Repeated Binary Logit: Analysing Variation in Behavioural Loyalty

Cam Rungie & Gilles Laurent

Abstract
Brand and store managers are anxious to obtain high “loyalty,” as operationalized by repeat purchase behaviour. In this paper, we introduce the Repeated Binary Logit (RBL) model, which analyses directly the impact of covariates on repeat purchase, and which can be described as an extension of traditional logistic regression. We present empirical applications of RBL, and we discuss its relationships to several classical models.

Keywords: Brand choice, Buyer behaviour, Choice models, Data mining, Marketing research, Segmentation

1. Introduction
Brand and store managers are anxious to obtain high “loyalty” from their consumers. And, very often, they equate “loyalty” and repeat purchase behaviour. Assessing the level of repeat purchase, in a category at large or for a specific brand, is a key step in diagnosing a market, and in defining a marketing strategy. We argue that market analysts must go beyond the sheer measurement of “loyalty” through repeat purchase. Rather, they should identify the causes of loyalty, and measure their impact, through a statistical analysis. In this paper, we introduce the Repeated Binary Logit (RBL) model that analyses directly the impact of covariates on loyalty. RBL can be described as an extension of traditional logistic regression. We present several empirical applications of RBL. Finally, we discuss its relationships to several classical models.

2. The Repeated Binary Logit (RBL) Model
2.1 Functional Form
We first describe the simple case where there are no covariates impacting loyalty. RBL is constructed from the beta binomial distribution. Let a randomly selected decision maker make \( k \) selections from a binary choice set containing alternatives \( A \) and \( B \) (\( A \) could be a specific brand \( A \), and \( B \) all the other brands; or \( A \) could be a specific store \( A \), and \( B \) all the other stores). Let the number of times alternatives \( A \) and \( B \) are selected be \( r_A \) and \( r_B \) where \( r_A + r_B = k \). Over the population of decision makers, these selection rates are random variables, \( R_A \) and \( R_B \). They have a beta binomial distribution (BBD) conditional on \( k \) and on the parameters \( \alpha_A \) and \( \alpha_B \) of the beta distribution. The probability density function for the BBD is:

\[
f_{\alpha_A, \alpha_B}(r_A, r_B \mid r_A + r_B = k) = \frac{\Gamma(\alpha_A + \alpha_B)k!}{\Gamma(\alpha_A + \alpha_B + k)!} \frac{\Gamma(\alpha_A + k)}{\Gamma(\alpha_A)} \frac{\Gamma(\alpha_B + k)}{\Gamma(\alpha_B)} \frac{r_A!}{r_A!} \frac{r_B!}{r_B!} \frac{1}{\Gamma(\alpha_A + \alpha_B + k)}
\]
2.2 Link to Marketing Statistics

For the BBD the mean, $\pi$, for the proportion of times alternative A is selected is (Johnson, Kotz, and Balakrishnan 1994; Johnson, Kotz, and Balakrishnan 1997; Kotz, Balakrishnan, and Johnson 2000):

Equation 2

$$\pi = E\left[ {\frac{R_A}{k}} \right] = \frac{\alpha_A}{\alpha_A + \alpha_B}$$

where $0 < \pi < 1$.

In many marketing applications this is equivalent to, or a close estimate of, the market share for alternative A.

A simple indicator of loyalty is the repeat rate, $\rho$. It is defined to be the probability of selecting alternative A at the next choice occasion, given it was selected at the last; i.e. $Pr(R_A=2|k=2)/Pr(R_A=1|k=1)$. (A choice occasion is an event where a selection from the choice set is made, e.g. a purchase in the category for a brand choice, or a shopping trip for a store choice.) Each decision maker makes two independent choices from the same choice set with replacement and under stationary conditions. From Equation 1:

Equation 3

$$\rho = \frac{1 + \alpha_A}{1 + \alpha_A + \alpha_B}$$

where $0 < \rho < 1$.

2.3 Introducing Covariates

Covariates are introduced by building a generalised linear model based on the BBD and the exponential linear link functions (McCullagh and Nelder 1989). These link functions are:

Equation 4

$$\alpha_A = \exp(\theta'_A x)$$

and

$$\alpha_B = \exp(\theta'_B x)$$

The vector of covariates is $x$ and there are two vectors of ‘regression’ coefficients, $\theta_A$ and $\theta_B$. The repeated binary logit (RBL) model is specified by inserting the link functions from Equation 4 into Equation 1. The resultant likelihood function is given below.

