Econometrics Working Paper EWP0007
ISSN 1485-6441

Stratified Sample Design for Fair Lending Binary Logit Models +

Judith A. Giles*
Department of Economics
University of Victoria
PO Box 1700
Victoria, BC, CANADA V8X 5A3
email: jgiles@uvic.ca

Marsha J. Courchane
Housing Economics & Financial Research
Federal Home Loan Mortgage Corporation
8200 Jones Branch Drive
MS 484, McLean, VA 22102 USA
email: marsha_courchane@freddiemac.com

July 2000

ABSTRACT

Logistic regressions are commonly used to assess for fair lending across groups of loan applicants. This paper considers estimation of the disparate treatment parameter when the sample is stratified jointly by loan outcome and race covariate. We use Monte Carlo analysis to investigate the finite-sample properties of two estimators of the disparate treatment parameter under six stratified sampling designs and three data generating processes; one estimator is consistent irrespective of sample design while the other is not. Unfortunately the inconsistent estimator is employed inadvertently in fair lending studies. We demonstrate the gains in using the consistent estimator as well as providing recommendations on sample design. We also study the effect of sample design on the empirical power of a test for statistical significance of the disparate treatment parameter. We recommend adopting a sample design that approximately balances by outcome and racial group, when using the estimator that adjusts for the stratification scheme. However, if the standard logit estimator is employed, then our results suggest a sample design that balances by outcome and allocates across racial groups proportionally to the population. Though our study is framed in terms of fair lending applications, our results apply generally to the estimation of logistic regressions that use stratified or choice-based sample designs.

JEL Classification: C49, G21, G28

Keywords: Logistic regression, design efficiency, stratified sampling, choice-based sampling, case-control studies, balanced sampling, Monte Carlo experiment, mean squared error.

---

* *Corresponding author*

+ This work was undertaken while Marsha Courchane was Senior Financial Economist at the Office of the Comptroller of the Currency (OCC). We are grateful to Linda Khalaf, Chris Nicol and Todd Vermilyea for their comments and suggestions, as well as to seminar participants at the OCC and the University of Victoria. The OCC kindly provided access to data. We also thank Betty Johnson and Andrew Kaplowitz for helpful research assistance. An earlier version of this paper was presented at the 2000 Canadian Economics Association Meeting.

1. Introduction

U.S. fair lending laws prohibit discrimination against loan applicants on prohibited bases including race, national origin, marital status, gender, and religion. Various government regulatory agencies monitor compliance with the statutes, including the Office of the Comptroller of the Currency (OCC), the Federal Reserve Board (FRB), Housing and Urban Development (HUD) and the Office of Thrift Supervision (OTS). Banking institutions covered by the Home Mortgage Disclosure Act (HMDA) are required to collect and disclose data on HMDA-reportable lending decisions that can then be used to assist regulators and the public in identifying discriminating practices, and to ensure enforcement of fair lending laws.

The Home Mortgage Disclosure Act (HMDA) mandates the collection and disclosure of data on home mortgages including: loan amount, census tract of property, purpose of the loan, loan type, application and action dates, race and gender of applicant, income relied on for loan, as well as loan disposition. Institutions are asked, but not all are required, to report the reasons why individual applications were denied. It is the only type of lending behavior for which race data may be collected legally.

Fair lending compliance regulation includes a review of a lender's loan policies and procedures to ascertain that these are not overtly discriminating against any one type of loan applicant, and to ascertain whether such policies and procedures are applied consistently across loan applicants. While enforcement varies by regulatory agency, a description of procedures can be found in the Federal Financial Institutions Regulatory Council (FFIEC) Inter-Agency Fair Lending Procedures issued in 1999 (http://www.ffiec.gov/fairlend.pdf). These procedures include detailed qualitative assessments and quantitative statistical analyses; the latter may include matched pair analysis and logistic regressions. Our focus is on the use of binary logit models to examine for disparate treatment, as discussed in and applied by, for example, Munnell et al. (1992, 1996), Carr and Megbolugbe (1993), Glennon and Stengel (1994), Horne (1994, 1997), Stengel and Glennon (1999), Harrison (1998), Day and Liebowitz (1998), Courchane, Nebhut and Nickerson (2000) and others. These logistic regressions are estimated with a binary outcome variable of whether a loan is approved or denied for an applicant as a function of covariates such as loan-to-value, debt-to-income ratio, income, one or more credit score variables, and various dummy variables to capture effects such as bad credit, insufficient funds to close, and race (e.g., white, black, Hispanic). The aim is to approximate the bank's complex underwriting criteria with the logit specification.

Ideally, the logistic regression is estimated using the population of loans, but this data, though collected, is not typically available in electronic form. Population data on the outcome variable (loan approved/denied) and some covariates (including race) is usually known. Accordingly, the logistic regressions are estimated from a sample taken from the population, which the government agencies collect, tabulate, clean and prepare for the statistical analysis. The approach is to examine for disparate treatment as a test of statistical significance for the race dummy variable after controlling for other effects.

In the fair lending studies, instead of sampling individuals unconditionally, or conditionally on a covariate vector x, and observing the outcome variable Y, choice-based or case-control stratified sampling is undertaken, for which a predetermined number of denied and approved loan applications are obtained and x is then recorded. In some cases, a further level of stratification is undertaken with the sampling stratified also by race; that is, the subjects for collection of additional data are stratified jointly by outcome (loan approved/denied) and covariate (race). Such a sampling procedure is sometimes called two-phase stratified sampling, stratified case-control sampling or stratified choice-based sampling, and it is very common in many fields including epidemiology and accountancy. By selecting a sample of a suitable size from each stratum it is possible to produce parameter estimates that are considerably more precise than that given by a simple random sample from the population. The gains from stratification will be larger the more marked the differences between the strata, and the more homogeneous the characteristics are within strata. Stratified random sampling will always increase precision over simple random sampling. How much we gain depends upon how well we carry out the stratification process.

In this paper we use Monte Carlo experiments to study the gains and losses in employing six stratified sampling designs, when estimating a racial group dummy variable parameter. We consider two estimators of this parameter: the standard logit estimator, which implicitly assumes simple random sampling, and an estimator that adjusts for the differences between strata allocations in the population and the sample. The former estimator is regularly used in fair lending studies; it is inconsistent. The latter estimator is consistent; a fact recognized in other fields.

Our simulation experiments show the reductions in bias and mean squared error that can be achieved by using the corrected estimator. We also show the impact of sample design on the finite-sample properties of both estimators, and on the sampling distribution of the t-statistic used to examine for statistical insignificance of the disparate treatment parameter. When using the inconsistent estimator, this statistic is not an asymptotic standard normal variate under the zero-value null hypothesis, and so we find that incorrect discrimination conclusions often arise.

Overall, our results suggest stratification that approximately balances by outcome and by racial group is favored when using the consistent estimator of the parameter of interest. However, if estimation uses the inconsistent standard logit estimator, then we recommend a sample design that approximately balances by outcome and allocates across racial groups to reflect population proportions.

The layout of our paper follows. In section 2 we discuss the estimation of logistic regressions with stratified samples. Section 3 provides brief descriptions of some fair lending studies that employ binary logit models to examine for disparate treatment. Our Monte Carlo design is explained in section 4, which also outlines our six sampling designs. Section 5 presents the simulation results. In the light of these results, in section 6 we illustrate the potential impacts on discrimination conclusions with three data sets from the OCC. Section 7 concludes.

2. Estimation of logistic regressions with stratified samples

We consider estimation of a logistic regression model for a binary categorical variable Y associated with a K-dimensional vector of covariates, denoted x, with parameter vector β. We assume that Y_t=0 when the t'th applicant's loan is denied while Y_t=1 when the t'th applicant's loan is approved. We suppose that the logit model is linear in β and that β includes a parameter for a minority status dummy variable, denoted DM, where DM_t=1 for a nonminority applicant. For simplicity, we express DM as our race covariate and we assume only two race categories; our analysis is easily extended to more than two strata. We suppose a finite population of N applicants with all subjects classified according to the binary outcome variable Y such that there are N_1 applicants whose loans have been approved, and N_0 applicants whose loans have been denied; $N_0+N_1=N$. We suppose that the population of N individuals is, or is regarded as, a random sample from the underlying joint data distribution. All subjects are also classified by the stratum covariate race; we assume N_N nonminority applicants and N_M minority applicants with $N_N+N_M=N$. We denote by N_{Nj} the number of nonminority applicants with Y=j (j=0,1). We likewise define N_{Mj}. The outcome variable and the race covariate now stratify the applicant population.

