Hurdle and non-hurdle specifications of finite mixture count data models for medical care demand

Hiroaki Masuhara†

abstract

This paper considers finite mixture count data models for medical care demand using hurdle and non-hurdle specifications. The National Medical Expenditure Survey (NMES) example shows that the standard finite mixture negative binomial model is not always the best choice. Moreover, we conclude that the performance of the models considered in this paper is relatively good.

Key words: Medical care demand; Count data; Finite mixture model

1. Introduction

The recent empirical studies on count data models for medical care demand compare the performance of the two most common approaches: the hurdle model and the finite mixture (FM) model. The hurdle (or two-part) model distinguishes the decision to seek care from the level of utilization, and the FM model, namely the finite mixture negative binomial (FMNB) model, discriminates between the frequent and infrequent users' behavior. The hurdle model is sometimes regarded as the approximation of the principal-agent hypothesis (Pohlmeier and Ulrich, 1996). The FMNB model is based on standard non-hurdle (or one-part) specifications. Deb and Trivedi (1997, 2002) showed that the FMNB model is a better approach for medical care demand. Jiménez-Martín *et al*. (2002) criticized the FMNB model on grounds that it is not based on economics but on statistical reasoning. Moreover, they also proposed that in the case of EU nations, the hurdle model is appropriate for visits to specialists and the FMNB model is appropriate for general practitioners.

However, despite the flexibility of the FM model, numerous studies (Deb and Trivedi, 1997, Deb and Trivedi, 2002, Jiménez-Martín *et al*., 2002, Bago d'Uva, 2005a) have only assumed negative binomial (NB) distribution for component densities of the FM model. Based on the models proposed by Bago d'Uva (2005b, 2006), this paper considers various flexible crosssectional finite mixture count data models for medical care demand including the hurdle FM model. In the empirical application, we compare the performance of the models, using the National Medical Expenditure Survey (NMES) data.

2. The model

Consider a finite mixture model. For simplicity, we assume two component densities. Let $y_i, i=1,...,N$ denote a count dependent variable that takes the values $0,1,2,...$ The probability density function of the FM model is expressed by $f(y_i)=\pi_1 f_1(y_i)+(1-\pi_1)f_2(y_i)$, where $f_j(y_i), j=1,2$ are component densities and $0\le\pi_1\le 1$ is a mixing probability.

In count data models, the component densities of the FM model are usually assumed to follow NB distribution, that is, the non-hurdle specification (Deb and Trivedi, 1997, Deb and Trivedi, 2002). This model is the FMNB model. Let x_i denote a $(K\times 1)$ vector of independent variables, and let β_{1j},β_{2j} denote vectors of unknown parameters. Then, the component densities of the FMNB take the following form:

$$f_j(y_i)=\frac{\Gamma(y_i+\psi_{j,i})}{\Gamma(\psi_{j,i})\Gamma(y_i+1)}\left(\frac{\psi_{j,i}}{\psi_{j,i}+\lambda_{j,i}}\right)^{\psi_{j,i}}\left(\frac{\lambda_{j,i}}{\lambda_{j,i}+\psi_{j,i}}\right)^{y_i},$$

where $\lambda_{j,i}=\exp(x_i'\beta_{2j}),\psi_{j,i}=\alpha_{2j}^{-1}\lambda_{j,i}$. The term $\Gamma(g)$ denotes the Gamma function, and α_{2j}^{-1} are overdispersion parameters.

For medical care demand, the FMNB model distinguishes between frequent and infrequent users, whereas the hurdle model focuses on the difference between users and non-users for medical care demand. The first part of the hurdle model specifies the decision to seek care, and the second part models the level of utilization, conditional on some use. There is no rule to use the only NB distribution for the component densities of the FM model and Bago d'Uva (2005b, 2006) has proposed a model using the hurdle specification for the component densities. We call this model the finite mixture hurdle negative binomial (FMHNB) model; its component densities are obtained by $f_j(y_i)=f_{1j}(0)^{d_i}\left[(1-f_{1j}(0))\times f_{Tj}(y_i\mid y_i>0)\right]^{1-d_i}$, where $d_i=1-\min(1,y_i)$. The first part of the hurdle model is of the following logistic type: $f_{1j}(0)=(1+\lambda_{1j,i})^{-1}$ and $\lambda_{1j,i}=\exp(x_i'\beta_{1j})$. The zero-truncated distribution $f_{Tj}(y_i\mid y_i>0)$ of the second part takes the following form:

