Implementing the BG/BB Model for Customer-Base Analysis in Excel

Peter S. Fader
www.petefader.com

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

January 2011

1. Introduction

This note describes how to implement the BG/BB model for customer-base analysis[1] using Microsoft Excel.

We first consider how to estimate the model parameters by "coding-up" the log-likelihood function. Next we show how to create three sets of plots used to evaluate the performance of the model for any given dataset: the in-sample model fit plot, the tracking plots, and the conditional expectations plots. Finally, we consider how to compute the posterior mean of P and DERT.

The specific steps are outlined in Sections 2–7 below. All these sections should be read in conjunction with the workbook `BGBB_2011-01-20.xlsx`. We strongly encourage interested readers to build the set of worksheets associated with this model "from scratch" for themselves, using this note and the Excel workbook as a guide.

This note does not show how to compute all the quantities associated with the BG/BB model. However, any reader that has worked through this note should find it easy to code up the associated equations in Excel.

[1] Fader, Peter S., Bruce G. S. Hardie, and Jen Shang (2010), "Customer-Base Analysis in a Discrete-Time Noncontractual Setting," *Marketing Science*, **29** (November–December), 1086–1108.

2. Parameter Estimation

The likelihood function for a randomly chosen customer with purchase history (x, t_x, n) is

$$L(\alpha,\beta,\gamma,\delta \,|\, x, t_x, n) = \frac{B(\alpha+x, \beta+n-x)}{B(\alpha,\beta)} \frac{B(\gamma,\delta+n)}{B(\gamma,\delta)} + \sum_{i=0}^{n-t_x-1} \frac{B(\alpha+x, \beta+t_x-x+i)}{B(\alpha,\beta)} \frac{B(\gamma+1,\delta+t_x+i)}{B(\gamma,\delta)}. \quad (1)$$

For a sample of K customers, where customer k's purchase history is denoted by (x_k, t_{x_k}, n_k), the sample log-likelihood function is given by

$$LL(\alpha,\beta,\gamma,\delta) = \sum_{k=1}^{K} \ln\left[L(\alpha,\beta,\gamma,\delta \,|\, x_k, t_{x_k}, n_k)\right]. \quad (2)$$

When $n_k = n$ for all k, as is this case for our empirical example, there is no need to loop over all the customers as in equation (2) above; we only need to loop over the $J = n(n+1)/2 + 1$ possible recency/frequency patterns, each containing f_j customers:

$$LL(\alpha,\beta,\gamma,\delta) = \sum_{j=1}^{J} f_j \ln\left[L(\alpha,\beta,\gamma,\delta \,|\, x_j, t_{x_j}, n)\right], \quad (3)$$

where x_j and t_{x_j} are the frequency and recency associated with each unique pattern. We now consider how to code up equations (1) and (3) in Excel.

- The recency/frequency summary of the annual donation behavior by the 1995 cohort of first-time supporters (Table 2 in the paper) is given in the worksheet `Table 2 data`. We start by making a copy of this worksheet — let's call it `Parameter Estimation` — and inserting seven rows at the top of the new worksheet.

- The BG/BB model has four parameters: α, β, γ,δ. In order to code up equations (1) and (3) in the spreadsheet without an error message appearing (e.g., `#NUM!` or `#DIV/0!`), we need some starting values for the four parameters. The exact values do not matter — provided they are within the defined bounds — so we start with 1.0 for all four parameters. We locate these parameter values in cells `B1:B4`.

- Looking at equation (1), we see that we will repeatedly use the quantities $B(\alpha,\beta)$ and $B(\gamma,\delta)$. We therefore compute them separately in cells `E1` and `E3` using

```
=EXP(GAMMALN(B1)+GAMMALN(B2)-GAMMALN(B1+B2))
```

and

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

respectively. (With starting values of $\alpha = \beta = \gamma = \delta = 1$, both of these quantities equal 1.)

