

Optimal estimation of treatment effects in the presence of confounding

David A. Stephens

Department of Mathematics and Statistics, McGill University

`d.stephens@math.mcgill.ca`



www.math.mcgill.ca/dstephens/ISM-Talk.pdf

The Role of the Research Statistician

- To find problems motivated by research in science and engineering;
- To formulate mathematical and probabilistic versions of these problems;
- To devise methods (theoretical, computational) for optimal use of experimental data;
- To study these methods theoretically;
- To study these methods computationally;
- To apply these methods in real practical settings.

The Need for Adjustment Methods

In many complex data collection contexts

- biostatistics
- econometrics
- psychology
- social sciences

there is difficulty in identifying the “true” effect of a variable of interest (a **treatment**, **exposure** or intervention-related variable) due to the inter-relation with other variables.

This is true in

- **Experimental studies:** where exposure is determined by the experimenter, but there is imperfect compliance or concordance depends on other factors.
- **Observational studies:** exposure/treatment received depends on other factors.

Both types of studies may also be prone to informative dropout.

BIOMETRICS 48, 479–495
June 1992

Estimating Exposure Effects by Modelling the Expectation of Exposure Conditional on Confounders

James M. Robins and Steven D. Mark

Harvard School of Public Health, 665 Huntington Avenue,
Boston, Massachusetts 02115, U.S.A.

and

Whitney K. Newey

Department of Economics, Massachusetts Institute of Technology,
Cambridge, Massachusetts 02139, U.S.A.

Example: The Six Cities Study

For concreteness, we shall attempt to estimate the effect of being a current cigarette smoker on the level of forced expiratory volume in one second (FEV1) in a cohort of 2,713 adult white male former and current cigarette smokers from the initial cross-sectional data collected in the Harvard Six Cities Study (Dockery et al., 1988). We shall estimate this effect while adjusting for the presence of the 22 potential confounding factors listed in Table 1 that include past smoking history, past respiratory symptoms, age, height, and coexistent heart disease. In this example the exposure of interest is dichotomous and we

Example: The Six Cities Study

Table 1

Twenty-two potential confounders of the effect of current smoking on FEV1

Age	History of emphysema
Age-squared	Past history of asthma
Height	Current asthma
Body mass index	Former cigar smoker
Chronic cough	Current cigar smoker level = hi
Recurrent bouts of coughing	Current cigar smoker level = medium
History of treatment for heart disease	Current cigar smoker level = lo
Chronic phlegm production	Former pipe smoker
Chronic wheeze	Current pipe smoker level = hi
Total years of cigarette smoking	Current pipe smoker level = medium
Lifetime pack-years smoked	Current pipe smoker level = lo

- A treatment variable, D , taking values in $\mathcal{D}$:
 - ▶ binary, $\mathcal{D} \equiv \{0, 1\}$.
 - ▶ categorical $\mathcal{D} \equiv \{d_0, d_1, \dots, d_k\}$.
 - ▶ continuous $\mathcal{D} \subseteq \mathbb{R}$.
- Response Y (binary, categorical, continuous)
- Covariates X

We wish to establish the extent to which changes in D would induce changes in Y **in the same individual**; we will define this as the **causal** influence of D on Y .

Inferential Objectives: Terminology

The usual statistical objective is to produce an estimator T_n of parameter θ that is

- **Unbiased**: $\mathbb{E}[T_n] = \theta$
- **Asymptotically unbiased/consistent**: As $n \rightarrow \infty$.

$$\mathbb{E}[T_n] \rightarrow \theta \quad \text{or} \quad T_n \xrightarrow{p} \theta$$

The **bias** of T_n is

$$b_{T_n}(\theta) = \mathbb{E}[T_n] - \theta$$

or its asymptotic limit.

- **Efficient**: $\text{Var}[T_n]$ is as small as possible, especially as $n \rightarrow \infty$.

We have the standard decomposition of the **Mean-Square Error** (MSE)

$$MSE_{T_n}(\theta) = \mathbb{E}[(T_n - \theta)^2] = \{b_{T_n}(\theta)\}^2 + \text{Var}[T_n]$$

via which to assess our inference.