The RBL model has the following properties:

Equation 5

$$\pi(x) = \frac{\exp(\theta'_A x)}{\exp(\theta'_A x) + \exp(\theta'_B x)}$$

Equation 6

$$\rho(x) = \frac{1 + \exp(\theta'_A x)}{1 + \exp(\theta'_A x) + \exp(\theta'_B x)}$$

Thus, RBL models the market share and repeat rate simultaneously.

Finally it is interesting to consider the special case where $k=1$, i.e. when each consumer makes a single choice. Obviously, in such a case, the study of loyalty loses its interest. However, we can observe that, in this condition, RBL reduces to:

Equation 7

$$Pr(R=1|x) = \frac{\exp(\theta'_A x)}{\exp(\theta'_A x) + \exp(\theta'_B x)}$$

which is the basis of MNL and logistic regression, a tool widely used in marketing for analysing binary choices (Ben-Akiva and Lerman 1985; Guadagni and Little 1983; Louviere, Hensher, and Swait 1999; McFadden 1974; McFadden 1984). Thus, from Equation 5 and Equation 7, RBL is an extension of logistic regression.

2.4 Estimation – The Likelihood Function

The likelihood function can be specified as follows. Let there be a sample of $n$ decision makers. For decision maker $i$, where $i=1,2,\ldots,n$:

Equation 8

$$\hat{\alpha}_{A,i} = \exp(\hat{\theta}'_{A,i} x_i) \quad \text{and} \quad \hat{\alpha}_{B,i} = \exp(\hat{\theta}'_{B,i} x_i)$$

From Equation 1 and Equation 8 and using the property of the gamma function where $k!=\Gamma(k+1)$ then the likelihood function is (Edwards 1976; Eliason 1993):

Equation 9

$$L = \prod_{i=1}^{n} \left( \frac{\Gamma(\hat{\alpha}_{A,i} + \hat{\alpha}_{B,i}) \Gamma(\hat{\alpha}_{A,i} + r_{A,i}) \Gamma(\hat{\alpha}_{B,i} + r_{B,i})}{\Gamma(\hat{\alpha}_{A,i} + \hat{\alpha}_{B,i} + k) \Gamma(r_{A,i} + 1) \Gamma(r_{B,i} + 1) \Gamma(\hat{\alpha}_{A,i})} \right)$$

Regarding implementation, the vectors of parameters $\theta_A$ and $\theta_B$ are identified provided that $k>1$ (and that, as in general linear regression models, the vector of covariates $x$ does not include excessive colinearity). Log gamma is widely available in computer packages. It is computationally more feasible to directly calculate the
log of Equation 9 rather than to calculate $L$ and then take the log. In optimizing $\log(L)$ local maxima can be encountered so it is wise to consider multiple start points. The statistical significance of each of the covariates can be established using the likelihood ratio test. A macro, written in MATLAB, for likelihood estimation and likelihood ratio testing is available for beta evaluation from the first author.

3. First Example: Store Loyalty and Brand Loyalty

We present an empirical example of RBL analysis, in which a covariate, while it has no impact on market share, has a significant impact on loyalty. Marketing Scan provided purchase data for one year in Angers, France. This is a “closed zone” panel, which follows all the supermarkets in the city and its suburbs. This means that the panel records all purchases by the panelists from supermarkets in the city, whatever the store in which the purchase takes place. For each purchase occasion by a panelist, we therefore record both the brand choice, and the supermarket where the purchase took place. There are 3,459 households followed over one year, who made a total of 21,982 purchases in the category while 313 households made no purchases. We analyse the binary choice between the leading brand, with 15.7% market share, and all other brands. The covariate we consider in this example is the proportion of each household’s purchases of all supermarket items, over a year, which is in their favourite supermarket, i.e. the individual share of category requirements for the favourite supermarket. This is a measure of their global loyalty to their favourite supermarket.

