Cox回归

在生存分析中，我们关注的是特定事件发生之前的时间这一随机变量的分布函数，例如在一次严重手术后死亡之前的时间、某设备故障之前的时间或前罪犯重新犯罪之前的时间。

多个因素可能会影响这个分布函数。例如，一名年轻女性在经历一次严重手术后死亡的概率可能低于年长女性，并且，如果前罪犯在重返社会后获得经济援助，相比于未获得援助的情况下，大概率他们会间隔一个更长的时间才重新犯罪。了解这些因素如何以及在多大程度上影响“生存期”是很重要的，这样我们可以确定更个性化的风险，并采取措施降低风险。例如，如果前罪犯在重返社会后面临经济困难时更容易重新犯罪，那么提供经济支持或帮助他们找到工作可能有助于这些人保持正确的生活轨迹。在这一应用中，我们将通过样本更深入地探讨生存分析。

我们接下来就使用这个例子来进行具体的说明.
前罪犯往往会重操旧业，并再次与警方和司法系统接触。假设我们想研究从释放到重新犯罪的时间分布函数，以及释放后提供经济支持是否对延长前罪犯安分守己的时间长度有积极影响。

首先，我们假设没有其他影响因素。

假设我们对 100 名前罪犯进行了为期一年的跟踪研究, 这里的时间设定是说跟踪每一个罪犯一年, 从其出狱开始。我们知道每个人在一年内是否重新犯罪，以及他们在释放后多少周重新犯罪。我们希望利用这些数据研究在 t 周内重新犯罪的前罪犯的百分比（$t \in [0, 52]$）。我们首先为这些数据建立一个统计模型。定义 Y_i^t 为指示变量，表示第 i 个前罪犯是否在 t 周内重新犯罪；如果没有犯罪，$y_i^t = 0$，如果犯罪了，$y_i^t = 1$。那么 $Y_1^t, \ldots, Y_{100}^t$ 是 Bernoulli 分布的，参数为 $p_t = P\left(Y_i^t = 1\right)$，即在 t 周内重新犯罪的概率。在假设 $Y_1^t, \ldots, Y_{100}^t$ 独立的情况下，统计模型已经固定。我们可以使用样本中的比例 $\sum_{i=1}^{100} y_i^t / 100$ 来“估计” p_t 的值。如果我们跟踪的前罪犯数量足够大（例如此处的 100 人），则根据大数定律，样本中找到的比例将接近真实比例 p_t。

研究通常会以另一种方式进行。我们可能不跟踪所有前罪犯一年，而是将研究的长度限制为一年, 这里所谓的研究有一个固定的持续时间, 正如之前所说, 超出研究时间段的数据可能会面临"删失"。我们跟踪在这一年中释放的罪犯，直到他们重新犯罪（如果他们犯罪了）或研究结束为止。我们在这样的研究中跟踪了 432 名前罪犯。下图显示了 5 名前罪犯的观察到的时间跨度。左图中的 x 轴代表从研究开始的时间（图中的竖线表示 0 周）到研究结束的时间（52 周的竖线）。y 轴上的数字是前罪犯的个人编号。第一个人在研究开始 10 周后释放，31 周后被逮捕。该人共自由了 $31 - 10 = 21$ 周。第二个人在研究开始 27 周后被释放，并在研究结束前未犯罪, 当然我们也不知道第二个人在研究结束后是否重新犯罪, 但我们知道的是在研究覆盖的52周的 $52 - 27 = 25$ 周内，他没有犯罪。对于第一个人，数据是完整的，而对于第二个人，我们只知道重新被逮捕的时间跨度的下限。我们称这种数据为**右删失数据**。

