J Real Estate Finan Econ (2010) 40:260–285
DOI 10.1007/s11146-008-9154-y

Discrimination and Mortgage Lending in Boston: The Effects of Model Uncertainty

Cullen F. Goenner

Published online: 8 October 2008

Abstract In 1992 the Federal Reserve Bank of Boston conducted an analysis that examined the effects of race on mortgage lending in the Boston Metropolitan Statistical Area. Collecting data on all the possibly relevant information used in the lending process, they find when controlling for a subset of this information that race has a statistically significant effect on the decision to reject a mortgage application. Other researchers, using the same dataset, have shown that analysis of alternative subsets of the variables significantly reduces the effects of race. While theory should guide variable selection, there is often no unique theory to explain social science. In such cases, uncertainty in model specification causes one to be uncertain as to the true effects of the variables of interest. This paper accounts for the effects of model uncertainty by using Bayesian model averaging and finds a reduced effect of race and weakened evidence concerning the statistical significance of the effect.

Keywords Mortgage lending · Discrimination · Variable selection · Bayesian model averaging

Introduction

Racial differences in homeownership rates are well documented. The Census Bureau reported that 75% of whites were homeowners in the third quarter of 2007, compared to only 47% of blacks and 50% of Hispanics. Such findings though are not surprising given the racial differences in rejection rates of mortgage loan applications. Data collected as part of the Home Mortgage Disclosure Act (HMDA) indicates for 2006 that white loan applicants were rejected 16% of the time, compared to 28% for black and 23% for Hispanic applicants. For some, these statistics alone suggest lenders discriminate against loan applications from

C. F. Goenner (✉)
University of North Dakota, Grand Forks, ND, USA
e-mail: cullen.goenner@und.nodak.edu

minorities. From an economic perspective there are several factors that influence the profitability of a loan, which may be correlated with race and the underlying cause of the disparity in rejection rates. For example minorities may tend to have weaker credit histories, which explain their higher rejection rate. Discrimination in the lending decision therefore exists when race influences the lending decision after controlling for all the relevant risk factors that influence the profitability of the loan.

Schafer and Ladd (1981) in an early empirical study of discrimination in mortgage lending attempt to control for these risk factors in their model's specification by including a number of variables characterizing the loan, borrower, property and neighborhood. Using loan application data from California and New York, they find statistical evidence of discrimination when controlling for differences in risk factors. Black applicants had statistically significant higher rejection rates than white applicants in 22 of the 30 areas examined in California and six of the ten areas in New York. Depending on the state and region, black applicants were 1.54 to 7.82 times more likely to have their loans rejected than similarly qualified whites. These results though were criticized as being statistically biased because the authors were unable to control for the applicant's credit history. It is well known that failing to include a variable positively correlated with race and negatively related to the lending decision, such as credit history, will bias upward the estimated effects of race.

To overcome the potential impact of omitted variables, researchers at the Federal Reserve Bank of Boston (Munnell et al. 1992, 1996) conducted an analysis of mortgage loans made in Boston during 1990. What made this study unique was that Munnell et al. (1996, 43) made the effort to obtain "every variable mentioned as important in numerous conversations with lenders, underwriters, and examiners" to the lending decision. Controlling for a subset of these factors, the authors find that race has a statistically significant effect on the decision to reject a mortgage loan, and that this result is robust across several model specifications. Black and Hispanic applicants are about eight percentage points more likely to be rejected for mortgage loans than white applicants with the same borrowing characteristics. The findings were widely discussed among the public, the banking industry, and regulators as indicating the existence of discrimination. Regulators increased exams and the justice department increased scrutiny of mergers and instituted prosecution.

The Boston Fed's results, while convincing for many, were subject to criticism.[1] Researchers (Day and Liebowitz 1998; Harrison 1998; Zandi 1993) were able to show that the magnitude and significance of race's effect is diminished when one adds variables to the specification that are included in the Boston Fed's dataset. The critics believe that Munnell et al's (1996) estimates of race are biased due to these omitted variables. The response of the Boston Fed (Browne and Tootell 1995; Tootell 1996) is that some of the variables collected should not be included due to their endogeneity. The debate, which has been waged over inclusion of two particular variables, highlights the more general issue of variable selection.

The difficulty faced here is common, where researchers have a large set of theoretically relevant variables to choose from, and the literature provides little guidance as to which measures to include (Perle et al. 1993). This may result in

[1] Ross and Yinger (2002) Chapter 5 provide a thorough discussion

researchers using different variables to come to disparate conclusions over the sign and significance for the coefficients of variables of interest. The Boston Fed and their critics assume, as typical in the literature, that they have strong prior information to which combination of variables is the "true" model that generates the data. In this paper this assumption is weakened. We assume that the researcher knows only the list of candidate variables that form the true model, but does not know which combination of these variables is the true model. The candidate variables for consideration here are those found in the Munnell et al. (1996) dataset. To incorporate our uncertainty of the model's specification into our estimates, we use Bayesian model averaging to average over the set of models supported by the data. While Bayesian model averaging is a method of variable selection, it does not determine or control for the endogeneity of variables. Therefore we examine separate model spaces that exclude and include the variables identified as potentially endogenous. We find across model spaces that there is significant uncertainty as to the true marginal effect of race. Averaging over models we find a reduced effect of race and weakened evidence concerning the statistical significance of this effect.[2] This result differs from previous findings as it is independent of the inclusion of potentially endogenous variables and miscoded observations.

Mortgage Lending Decision

The decision to grant a mortgage loan is based on a lender's desire to maximize expected returns, which is influenced by the interest rate and the expected cost from potential default. Until recently, lenders were largely unable to accurately assess and price the level of risk for mortgage loans. Lenders therefore did not alter mortgage rates based on the level of risk, but instead rationed credit. The market, which may or may not be competitive, determines the profit maximizing mortgage rate, from which the lender decides to grant mortgages to applicants who are the lowest risk. Stiglitz and Weiss (1981) theoretically motivate this type of credit rationing as due to asymmetric information problems in which the interest rate influences the probability of default. Their argument is that as the interest rate increases, adverse selection increases, attracting riskier borrowers that cannot be differentiated. Thus at the market interest rate the demand for credit may be greater than the supply, yet lenders will not charge higher interest rates as expected returns would fall after accounting for higher risk. Williamson (1986, 1987) also ties credit rationing to asymmetric information problems in lending. In Williamson's model, costly monitoring of loan contracts, rather than adverse selection and moral hazard, are influenced by the interest rate. Higher interest rates increase the probability of default, which increases the cost of monitoring, such that profit maximization need not occur at the interest rate that equates the demand and supply for funds.

[2] The statistical significance referred to here is based on Bayes factors, whereas previous studies reported significance based on traditional *p*-values. For a large sample and set standard of evidence, Bayes factors are more stringent than *p*-values (Raftery 1995). This implies that p-values lower than the typical cutoffs used in frequentist studies would be required to reject the null hypothesis of no effect

With interest rates fixed and credit rationed, the decision to reject a loan application is negatively related to the expected cost of default, which is the product of the probability and cost of default.[3] Influencing the probability of default are characteristics of the applicant, property, and the terms of the loan. Factors such as the applicant's income, wealth, occupation, age, and number of dependents influence the economic burden of loan payments, while the age and type of the property, along with neighborhood characteristics influence the market value of the collateral and the borrower's decision to default. Loans with higher down payments are less likely to default, as are loans with shorter terms, which build equity more quickly. The cost of default is a function of the value of the collateral (the home) and the terms of the loan. The probability of a lender rejecting a mortgage loan application is thus a function of applicant (A) and property characteristics (P), and the terms of the loan (T). $P(R)=f(A,P,T,M)$ Discrimination is said to exist if M (minority) applicants are more likely to be rejected than are whites when controlling for the relevant risk factors.

Boston Fed Dataset

Evaluation of discrimination in mortgage lending requires a wealth of information to adequately control for the characteristics that lenders may appropriately use to determine whether to grant a loan. Towards this end the Federal Reserve Bank of Boston (Munnell et al. 1992, 1996) decided to carefully reexamine all mortgage applications made by blacks and Hispanics and a random sample of whites for the Boston Metropolitan Statistical Area in 1990. They asked lenders to provide all the relevant information used in their lending decision, which included financial, employment, and property data. In all, 38 variables, which were noted by lenders, underwriters, and others as theoretically important, were collected for the 2,925 loans in the sample.

