GERALD J. ESKIN*

A depth of repeat model is presented that can forecast the demand for new consumer products. The relation of the model to other forecasting models is noted. Data analysis, estimation procedures, and the observed accuracy of forecasts are discussed.

Dynamic Forecasts of New Product Demand Using a Depth of Repeat Model

INTRODUCTION

A common problem encountered in new product marketing concerns the prediction of future sales, given observations for the first few periods of a product's life. Figure 1 illustrates this, where the projection problem can be defined: given values of the sales curve for all values of time up to a point *t*, determine which extension of the curve is most likely to occur. It must be determined if the product is to grow as in curve *a*, decay as in curve *b*, or "level out" as in curve *c*. Fate and other environmental conditions tend to lead to situations, as in Figure 1, where one is required to project "too early" in the sense that the product is still in the development or growth stage of the life cycle and hence direct curve extrapolation is foolhardy.

In the case of frequently purchased goods, consumer diary panels have been found useful in the projection process. Such panels allow the analyst to decompose sales and construct diagnostic measures that suggest the future course of events.

One popular device is to decompose sales volume (or transaction counts) into *trial* and *repeat* parts. Rules of thumb are then developed, high trial often being taken as a measure of initial success (e.g., first year sales) and high repeat being considered a measure of long-run potential.

Based on the trial–repeat partition, formal sales projection models have been developed, one of the earliest being the Fourt-Woodlock Model [7]. In that model, trial (cumulative proportion of a population buying at least once up to some time point) is described by a mathematical function based on the postulate that some fixed proportion of the population will eventually become triers, and that out of this group a fixed fraction of those persons that have not yet tried will do so in every period. Their mathematical representation of trial was found to fit the data well except in the upper tail (i.e., for large values of time). The poor fit was attributed to heavy buyers becoming triers earlier than light buyers. A trend factor was added to the equation to adjust for this discrepancy. Repeat is modeled by Fourt-Woodlock as a series of conversion ratios (i.e., the proportion that purchases J times, given that they have purchased $J - 1$ times). In estimating these ratios, account is taken of the average time between successive purchases. More recently work by Parfitt and Collins [15] and Ahl [1] have redirected the trial-repeat model for use in predicting share.

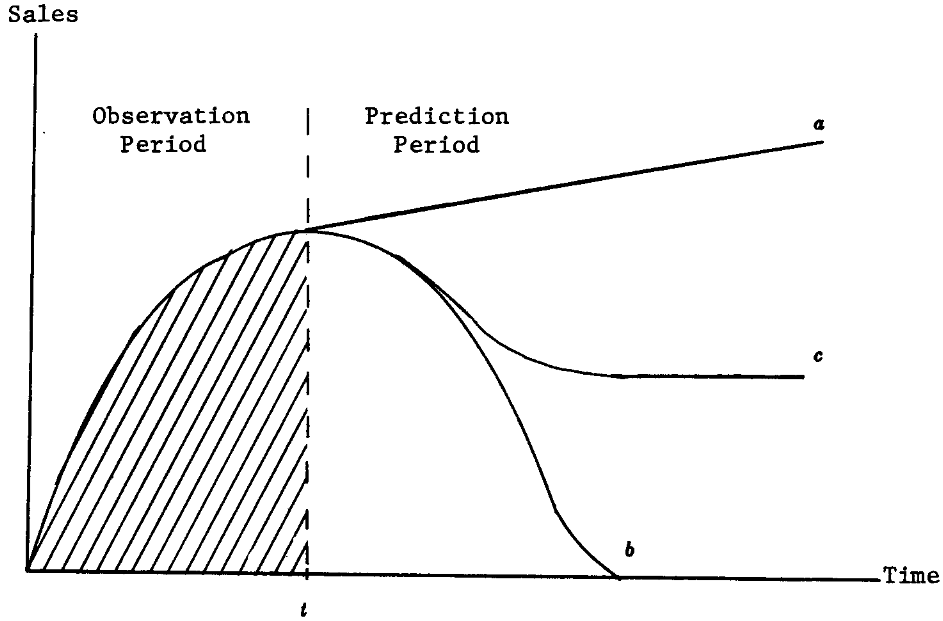
A sophisticated new product model has been developed by Massy [11, 12, 13] (called STEAM). It is a general and dynamic model presuming population heterogeneity. The model describes the propensity to enter various purchase classes (called depth of trial classes) for each consumer. These propensities depend on factors such as average product usage, timing of last purchase, and time since last purchase. Sophisticated data processing techniques are used in preparing for estimation. Estimation of the large number of parameters of the model is accomplished by search techniques; a simulation routine is then used to make projections.

The STEAM model appears to be a highly desirable prediction and monitoring device for situations where the quality and quantity of data is sufficient to parameterize its complex structure and whenever decision prob-

* Gerald J. Eskin is Assistant Professor of Marketing, The University of Iowa. He wishes to acknowledge the support provided by the Pillsbury Company to the research here reported. In particular Lynn Linn, Channing Stowell, and Dudley Ruch each made substantial contributions both to the development of the model and to subsequent efforts to verify and implement the results.

Figure 1

POSSIBLE EXTRAPOLATIONS OF SALES PATH



lems are of sufficient magnitude to warrant the relatively high cost of utilizing this type of model. Conversely the Fourt–Woodlock approach is simpler and possibly more understandable to the nontechnician. In general, it requires less complex data processing, these advantages being achieved at a cost in terms of diagnostic power and predictive accuracy.

Our objective here is to provide an "in-between" model with at least some of the desirable attributes of each. In particular we wish to maintain some of the simplicity and the intuitive structure of Fourt-Woodlock, while utilizing more complete data analysis within the framework of well-known estimation methods.

THE MODEL

Definitions

Our primary interest concerns the diagnosis and projection of cumulative sales S_t of a product up to some time point t. These sales arise by consumers engaging in a certain number of transactions and acquiring some number of units on each of these transaction occasions. Consumers are grouped in terms of these transaction numbers. We define $R_t(J)$ as cumulative number of consumers repeating at least J times by period t. $R(1)$ tells us how many consumers have repurchased at all (i.e., have made at least 2 purchases) and will be referred to as the repeat function. By 0 repeats ($J = 0$) we mean one purchase, thus $R(0)$ is the trial function.

Average units purchased on the Jth repurchase occasion will be denoted $U_t(J)$. Hence cumulative sales to t is:

$$S_t \equiv \sum_{J=0}^{\infty} R_t(J) U_t(J). \tag{1}$$

This equation merely states that volume purchased at a given cumulative repeat level is the product of the number of consumers purchasing at least that number of times and the average amount purchased, and that total sales up to time period t can be obtained by adding these values over all repeat levels.

In addition to classifying consumers by number of times purchased, it is possible to further subdivide based on the time at which a previous purchase was made. Thus we will consider $RI_{it}(J)$, the cumulative fraction that repeat a Jth time by period t given that the $J - 1$ purchase was made during period i. RI differs from R in that it is a fraction rather than an integer and in that it is conditional on the time of entry into the previous state. The algebraic relation between RI and R is:

$$R_t(J) \equiv \sum_{i=1}^{t} RI_{it}(J)\, \Delta R_i(J - 1) \qquad J = 1, 2, 3, \cdots, \tag{2}$$

where Δ means the first difference operator in time. The equation has no meaning for $J = 0$, thus only positive integers are included in the domain.

Equations (1) and (2) are definitional in nature. They provide a basis for sorting panel data in ways that may suggest the future course of events. These sorts are illustrated in Table 1 for hypothetical data covering four observation periods. Raw data counts are displayed in the left portion of the table, with resulting values of RI displayed on the right. Given trial values $R_t(0)$ and the

Table 1
CONSUMER DIARY PANEL
DATA SORTED BY REPEAT LEVEL AND TIME OF ENTRY

Time period	*Number of triers by period*	*Number of 1st repeaters by time of repeat t and time of previous purchase i*					*Values of RI(1)*				
							i				
			1	*2*	*3*	*4*	*t*	*1*	*2*	*3*	*4*
1	300	1	0	150	75	27	1	0	.5	.75	.875
2	150	2		0	37	18	2		0	.25	.375
3	100	3			0	25	3			0	.25
4	70	4				0	4				0

$R_4(0) = 620$: Cumulative Trial

Time period	*Number of repeaters by period*	*Number of 2nd repeats by time of repeat t and time of previous purchase i*					*Values of RI(2)*				
							i				
			1	*2*	*3*	*4*	*t*	*1*	*2*	*3*	*4*
1	0	1	—	—	—	—	1	—	—	—	—
2	150	2		0	75	37	2		0	.5	.75
3	112	3			0	56	3			0	.5
4	70	4				0	4				0

$R_4(1) = 332$: Cumulative 1st Repeat

Time period	*Number of 2nd repeats by period*	*Number of 3rd repeats by time of repeat t and time of previous purchase i*					*Values of RI(3)*				
							i				
			1	*2*	*3*	*4*	*t*	*1*	*2*	*3*	*4*
1	0	1	—	—	—	—	1	—	—	—	—
2	0	2		—	—	—	2		—	—	—
3	75	3			0	37	3			0	.5
4	93	4				0	4				0

$R_4(2) = 168$: Cumulative 2nd Repeat

Total transactions	1157
Trial transactions	620
Repeat transactions	537

conditional repeat proportion $RI_{tt}(1)$, the number of first repeats can be determined. This plus $RI(2)$ determines second repeat. Thus, knowing trial and $RI(J)$ for all J is sufficient to determine sales (ignoring for the present multiple purchases on a single transaction occasion). Projection can be thought of as involving filling in additional values for trial and adding new columns to the $RI(J)$ tables, then, using the definitional formulas. Values to be entered in the new columns should be suggested by the entry already in the table.

