Bayesian Causal Inference in Stan

Arman Oganisian

@StableMarkets

Division of Biostatistics
Department of Biostatistics, Epidemiology, and Informatics
University of Pennsylvania

Review/Tutorial paper

A Practical Introduction to Bayesian Estimation of Causal Effects: Parametric and Nonparametric Approaches

Arman Oganisian[1*] and Jason A. Roy[2]

[1] Division of Biostatistics
Department of Biostatistics, Epidemiology, and Informatics
University of Pennsylvania

[2]Department of Biostatistics and Epidemiology
Rutgers University

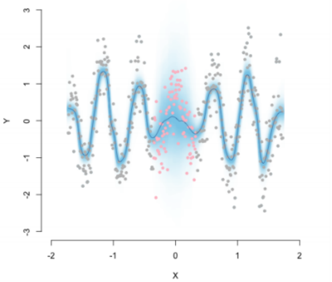
Review/Tutorial paper

A Practical Introduction to B... Effects:
Parametric and N...

Arman Ogani...

[1] Divis...
Department of Biostatis...
Univer...

[2]Department of B...
Ru...



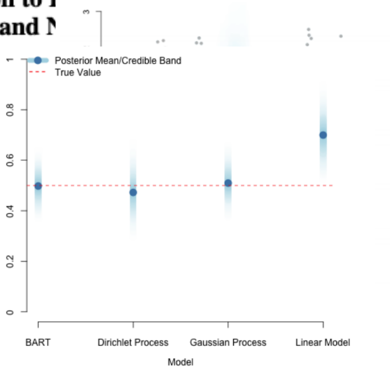
Review / Tutorial paper

A Practical Introduction to B... Effects:
Parametric and N...

Arma...

Department of

[2]Departm...



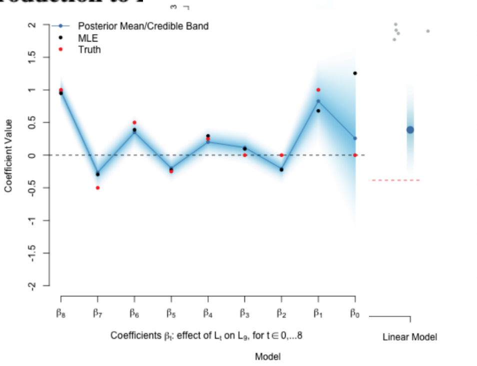
Review/Tutorial paper

A Practical Introduction to

Effects:

Par

Depart



What is Causal Inference?

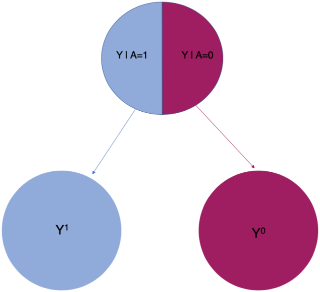
What would have happened had everyone in the target population if ...

- ... everyone took treatment 1 versus treatment 0?
- ... were vaccinated ?
- ... were enrolled in a job training program?

What is Causal Inference?

What would have happened had everyone in the target population if ...

- ... everyone took treatment 1 versus treatment 0?
- ... were vaccinated ?
- ... were enrolled in a job training program?



IDENTIFICATION VIA THE g-FORMULA

$D = \{Y_i, A_i, L_i, V_i\}_{1:n}$.
Define potential outcomes Y^a for $a \in \{0, 1\}$

Identification via the g-formula

$D = \{Y_i, A_i, L_i, V_i\}_{1:n}$.
Define potential outcomes Y^a for $a \in \{0, 1\}$

$$\Psi(v) = E[Y^1 - Y^0 \mid V = v]$$

Identification via the g-formula

$D = \{Y_i, A_i, L_i, V_i\}_{1:n}$.
Define potential outcomes Y^a for $a \in \{0, 1\}$

$$\Psi(v) = E[Y^1 - Y^0 \mid V = v]$$

Under some identification assumptions

$$E[Y^a \mid V = v] = \int_{\mathcal{L}} \underbrace{E[Y \mid A = a, V = v, L]}_{\text{Regression, } \mu(a,v,l)} \cdot \underbrace{dP(L \mid V = v)}_{\text{Confounder } P_v(L)}$$

Regression Modeling

- Parametric Approaches:

$$\mu(A, V, L) = g^{-1}(\beta_0 + \beta_1 A + \beta_2 V + \beta_3' L)$$

Need priors on βs.

- Nonparametric Approaches:

$$\mu(A, V, L) = g^{-1}(f(A, V, L))$$

Need prior for f.

Why Bayes?

- Priors can help us compute causal effects under sparsity.
- Avoid *ad hoc* approaches.
- Powerful suite of nonparametric models (BART, DP, GP, etc).
- Probabilistic sensitivity analyses.