Of these variables, Munnell et al. (1996) report estimates from several model specifications, which use a total of 25 different variables from the public use dataset.[4] Other researchers, using the Boston Fed's dataset, have added to their model specifications the applicant's years of education (Harrison 1998; Horne 1997), number of times application reviewed (Harrison 1998), the amount of liquid assets (Horne 1997), the presence of unverifiable information (Day and Liebowitz 1998; Harrison 1998; Horne 1997; Zandi 1993), and whether the applicant met the lender's credit standards (Carr and Megbolugbe 1993; Day and Liebowitz 1998; Horne 1997; Zandi 1993). Table 1 provides a brief description of the variables typically used.

[3] One could add to this framework, as Bostic (1997) does, factors influencing prepayment. Prepayment, similar to foreclosure, results in the need for lenders to reinvest, at potentially lower rates than the original loan terms. The underlying factors that influence prepayment are the same that influence default, though they may have different effects. For example, increasing a household's income may lower the probability of default, but increase the desire for a larger house resulting in prepayment

[4] Each of the five specifications also included lender and census tract dummies, which are unavailable to the public. One specification used a measure of the rental value of the census tract that is statistically significant, but is not in the public use dataset. Another variable unavailable to the public is the applicant's probability of unemployment, which was insignificant in the two specifications that it was included

Table 1 Variable description

debtinc	Debt to income ratio
concred	1 if no "slow pay" account; 2 if one to two slow pay; 3 if more than two; 4 insufficient credit history; 5 if 60 days past due; 6 if serious delinquencies with 90 days past due
pubrec	1 if any public record of credit problems; 0 otherwise
LTVmed	1 if loan to value <=.95 and loan to value > .8
LTVhigh	1 if loan to value > .95
pmideny	1 if applicant applied for and was denied PMI; 0 otherwise
nreview	Number of times application was reviewed by lender
unverify	1 if information on the application was unverified; 0 otherwise
selfemp	1 if applicant self employed; 0 otherwise.
housexp	1 if housing expense to income ratio > .3; 0 otherwise
dprop	1 if property 2-4 family home; 0 single family or condominium
race	1 if applicant African American or Hispanic; 0 otherwise
fixrate	1 if fixed rate loan; 0 otherwise
old	1 if applicant age >= MSA median; 0 if applicant age <= median
liqasset	Value of applicants liquid assets (in thousands)
single	1 if the applicant was unmarried; 0 otherwise
school	Years of education
uria	State unemployment rate for applicants industry in 1989
gift	1 if a gift or grant was part of down payment; 0 otherwise
term	Loan term in months
vacancy	1 if tract vacancy > MSA median; 0 otherwise
netw	Value of applicants net worth
mortcred	1 if no late payments; 2 if no payment history; 3 if one or two late payments; 4 if more than two
chval	Change in median value of property in census tract, 1980-1990
boardup	1 if boarded up value > MSA median; 0 otherwise
MHFA	1 if applicant applied under Massachusetts Housing Financing Authority program; 0 otherwise
cosigner	1 if cosigner; 0 otherwise
female	1 if applicant female; 0 otherwise
depend	Number of dependents
standard	1 if applicant met lender's credit standards; 0 otherwise

Inclusion of the "credit standards" and "unverified information" variables in the model's specification has received considerable theoretical and empirical attention. Day and Liebowitz (1998) show that the explanatory power of the Boston Fed's model increases dramatically when the two variables are added to the specification. The adjusted R squared from OLS regression increases from .29 to .49.[5] The marginal effect of race though is substantially reduced from the eight percentage point difference found by Munnell et al. (1996). Controlling for these two additional variables, minorities are only 2.8 percentage points more likely than similarly qualified whites to have their loans rejected, where the effect remains significantly different than zero. Ross and Yinger (2002) find similar results using a probit regression model. The pseudo R^2 increases from .51 to .61 and the marginal effect of race decreases from 7.7 to 3.3 percentage points when the two variables are added to Ross and Yinger's specification.

The Boston Fed researchers (Browne and Tootell 1995; Tootell 1996) believe that these two variables should not be used because they are potentially

[5] Day and Liebowitz (1998) note in footnote 20 that their OLS based results were "nearly identical" to those using logit, which were not reported

endogenous. Information on whether each loan application contained information that was unverified and whether each applicant met the lender's credit standards was collected via a survey of lenders a year after the disposition of the loans. A lender's response to these questions involved their ex post judgment, which may depend on their previous lending decision. The notion is applicants who are rejected, yet have no credit problems, could be reported by lenders as having failed to meet credit standards in order to justify denying the loan. In this case the variable would be endogenous and its inclusion unwarranted. Day and Liebowitz (1998) believe that this is not clear given 45% of rejected loan applications met the lender's credit standards. Inclusion of the variable they argue is important as it may capture a lender's use of an outside credit scoring system to evaluate loans. Ross and Yinger (2002) conclude, based on results from simultaneous equations models, this is not the case and the two variables are indeed endogenous.[6] They find unverified information, when treated as endogenous, has little to no impact on the magnitude of race's effect, relative to the single equation specification that excludes the variable entirely. The treatment of credit standards has a more substantial effect. Including this variable and treating it as endogenous reduces race's marginal effect from 7.7 to 6.5 percentage points.[7]

This debate, which is over the endogeneity of two variables in the Boston Fed's dataset, demonstrates the effect that variable selection more generally can have on the inferences drawn from variables of interest. Here, we see that the magnitude of race's affect varies substantially across different model specifications. Regardless of whether researchers include or exclude these two particular variables, it should be clear that one is still left with the choice of which of the many variables to include in the model's specification. A choice which Harrison (1998) notes is difficult to justify *a priori* given each variable in the Boston Fed dataset is by construction potentially relevant. The fear, Harrison (1998, 34) adds, is whether the model estimated will "adequately represent the set of inferences that are possible with the dataset and a different set of priors as to which variables 'ought' to be included in the final equation." Harrison's solution is to estimate a model that includes the "kitchen sink", which is to say that almost every variable in the dataset, excluding credit standards, is included in the model. The finding is that race does not have a statistically significant effect on the probability of rejection.

While Harrison is uncertain *a priori* of which variables to include in the model's specification, he makes the strong assumption that each of the variables should be included. The dataset collected by the Boston Fed included all the variables that could be relevant to the lending decision and not necessarily those that are relevant. It seems it would be more appropriate to assume *a priori* that each of the variables might be relevant, and allow our estimates to account for this uncertainty.

[6] Ross and Yinger (2002) estimate a recursive bivariate probit model, where one equation models the lending decision and includes a potentially endogenous variable and the other models the endogenous explanatory variable. Each equation includes the same exogenous variables and is identified.

[7] The marginal effect of race is 4.1 percentage points when the credit standards variable is included in Ross and Yinger's (2002) specification and not treated as endogenous.

Bayesian Model Averaging

Researchers who use the Boston Fed dataset have a large number of candidate variables to choose from as controls in their models of the mortgage lending decision. Given k candidate variables, there are 2^k different linear combinations of the variables that could be used to specify the model. With lenders able to weigh the importance of factors differently, it is theoretically difficult to know which variables to include. Researchers are then free to use different subsets of the variables in their models, which may result in disparate conclusions over the effects of variables of interest. This creates uncertainty as to which model and its results are the true model that generates the data.

A Bayesian approach provides a natural way of dealing with uncertainty as Bayesian parameters take into account our prior beliefs and are expressed in terms of posterior probabilities. This allows one to compare the extent, in terms of probability, to which the data support different model specifications. Rather than base our inferences on a single model specification, we use Bayesian model averaging (BMA) to incorporate our posterior uncertainty of the model's specification into our estimates. BMA estimates are a weighted average of estimates from each of the models examined, with weights determined by the posterior support that each specification receives from the data. BMA offers a theoretically appealing method of accounting for uncertainty in model specification and has been shown to have better predictive performance than other methods. Raftery et al. (1997) demonstrate using simulated data, where the true specification is known, that BMA is better able to discern the true model relative to stepwise methods, when one allows for uncertainty. BMA estimates have also been shown to improve out of sample predictive performance. For an excellent introduction to BMA see Raftery (1995) and Hoeting et al. (1999), while Brock and Durlauf (2001) and Fernaández et al. (2001) provide applications in economics.