The dependent variable, the market share for the leading brand on the market, does not change as the covariate, the loyalty to each consumer’s favourite supermarket, changes. A traditional logistic regression in the 21,982 purchases shows that the coefficient of the covariate is neither statistically significant ($LR=1.950$, $p=0.163$ with 1 df), nor managerially relevant. Given the large sample size, this is strong evidence that the leading brand’s market share does not vary according to shoppers’ loyalty to their favourite supermarket.

However, the application of the RBL model leads to a more complex diagnostic: a significant value of the parameters for the covariate leads to the prediction of no significant impact on the market share, but a significant impact on brand loyalty. The estimation of the model according to the procedure described earlier indicates that the covariate, the loyalty to favourite supermarket, is significant ($LR=19.1176$, $p=.00007$ with 2 df). For each distribution parameter $\theta_A$ or $\theta_B$ there are two estimated coefficients, a constant, and a coefficient for the covariate. The empirical estimates are: $\theta_A=[-1.0139,-0.9702]$ and $\theta_B=[0.6481,-1.0705]$.

These estimates predict no change in the market share of the leading brand when the covariate, the loyalty to the favourite supermarket, increases. This is illustrated in Figure 1, where the plot of the estimated change is practically a horizontal line (we also plot in Figure 1, for comparison purposes, the observed raw values for sub-samples based on the value of the covariate).

However, there is a considerable change in the repeat rate, $\rho$, as shown in Figure 2 (which plots the repeat rate predicted by the model, as well as the observed raw values for the sub-samples).

This example demonstrates the potential of RBL to produce differentiated diagnoses. The covariate of interest, loyalty to supermarket, has no impact on the market share of the leading brand, but it does relate to the loyalty for that leading brand. The market share of the leading brand does not increase with consumers’ loyalty to their preferred supermarket. However, consumers who are more loyal to their preferred supermarket are also more loyal to the leading brand.

This outcome results directly from the numerical values of the estimates $\theta_A=[-1.0139,-0.9702]$ and $\theta_B=[0.6481,-1.0705]$. The coefficients of the covariate are similar ($-0.9702$ and $-1.0705$). As a consequence, the predicted market share, given by $\alpha_A/(\alpha_A+\alpha_B)$, does not change much when the covariate increases. However, the predicted repeat rate, given by $(1+\alpha_A)/(1+\alpha_A+\alpha_B)$, increases markedly.

These results are very important from a managerial point of view. From a diagnosis approach, one should not interpret in the same manner the same observed repeat rate for the leading brand, depending on whether it is observed for a consumer who is highly loyal to one supermarket, or for a consumer who allocates her purchases across many supermarkets. From a decision-making point of view, the promotional program directed towards consumers of the leading brand should be very different, depending on whether or not they are loyal to their favourite supermarket.
4. Second Example: Diversified Factors of Brand Loyalty

We now consider all top six brands in the same category. They ranged from 15.7% down to 5.6% market share. For each of the brands, an RBL analysis was conducted, with five covariates (Table 1):

- Number of persons in the household.
- Average spend, by the household, in supermarkets, per week.
- Average number of visits per week to supermarkets for members of the household.
- Loyalty to favourite supermarket (SCR of favourite supermarket).
- Total purchase rate for the product category

After exploratory analysis the models were found to fit best by taking the natural logs of all covariates except loyalty to the favourite supermarket.

We use statistical tests to measure the impact of covariates. In each case, we compare the fit of the unconstrained model (including all five covariates) to that of a constrained model, in which one of the covariates has been removed. This is done separately for each brand and each covariate. The resulting significance tests are in Table 2.

A key result is that the significant covariates are not the same for all brands. For the leading brand, only one covariate is significant, loyalty to favourite supermarket. This is the result discussed in the first example. The covariate has a major impact on loyalty but no significant impact on market share.
For the third and fourth brands no covariate is significant. For these brands there is no difference in the purchase behaviour between shoppers from different sized households, with different household spends, etc. As a general result this is not overly surprising. A brand may well have category wide and market wide appeal. Generally, when searching for demographic differences between brands, it is the exception that a covariate will be significant and relevant, rather than the rule. Consequently, the interesting results are where there is a significant effect, i.e. brands 2, 5 and 6.

The impact of the covariates on the market share and repeat rate is given in Table 3 and Table 4. For each covariate the impact is calculated for a change from one standard deviation below its mean to one standard deviation above its mean. The results are only shown for the significant covariates.