Logistic Model For CATEs

Suppose Y is binary and $V \in \{1, 2, 3, 4, 5\}$ (e.g., race/ethnicity)

$$Y \mid A, V, L \sim Ber\big(\mu(A, V)\big)$$

Logistic Model For CATEs

Suppose Y is binary and $V \in \{1, 2, 3, 4, 5\}$ (e.g., race/ethnicity)

$$Y \mid A, V, L \sim Ber\big(\mu(A, V)\big)$$

Specify logistic regression

$$\mu(A, V, L) = g^{-1}(\beta_v + \beta_L' L + \theta_v A)$$

with parameters $\omega = (\beta_1 \ldots, \beta_5, \beta_L, \theta_1, \ldots, \theta_5)$

Prior for race effects

$$\mu(A, V, L) = g^{-1}(\beta_v + \beta_L' L + \theta_v A)$$

- Consider "partial pooling" prior:

Prior for race effects

$$\mu(A, V, L) = g^{-1}(\beta_v + \beta_L' L + \theta_v A)$$

- Consider "partial pooling" prior:

$$\theta_v \mid \theta^* \sim N(\theta^*, \phi)$$

- Shrinkage: shrinkage race effects towards common effect.
- Belief: the race effects shouldn't be *that* different.
- Causal intuition: small ϕ shrinks towards homogeneity.

Prior for race effects

$$\mu(A, V, L) = g^{-1}(\beta_v + \beta_L' L + \theta_v A)$$

- Consider "partial pooling" prior:

$$\theta_v \mid \theta^* \sim N(\theta^*, \phi)$$

- Shrinkage: shrinkage race effects towards common effect.
- Belief: the race effects shouldn't be *that* different.
- Causal intuition: small ϕ shrinks towards homogeneity.
- Note: implies

$$\theta_4 - \theta_5 \sim N(0, 2\phi)$$

As opposed to setting $\phi \approx 0$

$$\theta_4 - \theta_5 \sim \delta_0$$

Suppose we want to compute Causal Odds Ratio:

$$\Psi(v) = \frac{E(Y^1 \mid v)/[1 - E(Y^1 \mid v)]}{E(Y^0 \mid v)/[1 - E(Y^0 \mid v)]}$$

We have a data model...now what?

Suppose we want to compute Causal Odds Ratio:

$$\Psi(v) = \frac{E(Y^1 \mid v)/[1 - E(Y^1 \mid v)]}{E(Y^0 \mid v)/[1 - E(Y^0 \mid v)]}$$

Using g-computation,

$$E(Y^a \mid v) = \int_{\mathcal{L}} \mu(a, v, L) dP_v(L)$$

But what about model for $P(L)$?

The Bayesian Bootstrap

Let $S_v = \{i : V_i = v\}$ with size n_z

The Bayesian Bootstrap

Let $S_v = \{i : V_i = v\}$ with size n_z

- Frequentist estimate:

$$\hat{P}_v(l) = \sum_{j \in S_v} \frac{1}{n_v} \cdot \delta_{L_j}(l)$$

The Bayesian Bootstrap

Let $S_v = \{i : V_i = v\}$ with size n_z

- Frequentist estimate:

$$\hat{P}_v(l) = \sum_{j \in S_v} \frac{1}{n_v} \cdot \delta_{L_j}(l)$$

- Bayesian model:

$$P_v(l \mid p_v) = \sum_{j \in S_v} p_{vj} \cdot \delta_{L_j}(l)$$

The Bayesian Bootstrap

Let $S_v = \{i : V_i = v\}$ with size n_z

- Frequentist estimate:

$$\hat{P}_v(l) = \sum_{j \in S_v} \frac{1}{n_v} \cdot \delta_{L_j}(l)$$

- Bayesian model:

$$P_v(l \mid p_v) = \sum_{j \in S_v} p_{vj} \cdot \delta_{L_j}(l)$$

  - Unknown weight vector: $p_v = \{p_{vj} : j \in S_v\}$.

The Bayesian Bootstrap

Let $S_v = \{i : V_i = v\}$ with size n_z

- Frequentist estimate:

$$\hat{P}_v(l) = \sum_{j \in S_v} \frac{1}{n_v} \cdot \delta_{L_j}(l)$$

- Bayesian model:

$$P_v(l \mid p_v) = \sum_{j \in S_v} p_{vj} \cdot \delta_{L_j}(l)$$

  - Unknown weight vector: $p_v = \{p_{vj} : j \in S_v\}$.
  - Prior: $p_v \sim Dirichlet(0_{n_v})$

