Why Propensity Scores Should Be Used for Matching

Ben Jann

University of Bern, ben.jann@soz.unibe.ch

2017 German Stata Users Group Meeting
Berlin, June 23, 2017

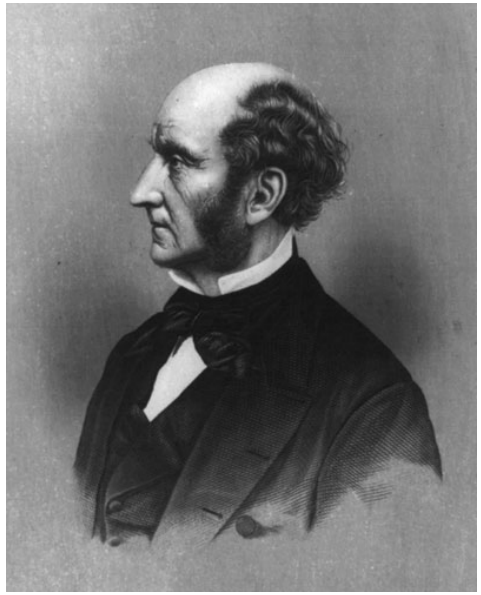
Contents

Counterfactual Causality (see Neyman 1923, Rubin 1974, 1990)

a.k.a. Rubin Causal Model a.k.a. Potential Outcomes Framework

- John Stuart Mill (1806–1873)

  Thus, if a person eats of a particular dish, and dies in consequence, that is, would not have died if he had not eaten of it, people would be apt to say that eating of that dish was the cause of his death. (Mill 2002[1843]:214)



Counterfactual Causality (see Neyman 1923, Rubin 1974, 1990)

a.k.a. Rubin Causal Model a.k.a. Potential Outcomes Framework

- Treatment variable D

$$D = \begin{cases} 1 & \text{treatment (eats of a particular dish)} \\ 0 & \text{control (does not eat of a particular dish)} \end{cases}$$

- Potential outcomes Y^1 and Y^0
  - Y^1: potential outcome with treatment ($D = 1$)
    - ★ If person i would eat of a particular dish, would she die or would she survive?
  - Y^0: potential outcome without treatment ($D = 0$)
    - ★ If person i would *not* eat of a particular dish, would she die or would she survive?
- Causal effect of the treatment for individual i:

causal effect = difference between potential outcomes

$$\delta_i = Y_i^1 - Y_i^0$$

Fundamental Problem of Causal Inference

- The causal effect of D on Y for individual i is defined as the difference in potential outcomes: $\delta_i = Y_i^1 - Y_i^0$
- However, the observed outcome variable is

$$Y_i = \begin{cases} Y_i^1 & \text{if } D_i = 1 \\ Y_i^0 & \text{if } D_i = 0 \end{cases}$$

- That is, only one of the two potential outcomes will be realized and, hence, only Y_i^1 or Y_i^0 can be observed, but never both.
- Consequence:

The individual treatment effect δ_i cannot be observed!

Average Treatment Effect

- Although individual causal effects cannot be observed, the *average* causal effect in a population (the so-called "Average Treatment Effect") can be identified comparing the expected values of Y^1 and Y^0:

$$ATE = E[\delta] = E[Y^1 - Y^0] = E[Y^1] - E[Y^0]$$

- Some other quantities of interest:
  - Average Treatment Effect on the Treated (ATT)

$$ATT = E[Y^1 - Y^0|D = 1] = E[Y^1|D = 1] - E[Y^0|D = 1]$$

  - Average Treatment Effect on the Untreated (ATC)

$$ATC = E[Y^1 - Y^0|D = 0] = E[Y^1|D = 0] - E[Y^0|D = 0]$$

Average Treatment Effect

- To determine the average effect, unbiased estimates of $E[Y^0]$ and $E[Y^1]$ are required.
- If the independence assumption

$$(Y^0, Y^1) \perp\!\!\!\perp D$$

applies, that is, if D is independent from Y^0 and Y^1, then

$$E[Y^0] = E[Y^0|D = 0]$$
$$E[Y^1] = E[Y^1|D = 1]$$

- In this case the average causal effect can be be measured by a simple group comparison (mean difference) of observations without treatment ($D = 0$) and observations with treatment ($D = 1$).
- Randomized experiments solve the problem: If the assignment of D is randomized, D is independent from Y^0 and Y^1 by design.

Conditional Independence / Strong Ignorability

- Can causal effects also be identified from "observational" (i.e. non-experimental) data?
- Sometimes it can be argued that the independence assumption is valid *conditionally* (conditional independence, "unconfoundedness"):

$$(Y^0, Y^1) \perp\!\!\!\perp D \,|\, X$$

- If, in addition, the overlap assumption

$$0 < \Pr(D = 1|X = x) < 1, \quad \text{for all } x$$

  is given, then the ATE (or ATT or ATC) can be identified by conditioning on X.
- For example:

$$ATE = \sum_x \Pr[X = x] \{E[Y|D = 1, X = x] - E[Y|D = 0, X = x]\}$$

Matching

- Matching is one approach to "condition on X" if strong ignorability holds.
- Basic idea:
  1. For each observation in the treatment group, find "statistical twins" in the control group with the same (or at least very similar) X values (and vice versa).
  2. The Y values of these matching observations are then used to compute the counterfactual outcome for the observation at hand.
  3. An estimate for the average causal effect can be obtained as the mean of the differences between the observed values and the "imputed" counterfactual values over all observations.

Matching

- Formally:

$$\widehat{ATT} = \frac{1}{N^{D=1}} \sum_{i|D=1} \left[Y_i - \hat{Y}_i^0 \right] = \frac{1}{N^{D=1}} \sum_{i|D=1} \left[Y_i - \sum_{j|D=0} w_{ij} Y_j \right]$$

$$\widehat{ATC} = \frac{1}{N^{D=0}} \sum_{i|D=0} \left[\hat{Y}_i^1 - Y_i \right] = \frac{1}{N^{D=0}} \sum_{i|D=0} \left[\sum_{j|D=1} w_{ij} Y_j - Y_i \right]$$

$$\widehat{ATE} = \frac{N^{D=1}}{N} \cdot \widehat{ATT} + \frac{N^{D=0}}{N} \cdot \widehat{ATC}$$

- Different matching algorithms use different definitions of w_{ij}.

Exact Matching

- Exact matching:

$$w_{ij} = \begin{cases} 1/k_i & \text{if } X_i = X_j \\ 0 & \text{else} \end{cases}$$

  with k_i as the number of observations for which $X_i = X_j$ applies.

- The result equivalent to "perfect stratification" or "subclassification" (see, e.g., Cochran 1968).

- Problem: If X contains several variables there is a large probability that no exact matches can be found for many observations (the "curse of dimensionality").

Multivariate Distance Matching (MDM)

- An alternative is to match based on a distance metric that measures the proximity between observations in the multivariate space of X.
- The idea then is to use observations that are "close", but not necessarily equal, as matches.
- A common approach is to use

$$MD(X_i, X_j) = \sqrt{(X_i - X_j)' \Sigma^{-1} (X_i - X_j)}$$

  as distance metric, where Σ is an appropriate scaling matrix.
  - Mahalanobis matching: Σ is the covariance matrix of X.
  - Euclidean matching: Σ is the identity matrix.
  - Mahalanobis matching is equivalent to Euclidean matching based on standardized and orthogonalized X.

Matching Algorithms

- Various matching algorithms can be employed to find potential matches based on MD, and determine the matching weights w_{ij}.
- Pair matching (one-to-one matching without replacement)
  - For each observation i in the treatment group find observation j in the control group for which MD_{ij} is smallest. Once observation j is used as a match, do not use it again.
- Nearest-neighbor matching
  - For each observation i in the treatment group find the k closest observations in the control group. A single control can be used multiple times as a match. In case of ties (multiple controls with identical MD), use all ties as matches. k is set by the researcher.
- Caliper matching
  - Like nearest-neighbor matching, but only use controls for which MD is smaller than some threshold c.

Mahalanobis Matching

- Radius matching
  - Use *all* controls as matches for which MD is smaller than some threshold c.
- Kernel matching
  - Like radius matching, but give larger weight to controls for which MD is small (using some kernel function such as, e.g., the Epanechnikov kernel).
- In addition, since matching is no longer exact, it may make sense to refine the estimates by applying regression-adjustment to the matched data (also known as "bias-adjustment" in the context of nearest-neighbor matching).

The Propensity Score Theorem (Rosenbaum and Rubin 1983)

- If the conditional independence assumption is true, then

$$\Pr(D_i = 1 | Y_i^0, Y_i^1, X_i) = \Pr(D_i = 1 | X_i) = \pi(X_i)$$

  where $\pi(X)$ is called the propensity score.

- That is,

$$(Y^0, Y^1) \perp\!\!\!\perp D \,|\, X$$

  implies

$$(Y^0, Y^1) \perp\!\!\!\perp D \,|\, \pi(X)$$

  so that under strong ignorability the average causal effect can be estimated by conditioning on the propensity score $\pi(X)$ instead of X.

- This is remarkable, because the information in X, which may include many variables, can be reduced to just one dimension. This greatly simplifies the matching task.

Propensity Score Matching (PSM)

- Instead of computing multivariate distances, we can thus simply match on the (one-dimensional) propensity score.
- Procedure
  - Step 1: Estimate the propensity score, e.g. using a Logit model.
  - Step 2: Apply a matching algorithm using differences in the propensity score, $|\hat{\pi}(X_i) - \hat{\pi}(X_j)|$, instead of multivariate distances.
- PSM is tremendously popular
  - https://scholar.google.ch/scholar?q="propensity+score"+AND+(matching+OR+matched+OR+match)

King and Nielsen

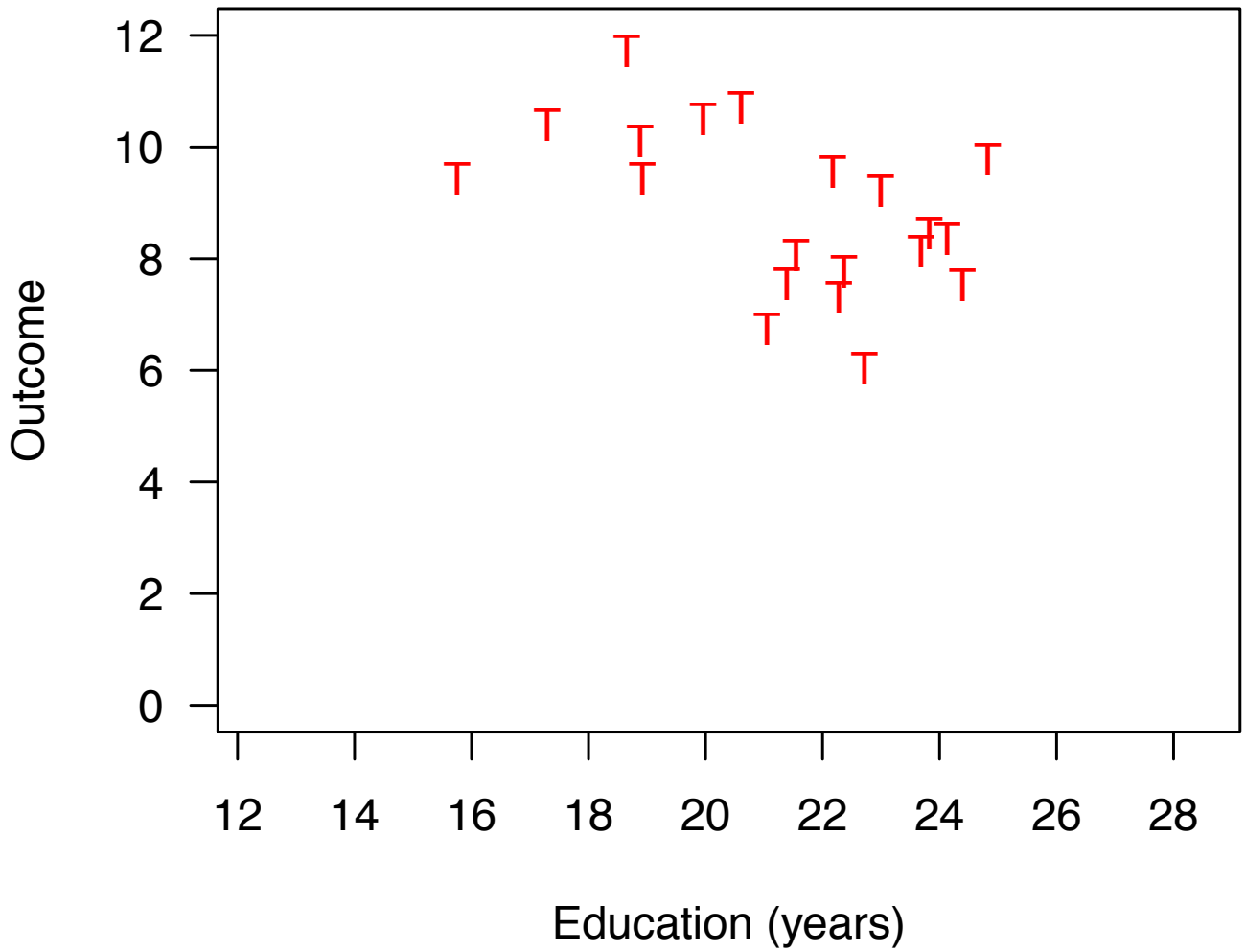
- In 2015/2016 Gary King and Richard Nielsen circulated a paper that created quite some concern among applied researchers.
- The basic message of the paper is that PSM is really, really bad and should be discarded.
- The paper
  - http://j.mp/1sexgVw
- Slides
  - https://gking.harvard.edu/presentations/why-propensity-scores-should-not-be-used-matching-6
- Watch it
  - https://www.youtube.com/watch?v=rBv39pK1iEs

King and Nielsen

- The story goes about as follows.
- Argument 1
  - Model dependence (i.e. dependence of results on modeling decisions made by the researcher) is bad because it leads to bias (people are selective in their decisions even if they try not to be).
  - Matching is good because it reduces model dependence.

Matching to Reduce Model Dependence

(Ho, Imai, King, Stuart, 2007: fig.1, *Political Analysis*)

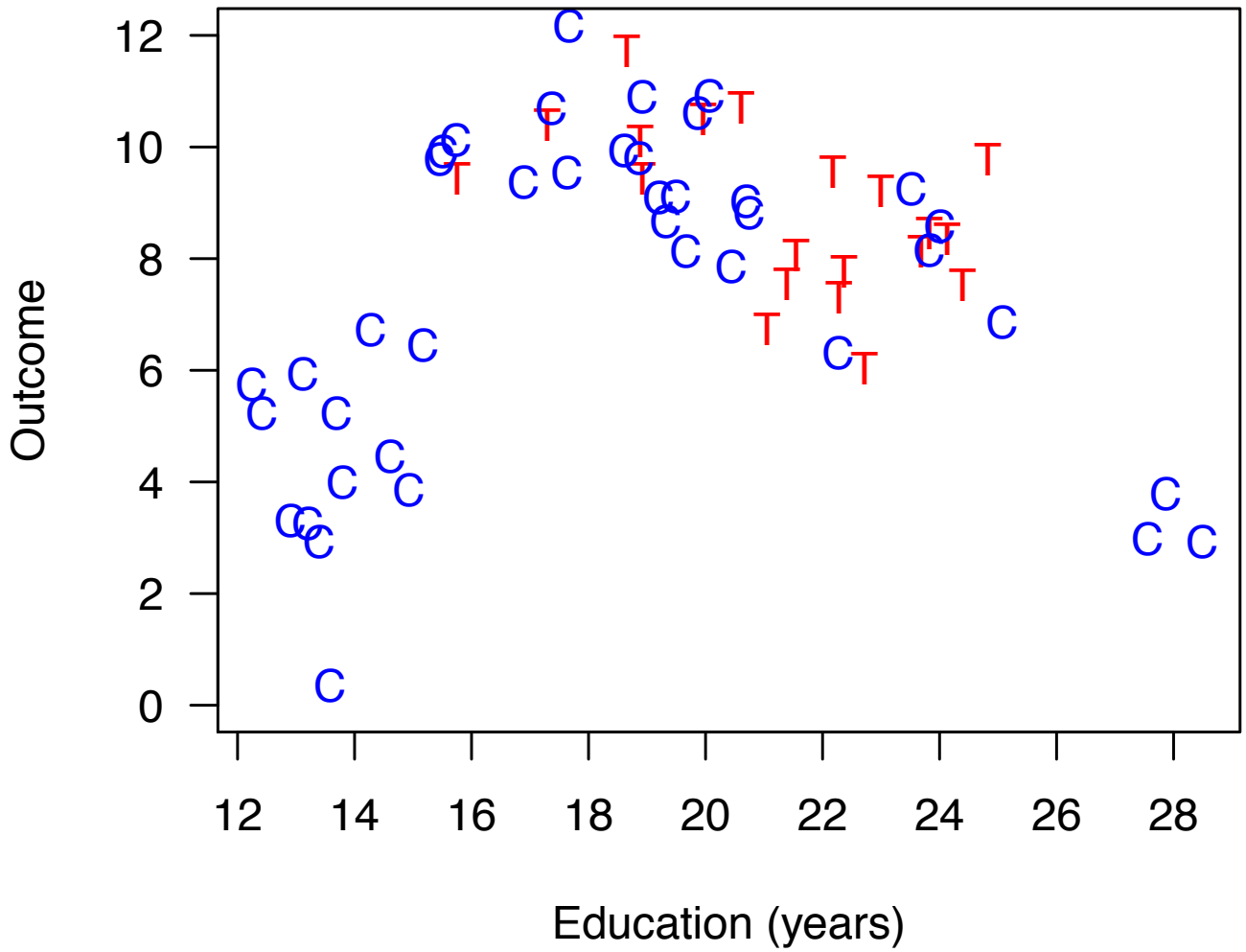


3/23

(slides by King and Nielsen)