Consider a situation where there are K different model specifications (M_1 .... M_K) and one is the true data generating process, which is unknown to researchers. Furthermore, the choice of specification influences the inferences made on a parameter of interest, such as race's estimated effect. Leamer (1978, 91) notes in such cases that "ambiguity over the model should dilute the information about regression coefficients, since part of the evidence is spent to specify the model." With K models of interest, we can use Bayesian model averaging to incorporate this ambiguity into our estimated coefficients, where their posterior distribution given data D is a weighted average of each model's posterior estimates, with weight being given by the posterior model probabilities $P(M_k/D)$.

$$P(\beta/D) = \sum_{k=1}^{K} P(\beta/M_k, D)P(M_k/D) \tag{1}$$

The posterior model probability represents for a given model specification the likelihood of that specification being the true model that generates the data. The sum

of which across models is equal to 1. By Bayes' rule and the law of total probability the posterior model probability is

$$P(M_k/D) = \frac{P(D/M_k)P(M_k)}{\sum_{l=1}^{K} P(D/M_l)P(M_l)} \tag{2}$$

where $P(D/M_k)$ is the likelihood and $P(M_k)$ is the prior probability that model M_k is the true model.

To implement Bayesian model averaging the researcher must specify a prior on the probability that each model is the true model. This is often perceived as a shortcoming of Bayesian analysis because many view the choice to be subjective rather than objective. As Raftery (1995) notes, the choice of priors has little influence on the posterior distribution in large samples. In the analysis below a uniform prior is used, which assumes that each of the K models is *a priori* equally likely and that $P(M_1) = \ldots P(M_K) = 1/K$. This implies that the prior probability of each variable being included is 50%. Fernández et al. (2001) note this is the standard choice when there is not strong prior information to suggest otherwise. With theory only providing a generalization of which variables to include and lenders having the freedom to weigh factors differently, it seems this is a relatively neutral choice without additional information. Assuming the prior is equal to $1/K$, the posterior model probability then simplifies to:

$$P(M_k/D) = \frac{P(D/M_k)}{\sum_{l=1}^{K} P(D/M_l)} \tag{3}$$

The integrated likelihood, also referred to as the marginal likelihood, is found by integrating over parameter vector β_k

$$P(D/M_k) = \int P(D/\beta_k, M_k)P(\beta_k/M_k)d\beta_k \tag{4}$$

where β_k is a vector of parameters (coefficients and variance), $P(D/\beta_k, M_k)$ is the likelihood and $P(\beta_k/M_k)$ is the prior likelihood assigned to model M_k. Raftery (1995) demonstrates using the Laplace method of integrals that the likelihood of model M_k can be approximated as a function, $\exp(-1/2\ \mathrm{BIC}_k)$, of the Bayesian information criterion (BIC) for model k. Schwarz (1978) shows that the BIC is

$$BIC_k = -2\,log\left(\hat{L}\right) + d_k\,log(N) \tag{5}$$

with $\hat{L}$ equal to the maximized likelihood under model k, d_k is the number of parameters in model k, and N the sample size.

With a large number of candidate variables and thus models to consider, computing Eqs. 1 and 3 requires a great deal of effort. Hoeting et al. (1999) describe

two means of reducing the computations. The first method, which is used below and by Brock and Durlauf (2001), appeals to Occam's Window to discard model specifications that are not supported by the data. That is, models whose predictions are significantly worse than the best are excluded from being averaged over.[8] Raftery (1995) suggests averaging over models where the relative odds in favor of the best model are less than or equal to 20 to 1. For those model specifications excluded there is strong evidence that the data favor an alternative specification. Excluding these models has little effect on the posterior estimates given the low weight that each of the models' estimates would receive if included. An alternative used by Fernández et al. (2001) is Markov chain Monte Carlo model composition (MC^3). The MC^3 methodology is adapted from Madigan and York (1995) to approximate the posterior distribution of the models based on the models visited by the Markov chain.

The Bayes factor (Jeffreys 1961; Kass and Raftery 1995) measures the probability in which the data support one hypothesis relative to another. Consider two hypothesized specifications, a null and alternative, where we have some preconceived beliefs as to the probable validity of each given by $P(H_0)$ and $P(H_1)$. The prior odds of H_0 being true is $P(H_0)/P(H_1)$. By Bayes theorem the posterior odds of observing (H_0/D) is given by

$$\frac{P(H_0/D)}{P(H_1/D)} = \frac{P(D/H_0)}{P(D/H_1)} \frac{P(H_0)}{P(H_1)}, \tag{6}$$

which is equal to the Bayes factor $B_{01}=P(D/H_0)/P(D/H_1)$ multiplied by the prior odds. The Bayes factor does not involve prior beliefs and therefore provides a comparison of how the two hypotheses predict the data alone. For example, a Bayes factor of three is interpreted as the null hypothesis being three times more likely than the alternative. The posterior probability of the null being true $P(H_0/D)$ is equal to $B_{01}/(1+B_{01})$. Thus a Bayes factor of three corresponds to a posterior probability of the null of 75%. The significance of which can be interpreted by the user. Jeffreys (1961) for instance cites Bayes factors between 1 and 3 as evidence in favor of the null, but not worth more than a mention, whereas Bayes factors greater than 100 are decisive towards the null.[9] Raftery (1995, 139) adopts a slightly stricter rule of thumb where the strength of the evidence is considered weak, positive, strong, and very strong based on the breakpoints .5, .75, .95, and .99 on the probability scale, which correspond to Bayes factors of 1, 3, 20, and 150 respectively.

Bayes factors are fundamentally different than p-values. A p-value measures the probability of obtaining a result more extreme than observed from the data, given the null hypothesis is true. It conditions on the null being correct, thus it is

[8] The procedure uses the leaps and bounds search algorithm (Furnival and Wilson 1974) to identify models in the model space with posterior model probabilities significantly worse than the model with the highest.

[9] Jeffreys (1961) characterizes the strength of evidence for the null as barely worth mentioning for Bayes factors between 1 to 3.2, substantial from 3.2 to 10, strong from 10 to 32, very strong from 32 to 100, and decisive above 100.

unable to directly test the probability of the null being false. A *p*-value of .05 in the Neyman–Pearson methodology indicates the null hypothesis, when true, is rejected 5% of the time. This though does not imply that only 5% of the evidence favors the null because it depends on the alternative. Consider an enlightening example by Edwards et al. (1963) of a two tailed test with many degrees of freedom. A frequentist statistician asks how likely a test statistic exceeds 1.96 if the null is true. The answer is 2.5%. Similarly for a test statistic of 2.58 the response is .5%. Therefore 2% of the time, when true null are tested, *t* lies between 1.96 and 2.58. A perhaps more interesting question is the probability of observing t between 1.96 and 2.58, when the null is false. This answer depends on our prior belief of the alternative. If we were to assume that values of t between −20 and 20 are equally likely, then t falls in the range above $(20-(20))*(2.58-1.96)=1.55$ % of the time. Based on our prior, the data in fact support the null.

It is well known that *p*-values can overstate evidence against the null in large samples (Greene 1997; Leamer 1978). More generally, it can be shown (Berger and Sellke 1987; Edwards et al. 1963; Goodman 1999) that *p*-values overstate evidence against the null. For a test of a normally distributed mean one can relate *p*-values to Bayes factors and the probability of the null.[10] If we assume a distribution that most supports the alternative hypothesis, we can establish a lower bound for the Bayes factor which depends on the *p*-value. $\underline{B}=e^{-t^2/2}$, where t corresponds to the critical value associated with the *p*-value of a two sided test. A *p*-value of .05, with t of 1.96, corresponds to a minimum Bayes factor of .15 and a lower bound on the posterior probability of the null equal to 13%. This implies that a *p*-value of .05 is lower than the corresponding probability of the null, which is at least equal to 13%. Under more objective priors Berger and Sellke (1987) show that the probability of the null is at least 30%.

The point is *p*-values can mislead one as to the evidence against the null. This is particularly true in cases involving specification searches (Freedman 1983; Leamer 1978). Freedman demonstrates this using an empirical example where the true model is known to be white noise. Construct 51 randomly distributed $N(0, 1)$ variables with 2,500 observations each. Define one variable as Y and the others as $X=x1.....x50$. Based on my own simulated data, stepwise regression of Y on X reveals that seven variables are significant at the 10% level and the overall F statistic is significant at less than the 1% level. These results would lead many to infer a strong relation between Y and X. The minimum Bayes factor of the null specification, which compares the probability of the null model (i.e. true specification) to the most likely alternative, is equal to 4.72 and indicates the lower bound on the probability of the null specification is 83%. The evidence for the null, while not conclusive, should indicate caution in using an alternative specification with this dataset. As seen from this example, *p*-values used without discretion can lead to strong support for an effect, when none in fact exists. A Bayesian approach provides a natural way of comparing the relative evidence of specifications and can avoid overstating the evidence in cases involving specification searches.

