Marginal effects for time-varying treatments

Erica E. M. Moodie

Department of Epidemiology, Biostatistics, & Occupational Health
McGill University
Montreal, QC, Canada

erica.moodie@mcgill.ca

Session goals

- What parameters does an MSM actually estimate?
- When and why is an MSM needed?
- How can the parameters of an MSM be estimated?
- Assumptions, cautions, caveats.

Road map

1. Marginal effects in a longitudinal treatment setting
   - Definition
   - Failure of standard approaches
2. Three approaches to estimations
   - Inverse weighting
   - Forwards regression (g-computation)
   - Recursive regression (g-estimation)
3. Assumptions for each of the approaches
4. Simple worked example
5. Further considerations if time permits

The Marginal Structural Model

In longitudinal studies we observe for each individual i a sequence of exposures

$$Z_{i1}, Z_{i2}, \ldots, Z_{iJ}$$

and confounders

$$X_{i1}, X_{i2}, \ldots, X_{iJ}$$

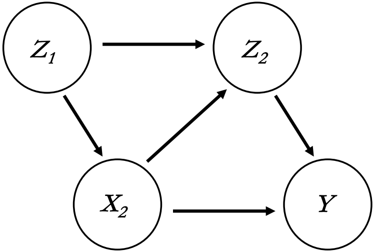
along with outcome $Y_i \equiv Y_{iJ}$ measured at the end of the study.

Intermediate outcomes $Y_{i1}, Y_{i2}, \ldots, Y_{i,J-1}$ also possibly available.

Some concepts, longitudinally

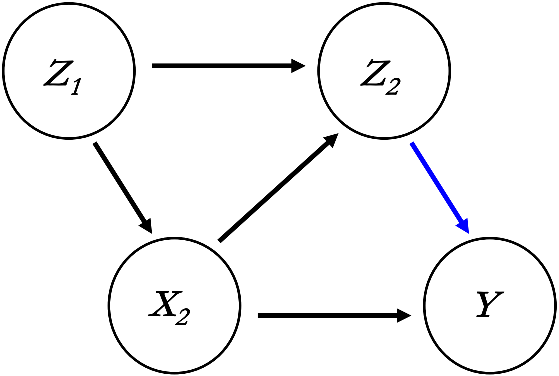
In a repeated measures or time-to-event setting, variables can be both intermediate and confounding.

Suppose we are interested in the total effects of treatments Z_1 and Z_2 on survival to time t, which we denote Y, in the presence of a time-dependent confounder X_2:



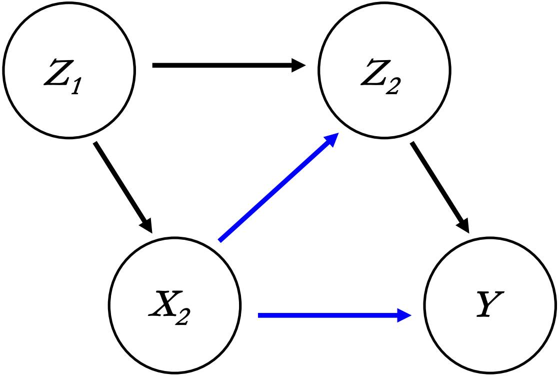
Some concepts, longitudinally

Z_2 affects Y directly:



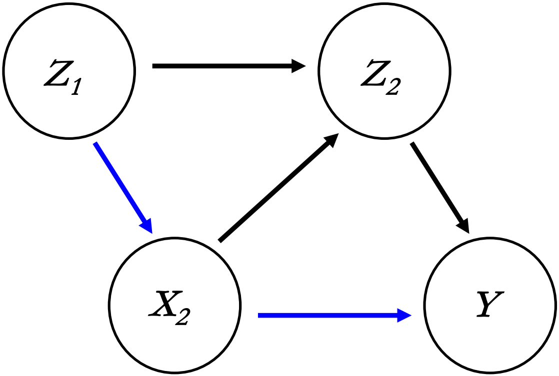
Some concepts, longitudinally

...and X_2 confounds the relationship between Z_2 and Y:



Some concepts, longitudinally

But Z_1 affects Y indirectly through X_2:

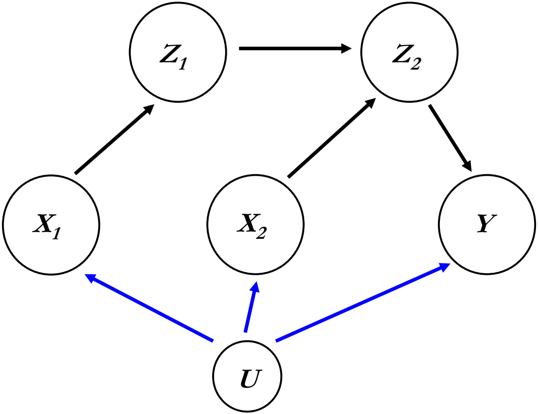


Some concepts, longitudinally

- Thus, if interested in the *total* effects of a *sequence* of treatment doses on an end-of-study response, standard regression models cannot be used.
- ...but what if there are no intermediate variables? Could we then condition on the time-depending confounders and use standard methods?
  - The answer in general is no.

