Causal Inference for time-dependent treatments: G-estimation of Structural Nested Models

Judith J Lok
Associate Professor of Mathematics and Statistics
jjlok@bu.edu

Catalyst, Harvard Medical School, 11/22/2019

Overview:

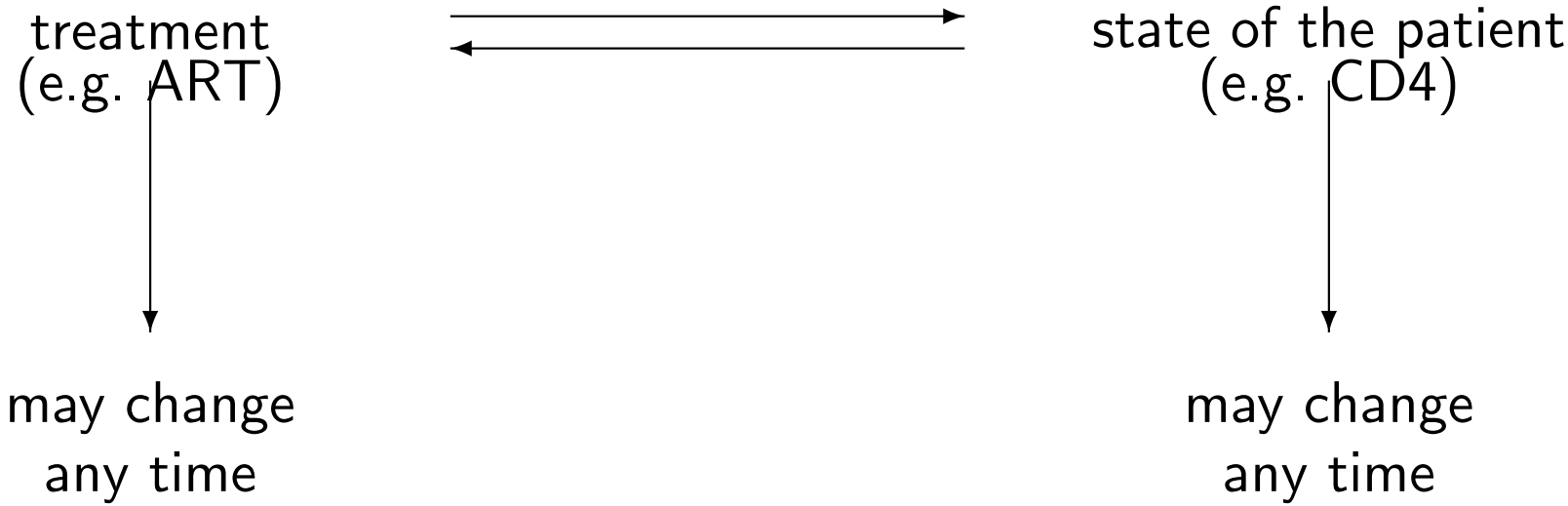
- Confounding by indication
- Bias of standard methods
- A few words on: Marginal Structural Models (MSMs)
- A few words on: G-computation Formula
- Structural Nested Mean Models (SNMMs)
- Inference for Structural Nested Mean Models
- Testing "treatment has no effect" without modeling treatment effect
- Illustration: coarse Structural Nested Mean Models with HIV application
- Marginal Structural Models versus Structural Nested Models
- Structural Nested Failure Time Models (SNFTMs, or maybe not)

The effect of time-dependent treatments

If treatment is repeatedly adapted to the state of the patient

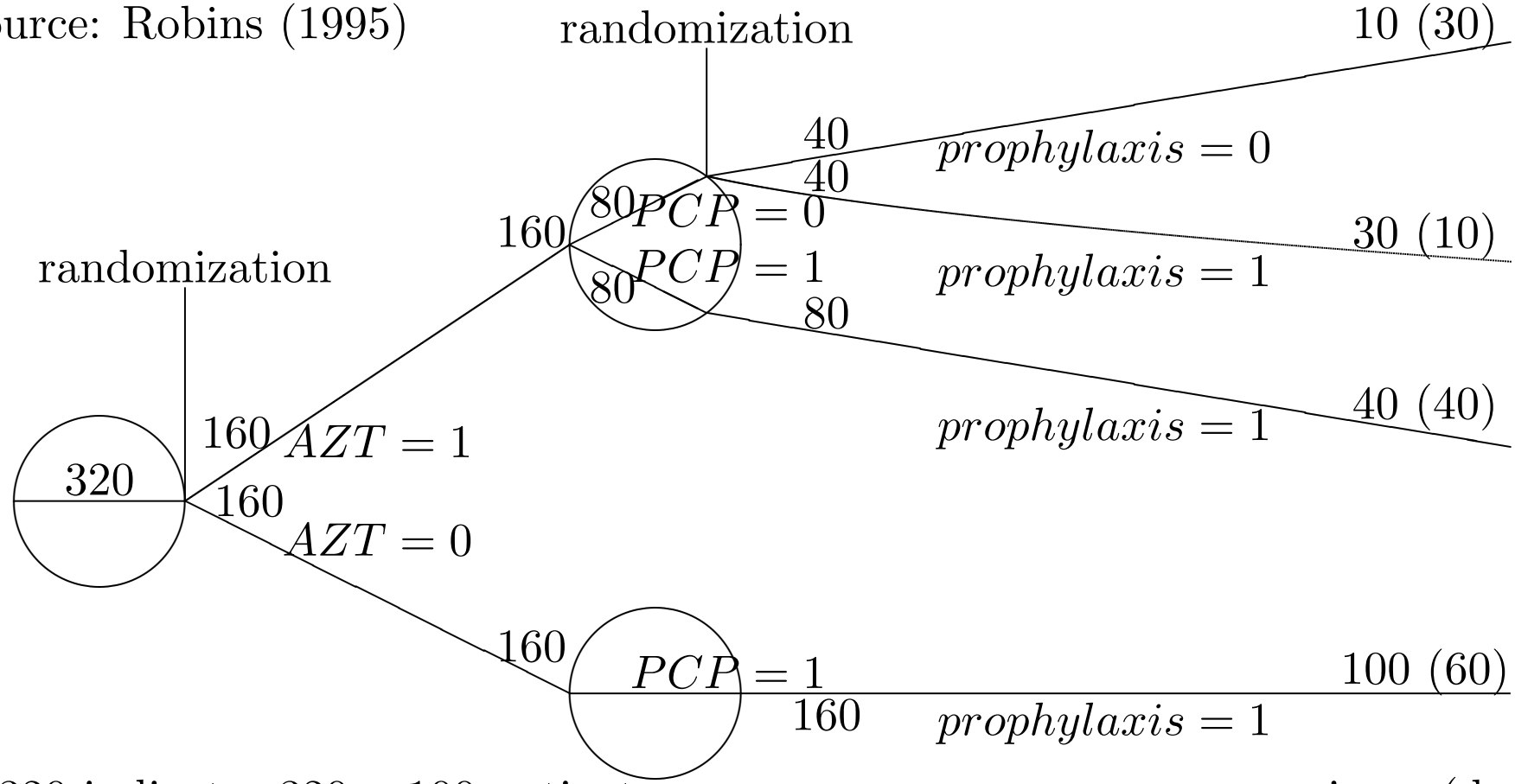
How can one estimate treatment effect?

Confounding by indication:



How to distinguish between treatment effect and the effect of the state of the patient (selection bias)?

source: Robins (1995)



320 indicates 320×100 patients

survivors (deaths)

Compare “AZT” versus “no AZT”

No adjustments:

Compare “AZT” versus “no AZT”

No adjustments:

- AZT:
  survived 80 (died 80)
- No AZT:
  survived 100 (died 60)

$\Rightarrow$ No AZT better.

Compare "AZT" versus "no AZT"

No adjustments:

- AZT:
  survived 80 (died 80)
- No AZT:
  survived 100 (died 60)

$\Rightarrow$ No AZT better.

BUT: in AZT group less prophylaxis use. And: prophylaxis helps (how can we see?). Unfair comparison?

Compare “AZT” versus “no AZT”

Prophylaxis helps, so let’s condition on prophylaxis=1:

Compare "AZT" versus "no AZT"

Prophylaxis helps, so let's condition on prophylaxis=1:

- AZT and prophylaxis:
  survived 70 (died 50)
- No AZT and prophylaxis:
  survived 100 (died 60)

$\Rightarrow$ No AZT better.

Compare "AZT" versus "no AZT"

Prophylaxis helps, so let's condition on prophylaxis=1:

- AZT and prophylaxis:
  survived 70 (died 50)
- No AZT and prophylaxis:
  survived 100 (died 60)

$\Rightarrow$ No AZT better.

BUT: prophylaxis use is predicted by PCP, and PCP is an independent predictor of death (how can we see?). Unfair comparison?

Compare “AZT” versus “no AZT”

Prophylaxis use is predicted by PCP, so need to adjust for PCP?

In “no AZT”, everyone developed PCP. So, let’s condition on PCP=1:

Compare "AZT" versus "no AZT"

Prophylaxis use is predicted by PCP, so need to adjust for PCP?

In "no AZT", everyone developed PCP. So, let's condition on PCP=1:

- AZT given PCP=1:
  survived 40 (died 40)
- No AZT given PCP=1:
  survived 100 (died 60)

$\Rightarrow$ No AZT better.

Compare "AZT" versus "no AZT"

Prophylaxis use is predicted by PCP, so need to adjust for PCP?

In "no AZT", everyone developed PCP. So, let's condition on PCP=1:

- AZT given PCP=1:
  survived 40 (died 40)
- No AZT given PCP=1:
  survived 100 (died 60)

$\Rightarrow$ No AZT better.

BUT: in "no AZT" group, ignored the group of patients with PCP=0. They have a better prognosis (how can we see?). Unfair comparison?

Compare "AZT" versus "no AZT"

Not adjusting for PCP leads to an unfair comparison.

Adjusting for PCP leads to an unfair comparison.

What can we do?

Time-dependent confounding by indication: AZT predicts PCP predicts prophylaxis, and PCP is an independent predictor of death.

What would have happened had EVERYONE taken “no AZT” followed by “prophylaxis”?

What would have happened had EVERYONE taken "no AZT" followed by "prophylaxis"?

This is what everyone did under "no AZT":

100 (60).

What would have happened had EVERYONE taken "AZT" followed by "prophylaxis"?

What would have happened had EVERYONE taken "AZT" followed by "prophylaxis"?

For most patients on AZT, we know what would have happened.

For the 40 on AZT who did not take prophylaxis: they would have fared the same as the patients with PCP=0 and prophylaxis, because of randomiation: 30 (10).

So, total:
30 (10)
30 (10)
40 (40)

Total:
100 (60).

The same as "no AZT", then prophylaxis!

What would have happened had EVERYONE taken "no AZT" followed by "no prophylaxis"?

Any ideas?

What would have happened had EVERYONE taken "no AZT" followed by "no prophylaxis"?

Any ideas?

Maybe some bounds, but no estimates without modeling assumptions.

"Positivity violation".

What would have happened had EVERYONE taken "AZT" followed by "no prophylaxis"?

Any ideas?

What would have happened had EVERYONE taken "AZT" followed by "no prophylaxis"?

Any ideas?

What would have happened to the 80 with PCP=1 had they not taken prophylaxis?

Maybe some bounds, but no estimates without modeling assumptions.

"Positivity violation".

What would have happened for a personalized treatment strategy:

"AZT" followed by:

- "no prophylaxis" if PCP=0
- "prophylaxis" if PCP=1.

What would have happened for a personalized treatment strategy:

"AZT" followed by:

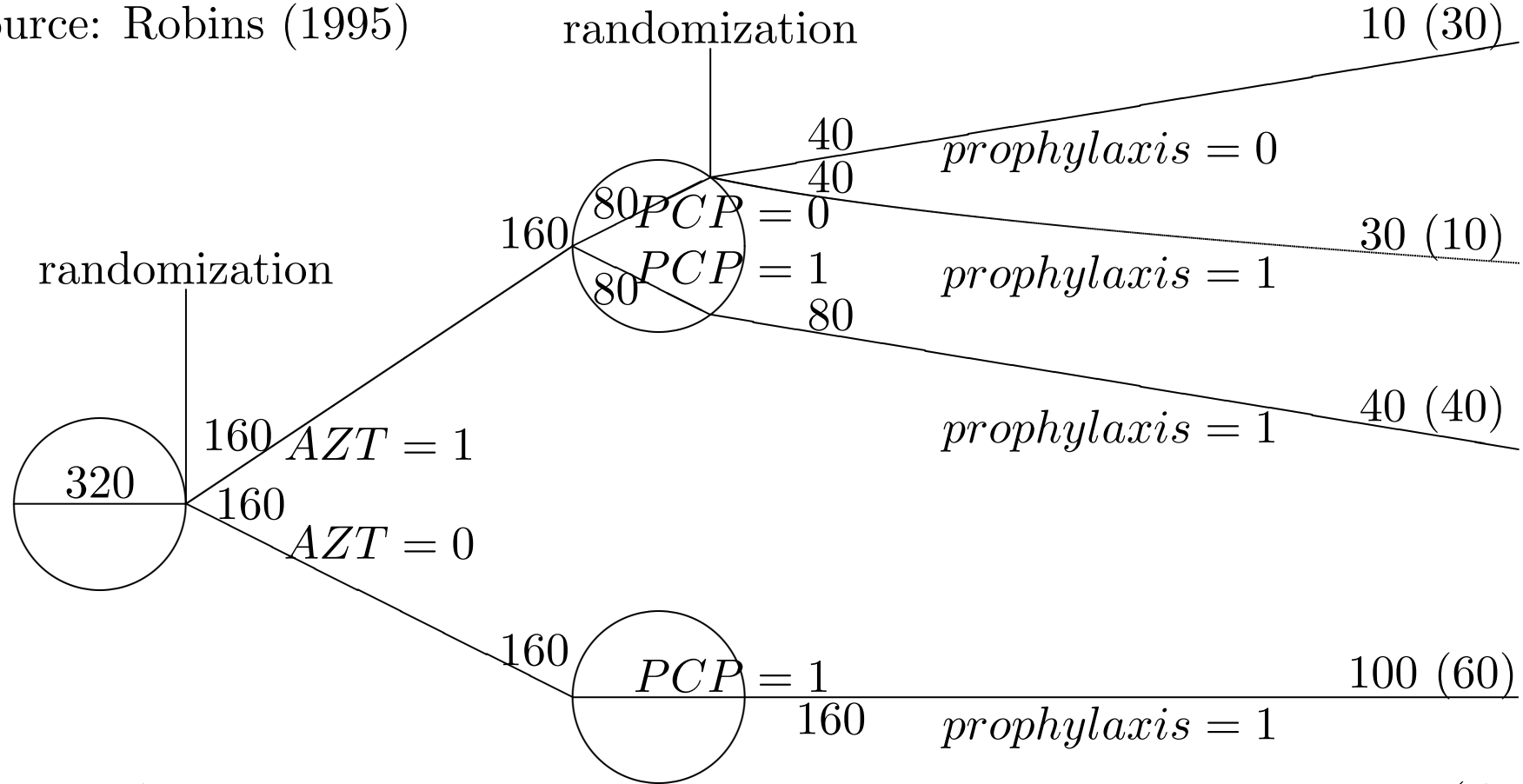
- "no prophylaxis" if PCP=0
- "prophylaxis" if PCP=1.

For most patients, we know what happened under this strategy. For the 40 who had PCP=0 and took prophylaxis: they would have fared the same as the 40 with PCP=0 who did not take prophylaxis: 10 (30).

So, total:
10 (30)
10 (30)
40 (40)

Total:
60 (100). Pretty bad. Expected.

source: Robins (1995)



320 indicates 320×100 patients

survivors (deaths)

What would have happened under other personalized treatment strategies?

Any thoughts on which personalized treatment strategies can be studied based on these data?

The Positivity Assumption

We noticed: if we are not making further assumptions, we cannot estimate the effect of treatment strategies unless they have been implemented for ALL types of patients.

E.g., AZT followed by "no prophylaxis" has not been implemented for anyone who developed PCP, so what would have happened to the 80.000 patients who were treated with AZT who developed PCP had they not been treated with prophylaxis??

We can only "guess" or "assume", informed by a comparison of patients who did not develop PCP with vs without prophylaxis.

The Positivity Assumption

Illustrates: Importance to talk with subject matter experts.

Illustrates: An "educated guess" would be extrapolation:

Conclusions would be based on data AND subject matter knowledge/(modeling) assumptions.

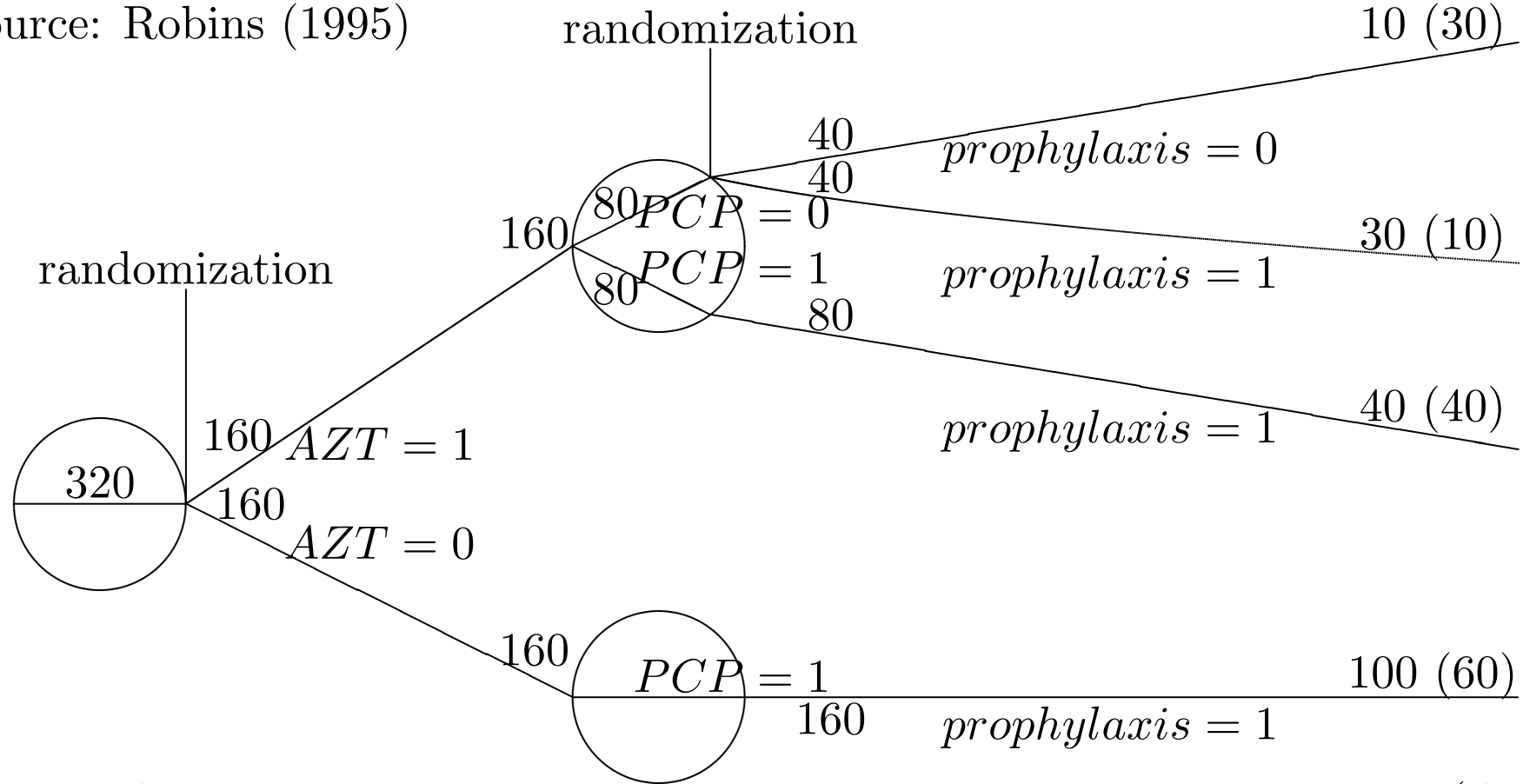
The time-dependent Cox-model

Problem with the **Cox model** here:

- Estimates: hazard of dying given the past.
- **BUT**: what would happen under a certain treatment regime??

Influence of PCP on prophylaxis and of prophylaxis on PCP later should be considered on top.

source: Robins (1995)



320 indicates 320×100 patients

survivors (deaths)

Inverse Probability of Treatment Weighting

What would have happened had EVERYONE taken "AZT" followed by "prophylaxis"?

In order to find out what had happened to the 40.000 patients who had AZT, no PCP, and no prophylaxis, had they taken prophylaxis, we copied the data from the 40.000 patients who had AZT, no PCP, and prophylaxis.

In other words, we gave the 40.000 patients who had AZT, no PCP, and prophylaxis a weight of 2.

That weight of 2 is the inverse of their probability to be treated with propylaxis, the treatment they actually took.

$\Rightarrow$ Inverse Probability of Treatment Weighting.

G-computation formula or multistate models

Model all transitions and combine. However:

- Computationally involved.
- Need to model each transition given the past.
- No specific parameter which indicates whether treatment affects the outcome of interest $\Rightarrow$ No standard test for treatment effect.

Recently used successfully in some applications by Jessica Young and Miguel Hernán et al, see e.g. Westreich et al. (2012) or Zhang et al. (2018).

Structural Nested Mean Models and Structural Nested Failure Time Models

Proposed by J.M. Robins
(Robins et al. (1992), Mark and Robins (1993), Robins (1998))
to estimate the effect of treatment on the final outcome.

