THE JOURNAL OF REAL ESTATE RESEARCH

# Self-Selection and Tests for Bias and Risk in Mortgage Lending: Can You Price the Mortgage If You Don't Know the Process?

*Robert F. Phillips**
*Anthony M. J. Yezer**

***Abstract.*** **There is increasing interest in understanding the determinants of mortgage rejection by lenders and default by borrowers. Although many researchers have proposed simple single-equation models of rejection and default, we argue that far more complex econometric specifications are needed. This paper focuses attention on problems of sample selection in the process creating a sample of applicants for conventional mortgages. We illustrate that corrections for sample selection bias may have a substantial effect on estimation results and hence should not be ignored in studies of mortgage rejection or default.**

## Introduction

The sample of mortgages reaching any advanced stage of the process, i.e., final approval by the lender or residing in the portfolio of an ultimate investor, is generated by a number of stages at which agents make decisions. These stages include: borrower selection of a mortgage type and lender, lender and perhaps approval of the borrower for private mortgage insurance (PMI), final acceptance by the borrower of the terms offered, and possible sale to secondary market. There is substantial interest in determining the behavior of particular agents involved in the mortgage transaction. First, regulators need to determine if lenders or private mortgage insurers, in the underwriting process, are discriminating against particular demographic groups. Second, a variety of actors need to determine the credit risk posed by lending to particular borrowers under various terms. These two questions are clearly related, and crucial to fair-lending regulation, because the justification of differential treatment of borrowers is potentially discriminatory unless justified by a business purpose, presumably credit risk.

In practice, we can only observe underwriting practices on the mortgage products offered by a limited fraction of all lenders for applications that are actually completed and only observe ultimate default losses for loans that are actually originated. In both cases, the sample that is observed is a small and likely nonrandomly selected subsample of all potential mortgage activity. Unfortunately, virtually all analysis of lending criteria and credit risk, including recent papers by Ferguson and Peters (1995) and by Hunter and Walker (1995), is based on single-equation models of applicant rejection, default, or default loss.[1] These single-equation models provide information that is conditional on the process in which the mortgages are originated. For some purposes, conditional estimates

*Department of Economics, The George Washington University, Washington, DC 20052.
Date Revised—June 1995; Accepted—July 1995.

are adequate but we will demonstrate that they may yield seriously biased estimates for two purposes, testing for discrimination and credit risk, for which there is increasing demand today. In addition, there has been almost no research on some aspects of the selection process that create the observed subsamples that form the basis for the single-equation models. Choice of FHA vs. conventional mortgages and of fixed- vs. adjustable-rate mortgages have received some attention but determinants of participation in special programs, partial completion of applications, application for and granting of private mortgage insurance, and refusal of mortgage terms offered have been largely ignored.

In this paper we discuss the nature of the econometric problems of sample selection bias complicating the estimation of equations to be used in testing for discrimination or credit risk. This involves initial consideration of the circumstances under which unconditional rather than conditional estimates are needed, followed by a review of the estimation problems involved in obtaining such unconditional estimates. Then we offer two illustrative examples and test for potential problems caused by selection bias. First, we consider testing for discriminating in mortgage lending by estimating a mortgage rejection equation to determine if demographic characteristics of the borrower are associated with differential underwriting outcomes. Sample selection bias in estimates of the rejection equation may arise if applicants self-select among alternative mortgage programs or lenders. Second, we examine the effects of sample selection on estimates of a mortgage refusal equation. Refusals can only be observed once the application has been approved. Finally, based on these two very different illustrations, we conclude with implications of these findings and suggestions for future research.

## When Are Unconditional Estimates Needed?

Conditional estimates of mortgage approval are obtained from single-equation models using the actual sample of applications received for a particular type of mortgage by one or more lenders, and conditional estimates of credit risk are based on the subsample of mortgages originated from the initial applicants. We refer to these estimates as conditional estimates of a behavioral equation because they apply to a subset of individuals, for example, all those who actually applied for a particular type of mortgage. Conditional estimates predict what happens as a result of the decision criteria of lenders in underwriting mortgages and borrowers in defaulting as long as the process generating these subsamples is unchanged. In contrast, there are a number of possible unconditional estimates of mortgage rejection that predict what would happen if some larger sample of potential applicants were to apply for a particular mortgage type offered by one or more lenders. The appropriate unconditional estimates of a rejection equation will depend on the purposes to which the inferences are put. For many regulatory purposes, estimates of the rejection equation unconditional to all applicants at a lender or group of lenders is needed.[2] Unconditional estimates of credit risk predict the losses if some expanded sample of potential applicants were approved.

A simple illustration of these differences is provided by the following three-equation system:

$$A^{*}=k_A+X_A\alpha_A+\varepsilon_A \tag{1}$$

$$R^{*}=k_R+X_R\beta_R+\varepsilon_R \tag{2}$$

$$D^{*}=k_{D}+X_{D}\theta_{D}+\varepsilon_{D}\ , \tag{3}$$

where $A^*$, $R^*$, $D^*$ are vectors of dependent variables reflecting the probability that actual mortgage applicants at one or more lenders apply for a particular mortgage type at a group of lenders, the probability of rejecting the application for endorsement, and the likelihood of future default respectively; the $k_J$ are constant terms, $\alpha_A$, $\beta_R$ and $\theta_D$ are vectors of coefficients to be estimated; $X_A$, $X_R$, and $X_D$ are matrices of independent variables, and the $\varepsilon_J$ $(J=A,\ R,\ D)$ are random error terms whose properties will be the object of some discussion. Estimates of equations (2) and (3) are the basis of attempts to detect the presence of bias in lending and credit risk, respectively.

