临床意义 - 单因素多因素 Cox 回归

Characteristics	Total(N)	HR(95% CI) Univariate analysis	P value Univariate analysis	HR(95% CI) Multivariate analysis
Age	228	1.020 (1.002 - 1.037)	0.025	1.011 (0.994 - 1.029)
Weight loss	214	1.001 (0.989 - 1.013)	0.828	
Sex	228		0.001	
Male	138	Reference		Reference
Female	90	0.588 (0.424 - 0.816)	0.001	0.582 (0.417 - 0.811)
Grade	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	227		< 0.001	
Stage1	63	Reference		Reference
Stage2	113	1.445 (0.979 - 2.133)	0.064	1.478 (0.990 - 2.206)

网址：https://www.xiantao.love

更新时间：2023.02.08

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

Cox 回归模块主要用于评估变量对于预后的影响，或者判断某个变量是否是独立预后因素（多因素中还有统计学意义）。一般在进行多因素 Cox 回归前，会先进行单因素 Cox 回归对每个变量逐个进行分析，将单因素有意义（p<0.1，这个一般不会设置 0.05）纳入到多因素中进行分析

分析流程

上传数据 ⟹ 数据验证 ⟹ 数据处理(清洗) ⟹ 单因素 Cox 分析 ⟹ (单因素分析 p 值是否满足设定的阈值) ⟹ 多因素 Cox 分析

- 数据格式：xlsx 格式
  - 第 1 列数据作为结局变量(事件发生情况)，需要是数值类型数据，用（0 和 1，0 表示未发生事件，1 表示发生了事件）或（1 和 2，1 表示未发生事件，2 表示发生了事件）表示，注：第 1 列(结局变量)不能都是删失

	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

- 第 2 列数据作为时间变量(具体时间/生存时间，必须以天作为单位，并且时间要长于 1 年以上)，需要是数值类型数据，注：第 2 列(时间变量)不能都是同一个时间，并且不能出现小于 0(负数)和非数值的情况

- 第 3 列以及以后为变量(支持单分类/二分类/多分类/数值(数值可以设置分组))；分组顺序可以在第 2 个 sheet 中设置

	A	B	C	D	E	F
1	Sex	Stage	Grade			
2	Male	Stage1	0			
3	Female	Stage2	1			
4		Stage3	2			
5						
6						
7						
8						
9						
10						
11						
12						
13						
14						
15						

data settings +

- 数据验证：
  - 先对数据第 1 列（事件）、第 2 列（时间）先进行基本格式要求验证（比如生存时间有小于 0 的情况、事件都是删失的情况等等）
  - 对上传数据的行数列数进行判断，需要满足分析数据的格式要求
- 数据清洗：

- 将第 1 列（事件）、第 2 列（时间）不满足的数据清理掉
- 再对其它列数据进行处理
  - 如果上传数据有个 sheet 且格式贴合样本数据，可以根据第 2 个 sheet 的数据对第 1 个 sheet 的数据进行相关调整（具体可以看数据格式部分）

➢ 单因素 Cox 回归分析：

- 构建预后 cox 回归模型：将清洗过的数据进行 cox 模型构建
- 通过模型得到模型所有变量的分析结果

➢ 多因素 Cox 回归分析

- 筛选变量：将单因素结果得到的 p 值，没有达到参数“阈值控制”p 值大小的变量筛选出来，不进行后续多因素 Cox 回归
- 多因素 Cox 回归分析

结果解读

Characteristics	Total(N)	HR(95% CI) Univariate analysis	P value Univariate analysis	HR(95% CI) Multivariate analysis
Age	228	1.020 (1.002 - 1.037)	0.025	1.011 (0.994 - 1.029)
Weight loss	214	1.001 (0.989 - 1.013)	0.828	
Sex	228		0.001	
Male	138	Reference		Reference
Female	90	0.588 (0.424 - 0.816)	0.001	0.582 (0.417 - 0.811)
Grade	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	227		< 0.001	
Stage1	63	Reference		Reference
Stage2	113	1.445 (0.979 - 2.133)	0.064	1.478 (0.990 - 2.206)

