

An empirical characterization of mortgage default in Colombia between 1997 and 2004

Juan Esteban Carranza

Dairo Estrada*

Abstract

This paper examines the relationship between mortgage default decisions and relevant observable variables in Colombia between 1997 and 2004. We estimate a discrete choice model of default using a panel of individual mortgage characteristics and payment information. We show that home prices and debt balances are the main determinants of mortgage default. More interestingly, we also show that such correlation can only be uncovered after controlling for the unobserved heterogeneity of debtors, which we do using Census data and simulation techniques.

Keywords: Mortgage default; Maximum simulated likelihood; Colombia

*University of Wisconsin and Banco de la República-Bogotá (Colombia), respectively. We are grateful to Oscar Leiva at Titularizadora Colombiana for providing the primary data set and to staff at Banco de la República in Cali and Bogotá for additional data support. We also thank Julio Escobar, Jean-Francois Houde, Salvador Navarro and seminar participants at the University of Wisconsin for their comments and insights. The corresponding author is Carranza: juanes@ssc.wisc.edu

1 Introduction

During the late 1990's the Colombian economy, similarly to several other emerging economies, experienced a severe financial crisis and economic slowdown. The effects of such crisis were fuelled by a dramatic increase in the default rates of mortgage holders, leading to the collapse of several major financial institutions and a major crisis that persisted for years.

During the crisis, the behavior of debtors was affected by several separate factors: on one hand, incomes fell, making it difficult for many households to fulfill their payment obligations; on the other hand, debt balances which were tied to a market interest rate increased as the monetary authority stepped in to contain the fall of the exchange rate. Simultaneously, and as the crisis ensued, the prices of real estate, which had risen to unprecedented levels since the mid 1990's, fell dramatically.

We use a random data set containing information on the basic characteristics and payment histories of individual mortgages to evaluate the importance of different factors in the determination of mortgage default. The raw correlations contained in the basic data set show that there is a positive relationship between home prices and mortgage default, which is counterintuitive and inconsistent with the conventional notion that default is more prevalent when home prices fall.

The source of this misleading correlation is the lack of information about the characteristics of debtors, specially income, that might be correlated with both default behavior and home prices. The presence of these unobserved correlated state variables creates an endogeneity problem. We correct the problem constructing a model of optimal default that accounts for the the unobserved heterogeneity of debtors. We control for such heterogeneity using non-matching census data and estimate the model using a method of

maximum simulated likelihood.

The results of the estimation indicate that home prices have a strong negative effect on mortgage default. The effect of the unobserved income variation is statistically significant but economically insignificant. Its inclusion in the model is nevertheless crucial for uncovering the true effect of home prices.

Given the behavioral model on which the estimation is based, the obtained estimates correspond to underlying structural parameters and can therefore be used for counterfactual computations. We use the estimates to predict default probabilities and show how sensitive the default behavior is to changes in observed and unobserved states.

The sharp identification of the model parameters relies partly on some singular institutional features of the Colombian mortgage market during the time span of the sample. First, the terms of the individual mortgages were not negotiated between debtors and financial institutions. In general, mortgage terms were negotiated between developers or construction companies and the mortgage banks, and the terms of the mortgage were transferred to any house buyers, who qualified according to a simple income rule. Second, while mortgage payments were based on a fixed rate on the balance of the loan, the balance was indexed to a market interest rate according to a predetermined formula set by the Central Bank¹. Therefore, the data contains enough exogenous variation to identify a detailed structural model.

The empirical literature on mortgage default is dominated by option mod-

¹Specifically, payments were made according to a fixed rate over the balance of the loan. But then the balance of the loan increased month by month according to a rate fixed by the Central Bank; it used to be that this rate of increase was tied to the rate of inflation but since the early 1990's it was tied to a market interest rate. In any case, the rate was set arbitrarily by the Central Bank.

els that are estimated using variations of duration models, as described in the influential paper by Deng, Quigley and Van Order et al (2002). In our case, we use a discrete choice framework, close in spirit to the empirical IO or labor literature. From an empirical point of view the advantage of duration models over models based on individual likelihood functions is that they can be estimated even if the default rates are very low. This is not an issue in our case, because default rates in Colombia during the time span of the sample were very high and allow a very precise estimation of the likelihood-based model. The use of structural microeconomic techniques to study the housing market is not common. It includes work on housing preferences by Bajari and Kahn (2005), Epple, Romer and Sieg (2001) and Bayer, Ferreira and McMillan (2007). The maximum simulated likelihood estimators are described in, for example, Train (2003).

The rest of the paper is organized as follows: in the following section we present a model of optimal default that generates default probabilities that can be used to construct a likelihood-based estimator. Then we discuss the data and the estimation. We present additional results and finalize with a discussion of the limitations of our approach and further research.

2 A behavioral model of mortgage default and the estimation strategy

The model below is to be implemented with data from the Colombian mortgage market between 1997 and 2004. The most salient feature of the Colombian mortgage financing system in those days was that all mortgages had variable rates tied to the market deposit rate, through a formula that was determined by the Central Bank. Therefore, refinancing was not a real op-

tion for most mortgage holders. The (total or partial) prepayment option will be ignored, as prepayments did not seem to be empirically significant (relative to defaulting) during the time-span of the sample and have less social implications than default².

