CER-ETH – Center of Economic Research at ETH Zurich

Handling excess zeros in count models for recreation demand analysis without apology

A. Martinez-Cruz

Working Paper 16/253
August 2016

Economics Working Paper Series

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Handling excess zeros in count models for recreation demand analysis without apology

Adan L. Martinez-Cruz*

July 21, 2016

Abstract

Hurdle and zero-inflated models are the two foremost methods used to deal with excess zeros. However, their reliance on the non-participation assumption is a drawback when applied to recreation demand analysis because there is not a theoretical framework convincingly explaining presence of non-participants. This paper discusses how latent class count models represent a theoretically consistent alternative to handle excess zeros. The theoretical model behind a latent class model does not require the non-participation assumption. Instead, excess zeros is explained as the accumulation of corner solutions from individuals belonging to different classes.

1 Introduction

Excess zeros is an empirical regularity in recreation demand analysis. Because standard count econometric models —i.e. Poisson and negative binomial models— can not adequately handle excess zeros, hurdle and zero-inflated models are widespread used instead. These models assume the existence of non-participants. In this way, a standard count model becomes a two-part model[1] in which behavior can be characterized by two data-generating processes: the participation decision and, conditional on participation, the trip demand decision. Accordingly, excess zeros is consequence of

*Senior researcher, Centre for Energy Policy and Economics (CEPE), ETH-Zurich, madan@ethz.ch

[1]This term is borrowed from health economics literature [e.g. Deb and Trivedi (2002)]

a large portion of the population deciding self-exclusion from the recreational market under study.

While the argument behind two-part models is intuitive, since their very seminal applications in recreation demand analysis researchers have apologized for the absence of a formal economic theory argument explaining the presence of non-participants (Gurmu and Trivedi, 1996) and the resulting arbitrariness in the specification of the participation and trip demand equations (Haab and McConnell, 1996). This arbitrariness in distinguishing non-participants from participants ultimately affect public policy recommendations because non-participants are excluded from welfare calculations (Phaneuf and Smith, 2005).

In this paper, latent class count models (LCCM) are shown to be a theoretically supported alternative to handle excess zeros. The theoretical framework supporting LCCM assumes individuals belong to exclusive groups. Groups are associated to an ordinal classification of preferences for trips — an empirically testable assumption. Conditional on membership, individuals maximize their utility function with respect to number of visits to the site under study. Individuals with corner solutions are potentially observed in every class. Large proportion of zeros can be interpreted as the accumulation of corner solutions from individuals belonging to different classes. In constrast to two-part models, LCCM do not require presence of non-participants to accommodate excess zeros, and therefore welfare estimations account for everyone in the sample, avoiding arbitrary exclusions.

Applications of LCCM to recreation demand analysis are very recent and have not yet discussed how LCCM represent a theoretically consistent alternative to two-part models when handling excess zeros.[2] English (2008) has proposed a strategy that explicitly incorporates nonparticipation into a behavioral model in the context of nested logit applications. While nested logit specifications have been used to model number of trips [e.g. (von Haefen et al., 2005)], implementation of these specifications requires researchers arbitrarily specify the number of choice occasions the individual faces during the period under study. Specification of choice ocassions raises conceptual and practical difficulties when estimating welfare (von Haefen and Phaneuf, 2003a). In constrast to those difficulties, welfare estimates from a LCCM are calculated simply as a weighted sum of groups' welfare.

Two features from the empirical illustration presented in this paper are

[2]See Evans and Herriges (2010); Scarpa et al. (2007); Thiene and Morey (2010).

underlined. First, the latent class specification outperforms two-part models in replicating empirical frequencies. Second, two-part and latent class specifications yield similar welfare estimates.

2 Optimization framework

2.1 Homogeneous individuals

The optimization framework for the homogeneous case was first presented by Hellerstein and Mendelsohn (1993). Assume individuals choose number of trips, T, and a vector of other goods, $\mathbf{C}$. $T \in I$ for $I = 0, 1, ...$; and $\mathbf{C}$ is a vector of goods that can be acquired in continuous quantities. The maximization problem is expressed as

$$\max_{T \in I} \left\{ \max_{C} U\left[(T, \mathbf{C}, \epsilon; \beta) \,|\mathbf{P}_C \mathbf{C} = Y - P_T T\right] \right\} \tag{1}$$

where P_k is the price of good of type $k = T, C$; Y is income; β is a vector of preference parameters; and ϵ stands for randomness specific to each individual. Optimization of (1) yields the marshallian demand for trips,

$$T^* = T(P_T, \mathbf{P}_C, Y, \epsilon, \beta) \tag{2}$$

T^* could be inferred if repeated observations on a single individual were available. Instead, only price variation across individuals is available. Thus conditional probabilities of observing a level of demand are estimated. A probability density function summarizing these probabilities is required. Given the nonnegative, integer nature of trips, the Poisson distribution is a natural candidate to summarize these probabilities.

In a Poisson regression, the expected value of the dependent variable is modeled as a function of prices, income, and other observable variables. Assuming an exponential form to keep the nonnegativity,

$$E(T^*) = \lambda = exp(\beta_0 + \mathbf{P}_k \beta_{P_k} + \beta_Y Y) \tag{3}$$

Given (3) and a change in P_T from P_T^a to the choke price, $\bar{P}_T$, expected value of the consumer surplus is estimated as

$$E_\epsilon\left[CS(T)\right] = \int_\epsilon \int_{P_T^a}^{\bar{P}_T} T(P_T, \mathbf{P}_C, Y, \epsilon; \beta) dp f(\epsilon) d\epsilon$$

$$= \int_{P_T^a}^{\bar{P}_T} \int_{\epsilon} T(P_T, \mathbf{P}_C, Y, \epsilon; \beta) f(\epsilon) d\epsilon dp$$

$$= \int_{P_T^a}^{\bar{P}_T} \lambda(P_T, \mathbf{P}_C, Y; \beta) dp = CS[E_\epsilon(T)] \tag{4}$$

As pointed out by Hellerstein (1991), reversal of integration in (4) is correct as long as estimation of $E(T)$ is unbiased. Unbiasness implies that demand function must be independent of the error term. Independence between demand function and errors is guaranteed when errors only reflect individual white noise. Thus $E[CS(T)] = CS[E(T)]$ as long as there is a non-systematic component in the error term.

