

基于Logistic回归模型构建疾病风险评分工具

作者：龚志忠 来源：医咖会

本文原地址：<https://www.iikx.com/news/statistics/22663.html>

本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！

基于Logistic回归模型构建疾病风险评分工具。1998年Framingham Heart Study发表了一个经典的冠心病10年风险预测评分工具，该风险函数成为美国国家胆固醇教育计划(NCEP)成人治疗方案(Adult Treatment Panel, ATP)计算冠心病风险的基础。

该模型不再使用原有的连续变量形式，而是将危险因素进行分层，对每一分层进行量化赋分，最后通过计算总分来对患者进行疾病风险评估，评分工具不仅有助于患者理解，也能够临床中得到较好的推广应用。

Step 1			Step 4					Step 8 (determine CHD risk from point total)							
Age			Blood Pressure					CHD Risk							
Years	LDL Pts	Chol Pts	Systolic (mm Hg)	Diastolic (mm Hg)				LDL Pts Total	10 Yr CHD Risk	Chol Pts Total	10 Yr CHD Risk				
30-34	-1	[-1]	<120	<80	80-84	85-89	90-99	≥100	<-3	1%	[-1]	[2%]			
35-39	0	[0]		0 [0]pts	0 [0]pts	1 [1]pts	2 [2]pts	3 [3]pts	-2	2%					
40-44	1	[1]	120-129	0 [0]pts	1 [1]pts	2 [2]pts	3 [3]pts	4 [4]pts	-1	2%	[1]	[3%]			
45-49	2	[2]	130-139	1 [1]pts	2 [2]pts	3 [3]pts	4 [4]pts	5 [5]pts	0	3%	[0]	[3%]			
50-54	3	[3]	140-159	2 [2]pts	3 [3]pts	4 [4]pts	5 [5]pts	6 [6]pts	1	4%	[1]	[3%]			
55-59	4	[4]	≥160	3 [3]pts	4 [4]pts	5 [5]pts	6 [6]pts	7 [7]pts	2	4%	[2]	[4%]			
60-64	5	[5]		4 [4]pts	5 [5]pts	6 [6]pts	7 [7]pts	8 [8]pts	3	6%	[3]	[5%]			
65-69	6	[6]		5 [5]pts	6 [6]pts	7 [7]pts	8 [8]pts	9 [9]pts	4	7%	[4]	[7%]			
70-74	7	[7]		6 [6]pts	7 [7]pts	8 [8]pts	9 [9]pts	10 [10]pts	5	9%	[5]	[8%]			
Step 2			Step 5					Step 6							
LDL-C			Diabetes					Smoker							
(mg/dl)	(mmol/L)	LDL Pts	LDL Pts		Chol Pts			LDL Pts		Chol Pts					
<100	<2.59	-3	No	0	[0]			No	0	[0]					
100-129	2.60-3.36	0	Yes	2	[2]			Yes	2	[2]					
130-159	3.37-4.14	0													
160-190	4.15-4.92	1													
≥190	≥4.92	2													
Cholesterol			Step 7 (sum from Steps 1-6)					Step 9 (compare to average person your age)							
(mg/dl)	(mmol/L)	Chol Pts	Adding up the points					Comparative Risk							
<160	<4.14	[-3]	Age	_____				Age (years)	Average 10 Yr CHD Risk	Average 10 Yr Hard* CHD Risk	Low** 10 Yr CHD Risk				
160-199	4.15-5.17	[0]	LDL-C or Chol	_____				30-34	3%	1%	2%				
200-239	5.18-6.21	[1]	HDL-C	_____				35-39	5%	4%	3%				
240-279	6.22-7.24	[2]	Blood Pressure	_____				40-44	7%	4%	4%				
≥280	≥7.25	[3]	Diabetes	_____				45-49	11%	8%	4%				
Step 3			Smoker	_____				50-54	14%	10%	6%				
HDL-C			Point total	_____				55-59	16%	13%	7%				
(mg/dl)	(mmol/L)	LDL Pts	Chol Pts									60-64	21%	20%	9%
<35	<0.90	2	[2]									65-69	25%	22%	11%
35-44	0.91-1.16	1	[1]									70-74	30%	25%	14%
45-49	1.17-1.29	0	[0]												
50-59	1.30-1.55	0	[0]												
≥60	≥1.56	-1	[-2]												

