx and passes through the point (x,y) or SD line.

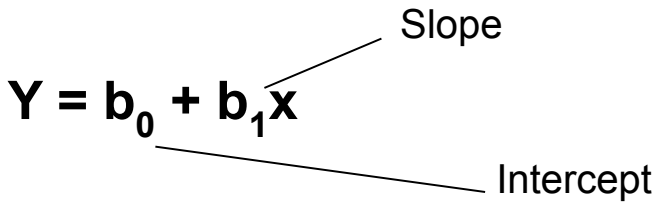
The Regression Line

In perfect linear relationships the line that fits exactly the data have slope Sy/Sx and passes through the point (x,y) or SD line. When there is not linear relationship:

$$\mathbf{Y = b_0 + b_1 x}$$

The Regression Line

In perfect linear relationships the line that fits exactly the data have slope Sy/Sx and passes through the point (x,y) or SD line. When there is not linear relationship:



The Regression Line

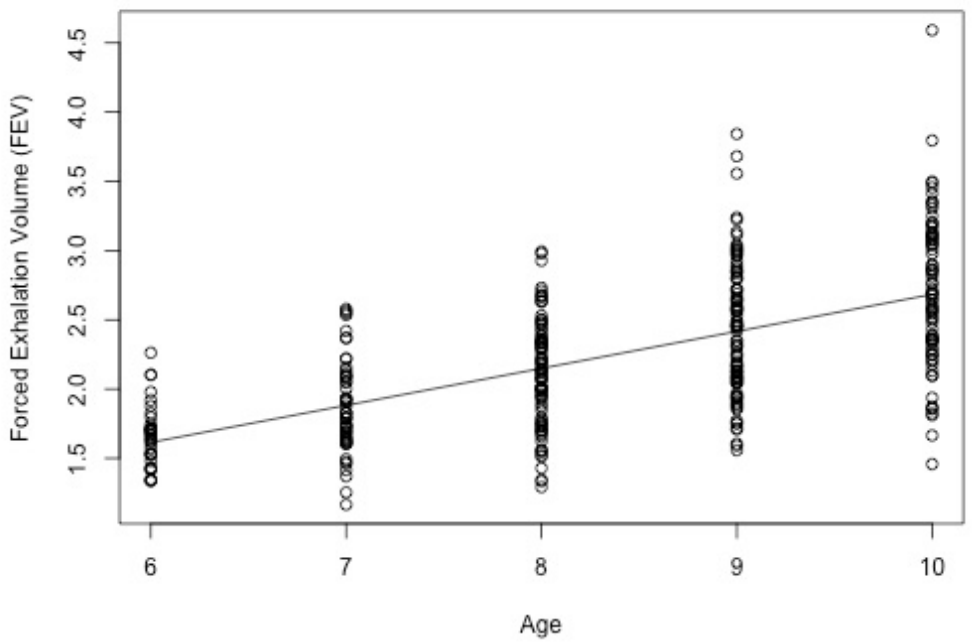
In perfect linear relationships the line that fits exactly the data have slope Sy/Sx and passes through the point (x,y) or SD line. When there is not linear relationship:

Slope

$$Y = b_0 + b_1 x$$

Intercept

Intercept = 0.01165 Slope = 0.26721



The Regression Line

In perfect linear relationships the line that fits exactly the data have slope Sy/Sx and passes through the point (x,y) or SD line. When there is not linear relationship:

