Flexible, Optimal Matching for Comparative Studies Using the `optmatch` package

Ben Hansen

Statistics Department
University of Michigan
`ben.b.hansen@umich.edu`
`http://www.stat.lsa.umich.edu/~bbh`

9 August 2007

Outline

- Matching and its role in statistics
  - Pair matching as an optimization problem
  - Recent history of pair matching in statistics
- Optimal matching of two groups
- A modern approach to “computerized” matching

Illustration: Hollywood matchmaking

Illustration: Hollywood matchmaking

- Lou Diamond Phillips!
- Boy George!
- Meg Ryan!
- Bo Derek!!! and...

Illustration: Hollywood matchmaking

- Lou Diamond Phillips!
- Boy George!
- Meg Ryan!
- Bo Derek!!! and. . .

Illustration: Hollywood matchmaking

- Lou Diamond Phillips!
- Boy George!
- Meg Ryan!
- Bo Derek!!! and…

Illustration: Hollywood matchmaking




- Lou Diamond Phillips!
- Boy George!
- Meg Ryan!
- Bo Derek!!! and. . .

Illustration: Hollywood matchmaking



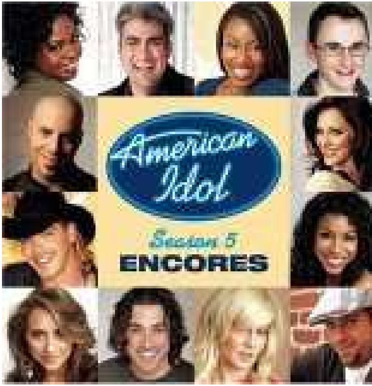


Winona Ryder!

Illustration: Hollywood matchmaking





Winona Ryder!

Matching based on a multivariate dissimilarity

Or multivariate “distance”

												
		–	0	–	1	2	–	1	–	0	–	–
	2	4	2	4	2	3	5	3	3	4	3	
	–		–	2	–	–	4	–	4	–	4	4
	–	3	–	4	–	–	4	–	5	–		2
	–	0	–	5	–	–		–	4	–	0	0

Matching based on a multivariate dissimilarity

Or multivariate “distance”

												
	0	–	0	–	1	2	–	1	–	0	–	–
	2	4	2	4	2	3	5	3	3	4	3	
	–		–	2	–	–	4	–	4	–	4	4
	–	3	–	4	–	–	4	–	5	–		2
	–	0	–	5	–	–		–	4	–	0	0

Matching based on a multivariate dissimilarity

Or multivariate “distance”

												
	0	–	0	–	1	2	–	1	–	0	–	–
	2	4	2	4	2	3	5	3	3	4	3	
	–		–	2	–	–	4	–	4	–	4	4
	–	3	–	4	–	–	4	–	5	–		2
	–	0	–	5	–	–		–	4	–	0	0

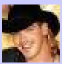
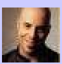
Matching based on a multivariate dissimilarity

Or multivariate "distance"

												
	0	–	0	–	1	2	–	1	–	0	–	–
	2	4	2	4	2	3	5	3	3	4	3	1
	–		–	2	–	–	4	–	4	–	4	4
	–	3	–	4	–	–	4	–	5	–		2
	–	0	–	5	–	–		–	4	–	0	0

Matching based on a multivariate dissimilarity

Or multivariate “distance”

												
	0	–	0	–	1	2	–	1	–	0	–	–
	2	4	2	4	2	3	5	3	3	4	3	1
	–		–	2	–	–	4	–	4	–	4	4
	–	3	–	4	–	–	4	–	5	–		2
	–	0	–	5	–	–		–	4	–	0	0

Matching based on a multivariate dissimilarity

Or multivariate “distance”

												
	0	–	0	–	1	2	–	1	–	0	–	–
	2	4	2	4	2	3	5	3	3	4	3	1
	–	1	–	2	–	–	4	–	4	–	4	4
	–	3	–	4	–	–	4	–	5	–		2
	–	0	–	5	–	–		–	4	–	0	0

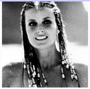
Matching based on a multivariate dissimilarity

Or multivariate “distance”

												
	0	–	0	–	1	2	–	1	–	0	–	–
	2	4	2	4	2	3	5	3	3	4	3	1
	–	1	–	2	–	–	4	–	4	–	4	4
	–	3	–	4	–	–	4	–	5	–		2
	–	0	–	5	–	–		–	4	–	0	0

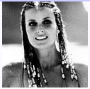
Matching based on a multivariate dissimilarity

Or multivariate “distance”

												
	0	–	0	–	1	2	–	1	–	0	–	–
	2	4	2	4	2	3	5	3	3	4	3	1
	–	1	–	2	–	–	4	–	4	–	4	4
	–	3	–	4	–	–	4	–	5	–	2	2
	–	0	–	5	–	–		–	4	–	0	0

Matching based on a multivariate dissimilarity

Or multivariate “distance”

												
	0	–	0	–	1	2	–	1	–	0	–	–
	2	4	2	4	2	3	5	3	3	4	3	1
	–	1	–	2	–	–	4	–	4	–	4	4
	–	3	–	4	–	–	4	–	5	–	2	2
	–	0	–	5	–	–		–	4	–	0	0

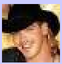
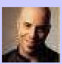
Matching based on a multivariate dissimilarity

Or multivariate “distance”

												
	0	–	0	–	1	2	–	1	–	0	–	–
	2	4	2	4	2	3	5	3	3	4	3	1
	–	1	–	2	–	–	4	–	4	–	4	4
	–	3	–	4	–	–	4	–	5	–	2	2
	–	0	–	5	–	–	4	–	4	–	0	0

Matching based on a multivariate dissimilarity

Or multivariate “distance”

												
	0	–	0	–	1	2	–	1	–	0	–	–
	2	4	2	4	2	3	5	3	3	4	3	1
	–	1	–	2	–	–	4	–	4	–	4	4
	–	3	–	4	–	–	4	–	5	–	2	2
	–	0	–	5	–	–	4	–	4	–	0	0

Outline

Matching and its role in statistics

- Pair matching as an optimization problem
- Recent history of pair matching in statistics

