부의 이항분포를 이용한 비정태적(Nonstationary) 구매빈도의 단기예측
(분류 : 마케팅, 구매발생 모형, NBD)

박흥수
연세대학교 상경대학 경영학과 부교수(마케팅 전공)
김동훈
연세대학교 상경대학 경영학과 부교수(마케팅 전공)

부의 이항분포(Negative Binomial Distribution, NBD)는 구매 과정이 정태적(stationary)인 경우에서의 구매 빈도에 대한 확률모형(stochastic model)으로 널리 사용되어져 왔다. 본 논문에서는 구매행동의 비정태적인 현상을 모형화하는데 있어서 과거 1기의 구매기록만 사용하는 것이 아니라 과거의 모든 구매기록을 사용하여 NBD의 모수를 추정해 내는 방법을 제시하고자 한다.
과거 t 기간동안의 구매기록을 이용한 NBD 조건기대치를 추정하는 방법으로서, 예측치와 실제치를 혼합하여 주는 최적 가중치 산정절차를 제시한다. 이러한 절차를 거쳐서 조건기대치와 비구매자 비율이 추정되면 Morrison(1969b)이 제안한 수열 추정 방법을 이용하여 비정태적 NBD의 모수를 산출한 후, 이 NBD로 평균 구매자 수와 비구매확률을 예측하고 그 결과로서 구매횟수에 따르는 구매확률들이 주어진다. 마지막으로 본 논문에서 제시된 모형으로부터의 예측치와 정태성을 가정한 모형으로부터의 예측치를 비교한다.

I. 서 언

부의 이항분포(Negative Binomial Distribution ; NBD)는 구매행동을 나타내는 모형으로써 널리 사용되어져 왔다(Chatfield 1969; Grahn 1969; Ehrenberg 1972; Frisbie 1980 참조). 어떤 제품군에서의 개인의 구매횟수가 포아송 과정(Poisson process)을 따르며 개인의 구매율의 차이는 감마(Gamma)분포에 의하여 포착되는 확률혼합모형(probability mixture model)이 NBD이다.

NBD모형은 구체적으로 어떤 기간 중에 한 개인의 구매횟수가 포아송분포를 따르고 포아송 분포의 모수 λ는 모수 α와 β를 갖는 감마분포를 따른다면, 특정 기간동안의 구매 빈도에 대한 분포는 NBD를 따른다는 것이다(Ehrenberg 1959). 그런데 NBD의 정태적(stationary) 속성 때문에 미래 구매에 대한 예측은 단일 기간을 활용하여 미래 구매빈도를 예측하여 왔다. 본 논문에서는 미래 구매를 예측하는데 있어서 단순히 현재의 구매 뿐만 아니라 과거 여러 기간동안의 구매 자료를 동시에 모두 고려하는 방법을 제시함으로서 구매의 비정태성(nonstationarity)을 도입하여 구매빈도가 변화하는 상황을 모형화하고자 한다.

NBD의 모수를 추정하는데 실무적으로 편리한 방법인 평균구매횟수와 비구매자 확률로 모수를 추

논문 접수일 : 96. 9 게재확정일 : 97. 1
본 논문에 대한 많은 도움말을 주신 두분의 익명의 심사자에 대해 감사드립니다.

정하여 왔는데 이 경우 이론적인 분산이 높아지는 경향이 있었다(Ehrenberg 1959; Chatfield, Ehrenberg and Goodhardt 1966). 이와 같은 문제를 해결하기 위하여 베타-이항분포 모형(Beta-Binomial Distribution, BBD)(Chatfield and Goodhardt, 1970)이 연구되어 왔다. BBD모형은 개인의 구매가 주어진 n기간(예: n주)동안에 k번 구매하는 이항분포를 하고 이항분포의 모수가 베타분포를 취한다는 것이다. 이 밖에도 구매빈도를 설명하기 위하여 대수정규분포(Lognormal distribution)(Lawrence, 1980)와 지수수열분포(Logarithmic Series Distribution) (Chatfield, Ehrenberg and Goodhardt, 1966)가 연구되었으며 구매자의 구매시간 간격관점에서 어랑분포(Erlang Distribution) (Chatfield and Goodhardt, 1973) 등 다양한 분포들이 연구되어 왔다.

사고발생 빈도에 대한 모형으로서 NBD를 이용한 Greenwood & Yule(1920)의 연구 이후로 NBD는 마케팅 분야에서 구매 빈도(Ehrenberg 1959), 재구매 애호도(Grahn 1969) 등을 설명하는데 많이 활용되어 왔다. Morrison(1969a)의 연구에서는 혼합분포(mixture distribution)가 $\lambda=0$인 점에서 정점을 갖게 하였고, Sehmittlein & Morrison(1983)의 연구에서는 압축(condensed) NBD를 제시하여, 이 모형의 조건기대구매빈도를 도출하였다. Goodhart & Ehrenberg(1967)의 연구에서 사용된 자료를 이용하여 Sehmittlein, Bemmaor, & Morrison(1985)은 현재 구매를 조건으로 미래 구매를 예측하는 데에 NBD가 유용하다는 것을 보여주었으며 정태적 NBD 모형으로부터의 구매예측치와 비교할 수 있는 벤치마킹을 기준으로 제시하였다. 그런데 이 연구에서 NBD 모형으로부터 조건기대치를 도출해 내는 데에는 정태성(stationarity)에 대한 가정이 현실적으로 적합하지 않은 경우가 많기 때문에, 비정태성(non-stationarity)을 추가함으로써 NBD 모형을 보다 확장하고 일반화시킬 필요성이 대두되었다.

본 연구에서 시도하는 NBD 모형의 확장은 두 가지 방향으로 이루어지고 있다. 첫째는 지난 한기의 구매만을 고려하는게 아니라, 과거 모든 기간동안의 구매기록을 이용하는 예측기법을 제안하는 것이다. 예를 들어 Sehmittlein, Bemmaor, & Morrison(1985)의 연구의 NBD 모형으로부터 도출되는 중요한 관리적 개념은 1기의 구매가 χ_1이라면 2기의 구매는 무엇인가($E(X_2|X_1=\chi_1)$)라는 질문에 대한 해답이었다. 그러나 만약 직전의 구매자료 뿐만 아니라, 과거 여러 기간동안의 구매기록이 있었다면 이를 이용하여 어떻게 위와 같은 조건기대값을 구할 수 있을까? 즉 $E(X_2|X_1=\chi_1, X_0=\chi_0)$의 값은 무엇이며, 과연 $E(X_2|X_1=\chi_1)$와 동일한 값을 가질 것인가? 뒤에서 설명되듯, 그렇지는 않다. 직관적으로 생각해 보아도 과거의 구매량과 구매패턴은 미래의 구매와 연관이 될 것으로 가상할 수 있다. 구매패턴을 무시하는 기존의 방법에서는 X_0을 무시하기 때문에 다음 기의 구매를 예측하는데 있어서 문제점을 야기시킬 가능성을 포함하고 있다. 모형확장의 둘째 방향은 과거 여러 기간의 구매자료를 이용하여 차기의 구매를 예측하는 것뿐만 아니라 미래 n번째 기의 구매를 예측하는 방법을 제시하는 것이다.