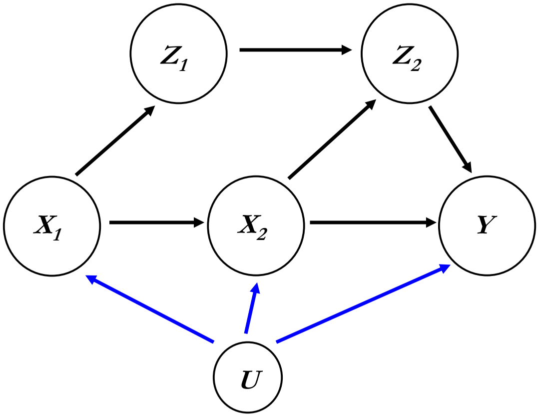
Some concepts, longitudinally

The potential for bias if there exists an unmeasured, underlying frailty:



Some concepts, longitudinally

Note that there are a variety of configurations that can lead to bias (including of the ‘collider-stratification’ variety):

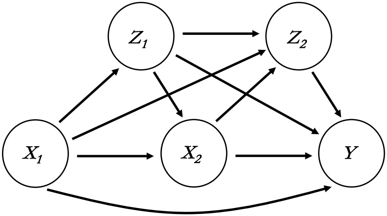


Marginal structural models

- Marginal Structural Models provide a powerful tool to assess the effects of exposures in longitudinal settings (can also be used for cross-sectional data).
  - Models are marginal because they pertain to population-average effects, structural because they describe causal (not associational) effects.
  - Most popular choice of model for data that exhibit time-varying confounding where the confounders are also mediators.

Potential outcomes: longitudinally

- Counterfactual or potential outcomes: the outcomes that would have been observed had a person been exposed to a particular treatment pattern.
- Consider a two-interval setting where data is collected at three times: baseline (t_0), t_1, and t_2, with covariates X_j measured at t_{j-1} ($j = 1, 2$), treatments Z_j taken between t_{j-1} and t_j ($j = 1, 2$), and outcome Y measured at t_2.



Potential outcomes

- There are four possible exposure patterns:
  - always exposed $(z_1, z_2) = (1,1)$,
  - never exposed $(z_1, z_2) = (0,0)$,
  - only exposed in one interval $(z_1, z_2) = (1,0)$, or $(0,1)$.

- We posit that each person has four responses (one corresponding to each exposure pattern), denoted $Y(1,1), Y(0,0), Y(1,0), Y(0,1)$, respectively.

Potential outcomes

- Suppose in reality an individual is treated in both intervals.
- We observed the outcome Y, which equals the counterfactual $Y(1,1)$ but we do not observe outcomes under the three other possible exposure patterns.
- Although we cannot observe most potential outcomes, we can use them to help formulate causal models.

Potential outcomes

- Rather than asking

  what is the average outcome among people who did receive treatment pattern $(z_1, z_2) = (1,1)$?

  we can ask

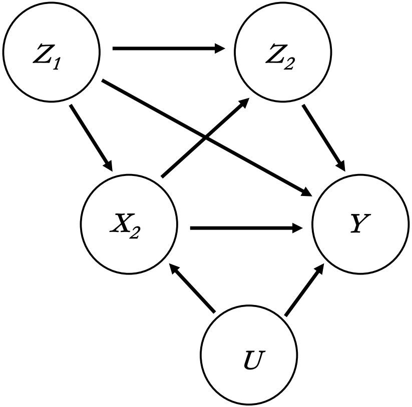
  what would be the average outcome among if everyone received treatment pattern $(z_1, z_2) = (1,1)$?

- An MSM is a model for $\mathbb{E}[Y(z_1, z_2)]$, i.e. the average outcome if the entire population was exposed to treatment pattern z_1, z_2, for each possible treatment pair.

Potential outcomes

- Since we can only ever observe one of the four counterfactuals, we can recast this as a missing data problem, and up- or down-weight individuals so as to create a population in which treatment receipt is not affected by time-varying covariates.
- Alternatively, we can again view this inverse probability weighting as an importance sampling approach.
- We create a pseudo-population of subjects in which the treatments Z_j and covariates X_j are unassociated, and therefore there exists no time-varying confounding.
    - Because there is no confounding, there is no need to condition on X_j.
    - By not conditioning on X_j, we do not block mediated pathways or induce collider-stratification bias.

Pseudo-population: What does it do?



Assumptions

What assumptions do we need to obtain an unbiased estimator of the marginal mean $\mathbb{E}[Y(z_1, z_2, \ldots, z_J)]$, via IPW, for some sequence of treatments $z_1, z_2, \ldots, z_J$?

- Correct model specification (of mean of Z_j given the past $\forall j$)
- No unmeasured confounding *at each interval* → sequential randomization
- Independence
- No extrapolation
- Well-defined exposure

Choice of weights

- Several options for the treatment weights. The simplest are unstandardized weights:

$$\begin{aligned} w &= \{\Pr(Z_1 = z_1 | X_1) \times \Pr(Z_2 = z_2 | X_1, Z_1, X_2)\}^{-1} \\ &= \frac{1}{\Pr(Z_1, Z_2 | X_1, X_2)}, \end{aligned}$$

i.e., each individual's weight is computed by taking the product of the estimated probability of receiving the treatment he actually received in each interval, conditional on past time-varying covariates (including, potentially, baseline covariates and previous treatment).

