# Bayesian Posterior Predictive Checks for Complex Models

SCOTT M. LYNCH
BRUCE WESTERN
*Princeton University*

*In sociological research, it is often difficult to compare nonnested models and to evaluate the fit of models in which outcome variables are not normally distributed. In this article, the authors demonstrate the utility of Bayesian posterior predictive distributions specifically, as well as a Bayesian approach to modeling more generally, in tackling these issues. First, they review the Bayesian approach to statistics and computation. Second, they discuss the evaluation of model fit in a bivariate probit model. Third, they discuss comparing fixed- and random-effects hierarchical linear models. Both examples highlight the use of Bayesian posterior predictive distributions beyond these particular cases.*

***Keywords:*** *Bayesian posterior predictive distributions; Bayesian approach; statistics; model fit*

Model diagnosis and comparison is often a difficult problem in quantitative research. Sometimes, a statistical model is fit to capture a substantively important feature of a data set, but goodness-of-fit statistics are uninformative about this theoretically important variation. For example, say we are interested in predicting a very rare event. The usual chi-square or $R$-square statistics may appear to be reasonable because the model can fit most data well by predicting that the event never occurs. In other cases, researchers may need to to compare nonnested models. In a regression analysis of panel data, for instance, we may want to choose between fixed- and random-effects

AUTHORS' NOTE: *Parts of this article were presented at the meetings of the Methodology Section of the American Sociological Association, Princeton, New Jersey, March 2002. We thank* SMR *reviewers and Scott Eliason for helpful comments on an earlier draft. Western's research was supported by grants from the Russell Sage Foundation and National Science Foundation grant SES-0004336.*

SOCIOLOGICAL METHODS & RESEARCH, Vol. 32, No. 3, February 2004 301-335
DOI: 10.1177/0049124103257303

models with different sets of covariates. There are no standard statistics to aid model choice in this situation.

In this article, we describe diagnostics for assessing model fit based on predictions generated by the model. Predictions inconsistent with the observed data provide evidence against the model assumptions. On the other hand, we would be disinclined to reject a model whose predictions resemble the data. To obtain the diagnostics, predictions are drawn from the Bayesian posterior predictive distribution. Comparing the predictive distribution to the observed data is generally termed a *posterior predictive check* (Rubin 1984:1165-71; Gelman et al. 1995:167-74). This approach has three advantages over standard applications of fit statistics. First, an extremely wide range of fit statistics can be defined, based on the distribution of predictions under a model. Researchers need not be confined to various forms of residual sums of squares. Second, the Bayesian basis for the statistic allows the calculation of $p$ values, describing the probability that the data arose by chance, given the model assumptions. Third, posterior predictive simulation explicitly accounts for the parametric uncertainty that is usually ignored by alternative approaches.

To describe and illustrate the application of posterior predictive checks, we briefly review the Bayesian approach to statistical inference. We then discuss posterior predictive distributions, describing how they can be simulated using Bayesian Markov chain Monte Carlo methods for posterior simulation. We then describe two applications—a bivariate probit model of causes of death and fixed- and random-effects models of wages.

## *REVIEW OF BAYESIAN PARADIGM*

Bayesian statistics has received considerable attention in statistics over the past decade. The Bayesian approach has made less headway, however, into mainstream sociological research, largely because of discomfort with Bayesian prior distributions and the seemingly steep learning curve required to perform Bayesian computation. Our goal in this study is to provide a brief introduction to Bayesian computation and show how these computations assist in calculating Bayesian posterior predictive distributions for assessing model fit and in comparing rival models. As we will see, there are numerous instances

in which classical methods fail, but the Bayesian approach offers a feasible method for assessing model fit.

The fundamental difference between Bayesian and more familiar likelihood approaches to statistics rests in the use of a prior distribution. Bayes's theorem states that

$$p(B|A) = \frac{p(A|B)p(B)}{p(A)}.$$

The theorem is a trivial result of applying a few basic probability axioms. However, substituting "parameters" for $B$ and "data" for $A$, essentially elevating data to the same level as parameters, incurs criticism. Upon substitution, the theorem states that a "posterior distribution" for parameters, after considering the data at hand, $p(B|A)$, is a function of a "prior" distribution, $p(B)$, representing prior uncertainty about the true value of the parameter and the likelihood function, $p(A|B)$. Although the latter term in the equation does not appear to be a likelihood function, standard likelihood analysis is based on the latter term: the sampling density for the data, given the parameter values. In likelihood analysis, the sampling density is rearranged so that the data are essentially treated as fixed and the parameter values that maximize the probability of observing the data are estimated. For Bayesians, both the parameters and the data are considered random quantities to which a probability distribution can be attached.

Most of the dispute between Bayesian and likelihood approaches arises from the inclusion of the prior distribution: It is argued that the prior injects too much subjectivity into the modeling process and may drastically affect the results. However, there are several responses to this criticism:

1. Priors can be diffuse (i.e., provide little or no information in the process of estimation).
2. Maximum likelihood estimates are a special case of the Bayesian paradigm in which all prior distributions are uniform.
3. Priors are generally asymptotically irrelevant (i.e., as $n$ gets large—as is typically the case in social science data sets—the prior has less effect on the results).
4. Examining model fit can reveal problems with unreasonable or unrealistic prior distributions.

5. Priors can be used to generate conservative tests (i.e., they can be conservative and force the data to overturn them).
6. Priors allows research to build on previous research rather than the typical approach of assuming ignorance and testing a null hypothesis that is uninteresting and often assumed to be untrue anyway.

In this article, we assume that one has decided to conduct a Bayesian analysis and relegate further discussion of anti-Bayesian arguments to other sources (e.g., Gelman et al., 1995).

## *BAYESIAN COMPUTATION*

A more serious problem with the inclusion of a prior distribution is that it tends to make each model unique and hence makes estimation the key hurdle to completing a Bayesian analysis. Furthermore, because the marginal distribution of the data (the denominator in the equation above) tends to be a complicated integral (the marginal for the data is the sum or integral of the probability of the data under all values of the parameter in the parameter space), it is often extremely difficult to compute. However, because the parameter does not depend on this denominator, Bayes's theorem is often reduced to

$$posterior \propto prior \times likelihood.$$

While this expression is theoretically easier to consider, it is nonetheless difficult to simulate from, and simulation of values from the posterior distribution is a primary goal of a Bayesian analysis.

Most of the advances in Bayesian methods in recent years have been in the area of model estimation. Many Bayesian models are analytically intractable. Thus, prior to 1990, estimation often rested on (multivariate) normal approximations to the posterior distribution—which may be suspect—and quadrature and other numerical methods that exceed the computational skills of most sociologists. Since 1990, however, the explosion in computing capability and the development of Markov chain Monte Carlo (MCMC) methods have enabled the computation of high-dimensional models that previously had been impossible, or nearly so, even for Bayesians. These methods have also allowed the generation of the techniques of investigating posterior

predictive distributions, which we discuss here, typically as a simple extension of the estimation algorithm.

The standard methods of MCMC estimation include the Gibbs sampler and the more general Metropolis-Hastings (MH) algorithm. The Gibbs sampler is a special case of the MH algorithm in which parameters are sampled from their full conditional distributions (rather than a relatively arbitrary proposal distribution). The typical MH algorithm consists of five steps:

1. Establish starting values for parameter(s) ($\theta_{j=0}$) (at maximum likelihood estimates, guesses, 0, or some other values).
2. Sample "candidate" parameter(s) ($\theta_c$) from a "proposal" distribution $\alpha(\theta)$. The "candidate" is so called because it is not automatically accepted as a draw from the posterior (see the next step). The "proposal" is so called because it is simply a distribution for generating candidates. The proposal generally can be any distribution from which it is easy to simulate draws.
3. Evaluate the posterior density at the candidate point and previous point and form the ratio $R = \frac{p(\theta_c)\alpha(\theta_{j-1}|\theta_c)}{p(\theta_{j-1})\alpha(\theta_c|\theta_{j-1})}$. The $\alpha(a|b)$ terms represent the probability that $a$ would be proposed as a value, given the chain is currently in state $b$, providing a way to compensate for asymmetric proposal densities. If the proposal density is symmetric, these terms drop from the ratio. For example, we may use a normal proposal that is centered over the previous value of the parameter, so that $\theta_c = \theta_{j-1} + N(0, d)$. That is, the candidate is a normal random variable with variance $d$ added to the previous value of the parameter. In this case, the proposal is symmetric because the probability of proposing $\theta_c$, given $\theta_{j-1}$, will be equal to the probability of proposing $\theta_{j-1}$ in the next iteration if the chain moves to $\theta_c$.
4. Accept the candidate parameter with probability $\min(1, R)$ by drawing a $U(0, 1)$ deviate and comparing it to $R$: If $R > U$, then accept the candidate ($\theta_j = \theta_c$); otherwise, reject it (set $\theta_j = \theta_{j-1}$).
5. Return to Step 2 and repeat until an adequate number of draws are obtained.

The prescribed algorithm generates a Markov chain of iterates that constitute a random walk over the posterior distribution, the results of which are equivalent to integrating the density over the data to obtain a sample from the posterior distribution for the parameters.

The method is very flexible, allowing various choices for proposal densities (indeed, if the proposals are full conditionals, the

algorithm is a "Gibbs" sampler, which accepts every candidate), sequential updating of parameters ("random-walk metropolis"), and mixing of Gibbs and metropolis steps for simulating different parameters in the model. For proof of the algorithm's convergence, see Gilks, Richardson, and Spiegelhalter (1996) in general and Tierney (1996) specifically.

