Covariate Balancing Propensity Score (CBPS)

Kosuke Imai

Princeton University

October 27 and 29, 2014

Seminars at Laval University and University of Montreal

Joint work with Marc Ratkovic and Christian Fong

References

1. Main Paper:
   "Covariate Balancing Propensity Score." (2014). *Journal of the Royal Statistical Society, Series B*, Vol. 76, No. 1, pp. 243–263.
2. Extensions:
   1. Non-binary treatments: "Covariate Balancing Propensity Score for General Treatment Regimes." working paper
   2. Longitudinal data: "Robust Estimation of Inverse Probability Weights for Marginal Structural Models." *Journal of the American Statistical Association*, Forthcoming.
3. Software: *CBPS: R Package for Covariate Balancing Propensity Score* available for download at the CRAN

These and other related materials available at
`http://imai.princeton.edu`

Motivation

- Causal inference is a central goal of scientific research
- Randomized experiments are not always possible
  $\rightsquigarrow$ Causal inference in observational studies
- Experiments often lack external validity
  $\rightsquigarrow$ Need to generalize experimental results to a target population
- Importance of statistical methods to adjust for confounding factors
- Distinction between observed and unobserved confounders

Overview of the Talk

1. **Review:** Propensity score
    - propensity score is a covariate balancing score
    - matching and weighting methods
2. **Problem:** Propensity score tautology
    - sensitivity to model misspecification
    - adhoc specification searches
3. **Solution:** **Covariate balancing propensity score (CBPS)**
    - Estimate propensity score so that covariate balance is optimized
4. **Evidence:** Reanalysis of two prominent critiques
    - Improved performance of propensity score weighting and matching
5. **Software:** R package `CBPS`
6. **Extension:** Non-binary treatments

Propensity Score

- Setup:
  - $T_i \in \{0, 1\}$: binary treatment
  - X_i: pre-treatment covariates
  - $(Y_i(1), Y_i(0))$: potential outcomes
  - $Y_i = Y_i(T_i)$: observed outcomes
- Definition: conditional probability of treatment assignment

$$\pi(X_i) \; = \; \Pr(T_i = 1 \mid X_i)$$

- Balancing property (without assumption):

$$T_i \perp\!\!\!\perp X_i \mid \pi(X_i)$$

Rosenbaum and Rubin (1983)

- Assumptions:
  1. Overlap:
  $$0 < \pi(X_i) < 1$$
  2. Unconfoundedness:
  $$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i \mid X_i$$

- Propensity score as a dimension reduction tool:
$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i \mid \pi(X_i)$$

- But, propensity score must be estimated (more on this later)

Use of Propensity Score for Causal Inference

- Matching
- Subclassification
- Weighting (Horvitz-Thompson):

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{T_i Y_i}{\hat{\pi}(X_i)} - \frac{(1-T_i)Y_i}{1-\hat{\pi}(X_i)}\right\}$$

  where weights are often normalized

- Doubly-robust estimators (Robins *et al.*):

$$\frac{1}{n}\sum_{i=1}^{n}\left[\left\{\hat{\mu}(1,X_i) + \frac{T_i(Y_i - \hat{\mu}(1,X_i))}{\hat{\pi}(X_i)}\right\} - \left\{\hat{\mu}(0,X_i) + \frac{(1-T_i)(Y_i - \hat{\mu}(0,X_i))}{1-\hat{\pi}(X_i)}\right\}\right]$$

- They have become standard tools for applied researchers

Propensity Score Tautology

- Propensity score is unknown
- Dimension reduction is purely theoretical: must model T_i given X_i

- Diagnostics: covariate balance checking
- In practice, adhoc specification searches are conducted
- Misspecification is possible especially for non-binary treatments

- Theory (Rubin *et al.*): ellipsoidal covariate distributions
  $\rightsquigarrow$ equal percent bias reduction
- Skewed covariates are common in applied settings

- Propensity score methods can be sensitive to misspecification

Kang and Schafer (2007, *Statistical Science*)

