Using V-Dem the Right Way: Monte Carlo Techniques for Regressing Random Variables

Dan Pemstein
NDSU NORTH DAKOTA STATE UNIVERSITY

V-Dem
VARIETIES OF DEMOCRACY

Workshop Goals

- (Re)familiarize you with Monte Carlo methods for estimating functions of random variables, integrating/marginalizing.
- (Re)familiarize you with how to work with the output of Markov chain Monte Carlo (MCMC) simulations.
- Introduce the V-Dem measurement model.
- Explain how to incorporate measurement uncertainty in V-Dem variables into statistical analyses (regressing random variables).

The Monte Carlo Principle

If we can sample many times from the density, $f(\theta)$, of a random variable, θ, we can learn anything we want to know about any computable function of that variable.

- $E(\theta) = \int \theta f(\theta) d\theta$.
- What if this integral is tricky to compute, but we can sample from $f(\theta)$?
- Sample $\theta^{(t)}$ for $t = 1, 2, \ldots, T$ draws from $f(\theta)$.
- $\sum_{t=1}^{T} \theta^{(t)} / T \rightarrow \int \theta f(\theta) d\theta$ as $T \rightarrow \infty$.

Example: Sums of Random Normal Variables

$$x_1 \sim \mathcal{N}(\mu_1, \sigma_1) \quad x_2 \sim \mathcal{N}(\mu_2, \sigma_2)$$

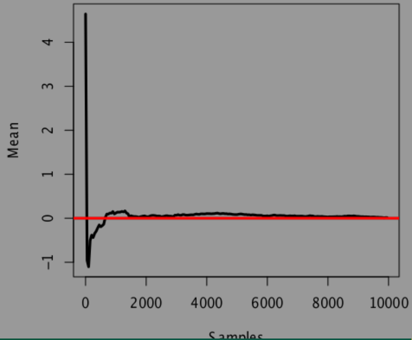
- What's the mean of $y = x_1 + x_2$. What's the SD?
    - $y \sim \mathcal{N}(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2})$
- Simulate:

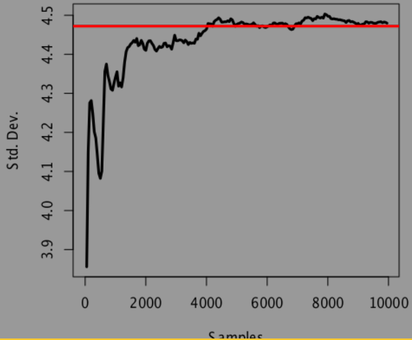
```
> T <- 10000
> MEAN <- c(-3, 3)
> SD <- c(2, 4)
> x1 <- rnorm(T, MEAN[1], SD[1])
> x2 <- rnorm(T, MEAN[2], SD[2])
> mean(x1 + x2)
[1] 0.01308404
> sd(x1 + x2)
[1] 4.476329
```

Example: Sums of Random Normal Variables

$$x_1 \sim \mathcal{N}(\mu_1, \sigma_1) \quad x_2 \sim \mathcal{N}(\mu_2, \sigma_2)$$

- What's the mean of $y = x_1 + x_2$. What's the SD?
  - $y \sim \mathcal{N}(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2})$





The Method of Composition

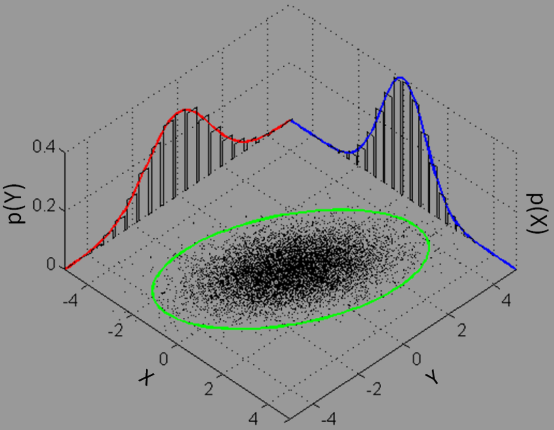
- Say you have a vector of n random variables $\boldsymbol{\theta} = \theta_1, \theta_2, \ldots, \theta_n$ and data vector $\mathbf{y}$.
- The joint posterior density of the random variables is $f(\boldsymbol{\theta}|\mathbf{y})$.
- You're interested in the marginal posterior density

$$f(\theta_1|\mathbf{y}) = \int f(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}_{-1} = \int f(\theta_1|\boldsymbol{\theta}_{-1}, \mathbf{y})f(\boldsymbol{\theta}_{-1}|\mathbf{y})d\boldsymbol{\theta}_{-1}$$

.

