Mark Lett
DOI 10.1007/s11002-013-9272-z

A simple heuristic for obtaining pareto/NBD parameter estimates

Pablo Marshall

Abstract In an influential study, Schmittlein et al. (1987) proposed the pareto/negative binomial distribution (P/NBD) model to predict purchase behavior of customers. Despite its recognized relevance, this model has some drawbacks as follows: (1) it does not allow a zero transaction rate, (2) it assumes convenient but not necessarily realistic gamma distributions for the transaction and drop-out rates across customers, and (3) the estimation procedure requires complicated computations. The purpose of this study is to relax the assumption that purchases and drop-out rates are distributed according to a gamma distribution and propose a simple estimation procedure for the individual parameters that can be applied even if the number of customers is large. A simulation exercise and empirical applications to real datasets compare the simple model proposed with the P/NBD model. The results show that the simple procedure is better in cases where the number of transactions and/or the observation period is large.

Keywords Prediction consumer behavior · Nonparametric model · Pareto NBD model

1 Introduction

The need to understand customer behavior and the interest of managers to focus on customers who can deliver long-term profits has changed the main purpose of marketers' activities from acquisition to retention. Many companies, nowadays, face a database containing information on the frequency and timing of transactions for a list of customers, for the purpose of forecasting their future purchasing behavior at the individual level. Manager decisions oriented to the growth and retention of customers are usually based on these predictions.

In an influential study, Schmittlein et al. (1987) proposed the pareto/negative binomial distribution (P/NBD) model to analyze and predict the purchase behavior of customers at the individual level in a non-contractual scenario. Under a non-contractual

P. Marshall (✉)
Escuela de Administración, Pontificia Universidad Católica de Chile, Av Libertador Bernardo O Higgins 340, Santiago, Chile
e-mail: pmarshall@uc.cl

setting (e.g., retail store, supermarket, mail order), consumers do not declare that they have become inactive, but simply stop doing transactions with the firm. Thus, the time at which a customer becomes inactive is unobserved by the firm and the single evidence for this is an unusually long interval since the last recorded purchase. The challenge facing the analyst is how to differentiate between those customers who have ended their relationship with the firm from those who are in the middle of a long interval between transactions.

The P/NBD model estimates the probability of being active and the expected number of future transactions for each customer. The model assumes that at time t=0 the individuals are active, and (1) while active, the number of transactions made by a customer follows a Poisson probability distribution, (2) the lifetime of a customer has an exponential probability distribution, (3) customers are heterogeneous in both their activity levels and their lifetimes, (4) transactions and the drop-out rates across individuals follow a gamma distributions, and (5) transactions and the drop-out rates are independent at the individual level.

Despite its recognized relevance for academics and practitioners, the P/NBD model has led, in recent years, to (1) several simplifications or alternative estimation procedures in order to avoid complicated computation (Fader et al. 2005; Ma and Liu 2007), and (2) some extensions in order to incorporate more realistic assumptions for the probability distribution of the transactions and drop-out rates (Singh et al. 2009; Bemmaor and Glady 2012).

The objective of this study is to propose an extremely simple nonparametric estimation procedure of the P/NBD model which leads to better in-sample and—also, under particular conditions—out-of-sample predictions. The model we consider in this study relaxes the assumption that purchases and drop-out rates are distributed according to a gamma probability distribution. In our specification, we do not assume any probability distribution for these coefficients and hence we call it simple nonparametric (simple-NP) model. We maintain the other assumptions of the P/NBD model as follows: at the individual level, transactions and lifetimes are independent and follow Poisson and exponential distributions with customer-specific coefficients. As a consequence, customers with purchase and/or drop-out rates equal to zero are allowed, something that cannot happen in the gamma model.

We expect a priori our simple model to fit the data best in scenarios where the estimation period is large because each customer has enough information to predict his/her behavior, and for customers with very low or very large purchase rates since the gamma model, which does not allow a zero value and has long tails on the right, might have a conflict and hence a poor fit. From a managerial point of view, our hypothesis has two implications as follows: an easier estimation procedure to fit the model in order to analyze large customers databases and an improvement in the prediction performance of the buying behavior of those premium customers who buy very frequently; that is, precisely the ones who are most important for managers.

