How Low Can You Go?
An Optimal Sampling Strategy for Fair Lending Exams

Jason Dietrich

OCC Economics Working Paper 2001-3
August 2001

How Low Can You Go?
An Optimal Sampling Strategy for Fair Lending Exams

Jason Dietrich

Office of the Comptroller of the Currency
Economic and Policy Analysis Working Paper

August 2001

Abstract: This study uses Monte Carlo simulation to examine the impact of nine sampling strategies on the finite sample performance of the maximum likelihood logit estimator. Empirical researchers face a tradeoff between the lower resource costs associated with smaller samples and the increased confidence in the results gained from larger samples. Choice of sampling strategy is one tool researchers can use to reduce costs yet still attain desired confidence levels. The nine sampling strategies examined in this study include simple random sampling and eight variations of stratified random sampling. Bias, mean-square-error, percentage of models that are feasibly estimated, and percentage of simulated estimates that differ statistically from the true population parameters are used as measures of finite sample performance.

The results show stratified random sampling by action (loan approval/denial) and race of the applicant, with balanced strata sizes and a bias correction for choice-based sampling, outperforms each of the other sampling strategies with respect to the four performance measures. These findings, taken together with supporting evidence presented in Scheuren and Sangha (1998) and Giles and Courchane (2000) make a strong argument for implementing such a sampling strategy in future fair lending exams.

The views expressed in this paper are those of the author alone, and do not necessarily reflect those of the Office of the Comptroller of the Currency or the Department of the Treasury. The author would like to thank Jeff Brown, Irene Fang, David Guilkey, Amber Jessup, Shanzi Ke, Dom Mancini, Todd Vermilyea, and Morey Rothberg for their insightful comments and editorial assistance.

Please address correspondence to Jason Dietrich, Economist, Risk Analysis Division, Office of the Comptroller of the Currency, 250 E. Street, S.W., Washington, DC 20219 (phone: 202-874-5119; e-mail: jason.dietrich@occ.treas.gov).

I. Introduction

Sample size and sampling strategy both affect the small sample bias of the maximum likelihood (ML) logit estimator and the precision of the subsequent estimates. Much work has been done showing that larger sample sizes reduce bias and increase precision. However, relatively little is known about the rates of these improvements. Further, even less is known about how alternative sampling strategies affect bias and precision, or the rates of improvement in these measures gained from larger samples. Considering the resource constraints present in most empirical analyses, as well as the high costs of data collection and entry, researchers face important cost/bias and cost/precision tradeoffs when making sample size and sampling strategy choices. Understanding these tradeoffs, including the rates of change of these tradeoffs, is therefore important to achieving reliable results with minimum resource costs.

This study uses Monte Carlo simulation to examine the sample-size dependent effects of nine sampling strategies on the small sample performance of the ML logit estimator.[1] The nine sampling strategies include simple random sampling and eight variations of stratified random sampling based on proportional and balanced strata allocations. Bias, mean-square-error (MSE), percentage of models that are feasibly estimated, and percentage of simulated estimates that differ statistically from the true population parameter are all used to measure small sample performance. For each

[1] A Monte Carlo simulation is a set of repeated trials of a partially random process and is used to characterize the process. As a very simple example, suppose we have a random number generator with an unknown distribution. To determine the characteristics of the distribution, we can take repeated draws and plot the results.

sampling strategy, the Monte Carlo simulation provides a stream of sample-size dependent estimates for each performance measure. Comparing these simulated streams of performance measures across sampling strategies will indicate which sampling strategies are feasible in small samples, which strategy leads to the best small sample performance for the ML logit estimator, and what sample size is needed to achieve accurate and precise parameter estimates. All of these Monte Carlo simulations are conducted in the context of a statistically modeled fair lending exam that the Office of the Comptroller of the Currency (OCC) conducted in 1998.

The remainder of the paper is laid out as follows. Section II provides a brief review of the literature. Section III discusses the nine sampling strategies examined in the Monte Carlo simulations. Section IV develops the data-generating and Monte Carlo processes. Section V contains the results and section VI concludes the discussion.

II. Background

Statisticians and epidemiologists have been the leaders in studying sample-size calculations and their subsequent effects on hypothesis tests for multivariate estimators. One common approach develops a minimal sample-size formula from a general test statistic, which achieves an assumed set of desirable goals or criteria. This test statistic is typically a function of a combination of the following: sample size, significance level, power, population distributions, and population parameters. Using available information, estimates, and assumptions, values of all unknowns except sample size are specified and the formula is solved. As an example, suppose we have a normally distributed random variable $\hat{\beta}$, which is the coefficient estimate from a logistic regression. What sample

size is needed to detect an analyst-specified linear trend in the log odds ratio? In other words, what sample size is needed to reject the null hypothesis that $\beta = \beta_0$ in favor of the alternative hypothesis that $\beta > \beta_0$? Following a typical analytical approach we could draw a sample, estimate $\hat{\beta}$, and compute the standard t-statistic, rejecting the null hypothesis if the t-statistic is greater than some critical value, Z_α,

$$\frac{\hat{\beta} - \beta_0}{\sigma_0 / \sqrt{n}} > Z_\alpha \tag{1}$$

The term σ_0 is the variance of β if the null hypothesis is true and α denotes the size of the test, or the probability of rejecting a true null hypothesis. Specifying values for β_0 and α, and creating some measure for σ_0, we can solve for n to calculate the minimal sample size needed to detect an analyst-specified linear trend in the odds ratio at the 100x(1-α) percent confidence level. This exercise is typically taken a step further to incorporate the power of the test into the minimal sample-size formula to achieve a desired power and size. The usual sample-size formula for a normal test statistic that takes into account both the size and power of the test statistic is,

$$n = \left[\frac{Z_\alpha \sigma_0 + Z_\rho \sigma_1}{\beta_1} \right]^2 \tag{2}$$

where "1" subscripts denote values under the alternative hypothesis and Z_ρ is the critical value for a test with a power of ρ (Bull (1993)).