At the completion of a run of an MH algorithm (or Gibbs sampler), initial iterates (prior to convergence) should be discarded (the "burn-in"), and simple descriptive statistics (e.g., mean, median, mode, variance, interquartile range, etc.) for the remaining iterates can be used to summarize posterior knowledge about the parameter. Furthermore, the interpretation of these results is more straightforward than the standard interpretation of likelihood-based confidence intervals and $t$ tests (given acceptance of the Bayesian paradigm of thinking). For example if 95 percent of the iterates for a parameter fall between 0 and 2, one can say that we are "95 percent certain that the parameter is between 0 and 2." A similar argument holds for hypothesis testing purposes: If the hypothesis were that the parameter was between 0 and 5, and 95 percent of the iterates fell in this region, we could say that the probability that the hypothesis were true was .95.

## *POSTERIOR PREDICTIVE DISTRIBUTIONS*

Aside from the descriptive statistics that can be produced from an MCMC-generated sample, other statistics can be generated to produce inference about other key quantities of interest and to assess model fit. One such quantity is the posterior predictive distribution:

$$p(y^{\text{rep}}|y) = \int \underbrace{p(y^{\text{rep}}|\theta)}_{\substack{\uparrow \\ \textit{Sampling} \\ \textit{Distribution}}} \underbrace{p(\theta|y)}_{\substack{\uparrow \\ \textit{Posterior} \\ \textit{Distribution}}} d\theta,$$

where $y^{\text{rep}}$ denotes replicated or hypothetical values of $y$. We can think of $y^{\text{rep}}$ as values that might have been observed if the conditions generating $y$ were reproduced. The integral defining the posterior predictive distribution has two parts. The first part gives the probability density of $y^{\text{rep}}$, given particular values of $\theta$. The form

of this density is given by the sampling distribution for $y$. The second part of the integral is the posterior distribution for $\theta$, $p(\theta|y) \propto l(\theta; y)p(\theta)$, where $l(\theta; y)$ is the likelihood. The posterior predictive distribution thus reflects two kinds of uncertainty: sampling uncertainty about $y$ given $\theta$ and parametric uncertainty about $\theta$.

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 should be 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 performs poorly.

The use of $y^{\text{rep}}$ to assess model fit is flexible and can be extended considerably. Aside from visual inspection of predictive simulations, discrepancy statistics can be computed that also help evaluate model fit. These discrepancy statistics can be chosen to capture substantively important features of the data. For example, in a one-sample problem, we may be interested in how well our model fits remote values. In this case, our discrepancy statistic might take the following form:

$$T(y) = y_{\text{min}} \quad \text{or} \quad T(y) = y_{\text{max}},$$

where $y_{\{}\text{min}\}$ and $y_{\{}max\}$ are the minimum and maximum observed values of $y$. Note that choosing sufficient statistics—such as the mean for the normal model—yields trivial results because the model fits these statistics exactly. In different applications, different patterns of variation will be substantively important, and different discrepancy statistics can be defined. We discuss this issue further in the applications below.

Model fit statistics can be assessed by comparing the observed $T(y)$ to the distribution of $T(y^{\text{rep}})$. A Bayesian $p$ value can be defined as

$$p = \Pr(\mathrm{T}(y^{\text{rep}}) \geq T(y)|y),$$

which can be interpreted as, conditional on the model, the probability of observing data at least as extreme as that actually observed. An extreme value for $p$ indicates that the data are unlikely under the model. Care should be taken in interpreting the $p$ value. In particular, the statistic 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 reflect the implausibility of the data under the model (and hence the lack of fit of the model to the data) and therefore suggest examining other models (Berkhof and van Mechelen 2000). The choice of $p$ value considered small enough to merit rejection of a model is a subjective determination (as any $p$ value criterion) but may be made based on conventional standards (e.g., $p < .05$, $p < .01$, etc.).

For the Bayesian $p$ value, the data are treated as fixed, and the distribution on $T(y^{\text{rep}})$ depends partly on uncertainty about model parameters, $\theta$. In contrast, the classical $p$ value treats the data as random, with parameters fixed at values given by a null hypothesis. Parameters might also be introduced into the definition of the Bayesian discrepancy statistic. Thus, Gelman, Meng, and Stern (1996) suggest an omnibus measure of goodness of fit based on a Bayesian residual sum of squares:

$$\chi^2(y, \theta) = \sum_i^n \frac{[y_i - E(y_i|\theta)]^2}{Var(y_i|\theta)}.$$

In the examples below, we use discrepancy statistics based just on the observed data, emphasizing the applied importance of calculations designed to capture substantively important patterns of variation.

With simulated values $\theta^*_{(j)}$ $(j = 1, \ldots, J)$ from the posterior distribution, yielding replicated discrepancy statistics, $T^j(y^{\text{rep}})$, the Bayesian $p$ value can be estimated by

$$\hat{p} = \frac{\#[T^j(y^{\text{rep}}) \geq T(y)]}{J},$$

the proportion of simulated discrepancy statistics at least as large as the observed value, $T(y)$.

To make these ideas concrete, we illustrate the use of posterior predictive checks in a linear regression, $y = X\beta + e$, where the error variance, $\sigma^2$, is known and the dependent variable, $y$, conditionally follows a normal distribution. With diffuse prior information, the posterior distribution of the vector of regression coefficients will be normal with posterior,

$$\beta \sim N(\hat{\beta}, \sigma^2(X'X)^{-1}), \tag{1}$$

**TABLE 1: Data and Predictive Statistics From a Bivariate Regression**

| *Observations* | $x$ | $y$ | $\hat{y}$ | $\hat{y} \pm 2s$ | *95% CI of* $y^{\text{rep}}$ |
|---|---|---|---|---|---|
| 1 | 1 | 0 | 1.01 | –3.37, 5.40 | –3.87, 5.89 |
| 2 | 2 | 0 | 1.89 | –2.50, 6.28 | –3.05, 6.46 |
| 3 | 3 | 1 | 2.77 | –1.62, 7.15 | –1.63, 7.45 |
| 4 | 4 | 3 | 3.64 | –0.74, 8.03 | –0.64, 8.14 |
| 5 | 3 | 4 | 2.77 | –1.62, 7.15 | –1.71, 7.17 |
| 6 | 5 | 2 | 4.52 | 0.13, 8.91 | 0.09, 9.23 |
| 7 | 4 | 6 | 3.64 | –0.74, 8.03 | –0.58, 8.37 |
| 8 | 6 | 7 | 5.40 | 1.01, 9.79 | 0.95, 9.77 |
| 9 | 5 | 8 | 4.52 | 0.13, 8.91 | 0.34, 8.88 |
| 10 | 7 | 7 | 6.28 | 1.89, 10.66 | 1.86, 10.95 |
| 11 | 5 | 6 | 4.52 | 0.13, 8.91 | 0.02, 9.26 |
| 12 | 9 | 5 | 8.03 | 3.64, 12.42 | 2.74, 13.21 |

NOTE: Column (5) gives the interval based on the standard error of the residuals, s.

where $\hat{\beta}$ is the least squares estimate, $(X'X)^{-1}X'y$. In this case, the posterior distribution has a known form, facilitating a one-step Gibbs sampler. If we take random draws from the normal distribution in (1), yielding $\beta^*_{(j)}$ $(j = 1, \ldots, J)$, the replicates are obtained as random draws,

$$y^{\text{rep}}_{(j)} \sim N(X\beta^*_{(j)}, \sigma^2).$$

Posterior predictive statistics for a bivariate regression with 12 observations are shown in Table 1. Posterior predictive statistics are generated by taking 1,000 random draws from the posterior distribution for the coefficients given in equation (1). The posterior predictive interval for each of the 12 observations is contrasted with the classical interval for $\hat{y}$, based on the residual standard error. The Bayesian predictive interval is wider than the classical interval because it reflects both parametric and sampling uncertainty, whereas the classical interval reflects only sampling uncertainty.

If we take as our discrepancy statistics the minimum and maximum values of $y$, we can compare the distribution of $T(y^{\text{rep}})$ to the observed $T(y)$ (Figure 1). The minimum can be written as $T(y) = (y_{(1)} + y_{(2)})/2$, the average of Observations 1 and 2. The maximum is $T(y) = y_{(9)}$, Observation 9. The top panel of the figure shows that the model predictions tend to exceed in the negative