The common procedure is to form conditional estimates of the determinants of the probabilities of application, rejection and default. These are obtained by sequential probit or logit techniques, i.e., by applying single-equation probit or logit estimation to each equation in turn. The actual values $J^*=A^*,\ R^*,\ D^*$ are not observed but an indicator $J=1$ if $J^*>0$ and 0 otherwise is observed. Of course, the second equation can only be observed for cases in which $A=1$ and the third equation is only observed when $R=0$. Thus the estimates are conditional on the prior stage decision.

Now consider attempts to test for discrimination in conventional mortgage lending by estimating the rejection equation, (2). Assume that there is an alternative government-sponsored program offered through these lenders that is attractive to high-risk borrowers. Then this type of applicant will self-select away from the conventional mortgage products and the threshold variable for applying for a conventional mortgage at these lenders, $A^*$, will be negatively associated with credit risk. If some of the risk factors that influence the mortgage product application and rejection decisions are unobservable to the econometrician, then $\rho_{AR}$, the correlation between the error terms, $\varepsilon_A$ and $\varepsilon_R$, is negative.[3] Given that $R=1$, 0 is only observed conditional on $A=1$, this implies that, in the subsample of applicants to conventional lenders, $E_C(\varepsilon_R)<0$, and hence the estimated constant term $\hat{k}_R$ will be biased downward.[4] Put another way, estimates of equation (2) will tend to predict lower levels of loan rejection than would occur if all applicants chose to apply for conventional mortgages. The argument is symmetric with regard to the self-selection process of applicants. If applicants with unobserved characteristics associated with lower risk are differentially attracted to the government-sponsored program, we would find $\rho_{AR}>0$, $E_C(\varepsilon_R)>0$, and $\hat{k}_R$ biased upward.

Now extend the thought experiment further by assuming that the government-sponsored mortgage program is only available to minority borrowers and that minority status has no effect on the probability of rejection by conventional lenders. These assumptions simplify the argument to the point where it can be made conceptually. Because self-selection into the government program is unavailable to non-minority applicants, the expectation of the error term in conditional estimates of the conventional rejection equation for non-minority applicants would be zero. However, the expectation for the error term in conditional estimates of the conventional rejection equation for minority applicants would be negative (positive) in cases where the government program was particularly attractive to applicants in the highest (lowest) risk categories in terms of unobserved factors in the error term. Conditional estimation of the rejection equation by ordinary probit techniques using a specification that incorporates a dummy variable for minority status will then result in a negative (positive) bias in the estimated minority coefficient, compared to its true value of zero, as the government program is differentially attractive to minority applicants in the highest (lowest) risk categories. In effect, the argument for a biased and inconsistent estimate of the constant term can be carried over to other parameters of the rejection equation when single-equation probit techniques are used. Clearly estimating a mortgage rejection equation with single-equation probit raises potentially serious problems of self-selection bias if regulators want to make inferences about what would happen if underwriting criteria for conventional mortgages were applied to all applicants. This confusion over choice of estimator may lead to perverse regulatory feedback. For example, lenders may respond to accusations of bias by increasing awareness of special loan programs. But the simple model above illustrates that this would increase selection bias and cause single-equation probit models of rejection for conventional mortgage products to indicate increased discrimination. Alternatively lenders could "improve" their fair-lending status without changing underwriting criteria simply by encouraging the best minority risks from the special mortgage programs to apply for conventional mortgages. Changing the selection process in which applicants choose mortgage products will change single-equation probit estimates of the rejection equation even if underwriting criteria are unchanged. Such changes in process do not change unconditional estimates. Thus unless there is information on the process generating the mortgage applicants, it may be misleading to rely on single-equation probit estimates of rejection equations to test for discrimination.

Although this detailed discussion has addressed problems of testing for discrimination with a rejection equation, the problem of partial observability and the potential selection bias created by relying on a subpopulation of endorsed mortgages when estimating default equations is likely more significant. Interest in determinants of default has been linked to regulatory examination of underwriting criteria by fair-lending requirements that variables used in rejection be justified by a legitimate business purpose. This justification requires estimates of default or default loss that are not conditional on the current underwriting criteria being used to reject a fraction of applicants. Unconditional estimates for the population of applicants are needed.[5] Default cannot be observed on mortgages that are rejected. Given that rejection is based on characteristics associated with default, it is likely that the same factors influencing the error term of the rejection equation also influence the error term of the default equation. Under these circumstances, the problem of partial observability results in estimates of a conditional default equation obtained using single-equation probit techniques that yield biased estimates of the parameters of the unconditional default equation. Berkovec et al. (1994) have recently demonstrated that the conditional default equation can be used to test for past discrimination, but estimates of the unconditional default equation are required to validate a credit scoring scheme.

