Bayesian Causal Inference in Political Science

Michael G. DeCrescenzo*

Updated March 14, 2019

Abstract

Causal inference and Bayesian analysis are two powerful and attractive methodological approaches for conducting empirical research, but almost never in political science does a single study employ both approaches. This stands in contrast to other fields—such as psychology, epidemiology, and biostatistics—where Bayesian analysis and causal inference methods interact regularly. In this paper I argue that the partition between these methodological schools in political science has no inherent basis in their fundamental goals, which are actually quite compatible: generating the best parameter estimates. In fact, Bayesian analysis provides a number of distinct advantages for improving causal inference designs. The methodological partition instead owes itself to informal norms surrounding each school in empirical political science. I discuss these sources of normative tension, discuss go-to practices doing Bayesian inference for the "skeptical causal inference audience," and demonstrate these practices using real examples from recent political science work. The purpose of the paper is *not* to convince all causal inference practitioners to adopt Bayesian estimation. The purpose is to show that Bayesian methods deserve space in the study of causal effects because they improve causal estimates and provide an appealing framework for evaluating causal evidence.

Keywords: experiments, causal inference, Bayesian statistics

*Ph.D. Candidate, Political Science, University of Wisconsin–Madison. Thanks to Matthew Blackwell, Barry Burden, William Christiansen, Andrew Heiss, Devin Judge-Lord, Michael Masterson, Anna Meier, Laura Meinzen-Dick, Evan Morier, Daniel Putman, Alex Tahk, Zach Warner, Chagai Weiss, and Dave Weimer for feedback and advice.

Contents

1 Introduction

Reader ability?

1.1 Causal Modeling with Potential Outcomes

Throughout this paper, I refer to "causal inference" as estimating the parameters of explicitly stated causal models whose assumptions allow researchers to interpret parameters as causal effects. In political science, these causal models typically flow from a Neyman-Rubin potential outcomes framework, but these are not the only models employed in causal inference. This section briefly reviews the intuition of the Neyman-Rubin model with notation that will be used throughout the paper.

cites?

Suppose we are interested in the causal effect of Z on Y, where $Z = 1$ indicates a "treatment" and $Z = 0$ indicates a "control." Let $Y_i(Z)$ represent potential outcomes for individual i given a value of Z, with $Y_i(Z = 1)$ being the potential outcome under treatment and $Y_i(Z = 0)$ being the potential outcome under control. We define the treatment effect T_i as the treatment effect on i, which is the difference between their potential outcomes: $T_i \equiv Y_i(1) - Y_i(0)$.

Causal inference is a body of methods for learning about T_i given that it is directly observed in real-world data. If both $Y_i(1)$ and $Y_i(0)$ were observed, calculating T_i would be trivial, but in any applied case an observation is assigned to only one treatment level, preventing the researcher from directly comparing i's potential outcomes. The inability to observe both multiple potential outcomes for the same observation is commonly called the "fundamental problem of causal inference," and it prevents the researcher from making inferences about the value of T_i (Holland 1986) without assumptions. Causal inference methods define the assumptions and estimators that enable researchers to estimate treatment effects *on average* from groups of individuals who receive only one treatment at a time. The average treatment effect (ATE) is the mean of the true treatment effects across all i. If we observed all potential outcomes for all N observations in our population of interest, we could calculate the ATE precisely.

cite

$$\bar{T} \equiv \mathrm{E}\left[Y_i(1) - Y_i(0)\right] \tag{1}$$

$$\equiv \frac{1}{N}\sum_{i=1}^{N}\left[Y_i(1) - Y_i(0)\right] \tag{2}$$

Because we don't observe all potential outcomes, we require additional assumptions.

