

第 30 章 久期分析

在实证研究中, 有时被解释变量为某种活动持续的时间 (duration, spell, time to event)。

比如: 病人存活的时间, 灯泡报废的时间, 失业持续的时间, 婚姻持续的时间, 罢工持续的时间, 战争持续的时间, 王朝的寿命, 刑满释放犯再次犯罪的时间等。

这类数据称为“久期数据” (duration data), 相应的分析方法称为“久期分析” (duration analysis)。

由于久期分析考察个体从某一状态转换到另一状态所花费的时间，故也称为“转换分析”(transition analysis)或“事件历史分析”(event history analysis)。

生物统计领域称其为“生存分析”(survival analysis)，运筹学领域称为“报废时间分析”(failure time analysis)，人口学领域称为“生命表分析”(life table analysis)，而保险领域称为“风险分析”(hazard analysis)。

久期分析的许多术语来自生物统计领域。

30.1 久期数据的处理方法

久期数据通常为横截面数据。

假设数据为 $\{T_i, \mathbf{x}_i\}_{i=1}^n$ ，其中 T_i 为被解释变量，即个体 i 的持续时间(寿命)；而 \mathbf{x}_i 为解释变量。

考虑用 OLS 估计以下模型：

$$T_i = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i$$

由于持续时间 $T_i \geq 0$ ，而由上式得到的预测值 $\hat{T}_i = \mathbf{x}_i' \hat{\boldsymbol{\beta}}$ 有可能为负数，故这是不现实的模型。

较为现实的建模方法将 $\ln T_i$ 作为被解释变量(假设 $T_i > 0$):

$$\ln T_i = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i$$

此对数线性模型的最大问题是，在我们抽样获得观测数据时，通常知道个体已经存活了一段时间。

而此方程却总是站在 $T_i = 0$ (即病情刚发作或确诊)的角度，无法纳入“个体已经存活了一段时间”这一信息。

我们常常更为关心给定个体已存活了一段时间的条件下，个体在下一个时刻死亡的概率，即下文的风险函数。

比如,我们关心已经失业三个月的失业者明天找到工作的概率。

另外,如果久期数据存在右归并(参见 14.3 节),或者随时间而变的解释变量 \mathbf{x}_{it} , 都很难通过 OLS 来处理。

久期分析常使用基于风险函数的一套特殊方法,形成自成体系的研究领域。

30.2 风险函数

记个体在某种状态中持续的时间(spell)或寿命为 $T \geq 0$, 其一个特定取值记为 t 。

假设 T 为连续型随机变量，并记其概率密度函数与累积分布函数分别为 $f(t)$ 与 $F(t)$ ，其中 $F(t)$ 也被称为“失效函数” (failure function)。

考虑“病人”存活期超过 t 的概率，称为“生存函数” (survivor function)：

$$S(t) \equiv P(T > t) = 1 - F(t), \quad t \geq 0$$

生存函数本质上相当于累积分布函数的“反函数” (reverse cumulative distribution function)。

由于累积分布函数 $F(t)$ 单调递增，故生存函数 $S(t)$ 单调递减。

假设病人已存活到时刻 t , 在 $[t, t + \Delta t)$ 期间 ($\Delta t > 0$) 死亡的概率为:

$$P(t \leq T < t + \Delta t | T \geq t) = \frac{P(t \leq T < t + \Delta t)}{P(T \geq t)} = \frac{F(t + \Delta t) - F(t)}{S(t)}$$

定义“风险率”(hazard rate)或“风险函数”(hazard function)为病人在时刻 t 的瞬间死亡率:

$$\begin{aligned}\lambda(t) &\equiv \lim_{\Delta t \rightarrow 0^+} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \\ &= \lim_{\Delta t \rightarrow 0^+} \frac{F(t + \Delta t) - F(t)}{\Delta t S(t)} \\ &= \frac{1}{S(t)} \lim_{\Delta t \rightarrow 0^+} \frac{F(t + \Delta t) - F(t)}{\Delta t} = \frac{f(t)}{S(t)}\end{aligned}$$

风险函数 $\lambda(t)$ 本质上是在给定存活至时刻 t 条件下的条件密度函数，故也称为“条件死亡率”(conditional failure rate)；而 $f(t)$ 为无条件密度函数。

如果 $f(t) = \phi(t)$ (标准正态密度)，则 $\lambda(t)$ 是反米尔斯比率(IMR)。

风险率的可能取值介于 0(无死亡风险)与 ∞ (必死无疑)之间。

在久期分析中，风险函数 $\lambda(t)$ 与生存函数 $S(t)$ 比密度函数 $f(t)$ 与累积分布函数 $F(t)$ 更为方便与常用。

也可以从风险函数 $\lambda(t)$ 出发，反推出生存函数 $S(t)$ 、累积分布函数 $F(t)$ 以及密度函数 $f(t)$ 。

首先，从上式可知，

$$\lambda(t) = -\frac{d \ln S(t)}{dt}$$

故 $d \ln S(t) = -\lambda(t) dt$ ，两边从 0 到 t 作定积分可得：

$$\ln S(t) = -\int_0^t \lambda(u) du$$

其中， u 为积分变量； $S(0) = 1$ (在初始时刻，所有个体都活着)。

$$F(t) = 1 - S(t) = 1 - \exp\left[-\int_0^t \lambda(u) du\right]$$

$$S(t) = \exp\left[-\int_0^t \lambda(u) du\right]$$

对方程两边求导可得,

$$f(t) = \lambda(t) \exp\left[-\int_0^t \lambda(u) du\right]$$

为度量截止时刻 t 的累积总风险, 定义“累积风险函数”(cumulative hazard 或 integrated hazard)为:

$$\Lambda(t) = \int_0^t \lambda(u) du = -\ln S(t)$$

累积风险函数的好处在于，它比风险函数可以更准确地估计。

如果知道累积风险函数，很容易计算生存函数：

$$S(t) = \exp[-\Lambda(t)]$$

例 假设 T 服从指数分布 (exponential distribution) ,
 $f(t) = \lambda e^{-\lambda t}$ ($\lambda > 0$) , 则

$$F(t) = \int_0^t \lambda e^{-\lambda s} ds = \int_0^t e^{-\lambda s} d(\lambda s) = -e^{-\lambda s} \Big|_0^t = 1 - e^{-\lambda t}$$

$$S(t) = e^{-\lambda t}, \quad \lambda(t) = \frac{f(t)}{S(t)} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda, \quad \Lambda(t) = \lambda t$$

指数分布的风险函数为常数，称为“无记忆性” (memoryless)。它意味着，瞬间死亡的概率并不依赖于已存活了多久。

可以证明，个体在 $(0, t_2)$ 区间死亡的概率等于在已知个体存活至时刻 t_1 的情况下，其在 $(t_1, t_1 + t_2)$ 区间死亡的概率。

另一方面，如果风险函数为常数 λ ，则其对应的密度函数为 $f(t) = \lambda \exp\left[-\int_0^t \lambda du\right] = \lambda e^{-\lambda t}$ ，一定服从指数分布。

指数分布的期望为 $E(T) = \frac{1}{\lambda}$ (风险率的倒数)，方差为 $\text{Var}(T) = \frac{1}{\lambda^2}$ 。

指数分布广泛应用于研究电子元器件的寿命问题。

指数分布只有一个参数，如果知道期望，则方差也确定(方差为期望的平方)，缺乏灵活性。

指数分布的无记忆性有时也不现实。它意味着，一个 20 岁的青年与一个 80 岁的老年不仅瞬间死亡率一样，而且在未来 10 年内死亡的概率也相同。

将指数分布拓展至两个参数的威布尔分布(Weibull distribution)。

定义 如果随机变量 T 的累积分布函数为 $F(t) = 1 - \exp(-\gamma t^p)$ ，其中 $\gamma > 0, p > 0$ ，则称其服从威布尔分布。