Choice of weights

- It is more common to use standardized weights:

$$sw = \frac{\Pr(Z_1 = z_1) \times \Pr(Z_2 = z_2 | Z_1)}{\Pr(Z_1 = z_1 | X_1) \times \Pr(Z_2 = z_2 | X_1, Z_1, X_2)}.$$

- These weights may still be quite variable, particularly if there are some individuals who received unusual treatments given their covariates $\rightarrow$ can normalize and/or truncate to further reduce variability.

MSM estimation

The MSM estimation procedure via IPW is straightforward:

1. Fit treatment models: fit a logistic regression model for the probability of being treated at each interval.
2. Determine the weights:
   (a) Use the models in step (1) to predict the probability that a person received the exposure pattern he did in fact receive, by taking the product of the probability of receiving the observed treatment in each interval.
   (b) Set each individual's weight to one over the probability computed in (2a). Optionally (recommended): stabilize, normalize, and/or truncate the weights.
3. Fit a response model: weighting each individual by the weights computed in (2b), use standard software to fit a regression model for the response given exposure and possibly baseline covariates.

MSM estimation

- All confounding covariates should be included in the treatment models; variables that predict only treatment (but not the outcome) can be omitted.
- Automated model selection (e.g., stepwise procedures) should not be used.

- The procedure outlined on the previous slide is valid for any type of outcome, including binary responses or time-to-event (survival) data.
- For time-to-event data, a weighted Cox model can be fit, or time can be discretized (e.g. into months) and a weighted logistic regression on status can be fit.

MSM: sample code for a two-interval example

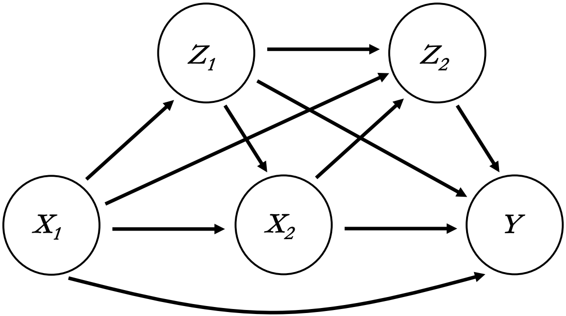
```
## First interval weight:
ps1 <- glm(Z1~X1,family=binomial)
w1 <- 1/ifelse(Z1==1,predict(ps1,type="response"),
                1-predict(ps1,type="response"))

## Second interval weight:
ps2 <- glm(Z2~X2+X1+Z1,family=binomial)
w2 <- 1/ifelse(Z2==1,predict(ps2,type="response"),
                1-predict(ps2,type="response"))

## Final weights, and MSM:
wt <- w1*w2  ## (unstabilized)
msm <- lm(Y~Z1+Z2,weights=wt)
#  msm <- lm(Y~Z1*Z2,weights=wt)
summary(msm)
```

MSM: (simulated) HIV example

- Suppose that researchers are interested in the effect of HAART interruptions on liver function in an HIV+ population.
- We simulate an example with $n = 100$ designed to follow the causal structure below:



MSM: HIV example

- Liver function is measured at baseline (X_1), six months (X_2), and 12 months (Y).
- Exposure Z_1 is a binary indicator of HAART interruption between baseline and month 6, and Z_2 the corresponding indicator for occlusion between months 6 and 12.
- Model 1 adjusts for baseline liver function (X_1) only; Model 2 adjusts for both baseline and six-month liver function (X_1 and X_2).

Table. Results from traditional regression models. True parameter values are -0.038, and -0.086.

	Model 1			Model 2		
Variable	$\hat{\beta}$	SE	% bias	$\hat{\beta}$	SE	% bias
Z_1	-0.036	0.0121	6.4	0.213	0.009	660.8
Z_2	-0.074	0.0121	14.2	-0.085	0.004	0.9

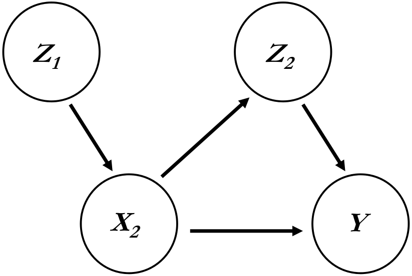
MSM: HIV example

- Results of the previous slide are based on a single data set of a modest size (n = 100).
- Repeating the simulation study 10,000 times:
  - the bias of the estimator of β for Z_2 in Model 1 is 6.0%
  - the bias of the estimator of β for Z_1 in Model 2 is 679%
- Repeating the simulation study 10,000 times for $n = 5000$:
  - the bias of the estimator of β for Z_2 in Model 1 is 6.3%
  - the bias of the estimator of β for Z_1 in Model 2 is 679%

  indicating that the bias does not diminish with increasing sample size.
- Using an MSM yields estimates (SE) of -0.039 (0.013) and -0.086 (0.013) for Z_1 and Z_2, respectively. Repeating the simulation 10,000 with $n = 100$: bias of less than 0.65% for each parameter.

