

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

Peter S. Fader
www.petefader.com

Bruce G. S. Hardie
www.brucehardie.com

Paul D. Beger[†]

June 2005

1. Introduction

This note describes how to implement the BG/BB model for customer-base analysis¹ using Microsoft Excel.

We first consider how to estimate the model parameters, doing this in two ways:

- i. an “informal” construction of the log-likelihood function, intended to illustrate the underlying logic of the model, and
- ii. a “formal” coding-up of the log-likelihood function as presented in the paper.

We then show how to compute the following three quantities:

- i. $P(\text{alive} | x, t_x, n)$, the probability that a customer with purchase history (x, t_x, n) will be “alive” at the $(n + 1)$ th transaction opportunity,
- ii. $E(X^* | n^*, x, t_x, n)$, the expected number of transactions across the next n^* transaction opportunities (i.e., in the interval $(n, n + n^*]$) by a customer with purchase history (x, t_x, n) , and
- iii. $DET(d | x, t_x, n)$, the present-value of the expected number of future transactions (discounted expected transactions, DET) for a customer with purchase history (x, t_x, n) .

[†]© 2005 Peter S. Fader, Bruce G. S. Hardie, and Paul D. Berger. This document and the associated spreadsheet can be found at <http://brucehardie.com/notes/010/>.

¹Fader, Peter S., Bruce G. S. Hardie, and Paul D. Berger (2005), “Customer-Base Analysis with Discrete-Time Transaction Data,” <http://brucehardie.com/papers/020/>.

The specific steps are outlined in Sections 2–6 below. All these sections should be read in conjunction with the Excel workbook `bgbb_2005-06.xls`. (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 `bgbb_2005-06.xls` as a guide.)

2. An Informal Construction of the Log-likelihood Function

In this section, we present an “informal” construction of the log-likelihood function that starts from first principles. As our primary objective is to illustrate the underlying logic of the model, any reader comfortable with the formal presentation of the likelihood function in the paper can jump straight to Section 3.

Our data are for a cohort of 6094 customers who took their first-ever cruise with “Joyful Voyages, Inc.” in 1993. We have information on their repeat-buying behavior for the period 1994–1997.

While there is a small number of customers who make more than one cruise a year, customer behavior has been summarized in terms of whether or not the customer took a cruise in each year. As such, the behavior of each customer is characterized by one of 16 binary strings of length four: from 1111 (for a customer who took a cruise in 1994, 1995, 1996, and 1997) to 0000 (for a customer who took no repeat cruises). The complete dataset is given in the worksheet `Raw Data`; a slightly modified version of the dataset is given in the worksheet `Sorted Raw Data`, in which the data are sorted by the year of last cruise. (See Figure 1 for the Excel dialogue box associated with this sorting operation.)



Figure 1: Sorting the data by year of last cruise

We start our informal construction of the model log-likelihood function by making a copy of the `Sorted Raw Data` worksheet—let’s call it `LL - Informal Construction`. As discussed in the paper, we are interested in

three summary statistics of a customer’s purchasing history: (x, t_x, n) , where x is the number of transactions that occurred in n transaction opportunities (i.e., “frequency”) and t_x ($0 \leq t_x \leq n$) is the time of the last transaction (i.e., “recency”). The first thing we need to do is create these summary measures for each of the sixteen possible purchase patterns. We do this for purchase pattern 1 in the following manner:

- The number of transactions (x) is simply the number of ones in the binary string. We compute this quantity by entering `=SUM(B2:E2)` in cell H2.
- Our recency measure (t_x) is the transaction opportunity on which the last transaction occurred. We determine this by entering the following expression in cell I2:

`=IF(E2=1,4,IF(D2=1,3,IF(C2=1,2,IF(B2=1,1,0)))`

- The number of transaction opportunities is the length of the binary string. While we could simply type the number 4 into cell J2, we use the following formula: `=COUNT(B2:E2)`.

We copy this block of three cells (H2:J2) down to row 17.