威布尔分布的生存函数为 $S(t) = 1 - F(t) = \exp(-\gamma t^p)$ 。

威布尔分布的密度函数为 $f(t) = F'(t) = \gamma p t^{p-1} \exp(-\gamma t^p)$ 。

威布尔分布的风险函数为

$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{\gamma p t^{p-1} \exp(-\gamma t^p)}{\exp(-\gamma t^p)} = \gamma p t^{p-1}$$

如果 $p=1$ ，则威布尔分布的 cdf 为 $F(t) = 1 - \exp(-\gamma t)$ ，就是指数分布的 cdf，故指数函数是威布尔分布的特例。

如果 $p > 1$ ，则风险函数 $\lambda(t)$ 单调递增，这意味着，活得越久则死亡概率越高(或许由于老年化过程)，被称为“正向久期依赖”(positive duration dependence);

如果 $p < 1$, 则风险函数 $\lambda(t)$ 单调递减(或许新生儿死亡概率最高); 被称为“负向久期依赖”(negative duration dependence)。

如果 $p = 1$ (即指数分布的情形), 则风险函数 $\lambda(t)$ 为常数(或许死亡由外在的随机因素所造成)。

参数 p 决定了风险函数 $\lambda(t)$ 的形状, 称为“形状参数”(shape parameter); 参数 γ 则决定其规模, 称为“规模参数”(scale parameter)。

30.3 久期数据的归并问题

久期数据常存在“右归并”(right censoring)。

当研究结束时，有些病人可能尚未死亡；或者有些失业者还未找到工作。

观测到个体存活时间从 0 直至某归并时间(censoring time) C^* 。

只知道个体的寿命属于区间 (C^*, ∞) ，不知道其具体取值。

导致右归并的原因还包括，个体中途退出研究，或研究者与个体失去联系，无法继续跟踪调查。

在存在右归并的情形，记个体 i 的真实寿命为 T_i^* (可能不可观测)，而归并时间为 C_i^* 。

实际观测到的 T_i 或为个体寿命 T_i^* ，或为归并时间 C_i^* ，取决于二者哪个更小：

$$T_i = \min(T_i^*, C_i^*)$$

以虚拟变量 d_i 来记录个体 i 的观测记录是否完整：

$$d_i = \mathbf{1}(T_i^* < C_i^*)$$

如果 $d_i = 1$ ，则有完整记录，无归并；如果 $d_i = 0$ ，存在右归并。

有时也会出现“左归并”(left censoring), 即只知道个体的寿命属于区间 $(0, C^*)$, 而不知道其具体取值。

另一种归并情形为“区间归并”(interval censoring), 即只知道个体的寿命属于区间 $[C_a^*, C_b^*)$, 也不知道其具体取值。

区间归并常因为数据的离散性而发生, 比如, 研究者只知道失业者在某一周找到工作, 而不知道其具体日期。

为保证久期分析的有效性, 常假设“独立归并”(independent censoring)或“无信息归并”(noninformative censoring), 即归并时间 C_i^* 的分布不包含任何有关个体寿命 T_i^* 分布的信息。

此时, 可将表示归并的虚拟变量 d_i 视为外生变量, 不必在意归

并究竟如何发生，也无须为“归并机制”(censoring mechanism)建模。

在久期样本中，每一个体开始活动(比如，开始生病或失业)的日历时间(calendar time)可以不同。

通常将“风险开始”(onset of risk)的时间标准化为 0 时刻。

以此度量的时间在 Stata 中称为“分析时间”(analysis time)。

久期分析的被解释变量 T_i 正是以分析时间来计算的。

30.4 描述性分析

常希望进行一些粗略的描述性分析，比如根据样本数据来估计生存函数、累积风险函数与风险函数，看它们的大致形状。

1. 生存函数

生存函数 $S(t)$ 为个体存活时间超过时刻 t 的概率。

如不存在归并，可定义 $S(t)$ 的估计量为，样本中存活时间超过时刻 t 的个体数目 r 占样本容量 n 的比例，即 $\frac{r}{n}$ 。

但此法在归并的情况下并不适用。

此时，一般使用 Kaplan-Meier 估计量(Kaplan and Meier, 1958, 简记 KM)，它在独立归并(independent censoring)的情况下依然是 $S(t)$ 的一致估计量。

记 $t_1 < t_2 < \cdots < t_j < \cdots < t_K$ 为样本中观测到的死亡时间。

记样本中在区间 $[t_{j-1}, t_j)$ 仍存活而面临危险(at risk)的个体数为 n_j 。

到了时间 t_j ，这些 n_j 个体的命运分为三种，即存活、死亡、归并(只知道其死亡时间大于 t_j ，但不再有观测数据)。

记在时间 t_j 死亡的人数为 m_j 。

给定存活至 t_{j-1} ，能进一步活至 t_j 的概率(频率)为 $\frac{n_j - m_j}{n_j}$ 。

活至 t_j 的无条件概率等于活过之前每一个区间的条件概率之乘积。Kaplan-Meier 估计量为

$$\hat{S}(t) \equiv \prod_{j|t_j \leq t} \left(\frac{n_j - m_j}{n_j} \right)$$

故 Kaplan-Meier 估计量也称为“连乘估计量”(product limit estimator)。

如果不存在归并，则 $n_{j+1} = n_j - m_j$ ，方程中的连乘可以错项相约，故 $\hat{S}(t) = r/n$ 。

2. 累积风险函数

累积风险函数 $\Lambda(t) = -\ln S(t)$ ，将 $\hat{S}(t)$ 的表达式代入，即可得 $\Lambda(t)$ 的估计量。

但此估计量的小样本性质不如 Nelson (1972) 与 Aalen (1978) 所提出的 Nelson-Aalen 估计量(简记 NA):

$$\hat{\Lambda}(t) \equiv \sum_{j|t_j \leq t} \left(\frac{m_j}{n_j} \right)$$

其中，每项 $\frac{m_j}{n_j}$ 为局部风险率，而上式则为局部风险率的加总。

3. 风险函数

可以用 $\hat{\lambda}_j \equiv m_j/n_j$ 作为风险率的估计量。

但此估计量为不光滑的阶梯函数。

另一方法是对累积风险函数求导，但 $\hat{\Lambda}(t)$ 也是阶梯函数，并不处处可导。

实践中，一般先通过核密度方法将阶梯形的累积风险函数光滑化，然后再以此生成风险函数。

30.5 久期模型的最大似然估计

真实持续时间 T^* 的概率分布可能依赖于某些解释变量 \mathbf{x} 。

记 T^* 的密度函数、累积分布函数、生存函数与风险函数为 $f(t|\mathbf{x},\boldsymbol{\theta})$ ， $F(t|\mathbf{x},\boldsymbol{\theta})$ ， $S(t|\mathbf{x},\boldsymbol{\theta})$ 与 $\lambda(t|\mathbf{x},\boldsymbol{\theta})$ ，其中 \mathbf{x} 为不随时间而变的解释变量(time-invariant covariates)， $\boldsymbol{\theta}$ 为待估计的未知参数。

比如， T^* 为失业持续的时间， \mathbf{x} 包括失业前的教育水平、工作经验、种族、婚姻状况、子女数目，以及政府发放的失业救济金额等；故 T 的分布依赖于 \mathbf{x} 。

