Two-Stage Estimation to Control for Unobservables in a Recreation Demand Model with Unvisited Sites

Richard T. Melstrom*
Department of Agricultural Economics
Agricultural Hall
Oklahoma State University
Stillwater, OK 74078

Deshamithra H. W. Jayasekera†
Department of Agricultural Economics
Agricultural Hall
Oklahoma State University
Stillwater, OK 74078

Selected Paper prepared for presentation for the 2016 Agricultural & Applied Economics Association, Boston, MA, July 31-August 2.

The authors thank Corey Jager, Kurt Kuklinski and John Polo for their assistance with the data. Financial support was provided by the Oklahoma Agricultural Experiment Station and NSF EPSCoR (Grant No. OIA-1301789).

*Assistant Professor. Email: melstrom@okstate.edu; Phone: 405-744-6171
†Graduate student. Email: jayasek@ostatemail.okstate.edu

Two-Stage Estimation to Control for Unobservables in a Recreation Demand Model with Unvisited Sites

Abstract

The role of unobserved site attributes is a growing concern in recreation demand modeling. One solution in random utility models (RUM) involves separating estimation into two stages, where the RUM model is estimated with alternative-specific constants (ASCs) in the first stage, and the estimated ASCs are regressed on the observed site attributes in the second stage. Prior work estimates the second stage with OLS and 2SLS regression. We present an application with censored regression in the second stage. We show OLS produces inconsistent parameters when there are unvisited sites with no estimable ASCs and that censored regression avoids this problem.

Keywords: Random utility model; non-market valuation; recreational fishing

JEL codes: C25; Q26; Q51

1 Introduction

There is growing recognition that unobserved site characteristics are a serious problem in random utility models (RUM) of recreation demand. Failure to control for unobservables in these models can lead to severely biased parameter and welfare estimates [1]. Similar issues have been noted in applications to differentiated consumer products, for example price endogeneity in modeling the supply and demand for automobiles [2]. Price endogeneity is also known to arise in recreation demand modeling [3]. However, the problem of unobservables in RUM models extends beyond endogeneity. Unobserved choice attributes independent of the included explanatory variables still produce biased standard errors, which effectively overstate the precision of the parameters [4]. Given the difficulty of measuring environmental quality, it may be challenging for RUM models of recreation demand to avoid the bias from unobservables.

To address problems of endogeneity and neglected heterogeneity, several papers with recreational RUM model applications use two-stage estimation [4, 5, 6]. The first stage of this procedure estimates the RUM model specified with a full set of alternative-specific constants (ASCs). The second stage regresses the estimated ASCs on the observed, alternative-specific characteristics not identified in the first stage.[1] Two-stage estimation is flexible in that a variety of estimators can be used in the second stage [4]. If the analyst is only concerned with neglected heterogeneity, then OLS will be sufficient [7, 8]. If an alternative-specific characteristic is thought to be endogenous, 2SLS can be used [5]. In general, the appropriate second-stage estimator will depend on the nature of the ASCs and the omitted variables problem.

In this paper we examine several second-stage regression models. Our motivation comes from the persistence of unvisited sites in RUM models of recreation demand—it is not unusual for several sites to receive no visits in a sample of trips. These occurrences do not preclude estimating RUM models, but in the context of two-stage estimation only the ASCs of visited sites are identified, so applying OLS in the second stage produces biased parameters. To our knowledge, only Timmins and Murdock [5] acknowledge this issue. Their solution adds small increments to the number of visits each site receives and applies a quantile estimator to the second stage to control for the fact that the ASCs for unvisited sites are arbitrarily small. In contrast, we use censored regression in the second stage. Rather than assigning arbitrarily small values to ASCs, this approach assumes the ASCs of unvisited sites are censored from below. Censored regression can be carried out in most statistical software packages, so analysts will find this procedure simple to perform. Specifically, we adopt censored Tobit regression in the second stage, which is preferrable to OLS and certain quantile regression strategies (such as median regression) because it remains a consistent estimator when there is a large number of unvisited sites in the data.

Our application is to recreational fishing in Oklahoma. One of our objectives was to derive welfare estimates for fishing in the state, which has about 150 fishable public lakes. Only secondary data were available for this task, which left about 45% of lakes with no visits from the sample. We also lacked a rich dataset on site characteristics, so neglected heterogeneity is likely to be a problem. In applying the two-stage correction procedure, the results show failing to account for the censored nature of the ASCs can lead the analyst to falsely conclude relevant measures of site quality are not important to anglers.

2 Empirical stategy

2.1 RUM model of recreational fishing

For a RUM model of recreational fishing we want to relate the demand for fishing sites to differences in site attributes such as travel cost and fish abundance. RUM models assume an individual chooses the alternative with highest utility. For an individual angler i, assume there are A alternatives, each associated with a utility level of U_{ij}, where $j = 1, \ldots, A$. The indirect utility level from choosing alternative j has the form:

$$U_{ij} = \boldsymbol{x_j}\boldsymbol{\beta} + \boldsymbol{z_{ij}}\boldsymbol{\gamma} + p_{ij}\rho + \nu_j + \epsilon_{ij} \tag{1}$$

which can be rewritten as

$$U_{ij} = \delta_j + \boldsymbol{z_{ij}}\boldsymbol{\gamma} + p_{ij}\rho + \epsilon_{ij} \tag{2}$$

where the term $\delta_j = \boldsymbol{x_j}\boldsymbol{\beta} + \nu_j$ is the component of utility that varies across alternatives but not across anglers. The vector $\boldsymbol{x_j}$ contains the observable site-specific characteristics and ν_j the unobservable site characteristics. The vector $\boldsymbol{z_{ij}}$ contains the site characteristics relevant to angler i at site j, p_{ij} is the travel cost, and ϵ_{ij} is the random part of utility. Anglers are assumed to choose the alternative j where $U_{ij} > U_{ik}$ for all $j \neq k$, although the researcher only observes the portion $V_{ij} = \boldsymbol{x_j}\boldsymbol{\beta} + \boldsymbol{z_{ij}}\boldsymbol{\gamma} + p_{ij}\rho$ and out of sample cannot predict with certainty the preferred fishing alternative for a given trip. Assuming $\nu_j = 0$ and ϵ_{ij} is distributed extreme value yields the conditional logit site choice model, where the probability of visiting site j is

$$prob_i(\text{choose } j) = \frac{e^{V_{ij}}}{\sum_{k=1}^{A} e^{V_{ik}}}. \tag{3}$$

