生存分析在顾客间隔购买时间研究中的应用

王 燕 王 高 赵 平

（清华大学经济管理学院，北京，100084）

摘 要 越来越多的零售商向顾客发放积分卡并积累了众多顾客购买行为的数据。这些数据对企业提高科学决策水平有重要价值，但是目前我国零售企业还很少对这些数据进行开发，主要原因之一就是缺乏方法上的支持。本文以一家购物中心客户积分卡的历史交易数据为基础，使用 Kaplan-Meier 非参数方法估计了顾客间隔购买时间，应用 Cox 比例风险模型对间隔购买时间的影响因素进行了分析。研究表明，间隔购买时间具有明显季节性，交易特征和人口统计特征都对间隔购买时间有显著的影响。

关键词 间隔购买时间 生存分析 Kaplan-Meier 非参数方法 Cox 比例风险模型

1 引 言

目前，许多零售商都实行了购物积分卡计划，这些积分卡详细记录了消费者的购买行为信息。如何利用这些丰富的数据，深入了解消费者购买行为，为零售企业的营销计划提供科学的决策支持就成为了一个重要的研究问题。在消费者购买过程中，有三个问题尤为重要：消费者间隔多长时间去商店购买一次（when），每次购买什么商品（what），每次购买的金额是多少(how much)。本文集中于回答第一个问题。了解消费者何时去商店购买，可以帮助零售企业更准确了解顾客的购买需求，确定补货时间，节约库存成本；了解消费者的购买规律，及时进行相应的促销活动，促使其购买，人为的调整消费者的购买需求。这些对于零售企业都是很重要的。

顾客隔多长时间去一次商店，也称间隔购买时间（inter-purchase timing）。二战之后随着世界经济的复苏和消费水平的提高，对这一问题的关注也开始增多。目前国外研究主要分为两类：一是构建随机模型（stochastic model）估计间隔购买时间，集中寻找合适的分布来拟和间隔购买时间，多是研究一类产品的购买规律；二是构建决定模型（deterministic model），研究间隔购买时间的影响因素。Reinartz 等的研究发现交易特征、人口统计特征对顾客间隔购买时间有显著的影响[6]。但是他们忽视了季节因素（如：购买的月份、购买的星期、是否在节假日等）的影响。而在国内几乎还没有学者对间隔购买时间问题进行研究。

生存分析是当前数理统计中最重要的分支之一，是与寿命、存活时间或者失效有关的数据的统计分析，主要研究事件发生时间的问题，在不同学科中有不同的称谓，社会学中（如人口迁移的时间）叫事件史分析（event analysis），工程研究（如灯泡的寿命）中叫失败分析（failure analysis），经济学（如罢工持续的时间）中叫持续时间分析（duration analysis）。虽然称谓不同，但研究问题的性质都是一致的：分析涉及到一定事件的发生和持续时间长度的数据，用以揭示时间发生和发展的规律。这一方法被广泛应用在医药学、工程学、人口学、金融、保险、质量控制等多个领域，本文以购物积分卡介绍其在零售业中的应用。

本文结构安排如下：首先介绍生存分析方法的原理、估计方法，然后结合一家零售商店的积分卡数据采用生存分析方法中的非参数方法估计顾客间隔购买时间，并构建比例风险模型分别分析交易特征、季节特征、人口统计特征对顾客间隔购买时间的影响，最后指出本文

的不足和未来的研究方向。

2 研究模型

2.1 生存分析方法

与传统的 OLS、logistic 回归分析方法相比，生存分析主要有两个优点：（1）可以很好的处理删失数据，减少估计偏差；（2）可以处理随时间变化的解释变量作用（time-varying variable）。而零售积分卡的数据一般是选取一段观察期，存在数据删失（censoring）现象，即在观察开始之前顾客可能已经发生购买而数据并未被记录，或者观察期结束时顾客仍然会继续购买而我们也无法观测到。而且某些解释变量是随时间变化的。所以用生存分析方法很适合分析零售积分卡数据。