假设存在右归并, 则其 MLE 估计类似于归并回归的 Tobit 模型。

对于未被归并的观测值, 其对似然函数的贡献为 $f(t | \mathbf{x}, \boldsymbol{\theta})$ 。

对于右归并的观测值, 我们只知道其持续时间超过 t , 故其对似然函数的贡献为 $P(T^* > t) = S(t | \mathbf{x}, \boldsymbol{\theta})$ 。

个体 i 对似然函数的贡献可写为

$$f(t_i | \mathbf{x}_i, \boldsymbol{\theta})^{d_i} S(t_i | \mathbf{x}_i, \boldsymbol{\theta})^{1-d_i}$$

d_i 为右归并虚拟变量, 即 $d_i = 1$ 表示无归并, $d_i = 0$ 表示右归并。

假设样本为 iid，将似然函数取对数，加总可得

$$\ln L(\boldsymbol{\theta}) = \sum_{i=1}^n [d_i \ln f(t_i | \mathbf{x}_i, \boldsymbol{\theta}) + (1 - d_i) \ln S(t_i | \mathbf{x}_i, \boldsymbol{\theta})]$$

由于 $f(t) = \lambda(t)S(t)$ ，故 $\ln f(t) = \ln \lambda(t) + \ln S(t)$ ，故

$$\ln L(\boldsymbol{\theta}) = \sum_{i=1}^n [d_i \ln \lambda(t_i | \mathbf{x}_i, \boldsymbol{\theta}) + \ln S(t_i | \mathbf{x}_i, \boldsymbol{\theta})]$$

如果直接对风险函数 $\lambda(t | \mathbf{x}, \boldsymbol{\theta})$ 建模，则此式更为方便。

如果似然函数正确，则 MLE 估计量一致、有效、渐近正态。

如果似然函数不正确，此 MLE 估计量通常不一致。

一个例外是 T^* 服从指数分布且不存在归并，则只要条件期望函数(conditional mean function)正确即可保证 MLE 估计量的一致性。

如果存在归并，即使 T^* 服从指数分布，MLE 估计量也不一致。

以上参数回归模型(parametric regression model)的最大缺点是缺乏稳健性，对概率分布的具体假设比较敏感。

如存在其他形式的归并，应对似然函数进行相应调整。

如果观测值存在左归并，只知道持续时间小于 t ，其对似然函数的贡献为

$$P(T^* < t) = F(t | \mathbf{x}, \boldsymbol{\theta})$$

如果观测值存在区间归并，只知道持续时间属于区间 $[C_a^*, C_b^*)$ ，故其对似然函数的贡献为

$$P(C_a \leq T^* < C_b) = S(C_a | \mathbf{x}, \boldsymbol{\theta}) - S(C_b | \mathbf{x}, \boldsymbol{\theta})$$

在经济应用中的久期数据常存在区间归并。比如，失业持续时间经常以周或月来计算，而指数分布、威布尔分布等都是连续型分布。此时，通常假设区间归并的影响甚微，作为连续数据处理。

30.6 比例风险模型

称风险函数 $\lambda(t; \mathbf{x})$ 为“比例风险”(Proportional Hazard, 简记 PH), 如果它可以分解为

$$\lambda(t; \mathbf{x}) = \lambda_0(t) h(\mathbf{x})$$

其中, $h(\cdot) > 0$, 而 $\lambda_0(t)$ 被称为“基准风险”(baseline hazard), 依赖于时间 t , 但不依赖于 \mathbf{x} 。

基准风险 $\lambda_0(t)$ 对于总体中的每一个体都相同, 而个体的风险函数则依据 $h(\mathbf{x})$ 与此基准风险 $\lambda_0(t)$ 成正比, 故名“比例风险”。

通常，令 $h(\mathbf{x}) = e^{\mathbf{x}'\boldsymbol{\beta}}$ ，则

$$\lambda(t; \mathbf{x}) = \lambda_0(t) e^{\mathbf{x}'\boldsymbol{\beta}}$$

其中， $e^{\mathbf{x}'\boldsymbol{\beta}}$ 被称为“相对风险” (relative hazard)，故 $\mathbf{x}'\boldsymbol{\beta}$ 也称为“对数相对风险” (log-relative hazard)。

如所有解释变量都为 0，风险函数 $\lambda(t; \mathbf{x})$ 就等于基准风险 $\lambda_0(t)$ 。

对方程两边取对数可得

$$\ln \lambda(t; \mathbf{x}) = \mathbf{x}'\boldsymbol{\beta} + \ln \lambda_0(t)$$

其中， $\boldsymbol{\beta}$ 可方便地解释为半弹性。

另一解释是, 如果 x 增加一单位, 则风险率为 $\lambda(t; \mathbf{x}) = \lambda_0(t)e^{x'\beta}$, 正好是原来风险率 $\lambda(t; \mathbf{x}) = \lambda_0(t)e^{x'\beta}$ 的 e^β 倍。

Stata 称 e^β 为 “风险比率” (Hazard Ratio, 简记 HR), 即 x 增加一单位, 将使新风险率变为原来风险率的 e^β 倍。

如果令基准风险 $\lambda_0(t) = e^a$, 其中 a 为待估参数(选择指数形式以保证 $e^a > 0$), 则为 “指数回归”, 因为指数分布的风险函数为常数。

也可从指数分布的密度函数出发, 即 $f(t) = \lambda e^{-\lambda t}$ ($\lambda > 0$), 而令 $\lambda = e^a e^{x'\beta} = e^{a+x'\beta}$, 其中 \mathbf{x} 不包含常数项(如果有常数项 e^{β_0} , 可与 e^a 合并)。

将指数分布的风险函数 λ 与生存函数 $S(t) = e^{-\lambda t}$ 代入可得：

$$\ln L(a, \boldsymbol{\beta}) = \sum_{i=1}^n \left[d_i (a + \mathbf{x}_i' \boldsymbol{\beta}) - e^{a + \mathbf{x}_i' \boldsymbol{\beta}} t_i \right]$$

如果令基准风险 $\lambda_0(t) = pt^{p-1}e^a$ ，其中 $p > 0$ ， a 为待估参数，则为“威布尔回归”(Weibull regression)，因为威布尔分布的风险函数为 γpt^{p-1} (此处令 $\gamma = e^{a + \mathbf{x}'\boldsymbol{\beta}}$)。

将指数分布的风险函数 γpt^{p-1} 与生存函数 $S(t) = \exp(-\gamma t^p)$ 代入可得：

$$\ln L(a, p, \boldsymbol{\beta}) = \sum_{i=1}^n \left\{ d_i [a + \mathbf{x}_i' \boldsymbol{\beta} + \ln p + (p-1) \ln t_i] - e^{a + \mathbf{x}_i' \boldsymbol{\beta}} t_i^p \right\}$$

当 $p = 1$ 时，威布尔分布退化为指数分布。

故可通过检验 “ $H_0: p = 1$ ” (或 $H_0: \ln p = 0$) 来确定应使用威布尔回归还是指数回归。

如果令基准风险 $\lambda_0(t) = e^{a+\gamma t}$ ，其中 a, γ 为待估参数，则为“冈珀茨回归” (Gompertz regression)，在人口学与保险精算应用广泛。

与威布尔分布相比，冈珀茨分布的风险函数也是单调函数，即当 $\gamma > 0$ 时单调递增，当 $\gamma < 0$ 时单调递减，而当 $\gamma = 0$ 时为常数(同样退化为指数分布)；二者的不同点在于，冈珀茨分布风险函数的变化率为指数形式($e^{\gamma t}$)，而威布尔分布风险函数的变化率为幂函数形式(t^{p-1})。

