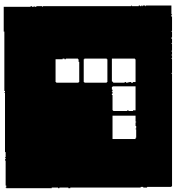
1998 V26 3: pp. 469–492
REAL ESTATE
ECONOMICS



Analyzing Real Estate Data Problems Using the Gibbs Sampler

John R. Knight,* C.F. Sirmans,** Alan E. Gelfand,*** and Sujit K. Ghosh****

Real estate data are often characterized by data irregularities: missing data, censoring or truncation, measurement error, etc. Practitioners often discard missing- or censored-data cases and ignore measurement error. We argue here that an attractive remedy for these irregularity problems is simulation-based model fitting using the Gibbs sampler. The style of the paper is primarily pedagogic, employing a simple illustration to convey the essential ideas, unobscured by implementation complications. Focusing on the missing-data problem, we show dramatic improvement in inference by retaining rather than deleting cases of partially observed data. We also detail Gibbs-sampler usage for other data problems.

Real estate data are messy and difficult to obtain. While the value of U.S. real estate assets is greater than that of stocks and fixed-income securities combined, real estate lags behind mainstream finance and economics in empirical research. This disparity no doubt is driven largely by differences in data availability and quality. Difficulty in obtaining timely and reliable data remains a major obstacle to examining the relationships among key real estate variables, a point neatly summarized with a single statement in Fogler, Granito and Smith (1985): "It is well known that statistical analysis using publicly available real estate data is tenuous at best."

Nevertheless, real estate research continues to expand in a number of directions. Examples include methodological advances in estimating residential and commercial real-estate values, comparisons of rates of returns and risk profiles for alternative real estate investments, pricing a variety of real estate options, such as the implicit default option in mortgages or the option to wait to develop land, and identifying the effects of institutional

*University of the Pacific, Stockton, California 95211 or JKnight@uop.edu.
**University of Connecticut, Storrs, CT 06269-2041 or CF@sbaserv.sba.uconn.edu.
***University of Connecticut, Storrs, CT 06269-2041 or Alan@stat.uconn.edu.
****North Carolina State University, Raleigh, NC 27695-8203 or Ghosh@stat.ncsu.edu.

practices. One common thread linking these and other diverse areas of real estate empirical research is messy data.

The nature of the real estate asset seriously impedes the ability to accurately estimate real-estate relationships. Properties are heterogeneous products that trade infrequently in markets that are highly localized. Moreover, data are often difficult to obtain, transaction records frequently contain empty data fields, and variables thought to influence relationships are routinely measured with error. Even when the researcher is able to overcome these problems, the ability to generalize estimates is hampered by the fact that transactions are infrequent, so that typically only a small portion of the population is represented in any cross section of data. Such data irregularities confound inference regarding real estate relationships using customary methods.

In this paper we introduce the Gibbs sampler, a simulation technique particularly well suited for fitting models in the presence of the above data irregularities. We demonstrate the effectiveness of the technique for a problem common to many real estate studies—missing data—and also describe its adaptation to several other frequently encountered situations.

The Gibbs sampler is a Monte Carlo sampling method which provides sample vectors, each approximately from a joint distribution of interest. It is implemented iteratively by making random draws from suitable conditional distributions. The sample obtained by using a particular component of these vectors is approximately from the marginal distribution of that component variable. From this description, connection of the Gibbs sampler to model fitting may not be apparent. We offer clarification in the sequel noting here that it has been applied to a wide variety of settings outside of real estate, and has been useful both in fitting complex models and in handling the challenges posed by data gathered outside of a laboratory environment. In the latter case the Gibbs sampler offers an alternative to other, sometimes *ad hoc*, methods of dealing with data inadequacies.

This paper is organized as follows. In the next section, we discuss a variety of data problems that confront real estate researchers and for which our technique is especially useful. Then, to ease the reader into the ensuing technical details, we provide an overview of simulation-based model fitting and the role of the Gibbs sampler. We next formalize details of the Gibbs sampler and its implementation in missing-data problems, and follow that with its application to a specific problem in real estate research: missing data in the hedonic estimation of house value. After demonstrating its effectiveness for this specific problem, we explain how the Gibbs sampler could be adapted to a number of other common real estate data inadequacies.

Date Inadequacies in Real Estate Problems

While data problems present a challenge to virtually all areas of empirical research (Morgenstern 1963), the real estate field is particularly rife with examples. Anyone who has gathered, processed or utilized real estate data appreciates the difficulties involved. Real estate research, therefore, provides a number of potential applications for the Gibbs sampler. We discuss below some examples of specific data inadequacies to which the technique has been beneficially applied in other contexts. The examples are not intended to be comprehensive, but rather to provide the reader with ideas for additional applications.

Missing Data

It is a rare data set that contains entries for all of the variables for each observation. Much more common are pieces of missing information for some observations where data are proprietary, unknown or otherwise unavailable, or simply inadvertently omitted at the point of data entry. Applications, surveys and questionnaires are often only partially completed by respondents. Also, when data for a study are collected from diverse sources, it is common for some data to be consistently missing from some of the sources.

The extent of damage caused by missing data depends on the number of records missing data relative to the number of complete records, and on the potential influence of the incomplete records on estimation. The conventional solutions are: (1) eliminate from the data set those records that are missing data, (2) interpolate or use some other measure of expected value in the data field or (3) return to the data source to attempt to determine the "true" value.

Eliminating observations is by far the most common method of dealing with missing data. This was the approach adopted in many well-known studies covering a wide array of empirical real estate research. Some specific examples from well-known studies include: separating consumption and investment effects on housing demand and tenure choice (Ioannides and Rosenthal 1994), estimating demand for the characteristics of housing (Palmquist 1984), investigating racial bias in mortgage lending (Munnell *et al.* 1996), predicting mortgage default (Vandell *et al.* 1993), and estimating the determinants of mortgage choice (Brueckner and Follain 1988). We choose these articles not as a criticism of the methodologies contained therein, but rather to emphasize two points: (1) missing data is an affliction that covers the gamut of real estate research; and (2) articles appearing in

top journals are not immune. In any event, the strategy of throwing out observations with missing data ignores potentially useful information in those records. The consequences are particularly adverse when the data set is small to begin with or when there is a correlation between those records missing data and the level of other variables to be estimated.

A second approach to the missing-data problem is to estimate what the value of the data field would be if it were available. In the case of time-series data, estimation frequently takes the form of interpolation between two data points that are observed. Interpolation, for example, is the approach employed by Engle, Lilien and Watson (1985) in their DYMIMIC model of housing price determination, and by Rosen and Smith (1983) in explaining the price adjustment process for rental housing. Missing data may also be replaced with simple or weighted averages of the available data in that data field, an approach adopted by Berkovec and Fullerton (1992), who used observations with enough data to impute remaining necessary variables. If imputation is an objective, such estimation is appropriate. But for model fitting, it serves no purpose. The imputed values are functions of the observed data, and hence can add nothing more to the fitting effort than what is provided by the observation.

Of course, the preferred means for dealing with missing data is to return to the source and find the "true" value for the data entry. Often, though, this is simply not possible or is prohibited by cost. The E–M algorithm (see Dempster, Laird and Rubin 1977) is widely used for missing-data problems. The Gibbs sampler offers an alternative solution with advantages detailed in the sequel.

