第 3 章

Logistic 回归分析

在实际的临床研究中，有多种类型的变量，如连续型变量或分类变量。响应变量除可能有连续的取值外，还可能会是二分类变量，即只有两种状态，例如某诊断结果是“阳性”或“阴性”，某结局事件是“生存”或“死亡”，某药物治疗效果是“有效”或“无效”等。此外，响应变量还可能会是多分类变量，即有多种状态，此时要综合考虑响应变量与其他变量之间的关系时，多元线性回归已不再适用，应采用 Logistic 回归来解决这类问题。

3.1 Logistic 回归分析的基本概念

在医学研究中，Logistic 回归是分析疾病与致病因子间联系的重要统计方法。它是以疾病发生概率为因变量，影响疾病发生的因子为自变量的一种回归方法。医学研究中的因变量有时并不是呈正态分布的连续型随机变量，其取值可能只有两个，如发病与未发病、阳性与阴性、暴露与未暴露等。此时，线性回归不再适用，而 Logistic 回归模型成功地解决了这一问题。根据响应变量（因变量）的类型，Logistic 回归可分为二分类响应变量的 Logistic 回归和多分类响应变量的 Logistic 回归。

3.2 Logistic 回归的模型结构

在多元线性回归 $\hat{Y}=\beta_0+\beta_1 x_1+\cdots+\beta_k x_k$ 中，Y 可以是任意的数值变量。若 Y 表示的是疾病发生的概率 P，则可以把 P 作为因变量并建立与各自变量 x_i 的回归方程。经过研究，如果把 P 转化为 $\ln\left(\frac{P}{1-P}\right)$，则会使回归方程的统计性能更好。此变换被称为 P 的 logit 转换。即：

$$\text{logit}(P)=\ln\frac{P}{1-P}$$

Logistic 回归的模型结构：设二分类因变量 Y（Y=1 或 Y=0），令 Y=1 的概率为 π，则 Y=0 的概率为 1−π。令 $\ln\frac{\pi}{1-\pi}=\text{logit}(\pi)$，则以 logit(π) 为因变量建立回归方程：

$$\text{logit}(\pi)=\beta_0+\beta_1X_1+\cdots+\beta_pX_p$$

进一步可推导出概率预报模型：

$$\pi=\frac{\exp(\beta_0+\beta_1X_1+\cdots+\beta_pX_p)}{1+\exp(\beta_0+\beta_1X_1+\cdots+\beta_pX_p)}=\frac{1}{1+\exp[-(\beta_0+\beta_1X_1+\cdots+\beta_pX_p)]}$$

实际上，Logistic 回归模型可以看作是多元线性回归模型的推广。

Logistic 回归系数通常采用最大似然估计法，此方法适用范围广，不需要自变量呈多元正态分布，但计算较为繁琐，不过有了计算机，已经变得很方便了。在 Logistic 回归模型中，通常采用 wald χ^2 检验对参数估计值进行假设检验，可检验参数 β_j 是否为 0。目前大多数软件都采用这种检验方法。将第 j 个自变量对应的优势比定义为：$OR_j=\exp(b_j)$。其中，优势比表示当其他自变量保持不变时，该自变量每增加一个单位，所引起的变化量。每个自变量对应的优势比 OR_j 的 95%的可信区间为：$\exp[b_j\pm1.96SE(b_j)]$。当自变量为连续型变量时，OR 值是指自变量每增加一个单位，其优势的变化量。当 OR>1，说明该因素是危险因素；而当 OR<1，说明该因素是保护因素。

3.3 应用实例 1：一般资料的 Logistic 回归

当前一些研究表明糖尿病与促甲状腺激素（TSH）血清水平有一定的潜在关系。加强对糖尿病患者甲状腺功能指标的检测可以早期诊断防治糖尿患者中无症状的甲状腺功能异常。

本例中将 126 名 65 岁以上糖尿病患者分为两组：TSH<4 组编码为 0，表示 TSH 水平正常；而 TSH>4 组编码为 1，表示 TSH 水平异常。将年龄、病程、糖化血红蛋白、空腹血糖、右眼底病变（分为 8 个等级，其中 0 表示正常，1～7 表示病变等级，等级越高表明病变越严重）作为自变量，采用 Logistic 回归分析 TSH 血清水平的影响因素。

这里采用 SPSS 软件进行分析。首先进行数据录入，如图 3.1 所示。

选择菜单 Analyze→Regression→Binary Logistic，如图 3.2 所示。

在 Logistic 回归对话框中，因变量“Dependent”放入“tsh 分组”，自变量“Covariates”放入“年龄、病程、血红蛋白、空腹血糖和右眼病变”。“Method”的主要方法如下。

（1）Enter：所有变量一次全部进入方程；

（2）Forward LR：逐步向前法，自变量根据似然比检验结果依次进入方程；

（3）Backward LR：后退法，自变量根据似然比检验结果依次移出方程。

	年龄	病程	血红蛋白	空腹血糖	右眼病变	tsh分组
1	84	3.00	6.90	9.60	1.00	.00
2	82	10.00	6.10	5.90	7.00	.00
3	82	20.00	10.00	11.00	3.00	.00
4	81	10.00	8.40	7.80	2.00	1.00
5	81	2.00	11.50	10.00	3.00	1.00
6	81	.15	13.50	17.50	2.00	.00
7	81	16.00	7.00	8.30	2.00	.00
8	80	8.00	5.90	6.00	.00	1.00
9	80	8.00	5.90	6.10	.00	1.00
10	80	8.00	5.90	6.50	.00	1.00
11	80	2.00	7.00	6.80	7.00	.00
12	80	20.00	8.30	7.00	.00	.00
13	79	6.00	6.70	9.00	4.00	1.00
14	79	13.00	9.10	7.80	.00	.00
15	79	14.00	10.30	11.90	2.00	.00
16	78	2.00	9.20	10.60	1.00	1.00
17	78	6.00	6.40	6.00	4.00	1.00
18	78	6.00	6.40	6.70	4.00	1.00
19	78	13.00	10.80	9.80	1.00	.00
20	78	20.00	9.20	10.00	1.00	.00