- Characteristic：变量以及分组
  - 如果变量是等级/分类变量，则紧接其后的变量的分组，其中第一个分组为参考组
  - 如果变量是数值类型，则该变量在表格中只有有一行结果
- Total(N)：数量情况。对应变量总所选分组总的数量以及各组的数量，此样本数量为进行单因素分析时变量和对应分组的数目
  - 由于可能包含缺失信息，所以不同的变量之间的总数可能是不同。（可在参数中选择“进行单因素前先过滤缺失样本”，就能保证变量的总数是一致的）
  - 这一列在文章中并不是一定需要的，可以不提供
- HR(95% CI) Univariate analysis：单因素分析得到的 HR（Hazard ratio，风险比）以及对应的置信区间。一般 HR>1 说明变量是危险因素，HR<1 为保护因素
  - 如果是等级/分类变量的参考组，则此分组 HR 为 Reference

- 如果该行对应的是等级/分类变量的变量名（非具体分组），则不会有 HR 值

➢ P value Univariate analysis：单因素分析得到的自变量对应的 p 值（一般是满足<0.1 就会纳入到模型中）

- 如果是等级/分类变量的参考组，则此分组单因素 p 值为空
- 如果该行对应的是等级/分类变量的变量名（非具体分组），则单因素 p 值为整个变量整体性检验的 p 值，这个 p 值影响该变量是否纳入到多因素。（该 p 值在文章中不需要报告）

➢ HR(95% CI) Multivariate analysis：（只有变量满足进入多因素 Cox 模型的 p 值阈值才会有值）多因素分析得到的 HR（Hazard ratio，风险比）以及对应的置信区间。一般 HR>1 说明变量是危险因素，HR<1 为保护因素

P value Multivariate analysis：（只有变量满足进入多因素 Cox 模型的 p 值阈值才会有值）多因素分析得到的自变量对应的 p 值。当纳入一定的变量时， 多因素仍然有意义则说明该变量可能是独立预后因素。

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

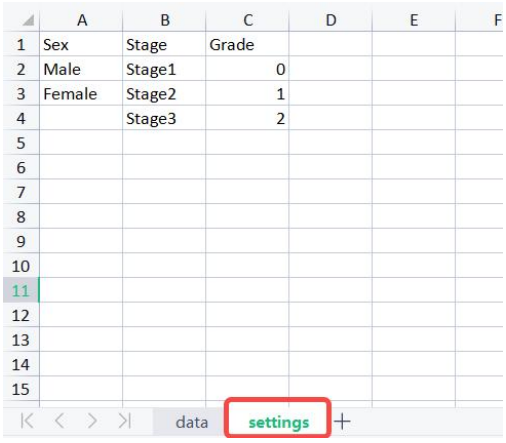
数据要求：表 1-分析数据

- 数据至少 3 列、预测变量个数*4（预测变量个数的 4 倍）行
- 最多支持 22 列（20 个预测变量）和 20000 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
- 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 ＜ 8 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，需要在 excel 的表 2 中设置等级参考顺序，否则该变量会以数值纳入，如果分类变量没有在 excel 表中设置等级参考顺序，则默认 以该列中不同分组出现的顺序作为等级顺序纳入。（等级超过 10 个将没办法纳入）

- 等级变量在不同等级之间的 HR 是不同的，比如结果表格中的 Stage 变量，可以看到 Stage2 和 Stage1 与 Stage4 和 Stage3 之间的 HR 是不同的。尤其要注意不要随意对一个等级资料编码为 0 1 2 3，如果在上传数据进行了此类编码，则这个变量会被认为是数值变量而产生上述数值变量的效果而出现错误。如果是进行了数字编码的等级变量，比如图中 Grade 变量，假设我们设置了 Grade 变量的等级是 0 1 2，可以在表 2 中设定该变量的等级顺序
- 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



数据要求:（表 2-可以不提供）：

- 对应（表 1）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）
  - 如果其取值跟表 1 预测变量完全一致，则会按照其顺序对上方对应的变量分类顺序进行分析。比如 Grade 变量在表 2 中各分类的顺序为 0、1、2，与表 1 的 Grade 变量中变量名还有具体值完全一样，则会按照表 2 变量法分类的顺序进行分析，如果不是则按照表 1 中变量分类的顺序进行分析。

