

**Net Impact Analysis for Program Evaluation
Modeling and SAS Programming**

Author

*Boqing Wang
Washington State Workforce Training and Education Coordinating Board
Olympia, WA 98504*

To be presented at The 15th Pacific Northwest SAS Users Conference at Seattle, June 1996

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.

Introduction

This paper discusses alternative methods for estimating the benefits of social programs with an application to Demand Side Management (DSM) in the utility industry. 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-Interactive Matrix Language (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 efficiency 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 efficiency equipment than non-participants would 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.

$$S_k = X_k \beta_k + \epsilon_k$$

where ϵ_k is a disturbance term. X_k . 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 do not participate in the program. More formally, this is:

$$E[S_p | X, P = 1] - E[S_p | X, P = 0]$$

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 $E(S_p | X, P=0)$

is not observable requires substituting $E(S_p | X, P=0)$ with the observable $E(S_{np} | X, P=0)$, where S_{np} is the savings realized by the non-participants. This, however, produces an evaluation bias equal to:

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

That is, the savings 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.

$$\begin{aligned} E[S_p | X, P = 1] - E[S_p | X, P = 0] &= \frac{1}{n} \sum_{i=1}^n [X_{pi} \hat{\beta} + \hat{\delta} * 1] - \frac{1}{n} \sum_{i=1}^n [X_{pi} \hat{\beta} + \hat{\delta} * 0] \\ &= \hat{\delta} \end{aligned}$$

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:

$$P^* = X_1\beta_1 + \gamma S + \varepsilon_1 \quad (1)$$

$$S = X_2\beta_2 + \delta P + \varepsilon_2 \quad (2)$$

$$P = 1 \text{ iff } P^* > 0$$

$$P = 0 \text{ iff } P^* \leq 0$$

where $X_1\beta_1$ is a set of features relating 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 . Let $\beta^*_2 = (\beta_2, \delta)$ and $W'_2 = (X_2, P)$. The estimate of β^*_2 is no longer unbiased as the ordinary least squares estimate is. The OLS estimate of β^*_2 is:

$$\hat{\beta}^*_2 = \beta^*_2 + (W'_2W_2)^{-1}W'_2\varepsilon_2$$

and the expectation is:

$$E[\hat{\beta}^*_2] = \beta^*_2 + E[(W'_2W_2)^{-1}W'_2\varepsilon_2]$$

where

$$E[\hat{\beta}_2^*] \neq \beta_2^* \quad \text{if } E[(W_2'X_2)^{-1}W_2'\varepsilon_2] \neq 0$$

Note that

$$E[(W_2'W_2)^{-1}W_2'\varepsilon_2] \neq E[(W_2'W_2)^{-1}W_2']E(\varepsilon_2) \quad \text{if } P \text{ in } W_2 \text{ is stochastically correlated with the error term } \varepsilon_2$$

OLS regression estimation produces a biased estimate for β_2^* .

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 two-stage method for censored sample regression.

This technique utilizes a reduced form of equations (1) and (2) as shown below.

$$P^* = X_1\beta_1 + \varepsilon_1^* \quad (1a)$$

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

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).

As will be shown, there are problems inherent in this process and the correction term itself.

The Heckman two-stage method decomposes the error term ε_2 into its expectation on subsamples and the deviation from their expectation. From equation (2), this can be written as follows:

$$\varepsilon_2 = P * E[\varepsilon_2 | P = 1] + (1 - P) * E[\varepsilon_2 | P = 0] + \varepsilon_2^*$$

where $E[\varepsilon_2 | \cdot]$ is a function of the probability of choosing to participate in the program. The deviation ε_2^* is due to factors that are unrelated to the decision to participate. The energy savings equation now becomes:

$$S = X_2 \beta_2 + \gamma P + P * E[\varepsilon_2 | P = 1] + (1 - P) * E[\varepsilon_2 | P = 0] + \varepsilon_2^*$$

Expressions for $E[\varepsilon_2 | \cdot]$ have been derived under various distribution assumptions. The following calculation derives the self selection correction term ($\hat{\lambda}$) when $P = 1$.

$$E[\varepsilon_2|P=1] = E[\varepsilon_2|\varepsilon_1^* > -X_1\beta_1]$$

A linear regression relationship is assumed for the conditional expectation of ε_2 given ε_1 :

$$E[\varepsilon_2|\varepsilon_1^* > -X_1\beta_1] = \theta_1 \varepsilon_1^*$$

The logistic distribution and many other distributions yield a linear representation¹. Therefore:

$$\begin{aligned} E[\varepsilon_2|\varepsilon_1^* > -X_1\beta_1] &= \theta_1 E[\varepsilon_1^*|\varepsilon_1^* > -X_1\beta_1] \\ &= \theta_1 \frac{1}{1 - F(-X_1\hat{\beta}_1)} \int_{-X_1\hat{\beta}_1}^{\infty} \varepsilon_1^* f(\varepsilon_1^*) d\varepsilon_1^* \end{aligned}$$

where $F(\cdot)$ is the cumulative probability function of ε_1^* and $f(\cdot)$ is probability density function of ε_1^* .

The correction term when $P = 0$ is thus:

$$\begin{aligned} E[\varepsilon_2|\varepsilon_1^* \leq -X_1\beta_1] &= \theta_0 E[\varepsilon_1^*|\varepsilon_1^* \leq -X_1\beta_1] \\ &= \theta_0 \frac{1}{F(-X_1\hat{\beta}_1)} \int_{-\infty}^{-X_1\hat{\beta}_1} \varepsilon_1^* f(\varepsilon_1^*) d\varepsilon_1^* \end{aligned}$$

¹ Kagan, A., T. Linnik and C. R. Rao (1973), Some Characterization Theorems in Mathematical Statistics, New York Wiley.

