Unobserved Heterogeneity in Models of Competing Mortgage Termination Risks

John M. Clapp,* Yongheng Deng** and Xudong An***†

Abstract

This paper extends unobserved heterogeneity to the multinomial logit model (MNL) framework in the context of mortgages terminated by refinance, move, or default. It tests for the importance of unobserved heterogeneity when borrower characteristics such as income, age and credit score are included to capture lender-observed heterogeneity. It does this by comparing the proportional hazard model (PHM) to MNL with and without mass-point estimates of unobserved heterogeneous groups of borrowers.

The mass point mixed hazard model (MMH) yields larger and more significant coefficients for several important variables in the move model, whereas the MNL model without unobserved heterogeneity performs well with the refinance estimates. The MMH clearly dominates the alternative models in-sample and out-of-sample. However, it is sometimes difficult to obtain convergence for the models estimated jointly with mass points.

JEL classification: G21; C25; C41; C52; D12

* School of Business Administration, University of Connecticut, Storrs CT, 06269-1041 or john.clapp@business.uconn.edu.
** School of Policy, Planning and Development, University of Southern California, Los Angeles CA, 90089-0626 or ydeng@usc.edu.
*** School of Policy, Planning and Development, University of Southern California, Los Angeles CA, 90089-0626 or xudongan@usc.edu.
† The authors are grateful for comments on earlier drafts by Real Estate Economics Editor David Ling, two anonymous referees and participants in the seminars at the University of Connecticut, Department of Statistics and University of Southern California, Lusk Center for Real Estate. Yongheng Deng and Xudong An acknowledge the financial support from the Lusk Center for Real Estate at the University of Southern California.

Introduction

Recent researches on mortgage borrower's behavior have proposed several models for the competing risks of mortgage termination by refinancing, moving and/or default (Deng, Quigley and Van Order 2000, Clapp *et al*. 2001, Deng and Quigley 2001).[1] Clapp *et al*. (2001) present evidence that a multinomial logit model (MNL) with restructured event history data is an attractive alternative to duration models such as the proportional hazard model (PHM). The MNL allows direct competition among the choices: the probabilities of termination risks, and the probability of continuing to pay, must sum to one. Thus, an increase in one termination probability must be offset by a decline in probability for one or more of the alternatives.

On the other hand, the MNL cannot allow correlations among the termination risks through unobservable variables, as implied by the independence from irrelevant alternatives (IIA) assumption.[2] In addition, the MNL requires the i.i.d. assumption for a given agent observed over time[3] − following standard practice, we stack the observations of historical events for each agent into our likelihood function. This logic also requires complicated formulation of variables measuring duration dependency. By way of contrast, the hazard function in a proportional hazard model (PHM) is constructed in a path dependent framework: i.e., the hazard rate of termination is conditioned on the subject surviving up to time t-1. Therefore, any event between t and t-1 is not an i.i.d. event. The full maximum likelihood estimation approach also allows researchers to estimate a PHM with correlated competing risks.[4]

Although the multinomial logit model (MNL) and proportional hazard model (PHM) differ in the above-mentioned perspectives, they are both widely used in the literature of mortgage termination risks and demonstrated to be effective in the studies. Since Green and Shoven (1986) first introduced the proportional hazard model (PHM) to analyze mortgage termination by refinance, there have been several major developments to improve the application of PHM to

[1] Strictly speaking, the default decision requires 3 months of nonpayment to be observed in the data.
[2] The Independence from Irrelevant Alternatives (IIA) property implied by MNL restricts the odds ratio of choice probabilities, *i* and *k*, so that they do not depend on any alternatives other than *i* and *k*. This in turn implies no correlation between the unobserved components of utility for alternatives. (See Train (1986) for a detailed discussion on MNL and IIA properties.)
[3] The logit model is obtained by assuming that each ε_{nj} follows an independently, identically distributed extreme value function. (See Train (2003) for a complete discussion.)

mortgage termination analysis. Recent applications include more sophisticated and realistic modeling frameworks. For example, Schwartz and Torous (1989) developed a contingent claim framework for valuation of GNMA mortgage-backed securities through the integration of an empirical PHM to estimate the aggregate GNMA mortgage pools' prepayment experience. Stanton (1995) extends the Schwartz and Torous (1989) model by allowing transaction cost of prepayment in the modeling of mortgage pools' rational prepayment behavior. The application of logit models to mortgage termination issues is well established. Mattey and Wallace (2001), Ambrose and Capone (1998), Berkovec et al. (1998), Archer, et al. (1996), Quigley and Van Order (1995), Philips et al. (1995), and Cunningham and Capone (1990) have used binomial logit or MNL models. The PHM is established in the literature, but to a lesser extent (See, Ambrose and Sanders 2003, Pavlov 2001, Bennett et al. 2001, Ambrose and Capone 2000, Vandell, et al. 1993, Schwatz and Torous 1989, and Green and Shoven 1986).[5]

Deng, Quigley and Van Order (2000) address competing risks of mortgage termination in a proportional hazard framework that allows correlated competing risks and accounts for the unobserved heterogeneity as discrete mass points. Their approach models individual mortgage borrowers as coming from two or more distinct groups with unobserved characteristics. The model cannot directly observe which group each individual belongs to, but it can estimate the discrete probability distribution that each type influences the hazard function. The technique assumes a discrete number of groups; the researcher obtains maximum likelihood estimators of the mass-point distribution, i.e., the idiosyncratic risk as well as the probability associated with such risk from each group.[6] Moreover, estimated mass-point parameters shift the baseline hazard function, allowing for a different hazard function for each unobserved group.

This idea is potentially important to mortgage lenders because borrower characteristics are observed only at the time of loan application. Any unobserved changes in borrower characteristics may have a large impact on default or prepayment rates. This is particularly relevant to the move decision, where changes in employment or family status are likely to play

[4] See Deng, Quigley, and Van Order (2000) for a discussion.
[5] Neither list is exhaustive of the articles using the two methods addressed in this paper.
[6] Deng and Quigley (2002) model unobserved heterogeneity as a continuous distribution, and they use three-stage maximum likelihood estimation (3SMLE) methods. But, obtaining convergence when estimating these models is

an important role. Therefore, a statistical method for modeling unobserved borrower characteristics may improve the power to predict mortgage terminations by move, refinance or default.

This paper develops a mass-point mixed multinomial logit model (MML) that accounts for unobserved heterogeneity. Our extension of unobserved heterogeneity to the MNL model is motivated by the advantages mentioned above, and by the extensive use of MNL in the literature. Previous literature shows that the mass-point mixed technique adds significantly to the proportional hazard model (PHM), so it is worth testing for its contribution to the MNL model. We want to test for improvements in model efficiency and predictive power associated with accounting for unobserved heterogeneity.

Part of our agenda is to develop and implement a framework for cross-model-validation of mortgage termination risks. This allows us to judge any improvement in predictive accuracy that might be associated with adding unobserved subgroups to any model of mortgage terminations. Finally, we compare proportional hazard model (PHM) and MNL in terms of estimated coefficients, statistical significance and out-of-sample predictive ability. Such comparisons allow judgment about the qualitative differences among the models. The predictive test is a particularly demanding standard for unobserved heterogeneity, where the number of unobserved groups, their location parameter, and their frequency are difficult to estimate from the micro loan history data.

Our extension of the mass point mixed framework to the MNL model can be positioned in the literature as follows:

Model	**No Heterogeneity**	**Unobserved Heterogeneity**
Proportional Hazard Model (PHM)	Han-Hausman	MMH model of DQVO
with Competing Risks	(1990)	(2000)
Multinomial logit model (MNL)	Clapp *et al.*	MML model
with Event History	(2001)	developed here

difficult, and commercial software is unavailable at this time for estimating such model.

The comparison of these four models will test for economically significant (i.e., important) differences and for out-of-sample predictive ability.

The remainder of this paper is organized as follows: The next section summarizes option theory as applied to mortgage termination and it develops observable variables that intervene in the termination decision; section 3 discusses empirical methods of model estimation, the role of unobserved heterogeneity, out-of sample prediction and cross-model validation; section 4 describes the data; section 5 discusses results and section 6 presents conclusions.

Observable Variables and Mortgage Termination Decisions

Each month the borrower must decide whether to make the next regularly scheduled payment, refinance the mortgage, move and prepay the mortgage or default. This section summarizes observable variables associated with the borrower's decision and provides an overview of theory for each choice.[7] The relevant variables can be classified as personal characteristics (income, age, etc.), loan characteristics (loan amount, loan-to-value ratio and note rate), financial market conditions (the stochastic path of market interest rates) and housing market conditions.

Hendershott and Van Order (1987) and Kau and Keenan (1995) showed that the right to refinance the mortgage provides the borrower a call option on the mortgage debt with a strike price equal to the unpaid mortgage balance. Viewing the problem narrowly as the decision to exercise a call option or not, the relationship between the market value of the loan and the unpaid mortgage balance is the primary determinant in the choice to refinance. When the default option is added, house prices and interest rates become the two observable variables of primary interest.