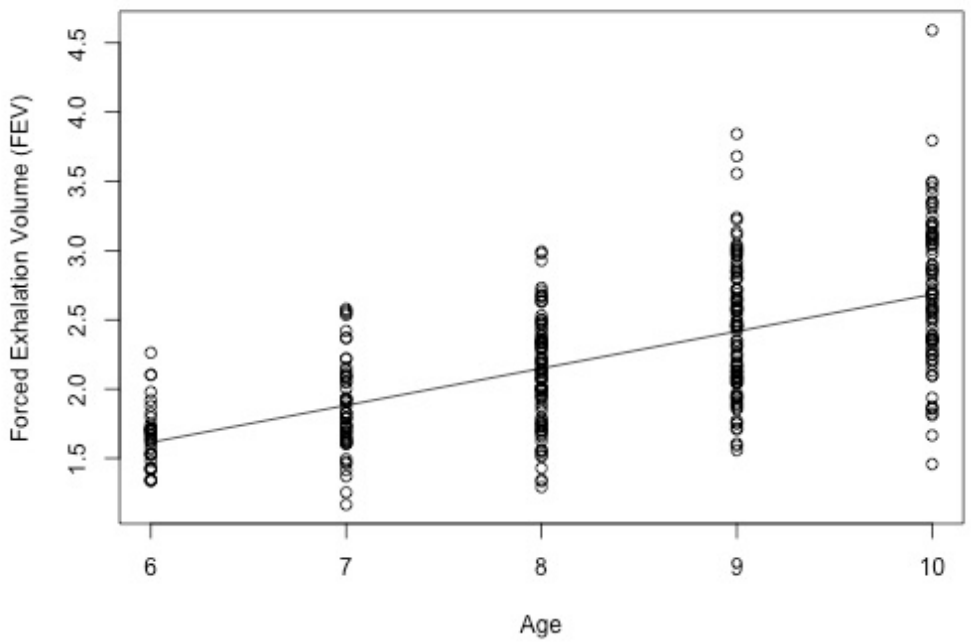
Slope

$$Y = b_0 + b_1 x$$

Intercept

Intercept = 0.01165 Slope = 0.26721

FEV = 0.01165 + 0.26721*Age



The Regression Line

In perfect linear relationships the line that fits exactly the data have slope Sy/Sx and passes through the point (x,y) or SD line. When there is not linear relationship:

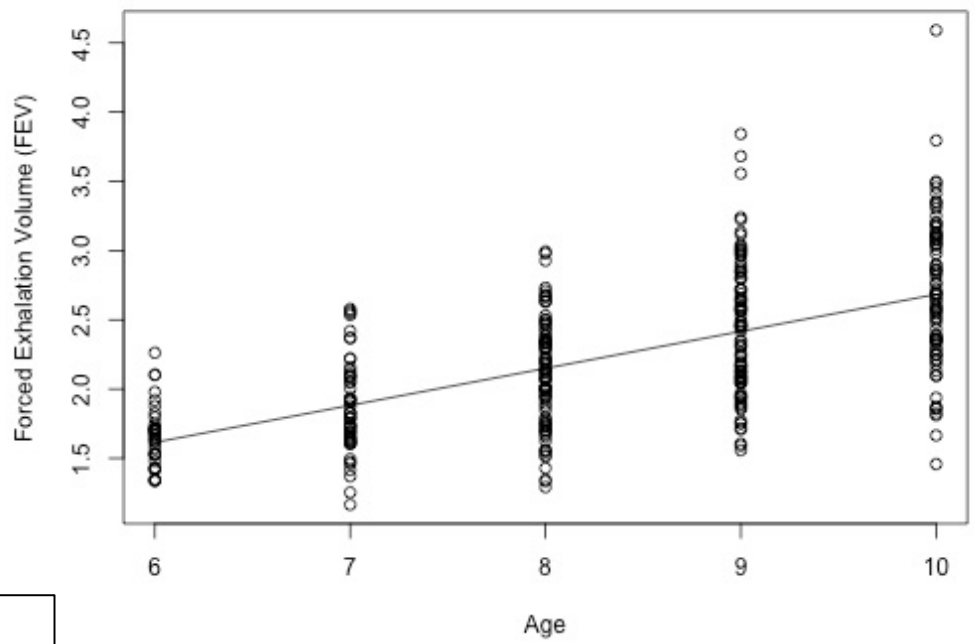
Slope

$$Y = b_0 + b_1 x$$

Intercept

Intercept = 0.01165 Slope = 0.26721

FEV = 0.01165 + 0.26721*Age



In R you can estimate slope & intercept:

lm(formula = Response ~ Explanatory, data = dataset)

(General) Linear Models

The General Linear Models are used to predict one Response variable from one or more Explanatory variables

Image taken from “Mixed Effects Models and Extensions in Ecology with R” Alain F. Zuur, Elena N. Ieno, Neil Walker, Anatoly A. Saveliev, Graham M. Smith - (2009, Springer-Verlag New York)

(General) Linear Models

The General Linear Models are used to predict one Response variable from one or more Explanatory variables

- Simple Regression ⟶ $Y = b_0 + b_1 x$

Image taken from “Mixed Effects Models and Extensions in Ecology with R” Alain F. Zuur, Elena N. Ieno, Neil Walker, Anatoly A. Saveliev, Graham M. Smith - (2009, Springer-Verlag New York)

(General) Linear Models

The General Linear Models are used to predict one Response variable from one or more Explanatory variables

- Simple Regression ⟶ $Y = b_0 + b_1x$
- Multiple Regression ⟶ $Y = b_0 + b_1x_1 + b_2x_2 + b_3x_3...$

Image taken from “Mixed Effects Models and Extensions in Ecology with R” Alain F. Zuur, Elena N. Ieno, Neil Walker, Anatoly A. Saveliev, Graham M. Smith - (2009, Springer-Verlag New York)