We study the behavior of mortgage holders (“debtors”) who live in the mortgaged piece of real estate (“home”). The utility that a debtor i gets from the home depends on a measure x_i of subjective home quality. It also depends on the difference between household income and mortgage payments $Y_{it} - R_{it}$ and an idiosyncratic preference shock ε_{it} which incorporates unobserved (to the econometrician) variables that affect default, e.g. home attributes that are only valued by its owner and other preference shocks that vary across consumers and time. The estimation of the model is based on the properties of these unobserved variables.

A debtor chooses to default on her mortgage if the utility of making the mortgage payments and staying in her home is lower than the utility generated by not making the loan payment at the time. This alternative, which we will broadly call “default”, gives rise to a complex scenario. Specifically, the individual may just be waiting to see whether the following period she can pay back her dues; she may try to sell the home and cash the difference between price and loan balance; she may return it to the bank to cover her obligation; finally, she could also just stop making payments indefinitely and face forfeiture or a renegotiation with the bank.

We write the decision problem of debtor i at time t as follows:

$$\max\{u_i(x_i, Y_{it} - R_{it}, \varepsilon_{it}) + V_{it+1}, W_{it}\} \quad (1)$$

where $u(\cdot)$ is the instant payoff from “consuming” the home at period t

²From the perspective of the lending institutions, prepayments are quite relevant. The discussed methodological framework can incorporate prepayments easily.

and V_{it+1} is the expected discounted continuation value for the debtor of delaying the default decision one period³. W_{it} is the value of default which is the weighted sum of payoffs across the complex set of random scenarios discussed above.

In general, the continuation payoff V_{it+1} would depend on the expected evolution of the state variables and could be computed taking advantage of its Bellman representation. We consider a model in which debtors are myopic in the sense that they ignore the dynamics of the state variables. Specifically, we assume that the difference $\tilde{W}_{it} = W_{it} - V_{it+1}$, which we call the net payoff of default, can be described using a reduced form.

Let the static utility be additively separable on observable and unobservable states:

$$u_i(x_i, Y_{it} - R_{it}, \varepsilon_{it}) = \theta_i + \gamma_i x_i + \alpha_i(Y_{it} - R_{it}) + \varepsilon_{0it} \quad (2)$$

where the coefficients of the utility function are, in principle, allowed to be heterogeneous across debtors. A debtor i chooses to continue paying her dues if the utility of doing so is higher than the utility of default, as follows:

$$\theta_i + \gamma_i x_i + \alpha_i(Y_{it} - R_{it}) + \varepsilon_{0it} \geq \tilde{W}_{it} + \varepsilon_{1it} \quad (3)$$

Let $N_{it} = 1$ be the event that debtor i does *not* default at time t . The individual probability of defaulting is the probability that (3) is true. By specifying a parametric distribution for ε we can obtain the individual choice probabilities:

$$\begin{aligned} Prob[N_{it} = 1] &= Prob[\theta_i + \gamma_i x_i + \alpha_i(Y_{it} - R_{it}) + \varepsilon_{0it} \geq \tilde{W}_{it} + \varepsilon_{1it}] \\ &= Prob[\theta_i + \gamma_i x_i + \alpha_i(Y_{it} - R_{it}) - \tilde{W}_{it} \geq \bar{\varepsilon}_{it}] \end{aligned} \quad (4)$$

³The difference between the continuation value and the value of default is the value of the non-default option.

We assume that the errors ε_{it} are *iid* draws from an extreme value distribution, so that the choice probabilities (4) have an analytical solution given by the usual logit form⁴:

$$Prob[N_{it}] = \frac{e^{\theta_i + \gamma_i x_i + \alpha_i (Y_{it} - R_{it}) - \tilde{W}_{it}}}{1 + e^{\theta_{i0} + \gamma_i x_i + \alpha_i (Y_{it} - R_{it}) - \tilde{W}_{it}}} \quad (5)$$

We parameterize the model and use the choice probabilities (5) to estimate it using a likelihood-based technique. There are two difficulties associated with such an approach. First, it requires matching data of all observable states at the micro level –specially it requires matching data on individual income over time, which is something that we don’t have. Second, it requires that we allow for unobserved “taste” heterogeneity –in particular, the empirical literature on mortgage default points out that debtors have a heterogeneous risk aversion.

In order to control for the unobserved income variation, we incorporate non-matching survey data into the estimation by integrating the predicted individual default probabilities over the empirical distribution of income, conditional on the observed states. In addition, the unobserved tastes are integrated out using a parametric assumption about their distribution. The model is then estimated using the method of maximum simulated likelihood.

3 The data

The model above is estimated with two separate non-matching panel data sets. The first (or “main”) data set contains information on 16000 random mortgages that were outstanding between 1997 and 2002. The monthly payment history of each mortgage, its original and current value and term of

⁴Other parametric distributions can also be adopted. For example, if ε are assumed to be standard normal, a standard probit model ensues.

the mortgaged home are included. On the other hand, the expected prices of individual homes at any point in time \bar{P}_{it} are computed using housing price indices constructed by the Colombian Central Bank. All data is aggregated into quarters, so that default observations are not confounded with missed payments or coding errors.

Since this main data set contains no information on the income of debtors over the span of the sample, survey data were collected with information on the joint distribution of households income and mortgage holdings (the “auxiliary” data set). Specifically, annual surveys conducted by the Colombian national statistics agency (DANE) contain large samples of individual household incomes and matching housing payments that can be used to simulate the time-varying joint distribution of income and the other state variables.

