Economics Bulletin

Volume 41, Issue 4

Multi-episode count data estimation for health care demand

Hiroaki Masuhara
Faculty of Economics and Law, Shinshu University

Abstract

This study proposes new multi-episode count data models for health care analysis. Using the Pólya-Aeppli distribution, a Poisson process for seeking medical care and a geometric process for the number of treatments are specified. Moreover, this paper introduces unobserved heterogeneities to both Poisson and geometric processes. Using the National Medical Expenditure Survey, the proposed models demonstrate good performance and the large differences in estimated coefficients compared with conventional hurdle and finite mixture count data models. It is useful to apply the multi-episode count data models proposed in this paper.

We would like to thank the anonymous referee and the journal Associate Editor, Gueorgui Kolev, who provided useful and detailed comments on an earlier version of the manuscript. This work was financially supported by JSPS KAKENHI Grant Number JP21K01481. The authors would like to thank Enago (www.enago.jp) for the English language review.

Citation: Hiroaki Masuhara, (2021) "Multi-episode count data estimation for health care demand", *Economics Bulletin* Vol. 41 No. 4 pp. 2281-2290.

Contact: Hiroaki Masuhara - masuhara@shinshu-u.ac.jp.

Submitted: September 10, 2021. **Published:** December 29, 2021.

1. Introduction

When analyzing health care demand using count data, econometricians usually compare the performance of the two most common approaches, a hurdle (two-part) model and a finite mixture (FM) model. The hurdle model, first discussed by Mullahy (1986), distinguishes the decision to seek care from the level of utilization, focusing on the difference between users and non-users, and is occasionally regarded as an approximation of the principal-agent hypothesis. Pohlmeier and Ulrich (1995) and Gerdtham (1997) apply this method to analyze health care demand.

FM models assume that data consist of a finite number of subpopulations and that each element is drawn from one of these latent subpopulations. These approaches are widely used in health econometrics given that they are semi-parametric and flexible (Heckman and Singer, 1984). Since FM models capture unknown health status and estimate *ex post* behaviors of both healthy and non-healthy individuals, many authors estimate health care demand using both hurdle and FM models. Deb and Trivedi (1997, 2002); Deb and Holmes (2000); and Gerdtham and Trivedi (2001) find the FM model to be a more desirable approach. Jemernéz-Martín *et al.* (2002) assert that the FM model is not based on economics but on statistical reasoning, and observe a good performance of the hurdle model in EU countries. Using panel FM and panel FM hurdle models, Bago d'Uva (2005, 2006) analyze demand for health care in Britain. Winkelmann (2004) expands the hurdle model based on bivariate normally distributed heterogeneity and compares the performance of the two models.

The third approach to estimate health care demand is the multi-episode model, first proposed by Santos Silva and Windmeijer (2001). This model conceptualizes a "spell" of illness as a set of consecutive medical services received by an individual patient by request. The total amount of medical services in a given period are broken down into two different decision-making process that include the individual's decision to seek medical care and the length of treatments by professionals (i.e., doctors, dentists, nurse practitioners, etc.). The consecutive processes of the medical treatment from the beginning to end is referred to as an *episode*. Although hurdle models distinguish non-users and users, a multi-episode model distinguishes the number of individuals seeking care and the number of treatments.

Although a multi-episode model enables better understanding of the patients' behavior, it has had limited use in health economics because the model's performance is inadequate compared to the hurdle and FM models. Santos Silva and Windmeijer (2001) assume a Poisson distribution for seeking medical care and a logarithmic distribution for treatments, demonstrating these compound process results in a negative binomial (NB) distribution. This model requires perfect specification because there is not unobserved heterogeneity. This assumption is relaxed in this study, introducing a log-normal distributed unobserved heterogeneity for the Poisson distribution. Moreover, as logarithmic distribution is difficult to interpret, a geometric distribution with a probit specification for treatments processes is also introduced. Thus, this paper proposes a new model, the Pólya-Aeppli (geometric-Poisson) distribution with normal distributed heterogeneities. Using the National Medical Expenditure Survey (Deb and Trivedi, 1997), the performance of the proposed model, hurdle models, and FM models are compared and the estimated results are examined.