The changes shown in Table 3 may look small, but they are very relevant. As household expenditure changes from the mean minus one standard deviation to the mean plus one standard deviation, the market share for brand 6 reduces by 2.3% share points. Since this brand has overall 5.6% market share, the change is from 6.7% to 4.5%, a very important change.

Table 3 displays results in terms of market shares. For example, it shows that bigger households which spend less are more likely to purchase brands 5 and 6. In a classic marketing approach based on market shares, the target for these brands might be defined as such households. However, Table 4 provides a different analysis, focused on changes in repeat rate.

In Table 4 the absolute level of change in the repeat rates is far greater than the absolute level of change in market shares (shown in Table 3). This reflects the fact that the typical levels are higher for repeat rates than for market shares. Given that repeat rates are higher there is greater potential for them to vary. Keeping this in mind, several of the results in Table 4 are still of particular relevance to

### Table 1: Descriptive Statistics for the Covariates

<table>
<thead>
<tr>
<th></th>
<th>Household Size</th>
<th>Household Spend</th>
<th>Visits to Supermarkets</th>
<th>Loyalty to Supermarket</th>
<th>Category Purchases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum</td>
<td>1</td>
<td>100</td>
<td>0.125</td>
<td>0.3</td>
<td>0</td>
</tr>
<tr>
<td>Maximum</td>
<td>5</td>
<td>900</td>
<td>4.000</td>
<td>1.0</td>
<td>90</td>
</tr>
<tr>
<td>Mean Standard Deviation</td>
<td>2.59</td>
<td>368</td>
<td>1.36</td>
<td>0.75</td>
<td>6.4</td>
</tr>
<tr>
<td>Deviation</td>
<td>1.28</td>
<td>213</td>
<td>0.83</td>
<td>0.21</td>
<td>7.0</td>
</tr>
</tbody>
</table>

### Table 2: Each Covariate is Significant for At Least One of the Top Six Brands

(Chi Square LR test, each with 2 df)

<table>
<thead>
<tr>
<th>Brand</th>
<th>Household Size</th>
<th>Household Spend</th>
<th>Visits to Supermarkets</th>
<th>Loyalty to Supermarket</th>
<th>Category Purchases</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.0042</td>
<td>2.7234</td>
<td>0.5308</td>
<td>19.1548***</td>
<td>0.5732</td>
</tr>
<tr>
<td>2</td>
<td>0.5228</td>
<td>9.5302**</td>
<td>13.2546**</td>
<td>2.2328</td>
<td>16.063***</td>
</tr>
<tr>
<td>3</td>
<td>3.8548</td>
<td>2.7666</td>
<td>1.1964</td>
<td>4.2566</td>
<td>2.8016</td>
</tr>
<tr>
<td>4</td>
<td>1.6546</td>
<td>0.003</td>
<td>3.9174</td>
<td>4.487</td>
<td>0.584</td>
</tr>
<tr>
<td>5</td>
<td>19.8304***</td>
<td>19.554***</td>
<td>1.6786</td>
<td>12.1882**</td>
<td>3.3948</td>
</tr>
<tr>
<td>6</td>
<td>16.243***</td>
<td>11.2026**</td>
<td>14.5012***</td>
<td>7.3396*</td>
<td>1.1688</td>
</tr>
</tbody>
</table>
For brand 1, we see again the earlier result, where households with more loyalty to their favourite supermarket are also more loyal to the leading brand. This also shows up for brands 5 and 6.

Smaller households show greater loyalty with regards to brand 6. Smaller spending households show more loyalty with regards to brand 5. Households shopping less-often show more loyalty with regards to brands 2 and 6.

How are these results to be interpreted? Brand 6 has a loyalty segment that is smaller households, which shop less often and are more loyal to their favourite supermarket. From Table 3 we see that in this segment there is not a strongly relevant change in market shares. This segment is not likely to purchase more, but it is more loyal in their choice to buy or not buy Brand 6. In this segment Brand 6 has less elasticity. It will have less sensitivity to deals. It will switch less. Once buyers are won in this market, they are more likely to stay. This will have an impact on the selection of the marketing mix for this brand. Amongst these shoppers it will focus more on attracting new buyers. The brand manager may well be able to devise ways of doing this. What can be done to attract new buyers amongst smaller households who shop less often and are more loyal to their favourite supermarket? The market analysis tool has helped