Optimal matching of two groups

A modern approach to "computerized" matching

Matching in Statistics: Cochran's School in the 1970s

- Matched sampling to focus data collection
  - *E.g.*, Althauser and Rubin (1970): prospective comparative study of effects of integration on black college graduates.
  - Problem: some info about many; get more info about some.
  - Many "controls" were not comparable to any black integrated-college graduates.
  - Solution: "computerized" matching procedures
- Multivariate distance matching (Cochran and Rubin, 1973; Rubin, 1976)
- Matched sampling as a way to make model-based analysis robust (Rubin, 1973, 1979)

Matching in Statistics: Cochran's School in the 1970s

- Matched sampling to focus data collection
  - *E.g.,* Althauser and Rubin (1970): prospective comparative study of effects of integration on black college graduates.
  - Problem: some info about many; get more info about some.
  - Many "controls" were not comparable to any black integrated-college graduates.
  - Solution: "computerized" matching procedures
- Multivariate distance matching (Cochran and Rubin, 1973; Rubin, 1976)
- Matched sampling as a way to make model-based analysis robust (Rubin, 1973, 1979)

Matching in Statistics: Cochran's School in the 1970s

- Matched sampling to focus data collection
  - *E.g.,* Althauser and Rubin (1970): prospective comparative study of effects of integration on black college graduates.
  - Problem: some info about many; get more info about some.
  - Many "controls" were not comparable to any black integrated-college graduates.
  - Solution: "computerized" matching procedures
- Multivariate distance matching (Cochran and Rubin, 1973; Rubin, 1976)
- Matched sampling as a way to make model-based analysis robust (Rubin, 1973, 1979)

Matching in Statistics: Cochran's School in the 1970s

- Matched sampling to focus data collection
  - *E.g.*, Althauser and Rubin (1970): prospective comparative study of effects of integration on black college graduates.
  - Problem: some info about many; get more info about some.
  - Many "controls" were not comparable to any black integrated-college graduates.
  - Solution: "computerized" matching procedures
- Multivariate distance matching (Cochran and Rubin, 1973; Rubin, 1976)
- Matched sampling as a way to make model-based analysis robust (Rubin, 1973, 1979)

Matching in Statistics: Cochran's School in the 1970s

- Matched sampling to focus data collection
  - *E.g.,* Althauser and Rubin (1970): prospective comparative study of effects of integration on black college graduates.
  - Problem: some info about many; get more info about some.
  - Many "controls" were not comparable to any black integrated-college graduates.
  - Solution: "computerized" matching procedures
- Multivariate distance matching (Cochran and Rubin, 1973; Rubin, 1976)
- Matched sampling as a way to make model-based analysis robust (Rubin, 1973, 1979)

Matching in Statistics: Cochran's School in the 1970s

- Matched sampling to focus data collection
  - *E.g.*, Althauser and Rubin (1970): prospective comparative study of effects of integration on black college graduates.
  - Problem: some info about many; get more info about some.
  - Many "controls" were not comparable to any black integrated-college graduates.
  - Solution: "computerized" matching procedures
- Multivariate distance matching (Cochran and Rubin, 1973; Rubin, 1976)
- Matched sampling as a way to make model-based analysis robust (Rubin, 1973, 1979)

Matching in Statistics: Cochran's School in the 1970s

- Matched sampling to focus data collection
  - *E.g.,* Althauser and Rubin (1970): prospective comparative study of effects of integration on black college graduates.
  - Problem: some info about many; get more info about some.
  - Many "controls" were not comparable to any black integrated-college graduates.
  - Solution: "computerized" matching procedures
- Multivariate distance matching (Cochran and Rubin, 1973; Rubin, 1976)
- Matched sampling as a way to make model-based analysis robust (Rubin, 1973, 1979)

Matching in Statistics: Cochran's School in the 1980s

- Propensity score
  - Close matches on multivariate **x** not needed if you can match closely on scalar $\phi(\mathbf{x})$ (Rosenbaum and Rubin, 1983, 1984).
  - Good to combine matching on **x** with matching on $\phi(\mathbf{x})$, privileging closeness on $\phi(\mathbf{x})$ (Rosenbaum and Rubin, 1985).
- Computerized matching $\longrightarrow$ optimal matching (Rosenbaum, 1989)

Matching in Statistics: Cochran's School in the 1980s

- Propensity score
  - Close matches on multivariate **x** not needed if you can match closely on scalar $\phi(\mathbf{x})$ (Rosenbaum and Rubin, 1983, 1984).
  - Good to combine matching on **x** with matching on $\phi(\mathbf{x})$, privileging closeness on $\phi(\mathbf{x})$ (Rosenbaum and Rubin, 1985).
- Computerized matching $\rightarrow$ optimal matching (Rosenbaum, 1989)

Matching in Statistics: Cochran's School in the 1990s

- Theoretical & methodological extensions of propensity scores (Rubin and Thomas, 1992, 1996)
- Theoretical & methodological extensions of optimal pair matching (Rosenbaum, 1991; Gu and Rosenbaum, 1993)
- Influential applications (Dehejia and Wahba, 1999; Connors et al., 1996)

Outline

Costs of nuclear plants

A small comparative study from a classic text

Costs of nuclear plants

A small comparative study from a classic text

Existing site		
	date	capacity
A	2.3	660
B	3.0	660
C	3.4	420
D	3.4	130
E	3.9	650
F	5.9	430
G	5.1	420

“date” is date of construction, in years after 1965; “capacity” is net capacity of the power plant, in MWe above 400.

New site		
	date	capacity
H	3.6	290
I	2.3	660
J	3.0	660
K	2.9	110
L	3.2	420
M	3.4	60
N	3.3	390
O	3.6	160
P	3.8	390
Q	3.4	130
R	3.9	650
S	3.9	450
T	3.4	380
U	4.5	440
V	4.2	690
W	3.8	510
X	4.7	390
Y	5.4	140
Z	6.1	730

Existing site		
	date	capacity
A	2.3	660
B	3.0	660
C	3.4	420
D	3.4	130
E	3.9	650
F	5.9	430
G	5.1	420

New site		
	date	capacity
H	3.6	290
I	2.3	660
J	3.0	660
K	2.9	110
L	3.2	420
M	3.4	60
N	3.3	390
O	3.6	160
P	3.8	390
Q	3.4	130
R	3.9	650
S	3.9	450
T	3.4	380
U	4.5	440
V	4.2	690
W	3.8	510
X	4.7	390
Y	5.4	140
Z	6.1	730

Example: 1:2 matching by a traditional, greedy algorithm.

