Survival Analysis and Causal Inference

Marginal Structural Models (MSM) for Survival analysis

Romain Neugebauer, Ph.D. candidate in Biostatistics
UCB Survival analysis class - March 8^{th} 2003
E-mail: romain@stat.berkeley.edu

Overview of the issues to be addressed

- Question of interest: defining a causal effect on a survival outcome
- Data structures: terminology and notations
- Survival causal parameter of interest: describing survival causal effects with MSMs
- How to identify and estimate MSM parameters: naive approach, IPTW estimator, assumptions, implementation, intuitive understanding
- Illustration with simulations
- Generalization of the approach to censored data and other estimators

$\longrightarrow$ Illustration with an example

Question of interest

What is the causal effect of a *treatment* on an survival outcome marginally or conditionally on a covariate?

- At the unit level, a causal effect can be defined by comparing treatment-specific survival outcomes.
- At the population level, it can be defined by the influence of a change in treatment values on the (conditional) distribution of treatment-specific survival outcomes or counterfactuals.

Two types of causal questions of interest corresponding with two data structures:

- point-treatment data structure: the treatment of interest is a random variable occurring at one time point.
- longitudinal data structure: the treatment of interest is a stochastic process, i.e. a collection of random variables measured over time.

$\longrightarrow$ The framework and notations for longitudinal data is more general.

Illustration

Causal effect of a chemotherapy drug on patients' survival.

The average marginal causal effect of the chemotherapy is the difference between:1) the average survival time if all patients are treated with chemotherapy and 2) the average survival time if all patients are not treated with chemotherapy.

The average adjusted causal effect of the chemotherapy per strata of sex in the population is define by the two following differences:

- 1) the average survival time if all male patients are treated with chemotherapy and 2) the average survival time if all male patients are not treated with chemotherapy.
- 1) the average survival time if all female patients are treated with chemotherapy and 2) the average survival time if all female patients are not treated with chemotherapy.

$\longrightarrow$ This example can be used as a longitudinal data example as well (different drug doses assigned over time).

Data structures

Whether the data are point-treatment data or longitudinal data, one can define two data sets:

- the full data: data from the ideal experiment in which all counterfactuals are collected for every subject.
  $\longrightarrow$ It is typically impossible to conduct such an ideal experiment in practice.
- the observed data: data that can be observed in practice in which one unique outcome is typically measured under one treatment per subject.
  $\longrightarrow$ It corresponds with a subset of the full data.

Note that causal effects are defined using the full data but only the observed data is available to evaluate these causal effects.

$\longrightarrow$ Notations used to represent both data sets for both point-treatment and longitudinal data structures.

Statistical framework for causal inference

Notations:

- Observed treatment:
  - for point-treatment data: A,
    possibly multivariate, i.e., $A = (A_1, \ldots, A_K)$
  - for longitudinal data (history): $A(t)$ for $t = 0, \ldots, T-1$:

$$\bar{A} = \bar{A}(T-1) = (A(0), \ldots, A(T-1)),$$

    generalize notation for treatment up to time t: $\bar{A}(t)$
  - Possible outcome values a or $\bar{a}$
  - Space of all possible treatments: $\mathcal{A}$, i.e., a or $\bar{a} \in \mathcal{A}$

Statistical framework for causal inference

- Observed covariate:
  - for point-treatment data: baseline covariates $W \supset V$ (possibly multivariate) and $Y(t) = I(T \leq t)$ for $t = 0, \ldots, T$:

$$\bar{Y}(T) = (Y(0), \ldots, Y(T)),$$

  - for longitudinal data: $L(t) = (W(t), Y(t) = I(T \leq t))$ for $t = 0, \ldots, T$:

$$\bar{L}(T) = (L(0), \ldots, L(T)),$$

  where $W(0) \supset V$ is/are baseline covariate(s).

Note that the survival outcome of interest, T, is included in the observed covariate through the variables $Y(t)$ and that there is no censoring.

Statistical framework for causal inference

- Counterfactuals:

  "variables/processes observed contrary to the fact, i.e. under a treatments which are not the observed treatment".

  By extension, treatment-specific variables/processes:

  – for point treatment data: T_a and $Y_a(t) = I(T_a \leq t)$ for $t = 0, \ldots, T_a$

  – for longitudinal data: $T_{\bar{a}}$ and $L_{\bar{a}}(t)$ for $t = 0, \ldots, T_{\bar{a}}$, i.e. $\bar{L}_{\bar{a}}(T_{\bar{a}})$

  Note the difference between $E(T \mid A = a)$ and $E(T_a)$.

