第 12 章　金融数据的计数模型

A. Colin Cameron 和 Dravin K. Trivedi

在一些金融研究中，因变量是一个计数（count），取非负的整数值，这样的例子包括对被兼并企业竞价收购（takeover bids）的次数、未支付的信用分期付款的次数（在信用评级时有用）、事故的次数或事故的索赔次数（在确定保险费时有用）、提前支付的抵押贷款的次数（在为有抵押保证的（mortgage-backed）证券定价时有用）。本文讨论诸如泊松模型和负二项式模型这样的计数模型，特别强调基本计算过程及其在持续期和对偶数据的联系。在金融应用的背景下，本文对标准计数模型的回归技术本身进行讨论。

1.引言

在计数回归中，主要关注的是协变量（covariate）对用非负整数值或数目衡量的一个事件的频率的影响。计数模型，例如泊松模型（Poisson）和负二项式模型（negative binomial）类似于 probit 模型 和 logit 模型这样的二元模型以及其他受限因变量模型——如著名的 tobit 模型——因变量的样本空间都受限。计数模型用于广泛的学科领域。对于早期在经济学上的应用和文献综述，参见 Cameron 和 Trivedi（1986）。对于更近期的发展，参见 Winkelmann(1994)以及 Winkelmann 和 Zimmerman（1995）。关于当代文献的一个完备的综述，参见 Gurmu 和 Trivedi（1994）。

计数数据的基准模型是泊松模型，假如离散随机变量 Y 是一个参数为 λ 的泊松分布，它具有密度函数 $e^{-\lambda}\lambda^{y}/y!$，均值 λ 和方差 λ。许多金融样本的频率、均值和方差见表 1。Jaggia 和 Thosar(1993)关于被兼并企业首次出价后竞价收购的次数的例子说明，在泊松模型的一个典型应用中，小计数占多数。Greene(1994)关于个人信用卡申请者的信用历史的主要信誉受损报告的次数说明有过度离中(overdispersion)，即与要求总体的均值和方差相等的泊松模型相比，样本方差比样本均值大得多，0 过多，因为观测到的 0 计数的比例为 0.804，比预测概率 $e^{-0.456}=0.633$ 大得多。将在下面定义的负二项式分布有适应这种过度离中的潜力。事实上，均值为 0.456、方差为 1.810 的负二项式模型给出的 0 的预测概率是 0.809。一个相关的例子是 Dionne，Antis 和 Guillen（1996）的数据。他们根据一家银行债权人无法收到的分期支付贷款的次数建模。Davutyan(1989)关于每年有多少银行倒闭的数据是一个时间序列，增加了复杂性。由于五个最大的次数是样本后期的最后五个观测值，数据可能是序列相关的。

在计数数据的计量经济学应用中，集中于分析回归元 X 的作用，X 是通过设定 $\lambda=\exp(X'\beta)$ 引入的，其中参数向量 β 可以用最大似然法估计。例如，对被兼并企业竞价的收购次数的平均次数可能与该企业的规模有关。

计数回归和持续期（或等待时间）模型之间有重要的联系。这些联系可从研究事件间的等待时间的基础随机过程来理解，这涉及到状态（state）、时段（spell）和事件（event）三个概念。*状态*是一个个人或一个金融实体在一个时点上的分类。*时段*由状态定义，是进入的时间和退出的时间。*事件*是从一个状态向另一个状态的瞬间转换。

表 1： 一些计数变量的频率

作者	Jaggia-Thosar	Greene	Guillen	Davutyan
计数变量	首次出价后竞价的次数	信誉受损报告次数	信用违约次数	银行倒闭次数
样本容量	126	1319	4691	40
均值	1.738	0.456	1.581	6.343
方差	2.051	1.810	10.018	11.820
计数…				
0	9	1060	3002	0
1	63	137	502	0
2	31	50	187	2
3	12	24	138	7
4	6	17	233	4
5	1	11	160	4
6	2	5	107	4
7	1	6	80	1
8	0	0	59	3
9	0	2	53	5
10	1	1	41	3
11	0	4	28	0
12	0	1	34	0
13	0	0	10	0
14	0	1	13	1
15	0	0	11	0
16	0	0	4	0
>17	0	0	28[a]	5[b]

[a] 大计数是 17（5 次），18（8），19（6），20（3），22（1），24（1），28（1），29（1）30（1），34（1）。

[b] 大计数是 17（1），42（1），48（1），79（1），120（1），138（1）。

*持续期*的回归模型涉及花费在一个特定状态的持续时间（非负的）长度与协变量集合之间的关系。持续期模型经常被改写成*险率*模型(models of hazard rate)。险率是从一个状态向另一个状态转换的瞬间速率。*计数*回归模型涉及到在一个固定时间区间上关注的事件发生的次数和一个协变集合之间的关系。

在实证研究工作中，采用哪一种方法将不仅取决于研究的目标，还取决于可获得的数据的形式。持续期或转换（transition）的经济计量模型为给定金融状态的持续时间建模提供了一个合适的框架；计数模型为每单位时间事件发生的频率建模提供了一个框架。本文与其他文章处理方法的不同之处在于强调计数回归与基础过程之间的联系，以及计数回归与持续期分析的联系。

为了使概念更清晰，以抵押贷款的预付为例，该例涉及从持有一个抵押贷款的状态退出，以及终止相关联的时段(spell)。如果可获得的数据提供了单个抵押贷款全期或非全期的样本信息，对于那些在某天开始或某天结束的抵押的样本信息，加上关于抵押贷款持有人和抵押

契约的特征的数据，持续期回归是分析协变量作用的一个自然的方法。[1] 现在，通常的情形是单个持续期区间的数据可能无法取得，但可能可以取得每某个时间单位里重复发生的频率数据，例如在某个日历时间期内提前支付的抵押贷款次数。这样加总过的数据，加上关于协变量的信息，可以形成计数回归的基础。仍有另一种数据情况——这种情况我们不讨论——是关于二元结果的样本信息，即在某个时间区间内，一笔抵押贷款是否被终止的样本信息，诸如 logit 模型或 probit 模型的二元回归模型是分析这种数据的自然方法。

持续期模型进一步的例子是：从敌意投标兼并一家企业开始到竞争这家公司的控制权结束之间的持续期；花在破产保护的时间；银行倒闭的时间；解散公共交易基金的时间区间；直到偿还借款第一次发生违约的时间。我们已经给出了在实证金融研究文献中计数模型的几个例子。我们重申，对每个例子，很容易想到这些以持续期的形式或计数形式出现的数据。

在第 2 节我们讨论持续期和计数计量经济模型之间的关系。第 3 节在金融应用的背景下，对第 3 节计数数据的回归技术本身进行讨论。第 4 节做总结性的评论。

2．计数数据和持续期数据的随机过程模型

从根本上说，持续期模型和计数模型彼此互为对偶。当基础数据生成过程服从一个平稳的（无记忆的）泊松过程的严格假定时，这种对偶关系最为明显。在这种情况下，很容易证明事件发生的频率服从泊松分布而持续时间服从指数分布。例如，如果被兼并企业竞价收购的次数遵循一个泊松过程，那么企业在一个给定的时间区间内的次数是泊松分布，而在竞价间流逝的时间是指数分布。在这个特殊情形下，持续期和计数的计量经济模型在衡量协变量（外生变量）的影响方面是等价的。

平稳性是一个较强的假定。通常基础的更新过程（underlying renewal process）是有依赖或有记忆的。花在一个状态的时间的长度，例如，自从最后收购出价以来的时间，可能会影响离开那个状态的可能性；或是，一个事件将来发生的频率，可能依赖于同一事件在过去发生的频率。在这些情况下，持续期模型和计数模型的信息内容可能相当不一样。然而，可以发现两种类型的模型都可以提供关于协变量对所关注事件作用的有用信息。本文余下部分主要关注计数模型。

2.1. 预备知识

我们观测一个长度为 t 的区间上的数据，对于非平稳过程，行为可能也依赖于区间的开始点，用 s 表示。特别关注的随机变量是 $N(s,s+t)$——代表在 $(s,s+t]$ 区间事件发生的次数——和 $T(s)$，$T(s)$ 代表给定一个事件在时间 s 发生后，到下一次事件发生之间的持续期。*事件次数*的分布通常用概率密度函数表示：

$$\Pr\{N(s,s+t)=r\},\quad r=0,1,2,\cdots$$

持续期的分布有几种表示方式，包括：

[1] 在抽样时，一个时段可能还在进行中（不完全的）。把这样经审查的（censored）观测值包括进回归分析中是持续期模型的一个主要特征。

$$F_{T(s)}(t) = \Pr\{T(s) < t\}$$

$$S_{T(s)}(t) = \Pr\{T(s) \geq t\}$$

$$f_{T(s)}(t) = \lim_{dt \to 0} \Pr\{t \leq T(s) < t + dt\}$$

$$h_{T(s)}(t) = \lim_{dt \to 0} \Pr\{t \leq T(s) < t + dt \mid T(s) \geq t\}$$

$$H_{T(s)}(t) = \int_s^{s+t} h_{T(s)}(u)du$$

这里函数 F ， S ， f ， h 和 H 分别称为*分布函数，生存函数，密度函数，险函数（hazard function）*和*积分险函数（integrated hazard function）*。

对于持续期，随机变量的分布通常用生存函数和险函数来设定，而不是用更常见的分布函数或密度函数，因为生存函数和险函数有更自然的、物理上解释。尤其是险函数给出了在直到该时刻仍未发生状态转移的情况下，从一个状态向另一状态转移的瞬间速率（或离散情况下的概率），而且险函数通过下式与密度函数、分布函数以及生存函数相联系：

$$h_{T(s)}(t) = \frac{f_{T(s)}(t)}{F_{T(s)}(t)} = \frac{f_{T(s)}(t)}{1 - S_{T(s)}(t)}$$

作为一个例子，考虑企业花费在破产保护下的时间长度。关注的是风险（hazard）如何随时间变化，如何随企业的特征变化。如果险函数随 t 递减，那么，企业处在破产保护的时间越长，脱离破产状态的概率就越低，而假如险函数随企业利息负担的增加而增加，则利息负担高的企业比利息负担低的企业更有可能脱离破产状态。

