Attributing Effects in Cluster Randomized Trials and Observational Studies: The Case of a Get Out The Vote Campaign

Jake Bowers[1] and Ben Hansen[2]

[1] Department of Political Science
University of Michigan

[2] Department of Statistics
University of Michigan

16 September 2005

Outline

Randomization Inference and Attributable Effects

- An Example: Asthma Outbreaks in New Orleans 1957–1959

Effects of a get-out-the-vote campaign: A role for propensity scores?

- The New Haven Vote ’98 Campaign
- A problem with the data, and a proposed solution
- A propensity adapted to per-protocol analysis

Full matching for effects attributable to Vote’98

- Example to illustrate full matching
- Restrictions for full matching via propensity-score calipers
- Effects attributable to telephone entreatments to vote

The revised NH data, and inference accounting for clustering

Outline

- Randomization Inference and Attributable Effects
  - An Example: Asthma Outbreaks in New Orleans 1957–1959
- Effects of a get-out-the-vote campaign: A role for propensity scores?
  - The New Haven Vote '98 Campaign
  - A problem with the data, and a proposed solution
  - A propensity adapted to per-protocol analysis
- Full matching for effects attributable to Vote'98
  - Example to illustrate full matching
  - Restrictions for full matching via propensity-score calipers
  - Effects attributable to telephone entreatments to vote
- The revised NH data, and inference accounting for clustering

Charity Hospital and the Grain Elevators[1]

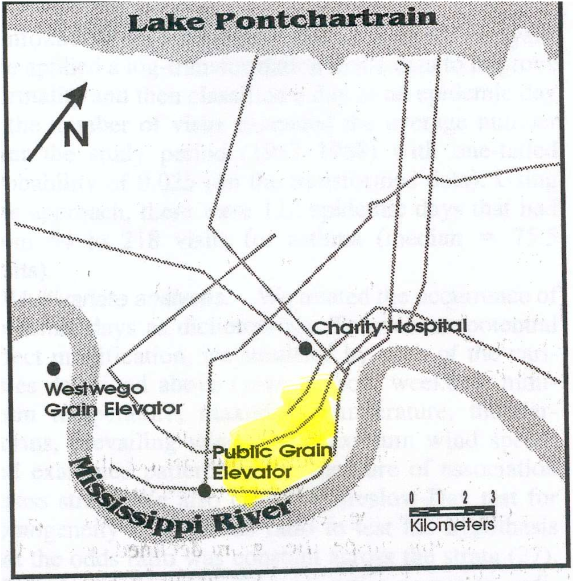


FIGURE 1. Map of the New Orleans harbor area, indicating the location of two grain elevators and Charity Hospital.

Epidemic Asthma in the City of New Orleans

	Days in 1957–59		
	Epidemic	Non-epid.	Total
S-days	32	336	368
N-days	33	694	727
total	65	1030	1095

Of 368 days with soy cargo at the harbor (‘S-days’), 8.7% were epidemic days, whereas only 4.5% of days without soy cargo (‘N-days’) had epidemic flare-ups.

Fisher’s exact test gives $p < .01$: evidently, at least *some* of the asthma epidemics on S-days are attributable to the presence of vessels loading soy cargo.

Epidemic Asthma in the City of New Orleans

	Days in 1957–59		
	Epidemic	Non-epid.	Total
S-days	32	336	368
N-days	33	694	727
total	65	1030	1095

Of 368 days with soy cargo at the harbor (‘S-days’), 8.7% were epidemic days, whereas only 4.5% of days without soy cargo (‘N-days’) had epidemic flare-ups.
Fisher’s exact test gives $p < .01$: evidently, at least *some* of the asthma epidemics on S-days are attributable to the presence of vessels loading soy cargo.

Fisher's Exact Test

- Ball-and-urn experiment, without replacement: see applet.
- Imagine a well-mixed urn containing 65 red balls, $r_i = 1$, and 1030 black balls, $r_i = 0$.
- If $\mathbf{s} \subseteq \{1, \ldots, 1095\}$ is a simple random sample of size 368, then

$$R \equiv \sum_{i \in \mathbf{s}} r_i$$

  has a fully specified, calculable probability distribution over the integers $\{0, 1, \ldots, 65\}$.
- E.g., $\mathbf{E}(R) = 21.8$ and $\sqrt{\mathrm{var}(R)} = 3.7$.
- Fisher's Exact Test rejects if the value of R (number of S-days that were also epidemic days) that is observed lies in the tails of this distribution.
- Here, the observed value of R is 32. . . .

Fisher's Exact Test

- Ball-and-urn experiment, without replacement: see applet.
- Imagine a well-mixed urn containing 65 red balls, $r_i = 1$, and 1030 black balls, $r_i = 0$.
- If $\mathbf{s} \subseteq \{1, \ldots, 1095\}$ is a simple random sample of size 368, then
$$R \equiv \sum_{i \in \mathbf{s}} r_i$$
has a fully specified, calculable probability distribution over the integers $\{0, 1, \ldots, 65\}$.
- E.g., $\mathbf{E}(R) = 21.8$ and $\sqrt{\mathrm{var}(R)} = 3.7$.
- Fisher's Exact Test rejects if the value of R (number of S-days that were also epidemic days) that is observed lies in the tails of this distribution.
- Here, the observed value of R is 32. . . .

Fisher's Exact Test

- Ball-and-urn experiment, without replacement: see applet.
- Imagine a well-mixed urn containing 65 red balls, $r_i = 1$, and 1030 black balls, $r_i = 0$.
- If $\mathbf{s} \subseteq \{1, \ldots, 1095\}$ is a simple random sample of size 368, then

$$R \equiv \sum_{i \in \mathbf{s}} r_i$$

  has a fully specified, calculable probability distribution over the integers $\{0, 1, \ldots, 65\}$.
- E.g., $\mathbf{E}(R) = 21.8$ and $\sqrt{\mathrm{var}(R)} = 3.7$.
- Fisher's Exact Test rejects if the value of R (number of S-days that were also epidemic days) that is observed lies in the tails of this distribution.
- Here, the observed value of R is 32. . . .

