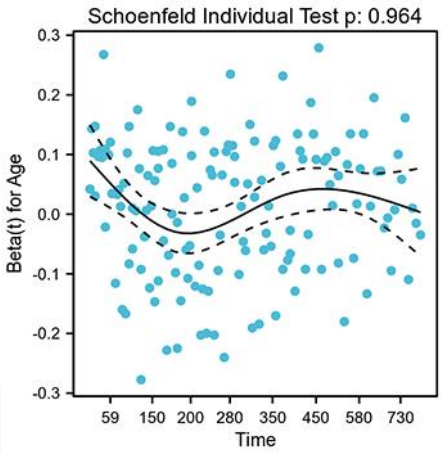
临床意义 - 风险比例图



网址：https://www.xiantao.love

更新时间：2023.01.30

目录

基本概念

- Cox 回归模型：又称为比例风险回归模型，是一种半参数回归模型。Cox 模型以生存结局和生存时间为因变量，分析众多自变量因素对生存期的影响
  - 数据要求
    - 结局建议用数字编码（0/1，1/2），其中最好用 0 代表删失或者未发生事件，1 表发生事件
    - 自变量（协变量）可以是数值或者分类变量。分类变量如果是含有等级的含义，则需要以等级资料纳入，需要设置参考组，其他组和这个参考组作对比；如果分类变量是无等级含义，一般是需要经过哑变量编码，但是经过哑变量编码后结果有可能不好解读，故无等级关系的分类变量也可以通过组合的方式形成二分类变量纳入。二分类的分类变量以等级或者非等级纳入的结果都是一致的（二分类分不分等级都一样）。数值变量可以直接以数值变量的形式纳入，亦可转换为等级资料或者二分类资料纳入
  - 条件假设：观测值独立，风险比不随时间改变（比例风险假设）。（模块内默认是满足此条件）
  - 对于回归模型的假设检验通常采用似然比检验、Wald 检验和记分检验
- PH 假设：比例风险（Proportional hazards）假定。Cox 模型应用的前提条件。基本假设为：协变量对生存率的影响不随时间的改变而改变，即风险比值 h(t)/h0(t)为固定值。而在实际进行生存分析的过程中，有些自变量对风险函数（事件发生概率）的影响会随时间的变化而变化，因此在构建 Cox 回归模型之前，必须对 PH 假定进行判定，只有 PH 假定得到满足时，Cox 回归模型的结果才有意义。

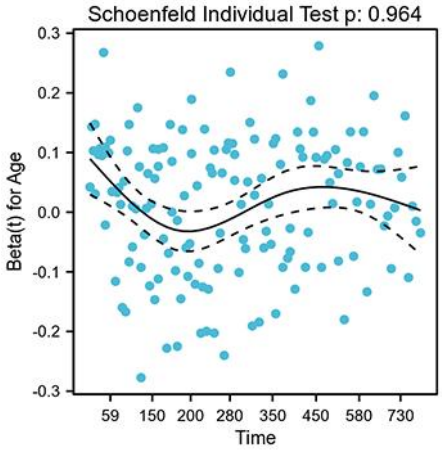
- 中位生存时间（半数生存期）：即当累积生存率为 50%时所对应的生存时间，表示有且只有 50%的生病个体可以活过这个时间。只有当分组内最终累积生存率低于 50%才会有中位生存时间

应用场景

风险比例图：用点和线来展示模型（或变量）在不同预后时间下的系数情况；

作用：验证模型的 cox 回归应用的前提假设中的比例风险假设的内容

结果解读

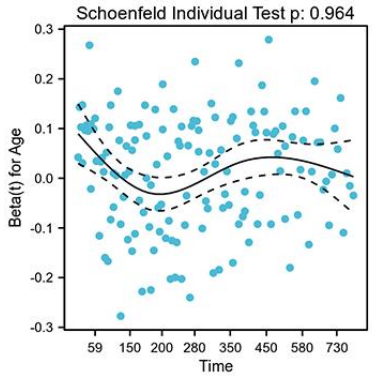


- 横坐标表示预后时间
- 纵坐标表示模型（模型中对应变量）的 cox 回归系数
- 点表示模型在不同预后时间下的系数值
  - 当系数值越聚集（接近于线性水平）且系数不会随时间变化很大，说明模型满足比例风险假设
  - 当系数值越分散且系数随时间变化很大，说明模型不满足比例风险假设
- 实线表示拟合的样条平滑曲线
  - 当曲线越平滑、曲线斜率越接近于 0，且系数不会随时间变化很大，说明模型满足比例风险假设
  - 当曲线偏离 2 个单位的标准差则表示不满足比例风险假设
- 虚线表示拟合曲线上下 2 个单位的标准差（置信区间）

