Bias Associated with Using Propensity Scores as a Regression Covariate

Erinn Hade and Bo Lu

Division of Biostatistics, College of Public Health
The Ohio State University

Present for 2011 Midwest Biopharmaceutical Statistics Workshop

Outline

Background

- A major goal of clinical/medical research is to assess the efficacy or effectiveness of an intervention.
- This is particularly challenging in observational studies when randomization is not feasible due to ethical or legal issues.
- Propensity score based adjustment is a popular approach to removing selection bias due to observed covariates.
- Propensity score is the conditional probability of receiving treatment given observed covariates.

Background

- Propensity score based adjustment works under two assumptions:
  - Stable Unit Treatment Value Assumption (SUTVA)
  - Strongly Ignorable Treatment Assignment
- Analytical use of propensity score includes:
  - Matching
  - Stratification
  - Regression Adjustment
  - Weighting
- Regression adjustment seems appealing to practitioners due to the simplicity.

Literature Review

- We conducted a systematic literature review to gauge how often each method is used in practice
  - Major medical/health journals
    The New England Journal of Medicine (NEJM), the Journal of the American Medical Association (JAMA), the American Journal of Public Health (AJPH), the American Journal of Preventive Medicine (AJPM), and the American Journal of Epidemiology(AJE)
  - Time period
    From January 2000 to December 2009, for NEJM, AJPH, AJPM, AJE.
    From January 2007 to December 2009, for JAMA, to avoid over-representation.
  - Key word search– "propensity score" in title, abstract, text.
    Including only intervention evaluation with a real data example.
    Pubmed and each journal's webiste
  - Total count: 80
    NEJM (27), JAMA (25), AJPH (11), AJPM (4), AJE (13).

Literature Review

Summary of Findings

Method (Propensity Score)	N (%)
Matching	25 (31%)
Linear covariate in regression	19 (24%)
Stratification	17 (21%)
Weighting	16 (20%)
Other	3 (4%)

The Problem

- Using propensity score as a regression covariate seems to be a popular choice.
- In practice,
  - Estimate the propensity score with a logistic regression model
  - Plug in the estimated propensity score into the original regression model, in addition to the treatment indicator, either with or without other covariates
  - Regard the coefficient of the treatment indicator as the treatment effect, controlling for selection bias
- The problem is that there is little justification or diagnostics regarding the appropriateness of such practice.

The Problem

- Austin et al. (2007) investigated the bias associated with using the propensity score as a covariate in nonlinear models, via a Monte Carlo simulation study
- Logistic regression model and Cox proportional hazard model were studied and the conditional odds ratio and hazard ratio were found to be biased (toward null)
- Our goal:
  - To investigate the bias in linear models under common assumptions
  - Compare the relative performance of different propensity score adjustment via an extensive simulation study, from a NEW data driven perspective

Outline

PS in Linear Models

As shown in Rosenbaum and Rubin (1983) and Wooldridge (2002), propensity scores can be used in covariance adjustment, with appropriate assumptions.

COROLLARY 4·3. *Covariance adjustment on balancing scores. Suppose treatment assignment is strongly ignorable, so that in particular,* $E\{r_t \mid z = t, b(x)\} = E\{r_t \mid b(x)\}$ *for balancing score* $b(x)$. *Further suppose that the conditional expectation of* r_t *given* $b(x)$ *is linear:*

$$E\{r_t \mid z = t, b(x)\} = \alpha_t + \beta_t b(x) \quad (t = 0,1).$$

Then the estimator

$$(\hat{\alpha}_1 - \hat{\alpha}_0) + (\hat{\beta}_1 - \hat{\beta}_0)\, b(x)$$

is conditionally unbiased given $b(x_i)$ $(i = 1, \ldots, n)$ *for the treatment effect at* $b(x)$, *namely*

PS in Linear Models

- It requires a very strong assumption:
  the conditional expectation of response is linear in the propensity score
  (in addition to other assumptions)

- This may present serious issues:
  - Propensity score is usually estimated with logistic or probit regression
  - Misspecified functional form of predictors may introduce bias
  - It tends to be ignored in practice

