CORRECTING SELF-SELECTION BIAS IN THE ESTIMATION OF AUDIT PROGRAM IMPACTS

by
Kenneth E. Train
Cambridge Systematics, Inc.
University of California, Berkeley

ABSTRACT

The fact that customers choose to participate in conservation actions leads to the possibility of self-selection bias in the evaluation of the impacts of conservation programs. This bias has been studied extensively in one context, namely, the analysis of energy consumption data. However, the bias also arises, and is perhaps more important, in models of customers' decisions of whether to take conservation actions and the impact of the audits on this decision. This bias has not previously been addressed. This paper provides intuitive examples of how the bias arises, and proposes and applies methods for correcting it.

The impact of audits on customers' decisions of whether to take conservation actions is typically estimated in a discrete choice model such as logit or probit. The dependent variable gives the probability that the customer takes an action. One of the explanatory variables is a dummy indicating whether or not the customer received an audit. The estimated coefficient of this dummy indicates the extent to which the audit influences the customers' decision to take action.

In reality, the audit dummy is self-selected since the customer chooses (or has some influence on) whether or not to be audited. Estimation without accounting for this fact leads to bias, the magnitude of which can be quite large. For example, consider an audit program that has no impact: that is, the audits do not induce customers to take actions that they would not otherwise take. Suppose further that any customer who plans to take a conservation action (for reasons other than the audit) requests an audit so as to obtain the audit information. Since each customer who is audited also takes an action, a model estimated on these data will give a large coefficient for the audit dummy, indicating, erroneously, that the audit program has a large impact on customers' decisions to take actions. Self-selection bias can also take the opposite form, indicating that a program has little effect when in actuality its impact is substantial.

This paper describes estimation procedures to correct for this bias. The methods are straightforward and can be implemented with standard packages such as SAS. The methods are applied to data from a commercial audit program, and the magnitude of the bias that occurs without the correction is measured.

SELF-SELECTION BIAS IN A NEW CONTEXT: ESTIMATING THE IMPACT OF CONSERVATION PROGRAMS ON MEASURE ADOPTION

by
Kenneth E. Train
Cambridge Systematics, Inc.
University of California, Berkeley

1. INTRODUCTION

Self-selection bias--how it arises and how to correct for it--has long been recognized as a vital issue in the evaluation of conservation program impacts (e.g., Williams and Walther, 1982). The concept is straightforward. Customers choose voluntarily to participate in conservation programs (that is, customers self-select participation). The customers that participate are generally different from those that do not participate: they are perhaps more energy conscious, or perhaps have a greater need to conserve since their energy bills are higher, or any number of things. Because of these differences, a comparison of participants in a program with a sample of nonparticipants does not provide an accurate estimate of program impacts. Any observed differences in the energy consumption and/or rate of measure adoption between participants and nonparticipants will be partly due to the program but will also be partly due to the fact that participants are different from nonparticipants independent of the program.

Self-selection bias has been examined extensively in one context, namely the analysis of energy consumption data. The analyst in this context compares consumption data for participants and nonparticipants. The standard procedure is to estimate a regression equation with energy consumption (or the change in energy consumption from before to after the program) as the dependent variable, using data from a sample of participants and nonparticipants. Explanatory variables include a dummy variable that indicates whether the customer is a participant plus other observed variables relating to the customer. The estimated coefficient of the dummy variable is intended to indicate the impact of the program, controlling for differences in observed characteristics. Self-selection bias arises because the participation dummy is endogenous, that is, is determined by the customers itself, and factors that affect a customer's choice to participate can be expected to be related to its energy consumption given that it participates.

Methods for correcting for self-selection bias in this context have been developed by Heckman (1978,1979) and Dubin and McFadden (1984). The methods have been applied by Trimble and Hirst (1983), Hirst et al (1983), Train and Ignelzi (1987), and Train (1988). These studies have shown that self-selection bias can (but need not)[1] be significant in this context and that correction for it is fairly simple.