When using Bayesian model averaging, the effect of a variable of interest, such as race's effect on the decision to reject a mortgage, can be summarized by its posterior

[10] Berger and Sellke (1987) and Goodman (1999) provide details of the derivation.

mean, variance, and effect probability. Raftery (1995) reports the posterior mean and variance for a particular coefficient, $\beta_{1,}$ can be approximated by

$$E(\beta_1/D,\beta_1 \neq 0) \approx \sum_{A1} \widehat{\beta}_1(k)P(M_k/D)Var(\beta_1/D,\beta_1 \neq 0)$$
$$\approx \sum_{A1} \left[Var(k) + \beta_1(k)^2\right]P(M_k/D) - E(\beta_1/D,\beta_1 \neq 0)^2 \quad (7)$$

where $\widehat{\beta}_1$ (k) and Var(k) are the maximum likelihood estimates and variance of β_1 under model k, and the summation is over models that include β_1 (set A_1). The posterior mean for a variable's effect is a weighted average of the ML estimates for the variable from each of the specifications, weighted by the probability that each specification is the true model. The posterior effect probability Pr[$\beta_1 \neq 0/D$] measures the posterior probability that a particular coefficient differs from zero, which is the sum of the posterior model probabilities for the models that include β_1. The posterior effect probability allows one to evaluate the relative evidence in favor of a variable having an effect. Bayes factors are used again, where this time the hypotheses involve comparing the probability of a non-zero effect, rather than the probability of a model specification. Similar rules of thumb to those discussed above can be used to evaluate the statistical significance of an effect. Under Raftery's guide for there to be strong evidence of an effect one must find posterior odds of 20–1, which corresponds to evidence of a non-zero effect of 95%.

Empirical Analysis

Researchers (Day and Liebowitz 1998; Harrison 1998; Horne 1997; Munnell et al. 1996; Zandi 1993) suggest a number of variables in the Boston Fed's dataset are both theoretically and empirically important to the mortgage lending decision and are thus likely candidates to be in the true model that generates the data. These variables are described in Table 1. Two of these variables, unverified information and credit standards, though are potentially endogenous and their use is at least suspect. A third variable, which measures whether primary mortgage insurance (PMI) was denied, was also identified by Munnell et al. (1996) as potentially endogenous. Munnell et al. include the variable in their base specifications, but show that excluding the variable has little impact on the effects of race, as does dropping observations where PMI was denied. A fourth variable, used by Harrison (1998), is the number of times the account was reviewed by the lender. This reflects a decision by the lender, which may not be independent of their lending decision, and is thus also potentially endogenous. Bayesian model averaging cannot account for endogeneity, therefore we will consider separate model spaces that exclude and include these variables.

For later comparison, we present here in Table 2 the logistic regression estimates, robust standard errors, and marginal effects from a few single model specifications. The first model replicates the specification used by Munnell et al. (1996) in column

Table 2 Logistic regression estimates of loan denial

	Boston fed specification			Kitchen sink—excluding potentially endogenous variables			Kitchen sink—excluding credit standards			Kitchen sink		
	Coefficient	Robust SE	Marginal effect	Coefficient	Robust SE	Marginal effect	Coefficient	Robust Std. Error	Marginal Effect	Coefficient	Robust Std. Error	Marginal Effect
constant	−6.200573	0.426315	−5.565762	0.895346			−5.565762	0.895346		−2.683281	1.031128	
housexp	0.449522	0.146001	0.552053	0.180690	4.59	5.21	0.552053	0.180690	4.59	0.451977	0.203539	3.06
debtinc	0.053857	0.008924	0.052296	0.011833	4.56	6.71	0.052296	0.011833	4.56	0.047480	0.011242	3.29
netw	0.000082	0.000035	0.000040	0.000071	0.29	0.16	0.000040	0.000071	0.29	−0.000006	0.000105	−0.03
concred	0.286246	0.032685	0.314286	0.039056	4.23	5.09	0.314286	0.039056	4.23	0.013524	0.054086	0.13
mortcred	0.279912	0.125329	0.206199	0.154616	0.78	1.38	0.206199	0.154616	0.78	0.039120	0.168798	0.12
pubrec	1.151542	0.175629	1.454016	0.202629	18.50	20.84	1.454016	0.202629	18.50	0.630257	0.253530	6.36
uria	0.071197	0.030096	0.044819	0.032682	0.67	1.17	0.044819	0.032682	0.67	0.047098	0.033046	0.57
selfemp	0.494858	0.198721	0.644729	0.220272	5.08	5.47	0.644729	0.220272	5.08	0.753213	0.242515	4.82
LTVmed	0.648604	0.139321	0.604097	0.166735	4.43	7.55	0.604097	0.166735	4.43	0.524330	0.188630	3.17
LTVhigh	1.760337	0.291418	1.874114	0.361157	19.14	25.86	1.874114	0.361157	19.14	1.898330	0.449302	15.35
dprop	0.497568	0.159669	0.449934	0.213893	3.91	5.26	0.449934	0.213893	3.91	0.641266	0.237113	4.50
race	0.528415	0.139818	0.412682	0.183575	3.95	5.68	0.412682	0.183575	3.95	0.223436	0.213482	1.66
boardup			−0.188146	0.162224	−1.39	−0.49	−0.188146	0.162224	−1.39	−0.279956	0.178713	−1.67
vacancy			0.155740	0.168910	1.19	1.15	0.155740	0.168910	1.19	0.272585	0.188675	1.67

Table 2 (continued)

	Boston fed specification			Kitchen sink—excluding potentially endogenous variables			Kitchen sink—excluding credit standards			Kitchen sink		
	Coefficient	Robust SE	Marginal effect	Coefficient	Robust SE	Marginal effect	Coefficient	Robust Std. Error	Marginal Effect	Coefficient	Robust Std. Error	Marginal Effect
chval			0.000660	0.000670	0.67	1.20	0.000660	0.000670	0.67	0.000742	0.000733	0.61
fixrate			0.499549	0.189156	3.21	3.89	0.499549	0.189156	3.21	0.719202	0.215603	3.58
MHFA			−0.718732	0.329607	−6.01	−1.25	−0.718732	0.329607	−6.01	−0.684536	0.374980	−3.93
term			−0.001315	0.001349	−0.53	−0.05	−0.001315	0.001349	−0.53	0.000908	0.001576	0.30
gift			−0.249963	0.220058	−1.87	−1.30	−0.249963	0.220058	−1.87	−0.129737	0.234614	−0.80
cosigner			−0.507841	0.528806	−2.98	−2.30	−0.507841	0.528806	−2.98	−0.468696	0.648880	−2.05
old			0.341890	0.162639	2.40	2.09	0.341890	0.162639	2.40	0.453150	0.185733	2.53
female			−0.372345	0.211707	−2.77	−2.79	−0.372345	0.211707	−2.77	−0.324208	0.245607	−1.89
depend			−0.011525	0.070665	−0.09	0.32	−0.011525	0.070665	−0.09	−0.053692	0.086361	−0.34
single			0.418349	0.180768	2.93	2.99	0.418349	0.180768	2.93	0.470093	0.210973	2.69
school			−0.042550	0.027127	−0.84	−0.89	−0.042550	0.027127	−0.84	−0.025088	0.032026	−0.41
liqasset			0.000780	0.000291	1.41	1.34	0.000780	0.000291	1.41	0.000745	0.000319	1.09
nreview			−0.309073	0.086493	−2.37		−0.309073	0.086493	−2.37	−0.315152	0.093494	−1.97
pmideny			4.761032	0.593428	53.29		4.761032	0.593428	53.29	4.646922	0.638897	42.55
unverify			3.377951	0.268018	42.61		3.377951	0.268018	42.61	3.168474	0.310669	31.76
standard										−3.368507	0.276457	−43.82

2 of their Table 4, which excludes the four potentially endogenous variables.[11] The coefficient for race is positive and statistically significant with a *p*-value equal .00016. Transforming the logistic coefficients to probabilities, we find quite similar to Munnell et al. (1996) that minority applicants are 7.2 percentage points more likely to have their loans rejected than white applicants with characteristics similar to minorities.[12] The second model specification includes the entire set of exogenous variables. Again, the coefficient for race is positive, where the marginal effect is 5.68 percentage points with a *p*-value of .008. Comparing just these two models, let alone others using a subset of these variables, one can see there is uncertainty as to the true effect of race. Adding the potentially endogenous variables to the model specification further reduces race's marginal effect and statistical significance. The third model includes all variables, except credit standards, and finds a marginal effect of race of 3.95 percentage points and a *p*-value of .025. Interestingly, the effect of race is insignificant at the 5% level, only when the credit standards variable is added to the specification. The marginal effect of race in the fourth model is 1.66 percentage points with a *p*-value of .29. Uncertainty as to the true effect of race exists whether or not the model space excludes the potentially endogenous variables.