Now let’s consider each of the 16 observed purchase patterns:

- With patterns 1–8, $t_x = 4$ (i.e., a transaction occurred in 1997); as noted in cells P2:P9, this means the associated customers must have been “alive” at all four transaction opportunities.
- With patterns 9–12, $t_x = 3$ (i.e., the individual’s last transaction occurred in 1996), which could be the result of one of two “alive” scenarios:
 - the customer was alive at all four transaction opportunities (as noted in cells P10:P13) but didn’t make a purchase in 1997, or
 - the customer was alive for just the first three transaction opportunities (as noted in cells Q10:Q13), and was therefore by definition couldn’t make a purchase in 1997.
- And so on, with there being five “alive” scenarios associated with pattern 16, ranging from being alive at all four transaction opportunities to being alive at none of them.

In constructing the log-likelihood function, we will first compute the beta-geometric probabilities of each of these “alive” scenarios (cells P21:T36) and then compute the beta-Bernoulli probabilities of the associated purchase string conditional on the assumed length of the customer’s lifetime (cells P40:T55).

These two sets of probabilities are functions of four parameters: α and β for the beta-Bernoulli probabilities, and γ and δ for the beta-geometric probabilities. In order to create the corresponding expressions 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 α , β , γ , and δ . We locate these parameter values in cells C19:C22, respectively.

Now the beta-geometric pmf and survivor function are, respectively,

$$P(T = t | \gamma, \delta) = \frac{B(\gamma + 1, \delta + t - 1)}{B(\gamma, \delta)}, \quad t = 1, 2, \dots$$

$$S(t | \gamma, \delta) = \frac{B(\gamma, \delta + t)}{B(\gamma, \delta)}, \quad t = 1, 2, \dots$$

We can therefore compute the probability of each “alive” scenario in the following manner:

- As we will be using it a number of times, we first compute the quantity $B(\gamma, \delta)$ separately in cell F21. Noting that

$$B(\gamma, \delta) = \frac{\Gamma(\gamma)\Gamma(\delta)}{\Gamma(\gamma + \delta)}$$

we use the following expression:

$$=\text{EXP}(\text{GAMMALN}(C21)+\text{GAMMALN}(C22)-\text{GAMMALN}(C21+C22))$$

(With starting values of $\gamma = 1$ and $\delta = 1$, this quantity equals 1.)

- The probability of being alive at all four transaction opportunities is the beta-geometric probability that “death” occurs sometime after period 4 (i.e., $S(4 | \gamma, \delta)$). We compute this by entering

$$=\text{EXP}(\text{GAMMALN}(\$C\$21)+\text{GAMMALN}(\$C\$22+4)-\text{GAMMALN}(\$C\$21+\$C\$22+4))/\$F\$21$$

in cell P21. With starting values of $\gamma = 1$ and $\delta = 1$, this probability equals 0.2. We copy this formula down to row 36.

- The probability of only being alive for the first three transaction opportunities is the beta-geometric probability of “dying” at the beginning of the fourth transaction opportunity. We compute this by entering

$$=\text{EXP}(\text{GAMMALN}(\$C\$21+1)+\text{GAMMALN}(\$C\$22+3)-\text{GAMMALN}(\$C\$21+\$C\$22+4))/\$F\$21$$

in cell Q29. With starting values of $\gamma = 1$ and $\delta = 1$, this probability equals 0.05. We copy this formula down to row 36.

- The probability of only being alive for the first two transaction opportunities is the beta-geometric probability of “dying” at the beginning of the third transaction opportunity. We compute this by entering

$$=EXP(GAMMALN(\$C\$21+1)+GAMMALN(\$C\$22+2)-GAMMALN(\$C\$21+\$C\$22+3))/\$F\$21$$

in cell R33. With starting values of $\gamma = 1$ and $\delta = 1$, this probability equals 0.0833. We copy this formula down to row 36.

