Causal Inference in Panel Data

Matthew Blackwell • Department of Government, Harvard University • blackwel@fas.harvard.edu

The Problem.

How do analyze causal effects when treatment varies over time? What if treatment responds to changes in time-dependent covariates?

(e.g.)

If we want to understand the effect of negative campaign advertising on election outcomes, what do we do about polls in the middle of the campaign?

Early Tone → Early Polls → Late Tone → Election

Polls are both a ***pre-treatment confounder*** for late negativity and a ***post-treatment consequence*** of early negativity. Standard statistical advice tells us to both include and exclude polls from our analysis.

- Can we identify causal effects in this environment?
- Can we identify the effects of ***conditional, dynamic*** treatments?

The (hypothetical) Data.

A simple example of the problem. Suppose the DNC provides us data for each of the 435 House races on the Democratic candidate's tone at different points in time, along with polling data for each race. Specifically:

Early Tone	X_1	Indicator of negative tone early in the campaign
Early Polls	Z_1	Indicator of leading (1) or trailing (0) status
Late Tone	X_2	Indicator of negative tone late in the campaign
Outcome	Y	Indicator of Democrat winning election

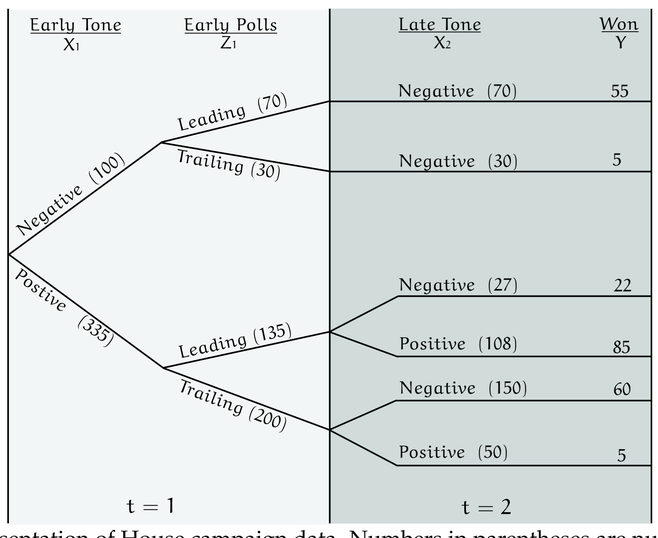


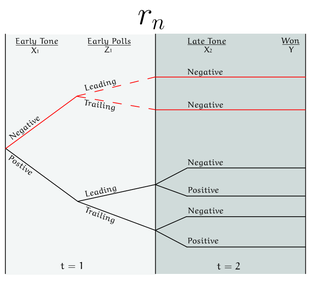
Figure 1: Tree representation of House campaign data. Numbers in parentheses are number of Democrats that follow that branch of the tree.

The Concepts.

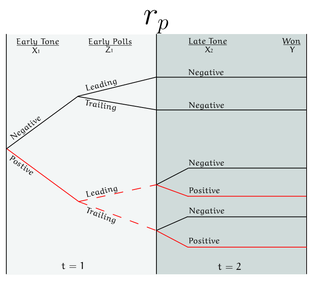
A *treatment regime*, r, is a rule for assigning treatment and it generalizes treatments to the time-series realm. For instance, regimes might be:

$$r_n = \{\text{stay negative always}\} = \{X_t = 1; \forall t\}$$
$$r_p = \{\text{stay positive always}\} = \{X_t = 0; \forall t\}$$
$$r_u = \{\text{stay positive unless trailing}\} = \{X_1 = 0, X_2 = 1 - Z_1\}$$

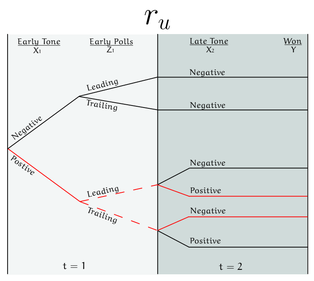
r_n



r_p



r_u



Each unit has a *potential outcome* under a given treatment regime.

$Y_i(r_p)$ = election outcome if i were to stay positive always
$Y_i(r_p)$ = election outcome if i were to stay negative always
$Y_i(r_u)$ = election outcome if i were to stay positive unless trailing

We also might want to identify *regime-observed subpopulations*:

$[Y|r_p]$ = outcomes among those observed to always be positive
$[Y|r_n]$ = outcomes among those observed to always be negative
$[Y|r_u]$ = outcomes among those observed to either {positive, trailing, negative} or {positive, leading, positive}

The Assumptions.

