RESEARCH SUPPORT
UNIVERSITY OF MICHIGAN BUSINESS SCHOOL

MARCH 1997

BAYESIAN BINARY REGRESSION WITH MEASUREMENT ERROR AND AUTOCORRELATION, WITH APPLICATION TO MORTGAGE DEFAULT MODELS

WORKING PAPER #9712-04

BY

MARTIN R. YOUNG
UNIVERSITY OF MICHIGAN
AND
DICKRAN KAZARIAN
LEHMAN BROTHERS

Bayesian Binary Regression with Measurement Error and Autocorrelation, with Application to Mortgage Default Models

Martin R. Young
University of Michigan Business School
Ann Arbor, MI 48109 U.S.A.
Email: myoung@umich.edu
Phone: 313-936-1332
Fax: 313-936-0274

Dickran Kazarian
Lehman Brothers

March, 1997

Summary

Accurate estimation of regression models for mortgage defaults is quite difficult, owing to the fact that a key explanatory variable, contemporaneous housing price, can only be observed with large measurement error. In this paper, we describe a binary regression model for mortgage defaults that accounts for measurement error in predictors, as well as possible autocorrelation in residuals, and develop a Markov chain Monte Carlo procedure for Bayesian estimation of the model parameters. Application to a large national database of mortgages shows that taking measurement errors and autocorrelation into account has a substantial effect on the estimates of binary regression coefficients.

Key Words: *Data Augmentation, Logistic Regression, Markov Chain Monte Carlo, Retrospective Sampling*

1 Introduction

There is over $3.8 trillion in mortgage debt outstanding in the United States; the mortgage debt market is larger than the U.S. treasury bond or corporate bond markets. A critical issue for the lending institutions that originate mortgage loans, and for the secondary investors who purchase bundles of mortgage contracts from the originating institutions, is the probability that any given borrower will default on the mortgage contract.

The default probability is known to depend on covariates pertaining to the loan. One of the covariates believed to most significantly influence default rate is the so-called loan to value (LTV) ratio (e.g., Quercia and Stegman 1992). The LTV ratio is the ratio of the current amount of principal due, to the current market value of the home. When the LTV ratio is high, owing to a decline in market value of the home, it may be in the borrower's economic interest to default on the loan, in effect exercising his or her implied put option to sell the underlying property back to the lender in exchange for eliminating the mortgage obligation.

Participants in the mortgage market have a need for a model relating LTV ratios, and perhaps other factors, to default rate. A natural choice for such a model would be a binary regression equation

$$\text{Prob}\,(\text{Default}) = \Phi(\boldsymbol{\beta}'\mathbf{x}), \tag{1}$$

where Φ is some cumulative probability function, $\mathbf{x}$ is a vector of covariates, including the LTV variable, and $\boldsymbol{\beta}$ is a vector of unknown parameters to be estimated. Choosing Φ as the normal or the logistic distribution leads to the probit and logistic regression models, respectively.

Software for estimating the parameters of the standard probit or logistic regression is, of course, widely available. However, for a few reasons it is the

case that typically available mortgage default data cannot be accurately fit into the standard formula (1). A major reason is that the crucial LTV variable is unobservable, since it depends on the market value of the home, which can be observed only when the home is actually sold. Several authors (e.g., Quigley and Van Order 1995) have used a regional price index as a proxy for the market value of homes within a given geographic region. However, there is known to be a large amount of variation within geographic regions with respect to changes in housing price (Case and Shiller 1989). For example, market value within a particular neighborhood may decline significantly over a year, even if the average price within the broad region on which the index is based has remained constant or risen. Thus, the regional price index measures the price of individual houses imperfectly, or with measurement error. If ignored, this measurement error in the independent variable will lead to biased estimates of regression coefficients (Fuller 1989; Carroll, Ruppert, and Stefanski 1995).To properly account for measurement error, it is necessary to quantify the amount of such error. The Weighted Repeat Sales (WRS) technique of Abraham and Schauman (1991) provides a means of assessing the amount of measurement error associated with a given aggregate housing price index. Thus, in principle it is possible to correct for the imperfect measurement of LTV ratios.

Another reason that the standard binary regression equation (1) may be inappropriate for modelling defaults is that default events appear to have complex time series properties, with defaults clustering temporally. This phenomenon can only partially be explained by introducing observable macro–economic covariates such as interest rate into the regression equation. What appears to be needed, in addition, is an autocorrelated error term; however, incorporating such a factor is beyond the capabilities of the usual logistic or probit regression program.

Finally, because defaults are relatively rare events, it is often necessary to

use retrospective, or case-control, sampling to obtain a database for mortgage modelling. Such a sampling scheme may include all of the defaulting loans in a given database, plus a small random sample of the non-defaulting loans. The alternatives to this approach are either to work with a very large database, which leads to an unacceptable computational burden, or to use a small database, which may have insufficient defaults represented to adequately estimate model parameters. This introduces a modeling constraint, in that any estimation procedure proposed for the default data must be able to accommodate retrospectively sampled data. McCullagh and Nelder (1989, page 111) state that "one important property of the logistic function not shared by the other link functions is that differences on the logistic scale can be estimated regardless of whether the data are sampled prospectively or retrospectively." Thus, subject to appropriate checks on goodness-of-fit, it will be often be desirable to perform specifically a *logistic* regression analysis of the binary default data.

In Section 2, we introduce a model for mortgage defaults which takes into account these various considerations. In Section 3, we describe a Markov chain Monte Carlo technique (Roberts and Smith 1993) for estimating the model parameters. The technique developed is an elaboration of the modelling strategy used in Albert and Chib (1993). In Section 4, we apply the modelling procedure to an existing database of loans, and show that ignoring the measurement error in the regressors and the autocorrelation in the residuals indeed leads to substantial bias in parameter estimates. In Section 5, we describe directions for future work in the area of mortgage model specification and estimation.

2 The Model

The goal of the analysis in this paper is to assess the relationship between loan–to–value ratio and other economic factors, and the probability of default on home loans. The mortgage data which are used for this analysis consist of the date of origination, the LTV ratio at origination, the date of loss to observation, the reason for the loss to observation, and the geographic region for each of M mortgages. The loss to observation for a particular mortgage may be due to default, or to prepayment, or to censoring of a mortgage that is still outstanding at the end of the observation period. Table 1 displays a subset of the observations in the mortgage database.