Following this general approach, a number of researchers have constructed minimal sample-size formulas for variations of the logit model. For example, Whittemore (1981) examined the logistic regression with small response probabilities;

Wilson and Gordon (1986) looked at the effects of confounding variables on minimal sample sizes in general linear models; Self and Mauritsen (1988) developed a minimal sample-size formula based on a score statistic used to conduct hypothesis tests for generalized linear models; and Bull examined multinomial logistic regression models with one categorical independent variable.[2]

The usefulness of this approach depends mainly on the type of information that is available, since the minimal sample size formulas are typically based on much unknown information. The three most important pieces are the sample size, the true population parameters, and the distribution of data. If more than one of these is unknown, the applicability of these formulas to empirical analysis is somewhat diminished. Further, other than that of Kao and McCabe, each of these studies looks only at the minimal sample size and not at the minimal sample allocation across strata. This is an important shortcoming since the small sample properties of the ML logit estimator differ with simple random and stratified random-sampling approaches.

Monte Carlo simulation is a second approach used to examine the effects of sampling on the small sample properties of multivariate estimators. Unlike the first approach, however, much less has been done in this area. Studies that have used Monte Carlo simulation to look at these sampling issues directly include Gordon *et al.* (1994), Breslow and Chatterjee (1999), Scheuren and Sangha, and Giles and Courchane. Gordon *et al.* used Monte Carlo simulation to examine the small sample properties of the probit and logit estimator in the context of female labor supply decisions with simple random sampling. The authors show that there is considerable variation in both the coefficient

[2] Also see Daganzo (1980), Donner (1984), Hsieh (1989), Phillips and Pocock (1989), Rochon (1989), Dupont and Plummer (1990), and Kao and McCabe (1991).

and standard error estimates up to nearly 2000 observations. In addition, for all 20 simulations they present, there was at least one rejection of the null hypothesis that the simulated estimate was equal to the true population parameter, for all sample sizes up to 12,000. Breslow and Chatterjee examine nonparametric maximum likelihood estimation of the logistic regression in the context of the Wilms National Tumor Study. They identify precision gains using this approach with a balanced-sampling scheme. Scheuren and Sangha, and Giles and Courchane both look specifically at fair lending models and take similar approaches. Scheuren and Sangha examine one data generating process (DGP) with two different sampling designs, while Giles and Courchane extend this to three DGPs and six sampling designs. Both studies find evidence supporting the use of balanced stratified sampling by action and race of the applicant, where action refers to a bank's underwriting decision on a credit application.

This study follows the second approach to studying sampling issues and extends the current research in three directions. First, it extends the number of sampling strategies to nine, looking at five of the six from Giles and Courchane plus simple random sampling, proportional sampling by race, proportional sampling by action, and the actual sampling used for the fair lending exam. The Kao and McCabe sampling approach with balanced race is the one approach from Giles and Courchane not included here.[3] Second, this study extends the performance results of the ML logit estimator to smaller samples. The smallest samples examined in the simulations by Gordon *et al.*, Breslow and Chatterjee, Scheuren and Sangha, Giles and Courchane are 734, 1142, 400,

[3] The Kao and McCabe formulation is developed on a premise of stratification of the outcome only. Since it is not clear how this approach, or formulation, would change if stratification by race was considered as well, I allow the racial strata sample sizes to independently vary and do not examine Kao and McCabe sampling with balanced race.

and 400, respectively, while the smallest sample examined in this study is 50. This is important since many empirical analyses use samples with fewer than 400 observations. As one example, of 16 statistically modeled fair lending exams the OCC has conducted since 1995, only six had samples larger than 400 applications, and five of these six examined more than two racial groups. With this precedence for sample size, along with knowledge of the resource constraints the OCC faces, performance results for smaller samples would be more useful to their statistically modeled fair lending program. Third, this study examines changes in performance measures for shorter sample intervals. For example, Giles and Courchane examine samples of size 400, 1200, and 2400. In contrast, this study examines all sample sizes from 50 to 1600 by 12. This will provide a better indication of the rates of improvement in bias and precision with larger samples, and will identify the minimum sample size needed to achieve precise results with a desired level of confidence.

III. Sampling Strategies

This section describes the nine sampling strategies examined in the Monte Carlo simulation. The first sampling strategy, simple random sampling, is the easiest to understand and apply. On average, simple random sampling yields a sample reflecting the true population distributions, with the likelihood of this occurring increasing with the sample size. For smaller samples, however, there is an increased risk that the model cannot be estimated because of limited variation in either the dependent or independent variables. In addition, the ML logit estimates will be less precise with simple random

sampling than for any of the second general group of sampling strategies, which employ stratified random sampling.

Stratified random sampling is most beneficial when the data are homogenous within strata and heterogeneous across strata. By incorporating information about the distributions of the data, one can potentially increase the precision of the estimates with fewer observations. This study analyzes eight variations of stratified random sampling. I consider three general levels of stratification: action, race, and action and race. For each of these three stratifications, I look at proportional and balanced sampling schemes. For proportional sampling, the sample strata sizes have the same proportions as the population strata. For balanced sampling, each sample strata contains an equal number of observations. Thus, two variations for each of the three stratifications yields six stratified random sampling strategies. The seventh stratified random sampling strategy follows Kao and McCabe who estimate strata sizes for stratified random sampling by action by minimizing the expected misclassification rates, or expected error regret (EER). This method provides a measure of strata allocation that lies between the extremes of proportional and balanced stratified random sampling. Unfortunately, there are currently no analogous techniques for stratification by race, or action and race. For the model used in this study, the EER was minimized for a sample with 62.5 percent approvals and 37.5 percent denials. This compares with 87.8 percent approvals and 12.2 percent denials in the population. The eighth and final stratified random sampling strategy is the one used during the actual exam. OCC staff examined two racial groups for this exam and used stratified sampling by action and race to create a sample with 54.9 percent race 1 approvals, 15.1 percent race 1 denials, 16.9 percent race 2 approvals, and 13.0 percent

race 2 denials.[4] Including this approach provides a nice comparison of how alternative sampling strategies might have affected the results for this exam.

