通俗阐释 比例风险(Cox) 回归模型

1. 引言

比例风险回归模型，又称Cox回归模型，是由英国 统计学 家D.R.Cox与1972年提出的一种**半参数回归模型**。模型可以用来描述了不随时间变化的多个特征对于在某一时刻死亡率的影响。它是一个在生存分析中的一个重要的模型。

比例风险回归模型是我在学习 广义线性模型 的时候看到的一个例子，出于好奇，就想学习一下它是啥玩意儿。我一直忿忿不平的是，写书写资料的人往往喜欢写一堆数学公式，却把建立模型Motivation的给去掉了，而 科学的创新往往就是来自这一点点灵感，公式的推导只是其中很机械化、很trivial、但也很重要的一部分工作。

关于Cox回归模型，笔者在学习时感到难以理解的有两点，一是Cox回归模型中比例风险假设的**现实意义和合理性**；二是Cox回归模型的**极大似然估计**有点Tricky，直接看公式会让人难以理解。

我写本文能为大家提供一点对模型的直观、通俗的理解。

2. *Motivation

假如你现在要研究一个人从出生开始，到t时刻时死亡的概率为多大。那么它会受什么影响呢？直观的来看：

- 一方面，它会受时间推移的影响。一个健康的人，随着年纪的增大，他死亡的概率也会不可避免的越来越大；
- 另一方面，它会受一些客观因素影响，比如,一个吸烟的人在某一时刻t死亡的概率，比一个不抽烟的同龄人概率会更大；再比如，一个富豪，每年都花大价钱为自己养生、雇佣营养师为自己控制饮食起居，那么他可能就比我这个穷屌丝死亡的概率更小。

综上所述，我们抽象出了两部分的因素，一部分受**时间**的影响，你可以理解为是理想情况下、不受任何外界影响下的死亡的概率、是一个基准；另一部分受**客观因素**的影响，这些因素会影响整体的概率，使得它在基准上增加或减少。

理解了这一部分，就不难理解模型的基本假设了。

3. 危险率

我前面用的说法“t时刻时死亡的概率”是非常不严谨、难以量化的。这个概念更应该是统计意义上的概率。我们不妨称它为**危险率**

在时刻t的危险率被定义为：$\frac{t时刻将要死掉的人数}{t时刻仍存活的总人数}$，可以理解为“某一时刻危险人群的比例”。举例来讲，假设在t时刻之前，原来有10个人，t时刻后有3人去世了，那么这时的危险率显然应该被定义为$\frac{3}{10}$。

在此我们要建模的，就是这个危险率 与 时间和客观因素 关系。

4. 模型基本假设

- 符号定义

1. X：客观因素，由m个影响因素组成：$X = (X_1, X_2, ..., X_m)$
2. t：时间
3. $h(t, X)$：当时间为t，客观因素为X的时候的危险率

- 假设的内容：

$$h(t, X) = \lambda_0(t)exp(\beta \cdot X)$$

在这里，$\lambda_0(t)$是一个仅与时间有关的函数，其选择具有充分的灵活度，一种可能的选择是采用概率论中的Weibull分布、指数分布等。

β是模型的参数。由于只要给定数据，就能够求出模型的参数β。

对公式进行变形，得到：

$$ln(h(t, X)) = \beta \cdot X + ln(\lambda_0(t))$$

分析这个公式，结论如下，模型中各危险因素与时间相互独立，同时，对数危险率与各个危险因素呈线性相关。这就是Cox回归中的两个基本假设。

4. 引理：与其他函数的联系

1. 生存函数，客观因素为X时，在t时刻仍然存活的概率：$S(t, X) = P(T > t, X)$，（tips：活着，真实寿命T比t长）
2. 死亡函数，客观因素为X时，在t时刻已经死亡的概率：$F(t, X) = P(T \leq t, X)$，（tips：死了，真实寿命T不超过t）
3. 死亡密度函数，客观因素为X时，在t时刻已经死亡的概率密度: $f(t, x) = lim_{\Delta t \to 0} \frac{P(t<T<t+\Delta t, X)}{\Delta t} = F(t, X)'$

> 推导
>
> $$h(t, X) = lim_{\Delta t \to 0} \frac{P(t < T < t + \Delta t | T > t, X)}{\Delta t}$$
>
> $$= \frac{1}{P(T > t, X)} lim_{\Delta t \to 0} \frac{P(t < T < t + \Delta t, X)}{\Delta t} = \frac{f(t, X)}{S(t, X)}$$

5. 参数的极大似然估计

（tips：建议结合阅读参考文献3、4中这部分的公式推导，其中的内容更加形式化。前面的例子能帮助你理解为什么这里用这样的方法做极大似然估计，它与概率论与数理统计中常用的解法不同）

极大似然估计的思想是，让已经发生的事件出现的可能性最大。那么，在当前的上下文中，时间出现的可能性最大的含义是什么呢?