본 논문의 구성은 다음과 같다. 다음 장에서는 과거 t-1 기간동안의 구매가 $X_1, X_2 \cdots X_{t-1}$이라는 조건하에서 X_t에 대한 확률모형으로서의 NBD 확률혼합모형(probability mixture model)을 제시하여 이로부터 조건기대값을 추정하는 방법론을 제시

하며 3장에서는 이를 일반화하여 t기까지의 자료를 가지고 t+n 번째 기의 구매를 예측하는 방법을 제시한다. 4장에서는 실제자료를 이용하여 이 모형의 예측능력을 평가하며, 마지막으로 5장에서는 향후 연구방향을 제시한다.

2. 모 형

비정태적 NBD모형을 설명하기 위하여 본 논문에서 활용되는 부호들을 먼저 설명하고자 한다.

부호(Notation)

t : 기간

X_t : t기의 구매횟수(확률변수)로서 포아송분포를 가정한다.

Λ_t : 일 개인의 t기의 평균구매횟수(확률변수)로서 개인간의 구매횟수의 차이를 나타내는데 감마 분포를 나타내는 것으로 가정한다.

α_t , β_t : 확률변수 Λ_t가 감마분포를 나타낼 때 감마분포의 모수로서 식(1)에서는 식(2)와의 차 이를 나타내기 위하여 a_t, b_t로 나타냈으며 식 (2)이후에서는 αt와 βt로 사용되었다.

$\Gamma(\cdot)$: 감마함수

$\bar{x}_t$: t기의 평균 구매횟수

P_t^0 : t기의 비구매자 비율 $= P(X_t=0)$

ω_{t+1} : t+1기의 평균구매자수를 추정하기 위하여 활용되는 가중치

D_t : t기의 평균구매횟수의 관찰치와 예측치의 차이 $= \bar{x}_t - \alpha_t\beta_t$

$\widetilde{P_t^0}$: t기의 비구매자 비율 예측치

υ_{t+1} : t+1기의 비구매자 비율을 추정하기 위하여 활용되는 가중치

C_t : t기의 비구매자 비율의 관찰치와 예측치의 차이 $= P_t^0 - \widetilde{P_t^0}$

s_t : t기에 이르기까지의 평균구매횟수의 평균비율

과거 1기부터 t-1기까지의 구매기록이 주어졌을 때 t기 구매빈도의 수는 다음의 정태적인 NBD로 표현될 수 있다.

$$
\begin{aligned}
&P(X_t = x \mid X_1, X_2, \cdots, X_{t-1}) \\
&\quad = P(X_t = x \mid X_{t-1}) = \int P(X_t = x \mid \Lambda_t) \\
&\qquad \cdot P(\Lambda_t = \lambda \mid X_{t-1}) d\lambda \\
&\quad = \int \frac{e^{-\lambda}\lambda^x}{x!}\left[\frac{1}{b_t^{a_t}\Gamma(a_t)}\lambda^{a_t-1}e^{-\lambda/b_t}\right]d\lambda \\
&\quad = \frac{\Gamma(x+a_t)}{\Gamma(x+1)\Gamma(a_t)}\left[\frac{b_t}{(1+b_t)}\right]^x(1+b_t)^{a_t} \quad (1) \\
&\qquad\qquad t=2,\ 3,\ \cdots \\
&\qquad\qquad x=0,\ 1,\ 2,\ \cdots
\end{aligned}
$$

여기서 $\Lambda_t=\lambda$는 구매빈도를 나타내는 포아송분포의 평균값이며, 개인간의 구매율의 차이를 표현하기 위하여 Λ_t는 (a_t, b_t)의 모수를 가지는 감마 분포에 따른다.

실제 구매행동을 관찰하면 정태적인 경우가 극히

드물 것이다. 예를 들어 새로 시장에 나온 저카페인 커피(decaffeinated coffee)를 생각해 보자. 커피시장의 전체규모는 정태적인데도 불구하고 새로운 형태의 커피제품인 저카페인 커피의 구매는 늘고(성장추세) 있을 수도 있을 것이다. 따라서 이러한 경우 전체 커피시장의 구매를 예측하고자 할 경우 정태성을 가정한 NBD가 적합할지도 모르지만, 저카페인 커피의 구매를 예측하는 데는 정태적인 NBD가 적합하지 않을 것이다. 만일 이러한 성장추세(즉, 비정태성)를 고려한다면 식 (1)의 첫 번째 단계는 적합하지 않게 되며 다음과 같이 바뀌어야 한다.

$$
\begin{aligned}
P(X_t = x \mid X_1, X_2, \cdots, X_{t-1}) &= \int P(X_t = x \mid \Lambda_t) \\
&\cdot P(\Lambda_t = \lambda \mid X_1, X_2, \cdots, X_{t-1}) d\lambda \\
&= \int \frac{e^{-\lambda}\lambda^x}{x!}\left[\frac{1}{\beta_t^{\alpha_t}\Gamma(\alpha_t)}\lambda^{\alpha_t-1}e^{-\lambda/\beta_t}\right]d\lambda \\
&= \frac{\Gamma(x+\alpha_t)}{\Gamma(x+1)\Gamma(\alpha_t)}\left[\frac{\beta_t}{(1+\beta_t)}\right]^x(1+\beta_t)^{\alpha_t} \quad (2) \\
&\qquad t = 2, 3, \cdots \\
&\qquad x = 0, 1, 2, \cdots
\end{aligned}
$$

식 (1)과 식 (2)의 최종단계가 유사하게 보이지만 중요한 차이점이 존재한다. 식 (1)에서는 NBD의 모수가 t-1기의 구매기록 하나에 의해 추정되지만 식 (2)에서의 NBD 모수는 과거 t-1 기간동안의 모든 구매자료에 의해 추정된다는 것이다. 이 추정방법을 설명하고자 한다.

