Survival Analysis

$Censored\ Data$ <删失数据>

- 生存分析主要处理一种特殊的时间数据，并且时间数据可能带有部分删失属性
- 盲目的删除 Censoring Data 会 **提高模型的方差**，以至于 **降低估计精度**
- 删失数据产生原因：

$$\left\{\begin{array}{l}\text{因实验终止而无法观测} \Longrightarrow \text{恢复较好} \\ \text{因情况变差而自动退出} \\ \text{因前往其他地区治疗而无法参与实验}\end{array}\right\} \Longrightarrow \text{恢复可能不理想}$$

- 删失数据如何带来偏差：

$$\text{不考虑病人提前退出的原因} \Longrightarrow \text{过高估计存活时间}$$

$T's\ pdf\quad f(t)$ <事件的概率密度函数>

$$f(t) = P(T = t) = p(t)$$

- 量化事件在 t 时间 **时** 发生的概率

$T's\ cdf\quad F(t)$ <事件的分布函数>

$$F(t) = P(T \leq t)$$

$Survival\ Function\quad S(t)$ <生存函数>

$$S(t) = 1 - F(t) = P(T > t)$$

- 量化事件在 t 时间 **之后** 发生的概率

$$\left\{\begin{array}{lll} Censored\ Data \Longrightarrow Died & Underestimate\ S(t) \\ Censored\ Data \Longrightarrow Alive & Overestimate\ S(t) \end{array}\right.$$

- 重要结论：

$$S'(t) = -F'(t) = -f(t) \leftrightarrow f(t) = -S'(t)$$

$Hazard\ Function$ $\lambda(t)$ <风险函数>

$$\begin{aligned}\lambda(t) &= \lim_{h \to 0^+} \underbrace{\frac{p(t < T \leq t+h | T > t)}{h}}_{\text{拆分条件概率}} \\ &= \frac{1}{P(T > t)} \lim_{h \to 0^+} \frac{P(t < T \leq t+h)}{h} \\ &= \frac{f(t)}{S(t)}\end{aligned}$$

- 量化事件在 t 时间 **之后瞬时发生的可能性（不是概率）**
- 有一个重要结论：$\lambda(t) = -\frac{d}{dt} lnS(t)$

$$\begin{aligned}\lambda(t) &= \frac{f(t)}{S(t)} = \frac{d}{dt}\int \frac{f(t)}{S(t)} dt \\ &= \frac{d}{dt}\int \frac{1}{S(t)} d\Big(-S(t)\Big) \\ &= -\frac{d}{dt} lnS(t)\end{aligned}$$

$Cumulative\ Hazard\ Function$ $\Lambda(t)$ <累积风险函数>

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

- 量化事件在 t 时间点前的风险和
- 有一个重要结论：$S(t) = \exp\{-\Lambda(t)\}$

$Mean\ Residual\ Life \quad r(t)$ <累积风险函数>

$$r(t) = E(T - t | T \geq t) = \frac{\int_t^\infty S(u)du}{S(t)}$$

- 量化事件在 t 时间 **后** 发生的期望值
- 证明过程:

$$Lemma\ 1:\ E(x) = \int_{-\infty}^{\infty} x \cdot p(x)dxdx = \int_0^\infty P(X \geq x)dx$$

$$\begin{aligned} Proof:\ E(x) &= \int_{-\infty}^{\infty} x \cdot p(x)dx \\ &= \int_{-\infty}^{\infty} \int_0^x p(y)dydx \\ &= \int_{-\infty}^{\infty} \int_x^{\infty} p(y)dydx = \int_0^\infty 1 - F(x)dx \ \square \end{aligned}$$

$$Lemma\ 2:\ E(x|y) = \int_{-\infty}^{\infty} x \cdot p(x|y)dx = \int_{-\infty}^{\infty} x \cdot \frac{p(x,y)}{p(y)}dx$$

$Proof:$ 显然

Proof : Mean Residual Life

$$r(t) = E(T-t|T>t) = \underbrace{\int_{-\infty}^{\infty} P(T-t>s|T>t)ds}_{\text{Lemma 1}}$$

$$= \underbrace{\int_{-\infty}^{\infty} \frac{P(T-t>s, T>t)}{P(T>t)}ds}_{\text{Lemma 2}}$$

$$= \int_{-\infty}^{\infty} \frac{P(T-t>s)}{P(T>t)}ds = \frac{1}{S(t)}\int_{-\infty}^{\infty} P(T-t>s)ds$$

$$= \frac{1}{S(t)}\int_{-\infty}^{\infty} P(T>t+s)ds$$

$$= \frac{1}{S(t)}\int_{t}^{\infty} P(T>t+s)d(t+s)$$

$$\underset{u=t+s}{\Longrightarrow} \frac{1}{S(t)}\int_{t}^{\infty} P(T>u)d(u) = \frac{\int_t^{\infty} S(u)du}{S(t)} \square$$

Relationship **<函数之间的关系>**

Likelihood **<删失数据的似然>**

Definition of Survival Data **<生存数据的定义>**

$$\{t_i, \delta_i, x_i\}_{i=1}^{n}$$

- t_i *fellowing-up time* 跟进时间:

$$\begin{cases} \text{左截断or删失:} & \max\{\underset{\text{生存时间}}{T_i}, \underset{\text{删失时间}}{C_i}, \underset{\text{截断时间}}{Tr_i}\} \\ \text{右截断or删失:} & \min\{\underset{\text{生存时间}}{T_i}, \underset{\text{删失时间}}{C_i}, \underset{\text{截断时间}}{Tr_i}\} \end{cases}$$

