Simple Sequential Procedure for Modeling of Item Non-Response in Econometric Analysis: Application to CV Survey Data

William M. Fonta[1,2], Elias T. Ayuk[1] and H. Eme Ichoku[2]

Item non-response occurs when respondents fail to provide answers to some or all of the questions posed during survey interviews. The standard procedure is to exclude such responses from the econometric analysis. This may be appropriate if the sample included does not differ significantly from those excluded in the analysis. If this is not the case, the econometric analyst faces a sample selection bias problem. The aim of this paper is to provide further evidence using a simple sequential procedure to deal with the problem when using non-randomly selected samples in social science research. The procedure entails different levels of estimation and diagnostic with the Ordinary Least Squares (OLS), Heckman's 2-step and Full Information Maximum Likelihood (FIML) estimators. In the application context, we found the FIML estimator to be more efficient in dealing with sample selection bias than the Heckman's 2-step approach.

Key Words*:* Survey data; Item non-response; Sample selection bias; Sequential procedure; OLS, Heckman 2-step and FIML estimators.

JEL Classification: C13, C25, C42, C51.

1.0 Introduction

One of the most widely encountered problems in sample survey is to record a large number of responses with zero or missing values (i.e., item non-responses). This may be ascribed to a variety of reasons such as free riding, adverse reaction to the interview in general, inadequate comprehension of the intent of the survey question or possibly, the lack of willingness or motivation to disclose the required information (Beatty and Herrmann, 2002; Krosnick, 2002; Strazzera *et al.*, 2003a; Amahia, 2010; Okafor, 2010; Fonta *et al.*, 2010). In most applied social science research work, the standard procedure for handling item non-response problem is to delete such responses from the econometric analysis. However, from a statistical point of view, this may be incorrect if the sub-sample that is excluded is systematically different from that which is included at least in terms of the covariates employed in the econometric analysis. When this is the case, the econometric analyst faces a sample selection bias problem. This could generate inconsistent parameter estimates for reasons similar to those described in Heckman (1976 and 1979), Madalla (1983), Amemiya (1984 and 1985), Vella (1992 and 1998), Melino (1992), Breen (1996), Fonta and Ichoku (2006), Fonta and Omoke (2008), and Fonta *et al.*, (2010). In such circumstance, a sample

[1]United Nations University Institute for Natural Resources in Africa (UNU-INRA), Accra, Ghana (emails: fontawilliam@gmail.com; ayuk@inra.unu.edu);

[2]Centre for Demographic and Allied Research (CDAR), Department of Economics, University of Nigeria, Nsukka, Enugu State, Nigeria (email: hichoku@yahoo.com)

selection model is required to detect, and if necessary, to produce correct estimates for the econometric parameters of the model (Heckman, 1979; Madalla, 1983; Strazzera *et al*., 2003a & 2003b).

The use of the Heckman's 2-step technique to detect and correct for sample selection bias problem, has largely dominated the econometric literature. Although widely used, it has been shown to sometimes perform poorly due to the presence of collinearity problems between the regressors of the 2-step equations (Winship *et al*., 1992 and Strazzera *et al*., 2003a). The main objective of this paper is therefore to illustrate with the aid of a survey data, a simple econometric procedure for simultaneously dealing with the problems of sample selection bias and collinearity when 'item non-responses' are excluded on ad hoc basis in econometrics estimation. The econometrics procedure involves different levels of estimation and diagnostics with the OLS, Heckman's 2-step and the Full Information Maximum Likelihood (FIML) estimators. The duration of the estimation procedure will depend on the diagnostic test results obtained at each stage of the modeling process (Fonta *et al*., 2010).

The rest of the paper is sub-divided as follows: in section II, we present the structural econometric models developed for the empirical estimation followed by the sequential guidelines. In section III, the empirical application is presented using a contingent valuation method (CVM) survey data. Section IV reports the empirical findings while sub-section V concludes the paper.

2.0 Econometric Models and Sequential Procedure

For empirical purposes, let us consider the following two-equation latent dependent variable model given by,

$$u_i^* = w_i'\alpha + \mu_i \tag{1}$$

$$v_i^* = y_i'\beta + \varepsilon_i \tag{2}$$

where w_i and y_i are k and j row vectors of exogenous explanatory variable that are assumed to be determinants of u_i and v_i, and α and β are k and j column vectors of parameters to be estimated for the model. In this simplistic model, it is assumed that,

$$u_i = \begin{cases} 1 & if \quad u_i^* > 0 \\ 0 & \quad u_i^* \le 0 \end{cases} \quad \text{and} \quad v_i = \begin{cases} v_i^* & if \quad u_i = 1 \\ 0 & \quad u_i = 0 \end{cases} \tag{3}$$

In words, we observe u_i a dummy variable, which is the realization of an unobserved (or latent) continuous variable u_i^* with error term μ_i. For values of $u_i = 1$, we observe v_i, which is the realization of a second latent variable v_i^* with

error ε_i. The joint distribution of (μ_i, ε_i) is assumed to be bivariate normal with zero means, variances equal to 1 and correlation ρ.

