Bayesian Analysis of Comparative Survey Data

Bruce Western[1]
Filiz Garip
Princeton University

April 2005

[1]Department of Sociology, Princeton University, Princeton NJ 08544. We thank Sara Curran for making the Nang Rong migration data available for this paper. This research was supported by the Princeton University Science Fund. `WinBUGS` code for the analysis reported in this paper is available at: `www.princeton.edu/~western`.

Abstract

Bayesian hierarchical models provide a useful way of analyzing multilevel survey data. The Bayesian estimates have good statistical properties, make good predictions, and realistically account for clustering in the data. Still the Bayesian estimates can be biased in the presence of omitted variables and fixed effect models might sometimes be preferable. Bayesian statistics for model comparison and evaluation—posterior predictive checks and the Deviance Information Criterion—assist an empirical approach to distinguishing between hierarchical models and their alternatives. These ideas are illustrated with an analysis of migration data from 22 villages in the Nang Rong district of Thailand.

Bayesian statistics can make a special contribution to comparative and historical social science. Comparative data are often not generated by a well-defined probability mechanism, so a researcher's uncertainty may be better described by a degree-of-belief than the frequency behavior of sample statistics (Berk et al. 1994). Comparative and historical researchers also unearth rich qualitative information about particular countries, regions, and historical periods. In a classicial analysis, this nonsample information generally provides informal guides to model choice or the post hoc interpretation of results. Bayesian prior distributions explicitly incorporate non-sample information that often influences data analysis in a more informal way (Western and Jackman 1994). Prior information can have a large effect in comparative analysis because data sets can be small and collinear. Under these conditions, the final results may also depend closely on the choice of models. Bayesian statistics can incorporate uncertainty about the model specification, pushing inference in a more conservative direction (Western 1995). Finally, a key message of comparative social science is that social and political processes vary across countries, regions, and time periods. Bayesian hierarchical models help us analyze these kinds of heterogeneity (Western 1998; Western and Kleykamp 2004).

This symposium on the analysis of multilevel survey data provides another opportunity to apply Bayesian methods to the special methodological problems of comparative research. Multilevel survey data are collected from, say, a dozen or more countries, perhaps at several points in time. This data structure shares some features with the pooled time series familiar to comparative researchers—observations are clustered by country and there is likely causal heterogeneity across countries. But unlike other comparative data, multilevel survey data provide enough information about each country to

conduct a country-level analysis. In this case, the data (within countries, at least) are usually generated by probability sampling and prior information will be less influential because sample sizes within countries are large.

Bayesian hierarchical models provide a useful way to study these clustered, causally heterogeneous, survey data. But Bayesian models can yield biased estimates when non-Bayesian alternatives do better. The Bayesian models can also be difficult to compare to non-Bayesian alternatives because models are non-nested and null hypotheses lie on the boundary of parameter spaces. In this paper we review the Bayesian hierarchical model for multilevel survey data. We argue that the merits of Bayesian and non-Bayesian models should be assessed empirically so we describe some Bayesian statistics for model comparison and evaluation. We illustrate these methods using survey data on migration from 22 villages in Thailand.

Bayesian Inference

Bayesian statistical inference pools sample data with nonsample information to make posterior probability statements about statistical parameters. Given a sampling distribution for the data, $p(y|\theta)$, and a prior distribution for the parameters, $p(\theta)$, the posterior inferences about θ given the data are made by applying Bayes rule:

$$p(\theta|y) \propto p(y|\theta) \times p(\theta)$$

If y is normally distributed with mean θ and variance, σ^2, and θ has a normal prior distribution with mean θ_0 and variance τ^2, then the posterior distribution for θ is also normal, where the posterior mean is the weighted average of the sample mean, $\bar{y}$, and the prior mean, θ_0,

$$\theta_1 = w\theta_0 + (1 - w)\bar{y}.$$

The weight, w, depends on the relative size of the prior variance and the variance of the sampling distribution for y,

$$w = \sigma^2/(\tau^2 + \sigma^2).$$

The posterior variance of θ is given by,

$$V(\theta|y) = 1/(\tau^{-2} + \sigma^{-2}).$$

If the prior variance is large, indicating great uncertainty about the location of θ before the data are observed, then w will be small and the posterior mean θ_1 will be near the sample mean $\bar{y}$. In practice, the variance of the sampling distribution, σ^2 must also be estimated and given a prior distribution in a Bayesian analysis. The main intuition is unaffected: the posterior mean is a compromise between the prior and sample mean that depends on the relative size of the prior and sample variance. Why do Bayesian inference? In this case, the posterior variance is smaller than the usual sampling variance. The gain in precision is not costless, however. The posterior mean, θ_1, will generally be biased and the choice of prior distribution is subjective, so different researchers may choose different values for θ_0 and τ^2.