비정태적인 시계열은 평균과 분산의 변화가 다양한 형태로 비정태적인 모습을 보일 수 있다 (Chatfield,1980). 일반적으로 실무자 입장에 쉽게 활용되어온 NBD모형은 구매자의 평균구매횟수와 비구매자의 확률을 활용하여 NBD모형의 모수 α_t와 β_t를 추정하여 왔다. NBD의 실무적인 활용에 따라서 모수추정에 활용되는 평균구매횟수와 비구매자 확률이 시기의 경과에 따라서 달라진다면 전기의 자료만을 활용하는데 그치지 않고 구매빈도의 비정태성을 포착할 수 있도록 과거의 평균구매횟수와 비구매자 확률에 관한 과거 자료들을 종합적으로 활용하여야 한다.

먼저 t기에 이르는 자료들의 평균구매횟수를 활용하여 t+1기의 평균구매횟수를 예측하는 과정을 토론하고자 한다. 자료에 의해 관찰된 평균 구매와 비구매자 비율을 각각 $\overline{x_t}$와 P_t^0라고 표기하자. 그러면 1기부터 t-1기까지의 구매자료를 조건으로 하는 X_t에 대한 예측치와 t기의 관찰된 평균치를 혼합하여 X_{t+1}에 대한 예측치를 도출하는 절차는 다음과 같이 주어진다.

$$
\begin{aligned}
E(X_{t+1} \mid X_1, X_2, \cdots, X_t) &= \alpha_{t+1}\beta_{t+1} \\
&= (1-\omega_{t+1})E(X_t \mid X_1, X_2, \\
&\quad \cdots, X_{t-1}) + \omega_{t+1}\overline{x}_t \qquad (3)
\end{aligned}
$$

여기서 ω_{t+1}는 추정되어야 하는 가중치이다. 위의 식을 순차적으로 적용시키면 다음과 같은 도출이 가능하다.

$$
\begin{aligned}
&E(X_{t+1} \mid X_1, X_2, \cdots, X_t) \\
&\quad = (1-\omega_{t+1})\alpha_t\beta_t + \omega_{t+1}\,\bar{x}_t \\
&\quad = (1-\omega_{t+1})[(1-\omega_t)\alpha_{t-1}\beta_{t-1} + \omega_t\,\bar{x}_{t-1}] + \omega_{t+1}\,\bar{x}_t \\
&\quad = (1-\omega_{t+1})(1-\omega_t)\alpha_{t-1}\beta_{t-1} + (1-\omega_{t+1})\omega_t\,\bar{x}_{t-1} + \omega_t\,\bar{x}_t \\
&\quad = \quad \vdots \\
&\quad = \alpha_{t-2}\beta_{t-2}\prod_{i=0}^{2}(1-\omega_{t-i+1}) + \sum_{j=0}^{2}\left[\omega_{t+1-j}\,\bar{x}_{t-j}\prod_{i=0}^{j-1}(1-\omega_{t+1-j})\right] \\
&\quad = \quad \vdots \\
&\quad = \alpha_2\beta_2\prod_{i=0}^{t-2}(1-\omega_{t-i+1}) + \sum_{j=o}^{t-2}\left[\omega_{t-j+1}\,\bar{x}_{t-j}\prod_{i=0}^{j-1}(1-\omega_{t-i+1})\right] \qquad (4) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad t = 2,\ 3,\ \cdots
\end{aligned}
$$

앞 식(3)에서의 결과는 위의 식(4)에서 나타난 바와 같이 $X_1, X_2, \cdots X_t$가 자료로서 주어진 상황에서 예측치는 기본적으로 평균구매 (식 (4)의 $\bar{x}_{t-j}$)와 가중치들로 구성되어 있다. 즉 가중치는 다음의 식 (5)에 의해서 도출될 수 있으며 1기부터 t기까지의 평균 구매자료가 모두 활용되고 있다.

위의 가중치는 가중 평방오차의 합을 최소화하는 방법으로 추정되어질 수 있다. t기와 t-1기에 있어서의 평균에 대한 예측치와 관찰된 평균간의 차이가 각각 $(\bar{x}_t - \alpha_t\beta_t)$와 $(\bar{x}_{t-1} - \alpha_{t-1}\beta_{t-1})$로 주어지기 때문에 가중치는 다음 문제에 대한 해를 구함으로써 추정될 수 있다.

$$
\begin{aligned}
&\underset{\omega_{t+1}}{\text{Min}}\ [\{(1-\omega_{t+1})(\bar{x}_{t-1} - \alpha_{t-1}\beta_{t-1})\}^2 \\
&+ \{\omega_{t+1}(\bar{x}_t - \alpha_t\beta_t)\}^2] \qquad (5) \\
&\text{s.t.} \\
&\qquad 0 \le \omega_{t+1} \le 1.
\end{aligned}
$$

$\bar{x}_t - \alpha_t\beta_t = D_t$ 라고 하면 위의 식은 다음과 같이 표시될 수 있다.

$$
\begin{aligned}
&\underset{\omega_{t+1}}{\text{Min}}\ [(D_t^2 + D_{t-1}^2)\omega_{t+1}^2 - 2D_{t-1}^2\omega_{t+1} + D_{t-1}^2] \\
&\text{s.t.} \qquad (6) \\
&\qquad 0 \le \omega_{t+1} \le 1.
\end{aligned}
$$

이 식은 오목함수이기 때문에 1차 미분에 의해

다음과 같은 최적해를 도출할 수 있다.

$$\omega_{t+1}=\frac{D_{t-1}^2}{D_t^2+D_{t-1}^2}. \qquad (7)$$

위 식을 관찰하여 보면, 전기의 자료에 의한 예측치의 오차가 클수록 그의 가중치는 작아지는 반면 샘플평균의 가중치가 커짐을 알 수 있다. 따라서 가중치의 산정절차는 습득을 통하여 과거 예측치를 수정하여 준다. 식 (4)에 가중치 식 (7)을 대입하면 다음식을 도출할 수 있다.

$$\begin{aligned}&E(X_{t+1}\mid X_1, X_2, \cdots, X_t)\\&=\alpha_2\beta_2\prod_{i=0}^{t-2}\left[\frac{D_{t-i}^2}{D_{t-i}^2+D_{t-i-1}^2}\right]\\&+\sum_{j=0}^{t-2}\left[\frac{D_{t-j-1}^2}{D_{t-j}^2+D_{t-j-1}^2}\right.\\&\left.\overline{X}_{t-j}\prod_{i=0}^{j-1}\frac{D_{t-i}^2}{D_{t-i}^2+D_{t-i-1}^2}\right]. \qquad (8)\\&\qquad t=2, 3, \cdots\end{aligned}$$