cleaner

Causal inference methods are particularly attentive to the specific mechanism that assigns units into treatment and control groups, as it's the researcher's model of the assignment mechanism that determines which causal effects can be identified from the study design (Rubin 1991). Most simply, causal effects can be identified if units are assigned to treatment groups by a mechanism that it uncorrelated with their potential outcomes. The most direct way to remove confounding between treatment status and potential outcomes is for the researcher to assign units randomly to treatment conditions. Under random assignment (and assumptions of "no interference" between units' treatment assignments or potential outcomes), the simple difference between treatment group mean ($\bar{y}_1$) and the control group mean ($\bar{y}_0$) is an unbiased estimator of the ATE on the outcome scale.

$$\bar{T} = \mathrm{E}\left[\bar{y}_1 - \bar{y}_0\right] \tag{3}$$

In situations where assignment is not random throughout the entire sample, additional assumptions allow researchers to identify allow *local* average treatment effects (confined to specific regions of the sample), *conditional* average treatment effects (controlling for other observed covariates), direct and indirect effects when mediating variables are observed, and even *average marginal component effects* that marginalize treatment effects across the values of other simultaneous treatments.

cites

The advantages to causal modeling are numerous and apparent. While it is commonly known that conventional regression modeling "merely" estimates a conditional mean with no necessary causal interpretation, regression coefficients are routinely discussed with causal language. This is despite the fact that conventional regression is liable to misspecification or other oversights in translating a causal diagram into a regression specification (see e.g. Keele 2015: or Samii (2016)). By identifying specific causal estimands, researchers are restricted by the nature of their research design about the language they can use to describe their estimates. Causal modeling and identification analysis introduce more difficulty into the research design process, but payoff is estimates that more closely adhere to the causal parameters at interest in the researcher's theoretical model of the world. Assuming that the researcher's quantity of interest is the causal parameter itself,[1] estimates from unconfounded research designs provide much more information about causal parameters than estimates from designs with an unknown (but presumably larger) potential for confounding (Gerber et al. 2014).

Applications?

[1] This assumption will be crucial for the remainder of the paper

1.2 Bayesian Inference

modeling?

Bayesian inference is often described using language that obscures its philosophical and practical appeal, making it "feel" incompatible with the principles and norms of causal inference methods. In this short overview of Bayesian inference, I hope to clarify its more difficult intuitions by using terminology that will make its compatibility with causal inference more apparent.

notation

First, we will define Bayesian inference using language from Gelman et al. (2013: p. 1). Bayesian inference is fitting a statistical model and analyzing the probability distribution of its unknown parameters. That is, which parameters are *likely*, and which are *unlikely*, based on the observed data? In a causal inference application, the parameter of interest is the ATE or a related causal estimand.

Mechanically, Bayesian models work by specifying a joint probability distribution over all variables in the model, where a "variable" could be observed (data, signified with y) or unobserved (parameters, signified θ).[2] This joint distribution includes a prior distribution over possible parameter values $p(\theta)$ and a distribution of observed data that depend on the parameter values, $p(y \mid \theta)$. The joint distribution could be written as

Do I?

Betancourt?

$$p(\theta, y) = p(\theta)p(y \mid \theta). \tag{4}$$

The prior $p(\theta)$ inevitably contains parameters that would be unsupported by data if an infinite amount of data were collected. Researchers learn which parameters are supported by data by conditioning the joint model on the data using Bayes' Theorem.

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{p(y)} \tag{5}$$

Conditioning on the data allows the researcher to "update" the distribution of parameters, paring down which values are more or less supported by the data. This updated distribution, the posterior distribution, is the distribution of parameter values that are most compatible with both the initial model and the data. The exact shape of the posterior is determined by the specificity of the prior and the strength of the signal from the data. When the data send a

[2] Specifying a probability distribution for every variable in the model "merely" applies the same distributional intuition to unobserved parameters as is commonly applied to observed data (McElreath 2015). I elaborate on this more below.

vague signal, the posterior relies more heavily on the prior distribution $p(\theta)$. When the data send a precise signal about likely and unlikely parameters, the posterior more heavily reflects the data, $p(y \mid \theta)$. When the researcher observes enough data, the posterior distribution converges toward $p(\theta \mid y)$ and the prior approaches irrelevance.[3] Crucially, the intuition of Bayesian updating is agnostic to whether the original model has a causal interpretation; if any model has parameters that can be conditioned on data, then posterior parameter inference is possible.