- Literature mainly focuses minimizing the variance of **consistent** estimators:
 - ▶ This is pragmatic statistical **theoretical** position...
 - ▶ ... procedures that focus on minimizing MSE are difficult to construct and study (in the frequentist world).
- Most of the time, researchers want to quote standard errors, confidence intervals, test results, so this is also a pragmatic **practical** position.
- Admitting **some** bias may allow lower MSE.
- For finite n , all bets are off (difficult to study the finite sample bias etc.); we just hope for procedures that are **stable**.

Regression Formulation: Additive model

Suppose in our study that D influences Y via the model

$$Y = \beta D + g(X; \alpha) + \epsilon$$

where $g(.; .)$ is a flexible function, and $\mathbb{E}[\epsilon|D, X] = 0$. Here, β is the causal **parameter of interest**; α is a **nuisance** parameter.

Suppose that

$$Y = \beta D + \sum_{k=0}^K \alpha_k g_k(X_i) + \epsilon.$$

Then the ordinary least-squares (OLS) estimator is

$$\hat{\beta}^{(A)} = \frac{\sum_{i=1}^n Y_i (D_i - \hat{D}_i)}{\sum_{i=1}^n D_i (D_i - \hat{D}_i)}$$

whereas $\hat{D}_i$ is the fitted value obtained from regressing D on $g(X)$.

Alternatively

Multiply through by $D - \mathbb{E}[D|X] = D - \pi(X)$, take expectations

$$\begin{aligned}\mathbb{E}[Y(D - \pi(X))] &= \beta \mathbb{E}[D(D - \pi(X))] + \mathbb{E}[(D - \pi(X))g(X; \alpha)] \\ &\quad + \mathbb{E}[\epsilon(D - \pi(X))] \\ &= \beta \mathbb{E}[D(D - \pi(X))]\end{aligned}$$

and so

$$\beta = \frac{\mathbb{E}[Y(D - \pi(X))]}{\mathbb{E}[D(D - \pi(X))]}$$

implies the moment-based estimator

$$\hat{\beta}^{(B)} = \frac{\sum_{i=1}^n Y_i(D_i - \hat{\pi}(X_i))}{\sum_{i=1}^n D_i(D_i - \hat{\pi}(X_i))}$$

By standard results

- $\hat{\beta}^{(A)}$ is consistent and efficient if $g(X; \alpha)$ is correctly specified.
- $\hat{\beta}^{(B)}$ is consistent and efficient if $\pi(X)$ is correctly specified.
- $\hat{\beta}^{(A)}$ and $\hat{\beta}^{(B)}$ are identical if D is modelled using a linear model.

Typically

$$\pi(X) = \mathbb{E}[D|X] = g_B(X; \alpha)$$

using a parametric specification.

So, which estimator you choose depends on whether you can model

$$g(X; \alpha) \quad \text{or} \quad g_B(X; \alpha)$$

more confidently.

Our causal parameter of interest is an expected difference in outcomes when the experimenter **intervenes** to change the treatment received.

Let $Y(d)$ denote the response variable when treated at $D = d$. In the causal literature, $Y(d)$ is called **counterfactual** outcome.

Parameters of interest then are

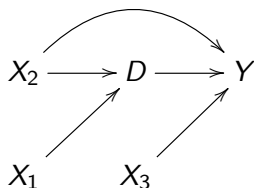
- for binary treatment: the **average treatment effect** (ATE) is

$$\mu = \mathbb{E}[Y(1) - Y(0)]$$

- for categorical and continuous treatments: $\mu(d) = \mathbb{E}[Y(d)]$

Note that the expectation is taken over the **joint distribution** of $(Y(d), X)$, for each fixed d .

