Quantitative Methods: The Quest for Causality

David A. Hughes, Ph.D.

Auburn University at Montgomery

david.hughes@aum.edu

August 19, 2021

Objectives for today

By the end of this section, students should be able to explain the following:

- The nature of political puzzles and how we use econometric models to attempt to address them, and
- The nature of causality, the factors that threaten valid causal inference, and how to best avoid these pitfalls.

How do we know what we know?

- The world is full of puzzles that cry out for answers.
- We're pretty good at spotting patterns but less talented at ignoring spurious relationships.
- In fact, it's our tendency to over-identify causal relationships that tend to get us into trouble.

An example: Individual annual income

- Suppose we want to learn about how different factors affect individual annual income.
- What are these factors, and what kinds of relationships do we expect them to have upon individual income?
- As a bit of terminology, we call these influential factors "independent variables," and we call the thing they influence the "dependent variables."
- How we measure these variables is referred to as "operationalization."

Association between variables: An economic model

- Suppose we think that education, age, and hours worked per week are the most important factors in determining one's annual income.
- Then we could represent the relationship between dependent and independent variables as:

$$\text{Income} = f(\text{education}, \text{age}, \text{hours}), \tag{1}$$

which we read as, "One's annual income is a function of one's level of education, age, and the number of hours one works per week."

- Equation 1 is an example of an economic model.

The difference between economic and econometric models

- Equation 1 is a fine example of an economic model, which one could derive via formal or informal means.
- But Equation 1 is not an econometric model as it does not specify the form of the equation nor the effect of unobserved disturbances.
- For example, how is Equation 1 able to account for something not already in it such as individual productivity?

An econometric model of individual annual income

- We can account for the ambiguities in Equation 1 by specifying an econometric model.
- As all of our variables are measured continuously, it makes sense that the form of our econometric model would be linear.
- As such let the following econometric model represent the linear effect on individual annual income:

$$\text{Income} = \beta_0 + \beta_1 \text{education} + \beta_2 \text{age} + \beta_3 \text{hours} + u, \qquad (2)$$

where each β represents the (constant) direction and size of effect each independent variable has on the dependent variable, and u represents unexplained variance (elsewhere termed the “disturbance” or “error” term).

A general form for a linear econometric model

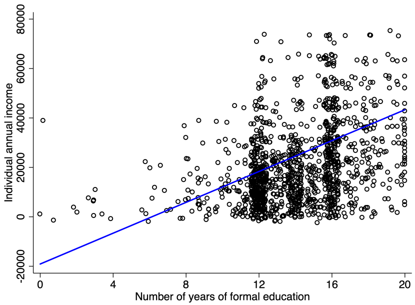
- If we let $Y \in \mathbb{R}$ represent some dependent variable and $\mathbf{X} = \{X_1, X_2, \ldots, X_k\}$ some set of independent variables, then we can generalize Equation 2 for any linear relationship:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + u_i. \tag{3}$$

- Note that the $i \in \{1, 2, \ldots, N\}$ subscript is simply an index of observations from the first (1) to last (N) observation.

A simple, bivariate example

- Suppose we just look at the relationship between education and income.
- We find: $\text{Income}_i = -7507 + 1986(\text{Educ}_i) + u_i$.
- Note: (1) The βs tell us predicted income given education, and (2) The unexplained variance, u_i tells a story about what variables have been omitted.



Source: GSS (2018)

Drawing valid causal inferences

- Generally, we can identify three criteria that, should we fulfill them, allow us to draw a causal inference. In order to conclude that $A \rightarrow B$:
  - A must precede B.
  - A must correlate with B.
  - Other factors, C, do not actually account for changes in A and B.
- Two major risks are associated with the third criterion of causality.