- Simulation study: the deteriorating performance of propensity score weighting methods when the model is misspecified
- Setup:
  - 4 covariates X_i^*: all are *i.i.d.* standard normal
  - Outcome model: linear model
  - Propensity score model: logistic model with linear predictors
  - Misspecification induced by measurement error:
    - $X_{i1} = \exp(X_{i1}^*/2)$
    - $X_{i2} = X_{i2}^*/(1 + \exp(X_{1i}^*) + 10)$
    - $X_{i3} = (X_{i1}^* X_{i3}^*/25 + 0.6)^3$
    - $X_{i4} = (X_{i1}^* + X_{i4}^* + 20)^2$
- Weighting estimators to be evaluated:
  1. Horvitz-Thompson
  2. Inverse-probability weighting with normalized weights
  3. Weighted least squares regression
  4. Doubly-robust least squares regression

Weighting Estimators Do Fine If the Model is Correct

		Bias		RMSE	
Sample size	Estimator	GLM	True	GLM	True
(1) Both models correct					
$n = 200$	HT	0.33	1.19	12.61	23.93
	IPW	-0.13	-0.13	3.98	5.03
	WLS	-0.04	-0.04	2.58	2.58
	DR	-0.04	-0.04	2.58	2.58
$n = 1000$	HT	0.01	-0.18	4.92	10.47
	IPW	0.01	-0.05	1.75	2.22
	WLS	0.01	0.01	1.14	1.14
	DR	0.01	0.01	1.14	1.14
(2) Propensity score model correct					
$n = 200$	HT	-0.05	-0.14	14.39	24.28
	IPW	-0.13	-0.18	4.08	4.97
	WLS	0.04	0.04	2.51	2.51
	DR	0.04	0.04	2.51	2.51
$n = 1000$	HT	-0.02	0.29	4.85	10.62
	IPW	0.02	-0.03	1.75	2.27
	WLS	0.04	0.04	1.14	1.14
	DR	0.04	0.04	1.14	1.14

Weighting Estimators are Sensitive to Misspecification

		Bias		RMSE	
Sample size	Estimator	GLM	True	GLM	True
(3) Outcome model correct					
$n = 200$	HT	24.25	−0.18	194.58	23.24
	IPW	1.70	−0.26	9.75	4.93
	WLS	−2.29	0.41	4.03	3.31
	DR	−0.08	−0.10	2.67	2.58
$n = 1000$	HT	41.14	−0.23	238.14	10.42
	IPW	4.93	−0.02	11.44	2.21
	WLS	−2.94	0.20	3.29	1.47
	DR	0.02	0.01	1.89	1.13
(4) Both models incorrect					
$n = 200$	HT	30.32	−0.38	266.30	23.86
	IPW	1.93	−0.09	10.50	5.08
	WLS	−2.13	0.55	3.87	3.29
	DR	−7.46	0.37	50.30	3.74
$n = 1000$	HT	101.47	0.01	2371.18	10.53
	IPW	5.16	0.02	12.71	2.25
	WLS	−2.95	0.37	3.30	1.47
	DR	−48.66	0.08	1370.91	1.81

Smith and Todd (2005, *J. of Econometrics*)

- LaLonde (1986; *Amer. Econ. Rev.*):
  - Randomized evaluation of a job training program
  - Replace experimental control group with another non-treated group
  - Current Population Survey and Panel Study for Income Dynamics
  - Many evaluation estimators didn't recover experimental benchmark
- Dehejia and Wahba (1999; *J. of Amer. Stat. Assoc.*):
  - Apply propensity score matching
  - Estimates are close to the experimental benchmark
- Smith and Todd (2005):
  - Dehejia & Wahba (DW)'s results are sensitive to model specification
  - They are also sensitive to the selection of comparison sample

Propensity Score Matching Fails Miserably

- One of the most difficult scenarios identified by Smith and Todd:
  - LaLonde experimental sample rather than DW sample
  - Experimental estimate: \$886 (s.e. $=$ 488)
  - PSID sample rather than CPS sample

- Evaluation bias:
  - Conditional probability of being in the experimental sample
  - Comparison between experimental control group and PSID sample
  - "True" estimate $= 0$
  - Logistic regression for propensity score
  - One-to-one nearest neighbor matching with replacement