The posterior distribution is the key philosophical payoff of Bayesian inference. It allows the researcher to make direct inferences about parameter values by evaluating their joint probability distribution. This stands in contrast to non-Bayesian methods where inference is performed indirectly, by fixing the value of unknown parameters (often at values that don't represent competing theories about causal effects) and calculating the probability of observing "more extreme" data under those assumed parameter values. A strict interpretation is that typical null hypothesis significance testing (NHST) says "these data are unlikely, given a model that I don't believe," while Bayesian inference says, "these parameters are likely, given data I have actually observed."

Overture to some clarifications? Posterior probability vs NHST. Random variable, likelihood and prior (it's likelihoods that are special cases of priors, not priors that are special or ad hoc). Information vs. belief

For parameters to have probability distributions, Bayesian statistics regards them as "random" variables. This terminology can be confusing, especially from a non-Bayesian perspective where the true parameter value is "fixed" but unknown. A more intuitive way to describe Bayesian parameters is to say that we don't (or can't) know the precise value of the true parameter—and indeed there could be one "true" value—but the information we have about the true value is characterized by a probability distribution. The parameter isn't random in the sense that it is fluctuating between ever-changing quantum states. Rather, it is random in the sense that it is an instantiation of the underlying process that produced the parameter. Because the precise value of the parameter is unknowable with a finite amount of data, as researchers we strive to obtain as specific information about that process as we can by revising our information about the parameter's probability distribution.

[3] As long as the prior does not assign zero probability to important regions of parameter space.

2 Shared Goals, Different Tactics

Do causal inference and Bayesian analysis "go together" in political science? I believe that they can, and the reason why is that both schools are fundamentally interested in the same thing: getting the best, most reliable parameter estimates that we can get, given the data that we have. They take different tactics

Before unpacking this further, it's worth asking whether this goal is unique to these schools, or if perhaps *every* statistical method has parameter estimation as its ultimate goal. My answer to that is a firm No. I'll elaborate using the prevailing statistical inference paradigm in political science: null hypothesis significance testing. 1st ref?

The objective of null hypothesis significance testing ("NHST") is to infer that research estimates are "statistically significant.". The researcher assumes a null hypothesis, calculates a test statistic, and compares the statistic to its sampling distribution to determine if the statistic is past some threshold of unlikeliness ($p < \alpha$) to reject the null hypothesis. Although NHST is *an* approach for making inferences about parameters, its true function is not to make inferences about what parameters actually are. Rather, its function is to judge, indirectly, that an assumed value of the parameter is unlikely. This is done without prejudice to what the pts true value of the parameter might be. Although researchers commonly interpret confidence intervals as best-guesses about where a parameter is likely to be, the mechanics of confidence intervals only allow the researcher to say that the interval contains parameter values that cannot be rejected at the requisite α level. Furthermore the logic of NHST is often misused fix? to make substantive inferences about *null* findings from the researcher's failure to reject the null, despite the widespread understanding that this isn't something NHST is equipped to do. It is worth highlighting the premise that causal inference and Bayesian analysis are both interested in *actual parameter values* because it stands in contrast to the statistical approach used almost all empirical political science research.

To answer this question, we should ask ourselves which elements of each method are in tension with each other and which are in harmony. I argue that

the underlying goals of both Bayesian analysis and causal inference are fundamentally compatible

Isn't this the goal of all methods. To be crystal clear: no.

- not NHST
- not RI

Use Gerber, Green, and Kaplan to justify the parameter intuition

- "First, under what conditions and to what extent should we update our prior beliefs based on experimental and observational findings?" (252)
- First statement of their project is "Suppose you seek to estimate the causal parameter M" (253)
- "In advance of gathering the data, you hold prior beliefs about the possible values of M." (253)
- *a strong interpretation of the GGK project is that learning about parameters is only made possible by specifying a prior and obtaining a posterior*
- A key citation in justifying experimental research relies on assumptions that we are learning about the posterior distribution of parameters
    - *it's worth pointing out for the record that the assumptions required to make observational research are ludicrous, but whatever*

But???