INSERT TABLE 1 HERE

To analyze the data, each loan under study was expanded into a set of independent loan–year observations, as in Table 2; for each loan–year, a binary response is recorded according to whether or not the given loan defaulted in the given year. Note from Table 2 that the LTV variable is known exactly in the year of origination, but not in subsequent years.

INSERT TABLE 2 HERE

If nothing were known about housing prices, and hence LTV values, after origination, then it would be difficult to infer the relationship between LTV and default. In fact, there are available data on housing index values within broad geographic regions. An example of such data are displayed in Table 3; these indices correspond to the geographic regions displayed in Figure 1. The trajectory of the indices provides information about the path of the LTV ratios for each loan. However, there is known to be substantial variation of housing

values within each region (Case and Shiller 1989). Thus, the housing indices can be a useful proxy for LTV, but there is considerable measurement error, and this measurement error is known to lead to bias in regression coefficient estimation, if ignored (Carroll, Ruppert, and Stefanski 1995).

The notation, and the model for defaults are now described: let $h_{ij}(t)$ denote the true market value of the home corresponding to loan i in region j, at time period t, and let $m_{ij}(t)$ denote the amount of money owed; $\ell_{ij}(t) = \log(m_{ij}(t)/h_{ij}(t))$ is the log of the LTV ratio. Let $c_j(t)$ denote the value of the regional housing price index for region j at time t. Let $x_{ij}(t)$ denote additional covariates pertaining to loan i, region j, time t, such as the age of loan (i,j) at time t (that is, the number of years since origination for the loan), or the prevailing lending rate at time t. It is assumed that covariates $x_{ij}(t)$ are measured without error. Let $\delta_j(t)$ denote an unobservable variable associated with region j, time t; it can be thought of as representing a regional economic factor that has not been incorporated into the set of covariates $x_{ij}(t)$. Finally, let $z_{ij}(t)$ denote a latent quantitative variable associated with loan (i,j) at time t; we will assume that $z_{ij}(t)$ represents "propensity to default", such that the loan will default if and only if $z_{ij}(t) \geq 0$. The model for defaults used in this paper presumes that $z_{ij}(t)$ is related to the covariates $\ell_{ij}(t)$ and $x_{ij}(t)$ according to the linear model

$$z_{ij}(t) = \beta_0 + \beta_1 \ell_{ij}(t) + \beta_2' x_{ij}(t) + \delta_j(t) + \epsilon_{ij}(t), \quad \epsilon_{ij}(t) \sim \mathrm{N}(0, \xi_{ij}(t)), \tag{2}$$

$$\delta_j(t) = \rho \delta_j(t-1) + u_j(t), \quad u_j(t) \sim \mathrm{N}(0, a), \tag{3}$$

$$i = 1, \ldots, I_j, \quad j = 1, \ldots, J, \quad t = t_{ij}^{\mathrm{b}}, \ldots, t_{ij}^{\mathrm{e}}$$

where $\mathrm{N}(\mu, \xi)$ denotes the normal distribution with mean μ and variance ξ, I_j denotes the number of loans from region j, J denotes the number of geographic regions, and $(t_{ij}^{\mathrm{b}}, t_{ij}^{\mathrm{e}})$ denote the calendar dates of origination and of loss to follow–

up for loan i in region j. The use of time values t_{ij}^{b} and t_{ij}^{e} in the notation for model (2)–(3) is made necessary by the fact that there are two relevant time scales in considering the experience of a loan in any given year: the time elapsed since origination of the loan, as well as the exact calendar year.

Equations (2)–(3) differ from the standard probit model in three respects. First, the independent variable $\ell_{ij}(t)$ is unobservable, since the market value of the house, $h_{ij}(t)$, can be measured only imperfectly via the regional housing price index $c_j(t)$. The weighted repeat sales technique used in Case and Shiller (1989) implies a particular relationship between the time series $c_j(t)$ and the unobservable series $h_{ij}(t)$; namely,

$$\log h_{ij}(t) = \log h_{ij}(t-1) - \log c_j(t) + \log c_j(t-1) + v_{ij}(t),$$
$$v_{ij}(t) \sim \mathrm{N}(0, \lambda_{ij}(t)). \tag{4}$$

Let $w_{ij}(t) = \log(m_{ij}(t)/m_{ij}(t-1)) - \log(c_j(t)/c_j(t-1))$; then

$$\ell_{ij}(t) = \ell_{ij}(t-1) + w_{ij}(t) + v_{ij}(t). \tag{5}$$

$w_{ij}(t)$ is thus the observed component of the change in $\ell_{ij}(t)$. The original loan to value ratio, $\ell_{ij}(t_{ij}^{\mathrm{b}})$ is known for each loan (i,j), since the housing value is observable at the time of loan origination t_{ij}^{b}. Correction for measurement error in regressors requires knowledge of the amount of error (Fuller 1989); the WRS technique (Case and Shiller 1989) provides estimates of the measurement error variances $\lambda_{ij}(t)$.[1]

[1]The results of Case and Shiller (1989) suggest that λ, the variance of housing price within regions, varies both across regions and over time, and in particular depends on the time since origination for the given loan: i.e., $\lambda_{ij}(t) = \lambda_j(t - t_{ij}^{\mathrm{b}})$.

The second respect in which model (2)–(3) differs from the standard probit model is in the inclusion of the auto-correlated variance component $\delta_j(t)$. This component is included to account for clustering of defaults within particular calendar years, beyond the extent which can be explained by the temporal variation in $\ell_{ij}(t)$ and $x_{ij}(t)$. The parameter a in equation (3) measures the degree to which such clustering occurs, and the parameter ρ measures the degree to which a year with an extraordinarily high number of defaults tends to be followed by another such high-default year.