In Heckman (1979), u_i^* expresses the desire of women to join the work force (i.e., participation equation) and v_i^* measures the observed wages of working women (i.e., outcome equation). Heckman showed that if we estimate the determinants of wages based only on the sub-sample of women working, it could be incorrect if there is bias introduced by self-selection of women into the work force as follows:

$$E[v_i^* | u_i = 1] = y_i'\beta + E(\mu_i | \varepsilon_i > -w_i'\alpha)$$
$$= y_i'\beta + \rho\varepsilon \frac{\phi(w_i'\alpha)}{\Phi(w_i'\alpha)} \quad (4)$$

where the term $\rho\varepsilon \frac{\phi(w_i'\alpha)}{\Phi(w_i'\alpha)}$, is the bias due to self-selection of female participants into the work force. Heckman termed it a simple specification error or omitted variable problem, which is akin to the problem of excluding item non-responses from econometric estimation on *ad hoc* basis (Heckman, 1979). Heckman therefore proposed a consistent 2-step estimator that will allow for the possible correction of the bias, and hence, produce correct estimates of the parameters of the models and the central tendency measures.

The Heckman procedure is carried out in two stages. First, note that the conditional expected value of v_i conditional on $u_i = 1$ and on the vector y_i is given by,

$$E[v_i | u_i = 1, y_i] = y_i'\beta + \rho\sigma_i\lambda_i(w_i'\alpha) \quad (5)$$

where, $\lambda(w_i'\alpha) = \frac{\phi(w_i'\alpha)}{\Phi(w_i'\alpha)}$ is the inverse of the Mills ratio, and ϕ and Φ are the standard normal density and standard normal distribution functions respectively. The first step of Heckman's proposal is to use a Probit model of equation (1) to obtain a consistent estimator of α and then use the estimated α to construct the variable λ (i.e., the inverse mills ratio). In the second step, including λ as a regressor in equation (2), allows us to estimate y and ρ consistently by OLS (Heckman, 1979). A by-product of the 2-step approach is a relatively simple test for identifying the presence of sample selection bias. Under the null hypothesis of no selection bias, (i.e., $\rho = 0$), the usual formula provides a consistent estimate of the covariance matrix of y. Under the alternative hypothesis $\rho \neq 0$, Heckman suggests the use of t-test of the coefficient on the λ variable as a test of sample selection bias (Heckman, 1979).

However, as earlier indicated, a well-known weakness of the 2-step approach is the problem of collinearity between the regressors of the Probit and OLS equations. Based on this, Strazzera *et al.*, (2003a) suggested the use of the following sequential guidelines to simultaneously deal with the problems of selectivity bias and collinearity: (a) First, estimate a two-part model for the separate equations (i.e., participation and outcome) using an OLS estimation technique (i.e., Craig's model), (b) Second, based on the two equations, estimate the models using Heckman's 2-step approach and control for the significance of the coefficient on λ (i.e., the inverse mills ratio), (c) Third, check for the presence of collinearity by regressing λ against the covariates of the outcome equation. If there are no collinearity problems (i.e., judging from the resultant R^2 from the OLS estimation procedure), and the coefficient on λ is not statistically significant, accept the plain OLS estimates obtained at the first stage. If there are no collinearity problems but the coefficient on λ is statistically significantly, accept the 2-step estimates obtained at the second stage. However, if there are some collinearity problems, proceed to the fourth stage as follows, (d) Based on the two equations, estimate a FIML sample selection model, and check for the presence of correlation by observing the significance of the parameter ρ. If ρ is not statistically significant, accept the plain OLS estimates obtained at stage one; otherwise, accept the estimates obtained from the FIML sample selection model.

The log-likelihood to maximize to obtain the FIML estimates is given by,

$$L = \sum_0 \ln(1-\Phi_i) + \sum_1 \frac{\ln 1}{\sqrt{2\pi\sigma^2_{\varepsilon_i}}} - \sum_1 \frac{1}{2\sigma^2_{\varepsilon_i}[v_i - y_i\beta]^2} + \sum_1 \ln \Phi_i \left(\frac{w_i'\alpha + \rho[(v_i - y_i\beta)/\sigma_{\varepsilon_i}]}{(1-\rho^2)^{\frac{1}{2}}} \right). \quad (6)$$

Maximization of this function produces simultaneous estimation of the parameters of both the participation and outcome equations.

