

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

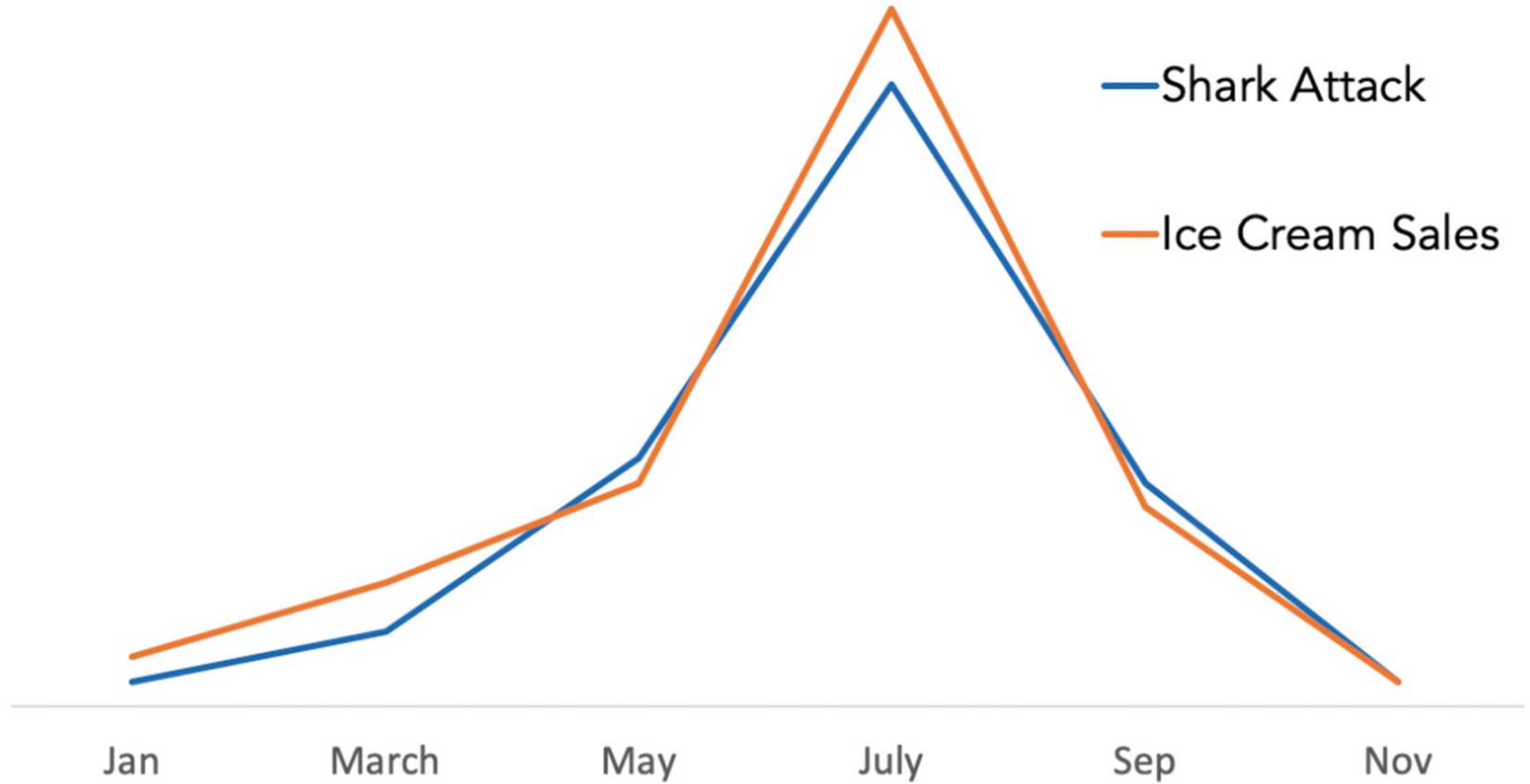
Correlation vs Causation

Correlation and causation

Correlation: A relationship between two variables.