- in the case where you have no prior about true mean or bias, the posterior is equal to an experimental estimate (you only learn from the experiment because you're infinite variance over the bias?)
- "When researchers lack prior information about the biases associated with observational research, they will The illusion of learning from observation research 253 assign observational findings zero weight and will never allocate future resources to it."
    - *this is a fucking abuse of Bayesian intuition*

Forceful argument

- The person who made these models often uses the Bayesian version because it makes more inferential sense (Rubin)
- Screeds against observational work in political science require Bayesian interpretation in order to for the idea of "learning from experiments" to be theoretically tractable, so you're fucking welcome

2.1 Potential Outcomes with Bayesian Language

This section demonstrates the mechanics of causal modeling under a Bayesian conceptual framework. It owes the key intuition to Rubin (1978) with some abuse of notation.

A key feature of Bayesian analysis is that unknown quantities are represented with probability distributions; we lack exact knowledge about their values, and the probability distribution represents our state of uncertainty. Functions of unknown variables, in turn, reflect uncertainty about the unknown parameters that compose it. This is relevant to causal inference because it changes our interpretation of the potential outcomes. An unobserved potential outcome $\tilde{y}_i$ in a Bayesian framework is represented by a *distribution* of unknown potential outcomes that reflects the posterior distribution of unknown parameters that create potential outcomes. We therefore regard the treatment effect τ_i by averaging over all unknown potential outcomes.

Suppose that $p(\tilde{y})$ is the *prior* predictive distribution of unobserved potential outcomes. We conduct causal inference by determining the posterior distribution of unobserved potential outcomes, that is by conditioning this distribution on the information conveyed by observed data y. This conditioning implies that we update model parameters θ in the process.

$$p(\tilde{y} \mid y) = \int p(\tilde{y}, \theta \mid y) d\theta. \tag{6}$$

Given our observed data y, the posterior predictive distribution of $\tilde{y}$ is the joint distribution of $\tilde{y}$ and the posterior θ, averaging across our posterior uncertainty about θ.

is this integral gibberish?

To fill out the modeling intuition, consider an experiment where individuals are randomly assigned to values of $z \in \{0, 1\}$. The difference of means is commonly estimated using a regression form…

$$y_i = \mu_0 + \tau z_i + \varepsilon_i, \tag{7}$$

where μ_0 is the control group mean, τ is the treatment effect of moving from $z = 0$ to $z = 1$, and ε_i is response-level error. For Bayesian estimation, we would specify the full probability model for all observed data and unobserved parameters. First, the response data are given a model implied by the parametric assumption of ε_i.

$$y_i \sim \text{Normal}\left(\mu_0 + \tau z_i, \sigma_{z[i]}\right) \tag{8}$$

This example assumes unequal variance across levels of z. We then include priors for the

model parameters.

$$\mu_0 \sim p(\mu_0) \tag{9}$$

$$\tau \sim p(\tau) \tag{10}$$

$$\sigma_z \sim p(\sigma_z) \tag{11}$$

The process of Bayesian estimation then conditions the parameters on the data to obtain updated distributions for μ_0, τ, and σ. If the priors are (improper) flat priors, then the posterior distribution for the parameters will be proportional to the shape of the likelihood.