The option-theoretic approach does not address the move decision.[8] The economic theory on

[7] In the residential mortgage market, default is a rare event comparing to a refinance or move. This is partly due to the high transactions costs (e.g., reputation costs) associated with mortgage default. In our sample of 1,985 loan records, there are only 27 default observations. Moreover, some defaults are worked out, and we do not have data on foreclosures. Therefore, in this study, we only focus on the competing risks of refinance and move. This analysis can be extended to the case of three competing risks of refinance, move and default, if the loan data contains sufficient observations of default events.

[8] The well developed option-theoretic approach in the literature assumes that the call and put options are exercised contingent upon the underlying market value of the mortgage contract and on property value.

household mobility points to a strong role for borrower characteristics in the move choice.[9] Clearly, a choice model with the move alternative needs more than the financial options related variables.

Observable Variables Explaining the Refinance Decision

Options pricing theory applied to mortgage refinancing implies that the borrower should exercise the option to call the debt whenever the market value of the mortgage exceeds the current balance by enough to cover the costs of refinancing. Transaction costs are treated as a constant increase in the strike price of the call option.

However, borrowers do not exercise the option to refinance as ruthlessly as do owners of other financial options (See, Green and LaCour-Little 1999, Deng, Quigley and Van Order 2000). This has led some researchers, such as Stanton (1995) and Green and LaCour-Little (1999), to treat transaction costs as varying across borrowers. However, in both studies, even implausibly high levels of transaction costs could not fully explain the observed prepayment behavior. Nevertheless, transactions costs suggest some observable variables that can be included in models of refinancing: For example, the larger the loan balance, then the greater the dollar amount of benefits from refinancing. This increases the probability of refinancing because fixed transactions costs (e.g., the time costs of refinancing) are more likely to be covered.

Since transaction costs alone seem insufficient to explain the under-exercise of the prepayment option, a number of researchers have incorporated the effects of institutional constraints on a borrower's ability to refinance. For example, Archer, Ling, and McGill (1996) used American Housing Survey data from 1985 and 1987 to examine the influence of post-origination income and collateral constraints on prepayment behavior. They found higher annual payment-to-income and loan-to-value ratios were negatively related to prepayments, after controlling for call option values. Deng, Quigley and Van Order (1996) found the importance of trigger events, such as unemployment and divorce, in affecting mortgage borrower's prepayment behavior. Caplin, Freeman, and Tracy (1997) found that regional recessions depressed prepayment rates by

[9] For reviews of the household mobility literature and its application to mortgage prepayment, see Clapp *et al*. (2000 and 2001), Pavlov (2001) and Quigley (1987).

as much as 50% in states with declining property markets. Mattey and Wallace (2001) and Downing, Stanton and Wallace (2001) found evidence that differences in house-price dynamics across regions are an important source of heterogeneity between mortgage pool performance. Bennett, *et al.* (2001) found strong evidence that poor credit history as well as high current LTV significantly reduced the probability of refinancing. These empirical findings are intuitive, for if collateral value declines below the loan balance, additional cash will be required to refinance. Similarly, a borrower whose income or financial position deteriorates may be unable to refinance due to payment-to-income or credit quality constraints.

In addition, making the right refinancing decision requires ongoing monitoring of market conditions and ready access to lenders. To the extent that particular demographic groups (e.g., minorities) have more limited access to information or lenders, we would expect that group to have higher transactions costs of refinancing (Deng and Gabriel 2004).

To summarize, the probability of refinancing is an increasing function of the market value of the loan, borrower income and the loan balance. It is a negative function of the current LTV, the probability of negative equity, the local unemployment rate, minority status and a low credit score dummy.

Observable Variables Explaining the Move Decision

Household mobility is a mechanism whereby households adjust their housing consumption to changes in circumstances (Rossi 1955). Theory says that a household's decision to move is based on housing "dissatisfaction", household characteristics and exogenous circumstances (e.g., job or family composition changes). The dissatisfaction that ultimately results in a move is the direct result of "changes in the needs of a household, changes in the social and physical amenities offered by a particular location, or a change in the standards used to evaluate these factors" (Speare 1974, p. 175).

Green and Shoven (1986) and Quigley (1987) documented a significant "lock-in" effect arising from below market rate financing. They found that homeowners with low mortgage rates (relative to current market rates) delayed moving. We extend this reasoning to an in-the-money

refinancing option, i.e., the borrower has a high mortgage rate relative to current market rates. In this case, the borrower has an added incentive to move to a new house, since the move effectively refinances the mortgage as well as dealing with housing dissatisfaction. Thus, we expect the market value of the mortgage[10] and the loan balance to be positively related to the move decision; the reasons are the same as for the expected positive signs in the refinancing equation.

Turning to borrower characteristics explaining the move decision, the age of the head of household has consistently been shown to have a strong, significant negative effect on household mobility (See, e.g., Quigley and Weinberg 1977, Myers, Choi and Lee 1997). A study by South and Crowder (1998) confirmed the importance of age and found that being married, having children and currently having a job significantly deterred household mobility and that household mobility increased with income. They also found that, controlling for these variables, African-Americans had lower household mobility than whites did.[11] Those studies that were able to track changes in family size, composition and income found that they were positively related to household mobility.[12] Unfortunately, our data only provides a snapshot description of the borrower at loan origination.

Pavlov (2001) found economic conditions to be important to move decisions, e.g. local unemployment rate was positively related to move because there might be attractive opportunities outside the local area, while the slope of the yield curve had a positive effect because it captured overall economic conditions. He also found that there was a burnout effect in move decision. In addition, variables related to the value of the mortgage were demonstrated to be insignificant to move, suggesting little synergy between the option to refinance and the move decision.

[10] In our sample period (1993–1998), the market variations in interest rates were much smaller than those studied by Green and Shoven (1986) and Quigley (1987) and so we do not expect as strong an influence from this term.

[11] The finding of lower household mobility for minorities has been reported by numerous earlier studies as well. See Quigley and Weinberg (1977). Yinger (1997) estimated that African Americans and Hispanics paid a discrimination "tax" of almost $4,000 every time they searched for a house to buy. Ross (1998) tested whether both race and job access had an independent effect on the probability of a joint residential move and job change. He found no evidence that race directly influenced the joint probability. However, because African-Americans are heavily concentrated in central cities, they had poorer job access and consequently lower job-related mobility.

In summary, we expect socioeconomic characteristics to have the effects found in previous literature: negative for age and minority status, positive for income. Given our inability to measure changes in demographic variables influencing demand, we expect time in the house to measure housing dissatisfaction and to be positively related to the probability of a move. This implies that the original refinance indicator (i.e., the loan was originated to refinance a previous loan) should have a positive sign because it indicates extra time in the house. Finally, the two variables from options pricing theory, the market value of the loan and the loan balance, should (in theory) be positively related to the move probability for the same reasons as for the refinancing decision.

Empirical Methods with Unobserved Borrower Characteristics

This section discusses the two dominant sets of modeling frameworks for competing risks of mortgage termination: the multinomial logit model (MNL) and the proportional hazard model (PHM). Both are estimated with and without unobserved heterogeneity using maximum likelihood methods. Our purpose here is to introduce unobserved borrower characteristics, evaluated with the mass-point mixed model, to the MNL framework. Since the use of this technique with the PHM has been established in previous literature, it is instructive to compare the four models.

What unobserved characteristics are likely to be most important to the move and refinance decision? We are interested here in personal characteristics of borrowers that can change in unobservable ways. Borrower age, minority status and sex are characteristics observed at the time of loan origination that can be projected with high accuracy beyond that time. Starting with the variables discussed in Section 2, we hypothesize that it will be difficult to observe or accurately predict changes in:

1) Marital status (single, married or divorced).

2) Births and deaths.

3) Income or job location, including labor force entry and exit and working at home part- or full-time.

[12] Elder and Rudolph (2000) found that change in job, divorce or the death of a spouse increased mobility.

This classification suggests a maximum of three unobservable groups.[13]

The idea behind the mass-point mixed model is to estimate unobserved characteristics as shifts in intercept, i.e., unobserved risk-spread, or "location", of the baseline hazard. The location of the first unobserved group is the constant term; the remaining groups shift the constant. These shifts are assumed to occur randomly, with fixed probabilities for the location of each group. Thus, the technique cannot classify individuals into groups, but only estimate the locations of groups and the associated probability of that location. The use of this technique can improve the fit of the model; presumably, this will improve predictive power, and we test for this.

The proportional hazard model (PHM) and MNL handle competing risks in very different ways. The PHM considers the joint survival probability and estimates the conditional probability of termination risks over time. It acknowledges that only the duration associated with the type which terminates first is observed and adjusts the equations for the probabilities of competing risks considering this effect. The PHM allows correlated risks. On the other hand, the MNL model directly models the probability of observing one risk versus another. At each observation point, the probabilities of refinance, move, default and continue to pay sum to one. An increase in one risk directly causes a decrease in at least one other risk. The MNL assumes that alternative termination risks are independent, which leads to the well-known IIA property.