图 3.1　糖尿病与血清水平的数据录入界面（部分数据）

本例采用“Enter”方法来进行分析（图 3.3）。

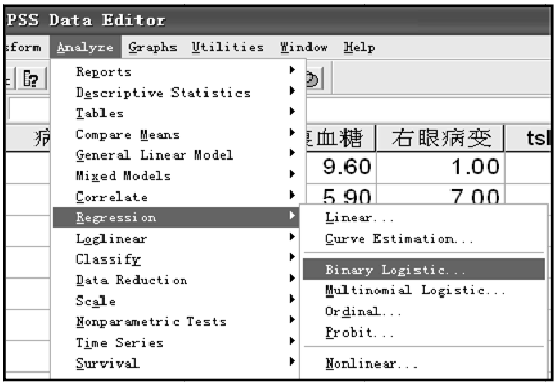


图 3.2　Logistic 回归分析菜单

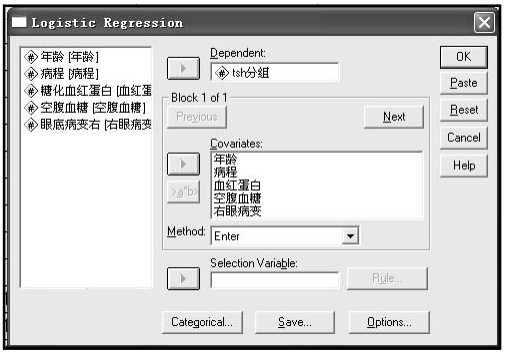


图 3.3　Logistic 回归分析主对话框

单击【Options…】按钮，打开如图 3.4 所示的对话框，勾选“CI for exp（B）”，可以计算出 OR 值的 95%置信区间。

主要输出结果如图 3.5～图 3.8 所示。

如图 3.5 所示的输出表输出的是模型总的全局检验，即似然比卡方检验，3 个结果分别为：Step 统计量为每一步与前一步相比的似然比检验结果；Block 统计量是指将 block1 与 block0 相比的似然比检验结果；而 Model 统计量则是上一个模型与现在方程中变量有变化后模型的似然比检验结果。

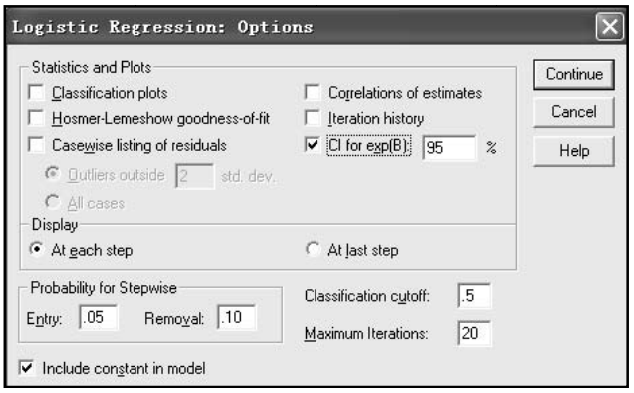


图 3.4　Logistic 回归 Options 子对话框

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	11.804	5	.038
	Block	11.804	5	.038
	Model	11.804	5	.038

图 3.5　Logistic 回归分析结果（全局检验）

如果采用 Enter 法，得到的 3 个统计量及假设检验结果是一致的，即 χ^2=11.804，P=0.038，说明模型有统计学意义。

如图 3.6 所示的输出表输出的“−2 Log likelihood”可用于拟合优度检验结果。而“Cox&Snell R Square”和“Nagelkerke R Square”类似于线性回归中的决定系数。

Model Summary

Step	−2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	159.681[a]	0.089	0.120

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than 0.001

图 3.6　Logistic 回归分析结果（Model Summary）

如图 3.7 所示的输出表是 Logistic 回归分析中最重要的表。表中的 B 为回归系数，S.E. 为标准误，Wald 为 Wald χ^2 值，df 为自由度，Sig.为 P 值，Exp(B)为 OR 值，95.0% C.I.for EXP(B)为 OR 值的 95%置信区间，其中 Lower 为置信区间的下限，Upper 为置信区间的上限。该回归分析结果说明，只有年龄是 TSH 水平的影响因素（P=0.021），而其他因素的回归系数均无统计学意义（P>0.05）。年龄的 OR 值（OR=1.106，95%CI：1.015～1.206）可解释为年龄每增加一岁，TSH 异常的风险增加到原来的 1.106 倍。Logistic 回归方程可以表示如下：