Much of the discussion of the problem of partial observability has been used on the bivariate probit model developed in articles by Poirier (1980), Farber (1983), Fishe, Trost and Lurie (1981), and Meng and Schmidt (1985) following the seminal enquiry by Zellner and Lee (1965). This literature analyzes the significant estimation problems present in cases of sequential partial observability. These problems arise if there is a cross-equation correlation among error terms. Such correlation is likely because a common set of factors that are difficult for the econometrician to observe are likely to be missing from several, if not all, of the equations. For example, expected future income and expected future collateral value are not observed by the econometrician, and one must be content with statistical constructs. These unobservable variables are considered by lenders in rejection decisions and play a role in the final observation of default or default loss. To the extent that the cross-equation correlation among error terms in equations (2) and (3) is positive, we can conclude that the expected value of the error in equation (3) conditional on loan approval, $E_C(\varepsilon_D)=E(\varepsilon_D \mid R=0)$, is negative and that the estimated constant term in a default equation estimated using single-equation probit on the subsample of endorsed mortgages will be biased downward. However, estimates of default in terms of the entire model represented by equations (1) through (3) would require us to consider two levels of self-selection, i.e., we would need to consider the possibility that neither $\rho_{RD}$ nor $\rho_{AD}$ is equal to zero. Similarly, the effects of selection bias on estimates of slope coefficients in the default equation is very complex and even the direction of the effect is not likely to be determined based on a priori considerations.

## Illustration of Selection Bias in Tests for Discrimination

Ultimately the importance of self-selection problems raised here is an empirical question which has been neglected in the literature. In this section, we illustrate the problem using a natural experiment from the literature on discrimination in mortgage lending. Specifically, we evaluate the difference between conditional and unconditional estimates of a rejection equation when the process generating applicants has features similar to that discussed in the previous section. In terms of the system presented as equations (1)–(3), we are interested in estimating equation (2) but recognize the problems created by prior selection due to equation (1).

The specific example selected for analysis is based on the data on mortgage applications taken in the Boston MSA during 1990 and collected by the Federal Reserve Bank of Boston (see Munnell et al., 1992).[6] The natural experiment arises because the sample included applications made on ordinary conventional loans and applications made through special programs, usually programs with direct government sponsorship. The most common special program in the dataset is offered by the Massachusetts Housing Finance Agency but a number of other smaller programs, particularly for first-time homebuyers, appear. These special mortgage programs have qualification criteria that include income, loan amount, demographic characteristics of the household, previous homeownership, and even location. Households self-select into the special mortgage programs based on their perception of the specific advantages of participation. Even among low income homebuyers requesting small loan amounts to buy inexpensive houses, a substantial proportion do not participate in these programs and apply for "regular" conventional mortgages.

Subsequent approval or rejection of the mortgage application is then observed for all applicants. However, we are unable to observe the underwriting process that would have been used for special program applicants if they had applied for conventional mortgages. There are a number of reasons to suspect that underwriting for special program participants is different than underwriting for conventional mortgages. First, some of these programs carry guarantees against credit risk. Second, special program applicants have taken extra steps to find programs, document eligibility, and comply with regulations. Third, underwriting for special programs places an extra burden on lenders to determine both program eligibility and creditworthiness. Following a detailed reexamination of banking files, Horne (1994) has reported that special program applicants may be rejected because of failure to meet conditions of the program. Such rejections may have little to do with underwriting on conventional mortgage applications.

The existence of special programs creates a problem of partial observability. We can only observe the regular mortgage rejection decisions for the subsample of applicants who seek regular conventional loans. In various studies using this data, the problem of partial observability for special program participants has been "solved" in three ways: totally ignore the problem (Horne, 1994), use all observations but insert a dummy variable for program participation (Glennon and Stengle, 1944), and estimate a rejection equation after casewise deletion of observations on special program applicants (Hunter and Walker, 1995). None of these approaches results in unbiased or consistent estimates of the parameters of the conventional loan rejection equation, shown as equation (2) in the previous section. The empirical importance of this estimation problem is illustrated by the reestimation conducted in this section.

Because the reliability of this dataset has been questioned in a number of sources (see discussion in Horne, 1994), some modest attempts at data-cleaning were undertaken on the sample of 2,814 observations before performing the estimation. First, applications that were clearly intended for the purchase of rental properties were dropped because underwriting investment properties should differ from owner-occupied units. Second, six cases in which the appraised value exceeded both the loan amount and the purchase price by exactly two million dollars were adjusted so that the appraised value equaled the purchase price. Third, based on alphabetic codes used to describe special program mortgages, a number of observations were deleted because they involved sale of real estate owned, construction loans, VA or FHA mortgages, second mortgages, and other categories that were clearly not special low income programs or regular conventional mortgages. Taken together, these steps reduced the sample size by 352 to 2,464 applications.

Compared to problems with the data noted in the literature, these adjustments are quite modest and still leave substantial room for errors in variables. For example, the final loan-to-value ratio in the sample used here has a mean of .75 and ranges from .03 to 6.68 with a standard deviation of .23. Clearly the final version of the data has substantial measurement error but this has not prevented estimation results based on this data from gaining substantial attention. Measurement error is less of a concern in this study because our focus is on illustration of estimation problems created by selection bias rather than specific hypothesis testing where the effects of measurement error could be consequential.

The variables used in the empirical analysis are described in Exhibit 1. Most of these variables are straightforward applications of the data and have been used in previous studies. Sales and appraised prices and income are expressed in thousands of dollars so that estimated coefficients are scaled more conveniently. *Loan/Value*, the loan-to-value ratio, used by lenders to make the reject decision, is based on appraised value because this is the appropriate decision variable for the lender's estimate of credit risk. The applicant's decision is based on contract price, $PRICE \times 10^{-3}$, because appraised value is not available until after the application decision has been made. The variable labeled *CREDIT* was constructed as the product of individual variables indicating repayment problems based on observations of the mortgage payment history, consumer credit record, and formal indicators of bankruptcy or charge-offs.[7] Higher values of *CREDIT* reflect greater problems with repayment and applicants with an unblemished credit record were assigned a value of zero for this variable.

