Handling protest responses in Contingent Valuation surveys

Mark Pennington PhD[1], Manuel Gomes PhD[2] and Cam Donaldson PhD[3]

[1] King's Health Economics, King's College London, London, UK

[2] Department of Health Services Research & Policy, London School of Hygiene & Tropical Medicine, London, UK

[3] Yunus Centre for Social Business and Health, Glasgow Caledonian University, Glasgow, UK

Corresponding author: Mark Pennington, PhD, King's Health Economics, P024 David Goldberg Centre, Institute of Psychiatry, Psychology & Neuroscience, King's College London, De Crespigny Park, London SE5 8AF. Tel: 020 7848 0589, Fax: 020 7848 0458

Email: mark.w.pennington@kcl.ac.uk

The study was undertaken whilst the first author was employed at the London school of Hygiene & Tropical Medicine.[1] The work that was not supported by funding from an external organisation. MG is recipient of a Medical Research Council Early Career Fellowship in Economics of Health. The views expressed are those of the authors alone.

Key words: Heckman Selection, Multiple Imputation, Contingent Valuation, missing data, EuroVaQ

Running title: Heckman and MI for protesters

Word count: 4644

[1] *No external financial support was provided for this study. The study was undertaken whilst MP was employed by the London School of Hygiene & Tropical Medicine. The authors had complete independence in designing the study, interpreting the data, writing, and publishing the report.*

Abstract

Objectives

The application of Contingent Valuation (CV) is growing in health economics, particularly to quantify the monetary value of health gains. Protest responses, whereby respondents refuse to state the value they place on the health gain, are commonly encountered in CV studies, and they tend to be excluded from analysis. Inferences based solely on non-protesters may be biased because protesters tend to differ from non-protesters on observed and unobserved characteristics that predict their responses. The Heckman selection model has been commonly used to adjust for protesters, but its underlying assumptions may be implausible in this context. We present a Multiple Imputation (MI) approach to appropriately address protest responses in CV studies, and compare it to the Heckman selection model.

Methods

This study exploits data from the multinational EuroVaQ study, which surveyed respondents' willingness-to-pay (WTP) for a Quality Adjusted Life Year (QALY). A simulation study assesses the relative performance of MI and Heckman selection models across different realistic settings grounded in the EuroVaQ study. We then illustrate the methods in the EuroVaQ study for estimating mean WTP for a QALY gain.

Results

We find that the MI provides lower bias and mean squared error compared to the Heckman approach across all scenarios considered, including different missing data mechanisms. The case study illustrates that, protesters are associated with a lower mean WTP for a QALY gain than non-protesters, but results differ according to method for handling protesters.

Conclusions

MI appears to be an appropriate method for addressing protest responses in CV studies.

Introduction

Contingent valuation (CV) surveys are one of the principal methods of valuing goods or services for which no market exists.[1] CV seeks the maximum willingness-to-pay (WTP) for a commodity or the minimum willingness to accept compensation for lack of a commodity through the presentation of hypothetical scenarios. Values are elicited from respondents in the form of an open response or the acceptance/rejection of a single or multiple values (bidding games). Valuation of commodities is an essential pre-requisite for Cost-Benefit Analysis,[2] and hence CV surveys are widely used in formulating environment and transport policy. Their application in health care is increasing in areas as diverse as diagnostic tests,[3] dental interventions,[4] and estimating the threshold value of a QALY for decision making within the cost-utility framework.[5]

There are well-documented challenges to the implementation of CV including strategic responses, anchoring or framing effects, and refusal to state a WTP value or indicate their willingness to pay a given value (protesting).[6-8] This paper focus on the specific issue of protesting. Respondents commonly refuse to state a WTP value or indicate their willingness to pay a given value in CV surveys. This may be because they place a zero value on the commodity. Alternatively, respondents may object to the principle of placing a monetary value on the commodity, or they may feel strongly that the responsibility for provision falls on another actor such as the Government.[9] Differentiation between zero values and protest responses is usually based on responses to a follow-up question requesting the selection of reason(s) for the refusal to respond from a menu of options. There is no universal agreement on the criteria for categorising responses as protest or zero values.[10] The number of protest responses can be sizeable. A recent review of 254 environmental CV studies indicates around 18% of respondents protested, but demonstrated considerable heterogeneity across studies.[11]

Protest responses are commonly excluded or assigned a zero value prior to estimating mean and median WTP.[10] Either approach may bias WTP estimates.[12] Zero is unlikely to reflect the value placed on the commodity by protesters. Excluding protesters relies on an assumption that the probability of protesting is independent of both observed and unobserved factors (analogous to missing completely at random, MCAR). If the differences between protesters and non-protesters can be explained by differences in the observed data, (protest) responses are said to be missing at random (MAR). In this case, bias caused by 'protesting' can be corrected by adjusting for observed factors that predict the likelihood of protesting. If the probability of protesting is associated with unobserved characteristics, then the responses are said to be missing not at random (MNAR), and conditioning on the observed data may not eliminate bias entirely.

Previous studies have considered the traditional Heckman selection model[13] to adjust for non-response (protesters) in contingent valuation studies.[14] The Heckman model addresses sample selection by adjusting the analysis (regression model) for the probability of being a protester (i.e. being selected to the sample). In other words, it recognises the possibility that the observed data (non-protesters) may not be a representative sample of the population of interest. An alternative approach to deal with sample selection is Multiple Imputation (MI).[15] This method was originally proposed to deal with non-response in surveys and has been applied in other areas such as biostatistics, epidemiology, and social sciences. With MI, the idea is to replace each missing (protest) response by a plausible value conditional on the observed data. The imputed values are often predicted from a regression model (imputation model) which includes all the variables associated with the response and the probability of being a protester.

Both Heckman and MI approaches can correct for the potential bias arising from protest responses by adjusting for observed differences between protesters and non-protesters. In principle, the standard Heckman model can also accommodate potential MNAR mechanisms, but this relies entirely parametric assumptions (about both model specification and distribution of the data).

A key distinction between Heckman and MI models is, therefore, the way these approaches deal with responses that are not Normally distributed. For example, the standard Heckman selection model assumes that the error terms for both the model for the probability of being a protester and the model for the observed data follow a bivariate Normal distribution. There is considerable evidence that the Heckman approach is highly sensitive to violations of this assumption.[16,17] While semiparametric[18] and non-parametric[19] extensions of the original Heckman model have been proposed, their implementation is challenging and not available in standard software. Alternatively, we can transform (Normalise) the response prior to estimation so that the Normality assumption is more plausible. However, this does not allow the response to be modelled in the original scale and requires back-transforming the parameter of interest which may be prone to issues such as heteroscedasticity. Unlike the Heckman approach, MI allows the imputation model to be estimated separately from the analysis model.[20] This provides MI with important advantages. Firstly, more plausible distributional assumptions can be made for the imputation model. For example, non-Normal responses can be normalised prior to imputation and back-transformed to the original scale before applying the analysis model to estimate the parameters of interest. Second, an appropriate model can be used to estimate the parameter of interest while maintaining the outcome of interest in the original scale. Third, both imputation and analysis can be modelled semi or non-parametrically in a relatively straightforward way.

This paper presents an MI approach to appropriately address protest responses in CV studies, and compares it to Heckman-selection models currently adopted in contingent valuation studies. We address this by comparing the methods in a simulation study across a range of realistic scenarios, and illustrating these approaches in the multinational EuroVaQ survey. The next section describes the motivating example. Then we introduce the statistical methods and the design for the simulation study. We then present the results of the simulation study and the case study. Finally, we consider the implications and limitations of the key findings.

