OXMEDICA

Driving Global Education

Linear regression

A motivating example: Education and future wages

See whiteboard

A simple functional form: World assumption

$$y^{(i)} = \alpha + \beta x_1^{(i)}$$

A simple functional form: The Linear regression model

$$\hat{y}^{(i)} = \hat{\alpha} + \hat{\beta} x_1^{(i)} + \varepsilon^{(i)}$$

Final Model

$$\hat{y}^{(i)} = \hat{\alpha} + \hat{\beta} x_1^{(i)}$$

Extending this to more features

$$\hat{y}^{(i)} = \hat{\alpha} + \hat{\beta}_1 x_1^{(i)} + \hat{\beta}_2 x_2^{(i)} + \hat{\beta}_3 x_3^{(i)} + \hat{\beta}_4 x_4^{(i)}$$

Discussion

- How good is this model?
- Why might it fail?
- …lots of other possible questions…!

5 Min Break

Individual Task: Worksheet

https://www.harrymayne.com/oxmedica

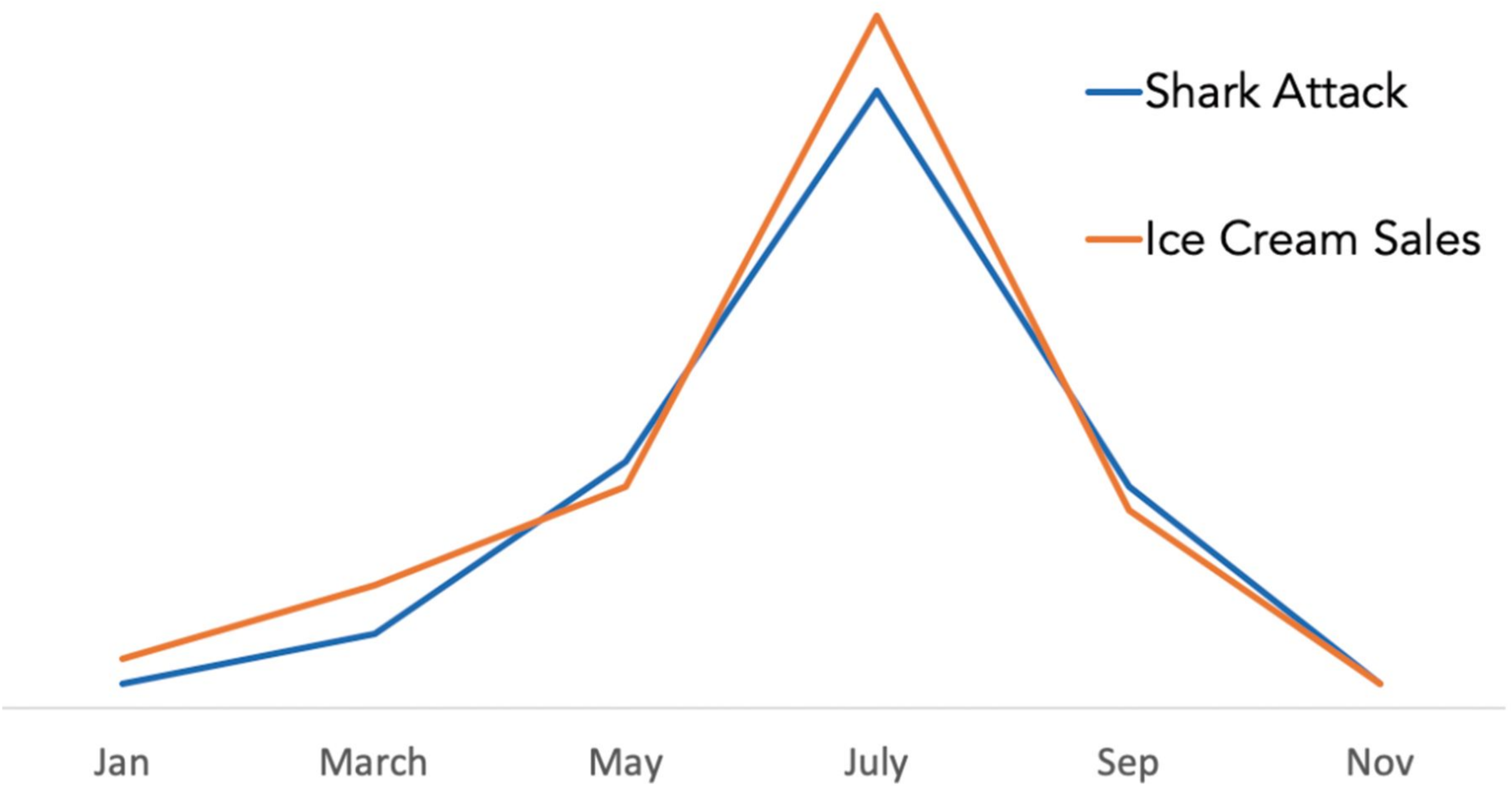
Part 2
Correlation vs Causation

Correlation and causation

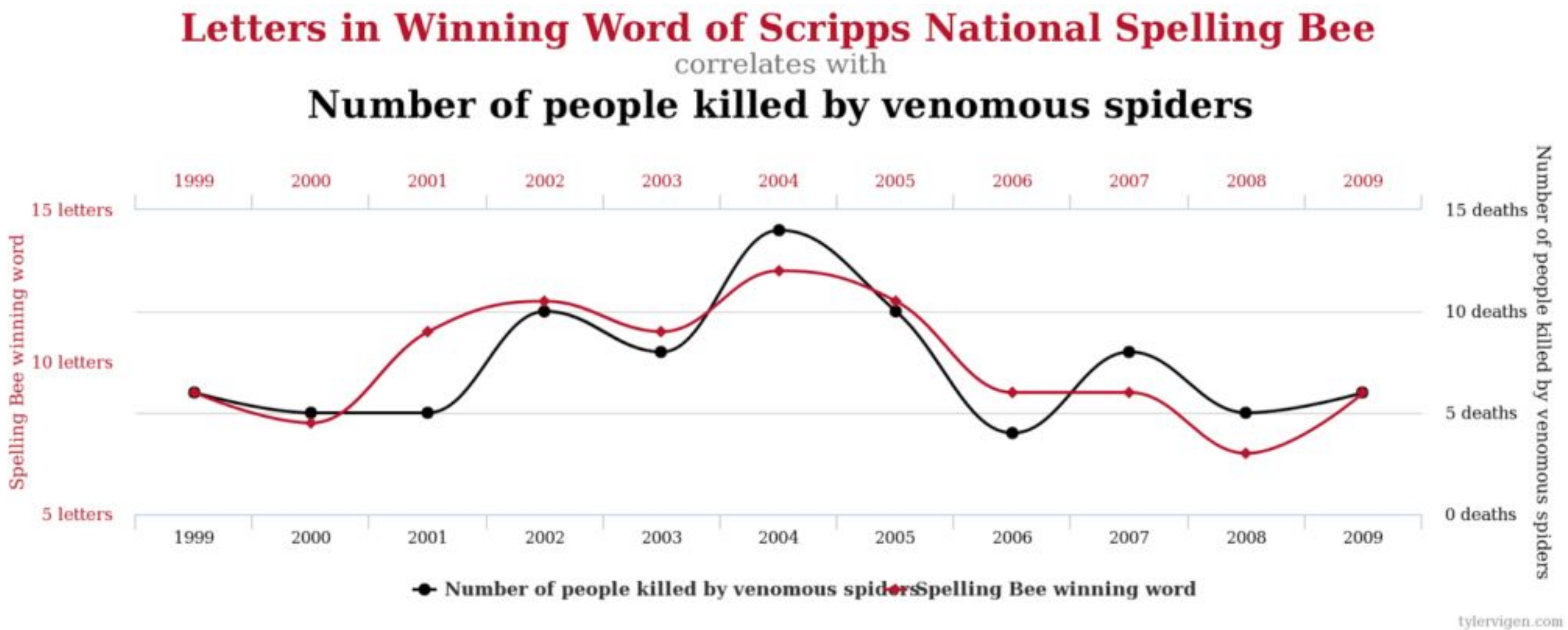
Correlation: A relationship between two variables.