Fisher's Exact Test

- Ball-and-urn experiment, without replacement: see applet.
- Imagine a well-mixed urn containing 65 red balls, $r_i = 1$, and 1030 black balls, $r_i = 0$.
- If $\mathbf{s} \subseteq \{1, \ldots, 1095\}$ is a simple random sample of size 368, then

$$R \equiv \sum_{i \in \mathbf{s}} r_i$$

  has a fully specified, calculable probability distribution over the integers $\{0, 1, \ldots, 65\}$.
- E.g., $\mathbf{E}(R) = 21.8$ and $\sqrt{\mathrm{var}(R)} = 3.7$.
- Fisher's Exact Test rejects if the value of R (number of S-days that were also epidemic days) that is observed lies in the tails of this distribution.
- Here, the observed value of R is 32. . . .

Fisher's Exact Test

- Ball-and-urn experiment, without replacement: see applet.
- Imagine a well-mixed urn containing 65 red balls, $r_i = 1$, and 1030 black balls, $r_i = 0$.
- If $\mathbf{s} \subseteq \{1, \ldots, 1095\}$ is a simple random sample of size 368, then

$$R \equiv \sum_{i \in \mathbf{s}} r_i$$

  has a fully specified, calculable probability distribution over the integers $\{0, 1, \ldots, 65\}$.
- E.g., $\mathbf{E}(R) = 21.8$ and $\sqrt{\mathrm{var}(R)} = 3.7$.
- Fisher's Exact Test rejects if the value of R (number of S-days that were also epidemic days) that is observed lies in the tails of this distribution.
- Here, the observed value of R is $32. \ldots$

Fisher's Exact Test

- Ball-and-urn experiment, without replacement: see applet.
- Imagine a well-mixed urn containing 65 red balls, $r_i = 1$, and 1030 black balls, $r_i = 0$.
- If $\mathbf{s} \subseteq \{1, \ldots, 1095\}$ is a simple random sample of size 368, then

  $$R \equiv \sum_{i \in \mathbf{s}} r_i$$

  has a fully specified, calculable probability distribution over the integers $\{0, 1, \ldots, 65\}$.
- E.g., $\mathbf{E}(R) = 21.8$ and $\sqrt{\mathrm{var}(R)} = 3.7$.
- Fisher's Exact Test rejects if the value of R (number of S-days that were also epidemic days) that is observed lies in the tails of this distribution.
- Here, the observed value of R is 32. . . .

Epidemic Days Attributable to Soy: 1957–59

To test $H_0 : A = 0$, apply Fisher's test to:

	Epidemic?		
	Yes	No	
S-days	32	336	368
N-days	33	694	727
total	65	1030	1095

$\Rightarrow p < .05$; rejected.

$H_0 : A = 16$ says: had there been no soy vessels, only a lesser epidemic, as follows, would have occurred.

	Epidemic?		
	Yes	No	
S-days	$32-\mathbf{16}$	$368+\mathbf{16}$	368
N-days	33	694	727
total	65-16	1030+16	1095

$\Rightarrow p = 1$; not rejected.

So the 95% CI must exclude 0 and include 16. (It is [5,24].)

Epidemic Days Attributable to Soy: 1957–59

To test $H_0 : A = 0$, apply Fisher's test to:

	Epidemic?		
	Yes	No	
S-days	32	336	368
N-days	33	694	727
total	65	1030	1095

$\Rightarrow p < .05$; rejected.

$H_0 : A = 16$ says: had there been no soy vessels, only a lesser epidemic, as follows, would have occurred.

	Epidemic?		
	Yes	No	
S-days	$32-\mathbf{16}$	$368+\mathbf{16}$	368
N-days	33	694	727
total	65-16	1030+16	1095

$\Rightarrow p = 1$; not rejected.

So the 95% CI must exclude 0 and include 16. (It is [5,24].)

Epidemic Days Attributable to Soy: 1957–59

To test $H_0 : A = 0$, apply Fisher's test to:

	Epidemic?		
	Yes	No	
S-days	32	336	368
N-days	33	694	727
total	65	1030	1095

$\Rightarrow p < .05$; rejected.

$H_0 : A = 16$ says: had there been no soy vessels, only a lesser epidemic, as follows, would have occurred.

	Epidemic?		
	Yes	No	
S-days	$32 - \mathbf{16}$	$368 + \mathbf{16}$	368
N-days	33	694	727
total	65-16	1030+16	1095

$\Rightarrow p = 1$; not rejected.

So the 95% CI must exclude 0 and include 16. (It is [5,24].)

Epidemic Days Attributable to Soy: 1957–59

To test $H_0 : A = 0$, apply Fisher's test to:

	Epidemic?		
	Yes	No	
S-days	32	336	368
N-days	33	694	727
total	65	1030	1095

$\Rightarrow p < .05$; rejected.

$H_0 : A = 16$ says: had there been no soy vessels, only a lesser epidemic, as follows, would have occurred.

	Epidemic?		
	Yes	No	
S-days	$32-\mathbf{16}$	$368+\mathbf{16}$	368
N-days	33	694	727
total	65-16	1030+16	1095

$\Rightarrow p = 1$; not rejected.

So the 95% CI must exclude 0 and include 16. (It is [5,24].)

Epidemic Days Attributable to Soy: 1957–59

To test $H_0 : A = 0$, apply Fisher's test to:

	Epidemic?		
	Yes	No	
S-days	32	336	368
N-days	33	694	727
total	65	1030	1095

$\Rightarrow p < .05$; rejected.

$H_0 : A = 16$ says: had there been no soy vessels, only a lesser epidemic, as follows, would have occurred.

	Epidemic?		
	Yes	No	
S-days	$32-\mathbf{16}$	$368+\mathbf{16}$	368
N-days	33	694	727
total	65-16	1030+16	1095

$\Rightarrow p = 1$; not rejected.

So the 95% CI must exclude 0 and include 16. (It is [5,24].)

Epidemic Days Attributable to Soy: 1957–68

Days in 1957–59[a]

	Epidemic?		
	Yes	No	
S	$32 - A_f$	$336 + A_f$	368
N	33	694	727
	$65 - A_f$	$1030 + A_f$	1095

[a] $A_f \equiv$ Soy-attributed asthma flare-ups during the fifties

Days in 1960–68[a]

	Epidemic?		
	Yes	No	
S	$43 - A_s$	$1424 + A_s$	1467
N	12	1809	1821
	$55 - A_s$	$3233 + A_s$	

[a] $A_s \equiv$ Soy-attributed asthma flare-ups during the sixties

Calculating a 95% CI for the Attributable Effect

generalizing the Fisher-test approach to multiple strata

By repetition of the Mantel-Haenszel test, $[29, 61]$. To wit:

$\vdots$ $\vdots$

$H_0 : A_f + A_s = 28$

$A_f =$	28 …	1	0
$A_s =$	0 …	27	28
p	.039 …	.006	.005

$H_0 : A_f + A_s = 29$

$A_f =$	29 …	1	0
$A_s =$	0 …	28	29
p	.054 …	.008	.007

$\vdots$ $\vdots$

$H_0 : A_f + A_s = 61$

$A_f =$	32	31 …	18
$A_s =$	29	30 …	43
p	.026	.028 …	.065

$H_0 : A_f + A_s = 62$

$A_f =$	32	31 …	19
$A_s =$	30	31 …	43
p	.016	.018 …	.041

$\vdots$ $\vdots$

…where

- $A_f = \#$ of epidemic days, 1957 to 1959, attributed to soy;
- $A_s = \#$ of epidemic days, 1960 to 1968, attributed to soy.