Matching to Reduce Model Dependence

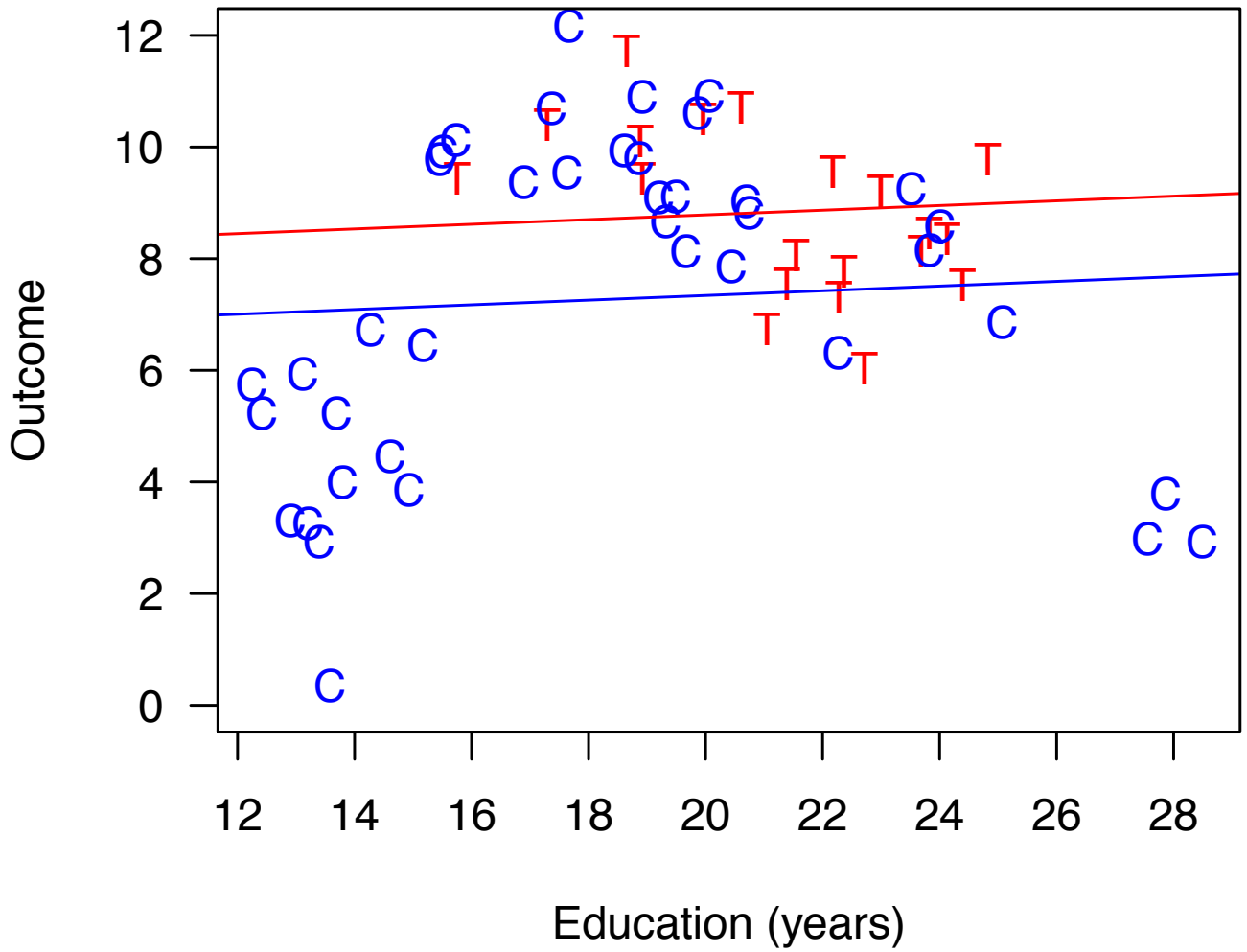
(Ho, Imai, King, Stuart, 2007: fig.1, *Political Analysis*)



(slides by King and Nielsen)

Matching to Reduce Model Dependence

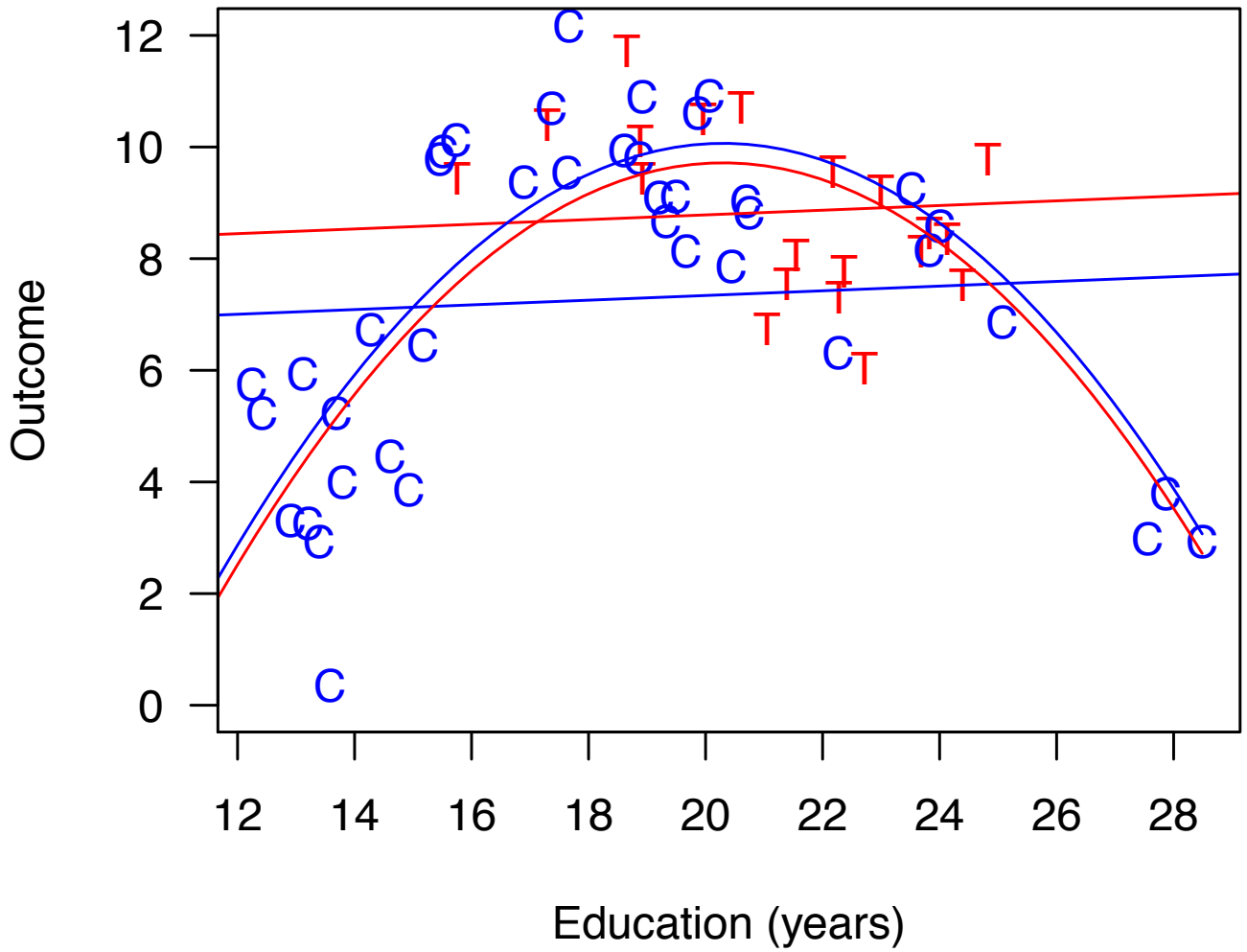
(Ho, Imai, King, Stuart, 2007: fig.1, *Political Analysis*)



(slides by King and Nielsen)

Matching to Reduce Model Dependence

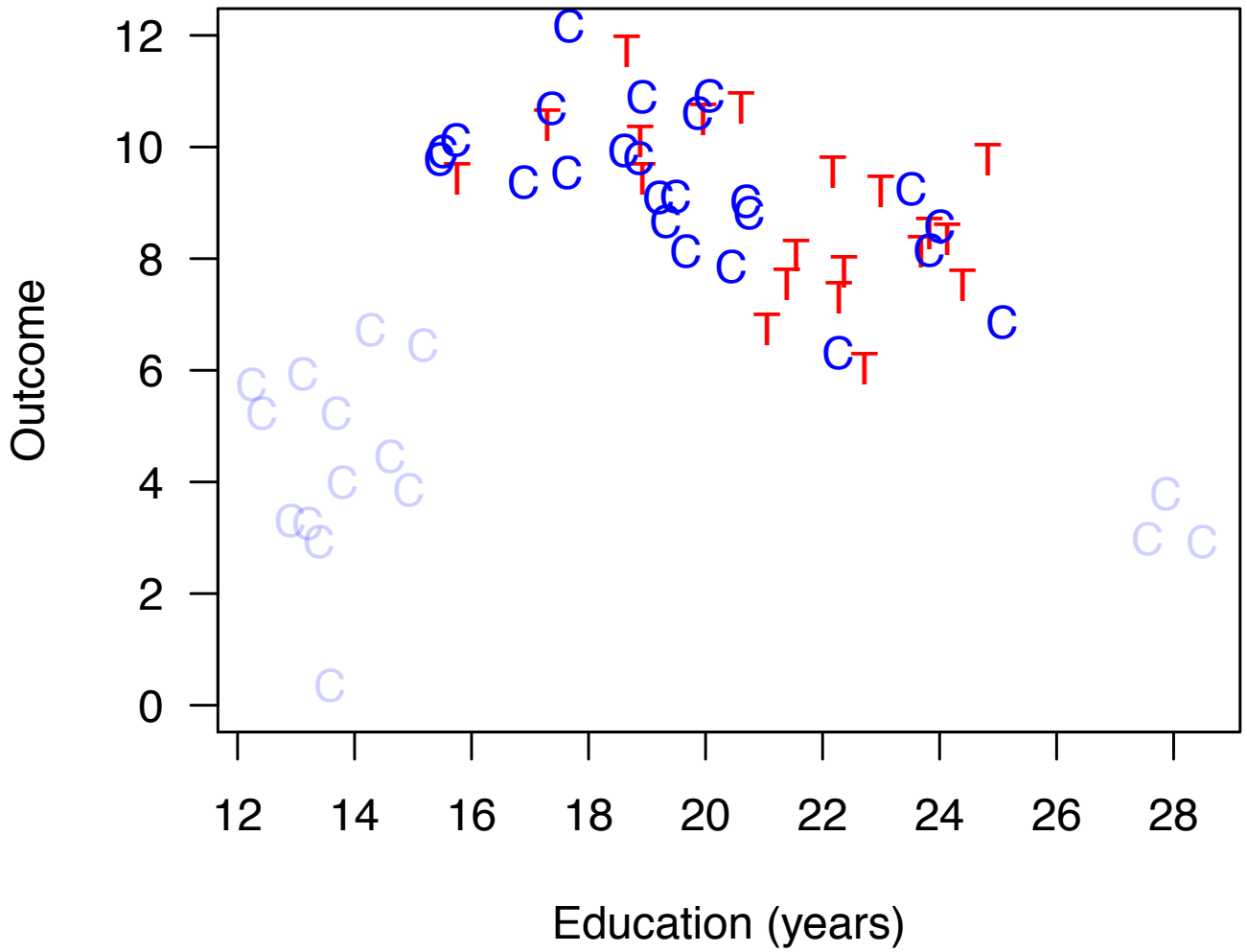
(Ho, Imai, King, Stuart, 2007: fig.1, *Political Analysis*)



(slides by King and Nielsen)

Matching to Reduce Model Dependence

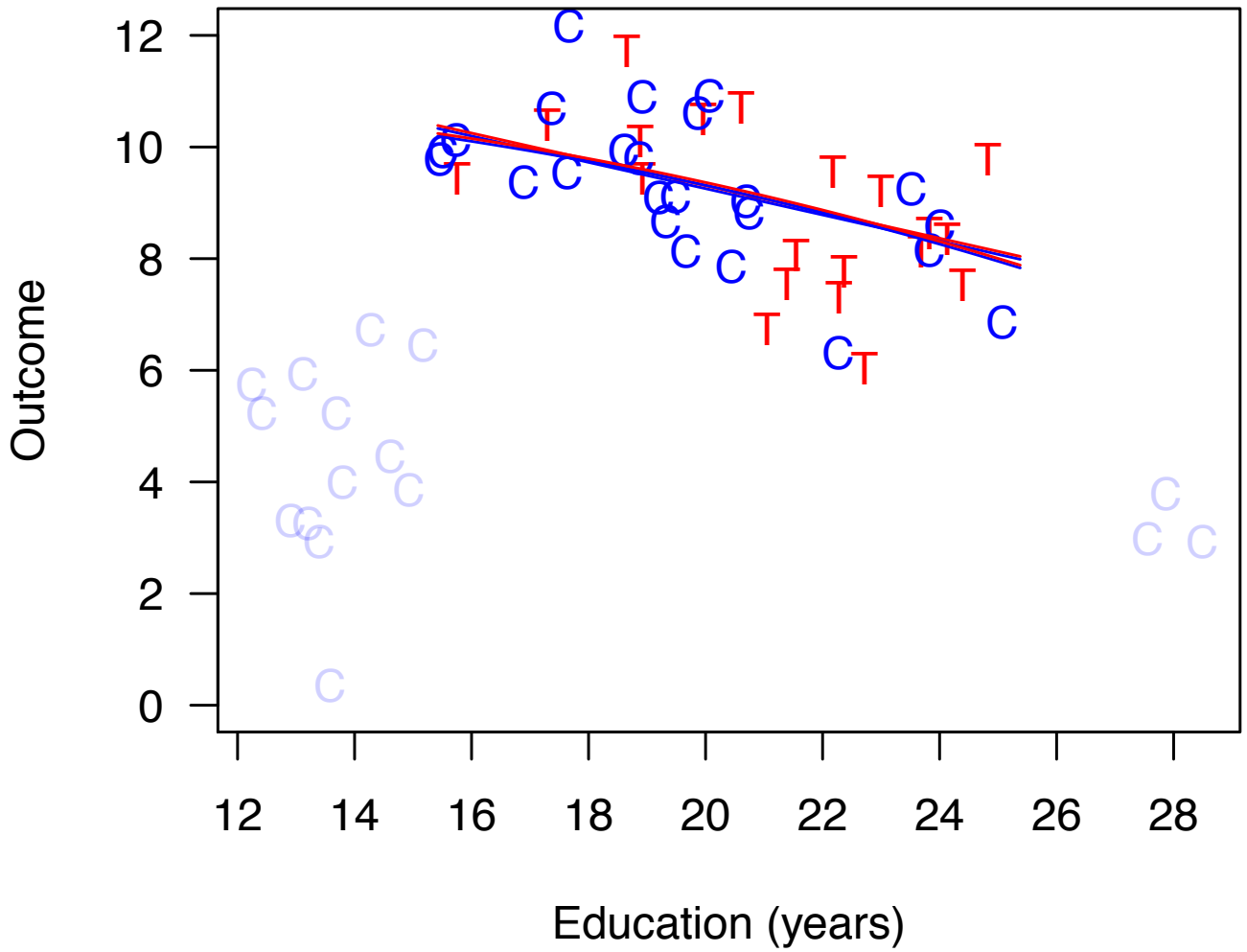
(Ho, Imai, King, Stuart, 2007: fig.1, *Political Analysis*)



(slides by King and Nielsen)

Matching to Reduce Model Dependence

(Ho, Imai, King, Stuart, 2007: fig.1, *Political Analysis*)



(slides by King and Nielsen)

King and Nielsen

- Argument 2
  - PSM approximates complete randomization.
  - Better are matching approaches that approximate fully blocked randomization, such as Mahalanobis matching, because complete randomization is less efficient than fully blocked randomization.