- δ_i *state* 状态:

$$\begin{cases} \text{删失} \begin{cases} \delta_i = 1 & \text{if } t_i = T_i \\ \delta_i = 0 & \text{if } t_i = C_i \end{cases} \\ \text{截断} \begin{cases} \delta_i = 1 & \text{if } t_i = T_i \\ \delta_i = 0 & \text{if } t_i = Tr_i \end{cases} \end{cases}$$

- $x_i \quad covariates$ 协变量: weight, height, blood pressure...

Likelihood of Survival Data <生存数据的似然>

- 似然函数：$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta) = \prod_{i=1}^{n} L_i$
- 生存数据的似然函数：$L(\theta) = \prod_{i=1}^{n} \Big(f(t_i;\theta)\Big)^{\delta_i} \Big(S(t_i;\theta)\Big)^{1-\delta_i}$
- 删失*or*截断类型不同，L_i 的形式也不同

$$\begin{cases} \text{正常数据}(T_i = t_i) & \to L_i = f(t_i;\theta) \\ \text{右删失数据}\ (T = t_i+)(T_i \geq t_i) & \to L_i = S(t_i;\theta) \\ \text{左删失数据}\ (T = t_i-)(T_i \leq t_i) & \to L_i = F(t_i;\theta) \\ \text{区间删失数据}\ (l_i \leq T_i \leq r_i) & \to L_i = F(r_i;\theta) - F(l_i;\theta) \\ \text{右截断数据}\ (T_i = t_i | T < u) & \to L_i = \dfrac{f(t_i)}{F(u)} \\ \text{左截断数据}\ (T_i = t_i | T > u) & \to L_i = \dfrac{f(t_i)}{S(u)} \end{cases}$$

- Ps: 不完全数据举例

$$\begin{cases} \text{右删失：实验结束之后才发病，但无法记录具体时间} \\ \text{左删失：实验开始之前就发病，但不记得具体时间} \\ \text{区间删失：实验只在两个周一之间提取数据，} \\ \qquad\qquad \text{患者在其间发病，不知道发病确切时间} \\ \text{右截断：实验只考虑30岁及以下的病人} \\ \text{左截断：实验只考虑50岁及以上的病人} \end{cases}$$

Type of Censored Data <删失数据的类型>

$$\begin{cases} \text{Type I Censoring: 删失时间固定。eg:实验只进行5天} \\ \text{Type II Censoring: 当失败样本达到一定比例后结束实验} \\ \text{Random Censoring: 样本删失情况是随机的} \end{cases}$$

- **以上三种删失都不会影响似然函数结果**
- **重要结论:**

$$\textit{If: } C \perp T|X \quad \textit{Then: } L_i(\theta) \propto f(t_i|\theta, X)^{\delta_i} S(t_i|\theta, X)^{1-\delta_i}$$

只要删失 C 与事件发生时间 T 无关, 似然函数 $L(\theta)$ 就不受影响

Kaplan Meier Estimator <K-M 估计>

Empirical Distribution <经验分布>

- *Empirical Distribution* :

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{X_i \le x} = \begin{cases} 0, & \text{if } x \le X_{(1)} \\ \dfrac{k}{n}, & \text{if } X_{(k)} < x \le X_{(k+1)} \\ 1, & \text{if } x > X_{(n)} \end{cases}$$

- 经验分布 **依概率1收敛(几乎处处收敛)** 于真实分布, **(格里汶科定理)**

$$p\left(\lim_{n\to\infty} F_n(x) = F(x)\right) = 1 \implies F_n(x) \xrightarrow{a.s.} F(x)$$

Kaplan Meier Estimator <K-M 估计>

- 使用经验函数来拟合事件的分布函数 $F(t)$，进而拟合出生存函数 $S(t)$。但删失数据会给估计带来偏差。
- 引入条件概率减小偏差:

$$S(t) = P(T > t)$$
$$\hat{S}(t) = P(T > t_i | T > t_{i-1}) \cdot \hat{S}(t_{i-1})$$

- n_j j 时刻剩余人数
  d_j j 时刻发生终点事件的人数
  n 总人数

$$\begin{aligned}
\because \hat{S}(t_j) &= P(T > t_j | T > t_{j-1}) \cdot \hat{S}(t_{j-1}) \\
&= \frac{P(T > t_j, T > t_{j-1})}{P(T > t_{j-1})} \cdot \hat{S}(t_{j-1}) \\
&= \underbrace{\frac{n_j - d_j}{n} \div \frac{n_j}{n}}_{\text{Empirical Distribution}} \cdot \hat{S}(t_{i-1}) \\
&= \frac{n_j - d_j}{n_j} \cdot \hat{S}(t_{i-1}) = \left(1 - \frac{d_j}{n_j}\right) \cdot \hat{S}(t_{i-1})
\end{aligned}$$

$$\because \hat{\lambda}_i = \frac{f(t_j)}{S(t_j)} = \frac{f(t_j)}{P(T > t_j)} = \underbrace{\frac{d_j}{n} \div \frac{n_j}{n}}_{\text{Empirical Distribution}} = \frac{d_j}{n_j}$$