Measurement Error

An errors-in-variables problem occurs whenever the true value of an explanatory variable is either unobservable, or observable but measured with error. Measurement error exists in almost all empirical work, manifesting itself when the data available to the researcher diverge from the variables prescribed by the model. The choice confronting the researcher is either to omit the variables for which the data are imperfect, or to accept some less than ideal substitute for the data. Omitting relevant variables, unless they are orthogonal to included variables, results in biased estimates of the included variables. Epple (1987) notes that orthogonality is especially unlikely in housing models because of the functional relationship suggested for suppliers, demanders and the product itself. While his comments are in the context of hedonic pricing models for housing, the same interactions

apply to many real estate relationships in addition to housing, a fact which suggests a more general interpretation may be appropriate.

Rather than omitting important variables when empirically testing theoretical models, researchers often utilize data known to only approximate the variables values. For example, Brueckner and Follain (1988) in their study of mortgage choice derive both borrower income and the age of the borrower from interval data. In estimating development costs in her test of real option pricing models, Quigg (1993) assigns the midpoint for the range of costs provided by Marshall Valuation Service to buildings based on their type, quality, location and time of construction. Clearly, in these cases as in countless others that employ interval data, the actual values of the variables would be preferred. Often, though, data are only collected on an interval basis, and sometimes proprietary and privacy concerns restrict the availability of actual values at the unit level.

Measurement error is also introduced when biased estimates of right-hand-side variables are substituted for actual values. For instance, Titman and Torous (1989) employ the values of mortgaged buildings as one of two state variables in their empirical test of the contingent-claims model's ability to explain default premia. They recognize that both sources of data for this variable produce biased estimates: commercial real-estate indices because they are based on appraisals rather than market prices [as discussed by Geltner, (1991)]; REITs because they are leveraged investments. Munnell *et al.* (1996), in their recent study of racial bias in mortgage lending, control for the risk associated with a particular property by using the ratio of rent to the value of the rental housing stock in the census tract where the property was located. The denominator of this ratio was unavailable and required estimation. Quigg (1993) uses the ratio of building price to development costs as the state variable to the option model and acknowledges that "both of these variables are likely to be observed with error." Measurement error is also possible when data from different data sets are merged.

The severity of the measurement error problem depends on the size of the error and the extent to which the error is contemporaneously correlated with the residual error term. However, because these aspects are never known, the errors-in-variables problem remains a thorny econometric issue. Weighted regressions and instrumental-variables approaches are the usual suggested refinements, but neither is completely satisfactory. By modeling the measurement error process, the Gibbs sampler offers an alternative approach to inference about the relationship between the response and the true value of an explanatory variable.

Censored Data

The data problems discussed up to this point have focused on the explanatory variables in the models. Sometimes data are available for independent variables but are missing for the dependent variable, and sometimes no data are observed unless the dependent variable falls in a certain category. Either situation results in biased estimates from ordinary least squares.

Within the real estate literature, the censored-sample problem has received considerable attention in the context of house price index construction. Clapp and Giaccotto (1992) have demonstrated sample selection bias in repeat sales indices, which use data only for properties that have sold more than once in the index period. Gatzlaff and Haurin (1997), also noting a repeat-sales-index bias for the locale of their study, go further to address the issue of censoring for all transactions-based house price indices; that is, transaction data are available only for homes that sell. This same censored-sample problem confronts assessors who perform mass appraisals for property tax allocation. Transaction prices are available only for homes that have changed hands, but values must be estimated for all homes in the taxing jurisdiction.

Of course, the censoring problem in real estate research extends beyond the estimation of house prices. Vandell *et al.* (1993), for example, note that specifying an arbitrary end point to a study of commercial mortgage defaults censors those loans that may yet default after the study period ends. Similarly, Zuehlke (1987) and Kluger and Miller (1990) point out difficulties in measuring the time-on-market variable in residential sales; both homes that sell after the data collection period ends and homes that are withdrawn from market are censored from the sample.

Maximum-likelihood fitting of logit, probit and Tobit models are conventional methods of correcting the bias in estimating models with limited dependent variables. The Gibbs sampler enables exact rather than approximate (based upon asymptotics) inference for such models.

Other Data Problems

The previous examples suggest the potential usefulness of the Gibbs sampler in addressing research issues that are clouded by inadequate data. However, the technique is also effective for classifying data by category when the category label is missing, and for predicting change points in cross-sectional or time-series data. Its flexibility and utility over a wide range of commonly encountered situations implies numerous opportunities for application to real estate problems. This should be still more apparent following our explanation

of its application to the missing-data problem, along with a specific example of its implementation which demonstrates a dramatic improvement in strength of inference.

Simulation-Based Model Fitting and the Gibbs Sampler

Though formal description of the Gibbs sampler for the data problems noted is straightforward, efficient implementation is somewhat of an art. Adopting more of a tutorial stance, we use standard Gaussian linear models in our illustration. As a result, technical detail is reduced to standard distribution theory, which we supply to assist the reader. To ease the transition, we first provide a conceptual picture of what the Gibbs sampler attempts to do and how it does it.

Usual statistical inference for real estate modeling is developed through analysis of the likelihood. In particular, estimation is carried out using the maximum-likelihood estimate (MLE) and its asymptotic (hence, approximate) standard error; hypothesis testing is carried out using a likelihood-ratio test (LRT) and its asymptotic chi-squared distributional approximation. In the case of Gaussian data, MLEs become least-squares estimates, and exact standard errors can often be obtained; LRTs provide familiar t and F statistics for identifying significant effects. In the case of no data irregularities there would be little reason to introduce an alternative inferential engine. However, in the presence of the aforementioned irregularities, it is not necessarily possible to implement the customary analysis without the unattractive discarding of incomplete cases.

As with likelihood analysis, we develop inference about model unknowns given the observed data. In fact, we view model unknowns as random variables. (After all, if quantities we can observe are assumed random, why not those we cannot observe?) But then, the model unknowns follow a joint distribution given the observed data, called the posterior distribution. Inference about model unknowns is expressed through features of this distribution. For example, for a particular component of the vector of model unknowns, if we seek a point estimate we might take a centrality measure of its posterior distribution such as the mean, median or mode. If we seek an interval estimate, we might take the 0.025 and 0.975 quantiles, whence, given the data, we have the direct interpretation that the probability is 0.95 that the unknown component will fall in this interval. Hypothesis testing for the component simply requires the probability that it falls in the hypothesized set given the data. The size of this probability, or its size relative to the probability of the complementary set (the odds), determines the strength of support for the hypothesis.

The posterior distribution of the unknown, through appropriate features, provides exact inference. We need not resort to distributional approximations such as asymptotic normality or to variability approximations such as asymptotic standard errors of MLEs obtained from the information matrix.

As we clarify in the next section, for problems of interest, the joint posterior distribution of the model unknowns will rarely be a standard distribution, and this will typically be the case for the marginal posterior distributions of a component as well. Hence, analytic calculation of a desired distributional feature is infeasible, and analytic approximations are usually of undeterminable accuracy. An alternative is to resort to the most fundamental of statistical ideas: To learn about a feature of the distribution or population, obtain samples from the distribution. In fact, by drawing a sample of arbitrarily large size we can achieve arbitrarily accurate estimation of the feature. This is precisely the simulation-based model-fitting procedure. Moreover, with the wide availability of high-speed computing, such an approach emerges as a very attractive (and possibly the only) technique for developing inference for complex models.