The utility specification in equations (1)-(2) allows for preference heterogeneity in the observable characteristics. Additional heterogeneity could be incorporated by using a random parameters logit model, which allows parameters to vary among individuals.[2]

The welfare impact associated with changing site characteristics is measured by the maximum amount an angler is willing to pay (WTP) to equate the utility they would get in the altered state to the utility they get in the current state. Following Haab and McConnell [9], WTP is expressed as

$$\text{WTP}_i = \frac{1}{\rho}\left[\ln\Big(\sum_{k=1}^{A} e^{V_{ik}^1}\Big) - \ln\Big(\sum_{k=1}^{A} e^{V_{ik}^0}\Big)\right] \tag{4}$$

where V^0 denotes utility in the current state and V^1 utility in the altered state.

As in any regression model, correlation between the observables $\boldsymbol{x_j}$ and $\boldsymbol{z_{ij}}$ and the unobservable ν_j leads to endogeneity [1]. However, ν_j will cause problems in RUM models even if it is uncorrelated with the observables, by creating attenuation bias in the standard errors. This bias can be significant even when the neglected heterogeneity is slight, producing standard errors that grossly overstate the precision of the parameters [4]. The solution is to estimate the utility function in equation (1) in two stages. First, estimate equation (2) as the discrete choice model, where δ_j is a constant for each site. The resulting set of ASCs capture all site-specific heterogeneity—observable and unobservable—so the first stage produces consistent parameters and standard errors. Second, regress the ASCs on the common observable utility component $\boldsymbol{x_j}$ to estimate $\boldsymbol{\beta}$. Assuming ν_j is uncorrelated with the observables, the parameters can be consistently estimated using OLS and reported with robust standard errors in the usual fashion. An instrumental variables strategy can be applied in the second stage to correct for the more serious case of endogeneity [5].

2.2 Second stage estimation

An important consideration in estimating the ASCs is whether any of the alternatives go unchosen. This problem often arises in RUM models of recreation demand unless sample sizes are very large. For each site the ASC is estimated based on the proportion of trips it receives in the sample. If a site receives no trips, then the ASC associated with it cannot be identified (besides knowing the constant must be small, so the number of sample visits predicted by the RUM model is zero). This is a problem in our application, in which about 45% of lakes were not visited in the sample.

There are several strategies to deal with this identification problem. First, unvisited sites can be aggregated with visited sites to ensure a non-zero share of trips go to all alternatives. Site aggregation should be done with caution because doing so can result in biased parameter estimates [10]. However, Lupi and Feather [11] demonstrate a partial site aggregation strategy can potentially avoid this bias by keeping a large fraction of the most popular sites disaggregated while aggregating the remaining sites into groups. Second, the choice set can be restricted so the model is applied only to visited sites. This strategy does not affect the properties of the first-stage estimates, but it will limit inference about the effects of the common utility component $\boldsymbol{\beta}$ estimated in the second-stage. Researchers employing this approach will have to be careful interpreting the role of $\boldsymbol{\beta}$, unless sites went unvisited arbitrarily [12]. This strategy has been adopted in a few cases of two-stage estimation [4, 7]. A third strategy is to assign suitably small values to the constants of unvisited sites (enough so that the predicted share of trips to these sites is zero) and use quantile regression in the second stage [5]. We consider each of these as a possible remedy in our own application below.

We propose a fourth strategy based on the intuition of Timmins and Murdock [5]. With δ denoting the baseline utility or ASC to each site, let $\hat{\delta}$ denote the "observable," estimated ASCs and $\tilde{\delta}$ the unidentified ASCs. Given the unidentified ASCs will tend to be smaller than any $\hat{\delta}$, we propose there exists some $\underline{\delta}$ where $\tilde{\delta} \leq \underline{\delta}$. We can then say $\hat{\delta}$ is censored at $\underline{\delta}$ and write the second-stage model as

$$\hat{\delta}_j = \boldsymbol{x_j}\boldsymbol{\beta} + \nu_j \tag{5}$$

$$\hat{\delta} = \max(\underline{\delta}, \delta) \tag{6}$$

Assuming ν_j is homoscedastic normal, equations (5) and (6) compose the standard censored Tobit model [13]. It is impossible to know the true $\underline{\delta}$, but Carson and Sun [14] and others [15] demonstrate that when the censoring threshold is unknown Tobit regression yields consistent parameters if the minimum order statistic of the observed sample is used as the threshold. The strategy we propose is to use this estimator in the second stage, with the smallest estimated ASC as the censoring point, $\underline{\delta} = \min\{\hat{\delta}\}$.

It is important to provide some remarks on the difference between a Tobit model for data censoring applications and a Tobit model for corner solutions. With corner solutions, the entire distribution of the data lies at and above some lower bound, and there is no data observability problem. Consequently, when Tobit models are applied to corner solutions the partial effect is computed conditional on the observed distribution of the data, which involves weighting $\boldsymbol{\beta}$ by the probability of an interior solution. Calculating partial effects and the meaning of $\boldsymbol{\beta}$ is much simpler in applications to censored data. Because δ is not inherently bounded, the Tobit parameters can be interpreted as they are written in equation (5), just as if there had been no censoring problem and the second stage was estimated by OLS [13].