Many of these estimators can be calculated with standard software,
by using this standard software in a non-standard way.

My contribution:

Understand and mathematically underpin this approach, especially when counterfactuals don't depend deterministically on the observed data (no "rank preservation"), and in continuous time.

And, proposed time-dependent coarse Structural Nested Mean Models, and underpin those, including optimal estimation, model fit, and sensitivity analyses.

Setting

Assumption: Covariates take values in discrete space.

Assumption (for now): Once treatment is initiated, it is not stopped; or, we study the effect of initiating treatment (intention to treat).

Assumption (for now): Treatment can be initiated at fixed times, 0, 1, ..., K. Covariates are measured at times 0, 1, ..., K.

Final outcome measured at time $K + 1$.

Setting and notation

- Time points $k \in \{0, 1, \ldots, K+1\}$: $K+1$ time the outcome is measured.
- Y: outcome of interest.
  Measured at time $K+1$, or survival time.
  Assume: continuously distributed.
- A_k: treatment in the interval $[k, k+1)$.
  $A_k = 1$ if treated, $A_k = 0$ if not.
- $\overline{A}_k = (A_0, \ldots, A_k)$.
- L_k: covariates at time k.
- $\overline{L}_k = (L_0, \ldots, L_k)$.
- Suppress subscript i for individual i.

Counterfactual outcomes

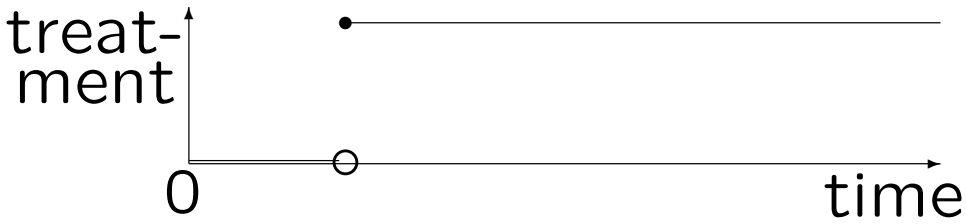
Write $Y^{(t)}$ for the outcome of the patient in case he/she would have been treated as in reality until time t and if treatment would have been stopped at t (for t in $[0, K+1]$).

Note: $Y^{(t)}$ only observed if treatment actually stopped before t.

Treatment
(0 indicates "no treatment")

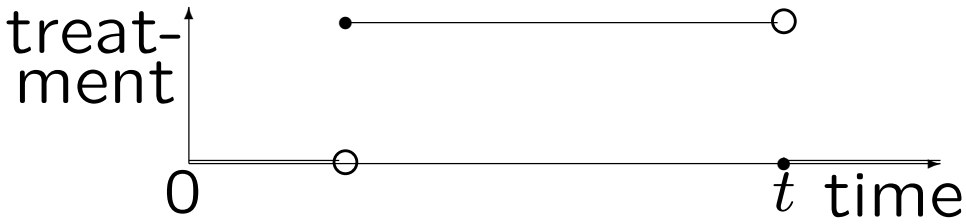
Outcome

Observed treatment



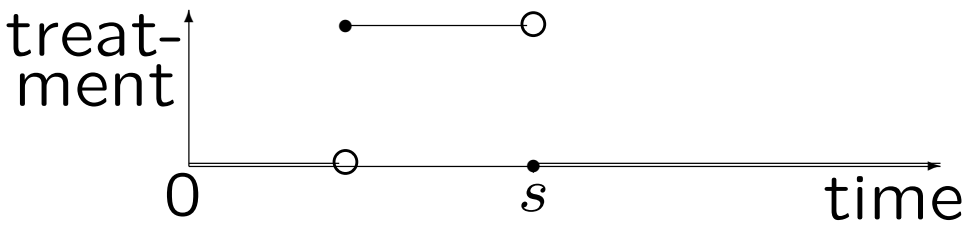
Observed
outcome (Y)

Treatment stopped at t



$Y^{(t)}$

Treatment stopped at s



$Y^{(s)}$

Structural Nested Models

Structural Nested Models model distributional relations between the different counterfactual outcomes.

No "rank preservation": We do not assume that different counterfactual outcomes are deterministically related.

Structural Nested Mean Models (SNMMs)

Estimate the mean difference between outcomes with treatment stopped a little later versus with treatment stopped now, given pre-treatment-stop patient characteristics.

Outcome often not survival.

Main assumption: Nu Unmeasured Confounding.

Notation

$(\overline{a}_k, \overline{0})$: treatment $(a_0, a_1, \ldots, a_k, 0, 0, \ldots, 0)$.

$Y^{(\overline{a}_k, \overline{0})}$: outcome under treatment $(\overline{a}_k, \overline{0})$.

Treatment effect, Structural Nested Mean Model (SNMM)

Definition (Treatment effect, SNMM).

$$\gamma_k\left(\bar{l}_k, \bar{a}_k\right) = E\left[Y^{(\bar{a}_k, \bar{0})} - Y^{(\bar{a}_{k-1}, \bar{0})} | \overline{L}_k = \bar{l}_k, \overline{A}_k = \bar{a}_k\right].$$

Compares the mean of $Y^{(\bar{a}_k, \bar{0})}$ and $Y^{(\bar{a}_{k-1}, \bar{0})}$, the outcome had treatment stopped at time $k+1$ versus at time k.

Obvious restriction: $\gamma_k = 0$ if $a_k = 0$.

Also called: blip function: for omitting the last blip of treatment.

Treatment effect, Structural Nested Mean Model (SNMM)

Definition (Treatment effect, SNMM).

$$\gamma_k\left(\bar{l}_k, \bar{a}_k\right) = E\left[Y^{(\bar{a}_k, \bar{0})} - Y^{(\bar{a}_{k-1}, \bar{0})} | \overline{L}_k = \bar{l}_k, \overline{A}_k = \bar{a}_k\right].$$

Let $\gamma_{k,\psi}$ be a **correctly specified model for** γ, with the true ψ being ψ^*.

Goal: Estimate ψ^*, the true ψ.

Main assumption for SNMM estimation: No Unmeasured Confounding

For identifiability: need information on all factors that both:

1. Influence treatment decisions and
2. Possibly predict an individual's prognosis with respect to the outcome of interest.

No Unmeasured Confounding (informal):

"Treatment decisions not based on more information about a patient's health prognosis than in database."

$Y^{(k)}$ indication of a patient's prognosis at time k.

No Unmeasured Confounding ("formal"):

A_k, the treatment decision at time k, is independent of $Y^{(k)}$ given *measured* data $(\overline{L}_k, \overline{A}_{k-1})$.

No Unmeasured Confounding (formal):

"Treatment decisions not based on more information about a patient's health prognosis than in database."

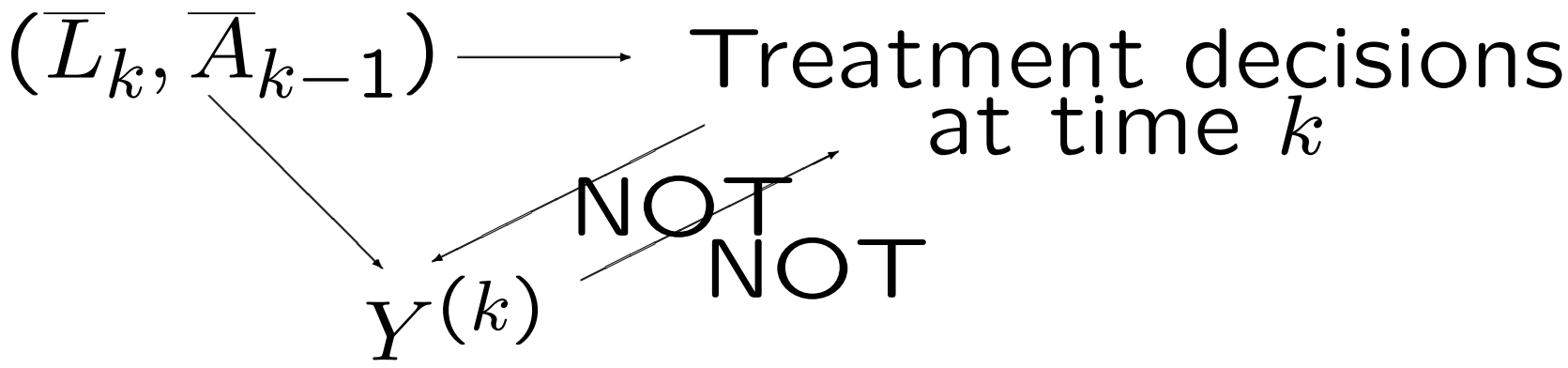
$Y^{(k)}$ indication of a patient's prognosis at time k.

No Unmeasured Confounding ("formal"):

$$A_k \perp\!\!\!\perp Y^{(k)} | \overline{L}_k = \overline{l}_k, \overline{A}_{k-1} = \overline{a}_{k-1}.$$

The notation $X \perp\!\!\!\perp Y | Z = z$, from Dawid (1979), means that the random variables X and Y are conditionally independent given the event $Z = z$.

No Unmeasured Confounding



Important remark: Y could not have been used instead of $Y^{(k)}$: Y depends on the future treatment decisions themselves!
$Y^{(k)}$ does not.

Note: $Y^{(k)}$ is usually not measured (assumption with counterfactuals).

Assumption: Consistency

Assumption: Consistency:

If $\overline{A}_K = \overline{a}_K$, $Y = Y^{(\overline{a}_K)}$.

Mimicking Counterfactual Outcomes

Definition. Define $H(K+1) = Y$, and for $k < K+1$ define recursively:

$$H(k) = H(k+1) - \gamma_k\left(\overline{L}_k, \overline{A}_k\right).$$

Also,

$$H_\psi(k) = H_\psi(k+1) - \gamma_{k,\psi}\left(\overline{L}_k, \overline{A}_k\right).$$

For the true ψ^*, $H_{\psi^*}(k) = H(k)$. $H(k)$ mimics a counterfactual outcome, with treatment "blipped off":

Mimickikng counterfactual outcomes

Theorem (Mimicking counterfactual outcomes.) Under Consistency,

$$E\left[H(k)|\overline{L}_k, \overline{A}_k = \overline{a}_k\right] = E\left[Y^{(\overline{a}_{k-1}, \overline{0})}|\overline{L}_k, \overline{A}_k = \overline{a}_k\right].$$

Corollary: Under No Unmeasured Confounding and Consistency, $E\left[H(k)|\overline{L}_k, \overline{A}_k = \overline{a}_k\right]$ doesn't depend on a_k.

Proof of Mimicking: Backwards induction, starting with $k = K$. See Appendix, or e.g. Lok and DeGruttola, Biometrics (2012).

G-estimation of SNMMs: Tool: Propensity score

Prediction of treatment given the past: Propensity score:

$$p(k) := P\left(A_k = 1 | \overline{A}_{k-1} = \overline{0}, \overline{L}_k\right),$$

for $k = 0, \ldots, K$.

In most cases, $p(k)$ is unknown and needs to be estimated.

$p_\theta(k)$: correctly specified model for $p(k)$, with $p(k) = p_{\theta^*}(k)$. Potentially easy to model. Flexible model: pooled logistic regression model: e.g., for $\overline{A}_{k-1} = \overline{0}$,

$$\text{logit} p_\theta(k) = \log(p_\theta(k)/(1 - p_\theta(k))) = \theta_0 + \theta_1 I_{\text{AZT}} + \theta_2 I_{\text{PCP}}(k),$$

or

$$p_\theta(k) = 1_{\overline{A}_{k-1}=\overline{0}} \frac{1}{1 + e^{-\theta_0 - \theta_1 I_{\text{AZT}} - \theta_2 I_{\text{PCP}}(k)}}.$$

Estimate with standard software for logistic regression.

G-estimation of SNMMs: strategy to estimate ψ^*

In the presence of confounding by indication, and assuming No Unmeasured Confounding, one uses a model to predict treatment initiation as a tool to estimate ψ.

Overview of G-estimation of ψ (Robins et al. (1992)):

- No Unmeasured Confounding: $Y^{(t)}$ does not add to the prediction model for treatment initiation at time t.
- Hope: $H_{\psi^*}(k)$, which mimics $Y^{(k)}$, does not add to the prediction model for treatment initiation at time k.
- Strategy: test this. $H_{\psi^*}(k)$ adds $\Rightarrow$ ψ^* not true parameter.

SNNMs: G-estimation of ψ^*

Proposed by Robins when he introduced SNMMs:

$\hat{\psi}$: the ψ so that H_ψ "predicts the least" to the prediction model for treatment initiation p_θ.

No Unmeasured Confounding implies that $H(k)$ does not add to the prediction model for treatment initiation p_θ, because $H(k)$ mimics $Y^{(\overline{a}_{k-1},\overline{0})}$, which does not add.

G-estimation of ψ: Example

Calculate the H_ψ on a grid of potential ψ's.

Add H_ψ to the model for treatment changes: Example:

$$\text{logit} p_{\theta,\alpha}(k) = \theta_0 + \theta_1 I_{\text{AZT}} + \theta_2 I_{\text{PCP}}(k) + \alpha H_\psi(k),$$

$$\text{or} \quad p_\theta(k) = 1_{\overline{A}_{k-1}=\overline{0}} \frac{1}{1 + e^{-\theta_0 - \theta_1 I_{\text{AZT}} - \theta_2 I_{\text{PCP}}(k) - \alpha H_\psi(k)}}.$$

Estimate $\hat{\theta}(\psi)$, $\hat{\alpha}(\psi)$ using standard software for logistic regression.

$\hat{\psi}$: the ψ that adds the least to the prediction model for treatment initiation: the ψ that generates $\hat{\alpha}(\psi) = 0$.

Or: $\hat{\alpha}(\psi) = 0 \Rightarrow \hat{\psi}$.

SNNMs: G-estimation of ψ using standard software

In general, finding ψ so that H_ψ "adds the least" to the prediction model for treatment initiation p_θ can be done as follows:

- Create a grid of potential ψ's.
- Add $\alpha H_\psi(k)$ to the treatment prediction model p_θ for each ψ separately.
- Estimate $\alpha(\psi)$ for each ψ separately, using standard software for logistic regression.
- Check: for which ψ does $\hat{\alpha}(\psi) = 0$?
- $\Rightarrow \hat{\psi}$.

G-estimation of ψ using standard software

$\hat{\alpha}(\psi) = 0 \Rightarrow \hat{\psi}$.

Proposed in e.g. Robins et al. (1992).

Unbiased Estimating Equations: Some Theory

Maximum Likelihood Estimation (MLE): often: solve the score equations. Proofs of consistency and asymptotic normality of MLEs often rely on the fact that, under regularity conditions, the score equations have expectation 0.

In general, there exists an extensive literature on "estimating equations": equations that one sets to 0 and solves for ψ. They usually rely on that those estimating equations have expectation 0: they are "**Unbiased Estimating Equations**". Overview: Van der Vaart (1998).

Under regularity conditions, estimators solving unbiased estimating equations are **consistent and asymptotically normal**.

Consistency and Asymptotic Normality of $\hat{\psi}$

It can be shown: $(\hat{\theta}, \hat{\psi})$ jointly solve Unbiased Estimating Equations.

Proof: see e.g. Lok and DeGruttola, Biometrics (2012).

Then, under regularity conditions, $\hat{\psi}$ is consistent and asymptotically normal:

$$\hat{\psi} \to^{P} \psi^*,$$

$$\sqrt{n}(\hat{\psi} - \psi^*) \to^{\mathcal{D}} \mathcal{N}(0, \Sigma),$$

for some covariance matrix Σ.

Testing "treatment has no effect" without modeling treatment effect

If there is no treatment effect, $\gamma_k \equiv 0$.

Then, $H(k) = Y$ for all k.

If there is no treatment effect, $H(k) = Y$, so $Y = H(k)$ does not add to the logistic regression model for treatment initiation.

$\Rightarrow$ **Can test whether Y adds to the logistic regression model for treatment initiation. If rejects: treatment affects the outcome.**

Testing "treatment has no effect" without modeling treatment effect

$\Rightarrow$ Can test whether Y adds to the logistic regression model for treatment initiation. If rejects: treatment affects the outcome.

Test can be carried out using standard software for logistic regression: just add Y to the model.

Proposed in e.g. Robins (1998).
Proof: e.g. Lok, Scandinavian Journal of Statistics (2007).

Testing "treatment has no effect" without modeling treatment effect: Example

If no treatment effect then adding Y to the model for treatment initiation should not help.

Model:

$$\text{logit} p_{\theta,\alpha}(k) = \theta_0 + \theta_1 I_{\text{AZT}} + \theta_2 I_{\text{PCP}}(k) + \alpha Y.$$

- Can estimate $(\theta_0, \theta_1, \theta_2, \alpha)$.
- If no treatment effect: true α equals 0.
- Test for treatment effect: test whether $\alpha = 0$, using standard software for logistic regression.
- Reject $\alpha = 0 \Rightarrow$ Treatment affects the outcome.

Test for other values of ψ

Similar to testing $\gamma = 0$:

If $\psi = \psi^*$ then adding $H_{\psi^*}(t)$ to the model for prediction of treatment in initiation should not help.

Standard software:

- Calculate the X_{ψ^*} for the different patients and time points.
- Test whether X_{ψ^*} adds to the prediction model for treatment changes.
- If it adds, reject ψ^*.

Confidence intervals for Structural Nested Mean Models

Can use duality between testing and confidence intervals: confidence interval contains all values ψ^* for which $H_0 : \psi = \psi^*$ is not rejected.

Resulting confidence regions for ψ turn out to be often "asymptotically equivalent" with $\hat{\psi} \pm 1.96\hat{se}$.

((where we defined asymptotic equivalence of (closed) confidence regions B_n and C_n as in Lok (2001).))

An alternative way to use standard software for Structural Nested Mean Models.

If γ is linear in ψ, this approach leads to a **linear restriction** on ψ once θ^* has been estimated $\Rightarrow$ closed form expression for $\hat{\psi}$.

$\Rightarrow$ Can use standard software to solve these equations for $\hat{\psi}$: e.g. PROC IML in SAS.

Then use bootstrap for confidence intervals for ψ_0.

See Lok and DeGruttola, Biometrics (2012).

Treatment regimes

Treatment regimes are usually denoted by g.

Notation: Let $\mathcal{L}_k$ be the space where L_k takes its values, and $\mathcal{A}_k$ be the space where A_k takes its values.

Definition: Treatment regimes. A treatment regime g is a vector $g = (g_0, \ldots, g_K)$ of functions $g_k : \overline{\mathcal{L}}_k \to \mathcal{A}_k$.

Notice: no need to include functions of $\overline{a}_{k-1}$: $\overline{a}_{k-1}$ also follows from g under the treatment regime.

Notation: Y^g is the outcome under treatment regime g.

Test for treatment effect based on SNMMs

We will restrict ourselves to baseline treatment regimes $\overline{0}$ that are **admissible**: after any treatment- and covariate history, there is a positive probability that the treatment is 0 in the next interval.

We will also restrict analysis to treatment regimes g that are **evaluable**: after any covariate history, if there is a positive probability that g was followed so-far, there is a positive probability that g will be followed in the next step.

No Unmeasured Confounding

Assumption: No Unmeasured Confounding, a little more general:

$$A_k \perp\!\!\!\perp Y^g | \overline{L}_k = \bar{l}_k, \overline{A}_{k-1} = g(\bar{l}_{k-1}).$$

Y^g: a patient's prognosis. No Unmeasured Confounding: Treatment decisions at time k based on previous observed history, and not further on a patient's prognosis.

Theorem: Under No Unmeasured Confounding, the mean outcome of interest is the same under any evaluable treatment regime g if and only if $\gamma_k \equiv 0$ for all k.

See Robins (2000) and Lok et al. (2004).

Doubly Robust Estimation of Structural Nested Mean Models

Instead of adding H_ψ to the model to predict treatment changes p_θ, add $H_\psi(k) - E\left[H_\psi(k)|\overline{L}_k, \overline{A}_{k-1} = \overline{0}\right]$ to the model to predict treatment changes p_θ.

Leads to unbiased estimating equations for ψ^* if either p_θ is correctly specified or $E\left[H_\psi(k)|\overline{L}_k, \overline{A}_{k-1} = \overline{0}\right]$ is correctly specified.

Proof of "**Double Robustness**" property: see Appendix.

An alternative way to use standard software for Structural Nested Mean Models.

If γ is linear in ψ, this approach leads to a **linear restriction** on ψ once θ^* (and maybe the censoring model) have been estimated $\Rightarrow$ closed form expression for $\hat{\psi}$.

$\Rightarrow$ Can use standard software to solve these equations for $\hat{\psi}$: e.g. PROC IML in SAS.

See Lok and DeGruttola, Biometrics (2012).

Conclusions SNMMs:

If there is No Unmeasured Confounding:

- SNMMs often lead to consistent, asymptotically normal estimators.
- Right censoring can be incorporated using IPCW (for MAR censoring, see later in the slide deck).
- SNMMs can often be estimated using standard software.

Illustration: Estimation of coarse Structural Nested Mean Models with application to initiating ART in HIV-infected patients

Literature:

- Lok and DeGruttola, Biometrics (2012).
- Lok, submitted, 2019.
- Yang and Lok, Biometrika (2016).
- Yang and Lok, Statistica Sinica (2018).
- Yang, Little, Smith, DeGruttola, and Lok, CROI (2016) (large HIV conference).

Motivation

Antiretroviral Therapy (ART) is recommended treatment for HIV-infected patients.

Evidence for best moment to start ART in early infection is not overwhelming.