参数说明

（说明：标注了颜色的为常用参数。）

阈值控制

- p 值（单因素进入多因素）：可以控制变量是否进入到多因素 Cox 模型中，常规阈值可选 0.1, 0.2。如果输入的是 1，则代表所有变量都纳入到多因素中

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，表示在经过单因素分析之后，通过变量缺失处理在进行多因素分析

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

结果说明

主要结果

单因素多因素Cox

· **Cox回归**: 比例回归风险模型，用于预后资料的分析。 **条件假设**：观测值独立，风险比不随时间改变（比例风险假设）
· **P值阈值**: 0.1

Characteristics	Total(N)	HR(95% CI) Univariate analysis	P value Univariate analysis	HR(95% CI) Multivariate analysis	P v
Age	228	1.020 (1.002 - 1.037)	0.025	1.011 (0.994 - 1.029)	
Weight loss	214	1.001 (0.989 - 1.013)	0.828		
Sex	228		0.001		
Male	138	Reference		Reference	
Female	90	0.588 (0.424 - 0.816)	0.001	0.582 (0.417 - 0.811)	
Grade	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	227		< 0.001		
Stage1	63	Reference		Reference	
Stage2	113	1.445 (0.979 - 2.133)	0.064	1.478 (0.990 - 2.206)	

Cox回归结果.xlsx　Cox回归结果.docx　Riskscore.xlsx

· Characteristics: 变量以及分组

补充结果：变量情况统计表

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	分类数量	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	2	0	纳入	
Grade	分类变量	3	0	纳入	
Stage	分类变量	3	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

这里提供变量情况统计表：

- 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

补充结果：中位生存时间表

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

Stage

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
Stage1	63	37	26	41.3%	394	348-574
Stage2	113	82	31	27.4%	306	268-429
Stage3	51	45	6	11.8%	183	153-288

备注: 中位生存时间的置信区间如果有?，则代表 分组中样本较少 或者是 随访时间不足 或者是 预后相对较好无法计算出来对应的上限或者下限

这里提供分类变量中位生存时间表：

- 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

补充结果：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	0.0253
Weight loss	数值变量	214	1.001	0.989 - 1.013	0.8282
Sex	等级变量	228			0.0011
Male		138	Reference		
Female		90	0.588	0.424 - 0.816	0.0015
Grade	等级变量	228			0.1705
0		40	Reference		
1		92	0.950	0.647 - 1.394	0.7923
2		96	0.696	0.450 - 1.077	0.1038
Stage	等级变量	227			0.0002
Stage1		63	Reference		
Stage2		113	1.445	0.979 - 2.133	0.0638
Stage3		51	2.537	1.637 - 3.931	3.11e-05

单因素中满足p < 0.1 就会纳入到多因素Cox回归中

这里提供单因素 cox 回归分析表：

- 表中所有变量都会纳入到多因素中

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	0.01114	1.011	0.994 - 1.029	0.2136
Sex				
Male		Reference		
Female	-0.54214	0.582	0.417 - 0.811	0.0014
Stage				
Stage1		Reference		
Stage2	0.39062	1.478	0.990 - 2.206	0.0558
Stage3	0.71291	2.040	1.173 - 3.548	0.0116
Score	-0.0090863	0.991	0.977 - 1.005	0.2011

模型常数/截距(Intercept): 0.031185
原始数据一共有228个, 变量信息缺失的样本有4个, 最终纳入的样本数: 224
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。
备注: 当如果多因素中出现HR异常大或者异常小时，说明这个变量的这个分类数量过少或者是存在共线性问题导致
·(分类/等级)变量(非分组)对应的单因素p值为对应变量单因素模型全局性检验的p值，该变量是否纳入取决于此p值
- 模型全局性统计检验情况：
·· 一致性(Concordance, C-index): 0.653(0.628-0.679)
·· Likelihood ratio test= 31.38 on 5 df, p=7.87e-06
·· Wald test = 30.87 on 5 df, p=9.93e-06

这里提供多因素 cox 回归分析表：

- 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
- 当多因素中出现 HR 异常大或者异常小时，说明这个变量的这个分类数量过少或者是存在共线性问题导致
- (分类/等级)变量(非分组)对应的单因素 p 值为对应变量单因素模型全局性检验的 p 值，该变量是否纳入取决于此 p 值
- 说明文本部分提供了 C 指数（一致性指数）

