Estimation of a Selectivity Model with Misclassified Selection

Vidhura S. B. W. Tennekoon

Department of Economics

IUPUI

Cavanaugh Hall 524

425 University Boulevard

Indianapolis, IN 46202

vtenneko@iupui.edu

Steven B. Caudill

Department of Economics

Florida Atlantic University

Kaye Hall, Room 141A

777 Glades Road

Boca Raton, FL: 33431

scaudill@fau.edu

Abstract

Despite the great interest in models of self-selection and models with misclassification, there have been few studies combining the two. Notable exceptions are given by McCarthy, Millimet, and Roy (2015) and Shiu (2016). None of these models have been developed in a contingent valuation setting that we are interested in. The goal of this note is to add to this literature by presenting a model for estimating willingness to pay using data collected through a contingent valuation survey. We examine the case of a selectivity model in which the *outcome* equation is interval censored but the *decision* indicator is not observed.

Keywords: contingent valuation, selectivity model, misclassification

JEL codes: C34, Q51

This is the author's manuscript of the article published in final edited form as:

Tennekoon, V. S. B. W., & Caudill, S. B. (2022). Estimation of a selectivity model with misclassified selection. Communications in Statistics: Case Studies, Data Analysis and Applications, 8(1), 1–14. https://doi.org/10.1080/23737484.2020.1827465

Estimation of a Selectivity Model with Misclassified Selection

1. Introduction

Since the seminal work of Heckman (1979) self-selection models have gained wide use and there have been many extensions. The basic selection model consists of a *decision* equation and an *outcome* equation. Selectivity models are commonly used in the contingent valuation literature.[1] Contingent valuation is a method for estimating values for goods for which markets do not exist. These studies are often used to estimate the value to society of preserving or improving some natural resource or estimating values for public goods. In the usual contingent valuation study, respondents provide information on personal and family characteristics and then answer a series of questions to elicit information about the amount they are willing to pay (WTP) to preserve or improve some non-market good. These WTP values are aggregated across respondents to estimate society's WTP. A selectivity model is a natural solution to the estimation problem presented by contingent valuation data. The *decision* equation models whether or not a respondent has positive WTP. For those responding affirmatively to this question, the *outcome* equation is used to model the amount a respondent is willing to pay. In the two-equation system; 1) model parameters can be estimated for the *decision* equation and the *outcome* equation, 2) the set of exogenous variables used in each model can differ, and 3) the correlation between the error terms in the two equations can be estimated. Selectivity models have been extended in many directions from this basic framework.[2]

---

[1] For examples of the use of selectivity models in contingent valuation studies see Whitehead (1991), Harpman, Welsh, and Sparling (2004), Brouwer, Beukering, and Sultanian (2008), Bellemare (2012), Soderberg and Barton (2014), Bonnichsen, Ole, and Soren (2016), and Skeie, Lindhjem, Skjeflo, and Navrud (2019).

[2] Extensions of the *decision* equation include polychotomous decisions by Trost and Lee (1984) and Malikov, Kumbhakar, and Sun (2016), sequential decisions by Caudill and Oswald (1993), an ordered decision function by Joyce (1994), a multinomial decision function by Schmertmann (1994), nonparametric selection by Ahn and Powell (1993), and ordered probit selection by Main and Reilly (1993). The form of the *outcome* equation has also been the subject of several extensions including a spatial autoregressive model by Hsieh and Lee (2016), semiparametric estimation of the equation by Malikov, Kumbhakar, and Sun

A separate strand of the literature examines misclassification. Usually these studies examine the consequences of using misclassified dummy dependent variables.[3] Despite the great interest in both models of self-selection and models with misclassification, there have been few studies combining the two. Notable exceptions are given by McCarthy, Millimet, and Roy (2015) and Shiu (2016). None of these models have been developed in a contingent valuation setting, the focus of our work.

Our goal is to add to this literature. We examine the case of a selectivity model in which the *outcome* equation is interval censored and the *decision* indicator is only partially observed, and, in the selectivity framework, can be considered misclassified. More specifically, we know the interval in which the reported WTP lies, but the true underlying distribution of WTP is a nonnegative mixture, censored from below at zero. The zero responses, essential for estimating the decision equation, are not observed; they are contained within the observed lowest interval. In addition, the true WTP may not be revealed by all respondents. We assume those not revealing their true WTP are also contained in the lowest interval causing a selection problem.

We use our model to estimate a selectivity model based on the rather peculiar contingent valuation dataset for the Kakadu nature park in Australia, previously examined by Carson, Wilks, and Imber (1994) and Werner (1999). Our extension is necessary in order to examine WTP in a selectivity model. Doing so provides some advantages over other methods. We can recover parameter estimates from *both* the *decision* and *outcome* equations under a joint bivariate

(2016), Poisson regression by Winkelmann (1998) and van Ophem (2000), nested logit by Falaris (1987), stochastic frontiers by Bravo-Ureta, Greene, and Solis (2012), and simultaneous Tobit models by Choe and Jeong (1993).

[3] By misclassified dummy dependent variables, we mean some of the ones are incorrectly coded as zero and vice versa. Studies examining the consequences include Meyer and Mittag (2017), Hyslop and Townsend (2017), Lu, Luo, and Xiao (2014), Fu, Gao, and Shi (2011), Hausman, Abrevaya, and Scott-Morton (1998), and Caudill, Ayuso, and Guillen (2005). Dustman and van Soest (2004) examine ordered response models with misclassification.

normal assumption. These equations may contain *different* sets of explanatory variables, and we are able to estimate the *correlation* between the error terms in the decision and outcome equations. As a full understanding of the Kakadu data set is critical to our analysis, we provide a detailed discussion in the next section.