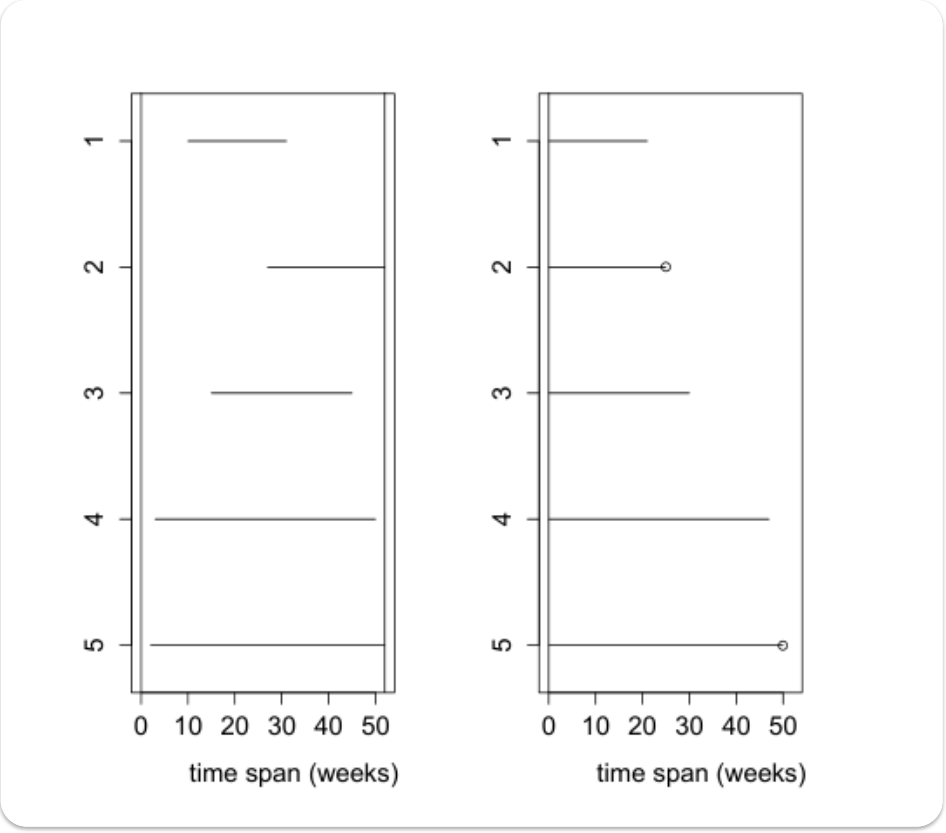

图 左图：5名前罪犯从释放到重新逮捕之间的时间（“时间跨度”）。x轴表示从研究开始的时间。右图：相同的数据，x轴表示从释放到重新逮捕或删失（研究结束或死亡）的时间。圆圈表示该人的数据是右删失的。

上面的右图显示了相同的信息，但方式不同。x轴现在表示从释放到重新逮捕的时间。对于在研究期间没有再次被逮捕的个体，我们只知道时间跨度的下限，这用一个小圆圈表示。

假设我们基于观察到的数据，想要估计在26周（半年）内重新被逮捕的前罪犯的百分比。例如，图中的第二个人在25周后仍然自由，但我们不知道25到26周之间的情况；这个人是右删失的。一个自然的解决方法是将所有右删失的人员从数据集中移除，并使用之前描述的相同估计方法。然而，这证明是一个糟糕的选择。一个前罪犯自由的时间越长，他被右删失的概率就越高，因此会被从数据集中移除。一个在释放51周后重新被逮捕的前罪犯只有在研究的第一周被释放时才会被纳入数据集。如果他在第二周被释放，他的数据将在研究结束时被删失。通过忽略右删失的个体，相对较多的长“生存期”将被移除，而在被简化的数据集中，前罪犯在t周内被重新逮捕的比例将（大大）偏高。例如，现在p_{51}的概率接近1。我们应该如何正确估计p_t呢？为了正确地做到这一点，我们首先为观察到的数据描述一个统计模型。

对于任意一个前罪犯，我们将从释放到重新犯罪的时间跨度定义为T。我们将T视为一个具有分布函数$t \mapsto F(t) = P(T \leq t)$和密度函数$f$的随机变量。生存函数$S$定义为$t \mapsto S(t) = 1 - F(t) = P(T > t)$，描述了在$t$周后尚未再次被逮捕的概率。我们可以假设分布函数或生存函数具有特定形式（例如，分布函数对应于指数分布或正态分布），但如果对分布形式没有先验知识，最好不要做任何假设。不正确的假设可能导致错误的结论。

对于一些前罪犯，在研究期间没有观察到时间跨度 T；在研究结束时，他们尚未被重新逮捕（或该人已经死亡）。因此，我们还为每个个体定义删失时间 C，即从释放到研究结束或死亡的时间跨度。如果 $T \leq C$，我们将观察到该个体重新犯罪；如果 $T \geq C$，我们观察到的不是 T，而是 C。因此，我们定义 $\tilde{T} = \min\{T, C\}$，这样我们就能观察到每个研究个体的 $\tilde{T}$。此外，我们定义指示函数 $\Delta = 1_{\{T \leq C\}}$；即，如果 $T \leq C$ 或 $\tilde{T} = T$，则 $\Delta = 1$，如果 $T > C$ 或 $\tilde{T} = C$，则 $\Delta = 0$。