$$f_{Tj}\left(y_i \mid y_i > 0\right) = \frac{f_{2j}\left(y_i\right)}{1 - f_{2j}\left(0\right)}$$
$$= \frac{\Gamma\left(y_i + \psi_{2j,i}\right)}{\Gamma\left(\psi_{2j,i}\right)\Gamma\left(y_i + 1\right)}\left[\left(\frac{\psi_{j,i}}{\psi_{j,i} + \lambda_{j,i}}\right)^{\psi_{j,i}} - 1\right]^{-1}\left(\frac{\lambda_{2j,i}}{\lambda_{2j,i} + \psi_{2j,i}}\right)^{y_i},$$

where $\lambda_{2j,i} = \exp\left(x_i'\beta_{2j}\right)$ and $\psi_{2j,i} = \alpha_{2j}^{-1}\lambda_{2j,i}$.

When we apply the FMHNB model to cross-sectional data, its FM logistic part leads to the non-identification problem (Teicher, 1960). To avoid this problem, Bago d'Uva (2005b, 2006) has proposed the *panel* FMHNB model. However, since this paper deals with *cross-sectional* FM models, we do not estimate the cross-sectional FMHNB model and consider alternative methods.[1)]

To avoid the non-identification of the FM logistic part, we integrate the first part of both the hurdle components. That is, the coefficients of the first part are the identical for both the components ($\beta_{11} = \beta_{12}$). This is the constrained FMHNB (CFMHNB) model, and it implies that the first part specifies the decision to seek care and the second part models the level of utilization, conditional on some use, *among* frequent and infrequent users.

Another approach is to assume different distributions for two component densities. That is, one component assumes the hurdle model, but the other, the NB model. The resultant model is the finite mixture hybrid model of the HNB and NB (FMHybrid). Although the FMHybrid and CFMHNB models contain a logistic part, these models do identify the parameters. Therefore, we apply all the models, except for the FMHNB model, to medical care data in the next section.[2)]

3. An application to medical care demand

We present the results of the application of the models, using the NMES data, which was originally used by Deb and Trivedi (1997). This data contains 4,406 observations. The dependent variable is the number of physician office visits (OFP); the mean OFP is 5.77 (15.5% of the observations are zero observations). The minimum and maximum values of this variable are 0 and 89, respectively. The explanatory variables are as follows: (1) health status measures—the number of chronic conditions (NUMCHRON), a dummy for self-perceived excellent health (EXCLHLTH), self-perceived poor health (POORHLTH), and the physical limitations in daily living (ADLDIFF); (2) socioeconomic variables—exact age (AGE), years of schooling (SCHOOL), log of annual family income (FAMINC), a dummy for northeastern residents (NOREAST), midwestern residents (MIDWEST), western residents (WEST), African-Americans (BLACK), male (MALE), marital status (MARRIED), employment status (EMPLOYED), private insurance status (PRIVINS), and public insurance status (MEDICAID).

Detailed descriptions of the variables and summary statistics can be found in Deb and Trivedi (1997).