Motivation

Benefits of postponing ART in HIV-infected patients:

- Postpones side effects of ART.
- Postpones the time patients develop drug resistance $\Rightarrow$ better long–term prognosis?
- Patients may get tired of taking drugs if they have to take them for a very long time.

Benefits of starting ART early:

- May prevent irreversible immune system damage.
- Less likely to transmit HIV.

Motivation for the theoretical development of time-dependent coarse Structural Nested Mean Models

Goal: How does the timing of ART initiation influence the effect of one year of treatment?

We study the effect of initiating ART, regardless of whether ART was discontinued within the year (intention to treat).

Motivation for the theoretical development of time-dependent coarse Structural Nested Mean Models

Goal: How does the timing of ART initiation influence the effect of one year of treatment?

Clinical trial data not available for this question.

$\Rightarrow$ Use observational data.

$\Rightarrow$ Confounding by indication.

Confounding by indication

Especially in resource limited settings, and historically also in the US, initiation of ART is postponed until some time after diagnosis of HIV. In clinical practice, decision to start ART based on:

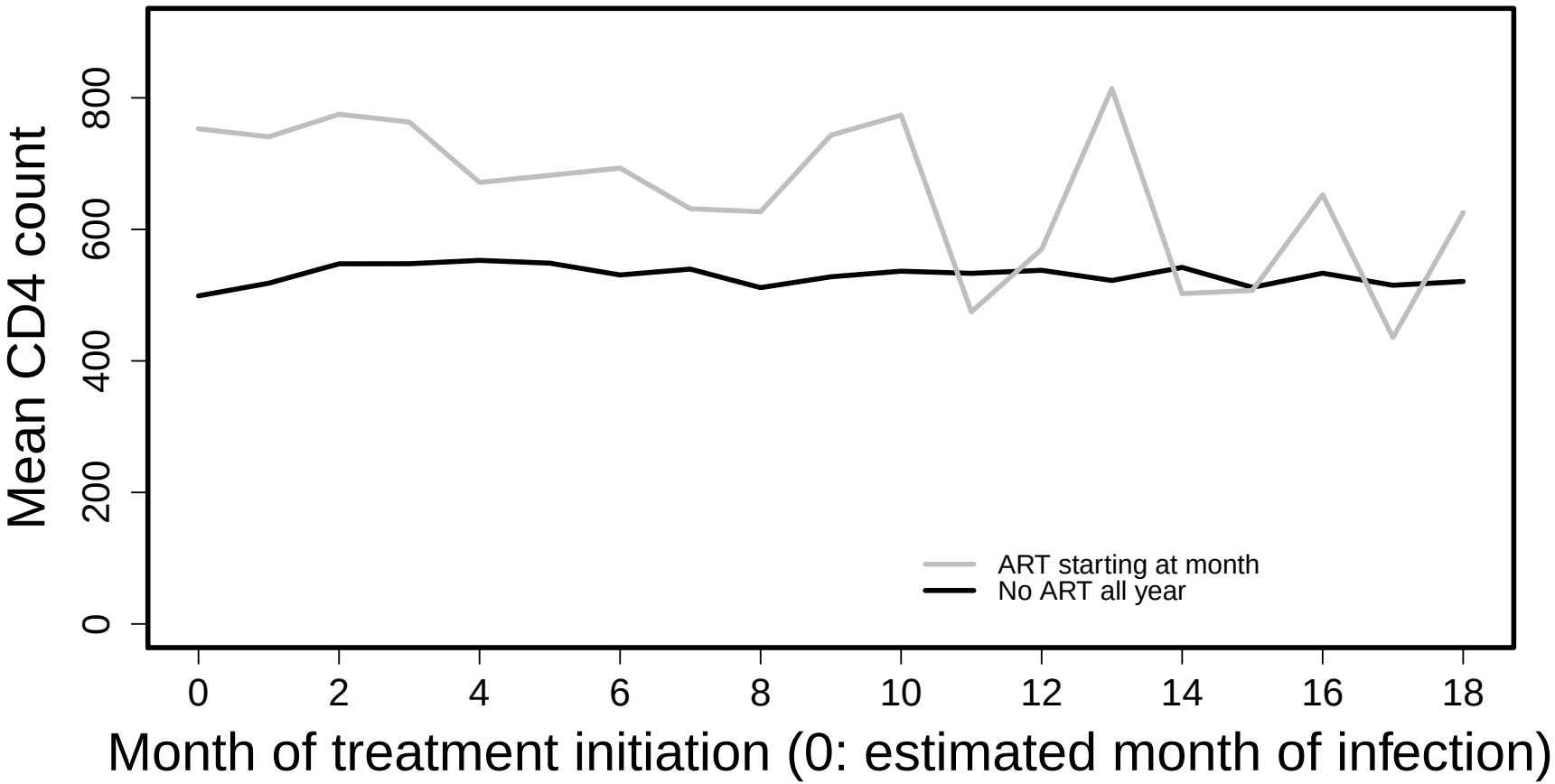
- CD4 count (often)
- viral load (sometimes)
- doctor's and patient's judgment (always)
- other factors?

Figure 1: ART in HIV-infected patients

Figure 1: ART in HIV-infected patients

- patients who start treatment from 0 to 18 months after infection: unadjusted mean values of CD4 count 1 year after treatment initiation
- for each such month: comparator group: patients who did not initiate treatment for at least one year: unadjusted mean values of CD4 count 1 year after

Figure 1: unadjusted mean CD4 count after 1 year, in subjects with visit at month, with and without ART



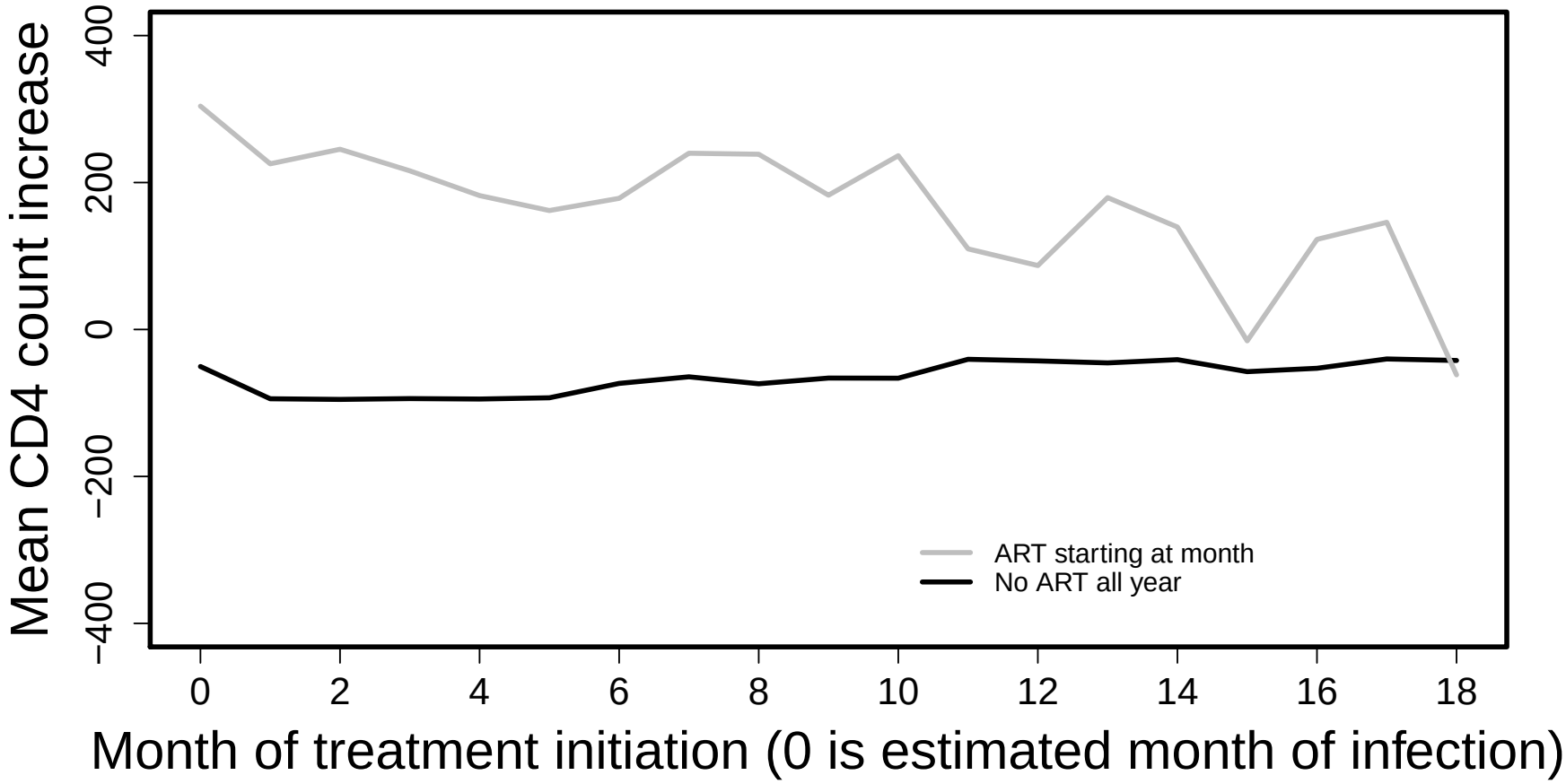
	N:																		
ART	96	216	205	91	46	37	23	26	14	8	6	7	12	6	5	3	5	7	6
No ART	46	83	278	285	249	254	230	179	173	130	152	97	118	78	106	68	78	75	63

Figure 1: ART in HIV-infected patients

There are months in which treated patients have even lower mean CD4 counts than those untreated.

Selection factor? Figure 2:

Figure 2: unadjusted mean CD4 increase after 1 year, in subjects with visit at month, with and without ART



Motivation for coarse Structural Nested Mean Models (SNMMs)

Confounding:
How to distinguish between treatment effect and the effect of the state of the patient (selection bias)?

Coarse Structural Nested Mean Models (coarse SNMMs)

$\Rightarrow$ **Coarse Structural Nested Mean Models (coarse SNMMs)**

Proposed for outcomes measured at the end of a study in Robins (1994).

Here: time-dependent outcomes.

Setting and notation: outcomes and covariates

All times measured in months since estimated date of infection: k is month k.

- Patients followed monthly, at month $0, 1, \ldots, K+1$, although visits may be missed.
- Y_k: outcome of interest at month k (e.g., CD4 count).
- $\overline{Y}_k$: outcome history until month k.

Setting and notation: treatment

- A_k: treatment at month k.
- Binary treatment, which is either given at month k ($A_k = 1$) or not ($A_k = 0$).
- Interest in effect of initiating treatment $\Rightarrow$ once $A_k = 1$, it stays 1 after k.
- T: month of treatment start (with $T = K + 1$ if treatment never started).
- Assumption: treatment can only start at non-missed visit months.
- L_k, $\overline{L}_k$, $\overline{A}_k$: as before.

Setting and notation: counterfactuals

- $\overline{Y}_k^{(\infty)}$: (counterfactual, not always measured) outcomes until month k had individual not been treated.
- $\overline{L}_k^{(\infty)}$: (counterfactual) covariates until month k had individual not been treated.
- $\overline{Y}_k^{(m)}$: (counterfactual) outcomes until month k had treatment started at month m.
- $\overline{L}_k^{(m)}$: (counterfactual) covariates until month k had treatment started at month m.

Baseline treatment $\overline{0}$

Notice: here $Y^{(m)}$ is with treatment *started* at month m.

We are considering here switching to "baseline treatment $\overline{0}$" where baseline treatment is "treat continuously".

Considering $Y^{(m)}$ avoids dealing with treatment regimes such as $(0, 1, 1, 0, 0)$, which according to our assumptions are never observed in clinical practice.

Model for treatment effect: coarse SNMMs

Model for treatment effect allows a different outcome, e.g. the CD4 count after one year, for each treatment initiation time:

Definition (treatment effect, time-dependent coarse Structural Nested Mean Model). For $k = m, \ldots, m + 12$,

$$\gamma_m^k\left(\bar{l}_m\right) = E\left[Y_k^{(m)} - Y_k^{(\infty)} | \bar{L}_m = \bar{l}_m, T = m\right].$$

Compares the mean of $Y_k^{(m)}$ and $Y_k^{(\infty)}$, the outcome at month k had treatment started at month m versus never starting treatment.

$\gamma > 0$: treatment increases the outcome; if outcome CD4: treatment is beneficial.

Effect of 1 year of ART

Definition (treatment effect, time-dependent coarse Structural Nested Mean Model). For $k = m, \ldots, m + 12$,

$$\gamma_m^k\left(\bar{l}_m\right) = E\left[Y_k^{(m)} - Y_k^{(\infty)} | \bar{L}_m = \bar{l}_m, T = m\right].$$

Effect of 1 year of treatment initiated at month m, given past covariate history $\bar{l}_m$, in individual initiating treatment at month m:

$$\gamma_m^{m+12}\left(\bar{l}_m\right).$$

Model specification: coarse SNMM

Definition (treatment effect). For $k = m, \ldots, m + 12$,

$$\gamma_m^k\left(\bar{l}_m\right) = E\left[Y_k^{(m)} - Y_k^{(\infty)} | \bar{L}_m = \bar{l}_m, T = m\right].$$

Example:

$$\gamma_{m,\psi}^k\left(\bar{l}_m\right) = \left(\psi_1 + \psi_2 m + \psi_3 m^2\right)(k - m),$$

$(k - m)$: duration of treatment from month m to month k. Effect of treatment linear in its duration. Coefficient depends on month of treatment start.

Model specification: coarse SNMM

Definition (treatment effect). For $k = m, \ldots, m + 12$,

$$\gamma_m^k\left(\bar{l}_m\right) = E\left[Y_k^{(m)} - Y_k^{(\infty)} | \overline{L}_m = \bar{l}_m, T = m\right].$$

Example:

Non-linear dependence on duration of treatment: could add terms like $\psi_4(k-m)^2$. Effect of treatment depends on its duration in linear and quadratic way.

Model specification: coarse SNMM

Definition (treatment effect). For $k = m, \ldots, m + 12$,

$$\gamma_m^k\left(\bar{l}_m\right) = E\left[Y_k^{(m)} - Y_k^{(\infty)} | \bar{L}_m = \bar{l}_m, T = m\right].$$

Examples:

Can extend the model by including terms like $\psi_4 lvl_m(k - m)$: effect of treatment is linear in its duration, with coefficient depending on month of treatment initiation as well as log viral load at treatment initiation, lvl_m.

Prediction of treatment given the past

Prediction of treatment given the past: tool to estimate treatment effect: propensity score:

$$p(m) := P\left(A_m = 1 | \overline{A}_{m-1} = \overline{0}, \overline{L}_m\right),$$

for $m = 0, \ldots, K$.

$p_\theta(m)$: correctly specified model for $p(m)$, with $p(m) = p_{\theta^*}(m)$. Potentially easy to model. Flexible model: pooled logistic regression model

$$p_\theta(m) = 1_{A_{m-1}=\overline{0}} 1_{visit}(m) \frac{1}{1 + e^{-\vec{\theta} \cdot \vec{f}(\overline{L}_m)}},$$

with $1_{visit}(m)$ an indicator of whether or not a visit took place at month m. We use maximum (partial) likelihood estimate, or "pooled logistic regression" (pooling over the times).

Assumptions

Assumption of No Unmeasured Confounding

For identifiability: need information on all factors that both:

1. Influence treatment decisions and
2. Possibly predict an individual's prognosis with respect to the outcome of interest.

Consistency Assumption

If an individual is not treated until month k, no difference in treatment between Y_k and $Y_k^{(\infty)}$. Assume: then $Y_k^{(\infty)} = Y_k$.

Mimicking Counterfactual Outcomes

Definition. For $12 \leq k \leq K+1$, on $T \geq k-12$, define

$$H(k) = Y_k - \gamma_T^k\left(\overline{L}_T\right).$$

$$H_\psi(k) = Y_k - \gamma_{T,\psi}^k\left(\overline{L}_T\right).$$

For the true ψ^*, $H_{\psi^*}(k) = H(k)$. $H(k)$ mimics a counterfactual outcome, with treatment "blipped off":

Theorem (Mimicking counterfactual outcomes.) Under Consistency, for $m \leq K$ and k with $12 \vee m \leq k \leq (m+12) \wedge (K+1)$,

$$E\left[H(k)|\overline{L}_m, \overline{A}_{m-1} = \overline{0}, A_m\right] = E\left[Y_k^{(\infty)}|\overline{L}_m, \overline{A}_{m-1} = \overline{0}, A_m\right].$$

Proof: backwards induction, starting with $m = K$, and differentiating between $A_m = 0$ and $A_m = 1$.

Unbiased estimating equations

Theorem (Unbiased Estimating Equations). Under No Unmeasured Confounding and Consistency, consider any $\vec{q}_m^{\,k}$ which are measurable and bounded. Then

$$E \sum_{m=0}^{K} \sum_{k=(m+1)\vee 12}^{(m+12)\wedge(K+1)} \vec{q}_m^{\,k}\left(\overline{L}_m\right) H\left(k\right) 1_{\overline{A}_{m-1}=\overline{0}} 1_{visit}(m)\left(A_m - p\left(m\right)\right) = 0,$$

with $1_{visit}(m)$ the indicator of whether a visit took place at month m. Thus if $p_\theta\left(m\right)$ and γ_ψ are correctly specified (parametric) models for $p\left(m\right)$ and γ, then

$$P_n \sum_{m=0}^{K} \sum_{k=(m+1)\vee 12}^{(m+12)\wedge(K+1)} \vec{q}_m^{\,k}\left(\overline{L}_m\right) H_\psi\left(k\right) 1_{\overline{A}_{m-1}=\overline{0}} 1_{visit}(m)\left(A_m - p_\theta\left(m\right)\right) = 0,$$

with P_n the empirical measure $P_n X = 1/n \sum_{i=1}^{n} X_i$, are unbiased estimation equations for both ψ^* and the (nuisance) parameter θ^*.

Unbiased estimating equations

Theorem (Unbiased Estimating Equations). If we restrict to k so that $k = m + 12$, the theorem becomes:

$$E \sum_{m=0}^{K-12} \vec{q}_m^{\,m+12}\left(\overline{L}_m\right) H\left(m+12\right) 1_{\overline{A}_{m-1}=\overline{0}} 1_{visit}(m)\left(A_m - p\left(m\right)\right) = 0,$$

with $1_{visit}(m)$ the indicator of whether a visit took place at month m. Thus if $p_\theta\left(m\right)$ and γ_ψ are correctly specified (parametric) models for $p\left(m\right)$ and γ, then

$$P_n \sum_{m=0}^{K-12} \vec{q}_m^{\,m+12}\left(\overline{L}_m\right) H_\psi\left(m+12\right) 1_{\overline{A}_{m-1}=\overline{0}} 1_{visit}(m)\left(A_m - p_\theta\left(m\right)\right) = 0,$$

with P_n the empirical measure $P_n X = 1/n \sum_{i=1}^n X_i$, are unbiased estimation equations for both ψ^* and the (nuisance) parameter θ^*.

Unbiased estimating equations

Stack the above estimating equations with partial likelihood estimating equations for θ^*.

Jointly: Unbiased Estimating Equations!

$\Rightarrow$ Under regularity conditions, $\hat{\psi}$ is consistent and asymptotically normal.

Solving the estimating equations

Corollary. If γ is linear in ψ, this approach leads to a **linear restriction** on ψ once θ has been estimated $\Rightarrow$ closed form expression for $\hat{\psi}$.

Can use standard software to solve these equations for $\hat{\psi}$.

Or, again, a grid search, searching for ψ so that $H_\psi(k)$, or $\vec{q}_m^{\,k}\left(\overline{L}_m\right) H_\psi(k)$, doesn't add to the treatment prediction model p_θ.

Censoring

Right Censoring: Assume Missing At Random and Positivity, and apply Inverse Probability of Censoring Weighting (IPCW).

Choice of estimating equations

Estimators and confidence intervals depend very much on the choice of unbiased estimating equations.

Which ones to use?

For now: focus on situation without censoring.

Double Robustness and more efficient estimators

Theorem (Double Robustness and more efficient estimators). The estimating equations:

$$P_n \sum_{m=0}^{K} \sum_{k=(m+1)\vee 12}^{(m+12)\wedge(K+1)} \vec{q}_m^{\,k}\left(\overline{L}_m\right)\left(H_\psi(k) - E\left[H_\psi(k)|\overline{L}_m, \overline{A}_{m-1} = \overline{0}\right]\right)$$

$$1_{\overline{A}_{m-1}=\overline{0}}\left(A_m - p_{\hat{\theta}}(m)\right) = 0$$

are unbiased and more efficient than the ones before. Moreover, they are doubly robust: stacked with the estimating equations for θ^*, they are unbiased for ψ^* if either p_θ is correctly specified or $E\left[H_\psi(k)|\overline{L}_m, \overline{A}_{m-1} = \overline{0}\right]$ is correctly specified.

Note: for model specification, note that

$$E\left[H(k)|\overline{L}_m, \overline{A}_{m-1} = \overline{0}\right] = E\left[Y_k^{(\infty)}|\overline{L}_m, \overline{A}_{m-1} = \overline{0}\right].$$

Choice of estimating equations

Assumption (Homoscedasticity). For $0 \leq m \leq K$ and $(m+1) \vee 12 \leq k, s \leq (m+12) \wedge (K+1)$,

$$Cov\left[H(k), H(s) | \overline{L}_m, \overline{A}_{m-1} = \overline{0}, A_m\right]$$

does not depend on A_m.