2.2 Heterogeneous individuals

Suppose individuals belong exlusively to one of G groups, $g = 1, ..., G$. Groups are assumed associated to an ordinal classification of preferences for trips, such that individuals in group g take more average trips than individuals in group $g - 1$. This assumption can be tested by comparing average trips across groups once the empirical model has been estimated, and groups have been indexed accordingly. Conditional on being member of group g, optimization of (1) yields

$$T_g^* = T_g(P_T, \mathbf{P}_C, Y, \epsilon, \beta) \tag{5}$$

and expected consumer surplus for group g is calculated as

$$E_\epsilon \left[CS_g(T_g)\right] = \int_{\epsilon} \int_{P_T^a}^{\bar{P}_{T,g}} T_g(P_T, \mathbf{P}_C, Y, \epsilon; \beta) dp f(\epsilon) d\epsilon$$

$$= \int_{P_T^a}^{\bar{P}_{T,g}} \int_{\epsilon} T_g(P_T, \mathbf{P}_C, Y, \epsilon; \beta) f(\epsilon) d\epsilon dp$$

$$= \int_{P_T^a}^{P_{T,g}} \lambda_g(P_T, \mathbf{P}_C, Y; \beta) dp = CS_g[E_\epsilon(T_g)] \tag{6}$$

where $\bar{P}_{T,g}$ is the choke price for group g. For given relative size of groups, π_g for $g = 1, ..., G$, expected consumer surplus for the entire population is

$$E_\epsilon \left[CS(T)\right] = \sum_{g=1}^{G} \pi_g E_\epsilon \left[CS_g(T_g)\right] =$$

$$\sum_{g=1}^{G} \pi_g \int_{P_T^a}^{\bar{P}_T} \lambda_g(P_T, \mathbf{P}_C, Y; \beta) dp = \sum_{g=1}^{G} \pi_g CS_g \left[E_\epsilon(T_g)\right] \tag{7}$$

Knowing individual's membership is a non-essential assumption. Let membership be unobserved. Because membership systematically explains number of trips, missing membership introduces a systematic component into the error term. Let η be the systematic error term, and ϵ remain as the white noise. In this case, welfare estimation is calculated as follows

$$
\begin{aligned}
E_{\eta,\epsilon}\left[CS(T)\right] &= E_{\eta,\epsilon}\left[\int_{P_T^a}^{P_T^b} T(P_T, \mathbf{P}_C, Y, \eta, \epsilon; \beta)dp\right] \\
&= \int_{\eta}\int_{P_T^a}^{P_T^b}\int_{\epsilon}\left[T(P_T, \mathbf{P}_C, Y, \eta, \epsilon; \beta)f(\epsilon)d\epsilon\right]dp f(\eta)d\eta \\
&= \int_{\eta}\int_{P_T^a}^{P_T^b}\lambda(P_T, \mathbf{P}_C, Y, \eta; \beta)dp f(\eta)d\eta \\
&= \int_{\eta} CS[E_\epsilon(T)] = E_\eta CS[E_\epsilon(T)] \qquad (8)
\end{aligned}
$$

As in the homogeneous case, reversal of integration with respect to the white noise integral is correct. However, η represents a systematic error term which makes $E(T)$ biased.[3] Thus reversal of integration with respect to the systematic error would be inconsistent.

Expressions (7) and (8) show that, in the heterogeneous case, welfare is measured as the expected consumer surplus of the expected trips. Preference parameters in a heterogeneous scenario can be recovered using a latent class count specification, as explained in the next section.

3 Latent class count models

Let $\mathbf{d}_i = (d_{i1}, ..., d_{iG})$ be an indicator matrix such that $d_{ig} = 1$ if T_i is drawn from the g^{th} group for $i = 1, ..., n$, and $\sum_g d_{ig} = 1$. Estimation of latent class specifications is carry out by treating membership, d_{ij}, as missing data which is consistent with the theoretical framework explained in section 2.2.[4]

3.1 Model

Derivation of a latent class model starts by assuming a complete-data scenario where $V_i = (T_i, \mathbf{d}_i)$. T_i is the observed number of trips and $\mathbf{d}_i$ is the

[3] See Mullahy (1997) for a mathematical proof.

[4] Detailed explanation of latent class models as incomplete-data problem is presented in Cameron and Trivedi (1998); McLachlan and Peel (2000).

membership matrix. Under a complete-data scenario, $(T_i|\mathbf{d_i}, \beta)$ is distributed with density

$$f(T_i|\mathbf{d}_i, \beta) = \sum_{g=1}^{G} d_{ig} f(T_i|\beta_g) = \prod_{g=1}^{G} f(T_i|\beta_g)^{d_{ij}} \tag{9}$$

where $\beta = [\beta_1, ..., \beta_G]$, and $f(T_i|\beta_g)$ is a count density, usually Poisson or negative binomial.