当 $\gamma = 0$ 时，冈珀茨分布退化为指数分布。

可检验“ $H_0: \gamma = 0$ ”来确定应使用冈珀茨回归还是指数回归。

威布尔分布与冈珀茨分布的风险函数都是单调函数，要么单调递增，要么单调递减，此性质有时也现实不符。

比如，对于人类的死亡率而言，婴儿与老人的死亡率较高，而中青年的死亡率较低，其风险函数并不具有单调性。

生物统计学家将人类死亡率的风险函数称为“浴缸风险”(bathtub hazard)，因为它的形状就像一个浴缸。

在比例风险模型的框架下，一解决方法是不设定基准风险 $\lambda_0(t)$ 的函数形式，参见下文的 Cox 模型。

另一解决方法是假设基准风险 $\lambda_0(t)$ 为阶梯函数，即 $\lambda_0(t)$ 在每个小区间内都是待估计的常数(正如指数模型的风险率为常数)，以更好地拟合实际的风险函数，称为“分段固定风险模型”(piecewise constant hazard model 或 piecewise constant exponential model)。

具体来说，分段固定风险模型的风险函数可写为

$$\lambda(t; \mathbf{x}) = \begin{cases} \lambda_1 e^{\mathbf{x}'\boldsymbol{\beta}} & t \in [0, \tau_1) \\ \lambda_2 e^{\mathbf{x}'\boldsymbol{\beta}} & t \in [\tau_1, \tau_2) \\ \vdots & \vdots \\ \lambda_M e^{\mathbf{x}'\boldsymbol{\beta}} & t \in [\tau_{M-1}, \tau_M) \end{cases}$$

为便于估计，上式可写为

$$\lambda(t; \mathbf{x}) = \begin{cases} e^{\ln \lambda_1 + \mathbf{x}'\boldsymbol{\beta}} & t \in [0, \tau_1) \\ e^{\ln \lambda_2 + \mathbf{x}'\boldsymbol{\beta}} & t \in [\tau_1, \tau_2) \\ \vdots & \vdots \\ e^{\ln \lambda_M + \mathbf{x}'\boldsymbol{\beta}} & t \in [\tau_{M-1}, \tau_M) \end{cases}$$

分段固定风险模型等价于在每个时间段分别引入一个截距项。

对应于每个时间段，可定义 M 个虚拟变量，并将 $(M - 1)$ 个虚拟变量作为解释变量引入指数模型(假设模型中包括截距项)。

【例】Lalive et al (2006)在研究失业持续时间时，将风险函数每四周分为一段(大多数失业者在 60 周内找到工作)。分段固定风险模型的缺点是，需要估计较多参数，故要求样本容量比较大。

对于比例风险模型 $\lambda(t; \mathbf{x}) = \lambda_0(t)e^{\mathbf{x}'\boldsymbol{\beta}}$ ，无论其基准风险 $\lambda_0(t)$ 如何设定(指数、威布尔、冈珀茨或分段固定)，对于参数 $\boldsymbol{\beta}$ 的经济含义解释都是一样的，即可将 $\boldsymbol{\beta}$ 解释为 x 对于风险函数的半弹性，或将 $e^{\boldsymbol{\beta}}$ 解释为风险比率。

30.7 加速失效时间模型

对于比例风险模型,通常分析的重点是解释变量 \mathbf{x} 对于风险函数 $\lambda(t; \mathbf{x})$ 的作用,不容易看出 \mathbf{x} 对于平均寿命 $E(T)$ 的影响。

考虑直接对 $\ln T$ 建模:

$$\ln T = \mathbf{x}'\boldsymbol{\beta} + u$$

其中, u 为扰动项。

由于 $\ln T$ 在 $(-\infty, \infty)$ 取值,故 u 服从取值于 $(-\infty, \infty)$ 的连续型分布。

由方程可得, $T = e^{\mathbf{x}'\boldsymbol{\beta}} v$, 其中 $v = e^u$ 。

记 v 的密度函数、累积分布函数、生存函数与风险函数分别为 $f_v(v)$, $F_v(v)$, $S_v(v)$ 与 $\lambda_v(v)$;

记 T 的密度函数、累积分布函数、生存函数与风险函数分别为 $f_T(t)$, $F_T(t)$, $S_T(t)$ 与 $\lambda_T(t)$, 则

$$S_T(t) = P(T > t) = P[e^{x'\beta} v > t] = P[v > e^{-x'\beta} t] = S_v(e^{-x'\beta} t)$$

因此,

$$f_T(t) = F_T'(t) = [1 - S_T(t)]' = -\frac{\partial S_v[e^{-x'\beta} t]}{\partial t} = f_v(e^{-x'\beta} t) e^{-x'\beta}$$

由此可知,

$$\lambda_T(t | \mathbf{x}) = \frac{f_T(t)}{S_T(t)} = \frac{f_v(e^{-x'\beta} t) e^{-x'\beta}}{S_v(e^{-x'\beta} t)} = \lambda_v(e^{-x'\beta} t) e^{-x'\beta}$$

如果 $e^{-x'\beta} > 1$ ，则意味着是对基准风险 $\lambda_v(t)$ 的加速；

如果 $e^{-x'\beta} < 1$ ，则意味着是对基准风险 $\lambda_v(t)$ 的减速。

这类模型被称为“加速失效时间模型” (Accelerated Failure Time Model, 简记 AFT)，尽管 $e^{-x'\beta} < 1$ 意味着减速。

如果扰动项 $u \sim N(0, \sigma^2)$ ，称为“对数正态模型” (log-normal model)。

如果 u 服从逻辑分布(logistic distribution)，称为“对数逻辑模型” (log-logistic model)。

如果 u 服从广义伽马分布 (generalized gamma

distribution) $\text{Gamma}(\beta_0, \kappa, \sigma)$, 称为“伽马模型”(gamma model)。

对数正态模型与对数逻辑模型的风险函数均为非单调函数(non-monotonic), 适用于风险率首先上升然后下降的情形。

包含三个参数的伽马模型, 其风险函数形状更为灵活。

如果 $\kappa=1$, 则伽马模型就是威布尔模型;

如果 $\kappa=1$ 而且 $\sigma=1$, 则伽马模型就是指数模型;

如果 $\kappa=0$, 则伽马模型就是对数正态模型。

对原假设“ $H_0: \kappa=1$ ”或“ $H_0: \kappa=0$ ”进行检验, 有助于选择模型适用的分布函数。

指数模型与威布尔模型也可以写成 AFT 形式,故它们既属于 PH 模型, 也属于 AFT 模型; 其他模型, 要么为 PH, 要么为 AFT。

对于 AFT 模型, 可进行 MLE 估计。

AFT 模型的参数 β 与 PH 模型的参数 β 的经济解释不同。

在 AFT 模型下, β 可以解释为当 x 边际增加时, 能使平均寿命增加的百分比(即半弹性)。

特别地, 对于指数模型而言, 在 PH 模型下, 风险率为 $\lambda = e^{a+x'\beta}$, 而平均寿命为 $E(T) = \frac{1}{\lambda} = e^{-(a+x'\beta)}$, 故其 AFT 模型为 $\ln T = -(a + \mathbf{x}'\beta) + u$ 。

由此可见,PH模型下参数 β 与AFT模型下的参数正好符号相反,而绝对值相等。在 Stata 中进行指数回归时,默认为 PH 模型;但如果加选择项“time”,则为 AFT 模型;二者的估计系数绝对值相等,但符号正好相反。

由于 PH 模型与 AFT 模型都对风险函数的形式作了具体假设,故属于“参数回归”(parametric regression)。

在进行参数回归时,可供选择的分布函数很多。如果进行抉择?