Table 1 contains some summary statistics of the main data set, which goes from the second quarter of 1997 to the second quarter of 2004⁵. Notice that the number of loans in the data set changes over time as loans are paid off completely or new loans start; this number fluctuates roughly between 5000 and 8000. Columns (3) and (4) of the table contain the percentage of loans in the data at each point in time with more than 3 and 6 months of past due payments, illustrating the dramatic prevalence of default during the crisis. After 2000 and until the end of the data set, more than 20% of all loans in the data set had past due payments of more than 3 months reaching 23% in the second quarter of 2003. The percentage of loans with past due payments of more than 6 months reaches its peak of more than 16% in the first quarter of 2003.

In the data it is observed that sometimes debtors temporarily stop making

⁵Since default is inferred from the change in the number of past due mortgage payments, the first observation in the first quarter of 1997 is dropped from the data set.

their payments. Therefore what ‘default’ means has to be defined. Specifically in the estimation below, loans that accumulate past due monthly payments of more than 3 months are assumed to be defaulted and are dropped off from the data set. Therefore, default is defined as the event in which the number of past due monthly payments in a loan history changes from 3 or less to more than 3 between two quarters. After a loan is defined to be defaulted, it is dropped off the sample⁶.

The default rate based on this definition (i.e. the number of defaults over the total number of outstanding loans) is displayed in column (5) of the table. This rate reached its peak of more than 6% in the midst of the financial crisis, during the earlier quarters of 2000. Notice, though, that this rate is generally decreasing over the time span of the sample, due to the fact that defaulted loans are dropped from the sample. A dramatic reflection of the depth of the crisis in these years is the fact that by the end of the sample debtors had defaulted on more than 80% of the loans included in this random sample according to our definition of default.

There is no direct information on the size of the required monthly payments. It is known though that they were directly tied to the rate of mortgage balance over the remaining term of the loan. The balance of the loan was itself tied to the market interest rate through a formula established by the Central Bank. Column (6) of Table 1 contains the average *real* value of the ratio of mortgage balance to remaining term among outstanding loans. Notice that this value increases throughout the whole span of the sample. This increase might have been partly driven by the price of new homes, specially

⁶The default rate based on this definition is highly correlated with default rates based on longer default periods. The 3-month threshold was chosen in order to observe as much default as possible and in order to capture *all* defaulted loans, including those that are terminated soon after default.

before 2000. Nevertheless, as can be seen in column (7) the real value of the homes in the sample is decreasing throughout the whole time span of the sample, in particular after 2000.

Table 2 characterizes the raw correlations contained in the data. Specifically, a linear probability model of non-default was estimated using the definition of default described above. The right hand side variables are the mortgage balance, the expected price of the collateral and the remaining term of the loan at each point in time. As expected, default is positively correlated with the balance of mortgages at any point in time and with their remaining term.

The crucial feature of the main data set is that it suggests that non-default is negatively correlated with the expected price of the collateral⁷. This correlation is counterintuitive in the sense that we would expect higher home prices to be associated with *less* default, or equivalently *more* non-default. Most likely, this correlation is induced by the presence of unobserved states that are correlated with prices and default. In particular, the variation of the unobserved income of individual debtors may be driving the results of the regressions. The last two columns of Table 2 contain results of the model including fixed time-effects that capture the component of the unobserved states that is common to all debtors. Notice that the magnitude and statistical significance of the correlations do not change much after the inclusion of the fixed time-effects, which suggests that the unobserved component of the error that is common to all debtors is not correlated with the observed variables included in the regression.

The estimates of the time effects, which are measured with respect to the

⁷Notice that the table reports the estimates of a regression of *non-default* on covariates, which is consistent with the specification of the structural model below.

constant in the second quarter of 1997, are mostly significant. The coefficient of correlation of these estimates and the average income of mortgage holders in the secondary data set is 0.41, consistent with the presumption that the time effects are capturing a lot of the common variation in household income.

The literature on mortgage default (e.g. Deng et al, 2003) has documented the fact that the initial loan to value (LTV) ratio of loans is correlated with the risk attitude of debtors who select themselves into different mortgage contracts. As seen on the left hand side columns of Table 2, there is a significant negative correlation between default and the initial LTV ratio, controlling for current home values and mortgage balances. As seen on the two right hand side columns of the table, this correlation is insignificant once we control for the time effects. This implies that the sharp correlation detected in the first set of estimates is not strong within time periods. It is suggestive, though, of the importance of debtor heterogeneity to explain observed default behavior.

4 Estimation

4.1 The empirical model

As indicated above, the data sets contain no information on the characteristics of the individual homes. Therefore, it is assumed that the unobserved “quality” of homes x_i is random:

$$x_i \equiv \kappa + \varepsilon_{it}^x \tag{6}$$

where ε_{it}^x is a random error that is potentially correlated over time and across debtors.

There is no information on the required monthly payments R_{it} of each

debtor. It is known, though, that payments are linear functions of mortgage balances K_{it} and remaining term L_{it} , with some random variation across debtors:

$$R_{it} = \rho_0 + \rho_1 K_{it} + \rho_2 L_{it} + \varepsilon_{it}^r \quad (7)$$

where ε_{it}^r is an error term.