In this complete-data scenario, $(d_{ij}|\pi)$ are assumed iid with multinomial distribution

$$\prod_{g=1}^{G} \pi_g^{d_{ig}}, \qquad 0 < \pi_g < 1, \qquad \sum_{g=1}^{G} \pi_g = 1 \tag{10}$$

where $\pi' = [\pi_1, ..., \pi_G]$. Expressions (9) and (10) together provide the distribution of the complete-data:

$$(T_i|\pi, \beta) \overset{iid}{\sim} \sum_{g=1}^{G} \pi_g^{d_{ig}} f_g(T; \beta_g), \quad 0 < \pi_g < 1, \quad \sum_{g=1}^{G} \pi_g = 1 \tag{11}$$

which is the canonical expression of a latent class model describing T_i as drawn from group g with probability π_g. The consequent likelihood function is

$$L(\beta, \pi|\mathbf{T}) = \prod_{i=1}^{n} \sum_{g=1}^{G} \pi_g^{d_{ig}} [f_g(\mathbf{T}; \beta_g)]^{d_{ig}} \tag{12}$$

While several approaches can be used in maximization of (12) (Bohning, 1995), explanation of the EM algorithm provides a neat description of how the incomplete-data problem is handled.

3.2 Estimation through the EM algortihm

The EM algorithm is an iterative algorithm that carries out two steps in each iteration: the expectation step (E-step) and the maximization step (M-step). The EM algorithm computes the maximum likelihood estimation (MLE) of an incomplete-data problem by formulating an associated complete-data problem, and exploiting the simplicity of the latter's MLE to compute the former's MLE (Ng et al., 2004). Both E- and M-steps have simple forms when the density functions belong to the exponential family (Ng et al., 2004) which is the case for the most common count distributions,

Poisson and negative binomial.[5] This simplicity relies on the linearity of the complete-data log likelihood with respect to the unobservable data.

Given an initial guess for π and β, (π^0, β^0), the EM algorithm maximizes expression (12) by treating d_{ig} as missing data. More specifically, the E-step deals with the incomplete-data problem by taking the conditional expectation of the complete-data log likelihood.

If d_{ig} were observable, the complete-data log likelihood would be

$$lnL(\beta|T, \mathbf{d}_i) = \sum_{i=1}^{n}\sum_{g=1}^{G} d_{ig}[lnf_g(T_i;\beta_g) + ln\pi_g] \tag{13}$$

Because the complete-data log likelihood is linear in d_{ig}, the E-step replaces d_{ig} by its expected value, $E(d_{ig})$. This strategy yields the expected log likelihood,

$$E[lnL(\beta|T, \mathbf{d}_i)] = \sum_{i=1}^{n}\sum_{g=1}^{G} \hat{z}_{ig}[lnf_g(T_i;\beta_g) + ln\pi_g] \tag{14}$$

where $E(d_{ig}) = \hat{z}_{ig}$, and $\hat{z}_{ig}$ denotes the posterior probability that observation T_i belongs to the group g. Given initial guesses, (π^0, β^0), and a vector of covariates, $\mathbf{X}_i$,

$$\hat{z}_{ig} = \frac{\pi_g^0 f_g(T_i|\mathbf{X}_i, \beta^0)}{\sum_{g=1}^{G} \pi_g^0 f_g(T_i|\mathbf{X}_i, \beta^0)} \tag{15}$$

Consistently, the M-step deals with the incomplete-data problem by replacing d_{ig} by its expected value, $\hat{z}_{ig}$, when maximizing the first order conditions of (13).

If d_{ig} were observable, the maximized first order conditions would be

$$\begin{aligned} \pi_g - \frac{\sum_{i=1}^{n} d_{ig}}{n} &= 0, \qquad g = 1, ..., G \\ \sum_{i=1}^{n}\sum_{g=1}^{G} d_{ig}\frac{\partial lnf_g(T_i;\beta_g)}{\partial \beta_g} &= 0 \end{aligned} \tag{16}$$

Instead, the M-step maximizes

$$\begin{aligned} \pi_g - \frac{\sum_{i=1}^{n} \hat{z}_{ig}}{n} &= 0, \qquad g = 1, ..., G \\ \sum_{i=1}^{n}\sum_{g=1}^{G} \hat{z}_{ig}\frac{\partial lnf_g(T_i;\beta_g)}{\partial \beta_g} &= 0 \end{aligned} \tag{17}$$

[5]The negative binomial is a member of the exponential family when the overdispersion parameter is treated as known and fixed (Congdon, 2005).

The E- and M-steps are alternated repeatedly until the change in the log likelihood is smaller than a user-defined value. Thus for a given number of groups, G, maximization yields probabilities for each individual's membership to each group, i.e. $N * G$ individual probabilities. Along with this probabilities, the relative size of each group, π_g, is also obtained. Consequently, the associated G sets of density parameters, β_g for $g = 1, ..., G$, are estimated.[6]

Convergence conditions are discussed in Wu (1983). Convergence to a global maximum is not guaranteed. Thus different starting values should be tried to find the global maximum. Peters and Walker (1978) and Redner and Walker (1984) provide regularity conditions in order to find consistent, efficient, and asymptotically normal estimates. The form of these conditions suggests they should hold for the distributions in the exponential family (McLachlan and Peel, 2000).

3.3 Welfare estimation

For the specific case of the latent class Poisson model, estimates are obtained for

$$f(T_i|P_{T_i}, \beta) = \sum_{g=1}^{G} \pi_g \frac{exp(-\lambda_g)\lambda_g^{T_i}}{T_i!} \tag{18}$$

where

$$\lambda_g = exp(\beta_{0,g} + \beta_{P_T,g} P_T) \tag{19}$$

With estimated parameters at hand, welfare estimates are carried out as follows

$$\begin{aligned} \sum_{g=1}^{G} \hat{\pi}_g \int_{P_T^a}^{\bar{P}_T} \hat{\lambda}_g(.)dp = \sum_{g=1}^{G} \hat{\pi}_g \int_{P_T^a}^{\bar{P}_T} exp(\hat{\beta}_{0,g} + \hat{\beta}_{P,g} P_T)dp \\ = E\{CS_g\left[E(T_g)\right]\} \end{aligned} \tag{20}$$

Expression (20) is the expected consumer surplus of the expected trips, just as in equations (7) and (8).

