Net Impact Analysis for Program Evaluation Modeling and SAS® Programming

Boqing Wang, Washington State Research and Data Analysis Olympia, WA 98504 USA, (360)902-0701

ABSTRACT

This paper discusses alternative methods for estimating the benefits of social programs with an application to Demand Side Management (DSM) in the utility industry.

INTRODUCTION

A vital issue in estimating program net impact is self-selection bias. Selection bias results from estimation on a subsample of individuals who have essentially selected themselves for estimation through their decision to participate in a particular program. Several techniques have been developed to correct for this bias, most notably a two-stage technique attributed to Heckman (1976). In the Heckman approach, a correction term is calculated from estimates in the first stage and used as a regressor in the second. While successful at correcting bias in the parameters, this approach has several shortcomings when compared to a Full Information Maximum Likelihood (FIML) approach. FIML estimation yields more efficient and robust parameter estimates relative to a two-stage method. In addition, a computer algorithm has been developed in SAS/IML®. by the author to perform FIML estimation on either a two or three equation system under logistic distribution assumptions. The remainder of this paper details major two-stage and Wang's FIML approaches and shows that the FIML estimation is superior in evaluating DSM program net impact. The *SAS* programs are attached in the appendix.

Description of the Problem

Consider a situation where a utility offers a program in which participants undertake defined measures in order to conserve energy. Some examples include installing high-efficientcy equipment (e.g. lighting, cooling, etc.) and improving structural insulation. Customers choose voluntarily to participate in the program. Because customers who participate in the program are generally different from those that do not, a comparison of participants in a program with a sample of non-participants does not provide an accurate estimate of program net impact. In other words, it is reasonable to assume that participants would be more likely to adopt high-efficientcy equipment than would non-participants even if the program did not exist. Therefore, the estimated difference in energy savings between participants and non-participants is a biased estimate for program net impact. The following equation relates energy savings of end use *k*, S_k, to a vector of exogenous characteristics X_k; such as market conditions, site characteristics like square footage, window area, or roof type, economic and demographic characteristics of the occupants including income, household size, and weather conditions.

$$\boldsymbol{S_k = X_k}\beta_{\boldsymbol{k}} + \varepsilon_{\boldsymbol{k}}$$

where ε_k is a disturbance term. Subscript *k* is omitted to make the context simpler. The net impact of participating in the program is defined as the difference in the expected savings of participants who participate in the program and participants who had not participated in the program. More formally, this is:

$$\boldsymbol{E[S_p | X, P = 1] - E[S_p | X, P = 0]}$$

where S_p is the savings achieved by participants and P is a dummy variable such that P = 1 if the individual participates in the program; P = 0, otherwise. $E(S_p | X, P=0)$ is the savings realized by participants if they had *not* participated in the program and is unobservable. The fact that $E(S_p \mid X, P=0)$ is not observable requires substituting $E(S_p \mid X, P=0)$ with the observable $E(S_{np} \mid X, P=0)$, where S_{np} is the savings realized by the non-participants. This, however, produces an evaluation bias equal to:

$$BIAS = E[\,S_p \mid X, P = 0\,] - E[\,S_{np} \mid X, P = 0\,]$$

That is, the savings that would be realized by an average non-participant are different from the savings realized by an average participant if that person had not participated in the program.

A model commonly employed in evaluating program impacts is the following:

$$S = X\beta + \delta P + \varepsilon$$

where P is the participation dummy variable defined above. The estimate of δ is interpreted as the program net impact.

The decision to participate or not participate, however, cannot be treated as an exogenous variable since participation is dependent upon individual self-selection. A person's expectation of savings, S, has an impact on her or his decision to participate.

A less restrictive form of the model can be represented by the following system of equations:

$$P^* = X_1\beta_1 + \gamma S + \varepsilon_1 \qquad (1)$$
$$S = X_2\beta_2 + \delta P + \varepsilon_2 \qquad (2)$$
$$P = 1 \;\; iff \;\; P^* > 0$$
$$P = 0 \;\; iff \; P^* \leq 0$$

where $X_1\beta_1$ is a set of features relating to the decision to participate and $X_2\beta_2$ is equivalent to $X\beta$ defined before. Equation (1) states that the expectation of savings realized by program participation may affect a person's decision to participate; that is, the decision to participate is endogenous in equation (2). P is correlated with the error term ε_2. The estimate of δ is no longer unbiased as is the ordinary least squares estimate.