Bayesian Inference for Comparative Analysis

How can this machinery be used for comparative research? If we are interested in estimating means, θ_i, from a number of surveys from countries, $i = 1, \ldots, I$, a mean could be calculated for each country, $\bar{y}_i$. If $\bar{y}_i \sim N(\theta_i, \sigma^2)$ we can obtain more precise estimates for each of I countries by supplying a prior distribution:

$$\theta_i \sim N(\theta_0, \tau^2).$$

With data from many countries, the prior parameters θ_0 and τ^2 for a particular θ_i can be estimated using the sample data from all countries. The prior parameters could be estimated by the mean of the country means, $\hat{\theta}_0 = I^{-1} \sum \bar{y}_i$, and variance of the country means, $\hat{\tau}^2 = \sum(\bar{y}_i - \hat{\theta}_0)^2/I$. This is the empirical Bayes estimator that yields the estimates of the posterior mean:

$$\hat{\theta}_i = \hat{w}\hat{\theta}_0 + (1 - \hat{w})\bar{y}_i$$

where $\hat{w} = \sigma^2/(\sigma^2 + \hat{\tau}^2)$. The empirical Bayes estimator has lower mean squared error and makes better forecasts than the maximum likelihood estimate, $\bar{y}_i$ (e.g., Carlin and Louis 2000, 69–72; Jusko and Shively make the point more rigorously in this issue). The superior performance of the empirical Bayes estimator results from the larger amount of information used for its estimation. The country-level mean, $\bar{y}_i$, uses just the sample data from country i. The empirical Bayes estimate, $\hat{\theta}_i$, uses the sample data from country i and additional information from all countries expressed by the hyperparameters $\hat{\theta}_0$ and $\hat{\tau}_2$.

Although empirical Bayes estimates are an improvement over maximum likelihood, the method does not acknowledge uncertainty about the prior parameters, θ_0 and τ^2. A fully Bayesian analysis puts prior distributions on these hyperparameters.

Multilevel Regression Models

The model for the mean, θ, can be generalized to a regression where y_i is a vector of observations on a dependent variable from country i ($i = 1, \ldots, I$) and X_i is a matrix of K covariates. A common Bayesian linear model gives a set of K regression for each of I countries. If the coefficients are in the $I \times K$

matrix, b, with rows, $b_{i\cdot}$ and columns $b_{\cdot k}$, then y_i has a normal likelihood,

$$y_i \sim N(X_i b_{i\cdot}, \sigma_i^2).$$

The regression coefficients, $b_{i\cdot}$, may be given a distribution that depends on country-level variables, z_i, that could describe, say, institutional or historical characteristics of national societies. The I coefficients for covariate k, $b_{\cdot k}$, are then given a normal distribution,

$$b_k \sim N(Z'\eta_k, \tau_k^2),$$

where Z is matrix of country-level covariates with rows, z_i. The empirical Bayes model is specified by the likelihood for y_i and the distribution for $b_{\cdot k}$. Because the coefficients, $b_{\cdot k}$, are given a distribution, they are often called "random effects." A fully Bayesian analysis adds prior distributions for η_k and τ_k^2. A conjugate analysis that yields normal posterior distributions assigns normal priors to the coefficients and gamma distributions to the inverse variances:

$$\begin{aligned} \eta_k &\sim N(c_{0k}, V_{0k}) \\ \tau_k^{-2} &\sim \text{Gamma}(a_0, b_0) \end{aligned}$$

With the η coefficients of covariate k given diffuse priors—like $c_{0k} = 0$, $V_{0k} = \text{diag}(10^6)$, and $a_0 = b_0 = .001$—Bayesian and empirical Bayes estimates will be approximately equal. Results for this conjugate model with diffuse priors can be sensitive to alternative noninformative specifications for the prior variances. A uniform distribution for the square root of the variance may provide a more robust distribution for vague prior information. Hierarchical model estimates of b will have lower mean squared error and will tend to make better forecasts than MLEs, $\hat{b}_i$, that use information from just one

country. The macro-level effects, η_k, will have larger standard errors than equivalent effects under a pooled model with interactions between X_i and z_i or in a regression of $\hat{b}_i$ on z_i. The large standard errors in the multilevel model result from the heterogeneity of the $b_{i\cdot}$ across countries and can be viewed as an adjustment for country-level clustering in the data.