Statistical framework for causal inference

- Full data = ideal data: $X \sim F_X$
  - for point treatment data: $X = (W, (\bar{Y}_a(T_a))_{a \in \mathcal{A}})$ or a simpler representation is $X = (W, (T_a)_{a \in \mathcal{A}})$
  - for longitudinal data: $X = (\bar{L}_{\bar{a}}(T_{\bar{a}}))_{\bar{a} \in \mathcal{A}}$ or a simpler representation is $X = (T_{\bar{a}}, \bar{W}_{\bar{a}}(T_{\bar{a}}))_{\bar{a} \in \mathcal{A}}$

  In particular, it includes all treatment-specific outcomes and baseline covariates.
- Observed data = only available data: $O \sim P$
  - for point treatment data: $O = (W, A, \bar{Y}(T))$ or a simpler representation is $O = (W, A, T)$
  - for longitudinal data: $O = (\bar{L}(T), \bar{A}(T-1)) = (L(0), A(0), L(1), A(1), \ldots, A(T-1), L(T))$ or a simpler representation is $O = (T, \bar{W}(T), \bar{A}(T-1))$

$\longrightarrow$ **The longitudinal data notations cover the point treatment data notations (more general)**

Defining the survival causal parameter of interest

A **(conditional) survival causal effect** can be described by the following parameters:

- at the subject/unit level by $\log T_{\bar{a}_1} - \log T_{\bar{a}_2}$,
- at the population level by
  - $\beta_{\bar{a}_1,\bar{a}_2}(V) = E(\log T_{\bar{a}_1} - \log T_{\bar{a}_2} \mid V) = E(\log T_{\bar{a}_1} \mid V) - E(\log T_{\bar{a}_2} \mid V)$,
  - $\beta_{\bar{a}_1,\bar{a}_2}(V) =$ median$(\log T_{\bar{a}_1} \mid V)$-median$(\log T_{\bar{a}_1} \mid V)$,
  - etc.

If $V = \emptyset$, $\beta_{\bar{a}_1,\bar{a}_2}$ describe marginal causal effects unlike $\beta_{\bar{a}_1,\bar{a}_2}(V)$ describe adjusted causal effects.
Typically it is the average causal effect that is of interest and it is described by the parameter:

$$\beta^* = \left(\beta_{\bar{a}_1,\bar{a}_2}(V)\right)_{V,\bar{a}_1,\bar{a}_2}, \text{ where } \beta_{\bar{a}_1,\bar{a}_2}(V) = E(\log T_{\bar{a}_1} \mid V) - E(\log T_{\bar{a}_2} \mid V)$$

Defining the survival causal parameter of interest

Such a parameter β^* is typically very high-dimensional and nonparametric estimation of β is then not possible with finite sample data or suffer from poor practical performance: curse of dimensionality.

A parametric model can be used to summarize a very high-dimensional parameter into a lower-dimensional parameter of interest β.

A parametric Marginal Structural Model describes average causal effects with a lower-dimensional parameter β using the mean feature of the counterfactuals distribution:

$$E(\log T_{\bar{a}} \mid V) = m(\bar{a}, V \mid \beta),$$

e.g. $\beta = (\beta_1, \beta_2, \beta_3)$ and $m(\bar{a}, V \mid \beta) = \beta_0 + \beta_1 \text{mean}(\bar{a}) + \beta_2 V + \beta_3 \text{mean}(\bar{a}) V$.

$\longrightarrow$ β is the survival causal parameter of interest

Defining the survival causal parameter of interest

We proposed to describe survival causal effects by comparing mean treatment-specific $\log T_{\bar{a}}$ per strata of V however survival causal effects can be described by comparing other mean treatment-specific survival outcomes like:

- the hazard of survival per strata of V: $\lambda_{\bar{a}}(t \mid V) = P(T_{\bar{a}} = t \mid T_{\bar{a}} > t, V)$
- the survival function per strata of V: $S_{\bar{a}}(t \mid V) = \Pi_{j \leq t}(1 - \lambda_{\bar{a}}(j \mid V)$.

Similarly , Marginal Structural Models modelling can be used to model these other expected treatment-specific outcomes and describe the average survival causal effects using different parameters of interest.

$\longrightarrow$ In this presentation, we only consider MSMs of the type $E(\log T_{\bar{a}} \mid V) = m(\bar{a}, V \mid \beta)$ but results can be generalized to more general MSMs

Naive approach to MSM estimation

Consider the estimation problem of the parameter $\beta = (\beta_0, \beta_1)$ defined by the following MSM for a point-treatment data set:

$$E(\log T_a) = \beta_0 + \beta_1 a$$

A naive estimation approach would consist in 1) performing a simple regression of $\log T$ on A using the association model $\alpha_0 + \alpha_1 A$ and 2) interpreting the resulting estimate $\hat{\alpha}$ as an estimate of β. This is saying that $E(\log T \mid A = a) = E(\log T_a)$ for all $a \in \mathcal{A}$.

This equality holds only if treatment A is randomized, i.e. $A \perp (Y_a)_{a \in \mathcal{A}}$.

In most cases, the effect of A on T is confounded and treatment A is thus not randomized. Using this naive estimation procedure leads to bias estimation of β.
$\longrightarrow$ This type of bias is similar to the bias induced by informative censoring, i.e. by missing data

How to estimate β: a missing data approach

The observed data O can be linked to the ideal data one would have liked to observe to investigate the average causal effect of interest.

