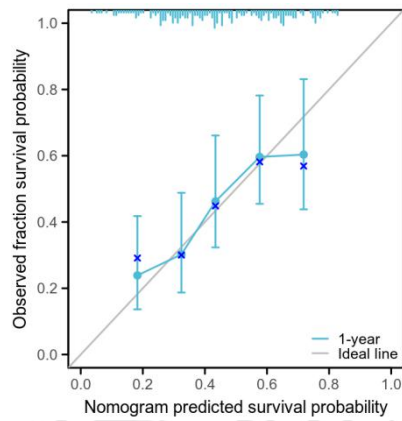


临床意义 - 预后 Calibration 校准曲线



网址: <https://www.xiantao.love>



更新时间: 2023.01.13

目录

基本概念.....	3
应用场景.....	4
结果解读.....	5
数据格式.....	6
参数说明.....	9
预测时间.....	9
数据处理.....	10
分析参数.....	11
线.....	12
点.....	13
图注.....	15
坐标轴.....	16
风格.....	16
图片.....	17
结果说明.....	18
主要结果.....	18
补充结果：变量情况统计表.....	19
补充结果：中位生存时间表.....	20
补充结果：单因素 cox 回归分析表.....	21
补充结果：多因素 cox 回归分析表.....	22
补充结果：PH 比例风险假设检验表.....	23
补充结果：方差膨胀因子表.....	24
方法学.....	25
如何引用.....	26
常见问题.....	27

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，其他组和其他这个参考组作对比；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 $h(t)/h_0(t)$ 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。

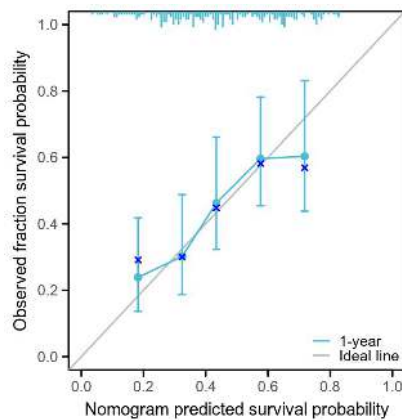
- 中位生存时间（半数生存期）：即当累积生存率为 50% 时所对应的生存时间，表示有且只有 50% 的生病个体可以活过这个时间。只有当分组内最终累积生存率低于 50% 才会有中位生存时间
- Calibration 校准曲线：分为诊断类型和预后类型两类。在 Calibration 图中，通过在图中绘制不同情况下实际概率和模型预测的概率的拟合情况，判断模型对实际结果预测效果的评估。简单来说，Calibration 图只要看线是否能够很好的拟合到对角线上



应用场景

预后 Calibration 图：描绘模型在不同时间点对应的预测概率与实际概率之间的差异情况，主要用于对 Cox 回归方法建立的模型与实际情况的拟合分析。一般在 Nomogram 之后也都会带上一个 Calibration 图来说明模型的准度情况

结果解读



- 横坐标为模型预测的生存概率
- 纵坐标为实际观测到的生存概率
- 图中一共有 2 条线，每条线分别代表：模型预测 1 年（1-year）生存情况与实际情况的对比，以及最理想的线（对角线，灰色）；模型预测的线越贴近对角线说明拟合情况越好（如何简单判断结果好坏：可以就看每根线跟对角线的拟合情况，越贴近对角线说明模型拟合效果越好）
- 校准曲线上的点代表模型预测的生存概率和实际观测到的生存概率情况（类似 Nomogram 中的最下部分中不同得分对应的概率）
- 校准曲线上的点对应的竖线代表该位置的置信区间
- 校准曲线上蓝色的叉代表每个点经过分层 Kaplan-Meier 校正后的结果
- 顶部的竖线代表具体样本对应的生存概率（生存率的分布情况），越密集说明越多样本的生存概率在这个概率

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	10
3	1	455	82	15	Male		0 Stage1	9
4	0	1010	42	15	Male		2 Stage1	9
5	1	210	57	11	Male		0 Stage2	6
6	1	883	60	0	Male		2 Stage1	9
7	0	1022	74	0	Male		2 Stage2	8
8	1	310	68	10	Female		0 Stage3	6
9	1	361	71	1	Female		2 Stage3	8
10	1	218	53	16	Male		1 Stage2	8
11	1	166	61	34	Male		0 Stage3	7
12	1	170	57	27	Male		1 Stage2	8
13	1	654	68	23	Female		1 Stage3	7
14	1	728	68	5	Female		0 Stage2	9
15	1	71	60	32	Male		0	7
16	1	567	57	60	Male		0 Stage2	7