The $RI(J)$ tables have rows corresponding to time of entry indices i for two reasons. First, it has been observed that early buyers behave differently than late buyers; hence one would not wish to predict sales for late buyers based on aggregates dominated by early buyers.[1] Variation due to time of entry might also be due to marketing factors such as changes in distribution and promotion. Another reason for sorting on i concerns the fact that each time of entry group has different amounts of time available in which to repurchase. This results in the triangular shape of the table where the number of periods available for repurchase is related to i (i.e., the available time is $t - i$).

Behavioral Assumptions

To predict future values of $RI(J)$ requires behavioral assumptions relating observed values of $RI(J)$ to future values. As a preliminary step, some established products were studied in hopes of discovering behavioral regularities which might carry over to the new product situation. This investigation is summarized in the final section; here we merely refer to a typical plot of RI (for one of the established products) to motivate the behavioral assumptions that will be specified below.

Figure 2 shows a plot of $RI_{\cdot T}(J)$ values for an established product. The "$\cdot$" notation in the first subscript position is to indicate that the data have been summed over all time of entry groups, and the second subscript ($T \equiv t - i$) indicates that time is measured from the time of entry into the J − 1st repeat class. The plot is given for six repeat levels ($J = 1$ to 6) and for 52 weeks from time of entry into the previous repeat class (i.e., $T = 1, 2, \cdots 52$).

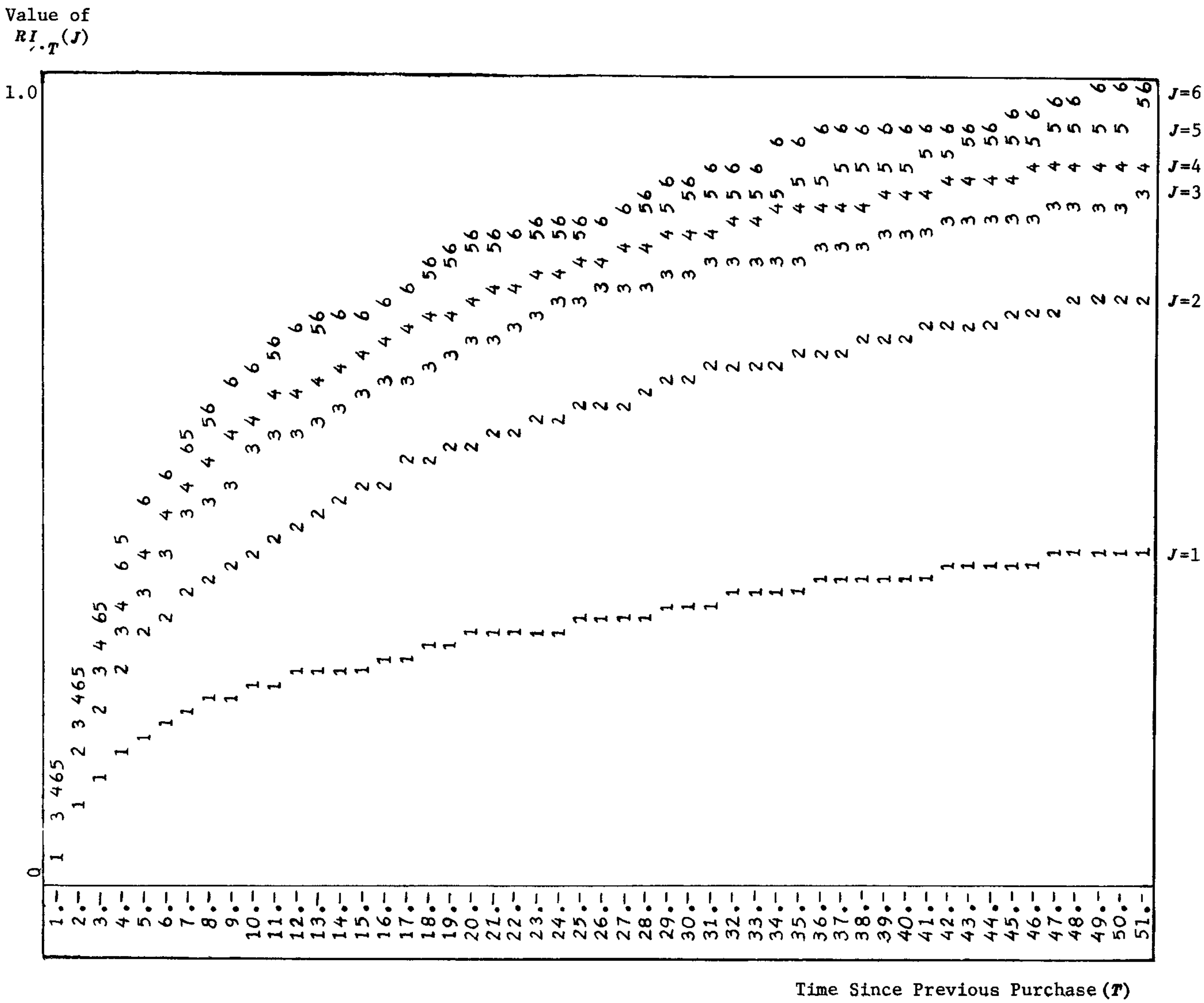
Casual inspection suggests:

1. Each curve rises rapidly at first with more buyers entering in the first period than in any other (i.e., the mode time to entry is one week). Latter portion of curve rises quite slowly but does not reach a limit within 52 weeks.
2. The six curves seem to be similar in that they are roughly proportional in the early part of the curve and parallel in the latter portion.
3. The proportions repurchasing within 52 weeks, $RI_{\cdot 52}(J)$, tend to increase in a monotone fashion over J. The increases tend to occur at a decreasing rate. For example, the difference between $RI(2)$ and $RI(1)$ is greater than the difference between $RI(3)$ and $RI(2)$.

[1] Fourt-Woodlock [7, p. 35] note that, "The first new buyers of an item are typically heavy buyers of a product class." Quantitative evidence on repeat by time of trial is given by Parfitt-Collins [15, p. 136].

Figure 2

CUMULATIVE PROPORTION REPURCHASING J TIMES (OUT OF THOSE PURCHASING $J - 1$ TIMES)



The following behavioral specifications are consistent with observations 1 through 3:

BEHAVIORAL ASSUMPTION 1: Geometric-Stretch Function

$$RI_{it}(J) = \alpha_{iJ}(1 - \gamma_{iJ}^{T}) + \delta_{iJ}T \tag{3}$$

for $T \equiv (t - i) \geqq 0$.

