生存分析

朱宇嘉

目录

1 生存数据

生存数据，又称为删失数据，是一种特殊的数据形式，其可由一个二元组表示。假设生存数据为 $\{(t_i, \delta_i)\}_{i=1}^N$，则对于 $\forall i = 1, 2, ..., N$，对于个体 i 的生存时间 T_i，我们有：

- 若 $\delta_i = 1$，则 $T_i = t_i$
- 若 $\delta_i = 0$，则 $T_i > t_i$(注意：此时表示的是右删失)

生存数据不是缺失数据。这是因为对于缺失数据，我们不仅没有观察到该数据的取值，而且我们没有关于该数据的任何信息。但对于删失数据，我们虽然没有观察到数据的取值，但我们拥有该数据的信息（例：我们仅知道 $T > 8$ 但不知道 T 的具体数值）。

如果我们将自变量也加入到数据中，则我们可将生存数据变为一个三元组 $\{(t_i, \delta_i, x_i)\}_{i=1}^N$。我们可以使用数据研究自变量 X 对生存时间 T 的影响。

1.1 生存时间 T 的分布

生存时间 T 是一个非负的随机变量，其表示个体的生存时间，除去概率密度函数 $f(t)$ 和累计分布函数 $F(t)$，对于生存分析还存在两个重要的分布函数。

生存函数 $S(t)$ 表示个体生存时间超过 t 的概率，$S(t) = P(T > t) = 1 - F(t)$。

风险函数 $\lambda(t)$ 表示个体在 t 时刻仍生存的情况下，该个体在时间 t 的去世或失效的程度，

$$\begin{aligned}\lambda(t) &:= \lim_{h\to 0^+} \frac{P(t \le T \le t+h | T \ge t)}{h} \\ &= \lim_{h\to 0^+} \frac{1}{h}\frac{P(t \le T \le t+h)}{P(T \ge t)} = \lim_{h\to 0^+} \frac{1}{hS(t)}\int_t^{t+h} f(s)ds = \frac{f(t)}{S(t)}\end{aligned}$$

累计风险函数 $\Lambda(t)$ 可以表示个体到当前时刻累计的风险，

$$\Lambda(t) = \int_0^t \lambda(s)ds$$

$f(t), S(t), \lambda(t), \Lambda(t)$ 之间存在如下关系：

- $f(t) = -S'(t), S(t) = \int_t^\infty f(s)ds$
- $\lambda(t) = -\frac{d}{dt}\log S(t)$
- $S(t) = \exp(-\Lambda(t)), \Lambda(t) = -\log S(t)$

平均剩余寿命是指个体在活过 t 后的平均剩余时间，即

$$
\begin{aligned}
r(t) :=& E(T-t|T\geq t)\\
=& \int_0^{\infty}(T-t)f_T(T|T\geq t)dT\\
=& \int_0^{\infty}(T-t)\frac{f(T)}{S(t)}dT\\
=& \frac{1}{S(t)}\int_0^{\infty}(T-t)dF(T)\\
=& \frac{1}{S(t)}\int_0^{\infty}(t-T)dS(T)\\
=& \frac{1}{S(t)}\left[(t-T)S(T)|_t^{\infty}+\int_t^{\infty}S(T)dT\right]\\
=& \frac{\int_t^{\infty}S(u)du}{S(t)}
\end{aligned}
$$

1.2 删失模式

假设我们现在拥有的是完全观测的数据 $t_1, t_2, ..., t_N$，则使用该组数据的似然函数为

$$L(\lambda)=\prod_{i=1}^{N}f(t_i|\lambda)$$

其中，λ 的估计可由 $\hat{\lambda}=\arg\max_{\lambda}L(\lambda)$。注意到之所以可以用 $f(t_i|\lambda)$ 表示 t_i 的贡献，是因为

$$f(t_i|\lambda)\approx\lim_{\Delta t\to 0}P(t_i-\Delta t<T<t_i+\Delta t|\lambda)$$

因此，对于以下删失机制，需要对似然函数做对应的改变

1. 右删失：在删失情况下，仅知道 $T>t_i$，需要用 $S(t_i|\lambda)=P(T>t_i|\lambda)$ 代替 $f(t_i|\lambda)$
2. 左删失：在删失情况下，仅知道 $T<t_i$，需要用 $F(t_i|\lambda)=P(T<t_i|\lambda)$ 代替 $f(t_i|\lambda)$
3. 区间删失：在删失情况下，仅知道 $T\in[L,U]$，需要用 $F(U|\lambda)-F(L|\lambda)=P(L<T<U|\lambda)$ 代替 $f(t_i|\lambda)$
4. 双侧删失：在删失情况下，仅知道 $T>t_i$ 还是 $T<t_i$，需要用 $F(t_i|\lambda)$ 或 $S(t_i|\lambda)$ 代替 $f(t_i|\lambda)$

其中，右删失在实际生活中最为常见。因此，在后续讨论中，我们主要专注于右删失数据。

2 生存函数的非参数估计

我们将使用数据 $\{t_i,\delta_i\}_{i=1}^N$ 对生存函数 $S(t)$ 进行估计。如果数据不存在删失，则我们可以通过经验估计得到 $S(t)$ 的估计值 $\hat{S}(t)=\hat{P}(T>t)=E[I(T>t)]=\frac{\#\{i:t_i>t\}}{N}$。不过当数据存在删失时，该方法便不再奏效。

2.1 Kaplan-Meier 估计

对于删失数据，Kaplan-Meier 估计是最为常用的估计方法。K-M 估计是一种非参数估计，其可通过最大化极大似然函数获得。假设我们的数据是右删失的，则数据 $\{t_i, \delta_i\}_{i=1}^N$ 的极大似然函数为

$$L(S) = \prod_{i=1}^{N} P(T_i = t_i)^{\delta_i} P(T_i > t_i)^{1-\delta_i} = \prod_{i=1}^{N} \left(S(t_i^-) - S(t_i)\right)^{\delta_i} S(t_i)^{1-\delta_i}$$

则生存函数可通过最大化极大似然函数获得：$\hat{S}(t) = \arg\max_S L(S)$。注意到 $L(S)$ 是极大似然函数，其输入也是一个函数 S。为了极大化似然函数 $L(S)$，我们需要首先承认两点：

1. 对于每个 $\delta_i = 1$ 的时刻 t_i，$S(t_i^-) - S(t_i) > 0$。否则 $S(t_i^-) - S(t_i) = 0, L(S) = 0$，似然函数无法取得最大值。
2. 对于未有观测到失效事件的时刻 (即 $\delta_i \neq 1$ 的时刻)，$S(t_i^-) - S(t_i) = 0$。否则 $S(t)$ 的取值会在未被观测到失效事件的时刻分走一部分，导致 $L(S)$ 的减小。

通过以上分析，我们知道，使得 $L(S)$ 极大化的估计 $\hat{S}$ 必然只在观测到失效事件的时刻点取值有下降，在其余时刻 $\hat{S}$ 取值不发生改变。下面定义一些符号，假定在数据集中共有 J 个失效时刻，分别为 $0 = t_0 < t_1 < ... < t_J < t_{J+1} = \infty$，记

- d_j 为时刻 t_j 时失效的个体数量
- c_j 为时刻 $[t_j, t_{j+1})$ 时间段内删失的个体数量
- n_j 为时刻 t_j^- 时在险的个体数量

因此，

$$L(S) = \prod_{i=1}^{N} \left(S(t_i^-) - S(t_i)\right)^{\delta_i} S(t_i)^{1-\delta_i} = \prod_{j=1}^{J} \left(S(t_i^-) - S(t_i)\right)^{d_j} S(t_i)^{c_j}$$

按照之前的分析，S 仅在 $t_1, t_2, ..., t_J$ 上有概率密度，因此可将 $S(t)$ 看做离散的生存函数，因此可将生存函数表示为 $S(t) = \prod_{t_j \le t}(1 - \lambda_j)$。因此有

$$S(t_j^-) = \prod_{i=1}^{j-1}(1 - \lambda_i), S(t_j) = \prod_{i=1}^{j}(1 - \lambda_i)$$

代入 $L(S)$，我们有

$$
\begin{aligned}
L(S) &= \prod_{j=1}^{J} \left(S(t_i^-) - S(t_i)\right)^{d_j} S(t_i)^{c_j} \\
&= \prod_{j=1}^{J} \left[\left[\prod_{i=1}^{j-1}(1-\lambda_i) - \prod_{i=1}^{j}(1-\lambda_i) \right]^{d_j} \left[\prod_{i=1}^{j}(1-\lambda_i) \right]^{c_j} \right] \\
&= \prod_{j=1}^{J} \left[\left[\left(\frac{1}{1-\lambda_j} - 1 \right) \prod_{i=1}^{j}(1-\lambda_i) \right]^{d_j} \left[\prod_{i=1}^{j}(1-\lambda_i) \right]^{c_j} \right] \\
&= \left(\prod_{j=1}^{J} \frac{\lambda_j^{d_j}}{(1-\lambda_j)^{d_j}} \right) \left(\prod_{j=1}^{J} \prod_{i=1}^{j} (1-\lambda_i)^{d_j} \right) \left(\prod_{j=1}^{J} \prod_{i=1}^{j} (1-\lambda_i)^{c_j} \right) \\
&= \left(\prod_{j=1}^{J} \frac{\lambda_j^{d_j}}{(1-\lambda_j)^{d_j}} \right) \left(\prod_{j=1}^{J} \prod_{i=1}^{j} (1-\lambda_i)^{d_j+c_j} \right) \\
&= \left(\prod_{j=1}^{J} \frac{\lambda_j^{d_j}}{(1-\lambda_j)^{d_j}} \right) \left(\prod_{i=1}^{J} \prod_{j=i}^{J} (1-\lambda_i)^{d_j+c_j} \right) \\
&= \left(\prod_{j=1}^{J} \frac{\lambda_j^{d_j}}{(1-\lambda_j)^{d_j}} \right) \left(\prod_{i=1}^{J} (1-\lambda_i)^{n_i} \right) \\
&= \prod_{j=1}^{J} \lambda_j^{d_j} (1-\lambda_j)^{n_j-d_j}
\end{aligned}
$$

因此，$\hat{\lambda}_j = \frac{d_j}{n_j}$，将其代入 $S(t)$ 的表达式，即可得到 Kaplan-Meier 统计量

$$
\hat{S}(t) = \prod_{t_j \le t} (1-\hat{\lambda}_j) = \prod_{t_j \le t} \frac{n_j - d_j}{n_j}
$$

需要注意，$\hat{S}(t)$ 仅仅在观察到的失效时间进行相乘，在删失的时间点不进行相乘，删失的数据仅会改变 n_j 的取值，从而影响 $\hat{S}(t)$ 的取值。

2.2 Kaplan-Meier 估计的置信区间

对数形式 由于 $\hat{S}(t) = \prod_{t_j \le t}(1-\hat{\lambda}_j)$，即 $\hat{S}(t)$ 是一个关于 $\hat{\lambda}_j$ 的函数。因此计算 $\hat{S}(t)$ 的方差可以转化为计算 $\hat{\lambda}_j$ 的方差。注意到得到 $\hat{\lambda}_j$ 是由最大化极大似然函数得到的，而 λ_j 在似然函数中的形式为 $\lambda_j^{d_j}(1-\lambda_j)^{n_j-d_j}$，其与二项分布的形式完全一致。根据二项分布的渐近性质，我们有

$$
\sqrt{n_j}(\hat{\lambda}_j - \lambda_j) \to^D N(0, \lambda_j(1-\lambda_j))
$$

因此

$$
\mathrm{Var}(\hat{\lambda}_j) = \frac{\lambda_j(1-\lambda_j)}{n_j} \approx \frac{\hat{\lambda}_j(1-\hat{\lambda}_j)}{n_j} = \frac{d_j(n_j-d_j)}{n_j^3}
$$