IV. Monte Carlo Simulation

This section outlines the Monte Carlo simulation procedure, as well as the DGP used to create the simulated population data. The main drawback of using Monte Carlo simulation is that the results necessarily depend on the underlying DGP assumptions. Because of this, I take a case study approach and develop a DGP that creates a simulated population closely reflecting the characteristics of one specific fair lending exam previously conducted by the OCC. At a minimum, then, conclusions can be drawn with some certainty about how different sampling strategies would have affected this exam's model results. Caution must be exercised, however, in generalizing these results to other analyses.

The OCC estimated the following model for this exam,

$$Denied_i^* = \beta_0 + \beta_1 LTV_i + \beta_2 DTI_i + \beta_3 Score_i + \beta_4 Derogs_i + \beta_5 Public_i + \beta_6 Race_i + \varepsilon_i \quad (3)$$

where $Denied^*$ = Unobserved continuous latent variable measuring the probability of being denied credit,
LTV = 0/1 Indicator conveying the loan-to-value > the policy cutoff,
DTI = Continuous debt-to-income ratio,
Score = 0/1 Indicator conveying the credit score > the policy cutoff,
Derogs = 0/1 Indicator conveying applicant does not have clean credit,
Public = 0/1 Indicator conveying at least one public record, and
Race = 0/1 Indicator conveying minority status.

The variable "denied" is the observed 0/1 counterpart to $denied^*$. With this model as the basis for the DGP, I attempted to re-create the following population characteristics:

[4] Dietrich (2001) summarizes the sampling approach the OCC uses for its statistically modeled fair lending exams, and presents specific population and sample sizes for this exam.

- Means and variances of explanatory variables
- Differential creditworthiness by race
- Correlations of explanatory variables
- Size
- Racial distribution of originations
- Parameters explaining the probability of being denied

For this exam, the OCC used choice-based sampling, stratifying by action and race. Using these sample strata sizes, along with the actual population strata sizes, one can construct consistent estimates of the population means and variances of the explanatory variables by race. Using these estimated means as cutoffs, I take random draws with replacement from a uniform distribution (U) to construct the five discrete explanatory variables. Different cutoffs are used to capture racial differences for LTV, Score, Derogs and Public found in the sample data. For the continuous variable DTI, I use random draws with replacement from a normal distribution (N) using the estimated population means and variances. Differences in means and variances by race are incorporated into the construction of DTI as well.[5] Data was generated for 2635 applicants with race 0 and 325 with race 1 to match the actual population data for the exam. Table 1 presents the moment estimates of the actual population that were used in the DGP process and the summary statistics of the resultant simulated population. The simulated population characteristics are relatively close to the true population in all instances.

The population correlations among explanatory variables are difficult to recreate since they involve a joint distribution of six variables. Therefore, the correlations were allowed to vary independently of the DGP. However, random draws creating the explanatory variable data were repeated until the correlation matrix of the simulated

[5] Except for a small percentage of applications with unusually high DTI values, the exam sample of DTI is approximately normally distributed.

Table 1: DGP and Simulated Population Data (N=2960)					
	LTV	DTI	Score	Derogs	Public
Race = 0					
DGP	U(0.335)	N(34.093,62.218*)	U(0.684)	U(0.147)	U(0.283)
Mean	0.341	33.810	0.676	0.143	0.277
Standard Deviation	0.474	8.369	0.468	0.350	0.448
Race = 1					
DGP	U(0.572)	N(35.486,73.047*)	U(0.265)	U(0.206)	U(0.580)
Mean	0.575	35.759	0.231	0.231	0.631
Standard Deviation	0.495	7.942	0.422	0.422	0.483
Total					
Mean	0.367	34.024	0.627	0.153	0.316
Standard Deviation	0.482	8.345	0.484	0.360	0.465
* This value denotes the variance of the distribution.					

population was qualitatively equal to the correlation matrix of the exam sample. Table 2 presents both of these correlation matrices. Although the correlations are categorically smaller for the simulated population, the direction of correlation is the same for all variable pairs.

The last stage of the DGP is to specify values for the β's in equation (3) and generate the dependent variable. The objective here is to construct a simulated population data set that has the same distribution of approvals and denials found in the true population and also yields the true population parameter values when estimating equation (3). The first step to achieving this objective is to determine unbiased and consistent estimates of the true population parameters. A natural place to start is with the actual exam estimates, but this is complicated by bias introduced from choice-based sampling and a small sample size. Using choice-based sampling by action and race, as was done for this exam, introduces bias into the constant and race coefficients. Scott and

Table 2: Correlation Matrices						
Correlation Matrix for the Actual Exam Data (N=284)						
	LTV	DTI	Score	Derogs	Public	Race
LTV	1.000					
DTI	0.027	1.000				
Score	-0.080	-0.212	1.000			
Derogs	-0.056	0.161	-0.350	1.000		
Public	0.100	0.145	-0.417	0.358	1.000	
Race	0.181	0.088	-0.395	0.121	0.297	1.000
Correlation Matrix for the Simulated Population Data (N=2960)						
LTV	1.000					
DTI	0.000	1.000				
Score	-0.037	-0.045	1.000			
Derogs	-0.023	0.027	-0.022	1.000		
Public	0.015	0.039	-0.077	0.021	1.000	
Race	0.152	0.073	-0.288	0.076	0.238	1.000

Wild (1986, 1991, 1997) provide a correction for this bias, which can be used to obtain consistent estimates of each of these parameters. Small sample bias is more problematic. Although approximations for small sample bias with a ML logit estimator are available, they depend on the true population parameter and are therefore of little use here.[6] Therefore, I simply take the Scott and Wild bias-corrected coefficients as the true population parameter values.

