Overcoming the BG/NBD Model's #NUM! Error Problem

Peter S. Fader
www.petefader.com

Bruce G. S. Hardie[†]
www.brucehardie.com

December 2013

The BG/NBD model has proven to be a popular and robust model for customer-base analysis in continuous-time noncontractual settings. The primary reason for its popularity (relative to, say, the Pareto/NBD model) is that it is easy to implement in Excel.

Hundreds of analysts and researchers have implemented the model without any problems. However, a few have come across the problem that they get a #NUM! error when evaluating the model's likelihood function. This typically occurs in settings where some customers have made a very large number of repeat transactions. This note documents two ways of overcoming this problem.

The Problem

- The BG/NBD likelihood function for a randomly chosen customer with purchase history (x, t_x, T) is

$$\begin{aligned} L(r, \alpha, a, b \,|\, x, t_x, T) \\ &= \frac{B(a, b+x)}{B(a, b)} \frac{\Gamma(r+x)\alpha^r}{\Gamma(r)(\alpha+T)^{r+x}} \\ &\quad + \delta_{x>0} \frac{B(a+1, b+x-1)}{B(a, b)} \frac{\Gamma(r+x)\alpha^r}{\Gamma(r)(\alpha+t_x)^{r+x}} \,. \end{aligned} \tag{1}$$

- To implement the model in Excel, we rewrite (1) as

$$L(r, \alpha, a, b \,|\, x, t_x, T) = \mathsf{A}_1 \cdot \mathsf{A}_2 \cdot (\mathsf{A}_3 + \delta_{x>0}\, \mathsf{A}_4)\,, \tag{2}$$

where

$$\mathsf{A}_1 = \frac{\Gamma(r+x)\alpha^r}{\Gamma(r)} \qquad \mathsf{A}_2 = \frac{\Gamma(a+b)\Gamma(b+x)}{\Gamma(b)\Gamma(a+b+x)}$$
$$\mathsf{A}_3 = \Big(\frac{1}{\alpha+T}\Big)^{r+x} \qquad \mathsf{A}_4 = \Big(\frac{a}{b+x-1}\Big)\Big(\frac{1}{\alpha+t_x}\Big)^{r+x}$$

- Taking logs gives us the following expression for the log-likelihood function:

$$LL(r,\alpha,a,b\,|\,x,t_x,T) = \mathsf{B}_1 + \mathsf{B}_2 + \ln[\exp(\mathsf{B}_3) + \delta_{x>0}\,\exp(\mathsf{B}_4)]\,, \quad (3)$$

where

$$\mathsf{B}_1 = r\ln(\alpha) + \ln[\Gamma(r+x)] - \ln[\Gamma(r)]$$
$$\mathsf{B}_2 = \ln[\Gamma(a+b)] + \ln[\Gamma(b+x)] - \ln[\Gamma(b)] - \ln[\Gamma(a+b+x)]$$
$$\mathsf{B}_3 = -(r+x)\ln(\alpha+T)$$
$$\mathsf{B}_4 = \ln(a) - \ln(b+x-1) - (r+x)\ln(\alpha+t_x)$$

This is the expression coded-up in the Excel spreadsheet reported in the original BG/NBD paper.

- To illustrate the problem, let us consider a customer with $x = 200$, $t_x = 38$ and $T = 40$, where the unit of time is one week. For $r = 0.25, \alpha = 4.00, a = 0.80, b = 2.50$, we get

$$\begin{aligned} \mathsf{B}_1 &= 858.3163 \\ \mathsf{B}_2 &= -3.5458 \\ \mathsf{B}_3 &= -757.7840 \\ \mathsf{B}_4 &= -753.9973 \end{aligned}$$

with LL = #NUM!. The problem is that $\exp(\mathsf{B}_3) = 7.92\text{E}{-330}$ and $\exp(\mathsf{B}_4) = 3.49\text{E}{-328}$, both of which are smaller than the smallest positive number that can be stored in Excel, 2.23E−308. As a result, Excel returns an answer of zero, and taking the log of zero in (3) gives us the `#NUM!` result.

Solution #1

- The first solution to the problem to change the unit of time. When we change the unit of time from, say, week to day, our estimates of r, a and b will remain exactly the same, and the value of α will be seven times its original value. Similarly, if we change the unit of time from week to quad-week, the value of α will be a quarter of its original value.