根据 Δ-方法 (若 $\sqrt{n}(X_n-\theta)\to^D N(0,\sigma^2)$，则 $\sqrt{n}(g(X_n)-g(\theta))\to^D N(0,g'^2(\theta)\sigma^2)$)，我们有

$$\mathrm{Var}(\log(1-\hat{\lambda}_j))=\left(\frac{1}{1-\lambda_j}\right)^2\mathrm{Var}(\hat{\lambda}_j)=\left(\frac{1}{1-\lambda_j}\right)^2\frac{\lambda_j(1-\lambda_j)}{n_j}=\frac{\lambda_j}{n_j(1-\lambda_j)}\approx\frac{\hat{\lambda}_j}{n_j(1-\hat{\lambda}_j)}=\frac{d_j}{n_j(n_j-d_j)}$$

若假定 λ_j 之间相互独立，则有

$$\mathrm{Var}(\log\hat{S}(t))=\mathrm{Var}\left(\sum_{t_j\le t}\log(1-\hat{\lambda}_j)\right)\approx\sum_{t_j\le t}\mathrm{Var}(\log(1-\hat{\lambda}_j))=\sum_{t_j\le t}\frac{d_j}{n_j(n_j-d_j)}$$

记 $\hat{s}^2(t)=\sum_{t_j\le t}\frac{d_j}{n_j(n_j-d_j)}$，由 Δ-方法，我们有

$$\sqrt{n}(\log\hat{S}(t)-\log S(t))\to^D N(0,\sigma_t^2)$$

其中，$\sigma_t^2=n\hat{s}^2(t)$。因此，$\log S(t)$ 的 95% 区间为 $[\log\hat{S}(t)-1.96\hat{s}(t),\log\hat{S}(t)+1.96\hat{s}(t)]$。$S(t)$ 的 95% 区间为 $[\hat{S}(t)e^{-1.96\hat{s}(t)},\hat{S}(t)e^{1.96\hat{s}(t)}]$。

Greenwood 形式 由于 $\sqrt{n}(\log\hat{S}(t)-\log S(t))\to^D N(0,\sigma_t^2)$，由 Δ-方法，有

$$\sqrt{n}(\hat{S}(t)-S(t))\to^D N(0,S^2(t)\sigma_t^2)$$

因此，$S(t)$ 的 95% 区间为 $[\hat{S}(t)-1.96\hat{S}(t)\hat{s}(t),\hat{S}(t)+1.96\hat{S}(t)\hat{s}(t)]$。

Log-log 形式 Log-log 变换为 $g(\theta)=\log(-\log(\theta))$，对于任意一个实数范围内的置信区间 C，均有 $g^{-1}(C)\in[0,1]$。$[0,1]$ 是生存函数 $S(t)$ 应该属于的范围，因此可考虑使用 log-log 函数进行逆变换，求得符合条件的置信区间。同样使用 Δ-方法，有

$$\sqrt{n}\left(g(\hat{S}(t))-g(S(t))\right)\to^D N\left(0,\frac{\sigma_t^2}{(\log S(t))^2}\right)$$

因此，$\log(-\log S(t))$ 的 95% 区间为 $[\log(-\log\hat{S}(t))-1.96\frac{\hat{s}(t)}{\log\hat{S}(t)},\log(-\log\hat{S}(t))+1.96\frac{\hat{s}(t)}{\log\hat{S}(t)}]$。变换后有 $S(t)$ 的 95% 区间为 $[\hat{S}(t)^{\exp\left(-1.96\frac{\hat{s}(t)}{\log\hat{S}(t)}\right)},\hat{S}(t)^{\exp\left(1.96\frac{\hat{s}(t)}{\log\hat{S}(t)}\right)}]$。

总结 对于 K-M 估计的置信区间，共有三种计算方法。置信区间的计算都基于

$$\hat{S}(t)=\prod_{t_j\le t}\frac{n_j-d_j}{n_j},\hat{s}^2(t)=\sum_{t_j\le t}\frac{d_j}{n_j(n_j-d_j)}$$

估计方法	95% 置信区间
Log 形式	$[\hat{S}(t)e^{-1.96\hat{s}(t)}, \hat{S}(t)e^{1.96\hat{s}(t)}]$
Greenwood 形式	$[\hat{S}(t)-1.96\hat{S}(t)\hat{s}(t), \hat{S}(t)+1.96\hat{S}(t)\hat{s}(t)]$
Log-log 形式	$\left[\hat{S}(t)^{\exp\left(-1.96\frac{\hat{s}(t)}{\log\hat{S}(t)}\right)}, \hat{S}(t)^{\exp\left(1.96\frac{\hat{s}(t)}{\log\hat{S}(t)}\right)}\right]$

Log-log 形式的优势是其计算出的置信区间一定落在 $[0,1]$ 区间内，而 Log 形式和 Greenwood 形式计算出的置信区间会落在 $[0,1]$ 外，需要通过截断来获得最终的置信区间。Log 形式在实际的估计中最为常用，因为在其方差估计中，仅含有 $\hat{s}(t)$ 而不含有 $\hat{S}(t)$，得到的置信区间较为准确。

3 Log-rank 检验

使用 Kaplan-Meier 估计，我们可以看出不同组之间的生存曲线的走势和重合度，但我们不能在统计意义上对两组生存曲线是否相同进行判断。

3.1 两组判断

我们有两组数据 $\{t_{1i},\delta_{1i}\}_{i=1}^{n_1}$ 和 $\{t_{2i},\delta_{2i}\}_{i=1}^{n_2}$。我们希望做检验：$H_0: S_1(t)=S_2(t)$。我们将两组数据中的失效时间汇总，假设汇总后的失效事件排序后为 $0=t_0<t_1<...<t_J<t_{J+1}=\infty$。定义

- d_{1j}, d_{2j} 分别为时刻 t_j 时第一组和第二组失效的个体数量
- c_{1j}, c_{2j} 为时刻 $[t_j, t_{j+1})$ 时间段内第一组和第二组删失的个体数量
- n_{1j}, n_{2j} 为时刻 t_j^- 时第一组和第二组在险的个体数量

在时刻 t_j，我们对第一组和第二组失效和在险的数量做出列联表，

	第一组	第二组	总和
失效个数	d_{1j}	d_{2j}	d_j
存活个数	$n_{1j}-d_{1j}$	$n_{2j}-d_{2j}$	n_j-d_j
总和	n_{1j}	n_{2j}	n_j

在 $S_1(t)=S_2(t)$ 的原假设下，d_{1j} 服从一个超几何分布 $d_{1j}\sim \mathrm{HG}(n_j, n_{1j}, d_j)$。因此 d_{1j} 的期望与方差分别为

$$E(d_{1j}) := e_{1j} = n_{1j}\frac{d_j}{n_j}$$
$$\mathrm{Var}(d_{1j}) := v_{1j} = n_{1j}\frac{d_j}{n_j}\frac{n_j - d_j}{n_j}\frac{n_j - n_{1j}}{n_j - 1}$$

d_{1j} 的和 $\sum_j d_{1j}$ 并不服从某一特定的分布，但我们可以用正态分布进行近似，得到 $\sum_j d_{1j}$ 的近似分布。对于 $w_j = d_{1j} - e_{1j}$，其近似服从 $w_j \sim N(0, v_{1j})$。因此，在原假设下，$W = \sum_{j=1}^{J} w_j$ 近似服从正态分布

$$W \sim^D N(0, V)$$

其中，$V = \sum_j v_{1j}$。因此，检验 $H_0 : S_1(t) = S_2(t)$ 可以使用统计量 $W \sim^D N(0, V)$ 或 $\frac{W^2}{V} \sim^D \chi^2(1)$ 进行检验。该检验方法即为 **Log-rank** 检验，log-rank 检验是生存分析中判断组之间生存函数是否相同最常用的方法。

加权 Log-rank 检验

Log-rank 检验的统计量具有形式：$\frac{(\sum_j w_j)^2}{\sum_j v_j}$。该形式表现了对于所有有失效事件的时刻，统计量给予的权重相同。我们可以将其扩展到加权的 log-rank 统计量，具有如下形式

$$\frac{(\sum_j \alpha_j w_j)^2}{\sum_j \alpha_j^2 v_j}$$

其中 α_j 为权重。有以下两种常见的加权方法：

1. 设置 $\alpha_j = n_j$。这种方法为 Gehan-Breslow 检验法。
2. 设置 $\alpha_j = \hat{S}(t_j)$。这种方法为 Peto 检验法。

这两种检验方法都对失效时间晚的个体给予了更大的权重。

分层 Log-rank 检验 有时候，我们害怕在检验 $S_1(t) = S_2(t)$ 时，该检验结果会受到混淆因子的干扰，从而导致检验结果的不正确。举个例子，假设存在变量 C，其共有 3 个取值 c_1, c_2, c_3，在水平 c_1, c_2 上有 $S_1(t) = S_2(t)$，但在水平 c_3 上则有 $S_1(t) \neq S_2(t)$。我们需要使用分层 Log-rank 检验对以上情况进行检验。分层 Log-rank 检验对混淆因子的每个水平进行拆分，假定 K 为混淆因子，我们得到以下统计量

$$\frac{W^2}{V} = \frac{(\sum_k \sum_j w_{jk})^2}{\sum_k \sum_j v_{jk}} \sim \chi^2(1)$$

3.2 多组判断

假设现在有 $K+1$ 组数据，其中一组为基准组或对照组。定义向量 $w_j = (d_{1j} - e_{1j}, ..., d_{Kj} - e_{Kj})^T$，$w_j$ 服从一个多元超几何分布，其期望为 $E(w_j) = 0$，方差矩阵为 V_j，其对角元素和非对角元素分别为

$$
\begin{aligned}
(V_j)_{ii} &= n_{ij}\frac{d_j}{n_j}\frac{n_j-n_{ij}}{n_j}\frac{n_j-d_j}{n_j-1} \\
(V_j)_{ik} &= -d_j\frac{n_{ij}}{n_j}\frac{n_{kj}}{n_j}\frac{n_j-d_j}{n_j-1}
\end{aligned}
$$

因此，定义 $w=\sum_j w_j, V=\sum_j V_j$，我们有

$$
w^T V^{-1} w \sim^D \chi^2(K)
$$

使用以上统计量即可完成多组的 log-rank 检验。事实上，将 $K+1$ 组都计算进 w, V 也是可以的，但在这种情况下 V 会产生不可逆的情况，我们需要使用广义逆代替。

4 生存数据的参数推断

我们对生存数据 $\{(t_i,\delta_i,x_i)\}_{i=1}^N$ 构建似然函数。假设 T 的概率密度函数为 $f(t;\theta)$，生存函数为 $S(t;\theta)$。则，对于数据 $\{(t_i,\delta_i)\}_{i=1}^N$，其似然函数为

$$
L(\theta)=\prod_{i=1}^N f(t_i;\theta)^{\delta_i} S(t_i;\theta)^{1-\delta_i}
$$

从似然函数中我们可以看到：对于观测到的失效事件，我们可直接像一般构建似然函数时直接使用概率密度函数；对于删失的事件，由于我们仅知道 $T>t$，因此我们需要使用生存函数代替概率密度函数。对数似然函数为

$$
l(\theta)=\sum_{i=1}^N [\delta_i \ln f(t_i;\theta) + (1-\delta_i)\ln S(t_i;\theta)]
$$

对对数似然函数求导即可得到得分函数

$$
U=\sum_{i=1}^N U_i = \sum_{i=1}^N \frac{\partial l_i(\theta)}{\partial \theta} = \sum_{i=1}^N \left[\delta_i \frac{\partial \ln f(t_i;\theta)}{\partial\theta} + (1-\delta_i)\frac{\partial \ln S(t_i;\theta)}{\partial\theta}\right]
$$

定理：对于生存数据，其得分函数仍满足 $E(U)=0, \mathrm{Var}(U)=E(U^2)=-E(U')=I$。

由于 t_i,δ_i 是删失性数据，这导致 t_i,δ_i 两者是不独立的。我们将其转换为独立的变量，假设第 i 个个体的真实寿命为 $\tilde{t}_i$，删失时刻为 c_i，则我们有 $t_i=\min(\tilde{t}_i,c_i), \delta_i=I(\tilde{t}_i\le c_i)$，$\tilde{t}_i, c_i$ 两个变量可以看做互相独立的。代入得分函数

$$
U_i = I(\tilde{t}_i \le c_i)\frac{\partial \ln f(\tilde{t}_i;\theta)}{\partial\theta} + I(\tilde{t}_i > c_i)\frac{\partial \ln S(c_i;\theta)}{\partial\theta}
$$