Example 2: marginal structural models "by hand"

Consider a simple example, where treatment at each interval is beneficial, and receipt of treatment in the second interval is strongly dependent on the intermediate outcome X_2 (with X_2, Y indicative of a negative outcome):



Example 2: the data

Z_1	0								1							
(n)	500								500							
X_2	0				1				0				1			
(n)	311				189				365				135			
Z_2	0		1		0		1		0		1		0		1	
(n)	194		117		71		118		227		138		51		84	
Y	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
(n)	142	52	96	21	27	44	59	59	166	61	113	25	19	32	42	42

Following Diggle, Heagerty, Liang & Zeger, §12.5.

Example 2: observed associations

$$
\begin{aligned}
\widehat{\mathbb{E}}[Y|z_1 = 1, z_2 = 1] &= (25 + 42)/(138 + 84) = 0.30 \\
\widehat{\mathbb{E}}[Y|z_1 = 1, z_2 = 0] &= (61 + 32)/(227 + 51) = 0.33 \\
\widehat{\mathbb{E}}[Y|z_1 = 0, z_2 = 1] &= (21 + 59)/(117 + 118) = 0.34 \\
\widehat{\mathbb{E}}[Y|z_1 = 0, z_2 = 0] &= (52 + 44)/(194 + 71) = 0.36
\end{aligned}
$$

The benefit of treatment is not evident here, with 30% experiencing the outcome when receiving treatment in both intervals, compared to 36% when treatment-free, giving an OR of 0.76 (=(0.3/0.7)/(0.36/0.64)).

Example 2: observed associations

Alternatively, we obtain the following regression coefficient estimates:

$$
\begin{aligned}
logit(\mathbb{E}[Y|z_1 = 1, z_2 = 1]) &= -0.56 - 0.13z_1 \\
&\quad -0.10z_2 - 0.04z_1z_2.
\end{aligned}
$$

Again, this leads to an OR of 0.76 ($= \exp(-0.13 - 0.10 - 0.04)$) for the always- vs never-treated comparison.

Example 2: an (impossible) controller

What if we could control treatment assignment, so that our design is experimental rather than observational? Then:

Z_1	0	1
(n)	0	1000

X_2	0	1	0	1
(n)	0	0	731	269

Z_2	0	1	0	1	0	1	0	1
(n)	0	0	0	0	0	731	0	269

Y	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
(n)	0	0	0	0	0	0	0	0	0	0	598	133	0	0	134	134

Example 2: an (impossible) controller

Similarly, if we could prevent the entire population from receiving treatment, we would expect to see:

Z_1	0								1							
(*n*)	1000								0							
X_2	0				1				0				1			
(*n*)	622				378				0				0			
Z_2	0		1		0		1		0		1		0		1	
(*n*)	622		0		378		0		0		0		0		0	
Y	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
(*n*)	455	167	0	0	143	235	0	0	0	0	0	0	0	0	0	0

Example 2: results found by the (impossible) controller

$$\widehat{\mathbb{E}}[Y(z_1 = 1, z_2 = 1)] = (134 + 133)/(731 + 269) = 0.267$$
$$\widehat{\mathbb{E}}[Y(z_1 = 0, z_2 = 0)] = (167 + 235)/(622 + 378) = 0.402$$

The benefit of treatment is now more evident here, with 27% experiencing the outcome when receiving treatment in both intervals, compared to 40% when treatment-free, giving an OR of 0.54.

Example 2: IPW

Since we cannot, in reality, control treatment receipt, let us instead perform an analysis that acknowledges the simultaneous roles of X_2 as confounder and mediator.

First, we have to construct weights. We will use stabilized weights, of the form:

$$sw = \frac{1}{\Pr(Z_1 = z_1)} \cdot \frac{\Pr(Z_2 = z_2 | Z_1 = z_1)}{\Pr(Z_2 = z_2 | X_2 = x_2, Z_1 = z_1)}.$$

where

$$\begin{aligned}
\widehat{\Pr}(Z_1 = 1) &= 0.5 \\
\widehat{\Pr}(Z_2 = 1 | Z_1 = 0) &= 0.47 \\
\widehat{\Pr}(Z_2 = 1 | Z_1 = 1) &= 0.444 \\
\widehat{\Pr}(Z_2 = 1 | X_2 = 0, Z_1 = 0) &= 0.367 \\
\widehat{\Pr}(Z_2 = 1 | X_2 = 1, Z_1 = 0) &= 0.624
\end{aligned}$$

etc.

Example 2: IPW

Z_1	0								1							
(n)	500								500							
X_2	0				1				0				1			
(n)	311				189				365				135			
Z_2	0		1		0		1		0		1		0		1	
(n)	194		117		71		118		227		138		51		84	
Y	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
(n)	142	52	96	21	27	44	59	59	166	61	113	25	19	32	42	42
sw	0.85	0.85	1.25	1.25	1.40	1.40	0.76	0.76	0.89	0.89	1.17	1.17	1.47	1.47	0.71	0.71
*n**	**120.7**	**44.2**	**120**	**26.3**	**37.8**	**61.6**	**44.8**	**44.8**	**147.7**	**54.3**	**132.2**	**29.3**	**27.9**	**47.0**	**29.8**	**29.8**