To evaluate the performance and fit of the three models, we apply the goodness of fit (GoF) test (Andrews, 1988, Deb and Trivedi, 2002), information criteria, and likelihood-based model selection tests. We consider that the range of y_i is broken into $\hat{J}$ cells, that is, $\{0\},\{1\}$, $,...,\{\hat{J}-1\}$ and $\{\hat{J},\hat{J}+1,...\}$. Let $d_{i\hat{j}}(y_i)$ be an indicator function with $d_{i\hat{j}}=1$ if y_i falls in the $\hat{j}$ th cell and $d_{i\hat{j}}=0$ otherwise. Moreover, let $p_{i\hat{j}}(x_i,\theta)$ denote the predicted probability that observation i falls in the $\hat{j}$ th cell, where θ is a known parameter vector. Then, under the null hypothesis that the density is correctly specified, that is, that $p_{i\hat{j}}(x_i,\theta)$ obtains the correct probabilities, the test static $\mathrm{GoF}=m(\hat{\theta})'\hat{V}_m^{-1}m(\hat{\theta})$ is chi-square distributed with $\hat{J}$ degrees of freedom, where $m(\hat{\theta})=\sum_{i=1}^{N}d_i(y_i)-p_i(x_i,\hat{\theta})\sim(\hat{J}+1)\times 1$, $d_i(y_i)=\left[d_{i0}(y_i),...,d_{i\hat{J}}(y_i)\right]'$, $p_i(x_i,\hat{\theta})=\left[p_{i0}(x_i,\hat{\theta}),...,p_{i\hat{J}}(x_i,\hat{\theta})\right]'$, $\hat{\theta}$ is the estimated parameter of θ, and $\hat{V}_m$ is the asymptotic variance matrix of $m(\hat{\theta})$. In the GoF test, presented in Table 1, the CFMHNB model shows good performance around zero but poor performance for high counts. Conversely, the FMNB model shows good performance for high counts but has poor performance around zero. The performance of the FMHybrid model is superior both around zero as well as for high counts. Deb and Trivedi (2002) report inferior performance of the hurdle model for high counts and inferior performance of the FMNB model around zero. This tendency is observed in our analysis as well; however, the FMHybrid model does not show inferior performance in a specific region. Next, in the information criteria presented in Table 2, the Akaike information criterion (AIC) value of the FMHybrid model, which shows good performance on the GoF test, is the smallest of the three. However, the minimum value of the Bayesian information criterion (BIC) is that of the FMNB model, which has the smallest parameters. Further, we use likelihood-based tests for model selection. The log-likelihood ratio (LR) test statistic for the FMHybrid model against the FMNB model is 85.424.

Since the CFMHNB model does not completely nest the FMNB and FMHybrid models, we do not apply this LR test and instead use Vuong's (1989) test. Let $f^{(1)}$ be the likelihood of model 1 and $f^{(2)}$ be that of model 2. Under the null hypothesis that both the models are equivalent ($E\left[\ln f^{(1)} - \ln f^{(2)}\right] = 0$), the test statistic

$$\frac{\sum_{i=1}^{N}\left[\ln f^{(1)} - \ln f^{(2)}\right]}{\sqrt{\sum_{i=1}^{N}\left[\left(\ln f^{(1)} - \ln f^{(2)}\right)^2 - \left(\sum_{i=1}^{N}\ln f^{(1)} - \ln f^{(2)}\right)^2 / N\right]}}$$

follows a standard normal distribution. If the statistic exceeds the critical value c, model $f^{(1)}$ is better than model $f^{(2)}$; if the statistic is smaller than $-c$, model $f^{(2)}$ is better than model $f^{(1)}$. The test statistic for the CFMHNB model against the FMNB (FMHybrid) model is 3.458 (−0.280), and its value at the significant level is 0.001 (0.616).[3)] Although, on the basis of the GoF, AIC, and the LR and Vuong's test results, the performance of the FMHybrid model is relatively good, there is no strong evidence to suggest that the FMHybrid model is the best model. However, we do conclude that the FMNB model is not always the best model to analyze this data.