3.0 Application to Contingent Valuation (CV) Survey Data

In 2008, the researchers conducted a study of the willingness to pay (WTP) of households to finance one aspect of the new National Health Insurance Scheme (i.e., community-based health insurance - CBHI) in the Nsukka Local Government Area (LGA) of Enugu State, Nigeria, using the contingent valuation method(CVM). The broad objective of the study was to design an improved planning technique that could help elicit information on the value placed by the Nsukka inhabitants on communal financing of the scheme, and decide appropriate household insurance premiums or levies. A key concept in such an improved planning technique is that of the WTP of households in the area to finance the scheme. Eliciting households' WTP, with the aid of the CVM, to inform the design of CBHI schemes is not a novelty in Africa. It has been used by Asenso-

Okyere *et al*., (1997); Asfaw and Braun (2004); Dong *et al*., (2003); Binam *et al.*, (2004); Fonta and Ichoku, (2005); Basaza et al., (2008); Ataguba *et al.,* (2008); Onwujekwe *et al*., (2009 and 2011); Fonta *et al.*, (2010 and 2011), to inform the design and initiation of CBHI schemes in Ghana, Ethiopia, Burkina Faso, Cameroon, Uganda and Nigeria, respectively.

The survey instrument was a pre-tested interviewer-administered structured questionnaire that was divided into two broad categories. The first category elicited information on households' socio-economic and demographic characteristics, health status, assets holding, housing and wealth information and community variables. The second mainly focused on the contingent valuation (CV) scenario under which the evaluation of the proposed CBHI scheme took place. This scenario detailed the nature of the new CBHI initiative being proposed in Nsukka, the current health service delivery situation in Nsukka, the institutional setting in which the proposed scheme will be provided, and how each household will have to pay to finance the scheme (i.e., quarterly contributions). The value elicitation formats used was the Dichotomous Choice (DC) format buttressed with open-ended follow-up and debriefing questions. Our choice of using the DC elicitation format is because of its incentive-compatibility feature compared to other formats (Mitchell and Carson, 1989). Five starting prices were used in the DC question as follows: N200, N400, N600, N800 and N1000. These bids were based on an earlier pilot study in the community. These prices were assigned randomly and roughly proportionately to the number of households in the study sample.

A two-stage selection procedure was adopted for the study design. The first stage was a random selection of five communities out of the 15 communities in Nsukka namely; Obukpa, Edem, Nsukka, Ibagwa-Ani and Ehalumona. From these five communities, the Federal Office of Statistics (FOS) now National Bureau of Statistics (NBS) enumeration-listing booklet was used to select four Enumeration Areas (EAs) from each of the five communities. In the second stage, a simple systematic random sampling technique was used to select 19 households from each of the EAs. This gave a total sample size of 380 households[3]. The sampled households were appropriately weighted during analysis. Under the weighting, each household selected from each EA was weighted to make it representative of the entire EA such that the sum of the weights for each EA equaled the approximate number of households in that EA.

During the CV interview, if a respondent said yes to the initial WTP bid proposed to him/her, a follow-up question was asked to elicit his/her maximum WTP

[3] This optimal size was obtained using the Taro Yamane (1967) specification. That is, $n = N / (1 + N(e)^2)$ where n equals to the sample size to be estimated, N stands for the population size (i.e., household size), and e represents the margin of error.

amount to finance the scheme. However, if the answer was no, another follow-up question was asked, to find out the respondent's actual WTP amount if different from that of the proposed bid. If no WTP amount was reported at this stage, a debriefing question was posed to the respondent to find out the reason(s) for not being willing to pay to finance the scheme. This was basically to distinguish 'item non-responses' or invalid responses from the valid responses. Overall, out of a total of 380 households randomly selected for interview, 235 (61.8%) provided valid responses to the valuation question, 74 respondents (19.5%) provided invalid responses (i.e., item non-responses)[4] to the valuation question, while about 71 households refused outright to be interviewed.

4.0 Empirical Results

4.1 Sample Statistics

Table 1 presents the summary statistics describing the sampled population. On average there are 6 members in a household living in an average of four rooms. Over 95% of thes households have bathrooms while only about 46% reported having toilet facilities. Also, most of the household heads interviewed (99%) are either employed in the formal sector by the Local Government Authority (though, mainly menial labourers and clerks) or the informal sector as craftsmen, petty-traders and farmers. Equally, most of the respondents were engaged in farming, which may not necessarily be as a full time occupation. This limited the direct observation of household income. Based on the pilot testing, a proxy measure of wealth was adopted as also suggested by Fonta (2006). Thus, the average income for the sample was calculated to be about NGN121,714.20 (US$936.3)[5] per annum or NGN10,142.85 (US$78) per month. By gender distribution, about 63% of the sampled respondents were male while only about 37% were females. In terms of age distribution, the average for the sample was about 51 years. The average distance from a household to the nearest health centre was estimated to be about 3.3km.

