商业银行零售贷款违约概率测算方法探讨

——非线性时变比例违约模型[1]

彭建刚，李樟飞，吕志华，周鸿卫

湖南大学 金融学院，长沙(410079)

Email：pengjiangang@hotmail.com

摘要：本文基于 Cox 模型，针对零售贷款的运行规律和违约因素影响违约行为的特点，提出了非线性时变比例违约模型测算商业银行零售贷款违约概率，该模型考虑了变量之间非线性关系和时间相依变量，使得其更符合客观实际，并通过实证分析和算例分析论证了这种模型在我国商业银行运用的科学性和可行性。

关键词：违约概率；零售贷款；非线性关系；时间相依变量

中图分类号：F832.21　　**文献标识码：**A

1.引言

2007 年爆发的美国次债危机给全球经济造成了重大冲击，至今还在继续。这次危机爆发的根源主要是银行次级住房按揭贷款的高违约率。次级住房按揭贷款是零售贷款的一种，银行如果能准确预测该类贷款的违约概率，那么可以大大降低危机爆发的可能性。我国商业银行现有大量的零售贷款，如何有效识别和度量其信用风险来防止此类危机爆发，是现在我国商业银行面临的一个重要挑战。本课题组（2008）在客户的信用评级基础上，提出了贷款违约表法，解决了线性判别模型、CPV模型因我国企业贷款能收集整理的数据不多而较难准确测算违约概率的问题。同时，贷款违约表法和Altman的死亡率模型相比，考虑了删失数据对违约概率的影响，提高了测算的准确性[1]。但是由于我国商业银行不对零售贷款客户进行信用评级，所以不能用贷款违约表法测算零售贷款违约概率。因此，本文的目的就是为我国商业银行零售贷款违约概率的测量提供一种有效的方法。

目前直接测算零售贷款违约概率模型主要有结构模型。Merton(1974)提出了结构模型理论，并将其应用于估计上市公司的违约概率[2]。Perli and Nayda（2004）运用该理论构建了零售循环贷款结构模型。他们认为当抵押资产价值低于临界值时，这笔贷款就发生了违约[3]。该模型建立在能够实时准确估计出抵押品价值的基础上，然而抵押品往往具有较低的流动性，很难准确估计出其价值，另外，许多零售贷款也并非是抵押贷款，因此，在实际操作中，应用该模型较难精确地测量出客户的违约概率。Andrade and Thomas（2006）试图解决上面问题，构建了零售贷款结构模型。他们提出了客户信誉值的概念，并以银行客户信用评分作为代理变量。当客户信誉值小于某个阀值时，他们就定义该笔贷款违约，然而不同客户的阀值很难估计[4]。

我国贷款抵押品的二手市场善未完善和商业银行没有对零售贷款客户每年进行信用评分，因此，将上述模型不能直接应用于我国商业银行。Cox（1972）最早提出了Cox模型，并提出了模型参数的估计方法。但是在存在时间节点的情况下，其提出的参数估计方法过于复杂[5]。Breslow（1974）提出了简便的估计参数的方法[6]。Narain（1992）最早将Cox模型应用于分析个人贷款的信用风险，建立信用评分模型[7]。本文在信用评分模型的基础上对基

1 本课题得到国家自然科学基金项目(编号: 70673021)和教育部博士点基金项目(编号: 20060532011)资助。

准危险率函数进行了修正，提出了可用于我国商业银行实际操作的非线性时变比例违约模型，并进行了实证分析和算例分析。

2.运用非线性时变比例违约模型测算违约概率的方法

2.1 Cox 模型的基本原理

$h(t)$ 是危险率函数，描述被观察个体在某时刻存活的条件下，在以后的单位时间死亡的概率。Cox(1972)针对 $h(t)$ 提出了 cox 模型，具体模型形式如下：

$$h(t\,|\,z)=h_0(t)\exp(\beta' z) \tag{1}$$

其中 $z=(z^1,z^2,\cdots,z^p)'$ 是协变量，可能是时间相依变量。$\beta=(\beta^1,\beta^2,\cdots,\beta^p)'$，$\beta^j$ 是模型系数。$h_0(t)$ 指 $z=0$ 时的危险率函数。$h(t\,|\,z)$ 指的是个体条件为 z 的危险率函数 [6]。