One practical consideration for Bayesian estimation is the parametric form of the regression. By estimating a treatment effect as the coefficient on a treatment indicator, it is very difficult to place equal prior uncertainty on both treatment group means. For this reason, it is more straightforward to write the conditional mean of y_i explicitly as the mean in each treatment group,

$$y_i \sim \text{Normal}\left(\mu_{z[i]}, \sigma\right). \tag{12}$$

Under this setup, the average treatment effect is $\tau = \mu_{z=1} - \mu_{z=0}$. From that starting point, it is straightforward to place equal priors on both treatment groups without affecting the behavior of the models for y_i and σ_j.

$$\mu_z \sim p(\boldsymbol{\mu}) \tag{13}$$

3 Epistemic Norms and Modeling Assumptions

Skepticism in different forms

- confounding
- model assumptions?
- "impossible" parameters

Bayes:

- measurement
- sticking to "hard problems"
- the demand for "value added" means that the benefits for "normal science" are swept aside (wrongly)

4 Practical Bayes for Skeptical Causal Inference

Another difficult concept in Bayesian inference is the role of *belief*. Priors are often described as a researcher's "beliefs" about the parameter before collecting data. While priors may be correlated with researchers' beliefs, equating the two is misleading. Priors are better thought of as explicit assumptions about plausible and implausible parameter values. They are mostly a way to guide estimated parameters to regions of parameter space that are consistent with plausible data that are possible to observe, depending on the parameterization of the model and the scale of the variables.[4]

in the norms section?

start with as little prior information as possible and work up

4.1 Philosophical Benefits even with Flat Priors Are Still Priors

What are the advantages of Bayesian view even if you kept priors flat and unrestricted?

There are distinct philosophical and practical benefits even with flat priors

- posterior inference
- proliferated and joint uncertainty
- computational convenience (in the sense of numerical approximation as an end-around analytic derivation) for functions of parameters, model evaluation
- multiple comparisons is second-nature

4.2 Weak prior information and predictive checks

Logistic regression presents a great example. Logit coefficients are typically thought to be normally distributed on the logit scale.

4.3 Regularization

- Informed priors *downweight* the probability of extreme effects and *upweight* the probability of small effects
- Flat priors *upweight* extreme effects

[4] A common example is logistic regression coefficients to predict a binary outcome. For a control group with a predicted probability of 50%, a treatment effect of 3.0 on the log odds scale is an effect of more than 40 percentage points. For most social science studies, a prior in a logistic regression can safely assume that coefficients of 3 or larger are nearly impossible.