A further assumption in applying a censored estimator in recreation demand applications is that the $\hat{\delta}$ must not be substantially influenced in the first-stage estimation by the $\tilde{\delta}$ (or lack thereof). Note that this is similar to the assumption Timmins and Murdock make in their procedure, i.e. the arbitrary amount added to the total number of visits in their model to estimate the ASCs is small enough not to affect the relative odds of any two choices with a positive number of visitors.[3]

We compare the results from two-stage estimation with censored Tobit to several alternative regressions. Initially, we ignore the potential influence of unobserved site characteristics and estimate a standard conditional logit. We then restrict the choice set to visited sites and estimate a conditional logit with ASCs. This first stage identifies the effects of the heterogeneous utility components and the ASCs of visited sites. OLS regression is then applied to the estimated ASCs in order to identify $\boldsymbol{\beta}$. Next, we adopt a partial site aggregation strategy in which the lakes with ≥ 2 visits from the sample are left disaggregated and the remaining lakes are aggregated into a few alternatives, each with at least one visit. We estimate this model with ASCs, and then OLS in the second stage to identify $\boldsymbol{\beta}$. Finally, we assign small values to the constants of unvisited sites and use quantile (median) regression in the second stage.

The five regressions examined can be summarized as:

- Case 1: Estimate equation (1) as a conditional logit.
- Case 2: Estimate equation (2) as a conditional logit. Estimate $\boldsymbol{\beta}$ by OLS from only the set of chosen alternatives.
- Case 3: Estimate equation (2) as a conditional logit after partial site aggregation. Estimate $\boldsymbol{\beta}$ by OLS from the set of partially aggregated alternatives.
- Case 4: Estimate equation (2) as a conditional logit. Estimate $\boldsymbol{\beta}$ by median regression from the complete choice set by assigning values $\tilde{\delta}$=ln$(0.001/A)$ to the ASCs of unvisited sites.
- Case 5: Estimate equation (2) as a conditional logit. Estimate $\boldsymbol{\beta}$ by censored Tobit regression from the complete choice set by assuming $\tilde{\delta} \leq \min\{\hat{\delta}\}$.

The degree of bias from discarding the unvisited sites in the second stage is assessed by comparing case (2) with case (5). Case (3) will be useful in determining whether partial site aggregation with ASCs can correct for unobserved heterogeneity when there are unvisited sites (in addition to estimating choice models with a large number of alternatives, for which partial site aggregation was originally intended) [11]. Case (4) mirrors an existing two-stage RUM estimation strategy to deal with unvisited sites [5].

2.3 Additional details

Along with the site characteristics that vary across individuals, we estimate the ASCs in the RUM model by maximum likelihood. All regressions are performed in Stata 13 [16]. Our partial site aggregation strategy is based on Parsons et al. [12] in that the least popular lakes (with < 2 visits) are aggregated into one of eight regional groups, which are then described by the average lake data among the group [11].[4] In the results, we report heteroscedasticity-robust standard errors for the OLS and quantile-estimated parameters.

3 Data

The fishing trip data were obtained from a survey conducted by the Oklahoma Department of Wildlife Conservation in 2014 of 3,000 randomly-selected resident fishing license holders. The department conducts the survey approximately every five years to collect data about angler opinions, attitudes and preferences. The questions in the 2014 survey were not designed for the purpose of this study. Individuals were initially contacted with a mailed packet containing a letter, a questionnaire and a pre-stamped return envelope. The letter described the purpose of the survey and informed anglers they could participate by returning the enclosed paper questionnaire, filling out a web-based questionnaire, or waiting several weeks to complete a follow-up phone survey. The response rate was 26%.

Among other items, the survey asked anglers to indicate their species preferences and recent fishing trip location. Removing those who did not fish in the past year (17% of respondents), did not report an identifiable destination (2% of respondents) or whose trip was not taken primarily for the purpose of fishing (14% of respondents) left 536 trips suitable to estimate the RUM model.

We developed a list of 146 lakes based on an index maintained by the Oklahoma Water Resources Board as well as the destination information provided by anglers in the survey. To account for potential substitution between fishing at lakes and other waterbodies [17], we added a two aggregated alternatives, one each for stream and pond sites. The complete choice set in our model therefore includes 148 alternatives.

Travel costs to each lake were calculated using information on travel distances, angler demographics and gasoline prices. Travel distances from an angler's home zip code to each lake were calculated using the PC*Miler program [18]. The opportunity cost of travel time was constructed from a wage proxy. This wage proxy was one-third the midpoint of an angler's income category from one of six possible categories on the questionnaire (or the observed mean for anglers who omitted a response) divided by 2000 (the approximate number of working hours in a year). We used a per-mile driving cost of $0.28, based on the marginal change in driving costs for a large sedan in AAA reports [19] that discounts depreciation costs [20]. The travel cost to each site was calculated as the sum of round-trip distance in miles times per mile driving costs, plus the opportunity cost of travel time assuming an average driving speed of 45 miles per hour, plus the access fee, if any.

Fishing quality at lakes was measured in terms of expected fish abundance. The wildlife department provided fish sampling data for select species and lakes. We focused on the abundance of black bass (largemouth and smallmouth bass), walleye, catfish and crappie, as these are considered important game fish in Oklahoma. Catfish, crappie and black bass are the most commonly targeted, with 65%, 59% and 55% of anglers reporting catfish, crappie and black bass, respectively, as one of their top species (in the survey, anglers ranked the three species they most preferred to catch in the past year). Walleye are less popular but are still considered a key game fish in the state. Black bass abundance is measured by spring electrofishing counts while walleye, catfish and crappie abundance are measured by fall gillnetting counts. We expect these measures are highly correlated with catch rates. Expected abundance was calculated at individual lakes using an exponential regression model with covariates for water quality and landscape features.[5] The mean of expected abundance was used to fill in for lakes that lacked data. The abundance measures were then interacted with dummy variables for whether the angler reported a preference for the species in question. The interactions are interpretable as targeted abundance, and allow us to focus on the importance abundance with species preferences held fixed [21].