나아가서 이와 유사한 절차를 거쳐서(1,…,t기의 자료를 조건으로 하는) t+1기의 비구매자비율을 추정할 수 있다. t기의 비구매자비율 예측치와 실제치를 각각 $\widehat{P}_t^0$, P_t^0라고 하자. 그러면

$$\begin{aligned}&P(X_{t+1}=0\mid X_1, X_2, \cdots, X_t)\\&=\left(\frac{1}{1+\beta_{t+1}}\right)^{\alpha_{t+1}}\\&=v_{t+1}P_t^0+(1-v_{t+1})\widehat{P}_t^0, \text{ 이며 여기서, } (9)\end{aligned}$$

v_{t+1}은 추정해야 할 가중치이다. 위를 순차적으로 적용시키면

$$\begin{aligned}&P(X_{t+1}=0\mid X_1, X_2, \cdots, X_t)\\&=\left[\frac{1}{1+\beta_{t+1}}\right]^{\alpha_{t+1}}\\&=(1-v_{t+1})\widehat{P}_t^0+v_{t+1}P_t^0\\&=\widehat{P}_2^0\prod_{i=1}^{t-2}(1-v_{t-i+1})+\\&\sum_{j=1}^{t-2}\left[v_{t-j+1}P_{t-j}^0\prod_{i=0}^{j-1}(1-v_{t-i+1})\right] \qquad (10)\\&\qquad t=2, 3, \cdots\end{aligned}$$

을 얻을 수 있다.

최적가중치는 비구매자비율의 예측치와 실제치의 가중평방오차를 최소화함으로써 앞의 식 (7)과 유사하게 도출되어진다. 그 결과 최적가중치는 다음과 같이 주어진다.

$$\begin{aligned}v_{t+1}&=(P_{t-1}^0-\widehat{P}_{t-1}^0)^2/\\&[(P_{t-1}^0-\widehat{P}_{t-1})^2+(P_t^0-\widehat{P}^0)^2].\end{aligned} \qquad (11)$$

이렇게 도출된 가중치를 식 (10)에 대입하면

$$\begin{aligned}&P(X_{t+1}=0\mid X_1, X_2, \cdots, X_t)\\&=\widehat{P}_2^0\prod_{i=1}^{t-2}\left[\frac{C_{t-j-1}^2}{C_{t-j-1}^2+C_{t-j}^2}\right]\\&+\sum_{j=1}^{t-2}\left[\frac{C_{t-j-1}^2}{C_{t-j-1}^2+C_{t-j}^2}\right.\\&\left.P_{t-j}^0\prod_{i=0}^{j-1}\left(\frac{C_{t-i}^2}{C_{t-i-1}^2+C_{t-i}^2}\right)\right] \qquad (12)\\&t=2, 3, \cdots\end{aligned}$$

이 되며 여기서 $C_t = P_t^0 - \widehat{P}_t^0$이다.

이상과 같은 추정절차에 실제자료를 적용시키는데 있어서 초기값을 산정하는 문제가 발생된다. 이를 위하여 1기가 끝난 시점에서 2기를 예측하고 다시 2기가 끝난 시점에서 3기를 예측하는 문제를 설명하고자 한다. 그 결과 초기값이 결정되면 반복적으로 위의 방법을 적용시킬 수 있을 것이다.

2기를 예측하는데는 1기에 대한 자료밖에 존재하지 않기 때문에 가중치 추정이 필요하지 않고 단순히 NBD 모형을 이용하여

$$E(X_2 \mid X_1) = \alpha_2\beta_2 = \bar{x}_1 \quad \text{와} \tag{13}$$

$$P(X_2=0 \mid X_1) = \frac{1}{(1+\beta_2)^{\alpha_2}} = P_1^0 \quad \text{를 계산한다.} \tag{14}$$

3기를 예측하는 데는 1기와 2기의 자료를 활용하게 된다.

$$E(X_3 \mid X_1, X_2) = \omega_3 \bar{x}_2 + (1-\omega_3)E(X_2 \mid X_1) = \omega_3 \bar{x}_2 + (1-\omega_3)\alpha_2\beta_2 = \omega_3 \bar{x}_2 + (1-\omega_3)\bar{x}_1. \tag{15}$$

그러나 3기의 가중치는 $\alpha_1\beta_1$가 필요하므로 위에서 설명한 가중치 추정절차가 적용될 수 없다. 따라서 일반적으로 사용되는 지수 가중법(exponential weighting)이 적용될 수 없다. 본 연구에서는 ω_3를 임의로 0.5의 값을 갖게 하였다.

2기와 3기의 예측치를 구한 후에는 본 연구에서 제시한 절차를 적용시킬 수가 있으며 그 알고리즘은 다음과 같다.

1단계 : 2기와 3기의 예측치를 위와 같이 도출하고

2단계 : 4기 부터는 식 (7)과 (11)에서와 같이 가중치를 추정한 후 식 (8)과 (12)를 이용하여 조건 평균구매와 비구매자 비율을 도출한다.

조건평균값과 비구매자 비율이 구해지면 수열 추정 방법(Morrison 1969b)을 이용하여 식 (2)에서 제시된 비정태적 NBD 모형의 모수를 추정한다.

3. 단기구매행동 예측

지금까지 t기 까지의 자료를 활용하여 t+1기의 NBD 모수 추정에 관한 내용을 토론하여 왔다. 그렇다면 t기까지의 자료가 주어지고 t+1기가 예측된 상황에서 t+2로 부터 시작되는 단기간의 NBD의 모수를 추정하는 방법에 대하여 설명하고자 한다.

t기까지의 자료에 의해서 t+1기가 α_{t+1}과 β_{t+1}의 모수를 지니고 있는 NBD이라면 t+2기 또한 그 이후의 NBD의 모수를 파악하여 구매빈도를 예측할 수 있다. 즉 전체 구매자의 이질성이 포착되는 Λ_{t+2}는 α_{t+2}와 β_{t+2}의 모수를 갖는 감마분포를 따르게 된다. 여기에서는 $X_1 \cdots\cdots X_t$에 이르는 자료들이 X_{t+1}기의 예측을 거쳐서 활용되기 때문에 다음과 같이 표시될 수 있다.