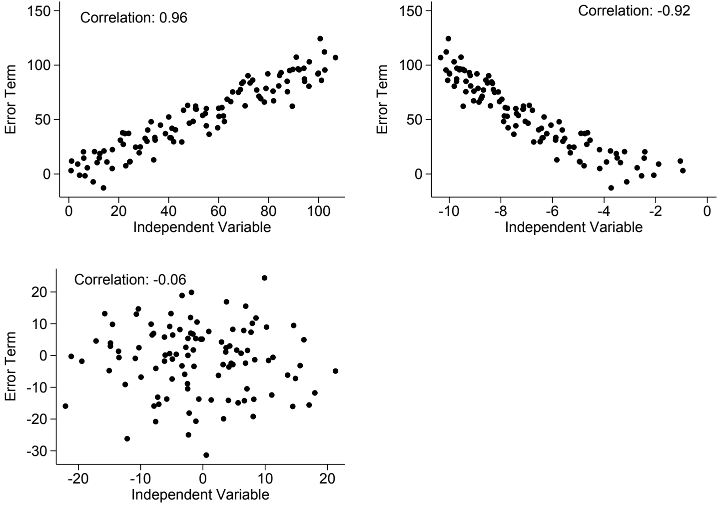
Threats to causality: Endogeneity

- If changes in the independent variables are related to changes in the disturbance, u_i, we have an endogeneity problem, which means we have a causal inference problem.
- If variance in the independent variables and u_i are exogenous, we are in a stronger position to draw a causal inference.

Assessing endogeneity

- We'd like to know how related an independent variable is with the error term, ϵ_i.
- We measure correlation on the scale of "-1" to "+1".
- Let's take a look at two endogenous and one exogenous set of relationships...

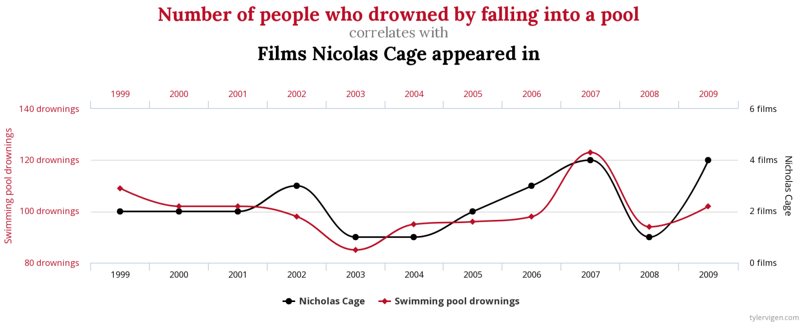
Visualizing endogenous correlations



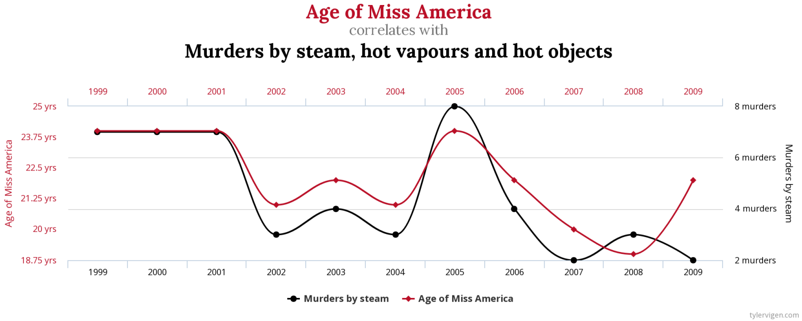
Threats to causality: randomness

- Suppose we have all the exogeneity we could ever need.
- We still must be wary of the truly random and bizarre.
- Put differently, we need a good reason to suspect that changes in Y are due to changes in X. Otherwise the relationship may be spurious.

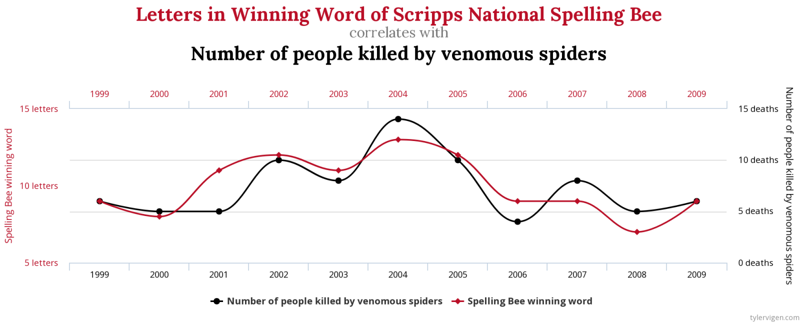
Random associations



Random associations



Random associations



Combating randomness and endogeneity

- What are the best methods we have at our disposal to combat the truly random or endogenous?
- The first is to have a sound theoretical reason for expecting A to cause B in the first place.
- The second is to have a rigorously tailored research design that uses the best practices of sampling and experimentation to arrive at sound conclusions.
- When experimental methods are not available, then our next best approach is to mimic these methods using econometric models—the subject nature of this course.

Conclusion

Today we learned a few critical things:

- We'd like to be able to make strong causal inferences.
- We can use the best practices of research design and econometric modeling to aid us in doing so.
- Even so, we must be mindful of the critical role that theory plays in helping to guide our expectations.