Note that sw, the stabilized weights, and n, the sample size in the reweighted pseudo-population, have been rounded. In practice, rounding should only be done on the final estimate.*

Example 2: IPW

Note that we are not weighting by the probability of receiving treatment in both intervals,

$$\frac{1}{\Pr(Z_1 = 1)} \cdot \frac{\Pr(Z_2 = 1 | Z_1 = z_1)}{\Pr(Z_2 = 1 | X_2 = x_2, Z_1 = z_1)}.$$

but rather by the probability of having received the observed treatment combination (z_1, z_2)

$$sw = \frac{1}{\Pr(Z_1 = z_1)} \cdot \frac{\Pr(Z_2 = z_2 | Z_1 = z_1)}{\Pr(Z_2 = z_2 | X_2 = x_2, Z_1 = z_1)}.$$

Using the reweighted sample, we now find
$\widehat{\mathbb{E}}[Y(1,1)] = (29.3 + 29.8)/(132.2 + 29.3 + 29.8 + 29.8) = 0.27$, and
$\widehat{\mathbb{E}}[Y(0,0)] = (44.2 + 61.6)/(120.7 + 44.2 + 37.8 + 61.6) = 0.40$.

Recap

We have now seen that:

1. When there exists a time dependent confounder, X_j, that acts as a mediator, standard regression models fail.
2. When there exists a time dependent confounder, X_j, that is not a mediator, but there exists an unmeasured variable U that affects both X_j and the outcome, standard regression models fail.
3. IPW can be used to estimate total effects in a marginal structural model.

Are there any alternatives?

G-computation

In IPW, the focus is on modelling the treatment process so as to obtain the inverse weights.

In g-computation, the focus is instead of modelling the intermediate covariates, and then to simulate the data forward under treatment regimes of interest.

The basis of g-computation is the “telescoping” sequence of conditional distributions:

$$f(Y, X|Z) = \prod_{j=1}^{J} f(Y_j|H_j) \times f(X_j|H_{j-1}, Y_{j-1})$$

where $H_j = (X_1, Z_1, X_2, ..., X_{j-1}, Z_{j-1}, Y_{j-1}, X_j, Z_j)$.

Let's return to the simple example. We have only 1 intermediate covariate, so g-computation requires only models for $\Pr(Y = 1|Z_1, X_2, Z_2)$ and $\Pr(X_2 = 1|Z_1)$.

Then we can compute

$$
\begin{aligned}
\widehat{\mathbb{E}}[Y(1,1)] &= \sum_{x_2} \Pr(Y = 1|Z_1 = 1, X_2 = x_2, Z_2 = 1) \cdot \\
&\qquad \Pr(X_2 = x_2|Z_1 = 1) \\
&= (25/138) * (365/500) + (42/84) * (135/500) \\
&= 0.267
\end{aligned}
$$

Example 2: G-computation

For the no-treatment scenario, we find:

$$
\begin{aligned}
\widehat{\mathbb{E}}[Y(0,0)] &= \sum_{x_2} \Pr(Y = 1 | Z_1 = 0, X_2 = x_2, Z_2 = 0) \cdot \\
&\qquad \Pr(X_2 = x_2 | Z_1 = 0) \\
&= (52/194) * (311/500) + (44/71) * (189/500) \\
&= 0.401
\end{aligned}
$$

Assumptions

What assumptions do we need to obtain an unbiased estimator of the marginal mean $\mathbb{E}[Y(z_1, z_2, \ldots, z_J)]$, via g-computation, for some sequence of treatments $z_1, z_2, \ldots, z_J$?

- Correct model specification (of the mean of Y given the past, and of the *distribution* of X_j given the past $\forall j$)
- No unmeasured confounding at each interval
- Independence
- No extrapolation
- Well-defined exposure

Note that the first assumption may be difficult to satisfy for moderate dimensionality of X_j, especially if some elements are continuous-valued.

G-estimation

- Yet a third approach to estimating marginal models is known as g-estimation.
- Unbiasedness in the simplest g-estimation method comes through modelling the expected treatment, though there is also a doubly-robust version.
- The focus in g-estimation is on *contrasts* between the treated and untreated.
- The contrasts are called blip functions, and may be simple, e.g. $\gamma(z; h, \psi) = \psi z$ or more complex, e.g. $\gamma(z; h, \psi) = z(\psi_0 + \psi_1 x)$.

Crash course in EEs

Additional considerations

Summary

G-estimation in one interval

In a one-interval setting, g-estimation proceeds as follows:

1. Specify a blip function, $\gamma(z; h, \psi)$ that parameterizes the effect of treatment Z on outcome Y (possibly modified by covariates X).
2. Specify a treatment model, $\mathbb{E}[Z|X; \alpha]$ and estimate its parameters (e.g. via logistic regression for binary Z).
3. Letting $S(z) = \frac{\partial}{\partial \psi}\gamma(z; h, \psi)$, solve the g-estimating equation:

$$U(\psi) = \sum_{i=1}^{n} \{[Y_i - \gamma(z_i; h_i, \psi)] \cdot [S(z_i) - \mathbb{E}(S(Z_i)|x; \alpha)]\} = 0.$$

Note that this g-EE is unbiased when the treatment model, $\mathbb{E}[Z|X; \alpha]$, is correctly specified.

