On-Site Sampling in Economic Valuation Studies

Termeh Shafie [1]

[1] Department of Statistics, Stockholm University, Sweden
e-mail: termeh.shafie@stat.su.se

Abstract

An overview of models suggested for analysis of on-site data will be presented, with a special focus on the problems of economic valuation. The purpose is partly to present an overview of models for analysis of on-site data and their relations to models for length biased samples. In addition, a new model for estimation of binary choice data from on-site samples is presented.

1 On-Site Sampling

On-site sampling can be an efficient sampling design when it is difficult to construct an effective sampling frame. This sampling design implies that sample inclusion probabilities depend on individual characteristics such as visiting frequency. This in turn leads to an estimation problem when using standard estimation techniques that do not account for the differences in inclusion probabilities.

When sampling is done on-site, the researcher surveys respondents on a specific site e.g., visitors to a park - during the surveying time frame. There are two immediate problems connected to this sampling design. First, the sampling inclusion probabilities will depend on the respondent's visiting frequency to the site. Second, the researcher may have a subjective influence when sampling indivuals at the site. Thus, care needs to be taken when using this sampling method to control for these two issues. Concerning the second issue, it is important to make sure that an objective sample mechanism is used and that randomness is somehow built in the sampling design. An example of how this can be done is by choosing the entrance at the park as a sampling location and using Bernoulli sampling to choose individuals that enter the park.

On-site sampling is often used in research fields such as marketing research, research that typically uses on-site data when applying consumer tests. One example is found in Keillor et al. (2001) where on-site samples are used to explore the notion that consumers around the globe are becoming more similar in terms of psychological consumer tendencies. Sudman(1980) identified procedures for shopping-center sampling that can improve the quality and reduce the biases in shopping center samples.

In a resource economics study that examined fishing methods, Pollock et al. (1994) proposed sampling anglers at fishing sites to collect data on angler effort and catch. This method intends to estimate the characteristics of the population of fishing trips during a season. The sampling unit is "fishing trips"; one trip represents one element in the population. Note that when sampling is done this way, we have a frame based sampling procedure and the estimation is not problematic.

Shaw (1988) also considers the problem of on-site sampling when surveying visitors at recreational sites. However, Shaw (1988) is interested in the sampled individuals and the underlying factors that affect the visiting frequency of each individual. Thus inference is done on the population of *visitors* instead of the population of *visits*. Other contributions to the literature on on-site sampling is given by Santos Silva (1997) and Nunes (2003).

2 Contingent Valuation Methods

In resource economics, contingent valuation methods (CVM) are stated preference methods that are used to estimate economic values for all kinds of environmental services (Bateman et al., 2002). It is called contingent valuation because it directly asks people about their willingness to pay (WTP) or willingness to accept compensation for a specific environmental service. It is performed by directly asking people, in a survey for their willingness to pay or willingness to accept as compensation for a specific environmetal service.

Several CVM methods can be used to to collect data about these kinds of economic valuations: open-ended studies (e.g., Duffield and Allan, 1988) and closed-ended studies (e.g., Cameron and James, 1987). One popular CVM is the binary choice model where dichotomous choice valuation questions are given to the respondents. Hanemann (1984) and McFadden (1973) model binary choice CV data with logit or probit models, providing a thorough discussion on the topic.

3 Modelling Economic Valuation with On-Site Samples

When using on-site sampling in CVM, the sampling inclusion probabilities may be correlated with respondent valuation and thus the results may become invalid.

Combined Poisson/Probit model

Nunes (2003) shows that the on-site conditional pdf of y_{1k} and y_{2k} given the explanatory variable x_k is given by

$$f^*(y_{1k}, y_{2k}|x_k) = f(y_{1k}, y_{2k}|x_k)\frac{y_{1k}}{E(y_{1k}|x_k)},$$

where the inclusion probability is proportionate to y_{1k} and y_{2k} is the variable we are interested in. The usual maximum likelihood method employed to estimate binary choice models such as the probit model may result in biased and inconsistent estimators when using on-site samples.