- If you can sample from $f(\theta_1|\boldsymbol{\theta}_{-1}, \mathbf{y})$ and $f(\boldsymbol{\theta}_{-1}|\mathbf{y})$, then you can simulate from the marginal density $f(\theta_1|\mathbf{y})$.
- for each $t \in 1, 2, \ldots T$ do
  1. sample $\boldsymbol{\theta}_{-1}^{(t)}$ from $f(\boldsymbol{\theta}_{-1}|\mathbf{y})$
  2. sample $\theta_1^{(t)}$ from $f(\theta_1|\boldsymbol{\theta}_{-1}^{(t)}, \mathbf{y})$
- $\theta_1^{(t)} \sim f(\theta_1|\mathbf{y})$.

The Method of Composition



Anatomy of a V-Dem Indicator

- We want to measure continuous latent traits (matrix of random variables, $\mathbf{Z}$).
- Coders have varying thresholds and reliabilities (coder parameters, vector of random variables, ϕ).
- Multiple coders per country-year provide observations on an ordinal scale (the data matrix, $\mathbf{R}$).
- We use MCMC methods to simulate latent traits from the marginal posterior density $f(\mathbf{Z}|\mathbf{R})$.
- We obtain a sample $\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, \ldots, \mathbf{Z}^{(T)}$:

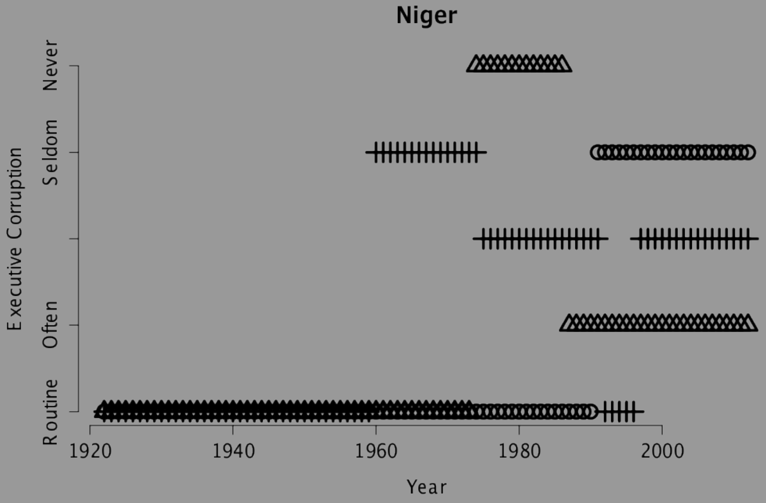
Afghanistan	1900	$z_{11}^{(1)}$	$z_{11}^{(2)}$	$\cdots$	$z_{11}^{(T)}$
Afghanistan	1901	$z_{12}^{(1)}$	$z_{12}^{(2)}$	$\cdots$	$z_{12}^{(T)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

V-Dem "C" Indicators are **Random Variables**

V-Dem VARIETIES OF DEMOCRACY

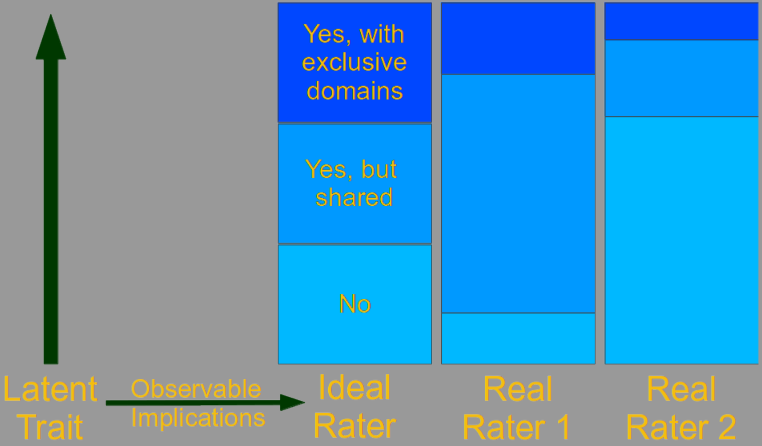
- Known only up to a density function.
- Working with point estimates throws out information.
  - This is true for both right and left-hand-side variables.
  - Hard to predict how measurement uncertainty will affect inferences.
    - Cross-correlations in draws.
    - Correlations with other variables may be robust across density.
- Standard "errors in variables" (EIV) model addresses a related, but different issue.
  - Data points in EIV model are country-years in $\mathbf{R}$.
  - Our measurement model addresses the EIV problem, while relaxing EIV assumptions about bias (a little bit).
- V-Dem point estimates are best estimates of latent values, but one shouldn't throw out our uncertainty around those estimates.