This study is organized as follows: Section 2 discusses multi-episode count data models with unobserved heterogeneities. Section 3 presents the information criteria, model specification tests, and estimated results using the National Medical Expenditure Survey. Section 4 concludes the paper.

2. Multi-episode count data modeling

Let V be the total number of visits to medical facilities in all episodes; S be the total number of episodes (or spells of illness); R_j, $j = 1, 2, \ldots, S$, be the number of visits in jth episode. Then, the total number of visits describes $V = S + \sum_{j=1}^{S} (R_j - 1)$. Santos Silva and Windmeijer (2001) identified S as the number of first treatments of medical professionals and R_j as the number of visits for that treatment. Given covariates $\mathbf{x} \sim K_x \times 1$, if the data of both R_j and S are observable, the conditional expectation of R_j and S can be easily estimated. However, in actuality, only the total number of medical care visits in a given period can be observed; that is, only V is available. Santos Silva and Windmeijer (2001) introduced a compound Poisson process to analyze multi-episode health care demand. Given the assumptions of conditional independence of R_j and S, and of $\mathrm{E}[R_j \mid \mathbf{x}] = \mathrm{E}[R \mid \mathbf{x}]$, V follows a compound Poisson distribution (also called a stopped-sum distribution). Using the law of iterated expectations, the following relation is obtained:

$$\mathrm{E}[V \mid \mathbf{x}] = \mathrm{E}_S[S \times \mathrm{E}_R[R_j \mid S, \mathbf{x}]] = \mathrm{E}[S \mid \mathbf{x}]\,\mathrm{E}[R \mid \mathbf{x}]. \tag{1}$$

When V takes discrete positive values, the probability mass function (PMF) of V elicits

$$\Pr(V = v) = \Pr(S = 0) + \sum_{j=1}^{\infty} \Pr(S = j)\, p^{(j)}(v), \tag{2}$$

where $p^{(j)}(v) = \underbrace{p * \cdots * p(v)}_{j}$ is jth convolution of distributions of independent random variables $v = \sum_j R_j$ that satisfies

$$p^{(1)}(v) = \Pr(R_1 = v),$$
$$p^{(j)}(v) = \Pr(R_1 + \cdots + R_j = v),$$
$$p^{(0)}(v) = \begin{cases} 0, & v \neq 0, \\ 1, & v = 0. \end{cases}$$

In the maximum likelihood estimation, the log-likelihood obtained by (2) is complicated and does not always have an explicit solution. The parameters are estimated by specifying the distributions of S and R that have moment generating functions and using the method of moments of (1).

Assuming specific distributions for S and R, V follows a simple distribution. For example, when S follows the Poisson distribution with a parameter λ and $R = 1, 2, \ldots$ follows the logarithmic (series) distribution with a parameter $0 < \theta < 1$, then V follows a NB distribution with parameters $(\theta, -\lambda/\ln(1-\theta))$. In a regression, the parameters λ and θ are usually specified as $\lambda = \exp(\mathbf{x}'\boldsymbol{\beta}_1)$ and $\theta = \Phi(\mathbf{x}'\boldsymbol{\beta}_2)$, where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are $K_x \times 1$ vectors of parameters and $\Phi(\cdot)$ is a cumulative distribution of the standard normal. Santos Silva and Windmeijer (2001) term this the Negbin$_X$ model. The maximum likelihood method easily estimates parameters of the Negbin$_X$ model since it is essentially the NB distribution.

If S follows a Poisson distribution with a parameter λ, and R follows a shifted geometric distribution $\Pr(R = r) = q^{r-1}(1-q)$ for $r = 1, 2, \ldots,$ where $0 < q < 1$ is a parameter for the geometric distribution, the PMF is directly obtained by

$$\Pr(V = 0) = \exp(-\lambda), \tag{3}$$
$$\Pr(V = v) = \sum_{j=1}^{v} \frac{\mathrm{e}^{-\lambda}\lambda^j}{j!} \binom{v-1}{j-1} q^{v-j}(1-q)^j. \tag{4}$$

Eqs. (3) and (4) are termed the Pólya-Aeppli distribution (Johnson *et al.*, 1992, Ch. 9, p. 378) or geometric-Poisson distribution. In a regression context, the parameter q is specified as $q = \Phi(\mathbf{x}'\boldsymbol{\beta}_2)$. In the Pólya-Aeppli distribution, using the latent variable j and the weighted sums of Poisson distribution, the calculation of this PMF is also feasible.