在生存分析的估计中我们把某个事件发生称为死亡。对个体而言，从开始到某事件发生的时间称为或“存活时间”，它是一个随机变量，可以通过四个函数来描述：生存函数 S(t)，概率函数 f(t)，分布函数 F(t)，风险函数 h(t)。四个函数之间的关系如下：

$$F(t)=\Pr(T\le t)=\int_{-\infty}^{t} f(y)dy=\int_{0}^{t} f(y)dy$$

$$S(t)=1-F(t)=\Pr(T\ge t)$$

$$h(t)=\frac{f(t)}{1-F(t)}=\frac{f(t)}{S(t)}$$

根据是否对参数的分布做出假设，常用的生存分析方法包括三类：非参数估计方法，不假设数据的分布，常用的方法有生命表分析和 Kaplan－Meier 方法；参数估计方法，已知数据服从的分布，常用的有威布尔（Weibull）分布，指数（Exponential）分布，对数正态（log normal）分布，对数逻吉斯蒂（log logistic)分布，龚珀兹（Gompertz）分布等；半参数估计方法，使用最多的是 Cox 比例风险模型（Cox proportional hazard model），可以研究各个协变量对生存状况的影响。因为事先不知道分布，而且参数分布的假定对函数的估计有很大影响，所以本文采用 Kaplan－Meier 非参数方法估计平均的间隔购买时间，再运用 Cox 比例风险模型检查不同变量对间隔购买时间的影响。

2.2 Kaplan-Meier 非参数方法

Kaplan-Meier 方法又称乘法极限法（product limit method），常被用来估计生存函数，是一种非参数的估计方法，不需要对理论分布做任何假设。用 $t_1<t_2<\cdots<t_n$ 代表 n 个个体的生存时间，然后用 Kaplan-Meier 方法对其生存函数 S（t）的估计如下：

$$S(t)=\prod_{t_r\le t}\frac{n-r}{n-r+1}$$

其中，r 从所有正整数中取那些 $t_r\le t$ 且 t_r 为无删截的。当所有观察都无删截，r 的值就是连续的整数列，否则 r 的取值就不是连续的。

2.3 Cox 比例风险模型

Cox（1972）提出了 Cox 比例风险模型，它的一般形式为：

$$h_t(t)=h_0(t)\exp(X'\beta)$$

其中 $h_0(t)$ 是基准风险函数，抓住了顾客的内在购买特征；$\exp(X'\beta)$ 是协变量函数，代表了其他变量的影响；X 是一组影响变量，本文中 X 表示交易特征、季节特征和人口统计特征；β 是影响系数。系数 β_i 为正，说明随着自变量的增加，再购买概率增加，顾客间隔购买时间缩短。$\exp(\beta_i)$ 表示相对风险率。对任意两个案例 i 和 j，两者的风险之比是个常数，只取决于两个体的特征，也就是两个体的相对风险率之比，与时间 t 无关。从这个意义上讲，Cox 模型又是成比例的。Cox 提出了用部分似然值法（partial likelihood）来估计生存函数，用似然比卡方检验方法来检验模型的拟和程度。

3 实证研究

3.1 数据

本研究的数据来自北京一家大型购物中心的 116 位积分卡客户 2003 年 6 月到 2005 年 2 月的 39341 条历史购买数据。在这 116 名顾客中，90 名为女性，26 名为男性。在这 21 个月中，每名顾客在这家购物中心平均购买了 40 次产品，最多购买次数为 102 次，最少购买次数为 1 次。每人每次平均的消费金额为 303.5 元。下面运用 SPSS11.5 中生存分析模块对顾客间隔购买时间进行分析。

3.2 间隔购买时间的估计

首先，我们应用 Kaplan-Meier 非参数方法对间隔购买时间进行估计。结果显示，间隔购买时间的均值是 15 天，95%置信区间是（14，15）天。即平均来说这家购物中心的积分卡客户每半个月就会光临一次商店。图 1 是利用 Kaplan-Meier 方法估计的顾客生存函数图，图中的线就是顾客的生存曲线上，每一个点就是具体发生的交易。可以看出，顾客生存的概率一直下降，特别是在 15 天到 65 天这一区间内，递降尤为明显，之后基本是缓慢下降，到 120 天时生存概率基本为零。可以粗略的说如果一名顾客在 120 天内没有购买，他基本就不会再购买了。