2 The models P/NBD and simple-NP

The P/NBD model (Schmittlein et al. 1987) assumes that, while active, customers make transactions with the firm of interest that are randomly distributed in time with some customer-specific transaction rate. Then, the number of transactions in a period of time has a Poisson probability distribution; an assumption with substantial support in consumer markets (Ehrenberg 1988). Customers are assumed to be active at the beginning of the observation period and remain active for some unobserved timeframe which is customer-specific. After that, the customer no longer purchases from the firm. This customer drop-out phenomenon is assumed to occur randomly in time; thus, the time spent as an active customer is an exponential random variable. The purchase rate and the drop-out rate are assumed to vary independently across customers.

Figure 1 represents a possible transaction scheme for a customer. The observation period is $(0, T)$ and the x represents transactions. The customer is active at time $t=0$ and makes transactions until time t_x. Between times t_x and T, the customer might become inactive but the exact time is not observable. The time at which a customer becomes inactive might be beyond T.

The P/NBD model also assumes that the parameters of the Poisson and exponential distributions, the frequency rate of the transactions, and the drop-out rate follow independent gamma distributions. Unlike the P/NBD model, however, we do not assume the intensity of the transactions and the drop-out rate to follow any probability distribution. We allow these rates to vary freely across customers. We expect the P/NBD model to perform better if customers carry out a small number of transactions, while the nonparametric specification in this paper may achieve better results if the amount of available information for each customer is large. The gamma distribution assumption has two drawbacks as follows: first, it does not allow a zero transaction rate although the evidence illustrates that in many applications a large proportion of customers' number of transactions is zero (Batislam et al. 2007); and second, if the number of transactions for each customer is large, there is no need for an a priori distribution, particularly when the number of customers is large. The obvious problem with the simple-NP model is that it is not parsimonious. If, as usual, the number of customers is very large, the number of coefficients that have to be estimated, the transactions, and the drop-out rates is also very large. In the following lines, we show a simple procedure that makes the estimation possible even if the number of customers is very large.

Let (x_i, t_{xi}, T_i) be the information of customer i, the number of transactions, the time of the last transaction, and the total observation period, respectively. Following the P/NBD model, we assume that the number of transactions x_i follows a Poisson distribution with parameter or intensity λ_i and the time until death or inactivity is a latent variable with probability distribution exponential and parameter μ_i. Some algebra shows that the likelihood function of the model has the following form (Fader and Hardie 2005).

$$p(x_i, t_{x_i}, T_i \mid \lambda_i, \mu_i) = \frac{\lambda_i^{x_i}}{\lambda_i + \mu_i} \left(\mu_i e^{-(\lambda_i + \mu_i) t_{x_i}} + \lambda_i e^{-(\lambda_i + \mu_i) T_i} \right) \tag{1}$$

A usual estimation procedure for the individual coefficients (λ_i, μ_i) maximizes (1). An analytic solution for these maximum likelihood estimators (MLE) for (λ_i, μ_i) is not possible but we can show, maximizing each of the two additive components in (1) and

Fig. 1 Transaction scheme of a customer

assuming x_i>0, that the MLE satisfy the conditions.

$$\frac{x_i}{T_i} \leq \quad \widehat{\lambda}_i \quad \leq \frac{x_i}{t_{x_i} + \frac{t_{x_i}}{x_i}}, \qquad 0 \leq \quad \widehat{\mu}_i \quad \leq \frac{1}{t_{x_i} + \frac{t_{x_i}}{x_i}} \tag{2}$$

The two lower bounds in Eq. (2) correspond to the case where customer i is still active at time T_i. In that case, the transaction rate is estimated as the number of transactions divided by the time span where the customer was active. As he/she was active for the whole observation period, the estimated drop-out rate is zero: the most probable value. The upper bounds in (2), on the other hand, can be interpreted as the solution when the lifetime of a customer is between t_{xi} and T_i. The denominator $\left(t_{x_i} + t_{x_i}/x_i\right)$ is an estimate of the lifetime of customer i, which is the time until the last transaction plus the expected time until a new transaction. That is, after the last transaction at t_{xi}, customer i is considered active until the expected value of a new transaction.

