A Brief History of Ecological Inference: the Good, the Bad, and the Ugly

Jeff Gill
University of California, Davis
Department: Political Science
Email: jgill@ucdavis.edu

♢ **Ecological Inference:** the process of using aggregate (ecological) data to infer discrete individual-level relationships of interest when individual-level data are unavailable.

♢ **History of the Ecological Inference Problem:**

1. The oldest recognized quantitative problem in political science: William Ogburn and Inez Goltra, in the very first multivariate statistical analysis of politics in a political science journal (PSQ 1919) made ecological inferences and recognized the problem. The big issue: are the newly enfranchised women going to take over the American political system? They regressed votes in an Oregon referendum on the percent of women in each precinct in a local election where women were allowed to participate before 1919. But they worried:

   > It is also theoretically possible to gerrymander the precincts in such a way that there may be a negative correlation even though men and women each distribute their votes 50 to 50 on a given measure...

2. One of the oldest methodological problems in the social sciences in general.
3. Robinson's (1950) article was the most influential, causing:
   (a) several literatures to retrench, including studies of local and regional politics through aggregate electoral statistics.
   (b) a new methodological literature to emerge devoted to solving the problem.
   (c) gave rise to survey research as the dominant data collection methodology in several fields, notably political science and sociology.

♢ **Voting in the Weimer Republic:** starting in the 1930 election in Germany a previously obscure party, the National Socialist German Worker's Party, began amassing electoral support and would eventually control the government. What type of person voted for this party? Were they predominantly from the lower middle class dramatically affected by failed economic policies? Were they from particular religious groups or geographic regions? Since most records were destroyed during the war, survey research is impractical and unreliable in this setting, and only aggregate data exists, then it is necessary to make ecological inferences.

♢ **Epidemiology and Public Policy:** the Appalachian region of the United States suffers from dramatically higher rates of certain cancers. While much data can be obtained by the medical records of living patients and some historical data can be obtained for patients treated in major medical facilities over the latter part of the last century, a great many cases were poorly diagnosed and produce no useful evidence about type of cancer and contributing causes. In this case ecological inferences can be made about covariates.

♢ **Education:** Some students attend elite private schools through voucher and assistance programs, whereas others can attend due to their parents resources. Do the former students perform as well as the latter? Since individual student performance records are protected by a host of privacy laws, only ecological inference can answer this question.

♢ **Marketing:** It is often the case that firms know the sales figures and the demographics for various regions but cannot specifically determine which subgroups of the population prefer their products. Ecological inference can indicate brand preference in relatively fine granulation.

Motivating "Case"

♢ **1965 Voting Rights Act:** along with its extensions (1970, 1975, 1982) prohibits electoral discrimination based on race, color, or language. Evidence of discrimination requires an indication that individuals are hampered or prevented from voting.

♢ **As amended in 1982**, the Act also forbids any practice that is shown to have a disparate effect on minority voting strength: in aggregating voting from the precinct level to the district level or higher, the votes of minorities are sufficiently diluted as to deny effective representation.

♢ **In Thornburg v. Gingles 1986** the U.S. Supreme Court ruled that discriminatory intent is not necessary, that it is sufficient for the plaintiff to show that absent electoral representation, three conditions are sufficient:

1. there exists a geographical district where the minority is a majority.
2. the minority group is politically cohesive.
3. the majority votes essentially as a unified racial voting block.

Motivating "Case"

(continued)

♢ **June 4, 1990** the U.S. District Court of the Central District of California filed an opinion in *Yolanda Garza et al. v. County of Los Angeles, Los Angeles Board of Supervisers et al.* in favor of the plaintiffs.

♢ **Background:** The Los Angeles County Board of Supervisors consists of 5 members elected to 4 year terms in non-partisan elections by district. An Hispanic had *never* been elected to the Board. In the ruling the Judge ruled that the plaintiffs had provided adequate statistical evidence through ecological inference that the Board of Supervisors had engaged in periodic redistricting that diluted Hispanic voting power.

♢ **The defining issue:** compact precincts of minority voters were shown to be aggregated together with larger numbers of non-minority precincts in all of the 5 districts.