U_i 是 $\tilde{t}_i, c_i$ 的函数，但直接对两个变量求积分存在一定困难，我们尝试使用条件期望的方式分两步求期望。首先求 $E(U_i|c_i)$,

$$\begin{aligned}
E(U_i|c_i) &= \int_0^{c_i} \frac{\partial \ln f(t;\theta)}{\partial \theta} f(t;\theta)dt + E\left[I(\tilde{t}_i > c_i)\right] \frac{\partial \ln S(c_i;\theta)}{\partial \theta} \\
&= \int_0^{c_i} \frac{1}{f(t;\theta)} \frac{\partial f(t;\theta)}{\partial \theta} f(t;\theta)dt + P(\tilde{t}_i > c_i) \frac{\partial \ln S(c_i;\theta)}{\partial \theta} \\
&= \int_0^{c_i} \frac{\partial f(t;\theta)}{\partial \theta} dt + S(c_i;\theta) \frac{1}{S(c_i;\theta)} \frac{\partial S(c_i;\theta)}{\partial \theta} \\
&= \frac{\partial}{\partial \theta} \int_0^{c_i} f(t;\theta)dt + \frac{\partial S(c_i;\theta)}{\partial \theta} \\
&= \frac{\partial}{\partial \theta} F(c_i;\theta) + \frac{\partial}{\partial \theta} S(c_i;\theta) \\
&= \frac{\partial}{\partial \theta} \left(F(c_i;\theta) + S(c_i;\theta)\right) \\
&= \frac{\partial}{\partial \theta} 1 = 0
\end{aligned}$$

因此 $E(U_i) = E(E(U_i|c_i)) = E(0) = 0$。所以 $E(U) = E(\sum_{i=1}^N U_i) = 0$。

下面计算 U 的方差。注意到

$$\begin{aligned}
\mathrm{Var}(U) = E(U^2) &= E\left[\left(\sum_{i=1}^N U_i\right)^2\right] \\
&= \sum_{i=1}^N E(U_i^2) + \sum_{i \neq j} E(U_i)E(U_j) \\
&= \sum_{i=1}^N E(U_i^2)
\end{aligned}$$

因此，为了证明 $E(U^2) = -E(U')$，我们仅需证明 $E(U_i^2) = -E(U_i')$。由于 $\tilde{t}_i, c_i$ 相独立，因此

$$\begin{aligned}
U_i^2 &= \left(I(\tilde{t}_i \le c_i) \frac{\partial \ln f(\tilde{t}_i;\theta)}{\partial \theta} + I(\tilde{t}_i > c_i) \frac{\partial \ln S(c_i;\theta)}{\partial \theta}\right)^2 \\
&= I(\tilde{t}_i \le c_i) \left(\frac{1}{f(\tilde{t}_i;\theta)}\right)^2 \left(\frac{\partial f(\tilde{t}_i;\theta)}{\partial \theta}\right)^2 + I(\tilde{t}_i > c_i) \left(\frac{1}{S(c_i;\theta)}\right)^2 \left(\frac{\partial S(c_i;\theta)}{\partial \theta}\right)^2
\end{aligned}$$

同样地，在此处我们也先求条件期望，

$$\begin{aligned}
E(U_i^2|c_i) &= \int_0^{c_i} \left(\frac{1}{f(t;\theta)}\right)^2 \left(\frac{\partial f(t;\theta)}{\partial\theta}\right)^2 f(t;\theta)dt + S(c_i;\theta)\left(\frac{1}{S(c_i;\theta)}\right)^2 \left(\frac{\partial S(c_i;\theta)}{\partial\theta}\right)^2 \\
&= \int_0^{c_i} \frac{1}{f(t;\theta)}\left(\frac{\partial f(t;\theta)}{\partial\theta}\right)^2 dt + \frac{1}{S(c_i;\theta)}\left(\frac{\partial S(c_i;\theta)}{\partial\theta}\right)^2
\end{aligned}$$

同样，对 U_i' 进行同样的操作，

$$\begin{aligned}
U_i' &= \left(I(\tilde{t}_i \le c_i)\frac{\partial \ln f(\tilde{t}_i;\theta)}{\partial\theta} + I(\tilde{t}_i > c_i)\frac{\partial \ln S(c_i;\theta)}{\partial\theta}\right)' \\
&= \left(I(\tilde{t}_i \le c_i)\frac{1}{f(\tilde{t}_i;\theta)}\frac{\partial f(\tilde{t}_i;\theta)}{\partial\theta} + I(\tilde{t}_i > c_i)\frac{1}{S(c_i;\theta)}\frac{\partial S(c_i;\theta)}{\partial\theta}\right)' \\
&= I(\tilde{t}_i \le c_i)\left(\frac{1}{f(\tilde{t}_i;\theta)}\frac{\partial^2 f(\tilde{t}_i;\theta)}{\partial\theta^2} - \frac{1}{f^2(\tilde{t}_i;\theta)}\left(\frac{\partial f(\tilde{t}_i;\theta)}{\partial\theta}\right)^2\right) \\
&\quad + I(\tilde{t}_i > c_i)\left(\frac{1}{S(c_i;\theta)}\frac{\partial^2 S(c_i;\theta)}{\partial\theta^2} - \frac{1}{S^2(c_i;\theta)}\left(\frac{\partial S(c_i;\theta)}{\partial\theta}\right)^2\right)
\end{aligned}$$

$$\begin{aligned}
E(U_i'|c_i) &= \int_0^{c_i}\left(\frac{1}{f(t;\theta)}\frac{\partial^2 f(t;\theta)}{\partial\theta^2} - \frac{1}{f^2(t;\theta)}\left(\frac{\partial f(t;\theta)}{\partial\theta}\right)^2\right) f(t;\theta)dt \\
&\quad + S(c_i;\theta)\left(\frac{1}{S(c_i;\theta)}\frac{\partial^2 S(c_i;\theta)}{\partial\theta^2} - \frac{1}{S^2(c_i;\theta)}\left(\frac{\partial S(c_i;\theta)}{\partial\theta}\right)^2\right) \\
&= \int_0^{c_i}\frac{\partial^2 f(t;\theta)}{\partial\theta^2}dt - \int_0^{c_i}\frac{1}{f(t;\theta)}\left(\frac{\partial f(t;\theta)}{\partial\theta}\right)^2 dt + \frac{\partial^2 S(c_i;\theta)}{\partial\theta^2} - \frac{1}{S(c_i;\theta)}\left(\frac{\partial S(c_i;\theta)}{\partial\theta}\right)^2 \\
&= -E(U_i^2|c_i) + \int_0^{c_i}\frac{\partial^2 f(t;\theta)}{\partial\theta^2}dt + \frac{\partial^2 S(c_i;\theta)}{\partial\theta^2} \\
&= -E(U_i^2|c_i) + \frac{\partial^2}{\partial\theta^2}F(c_i;\theta) + \frac{\partial^2}{\partial\theta^2}S(c_i;\theta) \\
&= -E(U_i^2|c_i)
\end{aligned}$$

因此 $E(U_i') = E(E(U_i'|c_i)) = E(-E(U_i^2|c_i)) = -E(U_i^2)$，所以 $E(U^2) = -E(U')$。

经此，我们完成了对该定理的证明。该定理说明，虽然生存数据的形式与一般数据不同，但在似然函数方面，生存数据却有着和一般数据相同的性质。因此，一些基于似然函数的统计量也可以用在生存数据的统计推断上。

由中心极限定理 $\frac{U-E[U]}{\mathrm{Var}(U)} \to^D N(0,1) \iff U \to^D N(0,I)$。对于参数是多元的情况，则有 $U^T I^{-1} U \to^D \chi^2(p)$。基于此，我们可以进行三种类型的检验。

Score 检验　检验 $H_0: \theta = \theta_0, H_1: \theta \neq \theta_0$。统计量为

$$U(\theta_0)^T I^{-1}(\theta_0) U(\theta_0) \to^D \chi^2(p)$$

Score 检验计算较为简单，但很难得到 θ 的置信区间。

Wald 检验 对得分函数 $U(\theta)$ 在 $\hat{\theta}$ 处泰勒展开，可得

$$U(\theta) = U(\hat{\theta}) + H(\theta^*)(\theta - \hat{\theta}) = H(\theta^*)(\theta - \hat{\theta})$$

因此 $\hat{\theta} - \theta = -H^{-1}(\theta^*)U(\theta)$。由于 $\theta^* \to^P \theta$，因此 $H(\theta^*) \to^P H(\theta) \to^P -I$ 且 $U(\theta) \to^D N(0, I)$。因此

$$\hat{\theta} - \theta \to^D N(0, I^{-1})$$

对于检验 $H_0 : \theta = \theta_0$，可使用统计量 $(\hat{\theta} - \theta_0)^T I(\hat{\theta})(\hat{\theta} - \theta_0) \to^D \chi^2(p)$ 进行检验。

似然比检验 似然比检验对似然函数 $l(\theta)$ 在 $\hat{\theta}$ 处进行泰勒展开，有

$$l(\theta) = l(\hat{\theta}) + U(\hat{\theta})(\theta - \hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^T H(\theta^*)(\theta - \hat{\theta}) = l(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^T H(\theta^*)(\theta - \hat{\theta})$$

因此，$2(l(\hat{\theta}) - l(\theta)) = -(\hat{\theta} - \theta)^T H(\theta^*)(\hat{\theta} - \theta)$。由于 $H(\theta^*) \to^P -I, \hat{\theta} - \theta \to^D N(0, I^{-1})$，因此可以使用统计量 $2(l(\hat{\theta}) - l(\theta_0)) \to^D \chi^2(p)$ 来检验 $H_0 : \theta = \theta_0$。

5 参数生存模型

我们开始对生存数据 $\{t_i, \delta_i, x_i\}$ 进行建模，构建生存时间 T 与自变量 X 之间的关系。本章首先介绍参数模型，参数模型假设生存时间 T 服从某个分布；而半参数模型则不对生存时间的分布进行假定。

5.1 指数比例风险模型

比例风险模型 比例风险模型是生存分析中常用的模型，其对风险函数 $\lambda(t)$ 建模。对第 i 个个体的风险函数 $\lambda_i(t)$，将其拆分为两个部分

$$\lambda_i(t) = \lambda_0(t) \exp\left(x_i^T \beta\right)$$

$\lambda_0(t)$ 是基准的风险函数，对于不同的个体，通过其自变量 x_i 给风险函数进行 $\exp(x_i^T \beta)$ 的放缩。注意到在比例风险模型中 $\frac{\lambda_i(t)}{\lambda_0(t)} = \exp(x_i^T \beta)$ 与 t 无关，其比例性便体现在此处，个体的生存风险仅与个体的自变量成比例。此外，比例风险模型是一类线性模型。

指数比例风险模型 指数比例风险模型假定生存时间 T 服从指数分布 $T \sim \mathrm{Exp}(\lambda)$。注意此处的生存时间是真实的生存时间，而不是删失观测下的生存时间。因此

$$\lambda_0(t) = \frac{f(t)}{S(t)} = \frac{\lambda \exp(-\lambda t)}{\exp(-\lambda t)} = \lambda$$

因此，指数比例风险模型的形式是

$$\lambda_i(t) = \lambda \exp(x_i^T \beta)$$

在 $x_i^T\beta$ 一项中不存在截距项 β_0，这是为了防止模型不可识别。若 $x_i^T\beta$ 中存在 β_0，则

$$\lambda(t) = \lambda \exp(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p) = [\lambda \exp(\beta_0)] \exp(\beta_1 x_1 + ... + \beta_p x_p)$$