Given this breakdown, we suppose that a stratified sample of size n is taken in which n_{N0}, n_{N1}, n_{M0}, and n_{M1} subjects are randomly selected from the N_{N0}, N_{N1}, N_{M0}, and N_{M1} available applicants in each of the defined strata and values x_{ijk} of a K-dimensional covariate vector are measured ($k=1,...,n_{ij}$; $i=N,M$; $j=0,1$); $n=n_{N0}+n_{N1}+n_{M0}+n_{M1}$. The stratified sample is taken to improve information content by taking account of data characteristics. We assume that the logistic regression model describes the association between outcome and covariates in the source population:

$$\Pr(Y=j|X=x) = \frac{\exp(jx^T\beta)}{1+\exp(x^T\beta)}, \qquad (1)$$

where x incorporates an intercept with coefficient β_0, and the coefficient for the variable of interest (here DM) is denoted β_r. There are two goals. First, for a given n, we desire efficient estimation of the regression coefficients by appropriately choosing n_{N0}, n_{N1}, n_{M0}, and n_{M1}. We define efficiency in terms of mean squared error relative to the β_r value that would have been obtained by fitting a logistic regression model with the same covariates to everyone. Secondly, in line with the practice of examining for disparate treatment as a test of statistical significance, we wish to choose the sample strata sizes, given n, so as to approximate as accurately as possible the decision that would have been obtained for this hypothesis test from the population logit analysis. In our case the null hypothesis of interest is H_0: $\beta_r=0$ against the alternative hypothesis H_A: $\beta_r>0$; the one-sided alternative reflects the belief that we are testing for discrimination towards the minority group[1].

[1] We are aware that care needs to be taken in binary response models when testing composite hypotheses, as classical (consistent) tests may have power that goes to zero in finite samples; Savin and Würtz (1999). It is our belief that this is not the case for the particular hypothesis we are examining, though this problem may occur with other coefficients in fair lending binary response models. Of particular interest, the power function goes to zero when testing the null hypothesis of a zero slope coefficient when the corresponding regressor has matching signs with the values of one of the other regressors; one such case is when the regressor is positive and there is an intercept in the model.

An attractive feature of using the logistic model under stratified sampling or choice-based sampling is that the sampling scheme can be ignored and (1) estimated using a standard logit program as if the data were collected using simple random sampling. That is, when the model contains a constant (intercept) term for each category, these intercept terms are the only coefficients affected by stratified sampling. The usual program output gives estimates of the non-intercept coefficients that are maximum likelihood; i.e., the estimators, under the appropriate regularity conditions, are consistent and asymptotically normal. Further, the standard approach consistently estimates the standard errors, which implies that the usual statistic for testing statistical significance is an asymptotic standard normal variate when the null hypothesis is valid. See Anderson (1972), Prentice and Pyke (1979), Cosslett (1981a,b).

However, the usual estimators of the intercept terms (which include stratum constants), as well as their corresponding elements in the variance-covariance matrix, are not consistent: Prentice and Pyke (1979), Scott and Wild (1986, 1991, 1997). That is, standard logit packages will not consistently estimate β_r as the inclusion of DM as a covariate, given our use of data stratified by race, allows for a separate intercept term for each race stratum; i.e., (1) is equivalently:

$$\Pr(Y=j|X=x,\ DM=i) = \frac{\exp[\,j(\beta_0 + i\beta_r + x_s^T\beta_s)]}{1+\exp(\beta_0 + i\beta_r + x_s^T\beta_s)}, \qquad i,j=0,1 \qquad (2)$$

where x_s is a (K-2)-dimensional covariate vector consisting of the columns in x excluding the two columns associated with the intercept parameter β_0 and the race dummy parameter β_r, and β_s is a correspondingly defined parameter vector. The logistic regression (2) implies an intercept for nonminority applicants $\beta_N=(\beta_0+\beta_r)$ and an intercept for minority applicants of $\beta_M=\beta_0$. The standard logit estimators of β_N and β_M can be corrected for their asymptotic bias if the population proportions are known, as it often is in fair lending studies.

Specifically, let b_N and b_M be, respectively, the usual logit estimators of β_N and β_M, with corresponding appropriate asymptotic variance estimators, $\hat{V}(b_N)$ and $\hat{V}(b_M)$. Consistent estimators of β_N and β_M, which we denote as $\hat{\beta}_N$ and $\hat{\beta}_M$ are (e.g., Scott and Wild, 1991, p501):

$$\hat{\beta}_N = b_N - \ln[(n_{N1}/N_{N1})/(n_{N0}/N_{N0})]$$

$$\hat{\beta}_M = b_M - \ln[(n_{M1}/N_{M1})/(n_{M0}/N_{M0})] .$$

Then, we can consistently estimate β_r by

$$\hat{\beta}_r = b_r - \ln[(n_{N1}/N_{N1})/(n_{N0}/N_{N0})] + \ln[(n_{M1}/N_{M1})/(n_{M0}/N_{M0})] \qquad (3)$$

where b_r is the standard logit estimator of β_r. Further, let $\hat{V}(\hat{\beta}_r)$ be the usual estimator of the asymptotic variance associated with b_r. Now

$$\hat{V}(\hat{\beta}_N) = \hat{V}(b_N) - n_{N1}^{-1}(1 - n_{N1}/N_{N1}) - n_{N0}^{-1}(1 - n_{N0}/N_{N0})$$

and

$$\hat{V}(\hat{\beta}_M) = \hat{V}(b_M) - n_{M1}^{-1}(1 - n_{M1}/N_{M1}) - n_{M0}^{-1}(1 - n_{M0}/N_{M0})$$

are, respectively, consistent estimators of the asymptotic variance of $\sqrt{n}(\hat{\beta}_N - \beta_N)$ and $\sqrt{n}(\hat{\beta}_M - \beta_M)$, from which we can construct a consistent estimator, denoted as $\hat{V}(\hat{\beta}_r)$, of the asymptotic variance of $\sqrt{n}(\hat{\beta}_r - \beta_r)$. These results imply that, with stratified sampling by race covariate, the t-ratio statistic to test the significance of the disparate treatment dummy variable coefficient should be formed as $t_r = \hat{\beta}_r / \sqrt{\hat{V}(\hat{\beta}_r)}$ not as $t_r = b_r / \sqrt{\hat{V}(b_r)}$; the latter is not an asymptotic standard normal variate under the null hypothesis of interest.

The inconsistency of b_r appears to have been missed by fair lending researchers. For instance, Harrison (1998, p34) conjectures, "The sample of blacks and Hispanics in the database is selected in a non-random and exhaustive manner...Although this causes no bias in the estimation procedure...” Munnell et al. (1996, p32) erroneously state “The logit model produces consistent estimates of the standard errors and efficient estimates of the coefficients...” It is clear that consistency is not assured, and efficiency depends on the form of the stratified sample. Several authors recognize that stratifying will affect estimation of the constant term, but then fail to realize that the inclusion of a racial group dummy variable results in separate stratum constants. In particular, reference is made to the discussions in Maddala (1983, pp90-91) and Maddala (1991, pp792-793), which relate to stratifying by outcome only; we need to extend the results when we also stratify by a dummy variable covariate.

3. Use of logistic regressions in fair lending analysis

If minority applicants with the same credit profile as non-minority applicants face a higher probability of denial, or, alternatively, a more stringent underwriting standard, then disparate treatment and discrimination exist. Previous research testing for evidence of discrimination in lending includes Munnell et al. (1992, 1996), Calem and Stutzer (1995), Stengel and Glennon (1999), Courchane, Golan and Nickerson (2000), Longhofer and Peters (1999) and others. Much of the debate in the discrimination literature concentrates on the issue of whether discrimination exists due to profit-motivated statistical discrimination or due to a Beckerian taste for discrimination (see Becker, 1993). Evidence for possible statistical discrimination in mortgage lending is presented in Munnell et al. (1996), known as the “Boston Fed” study. Even though this paper is frequently cited, there is ongoing debate whether these results are statistically meaningful as the race effects are highly sensitive to model and variable specification (e.g., Horne, 1997; Harrison, 1998; Stengel and Glennon, 1999). Ladd (1998, p59) claims “...While it is not clear whether the discrimination that emerges from the Boston Fed study is attributable to a taste for discrimination or to profit-motivated statistical discrimination, my guess is that a substantial part of it is statistical discrimination driven by the drive for profits. If so, market forces are not likely to eliminate it”. Recently, Heckman (1998) contributed to the debate by pointing out the distinction between the macro-level and the micro-level discrimination. Most of the previous research in lending discrimination concentrates on statistical, micro-level discrimination.

