Data Mining

Causality

Santiago Alonso-Díaz

Tecnólogico de Monterrey
EGADE, Business School

Photo: Dalle2

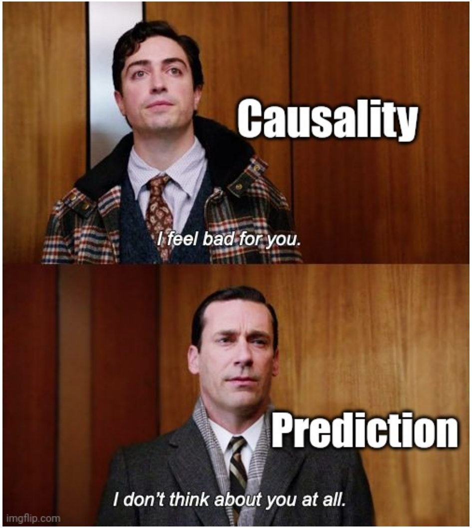


Figure: Source: Khoa Vu Twitter

Who cares?

Jonathan Larkin
@jonathanrlarkin

Unpopular opinion: Causality is **not relevant** in the majority of #quantfinance modeling applications! “Successful prediction does not require correct causal identification.”

Causal relationships are important if you want to **intervene** in a system. Quant traders are not intervening. Physicians and gov’t policy makers intervene — financial quants most often do not. Don’t believe me? Listen to the Bayesian causal GOAT:

Relevant people

Jonathan Larkin
@jonathanrlarkin

Follow

Investor @Columbia IMC; formerly CIO @quantopian, Global Head of Equities @ Millennium, Eq Derivs Trading @jpmorgan CIB | Kaggle Master | marketneutral.eth

Who cares?

- Predictions in uncertain context (e.g. stock markets) are hard to extrapolate.
- Causal relations alleviate uncertainty.
- Causal models guide interventions.
- Causal models can guide data fusion of multiple sources (Bareinboim & Pearl, 2016).

Who cares ... in business?

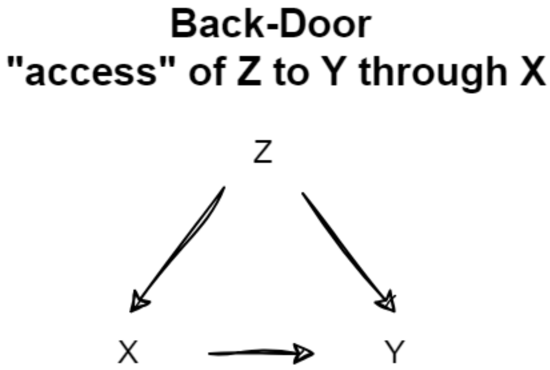
- Business intervene e.g. via ads, via revealing trading strategies, via hiding/showing info.
- Entrepreneurship, innovation, strategies, are uncertain
- Minimize resource spending
- Exploit effectively multiple data sources (Bareinboim & Pearl, 2016)

