Debiased Brier score estimation using TMLE

Anders Munch (with Thomas G., Helene, and Paul)

September 22, 2021

Disclaimer – unfinished work

Mostly theory and not so much implementation and “results”.

Disclaimer – unfinished work

Mostly theory and not so much implementation and “results”.

You have the opportunity to influence the project!

Validating risk prediction models

Full data

$X \in \mathbb{R}^p$	Static covariates measured at baseline ($t = 0$).
$T \in \mathbb{R}_+$	Time of event
$t \in \mathbb{R}_+$	Fixed time horizon
$r(t \mid X) \in [0, 1]$	Risk prediction at time t given baseline covariates

Validating risk prediction models

Full data

$X \in \mathbb{R}^p$ Static covariates measured at baseline ($t = 0$).

$T \in \mathbb{R}_+$ Time of event

$t \in \mathbb{R}_+$ Fixed time horizon

$r(t \mid X) \in [0, 1]$ Risk prediction at time t given baseline covariates

Risk prediction model

We assume the risk prediction model r is fixed (i.e., non-random); for instance, it could have been fitted on a separate data set. We want to use the data (X_i, T_i), $i = 1, \ldots, n$ to evaluate the performance of r.

Validating risk prediction models

Full data

$X \in \mathbb{R}^p$ Static covariates measured at baseline ($t = 0$).

$T \in \mathbb{R}_+$ Time of event

$t \in \mathbb{R}_+$ Fixed time horizon

$r(t \mid X) \in [0, 1]$ Risk prediction at time t given baseline covariates

Risk prediction model

We assume the risk prediction model r is fixed (i.e., non-random); for instance, it could have been fitted on a separate data set. We want to use the data (X_i, T_i), $i = 1, \ldots, n$ to evaluate the performance of r.

Example

Is a particular bio-marker relevant for predicting the risk of developing some disease within the next two years? Is it relevant when other risk factors are measured?

The Brier score

$Y(t) \in \{0, 1\}$ Event indicator at time t, i.e., $Y(t) := \mathbb{1}\{T \leq t\}$

The Brier score

$Y(t) \in \{0,1\}$ Event indicator at time t, i.e., $Y(t) := \mathbb{1}\{T \leq t\}$

The (average) Brier score of the risk model r is the (average of the) squared difference between $Y(t)$ and the predicted risk according to the model, i.e.,

$$\mathbb{E}\left[\{Y(t) - r(t \mid X)\}^2\right].$$

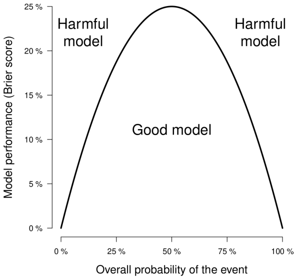
The Brier score

$Y(t) \in \{0, 1\}$ Event indicator at time t, i.e., $Y(t) := \mathbb{1}\{T \leq t\}$

The (average) Brier score of the risk model r is the (average of the) squared difference between $Y(t)$ and the predicted risk according to the model, i.e.,

$$\mathbb{E}\left[\{Y(t) - r(t \mid X)\}^2\right].$$

Benchmark prediction	Brier score
50% always	25%
Overall event prob.	See figure
Coin toss	50%
Uniform[0,1]	33%



[Gerds and Kattan, 2021]

Obtaining the Brier score from censored data

In many cases we do not get to observe the event time T.

Obtaining the Brier score from censored data

In many cases we do not get to observe the event time T.

Observed data

$X \in \mathbb{R}^p$ Static covariates measured at baseline ($t = 0$).

$\tilde{T} \in \mathbb{R}_+$ Observation time ($\tilde{T} := T \wedge C$)

$\Delta \in \{0, 1\}$ Event indicator ($\Delta := \mathbb{1}\{\tilde{T} = T\}$)

Let $(X, T) \sim Q$ and $O := (X, \tilde{T}, \Delta) \sim P$.

Obtaining the Brier score from censored data

In many cases we do not get to observe the event time T.

Observed data

$X \in \mathbb{R}^p$ Static covariates measured at baseline ($t = 0$).

$\tilde{T} \in \mathbb{R}_+$ Observation time ($\tilde{T} := T \wedge C$)

$\Delta \in \{0, 1\}$ Event indicator ($\Delta := \mathbb{1}\{\tilde{T} = T\}$)

Let $(X, T) \sim Q$ and $O := (X, \tilde{T}, \Delta) \sim P$.

Inverse probability of censoring weights (IPCW)

When $T \perp\!\!\!\perp C \mid X$ the Brier score is identifiable from the observed data[1]:

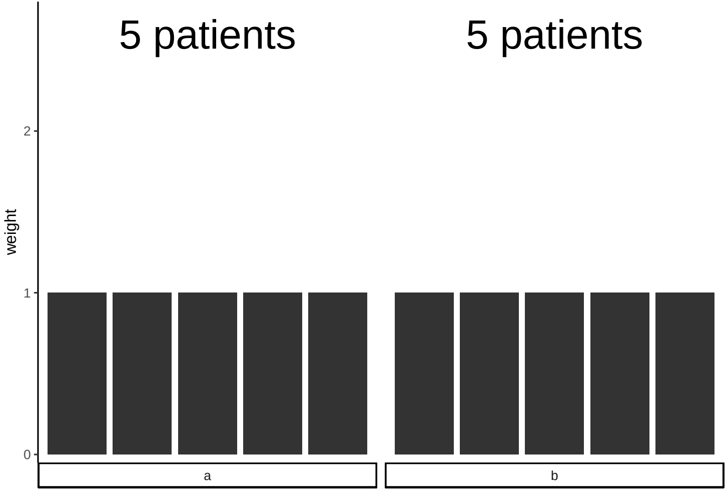
$$\mathbb{E}_Q\left[\{Y(t) - r(r \mid X)\}^2\right] = \mathbb{E}_P\left[W(t)\{Y(t) - r(r \mid X)\}^2\right],$$

with

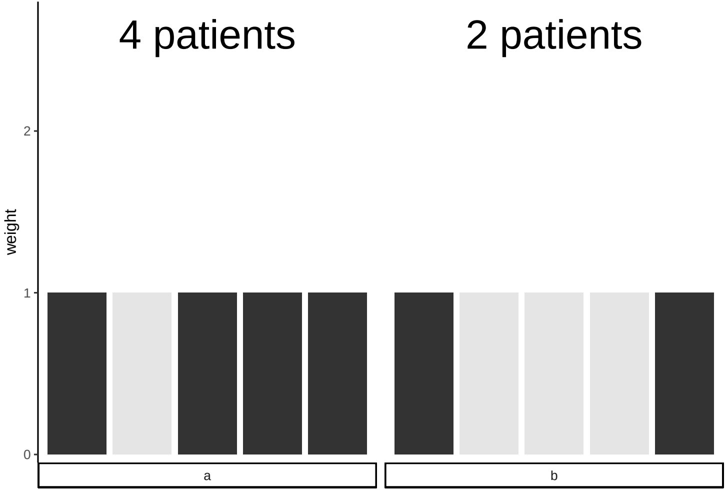
$$W(t) = \frac{\mathbb{1}(\tilde{T} > t)}{G(t \mid X)} + \frac{\mathbb{1}(\tilde{T} \leq t)\Delta}{G(\tilde{T} \mid X)}.$$

[1] Note that $W(t)\{Y(t) - r(t \mid X)\}^2$ is a function of the observed data, as $Y(t)$ is observed whenever $W(t)$ is non-zero.

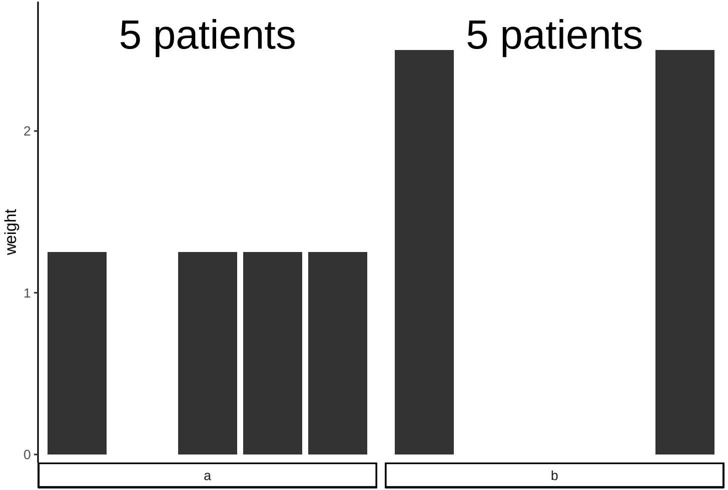
Visualizing the reweighting



Visualizing the reweighting



Visualizing the reweighting



Estimating the Brier score from censored data

By estimating the censoring distribution G we obtain the IPCW estimator

$$\widehat{W}_i(t) = \frac{\mathbb{1}(\tilde{T}_i > t)}{\hat{G}(t \mid X_i)} + \frac{\mathbb{1}(\tilde{T}_i \leq t)\Delta_i}{\hat{G}(\tilde{T}_i \mid X_i)}, \quad \hat{\theta}_n^t = \hat{\mathbb{P}}_n[\widehat{W}_i(t)\,\{Y_i(t) - r(t \mid X_i)\}^2]$$

Estimating the Brier score from censored data

By estimating the censoring distribution G we obtain the IPCW estimator

$$\widehat{W}_i(t) = \frac{\mathbb{1}(\tilde{T}_i > t)}{\hat{G}(t \mid X_i)} + \frac{\mathbb{1}(\tilde{T}_i \leq t)\Delta_i}{\hat{G}(\tilde{T}_i \mid X_i)}, \quad \hat{\theta}_n^t = \hat{\mathbb{P}}_n[\widehat{W}_i(t)\,\{Y_i(t) - r(t \mid X_i)\}^2]$$

- The parametric ($n^{-1/2}$) rate of convergence of $\hat{G}_n$ is obtainable under suitable assumptions on the censoring distribution (for instance random censoring or a Cox model), giving also $n^{-1/2}$ convergence and asymptotic normality of $\hat{\theta}_n^t$.