Assumptions for simple g-estimation

- Correct model specification (of mean of Z given X)
- No unmeasured confounding
- Independence
- No extrapolation
- Well-defined exposure

G-estimation in one interval: double robustness

In a one-interval setting, the g-estimation procedure can be made 'doubly robust' – and can yield more efficient estimators – by additional positing a model for the treatment-free outcome.

- Let $G(\psi) = Y - \gamma(z; h, \psi)$. The $G(\psi)$ is the (possibly counterfactual) treatment-free outcome.
- Let $\mathbb{E}[G(\psi)|h; \eta]$ parameterize a model for the expected value of $G(\psi)$. Note that we can re-write this to see that

$$\mathbb{E}[Y|h; \psi, \eta] = \mathbb{E}[G(\psi)|h; \eta] + \gamma(z; h, \psi).$$

- With $S(z) = \frac{\partial}{\partial \psi}\gamma(z; h, \psi)$, the following is a doubly-robust g-estimating equation:

$$\begin{aligned} U(\psi) &= \sum_{i=1}^{n} \left\{ [Y_i - \gamma(z_i; h_i, \psi) - \mathbb{E}(G(\psi)|h; \eta)] \cdot [S(z_i) - \mathbb{E}(Z_i|x; \alpha)] \right\} \\ &= 0. \end{aligned}$$

This g-EE is unbiased when either $\mathbb{E}[Z|X; \alpha]$ or $\mathbb{E}[G(\psi)|h; \eta]$ is correctly specified.

DR g-estimation in multiple intervals

In the multiple interval setting, we need to be careful in our specification of the blip.

We want it to parameterize the following:

$$\gamma_j(z_j; h_j, \psi_j) = \mathbb{E}[Y(z_1, ..., z_j, 0, ..., 0) - Y(z_1, ..., z_{j-1}, 0, ..., 0)],$$

i.e. it is a model for the ('true') effect of being treated in the j^{th} interval, given treatment history $z_1, ..., z_{j-1}$ and assuming no treatment in all subsequent intervals.

This particular form of blip is called a "zero-blip-to-zero" function.

DR g-estimation in multiple intervals

In a J-interval setting, g-estimation proceeds from the last interval to the first, recursively estimating the blip parameters:

1. At each interval, specify a blip function, $\gamma_j(z_j; h_j, \psi_j)$.
2. At each interval, specify a treatment model, $\mathbb{E}[Z_j|X_j; \alpha_j]$ and estimate its parameters.
3. At the last interval, J, set $G_J(\psi_J) = Y - \gamma_J(z_J; h_J, \psi_J)$, and specify $\mathbb{E}[G_J(\psi_J)|h_J; \eta_J]$.

DR g-estimation in multiple intervals (cont.)

4. Solve $U_J(\psi_J) = 0$.
5. For $j = J - 1, ..., 1$:
   i. Set $G_j(\psi_j) = Y - \sum_{k \geq j} \gamma(z_k; h_k, \psi_k)$, and specify $\mathbb{E}[G_j(\psi_j)|h_j; \eta_j]$.
   ii. Solve $U_j(\psi_j) = 0$.

$$\text{where } U_j(\psi_j) = \sum_{i=1}^{n} \Big\{ [Y_i - \sum_{k \geq j} \gamma(z_{ki}; h_{ki}, \psi_k) - \mathbb{E}(G_j(\psi_j)|h_j; \eta_j)] \cdot [S_j(z_{ji}) - \mathbb{E}[S(Z_{ji}|x_j; \alpha_j)]] \Big\} \text{ for } j = 1, ..., J.$$

DR g-estimation: further considerations

- Like IPW and g-computation, sequential randomization (i.e. no confounders at each interval) is required.
- The treatment models can be allowed to share parameters.
- The blip models can be allowed to share parameters, but the estimation is then more complicated: recursive estimation no longer appropriate and it is difficult to solve the g-EE for all intervals simultaneously.

DR g-estimation: further considerations

- Alternative blip models can also be specified to allow estimation of more complex treatment strategies, e.g. instead of all-or-nothing contrasts, we can specify ‘optimal’ blip functions that allow us to estimate optimal, personalized treatment strategies.
- Applying g-estimation to continuous exposures is straightforward.
- Applying g-estimation to binary or time-to-event outcomes is often not.
- `DRTreg` package in `R` can be used for estimation.

▸ Summary

Additional considerations: SEs and CIs

- All of the approaches considered (IPW/AIPW, g-computation, g-estimation) rely on substitution estimators.
  - In IPW, we plug in estimated weights; AIPW uses plug in weights and mean outcomes.
  - In g-computation, we simulate the outcome using estimated models.
  - In g-estimation, we plug in estimated weights and possibly also estimated treatment-free outcomes.
- We need to account for this when estimating standard errors and/or confidence intervals.
- Analytically derived asymptotic variances can also be used, but are not provided in standard software packages.
- The easiest approach is to bootstrap.

Additional considerations: missing data

- If data are missing intermittently, one can either impute or censor an individual at the first instance of missing data.
- Censored data (drop-out) can easily be handled by incorporating weights for censoring into the regression model or estimating equation for any of the three approaches that we have considered.