There remain unresolved modeling issues in the discrimination literature (e.g., Yinger, 1998; Ross and Yinger, 1999; Longhofer and Peters, 1998; Heckman, 1998; Ladd, 1998; and Courchane, Golan and Nickerson, 2000). The issue reflects the decision that must be made as to how best to represent the approval decision process at the bank level, and which econometric or statistical modeling approach is best able to capture differences in treatment. These choices may vary from bank to bank and are closely related to the availability of data.

The most widely employed statistical procedure used to model the bank's lending decision is an unordered discrete choice (logistic) modeling approach. The decision to approve or deny a loan from an applicant is based, primarily, on the individual's credit but may also include demographic, economic, and property-specific attributes. It is generally argued in the literature that the decision model should reflect the probability that an applicant will default – a conceptual framework that underlies the design of most mortgage credit scoring models. However, in most cases, the approval process involves judgmental decisions made by underwriters using established policy guidelines that are qualitatively related, but not quantitatively linked, to the likelihood of default. For example, it is generally accepted that the higher the debt-to-income ratio, the greater the likelihood of default (a qualitative relationship). However, few banks know what impact an increase in the total debt-to-income ratio from 32% to 36% (or 48%) has on the likelihood of default (a quantitative relationship). Under this type of underwriting process, it is possible that the underwriting (judgmental) guidelines may introduce differences in treatment of the different applicants. This may lead to a violation of the fair lending laws. For that reason, the purpose of the statistical model is not to determine the optimal weights an underwriter should use to assess the creditworthiness of the applicant, but rather, to determine if the (pre-determined) underwriting guidelines are being fairly applied. These models test the hypothesis that minority applicants with profiles (e.g., credit, employment, wealth, etc.)

similar to nonminority applicants face the same likelihood of approval. The statistical models, in this case, should be designed to assess the relative importance (beyond that associated with random chance) of any observed difference in the likelihood of approval for these different racial (minority) groups. We contribute to this goal by providing information on sample design and estimator choice to assist in determining this difference as accurately as possible within the commonly used logistic framework.

4. Monte Carlo design

One way to arrive at model (1) is to define an underlying (continuous) response variable Y_t^*

$$Y_t^* = x_t^T\beta + \varepsilon_t \qquad (3)$$

where $\{\varepsilon_t\}$ is i.i.d. logistically distributed with $E(\varepsilon_t)=0$ and $Var(\varepsilon_t)=\pi^2/3$. The binary random variable is then defined as $Y_t=1$ if $Y_t^* \geq 0$, and $Y_t=0$ otherwise. In this formulation $x_t^T\beta$ is termed the index function and Y_t^* the latent or hidden variable. The assumption of zero for the threshold is innocuous when the model contains a constant term.

Our Monte Carlo design uses the latent variable index function representation (3) to generate the data. We consider three population data generating processes (DGPs) to illustrate the impact of design matrix choice on the results; we denote these as DGP1, DGP2 and DGP3. For each DGP we simulate an underlying population of 10,000 applicants and then sample n=400, 1200 and 2400 according to one of either six stratified sampling designs; we denote the sample designs as S1 to S6. We repeat this for 2000 replications. The exception is for DGP2 for which we consider only n=400 due to the very small number of nonminority denied loans for this population. In each case we investigate a situation of no discrimination (NDIS), which corresponds to $\beta_r=0$, and of discrimination (DIS) towards minority applicants, which occurs when $\beta_r>0$. We report results when β_r is estimated using the inconsistent estimator b_r and the consistent estimator $\hat{\beta}_r$. So, the design of our experiment involves one hundred and sixty-eight basic situations.

Denote b_r(n, Sg, DGPl, G) as the usual, inconsistent, estimator of β_r for sample size n, sample design Sg (g=1,...,6), DGPl (l=1,2,3), disparate treatment outcome G=NDIS or DIS, and denote $\hat{\beta}_r$ (n, Sg, DGPl, G) conformably. We use our simulation study to estimate the bias, variance and mean squared error (MSE) of b_r(n, Sg, DGPl, G) and $\hat{\beta}_r$ (n, Sg, DGPl, G) relative to the value for β_r that would have been obtained by fitting the logistic regression model to the population of 10,000 applicants. We also test the null hypothesis H_0: $\beta_r=0$ against the alternative hypothesis H_a: $\beta_r>0$ for each scheme to enable us to estimate the rejection frequencies (associated with a 5% significance level) as the proportion of trials for which the observed t-ratio associated with b_r(n, Sg, DGPl, G), and $\hat{\beta}_r$ (n, Sg, DGPl, G) , is greater than the standard normal critical value of 1.645. The simulation sampling error for these rejection frequencies can be determined by noting the binomial nature of the empirical rejections. So, for example, the standard error associated with a rejection proportion of 0.023 is $\sqrt{0.023*(1-0.023)/2000} \approx 0.003$.

We now detail the three DGPs followed by the six sampling designs.

4.1. DGP descriptions

The key differences between the three DGPs are the form of the design matrix and the correlation between DM and the other variables in the covariate matrix.

4.1.1. DGP1

We generated data for the latent variable as

$$Y^{*} = \beta_0 + X_1 + X_2 + X_3 + X_4 + X_5 + \beta_r DM + \varepsilon \quad (3)$$

with $X_1 \sim N(0,1)$; $X_2 \sim U[0,2]$; $X_3 \sim N(0,2)$; $X_4 \sim U[0,4]$; $X_5 \sim N(0,3)$ and N and U denote normal and uniform random variates respectively. The X's are drawn independently from each other and from ε, and the race dummy variable is randomly assigned with DM=1 if a uniform random variate on the zero-one scale is greater than 0.7. Windmeijer (1995) also uses this DGP in his study of goodness-of-fit measures in binary logit models, aside from the inclusion of DM. The design matrix for DGP1 bears no resemblance to any used (to our knowledge) by regulatory agencies in checking for fair lending, though it is useful for two reasons. First, it provides information on the impact of stratified sampling design on an arbitrary design matrix as opposed to one used in a fair lending case. Secondly, the race dummy variable is orthogonal to the other covariates; this is not the case for DGP2 and DGP3 and so we can use DGP1 to assess the qualitative impact of correlation between the race dummy variable and the other covariates.

We set $\beta_r=0$ for our 'no disparate treatment' event and we set its value to two for the discrimination situation. The value of β_0 is then used to control the proportion of approved and denied loans in the population; we set the population denial rate at 0.30, resulting in $\beta_0=-1.97505$ for nondiscrimination and $\beta_0=-3.297$ for the discrimination case. Tables 1 and 2 provide the population values for N_{N0}, N_{N1}, N_{M0}, and N_{M1} as well as the population denial ratios across the two race categories.

4.1.2. DGP2

We generated data for this DGP to approximate that used by the OCC in their fair lending examinations of national banks. Each bank considers a wide-range of decision variables in deciding on conventional mortgage loan applications, some of which are common across banks for conformance to the secondary market, but many are bank-specific. As it is not feasible to include them all, we limited our attention to one bank, which we denote as Bank A, and three continuous valued variables income (INC), debt-to-income ratio (DTI), loan-to-value ratio (LTV) along with two dummy variables D1 and DM. The binary variable D1 is included to represent various bad credit variables.

Our examination of Bank A's data, collected by the OCC, indicated that the distribution of values for INC, DTI and LTV differs considerably across race stratum. For our study we joined Blacks and Hispanics to form the minority group, though we recognize the associated limitations. Accordingly, we generated separate race data for these three variables; so DM is correlated with INC, DTI and LTV, which differs from DGP1. For minority applicants we assumed INC ~ lognormal(3.15, 0.52), LTV ~ U[85,100] and DTI ~ U[30,60], while we generated INC ~ lognormal(3.94, 0.36), LTV ~ U[85,95] and DTI ~ U[20,55] for the nonminority cases. Table 3 presents the sample correlation coefficients for the raw Bank A data and those we generated; these statistic values suggest we are reasonably capturing the characteristics of the real data, though we recognize the simplicity of this measure. One feature we are ignoring is the correlation pattern between the covariates; the impact of this on our results remains for future work.