Calculating a 95% CI for the Attributable Effect

generalizing the Fisher-test approach to multiple strata

By repetition of the Mantel-Haenszel test, [29, 61]. To wit:

$\vdots$ $\vdots$

$H_0 : A_f + A_s = 28$

$A_f =$	28 . . .	1	0
$A_s =$	0 . . .	27	28
p	.039 . . .	.006	.005

$H_0 : A_f + A_s = 29$

$A_f =$	29 . . .	1	0
$A_s =$	0 . . .	28	29
p	.054 . . .	.008	.007

$\vdots$ $\vdots$

$H_0 : A_f + A_s = 61$

$A_f =$	32	31 . . .	18
$A_s =$	29	30 . . .	43
p	.026	.028 . . .	.065

$H_0 : A_f + A_s = 62$

$A_f =$	32	31 . . .	19
$A_s =$	30	31 . . .	43
p	.016	.018 . . .	.041

$\vdots$ $\vdots$

. . . where

- $A_f = \#$ of epidemic days, 1957 to 1959, attributed to soy;
- $A_s = \#$ of epidemic days, 1960 to 1968, attributed to soy.

Summary: Randomization Inference Workflow

1. Specify a null hypothesis about the attributable effect (*e.g.*, 100 people voted due to phone calls)
2. Enumerate corresponding “atomic” hypotheses (*e.g., these* 100 people voted due to phone)
3. Test each null with a probability model for treatment assignment which defines the exact sampling distribution of your test statistic.
4. The null regarding the a.e. is accepted if any of its atomic components are.
5. 95% Confidence Interval = a.e. values accepted at the .05 level.

We have seen the application of this to studies with binary outcomes and a treatment and a control group, with and without stratification. (2×2 and $2 \times 2 \times K$ tables.)

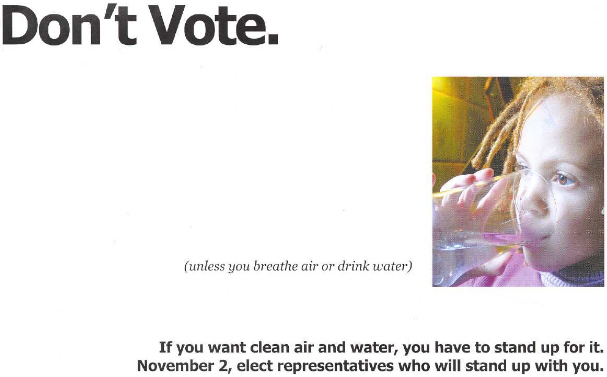
Outline

- Randomization Inference and Attributable Effects
  - An Example: Asthma Outbreaks in New Orleans 1957–1959
- Effects of a get-out-the-vote campaign: A role for propensity scores?
  - The New Haven Vote ’98 Campaign
  - A problem with the data, and a proposed solution
  - A propensity adapted to per-protocol analysis
- Full matching for effects attributable to Vote’98
  - Example to illustrate full matching
  - Restrictions for full matching via propensity-score calipers
  - Effects attributable to telephone entreatments to vote
- The revised NH data, and inference accounting for clustering

Outline

- Randomization Inference and Attributable Effects
  - An Example: Asthma Outbreaks in New Orleans 1957–1959
- Effects of a get-out-the-vote campaign: A role for propensity scores?
  - The New Haven Vote ’98 Campaign
  - A problem with the data, and a proposed solution
  - A propensity adapted to per-protocol analysis
- Full matching for effects attributable to Vote’98
  - Example to illustrate full matching
  - Restrictions for full matching via propensity-score calipers
  - Effects attributable to telephone entreatments to vote
- The revised NH data, and inference accounting for clustering

Example of a Nonpartisan Get-Out-The-Vote Mailer



Don't Vote.

(unless you breathe air or drink water)

**If you want clean air and water, you have to stand up for it.
November 2, elect representatives who will stand up with you.**

New Haven Vote ‘98 Campaign: A 2 × 2 × 2 Full Factorial Design

Telephone reminder		Mailings		
	Personal canvass	No Mail	Mail	Total
No Telephone				
	No Visit	11600	7800	19400
	Visit	2900	1900	4700
Telephone				
	No Visit	800	4700	5600
	Visit	200	1200	1400

cf. Gerber, A.S. and Green, D.P. (2000, 2005), *Am. Polit. Sci. Rev.*

Despite Random Assignment, Compliance Was Not Random

	Compliance Rate		
	No Mail Grp	Mail Group	Total
No Phone Assigned, Answered Door	33%	42%	37%
Answered Phone, No Visit Assigned	36%	40%	39%
Answered Phone, Answered door	20%	12%	13%

cf. Gerber, A.S. and Green, D.P. (2000, 2005), *Am. Polit. Sci. Rev.*

Outline

- Randomization Inference and Attributable Effects
  - An Example: Asthma Outbreaks in New Orleans 1957–1959
- Effects of a get-out-the-vote campaign: A role for propensity scores?
  - The New Haven Vote ’98 Campaign
  - A problem with the data, and a proposed solution
  - A propensity adapted to per-protocol analysis
- Full matching for effects attributable to Vote’98
  - Example to illustrate full matching
  - Restrictions for full matching via propensity-score calipers
  - Effects attributable to telephone entreatments to vote
- The revised NH data, and inference accounting for clustering

A Generalized Ball-And-Urn Comparison

With Which to Test for Balance on Pre-Treatment Variables

Let $x = (x_1, \ldots, x_n)^t$ be a pre-treatment variable measured on all subjects. If **t** is a simple random sample of m indices from $\{1, \ldots, n\}$, and **c** contains the $n - m$ indices left over, then

$$\begin{aligned}\bar{x}_{\mathbf{t}} - \bar{x}_{\mathbf{c}} &= \frac{1}{m}\sum_{i\in\mathbf{t}} x_i - \frac{1}{n-m}\sum_{i\in\mathbf{c}} x_i \\ &= \frac{n}{(n-m)}\left[\frac{1}{m}\sum_{i\in\mathbf{t}} x_i - \frac{1}{n}\sum_{i=1}^{n} x_i\right].\end{aligned}$$

— draw balls bearing numbers x_i from an urn, then shift and scale the result. This random quantity has expectation 0 and a variance v_x that is readily calculated. It is nearly Normal in moderate to large samples. Its magnitude ought to be no greater than a few times v_x.