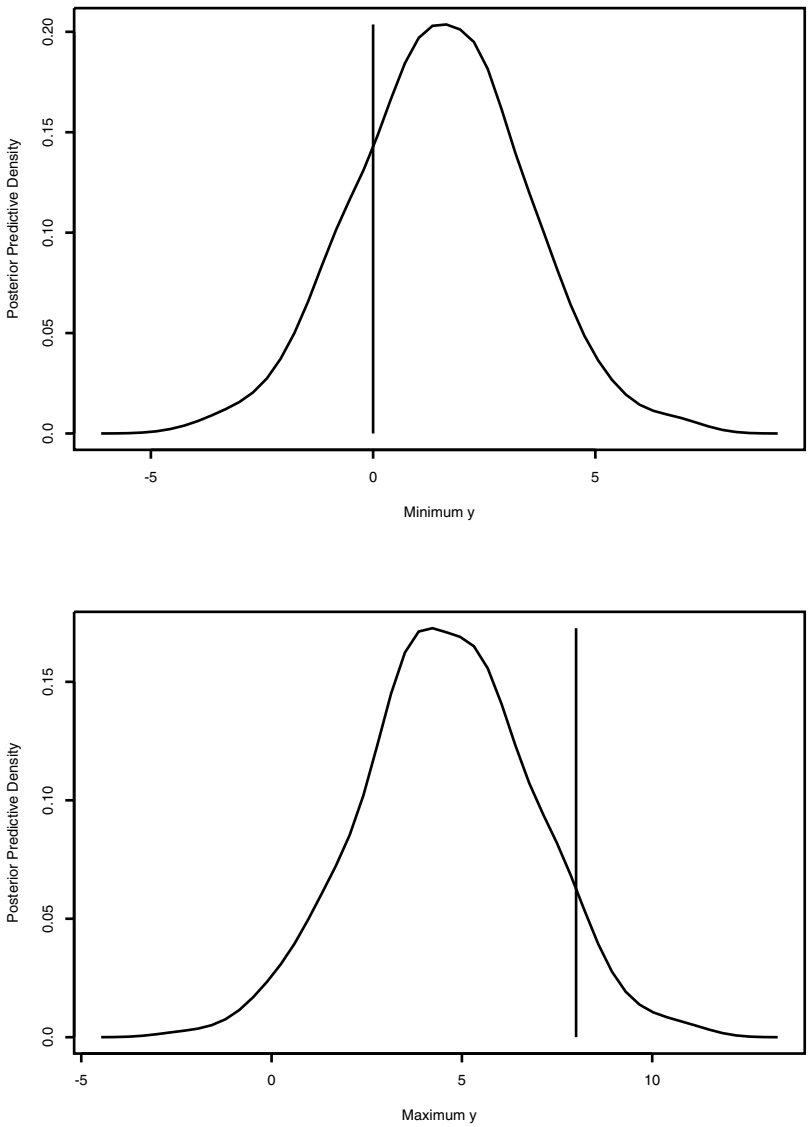


**Figure 1: Posterior Predictive Densities for the Regression Model of Data in Table 1**

NOTE: The top panel shows the predictive density for $(y_{(1)} + y_{(2)})/2$. The bottom panel shows the predictive density for the maximum of $y_{(9)}$.

direction the observed minimum, $y = 0$. With 209 out of 1,000 simulated values of $T(y^{rep})$ ($p = .209$), the data appear quite plausible under the model. However, evidence against the model is given by the sample maximum. The posterior predicted value of $y_{(9)}$ tends to be too small. The posterior predictions of the maximum cluster around 4, but the observed maximum is $y = 8$. In this case, the Bayesian $p$ value is just .062.

This example is extremely simple. The following applications apply Bayesian posterior predictive checks to more complex and realistic models.

## *EXAMPLE 1: BIVARIATE PROBIT MODEL FOR COMPETING RISKS OF CAUSE OF DEATH*

Competing risks models are used in demography to examine outcomes—especially death, in which an individual can succumb to only one outcome—although there are multiple outcomes that could claim the individual (Preston, Heuveline, and Guillot 2001). Often, these outcomes are modeled independently in a discrete time format in which the individual becomes a "censored" case in model $j$ when she or he experiences an event other than $j$ (see Allison 1984). This approach, however, assumes that the regression equations for each outcome are independent, but that is an empirical question. Ideally, we should model the outcomes in a multivariate model, allowing error correlation to exist between equations.

Assume that $n$ individuals alive at time $t$ can experience $k$ different outcomes by $t + m$. The possible transitions into these outcome states can be represented by a multinomial variable of dimension $k - 1$, with each dimension having binary response (1 = experienced this transition or 0 = did not). A 0 in all dimensions indicates no transition, and no one can have a 1 on more than one outcome (for simplicity in this example). A likelihood function can be established based on this multinomial variable, in which the latent propensities for individuals to achieve a 1 in any dimension (assumed to be $N(0, 1)$ in the probit model) are integrated out in the likelihood function.

An alternative strategy is to assume that individuals have latent propensities to die and to bring these propensities into the estimation

strategy. We can assume that these propensities to die from any one of $k - 1$ causes are distributed as follows:

$$Z_i \sim TruncMVN\,(\mu, \Sigma)\,.$$

Here, $Z_i$ is a vector of propensities for the $i$th individual. The propensities are distributed as a truncated multivariate normal variable (MVN), with the mean for each dimension ($a$) determined by a regression on a vector of covariates: $Z_i(a) = X(a)'b(a), \forall i, a$. The truncation point is defined at 0 (as in a standard probit model), with the individual's scores drawn from the cell region of the multinomial contingency table representing the observed transition. The dimensionality is $k - 1$ because if the individual lives, he or she will take a 0 in all outcome dimensions. $X'(a)$ is $1 \times q$; $b(a)$ is $q \times 1$, where $q$ represents the number of covariates used in modeling the response in dimension $a$.

These propensities are related to each other via the covariance matrix, $\Sigma$, which allows error correlations between the propensities that are not accounted for by the regression model. These correlations may exist and will likely be highly negative because of the structural aspect of the model that renders an outcome vector with more than one "1" impossible. Approaching a competing risk model from this perspective is more efficient, and it also allows for construction of multiple decrement (and multistate) life tables with empirical confidence intervals on state expectancies using the joint posterior distributions for all parameters (see e.g., Lynch, Brown, & Harmsen, 2003).

With a completed MVN set of propensities, the augmented likelihood function for the other parameters in the model becomes a multivariate normal likelihood:

$$L(b, \Sigma|Z) \propto \prod_{i=1}^{n} |\Sigma|^{-1/2} \exp\left(\omega' \Sigma^{-1} \boldsymbol{\omega}\right),$$

where the vector, $\boldsymbol{\omega}$, is a $k - 1$ dimensional residual, and $b$ and $\Sigma$ are regression coefficients and error covariances.

Here, we use a hybridized Gibbs/MH algorithm, specified as follows:

1. Simulate the latent data from $TruncMVN$ as discussed above. An individual $Z$ vector is drawn until it satisfies the truncation requirements discussed above.
2. Given these latent data, the likelihood function is now the multivariate normal likelihood discussed above—conduct MH steps to update $b$ and the components of $\Sigma$ over several additional iterations.
3. Return to Step 1.

This algorithm varies from a simple MH algorithm in that we have included a Gibbs sampling step (Step 1) in which the latent data are simulated rather than integrated out of the posterior distribution. A benefit to this approach is that it allows for greater leverage in diagnosing outliers, influential cases, and other problems. A limitation of standard analyses of discrete data is the relatively poor diagnostic measures available because many diagnostic measures assume normally distributed errors. Normal errors are impossible to obtain when the data are discrete, and so most diagnostic measures are ad hoc contortions of measures for linear models. In this model, on the other hand, pseudo-errors (latent residuals) can be obtained from the simulated distributions of the latent scores for each individual. Our focus here is on posterior predictive quantities, and so we do not emphasize this aspect of model fit.

After a completed MCMC run, we use the post-burn-in simulated iterates to simulate latent data for each observation. Unlike frequentist analyses, which assume that the parameter is fixed at its maximum likelihood estimation (MLE), the Bayesian approach explicitly accounts for parametric uncertainty in this fashion. Posterior simulation is accomplished by simulating $j$ vectors $Z \sim MVN(\mu, \Sigma)$, where $\mu(a) = X(a)'b(a)_j$ and $j$ references the simulated values of $b$ from the MCMC run. After this simulation, individuals are allocated to the appropriate multinomial cell based on their vector of latent traits and the known thresholds of 0 in each dimension. In this example, we assume two classes of causes of death: circulatory system diseases (CSD) and all other causes of death (OCD). An individual with vector $Z = [<0, <0]$ would be classified as a "survivor," an individual with a vector $Z = [>0, >0]$ would be double classified as dying from both CSD and OCD, an individual with a vector $Z = [<0, >0]$ would be classified as dying from OCD, and an individual with a vector $Z = [>0, <0]$ would be classified as dying from

CSD. In conducting this simulation, we do not impose the constraint that individuals cannot be double classified—if the model is specified appropriately (and, more specifically, error correlation is estimated appropriately), double classification should not occur.

The data for this example are from the National Health and Nutrition Examination Survey (NHANES) and its follow-ups (the National Health Epidemiologic Followup Survey—NHEFS). The baseline sample of 34,000 persons ages 25 to 74 was surveyed in 1971. A total of 14,400 individuals who were administered a detailed medical exam at baseline were administered follow-up questionnaires in 1982, 1987, and 1992. Of these, we use persons who were administered the health care supplement and were older than age 65 at baseline ($n = 1, 231$).

At the end of the survey period, a vital and tracing status file was compiled that included date and cause of death and final interview status. We excluded 29 (2.4 percent) individuals who did not have a final interview and whose status was unknown, leaving us with 1,202 respondents measured over a 22-year period. We use this information to construct a person-year data file encompassing the period 1971-1992 (final $N = 18, 005$).

Individuals surviving through the end of the period are treated as censored, while individuals who died ($n = 909$) were coded as dying from either CSD ($n = 497$) or OCD ($n = 412$), based on ICD9 codes from their death certificates (range: 390-459). CSDs include cardiovascular problems and events (e.g., myocardial infarction) and cerebrovascular problems and events (e.g., stroke), as well as structural defects of the heart, heart and renal problems induced by hypertension, embolisms, aneurysms, pulmonary circulation problems, peripheral vascular problems, and so forth. Many of these causes are related to risk factors (e.g., smoking, age, diet, race), and most individuals dying from these causes cluster around ischemic heart disease and stroke, two of the leading causes of death for adults in the United States.

We include age (baseline $M = 68.98$, $SD = 2.82$, range = 65-77), sex (male = 1, 48.7 percent), race (non-White = 1, 15.3 percent), southern residence (South = 1, 31.3 percent), baseline marital status (married = 1, 63.9 percent), smoking history (0 = nonsmoker, 51 percent; 1 = former smoker, 29.0 percent; 2 = current smoker,

20 percent), and education (in years, $M = 9.43$, $SD = 3.6$, range = 1-17+) as covariates predicting both outcomes. We expect age, race, gender, southern residence, and smoking to have positive effects on death due to either cause, while we expect marriage and education to be beneficial. We expect the effect of age to be stronger for CSDs than for OCDs because heart disease and stroke are the first and third leading causes of death.