Arguments Against the Multilevel Model

The parameters of the multilevel model have good statistical properties. They have low mean squared error and make good predictions within the sample and for new data. Should we always fit multilevel models when we have multilevel data? No. There will be analyses in which the multilevel model is undesirable. We describe two main objections that comparative researchers should consider: the Bayesian estimates are often biased and country-level random effects may not be exchangeable.

Estimates of random effects, $b_{i\cdot}$, are biased but an unbiased estimator is readily at hand: the MLE of $b_{i\cdot}$, based just on data from a particular country. A special case arises where the effects of covariates X_i are constrained to be the same across countries, but a separate intercept is estimated for each country:

$$E(y_i) = \alpha_i + X_i b.$$

The multilevel model assigns a normal distribution to α_i, whose parameters would also be given a hyperprior in a fully Bayesian analysis. Unbiased estimates of α_i could be obtained simply by pooling the data from all countries, and using country-level dummy variables to fit I different intercepts. In the analysis of panel data, this dummy variable specification yields the so-called fixed effects model which is distinguished from the (hierarchical Bayesian) random effects model.

With these kinds of models, researchers are often interested only in the slope coefficients, b, not the intercepts, α_i. In the presence of country-level omitted variables that are correlated with X_i, the misspecified fixed effects model will still yield unbiased estimates of b. The effects of omitted variables will be absorbed by the fixed effects, α_i. In the random effects model, however, estimates of b will be biased in the presence of country-level omitted variables correlated with X_i. In the limit, as the number of countries becomes large, the random effects estimates of b will go to the least squares estimates for the model $E(y_i) = \alpha + X_i b$ that includes no country-level intercepts.

Discussions of the sensitivity of the random effects estimates to this kind of model misspecification often arises in the econometrics of causal inference (e.g., Hsiao 2003, 41–48). The fixed effects model is offered as a useful tool for causal inference because the researcher need not specify any model for cross-country variation. With a misspecified model for country-level variation, the slope coefficients, b, will have smaller variance with random rather than fixed effects. However, the random effects estimates may have larger mean squared error because of their bias.

This argument in favor of fixed effects is now carrying the day in applied research in economics and sociology (Wooldridge 2002; Halaby 2004). Still, there are good arguments for random effects, particularly for comparative research. Often comparative researchers are interested in country-level variables, z_i. The country-level variables are linearly dependent on the fixed effects so their coefficients are not identified in the fixed effects model. For all its advantages, the fixed effects model treats cross-country differences as a nuisance that threatens causal inference, rather than as a substantively interesting source of variation.

Even in this case where we are interested in the effects of country-level

variables, there is a non-Bayesian alternative. We could simply fit a model that pools data from all countries and includes the coefficients of z_i as another parameter to be estimated. If the effects of z_i differ for subgroups of each country's population, we might also include interactions between z_i and X_i. However, coefficients for the pooled model that includes just a single intercept, α, will be optimistically small compared to the random effects estimates, because the pooled model makes no adjustment for clustering. Researchers now commonly account for clustering with a sandwich estimator (Huber 1967) for the coefficient covariance matrix of these pooled models. Beck and Katz (1995) popularized this option for pooled time series in comparative politics.

A second objection to the multilevel model challenges the treatment of parameters as realizations of an underlying probability distribution. The multilevel model assumes that the coefficients $b_{\cdot k}$ are exchangeable, meaning that their joint distribution, $p(b_{1k}, b_{2k}, \ldots, b_{Ik})$, is invariant under permutations of the indices, $i = 1, \ldots, I$. Researchers must assume that after conditioning on covariates, no information is available to distinguish one b_{ik} from another. The covariates exhaustively account for all the systematic variation in the random effects. With exchangeable b_{ik}, we will not expect the random effects from any particular country to be especially large or small. Because comparative researchers often have very rich and detailed information about countries, the assumption of exchangeability may be difficult to satisfy. The multilevel analysis can at least draw the researcher into studying the heterogeneity of effects and thus invites scrutiny of the exchangeability assumption.

Should we use Multilevel Models?

There are good statistical arguments for using multilevel models but there will be situations when pooled or fixed effects models are preferable. Ideally, alternative models should be evaluated empirically and not just on a priori grounds. We need statistics that will help us decide between alternative specifications. These statistics should let us compare fixed and random effects models, often in situations where the models are not nested, where they include different sets of predictors. Comparing fixed to random effects models involves testing a null hypothesis at the boundary of the parameter space where a variance component equals zero. In many cases a random effects model will include country-level predictors whose coefficients are not identified in fixed effects models, so the researcher must compare non-nested specifications. Conventional fit statistics—likelihood ratio chi-squares or specification tests based on F-statistics—cannot be applied to this task.