- The probability of only being alive for the first transaction opportunity is the beta-geometric probability of “dying” at the beginning of the second transaction opportunity. We compute this by entering

$$=EXP(GAMMALN(\$C\$21+1)+GAMMALN(\$C\$22+1)-GAMMALN(\$C\$21+\$C\$22+2))/\$F\$21$$

in cell S35. With starting values of $\gamma = 1$ and $\delta = 1$, this probability equals 0.1667. We copy this formula to cell S36.

- Finally, the probability of not being alive at any of the four transaction opportunities is the beta-geometric probability of “dying” at the beginning of the first transaction opportunity. We compute this by entering

$$=EXP(GAMMALN(\$C\$21+1)+GAMMALN(\$C\$22)-GAMMALN(\$C\$21+\$C\$22+1))/\$F\$21$$

in cell T36. With starting values of $\gamma = 1$ and $\delta = 1$, this probability equals 0.5.

Turning to the beta-Bernoulli probabilities of the associated purchase strings conditional on the assumed length of the customer’s lifetime, we note that the probability of any given pattern of x 1’s in a string of length y is

$$\frac{B(\alpha + x, \beta + y - x)}{B(\alpha, \beta)}.$$

- As we will be using it a number of times, we first compute the quantity $B(\alpha, \beta)$ separately in cell F19. Noting that

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$$

we use the following expression:

$$=EXP(GAMMALN(C19)+GAMMALN(C20)-GAMMALN(C19+C20))$$

(With starting values of $\alpha = 1$ and $\beta = 1$, this quantity equals 1.)

- For all 16 purchase patterns, it is possible that the customer was “alive” at all four transaction opportunities. In cells P40:P55 we compute the probability of observing the x transactions associated with each of the 16 purchase patterns for $y = 4$. For pattern 1, we enter the following expression in cell P40:

$$=EXP(GAMMALN(\$C\$19+\$H2)+GAMMALN(\$C\$20+4-\$H2)-GAMMALN(\$C\$19+\$C\$20+4))/\$F\$19$$

(With starting values of $\alpha = 1$ and $\beta = 1$, this quantity equals 0.2.) We copy this expression down to row 55, noting that these 16 cells sum to 1.0.

- For purchase patterns 9–16, we also consider the possibility that the customer was only “alive” for the first three transaction opportunities. In cells Q48:Q55 we compute the probability of observing the x transactions associated with each of these eight purchase pattern for $y = 3$. For pattern 9, we enter the following expression in cell Q48:

$$=EXP(GAMMALN(\$C\$19+\$H10)+GAMMALN(\$C\$20+3-\$H10)-GAMMALN(\$C\$19+\$C\$20+3))/\$F\$19$$

(With starting values of $\alpha = 1$ and $\beta = 1$, this quantity equals 0.25.) We copy this expression down to row 55, noting that these eight cells sum to 1.0.

- For purchase patterns 13–16, we also consider the possibility that the customer was only “alive” for the first two transaction opportunities. In cells R52:R55 we compute the probability of observing the x transactions associated with each of these four purchase pattern for $y = 2$. For pattern 13, we enter the following expression in cell R52:

$$=EXP(GAMMALN(\$C\$19+\$H14)+GAMMALN(\$C\$20+2-\$H14)-GAMMALN(\$C\$19+\$C\$20+2))/\$F\$19$$

(With starting values of $\alpha = 1$ and $\beta = 1$, this quantity equals 0.3333.) We copy this expression down to row 55, noting that these four cells sum to 1.0.

- Next, for purchase patterns 15–16, we also consider the possibility that the customer was only “alive” for the first transaction opportunity. In cells S54:S55 we compute the probability of observing the x transactions associated with each of these two purchase pattern for $y = 1$. For pattern 15, we enter the following expression in cell S54:

$$=EXP(GAMMALN(\$C\$19+\$H16)+GAMMALN(\$C\$20+1-\$H16)-GAMMALN(\$C\$19+\$C\$20+1))/\$F\$19$$

(With starting values of $\alpha = 1$ and $\beta = 1$, this quantity equals 0.5.) We copy this formula to cell S55, noting that these two cells sum to 1.0.