险函数建模应该把最初状态和目标状态考虑进去。两状态模型是最普遍的，但是在一些情况下，多状态模型在实证研究中可能是合适的。例如，一个现在处于破产保护的企业，可能其后被清偿，也可能继续它原来的运作。这些可能性要求使用三状态模型。

2.2. 泊松过程

定义常数 λ 为事件发生率，假如事件的发生是相互独立的，一个发生率为 λ 的（*纯*）*泊松过程*发生的概率等于 λ 乘以区间的长度。正式的表达是，当 $t \to 0$ 时，

$$\Pr\{N(s, s+t) = 0\} = 1 - \lambda t + o(t)$$

$$\Pr\{N(s, s+t) = 1\} = \lambda t + o(t)$$

而且 $N(s, s+t)$ 与 $(0, s]$ 区间上发生事件的次数和位置在统计上是相互独立的。注意取极限时，2 或更大的次数发生的概率是 0，而 0 和 1 事件发生的概率分别为 $(1 - \lambda t)$ 和 λt 。

对这个过程，可以推导出当 t 不受限制时，事件在区间 $(s, s+t]$ 上发生的次数是一个均值为 λt 的*泊松*分布，概率为

$$\Pr\{N(s,s+t)=r\}=\frac{e^{-\lambda t}(\lambda t)^r}{r!},\quad r=0,1,2,\cdots$$

而事件下一次发生的持续期是均值为 λ^{-1} 的指数分布，密度函数为

$$f_{T(s)}(t)=\lambda e^{-\lambda t}$$

对应的险率 $h_{T(s)}(t)=\lambda$ 是常数且不依赖于最近一次事件发生以来的时间，表现出所谓的泊松过程的无记忆特性。也要注意次数的分布和持续期的分布，都独立于开始的时间 s。

设 $s=0$，考虑单位长度的时间区间，那么 N ——在这个区间的事件的平均次数，其均值由下式给定

$$E[N]=\lambda$$

而 T 的均值——事件之间的持续期，由下式给出

$$E[T]=\frac{1}{\lambda}$$

从直觉上来说，每个时期的事件发生频率高，意味着在事件间的持续期平均来说较短。

通过用协变量 X 对 λ 参数化，例如 $\lambda=\exp(X'\beta)$，可得到回归模型的条件均值函数。可用最大似然估计或（非线性性）回归进行估计。对于泊松过程，为使非线性回归的估计更有效，用 $Var(N)=\lambda$ 或 $Var(T)=(1/\lambda)^2$ 进行估计。

泊松过程可能并不总是数据的合适模型。例如，1 发生的概率可能会增加进一步发生的可能性，那么泊松分布可能过多预测到数字 0，而非 0 的数字预测过少，而且方差超过均值。

2.3. 时间依赖的泊松过程

时间依赖的泊松过程，也称为非同质泊松过程或非平稳泊松过程，是一个非平稳的点过程，通过设定事件发生率依赖于过程开始后流逝的时间，即我们用 $\lambda(s+t)$[2] 代替 λ，将（纯）泊松过程一般化。[2]

则次数 $N(s,s+t)$ 的泊松分布的均值为 $\Lambda(s,s+t)$，这里

$$\Lambda(s,s+t)=\int_s^{s+t}\lambda(u)du$$

持续期 $T(s)$ 的分布的生存函数和密度函数

$$S_{T(s)}(t)=\exp(-\Lambda(s,s+t))$$

[2] 过程从时点 0 开始，而观测到的时间区间从时点 s 开始。

$$f_{T(s)}(t)=\lambda(s+t)\exp(-\Lambda(s,s+t))$$

因为 $h_{T(s)}(t)=\lambda(s+t)$，从而 $\lambda(.)$ 是险函数。同样有 $h_{T(s)}(t)=\Lambda(s,s+t)$，因此 $\Lambda(.)$ 是积分险函数。

函数形式的一个方便的选择是威布尔（Weibull）函数 $\lambda(s+t)=\lambda\gamma(s+t)^{\gamma-1}$，此时 $\Lambda(s,s+t)=\lambda(s+t)^{\gamma}-\lambda s^{\gamma}$。在这种情形下，$\lambda(.)$ 的时间依赖成分以指数为 $\gamma-1$ 相乘进入，参数 γ 代表持续期的依赖；$\gamma>1$ 代表正的持续期依赖，意味着当前状态的时段将终止的概率随着时段长度的增加而增大。$\gamma<1$ 代表负的持续期依赖。在 $(s,s+t]$ 区间上事件的平均次数也依赖于 s，当 $\gamma>1$ 或 $\gamma<1$ 时，随 s 的增加递增或递减。因此，这个过程是不平稳的。$\gamma=1$ 的情形是纯泊松过程，此时，威布尔分布退化为指数分布。持续期计量经济分析的标准参数化模型是威布尔回归模型。通过设定 λ 依赖于回归元——例如 $\lambda=\exp(X'\beta)$——且 γ 不依赖于回归元来构建回归模型。

这里有一个例子是比例险（proportional hazards）或比例强度因素分解；

$$\lambda(t,X,\gamma,\beta)=\lambda_0(t,\gamma)g(X,\beta) \qquad (2.1)$$

其中 $\lambda_0(t,\gamma)$ 是基准险函数，回归元的唯一作用是作为这个基准险函数的一个比例因子。这个因素分解使理解变简单了。因为对于一个 $X=X_1$ 的观测值，它离开这个状态的条件概率是 $\frac{g(X_1,\beta)}{g(X_2,\beta)}$ 乘上当 $X=X_2$ 时离开原状态的条件概率。估计也更简单了，因为回归元的作用可以和险函数随时间变化的方式分离开来。对单时段(single-spell)的持续数据，这是Cox(1972a)的偏似然估计的基础。当观测到的是多时段持续期时，这导致大部分信息来自计数的估计方法，参见 Lawless(1987)。相似的方法可用于分组计数数据，例如，Schwartz 和 Torous(1993)对在一个给定的时间区间上终止的未到期抵押贷款的次数建模。

2.4. 更新过程

*更新过程*是一个平稳的点过程，该过程在事件发生之间的持续期是独立同分布的。（纯）泊松过程是一个更新过程，但是时间依赖的过程不是一个更新过程，因为它不平稳。

对于一个更新过程 $f_{T(s)}(t)=f_{T(s')}(t)$，$\forall s,s'$，去除对 s 的依赖比较方便些。我们把 N_t 定义为在 $(0,t)$ 区间上发生的事件（更新）的次数，这在前面的表示方式是 $N(0,t)$，而且和

$N(s,s+t)$有相同的分布。我们也定义Tr为直到第r次更新的时间。

那么

$$\begin{aligned}\Pr\{N_t = r\} &= \Pr\{N_t < r+1\} - \Pr\{N_t < r\} \\ &= \Pr\{T_{r+1} > t\} - \Pr\{T_r > t\} \\ &= F_r(t) - F_{r+1}(t)\end{aligned}$$

这里Fr是Tr的累积分布函数。

上述等式的第 2 行表明有一种基于（或对偶于）所设定的持续期分布来导出N_t的参数化分布的有吸引力的方法。例如，有人可能想要一个对偶于威布尔分布的计数分布，因为威布尔分布能够潜在地适应一定类型的时间依赖。[3]不幸的是，这种方法在实践中经常不可行。

具体说来，Tr是r个独立同分布的持续期时间之和。用（逆）拉普拉斯变化，对矩生成函数的非负随机变量进行一个修正，相当容易求得Tr的分布。[4]当拉普拉斯变换简单且以闭形式(closed form)存在时，相当容易得到分析的结果。当持续期是独立同指数分布时，正如所预期的那样，N_t是泊松分布。当持续期是独立同厄朗分布（Erlangian distribution）时（厄朗分布是双参数伽马分布（gamma distribution）中第一个参数限制在只可取正整数时出现的特殊情形，参见 Feller（1966）和 Winkelmann（1995）），也可得到分析的结果。对于许多像威布尔这样的标准的持续期时间分布，Tr的分布的解析式不存在，从而N_t的分布的解析式也不存在。从原理上说，可以用数值法，但是迄今为止没有沿着这些路线的研究。

可以得到一些有用的渐近结果。如果事件之间独立同分布的持续期的均值为μ，方差为σ^2，那么随机变量

$$Z = \frac{N_t - t/\mu}{\sigma\sqrt{t/\mu^3}} \overset{\alpha}{\sim} N(0,1)$$

更新的*期望*次数 E[N_t]，称为更新函数，满足

$$E[N_t] = t/\mu + O(1)$$

当$t \to \infty$时，将持续期时间对半分将近似地使更新的平均次数变成两倍。因此，如果长时间观测一个更新过程，计数分析将提供很多关于平均持续期时间的信息。对于一个泊松过程，计数数据包含的信息和平均持续期时间包含的信息是一样的。

---

[3] 对于一个更新的韦布尔过程，发生率由上一个事件（那时事件被更新）以来的时间决定。对于一个时间依赖的韦布尔过程，则发生率是由过程开始以来的时间决定的。

[4] 如果$F(t)$是随机变量 T(T>0)的分布函数，那么 F 的拉普拉斯变换是$L(s) = \int_0^\infty e^{-st} dF(t) = E[e^{-sT}]$。如果$T = t_1 + t_2 + \cdots + t_n$，那么 T 的拉普拉斯变换是$L(s) = \prod_{i=1}^n L_i(s)$。对于任何一个拉普拉斯变换，都有唯一一个概率分布与之对应，在这个意义上说，拉普拉斯变换有唯一性。

更新过程的参数分析一开始是设定独立同分布的持续期的分布。因此，如果可以取得持续期长度的数据，更新过程的参数分析是直截了当的。更新过程的大多数计量经济分析集中于分析时段不完全或经审查（censored）的蕴涵，观测到的数据可能是后向递归时间，即最近的更新到固定时点t的时间长度，或是前向递归时间，即从t到下一个更新之间的时间，但没有完整时段的持续期——后向和前向递归时间之和，参见 Lancaster(1990，第 94 页)。