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	年龄	0.101	0.044	5.311	1	0.021	1.106	1.015	1.206
	病程	−0.002	0.025	0.008	1	0.929	0.998	0.949	1.049
	血红蛋白	−0.027	0.215	0.016	1	0.899	0.973	0.638	1.483
	空腹血糖	−0.090	0.185	0.238	1	0.625	0.914	0.637	1.312
	右眼病变	0.154	0.081	3.663	1	0.056	1.167	0.996	1.366
	Constant	−6.977	3.435	4.125	1	0.042	0.001		

a. Variable(s) entered on step 1: 年龄, 病程, 血红蛋白, 空腹血糖, 右眼病变

图 3.7　Logistic 回归分析结果（Variables in the Equation）

$$\ln\left(\frac{\pi}{1-\pi}\right) = -6.977 + 0.101 * 年龄$$

如果采用逐步向前法(Forward LR)，最终的输出结果如图 3.8 所示。该结果与采用 Enter 方法获得的结果基本相似，年龄仍然是唯一有统计学意义的变量（P=0.016），年龄的 OR 值（OR=1.106，95%CI：1.019～1.202）。

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	年龄	0.101	0.042	5.753	1	0.016	1.106	1.019	1.202
	Constant	−7.717	3.095	6.217	1	0.013	0.000		

a. Variable(s) entered on step 1: 年龄

图 3.8　Logistic 回归分析结果（Variables in the Equation）

3.4　应用实例 2：列联表资料的 Logistic 回归

列联表资料的数据除了可以用卡方检验进行比较之外，同样也可以采用 Logistic 回归进行分析。来看下面的例子，如表 3.1 所示。

表 3.1　COPD 患者与非患者的吸烟情况资料

	有吸烟史	无吸烟史	合计
患者	231	125	356
非患者	183	296	479
合计	414	421	835

下面建立 Logistic 回归模型方程分析病例与吸烟的关系。令 Y＝1 表示患者，Y＝0 表示非患者；X＝1 表示吸烟，X＝0 表示不吸烟。由此，可建立 Logistic 回归方程为：

$$\ln\frac{\pi}{1-\pi}=\beta_0+\beta_1 X$$

其中，π 表示 COPD 发病率。下面用 SPSS 软件完成 Logistic 回归分析。首先进行数据录入：病例否变量中 1 表示 COPD 患者，0 表示非患者；吸烟否变量中，1 表示吸烟，0 表示不吸烟（图 3.9）。

然后对人数进行加权，选择 Data→Weight Cases（图 3.10）。

病例否	吸烟否	人数
1.00	1.00	231.00
1.00	.00	125.00
.00	1.00	183.00
.00	.00	296.00

图 3.9　列联表资料的 Logistic 回归数据录入界面

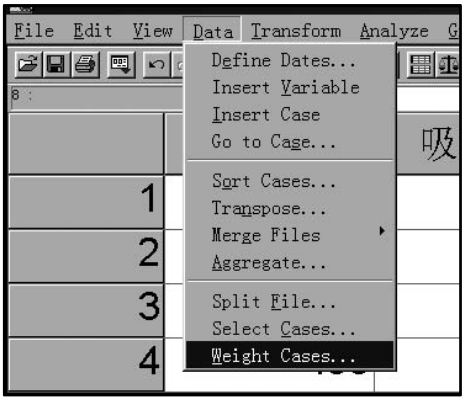


图 3.10　列联表资料的加权菜单

弹出对话框后，对人数进行加权（图 3.11）。

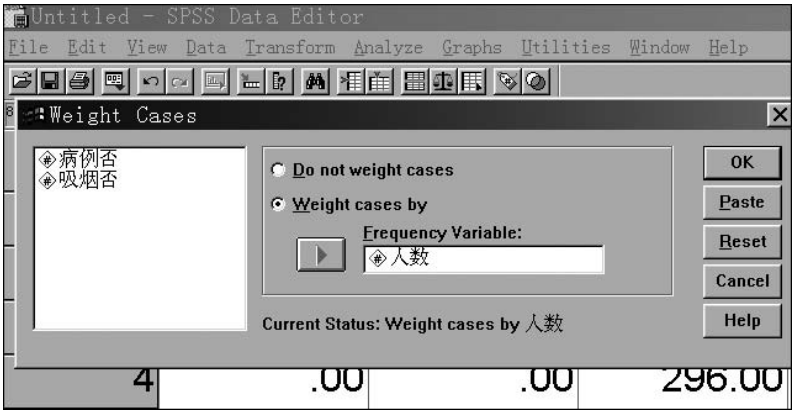


图 3.11　列联表资料的加权对话框

对人数进行加权后，选择菜单 Analyze→Regression→Binary Logistic，在弹出的对话框中选择因变量“Dependent”为“病例否”，协变量“Covariates”为“吸烟否”（图 3.12）。

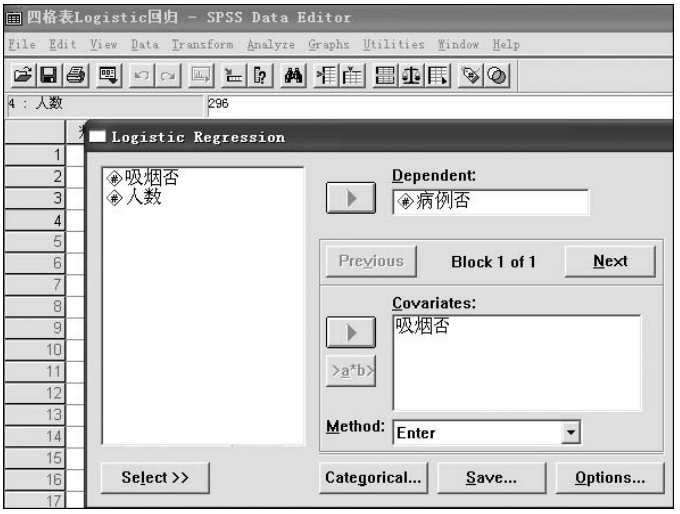


图 3.12　列联表资料的 Logistic 回归主对话框

主要的输出结果如图 3.13 所示。

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	吸烟否	1.095	0.146	56.631	1	0.000	2.989	2.247	3.976
	Constant	−0.862	0.107	65.310	1	0.000	0.422		

a. Variable(s) entered on step 1: 吸烟否

图 3.13　Logistic 回归分析结果（Variables in the Equation）

如图 3.13 所示的输出结果说明，吸烟是罹患 COPD 的影响因素（P<0.001，OR=2.989，95%CI：2.247～3.976）。如果写成方程，则可以写成：$\text{logit}(\pi) = -0.862 + 1.095$ 吸烟，说明吸烟对罹患 COPD 有影响，吸烟者患病的优势是不吸烟的近 3 倍。