- Finally, for purchase pattern 16, we have to consider the possibility of not being alive at any of the four transaction opportunities. With this scenario, the probability of making zero purchases equals 1, which we enter in cell T55.

We are now in a position to compute $L(\alpha, \beta, \gamma, \delta | x, t_x, n)$ for each of the 16 purchase patterns. For each purchase string, we multiply the conditional probability of the observed pattern of purchases (conditioned on a given “alive” scenario) by the probability of the particular “alive” scenario, and sum across the set of “alive” scenarios. This is achieved by entering =SUMPRODUCT(P21:T21,P40:T40) in cell M2, which we then copy down to row 17. (Note that cells M2:M17 sum to 1.0.)

The next step is to multiply the log of $L(\alpha, \beta, \gamma, \delta | x, t_x, n)$ for each of the 16 purchase patterns by the corresponding number of people who have that pattern. For the first purchase pattern we enter =F2*LN(M2) in cell L2; we then copy this expression down to cell L17. The sum of cells L2:L17 is found in cell C24; this is the value of the log-likelihood function given the values for the four model parameters in cells C19:C22. (With starting values of 1.0 for all four parameters, $LL = -7965.3$.)

Given these sample data, we find the maximum likelihood estimates of the four model parameters by maximizing the log-likelihood function. We do this using the Excel add-in Solver, available under the “Tools” menu. The *target cell* is the value of the log-likelihood, (cell C24. We wish to *maximize* this by *changing* cells C19:C22. 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 C19:C22 are \geq a small positive number (e.g., 0.00001) — see Figure 2.

Clicking the *Solve* button, Solver stops at the following parameter values: $\alpha = 0.69$, $\beta = 5.27$, $\gamma = 216.14$, $\delta = 2069.31^2$; the only problem is that the value of the log-likelihood function is #DIV/0!. What’s going on? With such large values of γ and δ , $B(\gamma, \delta)$ is so close to zero that it is being treated as zero when we compute the beta-geometric probabilities in cells P21:T36, hence the “divide by zero” error.

Setting δ to 1900 (and leaving α , β , γ at 0.69, 5.27, 216.14 respectively), we “fire-up” Solver again, and find that it converged to a solution where the maximum value of the log-likelihood function is -7130.7 , associated

²The exact values may vary depending on the computer being used.



Figure 2: Solver Settings

$\alpha = 0.66$, $\beta = 5.19$, $\gamma = 175.64$, $\delta = 1904.61$. Can we be sure that we have reached the maximum of the log-likelihood function? Using this solution as the set of starting values for the parameters, we “fire up” Solver again to see if it can improve on this solution. It can’t; we therefore conclude that the numbers given in cells C19:C22 are the maximum likelihood estimates of the model parameters. This solution is given in the worksheet LL - Informal Construction (II).

As an aside, we note that the magnitude of $\hat{\gamma}$ and $\hat{\delta}$ suggests that the beta distribution for θ is effectively a spike at $E(\theta) = \gamma/(\gamma + \delta) = 0.084$. In Appendix A, we estimate the parameters of a model that does not allow for heterogeneity in θ , the geometric/beta-Bernoulli (G/BB) model. We find that the value of this alternative model’s log-likelihood function is the same as that of the BG/BB model. This suggests that, for this dataset, there is no heterogeneity in the dropout parameter θ .

3. A Formal Construction of the Log-likelihood Function

Having estimated the model parameters using an informal construction of the model log-likelihood function, we now consider how to “code-up” the formal log-likelihood function as presented in the paper. Before doing so, let us first consider why we would want to do this, as we have already estimated the model parameters using the informal (albeit slightly cumbersome) method presented above. For settings where n , the number of transaction opportunities, is small, this informal approach is fine. However, as n grows, the number of purchase patterns we need to consider rapidly increases and the spreadsheet becomes very cumbersome. For example, if $n = 10$, there are $2^{10} = 1024$ different binary strings to consider! The formal likelihood function presented in the paper conditions on the a given recency/frequency pattern, rather than the entire purchase string. As there are $n(n + 1)/2 + 1$ possible recency/frequency patterns (56 when $n = 10$), this approach is far more manageable for large n .