2.5. 其他随机过程

许多其他的随机过程有应用于金融数据的潜力。随机过程的标准参考书是 Karlin 和 Taylor(1975)。象许多这类参考书一样，它没有考虑从这个理论上产生的统计模型的估计。Cox 的许多专著确实强调了统计应用，包括 Cox 和 Lewis(1966)以及 Cox（1962）。Lancaster(1990，第 86-87 页)推导了泊松过程的标准结果。Lancaster（1990，第 5 章）讲解了一些基本的随机过程理论，强调了更新理论以及该理论与持续期分析的密切关系，也参见 Winkelmann（1994，第 2 章)。

马尔可夫链是随机过程的一个子类，对计数数据建模特别有用。一个*马尔可夫链*是一个马尔可夫过程，即一个在给定当前状态全部知识的情况下，将来的行为不会因增加对过去行为的了解而改变的过程。马尔可夫过程仅在有限或可数的范围内取值，可以用从一个状态（离散值）向另一个状态的转移概率来刻画其特征。假如这些离散的数值是非负的整数，或者可以用新的尺度变成非负的整数值，马尔可夫链为计数描述了一个概率模型。这为计数开启了很大范围的一类模型，因为许多随机过程都是马尔可夫链。一个例子，一个分叉过程（branching process)，在 3.6 节讨论。

3．计数的计量经济模型

泊松回归是计数数据分析共同的出发点，并通过假设计数数据服从泊松过程而得到了很好的推动。然而，计数数据经常表现出重要的“非泊松”特征，包括：

1. 过度离中：条件方差超过条件均值，而泊松分布要求两者相等。
2. 过多的零：0（或其他某个整数值）出现的频率比给定均值的泊松过程所预测的频率高。
3. 左截断（Truncation from left)：小计数（特别是 0）被排除在外。
4. 右审查（censoring from right)：比某个设定的整数值更大的次数被归为一组。

在出现上述任何一个特征时使用泊松回归，都会损失有效性（有时还会损失一致性)，报告的标准误不正确，而且拟合差。这些考虑推动了除泊松之外的其他分布的使用。这些计数模型的设定通常很少考虑基础随机过程。

为了便于参考，表 2 给出了一些使用较普通的分布和它们的矩特性。每个子部分考虑一类计数模型，在考虑它们的应用和随机数据生成过程之前给出。表 3 概括了按文中列出的金融文献和模型顺序排列的例子。

表 2：标准参数计数分布和它们的矩

分布族	密度函数	计数	均值；方差
泊松	$f(y)=\frac{\exp(-\lambda)\cdot\lambda^{y}}{y!}$	$y=0,1,\cdots$	$\lambda;\lambda$
负二项式	$f(y)=\frac{\Gamma(y+v)}{\Gamma(v)\Gamma(y+1)}\left(\frac{v}{\lambda+v}\right)^{v}\left(\frac{\lambda}{\lambda+v}\right)^{y}$	$y=0,1,\cdots$	$\lambda;\lambda+\frac{1}{v}\lambda^{2}$
正计数	$f(y\mid y\geq 0)=\frac{f(y)}{1-F(0)}$	$y=1,2,\cdots$	随 f 变化
槛	$f(y)=f_1(0)$	$y=0$	随 f_1， f_2 变化
	$=\frac{1-f_1(0)}{1-f_2(0)}f_2(y)$	$y=1,2,\cdots$	
带零	$f(y)=f_1(0)+(1-f_1(0))f_2(y)$	$y=0$	随 f_1， f_2 变化
	$=(1-f_1(0))f_2(y)$	$y=1,2,\cdots$	

表 3：金融应用

例子	因变量	模型
1．Jaggia 和 Thosar	按被兼并企业计算的总竞价次数	泊松
2．Davutyan	每年银行倒闭次数	泊松
3．Dionne 和 Vanasse	每人事故次数	负二项式
4．Dean 等	事故索赔次数	泊松－逆高斯
5．Dionne 等	未支付的分期付款贷款的次数	截断的负二项式
6．Greene	信誉受损报告的次数	带零负二项式
7．Bandopadyaya	处于破产保护的时间	经审查的威布尔
8．Jaggia 和 Thosar	直到接受竞价收购的时间	经审查的威布尔－伽马
9．Green 和 Shoven	抵押贷款提前支付的次数	比例险
10．Schwartz 和 Torous	抵押贷款提前支付或违约的次数	分组比例险
11．Hausman 等	股票价格变动	有序的 probit
12．Epps	正规化的股票价格变动	泊松复合事件

3.1. 预备知识

应用研究工作中，典型的数据由 N 个观测值构成，第 i 个观测值是 (y_i, X_i)， $i=1,\cdots,n$，

其中纯量因变量 y_i 是所关注事件发生的次数，而 X_i 被认为是决定 y_i 大小的 $k\times1$ 协变量向量。除非特别指明，我们假设观测值之间是相互独立的。计数 y_i 的计量经济模型在参数上是非线性的。最大似然估计（ML）特别受青睐，虽然也可以使用与最大似然估计关系密切的、基于数据分布前两阶矩的估计方法。

关注的焦点是事件发生的平均次数怎样随着一个或多个回归元的变化而变化，条件均值最常见的设定是：

$$E\{y_i \mid X_i] = \exp(X_i^{'}\beta) \tag{3.1}$$

其中 β 是一个 $k\times1$ 的未知参数向量，这个设定保证了条件均值是非负的，而且由 $E\{y_i \mid X_i]/\partial X_{ij} = \exp(X_i^{'}\beta)\beta_j$，根据 β_j 的符号，保证了条件均值随 X_{ij} 严格单调增（或单调减）。其次，参数可以直接看成半弹性（系数），因为 β_j 给出了 X_{ij} 变化 1 个单位，条件均值变化的比例。最后，假如一个回归系数是另一个回归系数的两倍，那么对应的回归元变化 1 单位的效应是另一个的两倍。至此，我们已给出这个特殊的均值设定的结果。

作为一个例子，令 y_i 为第 i 个被兼并企业首次出价后竞价的次数，S_i 代表企业规模，用以十亿美元为单位的总资产帐面值度量。那么，使用和 Jaggia 及 Thosar(1993)相同的样本，将 y_i 对 S_i 进行泊松回归，产生的一个条件均值为 $E[y_i \mid S_i] = \exp(0.499 + 0.0375S_i)$，因此总资产增加十亿美元，竞价收购次数增加 3.7%。

有时，回归元是取过对数后进入（3.1）的。例如，我们可以有

$$\begin{aligned} E\{y_i \mid X_i] &= \exp(\beta_1 \log_e(X_{1i}) + X_{2i}^{'}\beta_2) \\ &= X_{1i}^{\beta_1}\exp(X_{2i}^{'}\beta_2) \end{aligned} \tag{3.2}$$

在这种情况下，β_1 是弹性。这种形式特别适用于当 X_{1i} 是对诸如汽车事故建模时，车开了多少英里这样的风险暴露（exposure）测度，在这种情况下，我们希望 β_1 接近于 1。

3.2. 泊松模型、负二项式模型和逆高斯模型

3.2.1. 最大似然估计

*泊松回归模型*假设 Y_i 在给定 X_i 时是泊松分布，密度函数为

$$f(y_i \mid X_i) = \frac{e^{-\lambda}\lambda_i^{y_i}}{y_i!},\ y_i = 0,1,2,\cdots \tag{3.3}$$

与（3.1）一样，均值参数 $\lambda_i = \exp(X_i'\beta)$。假定观测值之间相互独立，对数似然函数是

$$\log L = \sum_{i=1}^{n} \left\{ y_i X_i'\beta - \exp(X_i'\beta) - \log y_i! \right\} \qquad (3.4)$$

估计是简单直接的。对数似然函数是全局凹函数，许多统计软件包都有内置的泊松最大似然估计程序，或内置有 Newton-Raphson 算法，可用于反复迭代并再加权的 OLS。一阶条件为

$$\sum_{i=1}^{n} \left(y_i - \exp(X_i'\beta) \right) X_i = 0$$

或未加权的残差 $(y_i - \exp(X_i'\beta))$ 与回归元正交。应用常见的最大似然理论可得到 $\hat{\beta}$ 的渐近正态分布，均值为 β，使用 $E[\partial^2 \log L / \partial\beta\partial\beta'] = -\sum_{i=1}^{n} \exp(X_i'\beta) X_i X_i'$，有

$$Var(\hat{\beta}) = \left(\sum_{i=1}^{n} \exp(X_i'\beta) X_i X_i' \right)^{-1} \qquad (3.5)$$

泊松分布要求方差和均值相等。事实上，观测到的数据通常是过度离中的，即方差大于均值。如果均值的设定正确，即（3.1）成立，那么泊松的最大似然估计仍是一致的，但是它是无效的；报告的标准误是不正确的。[5]

通过对一个设定限制没有泊松那么多的密度函数进行最大似然估计，可以得到更有效率的参数估计。可以适应过度离中的、标准的双参数计数数据分布是均值为 λ_i，方差为 $\lambda_i + \alpha\lambda_i^2$ 的负二项式分布，密度函数为

$$f(y_i \mid X_i) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(y_i + 1)\Gamma(\alpha^{-1})} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \lambda_i} \right)^{\alpha^{-1}} \left(\frac{\lambda_i}{\alpha^{-1} + \lambda_i} \right)^{y_i} \quad y_i = 0,1,2,\ldots \qquad (3.6)$$

与（3，1）相同，均值参数 $\lambda_i = \exp(X_i'\beta)$，对数似然函数等于

$$\log L = \sum_{i=1}^{n} \{ \log \left(\frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(y_i + 1)\Gamma(\alpha^{-1})} \right)$$

$$- (y_i + \alpha^{-1}) \log(1 + \alpha \exp(X_i'\beta) + y_i \log\alpha + y_i X_i'\beta \} \qquad (3.7)$$

负二项式分布随方差函数的不同，有其他可供选择的参数化。上述这一种被 Cameron 和 Trivedi(1986)称为 Negbin2 模型，用 LIMDEP 作为例子计算过。设 $\alpha = 1$，它变成几何分布（Geometric）的一个特例。另一个可供选择的模型被称为 Negbin1 模型，方差为 $(1+\alpha)\lambda_i$，是均值的线性函数而不是二次函数。这个 Negbin1 模型很少有人用，不在此正式讨论。这两种模型都用最大似然法估计，$(\hat{\alpha}, \hat{\beta})$ 是方差阵为信息矩阵之逆的渐近正态。在过度离中参数 α