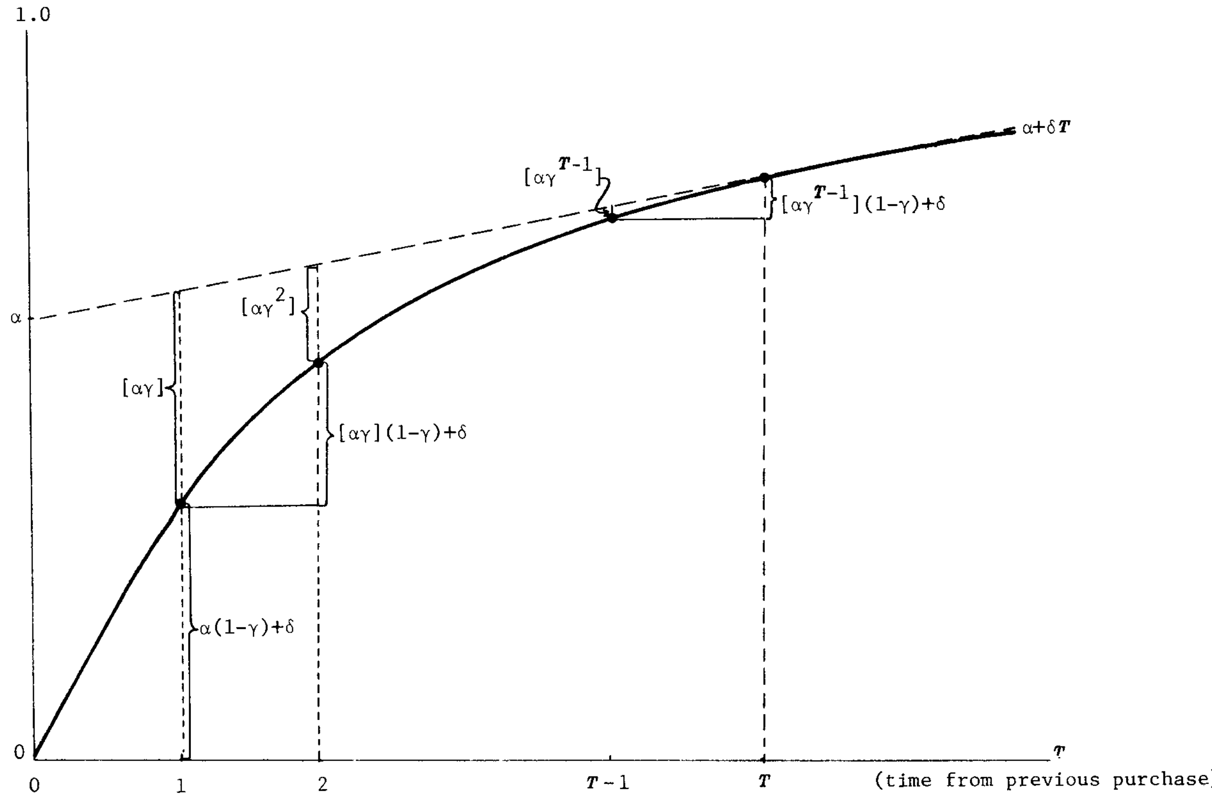
For $T < 0$, RI is defined as 0. For large T the RI function may exceed 1. In such cases RI is defined as the minimum of (3) or the number 1.

Equation (3) is equivalent to that proposed by Fourt–Woodlock as a model of trial. Here we apply it to trial and all repeat levels. It is graphed in Figure 3, where it can be seen that the curve is always below and approaches the ceiling line: $\alpha + \delta T$, the number of consumers entering the repeat class each period being equal to a constant δ plus a fixed proportion of the distance between the current value of the function and the limit line, the proportionality constant being $(1 - \gamma)$. Because γ measures how quickly the curve approaches the limit line, it shall be referred to as the *slope* factor. The effect of δ is to stretch out the function; hence it will be called the *stretch* factor. The function in general, i.e., (3), will be called the *Geometric-Stretch* function.[2]

The parameters of the Geometric-Stretch function may depend on various subscripts. The parameter α will

[2] The specification (3) implies that the maximum number of transactions per consumer within one unit of time is 1. This appears to be approximately true for our data, for time measured in weekly units. For estimation purposes, second purchases within a single week are coded in the following week.

Figure 3
EXAMPLE OF GEOMETRIC-STRETCH FUNCTION



depend on the repeat level J and α and γ may depend on time of entry i. Observation 2 above suggests that some parameters are constant for all repeat levels. Formally we assume:

BEHAVIORAL ASSUMPTION 2: Parameter Stability

(a) γ's are the same for all repeat levels (except trial), i.e., $\gamma_i = \gamma_{iJ}$ for all $J \geqslant 1$.

(b) δ's are the same for all repeat levels and time of entry cells, i.e., $\delta = \delta_{iJ}$ all $J \geqslant 1$ and all i.

Based on observation 3, we make the following assumption concerning the proportion repurchasing within one year.

BEHAVIORAL ASSUMPTION 3: Conversion Proportions

(4) $$RI_{.52}(J) = RI_{.52}(\infty)(1 - \lambda_1^J)$$

for $J = 2, 3 \cdots$.

This function is graphed in Figure 4. Trial and first repeat are excluded from the continuous part of the function, given their critical role in determining sales and given that they do not predict higher values of repeat well.

The functional form of (4) is similar to (3). In (4) no stretch factor is assumed; hence the sequence of $RI(J)$'s approaches a limit $RI(\infty)$. This limit, by the definition of RI, can be no larger than 1 but may be less than that value. For the type of products studied here and for forecasts of a year or less, measuring conversion proportion at $T = 52$ appears adequate. For other situations some modification of (4) may be required.

These three behavioral assumptions, along with the definitional equations (1) and (2), constitute a complete model in number of transactions.[3] To translate transactions into sales terms requires a specification of the units purchased per transaction. Based on analysis (not presented here), we take as a first approximation of the units per transactions relation the following:

BEHAVIORAL ASSUMPTION 4: Units Per Transaction

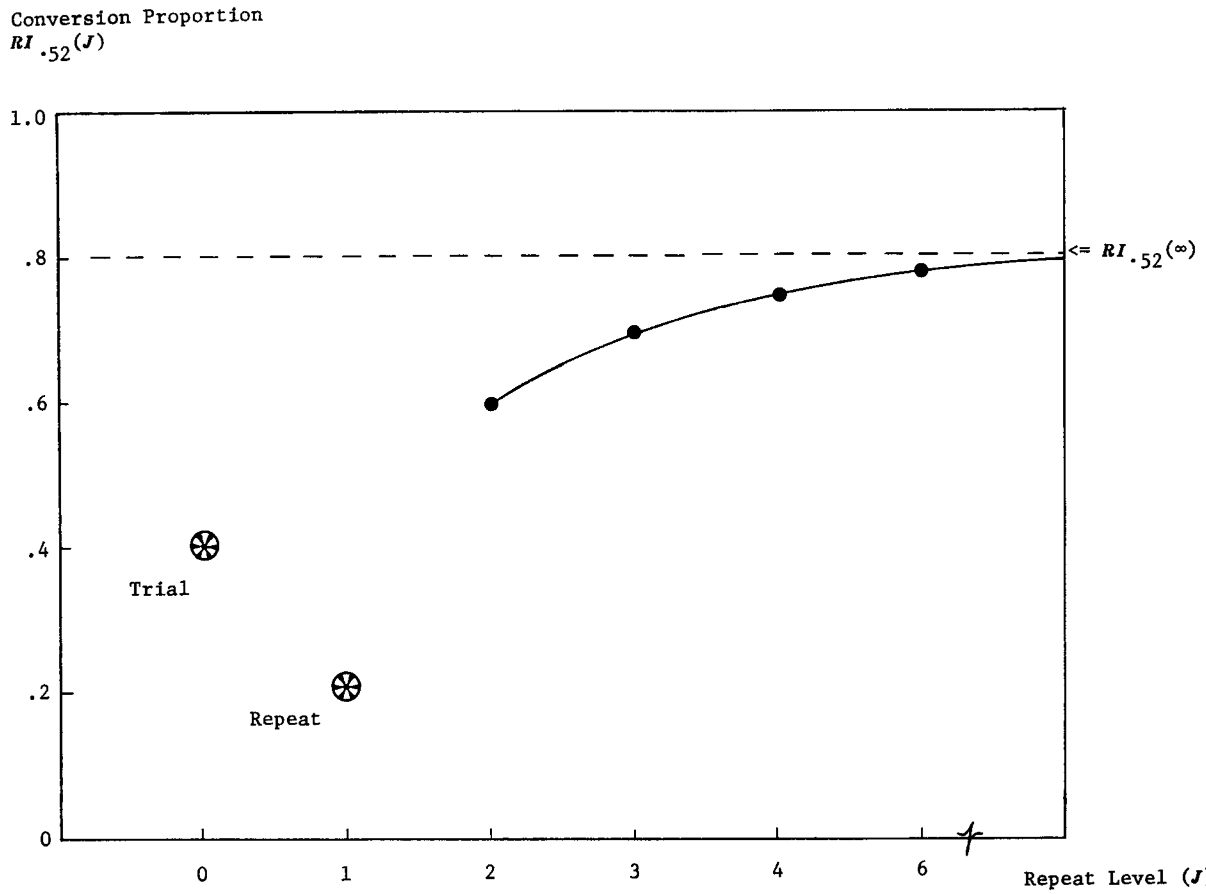
(5) $$U_t(J) = \min\left[\begin{matrix}\mu_o + \mu_1 J\\ \bar{U}\end{matrix}\right]$$

for $J = 1, 2, \cdots$.

