Fitting the sBG Model to Multi-Cohort Data

Peter S. Fader
www.petefader.com

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

January 2007

1 Introduction

Fader and Hardie (2007) introduce the shifted-beta-geometric (sBG) distribution as a model of customer contract duration in discrete-time contractual settings, and show how the model parameters can be estimated using data from a single cohort of customers.

In many situations, we will have data from more than one cohort of customers. When these cohorts are defined by the time of acquisition, we end up with a data structure of the form given in Table 1, where n_{ii} is the number of customers acquired in period i ("cohort i customers"), n_{ij} is the number of cohort i customers still active in period j ($j > i$), and $n_{.j}$ is the total number of active customers in period j.

Cohort	Calendar Time →				
1	n_{11}	n_{12}	n_{13}	...	n_{1I}
2		n_{22}	n_{23}	...	n_{2I}
3			n_{33}	...	n_{3I}
⋮				⋱	⋮
I					n_{II}
	$n_{.1}$	$n_{.2}$	$n_{.3}$	...	$n_{.I}$

Table 1: Structure of Multi-Cohort Data

Looking at the first cohort, we see that the firm acquired n_{11} customers in the first period; n_{12} were active in the second period, which means $n_{11} - n12$ customers did not renew their contract at the end of period one. And so so.

We could assume that each cohort is governed by its own process and fit separate sBG models for cohorts 1 to $I-2$. (We do not have enough cohort-specific data to be able to estimate separate models for the last two cohorts.) However the problem with this is that every new cohort has one less period of information than its temporal predecessor, which may result in less confidence in the model parameter estimates for the cohorts with fewer data points.

Therefore the natural starting point in such a situation is to pool the cohorts, assuming that each cohort is the realization of a common underlying contract duration process, and to estimate one set of parameters using all the data. When we have a dataset of the form given in Table 1, it is easy to compute the maximum likelihood estimates the two model parameters, even using Microsoft Excel — see Section 2.

In many settings, however, we do not have the full information matrix; we may only have subsets of the data, as illustrated in Table 2. In Section 3 we show how to estimate the two model parameters when faced with such limited information, illustrating how to do this in Excel.

Before discussing model estimation, let us briefly review the shifted-beta-geometric (sBG) distribution. Its probability mass function and survivor function are

$$P(T=t\,|\,\alpha,\beta)=\frac{B(\alpha+1,\beta+t-1)}{B(\alpha,\beta)}\,,\quad t=1,2,\ldots$$
$$S(t\,|\,\alpha,\beta)=\frac{B(\alpha,\beta+t)}{B(\alpha,\beta)}\,,\quad t=1,2,\ldots$$

It is not actually necessary for us evaluate beta functions in order to compute these quantities. The sBG probabilities can be computed using the following forward-recursion formula from $P(T=1)$:

$$P(T=t\,|\,\alpha,\beta)=\begin{cases}\dfrac{\alpha}{\alpha+\beta} & t=1\\[2ex] \dfrac{\beta+t-2}{\alpha+\beta+t-1}P(T=t-1) & t=2,3,\ldots\end{cases}\tag{1}$$

Once we have the $P(T=t)$, we can compute the survivor function using the following expression:

$$S(t\,|\,\alpha,\beta)=1-\sum_{i=1}^{t}P(T=i\,|\,\alpha,\beta)\,.\tag{2}$$

Case 1

Cohort	Calendar Time →				
1	n_{11}				n_{1I}
2		n_{22}			n_{2I}
3			n_{33}		n_{3I}
⋮				⋱	⋮
I					n_{II}

Case 2

Cohort	Calendar Time →				
1	n_{11}				
2		n_{22}			
3			n_{33}		
⋮				⋱	
I					n_{II}
	$n_{.1}$	$n_{.2}$	$n_{.3}$	...	$n_{.I}$

Case 3

Cohort	Calendar Time →				
1					n_{1I}
2					n_{2I}
3					n_{3I}
⋮					⋮
I					n_{II}
	$n_{.1}$	$n_{.2}$	$n_{.3}$	...	$n_{.I}$

Case 4

Cohort	Calendar Time →				
1				n_{1I-1}	n_{1I}
2				n_{2I-1}	n_{2I}
3				n_{3I-1}	n_{3I}
⋮					⋮
I					n_{II}