These analyses lead us to propose the following estimators of (λ_i, μ_i):

$$\widehat{\lambda}_i \quad = \frac{x_i}{t_{x_i} + \frac{t_{x_i}}{x_i}}, \quad \widehat{\mu}_i = \frac{1}{t_{x_i} + \frac{t_{x_i}}{x_i}} \qquad \text{if} \quad t_{x_i} + \frac{t_{x_i}}{x_i} < T_i,$$

and

$$\widehat{\lambda}_i \quad = \frac{x_i}{T_i}, \quad \widehat{\mu}_i = 0 \qquad \text{if} \quad t_{x_i} + \frac{t_{x_i}}{x_i} \geq T_i$$

These estimators have a clear intuition and are extremely easy to compute. Actually, they even need no optimization procedure. The special case x_i=0 means the customer was active at time 0 but became inactive immediately after, or the corresponding transaction rate of the customer is 0. For these cases, we assume $\widehat{\lambda}_i = \widehat{\mu}_i = 0$.

The proposed model, like the P/NBD model, estimates the probability of being active and the expected number of future transactions for each customer. Given an active customer at time t=0, and suppressing the customer index for convenience, the expected number of transactions in any interval of length t is as follows:

$$E[x(t)] = \begin{cases} \frac{\lambda}{\mu}\left[1 - e^{-\mu t}\right], & \text{if } \mu > 0 \\ \lambda\, t, & \text{if } \mu = 0 \end{cases} \tag{3}$$

While the probability that a customer is active or alive at time T is given as follows:

$$Pr[\text{Active at time}\, T] = 1/\left[1 + (\mu/(\lambda + \mu))\left(e^{(\lambda+\mu)(T-t_x)} - 1\right)\right] \tag{4}$$

To make predictions beyond the observation period, say (T, T+t), the product of (3) and (4) has to be considered since (3) assumes the customer is active at the beginning of the time interval. For more details on these formulas, see Fader and Hardie (2005).

In order to analyze the performance of the estimation procedure presented, we considered a simulation study similar to the one proposed by Fader et al. (2005). We simulated the buying behavior of customers assuming gamma distributions for the transaction and drop-out rates. For each of these two gamma distributions, we considered two values for each of the two parameters, namely the shape and the scale parameter. These combinations define a total of 16 scenarios and for the parameter values we used those proposed by Fader et al. (2005) in their simulation: for the shape parameters 0.25 and 0.75 and for the scale parameters 5 and 15. In each scenario, we simulated 5.000 customers and set the observation periods at 26, 52, and 78. The number of different and independent simulation scenarios was $2^4 \times 3 = 48$, and in each of these, we estimated the model and predicted a validation out-of-sample period.

In each simulation, we computed the mean absolute percent error (MAPE) for the out-of-sample predictions at 13, 26, 39, and 52 periods. The MAPE, across all simulated models, is 11.3 % with a minimum of 3.3 % and a maximum of 27.9 %. Furthermore, approximately one fifth of the cases have an MAPE of less than 5 % and half have an MAPE smaller than 10 %.

Table 1 presents the results concerning how MAPE varies with the period observed, the marginal expected values of λ, and the marginal expected values of μ which can be inferred from the shape and scale of the gamma distributions defined. As the observation period increases from 26 to 78, MAPE values decrease from 14.5 to 9.0 %. Similarly, as the mean value of λ increases from 0.02 to 0.15, MAPE decreases from 14.4 to 8.9 %. Finally, as the mean value of the μ parameter across customers increases from 0.02 to 0.15, MAPE decreases from 17.5 to 5.2 %. The mean value of the μ parameter is the most sensitive of these three components.

To understand these results, we also simulated previous scenarios with a drop-out rate equal to zero for all customers. The MAPE in these simulations was 24.8 % and again we observed that MAPE decreases when either the observation period or the transaction rate increases. What might be less intuitive is the result that MAPE decreases when the drop-out rate increases. Our understanding is that if the drop-out rate increases, the mean time until death decreases and a larger proportion of customers are inactive in the prediction period. Those inactive customers are easy to predict, since the number of transactions is zero by definition. In other words, the prediction exercise is less uncertain when more customers are inactive.

The results of these simulations are based on the assumption that the transaction and drop-out rates follow gamma distributions across customers. This might not be the case. The performance of the proposed estimators in actual datasets is considered in the following section.