Additional considerations: timing of the exposure

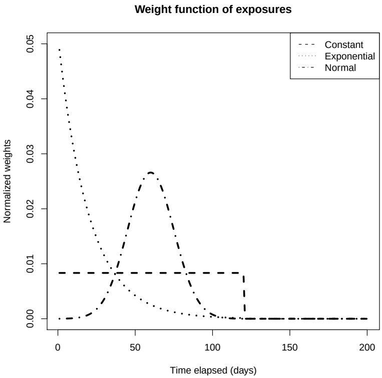
When covariates are time-dependent, there are several aspects of the analysis that require consideration

- Cumulative vs. current exposure
  - Driven by scientific/biological effect
  - E.g. Current smoking status or pack-years smoked?, dose of a medication since last visit or dose since start of study?, etc.
- Time lag
  - Again, driven by scientific/biological effect
  - E.g. Incubation period of pathogens, latency period for carcinogens/cancer, etc.

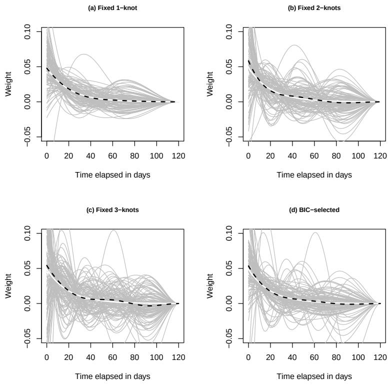
Additional considerations: timing of the exposure

- Whenever possible, use biological knowledge to inform the model.
- A commonly used approach is to cumulate the exposure (summing the number of exposed intervals).
- Although there are some models that try to learn about this lag from the data ("weighted cumulative exposure" models), these can be very unreliable.
  - WCE models include an indicator for exposure (yes/no) over many, many lagged time points and then attempts to learn from the data how exposure affects the outcome using a smooth function.

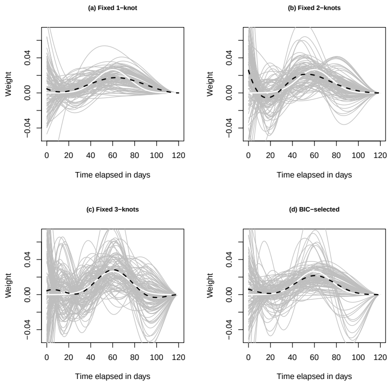
Idealized weight functions



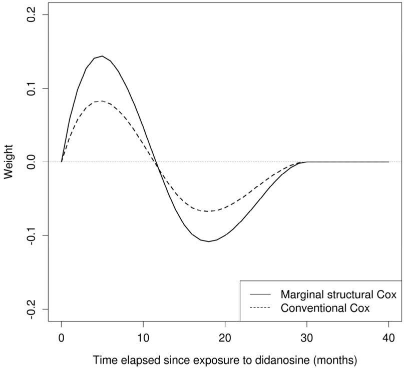
Realized weight functions (simulated data, exponential weights)



Realized weight functions (simulated data, normal weights)



An actual (estimated) weight function



Time-varying covariates & causal inference: Summary

- We have looked at three approaches to estimating marginal structural models:
  - Marginal structural models
  - G-computation
  - G-estimation
- The first of these provides marginal (population average) parameters, while the other two can provide results that are conditional on time-varying covariates.
- These approaches can be used where standard methods fail: in particular, when time-varying covariates exist and act as mediating variables.

Key points: Summary

- When exposures vary over time, there is a the potential for greater complexity in the data structure, particularly if variables act as both mediators and confounders.
- MSMs are straight-forward to compute and allow estimation of a range of parameters including population average effects under specific exposure patterns, and the decomposition of effects into the effect mediated through a variable and the “remainder” of the effect which is not.
- G-computation and G-estimation require a bit more statistical and computational expertise to implement, but also afford more flexibility.

Key points: Why use MSMs?

- Standard regression models yield biased estimates of treatment effects when:
  - (i) in a time-dependent exposure setting, some mediating variables are also confounders, or
  - (ii) in a time-dependent exposure setting, there exists an unmeasured variable that causes changes in a confounder and the outcome, or
  - (iii) in a mediation analysis if any of the confounders of mediator-outcome relationship are caused by the exposure.
- MSMs can also be used in other settings to adjust for confounding, e.g. a cross-sectional study (but typically aren't: why use them if a simpler approach will suffice?)
- MSMs are useful even in RCTs – not for an ITT analysis, but for secondary analyses when there is non-compliance or attrition.

Key points: Why use MSMs?

- MSMs are often criticized for their reliance on strong assumptions, however *all* statistical analyses rely on assumptions, many of which are the same.
- In a standard regression setting (e.g. cross-sectional data), let's quickly review the "MSM assumptions":
  - No unmeasured confounding.
  - Correct model specification (treatment and response).
  - Exposed and unexposed individuals at every covariate combination (positivity).
  - Exposures must be well-defined.
  - No interference between participants.

  Each of these is required to draw sensible interpretations from a standard regression model!

