KAIST Summer Session 2018

Module 1. Research Design for Data Analytics

The Right Tool for the Right Question

KAIST College of Business

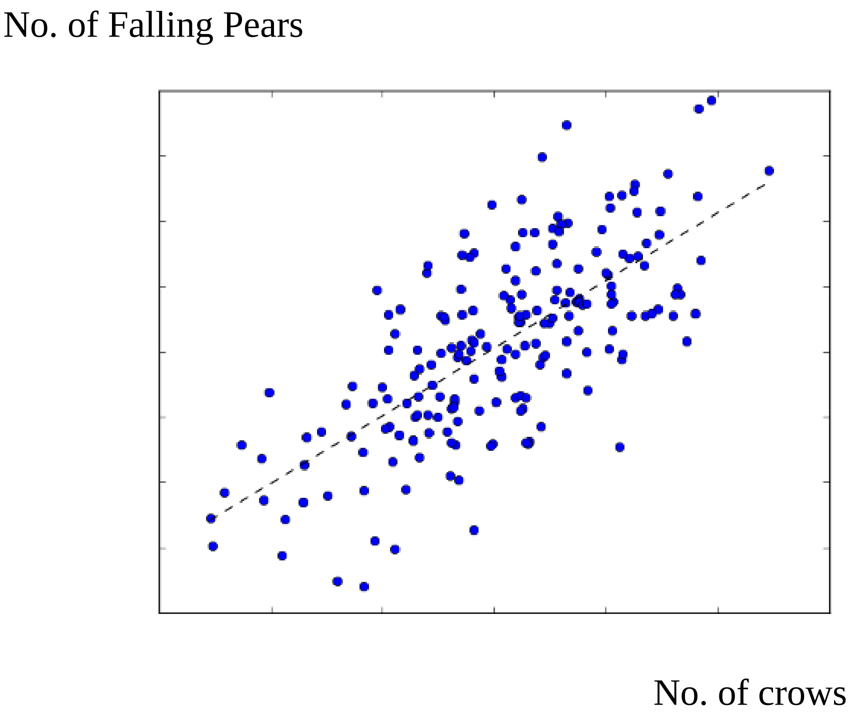
Jiyong Park

28 June, 2018

Why Does “Causality” Matter?

Why Does “Causality” Matter?

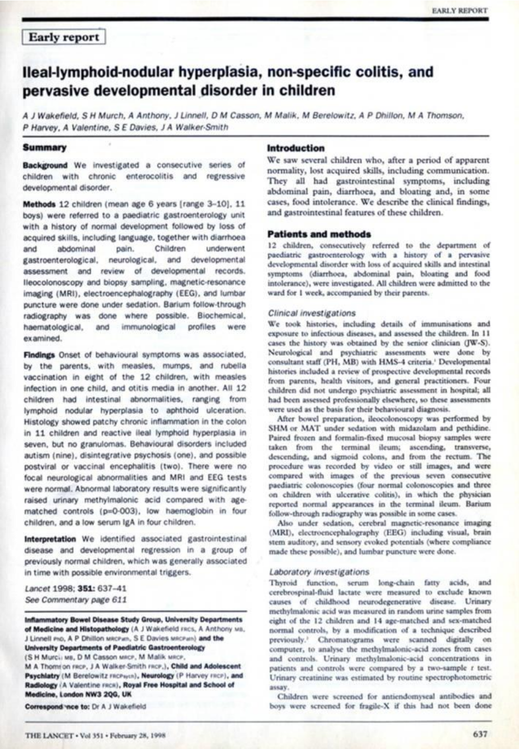
- “까마귀 날자 배 떨어진다.”



Do crows cause to make pears fall?

Why Does “Causality” Matter?

- Is the use of MMR—the vaccine that inoculates children against measles—responsible for the increase in reported autism cases?



EARLY REPORT

Early report

Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children

A J Wakefield, S H Murch, A Anthony, J Linnell, D M Casson, M Malik, M Berelowitz, A P Dhillon, M A Thomson, P Harvey, A Valentine, S E Davies, J A Walker-Smith

Summary

Background We investigated a consecutive series of children with chronic enterocolitis and regressive developmental disorder.

Methods 12 children (mean age 6 years [range 3–10], 11 boys) were referred to a paediatric gastroenterology unit with a history of normal development followed by loss of acquired skills, including language, together with diarrhoea and abdominal pain. Children underwent gastroenterological, neurological, and developmental assessment and review of developmental records. Ileocolonoscopy and biopsy sampling, magnetic-resonance imaging (MRI), electroencephalography (EEG), and lumbar puncture were done under sedation. Barium follow-through radiography was done where possible. Biochemical, haematological, and immunological profiles were examined.

Findings Onset of behavioural symptoms was associated, by the parents, with measles, mumps, and rubella vaccination in eight of the 12 children, with measles infection in one child, and otitis media in another. All 12 children had intestinal abnormalities, ranging from lymphoid nodular hyperplasia to aphthoid ulceration. Histology showed patchy chronic inflammation in the colon in 11 children and reactive ileal lymphoid hyperplasia in seven, but no granulomas. Behavioural disorders included autism (nine), disintegrative psychosis (one), and possible postviral or vaccinal encephalitis (two). There were no focal neurological abnormalities and MRI and EEG tests were normal. Abnormal laboratory results were significantly raised urinary methylmalonic acid compared with age-matched controls ($p=0·003$), low haemoglobin in four children, and a low serum IgA in four children.

Interpretation We identified associated gastrointestinal disease and developmental regression in a group of previously normal children, which was generally associated in time with possible environmental triggers.

Lancet 1998; **351:** 637–41
See Commentary page 611

Inflammatory Bowel Disease Study Group, University Departments of Medicine and Histopathology (A J Wakefield FRCS, A Anthony MB, J Linnell PhD, A P Dhillon MRCPath, S E Davies MRCPath) **and the University Departments of Paediatric Gastroenterology** (S H Murch MB, D M Casson MRCP, M Malik MRCP, M A Thomson FRCP, J A Walker-Smith FRCP,), **Child and Adolescent Psychiatry** (M Berelowitz FRCPsych), **Neurology** (P Harvey FRCP), **and Radiology** (A Valentine FRCR), **Royal Free Hospital and School of Medicine, London NW3 2QG, UK**

Correspondence to: Dr A J Wakefield

Introduction

We saw several children who, after a period of apparent normality, lost acquired skills, including communication. They all had gastrointestinal symptoms, including abdominal pain, diarrhoea, and bloating and, in some cases, food intolerance. We describe the clinical findings, and gastrointestinal features of these children.

Patients and methods

12 children, consecutively referred to the department of paediatric gastroenterology with a history of a pervasive developmental disorder with loss of acquired skills and intestinal symptoms (diarrhoea, abdominal pain, bloating and food intolerance), were investigated. All children were admitted to the ward for 1 week, accompanied by their parents.

Clinical investigations

We took histories, including details of immunisations and exposure to infectious diseases, and assessed the children. In 11 cases the history was obtained by the senior clinician (JW-S). Neurological and psychiatric assessments were done by consultant staff (PH, MB) with HMS-4 criteria.[1] Developmental histories included a review of prospective developmental records from parents, health visitors, and general practitioners. Four children did not undergo psychiatric assessment in hospital; all had been assessed professionally elsewhere, so these assessments were used as the basis for their behavioural diagnosis.

After bowel preparation, ileocolonoscopy was performed by SHM or MAT under sedation with midazolam and pethidine. Paired frozen and formalin-fixed mucosal biopsy samples were taken from the terminal ileum; ascending, transverse, descending, and sigmoid colons, and from the rectum. The procedure was recorded by video or still images, and were compared with images of the previous seven consecutive paediatric colonoscopies (four normal colonoscopies and three on children with ulcerative colitis), in which the physician reported normal appearances in the terminal ileum. Barium follow-through radiography was possible in some cases.

Also under sedation, cerebral magnetic-resonance imaging (MRI), electroencephalography (EEG) including visual, brain stem auditory, and sensory evoked potentials (where compliance made these possible), and lumbar puncture were done.

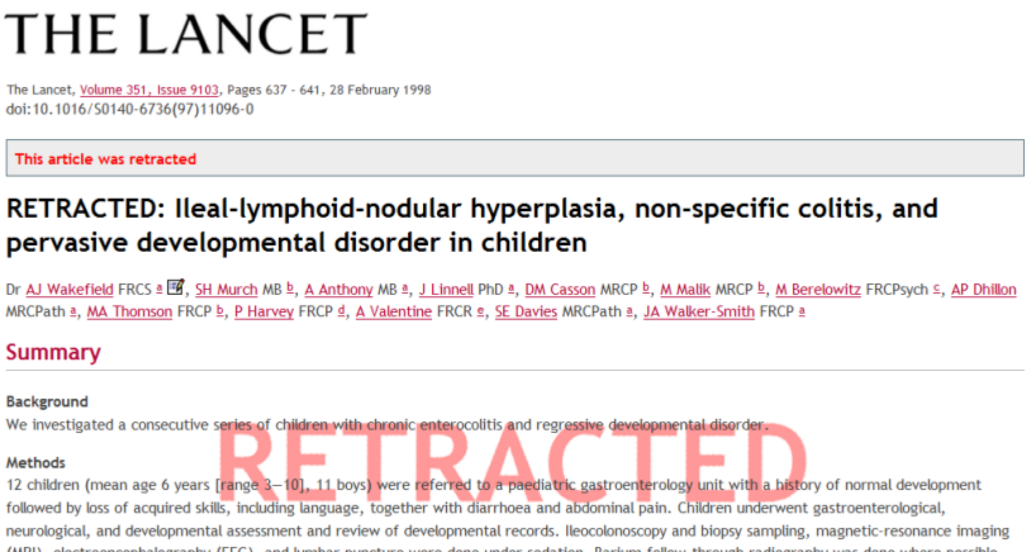
Laboratory investigations

Thyroid function, serum long-chain fatty acids, and cerebrospinal-fluid lactate were measured to exclude known causes of childhood neurodegenerative disease. Urinary methylmalonic acid was measured in random urine samples from eight of the 12 children and 14 age-matched and sex-matched normal controls, by a modification of a technique described previously.[2] Chromatograms were scanned digitally on computer, to analyse the methylmalonic-acid zones from cases and controls. Urinary methylmalonic-acid concentrations in patients and controls were compared by a two-sample t test. Urinary creatinine was estimated by routine spectrophotometric assay.

Children were screened for antiendomyseal antibodies and boys were screened for fragile-X if this had not been done

THE LANCET • Vol 351 • February 28, 1998

637



THE LANCET

The Lancet, Volume 351, Issue 9103, Pages 637 - 641, 28 February 1998
doi:10.1016/S0140-6736(97)11096-0

This article was retracted

RETRACTED: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children

Dr AJ Wakefield FRCS [a], SH Murch MB [b], A Anthony MB [a], J Linnell PhD [a], DM Casson MRCP [b], M Malik MRCP [b], M Berelowitz FRCPsych [c], AP Dhillon MRCPath [a], MA Thomson FRCP [b], P Harvey FRCP [d], A Valentine FRCR [e], SE Davies MRCPath [a], JA Walker-Smith FRCP [a]

Summary

Background

We investigated a consecutive series of children with chronic enterocolitis and regressive developmental disorder.

Methods

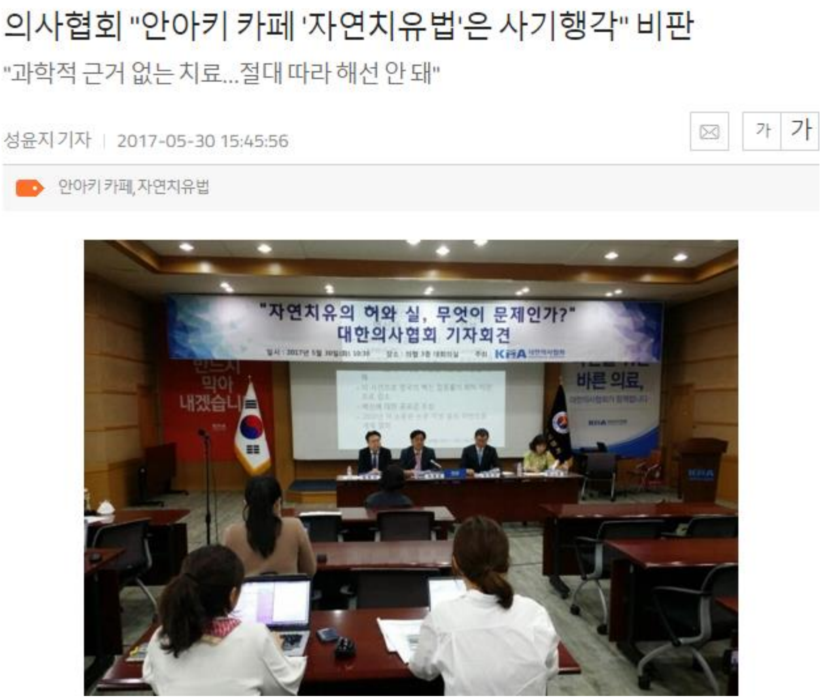
12 children (mean age 6 years [range 3–10], 11 boys) were referred to a paediatric gastroenterology unit with a history of normal development followed by loss of acquired skills, including language, together with diarrhoea and abdominal pain. Children underwent gastroenterological, neurological, and developmental assessment and review of developmental records. Ileocolonoscopy and biopsy sampling, magnetic-resonance imaging (MRI), electroencephalography (EEG), and lumbar puncture were done under sedation. Barium follow-through radiography was done where possible

The Lancet later retracted the paper with a further investigation uncovering falsification of data.

Wakefield, A. J. et al. 1998. RETRACTED: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *The Lancet*, 9103(351), 637-641.

Why Does "Causality" Matter?

- Is the use of MMR—the vaccine that inoculates children against measles—responsible for the increase in reported autism cases?