The ideal data corresponds with n i.i.d. observations of the full data $X = (\bar{L}_{\bar{a}}(T_{\bar{a}}))_{\bar{a}\in\mathcal{A}} = (T_{\bar{a}}, \bar{W}_{\bar{a}}(T_{\bar{a}}))_{\bar{a}\in\mathcal{A}} \sim F_X$ and we can link X to O as follows:

$$O = \Phi(\bar{A}, X) = (\bar{A}(T-1), \bar{L}_{\bar{A}(T-1)}(T)) = (\bar{A}(T-1), T_{\bar{A}(T-1)}, \bar{W}_{\bar{A}(T-1)}(T)).$$

The problem of estimating β using O can thus be treated as a missing data problem. The estimation methods developed for missing data problems can thus be applied to the estimation of causal parameters defined by MSMs.

$\longrightarrow$ Certain conditions are necessary, at least in practice, for the identification of β with O.

Conditions for identification of β

- Correct MSM specification: $\exists\ \beta \in \mathbb{R} \quad E(\log T_{\bar{a}} \mid V) = m(\bar{a}, V \mid \beta)$
- Existence of counterfactuals
- Time ordering: $\bar{L}_{\bar{a}}(t) = \bar{L}_{\bar{a}(t-1)}(t)$ (causal graph not deductible from the data only)
- Consistency assumption: $L(t) = L_{\bar{A}}(t)$
- Sequential Randomization Assumption (SRA):

$$A(t) \perp X \mid \bar{A}(t-1), \bar{L}(t)$$

This assumption is called the No Unobserved Confounder assumption or Randomization assumption (RA) for point-treatment data. It is necessary **in practice** for identification of β when one is not willing to make specific assumptions about F_X likely not to hold in most cases.

$\longrightarrow$ Under these assumptions, it is possible to estimate β consistently with the available data O (under additional assumptions).

The Sequential Randomization Assumption

The SRA implies that:

$$g(A(t) \mid \bar{A}(t-1), X) = g(A(t) \mid \bar{A}(t-1), \bar{L}(t)).$$

$$(\Longleftrightarrow g(A \mid X) = g(A \mid W) \text{ for point-treatment data, i.e. RA})$$

If the SRA is violated, it might still be possible to identify β in certain specific situations (instrumental variables).

However if one does not want to make assumptions only valid in certain specific scenarios likely not to hold, the SRA is required for identification of a causal effect.

Under the SRA we thus have:

$$g(\bar{A}(T-1) \mid X) = \prod_{t=0}^{T-1} g(A(t) \mid \bar{A}(t-1), \bar{L}(t))$$

Intuition behind the SRA

Intuitively the SRA means that: the way a treatment variable is assigned at each time point in *Reality* using the full data, also called the treatment mechanism, should only depend on **PAST OBSERVED** variables, i.e. in particular it cannot depend on unobserved confounders.

The SRA implies that we had enough information to predict at each time point the treatment received (based one the full data) using the past observed variables at that time only. Thus, the SRA is an assumption dealing with the information available in the observed data.

It can be viewed as one of the minimal assumptions that insures that there is enough information in the observed data O to identify a parameter (like β) defined using the unavailable full data X.

Estimating β: a missing data approach

Three estimators developed for missing data problems can be used to estimate causal effects like β under the SRA assumption. Under the SRA assumption, the likelihood factorizes into two parts:

$$\mathcal{L}(O) = \underbrace{f(L(0)) \underbrace{\prod_{j=1}^{k+1} f(L(j) \mid \bar{L}(j-1), \bar{A}(j-1))}_{Q_{F_X}}}_{F_X \text{ part}} \overbrace{g(\bar{A} \mid X)}^{g \text{ part}}.$$

Under the SRA, we denote the distribution of O, P, with $P_{F_X,g}$.

Thus the three estimators of β can rely on models for different part of the observed likelihood:

- Inverse Probability of Treatment Weighted (IPTW) estimator
- G-computation (G-comp) estimator
- Double Robust (DR) estimator.

$\longrightarrow$ We focus on the IPTW estimator of β in this presentation.

The IPTW estimator: definition

Definition:

The IPTW estimating function for β with nuisance parameter g is defined as:

$$D_h(O \mid g, \beta) = \frac{h(\bar{A}, V)\epsilon(\beta)}{g(\bar{A} \mid X)} \text{ where } \varepsilon(\beta) = \log T - m(\bar{A}, V \mid \beta).$$

Note that the IPTW estimating function is indeed a function of the observed data under the SRA.

We denote the estimator of the nuisance parameter g with g_n.

The IPTW estimator of β is defined as the solution of the estimating equation associated with the observed data O and the IPTW estimating function at g_n:

$$\sum_{i=1}^{n} D_h(o_i \mid g_n, \beta) = 0,$$

where o_i for $i = 1, \ldots, n$ represents the n i.i.d. observations in the observed data.