“date” is date of construction, in years after 1965; “capacity” is net capacity of the power plant, in MWe above 400.

Existing site

	date	capacity
A	2.3	660
B	3.0	660
C	3.4	420
D	3.4	130
E	3.9	650
F	5.9	430
G	5.1	420

New site

	date	capacity
H	3.6	290
I	2.3	660
J	3.0	660
K	2.9	110
L	3.2	420
M	3.4	60
N	3.3	390
O	3.6	160
P	3.8	390
Q	3.4	130
R	3.9	650
S	3.9	450
T	3.4	380
U	4.5	440
V	4.2	690
W	3.8	510
X	4.7	390
Y	5.4	140
Z	6.1	730

Example: 1:2 matching by a traditional, greedy algorithm.

“date” is date of construction, in years after 1965; “capacity” is net capacity of the power plant, in MWe above 400.

Existing site

	date	capacity
A	2.3	660
B	3.0	660
C	3.4	420
D	3.4	130
E	3.9	650
F	5.9	430
G	5.1	420

New site

	date	capacity
H	3.6	290
I	2.3	660
J	3.0	660
K	2.9	110
L	3.2	420
M	3.4	60
N	3.3	390
O	3.6	160
P	3.8	390
Q	3.4	130
R	3.9	650
S	3.9	450
T	3.4	380
U	4.5	440
V	4.2	690
W	3.8	510
X	4.7	390
Y	5.4	140
Z	6.1	730

Example: 1:2 matching by a traditional, greedy algorithm.

“date” is date of construction, in years after 1965; “capacity” is net capacity of the power plant, in MWe above 400.

Existing site

	date	capacity
A	2.3	660
B	3.0	660
C	3.4	420
D	3.4	130
E	3.9	650
F	5.9	430
G	5.1	420

New site

	date	capacity
H	3.6	290
I	2.3	660
J	3.0	660
K	2.9	110
L	3.2	420
M	3.4	60
N	3.3	390
O	3.6	160
P	3.8	390
Q	3.4	130
R	3.9	650
S	3.9	450
T	3.4	380
U	4.5	440
V	4.2	690
W	3.8	510
X	4.7	390
Y	5.4	140
Z	6.1	730

Example: 1:2 matching by a traditional, greedy algorithm.

“date” is date of construction, in years after 1965; “capacity” is net capacity of the power plant, in MWe above 400.

Existing site

	date	capacity
A	2.3	660
B	3.0	660
C	3.4	420
D	3.4	130
E	3.9	650
F	5.9	430
G	5.1	420

New site

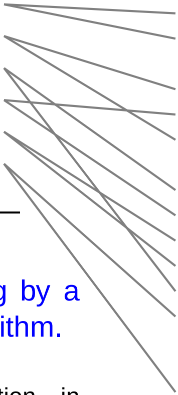
	date	capacity
H	3.6	290
I	2.3	660
J	3.0	660
K	2.9	110
L	3.2	420
M	3.4	60
N	3.3	390
O	3.6	160
P	3.8	390
Q	3.4	130
R	3.9	650
S	3.9	450
T	3.4	380
U	4.5	440
V	4.2	690
W	3.8	510
X	4.7	390
Y	5.4	140
Z	6.1	730

Example: 1:2 matching by a traditional, greedy algorithm.

“date” is date of construction, in years after 1965; “capacity” is net capacity of the power plant, in MWe above 400.

Existing site

	date	capacity
A	2.3	660
B	3.0	660
C	3.4	420
D	3.4	130
E	3.9	650
F	5.9	430
G	5.1	420



New site

	date	capacity
H	3.6	290
I	2.3	660
J	3.0	660
K	2.9	110
L	3.2	420
M	3.4	60
N	3.3	390
O	3.6	160
P	3.8	390
Q	3.4	130
R	3.9	650
S	3.9	450
T	3.4	380
U	4.5	440
V	4.2	690
W	3.8	510
X	4.7	390
Y	5.4	140
Z	6.1	730

Example: 1:2 matching by a traditional, greedy algorithm.

"date" is date of construction, in years after 1965; "capacity" is net capacity of the power plant, in MWe above 400.

Existing site

	date	capacity
A	2.3	660
B	3.0	660
C	3.4	420
D	3.4	130
E	3.9	650
F	5.9	430
G	5.1	420

New site

	date	capacity
H	3.6	290
I	2.3	660
J	3.0	660
K	2.9	110
L	3.2	420
M	3.4	60
N	3.3	390
O	3.6	160
P	3.8	390
Q	3.4	130
R	3.9	650
S	3.9	450
T	3.4	380
U	4.5	440
V	4.2	690
W	3.8	510
X	4.7	390
Y	5.4	140
Z	6.1	730

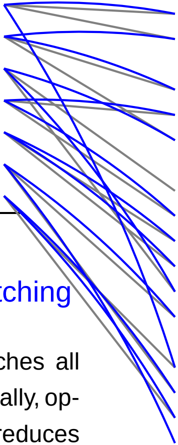
Example: 1:2 matching by a traditional, greedy algorithm.

“date” is date of construction, in years after 1965; “capacity” is net capacity of the power plant, in MWe above 400.