Types of Experiments

Balance Covariates:	*Complete Randomization*	*Fully Blocked*
Observed	On average	Exact
Unobserved	On average	On average

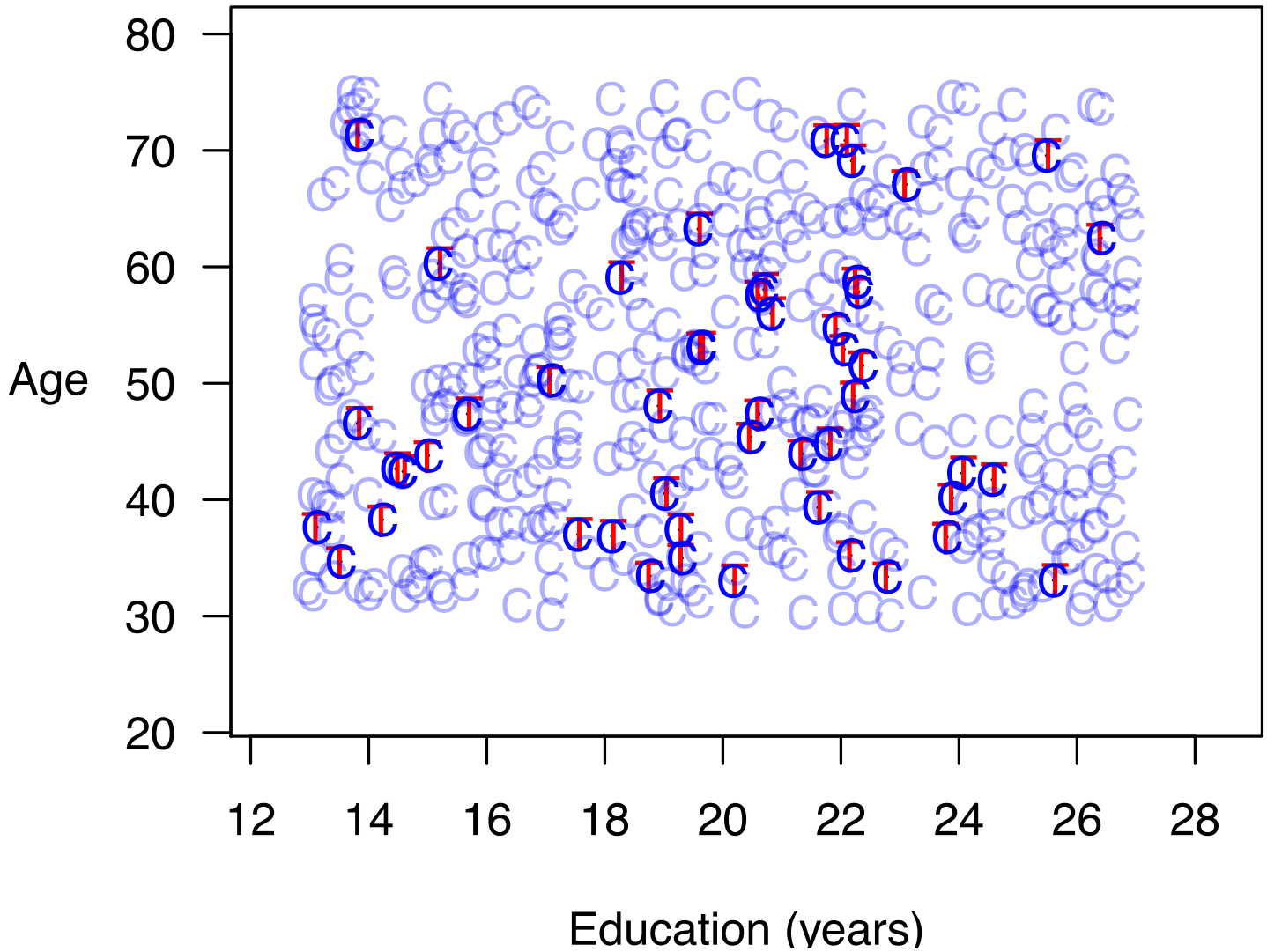
$\rightsquigarrow$ *Fully blocked* dominates *complete randomization* for: imbalance, model dependence, power, efficiency, bias, research costs, robustness. E.g., Imai, King, Nall 2009: SEs 600% smaller!

Goal of Each Matching Method (in Observational Data)

- PSM: *complete randomization*
- Other methods: *fully blocked*
- Other matching methods dominate PSM (wait, it gets worse)

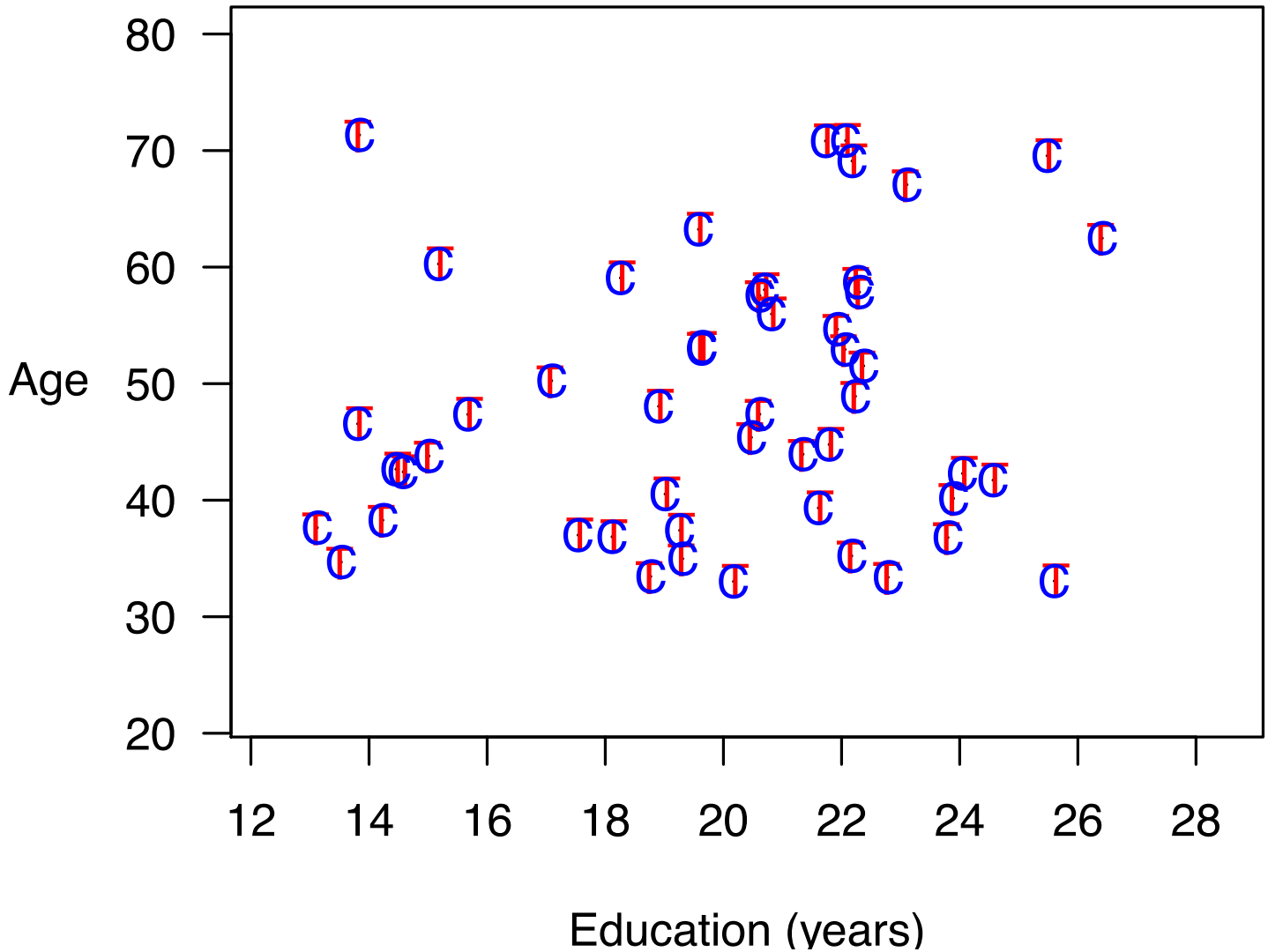
(slides by King and Nielsen)

Best Case: Mahalanobis Distance Matching



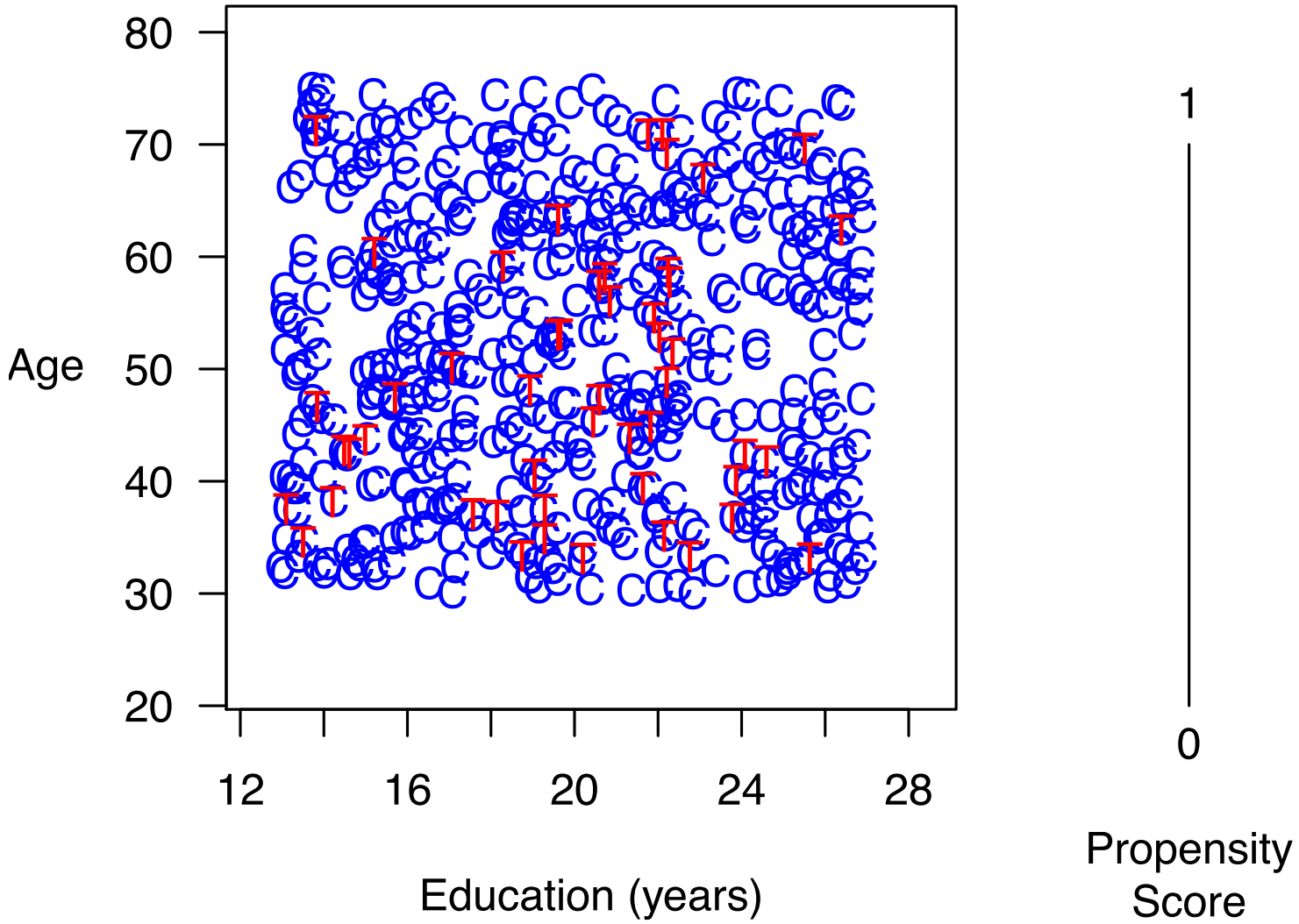
(slides by King and Nielsen)

Best Case: Mahalanobis Distance Matching



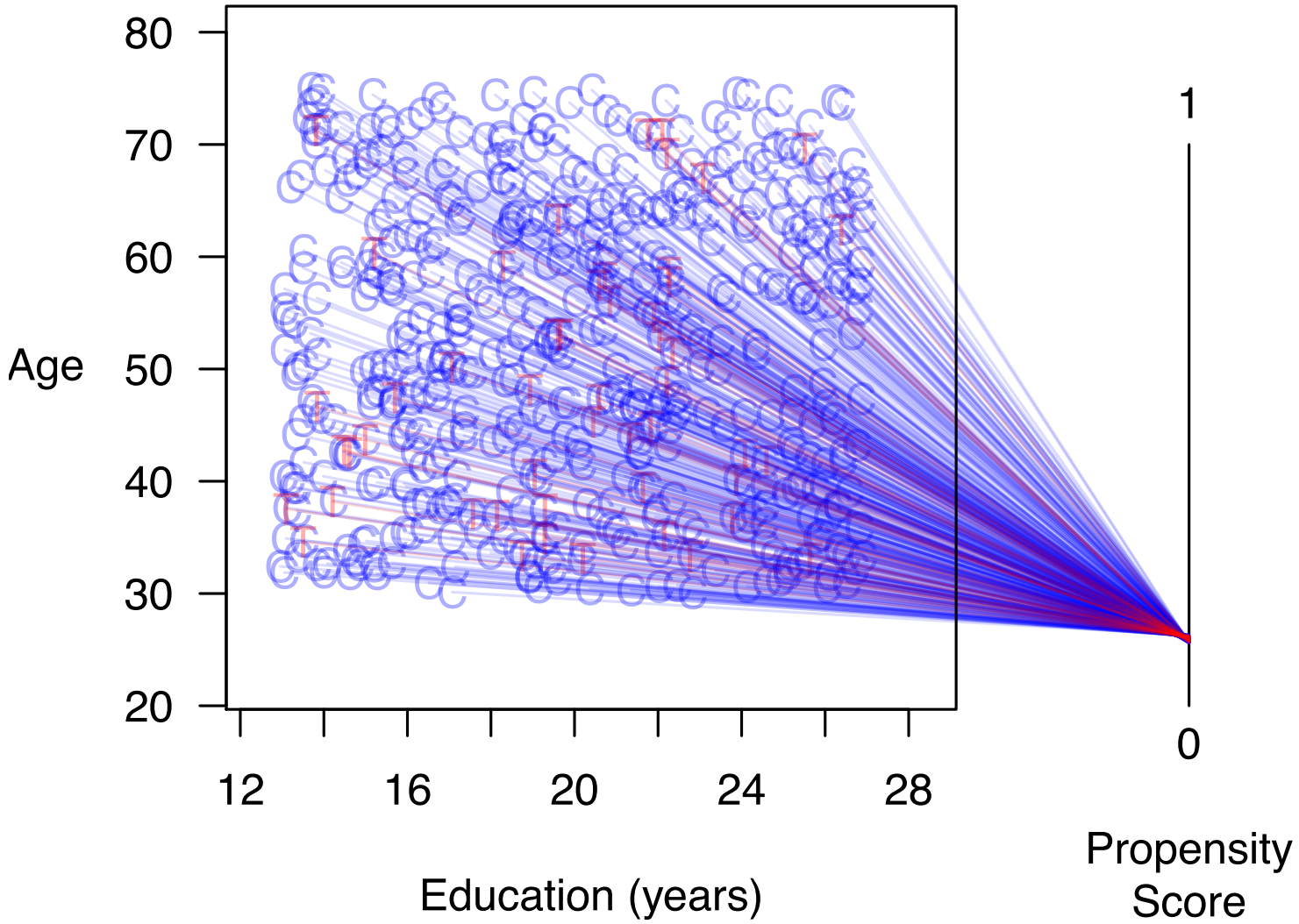
(slides by King and Nielsen)

Best Case: Propensity Score Matching



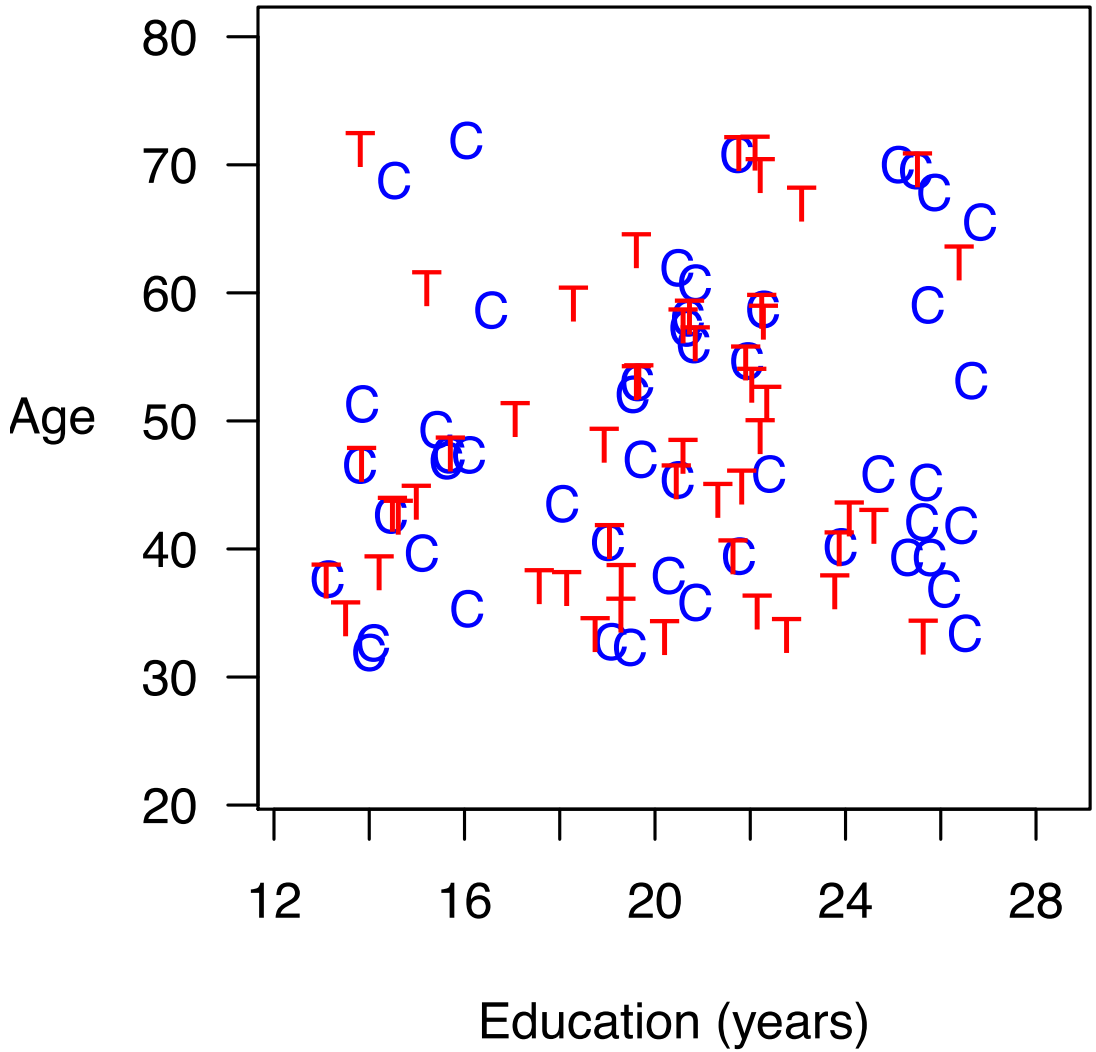
(slides by King and Nielsen)