Specification of the first-stage mortgage choice equation describing the decision to

**Exhibit 1**
**Glossary of Variables**

| Name | Description |
|---|---|
| *CONDO* | Dummy variable=1 if property is a condominium, mean=.29 |
| *COSIGN-GIFT* | Dummy variable=1 if application has cosigner or gift, mean=.184 |
| *CREDIT* | Index of credit history problems, mean=2.49 |
| *CreditOK* | Dummy variable=1 if applicant passed lender's credit test, mean=.91 |
| *DEPENDENTS* | Number of dependent children in household, mean=.78 |
| *Expense/Income* | Total fixed expense to income ratio in percent, mean=33 |
| *HiAgedTract* | Dummy variable=1 if average age in tract>median, mean=.40 |
| *HiEducatedTract* | Dummy variable=1 if average education in tract>median, mean =.76 |
| *HiIncomeTract* | Dummy variable=1 if average income in tract>median, mean=.84 |
| *HiMinorityTract* | Dummy variable=1 if % minority in tract>median, mean=.083 |
| *HiVacancyTract* | Dummy variable=1 if % vacant units in tract>median, mean=.43 |
| *INCOME* | Monthly income in dollars, mean=6,311 |
| $LOAN\times10^{-3}$ | Loan amount in thousands of dollars, mean=142.3 |
| *Loan/Value* | Ratio of loan amount to appraised value, mean=.78 |
| *MALE* | Dummy variable=1 if applicant is male (single or married), mean=.79 |
| *MARRIED* | Dummy variable=1 if applicants are married couple, mean=.61 |
| *MINORITY* | Dummy variable=1 if first applicant is Black or Hispanic, mean=.19 |
| *MULTIFAMILY* | Dummy variable=1 if unit is in a 2–4 family structure, mean=.01 |
| *PMISEEK* | Dummy variable=1 if applied for PMI, mean=.192 |
| $PRICE\times10^{-3}$ | Appraised value in thousands of dollars, mean=194.9 |
| *REVIEW* | Number of times application reviewed by lender, mean=1.2 |
| *SELFEMPLOY* | Dummy variable=1 if applicant self-employed, mean=.126 |
| *UNVERIFIED* | Dummy variable=1 if unverifiable information in loan file, mean=.048 |
| *WORK<2YRS* | Dummy variable=1 if applicant has worked at current job<2 years, mean=.083 |

apply for a regular conventional mortgage, as opposed to a special program, was based first on a belief that program parameters regarding income, type and location of housing purchase intended, and demographic characteristics would be important in the borrower's choice process. Accordingly various versions of the conventional mortgage choice function were estimated using the available information in the dataset reflecting these factors. Arguments of the mortgage choice function were limited to variables for which borrowers should have information before applying. This eliminated appraised value, and credit judgments by the lender. Selection of a final functional form was based on a process in which variables were dropped from an extended estimating equation until all included regressors had a $t$-ratio greater than 1.5, except for demographic variables that were kept in the equations.

The final estimation results for the conventional mortgage choice equation are displayed in Exhibit 2. As expected, income and house price vary directly with the index of conventional mortgage choice. The negative effect of a short work history likely arises because conventional underwriting criteria consider employment stability to be important. Minority applicants and property location in a census tract with high minority residential population are both negatively associated with conventional mortgage choice. Overall, characteristics of the census tract in which the property is located appear significant for this mortgage choice decision, perhaps reflecting a general

**Exhibit 2**
**Estimate of the Conventional Application Equation**

| Independent Variables | Conventional Mortgage Applicant |
|---|---|
| Constant | .113 |
| | (.64) |
| *INCOME*$\times 10^{-3}$ | .029* |
| | (2.39) |
| *PRICE*$\times 10^{-3}$ | 1.29* |
| | (2.09) |
| *WORK<2YRS* | −.210* |
| | (−1.77) |
| *CONDO* | −.288* |
| | (−3.38) |
| *HiMinorityTract* | −.919* |
| | (−6.90) |
| *HiIncomeTract* | .601* |
| | (5.55) |
| *HiAgedTract* | .213* |
| | (2.71) |
| *HiVacancyTract* | .223* |
| | (2.56) |
| *HiEducatedTract* | .249* |
| | (2.95) |
| *MINORITY* | .308* |
| | (−3.22) |
| *MALE* | .171* |
| | (1.85) |
| *MARRIED* | .108 |
| | (1.24) |
| *DEPENDENTS* | −.053 |
| | (−1.37) |
| *NOB* | 2,464 |
| $\chi^2$ | 488** |

*statistically significant at 10% level, two-tailed *t*-test
**overall fit statistically significant at the 1% level

locational component of the special programs.

As with the choice equation, relatively inclusive specifications of the rejection equation were initially tested using single-equation probit and variables whose estimated coefficients were even marginally significant were retained. However, because the rejection equation is often estimated to provide information on possible discrimination in mortgage lending, some additional procedures, common in the literature on testing for discrimination, were adopted. Specifically conditional estimates of the rejection equation were estimated for two subsamples, minority applicants only and non-minority applicants only.[8] The results are displayed in the first two columns of Exhibit 3. Based on the previous literature, there has been some controversy over the inclusion of the variables *CreditOK* and *UNVERIFIED* in the specification. Carr and Megbolugbe (1993) have argued that these variables are based almost entirely on lender discretion and hence