$$P(X_{t+2}=x \mid X_1, \cdots, X_t)$$
$$= \int P(X_{t+2}=x \mid \Lambda_{t+2})$$
$$P(\Lambda_{t+2}=\lambda \mid X_1, \cdots, X_t)d\lambda$$
$$= \int P(X_{t+2}=x \mid \Lambda_{t+2})$$
$$Gamma_{\lambda}(\alpha_{t+1}, s_t\beta_{t+1})d\lambda$$
$$= \frac{\Gamma(\alpha_{t+2}+x)}{\Gamma(x+1)\Gamma\alpha_{t+2}}\left[\frac{\beta_{t+2}}{(\beta_{t+2}+1)}\right]^{x}(1+\beta_{t+2})^{\alpha_{t+2}} \quad (16)$$

t+1기의 평균구매횟수와 t+2기의 평균구매횟수는 $E[X_{t+2}]=s_t E[X_{t+1}]$와 같이 표시될 수 있다. 즉 t+2기의 평균구매횟수는 t+1기의 구매횟수의 s_t배로 표시할 수 있다. 이 결과 감마분포의 성질에 따라서 감마분포의 형태모수(shape parameta)는 변화가 없게 되고 규모 모수(scale parameter)만 변화하게 된다(Mood, Graybill and Boes, 1974). 이를 표시하면 다음과 같다.

$$\alpha_{t+2}=\alpha_{t+1}, \ \beta_{t+2}=s_t\beta_{t+1} \quad (17)$$

일반적으로 t+n번째 기의 예측분포는 식 (17)의 관계를 반복적으로 적용함으로써 다음과 같이 도출될 수 있다.

$$P(X_{t+n}=x \mid X_1, \cdots, X_t)$$
$$= \int P(X_{t+n}=x \mid \Lambda_{t+n})$$
$$P(\Lambda_{t+n}=\lambda \mid X_1, \cdots, X_t)d\lambda$$
$$= \int P(X_{t+n}=x \mid (s_t)^{n-1}\Lambda_{t+1})$$
$$P((s_t)^{n-1}\Lambda_{t+1}=\lambda \mid X_1, \cdots, X_t)d\lambda$$
$$= \{\Gamma(\alpha_{t+n}+x)/\Gamma(x+1)\Gamma(\alpha_{t+n})\}$$
$$\{\beta_{t+n}/(\beta_{t+n}+1)\}^{x}\{1+\beta_{t+n}\}^{\alpha_{t+n}} \quad (18)$$

여기서

$$\alpha_{t+n}=\alpha_{t+1}$$
$$\beta_{t+n}=\{s_t\}^{n-1}\beta_{t+1}$$
$$\text{for } t=3, 4, \cdots \quad (19)$$
$$\text{for } n=1, 2, \cdots$$

처음 두 번째 기까지의 구매율자료가 우선 주어지면 $s_2=\bar{x}_2/\bar{x}_1$이며 평균기울기인 s_t는 다음과 같이 추정될 수 있다.

$$s_t=\left[\left(\frac{\bar{x}_t}{\bar{x}_{t-1}}\right)+(t-2)s_{t-1}\right]\Big/[t-1] \quad (20)$$

본 장과 이전 장에서 도출된 결과를 요약, 정리하면 〈표1〉과 같다.

〈표 1〉 예측 모형 요약

기간 (t)	활용가능 한자료	중요도 가중치 W	V	단기 예측치 모수	평균구매횟수	비구매자 비율	단기 예측치 기울기 모수
1	자료없음	-	-		-	-	-
2	1기자료	-	-	$\alpha_2,\ \beta_2$	$E[X_2 \mid X_3] = \overline{x}_1 = \alpha_2\beta_2$	$\widehat{P_2^0} = P_1^0 = \{1/(1+\beta_2)\}^{\alpha_2}$	-
3	1-2기 자료	1/2	1/2	$\alpha_3,\ \beta_3$	$E[X_3 \mid X_1, X_2] = \alpha_3\beta_3 = \omega_3 \overline{X}_2 + (1-\omega_3)\overline{X}_1$	$\widehat{P_3^0} = v_3 P_2^0 + (1-v_3)\widehat{P_0^2} = v_3 P_0^2 + (1-v_3)\{1/(1+\beta_2)\}^{\alpha_2}$	$s_2 = \overline{x}_2 / \overline{x}_1$ $\alpha_{3+n} = \alpha_3$ $\beta_{3+n} = \{s_2\}^{n-1}\beta_3$
⋮	⋮	⋮		⋮	⋮ ⋮		⋮ ⋮ ⋮ ⋮
$t+1$ $(t\geq 3)$	1기~t기	ω_{t+1}, v_{t+1}		$\alpha_{t+1}, \beta_{t+1}$	$E[X_{t+1} \mid X_1, X_2, \cdots, X_t] = \alpha_{t+1}\beta_{t+1} = (1-\omega_{t+1})\alpha_t\beta_t + \omega_{t+1}\overline{X}_t$	$P[X_{t+1}=0 \mid X_1, X_2, \cdots, X_t] = \{1/(1+\beta_{t+1})\}^{\alpha_{t+1}} = (1-v_{t+1})\widehat{P_t^0} + v_{t+1}P_t^0$	$s_t = \{\overline{X}_t / \overline{X}_{t-1} + (t-2)s_{t-1}\}/(t-1)$ $\alpha_{t+n} = \alpha_{t+1}$ $\beta_{t+n} = \{s_t\}^{n-1}\beta_{t+1}$

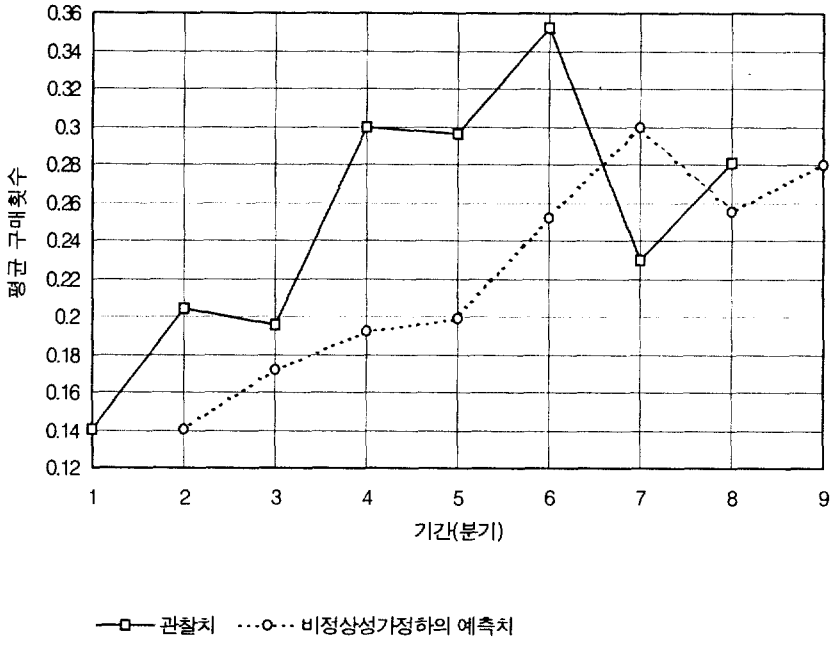
4. 실증분석

분석에 사용된 자료는 Market Research Corporation of America(MRCA)로부터 제공되었으며 이는 1932개의 패널가구의 2년간의 커피 구매자료이다. 커피시장은 저카페인 원두커피(decaffeinated coffee), 저카페인 인스턴트커피, 일반 원두커피, 일반 인스턴트커피 등 4가지 종류로 구분될 수 있다. 본 연구에서는 저카페인 원두커피 자료를 사용하였다.