The Proposition

Proposition

Covariance adjustment using propensity score in linear regression with one covariate.
Suppose treatment assignment is strongly ignorable and the conditional expectation of the continuous response has a linear relationship with the treatment indicator and some unknown function of a continuous covariate.

$$E[y_i|T_i, x_i] = \beta_0 + \beta_1 T_i + \beta_2 f(x_i)$$

Then running covariance adjustment replacing $f(x)$ with propensity score $p(x)$ produces biased treatment effect estimator and the bias depends on the correlation between T, $f(x)$ and $p(x)$:

$$E(\hat{\beta}_1^*) - \beta_1 = \frac{1}{1 - r_{12^*}^2} \sqrt{\frac{\sum[f(x_i) - \bar{f}(x_i)]^2}{\hat{p}(1-\hat{p})}} (r_{12} - r_{12^*} r_{22^*}) \beta_2$$

The Proposition

Where
$\hat{p}$ is the proportion of treated subjects
r_{12} is the correlation between T and $f(x)$
r_{12*} is the correlation between T and propensity score
r_{22*} is the correlation between $f(x)$ and propensity score

The Proposition

Implied by the proposition,

- The covariate is not predictive of the response given the treatment. This implies $\beta_2 = 0$ and the bias is zero. If the covariate is not associated with the response conditioning on the treatment assignment, such covariate is not a confounder.

- Treatment assignment is independent of the covariate. This implies $r_{12} = r_{12*} = 0$, hence the bias is zero. It is just like a randomized experiment.

- The covariate and the propensity score has a perfect correlation. This implies $r_{22^*} = 1$ and $r_{12} = r_{12^*}$, hence the bias is zero. The equality might not occur exactly in practice, but when the correlation is sufficiently large, the bias in covariance adjustment with propensity score might be negligible.

Corollaries

Similar results hold when there are multiple covariates.

Corollary

Covariance adjustment using propensity score in linear regression with multiple covariates.
Suppose treatment assignment is strongly ignorable and the conditional expectation of the continuous response given the treatment indicator and the covariates is linear.

$$E[y_i | T_i, x_{1i}, \cdots, x_{ki}] = \beta_0 + \beta_1 T_i + \beta_2 x_{1i} + \cdots + \beta_{k+1} x_{ki}$$

Then running covariance adjustment replacing x's with propensity score p(x) produces biased treatment effect estimator.

Corollaries

Corollary

Covariance adjustment using the logit of the propensity score in linear regression with multiple covariates.
Suppose treatment assignment is strongly ignorable and the conditional expectation of the continuous response given the treatment indicator and the covariates is linear.

$$E[y_i | T_i, x_{1i}, \cdots, x_{ki}] = \beta_0 + \beta_1 T_i + \beta_2 x_{1i} + \cdots + \beta_{k+1} x_{ki}$$

Then running covariance adjustment replacing $x_1, \cdots, x_k$ with the logit of the propensity score produces biased treatment effect estimator.

Outline

Goal of the Simulation Study

- Evaluate finite sample performance of the propensity score adjustment
- From a data driven perspective
- With various propensity score adjustment techniques
- With different response functions

Model Driven Perspective

Most empirical evaluations of propensity score methods take a model driven perspective:

- The true functional form of the propensity score model is pre-assumed
- The treatment assignment is randomly generated based on the known propensity score model
- Different propensity score adjustment strategies are compared with either correctly specified propensity scores or partially specified propensity scores

