Economics Bulletin

Volume 41, Issue 2

Double truncation in choice-based sample: An application of on-site survey sample

Kavita Sardana
TERI SAS

Abstract

In this paper, I derive the distribution of an on-site survey sample that is truncated from above. By bootstrapping the conditional expectation with and without upper truncation, I show that the differences are statistically significant. I conclude that ad-hoc truncation (upper truncation) might not be the best solution to homogenize the population, rather semi-parametric methods such as Latent Class Models should be used to identify different classes of homogeneous population.

Citation: Kavita Sardana, (2021) "Double truncation in choice-based sample: An application of on-site survey sample", *Economics Bulletin*, Vol. 41 No. 2 pp. 781-787.
Contact: Kavita Sardana - kavita.sardana@terisas.ac.in.
Submitted: January 27, 2021. **Published:** April 09, 2021.

I. Introduction

Choice-based samples are non-random samples based on stratification based on some attributes of the data generating process. A subsample of the population consisting of subjects with one outcome is collected. For example, the outcome could be participation by the user population when modelling an on-site sample. Data is then collected within the subsample with different attributes of the user population bringing the desired variation in the outcome variable (Sardana 2010). The primary reason for using choice-based samples is that the outcome variable is a rare event and using household survey data would require an immense amount of data collection effort, which in most cases is implausible and expensive. Choice-based samples provide economies of scale, which are not available with household surveys. For examples and benefits of choice-based sampling refer to the introduction by Manski and Lerman (1977). Predominant problems with choice-based samples are truncation, endogenous stratification, and non-negativity. Choice-based sampling has its drawbacks in that nonusers are not included in the sample which causes a truncated population. Truncation which commonly arises in on-site samples is when truncation limits are treated as exogenous or pre-determined[1]. For on-site samples, the left-truncation limit is set at zero. Sometimes, ad-hoc right side truncation is imposed to make the population homogeneous. For example, in the recreational demand literature, researchers routinely drop observations with high-frequency visits to rule out respondents who aren't really “visitors", but instead are best thought of as “residents" who report many visits due to the sites potentially very close proximity to their home. These residents who take frequent short-duration visits incur a lower travel cost and hence are willing to pay only a marginal price for these services. Englin and Shonkwiler (1995) drop observations with annual trips greater than 12, allowing one trip per month. Egan and Herriges (2006), and Bowker et al. (2009) drop observations with annual trips greater than 52, allowing one trip per weekend.

By dropping observations based on some arbitrary definition of “visitors”, researchers a-priori attempt to create a homogenous population of visitors. The right-side truncation affects the observed distribution of the endogenous variable that is different from the actual distribution. Not accounting for right-side truncation affects the associated probability mass function. In this paper, I derive the on-site distribution which is truncated from above at an ad-hoc truncation point. Our bootstraps results show that the expected value and therefore variance of an on-site Poisson distribution with and without upper truncation are statistically different. However, the statistical significance of the difference is sensitive to the point of ad-hoc truncation.

---

[1] Unlike when the truncation limits are endogenous. Random truncation occurs when a variable is observed only when its value lies within a random interval. Suitable correction of estimators is required because the sampling and population distributions are different (Moreira *et al*., 2016). This form of truncation is found in literature on public health, epidemiology, astronomy, and demography (Emura *et al*., 2015). The distribution function can be estimated through either non-parametric maximum likelihood estimation (NPMLE) or semi-parametric estimation where in the latter model, the distribution of truncation limits are assumed to belong to a parametric family (Moreira and de Una-Alvarez, 2010b; Shen, 2010b; Moreira *et al*., 2016).

II. Theoretical Model

Truncation affects the model in the following two ways: first, the model is truncated at zero because an on-site sampling procedure is used, and second, right-upper truncation where an ad-hoc ceiling is assumed on a strictly positive number of the discrete random variable. Endogenous stratification or the problem of recording higher frequency in the sampling process is addressed through the calculation of the on-site probability distribution (Shaw 1988, Pg. 214).