Estimating the Brier score from censored data

By estimating the censoring distribution G we obtain the IPCW estimator

$$\widehat{W}_i(t) = \frac{\mathbb{1}(\tilde{T}_i > t)}{\hat{G}(t \mid X_i)} + \frac{\mathbb{1}(\tilde{T}_i \leq t)\Delta_i}{\hat{G}(\tilde{T}_i \mid X_i)}, \quad \hat{\theta}_n^t = \hat{\mathbb{P}}_n[\widehat{W}_i(t)\{Y_i(t) - r(t \mid X_i)\}^2]$$

- The parametric ($n^{-1/2}$) rate of convergence of $\hat{G}_n$ is obtainable under suitable assumptions on the censoring distribution (for instance random censoring or a Cox model), giving also $n^{-1/2}$ convergence and asymptotic normality of $\hat{\theta}_n^t$.
- Such assumptions can be unpleasant, in particular if we are validating a risk prediction model for which we *do not* make similar assumptions (e.g., random forests or other "data-adaptive" estimator).

Estimating the Brier score from censored data

By estimating the censoring distribution G we obtain the IPCW estimator

$$\widehat{W}_i(t) = \frac{\mathbb{1}(\tilde{T}_i > t)}{\hat{G}(t \mid X_i)} + \frac{\mathbb{1}(\tilde{T}_i \leq t)\Delta_i}{\hat{G}(\tilde{T}_i \mid X_i)}, \quad \hat{\theta}_n^t = \hat{\mathbb{P}}_n[\widehat{W}_i(t)\{Y_i(t) - r(t \mid X_i)\}^2]$$

- The parametric ($n^{-1/2}$) rate of convergence of $\hat{G}_n$ is obtainable under suitable assumptions on the censoring distribution (for instance random censoring or a Cox model), giving also $n^{-1/2}$ convergence and asymptotic normality of $\hat{\theta}_n^t$.
- Such assumptions can be unpleasant, in particular if we are validating a risk prediction model for which we *do not* make similar assumptions (e.g., random forests or other "data-adaptive" estimator).
- When modeling G with flexible, data-adaptive methods we cannot expect $n^{-1/2}$-rate convergence, and hence the simple plug-in estimator $\hat{\theta}_n^t$ cannot be expected to be $n^{-1/2}$ consistent and asymptotically normal in this setting.

Estimating the Brier score from censored data

By estimating the censoring distribution G we obtain the IPCW estimator

$$\widehat{W}_i(t) = \frac{\mathbb{1}(\tilde{T}_i > t)}{\hat{G}(t \mid X_i)} + \frac{\mathbb{1}(\tilde{T}_i \leq t)\Delta_i}{\hat{G}(\tilde{T}_i \mid X_i)}, \quad \hat{\theta}_n^t = \hat{\mathbb{P}}_n[\widehat{W}_i(t)\{Y_i(t) - r(t \mid X_i)\}^2]$$

- The parametric ($n^{-1/2}$) rate of convergence of $\hat{G}_n$ is obtainable under suitable assumptions on the censoring distribution (for instance random censoring or a Cox model), giving also $n^{-1/2}$ convergence and asymptotic normality of $\hat{\theta}_n^t$.
- Such assumptions can be unpleasant, in particular if we are validating a risk prediction model for which we *do not* make similar assumptions (e.g., random forests or other "data-adaptive" estimator).
- When modeling G with flexible, data-adaptive methods we cannot expect $n^{-1/2}$-rate convergence, and hence the simple plug-in estimator $\hat{\theta}_n^t$ cannot be expected to be $n^{-1/2}$ consistent and asymptotically normal in this setting.

Can we construct an IPCW estimator using "flexible/data-adaptive" estimation of G?

Brier score estimation with "flexible" censoring modeling

By "flexible" we mean estimators of the censoring distribution not converging at parametric ($n^{-1/2}$) rate. The problem with the plug-in estimation using such nuisance parameter estimators is bias.

Brier score estimation with “flexible” censoring modeling

By “flexible” we mean estimators of the censoring distribution not converging at parametric ($n^{-1/2}$) rate. The problem with the plug-in estimation using such nuisance parameter estimators is bias.

DML [Chernozhukov et al., 2017] / one-step estimation

Construct estimators as the solution to the empirical efficient score equation. The obtained estimator is no longer an IPCW estimator.

Brier score estimation with “flexible” censoring modeling

By “flexible” we mean estimators of the censoring distribution not converging at parametric ($n^{-1/2}$) rate. The problem with the plug-in estimation using such nuisance parameter estimators is bias.

DML [Chernozhukov et al., 2017] / one-step estimation

Construct estimators as the solution to the empirical efficient score equation. The obtained estimator is no longer an IPCW estimator.

Undersmoothing

Challenging to do for general nuisance parameter estimators, but some recent work for the Highly Adaptive Lasso (HAL) estimator [Ertefaie et al., 2020, van der Laan et al., 2019]. Seems computationally challenging.

Brier score estimation with "flexible" censoring modeling

By "flexible" we mean estimators of the censoring distribution not converging at parametric ($n^{-1/2}$) rate. The problem with the plug-in estimation using such nuisance parameter estimators is bias.

DML [Chernozhukov et al., 2017] / one-step estimation

Construct estimators as the solution to the empirical efficient score equation. The obtained estimator is no longer an IPCW estimator.

Undersmoothing

Challenging to do for general nuisance parameter estimators, but some recent work for the Highly Adaptive Lasso (HAL) estimator [Ertefaie et al., 2020, van der Laan et al., 2019]. Seems computationally challenging.