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

- Kaplan Meier Estimator 可以有另一种视角来解释，即 MLE
  - 在该视角下，似然函数最大时，$\hat{\lambda}_j \backsim Binomial(n_j, p = \lambda_j)$

Delta Method

$$Assume: \ \sqrt{n}(\hat{\lambda_j} - \lambda_j) \xrightarrow{d} N(0, \sigma^2)$$
$$\text{一元} \implies \sqrt{n}(g(\hat{\lambda}_j) - g(\lambda_j)) \xrightarrow{d} N(0, [g'(\lambda_j)]^2 \sigma^2)$$
$$\text{多元} \implies \sqrt{n}(G(\hat{\Lambda}_j) - G(\Lambda_j)) \xrightarrow{d} N(0, \nabla G(\Lambda_j)^T \cdot \Sigma \cdot \nabla G(\Lambda_j))$$

$Proof:$

$Taylor$一阶展开 $\Longrightarrow g(\hat{\lambda_j}) = g(\lambda_j) + g'(\lambda_j)(\hat{\lambda_j} - \lambda_j)$

$\Longrightarrow g(\hat{\lambda_j}) - g(\lambda_j) = g'(\lambda_j)(\hat{\lambda_j} - \lambda_j) \xrightarrow{d} N(0, [g'(\lambda_j)]^2\sigma^2)$

$Confidence\ Interval\ of\ Kaplan\ Meier\ Estimator$

- 利用 $Delta\ Method$ 来求出 $K-M\ Estimator$ 的置信区间

$$\begin{cases} 1.\sqrt{n}(\log \hat{S}(t) - \log S(t)) \xrightarrow{d} N(0, \sigma_t^2) \Longrightarrow g(x) = \log x \\ 2.\sqrt{n}(\hat{S}(t) - S(t)) \xrightarrow{d} N(0, \sigma_t^2[S(t)]^2) \Longrightarrow g(x) = x \\ 3.\sqrt{n}(\hat{S}(t) - S(t)) \xrightarrow{d} N(0, \frac{\sigma_t^2}{[\log S(t)]^2}) \Longrightarrow g(x) = \log[-\log x] \end{cases}$$

- $\sigma_t^2 = n\sum_{t_j<t} \dfrac{d_j}{n_j(n_j - d_j)}$
- 置信区间:

$$-N_{\alpha/2} \leq \frac{g[\hat{S}(t)] - g(Survival\ Rate)}{\sqrt{Var\left[g(\hat{S}(t))\right]}} \leq N_{\alpha/2}$$

$Group\ Testing$ <分组检验>

$Contingency\ Table$ <列联表>

- 展示的为行数 R、列数 C 都为2时的列联表, 行数列数可增加

	C_1	C_2	$Total$
R_1	a	b	$a+b$
R_2	c	d	$c+d$
$Total$	$a+c$	$b+d$	$a+b+c+d=n$

$Pearson\ \chi^2\ Test$ <皮尔逊卡方检验>

- 皮尔逊卡方检验常检验 **列联表的拟合优度**$_1$ 和 **列联表行列是否相关**$_2$
- 利用了 **<u>大样本下渐近卡方分布的性质</u>**
- $1 \implies H_0$：理论频数与实际频数相同
  $2 \implies H_0$：行列因素相互独立
  统计量：

$$\chi^2 = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{(f_o - f_e)^2}{f_e}$$

$$if:\ \chi^2 > \chi^2_\alpha(df) \implies Then:\ refuse\ H_0$$

f_o：实际频数 $Eg: f_o = a$
F_e：理论频数 $Eg: f_e = \dfrac{(a+b)}{n}\cdot\dfrac{(a+c)}{n}\cdot n = \dfrac{(a+b)(c+d)}{n}$
df：自由度 $df = (R-1)\times(C-1)$

- 两个问题其实是等价的。若行列因素相互独立$_2$，则可通过相互独立两事件的积事件的定义 $P(AB) = P(A)P(B)$ 由边际分布求出联合概率密度。同时验证了同样用此方法求出的理论频数的正确性$_1$。

$Fisher's\ Exact\ Test$ <Fisher确切概率法>

- 利用了 **<u>超几何分布的定义</u>**
- $Fisher$ 确切概率法直接计算出 $p-value$

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{\binom{a+b}{b}\binom{c+d}{d}}{\binom{n}{b+d}}$$
$$= \frac{(a+b)!\ \ (c+d)!\ \ (a+c)!\ \ (b+d)!}{a!\ \ b!\ \ c!\ \ d!\ \ n!}$$

$$if:\ p < \alpha \implies Then:\ refuse\ H_0$$

- 若为多行多列数据，可使用 **多元超几何分布**
- Fisher确切概率法使用条件：
  某个格子的理论频数 $f_e < 5$ 或 $p - value \approx \alpha$ 时
  使用Fisher确切概率法

$McNemar's\ Test$ <McNemar检验>

- McNemar检验用来检验 **成对数据** 。即同一样本采用不同方法产生的数据,且只能在使用 2×2 列联表上使用

	$Test1\ pos$	$Test1\ neg$	$Total$
$Test2\ pos$	a	b	$a+b$
$Test2\ neg$	c	d	$c+d$
$Total$	$a+c$	$b+d$	$a+b+c+d$