数据格式

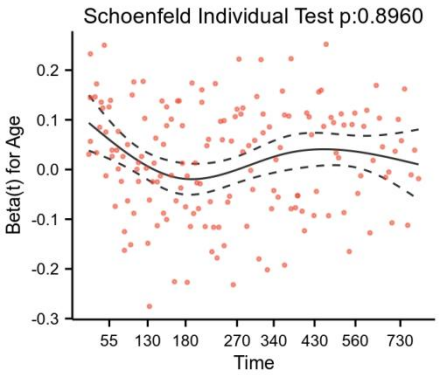
	A	B	C	D	E	F	G	H
1	event	time	Age	Weight los	Sex	Grade	Stage	Score
2	1	306	80		Male	0	Stage2	100
3	1	455	82	15	Male	0	Stage1	90
4	0	1010	42	15	Male	2	Stage1	90
5	1	210	57	11	Male	0	Stage2	60
6	1	883	60	0	Male	2	Stage1	90
7	0	1022	74	0	Male	2	Stage2	80
8	1	310	68	10	Female	0	Stage3	60
9	1	361	71	1	Female	2	Stage3	80
10	1	218	53	16	Male	1	Stage2	80
11	1	166	61	34	Male	0	Stage3	70
12	1	170	57	27	Male	1	Stage2	80
13	1	654	68	23	Female	1	Stage3	70
14	1	728	68	5	Female	0	Stage2	90

	A	B	C
1	Sex	Stage	Grade
2	Male	Stage1	0
3	Female	Stage2	1
4		Stage3	2
5		Stage4	



（第一组：模型含有多个变量，有两个上传数据表格）

	A	B	C
1	event	time	Age
2	1	306	80
3	1	455	82
4	0	1010	42
5	1	210	57
6	1	883	60
7	0	1022	74
8	1	310	68
9	1	361	71
10	1	218	53
11	1	166	61
12	1	170	57
13	1	654	68
14	1	728	68



（第二组：模型有一个（多个）变量，但只有一个表格）

数据要求：

第一组：

左侧表格：分析数据

- 数据至少 3 列、预测变量个数*8 行
- 最多支持 20 列（18 个预测变量）和 1500 行数据
- 第一列是事件发生情况，用 0 和 1 表示，0 表示未发生事件，1 表示发生了事件。例如，事件可以定义为死亡，当受试发生了死亡，该受试的事件就定义为 1，当受试未发生死亡（删失），该受试的事件就定义为 0
- 第二列是具体时间，必须以天作为单位，并且时间要长于 1 年以上
- 第三列及以后为预测的变量，可以是数值类型，也可以是分类类型
  - 如果变量是数值变量，请以数值纳入，只要含有非数值（除空值）外，则此列有可能没有办法纳入到分析
  - 数值变量如果其分类个数<10 个（如 Grade 变量只有 0 1 2）则会按照等级变量来处理
    - 数值变量在相同数值距离下的 HR 差值是一样。比如：假设 Age 年龄，从 40->50 和从 50->60 这两个数值距离都是 10，两个 HR 差值是一样的（风险增长是一样的）
    - 如果某个变量的风险确定是等比增加的，那么可以用数字倍数来进行编码，比如 1 2 4 8
  - 如果变量是等级变量，建议以具体的名字纳入，比如上图中的 Stage，也可以（类似 Grade）以数字 0 1 2 的形式纳入，但是，如果以数字编码的形式纳入并且数量超过 5 个分类以上，需要在 excel 的第二个表中设置等级参考顺序，否则该变量会以数值纳入（等级超过 10 个将没办法纳入）
  - 如果变量是分类变量，默认是以等级资料纳入。二分类变量以等级或者以分类资料或者数值纳入结果都是一样的。如果是多分类非等级资料，则需要以哑变量（暂不考虑）的形式纳入