New and refurbished nuclear plants: discrepancies in capacity and year of construction

Existing	New sites																		
	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	28	0	3	22	14	30	17	28	26	28	20	22	23	26	21	18	34	40	28
B	24	3	0	22	10	27	14	26	24	24	16	19	20	23	18	16	31	37	25
C	10	18	14	18	4	12	6	11	9	10	14	12	6	14	22	10	16	22	28
D	7	28	24	8	14	2	10	6	12	0	24	22	4	24	32	20	18	16	38
E	17	20	16	32	18	26	20	18	12	24	0	2	20	6	8	4	14	20	14
F	20	31	28	35	20	29	22	20	14	26	12	9	22	5	15	12	9	11	12
G	14	32	29	30	18	24	17	16	10	22	12	10	17	6	16	14	4	8	17

Existing site		
	date	capacity
A	2.3	660
B	3.0	660
C	3.4	420
D	3.4	130
E	3.9	650
F	5.9	430
G	5.1	420



New site		
	date	capacity
H	3.6	290
I	2.3	660
J	3.0	660
K	2.9	110
L	3.2	420
M	3.4	60
N	3.3	390
O	3.6	160
P	3.8	390
Q	3.4	130
R	3.9	650
S	3.9	450
T	3.4	380
U	4.5	440
V	4.2	690
W	3.8	510
X	4.7	390
Y	5.4	140
Z	6.1	730

Optimal vs. Greedy matching

By evaluating potential matches all together rather than sequentially, optimal matching (blue lines) reduces the sum of distances from 82 to 63.

(Match distance is to "optimal matching" as statistical model is to "maximum likelihood.")

Existing site

	date	capacity
A	2.3	660
B	3.0	660
C	3.4	420
D	3.4	130
E	3.9	650
F	5.9	430
G	5.1	420

New site

	date	capacity
H	3.6	290
I	2.3	660
J	3.0	660
K	2.9	110
L	3.2	420
M	3.4	60
N	3.3	390
O	3.6	160
P	3.8	390
Q	3.4	130
R	3.9	650
S	3.9	450
T	3.4	380
U	4.5	440
V	4.2	690
W	3.8	510
X	4.7	390
Y	5.4	140
Z	6.1	730

Optimal vs. Greedy matching

By evaluating potential matches all together rather than sequentially, optimal matching (blue lines) reduces the sum of distances from 82 to 63. (Match distance is to "optimal matching" as statistical model is to "maximum likelihood.")

Introducing restrictions on who can be matched to whom

With `optmatch`, matches are forbidden by placing ∞'s in the distance matrix. This is a way to exclude unwanted matches, or to reduce the number of controls.

Existing	New sites																		
	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	28	0	3	22	14	30	17	28	26	28	20	22	23	26	21	18	34	Inf	Inf
B	24	3	0	22	10	27	14	26	24	24	16	19	20	23	18	16	31	37	Inf
C	10	18	14	18	4	12	6	11	9	10	14	12	6	14	22	10	16	22	28
D	7	28	24	8	14	2	10	6	12	0	24	22	4	24	32	20	18	16	38
E	17	20	16	32	18	26	20	18	12	24	0	2	20	6	8	4	14	20	14
F	20	Inf	28	Inf	20	29	22	20	14	26	12	9	22	5	15	12	9	11	12
G	14	32	29	30	18	24	17	16	10	22	12	10	17	6	16	14	4	8	17

Outline

Matching and its role in statistics

Optimal matching of two groups

- Comparing nuclear plants: an illustration
- Generalizations of pair matching

A modern approach to “computerized” matching

Example # 2: Gender equity study for research scientists[1]

Women and men scientists are to be matched on grant funding.

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

[1] Discussed in Hansen and Klopfer (2006), Hansen (2004)

Full Matching[2] the Gender Equity Sample

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

- Combines with-replacement & multiple controls matching.
- In general, much better matches than with pair matching.
- Optional restrictions simplify matched sets' structure.

[2](Rosenbaum, 1991; Hansen and Klopfer, 2006)

Full Matching[2] the Gender Equity Sample

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

- Combines with-replacement & multiple controls matching.
- In general, much better matches than with pair matching.
- Optional restrictions simplify matched sets' structure.

[2](Rosenbaum, 1991; Hansen and Klopfer, 2006)

Full Matching[2] the Gender Equity Sample

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

- Combines with-replacement & multiple controls matching.
- In general, much better matches than with pair matching.
- Optional restrictions simplify matched sets' structure.

[2](Rosenbaum, 1991; Hansen and Klopfer, 2006)

Connection to propensity score matching

- Problem: compare a "treatment" group ($Z = 1$) to control ($Z = 0$), adjusting for covariates $X = (X_1, \ldots, X_k)$.
- Propensity score refers to $\phi(X) = \mathbf{E}(Z|X)$
- ... or to $\hat{\phi}(X)$.
- Propensity score≈linear discriminant.

Connection to propensity score matching

- Problem: compare a "treatment" group ($Z = 1$) to control ($Z = 0$), adjusting for covariates $X = (X_1, \ldots, X_k)$.
- Propensity score refers to $\phi(X) = \mathbf{E}(Z|X)$
- …or to $\hat{\phi}(X)$.
- Propensity score≈linear discriminant.

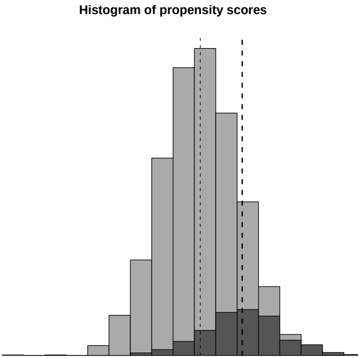
Connection to propensity score matching

- Problem: compare a "treatment" group ($Z = 1$) to control ($Z = 0$), adjusting for covariates $X = (X_1, \ldots, X_k)$.
- Propensity score refers to $\phi(X) = \mathbf{E}(Z|X)$
- ...or to $\hat{\phi}(X)$.
- Propensity score≈linear discriminant.

Connection to propensity score matching

- Problem: compare a "treatment" group ($Z = 1$) to control ($Z = 0$), adjusting for covariates $X = (X_1, \ldots, X_k)$.
- <u>Propensity score</u> refers to $\phi(X) = \mathbf{E}(Z|X)$
- $\ldots$or to $\hat{\phi}(X)$.
- Propensity score≈linear discriminant.

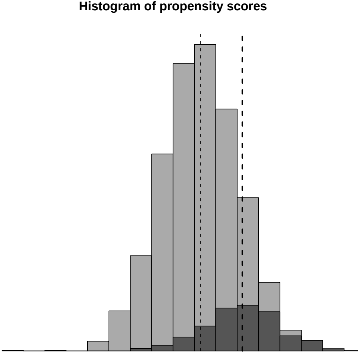
This is typical:



Connection to propensity score matching

- Problem: compare a "treatment" group ($Z = 1$) to control ($Z = 0$), adjusting for covariates $X = (X_1, \ldots, X_k)$.
- Propensity score refers to $\phi(X) = \mathbf{E}(Z|X)$
- ...or to $\hat{\phi}(X)$.
- Propensity score≈linear discriminant.