3.5　应用实例 3：多项 Logistic 回归分析

如果响应变量是多分类的，此时就需要采用多项 Logistic 回归分析。与二项 Logistic 回归不同，它是通过拟合广义 Logit 模型方法进行的。若响应变量有 K 个分类，则有一个参照类，其余每一类与参照类比较，拟合 K-1 个 Logit 模型。例如，因变量取 3 个值 a、b、c，如果以 a 为参照水平，则得到 2 个 Logistic 函数，一个是 b 与 a 相比，另一个是 c 与 a 相比。下面通过一个案例进行分析。

这里选择 R 软的流行病学软件包 epicalc 中的数据 Ectopic。先来看一下数据，在 R 窗

口中输入语句：

```
install.packages(pkgs="epicalc")          （安装 R 软件流行病学 epicalc 软件包）
library(epicalc)                          （加载 R 软件流行病学 epicalc 软件包）
data (Ectopic)                            （加载 Ectopic 数据集）
Ectopic                                   （输出 Ectopic 数据）
```

输出的结果如图 3.14 所示。

	id	outc		hia	gravi
1	1	Deli	ever	IA	1-2
2	2	Deli	ever	IA	3-4
3	3	Deli	never	IA	1-2
4	4	Deli	never	IA	1-2
5	5	Deli	never	IA	1-2
6	6	IA	ever	IA	1-2
7	7	IA	never	IA	1-2
8	8	IA	never	IA	1-2
9	9	IA	ever	IA	1-2
10	10	IA	ever	IA	1-2
11	11	EP	ever	IA	3-4
12	12	EP	ever	IA	>4
13	13	EP	ever	IA	1-2
14	14	EP	ever	IA	1-2
15	15	Deli	ever	IA	>4
16	16	IA	never	IA	3-4
17	17	EP	ever	IA	3-4
18	18	EP	never	IA	3-4
19	19	Deli	ever	IA	1-2
20	20	Deli	never	IA	1-2
21	21	IA	never	IA	1-2

图 3.14　epicalc 软件包中的数据 Ectopic

这里仅列出了部分数据。该数据是检验先前的人工流产是否为当前宫外孕的一个危险因素的病例对照研究。总样本数为 723 人。研究的患者（变量 outc）分为 3 组，分别为宫外孕患者（EP）、来做人工流产的孕妇（IA）和来分娩的孕妇（Deli）。变量 hia 表示患者的人工流产史：分为从未做过人工流产组（never IA）和曾经做过人工流产组（ever IA）。变量 gravi 表示患者的怀孕次数：分为怀孕 1～2 次（1～2），怀孕 3～4 次（3～4）和怀孕>4 次（>4）。下面以 outc 变量作为分类响应变量，hia 和 gravi 作为影响因素进行多项 logistc 回归分析。

由于 SPSS 软件做多项 Logistic 回归有一些不足，因此这里采用 R 软件的 nnet 软件包分析，简单又便于解释。在加载 epicalc 软件包时，会自动加载上 nnet 软件包。

由于在多项分类的 Logistic 回归中，响应变量都是高于 2 个水平。通常系统会以默认的第一个水平作为参照水平。但有时如果以第一个水平作为参照，可能不便于结果的解释，因此可以通过自行设定参照水平进行分析。

例如，想以 Deli 作为参照水平，可以在 R 窗口中输入如下语句：

```
outc<-Ectopic[,2]                         （从数据集中选择分类响应变量 outc）
```

```
EP<-outc=="EP"                                    （筛选出 outc 中的 EP 样本）
IA<-outc=="IA"                                    （筛选出 outc 中的 IA 样本）
Deli<-outc=="Deli"                                （筛选出 outc 中的 Deli 样本）
multi<-multinom(cbind(Deli,EP,IA)~hia+gravi,data=Ectopic)
                                                  （进行多项 logistic 回归分析，其中 cbind
                                                  表示将 outc 的三个水平合并，列在第一个的
                                                  Deli 被作为参照水平，如果想让 EP 作为参
                                                  照水平，只需要将 EP 列在第一位即可）
mlogit.display(multi)                             （显示多项 logistic 回归分析结果）
```

输出的结果如图 3.15 所示。

```
Outcome =cbind(Deli, EP, IA); Referent group = Deli
            EP                              IA
            Coeff./SE      RRR(95%CI)       Coeff./SE      RRR(95%CI)
(Intercept) -1.02/0.154*** -                -0.51/0.131*** -
hiaever IA  1.49/0.222***  4.44(2.88,6.86)  0.38/0.215     1.47(0.96,2.23)
gravi3-4    0.47/0.24      1.59(1,2.55)     0.85/0.237***  2.35(1.48,3.73)
gravi>4     0.7/0.366      2(0.98,4.11)     1.16/0.369**   3.2(1.55,6.61)

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual Deviance: 1489.17
AIC = 1505.17
```