**Exhibit 3**
**Alternate Estimates of the Rejection Equation**

| Independent Variables | Specification Test** Minority (1) | Non-Minority (2) | Conditional Rejection (3) | Unconditional Rejection (4) |
|---|---|---|---|---|
| Constant | −.74 | −1.67* | −1.42* | −1.14* |
| | (−.87) | (−4.56) | (−4.97) | (−3.25) |
| *Expense/Income* | .04* | .021* | .023* | .031* |
| | (3.20) | (4.98) | (5.80) | (5.45) |
| *Loan/Value* | .099 | .791* | .723* | .685* |
| | (.15) | (3.79) | (3.67) | (3.10) |
| *CreditOK* | −1.95* | −1.92* | −1.91* | −2.17* |
| | (−7.60) | (−11.9) | (−14.2) | (−12.4) |
| *UNVERIFIED* | 1.62* | 1.70* | 1.67* | 2.08* |
| | (3.83) | (9.08) | (9.76) | (8.69) |
| *CREDIT* | −.002 | .010* | .007* | .0062 |
| | (−.32) | (2.66) | (2.19) | (1.46) |
| *SELFEMPLOYED* | −.32 | .302* | .226* | .210 |
| | (−.81) | (2.19) | (1.75) | (1.29) |
| *MULTIFAMILY* | 1.21* | 1.62* | 1.288* | 1.15* |
| | (1.80) | (2.67) | (2.92) | (2.33) |
| *HiAgedTract* | −.09 | .229* | .159* | .17* |
| | (−.42) | (2.15) | (1.70) | (1.46) |
| *HiVacancyTract* | .82 | .370* | .309* | .48* |
| | (.34) | (3.41) | (3.36) | (3.55) |
| *MINORITY* | — | — | .214* | .152 |
| | | | (1.84) | (.92) |
| *MALE* | .06 | .338* | .232* | −.154 |
| | (.24) | (2.19) | (1.83) | (−.82) |
| *MARRIED* | −.20 | −.413* | −.323* | −.246* |
| | (−.09) | (−3.66) | (−3.26) | (−1.95) |
| *RHO* | | | | −.155 |
| | | | | (−.53) |
| *NOB* | 302 | 1,833 | 2,141 | 2,464 |
| $\chi^2$ | 166*** | 456*** | 669*** | 666*** |

*statistically significant at 10% level, two-tailed *t*-test
**probit estimates of rejection equation using only nonparticipants in special mortgage programs (minority applicants in column (1) and non-minority applicants in column (2))
***overall fit statistically significant at the 1% level

could be used to discriminate against minorities. Nevertheless, the separate estimates for minorities and non-minorities indicate that *CreditOK* and *UNVERIFIED* are extremely important and nearly identical in terms of sign and significance across equations. Exclusion of these variables from conditional estimates of the rejection equation would introduce extreme omitted variables bias and have a substantial effect on the problem of sample selection bias that we wish to study.

Conditional estimates for the final form of the rejection equation are shown in column (3) of Exhibit 3. The sign and significance of all variables on which there are strong a priori restrictions are in close agreement with these expectations. These results also resemble previous single-equation probit models estimated using this data. Inclusion of *CreditOK* and *UNVERIFIED* reduces but does not eliminate the positive and significant estimated coefficient for minority applicants. Note that the sex and marital status of the applicant are both statistically significant. Other than one table in which Munnell et al. (1992) report a significant estimated coefficient for marital status and nonsignificance for a male dummy, these other demographic characteristics of the household appear to have been neglected in the literature.

The equation system shown as (1) and (2) in the previous section and with the specific arguments shown in the conditional application and rejection estimates in Exhibits 2 and 3, respectively was estimated using bivariate probit techniques allowing *RHO,* the estimate of the correlation between error terms in the mortgage choice and rejection equations, to be non-zero. The specific estimator, bivariate probit with partial observability and self-selection, was developed by Poirier (1980) and is discussed in Green (1990). We expect *RHO* to be negative because unobservable applicant characteristics measuring credit-worthiness will cause more creditworthy borrowers to apply for conventional loans in the mortgage choice equation while this additional creditworthiness will make them less likely to be rejected in the second equation. Factors positively associated with choice of conventional loans should be negatively associated with rejection. To the extent that special programs are designed to serve marginally creditworthy households, loan officers should steer the better risks to conventional loans.

In addition to the problem of selection bias due to partial observability, single-equation estimates of mortgage rejection also may suffer from simultaneous equations bias. Specifically an applicant may choose the loan-to-value ratio, term-to-maturity, monthly payment, and/or cosigners based on the expectation that this choice will determine the likelihood of rejection. Concerns over endogeneity of loan terms are not new and there is substantial evidence that the effect of ignoring simultaneity is to generate false positives in testing for discrimination. Nevertheless papers ranging from Munnell et al. (1992) through Hunter and Walker (1995) and Ferguson and Peters (1995) have also ignored this problem.[9] Unfortunately, given that this dataset was taken largely from lender operating data used in underwriting decisions, identification of a separate applicant demand for loan terms is problematic.