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

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

For a sample of N customers, where customer i 's purchase history is denoted by (x_i, t_{x_i}, n_i) , the sample log-likelihood function is given by

$$LL(\alpha, \beta, \gamma, \delta) = \sum_{i=1}^N \ln [L(\alpha, \beta, \gamma, \delta | x_i, t_{x_i}, n_i)]. \quad (2)$$

When $n_i = n$ for all i , 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 set of 2^n possible binary strings that characterize all possible purchase patterns. In fact, we don't even have to loop over that many possible purchase patterns. Looking closely at equation (1), we see that the likelihood function is not conditioned on the complete binary string, it is only conditioned on the frequency (x) and recency (t_x) measures. We therefore only have to loop over the $n(n+1)/2+1$ possible recency/frequency patterns: for each pattern, we multiply the log of $L(\alpha, \beta, \gamma, \delta | x, t_x, n)$ by the number of people associated with that particular recency/frequency pattern.

Our first step is to create a recency/frequency summary of the dataset. We copy cells **B1:J17** from **LL - Informal Construction** and paste them into a new worksheet (let's call it **(x, t_x, n) Summary**). Sorting this whole block of cells by recency and frequency (see Figure 3 for the Excel dialogue box associated with this sorting operation), we see that some recency/frequency patterns appear multiple times; for example, there are three binary string patterns with $x = 3, t_x = 4$.

In rows 20–30, we manually create the 11 recency/frequency summaries of the 16 binary strings that characterize all possible repeat purchase patterns for 1994–1997. We copy this recency/frequency summary into a blank worksheet (let's call it **LL - Formal Construction**), leaving a few blank rows at the top. We are now in a position to “code-up” the formal log-likelihood function.

- As before, we need to specify starting values for the four model parameters; we start with 1.0 for α , β , γ , and δ , locating 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



Figure 3: Sorting the data by recency and frequency

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

and

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

respectively. (With starting values of $\gamma = 1$ and $\delta = 1$, 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 coded in cell I9 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 I19.

- 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 $n = 4$) is 3. In rows J8:M8 we enter the possible values that i could take on: 0, 1, 2, 3.

In cells J9:M19, we are going to enter an expression for the summand,

$$\frac{B(\alpha + x, \beta + t_x - x + i)}{B(\alpha, \beta)} \frac{B(\gamma + 1, \delta + t_x + i)}{B(\gamma, \delta)}. \quad (3)$$

However, we don't evaluate this for all values of i ; the upper limit depends on 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 H9 and copy this down to cells H19.

We then enter the following expression

$$\begin{aligned} &=IF(J\$8<=\$H9,EXP(GAMMALN(\$B\$1+\$A9) \\ &\quad +GAMMALN(\$B\$2+\$B9-\$A9+J\$8) \\ &\quad -GAMMALN(\$B\$1+\$B\$2+\$B9+J\$8))/\$E\$1 \\ &*EXP(GAMMALN(\$B\$3+1)+GAMMALN(\$B\$4+\$B9+J\$8) \\ &\quad -GAMMALN(\$B\$3+\$B\$4+\$B9+J\$8+1))/\$E\$3,0) \end{aligned}$$

in cell J9. This is evaluating equation (3) 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 M9, and then copy this block of cells down to row 19.

- Having computed all the elements in equation (1), we sum up all these elements by entering `=SUM(I9:M9)` 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 at the values of $\alpha, \beta, \gamma, \delta$ given in cells B1:B4. We copy this down to cell F19.
- Finally, we multiply the number of people associated with each of the 11 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 C19. The sum of these 11 cells is entered in cell B6: `=SUM(E9:E19)`. This is the value of the log-likelihood function given the values for the four model parameters in cells B1:B4. (With starting values of 1.0 for all four parameters, $LL = -7965.3$.)