Table 1: Goodness of fit test

	GoF		
cell	CFMHNB	FMHybrid	FMNB
0, 1+	0.000	0.786	5.150 *
0−1, 2+	2.270	0.827	5.191 *
0−2, 3+	2.413	2.974	9.388 *
0−3, 4+	2.766	3.069	10.108 *
0−4, 5+	5.313	3.607	10.119 *
0−5, 6+	13.988 *	7.688	11.444 *
0−6, 7+	19.106 *	9.580	11.577
0−7, 8+	20.478 *	9.919	11.727
0−8, 9+	21.533 *	10.483	11.734
0−9, 10+	24.777 *	13.998	12.930
0−10, 11+	24.899 *	14.199	12.969
0−11, 12+	25.563 *	15.465	13.349
0−12, 13+	25.994 *	15.555	13.687
0−13, 14+	27.369 *	15.888	14.030
0−14, 15+	27.768 *	17.702	15.986
0−15, 16+	28.053 *	17.705	16.003

Notes: GoF is a goodness of fit test statistic. * denotes statistical significance at the 10% level. CFMHNB, FMHybrid, and FMNB represent the constrained FMHNB, the finite mixture hybrid of the HNB and NB, and the finite mixture negative binomial models, respectively.

Table 2: Results of the three finite mixture count data models

	CFMHNB			FMHybrid			FMNB	
	0/1	1+		infrequent		frequent	infrequent	frequent
		infrequent	frequent	0/1	1+			
EXCLHLTH	−0.329 **	−0.312 ***	−1.095 **	−0.212	−0.367 ***	−0.698	−0.250 ***	−0.760
	(0.142)	(0.090)	(0.547)	(0.195)	(0.086)	(0.614)	(0.058)	(0.774)
POORHLTH	0.071	0.227 ***	0.633 **	0.284	0.279 ***	0.112	0.232 ***	0.068
	(0.169)	(0.053)	(0.252)	(0.405)	(0.054)	(0.632)	(0.067)	(0.723)
NUMCHRON	0.557 ***	0.153 ***	0.100 *	0.861 ***	0.151 ***	0.116	0.186 ***	0.137
	(0.053)	(0.013)	(0.059)	(0.112)	(0.012)	(0.105)	(0.013)	(0.110)
ADLDIFF	−0.188	0.028	0.445 **	−0.289	0.035	0.548 *	−0.015	0.534 **
	(0.130)	(0.048)	(0.191)	(0.214)	(0.046)	(0.309)	(0.043)	(0.272)
NOREAST	0.129	0.073	0.149	0.170	0.078	0.355	0.083 *	0.179
	(0.125)	(0.053)	(0.242)	(0.199)	(0.052)	(0.475)	(0.048)	(0.490)
MIDWEST	0.101	−0.002	−0.053	0.218	−0.013	0.130	0.017	0.058
	(0.115)	(0.046)	(0.293)	(0.183)	(0.044)	(0.369)	(0.039)	(0.346)
WEST	0.202	0.089 *	0.196	0.174	0.089 *	0.381	0.089 *	0.237
	(0.134)	(0.051)	(0.216)	(0.229)	(0.051)	(0.532)	(0.049)	(0.488)
AGE	0.190 **	−0.031	−0.241	0.514 ***	−0.027	−0.814 ***	0.028	−0.614 **
	(0.081)	(0.030)	(0.154)	(0.181)	(0.028)	(0.240)	(0.027)	(0.250)
BLACK	−0.327 **	−0.102	0.374	−0.273	−0.036	−1.052	−0.080	−1.067
	(0.133)	(0.069)	(0.279)	(0.211)	(0.063)	(0.801)	(0.070)	(1.069)
MALE	−0.464 ***	−0.063	0.066	−0.587 ***	−0.061	0.100	−0.136 ***	0.123
	(0.099)	(0.040)	(0.164)	(0.163)	(0.039)	(0.284)	(0.035)	(0.260)
MARRIED	0.247 **	−0.019	−0.232	0.408 **	−0.024	−0.478	0.047	−0.509
	(0.104)	(0.041)	(0.169)	(0.170)	(0.040)	(0.348)	(0.035)	(0.332)
SCHOOL	0.054 ***	0.009	0.085 ***	0.062 ***	0.012 **	0.128	0.014 ***	0.152 *
	(0.013)	(0.006)	(0.029)	(0.023)	(0.006)	(0.081)	(0.005)	(0.084)
FAMINC	0.007	0.001	−0.015	−0.004	−0.002	0.006	0.000	−0.005
	(0.018)	(0.007)	(0.022)	(0.024)	(0.007)	(0.018)	(0.005)	(0.020)
EMPLOYED	−0.012	−0.088	0.583 **	0.027	−0.065	0.213	−0.059	0.395
	(0.145)	(0.066)	(0.238)	(0.235)	(0.070)	(0.586)	(0.056)	(0.750)
PRIVINS	0.762 ***	0.204 ***	0.514	0.723 ***	0.222 ***	2.112	0.256 ***	2.976
	(0.117)	(0.061)	(0.339)	(0.183)	(0.055)	(1.598)	(0.053)	(2.639)
MEDICAID	0.554 ***	0.320 ***	−0.165	0.872 ***	0.300 ***	−2.108	0.353 ***	−3.160
	(0.181)	(0.071)	(0.246)	(0.324)	(0.066)	(1.819)	(0.062)	(3.375)
CONSTANT	−1.475 **	1.362 ***	2.657 *	−3.908 ***	1.340 ***	4.322 **	0.778 ***	1.918
	(0.646)	(0.252)	(1.508)	(1.435)	(0.232)	(1.738)	(0.222)	(2.544)
α		2.900 ***	9.368 ***		3.322 ***	22.225 ***	3.493 ***	16.969 **
		(0.188)	(1.932)		(0.227)	(1.258)	(0.288)	(6.992)
π		0.907 ***		0.907 ***			0.913 ***	
		(0.027)		(0.025)			(0.029)	
log-likelihood		−12,037.960			−12,035.281		−12,077.993	
AIC		24,183.920			24,178.561		24,229.987	
BIC		24,529.019			24,523.660		24,466.444	