It is also assumed that the payoff of default $\tilde{W}_{it}(\cdot)$ is a linear function of relevant states:

$$\tilde{W}_{it} = \omega_0 + \omega_1 Y_{it} + \omega_2 \bar{P}_{it} + \omega_3 K_{it} + \varepsilon_{it}^w \quad (8)$$

where ε_{it}^w is the structural error. Recall that this payoff is a linear combination of payoffs across random outcomes payoffs net of the continuation value; if these payoffs are linear functions of states, then the linear payoff function $\tilde{W}_{it}(\cdot)$ should be stable across counterfactual equilibria, as long as states don't affect the probabilities of individual outcomes. Notice that a careful interpretation of the function \tilde{W}_{it} is important because the usefulness of the model for counterfactual analysis relies on the assumption that this function will not change when we change the values or the transition probabilities of the state variables.

Substituting (6), (7) and (8) in condition (3), the non-default condition for debtor i at time t can be obtained:

$$\begin{aligned} \theta_0 + \gamma(\kappa + \varepsilon_{it}^x) + \alpha_i(Y_{it} - (\rho_0 + \rho_1 K_{it} + \rho_2 L_{it} + \varepsilon_{it}^r)) + \varepsilon_{it}^u \\ \geq \omega_0 + \omega_1 Y_{it} + \omega_2 \bar{P}_{it} + \omega_3 K_{it} + \varepsilon_{it}^w \end{aligned} \quad (9)$$

After grouping terms, the condition (9) can be rewritten as:

$$\zeta_0 + \zeta_1 \bar{P}_{it} + \zeta_2 Y_{it} + \zeta_3 K_{it} + \zeta_4 L_{it} + \bar{\varepsilon}_{it} \geq 0 \quad (10)$$

Therefore the non-default probability depends on the distribution of the error term $\bar{\varepsilon}_{it} \equiv \gamma \varepsilon_i^x - \alpha \varepsilon_{it}^r + \varepsilon_{it}^u - \varepsilon_{it}^w$. In order to allow a rich correlation across

choices we consider variations of the model in which the error is decomposed as follows:

$$\bar{\epsilon}_{it} = \xi_t + \mu_i + \epsilon_{it} \quad (11)$$

where the term μ_i is an individual-specific unobservable state and ϵ_{it} is the *iid* extreme value error. This specification allows individual choices to be correlated over time and across debtors; in addition, this unobserved heterogeneity can be allowed to depend on other observed states such as income which would be equivalent to a model with heterogenous ζ coefficients.

Given the extreme value assumption, the individual non-default probability (5) is given by:

$$Prob(N_{it} = 1) = \frac{e^{\zeta_0 + \zeta_1 \bar{P}_{it} + \zeta_2 Y_{it} + \zeta_3 K_{it} + \zeta_4 L_{it} + \xi_t + \mu_i}}{1 + e^{\zeta_0 + \zeta_1 \bar{P}_{it} + \zeta_2 Y_{it} + \zeta_3 K_{it} + \zeta_4 L_{it} + \xi_t + \mu_i}} \quad (12)$$

Again, $N_{it} = 1$ stands for the event of debtor i not defaulting on her mortgage at time t . Estimating the parameters ζ of the model above requires the maximization of the sample non-default likelihood predicted by the model. This likelihood is computed by multiplying the likelihood of observed choices across debtors and over time as follows:

$$L(\zeta) = \prod_{i \in S_t} \prod_{t \in T} (Prob(N_{it} = 1))^{N_{it}} (Prob(N_{it} = 1) - 1)^{1 - N_{it}} \quad (13)$$

where S_t is the random set of loans that are outstanding at time t .

The likelihood (13) cannot be computed directly due to the unavailability of matching income data. The non-matching income data from household surveys can be incorporated into the estimation above by integrating the likelihood over the empirical joint distribution of income and mortgage payments. Notice that the individual unobserved effects can also be incorporated into the estimation by assuming that they come from a known parametric distribution and integrating them out throughout the estimation.

Specifically, if we assume that the individual effects μ_i are distributed according to some known parametric distribution $\Phi(\sigma_\mu)$, the “expected” non default probability is:

$$\hat{Pr}ob[N_{it} = 1] = \int \frac{e^{\zeta_0 + \zeta_1 \bar{P}_{it} + \zeta_2 Y + \zeta_3 K_{it} + \zeta_4 L_{it} + \xi_t + \mu}}{1 + e^{\zeta_0 + \zeta_1 \bar{P}_{it} + \zeta_2 Y + \zeta_3 K_{it} + \zeta_4 L_{it} + \xi_t + \mu}} dG_t(Y | K) d\Phi(\sigma_\mu) \quad (14)$$

where $\xi = \{\xi_{t=1\dots T}\}$ is treated as a vector of fixed time-effects that can be estimated for each t . $G(. | K)$ is the empirical distribution of household income at time t , conditional on mortgage balances, which can be inferred from the survey data.

Given any set of parameters $\{\zeta, \xi, \sigma_\mu\}$ the probabilities above can be obtained via simulation and the simulated sample likelihood can be computed just like in (13) above:

$$\hat{L}(\zeta, \xi, \sigma_\mu) = \prod_{i \in S_i} \prod_{t \in T} (\hat{Pr}ob(N_{it} = 1))^{N_{it}} (\hat{Pr}ob(N_{it} = 1) - 1)^{1 - N_{it}} \quad (15)$$

4.2 Results

To estimate the model, the simulated likelihood (15) was maximized by computing the predicted probabilities (14) using simulation. The first issue to be addressed is the specification of the unobserved debtor heterogeneity. Following the previous empirical literature on mortgage default (e.g. Deng et al, 2002), debtor heterogeneity will be tied to the initial “loan to value” $LTV_i = K_{i0}/P_{i0}$ of the mortgage, where $t = 0$ stands for the moment at which the loan was first started. This specification presumes that debtors select themselves into mortgages with different LTV according to their attitude towards risk.

