Time Series Analysis

Lecture 5

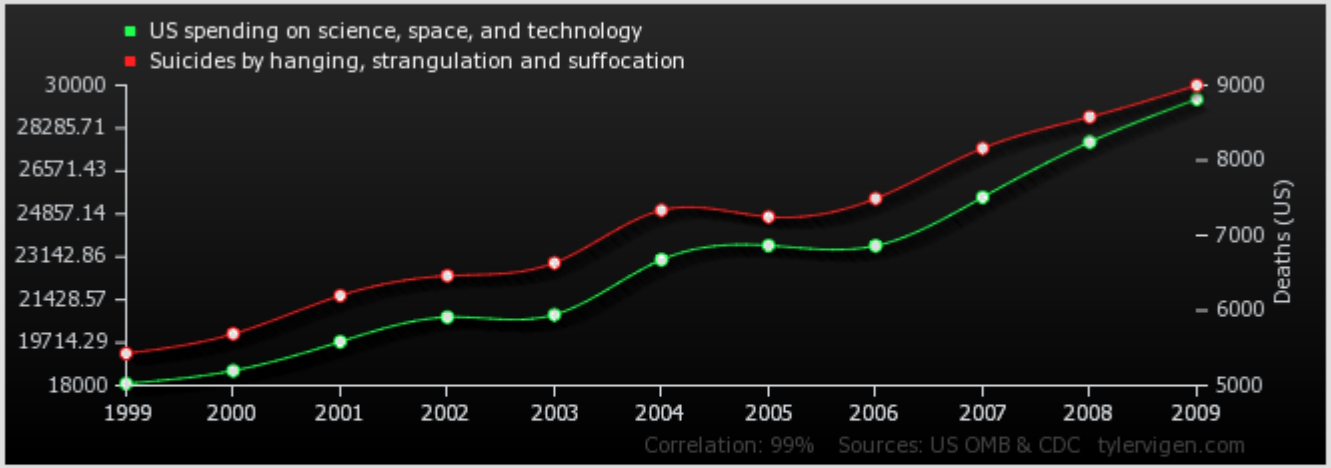
Vector Autoregressive (VAR) Models

datascience@berkeley

Spurious Correlation

Spurious Correlation: Example 1

US spending on science, space, and technology
correlates with
Suicides by hanging, strangulation and suffocation