3 Applications

We apply the proposed estimation procedure of the nonparametric model, a modification of the P/NBD model, to two datasets that have been used in previous studies in the literature. The first dataset corresponds to the purchasing of CDs from the online

Table 1 Mean absolute percent error (MAPE) simulation exercise

Observed period	MAPE	E(λ)	MAPE	E(μ)	MAPE
26	14.5	0.02	14.4	0.02	17.5
52	10.5	0.05	11.0	0.05	9.5
78	9.0	0.15	8.9	0.15	5.2
Mean	11.3	Mean	11.3	Mean	11.3

retailer CDNOW. This dataset focuses on a single cohort of new customers who made their first purchase at the CDNOW website in the first quarter of 1997. We have data covering their initial purchase as well as subsequent purchase occasions for the period from January 1997 through June 1998 (see Fader and Hardie 2001). For the purposes of this study, and to make it fully comparable with previous studies, we use a sample of 2,357 customers, 39 weeks for the estimation of the model, and 24 and 39 weeks for validation. We use these two horizons to validate the model in order to analyze possible differences.

The second dataset used in our analysis comes from a specific store of a large grocery retail in Turkey that offers a broad assortment of products. This dataset is used by Batislam et al. (2007) and by Jerath et al. (2011). Considering that people usually do their grocery shopping on a weekly basis, we use two versions of the dataset as follows: (1) transactions over a period of 52 weeks for the estimation and 39 weeks for validation, and (2) transactions over 78 weeks for the estimation period and 13 weeks for the validation period. Again, different estimation and validation periods allow comparisons and more fruitful conclusions. The number of customers in both cases is 5,473. In total, we have three datasets for the estimation of the model and four datasets for out-of-sample prediction. It is interesting to note that for the three cases considered, in the estimation sample, one CDNOW and two Retail, a large proportion of customers—between 40 and 45 %—perform no transactions after the initial transaction at time t=0. For those customers, we assume $\widehat{\lambda}_i = \widehat{\mu}_i = 0$.

Figure 2 presents the frequency distribution of the estimated parameters λ and μ across customers using the simple-NP and the P/NBD models. The bars correspond to the simple-NP model and the dashed line to the gamma distribution derived from the estimation of the P/NBD model. In the case of the λ parameter, the two distributions are very similar. In the μ coefficient, however, we observe some differences. Relative to the P/NBD model, the simple-NP model estimates a larger proportion of small μ values and hence a smaller proportion of large μ values.

Table 2 presents prediction results of the in-sample and out-of-sample prediction exercise using both the simple-NP and the P/NBD models. For each dataset and each out-of-sample period, we compare the percentage of cases where each method is best (% Best), the mean absolute difference (MAD) and the root of the mean square error (RMSE). We also compute the correlation coefficient between observed and predicted values.

In the in-sample exercise, the simple-NP model is better than the P/NBD for all cases and almost all statistics, the only exception being the correlation coefficient where both