Motivating example: The EuroVaQ Direct survey

The EuroVaQ study included two large CV surveys of over 37,000 people as part of a project to value a QALY.[21-23] Population sampling was broadly representative of the population distributions for age, sex, region of country and socio-economic status. The survey analysed here contained 13 questions and was split into four versions so that each respondent answered 4 or 5 questions. Data were obtained from 13,657 respondents in nine European countries. Respondents were allocated to a questionnaire version at random.

The format of the survey is described in detail elsewhere.[23] Respondents were initially asked to indicate their own health on a scale of 0 (death) to 100 (full health) and how long they expected to live for. The majority of the following CV questions assumed respondents maintained their current health state for their life expectancy if they purchased a treatment to avoid a health loss. The health increases from purchasing treatment were of predominantly one QALY in the form of improvements in quality of life (QOL) and gains in longevity. In this paper, we focus on responses to five 'key' questions. Each of these questions appeared in two of the four questionnaire versions. The questions described:

- Gain in QOL of 25 points over four years (used in simulation study)
- Gain in QOL of 10 points over ten years
- Gain in life expectancy of one QALY (at end of natural life)
- Avoidance of coma, duration equivalent to one QALY (longevity gain now)
- Postponement of death from terminal illness for one QALY (longevity gain now)

All five questions form the basis of this case study; simulation studies are performed using data from the first question.

Respondents provided open-ended WTP values constrained by a 'card sort' exercise. Prior to eliciting payment, respondents were asked whether they would be willing to pay for the health gain. Those who agreed to pay were presented with 15 cards containing values ranging from ca. 15USD to 460,000USD (in local currency) and asked to sort the cards into three categories: amounts they would pay, amounts they would not pay and amounts for which they were unsure. An open-ended maximum WTP value was then solicited within the range indicated by the respondent's card sort. Respondents unwilling to pay for the health gain were asked to select a reason. Consistent with previous analysis, we categorised these respondents as protesters if, from a menu of responses, they selected *solely* a statement that the Government should pay for health care. Respondents selecting any of the remaining statements (for example, a statement that they could not afford it) were assigned a WTP of zero.

Table 1 summarises response rates to the five key questions we analyse across each version of the questionnaire. Respondents choosing not to pay varied from 24% to 48% across questions in each of the versions. Between 6 and 10% of all respondents were classified as protesters. Protesters differed from non-protesters according to some observed characteristics, notably age, sex, social class and education level but not income (Table 2). Table 2 also reports mean WTP responses to the five questions according to whether the respondents chose to protest for one or more questions (but not always) or never protested. Values are reported in USD after conversion at purchasing power parity rates. With the exception of the Coma question, mean WTP values for respondents who sometimes protested were 53-86% lower than those who never protested. The distribution of WTP data was highly skewed with a long right tail and a spike at zero (Figure 1, Supplementary material). Log transformation reduced skewness and kurtosis but the resulting distribution was still far from Normal (Figure 2, Supplementary material).

Statistical methods to adjust for protesters

Heckman selection model

The Heckman model addresses sample selection by adjusting the analysis (regression model) for the probability of being a protester (i.e. being selected to the sample). The classical two-step approach involves using a probit regression to derive a correction factor (the inverse Mills ratio), which is included in a linear regression of the response.[24] This is often estimated by Limited information maximum likelihood (LIML), however, this approach is sensitive to collinearity between the inverse Mills ratio and the predictors of the response. Hence, fitting both models simultaneously using Full Information Maximum Likelihood (FIML) is recommended.[17] To help identification of the Heckman's model, estimated by either LIML or FIML, the selection model should include at least one variable that is predictive of the probability of response but unrelated to the response (exclusion restrictions).[25] A detailed description of the Heckman selection model is provided in the supplementary material.

A central assumption to the standard Heckman selection model is the bivariate Normality, and hence WTP data presents challenges to the application of this approach. The distribution of WTP data from open-ended responses is typically highly skewed with a spike at zero. Log transformation is commonly undertaken to reduce skew and generate approximately normal distributions. However, this has two limitations: 1) it does not allow the estimation and interpretation of the parameters of interest in the original scale; 2) the log transformation does not eliminate the spike at zero. To help address the latter, a Tobit specification can be used. This assumes the underlying values of the outcome $\mathrm{y_i}*$ are left censored at zero. More formally,

$$\mathrm{y_i} = \mathrm{y_i}* \text{ if } \mathrm{y_i}* > 0$$
$$\mathrm{y_i} = 0 \quad \text{if } \mathrm{y_i}* < 0$$

where $\mathrm{y_i}*$ is a latent variable $\mathrm{y_i} = \beta x_i + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$. The substitution of Tobit regression in place of OLS regression in the second step of the Heckman selection model has been advocated to allow for WTP data with a large proportion of zero values (Strazzera *et al.* 2003a).[14]

Multiple imputation

An alternative approach to deal with sample selection is multiple imputation (MI).[15] Briefly, MI involves replacing each missing (protest) response by a number of plausible values drawn from the posterior conditional distribution of the missing values given the observed data. After imputation, the outcome regression model is applied to each imputed dataset to estimate the parameters of interest. A detailed MI procedure is described in the supplementary material.

A flexible MI approach to address the distributional challenges inherent in WTP data is to use chained equations.[26] When variables are highly skewed or semi-continuous, semi-parametric imputation methods, such as Predictive Mean Matching (PMM) are recommended.[27] Rather than imputing values directly from a posterior Normal distribution, PMM replaces missing observations using the observed value whose linear prediction matches the closest linear prediction of the missing value. This guarantees that the imputed values are sampled only from the observed values, and respects the distribution of the data.

Simulation design

Missing data were simulated from the observed WTP responses to one of the five questions in the case study - the health gain of 25 points over 4 years. For the purposes of the simulation, we focused on the sub-sample of patients who responded to this question and assumed that the mean WTP derived from the observed responses (n = 7938) was the 'true value'. We then set some of the responses to missing, and assessed how well the estimates provided by the different adjustment methods compared to the 'true' values. This allowed us to assess the relative performance of the methods in a realistic case study rather than using stylised simulated data derived from parametric assumptions.

Briefly, we examined three broad settings in which missing data (protest responses) were simulated as MCAR, MAR and MNAR. For the MCAR setting we randomly replaced a proportion of WTP observations with missing values. For the MAR setting we simulated missing data using a model in which the chance of protesting was associated with WTP responses to other survey questions. For the MNAR settings, the probability of protesting was associated with the WTP response itself. In all three settings, we varied the proportion of missing data across the range 10-50%. Finally, we also generated missing responses for all respondents selected ‘government should pay’ as a reason for not electing to pay for any other health gain in the survey regardless of whether they also selected a reason taken to indicate a zero WTP value (18% of responses). Further details of the simulation mechanisms are provided in the supplementary material.

Selection and imputation models included predictors such as individual characteristics (e.g. age, gender, income, education, etc.), country indicators, and the WTP responses for other health gain questions. For each broad missing data mechanism (MCAR, MAR and MNAR) we investigated the performance of the methods considering the whole sample (base-case), two subsets of the observed responses (scenarios 1 and 2), and an additional scenario (3) with a different selection/imputation model. These scenarios considered 20% missing data. More specifically:

1. We deleted all respondents with missing household income data.
2. A common response to extremely high values in contingent valuation studies is to delete the top 1% of positive WTP responses. This ‘trim’ mitigates the potential for very high values to disproportionately influence mean WTP. We deleted the top 1% of WTP responses to the 25 point/4 year QOL gain question.
3. We excluded the WTP responses to the other health gain questions from both the selection and imputation models (mimicking a scenario with a single WTP question).