In recognition of the debate over inclusion of the four potentially endogenous variables, we present BMA results using two different model spaces for comparison. The first set of models consists of the 67 million different linear combinations of the 26 candidate variables, which excludes credit standards, unverified information, PMI denial, and number of reviews. The second set consists of the one billion specifications formed by all 30 variables. The dependent variable, whether a loan application was denied by the lender, is binary, so logistic models are specified. The SPlus program BIC.logit, written by Raftery and Volinsky (1996), is used to apply BMA to these model spaces. The program reports the model specifications supported by the data, their posterior model probabilities, the posterior mean and standard deviation of the coefficients, as well as the posterior effect probabilities.

For the model space that excludes the potentially endogenous variables the data support 78 different model specifications, which is to say that 78 of the 67 million possible models were within Occam's window. While perhaps a seemingly small number of models are supported by the data, the data in fact demonstrate a great deal of model uncertainty. The specification that receives the most support has a posterior model probability of only 9%, which implies that this specification has a 9% likelihood of

[11] Munnell et al. (1996) also include in their specification census and lender dummies not included in the publicly available dataset. The analysis here also uses a smaller sample of 2,562 observations, which excludes missing values of variables in the dataset and not used in this model. Munnell et al. (1996) do not calculate the marginal effect of race for the logistic specification used in Table 4, but do provide a comparable OLS estimate of 7 percentage points in column 5 of Table 4

[12] The marginal effects of the variables were calculated in the same manner as Munnell et al. (1996). For binary variables, such as race, the probability of being denied was calculated for each minority applicant and then averaged over. Using the same minority individuals' characteristics, we calculate the probabilities again for these individuals as if they were white and take the average. Comparing the two gives us the percentage point difference in average denial rates. For the continuous variables, such as net worth, the probability of denial was first calculated for each individual and then averaged over. Next, for each individual, one increases the variable of interest by one standard deviation and then recalculates their probability of denial using the original logistic regression estimates. Comparing the two averages, again, gives us the percentage point difference in average denial rates

being the true model. Basing inference on this specification or any of the 78 "best" specifications alone is not without uncertainty.

Table 3 provides BMA estimates of the posterior mean and standard deviation of the coefficients across the 78 models. Race, the primary variable of interest, has a positive effect on a loan being denied. Accounting for model uncertainty by averaging over the models, minority applicants are on average 6.55 percentage points more likely to be rejected than white applicants.[13] The variation of race's effect across models averaged over is significant. For the 15 models with the highest posterior model probability the marginal effect of race varies between 0 and 8.56 percentage points.[14] These 15 specifications appear in the "Appendix". We use an ogive in Fig. 1 to view the entire distribution of marginal effects for the set of 78 models.[15] An ogive is a graph of a cumulative distribution, where the horizontal axis lists in order of magnitude the possible data values and the vertical axis measures the cumulative relative frequency for each value. Using Fig. 1, we can find for any marginal effect of race, the corresponding probability that the effect is less than this value. Figure 1 shows it is very likely, 64%, that the effect is less than the 7.2 percentage points found here using the Boston Fed specification. Also noteworthy, is that there is a 10% likelihood that the effect is equal to zero. This shrinks the BMA estimate of race towards zero and increases the standard error relative to single model's estimates. Comparing BMA and the Boston Fed specification estimated here, we see that the marginal effect of race decreases by 9% and the standard error increases from .14 to .22.

The posterior effect probability, which measures the probability that a variable's effect differs from zero, is 90% for race. The corresponding Bayes factor of 9 indicates there are less than 10 to 1 odds of race having an effect. This implies that there is some evidence in favor of race having an effect, but that the strength of evidence is not strong according to either Jeffreys or Raftery's classifications. The evidence though is not strong enough to exclude this possibility. The Bayes factor indicates there is a 10% probability of race having no effect, which as noted in the discussion above, corresponds to a traditional p-value significantly less than 10 percentage points. The Bayesian approach therefore implies p-values smaller than the typical cutoffs are generally required to indicate strong evidence in favor of an effect. Given there is no single standard of evidence, the nearly 10 to 1 odds may be convincing enough for some to conclude that race has an effect on the decision to reject a mortgage application in Boston. Whether or not this is the case, it is clear

[13] To determine the marginal effect of a variable across models, we follow the steps described above, with one modification. Here we calculate the probability of denial for each individual using each of the 78 models' logistic regression coefficients. We then obtain a single value for each individual by multiplying the 78 probabilities by their corresponding posterior model probability and then summing the product. For the marginal effect of race this implies that for each individual we use a weighted average of the estimates from the 78 model specifications, weighted by the posterior model probability, to calculate the probability an individual is denied. We then compare the average denial rate of minorities with the denial rate of the same group of individuals if they had been white. That is, using the sample of minority individuals, calculate the average probability, treating them as white.

[14] The lowest non-zero effect of the 78 models is 5.42 percentage points.

[15] I would like to thank an anonymous referee for making this suggestion

Table 3 Logistic estimates of loan denial using Bayesian model averaging

	Model space excludes potentially endogenous variables				Model space includes potentially endogenous variables			
	Posterior mean	Posterior SD	Posterior effect prob.	Marginal effect	Posterior mean	Posterior SD	Posterior effect prob.	Marginal effect
constant	−5.98366	0.51204	100		−1.6746	0.52153	100	
housexp	0.29898	0.25495	64.3	3.69	0.02925	0.12103	6.7	0.20
debtinc	0.05712	0.00905	100	7.03	0.05453	0.00941	100	4.00
netw	–	–	–	–	–	–	–	–
concred	0.28314	0.03474	100	5.25	–	–	–	–
mortcred	0.08193	0.15126	26.1	0.44	–	–	–	–
pubrec	1.2206	0.17881	100	21.33	0.22791	0.34724	34.3	2.40
uria	0.01838	0.03549	24.3	0.39	–	–	–	–
selfemp	0.52649	0.30137	82.2	5.09	0.77383	0.30485	93.6	4.95
LTVmed	0.70593	0.14216	100	7.52	0.31132	0.29481	58.9	1.95
LTVhigh	1.84256	0.26885	100	25.97	1.72489	0.36175	100	14.46
dprop	0.3746	0.27139	72.3	4.74	0.77605	0.25896	96.8	5.31
race	0.48165	0.21554	90	6.55	0.00718	0.05591	2.1	0.05
boardup	–	–	–	–	–	–	–	–
vacancy	0.00419	0.0361	1.7	0.04	0.03085	0.11378	8.3	0.19
chval	0.00024	0.0005	21.3	0.35	–	–	–	–
fixrate	0.33169	0.223	76.2	2.95	0.4868	0.27582	83.4	2.55
MHFA	–	–	–	–	–	–	–	–
term	–	–	–	–	–	–	–	–
gift	–	–	–	–	–	–	–	–
cosigner	–	–	–	–	–	–	–	–
old	0.00166	0.0221	0.8	0.02	0.04537	0.13917	11.4	0.26
female	–	–	–	–	–	–	–	–
depend	–	–	–	–	–	–	–	–
single	0.00344	0.03235	1.5	0.03	0.07723	0.17664	18.8	0.46
school	-0.00109	0.00767	2.5	-0.03	–	–	–	–
liqasset	0.00001	0.00007	1.5	0.02	0.00001	0.00009	1.5	0.01
nreview					-0.28227	0.07847	100	-1.83
pmideny					4.46023	0.56631	100	41.29
unverify					3.03157	0.27504	100	4.84
standard					-3.56849	0.23559	100	-5.01

– These variables were not included in the models that were supported by the data

that the significance of race is much less certain and the impact smaller than that found by the Boston Fed.

A number of variables did receive strong support, including the debt to income ratio, consumer credit history, public record, and both loan to value measures. The coefficients for these variables had a sign consistent with theory. Each of the variables was *a priori* equally likely to be included in the models averaged over, but a number of them did not receive any posterior support from the data to be averaged over. These variables included the applicant's net worth, gender, and number of dependents, whether a gift was applied to the down payment, presence of a cosigner, whether the loan was part of a special program, the length of the loan, and the level of boarded up properties in the census tract. It may be the case that lenders do not use these variables because they are costly to verify (net worth), or prohibited (gender), or do not affect profitability of a loan (number of dependents).