The Problem

♢ District Level:

Race of Person	**Voting Decision** Democrat	Republican	Abstain	
$\underline{b}lack$	?	?	?	55,054
$\underline{w}hite$	?	?	?	25,706
	19,896	10,936	49,928	80,760

The 1990 Election to the Ohio State House, District 42.

♢ Precinct Level:

Race of Person	**Voting Decision** Democrat	Republican	Abstain	
$\underline{b}lack$	?	?	?	221
$\underline{w}hite$	?	?	?	484
	130	92	483	705

Precinct 1 of 131 in District 42.

Goodman's Regression

♢ Context: *Garza et al. v. County of Los Angeles et al.*

♢ Notation:

$y_i =$ percent received by Hispanic candidate from precinct i

$x_i =$ percent Hispanic voters in precinct i

$\epsilon_i =$ zero mean random error term

♢ Assumptions:

1. turnout across districts identical for minorities, and non-minorities.
2. percent minority voting independent of percent minority.

♢ Method: linearly regress $\mathbf{y}$ on $\mathbf{x}$:

$$y_i = \alpha + \beta x_i + \epsilon_i$$

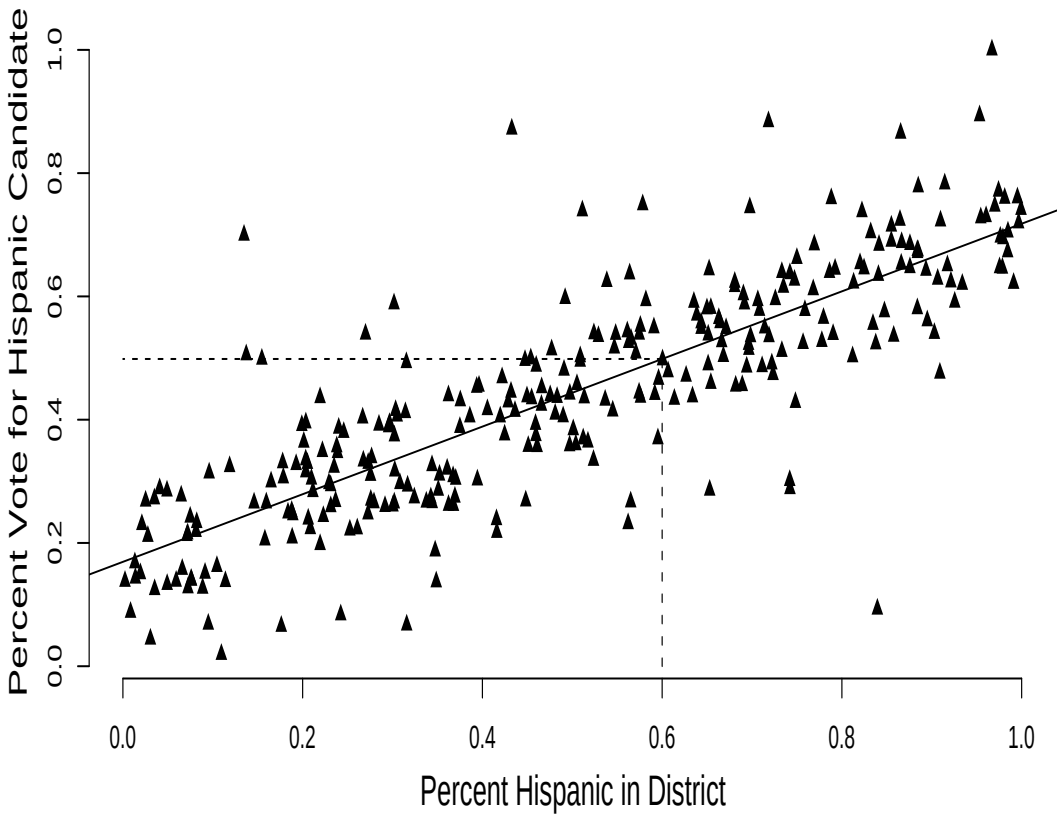
Interpretation:

1. α is the percentage of non-Hispanics voting for the Hispanic candidate.
2. $\alpha + \beta$ is the percentage of Hispanics.
3. model assumes α and β constant across all precincts.