Implementation

We estimated Heckman selection models using both FIML and LIML, considering a Tobit specification for the outcome regression. With the Heckman model, it is commonplace in the literature to log transform WTP data prior to modelling and then interpret the coefficients of the semi-Log regression model. It is rarely acknowledged that such inference concerns the geometric rather than the arithmetic mean. Conversion to the arithmetic mean is possible with the use of smearing factors,[28] but complicated by the presence of heteroskedasticity.[29] To avoid these issues and allow comparison of arithmetic means, we applied the Heckman approach to the log-transformed WTP response, but then back-transformed the predicted values to the original scale prior to estimating mean WTP.

Both MI approaches (with and without PMM) used a two-stage approach[30] to accommodate the spike in WTP values at zero: logistic regression to impute a binary variable indicating 0 or 1 (positive values) for the missing WTP; conditional on imputing the value 1, a linear regression to impute positive values for each missing response.

For each scenario, we created 500 bootstrap replicates of the EuroVaQ data and generated the missing data in each bootstrap sample. Bootstrapping is often preferred to a Monte Carlo approach when we wish to simulate from the empirical distribution of the data rather than simulating from a specific parametric distribution.[31] We then applied the methods to the 500 datasets and calculated bias and rMSE as:

1. $Bias = \frac{1}{N}\sum_{l=1}^{N}\hat{\theta}_l - \theta_l$
2. $rMSE = \sqrt{\frac{1}{N}\sum_{l=1}^{N}\left(\hat{\theta}_l - \theta_l\right)^2}$

Where θ denotes the true mean and $\hat{\theta}$ the estimate obtained from each method in the $l = 1, \dots, N$ replicated dataset, with $N = 500$. Briefly, biases closer to zero and lower rMSE indicate ‘better’ performance of the methods. While the bias assesses the deviations from the true value, the rMSE quantifies the overall accuracy of the method, which includes bias and variability.

In the appendix we tabulate the distribution of the observed data set to missing with the distribution of the predicted values derived from MI and Heckman selection models for the raw (not bootstrapped) data. This study has not considered confidence interval coverage since our primary concern is how well (least biased) each method performs compared to the true mean WTP, rather than Type-I or Type-II errors related to hypothesis testing.

Illustrating the methods in the case study

In the re-analysis of the case study, we applied the FIML Heckman selection model and MI with PMM to 'predict' WTP values for protesters for the five key questions: the 25 point/4 year and the 10 point/10 year QOL gains arising imminently, and the three gains in life expectancy. We applied these approaches the same way as in the simulations except that when undertaking MI we treated household income as a continuous variable and imputed missing values. Confidence intervals around both mean and median WTP values were derived from 1000 bootstrap replications.

This study was undertaken without external funding support.

Results

Simulation study

To simplify the presentation of the results, we focus the reporting on the performance of the FIML Heckman selection model and MI with PMM. Results for the LIML Heckman selection model, the Tobit variant and MI without PMM with 20% missing data are provided in the supplementary material. Table 3 reports the bias and rMSE derived from MI and Heckman models in each of the three base case settings (MCAR, MAR and MNAR). Overall, MI led to the least biased results and lowest rMSE compared with the Heckman selection model, irrespective of the missing data mechanism. For MI, in both the MCAR and MAR settings, bias and rMSE were consistently low across all missing data proportions; bias and rMSE were generally much higher with the Heckman selection models. Simulation results with the Heckman selection models showed a pattern in which mean WTP was either considerably underestimated or overestimated, and the proportion of simulations in which mean WTP was overestimated increased with the proportion of missing data. As a result, biases are negative at 10-20% missing data and positive at 40-50% missing data. As expected, MI performs poorly when the data is MNAR. However, bias is generally lower than that observed with the Heckman selection model and rMSE is always lower.

Table 4 reports bias and rMSE across the three additional scenarios within each broad missing data setting (MCAR, MAR and MNAR). MI continued to outperform the Heckman selection model in terms of bias and rMSE. Excluding respondents with missing income data had little impact on the performance of either the MI or Heckman methods. After trimming the top 1% of WTP responses, bias was considerably reduced for the Heckman selection models, but it remained larger than bias with MI; both methods performed much better in the MNAR scenario compared to the base case. Excluding covariate WTP data (other WTP questions) had a detrimental effect on bias and rMSE for both MI and Heckman selection models, but the impact was small in the MCAR and MAR scenarios with MI. Estimation of mean WTP for protesters in EuroVaQ

Table 5 reports the mean and median WTP values for all five ‘key’ health gain questions for protesters and for all respondents according to method. Overall, mean WTP values are modestly reduced after adjusting for protesters using MI or a Heckman selection model. Confidence intervals around mean WTP values indicate a significant difference between mean WTP for gains in QOL and gains in longevity in the coma scenario. A further premium is placed on gains in longevity in the terminal illness scenario. These results strengthen the findings of previous analysis which did not adjust for protesters.[23]

After MI, mean WTP values for protesters as a percentage of the mean for non-protesters ranged from 34% (25 point/4 year QOL gain) to 47% (increase in life expectancy). These ratios are similar to those observed when comparing mean WTP for respondents who sometimes protested with means for respondents who never protested (Table 2). After applying the FIML selection model, mean values for protesters were 4-10% of the corresponding means for non-protesters across the five questions. Figure 1 shows the distribution of the raw WTP data for non-protesters and predictions for protesters derived using MI and the Heckman selection model for three of the five questions.

Discussion

We assessed MI and Heckman selection models across a range of realistic settings in which empirical WTP data were set to missing completely at random (MCAR), missing dependent on observed respondents' WTP for other questions in the survey (MAR), and missing dependent on (unobserved) respondents' WTP (MNAR). Overall, MI using PMM resulted in lower bias and rMSE; mean WTP was consistently underestimated using a Heckman selection model at lower proportions of missing data and overestimated at higher proportions. The Heckman selection model erratically in simulations , possibly due to violations of Normality in the distribution of the log WTP data. This is mitigated after trimming the top 1% of WTP values. However, bias and rMSE associated with MI were still lower compared to the Heckman model. While in theory the Heckman approach may provide flexibility to accommodate data that are MNAR, our results suggest that the violation of the bivariate Normality assumption may outweigh those benefits. The limitations of this approach in non-Normal data are well documented.[20] We have illustrated the application of MI to open ended WTP data; application to dichotomous data and multiple bid data is relatively straightforward.

The MI approach performed well across all MAR scenarios and no worse than complete-case analysis when the data is MNAR. This relied on the inclusion of all observed covariates predicting missingness, notably the additional WTP response data. Exclusion of this data led to higher bias and rMSE, particularly where missingness was not at random. Future studies should carefully consider all variables associated with both the probability of protesting (missing) and the incomplete response, including other WTP responses if these exist, so that the MAR assumption is more plausible. Sensitivity analysis of the impact of potential departures from MAR is recommended.

This paper considered two MI approaches (two-step MI based on Normality or with PMM) that can make more plausible distributional assumptions of WTP responses in CV studies. Both are easily implemented within standard statistical software. Surprisingly, our simulations found that the MI under Normality performed nearly as well as MI with PMM. These findings corroborate previous studies which found that MI is relatively robust to departures from Normality.[32,33] Importantly, MI offers further advantages compared to the Heckman approach: missing data in covariates (such as household income) is naturally accommodated; it can be combined with a wide range of models to estimate the parameter of interest; and the ease of application of semi-parametric methods can avoid the need to transform the dependent variable.