补充结果：PH 比例风险假设检验表

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	0.036405	1	0.8487
Sex	2.3055	1	0.1289
Stage	3.6901	2	0.1580
Score	5.3024	1	0.0213
GLOBAL	7.6184	5	0.1786

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：

- 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
  - 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
  - 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

补充结果：方差膨胀因子表

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	1.0372
Sex	等级变量	
Male		Reference
Female		1.0053
Stage	等级变量	
Stage1		Reference
Stage2		1.6718
Stage3		2.4937
Score	数值变量	1.6342

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

这里提供方差膨胀因子表：

- 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
  - 当 1<VIF<10，不存在或存在较轻的多重共线性
  - 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[3.4.0]

处理过程：

(1) 使用 survival 包进行比例风险假设检验 并进行 Cox 回归分析

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. 为什么说明文本处只有 计数变量和等级变量？为什么没有分类变量？如何设置分类变量（二分类/多分类变量）？

答：所有的分类变量只有 2 个设置方法：设置哑变量或者设置等级变量。模块是默认 都设置为等级，如果要设置要哑变量，请看“数据格式”部分的说明。 二分类的分类变量，设置为哑变量或者等级变量或者数值变量，结果都是一样的，因为只有 2 个分类

2. 为什么不同变量的数量不同？

Characteristics	Total(N)	HR(95% CI) Univariate analysis
Sex	228	
Male	138	Reference
Female	90	0.588 (0.424 - 0.816)
Grade	228	
0	40	Reference
1	92	0.950 (0.647 - 1.394)
2	96	0.696 (0.450 - 1.077)
Stage	227	
Stage1	63	Reference

答：结果中的这个数量为进行单因素时的数量（如果是分组（而不是变量），则为对 应分组的数量）。由于可能包含缺失信息，所以不同的变量之间的总数可能是不 同。（可在参数中选择“进行单因素前先过滤缺失样本”，就能保证变量的总数是 一致的）。这一列在文章中并不是一定需要的，可以不提供

3. 为什么有一些变量没有在结果上?

答：可以查看补充结果中第 1 个部分的结果，里面会说明变量纳入情况。多分类变量类别过多（>10）不会进行分析。

4. 为什么有一些变量在多因素中统计学数值为 NA?

答：有可能是这么几种情况：

- 变量存在共线性
- 去除任一变量信息缺失后，某个变量的某个分组变成了 0

5．为什么结果里面的 risk score 值是如何计算得到的，如何应用到外部数据?

主要结果部分下载 riskscore 表：

	A	B	C	D	E	F	G
1	event	time	Age	Sex	Stage	Score	RiskScore
2	1	306	80	Male	Stage2	100	0.3948702
3	1	455	82	Male	Stage1	90	0.1200569
4	0	1010	42	Male	Stage1	90	-0.316818
5	1	210	57	Male	Stage2	60	0.5174909
6	1	883	60	Male	Stage1	90	-0.120224
7	0	1022	74	Male	Stage2	80	0.5162508
8	1	310	68	Female	Stage3	60	0.4046608
9	1	361	71	Female	Stage3	80	0.2505145
10	1	218	53	Male	Stage2	80	0.2868916
11	1	166	61	Male	Stage3	70	0.7707532
12	1	170	57	Male	Stage2	80	0.330579
13	1	654	68	Female	Stage3	70	0.3112049
14	1	728	68	Female	Stage2	90	-0.178738
15	1	71	60	Male		70	
16	1	567	57	Male	Stage2	70	0.424035
17	1	144	67	Male	Stage2	90	0.3463418
18	1	613	70	Male	Stage2	100	0.2856515
19	1	707	63	Male	Stage3	70	0.7925969

riskscore 多因素Cox模型情况 ... +

复制多因素 Cox 模型情况表到第一个表，然后通过下面的公式计算：（比如第一个患者，因为患者 Stage 是 Stage2，计算 riskscore 时直接加 Stage2 对应的系数即可）