Propensity score model	Estimates
Linear	-835
	(886)
Quadratic	-1620
	(1003)
Smith and Todd (2005)	-1910
	(1004)

Covariate Balancing Propensity Score

- Idea: Estimate the propensity score such that covariate balance is optimized

- Covariate balancing condition:

$$\mathbb{E}\left\{\frac{T_i\widetilde{X}_i}{\pi_\beta(X_i)} - \frac{(1-T_i)\widetilde{X}_i}{1-\pi_\beta(X_i)}\right\} = 0$$

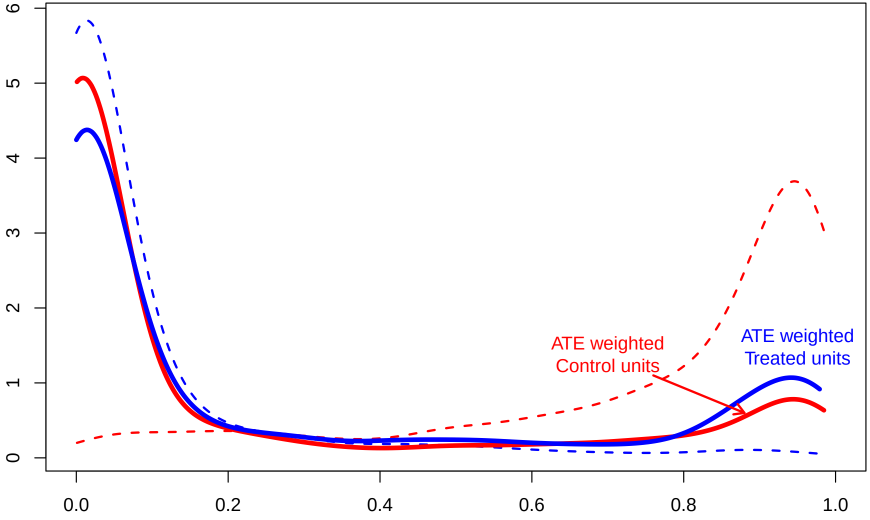
  where $\widetilde{X}_i = f(X_i)$ is any vector-valued function

- Score condition from maximum likelihood:

$$\mathbb{E}\left\{\frac{T_i\pi'_\beta(X_i)}{\pi_\beta(X_i)} - \frac{(1-T_i)\pi'_\beta(X_i)}{1-\pi_\beta(X_i)}\right\} = 0$$

Weighting to Balance Covariates

- Balancing condition: $\mathbb{E}\left\{\frac{T_i X_i}{\pi_\beta(X_i)} - \frac{(1-T_i)X_i}{1-\pi_\beta(X_i)}\right\} = 0$



Generalized Method of Moments (GMM) Framework

- Just-identified CBPS: covariate balancing conditions alone
- Over-identified CBPS: combine them with score conditions

- GMM (Hansen 1982):

$$\hat{\beta}_{\text{GMM}} = \underset{\beta \in \Theta}{\operatorname{argmin}} \ \bar{g}_\beta(T,X)^\top \Sigma_\beta(T,X)^{-1} \bar{g}_\beta(T,X)$$