The next step is to use these parameter values along with equation (3) and random draws from a logistic distribution to create the simulated dependent variable. Unfortunately, these parameter values will not result in the true population distribution of approvals and denials since the simulated correlation matrix of explanatory variables differs from the exam sample correlation matrix. Due to these smaller correlations, I adjust the parameters for LTV and race to achieve the desired distribution of approvals and denials while still retaining the qualitative characteristics of the actual exam

[6] See Amemiya (1980) and MacKinnon and Smith (1998)

estimation results. To match the true population size and racial distribution of originations exactly, I first construct a simulated population of 50,000 observations with approximately the distribution of approvals and denials desired and then use random sampling by action and race to pare this simulated dataset down to the final simulated population data set containing exactly the desired characteristics. Table 3 presents the actual exam estimates, the bias-corrected and correlation-adjusted β values used in the DGP, and the parameter estimates and racial distribution of originations using the final simulated population data. The simulated β values are taken as the true population parameters for the Monte Carlo simulation.

Table 3: Data Generating Process (DGP) and Simulated Population Data (N=2960)

$\varepsilon \sim$ Logistic(0, 3.295)

	Approved	Denied
Race = 0	2353	282
Race = 1	246	79

	Constant	LTV	DTI	Score	Derogs	Public	Race
Actual exam β's	-6.542	-0.398	0.129	-1.301	0.516	0.666	0.361
DGP β's	-6.350	-0.198	0.129	-1.301	0.516	0.666	0.184
Simulated β's	-6.597	-0.153	0.136	-1.411	0.657	0.519	0.195

The Monte Carlo simulation process using this simulated population data consists of five steps.

Step 1: For each of the nine sampling strategies, draw an initial total sample of size 50 using the particular sampling approach. Then estimate the model for each sample and save the results.

Step 2: Augment each of the nine initial samples with 12 additional observations, using the particular sampling strategy, sampling without replacement from the remaining population data. For simple random sampling, this simply entails adding 12 observations randomly chosen from the remaining population. For proportional sampling, the number of observations added to each stratum is proportional to the population stratum sizes. Non-integer increments are rounded to assure the sample is increased by 12. For balanced sampling, three observations are added to each sample strata. Once a particular population strata limit is reached, observations are added to the remaining strata in a balanced manner. For example, if strata 1 has reached its population limit, strata 2–4 are increased by four instead of three. Similarly, if two strata have reached their population limit, the remaining two strata are increased by six. Finally, the sample strata for the Kao and McCabe and actual exam sampling strategies are increased by the originally determined proportions. Similar to balanced sampling, when population strata limits are reached, the 12 additional observations are allocated among the remaining sample strata using the originally determined proportions. The model is then estimated using these nine newly augmented samples, and the results are saved.

Step 3: Repeat step 2 until the sample size reaches 1600. The coefficient and standard error estimates for all of the sampling strategies will converge to the population values as the sample size approaches the population size. However, due to resource constraints, it is unlikely that any fair lending exam will ever examine

more than 1600 observations. Once the sample size reaches 1600, there will be nine streams of estimates, one for each sampling strategy going from sample sizes of 50 to 1600 by 12.

Step 4: Repeat steps 1–3 10,000 times to create 10,000 streams of estimates for each of the nine sampling strategies.

Step 5: For each of the nine sampling strategies, calculate sample-size dependent streams for each of the four performance measures. This entails creating means of coefficient estimates, means of MSE estimates, percentages of models that could not be estimated, and percentages of rejections of the hypothesis that the simulated estimate equals the true population parameter.

V. Monte Carlo Simulation Results

Graphs 1–4 present all of the Monte Carlo simulation results. The first question of interest for this study is what sampling strategies are feasible for an ML logit estimator at small sample sizes. As one measure of feasibility, graph 1 presents the percentage of models that could not be estimated at each sample size for each sampling strategy. A model was deemed unable to be estimated if one or more coefficients could not be estimated.[7] Since the likelihood function for the logit estimator is concave, this is not a problem of convergence, but simply one of lack of variation. For example, with a particular sample, an indicator explanatory variable may have fewer values of 1 than there are parameters in the model. If this problem is isolated to variables available prior

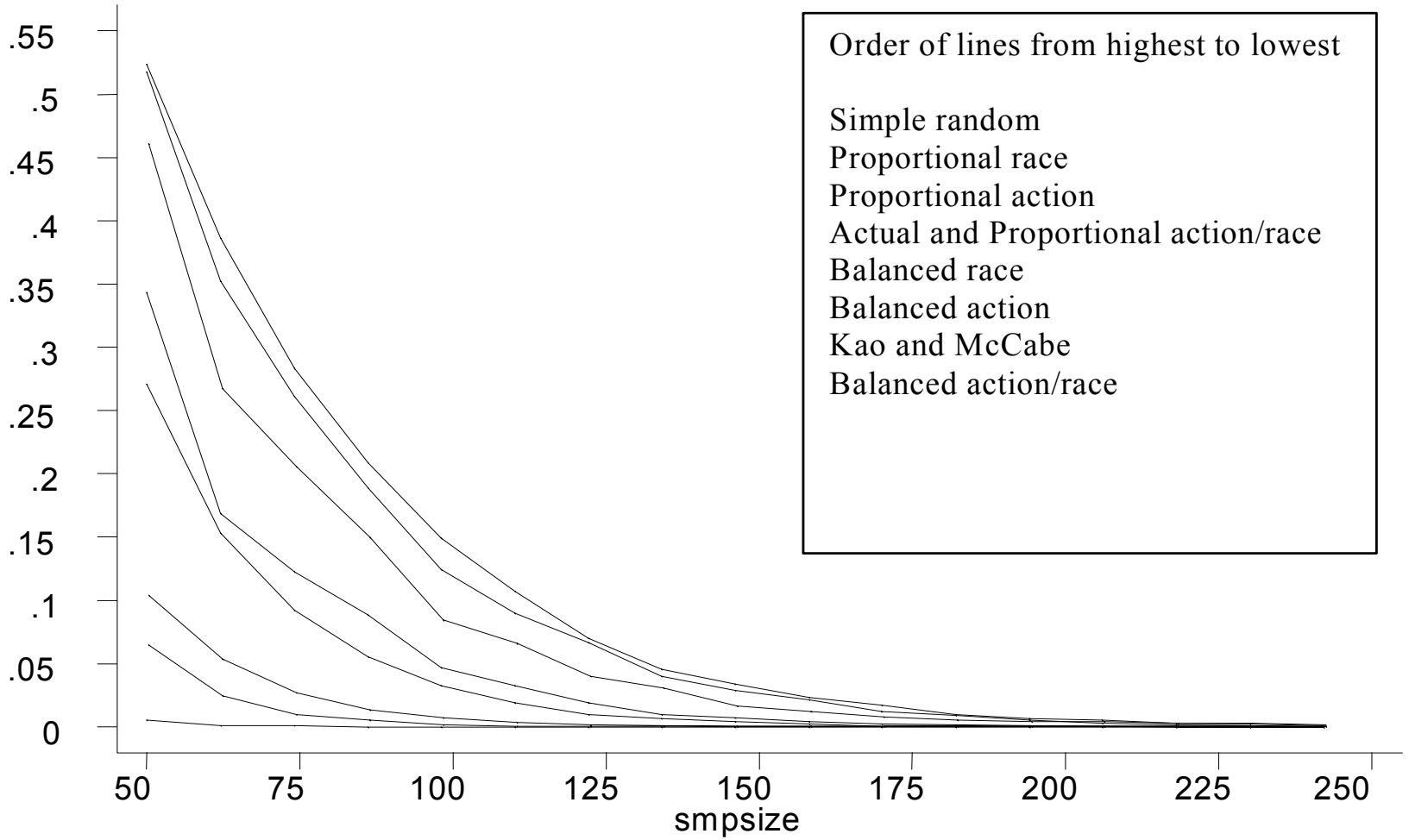
[7] Models with large and unstable estimates were considered able to be estimated.