- The first part of equation (1) does not depend on t_x:

$$\frac{B(\alpha + x, \beta + n - x)}{B(\alpha, \beta)} \frac{B(\gamma, \delta + n)}{B(\gamma, \delta)} .$$

For the first recency/frequency pattern, this formula is entered in cell `H9` as

```
=EXP(GAMMALN($B$1+A9)+GAMMALN($B$2+C9-A9)
-GAMMALN($B$1+$B$2+C9))/$E$1*EXP(GAMMALN($B$3)
+GAMMALN($B$4+C9)-GAMMALN($B$3+$B$4+C9))/$E$3
```

We copy this expression down to cell `H30`.

- The next step is to deal with the summation part of equation (1). This is slightly tricky as performing a looping operation in Excel (in this case, looping over i) is, at first glance, not easy without resorting to VBA code. The maximum upper limit of the summation is $n-1$ when $t_x = 0$ (for $x = 0$), which for this example (with $n = 6$) is 5. In cells `I8:N8` we enter the possible values that i could take on: $0, 1, 2, 3, 4, 5$.

  In cells `I9:N30`, we are going to enter an expression for the summand,

$$\frac{B(\alpha + x, \beta + t_x - x + i)}{B(\alpha, \beta)} \frac{B(\gamma + 1, \delta + t_x + i)}{B(\gamma, \delta)} . \tag{4}$$

  However, we do not evaluate this for all values of i; the upper limit depends on the recency value associated with each recency/frequency pattern. To determine the upper limit of the summation $(n - t_x - 1)$, we enter `=C9-B9-1` in cell `G9` and copy this down to cell `G30`.

  We then enter the following expression

```
=IF(I$8<=$G9,EXP(GAMMALN($B$1+$A9)
    +GAMMALN($B$2+$B9-$A9+I$8)
 -GAMMALN($B$1+$B$2+$B9+I$8))/$E$1
*EXP(GAMMALN($B$3+1)+GAMMALN($B$4+$B9+I$8)
  -GAMMALN($B$3+$B$4+$B9+I$8+1))/$E$3,0)
```

in cell `I9`. This is evaluating equation (4) while i is less than or equal to $(n - t_x - 1)$; if $i > n - t_x - 1$, it returns a value of 0.

We copy this across to cell `N9`, and then copy this block of cells down to row 30.

- Having computed all the elements of equation (1), we sum them up by entering `=SUM(H9:N9)` in cell `F9`. This gives us the value of the likelihood function $L(\alpha, \beta, \gamma, \delta \,|\, x, t_x, n)$ for the recency/frequency combination in row 9, as evaluated for the values of $\alpha, \beta, \gamma, \delta$ given in cells `B1:B4`. We copy this down to cell `F30`.

- Finally, we multiply the number of people associated with each of the 22 recency/frequency patterns by the log of the corresponding likelihood function value. We enter `=D9*LN(F9)` in cell `E9` and copy this down to cell `C30`. The sum of these 22 cells is entered in cell `B6`: `=SUM(E9:E30)`. This is the value of the sample log-likelihood function (equation (3)) given the values of the four model parameters in cells `B1:B4`. (With starting values of 1.0 for all four parameters, $LL = -37{,}232.0$.)

We find the maximum likelihood estimates of the four model parameters by maximizing this log-likelihood function using Solver. With reference to Figure 1, the *target cell* is the value of the log-likelihood, cell `B6`. We wish to *maximize* this by *changing* cells `B1:B4`. The *constraints* we place on the parameters are that α, β, γ, and δ are greater than 0. As Solver only offers us a "greater than or equal to" constraint, we *add* the constraint that cells `B1:B4` are $\geq$ a small positive number (e.g., 0.0001). Clicking the *Solve* button, Solver finds the values of the four model parameters that maximize the log-likelihood function.

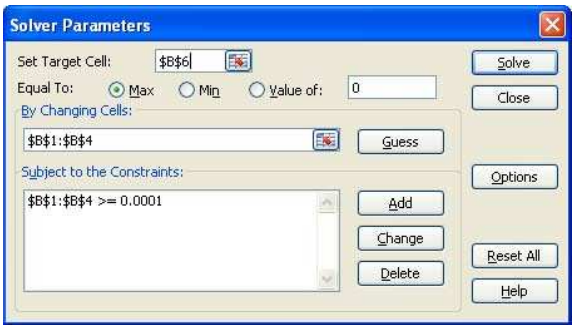

Figure 1: Solver Settings

Can we be sure that we have actually reached the maximum of the log-likelihood function? Using the solution given by Solver as the set of starting

values for the parameters, we "fire up" Solver again to see if it can improve on this solution. Once we are satisfied that the maximum has indeed been reached, we can say that the numbers given in cells `B1:B4` are the maximum likelihood estimates of the model parameters. As reported in Table 4 in the paper, the maximum value of the log-likelihood function is −33,225.6, associated with $\alpha = 1.204$, $\beta = 0.750$, $\gamma = 0.657$, $\delta = 2.783$.