The Bayesian Bootstrap

Let $S_v = \{i : V_i = v\}$ with size n_z

- Frequentist estimate:

$$\hat{P}_v(l) = \sum_{j \in S_v} \frac{1}{n_v} \cdot \delta_{L_j}(l)$$

- Bayesian model:

$$P_v(l \mid p_v) = \sum_{j \in S_v} p_{vj} \cdot \delta_{L_j}(l)$$

  - Unknown weight vector: $p_v = \{p_{vj} : j \in S_v\}$.
  - Prior: $p_v \sim Dirichlet(0_{n_v})$
  - Posterior: $p_v \mid L \sim Dirichlet(1_{n_v})$

$$E[p_{vj} \mid L] = 1/n_v$$

Full MCMC Inference

Inference or stratum v:

1. Obtain m^{th} set of posterior draws $\omega^{(m)}$ and for each $A = a$, for $j \in S_v$

$$\mu^{(m)}(a, v, L_j) = g^{-1}(\beta_v^{(m)} + \beta_L^{(m)} L_j + \theta_v^{(m)} a)$$

Full MCMC Inference

Inference or stratum v:

1. Obtain m^{th} set of posterior draws $\omega^{(m)}$ and for each $A = a$, for $j \in S_v$

$$\mu^{(m)}(a, v, L_j) = g^{-1}(\beta_v^{(m)} + \beta_L^{(m)} L_j + \theta_v^{(m)} a)$$

2. Draw Bayesian Bootstrap weights from posterior:

$$p_v^{(m)} \sim Dirichlet(1_{n_v})$$

Full MCMC Inference

Inference or stratum v:

1. Obtain m^{th} set of posterior draws $\omega^{(m)}$ and for each $A = a$, for $j \in S_v$

$$\mu^{(m)}(a, v, L_j) = g^{-1}(\beta_v^{(m)} + \beta_L^{(m)} L_j + \theta_v^{(m)} a)$$

2. Draw Bayesian Bootstrap weights from posterior:

$$p_v^{(m)} \sim Dirichlet(1_{n_v})$$

3. Integrate of confounder distribution

$$E^{(m)}(Y^a \mid v) = \int_{\mathcal{L}} \mu(a, v, L) dP_v(L) \approx \sum_{j \in S_v} \mu^{(m)}(a, v, L_j) \cdot p_{vj}^{(m)}$$

Full MCMC Inference

Inference or stratum v:

1. Obtain m^{th} set of posterior draws $\omega^{(m)}$ and for each $A = a$, for $j \in S_v$

$$\mu^{(m)}(a, v, L_j) = g^{-1}(\beta_v^{(m)} + \beta_L^{(m)} L_j + \theta_v^{(m)} a)$$

2. Draw Bayesian Bootstrap weights from posterior:

$$p_v^{(m)} \sim Dirichlet(1_{n_v})$$

3. Integrate of confounder distribution

$$E^{(m)}(Y^a \mid v) = \int_{\mathcal{L}} \mu(a, v, L) dP_v(L) \approx \sum_{j \in S_v} \mu^{(m)}(a, v, L_j) \cdot p_{vj}^{(m)}$$

4. Compute draw of causal odds ratio:

$$\Psi^{(m)}(v) = \frac{E^{(m)}(Y^1 \mid v)/[1 - E^{(m)}(Y^1 \mid v)]}{E^{(m)}(Y^0 \mid v)/[1 - E^{(m)}(Y^1 \mid v)]}$$

Implementation in Stan

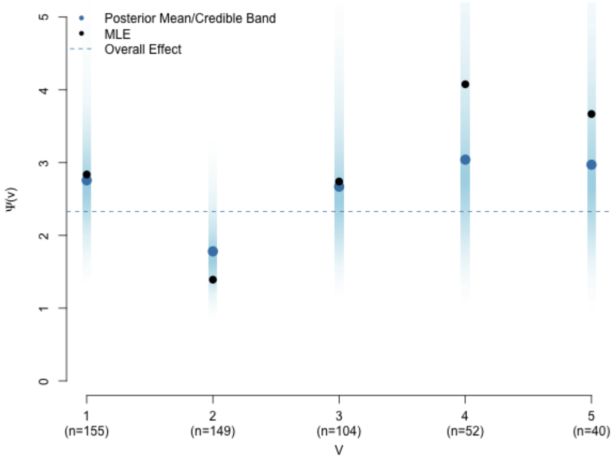
```
for(v in 1:Pv){
...
    // compute conditional means.
    cond_mean_y1 = inv_logit( Wv*beta_w + beta_v[v] + theta[v] );
    cond_mean_y0 = inv_logit( Wv*beta_w + beta_v[v] );

    // Bayesian bootstrap weights
    bb_weights = dirichlet_rng( rep_vector(1, nv) ) ;

    // taking average over bayesian bootstrap weights
    marg_mean_y1 = bb_weights' * cond_mean_y1;
    marg_mean_y0 = bb_weights' * cond_mean_y0;

    // compute odds ratio
    odds_1 = (marg_mean_y1/(1 - marg_mean_y1));
    odds_0 = (marg_mean_y0/(1 - marg_mean_y0));
    odds_ratio[v] = odds_1/odds_0;
...
}
```