On this basis, we test whether randomization succeeded in balancing each covariate x.

A Generalized Ball-And-Urn Comparison

With Which to Test for Balance on Pre-Treatment Variables

Let $x = (x_1, \ldots, x_n)^t$ be a pre-treatment variable measured on all subjects. If $\mathbf{t}$ is a simple random sample of m indices from $\{1, \ldots, n\}$, and $\mathbf{c}$ contains the $n - m$ indices left over, then

$$\begin{aligned} \bar{x}_{\mathbf{t}} - \bar{x}_{\mathbf{c}} &= \frac{1}{m} \sum_{i \in \mathbf{t}} x_i - \frac{1}{n-m} \sum_{i \in \mathbf{c}} x_i \\ &= \frac{n}{(n-m)} \left[\frac{1}{m} \sum_{i \in \mathbf{t}} x_i - \frac{1}{n} \sum_{i=1}^{n} x_i \right]. \end{aligned}$$

— draw balls bearing numbers x_i from an urn, then shift and scale the result. This random quantity has expectation 0 and a variance v_x that is readily calculated. It is nearly Normal in moderate to large samples. Its magnitude ought to be no greater than a few times v_x.

On this basis, we test whether randomization succeeded in balancing each covariate x.

Standardized Differences on Selected Covariates

A comparison of groups assigned to treatment and to control, ignoring non-compliance

	Standardized Bias (sds)	
persons1	.02	
persons2	-.02	
v96.abst	-.02	
v96.vote	.01	
majpty	.00	
age.Bspline1	.00	
age.Bspline6	-.02	
Ward2	.00	
Ward3	-.05	⋆ ⋆ ⋆
Ward8	.03	·
Ward12	.03	·
Ward15	.03	⋆
Ward17	.04	⋆⋆
Ward19	-.03	⋆
Ward30	.01	

χ^2 = 58, df = 38, p-value = .02

An As-Treated Analysis To Address Shortcomings of the Randomization[2]

Using Matching on the Basis of a Propensity Score

Telephone reminder	Number of Direct Mailings Sent					Contact Rates	
Personal canvass	0	1	2	3	Total	in-person	phone
No telephone							
No in-person	11000	2500	2700	2400	18500	-	-
In-person	2700	500	600	700	4500	28%	-
Telephone							
No in-person	~~800~~250	1400	1400	1500	5100	-	35%
In-person	200	400	300	400	1200	28%	34%

[2] Proposed by Imai (2005), *Am. Polit. Sci. Rev.*.

Outline

- Randomization Inference and Attributable Effects
  - An Example: Asthma Outbreaks in New Orleans 1957–1959

- Effects of a get-out-the-vote campaign: A role for propensity scores?
  - The New Haven Vote ’98 Campaign
  - A problem with the data, and a proposed solution
  - A propensity adapted to per-protocol analysis

- Full matching for effects attributable to Vote’98
  - Example to illustrate full matching
  - Restrictions for full matching via propensity-score calipers
  - Effects attributable to telephone entreatments to vote

- The revised NH data, and inference accounting for clustering

The Assignment Score: A Propensity Adapted to Retaining Treatment Assignment as an IV

By the *assignment score*, I mean the fitted propensity to belong to the group slated for telephone treatment, given covariates and also conditionally on assignment and receipt of complementary treatments (in-person and by-mail reminders). This differs from the propensity to have been assigned to treatment *and* to have received it, which Imai's (2005) analysis used.

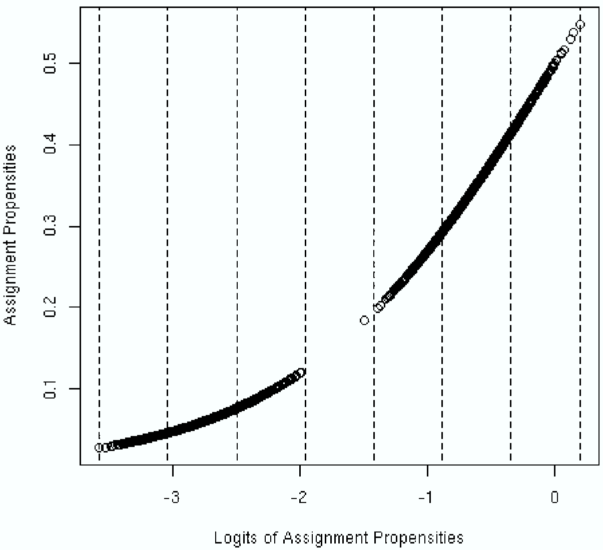
Propensity score specifications should err on the side of including terms in the name of comprehensiveness rather than excluding them in the interests of parsimony (Rubin and Thomas 1996, *Biometrics*), so I include interactions and squares of regressors in a logistic regression of treatment assignment on covariates and complementary treatments. But I reduce the number of second-order terms using stepwise regression and the AIC — a labor-saving variation of Rosenbaum and Rubin's (1984, *JASA*) method.

The Assignment Score: A Propensity Adapted to Retaining Treatment Assignment as an IV

By the *assignment score*, I mean the fitted propensity to belong to the group slated for telephone treatment, given covariates and also conditionally on assignment and receipt of complementary treatments (in-person and by-mail reminders). This differs from the propensity to have been assigned to treatment *and* to have received it, which Imai's (2005) analysis used.

Propensity score specifications should err on the side of including terms in the name of comprehensiveness rather than excluding them in the interests of parsimony (Rubin and Thomas 1996, *Biometrics*), so I include interactions and squares of regressors in a logistic regression of treatment assignment on covariates and complementary treatments. But I reduce the number of second-order terms using stepwise regression and the AIC — a labor-saving variation of Rosenbaum and Rubin's (1984, *JASA*) method.

