Covariate Balancing Propensity Score

Kosuke Imai

Princeton University

Joint work with Marc Ratkovic

November 3, 2012
Experiments in Governance and Politics Conference

Motivation

- Causal inference is a central goal of scientific research
- Randomized experiments are not always possible
  $\Longrightarrow$ Causal inference in observational studies
- Randomized experiments often lack external validity
  $\Longrightarrow$ Need to generalize experimental results
- Instrumental variables estimates are only applicable to compliers
  $\Longrightarrow$ Need to generalize to non-compliers
- **Common goal**: statistically adjust for confounding factors

Overview of the Talk

1. **Review:** Propensity score
   - conditional probability of treatment assignment
   - propensity score is a balancing score
   - matching and weighting methods
2. **Problem:** Propensity score tautology
   - sensitivity to model misspecification
   - adhoc specification searches
3. **Solution:** **Covariate balancing propensity score**
   - 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. **Extension:** Generalizing experimental estimates

Propensity Score

- Notation:
  - $T_i \in \{0, 1\}$: binary treatment
  - X_i: pre-treatment covariates
  - $(Y_i(1), Y_i(0))$: potential outcomes
  - $Y_i = Y_i(T_i)$: observed outcomes

- Dual characteristics of propensity score (without assumption):
  1. Predicts treatment assignment:

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

  2. Balances covariates:

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

Rosenbaum and Rubin (1983)

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

- The main result: Propensity score as a dimension reduction tool

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i \mid \pi(X_i)$$

Propensity Score Tautology

- Propensity score is unknown and must be estimated
  - Dimension reduction is purely theoretical: must model T_i given X_i
  - Diagnostics: covariate balance checking
- In theory: ellipsoidal covariate distributions
  $\Longrightarrow$ equal percent bias reduction
- In practice: skewed covariates and adhoc specification searches
- Model misspecification is always possible
- Propensity score methods can be sensitive to misspecification
- Tautology: propensity score methods only work when they work

Covariate Balancing Propensity Score (CBPS)

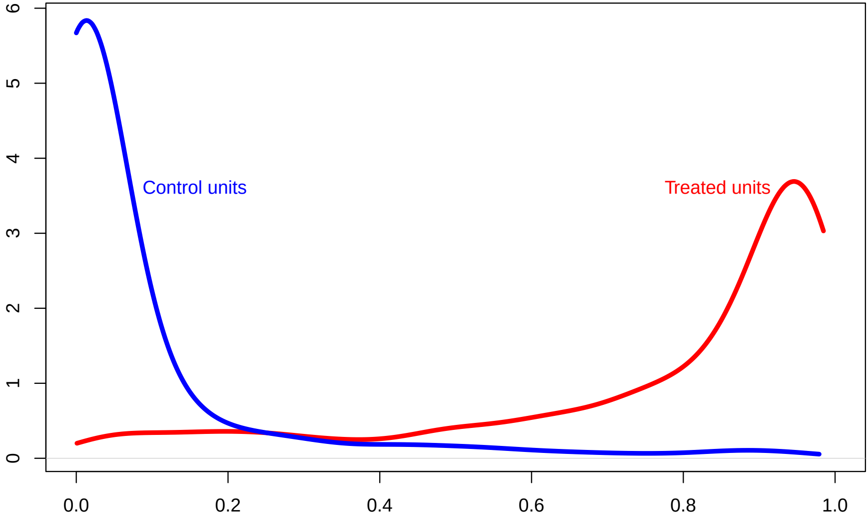
- Idea: take advantage of propensity score tautology
- Recall the dual characteristics of propensity score
  1. Predicts treatment assignment
  2. Balances covariates
- Implied moment conditions:
  1. Score condition: sets the first derivative of the log-likelihood to zero

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

  2. Balancing condition: sets weighted difference in means between treated an untreated observations to zero
- Score condition is a balancing condition
- CBPS uses the same propensity score model (e.g., logistic regression) but estimates it to best satisfy the above conditions