2. The Kakadu Data

In 1999 a contingent valuation study was conducted to estimate willingness to pay to preserve the Kakadu nature park in Australia. The study was unusual in that there was no independent information available through separate questions for the usual first stage *decision* equation. That is, at no time were respondents asked whether or not they were willing to pay *any* positive amount for Kakadu preservation.

The data in the questionnaire is collected by asking each respondent two questions about willingness to pay: a first question and then a follow-up question based on the response to the first question. For example, each individual is first asked whether they are willing to pay some amount, t_1, for Kakadu preservation. If the respondent answers *No*, the respondent is asked whether they are willing to pay a lower amount, t_{min}. If the respondent answers *Yes* to the first question, the respondent is asked whether they are willing to pay a higher amount, t_{max}. For the first question the value of t_1 changed from respondent to respondent, being randomly chosen from a set consisting of {5, 20, 50, 100}. The response to the follow-up question yielded the value of either t_{min} or t_{max}. Based on the answers to this pair of questions, an individual's willingness to pay can be converted into interval membership as indicated in Table 1.

As Table 1 indicates, we discuss the interval limits in general terms, [t_1, $t_{max,}$ $t_{min]}$ because the interval limits changed across respondents. The following values for [t_1, t_{max}, t_{min}] were used in the Kakadu design: [5, 20, 2], [20, 50, 5], [50, 100, 20] and [100, 250, 50]. Each triple is used to construct four response intervals, one of which contained the WTP for the i^{th} respondent. For example, the triple [5, 20, 2] results in WTP intervals: (0-2), (2-5), (5-20), and (20, +∞). Thus, each willingness-to-pay response consists of a lower bound (LB) and an upper bound (UB) on the respondent's willingness to pay. Information about respondent characteristics were also collected.

The pairs of questions described lead to three different types of intervals. The problematic first interval in this design, ($0 - t_{min}$), is bounded below by zero and consequently includes zero responders and those with small positive willingness to pay. Thus, the response to the *decision* question is hiding in the lowest interval. Worse yet, in contingent valuation studies there are two kinds of zero responders. There are those with zero WTP who do not wish to contribute but might like to free-ride on the contributions of others or have a negative WTP. There are also protest responders who do not reveal their true WTP, but rather report zero. Separating the two is important for obtaining good estimates of aggregate willingness to pay.

The remaining intervals present no particular econometric difficulties. The middle two intervals, ($t_{min} – t_1$) and ($t_1 – t_{max}$), are bounded by positive values above and below. The topmost interval is open ended as t_{max} is assumed to equal positive infinity. Estimation based on these types of intervals is routine.

3. Previous Econometric Solutions

With the main estimation problem arising due to the responses in the lowest interval, there are at least four possibilities for statistical analysis depending on the amount of information obtained on the treatment of the zero responders. Carson, Wilks, and Imber (1994), using the

Kakadu data, ignore the preponderance of zero responders in the lowest interval and treat the lowest interval like any other continuous interval. A consequence of this treatment is that the probability of a *zero* response in a continuous (0 - t_{min}) interval is zero. This is likely not the case with the Kakadu response data. A second approach ignores the difference in the two types of zero responders, considering them as a single group, and then directly addresses the inflated number of zero responses (see Kristom (1997), Del Saz-Salazar and Garcia-Mendez (2001), and Garcia and Rivera (2003)). A third approach attempts to separate the zero responses *after the fact* into *actual* zero and *protest* zero responses. This approach is used by Strazzera, Scarpa, Calia, Garrod, and Willis (2003) who extend a selectivity model for WTP by incorporating a mixture model to separate actual zero responders and protest zero responders. A related approach by Werner (1999) and Fernandez, Leon, Steel, and Vazquez-Polo (2004) analyses contingent valuation data using mixture models but not within a selectivity framework.

Our objective is to provide an alternative to the mixture approaches of Werner (1999) and Fernandez, Leon, Steel, and Vazquez-Polo (2004) based on self-selection. Our approach has several advantages over the approaches mentioned above. In particular, unlike the approaches of Werner (1999) and Fernandez, Leon, Steel, and Vazquez-Polo (2004), we do not assume independence of the "decision" and "outcome" equations. Unlike Strazzera, Scarpa, Calia, Garrod, and Willis (2003), we are able to estimate a selectivity model without observing the outcome in the "decision" equation.

4. Our Model

As an alternative to the logistic-generalized gamma mixture of Werner and the other special cases, we propose a bivariate normal distribution. Our approach is thus more like the selectivity models normally used to analyze WTP data. The main advantage of our approach over Werner's is that we are able to capture the dependence between the two decision processes: the *decision* equation and the *outcome* equation. In addition, our model recognizes two types of responders; protest responders and others who, justifiably, are not willing to pay anything.

There are firm economic foundations underlying our modeling approach to WTP based on separating *No* responders into two groups. An individual derives benefits and has a positive WTP for the preservation of a natural reserve for two reasons; either the utility gained by usage or potential usage of the reserve, or the utility due solely to the existence of the reserve. These components are called *use* value and *existence* value. One group of *No* responders is called protest respondents who do not state their true WTP in a contingent evaluation survey for whatever reason (Meyerhoff and Liebe, 2008). Protest respondents comprise about 20% of respondents across many studies (Frey and Pirscher, 2019). A second group of *No* responders are individuals who never anticipate using the reserve and do not care about its existence. For this group WTP is zero simply because preserving the reserve does not increase their utility. This group may also include people who do care about the fate of the reserve but are not necessarily willing to make any positive monetary contribution for preserving it--either expecting to free ride or obtain some benefit from an alternative policy option. This could be something like an economic benefit from future mining activities inside the reserve. Unlike Werner, we explicitly recognize both types of zero respondents.

