Tuning the tie-breaker design

Art B. Owen*, Stanford University

and

Hal Varian, Google

*Work (mostly) done for Google, not as part of my Stanford responsibilities.

Overview

Lots of problems come up when you combine data of different types.

From 40,000 feet

- Bayes
- Likelihood
- Empirical Bayes
- Transportability

At ground level

Specifics are interesting.

First I sketch some recent examples.

Then the work with Hal Varian.

Big data & small data

With Aiyou Chen and Minghui Shi of Google.

Small and good data set

$(x_i, y_i), \quad i \in S, \quad n$ obs

We want β for this small population.

Huge data set, possibly relevant

$(x_i, y_i), \quad i \in B, \quad N \gg n$ obs

Approach

Shrink $\hat{\beta}_S$ towards $\hat{\beta}_B$

Stein or Bayes

Related GWAS

With Edgar Dobriban, Stuart Kim, Kristen Fortney

- Tiny underpowered GWAS on centenarians
- Seek optimal weighting of the SNP hypotheses (inverse weight the p-values)
- Using huge GWAS on age-related illness (eg diabetes, hypertension)

We got some new longevity-associated genes.

Partial conjunction tests

Concept from Benjamini & Heller

Test same H_0 n data sets. Require at least r rejections.

Better reproducibility than meta-analysis.

2 papers lead by Jingshu Wang

Paper 1

Conditions for admissible testing of a weirdly composite "sparsity null".

Wang & O (2018) JASA

Paper 2

N genes in n studies

An $N \times n$ matrix of p-values

Filtering idea to do N PC tests at once.

Wang, Su, Sabatti, O (2018)

Propensity work

With Evan Rosenman and Michael Baiocchi and Hailey Banack (2018)

Does $W \in \{0, 1\}$ cause y?

Huge data base (W_i, x_i, y_i) for $i \in$ Obs.

W_i chosen in a way that could depend on x_i

Small randomized experiment (W_i, x_i, y_i) for $i \in$ Expt.

W_i chosen at random

First idea

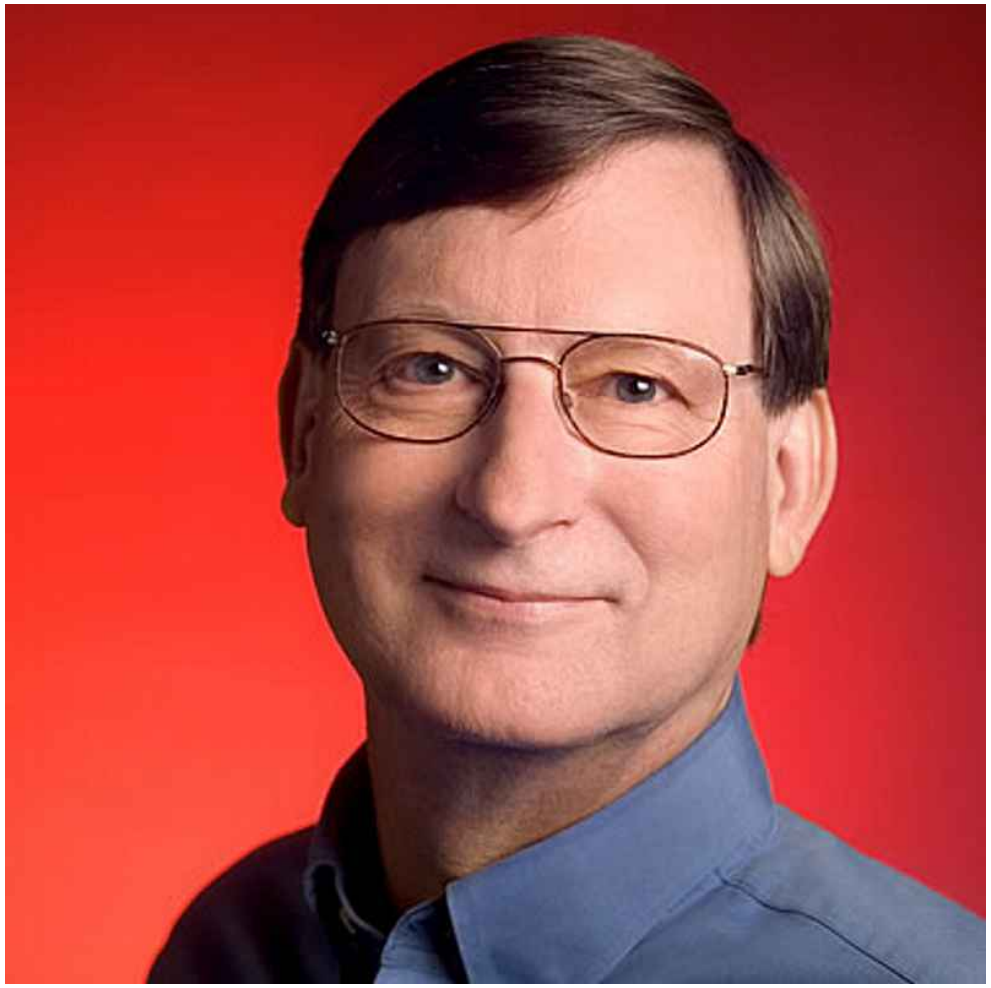
Put experimental subjects into a propensity bucket.

The one they would have occupied in the observational data.

Women's health initiative

Both kinds of data on hormone replacement vs coronary heart disease.

Hal Varian



Google chief economist

Customer loyalty plans

An airline can give an upgrade to n out of N customers. Who?

- The n most loyal customers?
- The n customers most likely to start flying / spending more?

Other examples

- Hotels & car rental companies
- E-commerce platforms, for their advertisers, reviewers, or content producers

Two goals

1) Get the most value from the offer

2) Measure the causal effect of the offer

Two acronyms

1) RDD = Regression Discontinuity Design

2) RCT = Randomized Controlled Trial

We will hybridize between these approaches.

The random variables

i	customer id
z_i	treatment, YES $= 1$, NO $= -1$
y_i	outcome, e.g., revenue one year later (or profit, or $\cdots$)
x_i	assignment variable (larger the better)

Assignment / running variable x

1) It could be past revenue, or

2) a machine learning prediction.

Some simplifications

Suppose at first that half of $z_i = 1$ and half are -1.

(undo later)

Rank transformation

Sort customers, $x_1 \leqslant x_2 \leqslant \cdots \leqslant x_N$, then

$$\text{re-define} \qquad x_i \leftarrow \frac{2i - N - 1}{N}$$

Now $-1 < x_i < 1$.

