Department of Economics
Harvard University

June 2001

NONPARAMETRIC APPLICATIONS OF BAYESIAN INFERENCE

Gary Chamberlain and Guido W. Imbens

ABSTRACT

The paper evaluates the usefulness of a nonparametric approach to Bayesian inference by presenting two applications. The approach is due to Ferguson (1973, 1974) and Rubin (1981). Our first application considers an educational choice problem. We focus on obtaining a predictive distribution for earnings corresponding to various levels of schooling. This predictive distribution incorporates the parameter uncertainty, so that it is relevant for decision making under uncertainty in the expected utility framework of microeconomics. The second application is to quantile regression. Our point here is to examine the potential of the nonparametric framework to provide inferences without relying on asymptotic approximations. Unlike in the first application, the standard asymptotic normal approximation turns out to not be a good guide.

NONPARAMETRIC APPLICATIONS OF BAYESIAN INFERENCE[1]

1. INTRODUCTION

The paper evaluates in the context of two applications the usefulness of a nonparametric approach to Bayesian inference. The basic approach is due to Ferguson (1973, 1974) and Rubin (1981). It has three key features. First, it has the basic benefits of Bayesian inference in providing a well-defined posterior distribution that is an important ingredient in many decision problems. Second, it has some of the advantages of semi-parametric models used in frequentist analyses by not relying on a tightly parameterized likelihood function, based, for example, on a normal distribution. Third, it avoids pitfalls arising in Bayesian analyses from using high-dimensional parameter spaces with flat or other conventional prior distributions by using a prior distribution that arguably reflects lack of prior knowledge. These three features will be illustrated in the two applications.

Our first application considers an educational choice problem. Specifically, we look at an individual's decision on the level of schooling when the individual is uncertain about the return to schooling. Following Angrist and Krueger (1991) we allow for endogeneity of the schooling measure by using a quarter of birth dummy as an instrumental variable. A standard parametric model would require distributional assumptions on the joint distribution of earnings and schooling given the instrument. On the other hand, standard instrumental variables methods that do not require such assumptions do not lead to the predictive earnings distributions required for the educational choice problem. The Bayesian nonparametric approach discussed in this paper allows us to obtain a predictive distribution for earnings corresponding to various levels of schooling that incorporates the parameter uncertainty, so that it is relevant for decision making under uncertainty in the expected utility framework of microeconomics. At the same time in this application this approach avoids strong distributional assumptions without introducing strong sensitivity to the prior distribution.

The second application is to quantile regression. Our point here is to examine the potential of the nonparametric Bayesian framework to provide inferences without making asymptotic approximations. Unlike in the first application, in this application the standard asymptotic normal distribution turns out to be a poor approximation to the sampling distribution of the estimator in some cases. If the standard normal distribution provides a good approximation to the finite sample distribution, posterior intervals obtained through the Bayesian nonparametric approach discussed in this paper are close to confidence intervals. When the large sample normal approximation fails to provide a good approximation to the finite sample distribution, the interpretation of our posterior distribution is not affected.

2. DIRICHLET PRIOR DISTRIBUTIONS

Here we present a concise review of the basic theory, extended to allow for parameters defined by moment restrictions, that is sufficient to follow the applications. For more details, see Ferguson (1973, 1974), Rubin (1981), Chamberlain and Imbens (1995) and Hirano (1999). There is a family of probability distributions $\{P_\theta : \theta \in \Theta\}$, and we observe $\{Z_i\}_{i=1}^n$, where the random variables Z_i are independently and identically distributed according to P_θ for some unknown value of θ in the parameter space Θ. To simplify notation, let Z denote a random variable that is distributed according to P_θ. We assume that the distributions P_θ have a common, finite support:

$$P_\theta(Z = a_j) = \theta_j \qquad (j = 1, \ldots, J),$$

where θ_j denotes the j^{th} component of θ and we take Θ to be the unit simplex in $\mathcal{R}^J$. Since J can be arbitrarily large, and our data are measured with finite precision, the finite support assumption is arguably not restrictive. In fact, Ferguson's (1973) discussion does not rely on discreteness. See also Hirano (1999).

Typically we are interested in some function of θ rather than elements of θ itself: $\beta = g(\theta)$, where the function $g(\cdot)$ may depend upon the points of support $\{a_j\}_{j=1}^J$. For example, we shall consider cases where $g(\cdot)$ is defined implicitly through moment restrictions,

$$E_\theta \psi(Z,\beta) = \sum_{j=1}^{J} \psi(a_j,\beta)\cdot\theta_j = 0, \tag{1}$$

where ψ is a given function with dimension equal to that of β, and there is a unique solution for all $\theta \in \Theta$. Although it may appear to be restrictive to limit this discussion to the case with the dimension of β equal to that of ψ, one can apply the same approach to overidentified gmm models where the dimension of ψ is higher than the dimension of β by augmenting the parameter vector and the moment functions. Specifically, let $\gamma = (\beta_0, \beta_1, \Gamma_0, \Gamma_1, \Delta)$, and let

$$\tilde{\psi}(Z,\gamma) = \begin{pmatrix} \Gamma_0'\psi(Z,\beta_0) \\ \mathrm{vec}(\partial\psi(Z,\beta_0)/\partial\beta' - \Gamma_0) \\ \mathrm{vec}(\psi(Z,\beta_0)\psi(Z,\beta_0)' - \Delta) \\ \Gamma_1'\Delta^{-1}\psi(Z,\beta_1) \\ \mathrm{vec}(\partial\psi(Z,\beta_1)/\partial\beta' - \Gamma_1) \end{pmatrix}.$$