Best Case: Propensity Score Matching



(slides by King and Nielsen)

Best Case: Propensity Score Matching is Suboptimal



15/23

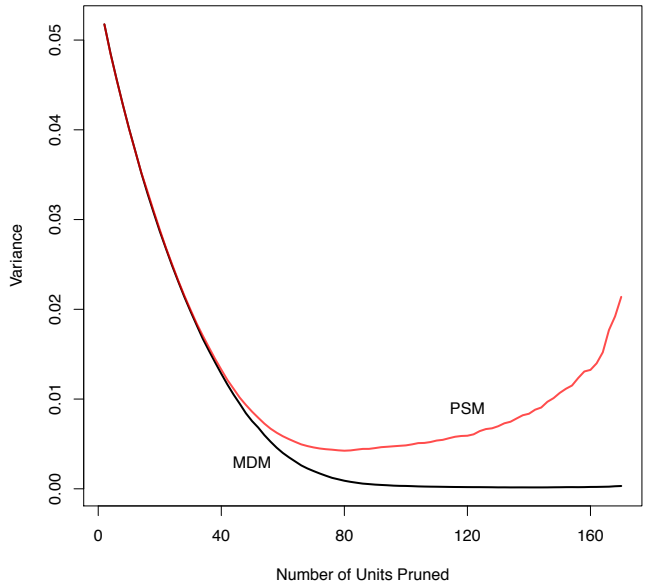
(slides by King and Nielsen)

King and Nielsen

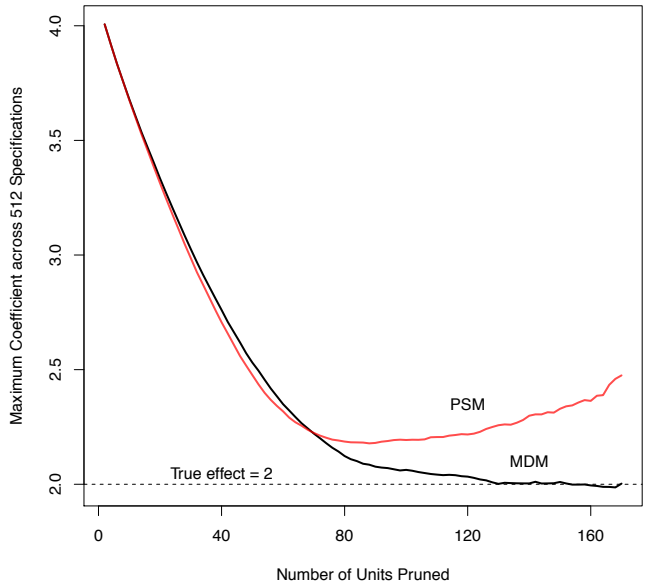
- Argument 3
  - Random pruning (deleting observations at random) increases imbalance. This is because the sample size decreases so that variance increases (large differences become more likely).
  - More imbalance/variance means more model dependence and researcher discretion.
  - Because PSM approximates complete randomization, it engages in random pruning.
  - PSM Paradox ("when you do 'better,' you do worse")
    - ★ When matching is made more strict (e.g., by decreasing the size of the caliper) PSM, like other matching methods, typically reduces imbalance. But soon the PSM Paradox kicks in, such that further pruning quickly increases imbalance.
    - ★ If the data is such that there are no big differences between treated and untreated to begin with, the PSM Paradox kicks in almost immediately.

PSM Increases Model Dependence & Bias

Model Dependence



Bias



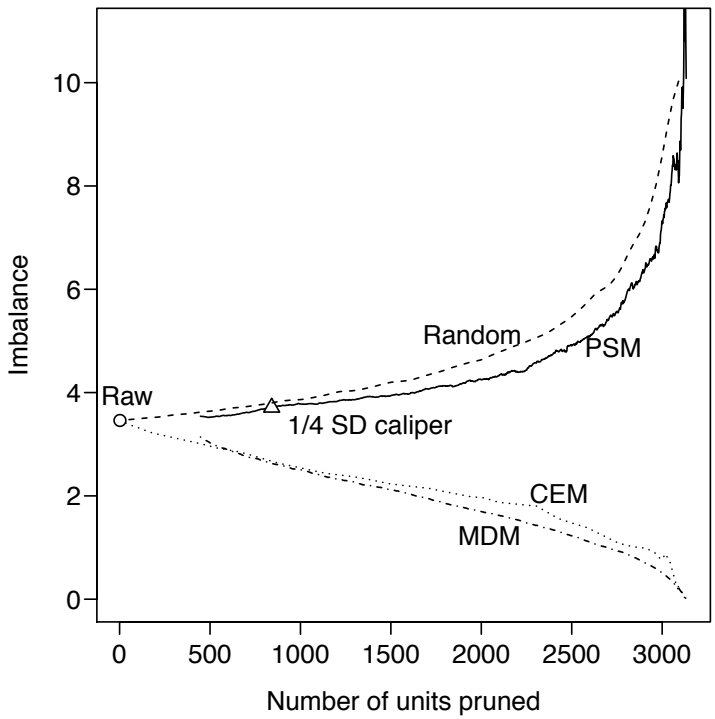
$$Y_i = 2T_i + X_{1i} + X_{2i} + \epsilon_i$$
$$\epsilon_i \sim N(0, 1)$$

20/23

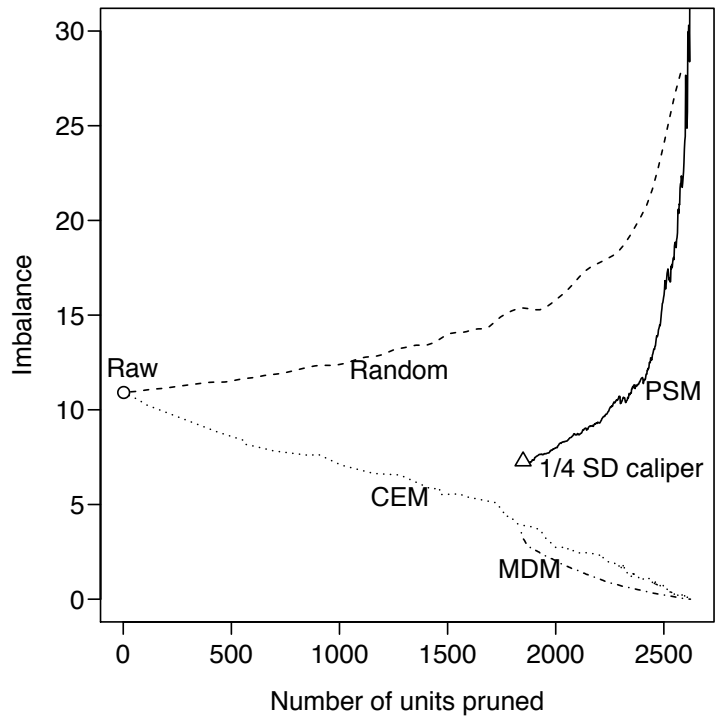
(slides by King and Nielsen)

The Propensity Score Paradox in Real Data

Finkel et al. (JOP, 2012)



Nielsen et al. (AJPS, 2011)



Similar pattern for $>$ 20 other real data sets we checked

(slides by King and Nielsen)

21/23

Are King and Nielsen right?

- Argument 1
  - Model dependence (i.e. dependence of results on modeling decisions made by the researcher) is bad because it leads to bias (people are selective in their decisions even if they try not to be).
  - Matching is good because it reduces model dependence.
- I fully agree!
- My view, however, may be somewhat less pessimistic. I believe that research results can be credible if researchers are well educated so that they know what they are doing and if modeling decisions are made transparent and robustness of results is evaluated (and documented).

Are King and Nielsen right?

- Argument 2
  - PSM approximates complete randomization.
  - Better are matching approaches that approximate fully blocked randomization, such as Mahalanobis matching, because complete randomization is less efficient than fully blocked randomization.
- That fully blocked randomization is more efficient than complete randomization – given the sample size – is of course true (how large the efficiency gains are depends on the strength of the relation between X and Y).
- However, if blocking reduces the sample size, it is not a priori clear whether estimates from the blocked sample are more efficient than estimates from the full sample (although often they will be).

Are King and Nielsen right?

- Argument 2
  - PSM approximates complete randomization.
  - Better are matching approaches that approximate fully blocked randomization, such as Mahalanobis matching, because complete randomization is less efficient than fully blocked randomization.
- That PSM approximates complete randomization is only partially true. PSM approximates complete randomization *within observations with the same propensity score*. Hence, PSM is somewhere between complete randomization and fully blocked randomization.
  - If the X variables have no relation to T (treatment), then all observations have the same propensity score. Hence we end up with complete randomization.
  - If the X variables have a strong effect on T, there is lots of blocking.

Are King and Nielsen right?

- Argument 3
  - Random pruning $\Rightarrow$ imbalance $\Rightarrow$ more model dependence.
  - PSM $\Rightarrow$ complete randomization $\Rightarrow$ lots of random pruning.
  - PSM Paradox: "when you do 'better,' you do worse"
- That random pruning makes things worse is, of course, true because it unnecessarily reduces the sample size (without changing anything else).
- As argued above, that PSM applies random pruning is only true for X variables unrelated to T (so that we are in a "local" complete randomization situation; although something similar can probably also happen if effects from several X's cancel each other out).
- Furthermore, it is only true if you employ a matching algorithm that throws away good matches! King and Nielsen's results seem to be based on the worst possible algorithm: one-to-one matching without replacement.

Are King and Nielsen right?

- Argument 3
  - Random pruning $\Rightarrow$ imbalance $\Rightarrow$ more model dependence.
  - PSM $\Rightarrow$ complete randomization $\Rightarrow$ lots of random pruning.
  - PSM Paradox: "when you do 'better,' you do worse"
- If you use a matching algorithm that does not throw away good matches, such as radius or kernel matching (or also nearest-neighbor matching as long as all ties are kept and observations are matched with replacement), random pruning can be avoided.
  - Such algorithms block (and hence prune) where it is necessary to prevent bias, but they average where such pruning is not necessary.
  - Hence, efficiency differences between PSM and multivariate matching should only be minor for such algorithms.

Are King and Nielsen right?

- Argument 3
  - Random pruning $\Rightarrow$ imbalance $\Rightarrow$ more model dependence.
  - PSM $\Rightarrow$ complete randomization $\Rightarrow$ lots of random pruning.
  - PSM Paradox: "when you do 'better,' you do worse"
- True is that post-matching modeling can do more harm with PSM than with MDM (because PSM leaves more "free" variance in X that can exploited by modeling decisions).
- In general, post-matching analyses are more limited for PSM than for MDM. For example, results from subgroup analyses will not be valid (you'd need to apply PSM stratified by subgroups in this case).

The Command

- `kmatch`: new matching software for Stata that has been written over the last few months; available from SSC (`ssc install kmatch`).
- Some key features:
  - Multivariate Distance Matching (MDM) and Propensity Score Matching (PSM) (or MDM and PSM combined).
  - Optional exact matching.
  - Optional regression-adjustment bias-correction.
  - Kernel matching, ridge matching, or nearest-neighbor matching.
  - Automatic bandwidth selection for kernel/ridge matching.
  - Flexible specification of scaling matrix for MDM.
  - Joint analysis of multiple subgroups and multiple outcome variables.
  - Various post-estimation commands for balancing and common-support diagnostics.
  - Computationally efficient implementation.

Some Examples

```
. // Use the NLSW data to estimate the "effect" of union membership on
. // wages, controlling for some covariated such as education, labor market
. // experience, or industry
. sysuse nlsw88, clear
(NLSW, 1988 extract)

. drop if industry==2
(4 observations deleted)

. // Mahalanobis-distance kernel matching
. kmatch md union collgrad ttl_exp tenure i.industry i.race south ///
>     (wage), nate att
(computing bandwidth ... done)
Multivariate-distance kernel matching           Number of obs     =      1,853
                                                Kernel            =       epan
Treatment : union = 1
Metric    : mahalanobis
Covariates: collgrad ttl_exp tenure i.industry i.race south
Matching statistics
```

	Matched			Controls			Band-
	Yes	No	Total	Used	Unused	Total	width
Treated	432	25	457	1105	291	1396	1.3394

Treatment-effects estimation

wage	Coef.
ATT	.6059013
NATE	1.432913

Some Examples

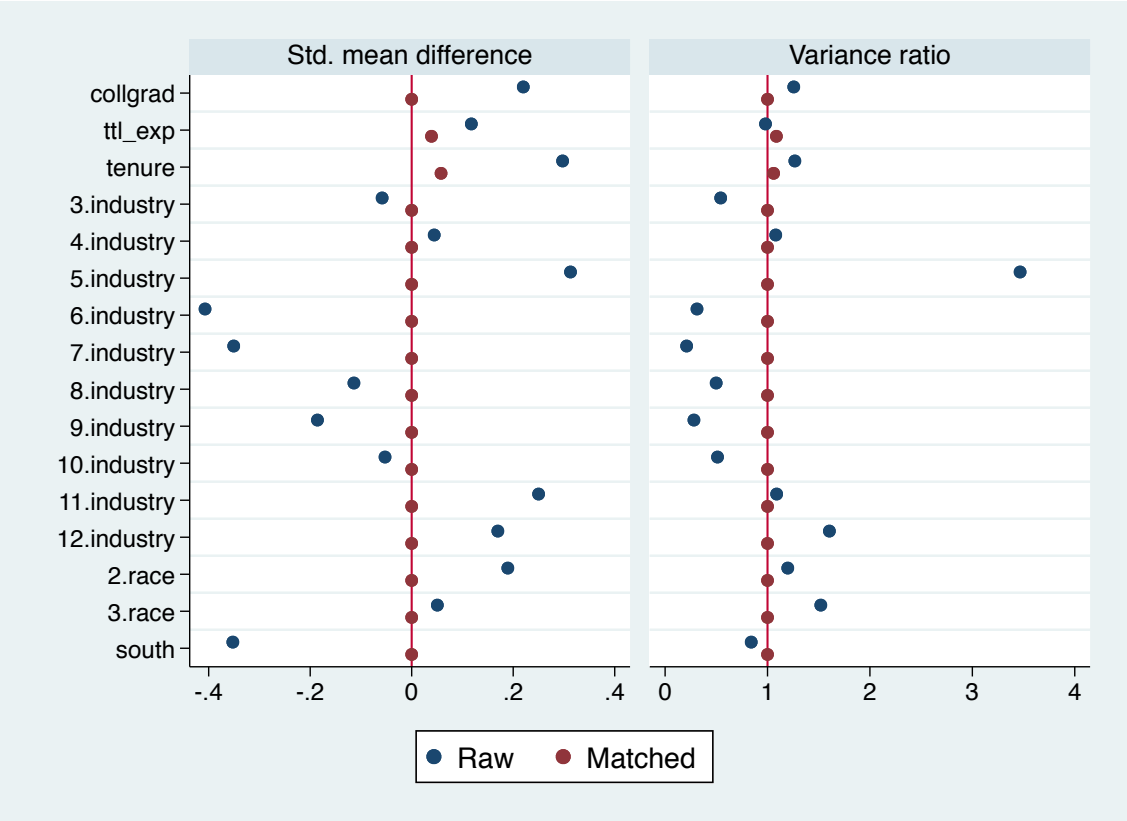
```
. // some balancing statistics
. kmatch summarize
(refitting the model using the generate() option)
```