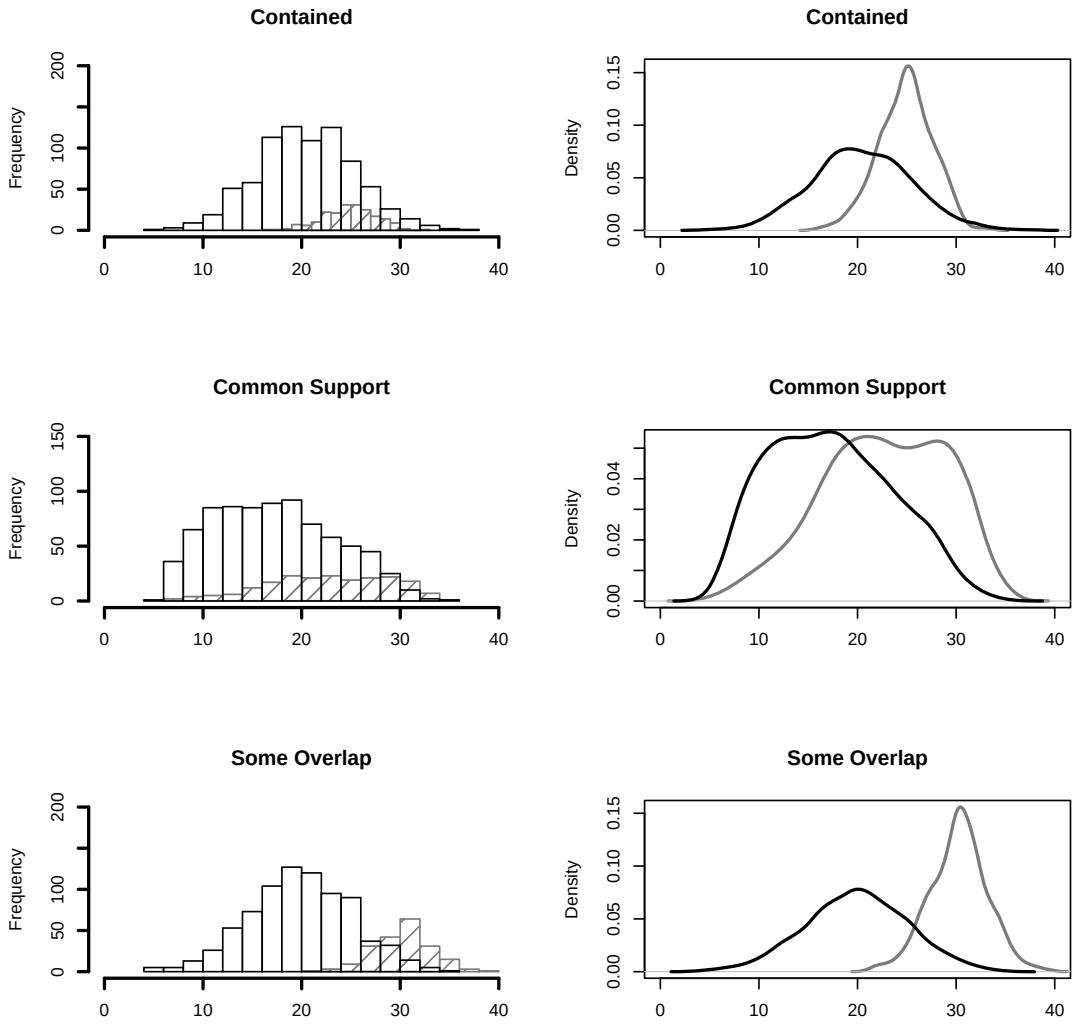
It offers theoretical insights regarding how different methods perform under the correctly/incorrectly specified propensity scores. But it pays little attention to the fact, in practice, most of the time, we have no clue what propensity score model looks like. The thing we know is the observed distribution of covariates between treated and control groups.

Data Driven Perspective

We propose to run the simulation from a data driven perspective:

- NOT assuming a known true propensity score model
- Instead, simulate the data by taking random draws from prespecified treated and control distributions separately
- The differential selection of the treatment is determined by varying overlaps between the treated and control distributions
  - Contained
  - Common Support
  - Some overlap

Three Distributional Scenarios (n=500)