We set N_M=2500 and N_N=7500, and we randomly selected twenty percent of minorities and nonminorities to each satisfy D1=1, our proxy bad credit variable. Finally, we specified the latent variable DGP as

$$Y^* = \beta_0 + 0.2INC - 0.03LTV - 0.12DTI - 0.5D1 + \beta_r DM + \varepsilon \qquad (4)$$

with β_0=4.65, β_r=0 for the nondiscrimination scenario, and β_0=4.05, β_r=0.8 for the discrimination case. These choices resulted in a population denial ratio of 0.10, similar to that for Bank A. The resulting values for N_{N0}, N_{N1}, N_{M0}, and N_{M1} and corresponding denial ratios for the race stratum are given in Tables 1 and 2. As the value of N_{N0} is small we could not undertake our experiments for several of the sampling designs for n>400. Hence, we generated results only for n=400 for DGP2.

4.1.3. DGP3

Our third DGP, DGP3, is a modified version of DGP2 to enable each of our sampling designs to be feasible for n=1200 and n=2400. Specifically, to give enough nonminority denials we changed the distributional assumptions for nonminorities to: INC ~ lognormal(3.55, 0.45), LTV ~ U[85,95] and DTI ~ U[25,55]. We also modified β_0 and β_r as follows: β_0=4.60 and β_r=0 for nondiscrimination, and β_0=4.45 and β_r=0.9 for discrimination. A rise in the population denial ratios resulted, except for minorities, which we detail in Tables 1 and 2.

4.2. Stratified Sampling Designs

Stratification jointly by outcome and covariate enhances efficiency compared with stratification based on outcome or covariate alone. Choosing sampling proportions that differ from the population may further improve precision, but may lead to estimation bias. We illustrate the potential trade-offs involved by studying six sampling designs, denoted as S1 to S6; they differ by balance and sample bias.

We define a sampling design to be balanced by outcome when $n_0=n_1$; i.e., there are equal numbers of approved and denied loan applicants in the sample. We say the design is balanced by covariate when $n_N=n_M$. Further, we denote a sample design Sg (g=1,...,6) as an unbiased sample when its sample denial odds-ratios are equal to the population denial odds-ratios; i.e., $d_i=D_i$, where $D_i=N_{i0}/N_i$ and $d_i=n_{i0}/n_i$, i=N,M.

Several studies suggest that efficiency gains may be obtained by using a balanced sampling design, while adopting an unbiased sampling design typically reduces estimation bias.

Anderson (1972, p34) suggests that balancing (by outcome) is a reasonable choice for logistic regressions. He writes "It is conjectured that for a given total sample size n, samples with balance give better estimates, on average, than those with unbalance".

Kao and McCabe (1991) determine optimal sample allocations based on minimizing the asymptotic expected error regret when sampling by outcome variable only. Expected error regret is the difference between the expected misclassification probability using a specified estimation procedure and the misclassification probability that would be obtained if all parameters were known. They show that balanced sampling minimizes the asymptotic expected error regret for logistic regressions when the population is also balanced, but that marginally unbalanced samples are preferable otherwise, with the optimal allocation depending on the number of covariates in the logit model. Their results provide guidance on specifying n_0 and n_1, but not on stratification allocations across race for a given n_0 and n_1. It would be interesting to extend their analytical analysis to stratification by outcome and covariate, though this is beyond the scope of this paper.

Breslow and Chatterjee (1999) advocate choosing sampling fractions that approximately result in equal numbers per stratum, in their work on the benefits of nonparametric maximum-likelihood estimation of the logistic regression. They illustrate the efficiency gains using data from the U.S. National Wilms Tumor Study.

The study by Scheuren and Sangha (1998) closely aligns with our research. They present results on two different sampling designs from a simulation study designed to represent the general attributes of a typical mortgage portfolio. They find that it is preferable to balance by loan decision and racial group such that there is an equal number of approved and denied loan files across racial groups. However, their data generating process is misspecified, and so caution is needed when interpreting their results; we discuss the specification error in their DGP in

Appendix 1. Further, Scheuren and Sangha do not adjust their estimate of β_r for stratification to ensure consistent estimation. We present corrected and uncorrected results.

The description of our six sampling designs follows:

1. S1 is balanced across outcome and race, and is a biased sample design. This is the design advocated by Scheuren and Sangha (1998), with $d_N = d_M$ and $n_{ij} = n/4$; i=N,M, j=0,1.
2. S2 is an unbiased sample design with the sample denial odds-ratios equal to the population denial odds-ratios. For DGP1, S1 is balanced across race but unbalanced across loan decision, while for DGP2 and DGP3, this sample design is unbalanced across outcome and race covariate.
3. S3 is unbalanced across outcome, balanced across race and is a biased sample design. This is the other sampling design studied by Scheuren and Sangha (1998). Here there are 50% more minority denials than approvals and an equal number of nonminority approvals and denials.
4. & 5. S4 and S5 are biased sample designs; both are based on the results from Kao and McCabe (1991) to determine n_0 and n_1. We use their Table 1 (p435) to set the sample ratio (n_1/n) for a given population proportion of (N_1/N) and number of covariates K. Kao and McCabe suggest (n_1/n) be set at 0.528 for DGP1, 0.585 for DGP2, and 0.575 for DGP3. Their analysis, however, offers no assistance in determining optimal allocations between racial groups. We maintain the population allocations between nonminorities and minorities, given n_i (i=N.M) for sampling design S4, while we set $n_{i0} = n_{i1}$ for sampling design S5. Hence, sample design S4 is unbalanced by race and loan outcome, while S5 is balanced by race stratum but unbalanced across loan decision.
6. The optimal ratios reported by Kao and McCabe are close to 0.5, which suggests that this may be a reasonable practical approximation. Accordingly in S6 we set $n_0 = n_1$, but choose the race

allocations to reflect those in the population; S6 is then unbalanced with respect to racial group, balanced with respect to the outcome variable and is a biased sample design.

5. Simulation results

5.1 Bias

Bias results from the estimated sampling distributions are shown in the third columns of Tables 4 to 15. By comparing these biases, we draw the following conclusions:

(i) The bias for S2, the unbiased sample design, is substantially smaller in magnitude than that for the biased sample designs, in particular for S1, S3 and S5, when using the inconsistent estimator b_r. The design matrix affects the magnitude of the bias; compare DGP1 and DGP2 for instance. These features are qualitatively similar for cases of discrimination and nondiscrimination. It is clear that it is preferable to use the unbiased sample design S2 when the usual inconsistent estimator of β_r is used. In terms of the biased sample designs our results do not support, when attempting to minimize estimation bias with b_r, use of balancing by covariate (S1, S3 and S5); it is better to maintain the population covariate ratios in the sample (S2, S4 and S6). While the estimation bias with S4 and S6 are generally similar, the former is to be slightly favored.

(ii) Correcting for asymptotic bias substantially reduces the estimation bias for the biased sample designs. We observe that the unbiased sample design S2 still has the smallest estimation bias when there is discrimination, but when there is no discrimination, other sample designs are sometimes preferable; in the latter case the bias results can be quite sensitive to the form of the design matrix. Generally, there is less estimation bias for S5 among the biased sample designs that balance by race (S1, S3, S5), and S4 results in smaller estimation bias when comparing S4 and S6, the two biased sample designs with sample race allocations that mirror those in the population. Sample designs S4 and S5 typically exhibit similar estimation biases, when using $\hat{\beta}_r$, though the latter is to be marginally preferred.

5.2 Variance

Estimated variances are shown in the fourth columns of Tables 4 to 15. We observe that:

(i) There is little difference between the estimated variances for $\hat{\beta}_r$ and b_r for our scenarios.

(ii) Choosing sample designs that balance by race (S1, S3, S5) results in significant gains in estimation precision. The variances for these sample designs is often half that obtained with those sample designs (S2, S4, S6) that have sample race allocations proportional to that in the population. Balancing by outcome, even approximately, is helpful in obtaining more precise estimates of β_r (e.g., often S4 and S6 are preferred to S2), but it is balancing by race that produces the large gains in precision (e.g., S5 always outperforms S4). At least within sampling variation, typically the variance associated with S2 is highest among the sampling designs we examined.

(iii) Not surprisingly, the precision gains are relatively greater for DGP2 and DGP3 than for DGP1, because the characteristics of the strata for the former DGPs were constructed to differ across race while those for DGP1 were generated orthogonal to race.

(iv) There is little to choose between S1, S3 and S5 in terms of estimation variance.