Notes: Standard errors are in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively. CFMHNB, FMHybrid, and FMNB represent the constrained FMHNB, the finite mixture hybrid of the HNB and NB, and the finite mixture negative binomial models, respectively. 0/1, 1+, frequent, and infrequent denote the logistic part, 0 truncated part, frequent users' component density, and infrequent users' component density, respectively. $AIC = -2\ln L + 2K$ and $BIC = -2\ln L + K\ln N$, where L is the maximized likelihood, K is the number of parameters, and N is the number of observations (N = 4,406).

Table 2 shows the estimated results. We need to be careful while interpreting the FMNB model and the other two models. For instance, although private insurance status (PRIVINS) does not affect the number of visits made by frequent users to the physician office, it does have a statistically positive effect on the behavior of infrequent users in all the models. However, the interpretation of this coefficient is not same; in the FMNB model, PRIVINS increases the total number of visits; in the FMHybrid model it increases the probability of seeking care and then increases the conditional positive number of visits; in the CFMHNB model it increases the probability of frequent and infrequent users seeking medical care and then increases the conditional positive number of infrequent users' visits. An interesting feature of the results is that the estimated component proportion of the three models is approximately 0.91. Moreover, the frequent users' behavior of the FMHybrid model closely resembles that of the FMNB model and the component proportion is around 0.09 for both models. This implies that the HNB specification of infrequent users' behavior in the FMHybrid model approximates the NB specification of the same in the FMNB model.

4. Concluding remarks

This paper considers alternative finite mixture models for medical care demand and compares the performance of the three models, using the NMES data employed by Deb and Trivedi (1997). The results are as follows: the GoF illustrates that the FMHybrid model shows good performance for both high and low count regions; the AIC value of the FMHybrid is the smallest, but its BIC value is not; Vuong's test implies that the FMNB model is strongly rejected. These results imply that although the FMHybrid shows good performance, it is difficult to determine the best model among the three. However, we do concludes that the FMNB model used in former studies is not always the best choice. Since the three models considered in this paper have other interpretations, we need to be careful while selecting the model. Therefore, not only is it important to compare the standard FMNB model with the HNB model but also to include the two alternative models in the comparison.

Notes

[†] This study is supported by a Grant-in-Aid for Young Scientists (B) 21730215 from the Ministry of Education, Culture, Sports, Science and Technology of the Japanese Government. All remaining errors are mine.