We also collected data on water quality and shoreline conditions. To measure water quality, Secchi disk depth was obtained from the state Water Resources Board.[6] Conditional on fish abundance, we expect anglers avoid lakes with high turbidity [22]. Shoreline length (in miles) was calculated from a GIS database. Following Train [23], we expect shoreline length to account for the fact that the number of fishable locations increases with the size of a lake. A dummy variable indicates whether a lake is associated with a Close to Home fishing agreement between the state wildlife department and municipalities to improve fishing opportunities and facilities at metro lakes.[7] Finally, the number of boat ramps at a site is included as a proxy for the number of access points at a lake [23]. To distinguish the effect of opening up access with one boat ramp and reducing congestion with additional ramps, we also include a dummy variable for whether there is at least one ramp at a lake. Table 1 provides descriptive statistics of the site characteristics used in the demand model.

4 Results and discussion

We report the RUM model estimates in two parts. In the first part, we focus on the effects of the individual heterogeneous utility components (varying across sites and anglers), that can be estimated in the first stage with ASCs. In the second part, we report the effects of the common utility components (the site-specific characteristics), which must be estimated in the second stage when there are ASCs. Following this, we demonstrate the implications of these different models for welfare analysis with several hypothetical valuation scenarios.

4.1 Effects varying across anglers estimated in the first stage

In this section we focus on the RUM estimates for *travelcost* and the nine interaction variables that vary across individuals and sites. Table 2 shows the results for the different RUM models. Note that the two-stage models estimated from the complete/dissaggregated choice set share the same first-stage, so their estimates at this point are identical (and, hence, not separately reported).

The models at this stage are largely similar in parameter signs and magnitudes. All models agree that smaller travel costs and greater catch rates positively affect utility.[8] For anglers who target black bass or walleye, the effect of fish abundance on site choice is statistically significant. For boat users, there is a clear preference for lakes with a boat ramp, and more ramps are preferred to fewer ramps. Bass anglers disproportionately prefer to fish at ponds, while trout anglers prefer rivers.

Only a few parameters are affected by the addition of ASCs, which improve the overall fit of the model significantly. A likelihood ratio test of the hypothesis that all characteristics are observed is rejected at the 5% level.[9] Most notably, the effect of *walleye* declines from 1.220 without ASCs to 0.758 with ASCs (or 0.790 in the case of partial site aggregation), suggesting that there may be unobserved lake characteristics correlated with walleye abundance which also influence site choice. The effects of $catfish$ and *crappie* increase with the ASCs, but remain statistically indistinguishable from zero. Partial site aggregation does an admirable job in this stage; with about one-third the number of observations, the estimates between the full choice set and the partially aggregated choice set are nearly identical. Importantly, the standard errors are all larger when the model includes ASCs. This result is consistent with the work of Murdock [4], who shows the standard errors of a conditional logit are biased downward when there are unobserved site characteristics.

4.2 Site-specific effects estimated in the second stage

Differences between estimation strategies are more apparent when comparing the site-specific effects. These results are reported in Table 3. All models predict that anglers prefer lakes with longer shorelines and are part of a Close to Home agreement, and that a significant share of anglers fish at rivers and ponds.[10] Nevertheless, between models there are important differences in the magnitudes of these effects and little agreement on the influence of Secchi depth.

Compare the RUM model that uses a standard conditional logit with the model that applies a second-stage OLS estimator on the identified ASCs. Two-stage OLS estimation produces smaller parameters and larger standard errors relative to the standard conditional logit. The standard errors in the latter model are biased downwards, so larger standard errors in the second stage is not suprising. It is mostly due to attenuation in the parameter estimates, though, that the *Secchidepth* effect is no longer statistically significant, and the $CloseToHome$ effect drops in significance from the 5% to the 10% level. The $shoreline$ parameter remains positive and highly significant but is about half as large as its counterpart in the standard conditional logit.

Now consider the parameters in the model developed with partial site aggregation. The effect of $Secchidepth$ is negative and statistically insignificant. The $Shoreline$ and $ClosetoHome$ effects are both smaller compared to the parameters in standard conditional logit, although still positive and statistically significant at the 5% level. The similarities between these parameters and those in the two-stage model estimated on visited sites is probably not a coincidence. In general, the impact of aggregating sites on the parameters is uncertain, but our strategy of aggregating the least-visited sites appears to have the same effect as removing the sites with no visits from the choice set. Actually, partial site aggregation may be worse in some sense, because it assumes the lake data used for the aggregated alternatives is fixed rather than averaged, which causes the standard errors to be biased downward.

Applying median regression in the second stage produces curious estimates. All of the parameters are at least several times larger than their counterparts estimated from the other regressions. Why is this? Recall that median regression is necessay because arbitrarily small values were assigned to the unidentified ASCs, $\tilde{\delta}$. Previous research states that median regression can consistently estimate site-specific effects in the presence of these assumed values as long as a majority of sites have a positive number of visitors [5], but our results show this is not true in general. In our application, 70 of 148 sites have zero visits, which is enough that the parameters are sensitive to the choice of $\tilde{\delta}$, and in Appendix A we demonstrate that the parameters change substantially with different assumptions about $\tilde{\delta}$. However, quantile regression can still consistently estimate the effects in the upper portions of the distribution of δ (i.e. lakes with above-average visitation); Appendix A provides a quantile regression example in which the parameters are robust the values of $\tilde{\delta}$ when estimated at the third quartile.

Finally, consider the RUM model that uses censored Tobit regression in the second stage. This strategy accommodates the complete choice set by assuming the unidentified ASCs are censored at the value of the smallest estimable constant. Overall, the censored Tobit regression parameters are similar to those from the standard conditional logit and appear much more reliable, with positive and significant parameters estimated for *shoreline*, *Secchidepth* and $ClosetoHome$. The standard errors are also larger compared to the standard conditional logit. Interestingly, censored Tobit regression fits the ASC data better than any of the other regressions.