So as to be more confident that we have reached the global maximum of the likelihood function, it is good practice to redo the optimization process using a completely different set of starting values. For example, using starting values of 0.01, 0.01, 0.01, 0.01 for cells `B1:B4`, repeatedly use Solver until you are satisfied that the maximum of the log-likelihood function has been reached. Are the corresponding values of the four model parameters equal to those given above? They should be.

In some empirical settings, we have found that Solver returns `#DIV/0!` as the value of the log-likelihood function, with γ and δ taking on "large" values (in the order of hundreds or thousands). See the Appendix for a discussion of this problem and possible solutions.

3. Creating an In-Sample Model Fit Plot

One way of assessing in-sample model fit is to compare the expected number of people making 0, 1, ..., 6 repeat transactions in the calibration sample period to the actual frequency distribution — as done in Figure 3 in the paper. The expected frequencies are computed using the following expression for the BG/BB pmf:

$$P(X(n) = x \mid \alpha, \beta, \gamma, \delta) = \binom{n}{x} \frac{B(\alpha + x, \beta + n - x)}{B(\alpha, \beta)} \frac{B(\gamma, \delta + n)}{B(\gamma, \delta)} + \sum_{i=x}^{n-1} \binom{i}{x} \frac{B(\alpha + x, \beta + i - x)}{B(\alpha, \beta)} \frac{B(\gamma + 1, \delta + i)}{B(\gamma, \delta)}. \quad (5)$$

Using logic very similar to that used for coding up the model likelihood function, equation (5) is coded up in cells `A8:I15` of the worksheet `In-Sample Fit`. (We leave it to the reader to reverse-engineer the exact logic.)

The expected number of people with frequency x in a sample of K customers is simply $K \times P(X(n) = x)$. This is computed in cells `C18:C24`. The actual frequency distribution (cells `B18:B24`) is determined by performing a pivot-table analysis on the data given in the worksheet `Table 2 data`, and is compared to the expected frequency distribution in the associated plot. (This is a replication of Figure 3 in the paper.)

4. Creating Tracking Plots

Another way to assess the performance of the model is to see how well it tracks repeat transactions over time. For a randomly chosen customer, the expected (cumulative) number of repeat transactions across n transaction opportunities is given by

$$E(X(n)\,|\,\alpha,\beta,\gamma,\delta) = \left(\frac{\alpha}{\alpha+\beta}\right)\left(\frac{\delta}{\gamma-1}\right)\left\{1-\frac{\Gamma(\gamma+\delta)}{\Gamma(\gamma+\delta+n)}\frac{\Gamma(1+\delta+n)}{\Gamma(1+\delta)}\right\}. \quad (6)$$

It follows that the expected number of repeat transactions from a sample of K customers across transaction opportunities $1,\ldots,n$ is $K \times E[X(n)]$.

With reference to the worksheet `Tracking Plots`, we start by evaluating equation (6) for $n = 1\,(1996), 2\,(1997), \ldots, 11\,(2006)$ in cells `A7:F17`. (The three terms of equation (6) are computed in cells `D7:F17`, with the product of these three terms (i.e., $E[X(n)]$) computed in cells `B7:B17`.) Given these numbers, we compute the expected number of (cumulative) repeat transactions made by the cohort of 11,104 people from 1996 up to 2006 in cells `J7:J17`. The corresponding actual numbers, as computed from the raw dataset (not presented in the workbook), are given in cells `I7:I17`. These data are plotted as both a line chart and a bar chart; the line chart is a replication of Figure 4a in the paper. Taking differences gives us the annual repeat sales numbers (cells `L7:M17`), which are plotted as both a line chart and a bar chart; the line chart is a replication of Figure 4b in the paper.

