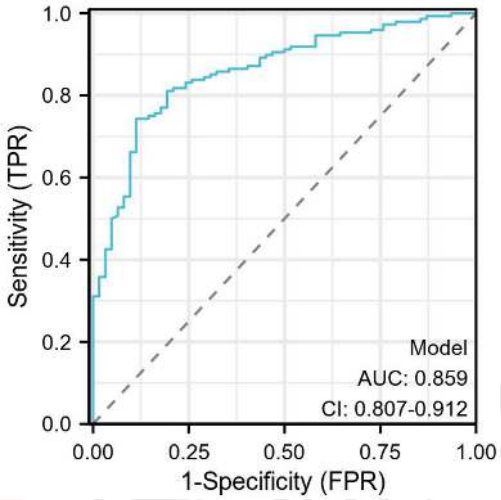
临床意义 - 诊断 ROC 曲线-联合指标



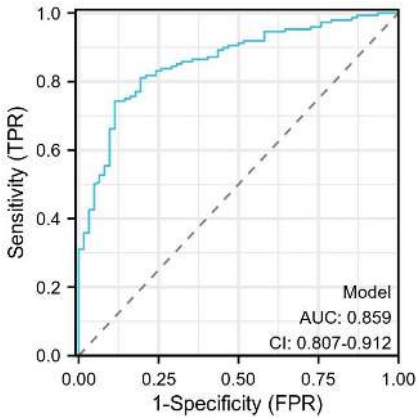
网址：https://www.xiantao.love

更新时间：2023.04.21

目录

基本概念

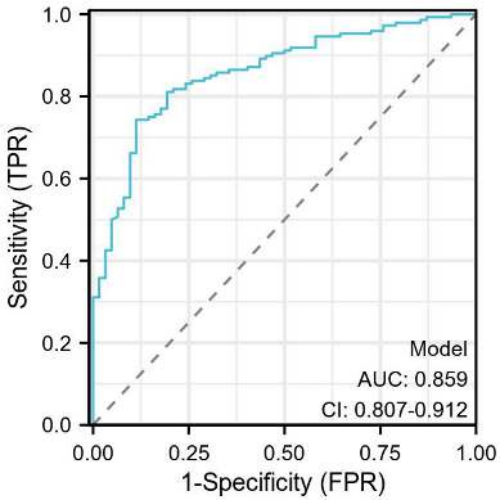
- 诊断 ROC 曲线：受试者工作特征曲线（Receiver Operating Characteristic Curve, ROC 曲线）和 ROC 曲线下的面积（Area Under ROC Curve，AUC）常用于诊断试验的评估，评估预测准确率情况。例如一组数据的结局为 group1 和 group2，变量为 a、b 和 c，也就是评估 a、b 和 c 在预测 group1 和 group2 上的结局，哪个的准确性更高。ROC 曲线图是反映敏感性与特异性之间关系的曲线。AUC 取值范围一般在 0.5 和 1 之间，使用 AUC 值作为评价标准是因为很多时候 ROC 曲线并不能清晰的说明哪个分类器的效果更好，而作为一个数值，对应 AUC 更大的分类器效果更好。
- 诊断 ROC-联合指标：通过数据中多个自变量构建 logistic 回归模型进行 ROC 分析，实现联合指标的效果
- 真阳率（True Positive Rate, TPR）| 敏感度（Sensitivity）：检测出来的真阳性样本数除以所有真实阳性样本数
- 假阳率（False Positive Rate, FPR）| 1-特异度：检测出来的假阳性样本数除以所有真实阴性样本数
- 真阴性率（特异度，Specificity）：检测出来的真阴性样本数除以所有真实阴性样本数
- 图形构成



应用场景

多应用在医学领域，判断某种因素对于某种疾病的诊断是否有诊断价值。

结果解读



诊断 ROC 曲线

- 横坐标 X 轴为 1 －特异性，也称为假阳性率， X 轴越接近零准确率越高
- 纵坐标 Y 轴称为敏感度，也称为真阳性率，Y 轴越大代表准确率越好
- AUC（Area Under Curve，AUC），ROC 曲线下的面积，常用于诊断试验的评估， AUC 取值范围一般在 0.5 和 1 之间， AUC 越接近于 1， 说明该变量在预测结局上诊断效果越好。图中 c 变量（AUC 面积为 0.924）相比于变量 a 和 b 诊断效果较好

数据格式

	A	B	C	D	E	F	G
1	outcome	Age	Weight loss	Sex	Grade	Stage	Score
2	0	42	15	Male	2	Stage1	90
3	1	80		Male	0	Stage2	100
4	1	82	15	Male	0	Stage1	90
5	1	57	11	Male	0	Stage2	60
6	1	60	0	Male	2	Stage1	90
7	0	74	0	Male	2	Stage2	80
8	1	68	10	Female	0	Stage3	60
9	1	71	1	Female	2	Stage3	80
10	1	53	16	Male	1	Stage2	80
11	1	61	34	Male	0	Stage3	70
12	1	57	27	Male	1	Stage2	80
13	1	68	23	Female	1	Stage3	70
14	1	68	5	Female	0	Stage2	90
15	1	60	32	Male	0		70
16	1	57	60	Male	0	Stage2	70