- H_0：两方法效果相同 $\implies Assume:\ \begin{cases} a+c=a+b \\ b+d=c+d \end{cases}$

  $\implies b=c$
- 统计量：

$$\chi^2 = \frac{(b-c)^2}{b+c}$$

$$if:\ \chi^2 > \chi_\alpha^2(1) \implies Then:\ refuse\ H_0$$

- **此时，两组数据之间不相互独立的，因此皮尔逊 χ^2 统计量失效。使用 $\frac{b+c}{2}$ 来作为理论频数，并使用 χ^2 统计量的计算方法即可得出此统计量。并证明 $b=c$**
- $\mathbf{Binomial} \xrightarrow{limit} \mathbf{Poisson} \xrightarrow{limit} \mathbf{Normal}$

$Cochran-Mantel-Haenszel\ Test$ <分层卡方检验>

$Layer\ i$	$Treat$	$Non-treat$	$Total$
$Case$	a_i	b_i	a_i+b_i
$Control$	c_i	d_i	c_i+d_i
$Total$	a_i+c_i	b_i+d_i	$a_i+b_i+c_i+d_i=n_i$

- CMH检验常用在多中心实验中，以去除实验中的中心效应。
- 几率 $Odds$: 是一种替代概率的概念，它没有上下界，因此可以用多种模型对其进行拟合。

$$Odds(A)=\frac{P(A)}{P(\bar{A})}=\frac{P(A)}{1-P(A)}$$

- 优势比 OR : 在不受实验组和对照组比例的条件下，得到因果之间的联系。

$$OR=\frac{\frac{a_i}{a_ic_i}}{\frac{c_i}{a_ic_i}}\Big/\frac{\frac{b_i}{b_id_i}}{\frac{d_i}{b_id_i}}=\frac{\frac{a_i}{c_i}}{\frac{b_i}{d_i}}=\frac{a_id_i}{b_ic_i}$$

- CMH检验使用多层优势比来去除中心效应。

$$OR=\frac{\sum_{i=1}^{T}\frac{a_id_i}{n_i}}{\sum_{i=1}^{T}\frac{b_ic_i}{n_i}}$$

- $H_0:OR=1$ 结局与治疗方案无关
- 统计量

$$\chi^2 = \frac{\left[\sum_{i=1}^{T}(a_i - \frac{(a_i+b_i)(a_i+c_i)}{n_i})\right]^2}{\frac{(a_i+b_i)(a_i+c_i)(b_i+c_i)(b_i+d_i)}{n_i^2(n_i-1)}}$$

相当于对a_i进行了Z-Score 标准化

$$if:\ \chi^2 > \chi^2_\alpha(1) \implies Then:\ refuse\ H_0\ \&\ calculate\ OR$$

- 显然，此时 a_i 服从超几何分布; 且注意，此时 $df = 1$ 是因为累加的是各实验中心的样本数，而并非单个特征的类数 (行数 or 列数)，只有行列增加时自由度是增加的，只增加样本数不改变自由度。
- 流程：进行CMH检验 $\longrightarrow$ 计算多层优势比 OR

$Log-Rank\ \ Test$ <Log-Rank 检验>

- $\begin{cases} H_0 : S_1 = S_2 & Group = 2 \\ H_0 : S_1 = \cdots = S_n & Group = n \end{cases}$

  与下式等价

  $\begin{cases} H_0 : \lambda_1(t) = \lambda_2(t) & Group = 2 \\ H_0 : \lambda_1(t) = \cdots = \lambda_n(t) & Group = n \end{cases}$
- $Group = 2$
  - 需指出, $Group = 2$ 时, Log-Rank test与CMH test 相同

$Layer\ t$	$Control$	$Case$	$Total$
$Treat$	d_{1t}	d_{2t}	d_t
$Non-treat$	$n_{1t}-d_{1t}$	$n_{2t}-d_{2t}$	n_t-d_t
$Total$	n_{1t}	n_{2t}	n_t

  - 此时，将存在终点事件的不同时间点 t 看作 CMH 检验中不同的层。

- 统计量

$$\chi^2 = \frac{W^2}{V}$$
$$W = \sum_{t=1}^{T} w_t; V = \sum_{t=1}^{T} v_t$$
$$w_t = d_{1t} - n_{1t}\frac{d_t}{n_t}$$
$$v_t = \frac{n_{1t}n_{2t}d_t(n_t - d_t)}{n_t^2(n_t - 1)}$$
$$if:\ \chi^2 > \chi^2_\alpha(1) \implies Then:\ refuse\ H_0$$

- 该统计量先对 d_{1t} 进行 Z-Score 标准化；则 w_t 渐进服从正态分布，再使用正态分布的可加性即可证明 W 服从正态分布，将其除于 V 后，即可证得 χ^2 渐进服从 χ^2 分布。