$S(t)$ 是生存函数，描述观察个体大于某时间存活的概率。当 t 是离散的，$h(t)$ 和 $S(t)$ 有如下关系： $S(t)=\prod_{t_j\le t}[1-h(t_j)]$ （2）

Breslow（1974）提出了估计 β 的简便的方法。我们能在 SAS 中找到该偏似然函数估计程序。他在估计出 β 的值 b 之后，接着使用轮廓似然函数构造方法，得到 $h_0(i)$ 估计值：

$$h_0(i)=\frac{d_i}{W(i;b)} \tag{3}$$

其中：$W(i;b)=\sum_{j\in R_i}\exp(b'Z_j)$，$Z_j$ 是训练样本中的 i 时间区间的影响因素向量，d_i 表示个体在 i 时间区间的死亡数，R_i 表示 i 时间区间初暴露于风险的所有个体的集合。

2.2 将Cox模型拓展为测算零售贷款违约概率方法的前提条件

Cox模型是研究危险率与危险因素之间关系的数学工具。本文基于该模型并且针对零售贷款的运行规律和违约因素影响违约行为的特点，提出了非线性时变比例违约模型测算零售贷款违约概率，其可行性基于以下前提条件：

1）基于零售贷款五级分类的违约界定。当零售贷款出现次级或以下状态（某银行客户分级系统出现棕色预警），本文将其定义为违约。这一定义与巴塞尔新资本协议关于违约的定义范围基本一致。贷款五级分类标准已经在我国商业银行得到了有效的实施，我们能够根据五级分类每月确定各笔贷款的违约情况。

2）变量的多项式形式的引入。一些违约概率对数与违约因素之间关系可能呈非线性，如：本文在实证中发现，借款人随着年龄的增长，其还款能力有弱变强然后变弱。模型如果用线性形式是无法准确拟合这一变化的，那么只能引入非线性形式。本文引入变量的多项式形式，能够使模型挖掘出更多的数据信息。

3）基准危险率函数的修正。Breslow（1974）提出了一个形式较为简单的基准危险率函数，但是如果将该函数形式引入模型直接测算零售贷款的违约概率，那么没有考虑在每个时间区间里提前偿还的贷款。该部分贷款在某个时间区间里可能并没有始终存在，则它们带给该时间区间样本总的风险应该更小。由于在每个时间点上，贷款提前偿还的可能性是相同的，因此，本文假设提前偿还贷款退出样本服从均匀分布，则：

$$PD_0 = h_0(i) = \frac{d_i}{W(i;b) - 0.5 * \sum_{m \in B_i} \exp(b' Z_m)} \tag{4}$$

其中：B_i 提前偿还类贷款的集合。

4）时间相依变量的引入。信用评分模型往往不考虑随时间变化变量对贷款违约行为的影响。对于期限较短的零售贷款，影响其违约行为因素的变化不大，银行利用不随时间变化的变量估计违约概率，简便也不失准确性。但对于期限较长的零售贷款，影响其违约行为因素的变化较大，如果银行还是不考虑随时间相依变量，那么模型估计的违约概率就不准确。但是银行为了节约风险管理成本，往往只对零售贷款债务人进行一次信用评估，那么意味着只能一次得到债务人和该笔贷款的特征。银行既要节约成本，又要随时有效地估计该笔贷款所面临的风险，那么就要去获取容易得到的数据信息。经济变量无非是最容易得到的数据信息，然后模型如果加入一些不随时间变化变量，那么预测违约概率的效果会更好。

2.3 测算零售贷款违约概率方法的具体内容

本文提出的商业银行零售贷款违约概率方法的步骤如下：

1）非线性时变理论模型的建立：按单位时间统计零售贷款生存时间，建立数学理论模型：

$$PD(i \mid z) = PD_0(i) \exp\left(\sum_{j=1}^{p} f(z_j) + \sum_{j=p+1}^{q} f(z_j(i))\right) \tag{5}$$

其中：$f(z_j) = \beta_{j1} z_j + \beta_{j2} z_j^{\ 2} + \cdots$；$f(z_j(i)) = \beta_{j1} z_j(i) + \beta_{j2} z_j^{\ 2}(i) + \cdots$。