Acknowledgements

Collaborator: David Stephens (McGill)

Selected references:

- Robins, Hernán, & Brumback (2000) Marginal structural models and causal inference in epidemiology. *Epidemiology*, **11**: 550–560.
- Hernán, Brumback, & Robins (2000) Marginal structural models to estimate the causal effect of Zidovudine on the survival of HIV-positive men. *Epidemiology*, **11**: 561–570.
- Bryan, Yu, & van der Laan (2004) Analysis of longitudinal marginal structural models. *Biostatistics*, **5**: 361–380.
- Bodnar, Davidian, Siega-Riz, & Tsiatis (2004) Marginal structural models for analyzing causal effects of time-dependent treatments: An application in perinatal epidemiology. *American Journal of Epidemiology*, **159**: 926–934.

Acknowledgements

References, continued:

- *Analysis of Longitudinal Data* (2nd ed., 2002) by Diggle, Heagerty, Liang and Zeger.
- Moodie & Stephens (2010) Using Directed Acyclic Graphs to detect limitations of traditional regression in longitudinal studies. *Int J Public Health*, **55**: 701–703.
- Moodie & Stephens (2011) Marginal structural models: Unbiased estimation for longitudinal studies. *Int J Public Health*, **56**: 117–119.
- Thorpe, Saeed, Moodie, and Klein (2011) Antiretroviral treatment interruption leads to progression of liver fibrosis in adults co-infected with HIV and Hepatitis C. *AIDS*, **25**: 967–975.
- Causal Inference (2016?) by Hernán & Robins, www.hsph.harvard.edu/miguel-hernan/causal-inference-book/

A crash course in estimating functions

Two main approaches to inference

Estimating Functions:

- A function of the parameter and data, $U(\theta, Y)$, of the same dimensionality as the parameter, for which $\mathbb{E}[U(\theta, Y)] = 0$ is considered.
- The EF estimator is then found as the solution to the estimating equation $U(\hat{\theta}, Y) = 0$.
- For inference, the frequency properties of the estimating function are derived and these are transferred to the resultant estimator.
- Often the estimating function is derived from a likelihood.

Bayesian:

- In addition to the likelihood $p(\boldsymbol{y}|\theta)$, specify a prior distribution $\pi(\theta)$.
- Then via Bayes theorem derive the posterior distribution $p(\theta|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\theta) \times \pi(\theta)}{p(\boldsymbol{y})}$.
- All inference follows from the posterior distribution.

Estimating functions

- An estimating function is a function

$$U_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} U(\theta, Y_i) \quad (1)$$

of the same dimension as θ for which

$$\mathbb{E}[U_n(\theta)] = 0 \quad (2)$$

for all θ.

- The estimating function $U_n(\theta)$ is a random variable because it is a function of Y.
- Maximum likelihood estimation is a special case of estimating equations with the score (deriv. of log likelihood) acting as the EF.

Estimating functions

The corresponding estimating equation that defines the estimator $\widehat{\theta}_n$ has the form

$$U_n(\widehat{\theta}_n) = \frac{1}{n}\sum_{i=1}^{n} U(\widehat{\theta}_n, Y_i) = {}_nU(\widehat{\theta}_n, Y_i) = 0. \qquad (3)$$

Suppose that $\widehat{\theta}_n$ is a solution to the estimating equation $U_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} U(\theta, Y_i) = 0$, i.e. $U_n(\widehat{\theta}_n) = 0$. Then

$$\begin{aligned} Var\left[U_n(\widehat{\theta}_n)\right] &= \mathbb{E}\left[(U(\theta, Y) - \mathbb{E}[U(\theta, Y)])^{\otimes 2}\right] \\ &= \mathbb{E}[U(\theta, Y)U(\theta, Y)^T] \end{aligned}$$

Now $U_n(\widehat{\theta}_n)$ is a sum of conditionally independent terms, so under regularity conditions (see Van der Vaart, 1998) we have

$$U_n(\widehat{\theta}_n) \sim \mathcal{N}\left(0, Var\left[U_n(\widehat{\theta}_n)\right]\right).$$

Then, using a first order Taylor expansion, we have

$$0 = U_n(\widehat{\theta}_n) = U_n(\theta) + \left(\frac{\partial U_n(\theta)}{\partial \theta} \right) (\widehat{\theta}_n - \theta) + o_p(1)$$

This gives

$$(\widehat{\theta}_n - \theta) =_d - \left(\frac{\partial U_n(\theta)}{\partial \theta} \right)^{-1} U_n(\theta).$$

Result 1: Suppose that $\widehat{\theta}_n$ is a solution to the estimating equation $U_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} U(\theta, Y_i) = 0$, i.e. $U_n(\widehat{\theta}_n) = 0$. Then $\widehat{\theta}_n \rightarrow_p \theta$ (consistency – see Crowder, 1994).

$$\sqrt{n}\,(\widehat{\theta}_n - \theta) \rightarrow_d \mathrm{N}_p(0, A^{-1}BA^{T-1}) \qquad (4)$$

(asymptotic normality) where

$$A = A(\theta) = -\mathbb{E}\left[\frac{\partial}{\partial\theta}U_n(\theta, Y)\right],$$

$$B = B(\theta) = \mathbb{E}[U_n(\theta, Y)U_n(\theta, Y)^T].$$

- The form of the variance in (4) has lead to it being called the sandwich estimator: $A^{-1}BA^{T-1}$.

Back to g-estimation