Four Functional Forms of the Response Model

1. $f(X)$ is linear, $f(X) = X$
2. $f(X)$ is exponential, $f(X) = exp(X/5)$
3. $f(X)$ is quadratic, $f(X) = 3(X - 25)^2$
4. $f(X)$ is a step function that included a quadratic component:

$$f(X) = \begin{cases} (X-25)^2 + 1 & \text{if } X \leq 25 \\ (X-25) * 14 + 1 & \text{if } X < 26 \\ (X-26) * (-5) + 15 & \text{if } X < 27 \\ 10 & \text{if } X < 29 \\ (X-29) * (-5) + 10 & \text{if } X < 30 \\ (X-30) * 15 + 5 & \text{if } X < 31 \\ 20 & \text{if else} \end{cases} \quad (1)$$

Three Types of Outcome

- Continuous
  Data are generated from a linear model with true treatment effect of 10, $Y = 1 + 10T + f(X) + \epsilon$
- Binary
  Probabilities are generated from a logit model with true treatment effect of 2
- Survival
  Survival times are generated from an exponential distribution with independent censoring and true treatment effect of 2
- For each simulated scenario, 1000 iterations were run

Eight Methods Compared

Eight commonly seen methods, without model selection or parameter fine tuning

- No adjustment (No adj)
  Just include treatment indicator in regression models, no adjustment on covariates
- ANCOVA using propensity score (ANCOVA)
  Include the estimated propensity score as a linear covariate in regression models
- Propensity score stratification (PS Strata)
  Adjust for propensity score quintile in regression models
- Propensity score stratification with regression adjustment (PS Strata and Reg)
  In addition to the above, also include an interaction between propensity score quintile and X in regression models

Eight Methods Compared (cont.)

- Spline with propensity score (Spline(PS))
  Using b-spline function bs() with the default setup in R, degree=3 for cubic spline
- Propensity score weighting with regression (IPTW and Reg)
  Inverse propensity score weighted regression model, adjusting for X
- Optimal propensity score matching (Opt Match)
  Optimal propensity score pair matching, analysis in the matched set only
- Optimal propensity score matching with regression adjustment (Opt Match and Reg)
  Optimal propensity score pair matching, regression adjusting for X in the matched set only

Simulation Results for Linear Models

	No Adj	ANCOVA	PS Strata	PS Strata and Reg	Spline(PS)	IPTW and Reg	Opt Match	Opt Match and Reg
Linear								
Contained	50%	6%	1%	0%	-1%	0.3%	0.6%	-0.1%
	0.03	0.02	0.02	0.02	0.02	0.03	0.03	0.03
Common Support	61%	-0.2%	6%	0.1%	0.1%	0.1%	4%	-0.2
	0.04	0.02	0.02	0.02	0.02	0.03	0.03	0.03
Some Overlap	100%	2%	12%	0.3%	-2%	0.3%	37%	0.3%
	0.02	0.02	0.02	0.02	0.02	0.04	0.02	0.07
Exponential								
Contained	879%	-235%	-87%	-24%	10%	-86%	39%	-8%
	0.58	0.54	0.58	0.17	0.12	1.17	0.23	0.08
Common Support	1197%	52%	29%	13%	0%	43%	299%	-24%
	0.83	0.32	0.52	0.12	0.04	0.19	0.73	0.14
Some Overlap	3913%	-257%	843%	121%	-6%	-1562%	2555%	-872%
	1.02	0.87	0.99	0.42	0.51	5.5	1.11	4.43
Quadratic								
Contained	-1236%	-449%	-82%	-10%	111%	-970%	9%	-4%
	0.46	0.42	0.20	0.06	0.27	0.66	0.07	0.04
Common Support	-1843%	23%	6%	4%	-10%	-144%	90%	-11%
	0.97	0.85	0.32	0.06	0.26	1.71	0.23	0.08
Some Overlap	-480%	-163%	248%	-23%	118%	-834%	788%	-642%
	0.42	0.25	0.29	0.09	0.38	2.03	0.33	1.67
Step								
Contained	25%	4%	-0.3%	-0.1%	1%	12%	0.4%	0%
	0.02	0.02	0.02	0.02	0.02	0.03	0.02	0.02
Common Support	31%	0.4%	4%	1%	0.3%	2%	4%	-0.7%
	0.03	0.02	0.02	0.02	0.02	0.03	0.03	0.03
Some Overlap	50%	-2%	7%	2%	-2%	-0.3%	21%	-27%
	0.02	0.02	0.02	0.03	0.02	0.04	0.02	0.08