当违约因素与违约概率对数之间关系都是线性时，那么模型简化为：

$$PD(i \mid z) = PD_0(i) \exp\left(\sum_{j=1}^{p} \beta_{j1} z_j + \sum_{j=p+1}^{q} \beta_{j1} z_j(i)\right) \tag{6}$$

当违约因素不随时间变化时，那么模型简化为：

$$PD(i \mid z) = PD_0(i) \exp\left(\sum_{j=1}^{p} f(z_j)\right) \tag{7}$$

设零售贷款的期限为 T 个单位时间，那么 $i = 1, 2, \cdots, T$ 。$PD(i \mid z)$ 表示在第 i 个单位时间初存在的零售贷款，在第 i 个单位时间的违约概率，即客户零售贷款的单位时间的条件违约概率，其中：z 是不随时间改变的变量，包括了债务人和贷款的特征（如：债务人学历、性别、家庭成员、过去信用状况、抵押贷款额比、利息类型等）；$z(i)$ 是时间依赖变量（如：贷款在时间 t 的利息、房地产价格、GDP 增长率和失业率等）。

2）贷款持续时间的确定：①违约贷款的持续时间等于贷款发放到违约时的时间，并用 1 作为违约数据的指示器；②提前偿还和到期偿还贷款的持续时间等于贷款发放到偿还时的时间，并用 0 作为没有违约数据的指示器；③统计时间结束时还没有到期贷款的持续时间等于贷款发放到统计结束时的时间，并用 0 作为没有违约数据的指示器。

3） 模型参数的估计：在 SAS 统计软件中首先输入样本中 z 、$z(i)$ 和它们多次方值，然后选择该模型偏似然函数估计方法，得出 β 估计值 b 。同时根据式（4）容易得出：

$$PD_0(i)=\frac{d_i}{W(i;b)-0.5*\sum_{m\in B_i}\exp(b'f(Z_m))} \tag{8}$$

4）累计违约概率和年违约概率的计算：根据式(2)可以得到零售贷款的累积违约概率：

$$\begin{aligned}CPD(i)&=1-S(i)\\&=1-\prod_{j=1}^{j=i}[1-PD(j)]\end{aligned} \tag{9}$$

$CPD(i)$ 表示零售贷款自发放日起至第 i 个单位时间末发生违约的概率

类似于式（9），按月统计零售贷款条件违约概率，若需要确定零售贷款期限 T 内任意一年的年违约概率，则可以运用该年内 12 个月的月条件违约概率 $(PD(i))$ 计算求得：

$$PD_1^m=1-[\prod_{j=m}^{j=m+11}(1-PD(j))] \tag{10}$$

假设违约因素不随时间变化。由于一般 PD_0 值非常小，则式（9）可以表示为：

$$\begin{aligned}CPD(i)&=1-S(i)\\&=1-[\prod_{j=1}^{i}(1-PD_0(j))]^{\exp(\sum_{j=1}^{p}b_{j1}f(z_j))}\end{aligned} \tag{11}$$

式（10）可以表示为：

$$PD_1^m=1-[\prod_{j=m}^{j=m+11}(1-PD_0(j)]^{\exp(\sum_{j=1}^{p}b_{j1}f(z_j))} \tag{12}$$

对于期限较短的零售贷款，如大部分个人消费类贷款，银行可以假定违约因素不随时间变化，然后估计模型参数。对于期限较长的零售贷款，如大部分个人住房贷款，假如商业银行数据库中该类到期贷款的数据非常有限，那么银行可以利用该类贷款已经经历时间的数据估计时间相依变量模型的参数，然后估计该类贷款发放后的前几年违约概率，比如，一批 10 年期个人住房贷款已经发放 5 年，可以利用 5 年数据估计出模型参数，并用其预测同类型贷款前 5 年的违约情况。随着银行数据仓库每年不断收集到期限较长零售贷款的数据，银行可以逐步将这些数据加入到模型参数估计中去。

3.实证分析

3.1 样本选取和变量确定

本文选取了我国某国有商业银行市支行 2005 年 1 月到 2007 年 2 月发放的 2 年期个人汽车消费类贷款 311 笔数据作为样本，其中训练样本 237 笔、检验样本 74 笔。为了保证样本数据的有效性，我们选取的数据信息包括：性别，学历，婚姻，每月客户家庭收入，分期还款额、家庭被抚养人口。另外本文将月收入偿付比作为衡量个人偿还贷款能力的一个指标。