(continued)

♢ Visual Interpretation:

Figure 1: 1998 Precinct Data: Sacramento, California



♢ Problems:

1. Assumes that votes determined by ethnicity not precinct-level factors: α and β constant across precincts, and ϵ identically distributed with mean zero.
2. Ecological data often heteroscedastically related to unit size.
3. Procedure often gives ridiculous answers.

(continued)

♢ True Example: Every Ohio State House district where and African American Democrat ran against a white Republican, 1986-1990.

Year	District	Estimated Percent of African Americans Voting for the Democratic Candidate
1986	12	95.65%
	23	100.06
	29	103.47
	31	98.92
	42	108.41
	45	93.58
1988	12	95.67
	23	102.64
	29	105.00
	31	100.20
	42	93.58
	45	97.49
1990	12	94.79
	14	97.83
	16	94.36
	23	101.09
	25	98.83
	29	103.42
	31	102.17
	36	101.35
	37	101.39
	42	109.63
	45	97.62

Source: statement of Gordon G. Henderson, presented as an exhibit in U.S. Federal District Court

♢ Object to constancy assumption of Goodman's Regression.

♢ Substitutes a *new constancy assumption*: the probability that the j^{th} Hispanic individual in precinct p votes equals the probability that the j^{th} non-Hispanic individual in precinct p, and differences due to the precincts.

♢ District heterogeneity/Precinct Homogeneity:

$$y_{jnp} = y_{jhp} = (\alpha + \beta)_p + \epsilon_j$$

♢ Aggregate over precinct:

$$y_p = (hy_{hjp} + ny_{jnp})/N = (\alpha + \beta)_p + \epsilon_p$$

♢ Now assume that differences in support for Hispanic candidate across precincts are due to proportion of Hispanics in district:

$$E[y_p] = (\alpha + \beta)_p = \alpha + \beta\frac{h}{N}$$

♢ Interpretation:

1. α is the percentage of votes for the Hispanic candidate *in an all non-Hispanic precinct*.
2. $\alpha + \beta$ is the percentage of votes for the Hispanic candidate *in an all Hispanic precinct*.
3. β is the precinct difference due to Hispanic proportion.

♢ Suppose the linear model produces:

$$y = 30\% + 40\%x$$

and we evaluate a district where $x_p = 0.6$

♢ Goodman's Regression:

30% non-Hispanics and 70% Hispanics vote for Hispanic candidate, so:

$$y_p = 30\%(0.4) + 70\%(0.6) = 54\%.$$

♢ Freedman et al.:

40% additional votes for the Hispanic candidate due to proportion of Hispanics in district:

$$y_p = 30\% + 40\%(0.6) = 54\%$$

♢ Although the two models differ in the underlying assumptions, they produce identical aggregate conclusions.

♢ Assumptions versus Conclusions.

(Duncan and Davis 1953, King 1997)

♢ Observed Variables:

$T_i =$ voter $\underline{T}urnout$ in precinct i

$X_i =$ Black proportion of voting age population in precinct i

♢ Unobserved Variables:

$\beta_i^b =$ proportion of $\underline{b}lacks$ voting in precinct i

$\beta_i^w =$ proportion of $\underline{w}hites$ voting in precinct i

♢ Table:

	Vote	Abstain	
$\underline{b}lack$	β_i^b	$1-\beta_i^b$	X_i
$\underline{w}hite$	β_i^w	$1-\beta_i^w$	$1-X_i$
	T_i	$1-T_i$	

♢ Accounting Identity:

$$\begin{aligned} T_i &= \beta_i^b X_i + \beta_i^w (1 - X_i) \\ &= \beta_i^w + (\beta_i^b - \beta_i^w) X_i \end{aligned}$$

(continued)