Then the solution to $\sum_{i=1}^{n} \tilde{\psi}(z_i,\gamma) = 0$ gives the standard optimal two-step gmm estimator for β, motivating our interest in the posterior distribution for the parameter defined as the solution to $E[\tilde{\psi}(Z,\gamma)] = 0$. Our proposed procedure will give a posterior distribution for this parameter given the data.

A second example concerns cases where $g(\cdot)$ is defined as the solution to an optimization problem:

$$\beta = \arg\min_t E_\theta[\rho(Z,t)] = \arg\min_t \sum_{j=1}^{J} \rho(a_j,t)\cdot\theta_j, \tag{2}$$

where ρ is a given scalar-valued function and there is a unique solution for all $\theta \in \Theta$. In both cases we obtain draws from the posterior distribution of β by first drawing from the posterior distribution of θ and then solving (1) or (2).

We limit ourselves to prior distributions in the Dirichlet family with density

$$p(\theta) \propto \prod_{j=1}^{J} \theta_j^{b_j-1} \quad \text{for} \quad \theta\in\Theta \qquad (b_j > 0), \tag{3}$$

which, with J free parameters b_j, is fairly flexible. Similar to the way the Beta distribution is the conjugate prior distribution for the parameter of a binomial distribution, the Dirichlet distribution is the conjugate prior distribution for the parameters of a multinomial distribution. Let $d = \{z_i\}_{i=1}^n$ denote the data, that is, the observed values of the Z_i, and let $n_j = \sum_{i=1}^n 1(z_i = a_j)$ be the number of sample observations equal to a_j. The posterior density is proportional to the product of the prior density and the likelihood function:

$$p_n(\theta \,|\, d) \propto \prod_{j=1}^J \theta_j^{n_j + b_j - 1},$$

and thus also Dirichlet with parameters $n_j + b_j$, $j = 1, \ldots, J$. Within this family of Dirichlet prior distributions we focus on the improper prior distribution with all the $b_j \to 0$. There are three important features of this improper prior distribution that we shall briefly comment on.

First, the improper prior distribution avoids the potential pitfall in using the Dirichlet prior with large J and all of the b_j bounded away from zero. Since we rely on J being large to make the model flexible, this would potentially be an important drawback of the method. To see the problem, let ϕ denote the probability that Z is in some set B: $\phi = \sum_{j:a_j \in B} \theta_j$. Then the posterior distribution for ϕ is a beta distribution with

$$E(\phi \,|\, d) = \sum_{j:a_j \in B} (n_j + b_j) \Bigg/ \sum_{j=1}^J (n_j + b_j)$$

$$\mathrm{Var}(\phi \,|\, d) = E(\phi \,|\, d)[1 - E(\phi \,|\, d)] \Bigg/ \left(1 + \sum_{j=1}^J (n_j + b_j) \right).$$

Suppose that the $b_j = \epsilon > 0$ for all j, and consider increasing the number of support points while keeping the data d fixed. Let the fraction of support points in B approach a limit r: $\frac{1}{J} \sum_{j=1}^J 1(a_j \in B) \to r$ as $J \to \infty$. Then $E(\phi \,|\, d) \to r$, $\mathrm{Var}(\phi \,|\, d) \to 0$, and both prior and posterior distribution of ϕ become concentrated at r, regardless of the data. In particular, this argument covers a flat prior for θ ($b_j \equiv 1$), suggesting that a flat prior distribution does not capture a lack of prior information very well when J is large.

The second point is computational. The algorithm for evaluation of $\beta = g(\theta)$ defined through moment functions takes a particularly simple form for the limiting posterior distribution that results from letting all the $b_j \to 0$ in (3). Then the θ_j corresponding to the support points a_j not observed in the sample are all zero with posterior probability one. Let $\{V_i\}_{i=1}^n$ be independently distributed according to a standard exponential distribution (i.e., the gamma distribution $\mathcal{G}(1,1)$). Then, for a given function $\lambda(\cdot)$,

$$\sum_{i=1}^{n} \lambda(z_i) V_i \Big/ \sum_{i=1}^{n} V_i = \sum_{j:n_j>0} \lambda(a_j) U_j \Big/ \sum_{j:n_j>0} U_j,$$