Table of contents

Concepts and diagrams

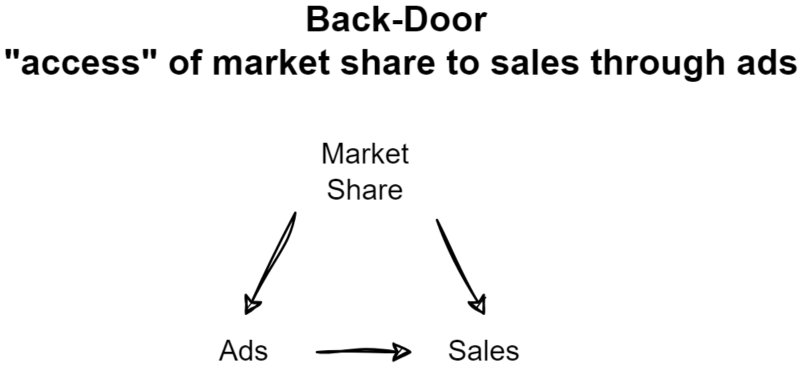
Back-door

Any path from X to Y that starts with an arrow pointing into X



Back-door

Any path from X to Y that starts with an arrow pointing into X

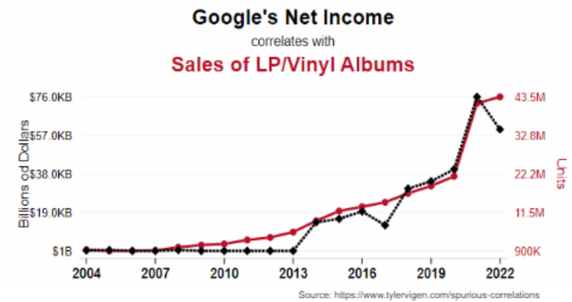


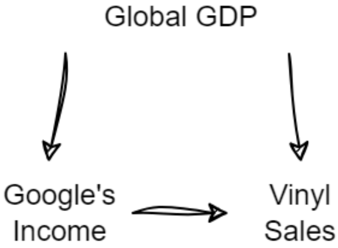
Back-door

Any path from X to Y that starts with an arrow pointing into X

Problems with back-doors

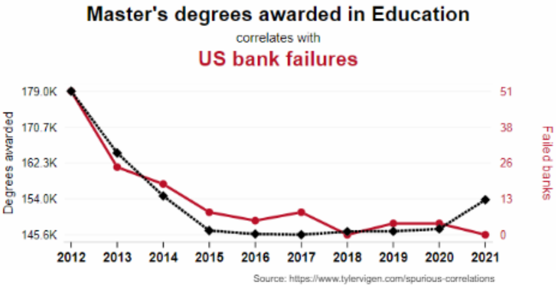
If not accounted, they enhance spurious correlations.

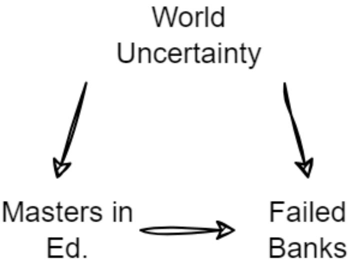




Problems with back-doors

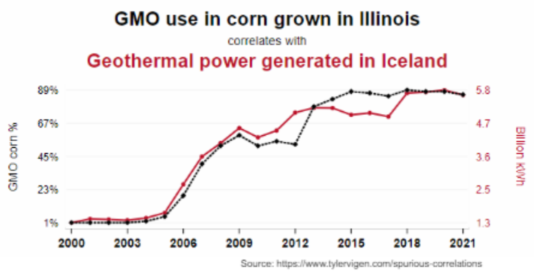
If not accounted, they enhance spurious correlations.

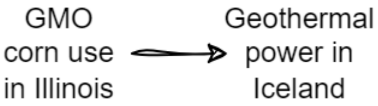


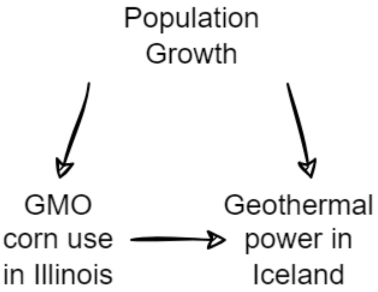


Problems with back-doors

If not accounted, they enhance spurious correlations.







Junctions

Chain

$X \rightarrow Y \rightarrow Z$

Fork

$X \leftarrow Y \rightarrow Z$

Collider

$X \rightarrow Y \leftarrow Z$

Junctions: Chain

A mechanism (need) mediates the relation with x (ad) and y (sales).

For instance, controlling-fixing for needs (mediator) cancels ad effects on sales.
Overcontrol: in this model, the only way to affect sales is through needs.