[3] In fact all data in Table 1 are generated by (1) through (4). With parameters: $\delta = 0$ for all i and t, $\gamma = .5$ for all i and t, and $\alpha = .5$ for $J = 1$, $i = 2$ and 3; $\alpha = 1$ otherwise.

Figure 4

FRACTION REPURCHASING J TIME WITHIN ONE YEAR OF A PREVIOUS PURCHASE



That is, the units function (5) covers all repeat levels except trial and is assumed to be linear up to some maximum value $\bar{U}$. Units on trial are some constant $U_t(0) = U(0)$. All U values are assumed to be independent of t.

The above specifications are encompassed in a computer program written in the BASIC language for time-sharing application. Sample interactive Input-Output for the program (called PANPRO for panel projection model) is exhibited in Figure 5.[4]

Characteristics of the Model

Insights into the behavior of the model can be obtained by examining the repeat buying behavior for all persons for which trial occurs in some specified period. First, ignoring the stretch factor (i.e., assume $\delta = 0$) and the possibility that everyone will not eventually repeat and differences in time of entry parameters, it can be shown that:

$$(6)\quad \Delta R_t(J) = \binom{t-1}{t-J}\gamma^{t-J}(1-\gamma)^J = \binom{\overset{*}{T}+J-1}{\overset{*}{T}}\gamma^{\overset{*}{T}}(1-\gamma)^J$$

for all $\overset{*}{T} \equiv t - J \geqq 0$; and all $J \geqq 1$,

where the large brackets denote the combinational operator; thus (6) is the Negative Binomial Distribution in $\overset{*}{T}$.[5] The resulting repeat function $R_t(J)$ is graphed in Figure 6, where it can be seen that $R(1)$ has the geo-

[4] The sample inputs (in Figure 5) are the Newprod estimates found in Tables 2 and 4 (based on 52 weeks of data). Thus the output generated is the detailed version of the projections summarized in Table 5.

[5] It is known that the J-fold convolution of the geometric distribution is the Negative Binomial distribution with parameter J [9, p. 6]. Equation (6) should not be confused with the Ehrenberg [5, 6] result. In (6) we have the Negative Binomial distribution over time $\overset{*}{T}$ where Ehrenberg's results involve a distribution over units. Grahn [8] has extended the Ehrenberg approach to a distribution over J.

metric shape of (3) (for $\alpha = 1$, $\delta = 0$). Higher level repeat curves exhibit an S-shape, with few repeaters entering the higher classes until relatively high values of t are reached. It can also be shown that for non-small values of t, the distribution of R across J is approximately Poisson.[6]

Reintroducing conversion ratios $RI_{.52}(J)$ less than 1 tends to increase the distance between the curves, while a nonzero stretch factor $\delta > 0$ will increase the slope in the later portion of each curve, affecting lower repeat levels more than higher ones.

ESTIMATION

In this section a method for estimating the parameters of (3) and (4) is considered. The reader primarily interested in the accuracy of prediction may wish to proceed to the next section.

First it will be shown that there exists a linear model, amenable to regression methods, such that knowledge of the parameters of that model allows determination of the parameters of (3) and (4).

For compactness in this section we drop the J and i subscript and use the notation RI_T to indicate values of RI measured from $T = 0$ (i.e., from time of entry). Equation (3) can then be written:

$$\begin{aligned} RI_T &= \alpha(1 - \gamma^T) + \delta T \\ &= \alpha(1 - \gamma\cdot\gamma^{T-1}) + \delta(T-1) + \delta, \end{aligned} \tag{7}$$

and $RI_{T-1} = \alpha(1 - \gamma^{T-1}) + \delta(T-1)$, hence

$$-\gamma^{T-1} = [RI_{T-1} - \alpha - \delta(T-1)]/\alpha. \tag{8}$$

Substituting (8) into (7)

$$\begin{aligned} RI_T &= \alpha + \gamma[RI_{T-1} - \alpha - \delta(T-1)] \\ &\quad + \delta(T-1) + \delta \\ &= [\alpha(1-\gamma) + \delta] + \gamma RI_{T-1} \\ &\quad + [\delta(1-\gamma)](T-1). \end{aligned} \tag{9}$$

Equation (9) has the linear autoregressive form:

$$\begin{aligned} RI_T &= \beta_o + \gamma RI_{T-1} + \beta_1(T-1) \\ T &= 1, 2, \cdots, \\ RI_o &= 0, \end{aligned} \tag{10}$$

where $\beta_1 = \delta(1-\gamma)$, thus:

$$\delta = \beta_1/(1-\gamma). \tag{11}$$

$\beta_o = \alpha(1-\gamma) + \delta$, thus

$$\begin{aligned} \alpha &= (\beta_o - \delta)/(1-\gamma) \\ &= [\beta_o - \beta_1/(1-\gamma)]/(1-\gamma). \end{aligned} \tag{12}$$

[6] This is due to the similarity of the geometric and exponential distributions. It is well-known that the time between purchases for a Poisson process is exponential.

Figure 5

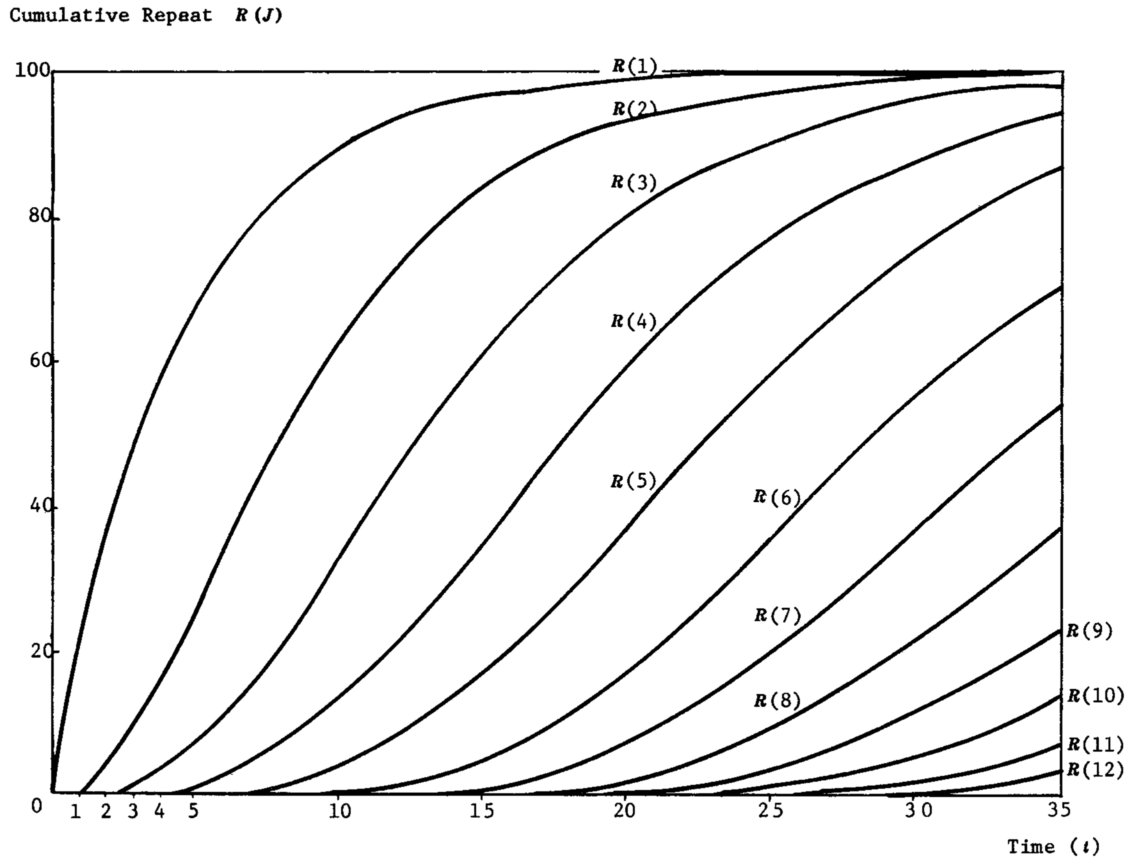
SAMPLE INPUT/OUTPUT FOR THE *PANPRO* MODEL

```
PANPRO

NUMBER OF TIME PERIODS IN PROJECTION          ?36
NUMBER OF TIME PERIODS IN EACH REPORT         ?12

TYPE 1 IF TRIAL IS TOBE DIRECTLY INPUT
TYPE 2 IF TOBE ESTIMATED BY A FUNCTION        ?2
SPEC. LIMIT, SLOPE AND STRETCH FACTORS FOR TRIAL?92,.88,.7

SPEC LIMIT FOR REPEAT LEVEL 1 AND NUMBER OF PERIODS
?.75,4
?.45,4
?.56,8
?.65,40
SPEC GENERAL SLOPE AND STRETCH FACTORS FOR REPEAT?.73,.0041
SPEC EARLY BUYER SLOPE AND PERIODS COVERED?.63,4
SPEC LIMIT FOR REPEAT LEVEL 2 AND MAX LIMIT?.8,1

INPUT AVE UNITS ON TRAIL,AU ON REPEAT 1,AND SLOPE?1.1,1.3,.1
SPEC LIMIT ON AVE UNITS        ?1.8

PERIOD           12

R( 1    )        20          .383201
R( 2    )        10          .502099
R( 3    )        6           .550691
R( 4    )        3           .490254
R( 5    )        1           .329937
TOTAL SALES      111.708
TRIAL SALES      57.2096
REPEAT SALES     54.4982     SALES PER REP   2.73451

PERIOD           24

R( 1    )        31          .447049
R( 2    )        19          .598596
R( 3    )        13          .708653
R( 4    )        10          .719794
R( 5    )        6           .654756
R( 6    )        4           .599657
R( 7    )        2           .537815
R( 8    )        1           .473843
TOTAL SALES      202.373
TRIAL SALES      76.7952
REPEAT SALES     125.578     SALES PER REP   4.02361

PERIOD           36

R( 1    )        40          .496751
R( 2    )        26          .645877
R( 3    )        20          .76331
R( 4    )        15          .783023
R( 5    )        12          .750836
R( 6    )        8           .725838
R( 7    )        6           .69533
R( 8    )        4           .659983
R( 9    )        2           .620378
R( 10   )        1           .577726
R( 11   )        1           .533674
TOTAL SALES      291.61
TRIAL SALES      88.2665
REPEAT SALES     203.343     SALES PER REP   5.10139
```

