A Note on Implementing the Fader and Hardie "CDNOW Model"

Peter S. Fader and Bruce G. S. Hardie[1]

(August 2001)

1. Introduction

This note describes how to implement Fader and Hardie's (2001) stochastic model of buyer behavior within a standard spreadsheet environment.

There are two key stages associated with the implementation of this model: (i) estimating the model parameters, and (ii) generating the sales forecast given these parameter estimates. The specific steps are outlined in sections 3 and 4 below. Section 2 provides the reader with a simple introduction to the process of estimating the parameters of a basic probability model within a spreadsheet environment. These three sections should be read in conjunction with the Excel spreadsheet `cdnow.xls`.

Let us first present some caveats for the reader who is interested in using this model in any given market setting. One should not blindly cut and paste a new dataset into this spreadsheet. It is important that the assumptions underlying the model are carefully examined, and thought be given as to whether they are appropriate for the dataset at hand. The two key assumptions are:

- The data being modeled are *counts* of relatively homogeneous units (e.g., CDs). This model must not be used to model dollar sales or counts of products that are not very similar (e.g., the number of products purchased at Amazon.com where the units include such disparate items as books, electronic equipment, lawn furniture, and so on).

- Use of the shifted-geometric/geometric distributions implies that the modal trial quantity is 1 unit and that the modal number of units purchased in subsequent weeks, conditional on being a "possible repeat buyer", is 0. If this is not that case, it will be necessary to change the underlying model structure. For example, the shifted beta-geometric model of trial counts could be replaced by the truncated or shifted NBD

(which can have a mode away from 1). Similarly, the beta-geometric repeat purchasing distribution could be replaced by the NBD (which can have a non-zero mode).

A number of other assumptions made, implicitly or explicitly, in the paper should also be acknowledged and taken into account (e.g., the independence of quantity decisions across transactions).

We strongly encourage interested readers to build the spreadsheet that implements the model for themselves "from scratch", using this note and the Excel spreadsheet `cdnow.xls` as a guide.

2. The Week 1 Trial Model

As a refresher (or primer) on estimating the parameters of a basic probability model using Excel, let us consider fitting the trial submodel to the week 1 data. As noted in Appendix A of the paper, the column of data in Table 1 corresponding to week 1 presents trial-week-only purchases by a group of 1574 customers. Our goal is to fit the shifted beta-geometric model, as given in equation (2) of the paper (p. S98), to these data.

The shifted beta-geometric model has two parameters, α_T and β_T. Maximum likelihood estimates of these two model parameters are found by maximizing the following log-likelihood function:

$$LL = \sum_{x=1}^{9} n_{1x} \ln\big[P(T_1 = x)\big] + \left(1574 - \sum_{x=1}^{9} n_{1x}\right) \ln\left[1 - \sum_{x=1}^{9} P(T_1 = x)\right]$$

where n_{1x} is the number of people making x purchases in week 1. We construct this log-likelihood function in an Excel worksheet in the following manner.

Consider the worksheet `Week 1 Trial (A)`. Cells `A5:B14` contain the relevant purchasing data from Table 1. The first thing we need to do is create expressions for the shifted beta-geometric probabilities of making x purchases ($x = 1, 2, \ldots, 9, 10+$), given α_T and β_T. These shifted beta-geometric probabilities can be computed by recursion using the expressions given in equation (7) (p. S100). In order to create the corresponding formulas in the spreadsheet without an error message appearing (e.g., `#NUM!` or `#DIV/0!`), we need some so-called starting values for α_T and β_T. Provided they are within the defined bounds ($\alpha_T, \beta_T > 0$), the exact values do not matter. We start with

1.0 for both parameters and locate these values in cells `B1:B2`. The formulas in cells `C5:C13` are a straightforward implementation of the expressions given in equation (7). The probability of making 10+ purchases in a trial week (cell `C14`) is simply $1 - \sum_{x=1}^{9} P(T_1 = x)$.

Now that we have the shifted beta-geometric probabilities, creating the log-likelihood function is simple. The individual elements of the above log-likelihood function are contained in cells `D5:D14`. The total is found in cell `D2`; this is the value of the log-likelihood function, given the values for the two model parameters in cells `B1:B2`.

Given these sample data, we find the maximum likelihood estimates of the shifted beta-geometric distribution by maximizing the log-likelihood function. We do this using the Excel add-in Solver. (Background information on Solver can be found in Lilien and Rangaswamy (1998) or Winston and Albright (1997).) The *target cell* is the value of the log-likelihood function (cell `D2`); we wish to *maximize* this by *changing* cells `B1:B2`. The *constraints* we place on the parameters are that both α_T and β_T are greater than 0. As Solver only offers us a "greater than or equal to" constraint, we *add* the constraint that cells `B1:B2` are $\geq$ a small positive number (e.g., 0.00001). Clicking the *Solve* button, Solver finds the values of α_T and β_T that maximize the log-likelihood function; these are the maximum likelihood estimates of the model parameters.