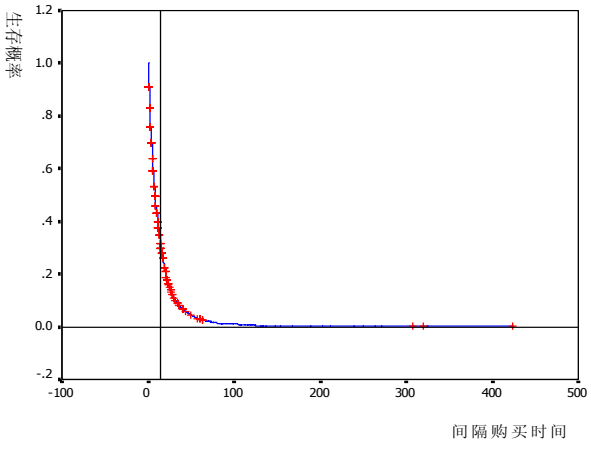

图 1　生存函数图

3.3 间隔购买时间的影响因素的研究

Reinartz[6]等人的研究和积分卡提供的数据，可以把顾客间隔购买时间的因素分为三类：交易特征变量（每次购买的种类、每次购买的金额、累计退货的数量）、季节特征变量（各

个月份的影响、不同节假日的影响）、人口统计变量（性别、年龄、月收入）。接下来，我们运用 Cox 比例风险模型研究各族变量的影响。我们采用嵌套模型方法（nested model method）逐步引入解释变量，比较分析其对间隔购买时间的影响，结果见表 1。三个模型的卡方检验 p 值都在 0.01 的显著性水平上显著，说明每一个模型对间隔购买时间都有显著的解释能力。而且－2log likelihood 逐步减少，说明每一组新加入的变量对于解释间隔购买时间有显著作用。下面分别分析每组变量的作用。

表 1　Cox 比例风险模型的系数估计

自 变 量	模型 1	模型 2	模型 3	模型 3 相对风险率 exp(B)
每次购买产品种类（单位：类）	-0.008＊＊＊	-0.009＊＊＊	-0.01＊＊＊	0.990
累计退货数量（单位：次）	0.125＊＊＊	0.135＊＊＊	0.13＊＊＊	1.139
每次购买总金额（单位：千元）	-0.073＊＊＊	-0.074＊＊＊	-0.072＊＊＊	0.931
购买月份（12 月为参考组）				
1 月		0.123＊	0.118＊	1.125
2 月		-0.306＊＊＊	-0.302＊＊＊	0.739
3 月		-0.115	-0.121	0.886
4 月		0.059	0.054	1.056
5 月		0.015	0.02	1.021
6 月		0.103	0.099	1.104
7 月		0.189＊＊＊	0.181＊＊＊	1.198
8 月		0.002	-0.002	0.998
9 月		0.148＊＊	0.147＊＊	1.158
10 月		-0.028	-0.028	0.972
11 月		-0.016	-0.016	0.984
是否节假日（平时为参考组）				
周末		-0.013	-0.011	0.989
春节（腊月二十三到正月初七）		0.246＊＊＊	0.249＊＊＊	1.282
五一（4 月 30 日到 5 月 7 日）		0.009	-0.002	0.998
十一（9 月 30 日到 10 月 7 日）		0.139	0.143	1.154
元旦（12 月 31 日到 1 月 3 日）		-0.033	-0.038	0.963
性别（男性为参考组）			-0.121＊	0.886
年龄（单位：岁）			0.003＊＊＊	1.003
月收入（1000 元以下为参考组）				
1000－3000 元			-0.08	0.923
3000－5000 元			0.109	1.115
5000－10000 元			-0.183＊	0.833

-2 log likelihood	65963.350	65881.410	65845.650	
卡方值	53.337	134.396	170.920	
自由度	3	25	30	

注：①＊P<0.10， ＊＊P<0.05， ＊＊＊P<0.01；②在周几购买的变量影响不显著，所以未汇报。

3.3.1 交易特征的影响

三个模型中三个交易特征变量都在 0.01 的显著性水平上显著。其中购买产品的种类、购买总金额系数为负，也就是说，一次购买的产品类别越多，购买的总金额越多，再购买的概率就越小，间隔购买时间就越长。大致说来，每次购买量×购买次数＝总需求，而人的需求基本是一定的，所以一次买的越多，购买的次数就越少，两次购买的间隔就长。而累计退货数量的影响为正，也就是说退货次数越多的消费者再购买的概率更大，间隔购买时间越短。这一方面可能因为商场对退货的处理增加消费者的满意度，加强了顾客再购买的概率；另一方面可能是经常购物的消费者买的商品更多，更容易发生退货。

