# Gov 2002: 13. Dynamic Causal Inference

Matthew Blackwell

December 19, 2015

1. Time-varying treatments

2. Marginal structural models

# **1/** Time-varying treatments

# Time-varying treatments

- Sometimes we want to know the effect of a treatment that varies over time.
- Example: negative advertising
  - Candidate decides whether to go negative based on polling.
  - Going negative affects future polling.
  - Which affects future negativity decisions.
  - Outcome: final voteshare.
  - Should we control for polling?

# Overarching themes

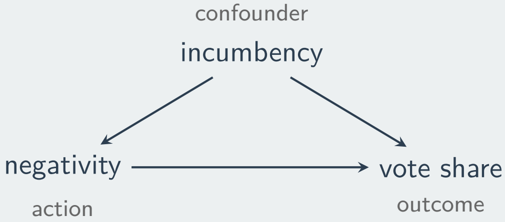
- Many possible effects to estimate!
- Conditioning methods (matching, regression) can't be used in an obvious way.
- Weighting methods will be useful, but highly sensitive.

# Notation

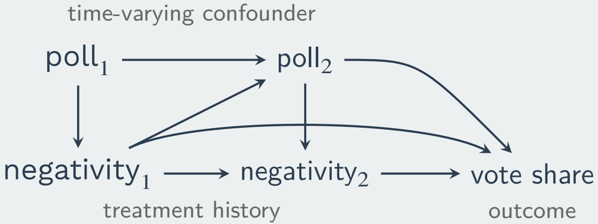
- $\underline{D}_{it} = (D_{i1}, \ldots, D_{it})$ is the partial history of treatment up to time $t$.
- $\underline{X}_{it}$ is the partial history of covariates up to $t$.
- $\underline{d}_t$ and $\underline{x}_t$ refer to specific values these vectors can take.
- $\underline{D}_i = (D_{i1}, \ldots, D_{iT})$, same for $\underline{X}_i$

# Example

- Single-shot causal inference:




- Dynamic causal inference:



# Potential outcomes

- Potential outcomes can be functions of the entire treatment history: $Y_i(\underline{d})$.
- Two-period example: $Y_i(d_1, d_2)$.
- These regimes are static:
    - $\underline{d}$ is a fixed sequence of negative/positive decisions.
    - No reaction to changing environment.

# Treatment regime

**Definition: Treatment regime**

A treatment regime is a mapping, $g(\cdot)$, from the history of time-varying covariates, $\underline{x}$, to a treatment history, $\underline{d}$, that $d_t = g_t(\underline{x}) = g_t(\underline{x}_t)$.

- Treatment regimes are rules that dictate what actions/treatments units should take given a certain covariate history.
- We enforce a no time-traveling assumption.
- Fairly complicated, but exactly the kind of effects we are often interested in.
- Static histories, $\underline{d}$, are (simple) treatment regimes.

# Treatment regimes and potential outcomes

- $Y_i(g)$ is the potential outcome under regime $g$.
- Consistency assumption connects the potential outcomes and the observed data:

$$Y_i = Y_i(g) \quad \text{if } \underline{D}_i = g(\underline{X}_i).$$

- This says that if a unit's observed history is equal to the perscription of the treatment regime, then the observed outcome equals the potential outcome under that regime.

# Estimands

- We would like to estimate the effects of these regimes. Something like the follwing:

$$\tau(g, g') = \mathbb{E}[Y_i(g) - Y_i(g')]$$

- In medical studies, the goal is often to estimate the "optimal" regime, which is the following:

$$g^* = \arg\max_g \mathbb{E}[Y_i(g)],$$

- Here, we are trying to find the regime that maximizes the outcome (assuming the outcome is beneficial).
- Either requires us to estimate the mean of the potential outcome under a given regime. How do we do that?