[5] 这完全类似于假定误差项是正态和同方差，而误差项事实上是非正态分布且异方差，但仍然有均值为零以使得条件均值被正确地设定时，用最大似然法估计进行线性回归的结果。

等于 0 这个特殊情形下，两种模型都退化成泊松模型。

负二项式模型的一个动机是假设 y_i 具有参数 $\lambda_i v_i$ 而不是 λ_i 的泊松分布，其中 v_i 代表不可观测的个体异质性。如果 v_i 的分布是均值为 1，方差为 α 的独立同伽马分布，那么 y_i 对 λ_i 和 v_i 的条件分布是泊松分布，仅对 λ_i 取条件分布时，它是均值为 λ_i 、方差为 $\lambda_i + \alpha\lambda_i^2$ 的负二项式分布（即 Negbin2）。这个用于推导负二项式分布的不可观测的异质性，假定基础随机过程是泊松过程。另一种负二项式分布的推导假定基础随机过程有一种特定形式的非平稳：一个事件的发生会进一步增大以后发生的概率。关于计数的截面数据本身不能充分地将两种假定区别开来。

显然，通过设定 v_i 的不同分布，可以生成一大类的模型，称作*混合*模型。一个这样的模型是 Dean 等人（1989）的*泊松—逆高斯*模型，假定 v_i 是逆高斯分布，这样形成的分布的尾比负二项式分布的尾更厚。没有多少证据表明这个可供选择的混合模型比负二项式模型更优越。

混合模型不能对不够离中(underdispersion)（方差比均值小）的数据建模，但是这个限制不算太大，因为大多数数据是过度离中的，不够离中数据的参数模型包括 Katz 系统，参见 King(1989)，以及广义泊松模型，参见 Consul 和 Famoye（1992）。

当数据是计数形式时，较好的做法是既用泊松模型估计，又用负二项式模型估计。泊松模型是负二项式模型在 $\alpha = 0$ 时的特殊情形。这可以用似然比来检验，用－2 乘上两个模型的拟合对数似然值之差，在没有过度离中的零假设下，其分布是 $\chi^2(1)$。另一种是用 Wald 检验，在负二项式模型中估计 α 并报告“t 统计量”，该 t 统计量在没有过度离中的零假设下具有渐近正态分布。当软件包中没有负二项式回归的程序时，第三种方法特别有吸引力。该方法是估计泊松模型，构建 $\hat{\lambda}_i = \exp(X_i'\hat{\beta})$，进行辅助 OLS 回归（没有常数项）

$$\left\{(y_i - \hat{\lambda}_i)^2 - y_i\right\} / \hat{\lambda} = \alpha\hat{\lambda}_i + u_i \qquad (3.8)$$

它所报告的 α 的 t 统计量在没有过度离中的零假设下，针对具有 Negbin2 形式的过度离中的备择假设，是渐近正态分布。这最后一个检验和泊松对负二项式分布的得分（score）检验或拉格朗日乘数（LM）检验一致，但是更一般化，因为它的动机是基于仅仅使用设定的均值和方差。它对有 Negbin2 形式的过度离中的任何一种备择分布都有效，而且它也可用来检验不够离中，参见 Cameron 和 Trivedi（1990）。为了检验 Negbin1 形式的过度离中，将（3.8）用下式代替

$$\left\{(y_i - \hat{\lambda}_i)^2 - y_i\right\} / \hat{\lambda} = \alpha + u_i \qquad (3.9)$$

3.2.2. 基于一阶矩的估计

到目前为此我们已经讨论了完全参数法（fully parametric approaches），另一种方法源自 Gourieroux，Montfort 和 Trognon(1984)，Cameron 和 Trivedi(1986)以及 McCullagh 和 Nelder(1989)，使用有关一阶矩或一阶和二阶矩信息的回归方法。最简单的方法是假设（3.1）

成立，用无效但一致的泊松最大似然估计法估计 β，用 $\hat{\beta}$ 代表*，并且计算正确的标准误。

如果假设方差是均值的 τ 倍，方差的计算就特别容易

$$Var(y_i \mid X_i) = \tau \exp(X_i^{'}\beta) \tag{3.10}$$

这是 Negbin1 形式的过度离中。那么，对泊松最大似然估计

$$Var(\hat{\beta}) = \tau\left(\sum_{i=1}^{n}\exp(X_i^{'}\beta)X_i X_i^{'}\right)^{-1} \tag{3.11}$$

这样，将那些由标准的泊松软件包报告的结果乘以（或除以）$\sqrt{\hat{\tau}}$，可得到正确的标准误（或 t 统计量），其中

$$\hat{\tau} = \frac{1}{n-k}\sum_{i=1}^{n}\frac{(y_i - \exp(X_i^{'}\hat{\beta}))^2}{\exp(X_i^{'}\hat{\beta})} \tag{3.12}$$

这通常可以从计算机输出的结果中计算，因为它就是简单地将皮尔逊统计量（Pearson statistic）（3.19）除以自由度。例如，如果 $\hat{\tau} = 4$，报告的 t 统计量需要除以 2。

如果方差不是均值的 τ 倍，而是均值的二次方，即

$$Var(y_i \mid X_i) = \exp(X_i^{'}\beta) + \alpha(\exp(X_i^{'}\beta))^2 \tag{3.13}$$

使用象

$$\hat{\alpha} = \frac{\sum_{i=1}^{n}(\exp(X_i^{'}\hat{\beta}))^2\{(y_i - \exp(X_i^{'}\hat{\beta}))^2 - \exp(X_i^{'}\hat{\beta})\}}{\sum_{i=1}^{n}(\exp(X_i^{'}\hat{\beta}))^4} \tag{3.14}$$

这样的 α 的一致估计，计算

$$Var(\hat{\beta}) = \left(\sum_{i=1}^{n}\exp(X_i^{'}\beta)X_i X_i^{'})\right)^{-1}\left(\sum_{i=1}^{n}\{\exp(X_i^{'}\beta) + \alpha(\exp(X_i^{'}\beta))^2\}X_i X_i^{'}\right)$$

$$\times\left(\sum_{i=1}^{n}\exp(X_i^{'}\beta X_i X_i^{'})\right)^{-1} \tag{3.15}$$

最后，一个相对不是那么严格的方法是用 Eicker-White 稳健估计量

$$Var(\hat{\beta}) = \left(\sum_{i=1}^{n}\exp(X_i^{'}\beta)X_i X_i^{'}\right)^{-1}\left(\sum_{i=1}^{n}\left((y_i - \exp(X_i^{'}\hat{\beta}))^2 X_i X_i^{'}\right)\right)$$

$$\times\left(\sum_{i=1}^{n}\exp(X_i^{'}\beta)X_i X_i^{'}\right)^{-1} \tag{3.16}$$

这种方法不需要对条件方差假定一个特定的模型。

---

* 原文是 β，有误——译者注。

当数据过度离中时，未能进行这样的修正，就会导致高估回归元的统计显著性。

3.2.3. 基于前两阶矩的估计

前面 3.2.2.小节仅在计算标准误时，利用了二阶矩的信息。在 β 的估计方法中直接使用这个信息可以提高有效性。

当方差是均值的倍数时，仅仅使用（3.1）和（3.10）是最有效的估计量。可以证明它等于用（3.11）和（3.12）正确计算标准误的泊松最大似然估计。

当方差是均值的二次方时，最有效的统计量是仅仅用（3.1）和(3.13)解一阶条件

$$\sum_{i=1}^{n}\frac{(y_i-\exp(X_i'\beta))}{\exp(X_i'\beta)+\hat{\alpha}(\exp(X_i'\beta))^2}\exp(X_i'\beta)X_i=0 \tag{3.17}$$

其中 $\hat{\alpha}$ 的估计值由（3.15）给定，渐近方差为

$$Var(\hat{\beta})=\left(\sum_{i=1}^{n}\{\exp(X_i'\beta)+\alpha(\exp(X_i'\beta))^2\}^{-1}\exp(X_i'\beta))^2X_iX_i'\right)^{-1} \tag{3.18}$$

这种基于前两阶矩的估计量，在统计文献中被称为*准似然估计量*（quasi-likelihood estimators），被 Gourieroux, Montfort 和 Trognon(1984)称为准广义伪最大似然估计量（quasi-generalized pseudo-maximum likelihood estimators）。

最后，我们指出，Delgado 和 Kniesner(1996)给出了一个调整的半参数估计量，仅仅需要设定一阶矩，但是和任何基于前两阶矩的知识的估计量一样有效。

3.2.4. 模型的评价

通过比较因变量这一计数变量的样本均值和样本方差可以得到不够离中和过度离中的大致幅度的指征，因为其后的泊松回归将在一定程度上使因变量的条件方差变小，但却不会改变条件均值的平均数（如果回归时包含常数项，由于泊松残差之和为 0，拟合均值的平均数等于样本均值）。如果样本方差比样本均值小，一旦引入回归元，数据甚至会更加不够离中，而如果样本方差超过样本均值的两倍，几乎可以肯定在包括回归元之后数据仍将会过度离中。

在 3.2.1 节，已经给出了对过度离中和不够离中以及区分泊松模型和负二项式模型的正式检验。在方差函数有不同设定——例如 Negbil 1 和 Negbin2——的负二项式模型中进行选择，可以根据最大似然值来决定，在不同的非嵌套混合模型（non-nested mixture models）中进行选择，也可以基于有最大似然值，或者当模型的参数次数不一样时，使用 Akaike 的信息准则来决定。

一个更具实质意义的选择是要用像负二项式模型这样的完全参数法，还是用仅仅使用了一阶矩和二阶矩的信息的估计量。在理论上，完全参数法的优点是有效率，缺点是对于模型的偏离不够稳健，即使正确地设定了均值，但若错误地设定了分布的其他方面，计数模型（泊松模型和 Negbin2 除外）的最大似然估计仍然是不一致的。在实际应用中，诸如 Cameron 和 Trivedi（1986）以及 Dean 等人（1989）的研究没有发现最大似然估计量和基于更弱的假定的估计量有多大差别。这种潜在的差别可以用作豪斯曼检验（Hausman test）的基础，例如，可参见 Dionne 和 Vanasse（1992）。而对于某些分析，例如要预测次数的概率而不仅仅是均值，分布的设定是必要的。有许多方法来评价模型的表现，一个标准的程序是比较*皮尔逊统计量*

$$P=\sum_{i=1}^{n}\frac{(y_i-\exp(X_i^{'}\hat{\beta}))^2}{v(X_i,\hat{\beta},\hat{\alpha})} \tag{3.19}$$