To Balance Covariates, Bin Widths $\leq .5s_p$ Are Needed



Outline

Randomization Inference and Attributable Effects

- An Example: Asthma Outbreaks in New Orleans 1957–1959

Effects of a get-out-the-vote campaign: A role for propensity scores?

- The New Haven Vote ’98 Campaign
- A problem with the data, and a proposed solution
- A propensity adapted to per-protocol analysis

Full matching for effects attributable to Vote’98

- Example to illustrate full matching
- Restrictions for full matching via propensity-score calipers
- Effects attributable to telephone entreatments to vote

The revised NH data, and inference accounting for clustering

Outline

Randomization Inference and Attributable Effects

- An Example: Asthma Outbreaks in New Orleans 1957–1959

Effects of a get-out-the-vote campaign: A role for propensity scores?

- The New Haven Vote ’98 Campaign
- A problem with the data, and a proposed solution
- A propensity adapted to per-protocol analysis

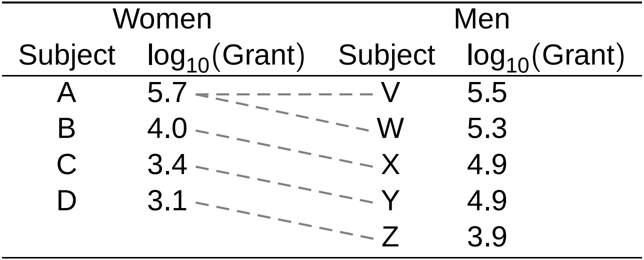
Full matching for effects attributable to Vote’98

- Example to illustrate full matching
- Restrictions for full matching via propensity-score calipers
- Effects attributable to telephone entreatments to vote

The revised NH data, and inference accounting for clustering

A Small Matching Problem from a Gender-Equity Study[3]

Women and men scientists are to be matched on grant funding. The solution that is optimal among 1:(1 or 2) matches is shown.

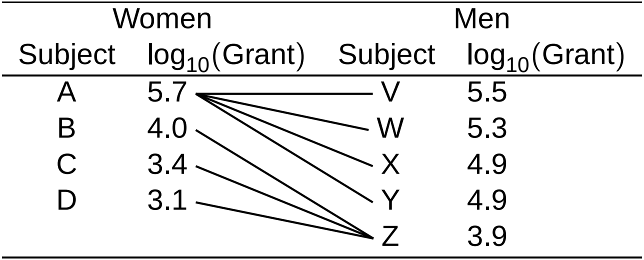


Women		Men	
Subject	$\log_{10}$(Grant)	Subject	$\log_{10}$(Grant)
A	5.7	V	5.5
B	4.0	W	5.3
C	3.4	X	4.9
D	3.1	Y	4.9
		Z	3.9

The optimal 1:(1 or 2) match has its largest matched discrepancy 1.5, at {C, Y}. The sum of its discrepancies is 3.8.

[3] Discussed in Hansen, B. and S. Olsen Klopfer (2005), ‘Optimal full matching and related designs via network flows’, U. of M. Statistics Dept. Technical Report; see also Hansen (2004), *JASA*.

An Optimal Full Matching for the Gender Equity Example



Women		Men	
Subject	$\log_{10}$(Grant)	Subject	$\log_{10}$(Grant)
A	5.7	V	5.5
B	4.0	W	5.3
C	3.4	X	4.9
D	3.1	Y	4.9
		Z	3.9

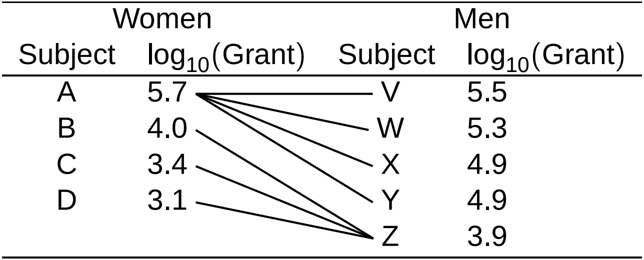
As compared to the optimal 1:(1 or 2) match, full matching:

- decreases the largest discrepancy from 1.5 to 0.8; and
- decreases the sum of discrepancies from 3.8 to 3.6.

In addition, it respects a *caliper*[4] of 1.0 on $\log_{10}$(Grant).
In global terms, it gives a tighter match. In local terms, it gives a *much* tighter match.

[4] Althauser & Rubin, 1971, *Am. J. Sociol.*

An Optimal Full Matching for the Gender Equity Example



Women		Men	
Subject	$\log_{10}$(Grant)	Subject	$\log_{10}$(Grant)
A	5.7	V	5.5
B	4.0	W	5.3
C	3.4	X	4.9
D	3.1	Y	4.9
		Z	3.9

As compared to the optimal 1:(1 or 2) match, full matching:

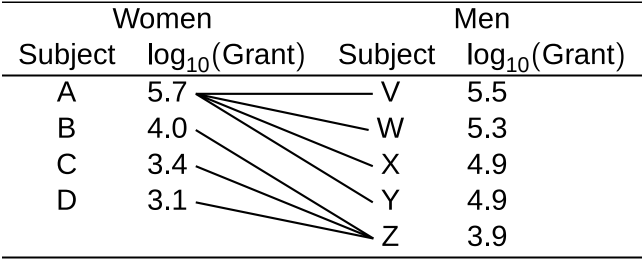
- decreases the largest discrepancy from 1.5 to 0.8; and
- decreases the sum of discrepancies from 3.8 to 3.6.

In addition, it respects a *caliper*[4] of 1.0 on $\log_{10}$(Grant).

In global terms, it gives a tighter match. In local terms, it gives a *much* tighter match.

[4] Althauser & Rubin, 1971, *Am. J. Sociol.*

An Optimal Full Matching for the Gender Equity Example



Women		Men	
Subject	$\log_{10}$(Grant)	Subject	$\log_{10}$(Grant)
A	5.7	V	5.5
B	4.0	W	5.3
C	3.4	X	4.9
D	3.1	Y	4.9
		Z	3.9

As compared to the optimal 1:(1 or 2) match, full matching:

- decreases the largest discrepancy from 1.5 to 0.8; and
- decreases the sum of discrepancies from 3.8 to 3.6.

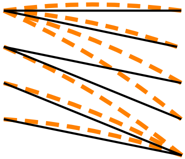
In addition, it respects a *caliper*[4] of 1.0 on $\log_{10}$(Grant).
In global terms, it gives a tighter match. In local terms, it gives a *much* tighter match.