Table 2 presents the results of the run of the model. The coefficients reported in the table are the posterior means for the bivariate probit coefficients, and they have an interpretation identical to those of a standard univariate probit model. The results reveal that age, being male, and smoking have significant and positive effects on death due to both classes of causes. Being non-White increases the probability of death due to OCDs but not CSDs. On the other hand, being from the South has a moderately significant positive effect on the probability of dying from CSDs but not OCDs. Being married significantly reduces the probability of dying from a CSD, but the effect is only moderately significant for OCDs. Education is only moderately significant in reducing the probability of death to either type of cause. Finally, the error correlation between equations is significantly different from 0 (= –.73).

The $p$ values for these coefficients were computed as the empirical proportion of simulated parameters that exceeded 0 (either positively or negatively). While a standard $t$ test could be constructed, based on the posterior standard deviations, this test assumes normality of the marginal posterior distributions—an unnecessary assumption to make in a Bayesian setting.

The third column of the table reports the probabilities that the coefficient for each parameter differs across equations. Specifically, we report the probability that the coefficient from the CSD equation is greater than the coefficient from the OCD equation. This test is useful for testing hypotheses that may be relevant in examining the effects of covariates on different CODs. For example, the empirical probability that the effect of age is stronger for CSDs than for OCDs is 1, as expected, given that CSDs constitute two of the three leading causes of death. There is also a fairly high probability (.87) that being from the South has a greater effect on CSD than OCD, and such may be attributable to dietary differences between the South and other regions

**TABLE 2: Bivariate Probit Model Results: Effects of Covariates on Death Due to Circulatory System Diseases (CSD) and Death Due to Other Causes (OCD)**

| | Outcome | | |
|---|---|---|---|
| Covariate | Death Due to CD | Death Due to OC | $P(\beta_{CD} > \beta_{OC})^b$ |
| *Intercept*[a] | –6.13(.23)**** | –5.15(.25)**** | 0 |
| Age | .05(.003)**** | .04(.003)**** | 1 |
| Male | .23(.046)**** | .13(.05)*** | .93 |
| Non-White | .06(.06) | .13(.06)** | .20 |
| South | .07(.04)* | –.006(.04) | .87 |
| Married | –.13(.05)*** | –.07(.05)* | .20 |
| Smoking | .10(.03)**** | .13(.03)**** | .22 |
| Education | –.0008(.006)* | –.007(.006)* | .46 |
| $\rho_e$ | –.73(.11)**** | | |

a. The $p$ values are the empirical probability that the parameter is either $> 0$ or $< 0$, depending on the sign of the parameter. These values are found by computing the proportion of the simulated parameter values that exceed 0 (either positively or negatively). The $p$ values reported are classified into the ranges represented by the asterisks.
b. The probabilities are the probabilities that the coefficient from the first equation is greater than the coefficient from the second equation, obtained by computing the number of posterior parameter values in the CSD equation that exceed the maximum value for the comparable parameter in the OCD equation.
$^*p < .1$. $^{**}p < .05$. $^{***}p < .01$. $^{****}p < .001$.

(e.g., greater fat intake from fried foods). Finally, the probability that the intercept is larger for CSD is 0. This finding might imply that the general propensity to die from CSDs would be less than the propensity to die from OCDs, if risk factors for CSDs as a whole had effects comparable to their effects on OCDs (or, put another way, that the distributions of risk factors in the population were comparable across causes).

The validity of these results obviously depends on the performance of the model. The three general posterior predictive tests that we use to examine the fit of the model include the following: the count of double classifications, a model chi-square goodness-of-fit statistic between the observed and expected cell frequencies, and the log of odds ratios for death due to CSD versus OCD.

We conduct two additional tests to determine whether gender and smoking are modeled appropriately. First, men and women have different mortality patterns, which may suggest that they should be modeled separately to allow the age dependence of their rates to vary. Thus, using simply a dummy variable for gender would be

inappropriate. We can easily test this hypothesis by examining whether the posterior predictive distribution for the odds ratio for gender for CSD and OCD outcomes is centered over the value in the sample. Second, the measurement of smoking in this model appears very crude. It is reasonable that former smokers may have quit due to poor health, thus making them more likely to die than current smokers. On the other hand, there is reason to suspect that the integer-level coding of smoking may be inappropriate because former smokers look more like never-smokers than smokers after sufficient time has passed to repair the damage done by smoking. The appropriateness of this measure can be tested in a similar fashion by examining the posterior predictive distribution of the various odds ratios that can be constructed comparing each level of smoking with other levels for each cause of death.

Figure 2 shows the distribution of posterior predictive data in 10,000 replicated predictive data sets by double classifications, that is, future observations about whom the model would classify as dying from both classes of causes. Obviously, this double classification should never occur because the model does not allow the latent simulated data to be simulated from that region of the contingency table. However, in the predictive data, it is possible for data to be classified in that cell on occasion since the predictive data are simply a replication from the (nontruncated) bivariate normal distribution. The figure, however, indicates that in the replicated data, we rarely see double classifications. In fact, we see 0 double classifications in 9,338 of the 10,000 replicated data sets. The majority of the remaining 662 replications have only 1 (out of 18,005) double-classified observations ($n = 505$). A $p$ value can be obtained by dividing the number of replicated cases that are more extreme than the observed data by the total number of replications. Because the data are discrete, we treat half of the 9,338 replicated data sets with no double classifications as being more extreme, yielding a $p$ value of .5331. By this standard, the model fits the data well. However, a limitation of this measure is that it does not tell us whether our model reproduces the correct marginals—that is, the correct number of counts for each cell. A model that simply classified everyone as a survivor, for example, would obviously have 0 double classifications but would also not produce the appropriate number of decedents.

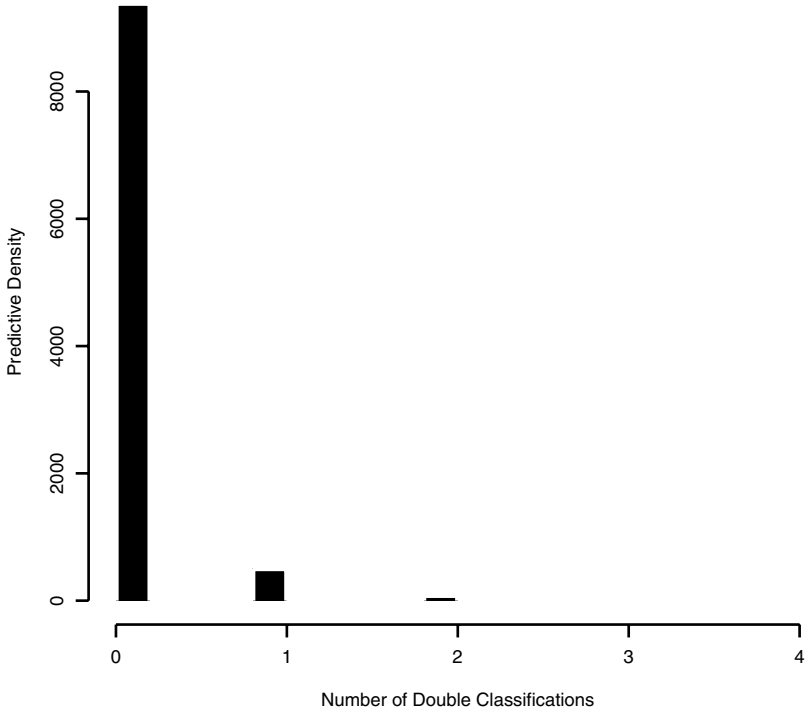


**Figure 2: Posterior Predictive Densities for the Number of Persons Classified as Dying From Both Classes of Causes**

Our second measure of model fit is a chi-square goodness-of-fit statistic computed for each of the 10,000 posterior predictive replications. This statistic has the following form:

$$\chi^2_{df-c-1} = \sum_{i=1}^{c} \frac{(O_i - E_i)^2}{E_i},$$

where the observed values are the cell counts in the posterior predictive data, the expected data are the cell counts in the original data, and the degrees of freedom are determined by the number of cells ($c$) in the table. In our case, we exclude the cell with a structural 0 count because there are 0 cases in it in most of the replicated data sets (refer to Figure 2), and the calculation cannot be made for that case.