Causality: A change in one variable causes a change in another variable.

- How are they different?
- Is this important?
- Examples of things which correlate well but are not causal?



Source


Letters in Winning Word of Scripps National Spelling Bee
correlates with
Number of people killed by venomous spiders
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
15 letters
10 letters
5 letters
Spelling Bee winning word
15 deaths
10 deaths
5 deaths
0 deaths
Number of people killed by venomous spiders
Number of people killed by venomous spiders
Spelling Bee winning word
tylervigen.com

[Source]

What does this really tell us?

$$\hat{y}^{(i)} = \hat{\alpha} + \hat{\beta} x_1^{(i)} + \varepsilon^{(i)}$$

EXTENSION: The Gauss-Markov Theorem

Gauss–Markov theorem

17 languages

Article Talk

Read Edit View history Tools

From Wikipedia, the free encyclopedia

Not to be confused with Gauss–Markov process.

"BLUE" redirects here. For queue management algorithm, see Blue (queue management algorithm). For the color, see Blue.

In statistics, the **Gauss–Markov theorem** (or simply **Gauss theorem** for some authors) [1] states that the ordinary least squares (OLS) estimator has the lowest sampling variance within the class of linear unbiased estimators, if the errors in the linear regression model are uncorrelated, have equal variances and expectation value of zero. [2] The errors do not need to be normal for the theorem to apply, nor do they need to be independent and identically distributed (only uncorrelated with mean zero and homoscedastic with finite variance).

The requirement for unbiasedness cannot be dropped, since biased estimators exist with lower variance and mean squared error. For example, the James–Stein estimator (which also drops linearity) and ridge regression typically outperform ordinary least squares. In fact, ordinary least squares is rarely even an admissible estimator, as Stein's phenomenon shows--when estimating more than two unknown variables, ordinary least squares will always perform worse (in mean squared error) than Stein's estimator.

Part of a series on

Regression analysis

Models

Linear regression · Simple regression · Polynomial regression · General linear model

Generalized linear model · Vector generalized linear model · Discrete choice · Binomial regression · Binary regression · Logistic regression · Multinomial logistic regression · Mixed logit · Probit · Multinomial probit · Ordered logit · Ordered probit · Poisson

Multilevel model · Fixed effects · Random effects · Linear mixed-effects model · Nonlinear mixed-effects model

Nonlinear regression · Nonparametric · Semiparametric · Robust · Quantile · Isotonic · Principal components · Least angle · Local ·

Reverse Causality

Reverse Causality: When a change in the y causes a change in the x in a causal way. I.e. we assumed the relationship was the other way round.

Reverse Causality

Mental Health and Social Media Use:

- What is the **perceived causality**
- What might the **reverse causality** be?

- **Perceived Causality:** Increased social media use leads to poor mental health.
- **Reverse Causality:** Individuals with poor mental health are more likely to spend more time on social media as a form of escapism or social connection.

Confounding variables

See whiteboard

A Debate

Task in groups of 5: Debate

You will be split into groups of 5. In your group of 5 you will tasked to defend one side of a debate. You will get **15 minutes** to prepare two arguments.

One side of the argument will argue that you should use regression models to address a problem, the other will argue that you should not. You and your group will have **15 minutes** to prepare two arguments for your side. You are welcome to do some research and use evidence in your arguments

After 20 minutes you will take it in turn to make one argument (uninterrupted by the other pair). Each argument can be a maximum of 2 minutes.

After each side has said their two arguments, you have 1 minute to freestyle a final response.

[Each person can do a maximum of one argument or response]

Task in groups of 5: Debate

Debate 1

"Crime rates are generally higher in areas with lower socioeconomic status. Given this, should law enforcement use socioeconomic data to predict future crime rates and allocate resources for crime prevention?"

Debate 2

"A study has shown that there is a strong correlation between the amount of time teenagers spend on social media and their reported levels of anxiety and depression. A linear regression model built from social media data predicts mental health outcomes and uses this to predict which students to offer support to first. Is this a good use of the linear regression model?"