Our approach is closely related to a sample selection model with a joint bivariate normal error distribution. We estimate a hierarchical system of two equations which mimics the data generating process. The first equation is the *decision* equation. Is one willing to report her true WTP? For those who do, our second equation, the *outcome* equation, explains the WTP amount which can be zero or positive.

As we note previously, our model is different from a standard sample selection model due two distinct features of the data generating process. The first feature is that those who are not willing to pay any amount (No responders) are mixed up in the lowest second-stage WTP interval resulting a misclassification problem. As Figure 1 shows, the lowest WTP interval contains both types of "No" responders as well as those with a low positive WTP. The second distinctive feature is that the participation regression model of WTP is an interval-censored or grouped data regression model with observations in all but the lowest interval correctly observed.[4]

Our general model is given by a hierarchical system of equations. The first stage is given by the following equation where P^* is the unobservable propensity to report the true WTP,

$$P_i{*}= \mathrm{Z}_i\delta + \eta_i \tag{1}$$

and a protest bid is observed when $P_i^* < 0$. Letting WTP_i^* represent the unobserved net utility from preserving the reserve and WTP_i represent the true willingness to pay, we have the following second stage equation when $P_i^* \geq 0$,

$$WTP_i^* = X_i\beta + \varepsilon_i \tag{2}$$

thus $WTP_i^* = WTP_i$ if $WTP_i^* \geq 0,$ otherwise $WTP_i = 0$ if $WTP_i^* < 0.$

The two error terms, ε_i and η_i, are assumed to have a joint bivariate normal distribution, $\Phi_2(\eta, \ \ \varepsilon)$, with zero means, standard deviations 1 and σ, respectively, and correlation

[4] Aside from the second feature, our model is similar to a partial observability model (Poirier, 1980) which has been used to model misclassified data (Nguimkeu et al., 2019; Tennekoon and Rosenman, 2014).

coefficient, ρ. We normalize the standard deviation of η to 1. Both sets of covariate vectors, Z_i *and* X_i, include constant terms which are identified along with the other parameters.

Note that WTP_i = 0 when either (i) P_i* < 0 or (ii) P_i*≥ 0 and WTP_i* < 0. However, neither of these two groups is identified in the Kakadu design but both inhabit the lowest WTP interval which includes zero. Again, the true value of WTP in this lowest response interval is either zero or a positive value less than the upper bound, UB.

Based on our assumed dgp, we have following four types of observations in the sample.

Type	DGP	Probability	Outcome
A1	P_i*< 0	$\Phi(-Z_i\delta)$	$LB_i = 0$
A2	P_i*≥ 0 and $WTP_i^* < 0$	$\Phi_2(Z_i\delta, -X_i\beta/\sigma, -\rho)$	$LB_i = 0$
A3	P_i*≥ 0 and $0 \leq WTP_i^* \leq UB_i$	$\Phi_2(Z_i\delta, UB_i/\sigma - X_i\beta/\sigma, -\rho)$ - $\Phi_2(Z_i\delta, -X_i\beta/\sigma, -\rho)$	$LB_i = 0$
B	P_i*≥ 0 and $LB_i < WTP_i^* \leq UB_i$	$\Phi_2(Z_i\delta, UB_i/\sigma - X_i\beta/\sigma, -\rho) -$ $\Phi_2(Z_i\delta, LB_i/\sigma - X_i\beta/\sigma, -\rho$	$LB_i > 0$

Types A1, A2 and A3 inhabit the lowest WTP interval which is bounded below by zero. Type A1 consists of protest responders who do not report their true WTP. Type A2 consists of those who are unwilling to pay any positive amount because of negative net utility and Type A3 consists of those who are willing to pay a small positive amount less than the interval upper bound. Type B contains all other responses-all lie in a WTP interval not bounded below by zero. The motivation behind protest responses also can be driven by negative net utility from the underlying valuation as for Type A2 but also can be a different reason, for example the disutility from the involuntary participation in the survey. This is why a two-equation model is needed to capture different motivations behind the two types of zero responses.

In the usual model the econometrician can uniquely identify the observation type. However, in our dataset, Types A1, A2 and A3 cannot be distinguished. We know only that an observation is either Type A1, Type A2 or Type A3 when we observe $LB_i = 0$. Type B respondents are of the usual type found in grouped-data or interval-censored regression models where their WTP responses are known up to interval strength.

Due to Kakadu data limitations, we cannot distinguish Type A1, Type A2 and Type A3 so we combine the first three cases above to obtain the probability that $LB_i = 0$ which is

$$\Pr(LB_i = 0) = \Phi(-Z_i\delta) + \Phi_2(Z_i\delta, UB_i/\sigma - X_i\beta/\sigma, -\rho). \tag{3}$$

This expression represents all respondents in the lowest response interval which is bounded below by zero.

For the remaining Type B observations, the probability that $LB_i > 0$, that is, the probability of reporting a WTP in any but the lowest response interval, is given by

$$\Pr(LB_i > 0\ ) = \Pr(LB_i > 0\ \&\ P_i{}^* \geq 0) = \Phi_2(Z_i\delta, UB_i/\sigma - X_i\beta/\sigma, -\rho) - \Phi_2\left(Z_i\delta, \frac{LB_i}{\sigma} - \frac{X_i\beta}{\sigma}, -\rho\right), \text{where } \mathrm{LB}_i \neq 0. \tag{4}$$

Based on the expressions given by (3) and (4) above, the following log-likelihood function for the model is obtained and can be estimated using observed data where *I* is the indicator function equal to one if the expression is true and zero otherwise,

$$logL = 1/\mathrm{N}\sum_{i=1}^{N} I(LB_i = 0)\ln(\Phi(-\mathrm{Z}_i\delta) + \Phi_2(\mathrm{Z}_i\delta, UB_i/\sigma - X_i\beta/\sigma, -\rho)) + I(LB_i > 0)\ln(\Phi_2(\mathrm{Z}_i\delta, UB_i/\sigma - X_i\beta/\sigma, -\rho) - \Phi_2(\mathrm{Z}_i\delta, LB_i/\sigma - X_i\beta/\sigma, -\rho)). \quad (5)$$