to sampling, one could simply redraw the sample if the number of observations included in the initial sample for any of these variables is less than the expected number of parameters in the model. Problems could still occur, however, since the variation of data gathered after sampling may be small as well. In addition, redrawing samples until achieving one with sufficient variation is essentially making the case for using an alternative sampling strategy.

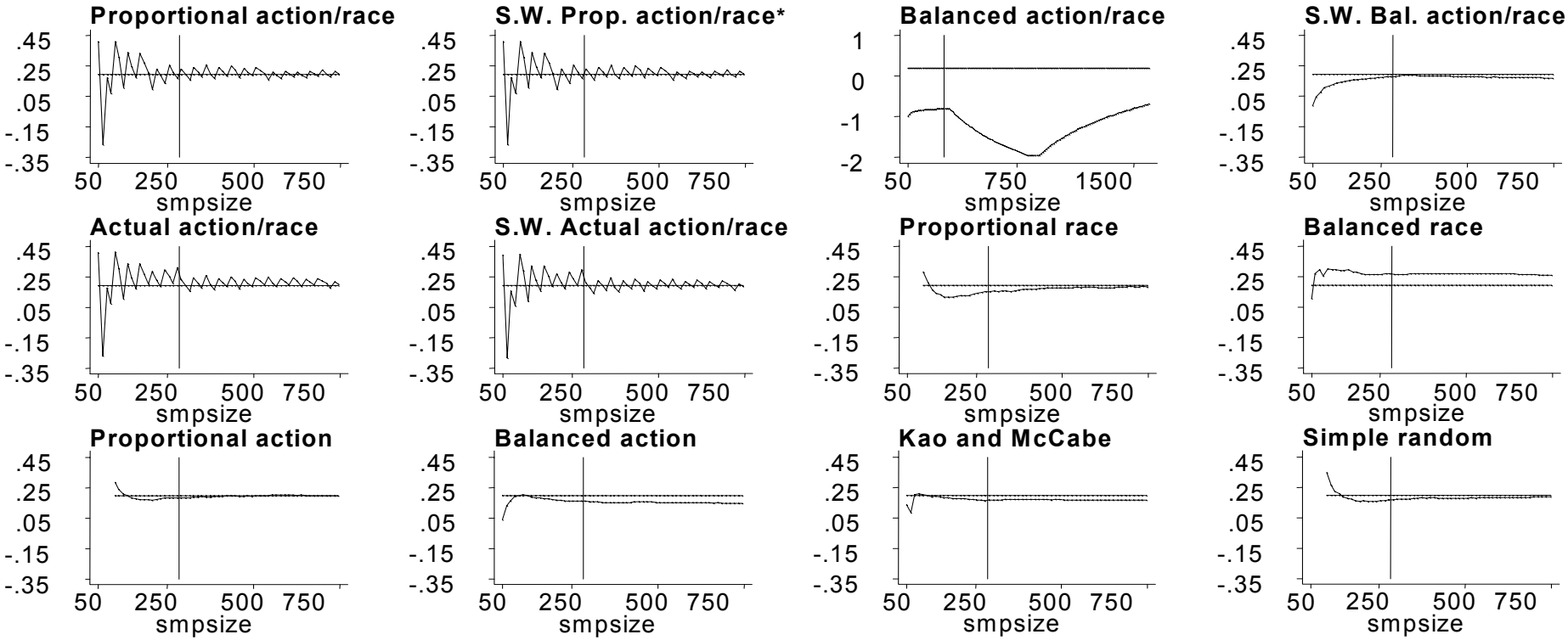
As graph 1 shows, simple random sampling and the proportional stratified random sampling strategies have the highest failure rates for small samples, while the balanced strategies have the lowest failure rates. This is what one would expect given the small number of population denials for race=1 (79) and the fact that balanced sampling includes larger numbers of these applications in the sample with more certainty than either simple or proportional stratified random sampling. This highlights one reason for using stratified random sampling over simple random sampling. A second item of note in graph 1 is that the inability to estimate models is basically eliminated for all sampling strategies by samples of size 200.