Figure 3 displays the distribution of these chi-square values for the replicated data. The figure shows that the majority (76.7 percent) of the chi-squares for the replicated data fall below the $\alpha = .05$ "critical value" of 5.991 (indicated by the reference line in the figure—also, 98.98 percent fall below the $\alpha = .01$ threshold). Unlike a standard likelihood analysis, which would only calculate this statistic at the MLE, the Bayesian approach takes uncertainty in the parameter vector into account. Even after considering uncertainty, however, the model appears to perform well, given the high percentage of replications that fall below the classical threshold for considering the difference between the model-implied data and the observed data to be "significant." Compared to a model in which everyone is classified

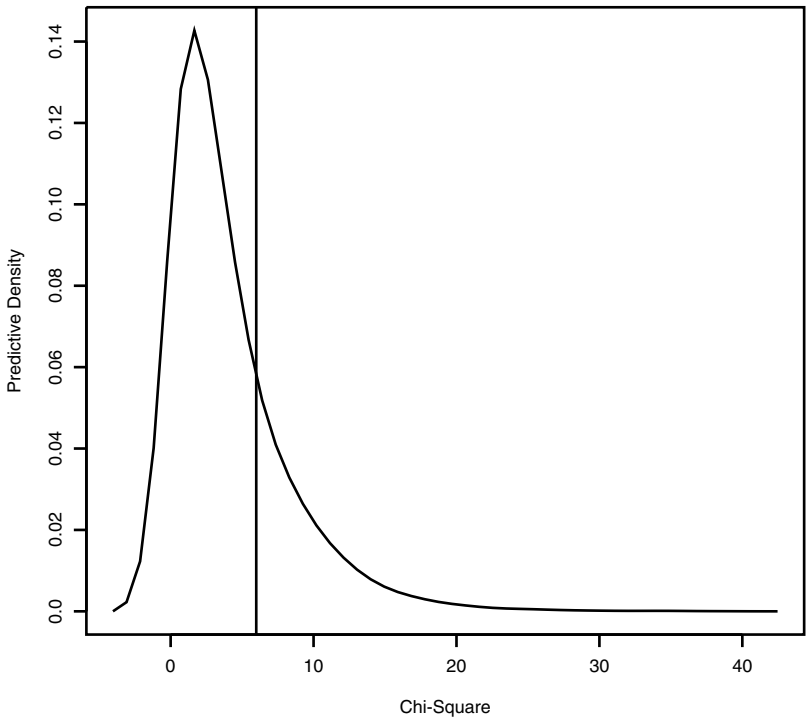


**Figure 3: Posterior Predictive Chi-Square Goodness-of-Fit Statistics**

as a survivor (which has a chi-square = 1042.33), this model looks much better.

Perhaps a better test of the model's ability to simultaneously fit each of the cells is the log of the odds ratio for death due to CSD versus OCD. The odds ratio is simply the ratio of the odds of dying from CSD to the odds of dying from OCD. The odds ratio is bounded at 0, so taking the log of the odds ratio makes the distribution more symmetric. For the observed data, the odds ratio is as follows:

$$\begin{aligned}&\text{log odds ratio}\\&= \log\left(\frac{497/18005}{(18005-497)/18005}\Big/\frac{412/18005}{(18005-412)/18005}\right)\\&= .19241.\end{aligned}$$

Figure 4 shows the distribution for the log odds ratio for the posterior predictive data. The value for the observed data falls squarely in the middle of the distribution here, with a $p$ value of .435. By this test, the observed data do not appear extreme, providing yet another indication that the model fits well. A model in which everyone was classified as surviving, on the other hand, would have a posterior predictive distribution at $-\infty$.

With the general model appearing to fit the data well, we turn our attention now to the checks of the specific covariates about which we were concerned. Figure 5 shows posterior predictive densities for gender for both causes of death. The figure suggests that our treatment of gender is appropriate in the model. The $p$ value for the odds ratio for CSD is .55, and the odds ratio for OCD is .46, indicating that the observed data and data replicated under the model assumptions are very close.

The results for smoking were not as promising. Figure 6 presents six histograms of the odds ratios for each level of smoking for both classes of causes of death. There are three odd ratios constructed for each level of smoking: 1 versus 0, 2 versus 0, and 2 versus 1. The first column of the figure reports the odds ratios for CSDs; the second column reports the odd ratios for OCDs. The posterior predictive distributions are reasonably centered over the true values for the OCD column. The (tail) $p$ values were .185, .149, and .131, suggesting that this measurement strategy is appropriate for the OCD outcome. On the other hand, the results for the CSD column clearly

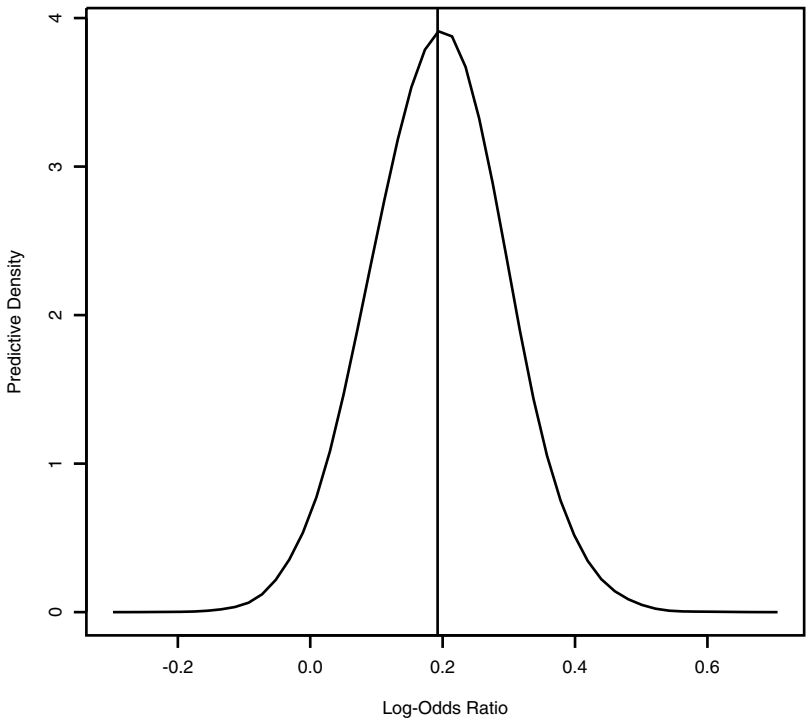


**Figure 4: Posterior Predictive Distribution of the Log Odds Ratio of Death Due to Circulatory System Diseases (CSD) Versus Other Causes of Death (OCD)**

indicate this is not an appropriate approach to measuring the effect of smoking on death due to CSDs. The (tail) $p$ values for these posterior predictive distributions were .003, .27, and .001. The results suggest that the hazards are not proportional across smoking levels measured in this fashion. Specifically, the posterior predictive distributions overpredict the odds of dying from a CSD for former smokers relative to nonsmokers (top histogram), and they underpredict the odds of dying from a CSD for current smokers versus former smokers (bottom histogram). Interestingly, however, the odds ratio for current smokers versus nonsmokers is estimated well. The results argue statistically for a revision of the measure that perhaps categorizes nonsmokers and former smokers together for this class of cause of death.

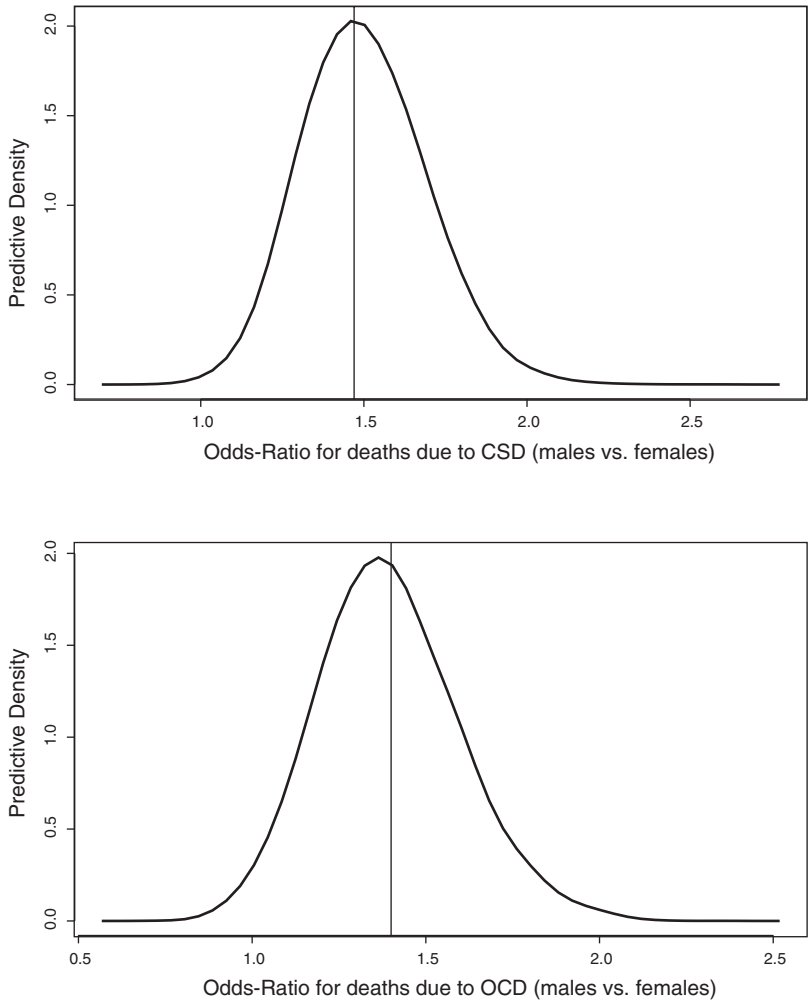


**Figure 5: Posterior Predictive Distribution of the Odds Ratio for Gender for Both Classes of Cause of Death**

Alternatively, we could simply replace the continuous measure of smoking with a set of dummy variables in both equations. Substantively, we may reconsider the fact that there is considerable heterogeneity within the former-smoker group, as well as within the current-smoker group, that may account for some of the poor fit.