As before, we find the maximum likelihood estimates of the four model parameters by maximizing the log-likelihood function using Solver.³

4. Computing P(alive)

Now that we've estimated the model parameters, we can turn our attention to computing several quantities of managerial interest. The first we will consider is the probability that a customer with purchase history (x, t_x, n) will be "alive" at the $(n + 1)$ th transaction opportunity. The formula for

³We may be wondering how well the model actually fits the data. We explore this in Appendix B, where we show how to create a plot of model fit.

this quantity is

$$P(\text{alive in period } n + 1 \mid x, t_x, n, \alpha, \beta, \gamma, \delta) \\ = \frac{B(\alpha + x, \beta + n - x)B(\gamma, \delta + n + 1)}{B(\alpha, \beta)B(\gamma, \delta)} \bigg/ L(\alpha, \beta, \gamma, \delta \mid x, t_x, n)$$

This is actually very easy to evaluate, as we've already created expressions for $B(\alpha, \beta)$, $B(\gamma, \delta)$, and $L(\alpha, \beta, \gamma, \delta \mid x, t_x, n)$ as part of the parameter estimation process.

- We first make a copy of the worksheet **LL - Formal Construction** (let's call it **P(alive)**) and insert two columns to the right of column G.
- Next we enter the following expression in cell H9

$$=EXP(GAMMALN(\$B\$1+A9)+GAMMALN(\$B\$2+C9-A9) \\ -GAMMALN(\$B\$1+\$B\$2+C9))*EXP(GAMMALN(\$B\$3) \\ +GAMMALN(\$B\$4+C9+1)-GAMMALN(\$B\$3+\$B\$4+C9+1))/(E\$1*E\$3)/F9$$

and copy it down to cell H19.

We now have the probability that a customer will be “alive” in 1998 for all 11 recency/frequency combinations.

5. Computing Conditional Expectations

The second quantity of interest is the expected number of transactions across the next n^* transaction opportunities for a customer with purchase history (x, t_x, n) . The formula for this *conditional expectation* is

$$E(X^* \mid n^*, x, t_x, n, \alpha, \beta, \gamma, \delta) = \frac{B(\alpha + x + 1, \beta + n - x)}{B(\alpha, \beta)B(\gamma, \delta)} \\ \times \frac{B(\gamma - 1, \delta + n + 1) - B(\gamma - 1, \delta + n + n^* + 1)}{L(\alpha, \beta, \gamma, \delta \mid x, t_x, n)}$$

As with the calculation of $P(\text{alive})$, this is very easy to evaluate as we've already created expressions for $B(\alpha, \beta)$, $B(\gamma, \delta)$, and $L(\alpha, \beta, \gamma, \delta \mid x, t_x, n)$ as part of the parameter estimation process.

- We first make a copy of the worksheet **LL - Formal Construction** (let's call it **Conditional Expectations**), insert two columns to the right of column G, and insert one row after row 4.
- We then need to specify n^* , the horizon over which we are computing the conditional expectation. For this example, we will compute the expected number of transactions in 1998–2001, so we enter a value of $n^* = 4$ in cell B5.