对于“嵌套模型”(nested models),可对相关参数进行 Wald 或似然比检验(比如,指数模型是威布尔模型与冈珀茨模型的特例;威布尔模型与对数正态模型是伽马模型的特例)。

对于“非嵌套模型”(non-nested models), 不存在这种嵌套关系, 一个模型并不是另一模型的特例, 可通过 AIC 信息准则比较模型拟合优度。

AIC 准则的定义为

$$\text{AIC} = -2\ln L + 2(K + c)$$

其中, $\ln L$ 为对数似然函数, K 为解释变量 \mathbf{x} 的维度, c 为概率分布的参数个数(比如, 指数分布的 $c = 1$; 威布尔分布的 $c = 2$; 伽马分布的 $c = 3$)。

30.8 Cox 模型

参数回归对风险函数的具体形式作了假设，如果模型设定正确，则 MLE 是最有效率的。

如果风险函数的设定错误，则 MLE 一般不一致。

我们对于风险函数的具体形式常常并无把握。

针对参数回归的缺点，在比例风险模型的框架下，Cox (1972, 1975)提出了以下“Cox 模型”或“Cox PH 模型”。

由于比例风险模型 $\lambda(t; \mathbf{x}) = \lambda_0(t)e^{x'\beta}$ 在形式上为乘法 (multiplicative), 故个体 i 与个体 j 的风险函数之比可以写为

$$\frac{\lambda(t; \mathbf{x}_i)}{\lambda(t; \mathbf{x}_j)} = \frac{\lambda_0(t)e^{x_i'\beta}}{\lambda_0(t)e^{x_j'\beta}} = e^{(x_i - x_j)'\beta}$$

个体 i 与 j 的风险函数之比不随时间而变, 只与 $(\mathbf{x}_i - \mathbf{x}_j)$ 有关。

故可不必假设基准风险 $\lambda_0(t)$ 的具体函数形式, 而依然得到对 β 的估计。

由于构成风险函数 $\lambda_0(t)e^{x'\beta}$ 的前半部分 $\lambda_0(t)$ 不需要估计参数, 而后半部分 $e^{x'\beta}$ 需要估计参数, 故为半参数回归。

1. 失效时间不重叠的情形(no tied failures)

首先考察失效时间均不相同的情形，即每一个体的“死亡”时间都不一样。

假设久期样本只包括以下四个观测值：

	subject	t	x
1.	1	2	4
2.	2	3	1
3.	3	6	3
4.	4	12	2

其中，*subject* 表示个体，*t* 表示失效时间(failure time)，而 *x* 为解释变量。记第 *j* 位个体的失效时间为 t_j ，则四位个体的失效时间分别为 $t_1=2$ ， $t_2=3$ ， $t_3=6$ ， $t_4=12$ ，完全不重叠。

记在时间 t_j 面临失效危险的所有个体所构成的集合为“风险集”(risk set) R_j 。如果个体已失效或被归并,则不再面临失效危险。

当 $t = 2$ 时, 风险集为 $R_1 = \{1, 2, 3, 4\}$, 实际观测到第 1 位个体失效。

当 $t = 3$ 时, 风险集为 $R_2 = \{2, 3, 4\}$, 实际观测到第 2 位个体失效。

当 $t = 6$ 时, 风险集为 $R_3 = \{3, 4\}$, 实际观测到第 3 位个体失效。

当 $t = 12$ 时, 风险集为 $R_4 = \{4\}$, 实际观测到第 4 位个体失效。

在以上四个失效时间(2, 3, 6, 12), 假设必然有一位个体会失效, 然后计算在此条件下实际失效的那位个体失效的条件概率, 分别记为 P_1, P_2, P_3, P_4 。

似然函数可写为

$$L(\beta) = P_1 P_2 P_3 P_4$$

下面分别来计算 P_1, P_2, P_3, P_4 。

当 $t = 12$ 时, 风险集只包含第4位个体, 故第4位个体失效的条件概率为1, 即 $P_4 = 1$ 。

当 $t = 6$ 时，给定第 3、4 位个体必然有一位失效，在此条件下第 3 位个体失效的条件概率为

$$P_3 = \frac{\lambda(6|x_3)}{\lambda(6|x_3) + \lambda(6|x_4)} = \frac{\lambda_0(6)e^{x_3\beta}}{\lambda_0(6)e^{x_3\beta} + \lambda_0(6)e^{x_4\beta}} = \frac{e^{x_3\beta}}{e^{x_3\beta} + e^{x_4\beta}}$$

其中， $\lambda(6|x_3)$ 表示 $t = 6$ ， $x = x_3$ 时的风险率。在上式中，由于基准风险 $\lambda_0(6)$ 已被约去，故 P_3 并不依赖于具体的失效时间 $t = 6$ 。

当 $t = 3$ 时，给定第 2、3、4 位个体必然有一位失效，在此条件下第 2 位个体失效的条件概率为

$$P_2 = \frac{\lambda(3|x_2)}{\lambda(3|x_2) + \lambda(3|x_3) + \lambda(3|x_4)} = \frac{e^{x_2\beta}}{e^{x_2\beta} + e^{x_3\beta} + e^{x_4\beta}}$$

当 $t=2$ 时, 给定第 1、2、3、4 位个体必然有一位失效, 在此条件下第 1 位个体失效的条件概率为

$$P_1 = \frac{\lambda(2 | x_2)}{\lambda(2 | x_1) + \lambda(2 | x_2) + \lambda(2 | x_3) + \lambda(2 | x_4)} = \frac{e^{x_1\beta}}{e^{x_1\beta} + e^{x_2\beta} + e^{x_3\beta} + e^{x_4\beta}}$$

似然函数可写为

$$L(\beta) = \prod_{j=1}^4 \left(\frac{e^{x_j\beta}}{\sum_{i \in R_j} e^{x_i\beta}} \right)$$

其中, R_j 为对应于 t_j (第 j 位个体失效时间) 的风险集。

一般地，如果样本中有 L 个失效时间 $\{t_1, \dots, t_L\}$ 以及多个解释变量 \mathbf{x} ，似然函数可写为

$$L(\boldsymbol{\beta}) = \prod_{j=1}^L \left(\frac{e^{\mathbf{x}'_j \boldsymbol{\beta}}}{\sum_{i \in R_j} e^{\mathbf{x}'_i \boldsymbol{\beta}}} \right)$$

上式称为“部分似然函数” (partial likelihood function)，因为它并非完整的似然函数。

部分似然函数只依赖于样本中个体失效时间的排序，不依赖于个体失效的具体时间。

由于同一个体多次出现于不同的风险集中，一般应使用 Lin and Wei (1989)提出的稳健标准误，在 Stata 中可由选择项 “`r`” 或 “`vce(robust)`” 来实现。

基于部分似然函数的 MLE 估计量一致且渐近正态，尽管其有效性不如基于完整似然函数的 MLE 估计量(前面的参数回归)。

但后者的一致性依赖于对风险函数形式的正确设定(我们对此一般无把握)，而前者的效率损失通常不大。

由于 Cox 模型的稳健性，在实践中十分流行，是半参数估计的成功例子。

2. 失效时间有重叠的情形(tied failures)

如果样本中的失效时间有重叠，则部分似然函数的计算更为复杂。

延续上面的简单例子，但假设第 3 位与第 2 位个体同时在 $t = 3$ 时失效，原始数据变为：

	subject	t	x
1.	1	2	4
2.	2	3	1
3.	3	3	3
4.	4	12	2

当 $t = 3$ 时，风险集为 $R_2 = \{2, 3, 4\}$ ，而实际观测到第 2、3 位个体同时失效。应如何计算在给定风险集中两位个体会失效的条件下，第 2、3 位个体失效的条件概率？