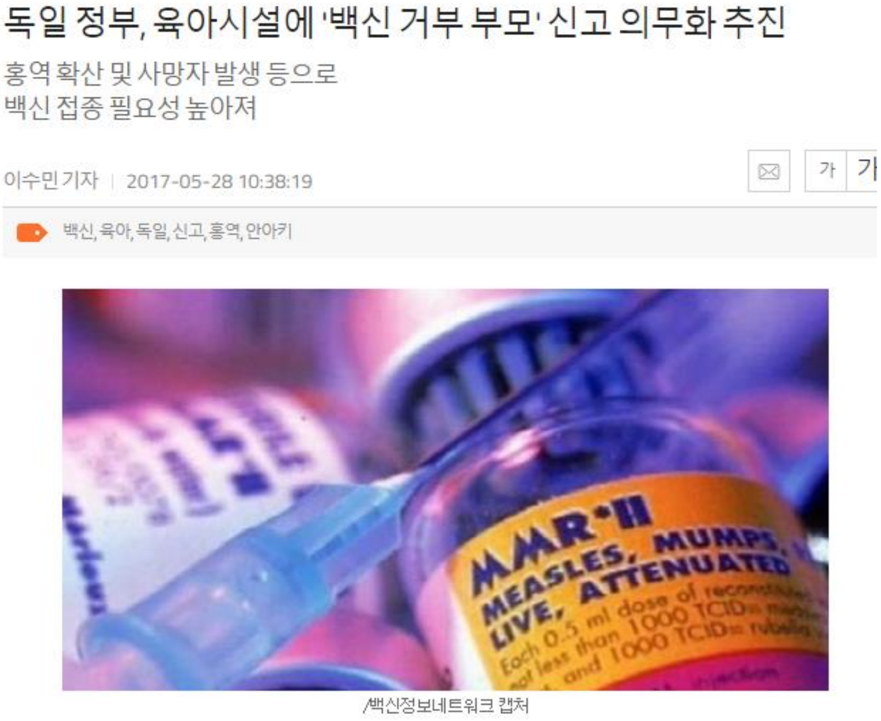
의사협회 "안아키 카페 '자연치유법'은 사기행각" 비판

"과학적 근거 없는 치료...절대 따라 해선 안 돼"

성윤지 기자 | 2017-05-30 15:45:56

안아키 카페, 자연치유법

서울경제 (2017.05.30)



독일 정부, 육아시설에 '백신 거부 부모' 신고 의무화 추진

홍역 확산 및 사망자 발생 등으로
백신 접종 필요성 높아져

이수민 기자 | 2017-05-28 10:38:19

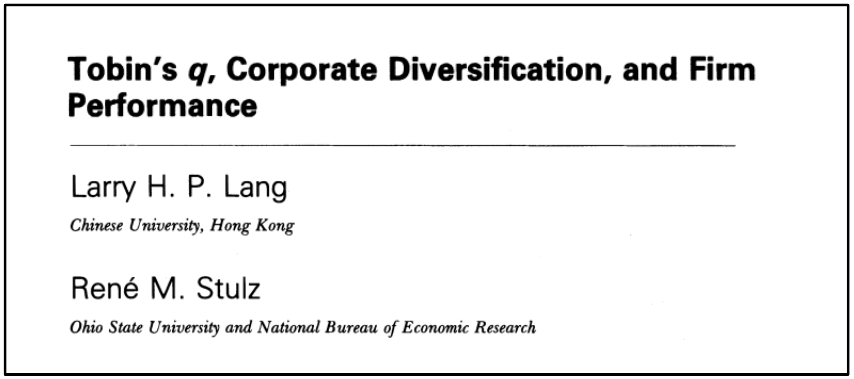
백신, 육아, 독일, 신고, 홍역, 안아키

/백신정보네트워크 캡처

서울경제 (2017.05.28)

Why Does “Causality” Matter?

- Does diversification destroy or enhance firm value?



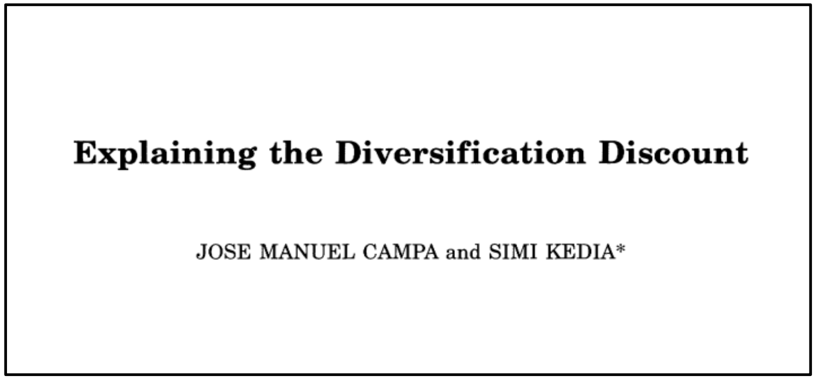
Tobin’s *q*, Corporate Diversification, and Firm Performance

Larry H. P. Lang

Chinese University, Hong Kong

René M. Stulz

Ohio State University and National Bureau of Economic Research



Explaining the Diversification Discount

JOSE MANUEL CAMPA and SIMI KEDIA*

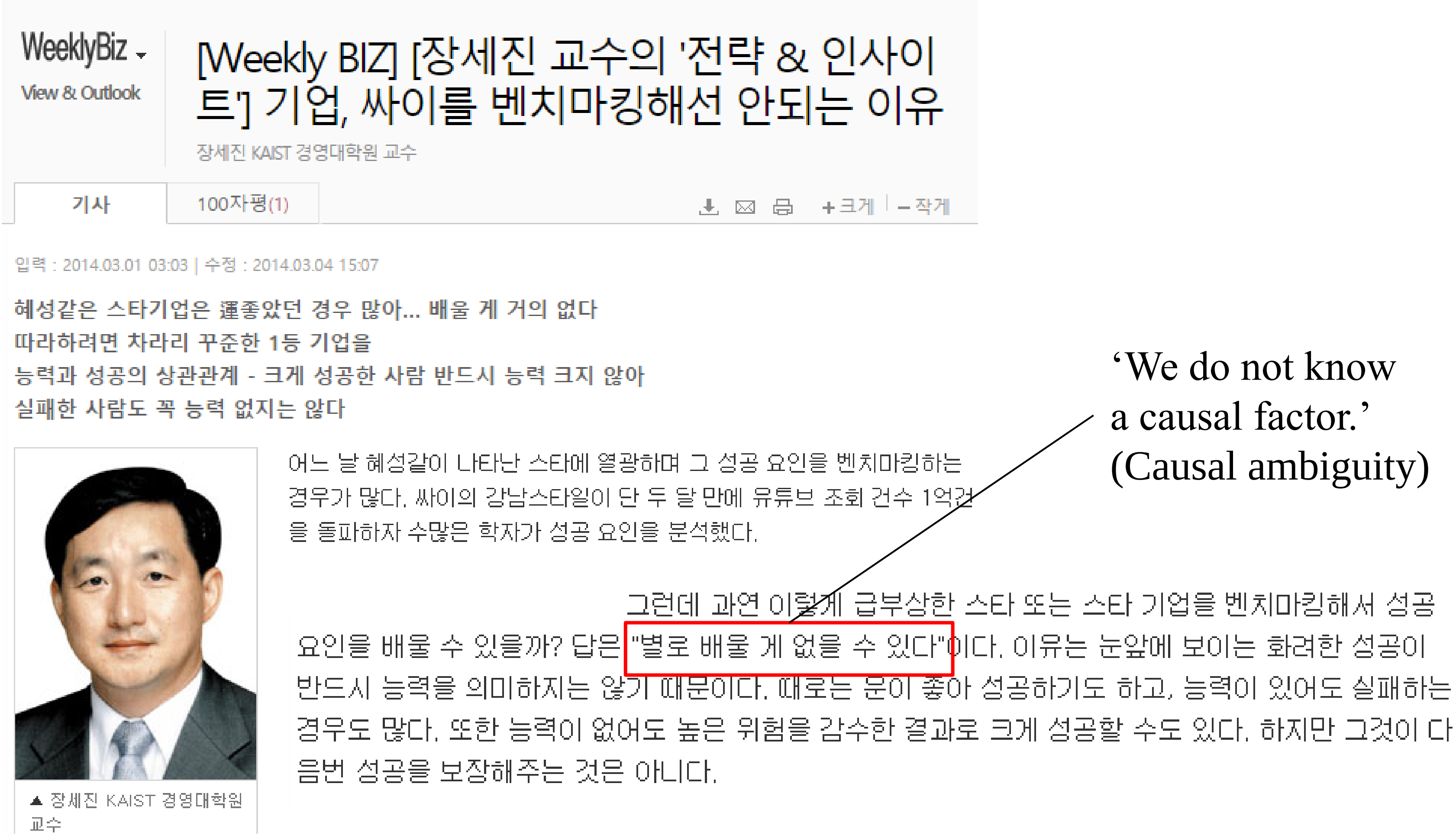
“Tobin’s q and firm diversification are negatively related throughout the 1980s.”

“We use three alternative econometric techniques to control for the endogeneity of the diversification decision…
the diversification discount always drops, and sometimes turns into a premium.”

Lang, L. H., & Stulz, R. M. 1994. Tobin's q, Corporate Diversification, and Firm Performance. *Journal of Political Economy*, *102*(6), 1248-1280.

Campa, J. M., & Kedia, S. 2002. Explaining the Diversification Discount. *Journal of Finance*, *57*(4), 1731-1762.

Why Does “Causality” Matter?



WeeklyBiz
View & Outlook

[Weekly BIZ] [장세진 교수의 '전략 & 인사이트'] 기업, 싸이를 벤치마킹해선 안되는 이유

장세진 KAIST 경영대학원 교수

입력 : 2014.03.01 03:03 | 수정 : 2014.03.04 15:07

혜성같은 스타기업은 運좋았던 경우 많아... 배울 게 거의 없다
따라하려면 차라리 꾸준한 1등 기업을
능력과 성공의 상관관계 - 크게 성공한 사람 반드시 능력 크지 않아
실패한 사람도 꼭 능력 없지는 않다

▲ 장세진 KAIST 경영대학원 교수

어느 날 혜성같이 나타난 스타에 열광하며 그 성공 요인을 벤치마킹하는 경우가 많다. 싸이의 강남스타일이 단 두 달 만에 유튜브 조회 건수 1억건을 돌파하자 수많은 학자가 성공 요인을 분석했다.

그런데 과연 이렇게 급부상한 스타 또는 스타 기업을 벤치마킹해서 성공 요인을 배울 수 있을까? 답은 "별로 배울 게 없을 수 있다"이다. 이유는 눈앞에 보이는 화려한 성공이 반드시 능력을 의미하지는 않기 때문이다. 때로는 운이 좋아 성공하기도 하고, 능력이 있어도 실패하는 경우도 많다. 또한 능력이 없어도 높은 위험을 감수한 결과로 크게 성공할 수도 있다. 하지만 그것이 다음번 성공을 보장해주는 것은 아니다.

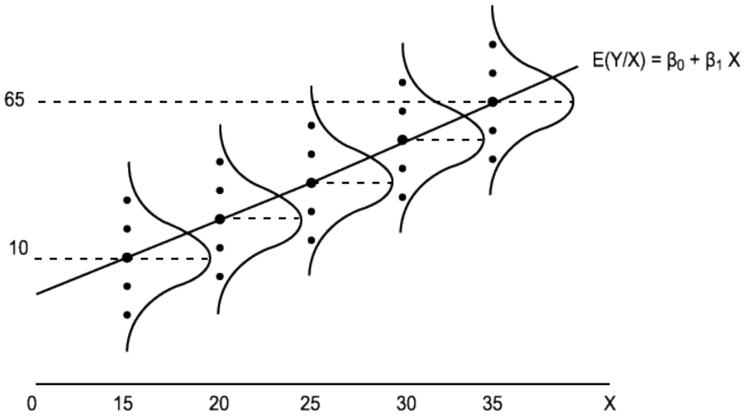
‘We do not know a causal factor.’
(Causal ambiguity)

조선비즈 (2014.03.04) http://biz.chosun.com/site/data/html_dir/2014/02/28/2014022801772.html

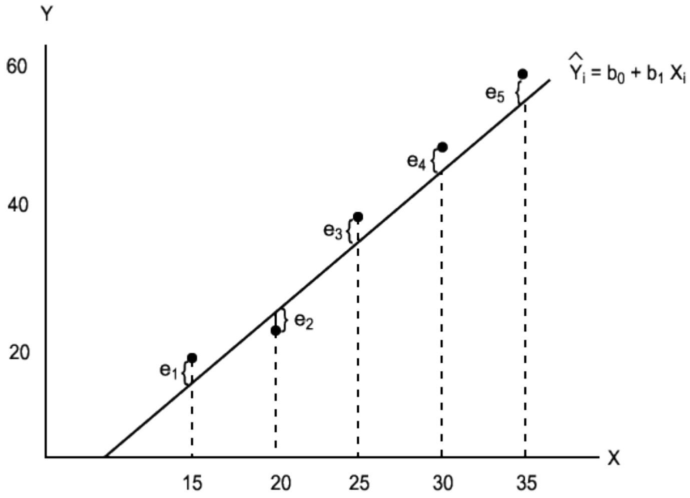
Regression and Causal Inference

Statistical Inference

- We are interested in b_1 (estimated from sample), thereby ultimately β_1 (true value in entire population).