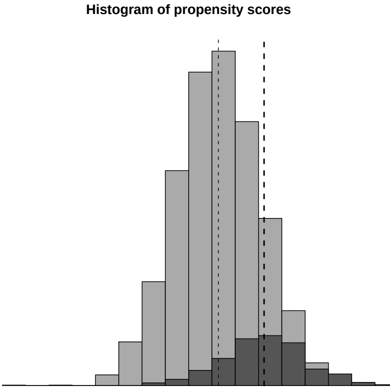
This is typical:



Among matching techniques, only full matching fully adapts...

Controlling the structure of matched sets

- Issue: v. different Tx:Ctl ratios at L and R of histogram.
- This arises because. . . (Hansen, 2004).
- Full matching accommodates this better, but maybe too well.
- Full matching with restrictions compromises between full matching and 1:k matching.

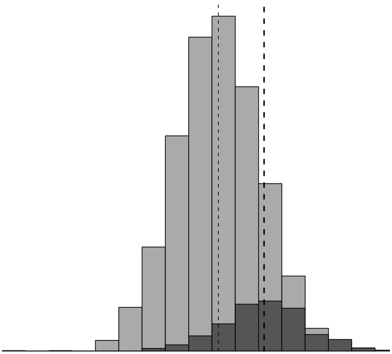


Controlling the structure of matched sets

- Issue: v. different Tx:Ctl ratios at L and R of histogram.
- This arises because. . . (Hansen, 2004).
- Full matching accommodates this better, but maybe too well.
- Full matching with restrictions compromises between full matching and 1:k matching.

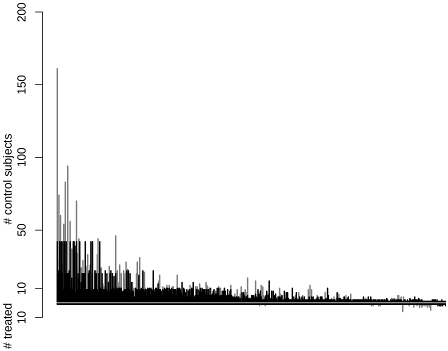
Controlling the structure of matched sets

- Issue: v. different Tx:Ctl ratios at L and R of histogram.
- This arises because. . . (Hansen, 2004).
- Full matching accommodates this better, but maybe too well.
- Full matching with restrictions compromises between full matching and 1:k matching.



Controlling the structure of matched sets

- Issue: v. different Tx:Ctl ratios at L and R of histogram.
- This arises because. . . (Hansen, 2004).
- Full matching accommodates this better, but maybe too well.
- Full matching with restrictions compromises between full matching and 1:k matching.



(Hansen, 2004)

Outline

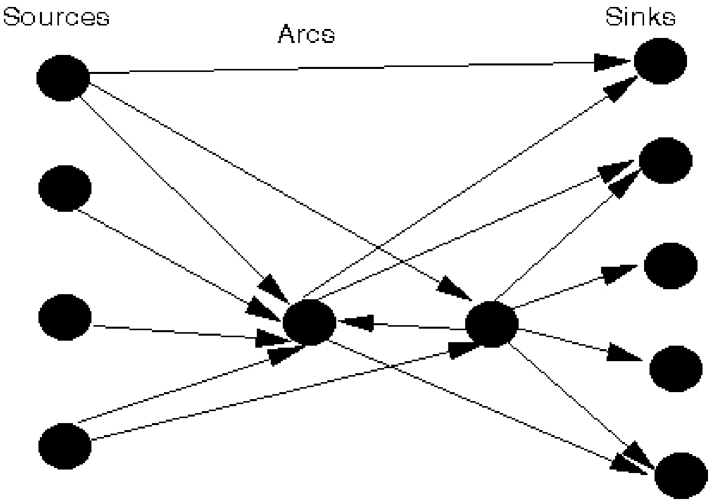
Matching and its role in statistics

Optimal matching of two groups

A modern approach to “computerized” matching

- Optimal bipartite matching via network flows
- Optimal bipartite matching in R

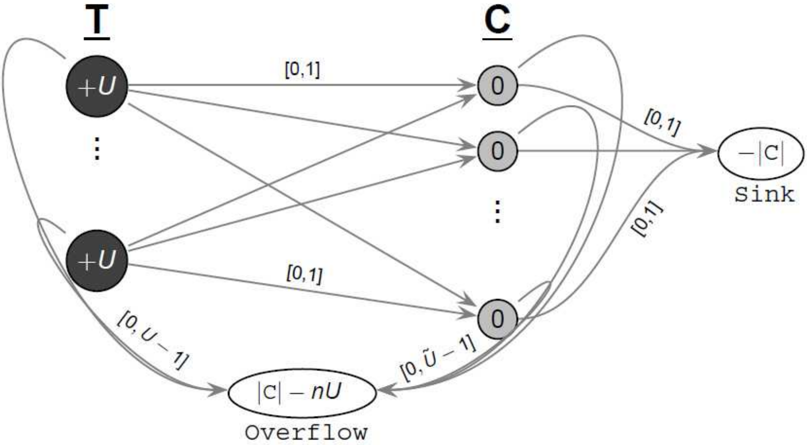
The min-cost flow optimization problem[3]



[3] Illustration from web notes by J. E. Beasley

Under the hood

Full matching via network flows[4]



[4](Hansen and Klopfer, 2006, Fig. 2). Time complexity of the algorithm is $O(n^3 \log(n \max(\text{dist})))$.

Outline

Matching and its role in statistics

Optimal matching of two groups

A modern approach to “computerized” matching

- Optimal bipartite matching via network flows
- Optimal bipartite matching in R

The `optmatch` add-on package: main functions

1. `pairmatch()`. Arguments:
   - distance: The argument demanding most attention from the user, b/c it defines "good" matches.
   - controls: The # *k* of controls, for 1:*k* matching. Defaults to 1.
2. `fullmatch()`. Arguments:
   - distance: (sole mandatory argument)
   - min.controls, max.controls: For controlling the structure of matched sets. *E.g.*, `min.c`=1/2, `max.c`=3 permits 2:1, 1:1, 1:2 and 1:3 matched sets. Default to 0 & ∞, permitting k:1 and 1:k ($\forall k$).
   - omit.fraction: To drop a specified # of controls, as in matched sampling. Defaults to 0, the appropriate value for matched adjustment.