If $UB_i = \infty$, then, the second term, $\Phi_2(\mathrm{Z}_i\delta, UB_i/\sigma - X_i\beta/\sigma, -\rho) - \Phi_2(\mathrm{Z}_i\delta, LB_i/\sigma - X_i\beta/\sigma, -\rho)$ reduces to $\Phi(\mathrm{Z}_i\delta) - \Phi_2(\mathrm{Z}_i\delta, LB_i/\sigma - X_i\beta/\sigma, -\rho) = \Phi_2(X_1\beta_1, X_2\beta_2/\sigma_2 - LB_i/\sigma_2, \rho)$.

The model can be identified uniquely when (i) each covariate vectors has at least one variable which need not be different; and (ii) either LB_i has at least two different non-zero values or UB_i has at least two different finite values. The first order conditions and the information matrix are given in the appendix. Once the model parameters are estimated, the proportion of observations with zero-WTP is estimated as Pr $(WTP = 0) = 1/N\sum_{i=1}^{N}\Phi\left(-\mathrm{Z}_i\hat{\delta}\right) + \Phi_2(\mathrm{Z}_i\hat{\delta}, -X_i\hat{\beta}/\hat{\sigma}, -\hat{\rho})$ and the mean WTP or $\overline{WTP} = 1/N\sum_{i=1}^{N}\max\left(X_i\hat{\beta}, 0\right)$. The estimation results are given in the next section.

5. Estimation Results

We estimate the same specification as Werner (1999) in order to facilitate comparisons and help establish the usefulness of our approach. The results from estimating our decision and outcome equations with misclassification are given in Table 2. Variable descriptions are given in column 2 of Table 2. The results from estimating the decision equation are given in column 3 of Table 2. The coefficients of *Recparks*, *Lowrisk*, *Aboriginal*, *Finben*, *Mineparks*, and major are statistically significant at the α = 0.10 level or lower. The results from estimating the outcome equation are given in column 4 of Table 1. All of the coefficients are significantly different from zero at the α = 0.05 level or lower (*Jobs*, *Finben*, *Mineparks*, *Moreparks*, *Envcon*, *Age*, and *Income*).

Given the scale differences between Werner's model and our own, we cannot directly compare parameter magnitudes. However, we are able to compare the statistical significance of coefficients and the maximized value of the likelihood function across models. In addition, all models provide estimates of the actual fraction of the respondents unwilling to pay any positive amount, along with estimates of the mean and median willingness to pay.

In comparison to Werner's results, the signs of our corresponding coefficients are generally similar. As far as statistical significance is concerned, we note that the absolute values of t-ratios in Werner's model are generally larger than ours in the *decision* equation but are very similar in the *outcome* equation.

Werner estimates that about 25% of the respondents have zero willingness to pay. Our estimate of the percentage of the respondents with zero willingness to pay is somewhat higher at 32.9%. Werner's approaches yield mean WTP estimates ranging from A$489 to A$639.[5] Our estimate is about A$155. Werner's median WTP estimates range from A$85 to A$87. Our estimate is about A$159. These differences might due, in part, to either our assumption of normality or our incorporation of correlated errors between our two equations. Finally, we do find evidence of a large negative and statistically significant correlation coefficient (-0.497)

[5] Werner estimates four different models, a univariate mixture model, a univariate model assuming all WTP are positive, a covariate mixture model, and a covariate model assuming the probability of a positive response is constant, all estimated using several different distributional assumptions. Across all models estimated by Werner, the range for mean is A$192.47 to A$63,956.95. For the median, the range is A$76.03 to A$95.93. Generally, a higher mean estimate is associated with a lower median estimate.

between the error terms in the *decision* and *outcome* equations. This parameter cannot be estimated with any of the other approaches currently in use.

6. Conclusions

This paper presents a complementary alternative to the mixture procedure of Werner (1999) and Fernandez, Leon, Steel, and Vasquez-Polo (2004). Our approach is based on an extension of the usual selectivity model in which the choices of some respondents are not directly observed and the WTP amounts are interval-censored. We apply this model to the estimation of the Kakadu WTP survey data examined by Werner in which individuals unwilling to pay any amount and those willing to pay a small positive amount are not identified as they are lumped together into a single interval response group. Werner solves this estimation problem by estimating a mixture model. We present a complementary approach based on the bivariate normal distribution. Compared to Werner, we estimate that a larger fraction of the respondents are unwilling to pay any amount. We also find a lower mean WTP, a higher median WTP, and a negative correlation between the errors in the willing to pay or not equation and in the WTP amount equation.

In addition to contingent valuation studies, a simplified version of our model is useful when one wishes to estimate a selectivity model of expenditures based on survey data. Surveys rely heavily on data collected in intervals to reduce collection time. Without additional questions on expenditures, this survey response data may contain a lower interval containing the zero responses which must be located to estimate the first stage in a selectivity model. A variation of our approach can be used to solve this problem.

References

Ahn, H. and J. L. Powell (1993) "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism." *Journal of Econometrics* 58, pp. 3–29.

Bellemare, M. F. (2012) "As You Sow, So Shall You Reap: The Welfare Impacts of Contract Farming." *World Development* 40, pp. 1418–34.

Bonnichsen, O., and S. B. Olsen (2016) "Correcting for Non-Response Bias in Contingent Valuation Surveys Concerning Environmental Non-Market Goods: An Empirical Investigation Using an Online Panel." *Journal of Environmental Planning and Management* 59, pp. 245–62.