即存在两组不同的参数 $(\lambda, \beta_0), (\lambda \exp(\beta_0), 0)$ 均满足比例风险模型。为避免模型不可识别，在指数比例风险模型中，我们将 λ 放入 $x_i^T\beta$ 中作为截距项。因此，指数比例风险模型的形式为

$$\lambda_i(t) = \exp(x_i^T \beta) = \exp(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p)$$

β 的估计 假设真实的生存时间在给定自变量的情况下满足分布指数分布 $T_i|X_i \sim \mathrm{Exp}(\lambda_i)$，其中 $\lambda_i = \exp(x_i^T\beta)$。对于生存数据 $\{(t_i, \delta_i, x_i)\}_{i=1}^N$，生存时间 T 的似然函数为

$$\begin{aligned} L(\beta) &= \prod_{i=1}^N [f(t_i, x_i, \beta)]^{\delta_i} [S(t_i, x_i, \beta)]^{1-\delta_i} \\ &= \prod_{i=1}^N \left[\lambda_i \exp(-\lambda_i t_i)\right]^{\delta_i} \left[\exp(-\lambda_i t_i)\right]^{1-\delta_i} \\ &= \prod_{i=1}^N \lambda_i^{\delta_i} \exp(-\lambda_i t_i) \end{aligned}$$

因此

$$l(\beta) = \sum_{i=1}^N \delta_i \ln \lambda_i - \lambda_i t_i = \sum_{i=1}^N \left(\delta_i x_i^T \beta - \exp(x_i^T \beta) t_i\right)$$

通过极大化 $l(\beta)$ 便可得到 β 的估计值 $\hat{\beta}$。由于似然函数无显式解，需要通过迭代方法对其进行求解。似然函数的得分函数为

$$U(\beta) = \frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^N \delta_i x_i - \exp(x_i^T \beta) t_i x_i = \sum_{i=1}^N \left(\delta_i - \exp(x_i^T \beta) t_i\right) x_i$$

记 $d = (\delta_1, \delta_2, ..., \delta_N)^T, \mu = (\exp(x_1^T\beta)t_1, \exp(x_2^T\beta)t_2, ..., \exp(x_N^T\beta)t_N)$。因此，

$$U(\beta) = \sum_{i=1}^N \left(\delta_i - \exp(x_i^T \beta) t_i\right) x_i = X^T(d - \mu)$$

β 的黑塞矩阵为

$$H(\beta)=\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta^T}=\frac{\partial U(\beta)}{\partial\beta^T}=\sum_{i=1}^{N}-\exp(x_i^T\beta)t_ix_ix_i^T$$

记 $W=\mathrm{diag}\left(\exp(x_1^T\beta)t_1,\exp(x_2^T\beta)t_2,...,\exp(x_N^T\beta)t_N\right)$，则

$$H(\beta)=-\sum_{i=1}^{N}\exp(x_i^T\beta)t_ix_ix_i^T=-X^TWX$$

根据 Newton-Raphson 算法，$\hat\beta$ 可按如下步骤得到

1. 初始化 β 的估计值 $\hat\beta^{(0)}$
2. 迭代计算 β 直至收敛 ($\|\hat\beta^{(m+1)}-\hat\beta^{(m)}\|<\epsilon$)。迭代公式为

$$\hat\beta^{(m+1)}=\hat\beta^{(m)}+\left(X^TW^{(m)}X\right)^{-1}X^T\left(d-\mu^{(m)}\right)$$

β 的推断　由于 $\hat\beta$ 是极大似然估计，因此 $\hat\beta\sim N(\beta,I^{-1})$。由于 $I=-H(\beta)=X^TWX$，有

$$\hat\beta\sim N\left(\beta,(X^TWX)^{-1}\right)$$

对于 β 的第 j 个分量 β_j，有 $\hat\beta_j\sim N(\beta_j,\mathrm{SE}_j^2)$，其中 $\mathrm{SE}_j=\sqrt{[(X^TWX)^{-1}]_{jj}}$。$\beta_j$ 的 $1-\alpha$ 置信区间为

$$\left[\hat\beta_j-z_{\frac{\alpha}{2}}\mathrm{SE}_j,\hat\beta_j+z_{\frac{\alpha}{2}}\mathrm{SE}_j\right]$$

β 的解释　在比例风险模型中，β 可通过风险比例 (hazard ratio, HR) 或相对风险 (relative risk, RR) 进行解释。假设对于个体 i 和 j，两者仅在变量 x_k 上有一个单位的区别 ($x_{ik}=x_{jk}+1$)，则 $\lambda_i(t)/\lambda_j(t)=\exp(\beta_j)$。

$\mathrm{HR}=e^{\beta_j}$ 即可解释为系数 β_j 的相对风险，即 X_j 每增加一个单位时，个体存活风险增加的倍数。

5.2 Weibull 比例风险模型

Weibull 分布　指数比例风险模型使用指数分布刻画生存时间，但用 Weibull 分布刻画生存时间更为合适。假设 $X\sim\mathrm{Exp}(\tau)$，令 $T=X^\sigma$，则

$$S_T(t)=P(T>t)=P(X^\sigma>t)=P(X>t^{\frac{1}{\sigma}})=\exp\left(-\tau t\frac{1}{\sigma}\right)$$

根据生存函数 $S_T(t)$，可以求得 T 的风险函数 $\lambda_T(t)=\frac{\tau}{\sigma}t^{\frac{1}{\sigma}-1}$。该风险函数的形式略显复杂，若记 $\gamma=\frac{1}{\sigma},\tau=\lambda^\gamma$，进行代换，则有

$$\lambda(t)=\lambda^\gamma\gamma t^{\gamma-1}=\lambda\gamma(\lambda t)^{\gamma-1}$$

若一个变量的风险函数具有如上形式，则我们称该变量 T 服从 Weibull 分布，分布的参数是 λ, γ，记为 $T \sim \text{Weibull}(\lambda, \gamma)$。注意到，当 $\gamma = 1$ 时，Weibull 分布便退化为了指数分布，因此 Weibull 分布是指数分布的拓展。此外，当 $\gamma = 1$ 时，$\lambda(t) = \lambda$，这表示在任何时刻，个体生存的风险均是一样的。当 $\gamma > 1$ 时，随 t 的增大，$\lambda(t)$ 也会变大。因此 $\gamma > 1$ 表示生存风险随时间的增加不断变大；而 $\gamma < 1$ 则表示生存风险随时间的增加不断减小。在这种性质下，我们可以通过控制 γ 的取值更好的刻画不同个体的风险函数。

Weibull 分布与极值分布 假设 $X \sim \text{Exp}(1)$，令 $Y = \log X$。易知 Y 的风险函数为 $\lambda(y) = e^y, f(y) = \exp(y - e^y)$。$Y$ 被称为标准极值分布，注意 Y 的取值不一定是正，可能是负的。

若 $X \sim \text{Exp}(1)$，令 $T = X^\sigma$，则 $Y = \log T = \log X^\sigma = \sigma \log X$。因此可以看出 Weibull 分布的参数 γ 在极值分布中是比例参数，控制分布的分布范围。

若 $X \sim \text{Exp}(\lambda)$，令 $Y = \log X$，则 $\lambda(y) = \lambda e^y = e^{y - \ln \lambda} := e^{y-\alpha}$。因此，相比标准极值分布 W，分布 Y 相当于分布 W 的左右平移 $Y = \alpha + W$。即对数指数分布的是极值分布的左右平移。

因此，对于 Weibull 分布 $T \sim \text{Weibull}(\lambda, \gamma)$，我们有 $Y = \log T = \alpha + \sigma W$。假设 $X \sim \text{Exp}(\tau), T = X^\sigma$，且 $\gamma = \frac{1}{\sigma}, \lambda^\gamma = \tau$，易知 T 服从 Weibull 分布 $T \sim \text{Weibull}(\lambda, \gamma)$。则

$$Y = \log T = \log X^\sigma = \sigma \log X = \sigma(-\log \tau + W) = \frac{1}{\gamma}(-\log \lambda^\gamma + W) = -\log \lambda + \frac{1}{\gamma} W$$

因此，对数 Weibull 分布具有形式 $Y = \log T = \alpha + \sigma W$，对数 Weibull 分布是标准极值分布的平移-比例族。

Weibull 比例风险模型 Weibull 比例风险模型具有如下形式

$$\lambda_i(t) = \lambda_0(t) \exp(x_i^T \beta) = \lambda \gamma (\lambda t)^{\gamma - 1} \exp(x_i^T \beta)$$

为了使模型可识别，此处我们规定 $x_i^T \beta$ 中不含有截距 β_0，对风险函数做变换

$$\begin{aligned}
\lambda_i(t) &= \lambda \gamma (\lambda t)^{\gamma-1} \exp(x_i^T \beta) = \lambda^\gamma \gamma t^{\gamma-1} \exp(x_i^T \beta) \\
&= \left[\lambda \exp\left(\frac{x_i^T \beta}{\gamma}\right)\right]^\gamma \gamma t^{\gamma-1} \\
&:= \lambda_i^\gamma \gamma t^{\gamma-1} = \lambda_i \gamma (\lambda_i t)^{\gamma-1}
\end{aligned}$$

Weibull 比例风险模型假设生存时间 $T_i | x_i \sim \text{Weibull}(\lambda_i, \gamma)$，其中 $\lambda_i = \lambda \exp(\eta_i), \eta_i = \frac{x_i^T \beta}{\gamma}$。其中自变量 x_i 仅影响 λ_i，而不影响 γ。

由于 $T_i | x_i \sim \text{Weibull}(\lambda_i, \gamma)$，我们考虑 $Y_i = \log T_i = \alpha_i + \sigma W_i$。其中 $\alpha_i = -\log \lambda_i = -\log(\lambda \exp(\eta_i)) = -\log \lambda - \eta_i, \sigma = \frac{1}{\gamma}$。因此

$$Y_i = -\log\lambda - \frac{x_i^T\beta}{\gamma} + \sigma W_i := \alpha + x_i^T\beta^* + \sigma W_i$$

其中 $\alpha = -\log\lambda, \beta^* = -\sigma\beta = -\frac{\beta}{\gamma}$。该表达形式与线性回归非常类似，区别在于在线性回归中 W_i 服从标准正态分布，而在此处 W_i 服从标准极值分布。我们同样可以通过极大似然估计得到 β 的估计。具有形式 $Y_i = x_i^T\beta + W_i$ 的生存模型被称为加速失效时间模型。当 W_i 服从极值分布时，其余 Weibull 比例风险模型等价。

5.3 加速失效时间模型

加速失效时间模型 (Accelerated Failure Time Model, AFT) 对生存时间 T 建模，令 $Y = \log T$，加速失效时间模型建立模型

$$Y_i = \log T_i = x_i^T\beta + W_i$$

其中 W_i 独立同分布于分布 f。当 f 已知时，模型是一个参数模型，当 f 未知时，模型是一个半参数模型。例如，当 $W_i \sim^{i.i.d} N(0,\sigma^2)$，则 $Y_i \sim N(x_i^T\beta, \sigma^2)$，生存时间服从一个对数正态分布。在这种情况下，我们可以使用最小二乘估计对没有删失的数据估计 β 的值。

同时，由于 $Y_i = \log T_i = x_i^T\beta + W_i$，因此 $T_i = e^{Y_i} = e^{\eta_i}e^{W_i} := T_0e^{\eta_i}$。其中 T_0 可认为基准生存时间，e^{η_i} 是生存时间的加速失效因子，因此该模型被称为加速失效时间模型。

比例风险模型和加速失效时间模型的关系 比例风险模型对风险函数 $\lambda_i(t)$ 进行建模，但加速失效时间模型对生存时间 T_i 进行建模。对于 AFT 模型，$T_i = T_0e^{\eta_i}$，假定 T_0 的生存函数和风险函数分别为 $S_0(t), \lambda_0(t)$，因此

$$S_i(t) = P(T_i > t) = P(T_0e^{\eta_i} > t) = P(T_0 > e^{-\eta_i}t) = S_0(e^{-\eta_i}t)$$