Assumption 1 (Consistency). *For any treatment regime, observed outcomes are equal to the potential outcome under the treatment regime actually observed. Formally, if unit i has a treatment history consistent with regime r, then $Y_i = Y_i(r)$.*

Assumption 2 (Sequential Ignorability). *For any treatment regime r, time t, treatment assignment is independent of the potential outcome conditional on observed information available at t. Formally,*

$$Y(r) \perp\!\!\!\perp X_t | (Z_{t-1}, ..., Z_1), (X_{t-1}, ..., X_1), \forall t.$$

These extend SUTVA and ignorability to the time-varying context.

The Analysis.

How do we compute the expected value of a potential outcome?

Omission Estimator:

$$\hat{\mathbb{E}}[Y(r_p)]_{omi} = \mathbb{E}[Y \mid r_p]$$

Naive Regression Estimator:

$$\begin{aligned}\hat{\mathbb{E}}[Y(r_p)]_{reg} = &\mathbb{E}[\,Y \mid r_p,\ \text{leading}\,] \cdot \mathbb{P}[\,\text{leading}\,] \\ &+ \mathbb{E}[\,Y \mid r_p,\ \text{trailing}\,] \cdot \mathbb{P}[\,\text{trailing}\,]\end{aligned}$$

g-estimator:

$$\begin{aligned}\hat{\mathbb{E}}[Y(r_p)]_{g} = &\mathbb{E}[\,Y \mid r_p,\ \text{leading}\,] \cdot \mathbb{P}[\,\text{leading} \mid r_p] \\ &+ \mathbb{E}[\,Y \mid r_p,\ \text{trailing}\,] \cdot \mathbb{P}[\,\text{trailing} \mid r_p]\end{aligned}$$

The Omission estimator simply collapses over poll status. The naive regression estimator directly adjusts for polls, but assumes that polls are unaffected by early tone. The g-estimator adjusts for both the confounding of polls on late tone and the confounding of early tone on polls.

$$\hat{\mathbb{E}}[Y(r_p)]_{omi} = 90/158 = 0.57$$
$$\hat{\mathbb{E}}[Y(r_p)]_{reg} = (85/108) \cdot (205/435) + (5/50) \cdot (230/435) = 0.42$$
$$\hat{\mathbb{E}}[Y(r_p)]_{g} = (85/108) \cdot (135/335) + (5/50) \cdot (200/335) = 0.33$$

We can also compute the expected outcome for time-dependent regimes, like the poll-watching regime, r_u:

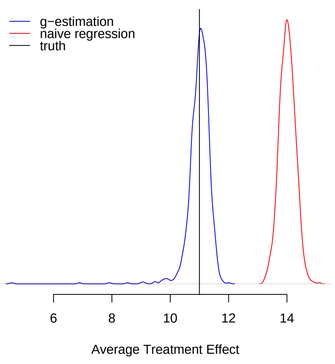
$$\hat{\mathbb{E}}[Y(r_u)]_{g} = (85/108) \cdot (135/335) + (60/150) \cdot (200/335) = 0.556$$

Thus, the causal effect of moving from the static always positive regime to the dynamic poll-watching regime is 0.556-0.33 = 0.226: a greater than 20% increase in the probability of winning the election.

The Simulations.

I generated a two-period time-series Monte Carlo study with the following specifications:

$$\begin{aligned} Z_1 &\sim \mathcal{N}(0,1) & & \\ X_1 &\sim \text{Bern}(p_1) & p_1 &= \Phi(Z_1) \\ Z_2 &\sim \mathcal{N}(\mu_2,1) & \mu_2 &= Z_1 + 2X_1 + 2X_1Z_1 \\ X_2 &\sim \text{Bern}(p_2) & p_2 &= \Phi(Z_1 + Z_2 + X_1 - .9X_1Z_2) \\ Y &\sim \mathcal{N}(\mu_y,1) & \mu_y &= 2X_1 + 5X_2 + 2X_2Z_2 \end{aligned}$$



I simulated 1000 datasets and computed two estimates of the effect of going from $(X_1 = 0, X_2 = 0)$ to $(X_1 = 1, X_2 = 1)$. One is the naive regression estimator that simply adjusts for Z_1 and Z_2. The other is the g-estimator. The root mean squared errors (RMSE) were

$$RMSE_{reg} = 3.03$$
$$RMSE_{g} = 0.42$$

The Discussion.

The g-estimation model was developed by Robins (1986) to handle sequential randomized trials and observational studies. The key to the analysis is controlling for time-dependent covariates, but controlling for them in a way that is consistent with the treatment regimes.

Potential Application:

- Campaign effects (going negative, spending, campaign stops)
- Economic interventions (interest rates, public funding)

Drawbacks:

- Identifying assumptions may be hard to meet as they require ignorability at every time period and no interference between units.
- We treat each time series as a unit, which leads to small sample sizes.
- With large numbers of time periods, as in political science, parametric models are required for treatment and outcome.