Estimating Inverse Probability Weights using Super Learner when weight-model specification is unknown in a Marginal Structural Cox Model context

Mohammad Ehsanul Karim (Ehsan)
Epidemiology, Biostatistics and Occupational Health, McGill University
Joint work with Robert Platt & The BeAMS Study Group

XXVIIIth International Biometric Conference

July 11th, 2016

Outline

1. Multiple sclerosis
   - Data description
   - Causal diagram for treatment effect
2. Marginal Structural Cox Model
   - Weight formula
   - Calculation using Super Learner
3. Results
   - Simulation
   - Data analysis
4. Discuss

- Retrospective study (1995-2008), BC (Karim ME, et al. 2014: AJE).
- 1,697 patients followed; 829 remained untreated.

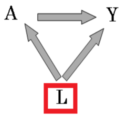
Variables under consideration:

- Treatment: β-interferons (time-dependent exposure A_t)
- Survival outcome (Y): time from baseline to irreversible disability (sustained EDSS 6).
- Confounders (L_0) measured at baseline:
  1. Disability status (measured by EDSS score)
  2. Disease duration
  3. Age
  4. Sex
- Time-dependent confounder (L_t): Relapse

Cox PH model with baseline and time-dependent confounder:

$$\lambda(t|L_0, L_t) = \lambda_{0t} \exp\left(\psi_1 A_t + \psi_2 L_0 + \psi_3 L_t\right)$$

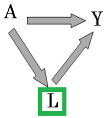
Multiple Sclerosis Data > Causal Diagram



Confounder

(L should be controlled)

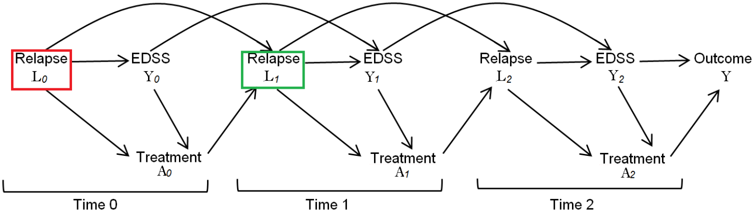
$\lambda_{0t} \exp\left(\psi_1 A_{it} + \psi_2 L_{i0} + \psi_3 L_{it}\right)$



Intermediate variable

(L should not be controlled)

$\lambda_{0t} \exp\left(\psi_1 A_{it} + \psi_2 L_{i0}\right)$



Time-dependent confounder (relapse) is affected by prior treatment.

Calculate weights (stabilized treatment IPW):

$$sw_t = \prod_{j=0}^{t} \frac{pr(A_j = a_j|\bar{A}_{j-1} = \bar{a}_{j-1}, L_0 = l_0)}{pr(A_j = a_j|\bar{A}_{j-1} = \bar{a}_{j-1}, L_0 = l_0, \bar{L}_j = \bar{l}_j)}$$

Outcome model (in the pseudo-population) with baseline confounder

$$\lambda(t|L_0) = \lambda_0(t) \exp\left(\psi_1 A_t + \psi_2 L_0\right)$$

Numerator weight model:

$$logit\ Pr(A_j = 1|\bar{A}_{j-1}, L_0, \boldsymbol{\alpha'}) = \alpha'_0(j) + \alpha'_1 A_{j-1} + \alpha'_2 L_0 \quad (1)$$

Denominator weight model:

$$logit\ Pr(A_j = 1|\bar{A}_{j-1}, L_0, \bar{L}_j, \boldsymbol{\alpha}) = \alpha_0(j) + \alpha_1 A_{j-1} + \alpha_2 L_0 + \alpha_3 L_j \quad (2)$$

Weights play a key role in the MSCM approach:

1. In practical applications, researchers are often unaware of the **true form of the weight model**:
   - non-linearity (e.g., quadratic or higher-order effects)
   - non-additivity (e.g., interaction terms)
2. MSCM estimates are highly **sensitive to the weight-model mis-specification**.

IPW Model	Parametric Regression	Data-adaptive Methods
Example	Logistic regression	Classification and regression trees
Pros	**Efficient** MSCM estimates.	Data-adaptively **detects** data features.
Cons	Assumptions may be too **restrictive**.	Possibly **inefficient** MSCM estimates.