从而

$$\lambda_i(t) = -\frac{d}{dt}\log S_i(t) = -\frac{S_0'(e^{-\eta_i}t)}{S_0(e^{-\eta_i}t)}e^{-\eta_i} = \lambda_0(e^{-\eta_i}t)e^{-\eta_i}$$

因此，比例风险模型和加速失效时间模型的风险函数不同，

- 比例风险模型的风险函数为 $\lambda_i(t) = \lambda_0(t)e^{\eta_i}$
- 加速失效时间模型的风险函数为 $\lambda_i(t) = \lambda_0(e^{-\eta_i}t)e^{-\eta_i}$

下面考虑两个模型对于 $Y_i = \log T_i$ 的风险函数，其对于两个模型之间的关系的刻画更为明显。对于比例风险模型，其 Y_i 生存函数为

$$S_i(y) = P(Y_i > y) = P(T_i > e^y) = \exp\left(-\int_0^{e^y}\lambda_i(u)du\right)$$

因此

$$\lambda_i(y) = -\frac{d}{dy}\log S_i(y) = \frac{d}{dy}\int_0^{e^y}\lambda_i(u)du = \lambda_i(e^y)e^y = \lambda_0(e^y)e^{\eta_i}e^y$$

记 $\tilde{\lambda}_0(y) = \lambda_0(e^y)e^y$，则对于比例风险模型，我们有

$$\log\lambda_i(y) = \log\lambda_0(e^y)e^{\eta_i}e^y = \log\tilde{\lambda}_0(y)e^{\eta_i} = \log\tilde{\lambda}_0(y) + \eta_i$$

对于加速失效时间模型，我们同样计算 Y_i 生存函数

$$S_i(y) = P(Y_i > y) = P(T_i > e^y) = \exp\left(-\int_0^{e^y}\lambda_i(u)du\right) = \exp\left(-\int_0^{e^y}\lambda_0(ue^{-\eta_i})e^{-\eta_i}du\right)$$

因此，

$$\lambda_i(y) = -\frac{d}{dy}\log S_i(y) = \frac{d}{dy}\int_0^{e^y}\lambda_0(ue^{-\eta_i})e^{-\eta_i}du = \lambda_0(e^{y-\eta_i})e^{y-\eta_i} = \tilde{\lambda}_0(y-\eta_i)$$

所以，

$$\log\lambda_i(y) = \log\tilde{\lambda}_0(y-\eta_i)$$

因此，对于比例风险模型和加速失效时间模型，Y_i 的风险函数的对数分别具有以下形式，

- 对于比例风险模型，$\log\lambda_i(y) = \log\tilde{\lambda}_0(y) + \eta_i$
- 对于加速失效时间模型，$\log\lambda_i(y) = \log\tilde{\lambda}_0(y-\eta_i)$

所以比例风险模型相当于对对数生存函数 $\log\tilde{\lambda}_0(y)$ 作上下平移，而加速失效事件模型相当于对对数风险函数做左右平移。在一般情况下，两者是不等价的 (因为分别为左右平移和上下平移)；不过当函数 $\log\tilde{\lambda}_0(y)$ 在坐标系中是一条直线时，左右平移和上下平移是等价的。下面说明 Weibull 模型属于这种情况。我们仅需说明，对于 Weibull 模型，有 $\log\tilde{\lambda}_0(y) = ay + b$ 即可。

由于 Weibull 模型的基准风险函数为 $\lambda_0(t) = \lambda\gamma(\lambda t)^{\gamma-1}$，因此 $\tilde{\lambda}_0(y) = \lambda_0(e^y)e^y = \lambda\gamma(\lambda e^y)^{\gamma-1}e^y = (\lambda e^y)^\gamma\gamma$，所以

$$\log\tilde{\lambda}_0(y) = \gamma y + \log(\lambda^\gamma\gamma) := ay + b$$

因此，对于 Weibull 模型，比例风险模型和加速失效时间模型是等价的。

加速失效时间模型的参数估计 加速失效时间模型为 $Y_i = x_i^T\beta + \sigma W_i$，其中参数为 β, σ，W_i 是随机变量，其概率密度函数和生存函数分别为 $f(.)$ 和 $S(.)$。因此 Y_i 的生存函数和概率密度函数分别为

$$
\begin{aligned}
&S_i(y)=P(Y_i>y)=P\left(x_i^T\beta+\sigma W_i>y\right)=P\left(W_i>\frac{y-x_i^T\beta}{\sigma}\right)=S\left(\frac{y-x_i^T\beta}{\sigma}\right):=S(w_i)\\
&f_i(y)=-S_i'(y)=f(w_i)\frac{1}{\sigma}
\end{aligned}
$$

根据生存数据似然函数的构建方法，我们有，

$$
L(\beta,\sigma|y_i,\delta_i)=\prod_{i=1}^N f_i(y)^{\delta_i}S_i(y)^{1-\delta_i}=\prod_{i=1}^N\left(\frac{1}{\sigma}f(w_i)\right)^{\delta_i}S(w_i)^{1-\delta_i}=\prod_{i=1}^N\left(\frac{1}{\sigma}\lambda(w_i)\right)^{\delta_i}S(w_i)
$$

对其取对数，有

$$
l(\beta,\sigma|y_i,\delta_i)=\sum_{i=1}^N l_i=\sum_{i=1}^N\delta_i\left[-\log\sigma+\log\lambda(w_i)\right]+\log S(w_i)
$$

分别对 β 和 σ 求导

$$
\begin{aligned}
&\frac{\partial l}{\partial\beta}=\sum_{i=1}^N\frac{\partial l}{\partial w_i}\frac{\partial w_i}{\partial\beta}=\sum_{i=1}^N-\frac{x_i}{\sigma}\frac{\partial l}{\partial w_i}:=-\sigma^{-1}X^T\alpha\\
&\frac{\partial l}{\partial\sigma}=\sum_{i=1}^N\left[-\frac{\delta_i}{\sigma}+\frac{\partial l}{\partial w_i}\frac{\partial w_i}{\partial\sigma}\right]=\sum_{i=1}^N\left[-\frac{\delta_i}{\sigma}-\frac{y_i-x_i^T\beta}{\sigma^2}\frac{\partial l}{\partial w_i}\right]=\sum_{i=1}^N\left[-\frac{\delta_i}{\sigma}-\frac{w_i}{\sigma}\frac{\partial l}{\partial w_i}\right]:=-\sigma^{-1}(\delta+W^T\alpha)
\end{aligned}
$$

其中 $\alpha=\left(\frac{\partial l}{\partial w_1},\frac{\partial l}{\partial w_2},...,\frac{\partial l}{\partial w_N}\right)^T,w=(w_1,w_2,...,w_N)^T,\delta=\sum_{i=1}^N\delta_i$。求解 β,σ 的极大似然估计等价于求解方程组

$$
\begin{cases}
-\sigma^{-1}X^T\alpha=0\\
-\sigma^{-1}(\delta+W^T\alpha)=0
\end{cases}
$$

我们可以使用 Newton-Raphson 算法对该方程组迭代求解。对于加速失效时间模型，我们还需注意以下几点，

1. α 与 W 的分布有关。例如，当 W 服从标准极值分布时，则有 $\lambda(w)=e^w,S(w)=\exp(-e^w)$，因此 $l_i=\delta_i(-\log\sigma+w_i)-e^{w_i},\alpha_i=\frac{\partial l_i}{\partial w_i}=\delta_i-e^{w_i}$。
2. 使用 Newton-Raphson 算法进行迭代时，可能在 σ 的估计值上取到负数，我们可以使用加权更新法保证 σ 始终为正，即 $\sigma_{m+1}=(1-\tau)\sigma_m+\tau\tilde{\sigma}$。

加速失效时间模型的参数推断 模型参数 $\theta=(\beta^T,\sigma)^T$ 的 Fisher 信息矩阵为

$$
I(\theta)=-\begin{pmatrix}\frac{\partial^2 l}{\partial\beta\partial\beta^T} & \frac{\partial^2 l}{\partial\beta\partial\sigma}\\ \frac{\partial^2 l}{\partial\sigma\partial\beta^T} & \frac{\partial^2 l}{\partial\sigma^2}\end{pmatrix}
$$

$\widehat{\beta}$ 的方差为 $[I^{-1}]_{1:p,1:p}$。注意，如果使用 $\left(-\frac{\partial^2 l}{\partial\beta\partial\beta^T}\right)^{-1}$ 作为方差的估计，则其会低估方差的大小。

加速失效时间模型的参数解释 注意到 $T_i = T_0 e^{x_i^T\beta}$。因此对于比例风险模型和加速失效时间模型，两者的系数解释不同，若 $e^{\beta_j}=2$，则

- 对于比例风险模型：当变量 X_j 增加一个单位时，个体的**相对存活风险**变为原来的两倍
- 对于加速失效时间模型：当变量 X_j 增加一个单位时，个体的**生存时间**变为原来的两倍

6 半参数生存模型

半参数模型一般可分为两部分：第一部分是参数部分，其度量因变量 (如：风险函数，生存时间等) 与自变量的关系，其结构是明确的 (如：线性等)；第二部分是非参数部分，其一般表示潜在的生存分布。由于模型由参数部分和非参数部分构成，我们称为半参数模型。与参数模型不同，对于潜在的生存分布，我们不知晓其分布形式。

6.1 半参数加速失效时间模型

加速失效时间模型具有形式 $\log T_i = x_i^T\beta + W_i$，在半参数条件下，$W_i$ 独立同分布于一个未知的分布 F。由于我们无法构建极大似然函数，仅可通过次序统计量或秩统计量对该问题进行求解。我们将此方法称为秩回归 (rank regression)。

在没有删失的情况下，假设我们有数据 $\{(x_1,y_1),(x_2,y_2),...,(x_N,y_N)\}$。假设数据 y 的排列顺序为 $y_{(1)} < y_{(2)} < ... < y_{(N)}$，$y_{(i)}$ 对应的 x 记为 $x_{(i)}$。对于数据 $\{(x_{(1)},y_{(1)}),(x_{(2)},y_{(2)}),...,(x_{(N)},y_{(N)})\}$，秩统计量为

$$\sum_{i=1}^{N}\left(i-\frac{N+1}{2}\right)\left(x_{(i)}-\bar{x}\right)$$

对于删失数据，我们将秩统计量扩展为

$$U = \sum_{j:\delta_j=1}\left(X_{(j)}-\bar{X}_{(j)}\right)$$

需要注意以下几点：

1. 秩统计量仅在未删失的数据上进行计算。
2. $X_{(j)}$ 是第 j 个失效事件的自变量，$\bar{X}_{(j)}$ 是在时刻 t_j 仍然在险的个体的自变量的平均值。

在假设条件 $\beta=0$ 下，我们有 $E(U)=0, \mathrm{Var}(U)=\sum_{j:\delta_j=1}\left(X_{(j)}-\bar{X}_{(j)}\right)^2$。在单变量的情况下，可使用以下统计量进行检验

$$\frac{U^2}{V}\sim\chi^2(1)$$

如果我们想检验 $H_0 : \beta = \beta_0$，我们可定义 $\tilde{Y}_i = Y_i - x_i^T \beta_0$，并对 $\tilde{Y}_i$ 进行同样的操作。同时我们可以看到，对于不同的 β_0，按如上操作得到的得到的统计量 $\frac{U^2}{V}$ 各不相同。因此 $\frac{U^2}{V}$ 是一个关于 β 的函数。我们可以通过最小化秩统计量得到 $\hat{\beta}$，

$$\hat{\beta} = \arg\min_{\beta} \frac{U^2}{V}$$

不过需要注意，由于秩统计量基于因变量的排序，因此 $\frac{U^2}{V}$ 不是 β 的连续函数 (这是因为，略为改变 β 的取值不会影响样本的排序，也不会改变统计量的值)。连续性的缺失导致我们无法使用 Newton-Raphson 算法求解 β 的估计值，也无法使用 Wald 法求得参数的置信区间。因此，半参数加速失效时间模型在实际中不常被使用。