The third respect in which equations (2)–(3) differ from the usual probit regression formulation is that the standard probit model assumes that the variances $\xi_{ij}(t)$ of the errors $\epsilon_{ij}(t)$ in (2) are all equal to 1. Here, we will retain more generality; this generality allows for the use of probability functions $\Phi(\cdot)$ in equation (1) other than the cumulative normal. Albert and Chib (1993) show that if the $\xi_{ij}(t)$ are randomly distributed according to an inverse gamma distribution, then the unconditional distribution of the $\epsilon_{ij}(t)$ is the Student-t, which may fit certain datasets better than does a normal model. As discussed in Section 1, a retrospective data collection approach, in which all defaulting loans but only a subset of non default loans are included in the analysis, may be required in order to obtain an adequate number of defaults with a manageable total sample size; this consideration leads to a preference for using the logistic link function. If the variables $1/2\xi_{ij}(t)$ are modelled as following the asymptotic Kolmogorov distribution, then the $\epsilon_{ij}(t)$, conditional on the $\xi_{ij}(t)$, are normally distributed, but the unconditional distribution of the $\epsilon_{ij}(t)$ is the logistic (Andrews and Mallows 1974). We show below that it is thus possible to combine the computational convenience of a (heteroscedastic) normal regression model with the interpretive convenience of the logistic model; see Mallick and Gelfand (1994) for a further discussion of alternative link functions in binary

regression.

Since the analysis proposed in this paper is Bayesian, prior distributions on the model parameters are required. We use non-informative, locally uniform, priors on all the parameters, except for the variances $\xi_{ij}(t)$, which for identification reasons must have a proper prior distribution. Given an analysis based on non-informative priors, the outputs of the analysis can be adjusted, using the methods of Smith and Gelfand (1992), to incorporate informative priors, should such be available.

3 Model Estimation

In this section, we demonstrate that the likelihood function for the coefficients β is given by a product of high-dimensional integrals, and that while this likelihood function cannot be easily maximized, draws from the posterior distribution of β can be obtained fairly straightforwardly using a Markov chain Monte Carlo approach.

To simplify the derivation of the log-likelihood, assume for the present time that the autocorrelated variance component $\delta_j(t)$ is not present, or alternatively that $\delta_j(t) = 0$ for all j and all t. Also, assume that the variances $\xi_{ij}(t)$ in (2) are all known and equal to 1; i.e., that the model is a simple probit regression. Then the model for defaults becomes

$$z_{ij}(t) = \beta_0 + \beta_1 \ell_{ij}(t) + \beta_2' x_{ij}(t) + \epsilon_{ij}(t), \quad \epsilon_{ij}(t) \sim \mathrm{N}(0,1), \tag{6}$$

$$\ell_{ij}(t) = \ell_{ij}(t-1) + w_{ij}(t) + v_{ij}(t). \tag{7}$$

Equations (6)–(7) together imply that

$$z_{ij}(t_{ij}^{\mathrm{b}}) = \beta_0 + \beta_1 \ell_{ij}(t_{ij}^{\mathrm{b}}) + \beta_2' x_{ij}(t_{ij}^{\mathrm{b}}) + \epsilon_{ij}(t_{ij}^{\mathrm{b}})$$

$$z_{ij}(t) = \beta_0 + \beta_1(\ell_{ij}(t_{ij}^{\mathrm{b}}) + \sum_{\tau=1}^{t} w_{ij}(t_{ij}^{\mathrm{b}} + \tau)) + \beta_2' x_{ij}(t) + \epsilon_{ij}(t) + \sum_{\tau=1}^{t} \beta_1^{\tau} v_{ij}(t_{ij}^{\mathrm{b}} + \tau), \quad t = t_{ij}^{\mathrm{b}} + 1, \ldots, t_{ij}^{\mathrm{e}}.$$

Thus, the set of latent variables $z_{ij}(t)$, $t = t_{ij}^{\mathrm{b}}, \ldots, t_{ij}^{\mathrm{e}}$ come from a multivariate normal distribution, and are non-independent, due to common dependence on the measurement errors $v_{ij}(t)$, $t = t_{ij}^{\mathrm{b}} + 1, \ldots, t_{ij}^{\mathrm{e}}$. Let μ_{ij} and Σ_{ij} denote the mean vector and covariance matrix of $z_{ij}(t)$, $t = t_{ij}^{\mathrm{b}}, \ldots, t_{ij}^{\mathrm{e}}$; these moments both depend on β. Then the likelihood function for the parameters β is given by

$$\mathcal{L} = \prod_{j=1}^{J} \prod_{i=1}^{I_j} \int_{A_{ij}} \Phi(\mathbf{z}; \mu_{ij}, \Sigma_{ij}) d\mathbf{z}, \tag{8}$$

where $\Phi(\mathbf{z}; \mu_{ij}, \Sigma_{ij})$ is the multivariate normal density function, and the region A_{ij} depends upon whether or not loan (i,j) is a default; if loan (i,j) is not a default, then $A_{ij} = \{\mathbf{z} | z_1 \leq 0, z_2 \leq 0, \ldots, z_{n_{ij}} \leq 0\}$, where $n_{ij} = t_{ij}^{\mathrm{e}} - t_{ij}^{\mathrm{b}}$, and if loan (i,j) has defaulted, then $A_{ij} = \{\mathbf{z} | z_1 \leq 0, z_2 \leq 0, \ldots, z_{n_{ij}} \geq 0\}$.

Equation (8) demonstrates that the likelihood function is a product of multidimensional definite integrals, one integral for each loan in the dataset. In the dataset used for this paper, there are over 4400 loans, with some observed for as long as 7 years, so that some of the integrals will be over a 7 dimensional region. Evaluation of the likelihood function, then, is very expensive (see, though, Geweke 1991), and numerical maximization of the likelihood function even more

of a challenge. Incorporation of the variance component $\delta_j(t)$ into the equation for loan defaults, and consideration of distributions other than the normal, will even further complicate the likelihood function. Thus, maximum likelihood estimation of the model parameters appears to be prohibitively time consuming in the setting of complex measurement error. However, exact finite sample inferences on the parameters of the binary regression model specified by equations (2), (3), and (4) can be estimated via data augmentation (Tanner and Wong 1987) and Gibbs sampling (Gelfand and Smith 1990).