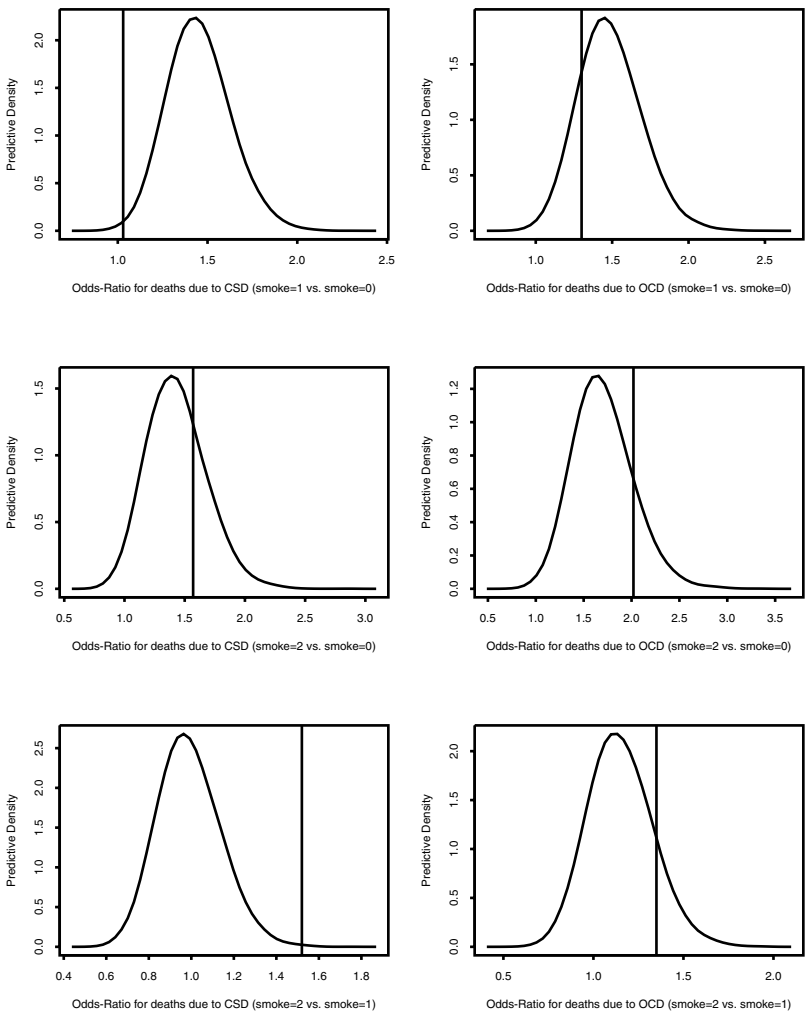


**Figure 6: Posterior Predictive Distributions of the Odds Ratio for Levels of Smoking for Both Classes of Cause of Death**

Respecification in this case, then, may require using an alternate (or additional) measure of smoking.

All the previous tests indicate that the joint assignment of individuals to each cell is relatively consistent across models—that there is

not a propensity to overassign or underassign individuals to one cell versus another in models when there is accurate placement of individuals in a third cell.

We have not resolved the issue of whether the model consistently misplaces specific individuals (possible outliers). We have developed a useful posterior predictive check (not a test) for such persistent misclassification. Specifically, for each individual in the data set, we compute the total number of times out of the 10,000 posterior predictive replications each observation is misclassified. If the model fit perfectly, then this distribution would be concentrated completely at 0. If the model had some slight error, then we might expect that the cluster of observations would be centered around 10 to 15 percent, implying that each individual was misclassified between $1/7$ and $1/10$ of the time. On the other hand, if the model were no better than random assignment of individuals to cells, we might expect the distribution to be centered around 75 percent, implying that each individual was misclassified $\frac{3}{4}$ of the time—there would only be a .25 probability that an individual would be classified correctly in any replicated data set. Figure 7 shows the distribution of these counts out of 10,000 replications, for all 18,005 individuals in the data set. The vast majority of the distribution is below 1,500, reflecting that 92 percent of the observations are classified correctly 85 percent of the time. The expectation of this distribution is 950, which means that the model, on average, misclassifies persons less than 10 percent of the time. If we compare this to random assignment, only about half of the observations would be classified more than 25 percent of the time, yielding an expectation of 7,500!

Beyond comparing hypothetical models, the most important feature of this posterior calculation is to locate outliers, which will tend to have very high misclassification rates. In Figure 7, we can see a handful of such persons at the extreme right edge of the histogram. Fortunately, before we began the analyses, we sorted the data by all the variables, starting with age. This allows us to pinpoint very quickly the characteristics of the outliers that may be causing the problem. In fact, it appears that the outliers are all persons who died—they are typically being misclassified as surviving. This may not be surprising, given that the probabilities of dying are very small—the data are highly skewed. However, the fact that the model is predicting the

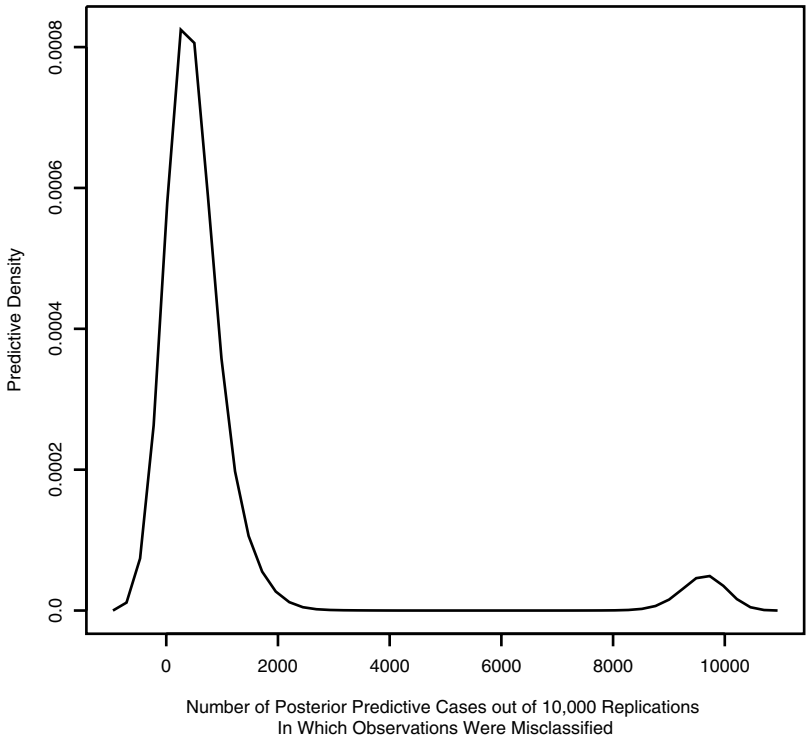


**Figure 7: Distribution of Replicated Observations by Frequency of Misclassification**

appropriate number of deaths in each category very well illustrates the ecological fallacy of assuming that the effects of the covariates hold for any particular observation.

In sum, the results suggest that the model fits the aggregate data fairly well, but the model does not fit very well to the individual cases. This finding may call for a revision of the model and further exploration of outliers.

### *EXAMPLE 2: ESTIMATING PRISON'S EFFECT ON WAGE GROWTH WITH PANEL DATA*

In this application, we consider the use of posterior predictive checks to study the fit of several models that are difficult to compare using

traditional methods. This application is motivated by the analysis of panel data in which we have observations on a dependent variable, $y$, for a unit (say, a survey respondent) $i$ at time points $t = 1, \ldots, T_i$. Inference for the usual least squares regression assumes that the observations are independent, conditional on a vector of covariates, $x_{it}$. This is unlikely to be true with this design because observations within units will tend to be correlated. A common model for panel data is written as follows:

$$y_{it} = \alpha_i + x'_{it}\beta + e_{it},$$

where the independence of the errors, $e_{it}$, within units is induced by fitting the unit-level effect, $\alpha_i$. For maximum likelihood estimation, $e_{it}$ is usually assumed to be normally distributed.

Assumptions about $\alpha_i$ define two alternative specifications, often termed *fixed- and random-effects models* (e.g., Hsiao 1986). The random-effects model is derived by assuming that $\alpha_i$ is a random variable, independently drawn from a normal distribution. With this distributional assumption, a likelihood can be written, and iterative methods can be used to find maximum likelihood estimates of $\alpha_i$ and the regression coefficients of key interest, $\beta$. The fixed-effects model makes no distributional assumptions and instead treats $\alpha_i$ as fixed parameters that may be correlated with other regressors. The fixed-effects model can be fit with least squares, either by introducing a dummy variable for each unit, $i$, or by transforming $x_{it}$ and $y_{it}$ into deviations from unit-level means.

The fixed-effects model is often preferred because it will yield unbiased estimates of $\beta$ in the presence of certain kinds of omitted variables. In particular, if there are other predictors, $z_i$, that vary across units and are correlated with the model covariates $x_{it}$, the effects of $z_i$ will be absorbed by the unit-level coefficients, $\alpha_i$. The random-effects estimates of $\beta$ will suffer from omitted variable bias in this situation.

Is there any reason, then, to prefer the random-effects model? The fixed-effects model fits an extremely large number of parameters and is thus less efficient than the random-effects model. This is reflected in relatively large standard errors for the fixed-effects estimates of the coefficients, $\beta$. In specific applications, the reduction in bias obtained by the fixed-effects model may be outweighed by the increase in variance. Although bias and variance considerations should guide

the issue of model choice, the criterion of unbiasedness has received greatest attention. Specification tests have been proposed that focus on the bias in the random-effects estimate of $\beta$. The most popular test, proposed by Hausman (1978), provides a chi-square statistic for the null hypothesis that the unit-level effects and the predictors are uncorrelated. Although textbooks indicate the importance of the bias-variance trade-off, the value of unbiased estimation appears to be a higher priority for applied researchers because fixed effects are often preferred to random effects in practice.

Although econometric analysis tends to argue for the utility of the fixed-effect model, the applied relevance of the analysis may be overstated. All discussions compare fixed- and random-effects estimators for models in which the predictors are identical. In practice, however, researchers will often have unit-level variables, $z_i$, that can be included in the random-effects models. Because the coefficients of $z_i$ are not identified for the fixed-effects estimator, researchers face the practical problem of comparing nonnested models. Consider an analysis of men's wages when there are unit-level data on the race, ethnicity, and cognitive test scores of survey respondents. These variables would be included in the random-effects model but not the fixed-effect model. Specification tests are not available to compare these two nonnested models.