The results of the bivariate probit estimates are shown in column (4) of Exhibit 3. *RHO* is negative but nonsignificant. Should the true correlation coefficient be negative, then applicants for conventional mortgages have unobserved characteristics that are associated with a lower tendency to reject or greater creditworthiness. If so, then applying single-equation probit without correcting for this correlation will result in downward bias in estimates of the constant term of the rejection equation. This follows because, holding observable factors constant, the probability of rejection for the subpopulation of applicants applying for conventional loans will be less than the probability of rejection for all applicants. The selection process causes conventional mortgage applicants to have unobserved characteristics that are associated with greater creditworthiness than the average applicant. Comparison of the conditional and unconditional estimates in columns (3) and (4) of Exhibit 3 shows that the constant in the unconditional estimates of the rejection equation is larger, $-1.14 > -1.42$. This is consistent with the negative *RHO.* There is a substantial increase in the coefficient for married applicants in going from columns (3) to (4) as indicated by the increase in the estimated coefficients. However, the coefficients of dummy variables for minority and male status are smaller in the unconditional estimates, column (4)<column (3). Taken together, the changes in estimated coefficients of the constant term and demographic dummies when we allow for self-selection suggest that the unobservable variables involved in producing these selection effects are different for different groups. Estimates of coefficients of non-demographic variables, particularly the continuous variables *Expense/Income, Loan/Value,* and *CREDIT*, are relatively unchanged between the conditional and unconditional estimates in columns (3) and (4), respectively.

These results indicate that, although the estimate of *RHO* is nonsignificant, use of an estimator that allows for the possibility of applicant self-selection may still have important implications for tests of discrimination. Overall, the difference, in rejection probability associated with various demographic groups is smaller and less significant in the unconditional estimates than it is in the conditional single-equation probit estimates that do not allow for self-selection.

## Illustration of Selection Bias in Estimates of Refusal

Before an approved mortgage can be originated, it must be endorsed by the borrower. Approximately 3.5% of all approved mortgages in this dataset were refused. Single-equation estimates of a refusal equation give the determinants of refusal conditional on the mortgage being approved. These may be contrasted with unconditional estimates of the refusal equation that apply to all applicants, whether they are rejected or approved. In this section, we investigate the possible presence of selection bias that may arise should single-equation probit be used to estimate a refusal equation. While the refusal equation has not been viewed as being particularly significant in the past, this illustration may be useful in considering the potential importance of selection bias in estimates of mortgage default equations.[10] Default occurs with approximately the same frequency as refusal and it is also only observed for mortgages that have not been rejected. Finally some of the factors that could generate refusal, such as discovery of a flaw in the property, or sudden financial reversals, could also prompt default if they occur after the mortgage is endorsed. It is, therefore, not unreasonable to suggest that selection bias may be even more consequential in a default equation than it is in a refusal equation.

The formal equation system characterizing the full refusal process examined here is:

$$A^* = k_A + X_A\alpha_A + \varepsilon_A \tag{1'}$$

$$R^* = k_R + X_R\beta_R + \varepsilon_R \,, \tag{2'}$$

where $A^*$ and $R^*$ are now interpreted as indexes that measure mortgage approval and refusal respectively. Actual refusal, $R$, is only observed conditional on actual acceptance, $A=1$, i.e., actual refusal is observed only for $A^*>0$.

Conditional estimates of equation (1′) are easily recovered by changing the signs of the estimated coefficients of the single-equation probit rejection equation in the previous section. Conditional estimates of the refusal equation using single-equation probit are shown in Exhibit 4. With little theory available to serve as a guide in specifying this equation, we initially tested extended functional forms and retained variables whose estimated coefficients where, at least, marginally significant. The final functional form

**Exhibit 4**
**Alternate Estimates of the Refusal Equation**

| Independent Variables | Conditional Refusal | Unconditional Refusal |
|---|---|---|
| Constant | .125 | −.54 |
| | (.69) | (1.00) |
| *Expense/Income* | −.013 | −.018 |
| | (−1.36) | (−1.59) |
| *LOAN*$\times 10^{-3}$ | .0053* | .0059* |
| | (3.03) | (2.82) |
| *PRICE*$\times 10^{-3}$ | −.0021* | −.0023 |
| | (−2.28) | (−1.42) |
| *CreditOK* | −1.269* | −.605* |
| | (−5.54) | (−3.06) |
| *PMISEEK* | −.49* | −.48* |
| | (−2.53) | (−2.01) |
| *COSIGN-GIFT* | −.328* | −.320 |
| | (−1.70) | (−1.50) |
| *REVIEW* | −.077 | −.073* |
| | (−1.40) | (−1.95) |
| *HiIncomeTract* | −.511* | −.48* |
| | (−3.21) | (−2.78) |
| *MARRIED* | −.178* | −.111 |
| | (−1.42) | (−.73) |
| *RHO* | | .997* |
| | | (9.22) |
| *NOB* | 1,881 | 2,141 |
| $\chi^2$ | 37.1** | 54.4** |

*statistically significant at 10% level, two-tailed *t*-test
**overall fit statistically significant at the 1% level

consists mainly of creditworthiness variables that entered the approval equation. New variables, not defined above and used in the application and rejection analysis, include: *PMISEEK*, a dummy variable equal to 1 if the borrower applied for private mortgage insurance; *COSIGN-GIFT,* a variable equal to 1 if the borrower had either a cosigner or gift and 2 if both were present; *REVIEW,* the number of times that the application was reviewed by the lender; and *HiIncomeTract,* a dummy variable indicating property location in a census tract with median income higher than the average for the entire Boston metropolitan area.

The conditional estimation results point to two factors influencing the applicant's decision to refuse the loan offer. First, factors associated with higher likelihood of rejection, particularly failure to pass the credit history test, larger loan amounts and being single, result in a higher index for refusal. These results suggest that individuals who believe rejection is likely apply at more than one lender. Second, factors that indicate that the applicant took extra steps to meet the lender's criteria, including applying for PMI, securing a cosigner and/or gift letter, and asking for the application to be reconsidered, were associated with lower refusal scores.