TMLE [van der Laan and Rose, 2011, van der Laan and Rubin, 2006]

TMLE constructs a plug-in estimator that solves the efficient score equation. Typically based on the G-formula, but theoretically this should not be important.

Different ways to the target

The target parameter (average Brier score) can be identified from the observed data using either the censoring or the survival distribution as nuisance parameter:

$$\mathbb{E}_Q\left[\{Y(t) - r(t \mid X)\}^2\right] = \mathbb{E}_P\left[\varphi^t_{\mathrm{IPCW}}(O; G)\right] = \mathbb{E}_P\left[\varphi^t_{\mathrm{alt}}(O; S)\right],$$

with

$$\varphi^t_{\mathrm{IPCW}}(O; G) = \left(\frac{\mathbb{1}(\tilde{T}_i > t)}{G(t \mid X_i)} + \frac{\mathbb{1}(\tilde{T}_i \le t)\Delta_i}{G(\tilde{T}_i \mid X_i)}\right)\{Y(t) - r(t \mid X)\}^2,$$

and

$$\varphi^t_{\mathrm{alt}}(O; S) = (1 - S(t \mid X))\{1 - 2r(t \mid X)\} + r(t \mid X)^2.$$

Representations of the efficient influence function

The functions $\varphi^t_{\mathrm{IPCW}}$ and φ^t_{alt} are influence functions in models where, respectively, G or S are known.

Representations of the efficient influence function

The functions $\varphi^t_{\mathrm{IPCW}}$ and φ^t_{alt} are influence functions in models where, respectively, G or S are known. Semiparametric efficiency theory then tells us that the efficient influence function φ^t (under the non-parametric model) can be obtained as either

$$\varphi^t = \varphi^t_{\mathrm{IPCW}} - \Pi\left[\varphi^t_{\mathrm{IPCW}} \mid \mathcal{T}_G\right],$$

or

$$\varphi^t = \varphi^t_{\mathrm{alt}} - \Pi\left[\varphi^t_{\mathrm{alt}} \mid \mathcal{T}_S\right],$$

where $\mathcal{T}_G$ and $\mathcal{T}_S$ are the orthogonal components of the tangent space corresponding to the parameters G and S, i.e., $\mathcal{T} = \mathcal{T}_G \oplus \mathcal{T}_S$.

Representations of the efficient influence function

The functions $\varphi^t_{\mathrm{IPCW}}$ and φ^t_{alt} are influence functions in models where, respectively, G or S are known. Semiparametric efficiency theory then tells us that the efficient influence function φ^t (under the non-parametric model) can be obtained as either

$$\varphi^t = \varphi^t_{\mathrm{IPCW}} - \Pi\left[\varphi^t_{\mathrm{IPCW}} \mid \mathcal{T}_G\right],$$

or

$$\varphi^t = \varphi^t_{\mathrm{alt}} - \Pi\left[\varphi^t_{\mathrm{alt}} \mid \mathcal{T}_S\right],$$

where $\mathcal{T}_G$ and $\mathcal{T}_S$ are the orthogonal components of the tangent space corresponding to the parameters G and S, i.e., $\mathcal{T} = \mathcal{T}_G \oplus \mathcal{T}_S$.

Both representation can be useful; and starting from one, it might not be complete straightforward to derive the alternative one.

The efficient influence function

The two representations give

$$\begin{aligned}\varphi^t(O; S, G) =& \varphi^t_{\mathrm{IPCW}}(O; G) \\ &+ \{1 - r(t \mid X)\}^2 \int_0^t \frac{M^C(\mathrm{d}s \mid X; G)}{G(s \mid X)} \\ &- \{1 - 2r(t \mid X)\} S(t \mid X) \int_0^t \frac{M^C(\mathrm{d}s \mid X; G)}{G(s \mid X) S(s \mid X)},\end{aligned}$$

and

$$\begin{aligned}\varphi^t(O; S, G) =& \varphi^t_{\mathrm{alt}}(O; S) \\ &+ \left[\int_0^t \frac{M(\mathrm{d}s \mid X; S)}{S(s \mid X) G(s \mid X)}\right] S(t \mid X)(1 - 2r(t \mid X)).\end{aligned}$$

Decomposition

We want to pick the nuisance estimator $\hat{G}_n$ such that the estimator

$$\hat{\theta}_n^t = \tilde{\Psi}^t(\hat{G}_n, \hat{\mathbb{P}}_n) = \hat{\mathbb{P}}_n \left[\varphi^t_{\mathrm{IPCW}}(O; \hat{G}_n)\right]$$

is asymptotically linear (and efficient), i.e., such that

$$\hat{\theta}_n^t - \theta = (\hat{\mathbb{P}}_n - P)[\varphi^t(O; G_P, S_P)] + o_P(n^{-1/2}).$$

Decomposition

We want to pick the nuisance estimator $\hat{G}_n$ such that the estimator

$$\hat{\theta}^t_n = \tilde{\Psi}^t(\hat{G}_n, \hat{\mathbb{P}}_n) = \hat{\mathbb{P}}_n \left[\varphi^t_{\mathrm{IPCW}}(O; \hat{G}_n)\right]$$

is asymptotically linear (and efficient), i.e., such that

$$\hat{\theta}^t_n - \theta = (\hat{\mathbb{P}}_n - P)[\varphi^t(O; G_P, S_P)] + o_P(n^{-1/2}).$$