where

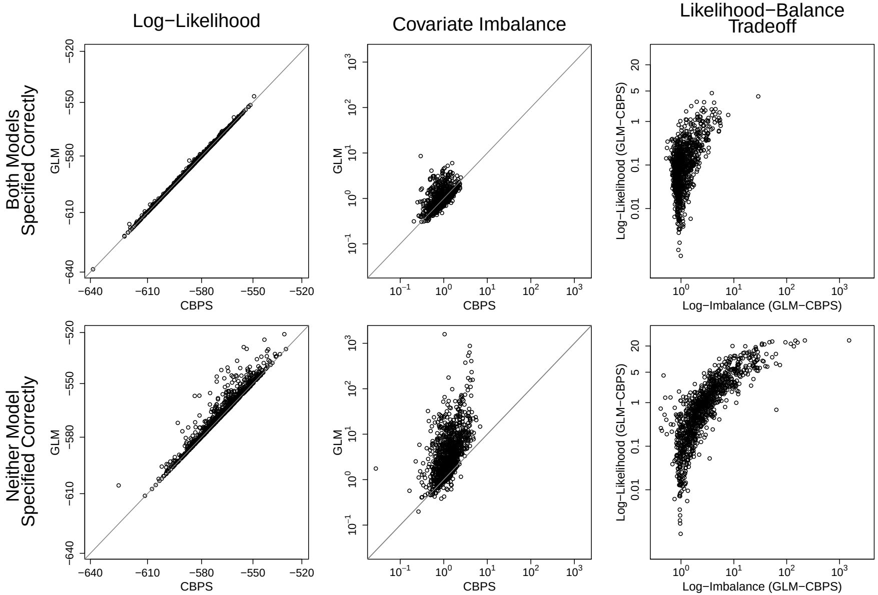
$$\bar{g}_\beta(T,X) = \frac{1}{N} \sum_{i=1}^{N} \underbrace{\begin{pmatrix} \text{score condition} \\ \text{balancing condition} \end{pmatrix}}_{g_\beta(T_i, X_i)}$$

- "Continuous updating" GMM estimator for Σ

CBPS Makes Weighting Methods Work Better

		Bias				RMSE			
	Estimator	logit	CBPS1	CBPS2	True	logit	CBPS1	CBPS2	True
(3) Outcome model correct									
$n = 200$	HT	24.25	1.09	−5.42	−0.18	194.58	5.04	10.71	23.24
	IPW	1.70	−1.37	−2.84	−0.26	9.75	3.42	4.74	4.93
	WLS	−2.29	−2.37	−2.19	0.41	4.03	4.06	3.96	3.31
	DR	−0.08	−0.10	−0.10	−0.10	2.67	2.58	2.58	2.58
$n = 1000$	HT	41.14	−2.02	2.08	−0.23	238.14	2.97	6.65	10.42
	IPW	4.93	−1.39	−0.82	−0.02	11.44	2.01	2.26	2.21
	WLS	−2.94	−2.99	−2.95	0.20	3.29	3.37	3.33	1.47
	DR	0.02	0.01	0.01	0.01	1.89	1.13	1.13	1.13
(4) Both models incorrect									
$n = 200$	HT	30.32	1.27	−5.31	−0.38	266.30	5.20	10.62	23.86
	IPW	1.93	−1.26	−2.77	−0.09	10.50	3.37	4.67	5.08
	WLS	−2.13	−2.20	−2.04	0.55	3.87	3.91	3.81	3.29
	DR	−7.46	−2.59	−2.13	0.37	50.30	4.27	3.99	3.74
$n = 1000$	HT	101.47	−2.05	1.90	0.01	2371.18	3.02	6.75	10.53
	IPW	5.16	−1.44	−0.92	0.02	12.71	2.06	2.39	2.25
	WLS	−2.95	−3.01	−2.98	0.19	3.30	3.40	3.36	1.47
	DR	−48.66	−3.59	−3.79	0.08	1370.91	4.02	4.25	1.81

CBPS Sacrifices Likelihood for Better Balance



Revisiting Smith and Todd (2005)

- Evaluation bias: "true" bias $= 0$
- CBPS improves propensity score matching across specifications and matching methods
- However, specification test rejects the null

	1-to-1 Nearest Neighbor			Optimal 1-to-N Nearest Neighbor		
Specification	GLM	CBPS1	CBPS2	GLM	CBPS1	CBPS2
Linear	−1209.15	−654.79	−505.15	−1209.15	−654.79	−130.84
	(1426.44)	(1247.55)	(1335.47)	(1426.44)	(1247.55)	(1335.47)
Quadratic	−1439.14	−955.30	−216.73	−1234.33	−175.92	−658.61
	(1299.05)	(1496.27)	(1285.28)	(1074.88)	(943.34)	(1041.47)
Smith & Todd	−1437.69	−820.89	−640.99	−1229.81	−826.53	−464.06
	(1256.84)	(1229.63)	(1757.09)	(1044.15)	(1179.73)	(1130.73)