♢ Precinct Example

	Vote	Abstain	
$\underline{b}lack$	β_i^b	$1-\beta_i^b$	221
$\underline{w}hite$	β_i^w	$1-\beta_i^w$	484
	222	483	705

$$T_i = \tfrac{222}{705} = 0.315$$

$$X_i = \tfrac{221}{705} = 0.314$$

$$\beta_i^b \in \left[\frac{0}{221}, \frac{221}{221}\right] = [0,1]$$

	Vote	Abstain	
$\underline{b}lack$	$[0, 221]$	$[0, 221]$	221
$\underline{w}hite$	$[1, 222]$	$[262, 483]$	484
	222	483	705

$$1-\beta_i^b \in \left[\frac{0}{221}, \frac{221}{221}\right] = [0,1]$$

$$\beta_i^w \in \left[\frac{1}{484}, \frac{222}{484}\right] = [0.002, 0.459]$$

$$1-\beta_i^w \in \left[\frac{262}{484}, \frac{483}{484}\right] = [0.541, 0.998]$$

♢ More "bite":

	hunting license	no hunting license	
$\underline{f}emale$	$[0.875, 1]$	$[0, 0.125]$	16
	$(14-16)$	$(0-2)$	
$\underline{m}ale$	$[0, 0.5]$	$[0, 0.5]$	4
	$(2-4)$	$(0-2)$	
	18	2	20

(continued)

♢ Formalized:

$$\max\left[0, \frac{T_i - (1 - X_i)}{X_i}\right] \leq \beta_i^b \leq \min\left[\frac{T_i}{X_i}, 1\right]$$

$$\max\left[0, \frac{T_i - X_i}{1 - X_i}\right] \leq \beta_i^w \leq \min\left[\frac{T_i}{1 - X_i}, 1\right]$$

♢ More realistic example: Second State Senate District, Precinct 52, Pennsylvania (1990).

$$X_{52} = \text{proportion Hispanic} = 0.88$$

$$T_{52} = \text{turnout proportion} = 0.19$$

$$\max\left[0, \frac{0.19 - (1 - 0.88)}{0.88}\right] \leq \beta_{52}^H \leq \min\left[\frac{0.19}{0.88}, 1\right]$$

$$0.07 \leq \beta_{52}^H \leq 0.21$$

$$\max\left[0, \frac{0.19 - 0.88}{1 - 0.88}\right] \leq \beta_{52}^N \leq \min\left[\frac{0.19}{1 - 0.88}, 1\right]$$

$$0 \leq \beta_{52}^N \leq 1$$

(continued)

♢ Recall the accounting identity:

$$T_i = \beta_i^b X_i + \beta_i^w (1 - X_i)$$

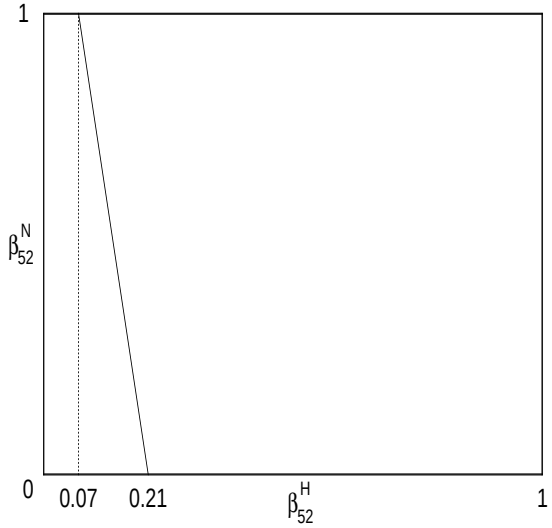
♢ Algebraically rearrange:

$$\beta_i^w = \underbrace{\left(\frac{T_i}{1 - X_i}\right)}_{intercept} - \underbrace{\left(\frac{X_i}{1 - X_i}\right)}_{slope} \beta_i^b$$

♢ Precinct 52:

$$\begin{aligned} \beta_{52}^N &= \left(\frac{0.19}{1 - 0.88}\right) - \left(\frac{0.88}{1 - 0.88}\right) \beta_{52}^H \\ &= 1.58 - 7.33 \beta_{52}^H \end{aligned}$$

Figure 2: Range of Possible Values



♢ Conceptualize with a tomography plot:

♢ Specify a truncated bivariate normal distribution generating the data over the unit square:

♢ Univariate inference produced from the "slice" determined by the accounting identity:

(continued)