Not necessary for estimating equations to be Doubly Robust.

True for both $H(k)$ and $H(s)$ separately, assumption about their product.

Possible to find optimal estimating equations under this assumption.

Optimal estimating equations

Suppose model for treatment effect is:

$$\begin{aligned}\gamma^{k}_{m,\psi}\left(\overline{L}_m\right) &= \left(\psi_1 + \psi_2 m + \psi_3 m^2 + \psi_4 k + \psi_5 k^2\right)(k-m) \\ &\quad + \left(\psi_6(k-m) + \psi_7 lvl_m\right)(k-m) + \psi_8 lvl_m.\end{aligned}$$

We find the optimal estimating equations as follows:

Optimal estimating equations

Theorem (optimal estimating equations)

With $Tr(m,k)$ the number of months of treatment between months m and k, define

$$\Delta_m(k) \quad = \quad E\left[\begin{pmatrix} Tr(m,k) \\ TTr(m,k) \\ T^2\,Tr(m,k) \\ kTr(m,k) \\ k^2\,Tr(m,k) \\ Tr^2(m,k) \\ lvl_T\,Tr(m,k) \\ lvl_T\,A_{k-1} \end{pmatrix} \mid A_m = 1, \overline{A}_{m-1} = \overline{0}, \overline{L}_m\right]$$

$-$ same with $A_m = 0$

Optimal estimating equations

so

$$\Delta_m(k) = \begin{pmatrix} k-m \\ m(k-m) \\ m^2(k-m) \\ k(k-m) \\ k^2(k-m) \\ (k-m)^2 \\ lvl_m(k-m) \\ lvl_m \end{pmatrix} - E\left[\begin{pmatrix} Tr(m,k) \\ TTr(m,k) \\ T^2\,Tr(m,k) \\ kTr(m,k) \\ k^2\,Tr(m,k) \\ Tr^2(m,k) \\ lvl_T\,Tr(m,k) \\ lvl_T\,A_{k-1} \end{pmatrix} \mid \overline{A}_m = \overline{0}, \overline{L}_m \right]$$

Optimal estimating equations

and, with $min = (m+1) \vee 12$, $max = m + 12 \wedge (K+1)$,

$$\vec{Cov}_m \left[HI_8 | \overline{L}_m, \overline{A}_{m-1} = \overline{0} \right]$$

$$= \begin{pmatrix} \Gamma^m_{min,min} I_8 & \Gamma^m_{min,min+1} I_8 & \cdot & \cdot & \Gamma^m_{min,max} I_8 \\ \Gamma^m_{min+1,min} I_8 & \cdot & \cdot & \cdot & \\ \cdot & \cdot & & & \\ \cdot & \cdot & & & \\ \Gamma^m_{max,min} I_8 & \cdot & \cdot & \cdot & \Gamma^m_{max,max} I_8 \end{pmatrix}$$

(I_8 the 8×8 identity matrix), and

$$\Gamma^m_{k,s} = \mathrm{Cov} \left[H(k), H(s) | \overline{L}_m, \overline{A}_{m-1} = \overline{0} \right].$$

Optimal estimating equations

Any $\vec{q}^{\,opt}$ satisfying the following equations is optimal: for $m = 0, \ldots, K$,

$$\begin{pmatrix} \Delta_m(min) \\ \Delta_m(min+1) \\ \cdot \\ \cdot \\ \Delta_m(max) \end{pmatrix} = \vec{Cov}_m\left[HI_8 | \overline{L}_m, \overline{A}_{m-1} = \overline{0}\right] \begin{pmatrix} \vec{q}_m^{\,opt,min}\left(\overline{L}_m\right) \\ \vec{q}_m^{\,opt,min+1}\left(\overline{L}_m\right) \\ \cdot \\ \cdot \\ \vec{q}_m^{\,opt,max}\left(\overline{L}_m\right) \end{pmatrix}.$$

Working identity covariance matrix

If instead of $\vec{Cov}_m \left[HI_8 | \overline{L}_m, \overline{A}_{m-1} = \overline{0} \right]$ we use the identity matrix, we get a result similar to what in Generalized Estimating Equations is called the result under an identity working correlation matrix.

Doubly Robust, not optimal.

Not really "estimating equations"

$\Rightarrow$ Doubly Robust Estimating Equations depend on ψ^* (through H).

$\Rightarrow$ Then, optimal choice of Estimating Equations depends on ψ^*.

$\Rightarrow$ Can start with preliminary estimator $\tilde{\psi}$ and plug that in $Cov\left[H(k), H(s)|\overline{L}_m, \overline{A}_{m-1} = \overline{0}\right]$ to obtain optimal Estimating Equations.

$\Rightarrow$ Plugging in $\tilde{\psi}$ does not affect unbiasedness of Estimating Equations, if $\tilde{\psi}$ results from Unbiased Estimating Equations.

$\Rightarrow$ Can be shown: Plugging in $\tilde{\psi}$ does not affect asymptotic variance of resulting estimators, if $\tilde{\psi}$ results from Unbiased Estimating Equations.

Simulations:

Truth: modeled based on the AIEDRP data: HIV-infected patients during acute and early infection, outcome: CD4 count.

$$\gamma^k_{m,\psi}\left(\overline{L}_m, \overline{A}_m\right) = (\psi_1 + \psi_2 m)(k - m),$$

$\psi_1 = 25$ and $\psi_2 = -0.7$.

We included two "naive" q's, not based on the theory:

$$\vec{q}_m^{\,m+12} = \left(\begin{array}{ccc} CD4_m & m & injdrug \end{array} \right)^\top \tag{1}$$

and

$$\vec{q}_m^{\,m+12} = \left(\begin{array}{ccc} CD4_m & injdrug & CD4_6 \end{array} \right)^\top. \tag{2}$$

1000 patients in 1000 datasets.

Simulations: Comparison of Mean Squared Errors for the various estimators: two parameters

	ψ_1	ψ_2
1a. q as in (1), not DR	12.6	0.084
1b. q as in (2), not DR	1.0×10^6	8.8×10^3
2. restricted approx. optimal[1], not DR	4.4	0.028
3. restricted approx. optimal[1], DR	4.4	0.028
4. working correlation matrices[2], DR	2.6	0.010
5. approximately optimal[3], DR	1.7	0.0062

[1] q approximately optimal within the class with $q_m^k = 0$ for $k \neq m + 12$.
[2] like the optimal estimator, but with working identity correlation matrices.
[3] optimal under correct specification of all models.

Simulations: Comparison of Mean Squared Errors for the various estimators: three parameters

	ψ_1	ψ_2	ψ_3
1a. q as in (1), not DR	1.3×10^7	6.1×10^5	1.4×10^3
1b. q as in (2), not DR	5.4×10^5	2.7×10^4	64
2. restricted approx. opt.[1], not DR	107	3.7	0.0071
3. restricted approx. opt.[1], DR	43.0	1.33	0.0024
4. working correlation matrices[2], DR	14.9	0.28	0.00029
5. approximately optimal[3], DR	9.9	0.17	0.00016

[1] q approximately optimal within the class with $q_m^k = 0$ for $k \neq m + 12$.
[2] like the optimal estimator, but with working identity correlation matrices.
[3] optimal under correct specification of all models.

The AIEDRP Core01 database with HIV-infected patients:

- 1762 patients.
- 1203 initiated treatment under follow-up, and 559 did not.
- Broad range of time from estimated date of infection to treatment initiation: interquartile spread 1.5 to 6 months.
- Viral Load range at time of treatment initiation. was very broad: interquartile spread 17,000 to 400,000 copies.
- CD4 count at time of treatment initiation: interquartile spread 340 to 630 cells/ml.

Effect of one year of ART following m months of infection

Effect of one year of ART following m months of infection on the mean one-year CD4 increase.

For the simplest model,

$$\gamma^k_{m,\psi}\left(\overline{L}_m, \overline{A}_m\right) = \left(\psi_1 + \psi_2 m\right)(k - m),$$

we chose as preliminary estimator $\vec{q}^{\,k}_m = 0$ if $k \neq m + 12$, and

$$\vec{q}_m^{\,m+12}\left(\overline{L}_m\right) = \begin{pmatrix} 12 - E\left[Tr(m, m+12) | A_m = 0, \overline{L}_m\right] \\ 12m - E\left[TTr(m, m+12) | A_m = 0, \overline{L}_m\right] \end{pmatrix},$$

($Tr(m, m+12)$: treatment between months m and $m+12$; estimated conditional expectations in prior regression step), $\Rightarrow$ 2 estimating equations for ψ after estimating θ. Any choice of q here is valid; this q for efficiency.

Effect of one year of ART following m months of infection

Effect of one year of ART following m months of infection on the mean one-year CD4 increase.

The analysis adjusted for confounding in the initiation of treatment and informative censoring.

We considered as predictors in the models for treatment initiation and censoring:

- demographic factors sex, injection drug use, age at infection.
- month, month=0, month=1 and month squared, first visit, second visit, days since last visit.
- time-varying measurements of CD4 count, CD4 count slope, log viral load, and (only for censoring) treated and firsttreated.

Results

Results (bootstrap confidence intervals: Efron's percentile method, based on 500 replicates):

AIEDRP data: Comparison of estimators and 95% CIs: two parameters

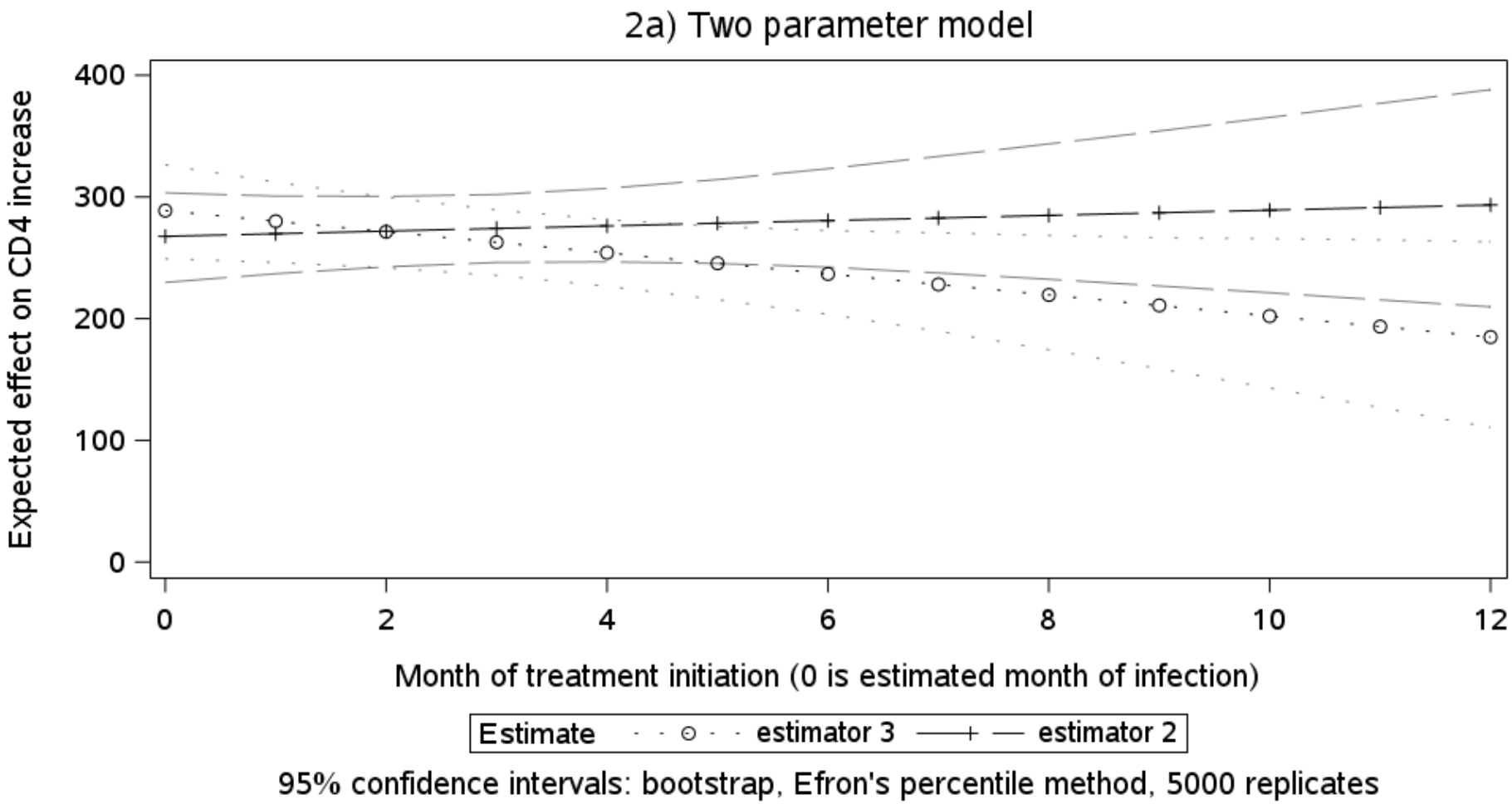
	ψ_1	ψ_2
1a. naive not DR	22.4 (18.9,25.8)	0.16 (-0.59,1.0)
1b. naive not DR	43 (-128,209)	-9 (-89,68)
2. restricted not DR[1]	22.3 (19.1,25.3)	0.18 (-0.53,1.00)
3. restricted DR[1]	24.1 (20.8,27.2)	-0.72 (-1.39,-0.02)
4. working corr DR[2]	25.3 (22.2,28.5)	-0.23 (-0.71,0.28)
5. optimal DR[3]	24.8 (20.2,29.0)	-0.44 (-2.1,1.3)

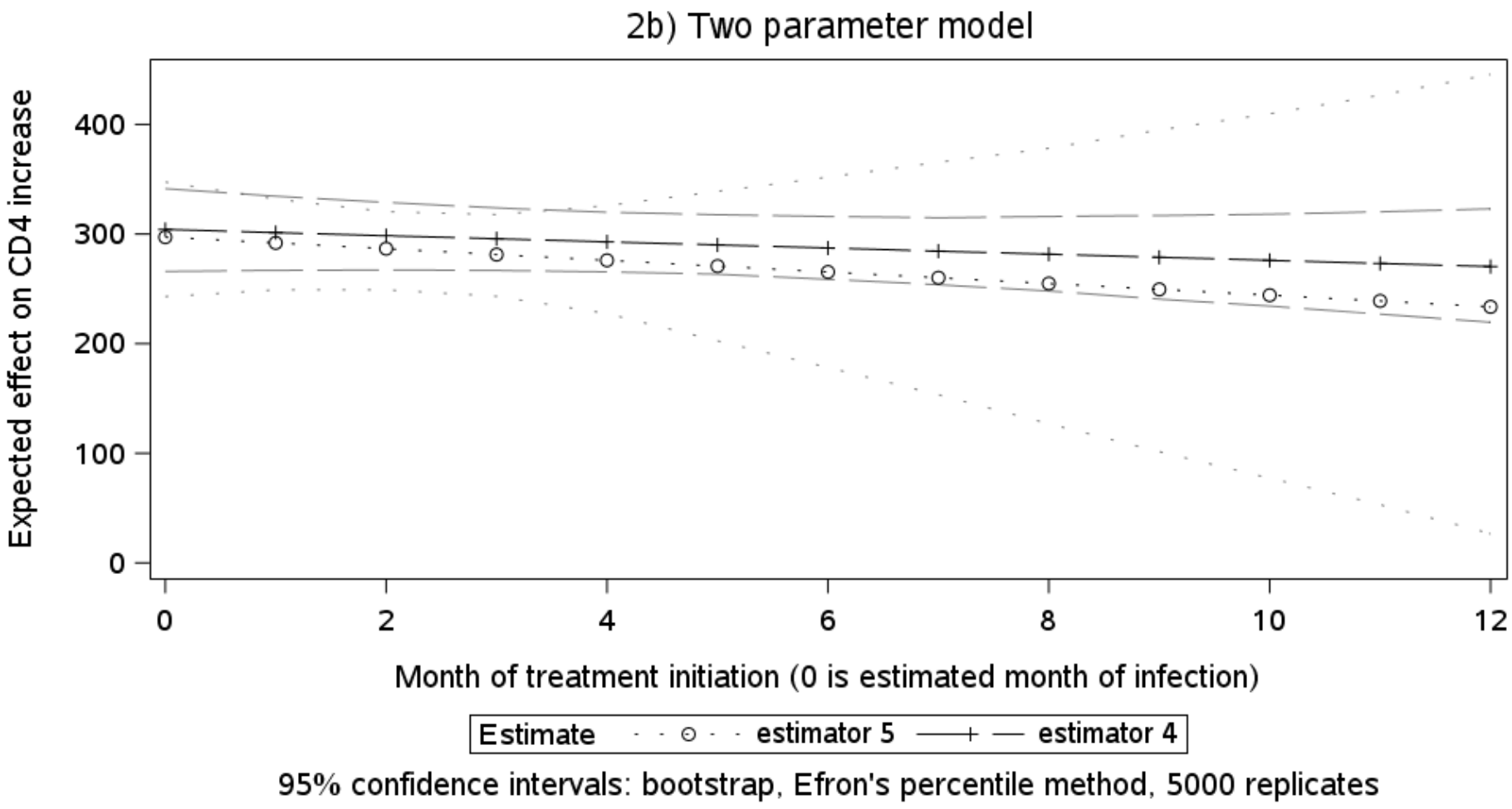
[1] q approximately optimal within the class with $q_m^k = 0$ for $k \neq m + 12$.
[2] like the optimal estimator, but with working identity correlation matrices.
[3] optimal under correct specification of all models.

AIEDRP data: Comparison of estimators and 95% CIs: three parameters

	ψ_1	ψ_2	ψ_3
1a.	39 (-27,175)	-11 (-99,33)	1.2 (-4.1,12)
1b.	38 (-126,208)	-11 (-132,107)	1.6 (-15,19)
2. [1]	19.5 (14.8,24.1)	2.0 (-0.3,4.4)	-0.20 (-0.43,0.02)
3. [1]	23.4 (18.5,28.0)	-0.2 (-2.6,2.3)	-0.06 (-0.32,0.19)
4. [2]	25.6 (21.7,29.5)	-0.35 (-1.8,1.2)	0.0095 (-0.09,0.11)
5. [3]	25.9 (19.0,31.8)	-0.87 (-3.8,2.7)	0.025 (-0.21,0.23)

[1] q approximately optimal within the class with $q_m^k = 0$ for $k \neq m + 12$.
[2] like the optimal estimator, but with working identity correlation matrices.
[3] optimal under correct specification of all models.


2a) Two parameter model
Expected effect on CD4 increase
400
300
200
100
0
0
2
4
6
8
10
12
Month of treatment initiation (0 is estimated month of infection)
Estimate
estimator 3
estimator 2
95% confidence intervals: bootstrap, Efron's percentile method, 5000 replicates


2b) Two parameter model
Expected effect on CD4 increase
400
300
200
100
0
0
2
4
6
8
10
12
Month of treatment initiation (0 is estimated month of infection)
Estimate
estimator 5
estimator 4
95% confidence intervals: bootstrap, Efron's percentile method, 5000 replicates

Conclusions coarse Structural Nested Mean Models:

- We developed time-dependent coarse SNMMs: J J Lok and V DeGruttola (2012). Impact of time to start treatment following infection with application to initiating HAART in HIV-infected patients. *Biometrics* 68: 745-754.
- We developed doubly robust estimators of coarse SNMMs.
- We studied which estimator of coarse SNMMs is optimal in the absence of censoring. The optimal estimating equations depend on the parameter(s) of the treatment effect model, which can be estimated in preliminary step.
- We programmed the estimators and CIs using the bootstrap.
- We applied coarse SNMMs to estimate the effect of one year of ART on CD4 increase, assuming No Unmeasured Confounding.

Additional work on coarse Structural Nested Mean Models

- S Yang and J J Lok (2016). A goodness-of-fit test for Structural Nested Mean Models. *Biometrika* 103(3): 734-741. Based on over-identification restrictions tests: there are more possible estimating equations than parameters.
- S Yang and J J Lok (2018). Sensitivity analysis for unmeasured confounding in coarse Structural Nested Mean Models. *Statistica Sinica* 28: 1703-1723.
  For outcome measured at the end of the study, need to specify plausible range for $E[Y^{(\infty)}|\overline{L}_m, \overline{A}_{m-1} = 0, A_m = 1]$ $-E[Y^{(\infty)}|\overline{L}_m, \overline{A}_{m-1} = 0, A_m = 0]$ (selection bias function).
- We presented our results to clinicians at the Conference on Retroviruses and Opportunistic Infections (CROI), 2016.

Right Censoring

In many observational studies, there is right censoring: after each visit time, some patients are no longer in follow-up.

If censoring is independent of the covariates and the treatments, can just restrict to the persons with complete data.

If censoring depends on (time-varying) covariates and treatments: Inverse Probability of Censoring Weighting (IPCW) (Robins et al. (1995)).

Rational for IPCW: if many patients with specific characteristics drop out, give more weight to similar patients in follow-up.

Inverse Probability of Censoring Weighting (IPCW)

Write $Q\left(Y, \overline{L}_K, \overline{A}_K, \theta, \psi\right)$ for the estimating equation based on the full data for 1 patient. Write $C_k = 1$ if a patient was censored at k, and $C_k = 0$ if not. Then:

$$E\left(\frac{1_{C_{K+1}=0}}{P(C_{K+1}=0|Y, \overline{L}_K, \overline{A}_K)} Q(Y, \overline{L}_K, \overline{A}_K, \theta^*, \psi^*)\right) = 0.$$

Follows by conditioning on $Y, \overline{L}_K, \overline{A}_K$. Need:

Assumption: Positivity. $P\left(C_{K+1}=0|Y, \overline{L}_K, \overline{A}_K\right) > 0$ for all possible values of $(Y, \overline{L}_K, \overline{A}_K)$.