其中，$v(X_i,\beta,\alpha)$ 等于 $Var(y_i\mid X_i)$ 与自由度（n-k）之比。当 $v(X_i,\beta,\alpha)=\exp(X_i^{'}\beta)$ 时，皮尔逊统计量对检验泊松模型的适当程度是有用的，但对于其他模型，其有用性就有限多了。特别地，如果设定 $v(X_i,\beta,\alpha)=\alpha\exp(X_i^{'}\beta)$，并且用（3.12）估计 α 值，那么 P 永远等于(n-k)。

Cameron 和 Windmeijer(1996)提出了计数模型的各种 R^2。对泊松模型，他们偏爱以偏差为基础的 R^2 的量度

$$R^2_{DEV,P}=\frac{\sum_{i=1}^{n}y_i\log(\exp(X_i^{'}\hat{\beta})/\bar{y}}{\sum_{i=1}^{n}y_i\log(y_i/\bar{y})} \tag{3.20}$$

其中，当 y=0 时，ylog y=0。如果软件包报告模型拟合的对数似然值，这个 R^2 可以用 $(l_{fit}-l_0)/(l_y-l_0)$ 计算，这里 l_{fit} 是模型拟合的对数似然值，l_0 是只有截距的模型的对数似然值，而 l_y 是均值等于实际值的模型的对数似然值，即 $l_y=\sum_{i=1}^{n}y_i\log(y_i)-y_i-\log(y_i!)$。各项都容易计算。同样的指标也适用于（3.10）的过度离中模型的估计。对于（3.13）过度离中的负二项式模型的估计即 Negbin2，对应的 R^2 是

$$R^2_{DEV,NB2}=1-\frac{\sum_{i=1}^{n}y_i\log(y_i/\hat{\lambda}_i)-(y_i+\hat{\alpha}^{-1})\log((y_i+\hat{\alpha}^{-1})/(\lambda_i+\hat{\alpha}^{-1}))}{\sum_{i=1}^{n}y_i\log(y_i/\bar{y})-(y_i+\hat{\alpha}^{-1})\log((y_i+\hat{\alpha}^{-1})/(\lambda_i+\hat{\alpha}^{-1}))}, \tag{3.21}$$

其中 $\hat{\lambda}_i=\exp(X_i^{'}\hat{\beta})$

一个粗略的诊断是用于计算拟合的频率分布，作为对观测值拟合每一计数的预测概率的平均，把这个分布与观测到的频率分布相比较。在这个指标上表现差，有理由拒绝这个模型，然而表现好未必是接受这个模型的理由。作为一个极端的例子，如果仅仅观测到 0 和 1 这两个数，用最大似然法估计一个带常数项的 logit 模型，可以发现平均拟合频率恰好等于观测到的频率。

3.2.5. 在金融数据上的某些应用

例 1-4 分别阐述泊松模型（两次）、负二项式模型和混合的泊松－逆高斯模型。

例 1. Jaggia 和 Thosar(1993)对 1978—1985 年间在第一次出价后至被实际收购的大约 52 周内 126 家美国被收购企业的竞价收购次数进行建模。因变量的计数变量 y_i 是被收购的企业在第一次收购出价之后竞价收购的次数，取值在表 1 中给出。Jaggia 和 Thosar 发现竞价收购的次数随目标企业的管理层采取防卫行动（通过提起法律诉讼进行法律防卫或邀请友好的

第三方来报价）而增加，随收购溢价（收购价除以收购前 14 个工作日内的价格）而减少。收购价的次数一开始随企业规模的增大而增大，然后随企业规模的增大而减小（是企业规模的二次方），不受联邦管理机构介入的影响。用（3.8）没有发现过度离中。

例 2. Davutyan(1989)根据表 1 概括的 1947 年到 1986 年期间美国银行每年倒闭的次数估计泊松模型，结果显示银行倒闭随银行整体盈利能力、公司盈利能力，以及银行从联邦储备银行借款的增多而减少。对泊松模型没有进行正式的检验。银行倒闭样本均值和方差分别为 6.343 和 11.820，因此，在回归后可能仍有中等程度的过度离中，t 统计量相应有一定程度的上偏。问题更大的是数据的时间序列性质。Davutyan 用杜宾－沃森序列相关检验方法检验泊松残差的自相关。但是当因变量是异方差时，这个检验是不合适的。检验一阶序列相关的一个更好的方法是基于标准化残差 $(y_t-\hat{\lambda}_t)/\sqrt{\lambda}$ 的一阶序列相关系数 r_1：在 y_t 没有序列相关的零假设下，Tr_1 有 $\chi^2(1)$ 的渐近分布，其中 T 是样本的大小，参见 Cameron 和 Trivedi(1993)。计数数据的时间序列回归还处于初始阶段，参见 Gurmu 和 Trivedi（1994）的简短评论。

例 3. Dionne 和 Vanasse（1992）使用魁北克地区 1982 年 8 月—1983 年 7 月间 19013 个司机向警察报告损坏超过 250 美元的交通事故的次数数据。其频率很低，样本均值为 0.070，样本方差为 0.078，与均值接近，但是 Negbin2 模型比泊松模型更好，因为离中参数在统计上是显著的，而且 Negbin2 模型拟合优度的卡方统计量要好得多。这篇文章的主要贡献是运用这些横截面的负二项式参数估计值导出预测的索赔频率，进而根据不同特征和不同记录的不同个体的数据预测保费。假定个体 i 在时间 $1,\cdots,T$ 的索赔次数 $(y_{i1},\cdots,y_{iT})$ 是相互独立的、均值为 $(\lambda_{i1}v_i,\cdots,\lambda_{iT}v_i)$ 的泊松分布，其中 $\lambda_i=\exp(X_i'\beta)$，而 v_i 是一个不随时间变化的不可观测的成分，服从均值为 1，方差为 α 的伽马分布。[6] 那么，已知过去索赔的经验、现在和过去的特征（但不知道不可观测的成分 v_i），第 i 个体在 $T+1$ 期索赔的次数的最优预测是 $\exp(X_{i,T+1}'\beta)\left[\dfrac{1/\alpha+\bar{Y}_i}{1/\alpha+\bar{\lambda}_i}\right]$，其中 $\bar{Y}_i=1/T\sum_{t=1}^{T}y_{it}$ 而 $\bar{\lambda}_i=1/T\sum_{t=1}^{T}\exp(X_{it}'\beta)$。这是对横截面负二项式估计的 $(\hat{\alpha},\hat{\beta})$ 求值计算的。当回归元是像年龄、性别、婚姻状况这样的变量且随时间的变化很容易衡量时，这一预测特别容易实行。

例 4，Dean 等人（1989）分析了出版在 Andrews 和 Herzberg(1985)中的、有关 1997 年瑞典 315 个风险组中，每组第三方的汽车保单中交通事故索赔次数。计数取值范围很大——中位数是 10 而最大值是 2127——因此显然有必要控制风险组的规模。通过定义均值等于 $T_i\exp(X_i'\beta)$ 来进行这一工作，其中 T_i 是这个组投保的汽车年数，这等价于把 $\log T_i$ 作为一个回归元而且限制它的系数等于 1 单位，参见（3.2）。甚至在包括了这一项以及其他回归元之

[6] 这隐含着在每个时间间隔，索赔次数服从 Negbin2 分布。

后，数据还是过度离中的。对于泊松模型的最大似然估计，皮尔逊统计量是 485.1，自由度是 296，对（3.10）形式的过度离中意味着使用（3.12），这样 $\hat{\tau}$ =1.638，远大于 1。Dean 等人通过用最大似然法估计混合泊松-逆高斯模型来控制过度离中，过度离中的形式是（3.13）。他们发现这些最大似然估计值落在仅仅使用第一、二阶矩解（3.1）得到的估计值的 1%范围以内。他们没有尝试将其与更方便的负二项式模型的估计值进行比较。

3.3. 截断、审查和修正的计数模型

在某些情况下，仅仅经历过所关注的事件个体才会被抽取，此时数据是*在 0 点左截断*，仅仅观测到正的次数。令 $f(y_i \mid X_i)$ 代表未截断的总体密度，通常是(3.3)定义的泊松或(3.6)定义的 Negbin2。那么截断过的密度用 $(1-f(0 \mid X_i)$ 正规化， y_i 超过零这一事件的条件概率是 $f(y_i \mid X_i)/(1-f(0 \mid X_i)$， $y_i = 1,2,3,\cdots$，而对数似然函数是

$$\log L = \sum_{i:y_i>0} [\log f(y_i \mid X_i) - \log(1-f(0 \mid X_i))] \qquad (3.22)$$

用最大似然法估计，对于泊松模型， $f(0 \mid X_i) = \exp(-\exp(X_i'\beta))$，而对于 Negbin2 模型，$f(0 \mid X_i) = -\alpha^{-1}\log(1+\alpha\exp(X_i'\beta))$。从原理上说，可以对截断均值进行非线性回归来估计模型，这样做而不用最大似然法估计，在计算上没有多少优点。Grogger 和 Carson(1991)以及 Gurmu 和 Trivedi（1992）讨论了其他直接的变化，例如在一个比 0 大的点左截断以及右截断。

比右截断更常见的是*右审查*，当计数在一个最大值——比如说 m 以上时，仅仅记录为范畴 m 或更多个范畴，则对数似然函数为：

$$\log L = \sum_{i:y_i<m} \log f(y_i \mid X_i) + \sum_{i:y_i \geq m} \log(1-\sum_{j=0}^{m-1} f(j \mid X_i)), \qquad (3.23)$$

