Comparing Causal Inference Estimators for Average Treatment Effect of Treated Units in Observational Studies

Kip Brown

kibrown@siue.edu
Department of Mathematics and Statistics
Southern Illinois University Edwardsville

March 6, 2018

Overview

1. Problem Set Up
   - Definitions and Assumptions
   - Propensity Score Framework
   - Estimating the Propensity Score
   - Covariate Balancing Propensity Score
2. Causal Inference Methods
   - Matching Methods
   - Stratification
   - Inverse Probability of Treatment Weighting
   - Entropy Balancing
3. Simulations
4. Empirical Study

Randomized vs. Observational Studies

- Randomized studies balance covariate distribution by design
- Observational studies may have unbalanced covariate distributions
- This leads to biased estimates

The Response Function

Definition

Suppose we have a random sample of size n from a population. For the i^{th} unit in the sample, let T_i denote which treatment was received, where $T_i = 0$ denotes the i^{th} unit receiving the control treatment, and $T_i = 1$ denote the i^{th} unit receiving the treatment of interest. Let $Y_i(0)$ and $Y_i(1)$ denote the outcomes of the control treatment and the treatment of interest, respectively. Let

$$Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0) \tag{1}$$

denote the response of the i^{th} unit.

Treatment Effects

Definition

Let the average treatment effect (ATE) be defined as

$$\tau = E[Y_i(1) - Y_i(0)].$$

Let the average treatment effect for the treated (ATT) be defined as,

$$\tau_t = E[Y_i(1) - Y_i(0) \mid T_i = 1]. \tag{2}$$

Major Assumptions

Assumption

(Unconfoundedness) For any unit $i = 1, \ldots, n$*,*

$$P(T_i = 1 \mid Y_i(0), Y_i(1), X_i) = P(T_i = 1 \mid X_i) \tag{3}$$

or, using conditional independence notation

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

Assumption

(Probabilistic Assignment) For any unit $i = 1, \ldots, n$*,*

$$0 < P(T_i = 1 \mid X_i) < 1$$

Major Assumptions Continued

Assumption

(Individualistic) For any unit $i = 1, \ldots, n$, the probability of treatment assignment can be written as a common function of the i^{th}'s unit potential outcome and observed covariates.

Balancing Scores

Definition

(Balancing Score) A balancing score $b(x)$ is a function of the covariates such that

$$T_i \perp\!\!\!\perp X_i \mid b(Xi).$$

This can also be represented as a probability,

$$P(T_i = 1 \mid X_i, b(X_i)) = P(T_i = 1 \mid b(X_i)). \tag{4}$$

- ie: X_i is a balancing score

A Better Balancing Score

Definition

(Propensity Score) The Propensity Score is the conditional probability that a unit with observed covariates, x, will be in treatment group 1. The Propensity Score $\pi(X_i)$ is then,

$$\pi(X_i) = P(T_i = 1 \mid X_i = x). \tag{5}$$

Propensity Score Theorems

Theorem

(Propensity Score is a balancing score) The propensity score $\pi(X_i) = P(T = 1|X_i = x)$ *is a balancing score.*

Theorem 1 Proof

Proof.

We must show that the propensity score is a balancing score, which by equation (4),

$$P(T_i = 1 \mid X_i, \pi(X_i)) = P(T_i = 1 \mid \pi(X_i)). \tag{6}$$

Starting with the left side of (6), we have

$$\begin{aligned} P(T_i = 1 \mid X_i, \pi(X_i)) &= P(T_i = 1 \mid X_i) \\ &= \pi(X_i). \end{aligned}$$

Proof Continued

Proof.