Comparison with the Experimental Benchmark

- LaLonde, Dehejia and Wahba, and others did this comparison
- Experimental estimate: \$866 (s.e. $=$ 488)
- LaLonde$+$PSID pose a challenge: e.g., GenMatch $-$571 (1108)

	1-to-1 Nearest Neighbor			Optimal 1-to-N Nearest Neighbor		
Specification	GLM	CBPS1	CBPS2	GLM	CBPS1	CBPS2
Linear	$-$304.92	423.30	183.67	$-$211.07	423.30	138.20
	(1437.02)	(1295.19)	(1240.79)	(1201.49)	(1110.26)	(1161.91)
Quadratic	$-$922.16	239.46	1093.13	$-$715.54	307.51	185.57
	(1382.38)	(1284.13)	(1567.33)	(1145.82)	(1158.06)	(1247.99)
Smith & Todd	$-$734.49	$-$269.07	423.76	$-$439.54	$-$617.68	690.09
	(1424.57)	(1711.66)	(1404.15)	(1259.28)	(1438.86)	(1288.68)

Software: R Package CBPS

```
## upload the package
library("CBPS")
## load the LaLonde data
data(LaLonde)
## Estimate ATT weights via CBPS
fit <- CBPS(treat ~ age + educ + re75 + re74 +
                    I(re75==0) + I(re74==0),
            data = LaLonde, ATT = TRUE)
summary(fit)
## matching via MatchIt
library(MatchIt)
## one to one nearest neighbor with replacement
m.out <- matchit(treat ~ 1, distance = fitted(fit),
                 method = "nearest", data = LaLonde,
                 replace = TRUE)
summary(m.out)
```

Extensions to Other Causal Inference Settings

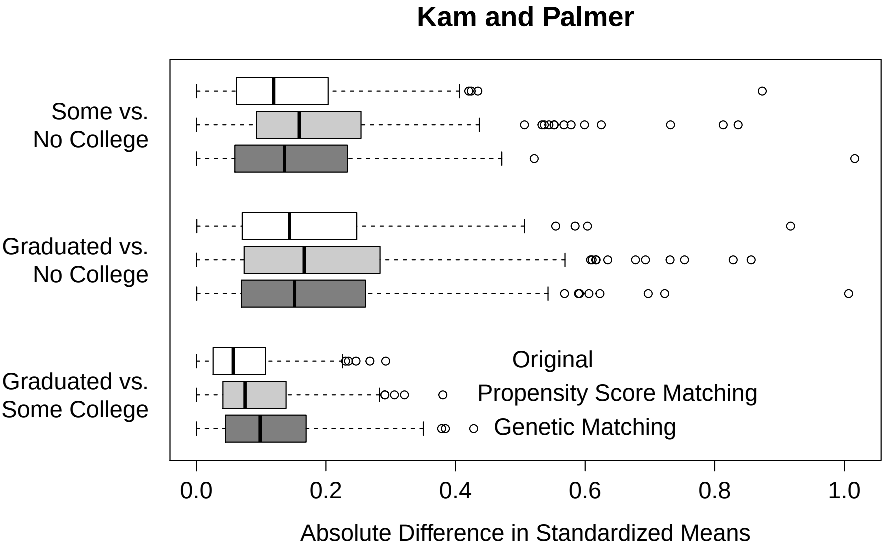
- Propensity score methods are widely applicable
- Thus, CBPS is also widely applicable
- Extensions of propensity score to general treatment regimes
  - Weighting (e.g., Imbens, 2000; Robins et al., 2000)
  - Subclassification (e.g., Imai & van Dyk, 2004)
  - Regression (e.g., Hirano & Imbens, 2004)
- But, propensity score is mostly applied to binary treatment
  - All existing methods assume correctly estimated propensity score
  - No reliable methods to estimate generalized propensity score
  - Harder to check balance across a non-binary treatment
  - Many researchers dichotomize the treatment
- Estimate the generalized propensity score such that covariate is balanced across *all* treatment groups