Positivity: No matter what a patient's full data, there is a positive probability of observing his/her full data.

Inverse Probability of Censoring Weighting (IPCW)

$$E\left(\frac{1_{C_{K+1}=0}}{P(C_{K+1}=0|Y,\overline{L}_K,\overline{A}_K)}Q\left(Y,\overline{L}_K,\overline{A}_K,\theta^*,\psi^*\right)\right)=0.$$

Re-write

$$P(C_{K+1}=0|Y,\overline{L}_K,\overline{A}_K)=\prod_{k=1}^{K+1}P(C_k=0|Y,\overline{L}_K,\overline{A}_K,C_{k-1}=0).$$

Proposed in Robins et al. (1995), see also Lok and Hughes, Statistics In Medicine (2016) .

Inverse Probability of Censoring Weighting (IPCW)

Assumption: (Missing At Random (MAR)). See e.g. Robins et al. (1995) or Rubin (1976).

$$C_k \perp\!\!\!\perp (Y, \overline{L}_K, \overline{A}_K) | \overline{L}_{k-1}, \overline{A}_{k-1}, C_{k-1} = 0.$$

Again, $\perp\!\!\!\perp$ indicates conditional independence Dawid (1979).

MAR: Censoring may depend on past observed information, but not further on prognosis. Prognosis: $(Y, \overline{L}_K, \overline{A}_K)$.

IPCW for Structural Nested Mean Models

Re-write under Missing At Random (MAR):

$$\begin{aligned} P(C_{K+1}=0|Y,\overline{L}_K,\overline{A}_K) &= \prod_{k=1}^{K+1} P(C_k=0|Y,\overline{L}_K,\overline{A}_K,C_{k-1}=0) \\ &= \prod_{k=1}^{K+1} P(C_k=0|\overline{L}_{k-1},\overline{A}_{k-1},C_{k-1}=0). \end{aligned}$$

Result: IPCW-weighted estimating equations for Structural Nested Mean Models:

$$P_n\left(\frac{1_{C_{K+1}=0}}{\prod_{k=1}^{K+1} P(C_k=0|\overline{L}_{k-1},\overline{A}_{k-1},C_{k-1}=0)}\, Q\left(Y,\overline{L}_K,\overline{A}_K,\theta,\psi\right)\right)=0.$$

$\Rightarrow$ Unbiased Estimating Equations, when stacked with estimating equations for a model to predict censoring given the past.

IPCW for Structural Nested Mean Models

Conclusion: Under regularity conditions, the IPCW Structural Nested Mean Models estimators are consistent and asymptotically normal.

Note: Estimators (and tests, confidence intervals) can be calculated by using standard software for *weighted* logistic regression.

Structural Nested Models or Marginal Structural Models?

Elaborate comparison: Robins (2000).

Marginal Structural Models (MSMs) are easier to interpret.

Structural Nested Models (SNMs) can estimate how treatment effects depend on pre-treatment patient characteristics.

MSMs, in contrast to SNMs, cannot be used if there exists a value of a possible covariate- and treatment history so that after that, only one treatment decision is possible. Example: a study of the effect of an occupational exposure on mortality if "on versus off work" needs to be in the l_k to make no unmeasured confounding plausible, and subjects off work can only receive exposure level $a_k = 0$.

Structural Nested Failure Time Models

Estimate the effect of treatment on a time-to-event outcome.

Same notation and setting as for SNMMs, different model for treatment effect.

Structural Nested Failure Time Models

Example: Y survival time, $A(t) = 1$ if treated, 0 if not. Maybe

$$Y^{(0)} \sim \int_0^Y e^{\psi A(t)} dt;$$

$\sim$: has the same distribution as; "survival time with treatment is multiplied by a factor e^{ψ} if treatment is withheld". Maybe also, if patient is alive at time t,

$$Y^{(t)} - t \sim \int_t^Y e^{\psi A(t)} dt \text{ given } (\overline{L}_t, \overline{A}_t)$$

(remaining survival time; same interpretation). Compare with Accellerated Failure Time Model (Cox and Oaks, 1984). However: we do *not* make assumptions like

$$Y^{(t)} = t + \int_t^Y e^{\psi A(t)} dt.$$

Structural Nested Failure Time Models

$$Y^{(t)} - t \sim \int_t^Y e^{\psi A(t)} dt \quad \text{given } (\overline{L}_t, \overline{A}_t).$$

Multiplication factor: interpreted as in distribution!

Example continued, without covariates:

Observed outcome Y, outcome without treatment $Y^{(0)}$.

- Observed: patient not treated at all and observed: Y.

$$Y^{(0)} \sim \int_0^Y e^{\psi \cdot 0} dt = Y.$$

- Observed: patient treated all the time and observed: Y.

$$Y^{(0)} \sim \int_0^Y e^{\psi \cdot 1} dt = e^{\psi}\, Y.$$

Previously: Important controversy in the Structural Nested Models literature:

Previously, counterfactuals were often assumed to depend deterministically on the observed data: given the model and the parameter values, all counterfactual outcomes $Y^{(t)}$ for each person can simply be calculated from the observed data. Robins (1998): local rank preservation (in most cases, this implies global rank preservation).

Treatment is then said *not* to affect the outcome of interest if the outcome for any particular person would have been exactly the same regardless of which treatment was given.

Related to the assumption of constant treatment effect in Holland (1986): the difference between counterfactual outcomes under different treatments is a constant identical for all patients.

Deterministic dependence/ (local) rank preservation has frequently been attacked.

Deterministic dependence does *not* hold, for example, if two persons with the same observed data (e.g., both receiving some prophylactic drug) could have had a different outcome had they not been treated starting from some time t (e.g., one might have contacted a virus and the other might not).

Deterministic dependence can never be tested.

$\Rightarrow$ Deterministic dependence should be avoided if at all possible.

Deterministic dependence/ (local) rank preservation

Lok et al. (2004) proved that it is not necessary to assume a deterministic treatment effect for discrete-time Structural Nested Models; Lok (2017b) proved this for continuous-time Structural Nested Models.

Structural Nested Failure Time Model

Treatment effect model (for now).

Y survival time, $A(t) = 1$ if treated, 0 if not.

We will assume:

$$Y^{(0)} \sim \int_0^Y e^{\psi A(t)} dt \quad \text{given } (L_0, A_0), \quad \text{and}$$

$$Y^{(t)} \sim \begin{cases} Y & \text{given } (\overline{L}_t, \overline{A}_t) \quad \text{if } t \geq Y \\ t + \int_t^Y e^{\psi A(t)} dt & \text{given } (\overline{L}_t, \overline{A}_t) \quad \text{if } t < Y; \end{cases}$$

"survival time with treatment is multiplied by a factor e^{ψ} if treatment is withheld", in distribution.

Accellerated Failure Time Models (Cox and Oaks, 1984).

Structural Nested Failure Time Model: Treatment effect model

Define

$$X_\psi(t) = \begin{cases} Y & \text{if } t \geq Y \\ t + \int_t^Y e^{\psi A(t)} dt & \text{if } t < Y. \end{cases}$$

Then the **treatment effect model** reads:

$$Y^{(t)} \sim X_\psi(t) \quad \text{given} \quad (\overline{L}_t, \overline{A}_t),$$

$$\text{or} \quad Y^{(t)} \sim t + \int_t^Y e^{\psi A(t)} dt \quad \text{given} \quad (\overline{L}_t, \overline{A}_t).$$

Mimicking counterfactual outcomes

$X_{\psi^*}(t)$ will play the role in SNFTMs that $H_{\psi^*}(k)$ played in SNMMs: it mimics a counterfactual outcome.

Example from the literature on Structural Nested Models

Survival of AIDS patients. Robins et al. (1992): AIDS clinical trial: AZT treatment $\Rightarrow$ survival in HIV-infected patients.

Time 0: enrollment in the study.

Pneumocystis Carinii Pneumonia, is an opportunistic infection that affects HIV-infected patients.

Robins et al. (1992) used continuous-time Structural Nested Failure Time Models to study the effect of PCP prophylaxis therapy on survival of HIV-infected patients, using these trial data in which PCP prophylaxis was not randomized.

Outcome Y: survival time. Treatment A_k: PCP prophylaxis.

Example from the literature on Structural Nested Models

Robins et al. (1992) estimated the effect of changes in the time the treatment is discontinued. The local rank preservation assumption of Robins et al. (1992):

$$Y^{(t)} - t = \int_t^Y e^{\psi 1_{\text{prophylaxis at } s}} ds. \tag{3}$$

Assumption (3) is very strong, because it requires that given the model parameter ψ and the observed outcome Y, all counterfactual outcomes $Y^{(t)}$ can be calculated from the observed data. It suffices to assume that

$$Y^{(t)} - t \sim \int_t^Y e^{\psi 1_{\text{prophylaxis at } s}} ds \tag{4}$$

conditional on $(\overline{L}_t, \overline{A}_t)$ and $Y > t$, where $\sim$ means "has the same distribution as".

Example from the literature on Structural Nested Models

Given $(\overline{L}_t, \overline{A}_t)$, both $Y^{(t)} - t$ and $\int_t^Y e^{\psi 1_{\text{prophylaxis at } s}} ds$ are random variables, depending on $Y^{(t)}$ and Y, respectively.

Assumption (4) does not impose

$$Y^{(t)} - t = \int_t^Y e^{\psi 1_{\text{prophylaxis at } s}} ds,$$

but only that the distribution of these two random variables is the same conditional on $(\overline{L}_t, \overline{A}_t)$ and $Y > t$.

$\Rightarrow$ Patients who have the exact same observed history over $[0, k+1]$, $((\overline{L}_K, \overline{A}_K), Y)$, do not necessarily have the same counterfactual outcomes $Y^{(t)}$.

No (local) rank preservation in this example:

In clinical practice, $Y^{(t)}$ may differ between two patients with the exact same observed history.

Example in this PCP setting:

- Suppose that two patients with the exact same observed history were both on PCP prophylaxis.
- If one of the two patients got in contact with pneumococcal bacteria (and therefore might have caught PCP without the preventive treatment, PCP prophylaxis), and the other did not get in contact with pneumococcal bacteria (and therefore might not have caught PCP, even without PCP prophylaxis), the outcomes for these two patients could have been different had they not taken PCP prophylaxis.

Example from the literature on Structural Nested Models

In equation (4),

$$Y^{(t)} - t \sim \int_t^Y e^{\psi 1_{\text{prophylaxis at } s}} ds$$

the part of the residual survival time, $Y - t$, that is treated gets multiplied by e^{ψ} to attain the same distribution as $Y^{(t)} - t$ (the residual survival time under "no treatment from t onwards"), conditional on $(\overline{L}_t, \overline{A}_t)$ and $Y > t$.

$\Rightarrow$ Analogous to Accelerated Failure Time Models (see e.g. Cox and Oaks, 1984), the multiplication factor e^{ψ} can be interpreted in a distributional way.

Structural Nested Failure Time Models

Treatment effect model reads:

$$Y^{(t)} \sim X_{\psi}(t) \quad \text{given} \quad (\overline{L}_t, \overline{A}_t).$$

Aim: to estimate ψ^*, the true ψ.

Main assumption: No Unmeasured Confounding.

Assumption of No Unmeasured Confounding

For identifiability: need information on all factors that both:

1. Influence treatment decisions and
2. Possibly predict an individual's prognosis with respect to the outcome of interest.

No Unmeasured Confounding (informal):

"Treatment decisions not based on more information about a patient's health prognosis than in database."

$Y^{(k)}$ indication of a patient's prognosis at time k.

No unmeasured confounding ("formal"):

A_k, the treatment decision at time k, is independent of $Y^{(k)}$ given *measured* data $(\overline{L}_k, \overline{A}_{k-1})$.

SNFTMs: G-estimation of ψ^*

In the presence of confounding by indication, and assuming No Unmeasured Confounding, one uses a model to predict treatment initiation as a tool to estimate ψ.

Overview of G-estimation (Robins et al. (1992)):

- No unmeasured confounding: $Y^{(t)}$ does not add to the prediction model for treatment initiation at time t.
- Hope: $X_{\psi^*}(t)$, which mimics $Y^{(t)}$, does not add to the prediction model for treatment initiation.
- G-estimation: test this. If $X_{\psi^*}(t)$ adds, then ψ^* not true parameter.
- G-estimation: choose estimate $\hat{\psi}$ so that $X_{\hat{\psi}}(t)$ "adds the least" to the prediction model for treatment initiation.

Prediction of treatment given the past

Prediction of treatment given the past: tool to estimate treatment effect: propensity score:

$$p(k) := P\left(A_k = 1 | \overline{A}_{k-1} = \overline{0}, \overline{L}_k\right),$$

for $k = 0, \ldots, K$.

In most cases, $p(k)$ is unknown and needs to be estimated.
Standard software.

Example of propensity score model:

Example:

$$\text{logit} p_\theta(k) = \log(p_\theta(k)/(1 - p_\theta(k))) = \theta_0 + \theta_1 I_{\text{AZT}} + \theta_2 I_{\text{PCP}}(k),$$

or

$$p_\theta(k) = 1_{\overline{A}_{k-1}=\overline{0}} \frac{1}{1 + e^{-\theta_0 - \theta_1 I_{\text{AZT}} - \theta_2 I_{\text{PCP}}(k)}}.$$

G-estimation of ψ

Add X_ψ to the model for treatment changes:

$$\text{logit} p_{\theta,\alpha}(k) = \log(p_{\theta,\alpha}(k)/(1 - p_{\theta,\alpha}(k))) = \vec{\theta} \cdot \vec{f}(\overline{L}_k) + \alpha X_\psi(k),$$

or

$$p_{\theta,\alpha}(k) = 1_{\overline{A}_{k-1}=\overline{0}} \frac{1}{1 + e^{-\vec{\theta} \cdot \vec{f}(\overline{L}_k) - \alpha X_\psi(k)}}.$$

Estimate $(\hat{\theta}(\psi), \hat{\alpha}(\psi))$: standard software for logistic regression.

G-estimation of ψ

$\hat{\psi}$: the ψ that adds the least to the prediction model for treatment initiation, so that generates $\hat{\alpha}(\psi) = 0$.

Example of G-estimation:

$$\text{logit} p_{\theta,\alpha}(k) = \theta_0 + \theta_1 I_{\text{AZT}} + \theta_2 I_{\text{PCP}}(k) + \alpha X_\psi(k).$$

$\Rightarrow$: $\hat{\psi}$: the ψ for which $\hat{\alpha}(\psi) = 0$.

Consistency and asymptotic normality of $\hat{\psi}$, for Structural Nested Failure Time Models

Lok, Annals of Statistics (2008) and Lok, Scandinavian Journal of Statistics (2007) have shown that $\hat{\psi}$ solves Unbiased Estimating Equations.

Then, under regularity conditions, $\hat{\psi}$ is consistent and asymptotically normal:

$$\hat{\psi} \rightarrow^{P} \psi^*,$$

$$\sqrt{n}(\hat{\psi} - \psi^*) \rightarrow^{\mathcal{D}} \mathcal{N}(0, \Sigma),$$

for some covariance matrix Σ.

Test for treatment effect using Structural Nested Failure Time Models

If $\psi^* \neq 0$: $Y^{(t)}$ has a different distribution than Y given $\overline{Z}_t$.

$\Rightarrow$ If $\psi^* \neq 0$: treatment affects the outcome of interest.

$\Rightarrow$ We can test

H_0: "Treatment doesn't affect the outcome of interest"

by testing whether $\psi^* = 0$.

Test for treatment effect using Structural Nested Failure Time Models

If "no treatment effect" then adding $X_{\psi^*=0}(t)$ to the model for treatment initiation should not help.

$X_{\psi^*=0}(t) = Y$! So, if no treatment effect then adding Y to the model for treatment initiation should not help: if

$$\text{logit} p_{\theta,\alpha}(k) = \theta_0 + \theta_1 I_{\text{AZT}} + \theta_2 I_{\text{PCP}}(k) + \alpha Y,$$

the true α is 0.

- Can estimate $(\theta_0, \theta_1, \theta_2, \alpha)$.
- If no treatment effect: true α equals 0.
- Test for treatment effect: test whether $\alpha = 0$.

Standard software: just add Y to the model for treatment initiation.

Structural Nested Failure Time Models: Alternative models for treatment effect: an example

Because the data in Robins et al. (1992) were from a clinical trial for AZT treatment, AZT treatment is described by a single variable R indicating the treatment arm the patient was randomized to (R equals 1 or 2). Let $PCP(t) = 1$ if the patient had PCP before or at time t *and* before prophylaxis treatment started; otherwise $PCP(t) = 0$.
The model described in Robins et al. (1992):

$$Y^{(t)} - t \sim \int_t^Y e^{1_{\{\text{prophylaxis at } s\}}(\psi_1+\psi_2 PCP(s)+\psi_3 R)} ds \text{ given } (\overline{L}_t, \overline{A}_t), \tag{5}$$

for $t < Y$.

Structural Nested Models: Alternative models for treatment effect

More generally, can have

$$Y^{(t)} \sim X_\psi(t) \quad \text{given} \quad (\overline{L}_t, \overline{A}_t).$$

Goal: estimate ψ^*, the true ψ.

Structural Nested Models: estimation: general models for treatment effect

Same procedure as before:

- Assume correctly specified treatment prediction model.
- If no unmeasured confounding: $Y^{(t)}$ doesn't help predict.
- Like before can prove: then $X_{\psi^*}(t)$ doesn't help predict.
- To estimate: choose $\hat{\psi}$ so that $X_{\hat{\psi}}$ "predicts the least".
- To test $H_0 : \psi = \psi^*$: test whether X_{ψ^*} predicts.
- Confidence interval: ψ^*'s which are not rejected.

Structural Nested Failure Time Models and standard software: testing

Testing: whether $\psi = \psi^*$ can be tested by testing whether αX_{ψ^*} adds to the prediction model for treatment initiation, which can be tested by testing whether $\alpha = 0$.

Can be done with standard software.

Proposed by Robins, see e.g. Robins et al. (1992); Robins (1998). Can be further understood using theory on partial likelihood (Lok (2007)).

Structural Nested Failure Time Models and standard software: estimation

Finding ψ so that X_ψ "adds the least" to the prediction model for treatment initiation can be done as follows:

- Create a grid of potential ψ's.
- Add αX_ψ to the prediction model for each ψ separately.
- Estimate $\alpha(\psi)$ for each ψ separately, using standard software.
- Check: for which ψ does $\hat{\alpha}(\psi) = 0$?
- $\Rightarrow \hat{\psi}$.

Proposed by Robins, see e.g. Robins et al. (1992); Robins (1998). Can be further understood using theory on partial likelihood (Lok (2007)).

Structural Nested Failure Time Models and standard software: estimation: alternative

If X_ψ is linear in ψ (or a function of ψ such as e^ψ), the estimating equations arising from a pooled logistic regression model are also linear in ψ (or a function of ψ such as e^ψ).

Estimating equations could then be solved using standard software to solve linear equations, such as PROC IML in SAS.

Proposed in Lok and DeGruttola, Biometrics (2012).

Administrative censoring and survival outcomes: Artificial censoring

In the case of a survival outcome, right censoring is common: after each visit time, a number of patients is no longer seen.

Often, there is administrative censoring: censoring due to end-of-follow-up because the study ends. Follow-up time C for each patient is time from enrollment until end-of-study.

Then, $X_\psi(t)$ not observed for censored outcomes ($T < C$), and whether $X_\psi(t)$ is observed may depend on whether or not treated (if treatment affects the outcome).

Administrative censoring and survival outcomes: Artificial censoring

Idea:

- Construct a variable that is a function of both X_ψ and C that is always observed
- Add that to the prediction model for treatment changes and find $\hat{\alpha} = 0 \Rightarrow \hat{\psi}$.

Proposed in e.g. Robins (1998), proof see Lok, Annals of Statistics (2008).

The artificial censoring estimator treats the censoring time C as a baseline covariate. This is justified in the case of censoring due to study closure, because then C only depends on the date a patient enrolled in the study.

Administrative censoring and survival outcomes: Artificial censoring

Instead of adding $X(t)$ or $X(0)$ to the model for predicting treatment changes, one could add a function $\tilde{X}(0) = g(X(0), C)$ of $X(0)$ and the censoring time C, which is observed for all patients.

Conditional on $(\overline{L}_k, \overline{A}_{k-1})$, functions of $X(k)$ and $(\overline{L}_k, \overline{A}_{k-1})$ are not predictive of treatment changes (No Unmeasured Confounding).

$\Rightarrow$ Conditional on $(\overline{L}_k, \overline{A}_{k-1})$, $\tilde{X}(0) = g(X(0), C)$ is not predictive of treatment changes either.

Administrative censoring and survival outcomes: Artificial censoring

$\Rightarrow$ Conditional on $(\overline{L}_k, \overline{A}_{k-1})$, $\tilde{X}(0) = g(X(0), C)$ is not predictive of treatment changes.

This produces an estimation procedure for ψ analogous to without censoring, but that allows for right censoring.

Which function $\tilde{X}(0) = g(X(0), C)$?

We slightly adapt Robins (1998), and propose to add a function $\tilde{X}(k)$ of $X(k)$ and C to the model for the prediction of treatment changes.