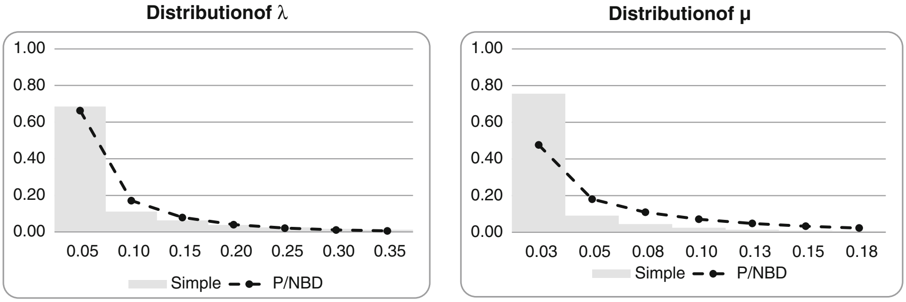

Fig. 2 Frequency distribution of parameters CDNOW

Table 2 In-sample and out-of-sample prediction error measures

Sample	Statistics	CDNOW 39/24[a]		CDNOW 39/39		Retail 78/13		Retail 52/39	
		Simple	P/NBD	Simple	P/NBD	Simple	P/NBD	Simple	P/NBD
In sample	% Best	88	12	88	12	82	18	82	18
	MAD	0.110	0.452	0.110	0.452	0.473	0.947	0.448	1.025
	RMSE	0.419	0.800	0.419	0.800	1.646	2.040	1.451	2.175
	Correlation	0.986	1.000	0.986	1.000	0.986	1.000	0.980	1.000
	Observed mean	1.042		1.042		4.534		3.289	
Out of sample	% Best	76	24	71	29	84	16	73	27
	MAD	0.513	0.543	0.747	0.755	0.371	0.398	1.325	1.318
	RMSE	1.372	1.310	1.817	1.603	0.966	0.977	3.391	3.174
	Correlation	0.575	0.587	0.613	0.630	0.888	0.823	0.754	0.771
	Observed mean	0.544		0.798		0.531		1.776	

[a] Estimation and validation period, respectively

models present values close to 1.0 and the differences between the models are minimal. The % Best of the simple-NP model is over 80 % in all cases. In the out-of-sample exercise, the % Best statistics of the simple-NP model is over 70 % in all cases. This high value is explained by the fact that the gamma distribution, and hence the P/NBD model, does not allow zero values for the purchase and/or drop-out rates, although a significant proportion of customers show no transactions in the observation period. The MAD and the RMSE statistics are similar between the two models although the P/NBD yields slightly better results. The MAD statistics of the simple-NP model is better in three out of four cases, while the RMSE is better in one out of four cases. The simple-NP model shows better results in cases where the estimation period is large and the mean purchase rate across customers is larger. This result is in line with the simulation exercise in Table 1.

An additional examination to the relative performance of the two models focuses on the predictions of individual-level transactions in the forecast period conditional on the number of observed transactions in the estimation period. In Fig. 3, we report these conditional expectations along with the average of the actual number of transactions in the forecast period for the CDNOW case, with 39 weeks in the estimation period and 39 weeks in the forecasting period, and the Retail case with 78 weeks in the estimation period and 13 in the forecasting period. In these two graphs, the observed value equal to 7 contains also the cases greater than 7, and the solid line represents actual values. Figure 3 confirms the results in Table 2 that the P/NBD performs better and our simple-NP model is a very good approximation. In the first panel and for t and more transactions, we observe a great difference between the two models. This phenomenon does not occur in the Retail graph where customers with a large number of transactions in the estimation period are well predicted by both models.

An observation concerning customers with a large number of transactions in the estimation period is the following: the simple-NP model uses only the information coming from his/her own transactions to forecast, while the P/NBD model uses the information from all customers and a gamma distribution for the transaction rate. If a

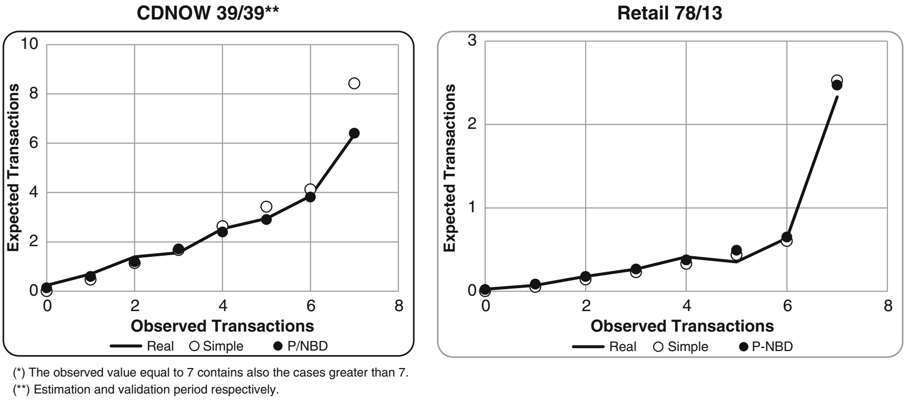

Fig. 3 Conditional expectations CDNOW and Retail*

customer is an outlier in both the estimation and the prediction period with a number of transactions far above average, the P/NBD model may not have a good performance since this customer might not fit well the gamma distribution and the simple-NP might be better. On the other hand, if the customer is an outlier only in the estimation period but not in the prediction period, the P/NBD model should do better. The big difference between the two models in the seven and more transactions in the CDNOW case is an example of this phenomenon. A few customers have many transactions in the estimation period and very few in the prediction period, although they were estimated as active. This is mainly explained by a few outlier observations which are better captured by the P/NBD although not by the simple-NP model, since the latter uses only the observation from the customer for forecasting.