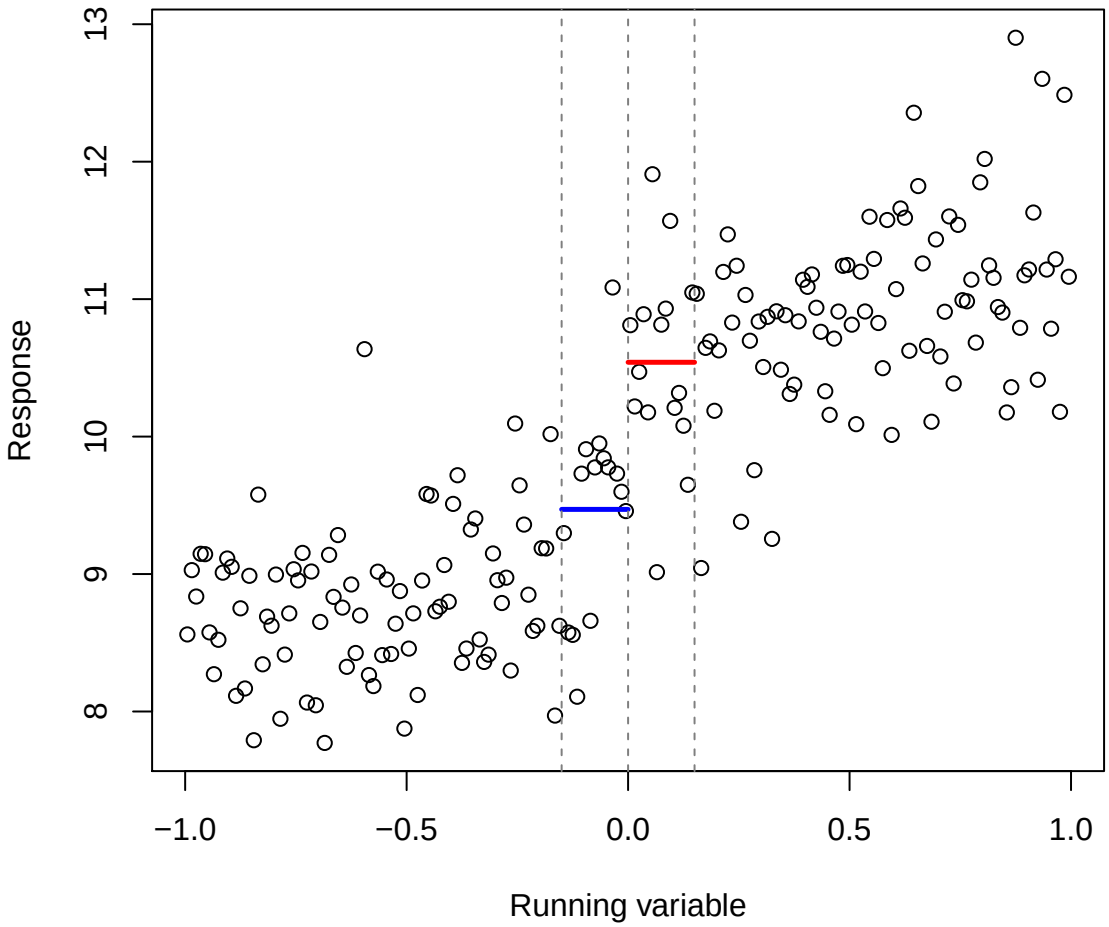
Two-line regression

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i + \varepsilon_i \qquad \varepsilon_i \sim (0, \sigma^2)$$

Other models are interesting, but we need to pick one, so this is it.

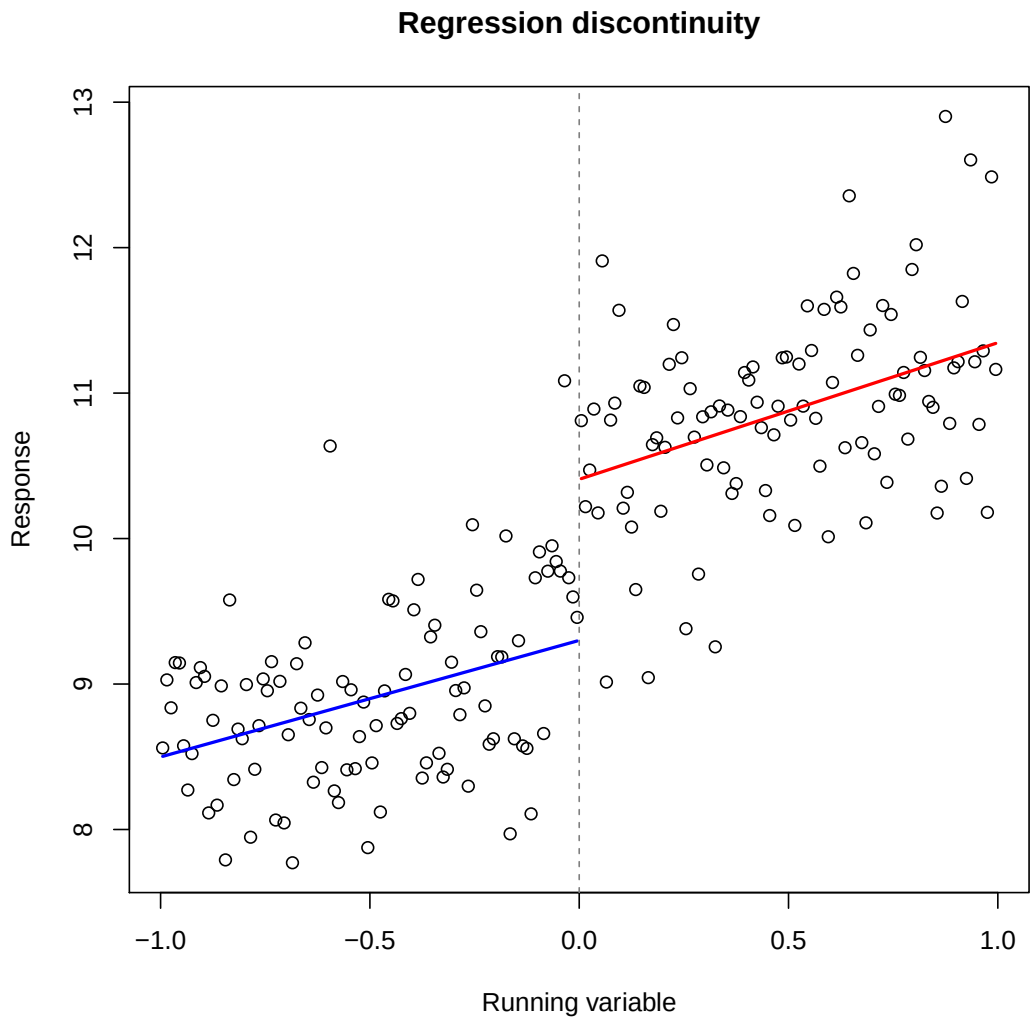
Regression discontinuity

Treatment IFF $x > 0$ Thistlethwaite & Campbell (1960)



People just left of the discontinuity should be comparable to those just right of it.