Under the assumption of additive nonautocorrelative error terms which are also uncorrelated with past values of RI, (10) can be fitted by ordinary least square (OLSQ), yielding consistent estimates of (β_o, γ, β_1). Estimates of (α, δ) may be obtained via (11) and (12). These estimates will also be consistent if the estimates of the parameters of (10) are consistent, though finite sample bias in γ will in general be present. If, however, the errors are correlated (i.e., autocorrelation is present) not even consistency is ensured. (See [10, Chapter 14] for a general treatment of the problems associated with the autoregressive model.) In this instance the presumption of uncorrelated error terms is tenuous. The fact that RI is defined as a cumulative proportion (and hence

Figure 6

MODEL FORECAST OF REPEAT $R(J)$ PER 100 NEW TRIERS[a]



[a] Plot assume $\gamma = .8$, $\delta = 0$, $RI_{.52}(J) = 1$ for all J, all trial occurs at $t = 0$.

nondecreasing) raises the possibility of positive autocorrelation.

In this situation it appears prudent to at least test for the existence of autocorrelation and possibly use special estimation procedures appropriate to such situations. An appropriate test within the context of the autoregressive model is the Durbin test of the h statistic [3] which is based on the Durbin-Watson [4] statistic.

In the presence of autocorrelation, consistent parameter estimates can be obtained via the use of a generalized least squares procedure. When estimating parameters for the established products, the Cochrane-Orcutt [14] (CORC) method is used (as well as ordinary least squares-OLSQ). In the CORC method, an initial estimate of the autocorrelative factor ρ is constructed based on a regression procedure and an iterative method is used to develop a final estimate of ρ and the other parameters of the model.

In situations where the RI curve is well developed (i.e., T is large), (10) may be estimated for each repeat level separately or for each time of entry class (within repeat level). During the introductory period for a new product, this cannot be done because of the limited number of observations and because of the truncation problem (exhibited in Table 1). In such situations, simultaneous estimation of all parameters for all time of entry groups within a single equation model is desirable. Equation (10) can be modified to estimate separate α's for each time of entry group but common γ and δ values. The modified equation is:

$$(13) \quad RI_{iT} = \sum_{i=1}^{I} \beta_o^{(i)} d_i + \gamma RI_{iT-1} + \beta_1(T-1)$$

$$\alpha^{(i)} = \beta_o^{(i)}/(1-\gamma)$$

$$d_i = \begin{Bmatrix} 1\text{: if time of entry is } i \\ 0\text{: otherwise} \end{Bmatrix}$$

$i = 1, 2, \cdots I$ (I = number of entry periods).

Modifications are also possible which allow separating by i cell the parameters γ and δ. An equation involving dummy variables as in (13) may also be used to construct estimates that pool over repeat levels J as well as over time of entry classes. By deleting β_1, (10) may also be used to estimate the parameters of the sequence of conversion proportions in (4).

Another estimation problem develops when only a short time series is available. For short series, the high degree of multicollinearity between RI_{T-1} and T result in unstable estimates of γ and β_1. Whenever an extraneous estimator of β_1 is available, it is possible to estimate the remaining parameters subject to this information. The method involves estimating the parameters (β_o, γ) in:

$$\{RI_T - \hat{\beta}_1(T-1)\} = \beta_o + \gamma(RI_{T-1}) \tag{14}$$

where $\hat{\beta}_1$ is the extraneous estimator.

VALIDATION ON A NEW PRODUCT

Here, data on a new product (called Newprod) will be examined and compared to model values to evaluate fit and predictive accuracy of the model. The data under study were generated by a consumer diary panel conducted during the test marketing of Newprod. Data were collected for the first 52 weeks that the product was available in the market. The panel contained 500 to 1,000 members, the exact size being withheld to ensure nondisclosure of proprietary information.

The analysis was conducted in two stages. First, the repeat buying component of the model was evaluated. For this purpose, actual trial $R(0)$ was taken as given, estimates of the repeat parameters were constructed, and projections were generated by the PANPRO model. These were then compared to actuals. The second portion of the analysis involved fitting trial and transaction-size functions to evaluate the total fit of the model.

In each case, the parameters of the model were fitted, using the first 12, 24, and 52 weeks of data. For each of these data sets, projections were made for $t = 12$, 24, 36, and 52. Thus sometimes our procedure generates a "fit" measure and sometimes a projection.

Repeat Buying

The estimation procedure used is basically that described in the estimation section. That is, a modified version of the least squares procedure, based on the autoregressive relation between $RI_T(J)$ and $RI_{T-1}(J)$, was used. Because of the sparseness of the data, no attempt was made to estimate repeat functions separately for each repeat level. Instead, one pooled estimate was constructed as in (13). The pooling is across all J (repeat levels) but allows separate estimates of conversion proportions for each repeat level and for various time of entry groups. The estimation method also allows for different timing of repurchases for "early buyers" via the construction of separate γ's for this group, here "early buyers" being defined as those that purchase in the first month.

As can be seen from Table 2 by examining the estimate based on all 52 weeks of data, early buyers do tend to repurchase faster than other buyers. This exhibits itself in the form of an estimated γ a full .1 lower than that for the rest of the population. First conversion rates $RI_{.52}(1)$ also differ by time of entry, those purchasing within the first month having the highest repurchase rate (.75). The rate then dips in the second month (to .45) and then partially recovers. This dip is concurrent with an out-of-stock problem that developed around the sixth week of the test. The out-of-stock situation was partially due to higher-than-expected sales during the first month. The pattern in the conversion ratios also follows in a rough fashion the trends in media spending. These had started at a high level, then were cut drastically in response to the out-of-stock condition. Approximately six months into the test, a second wave of high spending was inaugurated as the stocking difficulties abated.

Table 2

ESTIMATED REPEAT PARAMETERS FOR NEWPROD BASED ON THE FIRST 12, 24, AND 52 WEEKS IN TEST MARKET

Parameters	Number of weeks included in fit		
	12	24	52
Slope factor (γ_i)	.58	.61 (1–4)[a] .67 (5–52)	.63 (1–4) .73 (5–52)
Stretch factor (δ)	.0040[b]	.0027	.0041
First repeat conversion ratio ($RI_{i\,52}(1)$)	.76 (1–4) .39 (5–8) .58 (9–52)	.75 (1–4) .37 (5–12) .55 (13–52)	.75 (1–4) .45 (5–8) .56 (9–16) .65 (17–52)
Second repeat conversion ratio ($RI_{.52}(2)$)	.86	.68	.80
Equilibrium repeat conversion ratio ($RI_{.52}(\infty)$)	.97	.97	1.00

[a] Weeks included in each time of entry cell i are shown in parenthesis.
[b] Indicates that value shown is an extraneous estimate.