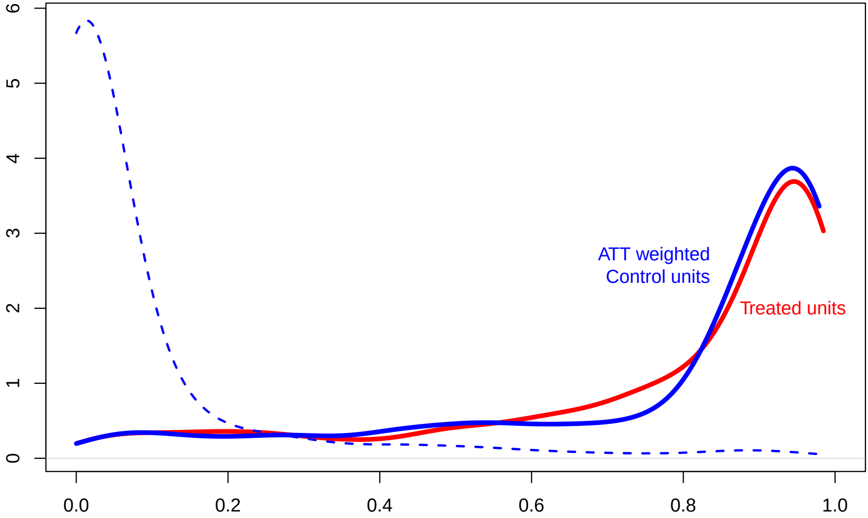
Weighting Control Group to Balance Covariates

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



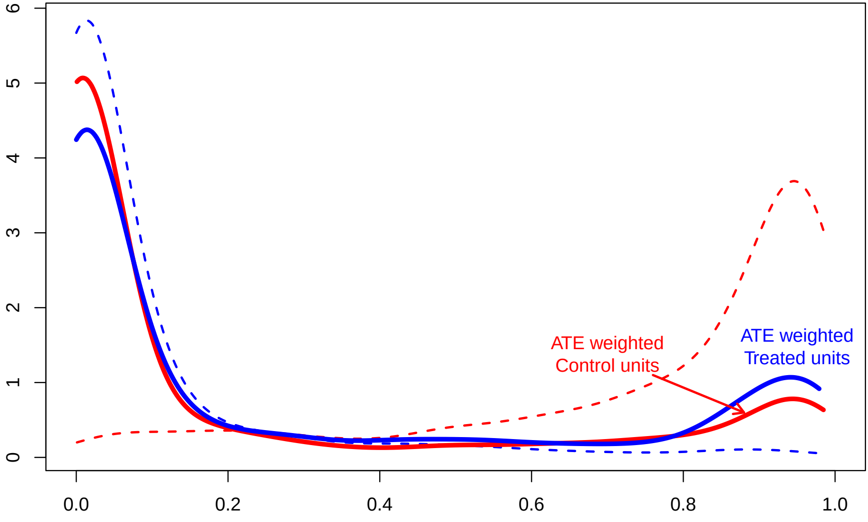
Weighting Control Group to Balance Covariates

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



Weighting Both Groups 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) Estimation

- Over-identification: more moment conditions than parameters
- GMM (Hansen 1982):

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

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 with the following Σ:

$$\Sigma_\beta(T, X) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}(g_\beta(T_i, X_i) g_\beta(T_i, X_i)^\top \mid X_i)$$

- Newton-type optimization algorithm with MLE as starting values

Specification Test and Optimal Matching

- CBPS is overidentified
- Specification test based on Hansen's J-statistic:

$$J = n\, \bar{g}_\beta(T, X)^\top \Sigma_\beta(T, X)^{-1} \bar{g}_\beta(T, X) \sim \chi^2_k$$

  where k is the number of moment conditions

- Can also be used to conduct "optimal" 1-to-N matching

Kang and Schafer (2007, *Statistical Science*)

- Simulation study: the deteriorating performance of propensity score weighting methods when the model is misspecified
- Can the CBPS save propensity score weighting methods?

- 4 covariates X_i^*: all are *i.i.d.* standard normal
- Outcome model: linear model
- Propensity score model: logistic model with linear predictors
- Nonlinear 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 Evaluated

1. Horvitz-Thompson (HT):

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

2. Inverse-probability weighting with normalized weights (IPW): Same as HT but with normalized weights
3. Weighted least squares regression (WLS): linear regression with HT weights
4. Doubly-robust least squares regression (DR): consistently estimates the ATE if *either* the outcome or propensity score model is correct