Table 2: Limited Information Data Structures

2 Parameter Estimation with Full Information

Given the data presented in Table 1, the corresponding sample log-likelihood function is

$$LL(\alpha,\beta\,|\,\text{data}) = \sum_{i=1}^{I-1}\left\{\sum_{j=i}^{I-1}(n_{ij}-n_{i(j+1)})\ln\left[P(T=j-i+1\,|\,\alpha,\beta)\right] + n_{iI}\ln\left[S(I-i\,|\,\alpha,\beta)\right]\right\}, \tag{3}$$

the maximum of which can be found using standard numerical optimization methods.

It is a simple exercise to "code up" (3) in Excel. To illustrate this, we use the data presented in the `Raw Data` worksheet in the Excel workbook `multi-cohort_sBG_estimation.xls`. This gives five years of customer data ($I = 5$) for a hypothetical setting where 10,000 customers were acquired each year.

Consider the worksheet `Full Information`:

- Given the parameter values located in cells `B9:B10`, we compute $P(T = t)$ for $t = 1,\ldots,4$ in cells `B14:E14` using the forward recursion given in (1). We then compute $S(T)$ for $t = 1,\ldots,4$ in cells `B15:E15` using (2).
- We compute the values of $n_{ij} - n_{i(j+1)}$, the number of customers not renewing their contracts each year, in cells `B17:E20`. We enter the corresponding values of n_{iI} in cells `F17:F20`.
- We enter the corresponding values of $P(T = j - i + 1\,|\,\alpha,\beta)$ in cells `B22:E25`, referring back to the appropriate entries in cells `B14:E14`. We enter the corresponding values of $S(I - i\,|\,\alpha,\beta)$ in cells `F22:F25`, referring back to the appropriate entries in cells `B15:E15`.
- We enter in cells `B27:F30` the product of each element of cells `B17:F20` and the natural log of the corresponding element of cells `B22:F25`. The sum of this block of numbers is located in cell `B11` and is the value of the log-likelihood function for the parameter values located in cells `B9:B10`.

We obtain the maximum likelihood estimates of α and β by using the Excel add-in Solver to find the values of α and β (cells `B9:B10`) that maximize the value of the log-likelihood function (cell `B11`), subject to the constraint that cells `B9:B10` are $\geq$ a small positive number (e.g., 0.0001). With starting values of $\alpha = 1, \beta = 1$, we find that the full-information maximum likelihood estimates of the model parameters are $\hat{\alpha} = 3.80$ and $\hat{\beta} = 15.19$.

3 Parameter Estimation with Limited Information

We now consider how to estimate the two model parameters when faced with limited information situations of the form presented in Table 2. (Note that it is not possible to estimate the two model parameters when we only have data on the total number of customers for each period (i.e., $n_{\cdot 1}$, $n_{\cdot 2}$, ..., $n_{\cdot I}$) or the number of members of each cohort active in the final observed period (i.e., n_{1I}, n_{2I}, ..., n_{II}).)

For Case 1, we can estimate the model parameters using maximum likelihood estimation; for Cases 2–4, we will estimate the model parameters using nonlinear least squares (NLS).

3.1 Case 1

The first case corresponds to the setting where we know the initial size of each cohort, as well as number of cohort members active in period I.

For period i ($i = 1, \ldots I-1$), we know that $n_{ii} - n_{iI}$ customers cancelled their contracts sometime in the first in the first $I - i$ periods, while n_{iI} customers renewed their contracts $I - i$ times. Since the joint probability of this is

$$[1 - S(I-i)]^{n_{ii}-n_{iI}} S(I-i)^{n_{iI}} ,$$

it follows that the sample log-likelihood function is

$$LL(\alpha, \beta \,|\, \text{data}) = \sum_{i=1}^{I-1} \Big\{ (n_{ii} - n_{iI}) \ln \big[1 - S(I-i \,|\, \alpha, \beta)\big] + n_{iI} \ln \big[S(I-i \,|\, \alpha, \beta)\big] \Big\} , \tag{4}$$

the maximum of which can be found using standard numerical optimization methods.