자료는 2년에 걸친 저카페인 원두커피의 개인별 구매자료를 가지고 13주를 관찰 단위기간으로 설정하였다. 여기에서 4분기 자료를 활용한 것은 월 단위로 자료를 관리할 경우 평균 구매회수가 너무 작아지는 결과를 초래하여 NBD가 너무 좌측으로 기울어지는 현상을 가져 왔다. 따라서 4분기 자료로서 8개의 관찰기간을 가지고서 저카페인 원두커피의 구매빈도를 분석하였다. 평균구매횟수와 비구매자비율(실제와 예측)은 각각 〈그림 1〉과 〈그림 2〉에 제시되어 있다.

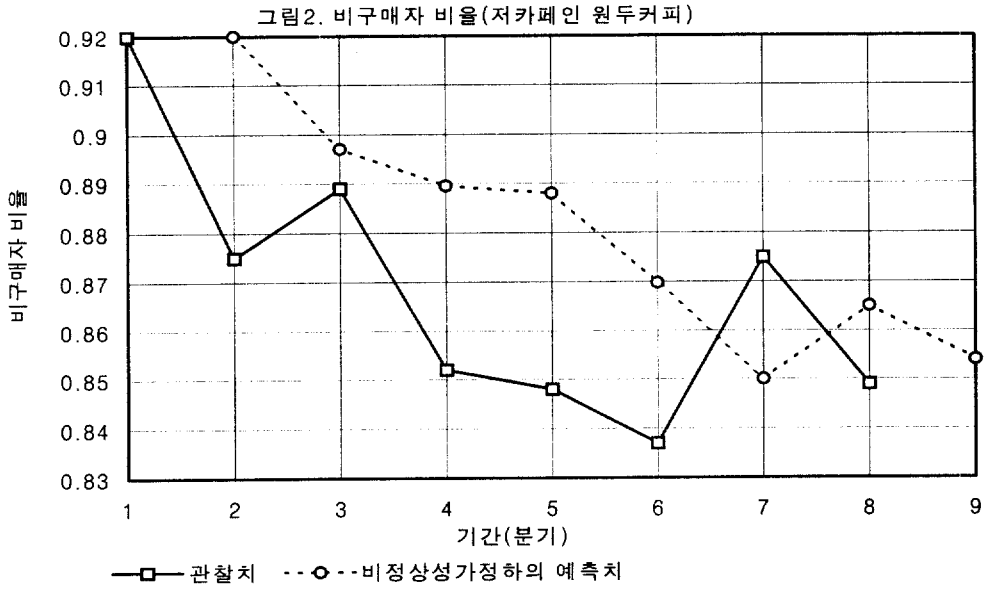
그림에서 보듯이 평균구매횟수는 성장추세를, 그리고 비구매자비율은 감소추세를 나타내고 있다. 즉, 정태성(stationarity) 가정이 맞지 않음을 볼

수 있는 것이다. 따라서 앞서 제시된 방법을 적용하여 저 카페인 원두커피 품목에서의 평균구매횟수와 비구매자비율을 예측하여 보았다. 첫 번째 기에서는 관찰치를 사용하여 제2기를 예측하였으며, 앞에서 설명한 바와 같이 3기부터는 먼저 중요성 가중치를 계산하고, 이를 이용하여 평균구매횟수와 비구매자비율을 조정한다 (식 (8)과 (12)). 이 수치들은 비정태(nonstationary)가정하의 예측치로서 〈표2〉와 〈표3〉에 제시되었다. 예를 들어 표2에서 제6기의 평균구매횟수관찰치는 0.35197로 되어 있으며, 6기의 예측치는 비정태성(nonstationarity) 가정 하에서 0.25176이다. 제6기에 대한 예측을 하는 과정에서 제4기의 예측치의 오차가 제5기의 예측치의 오차보다 컸기 때문에, 5기 예측치에 상대적으로 더 큰 가중치를 주게 되었다. 본 연구에서 제시한 알고리즘에 의해 비정태적 가정하의 예측치를 실제치와 비교해 보면, 6기 이후부터 정확해짐을 알 수 있다. 추가적으로 〈표3〉에서 보면, 6기의 비구매자비율 관찰치는 0.83644이며 예측치는 0.86911로 되어 있다. 마찬가지로 비구매자 비율을 파악할 수 있다. 지금까지는 $X_1 \cdots X_t$까지의 자료를 이용하여 다음기 t+1기의 평균구매횟수와 비구매자 비율을 예측하는 문제를 실증적으로 분석하였다. 다음에서는 t기 까지의 자료와 t+1기의 예측치를 가지고 t+2기 부터 t+5기에 이르는 다기간 예측을 살펴보고자 한다.



<그림 1> 평균 구매횟수(저카페인 원두커피)

그림2. 비구매자 비율(저카페인 원두커피)



〈표 2〉 비정태성 가정하에서의 평균구매 횟수 (저카페인 원두커피)

기간	관찰치	비정상성 가정하의 차기 예측치
1	0.13975	
2	0.20393	0.13975
3	0.19617	0.17184
4	0.30021	0.19311
5	0.29762	0.19837
6	0.35197	0.25176
7	0.22671	0.30138
8	0.28313	0.25337
9		0.27905

〈표 3〉 비정태성 가정하에서의 비구매자 비율 (저카페인 원두커피)

기간	관찰치	비정상성 가정하의 차기 예측치
1	0.91925	
2	0.87526	0.91925
3	0.88923	0.89726
4	0.85197	0.88949
5	0.84834	0.88785
6	0.83644	0.86911
7	0.87578	0.84971
8	0.84834	0.86564
9		0.85363

〈표4〉와 〈표5〉에는 제시된 방법을 적용한 후 예측된 NBD 모형으로부터 도출된 다기(multi-period) 예측치들이 제시되어 있다. 예를 들어 〈표5〉에서 제8기 비구매자비율 관찰치는 0.84834이다. 6기에서(n=2 : t-n=6) 다기예측을 통해 도출된 8기 예측치는 0.83249이며 5기에서의 8기 예측치는 0.83596이고 4기에서의 8기 예측치는 0.82548이고 3기에서의 8기 예측치는 0.83228이다. 단기 혹은 다기예측에 의해서 기별 평균구매횟수와 비구매자비율이 추정되면 구매횟수의 도수분포를 도출할 수 있게 된다. 이러한 분포는 모든 기에 대해 도출될 수 있으나 여기에서는 제6기의 분포를 예로 본다.