Weighting Estimators Do Fine If the Model is Correct

Sample size	Estimator	Bias GLM	Bias True	RMSE GLM	RMSE True
(1) Both models correct					
$n = 200$	HT	-0.01	0.68	13.07	23.72
	IPW	-0.09	-0.11	4.01	4.90
	WLS	0.03	0.03	2.57	2.57
	DR	0.03	0.03	2.57	2.57
$n = 1000$	HT	-0.03	0.29	4.86	10.52
	IPW	-0.02	-0.01	1.73	2.25
	WLS	-0.00	-0.00	1.14	1.14
	DR	-0.00	-0.00	1.14	1.14
(2) Propensity score model correct					
$n = 200$	HT	-0.32	-0.17	12.49	23.49
	IPW	-0.27	-0.35	3.94	4.90
	WLS	-0.07	-0.07	2.59	2.59
	DR	-0.07	-0.07	2.59	2.59
$n = 1000$	HT	0.03	0.01	4.93	10.62
	IPW	-0.02	-0.04	1.76	2.26
	WLS	-0.01	-0.01	1.14	1.14
	DR	-0.01	-0.01	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	24.72	0.25	141.09	23.76
	IPW	2.69	−0.17	10.51	4.89
	WLS	−1.95	0.49	3.86	3.31
	DR	0.01	0.01	2.62	2.56
$n = 1000$	HT	69.13	−0.10	1329.31	10.36
	IPW	6.20	−0.04	13.74	2.23
	WLS	−2.67	0.18	3.08	1.48
	DR	0.05	0.02	4.86	1.15
(4) Both models incorrect					
$n = 200$	HT	25.88	−0.14	186.53	23.65
	IPW	2.58	−0.24	10.32	4.92
	WLS	−1.96	0.47	3.86	3.31
	DR	−5.69	0.33	39.54	3.69
$n = 1000$	HT	60.60	0.05	1387.53	10.52
	IPW	6.18	−0.04	13.40	2.24
	WLS	−2.68	0.17	3.09	1.47
	DR	−20.20	0.07	615.05	1.75

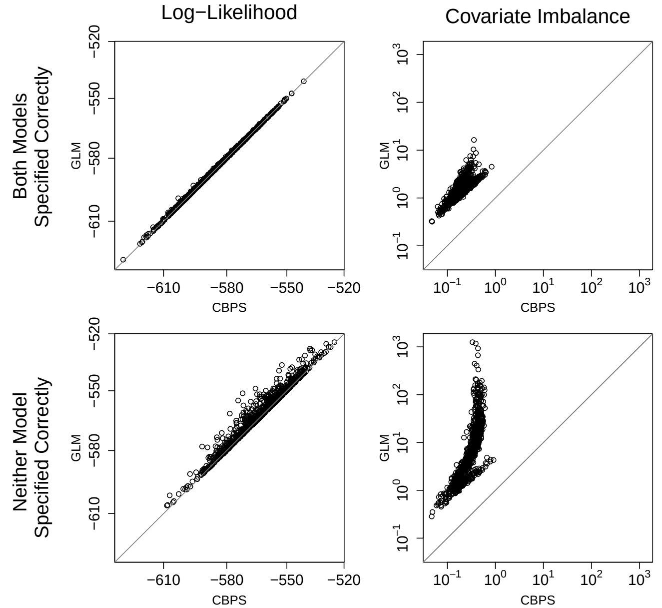
Revisiting Kang and Schafer (2007)

	Estimator	Bias GLM	Bias Balance	Bias CBPS	Bias True	RMSE GLM	RMSE Balance	RMSE CBPS	RMSE True
(1) Both models correct									
$n = 200$	HT	−0.01	2.02	0.73	0.68	13.07	4.65	4.04	23.72
	IPW	−0.09	0.05	−0.09	−0.11	4.01	3.23	3.23	4.90
	WLS	0.03	0.03	0.03	0.03	2.57	2.57	2.57	2.57
	DR	0.03	0.03	0.03	0.03	2.57	2.57	2.57	2.57
$n = 1000$	HT	−0.03	0.39	0.15	0.29	4.86	1.77	1.80	10.52
	IPW	−0.02	0.00	−0.03	−0.01	1.73	1.44	1.45	2.25
	WLS	−0.00	−0.00	−0.00	−0.00	1.14	1.14	1.14	1.14
	DR	−0.00	−0.00	−0.00	−0.00	1.14	1.14	1.14	1.14
(2) Propensity score model correct									
$n = 200$	HT	−0.32	1.88	0.55	−0.17	12.49	4.67	4.06	23.49
	IPW	−0.27	−0.12	−0.26	−0.35	3.94	3.26	3.27	4.90
	WLS	−0.07	−0.07	−0.07	−0.07	2.59	2.59	2.59	2.59
	DR	−0.07	−0.07	−0.07	−0.07	2.59	2.59	2.59	2.59
$n = 1000$	HT	0.03	0.38	0.15	0.01	4.93	1.75	1.79	10.62
	IPW	−0.02	−0.00	−0.03	−0.04	1.76	1.45	1.46	2.26
	WLS	−0.01	−0.01	−0.01	−0.01	1.14	1.14	1.14	1.14
	DR	−0.01	−0.01	−0.01	−0.01	1.14	1.14	1.14	1.14