Property of the IPTW estimator

The IPTW estimating function is unbiased at β:

$$E_{P_{F_X,g}} D_h(O \mid \beta, g) = 0,$$

if the Experimental Treatment Assignment (ETA) assumption holds:

$$max_{\bar{a} \in \mathcal{A}} \frac{h(\bar{a}, V)}{g(\bar{a} \mid X)} < \infty \quad F_X - ae$$

" at each time point, treatments are NOT DETERMINISTICALLY assigned according to observed covariate values"

$\longrightarrow$ More than required to hold in theory: it cannot be practically violated

If g_n is a consistent estimator of g and the ETA assumption holds for g then the IPTW estimator is asymptotically linear and thus consistent.

$\longrightarrow$ the consistency of the IPTW estimator relies on correct specification of the g part of the likelihood AND the ETA assumption

Property of the IPTW estimator

Proof:

$$
\begin{aligned}
E_{P_{F_X,g}}\left[D_h(O \mid g,\beta)\right] &= EE\left(\frac{h(\bar{A},V)\epsilon(\beta)}{g(\bar{A} \mid X)} \mid X\right) \\
&= E_{F_X}\left(\sum_{\bar{a}:g(\bar{a}|X)\neq 0} \frac{h(\bar{a},V)\epsilon_{\bar{a}}(\beta)}{g(\bar{a} \mid X)} g(\bar{a} \mid X)\right) \\
&\overset{\text{ETA}}{=} E_{F_X}\left(\sum_{\bar{a}} h(\bar{a},V)\epsilon_{\bar{a}}(\beta)\right) \\
&= \sum_{\bar{a}} E_{F_X}\left(h(\bar{a},V)\epsilon_{\bar{a}}(\beta)\right) \\
&= \sum_{\bar{a}\in\mathcal{A}} E_{F_V} E(h(\bar{a},V)\epsilon_{\bar{a}}(\beta) \mid V) \\
&= \sum_{\bar{a}\in\mathcal{A}} E_{F_V}(h(\bar{a},V)E(\epsilon_{\bar{a}}(\beta) \mid V)) \\
&= 0,
\end{aligned}
$$

where $\varepsilon_{\bar{a}}(\beta) = \log T_{\bar{a}} - m(\bar{a}, V \mid \beta)$ and $\varepsilon(\beta) = \varepsilon_{\bar{A}}(\beta)$ under the consistency assumption.

The IPTW estimator: implementation

The IPTW estimate of β can be obtained in practice by performing a weighted least squares regression of Y on $\bar{A}$ and V using the MSM and weights inversely proportional to the treatment mechanism:

$$w(\bar{A}, V) = \frac{\lambda(\bar{A}, V)}{g_n(\bar{A} \mid X)} \underline{\underline{SRA}} \frac{\lambda(\bar{A}, V)}{\prod_{t=0}^{K} g_n(A(t) \mid \bar{A}(t-1), \bar{L}(t))},$$

where λ can be any non-null function of $\bar{A}$ and V.

It can indeed be shown that the resulting estimate is a solution of the IPTW estimating equation where $h(\bar{a}, V) = \lambda(\bar{a}, V)\frac{d}{d\beta}m(\bar{a}, V \mid \beta)$.

The IPTW estimator: implementation

Robins, Hernan and Brumback recommended the following choice for λ: $\lambda(\bar{A}, V) = g^{'}(\bar{A} \mid V)$ where $g^{'}$ is the conditional distribution of $\bar{A}$ given V:

- to improve the efficiency of the IPTW estimator (more stable weights)
- to be consistent with the naive estimation approach we would use if we know the treatment is randomized per strata of V.

The resulting weights are called stabilized weights: $w(\bar{A}, V) = \frac{g_n^{'}(\bar{A}|V)}{g_n(\bar{A}|X)}$.

In addition, it was shown, see van der Laan and Robins (2002), that g should always be estimated even when g is known. As a result, the IPTW estimator may gain in efficiency by taking into account possible empirical confounding, without loss of consistency.

The IPTW estimator: implementation

Do not trust the standard errors provided by the weighted regression routines in standard statistical package.

They assume the weights provided are known, i.e. not estimated. The resulting confidence intervals would be conservative.

Use the bootstrap to obtain correct standard errors and confidence intervals or calculate them with the influence curve of the IPTW estimator you used.

The IPTW estimator

An intuitive understanding of the IPTW estimator?

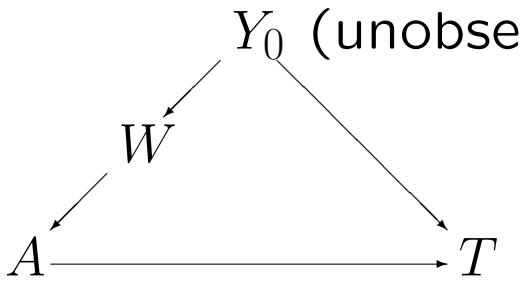
The IPTW estimator can be viewed intuitively as a "smart" weighted regression accomplishing two steps simultaneously:

1. Modify the original data such that the treatment is randomized in the ghost data artificially created by the weights: importance of the SRA, ETA assumptions and the models for the weights.