Accordingly, it is assumed that $\bar{\epsilon}_{it} = \xi_t + \sigma_\mu LTV_i \bar{\mu}_i + \epsilon_{it}$, so that the unobserved component of utility has a common element ξ_t that varies over time,

a consumer-specific component $\sigma_\mu LTV_i \mu_i$ and an extreme value consumer- and time-specific shock ϵ_{it} .

The consumer-specific shock $\sigma_\mu LTV_i \bar{\mu}_i$ is assumed to be correlated with the initial leverage of the mortgage, so that its distribution can be separated from the distribution of the idiosyncratic shock. Specifically, $\bar{\mu}_i$ is assumed to be a standard normal error, so that the consumer-specific error is normal with zero mean and variance $\sigma_\mu^2 LTV_i^2$. Higher absolute realizations of this unobserved error are associated with a higher initial LTV and are a consumer specific constant that shifts the individual utility function.

On the other hand, the common component of the error $\xi_{t=1,\dots,T}$ was estimated as a fixed time effect. Therefore, for any set of parameters $\{\zeta, \xi, \sigma_\mu\}$, a consistent estimator of such integral is given by:

$$Pr\hat{ob}[N_{it} = 1] = \frac{1}{J} \sum_{j=1}^J \frac{e^{\zeta_0 + \zeta_K K_{it} + \zeta_P \bar{P}_{it} + \zeta_L L_{it} + \zeta_Y Y_j + \xi_t + \sigma_\mu \bar{P}_{it} \bar{\mu}_i}}{1 + e^{\zeta_0 + \zeta_K K_{it} + \zeta_P \bar{P}_{it} + \zeta_L L_{it} + \zeta_Y Y_j + \xi_t + \sigma_\mu LTV_i \bar{\mu}_i}} \quad (16)$$

where μ_i are independent standard normal draws and Y_j are income draws taken from the empirical distribution of income, conditioned on housing payments, contained in yearly surveys. The average is taken over J simulations.

Specifically, the auxiliary survey data contains random observations of households' income and mortgage payments of homeowners⁸, while the main data set contains information on the balances and maturities of outstanding mortgages. It is assumed that the distribution of monthly mortgage payments is the same as the distribution of balances over the remaining maturity of mortgages K_{it}/L_{it} . Therefore, the quantiles of the distribution of K_{it}/L_{it} in the main data set correspond to the quantiles of the distribution of income for the households that are making mortgage payments. When computing (16),

⁸More precisely, the surveys ask whether people are financing the home they live in and how much they pay every month.

draws of $\{\bar{P}_{it}, K_{it}, L_{it}\}$ are therefore matched with random draws of Y_{it} from the same conditional distribution quantile. Given that surveys with mortgage payment data are only available at the yearly level, the distribution of income conditional on mortgage payments is interpolated to remaining quarters, by assuming that the income distribution was constant within years.

Four versions of the model were estimated with results reported in Table 3. In model 1 it is assumed that $\mu = \xi = 0$; in model 2 $\mu = 0$ and in model 3 $\xi = 0$. Model 4 is the full model as in (17). The displayed results of models 1 and 3 were obtained simulating 20 income draws for each observation from the corresponding income quantile in the auxiliary data set. Home prices, mortgage balances and income are measured in tens of millions of constant 1997 Colombian pesos. Due to the size of the involved matrices, models 2 and 4 were estimated using 10 income draws per observation; in addition, a random subsample of 1/4 of the simulated sample was taken to alleviate computer memory restrictions. To give an idea of the computational magnitude of the estimation, the size of the matrix of regressors after subsampling in model 4 was (572600x34). The reported standard errors were obtained using the standard formula and are robust to alternative specifications. Table 3 also reports the computed average marginal effects of each variable on the non-default probabilities.

The results have two salient features. The first is that, opposed to the raw correlations documented in section 3, there is a very significant and positive coefficient of home prices. That is, only after controlling for the unobserved heterogeneity of debtors can we uncover a positive price coefficient in the non-default probability. This coefficient is not only statistically significant but also economically significant. The reported marginal effects imply that a decrease in the price of home of COL\$10 million in 1998, which is less than the

average loss of housing values in the sample between 1998 and 1999, induces an increase in the default probability of around 0.2%. This magnitude is not insignificant, given the magnitude of the default rates, which is between 1% and 6%.

The second salient feature of the results is the income coefficients which, not surprisingly, imply that higher income increases non-default. Such effect is statistically significant but, perhaps surprisingly, economically insignificant. As indicated in table 3, an increase in monthly income of COL\$10 million in 1998 induces an increase in the non-default probability of 0.2%, comparable to the marginal effect of home price discussed above. This number is far above the mean monthly income of individuals in our Census sample of only COL\$0.6 million. To give some perspective, an increase in monthly income of COL\$10 million would have been equivalent to an increase in yearly income of around US\$100.000. The income coefficients are statistically significant in models 1 and 3. In models 2 and 4, the high standard errors are presumably due to the lower number of income draws used in its estimation.