# Sequential ignorability

- Sequential ignorability will help us identify effects:

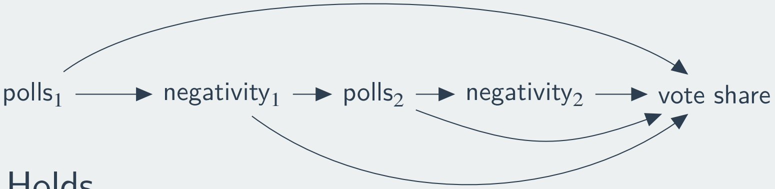
$$Y(g) \perp\!\!\!\perp D_t | \underline{X}_t = \underline{x}_t, \underline{D}_{t-1} = g_{t-1}(\underline{x}_{t-1})$$

- Similar to a sequential experiment, where the randomization can depend on the past.
- Positivity if $\mathbb{P}[\underline{D}_{t-1} = \underline{d}_{t-1}, \underline{X}_t = \underline{x}_t] > 0$, then

$$\mathbb{P}[D_t = d_t | \underline{X}_t = \underline{X}_t, \underline{D}_{t-1} = \underline{d}_{t-1}] > 0$$

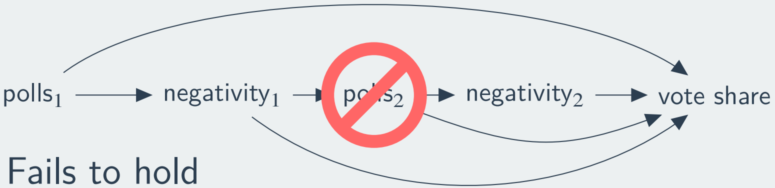
- If a covariate/treatment history is reachable, then any treatment is possible conditonal on that history.

# Sequential ignorability




Holds

No omitted variables in each period




Fails to hold

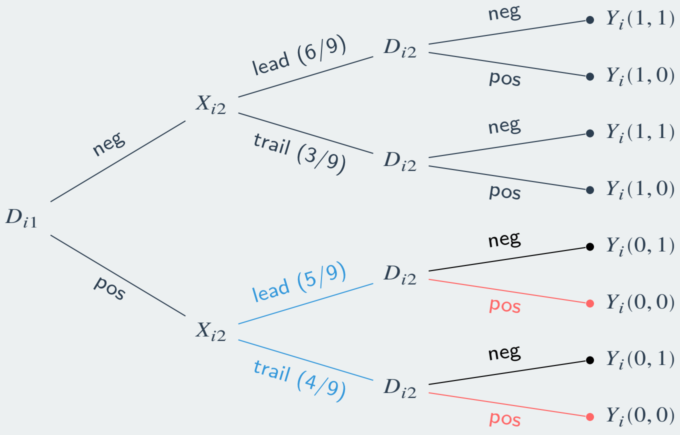
# g-computation

- How to get to marginal mean of $Y_i(g)$ in terms of observed $Y_i$?
- Jamie Robins's g-computational formula:

$$
\begin{aligned}
&E[Y(g)] \\
&= \int_{x_t} \cdots \int_{x_0} E[Y_i | \underline{X}_i = \underline{x}, \underline{D}_i = g(\underline{x})] \prod_{j=0}^{T} \left\{ f(x_j | \underline{X}_{j-1} = \underline{x}_{j-1}, \underline{D}_{j-1} = g_{j-1}(\underline{x}_{j-1})) dx_j \right\}
\end{aligned}
$$

- Right hand side here only has observeable quantities.
  - $E[Y_i | \underline{X}_i = \underline{x}, \underline{D}_i = g(\underline{x})]$ is the mean outcome for people following regime $g$ conditional on the history of covariates.
  - $f(x_j | \underline{X}_{j-1} = \underline{x}_{j-1}, \underline{D}_{j-1} = g(\underline{x}_{j-1}))$ is the density of the covariates at time $j$, conditional on the past.

# g-computation example




$$\begin{aligned}\mathbb{E}[Y_i(0,0)] = &\mathbb{E}[Y_i | D_i = (0,0), X_{i2} = 1] \times \mathbb{P}[X_{i2} = 1 | D_{i1} = 0] \\ &+ \mathbb{E}[Y_i | D_i = (0,0), X_{i2} = 0] \times \mathbb{P}[X_{i2} = 0 | D_{i1} = 0]\end{aligned}$$