我们观察到数据集中每个前罪犯的 $(\tilde{T}, \Delta)$ 对。数据集由432名前罪犯的 $\left(\tilde{t}_i, \delta_i\right)$ 的值组成，其中 $\tilde{t}_i$ 和 δ_i 是 $\tilde{T}_i$ 和 Δ_i 的观察值。

假设我们想估计任意一名前罪犯在26周内没有重新犯罪的概率，也就是说，我们想估计 $S(26)$。为了说明如何进行估计，我们假设现在我们只有5个人的数据，如图所示。我们有 $\tilde{t}_1 < \tilde{t}_2 < 26 < \tilde{t}_3 < \tilde{t}_4 < \tilde{t}_5$（参见上图）。第一个人是唯一一个在半年内被再次逮捕的。我们对第二个人知之甚少，他在半年内被删失。为了估计 $S(26)$，我们将 $S(26) = P(T > 26)$ 重写为

$$\begin{aligned} P(T > 26) &= P\left(T > 26 \mid \tilde{T} > \tilde{t}_1\right) P\left(\tilde{T} > \tilde{t}_1\right) \\ &= P\left(T > 26 \mid \tilde{T} > \tilde{t}_2, \tilde{T} > \tilde{t}_1\right) P\left(\tilde{T} > \tilde{t}_2 \mid \tilde{T} > \tilde{t}_1\right) P\left(\tilde{T} > \tilde{t}_1\right) \\ &= P\left(T > 26 \mid \tilde{T} > \tilde{t}_2\right) P\left(\tilde{T} > \tilde{t}_2 \mid \tilde{T} > \tilde{t}_1\right) P\left(\tilde{T} > \tilde{t}_1\right) \end{aligned}$$

需要注意的一点是在这个样本空间中 $\tilde{T} > \tilde{t}_1$ 的情况不对 $T > 26$ 产生影响. 所以第一个等号成立, 之后同理有第二个第三个等号.
我们不是直接估计$S(26)$，而是分别估计最后一行的三个因素。我们从最后开始，估计概率 $P\left(\tilde{T} > \tilde{t}_1\right)$。第一个人在第21周重新犯罪。在这五个人中，有四个人满足$\tilde{T} > \tilde{t}_1$。因此我们估计 $P\left(\tilde{T} > \tilde{t}_1\right)$的概率为$4/5$。为了估计第二个因素$P\left(T > \tilde{t}_2 \mid \tilde{T} > \tilde{t}_1\right)$，我们只使用满足条件$\tilde{T} > \tilde{t}_1$ 的个体的数据，这些个体是第2至第5个。第二个人在$\tilde{t}_2$时被删失（研究结束），因此在剩下的四名前罪犯中，只有三个人满足$\tilde{T} > \tilde{t}_2$。我们估计$P\left(\tilde{T} > \tilde{t}_2 \mid \tilde{T} > \tilde{t}_1\right)$的概率为$3/4$。我们以类似方式估计最后一个概率$P\left(T > 26 \mid \tilde{T} > \tilde{t}_2\right)$。对于满足$\tilde{T} > \tilde{t}_2$的三个人，他们的$T$值都大于26；因此我们估计这个概率为$3/3 = 1$。将这三个估计值相乘得到0.6。

这个方法可以进一步的推广计算别的概率或基于其他数据集进行估计。当我们以这种方式在0到52周之间估计函数 S 时，我们会得到一个阶梯函数，该函数仅在 $\delta_i = 1$ 的点 t_i 处跳跃（向下）。下图显示了基于完整数据集估计的生存曲线 S。如果观察数量增加，曲线保持常数的区间将会变短，跳跃的幅度也会减小。我们可以证明，如果观察数量趋于无穷大，在时间 t 的 S 估计值将（以概率）收敛于真实值 $S(t)$。如果我们假设 F 是一个已知的连续函数，例如对应于指数分布的分布函数，那么我们将不使用上述方法来估计 F，而是使用（参数化的）最大似然法。此时，生存曲线将由一个连续递减的曲线估计。然而，如果关于曲线形式的假设是错误的, 比如把连续性搞错了，那么在任意点 t 处的 S 估计值将不会（以概率）收敛于 $S(t)$。

下图还显示了如果移除所有被删失的数据所得到的估计曲线；在估计 $S(t)$ 时，所有在时间 t 之前被删失的个体都被移除。随着 t 的增加，两个曲线之间的差距增大。这是因为 t 越大，移除的值越多，所产生的误差也越大。这种方法会导致对生存曲线的低估。