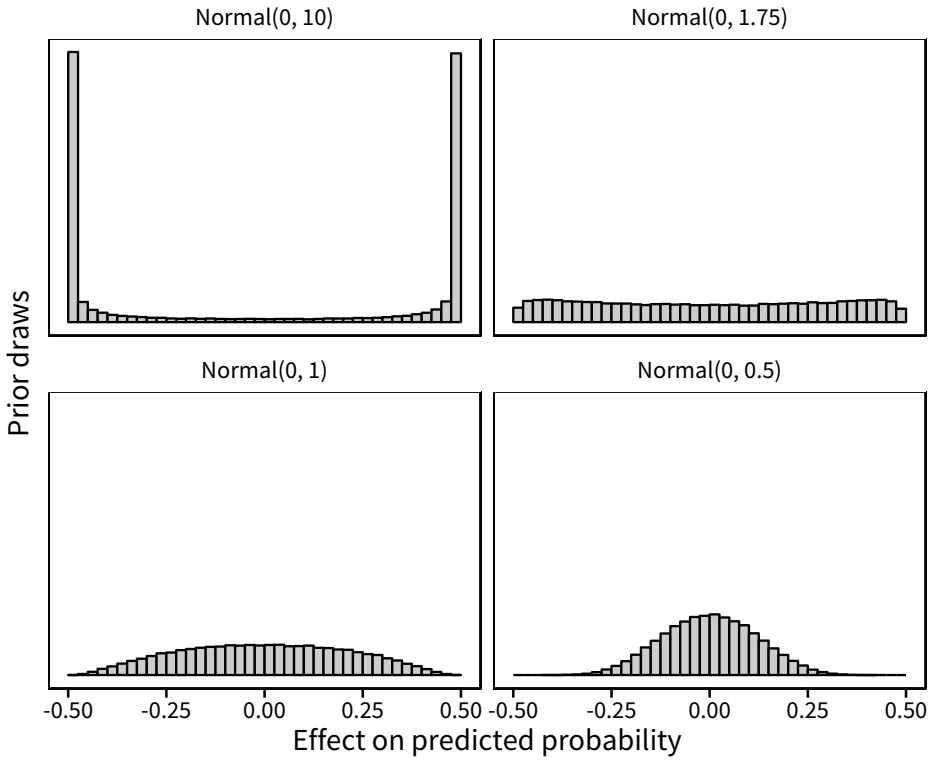


Figure 1: Priors simulations of logit coefficients on the probability scale

Fixed effects and OLS models are special cases of multilevel models, where subgroup information is either ignored completely ($\sigma^{\text{group}} = 0$) or estimated without any memory of other groups ($\sigma^{\text{group}} = \infty$).

5 Demonstration: Regression Discontinuity using Weakly Informative Priors

param

Hall's "local linear" specification is presented as…

$$\begin{aligned} y_{dpt} = \beta_0 + \beta_1 &Extremist\ Primary\ Win_{dpt} + \beta_3 Extremist\ Primary\ Margin_{dpt} \\ &+ \beta_3 \left(Extremist\ Win_{dpt} \times Extremist\ Margin_{dpt} \right) + \varepsilon_{dpt} \end{aligned} \tag{14}$$

where…

We start by re-writing the equation by indexing parameters by treatment status. The traditional interactive form makes it difficult to set priors. Let $w \in \{0, 1\}$ be the treatment status of district d for party p in year t, which equals 1 if the extremist candidate wins the primary. We index the parameters α_w and β_w to be constants and the effect of the running variable for control and treated units.

explain

$$y_{dpt} = \alpha_{w[dpt]} + \beta_{w[dpt]} Extremist\ Primary\ Margin_{dpt} + \varepsilon_{dpt} \tag{15}$$

The intuition of this equation form is that if the extremist candidate lost the primary in unit dpt, the parameters in the equation are α_0 and β_0, whereas the parameters are α_1 and β_1 if the extremist won the primary. The treatment effect at the discontinuity can be found by calculating the difference in the intercepts between extremist primary winners and losers: $\alpha_1 - \alpha_0$.

5.1 Prior Strategies

Goals for priors: Parameters only allow y values in the range of the data.

- No α_w value outside of $(0, 1)$ should be possible *a priori*.
- No β_w value should lead us to predict values of y outside of $(0, 1)$. This means that the bounds of β_w are themselves a function of constants α_w and the extremist primary margin for a given observation.

Let M_{dpt} represent the extremist's primary margin in district dpt. Formally stated, if predictions for y_{dpt} are defined by $\alpha_w + \beta_w M_{dpt}$, we want a prior for β such that $0 < \alpha_w + \beta_w M_{dpt} < 1$, subject to $\alpha \in (0, 1)$.

Practically, we can calculate these bounds by using the bandwidth of M_{dpt} around the threshold. Let the bandwidth be represented by ω. When the primary candidate wins, the farthest value from the threshold that M_{dpt} could take is $M_{dpt} = \omega$, and when the primary candidate loses, the farthest value from the threshold is $M_{dpt} = -\omega$. In order to constrain the slopes of the regression line, $\pm\omega$ is the farthest from the threshold that we need to consider.

When the extremist primary candidate loses, the regression line must satisfy

$$\alpha_0 + (-\omega)\beta_0 > 0 \tag{16}$$

$$\alpha_0 + (-\omega)\beta_0 < 1. \tag{17}$$

When the extremist primary wins, the slope must satisfy

$$\alpha_1 + \omega\beta_1 > 0 \tag{18}$$

$$\alpha_1 + \omega\beta_1 < 1. \tag{19}$$

Solving these inequalities for β_0 and β_1, respectively, we find that the following conditions must hold.

$$\frac{-\alpha_0}{-\omega} < \beta_0 < \frac{1 - \alpha_0}{-\omega} \tag{20}$$

$$\frac{-\alpha_1}{\omega} < \beta_1 < \frac{1 - \alpha_1}{\omega} \tag{21}$$

5.2 *Uniform bounded priors*

For notational convenience, we define x_w^* that equals $\min\left(M_{dpt}\right)$ when $w = 0$ and $\max\left(M_{dpt}\right)$ when $w = 1$.