Means	Raw Treated	Raw Untrea~d	Raw StdDif	Matched(ATT) Treated	Matched(ATT) Untrea~d	Matched(ATT) StdDif
collgrad	.321663	.224212	.219912	.319444	.319444	0
ttl_exp	13.2685	12.7323	.117584	13.3205	13.1425	.039036
tenure	7.89205	6.17658	.29735	7.91744	7.58347	.057888
3.industry	.006565	.012178	-.058246	.00463	.00463	0
4.industry	.183807	.166905	.044425	.185185	.185185	0
5.industry	.105033	.027937	.312944	.085648	.085648	0
6.industry	.045952	.169771	-.407129	.048611	.048611	0
7.industry	.019694	.102436	-.350657	.020833	.020833	0
8.industry	.017505	.035817	-.113785	.009259	.009259	0
9.industry	.010941	.040115	-.185669	.011574	.011574	0
10.industry	.004376	.008596	-.052551	.002315	.002315	0
11.industry	.479212	.356734	.250073	.506944	.506944	0
12.industry	.122538	.07235	.169707	.12037	.12037	0
2.race	.330416	.244986	.189418	.3125	.3125	0
3.race	.017505	.011461	.050566	.006944	.006944	0
south	.297593	.466332	-.352408	.291667	.291667	0

Variances	Raw Treated	Raw Untrea~d	Raw Ratio	Matched(ATT) Treated	Matched(ATT) Untrea~d	Matched(ATT) Ratio
collgrad	.218674	.174066	1.25628	.217904	.217904	1
ttl_exp	20.5898	21.0001	.980459	19.8177	18.2323	1.08696
tenure	37.2044	29.3629	1.26706	37.0399	34.9543	1.05966
3.industry	.006536	.012038	.542928	.004619	.004619	1
4.industry	.150351	.139148	1.08052	.151242	.151242	1
5.industry	.094207	.027176	3.46656	.078494	.078494	1
6.industry	.043936	.14105	.311496	.046355	.046355	1
7.industry	.019348	.092008	.210287	.020447	.020447	1
8.industry	.017237	.034559	.498769	.009195	.009195	1

Some Examples

```
. // make a graph of the balancing stats
. mat M = r(M)
. mat V = r(V)
. coefplot matrix(M[,3]) matrix(M[,6]) || matrix(V[,3]) matrix(V[,6]) || , ///
>     bylabels("Std. mean difference" "Variance ratio") ///
>     noci nolabels byopts(xrescale)
. addplot 1: , xline(0) norescaling legend(order(1 "Raw" 2 "Matched"))
. addplot 2: , xline(1) norescaling
```



Some Examples

```
. // Propensity-score kernel matching
. kmatch ps union collgrad ttl_exp tenure i.industry i.race south ///
>     (wage), nate att
(computing bandwidth ... done)
Propensity-score kernel matching                  Number of obs     =      1,853
                                                  Kernel            =       epan
Treatment : union = 1
Covariates: collgrad ttl_exp tenure i.industry i.race south
PS model  : logit (pr)
```

Matching statistics

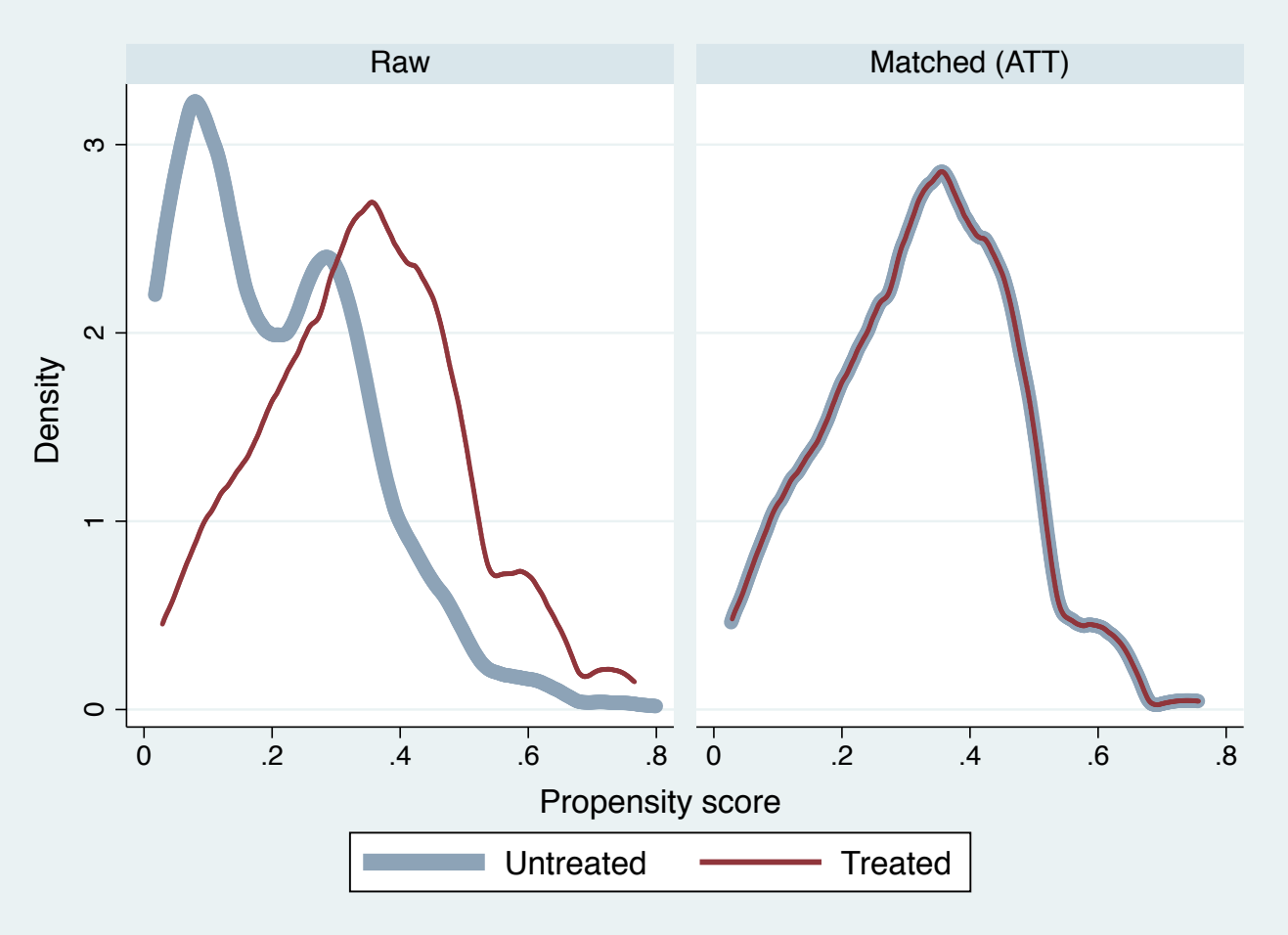
	Matched			Controls			Band-
	Yes	No	Total	Used	Unused	Total	width
Treated	431	26	457	1214	182	1396	.00188

Treatment-effects estimation

wage	Coef.
ATT	.3887224
NATE	1.432913

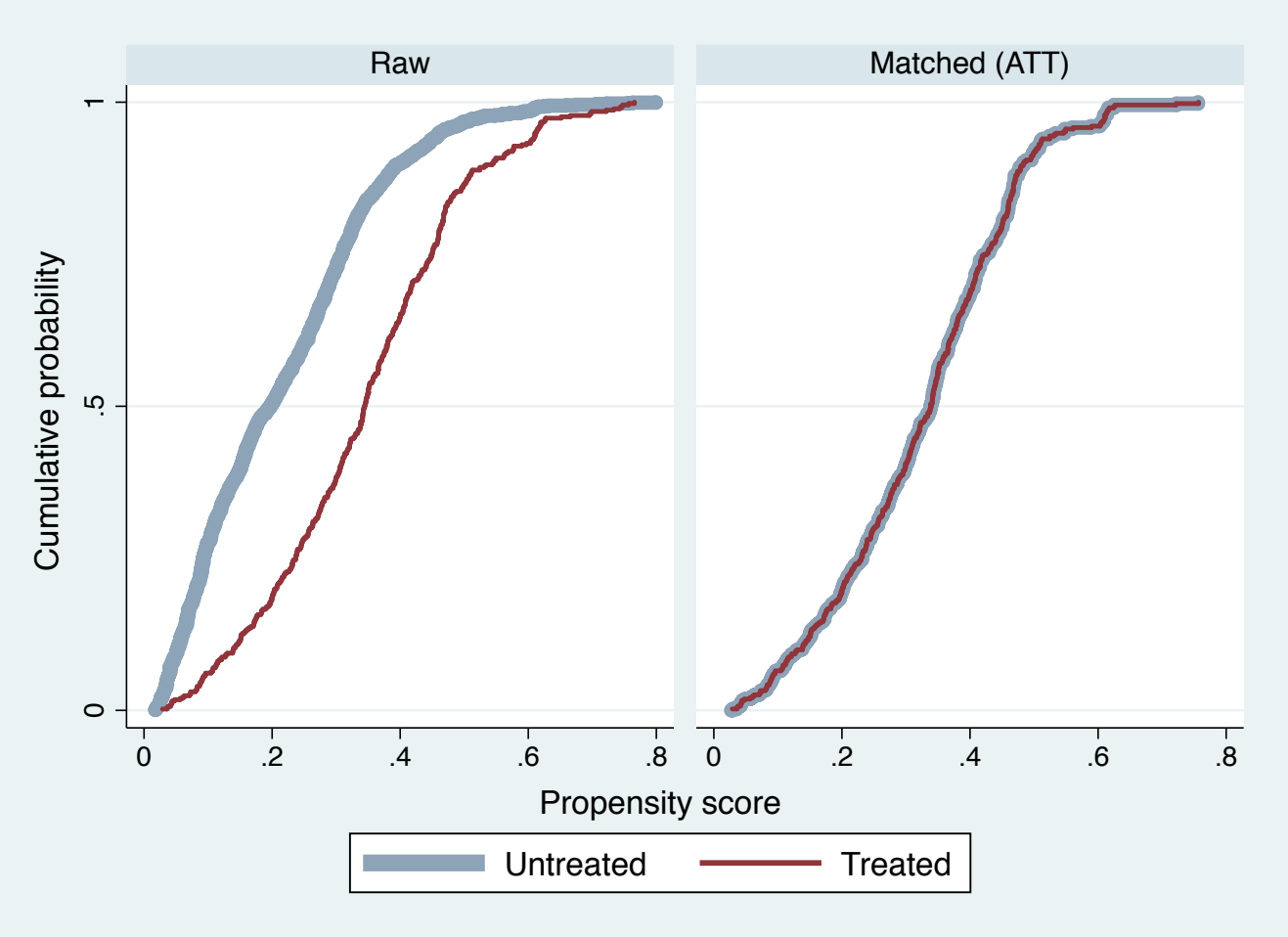
Some Examples

```
. // Kernel density balancing plot
. kmatch density, lw(*6 *2) lc(*.5 *1)
(refitting the model using the generate() option)
(applying 0-1 boundary correction to density estimation of propensity score)
(bandwidth for propensity score = .06803989)
```



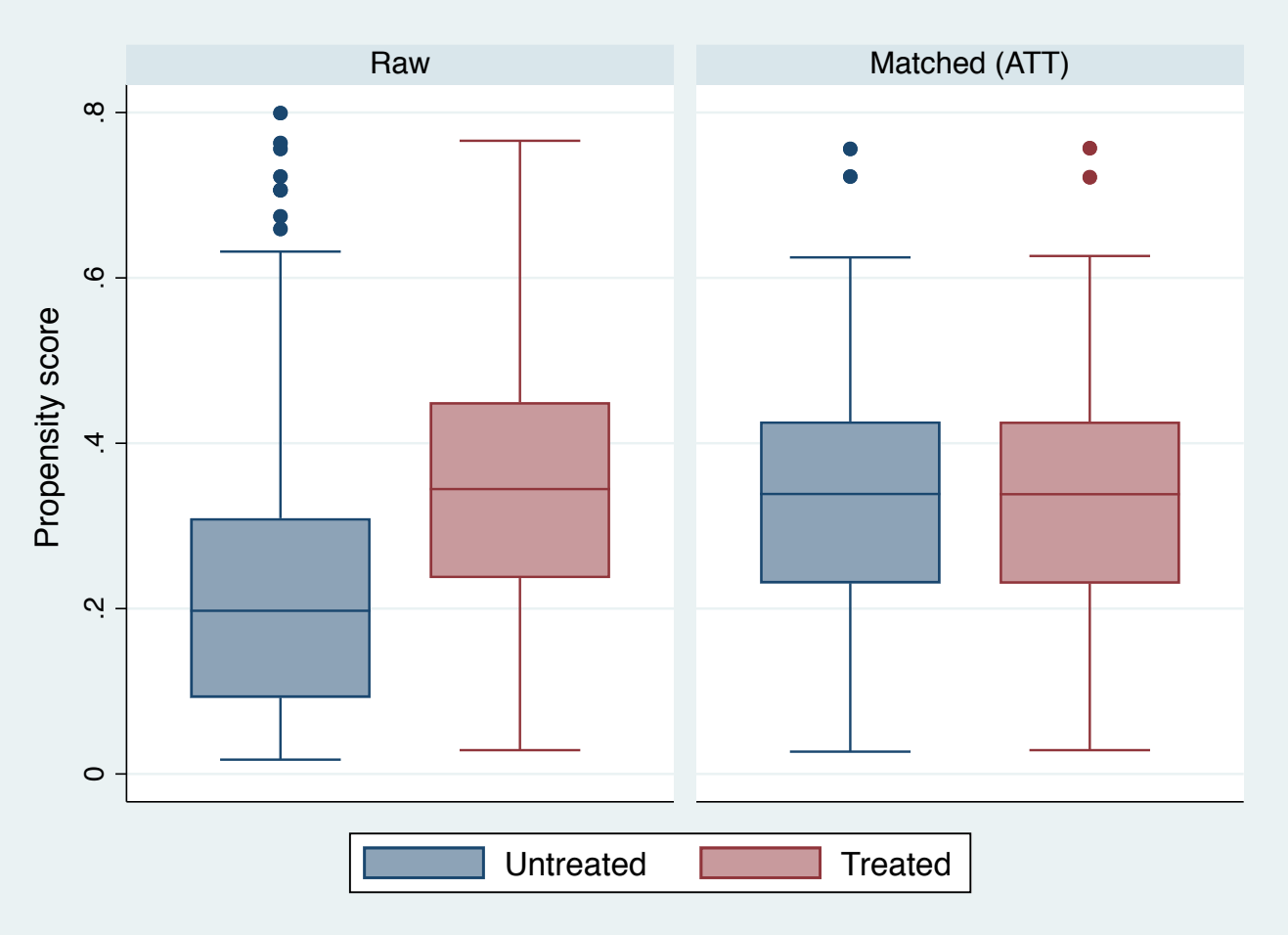
Some Examples

```
. // Cumulative distribution balancing plot
. kmatch cumul, lw(*6 *2) lc(*.5 *1)
(refitting the model using the generate() option)
```



Some Examples

```
. // Balancing box plot
. kmatch box
(refitting the model using the generate() option)
```



Some Examples

```
. // Standard errors
. kmatch md union collgrad ttl_exp tenure i.industry i.race south ///
>     (wage), nate ate att atc vce(bootstrap)
(computing bandwidth for treated ... done)
(computing bandwidth for untreated ... done)

(running kmatch on estimation sample)

Bootstrap replications (50)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50

Multivariate-distance kernel matching          Number of obs     =      1,853
                                               Replications      =         50
                                               Kernel            =       epan

Treatment : union = 1
Metric    : mahalanobis
Covariates: collgrad ttl_exp tenure i.industry i.race south

Matching statistics
```

	Matched			Controls			Band-
	Yes	No	Total	Used	Unused	Total	width
Treated	432	25	457	1105	291	1396	1.3394
Untreated	1386	10	1396	455	2	457	3.3975
Combined	1818	35	1853	1560	293	1853	.

Treatment-effects estimation

wage	Observed Coef.	Bootstrap Std. Err.	z	P>|z|	Normal-based [95% Conf.	Interval]
ATE	.4095729	.1920853	2.13	0.033	.0330928	.7860531
ATT	.6059013	.2472069	2.45	0.014	.1213846	1.090418
ATC	.3483797	.1893653	1.84	0.066	-.0227695	.7195289
NATE	1.432913	.2333282	6.14	0.000	.9755981	1.890228

Some Examples

```
. // Do some tests
. lincom ATT-NATE
 ( 1)  ATT - NATE = 0
```

wage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
(1)	-.8270117	.1810415	-4.57	0.000	-1.181847	-.4721768

```
. test ATT = ATC
 ( 1)  ATT - ATC = 0
           chi2(  1) =    2.42
         Prob > chi2 =    0.1200
```

Some Examples

```
. // Nearest-neighbor matching (1 neighbor)
. kmatch md union collgrad ttl_exp tenure i.industry i.race south (wage), att nn
Multivariate-distance nearest-neighbor matching
                                              Number of obs     =      1,853
                                              Neighbors:    min =          1
Treatment : union = 1                                       max =          1
Metric    : mahalanobis
Covariates: collgrad ttl_exp tenure i.industry i.race south
Matching statistics
```

	Matched			Controls			Band-
	Yes	No	Total	Used	Unused	Total	width
Treated	457	0	457	328	1068	1396	.