1. **Super Learner (SL)** selects from a set of user-specified candidate library that may include
   - parametric regression models
   - semi-parametric regression models
   - data-adaptive statistical learning methods
2. Using cross-validation, SL approach **optimally combines the predicted values** from each candidate learner.
3. Prediction-wise, SL generally **asymptotically outperforms** each of the candidate estimators in the library (in absence of the correct parametric model).
4. SL may offer a **better alternative** to logistic regression model or other data-adaptive statistical learning approaches (when true parametric weight-model specification unknown).

Candidate learners in the SL library

Learner	Description
Logistic regression	The main terms of the covariates
Stepwise logistic	Variables selected from quadratic terms and two-way interactions based on AIC criterion
Elastic net	Mixing parameter $= 0.5$
Bayesian logistic	Cauchy prior with scale $= 2.5$
CART	Complexity parameter $= 0.01$
Pruned CART	Complexity parameter chosen such that the cross-validated error rate is minimum
Bagged CART	Based on 100 replications
Boosted CART	Based on $5,000$ trees and interaction depth $= 3$
Random Forest	Based on $1,000$ trees
SVM	Polynomial kernel

Simulation specifications >

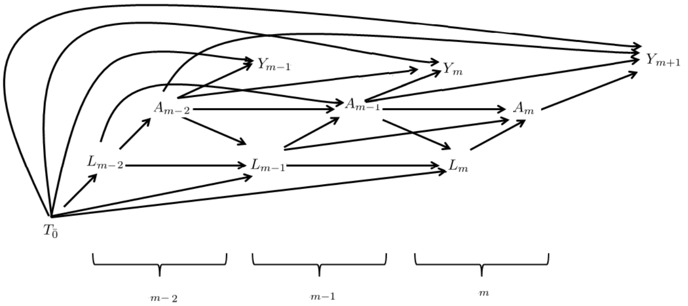


Figure: Causal diagram of MSCM data generation algorithm (similar to Young et al. 2010) where L_m is a continuous variable & a **time-dependent confounder**.

$$L_m = \beta_0 + \beta_1 \left(1/\log(T_0)\right) + \beta_2 A_{m-1} + \beta_3 L_{m-1}$$

$N = 1000$ datasets, $n = 2,500$ subjects, each followed for up to $m = 10$ subsequent monthly visits, $\lambda_0 = 0.01$ rate of monthly events

Possible forms of the true treatment model:

I. Additivity and linearity

$$logit(p_A) = \alpha_0 + \alpha_1 A_{m-1} + \alpha_2 L_m + \alpha_3 L_{m-1}.$$

II. Non-additivity

$$logit(p_A) = \alpha_0 + \alpha_1 A_{m-1} + \alpha_2 L_m + \alpha_3 L_{m-1} + \alpha_4 (A_{m-1} \times L_m).$$

III. Non-linearity

$$logit(p_A) = \alpha_0 + \alpha_1 A_{m-1} + \alpha_2 (L_m)^2 + \alpha_3 (L_{m-1})^2.$$

IV. Non-linearity and non-additivity

$$logit(p_A) = \alpha_0 + \alpha_1 A_{m-1} + \alpha_2 (L_m)^2 + \alpha_3 (L_{m-1})^2 + \alpha_4 (A_{m-1} \times L_m).$$

I. Additivity and linearity: ordered by MSE (ascending)

$$logit(p_A) \quad = \quad \alpha_0 + \alpha_1 A_{m-1} + \alpha_2 L_m + \alpha_3 L_{m-1}.$$

	Bias	**MSE**	SE	SD	Cov.Pr.
Super learner	-0.0719	0.0844	0.312	0.281	0.969
Elastic net	-0.1336	0.1031	0.308	0.292	0.934
Boosted CART	-0.1493	0.1039	0.314	0.286	0.951
Bayesian logistic	0.0195	0.1071	0.323	0.327	0.972
Logistic	0.0645	0.1218	0.329	0.343	0.972
Bagged CART	-0.2469	0.2749	0.386	0.463	0.837
Stepwise	0.1458	0.3750	0.346	0.595	0.950
CART	-0.4232	0.4221	0.397	0.493	0.722
Pruned CART	-0.6215	0.6246	0.342	0.488	0.507
SVM	0.3807	1.7024	0.502	1.248	0.601
Random Forest	-0.6002	2.4178	0.309	1.434	0.148

II. Non-additivity: ordered by MSE (ascending)

$$logit(p_A) = \alpha_0 + \alpha_1 A_{m-1} + \alpha_2 L_m + \alpha_3 L_{m-1} + \alpha_4 (A_{m-1} \times L_m).$$