Population



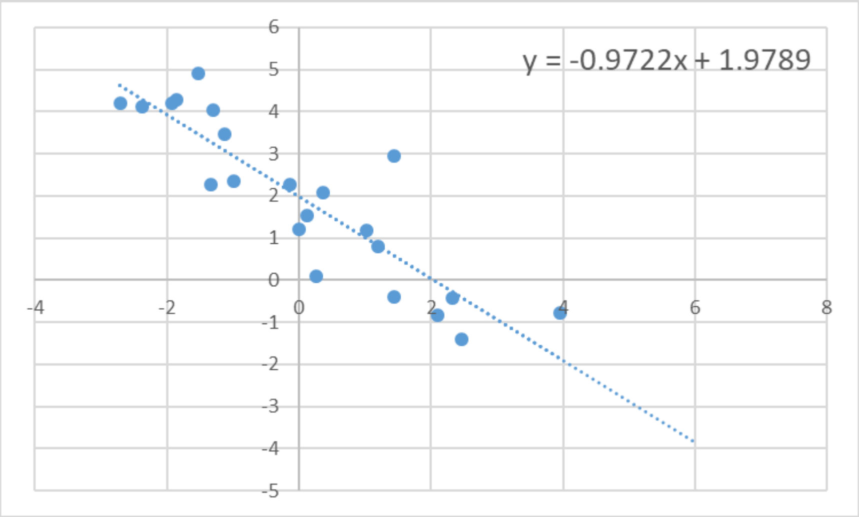
Sample

- How can we infer β_1 from b_1?
  - ➢ Collecting infinite samples
  - ➢ **Assuming about the unobserved (i.e., error term)**

Regression Analysis

- Regression is not a point estimator, rather it is an interval estimator.
  - (1) coefficients
  - (2) standard deviation of coefficients

- Ordinary Least Square (OLS)
  - Simplest and most powerful linear estimator

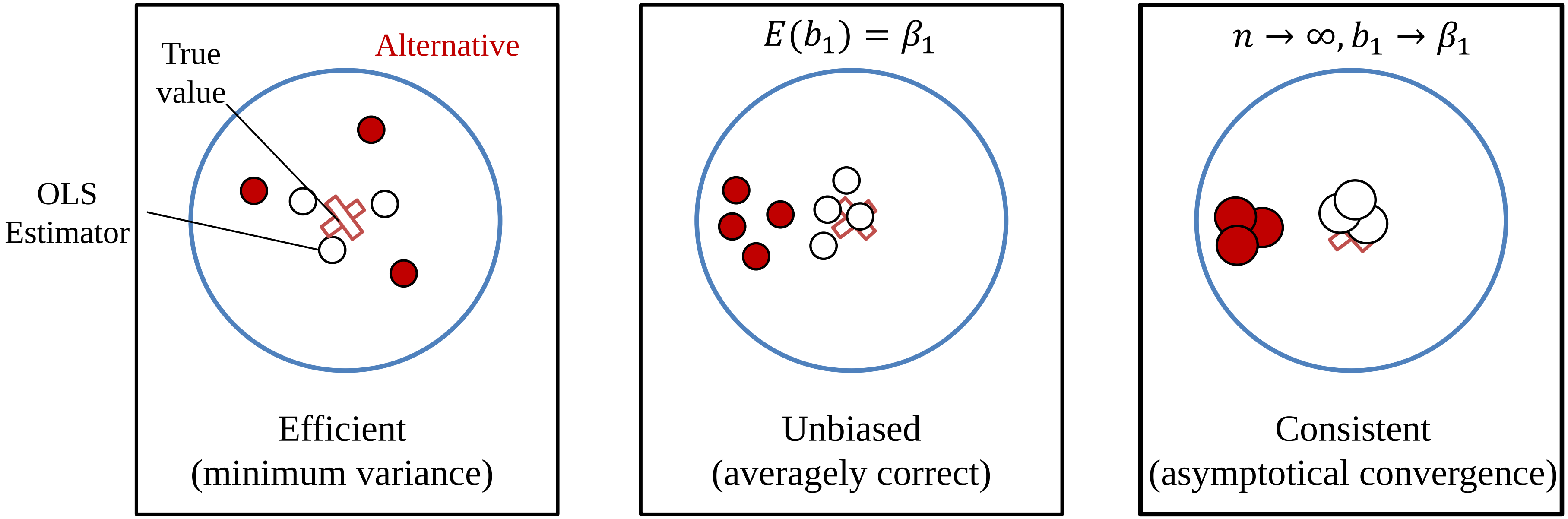


$$\hat{\beta}_{yx} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Why is OLS Ideal? (= Why is OLS not Realistic?)

- Gauss-Markov theorem
  - ➢ Under the Gauss-Markov assumptions, the OLS estimators will be BLUE (Best Linear Unbiased Estimator).

- What does BLUE mean? (OLS_Simulation.xlsx)



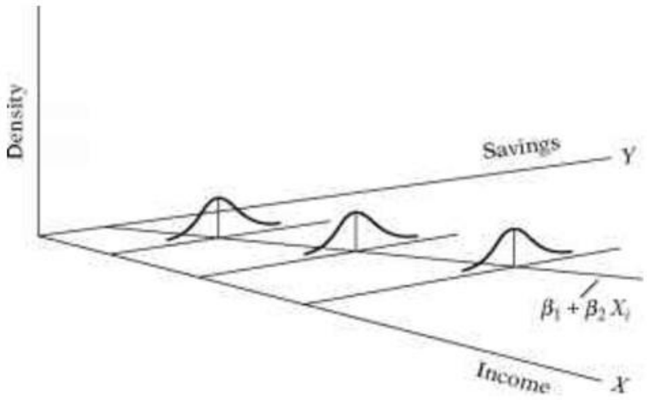
Efficient (minimum variance)

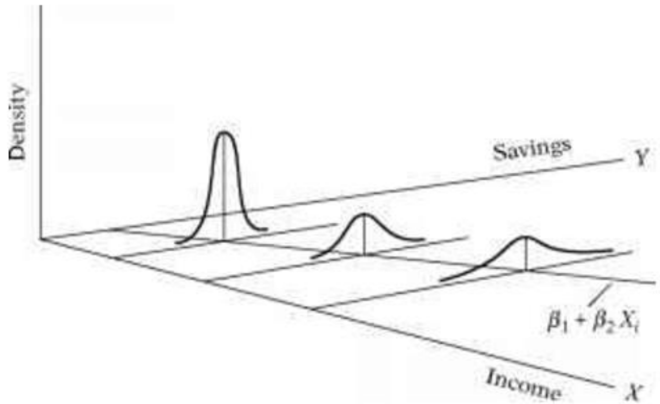
Unbiased (averagely correct)

Consistent (asymptotical convergence)

Why is OLS Ideal? (= Why is OLS not Realistic?)

- Gauss-Markov assumptions
  - (1) Linear relationship
  - (2) No perfect-multicollinearity
  - (3) Homoscedasticity (↔ heteroscedasticity)
    - $Var(\varepsilon_i) = \sigma^2 (constant)$
  - (4) No autocorrelation (serial correlation)
    - $Cov(\varepsilon_i, \varepsilon_j) = 0$
  - (5) Exogeneity (↔ endogeneity)
    - $Cov(X_i, \varepsilon_i) = 0$

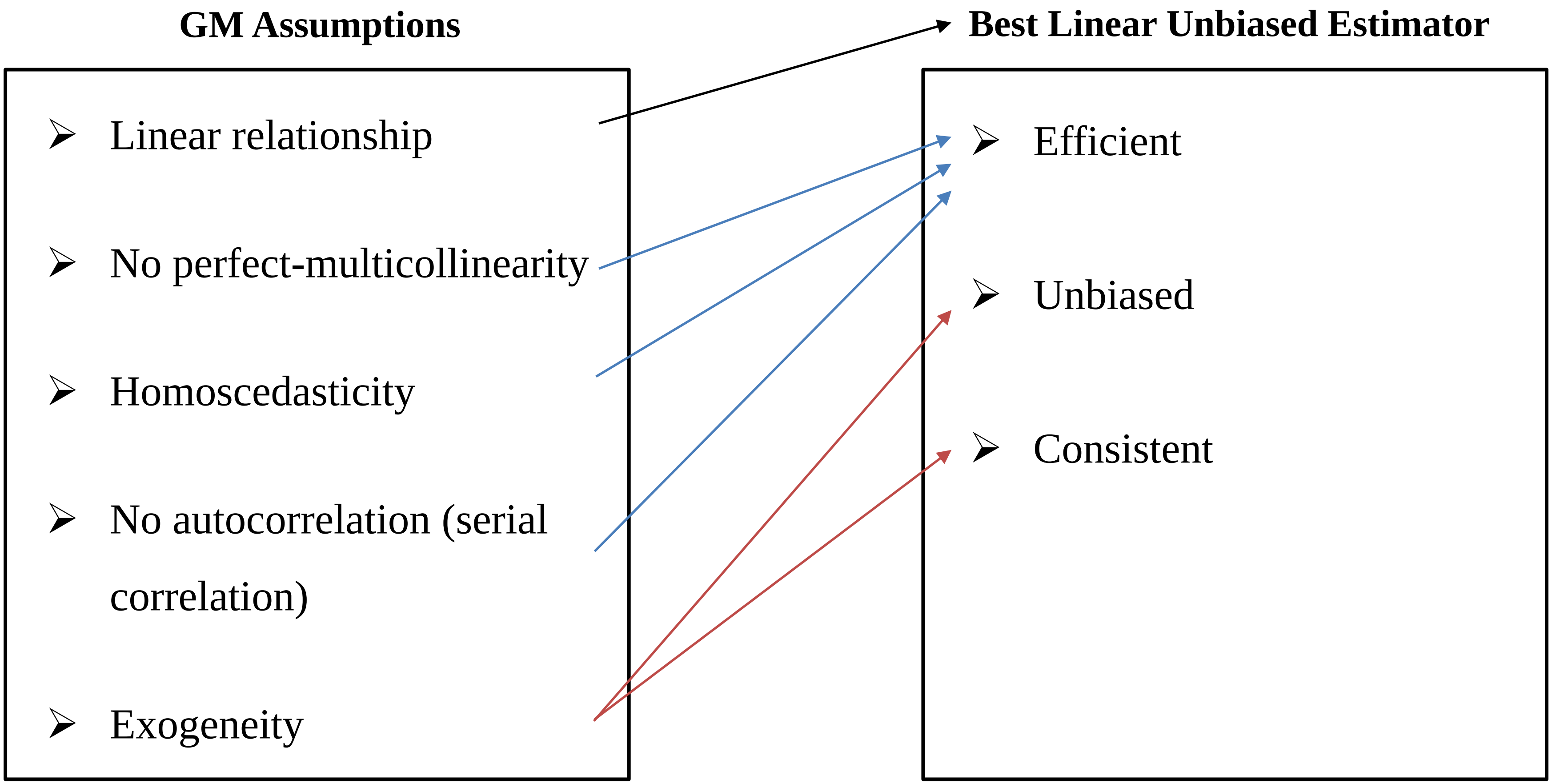




KAIST COLLEGE OF BUSINESS

What Happens if GM Assumptions are Violated?

- GM assumptions violations (OLS_Simulation.xlsx)



GM Assumptions

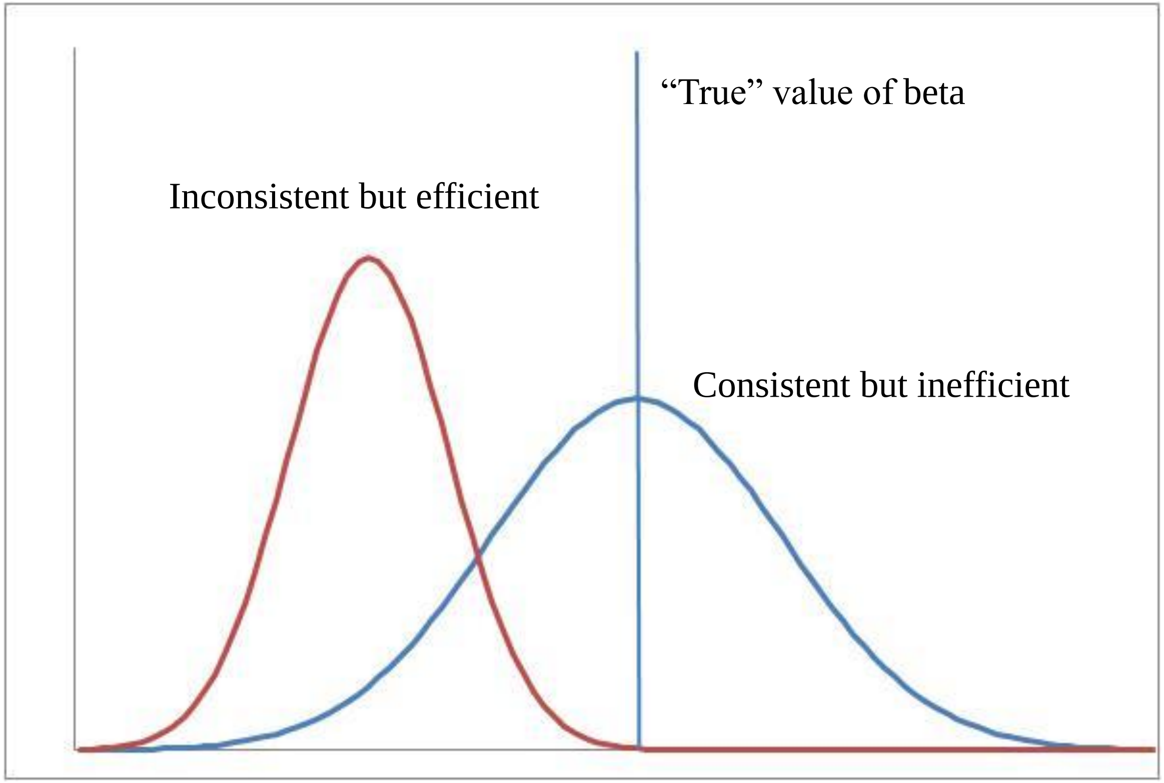
- Linear relationship
- No perfect-multicollinearity
- Homoscedasticity
- No autocorrelation (serial correlation)
- Exogeneity