图 3.15　多项 Logistic 回归分析结果

上述结果解释为：把分娩孕妇（Deli）作为参照水平，对有人工流产史的妇女（ever IA），在此次入院中出现宫外孕的风险增加为没有人工流产史妇女（never IA）的 4.44 倍（P<0.001，具有高度显著性）。把分娩孕妇（Deli）作为参照水平，对有人工流产史的妇女（ever IA），在此次入院中再次进行人工流产的风险是没有人工流产史妇女（never IA）的 1.47 倍（P>0.1，无统计学显著性）。

对于怀孕次数来说，把分娩孕妇（Deli）作为参照水平，怀孕 3～4 次的妇女（gravi3-4），在此次入院中出现宫外孕的风险增加为怀孕 1～2 次妇女（gravi1-2）的 1.59 倍（P>0.1，无统计学显著性）。怀孕超过 4 次（gravi>4）的妇女，在此次入院中出现宫外孕的风险增加为怀孕 1～2 次妇女（gravi1-2）的 2 倍（P>0.1，无统计学显著性）。同样，对于怀孕次数来说，把分娩孕妇（Deli）作为参照水平，怀孕 3～4 次的妇女（gravi3-4），在此次入院中再次人工流产的风险增加为怀孕 1～2 次妇女（gravi1-2）的 2.35 倍（P<0.001，具有高度显著性）。怀孕超过 4 次的妇女（gravi>4），在此次入院中再次人工流产的风险增加为怀孕 1～2 次妇女（gravi1-2）的 3.2 倍（P=0.001，具有高度显著性）。

研究者在进行疾病的风险因素分析时，常常会受到混杂因素的干扰，如年龄、性别、病程长短和病情轻重等。如果不对混杂因素加以控制，会使研究结果产生偏倚。Logistic

回归方法能够充分利用数据信息，有效地控制混杂因素，得到校正后风险因素的 OR 估计值和可信区间。

Logistic 回归与多元线性回归的区别在于多元线性回归模型的响应变量值是具体数值，因此模型的预测值具有实际意义。而 Logistic 回归模型的响应变量是分类变量，因此模型的预测更关注于样本的分类标识。Logistic 回归中的自变量既可以是连续变量，也可以是分类变量。但应用 Logistic 回归时应注意样本量的问题。一般样本量应是自变量个数的 20～30 倍。在样本量足够大，且自变量之间无相关性的情况下，可以将全部自变量进入回归方程中，并采用逐步回归方法筛选有统计学意义的变量。在样本含量不高的情况下，可以先采用单因素分析筛选出有统计学意义的变量，将这些变量进入到回归方程中再进行筛选。在作多项 Logistic 回归时，由于变量关系较为复杂，且涉及到参照组的选择，因此在作结果解释和下结论的时候要谨慎。

3.6 Logistic 回归模型的 Nomogram 图展示

诺莫（Nomogram）图，又称为列线图，是一种综合分析多个定量变量和定性变量以预测某特定事件发生的绘图法预测模型，它可以用一种直观的绘图来对个体患者进行风险评估。该模型可以基于 Logistic 回归模型和 Cox 回归模型，将其结果进行可视化的呈现。它根据模型回归系数的大小来制定评分标准，给每个自变量的每种取值赋值一个评分，对每个患者，均可计算得到一个总分，再通过得分与结局发生概率之间的转换函数来计算每个患者的结局时间发生的概率，其轴结构和风险点反映了各个变量对预测结果的影响和重要性。

目前在一些国家和地区，这种预测模型已经受到广大患者和临床医师的认可，患者和临床医师都被鼓励使用这种模型进行预后风险评估工作。下面通过一个案例讲解如何绘制 Logistic 回归模型的诺莫图。

首先安装 R 软件的 rms 软件包。在 R 窗口中输入语句：

```
install.packages("rms")                    （安装 R 软件的 rms 软件包）
library(rms)                               （加载 R 软件的 rms 软件包）
```

这里提供一个案例数据，其中 gene1 和 gene2 中的 0 表示基因低表达，1 表示基因高表达；stage 中的 0 表示肿瘤早的分期，1 表示肿瘤晚的分期；acid 表示某项临床指标；outcome 中的 0 表示存活，1 表示死亡。图 3.16 列出了数据集中的部分信息。数据保存为 nuomo.csv 文件，并存储于 D 盘中。

gene1	gene2	stage	age	acid	outcome
0	1	1	64	40	0
0	0	1	63	40	0
1	0	0	65	46	0
0	1	0	67	47	0
0	0	0	66	48	0
0	1	1	65	48	0
0	0	0	60	49	0
0	0	0	51	49	0
0	0	0	66	50	0
0	0	0	58	50	0
0	1	0	56	50	0
0	0	1	61	50	0
0	1	1	64	50	0
0	0	0	56	52	0
0	0	0	67	52	0
1	0	0	49	55	0
0	1	1	52	55	0
0	0	0	68	56	0
0	1	1	66	59	0
1	0	0	60	62	0
0	0	0	61	62	0
1	1	1	59	63	0
0	0	0	51	65	0
0	1	1	53	66	0
0	0	0	58	71	0
0	0	0	63	75	0
0	0	1	53	76	0
0	0	0	60	78	0
0	0	0	52	83	0
0	0	1	67	95	0
0	0	0	56	98	0
0	0	1	61	102	0
0	0	0	64	187	0

图 3.16　绘制诺莫图的案例数据（部分数据）

然后读入数据并构建数据列表。

```
read.table("D:\\nuomo.csv",header=TRUE,sep=",")->data    （读入数据）
ddist<-datadist(data)                                    （构建数据列表）
options(datadist="ddist")
```