5.3 MSE

(i) When using b_r, the inconsistent estimator of β_r, the bias distortions noted in section 5.1 typically dominate any potential variance gains discussed in section 5.2. Regardless of the form of DGP, the sample size or the population value of β_r, MSE is smallest for the sampling designs that mirror population covariate racial fractions in the sample (S2, S4, S6). However, an exception arises with DGP1 with nondiscrimination. Then, there is little difference between the MSEs for S1, S2, S4, S5 and S6, though use of sample design S3 is never favored. For the other cases, the choice between S2, S4 and S6 is sensitive to the form of the DGP, the sample size and the value of β_r, though at least within sampling variation S4 dominates S6. Sample design S2 dominates S4 for DGP1, for DGP2 with nondiscrimination and for DGP3 with discrimination. However, S4 dominates S2 for DGP2 with discrimination, for DGP3 with nondiscrimination when n=400, and the two designs produce similar MSEs for DGP3 with nondiscrimination when n=1200 or n=2400.

(ii) Correcting for asymptotic bias generally results in smaller MSEs for sample designs S1, S3 and S5, which balance across race. Regardless of the form of DGP, S5 is preferred to both S1 and S3, at least within sampling variation. An exception is that sample design S3 dominates S1 and S5 with DGP2 and nondiscrimination. However, it is clear that our simulations would not support adopting S1, the design advocated for by Scheuren and Sangha (1998). Generally, sample designs S4 and S6 produce smaller MSE than sample design S2, though exceptions include DGP1, and DGP3 with discrimination. Sample design S4 is typically favored to S6. It is clear that balancing, or near balancing, by outcome is useful, but not as crucial as approximately balancing by race.

(iii) The magnitudes of the MSE differences between the sampling designs are greater for DGP2 and DGP3 (and especially for DGP2) than for DGP1. These results accord with our prior expectation as the gains (and losses) from particular stratification patterns will be larger the more marked the differences between the strata.

Our results suggest the following practical prescriptions. First, when using the estimator b_r, the choice of sample design should reflect the race allocation in the population; balancing by covariate race is not recommended because of the estimation biases that can arise under this sampling strategy. Of the three sampling designs we investigate that satisfy this requirement, our recommendation is to use sample design S4, as there is little loss in MSE in using S4 when S2 is preferred, but there can be significant gains in MSE when S4 is favored. At least within sampling variation, sample design S4 always dominates sample design S6.

Second, if the corrected estimator $\hat{\beta}_r$ is adopted, then our results suggest a clear preference for sample designs S1, S3 and S5; that is, those sample designs that balance by race. Correcting for asymptotic estimation bias significantly reduces the finite-sample biases and allows the variance gains in these stratified sample designs to dominate. Of these three sample designs, our recommendation is to use sample design S5.

5.4 t-ratios

The tables report two summary statistics for the sampling distribution of t-ratios for testing H_0: β_r=0 versus H_A: β_r>0. For each of our scenarios we examine: (i) the mean $\bar{t}_r$ of the 2,000 trial values of t_r, and (ii) the rejection frequencies associated with a nominal one-sided 5% significance test with the (asymptotic) critical value obtained from a standard normal distribution. The first measure, $\bar{t}_r$, provides an indication of central tendency of the sampling distributions; we would like this value to be near to the t-ratio obtained from fitting a logit model to the loan applicant population. The rejection frequencies are empirical "pseudo powers", as they are not size-adjusted. That is, while these frequencies represent the actual ability of the test, associated with a particular sample design and estimator, to reject a false null hypothesis, we cannot strictly state that any one method has greater (true) power than any other procedure because the two approaches do not have equal finite-sample size or an appropriate ranking of size has not been possible. Further, we need to be mindful when interpreting the nondiscrimination rejection frequencies; they are not empirical sizes because the population value of β_r is not identically zero. Ideally, assuming that the "power" shapes are orthodox, we expect the nondiscrimination frequencies to be close to 0.05, allowing for finite-sample approximations and Monte Carlo sampling errors, while the discrimination rejection probabilities should be near as possible to one.

We remind the reader that the t-ratios formed using the uncorrected estimator of β_r, b_r, are not distributed as asymptotic standard normal variates under the null hypothesis because of the inconsistency of b_r. This explains some of the observed features, with other characteristics being due to the approximation error in using the standard normal critical value, differences arising from the stratified sampling designs, estimation biases, and sampling errors arising from the Monte Carlo analysis.

The following general results are apparent from the observed sampling distributions of the t-ratios. First, there is a tendency for "under-sizing"; i.e., we are not rejecting a true null as often as desired with the chosen significance level of 5%. This implies, in terms of the disparate treatment question, that we will conclude nondiscrimination in some cases that should be rejected. The critical value needs to be smaller than that associated with a standard normal distribution. We can conclude this feature from our results by noting that we present cases whose β_r value is marginally greater than zero. So, given the form of our alternative hypothesis and our decision rule, if our procedures had "true" size approximately equal to nominal size, we expect powers marginally greater than 0.05, assuming orthodox "power" shapes. However, we observe many pseudo powers well below 0.05, which suggests under-sizing. An exception is for sample design S3 with DGP1 for which there appears to be an over-rejection problem arising from positive bias in estimation; we conclude discrimination far often than desired.

Second, when using the inconsistent estimator b_r, the sample designs that balance by covariate (S1, S3 and S5) will often not detect discrimination when it is present (e.g., Tables 9 and 13). It is clear, when using b_r, that sample designs S2, S4 and S6 have greater ability to reject a false null, with our results suggesting a slight preference for S4 and S6.

Third, when using the consistent estimator $\hat{\beta}_r$, sample designs S1, S3 and S5 dominate those that do not balance by covariate (S2, S4 and S6). As our powers are not size-adjusted, we are not able to recommend one strategy over another, and our results do not suggest a clear winner in terms of size-uncorrected power.

6. Some illustrations

In this section we analyze some data for disparate treatment using the consistent estimator of β_r as well as the usual logit estimator b_r. It would be interesting to examine the impact of sample design, but this is infeasible given the available sample data. We show that there can be changes in the discrimination outcome once we correct the logit estimator for the stratified sample design.

We examine sample data supplied by the OCC for three banks, which we denote as Bank A, Bank B and Bank C; note that the data for Bank A is not that used in designing DGP2 as outlined in section 4.1.2. The OCC adopted stratified sample designs when collecting the data that were unbalanced across outcome and across racial groups, and they oversampled the denials in all groups. Bank A and Bank B each have two minority groups; we denote them as MA and MB respectively. Bank C has only one minority group, which we denote by M. Let β_{ri} be the parameter for the minority dummy variable; for Bank A and Bank B there are two such parameters; i=M, MA or MB.

Our interest is in examining the validity of the null hypothesis H_0: $\beta_{ri}=0$ versus the one-sided alternative hypothesis H_A: $\beta_{ri}>0$. We estimated bank-specific logistic regressions for the three banks using the inconsistent and consistent estimators of β_{ri} and obtained asymptotic P-values for the usual t-ratio of H_0, assuming a limiting standard normal null distribution. For confidentiality reasons we are unable to report details of the specific regressions, and we have also altered the OCC logit model specifications. That is, our regressions deliberately do not replicate actual regressions undertaken by the OCC. Consequently, any findings of disparate treatment should not be interpreted as evidence to suggest that these particular banks are discriminating against any racial group.

We provide the P-values in Table 16 for examining H_0 against H_A using the inconsistent estimator and that modified for the stratified sampling design. The results, though limited, indicate the impact of accounting for the sample design. Assuming a classical 5% significance level, there is no statistical support for disparate treatment by the banks when we use the standard, inconsistent, logit estimator, except for Bank B's minority group B. However, when using $\hat{\beta}_r$, this outcome changes for minority group B for Bank A, and for Bank C. The examples highlight the fact that established results based on the assumption of simple random sampling need to be reconsidered when the sample is stratified by outcome and covariate in the manner we have been investigating.

7. Concluding remarks

The results we present have some important implications for the use of logistic regressions in fair lending studies that examine for discrimination as a test of significance of a racial group dummy variable, when the sample data are obtained from samples of population data that has been stratified by loan outcome and racial group. Our results show the importance of using a consistent estimator of the disparate treatment parameter and the impact of the form of the stratified sample design on the finite-sample sampling distributions of two estimators of the parameter and the t-ratio for statistical significance.

We can make several practical recommendations from our results. First, if a standard logit package is used to estimate the disparate treatment binary variable parameter without correcting for asymptotic bias, then it is clear that sample designs that maintain the racial population allocations dominate those designs that balance sample racial numbers. Proportional racial allocations sampling schemes do not suffer as extreme estimation bias, and even though there is a loss in precision in using these sampling designs, there is still a gain in mean squared error. Further, we observed that the test of significance for these sample designs is more likely to reflect the outcomes we desire, though there is evidence to suggest under-sizing for all of the sample designs we investigated. Overall, our experiments suggest that sample design S4 is a good choice when using the usual inconsistent estimator.