where $U_j = \sum_{i:z_i=a_j} V_i \sim \mathcal{G}(n_j, 1)$, using the fact that a sum of independent exponential random variables has a gamma distribution. So in order to simulate the posterior distribution of β based on (1), instead of drawing from the posterior distribution of θ and then solving

$$\sum_{j=1}^{J} \psi(a_j, \beta) \cdot \theta_j = 0,$$

we draw sets of i.i.d. exponential random variables $\{V_i^{(l)}\}_{i=1}^n$ and solve

$$\sum_{i=1}^{n} \psi(z_i, \beta^{(l)}) V_i^{(l)} = 0, \tag{4}$$

and similarly for β based on (2) we solve

$$\beta^{(l)} = \arg\min_t \sum_{i=1}^{n} \rho(z_i, t) V_i^{(l)}. \tag{5}$$

Repeating this for $l = 1, \ldots, L$ gives us L independent draws from the posterior distribution of β. Rubin (1981) developed this simulation algorithm (using a representation for the ratio of exponentials to the sum of exponentials as gaps in order statistics from a uniform distribution), and it has been applied by Lancaster (1994) in the analysis of choice-based samples.

The third issue is that the improper prior distribution for θ does not imply a unique prior distribution for the parameter of interest. Although for proper prior distributions for θ the prior distribution for β is well-defined, the limiting prior distribution for β as the $b_j \to 0$ depends on the limits of the ratios b_j/b_l. To see this, consider the example discussed before where we are interested in ϕ, the probability that Z is in some set B: $\phi = \sum_{j:a_j \in B} \theta_j$. For fixed b_j the prior mean of ϕ is $E(\phi) = \sum_{j:a_j\in B} b_j \;/\; \sum_{j=1}^{J} b_j$. As we let the $b_j \to 0$, the limiting mean depends on the limit of the ratios of b_j/b_l. The posterior mean is $E(\phi \,|\, d) = \sum_{j:a_j\in B}(n_j + b_j) \;/\; \sum_{j=1}^{J}(n_j + b_j)$, which, after taking the limit $b_j \to 0$, equals $\sum_{j:a_j\in B} n_j \;/\; \sum_{j=1}^{J} n_j$, which does not depend on the limit of the ratios b_j/b_l. As this example illustrates, it is important to understand the implications of the choice of the limiting Dirichlet distribution. In order to measure the informativeness of the prior distribution for β, we propose calculating the expected posterior distribution given a small number m of observations, where we take the expectation over the empirical distribution. Let F_n denote the empirical distribution of our sample: $F_n(B) = \frac{1}{n}\sum_{i=1}^{n} 1(z_i \in B)$. Let $\pi_m^{\beta}(\cdot \,|\, \{t_i\}_{i=1}^m)$ denote the posterior distribution for β based on the m observations $Z_i = t_i$ (and assume for a moment that this posterior distribution is proper). The expected posterior distribution for β based on a random sample (with replacement) of size m from F_n is given by $\bar{\pi}_m^{\beta}(\cdot) = \int \pi_m^{\beta}(\cdot \,|\, \{t_i\}_{i=1}^m) \prod_{i=1}^m dF_n(t_i)$. In order to allow for the possibility of an improper posterior distribution, we modify this formula as follows:

$$\bar{\pi}_m^{\beta}(\cdot) = \tag{6}$$

$$\int \pi_m^{\beta}(\cdot \,|\, \{t_i\}_{i=1}^m) 1(\{t_i\}_{i=1}^m \in C_m) \prod_{i=1}^{m} dF_n(t_i) \Bigg/ \int 1(\{t_i\}_{i=1}^m \in C_m) \prod_{i=1}^{m} dF_n(t_i),$$

where the set C_m consists of the points $\{t_i\}_{i=1}^m$ such that $\pi_m^{\beta}(\cdot \,|\, \{t_i\}_{i=1}^m)$ is a proper distribution. If the prior distribution is not very informative for β, different small samples $\{t_i\}_{i=1}^m$ could potentially lead to very different posterior distributions, and thus the average posterior distribution should be relatively dispersed. If we find, therefore, that this average small sample posterior distribution is dispersed compared to the full posterior distribution, we interpret that as evidence that our prior distribution does not dominate the data.

3. INSTRUMENTAL VARIABLES

The first application illustrates how the general method described above can generate posterior distributions without tightly parametrized models. Such a posterior distribution is called for in order to include parameter uncertainty in the decision making formulation; see, for example, Rossi, McCulloch, and Allenby (1995), Kandel and Stambaugh (1996), and Barberis (2000). In this first example the large sample normal approximation to the sampling distribution can actually be used to approximate this posterior distribution fairly accurately. If, however, the objective is a posterior distribution for the parameter of interest, then our procedure is more direct than having to first approximate a sampling distribution by a normal distribution and then argue that this normal distribution can be used to approximate a posterior distribution.