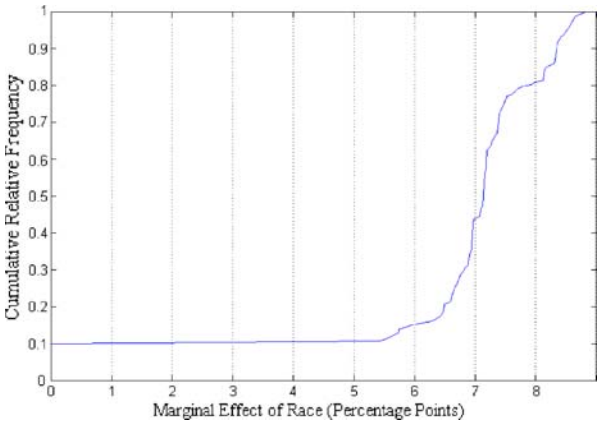

Fig. 1 Cumulative distribution of the marginal impact of race: Model space excludes potentially endogenous variables

Interestingly, we find that gender has no effect on the lending decision whatsoever, whereas Munnell et al. (1996) found female applicants to be significantly (5% level) more likely to get loans. This also points to Raftery's (1995) contention that *p*-values may overstate evidence of an effect in the presence of specification uncertainty.

Bayesian model averaging provides a method of selecting variables, but it does not determine whether variables are endogenous. Therefore we also present results that add the potentially endogenous variables (number of reviews, PMI deny, unverified information, and credit standards) to the model space to determine the sensitivity of our inferences to the treatment of these variables. The results indicate that 40 model specifications are within Occam's window and the specification with the most support has a posterior model probability of 15%.[16] Even when these variables are included, there remains a lot of uncertainty as to the true model's specification. Each of the four potentially endogenous variables receives strong support for being included in the model as does the type of property. The debt to income ratio and higher LTV ratio, remain important, while the consumer's credit history and public record are no longer important. The effects of the latter two variables may be reflected in the variable credit standards if based on an outside credit score. The marginal effect of race is negligible at .05 percentage points and the posterior effect probability is 2.1%. This case indicates there is strong evidence to suggest race does not have an effect. Figure 2 displays the cumulative distribution of the marginal effects of race, where 98% of its mass is at zero.

Including the credit standards variable in the model's specification is known to dramatically reduce the statistical significance of race as was seen in Table 2. Among the models chosen from the model space that included the potentially endogenous variables, each included the credit standards variable. Thus it might not be a surprise to conclude that race does not have an effect when using Bayesian model averaging

[16] The "Appendix" also includes posterior model probabilities and marginal effects of the 15 models with the highest posterior model probability for the model space that includes the potentially endogenous variables.

Table 4 Out of sample predictive performance

Model	Model space excludes potentially endogenous variables*				Model space includes potentially endogenous variables			
	Sensitivity (%)	Specificity (%)	Correctly classified (%)	Area under ROC	Sensitivity (%)	Specificity (%)	Correctly classified (%)	Area under ROC
BMA	32.28	97.25	87.67	0.803	65.61	96.70	92.12	0.912
Boston fed	35.45	96.52	87.51	0.787	–	–	–	–
Top PMP	34.92	96.61	87.51	0.790	62.96	96.43	91.49	0.910
Stepwise	33.33	96.52	87.20	0.781	62.96	96.89	91.88	0.910
Kitchen sink	32.28	97.25	87.67	0.792	63.49	96.89	91.96	0.913

The potentially endogenous variabiables are number of reviews, PMI was denied, unverified information, and credit standards

on this model space. Researchers, as seen in Table 2, though typically find race remains significant at the 5% level when the other potentially endogenous variables are included in the specification. When we exclude credit standards from the model space and apply BMA, there is little impact on our last conclusion. The marginal effect of race is found to be 1.65 percentage points and the posterior effect probability is 37%, which indicates evidence against an effect of race.[17]

To assess the predictive performance of BMA estimates relative to other model specifications, we compare the out of sample ability of different models to correctly classify individuals who in the data are actually denied or granted loans using both model spaces. For comparison we use the model specification that receives the highest posterior model probability, the specification selected by stepwise regression with backward elimination, the specification used by Munnell et al. (1996) in their Table 4, and the specification that includes each of the regressors, i.e. the "kitchen sink". The model specification with the highest PMP is similar to a model generated in a specification search, as the single specification that best fits a criterion is used to draw inferences. In the case of the highest PMP, the criterion is based on the Bayesian information criterion.

To examine out of sample predictions we randomly split the data in half, so each half has an equal number of loans denied. The first half of the data is used to build the models and obtain estimated coefficients. We use the coefficients from the build data along with the unused data to estimate the probability of each loan being denied. Performance is judged by the model's ability to correctly classify loans as either denied or approved using the prediction rule

$$\widehat{y} = 1 \; if \; P^* > .5 \; and \; 0 \; otherwise, \tag{8}$$

where $\widehat{y}$ is the classification of each loan and P^* is the estimated probability. The rule classifies a loan as rejected if it is more likely than not ($P^*>.5$) to be rejected.[18] The overall classification rate measures the model's ability to correctly classify loans, which is determined by the model's sensitivity and specificity. The model's

[17] Results are available upon request

[18] Greene (1997, 892:893) notes the threshold value is usually .5, but other values may be used

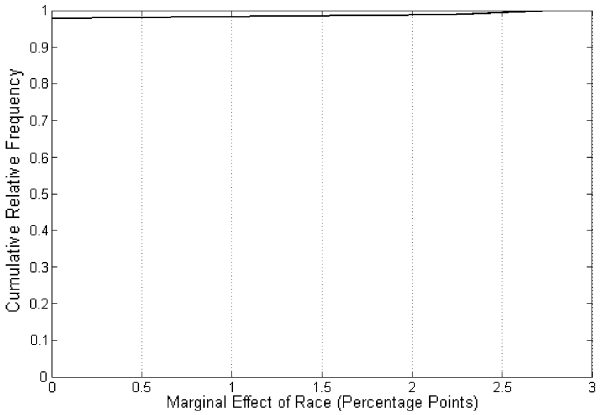

Fig. 2 Cumulative distribution of the marginal impact of race: Model space includes potentially endogenous variables

sensitivity is its ability to correctly classify loans that were denied by lenders. A higher sensitivity means fewer loans denied by lenders are misclassified by our predictions. Specificity, on the other hand, refers to the model's ability to correctly classify loans that were accepted. One can further compare the relationship between the model's sensitivity and specificity using the Receiver Operating Characteristics (ROC) curve. The ROC curve plots, for different cutoff values in the prediction rule, the relationship between the sensitivity of the predictions and the false positive rate (1-specificity). The area under the ROC can be used to summarize the relation, where a value equal to .5 indicates predictions equal to chance and a value of 1 indicates perfect prediction. Table 4 provides for each specification and model space the classification rate, sensitivity, specificity, and area under the ROC curve.

The out of sample predictions in Table 4 indicate that the overall predictive performance is quite similar across the specifications for each model space. BMA estimates produce minimally higher classification rates for both spaces. In each case the model's specificity is much more than its sensitivity. That is to say the model does a better job at predicting the loans actually accepted than the loans actually denied. Across model spaces, we see the models that include the potentially endogenous variables significantly outperform those excluding these variables as the sensitivity of the predictions improves. The area under the ROC curve suggests that BMA also outperforms the other specifications for the model space that excludes potentially endogenous variables and is second highest to the kitchen sink model for the space that includes all variables.

While the Boston Fed's dataset contains a wealth of information, several researchers including Carr and Megbolugbe (1993), Horne (1997), and Day and Liebowitz (1996, 1998) have identified several loan applications whose data values appear to be suspicious or inconsistent with each other. Carr and Megbolugbe (1993, 291) call particular attention to 53 observations that "were the worst examples of generally careless data construction." These were loans with LTV ratios greater than 3, loans with imputed interest rates either greater than 20% or less than 3%, and loans where the housing expense to income ratio was larger than the total expense to

income ratio.[19] Carr and Megbolugbe conclude that cleansing the data by dropping different sets of questionable observations does not reduce evidence of discrimination in the Boston Fed data. Day and Liebowitz (1996) strongly challenge this assertion. Using a narrower interest rate spread, they find that the effect of race is reduced by 24% when the sample includes loans with imputed rates between 7–12%.[20]

The Boston Fed's response (Tootell 1996) is largely that the criticism is much ado about nothing. Data outliers they note were double checked with the lender to ensure accuracy and lenders were warned that their data would be turned over to the appropriate regulators. Any "errors" that exist they argue are largely due to decimal errors, which they corrected for in their analysis but not in the released dataset, or are errors in variables not used in the analysis, such as whether the loan was sold. The values the Boston Fed notes, may in some cases be unusual, but are not impossible. They explain imputed interest rates can be above the market interest rate due to property taxes, insurance, and condo fees, which are added to monthly housing expense, while rates less than market are due to rental property income from multi family homes that offset expenses. This argument is hardly reassuring, as Day and Liebowitz (1998) find that taxes have a small impact on imputed interest rates and that rental income on multi-family properties is unlikely to be the source of low imputed rates.