In order to account for on-site sampling, Nunes (2003) proposes a combined Poisson and Binary Probit model. The model is such that the count variable y_1 has the following distribution,

$$y_{1k}|X_k, \epsilon_{1k} \sim Poisson(\lambda(\epsilon_{1k}, X_k)),$$

where $\lambda = exp((X_k'\alpha))v(\epsilon_{1k})$, α denotes a vector of unknown coefficients, $v(\epsilon_{1k}) = exp(\sigma\epsilon_{1k})$, X a vector of explanatory variables and ϵ_1 the error term. The binary choice model is defined by the random variable y_2. The conditional latent variable is defined as

$$y_{2k}^*|\epsilon_{1k} = X_k'\beta + \epsilon_{2k}^* + \rho\epsilon_{1k},$$

where the distribution for the error terms is

$$(\epsilon_{1k}, \epsilon_{2k}) \sim BN(0, 0, 1, 1, \rho).$$

The observable binary variable is defined as

$$y_{2k}|\epsilon_{1ik} = \begin{cases} 1 & \text{if} \quad x'_k\beta + \varepsilon^*_{2k} + \rho\epsilon_{1k} > 0 \\ 0 & \text{if} \quad x'_k\beta + \varepsilon^*_{2k} + \rho\epsilon_{1k} \leq 0. \end{cases}$$

The likelihood for this model is derived from $\prod f^*(y_{1k}, y_{2k}|x_k)$ where $f^*(y_{1k}, y_{2k}|x_k)$ is given in (1) with

$$\begin{aligned} f(y_{1k}, y_{2k}|X - k) &= \int_{-\infty}^{\infty} f(y_{2k}|x_k, \epsilon_{1k}) f(y_{1k}|x_k, \epsilon_{1k}) f(\epsilon_1) d\epsilon_{1k} \\ &= \int_{-\infty}^{\infty} \Phi\left(\frac{x'_{2k}\beta + \rho\epsilon_{1k}}{\sqrt{1-\rho^2}}\right) \frac{e^{-\lambda(x_k)v(\epsilon_{1k})}(\lambda(x_k)v(\epsilon_{1k}))^{y_{1k}}}{y_{1k}!} \frac{e^{-\epsilon_{1k}^2/2}}{\sqrt{2\pi}} d\epsilon_{1k}. \end{aligned}$$

A Binary Ordinal Probit Model[1]

A new proposal for modelling visit frequencies using an ordinal probit model is presented. Here, the frequency of visits is modelled using a latent continuous variable that on crossing specified thresholds determines the number of visits.

The proposed model here is such that there are two latent variables, one for the visit frequency and one for the binary choice. Let the latent variables be y^*_{1i} and y^*_{2i} and let them be related to explanatory variables as

$$\begin{aligned} y^*_{1k} &= x'_{1k}\beta_1 + \epsilon_{1k}, \\ y^*_{2k} &= x'_{2k}\beta_2 + \epsilon_{2k}, \end{aligned}$$

where β_1 and β_2 are vectors of unknown parameters, and ϵ_1 and ϵ_2 are two random terms with distribution $(\epsilon_{1k}, \epsilon_{2k}) \sim BN(0, 0, 1, 1, \rho)$. The latent variables are unobserved and observations are made on y_1 and y_2 where

$$\begin{aligned} y_{1k} &= 1 \quad \text{if} \quad y^*_{1k} \leq c_{11} \\ &= 2 \quad \text{if} \quad c_{11} < y^*_{1k} \leq c_{12} \\ &\vdots \\ &= I \quad \text{if} \quad c_{1I-1} < y^*_{1k}, \end{aligned}$$

and

$$\begin{aligned} y_{2k} &= 1 \quad \text{if} \quad y^*_{2k} \leq c_{21} \\ &= 2 \quad \text{if} \quad c_{21} < y^*_{2k} \leq c_{22} \\ &\vdots \\ &= J \quad \text{if} \quad c_{2J-1} < y^*_{2k}, \end{aligned}$$