5. Computing Conditional Expectations

As noted in the paper, a key examination of the predictive performance of the model focuses on the quality of the predictions of future behavior conditional on past behavior.

The expected number of future transactions across the next n^* transaction opportunities by a customer with purchase history (x, t_x, n) — the so-called *conditional expectation* — is

$$E(X(n, n+n^*)\,|\,\alpha,\beta,\gamma,\delta,x,t_x,n) = \frac{1}{L(\alpha,\beta,\gamma,\delta\,|\,x,t_x,n)}\frac{B(\alpha+x+1,\beta+n-x)}{B(\alpha,\beta)} \times \left(\frac{\delta}{\gamma-1}\right)\frac{\Gamma(\gamma+\delta)}{\Gamma(1+\delta)}\left\{\frac{\Gamma(1+\delta+n)}{\Gamma(\gamma+\delta+n)}-\frac{\Gamma(1+\delta+n+n^*)}{\Gamma(\gamma+\delta+n+n^*)}\right\}. \quad (7)$$

This is very easy to evaluate as we have already created expressions for $B(\alpha,\beta)$ and $L(\alpha,\beta,\gamma,\delta\,|\,x,t_x,n)$ as part of the parameter estimation process, and the last line is the same for all recency/frequency combinations (i.e., it is independent of x and t_x).

- We first make a copy of the worksheet `Parameter Estimation` (let's call it `Conditional Expectations (I)`), insert two columns to the right of column F, and two rows after row 4.

- We need to specify n^*, the horizon over which we are computing the conditional expectations. For this example, we will compute the expected number of transactions in 2002–2006, so we enter a value of $n^* = 5$ in cell `B6`.

- We compute the third line of equation (7) in cell `E6`, and compute the quantity $B(\alpha+x+1, \beta+n-x)/B(\alpha, \beta)$ in cells `H11:H32`. Combining terms in cells `G11:G32` gives us the expected number of transactions in 2002–2002 for each of the 22 recency/frequency combinations. These are the numbers reported in Table 5 in the paper.

Figure 6 in the paper plots the predicted versus actual conditional expectations of repeat transactions in 2002–2006 as a function of (a) frequency and (b) recency. These plots are created in the following manner.

- We insert a new worksheet which we call `Conditional Expectations (II)`. For each of the 22 recency/frequency combinations, we copy over the associated number of customers and the computed conditional expectation number. In cells `F2:F23` we compute the expected total amount of purchasing in 2002–2006 by the customers associated with each recency/frequency combination. The corresponding actual numbers, as computed from the raw dataset (not presented in the workbook), are given in cells `G2:G23`.

- Next we create a pivot table of the "actual total" numbers by frequency and recency, summarizing the data using the "Sum" calculation type (see `Pivot Table I`), a pivot table "summing" these "expected total" numbers by frequency and recency (see `Pivot Table II`) and a pivot table "summing" the "# donors" numbers by frequency and recency (see `Pivot Table III`). These three sets of numbers are copied back to `Conditional Expectations (II)`.

- The plot of predicted versus actual conditional expectations of repeat transactions in 2002–2006 as a function of frequency averages the conditional expectation numbers over customers with different values of t_x for each x. This is a weighted average, where the weights are the number of customers associated with each value of t_x. This is equivalent to dividing the row totals of the predicted and actual purchases by the number of people in each row — see cells `U3:V9`. The associated line chart is a replication of Figure 6a in the paper.

- Similarly, the plot of predicted versus actual conditional expectations of repeat transactions in 2002–2006 as a function of recency averages

the conditional expectation numbers over customers with different values of x for each t_x. This is a weighted average, where the weights are the number of customers associated with each value of x. This is equivalent to dividing the column totals of the predicted and actual purchases by the number of people in each column — see cells `U13:V19`. The associated line chart is a replication of Figure 6b in the paper.