# Two-period g-computation

$$\mathbb{E}[Y_i(0,0)] = \mathbb{E}[Y_i|D_i = (0,0), X_{i2} = 1] \times \mathbb{P}[X_{i2} = 1|D_{i1} = 0]$$
$$+ \mathbb{E}[Y_i|D_i = (0,0), X_{i2} = 0] \times \mathbb{P}[X_{i2} = 0|D_{i1} = 0]$$

$$\mathbb{E}[Y_i(1,1)] = \mathbb{E}[Y_i|D_i = (1,1), X_{i2} = 1] \times \mathbb{P}[X_{i2} = 1|D_{i1} = 1]$$
$$+ \mathbb{E}[Y_i|D_i = (1,1), X_{i2} = 0] \times \mathbb{P}[X_{i2} = 0|D_{i1} = 1]$$

- Implies that $\mathbb{E}[Y_i(1,1)] - \mathbb{E}[Y_i(0,0)]$ is not just within strata effects averaged over the distribution of the strata.
- Marginal means must be estimated separately unless first period treatment is the same.

# Complications

- As number of time periods grows or with continuous covariates, stratification becomes infeasible.
- Continuous covariates: requires integrating over their distribution.
- A couple of approaches:
  - Model-based: write down models for outcome, all time-varying covariates and use MLE/Bayesian methods.
  - Structural nested models: reparameterize the likelihood to fix some problems with directly using g-computation.
  - Weighting approach: avoid conditioning on time-varying covariates by weighting them away.

# 2/ Marginal structural models

# Marginal structural models

- Want to deal with time-varying confounders, but we don't want to model them.
- Ideally, we would run a regression-type model and read off coefficients as causal.
- A marginal structural model (MSM) is a model for the marginal mean of the potential outcome for a given treatment history:

$$E[Y_i(\underline{d})] = h(\underline{d}; \beta)$$

- Here $h$ is a link function and $\beta$ are a set of parameters.

# Curse of temporality

- With a binary treatment variable, there are $2^T$ possible treatment histories.
  - $T$ = 2 has 4 possible histories
  - $T$ = 10 has 1,024 possible histories
- Single-shot case ($T = 1$) $\rightsquigarrow$ non-parametrically estimate $E[Y_i(1)]$ and $E[Y_i(0)]$ using simple means.
- Dynamic case ($T = 10$) $\rightsquigarrow$ very few units following any treatment history.
- Need a model to say what features of the treatment history are relevant to the potential outcome.

# Models to reduce dimensionality

- How should we model this? It could be that the number of treated periods is all that matters:

$$E[Y_i(\underline{d})] = \beta_0 + \beta_1 \sum_{t=1}^{T} d_t$$

- Or it could be that the effect varies over time:

$$E[Y_i(\underline{d})] = \beta_0 + \beta_1 \sum_{t=1}^{T/2} d_t + \beta_2 \sum_{t=T/2+1}^{T} d_t$$

- Our model restricts certain treatment histories to have the same mean.
  - Could be wrong!

# Regression/matching?

- Can we use regression or matching to estimate the parameters? Unfortunately not.
- One model conditions on the time-varying confounders (polling):

$$E[Y_i|D_{i1}, D_{i2}, X_{i2}] = \alpha_0 + \alpha_1 D_{i1} + \alpha_2 D_{i2} + \alpha_3 X_{i2}$$

  - This model gets rid of the omitted variable bias for $D_{i2}$ but induces posttreatment bias for $D_{i1}$

- Maybe we omit polls and estimate this model:

$$E[Y_i|D_{i1}, D_{i2}] = \alpha_0 + \alpha_1 D_{i1} + \alpha_2 D_{i2}$$

  - Avoids posttreatment bias, induced omitted variable bias

# Time-varying covariates

- Basic message: time-varying covariates are dangerous when you have:
  - $D_{it}$ and $D_{i,t-1}$ in your regression.
  - A summary of $\underline{D}_{it}$ in your regression.
- TVCs are both pre- and post-treatment in this case.
  - If the effect of negativity early in the race flows through polls and we condition on polls, this is going to underestimate the effect of earlier negativity.
- Similar issue to the intermediate confounders in mediation/direct effects.
- Can avoid the posttreamtent bias via weighting approach.