即使计数是完全记录的，也可能有这种情况，即不是所有计数的值都来自相同的过程。特别地，由于 0 计数的某种阈值（threshold），0 计数的过程可能与为正的计数过程不一样。一个连续数据的例子是用于劳动力供给的样本选择模型，在这模型里某个人决定是否去工作，即工时是否为正的过程，与决定多少正工时的过程是不一样的。相似地，对于计数数据，决定信用分期付款是否不支付的过程可能与违约者决定不支付分期付款次数的过程是不一样的。*修正的计数模型*允许这样的不同过程，我们仅仅考虑 0 计数的修正，虽然这个方法也可扩展到其他计数模型。

一个修正的计数模型是 Mullahy(1986)的槛（hurdle）模型。假设 0 来自密度函数 $f_1(y_i \mid X_i)$，例如，以 X_{1i} 为回归元 Negbin2，参数为 α_1 和 β_1，而正数来自 $f_2(y_i \mid X_i)$，例如，以 X_{2i} 为回归元的 Negbin2，参数为 α_2 和 β_2。那么 0 值的概率清楚地是 $f_1(0 \mid X_i)$，而

为了保证概率之和等于 1，正计数的概率为$\frac{1-f_1(0\mid X_i)}{1-f_2(0\mid X_i)}f_2(y_i\mid X_i)$，$y_i=1,2,\cdots$,对数似然函数为

$$\log L=\sum_{i:y_i=0}\log f_1(0\mid X_i)+\sum_{i:y_i>0}\{\log(1-f_1(0\mid X_i))$$

$$-\log(1-f_2(0\mid X_i))+\log f_2(y_i\mid X_i)\} \qquad (3.24)$$

另一个可选择的修正是*“带零”*(with zero)模型，该模型用下述方式把二元过程和计数过程组合在一起。假如二元过程取值 0，譬如，一个事件发生的概率是$f_1(0\mid X_i)$，那么$y_i=0$；如果二元过程取值 1，一个事件发生的概率是$1-f_1(0\mid X_i)$，那么y_i可以取值$0,1,2,\cdots$的概率$f_2(y_i\mid X_i)$是由一个像泊松分布或负二项式分布这样的密度函数决定的。因此，0 值的概率是$f_1(0\mid X_i)+(1-f_1(0\mid X_i))f_2(0\mid X_i)$，而一个正数的概率是$(1-f_1(0\mid X_i))f_2(y_i\mid X_i)$，$y_i=1,2,\cdots$对数似然函数是

$$\log L=\sum_{i:y_i=0}\log\{f_1(0\mid X_i)+(1-f_1(0\mid X_i))f_2(0\mid X_i) \qquad (3.25)$$

$$+\sum_{i:y_i>0}\{\log(1-f_1(0\mid X_i)+\log f_2(y_i\mid X_i) \qquad (3.26)^{*}$$

这个模型也称作 0 膨胀(zero inflated)计数模型，尽管它可能也可以解释 0 计数太少。Mullahy(1986)提出了这个模型，他设$f_1(0\mid X_i)$等于一个常数，比如说β_1，而 Lambert(1992)和 Greene(1994)使用 logit 模型，在这种情况下，$f_1(0\mid X_i)=(1+\alpha\exp(-X_{1i}'\beta_1))^{-1}$。

由于仅仅报告因变量的均值和方差，0（或其他值）太少或太多的问题，很容易被遗漏。好的做法是同时报告频率，并把这些频率与拟合的频率进行比较。

例 5．在一个早期的版本中，Dionne 等（1996）分析一家西班牙银行授信的 4691 个个人样本中，未支付的分期付款贷款的次数。原始数据显示，有相当大的过度离中，均值为 1.581 而方差为 10.018。在包含年龄、婚姻状况、小孩的个数、每月净收入、房屋所有权、每月分期付款额、可否取到信用卡、要求贷款额等回归元后，过度离中现象仍然存在。对 Negbin2 模型，$\hat{\alpha}=1.340$。关注点是确定不利的信用风险，分开估计了截断的 Negbin2 模型（3.22）。如果决定 0 计数的过程和决定正计数的过程是一样的，那么仅仅估计正计数会导致效率损失。如果决定 0 计数的过程与决定正计数的过程是不同的，那么估计截断模型等于最大化槛对数似然（hurdle log-likelihood）（3.24）的子成分，没有效率损失。[7]

---

[*] 原书中，本式及下面有些式子都用了两个编号，有误。为了保持与原书一致，仍沿用这些编号——译者注。

[7] 槛对数似然对f_1和f_2是可加的，子成分f_2等于（3.22）而且如果f_1和f_2没有共同参数，信息矩阵是对角阵。

例 6，Greene（1994）分析了一个 1319 个个人主要信用卡申请者，60 天或更长时间清算的信用帐户的主要名誉受损的报告（major derogatory reports, MDR），发现 MDR 随支出收入比率（每月平均支出除以每年收入）的增大而减少，而年龄、收入、信用卡每月平均支出以及个人是否持有另外一张信用卡，在统计上不显著。数据过度离中，negbin2 要比泊松模型好很多，Greene 也估计了“带零”的 Negbin2 模型，用 logit 和 probit 模型估计 0，回归元是年龄、收入、房屋所有权、自我就业、受赡养者的数字、受赡养者的平均收入。带零模型可能是不必要的，因为标准的 Negbin2 模型预测了 1070 个 0，与观测到的 1060 个 0 很接近。带零的 Negbin2 模型的对数似然值是-1020.6，带有 7 个额外的参数，比 Negbin2 模型的对数似然值-1028.3 大不了多少。根据 Akaike 的信息准则，前一个模型更优。Greene 还另外估计了连续数据的标准样本选择模型的一个计数数据变形（variant）。

3.4 持续期数据的指数分布和威布尔分布

最简单的持续期数据模型是指数模型，使用纯泊松分布隐含了持续期分布的密度函数为 $\lambda e^{-\lambda t}$ 和常数险率 λ。如果数据完全观测到，并用指数模型来估计，而一个不同的模型例如威布尔模型是正确的，如果仍然正确设定了均值，那么指数的最大似然估计是一致、但无效的，而且通常的最大似然估计结果给出不正确的标准误。这与当负二项式模型是正确模型时，使用泊松模型来估计是相似的。然而，偏好比指数更一般的模型的更重要原因是数据的观测通常是不完全的。在这种情况下，不正确的分布选择会导致参数的估计不一致。例如，一个有限时期的观测可能意味着更长的时段而没有观测到它们结束，常数险率的限制对计量经济数据一般来说是不合适的，因而我们立即转到威布尔的分析，它嵌入指数分布作为它的一个特例。因为本文的焦点是计数而不是持续期，我们对威布尔的讨论是简短的。标准的参考文献包括 Kalbfleish 和 Prentice（1980）、Kiefer（1988）以及 Lancaster(1990)。

威布尔分布用它的险率 $\lambda(t)$——或用更早的表示方法 h(t)，等于 $\lambda\gamma t^{\gamma-1}$——相当容易定义。通过设定 λ 依赖于回归元，即 $\lambda=\exp(X_i'\beta)$，而 γ 不依赖于回归元，来构造回归模型。因而观测值 i 的风险是

$$\lambda_i(t_i \mid X_i)=\gamma t_i^{\gamma-1}\exp(X_i'\beta) \tag{3.27}$$

相应的密度函数为

$$f_i(t_i \mid X_i)=\gamma t_i^{\gamma-1}\exp(X_i'\beta)\exp(-t_i^{\gamma}\exp(X_i'\beta)) \tag{3.28}$$

这个过程的条件均值较为复杂

$$E[t_i \mid X_i]=(\exp(X_i'\beta))^{-1/\gamma}\Gamma(1+1/\gamma) \tag{3.29}$$

研究通常考虑回归元对险率而不是对条件均值的影响。如果 $\beta_j>0$，那么 X_{ij} 的增加会使险率增加，平均持续期减少，而如果 $\gamma>1$（或 $\gamma<1$），险率随持续期增大而增大（随持续期减小而减小）。

在许多应用中，持续期仅到上界才可以观测到。如果事件在这个时间之前没有发生，可以说这个时段是*不完全*的，更具体说是*右审查*的。对似然值的贡献是观测到一个至少持续到t_i的时段的概率，或生存函数

$$S_i(t_i \mid X_i) = \exp(-t_i^{\gamma} \exp(X_i^{'}\beta)) \tag{3.30}$$

组合后，若某些数据是不完全的，则对数似然函数是

$$\log L = \sum_{i:\ 完全} \left\{ \log\gamma + (\gamma - 1)\log t_i + X_i^{'}\beta - t_i^{\gamma} \exp(X_i^{'}\beta) \right\} \tag{3.31}$$

$$+ \sum_{i:不完全} - t_i^{\gamma} \exp(X_i^{'}\beta) \tag{3.32}$$

且γ和β都是用最大似然法估计的。

对不完全数据，如果模型没有正确设定，威布尔最大似然估计是不一致的。一个可能的错误设定是尽管t_i是威布尔分布，但参数是γ和$\lambda_i v_i$而不是γ和λ_i，其中v_i代表不可观测的个人异质性。如果v_i的分布是均值为 1，方差为α的 i.i.d 伽马分布，这就成了*威布尔－伽马*模型，生存函数为

$$S_i(t_i \mid X_i) = [1 + t_i^{\gamma-1} \exp(X_i^{'}\beta)]^{-1/\alpha} \tag{3.33}$$

从上式，可用通常的方式得到密度函数和对数似然函数。

持续期数据标准通用模型（standard general model）是（2.1）介绍的*比例险*模型或*比例强度*模型，它将险率分解为：

$$\lambda_i(t_i, X_i, \gamma, \beta) = \lambda_0(t_i, \gamma)\exp(X_i^{'}\beta) \tag{3.34}$$

其中$\lambda_0(t_i, \gamma)$是基准险函数，$\lambda_0(t_i, \gamma)$的不同选择对应于不同的模型，例如威布尔模型是$\lambda_0(t_i, \gamma) = \gamma t_i^{\gamma-1}$，而指数模型是$\lambda_0(t_i, \gamma) = 1$。回归元的唯一作用是作为这个基准险的尺度因子。险率的分解也导致对数似然的分解，子成分不依赖基准险，这对右审查数据特别有用。把$R(t_i) = \{j \mid t_j > t_i\}$定义为所有在时间$t_i$没有完成的时段的风险集合，那么 Cox(1972a)提出了最大化偏似然函数的估计量

$$\log L = \sum_{i=1}^{n} \left\{ X_i^{'}\beta - \log\left[\sum_{j \in R(t_i)} \exp(X_i^{'}\beta) \right] \right\} \tag{3.35}$$