Let $f^t := -\Pi(\varphi^t_{\mathrm{IPWC}} \mid \mathcal{T}_G)$ and consider the decomposition

$$\begin{aligned}
&\tilde{\Psi}^t(\hat{G}_n, \hat{\mathbb{P}}_n) - \tilde{\Psi}^t(G_P, P) \\
&= \hat{\mathbb{P}}_n \left[\varphi^t_{\mathrm{IPCW}}(O; \hat{G}_n)\right] - \tilde{\Psi}^t(G_P, P) \pm \hat{\mathbb{P}}_n \left[f^t(O; \hat{G}_n, \hat{S}_n)\right] \\
&= \hat{\mathbb{P}}_n \left[\varphi^t(O; \hat{G}_n, \hat{S}_n)\right] - \tilde{\Psi}^t(G_P, P) - \hat{\mathbb{P}}_n \left[f^t(O; \hat{G}_n, \hat{S}_n)\right] \\
&= (\hat{\mathbb{P}}_n - P) \left[\varphi^t(O; \hat{G}_n, \hat{S}_n)\right] + \mathrm{Rem}(\hat{G}_n, \hat{S}_n, P) - \hat{\mathbb{P}}_n \left[f^t(O; \hat{G}_n, \hat{S}_n)\right], \\
&=: (A) + (B) + (C)
\end{aligned}$$

with

$$\mathrm{Rem}(G, S, P) := P[\varphi^t(O; G, S)] - \tilde{\Psi}^t(G_P, P).$$

Donsker condition and remainder term

Donsker class conditions (or sample splitting) gives

$$(A) = (\hat{\mathbb{P}}_n - P)\left[\varphi^t(O; \hat{G}_n, \hat{S}_n)\right] = (\hat{\mathbb{P}}_n - P)\left[\varphi^t(O; G, S)\right] + o_P(n^{-1/2}).$$

Donsker condition and remainder term

Donsker class conditions (or sample splitting) gives

$$(A) = (\hat{\mathbb{P}}_n - P)\left[\varphi^t(O; \hat{G}_n, \hat{S}_n)\right] = (\hat{\mathbb{P}}_n - P)\left[\varphi^t(O; G, S)\right] + o_P(n^{-1/2}).$$

As $\mathrm{Rem}(P, \hat{G}_n, \hat{S}_n) = \tilde{\Psi}^t(\hat{G}_n, P) - \tilde{\Psi}^t(G_P, P) + P[f^t(O; \hat{G}_n, \hat{S}_n)]$ and $f^t(O; \hat{G}_n, \hat{S}_n)$ acts like the derivative of $G \mapsto \tilde{\Psi}^t(G, P)$, a functional Taylor expansion would suggest that

$$(B) = \mathrm{Rem}(P, \hat{G}_n, \hat{S}_n) = \mathcal{O}_P\left(\|(\hat{G}_n, \hat{S}_n) - (G, S)\|^2\right).$$

Donsker condition and remainder term

Donsker class conditions (or sample splitting) gives

$$(A) = (\hat{\mathbb{P}}_n - P)\left[\varphi^t(O; \hat{G}_n, \hat{S}_n)\right] = (\hat{\mathbb{P}}_n - P)\left[\varphi^t(O; G, S)\right] + o_P(n^{-1/2}).$$

As $\mathrm{Rem}(P, \hat{G}_n, \hat{S}_n) = \tilde{\Psi}^t(\hat{G}_n, P) - \tilde{\Psi}^t(G_P, P) + P[f^t(O; \hat{G}_n, \hat{S}_n)]$ and $f^t(O; \hat{G}_n, \hat{S}_n)$ acts like the derivative of $G \mapsto \tilde{\Psi}^t(G, P)$, a functional Taylor expansion would suggest that

$$(B) = \mathrm{Rem}(P, \hat{G}_n, \hat{S}_n) = \mathcal{O}_P\left(\|(\hat{G}_n, \hat{S}_n) - (G, S)\|^2\right).$$

Thus, when the Donsker condition holds, and $\|\hat{G}_n - G\| = o_P(n^{-1/4})$ and $\|\hat{S}_n - S\| = o_P(n^{-1/4})$, we have

$$\tilde{\Psi}^t(\hat{G}_n, \hat{\mathbb{P}}_n) - \tilde{\Psi}^t(G_P, P) = (\hat{\mathbb{P}}_n - P)\left[\varphi^t(O; G_P, S_P)\right] - \hat{\mathbb{P}}_n\left[f^t(O; \hat{G}_n, \hat{S}_n)\right] + o_P(n^{-1/2})$$

Donsker condition and remainder term

Donsker class conditions (or sample splitting) gives

$$(A) = (\hat{\mathbb{P}}_n - P)\left[\varphi^t(O; \hat{G}_n, \hat{S}_n)\right] = (\hat{\mathbb{P}}_n - P)\left[\varphi^t(O; G, S)\right] + o_P(n^{-1/2}).$$