Finally it should be pointed out that higher balances and longer remaining terms induce lower non-default probabilities and the associated marginal effects are similar to the results of the linear probability model. The results of the estimation are robust to the inclusion of fixed time-effects in models 2 and 4 and to the inclusion in models 3 and 4 of a normal unobserved heterogeneity that is correlated with the initial LTV and, at least indirectly, with current realizations of K_i and P_i . As indicated above, the addition in models 3 and 4 of the normal error that is correlated with LTV_i aims to capture the fact that even conditional on the unobservables, default rates might vary across consumer types who select themselves into homes with different prices.

Notice that the coefficient of the individual-specific error is significant and negative, which means that default rates of debtors are negatively correlated across initial *LTV* values due to underlying heterogeneity of preferences. The estimates of the fixed time-effects, which account for unobserved aggregate shocks that are not correlated with observed variables are large. In other words, the observed variables in the model cannot account for a big portion of the time variation of default behavior. This is not surprising given that most of the variation on which the identification of the model is based is cross sectional and dynamic effects were not modelled. The results are nevertheless reassuring of the econometric validity of the obtained estimates, in the sense that the the observed states variables do not appear to be correlated with the unobservables.

The main conclusion of the results above is that household income variation was not the driving force behind the dramatic increases in mortgage defaults during the late 1990's. This is consistent with a rational model of optimal default in which households default decisions are mainly based on the value of the involved assets, but may defy conventional wisdom. The variation in the price of the collateral and the increases in the size of the mortgage balances seemed to have played a more important role. In the following section a more precise evaluation of these impacts is performed.

4.3 Fit of the model and additional results

As seen in figure 1, the model can trace satisfactorily the aggregate default rates. Aggregated over time, the default rate reached 86%, which is a dramatic figure; it means that 86% of the household in this random sample

accumulated past due mortgage payments for more than 3 months⁹. The difference between this observed overall default rate and the rate predicted by the model is of less than 0.5%. Much of the time variation of default behavior is driven by the unobserved aggregate shock, equivalent to a time-changing model constant. In this sense, the model is better suited to understand the variation in default across debtors at any point in time.

In order to isolate more clearly the impact of individual factors on default probabilities, the model can be used to compute them directly. Table 4 contains default probabilities for an “average” debtor with different values of the observed states, keeping the other values at the average value they had at the beginning of the sample¹⁰. The predicted probabilities are computed for values of L , P and K that lie at the center of the quintile of their distribution.

Notice that predicted default probabilities for this “average” debtor are very sensitive to changes in the observed states. For example, at any point in time increasing the number of remaining periods of mortgage maturity, L , from the center of the lower quintile of the distribution of L ($L = 27$) to the center of the upper quintile of the distribution of L ($L = 57$) increases the predicted default probability by around 50%. Keeping everything else constant, increasing the balance of the mortgage K from the center of the lower quintile ($K = 0.36$) to the center of the upper quintile of its distribution ($K = 3.8$) increases the predicted default probabilities at any point in time by more than 60%. The effect of price is just as significant: increasing the

⁹If instead of defining default as the accumulation of 3 or more months of past due payments we had used a threshold of 6 months, the accumulated default rate over the time span of the sample would have been 49%, which is still a staggering figure.

¹⁰Specifically, default probabilities were computed for $\bar{K} = 1.614$, $\bar{P} = 4.493$, $\bar{L} = 42$ and $\bar{Y} = 0.0865$ which were the average values of these states in the 3rd quarter of 1997; recall that K , P and Y are measured in tens of millions of constant 1997 Colombian pesos.

price from the lower quintile ($P = 1.3$) to the upper quintile ($P = 10.8$) of its distribution decreases the predicted default probability by more than 50%.

Finally, we can use the simulated data to infer the default behavior of debtors across the distribution of income. This is important, because income is the most important individual random state that affects default behavior and that is not directly observed by banks or policy makers. As indicated above, the estimated effect of income on default is negligible, keeping fixed the other states. Nevertheless, income is correlated in the data with unobserved states that do affect default.

Figure 2 illustrates the percent difference in the default rates between debtors in the upper 20% and debtors in the lower 20% of the income distribution at each point in time as implied by the simulated data. It can be seen that, perhaps surprisingly, the inferred default rate of wealthier households is consistently higher than the predicted default rates of poorer households, despite the fact that income has a somewhat negative effect on the probability of default. This difference is almost 15% at the beginning of the sample period and tends to disappear over time as the pool of debtors shrinks. In fact, the predicted aggregate rate of default is 90% for debtors in the upper tail of the income distribution and 84% for debtors in the lower tail. For debtors located around the median of the distribution this rate was 86%. Notice that this inference is based only on the direct examination of the simulated data and doesn't rely on the estimates of the model.

We have already pointed out that income itself has no significant direct effect on the probability of default, which is mostly affected by home prices, mortgage term and balances. We conclude that this difference in default rates across income levels must have been induced by a disproportionate concentration of mortgages with long maturities in the hands of relatively wealthier

debtors. In other words, higher income households acquired a disproportionately big share of new mortgages during the 1990's in Colombia and then defaulted on their mortgages at higher rates than poorer households. This, in turn, is a reflection of the credit boom that preceded the time span of the sample.