Reversion between the price integral and the summation over groups in equation (20) would be inconsistent with theoretical framework explained in section 2.2. Reversal of integration is not correct if variation in choke

[6]Wedel et al. (1993) provides a step-by-step explanation for maximization of a latent class count model.

price systematically depends on variation of the unobserved factors (Haab and McConnell, 1996). This is the case here. To see why, recall that choke price is the minimum price driving an individual to take zero trips, i.e., $T(\bar{P}_T, \mathbf{X}, \eta; \beta) = 0$, or $\bar{P}_T = T^{-1}(\mathbf{X}, \eta; \beta)$. The order of integration can not be reversed when the upper bound of integration on the inner integral is a function of the variable of integration for the (discrete version of the) outer integral. In equation (20), choke price varies systematically from one group to other. This variation is driven by the variation in $\beta_{P_T,g}$, such that $\bar{P}_{T,g} = T^{-1}(\mathbf{X}; \beta_g)$ for $g = 1, ..., G$. This variation is a direct consequence of membership being missed which is what introduces a systematic error. Because withe noise —ϵ in section 2.2— does not induce systematic variation, reversal integration with respect to ϵ is implied in expression (20).

4 Two-part models

Hurdle and zero-inflated models are the two foremost methods used to deal with excess zeros (Hilbe, 2007). Both models assume that some portion of the population self-excludes from the recreational market. Non-participants may permanently not visit the site under study (hurdle model), or may be divided in permanent and temporal non-participants (zero-inflated model). Permanent non-participants will never visit the site under study, no matter how low price could get. Temporal non-participants are currently facing a price at or above their choke price, i.e., they are in a corner solution (Phaneuf and Smith, 2005).[7]

4.1 Hurdle model

In a hurdle model, trips are positive once a hurdle has been passed. Otherwise, zero trips are observed. This behavior is modeled sequentially. The first model explains why some individuals take positive trips. Conditional on taking positive trips, the second model explains number of trips. Statistically, these processes are modeled by a binary model and a zero-truncated

[7]For detailed description of these models, see Gurmu and Trivedi (1996); Haab and McConnell (1996); Shonkwiler and Shaw (1996). For details on econometric estimation, see Hilbe (2007).

count model, respectively. Welfare is estimated as

$$CS[E(T)] = \int_{P_T^a}^{\bar{P}_T} \frac{Prob_1(T>0)}{Prob_2(T>0)}\hat{\lambda}(.)dp = \frac{Prob_1(T>0)}{Prob_2(T>0)}\int_{P_T^a}^{\bar{P}_T}\hat{\lambda}(.)dp \quad (21)$$

where $Prob_1(T>0)$ is the probability of positive trips according to the binary model, and $Prob_2(T>0)$ is the probability of positive trips according to the count model.

4.2 Zero-inflated model

Unlike a hurdle model, a zero-inflated model allows zero counts be generated by both the binary process and the count process. In this way, zero-inflated models distinguish permanent non-participants —zeros in the binary model— from temporal non-participants —zeros in the count model. A zero-inflated model is a double-hurdle model in the sense that positive trips are observed if the individual overcomes the hurdle modeled by the binary model, and then overcomes the probability of taking zeros under the count model. Because the count model is not truncated, welfare is estimated as

$$CS[E(T)] = \int_{P_T^a}^{\bar{P}_T} Prob_1(T>0)\hat{\lambda}(.)dp = Prob_1(T>0)\int_{P_T^a}^{\bar{P}_T}\hat{\lambda}(.)dp \quad (22)$$

4.3 Drawbacks

Hurdle models do not explain presence of non-participants. In a hurdle model, the same vector of explanatory variables is included in both parts. In this sense, no specific reason to self-exclusion from the recreational market is offered.

Zero-inflated models offer a theroretically arbitrary explanation for non-participation. Unlike the hurdle model, the binary process in a zero-inflated model may include different predictors than in the count model. However, the estimation process typically faces difficulties discriminating between covariates affecting participation from covariates affecting number of trips (Haab and McConnell, 2002). This difficulties arise because of the lack of a theoretical framework guiding researchers in the understanding of how individuals choose to participate in a recreational market. However, a theoretical framework supporting separation of determinants of participation and demand

would need to deal with an inconsistency: separate determinants imply researchers are estimating different preference functions (Phaneuf and Smith, 2005).[8]

The ultimate concern from a public policy perspective is exclusion of the arbitrarily-classified as permanent non-participants from welfare estimations. Permanent non-participants are excluded in two ways. First, estimation of the impact of travel cost on number of trips (price parameter) is based on a subsample that does not include zero counts at all (hurdle model), or includes zero counts from only temporal non-participants (zero-inflated model). Second, welfare estimates are weighted such that only participants (hurdle model) or participants and temporal non-participants (zero-inflated model) are accounted for [see expressions (21) and (22)].

5 Empirical illustration

5.1 Data

Empirical strategies explained in previous sections are applied to recreation data from the 1997 Iowa Wetlands Survey conducted at Iowa State University. This mail survey gathered information from a sample of all Iowa residents on actual and hypothetical use of wetlands in the state. Sociodemographic characteristics and information to calculate travel costs were also collected. Focus in this study is on the actual visitation data, analyzed previously by Phaneuf and Herriges (1999) and von Haefen and Phaneuf (2003b).

From a sample of 6000 Iowa households, 2891 usable surveys were obtained. Individuals were provided with a copy of a map where Iowa was divided into 15 zones. Individuals were asked to record the number of trips made to wetlands in each zone during 1997. This study focuses on zone 11, the zone with the largest dispersion on trips —variance on trips ranges from 2.41 (zone 2) to 23.39 (zone 11). People taking zero trips to zone 11 represent 78.17% of the sample. The mean number of trips is 1.64 trips, with a maximum of 45 trips.