- $Group = n$
  - Log-Rank test在 $Group = n$ 时的情况将CMH test 拓展到了多分类的情况
  - 统计量

$$\chi^2 = W^T V^{-1} W$$
$$W = \sum_{t=1}^{T} w_t; V = \sum_{t=1}^{T} V_t$$
$$w_t = \left(d_{1t} - n_{1t}\frac{d_t}{n_t}, \cdots, d_{pt} - n_{pt}\frac{d_t}{n_t}\right)$$

$$\begin{cases} (V_t)_{ii} = \dfrac{(n_j - n_{it})n_{it}d_t(n_t - d_t)}{n_t^2(n_t - 1)} & \text{对角线元素} \\ (V_t)_{ij} = \dfrac{n_{it}n_{jt}d_t(n_t - d_t)}{n_t^2(n_t - 1)} & \text{非对角线元素} \end{cases}$$

$$if:\ \chi^2 > \chi^2_\alpha(p-1) \implies Then:\ refuse\ H_0$$

- **Weighted Log-Rank Test**

- 为了使 **样本多的时候更大的权重** 而引入。

$$W = \sum_{t=1}^{T} \alpha_t w_t; \quad V = \sum_{t=1}^{T} \alpha_t^2 v_t$$

$$\begin{aligned} \alpha_t &= n_j && \Longrightarrow \text{Gehan-Breslow Test} \\ \alpha_t &= \hat{S}(t) && \Longrightarrow \text{Peto-Prentice Test} \\ \alpha_t &= \left(\hat{S}(t)\right)^{\rho} && \Longrightarrow \text{Peto-Prentice Test} \end{aligned}$$

- **Stratified Log-Rank Test**
  - 为了 **控制不同变量** 而引入。
  - $H_0 : S(t|k) = S(t|k)$
    在固定因素k的情况下二者是否相等
  - 统计量:

$$\chi^2 = \frac{W^2}{V} = \frac{(\sum_k \sum_t w_{tk})^2}{\sum_t v_{tk}}$$

$$if:\ \chi^2 > \chi^2_\alpha(p-1) \implies Then:\ refuse\ H_0$$

Likelihood Testing <似然检验>

Score Function & Fisher Information

- 得分向量与 Fisher 信息阵与似然函数相关，下面将给出相关定义。
  - 单参数

$$\text{Likelihood Function:} \quad L(\theta) = \prod_i p(x_i, \theta)$$
$$l(\theta) = \ln L(\theta)$$
$$\text{Score Function:} \quad U(\theta) = \frac{\partial}{\partial \theta} l(\theta)$$
$$\text{Fisher Information:} \quad I(\theta) = -E_X(\frac{\partial}{\partial \theta} U(\theta))$$
$$= -E_X(\frac{\partial^2}{\partial \theta^2} l(\theta))$$

◦ 多参数

$$\text{Likelihood Function:} \quad L(\vec{\theta}) = \prod_i p(x_i, \vec{\theta})$$
$$l(\vec{\theta}) = \ln L(\vec{\theta})$$
$$\text{Score Function:} \quad U(\vec{\theta}) = \frac{\partial}{\partial \theta_i} l(\vec{\theta})$$
$$= \nabla_\theta l(\vec{\theta})$$
$$\text{Fisher Information:} \quad I(\vec{\theta}) = -E_X(\frac{\partial}{\partial \theta_j} U(\vec{\theta}))$$
$$= -E_X(\nabla_\theta U(\vec{\theta}))$$
$$= -E_X(\frac{\partial^2}{\partial \theta_i \theta_j} l(\vec{\theta}))$$
$$\text{(Matrix)}$$

- 下面给出几条 **重要结论**：
  i. $U(\theta)$ 其实也是关于 X 的函数
  ii. $E_X\Big[U(\theta)\Big] = 0$
  iii. $Var_X\Big[U(\theta)\Big] = I(\theta)$
  iv. if: X is iid $\Longrightarrow$
  $U(\theta) = \sum_i U(x_i; \theta)$ & $I(\theta) = \sum_i I_i(\theta)$
  v. 在计算时，可以使用近似值 $\tilde{I}(\theta)$ 替代 $I(\theta)$

$$\tilde{I}(X) \overset{E(x)=\frac{1}{n}\sum_x x_i}{=\!=\!=} \frac{\partial}{\partial\theta}U(-E_X(X);\theta)$$

vi. 由 $C-R$ 不等式 $Var[g(\hat{X})] \geq \frac{g'(\theta)}{I(\theta)}$ 可知, 当 $g(\hat{X})=\theta$ 时, 有 $Var[\theta] \geq \frac{1}{I(\theta)}$, 即 $Var[\hat{\theta}_{MLE}] = I^{-1}(\beta)$

Score Test

- $H_0 : \theta = \theta_0$
- 统计量: (理论来自于 $U(x;\theta)$ 的渐进分布)

$$\frac{U(\theta_0)}{\sqrt{I(\theta_0)}} \xrightarrow{d} N(0,1) \quad \text{单参数}$$

$$U(\vec{\theta_0})^T I(\vec{\theta_0})\, U(\vec{\theta_0}) \xrightarrow{d} \chi^2(p) \quad \text{多参数}$$

Wald Test

- $H_0 : \theta = \theta_0$
- 统计量: (理论来自于 $U(x;\theta)$ 的 Taylor 一阶展开)

$$\sqrt{I(\theta)}(\hat{\theta}-\theta_0) \xrightarrow{d} N(0,1) \quad \text{单参数}$$

$$(\hat{\theta}-\vec{\theta_0})^T I(\hat{\theta})\, (\hat{\theta}-\vec{\theta_0}) \xrightarrow{d} \chi^2(p) \quad \text{多参数}$$