Bravo-Ureta, B. E., W. Greene, and D. Solis (2012) "Technical Efficiency Analysis Correcting for Biases from Observed and Unobserved Variables: An Application to a Natural Resource Management Project." *Empirical Economics* 43, pp. 55–72.

Brouwer, Roy, Pieter van Beukering, and Elena Sultanian. 2008. "The Impact of the Bird Flu on Public Willingness to Pay for the Protection of Migratory Birds." *Ecological Economics* 64, pp. 575–85.

Carson, R. T., L. Wilks, and D. Imber (1994) "Valuing the Preservation of Australia's Kakadu Conservation Zone." *Oxford Economic Papers* 46, pp. 727–49.

Caudill, S. B., and S. L. Oswald. "A Sequential Selectivity Model of the Decisions of Arbitrators." *Managerial and Decision Economics* 14, pp. 261–67.

Caudill, S. B., M. Ayuso, and M. Guillen (2005) "Fraud Detection Using a Multinomial Logit Model with Missing Information." *Journal of Risk and Insurance* 72, pp. 539–50.

Choe, Y. S., and J. Jeong (1993) "Charitable Contributions by Low- and Middle-Income Taxpayers: Further Evidence with a New Method." *National Tax Journal* 46. pp. 33–39.

Del Saz-Salazar, S., and L. Garcia-Menendez (2001) "Willingness to Pay for Environmental Improvements in a Large City: Evidence from the Spike Model and from a Non-parametric Approach." *Environmental and Resource Economics*, *20,* pp. 103–112.

Dustmann, C., and A. van Soest. (2004) "An Analysis of Speaking Fluency of Immigrants Using Ordered Response Models with Classification Errors." *Journal of Business and Economic Statistics* 22, pp. 312–21.

Falaris, E. (1987). "A Nested Logit Migration Model with Selectivity." *International Economic Review* 28, pp. 429–43.

Fernandez, C., C. Leon, M. F. Steel, and F. Vazquez-Polo (2004) "Bayesian Analysis of Interval Data Contingent Valuation Models and Pricing Policies." *Journal of Business and Economic Statistics* 22, pp. 431–42.

Frey, U. J., and F. Pirscher (2019) Distinguishing protest responses in contingent valuation: A conceptualization of motivations and attitudes behind them. *PLoS ONE* 14(1): e0209872.

Fu, L., W. Gao, and N. Shi. (2011). "Estimation of Relative Average Treatment Effects with Misclassification." *Economics Letters* 111. pp. 95–98.

Garcia, D., and P. Riera. (2003). "Expansion versus Density in Barcelona: A Valuation Exercise." *Urban Studies* 40, pp. 1925–36.

Harpman, D. A., M. Welsh, and E. W. Sparling. (2004) "Unit Non-Response Bias in the Interval Data Model." *Land Economics* 80, pp. 448–62.

Hausman, J. A., J. Abrevaya, and F. Scott-Morton (1998) "Misclassification of the Dependent Variable in a Discrete-Response Setting." *Journal of Econometrics* 87, pp. 239–69.

Heckman, J.J. (1979) Sample Selection Bias as a Specification Error, *Econometrica*, 47, pp. 153-161.

Hsieh, C., and LF Lee. (2016). "A Social Interactions Model with Endogenous Friendship Formation and Selectivity." *Journal of Applied Econometrics* 31. pp. 301–19.

Hyslop, D. R., and W. Townsend. (2017). "Employment Misclassification in Survey and Administrative Reports." *Economics Letters* 155, pp. 19–23.

Joyce, T. (1994). "Self-Selection, Prenatal Care, and Birthweight among Blacks, Whites, and Hispanics in New York City." *Journal of Human Resources* 29, pp. 762–94

Kristrom, B. (1997). "Spike Models in Contingent Valuation." *American Journal of Agricultural Economics* 79, pp. 1013–23.

Lu, R., Y. Luo, and R. Xiao. (2014). "An MPEC Estimator for Misclassification Models." *Economics Letters* 125, pp. 195–99.

Main, B. G. M., and B. Reilly. (1993). "The Employer Size-Wage Gap: Evidence for Britain." *Economica* 60, pp. 125–42.

Malikov, E., Kumbhakar, S. C., & Sun, Y. (2016). Varying Coefficient Panel Data Model in the Presence of Endogenous Selectivity and Fixed Effects. *Journal of Econometrics*, *190, pp.* 233–251

McCarthy, I., D. Millimet, and M. Roy (2015). "Bounding Treatment Effects: A Command for the Partial Identification of the Average Treatment Effect with Endogenous and Misreported Treatment Assignment." *The Stata Journal*, 15, pp. 411–436.

Meyerhoff, J., & Liebe, U. (2008). Do protest responses to a contingent valuation question and a choice experiment differ?. *Environmental and Resource Economics*, *39*(4), 433-446.

Meyer, B. D., and N. Mittag. (2017). "Misclassification in Binary Choice Models." *Journal of Econometrics* 200, pp. 295–311.

Nguimkeu, P., Denteh, A., & Tchernis, R. (2019). On the estimation of treatment effects with endogenous misreporting. *Journal of Econometrics*, *208*(2), 487-506.

Poirier, D. J. (1980). Partial observability in bivariate probit models. *Journal of Econometrics*, *12*, pp. 209-217.

van Ophem, Hans. (2000). "Modeling Selectivity in Count-Data Models." *Journal of Business and Economic Statistics* 18, pp. 503–11.

Schmertmann, C. P. (1994). "Selectivity Bias Correction Methods in Polychotomous Sample Selection Models." *Journal of Econometrics* 60, pp. 101–32.

Shiu, J. (2016). "Identification and Estimation of Endogenous Selection Models in the Presence of Misclassification Errors." *Economic Modelling* 52, pp. 507–18.