通常有以下两种处理方法。

处理方法之一是假设第 2、3 位个体的失效时间其实并不完全相同，称为“精确边际计算法”(exact-marginal calculation)。

如果失效时间为连续变量，则第 2、3 位个体同时失效的概率为 0。之所以在样本中存在同时失效的个体，只是由于我们的观测时间不够细致(比如，以周或月为分析时间)。

共有两种可能情形。第一种情形是，第 2 位个体先失效，然后第 3 位个体失效，记此情形的条件概率为 P_{23} ；

第二种情形是，第 3 位个体先失效，然后第 2 位个体失效，记此情形的条件概率为 P_{32} ；而总的条件概率为 $(P_{23} + P_{32})$ 。

记 $r_j \equiv e^{x_j\beta}$ ，根据前面的推导逻辑可知：

$$P_{23} = \frac{r_2}{r_2 + r_3 + r_4} \cdot \frac{r_3}{r_3 + r_4}$$

$$P_{32} = \frac{r_3}{r_2 + r_3 + r_4} \cdot \frac{r_2}{r_2 + r_4}$$

如果多位个体同时失效，则精确边际法可能计算量过大。

比如，有 10 位个体同时失效，则共有 $10! = 3,628,800$ 种情形。

Breslow (1974)提出了以下近似计算方法，即在计算时，不调整每一项分母中的风险集(一直使用最初的风险集)。

根据 Breslow 的方法,

$$P_{23} \approx P_{32} \approx \frac{r_2 r_3}{(r_2 + r_3 + r_4)^2}$$

Efron (1977)提出了另一近似计算方法。

Efron 方法比 Breslow 更为精确, 但计算费时更长。

当 $t = 3$ 时, 风险集为 $\{3, 4\}$ 或 $\{2, 4\}$, 对应的条件概率之分母分别为 $(r_3 + r_4)$ 或 $(r_2 + r_4)$ 。如果使用二者的平均数作为分母, 可得

$$P_{23} \approx P_{32} \approx \frac{r_2 r_3}{(r_2 + r_3 + r_4) \left(\frac{r_2 + r_3}{2} + r_4 \right)}$$

Breslow 方法与 Efron 方法都是对精确边际法的近似。

处理方法之二是假设第 2、3 位个体确实同时失效，称为“精确部分计算法” (exact-partial calculation)。

给定风险集{2, 3, 4}中有两位个体同时失效，则共有三种情形，即第 2、3 位个体同时失效，第 2、4 位个体同时失效，以及第 3、4 位个体同时失效。

因此，第 2、3 位个体确实同时失效的条件概率为：

$$P_{23} = \frac{r_2 r_3}{r_2 r_3 + r_2 r_4 + r_3 r_4}$$

精确部分法与精确边际法的计算结果不相等，但一般差别不大。

究竟采取哪种方法，取决于将时间理解为连续型变量(精确边际法及其近似)还是离散型变量(精确部分法)。

在实践中，如果风险集较大或同时失效的个体较多，则精确法的计算时间可能很长，故 Stata 的默认方法为 Breslow 近似法。

2. 估计基准风险函数

虽然 Cox 模型不设定基准风险 $\lambda_0(t)$ 的函数形式，但估计 β 之后依然可得到对 $\lambda_0(t)$ 的估计。

根据 $S(t) = \exp[-\Lambda(t)]$ ，代入比例风险 $\lambda(t; \mathbf{x}) = \lambda_0(t)e^{\mathbf{x}'\beta}$ 可得：

$$\begin{aligned}
S(t) &= \exp[-\Lambda(t)] = \exp\left[-\int_0^t \lambda(u; \mathbf{x}) du\right] = \exp\left[-\int_0^t \lambda_0(u) e^{\mathbf{x}'\boldsymbol{\beta}} du\right] \\
&= \exp\left[\left(-\int_0^t \lambda_0(u) du\right) \left(e^{\mathbf{x}'\boldsymbol{\beta}}\right)\right] = \left[\exp\left(-\int_0^t \lambda_0(u) du\right)\right]^{e^{\mathbf{x}'\boldsymbol{\beta}}} \equiv S_0(t)^{e^{\mathbf{x}'\boldsymbol{\beta}}}
\end{aligned}$$

其中, $S_0(t) \equiv \exp\left(-\int_0^t \lambda_0(u) du\right)$ 为对应于基准风险 $\lambda_0(t)$ 的基准生存函数。根据 Cox 模型, 已知 $\boldsymbol{\beta}$ 的估计量 $\hat{\boldsymbol{\beta}}$; 根据 KM 估计量, 已知 $S(t)$ 的估计量 $\hat{S}(t)$; 故可得基准生存函数 $S_0(t)$ 的估计量 $\hat{S}_0(t)$ 。

累积风险函数可写为:

$$\Lambda(t) = \int_0^t \lambda(u) du = \int_0^t \lambda_0(u) e^{\mathbf{x}'\boldsymbol{\beta}} du = e^{\mathbf{x}'\boldsymbol{\beta}} \int_0^t \lambda_0(u) du \equiv e^{\mathbf{x}'\boldsymbol{\beta}} \Lambda_0(t)$$

其中, $\Lambda_0(t) \equiv \int_0^t \lambda_0(u) du$ 为对应于基准风险 $\lambda_0(t)$ 的基准累积风险函数。从 KM 估计量 $\hat{S}(t)$ 出发, 根据 $\Lambda(t) = -\ln S(t)$, 可得到 $\Lambda(t)$ 的估计量 $\hat{\Lambda}(t)$ 。故根据上述方程可得到 $\hat{\Lambda}_0(t)$ 。

根据关系式 $\Lambda_0(t) \equiv \int_0^t \lambda_0(u) du$, 可计算每个失效时间对于 $\hat{\Lambda}_0(t)$ 的边际贡献, 即 $\hat{\lambda}_0(t)$ 。

使用核密度方法将此边际贡献光滑化, 即得到对 $\lambda_0(t)$ 的估计。

根据对基准风险的估计 $\hat{\lambda}_0(t)$ 以及参数估计 $\hat{\boldsymbol{\beta}}$, 可计算在 $\mathbf{x} = \mathbf{x}^*$ 处的风险函数:

$$\hat{\lambda}(t; \mathbf{x}^*) = \hat{\lambda}_0(t) e^{\mathbf{x}^{*'} \hat{\boldsymbol{\beta}}}$$

通常，令 $\mathbf{x}^* = \bar{\mathbf{x}}$ ，即计算在解释变量平均值处的风险函数；

也可选择感兴趣的其他解释变量取值(特别当 x 为二值变量或离散变量时)。

30.9 比例风险模型的设定检验

比例风险模型(包括 Cox 模型)的重要假设是 $\lambda(t; \mathbf{x}) = \lambda_0(t)e^{\mathbf{x}'\boldsymbol{\beta}}$ 。

如果此假设不成立，则比例风险模型不能成立。

需对比例风险模型进行设定检验，Stata 称为“PH-assumption tests”。

1. 对数-对数图(log-log plot)

根据比例风险假定 $\lambda(t; \mathbf{x}) = \lambda_0(t)e^{\mathbf{x}'\boldsymbol{\beta}}$ ，可将累积风险函数写为

$$\Lambda(t) = \int_0^t \lambda(u; \mathbf{x}) du = \int_0^t \lambda_0(u) e^{\mathbf{x}'\boldsymbol{\beta}} du = e^{\mathbf{x}'\boldsymbol{\beta}} \int_0^t \lambda_0(u) du \equiv e^{\mathbf{x}'\boldsymbol{\beta}} \Lambda_0(t)$$