[4] The main reason for such invalid responses was because of 'protests' zeros and outliers. 'Protests' zeros, according to Freeman (1993:187), occur when respondents reject some aspect of the constructed CV market scenario by reporting a zero value even though they place a positive value on the amenity or resource being valued. On the other hand, outliers are determined by the researcher based on some measures such as the share of WTP in income or what Mitchell and Carson (1989: 226 –227) called α -trimmed mean where the analyst chooses the value of α.

[5] At the time of the survey, USD1 was approximately ₦130.

Table 1: Descriptive Statistics for the Sampled Households

Variable	Measurement/Definition	Mean/ Proportions	Std. Dev.
Age	Age of the respondent at the last birthday (in years)	51.69	12.56
Bathroom	1 if own a bathroom and, 0 otherwise	0.96	0.19
Bid	Starting prices in Naira	598.71	283.3
Borrowed_amount*	Amount borrowed for treatment in the last four weeks prior to survey	666.36	3,251
Distance	Km to nearest health centre	3.33	2.09
Dwelling	1 if building is constructed with cement/concrete and, 0 otherwise	0.85	0.36
Educ	Education attainment of household head and 1 if above primary school and 0, otherwise	0.89	0.95
Employed	1if employed and, 0 otherwise	0.89	0.11
Floor_material	Nature of floor material and 1if cement/tiles/concrete and 0, otherwise	0.82	0.39
HHnumber	No. of adults and children being fed	6.1	3.09
Hstate	Respondent's health status at time of the interview and 1 if good and, 0 otherwise	0.67	0.79
Know_insurance	1 if knowledgeable about health insurance, 0 otherwise	0.11	0.31
Male	1 if male and, 0 otherwise	0.63	0.48
Meanstreat	Means of seeking treatment during illness. 1 if orthodox means and, 0 otherwise	0.55	0.5
Numrooms	Number of sleeping room	4.13	1.61
Participation	1 if participated/participating in any health insurance scheme and, 0 otherwise	0.03	0.18
Qhcentre	Rating of the quality of the health centers. 1if judged as being good and, 0 otherwise	0.68	0.75
Sick	1 if sick two weeks prior to survey and, 0 otherwise	0.40	0.49
Toilet	1if own toilet, 0 otherwise	0.46	0.5
Treatamount*	Direct + indirect cost incurred in treatment a household member in the last four weeks prior to survey (Naira)	763.35	2,612
Trust	1 if confidence in trust fund and, 0 otherwise	0.78	0.82
Wealth_measure*	Assets and other household durables (in Naira)	121,714	114,741

Furthermore, about 40% of the respondents reported that a household member fell sick within the last two weeks prior to the survey. In terms of the cost of treatment, on the average, the rural households spend about ₦763 ($5.87) within four weeks. Equally, the amount borrowed for treatment including money realized from the sale of valuable assets and property was estimated to be about ₦666 (US$5.1). This is equivalent to over 87% of the amount spent on treatment across all respondents. In terms of health insurance knowledge, only about 11% of the sample were knowledgeable about what health insurance is and only 3% reported having ever participated in any form of insurance (not necessarily health related) in the past or at present.

Additionally, the literacy level of the respondents was quite low as over 77% of the respondents have not had more than 7 years of formal education. Conversely, about 78% of respondents expressed confidence in the proposed community trust fund where funds are to be pooled together and managed by the community. This gives a high indication of credibility for establishing such a scheme. Further still, more than half (60.2%) of the sampled household heads reported their health status as being better than 'Good' at the time of interview. In terms of household health seeking behavior, about (55%) of the sample reported seeking health care services from orthodox[6] health care providers while about 45% reported patronizing unorthodox health care providers. Finally, more than half (59%) of the respondents adjudged the quality of the health care centers nearest to them as being better than 'Good'.

4.2 Sample Selection Results

Having so far discussed the characteristics of the sample, we now turn to the econometric analysis. First, it is necessary to distinguish between responses that can be considered valid (i.e., $WTP \succ 0$) and those that appear invalid (i.e., $WTP \prec 0$). Of a total of 309 interviews that were actually completed, 74 respondents (19.5%) were considered to have invalid responses to the valuation question. As earlier indicated, the main reasons for such invalid responses were because of protests respondents (30) and outliers (44). It was therefore necessary in the analysis to determine whether excluding those with invalid responses from the econometric analysis would lead to a sample selection bias problem. As noted in Strazzera *et al.*, (2003a), Fonta and Ichoku (2006), and Fonta *et al.*, (2010), a preliminary test for the presence of sample selection bias is to compare the means of household covariates between the two groups (i.e., 'valid' and 'invalid' responses) using sample mean comparison test. Any significant difference between the two groups of responses is an early warning indicator of the presence

[6] Orthodox providers are categorized as clinics, maternity centres, dispensary, and hospitals. The unorthodox providers are categorized as patent medicine stores, traditional healers and herbalists, etc.

of sample selection bias and justifies the use of a sample selection model. For some of the variables (e.g. gender of respondent, the floor type, household size, respondent's health status, means of seeking treatment, nature of dwelling unit, confidence in trust fund and distance to health centres), the difference between the two groups (i.e. 'valid' and 'invalid' responses) are quite significant at 1 and 5% levels, respectively (Table 2). If these variables influence the respondents' WTP for the new social health insurance scheme in Nsukka, then we expect the final estimates obtained from the sub-sample of households with valid responses to be affected by selectivity bias.