The results of this optimization process are found in the worksheet `Week 1 Trial (B)`. In this worksheet, we also evaluate the fit of the model using the standard chi-squared goodness of fit test. We first have to compute the expected number of people buying $1, 2, \ldots, 9, 10+$ units in their trial week. We have $E(n_{1x}) = 1574 \times P(T_1 = x)$; these calculations are implemented in cells `F5:F14`. The chi-squared goodness of fit test statistic is computed as

$$\chi^2 = \sum_{x=1}^{10+} \frac{\left[n_{1x} - E(n_{1x})\right]^2}{E(n_{1x})}$$

Each element of this calculation is presented in cells `G5:G14`, with the total given in cell `G15`. The critical value can be computed using the `chiinv` command. As the value of the sample test statistic is less than the critical value (cell `G17`), we conclude that the shifted beta-geometric distribution adequately fits the data.

3. Calibrating the Full Model

We now turn our attention to the task of estimating the parameters of the full model presented in the body of the paper. Our goal is to construct the log-likelihood function — as given in equation (6) (p. S100) — in an Excel worksheet. At the heart of this log-likelihood function is $P(X_w = x)$, the probability that an eligible customer purchases x units in week w, as given in equation (1) (p. S98). As the sample data are in the form of a table documenting the number of people purchasing 0, 1, ... , 9, 10+ units (CDs) for each of the 12 weeks, we need to create a table that gives us $P(X_w = x)$, $x = 0, 1, \ldots, 9, 10+$ and $w = 1, 2, \ldots 12$, given values of the six model parameters $(\alpha_T, \beta_T, \alpha_R, \beta_R, \gamma, \delta)$.

From equation (1), we see that $P(X_w = x)$ is simply a weighted average of the week-of-trial-specific probabilities of purchasing x units in week w. As an intermediate step, we build twelve tables that give us the probability of purchasing x units in week w, one for each trial week. These will then be aggregated and the log-likelihood created. The exact steps are as follows — see the worksheet `Full Model (A)`.

Let us start with the week 1 triers. Our goal is to build a table that gives us the probability that any such person purchases x units in their trial week and in any of the subsequent eleven "repeat" weeks. We will create this table of probabilities in cells `C63:N73`. In order to create the corresponding formulas in the spreadsheet without any error messages appearing, we need some so-called starting values for the six model parameters. The exact values do not matter — provided they are within the defined bounds — so we start with 1.0 for α_T, β_T, α_R and β_R, and 0.2 and 0.1 for γ and δ respectively. We locate these parameter values in cells `B1:B6`.

We compute the shifted beta-geometric probabilities of making x purchases in the trial week, given α_T and β_T, by recursion using the expressions given in equation (7); the formulas found in cells `C64:C73` mirror those found in cells `C5:C14` of the worksheet `Week 1 Trial (A)`.

Next we need to create the expressions for the time-dependent, zero-inflated beta-geometric distribution probabilities that a week 1 trier makes a repeat purchase of x units in week w, i.e., $P(R_{w|1} = x)$, $x = 0, 1, \ldots, 9, 10+$, $w = 2, 3, \ldots 12$. Rather than directly use the recursive relationship given in equation (8) (p. S100), we take the following approach. We first compute the probability that a week 1 trier is a "possible repeat buyer" in weeks 2–12; following repeat model assumption (1) — see p. S99 — this is computed as

$\gamma(w-1)^{\delta}$. Cells `D49:N49` contain this formula, conditional on the parameter values in cells `B5:B6`.

Next we need to create the expressions for the beta-geometric probabilities of making x purchases ($x = 0, 1, \ldots, 9, 10+$), given α_R and β_R, for someone who is a "possible repeat buyer". The beta-geometric probabilities can be computed using the following recursive relationship:

$$P(R = x \,|\, \text{possible repeat buyer}) = \begin{cases} \dfrac{\alpha_R}{\alpha_R + \beta_R} & x = 0 \\ \dfrac{\beta_R + x - 1}{\alpha_R + \beta_R + x} P(R = x - 1) & x \geq 1 \end{cases}$$

The formulas in cells `D51:D60` are a straight-forward implementation of this expression. The probability of a "possible repeat buyer" making 10+ purchases (cell `D61`) is simply $1 - \sum_{x=0}^{9} P(R = x)$.

It follows that the probability of a week 1 trier making a repeat purchase of x units in week w is given by $P(R_{w|1}) = P(R = x \,|\, \text{possible repeat buyer}) \times P(\text{week 1 trier is a possible repeat buyer in week } w)$. Cells `D63:N73` contain this calculation for $x = 0, 1, \ldots, 9, 10+$ and $w = 2, 3, \ldots 12$.