Where does the Gibbs sampler come in? As we discuss in the sequel, generally the distribution we must sample is high-dimensional and is provided in the form of a non-normalized density. (That is, we do not know the constant required to make the density integrable to one, and to obtain this constant would require an infeasible high-dimensional integration.) The Gibbs sampler is a tool which enables us to obtain samples from such joint densities, and therefore is a tool to implement simulation-based fittings.

The Gibbs sampler breaks the so-called *curse of dimensionality* by replacing draws from the high-dimensional joint density with draws from suitable low-dimensional conditional densities. Often these low-dimensional distributions are standard, as they are in all of the illustrations in this paper. But, in any event, they are always densities known up to normalizing constants and, though blocking may be beneficial, can always be taken to be univariate if convenient. There is a wealth of literature on sampling non-normalized univariate distributions.

Operationally, from a starting vector of values for the model unknowns, the Gibbs sampler proceeds to update components for the vector using the draws from the low-dimensional densities. An iteration is complete after each component has been updated. Unfortunately, this resultant updated vector is not a draw from the desired joint distribution. However, we can implement this updating process again, completing another iteration. In fact we can do this repeatedly and obtain a sequence of updated vectors. The elegance of

the Gibbs sampler is that this sequence provides a random path or trajectory for a stationary Markov chain whose limiting or stationary distribution is precisely the desired joint distribution. Hence, iterations sufficiently far along the trajectory will have essentially this joint distribution. In this way, the Gibbs sampler obtains samples from the joint posterior distribution. Implicitly, the sequence associated with any component of the vector converges in distribution to the marginal distribution of that component, hence providing samples from marginal distributions. Such an algorithm is an example of a Monte Carlo or simulation method using Markov chains, and is thus referred to as an MCMC algorithm. Issues such as where to start the chain, how many chains to run in parallel, when to stop the chains, how to make the chains converge to their stationary distributions more quickly, etc., characterize the art of working with such simulation methods.

In this way, model fitting using the Gibbs sampler differs dramatically from customary model fitting. We do not obtain analytic forms for the estimates. Typically, we cannot use routine software but rather must develop customized programs tailored to each application in order to implement the sampling. Lastly, since inference is based upon randomly generated quantities, we obtain random values for distributional features; two individuals working with the same joint distribution will obtain slightly different answers. These concerns apply to simulation-based model fitting in general, suggesting that such techniques be viewed as last choices rather than first. However, this pessimism is offset by the fact that these techniques enable us to fit and infer about an enormous range of otherwise inaccessible problems. The data-irregularity examples we study here only scratch the surface of possibilities.

Gibbs Sampling in the Missing-Data Case

The subsection below provides a general description of the Gibbs sampler and related issues. Following that, we provide details for an illustrative missing-data case.

Basics of the Gibbs Sampler

The objective is to sample from this joint posterior distribution for all model unknowns. Letting the data be denoted generically by y and letting all model unknowns be denoted generically by θ, to specify the model requires provision of a joint density for y and θ. We customarily write this as

$$f(y|\theta)f(\theta) \qquad (1)$$

The first factor, when viewed as a function of θ for observed y, is referred to as the likelihood and is denoted by $L(\theta; y)$. The second factor is referred to as the prior density for θ, capturing our knowledge about θ prior to collecting data. We typically assume little prior information about θ, for instance taking $f(\theta)$ constant. Then inference about θ is driven by the data and is essentially based upon $L(\theta; y)$, so that it will resemble likelihood-based inference except that we may avoid high-dimensional maximization and possibly inappropriate asymptotics. In practice, we would typically investigate several specifications for $f(\theta)$ to assess the sensitivity of inference to these choices.

Assuming that (1) is integrable with respect to θ, we see that the posterior distribution of θ, $f(\theta|y)$, is only known up to proportionality, *i.e.*, $f(\theta|y) \propto L(\theta; y)f(\theta)$. We do not know the normalizing constant. Also, in complex models we will have many unknowns, hence θ will be high-dimensional. Thus, we require samples from a high dimensional, non-standard, non-normalized density.

The Gibbs sampler is a tool for implementing such sampling. It was proposed in the context of image reconstruction by Geman and Geman (1984) and introduced into the mainstream statistical literature by Gelfand and Smith (1990). Suppose we partition θ into r blocks, *i.e.*, $\theta = (\theta_1, \ldots, \theta_r)$. The blocks might, in fact, be just the univariate components of θ, but often it is advantageous to create some grouping of components. The Gibbs sampler provides a trajectory of a Markov chain. Starting from some initial value, $\theta^{(0)}$, we update to $\theta^{(1)}$ given $\theta^{(0)}$, then to $\theta^{(2)}$ given $\theta^{(1)}$, etc. If the current state of θ is $\theta^{(t)} = (\theta_1^{(t)}, \ldots, \theta_r^{(t)})$, then we make the transition to $\theta^{(t+1)}$ as follows:

draw $\theta_1^{(t+1)}$ from $f(\theta_1|\theta_2^{(t)}, \ldots, \theta_r^{(t)}, y)$

draw $\theta_2^{(t+1)}$ from $f(\theta_2|\theta_1^{(t+1)}, \theta_3^{(t)}, \ldots, \theta_r^{(t)}, y)$

$\vdots$

draw $\theta_r^{(t+1)}$ from $f(\theta_r|\theta_1^{(t+1)}, \ldots, \theta_{r-1}^{(t+1)}, y)$

The $f(\theta_i|\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_r, y)$ are known as the full conditional distributions. Under (1), they uniquely determine $f(\theta|y)$. The process of updating each of the r blocks as indicated updates the entire θ vector, yielding one complete iteration of the Gibbs sampler. Under mild conditions, the sequence of $\theta^{(t)}$ converges in distribution to a draw from $f(\theta|y)$.

Typically we run several chains past *burn-in*, *i.e.*, so that the distribution of $\theta^{(t)}$ is essentially $f(\theta|y)$, and thereafter *thin* within a chain (retaining, say, every k^{th} iteration) to reduce dependence in the retained draws. Ultimately we obtain the set θ_l^*, $1 = 1, 2, \ldots, B$, which is viewed as essentially a sample of size B from $f(\theta|y)$. Efficient implementation of a Gibbs sampler introduces a whole range of practical issues including selection of starting values, number of chains to run, diagnosis of convergence, how to thin, etc. The review paper by Gelfand (1996), and references therein, may prove useful, as well as the recent book by Gilks, Richardson and Spiegelhalter (1995).

Inference using the set of θ_l^* is apparent. For any function of interest, say $\eta = h(\theta)$ (perhaps a component, a difference, a ratio, an indicator of a set, a maximum), the θ_l^* provide the set $\eta_l^* = h(\theta_l^*, l = 1, \ldots, B$, which are essentially a sample from the posterior distribution of η. Hence, $\Sigma_{l=1}^{B} \eta_l^*/B$ provides a point estimate for η, and, after ordering the η_l^*, we can estimate quantiles of this distribution and hence develop interval estimates with any desired coverage probability. In fact, a histogram of the η_l^*'s, perhaps smoothed, provides an estimate of the posterior density for η.