Proportional Hazard Model (PHM)

Time to failure is the underlying random variable used in the proportional hazard model (PHM), one of the most frequently used models in the duration model framework, (Kalbfleisch and Prentice 2002). The model begins with a baseline time profile of the probability of termination conditional on the loan having survived to time *t, $h_0(t)$*. This baseline refinancing hazard can be shifted up or down by a factor that depends on the covariates, Z_{it} for observation *i* at time *t*:

$$h\left(t \mid Z_{it}, \eta_i\right) = h_0\left(t\right) \exp\left(Z_{it}\beta + \eta_i\right), \tag{1}$$

[13] Neighborhood characteristics might change in unobservable ways, and lenders are concerned with the evolution of house prices. In our work, prices are observed at the neighborhood level. This leaves maintenance of the individual house as a possible omitted category.

where η_i is the unobserved heterogeneity for individual *i*. Following Deng, Quigley and Van Order (2000), equation (1) can be generalized to the joint survivor function for refinance (*p*) and move (*m*). The joint survivor function is conditional on η_p, η_m, and *Z* (For simplicity, we omit the subscript *i*):

$$S\left(t_p, t_m \mid Z, \eta_p, \eta_m, \theta\right) = \exp\left\{-\eta_p \sum_{q=1}^{t_p} \exp\left(\gamma_{pq} + \beta_p^{'} Z\right) - \eta_m \sum_{q=1}^{t_m} \exp\left(\gamma_{mq} + \beta_m^{'} Z\right)\right\}, \tag{2}$$

where *Z* is a vector of covariates that have impacts on borrowers' prepayment and move decisions. Some of the covariates may be time-varying function of the contemporaneous market rate, *r*, and contemporaneous market value of the property, *H*. *Z* may also include other time invariant covariates, such as borrower characteristics (e.g., credit history, borrower age, income, ethnic background) and loan characteristics (e.g., original loan amount, length of the mortgage contract, initial discount points, and refinancing loan indicator). η_p and η_m are unobserved error terms associated with the hazard functions for prepayment and move respectively. θ is a vector of parameters (*e.g.*, γ and β) of the hazard function. γ_{jq} are parameters of the baseline hazard function, where *q* indexes discrete periods in the time dimension since the loan origination. The baseline may be estimated nonparametrically, following Han and Hausman (1990):

$$\gamma_{jq} = \log\left[\int_{q-1}^{q} h_{0j}(s)\, ds\right]; \quad j = p, m. \tag{3}$$

In order to construct the likelihood function, we need to first write down the probabilities of alternative termination risks as functions of the joint survival function. Then the likelihood function is the joint density function of competing risks for all observations[14]. It is noteworthy that the full maximum likelihood hazard model presented here differs from the Cox Partial Likelihood hazard model used by Clapp *et al*. (2001) and by Pavlov (2001): the Cox Partial Likelihood hazard model separates the baseline estimation from covariates estimation and its likelihood function is parallel to the MNL model with event history data.

The unobserved heterogeneity $\left(\eta_p, \eta_m\right)$ in equation (2) can be modeled and estimated in a mass point mixed hazard model framework following Deng, Quigley and Van Order (2000) by

[14] See Deng, Quigley and Van Order (2000) for details.

assuming the population of borrowers in the sample consists of L distinct groups[15]. The unobserved (η_p, η_m) are assumed to follow a joint mass point distribution characterized by the doublet of location parameters $(\eta_{p_l}, \eta_{m_l}), l = 1,2,...,L$, occur in the population with relative frequency $p_l, l = 1,2,...,L.$ The parameters shifting the constant (η_{p_l}, η_{m_l}) and mass point parameters p_l are estimated jointly with parameters of the survivor function, θ[16].

Multinomial Logit Model (MNL)

The multinomial logit model (MNL) provides an alternative approach to estimating a competing risks model. It treats the dependent variable as a polytomous qualitative choice variable. This model provides explicitly for competing risks, and it can be estimated with commercial software. But it requires a different assumption – the independence of irrelevant alternatives (IIA).

Consider a single prepayment risk. Previous literature shows that bivariate logit with a restructured data set provides a convenient method for dealing with prepayment risk of the mortgage borrower over time (see Clapp *et al*. 2000 and 2001)[17]. The information for each loan is restructured to include one observation for each time period in which that loan is active (i.e., from origination up to and including the period of termination). The restructured data and the use of $\gamma(t)$ to model the time varying intercept generalizes from the logit model. Generalize to multiple risks by letting Y_{it} represent the ith borrower's decision at time t. The log-likelihood function is:

$$\ln L = \sum_{t=1}^{T} \sum_{i=1}^{n_t} \sum_{j=0}^{2} d_{ijt} \ln \left(\Pr\left(Y_{it} = j\right) \right), \tag{4}$$

$$\Pr\left(Y_{it} = j\right) = \frac{e^{\gamma(t)+\beta_j' Z_{it}+\eta_i}}{1+\sum_{k=1}^{2} e^{\gamma(t)+\beta_k' Z_{it}+\eta_i}} \quad \text{j = 1, 2, and } \Pr\left(Y_{it} = 0\right) = \frac{1}{1+\sum_{k=1}^{2} e^{\gamma(t)+\beta_k' Z_{it}+\eta_i}} \tag{5}$$

[15] The proportional hazard model without unobserved heterogeneity can be viewed as a special case of the mass point mixed hazard model where L equals 1.

[16] See Deng, Quigley and Van Order (2000) for a detailed discussion of the estimation of a mass point mixed hazard model.

[17] Related literature includes Jenkins (1995); Bergström and Edin (1992); and Narendranathan and Stuart (1993).

In equations (4) and (5), n_t is the number of observations in the restructured data at time t (t = 1, … T), j indexes the possible choices (continue, refinance, move) and d_{ijt} is an indicator variable which takes on a value of one when the alternative j is chosen in the ith observation at time t, otherwise zero.

Once the data are restructured, the MNL is estimated using maximum likelihood by treating restructured discrete time information for each loan as i.i.d. records in the sample.[18] In equation (5), the η_i represent unobserved heterogeneity and can be estimated in a mass point mixed model framework. A full explanation of how to estimate a mass point mixed multinomial logit model (MML) is given in Appendix A.

Note that competing risks are included in Equations (4) and (5) by having probabilities that must sum to one. Thus, an increase in the probability of one risk must necessarily be associated with a decline in the probability of at least one other risk.

The MNL requires independence of irrelevant alternatives (IIA): the odds ratio for any pair of choices is assumed independent of any third alternative. Elimination of one of the choices should not change the ratios of probabilities for the remaining choices. Choices that are close, in the sense that their utilities are stochastically correlated, violate the IIA assumption. The MNL also assumes that choices at any point in time are independent of those at any other point in time. Limited path dependence may be introduced into the model by adding explanatory variables: e.g., a burnout variable to measure foregone opportunities to refinance at lower interest rates.[19]

Data[20]

Loan Histories

Table 1 describes data from a large loan servicer and originator; the data include information on 1,985 fixed-rate mortgages with both 30-year and 15-year maturities. Approximately 79% of the

[18] The maximum likelihood estimation approach for MNL with restructured discrete time period data does not account for potential autocorrelations among event history records for each individuals. Therefore, the estimates may be biased if such autocorrelation exist among restructured discrete time period records for each individual borrower.
[19] We experiment with the standard measures of mortgage burnout and do not find them to be significant.
[20] Clapp, et al. (2000 and 2001) provide more detail on the data sources and on manipulation of the data.

loans were originated to refinance an existing mortgage loan on the same property while 21% were loans for home purchases. The majority of the loans (64%) were originated by correspondents or brokers and purchased by the lender providing the data; the lender originated the remainder.

Loans were underwritten according to standard policies in effect during 1993 and 1994, including scoring loans using an internally developed mortgage credit scoring model that adds certain borrower and loan characteristics, including LTV, to traditional credit bureau measures, in order to estimate borrow creditworthiness.

Because of high housing costs in California, the loans had an average original loan balance of $167,600. Approximately 73% had original loan amounts below the GSE limits for 1993 and 1994 making them eligible for purchase by Fannie Mae and Freddie Mac.

Table 2 contains values for the estimated market price of the loans and other time-varying covariates. All data are quarterly, the smallest time interval common to all variables.[21] The table shows how these covariates change, on average, over time compared to the values at origination.

Data Used for House Price Indices and to Identify Refinances

We purchased six years of transactions data from the California Market Data Cooperative, Inc (CMDC). CMDC collects, verifies and, if necessary, corrects all property transactions from the county records. The sales for Contra Costa, Los Angeles and Orange counties are from the period from January 1993 through December 1998.

CMDC data contain a full street address for each property that sold as well as the date of sale, sales price, appraised value and recorded first mortgage loan. They also contain considerable detail on the property, including square footage, bathrooms, bedrooms and year built.

Identifying Movers and Refinancers

We match the full street address of the collateral underlying the loan, the origination date, loan

[21] Unemployment and neighborhood house price indices are estimated at the quarterly level to avoid excessive noise. Monthly data could be smoothed, but this would introduce time dependence. Details on estimation of neighborhood indices (used to estimate current loan to value and probability of negative equity) are given in Clapp *et al*. (2004).

amount and appraisal value to the housing transactions data to identify movers. When we find a house sale in the transaction data with the same address and a sale date close to the date of loan termination, we identify the prepayment as being the result of a move. When we find no match, we conclude that the prepayment was caused by a refinance.