其中， $\Lambda_0(t)$ 为对应于 $\lambda_0(t)$ 的累积风险函数。由上式可知，

$$\ln \Lambda(t) = \mathbf{x}'\boldsymbol{\beta} + \ln \Lambda_0(t)$$

代入 $\Lambda(t) = -\ln S(t)$ ，且方程两边同乘(-1)可得：

$$-\ln[-\ln S(t)] = -\mathbf{x}'\boldsymbol{\beta} - \ln[-\ln S_0(t)]$$

其中， $S_0(t)$ 为对应于 $\lambda_0(t)$ 的生存函数。由于 $0 < S(t) < 1$ ，故 $\ln S(t) < 0$ ， $-\ln S(t) > 0$ 。

函数 $-\ln[-\ln S(t)]$ 的斜率并不依赖于 \mathbf{x} 。

当 \mathbf{x} 取不同值时(比如， \mathbf{x} 为虚拟变量或离散变量)，函数 $-\ln[-\ln S(t)]$ 应为相互平行的曲线，只是截距项($-\mathbf{x}'\boldsymbol{\beta}$)不同。

据此画出的图被称为“对数-对数图”(log-log plot)。

如果对数-对数图中的曲线相互平行，则支持比例风险假设；如果不同曲线的斜率相差甚远，则意味着比例风险假设不成立。

对数-对数图的缺点是，如何确定曲线是否平行带有主观性。

2. 观测-预测图(observed versus expected plot)

检验比例风险假设的另一图示法为“观测-预测图”。针对每一解释变量 x ，分别画图。

假设 x 为离散变量，首先根据 x 的不同取值水平画其 Kaplan-Meier (KM)生存函数图，即观测图(observed plots)。

其次，估计 Cox 模型，以此计算基准生存函数 $S_0(t)$ ，然后代入 x 的不同取值水平，得到相应的生存函数图，即预测图(expected plot)。

最后，给定 x 的一个取值水平，比较其观测图与预测图是否足够接近。如果很接近，则表明变量 x 满足比例风险假设；反之，则比例风险假设不成立。观测-预测图的缺点同样是它的主观性。

3. 基于残差的检验(residual-based tests)

对于像久期分析这样的非线性模型，有多种残差的定义。

对于检验比例风险假设最有用的是“舍恩菲尔德残差”(Schoenfeld residuals)。

对于个体 j 与解释变量 x_k ($k = 1, \dots, K$, 假设共有 K 个解释变量), 可计算其对应的舍恩菲尔德残差如下:

$$r_{kj} = x_{kj} - \sum_{i \in R_j} \left(x_{ki} \cdot \frac{e^{x_i' \beta}}{\sum_{i \in R_j} e^{x_i' \beta}} \right)$$

其中, R_j 为当个体 j 失效时的风险集。

舍恩菲尔德残差 r_{kj} 为失效个体的解释变量观测值 x_{kj} 减去仍处于风险集中个体的解释变量之加权平均，而权重为“相对风险”(relative hazard) $e^{x_i'\beta}$ 。

如果比例风险假设成立，舍恩菲尔德残差不应随时间呈现出规律性的变化。

针对每个解释变量 x_k ，都可以将其舍恩菲尔德残差与时间画图，并考察其斜率是否为 0。

进一步，可以把舍恩菲尔德残差对时间回归，然后检验时间的系数是否为 0。

30.10 分层 Cox 模型

如果比例风险假设不满足，处理方法之一为“分层 Cox 模型”(stratified Cox model)。

不失一般性，假设变量 sex(性别)不满足比例风险假设，则可将样本中个体分为两组，第一组为男性，第二组为女性。

假设男性与女性的基准风险函数不同，但参数 $\boldsymbol{\beta}$ 都相同(此假设也可以检验)，个体的风险函数可写为

$$\lambda(t; \mathbf{x}_j) = \begin{cases} \lambda_{01}(t) e^{\mathbf{x}'_j \boldsymbol{\beta}}, & j \text{为男性} \\ \lambda_{02}(t) e^{\mathbf{x}'_j \boldsymbol{\beta}}, & j \text{为女性} \end{cases}$$

其中, $\lambda_{01}(t)$ 为男性组的基准风险, $\lambda_{02}(t)$ 为女性组的基准风险。

根据男性组的风险函数 $\lambda_{01}(t)e^{x_j'\beta}$, 可计算男性组的部分似然函数 L_1 ; 类似地, 可计算女性组的部分似然函数 L_2 。

整个样本的部分似然函数为 $L = L_1 \cdot L_2$ 。

由于变量 `sex` 不直接出现在似然函数中(`sex` 不满足比例风险假设, 无法直接引入 Cox 模型), 故无法估计其效应; 这是分层分析的代价。

进一步, 如果某分层变量有 k 个取值水平, 则可将样本中个体分为 k 层或 k 组, 然后进行分层估计。

更一般地，如果有多个分层变量，则应考虑它们的交叉情形。

比如，变量 x 有 3 种可能取值，变量 z 有 2 种可能取值，二者都不满足比例风险假设，故同为分层变量；此时，应将样本分为 6 组，使得每组内的 x, z 取值都相同。

分层 Cox 模型的一个重要假定为，不同组的参数 β 都相同。

为检验此假定，可以引入分层变量与留在 Cox 模型中变量的互动项进行分层分析，然后检验所有这些互动项的联合显著性。

如果都不显著，则支持“不同组参数 β 都相同”的原假设。

30.11 随时间而变的解释变量

如果比例风险假设不满足，处理方法之二为引入随时间而变的解释变量(Time-Varying Covariates, 简记 TVC)。

比如，在持续失业的过程中，宏观经济状况可能发生变化，从而影响就业概率。

引入 TVC 后，可能导致内生性。

不由个体控制的变量(比如宏观经济)是外生的；而个体可以影响的变量则可能为内生的。

例如，考察婚姻持续时间时，如果使用孩子的数量作为 TVC，可能导致内生性问题。由于稳定的婚姻会增加孩子的数量，故根据数据得出的“孩子越多，婚姻越持久”结论并不可靠。

下面，假设 TVC 为外生。

引入 TVC 后，可将风险函数写为

$$\lambda(t; \mathbf{x}(t)) = \lambda_0(t) e^{\mathbf{x}'(t)\boldsymbol{\beta}}$$

其中，解释变量 $\mathbf{x}(t)$ 可随时间而变，其系数 $\boldsymbol{\beta}$ 仍为固定。

一旦解释变量 $\mathbf{x}(t)$ 的取值变化，其风险函数也立即发生变化。

如要考虑滞后效应，可引入 $\mathbf{x}(t-1)$ 作为解释变量。

个体 i 与个体 j 的风险函数之比可写为：

$$\frac{\lambda(t; \mathbf{x}_i(t))}{\lambda(t; \mathbf{x}_j(t))} = \frac{\lambda_0(t) e^{\mathbf{x}_i'(t)\boldsymbol{\beta}}}{\lambda_0(t) e^{\mathbf{x}_j'(t)\boldsymbol{\beta}}} = e^{[\mathbf{x}_i(t) - \mathbf{x}_j(t)]'\boldsymbol{\beta}}$$

此风险函数之比是时间 t 的函数，不再满足比例风险假设，称为“扩展 Cox 模型” (extended Cox model)。

这对于基于似然函数(参数回归)或部分似然函数(Cox 模型)的估计并无实质影响。

这是因为，现实数据总是离散的(譬如年度数据、月度数据、日度数据)，故可将每位个体的单一记录分为几条记录，对应于几个时间段，使得在每个时间段内 TVC 变为常数，然后沿用解释变量不随时间而变的方法进行估计。