- $U(\theta)$ 的 Taylor 一阶展开:

$$U(\theta) \approx U(\hat{\theta}) - I(\hat{\theta})(\theta-\hat{\theta})$$

$$\Longrightarrow\ I(\hat{\theta})(\theta-\hat{\theta}) \xrightarrow{d} N(0, I(\hat{\theta}))$$

Likelihood Ratio Test

- $H_0 : \theta = \theta_0$
- 统计量: (理论来自于 $l(\theta)$ 的 Taylor 二阶展开)

$$-2(l(\theta_0)-l(\hat{\theta})) \xrightarrow{d} \chi^2(p)$$

- $l(\theta)$ 的 Taylor 二阶展开:

$$\begin{aligned} l(\theta) &= l(\hat{\theta}) + U(\hat{\theta})(\theta-\hat{\theta}) - \frac{1}{2}I(\hat{\theta})(\theta-\hat{\theta})^2 \\ &\Longrightarrow\ U(\hat{\theta}) = 0 \quad [MLE要求] \Longrightarrow \\ &\Longrightarrow\ (\theta-\hat{\theta})^T I(\hat{\theta})\,(\theta-\hat{\theta}) \xrightarrow{d} \chi^2(p) \end{aligned}$$

三种检验方法在大样本情况下渐近等价

Proportional Hazard Model <比例风险模型>

Newton − Raphson method<牛顿迭代法>

- 在比例风险模型中，由于指数函数的存在，使我们难以使用最小二乘来进行回归。因此我们将使用极大似然估计的方法对比例风险模型进行求解。

$$\begin{aligned} \text{MLE} \Longrightarrow \arg\max_{\theta} L(\theta) &\Longrightarrow \frac{\partial}{\partial\theta}\ln L(\theta) = 0 \\ &\Longrightarrow U(\theta) = 0 \end{aligned}$$

$$\begin{aligned} \text{进行Taylor展开:} \quad 0 = U(\theta) &\approx U(\hat{\theta}) - I(\hat{\theta})(\theta-\hat{\theta}) \\ \Longrightarrow\ \hat{\theta}^{(i+1)} &:= \hat{\theta^{(i)}} + I^{-1}(\hat{\theta}^{(i)})U(\hat{\theta}^{(i)}) \end{aligned}$$

Exponential Distribution<指数分布>

- pdf: $p(t) = \lambda e^{-\lambda t}, \quad \text{for } t \geq 0$

Weibull Distribution<韦布尔分布>

- pdf: $p(t) = \gamma\lambda(\lambda t)^{\gamma-1}e^{-(\lambda t)^{\gamma}}, \quad \text{for } t \geq 0$

$Proportional\ HazardModel$<指数回归模型>

- 形式：

$$\lambda_i(t) = \lambda(t)\exp\{x_i^T\beta\}, \quad \lambda(t) \geq 0$$

- 其中，$\lambda(t)$ 是 t 时刻的基线风险，即 **无协变量影响时的风险**

$Exponential\ Regression\ Model$<指数回归模型>

- 形式：

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

- 相关信息

$$\begin{aligned}
\lambda_i(t) &= \lambda\exp\{x_i^T\beta\} \\
\Lambda_i(t) &= \lambda t\exp\{x_i^T\beta\} \\
S_i(t) &= \exp\Big\{-\lambda t\exp\{x_i^T\beta\}\Big\} \\
f_i(t) &= \lambda\exp\{x_i^T\beta\}\exp\Big\{-\lambda t\exp\{x_i^T\beta\}\Big\} \\
T_i &\backsim Exp(\lambda\exp\{x_i^T\beta\})
\end{aligned}$$

- NR 迭代公式

$$\begin{aligned}
\hat{\beta}^{(i+1)} &= \hat{\beta}^{(i)} + \Big(X^TWX\Big)X^T(d-\mu) \\
\mu &= \{t_i\cdot\exp\{x_i^T\beta\}\}_{i=1}^n \quad \text{加权的事件发生时间} \\
W &= diag(\mu) \\
d &= \delta_{ii=1}^n \quad \text{样本删失情况}
\end{aligned}$$

- 估计置信区间

$$\because \hat{\beta} \backsim N(\beta, I^{-1}(\beta))$$

$$\therefore \text{置信区间为：}\quad \hat{\beta}_j \pm I_{jj}^{-1}$$

- 参数解释：

$$\begin{aligned}\frac{\lambda_1}{\lambda_2} &= \frac{\lambda(t)\exp\{x_1^T\beta\}}{\lambda(t)\exp\{x_2^T\beta\}} = \exp\{(x_1-x_2)^T\beta\} \\ &= \prod_{i=1}^{p}\exp\{\Delta x_i^T\beta\}\end{aligned}$$