不过我们需要指出，半参数模型仍有一定的优势。首先，我们不假定 W 的分布，不添加限制条件。此外，半参数模型比较稳定，与潜在的分布无关。

6.2 Cox 比例风险模型

Cox 比例风险模型是一个半参数模型，其对 $\lambda_0(t)$ 没有任何假设条件。因此，如果需要对该模型进行类似极大似然估计，求得 β 的估计值，需要消去未知的 $\lambda_0(t)$。

偏似然函数 我们首先考虑一种特殊情况，假设我们有两个观测 T_1 和 T_2（均未删失），其风险函数分别为 $\lambda_1(t)$ 和 $\lambda_2(t)$，则

$$\begin{aligned}
P(T_1 = t | T_{(1)} = t) &= \frac{P(T_1 = t, T_{(1)} = t)}{P(T_{(1)} = t)} = \frac{P(T_1 = t, T_2 > t)}{P(T_1 = t, T_2 > t) + P(T_1 > t, T_2 = t)} \\
&= \frac{f_1(t) S_2(t)}{f_1(t) S_2(t) + f_2(t) S_1(t)} = \frac{f_1(t)/S_1(t)}{f_1(t)/S_1(t) + f_2(t)/S_2(t)} \\
&= \frac{\lambda_1(t)}{\lambda_1(t) + \lambda_2(t)}
\end{aligned}$$

在 Cox 比例风险模型下，我们有

$$P(T_1 < T_2) = P(T_1 = t | T_{(1)} = t) = \frac{\lambda_1(t)}{\lambda_1(t) + \lambda_2(t)} = \frac{\exp(x_1^T \beta)}{\exp(x_1^T \beta) + \exp(x_2^T \beta)}$$

当我们将以上情况推至 J 个观测时，我们有

$$
\begin{aligned}
P(T_1 < T_2 < ... < T_J) &= P(T_1 = t, T_2 < T_3 < ... < T_J | T_{(1)} = t) = P(T_1 = t | T_{(1)} = t) P(T_2 < T_3 < ... < T_J | T_{(1)} = t) \\
&= \frac{\lambda_1(t)}{\lambda_1(t) + \lambda_2(t) + ... + \lambda_J(t)} P(T_2 < T_3 < ... < T_J) \\
&= ... \\
&= \prod_{j=1}^{J} \frac{\exp(x_j^T \beta)}{\sum_{k=j}^{J} \exp(x_k^T \beta)}
\end{aligned}
$$

注意到，经过以上的处理，我们将风险函数中未知的 $\lambda_0(t)$ 消去，在 $P(T_1 < T_2 < ... < T_J)$ 中未知的部分仅含有 β。事实上，$P(T_1 < T_2 < ... < T_J)$ 可以看做在给定失效时间条件下的 β 的似然函数 $L(\beta)$。当我们把删失情况也进行考虑时，似然函数将改写为

$$
L(\beta) = \prod_{j} \frac{\exp(x_j^T \beta)}{\sum_{k \in R(t_j)} \exp(x_k^T \beta)} = \prod_{i=1}^{N} \left(\frac{\exp(x_i^T \beta)}{\sum_{k \in R(t_i)} \exp(x_k^T \beta)} \right)^{\delta_i}
$$

在上式中，$R(t_j)$ 表示在 t_j 时刻仍然在险的个体，$k \in R(t_j)$ 当且仅当 $t_k \geq t_j$。此外，从上式中可以看到 $L(\beta)$ 仅在观测到的样本上进行计算，删失的样本不予计算。

$L(\beta)$ 被称为**偏似然函数 (partial likelihood function)**。不过事实上 $L(\beta)$ 不是给定 X 时，(t, δ) 的概率密度，因此严格来说，其并不是似然函数。但是，Cox 在 1972 年和 1975 年两次证明了，$L(\beta)$ 虽然严格来说不是似然函数，但其仍拥有似然函数的性质（0 均值，正定的黑塞矩阵），因此，我们仍然可以用基于似然函数的方法处理偏似然函数。

β 的估计　β 的估计值通过最大化偏似然函数得到，

$$
\hat{\beta} = \arg\max_{\beta} L(\beta) = \arg\max_{\beta} \prod_{i=1}^{N} \left(\frac{\exp(x_i^T \beta)}{\sum_{j \in R(t_i)} \exp(x_j^T \beta)} \right)^{\delta_i}
$$

记 $\eta_i = x_i^T \beta, w_i = \exp(x_i^T \beta), W_i = \sum_{j \in R_i} \exp(x_j^T \beta)$。$\sum_{j \in R(t_i)} \exp(x_j^T \beta) = \sum_{j=1}^{N} Y_j(t_i) \exp(x_j^T \beta)$，其中 $Y_j(t_i) = I(t_j \geq t_i)$。记 $\pi_{ij} = Y_i(t_j) \frac{w_i}{W_j}$。

由于 $L(\beta) = \prod_{i=1}^{N} (\frac{w_i}{W_i})^{\delta_i}$，因此对数似然函数 $l(\beta) = \ln L(\beta)$ 为

$$
l(\beta) = \sum_{i=1}^{N} \delta_i \ln w_i - \sum_{i=1}^{N} \delta_i \ln W_i = \sum_{i=1}^{N} \delta_i \eta_i - \sum_{i=1}^{N} \delta_i \ln W_i
$$

Cox 比例风险模型使用牛顿法最大化偏似然函数，下面分别求偏似然函数的梯度和黑塞矩阵，

$$
\frac{\partial l(\beta)}{\partial \eta_k} = \delta_k - \sum_{i=1}^{N} \delta_i \frac{1}{W_i} \frac{\partial W_i}{\partial \eta_k} = \delta_k - \sum_{i=1}^{N} \delta_i \frac{1}{W_i} I(t_k \geq t_i) w_k = \delta_k - \sum_{i=1}^{N} \delta_i \pi_{ki}
$$

因此 $U(\eta)=\frac{\partial l(\beta)}{\partial \eta}=\delta-P\delta$，其中 $\delta=(\delta_1,\delta_2,...,\delta_N)^T, P=(\pi_{ij})_{N\times N}$。根据链式法则，我们有

$$U(\beta)=\frac{\partial l(\beta)}{\partial \beta}=\sum_{k=1}^{N}\frac{\partial l(\beta)}{\partial \eta_k}\frac{\partial \eta_k}{\partial \beta}=\sum_{k=1}^{N}\left(\delta_k-\sum_{i=1}^{N}\delta_i\pi_{ki}\right)x_i=X^T(\delta-P\delta)$$

下面 β 的黑塞矩阵，同样我们还是先求关于 η 的黑塞矩阵，注意到 $\frac{\partial l(\beta)}{\partial \eta_k}=\delta_k-\sum_{i=1}^{N}\delta_i\pi_{ki}$。因此，

$$\begin{aligned}\frac{\partial^2 l(\beta)}{\partial \eta_k^2}&=-\sum_{i=1}^{N}\delta_i\frac{\partial \pi_{ki}}{\partial \eta_k}=-\sum_{i=1}^{N}\delta_i Y_k(t_i)\frac{w_kW_i-w_kY_k(t_i)w_k}{W_i^2}\\&=-\sum_{i=1}^{N}\left(\delta_iY_k(t_i)\frac{w_k}{W_i}-\delta_iY_k^2(t_i)\frac{w_k^2}{W_i^2}\right)\\&=-\sum_{i=1}^{N}\delta_i(\pi_{ki}-\pi_{ki}^2)\\&=-\sum_{i=1}^{N}\delta_i\pi_{ki}(1-\pi_{ki})\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 l(\beta)}{\partial \eta_k\partial \eta_j}&=-\sum_{i=1}^{N}\delta_i\frac{\partial \pi_{ki}}{\partial \eta_j}=-\sum_{i=1}^{N}\delta_iY_k(t_i)\frac{-w_kY_j(t_i)w_j}{W_i^2}\\&=\sum_{i=1}^{N}\delta_i\left(\frac{Y_k(t_i)w_k}{W_i}\right)\left(\frac{Y_j(t_i)w_j}{W_i}\right)\\&=\sum_{i=1}^{N}\delta_i\pi_{ki}\pi_{ji}\end{aligned}$$

因此，

$$H(\beta)=\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta^T}=\sum_{k=1}^{N}\sum_{j=1}^{N}\frac{\partial \eta_k}{\partial \beta}\frac{\partial^2 l(\beta)}{\partial \eta_k\partial \eta_j}\frac{\partial \eta_j}{\partial \beta^T}=\sum_{k=1}^{N}\sum_{j=1}^{N}x_k\frac{\partial^2 l(\beta)}{\partial \eta_k\partial \eta_j}x_j^T:=-X^TWX$$

其中 W 是一个 $N\times N$ 矩阵，其对角元素为矩阵 $P\mathrm{diag}(\delta_i)(1-P)$ 的对角元素，其非对角元素为矩阵 $-P\mathrm{diag}(\delta_i)P$ 的非对角元素。因此对于偏似然函数，我们有

$$U(\beta)=X^T(\delta-P\delta), H(\beta)=-X^TWX$$

β 的更新公式为

$$\hat{\beta}_{(m+1)}=\hat{\beta}_{(m)}+(X^TWX)^{-1}X^T(\delta-P\delta)$$

其中，W 和 P 都与 β 有关。我们可不断使用该更新公式，不断迭代直至收敛得到 $\hat{\beta}$。使用牛顿法迭代时，可能需要使用折半法确保算法的收敛性。

β **的推断** 由于黑塞矩阵为 $H = -X^TWX$，因此费舍尔信息矩阵为 $I = X^TWX$，所以

$$\hat{\beta} \to^D N\left(\beta, (X^TWX)^{-1}\right)$$

因此，β_j 的 $100(1-\alpha)\%$ 置信区间为 $(\hat{\beta}_j - z_{\frac{\alpha}{2}}\sqrt{(I^{-1})_{jj}}, \hat{\beta}_j + z_{\frac{\alpha}{2}}\sqrt{(I^{-1})_{jj}})$。

结的处理 结指两个个体在同一时间失效，那么偏似然函数需要作出一定调整。例如，个体 2 和个体 3 均在时刻 j 失效，个体 4 在时刻 j 仍然在险，则这三个个体的排序可以是 2，3，4 或者 3，2，4。两种排序下的偏似然函数分别为 $\frac{w_2}{\sum_{2,3,4} w_j}\frac{w_3}{\sum_{3,4} w_j}$ 和 $\frac{w_3}{\sum_{2,3,4} w_j}\frac{w_2}{\sum_{2,4} w_j}$。目前主要有三种方法对结进行处理

1. **平均似然法**：使用 $\frac{1}{2}\frac{w_2}{\sum_{2,3,4} w_j}\frac{w_3}{\sum_{3,4} w_j} + \frac{1}{2}\frac{w_3}{\sum_{2,3,4} w_j}\frac{w_2}{\sum_{2,4} w_j}$ 对两项进行代替。
2. **Breslow 近似**：平均似然法的优点在于计算结果较为精确，但其缺点在于在计算时非常耗时。Breslow 近似对分母进行进行近似 $\sum_{2,3,4} w_j \sum_{3,4} w_j \approx \sum_{2,3,4} w_j \sum_{2,4} w_j \approx (\sum_{2,3,4} w_j)^2$。对于数据中的所有结，均对其分母进行近似。
3. **Efron 近似**：Breslow 近似对分母进行了高估，因此对于多结的情况，偏似然函数的计算会不准确。Efron 近似中，对分母的取值进行调整，$\sum_{2,3,4} w_j \sum_{3,4} w_j \approx \sum_{2,3,4} w_j \sum_{2,4} w_j \approx (\sum_{2,3,4} w_j)(\sum_{2,3,4} w_j - \bar{w}_j)$。当出现 n 个个体在同一时刻失效，则在近似时，分别在分母处减去 $1, 2, ..., n-1$ 倍的均值。

非参数部分估计 下面对模型的非参数部分 $\lambda_0(t)$ 进行估计。与 Kaplan-Meier 估计类似，我们可以将非参数部分的似然函数写为

$$L(\lambda) = \prod_j \left(\prod_{i \in D_j} \lambda_{ij} \prod_{i \in R_j - D_j} (1 - \lambda_{ij})\right)$$