[1)] Although Bago d'Uva (2006) has referred to the following models in a panel data framework, the estimated results are not shown.

[2)] In the following analysis, since the hurdle model is statistically rejected, we omit the result of the hurdle model due to space constraints.

[3)] Since the Vuong's test statistics for the FMNB, FMHybrid, and CFMHNB models against the HNB model are 1.859, 4.457, and 4.253, respectively, the HNB model is strongly rejected.

References

Andrews, D.W.K. (1988) "Chi-square Diagnostic Tests for Econometric Models: Introduction and Applications", *Journal Econometrics*, vol. 37, No. 1, pp. 135-156.

Bago d'Uva, T. (2005a) "Latent Class Models for Use of Primary Care: Evidence from a British Panel", *Health Economics*, vol. 14, No. 9, pp. 873-892.
Bago d'Uva, T. (2005b) "Latent Class Models for Health Care Utilization", HEDG Working Paper 05/01, University of York.
Bago d'Uva, T. (2006) "Latent Class Models for Utilisation of Health Care", *Health Economics*, vol. 15, No. 4, pp. 329-343.
Deb, P. and P.K. Trivedi (1997), "Demand for Medical Care by the Elderly: a Finite Mixture Approach", *Journal of Applied Econometrics*, vol. 12, No. 3, pp. 313-336.
Deb, P. and P.K. Trivedi (2002) "The Structure of Demand for Health Care: Latent Class Versus Two-part Models", *Journal of Health Economics*, vol. 21, No. 4, pp. 601-625.
Jiménez-Martín, S., J.M. Labeaga and M. Martínez-Granado (2002) "Latent Class Versus Two-part Models in the Demand for Physician Services across the European Union", *Health Economics*, vol. 11, No. 4, pp. 301-321.
Pohlmeier, W. and V. Ulrich (1995) "An Econometric Model of the Two-part Decisionmaking Process in the Demand for Health Care", *Journal of Human Resources*, vol. 30, No. 2, pp. 339-361.
Teicher, H. (1960) "On the Mixture of Distributions", *Annals of Mathematical Statistics*, vol. 31, No. 1, pp. 55-73.
Ueda, N. and R. Nakano (1998) "Deterministic Annealing EM Algorithm", *Neural Networks*, vol. 11, No. 2, pp. 271-282.
Vuong, Q.H. (1989) "Likelihood Ratio Tests for Model Selection and Non-nested Hypotheses", *Econometrica*, vol. 57, No. 2, pp. 307-333.

Appendix

A. Deterministic annealing EM algorithm

A recommended procedure to estimate finite mixture models is the expectation-maximization (EM) algorithm. However, the EM algorithm is sensitive to initial values and contains problems of local maxima. Ueda and Nakano (1998) propose the deterministic annealing EM (DAEM) algorithm. This algorithm has advantages of moderate computation time and robustness to initial values, and attains the global maximum with a high probability. To avoid problems of local maxima, we use the DAEM algorithm. In the FMNB model, let the parameter vectors of each density be $\theta' = [\theta_1', \theta_2']$, where $\theta_j' = [\beta_{2j}', \alpha_{2j}]$ and $\Theta' = [\theta', \pi_1]$. Then, the DAEM algorithm is stated below.

1. Set $\delta^{(0)} \in (0,1]$, $\Theta^{(0)}$, and $t \leftarrow 0$, where $\delta^{(t)}$ is an annealing parameter at the t th iteration

step.

2. Iterate the following EM steps.

E-step. Let $\Theta^{(t)}$ be a parameter at the t th iteration step. Calculate the conditional expectation as follows:

$$Q^{(t)}\left(\Theta \mid \Theta^{(t)}\right)=\sum_{i=1}^{N} z_{i1} \ln\left[\pi_1 f_1\left(y_i;\theta_1\right)\right]+\left(1-z_{i1}\right)\ln\left[\left(1-\pi_1\right) f_2\left(y_i;\theta_2\right)\right],$$

where

$$z_{i1}=\frac{\left[\pi_1^{(t)} f_1\left(y_i;\theta_1^{(t)}\right)\right]^{\delta^{(t)}}}{\left[\pi_1^{(t)} f_1\left(y_i;\theta_1^{(t)}\right)\right]^{\delta^{(t)}}+\left[\left(1-\pi_1^{(t)}\right) f_2\left(y_i;\theta_2^{(t)}\right)\right]^{\delta^{(t)}}}$$

is a posterior density.