제6기 중 1회 구매가 이루어진 관찰도수는 157

〈표 4〉 평균 구매 횟수에 대한 다기예측결과

기간(t)	관찰치	예측치 : (t-n)기에 도출할 수 있는 t기에 대한 예측치			
		n=2	n=3	n=4	n=5
1	0.13975				
2	0.20393				
3	0.19617				
4	0.30021	0.25076			
5	0.29762	0.23378	0.36593		
6	0.35197	0.26129	0.28302	0.53398	
7	0.22671	0.31111	0.34416	0.34262	0.77922
8	0.28313	0.36923	0.38444	0.45332	0.41477
9		0.28587	0.45234	0.47507	0.59710
10		0.31965	0.32254	0.55416	0.58706
11			0.36616	0.36392	0.67891
12				0.41944	0.41060
13					0.48046

〈표 5〉 비구매자 비율에 대한 다기예측결과

기간(t)	관찰치	예측치 : (t-n)기에 도출할 수 있는 t기에 대한 예측치			
		n=2	n=3	n=4	n=5
1	0.91925				
2	0.87256				
3	0.88923				
4	0.85197	0.87107			
5	0.84834	0.87615	0.84227		
6	0.83644	0.86830	0.86210	0.81164	
7	0.87578	0.85289	0.84741	0.84744	0.77995
8	0.84834	0.83249	0.83596	0.82548	0.83228
9		0.85611	0.81464	0.81845	0.80283
10		0.84205	0.84631	0.79630	0.80052
11			0.83013	0.83629	0.77761
12				0.81790	0.82601
13					0.80541

이다. 이러한 도수분포로부터 NBD 모형을 추정하여 기대도수를 계산한 결과 166으로 나타났다(단기예측). 반면 2기에(n=4 ; ∴t-n=2) 다기예측을 통하여 예측한 1회구매도수는 160이며 3기에서의 예측(n=3)은 145이고 4기와 5기에서의 예측은 각각 142와 144로 나타났다. 이러한 수치들의 관리적 가치는 자명하다. 예를 들어 마케팅 관리자가 2기의 시점에서 6기에 대한 유연성 있는 광고예산을 계획하여야 하는 경우, 현재 존재하는 정태적인(stationary) 예측방법으로는 어려우나, 본 연구에서 제시한 방법은 이를 가능하게 해 줄 것이다.

다음의 〈표6〉은 6기에 있어서는 관찰치와 예측치의 구매횟수에 대한 빈도를 보여주고 있다. 여기에서 관찰도수는 6기의 실제자료를 의미하며 기대도수는 6기의 구매빈도와 비구매자율이 각각 0.35197과 0.83644인 상태에서 NBD에 적용시켰을때의 기대되는 빈도이다. 그리고 n=1일때는 5기에서 예측치이며 n=2일때는 4기 n=3일때는 3기에서 각각 앞에서 언급한 방법을 이용하여 예측한 빈도이다.

〈표 6〉 제6기의 관찰 및 예측 돗수분포 (T=6)

구매횟수	관찰도수	단기 기대도수	(6-n)기에 도출된 6기에 대한 예측도수			
			n=1	n=2	n=3	n=4
0	1616.	1616.	1679.	1678.	1666.	1568.
1	157.	166.	144.	142.	145.	160.
2	84.	67.	54.	54.	57.	73.
3	37.	34.	25.	26.	28.	42.
4	15.	19.	13.	14.	15.	26.
5	7.	11.	7.	8.	9.	18.
6	9.	7.	4.	4.	5.	12.
7	3.	4.	2.	3.	3.	9.
8	2.	3.	1.	2.	2.	6.
9	0.	2.	1.	1.	1.	5.
10	0.	1.	0.	1.	1.	3.
11	0.	1.	0.	0.	0.	2.
12	0.	0.	0.	0.	0.	2.
13	0.	0.	0.	0.	0.	1.
14	0.	0.	0.	0.	0.	1.
15	0.	0.	0.	0.	0.	1.
16	0.	0.	0.	0.	0.	1.
17	0.	0.	0.	0.	0.	0.

본 연구가 비정태적 구매행동을 설명하기에 충분한 기간의 자료들이 활용되었어야 할 것이다. 그러나 주어진 2년치의 자료를 활용하여 비정태성을 파악하기 위하여 모형화하여 분석하였는데 특히 커피자체가 생활 필수품이기 때문에 커피의 구매빈도가 정체적이라고 판단되어서 커피를 원두커피와 저카페인 커피로 분류하고 저카페인 원두커피에 대한 비정태성을 파악한 것이다.

이와 같은 상황하에서 앞에서 제시한 방법에 따라서 정태적 모형과 비정태적 모형하에서 7기말에서 예측한 결과를 나타내면 〈표7〉〈표8〉과 같다.

〈표7〉과 〈표8〉은 7기까지의 자료를 활용하여 8기에서부터 평균구매횟수와 비구매자 비율에 대한 정태모형과 비정태모형의 예측치를 보여주고 있다. 미래의 자료가 어떻게 전개될 것이냐에 따라서 두 모형의 타당성을 말해줄 것으로 보인다. 그러나 본 논문에서 주장하는 바는 정태모형일지라도 과거의 여러기간 동안의 자료를 종합하여 미래를 예측하여야 할 것이며 동시에 구매빈도가 비정태적이라고 한다면 더욱 더 과거의 관찰치를 종합하여 구매빈도를 예측하여야 할 것이다. 이와 같은 비정태모형의 구매빈도를 모형화하는 하나의 시도라 볼 수 있다.