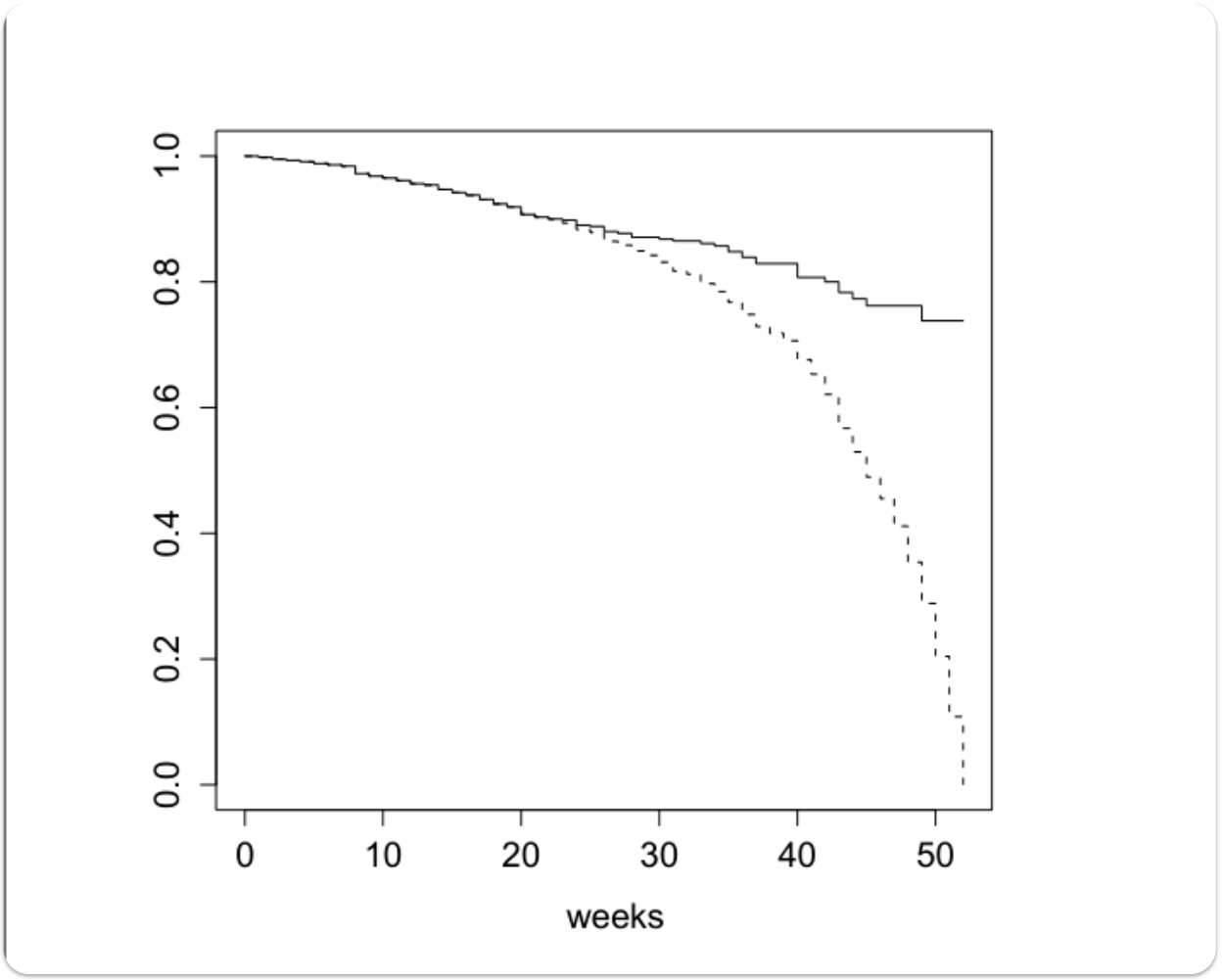

图:基于所有数据的估计生存曲线（实线）与移除删失观测值后数据集的估计曲线（虚线）。

另一种表示T的分布的方法是使用所谓的风险函数。与概率密度f和分布函数F相关的风险函数定义为:

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

如果我们将$f(t)dt$视为T位于区间$[t, t + dt)$的概率，那么$\lambda(t)dt$可以解释为:

$$\lambda(t)dt \approx \frac{P(t \leq T < t + dt)}{P(T > t)} = P(t \leq T < t + dt \mid T > t)$$