Although the Pólya-Aeppli distribution is not tractable, this distribution has a memoryless property. In acute treatment, if the doctor decides to end the treatment based only on the current diagnosis, this assumption is reasonable to interpret the behavior of health care demand. On the contrary, although the assumption of the logarithmic distribution for R is based mainly on mathematical tractability, this distribution is not memoryless. In chronic treatment, if the doctor provides treatments based on both the present diagnosis and on previous ones, the time-dependent assumption of the logarithmic distribution is more desirable for analyzing the total number of treatments.

In a compound Poisson distribution, the most important assumption is the conditional independence of S and R. Applying the techniques of Heckman and Singer (1984), Santos Silva and Windmeijer (2001) restricted the same coefficient parameters and modeled FM constant terms for correlated S and R. This method is difficult to estimate but is robust for the conditional independence assumption. However, it is not adequate for unobserved heterogeneity other than covariates, and full specification is required for λ and q (or θ). In this paper only unobserved heterogeneity is introduced to parameters, maintaining the conditional independence assumption. Since q (or θ) is specified as a probit model, the binary decision-making for whether the doctor continues treatment includes the normal distributed unobserved heterogeneity.[1] As in Winkelmann (2004), it is reasonable to introduce a normally distributed unobserved heterogeneity for λ.[2] That is, $\lambda\left(\varepsilon_1\right) = \exp\left(\mathbf{x}'\boldsymbol{\beta}_1 + \sigma_1\varepsilon_1\right)$, where ε_1 follows a standard normal distribution and σ_1 is a parameter for standard deviation. In this case, maintaining conditional independence, the PMF of the Pólya-Aeppli distribution takes the following form:

$$\Pr\left(V=0\right) = \int_{-\infty}^{\infty} \exp\left(-\lambda\left(\varepsilon_1\right)\right)\phi\left(\varepsilon_1\right)\mathrm{d}\varepsilon_1, \tag{5}$$

$$\Pr\left(V=v\right) = \int_{-\infty}^{\infty} \sum_{j=1}^{v} \frac{\mathrm{e}^{-\lambda(\varepsilon_1)}\lambda\left(\varepsilon_1\right)^j}{j!}\binom{v-1}{j-1} q^{v-j}\left(1-q\right)^j \phi\left(\varepsilon_1\right)\mathrm{d}\varepsilon_1, \tag{6}$$

where $\phi\left(\cdot\right)$ is a probability density function of a standard normal distribution. The above PMFs are easily evaluated using the Gauss-Hermite integration.

[1] Let $d^*_{\widetilde{r},\widetilde{r}+1}$, $1 \le \widetilde{r} \le r$ be a latent variable representing the decision to go to the $(\widetilde{r}+1)$-th visit. The hospital visit is determined by $d^*_{\widetilde{r},\widetilde{r}+1} = \mathbf{x}'\boldsymbol{\beta}_2 + \varepsilon_2$, and the unobserved heterogeneity ε_2 follows a standard normal distribution. The latent variable $d^*_{\widetilde{r},\widetilde{r}+1}$ cannot be observed, and instead $d_{\widetilde{r},\widetilde{r}+1} = 1$ is observable when $d^*_{\widetilde{r},\widetilde{r}+1} > 0$; otherwise, $d_{\widetilde{r},\widetilde{r}+1} = 0$. Corresponding to the geometric distribution, the termination condition $d_{r,r+1} = 0$ and $d_{\widetilde{r},\widetilde{r}+1} = 1$ for $1 \le \widetilde{r} \le r-1$ always hold. Therefore, the PMFs of (5) and (6) contain error terms that follow independent normal distributions.