其中 D_j 为在 j 时刻失效的个体集合，R_j 为在时刻 j^- 在险的个体集合。同时考虑到 $\lambda_i(t) = \lambda_0(t)\exp(x_i^T\beta)$，因此有

$$S_i(t) = \exp\left(-\int_0^t \lambda_i(s)ds\right) = \exp\left(-\int_0^t \lambda_0(s)\exp(x_i^T\beta)ds\right) = S_0(t)^{\exp(x_i^T\beta)} := S_0(t)^{w_i}$$

将 $\hat{S}_i(t) = \prod_{t_j \le t}(1 - \lambda_{ij})$ 和 $\hat{S}_0(t) = \prod_{t_j \le t}(1 - \lambda_{0j})$ 代入，可得到 $1 - \lambda_{ij} = (1 - \lambda_{0j})^{w_i}$，即 $\lambda_{ij} = 1 - (1 - \lambda_{0j})^{w_i}$。(注意：事实上，这步的推导是不严格的)

记 $\alpha_j = 1 - \lambda_{0j}$，则有

$$L(\lambda) = \prod_j \left(\prod_{i \in D_j} (1 - \alpha_j^{w_i}) \prod_{i \in R_j - D_j} \alpha_j^{w_i}\right)$$

当数据中不存在结时，我们可以最大化 $l(\lambda)$ 获得 α_j 的显式解 $\hat{\alpha}_j$。否则我们则需要通过迭代算法计算 $\hat{\alpha}_j$。对似然函数取对数，有

$$l(\lambda) = \sum_j \left[\sum_{i\in D_j} \ln(1-\alpha_j^{w_i}) + \sum_{i\in R_j - D_j} w_i \ln \alpha_j \right]$$

因此，

$$\frac{\partial l(\lambda)}{\partial \alpha_j} = \sum_{i\in D_j} \frac{-w_i \alpha_j^{w_i - 1}}{1-\alpha_j^{w_i}} + \sum_{i\in R_j - D_j} \frac{w_i}{\alpha_j}$$

当数据集中不存在结时，$|D_j| = 1$，记 $D_j = \{j\}$。因此若令 $\frac{\partial l(\lambda)}{\partial \alpha_j} = 0$，则有

$$\begin{aligned}
0 &= \frac{-w_j \alpha_j^{w_j - 1}}{1-\alpha_j^{w_j}} + \sum_{i\in R_j - D_j} \frac{w_i}{\alpha_j} \\
&= \frac{-w_j \alpha_j^{w_j - 1}}{1-\alpha_j^{w_j}} - \frac{w_j}{\alpha_j} + \sum_{i\in R_j} \frac{w_i}{\alpha_j} \\
&= \frac{1}{\alpha_j} \left(\frac{-w_j \alpha_j^{w_j}}{1-\alpha_j^{w_j}} - w_j + \sum_{i\in R_j} w_i \right) \\
&= \frac{1}{\alpha_j} \left(\frac{w_j}{1-\alpha_j^{w_j}} + W_j \right)
\end{aligned}$$

因此 $\hat{\alpha}_j = (1 - \frac{w_j}{W_j})^{1/w_j} = (1-\pi_{jj})^{1/w_j}$。代入生存函数，可得到生存函数的估计值

$$\hat{S}_i(t) = \hat{S}_0(t)^{\exp(x_i^T \beta)} = \prod_{t_j \le t} \hat{\alpha}_j^{\exp(x_i^T \beta)}$$

7 生存数据的机器学习模型

7.1 评判指标

在介绍机器学习模型之前，首先介绍生存模型的评价指标。

一致性指数 一致性指数 (Concordance Index, C-index) 是生存模型中最常用的评价指标。C-index 是 AUC 在生存数据上的扩展，它体现了模型预测的生存时间在排序上的正确情况。注意到 C-index 仅与

预测的排序有关，因此该指标对于比例风险模型非常有用 (这是因为比例风险的排序模型不会随时间变化，这使我们能够使用相对风险 $e^{x_i^T\beta}$ 直接进行比较)。C-index 的计算公式为

$$\text{C-index} = P\left(\hat{S}(T_j|x_j) < \hat{S}(T_i|x_i)|T_j < T_i, \delta_j = 1\right)$$

当模型是比例风险模型时，$\hat{S}(T_i|x_i) < \hat{S}(T_j|x_j)$ 等价于 $\eta_i = x_i^T\hat{\beta} > x_j^T\beta = \eta_j$，因此此时有

$$\text{C-index} = \frac{\sum_{i,j} I(T_j < T_i) I(\eta_j > \eta_i)\delta_j}{\sum_{i,j} I(T_j < T_i)\delta_j}$$

对于 C-index，当模型是最优时，所有预测的结果在排序上均为一致，C-index 等于 1；当模型的预测是随机时，C-index 为 0.5。注意到这一性质与 AUC 十分相似，因此 C-index 是 AUC 在生存数据上的拓展。此外还需注意到，在 C-index 的计算中，T_j 仅在未删失的数据上进行计算，但 T_i 在所有 T_j 时刻仍在险的个体上计算，包括删失的和未删失的。

Brier 分数 Brier 分数 (Breier Score, BS) 使用在二分类问题中时，其度量的是标签和预测概率的距离。假设对于一个样本个数为 N，其标签为 $y_i \in \{0,1\}$，预测第 i 个样本为 $y_i = 1$ 的概率为 p_i，则 BS 的计算公式为

$$\text{BS} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{p}_i)^2$$

当我们将 BS 拓展到生存数据上时，我们固定时间 t，对样本的存活时间是否大于或小于 t 做划分，并以此作为计算依据，BS 的计算公式为

$$\text{BS}(t) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{\hat{S}^2(t|x_i)I(T_i \le t)I(\delta_i = 1)}{\hat{G}(T_i)} + \frac{(1-\hat{S}(t|x_i))^2 I(T_i > t)}{\hat{G}(t)}\right]$$

其中 $\hat{G}(t)$ 为 t 时刻的 Kaplan-Meier 统计量，其对数据右删失情况下做校正。事实上，如果没有 $\hat{G}(t)$ 项的校正，该计算公式便会退化为原始的 BS 计算公式。对于生存时间 $T_i \le t$ 的个体，我们只计算未删失的数据；对于生存时间 $T_i > t$ 的个体，删失的和未删失的个体均需计算。对于一个有效的预测模型，对于任一时刻，BS 应小于 0.25。这是因为对于随机模型 $\hat{S}(t|x) = 0.5$，有 $BS(t) = 0.25, \forall t$。

$BS(t)$ 仅计算某一时刻 t 的的 BS，我们可以将其扩展到一段时间 $[t_1, t_2]$ 上，得到积分 Brier 分数 (IBS)

$$IBS = \frac{1}{t_2 - t_1}\int_{t_1}^{t_2} BS(s)ds$$

一般情况下，对上式求数值积分得到该值的近似值。我们仅在区间 $[t_1, t_2]$ 计算若干个时刻上的 BS，并使用类似梯形公式，辛普森公式等方法对积分计算近似值。

二项对数似然 二项对数似然 (Binomial Log-Likelihood, BLL) 对 BS 的计算公式进行微调，对生存时间使用如下公式进行计算，

$$\mathrm{BLL}(t)=\frac{1}{N}\sum_{i=1}^{N}\left[\frac{\log[1-\hat{S}(t|x_i)]I(T_i\leq t)I(\delta_i=1)}{\hat{G}(T_i)}+\frac{\log\hat{S}(t|x_i)I(T_i>t)}{\hat{G}(t)}\right]$$

同样地，二项对数似然也拓展到区间 $[t_1,t_2]$ 上，得到积分二项对数似然，

$$\mathrm{IBLL}=\frac{1}{t_2-t_1}\int_{t_1}^{t_2}\mathrm{BLL}(s)ds$$

7.2 随机生存森林

生存树 随机森林的构成元素是决策树。决策树的主要构件过程是节点的分裂过程。节点在分类时通过选取一个指标，该指标表示经过这次分裂，节点的子节点间的差异度的提升量。例如，ID3 算法使用信息增益，C4.5 算法使用信息增益比，CART 决策树使用 Gini 指数或 MSE 作为指标。当将决策树使用到生存数据上时，我们同样需要一个指标，该指标可以体现生存数据间的差异程度。

假设我们考虑一个二叉生存树，则该指标需要表示两个子节点所代表的两组生存数据的差异程度。Log-rank 统计量是最常用的指标，Log-rank 统计量表现了两组生存数据间的差异程度，该数值越大，表示两组数据间的差异越明显；该统计量越小，则两组生存数据间差异也越小。因此使用 Log-rank 统计量作为指标时，我们会对每一个变量 x 和该变量的所有取值 c 进行一次分割，按 $x<c$ 和 $x\geq c$ 切割成左右两个子树，并计算该分割下的 Log-rank 统计量。对于所有的 (x,c)，选择 Log-rank 统计量最大的 (x^*,c^*) 作为节点的分割标准。

该分割过程可迭代地在节点的左右节点上重复进行，直到满足一定条件停止分裂。这些条件可以是

- 该节点上仅一个未删失的数据 (必须停止)
- 树的深度达到给定的深度
- 该节点的样本个数达到给定的下限

生存树的叶子节点的预测可由以下方法得到。由于每个叶子节点必然含有至少一个未删失的样本，因此可以通过极大似然法得到该组数据的非参数估计，并将该非参数估计作为该节点所有样本的估计。假设叶子节点的数据为 $\{(t_1,\delta_1,x_1),(t_2,\delta_2,x_2),...,(t_N,\delta_N,x_N)\}$，则可使用节点上数据的 Kaplan-Meier 估计 $\hat{S}(t)=\prod_{t_j\leq t}\frac{n_j-d_j}{n_j}$ 或 Nelson-Aalen 估计 (也称为累计风险函数估计，Cumulative Hazard Function, CHF)$\hat{\Lambda}(t)=\prod_{t_j\leq t}\hat{\lambda}_j=\prod_{t_j\leq t}\frac{d_j}{n_j}$ 作为该节点的取值。注意到叶子节点的估计与 x_i 无关，x_i 的信息仅在树的构建过程中使用。

随机生存森林 随机森林通过种下多可各异的树，并对所有的树进行汇总，得到最终的输出结果。随机生存森林则通过集成多棵生存树，构建一个新的模型。为了使森林中的每棵树有一定的差异性，对每一棵树在训练时，同时对样本个数和变量个数均做随机筛选。样本个数通过 bootstrap 法随机选取，按照 bootstrap 法选取，每棵子树中大约包含 63% 的数据，另 37% 的数据在子树外，可用于验证模型使用。变量个数选择则表示在这棵子树划分时，仅可对选择出的变量进行划分，其余变量不予考虑。随机生存

森林中的生存树以累计风险函数估计作为其叶子节点的估计，并将每棵树的累计风险函数估计平均得到最终的累计风险函数估计。因此，随机生存森林的步骤如下：

1. 通过 bootstrap 法随机抽取得到 B 组不同的训练数据
2. 对每组数据，随机选取 p 个变量作为树中节点变量的可选集，通过 Log-rank 统计量不断对节点进行分裂，构建生存树。每棵树都要生长到最大的深度。
3. 对每棵生存树计算累计风险函数估计 $H_i(t|x)$，对它们求平均得到最终的累计风险函数估计

$$H^*(t|x) = \frac{1}{B}\sum_{i=1}^{B} H_i(t|x)$$

7.3 提升算法：CoxBoost

Boosting 方法 Boosting 方法先后构建多个基学习器，并使用加法模型融合这些学习器 $F(x) = \sum_{i=1}^{N} f_i(x)$。当构建 $f_i(x)$ 时，$f_1(x), f_2(x), ..., f_{i-1}(x)$ 均已构建完毕。假设此时的损失函数为 $L(y, F(x))$，则 $f_i(x)$ 可通过最小化损失函数获得

$$\hat{f}_i(x) = \arg\min_{f_i(x)} L\left(y, \sum_{j=1}^{i-1} \hat{f}_j(x) + f_i(x)\right)$$