A number of authors have proposed modifications to the Heckman selection model that allow relaxation of some of the distributional assumptions (Vella[18] provides a useful summary). Gallant and Nychka propose a semi-parametric method that relaxes the assumption of bivariate Normality in the error terms.[34] Two-step parametric methods have also been developed that sidestep the requirement for bivariate Normality including the use of copula functions to transform the error terms into bivariate Normal distributions.[35-37] However, semi-parametric methods are computationally intense and the two-step parametric approaches remain susceptible to collinearity problems in the absence of a strong instrument. Despite the availability of a rich source of covariate data, we did not identify any variable in the EuroVaQ data which was strongly predictive of missingness (protesting) but unrelated to observed WTP values. This is a common challenge when estimating selection models.

Both Heckman and MI approaches suggest that protesters place a lower value on health gains than non-protesters. The results after MI indicate mean WTP for protesters of roughly 40% of the corresponding mean WTP for non-protesters. This ratio is similar to the ratio between observed WTP for respondents who sometimes protest and respondents who never protest (Table 2) lending support to the results from MI. Studies in health care which have examined WTP for protesters are limited. Gervès-Pinquié *et al* report lower mean WTP for protesters after applying a Heckman selection model to data on the WTP for informal care.[38] In a study of WTP for colorectal cancer tests Whynes and colleagues characterised protesters using post valuation comments collected from all respondents.[39] They reported mean WTP 25-30% lower for protesters. Evidence of the relative value placed on environmental commodities by protesters compared to non-protesters is conflicting, with some studies reporting higher WTP[40,41] and others reporting lower WTP[42,43].

This study has some limitations. The survey was undertaken online which facilitated a large sample size. However, respondents may not have given the survey their full attention, potentially reducing the quality of the data. Respondents were offered a limited menu of responses after electing not to pay and protesters were narrowly defined. Whilst this gives some confidence that respondents classified as protesters were correctly identified, we may have misclassified respondents electing not to pay because they found the scenario implausible, or they disengaged from the survey. A further limitation is that the survey was not incentive compatible. The distribution of the EuroVaQ WTP data is highly skewed. Whilst we observed that MI outperformed Heckman selection models even after 'trimming' the top 1% of responses it is possible that these distributions, and the resulting poor performance of selection models, do not generalize beyond the valuation of health. For the purpose of comparing the methods, we have generated missing data from the empirical WTP responses in the EuroVAQ study. While the true data generation process is unknown, this allowed us to test the methods in a realistic setting. More importantly, we were able to control for the missing data mechanism, and applied the same analysis model (to estimate mean WTP) across all scenarios, so that any differences across the analytical methods could be attributed to their ability to handle the missing data.

Conclusions

Previous studies have used the Heckman selection model to correct for selection bias arising from protest responses in CV surveys. Our simulation studies found that MI outperformed selection models across all MCAR, MAR and MNAR settings. They provided further evidence that selection models are sensitive to the bivariate Normality assumption and this may result in misleading inferences in the context of CV. MI appeared to generate more plausible WTP values for protesters in EuroVaQ, a large contingent valuation survey of health gains, and indicated that protesters place a mean value on health gains approximately half that of non-protesters. MI is easy to implement and provides additional flexibility to accommodate missing covariates and zero WTP values. We recommend the use of MI to adjust for protest responses in the analysis of CV data.

Acknowledgments

MG is recipient of a Medical Research Council Early Career Fellowship in Economics of Health (MR/K02177X/1).

References

1. Mitchell RC, Carson RT. Using surveys to value public goods: the contingent valuation method. Resources for the Future. 1989. Washington, DC.
2. Boadway RW, Bruce N. Welfare economics. B. Blackwell; 1984.
3. Lin PJ, Cangelosi MJ, Lee DW, Neumann PJ. Willingness to pay for diagnostic technologies: a review of the contingent valuation literature. Value In Health. 2013;16(5):797-805.
4. Vernazza CR. The monetary value of oral health: willingness to pay for treatment and prevention. PhD thesis, Newcastle University. 2011. Available at: https://theses.ncl.ac.uk/dspace/bitstream/10443/1202/1/Vernazza11.pdf
5. Nimdet K, Chaiyakunapruk N, Vichansavakul K, Ngorsuraches S. A Systematic Review of Studies Eliciting Willingness-to-Pay per Quality-Adjusted Life Year: Does It Justify CE Threshold?. PloS one. 2015;10(4):e0122760.
6. Bateman IJ, Carson RT, Day B, Hanemann M, Hanley N, Hett T, Jones-Lee M, Loomes G, Mourato S, Özdemiroglu E, Pearce DW. Economic valuation with stated preference techniques: a manual. Economic valuation with stated preference techniques: a manual. 2002.
7. Hanley N, Ryan M, Wright R. Estimating the monetary value of health care: lessons from environmental economics. Health economics. 2003;12(1):3-16.
8. Klose T. The contingent valuation method in health care. Health policy. 1999;47(2):97-123.
9. Lindsey G. Market models, protest bids, and outliers in contingent valuation. Journal of Water Resources Planning and Management. 1994;120(1):121-9.

10. Jorgensen BS, Syme GJ, Bishop BJ, Nancarrow BE. Protest responses in contingent valuation. Environmental and resource economics. 1999;14(1):131-50.

11. Meyerhoff J, Liebe U. Determinants of protest responses in environmental valuation: A meta-study. Ecological Economics. 2010;70(2):366-74.

12. Donaldson C, Jones AM, Mapp TJ, Olson JA. Limited dependent variables in willingness to pay studies: applications in health care. Applied Economics. 1998;30(5):667-77.

13. Heckman JJ. Sample selection bias as a specification error. *Econometrica.* 1979;*47*(1): 153-61.

14. Strazzera E, Scarpa R, Calia P, Garrod GD, Willis KG. Modelling zero values and protest responses in contingent valuation surveys. Applied economics. 2003;35(2):133-8.

15. Rubin DB. Multiple imputation of nonresponses in survey data. 1987 Wiley, New York.

16. Manski CF. Anatomy of the selection problem. Journal of Human resources. 1989;24(3):343-60.

17. Puhani P. The Heckman correction for sample selection and its critique. Journal of economic surveys. 2000;14(1):53-68.

18. Vella F. Estimating models with sample selection bias: a survey. Journal of Human Resources. 1998;33(1):127-69.

19. Das M, Newey WK, Vella F. Nonparametric estimation of sample selection models. The Review of Economic Studies. 2003;70(1):33-58.

20. Little RJ, Rubin DB. *Statistical analysis with missing data*. John Wiley & Sons; 2002.

21. Donaldson C, Baker R, Mason H, Pennington M, Bell S, Lancsar E, Jones-Lee M, Wildman J, Robinson A, Bacon P, Olsen JA. European Value of a quality adjusted life year. Final Publishable Report. 2010. Available at: http://research.ncl.ac.uk/eurovaq/EuroVaQ_Final_Publishable_Report_and_Appendices.pdf

22. Robinson A, Gyrd-Hansen D, Bacon P, Baker R, Pennington M, Donaldson C, Team E. Estimating a WTP-based value of a QALY: the 'chained'approach. Social Science & Medicine. 2013;92:92-104.

23. Pennington M, Baker R, Brouwer W, Mason H, Hansen DG, Robinson A, Donaldson C. Comparing WTP values of different types of QALY gain elicited from the general public. Health economics. 2015;24(3):280-93.