As $\mathrm{Rem}(P, \hat{G}_n, \hat{S}_n) = \tilde{\Psi}^t(\hat{G}_n, P) - \tilde{\Psi}^t(G_P, P) + P[f^t(O; \hat{G}_n, \hat{S}_n)]$ and $f^t(O; \hat{G}_n, \hat{S}_n)$ acts like the derivative of $G \mapsto \tilde{\Psi}^t(G, P)$, a functional Taylor expansion would suggest that

$$(B) = \mathrm{Rem}(P, \hat{G}_n, \hat{S}_n) = \mathcal{O}_P\left(\|(\hat{G}_n, \hat{S}_n) - (G, S)\|^2\right).$$

Thus, when the Donsker condition holds, and $\|\hat{G}_n - G\| = o_P(n^{-1/4})$ and $\|\hat{S}_n - S\| = o_P(n^{-1/4})$, we have

$$\tilde{\Psi}^t(\hat{G}_n, \hat{\mathbb{P}}_n) - \tilde{\Psi}^t(G_P, P) = (\hat{\mathbb{P}}_n - P)\left[\varphi^t(O; G_P, S_P)\right] - \hat{\mathbb{P}}_n\left[f^t(O; \hat{G}_n, \hat{S}_n)\right] + o_P(n^{-1/2})$$

TMLE focuses on constructing $\hat{G}_n$ such that $(C) = \hat{\mathbb{P}}_n\left[f^t(O; \hat{G}_n, \hat{S}_n)\right] \approx 0.$[2]

Donsker condition and remainder term

Donsker class conditions (or sample splitting) gives

$$(A) = (\hat{\mathbb{P}}_n - P)\left[\varphi^t(O; \hat{G}_n, \hat{S}_n)\right] = (\hat{\mathbb{P}}_n - P)\left[\varphi^t(O; G, S)\right] + o_P(n^{-1/2}).$$

As $\mathrm{Rem}(P, \hat{G}_n, \hat{S}_n) = \tilde{\Psi}^t(\hat{G}_n, P) - \tilde{\Psi}^t(G_P, P) + P[f^t(O; \hat{G}_n, \hat{S}_n)]$ and $f^t(O; \hat{G}_n, \hat{S}_n)$ acts like the derivative of $G \mapsto \tilde{\Psi}^t(G, P)$, a functional Taylor expansion would suggest that

$$(B) = \mathrm{Rem}(P, \hat{G}_n, \hat{S}_n) = \mathcal{O}_P\left(\|(\hat{G}_n, \hat{S}_n) - (G, S)\|^2\right).$$

Thus, when the Donsker condition holds, and $\|\hat{G}_n - G\| = o_P(n^{-1/4})$ and $\|\hat{S}_n - S\| = o_P(n^{-1/4})$, we have

$$\tilde{\Psi}^t(\hat{G}_n, \hat{\mathbb{P}}_n) - \tilde{\Psi}^t(G_P, P) = (\hat{\mathbb{P}}_n - P)\left[\varphi^t(O; G_P, S_P)\right] - \hat{\mathbb{P}}_n\left[f^t(O; \hat{G}_n, \hat{S}_n)\right] + o_P(n^{-1/2})$$

TMLE focuses on constructing $\hat{G}_n$ such that $(C) = \hat{\mathbb{P}}_n\left[f^t(O; \hat{G}_n, \hat{S}_n)\right] \approx 0$.[2]

[2] Note that the exact same arguments would hold if we replaced $\varphi^t_{\mathrm{IPCW}}$ with φ^t_{alt} and used $f^t = -\Pi(\varphi^t_{\mathrm{alt}} \mid \mathcal{T}_S)$ instead of $f^t = -\Pi(\varphi^t_{\mathrm{IPWC}} \mid \mathcal{T}_G)$.

The TMLE strategy for controlling the remaining component

We construct the estimator $\hat{G}$ of G as $\hat{G} = e^{-\hat{\Lambda}_C}$ where $\hat{\Lambda}_C$ is the cumulative hazard of censoring.

The TMLE strategy for controlling the remaining component

We construct the estimator $\hat{G}$ of G as $\hat{G} = e^{-\hat{\Lambda}_C}$ where $\hat{\Lambda}_C$ is the cumulative hazard of censoring. We assume available (initial estimators) $\hat{\Lambda}_C^0$ and $\hat{\Lambda}$, where $\hat{\Lambda}$ is the cumulative hazard of the event of interest.

The TMLE strategy for controlling the remaining component

We construct the estimator $\hat{G}$ of G as $\hat{G} = e^{-\hat{\Lambda}_C}$ where $\hat{\Lambda}_C$ is the cumulative hazard of censoring. We assume available (initial estimators) $\hat{\Lambda}_C^0$ and $\hat{\Lambda}$, where $\hat{\Lambda}$ is the cumulative hazard of the event of interest. To construct $\hat{\Lambda}_C$ in the right way, we recursively construct fluctuation models

$$\mathcal{F}^k := \left\{ \hat{\Lambda}_C^k(\cdot; \varepsilon) \,:\, \varepsilon \in \mathbb{R} \right\} \subset \mathcal{F}, \quad k = 1, 2, \ldots,$$

and let ε_k^* denote the MLE of the fluctuation model $\mathcal{F}^k$, and $\hat{\Lambda}_C^k := \hat{\Lambda}_C^k(\cdot; \varepsilon_k^*)$ the model corresponding to the MLE.