Let X be a vector of independent variables for each individual i. As given by Shaw (1988), I assume that for n individuals in the population, X is the same and is given by X^0. The general form density function for the i[th] individual is given by $f(y_i^* | X^0)$ where y_i^* is the desired quantity demanded and the observable quantity demanded can be generated using $y_i = y_i^* \text{ if } y_i^* > 0$, i.e., according to a truncated density function (left truncated due to on-site nature of survey sample). Additionally, a right-side truncation can be imposed along with left-side truncation using the marginal distribution of observed quantity, y_i, when adjusted for double truncation, given by $g(y_i | X^0)$ in "(1)",

$$g(y_i | X^0) = \frac{f(y_i^* | X^0)}{P(a_1 < y_i^* < a_2)} \quad (1)$$

Where y_i is the observed quantity demanded, a_1 is the lower limit (0 trips for an on-site sample) and a_2 is the upper limit. Given the probability of y equal to a specific value t and X^0, the limiting on-site distribution with double-truncation is given by "(2)":

$$P(y = t) = \frac{t f(t | X^0)}{\sum_{t=1}^{a_2} t f(t | X^0)} \quad (2)$$

Where a_2 is the upper-right side truncation limit. This is the density function of an observation y=t given X= X^0 in the on-site population.

By substitution "(2)" in equation "(1)" and writing the equation for the i[th] individual, I get the density function for a double-truncated endogenously stratified discrete random variable as given in "(3)":

$$h(y_i | x_i) = \frac{y_i f(y_i | X_i)}{\sum_{t=1}^{a_2} t f(t | X_i)} \quad (3)$$

III. Data

Data for estimating the on-site Poisson model that is truncated from above were obtained from the U.S. Forest Service's National Visitor Use Monitoring (NVUM) program. The NVUM survey is based on a stratified random sampling

design (English et al. 2002). The data were collected from the fourth round, which began in 2012 for a period of five years, through 2016. Participating National Forests are sampled every five years. During the on-site interviews, information was collected from visitors on their annual number of trips to a sampled National Forest in the last 12 months, and also the number of trips to the sampled National Forest for the activity indicated as the respondent's primary activity.

Information on socio-economic variables was also collected in the NVUM survey, including the gender and age of the respondent. Unlike earlier rounds, primary information on self-reported income and distance was collected for one-third of the sample. Income for the household was recorded as the total annual income of the respondent. Information on travel distance as miles traveled from home to site was also collected from each respondent, and distance from home to substitute location was recorded.

For our analysis, a single-site recreational demand model was estimated using data collected for the George Washington & Jefferson National Forests. The George Washington & Jefferson National Forests are located in the southeastern region of the U.S. In 1995, the George Washington National Forest in west central Virginia and the Jefferson National Forest in southwest Virginia were grouped together to form the George Washington & Jefferson National Forests combined management unit.

IV. Empirical Model

The sampling unit for the NVUM survey is a "group," which can be a single person or a party of persons travelling together, such as a family (Zarnoch et al. 2005). The NVUM survey measures recreation visits to a National Forest on a 12-month basis. Following the TCM protocol, only visitors who were visiting for the primary purpose of recreation were included in our analysis. Our empirical demand equation was specified as,

$$Visits_i = f(TC_i, Income_i, Female_i, Age_i). \quad (4)$$

In "(4)"[2] the dependent variable ($Visits_i$) represents the annual number of trips from individual *i* to the sampled National Forest. Socio-economic variables include annual income ($Income_i$), age (Age_i), and an indicator for a female survey respondent ($Female_i$). $Income_i$ is represented by the total annual income of the household. The price of a recreational trip is equal to travel costs for individual *i* (TC_i) estimated as the sum of driving and time costs following "(5)"[3]:

[2] For detailed description of empirical model refer to Sardana et al. (2021).