The unconditional estimates of the refusal equation allowing for self-selection were obtained using the same bivariate probit with selection techniques discussed above, and are shown in the second column of Exhibit 4. The most striking result is the estimate for *RHO* of .997, not significantly different from 1.0. This estimation result was obtained from a specification that converged but many alternative versions of the refusal equation did not converge because the estimate of *RHO* crossed the limit of 1.0.[11] Consequently, we view these estimates of the refusal equation as illustrative only. They do indicate that partial observability is a serious problem for the refusal equation. Given that *RHO* is positive, it appears that borrowers with unobserved characteristics that indicate they should be approved are more likely to refuse a loan if offered. Thus holding observable factors constant, for these borrowers, the probability of refusal is higher than for applicants in general. Consequently the conditional estimate of the constant term for the refusal equation is higher and differs in sign compared to the estimate allowing for self-selection. The statistical significance of three of the estimated coefficients differs between the conditional and unconditional estimates although most of the parameter estimates are similar.

One explanation for the positive *RHO* is that applicants with unobservable characteristics associated with approval are less certain of approval and more likely to make multiple loan applications in order to avoid total rejection. This is consistent with the finding that included variables associated with higher rejection rates are also positively related to refusal.

## Conclusions

We have argued that partial observability may be an important problem in estimating models of mortgage approval and default used in testing for discrimination and evaluating credit risk. Although the data available for answering these questions are not ideal, we have demonstrated that correction for selection bias using appropriate bivariate probit techniques does make a difference in estimates of a rejection equation that has been widely used to test for discrimination. In the case of mortgage default, the data limitations are even more severe. Nevertheless, the test for mortgage refusal does indicate that conditional estimates based on ordinary probit models of refusal may differ substantially from unconditional estimates. These findings suggest the importance of efforts to expand the data used to estimate default models to include observations on rejected applicants. This is particularly important for estimates of default losses on conventional mortgages at a time when rejection criteria are changing. If you do not know the rejection process that creates the conditional sample of originated mortgages, then maximum likelihood probit estimates of the default equation may suffer from significant selection bias because these estimates do not consider the process that created the sample of originated mortgages. Thus we conclude that, to the extent that default is like refusal, if you don't know the process then you can't price the mortgage.

## Notes

[1]An elaborate review of the extensive empirical literature on mortgage default or default loss is provided by Quercia and Stegman (1992). All empirical studies in this review use single-equation estimation techniques. Examples of widely discussed mortgage discrimination studies based on single-equation estimates of an applicant rejection equation include Munnell et al. (1992) and

Siskin and Cupingood (1993).

[2]If lenders are to be examined regarding their treatment of actual applicants, then the relevant population for the unconditional inference is all households applying for mortgages at the institution. The existence of differentiated mortgage products with different underwriting criteria, requires estimation of separate rejection equations for each product type. However, conditional estimates of rejection equations using only the fraction of total applicants seeking a particular mortgage product will not yield the desired unconditional estimates appropriate for the entire population of applicants. For other regulatory purposes, there may be interest in the potential treatment of all mortgage applicants in the market area, regardless of the source of credit which they were seeking. In this case the relevant population is all potential homebuyers. The data requirements and modeling necessary to explain lender choice by all potential homebuyers are well beyond the current state of the literature.

[3]The term "unobservable" in this context refers to what the researcher or econometrician can observe in the data, often termed the deterministic part of the inference. The presence of error terms $\varepsilon_A$ and $\varepsilon_R$ reflects many factors that lead applicants to choose a particular mortgage product or lenders to reject a particular application that are not observed in the data or are observed with substantial error. Applicants and lenders, of course, observe these factors in making their choices but the econometrician must deal with them by adding a stochastic component to each choice equation.

[4]Note that the expected value of the error term in the conditional rejection equation $E_C(\varepsilon_R)$ is the expected value conditional on an application for a conventional loan being made, $E_C(\varepsilon_R)=E(\varepsilon_R\,|A=1)$. However, the unconditional expectation equals zero and is given by the sum of the expectation of the error term conditional on application and the expectation of the error term conditional on failing to apply for a conventional mortgage. Thus the unconditional expectation may be written as $E(\varepsilon_R)=E_C(\varepsilon_R)Pr(A=1)+E(\varepsilon_R|A=0)[1-Pr(A=1)]=0$. If the government program attracts individuals whose unobserved characteristics indicate high risk, then they would likely be rejected if they applied and $E(\varepsilon_R\,|\,A=0)>0$. From the expression for $E(\varepsilon_R)$, it follows that $E_C(\varepsilon_R)<0$ and estimates of the constant term in the rejection equation will be biased downward if single-equation probit is used because this estimation procedure forces estimates of the conditional mean of the error term to be zero. As a result, the probability of rejection for the entire applicant population will be underestimated.

[5]It is not clear whether the estimates should be unconditional with respect to all applicants for a particular mortgage product or with respect to all mortgage applicants at a particular lender or group of lenders being examined. In subsequent discussion, we assume the former case because it simplifies the presentation. Note that the need for unconditional estimates is well recognized in the consumer credit field. Indeed the problem of selection bias due to rejected applications is sometimes handled in the consumer credit field by secretly conducting experiments in which all applications are approved so that selection bias due to the underwriting process is eliminated. The costs of this procedure are modest if credit limits on new accounts are small but it is prohibitively expensive for mortgage lending. Therefore econometric solutions to the selection bias problem are needed by mortgage lenders and their regulators.