An intermediate result of the models—both the simple-NP and the P/NBD—is the calculation of the probability of a customer being active at time T which is computed from Eq. (4). Alternatively, Reinartz and Kumar (2002) propose a simple, not based on any parameter, formula for this probability.

$$P(Active) = \left(\frac{t_x}{T}\right)^x$$

Our simple model, however, presents probabilities of being active closer to the ones given by the P/NBD model than the probabilities obtained from this formula by Reinartz and Kumar (2002). The correlation coefficients between the probabilities from our model and the corresponding probabilities given by the P/NBD model vary between 0.64 and 0.97 in the cases studied.

4 Conclusions

This paper proposes a simple alternative to the P/NBD model that can be used to answer standard customer-base analysis questions in non-contractual settings. Our proposal has several advantages over the traditional P/NBD model. First, our specification allows a zero transaction rate, a condition which is not possible in the P/NBD model, even though in many applications, the proportion of customers with no activity in the observation period is large. Second, if the number of transactions for each customer is large, and the number of customers is also large, there is no need to assume an a priori probability distribution. The P/NBD model assumes a gamma a priori distribution for the transaction and drop-out rates and that is an unnecessary restriction under the above conditions. Finally, but not less important, we propose a simple estimation procedure that can be applied even if the number of customers is very large.

A simulation exercise shows that under parameter values typically found in applications, the simple-NP model produces out-of-sample mean prediction errors of 11.3 % in simulated datasets generated under the P/NBD model. This error diminishes in cases where the observation period is large, the transaction rate is large, and the drop-out rate is large.

We apply the proposed estimation procedure to two datasets that have been used in previous studies in the literature. The results show that at the aggregate level, our simple procedure performs almost as well as the P/NBD model. The take away for practitioners is an easier estimation procedure to fit the model in order to analyze large customer databases of transactions in non-contractual settings that perform better in cases where the observation period is large and/or the transaction rate is large.

References

Batislam, E. P., Denizel, M., & Filiztekin, A. (2007). Empirical validation and comparison of models for customer base analysis. *Internat. J. Res. Marketing, 24*(3), 201–209.

Bemmaor, A. C., & Glady, N. (2012). Modeling purchasing behavior with sudden death: a flexible customer lifetime model. *Management Science, 58*(5), 1012–1102.

Ehrenberg, A. S. C. (1988). *Repeat Buying* (2nd ed.). New York: Oxford University Press.

Fader, P.S., & Hardie, B.G.S. (2001). Forecasting Repeat Sales at CDNOW: A Case Study. Part 2 of 2. Interfaces 31(May–June) S94–S107.

Fader, P.S., & Hardie, B.G.S. (2005). A Note on Deriving the Pareto/NBD Model and Related Expressions. http://brucehardie.com/notes/009/.

Fader, P. S., Hardie, B. G. S., & Lee, K. L. (2005b). Counting your customers the easy way: an alternative to the Pareto/NBD model. *Marketing Science, 24*(3), 275–284.

Jerath, K., Fader, P. S., & Hardie, B. G. S. (2011). New perspectives on customer death. Using a generalization of the Pareto/NBD model. *Marketing Science, 30*(5), 866–880.

Ma, S.H., & Liu, J.L. (2007). The MCMC Approach for Solving the Pareto/NBD Model and Possible Extensions, Proceedings of the Third International Conference on Natural.

Reinartz, W., & Kumar, V. (2002). The Mismanagement of Customer Loyalty. Harvard Business Review (July), pp. 86–97.

Schmittlein, D. C., Morrison, D. G., & Colombo, R. (1987). Counting your customers: who are they and what will they do next? *Management Science, 33*(1), 1–24.

Singh, S. S., Borle, S., & Jain, D. C. (2009). A generalized framework for estimating customer lifetime value when customer lifetimes are not observed. *Quant Marketing Economy, 7*(2), 181–205.