Table 2: Comparison of Means by Groups of Respondents

Variable Name	**Valid WTP Responses**		**Invalid WTP Responses**		**Comparison** $(\mu_1 - \mu_0)$
	Mean(μ_1)	**Std. Dev.**	**Mean**(μ_0)	**Std. Dev.**	**t-stat.**
Male	0.70	0.46	0.42	0.5	4.54***
Floormaterial	0.85	0.36	0.72	0.45	2.65***
Numrooms	4.31	1.61	3.55	1.5	3.59***
Hstate	2.71	0.76	2.53	0.88	1.78*
Meanstreat	0.59	0.49	0.42	0.5	2.62***
Dwelling	1.14	0.43	1.24	0.43	-1.72*
Trust	3.13	0.82	2.95	0.79	1.72*
Distance	3.54	2.17	2.66	1.67	3.21***
Obs.		**235**		**74**	

*, **, *** Showing significance of parameter estimates at 10%, 5% and 1% levels respectively.

Tables 3 and 4 report the results of the econometric estimations of equations (1) and (2)[7] using different covariate specifications (i.e., reduced form models) related to the effects of households socio-economic characteristics listed in table 1. However, note that the tables report the parameter estimates for the best-fit specifications (most valid reduced form models) from the two equations (i.e., participation and outcome) selected by means of likelihood ratio tests.

Starting first with the Probit results (Table 3), to explain included versus excluded households in the participation equation (i.e., Probit estimation), the gender of the respondent seems to have an effect on the probability to participate or not to finance the scheme. In particular, being a male-headed household increases the probability to participate in financing the scheme. This could be linked to the

[7] Note that in our empirical context, equation (1) expresses the desire of households to participate in financing the scheme, while equation (2) measures the observed WTP amounts of households. w and y are the exogenous explanatory variable listed in Table 1, which are the determinants of u_i and v_i respectively.

roles of men in the community who have traditionally been charged with the responsibility of catering for the family financially. Similarly, falling sick two weeks prior to the survey increased the probability to participate. This may perhaps be because implementing the scheme in the area is expected to improve health care delivery services and hence, household health status.

Table 3: 2-steps (No Selection) and FIML Estimates (No Selection)

	Participation Equation			
Parameter	**Probit Estimates**		**FIML Estimates**	
	Estimates	**Std. Err.**	**Estimates**	**Std. Err.**
Constant	2.28	1.393*	2.28	1.351*
Male	0.821	0.185***	0.784	0.183***
Sick	1.183	0.704*	1.178	0.690*
Floor_material	0.485	0.245**	0.573	0.237**
Ln_Distance	-0.408	-0.138***	-0.432	-0.134**
Ln_wealthmeasure	0.278	0.106***	0.281	0.105***
Ln_Bid	-1.016	0.189***	-1.021	0.184***
% correctly predicted		94.49%		94.50%
Observation		**309**		**309**

*, **, *** Showing significance of parameter estimates at 10%, 5% and 1% levels respectively.

Equally, household income also had an effect on the probability to participation in financing the scheme and those with higher income had higher participation rates: possibly because, a higher-income earner apparently has a greater demand for better health care facilities than a lower-income earner. Finally, households that were farther away from the existing health care facilities in the community had a higher participation rate than those closest. Possibly because the farther away a household is from the nearest health center, the higher the cost of transportation and frequency of visits is lower. This may explain why such households are more willing to pay to finance the scheme than those living closer to existing healthcare facilities.

In Table 4 (i.e., the outcome model) where the observed WTP amounts of households' is the dependent variable, richer household heads were willing to pay higher amounts than poorer household heads (presumably for the same reason that they are also more willing to participate to finance scheme). Another important determinant of households WTP for the scheme is household knowledge about health insurance and, the more knowledgeable a household head is, the higher the stated amount for the scheme. Similarly, the health status of a household head was a significant determinant of the WTP amount. Heads of households with better health status were willing to pay less than those with poor health status. Education

was equally a significant determinant of household WTP for the scheme and the higher the educational attainment, the higher the stated amount for the scheme.