The second and third questions of interest for this study are which sampling strategy yields the best small sample performance for the ML logit estimator and how large of a sample is needed for sufficient confidence in the results. Graphs 2, 3, and 4 provide evidence to answer each of these questions. Graph 2 presents the simulated streams of sample-size dependent estimates of race. The horizontal line indicates the true population value and the vertical line shows the sample size drawn for the actual exam. In addition to the original nine sampling strategies, I also present Scott and Wild, bias-

Graph 1: Percent of Models Unable to be Estimated



Graph 2: Simulated Estimates of Race



* Scott and Wild (1986, 1991, 1997) data abbreviated as "S.W."

corrected results for the samples stratified by action and race using proportional, balanced, and actual strata proportions. Further, except for balanced sampling by action and race, which shows considerably higher bias than the other sampling strategies, all of the graphs have the same scale for ease of comparison, and only go to a sample size of 750.[8] A sample size of 750 is slightly larger than that drawn for any past fair lending exam conducted by the OCC, and resource constraints would make larger samples infeasible.

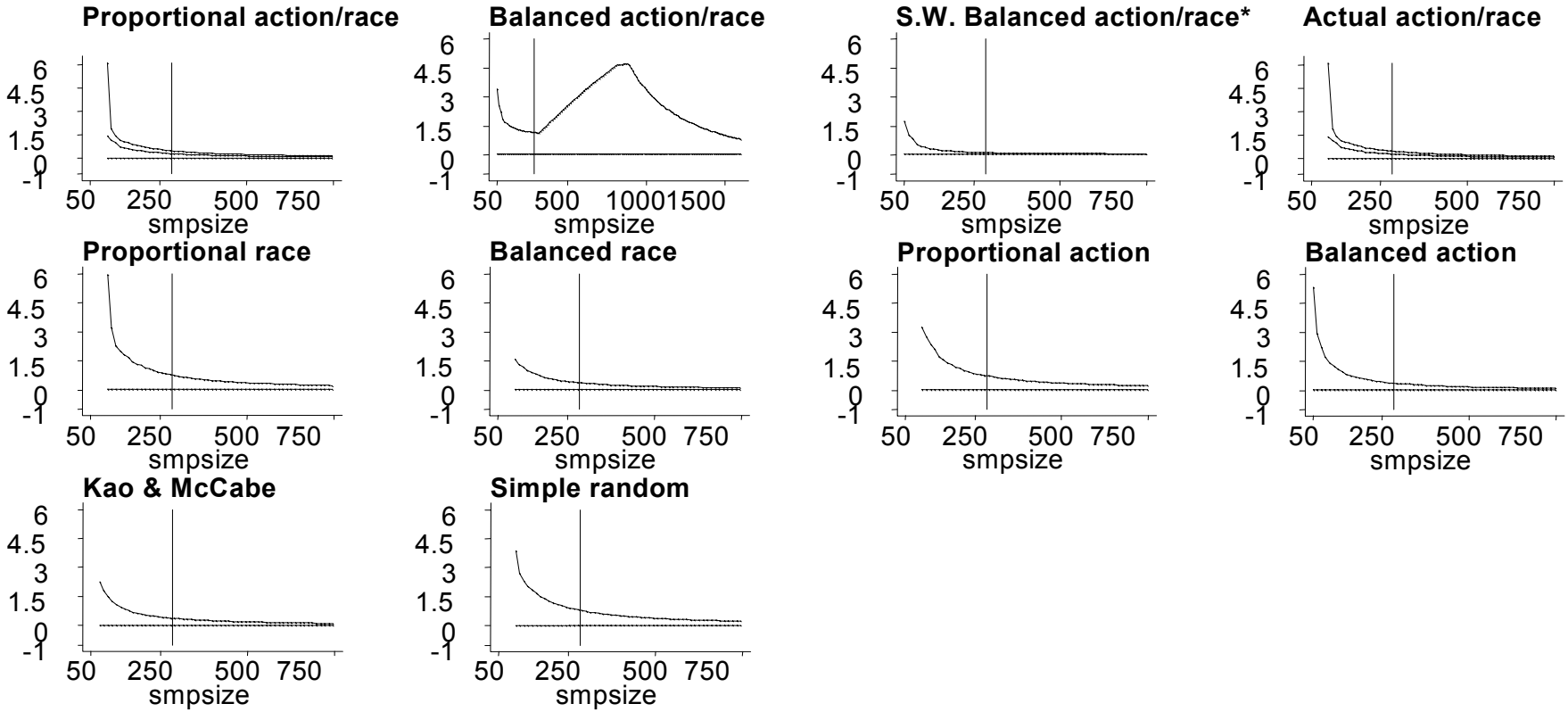
There are four items of note in graph 2. First, proportional sampling by action, and possibly bias-corrected balanced sampling by action and race show the least small-sample bias and quickest convergence to the true population parameter. Balanced sampling by action, proportional sampling by race, Kao and McCabe sampling, and simple random sampling all perform fairly strongly as well. Second, the results for balanced sampling by action and race differ considerably from the results for the other sampling strategies. The shape of the balanced action/race graph is a result of the sample augmentation process used in the Monte Carlo simulation. Each of the larger directional changes occurs when a population strata level is reached (79, 246, and 282) and the subsequent sets of 12 observations are allocated to the remaining strata. This result is relevant for current fair lending exams, since in some cases one or two strata are increased, independent of the overall sampling strategy, to reach 50 observations. This graph indicates that these increases may have a larger impact than one might expect. Third, the Scott and Wild bias correction has a considerable effect for balanced sampling. This is what one would expect, considering that balanced sampling by action and race is

[8] The graphs for proportional random sampling by race, proportional random sampling by action, and simple random sampling all start at 98 since estimates for smaller samples were much larger than any of the other sampling strategies making comparable axes impossible.

the most biased of all of the sampling strategies. Choice-based sampling with proportional strata allocation will not introduce any additional bias to the small sample bias of the ML logit estimator so nothing is gained by applying the Scott and Wild bias correction there. Similarly, the bias correction also shows little effect on the actual sample drawn since the strata allocations are nearly proportional. However, for all of these sampling strategies, there will be precision gains from applying the bias correction, which we will see later. Finally, except for balanced sampling by action and race and balanced sampling by race, each of the sampling strategies provides a good estimate of the true population parameter by 284 applications, the sample size used for the actual exam. It also appears that fewer applications could have been examined using alternative sampling strategies.