Separate linear regressions



Raises thorny extrapolation/linearity issues at large $|x_i|$.

Regression discontinuity

Famous example:

$x =$ test score

$z =$ merit scholarship iff $x \geqslant \tau$

$y =$ went to grad school

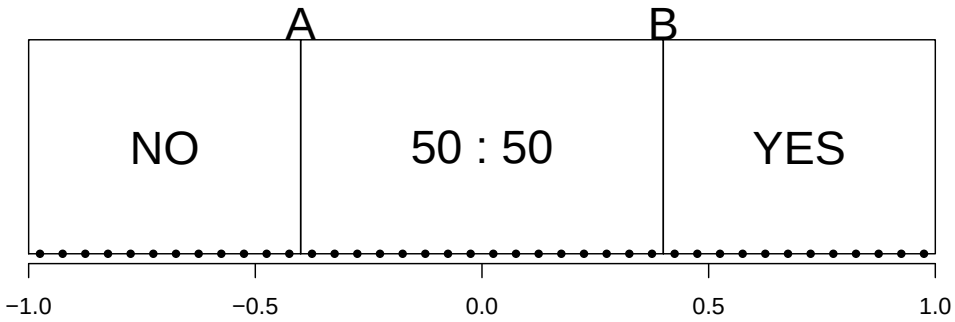
then logistic regression.

RDD is the second most believable causal inference method.

Tie-breaker design

Pick cutoffs $A \leqslant B$, then

$$z_i = \begin{cases} 1, & x_i \geqslant B \\ -1, & x_i \leqslant A \\ \text{random}, & A < x_i < B \end{cases}$$



Extreme cases

1) $x_1 < A = B < x_N \implies$ RDD

2) $A < x_1 \leqslant \cdots \leqslant x_N < B \implies$ RCT

Also called “cutoff designs” Cappelleri & Trochim

Examples

x	z	Ref
Reading ability	Remedial English class	Aiken et al. (1998)
Student ranking	Post secondary financial aid	Angrist et al (2014)
Composite prognostic	Inpatient rehab	Havassy

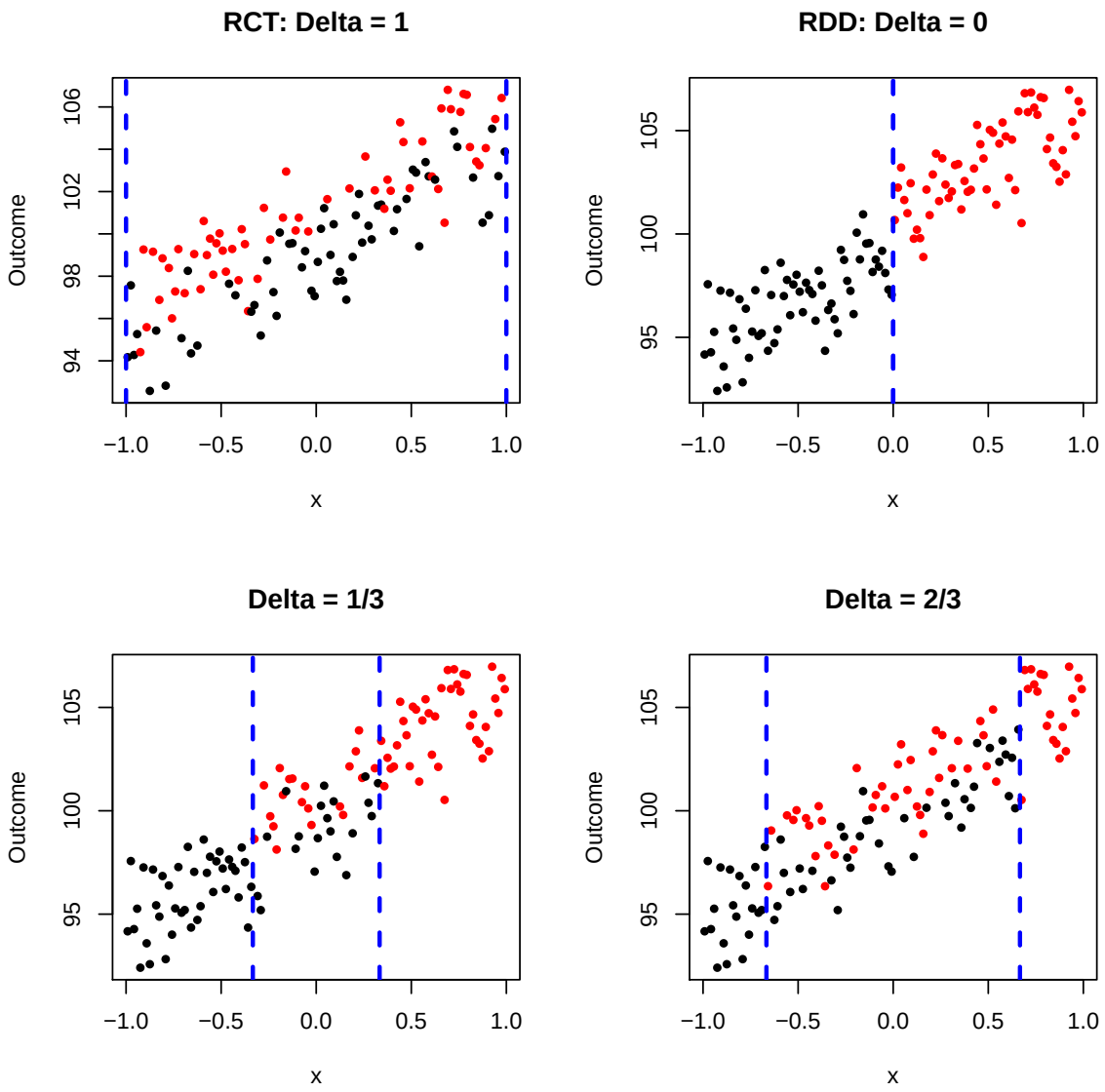
Lanarkshire milk experiment

Student (1931)

Maybe a tie-breaker would have worked.