模型变量的系数的估计需要数据信息参数化。按照相关文献的方法，本文将变量参数化的结果如表 1 所示：

表 1　变量参数化的结果

变量	变量简称	变量参数化
年龄	age	$0:age\le 25;1:26\le age\le 35;2:36\le age<50;3:age\ge 50$

学历	qua	小学=5，初中=4，高中或中专=3，大专=2，本科=1，研究生以上=0
性别	sex	男=1，女=0
婚姻	mar	未婚=1，已婚=0
月收入偿付比	Inp	每月客户家庭收入与每月偿还额的比值
家庭被抚养人口	Nump	家庭被抚养人口数目

注：表中一些变量赋值与一般文献赋值不一样，主要考虑到防止基准违约概率出现异常情况。

变量之间的高度相关性会使得模型参数的估计失去准确性，表 2 提供了相关系数不超过 0.5 的相关性矩阵。

表 2　变量相关性矩阵

	sex	qua	mar	age	inp	Nump
sex	1.000	-0.008	0.061	-0.016	-0.149	0.040
qua	-0.008	1.000	0.017	-0.288	-0.088	0.015
mar	0.061	0.017	1.000	-0.281	-0.115	0.246
age	-0.016	-0.288	-0.281	1.000	-0.140	-0.107
inp	-0.149	-0.088	-0.114	-0.140	1.000	0.024
Nump	0.040	0.015	0.246	-0.107	0.024	1.000

3.2 实证结果

3.2.1 模型变量的系数估计结果

本文通过 SAS9.1 运算和剔除模型中不显著变量，得出变量系数的估计结果如表 3 所示：

表 3 模型变量的系数估计结果

变量	参数估计	标准误差	卡方值	P 值	危险率
学历	0.879	0.389	5.114	0.023*	2.408
婚姻	1.217	0.530	5.274	0.021*	3.376
年龄	-2.193	0.776	7.972	0.004*	0.112
年龄^2	0.652	0.248	6.920	0.009*	1.920
收入月偿付比	-1.521	0.621	5.994	0.014*	0.218

注：＊5%情况下显著

从表 3 发现变量与违约概率之间的关系如下：

（1）客户的学历高低与违约概率大小呈反向关系。

学历是受教育水平高低的体现，学历越高往往意味着更稳定的工作，更高的收入，具有更强的偿还贷款能力，所以学历高往往代表着更好的信用。有些银行还针对高学历开发学历贷款品种。因此，学历高的客户比学历低的客户违约概率低。

（2）未婚客户有更高的违约概率。

婚姻往往意味着成熟和责任。与已婚客户相比，未婚客户不需太多考虑整个家庭情况，月偿还压力的大小也可能不会成为他们考虑贷款与否的主要原因。因此，未婚客户更有可能会不理性去贷款，从而导致违约事件的发生。

（3）客户年龄大小与违约概率大小呈先反向后同向关系。

年龄代表了成熟和责任。同时，年长者也可能出现收入的下降。因此，随年龄增长，客户贷款的违约概率先下降后上升。

（4）客户的月收入偿付比大小与违约概率大小呈反向关系。

月收入偿付比是最重要偿还能力变量，当客户有一个比较小的月收入偿付比时，每月用于偿还贷款占收入比就多，一旦出现意外支出时，就可能出现违约情况。因此，月收入偿付

比大的客户比小的客户违约概率低。

由上面估计的系数，式（11）可以表示成：

$$CPD(i)=1-[\prod_{j=1}^{j=i}(1-PD_0(i))]^{\exp(0.879\text{qua}+1.217\text{marry-}2.193\text{age}+0.652age^2-1.521inp)} \qquad (13)$$

式（12)可以表示成：

$$PD_1^m=1-[\prod_{j=m}^{j=m+11}(1-PD_0(i)]^{\exp(0.879\text{qua}+1.217\text{marry-}2.193\text{age}+0.652age^2-1.521inp)} \qquad (14)$$