[6]The Boston Fed data sample is taken from all mortgage applications at Boston MSA institutions reporting HMDA data for 1990 and having more than twenty-five mortgage applications. All 1,013 applications by homebuyers identified as Black or Hispanic were sampled along with a random sample of 2,340 out of 3,300 White applications. The data file made available for public use contained 2,816 observations with some observations eliminated due to data problems.

[7]When individual credit history variables reflecting consumer credit, mortgage repayment and past charge-offs were forced into the equations, their sign and significance were erratic. Measurement error appears to be a significant problem. For example, of fifty-three applicants participating in special programs for first-time homebuyers, fifty-two were coded as having a perfect record of previous mortgage repayment and only one was identified as lacking previous mortgage payment

history.

[8]Minority status of the applicant is based on the identification of the individual listed as the "applicant" as Black or Hispanic. In a small number of cases, the co-borrower's minority status differed from that of the borrower.

[9]See, for example, simultaneous estimation without correction for sample selection by Barth, Cordes and Yezer (1980), and extended discussion of both selection and simultaneous equations bias by Maddala and Trost (1982), Yezer, Phillips and Trost (1994) and Rachlis and Yezer (1993). The assumption that the distribution of credit risk among mortgage applicants can be drawn independently of underwriting criteria adopted by the lender, as in Ferguson and Peters (1995), implicitly assumes that applicants do not consider the likelihood of rejection when they request mortgage terms.

[10]Rosenblatt (1994) has studied withdrawal of the mortgage application. We could not observe withdrawn mortgages in this dataset and are not aware of other studies of refusal.

[11]Indeed, these estimates only converged when *LOANV* in the acceptance equation was replaced by loan amount. Other arguments of the acceptance equation are identical to those of the rejection equation in the previous section.

## References

Barth, J. R., J. J. Cordes and A. M. J. Yezer, Financial Institution Regulations, Redlining, and Mortgage Markets, in *The Regulation of Financial Institutions,* Conference Vol. 21, 101–43, Boston, Mass: 1980.

Berkovec, J., G. Canner, S. Gabriel, and T. Hannan, Race, Redlining and Residential Mortgage Loan Performance, *Journal of Real Estate Finance and Economics,* 1994, 9:3, 240–58.

Carr, J. H. and I. F. Megbolugbe, The Federal Reserve Bank of Boston Study on Mortgage Lending Revisited, *Journal of Housing Research,* 1993, 4, 277–314.

Farber, H. S., Worker Preferences for Union Representation, *Research in Labor Economics,* 1983, 2, 171–205.

Ferguson, M. F. and S. R. Peters, What Constitutes Evidence of Discrimination in Lending?, *Journal of Finance,* 1995, 50, 739–48.

Fishe, R. H. P., R. P. Trost and P. Lurie, Labor Force Earnings and College Choice Among Young Women: An Examination of Selectivity Bias and Comparative Advantage, *Economics of Education Review,* 1981, 1, 169–91.

Glennon, D. and M. Stengle, An Evaluation of the Federal Reserve Bank of Boston's Study of Racial Discrimination in Mortgage Lending, Economic and Policy Analysis Working Paper 94–2. Comptroller of the Currency, 1994.

Green, W. H., *Econometric Analysis,* New York: Macmillan, 1990.

Horne, D. K., Evaluating the Role of Race in Mortgage Lending, *FDIC Banking Review,* 1994, 1–15.

Hunter, W. C. and M. B. Walker, The Cultural Affinity Hypothesis and Mortgage Lending Decisions, Working Paper Series (WP–95–8), Chicago: Federal Reserve Bank of Chicago, 1995.

Maddala, G. S. and R. P. Trost, On Measuring Discrimination in Loan Markets, *Housing Finance Review,* 1982, 1, 245–66.

Meng, C. L. and P. Schmidt, On the Cost of Partial Observability in the Bivariate Probit Model, *International Economic Review,* 1985, 26, 71–85.

Munnell, A. H., L. E. Browne, J. McEneaney, and G. M. B. Tootell, Mortgage Lending in Boston: Interpreting the HMDA Data, Working Paper 92–7, Boston: Federal Reserve Bank of Boston, 1992.

Poirier, D. J., Partial Observability in the Bivarite Probit Model, *Journal of Econometrics,* 1980, 44, 210–17.

Quercia, R. G. and M. A. Stegman, Residential Mortgage Default: A Review of the Literature, *Journal of Housing Research,* 1992, 3, 341–79.

Rachlis, M. B. and A. M. J. Yezer, Serious Flaws in Statistical Tests for Discrimination in Mortgage Lending, *Journal of Housing Research,* 1993, 4, 315–36.

Rosenblatt, E., A Reconsideration of Discrimination in Mortgage Underwriting, unpublished Ph.D. dissertation, Rutgers University, 1994.

Siskin, B. R. and L. A. Cupingood, Use of Statistical Models to Provide Statistical Evidence of Discrimination in the Treatment of Mortgage Loan Applicants: A Study of One Lending Institution, paper presented at the Home Mortgage Lending and Discrimination Conference, Washington, DC, May 1993.

Yezer, A. M. J., R. F. Phillips and R. P. Trost, Bias in Estimates of Discrimination and Default in Mortgage Lending: Effects of Simultaneity and Self-Selection, *Journal of Real Estate Finance and Economics,* 1994, 9, 197–216.

Zellner, A. and T. H. Lee, Joint Estimation of Relationships Involving Discrete Random Variables, *Econometrica,* 1965, 33, 382–94.

*We appreciate comments from colleagues, discussants, and two referees. Final responsibility for the opinions expressed here are those of the authors. Data were supplied by the Research Department of the Federal Reserve Bank of Boston.*