Bayesian statistics are useful for evaluating and choosing between complex and non-nested models. The posterior predictive distribution provides diagnostics for assessing models as well as statistics for making inferences about the discrepancy between data and model. The posterior predictive distribution assigns probabilities to hypothetical or future values of y, written y^{rep}, integrating over uncertainty about the posterior distribution of the parameters (Gelman et al. 2004, 159–177):

$$p(y^{\text{rep}}|y) = \int \underset{\substack{\uparrow \\ \text{Sampling} \\ \text{Distribution}}}{p(y^{\text{rep}}|\theta)} \quad \underset{\substack{\uparrow \\ \text{Posterior} \\ \text{Distribution}}}{p(\theta|y)} \quad d\theta$$

The integral defining the posterior predictive distribution has two parts. The first part gives the probability density of y^{rep} given particular values of θ. The form of this density is given by the sampling distribution for y. The second part of the integral is the posterior distribution for the model parameters, θ. The posterior predictive distribution incorporates two kinds of uncertainty: sampling uncertainty about y given θ, and parametric uncertainty about θ. The posterior predictive distribution can be compared to the observed data to assess model fit. If a model fits the data well, the observed data are relatively likely under the posterior predictive distribution. On the other hand, large discrepancies between the observed data and the posterior predictive distribution indicate that the model fits poorly. The posterior predictive distribution is straightforward to simulate for Bayesian models estimated with Markov Chain Monte Carlo (MCMC) methods. Given random draws, θ^l from the posterior distribution and the sampling distribution, $p(y|\theta)$, y^{rep} can be generated by a random draw from $p(y|\theta^l)$.

The discrepancy between the model and the data can be assessed with a test statistic, $T(y)$, that summarize some substantively important feature of y. Model fit can be judged by comparing the observed $T(y)$ to the distribution of $T(y^{\text{rep}})$. A Bayesian p-value is defined by,

$$p = \Pr[T(y^{\text{rep}}) \geq T(y)|y].$$

The p-value describes, conditional on the model, the probability of observing data at least as extreme as that actually observed. An extreme value for p indicates the data are unlikely under the model. Note that the p-value does not describe the probability that a particular model is correct, nor does the p-value provide evidence against a null in favor of an alternative. Instead, small p-values indicate the implausibility of the data under the model and

the utility of examining other models (Berkhof and Hoitjtink 2000).

An alternative Bayesian approach to model evaluation uses penalized likelihood statistics of the form:

$$\text{Adequacy of a model} = \text{fit} - \text{complexity}.$$

These statistics weigh the model fit, usually measured by the log likelihood or the deviance (-2 times the log-likelihood). The fit is subject to a penalty for model complexity, often indexed by some function of the degrees of freedom, as for the Bayesian information criterion (BIC), or the number of parameters, as in the Akaike information criterion (AIC). Penalized likelihood statistics like the AIC and BIC reward simple models that fit the data well.

Penalized likelihood statistics are difficult to apply to hierarchical models because model complexity is difficult to measure. Each random effect consumes something less than a full degree of freedom. One measure of the discrepancy between the data and model that depends just on the data, y, conditions on a point estimate for θ,

$$D_{\hat{\theta}}(y) = D(y, \hat{\theta}[y]),$$

where the deviance might be evaluated at the posterior mean, $\hat{\theta} = E(\theta|y)$. Alternatively, we can account for uncertainty about the parameters by averaging over the posterior distribution for θ,

$$\bar{D}(y) = E[Dy, \theta)|y].$$

Like the posterior predictive statistics, the expectation of the deviance, $\bar{D}(y)$, lends itself to calculation with MCMC output. With draws, θ^l from the posterior distribution, we can estimate $\bar{D}(y)$ as the average of the deviance statistics. Speigelhalter and his colleagues (2002, 587) propose using the mean

deviance minus the deviance of the means as a measure of model complexity:

$$p_D = \bar{D}(y) - D_{\hat{\theta}}(y),$$

where p_D can be interpreted as the effective number of parameters in a Bayesian model. We can think of p_D as the number of unconstrained parameters in a model where a parameter scores: 1 if it is estimated with no constraints or prior information; 0 if it is fixed or fully specified by prior information; or between 0 and 1 if the parameter depends on the data and the prior (Gelman et al. 2004, 182).