Raw Data (R)


Niger
Executive Corruption
Never
Seldom
Often
Routine
1920
1940
1960
1980
2000
Year

Threshold Bias

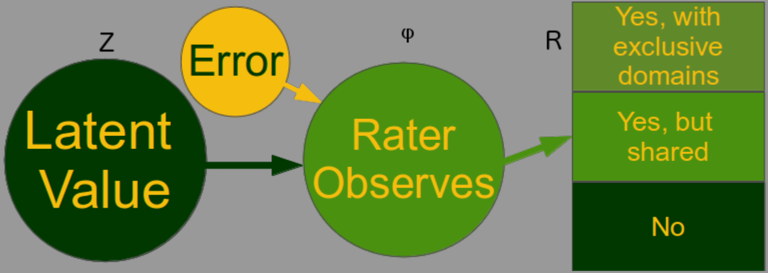
Can the head of government propose legislation in practice?



Modeling Framework: Core Assumptions

Multi-rater ordinal probit / ordinal item response theory model

- Rater perceptions equal reality + error
- Raters exhibit arbitrary thresholds
- Given thresholds, raters are correct on average (unbiased)
- Both thresholds and reliability vary across raters



Goal: For an arbitrary regression model, estimate the marginal posterior density of the coefficient vector $\boldsymbol{\beta}$, using V-Dem data, taking measurement uncertainty into account.

- Sample from the joint posterior density $f(\boldsymbol{\beta}, \mathbf{Z}|\mathbf{Y}, \mathbf{R})$, using the method of composition.
  - $\boldsymbol{\beta}$ is a vector of model coefficients.
  - $\mathbf{Y}$ is a matrix of data measured without uncertainty.
- We can take advantage of the decomposition $f(\boldsymbol{\beta}, \mathbf{Z}|\mathbf{Y}, \mathbf{R}) = f(\boldsymbol{\beta}|\mathbf{Z}, \mathbf{Y}, \mathbf{R})f(\mathbf{Z}|\mathbf{R}, \mathbf{Y})$.
- Assume:
  - $f(\boldsymbol{\beta}|\mathbf{Z}, \mathbf{Y}, \mathbf{R}) = f(\boldsymbol{\beta}|\mathbf{Z}, \mathbf{Y})$,
  - $f(\mathbf{Z}|\mathbf{R}, \mathbf{Y}) = f(\mathbf{Z}|\mathbf{R})$.
- We can now rewrite the decomposition as $f(\boldsymbol{\beta}, \mathbf{Z}|\mathbf{Y}, \mathbf{R}) = f(\boldsymbol{\beta}|\mathbf{Z}, \mathbf{Y})f(\mathbf{Z}|\mathbf{R})$.

Applying the Method of Composition

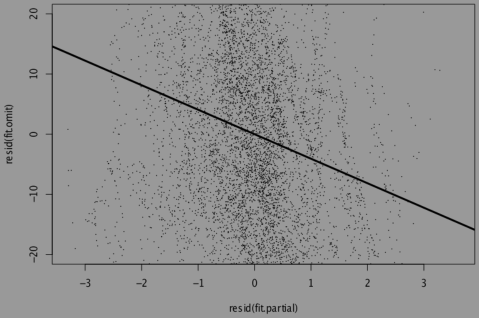
$$f(\boldsymbol{\beta}, \mathbf{Z}|\mathbf{Y}, \mathbf{R}) = f(\boldsymbol{\beta}|\mathbf{Z}, \mathbf{Y}) f(\mathbf{Z}|\mathbf{R})$$