As of December 31, 1998, 27 loans (1.4%) had terminated by default and 573 loans (28.9%) had terminated by prepayment. We estimate that moves triggered 252 of the prepayments and refinancing resulted in the remaining 321 prepayments. Since there were only 27 defaults, we did not estimate the default equation. Defaults become censored observations; they are no longer at risk for the other termination hazards. Models with move, refinance and default are discussed in Clapp *et al.* (2001) and Pavlov (2001).

Results

Table 3 compares the multinomial logit model (MNL) to the proportional hazard model (PHM); both are then estimated with unobserved heterogeneity: Models 3, the mass point mixed multinomial logit model (MML), and 4, the mass point mixed hazard model (MMH). We estimated these four models for two and three unobserved groups (L= 2 or 3): results for two groups are presented because three groups did not significantly improve the log likelihood.[22]

The table shows that the likelihood is significantly improved by using the mass point mixed version of the proportional hazard model (PHM) but not by the multinomial logit model (MNL).[23] The *A.I.C.* and pseudo R^2 are virtually unchanged when adding unobserved heterogeneity to the MNL model whereas the pseudo R^2 was improved substantially in the case of the PHM. The *B.I.C.* is unfavorable to MNL whereas it is neutral for PHM. The mass point for model 4 (one mass point is normalized to one, the other separately estimated) is strongly significant. Therefore, this specification (Model 4, MMH) will be evaluated and then compared

[22] All models reported in Table 3 are specified with Han-Hausman flexible baseline function (non-parametric baseline). As discussed in Appendix B, we adopted this baseline function for all alternative models because it substantially improved the likelihood value. Since we use the same baseline for all alternative models, we choose not to report the massive number of baseline estimates in Table 3. The estimated baseline functions are available upon request.

[23] The chi-square statistic is 3.80 for the logit model pair and 31.00 for the hazard models; the critical value is 7.82 with p=.05.

to Model 3 (MML).[24]

A few of the explanatory variables deserve explanation beyond the discussion in Section 2. Since house price indices were estimated at the neighborhood level, the current loan-to-value ratio gives the mean effect of house price appreciation and loan amortization at the property level on equity available for a move or refinance. The probability of negative equity indicator estimates a second order effect from the volatility of house prices; this variable is one if the probability of negative equity is greater than 90 percent, otherwise zero. We expect these variables to be negatively related to move and refinance.[25]

Borrower age should be negatively related to the probability of a move (see the discussion in Section 2.2). We hypothesized that this effect would be attenuated for borrowers who have experienced a lot of house price appreciation, because these borrowers have the equity necessary for a move. Hence, we developed the borrower age variables described in the tables.

For the refinance equation, all the significant coefficients in model 4 have the expected signs except the high credit score indicator. Since its t-value is close to the critical value, this coefficient might be the one out of twenty that represents a Type I error. Other signs in the refinance equation agree with theory and previous empirical literature. It is important that model 4 finds a strong positive sign on the market price of the loan and negative signs on current LTV, unemployment and minority status.

Model 4 reveals new information about the probability of a move. Most importantly, a negative sign is obtained for estimated discount points and a positive sign for the refinance loan indicator. The negative coefficient on discount points is what one would expect if paying higher discount points signals an intention of remaining in the house for a long time. When the mortgage is originated to refinance another loan, then the borrower has been in the house for a longer time

[24] All models are estimate with maximum likelihood methods. It was difficult to get convergence, especially for the MMH model. Various parameters were restricted for this purpose: The two tails of the baseline or the mass point in the moving function. These restrictions influence a few of the coefficients, generally in obvious ways, but most are insensitive to the restrictions chosen.

[25] Similarly, the obligation ratio should be negatively related to move or refinance because it indicates the stress placed on borrower income by fixed obligations at the time of loan origination.

than a purchase-money borrower with the same loan age. The positive sign is consistent with a higher level of dissatisfaction with the current bundle of housing characteristics.

Continuing with the move model, the current LTV has a positive sign whereas the probability of negative equity indicator has a strongly significant negative sign. A possible explanation is that declining house values signal neighborhood deterioration. The literature documents substantial turnover during these transitional periods. However, once equity has become negative, many households will have insufficient down payment to move to a new home: They are trapped in the declining area.

Comparison of Proportional Hazard Models with and without Unobserved Heterogeneity

The proportional hazard move model with unobserved heterogeneity (model 4) performed very well compared to the model without heterogeneity (model 2) in the sense that coefficients have plausible signs and there are more significant coefficients. Larger absolute coefficients together with larger t-values are obtained for discount points, original refinance indicator, CLTV, borrower age, minority indicator and borrower income. The only variable where model 2 has a larger coefficient is on the 15-year loan indicator, and this is marginally significant.

The larger absolute value of significant coefficients with heterogeneity indicates that different economic scenarios will have a bigger impact on predicted moves. This makes sense: the heterogeneity adjustment is intended to capture unobserved borrower characteristics. Changes in these move-related characteristics (e.g., changes in family income, wealth, credit rating, family status or family size) are difficult to observe in general, and specifically missing from our dataset. Several measured borrower characteristics (borrower age, minority status and income at the time of the loan application) become more significant in the proportional hazard model.

The introduction of unobserved heterogeneity makes little difference to the refinance estimates (Table 3, refinance column for model 4 compared to model 2). The original refinance indicator is higher in model 4, but other significant coefficients are very similar. Again, this makes sense in light of the four categories of variables influencing refinancing: personal characteristics, loan

characteristics, financial market conditions and housing market conditions.[26] Only the first category is likely to change in unobservable ways. The market price of the loan is the major driver of refinancing, and this is not sensitive to missing borrower characteristics.

The MNL Model

The mass point mixed MNL (model 3) is able to estimate two statistically significant groups of unobserved heterogeneity in borrowers refinance behavior. However, the MNL model is not significantly improved by the introduction of unobserved heterogeneity: The log likelihood is not significantly reduced as measured by a likelihood ratio test. Similarly, MNL refinance coefficients are not changed much by the mass point mixture method (model 3 compared to model 1).

The MNL refinance estimates with unobserved heterogeneity (MML) (Table 3, model 3, refinance column) can be compared to the corresponding proportional hazard model refinance estimates (MMH) (Table 3, model 4, refinance column). The MML model has more precisely measured baseline intercept estimates. The two models have about the same magnitudes for significant coefficients except for the minority indicator, which is about 2.5 times larger in the MML model. Also, the MML model is less sensitive to the high credit score indicator; this is a desirable characteristic since the sign is incorrect. The MNL refinancing model (model 1), has estimated coefficients that compare favorably to the MMH model. The MNL has larger signs and significance for original loan balance, original refinance, house value appreciation (age > 40), minority indicator and low credit score indicator; this compares to larger MMH coefficients for the market value of the loan and current LTV. Overall, one can conclude that the two sets of estimates are roughly similar, and that the addition of unobserved heterogeneity makes little difference to the refinance model. Thus, the MNL model might be preferable in some applications because it is easy to estimate with commercial software.

When the MNL move model with unobserved heterogeneity (MML) (Table 3, model 3, move column) is compared to the corresponding proportional hazard estimates (MMH) (Table 3, model 4, move column), the significant MML coefficients are generally smaller in absolute value.

[26] See the discussion in Section 2, especially Section 2.1.

Moreover, several important coefficients are disappointing in the MML move model: original refinance indicator and current LTV are insignificant, with marginal significance for discount points and borrower age. Overall, model 4, the mass pointed mixed hazard model dominates the rest of the models in goodness of fit measured by log likelihood value of the estimation.

Out-of-sample predictive accuracy (Cross-model-validation)

Table 4 presents the cross-model-validation results. The R-square for each individual model's out-of-sample prediction is very low.[27] To some extent this is an artifact of the sparse matrix of observations on the dependent variable. The dependent variable in any quarter is highly likely to be a zero (continue to pay) and any ones (prepay) are divided roughly equally between refinances and moves. Thus, the low R-squares do not tell us that all four models are poor in out-of-sample prediction. Rather, what we want to look at from these results is the relative performance of the four models. Actually the cross-model-validation procedure we implement here is designed just for this purpose.

For the refinance model, Model 4 (MMH) is clearly superior to the others. Thus, the model that minimizes the likelihood also does best out-of-sample. Also, the two models that account for unobserved heterogeneity have better out-of-sample predicting capacities than those not accounting for unobserved heterogeneity.

The move model does not show as clear a pattern as the refinance model. Model 4 does the best job. However, model 1 (MNL) and model 3 (MML) show no big difference, and model 2 (PHM) has poor performance. This may be partly due to the highly heterogeneous behavior of movers ─ it is hard to estimate mass-points when the sample is split into limited number of discrete groups.