Two-Stage Correction Term Methodology

The most widely used program evaluation methodology is the two-step process employing a correction term that was originally developed by the Heckman (1976) two-stage method for censored sample regression. This technique utilizes a reduced form of equations (1) and (2) as shown below.

$$P^* = X_1\beta_1 + \varepsilon^*{}_1 \qquad (1a)$$
$$S = X_2\beta_2 + \delta P + \theta\hat{\lambda} + \varepsilon^*{}_2 \qquad (2b)$$

The methodology can be summarized as follows:

- A binary participation variable P is estimated on the total sample of participants and non-participants in equation (1a). Logit or probit analysis is the most commonly used method in this estimation.
- The predicted participation probabilities are used to calculate a correction term, which is derived below.
- The correction term $\hat{\lambda}$ is then entered into the energy savings model (equation (2b)) as a regressor.
- The energy savings model is then estimated via ordinary least squares on the total sample and used to simulate the net impact of the program. The estimate of δ is the program net impact (See the appendix for SAS program).

Certain problems are inherent in this process and in the correction term itself. The estimate of the

Heckman correction term for P = 1 and P = 0 can be written as follows when ε^*_1 follows a logistical distribution:

$$\hat{\lambda} = P \cdot [(1+e^{(-X_1\hat{\beta}_1)})\, ln(1+e^{(X_1\hat{\beta}_1)}) - X_1\hat{\beta}]$$
$$+(1-P)\cdot[e^{(X_1\hat{\beta}_1)} \cdot X_1\hat{\beta}_1 - (1+e^{(X_1\hat{\beta}_1)})$$
$$ln(1+e^{(X_1\hat{\beta}_1)})]$$

The simple application of the Heckman two-stage method creates a problem in that the correction term ($\hat{\lambda}$) is a function of participation (P). This correlation creates problems when estimating the coefficients for participation (P) and the correction term ($\hat{\lambda}$) in the energy savings model. These problems include incorrect signs and implausible magnitudes for the estimate of net impact .

Other Two-Stage Methods

Train (1994) proposes an Instrumental Variable (IV) method to mitigate the endogeneity of the regressor P. The program participation equation is estimated at the first stage. At the second stage, the energy saving model is estimated via ordinary least squares with the regressor of the predicted P from the first stage (see the appendix for SAS program).

$$S = X_2\beta_2 + \delta\hat{P} + \varepsilon_2 \qquad (2b^*)$$

The estimate of δ is an unbiased estimate of net impact. Train's method does not provide a robust estimate because estimation error at the first stage directly adds to the net impact estimate at the second stage. That is, a poor prediction of P leads to a poor estimate of δ. In addition, the participation equation generally has few exogenous variables that are not also in the energy savings equation, a fact that does not help to mitigate the endogeniety of P in the energy savings equation. Thus, Train's method has limited use in practice.

Another approach for net impact analysis is to estimate the pre- and post-period difference and the fixed effect of each participant using a panel data regression (Jacobson and LaLonde,1996). can The Jacobson-LaLonde method is being used for the workforce training program evaluation in Washington state. To apply the Jacabson-LaLonde method to the DSM analysis, at the first stage a conditional demand model is estimated with the inclusion of high-efficientcy equipment adoption or structural insulation improvement.

$$KWH_{it} = \alpha_i + \sum_{k=1}^{K} UEC_{ikt}(ECM_{ikt}, SC_i, AF_i,$$
$$EDC_i, LS_t, WC_{it}, MC_t, \varepsilon_{ikt})D_{ikt} \quad (2c)$$

where UEC_{ikt} represents consumption of the kth end use and D_{ikt} is a binary variable reflecting the presence of the end use at the site i. α_i is an individual specific fixed effect. ECM_{ikt} is a set of variables representing the presence of energy conservation features such as high-efficientcy equipment or insulation. SC_i consists of a set of site characteristics like square footage, window area, or roof type. EDC_i is a vector of economic and demographic characteristics of the occupants, including income, household size, and other features. LS_t is a categorical variable to capture season load shapes. WC_{it} is an indicator of weather conditions. MC_{it} pertains primarily to energy prices. Note that non-participants may install high-efficientcy equipment k, and participants may install equipment k before the program exists. ECM_{ikt} thus is defined for both participants and non-participants and is less likely endogenous in equation (2c). Equation (2c) can be rewritten as

$$KWH_{it} = \alpha_i + \sum_{k}^{K} [\delta_k ECM_{ik} + \gamma_k LS_{ikt} + X_{i2}\beta_2$$
$$+\varepsilon_{ikt}]D_{ikt} \qquad (2c^*)$$

δ_k is explained as savings realized by high-efficientcy equipment installation or insulation adoption. At the second stage, an adoption equation is estimated for each installation of more efficient equipment or insulation.