右侧表格（可以不提供）：

- 对应左侧预测变量（分类类型）中各分类的顺序
  - 比如 Stage 想要设置 Stage1, Stage2, Stage3, Stage4 的顺序，就可以如上图设置。注意，设置了等级顺序后，多因素 Cox 回归的结果都是以第一个作为参考，其他的等级顺序与第一个等级进 行对比。另外，如果在表 1 中的分类变量没有设置等级顺序，则默认以在表 1 中各个分组出现的顺序作为等级顺序。此外，如果是以 0 1 2 编码的等级变量，如果没有在这个表中进行设置，则会以数值类型纳入（可见 Grade 列）
- 如果其取值跟左侧表格预测变量完全一致，则会按照其顺序对左侧对应的变量分类顺序进行分析。比如 Grade 变量在右侧表格中各分类的顺序为 0、1、2，与左侧表格的 Grade 变量中变量名还有具体值完全一样，则会按照右侧表格变量法分类的顺序进行分析，如果不是则按照第一个表格中变量分类的顺序进行分析。

第二组：

- 分析数据的处理（分类、数值变量处理，等级变量的处理会按照变量各分类出现的顺序进行后续分析）与第一组相同，
- 如上两组数据，第一组整个模型中的 Age 变量与第二组中整个 Age 变量作为整个模型，数据相同，所得比例风险假设结果不同。不同模型、不同数据同一个变量比例风险假设（统计显著性(p 值)）也会有所不同

参数说明

（说明：标注了颜色的为常用参数。）

模型

- 变量：可以根据整个比例风险模型选择其任意变量进行比例风险可视化。
- 类型：提供 Beta 系数或者 HR 值，影响 y 轴展示

.数据处理

- 缺失值处理：可以选择对数据中缺失值进行处理
  - 默认为单因素后多因素前处理变量缺失
  - 还可以选择单因素前统一处理缺失，则是在进行分析之前对全部的缺失值进行处理

线

- 颜色：可以修改图中拟合曲线（实线与虚线（置信区间））的颜色
- 粗细：可以修改图中拟合曲线的线条粗细
- 不透明度：可以修改图中拟合曲线的不透明度，默认为 1，表示几乎不透明，1 表示完全不透明，0 表示完全透明

点

- 填充色：点的填充色颜色选项。受配色方案全局性修改
- 描边颜色：点的描边色颜色选项。受配色方案全局性修改
- 样式：点的样式类型，可选择默认为圆形，还可以选择正方形、菱形、三角形、倒三角
- 大小：点的大小
- 不透明度：可以修改图中点的不透明度，默认为 1，表示完全不透明，0 表示完全透明

标题文本

- 大标题：大标题文本
- x 轴标题：x 轴标题文本
- y 轴标题：y 轴标题文本

补充：在要换行的中间插入\n。如果需要上标，可以用两个英文输入法下的大括号括住，比如 {{2}}；如果需要下标，可以用两个英文输入法下的中括号括住，比如 [[2]]