Treatment-effects estimation

wage	Coef.
ATT	.7246969

```
. teffects nnmatch (wage collgrad ttl_exp tenure i.industry i.race south) (union), atet
Treatment-effects estimation                  Number of obs      =      1,853
Estimator      : nearest-neighbor matching    Matches: requested =          1
Outcome model  : matching                                  min =          1
Distance metric: Mahalanobis                               max =          1
```

wage	Coef.	AI Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
ATET						
union (union vs nonunion)	.7246969	.2942952	2.46	0.014	.147889	1.301505

Some Examples

```
. // Nearest-neighbor matching (5 neighbors)
. kmatch md union collgrad ttl_exp tenure i.industry i.race south (wage), att nn(5)
Multivariate-distance nearest-neighbor matching
                                              Number of obs     =      1,853
                                              Neighbors:    min =          5
Treatment : union = 1                                       max =          5
Metric    : mahalanobis
Covariates: collgrad ttl_exp tenure i.industry i.race south
Matching statistics
```

	Matched			Controls			Band-
	Yes	No	Total	Used	Unused	Total	width
Treated	457	0	457	870	526	1396	.

Treatment-effects estimation

wage	Coef.
ATT	.5590823

```
. teffects nnmatch (wage collgrad ttl_exp tenure i.industry i.race south) (union), atet nn(5)
Treatment-effects estimation                Number of obs      =      1,853
Estimator      : nearest-neighbor matching  Matches: requested =          5
Outcome model  : matching                               min =          5
Distance metric: Mahalanobis                            max =          6
```

wage	Coef.	AI Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
ATET						
union (union vs nonunion)	.5590823	.2381752	2.35	0.019	.0922675	1.025897

Some Examples

```
. // Bias-correction / regression adjustment
. kmatch md union collgrad ttl_exp tenure i.industry i.race south ///
>     (wage = collgrad ttl_exp tenure i.industry i.race south), att nn(5)
Multivariate-distance nearest-neighbor matching
                                                Number of obs     =      1,853
                                                Neighbors:    min =          5
Treatment : union = 1                                         max =          5
Metric    : mahalanobis
Covariates: collgrad ttl_exp tenure i.industry i.race south
Matching statistics
```

	Matched			Controls			Band-
	Yes	No	Total	Used	Unused	Total	width
Treated	457	0	457	870	526	1396	.

```
Treatment-effects estimation
```

wage	Coef.
ATT	.5288023

```
adjusted for collgrad ttl_exp tenure i.industry i.race south
. teffects nnmatch (wage collgrad ttl_exp tenure i.industry i.race south) ///
>     (union), atet nn(5) biasadj(collgrad ttl_exp tenure i.industry i.race south)
Treatment-effects estimation                  Number of obs      =      1,853
Estimator      : nearest-neighbor matching    Matches: requested =          5
Outcome model  : matching                                  min =          5
Distance metric: Mahalanobis                               max =          6
```

wage	Coef.	AI Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
ATET						
union (union vs nonunion)	.5288023	.2420635	2.18	0.029	.0543666	1.003238

Some Examples

```
. // Mahalanobis-distance and propensity-score matching combined
. kmatch md union collgrad ttl_exp tenure (wage), att ///
>     psvars(i.industry i.race south) psweight(3)
(computing bandwidth ... done)
Multivariate-distance kernel matching           Number of obs     =      1,853
                                                Kernel            =       epan
Treatment : union = 1
Metric    : mahalanobis (modified)
Covariates: collgrad ttl_exp tenure
PS model  : logit (pr)
PS covars : i.industry i.race south
```

Matching statistics

	Matched			Controls			Band-
	Yes	No	Total	Used	Unused	Total	width
Treated	439	18	457	1258	138	1396	.83886

Treatment-effects estimation

wage	Coef.
ATT	.6408443

Some Examples

```
. // Exact matching
. kmatch md union collgrad ttl_exp tenure (wage), att ematch(industry race south)
(computing bandwidth ... done)
Multivariate-distance kernel matching           Number of obs     =      1,853
                                                Kernel            =       epan
Treatment : union = 1
Metric    : mahalanobis
Covariates: collgrad ttl_exp tenure
Exact     : industry race south
```

Matching statistics

	Matched			Controls			Band-
	Yes	No	Total	Used	Unused	Total	width
Treated	432	25	457	1103	293	1396	1.3013

Treatment-effects estimation

wage	Coef.
ATT	.6047374

Some Examples

```
. // Bandwidth selection: the default (based on distribution of distances in
. // one-nearest-neighbor matching)
. kmatch md union collgrad ttl_exp tenure i.industry i.race south (wage), att
(computing bandwidth ... done)
Multivariate-distance kernel matching           Number of obs     =      1,853
                                                Kernel            =       epan
Treatment : union = 1
Metric    : mahalanobis
Covariates: collgrad ttl_exp tenure i.industry i.race south
```

Matching statistics

	Matched			Controls			Band-
	Yes	No	Total	Used	Unused	Total	width
Treated	432	25	457	1105	291	1396	1.3394

Treatment-effects estimation

wage	Coef.
ATT	.6059013

Some Examples

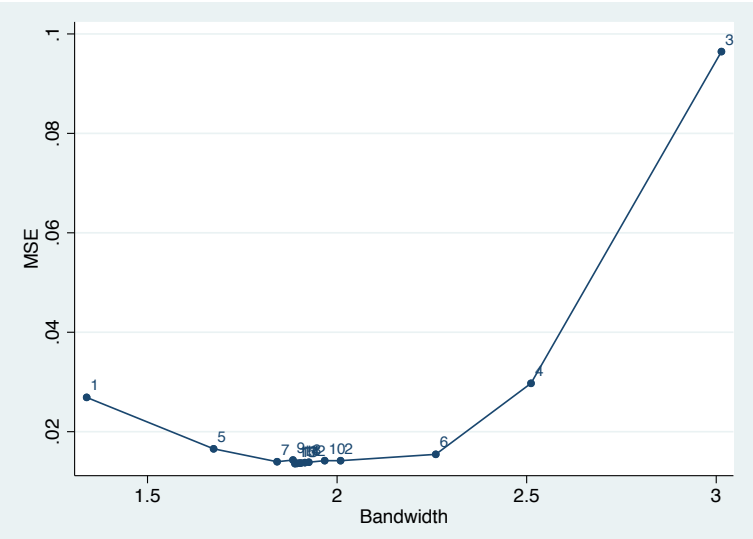
```
. // Bandwidth selection: cross validation with respect to X
. kmatch md union collgrad ttl_exp tenure i.industry i.race south (wage), ///
>     att bwidth(cv)
(computing bandwidth ..... done)
Multivariate-distance kernel matching           Number of obs     =      1,853
                                                Kernel            =       epan
Treatment : union = 1
Metric    : mahalanobis
Covariates: collgrad ttl_exp tenure i.industry i.race south
Matching statistics
```

	Matched			Controls			Band-
	Yes	No	Total	Used	Unused	Total	width
Treated	448	9	457	1184	212	1396	1.8888

Treatment-effects estimation

wage	Coef.
ATT	.6651578

```
. kmatch cvplot, ms(o) index mlabposition(1) sort
```



Some Examples

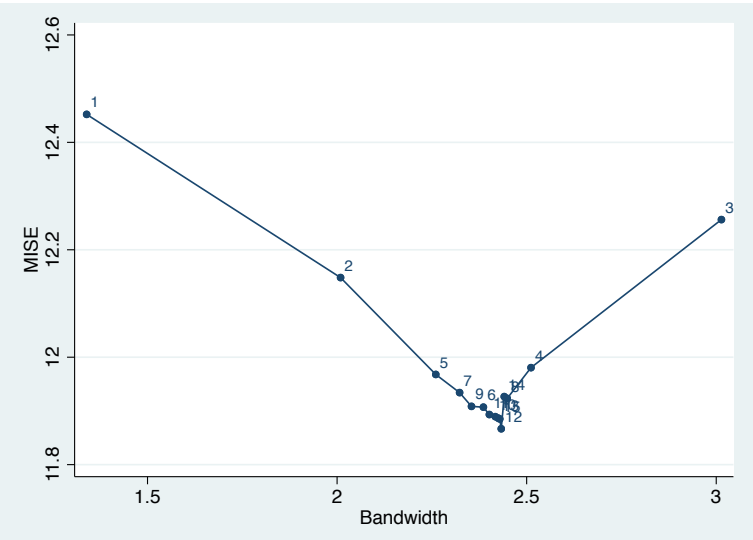
```
. // Bandwidth selection: cross validation with respect to Y
. kmatch md union collgrad ttl_exp tenure i.industry i.race south (wage), ///
>     att bwidth(cv wage)
(computing bandwidth ..... done)
Multivariate-distance kernel matching           Number of obs     =      1,853
                                                Kernel            =       epan
Treatment : union = 1
Metric    : mahalanobis
Covariates: collgrad ttl_exp tenure i.industry i.race south
Matching statistics
```

	Matched			Controls			Band-
	Yes	No	Total	Used	Unused	Total	width
Treated	453	4	457	1289	107	1396	2.433

Treatment-effects estimation

wage	Coef.
ATT	.6928956

```
. kmatch cvplot, ms(o) index mlabposition(1) sort
```



Some Examples

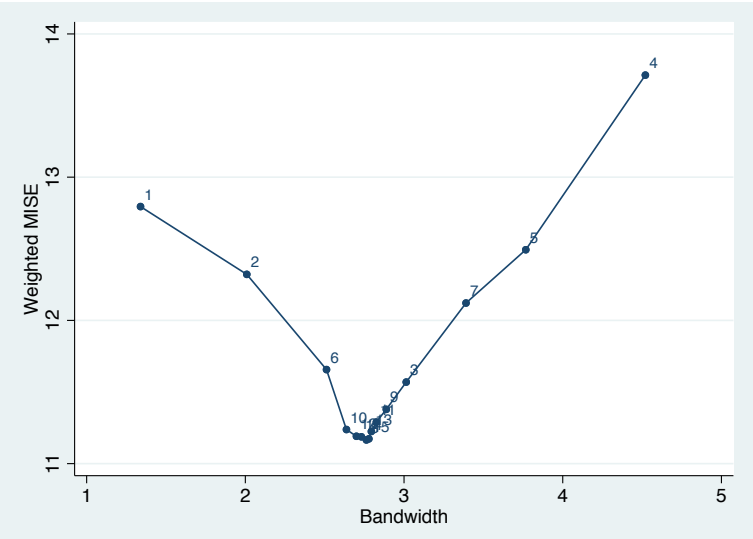
```
. // Bandwidth selection: weighted cross validation with respect to Y
. kmatch md union collgrad ttl_exp tenure i.industry i.race south (wage), ///
>     att bwidth(cv wage, weighted)
(computing bandwidth ..... done)
Multivariate-distance kernel matching           Number of obs     =      1,853
                                                Kernel            =       epan
Treatment : union = 1
Metric    : mahalanobis
Covariates: collgrad ttl_exp tenure i.industry i.race south
Matching statistics
```

	Matched			Controls			Band-
	Yes	No	Total	Used	Unused	Total	width
Treated	455	2	457	1356	40	1396	2.7626

Treatment-effects estimation

wage	Coef.
ATT	.7308166

```
. kmatch cvplot, ms(o) index mlabposition(1) sort
```



Some Examples

```
. // Common-support diagnostics
. kmatch md union collgrad ttl_exp tenure i.industry i.race south (wage), ///
>     att bwidth(0.5)
Multivariate-distance kernel matching           Number of obs     =      1,853
                                                Kernel            =       epan
Treatment : union = 1
Metric    : mahalanobis
Covariates: collgrad ttl_exp tenure i.industry i.race south
Matching statistics
```

	Matched			Controls			Band-
	Yes	No	Total	Used	Unused	Total	width
Treated	366	91	457	701	695	1396	.5

Treatment-effects estimation

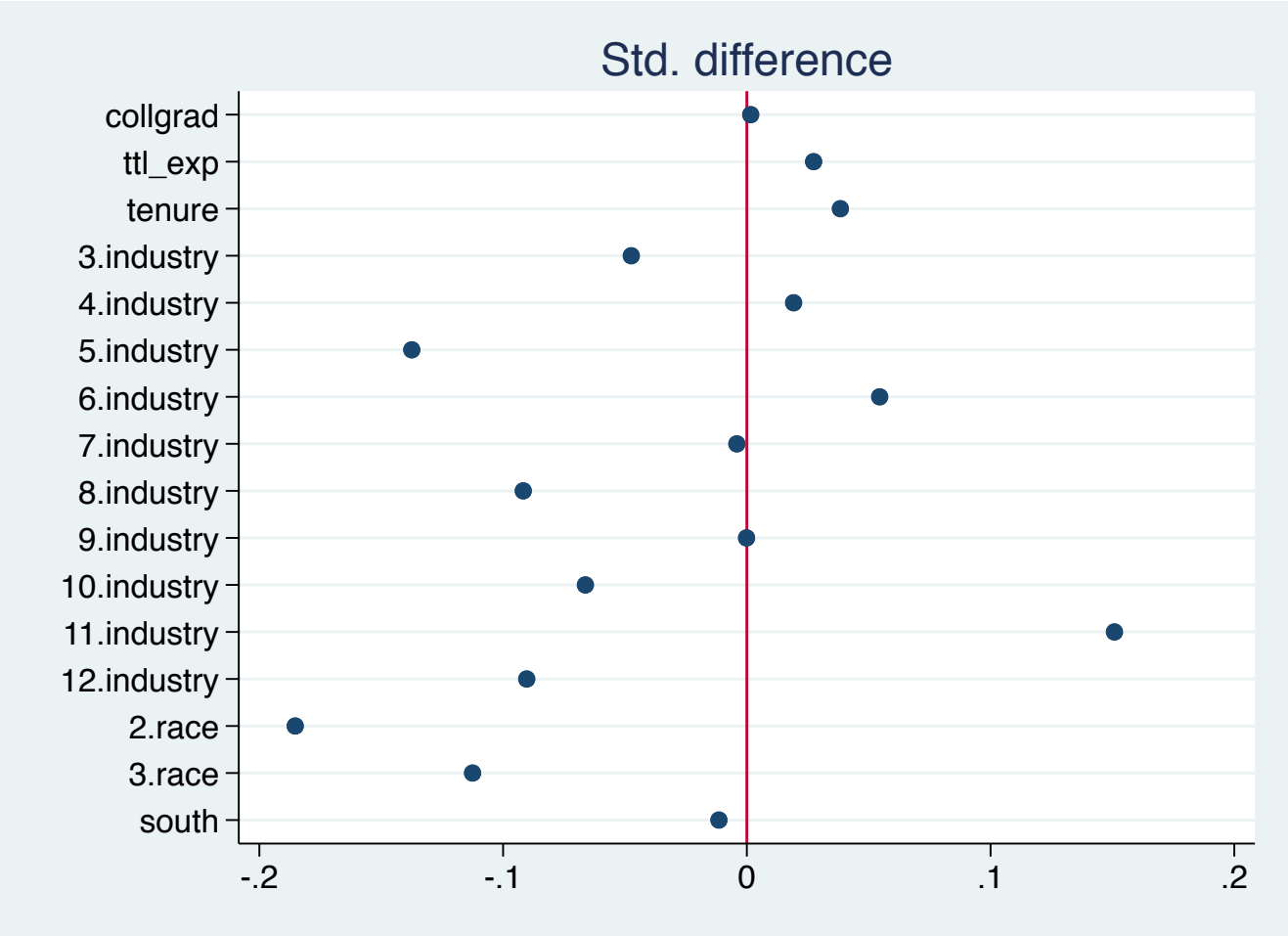
wage	Coef.
ATT	.3303161

```
. kmatch csummarize
(refitting the model using the generate() option)
```

	Common support (treated)			Standardized difference		
Means	Matched	Unmatc~d	Total	(1)-(3)	(2)-(3)	(1)-(2)
collgrad	.322404	.318681	.321663	.001585	-.006376	.007962
ttl_exp	13.3929	12.7682	13.2685	.027413	-.110253	.137666
tenure	8.12614	6.95055	7.89205	.038378	-.154356	.192734
3.industry	.002732	.021978	.006565	-.047404	.190657	-.238061
4.industry	.191257	.153846	.183807	.019212	-.077269	.096481
5.industry	.062842	.274725	.105033	-.137462	.552867	-.690329
6.industry	.057377	0	.045952	.054507	-.219225	.273732
7.industry	.019126	.021978	.019694	-.004083	.016423	-.020506
8.industry	.005464	.065934	.017505	-.091714	.368871	-.460585
9.industry	.010929	.010989	.010941	-.000115	.000462	-.000577
10.industry	0	.021978	.004376	-.066227	.266363	-.332589
11.industry	.554645	.175824	.479212	.15083	-.606636	.757467
12.industry	.092896	.241758	.122538	-.090299	.363181	-.45348
2.race	.243169	.681319	.330416	-.185284	.745209	-.930494
3.race	.002732	.076923	.017505	-.112525	.452572	-.565097

Some Examples

```
. // make a graph of the common-support stats
. mat M = r(M)
. coefplot matrix(M[,4]), title("Std. difference") noci nolabels xline(0)
```



Some Examples

```
. // Multiple outcome variables
. kmatch md union collgrad ttl_exp tenure i.industry i.race south ///
>     (wage hours), nate att
(computing bandwidth ... done)
Multivariate-distance kernel matching           Number of obs     =      1,852
                                                Kernel            =       epan
Treatment : union = 1
Metric    : mahalanobis
Covariates: collgrad ttl_exp tenure i.industry i.race south
```