To illustrate how to "code up" (4) in Excel for the case of our hypothetical dataset, consider the worksheet `Limited Information -- Case 1`:

- Given the parameter values located in cells `B9:B10`, we compute $P(T = t)$ for $t = 1, \ldots, 4$ in cells `B14:E14` using the forward recursion given in (1). We then compute $S(T)$ for $t = 1, \ldots, 4$ in cells `B15:E15` using (2).
- We compute the values of $n_{ii} - n_{iI}$ in cells `D18:D21`, and enter the corresponding values of n_{iI} in cells `E18:E21`.
- We enter the values of $S(I - i \,|\, \alpha, \beta)$ in cells `E24:E27`, referring back to the appropriate entries in cells `B15:E15`. We then compute the corresponding values of $1 - S(I - i \,|\, \alpha, \beta)$ in cells `D24:D27`.

- We enter in cells `D29:E32` the product of each element of cells `D18:E21` and the natural log of the corresponding element of cells `D24:E27`. The sum of this block of numbers is located in cell `B11` and is the value of the log-likelihood function for the parameter values located in cells `B9:B10`.

We then use Solver to find the maximum of the log-likelihood function; with starting values of $\alpha = 1, \beta = 1$, we find that this is located at $\hat{\alpha} = 3.79$ and $\hat{\beta} = 15.18$.

3.2 Case 2

The second case corresponds to the setting where we know the initial size of each cohort and the total number of customers for each period.

The estimation strategy is as follows:

1. Given n_{ii}, we compute an estimate of n_{ij} using

$$\widehat{n_{ij}} = n_{ii} S(j - i \,|\, \alpha, \beta) \tag{5}$$

   for $i = 1, \ldots, I-1, j = i+1, \ldots, I$.

2. We then compute an estimate of $n_{\cdot j}$ using

$$\widehat{n_{\cdot j}} = n_{jj} + \sum_{i=1}^{j-1} \widehat{n_{ij}} \tag{6}$$

   for $j = 2, \ldots, I$.

3. Our goal is to find the values of α and β that result in our estimates of the total number of customers for periods $2 \ldots, I$ $(\widehat{n_{\cdot 2}}, \ldots, \widehat{n_{\cdot I}})$ being as close as possible to the corresponding observed numbers $(n_{\cdot 2}, \ldots, n_{\cdot I})$. Formally, we seek to minimize the sum of squared errors

$$\text{SSE}(\alpha, \beta) = \sum_{j=2}^{I} (n_{\cdot j} - \widehat{n_{\cdot j}})^2 \,. \tag{7}$$

To illustrate how we can "code up" (5)–(7) in Excel for the case of our hypothetical dataset, consider the worksheet `Limited Information -- Case 2`:

- Given the parameter values located in cells `B9:B10`, we compute $P(T = t)$ for $t = 1, \ldots, 4$ in cells `B14:E14` using the forward recursion given in (1). We then compute $S(T)$ for $t = 1, \ldots, 4$ in cells `B15:E15` using (2).

- We enter in cells `B17:E20` the values of $S(j-i \,|\, \alpha, \beta)$ for $i = 1, \ldots, 4, j = i+1, \ldots, 5$, referring back to the appropriate entries in cells `B15:E15`.
- We enter the observed values of n_{ii} along the diagonal of cells `A22:E26`. We then compute the $\widehat{n_{ij}}$ above this diagonal using (5), referring back to the appropriate element of cells `B17:E20`.
- We compute the column totals of cells `B22:E26` in cells `B27:E27`, giving us $\widehat{n_{\cdot 2}}, \ldots, \widehat{n_{\cdot 5}}$.
- We then compute the associated squared error numbers, $(n_{\cdot j} - \widehat{n_{\cdot j}})^2$, in cells `B29:E29`. The sum of this block of numbers is located in cell `B11` and is the sum of squared errors for the parameter values located in cells `B9:B10`.

We then use Solver to find the values of α and β that minimize SSE; using starting values of $\alpha = 1, \beta = 1$, we find that they are $\hat{\alpha} = 3.77$ and $\hat{\beta} = 15.09$. (It is important to try out multiple starting values so as to ensure that the minimum of the function has been reached. For example, using starting values of $\alpha = 2, \beta = 2$ (and running Solver twice), we obtain a smaller SSE at $\hat{\alpha} = 3.80$ and $\hat{\beta} = 15.19$.)