Now with the right side of (6), we have

$$\begin{aligned}
P(T_i = 1 \mid \pi(X_i)) &= 1 \cdot P(T_i = 1 \mid \pi(X_i)) + 0 \\
&= 1 \cdot P(T_i = 1 \mid \pi(X_i)) + 0 \cdot P(T_i = 0 \mid \pi(X_i)) \\
&= E_T \left[T_i \mid \pi(X_i) \right] \\
&= E_X \Big[E_T \left[T_i \mid X_i, \pi(X_i) \right] \mid \pi(X_i) \Big] \\
&= E_X \Big[P(T_i \mid X_i, \pi(X_i)) \mid \pi(X_i) \Big] \\
&= E_X \Big[\pi(X_i) \mid \pi(X_i) \Big] \\
&= \pi(X_i).
\end{aligned}$$

Thus, $\pi(X_i)$ is a balancing score. □

Propensity Score Theorems

Theorem

(Unconfoundedness given any balancing score)
Suppose Assumption 1 is true. Then, treatment assignment is unconfounded given any balancing score,

$$P(T_i = 1 \mid Y_i(0), Y_i(1), b(X_i)) = P(T_i = 1 \mid b(X_i)) \tag{7}$$

or, using conditional independence notation

$$T_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \mid b(X_i).$$

Theorem 2 Proof

Proof.

Let Assumption 1 be true. We will start with the left side of (7), and show the right.

$$
\begin{aligned}
&P(T_i = 1 \mid Y_i(0), Y_i(1), b(X_i)) \\
&\quad = E_T\left[T_i \mid Y_i(0), Y_i(1), b(X_i)\right] \\
&\quad = E_X\Big[E_T\left[T_i \mid Y_i(0), Y_i(1), X_i, b(X_i)\right] \mid Y_i(0), Y_i(1), b(X_i)\Big] \\
&\quad = E_X\Big[E_T\left[T_i \mid X_i, b(X_i)\right] \mid Y_i(0), Y_i(1), b(X_i)\Big] \\
&\quad = E_X\Big[E_T\left[T_i \mid b(X_i)\right] \mid Y_i(0), Y_i(1), b(X_i)\Big] \\
&\quad = E_X\left[T_i \mid b(X_i)\right] \\
&\quad = 1 \cdot P(T_i = 1 \mid b(X_i)) + 0 \cdot P(T_i = 0 \mid b(X_i)) \\
&\quad = P(T_i = 1 \mid b(X_i)).
\end{aligned}
$$

Estimating the Propensity Score

- The true propensity score is unknown
- $\pi(X_i) = P(T_i = 1 \mid X_i)$ can be modeled with logistic regression

Definition

The binary logistic regression response function is

$$\pi(X_i) = \frac{\exp(X_i'\beta)}{1 + \exp(X_i'\beta)}, \tag{8}$$

where X_i is vector of covariates for the i^{th} unit, and β is the vector of parameters.

Likelihood Function

- T_i is a Bernoulli random variable

$$L(\beta) = \prod_{i=1}^{n} \pi(X_i)^{T_i} \cdot (1 - \pi(X_i))^{1-T_i}$$

$$\begin{aligned} l = \ln(L(\beta)) &= \ln\left(\prod_{i=1}^{n} \pi(X_i)^{T_i} \cdot (1 - \pi(X_i))^{1-T_i} \right) \\ &= \sum_{i=1}^{n} \ln\left(\pi(X_i)^{T_i} \cdot (1 - \pi(X_i))^{1-T_i} \right) \\ &= \sum_{i=1}^{n} \Big(T_i \cdot \ln(\pi(X_i)) + (1 - T_i) \cdot \ln(1 - \pi(X_i)) \Big). \end{aligned}$$

Estimated Propensity Score

Using the MLE method to estimate the parameters,

$$\hat{\pi}(X_i) = \frac{\exp(X_i'b)}{1 + \exp(X_i'b)}$$

or equivalently,

$$ln\left(\frac{\hat{\pi}(X_i)}{1 - \hat{\pi}(X_i)}\right) = X_i b$$

Measures of Model Accuracy

Definition

Let l_r and l_p be the log-likelihood functions for the reduced model and the proposed model respectively. Then the likelihood ratio statistic, D, is

$$D = 2[l_p - l_r]$$

Estimating the Propensity Score in R

1. Include all scientifically significant predictors
2. Include all statistically significant first order terms
3. Include all statistically significant second order terms