Gibbs sampling is a particular variant of the class of procedures known as Markov chain Monte Carlo methods (Roberts and Smith 1993), in which parameter vectors are randomly generated from a Markov chain constructed so that the chain's stationary distribution is equal to the joint posterior distribution of the model parameters. Summary statistics of the randomly generated parameters – means, variances, histograms – serve as posterior estimates of the model parameters.

Let Θ be the set of all the model parameters, and let $\mathcal{D}$ denote the data. The joint posterior distribution is $[\Theta \mid \mathcal{D}]$, with associated full conditional distributions $[\theta_1 \mid \theta_{(-1)}, \mathcal{D}]$, $[\theta_2 \mid \theta_{(-2)}, \mathcal{D}]$, $\ldots$, $[\theta_k \mid \theta_{(-k)}, \mathcal{D}]$, where $\theta_{(-l)}$ denotes the set of all parameters θ except θ_l. The Gibbs sampling procedure is initialized with some arbitrary vector $(\theta_1, \ldots, \theta_k)$. A new realization of θ_1 is then randomly generated from the conditional distribution $[\theta_1 \mid \theta_{(-1)}, \mathcal{D}]$, a realization of θ_2 from the conditional distribution $[\theta_2 \mid \theta_{(-2)}, \mathcal{D}]$, and so on, with the cycle repeated several – typically hundreds or thousands – of times. The stationary distribution of the vectors $(\theta_1, \ldots, \theta_k)$ generated in each cycle of this procedure has been shown to be equal to the joint distribution $[\Theta \mid \mathcal{D}]$ (Geman and Geman 1984). Because the process is ergodic (Gelfand and Smith 1990), the mean and variance of any variate θ_k can be estimated by the sample mean and variance

of θ_k over the realized stream of numbers generated by the Markov chain. In practice, one usually omits the first several hundred samples generated when computing sample means, variances, and histograms, to be sure that convergence to the desired density has occurred. Tanner (1993) and Tierney (1994) provide further information on the implementation of Markov chain methods, and on the convergence properties of these methods, Roberts (1992) and Robert (1995) offer discussions of convergence diagnostics, and Andrews, Berger, and Smith (1993) and Geweke and Keane (1996) provide detailed examples of econometric applications of the use of Markov chain Monte Carlo methods.

The data augmentation/Gibbs sampling method for estimating model (2)–(4) works by alternating between imputing values for the unobserved values $z_{ij}(t)$, $\ell_{ij}(t)$, $\delta_j(t)$, and $\xi_{ij}(t)$, and estimating model parameters β, ρ, and a. Given the imputed data, the parameter estimates are obtained easily, and given the parameter estimates, the data augmentations are also fairly direct. Appendix A describes the Gibbs sampling estimation algorithm in greater detail, and includes formulas for the conditional posterior distributions used to generate model parameters. Most of these distributions are obtained from established results in Bayesian econometrics (Zellner 1971). The exception is for the conditional posterior distribution of $\xi_{ij}(t)$, the latent scale parameters for the $z_{ij}(t)$; Section A.5 describes a Metropolis-Hastings method for generating the $\xi_{ij}(t)$ in the case when the $z_{ij}(t)$ have an unconditional logistic distribution, as is desired for the application in this paper.

4 Data Analysis

The mortgage data used for the analysis come from high initial LTV loans (initial LTV equal to 95%), which loans were sold to the Federal Home Loan

Mortgage Corporation (FHLMC). The loans, all of which were originated in 1983, are on single family homes, and all necessarily conform to FHLMC underwriting restrictions. Table 4 provides a summary of the default behavior of the loans; the geographic and temporal clustering of defaults is evident. 100% of the defaulting loans were included in the analysis, and 10% of the non-default loans were included; since most loans were non-defaults, this sub-sampling had only a minor effect on posterior standard deviations for model parameters, while significantly reducing computing time. The interpretation of the coefficient estimates is not affected by this non-random sampling scheme, due to the fact that the logistic link function was used for the binary regression model (McCullagh and Nelder 1989).

INSERT TABLE 4 HERE

The housing indices used for the study were the FHLMC Weighted Repeat Sales indices, which are based on the Case and Shiller (1989) index construction methodology; see Table 3. Figure 1 displays the five U.S. geographic regions corresponding to the indices. The weighted repeat sales methodology permits estimation of the amount of measurement error associated with the indices. Table 5 lists the R^2's for the indices; these correlation measures are defined as

$$R^2 = \text{Var}\left[\log[c_j(t)/c_j(0)]\right] / \text{Var}\left[\log[h_{ij}(t)/h_{ij}(0)]\right].$$

The low correlations indicate the considerable extent to which housing indices may mismeasure actual changes in individual housing price. These R^2 measures, though, may even be over-estimates; Case and Shiller (1989, page 127) study housing values in Atlanta, Chicago, Dallas, and San Francisco/Oakland, and find, using slightly different methodology, R^2 of .07 for Atlanta, .16 for Chicago, .12 for Dallas, and .27 for San Francisco/Oakland. Since the Case and Shiller

(1989) study uses data at a lower level of aggregation than does the present study (city versus multi–state region), this suggests that the measurement error in the FHLMC may be understated; this effect may be due to appraisal smoothing (Ross and Zisler 1991). To account for this possibility, a sensitivity analysis is performed, in which the analysis is conducted under several different assumptions concerning the amount of measurement error.

For forecasting the default status for each loan i in region j, time t, the covariates used, in addition to the LTV ratio, were the age of the loan at time t, and the interest rate (1 month treasury bill rate) prevailing at calendar time t. The age of a loan will be related to the true (unobservable) LTV, since older loans a) have payed a larger amount of their mortgage, and thus tend to have a lower amount of principal due, and b) have had more time to have their housing value depreciate (or appreciate). Analyses which ignore measurement error in LTV may erroneously attribute default behavior to aging ("seasoning") of loans, when really it may be increasing LTV's that are causing defaults. Thus, the present analysis, which accounts for measurement error in LTV, will afford a view of the actual relationship of age and default, conditional upon (true) LTV.