[8] "In a fully consistent model with non-participation, a comparison between the market price and the individual's reservation price, derived from all arguments in the utility problem, implies an extensive margin of choice between conditional utilities representing participation and non-participation. Importantly, the same factors determine participation and consumption" (Phaneuf and Smith (2005),p. 71).

Summary statistics of variables included in econometric specifications are presented in table 1. Travel costs were calculated as the sum of the estimated round trip travel distance multiplied by $0.21 and the estimated travel time valued at one-third the wage rate.[9] For this application, substitute site is the cheapest alternative for each individual.

5.2 Econometric specifications

Table 2 shows estimates from a latent class Poisson model with four classes. This model performs the best among the latent class specifications in terms of likelihood criteria (see table 3). Groups are indexed such that individuals in group g take more average trips than individuals in group $g-1$. According to comparisons presented in table 4, differences in average trips across groups are significant and consistent with the order implied by the index assigned.[10] Table 5 presents distribution of trips inside groups.

Thus the data can be described as a sample from a population that is composed by four groups of visitors. The first group represents 56% of the population and includes very infrequent visitors — most of them take zero trips, with a maximum of 5. The fourth group represents 11% of the population and includes the very frequent visitors — 50% of them take at least 23 trips, with a maximum of 45. Groups 2 and 3 represent 33% of the population together and include two groups of relatively frequent visitors — 50% of visitors in group 2 take at least 4 trips with a maximum of 15, and 50% of visitors in group 3 take at least 10 trips with a maximum of 30.

Table 6 shows hurdle and zero-inflated specifications. In contrast to the hurdle specification, zero-inflated specifications II and III describe trips with two different sets of variables, one for the binary process and other for the count process. The zero-inflated specification performing the best in terms of likelihood criteria is the third one. Both hurdle model specification and zero-inlfated III specification perform poorer than latent class poisson specification in terms of likelihood criteria.

Signs of travel cost parameters are theoretically consistent in hurdle, zero-inlfated and latent class specifications. However, magnitudes differ. While hurdle and zero-inflated specifications yield own travel cost parameters

[9]For further details, see von Haefen and Phaneuf (2003b).

[10]These comparisons correct for experiment-wise error implied by multiple comparisons, i.e., for the increase in probability of making a type I error when doing multiple comparisons [see Daniel (1990)].

around −0.18, latent class specification yields a similar magnitude (−0.19) only for the third group which represents 6% of the population. Impact from travel cost is considerable larger in groups 1 and 4 — −5.52 and −7.27, respectively.

Table 7 provides insights on how each specification replicates the empirical frequencies. The latent class specification outperforms both hurdle and zero-inflated specifications not only on replicating the empirical frequency of zeros but also the empirical frequencies in the upper tail.

5.3 Welfare estimations

Welfare estimations are presented in table 8. Average estimates from hurdle and zero-inflated specifications are larger than estimates from latent class specification. These differences, however, do not hold when confidence intervals are considered. These conclusions remain for intervals at 90% of confidence (not shown).

6 Conclusions and further research

In a latent class specification, corner solutions can potentially be originated in different groups of visitors. In this specific application, corner solutions are originated only in the group of very infrequent visitors.

Latent class specification outperforms two-part models in replicating empirical frequencies. Interestingly, improvement in frequency prediction does not translate into different welfare estimates. Whether this is a general result requires further research. Clearly, even when latent class specifications yield similar welfare estimates, an advantange in using latent class specifications is the chance to provide targeted public policy recommendations. Also, data requirement in a latent class model is smaller than in a zero-inflated model, in the sense that latent class models do not need to explain non-participation with variables different than those included in the count process.

For this specific application, we still can distinguish a tendency in two-part models to yield larger welfare estimates — in addition to larger average estimates, upper bounds in two-part models are larger than the upper bound in latent class specification by at least 34%. This tendency is a consequence of two-part models yielding a travel cost parameter that captures the behavior

of visitors from groups 2 and 3. These visitors obtain the largest per trip consumer surplus.

While this is a single-site application, results are relevant for multi-site applications. For instance, von Haefen and Phaneuf (2003b) find zero-inflated count model and Kuhn-Tucker econometric model yield similar welfare estimates in a multi-site application. Future research will focus on checking whether their result holds when using a latent class specification instead of a zero-inflated model.

References

D. Bohning. A review of reliable maximum likelihood algorithms for semiparametric mixture models. *Journal of Statistical Planning and Inference*, 47:5–28, 1995.

A. C. Cameron and P. K. Trivedi. *Regression analysis of count data*. Cambridge University Press, 1998.

P. Congdon. *Bayesian models for categorical data*. John Wiley and Sons, 2005.

W. Daniel. *Applied nonparametric statistics*. PWS KENT Publishing Company, second edition, 1990.

P. Deb and P. Trivedi. The structure of demand for health care: latent class versus two-part models. *Journal of Health Economics*, 21:601–625, 2002.

E. English. Recreation nonparticipation as a choice behavior rather than statistical outcome. *American Journal of Agricultural Economics*, 90(1):186–196, 2008.

K. Evans and J. A. Herriges. Rounding in recreation demand models: A latent class count model. Working paper, Department of Economics, Iowa State University, 2010.

S. Gurmu and P. K. Trivedi. Excess zeros in count models for recreational trips. *Journal of Business and Economic Statistics*, 14(4):469–477, 1996.

T. Haab and K. McConnell. Count data models and the problem of zeros in recreation demand analysis. *American Journal of Agricultural Economics*, 78: 89–102, 1996.

T. Haab and K. McConnell. *Valuing environmental and natural resources. The econometrics of non-market valuation*. Edward Elgar, 2002.

D. Hellerstein. Using count data models in travel cost analysis with aggregate data. *American Journal of Agricultural Economics*, 73(3):860–866, 1991.

D. Hellerstein and M. Mendelsohn. A theoretical foundation for count data models, with and application to a travel cost model. *American Journal of Agricultural Economics*, 75:604–611, 1993.