♢ Random Coefficients Model

- Start with the accounting identity:

$$T_i = \beta_i^b X_i + \beta_i^w (1 - X_i)$$

- Assumption 1:

$$\left[\beta_i^b, \beta_i^w\right] \sim TBVN(\mathbb{B}, \mathbf{\Sigma}) \qquad \mathbb{B} = \begin{bmatrix} \mathbb{B}^b \\ \mathbb{B}^w \end{bmatrix}, \mathbf{\Sigma} = \begin{bmatrix} \sigma_b^2 & \sigma_{bw} \\ \sigma_{bw} & \sigma_w^2 \end{bmatrix}$$

truncated parameter vector: $\psi = [\mathbb{B}^b, \mathbb{B}^w, \sigma_b, \sigma_w, \rho]$

- Assumption 2:

β_i^b, β_i^w are mean independent of X_i.

- Assumption 3:

$T_i|X_i$ are independent of $T_j|X_j, \forall i \neq j$.

- Assumption 4:

an unbiased error structure: $\beta_i^b = \mathbb{B}^b + \epsilon_i^b$

$(E[\epsilon_i^b] = 0, E[\epsilon_i^w] = 0)$ $\beta_i^w = \mathbb{B}^w + \epsilon_i^w$

(continued)

♢ Plug random coefficients into accounting identity:

$$\begin{aligned} T_i &= \beta_i^b X_i + \beta_i^w (1 - X_i) \\ &= (\mathbb{B}^b + \epsilon_i^b) X_i + (\mathbb{B}^w + \epsilon_i^w)(1 - X_i) \\ &= \mathbb{B}^b X_i + \mathbb{B}^w (1 - X_i) + \underbrace{\epsilon_i^b X_i + \epsilon_i^w (1 - X_i)}_{\epsilon_i} \end{aligned}$$

♢ Since $E[\epsilon_i | X_i] = 0$, then:

$$\mu_i = E[T_i | X_i] = \mathbb{B}^b X_i + \mathbb{B}^w (1 - X_i)$$

♢ And:

$$\epsilon_i = T_i - \mu_i = T_i - \mathbb{B}^b X_i - \mathbb{B}^w (1 - X_i)$$

♢ This is a linear model, so:

$$\begin{aligned} \sigma_i = \mathsf{var}[T_i | X_i] &= \mathsf{var}[\epsilon_i | X_i] \\ &= \mathsf{var}[\epsilon_i^b X_i + \epsilon_i^w (1 - X_i)] \\ &= E[(\epsilon_i^b X_i + \epsilon_i^w (1 - X_i))^2] - (E[\epsilon_i^b X_i + \epsilon_i^w (1 - X_i)])^2 \\ &= (\sigma_w^2) + (2\sigma_{bw} - 2\sigma_b^2) X_i + (\sigma_b^2 + \sigma_w^2 - 2\sigma_{bw}) X_i^2 \end{aligned}$$

(continued)

♢ These results demonstrate that the observed data can be used to obtain estimates of the parameters of the truncated bivariate normal, $TBVN[\mathbb{B}^b, \mathbb{B}^w, \sigma_b, \sigma_w, \rho]$:

$$\begin{aligned} E[T_i|X_i = 0] &= \mathbb{B}^w \\ E[T_i|X_i = 1] &= \mathbb{B}^b \\ \text{var}[T_i|X_i = 0] &= \sigma_w^2 \\ \text{var}[T_i|X_i = 1] &= \sigma_b^2 \\ \frac{\sigma_{bw}}{\sigma_b \sigma_w} &= \rho \end{aligned}$$

(Although this is poor way to do so.)