因此，$\lambda(t)$ 的值可以看作是在时间 t 时前罪犯尚未被重新逮捕的情况下，紧接着 t 时刻重新犯罪的条件概率。由于这种作为"瞬时概率"的解释，风险函数常用于建模生存数据。注意到一个数学上的处理是,风险函数是函数 $t \mapsto -\log(1 - F(t))$ 关于 t 的导数，给定风险函数 λ，我们可以通过公式 $F(t) = 1 - e^{-\Lambda(t)}$（其中 Λ 是累积风险函数，即 $\Lambda(t) = \int_0^t \lambda(s)ds$）恢复分布函数 F。密度 f 等于 $f(t) = \lambda(t)e^{-\Lambda(t)}$。对于生存曲线，我们现在可以假设风险函数具有特定形式。例如，如果我们假设风险函数是常数 $\lambda(t) \equiv \nu$，则相应的密度为 $f(t) = \lambda(t)e^{-\Lambda(t)} = \nu e^{-\nu t}$；换句话说，$T$ 服从指数分布。我们也可以不对形式做任何假设。为了估计 T 的分布函数，我们也可以借助风险函数和上述公式。

如果年龄、性别、教育等因素可能影响时间跨度，那么将这些因素纳入模型是明智的。通常，选择所谓的Cox模型。在该模型中，第i个前罪犯的风险函数与观察到的变量$X_i = x_i$相关，其形式为:

$$\lambda\left(t \mid X_i = x_i\right) = e^{\beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_K x_{iK}} \lambda_0(t)$$

其中x_{ik}是第i个人的第k个变量的值，K是模型中的变量数量。函数$t \mapsto \lambda_0(t)$称为基线风险函数，它等于所有预测变量都为0时的风险函数。根据Cox模型，具有预测变量x_i和x_j的两个前罪犯的风险函数是成比例的，这意味着：

$$\frac{\lambda\left(t \mid X_i = x_i\right)}{\lambda\left(t \mid X_j = x_j\right)} = e^{\beta^T (x_i - x_j)}$$

这个比例不依赖于 t。

这为参数 β 提供了一个简单的解释：它决定了与某些预测变量相关的相对风险大小。例如，假设两个前罪犯在所有预测变量上的得分相同，除了经济支持这个变量。前罪犯 i 接受了经济支持 $(x_{i1} = 1)$，而前罪犯 j 没有接受 $(x_{j1} = 0)$。那么，风险函数的比率简化为

$$\frac{\lambda\left(t \mid X_i = x_i\right)}{\lambda\left(t \mid X_j = x_j\right)} = e^{\beta_1 (x_{i1} - x_{j1})} = e^{\beta_1}$$

如果 β_1 的值为 -0.400，那么第 i 个前罪犯相对于第 j 个前罪犯的风险被再次逮捕的风险是 $e^{-0.400} = 0.670$。可以发现一个非常自然的点是相对风险因此与前罪犯释放后的时间无关。

在我们的例子中，Cox 模型包括以下预测变量：是否获得经济支持、年龄、种族、婚姻状况、先前定罪次数。基于 432 名前罪犯的数据，我们可以估计回归参数 $\beta = (\beta_1, \ldots, \beta_5)$ 和未知的基线风险函数 λ_0。第 7 章中会详细说明适用的方法。这里，我们只给出结果。表 1 显示了回归参数 $\beta_1, \ldots, \beta_5$ 的估计值。

	β_1	β_2	β_3	β_4	β_5
estimate	-0.400	-0.0425	0.282	-0.590	0.0977
exp (estimate)	0.670	0.958	1.326	0.554	1.103

表 1.1. $\beta_1, \ldots, \beta_5$ 的估计值；参数分别对应：是否获得经济支持（0：没有支持，1：有支持）、年龄、种族（0：其他；1：黑人）、婚姻状况（0：未婚，1：已婚）、先前定罪次数。

回归参数 β_1 的估计值为负，表明释放后获得经济支持有积极的效果。然而，已婚状态似乎比接受经济援助有更强的积极效果。如果这一影响是因果关系，那么帮助前罪犯找到伴侣可能比帮助他们找到工作更明智。

在上述分析中，我们对模型做出了一些假设。例如，我们假设了两个前罪犯的风险函数是成比例的，并且预测变量是可加的。当然，这些假设必须进行验证。例如，我们可以通过绘制合适的图表进行验证。更多信息可以参考相关文献。