An important caveat is that model misspecification in the second stage will affect the results. However, in our application none of the censored Tobit model assumptions are violated. In Appendix B, we show that the second-stage parameters from the Tobit are nearly identical to those obtained from censored least absolute deviations (CLAD) regression, suggesting the Tobit estimates are largely robust to the distributional assumptions of that estimator.[11] More importantly, we conducted a test of the normality and homoscedasticity assumptions by nesting the Tobit within a more general, Box-Cox specification, which could not be rejected at the 5% level.[12] Thus, the usual Tobit model assumptions are satisfied.

Although it may not be obvious, model misspecification is a problem in the model that applies OLS in the second stage using the identified ASCs. In general, with data censoring OLS is an inconsistent estimator [13], and in our application we cannot apply OLS to the subsample of visited sites and expect to consistently estimate the site-specific effects. Applying the model to only visited sites introduces a new source of error inherently correlated with the site-specific observables (since these variables affect visitation and therefore the probability of a site going unvisited), creating a source of endogeneity. This induced endogeneity explains the differences observed between the OLS and Tobit estimates.

Overall, we find that assuming the choice set includes only visited sites leads to selection bias. This is consistent with Parsons et al. [12], who found RUM model parameters were sensitive to excluding less popular lakes in the choice set. Parameter sensitivity to choice set definitions is also consistent with the results of Hicks and Strand [24]. Had we estimated the model on the restricted choice set to accommodate OLS estimation in the second stage, we would have falsely concluded turbidity and perhaps the Close to Home fishing program did not matter to anglers. Of course, some of the lakes without visits in the present model may be unknown to anglers, in which case excluding them from the choice set is appropriate [25], but we lack the data to determine which lakes anglers are aware of and which they are not. We expect the amount of bias from including unknown lakes is less than from excluding known but still unvisited lakes.

4.3 Welfare estimates

We consider five policy experiments involving elimination of the Close to Home fishing program, improvements in water clarity, and increases in fish populations at a popular lake in Oklahoma. We compare the per-trip benefits of increases in Secchi depth at lakes across the state versus those in the Grand River watershed because state conservation organizations are currently working to implement best management practices in that watershed. Table 4 presents the welfare effects of these site quality changes. Our preferred set of results comes from the RUM model that uses censored Tobit regression in the second stage—first, because two-stage estimation greatly improves the fit of the model and, second, because Tobit regression correctly uses information on the full choice set in the second stage. We find that for anglers the value of improving water clarity in the Grand River watershed is worth more per trip, on average, than at other lakes in the state. This is due to the relatively low Secchi depth levels and greater demand for fishing in the Grand River watershed compared to other regions. In terms of per trip WTP for increases in fish abundance, our estimates are similar to those found for other freshwater sport fisheries [21, 26, 27].

As one would expect based on the demand model results, two-stage estimation that omits unvisited sites undervalues the Close to Home fishing program and improvements in Secchi depth. Our preferred model predicts anglers have a WTP of \$0.40 per trip to maintain the Close to Home fishing program and \$7 per trip for a 50% increase ($\approx$ 1 foot) in water clarity, holding fishing quality fixed, while two-stage estimation applied to the ASCs of visited sites implies a WTP of \$0.16 per trip for the Close to Home fishing program and about \$2 per trip for the same increase in water clarity. Thus, the value of the Close to Home fishing program is about 2.5 times greater when one uses information on the full choice set versus naïve two-stage estimation. With 7,449,000 fishing trips taken in Oklahoma by residents in 2011 [28], our preferred model implies an annual benefit of \$3 million from this program.

The welfare estimates from median regression in the second stage are clearly biased. This model predicts essentially no welfare gains from improvements in fishing quality but huge gains from reductions in Secchi depth. This is because the unrealistically large site-specific effects estimated in the second stage dominate the contribution of the first stage parameters in the RUM model. Valid second stage estimates can be obtained by estimating the effects in the upper quantiles, although there is no guarantee that these estimates will reflect the relationship between sites at the center of the demand distribution and site characteristics.[13]

Finally, we note that there is also an economically meaningful difference in the WTP associated with improvements in fish abundance between the models, particularly for walleye. WTP is several times higher when one uses a standard conditional logit rather than two-stage estimation. This suggests that controlling for endogenous site characteristics may be important in our application.

5 Conclusion

Two-stage estimation is an innovative and flexible technique to control for unobserved choice attributes in RUM models. Prior recreation demand applications have used OLS, 2SLS and median regression in the second stage. To these we add censored regression. Using a censored estimator for site-specific effects was important in our application, which contained a large number of sites with no visits that precluded estimating a complete set of ASCs. In general, though, the appropriate second-stage regression depends on the nature of the ASCs, and the regressions mentioned here (OLS, Tobit, etc.) may not be suitable for every application.

In using two-stage estimation, a researcher may be tempted to restrict the chioce set to only the chosen alternatives in order to apply OLS in the second stage.

With ASCs, this restriction does not affect the first-stage parameters. However, the second-stage regression is then performed on a misspecified choice set, leading to incorrect parameter and welfare estimates. We found this to be true in our application to recreational fishing in Oklahoma. The technique proposed in this paper allows the researcher to use the original choice set in the presence of unvisited sites. Depending on the percentage of visited sites in the sample, the second-stage quantile regression strategy proposed by Timmins and Murdock [5] may also be valid.

Dropping unvisited sites from RUM models of recreation demand should be done cautiously. Doing so discards potentially useful information about influential site characteristics, and can produce misleading parameter estimates. This does not mean that a choice set of only visited sites is uninformative. There are many applications were all relevant alternatives are chosen at least once in the sample. In general, proper choice set definition is an unresolved issue in recreation demand modeling [29, 30]. Depending on the application, including unvisited sites may add only a small amount of information about the influence of observable site characteristics and be unnecessary.

References

[1] Klaus Moeltner and Roger Von Haefen. Microeconometric strategies for dealing with unobservables and endogenous variables in recreation demand models. *Annual Review of Resource Economics*, 3(1):375–396, 2011.