J. Hilbe. *Negative Binomial regression*. Cambridge University Press, 2007.

G. McLachlan and D. Peel. *Finite mixture models*. John Wiley and Sons, 2000.

J. Mullahy. Heterogeneity, excess zeros, and the structure of count data models. *Journal of Applied Econometrics*, 12:337–350, 1997.

A. S. K. Ng, T. Krishnan, and G. McLachlan. The EM algorithm. In J. E. Gentle, W. Hardle, and Y. Mori, editors, *Handbook of computational statistics: concepts and methods*, pages 137–168. Springer-Verlag, 2004.

B. C. Peters and H. F. Walker. An iterative procedure for obtaining maximum likelihood estimators ofthe parameters for a mixture of normal distributions. *SIAM Journal on Applied Mathematics*, 35:362–378, 1978.

D. Phaneuf and J. Herriges. Choice set definition issues in a Kuhn-Tucker model of recreation demand. *Marine Resource Economics*, 14:343–355, 1999.

D. Phaneuf and K. Smith. Recreation demand models. In K. Maler and J. Vincent, editors, *Handbook of Environmental Economics*, pages 671–751. Amsterdam:Elsevier, 2005.

R. A. Redner and H. F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26:195–239, 1984.

R. Scarpa, M. Thiene, and T. Tempesta. Latent class count models of total visitation demand: Days out hiking in the eastern Alps. *Environment and Resource Economics*, 2007. URL `doi:10.1007/S10640-007-9087-6`.

J. S. Shonkwiler and W. D. Shaw. Hurdle count-data models in recreation demand analysis. *Journal of Agricultural and Resource Economics*, 21(2):210–219, 1996.

M. Thiene and E. Morey. Explaining and predicting recreation participation and site selection with life-constraints/durables: kids, spouse, fat, fitness, skill, and bad habits. Paper presented at the Fourth World Congress of Environmental and Resource Economists, 28 June-2 July, Montreal Canada, 2010.

R. H. von Haefen and D. J. Phaneuf. A note on estimating nested constant elasticity of substitution preferences for outdoor recreation. *American Journal of Agricultural Economics*, 85(2):406–413, 2003a.

R. H. von Haefen and D. J. Phaneuf. Estimating preferences for outdoor recreation: A comparison of continuous and count data demand system frameworks. *Journal of Environmental Economics and Management*, 45:612–630, 2003b.

R. H. von Haefen, D. M. Massey, and W. L. Adamowicz. Serial nonparticipation in repeated discrete choice models. *American Journal of Agricultural Economics*, 87(4):612–630, 2005.

M. Wedel, W. Desarbo, J. Bult, and V. Ramaswamt. A latent class poisson regression model for heterogeneous count data. *Journal of Applied Econometrics*, 8:397–411, 1993.

C. F. J. Wu. On the convergence properties of the EM algorithm. *Annals of Statistics*, 11:95–103, 1983.

Table 1: Summary statistics (2891 observations).

Variable	Mean	Std Dev	Min	Max
Trips in 1997	1.65	4.83	0.00	45.00
Travel cost[a]	80.87	54.43	15.78	486.92
Cost to substitute[a]	31.08	11.18	10.44	112.41
Male[b]	0.74	0.44	0.00	1.00
Age	49.09	15.64	16.00	80.00
College[b]	0.28	0.45	0.00	1.00
Income[c]	43.26	28.03	5.00	175.00
Fishing license[b]	0.68	0.47	0.00	1.00

[a] 1997 $; [b] 1 if characteristic is observed; [c] thousands of 1997 $

Table 2: Latent class Poisson specification (4 classes)[a]

	Mean[b]	Group 1		Mean[b]	Group 2		Mean[b]	Group 3		Mean[b]	Group 4	
Constant	—	-0.47		—	2.31	***	—	3.61	***	—	3.81	***
	—	(0.54)		—	(0.13)		—	(0.15)		—	(0.12)	
Travel cost[c]	8.94	-5.52	**	3.99	-0.67	***	6.53	-0.19	***	3.55	-7.27	***
	(5.57)	(2.81)		(1.72)	(0.03)		(4.50)	(0.01)		(1.21)	(1.07)	
Substitute[c]	3.07	5.73	**	3.20	0.67	***	3.29	0.12	***	3.54	7.22	***
	(1.12)	(2.80)		(1.10)	(0.03)		(1.04)	(0.03)		(1.20)	(1.06)	
Male	0.73	1.67	***	0.78	0.52	***	0.79	0.35	***	0.74	0.40	***
	(0.45)	(0.38)		(0.42)	(0.07)		(0.41)	(0.07)		(0.44)	(0.06)	
Age[c]	4.96	-0.69	***	4.71	-0.24	***	4.69	-0.22	***	4.64	-0.16	***
	(1.58)	(0.07)		(1.49)	(0.02)		(1.42)	(0.02)		(1.41)	(0.02)	
College	0.28	0.25		0.30	0.12	*	0.27	-0.12	*	0.31	-0.18	***
	(0.45)	(0.18)		(0.46)	(0.06)		(0.45)	(0.06)		(0.47)	(0.06)	
Trips	0.05			4.40			9.69			22.88		
	(0.34)			(2.92)			(6.65)			(8.85)		
α_g[d]		—			—			—			—	
LL[e]		-242.17			-854.58			-417.76			-257.54	
π_g[f]		0.56			0.27			0.06			0.11	

[a] Standard errors in parentheses. [b] Average value based only on members of the class. Standard deviation in parentheses. [c] Scaled by 10. [d] Overdispersion parameter in Poisson specifications is restricted to zero. [e] Log-likelihood. [f] Relative size of groups. P-value * < 0.10, ** < 0.05, *** < 0.10 .