以 outcome 为因变量，以 gene1、gene2、age 和 stage 为自变量构建 Logistic 回归模型，在 R 窗口中输入如下语句：

```
model<-lrm(outcome~gene1+gene2+age+stage,data)      （构建 Logistic 回归模型）
```

输出结果如图 3.17 所示。

其中，在“Rank Discrim.Indexes”中的 C 为模型区分度评价的统计量，也就是 AUC=0.813，说明模型的区分度较好。gene1(P=0.0036)和 stage(P=0.0356)是 outcome 的影响因素。

下面构建诺莫图的绘图函数。在 R 窗口中输入如下语句：

```
nom<-nomogram(model,lp.at=seq(-2,4,by=0.5),   （设置坐标轴的范围，从-2 到 4，步长为 0.5）
fun=function(x)1/(1+exp(-x)),                 （转化为风险概率）
funlabel="Risk of Death",                     （给新的坐标轴设置名称）
fun.at=c(0.05,seq(0.1,0.9,by=0.1),0.95),      （给新的坐标轴设置范围）
```

```
conf.int=c(0.1,0.3))                    （显示置信区间）
nom                                     （输出 nom 函数结果）
```

```
Logistic Regression Model

 lrm(formula = outcome ~ gene1 + gene2 + age + stage, data = data)

                       Model Likelihood     Discrimination    Rank Discrim.
                          Ratio Test           Indexes           Indexes
 Obs            55    LR chi2     18.99    R2        0.395    C         0.813
  0             33    d.f.            4    g         1.665    Dxy       0.627
  1             22    Pr(> chi2) 0.0008    gr        5.283    gamma     0.631
 max |deriv| 3e-05                          gp        0.314    tau-a     0.306
                                            Brier     0.167

           Coef    S.E.   Wald Z Pr(>|Z|)
 Intercept  1.3858 3.1866  0.43  0.6636
 gene1      2.1971 0.7549  2.91  0.0036
 gene2      0.2837 0.7198  0.39  0.6935
 age       -0.0574 0.0537 -1.07  0.2849
 stage      1.5240 0.7252  2.10  0.0356
```

图 3.17　构建 Logistic 回归模型

nom 函数的输出结果如图 3.18 所示。

```
Points per unit of linear predictor: 45.51428
Linear predictor units per point   : 0.02197113


 gene1 Points
 0       0
 1     100


 gene2 Points
 0       0
 1      13


 age Points
 44  63
 46  57
 48  52
 50  47
 52  42
 54  37
 56  31
 58  26
 60  21
 62  16
 64  10
 66   5
 68   0


 stage Points
 0       0
 1      69


 Total Points Risk of Death
           52          0.20
           76          0.30
           96          0.40
          115          0.50
          133          0.60
          153          0.70
          178          0.80
          215          0.90
          249          0.95
```

图 3.18　nom 函数的输出结果

该输出结果中含有给每个变量的打分及总分（Total Points），并根据总分计算出相应的死亡风险概率。例如，如果 gene1 低表达就赋予 0 分，高表达则赋予 100 分。下面根据输出的 nom 函数结果进行绘图。在 R 窗口中输入语句：

```
plot(nom,lplabel="linear Predictor",          （绘制线性预测坐标轴）
fun.side=c(3,1,1,1,3,1,3,1,1,1,3),            （设置风险概率坐标轴刻度：1，3 表示刻度出现在轴的上下方）
label.every=2,                                （设置变量之间的间隔）
col.conf=c("red","green"),                    （设置置信区间颜色）
conf.space=c(0.1,0.3),                        （显示置信区间）
col.grid=gray(c(0.8,0.95)))                   （设置图片沿格线的灰度颜色）
```

绘制的诺莫图如图 3.19 所示。

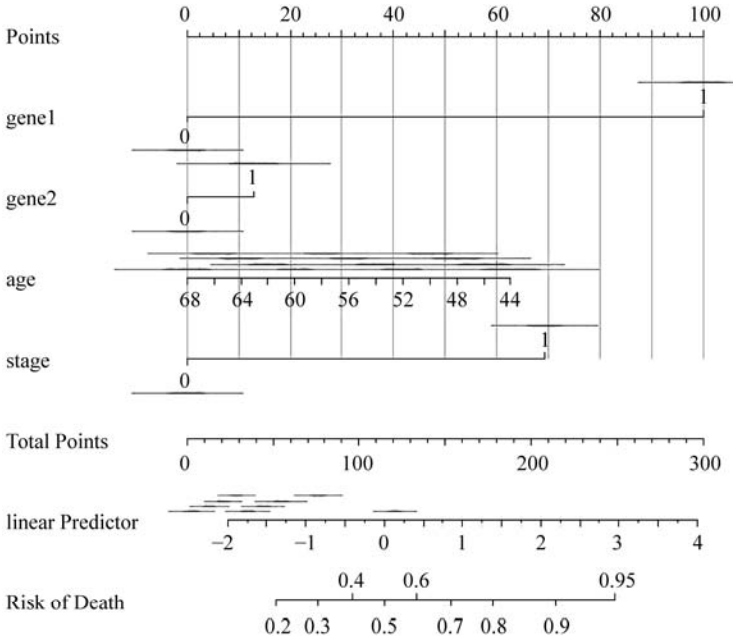


图 3.19　Logistic 回归模型的诺莫图

下面对诺莫图进行解释。在图 3.19 中，患者的取值都位于每个变量轴上，向上绘制一条直线以确定每个变量值对应的分数（Points）。所有变量获得的分数总和为 Total Points。将 Total Points 向下垂直延伸到 Risk of Death 概率轴，就可以获得患者发生死亡的风险概率了。例如，某个患者 gene1=1（高表达，Points=100），gene2=1（高表达，Points=12），age=64（Points≈10），stage=0（Points=0），Total Points=122；在 Total Points 轴上找到此数值并向