24. Wooldridge JM. Econometric analysis of cross section and panel data. MIT press; 2010. Cambridge, MA.

25. Leung SF, Yu S. On the choice between sample selection and two-part models. Journal of econometrics. 1996;72(1):197-229.

26. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Statistics in medicine. 2011;30(4):377-99.

27. Little RJ. Missing-data adjustments in large surveys. Journal of Business & Economic Statistics. 1988;6(3):287-96.

28. Duan N. Smearing estimate: a nonparametric retransformation method. Journal of the American Statistical Association. 1983;78(383):605-10.

29. Manning WG. The logged dependent variable, heteroscedasticity, and the retransformation problem. Journal of health economics. 1998;17(3):283-95.

30. Yu LM, Burton A, Rivero-Arias O. Evaluation of software for multiple imputation of semi-continuous data. Statistical Methods in Medical Research. 2007;16(3):243-58.

31. Cheng RC. Analysis of simulation output by resampling. International Journal of Simulation: Systems, Science & Technology. 2000;1:51-8.

32. Van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Statistical methods in medical research. 2007;16(3):219-42.

33. Lee KJ, Carlin JB. Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. American journal of epidemiology. 2010;171(5):624-32.

34. Gallant AR, Nychka DW. Semi-nonparametric maximum likelihood estimation. Econometrica: Journal of the Econometric Society. 1987:363-90.

35. Olsen RJ. A least squares correction for selectivity bias. Econometrica: Journal of the Econometric Society. 1980;48(7):1815-20.

36. Lee LF. Some approaches to the correction of selectivity bias. The Review of Economic Studies. 1982;49(3):355-72.

37. Lee LF. Generalized econometric models with selectivity. Econometrica: Journal of the Econometric Society. 1983;51(2):507-12.

38. Gervès-Pinquié C, Bellanger MM, Ankri J. Willingness to pay for informal care in France: the value of funding support interventions for caregivers. Health economics review. 2014;4(1):1-8*.*

39. Whynes DK, Frew E, Wolstenholme JL. A comparison of two methods for eliciting contingent valuations of colorectal cancer screening. Journal of health economics. 2003;22(4):555-74.

40. Brouwer R, Martín-Ortega J. Modeling self-censoring of polluter pays protest votes in stated preference research to support resource damage estimations in environmental liability. Resource and Energy Economics. 2012;34(1):151-66.

41. Garcia S, Harou P, Montagné C, Stenger A. Models for sample selection bias in contingent valuation: Application to forest biodiversity. Journal of Forest Economics. 2009;15(1):59-78.

42. Collins AR, Rosenberger RS. Protest adjustments in the valuation of watershed restoration using payment card data. Agricultural and Resource Economics Review. 2007;36(2):321.

43. Strazzera E, Genius M, Scarpa R, Hutchinson G. The effect of protest votes on the estimates of WTP for use values of recreational sites. Environmental and resource economics. 2003;25(4):461-76.

Tables

Question	Total respondents	Version 1 4255	Version 2 4435	Version 3 4447	Version 4 4310
25 pts/4 years	excluded#	0			54 (1.3%)
	missing*	20 (0.5%)			12 (0.3%)
	protesters	254 (5.9%)			287 (6.7%)
	zeros	576 (13.5%)			599 (13.9%)
	positives	3405 (80.0%)			3358 (77.9%)
10 pts/10 years	excluded#			185 (4.2%)	198 (4.6%)
	missing*			13 (0.3%)	11 (0.3%)
	protesters			247 (5.6%)	298 (6.9%)
	zeros			618 (13.9%)	736 (17.1%)
	positives			3384 (76.1%)	3067 (71.2%)
Extra year at the end of life	missing*		15 (0.3%)		10 (0.2%)
	protesters		273 (6.2%)		365 (8.5%)
	zeros		1679 (37.9%)		1703 (39.5%)
	positives		2468 (55.6%)		2232 (51.8%)
Coma	missing*		16 (0.4%)		11 (0.3%)
	protesters		318 (7.2%)		410 (9.5%)
	zeros		865 (19.5%)		963 (22.3%)
	positives		3236 (73.0%)		2926 (67.9%)
Terminal Illness	missing*		14 (0.3%)		10 (0.2%)
	protesters		304 (6.9%)		351 (8.1%)
	zeros		1163 (26.2%)		1113 (25.8%)
	positives		2954 (66.6%)		2836 (65.8%)

*respondents with a positive WTP for whom the questionnaire failed to record the final WTP value

#respondents excluded from a particular question due to very low health or low life expectancy

Table 1. EuroVaQ data: characterisation of responses to five ‘key’ questions across each version of the questionnaire.

	Never protests		Sometimes/ always protests		
Variable	n	Mean	n	Mean	p value*
male	15,096	0.49	2,351	0.52	**0.005**
Age	15,096	44.5	2,351	45.6	**0.007**
Social Class	15,096	3.45	2,351	3.70	**<0.0001**
Age left Education	7,864	22.9	1,234	22.2	**0.002***
Household size	15,096	1.71	2,351	1.70	0.17
Household inc. (PPP$)	13,058	48,557	1,909	50,469	0.41*
Personal income (PPP$)	7,865	8,227	1,030	8,642	0.76*
Health (0-100)	15,096	83.2	2,351	82.3	**0.009**
Respondent education level	15,096	2.25	2,351	2.10	**<0.0001**
Head of household education level	15,096	2.19	2,351	2.06	**<0.0001**
WTP, 25 pts/4 yrs (PPP$)	7,284	11,352	654	3,970	**<0.0001***
WTP, 10 pts/10 yrs (PPP$)	7,182	11,925	623	5,655	**<0.0001***
WTP, extension of life (PPP$)	7,423	11,519	659	1,581	**<0.0001***
WTP, Coma (PPP$)	7,421	19,141	569	17,711	**<0.0001***
WTP, Terminal illness (PPP$)	7,424	30,626	642	11,022	**<0.0001***

*t test on log transformed data, PPP = Purchasing Power Parity

Table 2. EuroVaQ data: respondent characteristics and responses to ‘key’ questions according to respondent categories

	MCAR		MAR		MNAR	
	FIML Heckman model	MI with PMM	Heckman model	MI with PMM	Heckman model	MI with PMM
10% missing data						
Bias	-6,025	-188	-9,705	196	-10,163	-3,041
rMSE	6,633	2,619	10,908	3,362	10,929	4,707
20% missing data						
Bias	-4,729	7	-7,204	-222	-10,878	-4778
rMSE	5,496	1,806	11,762	2,725	12,103	5527
30% missing data						
Bias	-2,776	-72	965	250	-9,363	-6,174
rMSE	5,091	1,710	18,988	2,407	13,882	6,642
40% missing data						
Bias	-253	52	14,974	20	-4,775	-7,043
rMSE	7,586	1,664	40,213	2,339	21,470	7,344
50% missing data						
Bias	5,936	-34	61,263	-109	7,884	-7,533
rMSE	15,150	1,520	104,472	2,318	49,249	7,779
Missing data for respondents selecting ‘government should pay’ in other WTP questions						
Bias			-2,987	-39		
rMSE			3,022	577		

rMSE: root mean square error.

Table 3. Simulation studies: Bias and rMSE according to method for estimating the mean WTP derived from bootstrap replicates when data are MCAR, MAR and MNAR.