Table 3: Likelihood criteria for latent class specifications

Number of classes					
Poisson	Negative Binomial	LL[a]	Parameters	AIC	BIC
2	0	-2605.71	12	5235.42	5252.95
1	1	-2740.11	13	5506.22	5525.21
0	2	-2608.05	14	5244.10	5264.55
3	0	-1894.31	18	3824.60	3850.89
2	1	-2267.56	19	4573.12	4600.87
1	2	-2420.52	20	4881.07	4910.22
0	3	-2340.72	21	4723.44	4754.12
4	**0**	**−1782.83**	**24**	**3613.60**	**3648.66**
3	1	-2189.77	25	4429.54	4466.06
2	2	-2062.09	26	4176.18	4214.16
1	3	-2269.95	27	4593.90	4633.34
0	4	-2140.48	28	4336.96	4377.86

[a] Log-likelihood

Table 4: Comparison of average trips across groups[a]

	95% confidence interval	
Comparison	lower bound	upper bound
$T_4 - T_1$	21.39	24.27
$T_4 - T_2$	16.92	20.04
$T_4 - T_3$	11.43	14.95
$T_3 - T_1$	8.56	10.72
$T_3 - T_2$	4.05	6.53
$T_2 - T_1$	3.64	5.06

Based on Tukey-Kramer method for multiple comparisons.

Table 5: Distribution of trips inside groups

Percentile	Sample	Group 1	Group 2	Group 3	Group 4
0	0	0	1	1	2
25	0	0	2	4	16
50	0	0	4	10	23
75	0	0	6	13	30
90	5	0	10	20	35
99	25	2	12	26	44
100	45	5	15	30	45

Table 6: Hurdle and zero-inflated negative binomial specifications[a]

	Hurdle model				Zero-inflated I				Zero-inflated II				Zero-inflated III			
	ZTNB		Logit		NB		Logit		NB		Logit		NB		Logit	
Constant	2.43	***	0.66	**	2.40	***	-0.85	***	2.43	***	-0.79	**	2.42	***	-0.88	***
	(0.22)		(0.26)		(0.22)		(0.30)		(0.22)		(0.35)		(0.22)		(0.29)	
Travel cost[b]	-0.20	***	-0.48	***	-0.18	***	0.46	***	-0.18	***	0.55	***	-0.18	***	0.55	***
	(0.02)		(0.02)		(0.02)		(0.03)		(0.02)		(0.03)		(0.02)		(0.03)	
Substitute[b]	0.22	***	0.52	***	0.20	***	-0.50	***	0.20	***	-0.00		0.20	***	—	
	(0.04)		(0.05)		(0.00)		(0.01)		(0.00)		(0.01)		(0.00)		—	
Male	0.20	*	0.49	***	0.20	*	-0.49	***	0.19	*	-0.14		0.21	**	—	
	(0.11)		(0.13)		(0.11)		(0.14)		(0.11)		(0.15)		(0.11)		—	
Age[b]	-0.13	***	-0.25	***	-0.13	***	0.24	***	-0.13	***	0.18	***	-0.13	***	0.18	***
	(0.03)		(0.03)		(0.03)		(0.04)		(0.03)		(0.04)		(0.03)		(0.04)	
College	-0.19	*	0.32	***	-0.19	*	-0.45	***	-0.20	**	-0.45	***	-0.20	**	-0.44	***
	(0.10)		(0.11)		(0.10)		(0.14)		(0.10)		(0.15)		(0.10)		(0.15)	
Income[c]	—		—		—		—		—		-0.31	***	—		-0.31	***
	—		—		—		—		—		(0.04)		—		(0.03)	
License	—		—		—		—		—		-1.06	***	—		-1.09	***
	—		—		—		—		—		(0.14)		—		(0.14)	
α[d]	1.08	***			1.01	***			0.99	***			0.99	***		
LL[e]	-2915.62				-2925.44				-2869.52				-2869.95			
AIC	5857.24				5876.88				5769.05				5765.91			
BIC	5871.61				5954.48				5858.59				5843.51			

[a] Standard errors in parentheses. Scaled by [b] 10, [c] 1000. [d] Overdispersion parameter. [e] Log-likelihood.
P-value * < 0.10, ** < 0.05, *** < 0.10 .

Table 7: Difference between predicted and empirical frequencies

	Sample	Difference in frequencies		
Trips	frequency	HNB	ZINB III	LCPM
0	78.17	6.61	-30.23	-1.28
1	2.87	-2.87	13.94	1.76
2	3.70	-3.70	6.61	-1.31
3	2.80	-2.80	4.43	-0.83
4	1.52	-1.52	4.05	0.93
5	1.83	-1.83	2.21	-0.07
6	1.25	-1.14	3.29	0.90
7	0.31	1.83	1.97	0.76
8	0.93	1.63	0.10	0.10
9	0.14	3.77	0.07	0.24
10	2.01	1.90	-2.01	-1.31
11	0.03	2.11	-0.03	0.55
12	0.76	-0.42	0.24	-0.55
13	0.07	0.03	-0.07	0.28
14	0.14	-0.14	-0.14	0.03
15	0.59	-0.59	-0.59	-0.31
16	0.10	-0.10	-0.10	0.17
17	0.10	-0.10	-0.10	0.17
18	0.00	0.00	0.00	0.17
19	0.03	-0.03	-0.03	0.17
20	1.07	-1.07	-1.07	-0.86
21	0.00	0.00	0.00	0.24
22	0.03	-0.03	-0.03	0.35
23	0.07	-0.07	-0.07	0.00
24	0.00	0.00	0.00	0.17
25	0.59	-0.59	-0.59	-0.55
26	0.03	-0.03	-0.03	0.10
27	0.00	0.00	0.00	0.31
28	0.03	-0.03	-0.03	0.17
29	0.00	0.00	0.00	0.10
30	0.48	-0.48	-0.48	-0.48
32	0.00	0.00	0.00	0.03
33	0.00	0.00	0.00	0.07
34	0.03	-0.03	-0.03	0.00
35	0.07	-0.07	-0.07	-0.03
40	0.17	-0.17	-0.17	-0.17
45	0.03	-0.03	-0.03	-0.03

Table 8: Consumer surplus per trip[a] (1997 $)

	HNB	ZINB III	LCPM
Group 1	—	—	0.18
			(0.07-1.09)
Group 2	—	—	1.49
			(1.37-1.63)
Group 3	—	—	5.28
			(4.77-5.86)
Group 4	—	—	0.14
			(0.10-0.19)
Total	1.65	1.46	0.89
	(1.37-2.04)	(1.18-1.84)	(0.73-1.37)

[a] Bootstrapped 95% confidence interval in parantheses (10,000 replications). [b] Weighted sum across groups. Weights are relative size of gropus.