第i个特征变动Δx_i $\Longrightarrow$ 风险函数变化$\exp\{\Delta x_i^T\beta\}$倍

Weibull Regression Model<韦布尔回归模型>

- 形式：

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

- 相关信息

$$\begin{aligned}\lambda_i(t) &= \lambda\gamma(\lambda t)^{\gamma-1}\exp\{x_i^T\beta\} \\ \Lambda_i(t) &= (\lambda t)^{\gamma}\exp\{x_i^T\beta\} \\ S_i(t) &= \exp\Big\{-(\lambda t)^{\gamma}\exp\{x_i^T\beta\}\Big\} \\ f_i(t) &= \lambda\gamma(\lambda t)^{\gamma-1}\exp\{x_i^T\beta\}\exp\Big\{-(\lambda t)^{\gamma}\exp\{x_i^T\beta\}\Big\} \\ T_i &\backsim Weibull\Big(\lambda\exp\{\frac{x_i^T\beta}{\gamma}\},\quad \gamma\Big)\end{aligned}$$

Accelerated Failure Time Model <AFT模型>

- AFT 假设:
  AFT模型直接对生存时间T进行建模。但时间为非负数据
  无法用线性回归直接建模。所以要对时间进行对数变换
  $Y = \log T$

  AFT Model: $Y = \log T = X^T\beta + \varepsilon$

  $$T = e^{X^T\beta+\varepsilon} = e^{X^T\beta}e^{\varepsilon} \underset{e^{\varepsilon}=T_0}{\Longrightarrow} e^{X^T\beta}T_0$$

  其中T_0是基线生存时间。显然 T_0 是由 ε 决定的，而 ε
  所服从的不同分布也决定着AFT模型的类型。

Extreme Value Distribution<极值分布>

- *Gumbel*分布: (极值I型分布 or SEVD)
  - 此处给出的是极小值格式

$$F(t) = 1 - e^{-e^t}$$
$$f(t) = e^{t-e^t}$$

Type of AFT Model<AFT模型类型>

- $\varepsilon \backsim Gumbel$
  - $y = \log t = x^T\beta + \varepsilon$　指数回归

$$\varepsilon = \log t - x^T\beta$$

$$\begin{aligned} F(t) &= 1 - \exp\Big\{ - \exp\{\log t - x^T\beta\}\Big\} \\ &= 1 - \exp\Big\{ - t \cdot \exp\{-x^T\beta\}\Big\} \\ &\Longrightarrow \ \lambda = exp\{-x^T\beta\} \\ &\Longrightarrow \ y \backsim Exp(\lambda) \end{aligned}$$

- $y = \log t = -\log\lambda - \frac{1}{\gamma}x^T\beta + \frac{1}{\gamma}\varepsilon$ 韦布尔回归

$$\varepsilon = \gamma\log t + \gamma\log\lambda + x^T\beta$$

$$\begin{aligned} F(t) &= 1 - \exp\Big\{ - \exp\{\gamma\log t + \gamma\log\lambda\}\Big\} \\ &= 1 - \exp\Big\{ - (\lambda t)^\gamma \exp\{x^T\beta\}\Big\} \\ &\Longrightarrow \ \tilde{\lambda} = \lambda\exp\{\frac{-x^T\beta}{\gamma}\} \\ &\Longrightarrow \ y \backsim Weibull(\tilde{\lambda}, \gamma) \end{aligned}$$

- $\varepsilon \backsim N(0, \sigma^2)$
  Lognormal-AFT Model

Semi Parametric AFT Model<半参数AFT模型>

- 保留线性部分的可解释性，放松残差(基线风险)的假设
  - Rank Statistics (秩统计量)

$$\begin{aligned} &if: y_{(i)}\text{为排序后第i大的因变量; } x_{(i)}\text{为其对应的自变量;} \\ &rank = \sum_{i=1}^{n}(i - \bar{i})(x_{(i)} - \bar{x}) \end{aligned}$$

秩统计量展现了x与y之间的相关关系(此时y的具体值不重要，重要的只有相对位置)

  - Rank Regression
    1 检测的是个体死亡顺序是否的确与解释变量x无关

(使用t_j排序)

$$H_0 : \beta = 0$$
$$\frac{U^2}{V} \backsim \chi_1^2$$
$$U = \sum_j (x_{(j)} - \bar{x}_{(j)})$$
$$V = \sum_j (x_{(j)} - \bar{x}_{(j)})^2$$

$x_{(j)}$代表了时间点t_j时死亡的个体，
$\bar{x}_{(j)}$代表了时间点t_j时仍然存活的个体。

2 检查残差中是否还含有解释变量的信息
(使用$W_j = Y_j - x_j^T \beta = \log(t_i) - x_j^T \beta$排序)

$$H_0 : \beta = \beta_0$$
$$\frac{U^2}{V} \backsim \chi_1^2$$

Cox Regression Model <Cox回归>

Partial Likelihood<偏似然>

- 偏似然的出现是为了解决在不知道基线风险分布的情况下计算似然函数
- 偏似然本质上与似然定义并不同：

$$\begin{cases} \textit{Likelihood}: \text{在已知变量分布族的情况下，} \\ \qquad\quad \text{求解在已知样本条件下出现概率最大的参数。} \\ \textit{Partial Likelihood}: \text{在未知变量分布族的情况下，} \\ \qquad\quad \text{通过归一化变量构造概率进行累乘。} \end{cases}$$

$$\begin{cases} \textit{Likelihood}: L(\theta) = \prod_i p(x_i;\theta) \\ \textit{Partial Likelihood}: L(\theta) = \prod_i q(x_i;\theta) = \prod_i \frac{x_i}{\sum_k x_k} \end{cases}$$

