The S Parameter in the Dirichlet NBD Model:
a Simple Interpretation.

John. A. Bound

Abstract

In fitting the NBD Dirichlet model to consumer purchase data the S parameter requires complicated calculation and has no obvious meaning unlike the other three required parameters.

Examination of consumer panel data for many product categories shows the S parameter to be highly correlated with the average number of brands bought. It may therefore be interpreted as a measure of diversity of brand purchasing behaviour in the category.

1. Introduction.

The Dirichlet Distribution was first described by Johann Peter Gustav Lejeune Dirichlet (1805-1859). It is the multivariate generalization of the beta distribution.

The patterns of buyer behaviour for brands – how many consumers buy at all, how often they buy, what else they buy – have been well documented over many years for many product categories (Ehrenberg and Uncles 2004). In all these the patterns of choice are well described by a single model of choice behaviour using the Dirichlet Distribution. The theory and application were first described in a paper read to the Royal Statistical Society in 1984 (Goodhardt, Ehrenberg and Chatfield 1984). The model is more exactly described as the Negative Binomial Distribution-Dirichlet. The distribution of the total number of purchases in the product category is assumed to be a Negative Binomial.

The model assumes consumers choose a small portfolio from the available options (split loyalty), with on-going as-if fixed propensities to choose any one entity (e.g. brand X six times out of ten). Propensities differ greatly from consumer to consumer but are highly predictable in total and can be summarised using standard distributions, which take a similar form from brand to brand. The only input needed relating to any entity (brand, size, flavour etc.) is its overall incidence of being chosen (share of choice occasions), which, with information on how many buy the category, how often on average, and an overall measure of multiple brand buying, is enough to calibrate the model. The theory then predicts a number of performance measures for various time periods, such as how many consumers will choose an entity at all (penetration), how often (frequency), how many only choose that entity (sole loyalty), what other entities are chosen by those choosing any one (duplication), etc. The theory also predicts phenomena such as Double Jeopardy (McPhee, 1963). The theory and development of the model and its application is fully described in Chapter 13 of 'Repeat Buying' by Ehrenberg (1972).

As applied to consumer panel data the NBD Dirichlet Model for a particular time period has four parameters, M, B, K and S. The meanings of B, M and K are clear. M is the product category purchasing rate, B is the proportion of the population buying the

product category, and K is the exponent of the NBD of the number of purchases in the product category.

The meaning of the S parameter is, however, obscure. There is no equation of closed form for its calculation when fitting the model. The calculation thus requires iteration and is difficult to do without specialist software.

There is little literature on the S parameter. A Ph. D. thesis submitted to Massey University by Dr. Zane Kearns in 1999 (Kearnes 1999) considered the variation of the S parameter in various circumstances. The S parameter has more recently been used as a measure of 'polarisation', or brand loyalty, by Jarvis, Rungie and Lockshin (2007).

This Note seeks to show that the S parameter has in practice quite a simple intuitively attractive meaning as a measure of brand purchasing diversity. It is closely associated with the average number of brands bought. A regression equation derived from consumer panel data for a wide range of grocery product categories is $S = 1.3N—1$, where N is the average number of brands purchased by purchasers of the category. Very broadly indeed it might be said that S equals N.

2. Data used.

The data used are for 62 non-overlapping product categories for the six years 2000 to 2005 inclusive giving a total of 372 observations and come from the British section of Worldpanel operated by Taylor Nelson Sofres (TNS). This panel continuously records purchases made by a 15,000 sample of households. TNS have made these data available for research to the Ehrenberg Centre for Research in Marketing at the London South Bank University.

A limitation of the data is that they are all derived from one panel and therefore any biases in the panel affect all the data. Data from other panels and other countries would increase the scope of the generalisation. The size of the panel means that sampling error as opposed to bias is of little importance.

3. Results.

Table 1 below shows the correlations between commonly tabulated consumer panel top-line variables. These are:

M: the product category purchasing rate

B: the proportion of the population buying the product category

K: the exponent of the NBD of the number of purchases in the product category.

S: the Dirichlet parameter.

 M, B, K and S are the four NBD-Dirichlet parameters discussed above,

A: the scaling coefficient of the NBD Distribution

W: the average rate of purchase for the category

D: the Duplication coefficient, describing the incidence of duplicated buyers in the category (buyers of brand *1* who also buy brand *2*, *3*, etc.)

D*: the Duplication coefficient adjusted for the penetration of the category (D* = DB)

N: the average number of brands purchased in the category by purchasers of the category.

Y: a serial number 1 to 6 distinguishing the data for the individual years 2000 to 2005. This is used to check whether there is any observable change in the relationship between S and N in individual years

Table 1 shows that the variables closely associated with S are W, M and N. These three variables are all closely correlated with one another, but the most closely correlated with S is N.