In addition, analysis generally focuses on random intercept models, but in some applications, it would also be useful to specify random slopes as well. Thus, for example, the analysis of wages by Bernhardt et al. (2001) supplement the usual random intercept with a random component for the effect of age on men's wages. This model may provide a better specification of the time-dependent character of observations than a model that assumes that the age-wage profile is identical for all respondents. Standard specification does not accommodate the comparison of fixed-effects models to models with multiple random components.

The fit of such nonnested models can be compared using posterior predictive checks. To explain the calculation of posterior predictive distributions, we begin by providing a Bayesian formulation of the fixed- and random-effects models. For all models, $y_{it}$ is conditionally normal:

$$y_{it}|x_{it} \sim N(\hat{y}_{it}, \sigma^2),$$

where the expectation is

$$\hat{y}_{it} = \alpha_i + x'_{it}\beta + z'_i\delta.$$

For the fixed-effect model, we specify a diffuse proper prior for $\alpha_i$ and each of the $k$ components of $\beta$:

$$\alpha_i \sim N(0, \sigma_\alpha^2),$$
$$\beta_k \sim N(0, \sigma_\beta^2),$$

where the prior variances are taken to be large—say, $\sigma_\alpha^2 = \sigma_\beta^2 = 10^6$. The effects of $z_i$ are not identified by the data and will be absorbed by the estimates of $\alpha_i$, so $\delta$ is set to zero for the fixed-effects model.

In the random intercept model, the priors for $\beta$ and $\delta$ are given the same diffuse prior given to $\beta$ in the fixed-effects model, but the prior for the unit-level effects has the following hierarchical structure:

$$\alpha_i \sim N(\alpha_0, \tau_\alpha^2)$$
$$\alpha_0 \sim N(0, \sigma_\alpha^2)$$
$$\tau_\alpha^2 \sim IG(a_0, b_0),$$

where $IG$ denotes the inverse gamma distribution for the variance parameter, and the prior parameters $a_0$ and $b_0$ are chosen to be large, representing diffuse prior information. From the Bayesian perspective, the fixed- and random-effects models are identical except in the choice of prior for $\alpha_i$. The fixed-effect model has a diffuse prior centered at zero; the random-effects model has an informative prior centered at $\alpha_0$.

The model can be elaborated further by allowing $\beta$ to vary across units. The so-called random slope model introduces a hierarchical structure for the regression coefficients:

$$\beta_i \sim N(\beta_0, \tau_\beta^2)$$
$$\beta_0 \sim N(0, \sigma_\beta^2)$$
$$\tau_\beta^2 \sim IG(a_0, b_0).$$

With a likelihood for $y_{it}$ and proper priors for all parameters, a Gibbs sampler can be written to simulate draws from the posterior distributions. If particular simulated values are written—$\beta^*$, $\alpha_i^*$, and

$\sigma^*$—these can be used to calculate the simulated conditional mean, $\mu^*_{it}$. The posterior predictive distribution can then be simulated by random draws from

$$y^{\text{rep}}_{it} \sim N(\mu^*_{it}, \sigma^{*2})$$

for simulated values of the parameters from each iteration of the Gibbs sampler.

We illustrate posterior predictive checks for models of panel data by analyzing data on wages for men, using data from the National Longitudinal Survey of Youth (NLSY). The NLSY is a panel survey of a national sample of men and women ages 14 to 21 in 1979. Survey respondents were interviewed annually until 1994 and then again in 1996 and 1998. Unlike many labor force surveys, the NLSY continued to interview respondents if they went to prison or jail. The incarceration status of respondents is recorded at the time of the interview. The data thus provide a useful source of information for studying the effects of incarceration on wage mobility. We analyze wage mobility among a high-risk sample of young Black NLSY men who report being in jail or prison or involved in juvenile delinquency or adult crime.

Interest centers on whether the age-wage gradient for men who have been to prison is different from wage growth among male nonprisoners. Some descriptive statistics on the growth in log hourly wages is reported in Table 3. Growth in median log wages between ages 22 and 35 is about 40 percent greater for nonprisoners than prisoners, providing preliminary evidence for slow wage growth among men with prison records.

Regression results for four models are reported in Table 4. Each model yields a statistically significant estimate for the interaction between age and incarceration status, indicating slow wage growth among ex-prisoners. The main effect of age, however, is sensitive to the model choice, suggesting different rates of wage mobility for prisoners and ex-prisoners. Which model should be preferred? The $R^2$ shows that the random slope and intercept model fits best. Which model best captures patterns of wage growth that we observe in the data?

To answer this question, we define a discrepancy statistic,

$$T(y) = P_c(y|\text{age} = 35) - P_c(y|\text{age} = 22),$$

**TABLE 3: Results for Regression of Log Hourly Wages on Incarceration Status and Other Characteristics, Black Men Reporting Crime or Incarceration, National Longitudinal Survey of Youth 1983-1998**

| | *Ordinary Least Squares* | *Fixed Effects* | *Random Intercepts* | *Random Slopes and Intercepts* |
|---|---|---|---|---|
| Intercept | .76 | .66 | .79 | .77 |
| | (12.69) | (4.24) | (8.36) | (7.40) |
| Was incarcerated ($P$) | –.02 | –.05 | –.03 | –.04 |
| | (.87) | (1.64) | (1.23) | (1.05) |
| Log age ($A$) | –.04 | .79 | .23 | .17 |
| | (.37) | (2.21) | (1.31) | (.87) |
| $P \times A$ | –.37 | –.30 | –.35 | –.42 |
| | (3.31) | (2.58) | (3.28) | (2.02) |
| Now incarcerated ($C$) | –.22 | –.23 | –.22 | –.18 |
| | (7.06) | (7.87) | (7.96) | (6.64) |
| Work experience | .59 | .64 | .57 | .59 |
| | (17.95) | (8.89) | (11.94) | (12.27) |
| Education | .30 | .63 | .35 | .34 |
| | (6.82) | (5.50) | (4.91) | (4.91) |
| Year | .01 | –.01 | .01 | .01 |
| | (3.87) | (1.16) | (.97) | (1.31) |
| AFQT | .56 | — | .47 | .48 |
| | (11.81) | | (5.09) | (5.04) |
| $R^2$ | .20 | .44 | .44 | .56 |

NOTE: Number of respondents = 533; 5,043 respondent-years. AFQT = Armed Forces Qualifying Test. Further details of this analysis are reported in Western (2002).

where the function $P_c$ returns the level of log wages at percentile $c$. For example, the median wage at age 35 is given by $P_{50}(y|\text{age} = 35)$. We examine three different discrepancy statistics for $c = 10, 50, 90$. Posterior predictive densities for growth in median log wages from age 22 to 25 is shown in Figure 8. Densities are shown for the fixed-effect model, the random intercept model, and the random slope and intercept model. The vertical line on the figure shows the observed difference in median wages. All three models accurately reproduce the difference in median wages and indeed yield very similar predictions.

At the tails of the wage distribution—at the 10th and 90th percentiles—however, we see some divergence between the model predictions (Figure 9). All three models tend to overestimate the difference in log wages at the 10 percentiles but underestimate wage

**TABLE 4: Location and Spread of Log Hourly Wages by Age and Incarceration Status, Black Men Reporting Delinquency or Incarceration, National Longitudinal Survey of Youth 1983-1998**

| | Age 22 (1) | Age 35 (2) | Difference (2) – (1) |
|---|---|---|---|
| | *Prisoners* | | |
| Mean | 1.461 | 1.589 | .128 |
| Median | 1.421 | 1.594 | .173 |
| 90-10 difference | .890 | 1.150 | .260 |
| | *Nonprisoners* | | |
| Mean | 1.473 | 1.804 | .331 |
| Median | 1.451 | 1.762 | .311 |
| 90-10 difference | .848 | 1.315 | .468 |

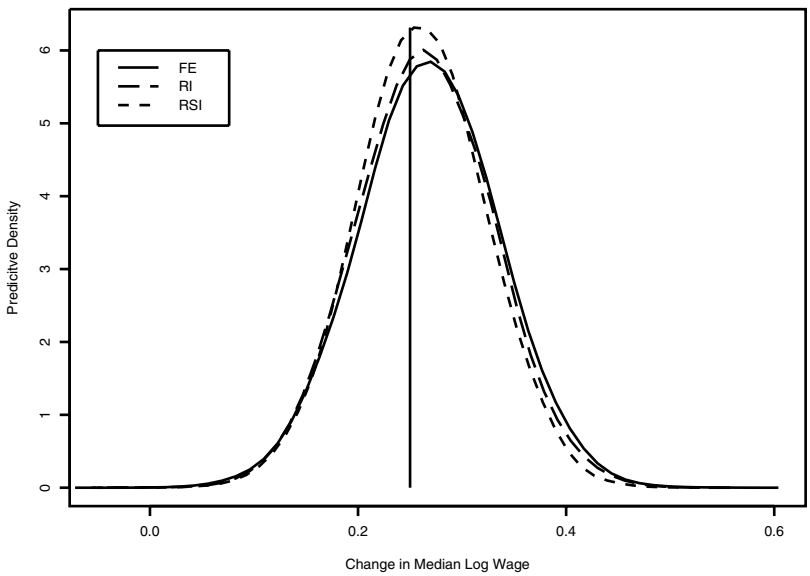


**Figure 8: Posterior Predictive Densities for Change in Log Median Wages From Ages 22 to 35 for Fixed-Effect (FE), Random Intercept (RI), and Random Slope and Intercept (RSI) Models**

NOTE: The vertical line indicates the observed change in median wages.