Risk of Death 概率轴作垂线，则可知该患者死亡风险概率为 0.5～0.6。

为了使诺莫图更加清晰，也可以将置信区间去掉。此时将构建的诺莫图的绘图函数和绘图语句可以修改如下：

```
nom<-nomogram(model,
lp.at=seq(-2,4,by=0.5),
fun=function(x)1/(1+exp(-x)),
funlabel="Risk of Death",
fun.at=c(0.05,seq(0.1,0.9,by=0.1),0.95))
plot(nom,lplabel="linear Predictor",
fun.side=c(3,1,1,1,3,1,3,1,1,1,3),
label.every=2,
col.grid=gray(c(0.8,0.95)))
```

修改后的诺莫图如图 3.20 所示，此时置信区间去掉，图片显得更为清晰。

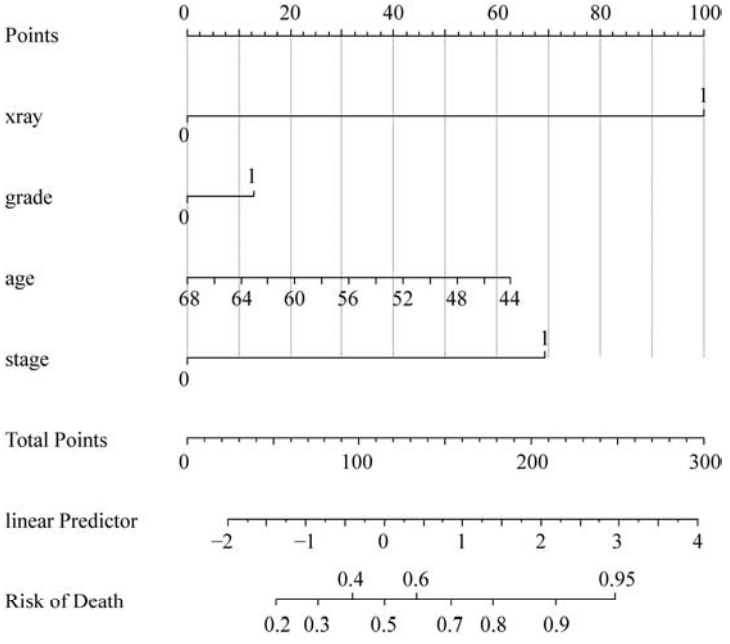


图 3.20　Logistic 回归模型的诺莫图（去除置信区间）

3.7　多个 Logistic 回归模型评价的决策曲线分析法

通常情况下，评价一种诊断方法是否好用一般是做 ROC（Receiver Operating Characteristic Curve）曲线并计算 AUC（Area Under ROC Curve）。但是，ROC 只是从该方法的特异性和

敏感性考虑，追求的是准确率。当通过某个生物标志物预测患者是否患了某种疾病，无论选取哪个值为临界值，都会遇到假阳性和假阴性的可能，有时避免假阳性时受益更大，而有时则希望能够避免假阴性。既然两种情况都无法避免，那就需要找到一个净收益最大的办法，而这个问题就是临床效用问题。对此，Andrew Vickers 博士等研究出另外一种评价方法，也就是决策曲线分析（Decision Curve Analysis，DCA）法。

DCA 分析涉及 3 个基本概念：**阈值概率**（Threshold Probability）、**净收益**（Net Benefit）和**权重因子**（Weighting Factor）。

（1）阈值概率为患者选择治疗的诊断确定性水平。阈值概率考虑患者治疗与否的相对价值，如果治疗效果高且风险低，则选择治疗的阈值概率低；如果治疗效果极低或有很大的风险，则选择治疗的阈值概率将很高。

（2）净收益是指治疗决策所带来的预期收益和预期伤害之和。

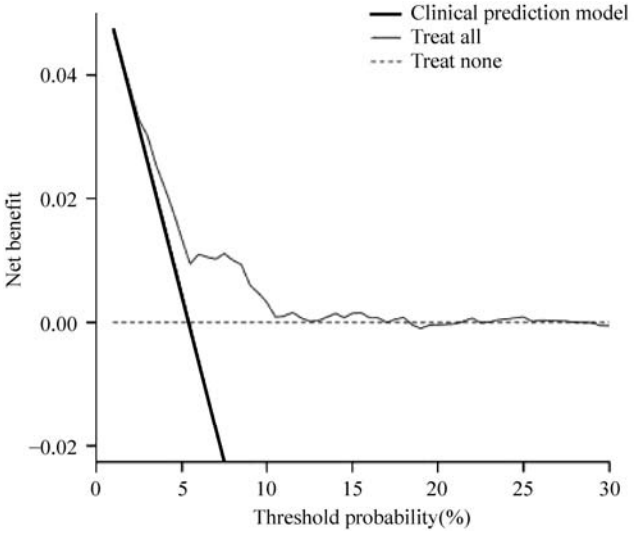
（3）权重因子衡量患者不接受治疗（或治疗不足）和过度治疗所带来的风险。

这三者之间的关系是：

$$\text{净收益} = \text{真阳性率} - (\text{假阳性率} \times \text{权重因子})$$

$$\text{权重因子} = \frac{\text{阈值概率}}{1 - \text{阈值概率}}$$

例如，发表在 2018 年 *Lancet Haematology* 杂志上的一篇论文就采用了决策曲线分析来评价临床预测模型。该论文用两个队列构建了肿瘤相关的静脉血栓栓塞的预测模型。其中，一个队列的模型评价采用了决策曲线分析，如图 3.21 所示。