The `optmatch` add-on package: main functions

1. `pairmatch()`. Arguments:
   - distance: The argument demanding most attention from the user, b/c it defines “good” matches.
   - controls: The # k of controls, for 1:k matching. Defaults to 1.
2. `fullmatch()`. Arguments:
   - distance: (sole mandatory argument)
   - min.controls, max.controls: For controlling the structure of matched sets. *E.g.*, `min.c`=1/2, `max.c`=3 permits 2:1, 1:1, 1:2 and 1:3 matched sets. Default to 0 & ∞, permitting k:1 and 1:k ($\forall k$).
   - omit.fraction: To drop a specified # of controls, as in matched sampling. Defaults to 0, the appropriate value for matched adjustment.

The `optmatch` add-on package: main functions

1. `pairmatch()`. Arguments:

   distance The argument demanding most attention from the user, b/c it defines "good" matches.

   controls The # k of controls, for 1:k matching. Defaults to 1.

2. `fullmatch()`. Arguments:

   distance (sole mandatory argument)

   min.controls, max.controls For controlling the structure of matched sets. *E.g.*, `min.c`=1/2, `max.c`=3 permits 2:1, 1:1, 1:2 and 1:3 matched sets. Default to 0 & ∞, permitting k:1 and 1:k ($\forall k$).

   omit.fraction To drop a specified # of controls, as in matched sampling. Defaults to 0, the appropriate value for matched adjustment.

The `optmatch` add-on package: main functions

1. `pairmatch()`. Arguments:

   distance The argument demanding most attention from the user, b/c it defines "good" matches.

   controls The # k of controls, for 1:k matching. Defaults to 1.

2. `fullmatch()`. Arguments:

   distance (sole mandatory argument)

   min.controls, max.controls For controlling the structure of matched sets. *E.g.*, `min.c`=1/2, `max.c`=3 permits 2:1, 1:1, 1:2 and 1:3 matched sets. Default to 0 & ∞, permitting k:1 and 1:k ($\forall k$).

   omit.fraction To drop a specified # of controls, as in matched sampling. Defaults to 0, the appropriate value for matched adjustment.

The `optmatch` add-on package: main functions

1. `pairmatch()`. Arguments:

   distance
   : The argument demanding most attention from the user, b/c it defines "good" matches.

   controls
   : The # k of controls, for 1:k matching. Defaults to 1.

2. `fullmatch()`. Arguments:

   distance
   : (sole mandatory argument)

   min.controls, max.controls
   : For controlling the structure of matched sets. *E.g.*, `min.c`=1/2, `max.c`=3 permits 2:1, 1:1, 1:2 and 1:3 matched sets. Default to 0 & ∞, permitting k:1 and 1:k ($\forall k$).

   omit.fraction
   : To drop a specified # of controls, as in matched sampling. Defaults to 0, the appropriate value for matched adjustment.

The `optmatch` add-on package: main functions

1. `pairmatch()`. Arguments:
   - distance: The argument demanding most attention from the user, b/c it defines "good" matches.
   - controls: The # k of controls, for 1:k matching. Defaults to 1.
2. `fullmatch()`. Arguments:
   - distance: (sole mandatory argument)
   - min.controls, max.controls: For controlling the structure of matched sets. *E.g.*, `min.c`=1/2, `max.c`=3 permits 2:1, 1:1, 1:2 and 1:3 matched sets. Default to 0 & ∞, permitting k:1 and 1:k ($\forall k$).
   - omit.fraction: To drop a specified # of controls, as in matched sampling. Defaults to 0, the appropriate value for matched adjustment.

The `optmatch` add-on package: main functions

1. `pairmatch()`. Arguments:

   distance The argument demanding most attention from the user, b/c it defines "good" matches.

   controls The # k of controls, for 1:k matching. Defaults to 1.

2. `fullmatch()`. Arguments:

   distance (sole mandatory argument)

   min.controls, max.controls For controlling the structure of matched sets. *E.g.*, `min.c`=1/2, `max.c`=3 permits 2:1, 1:1, 1:2 and 1:3 matched sets. Default to 0 & ∞, permitting k:1 and 1:k ($\forall k$).

   omit.fraction To drop a specified # of controls, as in matched sampling. Defaults to 0, the appropriate value for matched adjustment.

The `optmatch` add-on package: helper functions

1. `pscore.dist()`. Example:

   ```
   > pmodel <- glm(pr~.-(pr+cost), family=binomial,
   + data=nuclear)
   > pdist <- pscore.dist(pmodel)
   ```

2. `mahal.dist()`. Facilitates construction of Mahalanobis distances for matching. Example:

   ```
   > mdist <- mahal.dist(pr~date+cum.n, nuclear)
   ```

3. `makedist()`. Facilitates construction of arbitrary distances for matching. See help page for examples.

The `optmatch` add-on package: addressing likely problems

- Sequence is data frame ↦ distance matrix ↦ factor object encoding the match. Easy to scramble ordering of observations.
  My Solution: helper functions `pscore.dist`, `mahal.dist` and `makedist` carry metadata that `fullmatch` and `pairmatch` use to prevent this problem.
- Matching is slow for large problems. ($O(n^3 \log(n))$ flops.)
  My Solution: Match within subclasses. Example:

```
> mdist <- mahal.dist(pr~date+cum.n, nuclear,
pr~pt)
> fullmatch(mdist)
```

  This matches within levels of `pt`.

The `optmatch` add-on package: addressing likely problems

- Sequence is data frame $\mapsto$ distance matrix $\mapsto$ factor object encoding the match. Easy to scramble ordering of observations.
  My Solution: helper functions `pscore.dist`, `mahal.dist` and `makedist` carry metadata that `fullmatch` and `pairmatch` use to prevent this problem.
- Matching is slow for large problems. ($O(n^3 \log(n))$ flops.)
  My Solution: Match within subclasses. Example:

```
> mdist <- mahal.dist(pr~date+cum.n, nuclear,
pr~pt)
> fullmatch(mdist)
```

  This matches within levels of `pt`.