这个估计量不是完全有效率的，但它的优点是不管基准险的真实函数形式如何，它与那些用最大似然法软件包报告的正确的标准误是一致的。

例 7．Bandopadhyaya(1993)分析了 1979－1990 年间根据第 11 章破产保护下的 74 家美国企业的数据。31 家仍然处在破产保护下，在这种情况下，数据是不完全的，可以采用经审查的威布尔模型（3.30）的最大似然估计。因变量是处在破产保护下的天数，平均持续期（计

算完全或不完全的时段）是 714 天。现有利息额的系数是正的，意味着险率增大和破产保护的平均持续期减少。另一个在统计上显著的变量是一个生产能力利用指标，也对险率有正的效应。估计的$\hat{\alpha}$ =1.629，超过 1 单位，因此企业处在破产保护下的时间越长，越可能脱离破产保护。相应的标准误 0.385，导致检验指数$\alpha=1$的零假设的t统计量等于 1.63，在 5%显著性水平的单边检验是临界（borderline）的不显著。由于威布尔模型提供了最好的拟合，威布尔模型要优于指数模型和对数－logistic 模型。

例 8．Jaggia 和 Thosar(1995)分析了美国 1978-1985 年间 161 家受到管理层反抗的、被收购企业的数据，在 26 例中，竞标仍在继续，数据是经审查的，因变量是从公开宣称要收购至收购到所需股份数的时间长度（以周为单位）。平均持续期（计算完全和不完全的持续时段）是 18.1 周。该文对系列模型进行了估计和设定检验。对不同回归元的相对统计显著性，不同的模型给出了相似的结果。但是险率如何随自收购开始以后的时间变化，不同模型的结果不同。管理层反抗收购的行动——提起法律诉讼和提出要改变财务结构——成功地减少了风险，增加了到接受收购的平均持续期；而出价的相互竞争，增加了风险，减少了直到接受收购的平均持续期。较优的模型是经审查的威布尔－伽马（censored Weibll-gamma）模型（3.31）。估计的险，以$X_i=\hat{X}$取值，最初迅速增加，然后随t缓慢减少，而威布尔给出的是一个单调递增的险率。对威布尔－伽马模型这种模型的批评，是其假设所有的时段最终都会完成，而这里一些企业可能永远不会被收购。Jaggia 和 Thosar 对 Schmidt 和 Witte(1989)的分化总体模型（split-population model）——该模型允许不被收购的概率为正——的估计和拒绝进行了简短的讨论。该研究是其他相类似研究的一个好的范本，它使用的技术在 LIMDEP 中已经有了。

3.5 分组持续期数据的泊松模型

金融数据中状态转移的一个重要例子是从开始发放抵押贷款的状态向抵押贷款被终止的状态转移。终止的原因可以是提前还款，也可以是违约。实际上，这在对有抵押保证的证券定价时是重要的。从计量经济学的角度看，这涉及到对在开始发放抵押贷款到提前还款或违约之间的时间区间建模。特别关注对风险作为抵押贷款还款期间的函数的形状以及协变量的作用。在这种背景下，Cox 对于持续期的比例险模型被广泛使用，（参见 Green 和 Shoven（1986)），Lane 等（1986），Baek 和 Bandopadhyaya(1996)，也可选择把分组的持续期数据作为计数来分析（Schwartz 和 Torous(1993)）

例 9，Green 和 Shoven（1986）分析了在 1947 年和 1976 年间发行，在 1975 年和 1982 年间终止的 3938 个加州 30 年固定利率抵押贷款。2037 个抵押贷款付清了，关注的是估计提前还款抵押贷款对市场利率与一个抵押贷款的固定利率——所谓的“锁定的幅度”（lock-in magnitude）之差——的敏感性。可取得的数据很有限，锁定幅度的值是唯一的回归元，因此其他如家庭规模或收入的改变这样的与个人相关的因素都被忽略。（该文作者得到的唯一的个体水平的数据是房子租期，以及房子市场价值的衡量）。一个a_i到期的抵押贷款，这里$a_i=t_i-t_{0i}$，而t_{0i}代表抵押开始日，其转移概率由$\lambda_i(a_i,X,\beta)=\lambda_0(a_0,\gamma_i)\exp(X'\beta)$给

出。作者用 Cox 偏似然估计量来估计$(\beta,\gamma_i,\ i=1,\cdots,30)$。序列$\{\gamma_i, i=1,2,\cdots\}$的（非参数）估计值有点类似对应于每个抵押贷款到期日的种类变量（categorical variables）系数的估计值，得到基本险函数。1975-1978 和 1978-1982 分开处理以允许随着 1978 年法律规定禁止为了提高抵押利率的唯一目的而使用“到期销售”（due-on-sale）条款，[系数β]可能有结构性变化。该文作者能够展示平均抵押贷款提前还款期对利率变化的敏感性。

例 10，Schwartz 和 Torous（1993）把泊松回归方法和比例险结构组合在一起，提供了区别于 Green-Shoven 的另一有趣的方法。他们关于 1975-1990 年间固定利率抵押贷款的 Freddie Mac 数据，有超过 39000 次提前还款而超过 8500 次违约。他们使用按月份分组的提前还款和违约次数，两者分开建模。令n_j代表 j 季度初已知的发放抵押贷款的次数，y_i代表在该季提前还款的抵押贷款的次数，$X(j)$代表随时间变化的协变量集。令$\lambda(a,X(j),\beta)=\lambda_0(a,r)\exp(X(j)'\beta)$代表平均每月提前还款的比率，表示成外生变量$X(j)$的函数和另一种基准险函数$\lambda_0(a,r)$，那么季度提前还款抵押贷款的期望次数将是$n_j\lambda_0(a,r)\exp(X(j)'\beta)$，最大似然估计是基于泊松密度函数

$$f(y_i \mid n_j, X(j)) = \frac{[n_j \cdot \lambda_0(a,\gamma)\exp(X(j)'\beta]^{y_j}\ \exp(-n_j \cdot \lambda_0(a,\gamma)\exp(X(j)'\beta))}{y_j!} \quad (3.36)$$

作者使用地区、季度、提前还款贷款的到期年分布等虚拟变量，其他变量包括开始时借款与价值的比率，再融资的机会和地区房产收益率。他们的结果表明地区间有显著差异，以及在再融资机会中所起的重要作用。

3.6 其他计数模型

美国股票价格是以八分之一美元单位（或点数 tick）度量的，对短时期，股价应该明显地（explicitly）作为整数来建模。Hausman，Lo 和 Mackinlay(1994)详细研究的 6 只股票中，同一股票相邻的交易价格 60%没有变化，而 35%的变化仅是 1 点（1 tick）。甚至每日收盘价也仅能改变几个点。这种股票价格的离散通常被忽略了。虽然一些使用连续定价模型的研究，已经允许价格的离散（Gottlieb 和 Kalay（1985）以及 Ball（1988））。

一个可能的方法是把价格水平（用点数衡量）当作计数来建模，但是这样的计数将是高度序列相关的，而且计数的时间序列回归模型还没有得到很好的发展。更有成效的方法是把价格的变化（再用点数衡量）当作计数来建模，然而由于一些计数是负的，标准的计数模型不适用。

一个允许计数为负的模型是*有序的 Probit* 模型，例如在 Maddala（1983）中就有这个模型。令y_i^*代表一个衡量价格变化倾向的潜在（未观测到的）随机变量，这里$y_i^* = X_i'\beta+\varepsilon_i$，

ε_i 的分布是 $N(0,\sigma_i^2)$，而且通常 $\sigma_i^2=1$。y_i^* 的更高值和实际的、离散的价格变化 y_i 的更高值 j 以下列方式相联系：如果 $\alpha_j < y_i^* \le \alpha_{j+1}$，则 $y_i = j$。那么得到代数式

$$\begin{aligned}\Pr\{y_i = j\} &= \Pr\{\alpha_j - X_i'\beta < \varepsilon_i \le \alpha_{j+1} - X_i'\beta\} \\ &= \phi\left(\frac{\alpha_{j+1} - X_i'\beta}{\sigma_i}\right) - \phi\left(\frac{\alpha_j - X_i'\beta}{\sigma_i}\right)\end{aligned} \tag{3.37}$$

令 d_{ij} 为 $y_i = j$ 时等于 1，$y_i \ne j$ 时等于 0 的虚拟变量，对数似然函数可以写成

$$\log L = \sum_{i=1}^{n}\sum_{j} d_{ij}\log\left[\phi\left(\frac{\alpha_{j+1} - X_i'\beta}{\sigma_i}\right) - \phi\left(\frac{\alpha_j - X_i'\beta}{\sigma_i}\right)\right] \tag{3.38}$$

这个模型中可用于非负的计数数据，在这种情况下 $j = 0,1,2,\cdots,\max(y_i)$。当运用排序 probit 模型而不是泊松模型或负二项式模型时，对回归元在其应用中的重要性和显著性，Cameron 和 Trivedi（1986）得出性质上相似的结果。对于可能为负的、离散的价格变化数据，Hausman 等（1992）用排序 probit 模型，$j = -m,-m+1,\cdots,0,1,2,\cdots,m$，其中值 m 实际上是 m 或更大，而 $-m$ 实际上是 $-m$ 或更小。那么，要估计的参数是模型中的 σ_i^2，回归参数 β 以及阈值参数 $\alpha_{-m+1},\cdots,\alpha_m$，其中 $\alpha_{-m} = -\infty$，而 $\alpha_{m+1} = \infty$。

例子 11. Hausman 等（1992）用 1988 年纽约证券交易所和美国证券交易所 100 只股票跺时（time-stamped）（最接近于秒）交易的数据，详细报告了 6 只股票的结果。每只股票分开建模型，一只股票（IBM）有多达 206794 次交易。因变量是相邻交易的价格变化(以八分之一美元为单位)。用有序的 probit 模型估计，对大多数股票 $m = 4$。回归元包括从上一次交易流逝的时间、上一次交易时的出价/要价差额、价格变化的三阶滞后和股票交易额（美元）的三阶滞后，而方差 σ_i^2 是自上次交易以来流逝的时间和上次交易的出价/要价差的线性函数。这一设定不是基于随机过程理论，而是以算术布朗运动为指南。Hausman 等人得到的结论是交易顺序影响价格的变化，更大的交易对价格有更大的冲击。