Self-selection bias also arises in another context. It has not been studied previously in this context, and yet its importance is possibly greater than that in the traditional context. In particular, self-selection bias arises in the analysis of customers' decisions to adopt conservation measures and the impact of conservation programs on these decisions. Here the analyst estimates a discrete choice model (such as logit or probit) of the customer's choice of whether to adopt a particular conservation measure. The explanatory variables in the model include a dummy variable that indicates whether the customer participated in the conservation program. The estimated coefficient of this variable is intended to indicate the effect of the program on the customer's decision to adopt conservation measures. In particular, the estimated coefficient provides information on the amount by which the customer's probability of adopting a measure increases as a result of the program.

The difference between this situation and the traditionally studied one is that in this context the analyst is using a discrete choice model to examine customers' adoption decisions rather than a regression model of customers' consumption. However, the basis for self-selection bias is conceptually the same: the explanatory variable that indicates program participation is endogenous, chosen by the customer itself.

Self-selection bias in this less traditional context is the topic of this paper. We provide an intuitive explanation of how self-selection bias arises in this context and what forms it can take. We then describe a method for correcting for the bias. With data from an audit program, we use the method to estimate the impact of the program on customers' decisions to take a particular conservation measure. We compare the estimates obtained by this method with those obtained under standard procedures, and discuss the extent and form of the self-selection bias that occurs in this particular application.

2. DESCRIPTION OF THE PROBLEM

The purpose of a conservation program is (generally, at least) to induce customers to adopt conservation measures. To evaluate the program, the analyst therefore attempts to estimate the impact of the program on customers' decisions to adopt particular measures. This is generally accomplished by taking a sample of customers that participated in the program plus a sample of customers that did not participate, and comparing the rate

[1] Keating (1988) argues that self-selection in this context has been overemphasized.

of measure adoption between the participants and nonparticipants. Multivariate methods are used to control for other factors, such as differences in customer characteristics and in the cost and savings of the measure as faced by different customers. Discrete choice models in general, and specifically logit models, are used for this function. These models give the probability that the customer adopts a conservation measure as a function of various explanatory variables. Most prominent of the explanatory variables is a dummy variable that identifies whether the customer participated in the program. Other explanatory terms include the characteristics of the customer and the cost and savings of the measure.

For logit, the function takes the form:

$$(1) \qquad P_n = \frac{e^{cd_n + bx_n}}{1 + e^{cd_n + bx_n}}$$

where P_n is the probability that the customer n adopts the measure, d_n is a dummy variable that identifies whether customer n participated in the program, x_n is a vector of other explanatory variables, and c and b are parameters to be estimated. The parameter c reflects the impact of the program. If the program increases the customers' probability of adopting the measure, then c will be significantly positive.

Standard estimation procedures provide consistent estimates under the assumption that all of the explanatory variables are exogenous, that is, are determined independently of the conservation decisions of the customer (McFadden, 1973; Train, 1986). In the case of the participation dummy, this assumption is clearly violated. As stated in Part 1, the customer chooses to participate, and factors that affect the decision to participate can be expected to also affect the decision to adopt conservation measures. That is, the dummy d_n is self-selected, such that estimation with standard procedures leads to self-selection bias.

The bias can take several forms and can go in either direction, depending on how the decision to participate is related to the decision to adopt measures. For illustration, we describe two extreme situations that induce bias in opposite directions. First, we describe a situation in which a program actually has no effect on customers' decisions, but the estimated model, because of self-selection bias, will erroneously indicate that the program has a very large impact. We then describe a situation in which a program has a substantial impact on customers' decisions, but the estimated model indicates that it has a negative effect.

Consider first an audit program offered to customers on request. Suppose (for the sake of illustration) that customers decide whether or not to install conservation measures prior to any audit, but that customers who plan to install measures request an audit simply to obtain information on expected costs, and so on. That is: some customers decide to install conservation measures, and these customers request an audit to obtain the audit information. Other customers decide not to install conservation measures and as a result do not request an audit.