Second, if we correct the estimator for asymptotic bias, then our recommendations change, and those sample designs that balance by covariate are preferred. Correcting for asymptotic bias significantly reduces the finite-sample estimation biases and we can benefit from the gains in precision that are possible when balancing by covariate. Mean squared errors substantially reduce, as expected, when we use a consistent estimator of a parameter. In addition, we obtain sampling distributions for the test of significance statistic that do not reflect bias distortions, and so lead to desired disparate treatment decisions. Of the three sample designs we investigate that balance by race, we recommend sample design S5, when using the consistent estimator of the disparate treatment parameter.

There are several extensions to our study worthy of further research. Irrespective of the estimator we consider, it is clear that there are gains in using the optimal sample outcome allocations proposed by Kao and McCabe (1991). We would anticipate that a similar approximate balance by covariate would be preferable to the strict balancing by race that we have investigated. Some further work in this direction would be interesting.

It would also be interesting to consider methods of determining an optimal sample size (n), according to some chosen criterion. Some approaches to sample size suggestions explicitly for logistic regression coefficients are Whittemore (1981), Self and Mauritsen (1988) and Bull (1993). Another potential direction for further research is on alternative estimation principles. We focused on parametric maximum likelihood estimators of the disparate treatment parameter, and there may be benefits in exploring nonparametric methods. Breslow and Chatterjee (1999), for instance, find nonparametric maximum likelihood resulted in efficiency gains for logistic regression coefficients in their analysis of data from the U.S. National Wilms Tumor Study.

The use of bootstrap or Monte Carlo testing may be fruitful as such techniques may assist in eliminating or, at least, minimizing the size-distortions that seemed to be a feature of our study. Finally, we have limited our attention to the coefficient on the race binary variable, which in reality may or may not result in race significantly affecting the probability of mortgage approval, as the logistic parametric specification is not additive in the variables. The marginal effect of race depends on all parameters in the model and the specific values assumed for the covariate vector, and will therefore be affected by the choice of sample design and the estimator we use for the stratum constants. It remains for future research to investigate these impacts.

Appendix 1.

We show that the DGP used in the simulation study by Scheuren and Sangha (1998) is misspecified; we hereafter refer to the authors as SS. SS use a proxy variable 'score' to represent the underwriting criteria applied by a bank in deciding on loan applications. The generating process for score is as follows: for approved loans (Y=1), score = min(800U[0,1] +400U[0,1],800) with additional random variation to force score $\in$[200,800], while for denied loans (Y=0), score = max(800U[0,1]-400U[0,1],0) with additional random variation to force score $\in$[0,600]. Note that the variable score is overlapping for denied and approved loans.

The simulated loan applications are assigned the race covariate DM in a random manner for the no discrimination case, and, given the variables score and DM, SS estimate the logistic regression

$$\Pr(Y=j|\ DM=i) = \frac{\exp[\, j(\beta_0 + i\beta_r + \beta_s score)]}{1+\exp(\beta_0 + i\beta_r + \beta_s score)}\ , \qquad i,j=0,1 \qquad \text{(A.1)}$$

which implies an underlying latent variable or index function model

$$Y^* = \beta_0 + \beta_s score + \varepsilon \qquad \text{(A.2)}$$

when β_r=0. The latent variable is such that $Y^* \geq 0$ for Y=1, and $Y^* < 0$ for Y=0, so that score cannot be overlapping for the error term to have a standard logistic distribution with zero mean and the parameter values to be fixed constants. That is, for score to have overlapping values for denied and approved loans, DGP (A.2) must be misspecified, either from omitted variables or from functional form errors. The specification error results in inconsistent estimation of all parameters.

References

Anderson, J.A. (1972). Separate sample logistic discrimination, *Biometrika* 59, 19-35.

Becker, G. (1993). The evidence against banks doesn't prove bias, *Business Week*, April 19, 18.

Breslow, N.E. and Chatterjee, N. (1999). Design and analysis of two phase studies with binary outcome applied to Wilms tumour prognosis, *Applied Statistics* 48, 457-468.

Bull, S.B. (1993). Sample size and power determination for a binary outcome and an ordinal exposure when logistic regression analysis is planned, *American Journal of Epidemiology* 137, 676-684.

Carr, J.H. and Megbolugbe, I.F. (1993). The Federal Reserve Bank of Boston study on mortgage lending revisited, *Journal of Housing Research* 4, 277-313.

Calem, P. and Stutzer, M.(1995). The simple analytics of observed discrimination in credit markets, *Journal of Financial Intermediation* 4, 189-212.

Cosslett, S. (1981a). Maximum likelihood estimators for choice-based samples, *Econometrica* 49, 1289-1316.

Cosslett, S. (1981b). Efficient estimation of discrete-choice models. In *Structural Analysis of Discrete Data with Econometric Applications* (eds. Manski, C.F. and McFadden, D), Cambridge, Mass.: MIT Press, 51-111.

Courchane, M., Golan, A., and Nickerson, D. (2000) Estimation and evaluation of loan discrimination – an informational approach, *Journal of Housing Research* 11, forthcoming.

Courchane, M., Nebhut, D., and Nickerson, D. (2000) Lessons learned: application of statistical techniques to fair lending compliance examination, *Journal of Housing Research* 11, forthcoming.

Day, T.E. and Liebowitz, S.J. (1998). Mortgage lending to minorities: where's the bias?, *Economic Inquiry* 36, 3-28.

Kao, T-C. and McCabe, G.P. (1991). Optimal sample allocation for normal discrimination and logistic regression under stratified sampling, *Journal of the American Statistical Association* 86, 432-436.

Glennon, D. and Stengel, M. (1994). An evaluation of the Federal Reserve Bank of Boston's study of racial discrimination in mortgage lending. Office of the Comptroller of the Currency, Economic and Policy Analysis Working Paper 94-2.

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

Heckman, J. (1998), Detecting discrimination, *Journal of Economic Perspectives* 12, 101-116.

Horne, D.K. (1994). Evaluating the role of race in mortgage lending, *FDIC Banking Review* 7, 1-15.

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

Liebowitz, S.J. (1993). A study that deserves no credit, *Wall Street Journal*, September 1, A14.

Longhofer, S. and Peters, S. (1998), Self-selection and discrimination in credit markets, Federal Reserve Bank of Cleveland Working Paper 9809.

Longhofer, S. and Peters, S. (1999), Beneath the rhetoric: clarifying the debate on mortgage lending discrimination, Federal Reserve Bank of Cleveland, *Economic Review* 34, 2-13.

Maddala, G.S. (1983). *Limited Dependent and Qualitative Variables in Econometrics*. New York: Cambridge University Press.

Maddala, G.S. (1991). A perspective on the use of limited-dependent and qualitative variables models in accounting research, *The Accounting Review* 66, 788-807.

Munnell, A.H., Browne, L.E., McEneaney, J. and Tootell, G.M.B. (1992). Mortgage lending in Boston: interpreting HMDA data, working paper 92-7, Federal Reserve Bank of Boston.

Munnell, A.H., Tootell, G.M.B., Browne, L.E., and McEneaney, J. (1996). Mortgage lending in Boston: interpreting HMDA data, *American Economic Review* 86, 25-53.

Prentice, R.L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies, *Biometrika* 66, 403-411.

Ross, S. and Yinger, J. (1999). Does discrimination in mortgage lending exist? The Boston Federal study and its critics. In *Mortgage Lending Discrimination: A Review of Existing Literature* (eds. Turner, M. and Skidmore, F.), Washington, D.C.: The Urban Institute, 43-75.

Scheuren, F.J. and Sangha, B.S. (1998). Interaction of sample design and disparate treatment analysis in bank lending, mimeo., Policy Economics and Quantitative Analysis, Ernst & Young LLP.

Scott, A.J. and Wild, C.J. (1986). Fitting logistic models under case-control or choice based sampling, *Journal of the Royal Statistical Society, Series B* 23, 469-476.

Scott, A.J. and Wild, C.J. (1991). Fitting logistic regression models in stratified case-control studies, *Biometrics* 47, 497-510.

Scott, A.J. and Wild, C.J. (1997). Fitting regression models to case-control data by maximum likelihood, *Biometrika* 84, 57-71.

Self, S.G. and Mauritsen, R.H. (1988). Power/sample size calculations for generalized linear models, *Biometrics* 44, 79-86.

Stengel, M. and Glennon, D. (1999), Evaluating statistical models of mortgage lending discrimination: a bank-specific analysis, *Real Estate Economics* 27, 299-334.