M-step. Maximize $Q^{(t)}$ function. Further, $t \leftarrow t+1$.

3. Set $\delta^{(t)}$, where $\delta^{(t-1)}<\delta^{(t)}<1$.

4. Stop if Θ converges, and $\delta^{(t)} \simeq 1$; otherwise, repeat from step 2.

When $\delta^{(t)}=1$, the DAEM algorithm is the same as the EM algorithm. If $\delta^{(t)}$ is sufficiently small, the posterior density z_{i1} is the same across the entire sample, that is, the posterior density is like a uniform distribution. This uniform characteristic leads the parameters to the global maxima. We set $\delta^{(t)}>\delta^{(t-1)}$ in the next step. The M step of the EM algorithm seeks the neighborhood of the $t-1$ th maxima. If we choose the good $\delta^{(t)}$, this t th maxima is the global maxima. The name annealing is derived from this slowly increasing step of $\delta^{(t)}$. Continue this process until Θ converge and $\delta^{(t)}$ is almost one. The uniform feature of a posterior distribution ensures the global maximum with a high probability.

B. Monte Carlo experiment

This appendix presents some results of the Monte Carlo experiments. The data generating process of the NB component is obtained by $y \sim \text{NB1}$ with $\lambda_{2j}=\exp\left(x'\beta_{2j}\right)$ and α_{2j}, $j=1,2$, where $x=\left[1,x_1\right]'$ and $x_1 \sim N\left(0,0.25\right)$. The first part of the HNB component is $d\left(y\right) \sim \text{Logit}$ with $\lambda_{1j}=\exp\left(x'\beta_{1j}\right)$ and the second part is given by $y \mid y>0 \sim \text{truncated NB1}$ with $\lambda_{2j}=\exp\left(x'\beta_{2j}\right)$ and α_{2j}. The trued values of the parameters of the first component are $\beta_{11}=\left[1,1\right]'$, $\beta_{21}=\left[1,0.5\right]'$, and $\alpha_{21}=0.5$ and those of the second are $\beta_{12}=\left[1.5,1\right]'$,