We shall use a very simple model relating earnings and schooling with a constant, additive treatment effect, linear in years of schooling. An individual may choose schooling levels by maximizing expected lifetime discounted utility, with utility depending on earnings at various schooling levels as well as costs associated with schooling. Such a decision requires as one of the inputs the posterior distribution of earnings at the relevant schooling levels. The potential outcome with treatment level s is

$$Y_s = Y_0 + \gamma s,$$

where Y_0 is the potential outcome with treatment level 0, and γ is the unknown return to schooling, common to all individuals and common to all schooling levels. The actual treatment level is X, which gives an actual outcome Y of

$$Y = Y_0 + \gamma X.$$

Let α be the population mean of Y_0, and define the disturbance $U = Y_0 - \alpha$ so that $E_\theta(U) = 0$. The instrumental variable W satisfies $E_\theta(WU) = 0$ and $\mathrm{Cov}_\theta(W, X) \neq 0$. We are abstracting from the presence of exogenous covariates—they could be incorporated into the analyses presented below without any problems.

Let $Z = (Y, X, W)$ and $\beta' = (\alpha, \gamma)$. Then β satisfies the moment condition $E_\theta \psi(Z, \beta) = 0$

with

$$\psi(Z,\beta) = (Y - \alpha - \gamma X)\begin{pmatrix} 1 \\ W \end{pmatrix}.$$

Assuming finite support for the distribution of Z, we shall use the improper Dirichlet prior (with all the $b_j \to 0$ in (3)) for the parameters of this, and the posterior distribution of β can be simulated as in (4).

Our data is a subset of the data used by Angrist and Krueger (1991) containing males born in either the first or fourth quarters between 1930 and 1939. The sample size is $n = 162,515$. The outcome variable Y is the log of weekly earnings in 1979. The treatment X is years of schooling completed, and the instrumental variable W is an indicator equal to one if the individual was born in the fourth quarter and equal to zero otherwise.

First we evaluate the information content of the prior distribution for the parameter of interest γ. In order to do so, we shall calculate the expected posterior distribution $\bar{\pi}_m^\gamma$ as in (6), with $m = 10$ observations. We compare these expected posteriors with the actual posterior distribution based on the full sample with $n = 162,515$ observations. Here are some of the quantiles for the γ distributions:

quantile:	.025	.05	.25	.50	.75	.95	.975
$\bar{\pi}_{10}^\gamma$:	-2.43	-1.02	-.09	.07	.23	1.22	2.51
$\pi_n^\gamma(\cdot\,\vert\, d)$:	.047	.054	.075	.089	.104	.124	.132
$\mathcal{N}(.089, .021^2)$:	.048	.055	.075	.089	.103	.124	.130

It appears that the prior distribution is reasonably uninformative for γ, so that the posterior distribution is mainly reflecting the sample information.

The instrumental-variables estimate $\hat{\gamma}$ (i.e., the solution to $\sum_{i=1}^n \psi(z_i, \hat{\beta}) = 0$, where $\hat{\beta}' = (\hat{\alpha}, \hat{\gamma})$) is .089. An asymptotic approximation to its sampling distribution (allowing for heteroskedasticity of unknown form) gives a normal distribution with mean γ and standard deviation .021. A normal distribution with mean .089 and standard deviation .021 would in fact provide a good approximation to our posterior distribution.

4. QUANTILE REGRESSION

The second application illustrates how the posterior distribution can be well-defined when standard approximations to the sampling distribution are not appropriate. Let $Z = (X, Y)$, where Y is scalar and X is $K \times 1$. We can define a linear predictor corresponding to the τ^{th} quantile as follows: $E^*_\theta(Y \mid X = x) = \beta' x$, where

$$\beta = \arg\min_t E_\theta[c_\tau(Y - t'X)]$$

$$c_\tau(t) = |t| \cdot [(1-\tau) \cdot 1(t < 0) + \tau \cdot 1(t \geq 0)].$$

(β in general depends upon τ, but this should be clear from the context.) If $\tau = .5$, then this reduces to minimizing the mean absolute error: $\min_t E_\theta(|Y - t'X|)$. By weighting the absolute error differently for positive and negative values, the "check" function $c_\tau(\cdot)$ extends this notion of linear predictor to other quantiles. The role of the check function in quantile regression was developed by Koenker and Bassett (1978, 1982).

Our simulation procedure produces independent draws $\{\beta^{(l)}\}_{l=1}^{L}$ from the posterior distribution of β. To obtain $\beta^{(l)}$, first take i.i.d. draws $\{V_i^{(l)}\}_{i=1}^{n}$ from a standard exponential distribution. Then solve

$$\beta^{(l)} = \arg\min_t \sum_{i=1}^{n} V_i^{(l)} c_\tau(y_i - t'x_i)$$

(where the observed value of Z_i is $z_i = (x_i, y_i)$). The computations are simplified by exploiting the fact that $rc_\tau(t) = c_\tau(rt)$ if $r \geq 0$. So define $Y_i^{(l)} = V_i^{(l)} y_i$ and $X_i^{(l)} = V_i^{(l)} x_i$. Then

$$\beta^{(l)} = \arg\min_t \sum_{i=1}^{n} c_\tau(Y_i^{(l)} - t'X_i^{(l)}).$$

This is a linear programming problem, and we use the Barrodale-Roberts (1973) modification of the standard simplex algorithm.