3.2.2 基准月条件违约概率的估计结果

在估计出模型各个变量的系数之后，我们可以根据式（8），计算出基准月条件违约概率，其结果如表 4 所示：

表 4 基准月条件违约概率（%）

生存时间	1	2	3	4	5	6	7	8
基准月条件违约率（PD_0）	0.00	0.00	0.00	0.00	0.00	0.00	0.00	4.63
生存时间	9	10	11	12	13	14	15	16
基准条件月违约率（PD_0）	2.39	4.85	10.58 (9.98)	0.00	3.13	6.27	6.45	3.49
生存时间	17	18	19	20	21	22	23	24
基准条件月违约率（PD_0）	1.07	3.72	0.00	0.00	0.00	0.00	0.00	0.00

注：括号中的基准月条件违约概率是在没有修正基准危险函数的情况下得到的。

从表 4 的基准月条件违约概率中可以看出，在前 7 个月，基准条件违约概率为 0，这主要因为在刚刚发放贷款时，客户经理和客户本身能有效地估计到偿还贷款的能力。而第 19、20、21、22、23、24 月基准条件违约概率为 0，这主要因为最后几个月偿还的贷款额占整笔贷款额比例很小，客户会想尽办法偿还贷款，以免降低其在银行中的信用值，造成以后不能在银行取得贷款。

3.2.3 模型预测能力的检验

为了检验模型预测的准确性，我们首先利用式（13）或式（14）预测了训练样本每笔贷款发放后 1 年的违约概率，然后计算出违约概率的平均值为 $\overline{PD_1}=3.33\%$。如果没有修正基准违约函数，那么该模型测算的违约概率平均值为 $\overline{PD_1}=3.24\%$，而实际上，该贷款组合中贷款发放后一年的平均违约率为 4.10%。因此，可以看出，修正后的基准违约函数使得模型估计的违约概率更接近于实际观察值，这在一定程度上说明了我们修正后的模型具有更强的预测能力。

4.算例分析

实证分析部分说明了模型能较准确地预测违约概率，其原因之一是在 2 年内，影响违约行为的因素变化不大，这使得在每个时间区影响违为行为的因素与该贷款面临的风险具有一致性，但是对于期限较长贷款，必须考虑时间相依变量才能保持这一致性。

本文模拟了 300 笔 10 年期个人住房贷款数据用于算例分析。这一分析的目的是引入时间相依变量测算个人住房贷款的违约概率，说明本文提出的商业银行贷款违约概率测算方法

的可行性。

影响个人住房贷款违约的因素众多，我们选取了学历、首付、房价涨幅和年龄作为解释变量。学历和首付不随时间变化而变化，而房价涨幅和年龄是随贷款经历的时间区间不同而不同，其中首付和房价涨幅数据乘以 100。

SAS9.1 提供了估计模型系数的偏似然函数。在实际操作中，通过编写 SAS 程序，在每个时间区间里，调用每个随时间可变变量的取值。模型的变量系数估计结果如表 5 所示：

表 5 模型的变量系数估计结果

变量	参数估计	标准误差	卡方值	P 值	危险率
学历	1.096	0.247	19.769	0.000*	2.992
首付	-0.422	0.130	10.501	0.001*	0.655
年龄(i)	-3.504	0.942	13.821	0.000*	0.030
年龄(i)^2	0.995	0.289	11.811	0.000*	2.703
房价涨幅(i)	-0.107	0.024	20.118	0.000*	0.898

注： * 1%情况下显著。

估计个人住房贷款违约概率的模型可以表示成：

$$PD(i) = PD_0(i)\exp(1.096 * qua - 0.422 * sf - 3.504\mathrm{age(i)} + 0.995 age(i)^2 + 0.107 * hprice(i)) \tag{15}$$

在估计出模型各个变量的系数之后，我们假如要估计贷款发放后第 4 年房价涨幅与违约概率之间的关系，那么首先根据式（8）计算出基准年条件违约概率 $PD_0(4) = 15.32\%$，然后根据不同零售贷款的违约因素，计算违约概率，如：当 $qua = 2$，$sf = 7$，$age = 1$ 时，第 4 年房价涨幅与违约概率之间的关系如表 8 所示：