As indicated above, the model is not very well suited for predicting the variation of default probabilities over time, as much of it is explained by an unobserved aggregate shock which was estimated taking advantage of the panel structure of the data. The estimation results implied that this aggregate “error” is not correlated with the observed states. It is difficult to argue that these shocks are random, but it is also difficult to infer from the data what drives their evolution. Notice, though, that the estimated time-effects isolate the effects of aggregate variables on default. Therefore, they can potentially be used to construct a statistical model relating aggregate observed shocks to default, which is beyond the scope of this study¹¹.

5 Concluding remarks and further research

We have estimated a model of optimal mortgage default and the results indicate that default in Colombia between 1997 and 2004 was driven mostly by housing prices and mortgage balances. Moreover, income variation, though statistically significant, had a relatively small effect on the default probabilities.

These results are consistent with the notion that current income, if anything, only affects the marginal utility of housing consumption and should

¹¹For example, these aggregate shocks seem to be somewhat correlated with the interest rates and indicators of economic activity. Establishing any causality is difficult and requires further research.

have a small effect on default behavior. The results therefore imply that income support to indebted mortgage holders would have little effect on the likelihood of default, as long as the price of houses is unaffected.

Nevertheless, we showed that in order to uncover the underlying relationship between housing prices and default behavior, it was crucial to account for the unobserved variation in debtors' income, which we did using Census data and simulation techniques. Moreover, it is found that at any point in time and given the observed distribution of states, default rates were higher for higher income households.

We also accounted for the unobserved variation in risk attitudes among debtors and found that persistent debtor heterogeneity is statistically significant. The estimation also allows the estimation of an aggregate time-varying common shock that seems to be the driving force of the variation of default over time.

The estimation of the model relied on a simplified treatment of dynamics. Due to the size of the sample, estimating a fully dynamic model is complicated. It would require the computation of consumer-specific continuation payoffs along the estimation algorithm, which is computationally difficult and is left for future research.

References

- Bayer, P., F. Ferreira and R. McMillan (2007): "A Unified Framework for Measuring Preferences for Schools and Neighborhoods" *Journal of Political Economy*, August 2007.
- Deng, Y., J. Quigley and R. Van Order (2000): "Mortgage Terminations, Heterogeneity and the Exercise of Mortgage Options", *Econometrica*,

68(2), 275-307

Escobar J., C.A. Huertas, D.A. Mora and J.V. Romero (2006): “Indice de precios de la vivienda usada en Colombia - IPVU - Metodo de ventas repetidas” *Borradores Semanales de Economia* 368 Banco de la Republica, Bogota.

Epple, D., T. Romer and H. Sieg (2001): “Interjurisdictional Sorting and Majority Rule: An Empirical Analysis” *Econometrica*, 69(6), 1437-1465

Train, K. (2006): *Discrete choice methods with simulation*, Cambridge University Press.

Appendix: Summary of the estimation algorithm

The estimation of the model above is based on the computation of the simulated likelihood of the sample across observations:

- Organize the income data from highest to lowest housing payments using the surveys that contain both. Separate observations into quantiles; this joint distribution of income and housing payments is assumed to be equivalent to the joint distribution of income and K_{it}/L_{it} . Organize the observations of $\{\bar{P}_{it}, K_{it}, L_{it}, LTV_i\}$ from highest to lowest K_{it}/L_{it} and separate it into quantiles.
- For each loan in the sample generate a number J of standard normal draws ε_i that are constant over time. Match these draws with J random draws with replacement from the corresponding quantile of the

distribution of Y_{jt} , conditional on the mortgage payments. Keep these draws constant throughout the estimation.

- Set the vector of parameters $\{\hat{\zeta}_0, \sigma_{\mu 0}\}$. Compute \hat{L}_0 using (14) and (15) using the “simulated” sample described above.
- Look numerically for the set of parameters $\{\zeta^*, \sigma_\mu^*\}$ that maximize the likelihood of the sample. Compute the standard errors of the estimates using the usual methods.

Table 1: Summary statistics (main data set)

(1)	(2)	(3)	(4)	(5)	(6)	(7)
Quarter	Number of loans	Past due 3 months	Past due 6 months	Default rate	Balance/ term	Average price
1997 : 2	4965	6.4 %	1.6 %	4.0 %	370770	72482000
1997 : 3	4958	7.1 %	1.8 %	3.2 %	387160	73308000
1997 : 4	5101	8.1 %	2.4 %	3.0 %	399720	75518000
1998 : 1	7197	8.2 %	2.8 %	3.4 %	435920	64634000
1998 : 2	7365	7.9 %	3.1 %	2.3 %	447770	61548000
1998 : 3	7502	8.8 %	3.8 %	1.9 %	469250	58331000
1998 : 4	7569	10.6 %	4.6 %	2.8 %	493660	56400000
1999 : 1	7482	14.1 %	6.3 %	4.1 %	523170	57867000
1999 : 2	7809	16.3 %	7.8 %	4.7 %	504910	53608000
1999 : 3	8060	11.8 %	6.3 %	3.7 %	483630	47878000
1999 : 4	7827	19.0 %	8.3 %	6.3 %	495370	50559000
2000 : 1	8594	18.1 %	10.6 %	6.4 %	503100	47984000
2000 : 2	8020	16.1 %	9.3 %	5.3 %	478060	49014000
2000 : 3	7505	19.0 %	9.0 %	5.5 %	479880	48540000
2000 : 4	7053	19.5 %	10.6 %	3.5 %	481980	46981000

Continues in next page

Prices and balances are in 1997 COL\$.