Best Linear Unbiased Estimator

- Efficient
- Unbiased
- Consistent

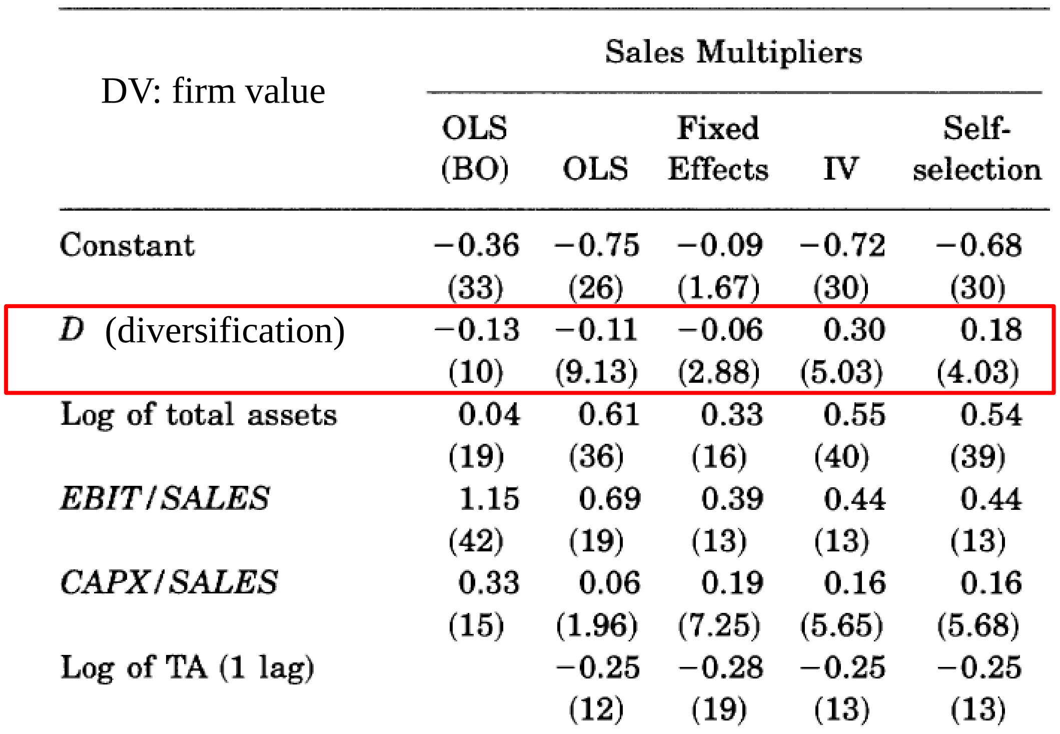
What Happens if GM Assumptions are Violated?

- Which problems are more serious in terms of causal inference?



What Happens if GM Assumptions are Violated?

- Which problems are more serious in terms of causal inference?
  - ➢ (Example) How endogeneity could mislead the conclusions



DV: firm value	Sales Multipliers				
	OLS (BO)	OLS	Fixed Effects	IV	Self-selection
Constant	−0.36	−0.75	−0.09	−0.72	−0.68
	(33)	(26)	(1.67)	(30)	(30)
D (diversification)	−0.13	−0.11	−0.06	0.30	0.18
	(10)	(9.13)	(2.88)	(5.03)	(4.03)
Log of total assets	0.04	0.61	0.33	0.55	0.54
	(19)	(36)	(16)	(40)	(39)
EBIT/SALES	1.15	0.69	0.39	0.44	0.44
	(42)	(19)	(13)	(13)	(13)
CAPX/SALES	0.33	0.06	0.19	0.16	0.16
	(15)	(1.96)	(7.25)	(5.65)	(5.68)
Log of TA (1 lag)		−0.25	−0.28	−0.25	−0.25
		(12)	(19)	(13)	(13)

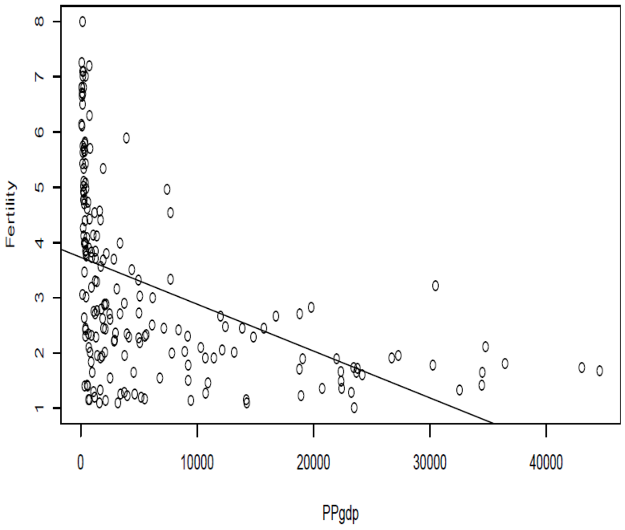
"We use three alternative econometric techniques to control for the endogeneity of the diversification decision… the diversification discount always drops, and sometimes turns into a premium." (Campa & Kedia 2002)

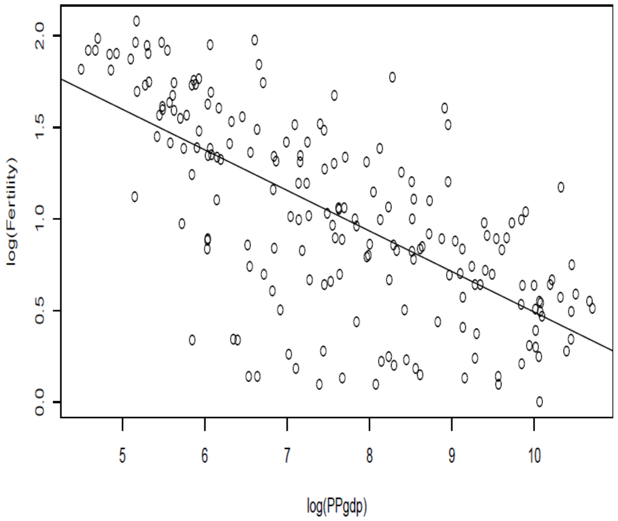
Campa, J.M. and Kedia, S., 2002. Explaining the Diversification Discount. *Journal of Finance*, 57(4), pp.1731-1762.

Remedy for GM Violations

(1) Linearity

- Why is log-specification preferred?
  - (1) To normalize the scale of variables
  - (2) (More importantly) To maintain a linear relationship
    - "the download data are heavily skewed, and hence log transformation provides a better fit." (Danaher et al. 2010, p. 1145)





(1) Linearity

- Why is log-specification preferred?
  - (3) For ease of interpretation

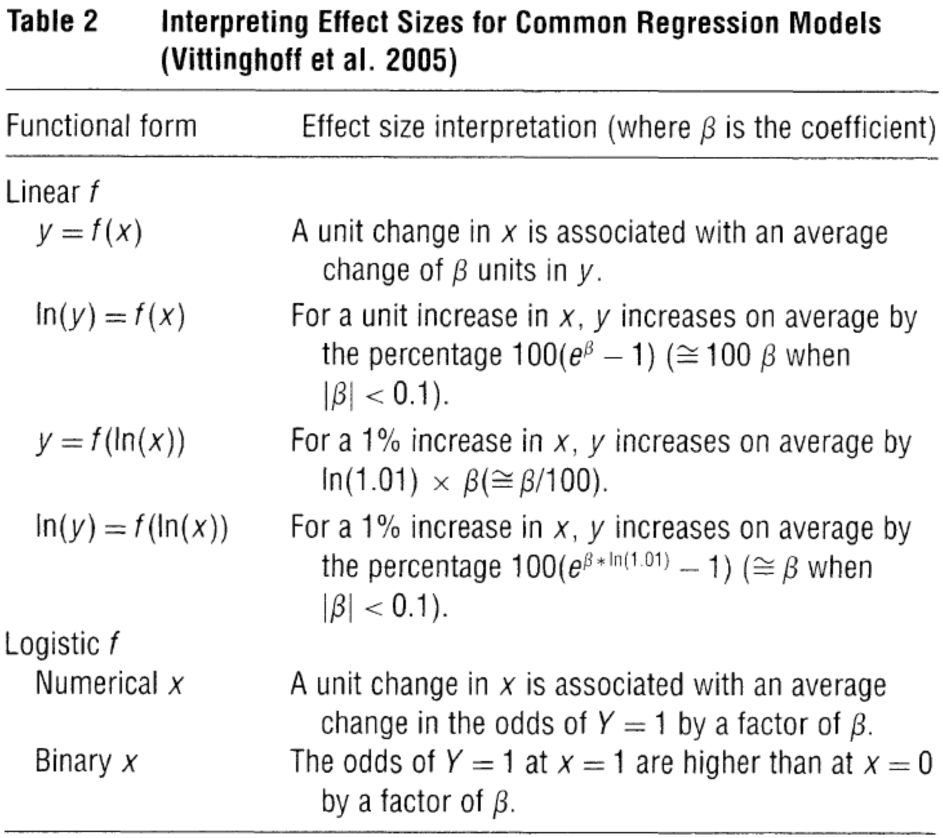


Table 2 Interpreting Effect Sizes for Common Regression Models (Vittinghoff et al. 2005)

Functional form	Effect size interpretation (where β is the coefficient)
Linear f	
$y = f(x)$	A unit change in x is associated with an average change of β units in y.
$\ln(y) = f(x)$	For a unit increase in x, y increases on average by the percentage $100(e^{\beta} - 1)$ ($\cong 100\,\beta$ when $\lvert\beta\rvert < 0.1$).
$y = f(\ln(x))$	For a 1% increase in x, y increases on average by $\ln(1.01) \times \beta (\cong \beta/100)$.
$\ln(y) = f(\ln(x))$	For a 1% increase in x, y increases on average by the percentage $100(e^{\beta * \ln(1.01)} - 1)$ ($\cong \beta$ when $\lvert\beta\rvert < 0.1$).
Logistic f	
Numerical x	A unit change in x is associated with an average change in the odds of $Y = 1$ by a factor of β.
Binary x	The odds of $Y = 1$ at $x = 1$ are higher than at $x = 0$ by a factor of β.

Lin, M., Lucas Jr, H.C. and Shmueli, G., 2013. Too Big to Fail: Large Samples and the p-Value Problem. *Information Systems Research*, 24(4), pp.906-917.

(2) Heteroscedasticity & Autocorrelation

- How to address heteroscedasticity and autocorrelation
  - ➢ (1) Linearly transforming the model
    - (Feasible) generalized least square (FGLS)
  - ➢ (2) Correcting standard errors
    - **Robust (clustered) standard errors** *(most common procedure)*

      "We estimate logit models using heteroscedasticity-consistent robust standard errors." (Lin and Viswanathan 2016, p. 1399)

      "To reduce heteroscedasticity concerns, we leverage robust standard errors clustered at the county level." (Greenwood and Wattal 2016, p. 170)
    - OLS with panel corrected standard errors (PCSE)
    - OLS with Driscoll-Kraay standard errors

(3) Multicollinearity

- How to detect multicollinearity
  - ➢ Statistically, check if variance inflation factor (VIF) is above 10.
  - ➢ In practice, it may be not a primary issue for causal inference.

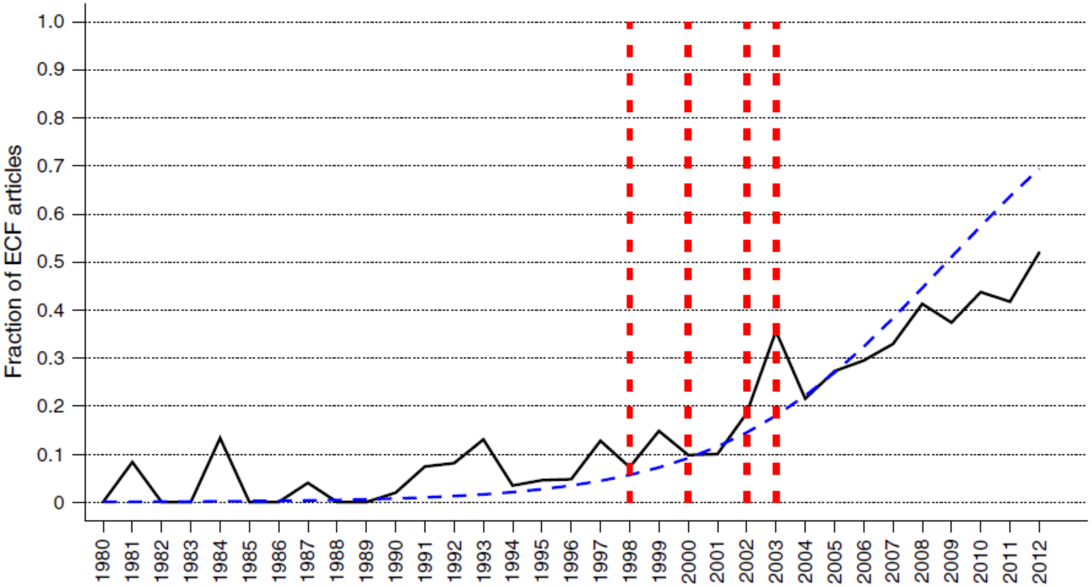
- How to address multicollinearity
  - ➢ (1) Ignore
  - ➢ (2) Drop the variables (e.g., LASSO)

    -> **However, omitted variables bias is more serious than multicollinearity.**
  - ➢ (3) Principal component analysis (PCA) (e.g., Atasoy et al. 2016)

Atasoy, H., Banker, R.D. and Pavlou, P.A., 2016. On the Longitudinal Effects of IT Use on Firm-Level Employment. *Information Systems Research*, 27(1), pp.6-26.

(4) Endogeneity

- Addressing endogeneity is not straightforward.
  - ➢ There is no mathematical solution. It's all about **research design**, so-called identification strategy.
  - ➢ (Example) Adoption of the identification strategies in top 3 finance journals*



Bowen III, D. E., Frésard, L. and Taillard, J. P. 2016. What's Your Identification Strategy? Innovation in Corporate Finance Research. *Management Science*. forthcoming.