The `optmatch` add-on package: addressing likely problems

- Sequence is data frame $\mapsto$ distance matrix $\mapsto$ factor object encoding the match. Easy to scramble ordering of observations.
  My Solution: helper functions `pscore.dist`, `mahal.dist` and `makedist` carry metadata that `fullmatch` and `pairmatch` use to prevent this problem.
- Matching is slow for large problems. ($O(n^3 \log(n))$ flops.)
  My Solution: Match within subclasses. Example:

```
> mdist <- mahal.dist(pr~date+cum.n, nuclear,
pr~pt)
> fullmatch(mdist)
```

  This matches within levels of `pt`.

The `optmatch` add-on package: addressing likely problems

- Sequence is data frame $\mapsto$ distance matrix $\mapsto$ factor object encoding the match. Easy to scramble ordering of observations.
  My Solution: helper functions `pscore.dist`, `mahal.dist` and `makedist` carry metadata that `fullmatch` and `pairmatch` use to prevent this problem.
- Matching is slow for large problems. ($O(n^3 \log(n))$ flops.)
  My Solution: Match within subclasses. Example:

```
> mdist <- mahal.dist(pr~date+cum.n, nuclear,
pr~pt)
> fullmatch(mdist)
```

  This matches within levels of `pt`.

The `optmatch` add-on package: addressing likely problems

- Distances of mixed type, *e.g.* Mahalanobis matching within propensity calipers (Rubin and Thomas, 2000), lead to messy code, particularly with large problems requiring matching within subclasses. **My Solution:** Defined arithmetic operations for matching distance objects. To Mahalanobis-match within levels of `pt` and with a propensity caliper of .2 pooled SDs,

```
> mdist <- mahal.dist(pr~date+cum.n, nuclear,
pr~pt)
> pmodel <- glm(pr~.-(pr+cost), family=binomial,
data=nuclear)
> pdist <- pscore.dist(pmodel, pr~pt)
> fullmatch(mdist/(pdist<.2))
```

The `optmatch` add-on package: addressing likely problems

- Distances of mixed type, *e.g.* Mahalanobis matching within propensity calipers (Rubin and Thomas, 2000), lead to messy code, particularly with large problems requiring matching within subclasses. **My Solution:** Defined arithmetic operations for matching distance objects. To Mahalanobis-match within levels of `pt` and with a propensity caliper of .2 pooled SDs,

```
> mdist <- mahal.dist(pr~date+cum.n, nuclear,
pr~pt)
> pmodel <- glm(pr~.-(pr+cost), family=binomial,
data=nuclear)
> pdist <- pscore.dist(pmodel, pr~pt)
> fullmatch(mdist/(pdist<.2))
```

Summary

- Matching has uses in design & analysis of observational studies.
- `optmatch` solves optimally such traditional problems as matched sampling, pair matching, and matching with *k* controls.
- `optmatch` can also solve matching problems more flexibly by way of full matching, with or without structural restrictions.
- Full matching combines particularly well w/ propensity scores.
- The effort required to articulate & code relevant algorithms seems to have dissuaded their widespread use. Now that we've made that effort, perhaps this situation can change! :)

Summary

- Matching has uses in design & analysis of observational studies.
- `optmatch` solves optimally such traditional problems as matched sampling, pair matching, and matching with k controls.
- `optmatch` can also solve matching problems more flexibly by way of full matching, with or without structural restrictions.
- Full matching combines particularly well w/ propensity scores.
- The effort required to articulate & code relevant algorithms seems to have dissuaded their widespread use. Now that we've made that effort, perhaps this situation can change! :)

Summary

- Matching has uses in design & analysis of observational studies.
- `optmatch` solves optimally such traditional problems as matched sampling, pair matching, and matching with k controls.
- `optmatch` can also solve matching problems more flexibly by way of full matching, with or without structural restrictions.
- Full matching combines particularly well w/ propensity scores.
- The effort required to articulate & code relevant algorithms seems to have dissuaded their widespread use. Now that we've made that effort, perhaps this situation can change! :)

Summary

- Matching has uses in design & analysis of observational studies.
- `optmatch` solves optimally such traditional problems as matched sampling, pair matching, and matching with k controls.
- `optmatch` can also solve matching problems more flexibly by way of full matching, with or without structural restrictions.
- Full matching combines particularly well w/ propensity scores.
- The effort required to articulate & code relevant algorithms seems to have dissuaded their widespread use. Now that we've made that effort, perhaps this situation can change! :)

Summary

- Matching has uses in design & analysis of observational studies.
- `optmatch` solves optimally such traditional problems as matched sampling, pair matching, and matching with *k* controls.
- `optmatch` can also solve matching problems more flexibly by way of full matching, with or without structural restrictions.
- Full matching combines particularly well w/ propensity scores.
- The effort required to articulate & code relevant algorithms seems to have dissuaded their widespread use. Now that we've made that effort, perhaps this situation can change! :)

Agresti, A. (2002), *Categorical data analysis*, John Wiley & Sons.

Althauser, R. and Rubin, D. (1970), "The Computerized Construction of a Matched Sample," *American Journal of Sociology*, 76, 325–346.

Cochran, W. G. and Rubin, D. B. (1973), "Controlling Bias in Observational Studies: A Review," *Sankhyā, Series A, Indian Journal of Statistics*, 35, 417–446.

Connors, A. J., Speroff, T., Dawson, N., Thomas, C., Harrell, F. J., Wagner, D., Desbiens, N., Goldman, L., Wu, A., Califf, R., Fulkerson, W. J., Vidaillet, H., Broste, S., Bellamy, P., Lynn, J., and Knaus, W. (1996), "The Effectiveness of Right Hearth Catheterization in the Initial Care of Critically Ill Patients. SUPPORT Investigators." *Journal of the American Medical Association*, 276, 889–97.

Cox, D. R. and Snell, E. J. (1989), *Analysis of Binary Data*, Chapman & Hall Ltd.

Dehejia, R. and Wahba, S. (1999), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94, 1053–1062.