[3] In "(4)", driving costs are a function of one-way distance (*Distance)* from an individual's origin to the destination, the average operating costs (variable costs) per mile for a typical sedan type car in 2016 of 14.54 cents/mile as defined by the American Automobile Association (AAA 2016), and the number of passengers per vehicle ($PeopleVeh_i$). Time costs are a function of travel time estimated by dividing the round-visit distance by an average speed of 40 mph (Rosenberger and Loomis 1999) and the opportunity cost of time, which was evaluated at one-third of the wage rate (Baerenklau 2010). The wage rate was estimated

$$TC = (2 * Distance \times \$0.1454 / mile) / PeopleVeh + 0.33 \times \frac{Income}{2000} \times \frac{2 * Distance}{40\,mph} \quad (5)$$

V. Results

The model estimation results from the On-Site Poisson distribution are summarized in Table 1 and Table 2. Table 1 provides the conditional expectation for Model 1 (without upper truncation) and Table 2 with Model 2 (with upper truncation) and the differences in expectation from upper truncated and un-truncated models with bootstrap standard errors and 95% confidence intervals. For empirical estimation of the upper truncated model, I assume the following four upper truncation limits: annual number of visits less than or equal to 12, 25, 50, and 75. Model estimation results vary depending on the point of ad-hoc truncation.

From "(3)", I show that the nature of double truncation impacts the on-site distribution through the upper limit of integration of the expected value. The expected value is integrated over the truncated upper limit (a2). As a result, the expected value, for higher values of the random variable (beyond a2), gets omitted from the conditional mean calculation. This assumption is restrictive- for higher values of the random variable that get omitted, the mass probability is small, but due to the disproportionately higher value that the random variable takes, the product, which is the expectation does not approximate to zero.

Table 1: Conditional Expectation Estimates from on-site Poisson distribution without Ad-hoc Truncation (Model 1) and Bootstrap Standard Errors (replications=100)[4]

No truncation	EV	Std.Error	p-value	CI
Model 1	8.50	0.30	0.00	7.90 - 9.09

This can be corroborated from our empirical estimation of on-site distribution with and without upper truncation. The expected value without truncation is eight trips per annum. However, in the non-truncated model, the expected trips are not representative of higher trips in the population- this is because the probability mass function of higher trips is proportionally smaller. The *statistically significant* expected values with ad-hoc truncation are 2.46, 3.72, 6.61, and 7.59 trips per annum at points 12, 25, 50, and 75 respectively. In the results, the *difference* in expected value is statistically significant at 1 % for the threshold of 12 and 25 visits, and after that, for the threshold of 50, the difference becomes statistically significant at 15%. This is because the *difference* in the population in the non-truncated and truncated starts to approximate the population in the non-truncated, as annual visits increase. The statistical significance is sensitive to the proportion of truncated to non-truncated sample in the data set- with a marginal change in this proportion the significance of the difference between truncated and un-truncated model changes. For our data, the proportion of truncated to non-truncated sample after the threshold of 75 annual visits is only 0.05.

---

by dividing the income variable per annum by 2000 (Hynes and Greene 2013). All three variables (round-visit distance, income, and time) are considered exogenous.

[4]We used standardized normal probability plot, to check normality of our bootstrap variables and found them to be normal so reported CI with normal approximation.

Table 2: Conditional Expectation Estimates from on-site Poisson distribution with Ad-hoc Truncation (Model 2), difference in the two-models, and Bootstrap Standard Errors (replications=100)

Point of Truncation	EV(Model 2)	Std. Error	p-value	CI	EV(Difference)	Std. Error	p-value	CI
Truncation Point 12	2.46	0.30	**0.00**	1.86 - 3.05	6.04	2.70	**0.03**	0.74 - 11.34
Truncation Point 25	3.72	0.56	**0.00**	2.63-4.82	4.77	2.59	**0.06**	-0.30-9.84
Truncation Point 50	5.39	0.92	**0.00**	3.58-7.19	3.10	2.20	**0.15**	-1.20-7.41
Truncation Point 75	7.59	1.55	**0.00**	4.54-10.64	0.90	1.73	**0.60**	-2.49-4.30

VI. Conclusion

Empirically, imposing an arbitrary cut-off homogenizes the data and hence, not account for heterogeneity that arises due to differences in preferences. Rather than imposing ad-hoc right side truncation, semi-parametric methods such as Latent Class Models should be used to identify different classes of a homogeneous population.