(General) Linear Models

The General Linear Models are used to predict one Response variable from one or more Explanatory variables

- Simple Regression ⟶ $Y = b_0 + b_1x$
- Multiple Regression ⟶ $Y = b_0 + b_1x_1 + b_2x_2 + b_3x_3...$

Assumption

- Linearity
- Normality of residuals
- Homoscedasticity
(Homogeneity of variance)

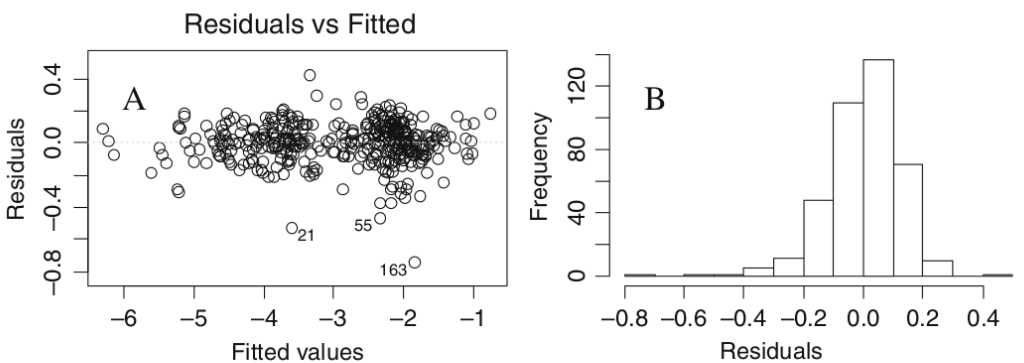


Image taken from “Mixed Effects Models and Extensions in Ecology with R” Alain F. Zuur, Elena N. Ieno, Neil Walker, Anatoly A. Saveliev, Graham M. Smith - (2009, Springer-Verlag New York)

(General) Linear Models

The General Linear Models are used to predict one Response variable from one or more Explanatory variables

Simple regression

(General) Linear Models

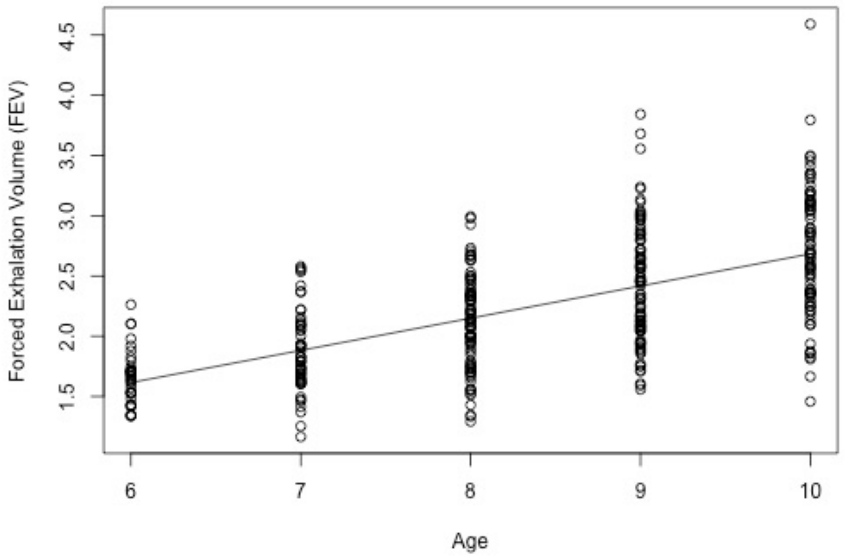
The General Linear Models are used to predict one Response variable from one or more Explanatory variables

Simple regression ⟶ One explanatory variables $Y = b_1X + b_0$

(General) Linear Models

The General Linear Models are used to predict one Response variable from one or more Explanatory variables

Simple regression ⟶ One explanatory variables $Y = b_1X + b_0$



(General) Linear Models

The General Linear Models are used to predict one Response variable from one or more Explanatory variables