3.3.2 季节因素的影响

研究发现，在一周之内哪天购买不会对间隔购买时间产生影响。但是购买的月份、购买是否是节假日对间隔购买时间有显著影响。五一、十一、元旦、周末与平时无大差异，但春节的再购买率比平时高出 28.8%。许多人在五一、十一时选择旅游而不是逛商场的方式来度过长假，而人们一般不会特意为元旦购买更多商品，再加上时间比较短，所以与平时没有太大差异。在春节这个中国最重要的节日人们一般会置办年货，更频繁的光顾商场，因而春节期间的间隔购买时间更短。抓住春节市场对于零售商是最重要的。

1-12 月中，1、7、9 月显著与 12 月不同，再购买概率比 12 月份分别高出 12.5%、19.8%、15.8%，而 2 月份再购买概率比 12 月低 26.1%。在调查期的一年半中，有两个春节，分别在 1 月底和 2 月初，人们在节前（也就是 1 月份）置办年货，频繁购买，而节后（也就是 2 月份）要消费节前自己储备的、亲戚朋友送的商品，所以间隔购买时间变长，说明一定程度上顾客在春节前发生了提前购买。7 月是北京最热的月份，许多市民把逛商场当作消夏的方式，去商场的概率也相应的高。从图 2 中可以清楚的看出，7 月份是平均购买总金额最少的，但是购买次数明显比前几个月有了大幅提高。而 9 月是夏秋换季，学生开学，购买概率相应增加，间隔购买时间相应缩短。

3.3.3 顾客特征的影响

与男性相比，女性的间隔购买时间更短，逛商场更频繁，这一方面是由于女性更愿意把逛商场作为放松的方式，另一方面整个家庭需要的商品的采购多是由女性（母亲）完成。年龄越大的顾客间隔购买时间越短。年轻人有更多消遣时间的方式，而且一般不会肩负为家庭成员选购商品的责任，所以年轻人的间隔购买时间更长。收入对间隔购买时间也有显著影响，月收入在 5000－10000 元的顾客再购买概率比月收入 1000 元以下的低 16.7%，高收入阶层消费者的时间成本比较高，相应去商场的间隔时间就较长，如果能缩短这些高收入阶层消费者的间隔购买时间，应该可以提高零售商的盈利能力。

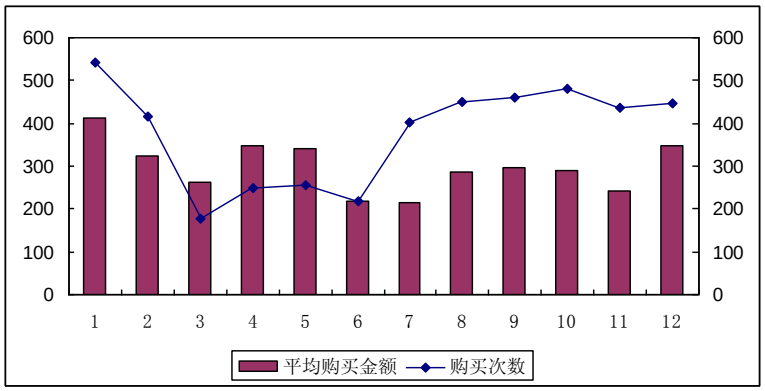

图 2　1－12 月平均购买金额与购买次数线柱图

4　结论及未来研究方向

本研究首先用 Kaplan-Meier 非参数方法估计出顾客的间隔购买时间，然后通过 Cox 比例风险模型分析交易特征、季节特征、人口统计特征对间隔购买时间的影响。研究发现，对于这家购物中心来说，顾客平均 15 天光顾一次商店。每次交易的总额、每次交易购买的商品数越多，间隔购买时间越长，累计退货的次数越多，间隔购买时间越短；季节因素对间隔购买时间有显著影响，尤其是春节和 7 月的炎热对间隔购买时间的影响显著；女性的间隔购买时间更短，年龄越大，间隔购买时间越短，月收入在 5000－10000 元的消费者间隔购买时间更长。