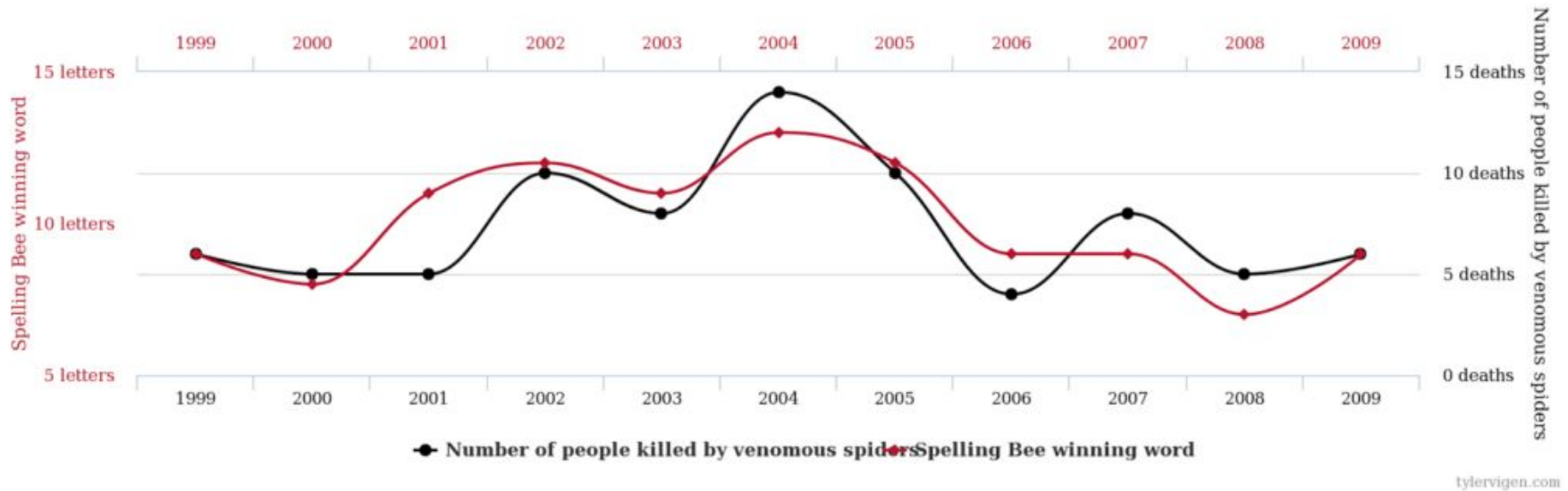
Causality: A change in one variable causes a change in another variable.

- 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



[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 be 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?”

EXTRA

A real example

Part II

Applications of Supervised Learning in Finance

Quant Finance

What is Quantitative (Quant) Finance

- Specific wing of finance concerned with using mathematics (often called **mathematical finance** or **financial mathematics**)
- Alternative is traditional trading

Difference Between Traditional & Quant Trading

Traditional Trading

Intuition based Decision Making



Emotional Biases



Based on Experience



Slow and Time Taking



Limited



Quant Trading

Data Driven Decision Making

Objective decisions

Mathematical Proficiency

Fast Paced

Scalable



[Source](#)

Financial objects

Stocks: A share of the ownership of a company. Also called shares or equity.

Bonds: Debt issued by a company or the government with the promise to pay the buyer interest payments (coupons) and repayment at maturity

Derivatives: Contracts whose value is derived from underlying financial assets. You can think of these as calculated bets on the price of other financial objects. Useful for hedging risk.

Risk and return

- Buying assets has the promise of **return**
- ... but comes at some **risk** of the return not materialising

Quantitative Finance

- Goal is to use statistics and machine learning to model the price of assets and therefore predict the best time to buy/sell
- Is this supervised or unsupervised?

Discussion: What is a basic model we could use?

Simple Moving Average (SMA)

- Average the prices over a period to predict the future price
- Simple as that!
- Common windows are 10-day, 50-day, or 200-day MAs.
- **Question:** What would a price above the moving average indicate and how might you act?

Exponential Moving Average (EMA):

- Slightly more complex model...
- Weights recent prices a little bit higher in predicting the next price