Administrative censoring and survival outcomes: Artificial censoring

In particular, for the model in equation (4),

$$Y^{(t)} - t \sim \int_t^Y e^{\psi 1_{\text{prophylaxis at } s}} ds$$

and for $min(Y, C) \geq t$, one could add to the prediction model of treatment changes the function $\tilde{X}(t, \psi) = \min(X_\psi(t), C(t, \psi))$, with

$$C(t, \psi) = \begin{cases} C & \text{if } \psi \geq 0 \\ t + e^{\psi}(C - t) & \text{if } \psi < 0. \end{cases}$$

As required, $\tilde{X}(t, \psi)$ is a function of $X_\psi(t)$ and C.

To be able to use: needs to be observed for all patients.

Administrative censoring and survival outcomes: Artificial censoring

We show that both for the case that $\psi \geq 0$ and for the case that $\psi < 0$, $\tilde{X}(t, \psi)$ is observed for all patients. This follows from the fact that $\tilde{X}(t, \psi) = min(X^*(t, \psi), C(t, \psi))$ with $X^*(t, \psi) = t + \int_t^{min(Y,C)} e^{\psi 1_{\text{prophylaxis at } s}} ds$, which is observed for all patients. For $\psi \geq 0$, this follows from

$$
\begin{aligned}
&\tilde{X}(t, \psi) \\
&= min(t + \int_t^Y e^{\psi 1_{\text{prophylaxis at } s}} ds, C) \\
&= min(t + \int_t^Y e^{\psi 1_{\text{no prophylaxis at } s}} ds, t + \int_t^C e^{\psi 1_{\text{no prophylaxis at } s}} ds, C) \\
&= min(X^*(t, \psi), C(t, \psi)),
\end{aligned}
$$

where for the second equality we used that for $\psi \geq 0$, $t + \int_t^C e^{\psi 1_{\text{prophylaxis at } s}} ds \geq C$.

Administrative censoring and survival outcomes: Artificial censoring

We show that both for the case that $\psi \geq 0$ and for the case that $\psi < 0$, $\tilde{X}(t, \psi)$ is observed for all patients. This follows from the fact that $\tilde{X}(t, \psi) = min(X^*(t, \psi), C(t, \psi))$ with $X^*(t, \psi) = t + \int_t^{min(Y,C)} e^{\psi 1_{\text{prophylaxis at } s}} ds$, which is observed for all patients. For $\psi < 0$, this follows from

$$\begin{aligned}
\tilde{X}(t, \psi) &= min(t + \int_t^Y e^{\psi 1_{\text{prophylaxis at } s}} ds, t + e^{\psi}(C - t)) \\
&= min(t + \int_t^Y e^{\psi 1_{\text{prophylaxis at } s}} ds, t + \int_t^C e^{\psi 1_{\text{prophylaxis at } s}} ds, \\
&\qquad t + e^{\psi}(C - t)) \\
&= min(X^*(t, \psi), C(t, \psi)),
\end{aligned}$$

where for the second equality we used that for $\psi < 0$, $t + \int_t^C e^{\psi 1_{\text{prophylaxis at } s}} ds \geq t + e^{\psi}(C - t)$.

Administrative censoring and survival outcomes: Artificial censoring

For the case that $\psi < 0$, some patients are "artificially" censored, since if $C > t$, $C(t, \psi) = t + e^{\psi}(C - t) < C$.

Robins (1998) suggests to also consider adding

$$\Delta(t, \psi) = 1_{\tilde{X}(t,\psi) \leq C(t,\psi)}$$

to the model for the prediction of treatment changes.

Both $\tilde{X}(t, \psi)$ and $C(t, \psi)$ are observed for all patients $\Rightarrow$ so is $\Delta(t, \psi)$.

$\Rightarrow$ The previous reasoning shows that adding $\tilde{X}(t, \psi)$ or $\Delta(t, \psi)$ to the model for treatment initiation also leads to consistent estimation of the treatment effect model.

Administrative censoring and survival outcomes: Artificial censoring

Artificial censoring can easily be adapted to for example the treatment effect model

$$Y^{(t)} - t \sim \int_t^Y e^{1_{\{\text{prophylaxis at } s\}}(\psi_1+\psi_2 PCP(s)+\psi_3 R)} ds \text{ given } (\overline{L}_t, \overline{A}_t),$$

for $t < Y$, by replacing $C(t, \psi)$ accordingly.

For that case one could use

$$C(t, \psi) = t + e^{min(\psi_1,0)+min(\psi_2,0)+min(\psi_3,0)}(C - t).$$

Conclusions SNFTMs

If there is No Unmeasured Confounding:

- SNFTMs often lead to consistent, asymptotically normal estimators.
- SNFTMs can often be estimated using standard software.
- Right censoring: artificial censoring for administrative censoring.

Bibliography I

Andersen, P. K., Ø. Borgan, R. D. Gill, and N. Keiding (1993). *Statistical models based on counting processes*. Springer series in statistics. New York: Springer–Verlag.

Dawid, A. P. (1979). Conditional independence in statistical theory (with discussion). *Journal of the Royal Statistical Society B 41*, 1–31.

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association 81*, 954–960.

Keiding, N. (1999). Event history analysis and inference from observational epidemiology. *Statistics in Medicine 18*, 2353–2363.

Lok, J. J. (2001, May). *Statistical modelling of causal effects in time*. Ph. D. thesis, Department of Mathematics, Free University of Amsterdam. http://www.math.vu.nl/research/theses/pdf/lok.pdf.

Bibliography II

Lok, J. J. (2007). Structural nested models and standard software: a mathematical foundation through partial likelihood. *Scandinavian Journal of Statistics 34*(1), 186–206.

Lok, J. J. (2008). Statistical modelling of causal effects in continuous time. *Annals of Statistics 36*(3), 1464–1507. arXiv: math.ST/0410271 at http://arXiv.org.

Lok, J. J. (2017a). Choice of optimal estimators in Structural Nested Mean Models with application to initiating HAART in HIV positive patients after varying duration of infection. Reject and resubmit from *Biometrika*.

Lok, J. J. (2017b). Mimicking counterfactual outcomes to estimate causal effects. *The Annals of Statistics 45*(2), 461–499. arXiv: http://arxiv.org/abs/math/0409045.

Bibliography III

Lok, J. J. and V. DeGruttola (2012). Impact of time to start treatment following infection with application to initiating HAART in HIV-positive patients. *Biometrics 68*, 745—-754.

Lok, J. J., R. D. Gill, A. W. van der Vaart, and J. M. Robins (2004). Estimating the causal effect of a time–varying treatment on time-to-event using structural nested failure time models. *Statistica Neerlandica 58*(3), 271–295.

Lok, J. J. and M. D. Hughes (2016). Evaluating predictors of competing risk outcomes when censoring depends on time-dependent covariates, with application to safety and efficacy of HIV treatment. *Statistics in Medicine 35*(13), 2183–2194.

Mark, S. D. and J. M. Robins (1993). Estimating the causal effect of smoking cessation in the presence of confounding factors using a rank preserving structural failure time model. *Statistics in Medicine 12*, 1605–1628.

Bibliography IV

Newey, W. K. and D. McFadden (1994). Large Sample Estimation and Hypothesis Testing. In R. F. Engle and D. L. McFadden (Eds.), *Handbook of Econometrics*, Volume 4, pp. 2111–2245. Elsevier. Edition 1, chapter 36.

Pearl, J. (2000). *Causality. Models, reasoning, and inference*. Cambridge: Cambridge University Press.

Robins, J. M. (1992). Estimation of the time–dependent accelerated failure time model in the presence of confounding factors. *Biometrika 79*, 321–334.

Robins, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics 23*, 2379–2412.

Robins, J. M. (1995). Causal inference from complex longitudinal data. In *Design and analysis of follow-up studies*. Rotterdam: The Netherlands Institute for Health Sciences. Lecture notes.

Bibliography V

Robins, J. M. (1998). Structural Nested Failure Time Models. In P. Armitage and T. Colton (Eds.), *Survival analysis*, Volume 6 of *Encyclopedia of Biostatistics*, pp. 4372–4389. Chichester, UK: John Wiley and Sons. Section Eds: P. K. Andersen and N. Keiding.

Robins, J. M. (2000). Marginal Structural Models versus Structural Nested Models as tools for causal inference. In M. E. Halloran and D. Berry (Eds.), *Statistical Models in Epidemiology: The Environment and Clinical Trials*, Volume 116, pp. 95–133. New York: Springer-Verlag.

Robins, J. M., D. Blevins, G. Ritter, and M. Wulfsohn (1992). G–estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of AIDS patients. *Epidemiology 3*(4), 319–336.

Bibliography VI

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association 90*(429), 106–121.

Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*, 581–592.

Van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge series in statistical and probabilistic mathematics. Cambridge: Cambridge University Press.

Westreich, D., S. R. Cole, J. G. Young, F. Palella, P. C. Tien, L. Kingsley, S. J. Gange, and M. A. Hernán (2012). The parametric g-formula to estimate the effect of highly active antiretroviral therapy on incident AIDS or death. *Statistics in Medicine 31*(18), 2000–2009.

Bibliography VII

Yang, S., S. Little, D. Smith, V. DeGruttola, and J. J. Lok (2016). Impact of each month delay in initiation on the effect of 1 year of ART on CD4 count. In *Conference on Retroviruses and Opportunistic Infections (CROI)*. Poster abstract 16-647.

Yang, S. and J. J. Lok (2016). A goodness-of-fit test for Structural Nested Mean Models. *Biometrika 103*(3), 734–741.

Yang, S. and J. J. Lok (2017). Sensitivity analysis for unmeasured confounding in coarse Structural Nested Mean Models. Accepted by *Statistica Sinica*, available on sinica.edu.tw: Statistica Sinica Preprint No: SS-2016-0133R1.

Zhang, Y., J. G. Young, M. Thamer, and M. A. Hernán (2018). Comparing the Effectiveness of Dynamic Treatment Strategies Using Electronic Health Records: An Application of the Parametric g-Formula to Anemia Management Strategies. *Health services research 53*(3), 1900–1918.

Acknowledgements

Acknowledgements

I am grateful to the patients who volunteered for AIEDRP, to the AIEDRP study team, and to Susan Little, Davey Smith, and Christy Anderson for their help and advice in interpreting the AIEDRP database. I would like to thank James Robins and Victor DeGruttola for insightful and fruitful discussions. This work was sponsored by the Milton Fund (PI: Lok), the Career Incubator Fund award from the Harvard School of Public Health (PI: Lok), and NIH grants NIAID R01AI100762 (PI: Lok), R01 AI51164 (PI: DeGruttola), R37 AI032475 (PI: Robins), AI43638 (PI: Little), AI74621 (PI: Little), and AI36214 (PI: Smith).

Not their fault!

The opinions expressed herein are solely those of the author and do not necessarily represent those of Harvard or the agencies that have funded this research.

Thanks for listening!!!

SNMMs: Mimicking Counterfactual Outcomes

Backwards induction, starting with $k = K$.

Proof for $k = K$:

$$\begin{aligned} & E\left[H(K)|\overline{L}_K = \overline{l}_K, \overline{A}_K = \overline{a}_K\right] \\ = \;& E\left[Y - \gamma_K\left(\overline{L}_K, \overline{A}_K\right)|\overline{L}_K = \overline{l}_K, \overline{A}_K = \overline{a}_K\right] \\ = \;& E\left[Y^{(\overline{a}_K, \overline{0})} - \gamma_K\left(\overline{l}_K, \overline{a}_K\right)|\overline{L}_K = \overline{l}_K, \overline{A}_K = \overline{a}_K\right] \\ = \;& E\left[Y^{(\overline{a}_{K-1}, \overline{0})}|\overline{L}_K = \overline{l}_K, \overline{A}_K = \overline{a}_K\right]. \end{aligned}$$

Equality 2: Consistency.

Proof for $k = K - 1$:

$$
\begin{aligned}
& E\left[H(K-1)|\overline{L}_{K-1} = \bar{l}_{K-1}, \overline{A}_{K-1} = \bar{a}_{K-1}\right] \\
&= E\left[H(K) - \gamma_{K-1}\left(\overline{L}_{K-1}, \overline{A}_{K-1}\right)|\overline{L}_{K-1} = \bar{l}_{K-1}, \overline{A}_{K-1} = \bar{a}_{K-1}\right] \\
&= E\left[E\left[H(K)|\overline{L}_K, \overline{A}_K\right]|\overline{L}_{K-1} = \bar{l}_{K-1}, \overline{A}_{K-1} = \bar{a}_{K-1}\right] \\
&\quad -\gamma_{K-1}\left(\bar{l}_{K-1}, \bar{a}_{K-1}\right) \\
&= E\left[E\left[Y^{(\bar{a}_{K-1},\bar{0})}|\overline{L}_K, \overline{A}_K\right]|\overline{L}_{K-1} = \bar{l}_{K-1}, \overline{A}_{K-1} = \bar{a}_{K-1}\right] \\
&\quad -\gamma_{K-1}\left(\bar{l}_{K-1}, \bar{a}_{K-1}\right) \\
&= E\left[Y^{(\bar{a}_{K-1},\bar{0})}|\overline{L}_{K-1} = \bar{l}_{K-1}, \overline{A}_{K-1} = \bar{a}_{K-1}\right] \\
&\quad -\gamma_{K-1}\left(\bar{l}_{K-1}, \bar{a}_{K-1}\right) \\
&= E\left[Y^{(\bar{a}_{K-2},\bar{0})}|\overline{L}_{K-1} = \bar{l}_{K-1}, \overline{A}_{K-1} = \bar{a}_{K-1}\right].
\end{aligned}
$$

Equality 3: Induction.

Proof for k smaller, given result holds for $k + 1$:

Similar.

Technical details for G-estimation of SNMMs

Estimate α: standard software solves (partial) score equations for this model: sum over all patients of

$$\sum_{k=0}^{K} \begin{pmatrix} 1 \\ I_{\mathrm{AZT}} \\ I_{\mathrm{PCP}}(k) \\ H_\psi(k) \end{pmatrix} (A_k - p_{\theta,\alpha}(k)).$$

Set to 0 and solve for (θ, α).
Write $P_n X_i$ for $\frac{1}{n}\sum_{i=1}^{n} X_i$: the empirical average over all patients.
Then, standard software solves

$$P_n \sum_{k=0}^{K} \begin{pmatrix} 1 \\ I_{\mathrm{AZT}} \\ I_{\mathrm{PCP}}(k) \\ H_\psi(k) \end{pmatrix} (A_k - p_{\theta,\alpha}(k)) = 0.$$

G-estimation of SNMMs

$\hat{\psi}$: the ψ that adds the least to the prediction model for treatment initiation, so that generates $\hat{\alpha} = 0$.

So, this solves the (partial) score equations for this model and generates $\hat{\alpha} = 0$, so it solves the equations with α set to 0:

$$P_n \sum_{k=0}^{K} \begin{pmatrix} 1 \\ I_{\mathrm{AZT}} \\ I_{\mathrm{PCP}}(k) \\ H_\psi(k) \end{pmatrix} (A_k - p_{\theta,\alpha=0}(k)) = 0,$$

so (!)

$$P_n \sum_{k=0}^{K} \begin{pmatrix} 1 \\ I_{\mathrm{AZT}} \\ I_{\mathrm{PCP}}(k) \\ H_\psi(k) \end{pmatrix} (A_k - p_\theta(k)) = 0.$$

G-estimation of SNMMs

So, the procedure solves

$$P_n \sum_{k=0}^{K} \begin{pmatrix} 1 \\ I_{\mathrm{AZT}} \\ I_{\mathrm{PCP}}(k) \\ H_\psi(k) \end{pmatrix} (A_k - p_\theta(k)) = 0.$$

The first equations generate $\hat{\theta}$, the maximum likelihood estimate for θ^*. The last equations, with $\hat{\theta}$ plugged in, determine $\hat{\psi}$.

G-estimation of SNMMs

First 3 equations determine $\hat{\theta}$ (note: don't depend on ψ). Plug that into last equation and solve: $\hat{\psi}$.

$\Rightarrow$ Then $(\hat{\theta}, \hat{\psi})$ jointly solve the estimating equations.

And indeed: the estimating equations for $\hat{\psi}$ are Unbiased Estimating Equations: for the true ψ and the true θ:

G-estimation of SNMMs

$$
\begin{aligned}
& E\left(\sum_{k=0}^{K} H_{\psi^*}(k)(A_k - p_{\theta^*}(k))\right) \\
= \quad & E\left(E\left[\sum_{k=0}^{K} H_{\psi^*}(k)(A_k - p_{\theta^*}(k))|\overline{L}_k, \overline{A}_k\right]\right) \\
= \quad & E\left(\sum_{k=0}^{K} E\left[H_{\psi^*}(k)|\overline{L}_k, \overline{A}_k\right](A_k - p_{\theta^*}(k))\right)
\end{aligned}
$$

First equality: Law of Iterated Expectations.

G-estimation of SNMMs

$$
\begin{aligned}
&= E\left(\sum_{k=0}^{K} E\left[H_{\psi^*}(k)|\overline{L}_k, \overline{A}_k\right](A_k - p_{\theta^*}(k))\right) \\
&= E\left(\sum_{k=0}^{K} E\left[H_{\psi^*}(k)|\overline{L}_k, \overline{A}_{k-1}\right](A_k - p_{\theta^*}(k))\right) \\
&= E\left(\sum_{k=0}^{K} E\left[E\left[H_{\psi^*}(k)|\overline{L}_k, \overline{A}_{k-1}\right](A_k - p_{\theta^*}(k))|\overline{L}_k, \overline{A}_{k-1}\right]\right) \\
&= E\left(\sum_{k=0}^{K} E\left[H_{\psi^*}(k)|\overline{L}_k, \overline{A}_{k-1}\right] E\left[(A_k - p_{\theta^*}(k))|\overline{L}_k, \overline{A}_{k-1}\right]\right) \\
&= E\left(\sum_{k=0}^{K} E\left[H_{\psi^*}(k)|\overline{L}_k, \overline{A}_{k-1}\right] \cdot 0\right) = 0.
\end{aligned}
$$

Second equality: No Unmeasured Confounding and Mimicking Counterfactual Outcomes!

Doubly Robust Estimating Equations

Doubly Robust Estimating Equations are e.g.:

$$P_n \sum_{k=0}^{K} \begin{pmatrix} 1 \\ I_{\mathrm{AZT}} \\ I_{\mathrm{PCP}}(k) \\ H_\psi(k) - E\left[H_\psi(k) | \overline{L}_k, \overline{A}_{k-1} = \overline{0}\right] \end{pmatrix} 1_{\overline{A}_{k-1}=\overline{0}}(k) \left(A_k - p_\theta(k)\right) =$$

$\hat{\theta}$ from first 3 equations, standard theory. Last entry: Unbiased Estimating Equation? Doubly Robust?

Suppose p_θ is correctly specified. Then, the term with H_{ψ^*} has expectation 0 as before. The term with $E\left[H_{\psi^*}(k) | \overline{L}_k, \overline{A}_{k-1} = \overline{0}\right]$ also has expectation 0, because $E\left[H_{\psi^*}(k) | \overline{L}_k, \overline{A}_{k-1} = \overline{0}\right]$ is a function of $\overline{L}_k$ only, so we can condition on $\overline{L}_k, \overline{A}_{k-1}$, and conclude: the term with $E\left[H_{\psi^*}(k) | \overline{L}_k, \overline{A}_{k-1} = \overline{0}\right]$ also has expectation 0.

Doubly Robust Estimating Equations

Suppose $E\left[H_\psi(k)|\overline{L}_k, \overline{A}_{k-1} = \overline{0}\right]$ is correctly specified. Then, if p_θ is wrong, θ is no longer going to be meaningful, but for any θ:

$$
\begin{aligned}
& E\left(\left(H_{\psi^*}(k) - E\left[H_{\psi^*}(k)|\overline{L}_k, \overline{A}_{k-1} = \overline{0}\right]\right) 1_{\overline{A}_{k-1}=\overline{0}}(k)\left(A_k - p_\theta(k)\right)\right) \\
= \quad & E\left(E\left[H_{\psi^*}(k) - E\left[H_{\psi^*}(k)|\overline{L}_k, \overline{A}_{k-1} = \overline{0}\right] 1_{\overline{A}_{k-1}=\overline{0}}(k)\left(A_k - p_\theta(k)\right)\right.\right. \\
& \qquad \left.\left.|\overline{L}_k, \overline{A}_k\right]\right) \\
= \quad & E\left(\left(E\left[H_{\psi^*}(k)|\overline{L}_k, \overline{A}_k\right] - E\left[H_{\psi^*}(k)|\overline{L}_k, \overline{A}_{k-1} = \overline{0}\right]\right)\right. \\
& \qquad \left. 1_{\overline{A}_{k-1}=\overline{0}}(k)\left(A_k - p_\theta(k)\right)\right) \\
= \quad & E\left(0 \cdot 1_{\overline{A}_{k-1}=\overline{0}}(k)\left(A_k - p_\theta(k)\right)\right) = 0.
\end{aligned}
$$

Last equality: Mimicking Counterfactual Outcomes and No Unmeasured Confounding.

Doubly Robust Estimation of Structural Nested Mean Models

We don't have $E\left[H_\psi(k)|\overline{L}_k, \overline{A}_{k-1} = \overline{0}\right]$, but we can use that

$$H(k) = Y - \gamma_k\left(\overline{L}_k, \overline{A}_k\right) - \gamma_{k+1}\left(\overline{L}_{k+1}, \overline{A}_{k+1}\right) - \ldots - \gamma_K\left(\overline{L}_K, \overline{A}_K\right).$$

If $\gamma_{k,\psi}$ is linear in ψ, we can condition each term on $(\overline{L}_k, \overline{A}_{k-1} = \overline{0})$ and obtain Doubly Robust Estimating Equations.