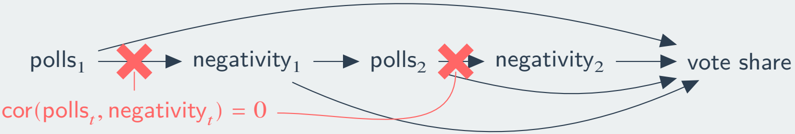
# Ideal, balanced sample

stay positive when leading

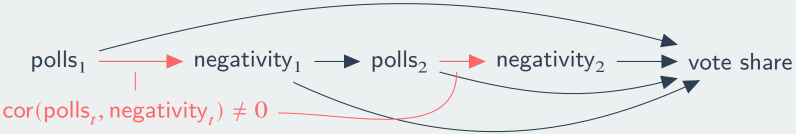
go negative when leading

go negative when trailing

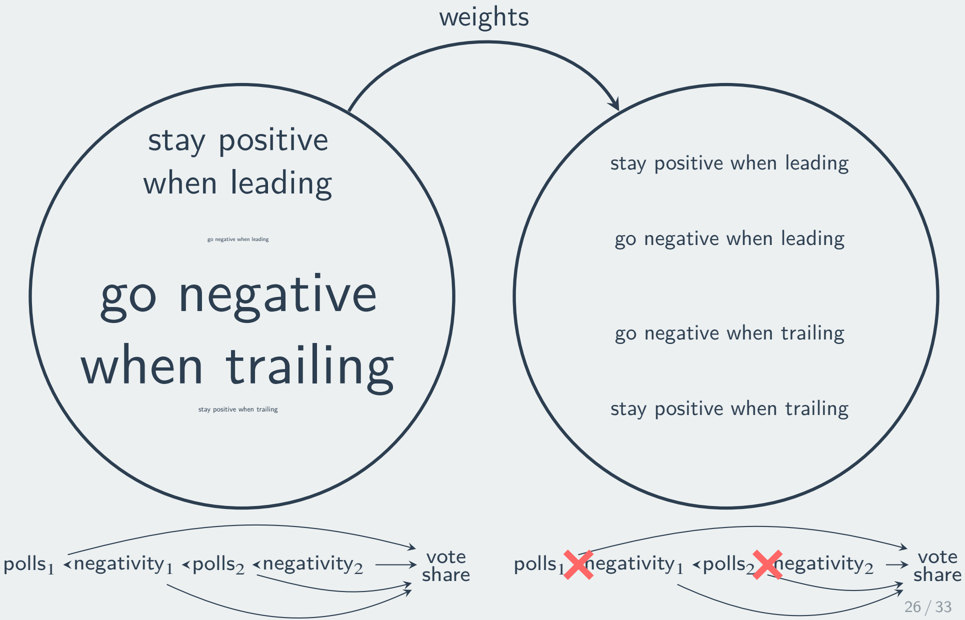
stay positive when trailing



# Messy real world data

# Fixing an imbalance sample



# IPTW with single-shot treatment

- How do we weight?
- With a single-shot treatment, we can weight by the inverse of the propensity score:

$$W_i = \frac{D_i}{\mathbb{P}[D_i = 1|X_i]} + \frac{1 - D_i}{\mathbb{P}[D_i = 1|X_i]}$$

- With a dynamic treatment, the propensity scores are more complicated because the treatment is more complicated.

# IPTW with time-varying treatment

- We have to weight by the probability of observing the entire history: $\mathbb{P}[\underline{D}_i | \underline{X}_i]$.
- Easiest to build up over time using factorization:

$$\begin{aligned} \mathbb{P}(\underline{D}_i | \underline{X}_i) = & \mathbb{P}(D_{i1} | X_{i1}) \times \mathbb{P}(D_{i2} | D_{i1}, \underline{X}_{i2}) \\ & \times \mathbb{P}(D_{i3} | \underline{D}_{i2}, \underline{X}_{i3}) \times \cdots \times \mathbb{P}(D_{iT} | \underline{D}_{i,T-1}, \underline{X}_{iT}) \end{aligned}$$