Tie-breakers

Δ = Fraction in RDD between Blue dashed lines



Two-line regression

$$\mathbb{E}(y) = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz$$

$$\mathcal{X} = \begin{pmatrix} 1 & x_1 & z_1 & x_1 z_1 \\ 1 & x_2 & z_2 & x_2 z_2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_N & z_N & x_N z_N \end{pmatrix}$$

$$\mathrm{Var}(\hat{\beta}) = (\mathcal{X}^\mathsf{T}\mathcal{X})^{-1}\sigma^2$$

$$\Pr(z_i = 1) = \begin{cases} 0, & x_i \leqslant -\Delta \\ 1/2, & |x_i| < \Delta \\ 1, & x_i \geqslant \Delta \end{cases}$$

Integral approximation

$$\frac{1}{N}\mathcal{X}^\mathsf{T}\mathcal{X} \approx \begin{array}{c} \\ 1 \\ x \\ z \\ xz \end{array} \begin{array}{c} \begin{array}{cccc} 1 & x & z & xz \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & \phi(\Delta) \\ 0 & 1/3 & \phi(\Delta) & 0 \\ 0 & \phi(\Delta) & 1 & 0 \\ \phi(\Delta) & 0 & 0 & 1/3 \end{pmatrix} \end{array}$$

where

$$\begin{aligned} \phi(\Delta) &\equiv \frac{1}{2}\int_{-1}^{1} x\mathbb{E}(z \mid x)\,\mathrm{d}x \\ &= \frac{1}{2}\int_{-1}^{-\Delta}(-x)\,\mathrm{d}x + \frac{1}{2}\int_{-\Delta}^{\Delta} 0\,\mathrm{d}x + \frac{1}{2}\int_{\Delta}^{1} x\,\mathrm{d}x \\ &= \frac{1-\Delta^2}{2} \end{aligned}$$

The error above is $O_p(1/\sqrt{N})$.

Even less under stratification.

Rearrange $\mathcal{X}^\mathsf{T}\mathcal{X}/N$

$$\begin{array}{c} \\ 1 \\ zx \\ z \\ x \end{array} \begin{array}{c} \begin{array}{cccc} 1 & zx & z & x \end{array} \\ \begin{pmatrix} 1 & \phi & \cdot & \cdot \\ \phi & 1/3 & \cdot & \cdot \\ \cdot & \cdot & 1 & \phi \\ \cdot & \cdot & \phi & 1/3 \end{pmatrix} \end{array}$$

(using $\cdot$ for 0)

$$N \times \mathrm{Var}\begin{pmatrix} \begin{pmatrix} \hat\beta_0 \\ \hat\beta_3 \\ \hat\beta_2 \\ \hat\beta_1 \end{pmatrix} \end{pmatrix} = \frac{1}{1/3-\phi^2} \begin{pmatrix} 1/3 & -\phi & \cdot & \cdot \\ -\phi & 1 & \cdot & \cdot \\ \cdot & \cdot & 1/3 & -\phi \\ \cdot & \cdot & -\phi & 1 \end{pmatrix} \sigma^2$$

$$\phi = \phi(\Delta) = \frac{1-\Delta^2}{2}$$

Normalization

The design choice is which Δ to use.

That comes down to

$$\frac{\mathrm{Var}(c^\mathsf{T}\hat{\beta};\Delta_1)}{\mathrm{Var}(c^\mathsf{T}\hat{\beta};\Delta_0)}$$

for various vectors c.

Cancellation

σ^2 cancels in this ratio.

So we fix $\sigma^2 = 1$.

Impact

Changing z from -1 to $+1$ increases $\mathbb{E}(y)$ by

$$\begin{aligned}&\big(\beta_0 + \beta_1 x + \beta_2 + \beta_3 x\big) - \big(\beta_0 + \beta_1 x - \beta_2 - \beta_3 x\big)\\ &= 2(\beta_2 + x\beta_3)\end{aligned}$$

So β_2 and β_3 are important.

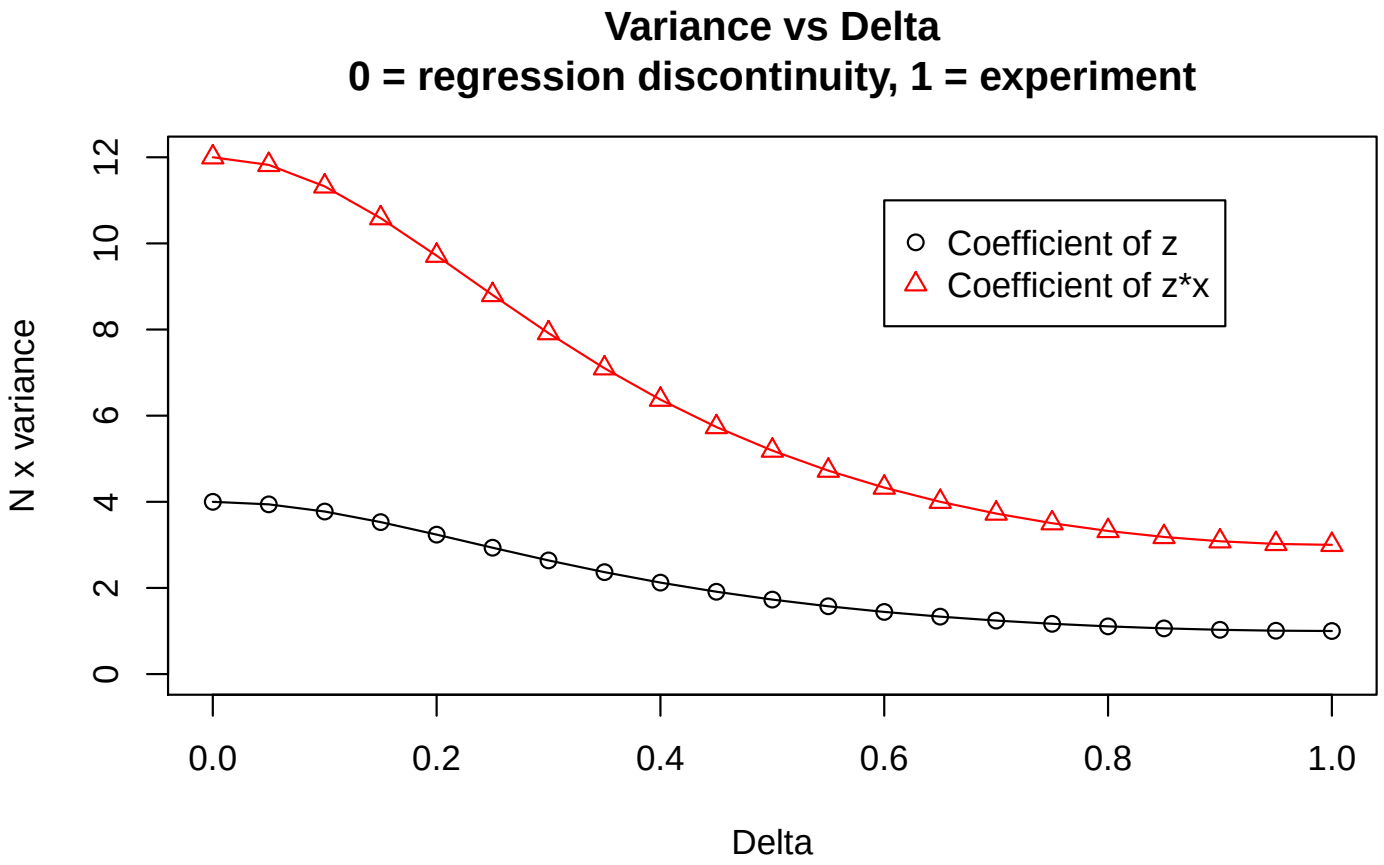
So is x.