	Bias	**MSE**	SE	SD	Cov.Pr.
Super learner	0.00825	0.0312	0.185	0.176	0.970
Boosted CART	0.02492	0.0316	0.187	0.176	0.965
Bagged CART	-0.00614	0.0325	0.193	0.180	0.965
Stepwise	0.03801	0.0654	0.223	0.253	0.966
Random Forest	0.03017	0.0741	0.294	0.270	0.973
CART	0.02451	0.0769	0.229	0.276	0.914
Pruned CART	-0.04692	0.0849	0.222	0.288	0.867
Elastic net	0.21918	0.0881	0.207	0.200	0.839
Bayesian logistic	0.24436	0.1011	0.210	0.203	0.822
Logistic	0.25562	0.1083	0.213	0.207	0.814
SVM	0.16572	0.2104	0.240	0.428	0.845

III. Non-linearity: ordered by MSE (ascending)

$$logit(p_A) \quad = \quad \alpha_0 + \alpha_1 A_{m-1} + \alpha_2 (L_m)^2 + \alpha_3 (L_{m-1})^2.$$

	Bias	**MSE**	SE	SD	Cov.Pr.
Super learner	0.1059	0.259	0.468	0.498	0.9667
CART	0.2217	0.324	0.473	0.524	0.9170
Bagged CART	0.3249	0.342	0.491	0.486	0.9410
Boosted CART	0.3369	0.357	0.498	0.493	0.9157
Pruned CART	-0.0903	0.411	0.472	0.634	0.8550
Stepwise	0.2332	0.594	0.421	0.735	0.7655
Elastic net	0.3813	0.601	0.540	0.675	0.8960
Bayesian logistic	0.4147	0.602	0.571	0.656	0.9000
Logistic	0.3290	0.695	0.488	0.766	0.8847
Random Forest	-1.1129	1.402	0.434	0.405	0.2420
SVM	2.1906	6.128	0.323	1.153	0.0287

IV. Non-linearity and non-additivity: ordered by MSE (ascending)

$$logit(p_A) \quad = \quad \alpha_0 + \alpha_1 A_{m-1} + \alpha_2 (L_m)^2 + \alpha_3 (L_{m-1})^2 + \alpha_4 (A_{m-1} \times L_m).$$

	Bias	**MSE**	SE	SD	Cov.Pr.
Super learner	0.0952	0.241	0.468	0.481	0.9688
CART	0.2038	0.305	0.471	0.513	0.9230
Bagged CART	0.3018	0.315	0.487	0.473	0.9470
Boosted CART	0.3386	0.356	0.497	0.491	0.9137
Pruned CART	-0.1086	0.402	0.470	0.624	0.8560
Bayesian logistic	0.3522	0.506	0.562	0.618	0.9190
Elastic net	0.3490	0.551	0.537	0.655	0.9060
Stepwise	0.2186	0.555	0.428	0.712	0.7928
Logistic	0.3110	0.660	0.490	0.751	0.8933
Random Forest	-1.1216	1.423	0.434	0.406	0.2320
SVM	2.1899	6.137	0.324	1.158	0.0289

Table: The estimated causal effect of β-IFN on reaching sustained EDSS 6 for BC MS patients (1995-2008).

Estimated weights sw generated via SL		Treatment effect estimate		
Mean (log-SD)	Min-Max	HR	SE	95% CI
1.056 (-0.771)	0.392 - 2.379	1.349	0.316	0.853 - 2.134

- In our **Multiple Sclerosis application**, the hazard ratio estimates from the super learning approach is 1.349, and this effect estimate was not significant (95% CI $0.853 - 2.134$).
- This conclusion is consistent with those of the previous studies.

Summary >

- When stabilized weights were computed via this SL, the resulting MSCM estimates computed from **SL generally performed better in terms of MSE** compared to individual candidate learners.

- These simulations shows the **utility of using SL** approach with rich set of candidate learners in practical scenarios when the form of the **treatment decision model is unknown** and may deviate from linearity, additivity or both.

- However, these tools are not meant to replace **subject-matter knowledge and expert-opinion**:
  1. bias amplification (controlling IV),
  2. under-adjustment (omitting potential confounder).

Thank You!

Contact e-mail: ehsan.karim@mail.mcgill.ca