Our application is based on "Workers' Compensation and Injury Duration: Evidence from a Natural Experiment," by Meyer, Viscusi, and Durbin (1995). The authors (MVD) obtained data for two states, Kentucky and Michigan, on a random sample of indemnity claims. We shall focus on Kentucky. The claims were filed by workers seeking compensation for work-related injury or illness. MVD concentrate on temporary total disability claims. Such a claim is filed when the person is unable to work but is expected to recover fully and return to work. The data include date injured, duration of temporary total benefits, total medical costs, previous wage, weekly benefit amount, type of injury (body part affected and the type of damage), age, sex, marital status, and an industry code.

The amount of the weekly benefit is based on a schedule that determines the benefit as a function of previous earnings. The schedule has a ceiling, with earnings levels above a threshold corresponding to the same weekly benefit. Kentucky raised the maximum benefit from \$131 to \$217 per week on July 15, 1980.

MVD work with claims that have injury dates during the year before or the year after the change in the benefit schedule. They also limit the sample to a high earnings group and a low earnings group. The weekly benefit amount for the high earnings group was affected by the increase in the benefit ceiling, whereas the benefit amount for the low earnings group was not affected. So the low earnings group can provide a control for period effects. The basic specification in MVD is

$$E_\theta(Y \mid X = x) = \beta_1 + \beta_2 \cdot x_2 \cdot x_3 + \beta_3 \cdot x_2 + \beta_4 \cdot x_3. \tag{7}$$

($x_1 \equiv 1$ denotes a constant.) Here Y = log of duration, with duration measured by weeks of temporary total benefits paid; $x_2 = 1$ if injured after the benefit increase, $x_2 = 0$ otherwise; $x_3 = 1$ if high earnings group, $x_3 = 0$ otherwise. The key coefficient is β_2, measuring the effect of the benefit increase on time out of work, with controls for period and for the earnings group:

$$\begin{aligned}\beta_2 = &\Big[\, E_\theta(Y \mid x_2 = 1, x_3 = 1) - E_\theta(Y \mid x_2 = 0, x_3 = 1)\,\Big] \\ &- \Big[\, E_\theta(Y \mid x_2 = 1, x_3 = 0) - E_\theta(Y \mid x_2 = 0, x_3 = 0)\,\Big] .\end{aligned}$$

An appealing aspect of the MVD analysis is that it is plausible to regard the injury date, and hence the applicable benefit schedule, as if it were randomly assigned.

To account for possible changes in the composition of the sample after the benefit increase, MVD also include regression controls for attributes of the individual, the job, and the injury—sixteen regressors in addition to the four in (7). The last column of Table 1 presents least-squares estimates (and conventional standard errors) corresponding to Table 6 in MVD. The first five columns of Table 1 present estimates of the linear predictor coefficients corresponding to the .10, .25, .50, .75, and .90 quantiles. These estimates are based on the simulation procedure described above. The point estimates are posterior medians and the "standard errors" in parentheses are constructed so that the point estimate plus or minus 1.96 standard errors gives an interval with a .95 posterior probability. The key coefficients (corresponding to β_2 in (7)) are in the second row. The effect of the benefit increase is fairly constant across the quantiles, suggesting a location model in which the distribution of log duration shifts rigidly in response to the benefit increase.

Table 2 presents results using duration out of work (in weeks) instead of its logarithm. Now the estimates show a substantial increase as we go from low to high quantiles, suggesting that the effect of the benefit increase is concentrated on the upper half of the duration distribution. The estimated effect on the median of the distribution is .87 weeks, with a standard error of .23. In contrast, the least-squares estimate of the effect on the mean of the distribution is quite imprecise, with a point estimate of 1.66 and a standard error of 1.04.

The histogram of the draws from the posterior distribution of β_2 is shown in Figure 1 for $\tau = .5$, using duration in weeks. The posterior mean is .87, and the posterior standard deviation is .23. So assuming the posterior distribution is normal and using $.87 \pm 1.96 \times .23$ gives a probability interval close to the one we constructed without assuming normality.

We examine the influence of the prior distribution by calculating the expected posterior distribution $\bar{\pi}^{\beta}_{m}$ as in (6), for $m = 21$ observations, and comparing this distribution with the posterior distribution $\pi^{\beta}_{n}(\cdot \,|\, d)$ based on the full sample with $n = 5349$ observations. Here are some of the quantiles of the β_2 distributions for $\tau = .5$, using duration in weeks:

quantile:	.025	.05	.25	.50	.75	.95	.975
$\bar{\pi}_{21}^{\beta_2}$:	-290	-157	-20.4	1.01	24.3	184	323
$\pi_n^{\beta_2}(\cdot \mid d)$:	.41	.49	.71	.87	1.03	1.25	1.32

The prior distribution is dominated by the sample information.