Matching statistics

	Matched			Controls			Band-
	Yes	No	Total	Used	Unused	Total	width
Treated	432	25	457	1104	291	1395	1.3392

Treatment-effects estimation

	Coef.
wage	
ATT	.6021049
NATE	1.430823
hours	
ATT	1.263759
NATE	1.450303

Some Examples

```
. // Multiple outcome variables with different regression-adjustment
. // equations
. kmatch md union collgrad ttl_exp tenure i.industry i.race south ///
>     (wage = collgrad ttl_exp tenure) ///
>     (hours = i.industry i.race), nate att
(computing bandwidth ... done)
Multivariate-distance kernel matching           Number of obs     =      1,852
                                                Kernel            =       epan
Treatment : union = 1
Metric    : mahalanobis
Covariates: collgrad ttl_exp tenure i.industry i.race south
Matching statistics
```

	Matched			Controls			Band-
	Yes	No	Total	Used	Unused	Total	width
Treated	432	25	457	1104	291	1395	1.3392

Treatment-effects estimation

	Coef.
wage	
ATT	.5152752
NATE	1.430823
hours	
ATT	1.263759
NATE	1.450303

wage: adjusted for collgrad ttl_exp tenure
hours: adjusted for i.industry i.race

Some Examples

```
. // Treatment effects by subpopulation
. kmatch md union collgrad ttl_exp tenure i.industry i.race (wage), ///
>     att vce(boot) over(south)
(south=0: computing bandwidth ... done)
(south=1: computing bandwidth ... done)

(running kmatch on estimation sample)

Bootstrap replications (50)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50

Multivariate-distance kernel matching          Number of obs     =       1,853
                                               Replications      =          50
                                               Kernel            =       epan
Treatment : union = 1
Metric    : mahalanobis
Covariates: collgrad ttl_exp tenure i.industry i.race

         0: south = 0
         1: south = 1

Matching statistics
```

	Matched			Controls			Band-
	Yes	No	Total	Used	Unused	Total	width
0							
Treated	306	15	321	625	120	745	1.3199
1							
Treated	126	10	136	473	178	651	1.3398

Treatment-effects estimation

wage	Observed Coef.	Bootstrap Std. Err.	z	P>|z|	Normal-based [95% Conf.	Interval]
0						
ATT	.4586332	.2763358	1.66	0.097	-.082975	1.000241
1						
ATT	.9518705	.406903	2.34	0.019	.1543553	1.749386

```
. test [0]ATT = [1]ATT
 ( 1)  [0]ATT - [1]ATT = 0

           chi2(  1) =    1.23
         Prob > chi2 =    0.2679
. lincom [1]ATT - [0]ATT
 ( 1)  - [0]ATT + [1]ATT = 0
```

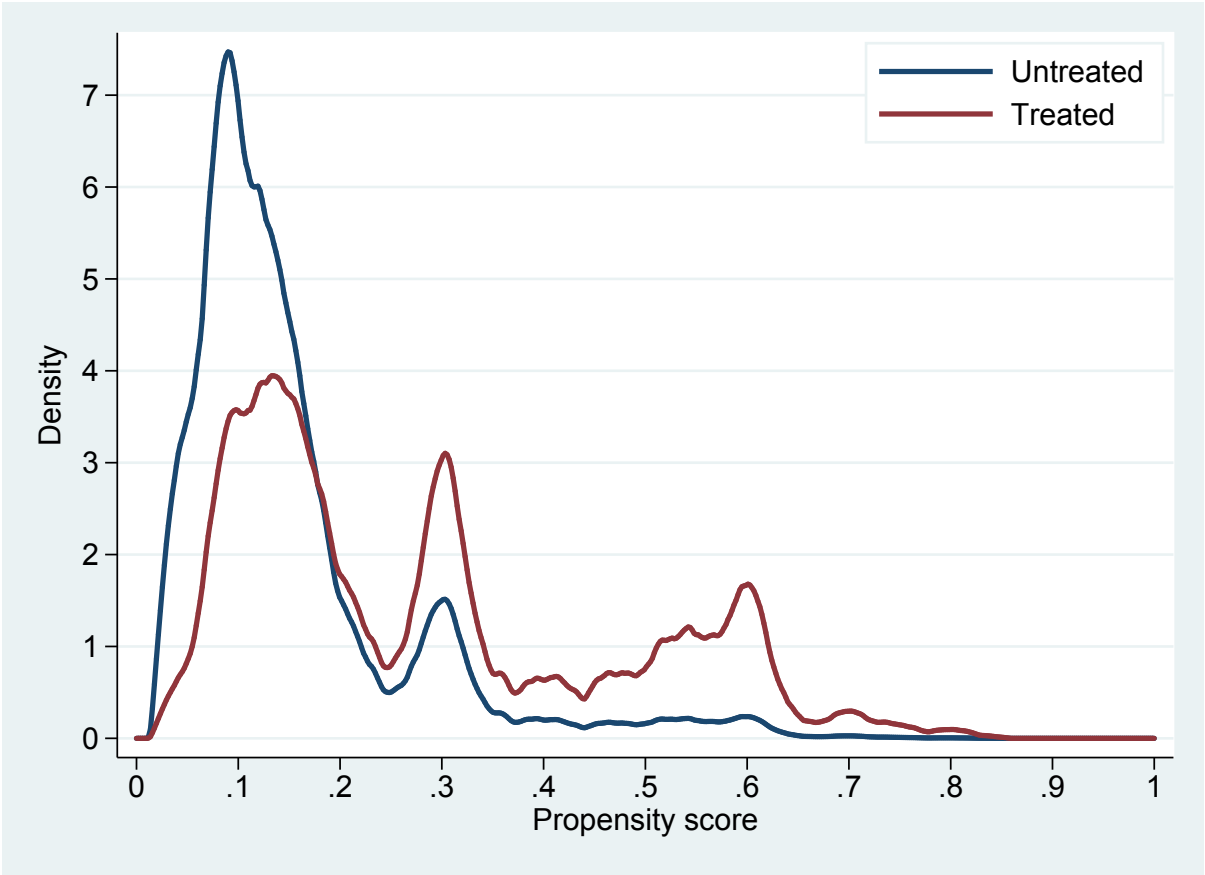
wage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
(1)	.4932373	.4452343	1.11	0.268	-.379406	1.365881

Simulation

- Population data from Swiss census of 2000.
- Outcome: Treiman occupational prestige (recoded from ISCO codes of the current job using command `iskotrei` by Hendrickx 2002) (values from 6 to 78; mean 44).
- Estimand: ATT of nationality on occupational prestige, with resident aliens as the treatment group and Swiss nationals as the control group.
- Control variables: gender, age, and highest educational degree.
- Population restricted to people between 24 to 60 years old who are working.
- 2'308'006 individuals, of which 17.5% belong to the treatment group.
- Draw random samples ($N = 500$, 1000, or 5000) from population and compute various matching estimators.

Simulation

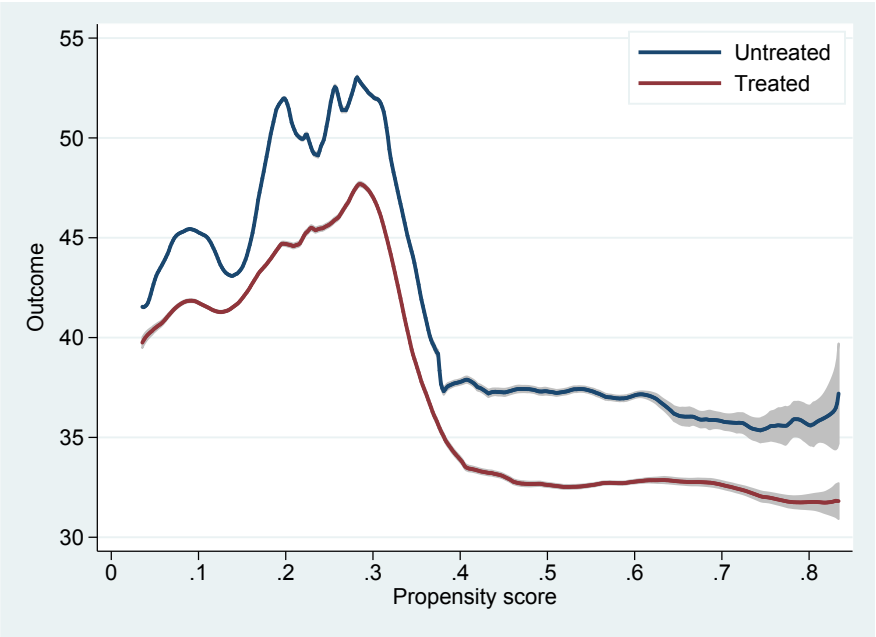
- Substantial differences between resident aliens and Swiss nationals on all three covariates.
- Propensity score in population (computed from fully stratified data)

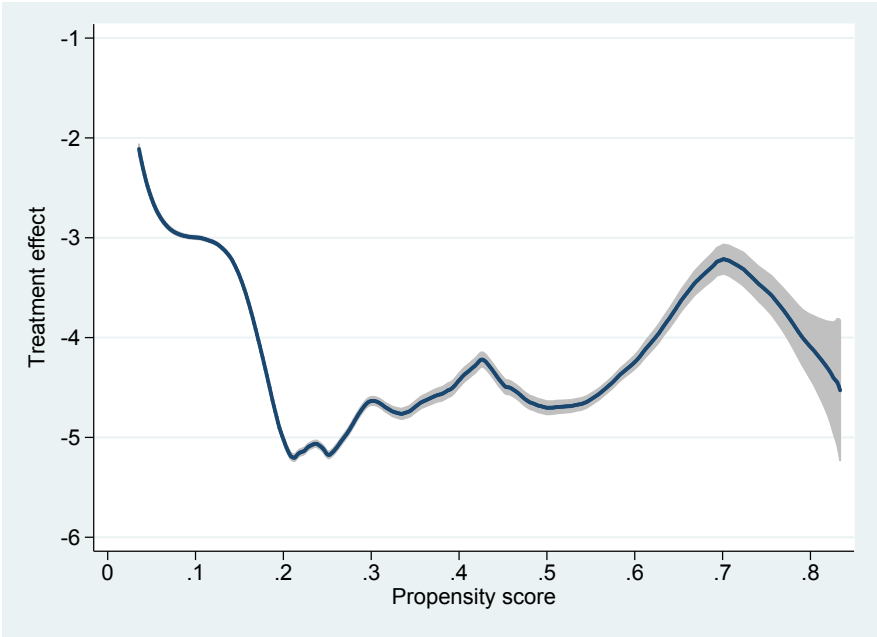


- McFadden $R^2 = 0.121$

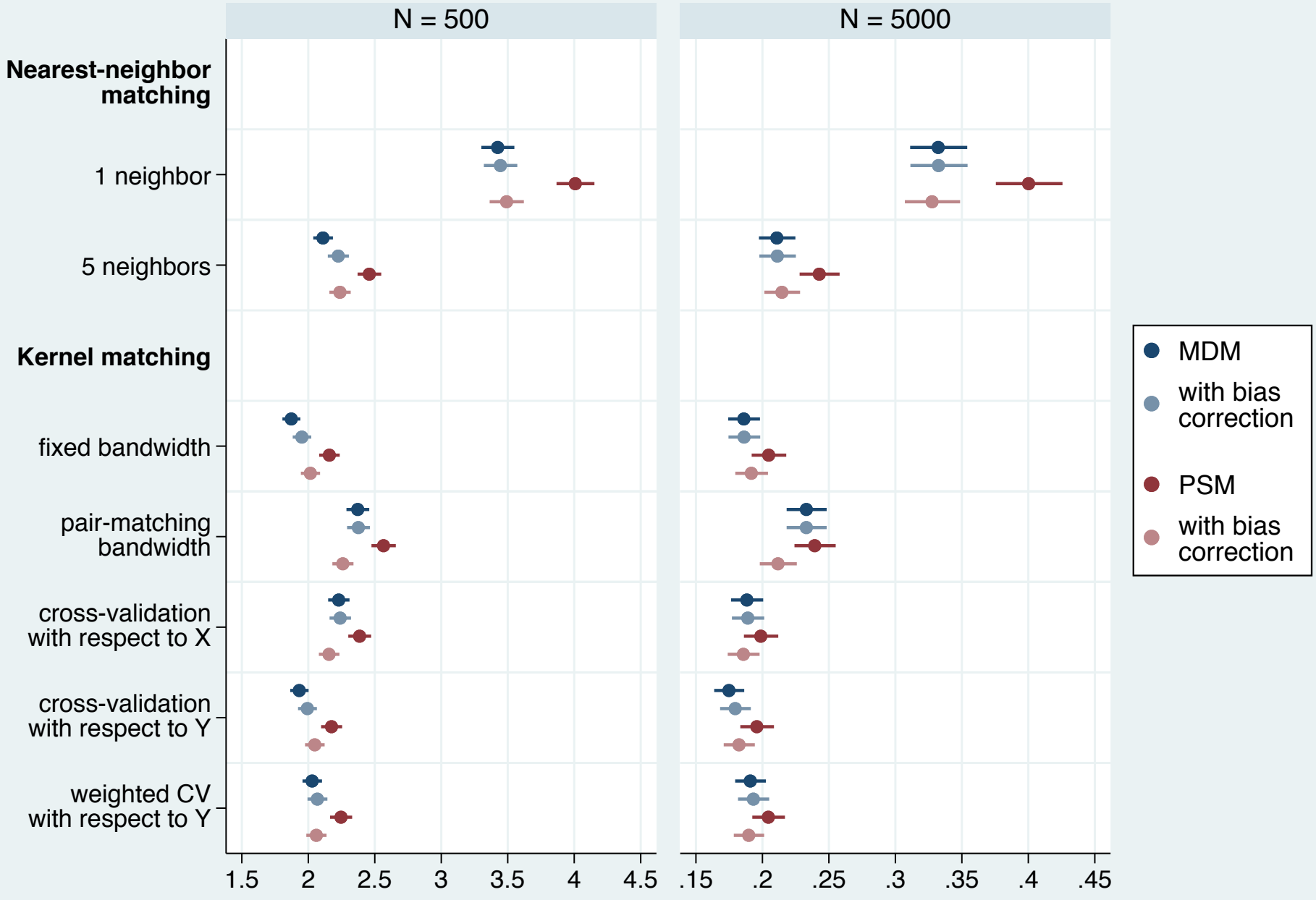
Simulation

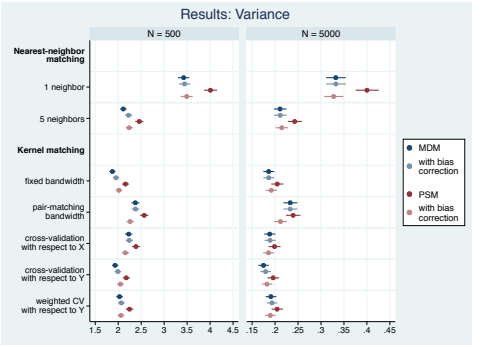
- Raw mean difference in occupational prestige (NATE): −4.79
- Population ATT (computed from fully stratified data): −3.96
- There is some treatment effect heterogeneity (ATE = −3.51, ATC = −3.41)