6. Computing the Posterior Mean of P

For $l, m = 0, 1, 2, \ldots$, the (l, m)th product moment of the joint posterior distribution of P and Θ is

$$E(P^l \Theta^m \,|\, \alpha, \beta, \gamma, \delta, x, t_x, n) = \frac{B(\alpha + l, \beta)}{B(\alpha, \beta)} \frac{B(\gamma + m, \delta)}{B(\gamma, \delta)} \frac{L(\alpha + l, \beta, \gamma + m, \delta \,|\, x, t_x, n)}{L(\alpha, \beta, \gamma, \delta \,|\, x, t_x, n)}, \quad (8)$$

where $L(\alpha + l, \beta, \gamma + m, \delta \,|\, x, t_x, n)$ is simply equation (1) evaluated using $\alpha + l$ in place of α and $\gamma + m$ in place of γ.

- We start by making a copy of the worksheet `Parameter Estimation` (let's call it `E(P^l,Theta^m)`). We delete the contents of cells `A6:B6` and `E9:E30`, and insert one row after row 4.

- We are interested in computing the mean of the marginal posterior distribution of P; we therefore enter the value of $l = 1$ in cell `B6` and $m = 0$ in cell `B7`. (If we want to compute other product moments of the joint posterior distribution of P and Θ, we simply change the numbers in cells `B6:B7`.)

- We need to compute $L(\alpha + l, \beta, \gamma + m, \delta \,|\, x, t_x, n)$, which is simply the model likelihood function evaluated using $\alpha + l$ in place of α and $\gamma + m$ in place of γ. We also need to compute $B(\alpha, \beta)$ and $B(\gamma, \delta)$. We copy the original parameter estimates to cells `H1:H4` and enter the expressions for $B(\alpha, \beta)$ and $B(\gamma, \delta)$ in cells `K1` and `K3`, respectively. We enter `=H1+B6` in cell `B1` (giving us $\alpha + l$) and `=H3+B7` in cell `B3` (giving us $\gamma + m$). Cells `F10:F31` now contain $L(\alpha + l, \beta, \gamma + m, \delta \,|\, x, t_x, n)$.

- Finally we enter

  ```
  =$E$1/$K$1*$E$3/$K$3*F10/'Parameter Estimation'!F9
  ```

  in cell `E10`, which evaluates equation (8), and copy it down to cell `E31`.

This gives us the mean of the marginal posterior distribution of P for each of the 22 recency/frequency combinations. These are the numbers reported in Table 7 in the paper.

7. Computing DERT

The number of discounted expected residual transactions ($DERT$) is the present value of the expected future transaction stream for a customer with purchase history (x, t_x, n). The formula for this quantity under the BG/BB model for a specified discount rate d is

$$\begin{aligned} &DERT(d \,|\, \alpha, \beta, \gamma, \delta, x, t_x, n) \\ &\quad = \frac{B(\alpha + x + 1, \beta + n - x)}{B(\alpha, \beta)} \frac{B(\gamma, \delta + n + 1)}{B(\gamma, \delta)(1 + d)} \\ &\qquad \times \frac{{}_2F_1\big(1, \delta + n + 1; \gamma + \delta + n + 1; \frac{1}{1+d}\big)}{L(\alpha, \beta, \gamma, \delta \,|\, x, t_x, n)}\,, \end{aligned} \tag{9}$$

where ${}_2F_1(\cdot)$ is the Gaussian hypergeometric function. (While the presence of the Gaussian hypergeometric function complicates the evaluation of this formula, it is worth emphasizing that the function only needs to be evaluated once for any given value of n (i.e., only once per cohort, not for every recency/frequency pattern).)

The Gaussian hypergeometric function is the power series of the form

$${}_2F_1(a, b; c; z) = \sum_{j=0}^{\infty} \frac{(a)_j (b)_j}{(c)_j} \frac{z^j}{j!}\,, \quad c \neq 0, -1, -2, \ldots,$$

where $(a)_j$ is Pochhammer's symbol, which denotes the ascending factorial $a(a+1)\cdots(a+j-1)$. (Note that an ascending factorial can be represented as the ratio of two gamma functions, $(a)_j = \Gamma(a+j)/\Gamma(a)$.) The series converges for $|z| < 1$ and is divergent for $|z| > 1$; if $|z| = 1$, the series converges for $c - a - b > 0$.