- We want the marginal distribution of $\boldsymbol{\beta}, f(\boldsymbol{\beta}|\mathbf{Y})$.
- Remember:
  $f(\theta_1|\mathbf{y}) = \int f(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}_{-1} = \int f(\theta_1|\boldsymbol{\theta}_{-1}, \mathbf{y}) f(\boldsymbol{\theta}_{-1}|\mathbf{y}) d\boldsymbol{\theta}_{-1}$.
- So, given our assumptions: $f(\boldsymbol{\beta}|\mathbf{Y}) = \int f(\boldsymbol{\beta}|\mathbf{Z}, \mathbf{Y}) f(\mathbf{Z}|\mathbf{R}) d\mathbf{Z}$.
- The V-Dem modeling team already simulated T=900 draws where $\tilde{\mathbf{Z}}^{(t)} \sim f(\mathbf{Z}|\mathbf{R})$ [first MoC step].
- Regression coefficients are distributed $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. To apply the method of composition, for each $t \in 1, 2, \ldots, T$:
  1. Fit your arbitrary regression model to data $\mathbf{Y}$ and $\mathbf{Z}^{(t)}$, yielding partial likelihood estimates $\hat{\boldsymbol{\mu}}^{(t)}$ and $\hat{\boldsymbol{\Sigma}}^{(t)}$.
  2. Sample $\tilde{\boldsymbol{\beta}}^{(t)} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}^{(t)}, \hat{\boldsymbol{\Sigma}}^{(t)})$.

$$\text{infant mortality}_{cy} = \text{free discussion women}_{cy} + ln(\text{GDPpc})_{cy}$$

Afghanistan	1900	$\bar{z}_{11}$	$z_{11}^{(1)}$	$z_{11}^{(2)}$	$\cdots$	$z_{11}^{(T)}$
Afghanistan	1901	$\bar{z}_{12}$	$z_{12}^{(1)}$	$z_{12}^{(2)}$	$\cdots$	$z_{12}^{(T)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

1. Partial regression, point estimate for $\mathbf{z}$ (discussion women).

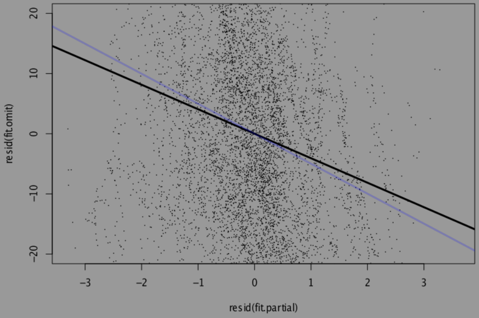


V-Dem Example

$$\text{infant mortality}_{cy} = \text{free discussion women}_{cy} + \mathit{ln}(\text{GDPpc})_{cy}$$

Afghanistan	1900	$\bar{z}_{11}$	$z_{11}^{(1)}$	$z_{11}^{(2)}$	$\cdots$	$z_{11}^{(T)}$
Afghanistan	1901	$\bar{z}_{12}$	$z_{12}^{(1)}$	$z_{12}^{(2)}$	$\cdots$	$z_{12}^{(T)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

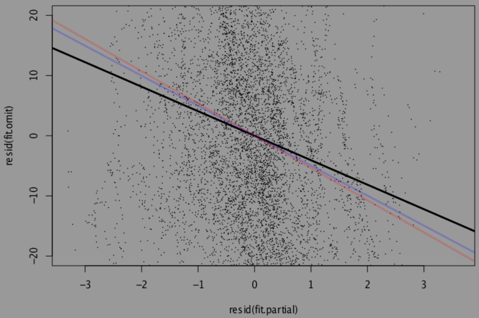
2. Fit model using a draw from the marginal density of $\mathbf{z}$.



$$\text{infant mortality}_{cy} = \text{free discussion women}_{cy} + ln(\text{GDPpc})_{cy}$$

Afghanistan	1900	$\bar{z}_{11}$	$z_{11}^{(1)}$	$z_{11}^{(2)}$	$\cdots$	$z_{11}^{(T)}$
Afghanistan	1901	$\bar{z}_{12}$	$z_{12}^{(1)}$	$z_{12}^{(2)}$	$\cdots$	$z_{12}^{(T)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

3. Sample $\tilde{\beta}^{(t)} \sim \mathcal{N}(\hat{\mu}^{(t)}, \hat{\Sigma}^{(t)})$.



V-Dem Example

$$\text{infant mortality}_{cy} = \text{free discussion women}_{cy} + ln(\text{GDPpc})_{cy}$$

Afghanistan	1900	$\bar{z}_{11}$	$z_{11}^{(1)}$	$z_{11}^{(2)}$	$\cdots$	$z_{11}^{(T)}$
Afghanistan	1901	$\bar{z}_{12}$	$z_{12}^{(1)}$	$z_{12}^{(2)}$	$\cdots$	$z_{12}^{(T)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$