Prediction is often an important inference goal. Assume a regression setting where each independent response y_i has an associated fixed vector of explanatory variables x_i. Collecting these x_i's into a matrix X, in (1) $f(y|\theta)$ becomes $f(y|\theta, X)$. Otherwise none of the above discussions is changed, though we might write the posterior as $f(\theta|y, X)$. Now suppose we wish to predict y_0 at a given covariate vector x_0. From the simulation perspective we seek draws from the conditional distribution of y_0 given x_0; *i.e.*, $f(y_0|y, x_0)$, the *predictive* distribution for y_0. But

$$f(y_0|y, x_0) = \int f(y_0|\theta, x_0) f(\theta|y, X)\, d\theta \tag{2}$$

So, given θ_l^* as above, if we draw y_{0l}^*, from $f(y_0|\theta_l^*, x_0)$, the set of y_{0l}^*, $l = 1, \ldots, B$ is a sample from $f(y_0|y, x_0)$. As with the η_l^*, this sample enables inference about any feature of $f(y_0|y, x_0)$.

We have replaced the sampling of a high-dimensional density with sampling of r lower-dimensional, possibly univariate densities. Moreover, the $f(\theta_i|\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_1; y)$ are proportional to $L(\theta; y)f(\theta)$ when the latter is viewed as a function of θ_i with all other arguments fixed. Hence, we always know the full conditional distributions at least up to normalization. Gelfand (1996) and references therein discuss a variety of methods for efficiently sampling the full conditional distributions. However, for the

problems described in the introductory section, through the use of augmentation we can often obtain full conditional densities which can be recognized as standard distributions, *e.g.*, normals, gammas, Bernoulli trials, and uniforms. Sampling from such distributions is *canned* and is available in many software packages, so that a Gibbs sampler can be developed without the need to resort to writing code to implement alternative random generation techniques.

We now elaborate the above ideas in the context of missing-data settings for Gaussian linear models. Missing data in non-normal linear models (*e.g.*, binomial or Poisson regression) can be handled similarly.

Details for the Missing-Data Case

To simplify details, we illustrate in the case of a multiple linear regression with two covariates where it is routine to develop the required distribution theory. Hence, let $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$, $i = 1, \ldots, n$, where the ϵ_i are independent $N(0, \sigma^2)$ and the x_{1i} and x_{2i} are assumed for now to be fixed continuous regressors. First, suppose no missing data. Then $\theta = (\beta_0, \beta_1, \beta_2, \sigma^2)$ and $L(\theta; y) = (\sigma^2)^{-n/2} \exp[-\Sigma(y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i})^2/2\sigma^2]$. Suppose we take as our prior information $f(\theta) = f(\beta_0, \beta_1, \beta_2) f(\sigma^2)$ with $f(\beta_0, \beta_1, \beta_2) = 1$, *i.e.*, all $(\beta_0, \beta_1, \beta_2)$ equally likely, a noninformative specification. For σ^2, from the form of likelihood it is convenient to assume that $1/\sigma^2$ comes, *a priori,* from a gamma distribution; *i.e.*, σ^2 comes from an inverse gamma distribution. In our notation, the inverse gamma density is $f(\sigma^2) = b^a \exp(-b/\sigma^2)/\sigma^{2(a+1)} r(a)$ for $a > 0$, $b > 0$. If $a > 1$, the mean exists and is $b/(a-1)$. If $a > 2$, the variance exists and is $b^2/(a-1)^2(a-2)$. Hence, a rather imprecise or vague prior specification for σ^2 could set $a = 2$ (infinite variance) and b equal to the square of one-sixth of the anticipated range of the y_i (since for a Gaussian model the data range is approximately 6σ). We could pass to a limit and take $f(\sigma^2) = 1/\sigma^2$, which corresponds to a constant prior on $\log \sigma^2$. We would use a more informative choice if we did have useful prior information about σ^2, say from an earlier data set. In the sequel we will work with a generic a and b, writing $\sigma^2 \sim IG(a, b)$. We make a draw from $IG(a, b)$ by drawing z from $Ga(a, b)$ and setting $\sigma^2 = 1/z$.

Hence, following the discussion below (1),

$$f(\theta|y, X) \propto \frac{1}{\sigma^n} \exp\left(-\frac{\Sigma[y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})]^2}{2\sigma^2}\right) \cdot \frac{1}{\sigma^{2(a+1)}} \exp\left(-\frac{b}{\sigma^2}\right) \tag{3}$$

From (3), with a little algebra, it is apparent that

$$f(\beta_0|\beta_1, \beta_2, \sigma^2, y) = N\left(\frac{\Sigma(y_i - \beta_1 x_{1i} - \beta_2 x_{2i})}{n}, \frac{\sigma^2}{n}\right) \tag{4a}$$

$$f(\beta_1|\beta_0, \beta_2, \sigma^2, y) = N\left(\Sigma x_{1i}(y_i - \beta_0 - \beta_2 x_{2i}), \frac{\sigma^2}{\Sigma x_{1i}^2}\right) \tag{4b}$$

$$f(\beta_2|\beta_0, \beta_1, \sigma^2, y) = N\left(\Sigma x_{2i}(y_i - \beta_0 - \beta_1 x_{1i}), \frac{\sigma^2}{\Sigma x_{2i}^2}\right) \tag{4c}$$

$$f(\sigma_2|\beta_0, \beta_1, \beta_2, y) = IG\left(a + \frac{n}{2}, b + 0.5\,\Sigma(y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i})^2\right) \tag{4d}$$

All of the full conditional distributions for the Gibbs sampler are standard. Implementation of the required sampling is routine.

Now suppose we introduce missing data. For a given i we could be missing one or two of y_i, x_{1i} or x_{2i} (but not all three). Suppose, for instance, we are missing y_1 and x_{12}. Let us *augment* θ to $\theta' = (\beta_0, \beta_1, \beta_2, \sigma^2, y_1, x_{12})$ and view (3) as a function of θ'. The full conditional distributions for β_0, β_1, β_2 and σ^2 are precisely those in (4). We only need the full conditional distributions for y_1 and for x_{12}. From (3) it is clear that for y_1 we obtain $N(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}, \sigma^2)$, while for x_{12} we obtain $N((y_2 - \beta_0 - \beta_2 x_{22})/\beta_1, \sigma^2/\beta_1^2)$. Again implementation of the Gibbs sampler is routine.

In (3) we could delete cases $i = 1$ and $i = 2$ (as is typically done in practice) and work with the resulting reduced data, $D_{\text{red}} = \{(x_{1i}, x_{2i}, y_i), i = 3, \ldots, n\}$. How is our learning increased by retaining the observed (x_{11}, x_{21}) and (x_{22}, y_2)? Since we can integrate over the unobserved y_1 in (3), case $i = 1$ contributes no information regarding θ and should be deleted. This is not surprising. Without observing y_1, we obtain no information about the stochastic mechanism which generates y's from fixed x's. Expressed in a different way, since, as in (3), our analysis is conditional on the y we observed, regardless of the reason why y_1 was not observed, the $i = 1$ term in the likelihood should be discarded. The $i = 2$ case is different. We can integrate over x_{12}, but in doing so we acquire a $\sqrt{\beta_1^2}$ contribution to the denominator which revises the posterior from that for θ given D_{red}. There is information in the $i = 2$ term in (3). The general rule for fixed x's is that only cases with a missing response should be deleted.