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:

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

The simple application of the Heckman two-stage method creates a problem in that the correction term ($\hat{\lambda}$) is a linear function of participation (P). This linear 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).

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

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 which are also not in the energy savings equation, a fact that does not help to mitigate the endogeneity of P in the energy savings equation. Thus, Train's method has limited use in practice.

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

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

where UEC_{ikt} represents consumption of the k^{th} end use and D_{ikt} is a binary variable reflecting the presence of the end use at the site in equation. 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 efficiency 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_i 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 efficiency 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

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

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

$$\begin{aligned} P^*_k &= X_{ik} \beta_{k1} + \varepsilon_{ik} \\ P_k &= 1 \quad \text{iff } P^*_k > 0 \\ P_k &= 0 \quad \text{iff } P^*_k \leq 0 \end{aligned} \quad (1c)$$

where $P_k = 1$ if a participant adopts high efficiency equipment k , $P_k = 0$, otherwise; $k=1, \dots, K$. Finally, the net program impact on end use k is computed by (See the appendix for SAS program)

$$\begin{aligned} \text{Impact} &= \text{Prob}(\text{participation} \cap \text{adoption of Equipment } k) * \text{Savings of equipment } k \\ &= \hat{P}^*_k * \hat{d}_k \end{aligned}$$

This model provides more robust estimates than the previous two-stage methods do. 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. If there are not enough observations to estimate the measured individual fixed effect α_i , α_i is deleted from equation (2c*) and the individual effect is measured by the error term.

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:

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

Let $V_1 = \frac{X_1\beta_1 + \gamma X_2\beta_2}{1-\gamma\delta}$; $V_2 = \frac{X_2\beta_2 + \delta X_1\beta_1}{1-\gamma\delta}$

$$v_1 = \frac{\gamma\varepsilon_2 + \varepsilon_1}{1-\gamma\delta} ; \quad v_2 = \frac{\delta\varepsilon_1 + \varepsilon_2}{1-\gamma\delta}$$

So $P^* = V_1 + v_1$
 $S = V_2 + v_2$

Error terms v_1 and v_2 are assumed 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, v_1 and v_2 are assumed follow a bivariate logistical distribution. Thus,

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

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 ($\beta_1, \beta_2, \delta, \gamma$). Since this simultaneous estimation models the mutual impact process of savings expectation and participation decision-making, the FIML estimate is efficient and robust relative a two-stage method². 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

² Monte Carlo experiment for FIML and a two-stage method: Wang, B (1994), "Maximum Likelihood Estimation with Sample Selection", Ph.D. Dissertation, Washington State University.

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.
- Greene, H. W. (1993), "Econometric Analysis", Macmillan.
- Griffin, J. (1993), "Methodological Advances in Energy Modeling: 1970-1990" *The Energy Journal* 14(1):111-124
- Hausman, J. A.(1986), "Specification and Estimation of Simultaneous Equation Models", *Handbook of Econometrics*, Vol. III New York: North-Holland.
- 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.
- Heckman, J. J.(1994), "Evaluating the Impact of Training on the Earnings and Labor Force Status of Young Women: Better Data Help a Lot ", Working Paper, University of Chicago.
- Jacobson, L, R. J. Ladonde, D. Sullivan (1993) "Earnings Losses of Displaced Workers", *American Economic Review*, September 1993, 685-709.
- Jacobson, L, R. J. Ladonde, etc. (1996) "Evacuation of Publicly Funded Training for Unemployed Workers in Washington", working paper, Washington State Workforce Training and Education Coordinating Board.
- Lee, L. F., G. S. Maddala and R. P. Trost (1980), "Asymptotic Covariance Matrices of Two-Stage Probit and Two-Stage Tobit Methods for Simultaneous Equation Models with Selectivity", *Econometrica*, Vol 49, 491-503.

Maddala, G.S. (1983), "Limited-Dependent and Qualitative Variables in Econometrics", Cambridge University Press.

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.

Train, Kenneth (1986), "Qualitative Choice Analysis", MIT Press.

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

Zabel, J. E. (1993) "The Relationship between Hours of Work and Labor Force Participation in Four Models of Labor Supply Behavior", *Journal of Labor Economics*, 1993, Vol. 11, No..

Appendix

A SAS Program Example for Heckman Two-Stage Method

```
data savings;

set  savings;

proc logistic data=savings;

    model pk= totsqt 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=totsqt type weather p lambda ;

    output out=saving p=psavingsk;
```

A SAS Program Example for Train's IV Method

```
proc logistic data=savings;

    model pk= totsqt type weather utility/link=logit ;

output out=logit p=pkhat;

data logit;

set  logit;

pkhat=1-pkhat;
```

```
proc reg data=logit outest=param;

    model savingsk=totsqft type weather phatk;

    output out=saving p=psavingsk;
```

A SAS Program Example for Jacobson and 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;

    model pk= totsqft type weather utility/link=logit ;

    output out=logit p=phatk ;

data logit;

set logit;

phatk=1-phatk;

proc summary data=logit;

    var phat;

    output out=mean mean=phatmean;

data param;

merge param mean;

impactk=ecmk*phatmean;
```

SAS *PROC MIXED* can be used for unbalanced time series data. SAS program examples are available on request.

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