(If we didn't already know)

Variance

$$\mathrm{Var}(2(\hat{\beta}_2 + x\hat{\beta}_3)) = \cdots = \frac{16(1+3x^2)}{1+3\Delta^2(2-\Delta^2)}$$

Variance vs Δ



$\mathrm{Var}(\hat{\beta}_2) = 3\mathrm{Var}(\hat{\beta}_3)$ all Δ.

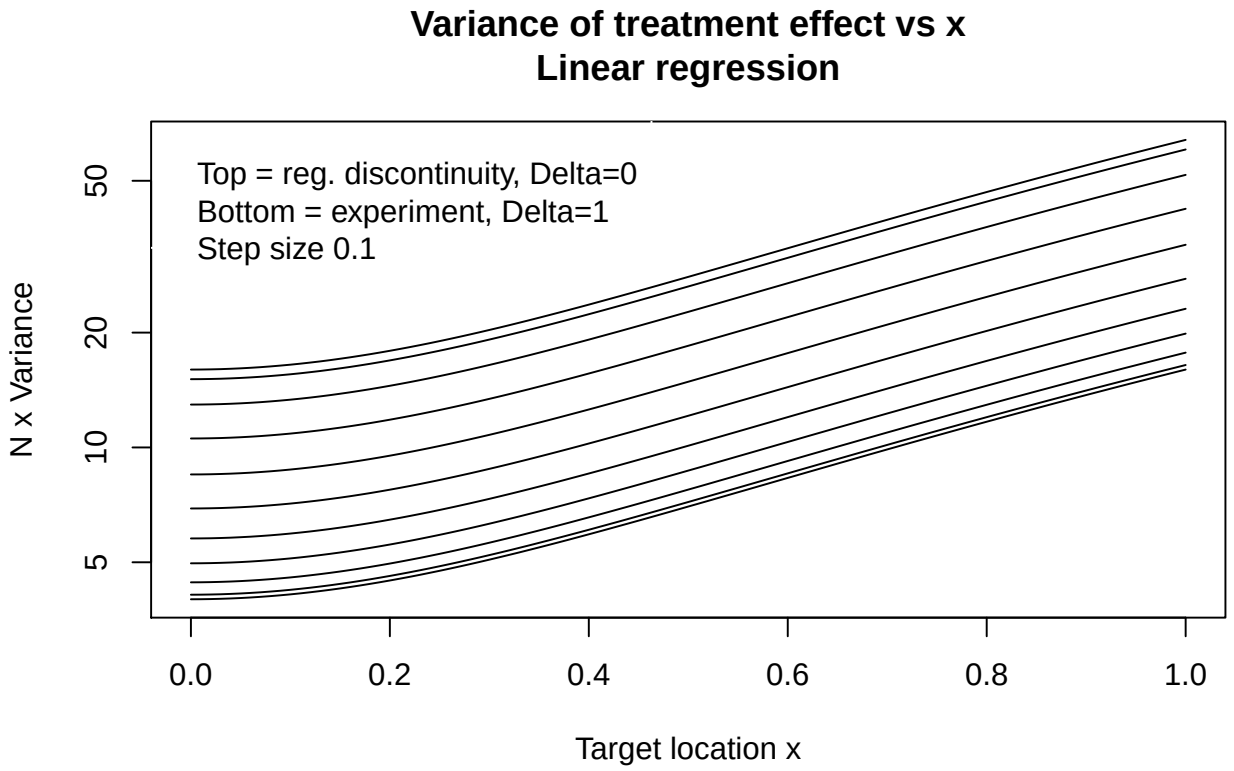
RCT vs RDD

Method	Δ	$\mathrm{Var}(\hat{\beta}_2)$	$\mathrm{Var}(\hat{\beta}_3)$
Regression discontinuity	0	$4/N$	$12/N$
Experiment	1	$1/N$	$3/N$

An RDD with N observations is as good as an RCT with $N/4$ observations.

Section 6 of Jacob, Zhu, Somers & Bloom (2012) has this and more observations.

$$\mathrm{Var}(2(\hat{\beta}_2 + x\hat{\beta}_3))$$



The worst RCT (at $x = 1$) is better than the best RDD (at $x = 0$).