Three types of covariates



- I covariates which are just related to the treatment (X_1)
- II covariates which are related to the outcome as well as the treatment (confounders) (X_2)
- III those which are just related to the outcome variable (X_3)

Under the assumption of **no unmeasured confounders**

$$Y(d) \perp D | X$$

where $X = (X_1, X_2, X_3)$, the ATE can be computed

$$\begin{aligned}\mu &= \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}_X[\mathbb{E}[Y(1)|X] - \mathbb{E}[Y(0)|X]] \\ &= \mathbb{E}_X[\mathbb{E}[Y|D = 1, X] - \mathbb{E}[Y|D = 0, X]]\end{aligned}$$

Confounding and Balancing

The quantity $\pi(X) = \Pr(D = 1|X)$ is termed the **propensity score**, and it follows that

$$D \perp X | \pi(X)$$

that is, at a fixed value of scalar π , D and X are uncorrelated; $\pi(X)$ is a **balancing score**, and we have unconfoundedness given π

$$Y(d) \perp D | \pi(X).$$

Hence the equivalent result

$$\mu = \mathbb{E}_\pi[\mathbb{E}[Y|D = 1, \pi] - \mathbb{E}[Y|D = 0, \pi]]$$

follows, where the internal expectation can be estimated from the observed data.

The Propensity Score

$\pi(X)$ defined in this way by Rosenbaum and Rubin (1983) is the **coarsest** balancing score; if $T(X)$ is any other balancing score

$$D \perp X | T(X)$$

then $\pi(X)$ is a function of $T(X)$.

Remark: RR83 refer to X as “covariates” (the implication seems to be “all observed covariates”). This interpretation persists in the literature.

However, it is evident that X can be reduced to contain only those predictors that are **confounders**, that is, we can use

$$\pi(x_2) = \Pr(D = 1 | X_2 = x_2).$$

This is still a balancing score with respect to the confounders. For whatever score T we use, we require the validity of the step

$$\mathbb{E}[Y(d) | T] = \mathbb{E}[Y | D = d, T].$$

Propensity Score Regression

Propensity Score Regression (PSR) is the two-step strategy;

- (i) build a model for conditional expectation D given X , $\pi(X)$,
- (ii) build a regression model for Y given D and $\hat{\pi}(X)$.

$$Y_i = \mu(D_i, \pi_i, \beta) + \epsilon_i, \quad \mathbb{E}[\epsilon_i | D_i, \pi_i] = 0$$

We need to

- select the form of each model,
- decide **which covariates** enter into each component.

Typically $\pi(X)$ is specified via a finite dimensional parameter α .

Scenario 1: Including Type III variable in the PS model

True model: Suppose $D|X \sim \text{Bernoulli}(\pi(X))$, with

$$\text{logit } \pi(x) = \alpha_0 + \alpha_2 x_2$$

and that

$$\mathbb{E}[Y|D = d, X = x] = \beta_0 + \beta d + \beta_2 x_2 + \beta_3 x_3,$$

Fitted model: Assume $D|X \sim \text{Bernoulli}(\pi(X))$ with

$$z_{2,3} = \text{logit } \pi(x_2, x_3) = \alpha_0 + \alpha_2 x_2 + \alpha_3 x_3$$

and that

$$\mathbb{E}[Y|D = d, X = x] = \beta_0 + \beta d + \theta \hat{z}_{2,3},$$

The presence of X_3 in the fitted PS model increases the covariance between $Z_{2,3}$ and Y , which **reduces** the standard error of $\hat{\beta}$.

It **does not** increase the covariance between $Z_{2,3}$ and D relative to the model with X_3 omitted.

Scenario 2: Including Type I variable in the PS model

True model: Suppose $D|X \sim \text{Bernoulli}(\pi(X))$, with

$$\text{logit } \pi(x) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2$$

and that

$$\mathbb{E}[Y|D = d, X = x] = \beta_0 + \beta d + \beta_2 x_2$$

Fitted model 1: Assume $D|X \sim \text{Bernoulli}(\pi(X))$ with

$$z_{1,2} = \text{logit } \pi(x_1, x_2) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2$$

and that

$$\mathbb{E}[Y|D = d, X = x] = \beta_0 + \beta d + \theta \hat{z}_{1,2}$$

Fitted model 2: Assume $D|X \sim \text{Bernoulli}(\pi(X))$ with

$$z_2 = \text{logit } \pi(x_2) = \alpha_0 + \alpha_2 x_2$$

and

$$\mathbb{E}[Y|D = d, X = x] = \beta_0 + \beta d + \theta \hat{z}_2,$$

Scenario 2: Including Type I variable in the PS model

Consider estimation of β using the two models

$$Y = \beta_0 + \beta d + \theta \widehat{z}_{1,2} + \epsilon \quad (1a)$$

$$Y = \beta_0 + \beta d + \theta \widehat{z}_2 + \epsilon \quad (1b)$$

Both models yield consistent estimators of β .