Gu, X. and Rosenbaum, P. R. (1993), "Comparison of Multivariate Matching Methods: Structures, Distances, and Algorithms," *Journal of Computational and Graphical Statistics*, 2, 405–420.

Hansen, B. B. (2004), "Full matching in an observational study of coaching for the SAT," *Journal of the American Statistical Association*, 99, 609–618.

Hansen, B. B. and Klopfer, S. O. (2006), "Optimal full matching and related designs via network flows," *Journal of Computational and Graphical Statistics*, 15, 609–627.

Raudenbush, S. W. and Bryk, A. S. (2002), *Hierarchical Linear Models: Applications and Data Analysis Methods*, Sage Publications Inc.

Rosenbaum, P. R. (1989), “Optimal Matching for Observational Studies,” *Journal of the American Statistical Association*, 84, 1024–1032.

— (1991), “A Characterization of Optimal Designs for Observational Studies,” *Journal of the Royal Statistical Society*, 53, 597– 610.

— (2002a), “Attributing effects to treatment in matched observational studies,” *Journal of the American Statistical Association*, 97, 183–192.

— (2002b), “Covariance adjustment in randomized experiments and observational studies,” *Statistical Science*, 17, 286–327.

— (2002c), *Observational Studies*, Springer-Verlag, 2nd ed.

Rosenbaum, P. R. and Rubin, D. B. (1983), “The Central Role of the Propensity Score in Observational Studies for Causal Effects,” *Biometrika*, 70, 41–55.

— (1984), “Reducing Bias in Observational Studies using Subclassification on the Propensity Score,” *Journal of the American Statistical Association*, 79, 516–524.

— (1985), “Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score,” *The American Statistician*, 39, 33–38.

Rubin, D. B. (1973), “The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies,” *Biometrics*, 29, 185–203.

— (1976), “Multivariate Matching Methods That Are Equal Percent Bias Reducing. I: Some Examples (Corr: V32 P955),” *Biometrics*, 32, 109–120.

— (1979), “Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies,” *Journal of the American Statistical Association*, 74, 318–328.

Rubin, D. B. and Thomas, N. (1992), “Characterizing the Effect of Matching Using Linear Propensity Score Methods With Normal Distributions,” *Biometrika*, 79, 797–809.

— (1996), “Matching Using Estimated Propensity Scores: Relating Theory to Practice,” *Biometrics*, 52, 249–64.

— (2000), “Combining Propensity Score Matching with Additional Adjustments for Prognostic Covariates,” *Journal of the American Statistical Association*, 95, 573–585.

Smith, H. (1997), “Matching with Multiple Controls to Estimate Treatment Effects in Observational Studies,” *Sociological Methodology*, 27, 325–353.

Example with propensity scores and stratification prior to matching

```
>nuclear$pscore <- glm(pr~.-cost,
+ family=binomial,data=nuclear)$linear.predictors

> pscorediffs <- function(trtvar,data) {
+ pscr <- data[names(trtvar), 'pscore']
+ abs(outer(pscr[trtvar],pscr[!trtvar], '-'))
+ }

> psd2 <- makedist(pr~pt, nuclear, pscorediffs)

> fullmatch(psd2)

> fullmatch(psd2, min.controls=1, max.controls=3)
> fullmatch(psd2, min=1, max=c('0'=3, '1'=2))
```

RItools package provides diagnostics. . .

Example with propensity scores and stratification prior to matching

```
>nuclear$pscore <- glm(pr~.-cost,
+ family=binomial,data=nuclear)$linear.predictors

> pscorediffs <- function(trtvar,data) {
+ pscr <- data[names(trtvar), 'pscore']
+ abs(outer(pscr[trtvar],pscr[!trtvar], '-'))
+ }

> psd2 <- makedist(pr~pt, nuclear, pscorediffs)

> fullmatch(psd2)

> fullmatch(psd2, min.controls=1, max.controls=3)
> fullmatch(psd2, min=1, max=c('0'=3, '1'=2))
```

RItools package provides diagnostics. . .

Example with propensity scores and stratification prior to matching

```
>nuclear$pscore <- glm(pr~.-cost,
+ family=binomial,data=nuclear)$linear.predictors

> pscorediffs <- function(trtvar,data) {
+ pscr <- data[names(trtvar), 'pscore']
+ abs(outer(pscr[trtvar],pscr[!trtvar], '-'))
+ }

> psd2 <- makedist(pr~pt, nuclear, pscorediffs)

> fullmatch(psd2)

> fullmatch(psd2, min.controls=1, max.controls=3)
> fullmatch(psd2, min=1, max=c('0'=3, '1'=2))
```

RItools package provides diagnostics. . .

Example with propensity scores and stratification prior to matching

```
>nuclear$pscore <- glm(pr~.-cost,
+ family=binomial,data=nuclear)$linear.predictors

> pscorediffs <- function(trtvar,data) {
+ pscr <- data[names(trtvar), 'pscore']
+ abs(outer(pscr[trtvar],pscr[!trtvar], '-'))
+ }

> psd2 <- makedist(pr~pt, nuclear, pscorediffs)

> fullmatch(psd2)

> fullmatch(psd2, min.controls=1, max.controls=3)
> fullmatch(psd2, min=1, max=c('0'=3, '1'=2))
```

`RItools` package provides diagnostics...

Modes of estimation for treatment effects

Preferred mode of inference	Type of outcome: Categorical	Continuous
Randomization	Agresti (2002), Categorical Data Analysis; Rosenbaum (2002a), "Atributing effects to treatment . . . "	Rosenbaum (2002c), Observational Studies; Rosenbaum (2002b), "Covariance adjustment . . . ."
Conditional [a]	Agresti (2002); Cox and Snell (1989), Analysis of binary data	ordinary OLS[b] is fine; see also Rubin (1979), "Using multivariate matched. . . ."
Bayes/Empirical Bayes, esp. hierarchical linear models [c]	Agresti (2002)	Smith (1997), "Matching with multiple controls. . . "; Raudenbush and Bryk (2002), Hierarchical linear models

[a] Uses a fixed effect for each matched set.

[b] i.e., OLS with a fixed effect for each matched set plus treatment effect(s)

[c] Uses a random effect for each matched set.