Simulation Results for Linear Models

	ANCOVA	Spline(PS)	IPTW and Reg	Opt Match	Opt Match and Reg
Exponential					
Contained	-235%	10%	-86%	39%	-8%
	0.54	0.12	1.17	0.23	0.08
Common Spt	52%	0%	43%	299%	-24%
	0.32	0.04	0.19	0.73	0.14
Some Overlap	-257%	-6%	-1562%	2555%	-872%
	0.87	0.51	5.5	1.11	4.43
Quadratic					
Contained	-449%	111%	-970%	9%	-4%
	0.42	0.27	0.66	0.07	0.04
Common Spt	23%	-10%	-144%	90%	-11%
	0.85	0.26	1.71	0.23	0.08
Some Overlap	-163%	118%	-834%	788%	-642%
	0.25	0.38	2.03	0.33	1.67

Simulation Results for Logistic Models

	No Adj	ANCOVA	PS Strata	PS Strata and Reg	Spline(PS)	IPTW and Reg	Opt Match	Opt Match and Reg
Linear								
Contained	65%	17%	10%	13%	11%	11%	17%	18%
	0.03	0.03	0.03	0.03	0.03	0.03	0.04	0.04
Common Support	80%	6%	12%	7%	6%	5%	17%	13%
	0.04	0.03	0.03	0.03	0.03	0.03	0.04	0.04
Some Overlap	159%	23%	27%	16%	14%	65%	59%	15%
	0.05	0.04	0.04	0.04	0.04	0.14	0.04	0.04
Exponential								
Contained	69%	11%	9%	10%	8%	8%	7%	9%
	0.03	0.02	0.02	0.02	0.02	0.03	0.02	0.02
Common Support	67%	4%	9%	5%	4%	5%	4%	7%
	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.03
Some Overlap	210%	10%	31%	15%	13%	40%	71%	14%
	0.108	0.03	0.04	0.04	0.04	0.08	0.05	0.04
Quadratic								
Contained	-79%	-55%	-17%	6%	16%	-60%	-7%	-6%
	0.003	0.01	0.02	0.02	0.03	0.01	0.02	0.02
Common Support	-71%	14%	16%	11%	13%	-4%	-6%	11%
	0.004	0.03	0.02	0.03	0.03	0.02	0.02	0.02
Some Overlap	-27%	-28%	51%	15%	102%	-68%	282%	5%
	0.01	0.02	0.04	0.03	0.09	0.02	0.09	0.03
Step								
Contained	-77%	-38%	-8%	2%	9%	-45%	-4%	-2%
	0.003	0.01	0.02	0.02	0.02	0.01	0.02	0.02
Common Support	-76%	23%	0.2%	6%	6%	3%	-16%	9%
	0.004	0.04	0.02	0.02	0.02	0.03	0.01	0.06
Some Overlap	-67%	-10%	32%	8%	49%	-37%	77%	6 %
	0.005	0.02	0.03	0.03	0.05	0.03	0.04	0.03

Simulation Results for Logistic Models

	ANCOVA	Spline(PS)	IPTW and Reg	Opt Match	Opt Match and Reg
Exponential					
Contained	11%	8%	8%	7%	9%
	0.02	0.02	0.03	0.02	0.02
Common Spt	4%	4%	5%	4%	7%
	0.02	0.02	0.02	0.02	0.03
Some Overlap	10%	13%	40%	71%	14%
	0.03	0.04	0.08	0.05	0.04
Quadratic					
Contained	-55%	16%	-60%	-7%	-6%
	0.01	0.03	0.01	0.02	0.02
Common Spt	14%	13%	-4%	-6%	11%
	0.03	0.03	0.02	0.02	0.02
Some Overlap	-28%	102%	-68%	282%	5%
	0.02	0.09	0.02	0.09	0.03