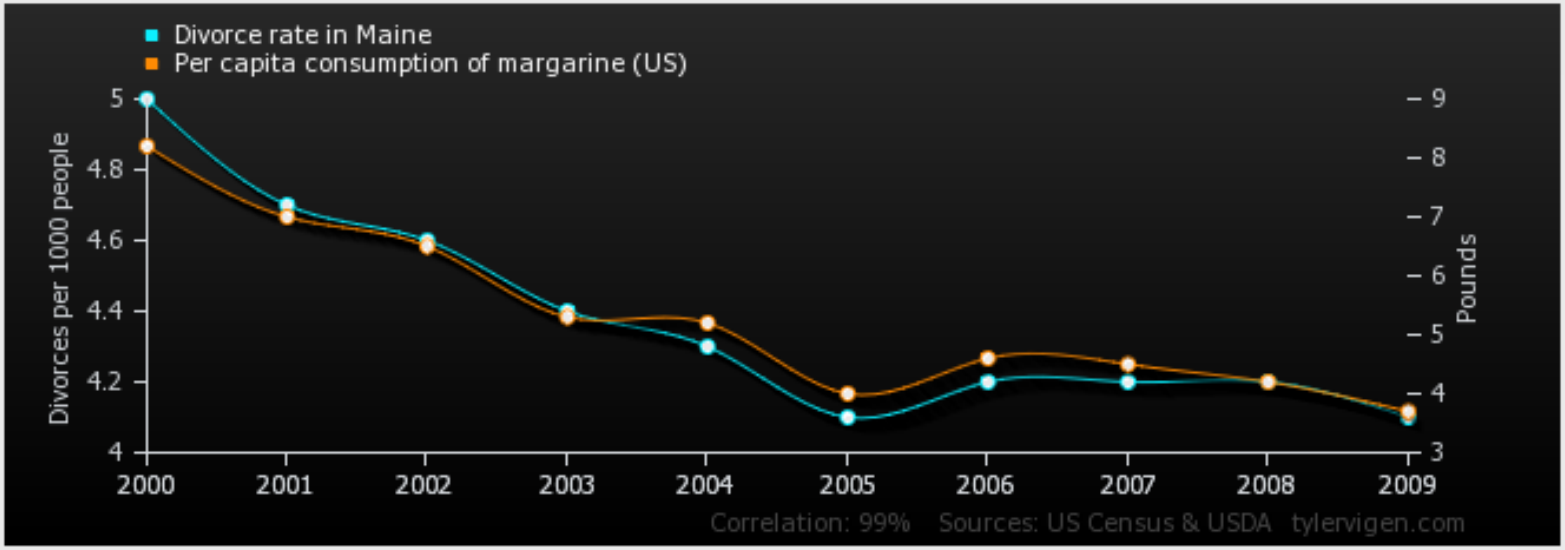


	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
US spending on science, space, and technology *Millions of todays dollars (US OMB)*	18,079	18,594	19,753	20,734	20,831	23,029	23,597	23,584	25,525	27,731	29,449
Suicides by hanging, strangulation and suffocation *Deaths (US) (CDC)*	5,427	5,688	6,198	6,462	6,635	7,336	7,248	7,491	8,161	8,578	9,000

Correlation: 0.992082

Spurious Correlation: Example 2

Divorce rate in Maine
correlates with
Per capita consumption of margarine (US)



	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Divorce rate in Maine *Divorces per 1000 people (US Census)*	5	4.7	4.6	4.4	4.3	4.1	4.2	4.2	4.2	4.1
Per capita consumption of margarine (US) *Pounds (USDA)*	8.2	7	6.5	5.3	5.2	4	4.6	4.5	4.2	3.7

Correlation: 0.992558

Spurious Correlation

- We just illustrated that two independent series (each with a stochastic trend) can produce high correlation!
- We also demonstrate that completely unrelated time series can generate high correlation.
- This is called “spurious correlation,” which in general is used to describe a situation in which correlation between two variables is driven by some underlying common driver or that the correlation be “coincidental.”
- You may find it trivial and say who would do that in practice, but you may be surprised how frequently you may encounter practitioners telling you how high the correlation is between two (trending) time, such as some revenue drivers of a company and macroeconomic time series.

Correlation Revisit: Mathematical Form

- Whenever we examine a trending time series, is correlation a good measure of the dependency of the two series?
- For that matter, is a sample mean a good measure of the "average" of a trending time series?
- Let's examine their mathematical forms of the sample estimates of mean, variance, covariance, and correlation. What is a key underlying assumption when applying these formulas?

Sample Mean: $$m(\boldsymbol{x}) = \frac{1}{n}\sum x_i, \quad m(\boldsymbol{y}) = \frac{1}{n}\sum y_i$$

Sample Variance: $$s^2(\boldsymbol{x}) = \frac{1}{n-1}\sum_{i=1}^{n}[x_i - m(\boldsymbol{x})]^2, \quad s^2(\boldsymbol{y}) = \frac{1}{n-1}\sum_{i=1}^{n}[y_i - m(\boldsymbol{y})]^2$$

Sample Covariance: $$s(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{n-1}\sum_{i=1}^{n}[x_i - m(\boldsymbol{x})][y_i - m(\boldsymbol{y})]$$

Sample Correlation: $$r(\boldsymbol{x},\boldsymbol{y}) = \frac{s(\boldsymbol{x},\boldsymbol{y})}{s(\boldsymbol{x})s(\boldsymbol{y})}$$

Correlation and Causation

- Also, as you've heard numerous times in this program, **correlation does not imply causation!**

- All of these examples clearly show that these (high) correlations really mean nothing and may even be misleading.

- However, in reality, spurious correlation between any two variables may not be as obvious. As data scientists, your bosses, clients, or colleagues may ask you to build a regression of trending time series. The concepts and techniques discussed in this lecture will be useful in dealing with these situations.

Berkeley SCHOOL OF INFORMATION