以往的研究没有特别关注季节性的影响，本文研究发现季节性对间隔购买时间有显著影响，零售商可以根据这一发现在不同时期为顾客准备相应商品。比如：在春节前准备充足的年货，在 7 月份准备更多的消暑产品，在 9 月准备学生开学用的商品和秋季换季的商品，避免出现断货现象。而且利用我们的模型可以根据消费者的购买特点预测顾客下次的购买时间，针对不同客户特点在细分市场上促销，如通过短信方式发送促销信息提醒消费者等，以增加顾客的惠顾率。如果发现有些顾客超过平均购买时间还没有发生购买，可以通过手机短信、寄送活动宣传册的方式提醒其购买。

限于积分卡的数据有限，未能包括如促销、广告等营销变量，所以本研究无法分析这些营销变量对间隔购买时间的影响，这是本研究的主要不足之处。在未来对间隔购买时间问题进行研究时，我们认为有以下几个主要方面可以改进：（1）估计个体层面的模型系数，了解不同变量对每一名顾客的影响大小。了解了个人的购买行为规律，可以帮助零售商做好一对一的营销，有的放矢地实行营销策略，我们可以应用贝叶斯方法来达到这个目的；（2）将间隔购买时间与购买金额、购买产品类别综合考虑，这三者是消费者决策过程中一个连贯的总体，是一个问题的三个方面，综合考虑会有更多的结论；（3）增加数据的时间长度，如果要研究季节性规律，我们需要每一个季节有多个数据点，增加时间长度，使得我们关于季节性影响的研究结果更为稳健。

参 考 文 献

[1]Sunil gupta (1991), Stochastic Models of Interpurchase Time With Time-Dependent Covariates. *Journal of Marketing Research*, 28(1):1-15.

[2]Chatfield, C. and G. J. Goodhardt(1973), A Consumer Purchasing Model with Erlang InterPurchase Times. *Journal of the American Statistical Association*, 68(4): 828-835.

[3]Dunn, R., S. Reader and N. Wrigley(1983), An Investigation of the Assumptions of the NBD Model as Applied to Purchasing at Individual Stores. Applied Statistics,32(3): 249-259.

[4]Dipak C. Jain & Naufel Vilcassim (1991), Investigating Household Purchase Timing Decision: a Conditional Hazard Function Approach. *Marketing Science*,10(1): 1-23.

[5]P B Seetharaman, Pradeep K Chintagunta(2003), The proportional hazard model for purchase timing: A comparison of alternative specifications. *Journal of Business & Economic Statistics*, 21(3): 368-382.

[6]Reinartz, Werner and V. Kumar(2003), The Impact of Customer Relationship Characteristics on Profitable Lifetime Duration. *Journal of Marketing*, 67(1): 77–99.

[7]王元月、马驰骋（2005）:《失业保险给付期限差异下的失业持续时间研究》，载于《中国管理科学》第 13 卷第 6 期。

[8]郭志刚（1999）:《社会统计分析方法—SPSS 软件应用》，中国人民大学出版社。

Application of Survival Analysis in the Study of Inter-Purchase Timing

WANG Yan WANG Gao ZHAO Ping

（School of Economics and Management, Tsinghua University, Beijing, 100084）

Abstract More and more retailers issue frequent shopper cards to their customers and accumulate numerous data of customer purchase behavior. The data is of great value for the enterprises to improve scientific decision making, but few retailers in China has used the data. One of the main reasons is the lack of methodological support. Based on the transaction data of frequent shopper card of a shopping center, this article applies the Kaplan-Meier non-parametric method to estimate the inter-purchase timing, and applies the Cox proportional hazard model to analyze the effects on inter-purchase timing of several factors. The results show that the inter-purchase timing has obvious seasonality, the transaction characteristics and demographic characteristics also have significant impacts on inter-purchase timing.

Keywords Inter-purchase timing, Survival analysis, Kaplan-Meier non-parametric method, Cox proportional hazard model