The TMLE strategy for controlling the remaining component

We construct the estimator $\hat{G}$ of G as $\hat{G} = e^{-\hat{\Lambda}_C}$ where $\hat{\Lambda}_C$ is the cumulative hazard of censoring. We assume available (initial estimators) $\hat{\Lambda}_C^0$ and $\hat{\Lambda}$, where $\hat{\Lambda}$ is the cumulative hazard of the event of interest. To construct $\hat{\Lambda}_C$ in the right way, we recursively construct fluctuation models

$$\mathcal{F}^k := \left\{ \hat{\Lambda}_C^k(\cdot\,; \varepsilon) \,:\, \varepsilon \in \mathbb{R} \right\} \subset \mathcal{F}, \quad k = 1, 2, \dots,$$

and let ε_k^* denote the MLE of the fluctuation model $\mathcal{F}^k$, and $\hat{\Lambda}_C^k := \hat{\Lambda}_C^k(\cdot\,; \varepsilon_k^*)$ the model corresponding to the MLE. These should be constructed such that

1. At $\varepsilon = 0$, $\hat{\Lambda}_C^1(\cdot\,; 0) = \hat{\Lambda}_C^0$ and $\hat{\Lambda}_C^{k+1}(\cdot\,; 0) = \hat{\Lambda}_C^k = \hat{\Lambda}_C^k(\cdot\,; \varepsilon_k^*)$.

The TMLE strategy for controlling the remaining component

We construct the estimator $\hat{G}$ of G as $\hat{G} = e^{-\hat{\Lambda}_C}$ where $\hat{\Lambda}_C$ is the cumulative hazard of censoring. We assume available (initial estimators) $\hat{\Lambda}_C^0$ and $\hat{\Lambda}$, where $\hat{\Lambda}$ is the cumulative hazard of the event of interest. To construct $\hat{\Lambda}_C$ in the right way, we recursively construct fluctuation models

$$\mathcal{F}^k := \left\{ \hat{\Lambda}_C^k(\cdot\,; \varepsilon) \,:\, \varepsilon \in \mathbb{R} \right\} \subset \mathcal{F}, \quad k = 1, 2, \ldots,$$

and let ε_k^* denote the MLE of the fluctuation model $\mathcal{F}^k$, and $\hat{\Lambda}_C^k := \hat{\Lambda}_C^k(\cdot\,; \varepsilon_k^*)$ the model corresponding to the MLE. These should be constructed such that

1. At $\varepsilon = 0$, $\hat{\Lambda}_C^1(\cdot\,; 0) = \hat{\Lambda}_C^0$ and $\hat{\Lambda}_C^{k+1}(\cdot\,; 0) = \hat{\Lambda}_C^k = \hat{\Lambda}_C^k(\cdot\,; \varepsilon_k^*)$.
2. The score function of the model $\hat{P}_n^{k+1}(\cdot\,; \varepsilon)$ equals f^t, i.e.,

$$\frac{\partial}{\partial \varepsilon}\bigg|_{\varepsilon = 0} \log \mathrm{d}\hat{P}_n^{k+1}(\cdot\,; \varepsilon) = f^t(\cdot\,; e^{-\hat{\Lambda}_C^k}, e^{-\hat{\Lambda}}).$$

The obtained estimator works

If the procedure converges after some K, we set $\hat{\Lambda}_C = \hat{\Lambda}_C(\cdot\ ; \varepsilon_K^*)$; then

$$\begin{aligned}
\hat{\mathbb{P}}_n[f^t(O; e^{-\hat{\Lambda}_C}, e^{-\hat{\Lambda}})] &\approx \hat{\mathbb{P}}_n[f^t(O; e^{-\hat{\Lambda}_C^{K-1}}, e^{-\hat{\Lambda}})] \\
&= \frac{\partial}{\partial \varepsilon}\Big|_{\varepsilon=0} \hat{\mathbb{P}}_n \left[\log \mathrm{d}\hat{P}_n^K(\cdot; \varepsilon)\right] \\
&\approx \frac{\partial}{\partial \varepsilon}\Big|_{\varepsilon=\varepsilon_K^*} \hat{\mathbb{P}}_n \left[\log \mathrm{d}\hat{P}_n^K(\cdot; \varepsilon)\right] = 0,
\end{aligned}$$

as $\varepsilon_K^* \approx 0$ because the procedure is converging.

Fluctuation model

We can choose a multiplicative update step to get the fluctuation model

$$\mathcal{F}^{k+1} := \left\{ \hat{\Lambda}_C^{k+1}(\cdot\,;\varepsilon) \;\middle|\; \hat{\Lambda}_C^{k+1}(\mathrm{d}s \mid x;\varepsilon) := e^{\varepsilon g(s,x;\hat{\Lambda}_C^k,\hat{\Lambda})}\hat{\Lambda}_C^k(\mathrm{d}s \mid x), \varepsilon \in \mathbb{R} \right\},$$

where

$$g(s,x;\Lambda_C,\Lambda) := \mathbb{1}(s \leq t)\left\{ \frac{\{1 - r(t \mid x)\}^2}{e^{-\Lambda_C(s|x)}} - \frac{\{1 - 2r(t \mid x)\}e^{-\Lambda(t|x)}}{e^{-\Lambda_C(s|x) - \Lambda(s|x)}} \right\}.$$