As we attempt to estimate parameters based on shorter time slices ($t = 12$ and 24), difficulties are encountered. These are most extreme in the case of $t = 12$. At this time point it is not possible to clearly separate out "early" and "late" buyer behavior as was possible for $t = 52$. The data are essentially dominated by early buyers. This can be seen in the first column of Table 2 where only one γ is shown, and the gradations of $RI_{it}(1)$ are not as fine. Estimation of the stretch factor δ is also difficult at this time point. Formally the difficulty is related to the fact that there is a high degree of multicollinearity between the sequence RI_{T-1} and T for small values of T. This makes it virtually impossible to estimate both γ and δ accurately. Our approach was to estimate σ based on an extraneous estimator of δ using

Table 3

COMPARISON OF ACTUAL AND ESTIMATED REPEAT FOR VARIOUS ESTIMATION AND FORECAST PERIODS

Repeat level	*Prediction to 12 weeks*				*Prediction to 24 weeks*				*Prediction to 36 weeks*				*Prediction to 52 weeks*			
	Actual	*Fit period*			*Actual*	*Fit period*			*Actual*	*Fit period*			*Actual*	*Fit period*		
		12	*24*	*52*		*12*	*24*	*52*		*12*	*24*	*52*		*12*	*24*	*52*
1	19	19	20	19	30	31	31	31	40	38	38	40	46	47	46	49
2	10	11	10	10	17	20	17	19	27	27	22	25	34	35	28	34
3	6	7	6	6	11	15	12	13	17	21	16	19	24	30	22	28
4	4	4	3	3	9	11	9	10	14	16	13	15	21	25	19	24
5	3	2	1	1	6	8	7	6	8	13	11	12	14	20	16	19
6	2	1			4	5	5	4	6	10	8	9	14	16	13	15
7	1				4	3	3	2	5	7	7	6	9	13	11	12
8					2	2	2	1	5	5	5	4	9	10	9	9
9					1	1	1	1	3	4	4	2	8	8	7	7
10					1	1			3	3	2	1	6	6	6	5
11					1				2	2	2	1	5	4	4	3
12					1				2	1	1		5	3	3	2
Remainder									3	1	1		12	8	7	4
Total transactions	46	44	40	39	87	97	87	87	135	148	130	134	207	225	191	211

(14). The extraneous estimator used was the average value of δ for the six established products.[7]

Based on the estimates in Table 2, projections of transactions counts by repeat level were made for 12, 24, 36, and 52 weeks using the PANPRO model (see Table 3).

The projections based on 52 weeks of data are all "fits" (as opposed to projections into the future) given that all data up to $t = 52$ was used in the estimation procedure. Without resorting to formal tests, our reading of the results is that the fit to total transactions is quite good at 52 weeks (211 vs. 207 actual) and that at other points the fit is adequate. The model tends not to fit the distribution over J as precisely, exhibiting a tendency to underestimate the number of buyers at high repeat levels.

The estimates based on 12 weeks of data constitute "fit" values at the 12-week time point, and projections into the future at the 24-, 36-, and 52-week points. As expected, the best fit occurs at 12 weeks (46 vs. 44 for total repeat transactions). The propensity to overproject at the other time points is partially due to the different behavioral patterns exhibited by "early" and "late" buyers, the 12-week estimates being dominated by early buyers.

Total Fit

To assess the total fit of the model, the trial function is estimated using (10), and the parameters of the units per transaction functions (μ_o, μ_1) are estimated by ordinary least squares. The value of the maximum transaction size parameter $\bar{U}$, was selected by direct inspection of the raw data. These estimates are displayed in Table 4.

Again, estimation of the stretch factor δ is difficult with only 12 weeks of data. For the case of repeat levels greater than zero, it appeared reasonable to impute a value of δ based on established products. In the case of trial, this procedure cannot be used; one cannot expect the number of persons becoming triers, say in the second half of the first year of a product's life, to bear a fixed relationship to the same statistic for an established product. Given that estimates of δ were not available for other new products, it was necessary to estimate under the assumption $\delta = 0$, realizing that this clearly erroneous assumption should produce underprojection.

The impact of the misspecification can be seen in the estimate of the number of persons purchasing at least once within the first year $R_{52}(0)$. The estimated number of such persons is 92, based on 52 weeks of data, while the estimated number based on the 12-week data is 60. There also appears to be a bias in the parameters of the volume per transaction function where all 4 parameters based on the 12-week data are higher than at the other time points. This also appears to be related to an "early" buyer problem where not only do such buyers repurchase in larger numbers, but they also buy more on each purchase occasion.

Projections based on these estimates are displayed in Table 5. Based on 52 weeks of data, the trial function appears to fit adequately, the model projections never being off from actual by more than 8% at any time point. This, coupled with the repeat fit exhibited in Table 3, results in a good fit for total sales.

Based on the first 12 weeks of data the report is more mixed. Trial units fit well at 12 weeks, then tend to di-

[7] More precisely an extraneous estimate of β_1 is required. These were obtained by estimating (10). These same β_1's were used to calculate the δ values shown in Table 8.

Table 4

ESTIMATED TRIAL AND UNITS PER TRANSACTION PARAMETERS FOR THE NEWPROD PRODUCT, BASED ON THE FIRST 12, 24, AND 52 WEEKS IN TEST MARKET

	Number of weeks included in fit		
	12	*24*	*52*
Parameter of the trial function			
Slope factor γ	.87	.86	.88
Stretch factor δ		1.0	.7
Trial at 1 year $R_{52}(0)$	60	102	92
Parameters of the units per transaction function			
Units purchase on trial occasion U(0)	1.2	1.1	1.1
Units on 1st repeat μ_0	1.4	1.3	1.3
Slope factor μ_1	.2	.1	.1
Maximum units purchased $\bar{U}$	2.0	1.7	1.8

verge from actual. At 52 weeks the understatement is one-third, the result being caused by the incorrect assumption on δ. On the other hand, total repeat transactions fit well at 12, 24, and 36 weeks and diverge from actual by 10% at 52 weeks, this good fit being achieved by repeat parameters that tend to overestimate repeat given trial and a trial function which underestimates the number of people available to repeat. Total repeat sales are overstated due to positive bias in the units per transaction parameters. Happily these "high" repeat sales are offset by "low" trial sales, resulting in rather accurate total sales estimates. One cannot expect such coincidences to be reproducible. On the other hand, better estimates of the trial function should be possible as knowledge accumulates concerning typical values for the trial stretch parameter.

STRUCTURAL FIT FOR SOME ESTABLISHED PRODUCTS

Here the compatibility of the behavioral propositions (stated in the model section) and observed data for six

Table 6

ROOT MEAN SQUARE ERROR[a] BASED ON MODEL FIT FOR SIX ESTABLISHED PRODUCTS

Product Number	*Repeat Level*					
	1	*2*	*3*	*4*	*5*	*6*
1	.005	.012 (.007)	.007	.023 (.014)	.010	.009
2	.020 (.012)	.007	.005	.009	.021 (.011)	.010
3	.011	.014	.010	.011	.026	.010
4	.017 (.012)	.014	.013	.008	.033 (.016)	.011
5	.010	.009	.012	.014	.021 (.013)	.018
6	.013	.017	.011	.028	.026 (.015)	.019

[a] Estimates shown are the CORC Estimates. Whenever the OLSQ differs from the CORC estimates by more than .003, the OLSQ estimates are shown in parentheses.