### Table 3: Changes in Market Share

(For example, as household spend increases from the mean minus one standard deviation to the mean plus one standard deviation, the market share for brand 2 increases by 3.8% share points)

<table>
<thead>
<tr>
<th>Brand</th>
<th>Market Share</th>
<th>Household Size</th>
<th>Household Spend</th>
<th>Visits to Supermarkets</th>
<th>Loyalty to Supermarket</th>
<th>Category Purchases</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15.7%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.5%</td>
</tr>
<tr>
<td>2</td>
<td>9.8%</td>
<td>3.8%</td>
<td>-2.3%</td>
<td></td>
<td></td>
<td>0.6%</td>
</tr>
<tr>
<td>3</td>
<td>9.5%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>8.9%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>8.6%</td>
<td>3.3%</td>
<td>-2.6%</td>
<td></td>
<td>-0.1%</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>5.6%</td>
<td>1.2%</td>
<td>-2.3%</td>
<td>-0.2%</td>
<td>-0.5%</td>
<td></td>
</tr>
</tbody>
</table>

### Table 4: Changes in Repeat Rate

(For example, as household spend increases from the mean minus one standard deviation to the mean plus one standard deviation, the repeat rate for brand 2 increases by 8.4% points; i.e. from 49% to 57%)

<table>
<thead>
<tr>
<th>Brand</th>
<th>Repeat Rate</th>
<th>Household Size</th>
<th>Household Spend</th>
<th>Visits to Supermarkets</th>
<th>Loyalty to Supermarket</th>
<th>Category Purchases</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>58%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>9.3%</td>
</tr>
<tr>
<td>2</td>
<td>53%</td>
<td>8.4%</td>
<td>-15.5%</td>
<td></td>
<td></td>
<td>15.4%</td>
</tr>
<tr>
<td>3</td>
<td>56%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>45%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>42%</td>
<td>3.5%</td>
<td>-8.4%</td>
<td></td>
<td></td>
<td>9.3%</td>
</tr>
<tr>
<td>6</td>
<td>40%</td>
<td>-9.3%</td>
<td>0.7%</td>
<td>-16.6%</td>
<td></td>
<td>6.7%</td>
</tr>
</tbody>
</table>
identify the marketing objectives. The task is now for the brand manager to find the appropriate marketing tactics.

The analysis shows that targeting on the basis of loyalty is a possible alternative to the traditional targeting on the basis of market share.

- One can identify high loyalty segments where there is low elasticity and larger returns from attracting new buyers.
- Conversely one can identify low loyalty segments where there is considerable switching. As buyers are often switching, small initiatives can generate relevant results but these gains can also be quickly lost. Strategies should focus on regularly implementing inexpensive activities.

The market analysis tools presented here have the potential to open up a whole new array of marketing strategies based on loyalty segmentation.

5. Data Specification

The data required for RBL is repeated choices from the same binary choice set. The data must record the repeated choices for each decision maker; at least two choices but more is better. All decision makers must be presented with the same dichotomous choice set. This is the form of panel data that is often encountered in purchase records, sales databases, scanner data and media surveys. There is a substantial supply of data to which RBL could be usefully applied.

The covariates must be constant over the period during which the choice are made. This is typical of panel surveys where there is one baseline survey and then repeated follow-up surveys. The baseline collects the data that doesn’t change or which the researcher accepts can only be measured once. This is the data for the covariates. The follow-up surveys record purchases, viewing etc. This is the repeated choice data.

RBL is for stationary markets. Over the term of the data collection there can and almost always will be random fluctuations in discrete choice and in the influences on this discrete choice. A stationary market can still fluctuate. It is the nature of these fluctuations which RBL analyses. Many established markets are sufficiently stationary. However, the model is not intended for markets where there are substantial trends such as major evolving brands. Where trends exist, the smaller the term of the data the less the trend, so the problem may be solved, for example, by analysing quarterly data rather than annual. But generally RBL is a model for more established or mature markets.