	MCAR		MAR		MNAR	
	FIML Heckman model	MI with PMM	Heckman model	MI with PMM	Heckman model	MI with PMM
Scenario 1: Individuals with missing income data are deleted						
Bias	-4,496	62	-5,356	28	-9,987	-4,629
rMSE	5,383	2,129	11,002	2,834	11,543	5,625
Scenario 2: Top 1% of WTP responses are deleted.						
Bias	-1,556	18	-1,044	131	-4,200	-1,680
rMSE	2,500	420	8,022	647	5,725	1,788
Scenario 3: WTP responses to other health gain questions are excluded from selection/imputation model						
Bias	-7,823	-201	-9,105	748	-15,855	-6,926
rMSE	8,151	1,959	9,341	1,791	16,095	7,391
Base case (20% missing)						
Bias	-4,729	7	-7,204	-222	-10,878	-4778
rMSE	5,496	1,806	11,762	2,725	12,103	5527

rMSE: root mean square error.

Table 4. Simulation studies: Bias and rMSE for the alternative methods across different sensitivity scenarios, compared to the base case, when data are MCAR, MAR and MNAR.

			FIML Heckman model		MI with PMM	
Question	Respondent category	n	Mean (PPP$)*	Median (PPP$)*	Mean (PPP$)*	Median (PPP$)*
WTP – 25 points 4 years	Protester	529	416	89	3,652	55
	Observed	7,751	10,807	1,468	10,807	1,468
	All [95% CI]	8,280	10,143 [8,703 - 12,032]	1,138 [1,098 - 1,150]	10,354 [9,192 - 12,256]	1,150 [1,150 - 1,468]
WTP – 10 points 10 years	Protester	543	488	84	4,327	77
	Observed	7,751	11,466	1138	11,466	1,138
	All [95% CI]	8,294	10,747 [9,263 - 12,717]	1,032 [854 - 1,078]	11,004 [9,650 - 13,467]	1,075 [1,072 - 1,138]
WTP - extension of life	Protester	619	1016	62	5,095	179
	Observed	7,832	10,785	160	10,785	160
	All [95% CI]	8,451	10,069 [8,063 - 12,527]	143 [114 - 182]	10,352 [8,410 - 13,018]	160 [149 - 200]
WTP - Coma	Protester	702	1825	211	7,798	323
	Observed	7,749	19,194	2,149	19,194	2,149
	All [95% CI]	8,451	17,751 [15,949 - 20,033]	1,548 [1,510 - 1,976]	18,255 [16,684 - 20,838]	1,647 [1,647 - 2,157]
WTP - Terminal illness	Protester	637	2,391	1,087	11,945	329
	Observed	7,814	29,246	1,176	29,246	1,176
	All [95% CI]	8,451	27,222 [24,239 - 31,580]	1,803 [1,617 - 2,276]	27,969 [25,444 - 32,193]	1,976 [1,791 - 2,298]

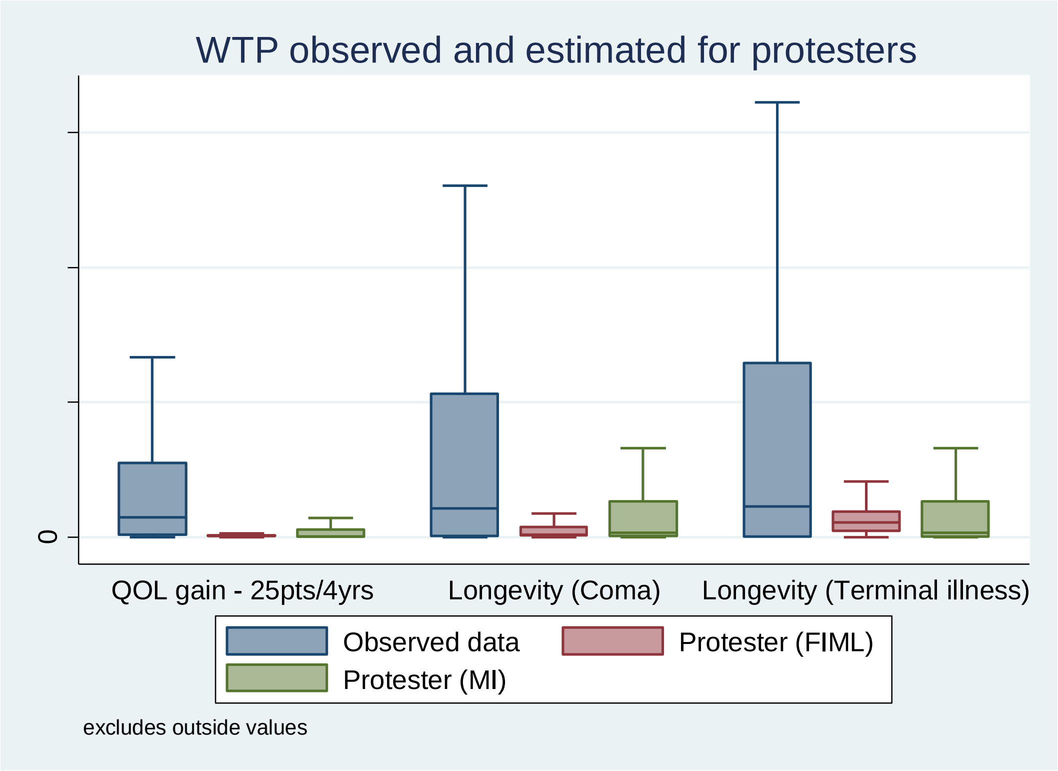
* PPP = Purchasing Power Parity

Table 5. Case study: Mean and Median WTP values for observed (non-protesters), protesters and the entire sample after adjusting for protest responses using Heckman selection models and MI using PMM, for each of the five questions in the EuroVaQ survey.

Figure Legend

Figure 1. Mean WTP and dispersion of WTP values for non-protesters (observed) and protesters according to method.

Figure


WTP observed and estimated for protesters
0
QOL gain - 25pts/4yrs
Longevity (Coma)
Longevity (Terminal illness)
Observed data
Protester (FIML)
Protester (MI)
excludes outside values

Supplementary material

Statistical methods to adjust for protesters.

Heckman selection model

Heckman's sample selection correction[1] is essentially a two-step process which treats sample selection as a form of omitted-variable bias. This approach involves a regression model (1), which relates the response y_i (say, the WTP) with the explanatory variables x_i, and the selection model (2) which relates a latent variable w_i with the variables z_i that predict the probability of being observed (non-protester), as described below,

$$y_i = \beta x_i + \varepsilon_i \tag{1}$$

$$w_i = \gamma z_i + v_i \tag{2}$$

The first step of Heckman's approach estimates the probability of observing the response, $P(r_i = 1)$, using a probit regression model:

$$probit\ P(r_i = 1|\ z_i) = probit\ P(w_i > 0|\ z_i) = \gamma z_i \tag{1}$$

where $r_i = 1$ if the response of individual i is observed, and $r_i = 0$ otherwise. Then, the predicted values are used to estimate a 'correction' factor, the inverse Mills ratio, $\lambda_i = \varphi(-\gamma z_i)/(1 - \Phi(-\gamma z_i))$. φ and Φ are the standard normal density and standard normal cumulative distribution functions, respectively. In the second step, the correction factor (λ_i) is included as an additional explanatory variable in the regression model,

$$E(y_i|x_i, \lambda_i) = \beta x_i + \rho\sigma_\varepsilon \lambda_i(-\gamma z_i) \tag{2}$$

$$\begin{pmatrix} v_i \\ \varepsilon_i \end{pmatrix} \sim BVN\left(\mathbf{0}, \mathbf{\Omega} = \begin{pmatrix} \sigma_v^2 & \rho\sigma_v\sigma_\varepsilon \\ & \sigma_\varepsilon^2 \end{pmatrix}\right)$$

where ρ is the correlation between unobserved determinants of the probability of response and unobserved predictors of the response itself, (v, ε) is Normally distributed and independent of z and x, and $\sigma_v^2 = 1$. Hence, conditional on the first step, model (2) assumes that individuals with observed responses are a random sample of the population, $E(y_i|r = 1, x_i) = \beta x_i + E(\varepsilon_i|\gamma z_i + v_i > 0)$.