For each model, and for large n , the variance of $\widehat{\beta}$ simplifies to

$$\frac{\sigma^2}{n\text{Var}[D](1 - \text{Corr}(D, Z)^2)}$$

which is a monotonic **increasing** function of $|\text{Corr}(D, Z)|$; X_1 is a predictor of D , so $\widehat{Z}_{1,2}$ is more correlated with D than is $\widehat{Z}_2$, and hence the variance of $\widehat{\beta}$ is **larger** when model (1a) is fitted.

The extent of the difference in correlation depends on the magnitude of α_1 .

Propensity Score Regression (PSR): Model construction

Key facts: Omitting important confounders from the PS model results in an inconsistent estimator; including Type-III covariate in the PS model results in an efficiency gain; including Type-I covariates in the PS model inflates the variance of the estimator.

What strategy should we adopt ?

- Ignore the response and apply a covariate selection technique for the PS model. **This can include useless Type I variables, and ignores the Type III covariates altogether.**
- Fit a regression mean model by considering the treatment and the covariates as regressors and perform a covariate selection technique. Include the selected covariates in the PS model. **Type III covariates can dominate the selection process.**

To avoid these shortcomings, we need to introduce a covariate selection technique based on the **penalized likelihood** which utilizes response and treatment assignment models simultaneously.

Penalized Estimators

Maximum penalized likelihood estimator (MPLE) given by

$$\hat{\eta}_{ml} = \arg \min_{\eta} \left\{ l_m(\eta) + n \sum_{j=1}^p p_{\lambda_n}(|\eta_j|) \right\},$$

where $l_m(\eta)$ is the minus log-likelihood and $p_{\lambda}(\cdot)$ is a penalty function.

The most well-known penalty functions are:

- LASSO (Tibshirani (1996)): $p_{\lambda}(|\beta_j|) = \lambda|\beta_j|$.
- SCAD (Fan & Li (2001)):

$$p'_{\lambda}(|\beta_j|) = \lambda \left\{ I(|\beta_j| < \lambda) + \frac{(a\lambda - |\beta_j|)_+}{(a-1)\lambda} I(|\beta_j| > \lambda) \right\}$$

for $a > 2$.

- The joint density of (Y, D, π) can be written as

$$f(y, d, x; \eta) = f(y_i | d_i, \pi_i, g(x_i; \alpha), \beta) \\ p(D = 1 | x_i, \alpha)^{d_i} p(D = 0 | x_i, \alpha)^{1-d_i}$$

- The joint penalized (pseudo) density of (Y, D, π) can be defined as

$$f_p(z_i, \eta) = f(y_i, d_i, \pi_i; \eta) f(\alpha)$$

where $\eta = (\beta, \alpha)$, and $f(\alpha) = \exp\{-np_{\lambda_n}(\alpha)\}$.

- Only penalize components of α .

Penalized estimation strategy

1. Fit a saturated PS model, $\hat{\pi}(X)$.
2. Define $S_i = D_i - \hat{\pi}(X)$ and fit the conditional mean model

$$\mathbb{E}[Y_i|D = d_i, X = x_i] = \beta_0 + \beta s_i + \theta g(x_i, \alpha)$$

where $g(x; \alpha)$ is a flexible function of x , linear in α say.

3. Estimate the vector of parameters $\eta = (\beta, \alpha)$ such that

$$\prod_{i=1}^n f_p(y_i, s_i, \pi_i; \hat{\eta}_n) = \sup_{\eta} \prod_{i=1}^n f_p(y_i, s_i, \pi_i; \eta), \quad (2)$$

where $f_p(\cdot)$ is the joint penalized density of (Y, S, π) .

4. Take the covariates corresponding to non-zero α 's, fit a PS model and use the model in 2. to estimate β .

Properties of the penalty functions

A good penalty function should result in estimators with three properties:

1. **Asymptotic unbiasedness**
2. **Sparsity**: small estimated coefficients driven to zero.
3. **Continuity**: continuous in data to avoid instability in model prediction.