The corresponding purchase probabilities for a week 2 trier are simply the week 1 trier numbers lagged by one week; these are given in cells `D75:N85`. The purchase probabilities for a week 3 trier are simply the week 2 trier numbers lagged by one week (cells `E87:N97`), and so on.

All that we need to do in order to create the table of $P(X_w = x)$ is to take a weighted average of these twelve sets of probability tables. This calculation is performed in cells `C37:N47`; the corresponding formulas are simply the implementation of equation (1).

Now that we have this table of probabilities, creating the log-likelihood function is straightforward. The individual elements of the log-likelihood function, equation (6), are contained in cells `C25:N35`. The total is given in cell `E1`; this is the value of the log-likelihood function, given the values for the six model parameters in cells `B1:B6`.

Given the sample data (i.e., cells `C10:N20`), we find the maximum likelihood estimates of the model parameters by maximizing the log-likelihood function. We do this using Solver. The *target cell* is the value of the log-likelihood (cell `E1`); we wish to *maximize* this by *changing* cells `B1:B6`. The *constraints* we place on the parameters are that α_T, β_T, α_R, β_R and γ are

greater than 0. As Solver only offers us a "greater than or equal to" constraint, we *add* the constraint that cells `B1:B5` are $\geq$ a small positive number (e.g., 0.00001). Clicking the *Solve* button, Solver finds the values of the six model parameters that maximize the log-likelihood function. But can we be sure that we have 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 been reached, we can say that the numbers given in cells `B1:B6` are the maximum likelihood estimates of the model parameters.

So as to be sure that these are indeed the maximum likelihood estimates of the model parameters, it is good practice to redo the optimization process using a completely different set of starting values. For example, using starting values of $\{0.01, 0.01, 0.01, 0.01, 0.01, 0\}$ for cells `B1:B6`, repeatedly use Solver until you are satisfied that the maximum of the log-likelihood function has been reached. Are the corresponding values of the six model parameters equal to those given in the paper? (They should be!)

The results of this optimization process are found in the worksheet `Full Model (B)`. In this worksheet, we also evaluate the fit of the model using the standard chi-squared goodness of fit test. We first have to compute the expected number of people buying $0, 1, \ldots, 9, 10+$ units in each week ($w = 1, 2, \ldots, 12$). We have $E(n_{wx}) = P(X_w = x)\times$ the number of eligible cohort members in week w (i.e., $\sum_{i=1}^{w} n_i$). These calculations are implemented in cells `Q10:AB20`. The chi-squared goodness of fit test statistic is computed as

$$\chi^2 = \sum_{x=1}^{10+} \frac{\left[n_{1x} - E(n_{1x})\right]^2}{E(n_{1x})} + \sum_{w=2}^{12} \sum_{x=0}^{10+} \frac{\left[n_{wx} - E(n_{wx})\right]^2}{E(n_{wx})}$$

Each element of this calculation is presented in cells `Q22:AB32`, with the total given in cell `AB34`. The critical value can be computed using the `chiinv` command. As the value of the sample test statistic is less than the critical value (cell `AB35`), we conclude that the model adequately fits the data.

4. Creating the Sales Forecast

Now that we have estimates of the six model parameters, creating the sales forecast is a simple exercise. The expression for the expected total number of units sold in any given week is given in equation (9) (p. S101). At the heart

this are expressions for the expected number of units purchased in the trial week and the expected number of units purchased $w-i$ weeks after trial. This is implemented in the worksheet `Sales Forecast` in the following manner.

We first consider the case of the week 1 triers. Cells `C11:C62` contain the expected number of units purchased in weeks 1–52 by any given week 1 trier. Cell `C11` is the expected number of units purchased in the trial week, using the formula given in equation (3) (p. S98). Expected repeat sales in subsequent weeks are given in the remaining cells, which contain the formula for $E(R_{w|i})$ $(w = 2, 3, \ldots, 52; i = 1)$ as given in equation (5) (p. S100).

The corresponding numbers for a week 2 trier are simply the week 1 trier numbers lagged by one week; they are given in cells `D12:D62`. The expected unit sales for a week 3 trier are simply the week 2 trier numbers lagged by one week, and so on. Given these sets of mean weekly unit purchases for any week i trier $(i = 1, 2, \ldots 12)$, the expected total number of units sold in weeks 1–52, as computed using equation (9), are given in cells `P11:P62`. And that's it!

References

Fader, Peter S. and Bruce G. S. Hardie, (2001), "Forecasting Repeat Sales at CDNOW: A Case Study," *Interfaces*, **31** (May–June), Part 2 of 2, S94–S107.

Lilien, Gary L. and Arvind Rangaswamy (1998), *Marketing Engineering: Computer-Assisted Marketing Analysis and Planning*, Reading, MA: Addison-Wesley.

Winston, Wayne L. and S. Christian Albright (1997), *Practical Management Science: Spreadsheet Modeling and Applications*, Belmont, CA: Duxbury Press.