Skeie, M. A., H. Lindhjem, S. Skjeflo, and S. Navrud (2019). "Smartphone and Tablet Effects in Contingent Valuation Web Surveys--No Reason to Worry?" *Ecological Economics* 165 (November).

Soderberg, M. and D. N. Barton (2014). "Marginal WTP and Distance Decay: The Role of 'Protest' and 'True Zero' Responses in the Economic Valuation of Recreational Water Quality." *Environmental and Resource Economics* 59, pp. 389–405.

Strazzera, E., R. Scarpa, P. Calia, G. Garrod, and K. Willis (2003). "Modelling Zero Values and Protest Responses in Contingent Valuation Surveys." *Applied Economics* 35, pp. 133–38.

Tennekoon, V., & Rosenman, R. (2014). 'Behold, a virgin is with HIV!'misreporting sexual behavior among infected adolescents. *Health economics*, *23*, pp. 345-358.

Trost, R. P., and LF Lee. (1984). "Technical Training and Earnings: A Polychotomous Choice Model with Selectivity." *Review of Economics and Statistics* 66, pp. 151–56.

Werner, M. (1999) "Allowing for Zeros in Dichotomous-Choice Contingent-Valuation Models." *Journal of Business and Economic Statistics* 17, pp. 479–86.

Whitehead, J. C. (1991) "Environmental Interest Group Behavior and Self-Selection Bias in Contingent Valuation Mail Surveys." *Growth and Change* 22, pp. 10–21.

Winkelmann, R. (1998). "Count Data Models with Selectivity." *Econometric Reviews* 17, pp. 339–59.

Figure 1. Our Decision Tree

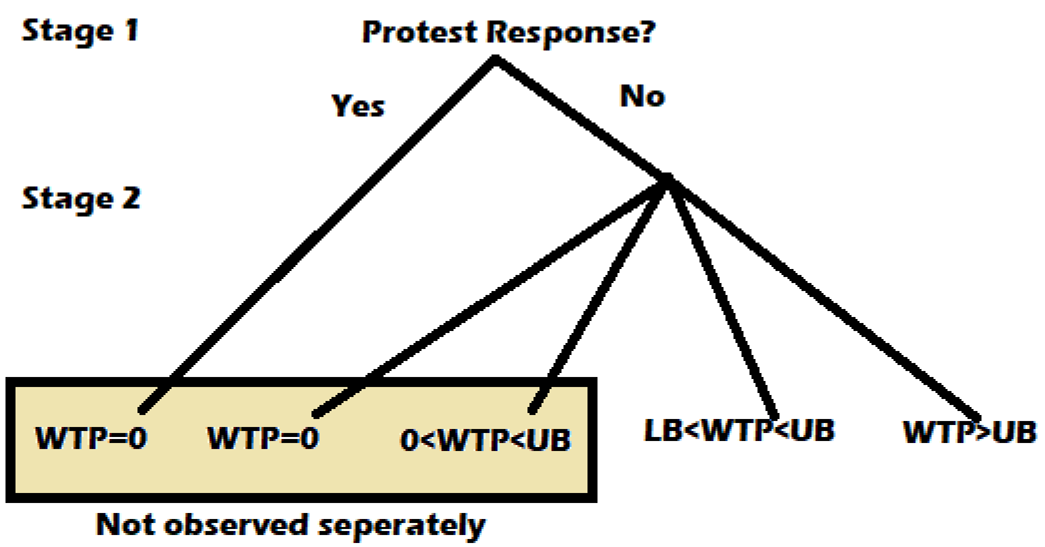

Table 1. Kakadu Elicitation Framework

Response (first answer/second answer)	Lower Bound (LB)	Upper Bound (UB)
(no/no)	0	t_{min}
(no/yes)	t_{min}	t_1
(yes/no)	t_1	t_{max}
(yes/yes)	t_{max}	∞

Table 2. Estimation Results for the Selectivity Model with Misclassification

Variable	Description	Decision Equation	Outcome Equation
Constant	-------	0.572 (0.90)	223.836*** (4.90)
Recparks	Higher values indicate higher value placed on nature preserves	0.206*** (3.14)	-------
Lowrisk	Higher values indicate greater acceptance of low-risk mining	-0.604*** (4.81)	-------
Aboriginal	Higher values indicate more weight given to importance of Kakadu to Aboriginals when making environmental decisions	0.113* (1.94)	-------
Jobs	Higher values indicate jobs are an important factor in making natural resource decisions	-------	-31.637*** (6.44)
Finben	Higher values indicate greater importance of financial benefits when making environmental decisions	-0.192** (2.43)	-28.272*** (4.92)
Mineparks	Higher values indicate a belief that mining in national parks reduces their value	0.470*** (5.64)	19.968*** (2.93)
Moreparks	Higher values indicate respondent favors more parks	0.088 (1.29)	14.661** (2.42)
Envcon	Equal 1 if respondent recycles and buys environmentally friendly products	-------	28.503*** (2.58)
Age	Respondent's age	0.004 (0.66)	-1.904*** (5.12)
Major	Equal to 1 if respondent received the major impact version of questionnaire	0.369** (2.25)	-------
Income	Respondent's annual income in thousands of dollars	-------	1.228*** (3.55)
Sigma	-------	-------	149.440*** (18.22)
Rho	-------	-------	-0.497*** (2.58)
Mean	-------	-------	155.15
Median	-------	-------	158.97
Log likelihood	-------	-------	-1506.49
WTP = 0	-------	-------	32.90%

[a]Figures in parentheses are absolute values of t-ratios.

***Indicates statistical significance at the $\alpha = 0.01$ level.

**Indicates statistical significance at the $\alpha = 0.05$ level.

*Indicates statistical significance at the $\alpha = 0.10$ level.