例 12，Epps（1993）直接把离散股票价格水平（而不是价格的变化）作为随机过程建模。假设在离散的时间 t 的股票价格 P_t 是 Galton-Watson 过程的实现。Galton-Watson 过程是一个标准的分叉过程，复杂之处在于代（generations）的次数也是随机的。给定 P_{t-1} 时，P_t 的条件密度，（或转移概率）容易用解析式表示。但是当它涉及到卷积（convolution）时，很难计算。这使得估计即使不是不可能也是困难的。Epps 用一个近似来代替，对（连续的）正规

化的价格变化 $(P_t - P_{t-1})/\sqrt{P_{t-1}}$ 建模，可以证明 y_t 是泊松复合事件分布的实现。Epps（1993）分析了从 1962 年到 1987 年单只股票的日收盘价数据。50 家公司每个分别进行分析，用矩程序的方法估计。这个模型的优点是收益率的条件分布包含了它对肥尾分布的预测。

4.总结性评论

基本的泊松模型和负二项式计数模型（以及其他泊松混合模型）可用现成的软件直接估计，而且在许多情况下是合适的。泊松回归模型估计后，应该使用辅助回归（3.8）或（3.9）对不够离中或过度离中进行正式的检验。如果这些检验拒绝中度离中(equidispersion)，那么标准误就该用（3.11）、(3.14)或(3.16)计算。如果数据过度离中，最好是用 Negbin2 模型（3.6）的最大似然估计。然而，需要指出的是，过度离中检验对检测其他形式的错误设定，例如，未能考虑到过多 0，是有功效的。

一种常见的情况是，当决定 0 计数的过程不同于决定正计数的过程时，这些模型就不适用。这种情况可以通过比较拟合的频率和观测到的频率来诊断。那么修正的计数模型，诸如槛模型或带零模型，或截断或审查模型都是合适的。

本文强调了计数模型和持续期模型的共同基础。当持续期和计数数据都可用时，特别是可以取得给定个体的多个时段的数据或是数据被分组时，对后者建模可能为回归元的作用提供更多的信息。用统一的时间区间分组是方便的，但是有时计数数据不从属于同样的区间。可能可以得到不同时间区间事件数的时间序列数据。通过使用比例强度泊松过程数据回归模型（Lawless(1987)），可以适应这样的复杂性。

在处理金融数据时，假设最简单的随机过程有时是不合适的。一个例子是交易或金融交易的次数可能在每个小时间单位内执行。在这种情况下，事件相互独立将不是一个令人信服的假定，因此更新理论是不合适的。体现相互依赖的一个方法是使用调制过的更新过程（modulated renewal process)(Cox(1972b))。对于持续期而不是计数的时间序列数据，Engle 和 Russell（1994）引入自回归条件持续期模型，该模型是类似 GARCH 的持续期数据模型。这个模型成功地解释了纽约证券交易所 IBM 股票连续两次交易之间秒数据的自相关。

除了相当有限的纯粹时间序列的情形外，时间序列计数回归模型还很不成熟。事实上，计量经济学家考虑处理的大多数标准的复杂性——例如同时性（simultaneity）和选择性偏差数据的技术，与连续数据相比，在计数数据中的发展还很不够。一个有用的出发点在 Gurmu 和 Trivedi(1994)的文献中综述。

致谢

作者感谢 Arindam Bandopadhyaya，Sanjiv Jaggia，John Mullahy 和 Per Johansson 对本文初稿的评论。

参考文献

Andrews, D. F. and A. M. Her'zberg (1985). Data. Springer-Verlag, New York.

Baek, I-M. and A. Bandopadhyaya(1996). The determinants of the duration of commercial bank debt rencgotiation for sovereigns. J. Banking Finance 20, 673-685.

Ball, C. A. (1988). Estimation bias induced by discrete security prices. J. Finance 43, 841-865.

Bandopadhyaya, A. (1994)- An estimation of the hazard rate of firms under chapter 11 protection. Rev. Econom. Statist. 76, 346-350.

Cameron, A. C. and P. K. Trivedi (1986). Econometric models based on count data: Comparisons and applications of some estimators and tests. J. Appl. Econom. 1 (I), 29-54.

Cameron, A. C. and P.K. Trivedi (1990). Regression based tests for overdispersion in the Poisson model. J. Econometrics 46 (3), 347-364.

Cameron, A. C. and P. K. Trivedi (1993). Tests of independence in parametric models with applications and illustrations. J. Business Econom. Statist. 11, 2943.

Cameron, A. C. and F. Windmeijer (1995). R-Squared measures for count data regression models with applications to health care utilization. J. Business Econom. Statist. 14(2), 209-220.

Consul, P. C. and F. Famoye (1992). Generalized Poisson regression model. Communications in statistics.' Theory and method 21 (1), 89-109.

Cox, D. R. (1962). Renewal Theory. Methuen, London.

Cox, D. R. (1972a). Regression models and life tables. J. Roy. Statist. Soc. Ser. B. 34, 187-220.

Cox, D. R. (1972b). The statistical analysis of dependencies in point processes. In: P.A.W. Lewis ed., Stochastic Point Processes. John Wiley and Sons, New York.

Cox, D. R. and P. A. W. Lewis (1966). The Statistical Analysis of Series of Events. Methuen, London.

Davutyan, N. (1989). Bank failures as Poisson variates. Econom. Lett. 29 (4), 333-338.

Dean, C., J. F. Lawless, and G. E. Wilmot (1989). A mixed Poisson-inverse Gaussian regression Model. Canad. J. Statist. 17 (2), 171-181.

Delgado, M. A. and T. J. Kniesner (1996). Count data models with variance of unknown form: An application to a hedonic model of worker absenteeism. Rev. Econom. Statist., to appear.

Dionne, G., M- Artis and M. Guillen (1996). Count data models for a credit scoring system. j. Empirical Finance, to appear.

Dionne, G. and C. Vanasse (1992). Automobile insurance ratemaking in the presence of asymmetric information. J. Appl. Econometrics 7 (2), 149-166.

Engle, R. F. and J. R. Russell (1994). Forecasting transaction rates: The autoregressive conditional duration model. Working Paper No. 4966, National Bureau of Economic Research, Cambridge, Massachusetts.

Epps, W. (1993). Stock prices as a branching process. Department of Economics, University of Virginia, Charlottesville.

Feller, W. (1966). An Introduction to Probabdity Theory, Vol Il. New York: Wiley.

Gottlieb, G. and A. Kalay (1985). Implications of the discreteness of observed stock prices. J. Finance 40 (1), 135-153.

, C., A. Monfort and A. Trognon (1984). Pseudo maximum likelihood methods: Applications to Poisson models. Econometrica 52 (3), 681-700.

Green, J. and J. Shoven (1986). The effects of interest rates on mortgage prepayments. J. Money, Credit and Banking 18 (1), 41-59.

Greene, W. H. (1994). Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. Discussion Paper EC-94-10, Department of Economics, New York University, New York.

Grogger, J. T. and R. T. Carson (1991). Models for truncated counts. J. Appl. Econometrics 6 (3), 225-238.

Gurmu, S. and P. K. Trivedi (1992). Overdispersion tests for truncated Poisson regression models. J. Econometrics 54, 347-370.

Gurmu, S. and P. K. Trivedi (1994). Recent developments in models of event counts: A Survey. Discussion Paper No.261, Thomas Jefferson Center, University of Virginia, Charlottesville.

Hausman, J. A., A. W. Lo and A. C. MacKinlay (1992). An ordered probit analysis of transaction stock prices. J. Financ. Econom. 31,319-379.

Jaggia, S., and S. Thosar (1993). Multiple bids as a consequence of target management resistance: A count data approach. Rev. Quant. Finance Account. December, 447--457.

Jaggia, S. and S. Thosar (1995). Contested tender offers: An estimate of the hazard function. J. Business Econom. Statist. 13 (1), 113-119.

Kalbfleisch, J. and R. Prentice (1980). The Statistical Analysis of Failure Time Data. John Wiley and Sons, New York.

Karlin, S. and H. Taylor (1975). A First Course in Stochastic Processes, 2nd. ed., Academic Press, New York.

_Kiefer, N. M. (1988). Econometric duration data and hazard functions. J. Econom. Literature 26 (2), 646679.

King, G. (1989). Variance specification in event count models: From restrictive assumptions to a generalized estimator. Amer. J. Politic. Sci. 33, 762-784.

Lambert, D. (1992).Zero-inflated Poisson regression with an application to defects in manufacturing. Technometrics 34, 1-14.

Lancaster, T. (1990). The Econometric Analysis of Transition Data. Cambridge University Press, Cambridge.

Lane, W., S. Looney and 1. Wansley (1986). An application of the Cox proportional hazard model to bank failures. J. Banking Finance 18 (4), 511-532.

Lawless, J. F. (1987). Regression methods for Poisson process data. J. Amer. Statist. Assoc. 82 (399), 808-815.

Maddala, G. S. (1983). Limited-Dependent and Qualitative Variables in Econometrics. Cambridge University Press, Cambridge.

Mccullagh, P. and J. A. Nelder (1989). Generalized Linear Models. 2nd ed., Chapman and Hall, London.

Mullahy, J. (1986). Specification and testing of some modified count data models. J. Econometrics 33 (3), 341-365.

Schmidt, P. and A. Witte (1989). Predicting criminal recidivism using split population survival time models. J. Econometrics 40 (1), 141-159.

Schwartz, E. S. and W. N. Torous (1993). Mortgage prepayment and default decisions: A Poisson regression approach. AREUEA Journal: J. American Real Estate Institute 21 (4), 431-449.

Winlelmann, R. (1995). Duration dependence and dispersion in count-data models. J. Business and

Econom. Statist. 13, 467-474.

Winkelmann, R. (1994). Count Data Models: Econometric Theory and an Application to Labor Mobility. Springer-Verlag, Berlin.

Winkelmann, R. and K. F. Zimmermann (1995). Recent developments in count data modeling: Theory and application. J. Econom. Surveys 9, 1-24.