Table 1, continued

(1)	(2)	(3)	(4)	(5)	(6)	(7)
2001 : 1	6786	20.4 %	12.1 %	2.6 %	488410	46750000
2001 : 2	6601	22.1 %	13.8 %	2.7 %	512340	38483000
2001 : 3	6416	22.9 %	14.7 %	2.6 %	520730	40771000
2001 : 4	6253	22.8 %	15.1 %	2.2 %	525090	36298000
2002 : 1	6140	22.2 %	15.3 %	1.8 %	528920	32062000
2002 : 2	6060	22.0 %	15.6 %	1.6 %	541360	34959000
2002 : 3	6028	23.4 %	16.5 %	2.5 %	553380	33041000
2002 : 4	5891	22.6 %	15.9 %	1.7 %	554660	36579000
2003 : 1	5862	23.0 %	16.6 %	1.6 %	563210	32067000
2003 : 2	5816	23.2 %	16.4 %	1.8 %	581450	32043000
2003 : 3	5580	22.6 %	16.4 %	1.3 %	584720	31138000
2003 : 4	5666	23.3 %	16.9 %	1.8 %	576500	31256000
2004 : 1	5553	22.7 %	16.8 %	1.2 %	571580	29534000
2004 : 2	5450	22.0 %	16.4 %	1.0 %	582490	31386000

Prices and balances are in 1997 COL\$.

Table 2: Linear Probability Regressions

Variable	Est.	t-stat	Est	t-stat
Constant	0.0239	14.8909	0.0271	7.8372
Balance	0.0059	12.1878	0.0053	10.7226
Price	-0.0016	-10.0016	-0.0013	-8.3808
Term	0.0004	10.8136	0.0003	5.6982
LTV	-0.0075	-3.2854	-0.0029	-1.2497

Time-effects No Yes

Prices, balances and income are measured in tens of millions of 1997 COL\$.

Table 3: Estimation Results (Standard errors in parenthesis)

	Model 1	Marginal	Model 2	Marginal
Coefficient		effects		effects
Constant	3.7933 (0.0081)	0.1144	3.6990 (0.0630)	0.1079
Balance	-0.1776 (0.0031)	-0.0054	-0.1693 (0.0089)	-0.0049
Price	0.0589 (0.0013)	0.0018	0.0492 (0.0037)	0.0014
Term	-0.0128 (0.0003)	-0.0004	-0.0127 (0.0010)	-0.0004
Income	0.0847 (0.0252)	0.0026	0.0859 (0.0723)	0.0025
σ_μ	0		0	

	Model 3	Marginal	Model 4	Marginal
Coefficient		effects		effects
Constant	3.7911 (0.0081)	0.1144	3.0973 (0.0569)	0.1011
Balance	-0.1738 (0.0031)	-0.0052	-0.1038 (0.0091)	-0.0034
Price	0.0579 (0.0013)	0.0017	0.0645 (0.0034)	0.0021
Term	-0.0129 (0.0003)	-0.0004	-0.0007 (0.0009)	-0.00002
Income	0.0825 (0.0251)	0.0025	0.0453 (0.0631)	0.0015
σ_μ	-0.0169 (0.0051)		-0.264 (0.0128)	

Models 2 and 4 contain time-effects (not shown). Prices, balances and income are measured in tens of millions of 1997 COL\$.

Table 4: Predicted Default Probabilities (selected quarters; evaluated at mean values as of 1997:3)

Variable	1997:3	1998:3	1999:3	2000:3	2001:3	2002:3	2003:3
<i>L</i>							
27	0.0435	0.0331	0.0665	0.0586	0.0106	0.0023	0.0121
40	0.0504	0.0384	0.0769	0.0678	0.0123	0.0026	0.0141
45	0.0533	0.0407	0.0812	0.0717	0.0131	0.0028	0.0149
51	0.0571	0.0436	0.0868	0.0766	0.0140	0.0030	0.0160
57	0.0611	0.0467	0.0926	0.0818	0.0151	0.0032	0.0172
<i>K</i>							
0.3644	0.0439	0.0305	0.0588	0.0509	0.0089	0.0019	0.0099
0.7956	0.0468	0.0325	0.0626	0.0542	0.0095	0.0020	0.0106
1.1973	0.0496	0.0344	0.0663	0.0574	0.0101	0.0021	0.0112
1.7831	0.0539	0.0375	0.0720	0.0624	0.0110	0.0023	0.0122
3.8742	0.0727	0.0509	0.0964	0.0838	0.0151	0.0032	0.0168
<i>P</i>							
1.3044	0.0639	0.0433	0.0778	0.0652	0.0111	0.0023	0.0121
2.1439	0.0607	0.0411	0.0740	0.0620	0.0105	0.0022	0.0114
3.1892	0.0570	0.0385	0.0694	0.0581	0.0098	0.0021	0.0107
4.9138	0.0512	0.0346	0.0625	0.0523	0.0088	0.0019	0.0096
10.8746	0.0353	0.0237	0.0432	0.0360	0.0060	0.0013	0.0065

The probabilities were evaluated at $\bar{K} = 1.614$, $\bar{P} = 4.493$, $\bar{L} = 42$ and $\bar{Y} = 0.0865$; prices, balances and income are measured in tens of millions of 1997 COL\$.