如果说只是单纯的构建疾病风险预测模型，相信大家应该都比较熟悉，至少大家也都用过Logistic

回归、Cox回归等这些常用的回归模型。但是如何将自己构建的预测模型转化为风险评分工具，似乎成了一个难题。

今天小咖就以多因素Logistic回归模型为例，给大家逐步解密，教会大家也能做出这种高大上的疾病评分工具。

研究实例.

我们仍以Framingham Heart Study为例来进行说明，研究共纳入9443名健康人群，年龄为30-79岁，对研究对象随访5年，观察他们冠心病硬终点的发病情况，包括心肌梗死、死亡。

假定我们要考虑的主要危险因素包括：年龄、性别、收缩压、吸烟，研究人群各个危险因素的分布情况如下表所示，我们在此研究的基础上来制作疾病风险评分工具。

Table III. Summary statistics on risk factors and outcome: Example 5.1.			
Risk factor	Total (n = 9443)	Men (n = 5182)	Women (n = 4261)
Sex: per cent male	55 per cent	—	—
Age, years	51.4 (12.7)	50.6 (12.4)	52.1 (12.9)
Systolic blood pressure, mm Hg	128.9 (19.5)	130.4 (17.8)	127.8 (20.6)
Per cent current smoker	33 per cent	34 per cent	32 per cent
Number of events (per cent) over 5 Years	237 (2.5 per cent)	171 (3.3 per cent)	66 (1.5 per cent)

步骤.

1. 构建多因素Logistic回归模型.

通过构建多因素Logistic回归模型，将我们主要考虑的危险因素纳入到回归模型中，从而估计各个危险因素的回归系数 β ，OR值及其95% CI，回归结果如下表所示。

Risk factor	Regression coefficient	P	OR	95 per cent CI for OR
Intercept	-10.5161	0.0001	—	
Age	0.0575	0.0001	1.059	1.046-1.073
Male sex	1.3078	0.0001	3.698	2.579-4.956
Systolic blood pressure	0.0185	0.0001	1.109	1.012-1.025
Current smoker	0.9456	0.0001	2.574	1.958-3.384

2. 将各个危险因素进行分类，并指定每组的参考值Wij.

我们按照临床意义或使用习惯将各个危险因素进行分组，并在每个分组中选择合适的数值作为参考值Wij，通常选择组内的中间值作为参考值。

例如在本例中，研究人群的年龄范围为30-79岁，通常我们按照10岁一个年龄段来进行划分，将

其分为5组，每组选择中间值为参考值 W_{ij} ，例如30-39岁这一组的参考值 W_{ij} 为 $(30+39) / 2 = 34.5$ 。

本研究中血压的范围为78-240mmHg，我们按照高血压指南的诊断切点来进行划分，每10mmHg为一组，共分为5组，每组选择中间值为参考值 W_{ij} 。但这里需要注意的是，第一组(<120)和最后一组(≥ 160)都是半闭合区间，此时应该如何设置 W_{ij} 呢？

通常我们选择第1百分位和第99百分位的数值来辅助进行计算，在本研究中收缩压第1百分位和第99百分位的数值分别为94mmHg和190mmHg，因此第一组的参考值 W_{ij} 为 $(94+120) / 2=107$ ，最后一组的参考值 W_{ij} 为 $(160+190) / 2 = 175$ 。

对于分类变量，如性别，此时我们可以设置女性为参照，即参考值 W_{ij} 为0，那么男性就自然赋值为1，同理，不吸烟设置为0，吸烟为1。

Risk factor	Categories	Reference value (W_{ij})
Age*	30–39	34.5
	40–49	44.5
	50–59	54.5
	60–69	64.5
	70–79	74.5
Sex	Female	0
	Male	1
Systolic blood pressure**	< 120	107
	120–129	125
	130–139	135
	140–159	150
	≥ 160	175
Current smoker	No	0
	Yes	1

3. 确定各个危险因素的基础分险参考值 $WiREF$.