* Journal of Finance, Journal of Financial Economics, Review of Financial Studies

What Happens if GM Assumptions are Violated?

- Summary

	Severity for Causal Inference	**Level of Difficulty**
Linearity	High	Easy
Multicollinearity	Low	Not recommended for causal inference (There is a tradeoff with endogeneity)
Heteroscedasticity	Low	Easy
Autocorrelation	Low	Easy
Endogeneity	**High**	**Difficult**

Identification Strategy

Why is Predictive Analytics not Well-Suited for Causal Inference?

- Although causal inference aims at unbiased estimation, predictive analytics focuses more on low variance (out-of-sample prediction), rather than low bias (in-sample unbiased estimation).

$$Expected\ Estimation\ Error = E\left[\left(Y - \hat{f}(x)\right)^2\right]$$

$$= E\left[Y^2 - 2Y\hat{f} + \hat{f}^2\right] = \mathrm{E}[\mathrm{Y}^2] + \mathrm{E}\left[\hat{f}^2\right] - E\left[2Y\hat{f}\right]$$

$$= Var(Y) + E[Y]^2 + \mathrm{Var}\left(\hat{f}\right) + \mathrm{E}\left[\hat{f}\right]^2 - 2f\mathrm{E}\left[\hat{f}\right]$$

$$= Var(Y) + \underline{\left\{E\left[\hat{f}(x)\right] - f(x)\right\}^2} + \underline{E\left[\left(\hat{f}(x) - E\left[\hat{f}(x)\right]\right)^2\right]}$$

$$= \underline{Var[\varepsilon] + E[\varepsilon]^2} + Bias^2 + Var\left[\hat{f}(x)\right]$$

(Untestable assumption) Error term is the source of endogeneity

Why is Predictive Analytics not Well-Suited for Causal Inference?

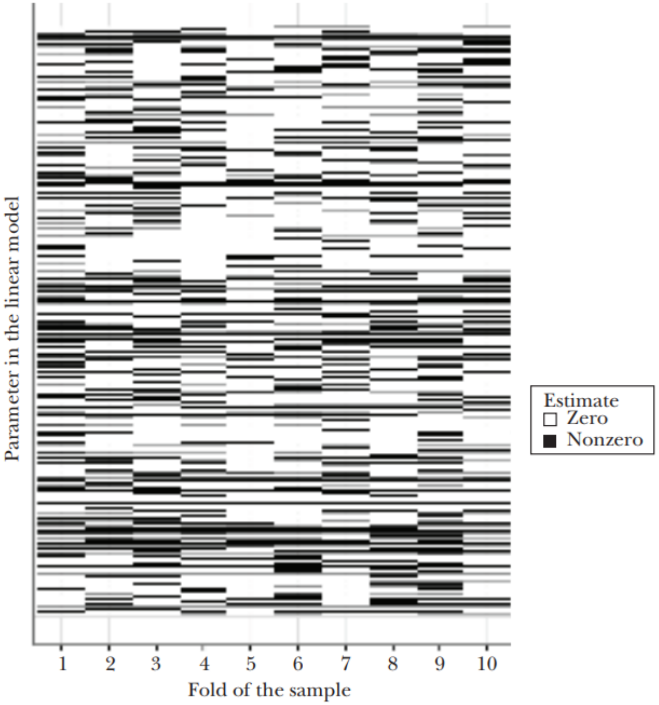
- Automatic feature selection to avoid overfitting and minimize variance (i.e., regularization) could drop several important variables.
  - ➢ (Example) LASSO (least absolute shrinkage and selection operator)

$$\hat{f}_{OLS} = \underset{f_\beta \in \mathcal{F}_{lin}}{\arg\min} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

$$\hat{f}_{ML} = \underset{f \in \mathcal{F}}{\arg\min} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda R(f)$$

  - ➢ LASSO model yields instable patterns of variable selection. (e.g., a focal variable is included in a model, but not in another model) (Mullainathan and Spiess 2017)

Selected Coefficients (Nonzero Estimates) across Ten LASSO Regressions



Mullainathan, S. and Spiess, J., 2017. Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, 31(2), pp.87-106.

Identification Strategy is an Antidote to Endogeneity

- Identification strategy is a **research design** which is intended to solve the endogeneity problem, resting on identification assumption that identifies a proper counterfactual for the treated outcome.
  - [3rd session] Randomized Experiment
  - [3rd session] Quasi-Experiment (Difference-in-Differences)
  - [3rd session] Regression Discontinuity
  - [4th session] Instrument Variable

- Key questions for causal inference
  - How can we leverage the research design to identify the causal relationship?
  - How can we justify the identification assumption in the empirical context?

Predictive Analytics

Prediction is based on Correlation

- Prediction does not require causality. Correlation is enough for prediction.
  - ➢ (Example) Prediction of influenza using Google search query (Ginsberg et al. 2009)
    - Internet search for ‘influenza complication’ is correlated to occurring influenza.
    - But, searching for ‘influenza complication’ does not cause the influenza.

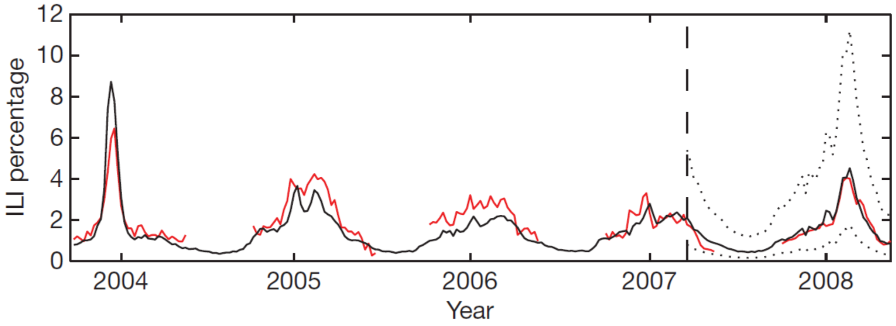
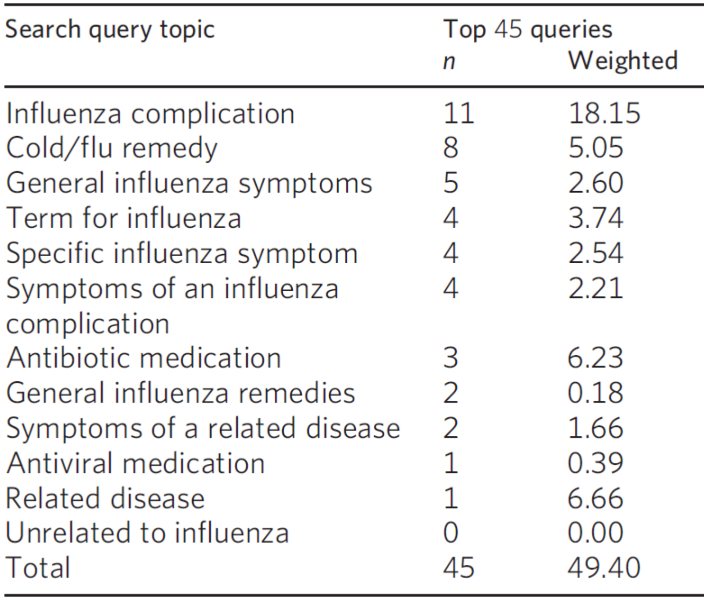


Figure 2 | A comparison of model estimates for the mid-Atlantic region (black) against CDC-reported ILI percentages (red), including points over which the model was fit and validated. A correlation of 0.85 was obtained over 128 points from this region to which the model was fit, whereas a correlation of 0.96 was obtained over 42 validation points. Dotted lines indicate 95% prediction intervals. The region comprises New York, New Jersey and Pennsylvania.

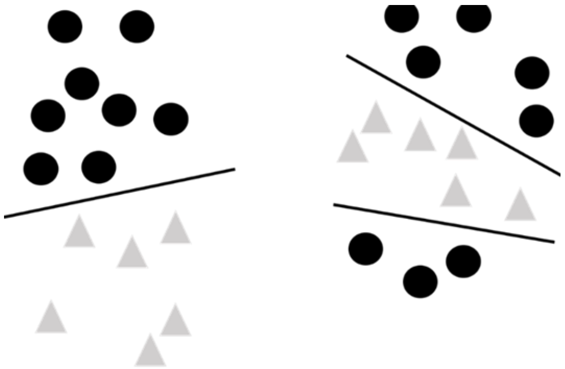


Search query topic	Top 45 queries	
	n	Weighted
Influenza complication	11	18.15
Cold/flu remedy	8	5.05
General influenza symptoms	5	2.60
Term for influenza	4	3.74
Specific influenza symptom	4	2.54
Symptoms of an influenza complication	4	2.21
Antibiotic medication	3	6.23
General influenza remedies	2	0.18
Symptoms of a related disease	2	1.66
Antiviral medication	1	0.39
Related disease	1	6.66
Unrelated to influenza	0	0.00
Total	45	49.40

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S. and Brilliant, L., 2009. Detecting Influenza Epidemics using Search Engine Query Data. *Nature*, 457(7232), p.1012.

Why Does Machine Learning Outperform in Prediction?

- Machine learning (e.g., neural networks, random forests) can represent non-linearity, which is not possible in linear regressions.



Linearly separable problems

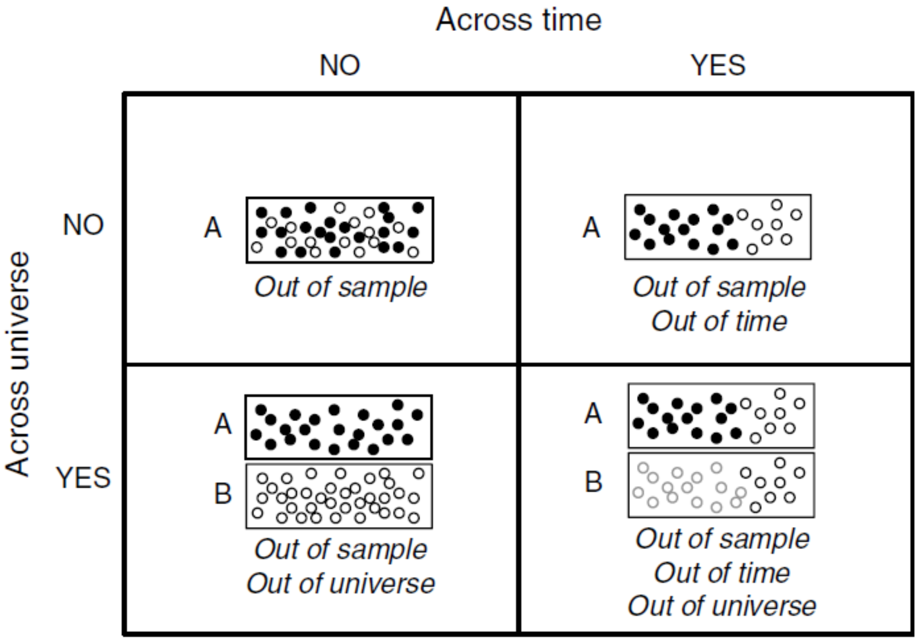
Non-linearly separable problems

- Automatic feature selection
  - Machine learning searches for model specification automatically to achieve the best model fit. (e.g., which variables or interactions should be included?) (Mullainathan and Spiess 2017)

Mullainathan, S. and Spiess, J., 2017. Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, 31(2), pp.87-106.

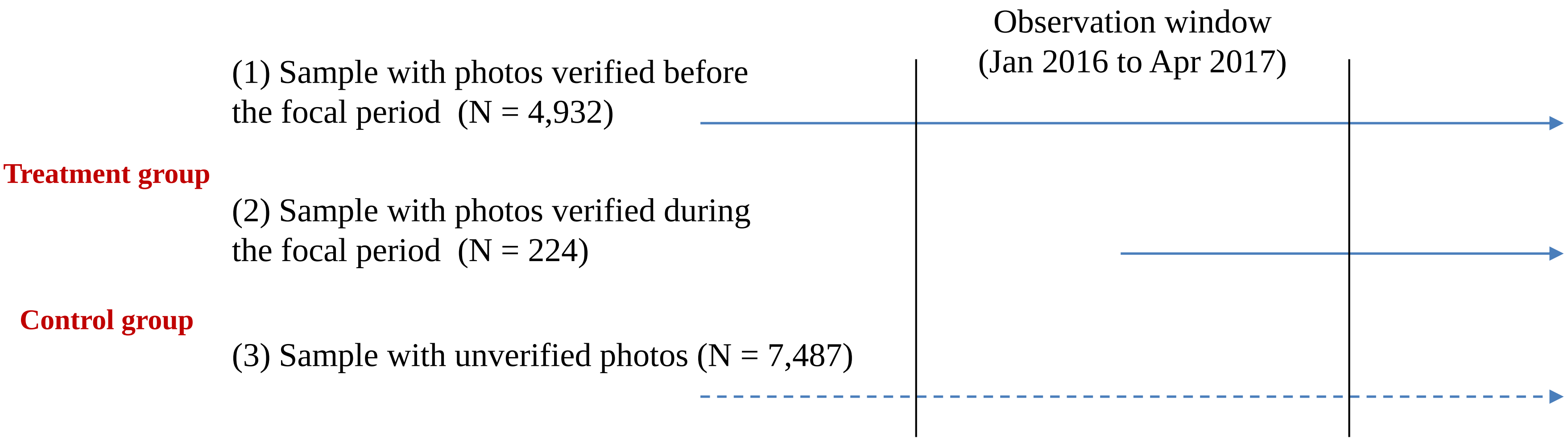
Prediction for Out-of-Sample

- Predictive analytics aims at higher predictive power for out-of-sample
  - ➢ "A dumb algorithm with lots and lots of data beats a clever one with modest amounts of it."
  - ➢ Sometimes, external validity may be ensured at the expense of internal validity or transparency (explainability).