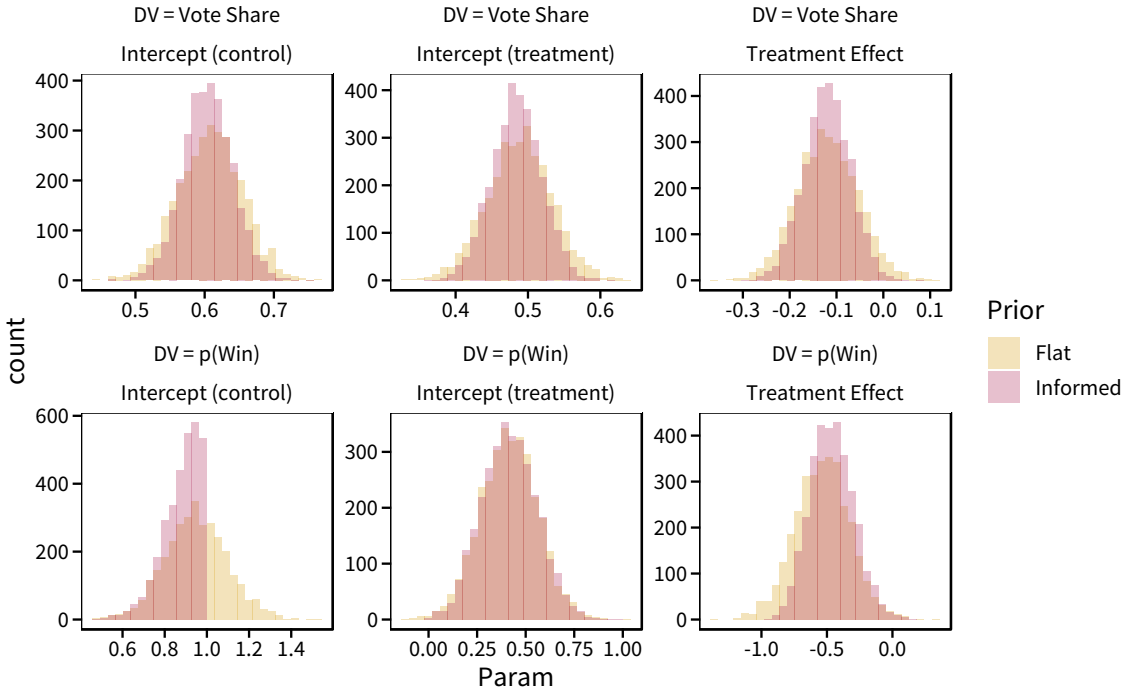
$$y_{dpt} \sim \text{Normal}\left(\mu_{dpt}, \sigma\right) \tag{22}$$