[2] Steven Berry, James Levinsohn, and Ariel Pakes. Automobile prices in market equilibrium. *Econometrica*, 63(4):841–890, 1995.

[3] George R. Parsons. A note on choice of residential location in travel cost demand models. *Land Economics*, 67(3):360–364, 1991.

[4] Jennifer Murdock. Handling unobserved site characteristics in random utility models of recreation demand. *Journal of Environmental Economics and Management*, 51(1):1–25, 2006.

[5] Christopher Timmins and Jennifer Murdock. A revealed preference approach to the measurement of congestion in travel cost models. *Journal of Environmental Economics and management*, 53(2):230–249, 2007.

[6] Babatunde O. Abidoye, Joseph A. Herriges, and Justin L. Tobias. Controlling for observed and unobserved site characteristics in rum models of recreation demand. *American Journal of Agricultural Economics*, 24(2):227–301, 2012.

[7] Richard T. Melstrom, Frank Lupi, Peter C. Esselman, and R. Jan Stevenson. Valuing recreational fishing quality at rivers and streams. *Water Resources Research*, 51(1):140–150, 2015.

[8] Matthew A. Weber, Pallab Mozumder, and Robert P. Berrens. Accounting for unobserved time-varying quality in recreation demand: An application to a sonoran desert wilderness. *Water Resources Research*, 48(5), 2012.

[9] Timothy C. Haab and Kenneth E. McConnell. *Valuing Environmental and Natural Resources: The Econometrics of Non-Market Valuation*. Edward Elgar Publishing, 2002.

[10] George R. Parsons and Michael S. Needelman. Site aggregation in a random utility model of recreation. *Land Economics*, 68(4):418–433, 1992.

[11] Frank Lupi and Peter M. Feather. Using partial site aggregation to reduce bias in random utility travel cost models. *Water Resources Research*, 34(12):3595–3603, 1998.

[12] George R. Parsons, Andrew J. Plantinga, and Kevin J. Boyle. Narrow choice sets in a random utility model of recreation demand. *Land Economics*, 76(1):86–99, 2000.

[13] Jeffrey M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. MIT press, 2002.

[14] Richard T. Carson and Yixiao Sun. The tobit model with a non-zero threshold. *The Econometrics Journal*, 10(3):488–502, 2007.

[15] Thomas W. Zuehlke. Estimation of a tobit model with unknown censoring threshold. *Applied Economics*, 35(10):1163–1169, 2003.

[16] StataCorp Stata. Release 13, 2013. Statistical software. College Station. Texas, USA: StataCorp LP.

[17] Carol A. Jones and Frank Lupi. The effect of modeling substitute activities on recreational benefit estimates. *Marine Resource Economics*, 14(4):357–374, 1999.

[18] PC Miler, Version 27. ALK Technologies, Inc, Princeton, N.J.

[19] AAA (American Automobile Association). Your driving costs: How much are you really paying to drive, 2014. AAA Association Communication, Heathrow, Florida.

[20] Derrick Hang, Daniel McFadden, Kenneth Train, and Ken Wise. Is vehicle depreciation a component of marginal travel cost? A literature review and empirical analysis. *Journal of Transport Economics and Policy*, 50(2), 2016.

[21] David Scrogin, Kevin Boyle, George Parsons, and Andrew J. Plantinga. Effects of regulations on expected catch, expected harvest, and site choice of recreational anglers. *American Journal of Agricultural Economics*, 86(4):963–974, 2004.

[22] Kevin J. Egan, Joseph A. Herriges, Catherine L. Kling, and John A. Downing. Valuing water quality as a function of water quality measures. *American Journal of Agricultural Economics*, 91(1):106–123, 2009.

[23] Kenneth E. Train. Recreation demand models with taste differences over people. *Land Economics*, 74(2):230–239, 1998.

[24] Robert L. Hicks and Ivar E. Strand. The extent of information: Its relevance for random utility models. *Land Economics*, 76(3):374–385, 2000.

[25] Timothy C. Haab and Robert L. Hicks. Accounting for choice set endogeneity in random utility models of recreation demand. *Journal of Environmental Economics and Management*, 34(2):127–147, 1997.

[26] Len M. Hunt, Peter C. Boxall, and Barry Boots. Accommodating complex substitution patterns in a random utility model of recreational fishing. *Marine Resource Economics*, 22(2):155–172, 2007.

[27] Richard T. Melstrom and Frank Lupi. Valuing recreational fishing in the Great Lakes. *North American Journal of Fisheries Management*, 33(6):1184–1193, 2013.

[28] U.S. Fish, Wildlife Service, and U.S. Census Bureau. *2011 National Survey of Fishing, Hunting, and Wildlife-associated Recreation*. US Department of the Interior, Fish and Wildlife Service., 2014.

[29] Timothy C. Haab and Robert L. Hicks. Choice set considerations in models of recreation demand: History and current state of the art. *Marine Resource Economics*, 14(4):271–281, 1999.

[30] Daniel J. Phaneuf and V. Kerry Smith. Recreation demand models. In K.G. Maler and J.R. Vincent, editors, *Handbook of Environmental Economics*, pages 671–761. North Holland Publishing, 2005.

[31] Victor Chernozhukov, Ivan Fernandez-Val, Sukjin Han, and Amanda Kowalski. CQIV: Stata module to perform censored quantile instrumental variables regression. 2012. Statistical Software Components.

[32] Roger H. Von Haefen and Daniel J. Phaneuf. Identifying demand parameters in the presence of unobservables: A combined revealed and stated preference approach. *Journal of Environmental Economics and Management*, 56(1):19–32, 2008.

[33] Katherine M. Harris and Michael P. Keane. A model of health plan choice:: Inferring preferences and perceptions from a combination of revealed preference and attitudinal data. *Journal of Econometrics*, 89(1):131–157, 1998.

[34] H. Allen Klaiber and Roger H. von Haefen. Incorporating random coefficients and alternative specific constants into discrete choice models: Implications for in-sample fit and welfare estimates, 2008. Working paper.