Whittemore, A. (1981). Sample size for logistic regression with small response probability, *Journal of the American Statistical Association* 76, 27-32.

Windmeijer, F.A.G. (1995). Goodness-of-fit measures in binary choice models, *Econometric Reviews* 14, 101-116.

Yinger, J. (1998), Housing discrimination is still worth worrying about, *Housing Policy Debate* 9, 893-927.

Table 1: Population characteristics by outcome and race with no discrimination

(N_{Nj} & N_{Mj}, j=0,1)

DGP	Race Category	Loan Outcome		Denial Ratios D
		Approved (j=1)	Denied (j=0)	
DGP1	Nonminority	4900	2075	0.30
	Minority	2100	925	0.30
	Total	7000	3000	0.30
DGP2	Nonminority	7340	160	0.02
	Minority	1785	715	0.29
	Total	9003	997	0.10
DGP3	Nonminority	6498	1002	0.13
	Minority	1771	729	0.29
	Total	8269	1731	0.17

Table 2: Population characteristics by outcome and race with discrimination (N_{Nj} & N_{Mj}, j=0,1)

DGP	Race Category	Loan Outcome		Denial Ratios
		Approved (j=1)	Denied (j=0)	
DGP1	Nonminority	5392	1583	0.23
	Minority	1608	1417	0.47
	Total	7000	3000	0.30
DGP2	Nonminority	7374	126	0.02
	Minority	1628	872	0.35
	Total	9002	998	0.10
DGP3	Nonminority	6869	631	0.08
	Minority	1731	769	0.31
	Total	8600	1400	0.14

Table 3: Sample simple correlation coefficients between Bank A and generated data for DGP2

	Minorities	Nonminorities
INC	0.997	0.982
LTV	0.764	0.804
DTI	0.919	0.857

Table 4. Summary statistics for estimated sampling distributions of b_r (n, Sg, DGP1, NDIS), g=1...6, n=400,1200,2400. Population coefficient value is 0.100 with a t-ratio of 1.249.

Sampling Design	n	$\hat{Bias}$	$\hat{Var}$	$\hat{MSE}$	$\bar{t}_r$	Rejection Frequency
S1	400	-0.089	0.106	0.114	0.030	0.027
	1200	-0.086	0.030	0.037	0.064	0.026
	2400	-0.081	0.016	0.023	0.127	0.037
S2	400	-0.028	0.113	0.114	0.183	0.034
	1200	-0.036	0.034	0.035	0.291	0.053
	2400	-0.030	0.016	0.017	0.453	0.071
S3	400	0.318	0.107	0.208	1.089	0.259
	1200	0.309	0.032	0.127	1.907	0.636
	2400	0.312	0.016	0.113	2.730	0.900
S4	400	-0.041	0.118	0.120	0.143	0.033
	1200	-0.043	0.035	0.037	0.250	0.045
	2400	-0.039	0.017	0.019	0.386	0.055
S5	400	-0.058	0.106	0.109	0.111	0.034
	1200	-0.074	0.031	0.036	0.121	0.030
	2400	-0.073	0.016	0.021	0.180	0.041
S6	400	-0.038	0.116	0.117	0.151	0.034
	1200	-0.044	0.034	0.036	0.249	0.046
	2400	-0.038	0.017	0.018	0.390	0.062

Table 5. Summary statistics for estimated sampling distributions of b_r (n, Sg, DGP1, DIS), g=1...6, n=400,1200,2400. Population coefficient value is 3.020 with a t-ratio of 31.176.

Sampling Design	n	$\hat{Bias}$	$\hat{Var}$	$\hat{MSE}$	$\bar{t}_r$	Rejection Frequency
S1	400	-0.884	0.164	0.945	4.806	1.000
	1200	-0.956	0.048	0.962	8.353	1.000
	2400	-0.965	0.024	0.955	11.857	1.000
S2	400	0.161	0.235	0.261	6.150	1.000
	1200	0.047	0.068	0.070	10.758	1.000
	2400	0.025	0.030	0.031	15.255	1.000
S3	400	-0.406	0.190	0.355	5.609	1.000
	1200	-0.489	0.054	0.293	9.775	1.000
	2400	-0.503	0.026	0.279	13.871	1.000
S4	400	0.350	0.241	0.364	6.533	1.000
	1200	0.240	0.069	0.127	11.412	1.000
	2400	0.215	0.034	0.080	16.180	1.000
S5	400	-0.891	0.162	0.956	4.815	1.000
	1200	-0.976	0.047	1.000	8.320	1.000
	2400	-0.992	0.025	1.009	11.771	1.000
S6	400	0.358	0.247	0.375	6.515	1.000
	1200	0.250	0.073	0.136	11.388	1.000
	2400	0.227	0.033	0.085	16.145	1.000

Table 6. Summary statistics for estimated sampling distributions of $\hat{\beta}_r$ (n, Sg, DGP1, NDIS), g=1...6, n=400,1200,2400. Population coefficient value is 0.100 with a t-ratio of 1.249.

Sampling Design	n	$\hat{\text{Bias}}$	$\hat{\text{Var}}$	$\hat{\text{MSE}}$	$\bar{t}_r$	Rejection Frequency
S1	400	-0.049	0.106	0.108	0.135	0.036
	1200	-0.047	0.030	0.032	0.253	0.048
	2400	-0.042	0.016	0.018	0.400	0.075
S2	400	0.011	0.113	0.113	0.285	0.047
	1200	0.003	0.034	0.034	0.475	0.088
	2400	0.009	0.016	0.016	0.723	0.133
S3	400	-0.048	0.107	0.109	0.128	0.033
	1200	-0.057	0.032	0.035	0.199	0.038
	2400	-0.055	0.016	0.019	0.304	0.059
S4	400	-0.041	0.118	0.120	0.136	0.026
	1200	-0.043	0.035	0.037	0.238	0.033
	2400	-0.039	0.017	0.019	0.370	0.050
S5	400	-0.038	0.106	0.107	0.163	0.038
	1200	-0.042	0.031	0.033	0.280	0.048
	2400	-0.034	0.016	0.017	0.455	0.090
S6	400	-0.046	0.116	0.118	0.125	0.024
	1200	-0.051	0.034	0.037	0.204	0.032
	2400	-0.046	0.017	0.019	0.327	0.046

Table 7. Summary statistics for estimated sampling distributions of $\hat{\beta}_r$ (n, Sg, DGP1, DIS), g=1...6, n=400,1200,2400. Population coefficient value is 3.020 with a t-ratio of 31.176.

Sampling Design	n	$\hat{Bias}$	$\hat{Var}$	$\hat{MSE}$	$\bar{t}_r$	Rejection Frequency
S1	400	0.215	0.164	0.210	8.070	1.000
	1200	0.143	0.048	0.068	14.059	1.000
	2400	0.134	0.024	0.042	19.581	1.000
S2	400	0.164	0.235	0.262	6.795	1.000
	1200	0.050	0.068	0.071	11.852	1.000
	2400	0.028	0.030	0.031	16.587	1.000
S3	400	0.288	0.190	0.272	7.777	1.000
	1200	0.205	0.054	0.096	13.551	1.000
	2400	0.191	0.026	0.062	18.916	1.000
S4	400	0.343	0.241	0.359	7.119	1.000
	1200	0.233	0.069	0.123	12.396	1.000
	2400	0.208	0.034	0.077	17.347	1.000
S5	400	0.188	0.162	0.197	8.057	1.000
	1200	0.116	0.047	0.060	14.032	1.000
	2400	0.107	0.025	0.036	19.543	1.000
S6	400	0.369	0.247	0.383	7.142	1.000
	1200	0.261	0.073	0.141	12.432	1.000
	2400	0.238	0.033	0.090	17.399	1.000

Table 8. Summary statistics for estimated sampling distributions of b_r (400, Sg, DGP2, NDIS), g=1...6. Population coefficient value is –0.189 with a t-ratio of -1.212.

Sampling Design	n	$\hat{Bias}$	$\hat{Var}$	$\hat{MSE}$	$\bar{t}_r$	Rejection Frequency
S1	400	-2.865	0.312	8.520	-4.984	0.000
S2	400	-0.029	0.632	0.633	-0.198	0.011
S3	400	-2.480	0.306	6.456	-4.336	0.000
S4	400	-0.690	0.324	0.800	-1.360	0.001
S5	400	-2.881	0.318	8.618	-5.064	0.000
S6	400	-0.935	0.331	1.205	-1.771	0.001

Table 9. Summary statistics for estimated sampling distributions of b_r (400, Sg, DGP2, DIS), g=1...6. Population coefficient value is 1.224 with a t-ratio of 7.758.