Junctions: Fork

A common cause (need) explains two downstream variables (emotions and sales)

For instance, controlling-fixing for needs (confounder) cancels any spurious correlation between emotion & sales. In this model, emotions and sales are not connected. If I do not know needs, emotions and sales would look as if related due to the common source (e.g. needs go up, both emotions and sales change).

Junctions: Collider

Two variables (emotions and needs) affect a variable (sales).

For instance, controlling-fixing for sales (collider) connects emotions and needs artificially. For a fixed level of sales, we need to "open" emotions and sales because both cause sales. If I increase emotions, I need to modify needs to obtain that fixed level of sales. This creates an illusion that they are related, but just because we fixed sales.

Issues of controlling-fixing a collider

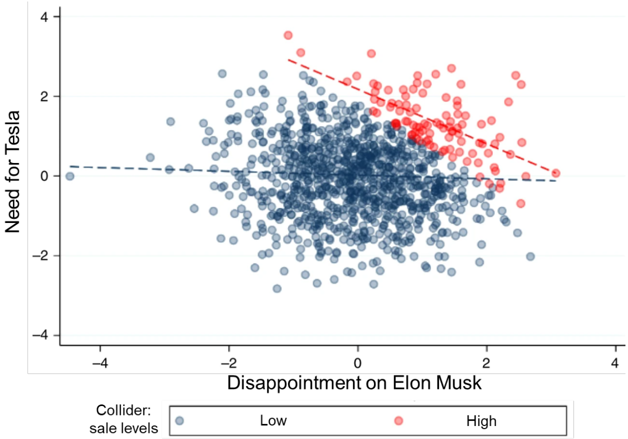


Figure: In loyal clients, needs and emotions are related (red dots). In the general population they are not (red + blue dots). Adapted from Griffith et al., 2020

Issues of controlling-fixing a collider

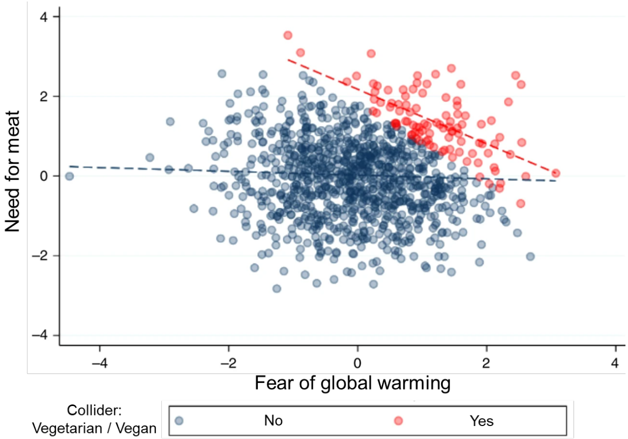


Figure: In vegans, needs and emotions are related (red dots). In the general population they are not (red + blue dots). Adapted from Griffith et al., 2020

When to control?

- Control-fix-condition on confounders to avoid omitted variable bias.
- Do not control for mediators. This could erase the path (overcontrol bias).
- Do not control for colliders due to overcontrol bias (colliders as mediators), spurious correlations, or could open a backdoor path.

Issues with the traditional regression

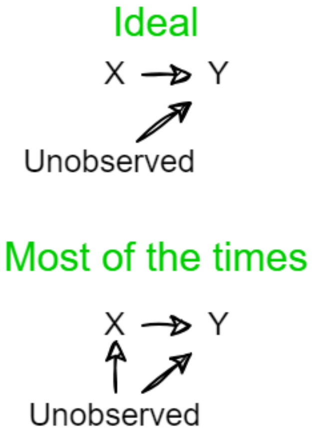


Figure: Most of the times unobserved variables have back door access to the outcome Y, biasing the effects of X (and some x may be colliders or mediators but here we assume the X of interest)

Issues with the traditional regression

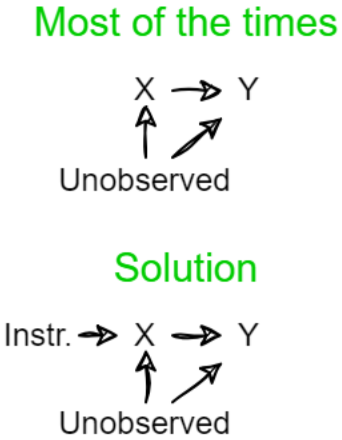


Figure: To isolate the effect of X on Y, we need an instrumental variable that changes X but not the unobserved.