[2] Dhaene and Santos Silva (2012) proposed a Poisson model that has more general unobservable heterogeneity, including the normal distribution as a special case. In this paper, the unobservable heterogeneity is restricted to the normal distribution because it does not converge when applied to the multi-episode model with log-normal heterogeneity.

3. Empirical application

This section uses the same data from the National Medical Expenditure Survey (NMES) of the United States in 1987 and 1988 that was analyzed by Deb and Trivedi (1997) to examine the estimated results of the proposed models. The NMES interviews include health insurance coverage, services, and costs *quarterly* of more than 38,000 individuals. This analysis examines a subsample of females over 66 years of age, all of whom are covered by Medicare and not covered by Medicaid. The number of observations examined is 2,308.

As noted by Deb and Trivedi (1997), although the data contain six count variables, the outcome of visits to a physician's office (`OFP`) is used. The mean of `OFP` is 5.86, standard deviation is 6.504, maximum value is 61, and the proportion of 0 is 13.52%. The covariates are the dummy variable for whether self-perceived health is excellent (`EXCLHLTH`, with an average of 7.84%); the dummy variable for whether self-perceived health is poor (`POORHLTH`, with an average of 10.96%); the number of chronic conditions (`NUMCHRON`, with an average of 1.488 and a standard deviation of 1.300); the dummy variable for whether the individual is covered by private health insurance (`PRIVINS`, with an average of 84.62%).

Following Deb and Trivedi (1997), Santos Silva and Windmeijer (2001) and Winkelmann (2004), various models are estimated to specify the model: conventional count data models, such as, NB1 and NB2 models, a Poisson log-normal (PLN) model; hurdle count data models, such as hurdle NB1 and NB2 (HNB); a probit Poisson log-normal (PPLN) model; FM count data models, such as, two components FM NB1 and NB2 (FM2-NB) models; multi-episode count data models, such as the Negbin$_X$ (NBX) model, the Negbin$_X$ with log-normal distributed unobserved heterogeneity (NBX-LN), the Pólya-Aeppli (geometric-Poisson, GP) model, and the Pólya-Aeppli model with log-normal distributed unobserved heterogeneity (GP-LN).

Table I presents the log-likelihood, Akaike's information criteria (AIC), Bayesian information criteria (BIC), and the number of parameters of the above models. From Table I, the maximum log-likelihood is the GP-LN model, the second is the NBX-LN, and the third is the PPLN model. The minimum AIC and BIC is the GP-LN model, the second is the NBX-LN, and the third is the PPLN ($\rho = 0$) model. Information criteria demonstrates that multi-episode models with log-normal distributed unobserved heterogeneities dominate the other models.

Next, following Santos Silva and Windmeijer (2001), a specification test for hurdle and multi-episode models is applied. The parametric test uses a Wald test of whether the coefficients $\widehat{\boldsymbol{\beta}}_1$ of the Poisson distribution part of the multi-episode model are equal to the coefficients $\widehat{\boldsymbol{\beta}}_1^*$ estimated on the

Table I: Log-likelihood, AIC, BIC, and GoF

	log-likelihood	AIC	BIC	K	GoF
Conventional models					
NB1	−6,420.658	12,853.317	12,887.782	6	19.997
NB2	−6,445.669	12,903.338	12,937.803	6	41.242
PLN	−6,440.570	12,893.139	12,927.604	6	82.479
Hurdle models					
HNB1	−6,404.206	12,830.412	12,893.597	11	21.782
HNB2	−6,423.199	12,868.398	12,931.584	11	37.415
PPLN	−6,390.570	12,805.141	12,874.070	12	28.641
PPLN ($\rho = 0$)	-6,390.626	12,803.251	12,866.437	11	27.224
FM models					
FM2-NB1	−6,391.456	12,808.912	12,883.586	13	11.447
FM2-NB2	−6,411.633	12,849.266	12,923.939	13	43.799
Multi-episode models					
NBX	−6,406.833	12,833.666	12,891.107	10	22.747
NBX-LN	−6,389.897	12,801.794	12,864.979	11	10.732
GP	−6,451.241	12,922.481	12,979.923	10	73.502
GP-LN	−6,388.494	12,798.987	12,862.173	11	21.465