Conclusions

This paper disaggregates the prepayment decision into moving and refinancing. Two models for estimating these hazards are compared: The proportional hazard model (PHM) and multilnomial logit model (MNL) with event history. Both models are developed to account for unobserved heterogeneity; with the heterogeneity adjustment, missing information on borrower

[27] The details for the construction of Table 4 are contained in Appendix B.

characteristics is modeled in terms of borrowers randomly chosen from different probability distributions.

House sales prices for three counties in California are used to identify moves and to estimate neighborhood house price appreciation, current loan to value and probability of negative equity. This information is combined with loan histories for nearly 2,000 mortgages originated in 1993 and 1994 in the three counties. Few defaults were observed, so the models were estimated with two competing termination hazards: refinancing and moving.

Unobserved heterogeneity combined with the proportional hazard model (mass point mixed hazard model, MMH) produces the smallest log likelihood value, but it can be difficult to obtain convergence with the maximum likelihood procedure. Within the proportional hazard framework, unobserved heterogeneity makes a big difference to the move coefficients, less so to the refinance coefficients. This is plausible given that unobserved heterogeneity is a way of dealing with missing borrower characteristics. A list of missing characteristics that might influence the move decision includes changes in family income, wealth, credit rating, family status or family size. It would appear that changes in one or more of these characteristics would occur frequently, so that unobserved heterogeneity should matter a lot to the move decision.[28] On the other hand, it is a subset of these characteristics that matter to loan refinancing, and the financial value of the refinancing option (measured here by the estimated market price of the loan) would plausibly overwhelm demographic characteristics.

Comparison of the multilnomial logit model (MNL) refinancing model with the proportional hazard model (PHM) shows little difference in the coefficients, regardless of unobserved heterogeneity. All four models have similar significant coefficients with plausible signs. This suggests that commercial software can be used to estimate a MNL refinancing model without unobserved heterogeneity.

[28] Most of the improvements in the move function after controlling for unobserved heterogeneity are focused on borrower characteristics observed at the time of loan origination (greater significance for borrower age, minority indicator, borrower income and high credit score indicator).

Estimating the move equation with the proportional hazard model (PHM) and unobserved heterogeneity produce many more significant coefficients than the MNL model. Thus, we conclude that the proportional hazard model is the preferred method for learning about the move relationship.

Our cross-model-validation through the comparison of out-of-sample predicting capabilities shows results consistent with those from estimated model coefficients. The mass point mixed hazard model (MMH) has the best out-of-sample prediction. Our stratification method (See Appendix B and Table 4) provides an important approach for cross-model-validation of mortgage termination risk modeling.

Appendix A. Estimating the MNL Model with Unobserved Heterogeneity

The log likelihood function for a multinomial logit model (MNL) can be expressed as:

$$\log L = \sum_{i=1}^{n}\sum_{j=0}^{J} y_{ij} \log \pi_{ij}, \tag{A1}$$

where y_{ij} = 1 if individual *i* terminates for reason *j*, otherwise zero; *j*= 0 for continue to pay, 1 for refinance and 2 for move. Define:

$$\pi_{ij} \frac{\exp\left(\gamma_j + z_i^{'}\beta_j\right)}{\sum_{k=0}^{J}\exp\left(\gamma_k + z_i^{'}\beta_k\right)}, \tag{A2}$$

and γ_j (a $T \times 1$ vector) is the log transform of group baseline (in our case it is a step function of discrete periods of duration since the loan origination, and total number of discrete period is *T*). Following Han and Hausman (1990):

$$\gamma_{jq} = \log\left[\int_{q-1}^{q} h_{0j}(s)\,ds\right]; \quad j = 1,...,J; \quad q = 1,...,T. \tag{A3}$$

where $\gamma_{jq}, q = 1,...,T,$ are the T elements of γ_j, and they are estimated jointly with the parameter of the multinomial logit moel.

Following McFadden and Train (2000), the modified log likelihood function for mass point mixed multinomial logit model (MML) can be expressed as following:

$$\log L = \sum_{i=1}^{n}\sum_{j=0}^{J} y_{ij} \log \Pi_{ij}, \tag{A4}$$

where

$$\Pi_{ij} = \sum_{l=1}^{L} p_l \pi_{ijl}, \tag{A5}$$

$$\begin{aligned}\pi_{ijl} &= \frac{\eta_{jl}\exp\left(\gamma_j + z_i^{'}\beta_j\right)}{\sum_{k=0}^{J_i}\eta_{kl}\exp\left(\gamma_k + z_i^{'}\beta_k\right)} \\ &= \frac{\exp\left(\gamma_j + z_i^{'}\beta_j + \log\left(\eta_{jl}\right)\right)}{\sum_{k=0}^{J}\exp\left(\gamma_k + z_i^{'}\beta_k + \log\left(\eta_{kl}\right)\right)},\end{aligned} \tag{A6}$$

η_{jl} is the location parameter (that can be interpreted as idiosyncratic risk added to the baseline hazard function) associated with risk type j for the lth unobserved heterogeneous group; p_l is the mass-point parameter (frequency) for the lth group, $l = 1, \ldots, L$. The location parameters η_{jl}, and the mass-point parameters p_l are estimated jointly with coefficients of the proportional function β and the flexible baseline function γ_{jq}. Note that we normalize the mass-point parameter associated with the first group to 1, so that the probability of individual belonging to first group is *1/(1+p)*, and the probability of individual belonging to second risk group is *p/(1+p)*.

Estimation with unobserved heterogeneity is achieved with any software that can maximize a likelihood function. The η_{jl}, γ_{jq}, and β terms are used to evaluate equation (A6); the result is multiplied by the probabilities, p_l, summed over j, and inserted into equation (A5). The mass point mixed hazard model (MMH) of Deng, Quigley and Van Order (2000), equation(2), is estimated in essentially the same way.

Appendix B. Choice of the Baseline Function and Out-of-sample Tests

The purpose of this appendix is to present pertinent findings that would have detracted had they been presented in the main text of this paper.

Choice of the Han-Hausman Baseline Hazard Function

Table B.1 compares the maximum likelihood estimation of the proportional hazard model (PHM) specified with three different baseline functions. We choose the Han-Hausman flexible baseline specification, which is a step function taking T different values for the series γ_q. The flexible baseline is estimated non-parametrically. Therefore it is less restrictive compared to the other two parametric baseline forms, e.g., the PSA schedule and the 5th order polynomial function, respectively.

Table B.1 shows that the likelihood is substantially improved by using the Han-Hausman flexible baseline function to estimate the PHM. Therefore, this baseline specification is retained in all

tables in the paper for the purposes of comparing this model to the others with and without unobserved heterogeneity.

Out-of-sample Tests

For out-of-sample tests, we followed a procedure presented by Hosmer and Lemenshow (2000) and by Pampel (2000). The method can be outlined as follows:

Sampling (Estimation sub-sample and validation sub-sample formation)

1) Use the full sample (1,985 loans) to estimate a model 2 (PHM);

2) Use the above estimated model 2 (PHM) to predict the refinance (move) probability at termination point of each of the 1,985 loans;

3) Sort the 1,985 loans by the above predicted refinance (move) probabilities;

4) Divide the above-sorted sample evenly into ten sub-samples;

5) From each of the above ten sub-samples, randomly draw 90% (Uniform Distribution), then stack the ten 90% sub-samples together to form the estimation sub-sample;

6) Use the left 10% loans as the validation sub-sample.

Estimation and prediction

1) Use the estimation sub-sample to estimate the four models separately;

2) Predict the refinance and move probability at termination with the above estimated models.

Cross-model validation

1) Regress the real events on the predicted hazard rate (refinance and move are done separately)

2) Compare the R-squares of the above regressions.

Given that refinance and move are done separately during the validation, we have two panels of Table 4 in the body of the paper. Table B.2 contains the estimates based on the 90% sample.

References

Ambrose, B.W. and C.A. Capone, Jr. 1998. Modeling the Conditional Probability of Foreclosure in the Context of Single-Family Mortgage Default Resolutions. *Real Estate Economics* 26(3): 391-430.

Ambrose, B.W. and C.A. Capone, Jr. 2000. The Hazard Rates of First and Second Defaults. *Journal of Real Estate Finance and Economics* 20(3): 275-293.

Ambrose, B.W. and A. B. Sanders. 2003. Commercial Mortgage Backed Securities: Prepayment and Default. *Journal of Real Estate Finance and Economics* 26(2/3): 179-196.

Archer, W., D. Ling and G. McGill. 1996. The Effect of Income and Collateral Constraints on Residential Mortgage Terminations. *Regional Science and Urban Economics* 26(3/4): 235-261.

Berkovec, J.A., G.B. Canner, S.A. Gabriel and T. Hannan. 1998. Discrimination, Competition, and Loan Performance in FHA Mortgage Lending. *Review of Economics and Statistics* 80(2): 241-250.

Bennett, P., R. Peach and S. Peristiani. 2001. Structural Change in the Mortgage Market and the Propensity to Refinance. *Journal of Money, Credit and Banking* 33(4): 955-975.

Bergström, R. and P. A. Edin. 1992. Time Aggregation and the Distributional Shape of Unemployment Duration. *Journal of Applied Econometrics* 7(1): 5-30.