CBPS Makes Weighting Methods Work Better

	Estimator	Bias GLM	Bias Balance	Bias CBPS	Bias True	RMSE GLM	RMSE Balance	RMSE CBPS	RMSE True
(3) Outcome model correct									
$n = 200$	HT	24.72	0.33	−0.47	0.25	141.09	4.55	3.70	23.76
	IPW	2.69	−0.71	−0.80	−0.17	10.51	3.50	3.51	4.89
	WLS	−1.95	−2.01	−1.99	0.49	3.86	3.88	3.88	3.31
	DR	0.01	0.01	0.01	0.01	2.62	2.56	2.56	2.56
$n = 1000$	HT	69.13	−2.14	−1.55	−0.10	1329.31	3.12	2.63	10.36
	IPW	6.20	−0.87	−0.73	−0.04	13.74	1.87	1.80	2.23
	WLS	−2.67	−2.68	−2.69	0.18	3.08	3.13	3.14	1.48
	DR	0.05	0.02	0.02	0.02	4.86	1.16	1.16	1.15
(4) Both models incorrect									
$n = 200$	HT	25.88	0.39	−0.41	−0.14	186.53	4.64	3.69	23.65
	IPW	2.58	−0.71	−0.80	−0.24	10.32	3.49	3.50	4.92
	WLS	−1.96	−2.01	−2.00	0.47	3.86	3.88	3.88	3.31
	DR	−5.69	−2.20	−2.18	0.33	39.54	4.22	4.23	3.69
$n = 1000$	HT	60.60	−2.16	−1.56	0.05	1387.53	3.11	2.62	10.52
	IPW	6.18	−0.87	−0.72	−0.04	13.40	1.86	1.80	2.24
	WLS	−2.68	−2.69	−2.70	0.17	3.09	3.14	3.15	1.47
	DR	−20.20	−2.89	−2.94	0.07	615.05	3.47	3.53	1.75

CBPS Sacrifices Likelihood for Better Balance



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):
  - LaLonde experimental sample rather than DW sample
  - Dehejia & Wahba (DW)'s results are sensitive to model specification
  - They are also sensitive to the selection of comparison sample