Table B1: Monte Carlo results

		FMHNB			CFMHNB			FMHybrid			FMNB		
	Truth	$N=500$	$N=1{,}000$	$N=2{,}000$	$N=500$	$N=1{,}000$	$N=2{,}000$	$N=500$	$N=1{,}000$	$N=2{,}000$	$N=500$	$N=1{,}000$	$N=2{,}000$
β_{11}	1	0.492	0.571	0.414	-0.009	0.013	-0.013	0.071	0.009	0.002			
		(1.411)	(1.269)	(0.957)	(0.112)	(0.078)	(0.049)	(0.341)	(0.254)	(0.168)			
	1	0.589	0.347	0.264	0.001	-0.012	-0.009	0.030	-0.002	-0.010			
		(1.649)	(1.298)	(0.906)	(0.220)	(0.158)	(0.107)	(0.549)	(0.306)	(0.232)			
β_{12}	1	-0.010	0.006	0.026	0.073	0.013	0.025	0.078	0.015	0.015	-0.020	0.013	0.011
		(0.317)	(0.216)	(0.186)	(0.252)	(0.186)	(0.126)	(0.293)	(0.201)	(0.139)	(0.269)	(0.162)	(0.114)
	0.5	0.084	0.025	-0.001	-0.024	-0.007	-0.012	-0.028	-0.004	0.004	0.029	-0.003	-0.013
		(0.422)	(0.214)	(0.149)	(0.268)	(0.174)	(0.104)	(0.282)	(0.183)	(0.136)	(0.187)	(0.130)	(0.086)
α_{21}	0.5	0.123	0.111	0.095	0.320	0.074	0.098	0.259	0.172	0.085	0.103	0.048	0.035
		(0.847)	(0.721)	(0.582)	(1.198)	(0.656)	(0.568)	(1.121)	(0.838)	(0.473)	(0.774)	(0.450)	(0.255)
β_{21}	1.5	0.504	0.155	0.012									
		(1.395)	(0.828)	(0.470)									
	1	0.164	0.043	0.074									
		(1.051)	(0.689)	(0.399)									
β_{22}	2.5	0.009	0.004	0.007	0.010	0.004	0.002	0.011	0.004	0.002	-0.002	0.003	0.004
		(0.065)	(0.047)	(0.045)	(0.064)	(0.049)	(0.028)	(0.048)	(0.035)	(0.023)	(0.057)	(0.035)	(0.027)
	0.5	0.019	0.006	-0.006	0.002	-0.001	-0.001	0.000	0.005	0.005	0.000	-0.002	-0.003
		(0.089)	(0.056)	(0.043)	(0.079)	(0.040)	(0.033)	(0.055)	(0.037)	(0.027)	(0.055)	(0.038)	(0.027)
α_{22}	1.5	-0.043	-0.028	-0.056	-0.108	-0.029	-0.024	-0.079	-0.008	-0.014	-0.008	-0.015	-0.030
		(0.386)	(0.263)	(0.248)	(0.401)	(0.332)	(0.184)	(0.320)	(0.207)	(0.171)	(0.402)	(0.224)	(0.193)
π_1	0.3	0.018	0.007	0.003	0.032	0.007	0.007	0.020	0.010	0.004	0.005	0.007	0.005
		(0.080)	(0.063)	(0.059)	(0.094)	(0.061)	(0.039)	(0.075)	(0.045)	(0.032)	(0.087)	(0.044)	(0.032)

Notes: The mean bias (BIAS) values are the figures without parenthesis. The root mean squared errors (RMSE) are presented within parenthesis. FMHNB, CFMHNB, FMHybrid, and FMNB are the finite mixture hurdle negative binomial, the constrained FMHNB, the finite mixture hybrid of the HNB and NB, and the finite mixture negative binomial models, respectively.

$\beta_{22} = [2.5, 0.5]'$, and $\alpha_{22} = 1.5$. The mixing probability of the first component π_1 is 0.3.

Table B1 presents the Monte Carlo results of four finite mixture models for $N = 500$, 1,000, and 2,000. For each set of parameter estimators, we make 100 replications and find the mean and standard deviation of the parameter estimates. The results of the three models—FMNB, FMHybrid, and CFMHNB—show that the parameter estimates are unbiased. However, the estimates of overdispersion parameters α_{2j} vary somewhat in a small sample. The bias of α_{2j} decreases with an increase in the sample size. The results of the FMHNB show that the parameter estimates of the second part (β_{21} and β_{22}) of each component are unbiased; however, those of the first part (β_{11} and β_{12}) are not always unbiased for all the components.

Moreover, the bias is rather large in a small sample, especially in the case of $N = 500$. This is an identification problem of a finite mixture binary model, as pointed out by Teicher (1963) and Bago d'Uva (2006). It is widely known that the parameters of the cross-section finite mixture binary model have not been estimated well for the last five decades. At a glance, in the case of $N = 2{,}000$, we may observe the beginning of unbiasedness and that convergence is very slow. However, this result is obtained by setting true values for the initial values of coefficients and by using the simulated annealing (SA) algorithm. Although the SA algorithm is robust to initial values, the searching process is limited to a local area. In Monte Carlo simulations, we search a wide area by setting a big scale parameter. However, the result of the FMHNB model is not stable and does not reject the possibility of the identification problem. In fact, when we regress the FMHNB model using NMES data, the absolute value of the binary part is rather large. Therefore, we treat the binary part of the FMHNB model as one component of the CFMHNB model.

（ますはら　ひろあき　医療経営学科　講師）