Example strategy

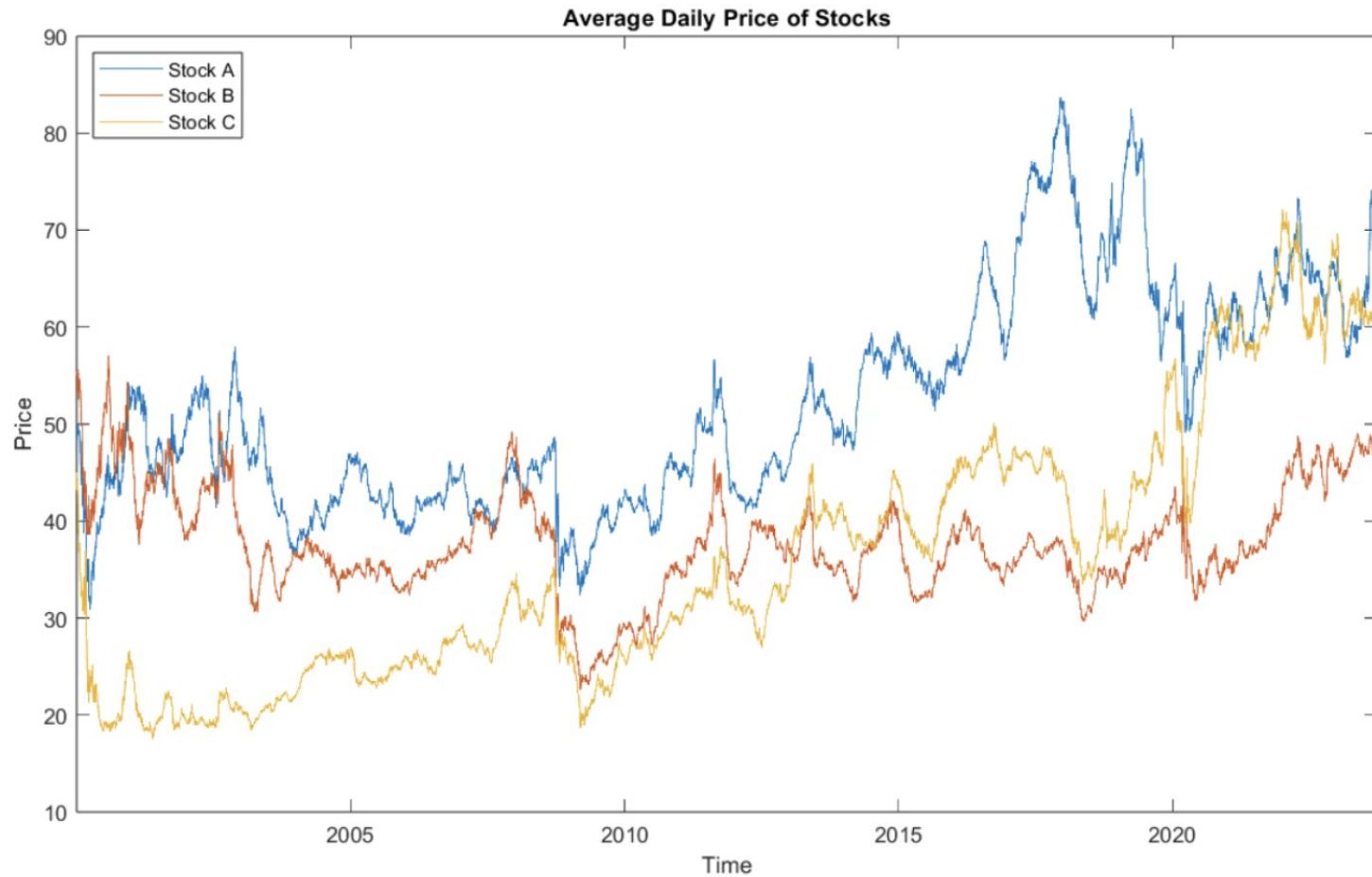
- Calculate the 200 and 50-day moving averages
- When the 50-day MA $>$ 200-day MA we call this the “Golden Cross”
-> What should we do here?
- When the 50-day MA $<$ 200-day MA we call this the “Death Cross”.
What should we do here?

Discussion in pairs: What are the strengths and limitations of this method?

Strengths and weaknesses of this strategy

- ✓ Simple strategy which may make money long-term
- ✗ The cross is likely to indicate a past change in the trend so it might have already happened by the time you make your move!

Advanced Machine Learning Models



[Source](#)

Advanced Machine Learning Models: Deep Learning

- Use **deep learning models** to predict future stock price.
- Is this supervised or unsupervised learning? Why?
- What might be useful input features?

Discussion in pairs:

Can you think of any limitations of quantitative approaches to finance?

Limitations

1. **Overfitting** on past data.
2. Complexity
3. Interpretability and explainability.
4. Herding
5. Non-quantitative factors like “mood” and “expectations”
6. Events not captured by the training data e.g. 2008

Case Study

Individual Task

Long-Term Capital Management

- Read this overview of the hedge fund Long-Term Capital Management, then we're going to discuss the text.
- <https://www.bauer.uh.edu/rsusmel/7386/ltdcm-2.htm>
- Discussion