[35] Christian Gourieroux, Alain Monfort, and Alain Trognon. Pseudo maximum likelihood methods: Applications to poisson models. *Econometrica*, pages 701–720, 1984.

Table 1: Site characteristics used in the recreational fishing RUM models

Characteristic	Description	Mean	St. Dev.
$Shoreline$	Natural log of shoreline length in miles	2.478	1.473
$Secchidepth$	Natural log of Secchi disk depth in centimeters	3.972	0.833
$ClosetoHome$	Dummy = 1 if lake is part of the state's Close to Home fishing program	0.020	0.141
$Ramp_{dum}$	Dummy = 1 if there is at least one boat ramp at a lake	0.932	0.252
$Ramp_{num}$	Number of boat ramps at lake	3.236	7.231
$Blackbass$	Predicted abundance based on fish counts from overnight electrofishing surveys (reported as average fish per attempt)	67.595	31.543
$Walleye$	Predicted abundance based on fish counts from overnight gill-netting surveys (reported as average fish per net)	0.143	0.286
$Catfish$	Predicted abundance based on fish counts from overnight gill-netting surveys (reported as average fish per net)	2.894	2.475
$Crappie$	Predicted abundance based on fish counts from overnight gill-netting surveys (reported as average fish per net)	1.784	1.501
$Pond$	Dummy = 1 for pond aggregated alternative	0.007	0.082
$River$	Dummy = 1 for river aggregated alternative	0.007	0.082

Table 2: Results of the recreational fishing RUM models – angler-varying effects

Variables	Standard CL	CL with ASCs Full choice set	CL with ASCs Partial aggregation
$Travelcost$	-0.019**	-0.020**	-0.019**
	(0.001)	(0.001)	(0.001)
$Ramp_{dum} \times boat\ user$	1.733**	1.736**	1.808**
	(0.263)	(0.265)	(0.270)
$Ramp_{num} \times boat\ user$	0.016**	0.017**	0.016**
	(0.004)	(0.006)	(0.005)
$Blackbass \times bass\ angler$	0.005**	0.007**	0.006**
	(0.002)	(0.002)	(0.002)
$Walleye \times walleye\ angler$	1.220**	0.758**	0.790**
	(0.270)	(0.377)	(0.377)
$Catfish \times catfish\ angler$	0.000	0.027	0.060
	(0.027)	(0.047)	(0.054)
$Crappie \times crappie\ angler$	0.041	0.105	0.129
	(0.055)	(0.076)	(0.081)
$River \times trout\ angler$	1.741**	1.641**	1.611**
	(0.356)	(0.361)	(0.362)
$Pond \times bass\ angler$	2.028**	2.093**	2.091**
	(0.334)	(0.341)	(0.342)
$Pond \times catfish\ angler$	-0.240	-0.152	-0.107
	(0.273)	(0.283)	(0.287)
ASCs	No	Yes	Yes
Observations	79328	79328	28944
Log-likelihood	-1492.102	-1378.340	-1337.671

Standard errors in parentheses below parameters. *Significant at the 10% level. **Significant at the 5% level.

Table 3: Results of the recreational fishing RUM models – site-specific effects

Variables	Standard CL	2SCL – OLS Visited sites only	2SCL – OLS Partial aggregation	2SCL – Median regression	2SCL – Tobit regression
Shoreline	0.945**	0.525**	0.524**	4.145**	0.985**
	(0.042)	(0.077)	(0.065)	(0.291)	(0.078)
Secchidepth	0.208*	0.128	-0.062	1.496**	0.522*
	(0.085)	(0.156)	(0.085)	(0.684)	(0.147)
ClosetoHome	2.339**	0.801*	1.436**	18.072**	2.999**
	(0.612)	(0.421)	(0.150)	(2.095)	(0.355)
River	5.299**	3.344**	2.286**	26.553**	7.236**
	(0.477)	(0.742)	(0.331)	(4.412)	(0.670)
Pond	4.770**	2.721**	1.629**	25.930**	6.613**
	(0.564)	(0.742)	(0.331)	(4.412)	(0.670)
Observations in second stage	-	78	54	148	148
R^2 in second stage	-	0.480	0.571	0.433	0.611

Standard errors in parentheses below parameters. *Significant at the 10% level. **Significant at the 5% level. The R2 is calculated as the squared correlation coefficient between the observed and predicted values of the dependent variable.

Table 4: Mean per trip WTP ($) for changes in selected site characteristics

Scenario	Standard CL	2SCL – OLS Visited sites only	2SCL – OLS Partial aggregation	2SCL – Median regression	2SCL – Tobit regression
Eliminate Close to Home program[a]	-0.26 (-0.65, 0.02)	-0.16 (-0.87, 0.00)	-0.71 (-1.59, -0.22)	-0.23 (-1.18, 0.00)	-0.40 (-0.94, 0.00)
50% increase in Secchi depth at all lakes[a]	3.07 (1.14, 5.77)	1.90 (-1.61, 5.46)	-0.90 (-3.88, 1.18)	29.26 (15.34, 50.12)	7.25 (4.67, 14.46)
50% increase in Secchi depth in Grand River watershed[b]	3.91 (1.29, 7.56)	1.84 (-1.47, 5.42)	-1.20 (-5.19, 1.67)	46.91 (22.19, 98.56)	11.40 (7.26, 27.69)
50% increase in black bass at Canton Lake[b]	0.78 (0.23, 1.51)	1.09 (0.29, 2.77)	2.16 (0.45, 4.55)	0.00 (0.00, 0.00)	0.55 (0.07, 1.08)
50% increase in walleye at Canton Lake[b]	4.81 (0.97, 8.91)	2.04 (-0.03, 9.22)	3.98 (0.06, 14.57)	0.00 (0.00, 0.01)	1.24 (-0.01, 5.27)

95% confidence intervals in parentheses below parameters. Confidence intervals are from bootstrapping with 200 resamples. [a]WTP is denominated per trip for all trips. [b]WTP is denominated per trip for trips to the affected lake.