Table 4: OLS (No Selection), 2-step (Selection) and FIML (Selection)

	Outcome Equation					
Parameter (1)	**OLS (No selection)**		**2-Step Estimates**		**FIML Estimates**	
	Est. (2)	**S. Err. (3)**	**Est. (4)**	**S. Err. (5)**	**Est. (6)**	**S. Err. (7)**
Constant	2.614	0.904***	2.212	0.642***	2.277	0.640***
Age	-0.005	0.004	-0.007	0.003*	-0.007	0.003**
Knowinsurance	0.335	0.153**	0.386	0.122***	0.381	0.122***
Hstate	-0.084	0.068	-0.112	0.058*	-0.116	0.058**
Floor	-0.300	0.135**	-0.255	0.113**	-0.238	0.113**
Toilet	0.271	0.100***	0.375	0.082***	0.363	0.082***
Ln_Wealth measure	0.152	0.056***	0.137	0.048***	0.139	0.048***
Ln_Bid	0.392	0.078***	0.437	0.070***	0.43	0.069***
Mills lambda (λ)			0.357	0.146**		
Rho (ρ)					0.470	0.177***
Sigma (σ)				0.621		0.611
Adjusted R^2		0.26		0.15		
Observation		309		235		235
Log-Likelihood	-337.674					

*, **, *** Showing levels of significance of parameter estimates at 10%, 5% and 1% respectively.

Also, male headed households are more willing to pay higher amounts than their female counterparts, which might be as a result of cultural reasons where the males are responsible for most financial decisions within the household. Other important determinants of household WTP for the scheme includes; the household size, distance to health care facilities, and the number of rooms in a household.

4.3 Implications of Sequential Procedure on Mean WTP Estimates

Having analysed the determinants of households WTP for the scheme in the light of sample selection bias, we now turn our attention to the empirics of the different estimation techniques (i.e., the OLS, 2-step and FIML estimators). Columns 2 and 3 of table 4 report the parameter estimates obtained from the plain OLS estimation technique at stage (a) of the modelling process. As observed, the parameter estimates are slightly higher than those obtained using Heckman's 2-step approach and the FIML estimator. This is partly as a result of including all the observation (i.e., 'valid' and 'invalid') in the estimation procedure without correcting for sample selection bias. Since there is no way to judge *a priori* from

the OLS estimates any evidence of sample selection bias, wrong conclusions can be deduced that excluding 'invalid responses' from the analysis may have little or no effects on the final WTP estimates obtained from only the sub-sample of households with valid responses. However, when we considered the 2-step estimates obtained at stage (b) of the sequential procedure when sample selection bias correction took place (i.e., columns 4 and 5 of same table), the interpretation becomes slightly different. The standard errors for the coefficient estimates showed higher levels of significance with also more significant parameter estimates. However, besides this information, the result gives us no additional clue about the degree of correlation between the regressors of the participation and outcome equations: a well-known weakness of the method and a critical assumption of econometric estimation in general (Fonta *et al.*, 2010).

To therefore check for the presence of collinearity in the 2-step estimates obtained at stage (b), we ran an OLS regression of mills lambda (i.e., λ) against the covariates of the outcome equation as suggested in the sequential guidelines[8]. The resulting $R^2 = 0.51$ from the estimation procedure indicates a moderate level of correlation. Since the 2-step estimates suffer from collinearity problems, we proceeded to stage (d) by estimating a FIML sample selection model. The regression results are reported in columns 6 and 7 of Table 4. As expected, ρ is statistically significant indicating a high level of correlation between the regressors of the two equations. However, note that if the coefficient on ρ was not statistically significant, the plain OLS estimates obtained at stage (a) would have been preferred to the FIML estimates obtained at stage (d). Equally, if the 2-step estimates obtained at stage (2) were somehow free from collinearity problems, the results would have been as efficient as those obtained with the FIML estimates at stage (d).

Table 5: Descriptive Stats of Quarterly Mean WTP Estimates for the Scheme

Modeling Method	Obs.	Mean	95% Conf. Int.
All Respondents (OLS)	309	392.20($3.0)*	337.0 – 447.5 ($2.6 – 3.4)
OLS (Selection)	235	509.94($3.9)	485.1 – 534.8 ($3.7 – 4.1)
Heckman's Model	235	458.67($3.5)	434.3 – 483.0 ($3.3 – 3.7)
FIML Estimator	235	466.68 ($3.6)	442.2 – 491.2 ($3.4 – 3.8)

* The figures in parenthesis represent the US Dollars equivalence

Although the parameter estimates obtained using the Heckman's 2-step estimator are not much different from that of the FIML estimator, differences normally occur when calculating the final mean WTP estimates for project or policy

[8] The regression results are however not reported here but the procedure for doing this could be obtained from the authors on request.