It is noteworthy that, even though x_{12} is thought of as "fixed," when it is missing it is treated as random. In fact, implicit in the above full conditional distribution for x_{12} is a constant prior. In practice, we can instead use a proper but rather vague prior based upon the feasible range of the variable. For instance, we could assume, *a priori*, that $x_{12} \sim N(\mu_1, c\sigma_1^2)$, where μ_1 is the midpoint of the range with $6\sigma_1$ equal to the range and c is a constant greater than 1. By completing the square one can check that the full conditional for x_{12} is modified to

$$N\left(\frac{c\sigma_1^2\beta_1(y_2 - \beta_0 - \beta_2 x_{22}) + \sigma^2\mu_1}{c\sigma_1^2\beta_1^2 + \sigma^2}, \frac{c\sigma_1^2\sigma^2}{c\sigma_1^2\beta_1^2 + \sigma^2}\right).$$

The situation is changed if the covariates are viewed as stochastic regressors. Now the model specification provides a joint normal distribution for x_1, x_2 and y, resulting in a θ which has nine components (three mean parameters and six variance–covariance parameters). The likelihood becomes $L(\theta; y, X) = \Pi_{i=1}^n f(x_{1i}, x_{2i}, y_i|\theta)$, which we can factor say as $\Pi_{i=1}^n f(y_i|x_{1i}, x_{2i}, \theta)f(x_{1i}, x_{2i}|\theta)$. Hence, if y_i is not observed and we integrate over it, $f(x_{1i}, x_{2i}|\theta)$ remains to contribute to the likelihood. Thus, with stochastic regressors, the general rule is that *no* cases should be deleted.

Often x_1 is discrete, *i.e.*, a nominal or ordinal categorical variable. Returning to our illustration above, how would we now sample x_{12}? Suppose the possible values for x_1 are $a_1, \ldots, a_k$. Then assuming, *a priori,* that the a_l, $l = 1, \ldots, k$, are equally likely values for x_{12}, the full conditional distribution for x_{12} places mass q_l on a_l, where

$$q_l = \frac{\exp[-(y_2 - \beta_0 - \beta_1 a_l - \beta_2 x_{22})^2/2\sigma^2]}{\Sigma_{i=1}^k \exp[-(y_2 - \beta_0 - \beta_1 a_l - \beta_2 x_{22})^2/2\sigma^2]}, \qquad l = 1, 2, \ldots, k \quad (5)$$

We may trivially sample such a distribution by partitioning the unit interval as $(0, q_1]$, $(q_1, q_1 + q_2]$, etc. If we draw a uniform variable on the interval (0,1], we realize $x_{12} = a_l$ if the uniform observation belongs to $(\Sigma_{j=0}^{l=1} q_j, \Sigma_{j=0}^{l} q_j]$, where $q_0 = 0$.

Extension to more general multiple linear regression is clear. As above, we augment the parameter vector with all of the missing data (again, only explanatory variables with non-stochastic regressors). Depending upon whether the covariate is continuous or discrete, the associated full conditional distribution will either be normal or have a form like that in (5). We show in the next section that dramatically better inference results if entire cases are not deleted.

The output of the Gibbs sampler will be a sample of β's and σ's drawn from the posterior distribution given all the retained cases. These samples can be used to provide any desired inference summaries, *e.g.*, point and interval estimates for the parameters, for the mean at new x, or for the observation at a new x. In fact, entire posterior distributions are available for these unknowns as well as for any of the missing data, should imputation be of interest.

As a related remark, we can view the Gibbs sampler in the missing-data context as a simulation-based analog of the E–M algorithm (see Dempster, Laird and Rubin, 1977; Little and Rubin, 1989; Tanner 1993). Sampling the missing data given the incomplete data corresponds to the E-step; sampling θ given both the incomplete data and the draws from the missing data corresponds to the M-step.

In the present missing data illustration, we conclude that the E–M algorithm is an attractive alternative. It is easy to implement with less reliance on "art" and produces familiar MLEs (or, more generally, posterior modes) along with associated asymptotic variance estimates (Louis 1982). In fact, for a finance-related missing-data application, see Warga (1992). The Gibbs sampler is applied using this illustrative setting chiefly to provide simplified illumination of the basic ideas, unobscured by implementation complications.

Although our primary objective is to fit the model in (3), an advantage to using Gibbs sampling is the opportunity to attach inference to the imputation of the missing values. For instance, the samples of x^*_{12l}'s which would be generated with the Gibbs sampler provide an entire posterior distribution for this missing observation.

Advantage accrues to the Gibbs sampler particularly in more challenging modeling introducing non-Gaussian likelihood, nonlinear model specification, random effects, constrained parameters and data, etc. In such cases the E–M algorithm will not be viable because likelihood inference for the full model (without missing data) will not be feasible.

An additional advantage to Gibbs sampling is that the resulting inference is not based upon possibly inappropriate asymptotics. That is, likelihood-based inference relies upon approximate normality of the MLE, so that with a suitable asymptotic standard error, symmetric interval estimates about the MLE result. Simulation-based model fitting provides interval estimates which avoid asymptotics, are arbitrarily accurate and reflect any asymmetry in the posterior distribution.

A further advantage is that we can handle inference for arbitrary functions of θ. As noted earlier, the output of the Gibbs sampler provides posterior distributions for any $\eta = h(\theta)$. While the MLE for η is immediate, standard errors based upon the delta method and interval estimates based upon approximate normality may not be attractive.

An Application to Residential Sales Data

To illustrate the use of the Gibbs sampler for a missing-data application, we consider a sample from a large data set consisting of sales of detached single-family dwellings in Baton Rouge, Louisiana. Other researchers; *e.g.*, Knight, Sirmans and Turnbull (1994) and Gelfand *et al.* (1998), have also analyzed parts of this data set. The sample we selected consists of all sales in 1992 for ten subdivisions, a total of 621 transactions. From these, a total of 50 transactions, 5 at random from each subdivision, were held out for prediction.

Our simplified model takes log selling price as y with covariates square feet of living area, square feet of other covered area (garages, porches, etc.), and age. In addition, to attempt to capture additional heterogeneity across subdivisions, we introduce individual subdivision random effects. Such heterogeneity is a surrogate for other unmeasured explanatory variables. In particular we might anticipate that location is consequential, and thus include spatial effects in place of or in addition to the heterogeneity effects. This has been done in Gelfand *et al.* (1998), utilizing all 49 subdivisions in the data set. With a subset of only ten it will be difficult to capture spatial patterns; we confine our modeling to just the random effects, which we denote by α_i. Hence, for the j^{th} property in the i^{th} subdivision we set

$$y_{ij} = \alpha_i + \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \epsilon_{ij} \tag{6}$$

where y_{ij} is the log selling price, x_{1ij} is the square footage of living area, x_{2ij} is the square footage of additional living area and x_{3ij} is the age of the house. We assume the regressors are fixed in the ensuing analysis.

As is customary for random effects, we introduce a variance component and assume that the α_i are independent $N(0, \sigma_\alpha^2)$. We assume a constant prior on β_0, β_1, β_2 and β_3 and, following the discussion above (3), inverse gamma priors for σ^2 and σ_α^2, namely, $\sigma^2 \sim IG(2, 0.0159)$ and $\sigma_\alpha^2 \sim IG(2, 0.1547)$. These specifications arise on setting the prior mean for σ^2 equal to the mean squared error under usual least-squares fitting of (6) with infinite *a priori* variance, similarly for σ_α^2. The inference is not very sensitive to the choice

of prior means. Using a value based upon the range of the y's yields similar results.