对于每一个危险因素，我们需要选择一个合适的分组来作为基础风险参考值 $WiREF$ ，在后续构建评分工具时，该组分值将记为0分，危险因素的值高于 $WiREF$ 时记正分，得分越高则风险越高，相反低于 $WiREF$ 时记负分。

在本研究中，我们选择年龄30-39岁、女性、血压120-129mmHg和不吸烟对应的参考值 W_{ij} ，作为每个危险因素的基础风险参考值 $WiREF$ 。

Risk factor	Categories	Reference value (W_{ij})
Age*	30–39	34.5 = W_{1REF}
	40–49	44.5
	50–59	54.5
	60–69	64.5
	70–79	74.5
Sex	Female	0 = W_{2REF}
	Male	1
Systolic blood pressure**	< 120	107
	120–129	125 = W_{3REF}
	130–139	135
	140–159	150
	≥ 160	170
Current smoker	No	0 = W_{4REF}
	Yes	1

4. 计算每一个危险因素的分组与基础风险参考值之间的距离D.

结合多因素Logistic回归模型估计的回归系数 β_i ，以及危险因素各组的参考值 W_{ij} ，来计算危险因素的每一分组与基础风险参考值 W_{iREF} 之间的距离D，计算公式为 $D = (W_{ij} - W_{iREF}) * \beta_i$

例如在本研究中，年龄的基础风险参考值 W_{iREF} 为34.5，年龄在Logistic回归中对应的回归系数 β_i 为0.0575，那么对于70-79岁组，其参考值 W_{ij} 为74.5，该组与基础风险参考值的距离即为 $(74.5 - 34.5) * 0.0575 = 2.3000$ 。

同理，其他危险因素也按照上述公式计算每个分组到基础风险参考值的距离D。

Risk factor	Categories	Reference value (W_{ij})	β_i	$\beta_i(W_{ij} - W_{iREF})$
Age			0.0575	
	30–39	$34.5 = W_{1REF}$		0
	40–49	44.5		0.5750
	50–59	54.5		1.1500
	60–69	64.5		1.7250
	70–79	74.5		2.3000
Sex			1.3078	
	Female	$0 = W_{2REF}$		0
	Male	1		1.3078
Systolic blood pressure			0.0185	
	< 120	107		–0.3330
	120–129	$125 = W_{3REF}$		0
	130–139	135		0.1850
	140–159	150		0.4625
	≥ 160	170		0.8325
Current smoker			0.9456	
	No	$0 = W_{4REF}$		0
	Yes	1		0.9456

5. 设定评分工具中1分对应的常数B.

我们需要设定评分工具中每记1分时，对应的各个危险因素变化的常数。例如本例中，如果设定年龄每增加5岁时记为1分，那么此时常数 $B = 5 * \beta_i = 5 * 0.0575 = 0.2875$ 。

6. 计算危险因素每个分类对应的分值Points_{ij}.

在第5步确定常数B的基础上，计算危险因素每一个分类所对应的分值，计算公式为 $Points_{ij} = D/B = (W_{ij} - W_{iREF}) * \beta_i / B$ ，最后将计算出来的数值四舍五入取整，即为该组对应的分值。

例如本研究中收缩压 ≥ 160 组，其计算的分值为 $0.8325/0.2875=2.8957$ ，四舍五入取整为3分。

Risk factor	Categories	Reference value (W_{ij})	β_i	$\beta_i(W_{ij} - W_{iREF})$	Points _{ij} = $\beta_i(W_{ij} - W_{iREF})/B$
Age			0.0575		
	30–39	34.5 = W_{1REF}		0	0
	40–49	44.5		0.5750	2
	50–59	54.5		1.1500	4
	60–69	64.5		1.7250	6
	70–79	74.5		2.3000	8
Sex			1.3078		
	Female	0 = W_{2REF}		0	0
	Male	1		1.3078	5
Systolic blood pressure			0.0185		
	< 120	107		-0.3330	-1
	120–129	125 = W_{3REF}		0	0
	130–139	135		0.1850	1
	140–159	150		0.4625	2
	≥ 160	170		0.8325	3
Current smoker			0.9456		
	No	0 = W_{4REF}		0	0
	Yes	1		0.9456	3

7. 计算总分与风险预测概率的对应表.