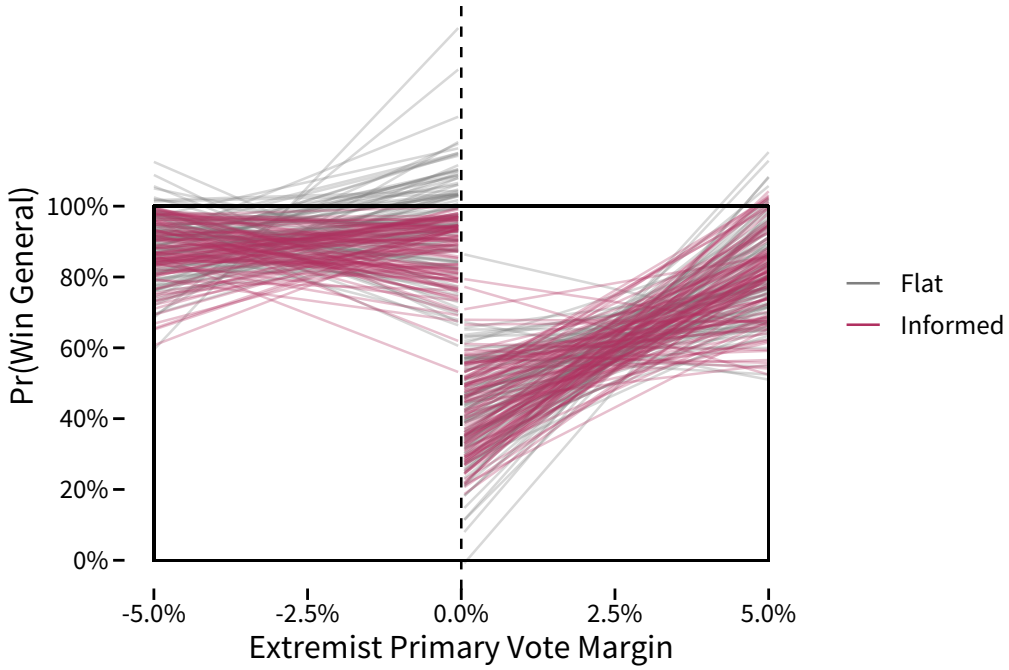
$$\mu_{dpt} = \alpha_{w[dpt]} + \beta_{w[dpt]} M_{dpt} \tag{23}$$

$$\alpha_w \sim \text{Unif}\left(0, 1\right) \tag{24}$$

$$\beta_w \sim \text{Unif}\left(\frac{-\alpha_w}{x_k^*}, \frac{1-\alpha_w}{x_k^*}\right) \tag{25}$$

$$\sigma \sim \mathcal{H}(\cdot) \tag{26}$$





5.3 Logistic Modeling with Priors

The idea:

- the logistic model probably weakens the evidence for a treatment effect.
- BUT logistic model combined with sensible priors could easily tighten this up, create more realistic model results

6 Demonstration: Conjoint Design with Partial Pooling for Regularization

6.1 Prior PCs

7 Unresolved Issues

- Thinking harder about the information you have
  - Is an advantage of experiments that you can offload these demands?
    * estimates may be inefficient, but that's conservative which we like
    * What's the goal? Parameter estimation or minimizing type-1 error?

    - Thinking harder is a benefit
        * it reduces the risk of specifying 100 models
    - Imagine that your "minimal" model gives you something that you don't really believe, but your maximal model gives you something more realistic looking. What do you do?

- Pre-analysis plans
    - but Prior PCs

- Non-Bayesian methods with similar benefits
    - Regularization: cross-validation and ridge regression
    - Computation (inference?): randomization inference
        * RI is NHST-oriented and so should be treated with skepticism
        * Keele, McC and White say that randomization inference "directly estimates the quantity of interest" in reference to the p-value. The quantity of interest is the causal effect.

References

Gelman, Andrew, Hal S Stern, John B Carlin, David B Dunson, Aki Vehtari, and Donald B Rubin. 2013. *Bayesian data analysis*. Chapman and Hall/CRC.

Gerber, Alan S, Donald P Green, Edward H Kaplan, Ian Shapiro, Rogers M Smith, and Tarek Massoud. 2014. "The illusion of learning from observational research." *Field experiments and their critics: Essays on the uses and abuses of experimentation in the social sciences* , 9–32.

Holland, Paul W. 1986. "Statistics and causal inference." *Journal of the American statistical Association* 81(396): 945–960.

Keele, Luke. 2015. "The statistics of causal inference: A view from political methodology." *Political Analysis* 23(3): 313–335.

McElreath, Richard. 2015. *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC.

Rubin, Donald B. 1991. "Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism." *Biometrics* , 1213–1234.

Samii, Cyrus. 2016. "Causal empiricism in quantitative research." *The Journal of Politics* 78(3): 941–955.