Appendix:

I. Contingent Valuation Studies

Contingent valuation (CV) involves asking individuals questions about their WTP (willingness to pay) for a good or their WTA (willingness to accept) compensation for a loss. WTP is the maximum price at or below which a consumer will buy one unit of a product. WTP cannot be observed directly. Purchase decisions of consumers of *marketed products* reveal some information about their WTP. If a consumer purchases a product at a given price, her WTP must be equal or larger than the price. This helps us to find one point of the CDF of WTP. If we can observe consumer choices at many different prices, we can identify several points along the CDF of WTP, which helps us to estimate the demand curve.

Unlike the case of marketed products, we have no information to infer the WTP of a non-marketed product, for example a new product to be introduced or a non-excludable common good. Contingent valuation surveys (CVS) are used to measure WTP of a non-marketed product, either using the direct method (stated preferences) or the indirect method (revealed preferences). Indirect method attempts to simulate an actual purchase decision.

Early contingent valuation surveys attempted to collect stated preferences using open-ended questions of the form "how much compensation would you demand for the destruction of X area?" or "how much would you pay to preserve X?". Such open-ended questions were identified to have several shortcomings: strategic behavior, protest answers, social desirability bias and respondents ignoring income constraints. Indirect method is the preferred method now.

In response to criticisms of CVSs, a panel of six eminent economists which included two Nobel laureates was appointed in 1992 by Bush Administration's National Oceanic and Atmospheric Administration (NOAA) to inquire into the validity of CV measures of "non-use value". The panel heard evidence from 22 expert economists and made recommendations on designing of CVSs (Arrow et. al., 1993). Among other recommendations, the panel suggested that CVSs should be designed in a yes or no referendum format instead of attempting to collect stated preferences. The CVS questions, thus, take the form "If you are compensated $P would you demand for the destruction of X area?" or "Would you pay $P to preserve X?".

By using this referendum format, similar to observing purchase decisions at a given price in case of marketed goods, we can identify two groups; those who have a valuation of at least $P and those who have a valuation less than $P. Unlike in the open-ended surveys, the referendum format does not help to identify many points along the distribution of WTP. Therefore, deriving the CDF of WTP requires additional assumptions, usually functional form assumptions.

In order to identify a two-parameter distribution, we need to know at least two points on the curve. Double bound format of CVSs involves an additional follow-up question of the referendum format and that allows to identify two points on the CDF of WTP.

Even with the single bound approach, several points along the CDF of WTP can be identified by randomly assigning different threshold values ($P) to participants. The single bound procedure is easier to implement than the double bound procedure, but the double bound estimator is more efficient than the single bound estimator. Data from a double bound CVS can be used to estimate the CDF of WTP, assuming a functional form with unknown parameters. The mean and median WTP, usually the values of interest, are sensitive to the choice of functional form. If a two-parameter distribution is assumed, a single fixed pair of threshold values is sufficient to identify the distribution as long as there are some observations in each of the three categories defined by these threshold values. The purpose of randomly assigning one of several threshold values is to make sure that the chosen values are neither too high nor too low.

Early research used a normal CDF due to convenience, but this choice was criticized later based on the argument that WTP should be non-negative. Consequently, non-negative CDFs such as log-normal, Gamma, and Weibull distributions became popular. The lowest interval generated by the *no-no* pair of WTP responses includes respondents who are unwilling to pay anything and some who are willing to pay a small positive amount. A CVS of referendum format does not identify these two groups separately. An additional complication arises due to protest responses which are observed as *no-no* pairs of WTP responses, but those responses do not reveal the true preferences of respondents. Two stage hierarchical structures (as in Heckman, 1974) and zero inflated structures including Tobit-like models have been imposed on CVS data to model these complexities. The limitation of these structures is that they do not differentiate protest zeros from true zeros. The proposed structure assumes that these two types of zeros are generated by two different processes while not demanding any additional information beyond what is collected through standard double bounded approach.

References

Arrow, K., Solow, R., Portney, P. R., Leamer, E. E., Radner, R., & Schuman, H. (1993). Report of the NOAA panel on contingent valuation. *Federal register*, 58(10), 4601-4614.

Heckman, J. (1974). Shadow prices, market wages, and labor supply. *Econometrica* 679-694.

II. Technical details

The log likelihood function given by equation (5) can be expressed as,

$$\mathcal{LL} = 1/\mathrm{N}\sum_{i=1}^{N} I(LB_i = 0)\ln\left(P_{0,i}\right) + I(LB_i > 0)\ln\left(P_{1,i}\right)$$

where $P_{0,i} = \Phi(-\mathrm{Z}_i\delta) + \Phi_2(\mathrm{Z}_i\delta, UB_i/\sigma - X_i\beta/\sigma, -\rho)$ and $P_{1,i} = \Phi_2(\mathrm{Z}_i\delta, UB_i/\sigma - X_i\beta/\sigma, -\rho) - \Phi_2(\mathrm{Z}_i\delta, LB_i/\sigma - X_i\beta/\sigma, -\rho)$.