Gain from the sample

The expected payoff per customer in the data set is

$$\frac{1}{N}\sum_{i=1}^{N}\Big(\beta_0+\beta_1 x_i+\beta_2\mathbb{E}(z_i)+\beta_3 x_i\mathbb{E}(z_i)\Big)$$

$$\mathbb{E}(z_i)=\begin{cases}-1, & x_i<-\Delta\\ 0, & |x_i|\leqslant\Delta\\ 1, & x_i>\Delta\end{cases}$$

So plan with

$$\begin{aligned}g(\Delta)\equiv{}&\frac{1}{2}\int_{-1}^{-\Delta}(\beta_0+\beta_1x-\beta_2-\beta_3x)\,\mathrm{d}x+\frac{1}{2}\int_{-\Delta}^{\Delta}(\beta_0+\beta_1x)\,\mathrm{d}x\\&+\frac{1}{2}\int_{\Delta}^{1}\beta_0+\beta_1x+\beta_2+\beta_3x\,\mathrm{d}x\\={}&\beta_0+\beta_3(1-\Delta^2)/2.\end{aligned}$$

The tradeoff

Short term gain per customer

$$g(\Delta) = \beta_0 + \frac{\beta_3(1-\Delta^2)}{2}$$

Define the information gain per customer

$$\text{info}(\Delta) \equiv \frac{1}{N\text{Var}(\hat{\beta}_3)} = \frac{1}{3} - \frac{(1-\Delta^2)^2}{4}$$

Balance

$$\begin{aligned} v(\Delta) &\equiv g(\Delta) + \lambda \times \text{info}(\Delta) \\ &= \beta_0 + \beta_3 \frac{1-\Delta^2}{2} + \lambda\Big(\frac{1}{3} - \frac{(1-\Delta^2)^2}{4}\Big) \end{aligned}$$

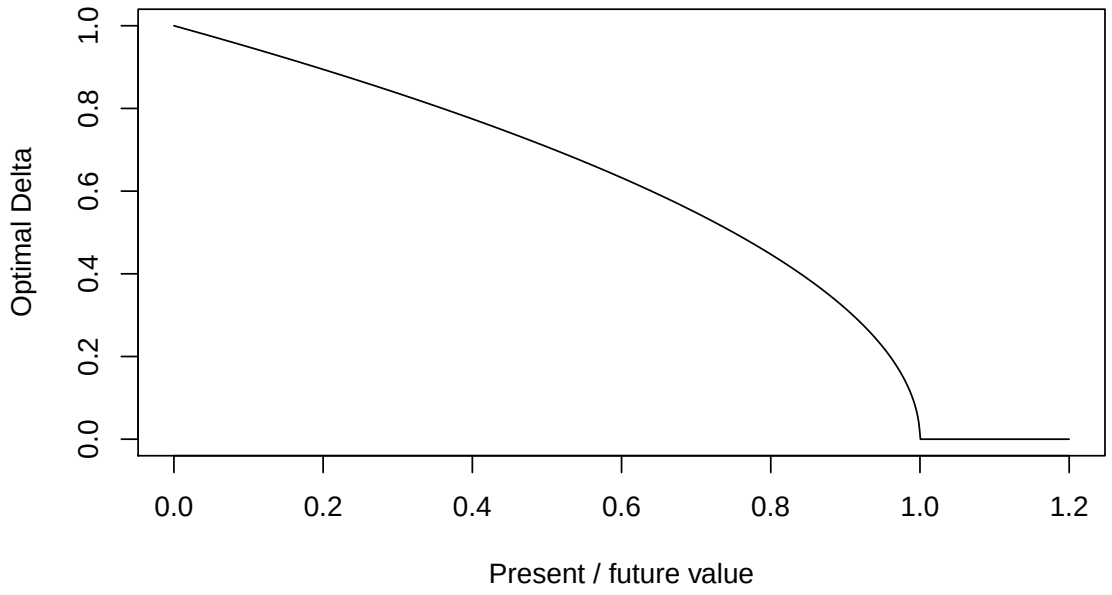
NB: β_0 does not affect our choice of Δ.

Optimal Δ

β_3 is the coefficient of $x_i z_i$

λ is value of information

$$\Delta_* = \begin{cases} 1, & \beta_3/\lambda \leqslant 0 \\ \sqrt{1-\beta_3/\lambda}, & 0 \leqslant \beta_3/\lambda \leqslant 1 \\ 0, & 1 \leqslant \beta_3/\lambda. \end{cases}$$



Value of future information

It is really hard to quantify the value of that information.

Maybe harder than eliciting a prior.

Simpler approach

Let Δ_0 be smallest Δ with efficiency ρ vs RCT $\Delta = 1$.

We know that $1/4 \leqslant \rho \leqslant 1$.

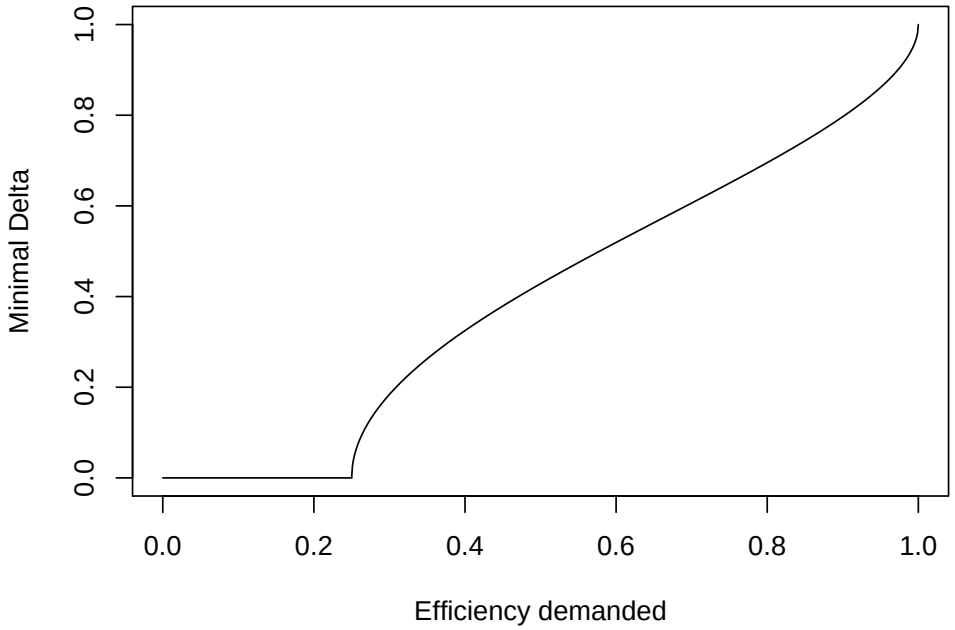
$$\rho = \frac{\mathrm{Var}(2(\hat{\beta}_2 + x\hat{\beta}_3) \mid \Delta = 1)}{\mathrm{Var}(2(\hat{\beta}_2 + x\hat{\beta}_3) \mid \Delta = \Delta_0)} = \cdots = \frac{1 + 3\Delta_0^2(2 - \Delta_0^2)}{1 + 3(2 - 1)}$$