[4] Althauser & Rubin, 1971, *Am. J. Sociol.*

Full Matching versus Full Matching with Restrictions

Full match *with restrictions* for the gender equity example, solid lines, as compared to the full match without restrictions, dashed lines. In this example, the restrictions are that control-to-treatment ratios be constrained to lie between 1/2 and 2.

Women		Men	
Subject	$\log_{10}$(Grant)	Subject	$\log_{10}$(Grant)
A	5.7	V	5.5
B	4.0	W	5.3
C	3.4	X	4.9
D	3.1	Y	4.9
		Z	3.9



In terms of both local and local discrepancy, full matching with restrictions is intermediate to 1:(1 or 2) and full matching. Locally, it is nearly as good as unrestricted full matching.

Outline

Randomization Inference and Attributable Effects

- An Example: Asthma Outbreaks in New Orleans 1957–1959

Effects of a get-out-the-vote campaign: A role for propensity scores?

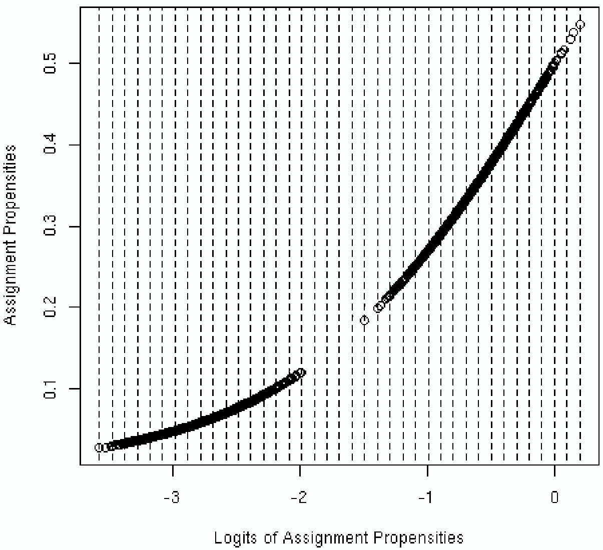
- The New Haven Vote '98 Campaign
- A problem with the data, and a proposed solution
- A propensity adapted to per-protocol analysis

Full matching for effects attributable to Vote'98

- Example to illustrate full matching
- Restrictions for full matching via propensity-score calipers
- Effects attributable to telephone entreatments to vote

The revised NH data, and inference accounting for clustering

To Balance the Score Itself, Widths $\leq .1s_p$ Are Needed



To Get Matched Sets Spanning No More Than .1s_p, We Use a Caliper of .05s_p

Discrepancies associated with each potential match are represented in a #{Treatment} by #{Control} matrix, `disc`, say. With ‘optmatch’, my add-on package for R, one forbids matches by placing `Inf` or NA at the appropriate locations in the discrepancy matrix. To have matching be done so as to minimize discrepancies encoded in `disc`, but also respecting a caliper of .05 on a discrepancy measure `caldisc`, one creates a discrepancy matrix encoding these aims as follows:

```
> disc2 <- ifelse(caldisc<.05, disc, Inf)
```

(To match without regard to discrepancies but respecting the caliper requirements, one would set `disc` equal to a constant.)

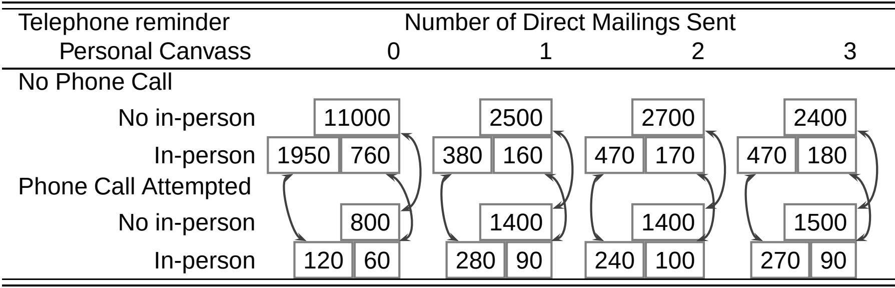
To Get Matched Sets Spanning No More Than $.1s_p$, We Use a Caliper of $.05s_p$

Discrepancies associated with each potential match are represented in a #{Treatment} by #{Control} matrix, `disc`, say. With ‘optmatch’, my add-on package for R, one forbids matches by placing `Inf` or NA at the appropriate locations in the discrepancy matrix. To have matching be done so as to minimize discrepancies encoded in `disc`, but also respecting a caliper of .05 on a discrepancy measure `caldisc`, one creates a discrepancy matrix encoding these aims as follows:

```
> disc2 <- ifelse(caldisc<.05, disc, Inf)
```

(To match without regard to discrepancies but respecting the caliper requirements, one would set `disc` equal to a constant.)

To Isolate Effects from Those of Complementary Treatments, I Subclassify Prior to Matching



Telephone reminder Personal Canvass	Number of Direct Mailings Sent 0		1		2		3	
No Phone Call								
No in-person		11000		2500		2700		2400
In-person	1950	760	380	160	470	170	470	180
Phone Call Attempted								
No in-person		800		1400		1400		1500
In-person	120	60	280	90	240	100	270	90

Because of the caliper requirement and because of matching within subclasses, nine treatment subjects and 116 controls lack a suitable match. The default in 'optmatch' is to exclude such subjects from the matching. As these 125 are only .4% of the sample, that is what I did. Alternatively, Rosenbaum and Rubin (1985, *Amer. Statist.*) would have relaxed the caliper requirement for the nine treatment subjects in order to include them too.

To Isolate Effects from Those of Complementary Treatments, I Subclassify Prior to Matching

Telephone reminder	Number of Direct Mailings Sent							
Personal Canvass	0		1		2		3	
No Phone Call								
No in-person		11000		2500		2700		2400
In-person	1950	760	380	160	470	170	470	180
Phone Call Attempted								
No in-person		800		1400		1400		1500
In-person	120	60	280	90	240	100	270	90

Because of the caliper requirement and because of matching within subclasses, nine treatment subjects and 116 controls lack a suitable match. The default in ‘optmatch’ is to exclude such subjects from the matching. As these 125 are only .4% of the sample, that is what I did. Alternatively, Rosenbaum and Rubin (1985, *Amer. Statist.*) would have relaxed the caliper requirement for the nine treatment subjects in order to include them too.