Simple regression ⟶ One explanatory variables $Y = b_1X + b_0$

Linear Regression

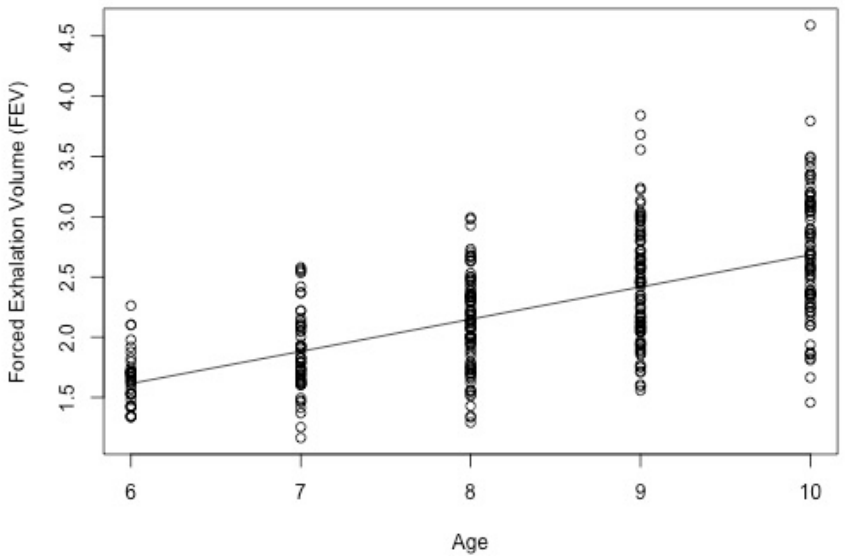
X

↘

Y

X explain Y

X ~ Y



(General) Linear Models

The General Linear Models are used to predict one Response variable from one or more Explanatory variables

Multiple regression

(General) Linear Models

The General Linear Models are used to predict one Response variable from one or more Explanatory variables

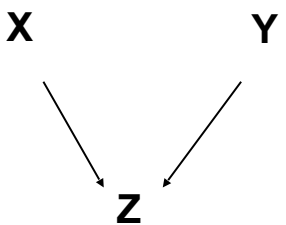
Multiple regression ⟶ Have multiple explanatory variables $Y = b_1X_1 + b_2X_2 + b_3X_3 + \ldots. + b_0$

(General) Linear Models

The General Linear Models are used to predict one Response variable from one or more Explanatory variables

Multiple regression ⟶ Have multiple explanatory variables $Y = b_1X_1 + b_2X_2 + b_3X_3 + \ldots. + b_0$

Additive independent effects



X and Y explain the variation in Z independently

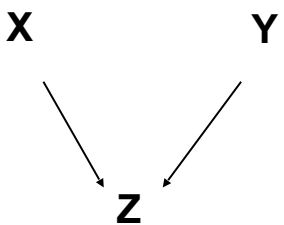
$Z \sim X + Y$

(General) Linear Models

The General Linear Models are used to predict one Response variable from one or more Explanatory variables

Multiple regression ⟶ Have multiple explanatory variables $Y = b_1X_1 + b_2X_2 + b_3X_3 + \ldots. + b_0$