- Thus, we can create weights like so:

$$W_i = \prod_{t=1}^{T} W_{it}$$

- Here, the weight in period $t$ is the probability of recieving treatment at time $t$ conditional on the past:

$$W_{it} = \frac{1}{\Pr(D_{it} | \underline{D}_{it-1}, \underline{X}_{it})}.$$

# Weights

$$W_i = \prod_{t=1}^{T} W_{it}$$

$$W_{it} = \frac{1}{\Pr(D_{it} | \underline{D}_{it-1}, \underline{X}_{it})}$$

- Weight unit by the probability of receiving the treatment history they did, conditional on the past.
- Two-period example:
  - Race $i$ is positive in the first period, then they are trailing, then they negative later in the race. The weights we would calculate would be:

$$W_i = \frac{1}{\mathbb{P}(\text{pos}_1)} \cdot \frac{1}{\mathbb{P}(\text{neg}_2 | \text{trail}, \text{pos}_1)}.$$

# Why weighting works

- Why does the weighting work?
- Essentially replaces $D_{it}$ with a version that is unaffected by time-varying confouders
- Reweighted $D_{it}$ still has the same effect on $Y_i$
- In reweighted data, $X_{it}$ is no longer a confounder, don't have to control for it.
  - Weighting takes care of omitted variable bias
  - Leave $X_{it}$ out of the outcome model to remove posttreatment bias.

# Weighting to achieve balance

| Race | $\text{polls}_1$ | $\text{negativity}_1$ | vote share |
|---|---|---|---|
| 1 | trailing | negative | 0.45 |
| 2 | trailing | negative | 0.45 |
| 3 | trailing | positive | 0.4 |
| 4 | leading | negative | 0.6 |
| 5 | leading | positive | 0.55 |
| 6 | leading | positive | 0.55 |

trailing & negative > trailing & positive

Weights →

| Race | $\text{polls}_1$ | $\text{negativity}_1$ | vote share |
|---|---|---|---|
| 1 | trailing | negative | 0.45 |
| 1 | trailing | negative | 0.45 |
| 1 | trailing | negative | 0.45 |
| 2 | trailing | negative | 0.45 |
| 2 | trailing | negative | 0.45 |
| 2 | trailing | negative | 0.45 |
| 3 | trailing | positive | 0.4 |
| 3 | trailing | positive | 0.4 |
| 3 | trailing | positive | 0.4 |
| 3 | trailing | positive | 0.4 |
| 3 | trailing | positive | 0.4 |
| 3 | trailing | positive | 0.4 |
| 4 | leading | negative | 0.6 |
| 4 | leading | negative | 0.6 |
| 4 | leading | negative | 0.6 |
| 4 | leading | negative | 0.6 |
| 4 | leading | negative | 0.6 |
| 4 | leading | negative | 0.6 |
| 5 | leading | positive | 0.55 |
| 5 | leading | positive | 0.55 |
| 5 | leading | positive | 0.55 |
| 6 | leading | positive | 0.55 |
| 6 | leading | positive | 0.55 |
| 6 | leading | positive | 0.55 |

trailing & negative = trailing & positive

# Estimating the weights

- Need to estimate $\mathbb{P}[D_{it}|\underline{D}_{i,t-1}, \underline{X}_{it}]$ for all time periods
  - Easy, but not robust: logit models.
  - More complicated, but robust: Covariate Balancing Propensity Score (Imai, Ratkovic)
- Hard to include all past treatments, confoudners. Possible strategies:
  - Last period confounders, perhaps a few lags for important confounders.
  - Use GAMs to smooth functional form for key variables.
  - Last few lags of treatment and/or a summary of cumulative treatment.
  - Time trend in $t$.
- Use the cumulative product of predicted values from this model to get weights.

# Outcome MSM

$$
\begin{aligned}
E[Y_i(\underline{d})] =& \beta_0 + \beta_1 \left( \sum_{T-5}^{T} D_{it} \right) + \beta_2 Z_i \\
& + \beta_3 Z_i \left( \sum_{T-5}^{T} D_{it} \right) + \beta_4 \underline{D}_{iT-6} \\
& + \beta_5 Z_i \underline{D}_{iT-6} + \beta_6 X_i
\end{aligned}
$$

- Model from Blackwell (2014)
  - Allows for different effects of negativity early and late in campaign.
  - Interaction with baseline covariate, $Z_i$, incumbency status.
  - Controls for other baseline covariates.
- Baseline covariates are ok to include—never posttreatment.
- Bootstrap for SEs.