Now consider dropping all the predictor variables except for the four that appear in (7): 1, $x_2 \cdot x_3$, x_2, x_3. We shall compare the expected posterior distribution for $m = 5$ observations with the posterior distribution based on the full sample. Here are quantiles of these distributions for β_2 with $\tau = .5$, using duration in weeks:

quantile:	.025	.05	.25	.50	.75	.95	.975
$\bar{\pi}_{5}^{\beta_2}$:	-121	-36	-6	1	9	59	110
$\pi_n^{\beta_2}(\cdot \mid d)$:	0	0	1	1	2	2	2

The posterior histogram for β_2 is in Figure 2. It is concentrated on just four points: -1, 0, 1, and 2 weeks, with posterior probabilities of .01, .14, .55, and .30. This reflects the discreteness of the benefit duration distribution. The upper tail of that distribution is somewhat continuous, but 56% of the distribution is concentrated on the integers from 0 to 4 weeks. The $(.5, .75, .9, .95, .975)$ quantiles are $(4, 8, 15, 25, 49)$ weeks. Including the long list of predictor variables smoothes out this discreteness in the outcome variable, in the sense of producing a residual distribution (for $Y - \beta' X$) that is much closer to being continuous.

Here are the quantiles of the β_2 distributions for $\tau = .9$, using just the four regressors in (7) and duration in weeks:

quantile:	.025	.05	.25	.50	.75	.95	.975
$\bar{\pi}_{5}^{\beta_2}$:	-145	-41	-7	1	10	72	124
$\pi_n^{\beta_2}(\cdot \mid d)$:	2	3	5	7	8	11	12

The posterior histogram for β_2 is in Figure 3. This is closer to a normal distribution, corresponding to the continuity in the upper tail of the duration distribution.

The standard asymptotic distribution theory for quantile regression requires that the distribution of the residual $Y - \beta' X$ (conditional on θ) be absolutely continuous with a positive density in a neighborhood of zero. This requirement may be satisfied because the distribution of Y conditional on X is continuous. Alternatively, even if Y is discrete, it may be satisfied because $X'\beta$ is continuous. For example, with Y binary and X uniform on $[0, 1]$, and $E[Y|X] = X$, the limiting distribution of the coefficient in a quantile regression is normal despite the binary nature of Y. In our example Y is discrete with most mass concentrated on a few values. With only three binary regressors, the resulting distribution of the residual is still highly discrete. With the long list of regressors, even though many of them are discrete, the continuity requirement for the residual is much closer to being satisfied, and the standard large sample approximation to the sampling distribution is more accurate. In contrast, our posterior distributions provide straightforward inferences that do not rely upon the approximate normality of a sampling distribution.

5. CONCLUSION

The Bayesian approach to inference provides an attractive conceptual framework due to its connection with optimization concepts in decision theory and its lack of reliance on large-sample approximations. In practice, its use has been limited by the requirement of a fully specified parametric model since many econometric models are only partly specified. In this paper we have presented two applications of a less parametric Bayes approach due to Ferguson (1973, 1974) and Rubin (1981). In the first application, the decision-theoretic nature of the underlying question forces the use of posterior distributions rather than sampling distributions. In the second application, the assumptions underlying the asymptotic normality of the sampling distributions are clearly violated, but inference based on posterior distributions is straightforward.

FOOTNOTES

[1] The authors thank David Cox, Jinyong Hahn, and Neil Shephard for helpful comments, and Alan Krueger and Bruce Meyer for making their data available to us. The National Science Foundation provided financial support.

REFERENCES

Angrist, J., and Krueger, A. (1991): "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics*, 106, 979–1014.

Barberis, N. (2000): "Investing for the Long Run when Returns are Predictable," *Journal of Finance*, 55, 225–264.

Barrodale, I., and Roberts, F. (1973): "An Improved Algorithm for Discrete l_1 Linear Approximation," *SIAM Journal of Numerical Analysis*, 10, 839–848.

Chamberlain, G., and Imbens, G. (1995): "Semiparametric Applications of Bayesian Inference," Harvard Institute of Economic Research, Discussion Paper No. 1716.

Ferguson, T. (1973): "A Bayesian Analysis of Some Nonparametric Problems," *The Annals of Statistics*, 1, 209–230.

Ferguson, T. (1974): "Prior Distributions on Spaces of Probability Measures," *The Annals of Statistics*, 2, 615–629.

Hirano, K. (1999): "Semiparametric Bayesian Models for Dynamic Earnings Data", working paper, UCLA.

Kandel, S., and Stambaugh, R. (1996): "On the Predictability of Stock Returns: An Asset-Allocation Perspective," *Journal of Finance*, 51, 385–424.

Koenker, R., and Bassett, G. (1978): "Regression Quantiles," *Econometrica*, 46, 33–50.

Koenker, R., and Bassett, G. (1982): "Robust Tests for Heteroscedasticity Based on Regression Quantiles," *Econometrica*, 50, 43–61.