Caplin, A., C. Freeman and J. Tracy. 1997. Collateral Damage: Refinancing Constraints and Regional Recessions. *Journal of Money, Credit and Banking* 29(4): 497-516.

Clapp, J. M. 2004. A Semi Parametric Method for Estimating Local House Price Indices. *Real Estate Economics* 32(1): 127-160.

Clapp, J. M., J. Harding and M. LaCour-Little. 2000. Expected Mobility: Part of the Prepayment Puzzle. *Journal of Fixed Income* 10(1): 68-78.

Clapp, J. M., G. Goldberg, J. Harding and M. LaCour-Little. 2001. Movers and Shuckers: Interdependent Prepayment Decisions. *Real Estate Economics* 29(3): 411-450.

Cox, J. C., J. E. Ingersoll and S. A. Ross. 1985. An Intertemporal General Equilibrium Model of Asset Prices. *Econometrica* 53(2): 363-384.

Cunningham, D., and C. Capone. 1990. The Relative Termination Experience of Adjustable to Fixed-Rate Mortgages. *Journal of Finance* 45(5): 1687-1703.

Deng, Y., J. M. Quigley and R. Van Order. 1996. Mortgage Default and Low Down Payment Loans: The Cost of Public Subsidy. *Regional Science and Urban Economics* 26(3/4): 263-285.

Deng, Y., J. M. Quigley and R. Van Order. 2000. Mortgage Terminations, Heterogeneity and the Exercise of Mortgage Options. *Econometrica* 68(2): 275-307.

Deng, Y. and J. M. Quigley. 2002. Woodhead Behavior and the Pricing of Residential Mortgages. University of Southern California, CA, Lusk Center Working Paper, No. 2003-1005.

Deng, Y. and S. A. Gabriel. 2004. Are Underserved Borrowers Lower Risks? New Evidence on the Performance and Pricing of FHA-Insured Mortgages. University of Southern California, CA, Finance and Business Economics Working Paper, No. 2004-1004.

Downing, C., R. Stanton and N. Wallance. 2003. An Empirical Test of a Two-Factor Mortgage Prepayment and Valuation Model: How Much Do House Prices Matter? FEDS Working Paper, No. 2003-42.

Elder, H. and P. M. Rudolph. 2000. Mobility and Housing Tenure Transitions of Older Americans. University of Alabama, Tuscaloosa, AL, draft paper.

Green, R. K. and M. LaCour-Little. 1999. Some Truths About Ostriches: Who Never Refinances Their Mortgage and Why They Don't. *Journal of Housing Economics* 8(3): 233-248.

Green, J. and J. B. Shoven. 1986. The Effect of Interest Rates on Mortgage Prepayment. *Journal of Money, Credit and Banking* 36(1): 41-58.

Han, A. and J. A. Hausman. 1990. Flexible Parametric Estimation of Duration and Competing Risk Models. *Journal of Applied Econometrics* 5(1): 1-28.

Harding, J. 1994. Rational Mortgage Valuation Using Optimal Intertemporal Refinancing Strategies and Homogeneous Borrowers. University of California, Berkeley, CA, Ph.D. Dissertation.

Hendershott, P. and R. Van Order. 1987. Pricing Mortgages: An Interpretation of Models and Results. *Journal of Financial Services Research* 1(1): 19-55.

Hosmer, D. W. and S. Lemeshow. 2000. *Applied Logistic Regression* (2nd edition). New York: John Wiley & Sons.

Jenkins, S. P. 1995. Practitioner's Corner: Easy Estimation Methods for Discrete-Time Duration Models. *Oxford Bulletin of Economics and Statistics* 57(1): 129-138.

Kalbfleisch, J. D. and R. L. Prentice. 2002. *The Statistical Analysis of Failure Time Data* (2nd edition). New York: John Wiley & Sons.

Kau, J.B. and D.C. Keenan. 1995. An Overview of the Option-Theoretic Pricing of Mortgages. *Journal of Housing Research* 6(2): 217-244.

Mattey J. and N. Wallace. 2001. Housing-price cycles and prepayment rates of US mortgage pools. *Journal of Real Estate Finance and Economics* 23(2): 161-184.

McFadden, D. and K. Train. 2000. Mixed MNL Models of Discrete Response. *Journal of Applied Econometrics* 15(5): 447-470.

Myers, D., S. S. Choi and S. W. Lee. 1997. Constraints of Housing Age and Migration on Residential Mobility. *Professional Geographer* 49(1): 14-28.

Narendranathan, W. and M. B. Stewart. 1993. Modelling the Probability of Leaving Unemployment: Competing Risks Models with Flexible Base-line Hazards. *Applied Statistics* 42(1): 63-83.

Pampel, F. C. 2000. *Logistic Regression: A Primer*. Thousand Oaks: Sage.

Pavlov, A. 2001. Competing Risks of Mortgage Termination: Who Refinances, Who Moves, and Who Defaults? *Journal of Real Estate Finance and Economics* 23(2): 185-211.

Philips, R. A., E. Rosenblatt and J. H. VanderHoff. 1995. The Probability of Fixed and Adjustable Rate Mortgage Termination. *Journal of Real Estate Finance and Economics* 13(2): 95-104.

Quigley, J. M. 1987. Interest Rate Variations, Mortgage Prepayments, and Household Mobility. *Review of Economics and Statistics* 69(4): 636-643.

Quigley, J.M., and R. Van Order. 1995. Explicit Tests of Contingent Claims Models. *Journal of Real Estate finance and Economics* 11(2): 99-117.

Quigley, J. M. and D. H. Weinberg. 1977. Intra-Urban Residential Mobility: A Review and Synthesis. *International Regional Science Review* 2(1): 42-66.

Ross, S. L.1998. Racial Differences in Residential and Job Mobility: Evidence Concerning the Spatial Mismatch Hypothesis. *Journal of Urban Economics* 43(1): 112-135.

Rossi, P. H. 1955. *Why Families Move*. Glencoe: The Free Press.

South, S. J. and K. D. Crowder. 1998. Leaving the Hood: Residential Mobility Between Black, White, and Integrated Neighborhoods. *American Sociological Review* 63: 17-26.

Speare, A. Jr. 1974. Residential Satisfaction as an Intervening Variable in Residential Mobility. *Demography* 11: 173-188.

Stanton, R. 1995. Rational Prepayment and Valuation of Mortgage-Backed Securities. *Review of Financial Studies* 8(3): 677-708.

Schwartz, E. S. and W. N. Torous. 1989. Prepayment and the Valuation of Mortgage-Backed Securities. *Journal of Finance* 44(2): 375-392.

Train, K. 1986. *Qualitative Choice Analysis*. Cambridge, Massachusetts: The MIT Press.

Train, K. 2003. *Discrete Choice Methods with Simulation*. Cambridge, UK: Cambridge University Press.

Vandell, K.D., W.C. Barnes, D.J. Hartzell, D. Kraft and W. Wendt. 1993. Commercial Mortgage Defaults: Proportional Hazards Estimation Using Individual Loan Histories. *Journal of American Real Estate and Urban Economics Association* 20(4): 55-88.

Yinger, J. 1997. Cash in Your Face: The Cost of Racial and Ethnic Discrimination in Housing. *Journal of Urban Economics* 42(3): 339-365.

Table 1 Means and standard deviations of variables from the loan application

Variables	Means (STDs)	Description
Original loan balance ($000)	167.63 (121.83)	Face amount of the mortgage at the date of origination (in 1993 or 1994), thousands of dollars.
15-year loan indicator	0.31 (0.46)	Indicator variable is one if the loan has a 15 year maturity, zero if a 30 year maturity.
Discount points (estimated)	2.00 (1.43)	We regressed the loan coupon rate on current treasury rates and on loan and borrower characteristics. The residuals from this equation provide a measure of discount points since a borrower paying a rate substantially below the predicted rate must have "bought down" the rate by paying above average points.
Original refinance indicator	0.79 (0.41)	One if the mortgage at the date of origination (in 1993 or 1994) was to refinance a previous mortgage, zero if it was to purchase the home.
Borrower age	46.74 (11.18)	Age of the borrower in years, from the loan application.
Minority indicator	0.23 (0.42)	Indicator variable equal to one if the application classifies borrowers into any one of three minority groups, otherwise zero.
Borrower income ($000)	8.08 (9.03)	Monthly household income at the time of origination.
Obligation ratio (%)	30.12 (9.60)	The ratio of fixed expenses to borrower income. This is the standard ratio used by lenders when evaluating loan applications.
High credit score indicator	0.64 (0.48)	The high score indicator flags borrowers with credit scores greater than 1000 on the lender's proprietary scale.
Low credit score indicator	0.103 (0.30)	The low score indicator flags borrowers with credit scores less than 800 on the lender's proprietary scale.
Number of observations	1,985	Number of loans with data on all variables.

Notes:

1. Standard deviations are in parentheses.
2. We use standardized value for each continuous variable during the estimation, i.e. all continuous variables have zero mean and unit variance.