A Deviance Information Criterion (DIC) is then given by:

$$\begin{aligned} DIC &= \bar{D}(y) + p_D \\ &= 2\bar{D}(y) - D_{\hat{\theta}}(y) \end{aligned}$$

where smaller values of the DIC indicate a better fitting model. The DIC can be regarded as a generalization of AIC. For non-hierarchical models the two statistics will be approximately equal. In the following application we consider a variety of models, and compare them using posterior predictive checks and DIC statistics.

An Empirical Application: Migration from Thai Villages

To illustrate the Bayesian analysis of multilevel survey data, we study data on migration for men and women aged 18 to 25, from 22 villages in the Nang Rong district of Northeastern Thailand (Curran, et al. 2005). Like the survey data analyzed in other papers in this issue, we have a relatively large number of observations from each unit (villages in our case, but countries in the other papers). The Thai village survey collects information on whether a respondent has spent longer than two months away from the village in the

survey year, 1990. (Some respondents are living away from the village at the time of the survey, and they followed up and interviewed.) Migration is coded as a binary variable where $y = 1$ if the respondent was away from the village for at least two months, and $y = 0$ otherwise. Because remittances from migrants make an important contribution to village economies, researchers are interested in the characteristics of those that leave. We predict migration in the survey year as a function of age, sex, education, whether the respondent has migrated in the past. We are also interested in the effects of village characteristics—the average level of migration in the village and village-level inequality in migration experiences.

If for respondent i $(1 = 1, \ldots, n_j)$ in village j $(j = 1, \ldots, 22)$, individual-level covariates are collected in the vector, x_{ij}, and village-level covariates are in the vector, z_j, the probability of migration, $p_{ij} = E(y_{ij})$, might be written in four alternative models:

(1) The Pooled Model:
$\text{logit}(p_{ij}) = \alpha + x'_{ij}\beta + z'_j\gamma$

(2) The Fixed Effect model:
$\text{logit}(p_{ij}) = \alpha_j + x'_{ij}\beta$

(3) The Random Intercept Model:
$\text{logit}(p_{ij}) = \alpha_j + x'_{ij}\beta + z'_j\gamma, \qquad \alpha_j \sim N(\mu_\alpha, \tau^2_\alpha)$, and

(4) The Random Slope and Intercept Model:
$\text{logit}(p_{ij}) = \alpha_j + x'_{ij}\beta_j + z'_j\gamma, \qquad \alpha_j \sim N(\mu_\alpha, \tau^2_\alpha), \beta_j \sim N(\mu_\beta, \tau^2_\beta),$

To complete the Bayesian specification, the hyperparameters—$\mu_\alpha, \tau^2_\alpha, \mu_\beta$, and τ^2_β—are all given diffuse prior distributions. In our application the hypermeans are given a prior zero with an arbitrarily large variance (10^6), and the hypervariances are given a diffuse inverse gamma distribution. The fixed effect model is a special case of the random effects model, where in the fixed effect specification, the village effects, α_j, are given a diffuse proper prior.

The fixed effects model drops the village-level effects because they are not identified by the sample data. All other parameters in all models were given diffuse proper priors. All models were estimated with MCMC methods. Slow mixing in the chain was corrected by standardizing the continuous predictors. For each model, two parallel chains were run for 50,000 iterations after a burn-in of 5,000 iterations. Diagnostics and trace plots indicated convergence.

Heterogeneity in the sample data is shown by the descriptive statistics (Table 1). The sample sizes for each village are smaller than the sample sizes for countries in comparative data sets like the ISSP or Eurobarometer. Still, enough data has been collected to estimate separate migration equations for each village. Table 1 shows large differences in gender-specific migration rates. In village 6, 57 percent of young women had been absent for at least two months in the year. In villages 10 and 15, around 70 percent of young men migrated at some point in the year. Village 10 is especially unusual, because the high level of men's migration accompanies a low level of women's migration, just 15 percent. Posterior predictive checks will show that this feature of the data contributes to poor predictions with a fixed effect model.

Table 2 reports regression results from the four migration models. The pooled model combines all the data and fits the effects of our six covariates. The model takes no account of the hierarchical structure of the data (clustered within villages) and assumes that the effects of sex, age, education, and an individual's earlier trips do not vary across villages. The pooled model provides strong evidence for gender and education differences in migration. People who have spent long periods out of their village in the past are also likely to migrate in the current year. At the village level, great inequality in migration experiences is associated with a low probability of migration

Table 1. Proportion of men and women migrating in 22 villages, Nang Rong, Thailand, 1990.