Two examples have been presented above. The first, with only one covariate, demonstrates how a characteristic of shoppers does not relate to their likelihood to purchase the brand but it still relates to their purchase behaviour. A simple descriptive analysis comparing the profile of the shoppers of the brands would miss this point. Nevertheless a descriptive analysis that included statistics, such as penetrations or purchase frequencies, could identify the important variations in behaviour. The RBL model is not necessary when there is only one covariate. Descriptive analyses will suffice. However, in revealed preference data there are often multiple covariates that are also often correlated. In the second example, not surprisingly, there is a correlation between household size and category purchase rate ($r = 0.5$). As the model analyses both covariates simultaneously the relative importance of each is determined. This type of pattern in the data, particularly when there are many covariates, cannot be easily investigated with descriptive analyses. This is one of the great strengths of the RBL model. It can analyse multiple covariates.

6. Comparison of RBL with other Discrete Choice Models

6.1 Discrete Choice Models

The RBL model is an extension of the beta distribution. However, it has structural links to some of the major discrete choice models: the NBD/BBD, the Dirichlet model, logistic regression and multinomial logit.

NBD/BBD. The NBD/BBD model is a simple extension, albeit without covariates, of the model described above. The BBD is applied to the binary choice, the selection between purchasing the specified alternative and all others, conditional on the total purchase rate for all alternatives (often, all brands in a category) over some time period. The negative binomial distribution (NBD) is used to model this total number of purchases made by each of the shoppers in the population.

The NBD/BBD model serves to demonstrate that there are two forms of heterogeneity: (1) the variance in the NBD which reflects how much difference there is between heavy and light buyers of the category and (2) the variance in the BBD which reflects behavioural loyalty.
Dirichlet. The Dirichlet Model is an extension of the NBD/BBD to cover more than binary choice. The BBD is replaced by its multivariate alternative, the Dirichlet multinomial distribution (DMD). Thus the Dirichlet is an NBD/DMD model. The NBD depicts the distribution for the category purchase rate. The DMD depicts the distribution for the repeated choices of the brands in the category conditional on the category purchase rate.

It may sound attractive to generalize the Dirichlet Model to incorporate covariates. Both the NBD (Cameron and Trivedi 1998; Winkelmann 1997) and the DMD can be expanded in a similar manner as has been done in this paper with the BBD. Thus, covariates can be introduced to each of the two parts of the NBD/DMD. The outcome is a model which fits covariates simultaneously to all the brand performance measures for all the brands in the product category. Conceptually this is simple, but the devil is in the computing code. Besides, while the approach appears to be potentially useful, there is a self-defeating dimension to adding covariates to the Dirichlet Model without a critical perspective. This approach requires fitting covariates to the NBD and DMD separately but then combining the results. It is better to keep the analyses separate: fit covariates to the NBD to investigate changes in the category purchase rates; fit covariates to the DMD to investigate changes in the brand preferences. This allows analysis of the covariates directly and separately on each of the heterogeneities; (1) variance in NBD and (2) variance in DMD.

Logistic Regression. RBL can be seen as an extension of logistic regression. In logistic regression, only one binary choice is observed for each respondent, vs. a series of choices in RBL. If data are available on a single choice per respondent, the likelihood functions for logistic regression and RBL are identical. This is not just coincidence. The theoretical foundations of logistic regression are in loglinear modelling and random utility theory whereas the foundation of RBL is in the beta binomial distribution (BBD). These bodies of theory are related, and from the theoretical and practical perspective, it is both encouraging and not surprising that they generate the same functional form and likelihood function.

Multinomial Logit. Random utility theory, as used in multinomial logit (MNL), is an excellent model for fitting covariates to the conditional choice between alternatives such as the brands in a product category. It models the mean choice probability. It has an excellent ability to model varying choice sets and the attributes of each alternative. Obviously, RBL cannot do this as it analyses binary choice only. However, RBL incorporates repeated purchases, and it can therefore analyse variations in heterogeneity and loyalty directly. Expansion of the RBL model to incorporate more than two alternatives, the attributes of the alternatives and varying choice sets is conceptually possible, but seems exceedingly complex.

6.2 Identification of Stochastic Processes

The comparison of the RBL and MNL models can be further developed by examining their stochastic forms. While the models have strong similarities in their likelihood functions they differ in their conceptualisation of the random influences on discrete choice.