purposes. Table 5 reports the calculated mean WTP estimates using the three different estimators. The first row of table 5 reports the mean WTP estimates calculated for all the respondents (i.e., 'valid' and 'invalid' responses) based on the plain OLS estimation at stage (a). As shown, for all the respondents, the mean quarterly WTP estimate for scheme is about NGN392.20 ($3.0) with associated confidence intervals of 337.0 – 447.5 ($2.6 – 3.4). The second row reports the estimates calculated for only the sub-sample of respondents with valid responses without correcting for sample selection bias. As observed, the estimates are quite high when compared to the other mean WTP estimates obtained from the different estimation methods. It is biased upwards as equally suggested by the positive sign of ρ. Note that if ρ had been negatively signed; the WTP estimates obtained from plain OLS without sample selection bias correction, would have been biased downwards. The third and fourth rows report WTP estimates calculated using Heckman's 2-step approach and the FIML sample selection model when sample selection bias correction took place. As further observed, the two estimates are slightly different although the parameter estimates of the two estimators are not much different from each other. The same can be said about their confidence interval estimates; those of the FIML estimators are slightly higher than those of the 2-step approach. This obviously suggests that the choice of the estimation technique in econometric analysis of survey data can significantly affect the final parameter estimates obtained from a given sample for welfare estimates and policy conclusions. This is, if care is not taken to address peculiar sample survey problems that might arise from modeling of item non-response such as selectivity bias and the problem of collinearity.

5.0 Conclusion

This paper had several motivations. Firstly, the study was motivated by the need to highlight the importance of choosing appropriate econometric techniques when using non-randomly selected samples to estimate behavioral relationships in applied social sciences research works. Secondly, it was equally motivated by the need to design an improved planning methodology that could help elicit information on the value placed by rural households in Nigeria to finance one aspect of the new National Health Insurance Scheme (i.e., the community-based social health insurance scheme).

In the application context, some important methodological and policy findings we equally arrived at towards the study objectives. Firstly, the study found out that when item non-responses are excluded from econometric estimation on *ad hoc* basis, the social science researchers may encounter a sample selection bias problem, which may have two consequences namely; (a) the empirical analysis may generate inconsistent parameter estimates for reasons similar to those described in Heckman (1976 & 1979), and the final estimate obtained for policy purposes from the included sub-sample is likely to be biased. Secondly, the study

also revealed that in the absence of any collinearity problems between the regressors of a two-equation latent dependent variable models, the Heckman's 2-step estimator would produce parameters estimates that are equally as efficient as the FIML sample selection estimator. Thirdly, the results further revealed that the CV survey device can be successfully used to support the design and implementation of CBHIS in rural Nigeria.

Acknowledgements

This paper was prepared while the first author was a visiting senior research fellow at UNU-INRA, under a fellowship grant from the Centre of Environmental Economics and Policy in Africa (CEEPA) at the University of Pretoria. We acknowledged the helpful comments from two anonymous reviewers for the journal. All other declaimer applies.

References

Asenso-Okyere,W.K., Osei-Akoto I., Anum, A. and Appiah, E.N. (1997). Willingness to Pay for Health Insurance in a Developing Economy: A Pilot Study of the Informal Sector of Ghana using Contingent Valuation. *Health Policy,* 42: 223-237.

Asfaw, A. and Braun, J. (2004). Can Community Health Insurance Schemes Shield the Poor against the Downside Health Effects of Economic Reforms? The Case of Rural Ethiopia. *Health Policy,* 70: 97-108.

Amahia, G.N. (2010). Factors, Preventions and Correction Methods for Non-Response in Sample Survey. *CBN Journal of Applied Statistics,* 1(1): 79-89.

Amemiya, T. (1984). Tobit Models: A Survey. *Journal of Econometrics*, **24**:3-61.

Amemiya, T. (1985). *Advanced Econometrics*. Oxford, Basic Blackwell.

Ataguba, A. J., Ichoku, E. H. and Fonta, M.W. (2008). Estimating the Willingness to Pay for Community Healthcare Insurance in Rural Nigeria. PMMA Working Paper No. 2008-10, University of Laval, Canada.

Beatty and Herrmann, (2002). To Answer or Not to Answer: Decision Processes Related to Survey Item Non-response. Eds. R. Groves et al., John Wiley & Sons, New York.

Basaza, R., Criel, B., and Vander-Stuyft, P. (2008). Community Health Insurance in Uganda: Why Does Enrolment Remain Low? A View from Beneath. *Health Policy,* 87: 172-184.

Binam, J., Nkama A. and Nkenda, R. (2004). Estimating the Willingness to Pay for Community Health Prepayment Schemes in Rural Area: A Case Study of the use of Contingent Valuation Surveys in Centre Cameroon. Retrieved, August 12, 2005, *from: http:// www.csae.ox.ac.uk/conferences/2004-GPRaHDiA /papers/4h-Binam-CSAE2004.pdf*.

Breen, R. (1996). *Regression Models, Censored, Sample-Selected or Truncated Data*. Sage Publications, Thousand Oaks, London.