Ross and Yinger (2002) consider Day and Liebowitz's (1996) findings to be one of the few significant challenges to those of the Boston Fed. The difficulty they note is that one is unable to determine the actual interest rate without the monthly mortgage expense, which differs from the housing expense found in the data by the amount of property taxes, insurance, condo fees, and rental income. Ross and Yinger try to isolate mortgage expenses by eliminating loans made on condominiums and multi-family properties from their sample. This eliminates loans where condo fees and rental incomes may contribute to or offset housing expenses. Mortgage expenses for the remaining loans are then equal to housing expenses minus insurance and taxes, which Ross and Yinger assume are a fraction of the home's price. For robustness, they calculate interest rates for each loan under the assumption that taxes and insurance are 2%, 2.5%, or 3% of the purchase price and use interest rate cutoffs of 7% to 12% and 5% to 14%.[21]

Ross and Yinger find that the marginal effect of race varies with their assumption on the amount of taxes and insurance and the interest rate cutoff used to exclude observations. The most confusing aspect they note is that race's effect is weakest under the 7 to 12 cutoff and strongest under the 5 to 14 cutoff, when they assume taxes and insurance are 2.5% and exclude the credit standards variable. This confusion Ross and Yinger believe is fundamentally

[19] The interest rate can be determined using the monthly mortgage expense, loan term, and loan amount. The Boston Fed dataset contains the variable monthly housing expense, which adds property taxes, insurance, condo fees, while deducting any rental income. Thus the imputed interest rate differs from the effective rate depending on the variation in mortgage and housing expenses

[20] The average mortgage rate was approximately 10% in 1990 at the time the loans were made

[21] Day and Liebowitz (1998, 22) estimate the value of insurance and taxes is 1.8%, based on the fact that the property tax in Boston for 1993 was 1.4%, which was higher than the rate in 1990, and their assumption of an insurance rate of .5%.

Table 5 Impact of minority status on loan rejection (percentage points)

Insurance (%)	Model space excludes potentially endogenous variables*				Model space includes potentially endogenous variables			
	7% to 12% Sample		5% to 14% Sample		7% to 12% Sample		5% to 14% Sample	
	Marginal effect	Posterior effect prob.	Marginal effect	Posterior effect prob.	Marginal effect	Posterior effect prob.	Marginal effect	Posterior effect prob.
2	3.86	44.2	0.51	9	0.05	2.1	0.03	1.2
2.50	5.9	61.1	1.34	19.4	0.68	16.9	0.14	5.1
3	10.11	84.6	4.55	50.9	5.5	75.2	0.32	9.8

The potentially endogenous variabiables are number of reviews, PMI was denied, unverified information, and credit standards

caused by errors in the imputation procedure and not necessarily the data itself.[22] Another possibility is that it is not imputation error, but specification uncertainty, which is the cause of the confounding results across samples. To test this theory we recreate the samples used by Ross and Yinger and apply BMA to the model spaces that include and exclude the potentially endogenous variables. A clear pattern appears in Table 5 when one accounts for specification uncertainty. Across both interest rate cutoffs, the impact and significance of race rises with the amount of insurance and taxes. It is also quite clear that the effect of race is larger when the 7% to 12% cutoff is used. Despite this, for each of the model spaces and samples, the posterior effect probabilities indicate little evidence of race affecting mortgage denial. Under the assumption that taxes and insurance are 2%, which is closest to Day and Liebowitz's (1998) estimate of 1.8%, there is evidence against an effect for both samples and model spaces.

Conclusion

The decision to deny a mortgage loan is complex, as few mortgage applications are perfect. Lenders thus often have the ability to use their own judgment, which may result in different lenders emphasizing different characteristics of the borrower or

[22] Ross and Yinger (2002, 135) also speculate that screening imputed interest rate values may potentially bias the impact of race. "If minorities tend to live in areas with relatively high property taxes and insurance rates, for example, then assuming a low mill rate for tax and insurance costs may disproportionately—and inappropriately—filter out minority applications." This does not appear to be the case with the data used here. Minorities represent 14.5% of the unfiltered sample. Assuming taxes and insurance are zero and using a 5–14% interest rate filter, the remaining sample is 15.5% minority. Similarly, under a 2%, 2.5%, and 3% insurance and tax rate the sample is 15%, 15.1%, and 15.7% minority respectively. The limited impact of taxes and insurance on the sample of minorities is potentially due to offsetting effects of minority status on insurance and tax rates. Our analysis of tax data (Harrison and Rubinfeld 1978) from 506 census tracts in the Boston Metropolitan Statistical Area for 1970 reveals a strong negative relation between the proportion of blacks and the property tax rate. The t statistic of a bivariate regression is -11.06. Higher property insurance rates in minority areas though are likely, which possibly offset the lower tax rates

property in their decision to lend. This implies that there are a large number of theoretically relevant variables that lenders may use. For researchers the large number of candidate variables then implies a rather large space of models for consideration. With 30 potentially relevant regressors there are more than one billion different linear combinations of these variables that researchers may use for their model specification. As is often typical, researchers such as Munnell et al. (1996) report the results from a small number of model specifications, which largely ignores the effects of model uncertainty.

Bayesian model averaging allows researchers a formal treatment of the issue of model uncertainty and can improve predictive performance. With respect to mortgage lending, we find that there is a great deal of uncertainty in the true model specification that generates the data, which is independent of the decision whether to include the potentially endogenous variables. Excluding these variables from the model set, we find 78 models are supported by the data, where the model with the highest posterior model probability explains only 9% of the total model probability. Including the potentially endogenous variables in the set of models examined, results in 40 models that are supported by the data, where the model with the highest posterior model probability explains 15% of the total model probability. Basing our inferences on any of these specifications or the Boston Fed's specification alone creates uncertainty as to the true effect of variables of interest.

Accounting for model uncertainty by using Bayesian model averaging, the results indicate the marginal effect of race is likely to be less than the 7.2 percentage points found here from using the Boston Fed's specification. In addition, the statistical significance of race's effect is dramatically diminished when accounting for model uncertainty. Unlike previous findings, these two conclusions do not depend on the inclusion of the potentially endogenous variables or on excluding a large number of observations. Excluding the potentially endogenous variables from the model space, the marginal effect of race is 6.55 percentage points with a probability of 90% that the effect is non-zero. Based on this level of evidence, some may conclude the Boston Fed was correct and discrimination exists in the data. If this is the case, then it should be recognized that the statistical significance of race is much less certain and the impact smaller than that found by the Boston Fed. It is crystal clear though that race has no effect when one includes the potentially endogenous variables in the model space under consideration. The marginal effect of race is .05 percentage points and the probability of a non-zero effect is 2.1%. Filtering out loans, where imputed interest rates were unreasonable, also showed little evidence of race having an effect on lending when accounting for model uncertainty.

Acknowledgements I would like to thank seminar participants at the FDIC, Fannie Mae, Freddie Mac, JPMorgan/Chase, University of Manitoba, and Holy Cross for valuable comments and discussions. I also thank anonymous reviewers for their contributions.