2. Estimate β using a simple (unweighted) mean regression of Y on A using the MSM and the ghost data.

Importance of the ETA assumption: violation AND practical violation

Illustration by simulations:



MSM: $E(\log T_a) = 2 - 5a$

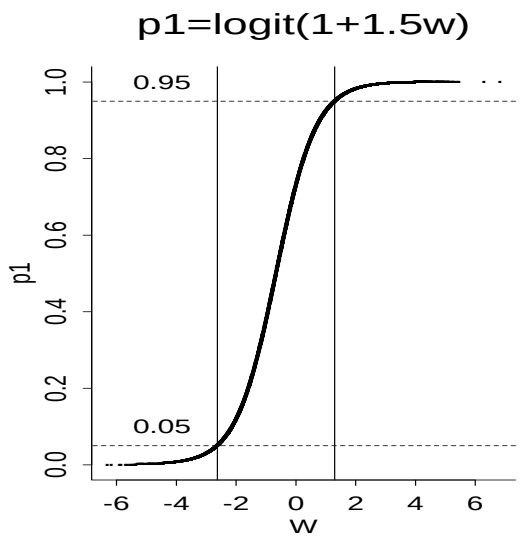
To obtain one data set with N observations, repeat N times:

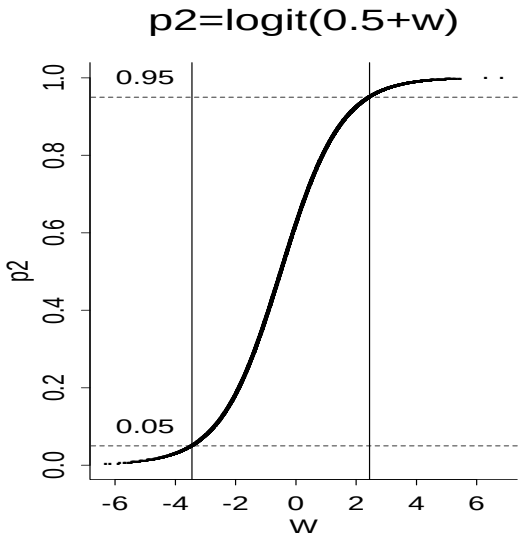
1. Generate $Y_0 \sim \mathcal{U}[-10, 10]$
2. Generate $W \sim \mathcal{N}(\frac{Y_0}{3}, 1)$
3. Generate A using $g(A \mid W)$
4. Generate $\log T \sim \mathcal{N}(2 + 4Y_0 - 5A, 1)$

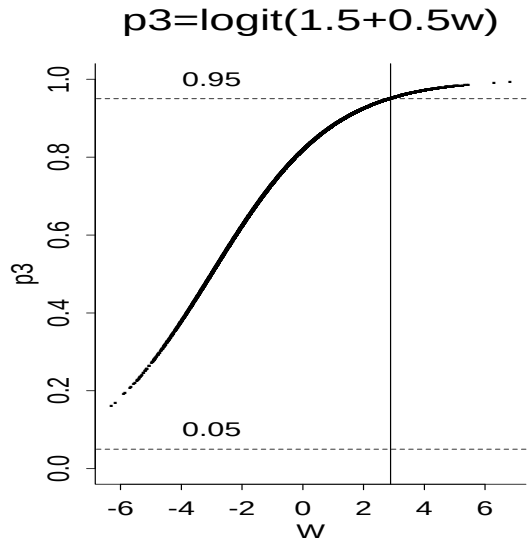
One generates 500 data sets for each g considered and each sample size $N = 100,\ 200,\ 300,\ 400,\ 500,\ 1000,\ 2000,\ 100000$.

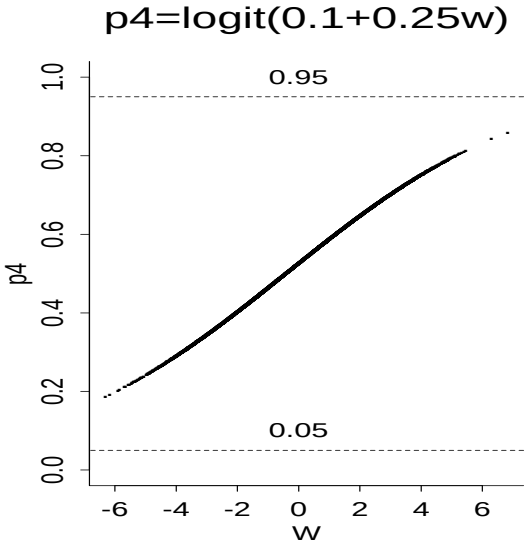
$\longrightarrow$ For each data set, one estimates $\beta = (2, 5)$ using the IPTW and reports the mean estimates per g and N considered.