Examples

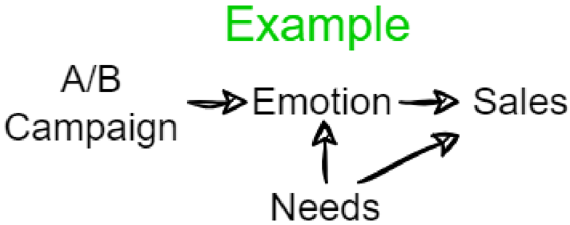


Figure: Randomization as the gold standard. Group A sees an emotional video, group B sees a plain video. Assuming perfect randomization, the effect of the videos on unobserved variables, such as needs, is similar across the A and B group. The videos, by construction, only affect the measured emotions

Examples

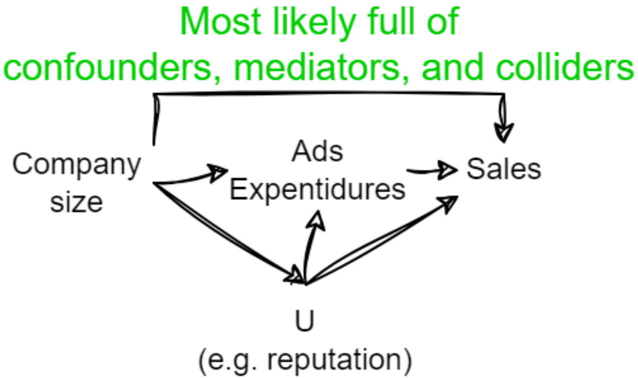


Figure: Usually we have messy (non-random) observational data to answer our questions.

Examples

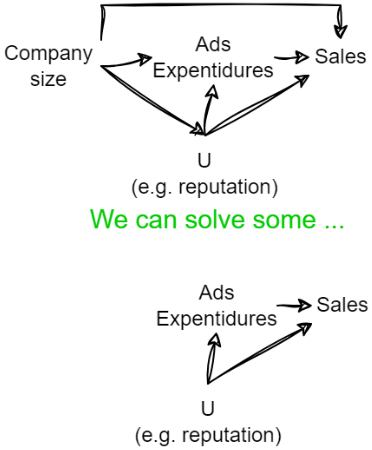


Figure: Some complexity is easier to manage. Given the position of company size in the DAG, we can separate the analysis for small and large companies.

Examples

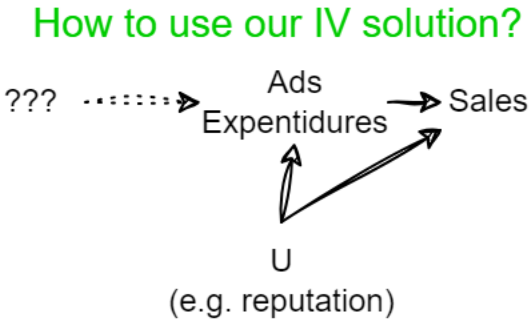


Figure: What variable affects ads expenditures but not unobserved variables or sales directly?

Examples

Across colors, people see different advertisements. We assume that at frontiers county demographics across each side are similar.