Lancaster, T. (1994): "Bayes WESML: Posterior Inference from Choice-Based Samples," unpublished manuscript, Brown University.

Meyer, B., Viscusi, W. K., and Durbin, D. (1995): "Workers' Compensation and Injury Duration: Evidence from a Natural Experiment," *American Economic Review*, 85, 322–340.

Rossi, P., McCulloch, R. and Allenby, G. (1995): "Hierarchical Modelling of Consumer Heterogeneity: An Application to Target Marketing," in *Case Studies in Bayesian Statistics*, Volume II, *Lecture Notes in Statistics*, 105, eds. C. Gatsonis, J. Hodges, R. Kass, and N. Singpurwalla, New York: Springer-Verlag, 323–349.

Rubin, D. (1981): "The Bayesian Bootstrap," *The Annals of Statistics*, 9, 130–134.

TABLE 1

Quantile Regression Coefficients for Log of Duration, Kentucky
High and Low Earnings Groups Pooled

	Quantile					
Variables	.10	.25	.50	.75	.90	OLS
Intercept	-5.555	-3.067	-1.749	-0.811	-1.239	-1.994
	(0.817)	(0.497)	(0.403)	(0.490)	(0.692)	(0.410)
After increase	0.136	0.141	0.164	0.170	0.137	0.145
∗High earnings group	(0.102)	(0.057)	(0.053)	(0.060)	(0.088)	(0.051)
After increase	-0.008	-0.039	-0.029	0.013	0.074	0.000
	(0.073)	(0.042)	(0.034)	(0.040)	(0.057)	(0.033)
High earnings group	1.755	0.525	0.024	-0.792	-3.191	-0.696
	(1.352)	(0.931)	(0.771)	(1.014)	(1.692)	(0.806)

Note: The dependent variable is ln(.5 + duration). The sample size is 5349. The additional regressors are Ln(previous wage), Ln(previous wage)∗High earnings group, Male, Married, Ln(age), Ln(total medical costs), Hospital stay indicator; *Industry indicators*: Manufacturing, Construction; *Injury type indicators*: Head, Neck, Upper extremities, Trunk, Low back, Lower extremities, Occupational diseases. The omitted industry is other industries, and the omitted injury is other injuries.

TABLE 2

Quantile Regression Coefficients for Duration, Kentucky
High and Low Earnings Groups Pooled

	Quantile					
Variables	.10	.25	.50	.75	.90	OLS
Intercept	-6.199	-7.258	-8.972	-11.566	-19.848	-25.886
	(1.157)	(1.441)	(1.779)	(3.310)	(7.254)	(8.412)
After increase	0.229	0.302	0.873	1.351	2.661	1.665
*High earnings group	(0.143)	(0.165)	(0.230)	(0.554)	(1.339)	(1.043)
After increase	-0.052	-0.032	-0.116	0.122	0.498	0.457
	(0.085)	(0.097)	(0.138)	(0.289)	(0.629)	(0.674)
High earnings group	0.051	-0.356	-1.655	-11.541	-56.802	-41.783
	(2.546)	(2.848)	(3.528)	(9.299)	(27.400)	(16.539)

Note: The dependent variable is duration (in weeks). The sample size is 5349. The additional regressors are the same as in Table 1. The omitted industry is other industries, and the omitted injury is other injuries.

Figure 1. Posterior Histogram, q = .5 (long list for x)

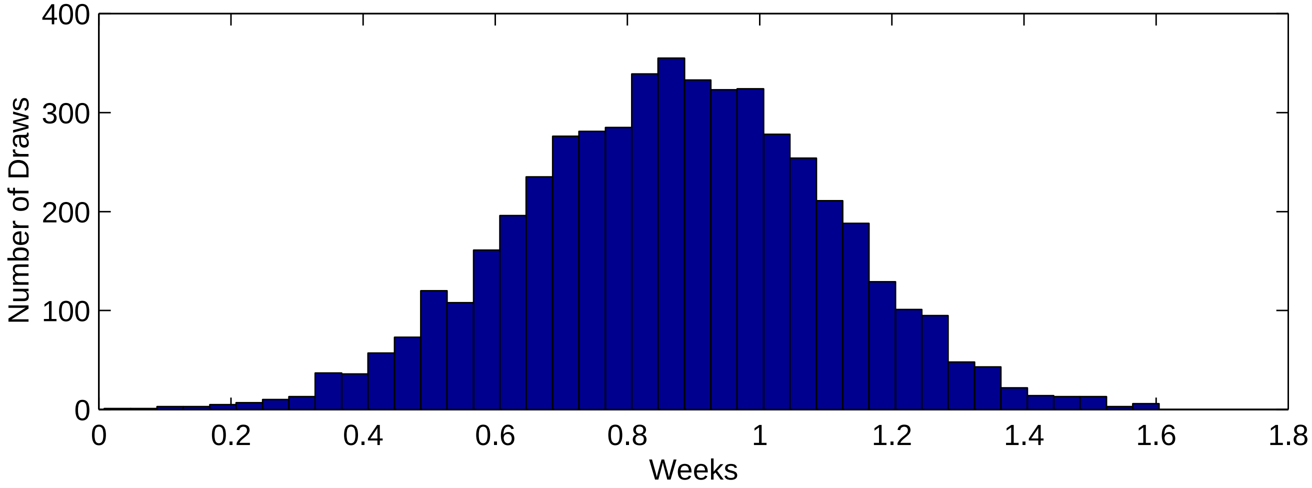


Figure 2. Posterior Histogram, q = .5 (short list for x)

