Анализ выживаемости

1 Основные понятия

$Z \geq 0$ — время жизни чего-либо (например, время безаварийной работы какой-нибудь машины), случайная величина, распределение которой нам нужно оценить.

$S(t) = \mathbf{P}(Z > t)$ — *функция выживания*. Понятно, что $S(t) = 1 - F(t)$, где $F(t) = \mathbf{P}(Z \leq t)$ — функция распределения Z.

Будем предполагать, что у величины Z есть плотность f, соответственно

$$f(t) = -\frac{dS(t)}{dt}.$$

Функция риска (функция интенсивности отказов, hazard function)

$$\lambda(t) = \frac{f(t)}{S(t)} = \lim_{\Delta t \to 0} \frac{\mathbf{P}(t \leq Z < t + \Delta t \mid Z \geq t)}{\Delta t},$$

если $S(t) > 0$; и $\lambda(t) = 0$, если $S(t) = 0$.

Утверждение. Если $S(t) > 0$ для некоторого $t \geq 0$, то

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

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

1.1 Цензурированные наблюдения

Пусть $Z_1, ...Z_n$ — (истинные) времена жизни некоторых объектов, ненаблюдаемые неотрицательные одинаково распределенные случайные величины. Пусть также заданы «времена слежения за объектами» $C_1, ..., C_n$, ненаблюдаемые неотрицательные одинаково распределенные случайные величины. Предполагаем, что случайные величины $Z_1, ..., Z_n,\ C_1, ..., C_n$ независимы. Мы наблюдаем следующие величины

$$T_j = \min\{Z_j,\, C_j\}, \quad \Delta_j = I(T_j \leq C_j),$$

$j = 1, ..., n$. Если $C_j < T_j$, т.е. $\Delta_j = 0$, то j-е наблюдение *цензурировано*. Нам нужно оценить распределение случайных величин Z_j по наблюдениям $(T_j,\ \Delta_j)$.

Оценка Каплана-Мейера:

$$S^*(t) = \prod_{j:\ t_j \le t} \left(1 - \frac{d_j}{n_j}\right).$$

t_j — это времена, в которые происходят наблюдаемые события (например, смерти или поломки), т.е. точки времени, в которых есть наблюдения $(T_j = t_j, 1)$. d_j — это количество наблюдаемых событий в момент времени t_j:

$$d_j = \#\{j :\ T_j = t_j, \Delta_j = 1\}.$$

n_j — количество объектов, за которыми мы еще следим в момент времени t_j:

$$n_j = \#\{j :\ T_j \ge t_j\}.$$

Также для *кумулятивной функции риска*

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

известна *оценка Нельсона-Аалена*

$$\Lambda^{**}(t) = \sum_{j:\ t_j \le t} \frac{d_j}{n_j}.$$

Соответственно можно построить оценку $S^{**}(t) = \exp\left(-\Lambda^{**}(t)\right)$. Но далее мы эти оценки рассматривать не будем.

KM-оценка есть оценка максимального правдоподобия (ОМП), если рассматривать функцию правдоподобия следующего вида. Для функции выживания вида

$$S(t) = \prod_{j:\ t_j \le t} (1 - h_j)$$

положим

$$L(t) = \prod_{j:\ t_j \le t} h_j^{d_j} (1 - h_j)^{n_j - d_j}.$$

(На самом деле такая функция правдоподобия имеет смысл только когда $Z_1, ..., Z_n,\ C_1, ..., C_n$ независимы.) Далее

$$\ln L(t) = \sum_{j:\ t_j \le t} d_j \ln h_j + (n_j - d_j) \ln(1 - h_j).$$

Максимум этой функции достигается при $h_j^* = d_j/n_j$.

При $n \to \infty$ и фиксированном t дисперсия оценки $S^*(t)$

$$\mathbf{D}S^*(t) \sim (S^*(t))^2 \sum_{j:\ t_j \le t} \frac{d_j}{n_j(n_j - d_j)}.$$

Это формула Гринвуда. С помощью нее можно строить доверительные интервалы для KM-оценок.

1.2 Логранговый тест

Логранговый тест - самый распространенный критерий сравнения наблюдений двух групп.

Пусть d_{1j}, d_{2j} и n_{1j}, n_{2j} — количества событий и количества объектов, за которыми мы еще следим, в момент времени $t_j,\ j = 1, ..., N$. Положим

$$e_{1j} = n_{1j}\frac{d_{1j} + d_{2j}}{n_{1j} + n_{2j}}, \quad e_{2j} = n_{2j}\frac{d_{1j} + d_{2j}}{n_{1j} + n_{2j}}.$$

Это ожидаемые количества событий в первой и второй группе в момент времени t_j. Разность количеств наблюдаемых и ожидаемых событий для группы i:

$$D_i - E_i = \sum_j (d_{ij} - e_{ij}).$$

Статистика логрангового теста

$$\frac{(D_i - E_i)^2}{V_i},$$

где V_i – оценка для дисперсии $D_i - E_i$

$$V_i = \sum_j \frac{n_{1j}n_{2j}(d_{1j} + d_{2j})(n_{1j} + n_{2j} - d_{1j} - d_{2j})}{(n_{1j} + n_{2j})^2(n_{1j} + n_{2j} - 1)}$$

не зависит от i. Это оценка дисперсии гипергеометрического распределения.

Если количества наблюдений в каждой группе стремятся к бесконечности, распределение статистики $(D_i - E_i)^2/V_i$ слабо сходится к распределению χ^2 с одной степенью свободы (распределение квадрата стандартной нормальной случайной величины).

1.3 Модель Кокса пропорциональных рисков

Пусть $\boldsymbol{X} = (X_1, ..., X_p)$ набор некоторых параметров. Регрессионная модель следующего вида

$$\lambda(t|\boldsymbol{X}) = \lambda_0(t)\exp(\boldsymbol{X}\boldsymbol{\beta}),$$

где $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ — вектор коэффициентов, $\lambda_0(t)$ — базовая функция риска.

Если у нас есть наблюдения $(T_j, \Delta_j, \boldsymbol{X}_j)$, то функция правдоподобия имеет вид

$$L(\boldsymbol{\beta}) = \prod_{j:\ \Delta_j=1} \frac{\lambda(T_j|\boldsymbol{X}_j)}{\sum_{k:\ T_k \geq T_j} \lambda(T_j|\boldsymbol{X}_k)} = \prod_{j:\ \Delta_j=1} \frac{\lambda_0(T_j)\exp(\boldsymbol{X}_j\boldsymbol{\beta})}{\sum_{k:\ T_k \geq T_j} \lambda_0(T_j)\exp(\boldsymbol{X}_k\boldsymbol{\beta})}$$

$$= \prod_{j:\ \Delta_j=1} \frac{\exp(\boldsymbol{X}_j\boldsymbol{\beta})}{\sum_{k:\ T_k \geq T_j} \exp(\boldsymbol{X}_k\boldsymbol{\beta})}.$$