Notes: AIC $= -2\ln L + 2K$, BIC $= -2\ln L + K\ln T$, where L is the maximized likelihood, K is the number of parameters of maximum likelihood estimation, T is the number of observations, and GoF indicates the goodness-of-fit test statistics.

binary data with $V = 0$ and $V \geq 1$. The test statistics are 189.589 for the GP-LN, 19.160 for the GP, 69.424 for the NBX-LN, and 2.771 for the NBX. The null hypothesis $\widehat{\boldsymbol{\beta}}_1 = \widehat{\boldsymbol{\beta}}_1^*$ is rejected at the 1% significance level, except for the NBX. Next, we test the single period hypothesis.[3] This tests $\mathrm{E}\left(V - \mathrm{E}\left(R \mid \mathbf{x}, \boldsymbol{\beta}_2\right) \mid V > 0\right) = 0$. The test statistic is 0.755 and its p-value is 0.000; thus, the single period hypothesis is rejected at the 1% significance level.

Finally, the goodness-of-fit (GoF) test used in Deb and Trivedi (1997) is performed. Since the sample with `OFP` greater than 12 days is 11.48% of the total, the test was performed with each cell from 0 to 12 days and 13 days or more combined as one cell (degrees of freedom is 13), as in Deb and Trivedi (1997). From the results in the rightmost column of Table I, for the GP-LN, NBX-LN, NB1, HNB1, and FM-NB1, the null hypotheses are not rejected at the 5% significance level. The results of the information criteria,

[3] This test is based on conditional moment tests (Newey, 1985) using the parameters obtained by the generalized method of moments. Estimation and testing are calculated by Aptech's Gauss 16, but this test is only based on the calculation of the stata ado file by Andrews *et al.* (2017).

Table II: Estimated results

	GP-LN			NBX-LN		
Episodes						
EXCLHLTH	−0.129	(0.115)		−0.158	(0.117)	
POORHLTH	−0.140	(0.115)		−0.084	(0.107)	
NUMCHRON	0.273	(0.023)	***	0.274	(0.023)	***
PRIVINS	0.765	(0.094)	***	0.738	(0.089)	***
constant	0.012	(0.092)		0.037	(0.088)	
σ	0.722	(0.025)	***	0.709	(0.029)	***
Visits						
EXCLHLTH	−1.189	(0.751)		−1.064	(0.577)	*
POORHLTH	0.590	(0.162)	***	0.576	(0.192)	***
NUMCHRON	−0.166	(0.053)	***	−0.200	(0.063)	***
PRIVINS	−0.494	(0.127)	***	−0.519	(0.146)	***
constant	0.179	(0.114)		0.774	(0.134)	***
log-likelihood	−6,388.494			−6,389.897		

Notes: Standard errors are in parentheses; statistically significant at the 1% (***), 5% (**), and 10% (*) levels.

the test statistic for the single period hypothesis, and the GoF test indicate that there is no clear evidence that the GP-LN and NBX-LN are inferior to the other models, but rather superior. Therefore, at least for this data, multi-episode models, particularly the GP-LN model presented in this paper, cannot be ignored.

Table II presents the estimated results of multi-episode count data models (GP-LN and NBX-LN). In multi-episode models, the estimated coefficients resemble one another. The most attractive feature is the interpretation of the six models. In multi-episode models (GP-LN and NBX-LN), the variable NUMCHRON (the number of chronic conditions) increases the number of the first treatment of medical professionals (S), but decreases the number of visits for that treatment (R). Although the estimation results are omitted due to space limitations, in the hurdle models (PPLN with or without correlation), the variable NUMCHRON positively affects the 0/1 decision-making for an individual visiting a physician, and also positively affects the number of visits following the first contact. In the FM models (FM2-NB1 or FM2-NB2), the variable NUMCHRON increases the number of doctor visits of both frequent and infrequent patients.

