Creating a Fit Histogram for the BG/NBD Model

Peter S. Fader
www.petefader.com

Bruce G. S. Hardie
www.brucehardie.com

Ka Lok Lee†
www.kaloklee.com

January 2007

1. Introduction

One way of assessing the fit of a model of repeat buying is to compare the actual frequency distribution of transaction counts (how many people made 0, 1, 2, ... repeat transactions) with that predicted by the model. This note describes how to create such a plot for the BG/NBD model (Fader et al. 2005a) in Excel. It is assumed that the reader is familiar with the material presented in Fader et al. (2005b) and has access to the associated Excel workbook (`bgnbd.xls`).

Section 2 describes the mechanics of "coding up" the BG/NBD expression for $P(X(t) = x)$. This is extended in Section 3 to the case where the time period over which repeat transactions could have occurred varies across customers, as is the case in the CDNOW example. These sections should be read in conjunction with the Excel workbook `bgnbd_fit_histogram.xls`. (We strongly encourage interested readers to build the spreadsheet "from scratch" for themselves, using this note and the associated Excel workbook as a guide.)

2. Basic Coding

As derived in Fader et al. (2005a), the BG/NBD probability of observing x purchases in a time period of length t is

$$P(X(t)=x|r,\alpha,a,b) = \frac{B(a,b+x)}{B(a,b)}\frac{\Gamma(r+x)}{\Gamma(r)x!}\left(\frac{\alpha}{\alpha+t}\right)^r\left(\frac{t}{\alpha+t}\right)^x + \delta_{x>0}\,\frac{B(a+1,b+x-1)}{B(a,b)} \times\left[1-\left(\frac{\alpha}{\alpha+t}\right)^r\underbrace{\left\{\sum_{j=0}^{x-1}\frac{(\Gamma(r+j)}{\Gamma(r)j!}\left(\frac{t}{\alpha+t}\right)^j\right\}}_{\mathsf{A}}\right]. \quad (1)$$

The first two lines of this expression are easy (albeit tedious) to "code up" in Excel. At first glance the third line may appear to be more of a coding challenge, given the summation denoted by A. However, as we shall see, there is a simple way of managing this summation in a basic worksheet.

To illustrate this, let us compute the probability that a CDNOW customer who made his first purchase on 1997-01-01 makes $0, 1, \ldots, 6, 7+$ repeat transactions in the following $39 - \frac{1}{7} = 38.86$ weeks.

Starting with a blank worksheet (which we name `Basic Coding`), we copy the maximum likelihood estimates of the four model parameters ($r = 0.243$, $\alpha = 4.414$, $a = 0.793$, and $b = 2.426$) from the `BGNBD Estimation` worksheet in the Excel workbook `bgnbd.xls` to cells `B1:B4`. Next we specify t (cell `B6`), the length of the period of time over which the customer could have made repeat purchases; for this example, $t = 38.86$. We then specify the values of x for which we wish to compute $P(X(t) = x)$, entering $0, 1, \ldots, 6, 7+$ in cells `A9:A16`. As a final set-up step, we note that

$$P(X(t) \geq 7) = 1 - \sum_{x=0}^{6} P(X(t) = x)\,,$$

and therefore enter `=1-SUM(B9:B15)` in cell `B16`. (At this stage, 1.0000 should appear in this cell.)

Since it appears twice in (1), we compute the quantity $B(a, b)$ separately in cell `E1`. Noting that

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}\,, \quad (2)$$

we compute this quantity using the following expression:

```
=EXP(GAMMALN(B3)+GAMMALN(B4)-GAMMALN(B3+B4))
```

For the case of $x = 0$, we only need to evaluate

$$\frac{B(a,b+x)}{B(a,b)}\frac{\Gamma(r+x)}{\Gamma(r)x!}\left(\frac{\alpha}{\alpha+t}\right)^r\left(\frac{t}{\alpha+t}\right)^x.$$

Recalling (2) and noting that $x! = \Gamma(x+1)$, we compute this quantity by entering the following formula in cell `B9`:

```
=EXP(GAMMALN($B$3)+GAMMALN($B$4+A9)-GAMMALN($B$3+$B$4+A9))/
$E$1*EXP(GAMMALN($B$1+A9)-GAMMALN($B$1)-GAMMALN(A9+1))*
($B$2/($B$2+$B$6))^$B$1*($B$6/($B$2+$B$6))^A9
```

For $x > 0$, we need to evaluate the summation denoted by A in (1). We will compute this separately for $x - 1 = 0, 1, \ldots, 5$ in cells `B19:B24` (why do we stop at $x - 1 = 5$?), referring to the relevant cell when evaluating (1).