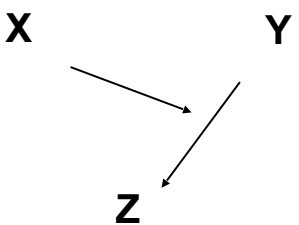
Additive independent effects



X and Y explain the variation in Z independently

Z ~ X + Y

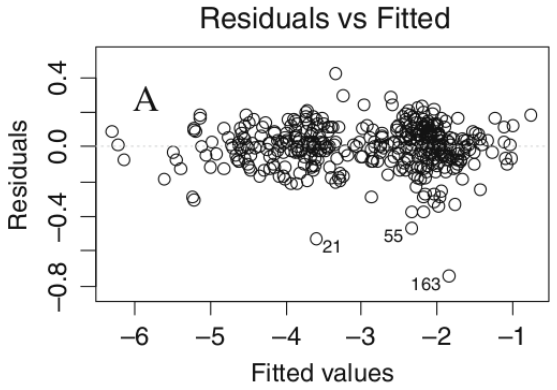
Interaction among variable



X modifies how Y affects Z

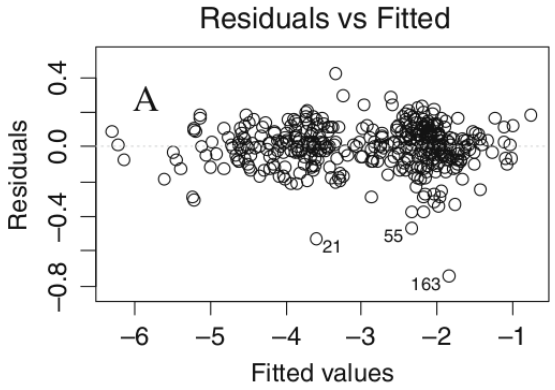
Z ~ X + Y + X*Y

Residuals

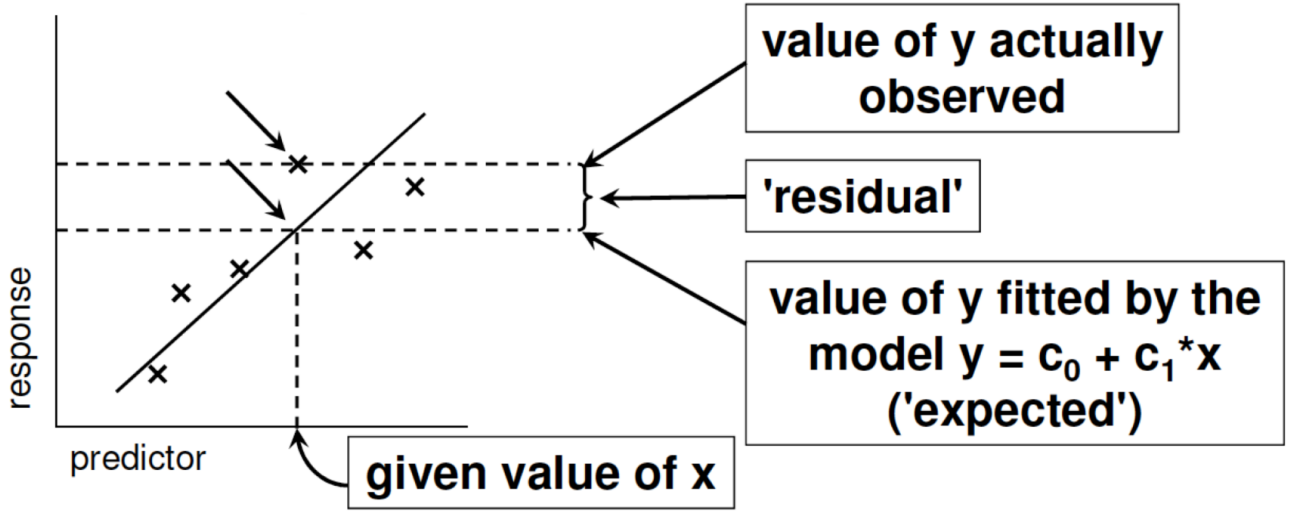


- *Residuals:* Difference between observation and fitted values
- *Fitted values: Estimation of an observation using all previous ones*

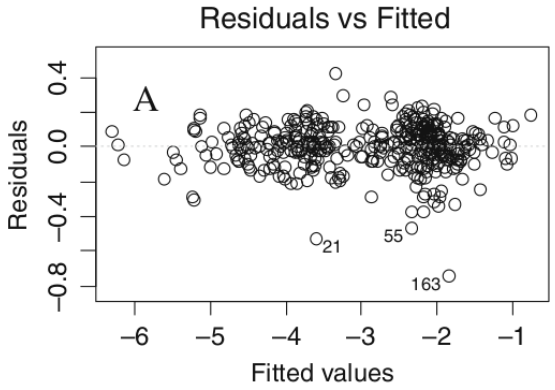
Residuals



- *Residuals:* Difference between observation and fitted values
- *Fitted values: Estimation of an observation using all previous ones*



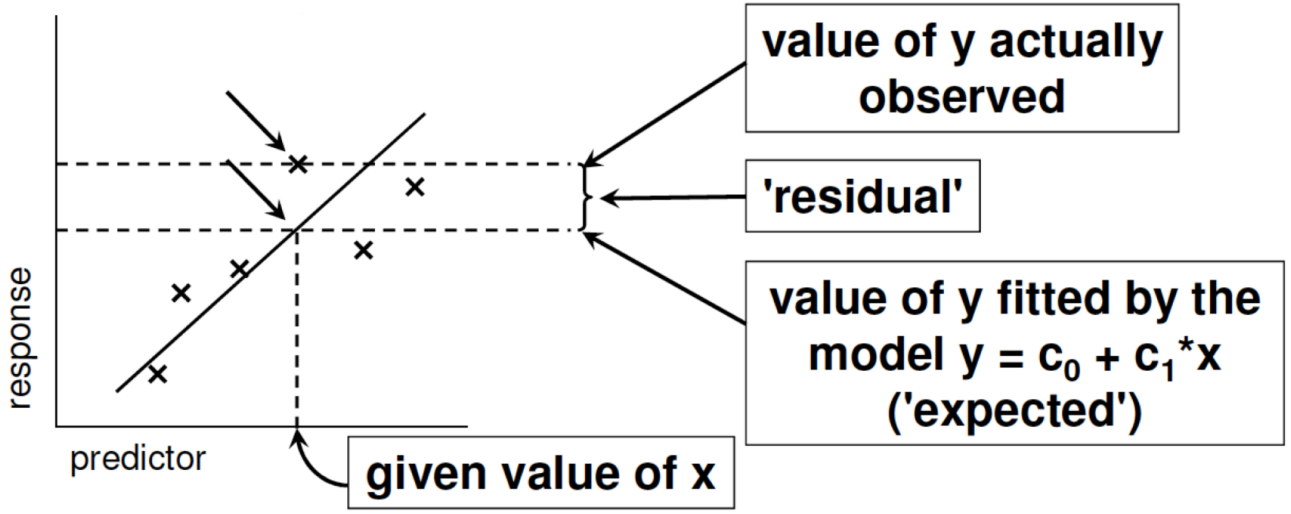
Residuals



- *Residuals:* Difference between observation and fitted values
- *Fitted values: Estimation of an observation using all previous ones*