frequently purchased grocery products is considered. All six products were in national distribution for at least five years and are measured by the Market Research Corporation of America diary panel. The analysis centers on the value of the repeat curves $RI_{it}(J)$ over time (t being measured in weeks). To avoid the truncation problem (exhibited in Table 1), a five-year time slice is used, repeat curves for various repeat levels being constructed based on all repurchases that accrue within one year of a previous purchase. These series are then summed over all time of entry groups, yielding:

$$RI._T(J) = \sum_{i=t-T=1}^{260-T} \overset{*}{RI}_{it}(J) \tag{15}$$

for $T = 0, 1 \cdots 52,$
$J = 1, 2 \cdots 14.$

The "*" on RI is to indicate that the raw data is first

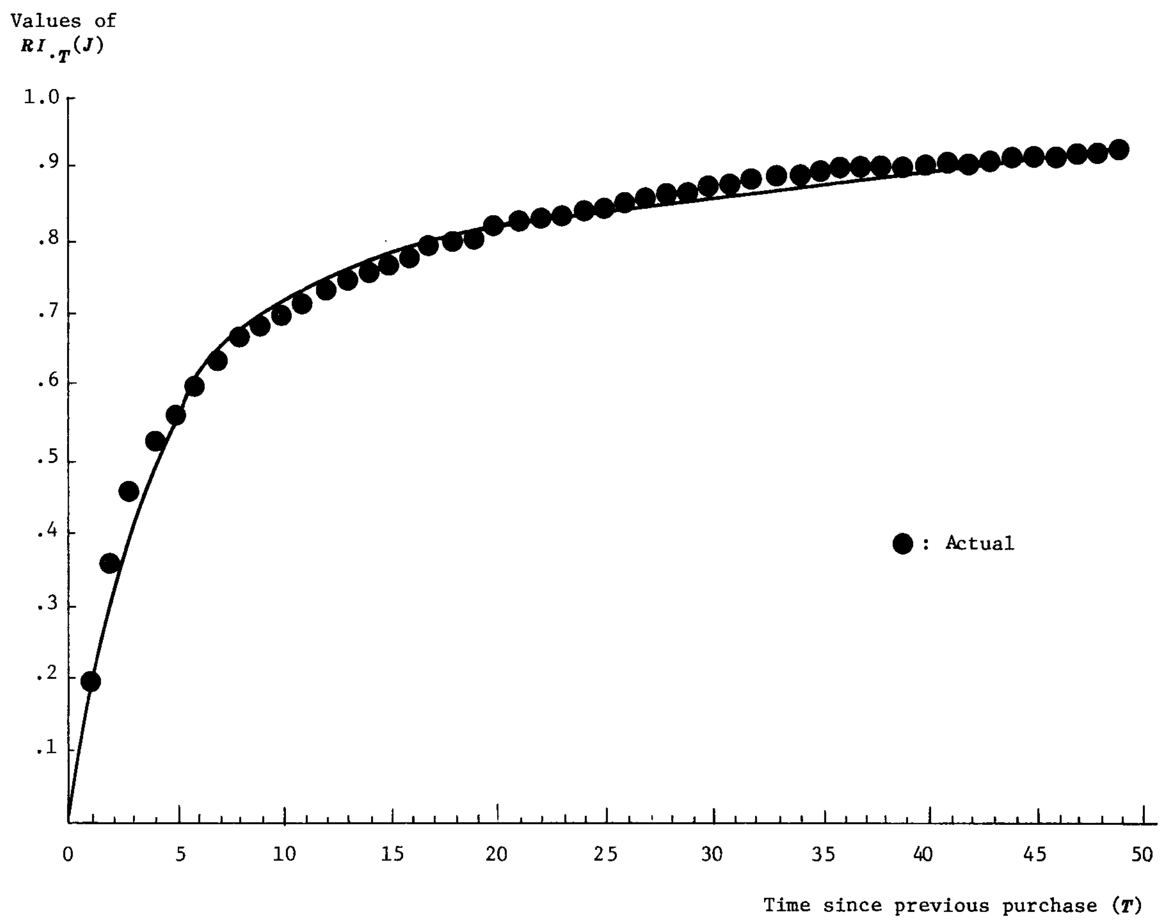
Table 5

COMPARISON OF ACTUAL AND ESTIMATED SALES AND TRANSACTION COUNTS FOR VARIOUS ESTIMATION AND FORECAST PERIODS

	Prediction to 12 weeks				*Prediction to 24 weeks*				*Prediction to 36 weeks*				*Prediction to 52 weeks*			
	Actual	*Fit period 12*	*24*	*52*	*Actual*	*Fit period 12*	*24*	*52*	*Actual*	*Fit period 12*	*24*	*52*	*Actual*	*Fit period 12*	*24*	*52*
Total sales	131	132	122	111	217	227	212	203	310	303	298	291	418	404	428	430
Trial sales	58	58	59	57	81	69	80	77	95	72	94	88	103	72	112	101
Repeat sales	73	74	63	54	136	158	132	126	215	231	204	203	315	332	316	329
Trial transactions	50	48	54	52	71	58	73	70	86	59	85	80	92	60	101	92
Repeat transactions	46	46	46	40	87	93	89	86	135	133	137	135	207	185	209	212

Figure 7

"TYPICAL" PLOT OF ACTUAL VS. ESTIMATED REPEAT PROPORTION[a]



[a]Data shown is for Product 3, Repeat level 5.

summed, then proportioned, rather than vice versa. In other words, the data consist of the cumulative proportion of consumers who have purchased at least J times within T weeks of entry into the $J - 1$ repeat class, this without regard to time of entry into the $J - 1$ repeat class, such curves being produced for repeat value of up to 14, for repurchases within 52 weeks of a previous purchase and for 6 products. There then are a total of 84 repeat purchase sequences of 52 points each under analysis. Primary consideration will be given to the first six repeat levels, this being a cutoff point such that the sample size for all sequences is at least 100 consumers.

Estimation

As mentioned in the estimation section earlier, there is a reasonable supposition that autocorrelation may be present in a data series obtained by calculations as in (15), and that the presence of this may have serious consequences.

The residuals from each of the regression runs on each product and for each of the first six repeat levels were tested for autocorrelation using the Durbin test. These Durbin tests strongly suggest that a positive, first order autocorrelation disturbance could in fact be present in our data. Out of the 36 Durbin statistics calculated, 21 values exceeded the number 2. In the absence of autocorrelation the asymptotic distribution of the Durbin statistic is normal with mean zero and variance equal to one. Thus the above result is improbable under the hypothesis that autocorrelation is absent.

Given these findings, it appears prudent to use estimation procedures which account for autocorrelative disturbances. Here we use the Cochrane-Orcutt variation of the generalized least squares procedures. Whenever Ordinary Least Squares differ substantially from the

Table 7

ESTIMATED VALUES OF SLOPE FACTOR γ^a FOR SIX ESTABLISHED PRODUCTS

Repeat Level	*Product Number*					
	1	*2*	*3*	*4*	*5*	*6*
1	.76	.82 (.74)	.80	.81 (.76)	.83	.81
2	.84	.74	.80	.82	.85 (.80)	.74
3	.82	.76	.82	.78	.87	.78
4	.83	.78	.82	.82	.86 (.79)	.75
5	.81	.82	.83	.84	.84 (.79)	.76
6	.83	.80	.80	.81	.85 (.80)	.76
Average for 1st 6 repeat levels	.81	.79	.82	.81	.85 (.80)	.70
Average for 1st 14 repeat levels	.80	.81	.82	.81	.84 (.77)	.77

[a] Whenever the OLSQ differs from the CORC estimates by more than .04, the OLSQ estimates are shown in parentheses.

Table 8

ESTIMATED VALUES OF STRETCH FACTOR δ^a FOR SIX ESTABLISHED PRODUCTS

Repeat Level	*Product Number*					
	1	*2*	*3*	*4*	*5*	*6*
1	.0031	.0037	.0044	.0030	.0051	.0038
2	.0047	.0051	.0047	.0031 (.0035)	.0048 (.0055)	.0047
3	.0047	.0051	.0052	.0042	.0045 (.0055)	.0045
4	.0043	.0055	.0054	.0042	.0037 (.0050)	.0045
5	.0049	.0047 (.0043)	.0045	.0029 (.0036)	.0041 (.0050)	.0041
6	.0049	.0065 (.0061)	.0054	.0041	.0032 (.0043)	.0041
Average for 1st 6 repeat levels	.0044	.0051	.0049	.0036	.0042	.0043
Average for 1st 14 repeat levels	.0038	.0042	.0045	.0036	.0034	.0033

[a] Whenever the OLSQ differs from the CORC estimates by more than .0003, the OLSQ estimates are shown in parentheses.

CORC estimates, the OLSQ estimates will be noted parenthetically.

Behavioral Propositions

Behavioral Assumption 1 asserts that the Geometric-Stretch model will describe well values of the *RI* function over products and repeat levels. A method of amassing evidence on this issue consists of estimating the function over various products and repeat levels using the methods discussed in the estimation section. If the model is "correct," then (*i*) some coefficients of the model (i.e., β_o, β_1, and/or γ) should be large compared to their error terms and (*ii*) the model should fit the data well. Often (*ii*) is measured by the coefficient of determination R^2. In the present instance R^2 is a misleading measure tending to overstate the degree of fit.[8] Here we take as a measure of general fit a sum of squares based on the difference between actual values of *RI* and the values obtaining by estimating (10), then regenerating *RI* via (11), (12), and (3). This is the Root Mean Square Error (RMSE) calculation exhibited in Table 6. Here it can be seen that 29 of the 36 values are below or equal to .02 (where $0 \leqq RI \leqq 1$). The general degree of fit can also be seen by reference to Figure 7, which is a plot of estimated vs. actual, for one repeat level on one product; the RMSE for this particular product and repeat level is the median value out of the 36 values shown in Table 6. Values of the *t* statistics also tend to support the assertion of good fit with all 84 value of t_γ and 74 of the 84 value of t_{β_1} equaling or exceeding 2. The average value of t_γ is 50, while the average t_{β_1} value is 8. For the reader interested in R^2, the average value for this statistic over all 84 equations is .98.