Matching Restrictions From a Caliper

First, a line search within each of the 12 subclasses for the largest number that can limit from below the number of controls per treatment subject in a matched set:

```
> minc <- minControlsCap(disc2)
```

Next, another line search within each subclass for the smallest feasible upper limit on the number of controls per treatment subject, per matched set. This line search takes into account the results of the first.

```
> maxc <- maxControlsCap(disc2, min.controls=
+             minc$strictest.feasible.min.controls)
```

A full match with both restrictions may then be obtained by:

```
> fmr <- fullmatch(disc2,min.controls=
+                 minc$strictest,
+                 max.controls=maxc$strictest)
```

Matching Restrictions From a Caliper

First, a line search within each of the 12 subclasses for the largest number that can limit from below the number of controls per treatment subject in a matched set:

```
> minc <- minControlsCap(disc2)
```

Next, another line search within each subclass for the smallest feasible upper limit on the number of controls per treatment subject, per matched set. This line search takes into account the results of the first.

```
> maxc <- maxControlsCap(disc2, min.controls=
+             minc$strictest.feasible.min.controls)
```

A full match with both restrictions may then be obtained by:

```
> fmr <- fullmatch(disc2,min.controls=
+                  minc$strictest,
+                  max.controls=maxc$strictest)
```

Matched Set Configurations To Result from Full Matching With Restrictions

	Number of direct mailings sent			
	0	1	2	3
Personal canvass not attempted				
	7 to 14	1 to 3	1 to 3	1 to 3
Assigned to receive personal canvass				
Contact occurred	7 to 14	$\frac{1}{2}$ to 3	1 to 6	$\frac{1}{2}$ to 4
Not contacted	11 to 23	1 to 6	$\frac{1}{2}$ to 5	1 to 3

The table indicates, by subclass, the number of controls matched to each treatment subject. (Where it reads $\frac{1}{2}$, pairs of treatment subjects had to share a control in order to meet requirements imposed by the caliper.)

Outline

Randomization Inference and Attributable Effects

- An Example: Asthma Outbreaks in New Orleans 1957–1959

Effects of a get-out-the-vote campaign: A role for propensity scores?

- The New Haven Vote ’98 Campaign
- A problem with the data, and a proposed solution
- A propensity adapted to per-protocol analysis

Full matching for effects attributable to Vote’98

- Example to illustrate full matching
- Restrictions for full matching via propensity-score calipers
- Effects attributable to telephone entreatments to vote

The revised NH data, and inference accounting for clustering

Can The Hypothesis of No Attributable Effect Be Rejected?

Under the hypothesis that $A = 0$, precisely those treatment-group subjects who actually voted would have voted in the absence of treatment: for each subject i, $y_{ci} = y_i$. The plausibility of this can be assessed with a Mantel-Haenszel test for independence of treatment assignment and voting in the 1998 election, conditional upon matched sets.

The z-statistic that results is $-.6$ ($p = .53$). The hypothesis of no effect cannot be rejected.

Can The Hypothesis of No Attributable Effect Be Rejected?

Under the hypothesis that $A = 0$, precisely those treatment-group subjects who actually voted would have voted in the absence of treatment: for each subject i, $y_{ci} = y_i$. The plausibility of this can be assessed with a Mantel-Haenszel test for independence of treatment assignment and voting in the 1998 election, conditional upon matched sets.

The z-statistic that results is $-.6$ ($p = .53$). The hypothesis of no effect cannot be rejected.

A Two-Sided Confidence Interval for the Attributable Effect

A confidence interval for the attributable effect requires inverting a family of hypothesis tests. Each of the hypotheses has the form $H_0 : A = A_0$, and comprises a number of hypotheses that could be tested using Mantel-Haenszel. For $H_0 : A = 0$, there is just one embedded hypothesis, but for all others there are many more. Using asymptotic separability, as explained by Rosenbaum (2002, *JASA*), we find that $H_0 : A = 0, 1, \ldots, 84$ are not rejected at the 5% level, two-sided, while hypotheses of 85 or more votes caused by telephone reminders are rejected. With 95% confidence, at most 3.8% of those contacted by telephone were induced to vote because of it.

(By supposing that telephone reminders suppressed voting rather than encouraging it, negative attributions can be considered. The 95% confidence interval that results from this spans -159 through 0 — which would allege that as many as 7.3% of those called may have been induced *not* to vote by the telephone reminder.)

A Two-Sided Confidence Interval for the Attributable Effect

A confidence interval for the attributable effect requires inverting a family of hypothesis tests. Each of the hypotheses has the form $H_0 : A = A_0$, and comprises a number of hypotheses that could be tested using Mantel-Haenszel. For $H_0 : A = 0$, there is just one embedded hypothesis, but for all others there are many more. Using asymptotic separability, as explained by Rosenbaum (2002, *JASA*), we find that $H_0 : A = 0, 1, \ldots, 84$ are not rejected at the 5% level, two-sided, while hypotheses of 85 or more votes caused by telephone reminders are rejected. With 95% confidence, at most 3.8% of those contacted by telephone were induced to vote because of it.

(By supposing that telephone reminders suppressed voting rather than encouraging it, negative attributions can be considered. The 95% confidence interval that results from this spans -159 through 0 — which would allege that as many as 7.3% of those called may have been induced *not* to vote by the telephone reminder.)

Outline

Randomization Inference and Attributable Effects

- An Example: Asthma Outbreaks in New Orleans 1957–1959

Effects of a get-out-the-vote campaign: A role for propensity scores?

- The New Haven Vote '98 Campaign
- A problem with the data, and a proposed solution
- A propensity adapted to per-protocol analysis

Full matching for effects attributable to Vote'98

- Example to illustrate full matching
- Restrictions for full matching via propensity-score calipers
- Effects attributable to telephone entreatments to vote

The revised NH data, and inference accounting for clustering

Another variation on the Ball-And-Urn Experiment

To account for group random assignment

The Importance of Adjusting for Clustering

	Standardized Bias (sds)			
	Accounting for Clustering?			
Covariate	No		Yes	
persons1	.02		.02	
persons2	-.02		-.02	
v96.abst	-.02		-.04	
v96.vote	.01		.01	
majpty	.00		-.05	
age.Bspline1	.00		-.02	
age.Bspline6	-.02		-.02	
Ward2	.00		-.00	
Ward3	-.05	⋆ ⋆ ⋆	-.12	⋆⋆
Ward8	.03	·	.05	
Ward12	.03	·	.05	
Ward15	.03	⋆	.05	
Ward17	.04	⋆⋆	.08	⋆
Ward19	-.03	⋆	-.07	·
Ward30	.01		.01	