Choosing Statistically Significant Terms

Suppose there are *p* variables in the data set,

1. Fit a base model with all *a* scientifically significant predictors
2. Fit *p* - *a* new models, each with the scientifically significant predictors, plus 1 of the remaining variables.
3. Calculate the likelihood ratio statistic, D, for each model
4. If any $D \geq 1$, add that predictor variable to the base model, and repeat steps 2 - 4
5. When all $D < 1$, add no more first order predictor variables

Choosing Statistically Significant Terms

- Choosing second order terms follows a very similar logic
- Only consider second order terms that include variables already in the model
- If any $D \geq 2.71$, add it to the model, if not move on

Issues with the Propensity Score

- The propensity score must be correctly modeled
- The goal of a logistic regression model is accurate prediction of $P(T_i = 1)$
- Iterative processes can be lengthy and difficult

Covariate Balancing Propensity Score

- The goal is to balance the covariates
- This is not an iterative process
- Parameter estimates are derived by the MLE method with the following balancing condition

$$\frac{1}{n_1}\sum_{i=1}^{n}\left(T_i - \frac{(1-T_i)\pi(X_i)}{1-\pi_(X_i)}\right)\tilde{X}_i = 0$$

- $\tilde{X}_i = X_i$ or $\tilde{X}_i = (X_i^T(X_i^2)^T)^T$

Estimation Methods

1. Matching Methods
2. Stratification Methods
3. Inverse Probability Methods
4. Entropy Balancing Methods

Matching Logic

- Similar treatment and control units are directly compared
- ”Most similar” is determined by some distance metric based on the covariates
- Once treatment and control units are matched, responses can be compared

Distance Metrics

Definition

(Mahalanobis Distance): The mahalanobis distance is defined as

$$D_{ij} = (\mathbf{X_i} - \mathbf{X_j})' \Sigma^{-1} (\mathbf{X_i} - \mathbf{X_j}),$$

where $\mathbf{X_i}, \mathbf{X_j}$ are $p \times 1$ vectors of covariates for the i^{th} and j^{th} units respectively, and Σ is the variance-covariance matrix of the covariates.

Distance Metrics

Definition

(Absolute Propensity Score Difference): This difference is defined as,

$$D_{ij} = |\pi(X_i) - \pi(X_j)|,$$

where $\pi(X_i), \pi(X_j)$ are the estimated propensity scores for the i^{th} and j^{th} units respectively.

Matching Processes

- Nearest Neighbor Matching
- 1 - 1 matching
- 1 - n matching
- greedy matching
- optimal matching

Matching Algorithm

Focus on 1 - 1 nearest neighbor greedy matching

1. Estimate the propensity score
2. Select a treatment unit at random, find the closest control unit
3. Find the difference in the responses
4. Discard both units
5. Repeat steps 2 - 4 until all treatment units have been discarded
6. Take the mean of the responses

Calipers and Efficiency Bounds

- Calipers are set as maximum allowed distance between matched units
- Efficiency bounds trim extreme propensity scores
- Trim controls with high propensity scores and treatments with low propensity scores

Optimal Trimming

Theorem

Let α be defined such that we only consider propensity score values $\alpha \leq \pi(X_i) \leq 1 - \alpha$. If

$$\sup_{x \in \mathbb{X}} \frac{1}{\pi(X_i)1 - \pi(X_i)} \leq 2E\left[\frac{1}{\pi(X_i)1 - \pi(X_i)}\right],$$

then $\alpha = 0$. Otherwise, α is a solution to

$$\frac{1}{\alpha(1-\alpha)} = 2E\left[\frac{1}{\pi(X_i)1 - \pi(X_i)} \,\middle|\, \frac{1}{\pi(X_i)1 - \pi(X_i)} \leq \frac{1}{\alpha(1-\alpha)}\right].$$