LASSO estimator exhibits 2. and 3.; SCAD estimator has all of them.

The LASSO shrinkage estimator has more bias but less variation. So the estimates are not consistent (in general); could

- run a LASSO, identify the non-zero set and then fit the LS on those.
- run a LASSO, identify the non-zero set and then fit the LASSO again.

Required Assumptions for a Penalty Function

- P1.** For all n , $p_{\lambda_n}(0) = 0$, $p_{\lambda_n}(\alpha) \geq 0$ and $p_{\lambda_n}(\alpha)$ is symmetric; it is twice differentiable with derivatives $p'_{\lambda_n}(\cdot)$ and $p''_{\lambda_n}(\cdot)$ and nondecreasing for $\alpha \neq 0$ with at most a few exceptions.

P1 is used to prove consistency;

- P2.** As $n \rightarrow \infty$,

$$\max_{\alpha \neq 0} \{p''_{\lambda_n}(\alpha)\} \rightarrow 0 \quad \text{and} \quad \max_{\alpha \neq 0} \{\sqrt{np}'_{\lambda_n}(\alpha)\} \rightarrow 0.$$

P2 prevents the j th element of the penalized likelihood from being dominated by the penalty function since it vanishes when $n \rightarrow \infty$.

- P3.** For $N_n \equiv (0, M_n)$, where $M_n \rightarrow 0$ as $n \rightarrow \infty$,

$$\liminf_{\alpha \in N_n} \sqrt{np}'_{\lambda_n}(\alpha) = \infty.$$

If $\alpha_j = 0$, P3 allows the penalty function to dominate the penalized likelihood which leads to sparsity.

Asymptotic Properties

LASSO: The sparsity and consistency require $\sqrt{n}\lambda_n \rightarrow 0$ and $\sqrt{n}\lambda_n \rightarrow \infty$ which can not be satisfied simultaneously.

SCAD: The sparsity and consistency properties can be achieved if $\sqrt{n}\lambda_n \rightarrow \infty$ and $\lambda_n \rightarrow 0$, respectively. **SCAD exhibits these oracle properties**

Theorem (Asymptotic Normality)

If $\det I(\eta) \neq 0$ for $\eta \in \Xi$, then

$$\frac{1}{\sqrt{n}} \left\{ \sum_{i=1}^n \left[\frac{\partial \ln f(z, \eta_{t1})}{\partial \eta_{t1}} \right] - np'_{\lambda_n}(\alpha_{t1}) \right\} \xrightarrow{d} \text{Normal}(0, I(\eta_{t1}))$$

where $\eta_{t1} = (\beta, \alpha_{t1})$ and $\alpha_{t1} = (\alpha_1, \dots, \alpha_s)$.

If $\lim_{n \rightarrow \infty} \sqrt{n}p'_{\lambda_n}(\alpha) = 0$, the MPL estimator behaves similar to the ML estimator.

Simulation study: 1000 replicate data sets

We assume the following model:

$$D \sim \text{Bernoulli} \left(\frac{\exp\{0.2x_1 - x_2 + 3x_8 - x_9 + x_{10}\}}{1 + \exp\{0.2x_1 - x_2 + 3x_8 - x_9 + x_{10}\}} \right)$$

with

$$Y \sim \text{Normal}(d + 2x_1 + 0.5x_2 + 5x_3 + 5x_4, \sigma^2)$$

where X_k for $k = 1, \dots, 10$ are normal random variables each with mean $\mu = 1$, variance σ^2 .

Table below shows the results for different values of σ and ρ where $\rho = \text{Corr}(X_i, X_j)$ for $i \neq j$.

- **Oracle** refers to the model which includes just covariates related to the response in the propensity score model (that is, $x_1, \dots, x_4$)
- **Saturated** refers to the propensity score model including all the covariates ($x_1, \dots, x_{10}$).