3.3 Case 3

The third case corresponds to the setting where we know the total number of customers for each period and the number of members of each cohort active in the final observed period.

The estimation strategy is as follows:

1. Given n_{iI}, we compute an estimate of n_{ii} using

$$\widehat{n_{ii}} = n_{iI} S(I - i \,|\, \alpha, \beta) \tag{8}$$

   for $i = 1, \ldots, I-1$.

2. Given $\widehat{n_{ii}}$, we compute an estimate of n_{ij} using

$$\widehat{n_{ij}} = \widehat{n_{ii}} S(j - i \,|\, \alpha, \beta) \tag{9}$$

   for $i = 1, \ldots, I-2, j = i+1, \ldots, I-1$.

3. We then compute an estimate of $n_{\cdot j}$ using

$$\widehat{n_{\cdot j}} = \sum_{i=1}^{j} \widehat{n_{ij}} \tag{10}$$

   for $j = 1, \ldots, I-1$.

4. Our goal is to find the values of α and β that result in our estimates of the total number of customers for periods $1\ldots, I-1$ ($\widehat{n_{.1}}, \ldots, \widehat{n_{.I-1}}$) being as close as possible to the corresponding observed numbers ($n_{.1}, \ldots, n_{.I-1}$). Formally, we seek to minimize the sum of squared errors

$$\text{SSE}(\alpha, \beta) = \sum_{j=1}^{I-1} (n_{.j} - \widehat{n_{.j}})^2 \,. \tag{11}$$

(Since $n_{.1} = n_{11}$, we could instead perform the above calculations for $i = 2, \ldots, I-1$, use n_{11} to compute $\widehat{n_{1I}}$, and add $(n_{1I} - \widehat{n_{1I}})^2$ to our expression for SSE. We feel that these two approaches are equivalent and therefore use the first approach since it is "cleaner".)

To illustrate how we can "code up" (8)–(11) in Excel for the case of our hypothetical dataset, consider the worksheet `Limited Information -- Case 3`:

- Given the parameter values located in cells `B9:B10`, we compute $P(T = t)$ for $t = 1, \ldots, 4$ in cells `B14:E14` using the forward recursion given in (1). We then compute $S(T)$ for $t = 1, \ldots, 4$ in cells `B15:E15` using (2).
- We enter in cells `B17:E20` the values of $S(j-i \,|\, \alpha, \beta)$ for $i = 1, \ldots, 4, j = i+1, \ldots, 5$, referring back to the appropriate entries in cells `B15:E15`.
- We enter the observed values of n_{iI} in cells `E22:E26`.
- We use (8) to compute the $\widehat{n_{ii}}$ along the diagonal of cells `A22:D25`. Above this diagonal, we compute the $\widehat{n_{ij}}$ using (9), referring back to the appropriate elements of cells `B17:E20`.
- We compute the column totals of cells `A22:D25` in cells `A27:D27`, giving us $\widehat{n_{.1}}, \ldots, \widehat{n_{.4}}$.
- We then compute the associated squared error numbers, $(n_{.j} - \widehat{n_{.j}})^2$, in cells `A29:D29`. The sum of this block of numbers is located in cell `B11` and is the sum of squared errors for the parameter values located in cells `B9:B10`.

We then use Solver to find the values of α and β that minimize SSE; using starting values of $\alpha = 1, \beta = 1$ (and running Solver twice), we find that they are $\hat{\alpha} = 3.80$ and $\hat{\beta} = 15.20$. (It is important to try out multiple starting values so as to ensure that the minimum of the function has been reached.)

3.4 Case 4

The fourth case corresponds to the setting where we know the number of members of each cohort active in the last two observed periods.

The estimation strategy is as follows:

1. Given n_{iI-1}, we compute an estimate of n_{ii} using

$$\widehat{n_{ii}} = n_{iI-1} S(I - i - 1 \,|\, \alpha, \beta) \tag{12}$$

   for $i = 1, \ldots, I - 2$.