Village	Sample Size (n_j)	Proportion Migrating	
		Women	Men
1	104	.38	.58
2	82	.45	.54
3	52	.50	.63
4	80	.56	.48
5	87	.38	.48
6	40	.57	.61
7	83	.37	.33
8	75	.18	.26
9	89	.26	.45
10	89	.15	.69
11	59	.46	.48
12	44	.22	.43
13	75	.26	.27
14	50	.20	.30
15	76	.53	.70
16	117	.25	.40
17	107	.38	.59
18	59	.29	.38
19	63	.50	.59
20	102	.17	.30
21	57	.33	.43
22	96	.26	.34
Average	77	.35	.47
Minimum	40	.17	.26
Maximum	117	.57	.70

Note: To maintain the confidentiality of the data, villages are assigned arbtirary identification numbers.

Table 2. Posterior means of parameters in four models of migration in 22 Thai villages. (Posterior standard deviations in parentheses.)

	Pooled (1)	Fixed Effect (2)	Random Intercept (3)	Random Slope and Intercept (4)
Intercept	-.80	-.66	-.79	-.76
	(.12)	(.10)	(.13)	(.14)
Male	.32	.38	.33	.30
	(.14)	(.14)	(.14)	(.15)
Age	-.13	-.15	-.13	-.13
	(.08)	(.08)	(.08)	(.09)
Years of education	.39	.38	.39	.39
	(.07)	(.08)	(.07)	(.08)
Prior trips (individual)	1.08	1.13	1.08	1.22
	(.09)	(.10)	(.09)	(.16)
Prior trips (village)	-.61	-	-.59	-.69
	(.37)		(.43)	(.46)
Gini of prior trips (village)	-.62	-	-.60	-.61
	(.21)		(.24)	(.25)
DIC	1259.41	1273.29	1259.62	1247.55
p_D	7.02	26.24	10.93	30.58

among individuals. This effect of migration inequality is not simply an artifact of a low general level of migration in highly unequal villages, because the village mean migration rate is controlled.

This pattern of results is stable across different models. The age, gender, and education effects are insensitive to whether village heterogeneity is modelled with fixed or random effects. The effect of an individual's migration history is estimated to be about 10 percent large under the random slope model than under the other models, and uncertainty about this larger effect is also relatively large. Although the village-level coefficients are not identified by the sample data for the fixed effects model, the random coef-

ficient models—like the pooled model—offer strong evidence that migration is less likely in villages where only a few residents have substantial migratory experience. As we would expect, standard errors are generally larger in the multilevel models than the pooled model because of the additional heterogeneity due to the random effects.

In this application, the performance of any one model is not clearly superior to the others. Are there other empirical criteria we might use to decide between the models? The DIC statistic provides a way of comparing these non-nested and hierarchical models. The measure of model complexity, p_D, shows that the pooled model is the most parsimonious. Although there are only 7 effects to be estimated, the effective number of parameters, p_D is slightly larger because of the small influence of prior information on the posterior distribution of the regression coefficients. The random intercept model is only slightly more complex although the model adds an intercept term for each village in the sample. The most complex model includes random slopes and intercepts, a total of 110 random effects for the 22 villages. Because of its hierarchical structure, the random slopes model is only slightly more highly parameterized than the fixed effects specification that adds a fixed intercept term for each village. The DIC statistic of 1247.55 for the random slope model is about 12 points lower than for either the pooled model or random intercept model. The DIc indicates that fixed effects model does not yield a sufficient improvement in fit to justify its parametric complexity.

The models can also be assessed with posterior predictive checks. This approach yields multilevel diagnostics, similar in spirit to the "HLM EDA" recommended elsewhere in this issue. Calculating posterior predictive statistics first involves defining a test statistic that records discrepancies between a model and the sample data. We take the sex-specific village mean, $T(y) = \bar{y}_{jk}$