Evaluation Design

Observed Data

Experimental (LaLonde) Sample

Treated, $n = 297$

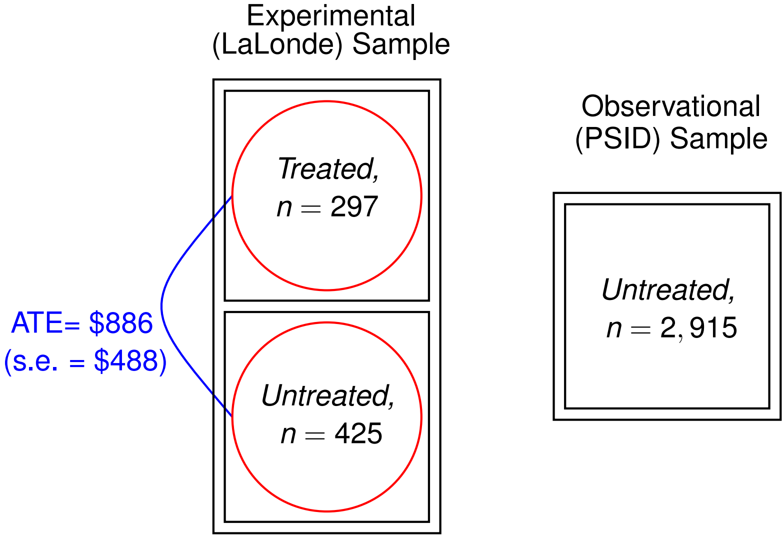
Untreated, $n = 425$

Observational (PSID) Sample

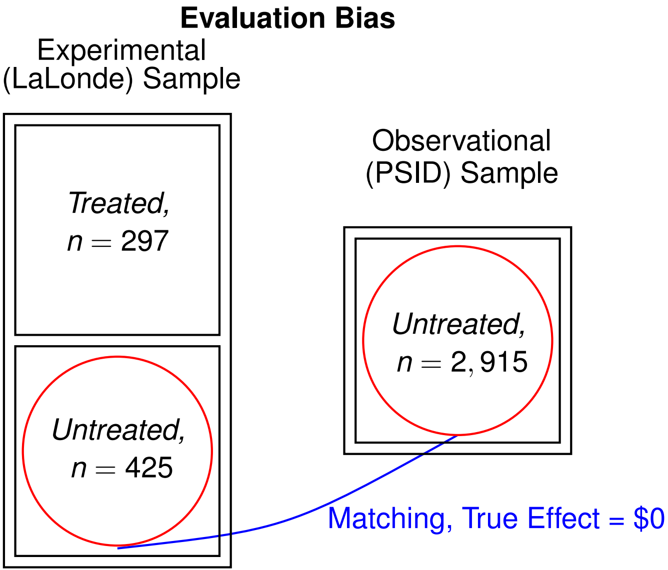
Untreated, $n = 2,915$

Evaluation Design

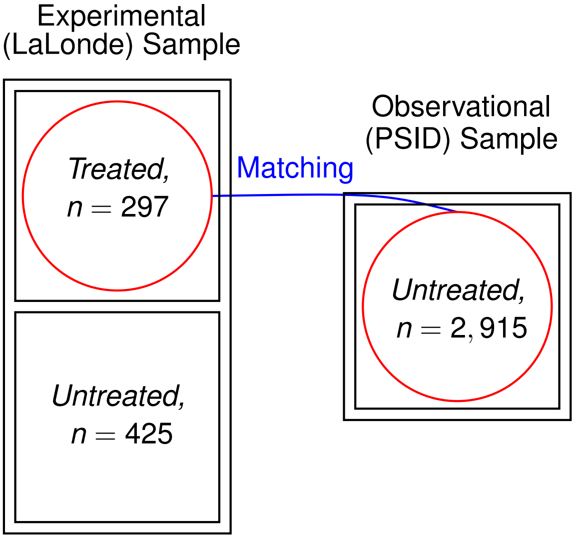
Experimental Benchmark



Evaluation Design



Matching Estimator



Evaluation Bias

- Propensity score:
  - Conditional probability of being in the experimental sample
  - Logistic regression for propensity score
- "True" estimate $= 0$
- Nearest neighbor matching with replacement
- CBPS reduces bias:

	1–to–1 Matching			Optimal 1–to–N Matching		
Specification	GLM	Balance	CBPS	GLM	Balance	CBPS
Linear	−835	−568	−302	−1022	−265	−67
	(886)	(898)	(869)	(499)	(492)	(487)
Quadratic	−1570	−950	−1036	−1209	−950	−480
	(1003)	(882)	(831)	(558)	(617)	(512)
Smith & Todd (2005)	−1859	−1074	−1298	−1810	−1164	−419
	(1004)	(860)	(800)	(500)	(485)	(464)

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:
  - GenMatch: $-$\$412 (s.e. $=$ 553)
  - CEM: $-$\$29 (s.e. $=$ 452)
  - ebal: $-$\$203 (s.e. $=$ 256)
- CBPS gives estimates closer to experimental benchmark:

	1–to–1 Matching			Optimal 1–to–N Matching		
Model specification	GLM	Balance	CBPS	GLM	Balance	CBPS
Linear	−835	−568	−302	−430	507	123
	(1374)	(1811)	(1849)	(749)	(822)	(799)
Quadratic	−919	−379	−379	−419	193	439
	(1245)	(1219)	(1140)	(558)	(617)	(512)
Smith & Todd (2005)	−811	−507	−131	−811	−487	289
	(1225)	(1189)	(1058)	(1225)	(676)	(673)

Extensions to Other Causal Inference Settings

- Propensity score methods are widely applicable
- Thus, CBPS is also widely applicable
- Extensions in progress:
  1. Non-binary treatment regimes
  2. Causal inference with longitudinal data
  3. Generalizing experimental estimates
  4. Generalizing instrumental variable estimates
- In many of these situations, balance checking is difficult

Generalizing Experimental Estimates

- Lack of external validity for experimental estimates
- Target population $\mathcal{P}$
- Experimental sample: $S_i = 1$ with $i = 1, 2, \ldots, N_e$
- Non-experimental sample: $S_i = 0$ with $i = N_e + 1, \ldots, N$
- Sampling on observables: $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp S_i \mid X_i$
- Propensity score: $\pi_\beta(X_i) = \Pr(S_i \mid X_i)$
- Score equation: logistic likelihood
- Balancing between experimental and non-experimental sample:

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

- Can also balance weighted treatment and control groups

Concluding Remarks

- Covariate balancing propensity score:
    1. simultaneously optimizes prediction of treatment assignment and covariate balance under the GMM framework
    2. is robust to model misspecification
    3. improves propensity score weighting and matching methods
    4. can be extended to various situations

- Open-source software, `CBPS: R Package for Covariate Balancing Propensity Score`, is available at CRAN