Stratification Logic

- The data is broken into strata where each unit has a similar propensity score
- Inside each strata, units will have a better covariate balance
- The average response for the treatment and controls are compared within each strata
- A weighted average of the strata specific difference is an estimate for average treatment effect

Stratification Methods

- The researcher defines some J strata by a predetermined method
- **Method 1**: Divided such that there are a roughly equal number of units in each strata
- **Method 2**: Divided such that there are equal ranges of propensity scores in each strata

Stratification Algorithm

1. Estimate the propensity score
2. Divide the data into J strata by chosen method
3. Stratum specific average differences are calculated

$$\tau_{diff}(j) = \bar{Y}_t(j) - \bar{Y}_c(j),$$

where

$$\bar{Y}_t(j) = \frac{1}{N_t(j)} \sum_{i=1}^{n} T_i \cdot B_i(j) \cdot Y_i$$

and

$$\bar{Y}_c(j) = \frac{1}{N_c(j)} \sum_{i=1}^{n} (1 - T_i) \cdot B_i(j) \cdot Y_i,$$

Stratification Algorithm

The stratification estimator then becomes,

$$\tau_{strat} = \sum_{j=1}^{J} q(j) \cdot \tau_{diff}(j)$$

Inverse Probability Weighting

- Responses are re-weighted via the following estimator

$$\tau_{W,ATT} = \frac{1}{n}\sum_{i=1}^{n} T_i Y_i - \frac{1}{n}\sum_{i=1}^{n} \frac{(1-T_i)Y_i\pi(X_i)}{1-\pi(X_i)},$$

Entropy Balancing Logic

- The responses are each re-weighted
- A loss function is defined, and constraints are set to balance the covariates
- The method of Lagrange multipliers is used to derive the weights.

Entropy Balancing

- **Loss Function:** $h(w_i) = w_i \cdot \ln(\frac{w_i}{q_i})$
- **Constraint:** $\sum_{i|T=0} w_i \cdot c_{ri}(X_i) = m_r \quad \text{with} \quad r \in 1, ..., R.$
- **Constraint:** $\sum_{i|T=0} w_i = 1 \quad \text{and} \quad w_i \geq 0.$

Entropy Balancing: Deriving the Weights

- $L = \sum_{i|T=0} w_i \cdot \ln(\frac{w_i}{q_i}) + \sum_{r=1}^{R} \lambda_r \Big(\sum_{i|T=0} w_i \cdot c_{ri}(X_i) - m_r \Big) + (\lambda_0 - 1) \Big(\sum_{i|T=0} w_i - 1 \Big)$
- $\frac{\partial L}{\partial w_i} = \ln(\frac{w_i}{q_i}) + 1 + \Big(\sum_{r=1}^{R} \lambda_r c_{ri}(X_i) \Big) + (\lambda_0 - 1) = 0$
- $w_i = q_i \cdot \exp\Big(- \sum_{r=1}^{R} \lambda_r c_{ri}(X_i) \Big) \cdot \exp(-\lambda_0)$
- $w_i^* = \frac{q_i \cdot \exp(-\lambda_1 c_{1i}(X_i) - \lambda_2 c_{2i}(X_i))}{\Sigma_{\{i|T=0\}} q_i \cdot \exp(-\lambda_1 c_{1i}(X_i) - \lambda_2 c_{2i}(X_i))}$

Simulation Design

	Description
A	50 Treated, 100 Control, equal variance-covariance
B	250 Treated, 250 Control, equal variance-covariance
C	50 Treated, 100 Control, unequal variance-covariance
D	250 Treated, 250 Control, unequal variance-covariance

- 3 Pre-treatment variables
- 1000 iterations

Design A and B

- $\Sigma_{T=1} = \Sigma_{T=0} =$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

- $\mu_{T=1} = [0, 0, 0]'$
- $\mu_{T=0} = [0.4, 0.4, 0.4]'$

Design C and D

- $\Sigma_{T=1} =$

$$\begin{bmatrix} 1.5 & 0 & 0 \\ 0 & 1.5 & 0 \\ 0 & 0 & 1.5 \end{bmatrix}$$

- $\Sigma_{T=0} =$

$$\begin{bmatrix} 0.5 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0.5 \end{bmatrix}$$