让我们来举一个例子说明，假如有3个输入样例，分别在时间t=1，3，7死去。我们希望我们的模型预测的结果是，当t=1时，第1个人死了，其它2个人活着，同时第1个人死掉的概率最大；当t=3时，第1，2个人死了，其它1个人活着,当t=5时，第3个人也死了。

如何达到上述目标呢？以t=1时为例，要想达到上述目标，一个可行的方法是，试图让第一个人死去的概率变大，让剩下的2人活着的概率变大（死去的概率变小），即：

$$max\ h(1,X^{(1)}), min\ h(1,X^{(2)})+h(1,X^{(3)})$$

为了将这两个目标统一起来，我们得到：

$$max\ \frac{h(1,X^{(1)})}{h(1,X^{(2)})+h(1,X^{(3)})}$$

以此类推，得到t=3时的目标为：$max\ \frac{h(3,X^{(2)})}{h(3,X^{(3)})}$，当t=5时，遇到问题了，因为没有其它人活着了，第二个目标不存在，分母为0。为了解决这个问题，我们在分母上加上分子这一项用来平滑。

所以最终三个目标为：$max\ \frac{h(1,X^{(1)})}{h(1,X^{(1)})+h(1,X^{(2)})+h(1,X^{(3)})}$,$max\ \frac{h(3,X^{(2)})}{h(3,X^{(2)})+h(3,X^{(3)})}$,
$max\ \frac{h(5,X^{(3)})}{h(5,X^{(3)})}$

似然函数为：

$$L(\beta)=\frac{h(1,X^{(1)})}{h(1,X^{(1)})+h(1,X^{(2)})+h(1,X^{(3)})}\frac{h(3,X^{(2)})}{h(3,X^{(2)})+h(3,X^{(3)})}\frac{h(5,X^{(3)})}{h(5,X^{(3)})}$$

化简消去$\lambda_0(t)$,得到：
$L(\beta)=\frac{exp(\beta\cdot X^{(1)})}{exp(\beta\cdot X^{(1)})+exp(\beta\cdot X^{(2)})+exp(\beta\cdot X^{(3)})}\frac{exp(\beta\cdot X^{(2)})}{exp(\beta\cdot X^{(3)})+exp(\beta\cdot X^{(3)})}\frac{exp(\beta\cdot X^{(3)})}{exp(\beta\cdot X^{(3)})}$

（tips：我们称之为“比例”风险模型，就是因为$\lambda_0(t)$能被消去，参数估计时与它无关）

公式化

以下对上述讨论进一步推广、泛化。设共有N个事件，第i个事件的风险特征为$X^{(i)}$,发生的时间为t_i,由此我们得到极大似然函数为：

$$L(\beta)=\prod_{i=1}^{N}\frac{exp(\beta\cdot X^{(i)})}{\sum_{j:\ t_j\geq t_i}exp(\beta\cdot X^{(j)})}$$

对数似然函数为：

$$l(\beta)=ln\,L(\beta)=\sum_{i=1}^{N}[\beta\cdot X^{(i)}-ln(\sum_{j:\ t_j\geq t_i}exp(\beta\cdot X^{(j)}))]$$

梯度为：

$$\frac{\partial l(\beta)}{\partial\beta}=\sum_{i=1}^{N}[X^{(i)}-\frac{\sum_{j:\ t_j\geq t_i}X^{(j)}\cdot exp(\beta\cdot X^{(j)})}{\sum_{j:\ t_j\geq t_i}exp(\beta\cdot Xj)}]$$

接下来，就可以采用梯度下降法等方法对参数进行估计。

Reference

1. 原论文出处：Cox D R. Regression models and life-tables[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1972, 34(2): 187-202.

2. 公式部分，参考了Wikipedia：Proportional hazards model
3. 基本假设部分，参考了百度百科：COX回归模型
4. 极大似然部分的例子，参考了科学网：关于Cox回归模型你需要知道的数学，邵斌的博文

转载请指明出处。