VII. References

1. American Automobile Association (AAA). 2016. Your Driving Cots. https://exchange.aaa.com/wp-content/uploads/2017/05/2016-YDC-Brochure.pdf

   accessed March 18, 2020.

2. Baerenklau, Kenneth A. (2010): "A latent class approach to modeling endogenous spatial sorting in zonal recreation demand models." *Land Economics 86(4),* 800-816.

3. Bowker, J. Michael, C. Meghan Starbuck, Donald BK English, John C. Bergstrom, R. S. Rosenburger, and D. C. McCollum. (2009). "Estimating the net economic value of national forest recreation: an application of the National Visitor Use Monitoring database." Faculty Series Working Paper, FS 09-02, September 2009; The University of Georgia, Department of Agricultural and Applied Economics, Athens, GA 30602

   http://ageconsearch.umn.edu/handle/59603

4. Egan, Kevin, and Joseph Herriges. (2006): "Multivariate count data regression models with individual panel data from an on-site sample." *Journal of*

environmental economics and management 52(2), 567-581.

5. Emura, Takeshi, Yoshihiko Konno, and Hirofumi Michimae. (2015). "Statistical inference based on the nonparametric maximum likelihood estimator under double-truncation." *Lifetime data analysis* 21(3), 397-418.

6. Englin, Jeffrey, and J. Scott Shonkwiler. (1995). "Estimating social welfare using count data models: an application to long-run recreation demand under conditions of endogenous stratification and truncation." The Review of Economics and statistics *77(1),* 104-112.

7. English, D., et al. (2002). "Forest Service national visitor use monitoring process." USDA Forest Service, General Technical Report SRS-57, Southern Research Station, Ashville, NC

8. Hynes, Stephen, and William Greene. (2013). "A panel travel cost model accounting for endogenous stratification and truncation: A latent class approach." *Land Economics 89(1),* 177-192.

9. Sardana, Kavita. 2010. *Modeling demand for outdoor recreation with choice-based samples*. Doctoral Dissertation, UGA.

10. Manski, Charles F., and Steven R. Lerman. (1977). "The estimation of choice probabilities from choice based samples." *Econometrica: Journal of the Econometric Society*, 1977-1988.

11. Moreira, Carla, and Jacobo de Una-Alvarez. (2010). "Bootstrapping the NPMLE for doubly truncated data." *Journal of Nonparametric Statistics 22(5),* 567-583.

12. Moreira, Carla, Jacobo de Uña-Álvarez, and Luís Meira-Machado. (2016). "Nonparametric regression with doubly truncated data." *Computational Statistics & Data Analysis* 93, 294-307.

13. Rosenberger, Randall S., and John B. Loomis. (1999). "The value of ranch open space to tourists: combining observed and contingent behavior data." *Growth and change 30(3),* 366-383.

14. Sardana, K., Bergstrom, J. C., & Bowker, J. M. (2021). Effects of Ad-hoc Data Truncation and Homogeneous Preferences on Recreational Demand and Values: An Application to the George Washington and Jefferson National Forests. *Journal of Agricultural and Applied Economics*, 1-15.

15. Shaw, Daigee. (1988). "On-site samples' regression: Problems of non-negative integers, truncation, and endogenous stratification." *Journal of Econometrics 37(2),* 211-223.

16. Shen, Pao-sheng. (2010). "Nonparametric analysis of doubly truncated data." *Annals of the Institute of Statistical Mathematics 62(5),* 835-853.

17. Zarnoch, Stanley J., Donald BK English, and Susan M. Kocis. (2005). "An outdoor recreation use model with applications to evaluating survey estimators." Res. Pap. SRS-37 Asheville, NC: US Department of Agriculture, Forest Service, Southern Research Station 15 p37.