Binary treatment and logistic treatment mechanism



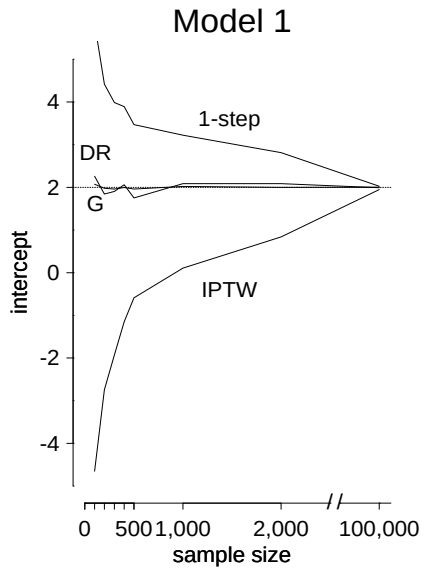


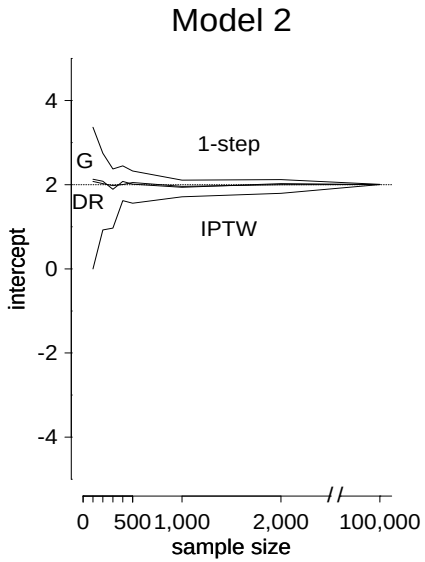


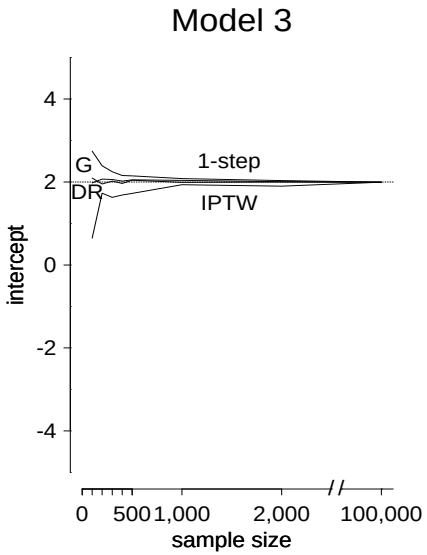


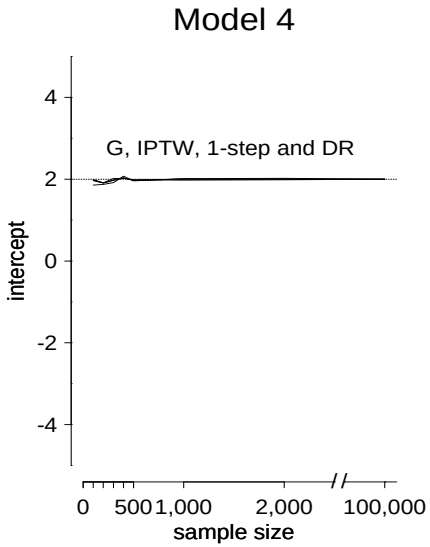
$\longrightarrow$ the ETA assumption becomes less practically violated as we go from model 1 to 4