Why is Identification Strategy not Well-Suited for Prediction?

- Identification strategy aims at in-sample “unbiased” estimations
  - ➢ Researchers often trade external validity for internal validity.
  - ➢ (Example) Verified photo and Airbnb demand (Zhang et al. 2017)
    - Zhang et al. (2017) estimate the effect of verified photo in Airbnb on property demand
    - Of over 13,000 listings in Airbnb, some observations are excluded to satisfy identification assumptions (for difference-in-differences, in this context).



Observation window
(Jan 2016 to Apr 2017)

Treatment group

(1) Sample with photos verified before the focal period (N = 4,932)

(2) Sample with photos verified during the focal period (N = 224)

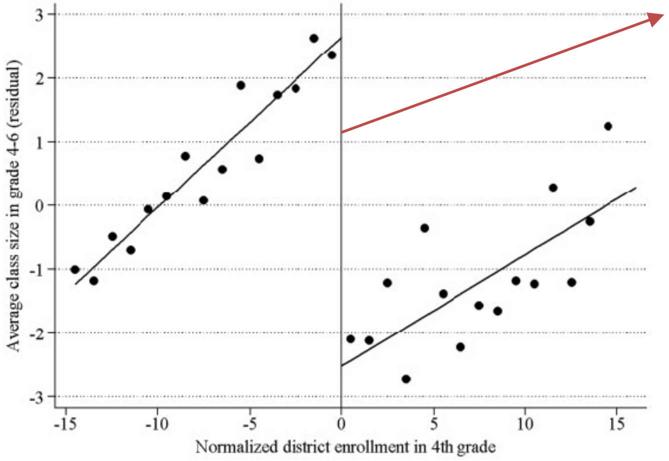
Control group

(3) Sample with unverified photos (N = 7,487)

Zhang, S., Lee, D., Singh, P. V. and Srinivasan, K. 2017, How Much Is an Image Worth? Airbnb Property Demand Estimation Leveraging Large Scale Image Analytics. *CMU Working Paper*.

Why is Identification Strategy not Well-Suited for Prediction?

- Identification strategy aims at in-sample "unbiased" estimations
  - ➢ Researchers often trade external validity for internal validity.
  - ➢ (Example) Long-term effects of class size (Fredriksson et al. 2012)
    - Regression discontinuity (RD) estimate only applies to some limited range of units above and below the threshold, but not to all units
    - Fredriksson et al. (2012) employ the RD design using the policy of 25 maximum class size. Then, how about 10 or 40 cutoffs?

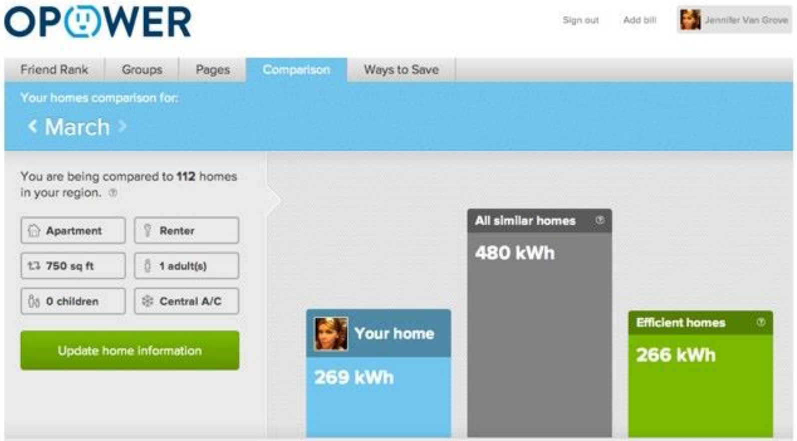


Maximum class size cutoff (25, by law)

Fredriksson, P., Öckert, B. and Oosterbeek, H., 2012. Long-Term Effects of Class Size. *Quarterly Journal of Economics*, 128(1), pp.249-285.

Why is Identification Strategy not Well-Suited for Prediction?

- Experimental estimates are not necessarily stable across settings.
  - ➢ (Example) Opower effectiveness in energy saving (Allcott 2015)
    - Allcott (2015) estimates the impact of home energy reports (Opower case) for 111 markets separately.
    - Effects of Opower treatment are larger for “early adopter” utilities, possibility due to more environmentally-favored customers in its early stages.



Your homes comparison for:

‹ March ›

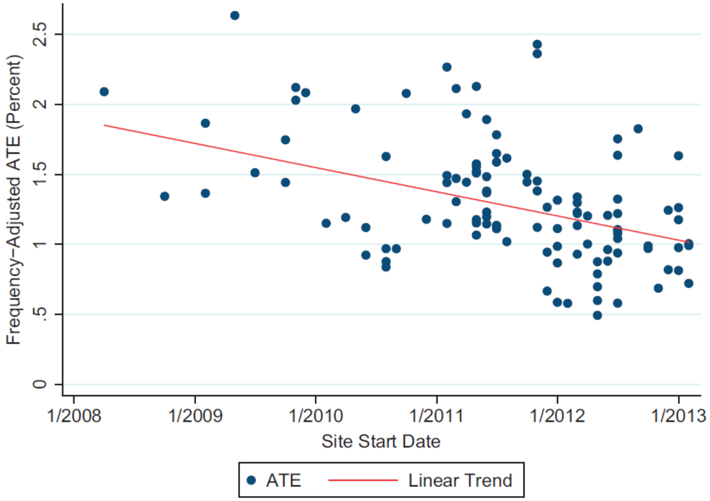
You are being compared to **112** homes in your region.

Apartment | Renter
750 sq ft | 1 adult(s)
0 children | Central A/C

Your home 269 kWh

All similar homes 480 kWh

Efficient homes 266 kWh

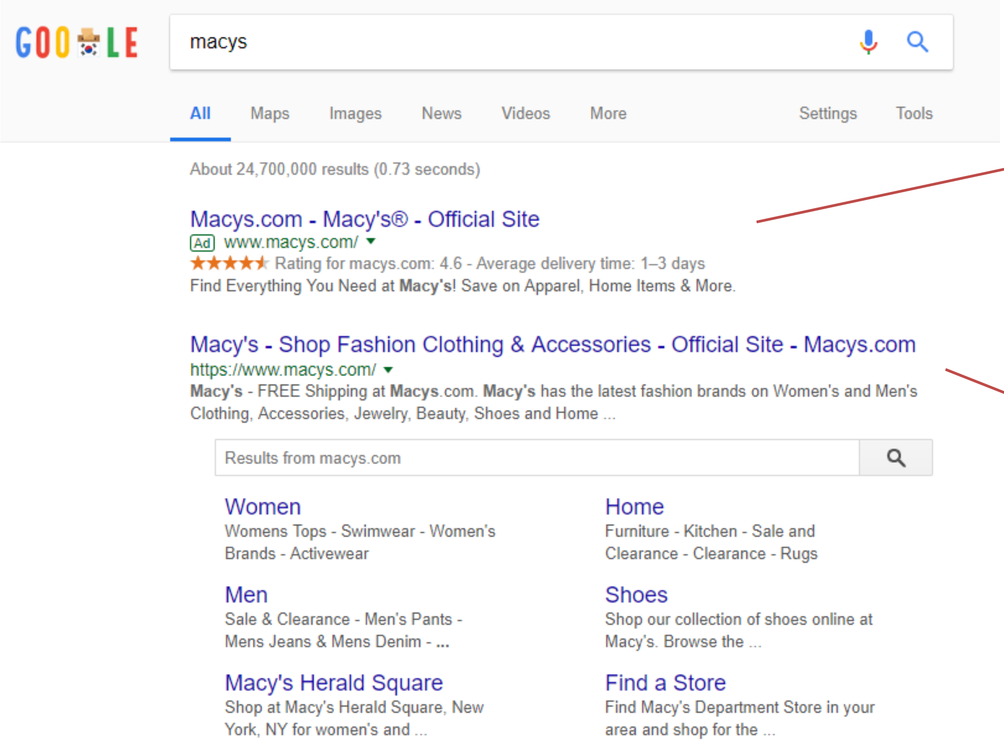


Allcott, H., 2015. Site Selection Bias in Program Evaluation. *Quarterly Journal of Economics*, 130(3), pp.1117-1165.

Example: Search Advertising and Sales

Search Advertising and Sales

- Is paid search advertising worthwhile?



About 24,700,000 results (0.73 seconds)

Macys.com - Macy's® - Official Site
Ad www.macys.com/ ▾
★★★★★ Rating for macys.com: 4.6 - Average delivery time: 1–3 days
Find Everything You Need at **Macy's**! Save on Apparel, Home Items & More.

Macy's - Shop Fashion Clothing & Accessories - Official Site - Macys.com
https://www.macys.com/ ▾
Macy's - FREE Shipping at **Macys**.com. **Macy's** has the latest fashion brands on Women's and Men's Clothing, Accessories, Jewelry, Beauty, Shoes and Home ...

Results from macys.com

Women
Womens Tops - Swimwear - Women's Brands - Activewear

Home
Furniture - Kitchen - Sale and Clearance - Clearance - Rugs

Men
Sale & Clearance - Men's Pants - Mens Jeans & Mens Denim - ...

Shoes
Shop our collection of shoes online at Macy's. Browse the ...

Macy's Herald Square
Shop at Macy's Herald Square, New York, NY for women's and ...

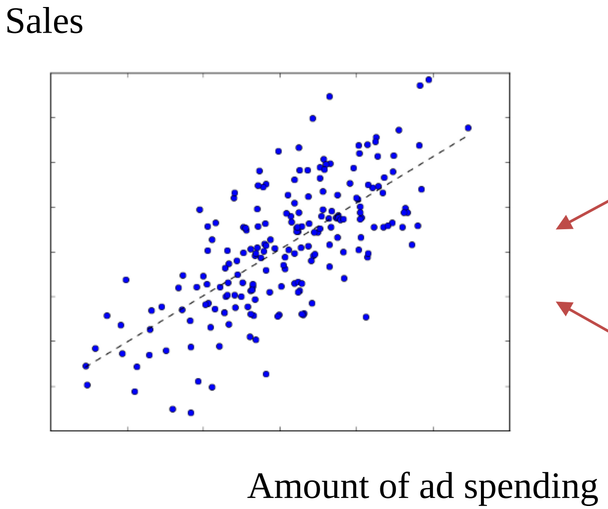
Find a Store
Find Macy's Department Store in your area and shop for the ...

Paid search advertising

Organic (natural) search results

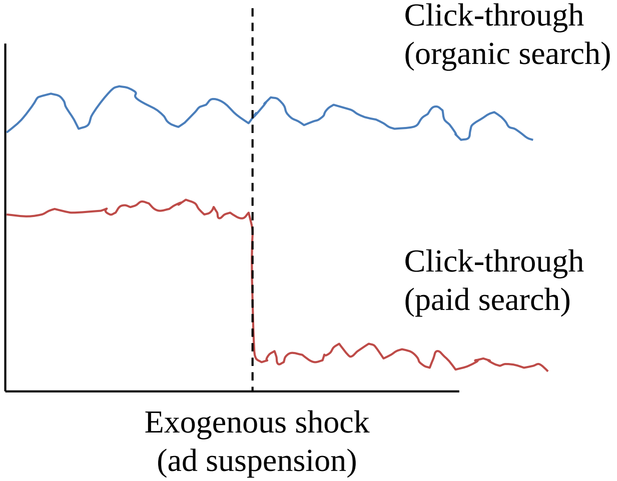
Search Advertising and Sales

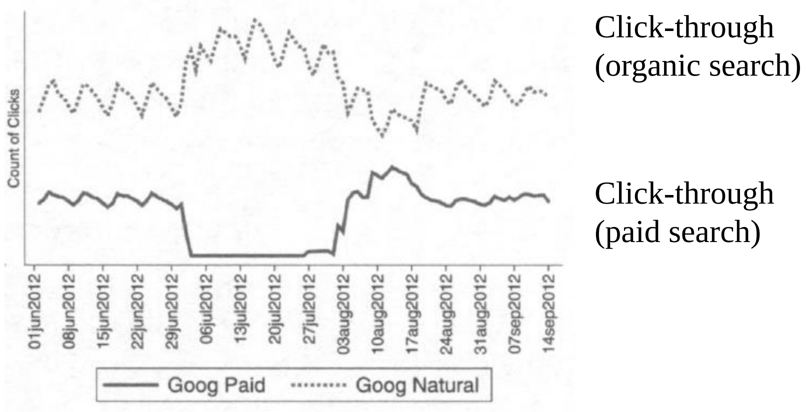
- Positive correlation does not have to result from causality.