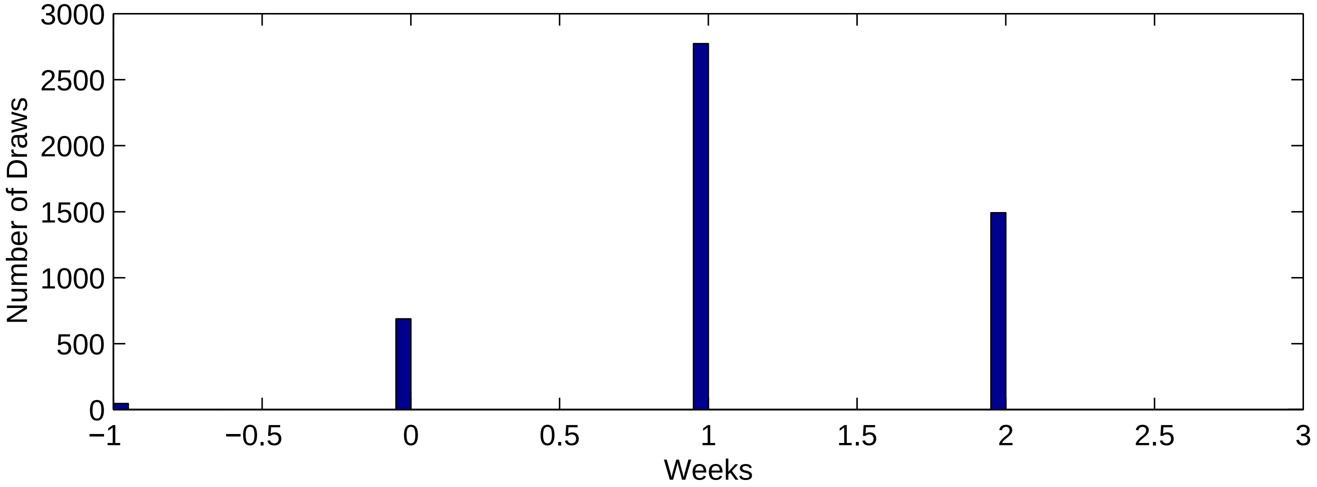
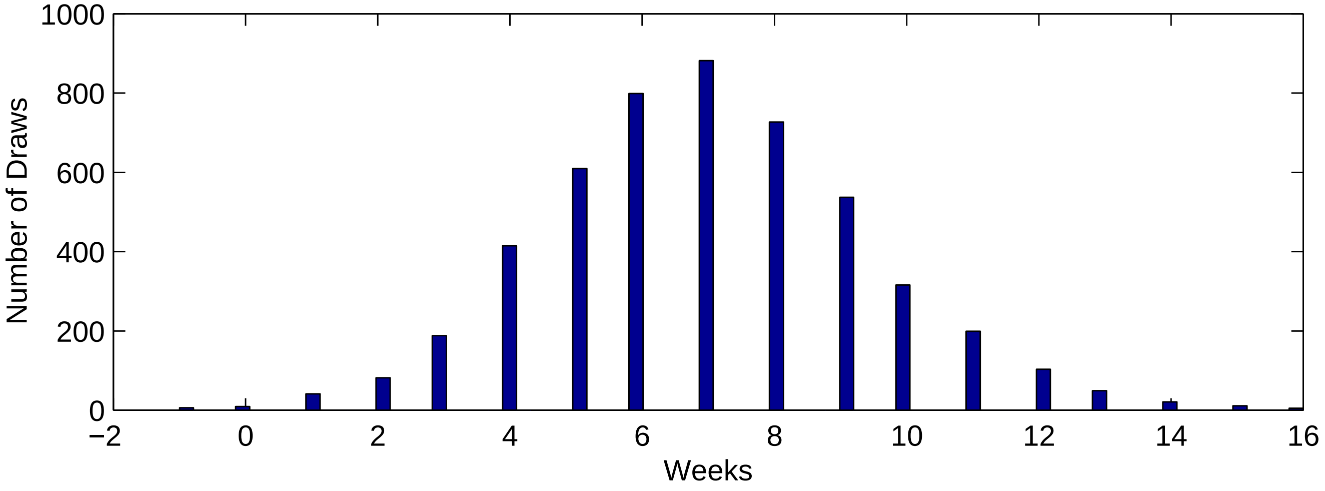


Figure 3. Posterior Histogram, q = .9 (short list for x)


Number of Draws
Weeks