Working Papers of the Center of Economic Research at ETH Zurich

(PDF-files of the Working Papers can be downloaded at www.cer.ethz.ch/research/working-papers.html).

16/253 A. Martinez-Cruz
Handling excess zeros in count models for recreation demand analysis without apology

16/252 M.-C. Riekhof and F. Noack
Informal Credit Markets, Common-pool Resources and Education

16/251 M. Filippini, T. Geissmann, and W. Greene
Persistent and Transient Cost Efficiency - An Application to the Swiss Hydropower Sector

16/250 L. Bretschger and A. Schaefer
Dirty history versus clean expectations: Can energy policies provide momentum for growth?

16/249 J. Blasch, M. Filippini, and N. Kumar
Boundedly rational consumers, energy and investment literacy, and the display of information on household appliances

16/248 V. Britz
Destroying Surplus and Buying Time in Unanimity Bargaining

16/247 N. Boogen, S. Datta, and M. Filippini
Demand-side management by electric utilities in Switzerland: Analyzing its impact on residential electricity demand

16/246 L. Bretschger
Equity and the Convergence of Nationally Determined Climate Policies

16/245 A. Alberini and M. Bareit
The Effect of Registration Taxes on New Car Sales and Emissions: Evidence from Switzerland

16/244 J. Daubanes and J. C. Rochet
The Rise of NGO Activism

16/243 J. Abrell, Sebastian Rausch, and H. Yonezawa
Higher Price, Lower Costs? Minimum Prices in the EU Emissions Trading Scheme

16/242 M. Glachant, J. Ing, and J.P. Nicolai
The incentives to North-South transfer of climate-mitigation technologies with trade in polluting goods

16/241 A. Schaefer
Survival to Adulthood and the Growth Drag of Pollution

16/240 K. Prettner and A. Schaefer
Higher education and the fall and rise of inequality

16/239 L. Bretschger and S. Valente
Productivity Gaps and Tax Policies Under Asymmetric Trade

16/238 J. Abrell and H. Weigt
Combining Energy Networks

16/237 J. Abrell and H. Weigt
Investments in a Combined Energy Network Model: Substitution between Natural Gas and Electricity?

16/236 R. van Nieuwkoop, K. Axhausen and T. Rutherford
A traffic equilibrium model with paid-parking search

16/235 E. Balistreri, D. Kaffine, and H. Yonezawa
Optimal environmental border adjustments under the General Agreement on Tariffs and Trade

16/234 C. Boehringer, N. Rivers, H. Yonezawa
Vertical fiscal externalities and the environment

16/233 J. Abrell and S. Rausch
Combining Price and Quantity Controls under Partitioned Environmental Regulation

16/232 L. Bretschger and A. Vinogradova
Preservation of Agricultural Soils with Endogenous Stochastic Degradation

16/231 F. Lechthaler and A. Vinogradova
The Climate Challenge for Agriculture and the Value of Climate Services: Application to Coffee-Farming in Peru

16/230 S. Rausch and G. Schwarz
Household heterogeneity, aggregation, and the distributional impacts of environmental taxes

16/229 J. Abrell and S. Rausch
Cross-Country Electricity Trade, Renewable Energy and European Transmission Infrastructure Policy

16/228 M. Filippini, B. Hirl, and G. Masiero
Rational habits in residential electricity demand

16/227 S. Rausch and H. Schwerin
Long-Run Energy Use and the Efficiency Paradox

15/226 L. Bretschger, F. Lechthaler, S. Rausch, and L. Zhang
Knowledge Diffusion, Endogenous Growth, and the Costs of Global Climate Policy

15/225 H. Gersbach
History-bound Reelections

15/224 J.-P. Nicolai
Emission Reduction and Profit-Neutral Permit Allocations

15/223 M. Miller and A. Alberini
Sensitivity of price elasticity of demand to aggregation, unobserved heterogeneity, price trends, and price endogeneity: Evidence from U.S. Data

15/222 H. Gersbach, P. Muller and O. Tejada
Costs of Change, Political Polarization, and Re-election Hurdles

15/221 K. Huesmann and W. Mimra
Quality provision and reporting when health care services are multi-dimensional and quality signals imperfect

15/220 A. Alberini and M. Filippini
Transient and Persistent Energy Efficiency in the US Residential Sector: Evidence from Household-level Data

15/219 F. Noack, M.-C. Riekhof, and M. Quaas
Use Rights for Common Pool Resources and Economic Development

15/218 A. Vinogradova
Illegal Immigration, Deportation Policy, and the Optimal Timing of Return

15/217 L. Bretschger and A. Vinogradova
Equitable and effective climate policy: Integrating less developed countries into a global climate agreement

15/216 M. Filippini and L. C. Hunt
Measurement of Energy Efficiency Based on Economic Foundations

15/215 M. Alvarez-Mozos, R. van den Brink, G. van der Laan and O. Tejada
From Hierarchies to Levels: New Solutions for Games with Hierarchical Structure

15/214 H. Gersbach
Assessment Voting

15/213 V. Larocca
Financial Intermediation and Deposit Contracts: A Strategic View