Solve a quadratic equation for Δ_0^2

$$3\Delta_0^4 - 6\Delta_0^2 + 4\rho - 1 = 0$$

$$\implies \quad \Delta_0 = \sqrt{1 - \sqrt{1 - (4\rho - 1)/3}}$$

Minimal Δ for efficiency ρ



ρ	Δ_0
0.99	0.94
0.9	0.80
0.8	0.70
0.7	0.61
0.6	0.52

Gaussian running variable

For $x_i = \Phi^{-1}\left(\frac{i-1/2}{n}\right)$,
experiment on central ΔN observations, then

$$\frac{1}{N}\mathcal{X}^\mathsf{T}\mathcal{X} \approx \begin{array}{c} \\ 1 \\ zx \\ z \\ x \end{array} \begin{array}{c} \begin{array}{cccc} 1 & zx & z & x \end{array} \\ \begin{pmatrix} 1 & \phi_G & 0 & 0 \\ \phi_G & 1 & 0 & 0 \\ 0 & 0 & 1 & \phi_G \\ 0 & 0 & \phi_G & 1 \end{pmatrix} \end{array}$$

$$\phi_G = \mathsf{avg}(x_i z(x_i)) = \cdots = 2\varphi\left(\Phi^{-1}\left(\frac{1+\Delta}{2}\right)\right)$$

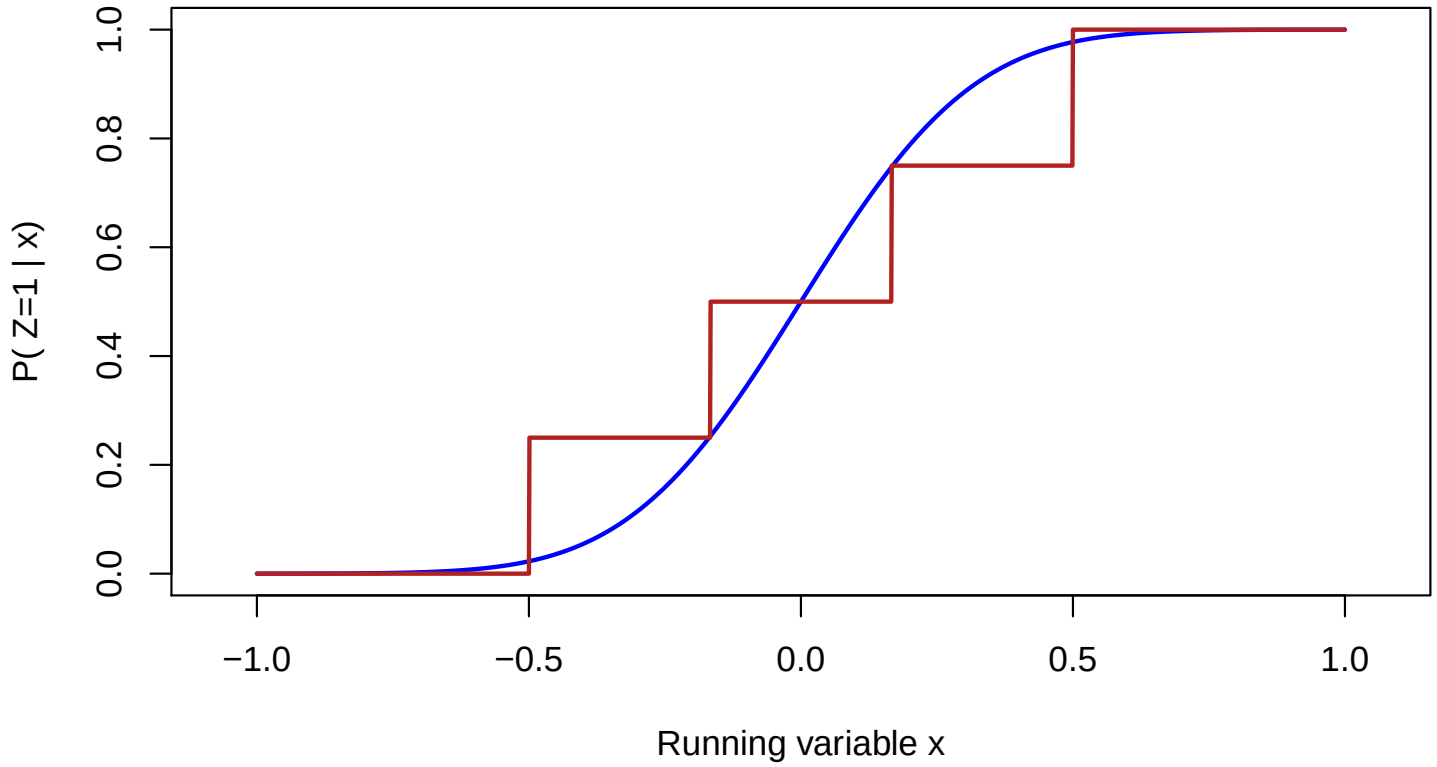
After some algebra, the RDD efficiency vs RDD is

$$\frac{\pi}{\pi - 2} \doteq 2.75$$

Goldberger (1972).

Carpentry

We don't have to keep $p(x) \equiv \Pr(Z = 1 \mid x) \in \{0, 1/2, 1\}$.



Carpentry doesn't really help

(or hurt).

Carpentry ctd

Under symmetry,

$$p(-x) = 1 - p(x),$$

the shape of the curve doesn't matter, only

$$\overline{zx} \equiv \frac{1}{2}\int_{-1}^{1} x\mathbb{E}(Z \mid x)\,\mathrm{d}x = \frac{1}{2}\int_{-1}^{1} x(2p(x)-1)\,\mathrm{d}x > 0$$

short term gain is

$$\beta_0 + \beta_3\overline{zx}$$

information proportional to

$$\frac{1}{3} - \overline{zx}^2$$

Asymmetric p

Replace by a symmetric one. That reduces

$$\mathrm{diag}(\mathrm{Var}(\hat{\beta}))$$

keeping gain the same.