In this situation a standard comparison of the adoption rates of audited and nonaudited customers will lead to very misleading results. All of the audited customers will be observed to take conservation actions (since they had requested the audit after deciding to take the actions), and all nonaudited customers will be observed not to take conservation actions. Since the adoption rate is 100% among audited customers and 0% among nonaudited customers, the standard methods will suggest that the program was tremendously effective in inducing customers to take actions, when in actuality it was not effective at all (the program provided information but did not affect any customers' decisions.) In this case, self-selection bias is in the direction of making a program appear more effective than it really is.

The opposite direction of bias is also possible. Consider again an audit program, but now suppose that some customers decide, prior to an audit, to take conservation measures and that these customers do not request an audit since they know already that they are going to adopt the measures. Other customers do not decide prior to an audit; suppose that all of these customers request audits to aid in their decisions. Finally, suppose that the audit is very effective, convincing two-thirds of the audited customers to take the measures.

In this situation, two-thirds of the audited customers are observed to take conservation measures. However, all nonaudited customers are observed to take conservation measures (since they did not request an audit because they had decided previously to take the actions). The adoption rate is therefore 67% for audited and 100% for nonaudited customers. Since the adoption rate is lower for audited customers than nonaudited customers, standard methods will suggest that the audit program has a negative impact, when in actuality the program has a very large impact on those customers who request an audit.

Both of these situations are obviously extreme. For example, in the second situation it is more likely that the estimated impact would be smaller than the actual impact, but not that it will result in a negative impact. However, the story told in each can be expected to occur, to some degree, in most conservation programs. In the next section we propose a method for estimating the true impacts of programs in the face of this self-selection.

3. METHOD FOR CONSISTENT ESTIMATION

The task is to estimate the choice model in Equation 1 given that the dummy variable d_n indicating participation in the program is self-selected. A method of moments (MOM) estimator for this model is the value of the parameters that satisfy the following first order equation:

$$(2) \qquad \sum_n (k_n - P_n) * W_n = 0$$

where k_n is a dummy that indicates whether customer n adopted the conservation measure, P_n is the probability given by the model in Equation 1 for this customer, and W_n is a vector of weights, or "instruments," relating to customer n. Amemiya (1983) discusses the properties of MOM estimators. As motivation, note that if all the explanatory variables in the choice model are exogenous and the instruments are specified to be these explanatory variables, then MOM is equivalent to maximum likelihood estimation, which is the standard procedure for estimating choice models. That is, the standard procedure for estimating the model of Equation 1 is to set $W_n = (d_n, x_n)$ and find the set of parameters that satisfy Equation 2. MOM is therefore a generalization of maximum likelihood in the case of all exogenous variables. However, it adapts more readily to situations with endogenous (self-selected) variables.

If the instruments W_n are uncorrelated with residuals $k_n - P_n$, then the MOM estimator is consistent. Efficiency is increased to the extent that the instruments are close to the explanatory variables (Amemiya, 1983). Instruments that satisfy these conditions can be obtained in our situation by using the probability of participating as an instrument instead of the dummy indicating participation. That is, if d_n were self-selected, then the appropriate instruments would be $W_n = (d_n, x_n)$; however, since d_n is endogenous, we use instruments $W_n = (R_n, x_n)$, where R_n is the probability of participating in the program.

The probability of participating is obtained by estimating a model of the choice to participate or not, using the sample of participants and nonparticipants. This model, if specified as logit, takes the form:

$$(3) \qquad R_n = \frac{e^{hz_n}}{1 + e^{hz_n}}$$

where z_n is a vector of exogenous variables, including customer characteristics. Since R_n is a function of exogenous variables, R_n itself is exogenous and hence is eligible as an instrument in the estimation of Equation 1.