Estimated values for γ are displayed in Table 7. Table 8 displays the estimated values of δ. Our reading of Table 7 is that γ does not fluctuate excessively nor does it exhibit a strong trend. Support for the latter assertion is contained in a comparison of the average values of γ for the first 6 repeat levels and the average for the first 14 repeat levels. Here it can be seen that these averages are quite close. The estimated values of δ also tend to vary in a relatively small range but do tend to exhibit a negative trend over *J*. This can be most clearly seen by again comparing the 6 and 14 level averages. The existence of trends in δ of this magnitude have little effect on the distribution of number of repeaters $R_t(J)$ for short periods (say $t \leqq 52$ weeks). This is because δ plays only a strong role for low values of *J*.

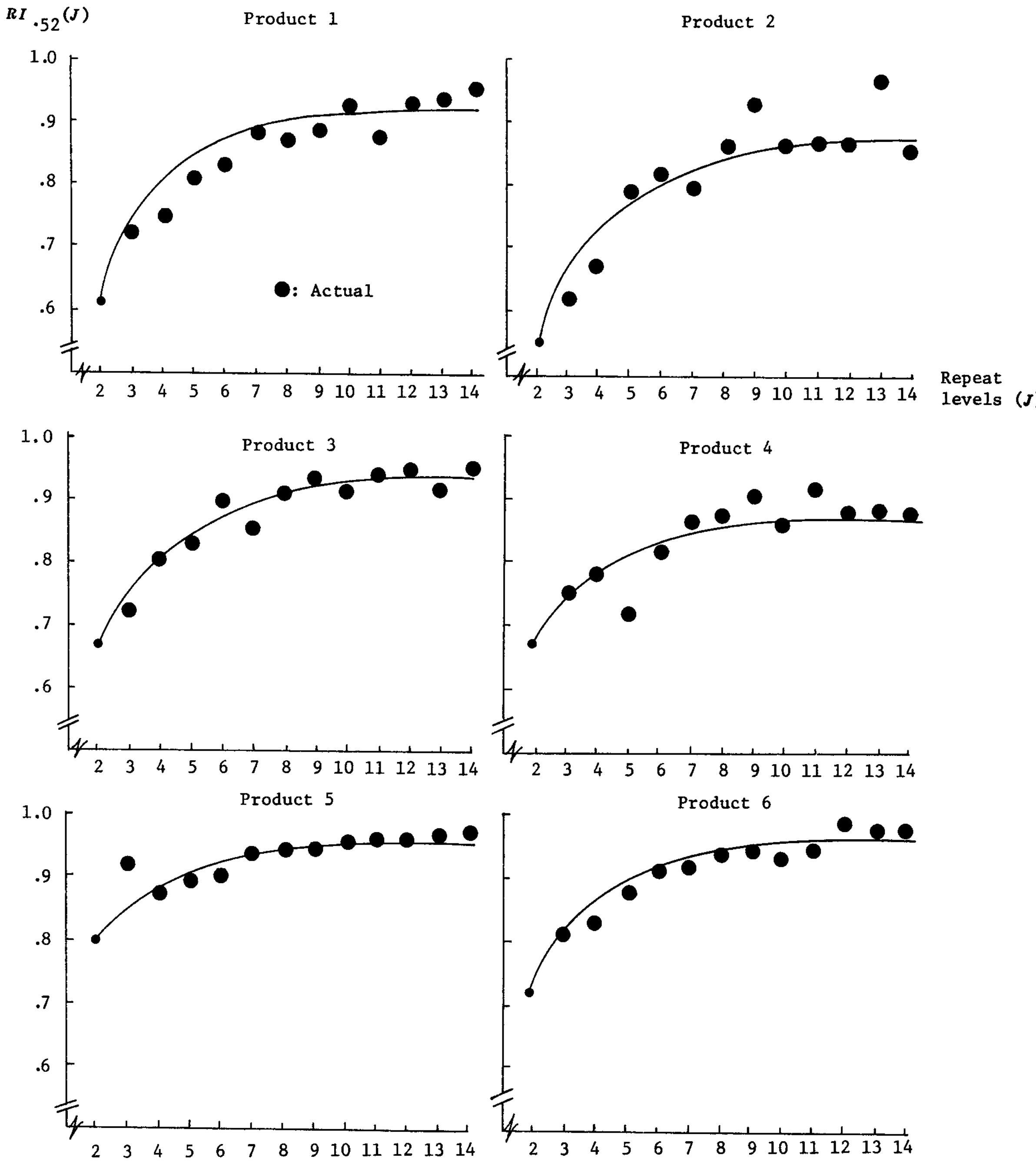
In order to evaluate the consistency of Behavioral Assumption 3 and the data for the 6 established products, the end point values $RI_{.52}(J)$ (i.e., the proportion repurchasing within one year of entry into the $J-1$ class) are regressed over repeat levels *J* as in (10), but without the linear term in *T*. A comparison of estimated vs. actual based on this operation is displayed in Figure 8.

It is worth noting that although conversion to an equilibrium repurchase proportion is relatively rapid, the equilibrium conversion proportions $RI_{.52}(\infty)$, although high, does not quite reach 1.

[8] Overstatement in R^2 is due to the monotone nondecreasing nature of RI which precludes $R^2 = 0$ (even for a purely random process) and the difference in error terms as measured by (3) vs. (10).

Figure 8

ACTUAL VS. ESTIMATED FRACTION REPURCHASING WITHIN ONE YEAR OF A PREVIOUS PURCHASE



REFERENCES

1. Ahl, David H. "New Product Forecasting Using Consumer Panels," *Journal of Marketing Research,* 7 (May 1970), 160–7.
2. Chatfield, C., A. S. C. Ehrenberg, and G. J. Goodhardt. "Progress on a Simplified Model of Stationary Purchasing Behavior," *Journal of the Royal Statistical Association,* Series A, Part 3, 129 (1966), 317–67.
3. Durbin, J. "Testing for Serial Correlation in Least Squares Regression When Some of the Regressors Are Lagged Variables," *Econometrica,* 38 (May 1970), 410–21.
4. ——— and G. S. Watson. "Exact Tests of Serial Correlation Using Non-circular Statistics," *Annals of Mathematical Statistics,* 22 (September 1951), 446–51.
5. Ehrenberg, A. S. C. and G. J. Goodhardt. "A Comparison of American and British Repeat-Buying Habits," *Journal of Marketing Research,* 5 (February 1968), 29–33.

6. ———. "Conditional Trend Analysis: A Breakdown by Initial Purchasing Level," *Journal of Marketing Research,* 4 (May 1967), 155–62.
7. Fourt, Louis A. and Joseph W. Woodlock. "Early Prediction of Market Success for New Grocery Products," *Journal of Marketing,* 25 (October 1960), 31–8.
8. Grahn, Gary L. "The Negative Binomial Distribution Model of Repeat-Purchase Loyalty: An Empirical Investigation," *Journal of Marketing Research,* 6 (February 1969), 72–8.
9. Herniter, Jerome D. "Notes on the Negative Binomial Distribution and Purchase Behavior," working paper, Marketing Science Institute, 1–17.
10. Malinvaud, E. *Statistical Methods of Econometrics,* second revised edition. Amsterdam: North Holland, 1970.
11. Massy, William F. "Forecasting the Demand for New Convenience Products," *Journal of Marketing Research,* 6 (November 1969), 405–12.
12. ———. "Stochastic Models for Monitoring New Product Adoptions," in Frank Bass, Charles King, and Edgar Pessemier, eds., *Applications of the Sciences in Marketing Management.* New York: John Wiley & Sons, 1968, 85–111.
13. ———, David B. Montgomery, and Donald G. Morrison. *Stochastic Models of Buying Behavior.* Cambridge, Mass.: The M.I.T. Press, 1970.
14. Orcutt, Guy H. and D. Cochrane. "A Sampling Study of the Merits of Autoregressive and Reduced Form Transformations in Regression Analysis," *Journal of the American Statistical Association,* 44 (1949), 356–72.
15. Parfitt, J. H. and B. J. K. Collins. "Use of Consumer Panels for Brand-Share Prediction," *Journal of Marketing Research,* 5 (May 1968), 131–45.