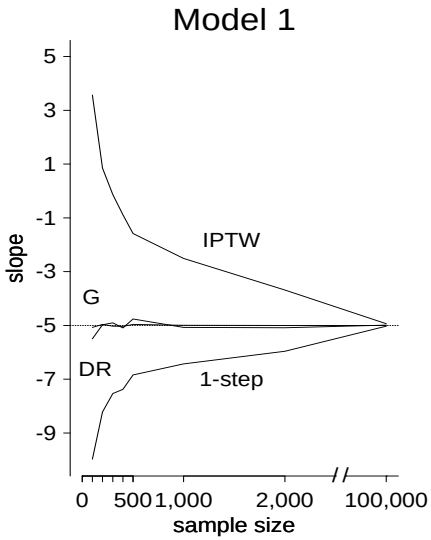
Comparison of the four estimators: bias

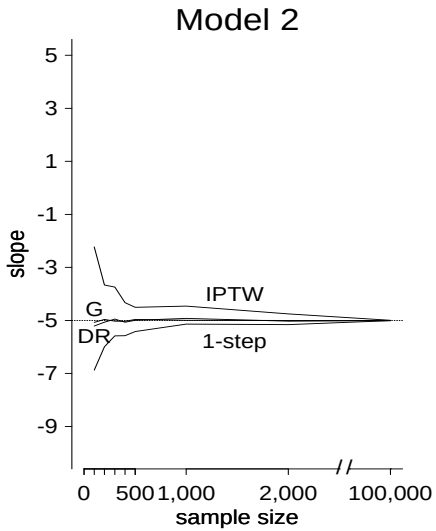


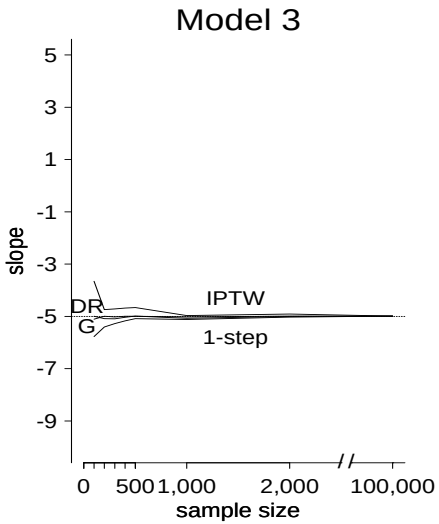


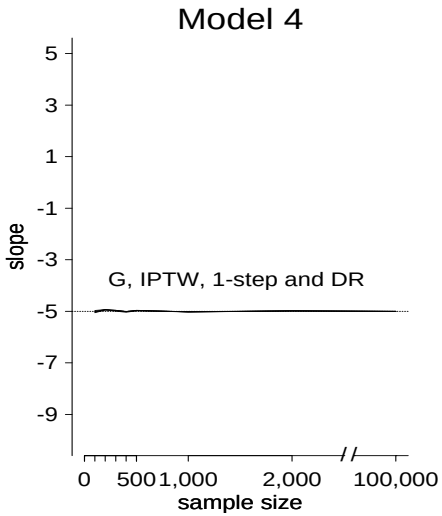






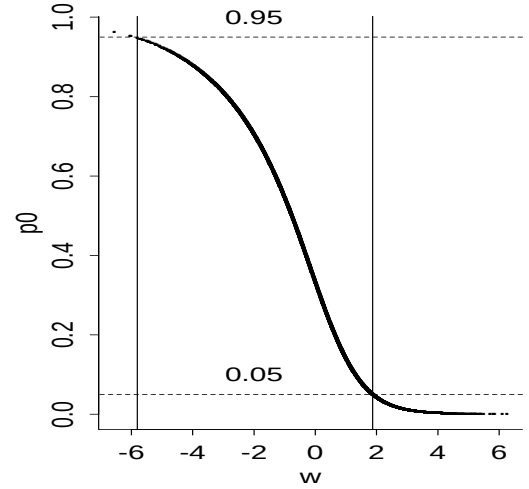




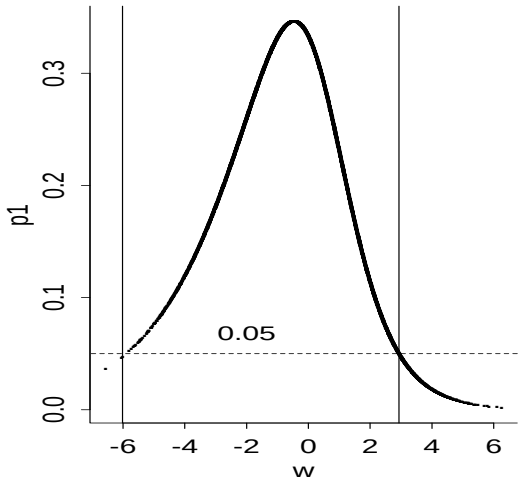


Categorical treatment and multinomial treatment mechanism

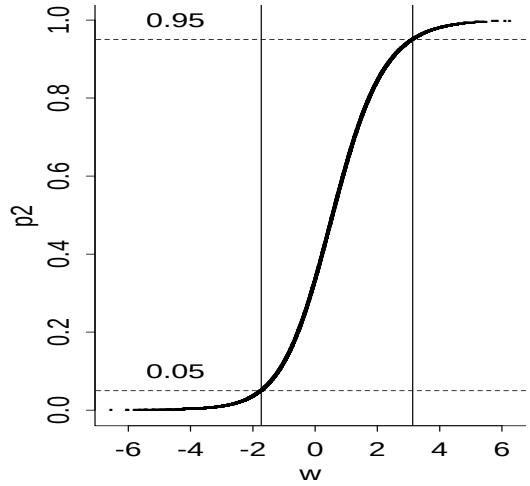
p0=1/[1+exp(0.5w)+exp(1.5w)]



p1=p0exp(0.5w)