Table 6 lists the posterior means and standard deviations for the regression coefficients. The column labeled "Measurement Error" denotes the assumption made about the extent of of measurement error in the housing indices. The first analysis, with "0%" error, assumes that the housing indices are perfect measures of the actual housing value for each individual loan. The second analysis, with "100%" error, assumes that the R^2 values cited in Table 5 correctly quantify the amount of variation in price within the geographic regions. The analyses with "200%" and "300%" measurement error, are based on the assumption that the variance within regions is 2 times and 3 times larger, respectively, than that obtained by the WRS method; the value corresponding to 300% is consistent with

the results Case and Shiller (1989) obtained for their more disaggregate housing index data, and may be the most believable estimate. The parameter estimates were obtained by running the Markov chain Monte Carlo procedure for 50,000 iterations, and discarding the samples from the first 25,000 iterations.

INSERT TABLE 6 HERE

As anticipated, the analysis which does not account for measurement error has an estimate for the coefficient for LTV which is substantially smaller than those obtained when the measurement error is acknowledged. This implies that default rate is more sensitive to changes in housing prices than the naive analysis suggests. It is also interesting to note that, even after accounting for the measurement error in LTV ratios, the age of the loans has a statistically significant effect on default probability, and indeed the coefficient for age increases in magnitude. It appears, then, that the seasoning effect cannot be explained by the imperfect measurement of LTV; the age of the loan may be acting as a proxy for some economically significant characteristic of the borrower. The coefficient estimate for interest rate, on the other hand, decreases in magnitude when the measurement error is taken into account. As expected, the posterior standard errors for the parameter estimates are larger for the analyses which acknowledge the measurement error in the housing index numbers.

This example suggests the complex way that measurement error for just a single instrument can affect regression coefficients in a setting with multiple predictor variables. In practice, the estimated model for default rates may be used as part of a scenario analysis program for managing risk with mortgage portfolios. The effect of the analysis in this paper is to demonstrate that loan default behavior is more sensitive to changing housing prices than had previously been shown, and also to make default behavior less sensitive to interest rates

than would appear from an analysis which ignores the imperfect measurement of housing price in the regression data.

5 Conclusion

The U.S. mortgage market is a large component of the national economy, and understanding this market is important both to participants in the primary and secondary mortgage markets, and to academics who wish to understand the macroeconomy. While stock markets and treasury bond markets have been very intensively studied by financial economists, studies of the housing market have been relatively less common; this is largely because real estate is very thinly traded, and thus price data are much less easily available than are data for stock and treasury bond markets. The recent development of real estate price indices (Case and Shiller 1989; Abraham and Schauman 1991; Clapp and Giancotto 1991; Quigley 1995) promises to help lead to greater understanding of housing markets; however, as this article has shown, using regional price indices without accounting for the variation of prices within regions can, in some circumstances, lead to significantly biased inferences of housing market model parameters. The data augmentation method described in this paper provides a way of correcting for this inequivalence between housing index values and actual housing prices. The method will be useful in the estimation of default models, prepayment models, and any other model in which housing value may be an important covariate. Measurement error in binary regression models has been previously examined, for example in Stefanski and Carroll (1985) and Carroll, Ruppert, and Stefanski (1995). The measurement error process examined in the present paper differs substantially from these earlier treatments, in that the measurement error process considered in equation (4) itself has time series properties, and thus a

single measurement error will effect multiple observations in the binary regression equation (2).

In addition to treating the measurement error problem, the current paper also demonstrates the treatment of autocorrelation, and shows how a logistic regression analysis can be performed by modeling the logistic distribution as a scale mixture of normals, augmenting the data with the continuous probits, and then using standard normal regression theory.

Future research will involve joint modeling of defaults and prepayments, and will also involve the development of a fully Bayesian version of the weighted repeat sales method for construction of housing price indices.

A The Markov Chain Monte Carlo Estimation Algorithm

The Gibbs sampling algorithm for estimating the parameters in model (2)–(4) consists of the following sequence of steps, which are performed repeatedly:

1. generate β;
2. impute the missing LTV ratios $\ell_{ij}(t)$;
3. impute the autocorrelated random effects $\delta_j(t)$;
4. generate scale parameters $\xi_{ij}(t)$;
5. generate autocorrelation parameters ρ and a.

In each step, the generation/imputation is done using the respective conditional posterior distribution for the particular parameter; these distributions are derived below. The Gibbs sampling procedure can be initialized as described in section A.7.

A.1 Generating β

Conditional on the latent variates $z_{ij}(t)$, the LTV ratios $\ell_{ij}(t)$, the autocorrelated components $\delta_j(t)$, and the variances $\xi_{ij}(t)$, the distribution for $\beta = (\beta_0, \beta_1, \beta_2)$ can be obtained using the standard results for Bayesian linear models (e.g., Zellner (1971)). Let $\tilde{z}_{ij}(t) = z_{ij}(t) - \delta_j(t)$, and let $\tilde{\mathbf{z}}$ denote the quantities $\tilde{z}_{ij}(t)$ strung into a vector of length $n = \sum_{j=1}^{J} \sum_{i=1}^{I_j} (t_{ij}^{\mathrm{e}} - t_{ij}^{\mathrm{b}})$. Let $\boldsymbol{\Xi}$ be a $n \times n$ diagonal matrix with elements $\xi_{ij}(t)$ on the diagonal, and let $\mathbf{X}$ be the $n \times (2 + p)$ matrix whose first column's elements are 1's, second column's elements are the $\ell_{ij}(t)$ elements arrayed as a vector, and last p columns are

the $x_{ij}(t)$ similarly arranged. Then the full conditional distribution for β is $\mathrm{N}\left(\left(\mathbf{X}'\Xi^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\Xi^{-1}\tilde{\mathbf{z}}, \left(\mathbf{X}'\Xi^{-1}\mathbf{X}\right)^{-1}\right)$.

A.2 Generating $z_{ij}(t)$

As in Albert and Chib (1993), the $z_{ij}(t)$ are generated from truncated normal distributions. If loan i of region j defaulted at time t, then the conditional distribution for $z_{ij}(t)$ is $\mathrm{N}\left(\beta_0 + \beta_1\ell_{ij}(t) + \beta_2'x_{ij}(t) + \delta_j(t), \xi_{ij}(t)\right) I\left(z_{ij}(t) \geq 0\right)$. If this observation does not correspond to a default, then the full conditional distribution is $\mathrm{N}\left(\beta_0 + \beta_1\ell_{ij}(t) + \beta_2'x_{ij}(t) + \delta_j(t), \xi_{ij}(t)\right) I\left(z_{ij}(t) < 0\right)$. Clearly, $z_{ij}(t)$ cannot be positive unless $t = t_{ij}^{\mathrm{e}}$, and is positive then only if the loan is lost to observation due to default, and not to prepayment or censoring. Gelfand, Smith, and Lee (1992) provides information on generating from a truncated normal distribution.