- Cox回归的偏似然：

$$\begin{cases} L(\beta) = \dfrac{\exp\{x_i^T\beta\}}{\sum_k \exp\{x_k^T\beta_i^T\beta\}} & \text{不考虑删失} \\ L(\beta) = \dfrac{\exp\{x_i^T\beta\}}{\sum_{k\in R(t_i)} \exp\{x_k^T\beta\}} & \text{考虑删失} \end{cases}$$

- Cox回归的$NR-method$：

$$\begin{aligned} \hat{\beta}^{(r+1)} &:= \hat{\beta}^{(r+1)}(X^TWX)^{-1}X^T(d-Pd) \\ w_i &= \exp\{x_i^T\beta\} \\ \pi_{ij} &= \frac{w_i}{\sum_{k\in R(t_j)}^n w_k} = \frac{Y_i(t_j)w_i}{\sum_{k=1}^n Y_k(t_j)w_k} \\ P &= \{\pi_{ij}\} \\ W_{kk} &= -\sum_i \delta_i\pi_{ki}(1-\pi_{ki}) \\ W_{kj} &= -\sum_i \delta_i\pi_{ki}\pi_{ji} \end{aligned}$$

各参数意义：

$Y_i(t_j)$：是一个个体i在某时间t_j生存的示性函数

P：代表某个个体i在时间t_j时的相对死亡风险
因此我们只考虑还活着的个体

W_{kk}：代表个体的加权情况。每当一个其余个体i死亡，
如果k仍然存活(Y决定)，则其样本权重增加。

W_{kj}：代表个体k,j的交互情况。每当一个其余个体i死亡，
如果k,j仍然存活(Y决定)，其交互权重增加

- 相关检验：
  Wald:

  理论依据: $\hat{\beta} \xrightarrow{d} N(\beta, (X^T W X)^{-1})$

  LRT:

  理论依据: $-2(l(\hat{\beta}_1) - l(\hat{\beta}_2)) \xrightarrow{d} \chi^2(\dim_{\hat{\beta}_1} - \dim_{\hat{\beta}_2})$

  Score-test:

  $U(\beta_0)^T I^{-1}(\beta_0) U(\beta_0) \xrightarrow{d} \chi^2(p)$

  当 X 为分组变量时Score-test=Log-Rank test

- 拟合基线风险

  Lemma1: $S_i(t) = S_0(t)^{\exp\{x_i^T \beta\}}$

  $\because \lambda(t) = -\frac{d}{dt} \ln S(t) \quad \therefore \begin{cases} -\int \lambda_0(t)dt = \ln S_0(t) \\ -\int \lambda_i(t)dt = \ln S_i(t) \end{cases}$

  $\because \lambda_i(t) = \lambda_0(t) e^{x_i^T \beta}$

  $\therefore \ln S_i(t) = -\int \lambda_i(t) e^{x_i^T \beta} dt = e^{x_i^T \beta}(-\int \lambda_0(t)dt)$

  $\therefore S_i(t) = \exp\left\{-\int \lambda_0(t)dt\right\}^{e^{x_i^T \beta}} = S_0(t)^{\exp\{x_i^T \beta\}}$ □

  Lemma2: $\lambda_{ij} = 1 - (1 - \lambda_{0j})^{\exp\{x_i^T \beta\}}$

  ∵ 离散假设下: $S(t) = \prod_{t \le t_j} (1 - \lambda_j)$

  ∵ 由 Lemma1 可知: $S_i(t) = S_0(t)^{\exp\{x_i^T \beta\}}$

  $\therefore 1 - \lambda_{ij} = (1 - \lambda_{0j})^{\exp\{x_i^T \beta\}}$

  $\therefore \lambda_{ij} = 1 - (1 - \lambda_{0j})^{\exp\{x_i^T \beta\}}$ □

Theory: $S_i(t) = \prod_{t_i \le t} \hat{\alpha}_j^{\exp\{x_i^T \beta\}}$

$$\because L(\beta) = \prod_j \Big\{ \prod_{i \in D_j} \lambda_{ij} \prod_{i \in R_j - D_j} (1 - \lambda_{ij}) \Big\}$$

$$\xrightarrow{\alpha_j = 1 - \lambda_{0j} \ \& \ Lemma2} L(\beta) \prod_j \Big\{ \prod_{i \in D_j} 1 - \alpha_j^{\exp\{x_i^T \beta\}} \prod_{i \in R_j - D_j} \alpha_j^{\exp\{x_i^T \beta\}} \Big\}$$

$$\Longrightarrow \ \hat{\alpha}_j = (1 - \pi_{jj})^{\frac{1}{\exp\{x_i^T \beta\}}} \text{ (无打结情况下)}$$

$$\Longrightarrow \ S_i(t) = \prod_{t_i \le t} \hat{\alpha}_j^{\exp\{x_i^T \beta\}} \quad \square$$

- Tie Problem 打结问题
  - 若在一个时间点内有多个个体死亡，则称为打结问题。
  - 处理方法: $\begin{cases} 1 \text{平均偏似然} \\ 2 Efron \\ 3 Breslow \end{cases}$
  - 若在一个时间点内有死亡个体过多，则考虑使用离散Cox回归

$$\frac{\lambda_{ij}}{1 - \lambda_{ij}} = \frac{\lambda_{0j}}{1 - \lambda_{0j}} \exp\{x_i^T \beta\}$$