风格

- 外框：是否添加外框，默认添加
- 网格：是否添加网格，默认不添加
- 文字大小：控制整体文字大小，默认为 7pt

图片

- 宽度：图片横向长度，单位为 cm
- 高度：图片纵向长度，单位为 cm
- 字体：可以选择图片中文字的字体

结果说明

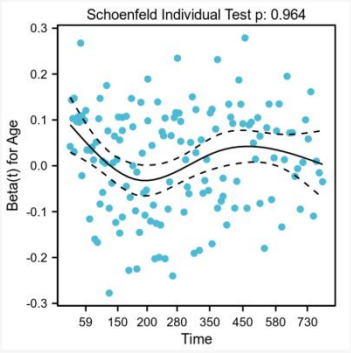
主要结果

主要结果　补充结果　方法学　保存结果　下载整份报告

预后风险比例图

风险比例图: 用点和线来展示比例风险模型(Cox模型)中变量随时间变动下统计量变化情况, 验证Cox回归应用前提假设中的比例风险假设

注意: 不同模型、同一份数据下同一个变量结果也会有所不同



风险比例图.pdf　风险比例图.tiff

· 横坐标表示预后时间, 纵坐标表示模型的Cox回归系数或者HR值

· 点表示模型在不同预后时间下的系数值

· 当系数值越聚集（接近于线性水平）且系数不会随时间变化很大，说明模型满足比例风险假设

补充结果

变量情况

变量情况

各个变量识别出来的类型 以及 是否纳入 进行分析

变量	类型	缺失数量	是否纳入分析	补充说明
event	数值变量	0	纳入	
time	数值变量	0	纳入	
Age	数值变量	0	纳入	
Weight loss	数值变量	14	纳入	
Sex	分类变量	0	纳入	
Grade	分类变量	0	纳入	
Stage	分类变量	1	纳入	
Score	数值变量	3	纳入	

总样本数: 228

· 如果某个分类变量的分类>10，将无法识别为分类变量/等级变量

· 如果变量的分组是以 0 1 2此类进行编码，如果分类数量<5, 会被识别为分类变量；如果>5, 会被识别为数值变量

· 如果数据中含有无穷值，无穷值会被当做缺失处理

补充说明: 单因素分析前，会先去掉 结局和时间列 中的缺失的样本(时间或者结局缺失的样本是无法纳入进行分析的)

缺失处理策略: 单因素后多因素前处理变量缺失

这里提供数据类型统计表:

- 如果是数值类型变量，当变量的个数<5 个时将按照分类类型进行处理
- 如果是数值类型+分类类型，当数值类型变量个数>数据行数的 80%，将按照数值类型进行处理
- 如果是分类类型变量，当变量的个数>10 个时将不做处理
- 如果是分类类型变量，当变量的个数只有一个（单分类类型），将不做处理
- 当变量的异常值过多时（>数据行数 80%）该变量将不作处理

中位生存时间

中位生存时间

中位生存时间只针对分类变量进行，数值变量无法统计中位生存时间

Sex

分组	数目	总事件数	总删失数	总删失比例	中位生存时间	中位生存时间置信区间
Male	138	112	26	18.8%	270	212-310
Female	90	53	37	41.1%	426	348-550

Grade

分组	数目	总事件数	总删失数	总删失比例	中位生存时间	中位生存时间置信区间
0	40	39	1	2.5%	308	153-473
1	92	80	12	13.0%	267	212-363
2	96	46	50	52.1%	340	286-457

- 数值类型将无法进行统计描述（事件情况，中位生存时间等）

单因素 cox 回归分析表

单因素Cox

变量	类型	数量	HR	置信区间	p值
Age	数值变量	228	1.020	1.002 - 1.037	0.025
Weight loss	数值变量	214	1.001	0.989 - 1.013	0.828
Sex	等级变量	228			0.001
Male		138	Reference		
Female		90	0.588	0.424 - 0.816	0.001
Grade	等级变量	228			0.170
0		40	Reference		
1		92	0.950	0.647 - 1.394	0.792
2		96	0.696	0.450 - 1.077	0.104
Stage	等级变量	227			< 0.001
Stage1		63	Reference		
Stage2		113	1.445	0.979 - 2.133	0.064
Stage3		51	2.537	1.637 - 3.931	< 0.001

表中所有变量都会纳入到多因素中

这里提供单因素 cox 回归分析表：

- 如果出现了 NA，说明这个变量分组是参考组(ref)或者是这个变量分组在去除变量信息缺失后数目过少或者只有单分类导致没办法计算
- 每一个变量的缺失情况可以查看第一个表（数据类型统计）
- 分类类型变量的各分组数量是经过缺失值处理之后得到的，不是原始的

补充结果：多因素 cox 回归分析表

多因素Cox

变量	系数β	HR	置信区间	p值
Age	0.013155	1.013	0.995 - 1.032	0.159
Weight loss	-0.012988	0.987	0.974 - 1.001	0.064
Sex				
Male		Reference		
Female	-0.65399	0.520	0.364 - 0.742	< 0.001
Grade				
0		Reference		
1	0.22602	1.254	0.826 - 1.903	0.289
2	-0.19297	0.825	0.514 - 1.322	0.423
Stage				
Stage1		Reference		
Stage2	0.4457	1.562	1.031 - 2.365	0.035
Stage3	0.81334	2.255	1.257 - 4.046	0.006