$$\begin{aligned} P^*{}_k &= X_{1k}\beta_{1k} + \varepsilon_{1k} \\ \mathrm{P_k} &= 1 \quad \text{iff } P^*{}_k > 0 \\ \mathrm{P_k} &= 0 \quad \text{iff } P^*{}_k \le 0 \end{aligned} \tag{1c}$$

where $\mathrm{P_k}$ =1 if a participant adopts high-efficientcy equipment k, and $\mathrm{P_k}$ =0, if a non-participant adopts high-efficientcy equipment; k=1, ... K. Finally, the net program impact on end use k is computed by (see the appendix for SAS program):

$$\begin{aligned} &\mathit{Im\,pact} = \\ &\mathit{Pr\,ob}(\mathit{participation} \cap \mathit{adoption\ of\ \ equipment\ k}) \\ &\times \mathit{Savings\ of\ \ Equipment\ k} \\ &= \bar{\hat{P}}^*{}_k \times \hat{\delta}_k \end{aligned}$$

This model provides more robust estimates than do the previous two-stage methods. However, this two-stage method does not lead to an efficient estimate for net impact because it does not capture the mutual impact process of expectation for energy savings and decision to participate.

FIML Estimation

Since a person's expectation for savings affects the decision to participate and participation affects her energy savings, a structural form of the model for this process is represented by equations (1) and (2) . The reduced form is:

$$\begin{aligned} P^* &= \frac{X_1\beta_1 + \gamma X_2\beta_2}{1-\gamma\delta} + \frac{\gamma\varepsilon_2 + \varepsilon_1}{1-\gamma\delta} \\ S \ &= \frac{X_2\beta_2 + \delta X_1\beta_1}{1-\gamma\delta} + \frac{\delta\varepsilon_1 + \varepsilon_2}{1-\gamma\delta} \end{aligned}$$

$$\begin{aligned} \mathit{Let} \qquad \mathrm{V}_1 &= \frac{X_1\beta_1 + \gamma X_2\beta_2}{1-\gamma\delta}\ ; \\ \mathrm{V}_2 &= \frac{X_2\beta_2 + \delta X_1\beta_1}{1-\gamma\delta} \\ \upsilon_1 &= \frac{\gamma\varepsilon_2 + \varepsilon_1}{1-\gamma\delta}\ ; \qquad \upsilon_2 = \frac{\delta\varepsilon_1 + \varepsilon_2}{1-\gamma\delta} \end{aligned}$$

$$\begin{aligned} \mathit{So} \qquad P^* &= V_1 + \nu_1 \\ S \ &= V_2 + \nu_2 \end{aligned}$$

Error terms ν_1 and ν_2 are assumed to follow a bivariate probability distribution to model the mutual impact process of the expectation of savings and program participation. Since the logistical distribution can provide a robust estimate relative to normal distribution, ν_1 and ν_2 are assumed to follow a bivariate logistical distribution. Thus,

$$\begin{aligned} \mathbf{Prob(P^* \ge 0,}\ S \ge s) &= \frac{e^{\frac{V_1}{\lambda}}}{e^{\frac{V_1}{\lambda}} - e^{\frac{(V_2-s)}{\lambda}}} \cdot \frac{1}{1+e^{-(V_2-s)}} \\ &- \frac{e^{\frac{(V_2-s)}{\lambda}}}{e^{\frac{V_1}{\lambda}} - e^{\frac{(V_2-s)}{\lambda}}} \cdot \frac{1}{1+e^{-V_1}} \end{aligned}$$

where $\lambda = \sqrt{1\text{-}\rho^2}$, and ρ is the correlation coefficient of ν_1 and ν_2.

$$f(P^* \ge 0,\ S = s) = -\frac{\partial\, \mathit{Pr\,ob}(P^* \ge 0,\ S \ge s)}{\partial s}$$

$$= \frac{1}{l}\left(\frac{e^{\frac{(V_1+V_2-s)}{l}}}{\left(e^{\frac{V_1}{l}} - e^{\frac{(V_2-s)}{l}}\right)^2} \left(\frac{1}{1+e^{-(V_2-s)}} - \frac{1}{1+e^{-V_1}} \right) \right)$$

$$+ \left(\frac{e^{\frac{V_1}{l}}}{e^{\frac{V_1}{l}} - e^{\frac{(V_2-s)}{l}}} \right) \cdot \left(\frac{e^{-(V_2-s)}}{\left(1+e^{-(V_2-s)}\right)^2} \right)$$

$$f(P^* < 0, S = s) = \frac{e^{-(V_2 - s)}}{\left(1 + e^{-(V_2 - s)}\right)^2} - f(P^* \geq 0, S = s)$$

The log-likelihood function is:

$$\sum_{i=1}^{N} [P_i \cdot log[\ f(P_i = 1, S = s_i)] + (1 - P_i) \cdot log[\ f(P_i = 0, S = s_i)]]$$

This likelihood function is maximized to simultaneously estimate the coefficients of both the participation and the savings equation (β_1, β_2, δ, γ). Since this simultaneous estimation models the mutual impact process of savings expectation and participation decision-making, the FIML estimate is efficient and robust relative to a two-stage method[1]. The author has developed SAS programs of IML and NLIN to maximize a likelihood function.