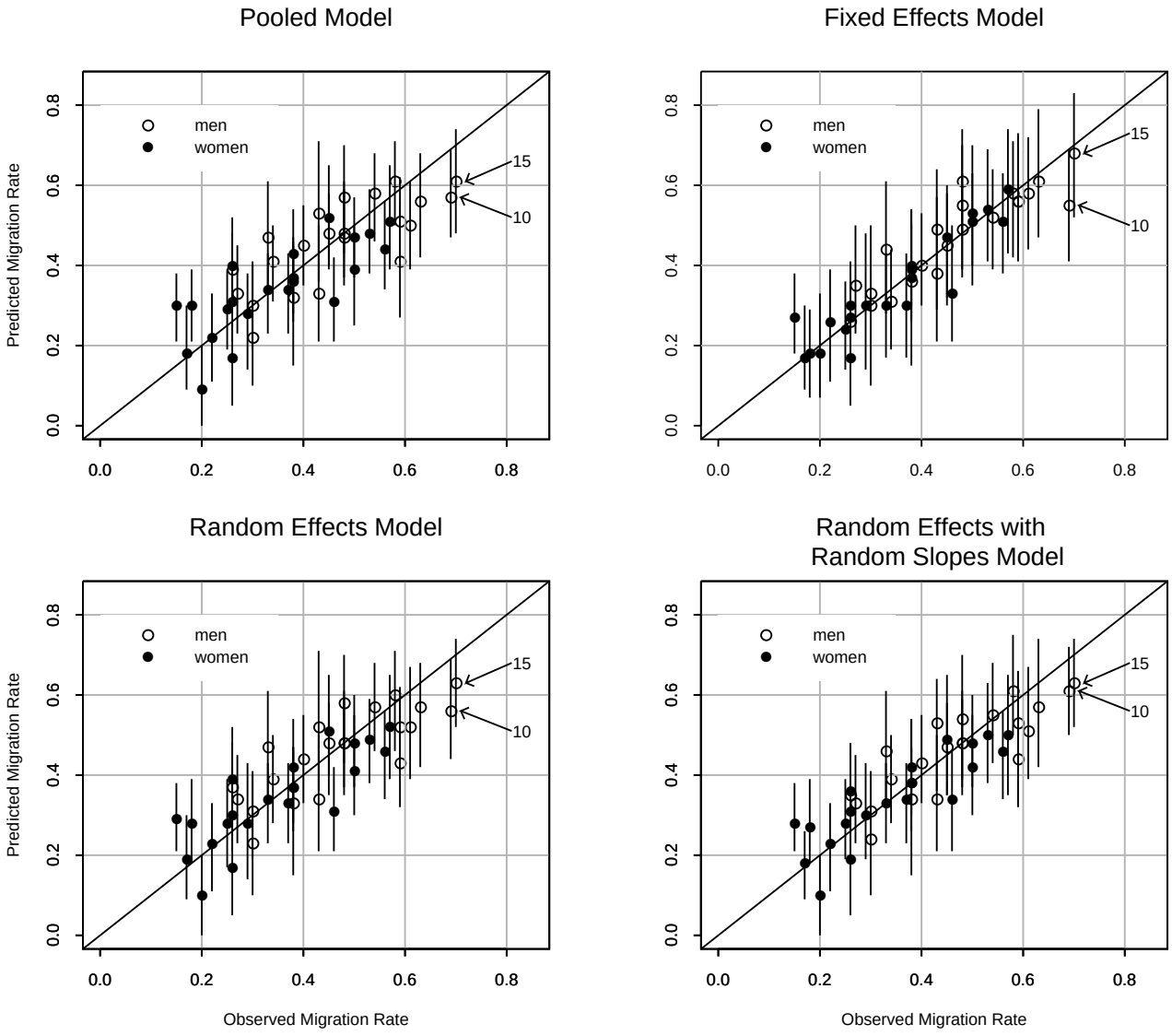

Figure 1. Observed and predicted proportions of men and women, migrating from 22 villages in Nang Rong district, Thailand. Vertical lines indicate 80 percent confidence intervals. Villages 10 and 15 are indicated on the plot.

(for village $j = 1, \ldots, 22$, and k =male or female), to assess the adequacy of the model. (The village mean might be an alternative test statistic but in this case the fixed effect specification provides a model of the migration rate within villages.) The posterior predictive distribution yields a distribution for the test statistic that can be compared to the sample data.

Figure 1 shows the mean levels of migration for men and women in each of the 22 villages. The horizontal axis shows the observed means and the vertical axis shows the means predicted by the different models. Vertical lines show an 80 confidence interval. If the model accurately predicts the observed mean, the confidence interval will cross the 45 degree line. The predictions from the pooled model are generally more dispersed than those from the other specifications that allow for heterogeneity across villages. Predictions for the fixed effect model are much more tightly clustered, often falling close to the 45 degree, indicating a close match between model and data. Predictions from the two random effects models are slightly more dispersed than under the fixed effects model. Predictive performance varies noticeably in the two villages with high men's migration rates. Villages 10 and 15, where 70 percent of the men report absences of at least 2 months, are quite accurately predicted by the random slope model. By contrast, the fixed effect model only accurately predicts men's migration rates for village 5. The fixed effect model does more poorly for village 10, where women's migration is unusually low. Indeed the predictive interval for the fixed effect model indicates that the high rate of men's migration in village 10 is highly unlikely.