Under these specifications, the resulting full conditionals are as follows:

$f(\beta_0|\beta_1, \beta_2, \beta_3, \alpha, \sigma^2, \sigma_\alpha^2; y)$ is

$$N\left(\frac{\Sigma_i\Sigma_j(y_{ij} - \alpha_i - \beta_1 x_{1ij} - \beta_2 x_{2ij} - \beta_3 x_{3ij})}{n}, \frac{\sigma^2}{n}\right)$$

$f(\beta_1|\beta_0, \beta_2, \beta, \alpha, \sigma^2, \sigma_\alpha^2; y)$ is

$$N\left(\frac{\Sigma_i\Sigma_j x_{1ij}(y_{ij} - \alpha_i - \beta_0 - \beta_2 x_{2ij} - \beta_3 x_{3ij})}{n}, \frac{\sigma^2}{\Sigma_{ij} x_{1ij}^2}\right)$$

$f(\beta_2|\beta_0, \beta_1, \beta_3, \alpha, \sigma^2, \sigma_\alpha^2; y)$ is

$$N\left(\frac{\Sigma_i\Sigma_j x_{2ij}(y_{ij} - \alpha_i - \beta_0 - \beta_1 x_{1ij} - \beta_3 x_{3ij})}{n}, \frac{\sigma^2}{\Sigma_{ij} x_{2ij}^2}\right)$$

$f(\beta_3|\beta_0, \beta_1, \beta_2, \alpha, \sigma^2, \sigma_\alpha^2; y)$ is

$$N\left(\frac{\Sigma_i\Sigma_j x_{3ij}(y_{ij} - \alpha_i - \beta_0 - \beta_1 x_{1ij} - \beta_2 x_{2ij})}{n}, \frac{\sigma^2}{\Sigma_{ij} x_{3ij}^2}\right)$$

$f(\alpha_i|\beta_0, \beta_1, \beta_2\beta_3, \alpha_l, l \neq i, \sigma^2, \sigma_\alpha^2; y)$ is

$$N\left(\frac{\sigma_\alpha^2\Sigma_j(y_{ij} - \beta_0 - \beta_1 x_{1ij} - \beta_2 x_{2ij} - \beta_3 x_{3ij})}{\sigma^2 + n_i\sigma_\alpha^2}, \frac{\sigma^2\sigma_\alpha^2}{\sigma_2 + n_i\sigma_\alpha^2}\right)$$

$f(\sigma^2|\beta_0, \beta_1, \beta_2, \beta_3, \alpha_i, \sigma_\alpha^2; y)$ is

$$IG\left(2 + \frac{n}{2}, 0.0159 + 0.5\,\Sigma_i\Sigma_j(y_{ij} - \alpha_i - \beta_0 - \beta_1 x_{1ij} - \beta_2 x_{2ij} - \beta_3 x_{3ij})^2\right)$$

$f(\sigma_\alpha^2|\beta_0, \beta_1, \beta_2, \beta_3, \alpha_i, \sigma^2; y)$ is $IG(2 + 5, 0.1547 + 0.5\,\Sigma_i\alpha_i^2)$

where $i = 1, \ldots, 10, j = 1, \ldots, n_i$, the number of transactions in subdivision i, and $n = \Sigma n_i = 571$, the total number of transactions in the analysis.

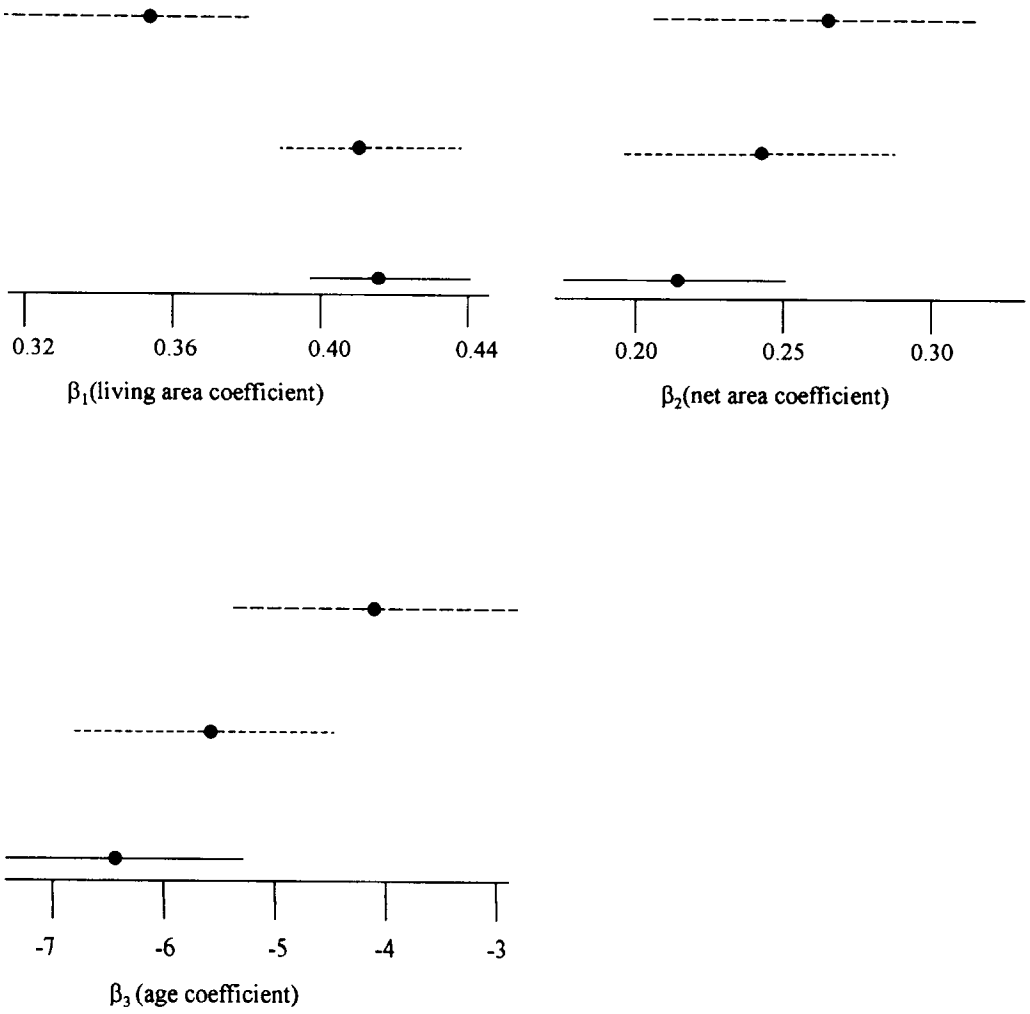
Having a complete data set allows us the opportunity to introduce missing data and then to compare inference under missing-data case deletion and

under augmentation using the Gibbs sampler with that for the complete data. In particular, we carried out the customary analysis of the full set of 571 observations, a total of $571 \times 4 = 2284$ data values. We then randomly selected 40% of the 571 cases. For each case selected we randomly assumed one of its four data values was missing, a total of 228 values. Hence 10% of the data are missing.