2. We then compute an estimate of n_{iI} using

$$\widehat{n_{iI}} = \widehat{n_{ii}} S(I - i \,|\, \alpha, \beta) \tag{13}$$

   for $i = 1, \ldots, I - 2$, and

$$\widehat{n_{iI}} = n_{ii} S(I - i \,|\, \alpha, \beta) \tag{14}$$

   for $i = I - 1$.

3. Our goal is to find the values of α and β that result in our estimates of the number of cohort i customers active in period I $(\widehat{n_{1I}}, \ldots, \widehat{n_{(I-1)I}})$ being as close as possible to the corresponding observed numbers $(n_{1I}, \ldots, n_{(I-1)I})$. Formally, we seek to minimize the sum of squared errors

$$\text{SSE}(\alpha, \beta) = \sum_{i=1}^{I-1} (n_{iI} - \widehat{n_{iI}})^2 \,. \tag{15}$$

To illustrate how we can "code up" (12)–(15) in Excel for the case of our hypothetical dataset, consider the worksheet `Limited Information -- Case 4`:

- Given the parameter values located in cells `B9:B10`, we compute $P(T = t)$ for $t = 1, \ldots, 4$ in cells `B14:E14` using the forward recursion given in (1). We then compute $S(T)$ for $t = 1, \ldots, 4$ in cells `B15:E15` using (2).

- Using (12), we compute the $\widehat{n_{ii}}$ $(i = 1, 2, 3)$ in cells `D18:D20`, referring back to the appropriate entries in cells `D2:D4` and cells `B15:E15`. We then enter the observed value of n_{44} in cell `D21`.

- Using (13) and (14), we then compute $\widehat{n_{15}}, \ldots, \widehat{n_{45}}$ in cells `E18:E21`.

- We then compute the associated squared error numbers, $(n_{iI} - \widehat{n_{iI}})^2$, in cells `F18:F21`. The sum of this block of numbers is located in cell `B11` and is the sum of squared errors for the parameter values located in cells `B9:B10`.

We then use Solver to find the values of α and β that minimize SSE; using starting values of $\alpha = 1, \beta = 1$, we find that they are $\hat{\alpha} = 3.79$ and $\hat{\beta} = 15.17$. (As before, it is important to try out multiple starting values so as to ensure that the minimum of the function has been reached.)

4 Too Little Data?

Let us conclude by considering two data-related issues.

- In order to perform the calculations outlined in Sections 2 and 3 above, it is necessary to have at least three cohorts worth of data ($I = 3$). However, we encourage analysts to make sure that more data are at their disposal.

- It was claimed at the beginning of Section 3 that it is not possible to estimate the two model parameters when we only have data on the total number of customers for each period (i.e., $n_{\cdot 1}$, $n_{\cdot 2}$, ..., $n_{\cdot I}$) or the number of members of each cohort active in the final observed period (i.e., n_{1I}, n_{2I}, ..., n_{II}).

  To get a sense of why is this the case, consider the situation where we only have the total number of customers for each period (i.e., $n_{\cdot 1}$, $n_{\cdot 2}$, ..., $n_{\cdot I}$). Clearly we need to arrive at estimates of $(\widehat{n_{\cdot 2}}, \ldots, \widehat{n_{\cdot I}})$ in order to compute the SSE function. Since $n_{\cdot 1} = n_{11}$, we can compute $\widehat{n_{12}} = n_{11} S(1 \,|\, \alpha, \beta)$. However unless we are willing to assume that $n_{22} = k n_{11}$, where k is predefined by the analyst, we cannot compute the required $\widehat{n_{\cdot 2}} = n_{22} + \widehat{n_{12}}$. (If we compute $\widehat{n_{22}} = n_{\cdot 2} - \widehat{n_{12}}$, we no longer have a value of $\widehat{n_{\cdot 2}}$ that can be compared to $n_{\cdot 2}$ for the calculation of SSE.) And so on for n_{33}, etc. As it is unacceptable to make such assumptions about the unobserved n_{ii}, we can conclude that it is not possible to estimate the two model parameters when we only have data on the total number of customers for each period. (A similar logic applies in the situation where we only have data on the number of members of each cohort active in the final observed period.)

References

Fader, Peter S. and Bruce G. S. Hardie (2007), "How to Project Customer Retention," *Journal of Interactive Marketing*, **21** (Winter), 76–90.