[1] Joined work with Thomas Laitila, thomas.laitila@esi.oru.se

The cutoffs satisfy the condition that $c_{11} < c_{12} < \cdots < c_{1I-1}$ and $c_{21} < c_{22} < \cdots < c_{2J-1}$, $c_{10} = c_{20} = -\infty$ and $c_{1I} = c_{2J} = \infty$. Note that in this model $J = 2$ and thus y_2 is the binary choice variable. The general presentation for the binary reponse is given since there are binary choice models with more than two classes.

The probability that $y_{1k} = j$ and $y_{2k} = k$ given x is

$$\begin{aligned} P_x(y_{1k} = i, y_{2k} = j) =& P_x(c_{1k-1} < y^*_{1k} \leq c_{1i}, c_{2j-1} < y^*_{2k} \leq c_{2k}) \\ =& P_x(y^*_{1k} \leq c_{1i}, y^*_{2k} \leq c_{2j}) \\ &- P_x(y^*_{1k} \leq c_{1i-1}, y^*_{2k} \leq c_{2j}) \\ &- P_x(y^*_{1k} \leq c_{1i}, y^*_{2k} \leq c_{2j-1}) \\ &+ P_x(y^*_{1k} \leq c_{1i-1}, y^*_{2k} \leq c_{2j-1}). \end{aligned} \tag{1}$$

Kim (1995) considers the ML estimation of the bivariate ordinal probit model under random (SRS) sampling. The log-likelihood is then given by

$$\ell = \sum_{k=1}^{N} \sum_{i=1}^{I} \sum_{j=1}^{J} I(y_{1k} = i, y_{2k} = j) \text{ln} P_x(y_{1k} = i, y_{2k} = j),$$

where $I(y_{1k} = i, y_{2k} = j)$ is an indicator function.

When sampling is done on-site, the pdf will instead be given by the expression in eq. (1), that is

$$f^*(y_{1k}, y_{2k}|x_k) = f(y_{1k}, y_{2k}|x_k) \frac{y_{1k}}{E(y_{1k}|x_k)}.$$

For the proposed model here, $f(y_{1k}, y_{2k}|x_k)$ is given in eq. (1) and

$$E(y_{1k}|x_k) = \sum_{i}^{T} i \left[\Phi(c_{1i} - x'_{1k}\beta) - \Phi(c_{1i-1} - x'_{1k}\beta), \right]$$

where Φ is the standard bivariate normal distribution function and T is the largest possible value for y_1. An example is the number of days visits are made to recreation parks and fishing sites during a certain time period.

The log likelihood for the on-site sample is given by

$$\ell^* = \sum_{k=1}^{N} \sum_{i=1}^{I} \sum_{j=1}^{J} I(y_{1k} = i), y_{2k} = j)\{\ln f(y_{1k}, y_{2k}|x_k) + \ln y_{1k} - \ln E(y_{1k}|x_k)\}.$$

A Simulation Study

This section reports on a small simulation study on the properties of the estimators associated with the new model proposed and the model proposed by Nunes (2003). The simulation is built on generation of data from two different sets of population models. The first set of models is based on the model used for the simulations in Nunes (2003). Frequency of visits are generated from the model

$$y_{1k}|X_k, \epsilon_{1k} \sim Poisson(\lambda(\epsilon_{1k}, X_k)),$$

where $\lambda = exp(\alpha_0 + \alpha_1 X) exp(\sigma\epsilon_{1k})$. The binary choice model is defined by the random variable

$$y^*_{2k} = \beta_0 + \beta_1 X + \epsilon_{2k},$$

where

$$(\epsilon_{1k}, \epsilon_{2k}) \sim BN(0, 0, 1, 1, \rho).$$

Simulations at sample size $N = 500$ are made for $\rho = (-.7, -.5, 0, .5, .7)$. The remaining parameters of the model are kept fixed at $\alpha_0 = -0.3665, \alpha_1 = -1.5, \beta_0 = -2, \beta_1 = 1$, and $\sigma = 0.8561$. The explanatory variable are generated from a uniform distribution over [1,3]. This set of models is below named as the Population 1 models.