Dong, H., Kouyate, B., Cairns, J., Mugisha, F. and Sauerborn, R. (2003). Willingness-to-Pay for Community-based Insurance in Burkina Faso. *Health Economics*, **12**: 849-862.

Fonta, M.W. and Ichoku, E.H. (2005). The Application of the Contingent Valuation Method to Community-led Financing Schemes: Evidence from Rural Cameroon. *Journal of Developing Areas,* 39 (1): 106-126.

Fonta, M.W. and Ichoku, E.H. (2006). Evaluating the Statistical Efficiency of the OLS, Heckman's 2-step and FIML Estimators in Addressing Sample Selection Bias in Social Science Research. Paper presented at the 11th Annual African Econometric Society Conference, July 5-7, 2006, Dakar, Senegal.

Fonta, M. W. and Omoke, P.C. (2008). Testing and Correcting for Sample Selection Bias in Social Science Research: Application to Contingent Valuation Method (CVM) Survey Data. *European Journal of Social Sciences, 6(2):* 232-243

Fonta, M. W., Ichoku, E. H., Ogujiuba, K.K. and Chukwu, J. (2008). Using a Contingent Valuation Approach for Improved Solid Waste Management Facility: Evidence from Enugu State, Nigeria. *Journal of African Economies,* 17(2):277-304.

Fonta, M.W., Ichoku, E. H. and J. K-Mariara, (2010). The Effect of Protest Zeros on Estimates of Willingness to Pay in Healthcare Contingent Valuation Analysis. *Applied Health Economics and Health Policy,* 8(4):225-237.

Fonta, M.W., Ichoku, E.H. and Nwosu, E.O. (2011). Contingent Valuation in Community-based Project Planning: The Case of Lake Bamendjim fishery

Re-stocking in Cameroon. *AERC Research Paper No. 210, African Economic Research Consortium, Nairobi, Kenya.*

Freeman, A.M. (1993). *The Measurement of Environmental and Resource Values: Theory and Methods.* Resources for the Future, Washington D.C.

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**: 475-492.

Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, **47**:153-162.

Krosnick, J. (2002). The Causes of No Opinion Responses to Attitude Measures in Surveys. *Survey Non-Response*, 87-100, Wiley, N.Y.

Leung, S. F. and Yu, S. (1996). On the Choice between Sample Selection Models and Two-part Models. *Journal of Econometrics*, 72: 197-229.

Leung, S. F. and Yu, S. (2000). Collinearity and 2-step Estimation of Sample Selection Models: Problems, Origins, and Remedies. *Computational Economics*, 15: 173-99.

Madalla, G. S. (1983). *Limited-Dependent and Quantitative Variables in Econometric*. Econometric Society Monographs in Quantitative Economics, Cambridge University Press, Cambridge.

Melino, A. (1982). Testing for Sample Selection Bias. *The Review of Economic Studies*, **9**(1): 151-153.

Mitchell R.C. and Carson, R. (1989). *Using Surveys to Value Public Goods: The Contingent Valuation Method*. Resources for the Future, Washington D.C.

Okafor, F.C. (2010). Addressing the Problem of Non-Response and Response Bias. *CBN Journal of Applied Statistics*, 1(1): 91-97.

Onwujekwe. O., Onoka. C. and Uzochukwu. B. (2009). Community-based Health Insurance as Equitable Strategy for Paying for Healthcare? Experiences from Southeast Nigeria. *Health Policy*, 92: 96-102.

Onwujekwe, O., Uzochukwu, B. and Kirigia, J. (2011). Basis for Effective Community-based Health Insurance Schemes: Investigating Inequities in Catastrophic Out-of-Pocket Health Expenditures, Affordability and Altruism. *African Journal of Health Economics*, 0001:1-11.

Strazzera, E., Scarpa, R., Calia, P., Garrod, D.G. and Willis, G. K. (2003a). Modeling Zero Values and Protest Responses in Contingent Valuation Surveys. *Applied Economics*, 35:133-138.

Strazzera, E., Genius, M., Scarpa, R. and Hutchinson, G. (2003b). The Effect of Protest Votes on the Estimates of WTP for Use Values of Recreational Sites. *Environmental and Resource Economics*, 25:461-476.

Taro, Y. (1967). *Elementary Sampling Theory*. Prentice-Hall, Network, USA.

Vella, F. (1992). Simple Test for Sample Selection Bias in Censored and Discrete Choice Models. *Journal of Applied Econometrics*, **7**(4), 413 - 421.

Vella, F. (1998). Estimating Models with Sample Selection Bias: A Survey. *Journal of Human Resources*, **33**(1): 127 -169.

Winship, C. and Mare, D.R. (1992). Models for Sample Selection Bias. *Annual Review of Sociology*, **18**:327-350.