	A	B	C
1	Sex	Stage	Grade
2	Male	Stage1	0
3	Female	Stage2	1
4		Stage3	2
5		Stage4	
6			

数据要求：

上方表格：（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）且最低 20 行
- 最多支持 22 列（20 个预测变量）和 30000 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，如果种类超过 6 个，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（等级超过 8 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

下方表格：（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 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。)

预测时间

预测时间

时间1

1

时间2

请输入数字

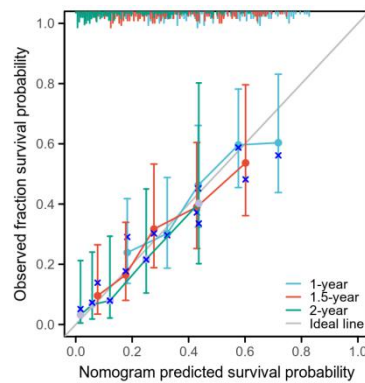
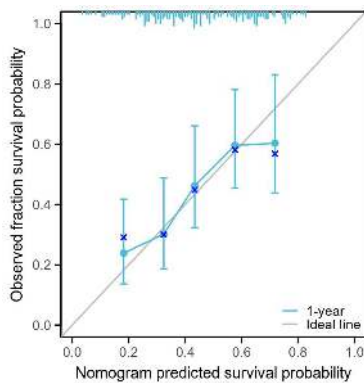
时间3

请输入数字

时间单位

年

- **时间 1**：第一个时间点，数字，单位根据选择的预测时间单位
 - **时间 2**：第二个时间点，数字，单位根据选择的预测时间单位
 - **时间 3**：第三个时间点，数字，单位根据选择的预测时间单位
 - **时间单位**：可以选择上传数据预测时间列的单位，默认以年为单位，可以选择月、天为单位
- 如下所示：左侧为只设置一个预测时间的时候，右侧为设置了多个时间的情况



预测时间

时间1

1

时间2

2

时间3

1.5

时间单位

年

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



分析参数

分析参数 

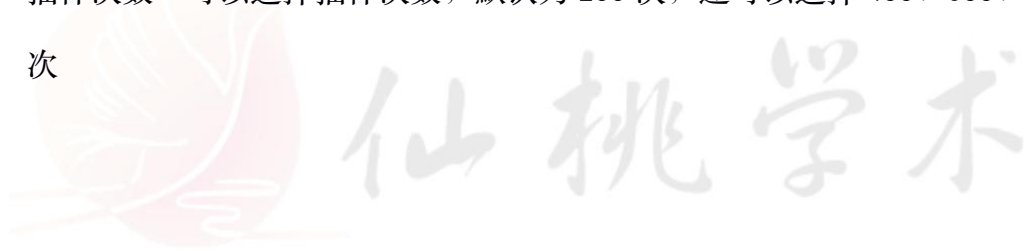
每次重复抽样的
样本量

40

抽样次数

200 

- 每次重复抽样的样本量：可以修改每次重复抽样的样本量
 - 每次抽样的样本量不能超过样本(行数)的一半
 - 每次抽样的样本量不能少于 20 个
- 抽样次数：可以选择抽样次数，默认为 200 次，还可以选择 400、600、800 次



线

线

颜色

线条类型

实线

线条粗细

0.75pt

不透明度

1

- 颜色：可以修改不同预测时间点对应的校准曲线的颜色
- 线条类型：可以选择校准曲线的线条类型，默认为实线，还可以选择虚线类型
- 线条粗细：可以选择校准曲线的线条粗细，默认为 0.75pt
- 不透明度：可以修改校准曲线的不透明度，默认为 1，表示完全不透明

点

点

填充色

描边颜色

样式

圆形

大小

1

不透明度

1

- 填充色：可以修改校准曲线上各点（模型预测的生存概率和实际观测到的生存概率情况）的填充色
- 描边颜色：可以修改校准曲线上各点的描边颜色
- 样式：可以选择校准曲线上各点的形状/样式，默认为圆形，还可以选择正方形、菱形、三角形、倒三角形
- 大小：可以修改校准曲线上各点的大小，默认为 1
- 不透明度：可以修改校准曲线上各点的不透明度，默认为 1，表示完全不透明

标题文本

标题	
大标题	大标题内容
x轴标题	x轴标题内容
y轴标题	y轴标题内容

- 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]]

图注

图注

是否展示

☒

图注标题

图注标题内容

- 是否展示：可以选择是否展示图注信息，默认展示（如左图），也可以不展示（如右图）
- 图注标题：可以修改图注对应的标题内容，如果需要换行可以在需要换行的位置插入“\n”，如下图：