Table 2 Means and standard deviations of time-varying variables at origination and termination

Variables	At Origination				At Termination	
	All Loans	Refinance	Move	Other	Refinance	Move
Market price of loan	100.00 (1.86)	99.85 (1.43)	99.89 (1.33)	100.06 (2.02)	101.36 (3.18)	98.53 (3.12)
Prob. Negative Equity > 90 Percent Indicator (0/1)	0.15 (0.36)	0.10 (0.30)	0.08 (0.27)	0.18 (0.38)	0.09 (0.29)	0.04 (0.20)
Current loan-to-value (%)	60.17 (22.57)	52.93 (22.03)	54.03 (21.34)	62.91 (22.33)	54.63 (23.41)	55.38 (22.14)
House price appreciation* Age*indicator (Age<40) $,000	-389.73 (2576.52)	-104.38 (805.56)	-264.68 (1625.00)	-476.93 (2947.62)	6831.10 (25392.27)	3383.82 (14936.35)
House price appreciation* Age*indicator (Age>40) $,000	-623.00 (4502.84)	-338.88 (1597.36)	-247.63 (643.59)	-754.58 (5272.20)	16542.46 (42970.32)	4146.33 (28597.13)
Unemployment rate	8.28 (1.67)	8.22 (1.64)	8.35 (1.62)	8.28 (1.68)	5.69 (1.80)	6.01 (1.79)
Number of Observations	1,985	321	252	1,412	321	252

Notes:

1. Standard Deviations are in parentheses.
2. The data are structured as a panel dataset, with one observation for each quarter for each loan during the observation period. Loans were observed from origination to termination, or 12/31/1998, whichever was earlier. The market price of the loan is the present value of the remaining payments at the current interest rate: an adjustment is made for the option to terminate the loan early. The current loan balance was estimated from the original balance and the amortization formula. The local regression model described in Clapp *et al.* (2001) estimated the value of the house and its standard deviation at each point in time. The house value was compared to the current loan balance and the normal distribution was used to estimate the probability of negative equity. This was converted to a 0/1 indicator if the probability was above 90%. Current loan-to-value: the estimated value of the house at each point in time was divided into the current loan balance estimated from the original face amount and the amortization formula. The appreciation variables were constructed as follows: house price appreciation in thousands of dollars is multiplied by an indicator of borrower age and by borrower age. The unemployment rate is for the county of residence in each quarter.
3. Other includes those outstanding at the end of the observation period.
4. We use standardized values for each continuous variable during the estimation, i.e. all continuous variables have zero mean and unit variance.

Table 3 Estimates for competing risks of mortgage terminations by refinance and move

	Model 1 MNL		Model 2 PHM		Model 3 MML		Model 4 MMH	
	Refi.	Move	Refi.	Move	Refi.	Move	Refi.	Move
Market value of loan	0.869 (9.88)	-0.165 (1.39)	1.033 (11.80)	-0.115 (0.86)	1.068 (5.68)	0.235 (0.83)	0.985 (11.26)	0.110 (0.71)
Original loan balance ($00,000)	0.344 (5.21)	-0.113 (1.30)	0.303 (4.04)	-0.104 (1.12)	0.365 (3.36)	-0.119 (0.70)	0.328 (4.33)	-0.227 (1.73)
15-year loan indicator	-0.007 (0.05)	-0.288 (1.77)	0.026 (0.17)	-0.303 (1.86)	-0.048 (0.23)	-0.612 (1.85)	0.003 (0.02)	0.094 (0.41)
Discount points (estimated)	-0.030 (0.48)	-0.200 (2.51)	0.015 (0.24)	-0.191 (2.41)	-0.060 (0.70)	-0.293 (1.84)	0.011 (0.17)	-0.350 (3.36)
Original refinance indicator	-0.569 (3.63)	0.221 (1.04)	-0.419 (2.79)	0.150 (0.64)	-0.576 (2.87)	0.284 (0.63)	-0.532 (3.55)	0.927 (2.90)
Current loan-to-value	-0.343 (3.43)	0.108 (1.01)	-0.359 (3.38)	0.082 (0.77)	-0.343 (2.50)	0.163 (0.73)	-0.383 (3.61)	0.310 (1.99)
Prob. Negative Equity > 90 percent indicator	0.204 (0.91)	-0.865 (2.53)	0.086 (0.49)	-0.626 (2.47)	0.026 (0.08)	-1.487 (2.25)	0.077 (0.44)	-1.574 (4.26)
House price appreciation (Age < 40, $,000)	0.000 (0.16)	0.000 (0.67)	0.002 (0.81)	0.004 (0.71)	0.000 (0.05)	0.000 (0.25)	0.001 (0.71)	0.001 (0.23)
House price appreciation (Age > 40, $,000)	-0.000 (1.94)	-0.000 (2.59)	-0.001 (0.75)	-0.005 (1.93)	-0.000 (1.68)	-0.000 (1.63)	-0.001 (0.82)	-0.004 (1.05)
Unemployment rate	-0.171 (2.50)	-0.094 (1.21)	-0.188 (2.62)	-0.076 (0.95)	-0.217 (2.20)	-0.244 (1.46)	-0.188 (2.63)	-0.025 (0.22)
Borrower age	-0.029 (0.41)	-0.160 (2.02)	-0.036 (0.48)	-0.185 (1.85)	-0.093 (0.94)	-0.397 (1.85)	-0.024 (0.31)	-0.901 (5.35)
Minority indicator	-0.670 (4.15)	-1.075 (4.97)	-0.636 (3.61)	-1.121 (4.51)	-1.46 (3.93)	-2.618 (5.00)	-0.582 (3.29)	-3.624 (7.94)
Borrower income	-0.144 (1.86)	0.163 (3.12)	-0.115 (1.22)	0.149 (2.15)	-0.132 (1.05)	0.266 (2.68)	-0.149 (1.52)	0.258 (2.83)
Obligation ratio	0.011 (0.18)	0.064 (0.90)	0.013 (0.21)	0.057 (0.76)	0.028 (0.33)	0.104 (0.69)	0.001 (0.02)	0.064 (0.56)
High credit score indicator	-0.169 (1.22)	0.253 (1.60)	-0.221 (1.51)	0.189 (1.15)	-0.139 (0.76)	0.265 (0.81)	-0.298 (2.04)	0.443 (1.86)
Low credit score indicator	-0.503 (1.99)	-0.177 (0.51)	-0.455 (1.81)	-0.161 (0.43)	-0.727 (2.07)	-0.957 (1.32)	-0.465 (1.83)	-0.732 (1.40)

Table 3 (Continued) Estimates for competing risks of mortgage terminations by refinance and move

	Model 1 MNL		Model 2 PHM		Model 3 MML		Model 4 MMH	
	Refi	Move	Refi	Move	Refi	Move	Refi	Move
Baseline Intercept (Group 1)					0.000	0.000	3.162	0.005
					(8.27)	(0.00)	(2.10)	(0.78)
Baseline Intercept (Group 2)					0.075	0.050	0.700	1.520
					(4.29)	(4.15)	(0.00)	(0.77)
Mass Point (Group 2)					0.017		0.117	
					(3.75)		(10.01)	
Log Likelihood	-3001.50		-2914.13		-2999.60		-2898.63	
A.I.C.	0.1576		0.1530		0.1576		0.1524	
B.I.C.	6171.85		5997.11		6199.71		5997.77	
McFadden Pseudo R-Square	0.1093		0.1190		0.1098		0.1237	

Notes:

1. The four models 1-4 are: multilnomial logit model (MNL), proportional hazard model (PHM), mass point mixed logit model (MML), mass point mixed hazard model (MMH). The data are structured as a panel dataset, with one observation for each quarter for each loan during the observation period.
2. The t-ratios are in parentheses. Refinance and move functions are considered as correlated competing risks and they are estimated jointly. Baseline functions for refinance and move are specified as Han-Hausman flexible functions in all four models. In Models 3 and 4, the mass point for group 1 was normalized into 1.0 during the estimation.
3. We estimated mass point mixed models using three mass points, with only one normalized to unity. These were dropped because they did not significantly improve the likelihood. They are available from the authors on request.

Table 4 Cross-model-validation for the four models

Refinance	Model 1 MNL	Model 2 PHM	Model 3 MML	Model 4 MMH
R-Square	0.001582	0.014051	0.008349	0.019346
Move	Model 1 MNL	Model 2 PHM	Model 3 MML	Model 4 MMH
R-Square	0.006818	0.000930	0.006398	0.014086

Notes:

1. The four models 1-4 are: multilnomial logit model (MNL), proportional hazard model (PHM), mass point mixed logit model (MML), mass point mixed hazard model (MMH).
2. The cross-model validation follows the out-of sample approach (See Appendix B). The full sample is split into two sub-samples: one for estimation, which has 90% of the full sample; the other for validation, which has the remaining 10% of the full sample. Then the estimates based on the 90% sub-sample are used to predict the 10% sub-sample. Finally, for the 10% sub-sample, we regress the real event on the predicted hazard rate for model 1, model 2, model 3 and model 4 respectively. The above table shows the R-squares of the four regressions, which indicate the goodness of fit of four different models.
3. Given that refinance and move are done separately during the validation, we have two panels in the table.