Conclusion

This paper has extended the use of Bayesian statistics in comparative research to consider survey data in which samples from each country are large enough

to estimate the parameters of a regression. Bayesian hierarchical models provide an appealing framework for data with this structure. The Bayesian estimates have lower mean squared error than either the pooled estimate that combines data and fits just one effect for all countries, or the unpooled (fixed effect) estimate that fits one effect for each country. Lower mean squared error of the Bayes estimator is reflected in better within sample and out-of-sample predictions than those of either the pooled or unpooled estimators.

Despite these desirable properties, the Bayesian hierarchical model yields biased estimates in the presence of country-level omitted variables that are correlated with predictors. If causal inference is the main objective, researchers may prefer a fixed effect estimator that is unbiased for this specific model misspecification. In comparative analysis, however, we will often be interested in studying the effect of country-level variables. In the absence of good instruments, the effects of country-level variables will not identified in fixed effects models. If interest centers on these country-level variables, the researcher will have little choice but to fit a pooled model or a Bayesian hierarchical model. Because standard errors of the pooled model will be too small in the presence of country-level clustering, the Bayesian model is preferable.

Given the advantages and disadvantages of Bayesian models for comparative survey data we have argued that the merits of different models should be assessed empirically. Statistics for model evaluation should ideally be able to compare models with fixed and random effects, and models that are non-nested. Two Bayesian approaches provide some empirical standards for model comparison and evaluation. A deviance-based measure, p_D, provides a useful way of comparing the parametric complexity of fixed and random effects models. A penalized likelihood approach motivates the DIC statistic that weighs goodness of fit and parametric complexity in assessing the ade-

quacy of a model. Posterior predictive checks allow the construction of test statistics that can be tailored to measure substantively important patterns of variation that arise in specific applications. In our analysis of the Thai migration data, we found that estimates from a fixed effects model were often very accurate, but one extreme data point was more accurately predicted with random effects.

In sum, Bayesian methods provide useful tools for analyzing multilevel survey data in comparative research. In applications, however, a number of non-Bayesian alternatives will often be available. We recommend an empirical approach to the problem of model evaluation in which preferred models yield not just significant coefficients for our favorite hypotheses, but also capture substantively important patterns of variation in the data.

References

Berk, Richard A., Bruce Western, and Robert Weiss. 1995. "Statistical Inference for Apparent Populations" (with discussion). *Sociological Methodology* 25:421–485.

Beck, Nathaniel and Jonathan N. Katz. 1995. "What to do (and not to do) with Time-Series Cross-Section Data." *American Political Science Review* 89:634-647.

Berkhof, Johannes, Iven van Mechelen and H. Hoijtink. 2000. "Posterior predictive checks: Principles and discussion." *Computational Statistics* 15:337–54.

Carlin, Bradley P. and Thomas A. Louis. 2000. *Bayes and Empirical Bayes Methods for Data Analysis*. Second Edition. New York: Chapman Hall.

Curran, Sara R., Filiz Garip, Chang Chung, and Kanchana Tangchonlatip. 2005. "Gendered Migrant Social Capital: Evidence from Thailand." *Social Forces*.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. rubin. 2004. *Bayesian Data analysis*. Second edition. New York: Chapman Hall.

Halaby, Charles N. 2004. "Panel Models in Sociological Research: Theory into Practice." *Annual Review of Sociology* 30:507-544.

Hsiao, Chen 2004. *The Analysis of Panel Data*. Second edition. New York: Cambridge University Press.

Huber, Peter J. 1967. "The Behavior of Maximum Likelihood Estimates Under Non-Standard Conditions." *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 1:221–233.

Spiegelhalter, David J., Nicola G. Best, Bradley P. Carlin, and Angelika van der Linde. 2002. "Bayesian Measures of Model Complexity and Fit" (with discussion). *Journal of the Royal Statistical Society* Series B 64:583–639.

Western, Bruce and Simon Jackman. 1994. "Bayesian Inference for Comparative Research." *American Political Science Review* 88:412–423.

Western, Bruce and Meredith Kleykamp. 2004. "A Bayesian Change Point Model for Historical Time Series Analysis." *Political Analysis* 12:354–374.

Wooldridge, Jeffrey. 2002. *Econometric Analysis of Cross Section and Panel Data*. Boston, MA: MIT Press.