Results: Variance

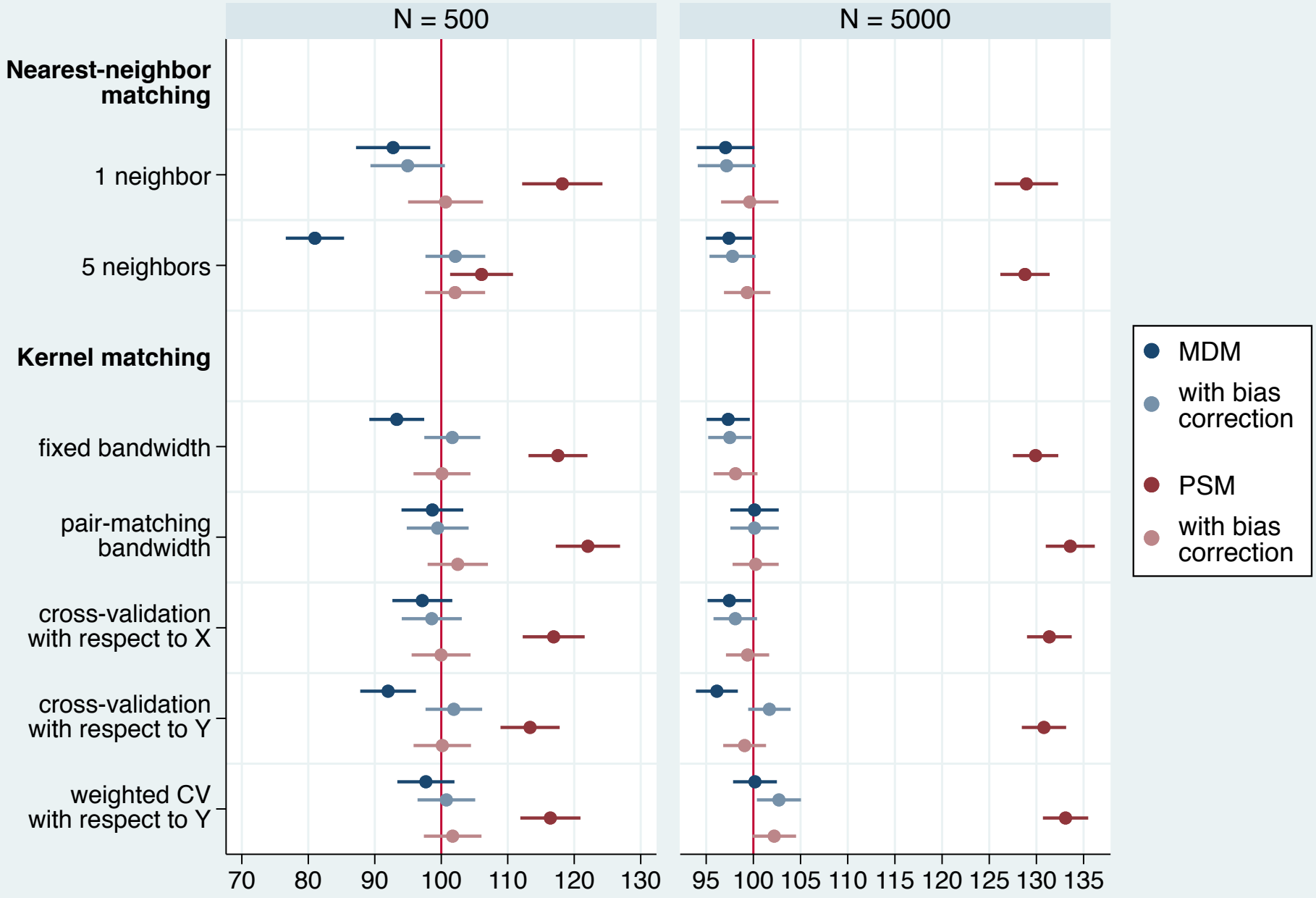


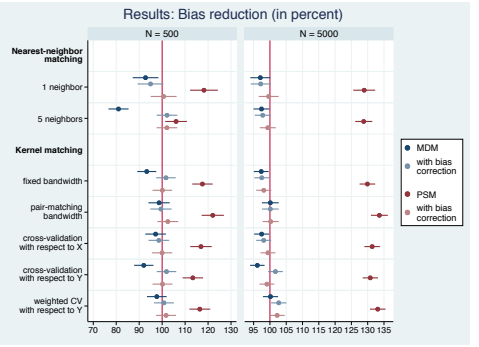


In this slide we can see that for the same algorithm PSM typically is somewhat less efficient than MDM, but that across algorithms PSM can also be much more efficient than MDM. For example, kernel matching PSM has a much smaller variance than 1-nearest-neighbor MDM. That is, the choice of algorithm matters much more than the choice between PSM and MDM.

For kernel matching the efficiency differences between PSM and MDM are only small; additional post-matching regression adjustment further reduces the differences.

Results: Bias reduction (in percent)

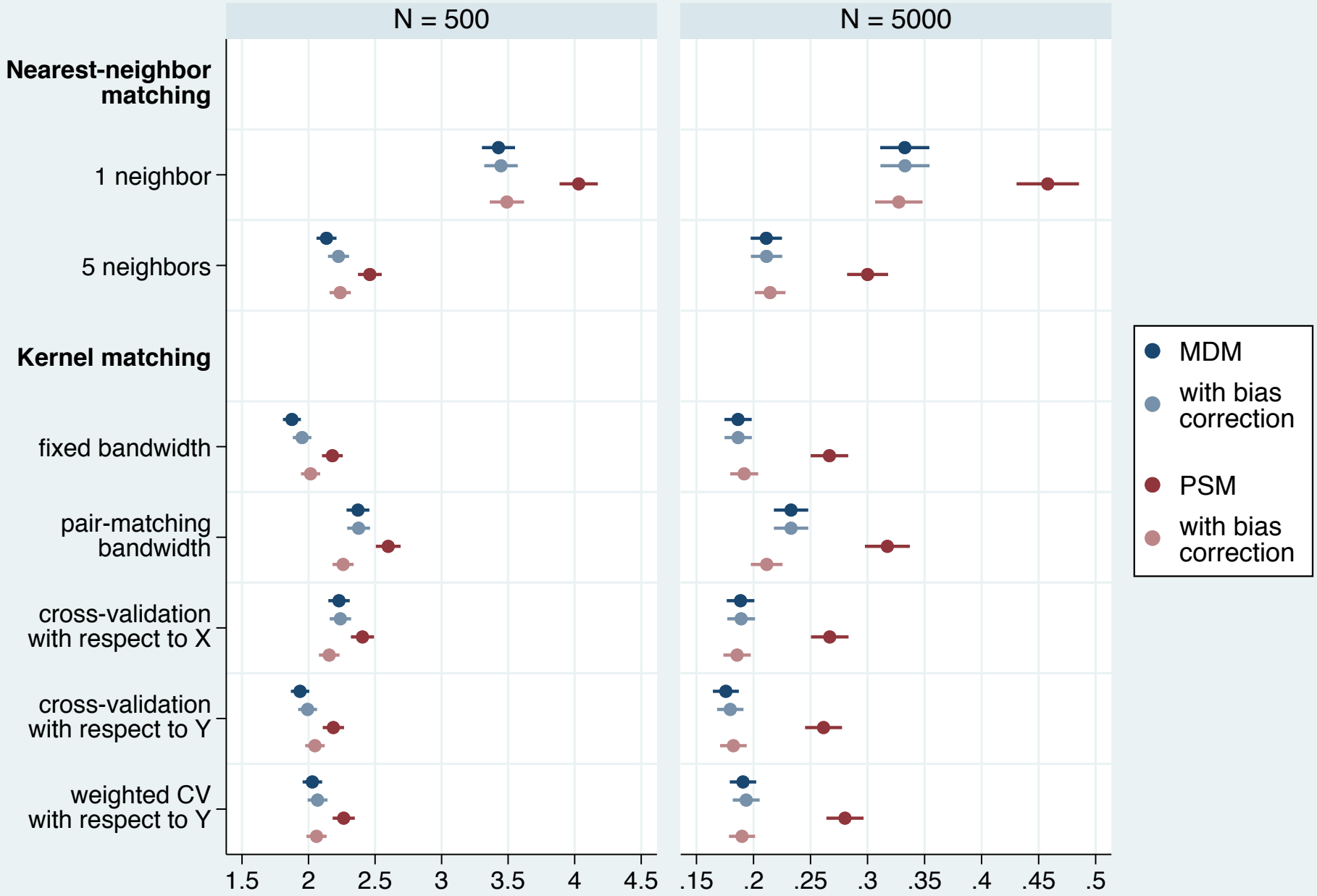




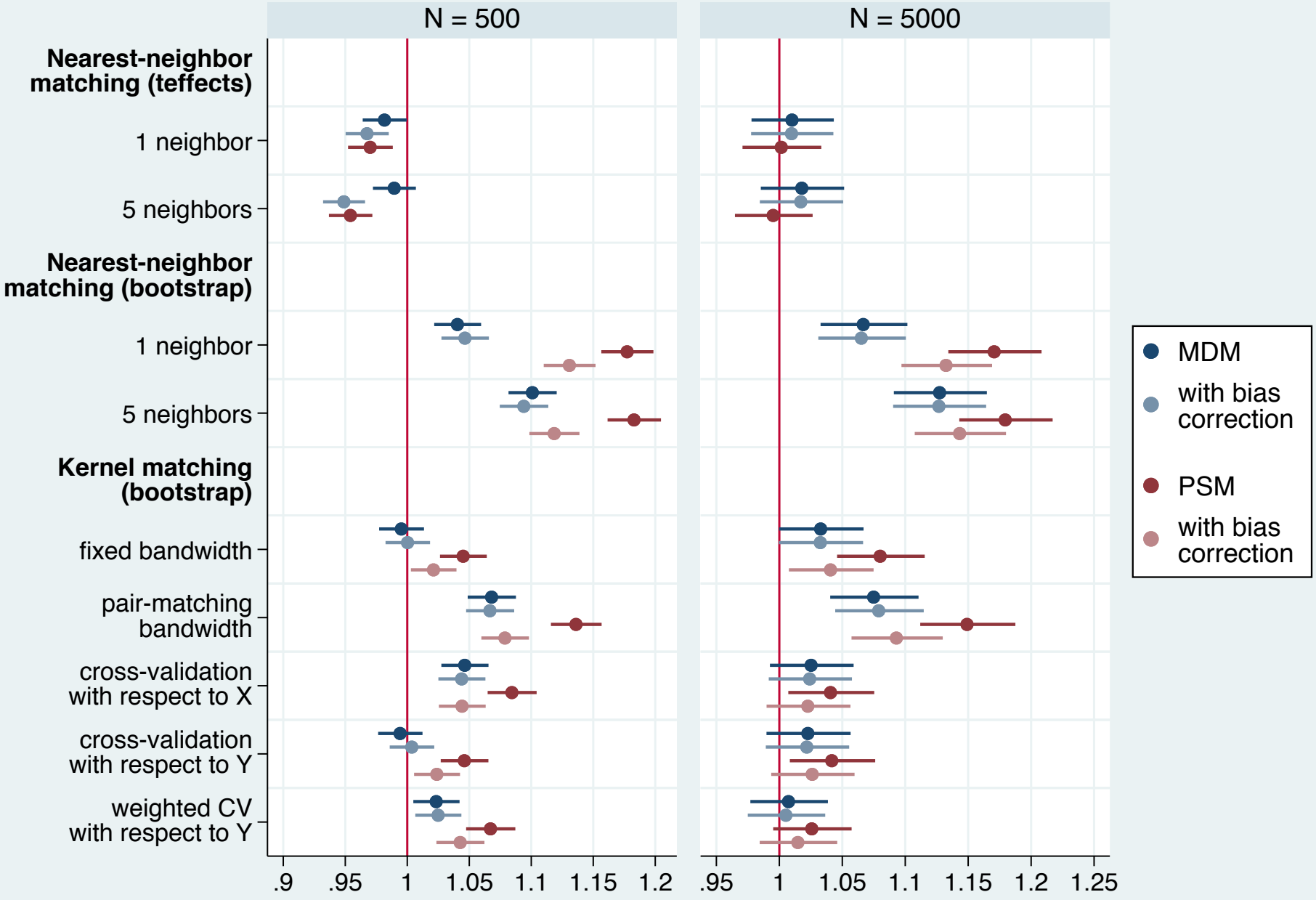
Here we see that PSM has a bias that does not vanish as the sample size increases. The reason is that the same propensity-score model specification is used for both sample sizes. The model is rather simple (linear effect of age, no interactions) and due to the specific pattern of the data (in particular, the sharp drop in the outcome variable after propensity score 0.3) small imprecisions can have substantial effects on the results. In practice, one would probably use a more refined specification in the large-sample situation, which would reduce bias.

The bias also vanishes once post-matching regression adjustment is applied.

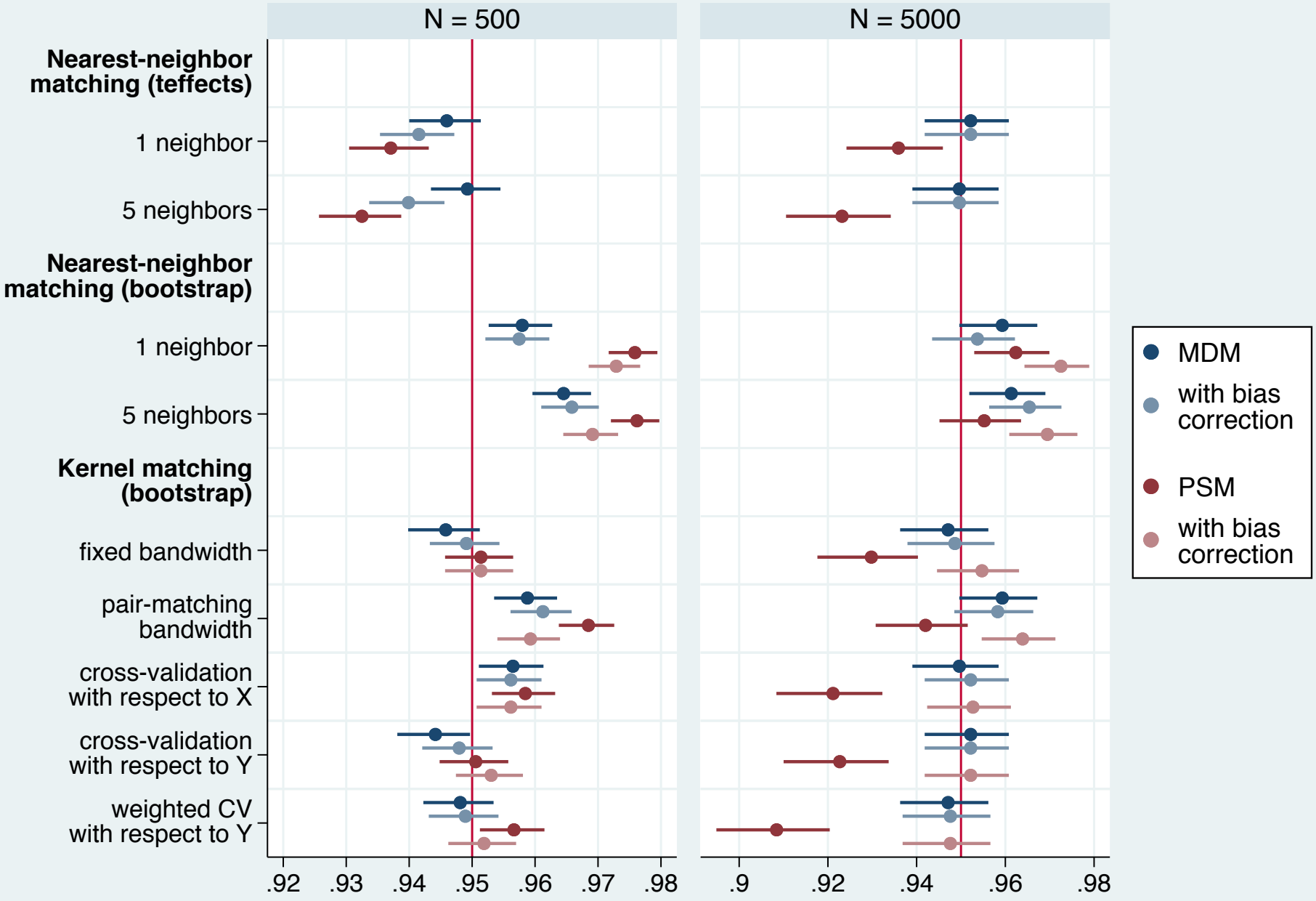
Results: Mean squared error



Results: Relative standard error



Results: Coverage of 95% CIs



Conclusions

- The arguments brought forward by King and Nielsen against Propensity Score Matching are valid, but they mostly apply to one specific form of PSM: pair matching (one-to-one matching without replacement).
- Other PSM matching algorithms perform much better because they are less affected by the random pruning problem.

Conclusions

- Overall, I agree that MDM has advantages over PSM, but it also has some disadvantages. In applied research the choice may not be that clear.
  - 👍 MDM leaves less scope for post-matching modeling decision biases.
  - 👍 Theoretical results (see, e.g., Frölich 2007) suggest that MDM will generally tend to outperform PSM in terms of efficiency (but differences are likely to be small).
  - 👍 Less restrictions in terms of possible post-matching analyses.
  - 👎 Choice of scaling matrix largely arbitrary; various suggestions in the literature (somewhat unclear, e.g., how categorical variables should be treated).
  - 👎 Computational complexity.

- One clear conclusion we can draw, however, is:

  Do not use propensity scores for pair matching!

  (But don't use pair matching anyhow.)

Conclusions

- Some conclusions from the simulation
  - For PSM, application of regression-adjustment seems like a great idea (reduction of bias and variance); for MDM the advantages of regression-adjustment are less clear.
  - Bootstrap standard error/confidence interval estimation seems to be mostly ok for kernel/ridge matching; this is in contrast to nearest-neighbor matching, where bootstrap standard errors are clearly biased.
- To do
  - Run some simulations comparable to the ones by King and Nielsen using various matching algorithms.

References I

- Cochran, W.G. 1968. The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies. *Biometrics* 24(2):295–313.
- Frölich, M. 2004. Finite-sample properties of propensity-score matching and weighting estimators. The Review of Economics and Statistics 86(1):77–90.
- Frölich, M. 2007. On the inefficiency of propensity score matching *AStA* 91:279–290.
- Hendrickx, J. 2002. ISKO: Stata module to recode 4 digit ISCO-88 occupational codes. Statistical Software Components S425802, Boston College Department of Economics.
- King, G., R. Nielsen. 2016. Why Propensity Scores Should Not Be Used for Matching. Working Paper. Available from http://j.mp/1sexgVw.
- Mill, J.S. 2002. *A System of Logic*. Reprinted from the 1981 edition (first published 1843). Honolulu, Hawaii: University Press of the Pacific.

References II

- Neyman, J. 1990[1923]. On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9 (Translated and edited by D.M. Dabrowska and T.P. Speed from the Polish original). *Statistical Science* 5(4):465–472.
- Rosenbaum, P.R., D.B. Rubin. 1983. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* 70:41–55.
- Rubin, D.B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66(5):688–701.
- Rubin, D.B. 1990. Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies. *Statistical Science* 5(4):472–480.