Appendix

Table 6 Marginal effects (percentage points) of the 15 top models—model space excludes potentially endogenous variables

Variable	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12	Model 13	Model 14	Model 15
housexp	5.74	–	5.54	–	5.61	5.72	5.78	–	6.02	–	6.06	5.39	5.81	6.06	5.65
debtinc	6.51	7.94	6.64	8.02	6.56	6.39	6.48	7.94	6.35	8.03	6.46	6.7	6.69	6.5	6.51
netw	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
concred	5.29	5.2	5.05	4.96	5.31	5.43	5.35	5.23	5.46	4.97	5.32	5.06	5.47	5.24	5.22
mortcred	–	–	1.63	1.72	–	–	–	–	–	1.8	–	1.72	–	–	–
pubrec	21.49	21.55	21.26	21.3	21.36	20.71	21.34	21.41	20.77	21.15	21.47	21.12	22.29	21.71	21.63
uria	–	–	–	–	–	–	1.42	–	1.81	–	1.72	–	–	–	–
selfemp	6.18	6.47	6.18	6.47	6.43	5.69	5.58	6.71	–	6.74	–	6.45	6.06	–	6.37
LTVmed	7.5	7.5	7.2	7.19	7.94	7.33	7.4	7.93	7.02	7.59	7.17	7.61	8.2	7.26	7.64
LTVhigh	25.97	25.84	25.61	25.43	26.72	25.53	25.71	26.54	25.49	26.09	25.9	26.31	26.79	26.29	25.94
dprop	6.61	6.44	6.43	6.24	–	6.82	6.35	–	6.74	–	6.6	–	6.91	6.91	6.07
race	7.16	7.4	6.97	7.19	8.36	6.92	7.48	8.56	6.94	8.32	7.09	8.13	–	6.66	5.74
boardup	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
vacancy	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
chval	–	–	–	–	–	–	–	–	–	–	–	–	1.91	–	1.29
fixrate	3.93	3.91	4.01	4	4.04	–	3.75	4.01	–	4.09	3.43	4.11	3.7	3.6	3.85
MHFA	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
term	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
gift	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
cosigner	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
old	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
female	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
depend	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
single	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
school	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
liqasset	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
PMP	0.0927	0.0536	0.0435	0.0351	0.0342	0.0276	0.0276	0.0266	0.0237	0.0233	0.0214	0.021	40.021	0.0205	0.0176

Table 7 Marginal effects (percentage points) of the 15 top models—model space includes potentially endogenous variables

Variable	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12	Model 13	Model 14	Model 15
housexp	–	–	–	–	–	–	–	–	–	–	–	–	–	2.91	–
debtinc	4.12	4.16	4.08	4.12	4.05	4.00	3.97	3.91	4.07	4.05	4.00	3.91	3.96	3.46	4.12
netw	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
concred	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
mortcred	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
pubrec	–	–	6.78	7.33	–	–	–	–	–	–	7.44	–	6.88	–	–
uria	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
selfemp	5.50	5.05	5.44	5.00	5.30	5.48	4.91	5.33	5.69	5.06	5.01	5.27	5.41	5.46	5.28
LTVmed	3.33	–	3.13	–	3.68	3.27	–	3.09	3.20	–	–	3.68	3.05	3.28	–
LTVhigh	15.44	13.31	15.03	13.06	15.65	15.49	12.95	14.88	15.23	13.40	13.15	15.76	15.07	15.57	13.18
dprop	5.30	5.96	5.23	5.87	5.23	5.12	5.98	5.36	4.86	5.79	5.70	5.02	5.06	5.26	5.49
race	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
boardup	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
vacancy	–	–	–	–	–	–	–	–	2.22	–	–	–	–	–	2.39
chval	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
fixrate	3.07	2.89	3.22	3.06	3.09	3.00	–	–	3.20	2.81	2.98	3.02	3.14	3.09	3.04
MHFA	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
term	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
gift	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
cosigner	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
old	–	–	–	–	2.22	–	–	–	–	–	–	2.52	–	–	–
female	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
depend	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
single	–	–	–	–	–	2.31	–	–	–	2.38	2.44	2.63	2.36	–	–
nreview	–1.85	–1.78	–1.87	–1.81	–1.83	–1.83	–1.86	–1.92	–1.85	–1.77	–1.79	–1.81	–1.85	–1.91	–1.79
school	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
liqasset	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
pmideny	40.56	41.95	40.56	41.87	40.50	41.00	42.40	41.19	40.39	42.37	42.27	41.00	40.98	40.85	41.71
unverify	4.80	4.86	4.86	4.93	4.78	4.82	4.82	4.76	4.78	4.88	4.95	4.80	4.88	4.82	4.84
standard	–5.09	–5.17	–4.80	–4.87	–5.07	–5.09	–5.14	–5.05	–5.07	–5.17	–4.87	–5.08	–4.80	–5.07	–5.15
PMP	0.146	0.078	0.074	0.063	0.04	0.04	0.039	0.033	0.027	0.025	0.024	0.024	0.023	0.023	0.022

References

Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association*, *82*, 112–122. doi:10.2307/2289131.

Bostic, R. W. (1997). *The role of race in mortgage lending*: *Revisiting the Boston fed study*. Washington, DC: Federal Reserve Board of Governors Finance and Economics Discussion Series 1997-2, December.

Brock, W. A., & Durlauf, S. N. (2001). Growth empirics and reality. *World Bank Economic Review*, *15*(2), 229–272. doi:10.1093/wber/15.2.229.

Browne, L. E., & Tootell, G. M. B. (1995). Mortgage lending in Boston—A response to the critics. *New England Economic Review*, *September/October*, 53–78.

Carr, J. H., & Megbolugbe, I. F. (1993). The federal reserve bank of Boston study on mortgage lending revisited. *Journal of Housing Research*, *4*(2), 277–313.

Day, T. E., & Liebowitz, S. J. (1996). Mortgages, minorities, and HMDA. Paper presented at the Federal Reserve Bank of Chicago, April.

Day, T. E., & Liebowitz, S. J. (1998). Mortgage lending to minorities: Where's the bias. *Economic Inquiry*, *36*, 3–28.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*(3), 193–242. doi:10.1037/h0044139.

Freedman, D. A. (1983). A note on screening regression equations. *American Statistician*, *37*(2), 152–155. doi:10.2307/2685877.

Fernandez, C., Ley, E., & Steel, M. F. J. (2001). Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics*, *16*(5), 563–576.

Furnival, G. M., & Wilson, R. W. (1974). Regression by leaps and bounds. *Technometrics*, *16*(4), 499–511.

Goodman, S. N. (1999). Toward evidence-based medical statistics. 2: The Bayes factor. *Annals of Internal Medicine*, *130*(12), 1005–1013.

Greene, W. H. (1997). *Econometric Analysis* (3rd ed.). Upper Saddle River: Prentice hall.

Harrison, G. W. (1998). Mortgage lending in Boston: A reconsideration of the evidence. *Economic Inquiry*, *36*, 29–38.

Harrison, D., & Rubinfeld, D. L. (1978). Hedonic prices and the demand for clean air. *Journal of Environmental Economics and Management*, *5*(1), 81–102.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*(4), 382–417.

Horne, D. K. (1997). Mortgage lending, race, and model specification. *Journal of Financial Services Research*, *11*, 43–68.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford: Oxford university press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Leamer, E. E. (1978). *Specification searches: Ad hoc inference with non-experimental data*. New York: Wiley.

Madigan, D., & York, J. (1995). Bayesian graphical models for discrete data. *International Statistical Review*, *63*(2), 215–232.

Munnell, A. H., Tootell, G. M. B., Browne, L. E., & McEneaney, J. (1992). *Mortgage lending in Boston: Interpreting HMDA data*. Boston: Federal Reserve Bank of Boston Working Paper 92-07.

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

Perle, E. D., Lynch, K., & Horner, J. (1993). Model specification and local mortgage market behavior. *Journal of Housing Research*, *4*(2), 225–243.

Raftery, A. E. (1995). Bayesian model selection in social research. In P. V. Marsden (Ed.), *Sociological methodology 1995* (pp. 111–195). Cambridge: Blackwells.

Raftery, A. E., Madigan, D., & Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, *92*, 179–191.

Raftery, A. E., & Volinsky, C. T. (1996). Splus function biclogit, version 2.0. (http://www.research.att.com/~volinsky/bma.html)

Ross, S., & Yinger, J. (2002). *The color of credit: Mortgage discrimination, research methodology, and fair-lending enforcement*. Cambridge: MIT.

Schafer, R., & Ladd, H. (1981). *Discrimination in mortgage lending*. Cambridge: MIT.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.
Stiglitz, J. E., & Weiss, A. (1981). Credit rationing in markets with imperfect information. *The American Economic Review*, *71*(3), 393–410.
Tootell, G. M. B. (1996). Turning a critical eye on the critics. In J. Goering, & R. Wienk (Eds.), *Mortgage lending, racial discrimination, and federal policy* (pp. 143–182). Washington, DC: The Urban Institute Press.
Williamson, S. D. (1986). Costly monitoring, financial intermediation, and equilibrium credit rationing. *Journal of Monetary Economics*, *18*, 159–179.
Williamson, S. D. (1987). Costly monitoring, loan contracts, and equilibrium credit rationing. *The Quarterly Journal of Economics*, *102*(1), 135–145.
Zandi, M. (1993). Boston fed's study was deeply flawed. *American Banker*, August 19, p. 13.