♢ Likelihood for $\psi = [\mathbb{B}^b, \mathbb{B}^w, \sigma_b, \sigma_w, \rho]$:

$$\begin{aligned} L(\psi|\mathbf{T}, \mathbf{X}) &\propto \Pi_{X_i \in [0,1]} P(T_i|\psi) \\ &= \Pi_{X_i \in [0,1]} N(T_i|\mu_i, \sigma_i^2) \frac{S(\mathbb{B}, \mathbf{\Sigma})}{R(\mathbb{B}, \mathbf{\Sigma})} \end{aligned}$$

where: $$R(\mathbb{B}, \mathbf{\Sigma}) = \int_0^1 \int_0^1 BN(\beta_i^b, \beta_i^w|\mathbb{B}, \mathbf{\Sigma}) d\beta^b d\beta^w$$

$$S(\mathbb{B}, \mathbf{\Sigma}) = \int_{\max\left(0, \frac{T_i - (1 - X_i)}{X_i}\right)}^{\min\left(1, \frac{T_i}{X_i}\right)} N\left(\beta^b|\mathbb{B}^b + \frac{w_i}{\sigma_i^2}, \sigma_b^2 - \frac{w_i}{\sigma_i^2}\right) d\beta^b$$

and: $$w_i = \sigma_b^2 X_i + \sigma_{bw}(1 - X_i)$$

(continued)

♢ Interpretation:

$\mathbb{B}^b$ is the unconditional average across precincts.

$$E[\beta_i^b|T_i, X_i] = \mathbb{B}^b + \frac{w_i}{\sigma_i^2}\epsilon_i$$

$\frac{w_i}{\sigma_i^2}$ comes from plugging the random effects into the accounting identity; the estimated proportion of ϵ_i variation is now systematic.

σ_b is the variance without conditioning on T.

$$\text{var}[\beta_i^b|T_i, X_i] = \sigma_b - \frac{w_i^2}{\sigma_i^2}$$

$\frac{w_i^2}{\sigma_i^2}$ is the variance shrinkage from conditioning on T. When X_i large, σ_b dominates, and when X_i small then we have to "borrow more strength" from covariance.

♢ Unconditional posterior of β_i^b:

$$P(\beta_i^b|X_i, T_i) \propto \int_{-\infty}^{+\infty} P(\psi) \prod_{i=1}^{P} N(T_i|\mu_i, \sigma_i^2) \frac{S(\mathbb{B}, \mathbf{\Sigma})}{R(\mathbb{B}, \mathbf{\Sigma})} \times TBVN\left(T_i|\mathbb{B}_i^b + \frac{w_i}{\sigma_i^2}\epsilon_i, \sigma_b^2 - \frac{w_i^2}{\sigma_i^2}\right) d\psi$$

(continued)

♢ The unconditional marginal distribution of β_i^b has no analytical solution.

♢ Estimation Steps for β_i^b:

1. Obtain maximum likelihood estimates of TBVN parameters: $\hat{\psi} = [\hat{\mathbb{B}}^b, \hat{\mathbb{B}}^w, \hat{\sigma}_b^2, \hat{\sigma}_w^2, \hat{\sigma}_{bw}]$.
2. Draw a 5-dimensional MVN vector on the untruncated scale with mean and variance from the maximum likelihood estimation: $\tilde{\psi} = [\tilde{\mathbb{B}}^b, \tilde{\mathbb{B}}^w, \tilde{\sigma}_b^2, \tilde{\sigma}_w^2, \tilde{\sigma}_{bw}]$.
3. Improve the normal approximation with importance sampling.
4. For each i calculate:

$$\tilde{\sigma}_i^2 = \text{var}(T_i|X_i) = \tilde{\sigma}_b^2 X_i^2 + \tilde{\sigma}_w^2 (1 - X_i)^2 + \tilde{\sigma}_{bw} 2 X_i (1 - X_i)$$
$$\tilde{\epsilon}_i = T_i - \mu_i = T_i - \tilde{\mathbb{B}}^b X_i - \tilde{\mathbb{B}}_w (1 - X_i)$$
$$\tilde{w}_i = \tilde{\sigma}_b^2 X_i + \tilde{\sigma}_{bw} (1 - X_i)$$

5. Substitute these values into the *truncated* conditional normal posterior for β_i^b:

$$P(\beta_i^b | T_i, \hat{\psi}) = N\left(\tilde{\beta}^b | \tilde{\mathbb{B}}^b + \frac{\tilde{w}_i}{\tilde{\sigma}_i^2}, \tilde{\sigma}_b^2 \frac{\tilde{w}_i}{\tilde{\sigma}_i^2}\right) \frac{I_{\beta_i \in [0,1]}}{S(\mathbb{B}, \mathbf{\Sigma})}$$