- We enter $0, 1, \ldots, 5$ in cells `A19:A24`.
- For $x - 1 = 0$, the summand (and therefore the sum) equals 1.0, which we enter in cell `B19`.
- Next we enter

  ```
  =B19+EXP(GAMMALN($B$1+A20)-GAMMALN($B$1)
  -GAMMALN(A20+1))*($B$6/($B$2+$B$6))^A20
  ```

  in cell `B20`, and copy this formula down to cell `B24`.

We are now in a position to compute $P(X(t) = x)$ for $x > 0$:

- We first copy `B9` to `B10` and then *add* the following to the end of the formula currently in `B10`

  ```
  +EXP(GAMMALN($B$3+1)+GAMMALN($B$4+A10-1)
  -GAMMALN($B$3+$B$4+A10))/$E$1
  *(1-($B$2/($B$2+$B$6))^$B$1*B19)
  ```

- We copy the contents of `B10` down to cell `B15`.

And that's it!

3. Predicted Distribution of Transactions for the Cohort

Now that we know how to evaluate (1) in Excel, we consider how to compute the predicted distribution of transactions for the cohort of 2357 customers who made their first purchase at CDNOW in the first quarter of 1997 (and for whom the time period over which repeat transactions could have occurred varies across customers). In order to do this, we need to copy the `Raw Data` worksheet from `bgnbd.xls`.

Let f_x denote the number of people making x repeat transactions in the 39-week model calibration period (x = 0, 1, 2, ...). The actual frequency distribution of repeat transaction counts can easily be determined using the "pivot table" feature in Excel. Starting in the `Raw Data` worksheet, we select the *PivotTable and PivotChart ...* under the *Data* menu. We use x as the *row field* and use ID as the *data item*. The resulting table is reported in the

`Actual Frequency Distribution` worksheet. Using this full distribution, we create a right-censored distribution in which counts greater than 7 are collapsed into a 7+ bin.

As noted above, the task of computing the expected frequency distribution for this example is complicated by the fact that the time period over which repeat transactions could have occurred varies across customers. Let n_s is the number of customers who made their first purchase at CDNOW on day s of 1997 (and therefore have $t - \frac{s}{7}$ weeks within which to make repeat purchases). It follows that the expected number of people in this cohort of new customers with x repeat transactions is computed in the following manner:

$$E(f_x) = \sum_{s=1}^{84} n_s P(X(t - \tfrac{s}{7}) = x)\,,\ x = 0, 1, 2, \ldots \tag{3}$$

The first step is to determine n_s, which we do using Excel's pivot table tool. We start by making a copy of the `Raw Data` worksheet — let's call it `n_s`. Given T, the number of days (in weeks) during which *repeat* transactions could have occurred in the 39-week calibration period, it follows that the "time of first purchase" (column `D`) is simply $39 - T$. Selecting the *PivotTable and PivotChart ...* under the *Data* menu, we use "time of first purchase" as the *row field* and use ID as the *data item*. We see that 18 people made their first-ever purchase at CDNOW on the first day of the first week of 1997, 22 on the second day of the first week of 1997, ..., and 30 on the seventh day of the twelfth week of 1997.

We copy (transpose) cells `G7:H90` into a new worksheet (which we call `Histogram`), starting at cell `F6`. In cell `F8` we compute the time horizon over which the 18 people who made their first-ever purchase at CDNOW on 1997-01-01 could have made repeat purchases using the formula `=39-F6`, and then copy this across to cell `CK8`.

In rows `10`–`24` of columns `F`–`CK`, we enter the formulae required to compute $P(X(T) = x)$ ($x = 0, 1, \ldots, 7+$) for each of the 84 possible values of T (using the structure developed in Section 2 above).

Finally, the $E(f_x)$ are computed in cells `C10:C17` using the Excel function `=SUMPRODUCT()` to perform the calculation give in (3). (The chi-square goodness-of-fit test statistic is computed in cell `D18`.)

References

Fader, Peter S., Bruce G. S. Hardie, and Ka Lok Lee (2005a), ""Counting Your Customers" the Easy Way: An Alternative to the Pareto/NBD Model," *Marketing Science*, **24** (Spring), 275–284.

Fader, Peter S., Bruce G. S. Hardie, and Ka Lok Lee (2005b), "Implementing the BG/NBD Model for Customer Base Analysis in Excel." <http://brucehardie.com/notes/004/>