Sampling Design	n	$\hat{Bias}$	$\hat{Var}$	$\hat{MSE}$	$\bar{t}_r$	Rejection Frequency
S1	400	-3.069	0.227	9.646	-3.283	0.000
S2	400	0.151	0.965	0.988	1.514	0.468
S3	400	-2.642	0.227	7.207	-2.492	0.000
S4	400	0.337	0.337	0.487	2.309	0.798
S5	400	-3.136	0.237	10.071	-3.429	0.000
S6	400	0.386	0.390	0.539	2.333	0.808

Table 10. Summary statistics for estimated sampling distributions of $\hat{\beta}_r$ (400, Sg, DGP2, NDIS), g=1...6. Population coefficient value is –0.189 with a t-ratio of -1.212.

Sampling Design	n	$\hat{Bias}$	$\hat{Var}$	$\hat{MSE}$	$\bar{t}_r$	Rejection Frequency
S1	400	0.046	0.312	0.314	-0.135	0.031
S2	400	-0.115	0.632	0.645	-0.333	0.016
S3	400	0.025	0.306	0.307	-0.177	0.024
S4	400	-0.053	0.366	0.369	-0.364	0.017
S5	400	0.030	0.318	0.319	-0.163	0.027
S6	400	-0.067	0.390	0.394	-0.378	0.019

Table 11. Summary statistics for estimated sampling distributions of $\hat{\beta}_r$ (400, Sg, DGP2, DIS), g=1...6. Population coefficient value is 1.224 with a t-ratio of 7.758.

Sampling Design	n	$\hat{Bias}$	$\hat{Var}$	$\hat{MSE}$	$\bar{t}_r$	Rejection Frequency
S1	400	0.376	0.227	0.368	3.128	0.903
S2	400	0.138	0.965	0.984	1.826	0.603
S3	400	0.398	0.227	0.385	3.129	0.905
S4	400	0.330	0.373	0.482	2.526	0.835
S5	400	0.309	0.237	0.332	3.042	0.884
S6	400	0.368	0.390	0.525	2.513	0.846

Table 12. Summary statistics for estimated sampling distributions of b_r (n, Sg, DGP3, NDIS), g=1...6, n=400,1200,2400. Population coefficient value is 0.084 with a t-ratio of 0.872.

Sampling Design	n	$\hat{Bias}$	$\hat{Var}$	$\hat{MSE}$	$\bar{t}_r$	Rejection Frequency
S1	400	-1.051	0.136	1.241	-2.270	0.001
	1200	-1.028	0.041	1.098	-3.964	0.000
	2400	-1.028	0.021	1.078	-5.646	0.000
S2	400	-0.127	0.214	0.230	-0.060	0.018
	1200	-0.127	0.063	0.065	-0.137	0.018
	2400	-0.123	0.029	0.044	-0.183	0.011
S3	400	-0.621	0.124	0.510	-1.268	0.000
	1200	-0.606	0.040	0.407	-2.208	0.000
	2400	-0.604	0.019	0.384	-3.137	0.000
S4	400	-0.135	0.173	0.191	-0.109	0.020
	1200	-0.135	0.052	0.070	-0.193	0.017
	2400	-0.138	0.026	0.045	-0.294	0.018
S5	400	-1.052	0.133	1.240	-2.293	0.000
	1200	-1.037	0.040	1.115	-4.029	0.000
	2400	-1.036	0.020	1.093	-5.731	0.000
S6	400	-0.110	0.179	0.191	-0.059	0.023
	1200	-0.118	0.052	0.065	-0.130	0.018
	2400	-0.125	0.027	0.043	-0.219	0.018

Table 13. Summary statistics for estimated sampling distributions of b_r (n, Sg, DGP3, DIS), g=1...6, n=400,1200,2400. Population coefficient value is 1.440 with a t-ratio of 13.691.

Sampling Design	n	$\hat{Bias}$	$\hat{Var}$	$\hat{MSE}$	$\bar{t}_r$	Rejection Frequency
S1	400	-1.366	0.146	2.012	0.160	0.050
	1200	-1.383	0.045	1.958	0.230	0.057
	2400	-1.385	0.021	1.939	0.324	0.066
S2	400	0.108	0.273	0.285	2.716	0.913
	1200	0.046	0.071	0.073	4.774	1.000
	2400	0.041	0.035	0.037	6.815	1.000
S3	400	-0.925	0.161	1.017	1.159	0.288
	1200	-0.947	0.048	0.945	2.007	0.662
	2400	-0.949	0.023	0.924	2.850	0.913
S4	400	0.310	0.216	0.312	3.477	0.993
	1200	0.233	0.068	0.122	6.037	1.000
	2400	0.212	0.031	0.076	8.538	1.000
S5	400	-1.416	0.143	2.148	0.045	0.037
	1200	-1.427	0.044	2.080	0.051	0.034
	2400	-1.432	0.021	2.072	0.044	0.036
S6	400	0.355	0.239	0.365	3.499	0.993
	1200	0.287	0.069	0.151	6.106	1.000
	2400	0.273	0.033	0.108	8.663	1.000

Table 14. Summary statistics for estimated sampling distributions of $\hat{\beta}_r$ (n, Sg, DGP3, NDIS), g=1...6, n=400,1200,2400. Population coefficient value is 0.084 with a t-ratio of 0.872.

Sampling Design	n	$\hat{Bias}$	$\hat{Var}$	$\hat{MSE}$	$\bar{t}_r$	Rejection Frequency
S1	400	-0.147	0.136	0.158	-0.123	0.029
	1200	-0.124	0.041	0.056	-0.165	0.027
	2400	-0.123	0.021	0.036	-0.249	0.023
S2	400	-0.159	0.214	0.239	-0.135	0.034
	1200	-0.141	0.063	0.083	-0.218	0.035
	2400	-0.140	0.029	0.049	-0.319	0.019
S3	400	-0.137	0.137	0.156	-0.113	0.030
	1200	-0.133	0.043	0.061	-0.184	0.026
	2400	-0.130	0.020	0.037	-0.281	0.019
S4	400	-0.114	0.173	0.186	-0.112	0.035
	1200	-0.130	0.052	0.069	-0.197	0.033
	2400	-0.120	0.026	0.040	-0.293	0.023
S5	400	–0.148	0.133	0.154	-0.127	0.036
	1200	-0.133	0.040	0.058	-0.207	0.023
	2400	-0.132	0.020	0.037	-0.296	0.020
S6	400	-0.126	0.179	0.195	-0.082	0.035
	1200	-0.124	0.052	0.067	-0.171	0.027
	2400	-0.121	0.027	0.042	-0.273	0.023

Table 15. Summary statistics for estimated sampling distributions of $\hat{\beta}_r$ (n, Sg, DGP3, DIS), g=1...6, n=400,1200,2400. Population coefficient value is 1.440 with a t-ratio of 13.691.

Sampling Design	n	$\hat{Bias}$	$\hat{Var}$	$\hat{MSE}$	$\bar{t}_r$	Rejection Frequency
S1	400	0.211	0.146	0.191	4.385	0.995
	1200	0.193	0.045	0.082	7.555	1.000
	2400	0.191	0.021	0.057	10.352	1.000
S2	400	0.086	0.273	0.280	3.156	0.940
	1200	0.024	0.071	0.072	5.533	1.000
	2400	0.019	0.035	0.035	7.722	1.000
S3	400	0.245	0.161	0.221	4.322	0.995
	1200	0.224	0.048	0.098	7.467	1.000
	2400	0.221	0.023	0.072	10.251	1.000
S4	400	0.311	0.216	0.313	3.879	0.997
	1200	0.234	0.068	0.123	6.656	1.000
	2400	0.213	0.031	0.076	9.165	1.000
S5	400	0.160	0.143	0.169	4.330	0.994
	1200	0.150	0.044	0.067	7.475	1.000
	2400	0.144	0.021	0.042	10.217	1.000
S6	400	0.344	0.239	0.357	3.866	0.994
	1200	0.276	0.069	0.145	6.666	1.000
	2400	0.262	0.033	0.102	9.217	1.000

Table 16. Asymptotic P-values for the OCC Banks A, B and C.

Bank and Minority Group		Using b	Using $\hat{\beta}$
Bank A	MA	0.102	0.125
	MB	0.165	0.025
Bank B	MA	0.203	0.512
	MB	0.002	<0.001
Bank C	M	0.257	0.009