- $\mu_{T=1} = [0, 0, 0]'$
- $\mu_{T=0} = [0.4, 0.4, 0.4]'$

Design A Results

	RAW	PSM	PSMC	PSTMC	PSTMO
Bias	-121.26	-2.75	0.53	0.44	0.72
MSE	155.96	1.76	1.53	1.66	1.61
	WPS	CBMat1	CBMat2	CBMat3	CBMat4
Bias	-0.53	-3.95	-1.20	-1.15	-1.36
MSE	3.08	1.84	1.69	1.82	1.77
	MD	STRT1	STRT2	CBWPS	EB
Bias	-18.26	-11.96	-12.66	0.04	-0.95
MSE	4.84	3.05	3.18	0.24	0.26

Table: Design A Results

Design B Results

	RAW	PSM	PSMC	PSTMC	PSTMO
Bias	-120.17	-1.52	-0.10	-0.24	0.06
MSE	146.74	0.26	0.19	0.19	0.18
	WPS	CBMat1	CBMat2	CBMat3	CBMat4
Bias	-0.02	-1.93	-0.53	-0.58	-0.43
MSE	1.25	0.27	0.19	0.20	0.20
	MD	STRT1	STRT2	CBWPS	EB
Bias	-12.09	-10.33	-12.27	0.01	-0.12
MSE	1.70	1.38	1.84	0.07	0.07

Table: Design B Results

Design C Results

	RAW	PSM	PSMC	PSTMC	PSTMO
Bias	-121.05	-23.04	0.24	-2.58	0.82
MSE	156.31	9.32	1.42	2.50	1.57
	WPS	CBMat1	CBMat2	CBMat3	CBMat4
Bias	-49.79	-25.07	-1.03	-2.28	-1.33
MSE	28.36	10.64	1.91	2.65	2.22
	MD	STRT1	STRT2	CBWPS	EB
Bias	-43.61	1.36	-13.60	-0.09	-0.83
MSE	22.23	1.55	3.36	0.36	0.37

Table: Design C Results

Design D Results

	RAW	PSM	PSMC	PSTMC	PSTMO
Bias	-119.56	-15.78	-0.56	-1.45	-0.18
MSE	145.44	3.60	0.21	0.38	0.19
	WPS	CBMat1	CBMat2	CBMat3	CBMat4
Bias	-57.87	-15.62	-0.46	-0.77	0.50
MSE	34.45	3.61	0.26	0.28	0.23
	MD	STRT1	STRT2	CBWPS	EB
Bias	-34.54	-9.54	-12.63	-0.09	-0.20
MSE	12.75	1.31	1.96	0.10	0.10

Table: Design D Results

Simulation Conclusions

- IPW with CBPS
- Entropy Balancing
- Propensity score matching with a caliper = 0.1 and trimming

Mastectomies versus Breast Conservation Methods

- 1980's study conducted by the the GBCSG
- **Treatments:** Mastectomies, Breast Conservation Methods (BCM)
- **Responses:** Physical and Emotional Status
- $n = 646$ patients, $n_0 = 479$ mastectomies, $n_1 = 167$ BCM's

Table: Head of the stu1 data

	klinik	tmass	therapie	alter	tgr	age	ewb	pst	mp
1	3	-7.76	0	-11.91	1	1	63.46	81.25	0
2	3	-4.76	0	-4.91	1	1	90.38	93.75	0
3	4	-3.76	0	-14.91	1	1	73.08	93.75	1
4	6	-7.76	0	-0.91	1	1	75.00	81.25	1
5	6	-3.76	0	-1.91	1	1	34.62	56.25	1

