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).
If we can sample many times from the density, $f(\theta)$, of a random variable, $\theta$, 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?
  - \( y \sim \mathcal{N}(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}) \)
Say you have a vector of $n$ random variables $\theta = \theta_1, \theta_2, \ldots, \theta_n$ and data vector $y$.

The joint posterior density of the random variables is $f(\theta | y)$.

You’re interested in the marginal posterior density

$$f(\theta_1 | y) = \int f(\theta | y) d\theta_{-1} = \int f(\theta_1 | \theta_{-1}, y) f(\theta_{-1} | y) d\theta_{-1}.$$

If you can sample from $f(\theta_1 | \theta_{-1}, y)$ and $f(\theta_{-1} | y)$, then you can simulate from the marginal density $f(\theta_1 | y)$.

for each $t \in 1, 2, \ldots T$ do

1. sample $\theta_{-1}^{(t)}$ from $f(\theta_{-1} | y)$
2. sample $\theta_1^{(t)}$ from $f(\theta_1 | \theta_{-1}^{(t)}, y)$

$\theta_1^{(t)} \sim f(\theta_1 | y)$. 
The Method of Composition

Using V-Dem the Right Way
Anatomy of a V-Dem Indicator

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

<table>
<thead>
<tr>
<th>Country</th>
<th>Year</th>
<th>$Z_{11}^{(1)}$</th>
<th>$Z_{11}^{(2)}$</th>
<th>$\ldots$</th>
<th>$Z_{11}^{(T)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afghanistan</td>
<td>1900</td>
<td></td>
<td></td>
<td>$\ldots$</td>
<td></td>
</tr>
<tr>
<td>Afghanistan</td>
<td>1901</td>
<td></td>
<td></td>
<td>$\ldots$</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$Z_{12}^{(1)}$</td>
<td>$Z_{12}^{(2)}$</td>
<td>$\ldots$</td>
<td>$Z_{12}^{(T)}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
<td>$\vdots$</td>
</tr>
</tbody>
</table>
V-Dem “C” Indicators are Random Variables

- 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 \( \mathbb{R} \).
  - 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.
Using V-Dem the Right Way

Niger

Year

Executive Corruption

Routine

Often

Seldom

Never
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 $\beta$, using V-Dem data, taking measurement uncertainty into account.

- Sample from the joint posterior density $f(\beta, Z|Y, R)$, using the method of composition.
  - $\beta$ is a vector of model coefficients.
  - $Y$ is a matrix of data measured without uncertainty.

- We can take advantage of the decomposition
  \[ f(\beta, Z|Y, R) = f(\beta|Z, Y, R)f(Z|R, Y). \]

- Assume:
  - $f(\beta|Z, Y, R) = f(\beta|Z, Y),$
  - $f(Z|R, Y) = f(Z|R).

- We can now rewrite the decomposition as
  \[ f(\beta, Z|Y, R) = f(\beta|Z, Y)f(Z|R) . \]
Applying the Method of Composition

\[ f(\beta, Z|Y, R) = f(\beta|Z, Y)f(Z|R) \]

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

\begin{align*}
\text{Afghanistan} & \quad 1900 & \bar{z}_{11} & z_{11}^{(1)} & z_{11}^{(2)} & \cdots & z_{11}^{(T)} \\
\text{Afghanistan} & \quad 1901 & \bar{z}_{12} & z_{12}^{(1)} & z_{12}^{(2)} & \cdots & z_{12}^{(T)} \\
\vdots & \quad \vdots & \vdots & \vdots & \vdots & \cdots & \vdots 
\end{align*}

1. Partial regression, point estimate for \( z \) (discussion women).
infant mortality_{cy} = free discussion women_{cy} + ln(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 \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots

2. Fit model using a draw from the marginal density of z.
infant mortality_{cy} = \text{free discussion women}_{cy} + \ln(\text{GDPpc})_{cy}

\begin{align*}
\text{Afghanistan} & \quad 1900 \quad \bar{z}_{11} \quad z_{11}^{(1)} \quad z_{11}^{(2)} \quad \cdots \quad z_{11}^{(T)} \\
\text{Afghanistan} & \quad 1901 \quad \bar{z}_{12} \quad z_{12}^{(1)} \quad z_{12}^{(2)} \quad \cdots \quad z_{12}^{(T)} \\
\vdots & \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots
\end{align*}

3. Sample \( \tilde{\beta}^{(t)} \sim \mathcal{N}(\hat{\mu}^{(t)}, \hat{\Sigma}^{(t)}) \).
infant mortality_{cy} = free discussion women_{cy} + ln(GDPpc)_{cy}

Affghanistan 1900 \bar{z}_{11} \ z^{(1)}_{11} \ z^{(2)}_{11} \cdots \ z^{(T)}_{11}

Affghanistan 1901 \bar{z}_{12} \ z^{(1)}_{12} \ z^{(2)}_{12} \cdots \ z^{(T)}_{12}

... ... ... ... ... ...