RBL identifies that each shopper has a fixed probability of selecting A or B over the choice occasions. Consider a subset of shoppers who all have the same values for each of the covariates; i.e. a common x. These shoppers do not all have the same probability. Each shopper i has a probability, \( p_{A,i} \), which is a separate observation from a beta process; i.e. \( P_{A,i} \) is an observation of a random variable \( P \) with a beta distribution with parameters \( \alpha_A(x) \) and \( \alpha_B(x) \). Within the subset with common x all shoppers will have the same parameters for this beta process and the same expected value as in Equation 5 but they will not have the same probability \( p_{A,i} \).

The observed repeated discrete choice generates a count of the number of times shopper i selects Alternative A. This count, \( r_{A,i} \), is an observation from a binomial process; i.e. \( r_{A,i} \) is an observation of a random variable \( R \) with a binomial distribution based on \( k \) trials and the probability \( p_{A,i} \).

Thus, in the RBL model there are two stochastic processes. There are two terms which reflect the unobserved influences on shoppers’ discrete choices.

The first is reflected in the variation between shoppers in their latent probability, \( p_{A,i} \), even between shoppers with the same x. In the beta process this ‘error’ is the difference between \( P \) and \( E[P] \).

The second is reflected in the variation within shoppers in their succession of \( k \) choices. Most shoppers do not always select the same alternative. In the binomial process this ‘error’ is the difference between \( R/k \) and \( E[R/k]=P \).

Now it is also possible to discuss how RBL represents randomness in decision making. Each shopper’s discrete
choices are affected by a range of influences. These might include household’s requirements and the marketing mix for the alternatives. Each shopper has a probability $p_{A,i}$. This is nothing more than his/her long run stationary chance, or proportion of times, A would be selected. It is an aggregation of all the influences on the shopper. The differences between shoppers in the aggregation of these influences can be (1) observed and captured through $x$, or (2) unobserved and captured through the beta process. However, at successive choice occasions each shopper varies in his/her choice. This is because of fluctuations in some of the influences and is captured through the binomial process. The nature of the RBL model is that the observed influences must be constant over the period of the data collection and any fluctuation in the unobserved influences must be stationary.

For the user of RBL this stochastic structure delivers an additional benefit which is theoretically a little complex. MNL is constrained because the variances of the error terms are not identified and consequently parameter values from similar studies on different data sets cannot be directly compared. With RBL this problem does not exist. The combination of the beta and binomial processes described above leads to the beta binomial process. The RBL model identifies the variance of the beta and the beta binomial processes as functions of $x$. Thus, all the variances of the stochastic processes are identified. This is in stark contrast to the MNL model where the stochastic error term in the utility has an unidentified Gumbel distribution. There is no ambiguity, no lack of identification, in the RBL model. Consequently, the parameter values for comparative studies using RBL and different data sets can be directly compared.

6.3 Revealed Preference

RBL has been presented as a model for revealed preference data. This is data which records shoppers’ actual purchases. In the analysis of discrete choice, stated preference data is often used and is generated through respondents in surveys making hypothetical choices. The benefits of revealed preference data include (1) it can be inexpensive to obtain as it is automatically collected in customer databases and (2) it reflects shoppers’ actual choices and is not hypothetical. The benefits of stated preference data include (1) it can examine new attributes currently not available on the market and (2) it can be used to conduct experiments. This last point is critical. Revealed preference data is usually observational, stated preference is usually experimental.

Regardless of the type of data, be it discrete choice or interval/ratio data, the differences between observational and experimental data must be kept in mind. Firstly experimental data is much stronger in implying causation. Secondly with observational data there are limitations arising from omitted variables. Observational data can and will contain correlations in the covariates; i.e. colinearity. In stated preference data factorial designs can remove most colinearity. These correlations in revealed preference data lead to biased results particularly when there are omitted covariates. For example, Table 2 to Table 4 show that for two brands visits to supermarkets is significant and relevant. It may be that there is a brand targeted on this basis so the marketing mix for that brand is correlated with the visits to supermarket. By excluding marketing mix as covariates and including visits to supermarket the tables might be interpreted incorrectly. The optimum solution is to select covariates intelligently, observe as many covariates as possible and to use a multivariate model. RBL is one such model as it allows for many covariates. When applying RBL to revealed preference data, as with all observational data, care must be taken to account for missing covariates.