Simulation study: ATE estimators

		$n = 500$			$n = 5000$		
$\sigma = 2$	ρ	Bias	S.D.	MSE	Bias	S.D.	MSE
SCAD	0	0.012	0.230	0.053	0.000	0.065	0.004
	0.5	0.003	0.233	0.055	0.005	0.064	0.004
LASSO	0	0.003	0.254	0.065	0.002	0.065	0.004
	0.5	0.023	0.246	0.061	0.005	0.064	0.004
Oracle	0	0.006	0.195	0.038	0.000	0.065	0.004
	0.5	0.010	0.187	0.035	0.001	0.064	0.004
Saturated	0	0.001	0.326	0.106	0.002	0.123	0.015
	0.5	0.031	0.329	0.109	0.003	0.125	0.016
$\sigma = 4$							
SCAD	0	0.025	0.541	0.290	0.006	0.127	0.016
	0.5	0.009	0.537	0.288	0.007	0.124	0.016
LASSO	0	0.055	0.618	0.385	0.003	0.127	0.016
	0.5	0.090	0.562	0.316	0.008	0.124	0.015
Oracle	0	0.021	0.387	0.150	0.008	0.127	0.016
	0.5	0.012	0.372	0.139	0.002	0.121	0.015
Saturated	0	0.039	0.899	0.810	0.008	0.332	0.111
	0.5	0.026	0.902	0.814	0.014	0.343	0.118

Results

- For moderate sample sizes, the covariate selection technique results in a lower MSE than the saturated model by decreasing both bias and standard error.
- The performance of LASSO and SCAD are very similar for large sample sizes.
- For moderate sample sizes SCAD outperforms LASSO in terms of overall fit for the both noise levels.
- The proposed technique converges to the ORACLE model as sample size increases.

The National Supported Work Demonstration (NSWD)

The NSWD aimed to provide work experience to individuals who had lack of basic skills.

- A total of 2434 workers were sampled at random from different cities across the US by the Manpower Demonstration Research Corporation from 1975 to 1980
- Qualified individuals were assigned to receive training intervention (230 treated).

- Predictors are

X_1 : Age

X_2 : Education

X_3 : African American

X_4 : Hispanic

X_5 : Married

X_6 : No degree

X_7 : Earning in 1975

- Based on PS values, sample reduced to 551 individuals (335 not treated, 216 treated); $\sim 40\%$ of untreated had $\hat{\pi}(x) < 10^{-3}$.

Method	ATE	S.D.	C.I.	X_1	X_2	X_3	X_4	X_5	X_6	X_7
SCAD	-0.552	0.091	(-0.734,-0.370)		•		•	•	•	
LASSO	-0.565	0.094	(-0.753,-0.377)		•		•	•	•	
Saturated	-0.502	0.107	(-0.716,-0.288)	•	•	•	•	•	•	•
Unadjusted	-1.376	0.072	(-1.520,-1.232)							

- The penalized covariate selection technique keeps “Education”, “Hispanic”, “Married Status” and “No degree” as the only significant confounders in the propensity score model.
- 27% reduction in variance compared with the saturated model.

Summary: Key Results

In the standard causal setting, under the usual assumptions:

1. The propensity score (for balancing, inverse probability weighting, propensity score regression) can be constructed from **predictors of outcome**, not predictors of treatment, and will still be a balancing score.
2. Pure predictors of treatment should be **excluded**. Including them will **increase the variance and MSE** of the resulting effect estimator.
3. In constructing the propensity score, **you cannot ignore outcome**.
4. **Penalized regression**-based estimation seems a good way to proceed for choosing the covariates that enter the models.
5. Model/variable selection should be carried out in **both component models simultaneously**.

Optimality of adjustment methods.

The propensity score is also used in weighting-based estimators: the **inverse probability of treatment weighted** (IPTW) estimator for binary doses is

$$\hat{\beta} = \sum_{i=1}^n \frac{D_i(Y_i - \mu_1(X_i))}{\pi(X_i)} - \sum_{i=1}^n \frac{(1 - D_i)(Y_i - \mu_0(X_i))}{1 - \pi(X_i)}$$

- **first** term is from those who **received** treatment
- **second** term is from those who **did not receive** treatment
- weights

$$\frac{1}{\pi(X_i)} \quad \frac{1}{1 - \pi(X_i)}$$

create a pseudo-population in which treatment is not confounded with X .