Causality

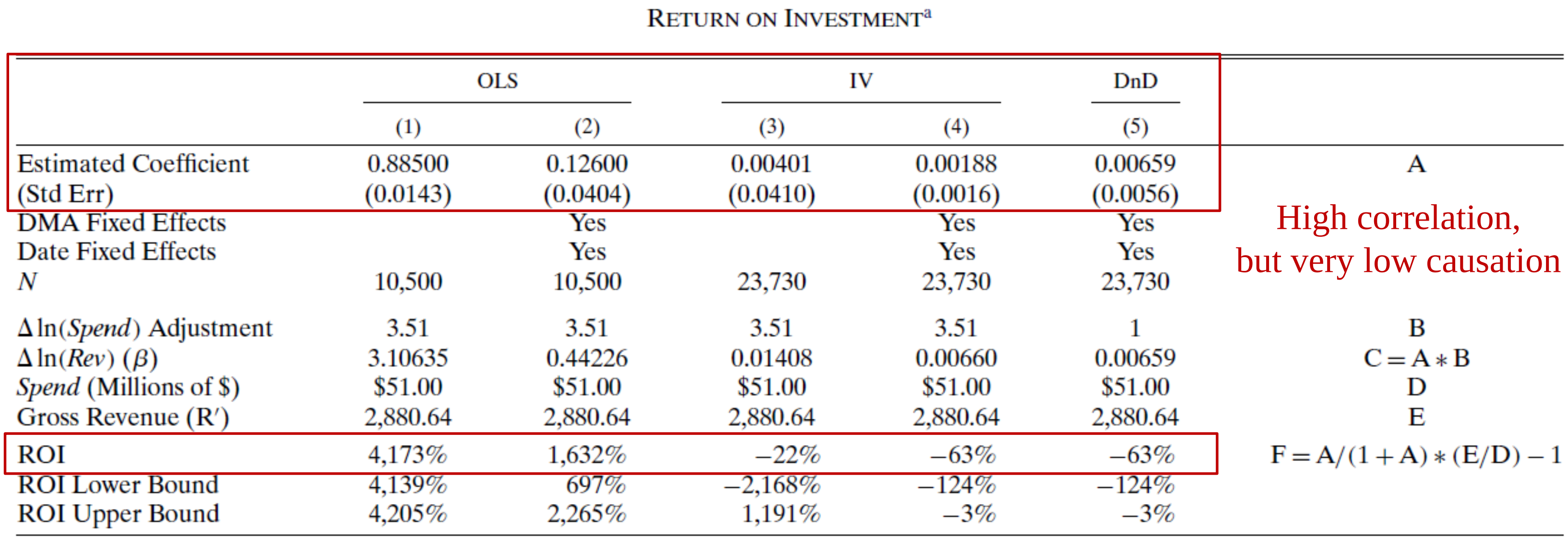
Correlation





Search Advertising and Sales

- A positive correlation does not necessarily guarantee a positive causation.
  - Blake et al. (2015) show that OLS and identification strategies (IV and DID, based on a filed experiment) yield quite different estimates on the effect of search advertising spending on sales.



RETURN ON INVESTMENT[a]

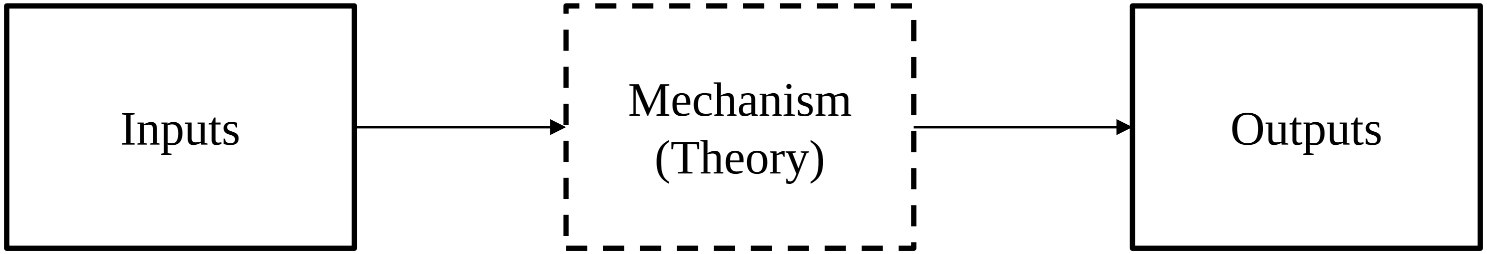
	OLS		IV		DnD	
	(1)	(2)	(3)	(4)	(5)	
Estimated Coefficient	0.88500	0.12600	0.00401	0.00188	0.00659	A
(Std Err)	(0.0143)	(0.0404)	(0.0410)	(0.0016)	(0.0056)	
DMA Fixed Effects		Yes		Yes	Yes	
Date Fixed Effects		Yes		Yes	Yes	
N	10,500	10,500	23,730	23,730	23,730	
$\Delta\ln(Spend)$ Adjustment	3.51	3.51	3.51	3.51	1	B
$\Delta\ln(Rev)$ (β)	3.10635	0.44226	0.01408	0.00660	0.00659	$C = A * B$
Spend (Millions of $)	$51.00	$51.00	$51.00	$51.00	$51.00	D
Gross Revenue (R′)	2,880.64	2,880.64	2,880.64	2,880.64	2,880.64	E
ROI	4,173%	1,632%	−22%	−63%	−63%	$F = A/(1 + A) * (E/D) - 1$
ROI Lower Bound	4,139%	697%	−2,168%	−124%	−124%	
ROI Upper Bound	4,205%	2,265%	1,191%	−3%	−3%	

High correlation, but very low causation

Blake, T., Nosko, C. and Tadelis, S., 2015. Consumer Heterogeneity and Paid Search Effectiveness: A Large-Scale Field Experiment. *Econometrica*, 83(1), pp.155-174.

Input-Output Framework

- Is paid search advertising worthwhile?



Search advertising spending

Sales

- ➢ Based on causality (estimated by IV and DID), the return on investment for advertising was negative.
- ➢ For the **marketing intervention**, correlation is definitely not enough.

  “*Should we spend more money on search advertising?*”

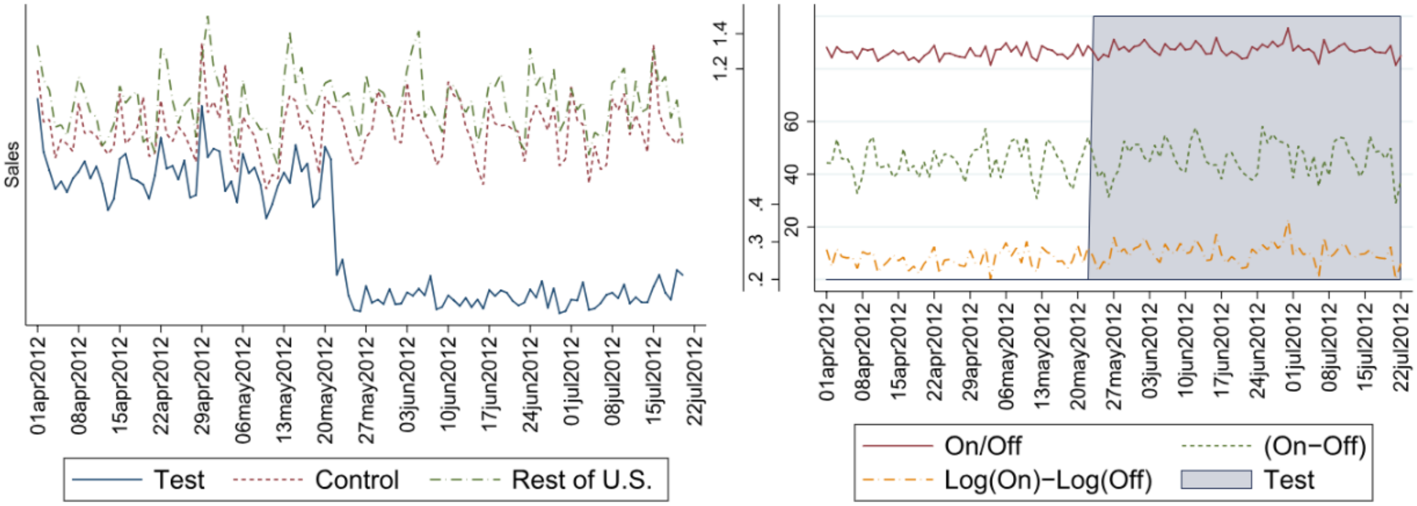
- ➢ Based on correlation (OLS), the return on investment for advertising was about 1,600%.
- ➢ For the **sales prediction**, that’s enough.

  “*When a company spends $1M on search advertising, how much would their short-term sales be?*”

Blake, T., Nosko, C. and Tadelis, S., 2015. Consumer Heterogeneity and Paid Search Effectiveness: A Large-Scale Field Experiment. *Econometrica*, 83(1), pp.155-174.

Search Advertising Effectiveness

- Using a field experiment in corporation with eBay, Blake et al. (2015) find that there is no noticeable difference between the pre- and post-experimental period, demonstrating the muted overall effect of paid search.



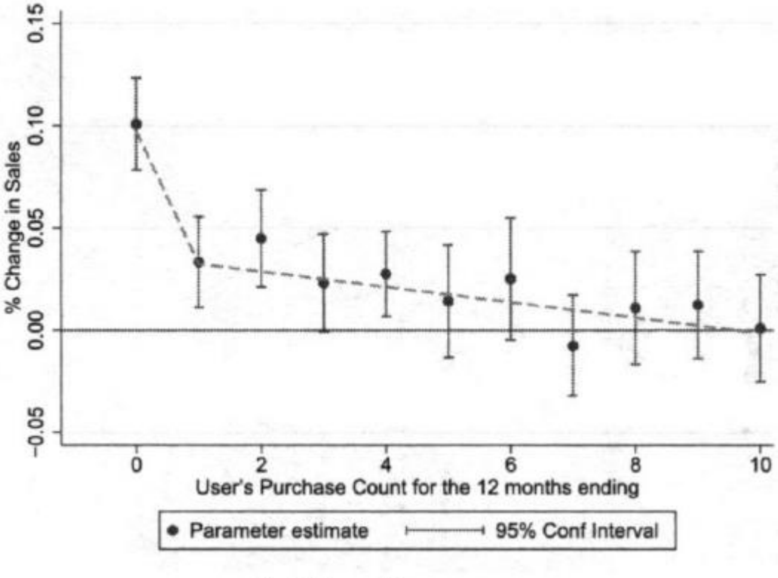
(a) Attributed Sales by Region

(b) Differences in Total Sales

Blake, T., Nosko, C. and Tadelis, S., 2015. Consumer Heterogeneity and Paid Search Effectiveness: A Large-Scale Field Experiment. *Econometrica*, 83(1), pp.155-174.

Search Advertising Effectiveness

- Delving deeper into the mechanism sheds light on how advertising strategy should be designed. (Blake et al. 2015)
  - ➢ "Targeting uninformed users is a critical factor for successful advertising." (p. 155)
  - ➢ "Arguments have been made that brand keyword advertising acts as a defense against a competitor bidding for a company's brand name." (Prisoner's Dilemma) (p. 171)



(a) User Frequency

Blake, T., Nosko, C. and Tadelis, S., 2015. Consumer Heterogeneity and Paid Search Effectiveness: A Large-Scale Field Experiment. *Econometrica*, 83(1), pp.155-174.

Sales Prediction

- Predicting more accurately requires more data (not necessarily based on causation).
  - ➢ (Example) Geva et al. (2017) use a new source of online information (i.e., search volume and social media mentions) to predict offline sales.

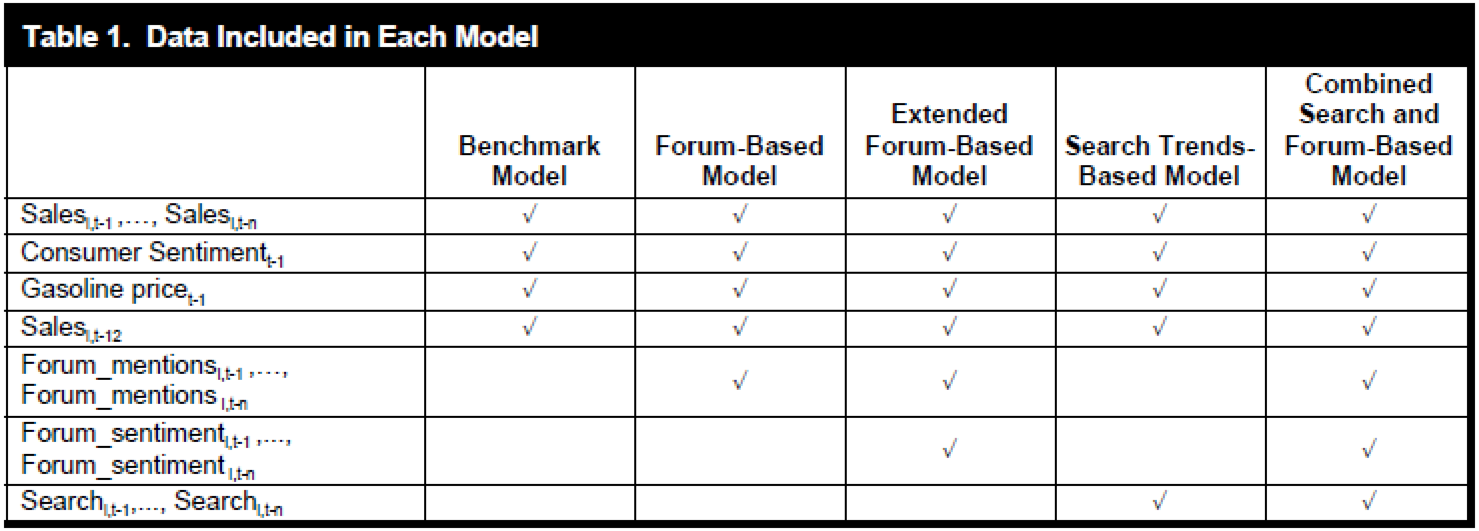


Table 1. Data Included in Each Model

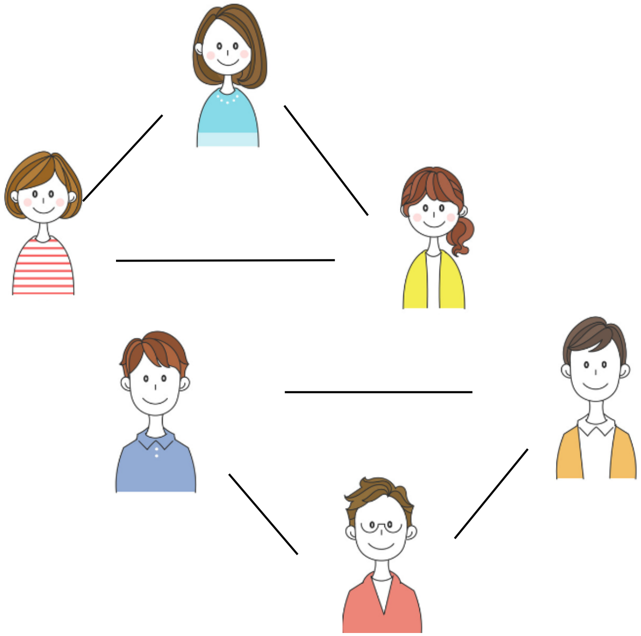
	Benchmark Model	Forum-Based Model	Extended Forum-Based Model	Search Trends-Based Model	Combined Search and Forum-Based Model
Sales$_{i,t-1}$,…, Sales$_{i,t-n}$	√	√	√	√	√
Consumer Sentiment$_{t-1}$	√	√	√	√	√
Gasoline price$_{t-1}$	√	√	√	√	√
Sales$_{i,t-12}$	√	√	√	√	√
Forum_mentions$_{i,t-1}$,…, Forum_mentions $_{i,t-n}$		√	√		√
Forum_sentiment$_{i,t-1}$,…, Forum_sentiment $_{i,t-n}$			√		√
Search$_{i,t-1}$,…, Search$_{i,t-n}$				√	√

Geva, T., Oestreicher-Singe, G., Efron, N. and Shimshoni, Y., 2017. Using Forum and Search Data far Sales Prediction of High-Involvement Projects. *MIS Quarterly*, 41(1), pp.65-82.