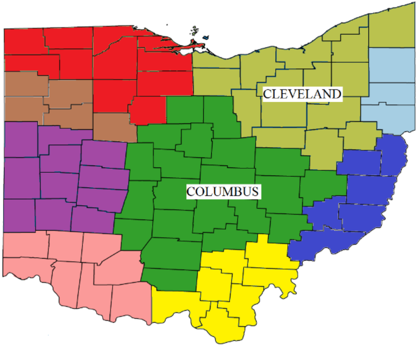


Figure: Advertising frontiers (Shapiro, 2018)

Examples

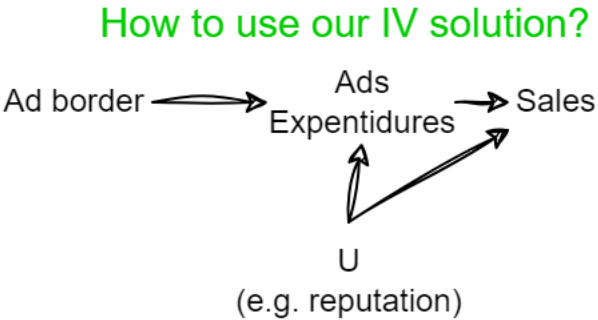


Figure: Ads change at borders. We assume that unobserved are balanced, due to county similarity. Also, no considerable exchange of info. across county's (Shapiro, 2018)

Examples

With this empirical strategy (and others), researchers demonstrated close to 0 effect of advertisements for most brands.

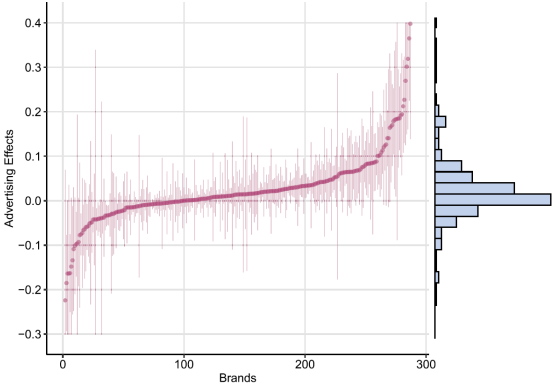


Figure: (Shapiro et al., 2021)

Discussion examples

Board independence. Is this endogenous? Bi-direction? Unobserved variables?

Figure: (Duchin et al., 2010)

Discussion examples

Is a change in law a good IV? Exclusion criteria: not correlated with Y (only through X) nor U. Relevance: correlates with X.

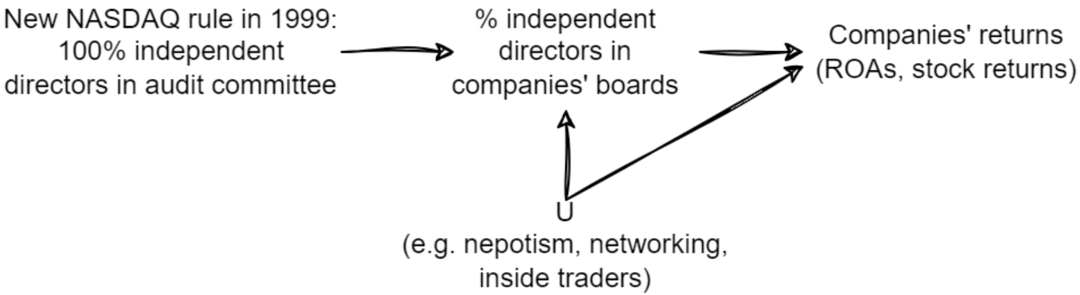


Figure: (Duchin et al., 2010)

Discussion examples

Hotel prices in Internet. Is this endogenous? Bi-direction? Unobserved variables?

$$Prices_{competition} \longrightarrow Prices_{own}$$

Figure: Li et al., 2018

Discussion examples

Is the hotel exposure in websites a good IV? Exclusion criteria: not correlated with Y (only through X) nor U. Relevance: correlates with X.