表格 1：变量预测数据

- 第 1 列结局变量（必须是二分类），缺失值不能超过第一列长度的 85%。第 1 列中分类的前后出现的顺序会被参考的顺序，先出现的分类会被当做参考组
- 至少需要 2 列数据，最多不能超过 20 列，最少需要 20 行，最多不能超过 30000 行，样本量需要至少 4 倍以上变量数量，样本过少拟合模型效果相对较差
- 第二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）或者是无穷值外，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
  - 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 5 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 OR 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 OR 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
- ■ 数值变量

➢ 注意：尽量按照示例数据格式整理数据，否则有可能会导致验证数据失败。

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

表格 2：预测变量各分类等级

➢ 对应（表 1）预测变量（分类类型）中各分类的顺序（可以不提供）

- ■ 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 Logistic 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。）

数据处理

- 缺失值处理：默认是单因素后多因素前处理变量缺失，也可以选择单因素分析前统一删除缺失值

统计

- 方向：可以选择自动、正向或者反向

置信区间

- 计算方法：可以选择在分析过程中对相关内容(OR 值)进行置信区间计算的方法，包含有：Wald 方法、profile 方法(MASS 包)、传统计算方法。其 中，Wald 方法得到的置信区间是和 SPSS 是一致的，传统计算方法为（OR±1.96*SE），传统计算方法对应原本生成置信区间的方式。建议选择 Wald 方法。

线

- 颜色：每条曲线的颜色
- 样式：默认是实线，也可以选择虚线
- 粗细：默认是 0.75pt
- 不透明度：默认是 1，0 是完全透明，1 是完全不透明

点

- 展示：是否展示曲线上的点
- 填充色：点的填充色
- 描边色：点的描边色
- 样式：圆形、三角形等形状选择
- 大小：点的大小，默认 0.3
- 不透明度：默认是 1，0 是完全透明，1 是完全不透明

曲线下面积

- 是否展示：是否展示出每个变量曲线下的面积
- 不透明度：如果展示曲线下面积，可以设定面积的不透明度

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

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

- 是否展示：图注内容是否展示
- 图注标题：可以填入图注标题
- 图注位置：默认是右下，还可以选右

风格

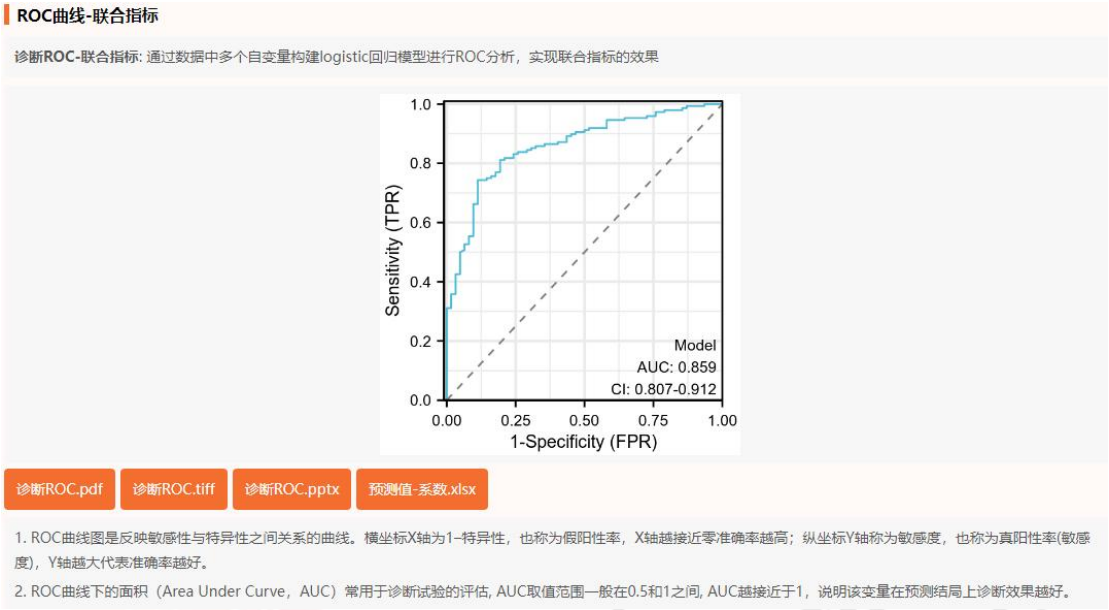
- 边框：是否在图中添加边框
- 网格：是否在图中添加网格线
- 文字大小：图中的文字部分的大小（包括标签文字和刻度数），默认是 7pt

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：可以选择图中文本内容字体

结果说明

主要结果



主要结果格式为图片格式，提供 PDF、TIFF 、PPT 格式下载。同时提供了预测值-系数表格，表格里面对应的 linear_predictor 代表模型的线性预测值，predicted_probability 代表模型预测对应样本的概率值，可视化主要是用这个来代表模型进行联合指标可视化。