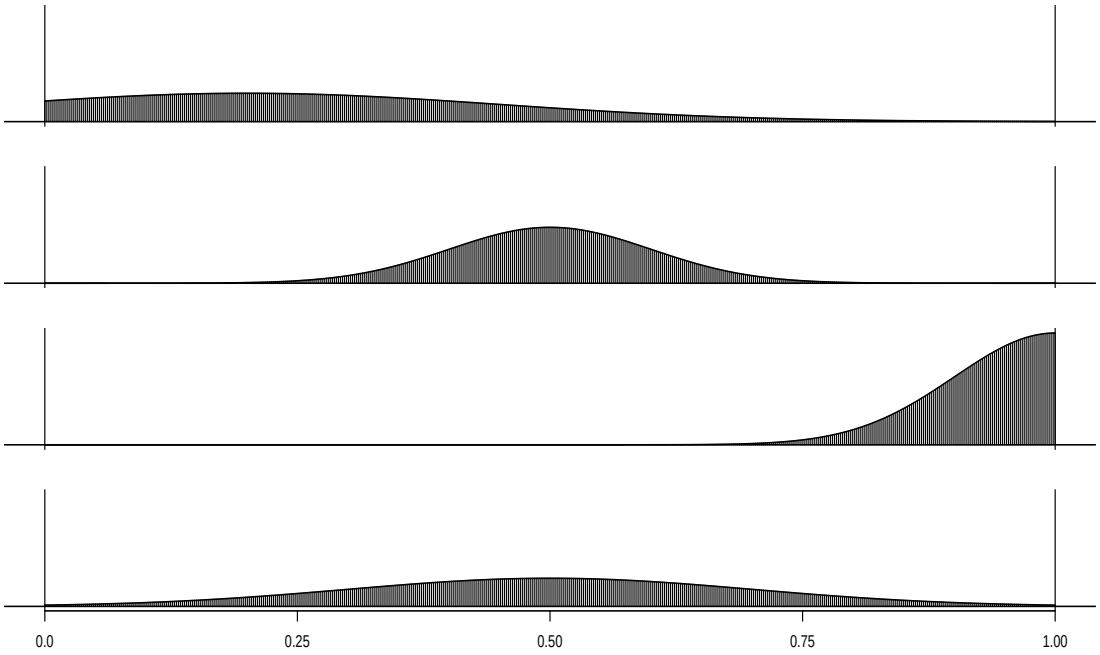
   and randomly draw a value of β_i from this distribution.
6. Repeat steps 2-5 many times. Now there is a complete posterior distribution for each i to produce empirical summaries (a "slice" not a marginal).

(continued)

♢ Okay, now what do we do?

1. Take many (K) draws from $P(\beta_i^b|T_i, \hat{\psi})$ for each precinct (i) of interest: $\tilde{\beta}_i^{b(K)}$
2. Plot posterior density estimates.

Figure 3: Posterior Distributions of Precinct Parameters: β_i^b, Pennsylvania 8th State Senate,1990.



3. Calculate point estimates:

$$\hat{\beta}_i^b = \frac{1}{K}\sum_{k=1}^{K}\tilde{\beta}_i^{b(k)} \qquad \text{var}[\beta_i^b] = \frac{1}{K}\sum_{k=1}^{K}(\tilde{\beta}_i^{b(k)} - \hat{\beta}_i^b)^2$$

4. Make confidence intervals from empirical draws:
   (a) sort the k draws: $\tilde{\beta}^{b(1)}, \tilde{\beta}^{b(2)}, \ldots, \tilde{\beta}^{b(K)}$.
   (b) 90% CI: $\left[\tilde{\beta}^{b(\lfloor 0.05k \rfloor)} : \tilde{\beta}^{b(\lceil 0.95k \rceil)}\right]$.

(continued)

♢ Extensions

1. Adding covariates through the mean function.
2. Nonparametric density estimation of multimodal posteriors (King 1997)
3. EM estimation of TBVN (Mattos & Veiga 2000)

♢ Criticisms

1. Freedman
2. Tam-Cho
3. Rivers