A.3 Generating $\ell_{ij}(t)$

The conditional posterior for the LTV ratio $\ell_{ij}(t)$ is proportional to the prior times the likelihood. The prior is defined by (5), and the likelihood by (2). The log–posterior is thus equal, up to a constant, to

$$
\begin{aligned}
&-\tfrac{1}{2\xi_{ij}(t)}\left(z_{ij}(t) - \beta_0 - \beta_1\ell_{ij}(t) - \beta_2'x_{ij}(t) - \delta_j(t)\right)^2 \\
&-\tfrac{1}{2\lambda_{ij}(t)}\left(\ell_{ij}(t) - \ell_{ij}(t-1) - w_{ij}(t)\right)^2 \\
&-\tfrac{1}{2\lambda_{ij}(t+1)}\left(\ell_{ij}(t+1) - \ell_{ij}(t) - w_{ij}(t+1)\right)^2
\end{aligned}
$$

Rearranging shows that the full conditional for the unobserved $\ell_{ij}(t)$ is $\mathrm{N}\left(\left(\omega_1+\omega_2+\omega_3\right)^{-1}\left(\omega_1\alpha_1+\omega_2\alpha_2+\omega_3\alpha_3\right), \left(\omega_1+\omega_2+\omega_3\right)^{-1}\right)$, where $\omega_1 = \left(\xi_{ij}(t)\beta_1\right)^{-1}$, $\omega_2 = \left(\lambda_{ij}(t)\right)^{-1}$, $\omega_3 = \left(\lambda_{ij}(t+1)\right)^{-1}$, $\alpha_1 = \left(z_{ij}(t) - \beta_0 - \beta_2'x_{ij}(t) - \delta_j(t)\right)/\beta_1$, $\alpha_2 = \ell_{ij}(t-1) + w_{ij}(t)$, and

$\alpha_3 = \ell_{ij}(t+1) - w_{ij}(t+1)$. A simple modification of this expression is required for the endpoint $t = t^{\mathrm{e}}_{ij}$, in which case the full conditional is $\mathrm{N}\left((\omega_1+\omega_2)^{-1}(\omega_1\alpha_1+\omega_2\alpha_2),(\omega_1+\omega_2)^{-1}\right)$.

A.4 Generating $\delta_j(t)$

Let $S(j,t)$ denote the set of all loans in region j that are at risk during calendar year t; i.e., $S(j,t) = \{i,j|t^{\mathrm{b}}_{ij} \leq t \leq t^{\mathrm{e}}_{ij}\}$. The posterior of $\delta_j(t)$ depends on the experience of these loans during year t; if there are many defaults, then $\delta_j(t)$ would appear, a posteriori, to be high. The likelihood and prior for $\delta_j(t)$ are obtained from equations (2) and (3), respectively, the full conditional is thus $\mathrm{N}(D_j(t)^{-1}d_j(t), D_j(t)^{-1})$, with

$$\begin{aligned} d_j(t) &= \sum_{i\in S(j,t)} \xi_{ij}(t)^{-1}\left(z_{ij}(t) - \beta_0 - \beta_1\ell_{ij}(t) - \boldsymbol{\beta}_2'x_{ij}(t)\right) + \\ &\quad a^{-1}\rho\delta_j(t-1) + a^{-1}\rho\delta_j(t+1), \\ D_j(t) &= \sum_{i\in S(j,t)} \xi_{ij}(t)^{-1} + 2a^{-1}. \end{aligned}$$

Simple modifications for $d_j(t)$ and $D_j(t)$ are appropriate for the first and last calendar years in the database. In the former case,

$$\begin{aligned} d_j(t) &= \sum_{i\in S(j,t)} \xi_{ij}(t)^{-1}\left(z_{ij}(t) - \beta_0 - \beta_1\ell_{ij}(t) + \boldsymbol{\beta}_2'x_{ij}(t)\right) + a^{-1}\rho\delta_j(t+1), \\ D_j(t) &= \sum_{i\in S(j,t)} \xi_{ij}(t)^{-1} + a^{-1}; \end{aligned}$$

in the latter case,

$$d_j(t) = \sum_{i \in S(j,t)} \xi_{ij}(t)^{-1} \left(z_{ij}(t) - \beta_0 - \beta_1 \ell_{ij}(t) + \beta_2' x_{ij}(t)\right) + a^{-1}\rho\delta_j(t-1),$$

$$D_j(t) = \sum_{i \in S(j,t)} \xi_{ij}(t)^{-1} + a^{-1}.$$

A.5 Generating $\xi_{ij}(t)$

In the standard probit regression setting, the $\xi_{ij}(t)$, the variances of the residuals $\epsilon_{ij}(t)$, are all fixed and equal to 1, and hence do not need to be generated during the Markov chain. Let $\gamma_{ij}(t) = [\xi_{ij}(t)]^{-1/2}$; Andrews and Mallows (1974) show that if the density of $\gamma_{ij}(t)$ is given by

$$f_\Gamma(\gamma) = 2\sum_{k=1}^{\infty}(-1)^{k-1}k^2\gamma^{-3}\exp\left(-k^2/2\gamma^2\right), \tag{9}$$

then the unconditional distribution of the deviates $\epsilon_{ij}(t)$ is the standard logistic. A logistic regression analysis can thus be performed by modeling the residuals as a particular scale mixture for normals. The prior for $\xi_{ij}(t)$ is given by

$$f_\Gamma((\xi_{ij}(t))^{-1/2}) \cdot \frac{1}{2}(\xi_{ij}(t))^{-3/2}, \tag{10}$$

where $\frac{1}{2}(\xi_{ij}(t))^{-3/2}$ is the Jacobian of the transformation from $\gamma_{ij}(t)$ to $\xi_{ij}(t)$.