The Population 2 models are generated from a bivariate ordinal probit model. The model for the discrete choice variable y_2 is the same as in the Population 1 models. The latent model for the visit frequency variable is defined as $y_1^* = \alpha_0 + \alpha_1 X_k + \epsilon_{1k}$ where y_{1k}^* is observed as

$$\begin{array}{lll}
y_{1k} = 1 & \text{if} & y_{1k}^* < 1.4 \\
y_{1k} = 2 & \text{if} & 1.4 \leq y_{1k}^* < 2.0 \\
y_{1k} = 3 & \text{if} & 2.0 \leq y_{1k}^* < 2.6 \\
y_{1k} = 4 & \text{if} & 2.6 \leq y_{1k}^* < 3.2 \\
y_{1k} = 5 & \text{if} & 3.2 \leq y_{1k}^* < 3.6 \\
y_{1k} = 6 & \text{if} & 3.6 \leq y_{1k}^*
\end{array}$$

The random terms ϵ_1 and ϵ_2 are generated from the $BN(0, 0, 1, 1, \rho)$ distribution. The parameters used are $\alpha_0 = 4, \alpha_1 = -1, \beta_0 = -2$, and $\beta_1 = 1$ and the explanatory variable is again generated from a uniform distribution over [1,3].

The log-likelihood for the Nunes' model is calculated using a 16-point Gausse Hermite quadrature approximation of the joint probability function $f(y_{1k}, y_{2k}|x_k)$. For both models considered, the log-likelihood is maximized using a direct search using the simplex algorithm AMOEBA (Press et al., 1986). For each population simulated, estimates are derived for 500 replications. Table 1 and table 2 present the bias and MSE estimates for β_0 and β_1.

Table 1: Bias and MSE in Population 1 models.

	Nunes' ML				Bivariate Probit ML			
	Bias		MSE		Bias		MSE	
ρ	β_0	β_1	β_0	β_1	β_0	β_1	β_0	β_1
-0.7	0.0080	-0.0094	0.1480	0.0252	-0.4895	-0.0449	0.3699	0.0437
-0.5	-0.0005	-0.0005	0.0963	0.0169	-0.4605	0.0309	0.2809	0.0599
0	-0.0180	0.0030	0.0789	0.0189	-0.0231	0.0118	0.1085	0.0395
0.5	-0.0124	0.0089	0.0516	0.0181	0.2793	0.0746	0.1579	0.0365
0.7	-0.0217	0.0083	0.0793	0.0223	0.3901	0.1081	0.2276	0.0012

The results presented in Table 1 replicates the simulation results presented by Nunes (2003) for his ML estimator. Bias and MSE are both very low for all of the correlation levels considered. The ML estimator of the bivariate ordinal probit model does not work equally well for the Population 1 models. For zero correlation, biases are small. For non-zero correlations, bias increases with the magnitude

Table 2: Bias and MSE in Population 2 models.

	Nunes' ML				Bivariate Probit ML			
	Bias		MSE		Bias		MSE	
ρ	β_0	β_1	β_0	β_1	β_0	β_1	β_0	β_1
-0.7	1.1888	-0.5148	2.3172	0.4155	-0.3792	0.1156	0.5595	0.0207
-0.5	0.6011	-0.2728	1.2396	0.2233	-0.1716	0.0736	0.2568	0.0589
0	0.5067	-0.0997	0.5048	0.0634	-0.0194	0.0091	0.1278	0.0355
0.5	0.0460	-0.0078	0.2236	0.0765	-0.0328	0.0244	0.1307	0.0370
0.7	0.3516	-0.2126	0.6245	0.2407	-0.1622	0.1078	0.3039	0.0992

of the correlation. The intercept estimator β_0 is more biased than the slope coefficient estimator β_1 (Table 1).