In Figure 1 we present point and 95% interval estimates for each of the three regression coefficients (ignoring β_0) under the full data set, under deletion of 40% of the transactions and under fitting using the Gibbs sampler

Figure 1 ■ Confidence interval estimates.

For the residential sales data example of section 5, this figure shows a comparison of 95% interval estimates for the regression coefficients associated with living area, additional or net area, and age obtained under three data conditions. ——— = full data; – – – – – = 40% of cases deleted; --------- = retained incomplete cases using the Gibbs sampler.



with missing data. In fact, we fitted each of these three situations using the Gibbs sampler, thereby obtaining for each one a posterior sample and hence for β_1 (say) a sample of β_{1l}^*'s. The point estimate for β_1 is the mean of this sample (though the sample median could be used), while the interval estimate is determined by the lower 0.025 and upper 0.025 quantiles of the sample. We see that when 40% of the data are discarded, point estimates are shifted and interval estimates are inappropriately centered and much longer than for the full data analysis. However, the Gibbs sampler with one in ten data missing performs remarkably well. Similar findings occur for the estimated regression lines and for the predictive distributions at new x's. We conclude that learning is dramatically increased by retaining the useful information in the missing-data transactions.

Turning to prediction we use the holdout sample of 50 cases, five from each subdivision. Following the discussion after (2), we obtain a sample of size B from each of the 50 predictive distributions under each of the three fitting situations. By exponentiation we can restore these samples from the log scale to the dollar scale. On the selling price scale, for a particular $\exp(y_0)$, denote the mean, median and variance of the predictive distribution by $\mu_0(y)$, $m_0(y)$ and $\sigma_0^2(y)$, respectively. The argument y is intended to remind the reader that the data used in fitting the model in (6) changes over the three fitting situations; hence, so do these features of the predictive distribution.

We can compute a version of a mean squared error (MSE) and a mean absolute error (MAE) for the 50 holdout transactions by calculating

$$\text{RMSE} = \sqrt{\frac{1}{50}\sum_{j=1}^{50}[\exp(y_{0j,\text{obs}}) - \mu_{0j}(y)]^2}$$

and

$$\text{MAE} = \frac{1}{50}\sum_{j=1}^{50}|\exp(y_{0j,\text{obs}}) - m_{0j}(y)|$$

respectively. We can also obtain an average predictive standard deviation (APSD) by computing

$$\text{APSD} = \frac{1}{50}\sum_{j=1}^{50}\sigma_{0j}(y)$$

Results are presented in Table 1. Note that, on retaining the useful information in the missing-data transactions, the predictive performance

Table 1 ■ Comparative predictive performance.

	RMSE	MAE	APSD
Full data	14296.26	10275.93	13195.73
40% of the data discarded	22298.85	17988.06	16585.26
Fitting using Gibbs sampler	14934.21	11061.61	13850.58

This table demonstrates potential benefits of the Gibbs sampler over the customary procedure of deleting incomplete data records. Under three common predictive performance criteria, the Gibbs sampler restores an impressive portion of the predictive ability. The predictive criteria are:

$$\text{Root mean square error (RMSE)} = \sqrt{\frac{1}{50}\Sigma_{j=1}^{50}\,[\exp(y_{0j,\text{obs}}) - \mu_{0j}(y)]^2}$$

$$\text{Mean absolute error (MAE)} = \frac{1}{50}\Sigma_{j=1}^{50}\,|\exp(y_{0j,\text{obs}}) - m_{0j}(y)|$$

$$\text{Average predictive standard deviation (APSD)} = \frac{1}{50}\Sigma_{j=1}^{50}\,\sigma_{0j}(y)$$

The holdout sample consists of fifty cases. $Y_{0j,\text{obs}}$ is the log of the selling price for the jth holdout case, and $u_0(\text{Y})$, $m_0(\text{Y})$, and $\sigma_0^2(Y)$ are the mean, median, and variance of the predictive distribution, respectively.

nearly matches that of the full data. On discarding 40% of the data, the predictive performance degrades dramatically.

The substantial advantage to retaining the missing-data cases as opposed to deleting them is apparent for both estimates of the relationship and prediction.

Other Applications

The Gibbs sampler is useful for a wide range of data irregularities. Here we briefly describe how the Gibbs sampler may be used to handle censored data and errors-in-variables problems. As we shall see, both contexts utilize the augmentation idea previously discussed. Missing-data problems can be accommodated in conjunction with either of these two contexts by additional augmentation, further demonstrating the versatility of the Gibbs sampler for handling a wide range of practical data analysis problems for real estate models.

Censored Data

We illustrate in the setting of a standard Tobit model with right censoring of the dependent variable. Other forms of censoring such as interval censoring (scoring) can be handled similarly. Suppose then that given β and σ^2 we have independent $y_i \sim N(x_i^T\beta, \sigma^2)$. As before, let $\theta = (\beta, \sigma^2)$ and assume a constant prior on β with an $IG(a, b)$ prior for σ^2. In the absence of censoring, the Gibbs sampler would proceed by sampling full conditionals analogous to those in (4). Suppose y_1 is censored so that we only know that $y_1 > c_1$, a known value. If we again think of y_1 as "missing" and augment θ to $\theta' = (\theta, y_1)$, then, given y_1, the full conditional distribution for the components of θ arise as if there were no censoring. What about the full conditional distribution for y_1? We know that given θ and c_1, y_1 follows a normal distribution with mean $x_1^T\beta$ and variance σ^2 restricted to the interval $[c_1, \infty)$. To sample such a y_1 we need only make a draw from the conditional normal density, retaining it if it exceeds c_1. This so-called rejection sampling method, while not necessarily most efficient, is adequate for most applications. As in the section "Gibbs Sampling in the Missing-Data Case," implementing the Gibbs sampler will produce θ_l'*'s, and hence θ_l*'s, from the posterior given both the censored and uncensored cases. In general, we augment θ with a *latent* y_i for each censored response. The full conditional distribution for each such y_i will be Gaussian over a restricted right tail set and can be sampled using the above rejection approach. By not deleting any censored cases we obtain stronger inference for θ. Additionally, we obtain the advantages over a likelihood analysis, as previously noted.

Errors-in-Variables Models

Measurement error frequently occurs in the explanatory variables. If we model the measurement error process, the Gibbs sampler may be used to infer about the relationship between the response and the true value of the explanatory variables. This contrasts with classical approaches, where an external validation sample is usually required. We illustrate in the case of the multiple regression model. Here we assume that the x_{1i}'s are observed without error but that the actual x_{2i}'s, $x_{2i,\text{act}}$, are not observed. Still one wants to infer about the relationship between y, x_1 and $x_{2,\text{act}}$. We assume that $x_{2i,\text{act}}$ varies randomly about the observed x_{2i}, $x_{2i,\text{obs}}$, the so-called Berkson specification; *i.e.*, $x_{2i,\text{act}} \sim N(x_{2i,\text{obs}}\ \tau^2)$.