Fluctuation model

We can choose a multiplicative update step to get the fluctuation model

$$\mathcal{F}^{k+1} := \left\{ \hat{\Lambda}_C^{k+1}(\cdot\,; \varepsilon) \;\middle|\; \hat{\Lambda}_C^{k+1}(\mathrm{d}s \mid x; \varepsilon) := e^{\varepsilon g(s,x;\hat{\Lambda}_C^k,\hat{\Lambda})} \hat{\Lambda}_C^k(\mathrm{d}s \mid x), \varepsilon \in \mathbb{R} \right\},$$

where

$$g(s,x;\Lambda_C,\Lambda) := \mathbb{1}(s \leq t) \left\{ \frac{\{1 - r(t \mid x)\}^2}{e^{-\Lambda_C(s|x)}} - \frac{\{1 - 2r(t \mid x)\} e^{-\Lambda(t|x)}}{e^{-\Lambda_C(s|x) - \Lambda(s|x)}} \right\}.$$

One can then verify that $\hat{\Lambda}_C^{k+1}(\cdot\,; 0) = \hat{\Lambda}_C^k$ and

$$\begin{aligned} \frac{\partial}{\partial \varepsilon}\bigg|_{\varepsilon=0} \log \mathrm{d}\hat{P}_n^{k+1}(O; \varepsilon) &= \int g(s, X; \hat{\Lambda}_C^k, \hat{\Lambda}) M_C(\mathrm{d}s \mid X; \hat{\Lambda}_C) \\ &= f^t(O; e^{-\hat{\Lambda}_C^k}, e^{-\hat{\Lambda}}). \end{aligned}$$

Final algorithm

Algorithm 1: TMLE-based IPCW estimator of the average Brier score.

Input : Data O_i, $i = 1, \ldots, n$, risk prediction model r, and estimates $\hat{\Lambda}_C^0$ and $\hat{\Lambda}$

Output: Estimate of the average Brier score

$\varepsilon^* \leftarrow \infty$

$\hat{\Lambda}_C \leftarrow \hat{\Lambda}_C^0$

while $\varepsilon^* \not\approx 0$ **do**

$\quad g(s, x; \hat{\Lambda}_C, \hat{\Lambda}) \leftarrow \mathbb{1}(s \leq t) \left\{ \frac{\{1-r(t|x)\}^2}{e^{-\hat{\Lambda}_C(s|x)}} - \frac{\{1-2r(t|x)\}e^{-\hat{\Lambda}(t|x)}}{e^{-\hat{\Lambda}_C(s|x)-\hat{\Lambda}(s|x)}} \right\}$

$\quad \hat{\Lambda}_C^\dagger(\mathrm{d}s \mid x; \varepsilon) \leftarrow e^{\varepsilon g(s,x;\hat{\Lambda}_C,\hat{\Lambda})} \hat{\Lambda}_C(\mathrm{d}s \mid x)$

$\quad \varepsilon^* \leftarrow \mathrm{argsmax}_\varepsilon \sum_{i=1}^n \left\{ (1 - \Delta_i) \log(\mathrm{d}\hat{\Lambda}_C^\dagger(\tilde{T}_i \mid X_i; \varepsilon)) - \hat{\Lambda}_C^\dagger(\tilde{T}_i \mid X_i; \varepsilon) \right\}$

$\quad \hat{\Lambda}_C \leftarrow \hat{\Lambda}_C(\cdot; \varepsilon^*)$

$\hat{G}(s \mid x) \leftarrow e^{-\hat{\Lambda}_C(s|x)}$

$\widehat{W}_i \leftarrow \frac{\mathbb{1}(\tilde{T}_i > t)}{\hat{G}(t|X_i)} + \frac{\mathbb{1}(\tilde{T}_i \leq t)\Delta_i}{\hat{G}(\tilde{T}_i|X_i)}$, for $i = 1, \ldots, n$

$\hat{\theta}_n^t \leftarrow \frac{1}{n} \sum_{i=1}^n \widehat{W}_i \{r(t \mid X_i) - Y_i\}^2$

return $\hat{\theta}_n^t$

Next steps and discussion

- Implement the estimator...
- Construct both type of TMLE plug-in estimators – is the are finite sample difference, and are they more or sensitive to mis-specification of which nuisance model?
- Compare with undersmoothing – should be quite similar, and maybe they don't have to construct a fluctuation model?
- Extend to time-dependent covariates.
- General discussion about cross validation in the presence of censoring.

Thank you!

Thought and comments?

References

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey. Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–65, 2017.

A. Ertefaie, N. S. Hejazi, and M. J. van der Laan. Nonparametric inverse probability weighted estimators based on the highly adaptive lasso. *arXiv preprint arXiv:2005.11303*, 2020.

T. A. Gerds and M. W. Kattan. *Medical Risk Prediction Models: With Ties to Machine Learning*. CRC Press, 2021.

M. J. van der Laan and S. Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

M. J. van der Laan and D. Rubin. Targeted maximum likelihood learning. *The international journal of biostatistics*, 2(1), 2006.

M. J. van der Laan, D. Benkeser, and W. Cai. Efficient estimation of pathwise differentiable target parameters with the undersmoothed highly adaptive lasso. *arXiv preprint arXiv:1908.05607*, 2019.