Table 2 presents the performance of the estimators under the Population 2 models. The bivariate ordinal probit ML estimator have relatively small biases for correlations -.5, 0 and .5. For the larger correlations, -.7 and .7, the biases are notably larger. The biases for the Nunes' ML estimator is in general very large for the Population 2 models. Somewhat surprisingly, the biases are small in the $\rho = 0.5$ case while large in the zero correlation case.

4 Future research

The different developed analysis techniques for on-site sampling reviewed here are specific for their application field. Thus future research should develop a more general approach that can be used across scientific fields.

The model proposed in paper here was chosen so results could be compared with those presented by Nunes (2003). Even though the simulation results in show the proposed estimator to be an interesting alternative, the population models used in the simulation study are not realistic from the economic valuation point of view. Thus the estimator should be evaluated in a variety of population models and the asymptotic properties for these estimators should be further examined. Finally, future studies should examine the behavior of the of the ratio of coefficients because they measure willingness to pay.

References

Bateman, I.J., Carson, R.T., Day, B., Hanemann, M., Hanley, N., Hett, T., Jones-Lee, M., Loomes, G., Mourato, S., Özdemiroglu, E., Pearce, D.W., Sugden, R., and Swanson, J. (2002) *Economic Valuation with Stated Preference Techniques: A Manual*, Edward Elgar, Cheltenham, UK.

Cameron, T. A. and James, M.D. (1987) Efficient Estimation Methods for ”Close-Ended” Contingent Valuation Surveys. *The Review of Economics and Statistics* **69**, 269-276.

Duffield, J.V. and Allan, S. (1988) *Contingent Valuation of Montana Trout Fishing by River and Angler Subgroups.* Montana Department of Fish, Wildlife and Parks, Helena.

Hanemann, W. M. (1984) Welfare Evaluations in Contingent Valuation Experiments with Discrete Responses. *American Journal of Agricultural Economics*, Vol. 66, No. 3 (Aug., 1984), 332-341.

Keillor, B.D., D'Amico, M., Horton, V. (2001) Global Consumer Tendencies, *Psychology and Marketing* **18**, 1-19.

Kim, K. (1995) *A Bivariate Cumulative Probit Regression Model For Ordered Categorical Data.* Statistics in Medicine, Vol. **14**, 1341-1352.

Nunes, L.C. (2003) *Estimating Binary Choice Models With On-Site Samples* Faculdade de Economia, Universidade Nova de Lisboa.

Pollock, K.H., Jones, C.M. and Brown T.L., (1994). Angler Survey Methods and Their Applications in Fisheries Management, *American Fisheries Society Special Publication 25*, American Fisheries Society, Bethseda.

Press, W.H., Flannery, B. P., Teukolsky, S. A., Vetterling, W. T., (1988) *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge University Press, Cambridge.

Santos Silva, J.M.C. (1997) Unobservables in Count Data Models for On-Site Samples. *Economics Letters* **54**, 217-220.

Shaw, D. (1988) On-Site Samples' Regression, Problems of Non-negative Integers, Truncation, and Endogenous Stratification. *Journal of Econometrics* **37**, 211-223.

Sirken, M.G., Levy, P.S. (1974) Multiplicity Estimation of Proportions Based on Ratios of Random Variables, *Journal of American Statistical Association*, Vol. 69, No. 345., 68-73

Sudman, S. (1980) Improving the Quality of Shopping Center Sampling. *Journal of Marketing Research* **17**, 1980, 423-431.