- Finally we enter the following expression in cell H10

$$\begin{aligned}
&= (\text{EXP}(\text{GAMMALN}(\text{\$B\$3}-1) + \text{GAMMALN}(\text{\$B\$4} + \text{C10} + 1) \\
&\quad - \text{GAMMALN}(\text{\$B\$3} + \text{\$B\$4} + \text{C10})) - \text{EXP}(\text{GAMMALN}(\text{\$B\$3}-1) \\
&\quad + \text{GAMMALN}(\text{\$B\$4} + \text{C10} + \text{\$B\$5} + 1) - \text{GAMMALN}(\text{\$B\$3} + \text{\$B\$4} + \text{C10} + \text{\$B\$5}))) \\
&\quad * \text{EXP}(\text{GAMMALN}(\text{\$B\$1} + \text{A10} + 1) + \text{GAMMALN}(\text{\$B\$2} + \text{C10} - \text{A10}) \\
&\quad - \text{GAMMALN}(\text{\$B\$1} + \text{\$B\$2} + \text{C10} + 1)) / (\text{E\$1} * \text{E\$3}) / \text{F10}
\end{aligned}$$

and copy it down to cell H20.

We now have computed the expected number of transactions in 1998–2001 for each of the 11 recency/frequency combinations.

5. Computing DET

Finally, we consider how to compute the present value of the expected number of future transactions (discounted expected transactions, *DET*) for a customer with purchase history (x, t_x, n) . The formula for this, with a specified discount rate d , is

$$\begin{aligned}
&DET(d | x, t_x, n, \alpha, \beta, \gamma, \delta) \\
&= \frac{B(\alpha + x + 1, \beta + n - x)B(\gamma, \delta + n + 1)}{B(\alpha, \beta)B(\gamma, \delta)(1 + d)} \\
&\quad \times \frac{{}_2F_1(1, \delta + n + 1; \gamma + \delta + n + 1; \frac{1}{1+d})}{L(\alpha, \beta, \gamma, \delta | x, t_x, n)}, \tag{4}
\end{aligned}$$

where ${}_2F_1(\cdot)$ is the Gaussian hypergeometric function. This function is the power series of the form

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

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

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

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

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

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., $j = 0, 1, \dots, 150$).

- As before, we start by making a copy of the worksheet **LL - Formal Construction** (let’s call it **DET**), insert two columns to the right of column **G**, and insert 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 **K25**. The ‘ b ’ parameter is $\delta + n + 1$, entered as **=B4+C11+1** in cell **K26**. The ‘ c ’ parameter is $\gamma + \delta + n + 1$, so we enter as **=B3+B4+C11+1** in cell **K27**. Finally, the ‘ z ’ argument of the function is $1/(1 + d)$, which is entered as **=1/(1+B6)** in cell **K28**.

Starting at cell **M24**, we compute each term of the series for $j = 0, \dots, 150$. (The values of the index j are given in cells **L24:L174**.) As noted above, the value of u_0 is 1 (cell **M24**). 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

$$\begin{aligned} &=M24*(\$K\$25+L25-1)*(\$K\$26+L25-1) \\ &\quad \$K\$28/((\$K\$27+L25-1)*L25) \end{aligned}$$

in cell **M25**. We copy this formula down to cell **M174**, 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 **K24**).

- Finally we enter

$$\begin{aligned} &=EXP(GAMMALN(\$B\$1+A11+1)+GAMMALN(\$B\$2+C11-A11) \\ &\quad -GAMMALN(\$B\$1+\$B\$2+C11+1))*EXP(GAMMALN(\$B\$3 \\ &\quad +GAMMALN(\$B\$4+C11+1)-GAMMALN(\$B\$3+\$B\$4+C11+1)) \\ &\quad *\$K\$24/(E\$1*E\$3*(1+\$B\$6))/F11 \end{aligned}$$

in cell H11, which evaluates equation (4), and copy it down to cell H21.

This gives us the present value of the expected number of future transactions for each of the 11 recency/frequency combinations.

Appendix A: The G/BB Model

As noted in Section 2, the magnitude of $\hat{\gamma}$ and $\hat{\delta}$ suggests that the beta distribution for θ is effectively a spike at $E(\theta) = \gamma/(\gamma + \delta) = 0.084$. In this appendix we consider a constrained version of the BG/BB model that does not allow for heterogeneity in the underlying “death” parameter θ ; we call this the geometric/beta-Bernoulli (G/BB) model.