Two Motivating Examples

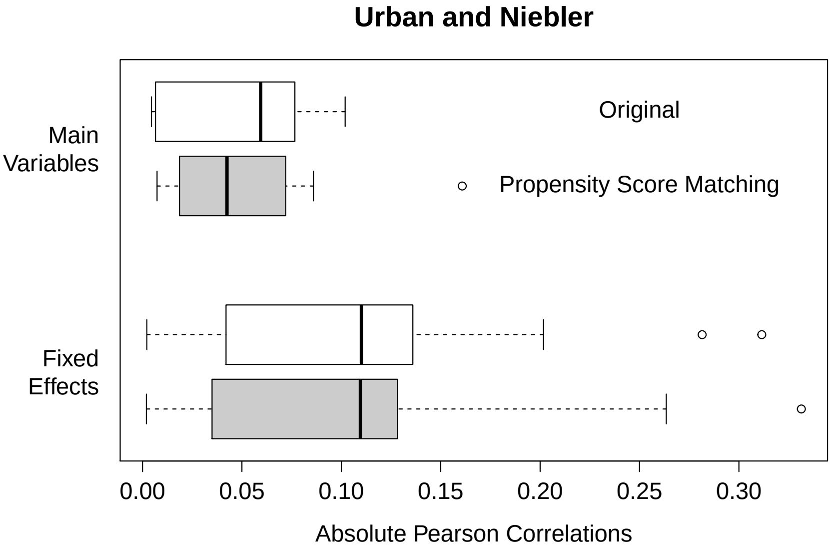
1. Effect of education on political participation
   - Education is assumed to play a key role in political participation
   - T_i: 3 education levels (graduated from college, attended college but not graduated, no college)
   - Original analysis $\rightsquigarrow$ dichotomization (some college vs. no college)
   - Propensity score matching
   - Critics employ different matching methods

2. Effect of advertisements on campaign contributions
   - Do TV advertisements increase campaign contributions?
   - T_i: Number of advertisements aired in each zip code
   - ranges from 0 to 22,379 advertisements
   - Original analysis $\rightsquigarrow$ dichotomization (over 1000 vs. less than 1000)
   - Propensity score matching followed by linear regression with an original treatment variable

Covariates are Not Balanced for Original Treatment



Covariates are Not Balanced for Original Treatment



CBPS for a Multi-valued Treatment

- Consider a 3 treatment value case as in our motivating example
- Generalized propensity score:
  1. $\pi^1_\beta(X_i) = \Pr(Y_i = 1 \mid X_i)$
  2. $\pi^2_\beta(X_i) = \Pr(Y_i = 2 \mid X_i)$
- Standard estimation: multinomial logit regression
- Sample balance conditions with orthogonalized contrasts:

$$\bar{g}_\beta(T, X) = \frac{1}{N} \sum_{i=1}^{N} \begin{pmatrix} \frac{\mathbf{1}\{T_i=1\}}{\pi^1_\beta(X_i)} - \frac{\mathbf{1}\{T_i=2\}}{\pi^2_\beta(X_i)} \\ \frac{\mathbf{1}\{T_i=1\}}{\pi^1_\beta(X_i)} + \frac{\mathbf{1}\{T_i=2\}}{\pi^2_\beta(X_i)} - 2\frac{\mathbf{1}\{T_i=0\}}{1-\pi^1_\beta(X_i)-\pi^2_\beta(X_i)} \end{pmatrix} X_i$$

- GMM estimation as before

CBPS for a Continuous Treatment

- Generalized propensity score: $f(T_i \mid X_i)$
- Standard model: linear regression
- The stabilized weights (Robins *et al*.):

$$\frac{f(T_i)}{f(T_i \mid X_i)}$$

- Covariate balancing condition:

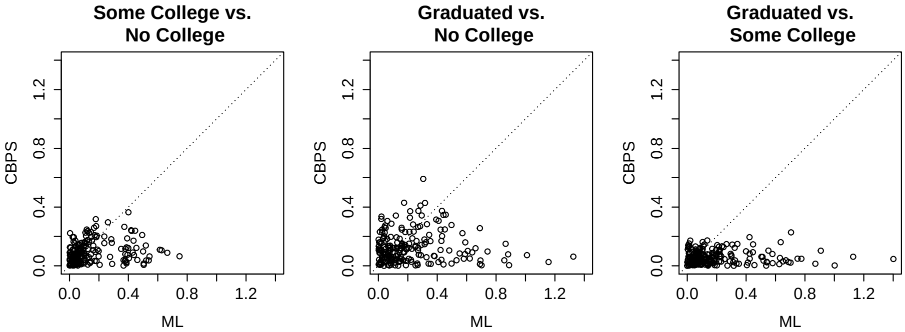
$$\begin{aligned}\mathbb{E}\left(\frac{f(T_i^*)}{f(T_i^* \mid X_i^*)} T_i^* X_i^*\right) &= \int \left\{\int \frac{f(T_i^*)}{f(T_i^* \mid X_i^*)} T_i^* dF(T_i^* \mid X_i^*)\right\} X_i^* dF(X_i^*) \\ &= \mathbb{E}(T_i^*)\mathbb{E}(X_i^*) = 0.\end{aligned}$$

  where T_i^* and X_i^* are centered versions of T_i and X_i

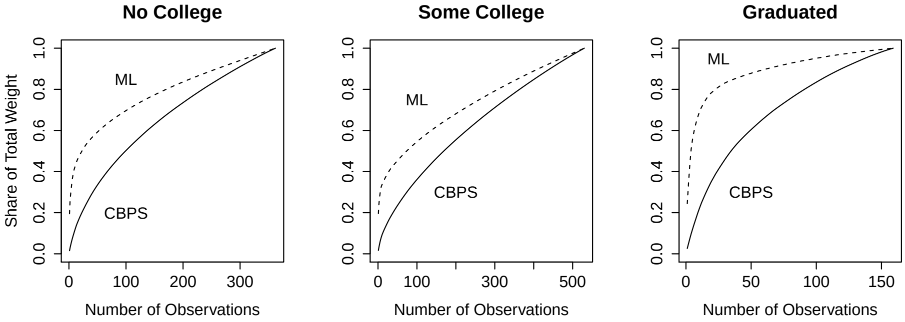
- Again, estimate the generalized propensity score such that covariate balance is optimized

Back to the Education Example: CBPS vs. ML

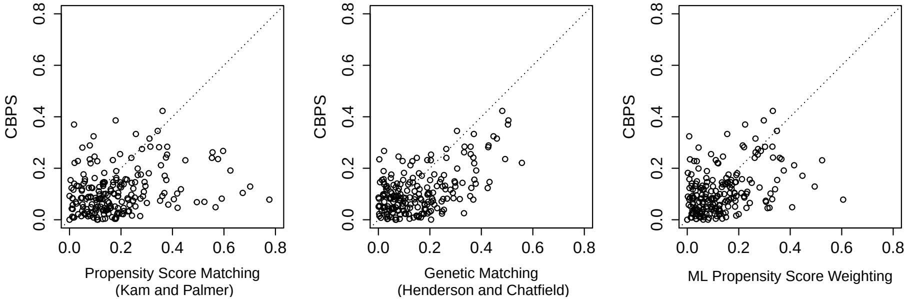
- CBPS achieves better covariate balance