Summary Statistics

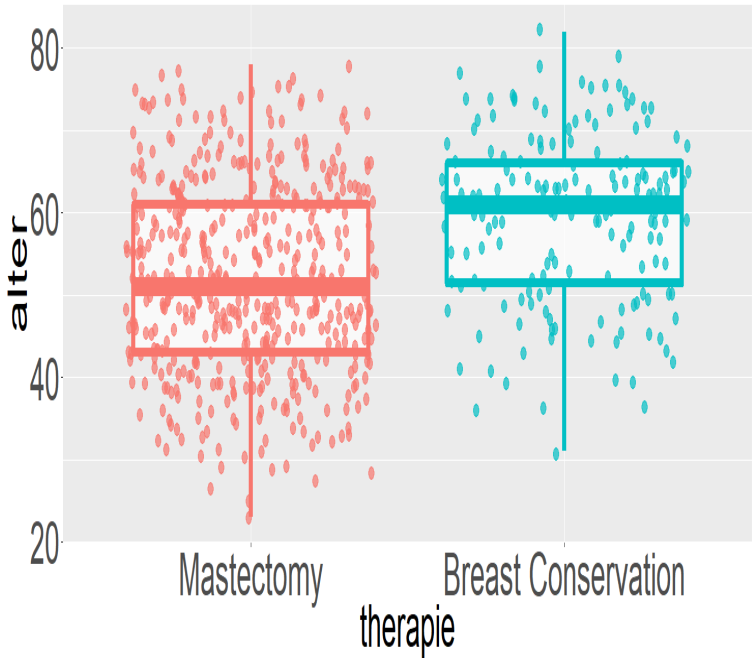
Table: Sample Means and Standard Deviations

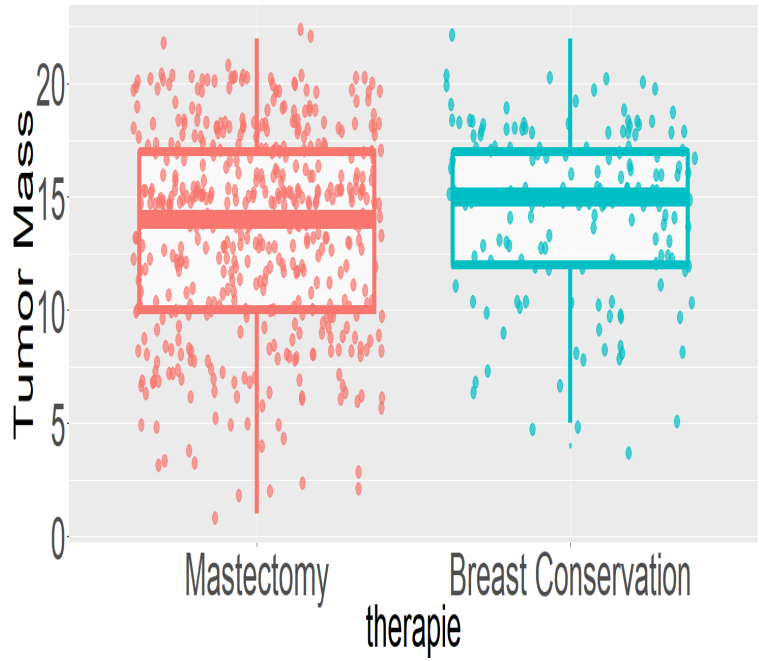
	$\bar{x}_{BC}$	$\bar{x}_{M}$	S_{BC}	S_{M}
tmass	14.47	13.51	4.40	3.64
alter	59.41	52.00	11.50	10.40

Table: Sample Conditional Proportions of Categorical Variables

	mp		tgr		age	
	< 15	≥ 15	≤ 10mm	> 10mm	≤ 55	> 55
BC	0.5449	0.4551	0.1796	0.8204	0.3353	0.6647
M	0.4405	0.5595	0.2714	0.7286	0.6138	0.3862

Continuous Plots





Estimating the Propensity Scores

- Include scientifically significant predictors
- Include statistically significant linear predictors