根据第6步的结果，将每个危险因素的分值相加起来计算总分，理论上每个危险因素取最低值时，可以得到总分最低值为 $0+0+(-1)+0=-1$ ，同理可得到总分最高值为 $8+5+3+3=19$ ，因此总分的范围为：-1~19分。

然后再根据多因素logistic回归模型的方程，来计算每一分值对应的风险预测概率值，计算公式如下：

其中 $\sum_{i=0}^p \beta_i X_i \approx \text{常数项} + \beta_i * W_{ij} + B * \text{总分}$

在本研究中， $\sum_{i=0}^p \beta_i X_i \approx -10.5161 + 0.0575 * 34.5 + 1.3078 * 0 + 0.0185 * 125 + 0.9456 * 0 + 0.2875 * \text{总分} = -6.21985 + 0.2875 * \text{总分}$

例如，当总分为 10 分时， $\sum_{i=0}^p \beta_i X_i \approx -6.21985 + 0.2875 * 10 = -3.34485$ ，对应的风险概率值即为 $\hat{p} = \frac{1}{1 + \exp(3.34485)} = 0.0341 = 3.41\%$

以此类推，即可算出总分与风险预测概率的对应表，如下表所示。

Point total	Estimate of risk	Point total	Estimate of risk
-1	0.0015	10	0.0341
0	0.0020	11	0.0449
1	0.0026	12	0.0590
2	0.0035	13	0.0771
3	0.0047	14	0.1002
4	0.0062	15	0.1293
5	0.0083	16	0.1652
6	0.0110	17	0.2088
7	0.0147	18	0.2602
8	0.0195	19	0.3192
9	0.0258		

至此，一个基于多因素Logistic回归模型构建疾病风险预测的评分工具就已经做好了，是不是感觉又get了一项高大上的技能呢!虽然看上去步骤有点多，过程有点复杂，但是如果按照小咖讲的一步一步做下来，相信这么聪明的你一定不会被难倒的。

评分工具.

与Logistic回归模型结果比较.

风险评分工具已经新鲜出炉，为了进一步验证它的准确性，我们举一个实例，来比较一下评分工具与原始Logistic回归模型预测结果之间的差距。

假设有一位男性患者，75岁，收缩压150mmHg，否认吸烟，来预测他未来5年冠心病的发生风险。

首先，我们根据评分工具里各个危险因素的分值，分别记为5、8、2和0分，总分为15分，查表对应的风险概率为12.93%。

然后，我们再根据多因素logistic回归模型进行一次计算：

$$\sum_{i=0}^p \beta_i X_i \approx -10.5161 + 0.0575 * 75 + 1.3078 * 1 + 0.0185 * 150 + 0.9456 * 0 = -2.1208$$

$$\text{风险预测概率为 } \hat{p} = \frac{1}{1 + \exp(2.1208)} = 0.0338 = 10.71\%$$

可以看出，评分工具与Logistic回归模型预测结果之间仅仅只相差2%，足够满足疾病风险预测评估的要求，而且应用起来也很直观和便捷。

本期我们以多因素Logistic回归模型为例对评分工具进行了介绍，先留给大家慢慢消化一下。往往在很多时候，我们在研究中还应考虑事件发生的时间因素，需要用Cox回归来构建疾病预测模型，此时又该如何将其转化为风险评估工具呢？小咖会在后续的内容中来向大家继续进行讲解。

更多统计方法 请访问 <https://www.iikx.com/news/statistics/>

本文版权归原作者所有，请勿用于商业用途，[爱科学iikx.com](https://www.iikx.com)转发