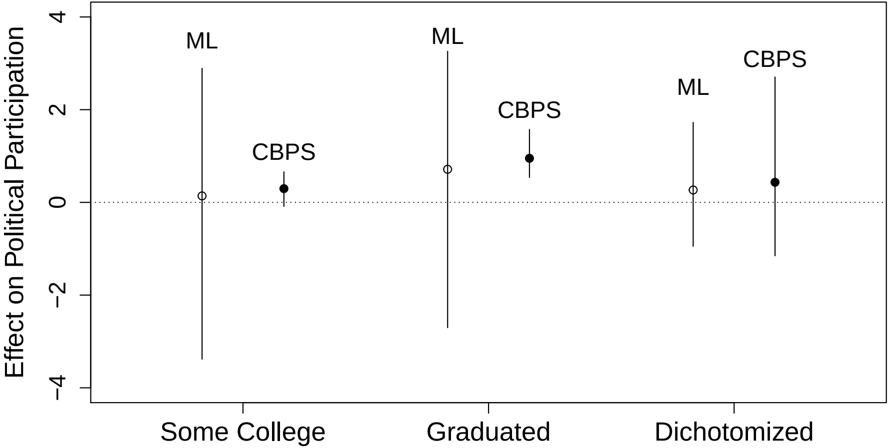
CBPS Avoids Extremely Large Weights



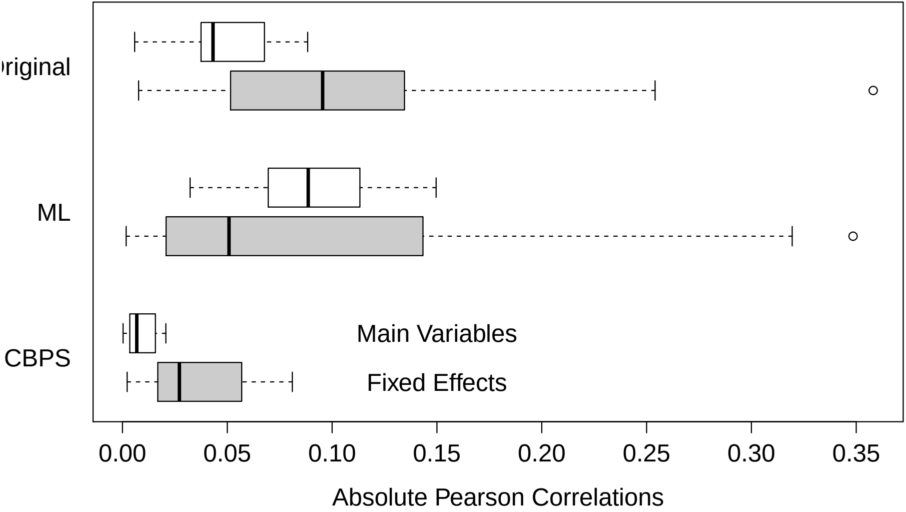
CBPS Balances Well for a Dichotomized Treatment



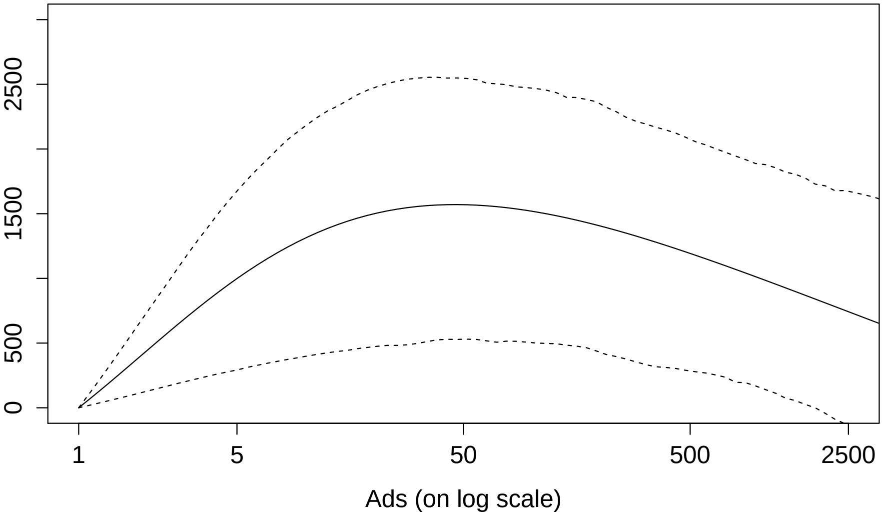
Empirical Results: Graduation Matters, Efficiency Gain



Onto the Advertisement Example



Empirical Finding: Some Effect of Advertisement



Concluding Remarks

- Numerous advances in generalizing propensity score methods to non-binary treatments
- Yet, many applied researchers don't use these methods and dichotomize non-binary treatments

- We offer a simple method to improve the estimation of propensity score for general treatment regimes
- Open-source R package: CBPS: Covariate Balancing Propensity Score available at CRAN

- Ongoing extensions:
    1. nonparametric estimation via empirical likelihood
    2. generalizing instrumental variables estimates
    3. spatial treatments