event	time	Age	Sex	Stage	Score	RiskScore		多因素模型变量	系数		L
1	306	80	Male	Stage2	100	0.3948702		常数(Intercept)	0.065567		* J10
1	455	82	Male	Stage1	90	0.1200569		Age	0.0109219		
0	1010	42	Male	Stage1	90	-0.316818		Sex=Male	0		
1	210	57	Male	Stage2	60	0.5174909		Sex=Female	-0.536001		
1	883	60	Male	Stage1	90	-0.120224		Stage=Stage1	0		
0	1022	74	Male	Stage2	80	0.5162508		Stage=Stage2	0.390113		
1	310	68	Female	Stage3	60	0.4046608		Stage=Stage3	0.6931437		
1	361	71	Female	Stage3	80	0.2505145		Stage=Stage4	1.830114		
1	218	53	Male	Stage2	80	0.2868916		Score	-0.009346		
1	166	61	Male	Stage3	70	0.7707532					

event	time	Age	Sex	Stage	Score	RiskScore		多因素模型变量	系数		L
1	306	80	Male	Stage2	100	0.3948702		常数(Intercept)	0.065567		0.39487016
1	455	82	Male	Stage1	90	0.1200569		Age	0.0109219		
0	1010	42	Male	Stage1	90	-0.316818		Sex=Male	0		
1	210	57	Male	Stage2	60	0.5174909		Sex=Female	-0.536001	手动计算	
1	883	60	Male	Stage1	90	-0.120224		Stage=Stage1	0		
0	1022	74	Male	Stage2	80	0.5162508		Stage=Stage2	0.390113		
1	310	68	Female	Stage3	60	0.4046608		Stage=Stage3	0.6931437		
1	361	71	Female	Stage3	80	0.2505145		Stage=Stage4	1.830114		
1	218	53	Male	Stage2	80	0.2868916		Score	-0.009346		

可以看到手动计算的 riskscore 值 和 G 列的 riskscore 只有小数位后几位的差别（因为约数的问题，可以忽略不计）

如果是要应用这个模型到外部数据，也是同样的计算方法。

6. 多分类变量如何计算这个变量对应的得分？

	A	B	C	D
1	多因素模型	系数		
2	常数(Interce	0.03118535		
3	Age	0.01114027		
4	SexFemale	-0.5421407		
5	StageStage2	0.390623		
6	StageStage3	0.71291434		
7	Score	-0.0090863		
8				
9				
10				
11				
12				
13				
14				
15				

riskscore 多因素Cox模型情况

由于多因素 Cox 模型是广义线性模型的一种，也是存在有常数项的。如果在文章中是有可能看到没有写这个常数项的情况，这种情况有这么几种可能：

1．在 R 里面(print)输出 cox 模型的表是不带常数项的，只有变量和系数，所以就忽略了这个常数项的情况

2．常数项本身在模型里面也不是关键的，因为是常数，不会对结果有什么影响。

另外补充一点多分类（等级）变量的问题：

可以看到上面的表中的 Stage 是多分类，其中 Stage1 的系数为 0，这个说明这个变量是作为参考变量，Stage2 的系数是 0.390，Stage3 的系数是 0.693，这个就是和 Stage1 比的结果。放到原数据中，如果一个患者是 Stage1，那么这个部分就为 0，如果一个患者是 Stage2，那么这个患者在 Stage 这个部分就记 0.390，以此类推。

7．为什么文章里面的多因素 Cox 模型没有常数?

答：

由于多因素 Cox 模型是广义线性模型的一种，也是存在有常数项的。如果在文章中是有可能看到没有写这个常数项的情况，这种情况有这么几种可能：

1．在 R 里面(print)输出 cox 模型的表是不带常数项的，只有变量和系数，所以就忽略了这个常数项的情况

2．常数项本身在模型里面也不是关键的，因为是常数，不会对结果有什么影响。

8．为什么文章里面的模型是放的多因素 p 值有意义的变量，而工具却给的是多因素纳入的变量?

答:

首先可以明确的是，多因素模型中的自变量就是纳入到多因素模型的所有变量，包括进入多因素模型后 p 值没有意义的变量，肯定不是只要多因素 p 值有意义的变量才是多因素模型的变量。

多因素模型里面正是这些所有变量在模型里面经过变量之间和混杂因素分析后才得到的每个变量的校正后的情况。

如果是提取了多因素 p 值有意义的变量再构建一个新的多因素模型，那么这些变量的系数肯定不是用的之前的那个模型，肯定是来自一个这个新的模型，而且这些在上一个多因素中 p 值有意义的变量在新的多因素模型中未必还都是 p 值有意义的。