Given instruments, the MOM estimator is calculated fairly simply. Equation 2 is equivalent to the first order condition for the two-stage nonlinear least squares (2SNLS) estimator of the model

$$d_n = P_n + e_n$$

with instruments W_n. That is, the parameters that minimize the expression

$$(\Sigma\ (d_n - P_n)\ W_n')(\Sigma\ (d_n - P_n)\ W_n\ )$$

are the 2SNLS estimators, and the first order condition for minimizing this expression is equivalent to Equation 2. Standard estimation packages, such as SAS, contain 2SNLS estimation and can be used to estimate the choice model, Equation 1.

In summary, the following two steps are taken to estimate the choice model given that the participation dummy is self-selected:

1. Estimate a logit model of customers' decision to participate or not, using standard routines for logit estimation.

2. Estimate a logit model of customers' decision to adopt conservation measures. Include as an explanatory variable in this model a dummy variable that indicates whether the customer participated in the program, plus other variables such as measure costs and savings and customer characteristics. Estimate this model by 2SNLS, using as instruments the probability of participating from step 1 (instead of the participation dummy itself), plus all the other explanatory variables in the model.

4. APPLICATION

In this section we apply the method to data from an audit program. Audits were offered on request to commercial buildings. The audits were intended to promote cost-effective conservation actions of various types, including insulation, HVAC modifications, hot water heating improvements, more efficient refrigeration, automatic energy controls, and other actions. We concentrate in this paper on hot water measures, using the method described above to estimate the impact of the program on customers' decisions to adopt efficient water heating measures. Similar analysis can be performed on each type of conservation action.

We examine a sample of audited customers plus a random sample of customers that had not been audited. For each customer, information was obtained on whether the customer had adopted efficient water heating measures (subsequent to the audit for audited customers, and during the past year for nonaudited customers). To control for differences across customers in other factors, data were obtained on the building characteristics, business type, operation characteristics, perceived energy efficiency, and whether the owner plans to sell or renovate the building. The list of explanatory variables that enter the models is given in Table I.

We first estimate a logit model of the customer's decision of whether to request an audit. This model gives the probability of the customer requesting an audit as a function of the customer's characteristics. It is estimated with standard logit estimation packages on the sample of audited and non-audited customers; the estimation routine determines the characteristics that most readily differentiate the customers that were audited from those that were not. The results are given in Table II. The most significant variable is BUSSIZE, which captures the size of the customer in terms of number of employees, number of apartments, or number of office units, whichever is most appropriate for the customer's business (the exact definition is given in Table I). Since this variable enters with a positive sign, the model indicates that customers with numerous employees/apartments/offices are more likely to request an audit than comparable buildings with fewer of these. Note that SQFT, which denotes the square footage of the building, enters with a negative sign, indicating that large buildings are less likely to request a audit than smaller buildings, if all other things (including number of employees/apartments/offices) are held constant. These two results combined suggest that buildings with more employees per square foot, or smaller apartments and offices, are more likely to request an audit that those with fewer employees per square foot, or with more spacious apartments or offices.

This model is used to calculate for each sampled customer, using the formula in Equation 3, the probability that the customer requests an audit. Since the explanatory variables in this model can be considered exogenous, the calculated probability, which is a function of these variables, is also exogenous. This probability is used as an instrument in the estimation of a model of the customer's decision to take water heating measures.

The model of whether to adopt water heating measures is specified as logit, as given in Equation 1, and includes as an explanatory variable a dummy that indicates whether the customer was audited. Since this variable is self-selected, the model is not estimated by standard procedures for logit models. Rather, it is estimated by 2SNLS with the instruments being the probability of requesting an audit plus all the other explanatory variables in the model. The results are given in Table III, column a. For comparison, the model estimated with standard procedures, which do not account for self-selection, are given in column b.

The variable of interest is the dummy that indicates that the customer was audited. Without correction for self-selection, this variable enters with a positive coefficient that is significant at the 90% confidence level. This would suggest that the audit had a definite, positive impact on the decision to take water heater measures. That is, the result would suggest that we could state with 90% confidence that the audit increases customers' likelihood of taking these conservation measures. However, these results are biased, due to self-selection in the decision to be audited.