Variance of IPTW estimators

Hahn (1998) provides a **semiparametric variance bound** for IPTW estimators for binary dose case; corrected by Hirano et al. (2003), this bound is

$$\mathbb{E}_X \left[\frac{\sigma_1^2(X)}{\pi(X)} + \frac{\sigma_0^2(X)}{1 - \pi(X)} + (\mu(X) - \mu)^2 \right]$$

where the expectation is over the distribution of X , for $d = 0, 1$
 $\sigma_d^2(X) = \text{Var}[Y(d)|X]$, and

$$\mu(X) = \mathbb{E}[Y(1) - Y(0)|X].$$

Note that in the homoscedastic case, where $\sigma_0^2(X) = \sigma_1^2(X) = \sigma^2$ say, then the bound is

$$\mathbb{E}_X \left[\frac{\sigma^2}{\pi(X)(1 - \pi(X))} + (\mu(X) - \mu)^2 \right]$$

Variance of IPTW estimators

How do we compute this ?

- consider the relevant Hilbert space of estimating functions $\{\varphi\}$ based on (Y, D, X) .
- define a suitable inner product: for two random vectors V_1, V_2

$$\langle V_1, V_2 \rangle = \text{Cov}(V_1^T, V_2^T) = \mathbb{E}[(V_1 - \mathbb{E}[V_1])^T (V_2 - \mathbb{E}[V_2])]$$

- find the length of the projection from the data space onto the subspace spanned by the estimating functions; this length is the **smallest possible variance** of estimators of β .

This is the analogue of the **Cramér-Rao variance bound** in parametric models; here we utilize a **semiparametric** framework that does not specify the probability densities, merely (say) distributional moments.

Variance of IPTW estimators

The relevant Hilbert space comprises linear combinations of (quasi)-score functions, that is, first partial derivatives of the log of a (quasi)-likelihood function.

The score function corresponding to the IPTW bound is

$$S_{\theta}(d, y, \pi|\theta) = dS_{1\theta}(y|\pi, \theta) + (1 - d)S_{0\theta}(y|\pi, \theta)$$

and the corresponding quasi-likelihood is

$$\{f_1(y|\pi, \theta)\}^d \{f_0(y|\pi, \theta)\}^{1-d}$$

Here

$$S_{d\theta}(y|\pi, \theta) = \frac{\partial \log f_d(y|\pi, \theta)}{\partial \theta}$$

Variance of PS estimators

Suppose that the causal parameter of interest $\mu(\theta)$ can be written

$$\begin{aligned}\mu(\theta) &= \int \int y f_1(y|\pi, \theta) f(\mathbf{x}) dy d\mathbf{x} - \int \int y f_0(y|\pi, \theta) f(\mathbf{x}) dy d\mathbf{x} \\ &= \int \int y f_1(y|\pi, \theta) f_\pi(\pi) dy d\pi - \int \int y f_0(y|\pi, \theta) f_\pi(\pi) dy d\pi\end{aligned}$$

where $\pi \equiv \pi(X)$ is the propensity score r.v.. Let

$$\begin{aligned}\mu_d(\pi, \theta) &= \mathbb{E}[Y(d)|\pi, \theta] \quad d = 0, 1 \\ \epsilon_d &= Y - \mu_d(\pi, \theta) \quad d = 0, 1 \\ \epsilon &= d\epsilon_1 - (1 - d)\epsilon_0\end{aligned}$$

Let

$$\mu_{d\theta}(\pi, \theta) = \frac{\partial \mu_d(\pi, \theta)}{\partial \theta}$$

Consider propensity score regression estimators: consider the score function

$$S_{\theta}(d, y, \pi|\theta) = \frac{d}{\pi} S_{1\theta}(y|\pi, \theta) + \frac{1-d}{1-\pi} S_{0\theta}(y|\pi, \theta)$$

and the corresponding (pseudo) likelihood based on

$$\{f_1(y|\pi, \theta)f_{\pi}(\pi)\}^{d/\pi} \{f_0(y|\pi, \theta)f_{\pi}(\pi)\}^{(1-d)/(1-\pi)}$$