表 6 个人住房贷款违约概率表（%）

hprice（4）	违约概率
10	0.60
5	1.02
-5	2.97
-10	5.07

从表 6 可以看出，随着房价的下跌，该笔贷款的违约概率从 0.60%上升到 5.07%。

5.结论

1）Cox 原模型假定比例危险的自然对数与危险因素呈线性关系，因此，该模型不能体现两者之间的非线性关系。非线性时变比例违约模型引入了违约因素的多项式形式，使得其能适应更复杂的变量之间关系。本文实证分析部分拟合了年龄的二次多项式形式，并且经检验发现，统计量具有显著性；并非所有违约因素与违约概率自然对数之间的关系都是线性的，这对确定我国商业银行构建测算零售贷款违约概率测算方法具有重要的启示。

2）根据商业银行零售贷款数据仓库收集到数据的情况，本文提出了切实可行的零售贷款违约概率测算方法。针对期限较短的贷款，本文不考虑变量随时间变化；针对期限较长的零售贷款，本文提出的方法引入了时间相依变量，使得变量之间关系更符合客观实际。算例分析表明该方法在期限较长零售贷款中应用是可行的。

3）样本中贷款在每个时间区间里可能出现提前偿还情况。本文在考虑了这一情况的基础上，修正了基准危险率函数。实证结果表明这一修正提高了测算违约概率准确度，这有利

于我国商业银行准确地估计零售贷款的风险。

4）本文提出的测算零售贷款违约概率的方法具有普适性。该方法不仅可以应用于测算零售贷款违约概率，还可以为本课题组提出的贷款违约表法中的企业贷款信用等级的准确划分提供有力的工具。银行可以首先将影响企业违约行为的因素输入模型，得到初始违约概率，并根据其划分一个初始信用等级，这比由主观性很强的权重方法得到的信用评级更加科学、客观，接着结合专家意见对初始信用等级进行调整得到最终信用等级，最后利用我们课题组提出的贷款违约表法测算出最终的违约概率，这会大大提高银行测算企业贷款违约概率的准确度。

参考文献

[1] 彭建刚,易宇,李樟飞,商业银行贷款违约概率测算方法探讨：贷款违约表法.管理学报（6）,2008.

[2] R.Merton.On the Pricing of Corporate Debt:the Risk Structure of Interest Rates[J].Journal of Finance,1974(29):449-470.

[3] R. Perli,W. I.Nayda.Economic and regulatory capital allocation for revolving retail exposures[J]. Journal of Banking and Finance，2004（28）:89–809.

[4] F.W. M.Andrade.L.Thomas.Structural models in consumer credit[J]. European Journal of Operational Research,2007（183）:1569-1581.

[5] D.R.Cox.Regression Model and Life Tables(with Disscussion).Journal of the Royal Statistical Society B,1972（30）:187-220.

[6] N. E.Breslow. Covariance Analysis of Censored Survival Data.Biometrics,1974（30）:89-99.

[7] Narain, B. 1992. Survival analysis and the credit granting decision.L. C. Thomas, J. N. Crook, D. B. Edelman, eds. Credit Scoring and Credit Control. OUP, Oxford, U.K., 109–121.

Discussion on Measuring the Retail Loan Default Possibility of Commercial Banks—the Non-linear Proportional Default Model with Time-dependent Variables

PENG Jian-gang ，LI Zhang-fei，LV Zhi-hua，ZHOU Hong-wei

College of Finance,Hunan University,Changsha（410079)

Abstract

In view of the operation rules of retail loans and the characteristics of default factors affecting default Behavior, this paper puts forward Non-linear proportional default model with Time-dependent variables to calculate the retail loan default probability of China's commercial banks based on the Cox model. This model takes into account the non-linear relationship between variables and time-dependent variables，which will be more in line with its objective reality. This paper demonstrates this model is scientific and feasible in practice of China's commercial banks through empirical analysis and example analysis.

Keywords: the probability of default; Retail loans; Non-linear relationship；Time-dependent variables

作者简介：彭建刚，男，湖南长沙人，经济学博士，湖南大学研究院副院长、金融管理研究中心主任，湖南大学金融学院教授、博士生导师，研究方向：金融管理与金融工

程。
李樟飞，男，江西上饶人，硕士研究生，湖南大学金融学院，研究方向：金融管理与金融工程。
吕志华，男，江西抚州人，博士研究生，湖南大学金融学院，研究方向：金融管理与金融工程。
周鸿卫，男，湖南祁东人，经济学博士，湖南大学金融管理研究中心副主任，湖南大学金融学院副教授，研究方向：金融管理与金融工程。