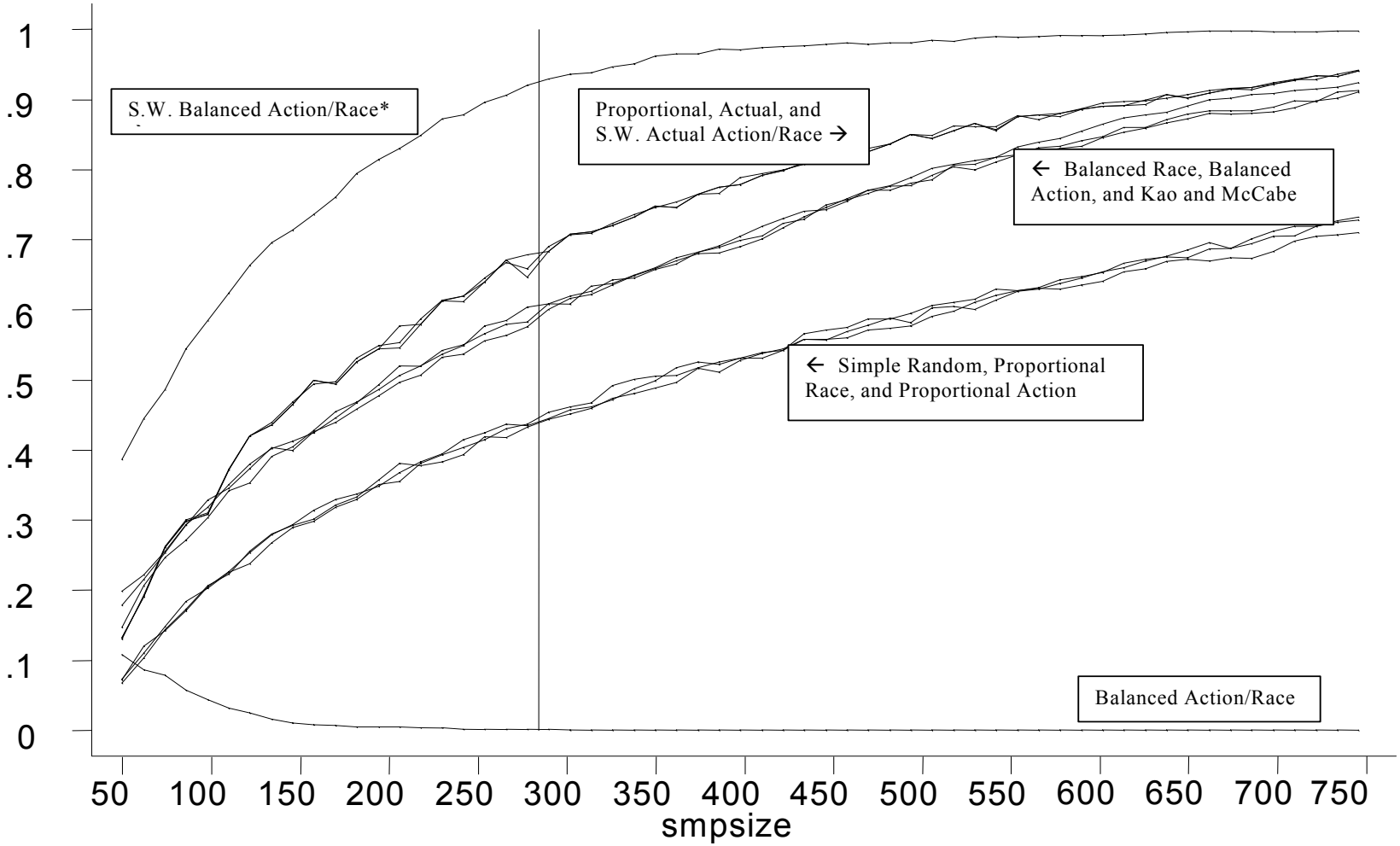
Graph 3 presents MSE results, which take into account the estimated standard errors. Similar to graph 2, a horizontal line indicates the true variance of the race parameter and a vertical line indicates the actual sample size used for the exam. In addition, the Scott and Wild results for proportional and actual sampling by action and race are combined with their uncorrected counterparts, and all scales are now equivalent. The bias-corrected balanced sampling by action and race again appears to perform well relative to the remaining sampling strategies, with balanced sampling by race, Kao and McCabe, and the other two bias-corrected approaches close behind. Looking at the bias-corrected results, we now see the efficiency gains mentioned earlier. For proportional sampling by race and action, MSE equals 1.88 at a sample size of 110. With the bias correction, MSE equals 1.18, a 37.2 percent decrease. Looking finally at sample size, the

Graph 3: Simulated MSE for Race



* Scott and Wild (1986, 1991, 1997) data abbreviated as "S.W."

Graph 4: Non-Rejection Percentage of
H0: Estimated Race Coefficient Equals the True Population Value



* Scott and Wild (1986, 1991, 1997) data abbreviated as "S.W."

actual sample again appears to have been suitable, but smaller samples might have been sufficient.