假设个体 1 从时间 0 至 T 一直处于失业状态，并在时间 T 找到工作。假设 $0 < t_1 < t_2 < T$ 。

共有两个解释变量，其中 x_1 不随时间而变(假设为虚拟变量，比如性别)；而 $x_2(t)$ 可以随时间而变，在区间 $[0, t_1)$ ， $[t_1, t_2)$ ， $[t_2, T)$ 的取值分别为 $x_2(t_1)$ ， $x_2(t_2)$ 与 $x_2(T)$ ，故为阶梯函数。则个体 1 的信息由三条记录构成，参见表 14.1。

表 14.1 个体 1 的相关记录

个体 ID	持续时间	x_1	$x_2(t)$	归并虚拟变量 d
1	t_1	1	$x_2(t_1)$	0
1	t_2	1	$x_2(t_2)$	0
1	T	1	$x_2(T)$	1

个体 1 的完整信息被分为三个部分，对应于三条记录。在第 1、2 条记录，个体 1 均未失效，而以归并结束，故归并虚拟变量为 0；在第 3 条记录，个体 1 在时间 T 失效，故归并虚拟变量为 1。

扩展 Cox 模型的用途之一是用来检验比例风险假设。

回到解释变量不随时间而变的情形，可将比例风险假设解释为风险函数 $\lambda(t; \mathbf{x}) = \lambda_0(t)e^{\mathbf{x}'\boldsymbol{\beta}}$ 的参数 $\boldsymbol{\beta}$ 不随时间而变。

考虑一维情形。假设 β 随时间而变，并将其写为 $\beta + \gamma g(t)$ ，其中 $g(t)$ 为时间的函数(比如， t 或 $\ln t$)，则风险函数可写为

$$\lambda(t; x) = \lambda_0(t)e^{x[\beta + \gamma g(t)]} = \lambda_0(t)e^{x\beta + \gamma xg(t)}$$

这相当于是原模型中引入了一个 TVC，即 $x \cdot g(t)$ ，由不随时间而变的解释变量 x 与时间的函数 $g(t)$ 的乘积构成，是经人为定义而生成的 TVC。

如果 $\gamma = 0$ ，则简化为标准 Cox 模型的风险函数。

可通过检验原假设“ $H_0: \gamma = 0$ ”来检验比例风险假设。

通常选择 $g(t)$ 为 t 或 $\ln t$ 。

在多维情形下，则引入所有解释变量 \mathbf{x} 与 $g(t)$ 的乘积，然后检验所有这些乘积项的系数联合显著性。

在作此检验时，由于引入了原模型中所没有的 TVC，故需要调整数据格式，将个体的信息根据需要分为多条记录。

30.12 不可观测的异质性

在进行久期分析时，样本中的个体可能具有“不可观测的异质性”(unobserved heterogeneity)，也称为“弱质”(frailty)。

比如，在研究病人的存活时间时，病人的体质或弱质并不能完全观察，但却影响病人的死亡风险。

在线性模型中，忽略不可观测的异质性相当于遗漏变量；如果遗漏变量与解释变量不相关，则不影响估计量的一致性。

在久期分析的非线性模型中，即使不可观测的异质性与解释变量不相关，仍会导致不一致。

【例】对于失业持续时间的研究。经验数据表明，失业时间越长，则找到工作的概率越低，即风险率随着时间而下降。

如果样本中的每位个体完全一样，则意味着负向久期依赖(negative duration dependence)。但事实可能并非如此。

假设失业人口可以分为两类，即 F 类(fast，其风险函数恒等于 0.4)与 S 类(slow，其风险函数恒等于 0.1)，这两类各占一半。

由于 F 类的风险率高于 S 类，故 F 类平均地比 S 类更快找到工作而退出样本，导致样本中 F 类占比随时间而下降，引起整体的风险率下降。

假设样本中有 200 人，其中 100 人为 F 类，另 100 人为 S 类。

在 100 位 F 类人中，第 1 期将有 40 人找到工作，第 2 期将有 24 人找到工作，而第 3 期将有 14.4 人找到工作。

在 100 位 S 类人中，第 1 期将有 10 人找到工作，第 2 期将有 9 人找到工作，而第 3 期将有 8.1 人找到工作。

就整个样本而言，第 1 期的风险率为 $(40+10)/200 = 0.25$ ，第 2 期的风险率为 $(24+9)/150 = 0.22$ ，第 3 期的风险率为 $(14.4+8.1)/117 = 0.192$ ，呈逐渐下降的趋势。

整体风险率的逐渐下降，只是由于 F 类与 S 类的结构变化而造成，而个体的风险函数始终为常数。

此效应被称为“弱质效应” (frailty effect)。

对于威布尔模型，风险函数为 $\lambda(t) = \gamma p t^{p-1}$ ，其中参数 p 决定风险函数的斜率，表示久期依赖的方向与程度。

如果存在不可观测的异质性而被忽略，则久期依赖参数 p 将被系统地低估，得不到一致估计。

解决方法是直接将不可观测的异质性引入模型中。

假设个体 j 的风险函数为

$$\lambda(t; \mathbf{x}_j, v_j) = \lambda_0(t) e^{\mathbf{x}_j' \boldsymbol{\beta}} v_j, \quad v_j > 0$$

其中，不可观测的异质性 $v_j > 0$ 以乘积的形式进入风险函数。 v_j 越大，则个体 j 越弱质，失效的风险越高。

$\lambda(t; \mathbf{x}_j, v_j)$ 称为“条件风险函数” (conditional hazard function)，因为它是给定 v_j 条件下的风险函数。

生存函数为：

$$S(t; \mathbf{x}_j, v_j) = e^{-\Lambda(t)} = \exp\left[-\int_0^t \lambda_0(u) e^{\mathbf{x}'_j \boldsymbol{\beta}} v_j du\right] = \left[\exp\left(-\int_0^t \lambda_0(u) e^{\mathbf{x}'_j \boldsymbol{\beta}} du\right)\right]^{v_j} = S(t; \mathbf{x}_j)^{v_j}$$

其中， $S(t; \mathbf{x}_j)$ 为不含异质性的生存函数。

由于 v_j 不可观测，故需要假设其概率分布，然后将 v_j 积分掉，得到无条件的风险函数与生存函数。

假设 v_j 的概率密度函数为 $g(v_j)$ ，则无条件风险函数与生存函数分别为

$$\lambda(t; \mathbf{x}_j) = \int_0^\infty \lambda(t; \mathbf{x}_j, v_j) g(v_j) dv_j$$

$$S(t; \mathbf{x}_j) = \int_0^\infty S(t; \mathbf{x}_j, v_j) g(v_j) dv_j$$

为使方程有解析解，常假设原模型为威布尔模型(指数模型为其特例)，而 v_j 服从 $Gamma(\delta, \delta)$ 分布或“逆正态分布”(Inverse Gaussian，简记 IG)；这类模型称为“混合模型”(mixture model)。

将 v_j 的期望值标准化为 1(故 $v_j = 1$ 代表平均弱质, average frailty), 并记其方差为 θ 。

如果 $\theta = 0$, 则 v_j 退化为常数, 不存在异质性;

可通过检验 “ $H_0: \theta = 0$ ” 来判断是否存在异质性。

由于不可观测的异质性不再出现在无条件风险函数与生存函数, 故可对其进行 MLE 估计。

还可考虑依某变量的取值将个体分组，然后假设组内个体为同质，而不同组的个体具有异质性，这被称为“共享异质性”(shared frailty)，即该异质性为同组内成员所共有。

上文的个体异质性则被称为“非共享异质性”(unshared frailty)。