〈표 7〉 비구매자 비율에 대한 7기말에서의 예측치 비교

기간	관찰치	정태모형	비정태모형
1	0.91925		
2	0.87526		
3	0.88923		
4	0.85197		
5	0.84834		
6	0.83644		
7	0.87578		
8	0.84834	0.87578	0.86564
9	-	0.87578	0.85611
10	-	0.87578	0.84631
11	-	0.87578	0.83629

〈표 8〉 평균구매 횟수에 대한 7기말에서의 예측치 비교

기간	관찰치	정태모형	비정태모형
1	0.13975		
2	0.20393		
3	0.19617		
4	0.30021		
5	0.29762		
6	0.35197		
7	0.22671		
8	0.28313	0.22671	0.25337
9	-	0.22671	0.28587
10	-	0.22671	0.32254
11	-	0.22671	0.36392

5. 결 론

어떤 새로운 제품군이 시장에 소개되는 경우(예를 들어 저카페인 커피), 이러한 제품의 소비율은 증가하면서 기존제품의 소비율은 감소하는 경우가 많을 것이다. 즉, 소비의 변화는 어떤 추세를 따를 수 있고, 결과적으로 구매행동에도 추세가 나타날 것이다. 이렇듯 구매행동에 증가 혹은 감소추세가 내재되어 있다면, 정태성(stationarity)의 가정은 부정확한 예측을 초래하게 될 것이다. 본 연구에서는 비정태성(nonstationarity)을 가정하는 구매행동 예측기법을 소개하였다. 기존의 예측방법들은 현재의 관찰치만을 토대로 차기에 대한 예측을 하는 반면, 본 논문에서 소개된 방법은 가중치할당기법을 통하여 과거의 모든 자료의 이용이 가능하다. 이 연구의 또다른 기여는 n기 이후의 구매까지도 미리 예측할 수 있는 방법을 제시하였다는 것이다.

NBD모형은 구매빈도를 모형화 하는데 있어서 개인의 구매빈도와 개인간의 구매빈도를 포착하는 확률혼합모형(probability mixture model)으로서 널리 활용되어 왔다. 뿐만 아니라 모형의 단순성과 쉽게 NBD의 모수를 추정할 수 있다는 이점으로 인하여 실무적인 차원에서 또한 널리 사용된 것이 사실이다. 그러나 본 논문에서 제시한 비정태성(nonstationary)모형은 매 기간마다 자료가 입수되는 데로 NBD모형의 모수들을 추정하고 그에 따른 예측을 하여야 하기 때문에 추정과정에 더욱 복잡하게 된 것은 사실이다. 이를 위하여 간편하게 실무적으로 활용하기 위하여 응용프로그램들이 개발되어야 할 것이다.

마지막으로 시계열자료에는 추세, 순환, 계절, 그리고 비규칙 등의 네 가지 요인이 포함되어 있을 수 있다. 소개된 방법은 이 중 추세만을 고려한다. 그러나 만약 주어진 시계열자료에 추세변동 이외에 순환, 혹은 계절변동이 존재한다면 제시된 방법 역시 비정태성을 효과적으로 반영하지 못하며, 이러한 경우 박스-젠킨스(Box-Jenkins)방법과 같은 다른 시계열분석방법과 병행하여 사용되어야 할 것이다.

참 고 문 헌

Chatfield, C. (1969), "On Estimating the Parameters of the Logarithmic Series and Negative Binomial Distributions," *Biometrika*, 56, 2, 411-14.

Chatfield, C. (1980), *The Analysis of Time Series: An Introdution*, New York, Chapman and Hall.

Chatfield, C., A.S.C. Ehrenberg and G.J. Goodhardt (1966), "Progress on a Simplified Model of Stationary Purchasing Behavior," *Journal of Royal Statistical Society*, A, 129, 317-367.

Chatfield, C. and G.J. Goodhardt (1970), "The Beta-Binominal Model for Consumer Purchasing Behavior," Applied Statistics, 19, 3, 240-50.

Chatfield, C. and G.J. Goodhardt (1973), "A Consumer Purchasing Model with Erlang Intre-Purchase Times," *Journal of the American Statistical Association*, 68, 344, 828-835.

Ehrenberg, A. S. C. (1959), "The Patterns of Consumer Purchase," *Applied Statistics*, 8, 26-41.

Ehrenberg, A.S.C. (1972), *Repeat-Buying: Theory and Applications*, North-Holland Publishing Company, Amsterdam.

Frisbie, G. A. Jr. (1980), "Ehrenberg's Negative Binomial Model Applied to Grocery Store Trips," *Journal of Marketing Research*, 17, 385-90.

Grahn, G.L. (1969), "NBD Model of Repeat-Purchase Loyalty: An Emprical Investigation," *Journal of Marketing Research*, 6, 72-79.

Goodhardt, G. J. and A. S. C. Ehrenberg (1967), "Conditional Trend Analysis: A Breakdown by Initial Purchasing Level," *Journal of Marketing Research*, 4, 155-162.

Greenwood, M. and G. U. Yule (1920), "An Enquiry into the Nature of Frequency Distributions Representative of Disease or Repeated Accidents," *Journal of Royal Statistical Society*, A, 83, 255-79.

Lawrence, R. J. (1980), "The Lognormal Distribution of Buying Frequency Rate," *Journal of Marketing Research*, 17, 212-220.

Mood, A. M., F. A. Graybill, and D. C. Boes (1974), *Introduction to the Theory of Statistics*, McGraw-Hill, New York.

Morrison, D. G. (1969a), "Conditional Trend Analysis: A Model that Allows for Nonusers," *Journal of Marketing Research*, 11, 342-46.

Morrison, D. G. (1969b), "A Series Approximation for Negative Binomial Parameter Estimation," *Journal of Marketing Research*, 11, 355-56.

Morrison, D. G. and D. C. Schmittlein, (1981), "Predicting Future Random Events Based on Past Performance," *Management Science*, 27, 1006-1023.

Schmittlein, D. C., A. C. Bemmaor, and D.G. Morrison, (1985), "Why Does the NBD Model Work? Robustness in Representing Product Purchases, Recorded Purchases," *Management Science*, 4, 255-66.

Schmittlein, D. C. and D. G. Morrison, (1983), "Prediction of Future Random Events with the Condensed Negative Binomial Distribution," *Journal of the American Statistical Association*, 78, 449-56.

Short Term Forecasting of Nonstationary Purchase Incidence Using Negative Binomial Distribution

Heung Soo Park, Donghoon Kim*

Abstract

Negative binomial distribution (NBD) has been widely used as a stochastic model of purchase incidence, where the purchase process is assumed to be stationary. This paper proposes a way to incorporate the nonstationarity behavior by providing a method to estimate the parameters of NBD given the purchase history of past periods instead of just one past period.

An optimum weighting scheme which combines the forecasted value and the actual value is the basis for obtaining the NBD conditional expectation given t period purchase history. Once the conditional expectation and the proportion of nonbuyers are estimated using the weighting scheme, the parameters of the nonstationary NBD are calculated using the series approximation method proposed by Morrison(1969b). The NBD so obtained is used to predict the mean number of buyers and the probability of nonbuyers of ground decaffeinated coffee. These values are compared with the values obtained with the stationarity assumption.

*Associate Professor, Dept. of Business Administration, Yonsei University, Seoul, Korea