Two quadratics

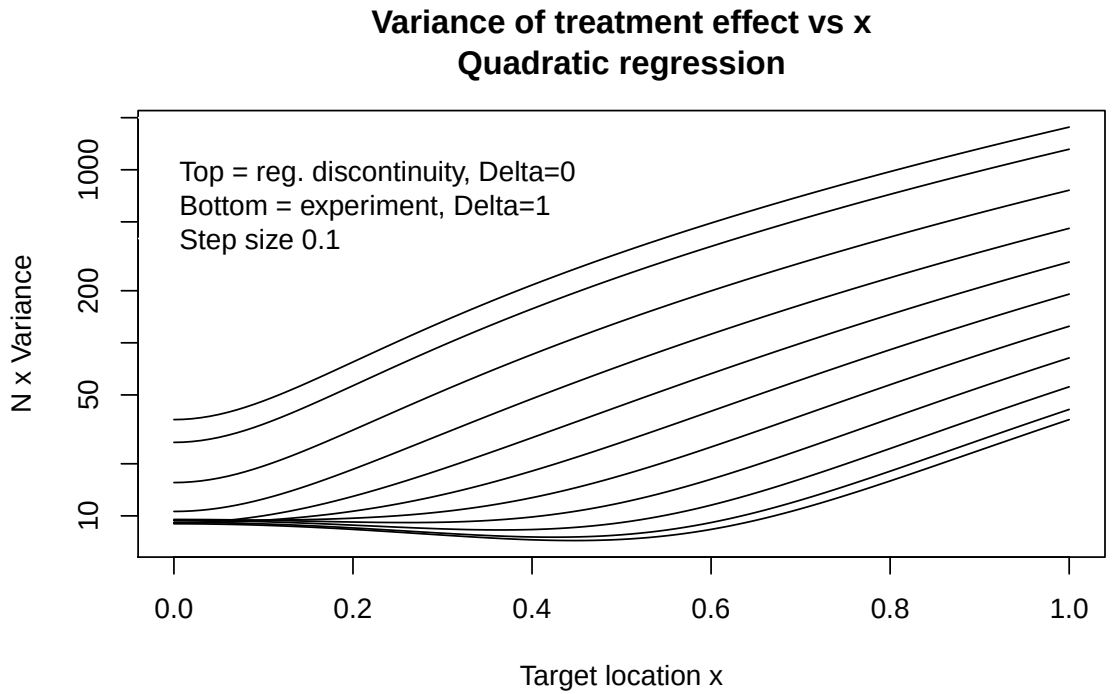
$$\mathbb{E}(Y) = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz + \beta_4 x^2 + \beta_5 x^2 z$$

$$\frac{1}{N}\mathcal{X}^\top\mathcal{X} \doteq \begin{array}{c} \\ 1 \\ zx \\ x^2 \\ z \\ x \\ zx^2 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & zx & x^2 & z & x & zx^2 \end{array} \\ \begin{pmatrix} 1 & \phi_1 & 1/3 & \cdot & \cdot & \cdot \\ \phi_1 & 1/3 & \phi_3 & \cdot & \cdot & \cdot \\ 1/3 & \phi_3 & 1/5 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & 1 & \phi_1 & 1/3 \\ \cdot & \cdot & \cdot & \phi_1 & 1/3 & \phi_3 \\ \cdot & \cdot & \cdot & 1/3 & \phi_3 & 1/5 \end{pmatrix} \end{array}$$

$$\phi_1 = (1 - \Delta^2)/2 \qquad \phi_3 = (1 - \Delta^4)/4$$

Treatment effect

$$\mathbb{E}(y \mid x, z = 1) - \mathbb{E}(y \mid x, z = -1) = 2(\beta_2 + x\beta_3 + x^2\beta_5)$$



Note log scale.

Gelman & Imbens (2017) warn against polynomial RDD.

More elaborate models

For a feature vector $F = F(\boldsymbol{x}) \in \mathbb{R}^d$ including intercept

$$\mathbb{E}(y) = F^{\mathsf{T}}\beta + zF^{\mathsf{T}}\gamma$$

take

$$z_i = \begin{cases} 1, & \theta^{\mathsf{T}}F_i \geqslant \Delta \\ \text{random}, & |\theta^{\mathsf{T}}F_i| < \Delta \\ -1, & \theta^{\mathsf{T}}F_i \leqslant -\Delta. \end{cases}$$

Now

$$\mathcal{X}^{\mathsf{T}}\mathcal{X} = \begin{pmatrix} A & B \\ B & A \end{pmatrix}, \quad A = \sum_i F_i F_i^{\mathsf{T}}, \quad B = \sum_i w_i F_i F_i^{\mathsf{T}},$$

for

$$w_i = \mathbb{E}(z_i \mid F_i) = \begin{cases} 1, & \theta^{\mathsf{T}}F_i \geqslant \Delta, \\ 2p-1, & |\theta^{\mathsf{T}}F_i| < \Delta, \\ -1, & \theta^{\mathsf{T}}F_i \leqslant -\Delta. \end{cases}$$

Inverting block matrices

$$\mathrm{Var}(\hat{\gamma}) = \mathrm{Var}(\hat{\beta}) = (A - BA^{-1}B)^{-1}\sigma^2$$
$$\mathrm{Cov}(\hat{\gamma}, \hat{\beta}) = -A^{-1}B(A - BA^{-1}B)^{-1}\sigma^2$$

We could pick θ, F and p by brute force search with Monte Carlo as an inner loop.

Here matrix algebra can replace the inner Monte Carlo.

Big Δ better

For large enough Δ we get $B = 0$.
Smaller Δ raises $BA^{-1}B$ and hence $\mathrm{Var}(\hat{\beta})$.

Non-central regions

The airline won't give upgrades to half of their passengers.

They are more likely to do:

$$z = \begin{cases} 1, & \text{top few} \\ \text{random}, & \text{next few} \\ -1, & \text{majority.} \end{cases}$$

Of the majority, only retain those where the linear model is ok.

Two lines

Experiment in range (A, B):

Method	A	B	$\mathrm{Var}(\hat{\beta}_3)$
Full experiment	-1.00	1.00	$3.00/N$
RDD	0.00	0.00	$12.00/N$
Expt on bottom 50%	-1.00	0.00	$13.09/N$
Expt on second 10%	0.60	0.80	$137.56/N$
Top 10% only	0.80	0.80	$751.03/N$
Top 15% only	0.70	0.70	$223.44/N$
Top 20% only	0.60	0.60	$95.21/N$

Followup directions

- This x can be the output of a prediction algorithm based on many variables.

  So how does the sampling plan help fit the next model?

  I.e., how to handle concomitants?

- What about binary responses y?

  Logistic regression efficiency actually depends on the underlying β.

  Usual approaches are Bayesian.

Thanks

- Hal Varian, co-author
- Google, environment