When the model is estimated consistently, with self-selection accounted for, then the results are somewhat different. Consider the estimated coefficient of the audit dummy first, then its t-statistic. The estimated coefficient is positive; in fact, it is more positive than without correction for self-selection. This indicates that, in this situation, self-selection bias takes the form of making the program look less effective than it actually is. Following the two examples of self-selection bias given in Part 2, it seems that the second type is more applicable in this particular situation.

The t-statistic indicates, however, that the estimated coefficient is not significant at any reasonable level of confidence. This suggests that, while our best estimate of the audits' impact is positive, we cannot be very confident that the audits actually had an impact. In fact, a comparative lack of confidence seems quite reasonable: customers' decisions to take conservation actions are closely related to their decisions to request an audit, such that confidently estimating the impact of one on the other requires more information than would be needed if customers did not have a choice of whether to be audited. The high level of significance, or confidence, that is suggested by the standard estimation procedure overestimates our ability to isolate the impact of audits on conservation actions. More information, such as larger sample sizes and greater understanding of the customer's decision to be audited, are required to attain the same level of confidence that would be possible if there were no self-selection.

In summary, we have presented and applied a method for estimating the impact of audits on conservation actions when the customer decides whether or not to be audited. We have compared our estimates with those obtained by standard estimation methods, which are biased in this situation since they do not account for the fact that the customer chooses to be audited. We found in this particular application that the point estimates imply that the audits were actually more effective than indicated by the standard methods. That is, the self-selection bias in this situation is downward. The t-statistics indicate, however, that we have considerably less confidence that the audits have an impact than would be suggested by the standard methods.

Table I. Definitions of variables.

Variable Group	Variable Name	Definition
Audit Variables	AUDIT	Was an audit performed
Size Variables	SQFT	Square footage of building, in thousands
	STORIES	Number of stories in building
	BASEMENT	Is there a basement in building
	BUSSIZE	Size of business (small, medium, or large; based on number of apartments and offices for apartment and office buildings, based on number of employees for other buildings)
	APTDUMMY	Was number of apartments or offices left blank on survey form?
Building Type	OFFICE	An office building
	APT	An apartment building
	RETAIL	A retail building
	RELIG	A religious building
	SERVICE	A service building
	REST	A restaurant
Energy Factors	ENCOSTS	The percent of total operating costs spent on energy
	BLGEE	Perception of building energy efficiency (1—very efficient, 5—very inefficient)
	MONITOR	Is there a person to monitor energy use?
Operation Characteristics	OWNBLDG	Own or lease the building
	OWNDEC	Is the owner the decision maker for energy expenditures
	HRSOPEN	Number of hours per week open
	HRSDUMMY	Was HRSOPEN left blank
	TENURE	Number of months in building
Plans	PLNSELL	Are there plans to sell the building in the next few years
	PLNRENO	Are there plans to renovate the building in the next few years

Table II. Logit model results of whether building was audited.

	Variable Name	Coefficient	t-Statistic
Size Variables	SQFT	-0.060	-5.22
	STORIES	0.174	1.24
	BASEMENT	-0.129	-0.79
	BUSSIZE	2.140	14.50
Building Type	OFFICE	-0.216	-0.75
	APT	-4.978	-8.87
	RETAIL	0.698	2.34
	RELIG	0.177	0.52
	SERVICE	0.068	0.23
Energy Factors	ENCOSTS	0.018	3.79
	BLGEE	0.216	3.25
	MONITOR	0.882	5.44
Operation Characteristics	OWNBLDG	-0.092	-0.47
	OWNDEC	0.274	1.59
	HRSOPEN	-0.004	-1.47
	HRSDUMMY	5.173	9.42
	TENURE	-0.000	-1.89
Plans	PLNSELL	-0.042	-0.18
	PLNRENO	0.157	0.79
Constant	CONSTANT	-3.550	-7.34

-LOG LIKELIHOOD	
AT 0	926.7
AT CONVERGENCE	591.2
NUMBER OF CASES	
AUDITED	793
NOT AUDITED	544
TOTAL	1337

Table III. Logit model of whether customer took water heating measure.