Error (so called Residual)

$$Y = b_1X + b_0 + e$$



Generalized Linear Models

If we don’t have the normality of residuals, we can use the Generalized Linear Models (GLM).

Generalized Linear Models

If we don’t have the normality of residuals, we can use the Generalized Linear Models (GLM).

- Can be used with residuals with distribution normal, binomial, poisson...
- Have the same features of General Linear Models

Generalized Linear Models

If we don’t have the normality of residuals, we can use the Generalized Linear Models (GLM).

- Can be used with residuals with distribution normal, binomial, poisson...
- Have the same features of General Linear Models

In R you can fit your data in a General Linear Model:

*lm(formula = Response ~ Explanatory + Z + Z*Y, data = dataset)*

In R you can fit your data in a GLM:

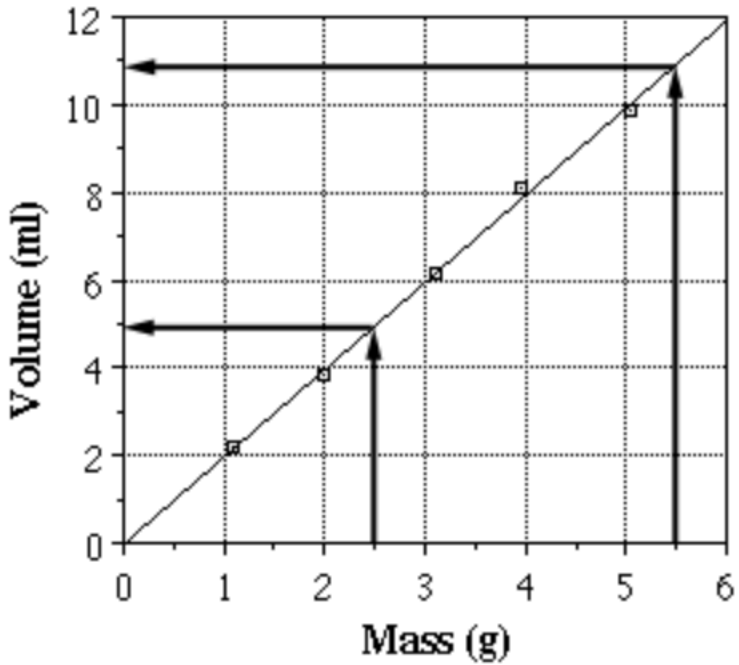
*glm(formula = Response ~ Explanatory + Z + Z*Y, family = binomial, data = dataset)*

Summary

Model	Variables	Distribution	R code
Linear Regression	$Y = b_0 + b_1x$	Normal	*lm(formula, data)*
General Linear Models	$Y = b_0 + b_1x_1 + b_2x_2 + ...$	Normal	*lm(formula, data)*
Generalized Linear Models (GLM)	$Y = b_0 + b_1x_1 + b_2x_2 + ...$	Any	*glm(formula, family, data)*

Interpolation and extrapolation

- Interpolation: predict values that fall **within** the range of data points taken
- Extrapolation: predict values **outside** of the range of data points taken



http://dept.harpercollege.edu/chemistry/chm/100/dgodambe/thedisk/labtech/data7.htm