其中 $\hat{f}_j(x), j = 1, 2, ..., i-1$ 是之前通过同样方法得到的估计。记 $F_i(x) = \sum_{j=1}^{i} \hat{f}_j(x)$，则损失函数为 $L(y, F_{i-1}(x) + f_i(x))$。由于梯度方向是函数值下降最快的方向，因此 $f_i(x)$ 可使用梯度的负方向进行拟合

$$\hat{f}_i(x) = -\frac{\partial L(y, F(x))}{\partial F(x)}|_{F(x)=F_{i-1}(x)}$$

以上所述的优化方法为基于梯度的 Boosting 方法，称为梯度提升法。梯度提升法的步骤可如下所示，

1. 初始化估计量，如 $\hat{F}(x) = 0$
2. 计算损失函数的负梯度 $g = -\frac{\partial L(y, F(x))}{\partial F(x)}|_{F(x)=\hat{F}(x)}$
3. 使用模型 $h(x;\theta)$ 拟合 g，得到 $\hat{h}(x;\theta)$
4. 更新估计 $\hat{F}(x) = \hat{F}(x) + \gamma\hat{h}(x;\theta)$

CoxBoost Cox 比例风险模型构建模型 $\lambda_i(x) = \lambda_0(x)\exp(x_i^T\beta)$，事实上相对风险部分 $x_i^T\beta$ 可以写为关于 x_i 的函数 $F(x_i;\beta)$。当我们将其表示为一个函数后，对应的对数偏似然函数变为

$$l(\beta) = \sum_{i=1}^{N} \delta_i \left[F(x_i;\beta) - \log \sum_{j\in R(t_i)} \exp(F(x_j;\beta))\right]$$

由于需要最大化 $l(\beta)$，因此 $-l(\beta)$ 可认为是损失函数。CoxBoost 对以下函数进行优化，

$$\hat{F}(x;\beta) = \arg\min_{F(x;\beta)} -\sum_{i=1}^{N} \delta_i \left[F(x_i;\beta) - \log \sum_{j\in R(t_i)} \exp(F(x_j;\beta)) \right]$$

对以上优化问题，可使用一般的梯度提升算法得到 $\hat{F}(x;\beta)$ 的求解。CoxBoost 仍然假定 $F(x_i;\beta)$ 是一个线性函数，不过我们也可简单的将其改为非线性函数，以提升模型的表达能力。此外 CoxBoost 还对 β 进行了正则化操作，因此最终的优化问题为

$$\hat{\beta} = \arg\min_{\beta} -\sum_{i=1}^{N} \delta_i \left[x_i^T\beta - \log \sum_{j\in R(t_i)} \exp(x_j^T\beta) \right] + \frac{\lambda}{2}\beta^T\beta$$

在此优化函数下，CoxBoost 方法与 Cox 比例风险模型十分类似，CoxBoost 方法只使用一阶导数求解 β，而 Cox 比例风险模型使用到了二阶导数。但是 CoxBoost 方法可以简单的将其拓展到非线性函数拟合上，提升模型的表示能力。

7.4 神经网络

Cox 比例风险模型假定了风险函数和自变量间的线性关系，但是线性关系的假定降低了模型的表示能力。部分变量的非线性关系无法很好的度量，变量间的交互效应也无法很好的展现。因此，一个自然的想法是对线性关系进行改进，将其升级为一个非线性的关系。如果使用神经网络刻画非线性关系，则 Cox 比例风险模型的形式会变为

$$\lambda_i(t) = \lambda_0(t)\exp\left(F_\theta(x_i)\right)$$

$F_\theta(x)$ 为神经网络，其参数为 θ，θ 在优化时同样可通过优化对数偏似然得到

$$\hat{\theta} = \arg\min_{\theta} -\sum_{i=1}^{N} \delta_i \left[F_\theta(x_i) - \log \sum_{j\in R(t_i)} \exp(F_\theta(x_j)) \right]$$

由于 $F_\theta(x)$ 的结构比较复杂 (其结构取决于自身的设计)，我们一般使用梯度下降法对 θ 进行求解，记

$$l(\theta) = -\sum_{i=1}^{N} \delta_i \left[F_\theta(x_i) - \log \sum_{j\in R(t_i)} \exp(F_\theta(x_j)) \right]$$

则 θ 的更新过程为 $\theta := \theta - \alpha\frac{\partial l(\theta)}{\partial\theta}$。

随机梯度下降法 注意到之前的梯度下降法的更新方式直接在所有 N 个样本上对梯度进行求和，这种梯度下降法使用的是非随机梯度下降法，但是对于神经网络，最为常见的训练方式为随机梯度下降法。

随机梯度下降法的方法是，将全体数据分为若干的批次 (批次的常见大小为 32，64，128 等)，在每一个批次内更新参数时，用该批次内样本的梯度近似整体的梯度。随机梯度下降法可以在每次参数更新时减少计算量，并通过并行加快训练的速度。

不过使用随机梯度下降法训练 Cox-神经网络模型时仍然存在一定问题。首先，对于每项的梯度计算，需要计算在险集合 $R(t_i)$ 上的所有梯度。该集合可能非常大，甚至可能和整体样本的个数一样大，这在计算过程中需要消耗大量的时间。此外，由于一般神经网络模型较大，其计算时间较大，因此对于在险集合 $R(t_i)$ 内的所有样本进行计算也需要很多的时间。综上所述，如果不对普通的随机梯度下降法进行修改，该方法难以在该模型的优化中表现良好。尤其对于数据量大且复杂，网络结构复杂的情况下，我们更需要找到解决的方法。

第一种方法是使用批次内的样本 $\tilde{R}(t_i)$ 代替在险集合 $R(t_i)$。假设批次的大小为 B，此时批次内的损失函数使用以下函数进行近似

$$\tilde{l}(\theta) = -\sum_{i=1}^{B} \delta_i \left[F_\theta(x_i) - \log \sum_{j \in \tilde{R}(t_i)} \exp(F_\theta(x_j)) \right]$$

对参数更新时使用 $\theta := \theta - \alpha \frac{\partial \tilde{l}(\theta)}{\partial \theta}$ 即可。从中可以看到此方法的核心为，使用抽样得到的批次数据近似全体数据，从而使用批次内的在险集合近似整体的在险集合，降低计算量。

第二种方法仍然使用抽样的方法。我们知道在 Cox-神经网络模型中，损失函数为

$$\begin{aligned}
l(\theta) &= -\sum_{i=1}^{N} \delta_i \left[F_\theta(x_i) - \log \sum_{j \in R(t_i)} \exp(F_\theta(x_j)) \right] \\
&= -\sum_{i=1}^{N} \delta_i \log \frac{\exp(F_\theta(x_i))}{\sum_{j \in R(t_i)} \exp(F_\theta(x_j))} \\
&= \sum_{i=1}^{N} \delta_i \log \left[\sum_{j \in R(t_i)} \frac{\exp(F_\theta(x_j))}{\exp(F_\theta(x_i))} \right] \\
&= \sum_{i=1}^{N} \delta_i \log \left[\sum_{j \in R(t_i)} \exp\left[F_\theta(x_j) - F_\theta(x_i) \right] \right]
\end{aligned}$$

因此，损失函数等价于

$$\begin{aligned}
\text{loss} &= \frac{1}{N} \sum_{i=1}^{N} \delta_i \log \left[\sum_{j \in R(t_i)} \exp\left[F_\theta(x_j) - F_\theta(x_i) \right] \right] \\
&= \frac{1}{N} \sum_{\delta_i = 1} \log \left[1 + \sum_{j \in R(t_i) - \{i\}} \exp[F_\theta(x_j) - F_\theta(x_i)] \right]
\end{aligned}$$

此时对于损失函数的计算，大部分计算量位于 $\sum_{j\in R(t_i)-\{i\}}\exp[F_\theta(x_j)-F_\theta(x_i)]$。我们可以通过对集合 $R(t_i)-\{i\}$ 采样降低计算时间，研究表明每次从集合中抽取一个样本便可得到不错的优化效果，因此有

$$\text{loss}=\frac{1}{N}\sum_{\delta_i=1}\log\left[1+\exp[F_\theta(x_j)-F_\theta(x_i)]\right],\quad j\in R(t_i)-\{i\}$$

参数的更新方法变为 $\theta:=\theta-\alpha\frac{\partial\text{loss}}{\partial\theta}$。

8 残差和模型诊断

模型诊断可以用来评估模型的合理程度，尤其对于统计模型，其可检验模型的假设是否成立。

对于生存模型的残差，我们基于一个非常重要的定理。假定 T 是一个连续非负的随机变量，其累计生存函数为 Λ，则随机变量 $Y=\Lambda(T)$ 服从参数为 1 的指数分布。这是因为

$$\begin{aligned}S_Y(y)&=P(Y>y)=P(\Lambda(T)>y)=P(T>\Lambda^{-1}(y))\\&=S_T(\Lambda^{-1}(y))=\exp\left(-\Lambda(\Lambda^{-1}(y))\right)=e^{-y}\end{aligned}$$

因此 $Y\sim\text{Exp}(1)$。通过这一定理，我们可以通过比较 $\hat{\Lambda}(t_i)$ 和 $\text{Exp}(1)$ 来判断模型的可行性。

Cox-Snell 残差 Cox-Snell 残差定义为 $e_i:=\hat{\Lambda}(T_i)$。如果模型满足比例风险模型，则 $e_i=\hat{\Lambda}_0(t)\exp(x_i^T\beta)$。使用 Cox-Snell 残差，我们可以通过画出 $\{e_i\}$ 和标准指数分布进行检验。

此外，我们可以画出 Nelson-Aalen 估计，如果模型恰当，则 Nelson-Aalen 估计应该位于直线 $y=x$ 附近。不过使用此方法时，在 t 较小时的残差可能不宜察觉出其是否偏离了直线。

Cox-Snell 残差在分布上有较好的性质，但其数值大小不具有可比较性，我们无法通过数值判断出哪个样本违反了模型假定。

鞅残差 鞅残差的定义为 $m_i:=\delta_i-\hat{\Lambda}(t_i)=\delta_i-\hat{e}_i$。鞅残差在数值上为正时表示该个体比预期失效的时间早 (根据模型)，数值为负时表示该个体的存活时间比预期的多。

下面解释为什么 m_i 被称为鞅残差。假设数据为 $\{T_i,\delta_i\}$，定义 $Y_i(t)=I(t_i\geq t),N_i(t)=\delta_iI(T_i\leq t)$。定义 $M_i(t)=N_i(t)-\int_0^tY_i(s)d\Lambda_i(s)$，则

$$
\begin{aligned}
M_i(t_i) &= \delta_i I(T_i \le T_i) - \int_0^{T_i} Y_i(s) d\Lambda_i(s) \\
&= \delta_i - \int_0^{T_i} I(T_i \ge s) d\Lambda_i(s) \\
&= \delta_i - \int_0^{\max(T_i, T_i)} d\Lambda_i(s) \\
&= \delta_i - \Lambda_i(t_i) + \Lambda(0) \\
&= \delta_i - \Lambda_i(t_i)
\end{aligned}
$$

而 $M_i(t)$ 符合鞅的两条性质（$1.E[M(t)] = 0$; 2. 对任意 $s < t$，有 $E[M(t)|M(s)] = M(s)$），因此将 $M_i(t_i) = \delta_i - \Lambda_i(t_i)$ 称为鞅残差。鞅残差具有上界 1，但没有下界，且删失时，鞅残差小于 0，因此鞅残差的对称性不佳。

Deviance 残差 Deviance 残差的定义为 $\hat{d}_i = \mathrm{sign}(\hat{m}_i)\sqrt{2(\tilde{I}_i - I_i)}$，其中 $\tilde{I}_i$ 为饱和模型的偏似然函数值，I_i 为该个体的偏似然函数值。经过化简，有

$$
\hat{d}_i = \mathrm{sign}(\hat{m}_i)\sqrt{-2[\hat{m}_i + \delta_i \ln(\delta_i - \hat{m}_i)]}
$$

Deviance 残差对称性较好并对检测异常点较有帮助。