Table 1: Correlation Matrix

	S	N	M	W	A	D*	B	K	D	Y
S	1.0									
N	0.8	1.0								
M	0.7	0.9	1.0							
W	0.7	0.9	0.9	1.0						
A	0.5	0.6	0.5	0.8	1.0					
D*	0.4	0.6	0.4	0.5	0.4	1.0				
B	0.3	0.6	0.6	0.4	0.0	0.4	1.0			
K	0.1	0.3	0.4	0.1	-0.2	0.1	0.9	1.0		
D	-0.1	-0.3	-0.3	-0.1	0.1	0.1	-0.7	-0.7	1.0	
Y	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	1.0

The simple correlation r in the whole dataset between the value of S and N (the number of brands bought) is 0.81. Additional predictors add little as is shown by the multiple correlation R^2 with N and additional predictors, in Table 2.
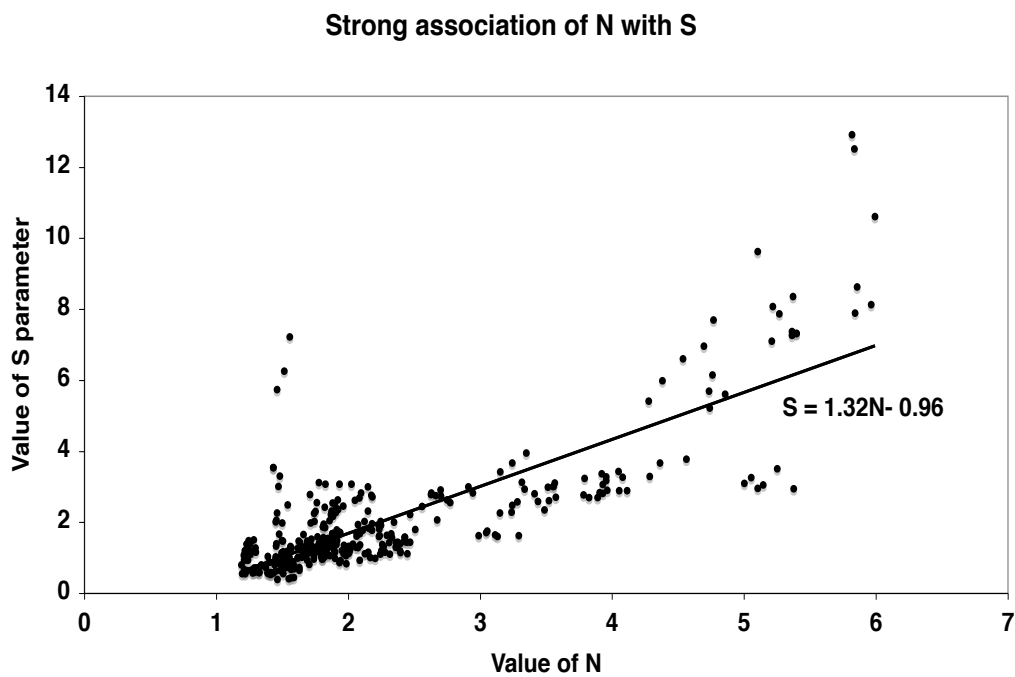
Table 2: Values of R^2

S with N	0.650
S with N and M	0.665
S with N, W and M	0.666
S with all 9 tabulated descriptors	0.780

4.2 The relation of S to the number of brands bought (N)

Chart 1 shows the strong approximately linear association of S and N with some notable outliers. For example, the linear equation under-predicts Sugar, Chocolate Confectionery and Cat Food (categories for which the rate of purchase is very high), but over-predicts Wrapped Bread and Carbonated Drinks such as Coca Cola (categories with very high annual penetration).

Chart 1

Strong association of N with S

$S = 1.32N - 0.96$

5. Applications

The S parameter has been used to calculate φ, a measure of polarisation, which has been defined as a brand loyalty measure. Jarvis, Rungie and Lockshin (2007) define it as φ = 1/ 1+ S. In that paper φ is used to compare the brand metrics of product categories of repertoire and subscription type categories.

Since the concept of the average number of brands purchased is so much simpler it might usefully be used as a criterion of brand loyalty for analysis of product categories categorised in any way. An objection to this simple model is that S is associated with the penetration (r = 0.3) and volume of purchase of the category (r = 0.6) and as noted above, a three parameter linear model increases the R^2 between observed and predicted S to 0.8. The simple model is in the other hand intuitively meaningful and easy to apply.

It should also be noted that the value of N depends on the length of the analysis period. Over a longer time period people buy a greater number of competing brands. The present data are for 52 weeks and for fmcg categories that seems likely to be close to a maximum value for N. N may be regarded as the average repertoire size for the category. It may be noted that if few brands are analysed separately and many are grouped in an 'all other' category or similar ranges N is reduced. S and hence φ are not affected by the length of the analysis period.

It remains more appropriate to use φ as a criterion variable when the analysis, as in Jarvis, Rungie and Lockshin (2007), is of individual brands for which the values of S and hence of φ vary within a category. The S value used for the analysis by categories above is an average of all the individual brand S values within a category, weighted by brand share.

References:

Ehrenberg A.S.C. Repeat Buying Charles Griffin & Co. Ltd., London, Oxford University Press, New York, 1972, new edition 1988.
(This is now out of print but the complete text is available at:
http://www.empgens.com/index.htmlo)

Goodhardt, G. J., Ehrenberg, A.S.C. and Chatfield, C. (1984), "The Dirichlet: A Comprehensive Model of Buying Behaviour", *Journal of the Royal Statistical Society Series A,* **147**, 621-655.

Jarvis, W., Rungie C. M. and Lockshin L (2007) The polarisation method for merging data files and analysing loyalty to product attributes, prices and brands in revealed preference, International Journal of Market Research, **49,** 4: Data Integration Special Issue, pp. 487-513.

Kearns, Z. A systematic investigation of the estimation of the Dirichlet model : a thesis presented in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Marketing at Massey University, 1999.