The efficient score is this case is

$$S_{\text{eff}} = \frac{d\epsilon_1\mu_{1\theta}(\pi, \theta) - (1-d)\epsilon_0\mu_{0\theta}(\pi, \theta)}{\pi\sigma_1^2(\pi) + (1-\pi)\sigma_0^2(\pi)}.$$

The semiparametric variance bound is

$$\begin{aligned}\mathbb{V} &= \mu_{\theta}^{\text{T}}(\theta) \mathbb{E}[S_{\text{eff}}^{\text{T}} S_{\text{eff}}]^{-1} \mu_{\theta}(\theta) \\ &= \mu_{\theta}^{\text{T}}(\theta) \mathbb{E} \left[\frac{\pi\mu_{1\theta}(\pi, \theta)\mu_{1\theta}(\pi, \theta)^{\text{T}} + (1-\pi)\mu_{0\theta}(\pi, \theta)\mu_{0\theta}(\pi, \theta)^{\text{T}}}{\pi\sigma_1^2(\pi) + (1-\pi)\sigma_0^2(\pi)} \right]^{-1} \mu_{\theta}(\theta)\end{aligned}$$

Variance of PS estimators

In the simplest case, we might specify

$$\mu_d(\pi, \theta) = \theta_1 d + \theta_2 \pi$$

and assume that $\sigma_1^2(\pi) = \sigma_0^2(\pi) = \sigma^2$. Then the bound becomes

$$\frac{\sigma^2}{\mathbb{E}_X [\pi(X)(1 - \pi(X))]}$$

Note that this equals

$$\frac{\sigma^2}{\mathbb{E}_X [\text{Var}_{D|X}[D|X]]}$$

which is a well-known semiparametric variance bound

- see Chamberlain (1987), Newey (1990), Robins et al. (1992).

Variance of PS estimators: Implications

Clearly

$$\frac{\sigma^2}{\mathbb{E}_X [\pi(X)(1 - \pi(X))]} \leq \mathbb{E}_X \left[\frac{\sigma^2}{\pi(X)(1 - \pi(X))} \right]$$

so that the PS variance bound is **lower** than the IPTW variance bound.

The PS variance bound corresponds to the estimator derived from the likelihood/efficient score above.

Note also that the variance bound is the same for $\pi(X)$ as for X ; hence asymptotically, regression on π and on X have the same variance bound (in the semiparametric setting, where the functional relationship between Y and X is completely unspecified).

Variance of PS estimators: Implications

Recall the earlier result: in the regression model the variance of $\hat{\beta}$ simplifies to

$$\frac{\sigma^2}{n\text{Var}[D](1 - \text{Corr}(D, Z)^2)}$$

where Z is a balancing score. We know that $Z = \pi(X) = \mathbb{E}[D|X]$ is a balancing score, in which case

$$\text{Corr}(D, \pi(X))^2 = \frac{\text{Cov}(D, \pi(X))^2}{\text{Var}[D]\text{Var}[\pi(X)]}$$

But

$$\begin{aligned}\text{Cov}(D, \pi(X)) &= \mathbb{E}[D\pi(X)] - \mathbb{E}[D]\mathbb{E}[\pi(X)] \\ &= \mathbb{E}_X[\mathbb{E}[D|X]\pi(X)] - \mathbb{E}_X[\mathbb{E}[D|X]]\mathbb{E}[\pi(X)] \\ &= \text{Var}[\pi(X)] = \text{Var}[\mathbb{E}[D|X]]\end{aligned}$$

Variance of PS estimators: Implications

Therefore the variance is

$$\frac{\sigma^2}{n(\text{Var}[D] - \text{Var}[\mathbb{E}[D|X]])} = \frac{\sigma^2}{n\mathbb{E}[\text{Var}[D|X]]}$$

or equivalently

$$\frac{\sigma^2}{n\mathbb{E}[\pi(X)(1 - \pi(X))]}$$

which equals the variance bound we found earlier.

BUT: If Z is any other balancing score, it follows that

$$\text{Corr}(D, Z)^2 \leq \text{Corr}(D, \pi(X))^2$$

so we should be seeking balancing scores that are not equal to $\pi(X)$, as regressing on Z rather than $\pi(X)$ will further reduce the variance of $\hat{\beta}$.