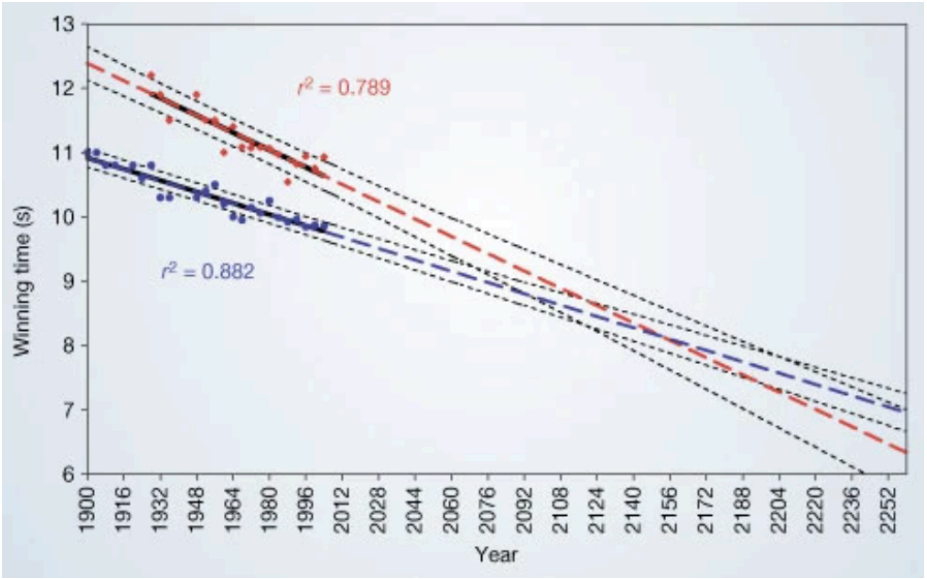
Momentous sprint at the 2156 Olympics?

Andrew J. Tatem, Carlos A. Guerra, Peter M. Atkinson & Simon I. Hay

Nature **431**, 525(2004) | Cite this article

Women sprinters are closing the gap on men and may one day overtake them.

- The 2004 Olympic women's 100-metre sprint champion, Yuliya Nesterenko, is assured of fame and fortune. But we show here that — if current trends continue — it is the winner of the event in the 2156 Olympics whose name will be etched in sporting history forever, because this may be the first occasion on which the race is won in a faster time than the men's event.



blue line: men
red line: women, dotted black lines: 95% confidence intervals

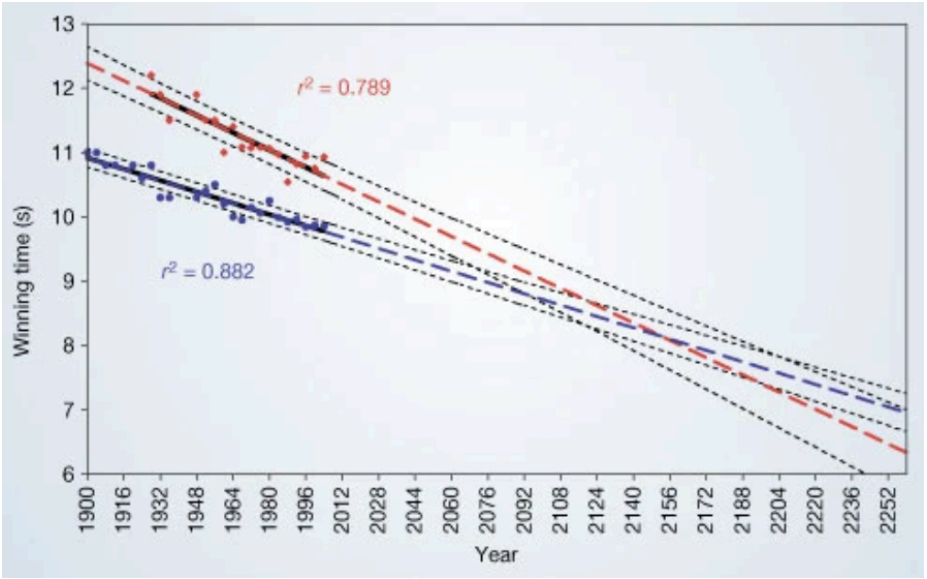
Momentous sprint at the 2156 Olympics?

Andrew J. Tatem ✉, Carlos A. Guerra, Peter M. Atkinson & Simon I. Hay

Nature **431**, 525(2004) | Cite this article

Women sprinters are closing the gap on men and may one day overtake them.

- The 2004 Olympic women's 100-metre sprint champion, Yuliya Nesterenko, is assured of fame and fortune. But we show here that — if current trends continue — it is the winner of the event in the 2156 Olympics whose name will be etched in sporting history forever, because this may be the first occasion on which the race is won in a faster time than the men's event.



"a hilarious extrapolation error"
(Winter 2019:72)


nature

Explore our content ⌄ Journal information ⌄

nature › correspondence › article

Published: 10 November 2004

Sprint research runs into a credibility gap

Kenneth Rice

Nature **432**, 147(2004) | Cite this article

439 Accesses | **10** Citations | **6** Altmetric | Metrics

- Sir:

  A. J. Tatem and colleagues calculate that women may outsprint men by the middle of the twenty-second century (*Nature* 431, 525; 200410.1038/431525a). They omit to mention, however, that (according to their analysis) a far more interesting race should occur in about 2636, when times of less than zero seconds will be recorded.