Writing

$${}_2F_1(a, b; c; z) = \sum_{j=0}^{\infty} u_j\,, \text{ where } \; u_j = \frac{(a)_j (b)_j}{(c)_j} \frac{z^j}{j!}\,,$$

we have the following recursive expression for each term of the series:

$$\frac{u_j}{u_{j-1}} = \frac{(a + j - 1)(b + j - 1)}{(c + j - 1)j} z\,, \; j = 1, 2, 3, \ldots$$

where $u_0 = 1$. This lends itself to a simple (and relatively robust) numerical method for the evaluation of the Gaussian hypergeometric function: continue adding terms to the series until u_j is less than "machine epsilon" (the smallest number that a specific computer recognizes as being bigger than zero). However, when "hard-coding" this in a worksheet (as opposed to, say, creating a custom function using VBA), it is easier to compute the series to a fixed number of terms; in this case, we will evaluate the first 151 terms (i.e., $j = 0, 1, \ldots, 150$).

- We start by making a copy of the worksheet `Parameter Estimation` (let's call it `DERT`), and insert one column to the right of column F, and two rows after row 4.

- We then need to specify the discount rate d. For this example, we will assume an annual rate of 10%, so we enter a value of 0.1 in cell `B6`.

- Next we evaluate the Gaussian hypergeometric function. The 'a' parameter is simply 1, which we enter in cell `D36`. The 'b' parameter is $\delta + n + 1$, entered as `=B4+C11+1` in cell `D37`. The 'c' parameter is $\gamma + \delta + n + 1$, so we enter as `=B3+B4+C11+1` in cell `D38`. Finally, the 'z' argument of the function is $1/(1+d)$, which is entered as `=1/(1+B6)` in cell `D39`.

  Starting at cell `F35`, we compute each term of the series for $j = 0, \ldots, 150$. (The values of the index j are given in cells `E35:E185`.) As noted above, the value of u_0 is 1 (cell `F35`). To compute the value of u_1, we multiply u_0 by

  $$\frac{(a+j-1)(b+j-1)}{(c+j-1)j} z$$

  evaluated at $j = 1$. We therefore compute u_1 by entering

  ```
  =F35*($D$36+E36-1)*($D$37+E36-1)
     $D$39/(($D$38+E36-1)*E36)
  ```

  in cell `F36`. We copy this formula down to cell `F185`, which corresponds to u_{150}. Summing these 151 terms gives us the numerical value of the Gaussian hypergeometric function for this set of function parameters (cell `D35`).[2]

- Finally we enter

  ```
  =EXP(GAMMALN($B$1+A11+1)+GAMMALN($B$2+C11-A11)
   -GAMMALN($B$1+$B$2+C11+1))*EXP(GAMMALN($B$3)
  +GAMMALN($B$4+C11+1)-GAMMALN($B$3+$B$4+C11+1))/
           ($E$1*$E$3*(1+$B$6))*$D$35/F11
  ```

  in cell `G11`, which evaluates equation (9), and copy it down to cell `G32`.

This gives us the present value of the expected number of future transactions for each of the 22 recency/frequency combinations. (These numbers are not reported in the paper.)

[2] In some settings, we may find that the value of u_{150} is not approaching zero. This means we need to add more terms to the series: we simply add values of the index j below cell `E185`, copy down cell `F185`, and make the required changes to the formula in cell `D35`.

Appendix: Potential Estimation Problem

When Solver returns `#DIV/0!` as the value of the log-likelihood function with γ and δ taking on "large" values (in the order of hundreds or thousands), the large values of γ and δ mean that $B(\gamma, \delta)$ takes on a value of zero (which causes the error when we compute the model likelihood function). These large values for γ and δ mean that the beta distribution for θ is effectively a spike at $E(\Theta) = \gamma/(\gamma + \delta)$; that is, there is no heterogeneity in θ.

While we could reformulate the underlying model — removing the beta heterogeneity associated with the geometric distribution, giving us the G/BB (geometric/beta-Bernoulli) model — it is easier to place a constraint on the potential magnitude of γ and δ. Instead of constraining γ and/or δ directly, we place a constraint on $\gamma + \delta$; a limit of 1000 should suffice. To implement this, we enter `=B3+B4` in cell `B5`, *Add* `$B$5 <= 1000` to the existing Solver constraint (`$B$1:$B$4 >= 0.0001`) and click the *Solve* button.