First we make a copy of the LL - Informal Construction worksheet — let’s call it **G_BB Model**. We need to replace the beta-geometric probabilities of each of the “alive” scenarios (cells P21:T36) with the equivalent geometric distribution probabilities. The geometric distribution has just one parameter, θ , which we locate in cell C21. (We delete the old δ parameter as well as the expression for $B(\gamma, \delta)$.) As starting values, we choose 1.0 for α and β , and 0.5 for θ .

The geometric pmf and survivor function are, respectively,

$$P(T = t | \theta) = \theta(1 - \theta)^{t-1}, \quad t = 1, 2, \dots$$
$$S(t | \theta) = (1 - \theta)^t, \quad t = 1, 2, \dots$$

We can therefore compute the probability of each “alive” scenario in the following manner:

- The probability of being alive at all four transaction opportunities is the geometric probability that “death” occurs sometime after period 4 (i.e., $S(4 | \theta)$). We compute this by entering $= (1 - \theta)^4$ in cell P21. With a starting value of $\theta = 0.5$, this probability equals 0.0625. We copy this formula down to row 36.
- The probability of only being alive for the first three transaction opportunities is the geometric probability of “dying” at the beginning of the fourth transaction opportunity. We compute this by entering $= (1 - \theta)^3 * \theta$ in cell Q29. With a starting value of $\theta = 0.5$, this probability equals 0.0625. We copy this formula down to row 36.
- The probability of only being alive for the first two transaction opportunities is the geometric probability of “dying” at the beginning of the third transaction opportunity. We compute this by entering $= (1 - \theta)^2 * \theta$ in cell R33. With a starting value of $\theta = 0.5$, this probability equals 0.125. We copy this formula down to row 36.
- The probability of only being alive for the first transaction opportunity is the geometric probability of “dying” at the beginning of the second transaction opportunity. We compute this by entering $= (1 - \theta) * \theta$ in cell S35. With a starting value of $\theta = 0.5$, this probability equals 0.25. We copy this formula to cell S36.

- Finally, the probability of not being alive at any of the four transaction opportunities is the geometric probability of “dying” at the beginning of the first transaction opportunity. We compute this by entering =C21 in cell T36, which is simply the value of the θ parameter.

Everything else that makes up the log-likelihood function remains the same. Using Solver, we find that the maximum value of the log-likelihood function is -7130.7 , associated $\alpha = 0.66$, $\beta = 5.19$, $\theta = 0.084$. This value of the log-likelihood function is the same as that of the BG/BB model, which suggests that, for this dataset, there is no heterogeneity in the individual-level dropout parameter θ .

Appendix B: Creating a Plot of Model Fit

How well does the BG/BB model fit the actual data? In order to assess this, we would like to create a plot of the predicted versus actual trip frequencies.

- We start by copying the *values* of cells F1:F17, H1:H17, and M1:M17 from LL - Informal into a blank worksheet (which we call Plot of Model Fit). (Note that $L(.|x,t_x,n)$ numbers are the probabilities of observing each of the 16 possible repeat purchase patterns. As such, they obviously sum to 1.0.)
- We first determine the number of people who took a repeat cruise in 0, 1, 2, 3, or all 4 of the four years in which a repeat cruise could have taken place. We do this by performing a pivot table analysis on these data where the row field is **x** and the data item is # cust..
- We compute the expected number of people taking a repeat cruise in 0, 1, 2, 3, or all 4 of the four years in which a repeat cruise could have taken place in the following manner:
 - We perform another pivot table analysis on these data where the row field is **x** and the data item is $L(.|x,t_x,n)$. This gives us the probability of a randomly-chosen customer taking a repeat cruise in 0, 1, 2, 3, or all 4 of the four years in which a repeat cruise could have taken place.
 - Multiplying these probabilities by the size of the customer base (6094) gives us the expected number of customers associated with each of the five repeat cruise-year categories.
- We plot the predicted and actual numbers, observing that the predicted distribution of repeat cruise-years is extremely close to the actual distribution.