Graph 4 shows, for each sampling strategy, the percentage of times I could not reject the null hypothesis that the estimated race coefficient equals the true population value at the 95 percent confidence level. The point of statistics is to estimate some unknown population value. This graph shows how well each of the sampling strategies performs on this point. Bias-corrected balanced sampling by action and race is clearly superior on this count, reaching a success rate of just over 90 percent by 284 applications. Proportional sampling by action and race, and both actual and bias-corrected actual sampling by action and race all perform second best. Each of this second group performed poorly in graph 2 because of the high variance of estimates, not because of a systematic bias. Their relatively high success rates at predicting the true population value suggest the lower precision is not necessarily a problem. Other than balanced sampling by action and race, which is highly biased, simple random sampling, proportional random sampling by race and proportional random sampling by action all performed worst on this performance measure. As for sample sizes, other than the bias-corrected balanced sampling by action and race, the results do not build high levels of confidence in the logit results until 600 or 700 observations.

VI. Conclusion

Given the resource constraints faced by most empirical analyses, it is important to identify sampling and estimation techniques, which produce unbiased and precise parameter estimates at minimal resource cost. This study examines how nine sampling

strategies affect the small sample performance of the ML logit estimator in the context of a fair lending exam the OCC conducted in 1998. The results indicate that balanced sampling by action and race with a bias correction for choice-based sampling outperforms all of the other sampling strategies. Bias and MSE are both low and converge quickly to zero and the population variance, respectively. Further, the percentage of models able to be estimated at small samples is high and the percentage of rejections of the null hypothesis that the estimated coefficient for race equals the true population value is low. A second important result of this study is that the sample size used for the actual exam appears to have been adequate to accurately estimate the true population value. However, a smaller sample size could have been used to achieve the same outcome with the bias-corrected balanced sampling approach.

All of the results presented in this study are based on Monte Carlo simulation and are therefore dependent on the assumptions of the DGP. However, these results, taken together with similar findings in Scheuren and Sangha, and Giles and Courchane, provide ever-increasing evidence for using balanced sampling by action and race with a correction for choice-based sampling for future fair lending exams.

References

Amemiya, Takeshi. 1980. The n^{-2}-order Mean Squared Errors of the Maximum Likelihood and the Minimum Logit Chi-Square Estimator. *The Annals of Statistics* 8(3):488-505.

Avery, Robert A., Patricia E. Beeson, and Paul S. Calem. 1997. Using HMDA Data as a Regulatory Screen for Fair Lending Compliance. *Journal of Financial services Research* 11:9-42.

Black, Harold A., Thomas P. Boehm, and Ramon P. DeGennaro. 1999. Overages in Mortgage Pricing. Federal Reserve Bank of Chicago Proceedings from the Conference on Bank Structure and Competition. 255-80.

Breslow, N.E., and N. Chatterjee. 1999. Design and Analysis of Two Phase Studies with Binary Outcome Applied to Wilms Tumour Prognosis. *Applied Statistics* 48:457-68.

Bull, S.B. 1993. Sample Size and Power Determination for a Binary Outcome and Ordinal Exposure when Logistic Regression Analysis is Planned. *American Journal of Epidemiology* 137(6):676-84.

Cochran, W.G. 1977. *Sampling Techniques*. 3rd Ed. New York: Wiley.

Courchane, Marsha J., and David Nickerson. 1997. Discrimination Resulting From Overage Practices. *Journal of Financial Services Research* 11:133-52.

Daganzo, Carlos F. 1980. Optimal Sampling Strategies for Statistical Models with Discrete Dependent Variables. *Transportation Science* 14(4):324-45.

Dietrich, Jason. 2001. The Effects of Choice-based Sampling and Small Sample Bias on Past Fair-lending Exams. Mimeo. Office of the Comptroller of the Currency.

Donner, A. 1984. Approaches to Sample Size Estimation in the Design of Clinical Trials -- A Review. *Stat Med* 3:199-214.

Dupont, W.D., and W.D. Plummer, Jr. 1990. Power and Sample Size Calculations: A Review and Computer Program. *Controlled Clinical Trials* 11:116-28.

Giles, Judith A., and Marsha J. Courchane. 2000. Stratified Sample Design for Fair Lending Binary Logit Models. Econometrics Working Paper EWP000, Freddie Mac.

Gordon, Daniel V., Zhengxi Lin, Lars Osberg, and Shelley Phipps. 1994. Predicting Probabilities: Inherent and Sampling Variability in the Estimation of Discrete-Choice Models. *Oxford Bulletin of Economics and Statistics* 56(1):13-31.

Hsieh, F.Y. 1989. Sample Size Tables for Logistic Regression. *Stat Med* 8:795-802.

Kao, T.C., and G.P. McCabe. 1991. Optimal Sample Allocation for Normal Discrimination and Logistic Regression Under Stratified Sampling. *Journal of the American Statistical Association* 86:432-36.

MacKinnon, James G., and Anthony A. Smith, Jr. 1998. Approximate Bias Correction in Econometrics. *Journal of Econometrics* 85:205-30.

Phillips, A.N., and S.J. Pocock. 1989. Sample Size Requirements for Prospective Studies with Examples for Coronary Heart Disease. *Journal of Clinical Epidemiology* 42:639-48.

Prentice, R.L., and R. Pyke. 1979. Logistic Disease Incidence Models and Case-Control Studies. *Biometrika* 66:403-11.

Rochon, J. 1989. The Application of the GSK Method to the Determination of Minimum Sample Sizes. *Biometrics* 45:193-205.

Scheuren, F.J., and B.S. Sangha. 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 C. J. Wild. 1986. Fitting Logistic Models under Case-control or Choice Based Sampling. *Journal of the Royal Statistical Society* Series B 48(2):170-82.

Scott, A.J., and C. J. Wild. 1991. Fitting Logistic Models in Stratified Case-Control Studies. *Biometrics* 47:497-510.

Scott, A.J., and C. J. Wild. 1997. Fitting Regression Models to Case-Control Data by Maximum Likelihood. *Biometrika* 84:57-71.

Self, Steven G., and Robert H. Mauritsen. 1988. Power/Sample Size Calculations for Generalized Linear Models. *Biometrics* 44:79-86.

Stengel, Mitch, and Dennis Glennon. 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.

Wilson, S.R., and I. Gordon. 1986. Calculating Sample Sizes in the Presence of Confounding Variables. *Applied Statistics* 35:207-13.