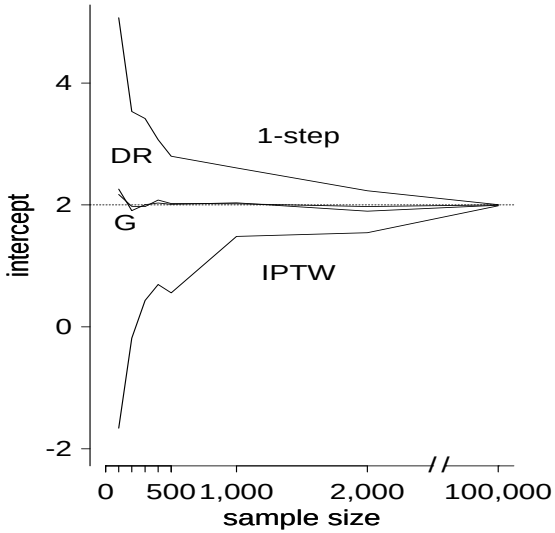


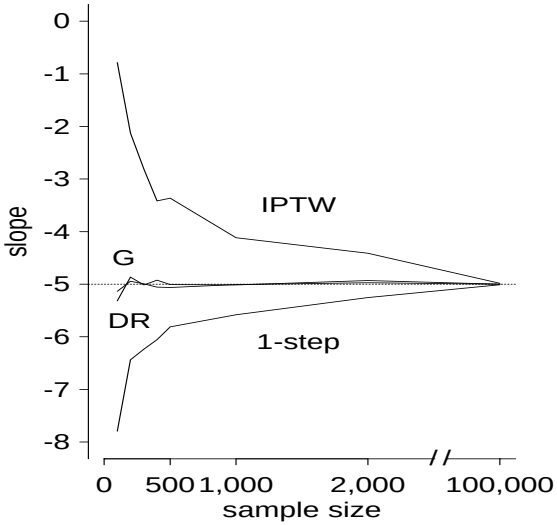
p2=p0exp(1.5w)



$\longrightarrow$ the ETA assumption is practically violated

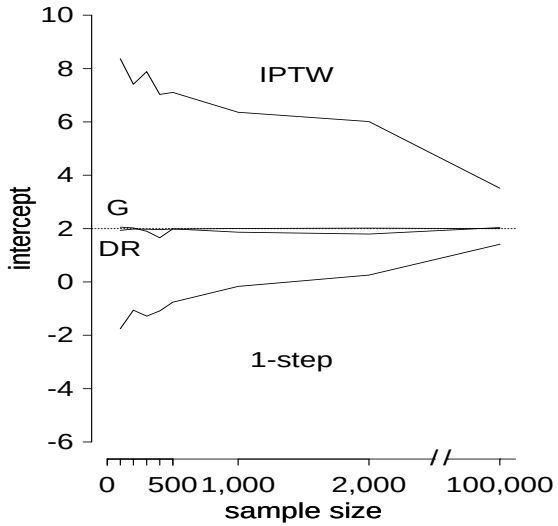
Comparison of the four estimators: bias

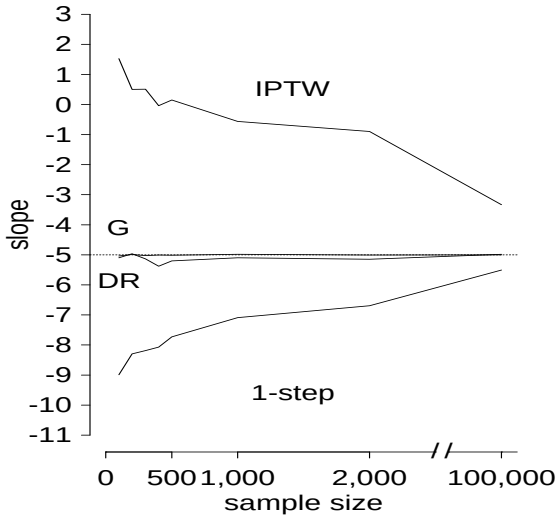




Continuous treatment and gaussian treatment mechanism

$$A \sim \mathcal{N}(W - 1, 1)$$





$\longrightarrow$ The IPTW estimates are still biased at $N = 100000$!

Importance of the ETA assumption

These simulation results should convince you of the importance of checking the validity of the ETA assumption in practice.

A visual check of the validity of the ETA assumption can be performed by plotting y) the observed treatment or predicted probability of treatment against x) the linear part of the treatment mechanism model.

The resulting plot should demonstrate that for any value of the covariates in the treatment mechanism model all treatment regimen are possible with a probability different enough from 0 or 1 (e.g. ≥ 0.1 and ≤ 0.9).

Generalization of the approach to censored data

We assume in this presentation that all events are observed for all subject in the data set.

In practice, censoring makes the problem more complex however the approach and the tools used to estimate causal effects remain the same.

Example of right censoring:
Censoring, C, is treated as another treatment variable and the counterfactuals are defined using a joint treatment $A(t) = (A_1(t), A_2(t))$ where $A_1(t)$ corresponds with the treatment of interest and $A_2(t) = I(C \leq t)$ and the MSM becomes:

$$E(\log T_{\bar{a}_1, \bar{a}_2=0} \mid V) = m(\bar{a}_1, V \mid \beta)$$

Other estimators

With or without censoring, the IPTW estimator fails to provide unbiased estimates of causal parameters of interest when the ETA assumption is violated or practically violated.

Other estimators should then be used. Two alternate estimators are possible: the G-computation and Double Robust (DR) estimators. They rely on the other part of the likelihood: the F_X part (and on the g part of the likelihood as well for the DR).