No Cluster: χ^2 = 58, df = 38, p-value = .02

Yes Cluster: χ^2 = 39.6, df = 38, p-value =.40

The Importance of Adjusting for Clustering

	Standardized Bias (sds)			
	Accounting for Clustering?			
Covariate	No		Yes	
persons1	.02		.02	
persons2	-.02		-.02	
v96.abst	-.02		-.04	
v96.vote	.01		.01	
majpty	.00		-.05	
age.Bspline1	.00		-.02	
age.Bspline6	-.02		-.02	
Ward2	.00		-.00	
Ward3	-.05	⋆⋆⋆	-.12	⋆⋆
Ward8	.03	·	.05	
Ward12	.03	·	.05	
Ward15	.03	⋆	.05	
Ward17	.04	⋆⋆	.08	⋆
Ward19	-.03	⋆	-.07	·
Ward30	.01		.01	

No Cluster: χ^2 = 58, df = 38, p-value = .02

Yes Cluster: χ^2 = 39.6, df = 38, p-value =.40

The Importance of Adjusting for Clustering

	Standardized Bias (sds)			
	Accounting for Clustering?			
Covariate	No		Yes	
persons1	.02		.02	
persons2	-.02		-.02	
v96.abst	-.02		-.04	
v96.vote	.01		.01	
majpty	.00		-.05	
age.Bspline1	.00		-.02	
age.Bspline6	-.02		-.02	
Ward2	.00		-.00	
Ward3	-.05	⋆⋆⋆	-.12	⋆⋆
Ward8	.03	·	.05	
Ward12	.03	·	.05	
Ward15	.03	⋆	.05	
Ward17	.04	⋆⋆	.08	⋆
Ward19	-.03	⋆	-.07	·
Ward30	.01		.01	

No Cluster: χ^2 = 58, df = 38, p-value = .02

Yes Cluster: χ^2 = 39.6, df = 38, p-value =.40

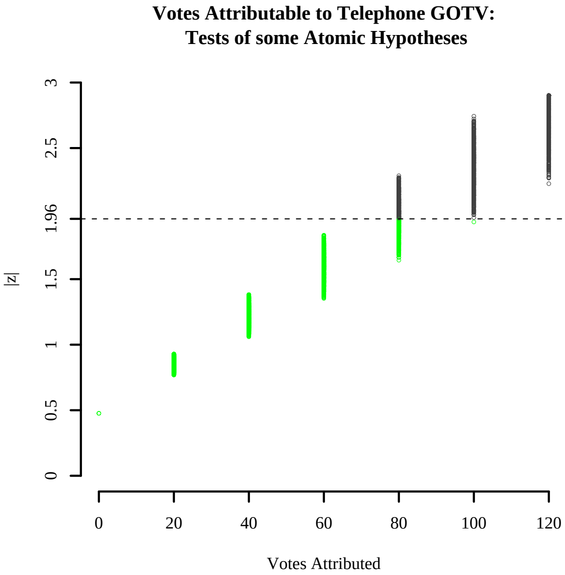
The Importance of Adjusting for Clustering

	Standardized Bias (sds)			
	Accounting for Clustering?			
Covariate	No		Yes	
persons1	.02		.02	
persons2	-.02		-.02	
v96.abst	-.02		-.04	
v96.vote	.01		.01	
majpty	.00		-.05	
age.Bspline1	.00		-.02	
age.Bspline6	-.02		-.02	
Ward2	.00		-.00	
Ward3	-.05	⋆⋆⋆	-.12	⋆⋆
Ward8	.03	·	.05	
Ward12	.03	·	.05	
Ward15	.03	⋆	.05	
Ward17	.04	⋆⋆	.08	⋆
Ward19	-.03	⋆	-.07	·
Ward30	.01		.01	

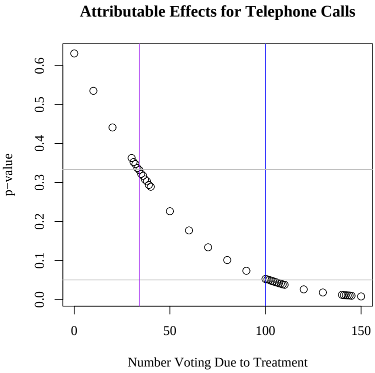
No Cluster: χ^2 = 58, df = 38, p-value = .02

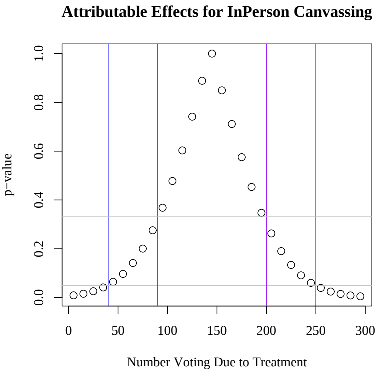
Yes Cluster: χ^2 = 39.6, df = 38, p-value =.40

Attributing Effects to a 2 × 2 × 4 Table



Attributed Effects[5]





[5] Blue vertical lines show the 95% CI; Purple vertical lines show the 66% CI. Gray horizontal lines indicate the relevant p-values.

Summary

- Randomization inference requires very little in the way of assumptions from the data analyst: no large samples, no distributions for Y or θ.
- And with modern computers, it need not be restricted to simple experiments.
- We have shown how to it can be extended to handle cluster-randomized studies.
- It seamlessly produces confidence intervals from per-protocol comparisons.
- It combines naturally with matching and propensity scores to analyze observational studies.
- For offprints, working papers, and software, see `www.stat.lsa.umich.edu/~bbh/`.

Some References

- Bowers, J. and B. Hansen (2005),‘Attributing Effects to A Cluster Randomized Get-Out-The-Vote Campaign: An Application of Randomization Inference Using Full Matching.’ Working paper.
- Hansen, B. (2004), ‘Full matching in an observational study of coaching for the SAT.’ *JASA* **99** (September issue), 609–618.
- Hansen, B. (2005), ‘Optmatch: an add-on package for R.’ `www.stat.lsa.umich.edu/~bbh/optmatch/`.
- Hansen, B. and S. Olsen Klopfer (2005), ‘Optimal full matching and related designs via network flows.’ Tech. Rept. # 416, Statistics Dept., University of Michigan.
- Rosenbaum, P. R. (2002), ‘Attributing effects to treatment in matched observational studies.’ *JASA* **97**, 183–192.