Synthetic Example



Sensitivity Analysis

Identification requires conditional ignorability

$$Y^a \perp A \mid L, V = v$$

But, what if ignorability is violated?

$$E[Y^a \mid A = 1, L, v] \neq E[Y^a \mid A = 0, L, v]$$

Consequence of Violation

Define,

$$\Delta^a(L) = E[Y^a \mid A = 1, L, v] - E[Y^a \mid A = 0, L, v]$$

Consequence of Violation

Define,

$$\Delta^a(L) = E[Y^a \mid A = 1, L, v] - E[Y^a \mid A = 0, L, v]$$

$$\int \mu(1, v, L) - \mu(0, v, L)\, dP_v(L) = E[Y^1 - Y^0 \mid v] + \xi$$

Estimate of risk difference is biased by ξ.

Form of Violation

Trade-offs involved in sensitivity analyses

$$\xi = \int \Delta^1(L)(1 - \pi_v(L)) + \Delta^0(L)\pi_v(L)\, dP_v(L)$$

Where $\pi_v = P(A = 1 \mid L, v)$ is the propensity score.

Form of Violation

Trade-offs involved in sensitivity analyses

$$\xi = \int \Delta^1(L)(1 - \pi_v(L)) + \Delta^0(L)\pi_v(L)\, dP_v(L)$$

Where $\pi_v = P(A = 1 \mid L, v)$ is the propensity score.
Simplify $\Delta := \Delta^1 = \Delta^0$ and $\Delta \perp L, V$.

Form of Violation

Trade-offs involved in sensitivity analyses

$$\xi = \int \Delta^1(L)(1 - \pi_v(L)) + \Delta^0(L)\pi_v(L)\, dP_v(L)$$

Where $\pi_v = P(A = 1 \mid L, v)$ is the propensity score.
Simplify $\Delta := \Delta^1 = \Delta^0$ and $\Delta \perp L, V$.
Then,

$$\xi = \Delta$$

Now we can specify priors over Δ.

Priors over bias

Note that $-1 < E[Y^1 - Y^0 \mid v] < 1$ and recall:

$$\Delta = E[Y^a \mid A = 1, L, v] - E[Y^a \mid A = 0, L, v]$$

Priors over bias

Note that $-1 < E[Y^1 - Y^0 \mid v] < 1$ and recall:

$$\Delta = E[Y^a \mid A = 1, L, v] - E[Y^a \mid A = 0, L, v]$$

- Treated patients systematically worse:

$$\Delta \sim U(0, 1)$$

Priors over bias

Note that $-1 < E[Y^1 - Y^0 \mid v] < 1$ and recall:

$$\Delta = E[Y^a \mid A = 1, L, v] - E[Y^a \mid A = 0, L, v]$$

- Treated patients systematically worse:

$$\Delta \sim U(0, 1)$$

- Treated patients systematically better:

$$\Delta \sim U(-1, 0)$$

Priors over bias

Note that $-1 < E[Y^1 - Y^0 \mid v] < 1$ and recall:

$$\Delta = E[Y^a \mid A = 1, L, v] - E[Y^a \mid A = 0, L, v]$$

- Treated patients systematically worse:

$$\Delta \sim U(0, 1)$$

- Treated patients systematically better:

$$\Delta \sim U(-1, 0)$$

- Biased with uncertain direction:

$$\Delta \sim U(-1, 1)$$

Modified MCMC Inference

In Step 3 at m^{th} iteration:
Draw $\Delta^{(m)}$ from the prior and compute,

$$E^{(m)}(Y^1 \mid v) - E^{(m)}(Y^0 \mid v) = \sum_{j \in S_v} \left\{ \mu^{(m)}(1, v, L_j) - \mu^{(m)}(0, v, L_j) \right\} \cdot p_{vj}^{(m)}$$

Subtract off bias from prior,

$$\left\{ E^{(m)}(Y^1 \mid v) - E^{(m)}(Y^0 \mid v) \right\} - \Delta^{(m)}$$

Implementation in Stan

- Could specify prior for Δ in "model" block. Manipulate in "generated quantities".
- Could draw Δ from specified distribution in "generated quantities" block.

Some Resources

- A Practical Introduction to Bayesian Estimation of Causal Effects: Parametric and Nonparametric Approaches `https://arxiv.org/pdf/2004.07375.pdf`
- Companion GitHub repo for paper: `https://github.com/stablemarkets/intro_bayesian_causal`
- GitHub Repo for this talk: `https://github.com/stablemarkets/StanCon2020_BayesCausal`

THANK YOU!