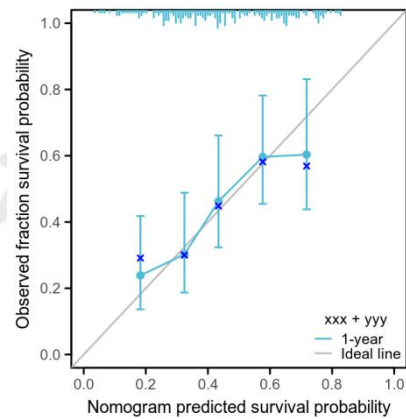
图注

是否展示

☒

图注标题

xxx + yyy



坐标轴



坐标轴

y轴范围 逗号隔开

x轴范围 逗号隔开

- y 轴范围：可以控制 y 轴范围，需要提供 2 个值来控制范围。形如 0.1, 0.3
(最小值不能低于-0.5，最大值不能大于 1.5，如果调整过大可能会无作用)
- x 轴范围：可以控制 x 轴范围，需要提供 2 个值来控制范围。形如 0.1, 0.3
(最小值不能低于-0.2，最大值不能大于 1.2，如果调整过大可能会无作用)

风格



风格

边框 ☒

网格 ☐

文字大小 7pt

- 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

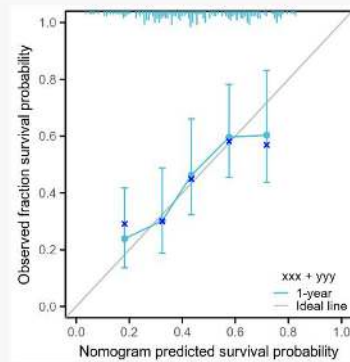


结果说明

主要结果

预后Calibration

预后Calibration: 描绘模型在不同时间点对应的预测概率与实际概率之间的差异情况



预后校准曲线.pdf

预后校准曲线.tif

预后校准曲线.pptx

- 横坐标为模型预测的生存概率；纵坐标为实际观测到的生存概率
- 每条线代表对应时间点的生存情况与实际情况的对比，以及最理想的线（对角线：灰色）；越贴近对角线说明拟合情况越好
- 每条线的点代表模型预测的生存概率和实际观测到的生存概率情况（类似 nomogram 中的最下部分中不同得分对应的概率）
- 每条线的点对应的竖线代表该位置的置信区间

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	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
Weight loss	数值变量	214	1.001	0.989 - 1.013	0.828
Sex	等级变量	228			< 0.01
Male		138	Reference		
Female		90	0.588	0.424 - 0.816	< 0.01
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
Score	数值变量	225	0.980	0.970 - 0.991	< 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：

- 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
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	< 0.05
Stage3	0.81334	2.255	1.257 - 4.046	< 0.01
Score	-0.013034	0.987	0.972 - 1.002	0.087

多因素Cox.xlsx

模型常数/截距(Intercept): 0.34852

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

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	0.0020385	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：

- 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	1.0458
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 $10 \leq VIF < 100$ ，存在较强的多重共线性; 当 $VIF \geq 100$ 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 $10 \leq VIF < 100$ ，存在较强的多重共线性
- 当 $VIF \geq 100$ 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], rms 包 (6.3-0)

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,
- (2) 使用 rms 包进行 Calibration 分析与可视化

补充说明:

- (1) 每次重复抽样的样本量: 40
- (2) 抽样次数: 200



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. 为什么不能改参数提交后直接查看结果?

答：如果每选择一次分析参数进行修改的时候，如预测时间、或数据处理、或分析参数修改，因为分析过程需要花费一定时间（比如当数据量比较大的时候），所以每修改一次都会在后台进行，完成后在“历史记录”查看结果。如果是可视化参数之类的话，可以在历史记录找到记录并选择修改

2. 如何理解结果（图）

答：简单来看，就是看每根线跟对角线的拟合情况，越贴近对角线说明模型拟合效果越好。具体详细结果解读可以看本文档的结果解读部分

3. 为什么所有的变量都进行了单因素分析和多因素分析?

答：一般情况下，是通过对变量进行单因素分析，在对其结果进行筛选，选择单因素变量统计学 p 值大于 0.1（常用）作为筛选条件，满足则对这些变量进行多因素分析，不满足的这分析。但是不能避免有时候上传的数据所有变量都不满足（或条件太过于苛刻）导致无法分析，所以就不进行筛选，直接通过单因素和多因素分析进行计较就行。