The likelihood function for $\xi_{ij}(t)$ is proportional to

$$(\xi_{ij}(t))^{-\frac{1}{2}} \exp\left(-\frac{1}{2\xi_{ij}(t)}\left(z_{ij}(t) - \beta_0 - \beta_1\ell_{ij}(t) - \beta_2' x_{ij}(t) - \delta_j(t)\right)^2\right), \tag{11}$$

and the full conditional posterior for $\xi_{ij}(t)$ is just the product of the prior and the likelihood function. This full conditional distribution is not of a standard form. To sample from the distribution, one can use a Metropolis-Hastings

technique (Tierney 1994; Hastings 1970), which involves generating from some approximation to the full conditional, and then accepting this sample with some specified probability, in order to correct for the approximation.

To obtain a suitable generating density, the prior density in (10) can be approximated by an inverse gamma density, by matching the first two moments; combining this inverse gamma approximate prior with the likelihood function (11) gives an inverse gamma approximate posterior, which empirically is seen to lead to very high acceptance probabilities. Use of the Metropolis–Hastings method requires frequent evaluation of the function in equation (9); while the sum is convergent, the evaluation is expensive. Computational time can be significantly reduced by building an accurate spline approximation to the log of the density in equation (9) at the start of the program, and then using the spline function, rather than the asymptotic summation, at each Metropolis–Hastings step.

A.6 Generating ρ, a

Given the values of $\delta_j(t)$, the parameters ρ and a are obtained using results from Bayesian analysis of AR models (Chib 1993). Let t_f and t_e denote the first and last calendar years during which loans in the database are at risk; $t_f = \min_{i,j}\{t_{ij}^{\mathrm{b}}\}$ and $t_e = \max_{i,j}\{t_{ij}^{\mathrm{e}}\}$. Then the full conditional for ρ is

$$\mathrm{N}\left(\sum_{j=1}^{J}\sum_{t=t_f+1}^{t_e}\delta_j(t-1)\delta_j(t)/\sum_{j=1}^{J}\sum_{t=t_f}^{t_e}\delta_j(t-1)^2, a(\sum_{j=1}^{J}\sum_{t=t_f+1}^{t_e}\delta_j(t-1)^2)^{-1}\right),$$

and the full conditional for a, is

$$\mathrm{IG}\left(\frac{1}{2}J(t_e-t_f-1), \frac{1}{2}\sum_{j=1}^{J}\sum_{t=t_f}^{t_e}(\delta_j(t)-\rho\delta_j(t-1))^2\right),$$

where IG (a_0, a_1) denotes the inverse gamma distribution.

A.7 Initialization of the Gibbs Sampler

The $z_{ij}(t)$ can be initialized to equal 1 for default observations, and -1 for non-default observations; the $\xi_{ij}(t)$ can be initialized at the mean of the prior density in (10); the $\ell_{ij}(t)$ can be initialized to $w_{ij}(t)$, the unbiased prior means based on the regional indices; and the $\delta_j(t)$ can be initialized to small random values. The coefficients $\boldsymbol{\beta}$ and the parameters ρ and a can then be estimated from these initialized data.

Figure 1: Regional composition of U.S. mortgage indices

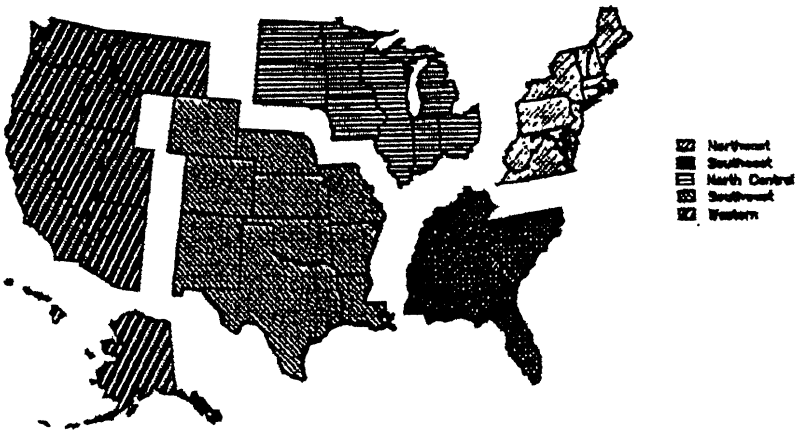

Table 1: Sample observations from mortgage database

Loan	Region	Initial LTV	Year Orig.	Year Term.	Reason Term.*	Coupon	Maturity
1	1	95	83	86	0	12	360
2	3	95	83	89	1	12	360
3	1	95	83	88	0	12	360
4	2	95	83	90	2	12	360
5	1	95	83	85	0	12	360
6	3	95	83	86	0	12	360
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
4416	1	95	83	86	0	12	360

* Reason for termination: 0 = censoring, 1 = default, 2 = prepayment.

Table 2: Sample observations from mortgage database after expansion

Loan	Region	Year	Default	Z	LTV	Age	Interest Rate
1	1	83	0	< 0	95	1	8.34
1	1	84	0	< 0	?	2	8.97
1	1	85	0	< 0	?	3	6.98
1	1	86	0	< 0	?	4	5.54
2	3	83	0	< 0	95	1	8.34
2	3	84	0	< 0	?	2	8.97
2	3	85	0	< 0	?	3	6.98
2	3	86	0	< 0	?	4	5.54
2	3	87	0	< 0	?	5	4.83
2	3	88	0	< 0	?	6	5.81
2	3	89	1	≥ 0	?	7	7.68
3	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 3: Freddie Mac regional housing indices, 1981–92

Year	Region				
	1	2	3	4	5
81	89.23	76.74	89.31	95.19	88.63
82	94.73	76.48	90.91	99.38	91.70
83	96.84	92.52	93.83	103.94	94.89
84	98.66	96.52	97.91	99.09	98.19
85	103.41	114.38	103.68	101.76	104.39
86	111.12	133.77	110.16	104.14	112.55
87	119.48	149.48	117.21	101.04	128.12
88	126.1	163.0	120.3	98.8	146.0
89	134.4	170.2	123.9	102.2	172.2
90	138.7	165.7	124.5	101.9	178.0
91	144.9	168.7	128.5	106.9	181.5
92	150.35	169.19	133.66	110.95	179.32

Table 4: Summary statistics for 4,416 FHLMC loans originated in 1983 with initial LTV of 95%, by region. NC=North Central, NE=Northeast, SE=Southeast, SW=Southwest, W=West.