Continuous-time Structural Nested Models

Continuous-time Structural Nested Models: method proposed in Robins (1998).

If treatment can be initiated at any time, could fit a hazard model for treatment initiation.

Adding $Y^{(t)}$ to that model would not help predict treatment initiation, under No Unmeasured Confounding.

Can be shown: also adding X_{ψ^*} to that model would not help predict treatment initiation.

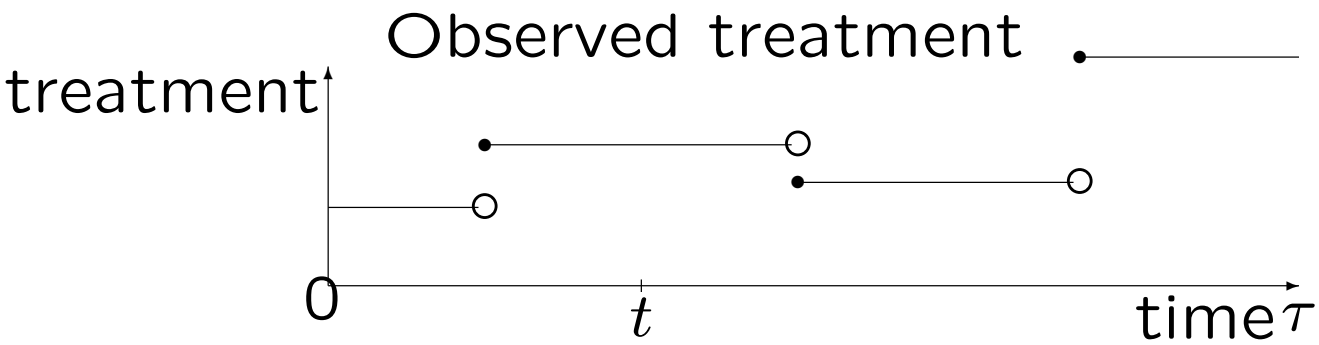
Leads to consistent, asymptotically normal, estimators for ψ as before.

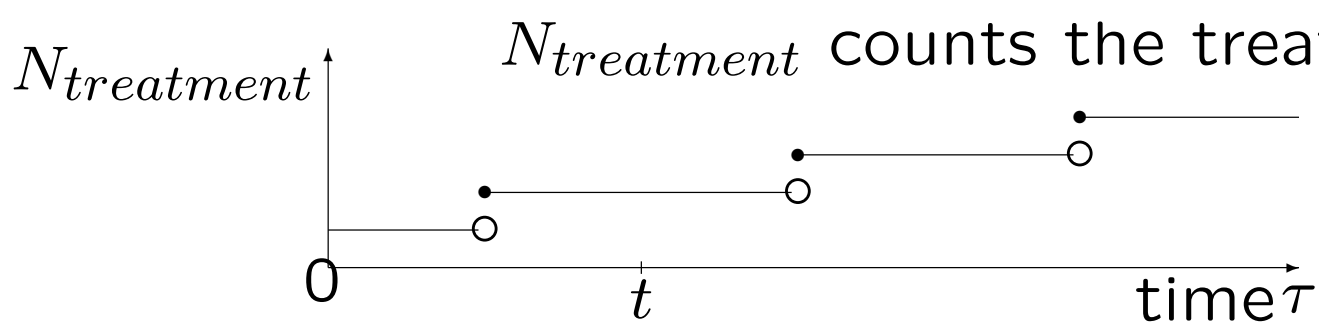
Proofs: Lok (2008) (based on martingale theory).

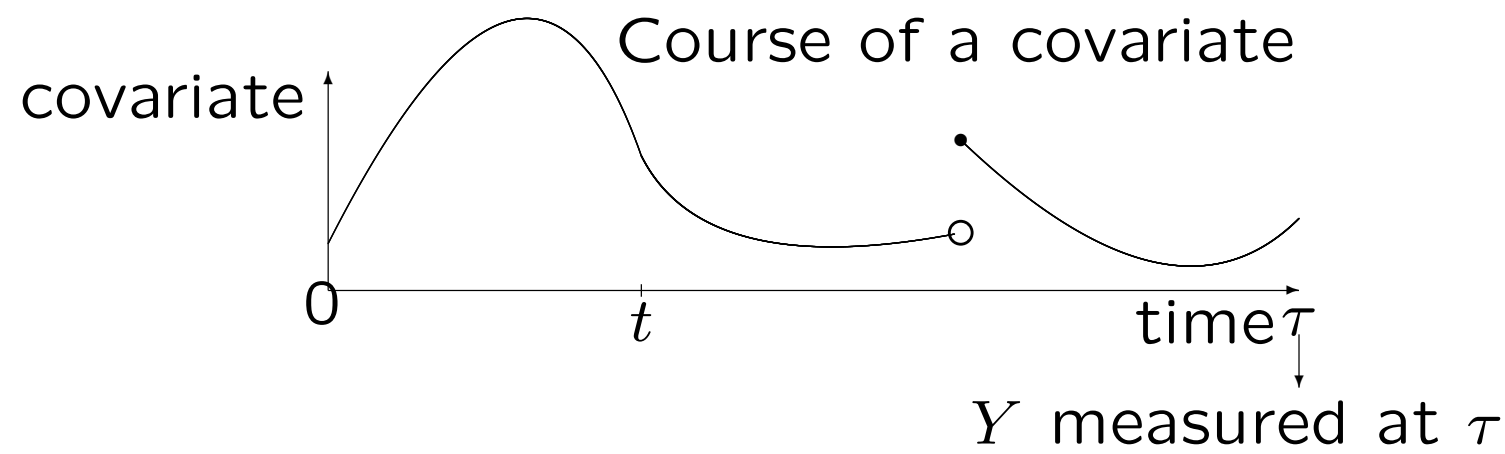
Continuous-time setting:

$\overline{Z}_t$: covariate- and treatment history until time t.

Continuous-time setting:







Model for treatment effect ("D")

Definition:

$$D\left(y, t; \overline{Z}_t\right) = \left.\frac{\partial}{\partial h}\right|_{h=0} \left(F^{-1}_{Y^{(t+h)}|\overline{Z}_t} \circ F_{Y^{(t)}|\overline{Z}_t}\right)(y)$$

Underlying idea: probably "no treatment effect" means "$D \equiv 0$" (and: YES)

Example:

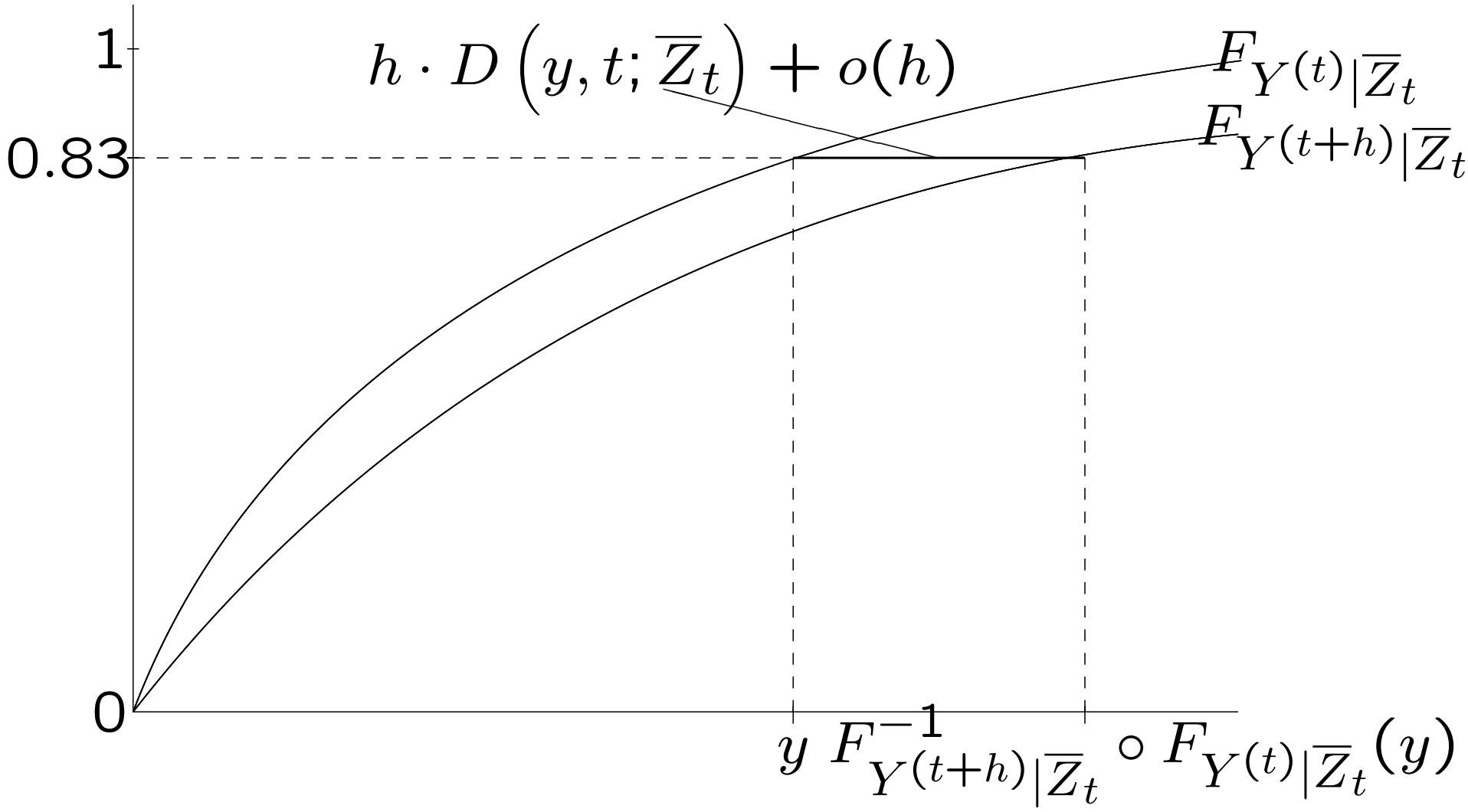
E.g.,

$$A(t) = \text{treatment at time } t$$

$$D_\psi\left(y, t; \overline{Z}_t\right) = \left(1 - e^{\psi A(t)}\right) 1_{\{\text{alive at } t\}}$$

$$\text{“ } Y^{(t+\Delta t)} \sim Y^{(t)} + \underbrace{\left(1 - e^{\psi^* A(t)}\right)}_{\text{infinitesimal effect of treatment } A(t)\Delta t} 1_{\{\text{alive at } t\}} \Delta t \text{ ”}$$

Illustration of the infinitesimal shift-function D



Estimation of treatment effect, D

Suppose: D known up to a parameter $\psi \in \mathbb{R}^d$:

$$D = D_{\psi^*}.$$

To estimate: ψ^*

- No unmeasured confounding: $Y^{(t)}$ does not add to the prediction model for treatment changes, λ
- Construct $X_\psi(t)$ to mimic $Y^{(t)}$ for $\psi = \psi^*$
- Hope: X_{ψ^*} does not add to the prediction model for treatment changes, λ.

Mimicking counterfactual outcomes

Definition: $X(t)$ is the continuous solution to the differential equation

$$X'(t) = D\left(X(t), t; \overline{Z}_t\right)$$

with final condition $X(\tau) = Y$.

Mimicking counterfactual outcomes (under regularity conditions): $X(t)$ has the same distribution as $Y^{(t)}$ given $\overline{Z}_t$ (!!)

Mimicking counterfactual outcomes

Under regularity conditions): $X(t)$ has the same distribution as $Y^{(t)}$ given $\overline{Z}_t$ (!!)

Example of $X(t)$: Y survival time, $A(t) = 1$ if treated, 0 if not. If

$$D\left(y, t; \overline{Z}_t\right) = \left(1 - e^{\psi A(t)}\right) 1_{\{\text{alive at } t\}}$$

then

$$X(0) = \int_0^Y e^{\psi A(t)} dt, \quad \text{and}$$

$$X(t) = \begin{cases} Y & \text{if } t \geq Y \\ t + \int_t^Y e^{\psi A(t)} dt & \text{if } t < Y. \end{cases}$$

Remark: D known $\Rightarrow$ X known (known function of data).

No unmeasured confounding: formal

Define $\overline{Y}^{(t)} = \left(Y^{(s)}\right)_{0 \leq s \leq t}$.

No unmeasured confounding:

rate λ with which $N_{treatment}$ jumps given past treatment- and covariate history $\overline{Z}_{t-}$

equals

rate λ with which $N_{treatment}$ jumps given past treatment- and covariate history $\overline{Z}_{t-}$ AND $\overline{Y}^{(t-)}$

Example: effect of PCP prophylaxis on survival in HIV-infected patients

$N(t) = 1$ if prophylaxis treatment for PCP started at or before time t, and $N(t) = 0$ otherwise. Suppose initiation of prophylaxis treatment has Weibull proportional hazard

$$\lambda_{\xi,\gamma,\theta}(t) = 1_{\{\text{at risk for prophylaxis at } t\}} \xi\gamma t^{\gamma-1} e^{\theta_1 I_{\mathrm{AZT}} + \theta_2 I_{\mathrm{PCP}}(t)}$$

with $I_{\mathrm{PCP}}(t) = 1$ if PCP before t and $I_{\mathrm{PCP}}(t) = 0$ otherwise and I_{AZT} indicating AZT treatment arm.

- can estimate $(\xi, \gamma, \theta_1, \theta_2)$; unbiased estimating equations from maximum likelihood approach are

$$P_n \int_0^\tau \begin{pmatrix} \frac{1}{\xi} \\ \frac{1}{\gamma} + \log t \\ I_{\mathrm{AZT}} \\ I_{\mathrm{PCP}}(t) \end{pmatrix} (dN(t) - \lambda_{\xi,\gamma,\theta}(t)dt) = 0.$$

Example (continued): effect of PCP prophylaxis on survival in HIV-infected patients

Add $Y^{(t)}$ to the model:

$$\lambda_{\xi,\gamma,\theta,\alpha}(t) = 1_{\{\text{at risk for prophylaxis at } t\}} \xi\gamma t^{\gamma-1} e^{\theta_1 I_{\text{AZT}} + \theta_2 I_{\text{PCP}}(t) + \alpha Y^{(t-)}}.$$

If $Y^{(t)}$ would be observed:

- can estimate $(\xi, \gamma, \theta_1, \theta_2, \alpha)$
- true α equals 0, so expect that $\hat{\alpha}$ tends to 0
- and indeed

$$E \int_0^\tau \begin{pmatrix} \frac{1}{\xi} \\ \frac{1}{\gamma} + \log t \\ I_{\text{AZT}} \\ I_{\text{PCP}}(t) \\ Y^{(t-)} \end{pmatrix} (dN(t) - \lambda_{\xi,\gamma,\theta,\alpha=0}(t)dt) = 0.$$

Result central to the proof of unbiased estimating equations

Theorem: No unmeasured confounding and mimicking counterfactuals and "no instantaneous treatment effect"[1] and regularity conditions:

rate λ with which $N_{treatment}$ jumps given past treatment- and covariate history $\overline{Z}_{t-}$

equals

rate λ with which $N_{treatment}$ jumps given $\overline{Z}_{t-}$ AND $\overline{Y}^{(t-)}$

equals

rate λ with which $N_{treatment}$ jumps given $\overline{Z}_{t-}$ AND $X(t)$

[1] $P\left(\exists t : \Delta N_{treatment}(t) = 1 \text{ and } Y^{(t)} \text{ jumps}\right) = 0.$

Unbiased estimating equations

Theorem: Suppose $N_{\text{treatment}}$ has bounded hazard λ, no unmeasured confounding, $Y^{()}$ cadlag, and no instantaneous treatment effect. Suppose for every $t \in [0, \tau]$, $X(t)$ has the same distribution as $Y^{(t)}$ given $\overline{Z}_t$. Then

$$E \int_0^\tau h_t \left(X(t), \overline{Z}_{t-} \right) \left(dN(t) - \lambda(t)\, dt \right) = 0$$

for each h_t satisfying regularity conditions. Thus if D_ψ and λ_θ are correctly specified parametric models for D and λ, respectively,

$$P_n \int_0^\tau h_t \left(X_\psi(t), \overline{Z}_{t-} \right) \left(dN(t) - \lambda_\theta(t)\, dt \right) = 0$$

is an unbiased estimating equation for (θ^*, ψ^*). h_t here is allowed to depend on ψ and θ, as long as it satisfies regularity conditions.

Restriction on estimating equations

Restriction: When we consider functions h_t from $\mathbb{R} \times \overline{\mathcal{Z}}_{t-}$ to $\mathbb{R}^k$, we assume that they are measurable and satisfy

a) h_t bounded by constant not depending on t and $\overline{Z}$

b) for all $t_0 \in [0, \tau]$, $y_0 \in \mathbb{R}$ and $\omega \in \Omega$,
$h_t\left(y, \overline{Z}_{t-}(\omega)\right) \to h_{t_0}\left(y_0, \overline{Z}_{t_0-}(\omega)\right)$ when $y \to y_0$ and $t \uparrow t_0$.

Consistency and asymptotic normality

Theorem:

for $N_{treatment}$ which only jumps at fixed times and
for $N_{treatment}$ with rate (intensity process) λ:

under no unmeasured confounding,
"no instantaneous treatment effect" and
regularity conditions:

the resulting $\hat{\psi}$ is
consistent
and
asymptotically normal

Partial likelihood

Suppose that $N_{treatment}$ can only jump at time

$$0 = \tau_0 < \tau_1 < \ldots < \tau.$$

"Split up" the covariate- and treatment process $\overline{Z}$ and the outcome Y:

$$\left(\left(\overline{Z}_{\tau_0-}, X\left(\tau_0\right)\right), \Delta N\left(\tau_0\right), \ldots, \left(\overline{Z}_{\tau_k-}, X\left(\tau_k\right)\right), \Delta N\left(\tau_k\right), \left(\overline{Z}_{\tau}, Y\right)\right),$$

and define

$$\tilde{p}_{\theta,\alpha}(k) = P\left(\Delta N\left(\tau_k\right) = 1 | \overline{Z}_{k-}, X\left(\tau_k\right)\right),$$

with θ modelling the dependence on $\overline{Z}_{k-}$ and α (fake-)modelling the dependence on $X\left(k\right)$. A partial likelihood for (θ, α):

$$L\left(\theta, \alpha\right) = \prod_{k=0}^{K} \tilde{p}_{\theta,\alpha}(k)^{\Delta N(\tau_k)} \left(1 - \tilde{p}_{\theta,\alpha}(k)\right)^{1-\Delta N(\tau_k)}.$$

Partial likelihood in continuous time

$\tilde{\lambda}_{\theta,\alpha}(t)$ rate at which $N(t)$ jumps given $\overline{Z}_{t-}$ (parameter θ) and $X(t)$ (parameter α).

$$\begin{aligned} L(\theta,\alpha) &= \prod_{t\in[0,\tau]} \tilde{\lambda}_{\theta,\alpha}(t)^{\Delta N(t)} \left(1-\tilde{\lambda}_{\theta,\alpha}(t)dt\right)^{1-\Delta N(t)} \\ &= \left(\prod_{t:\Delta N(t)=1} \tilde{\lambda}_{\theta,\alpha}(t)\right) e^{-\int_0^\tau \tilde{\lambda}_{\theta,\alpha}(t)dt}. \end{aligned}$$

$$\log L(\theta,\alpha) = \int_0^\tau \log \tilde{\lambda}_{\theta,\alpha}(t)\, dN(t) - \int_0^\tau \tilde{\lambda}_{\theta,\alpha}(t)\, dt.$$

$$\frac{\partial}{\partial\theta,\alpha} \log L(\theta,\alpha) = \int_0^\tau \begin{pmatrix} \frac{\partial}{\partial\theta}\log\tilde{\lambda}_{\theta,\alpha}(t) \\ \frac{\partial}{\partial\alpha}\log\tilde{\lambda}_{\theta,\alpha}(t) \end{pmatrix} \left(dN(t) - \tilde{\lambda}_{\theta,\alpha}(t)\, dt\right).$$

Would lead to unbiased estimating equations for θ and α if X were known.

Partial likelihood in continuous time, continued

However, $\alpha = 0$ is known and X is unknown. Plugging in $\alpha = 0$ and replacing X by X_ψ usually leads to unbiased estimating equations for θ and ψ:

$$P_n \int_0^\tau \begin{pmatrix} \frac{\partial}{\partial \theta} \log \lambda_\theta(t) \\ \frac{\partial}{\partial \alpha}\big|_{\alpha=0} \log \tilde{\lambda}^\psi_{\theta,\alpha}(t) \end{pmatrix} (dN(t) - \lambda_\theta(t)dt) = 0,$$

with $\frac{\partial}{\partial \alpha}\big|_{\alpha=0} \log \tilde{\lambda}^\psi_{\theta,\alpha}(t)$ a function of θ and X_ψ. They are martingales for the true θ and ψ: integral of predictable process with respect to martingale.

Example

Example: Add X_ψ to the model (for true ψ this should not help!):

$$\tilde{\lambda}_{\xi,\gamma,\theta,\alpha}(t) = 1_{\{\text{at risk for prophylaxis at } t\}} \xi\gamma t^{\gamma-1} e^{\theta_1 I_{\mathrm{AZT}} + \theta_2 I_{\mathrm{PCP}}(t) + \alpha X_\psi(t)}.$$

- can estimate $(\xi, \gamma, \theta_1, \theta_2, \alpha) = (\xi, \gamma, \theta_1, \theta_2, \alpha)(\psi)$
- for $\psi = \psi^*$: true α equals 0
- plug in $\alpha = 0$ by looking for $\hat{\alpha}(\psi) = 0$
- $\Rightarrow \hat{\psi}$.