- If we change the unit of time in the above example from week to quad-week, B_3 and B_4 will be less negative and we will typically overcome the log of zero problem. (Note that some users of the model in high-frequency environments feel the need to use day as the unit of time. This is not necessary and only increases the likelihood of facing the `#NUM!` problem.)

- To illustrate this, changing the unit of time from week to quad-week sees the customer's behavior coded as $x = 200$, $t_x = 9.5$ and $T = 10.0$. With α rescaled to 1.00, we get

$$\begin{aligned} \mathsf{B}_1 &= 857.9698 \\ \mathsf{B}_2 &= -3.5458 \\ \mathsf{B}_3 &= -480.1785 \\ \mathsf{B}_4 &= -476.3918 \end{aligned}$$

with $LL = 378.0546$. Problem solved! (Don't worry about the positive value of the log-likelihood function; we are not taking the log of a probability so it does not have to be negative.)

Solution #2

- Reflecting on (2) and (3), the basic problem is that we need to evaluate

$$\ln\left[\left(\frac{1}{\alpha+T}\right)^{r+x} + \delta_{x>0}\left(\frac{a}{b+x-1}\right)\left(\frac{1}{\alpha+t_x}\right)^{r+x}\right].$$

For large values of x, both $1/(\alpha+T)^{r+x}$ and $1/(\alpha+t_x)^{r+x}$ are computed as zero. If we factor out $1/(\alpha+t_x)^{r+x}$ before taking the logs, we get

$$-(r+x)\ln(\alpha+t_x) + \ln\left[\left(\frac{\alpha+t_x}{\alpha+T}\right)^{r+x} + \delta_{x>0}\left(\frac{a}{b+x-1}\right)\right].$$

Since it is extremely unlikely that $[(\alpha+t_x)/(\alpha+T)]^{r+x}$ will be computed as zero for large values of x, the log of zero problem is avoided.

- Therefore, our second solution to the problem is to rewrite (1) as

$$L(r,\alpha,a,b\,|\,x,t_x,T) = \mathsf{C}_1 \cdot \mathsf{C}_2 \cdot (\mathsf{C}_3 + \delta_{x>0}\,\mathsf{C}_4)\,,$$

where

$$\mathsf{C}_1 = \frac{\Gamma(r+x)}{\Gamma(r)}\frac{\Gamma(a+b)\Gamma(b+x)}{\Gamma(b)\Gamma(a+b+x)} \qquad \mathsf{C}_2 = \frac{\alpha^r}{(\alpha+t_x)^{r+x}}$$

$$\mathsf{C}_3 = \left(\frac{\alpha+t_x}{\alpha+T}\right)^{r+x} \qquad \mathsf{C}_4 = \left(\frac{a}{b+x-1}\right)$$

- Taking logs gives us the following expression for the log-likelihood function:

$$LL(r, \alpha, a, b \,|\, x, t_x, T) = \mathsf{D}_1 + \mathsf{D}_2 + \ln(\mathsf{C}_3 + \delta_{x>0}\,\mathsf{C}_4)\,, \tag{4}$$

  where

$$\begin{aligned}\mathsf{D}_1 &= \ln[\Gamma(r+x)] - \ln[\Gamma(r)] + \ln[\Gamma(a+b)] + \ln[\Gamma(b+x)] \\ &\qquad - \ln[\Gamma(b)] - \ln[\Gamma(a+b+x)] \\ \mathsf{D}_2 &= r\ln(\alpha) - (r+x)\ln(\alpha+t_x)\end{aligned}$$

- Let us reconsider the customer with $x = 200$, $t_x = 38$ and $T = 40$. For $r = 0.25, \alpha = 4.00, a = 0.80, b = 2.50$, we get

$$\begin{aligned}\mathsf{D}_1 &= 854.4240 \\ \mathsf{D}_2 &= -748.1218 \\ \mathsf{C}_3 &= 9.00\text{E}{-05} \\ \mathsf{C}_4 &= 3.97\text{E}{-03}\end{aligned}$$

  with $LL = 100.7957$. Problem solved!

In Closing ...

- To summarize, the problem of getting the `#NUM!` error when evaluating the BG/NBD model likelihood function can be solved by rescaling time (i.e., using a longer unit of time) or using (4) instead of (3).
- The spreadsheet `alt_bgnbd_ll.xlsx` illustrates these two solutions for the standard CDNOW dataset.
- Note that the Pareto/NBD model suffers from exactly the same problem, which can also be overcome by either rescaling time or factoring out $1/(\alpha + t_x)^{r+x}$ in the likelihood function.