Alternative specifications include replacing the mean $x_{2i,\text{obs}}$ with a monotone parametric function $g(x_{2i,\text{obs}}, \delta)$. Such a calibration function allows the possibility of capturing a situation where, for example, actual values tend to be above observed values when observed values are large, but below when

observed values are small. One could also envision $x_{2i,\text{act}}$ as a function of observed surrogate variables, leading to a regression form for the mean of $x_{2i,\text{act}}$. One could also exchange the roles, having $x_{2i,\text{obs}}$ vary about $x_{2i,\text{act}}$. The book of Fuller (1987) provides detailed discussion of all of these possibilities.

For our choice suppose again we augment θ to $\theta' = (\theta, \tau^2, x_{21,\text{act}}, \ldots, x_{2n,\text{act}})$. In addition to the earlier prior for θ, the above Berkson specification provides a prior for each $x_{2i,\text{act}}$. Adding an inverse gamma prior for τ^2 [*i.e.*, $\tau^2 \sim IG(c, d)$] completes the model specification. Turning to the full conditional distributions, those for the components of θ are as in (4) with x_{2i} taken as $x_{2i,\text{act}}$. For the $x_{2i,\text{act}}$, by completing the square, we obtain

$$N\left(\frac{\tau^2\beta_2(y_i - \beta_0 - \beta_1 x_{1i}) + \sigma^2 x_{2i,\text{obs}}}{\tau^2\beta_2^2 + \sigma^2}, \frac{\tau^2\sigma^2}{\tau^2\beta_2^2 + \sigma^2}\right).$$

Finally, for τ^2 we obtain an updated inverse gamma, $\text{IG}(c + n/2, d + \Sigma(x_{2i,\text{act}} - x_{2i,\text{obs}})$.

Conclusions

As we have shown, the Gibbs sampler offers advantages over existing methods for handling a number of data problems common to empirical research in real estate. In particular, relative to discarding observations with incomplete covariate information, we have demonstrated impressive gains in both estimation and prediction precision by using the Gibbs sampler to deal with missing data in a hedonic house price model. With the increasing availability and economy of high-speed computing, implementation of the Gibbs sampler is within the reach of most data analysts, thus reducing the justification for throwing away information.

Our purpose in this paper is to introduce the real estate academic community to a model-fitting technique that has been successfully used in other fields to accommodate problems that are very familiar in real estate. For pedagogic purposes, our demonstration of the approach is limited to a simple missing-data problem, but its potential for use beyond this problem is apparent.

References

Berkovec, J. and D. Fullerton. 1992. A General Equilibrium Model of Housing, Taxes, and Portfolio Choice. *Journal of Political Economy* 100(2): 390–429.

Brueckner, J.K. and J.R. Follain. 1988. The Rise and Fall of the ARM: An Econometric Analysis of Mortgage Choice. *The Review of Economics and Statistics* 70(1): 93–102.

Clapp, J. and C. Giaccotto. 1992. Estimating Price Indices for Residential Property: A Comparison of Repeat Sales and Assessed Value Methods. *Journal of the American Statistical Association* 87(418): 300–306.

Dempster, A.P., N.M. Laird and D.B. Rubin. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B Methodological* 39(1): 1–38.

Engle, R.F., D.M. Lilien and M. Watson. 1985. A DYMIMIC Model of Housing Price Determination. *Journal of Econometrics* 28(3): 307–326.

Epple, D. 1987. Hedonic Prices and Implicit Markets: Estimating Demand and Supply Functions for Differentiated Products. *Journal of Political Economy* 95(1): 59–80.

Fogler, H.R., M.R. Granito and L.R. Smith. 1985. A Theoretical Analysis of Real Estate Returns. *The Journal of Finance* 40(3): 711–719.

Fuller, W.A. 1987. *Measurement Error Models*. John Wiley and Sons: New York.

Gatzlaff, D.H. and D.R. Haurin. 1997. Sample Selection Bias and Repeat-Sales Index Estimates. *Journal of Real Estate Finance and Economics* 14(1): 33–50.

Gelfand, A.E. 1996. Gibbs Sampling. J. Kotz, C. Reed and D. Banks, editors. *The Encyclopedia of Statistical Sciences* (update). John Wiley and Sons: New York, 283–292.

Gelfand, A.E. and A.F.M. Smith. 1990. Sampling Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association* 85(410): 398–409.

Gelfand, A.E., S.K. Ghosh, J. Knight and C.F. Sirmans. 1998. Spatio-temporal Modeling of Residential Sales Data. *Journal of Business and Economic Statistics* 16(3): 312–321.

Geltner, D.M. 1991. Smoothing in Appraisal-Based Returns. *Journal of Real Estate Finance and Economics* 4(3): 327–345.

Geman, S. and D. Geman. 1984. Stochastic Relaxations, Gibbs Distributions and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6(6): 721–741.

Gilks, W.R., S. Richardson and D.G. Spiegelhalter. 1995. *Markov Chain Monte Carlo in Practice*. Chapman and Hall: London.

Ioannides, Y. and S. Rosenthal. 1994. Estimating the Consumption and Investment Demands for Housing and Their Effect on Housing Tenure Status. *The Review of Economics and Statistics* 76(1): 127–141.

Kluger, B. and N. Miller. 1990. Measuring Residential Real Estate Liquidity. *Journal of the American Real Estate and Urban Economics Association* 18(2): 145–159.

Knight, J.R., C.F. Sirmans and G. Turnbull. 1994. List Price Signaling and Buyer Behavior in the Housing Market. *The Journal of Real Estate Finance and Economics* 9(3): 177–192.

Little, R.J.A. and D.B. Rubin. 1989. *Statistical Analysis with Missing Data*. John Wiley and Sons: New York.

Louis, T.A. 1982. Finding the Observed Information Matrix When Using the E–M Algorithm. *Journal of the Royal Statistical Society. Series B Methodological* 44(2): 226–233.

Morgenstern, O. 1963. *On the Accuracy of Economic Observations*. Princeton University Press: Princeton, NJ.

Munnell, A.H., G.M.B. Tootell, L.E. Browne and J. McEneaney. 1996. Mortgage Lending in Boston: Interpreting HMDA Data. *The American Economic Review* 86(1): 25–53.

Palmquist, R.B. 1984. Estimating the Demand for the Characteristics of Housing. *The Review of Economics and Statistics* 66(3): 394–404.

Quigg, L. 1993. Empirical Testing of Real Option-Pricing Models. *The Journal of Finance* 48(2): 621–640.

Rosen, K.T. and L.B. Smith. 1983. The Price-Adjustment Process for Rental Housing and the Natural Vacancy Rate. *The American Economic Review* 73(4): 779–786.

Tanner, M. 1993. *Tools for Statistical Inference*. Springer-Verlag, New York.

Titman, S. and W. Torous. 1989. Valuing Commercial Mortgages: An Empirical Investigation of the Contingent-Claims Approach to Pricing Risky Debt. *The Journal of Finance* 44(2): 345–373.

Vandell, K., W. Barnes, D. Hartzell, D. Kraft and W. Wendt. 1993. Commercial Mortgage Defaults: Proportional Hazards Estimation Using Individual Loan Histories. *Journal of the American Real Estate and Urban Economics Association* 21(4): 451–480.

Warga, A. 1992. Bond Returns, Liquidity, and Missing Data. *Journal of Financial and Quantitative Analysis* 27(4): 605–615.

Zuehlke, T.W. 1987. Duration Dependence in the Housing Market. *The Review of Economics and Statistics* 69(4): 701–704.