Example (continued):

Resulting estimating equations: plug in $\hat{\alpha} = 0$ in mle estimating equations in this example:

$$P_n \int_0^\tau \begin{pmatrix} \frac{1}{\xi} \\ \frac{1}{\gamma} + \log t \\ I_{\mathrm{AZT}} \\ I_{\mathrm{PCP}}(t) \\ X_\psi(t) \end{pmatrix} (dN(t) - \lambda_{\xi,\gamma,\theta,\alpha=0}(t)dt).$$

Those estimating equations have expectation 0 for the true ψ^*.

Estimation with standard software

For ψ fixed standard software estimates $\left(\hat{\theta}(\psi), \hat{\alpha}(\psi)\right)$ from

$$P_n \int_0^\tau \begin{pmatrix} \frac{\partial}{\partial \theta} \log \tilde{\lambda}^{\psi}_{\theta,\alpha}(t) \\ \frac{\partial}{\partial \alpha} \log \tilde{\lambda}^{\psi}_{\theta,\alpha}(t) \end{pmatrix} \left(dN(t) - \tilde{\lambda}^{\psi}_{\theta,\alpha}(t)\, dt \right) = 0.$$

When searching for $\hat{\alpha}(\psi) = 0$ we solve

$$P_n \int_0^\tau \begin{pmatrix} \frac{\partial}{\partial \theta} \log \lambda_\theta(t) \\ \frac{\partial}{\partial \alpha}\Big|_{\alpha=0} \log \tilde{\lambda}^{\psi}_{\theta,\alpha}(t) \end{pmatrix} (dN(t) - \lambda_\theta(t)\, dt) = 0,$$

which are the same estimating equations as resulted from the partial likelihood approach.

Example (continued):

$$\lambda_{\xi,\gamma,\theta}(t)dt = 1_{\{\text{at risk for prophylaxis at } t\}}\xi\gamma t^{\gamma-1}e^{\theta_1 I_{\mathrm{AZT}}+\theta_2 I_{\mathrm{PCP}}(t)}.$$

Estimating equations for θ from the likelihood:

$$P_n \int_0^\tau \begin{pmatrix} \frac{1}{\xi} \\ \frac{1}{\gamma} + \log t \\ I_{\mathrm{AZT}} \\ I_{\mathrm{PCP}}(t) \end{pmatrix} (dN(t) - \lambda_{\xi,\gamma,\theta}(t)dt).$$

Add αX_ψ to the model:

$$\tilde{\lambda}_{\xi,\gamma,\theta,\alpha}(t)dt = 1_{\{\text{at risk for prophylaxis at } t\}}\xi\gamma t^{\gamma-1}e^{\theta_1 I_{\mathrm{AZT}}+\theta_2 I_{\mathrm{PCP}}(t)+\alpha X_\psi(t)}.$$

Estimating equations using standard software

Estimating equations software will use for θ, α are

$$P_n \int_0^\tau \begin{pmatrix} \frac{1}{\xi} \\ \frac{1}{\gamma} + \log t \\ I_{\mathrm{AZT}} \\ I_{\mathrm{PCP}}(t) \\ X_\psi(t) \end{pmatrix} \left(dN(t) - \tilde{\lambda}_{\xi,\gamma,\theta,\alpha}(t)dt \right).$$

Plug in $\hat{\alpha} = 0$:

$$P_n \int_0^\tau \begin{pmatrix} \frac{1}{\xi} \\ \frac{1}{\gamma} + \log t \\ I_{\mathrm{AZT}} \\ I_{\mathrm{PCP}}(t) \\ X_\psi(t) \end{pmatrix} (dN(t) - \lambda_{\xi,\gamma,\theta}(t)dt).$$

Unbiased estimating equations for (θ, ψ).

Testing "treatment has no effect" without modeling treatment effect

Definition (simple treatment regimes):
Follow the distribution of treatment as in reality, but stop treatment at a stopping time. This stopping time with respect to $\sigma\left(\overline{Z}_t\right)$ should be finite-valued.

Consistency Assumption:
Suppose S is a stopping time which can take values $\tau_1, \ldots, k$ and ∞. Y^S: outcome under simple treatment regime described by S. Then

$$Y^S = \begin{cases} Y^{(k)} & \text{on the event } \{S = k\} \\ Y & \text{on the event } \{S = \infty\}. \end{cases}$$

Testing "treatment has no effect" without modeling treatment effect

Theorem:
Under regularity conditions $D \equiv 0$ if and only if all simple treatment regimes lead to the same distribution of the outcome,

$\Rightarrow$ under regularity conditions, a test for whether $D \equiv 0$ (or for $\psi = 0$) is a **test for treatment effect**.

Test whether treatment affects the outcome by testing whether $D \equiv 0$, or whether $X(t) \equiv Y$.

Or more generally: test whether $D = D_0$ by testing whether $X(t) = X_{D_0}(t)$.

Testing with standard software: idea (J.M. Robins):

- For the true D: $\alpha(D)$ equals 0.
- Test for $\alpha(D)$=0:
  - Rejects $\Rightarrow$ reject D.
  - Does not reject $\Rightarrow$ do not reject D.

Theorem: the partial "score"-test, and the partial "likelihood ratio" test for whether $\alpha(D) = 0$ often have asymptotically correct level.

Testing with standard software

If one has calculated $X_D(t)$ for different D this can be done with standard software (if one uses standard models for prediction of treatment).

Partial "score" test with standard software; does treatment affect the outcome?

Partial "score" test computer programs perform: based on the asymptotic χ^2 distribution of

$$W = \sqrt{n}\big(P_n \tilde{U}_\alpha^Y(\hat{\theta})\big)^\top \big(I_{\alpha\alpha} - I_{\alpha\theta} I_{\theta^*}^{-1} I_{\alpha\theta}^\top\big)\widehat{\ }^{-1} \sqrt{n} P_n \tilde{U}_\alpha^Y(\hat{\theta})$$

under the null hypothesis of no treatment effect, with

$$\tilde{U}_\alpha^Y(\theta) = \int_0^\tau \frac{\partial}{\partial \alpha}\bigg|_{\alpha=0} \log \tilde{\lambda}_{\theta,\alpha}^Y(t)\,(dN(t) - \lambda_\theta(t)dt)$$

and $\hat{\theta}$ solving

$$P_n U_\theta(\theta) = P_n \int_0^\tau \frac{\partial}{\partial \theta} \log \lambda_\theta(t)\,(dN(t) - \lambda_\theta(t)dt) = 0$$

and with

$$I = \mathrm{COV}\begin{pmatrix} U_\theta(\theta^*) \\ \tilde{U}_\alpha^Y(\theta^*) \end{pmatrix}.$$

Partial "score" test with standard software; does treatment affect the outcome? (continued)

But we have that asymptotic χ^2 distribution, too:

$$\begin{aligned}
\sqrt{n}P_n\tilde{U}_\alpha^Y\left(\hat{\theta}\right) &= \sqrt{n}P_n\tilde{U}_\alpha^Y\left(\theta^*\right) + P_n\left.\frac{\partial}{\partial\theta}\right|_{\theta=\tilde{\theta}}\tilde{U}_\alpha^Y\left(\theta\right)\sqrt{n}\left(\hat{\theta}-\theta^*\right) \\
&= \sqrt{n}P_n\tilde{U}_\alpha^Y\left(\theta^*\right) + \left(E\left.\frac{\partial}{\partial\theta}\right|_{\theta=\theta^*}\tilde{U}_\alpha^Y\left(\theta\right)\right)I_{\theta^*}^{-1}\sqrt{n}P_nU_\theta\left(\theta^*\right) + o_P(1) \\
&= \begin{pmatrix} -I_{\alpha\theta}I_{\theta^*}^{-1} & \mathrm{Id}_{\dim\alpha} \end{pmatrix}\sqrt{n}P_n\begin{pmatrix} U_\theta\left(\theta^*\right) \\ \tilde{U}_\alpha^Y\left(\theta^*\right) \end{pmatrix} + o_P(1) \\
\to_{\mathcal{D}} & \mathcal{N}\left(0, \begin{pmatrix} -I_{\alpha\theta}I_{\theta^*}^{-1} & \mathrm{Id}_{\dim\alpha} \end{pmatrix}\begin{pmatrix} I_{\theta^*} & I_{\alpha\theta}^\top \\ I_{\alpha\theta} & I_{\alpha\alpha} \end{pmatrix}\begin{pmatrix} -I_{\theta^*}^{-1}I_{\alpha\theta}^\top \\ \mathrm{Id}_{\dim\alpha} \end{pmatrix}\right) \\
&= \mathcal{N}\left(0, I_{\alpha\alpha} - I_{\alpha\theta}I_{\theta^*}^{-1}I_{\alpha\theta}^\top\right).
\end{aligned}$$

Parameterization of continuous-time SNMs without rank preservation: an example

In this parameterization example, no one is treated at time zero, and once treatment is initiated, it is never stopped.

$Y^{(t)}$: counterfactual outcome had treatment been as given in reality until time t, and continued or initiated after that.

For example, if treatment was initiated by time t for a particular patient, $Y^{(t)}$ is the observed outcome for that patient, since he or she was already treated at time t and treatment is never stopped.

On the other hand, if treatment was not initiated by time t, $Y^{(t)}$ is the outcome had treatment been initiated at time t.

Parameterization of continuous-time SNMs without rank preservation: an example

$Y^{(t)}$: counterfactual outcome had treatment been as given in reality until time t, and continued or initiated after that.

Thus, in the definition of $Y^{(t)}$, the switch at time t to "some kind of baseline treatment regime $\overline{0}$" is, in this case, "treat continuously" from time t onwards.

Parameterization of continuous-time SNMs without rank preservation: an example

Let's study a setting with $t \in [0, 2]$. The subscript $_t$ indicates the treatment initiation time, so for example $L_{1,t}$ indicates (counterfactual) covariates at time 1 under "treatment started at time t". Similarly, the subscript $_\infty$ indicates (counterfactual) variables under no treatment. For example, $L_{2,\infty}$ indicates (counterfactual) covariates at time 2 under no treatment.

In this example, treatment can be initiated in continuous time, but the covariates are only measured at times 0, 1, and 2, so that the treatment and covariate history up to time t, $\overline{Z}_t$, consists of the treatment information up to time t and L_0, (L_0, L_1), or (L_0, L_1, L_2), depending on whether $t \in [0, 1)$, $t \in [1, 2)$, or $t = 2$.

Counterfactual covariates

Counterfactual covariates L:

$$
\begin{aligned}
L_0 &= \tilde{L}_0 + e_0, \\
L_{1,\infty} &= \tilde{L}_0 - \beta_0 + e_{1,\infty}, \\
L_{2,\infty} &= \tilde{L}_0 - 2\beta_0 + e_{2,\infty} \\
L_{1,t} &= \tilde{L}_0 - \beta_0 + \theta(1-t) + e_{1,t} \text{ for } t \in [0,1], \text{ and } L_{1,\infty} \text{ otherwise} \\
L_{2,t} &= \tilde{L}_0 - 2\beta_0 + \psi(2-t) + e_{2,t}.
\end{aligned}
$$

$\tilde{L}_0$ and the $e_{j,t}$: random variables with values in $\mathbb{R}$. $(1-t)$ and $(2-t)$: the durations of treatment until the respective covariate measurements.

Assume: the $e_{j,t}$ $(j = 0, 1, 2)$ are independent of $\tilde{L}_0$, and the $e_{2,t}$ have a distribution function which does not depend on t.
Assume: the $e_{2,t}$ are independent of all previous variables (and of the treatment initiation time, T, described below).

Counterfactual covariates

Counterfactual covariates L:

$$
\begin{aligned}
L_0 &= \tilde{L}_0 + e_0, \\
L_{1,\infty} &= \tilde{L}_0 - \beta_0 + e_{1,\infty}, \\
L_{2,\infty} &= \tilde{L}_0 - 2\beta_0 + e_{2,\infty} \\
L_{1,t} &= \tilde{L}_0 - \beta_0 + \theta(1-t) + e_{1,t} \text{ for } t \in [0,1], \text{ and } L_{1,\infty} \text{ otherwise} \\
L_{2,t} &= \tilde{L}_0 - 2\beta_0 + \psi(2-t) + e_{2,t}.
\end{aligned}
$$

Define $Y_t = L_{2,t}$, the counterfactual outcome with treatment initiated at time t. Could potentially be observed at time 2.

Parameterization of continuous-time SNMs without rank preservation: an example

$$
\begin{aligned}
L_0 &= \tilde{L}_0 + e_0, \\
L_{1,\infty} &= \tilde{L}_0 - \beta_0 + e_{1,\infty}, \\
L_{2,\infty} &= \tilde{L}_0 - 2\beta_0 + e_{2,\infty} \\
L_{1,t} &= \tilde{L}_0 - \beta_0 + \theta(1-t) + e_{1,t} \text{ for } t \in [0,1], \text{ and } L_{1,\infty} \text{ otherwise} \\
Y_t = L_{2,t} &= \tilde{L}_0 - 2\beta_0 + \psi(2-t) + e_{2,t}.
\end{aligned}
$$

The outcome processes adopted in this example are not rank preserving.

((because with probability one, two patients with the same observed data do not have the same value of $\tilde{L}_0$))

Hazards: example

Suppose that the hazard of the treatment initiation time, T, given the covariate history at time t and given that treatment was not initiated before time t, is piecewise constant as follows:

$$\lambda_T(t) = \begin{cases} \lambda_0^{(0)} & \text{if } L_0 > c_0 \text{ and } t \in [0,1] \\ \lambda_1^{(0)} & \text{if } L_0 \leq c_0 \text{ and } t \in [0,1] \\ \lambda_0^{(1)} & \text{if } L_{1,\infty} > c_1 \text{ and } t \in (1,2] \\ \lambda_1^{(1)} & \text{if } L_{1,\infty} \leq c_1 \text{ and } t \in (1,2], \end{cases}$$

for constants c_0 and c_1 in $\mathbb{R}$. Notice that T depends on $\tilde{L}_0$, $e_{0,\infty}$, and $e_{1,\infty}$, if $\lambda_0^{(0)} \neq \lambda_1^{(0)}$ or $\lambda_0^{(1)} \neq \lambda_1^{(1)}$.

Treatment affects later outcomes, and time-dependent covariates (L_1) which depend on previous treatment also predict future treatment and the outcome of interest.

$\Rightarrow$ Type of setting Structural Nested Models were developed for.

Treatment effect model D: Example

Lok (2017b) shows that for this data generating mechanism,

$$D(y, t; \overline{Z}_t) = -\psi 1_{\text{untreated at } t}.$$

Then, it follows from the definition of X_ψ that

$$X_\psi(t) = Y + \psi(min(T, 2) - t)1_{T>t},$$

where $(min(T, 2) - t)1_{T>t}$ is the duration of the patient not being on treatment between time t and time 2.

Simulation study

In the simulation study, we calibrated the distributions of the variables and the parameter values to HIV/AIDS data, perhaps the most salient example of application of Structural Nested Models in the empirical literature.

We focus on the first two years since HIV diagnosis. Time 0 is the time of HIV diagnosis. The outcome variable is the CD4 count, a commonly used marker of the state of the immune system of HIV-infected patients.

Simulation study

The usual treatment for HIV-infected patients is ART, antiretroviral treatment. ART is not always initiated immediately after diagnosis. ART initiation time often depends on the last measured CD4 count. When the CD4 count is at or below 350 copies/ml, HIV-infected patients are much more likely to initiate ART than when the CD4 count is above 350 copies/ml.

We generated the data as in the parameterization example above.

Lok (2017b) describes how we generated the data for the simulation study in detail, including distributions and parameter values.

Simulations. Mean Squared Errors (MSE) and bias. 5000 repetitions each.

n	setting 1 MSE	$\frac{\text{MSE}\times n}{1000}$	bias	setting 2 MSE	$\frac{\text{MSE}\times n}{1000}$	bias
100	747	75	-0.28	1907	191	-0.040
500	146	73	-0.22	356	178	-0.29
1000	72	72	-0.10	176	176	-0.068
2000	35	70	-0.11	89	179	0.051
5000	14	69	-0.067	35	175	-0.0040
10000	6.6	66	-0.066	18	178	-0.022

Simulations. Mean Squared Errors (MSE) and bias. 5000 repetitions each.

n	setting 3 MSE	$\frac{\text{MSE}\times n}{1000}$	bias
100	2875	287	-0.39
500	542	271	-0.61
1000	268	268	-0.23
2000	138	275	-0.08
5000	54	268	-0.05
10000	27	270	-0.06

Simulations

In this simulation study, both for small and large samples, the bias of the estimators is small. In all three settings and for all sample sizes considered (including the small sample size $n = 100$), the MSE of the estimators arises mostly from the variance, not from the bias.

If the true parameter ψ equals 300 as in this simulation study, for $n = 500$, $\sqrt{MSE}/\psi = 0.04$ in setting 1, and 0.08 in setting 3. Thus, the estimates are already precise in relatively small samples.

The MSE in this simulation study does not depend on the true parameter, $\psi \Rightarrow$ a larger sample size would be required to obtain precise estimators of small true parameter values ψ.

Simulation study

In this simulation study, continuous-time structural nested models perform extremely well.

Coarse Structural Nested Mean Models: proof for optimal estimator

1. Known propensity scores: making the estimator doubly robust decreases the estimator's asymptotic variance.
2. For non-doubly robust estimators, estimating propensity scores may decrease the asymptotic variance of the treatment effect (was also seen in e.g. Robins (1994) and Lok (2008), for different Structural Nested Models).
3. Non-doubly robust estimator with estimated propensity scores has a larger asymptotic variance than the corresponding doubly robust estimator.
4. Doubly robust estimators: estimating propensity scores does not change the asymptotic variance.

Sketch of proof of optimal estimator, continued

5. $\Rightarrow$ The doubly robust estimators are asymptotically preferable, not only because they are doubly robust but also because of their increased precision.
6. Finally, we use Theorem 5.3 from Newey and McFadden (1994) to derive an optimal estimator under the homoscedasticity condition. Homoscedasticity could be replaced by assuming rank preservation as defined in e.g. Robins (1998), see also Lok (2017b).
7. Note: also without homoscedasticity and rank preservation, the resulting estimator is doubly robust, although likely not optimal.

Plugging in a stage 1 estimator

Theorem. Suppose $\tilde{\psi}$ is a preliminary estimator of ψ_* which is the result of unbiased estimating equations $P_n\tilde{G}(\psi) = 0$, $\hat{\theta}$ is an estimator of θ_* from a correctly specified pooled logistic regression model with estimating equations $P_n U(\theta) = 0$, and $E_{\xi^0}\left[H_{\psi_*}(k)|\overline{L}_m, \overline{A}_{m-1} = \overline{0}\right]$ and $q^{opt}_{\psi_*,\xi}$ are parameterized by ξ, which can be estimated using estimating equations $P_n J(\xi, \psi) = 0$ with $EJ(\xi_*, \psi_*) = 0$. Then, under regularity conditions, solving $\hat{\psi}$ from the unbiased estimating equations

with $P_n\left(\begin{array}{cccc} G^*(\psi_1, \psi_2, \xi, \theta) & \tilde{G}(\psi_2) & J(\xi, \psi_2) & U(\theta) \end{array}\right) = 0$

$G^*(\psi_1, \psi_2, \xi, \theta)$

$$= \sum_{m=0}^{K} \sum_{k=(m+1)\vee 12}^{(m+12)\wedge(K+1)} \vec{q}^{\,k,opt}_{m,\psi_2,\xi}(\overline{L}_m)\left(H_{\psi_1}(k) - E_\xi\left[H_{\psi_2}(k)|\overline{L}_m, \overline{A}_{m-1} = \overline{0}\right]\right.$$

$$1_{\overline{A}_{m-1}=\overline{0}}(A_m - p_\theta(m))$$

results in the same asymptotic variance for $\hat{\psi}$ as using the true (but unknown) q^{opt}.

Plugging in a stage 1 estimator: stacked estimating equations

Notice that solving these estimating equations simultaneously leads to the same estimator $\hat{\psi}$ as plugging in $\tilde{\psi}, \hat{\theta}, \hat{\xi}$ into the estimating equations for ψ_* and then solving for $\hat{\psi}$.

Theorem 5.3 from Newey and McFadden (1994): Sufficient optimality criterion with θ_* known: If q^{opt} satisfies

$$E\,\partial/\partial\psi|_{\psi_*}\,G^*\left(\psi,\theta_*,q\right)=E\left(G^*\left(\psi_*,\theta_*,q\right)G^*\left(\psi_*,\theta_*,q^{opt}\right)^\top\right) \tag{6}$$

then no other q satisfying our regularity conditions within this class leads to an estimator for ψ_* with a smaller asymptotic variance than q^{opt}. The estimator resulting from q^{opt} has asymptotic variance equal to the inverse of
$EG^*\left(\psi_*,\theta_*,q^{opt}\right)G^*\left(\psi_*,\theta_*,q^{opt}\right)^\top$. There is a unique (in $L_2(P)$-sense) optimal solution to (6) within this class of estimating equations.

The co-homoscedasticity assumption

When programming the optimal estimator, we assumed:

H is co-homoscedastic: for $k, s = min, \ldots, max$,

$$\mathrm{Cov}\left[H(k), H(s) | \overline{L}_m, \overline{A}_{m-1} = \overline{0}, \text{visit at } m\right]$$

does not depend on $\overline{L}_m$.

For the theory, this is not necessary. For double robustness, this is also not necessary.