	Region				
	NC	NE	SE	SW	W
	Number of loans originated in 1983				
	504	1200	632	1196	884
Year	Number of loans that defaulted				
1984	0	1	3	1	3
1985	0	5	4	4	11
1986	4	4	4	24	18
1987	4	6	3	45	10
1988	2	3	9	93	7
1989	4	0	4	32	3
1990	3	2	4	15	0

Table 5: Measurement error (R^2) in FHLMC WRS housing price indices. The R^2 is defined as the variance of the change in the index divided by the variance of the change in individual home prices, over various horizons

	R^2			
Region	1 year	5 year	10 year	20 year
NC	.289	.296	.320	.387
NE	.279	.292	.334	.477
SE	.344	.351	.333	.298
SW	.398	.378	.385	.406
W	.315	.370	.435	.653

Table 6: Parameter estimates (posterior means and standard deviations) for logistic regression coefficients, under different assumptions for measurement error

	Measurement error (as multiple of WRS estimate)							
	0%		100%		200%		300%	
Intercept	-1.577	(0.347)	-2.145	(0.385)	-2.819	(0.472)	-3.472	(0.623)
LTV	3.619	(0.356)	4.228	(0.529)	4.646	(0.582)	5.031	(0.771)
AGE	0.550	(0.037)	0.612	(0.047)	0.673	(0.058)	0.748	(0.078)
Int. Rate	-0.400	(0.054)	-0.351	(0.057)	-0.292	(0.063)	-0.253	(0.071)

References

Abraham, J. and W. Schauman (1991), 'New evidence on home prices from Freddie Mac repeat sales', *AREUEA Journal*, **19**, 333–352.

Albert, J. and S. Chib (1993), 'Bayesian analysis of binary and polychotomous response data', *Journal of the American Statistical Association*, **88**, 669–679.

Andrews, D. and C. Mallows (1974), 'Scale mixtures of normality', *Journal of the Royal Statistical Society, Series B*, **36**, 99–102.

Andrews, R. W., J. O. Berger, and M. H. Smith (1993), 'Bayesian estimation of manufacturing effects in a fuel economy model', *Journal of Applied Econometrics*, **8**, 5–18.

Carroll, R., D. Ruppert, and L. Stefanski (1995), *Measurement Error in Nonlinear Models*, Chapman & Hall, London, UK.

Case, K. and R. Shiller (1989), 'The efficiency of the market for single family homes', *American Economic Review*, **79**, 125–137.

Chib, S. (1993), 'Bayes inference in the AR model', *Journal of Econometrics*, **51**, 79–99.

Clapp, J. and D. Giancotto, C. andd Tirtiroglu (1991), 'Housing price indices based on all transactions compared to repeat subsamples', *AREUEA Journal*, **19**, 270–286.

Fuller, W. (1989), *Measurement Error Models*, John Wiley and Sons, New York, NY.

Gelfand, A. and A. Smith (1990), 'Sampling based approaches to calculating marginal densities', *Journal of the American Statistical Association*, **85**, 398–409.

Gelfand, A., A. Smith, and T.-M. Lee (1992), 'Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling', *Journal of the American Statistical Association*, **87**, 523–531.

Geman, S. and D. Geman (1984), 'Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Geweke, J. (1991), 'Efficient simulation from the multivariate normal and student–t distributions subject to linear constraints', In E. Keramidas (ed.) *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pp. 571–578.

Geweke, J. and M. Keane (1996). 'An empirical analysis of the male income dynamics in the PSID: 1968–1989'. Technical Report 94–10, University of Minnesota.

Hastings, W. (1970), 'Monte Carlo sampling methods using Markov chains and their applications', *Biometrika*, **57**, 97–109.

Mallick, B. and A. Gelfand (1994), 'Generalized linear models with unknown link functions', *Biometrika*, **81**, 237–245.

McCullagh, P. and J. Nelder (1989), *Generalized Linear Models* (2nd ed.)., Chapman & Hall, London, UK.

Quercia, R. and M. Stegman (1992), 'Residential mortgage default: A review of the literature', *Journal of Housing Research*, **3**, 341–379.

Quigley, J. (1995), 'A simple hybrid model for estimating real estate price indexes', *Journal of Housing Economics*, **4**, 1–12.

Quigley, J. and R. Van Order (1995), 'Explicit tests of contingent claims models of mortgage default', *Journal of Real Estate Finance and Economics*, **11**,

99–117.

Robert, C. (1995), 'Convergence control methods for Markov chain Monte Carlo algorithms', *Statistical Science*, **10**, 231–253.

Roberts, G. O. (1992), 'Convergence diagnostics of the Gibbs sampler', In J. Bernardo, J. Berger, A. Dawid, and A. Smith (eds.) *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting*, pp. 775–782. Oxford University Press.

Roberts, G. O. and A. F. M. Smith (1993), 'Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods', *Journal of the Royal Statistical Society, Series B*, **55**, 3–23.

Ross, S. and R. Zisler (1991), 'Risk and return in real estate', *Journal of Real Estate Finance and Economics*, **4**, 175–190.

Smith, A. and A. Gelfand (1992), 'Bayesian statistics without tears: A sampling–resampling perspective', *American Statistician*, **46**, 84–88.

Stefanski, L. A. and R. J. Carroll (1985), 'Covariate measurement error in logistic regression', *Annals of Statistics*, **13**, 1335–1351.

Tanner, M. (1993), *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions* (2nd ed.)., Springer Verlag, New York, NY.

Tanner, M. A. and W. Wong (1987), 'The calculation of posterior distributions by data augmentation', *Journal of the American Statistical Association*, **82**, 528–550.

Tierney, L. (1994), 'Markov chains for exploring posterior distributions (with discussion)', *Annals of Statistics*, **4**, 1701–1762.

Zellner, A. (1971), *An Introduction to Bayesian Inference in Econometrics*, John Wiley and Sons, New York, NY.