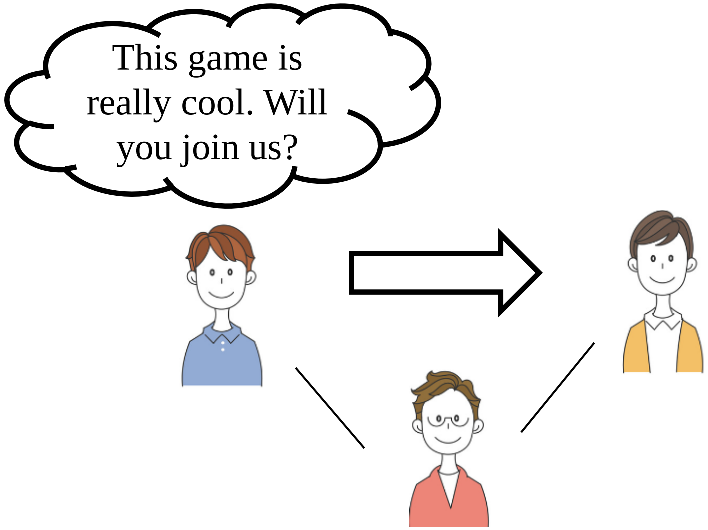
Example: Social Network and Word-of-Mouth

Homophily versus Peer Influence

- Similar patterns of connected nodes in networks might stem from either homophily or peer influence.
  - ➢ Similar patterns from homophily are based on correlation, and those from peer influence are based on causality.



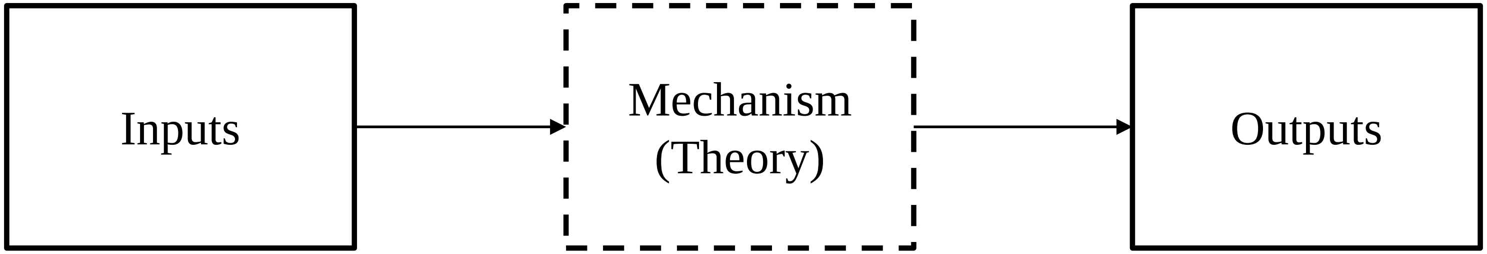
Homophily



Peer influence

Input-Output Framework

- Social network and word-of-mouth



Social network

(1) Homophily
(2) Peer influence

Sales (word-of-mouth)

"The establishment of social influence, however, does not provide an answer to the question of whether social network data will improve targeting and prediction in practice."

"And **even when individuals are known not to influence one another, social ties may be still be predictive**, as observed in the sociological research on homophily."

For a goal of prediction, correlation is sufficient

"On the contrary, **when nodes are known to influence each other, marketers may wish to affect diffusion patterns**, for instance, by seeding influential nodes with marketing actions in order to encourage them to adopt early."

(Goel and Goldstein 2014, p. 83-84)

For a goal of intervention, causal inference is required

Goel, S., & Goldstein, D. G. (2014). Predicting Individual Behavior with Social Networks. *Marketing Science*, 33(1), 82-93.

Identifying (Causal) Peer Influence

- Quasi-experiment on social media (Adamopoulos et al. 2018, p. 8)

Possibility of confounds (or omitted variables)

"However, users who are connected in the platform (through a nonreciprocal or reciprocal relationship) tend to have similar preferences and idiosyncrasies."

"Hence, if one simply employs observational data under the aforementioned research design, it would be difficult to distinguish the actual effect a WOM instance might exert from simple correlations in users' behaviors and homophily."

Endogneiety problem due to correlation (homophily)

However, correlation is sufficient for prediction

"Nonetheless, the discovery of correlations among latent personality traits and the effectiveness of WOM would be sufficient for forecasting objectives and practical marketing strategies."

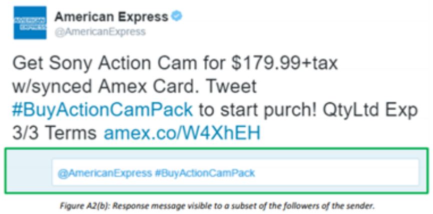
"Such a quasi-experiment creates an exogenous source of variation in the explanatory variables and allows us to identify the various effects related to being exposed to a friend's actual purchase and advocacy"

Identifying peer influence requires an experiment

Adamopoulos, P., Ghose, A. and Todri, V., 2018. The Impact of User Personality Traits on Word of Mouth: Text-Mining Social Media Platforms. *Information Systems Research*. forthcoming

Identifying (Causal) Peer Influence

- Quasi-experiment on social media (Adamopoulos et al. 2018)

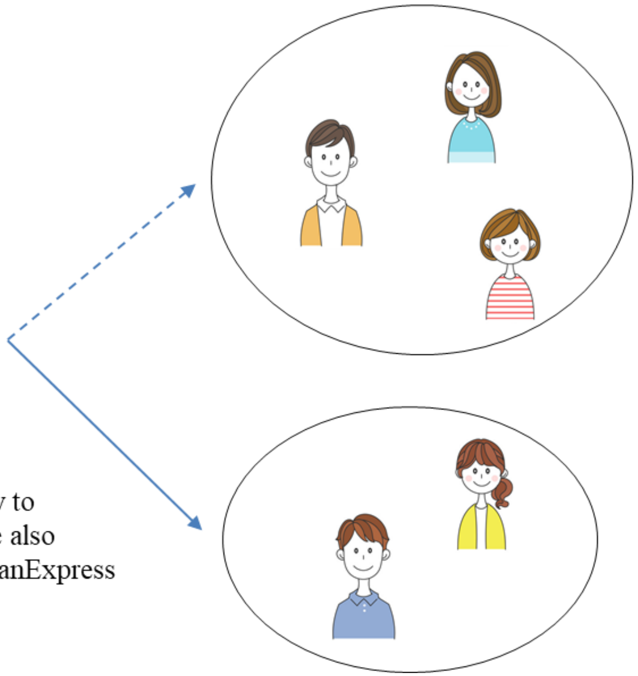


American Express
@AmericanExpress

Get Sony Action Cam for $179.99+tax w/synced Amex Card. Tweet #BuyActionCamPack to start purch! QtyLtd Exp 3/3 Terms amex.co/W4XhEH

@AmericanExpress #BuyActionCamPack

Figure A2(b): Response message visible to a subset of the followers of the sender.



Broadcasting only to followers who are also following AmericanExpress

Control Group: Followers who are not following AmericanExpress

Treatment Group: Followers who are also following AmericanExpress

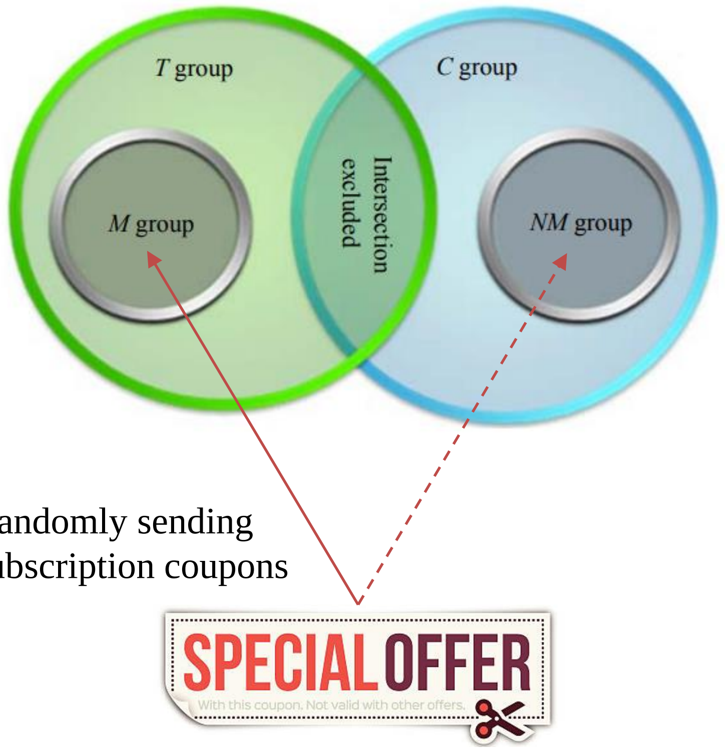
Adamopoulos, P., Ghose, A. and Todri, V., 2018. The Impact of User Personality Traits on Word of Mouth: Text-Mining Social Media Platforms. *Information Systems Research*. forthcoming

Identifying (Causal) Peer Influence

- Randomized field experiment on social media (Bapna and Umyarov 2015)

last.fm

Figure 1 (Color online) Venn Diagram of Experimental Design



Treatment group
= Friends of M groups

Control group
= Friends of NM groups

Randomly sending
subscription coupons

Bapna, R. and Umyarov, A., 2015. Do Your Online Friends Make You Pay? A Randomized Field Experiment on Peer Influence in Online Social Networks. *Management Science*, 61(8), pp.1902-1920.

Prediction Using Social Network

- Are social data useful for predicting clicks on ads?
  - (Example) Goel and Goldstein (2014) find that people whose social contacts clicked on the ad were more than 10 times (for *Movie 1*) as likely to click than those without contacts who clicked (note that these results are meaningful, regardless of whether they are directly influenced or not).

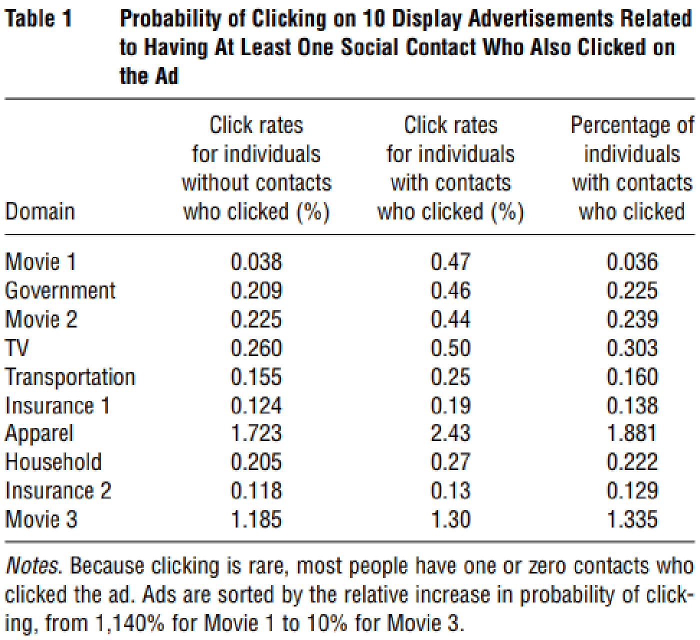


Table 1 **Probability of Clicking on 10 Display Advertisements Related to Having At Least One Social Contact Who Also Clicked on the Ad**

Domain	Click rates for individuals without contacts who clicked (%)	Click rates for individuals with contacts who clicked (%)	Percentage of individuals with contacts who clicked
Movie 1	0.038	0.47	0.036
Government	0.209	0.46	0.225
Movie 2	0.225	0.44	0.239
TV	0.260	0.50	0.303
Transportation	0.155	0.25	0.160
Insurance 1	0.124	0.19	0.138
Apparel	1.723	2.43	1.881
Household	0.205	0.27	0.222
Insurance 2	0.118	0.13	0.129
Movie 3	1.185	1.30	1.335

Notes. Because clicking is rare, most people have one or zero contacts who clicked the ad. Ads are sorted by the relative increase in probability of clicking, from 1,140% for Movie 1 to 10% for Movie 3.

Goel, S. and Goldstein, D.G., 2013. Predicting Individual Behavior with Social Networks. *Marketing Science*, 33(1), pp.82-93.

Conclusion

The Right Tool for the Right Question

- Identification strategy and predictive analytics do aim different goals
  - ➢ Identification strategy aims to yield unbiased estimates for causal inference.
  - ➢ Predictive analytics aims to make the most of correlations, while avoiding overfitting (ensuring out-of-sample prediction)

- Think first what question you want to solve
  - ➢ On one hand, identification strategies for causal inference are well-suitable for "intervention-oriented" research.
  - ➢ On the other hand, predictive analytics is well-suitable for "solution-oriented" research.

End of Document