First order conditions can be expressed as,

$$\begin{bmatrix} \frac{\partial \mathcal{LL}}{\partial \delta} \\ \frac{\partial \mathcal{LL}}{\partial \beta} \\ \frac{\partial \mathcal{LL}}{\partial \sigma} \\ \frac{\partial \mathcal{LL}}{\partial \rho} \end{bmatrix} = 1/\mathrm{N} \sum_{i=1}^{N} \frac{I(LB_i=0)}{P_{0,i}} . C_{0,i} + \frac{I(LB_i>0)}{P_{1,i}} . C_{1,i} = 0 \text{ where}$$

$$C_{0,i} = \begin{bmatrix} -\varphi(\mathrm{Z}_i\delta)\mathrm{Z}_i + \varphi\left(\frac{\left(\frac{UB_i}{\sigma} - \frac{X_i\beta}{\sigma}\right) + \rho \mathrm{Z}_i\delta}{\sqrt{(1-\rho^2)}}\right)\varphi(\mathrm{Z}_i\delta)\mathrm{Z}_i \\ -\varphi\left(\frac{\mathrm{Z}_i\delta + \rho\left(\frac{UB_i}{\sigma} - \frac{X_i\beta}{\sigma}\right)}{\sqrt{(1-\rho^2)}}\right)\varphi\left(\frac{UB_i}{\sigma} - \frac{X_i\beta}{\sigma}\right)\frac{X_i}{\sigma} \\ -\varphi\left(\frac{\mathrm{Z}_i\delta + \rho\left(\frac{UB_i}{\sigma} - \frac{X_i\beta}{\sigma}\right)}{\sqrt{(1-\rho^2)}}\right)\varphi\left(\frac{UB_i}{\sigma} - \frac{X_i\beta}{\sigma}\right)\left(\frac{UB_i}{\sigma^2} - \frac{X_i\beta}{\sigma^2}\right) \\ -\Phi_2(\mathrm{Z}_i\delta, UB_i/\sigma - X_i\beta/\sigma, -\rho) \end{bmatrix}$$

and

$$C_{1,i} = \begin{bmatrix} \varphi\left(\frac{\left(\frac{UB_i}{\sigma} - \frac{X_i\beta}{\sigma}\right) + \rho \mathrm{Z}_i\delta}{\sqrt{(1-\rho^2)}}\right)\varphi(\mathrm{Z}_i\delta)\mathrm{Z}_i - \varphi\left(\frac{\left(\frac{LB_i}{\sigma} - \frac{X_i\beta}{\sigma}\right) + \rho \mathrm{Z}_i\delta}{\sqrt{(1-\rho^2)}}\right)\varphi(\mathrm{Z}_i\delta)\mathrm{Z}_i \\ -\varphi\left(\frac{\mathrm{Z}_i\delta + \rho\left(\frac{UB_i}{\sigma} - \frac{X_i\beta}{\sigma}\right)}{\sqrt{(1-\rho^2)}}\right)\varphi\left(\frac{UB_i}{\sigma} - \frac{X_i\beta}{\sigma}\right)\frac{X_i}{\sigma} + \varphi\left(\frac{\mathrm{Z}_i\delta + \rho\left(\frac{LB_i}{\sigma} - \frac{X_i\beta}{\sigma}\right)}{\sqrt{(1-\rho^2)}}\right)\varphi\left(\frac{LB_i}{\sigma} - \frac{X_i\beta}{\sigma}\right)\frac{X_i}{\sigma} \\ -\varphi\left(\frac{\mathrm{Z}_i\delta + \rho\left(\frac{UB_i}{\sigma} - \frac{X_i\beta}{\sigma}\right)}{\sqrt{(1-\rho^2)}}\right)\varphi\left(\frac{UB_i}{\sigma} - \frac{X_i\beta}{\sigma}\right)\left(\frac{UB_i}{\sigma^2} - \frac{X_i\beta}{\sigma^2}\right) + \varphi\left(\frac{\mathrm{Z}_i\delta + \rho\left(\frac{LB_i}{\sigma} - \frac{X_i\beta}{\sigma}\right)}{\sqrt{(1-\rho^2)}}\right)\varphi\left(\frac{LB_i}{\sigma} - \frac{X_i\beta}{\sigma}\right)\left(\frac{LB_i}{\sigma^2} - \frac{X_i\beta}{\sigma^2}\right) \\ -\Phi_2(\mathrm{Z}_i\delta,\ UB_i/\sigma - X_i\beta/\sigma, -\rho) + \Phi_2(\mathrm{Z}_i\delta,\ LB_i/\sigma - X_i\beta/\sigma, -\rho) \end{bmatrix}.$$

The information matrix is,

$$I = -E\begin{bmatrix} \frac{\partial^2 \mathcal{LL}}{\partial\delta\partial\delta'} & \frac{\partial^2 \mathcal{LL}}{\partial\delta\partial\beta'} & \frac{\partial^2 \mathcal{LL}}{\partial\delta\partial\sigma} & \frac{\partial^2 \mathcal{LL}}{\partial\delta\partial\rho} \\ \frac{\partial^2 \mathcal{LL}}{\partial\beta\partial\delta'} & \frac{\partial^2 \mathcal{LL}}{\partial\beta\partial\beta'} & \frac{\partial^2 \mathcal{LL}}{\partial\beta\partial\sigma} & \frac{\partial^2 \mathcal{LL}}{\partial\beta\partial\rho} \\ \frac{\partial^2 \mathcal{LL}}{\partial\sigma\partial\delta'} & \frac{\partial^2 \mathcal{LL}}{\partial\sigma\partial\beta'} & \frac{\partial^2 \mathcal{LL}}{\partial\sigma^2} & \frac{\partial^2 \mathcal{LL}}{\partial\sigma\partial\rho} \\ \frac{\partial^2 \mathcal{LL}}{\partial\rho\partial\delta'} & \frac{\partial^2 \mathcal{LL}}{\partial\rho\partial\beta'} & \frac{\partial^2 \mathcal{LL}}{\partial\rho\partial\sigma} & \frac{\partial^2 \mathcal{LL}}{\partial\rho^2} \end{bmatrix} = 1/\mathrm{N}\sum_{i=1}^{N}\frac{I(LB_i=0)}{P_{0,i}} . C_{0,i}\, C'_{0,i} + \frac{I(LB_i>0)}{P_{1,i}} . C_{1,i}C'_{1,i}.$$