Table B.1 Maximum likelihood estimates of proportional hazard model (PHM) with alternative baseline functions

	Model 1		Model 2		Model 3	
	5th Order Poly		100% PSA		Han-Hausman Flex.	
	Refi.	Move	Refi.	Move	Refi.	Move
Market price of loan	0.746	-0.519	1.136	0.090	1.030	-0.135
	(8.89)	(4.09)	(17.09)	(1.01)	(12.75)	(1.03)
Original loan balance	0.310	-0.083	0.304	-0.115	0.323	-0.113
($00,000)	(4.27)	(0.93)	(4.32)	(1.24)	(4.36)	(1.21)
15-year loan indicator	-0.137	-0.374	-0.048	-0.368	0.043	-0.292
	(0.92)	(2.37)	(0.33)	(2.38)	(0.28)	(1.80)
Loan points (estimated)	-0.052	-0.331	0.113	-0.099	0.016	-0.193
	(0.89)	(4.39)	(1.92)	(1.45)	(0.25)	(2.45)
Original refinance	-0.964	-0.304	-0.609	-0.147	-0.414	0.142
indicator	(7.05)	(1.39)	(5.03)	(1.08)	(2.85)	(0.62)
Current loan-to-value	-0.442	0.025	-0.434	-0.024	-0.359	0.080
	(4.36)	(0.25)	(4.33)	(0.23)	(3.39)	(0.76)
Prob. Negative Equity >	0.012	-0.643	0.075	-0.622	0.076	-0.611
90 percent indicator	(0.07)	(2.58)	(0.44)	(2.54)	(0.44)	(2.41)
House price appreciation	0.001	0.003	0.002	0.005	0.002	0.004
(Age < 40, $,000)	(0.28)	(0.58)	(1.08)	(0.97)	(0.81)	(0.72)
House price appreciation	-0.002	-0.005	-0.000	-0.004	-0.001	-0.005
(Age > 40, $,000)	(1.15)	(2.28)	(0.16)	(1.45)	(0.76)	(1.90)
Unemployment rate	-0.221	-0.150	-0.187	-0.156	-0.179	-0.082
	(3.21)	(1.94)	(2.92)	(2.11)	(2.50)	(1.02)
Borrower age	-0.017	-0.143	-0.069	-0.175	-0.037	-0.167
	(0.24)	(1.49)	(0.94)	(1.84)	(0.49)	(1.67)
Minority indicator	-0.764	-1.141	-0.747	-1.133	-0.627	-1.081
	(4.43)	(4.77)	(4.48)	(4.75)	(3.59)	(4.36)
Borrower income	-0.132	0.153	-0.167	0.114	-0.146	0.147
	(1.49)	(2.30)	(1.77)	(1.66)	(1.51)	(2.12)
Obligation ratio	-0.020	0.046	-0.040	0.012	0.010	0.057
	(0.37)	(0.63)	(0.67)	(0.16)	(0.15)	(0.76)
High credit score indicator	-0.576	-0.003	-0.260	0.001	-0.220	0.189
	(4.28)	(0.02)	(2.20)	(0.01)	(1.57)	(1.16)
Low credit score indicator	-0.609	-0.290	-0.967	-0.662	-0.436	-0.144
	(2.47)	(0.82)	(4.15)	(2.01)	(1.73)	(0.39)
Log Likelihood	-3,062.27		-2,994.90		-2918.95	

Note:
T-ratios are in parentheses. Baseline functions are specified as 5th order polynomial functions, 100% PSA functions, and Han-Hausman flexible functions in Model 1, Model 2, and Model 3, respectively. Refinance and move functions are considered as correlated competing risks and they are estimated jointly. The data are structured as a panel dataset, with one observation for each quarter for each loan during the observation period.

Table B.2 Estimates for refinance and move models based on the 90% sample

	Model 1 MNL		Model 2 PHM		Model 3 MML		Model 4 MMH	
	Refi.	Move	Refi.	Move	Refi.	Move	Refi.	Move
Market value of loan	0.894	-0.165	1.062	-0.192	1.015	0.056	1.015	-0.053
	(9.50)	(1.39)	(11.94)	(1.40)	(4.65)	(0.18)	(10.47)	(0.32)
Original loan balance ($00,000)	0.333	-0.113	0.314	-0.100	0.342	-0.081	0.316	-0.194
	(4.79)	(1.30)	(3.99)	(1.00)	(2.87)	(0.48)	(3.94)	(1.41)
15-year loan indicator	-0.037	-0.288	0.004	-0.244	-0.052	-0.433	-0.041	0.255
	(0.23)	(1.77)	(0.02)	(1.41)	(0.25)	(1.36)	(0.25)	(1.04)
Loan points (estimated)	-0.032	-0.200	0.013	-0.247	-0.078	-0.373	0.009	-0.495
	(0.48)	(2.51)	(0.19)	(3.02)	(0.86)	(2.45)	(0.14)	(4.34)
Original refinance indicator	-0.528	0.221	-0.363	0.030	-0.559	0.216	-0.481	0.739
	(3.11)	(1.04)	(2.33)	(0.12)	(2.60)	(0.46)	(2.92)	(2.14)
Current loan-to-value	-0.420	0.108	-0.457	0.078	-0.428	0.154	-0.481	0.371
	(3.91)	(1.01)	(3.88)	(0.70)	(2.83)	(0.69)	(4.07)	(2.30)
Prob. Negative Equity > 90 percent indicator	0.183	-0.865	0.136	-0.554	0.080	-1.401	0.140	-1.541
	(0.75)	(2.53)	(0.72)	(2.04)	(0.23)	(1.97)	(0.73)	(3.71)
House price appreciation (Age < 40, $,000)	0.000	0.000	0.001	0.004	0.000	0.000	0.001	0.004
	(0.04)	(0.67)	(0.56)	(0.90)	(0.035)	(0.38)	(0.50)	(0.66)
House price appreciation (Age > 40, $,000)	-0.000	-0.000	-0.001	-0.004	-0.000	-0.000	-0.001	-0.003
	(1.60)	(2.59)	(0.51)	(1.30)	(1.22)	(1.49)	(0.54)	(0.59)
Unemployment rate	-0.156	-0.094	-0.174	-0.068	-0.181	-0.183	-0.186	0.071
	(2.13)	(1.21)	(2.27)	(0.80)	(1.77)	(1.15)	(2.43)	(0.61)
Borrower age	-0.061	-0.160	-0.074	-0.124	-0.111	-0.285	-0.059	-0.853
	(0.81)	(2.02)	(0.89)	(1.18)	(1.04)	(1.37)	(0.70)	(4.94)
Minority indicator	-0.670	-1.075	-0.597	-1.035	-1.210	-2.071	-0.555	-3.538
	(3.75)	(4.97)	(3.22)	(3.97)	(3.01)	(3.83)	(2.96)	(7.16)
Borrower income	-0.140	0.163	-0.144	0.130	-0.121	0.244	-0.145	0.236
	(1.73)	(3.12)	(1.39)	(1.69)	(0.93)	(2.35)	(1.36)	(2.61)
Obligation ratio	0.029	0.064	0.029	0.042	0.048	0.079	0.021	0.023
	(0.45)	(0.90)	(0.44)	(0.53)	(0.55)	(0.54)	(0.32)	(0.19)
High credit score indicator	-0.048	0.253	-0.083	0.227	0.038	0.479	-0.173	0.753
	(0.31)	(1.60)	(0.55)	(1.30)	(0.19)	(1.54)	(1.07)	(2.98)
Low credit score indicator	-0.414	-0.177	-0.361	-0.107	-0.527	-0.577	-0.413	0.111
	(1.53)	(0.51)	(1.31)	(0.27)	(1.49)	(0.83)	(1.49)	(0.20)

Table B.2 (Continued) Estimates for refinance and move models based on the 90% sample

	Model 1		Model 2		Model 3		Model 4	
	MNL		PHM		MML		MMH	
	Refi	Move	Refi	Move	Refi	Move	Refi	Move
Baseline Intercept (Group 1)					0.000	0.000	2.953	0.001
					(6.33)	(0.00)	(2.22)	(0.76)
Baseline Intercept (Group 2)					0.035	0.019	0.700	0.455
					(4.39)	(4.08)	(0.00)	(0.75)
Mass Point (Group 2)					0.018		0.117	
					(2.27)		(9.65)	
Log Likelihood	-2647.55		-2576.60		-2647.06		-2559.03	

Notes:

1. The four models 1-4 are: Multilnomial logit model (MNL), proportional hazard model (PHM), mass point mixed logit model (MML), mass point mixed hazard model (MMH). The data are structured as a panel dataset, with one observation for each quarter for each loan during the observation period.
2. The t-ratios are in parentheses. Refinance and move functions are considered as correlated competing risks and they are estimated jointly. Baseline functions for refinance and move are specified as Han-Hausman flexible functions in all four models. In Model 3 and 4, Mass point for group 1 was normalized into 1.0 during the estimation respectively.