Simulation Results for Cox PH Models

	No Adj	ANCOVA	PS Strata	PS Strata and Reg	Spline(PS)	IPTW and Reg	Opt Match	Opt Match and Reg
Linear								
Contained	146%	12%	-13%	4%	4%	3%	-26%	4%
	0.03	0.01	0.01	0.01	0.01	0.01	0.01	0.02
Common Support	153%	5%	10%	4%	4%	4%	-32%	4%
	0.03	0.01	0.02	0.01	0.01	0.01	0.01	0.02
Some Overlap	1128%	8%	42%	4%	4%	6%	192%	4%
	0.19	0.02	0.02	0.02	0.02	0.02	0.05	0.02
Exponential								
Contained	51%	5%	1%	3%	2%	-2%	0%	3%
	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Common Support	57%	2%	9%	3%	2%	2%	-2%	4%
	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Some Overlap	165%	7%	18%	5%	3%	5%	54%	2%
	0.03	0.02	0.02	0.02	0.02	0.03	0.02	0.02
Quadratic								
Contained	-34%	-15%	-2%	3%	6%	-16%	3%	3%
	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.02
Common Support	-50%	-8%	-0.3%	4%	3%	-4%	0.2%	6%
	0.01	0.01	0.01	0.02	0.01	0.02	0.02	0.02
Some Overlap	-17%	-12%	12%	7%	12%	-10%	33%	-1%
	0.01	0.01	0.01	0.02	0.02	0.4	0.02	0.02
Step								
Contained	-34%	-11%	0.2%	3%	7%	-11%	4%	5%
	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.02
Common Support	-51%	-8%	3%	3%	2%	-5%	-2%	5%
	0.01	0.01	0.01	0.02	0.01	0.02	0.02	0.02
Some Overlap	-32%	-6%	7%	7%	10%	9%	10%	5%
	0.01	0.01	0.02	0.02	0.02	0.4	0.02	0.02

Outline

Findings

- Even for linear models, propensity score adjustment methods could be biased
- The overlap of covariate distributions matters
  Compared to contained/common support scenarios, some overlap scenario tends to produce a lot bias in almost all analyses.
- Matching works well for contained scenario
  Since it is easier to find high quality matched pairs
- Regression based method/Stratification work better for common support scenario
  The extrapolation is bounded in a "reasonable" range

Further thoughts

- When there is not much overlap of the distributions, we need to be extremely careful with using any propensity score adjustment technique.
  Some balance test between the treatment and control groups would be very helpful (Rosenbaum 2005; Hansen and Bower, 2008; etc.)
- The performance of spline is generally pretty good.
- The performance of IPTW estimator is under expectation.
  - When both response model and propensity score model are not correctly specified, the results could be seriously biased.
  - The doubly robustness property might not work well for small samples
  - An ongoing simulation study is looking at the relative performance of those methods, when the treatment assignment probability is generated from a known model.

Thank you!
Questions ?

References

- Rosenbaum, P. and Rubin, D. B. (1983), "The central role of the propensity score in observational studies for causal effects", *Biometrika*, 70: 41-55.
- Wooldridge, J. (2002), "Econometric Analysis of Cross Section and Panel Data", *The MIT Press*.
- Rosenbaum, P.R. (2005). "An Exact Distribution-Free Test Comparing Two Multivariate Distributions Based on Adjacency", *Journal of the Royal Statistical Society B*, 67: 515-530.
- Austin, P.C., Grootendorst, P., Sharon-Lise, N., and Anderson, G.M. (2007) "Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: A Monte Carlo study", *Statistics in Medicine*, 26: 754-768.
- Hansen, B.B and Bowers, J. (2008), "Covariate Balance in Simple, Stratified and Clustered Comparative Studies", *Statistical Science*, 23: 219-236.
- Lu, B., Greevy, R., Xu, X., and Beck, C. (2011) "Optimal Nonbipartite Matching and Its Statistical Applications", *The American Statistician*, 65: 21-30.