A1 fx outcome

	A	B	C	D	E	F	G	H	I	J
1	outcome	Age	Weight loss	Sex	Grade	Stage	Score	linear_predict	predicted_probability	
2	0	42	15	Male	2	Stage1	90	-0.70139188	0.331503702	
3	1	80		Male	0	Stage2	100			
4	1	82	15	Male	0	Stage1	90	3.59685681	0.973321509	
5	1	57	11	Male	0	Stage2	60	4.117700119	0.983978936	
6	1	60	0	Male	2	Stage1	90	-0.12725622	0.468228808	
7	0	74	0	Male	2	Stage2	80	0.77071235	0.683674969	
8	1	68	10	Female	0	Stage3	60	3.842836883	0.979017009	

补充结果

1. 变量情况表

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
outcome	分类变量	2	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 结局列中的缺失的样本(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

2. 单因素 Logistic

单因素Logistic

变量	类型	数量	OR	置信区间	p值
Age	数值变量	228	1.039	1.006 - 1.072	0.0205
Weight loss	数值变量	214	1.006	0.983 - 1.029	0.6088
Sex	等级变量	228			
Male		138	Reference		
Female		90	0.333	0.183 - 0.605	0.0003
Grade	等级变量	228			
0		40	Reference		
1		92	0.171	0.021 - 1.362	0.0953
2		96	0.024	0.003 - 0.179	0.0003
Stage	等级变量	227			
Stage1		63	Reference		
Stage2		113	1.859	0.970 - 3.560	0.0615
Stage3		51	5.270	1.961 - 14.163	0.0010

表中所有变量都会纳入到多因素中

3．多因素 logistic

多因素logistic

模型对应二分类结局(因变量): 1 vs. 0 (其中参考组: 0)

变量	系数β	OR	置信区间	p值
Weight loss	-0.021661	0.979	0.951 - 1.007	0.1386
Sex				
Male		Reference		
Female	-1.3773	0.252	0.116 - 0.550	0.0005
Grade				
0		Reference		
1	-1.4708	0.230	0.027 - 1.925	0.1751
2	-3.7444	0.024	0.003 - 0.191	0.0004
Stage				
Stage1		Reference		
Stage2	0.66601	1.946	0.831 - 4.559	0.1251
Stage3	1.5945	4.926	1.203 - 20.175	0.0266
Score	-0.0038113	0.996	0.967 - 1.026	0.8023

多因素logistic.xlsx

模型常数/截距(Intercept): 3.1294

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

4．AUC 结果表

AUC结果表

预测变量	预测结局	曲线下面积(AUC)	置信区间(CI)
Model	1 vs 0	0.859	0.807 - 0.912

预测结局中, vs后面的结局事件为参考组(影响真/假阳性和真/假阴性的区分)(如果统计-方向参数选择的是“自动”，则会对结局的方向会进行调整保证曲线都是往上凸(pROC包提供))

在AUC > 0.5的情况下，AUC越接近于1，说明该变量在预测结局上诊断效果越好。

AUC在0.5～0.7时有较低准确性, AUC在0.7～0.9时有一定准确性, AUC在0.9以上时有较高准确性。

AUC = 0.5时，说明该变量不起作用，无诊断价值。

5．ROC 信息表

ROC信息表

预测变量	cut-off值	灵敏度	特异度	准确率	真阳个数	真阴个数	假阳个数	假阴个性	阳性预测值	阴性预测值	约
Model	0.90265	0.74324	0.8871	0.78571	110	55	7	38	0.94017	0.5914	0.6

各预测变量在各自最佳cut-off值下部分ROC相关信息和数据。

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：pROC[1.18.0] 用于 ROC 分析和 ROC 检验

1. 数据清洗后，利用 glm 构建多因素 logistic 回归模型，并对模型使用 pROC 包进行 ROC 分析

2. 结果用 ggplot2 进行可视化

3. pROC 包默认会对数据的结局顺序进行校正(保证结果是往上凸)

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. AUC 会出现< 0.5 的情况吗?

答：一般情况下，pROC 分析结果中 AUC 面积是在 0.5-1 之间。

2. 为什么没有给出统计学检验的 p 值?

答:

ROC 一般是看 AUC 的大小的,只有当存在有多个曲线的时候才会进行检验比较。如果只有 1 条曲线，是没办法进行统计检验的，除非是跟 0.05 的对角线比，这种比较其实是没有意义的，这种只要 AUC 的下限没有跨过 0.5，那么这个曲线肯定是有意义的，所以单个曲线是没有统计学比较的意义。

3. 数据的结局是以哪个作为阴性（参考）？哪个作为阳性（实验）？

答:

默认上传数据的第一列（二分类）以第一个出现的分类组参考，后出现的分类作为实验。这个方向会影响最终的真阳、真阴、假阳、假阴个数。如果需要反过来，可以在<统计>-<方向>参数中进行修改。