Multiple imputation

With MI each missing value is replaced with a set of M plausible values.[2] Each of these values is drawn, in a Bayesian manner, from the conditional distribution of the missing observations given the observed data, so that the set of imputed values reflects the uncertainty associated with both the missing data and the estimation of the parameters in the imputation model. The regression model (2) is then applied to these multiple imputed datasets to estimate the parameters of interest. These M sets of estimates and accompanying measures of uncertainty are then combined using Rubin's rules[2] to properly reflect the variation both within and between imputations.

A popular approach to conduct MI is via fully-conditional specification (FCS) or chained equations, where missing values are imputed for one variable at a time.[3] When imputing a continuous variable, say y_i, standard implementation of MI via FCS typically draws values from the posterior Normal distribution:

$$y_i | w_i \sim N(\beta w_i, \sigma_\varepsilon^2) \tag{3}$$

The algorithm to impute the missing observations is as follows:

Step 1. Fit model (3) to the complete data to obtain $\hat{\beta}$, $\hat{\sigma}$ and the covariance matrix of these (Λ). w_i should include explanatory variables in model 2 (x_i) plus other auxiliary variables that are associated with the probability of observing the response and the missing values.

Step 2. Draw σ_ε^* and β^* from the joint posterior distribution of (σ_ε,β), where $\sigma_\varepsilon^* = \hat{\sigma}_\varepsilon \sqrt{(n_{obs} - k)/g}$ and $\beta^* = \hat{\beta} + (\sigma_\varepsilon^* / \hat{\sigma}_\varepsilon) u_1 \Lambda^{1/2}$. g is a random draw from a distribution on $n_{obs} - k$ degrees of freedom, u is the vector of random draws from $N(0,1)$ and $\Lambda^{1/2}$ is the Cholesky decomposition of Λ.

Step 3. Replace each missing observation by $y_i^* = \beta^* x_i + u_i \sigma^*$.

Step 4. Repeat steps 1-3 M times and apply model (2) to the multiple imputed datasets and obtain parameters of interest.

Step 5. Combine M multiple estimates using Rubin's rules.

When variables are highly skewed or semi-continuous (high proportion of zeros) such as the WTP responses it may not be possible to find a plausible transformation. In these cases, the use Predictive Mean Matching (PMM) may be more plausible.[4] The PMM procedure starts by estimating missing values using a linear regression for the incomplete variable (step 1). However, rather than imputing directly from posterior Normal distribution with mean $\beta^{*}x_i$, the linear prediction for each missing value is matched to the closest linear prediction for an observed value, with that observed value being used to fill in the missing observation. This guarantees that the imputed values are sampled only from the observed values of y_i, which may be desirable when a distribution is truncated or 'lumpy'.

Generation of Missing data in simulation studies

We generated missing data from the observed responses for the health gain of 25 points over 4 years. We chose this question because protest (missing) responses to this question were lowest across all questions in the survey. Where respondents had protested for this question, and the response was missing we dropped their data for the purposes of the simulation. Consequently, we knew the true response for each datum. Simulations were then performed by selecting data and changing values to missing (protest) responses. This approach avoids the need for parametric assumptions to mimic the likely distribution of the data – instead we utilize a large sample of real WTP data.

To simulate a situation in which protest responses are missing completely at random (MCAR) we simply selected n% of the data, at random, and changed values to missing. To simulate a situation in which responses are missing not at random (MNAR) we allowed the magnitude of the WTP response to influence data selection, with higher values having an increased probability of selection. Data were selected using the criteria shown below:

Let ρ_i = exp(log(δ/(1-δ) – αy_i^* + αy_i)/(1 + exp(log(δ/(1-δ)– αy_i^* + αy_i))

Let π_i = random number between 0 and 1

Replace y_i = missing if ρ_i < π_i, where y_i is the log of the WTP response, y_i^* is the mean of the log of the WTP response, and the variables α and δ take the values 0.1 and 0.095, 0.15

and 0.19, 0.2 and 0.285, 0.25 and 0.38, and 0.3 and 0.49 for missing data at 10%, 20%, 30%, 40% and 50%, respectively

To simulate a situation in which responses are missing at random (MAR) we allowed the magnitude of the response to other WTP questions to influence data selection, with higher values having an increased probability of selection. Data were selected using the criteria shown below:

Let ρ_i = exp(log(δ/(1-δ) – 0.5y_i^* + 0.5y_i)/(1 + exp(log(δ/(1-δ) – 0.5y_i^* + 0.5y_i))

Let π_i = random number between 0 and 1

Replace y_i = missing if $\rho_i < \pi_i$, where y_i is the mean of the quintile for each of the other WTP questions answered by the respondent, y_i^* is the intra-respondent mean of the respondent mean WTP quintile, and the variable δ takes the values 0.085, 0.18, 0.287, 0.39, and 0.5 for missing data at 10%, 20%, 30%, 40% and 50%, respectively

Specification of covariates and model optimisation

We included covariates for country; sex; age; social class; respondents' education level; head of household education level; household income; household size; health; working status and profession of respondent; working status and profession of head of household; and question order for questions appearing as part of a pair in which the order was randomised. These prognostic variables were fully observed with the exception of the household income (14.2% missing). To facilitate comparisons across methods in the simulation study, we used an ad-hoc approach to deal with missing data, by dividing values into quintiles and including an additional category for missing responses. For the simulation studies and all selection models household income was divided into quintiles at country level and a missing category created for missing income. Likewise, protest responses for each WTP question created missing data where such data was included as a prognostic variable. For the simulation studies we calculated the quintile value for responses to each question other than the 25 point/4 year QOL gain question answered by the respondent. We then calculated the mean across questions for each respondent. This mean value was missing for respondents who protested to all other questions. Hence, we specified this value as quintiles and added a category for missing responses when including in MI or Heckman selection models. Respondents' education was assigned to three levels representing compulsory only, some additional education and university degree or above. Social class was assigned using the ESOMAR algorithm to six classes (A, B, C1, C2, D, E). For the simulation studies we created two variables which reported the proportion of the questions offered (other than the 25 point/4 year QOL gain question) for which the respondent, firstly, elected not to pay and selected the 'govt should pay' option or, secondly, elected not to pay and selected one or more of the remaining reasons indicating a zero valuation. These variables were specified using fractional polynomials where this improved model fit as assessed by AIC.

In the final case study we did not use ad hoc methods to accommodate missing covariate data when applying MI. All missing data were imputed.

References

1. Heckman JJ. Sample selection bias as a specification error. *Econometrica.* 1979;*47*(1): 153-61.
2. Rubin DB. Multiple imputation of nonresponses in survey data. 1987 Wiley, New York.
3. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Statistics in medicine. 2011;30(4):377-99.
4. Little RJ. Missing-data adjustments in large surveys. Journal of Business & Economic Statistics. 1988;6(3):287-96.