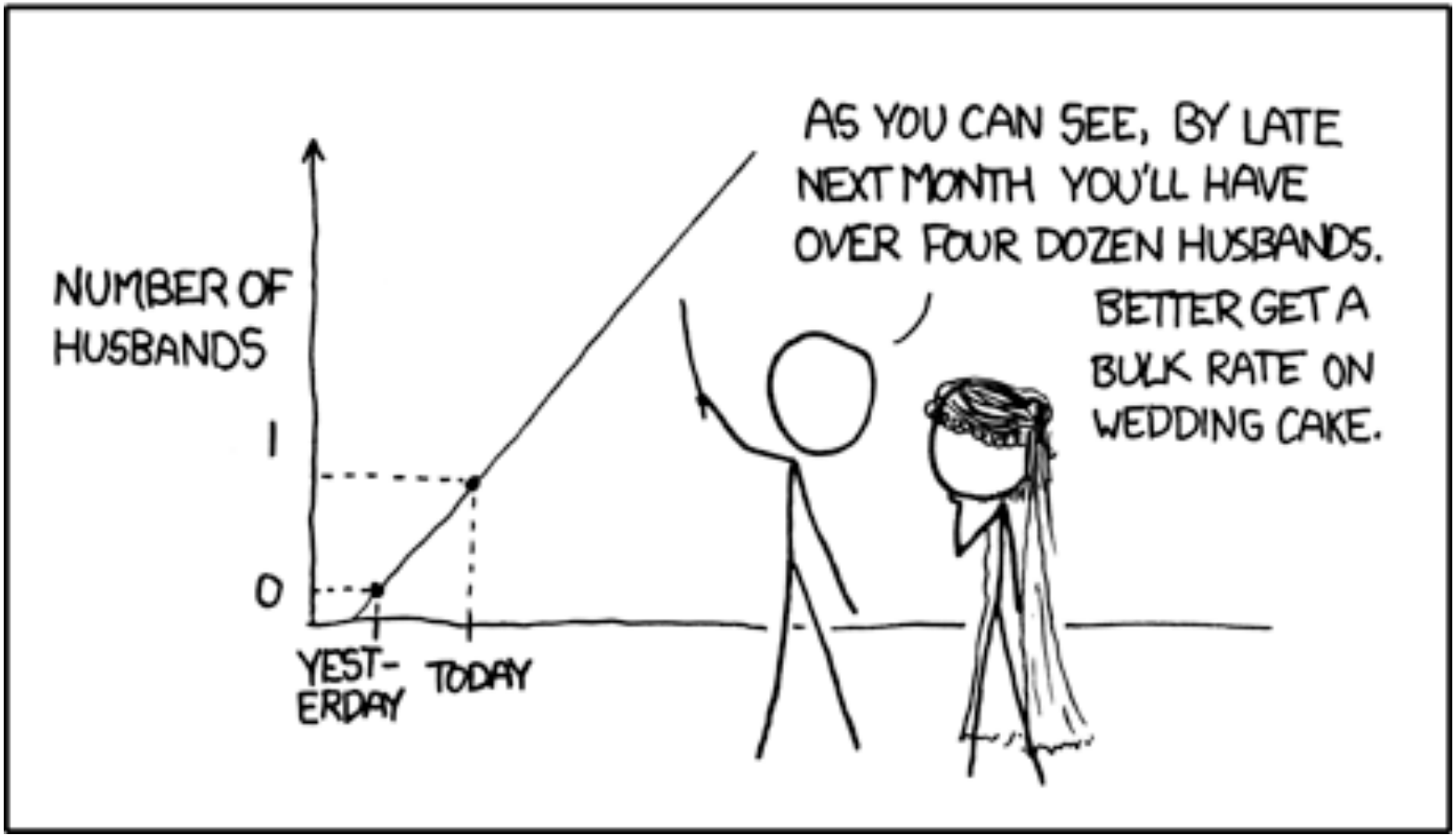
  In the intervening 600 years, the authors may wish to address the obvious challenges raised for both time-keeping and the teaching of basic statistics.

Extrapolation

- Extrapolation beyond the scope of the model occurs when one uses an estimated regression equation to predict a response *y* for new *x* values not in the range of the sample data used to determine the estimated regression equation.
- In general, it is dangerous to extrapolate beyond the scope of model.
- Extrapolation is subject to greater **uncertainty** and a higher risk of producing **meaningless results**: the trend in the data as summarized by the estimated regression equation does not necessarily hold outside the scope of the model.

https://xkcd.com/605/

MY HOBBY: EXTRAPOLATING



If she loves you more each and every day,
by linear regression she hated you before you met.


datacamp


R