Similar results are found for the other variables. The variable PRIVINS (a private health insurance dummy) increases the number of first treatments but decreases the number of visits for that treatment. In the hurdle models, PRIVINS significantly increases the probability of the first visit and has a positive effect on successive visits. In the FM models, PRIVINS increases the number of doctor visits of both frequent and infrequent patients. The negative effect of PRIVINS is not found in the hurdle or FM models.

4. Conclusion

For the assessment of health care demand using count data analysis, this paper proposes new multi-episode count data models. Based on Santos Silva and Windmeijer (2001), the Pólya-Aeppli distribution, which assumes a Poisson distribution, is applied for seeking medical care and assumes geometric distribution for treatments. Moreover, this paper introduces normal distributed unobserved heterogeneities for both the Poisson and geometric distributions. Using the USA NMES, the results of the information criteria, model specification tests, and GoF tests demonstrate that the performance of the proposed model is not inferior to conventional models, but rather superior. The estimated coefficients differ from those of the conventional models and the number of treatments decreases in some variables. Of course, it cannot be denied that FM, hurdle, and the multi-episode models all rely on strong distributional assumptions, and determining which one is better depends on the data and the application. Although it is difficult to discern the unique model in health care demand analyses, considering that the same variable can be interpreted differently in different models, it is useful and informative to apply the multi-episode models proposed in this paper.

References

Andrews, D. W. K., W. Kim, and X. Shi (2017) "Commands for testing conditional moment inequalities and equalities" *The Stata Journal* **17 (1)**, 56–72.

Bago d'Uva, T. (2005) "Latent class models for use of primary care: Evidence from a British panel" *Health Economics* **14 (9)**, 873–892.

——— (2006) "Latent class models for utilisation of health care" *Health Economics* **15 (4)**, 329–343.

Deb, P. and A. M. Holmes (2000) "Estimates of use and costs of behavioural health care: A comparison of standard and finite mixture models" *Health Economics* **9 (6)**, 475–489.

Deb, P. and P. K. Trivedi (1997) “Demand for medical care by the elderly: A finite mixture approach” *Journal of Applied Econometrics* **12 (3)**, 313–336.

——— (2002) “The structure of demand for health care: Latent class versus two-part models” *Journal of Health Economics* **21 (4)**, 601–625.

Dhaene, G. and J. M. C. Santos Silva (2012) “Specification and testing of models estimated by quadrature” *Journal of Applied Econometrics* **27 (2)**, 322–332.

Gerdtham, U. (1997) “Equity in health care utilization: Further tests based on hurdle models and Swedish micro data” *Health Economics* **6 (3)**, 303–319.

Gerdtham, U. and P. K. Trivedi (2001) “Equity in Swedish health care reconsidered: New results based on the finite mixture model” *Health Economics* **10 (6)**, 565–572.

Heckman, J. and B. Singer (1984) “A method for minimizing the impact of distributional assumptions in econometric models for duration data” *Econometrica* **52 (2)**, 271–320.

Jemernéz-Martín, S., J. M. Labeaga, and M. Matínez-Granado (2002) “Latent class versus two-part models in the demand for physician services across the European union” *Health Economics* **11 (4)**, 301–321.

Johnson, N., S. Kotz, and A. Kemp (1992) *Univariate Discrete Distributions*, Wiley series in probability and mathematical statistics: Probability and mathematical statistics: John Wiley & Sons.

Mullahy, J. (1986) “Specification and testing of some modified count data models” *Journal of Econometrics* **33 (3)**, 341–365.

Newey, W. K. (1985) “Maximum likelihood specification testing and conditional moment tests” *Econometrica* **53 (5)**, 1047–1070.

Pohlmeier, W. and V. Ulrich (1995) “An econometric model of the two-part decisionmaking process in the demand for health care” *The Journal of Human Resources* **30 (2)**, 339–361.

Santos Silva, J. M. C. and F. A. G. Windmeijer (2001) “Two-part multiple spell models for health care demand” *Journal of Econometrics* **104 (1)**, 67–89.

Winkelmann, R. (2004) “Health care reform and the number of doctor visits—an econometric analysis” *Journal of Applied Econometrics* **19 (4)**, 455–472.