	Step			
$tmass$	6.53	4.28	5.11	-
$alter$	52.07	-	-	-
$many\_pat$	5.42	6.44	-	-
tgr	5.87	4.14	4.39	0.25
age	38.96	0.05	0.03	0.01

Table: Likelihood Ratio Statistics

Estimating the Propensity Scores

- Include statistically significant second order predictors

	Step	
$alter^2$	0.86	0.76
$alter \cdot many\_pat$	0.38	0.38
$alter \cdot tmass$	0.19	0.17
$many\_pat \cdot tmass$	0.06	0.22
$tmass^2$	3.98	-

Table: Likelihood Ratio Statistics

Estimated Propensity Scores

Logit Propensity Score Model:

$$\ln\left(\frac{\pi(X_i)}{1-\pi(X_i)}\right) =$$

$$-0.756 + 0.059 \cdot alter - 0.495 \cdot mp + 0.041 \cdot tmass - 0.010 \cdot tmass^2$$

Logit CBPS Model:

$$\ln\left(\frac{\pi_{CB}(X_i)}{1-\pi_{CB}(X_i)}\right) =$$

$$-0.780 + 057 \cdot alter - 0.476 \cdot mp + 0.044 \cdot tmass - 0.009 \cdot tmass^2$$

ATT Estimates

	RAW	PSTMO	CBWPS	EB
ATT	-1.59	-1.67	-1.15	1.73

Table: `ATT` Estimates for pst

	RAW	PSTMO	CBWPS	EB
ATT	0.09	0.44	0.19	0.33

Table: `ATT` Estimates for ewb

Conclusion

- Propensity score can be modeled with logistic regression
- CBPS takes away the iterative process
- Matching calipers and trimming
- IPW and entropy balancing
- Codes at https://github.com/kbrown1224/Thesis-Codes

- Austin, P. (2011), "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies," *Multivariate Behavioral Research,* 46: 399 - 424.
- Crump et. al. (2009), "Dealing with limited overlap in estimation of average treatment effects" *Biometrika,* Vol. 96, No. 1, 187 - 199.
- Dawid, A. (1979), "Conditional Independence in Statistical Theory (with discussion)" *J.R. Statist Soc.,* B41, 1-31.
- R. Dehejia and S. Wahba (1999), "Causal Effects in Nonexperimental Studies: Reevaluationg the Evaluation of Training Programs," *Journal of the American Statistical Association,* Vol. 94, No. 448 1053 - 1062.
- K. Hirano and G. Imbens (2001), "Estimation of Causal Effects using Propensity Score Weighting: An Application to Data on Right Heart Catherization," *Health Services & Outcomes Research Methodology,* 2: 259 - 278.

- P. Rosenbaum and D. Rubin. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika,* Vol. 70, No. 1, 41 - 55.
- P. Rosenbaum and D. Rubin. (1984), "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score" *Journal of the American Statistical Association,* Vol. 79, No. 387 516 - 524.
- J. Hahn. (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica,* Vol. 66, No. 2, 315 - 331.
- J. Hainmueller. (2012), "Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies," *Oxford Journals,* Vol. 20, No. 1, 25 - 46.
- K Imai and M Ratkovic. (2014) "Covariate balancing propensity score," *Journal of the Royal Statistical Society,* Vol. 76, Part 1, 243 - 263.

- M. Kutner, C. Nachtsheim, and J. Neter. (2004), ”Applied Linear Regression Models”, *McGraw-Hill/Irwin* 4th Ed., 570 - 572.
- E. Stuart. (2010), ”Matching Methods for Causal Inference: A Review and A Look Forward”, *Statistical Science,* Vol. 25, No. 1, 1 - 21.
- G. Imbens and D. Rubin. (2015) ”Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction”, *Cambridge University Press*
- S. Senn, E. Graf, and A. Caputo. (2007) ”Stratification for the propensity score compared with linear regression techniques to assess the effect of treatment or exposure”, *Wiley InterScience* Vol. 26, 5529 - 5544.