CONCLUSION

The benefit of estimating the parameters simultaneously is that there is no information loss. This is important in the evaluation of DSM programs because the decision to participate is based on the expected energy savings associated with participating in the program. Energy savings (S) and program participation (P) are both endogenous in equations (1) and (2), and should therefore be estimated simultaneously. FIML estimation accurately captures this simultaneity. It also avoids the estimation problems encountered when using a two-stage approach. The approach introduced and developed in this paper can be applied to other program/policy impact analysis. In fact, the Heckman and Jacobson-LaLonde methods originally were applied to female labor supply and education program evaluation.

REFERENCES

Cardell, N. S. (1989), "Extensions of Multinomial Logit: the Hedonic Demand Model, the Non-independent Logit Model, and the Ranked Logit Model", Ph.D. dissertation, Harvard University.

Griffin, J. (1993), “Methodological Advances in Energy Modeling: 1970-1990” *The Energy Journal*14(1):111-124

Heckman, J. J. (1976), "The Common Structure of Statistical Models of Truncation Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models", *Annals of Economic and Social Measurement,* 5/4, 1976.

Jacobson, L, R. J. Ladonde, D. Sullivan (1993) “Earnings Losses of Displaced Workers”, *American Economic Review,* September 1993, 685-709.

SAS Institute Inc. (1993), *SAS/ETS User’s Guide, Version 6 Second Edition,* Cary, NC*: SAS* Institute Inc.
SAS Institute Inc. (1990), *SAS/IML Software, Version 6 First Edition,* Cary, NC*: SAS* Institute Inc.

SAS Institute Inc. (1990), *SAS/STAT User’s Guide, Version 6 Fourth Edition*, Cary, NC*: SAS* Institute Inc.

SAS Institute Inc. (1990), *SAS Technical Report P-229,Release 6.07,* Cary, NC:*SAS* Institute Inc.

Train, K. E.(1994)"Self-Selection Bias in A new Context: Estimating the Impact of Conservation Programs on Measure Adoption", Working Paper, U.C. Berkeley.

[1] Wang (1984).

Wang, B(1994) " Maximum Likelihood Estimation with Sample Selection", Ph.D. Dissertation, Washington State University.

ACKNOWLEDGMENTS

The author gratefully acknowledges the valuable advice and assistance of Ernst Stromsdorfer and Scott Cardell at Washington State University, Fred Sebold at Regional Economics Research, Inc., and Paul Zak at Claremont Graduate School.

APPENDIX

A SAS Program Example for Heckman Two-Stage Method

```
data savings;
set   savings;
proc logistic  data=savings descending;
       model pk= totsqft type weather
                         utility/link=logit ;
      output out=logit xbeta=xbetahat;
data  logit;
set    logit;
t=exp(xbetahat);
if pk=1 then   lambda=(1+1/t)*log(1+t)-log(t);
else if pk=0 then lambda=t*log(t)-(1+t)
*log(1+t);
proc reg  data=logit outest=param;
       model savingsk=totsqft type weather  p
                                 lambda ;
       output out=saving p=psavingks;
```

A SAS Program Example for Train’s IV Method

```
proc logistic  data=savings descending;
       model pk= totsqft type weather
                         utility/link=logit ;
output out=logit p=pkhat;
proc reg  data=logit outest=param;
       model savingsk=totsqft type weather
                             phatk;
      output out=saving p=psavingsk;
```

A SAS Program Example for Jacobson-LaLonde Method

```
proc tscsreg  data=savings outest=param;
      model svaings=alpha1-alphan ls sqft type
                           weather ecm1-ecmk ;
      id id month;
proc logistic  data=savings decending;
      model pk= totsqft type weather
                      utility/link=logit ;
      output out=logit p=phatk ;
 where  ecmk=1;
proc summary data=logit;
       var phatk;
      output out=mean mean=phatmean;
where  ecmk=1;
data param;
merge param mean;
impactk=ecmk*phatmean;
```

SAS *PROC MIXED* can be used for unbanlanced time series data.

SAS Programs for Wang’s FIML Model (on request)