多因素Cox.xlsx

模型常数/截距(Intercept): 0.34852

原始数据一共有228个, 变量信息缺失的样本有18个, 最终纳入的样本数: 210

备注: 如果出现纳入了多因素但是对应的统计量为空的情况，说明(1)这个变量在去除变量信息缺失后某个分类数目过少(只有1个或者0个)或者是(2)存在严重共线性导致这个变量导致没办法计算。

这里提供多因素 cox 回归分析表：

- 如果出现了 NA，说明这个变量分组是参考组(ref)或者是这个变量分组在去除变量信息缺失后数目过少或者只有单分类导致没办法计算

PH 假设检验表

比例风险假设(PH)

Cox回归应用的前提是要求自变量满足等比例风险假设(P > 0.05)，即自变量的风险不会随着时间改变而改变，若不满足，则不适合用Cox回归进行检验。

这里只对多因素模型以及纳入的变量进行ph假设检验

备注: (1)单个变量直接PH假设和在模型里面这个变量的PH假设的结果是不一样的; (2)同一份数据不同Cox模型中同一个变量的PH假设的结果也是不一样的

变量	统计量(卡方值)	自由度(df)	p值
Age	0.0020385	1	0.9640
Weight loss	0.0082702	1	0.9275
Sex	2.093	1	0.1480
Grade	0.22192	2	0.8950
Stage	2.5408	2	0.2807
Score	4.3366	1	0.0373
GLOBAL	8.1075	8	0.4230

如果全局(GLOBAL)满足p > 0.05，可以认为多因素模型满足比例风险假设

这里提供 PH 假设检验表：全局假设 $p > 0.05$ 整个模型满足比例风险假设

方差膨胀因子

方差膨胀因子(VIF)

方差膨胀因子可用于分析模型中的变量是否存在多重共线性问题

变量	类型	VIF
Age	数值变量	1.0458
Weight loss	数值变量	1.211
Sex	等级变量	
Male		Reference
Female		1.0597
Grade	等级变量	
0		Reference
1		1.6531
2		1.5997
Stage	等级变量	
Stage1		Reference
Stage2		1.6513
Stage3		2.4823

一般认为，当0 < VIF < 10，不存在多重共线性(补充: 也有认为VIF > 4就存在多重共线性); 当10 ≤ VIF < 100，存在较强的多重共线性; 当VIF >= 100或者是出现NaN，多重共线性非常严重

方法学

统计分析和可视化均在 R 4.2.1 版本中进行

涉及的 R 包：survival 包 ggplot2 包

处理过程：

使用 survival 包进行进行 Cox 回归分析，对构建的模型进行比例风险假设检验并用 ggplot2 进行可视化

如何引用

生信工具分析和可视化用的是 R 语言，可以直接写自己用 R 来进行分析和可视化即可，可以无需引用仙桃，如果想要引用仙桃，可以在致谢部分（Acknowledge）致谢仙桃学术（www.xiantao.love）。

方法学部分可以参考对应说明文本中的内容以及一些文献中的描述。

常见问题

1. 如何理解结果（图）？

答：简单来看：

（1）当系数值越聚集（接近于线性水平）且系数不会随时间变化很大，说明模型满足比例风险假设

（2）当曲线越平滑、曲线斜率越接近于 0，且系数不会随时间变化很大，说明模型满足比例风险假设

2. 为什么结果与 survminer 包可视化的结果不大一样?

答：此分析是在 base 包基础上进行可视化的，分析数据一样，只是此可视化结果绘制的是误差线（虚线），而 survminer 包则不是；其可视化结果对应坐标轴的刻度取值不相同（完全不影响）

3．风险比例假设 ph < 0.05 该怎么办？是不是就不能进行 cox 回归分析了？还是说在模块中会进行矫正?

答：主要看整个模型和单个变量模型中的 ph 情况：

① 当整个模型(全局) ph < 0.05 时，所有的变量将不会被纳入进行多因素分析；

② 当单个变量模型 ph < 0.05，但整个模型(全局) ph > 0.05 时，则不会影响到整个模型的分析