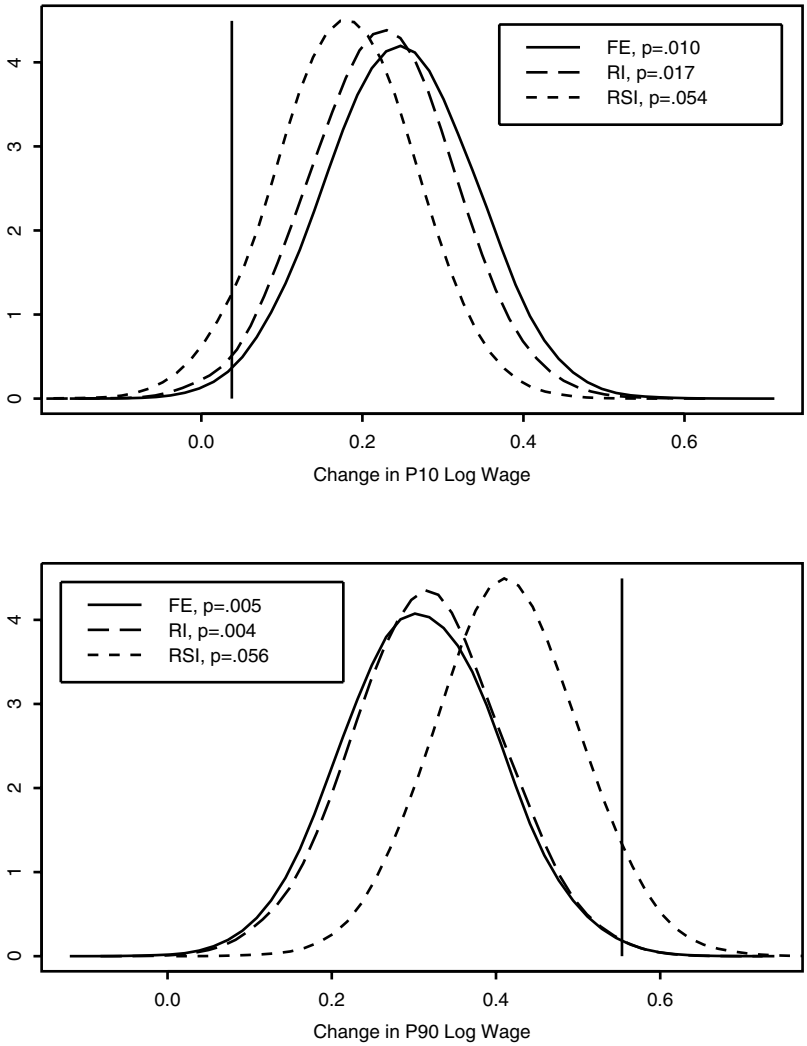


**Figure 9: Posterior Predictive Densities for Change in Wages From Ages 22 to 35 for Fixed-Effect (FE), Random Intercept (RI), and Random Slope and Intercept (RSI) Models**

NOTE: The top panel shows the change in the 10th percentile of the wage distribution. The lower panel shows the change in the 90th percentile of the wage distribution.

differences at the top. The wage distribution tends to fan out as the NLSY respondents age, and none of the models are able to successfully reproduce this pattern of variation. Although all three models tend to fit rather poorly in this respect, the random slope and intercept specification tends to do somewhat better than the other two. Bayesian $p$ values for wage differences at the 10th and 90th percentiles for the random slope and intercept (RSI) model are around .05 compared to $p$ values of under .01 for the model with just unit-level effects.

What conclusions can be drawn from the posterior predictive analysis? The posterior predictive distributions are diagnostic tools, not statistics for model selection. The superior fit of the RSI model to patterns of wage growth at the 10th and 90th percentiles indicates that this model captures the variety of wage trajectories somewhat better than alternative models. Given that other research has emphasized the importance of the increasing dispersion of wage profiles (Bernhardt et al. 2000), the RSI model can be understood to reproduce a substantively important pattern of variation. Although it is tempting to use this as a basis for preferring the RSI model over the fixed-effect and random-effect alternatives, other discrepancy statistics could also be defined. For example, we might also study the capacity of models to predict the difference in wage dispersion between ex-inmates and noninmates or among Black and White men. Discrepancy statistics can be calculated to describe the realism of the models in capturing these other features of the data. The substantive context of the research problem should govern the choice of discrepancy statistics.

What if posterior predictive distributions indicate that none of the models under consideration accurately capture important patterns of variation? In this case, we resort to the usual tools for model reformulation: adding predictors, experimenting with new functional forms, data transformations, and so on. Despite statistically significant regression coefficients, posterior predictive checks may indicate a qualitatively poor fit of the model to the data. Under these circumstances, our coefficients may not provide a substantively relevant description of the data or the data-generating mechanism, and alternative models should be examined.

## *DISCUSSION*

The posterior predictive distribution provides a useful set of statistics for assessing model fit. Unlike standard goodness-of-fit statistics, discrepancy statistics formed from the posterior predictive distribution can be designed to assess a model's fit to substantively important patterns of variation. It is difficult to generalize about the form that discrepancy statistics should take, but in our research, we have found it useful to examine conditional odds ratios in analyzing discrete data and conditional quantiles in examining continuous data. Arising in a Bayesian framework, the discrepancy statistics have an inferential as well as diagnostic interpretation. Discrepancy statistics allow the calculation of $p$ values that indicate the plausibility of observed data under a particular model.

Because posterior predictive distributions can be approximated using simulations from the posterior distribution of model parameters, the diagnostic offers a useful approach for studying models fit by simulation methods such as MCMC. Such models often have a hierarchical structure and can often consist of very large numbers of parameters. In many cases, rival models will be nonnested. Under these conditions, standard fit statistics and diagnostics will be difficult to apply. Furthermore, in many cases, the model(s) under consideration may simply not have any (or many) established diagnostic tests, or the standard diagnostic tests may be based on assumptions (e.g., normality of the error) that are unreasonable for the particular model. Posterior predictive checks, however, provide an easily implemented approach to model evaluation and comparison in this context.

A Bayesian approach to monitoring model fit by comparing model predictions to observed data was introduced by Rubin (1981, 1984). A modern application of posterior predictive checks in the context of MCMC simulation is provided by Gelman et al. (1996). A discussion of the Bayesian $p$ values, as well as a comparison to classical hypothesis testing, is provided by Berkhof and van Mechelen (2000). A good didactic discussion of posterior predictive checks can be found in Gelman et al. (1995, chap. 6). In a sociological context, the merits of posterior predictive checks in comparison to the calculation of Bayes factors for model comparison have been discussed by Raftery (1995a, 1995b) and Gelman and Rubin (1995).

## REFERENCES

Allison, Paul D. 1984. *Event History Analysis: Regression for Longitudinal Event Data*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-046. Beverly Hills, CA: Sage.

Berkhof, Johannes and Iven van Mechelen. 2000. *Computational Statistics* 15:337-54.

Bernhardt, Annette, Martina Morris, Mark S. Handcock, and Marc A. Scott. 2000. *Divergent Paths: Economic Mobility in the New American Labor Market*. New York: Russell Sage Foundation.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 1995. *Bayesian Data Analysis*. London: Chapman & Hall.

Gelman, Andrew, Xiao-Li Meng, and Hal S. Stern. 1996. "Posterior Predictive Assessment of Model Fitness Via Realized Discrepancies." *Statistica Sinica* 6:733-807.

Gelman, Andrew and Donald B. Rubin. 1995. "Avoiding Model Selection in Bayesian Social Research." *Sociological Methodology* 25:165-74.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter, eds. 1996. *Markov Chain Monte Carlo in Practice*. Boca Raton, FL: Chapman & Hall/CRC.

Hausman, J. A. 1978. "Specification Tests in Econometrics." *Econometrica* 46:1251-71.

Hsiao, Cheng. 1986. *Analysis of Panel Data*. Cambridge, UK: Cambridge University Press.

Lynch, Scott M., J. Scott Brown, and Katherine G. Harmsen. (2003). "The Effect of Altering ADL Thresholds on Active Life Expectancy Estimates for Older Persons." *The Journals of Gerontology: Social Sciences* 58(3):S171-178.

Preston, Samuel H., Patrick Heuveline, and Michel Guillot. 2001. *Demography: Measuring and Modeling Social Processes*. Oxford, UK: Blackwell.

Raftery, Adrian E. 1995a. "Bayesian Model Selection in Social Research." *Sociological Methodology* 25:111-64.

———. 1995b. "Rejoinder: Model Selection Is Unavoidable in Social Research." *Sociological Methodology* 25:185-96.

Rubin, Donald B. 1981. "Estimation in Parallel Randomized Experiments." *Journal of Educational Statistics* 6:377-401.

———. 1984. " Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician." *Annals of Statistics* 12:1151-72.

Tierney, Luke. 1996. "Introduction to General State-Space Markov Chain Theory." Pp. 59-74 in *Markov Chain Monte Carlo in Practice*, edited by W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. Boca Raton, FL: Chapman & Hall/CRC.

Western, Bruce. (2002). "The Impact of Incarceration on Wage Mobility and Inequality." *American Sociological Review* 67:526-546.

*Scott M. Lynch is an assistant professor of sociology at Princeton University. His recent substantive research examines inequality and health over the life course. His recent methodological work involves developing a Bayesian approach to estimating multistate life tables with covariates.*

*Bruce Western is a professor of sociology at Princeton University. His recent research examines labor market inequality and social impact of the growth of incarceration in the United States.*