Appendix A

Median regression estimates under alternative values about the unidentified ASCs are shown Table A1. Third-quartile regression estimates under alternative values about the unidentified ASCs are shown Table A2.

Table A1. Site-specific effects estimated from median regression under alternative values of $\tilde{\delta}$.

Variables	$\tilde{\delta} = ln(10^{-1}/536)$	$\tilde{\delta} = ln(10^{-3}/536)$	$\tilde{\delta} = ln(10^{-5}/536)$
Shoreline	2.949	4.145	5.311
	(0.206)	(0.291)	(0.425)
Secchidepth	1.079	1.496	2.159
	(0.469)	(0.684)	(1.050)
ClosetoHome	12.241	18.072	23.987
	(1.401)	(2.095)	(2.927)
River	19.042	26.553	35.141
	(2.987)	(4.412)	(6.394)
Pond	18.418	25.930	34.518
	(2.987)	(4.412)	(6.394)

Standard errors in parentheses below parameters.

Table A2. Site-specific effects estimated from third-quartile regression under alternative values of $\tilde{\delta}$.

Variables	$\tilde{\delta} = ln(10^{-1}/536)$	$\tilde{\delta} = ln(10^{-3}/536)$	$\tilde{\delta} = ln(10^{-5}/536)$
Shoreline	1.075	1.075	1.075
	(0.355)	(0.355)	(0.355)
Secchidepth	0.563	0.563	0.563
	(0.703)	(0.703)	(0.703)
ClosetoHome	3.184	3.184	3.184
	(1.715)	(1.715)	(1.715)
River	7.134	7.134	7.134
	(3.800)	(3.800)	(3.800)
Pond	6.510	6.510	6.510
	(7.134)	(7.134)	(7.134)

Standard errors in parentheses below parameters.

Appendix B

The results of the second-stage censored least absolute deviations (CLAD) estimator are reported in Table B1. CLAD regression is generally considered more robust to distributional misspecification than the Tobit, which relies on the homoscedastic normal assumption. Estimation was implemented in Stata using the `cqiv` program [31].

Table B1. Site-specific effects estimated from CLAD and censored Tobit regressions.

Variables	CLAD	Tobit
Shoreline	1.042	0.985
	(0.119)	(0.078)
Secchidepth	0.416	0.522
	(0.136)	(0.147)
ClosetoHome	3.054	2.999
	(1.704)	(0.355)
River	7.062	7.236
	(0.768)	(0.670)
Pond	6.438	6.613
	(0.785)	(0.670)

Standard errors in parentheses below parameters.

Notes

[1]Von Haefen and Phaneuf [32] demonstrate how to use stated preference data to avoid the identification problem brought on by the ASCs, circumventing the second regression.

[2]We do not report estimates from a random parameters logit for two reasons. First, applications of random parameter logits typically involve thousands of observations with several observations per individual, while in our data there are 536 trips with only one trip per angler. Random parameters have been shown to be poorly identified when estimated with limited cross section data [33], a fact born out in our trial applications. Second, as noted by Klaiber and von Haefen [34], the ASCs from random parameter logits may poorly reflect in-sample visitation patterns when the model is misspecified (e.g. due to misspecification of the parameter distributions). In contrast, by including ASCs the conditional logit will predict visitation patterns that match the data perfectly. This is because the conditional logit is part of the linear exponential family of distributions, which guarantees consistent parameter estimates if the conditional mean is correctly specificied, regardless of higher order misspecification [35].

[3]This point is discussed in the first part of section 5 in their paper [5].

[4]These regions correspond to Oklahoma's designated fisheries management zones.

[5]The abundance of species s at site j was modeled as $A_{sj} = exp(q_j \alpha_s + \gamma_{sg}) + \omega_{sj}$ where q_j is a vector of lake characteristics, γ_{sj} is a vector of watershed-specific dummies to control for systematic differences in regional abundance and ω_{sj} is the error. We found the models for black bass and walleye predicted relatively well—with R^2s of 0.60 and 0.54, respectively—compared to catfish and crappie—which had R^2s of 0.38 and 0.36, respectively.

[6]The Oklahoma Water Resources Board rotates water quality sampling at lakes. When a lake is sampled, water quality measurements, including Secchi disk depth, are taken at stratified dates and locations. We used the lakewide average for the most recent available year. The observed group mean was used for a dozen lakes which lacked Secchi disk data.

[7]The Close to Home program was set up in 2002 through cooperative agreements between the Oklahoma Department of Wildlife Conservation and Oklahoma municipalities to manage small, urban lakes for recreational fishing. Close to Home lakes support a variety of warm water species and are regularly stocked with channel catfish, and in the winter some are stocked with trout. Bag limits at these sites are more restrictive than regulations statewide.

[8]Technically, parameter estimates are not comparable across logit models, so it is common practice to compare the implied marginal rates of substitution. However, given the similarity in the travel cost parameters between models, we opt not to report marginal rates of substitution because dividing the other effects by essentially equivalent travel cost parameters suggests the two sets of estimates are comparable. In any case, later in this section we compute the welfare effects of site quality changes, which converts the results into comparable units.

[9]The χ^2 statistic for this test is 228, compared to a critical value of 175 when there are 146

degrees of freedom.

[10]Although the ODWC does not restrict the types of municipal lakes that can enroll, self-selection means the Close to Home fishing program may be endogenous. The most likely scenario is that municipalities enroll their less popular lakes to boost use, which would bias the Close to Home effect downward. Including a dummy variable for municipal lakes does not affect the magnitude and significance level of the parameter reported here.

[11]This similarity suggests at a minimum that the distribution of the data is symmetric, which is of course required for normality.

[12]The LM statistic for this test was 0.882, compared to a critical value of 6.483.

[13]Nevertheless, in our application the parameter estimates from censored Tobit and third-quartile regressions are similar. Compare Tables A2 and B1 in the appendices.