Figures - Distribution of raw and log transformed data

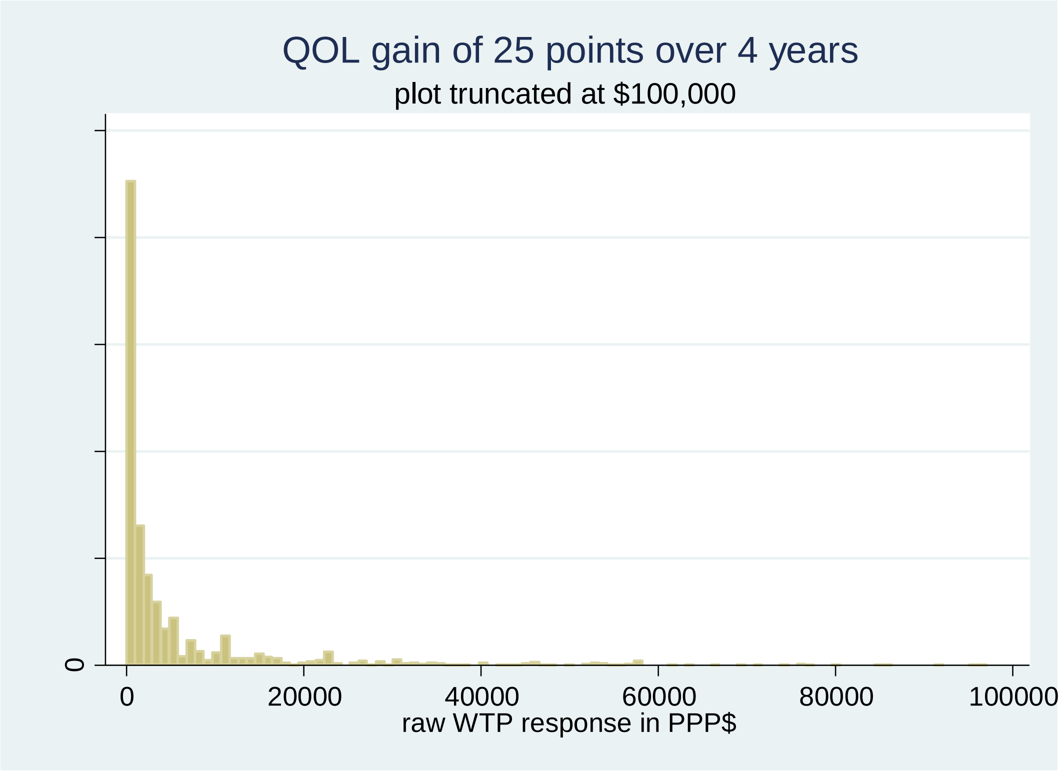


Figure 1. Distribution of raw WTP data

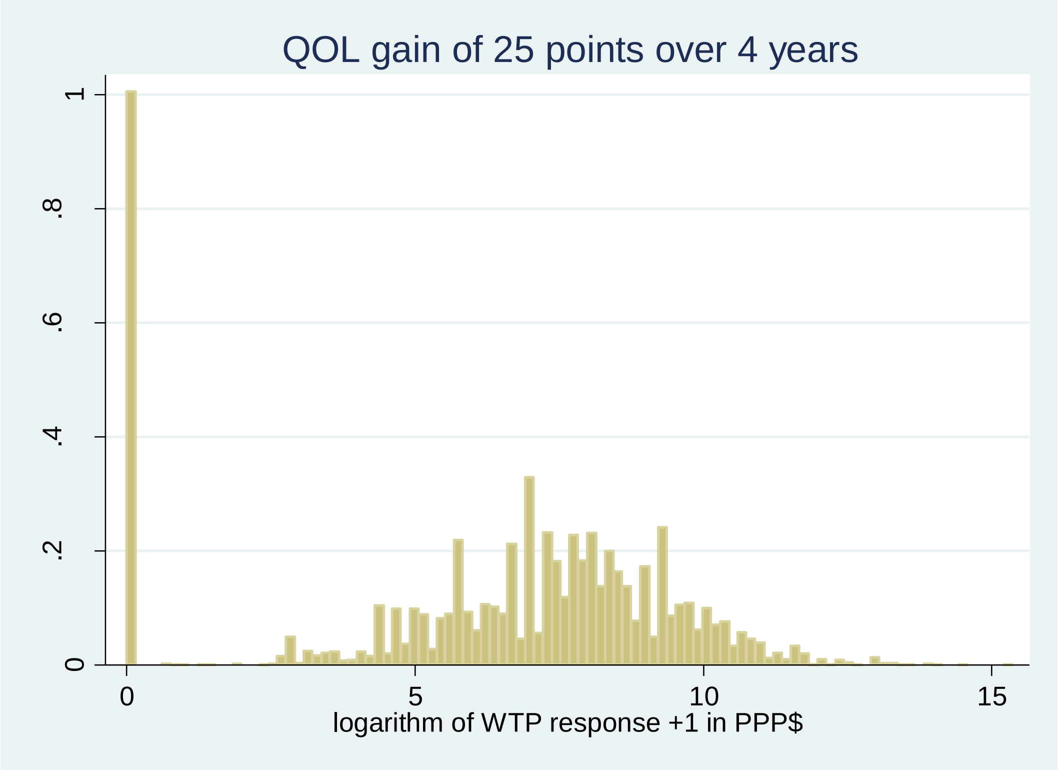


Figure 2. Distribution of logarithm of WTP values

Tables

	MCAR			MAR			MNAR		
Percentile	Observed (n = 1625)	Heckman model	MI with PMM	Observed (n = 1545)	Heckman model	MI with PMM	Observed (n = 1572)	Heckman model	MI with PMM
1%	0	4	0	0	1	0	0	3	0
5%	0	11	0	0	6	0	0	12	0
10%	0	23	0	72	66	34	77	100	0
25%	162	338	176	613	378	667	571	521	329
50%	1,150	1,963	1,489	3,069	1,881	3,073	2,596	3,384	2,196
75%	5,319	9,238	5,489	11,382	5,516	11,499	10,765	17,180	7,881
90%	20,013	26,847	17,248	30,732	9,604	30,958	28,747	37,654	22,998
95%	34,146	40,009	34,497	51,597	12,614	57,495	56,910	54,103	45,996
99%	133,417	64,123	126,488	229,978	18,203	234,018	453,617	88,934	161,745
Mean	10,288	8,230	11,329	15,996	3,641	17,586	20,673	12,594	12,001

Table S1. Simulation results for the base case MCAR, MAR and MNAR settings: distribution of WTP responses for the observed sample ('True') with 20% missing data, and those predicted by FIML Heckman selection model and MI with PMM; bias and rMSE according to method for estimating the mean WTP derived from bootstrap replicates

	MCAR					MAR					MNAR				
	Heckman			MI		Heckman			MI		Heckman			MI	
	FIML	LIML	Tobit	PMM	No PMM	FIML	LIML	Tobit	PMM	No PMM	FIML	LIML	Tobit	PMM	No PMM
Bias	12,047	-4,729	-6,677	7	-638	-4,959	-7,204	-10,728	-222	-869	-10,587	-10,878	-12,948	-4778	-5,269
rMSE	126,880	5,496	6,928	1,806	1,963	44,410	11,762	10,968	2,725	2,515	22,752	12,103	13,210	5527	5,903

rMSE: root mean square error.

Table 3. Simulation studies: Bias and rMSE according to method for estimating the mean WTP from all models with 20% missing data either MCAR, MAR and MNAR.