	(a) Consistent Estimation		(b) Biased Estimation	
Variable	Coefficient	t-Statistic	Coefficient	t-Statistic
CONSTANT	-2.668	-5.57	-2.668	-5.70
AUDIT	0.538	0.13	0.474	1.75
OFFICE	0.270	0.72	0.292	0.85
APT	0.760	0.32	0.750	1.54
RETAIL	0.637	1.10	0.639	2.04
RELIG	0.998	2.68	1.000	2.79
SERVICE	0.493	1.39	0.496	1.55
STORIES	0.110	0.80	0.110	0.94
ENCOSTS	-0.001	-0.10	-0.000	-0.21
OWNMAN	-6E-5	-0.118	-6E-5	-0.19
SQFT	0.011	0.42	1E-5	1.44
MONITOR	0.489	1.00	0.492	3.38
OWNDEC	0.137	0.70	0.139	0.85
BASEMENT	-0.029	-0.16	-0.030	-0.19
HRSOPEN	0.004	1.34	0.004	1.68
SIZE	0.187	0.17	0.192	1.73
OWNBLDG	-0.320	-1.73	-0.322	-1.76
BLGEE	-0.080	-0.62	-0.079	-1.23
PLNSELL	-0.175	-0.71	-0.176	-0.71
PLNRENO	0.068	0.33	0.068	0.36
HRSDUMMY	-0.638	-0.21	-0.584	-1.34
APTDUMMY	-0.136	-0.30	-0.090	-0.33

References

Amemiya, T., 1983, "Nonlinear Regression Model", in Griliches and Intriligator (eds.), Handbook of Econometrics, Vol. 1, Ch. 6, North-Holland.

Dubin, J. and D. McFadden, 1984, "An Econometric Model of Residential Electric Appliance Holdings and Consumption," Econometrica, Vol. 52, No. 2, pp. 345-362.

Heckman, J., 1978, "Dummy Endogenous Variables in a Simultaneous Equation System," Econometrica, Vol. 46, No. 6, pp. 931-959.

Heckman, J. 1979, "Sample Selection Bias as a Specification Issue," Econometrica, Vol. 47, pp. 153-62.

Hirst, E., B. Bronfman, R. Goeltz, J. Trimble, and D. Lerman, 1983, "Evaluation of the BPA Residential Weatherization Pilot Program," Oak Ridge National Laboratory, paper number ORN CON-124.

Keating, K., 1988, "Self-Selection Bias: Are We Beating a Dead Horse?" Evaluation Program Planning, forthcoming.

McFadden, D., 1973, "Conditional Analysis of Qualitative Choice Behavior." Frontiers in Econometrics, (ed.) P. Zarembka, pp. 105-42. New York: Academic Press.

Trimble, J. and E. Hirst, 1983, "Energy Use in Institutional Buildings: Estimates from State Energy Audit Surveys," Journal of Business and Economic Statistics, Vol. 1, No. 4, pp. 337-347.

Train, K., 1986, Qualitative Choice Analysis, MIT press.

Train, K. and P. Ignelzi, 1987, "The Economic Value of Energy-Saving Investment by Commercial and Industrial Firms," Energy, Vol. 12, No. 7, pp. 543-553.

Train, K., 1988, "Incentives for Energy Conservation in the Commercial and Industrial Sectors," Energy Journal, in Press.

Williams, M. and R. Walther, 1982, "Issues in the Evaluation of Conservation Programs: Nonrandom Data and Participation Bias," in E. Hirst (ed.), Workshop Proceedings: Measuring the Effects of Utility Conservation Programs, Electric Power Research Institute Report No. EA-2496, Palo Alto, CA.