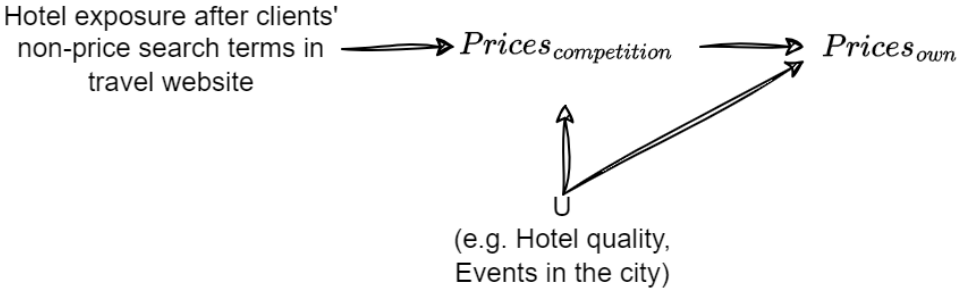


Figure: (Li et al., 2018)

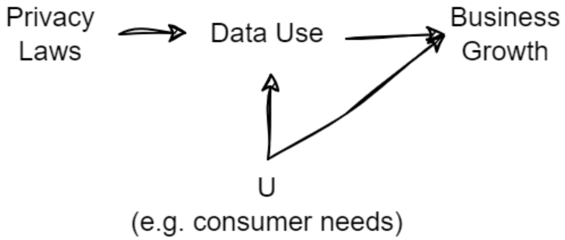
Discussion examples

Data analytics. Is this endogenous? Bi-direction? Unobserved variables?

Figure: (Li et al., 2018)

Discussion examples

Is a change in law a good IV? Exclusion criteria: not correlated with Y (only through X) nor U. Relevance: correlates with X.



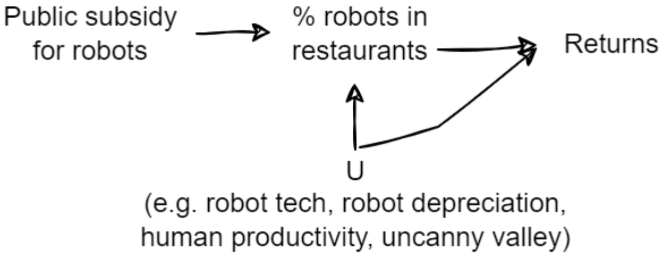
Discussion examples

Robots in restaurants. Is this endogenous? Bi-direction? Unobserved variables?

Figure: (Li et al., 2018)

Discussion examples

Is a government subsidy a good IV? Exclusion criteria: not correlated with Y (only through X) nor U. Relevance: correlates with X.



Python

Python

Let's see IV and other techniques in Python DM_Causality.ipynb

References

Bareinboim, E., & Pearl, J. (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, *113*(27), 7345–7352.

Duchin, R., Matsusaka, J. G., & Ozbas, O. (2010). When are outside directors effective? *Journal of financial economics*, *96*(2), 195–214.

Griffith, G. J., Morris, T. T., Tudball, M. J., Herbert, A., Mancano, G., Pike, L., Sharp, G. C., Sterne, J., Palmer, T. M., Davey Smith, G., et al. (2020). Collider bias undermines our understanding of covid-19 disease risk and severity. *Nature communications*, *11*(1), 5749.

Li, J., Netessine, S., & Koulayev, S. (2018). Price to compete... with many: How to identify price competition in high-dimensional space. *Management Science*, *64*(9), 4118–4136.

Shapiro, B. T. (2018). Positive spillovers and free riding in advertising of prescription pharmaceuticals: The case of antidepressants. *Journal of political economy*, *126*(1), 381–437.

Shapiro, B. T., Hitsch, G. J., & Tuchman, A. E. (2021). Tv advertising effectiveness and profitability: Generalizable results from 288 brands. *Econometrica*, *89*(4), 1855–1879.