6.4 Further Research

RBL uses repeated binary choices, and covariates which are constant, to examine variations in market shares and repeat rates. It has been presented as a tool for revealed preference data. On this basis alone it is very useful for segmenting on repeat rates. However, there are clear benefits in expanding the model to include more than binary choice, covariates which change and experimental data. Already RBL can be used for comparisons between revealed preference data sets. Eventually the method will be expanded to stated preference where the replications involve different choice sets. Then it will be possible to compare revealed and stated preference, building on the work of Louviere, Hensher and Swait. This is the subject of further research.

7. Conclusion

The paper introduces RBL, a model for repeated binary discrete choice data. RBL identifies the characteristics of decision makers that have an impact on classic measures of loyalty such as repeat purchase. Thus, it provides a
methodology for targeting on the basis of loyalty. It has
the potential to identify empirical generalizations
regarding the roles of covariates in loyalty, or to show
that these roles vary between brands. It is a surprisingly
simple model to use. Finally, RBL's applications
potentially cover many forms of binary choice, not just
brand choice.

References

Analysis: Theory and Application to Travel Demand. The
MIT Press, Boston MA.

Cameron, A.C. and Trivedi, P.K., 1998. Regression
Analysis of Count Data. Cambridge University Press,
Cambridge and New York.

Edwards, A.W. F., 1976. Likelihood - An Account of the
Statistical Concepts of Likelihood and its Application to
Scientific Inference. Cambridge University Press,
Cambridge and New York.

Eliason, S.R., 1993. Maximum Likelihood Estimation -
Logic and Practice, Sage Publications, California.

Guadagni, P.M. and Little, J.C., 1983. A logit model of
brand choice calibrated on scanner data. Marketing
Science 2 (Summer), 203-38.

Continuous Univariate Distributions, Volume 2 (2 ed.).

Discrete Multivariate Distributions. John Wiley & Sons,

Continuous Multivariate Distributions, Volume 1:
Models and Applications (2 ed.). John Wiley & Sons,

Choice Methods Analysis and Application. Cambridge


McFadden, D., 1984. Conditional logit analysis of
qualitative choice behavior. In: Griliches, Z., Intriligator,
M.D. (Eds.), Handbook of Econometrics. North-
Holland.

McFadden, D. 1974. Conditional logit analysis of
qualitative choice behaviour. In: Zarembka, (Ed.),

Data (2nd ed.). Springer-Verlag, Berlin.

Acknowledgements

The authors wish to thank MarketingScan and GfK, and
especially Laurent Battais, Raimund Wildner, and
Gérard Hermet, for providing the data and for their
assistance in this study.

Biographies

Cam Rungie is a Senior Lecturer in Marketing at the
University of South Australia. He completed his PhD on
Discrete Choice at the University of South Australia in
2000, having previously completed an honours science
degree in mathematics at the University of Adelaide and
a Masters of Business Administration at Queen’s
University, Canada. He spent 15 years in commercial
market research and consulted to WHO on survey
methods throughout Asia. His research interests include
the analysis of purchase and audience data. He is
consulting with OzTAM and ACNielsen and won an
ESOMAR best paper award in 2000 with ACNielsen.

Gilles Laurent is Carrefour Professor of Marketing at
HEC School of Management, Paris. His Ph.D. is from
MIT (1978). He has published papers in a variety of
journals (JM, JMR, MkgS, MgtS, JCR, IJRM, MkgLet,
etc.) on somewhat unrelated topics, including: the
structure and measurement of consumer involvement,
choice by older consumers, brand awareness, evoked
sets, sales promotion, the reliability of brand attitude
measures, consumers’ immediate memory for prices,
linear programming, etc. He is a former editor of the
International Journal of Research in Marketing, and a
former academic trustee of the Marketing Science
Institute.

Correspondence Addresses

Cam Rungie, Senior Lecturer, School of Marketing, University
of South Australia, GPO Box 2471, Adelaide SA 5001,
Australia. Telephone: +61 (0)8 8302 0768. Facsimile: +61 (0)8
8302 0442. Email: cam.rungie@unisa.edu.au.; Gilles Laurent,
Carrefour Professor of Marketing, Marketing Department,
HEC School of Management, 78350 Jouy-en-Josas, France.
Telephone: +33 (0)1 39 67 74 80. Facsimile: +33 (0)1 3967 7087.
Email: laurent@hec.fr.