扫码看彩图

图 3.21　决策曲线评价临床预测模型（图片来自 *Lancet Haematology* 期刊）

图 3.21 中的横坐标为阈值概率，纵坐标为净增益。Treat all 和 Treat none 表示两种极端情况。Treat none 表示不治疗，所有样本净获益均为 0；Treat all 表示全面治疗，其净获益是斜率为负值的反斜线。从图中可以看到，相比不治疗或全面治疗，该预测模型有更好的临床应用价值。

下面应用 R 软件来绘制决策曲线分析图。首先需要安装和加载 R 软件的 rmda 软件包。这里采用该软件包自带的数据 dcaData，数据结构如图 3.22 所示。该数据中含有年龄 Age（连续变量）、性别 Female（二分类变量：1 为女性，0 为男性）、Smokes（二分类变量：FALSE 为不吸烟，TRUE 为吸烟）、Marker1 和 Marker2（两个肿瘤标记物，连续变量）、Cancer（二

Age	Female	Smokes	Marker1	Marker2	Cancer
33	1	FALSE	0.245311	1.021085	0
29	1	FALSE	0.942966	-0.25576	0
28	1	FALSE	0.773594	0.331844	0
27	0	FALSE	0.406359	-0.00569	0
23	1	FALSE	0.507515	0.207533	0
35	1	FALSE	0.185671	1.41251	0
34	1	FALSE	0.621037	0.615094	0
29	1	FALSE	0.401515	1.15764	0
35	1	FALSE	0.389584	1.38444	0
27	1	FALSE	0.150983	2.438673	0
25	0	FALSE	0.928431	0.824993	0
30	1	FALSE	0.672424	0.588465	0
28	1	FALSE	0.571275	-0.41087	0
24	1	TRUE	0.516645	-1.66893	1
33	1	FALSE	0.671228	1.374315	0
29	1	FALSE	0.543388	0.176627	0
24	1	FALSE	0.573114	-0.06622	0
23	0	TRUE	0.957571	-0.80387	0
24	0	TRUE	0.802387	0.027684	0
27	1	FALSE	0.531213	-0.58641	1
20	1	FALSE	0.529426	0.556889	0
30	0	FALSE	0.55679	-2.72032	1
22	1	FALSE	0.187532	0.168387	0
20	0	TRUE	0.961829	0.315848	0
28	0	FALSE	0.048032	-1.43479	0
26	0	TRUE	0.362291	0.033908	0
21	0	FALSE	0.48976	0.531797	0
27	1	FALSE	0.953092	-1.12645	1
32	1	FALSE	0.171899	1.084137	0
20	0	TRUE	0.772615	1.498544	0
25	1	FALSE	0.364142	0.89579	0
26	0	FALSE	0.512712	1.26453	0
30	1	FALSE	0.357476	0.293341	0
31	1	FALSE	0.581704	0.663437	0
31	1	FALSE	0.739045	1.239507	0
24	1	TRUE	0.936645	0.347186	0
24	1	TRUE	0.270908	1.425537	0
35	1	FALSE	0.911283	1.992642	0
30	0	FALSE	0.454406	1.318215	0
25	0	FALSE	0.735642	-0.00131	0
34	1	FALSE	0.931226	-1.20328	1

图 3.22　决策曲线分析的数据结构

分类变量：0 为非癌患者，1 为癌症患者）。在 R 窗口中输入语句：

```
install.packages("rmda")                    （安装 R 软件的 rmda 软件包）
library(rmda)                               （加载 R 软件的 rmda 软件包）
data(dcaData)                               （加载软件包自带数据 dcaData）
```

下面构建两个 Logistic 回归模型，其中一个是以癌症（Cancer）为因变量，以年龄（Age）、性别（Female）和吸烟（Smokes）为自变量的模型 model1；另一个是以癌症（Cancer）为因变量，以年龄（Age）、性别（Female）、吸烟（Smokes）以及肿瘤标志物 1（Marker1）和肿瘤标志物 2（Marker2）为自变量的模型 model2，在 R 窗口中输入语句：

```
model1=Cancer~Age+Female+Smokes                     （构建模型 model1）
model2=Cancer~Age+Female+Smokes+Marker1+Marker2     （构建模型 model2）
```

下面采用 decision_curve()函数分别以公式 model1 和 model2 构建基于 Logistic 回归的 DCA 模型，在 R 窗口中输入语句计算 model1 的决策曲线：

```
baseline_model<-decision_curve(formula = model1,    （model1 作为基线模型）
data=dcaData,
family=binomial(link='logit'),                      （构建 Logistic 回归模型）
thresholds=seq(0,1,by=0.01),                        （以 0.01 为步长取阈值概率）
confidence.intervals=0.95,                          （计算 95%置信区间）
study.design = "cohort")                            （研究设计为队列研究）
```

在该语句中 family=binomial(link='logit')表示使用 Logistic 模型进行拟合；threshold 设置横坐标阈值概率范围，一般是 0～1，但是如果遇到某种具体情况，在临床上一致认为阈值概率达到某个值以上，如 40%，则必须采取干预措施，那么 0.4 以后的研究就没什么意义了，此时可以设阈值概率为 0～0.4；study.design 为研究类型，根据队列研究和病例对照研究，此值分别可设为“cohort”或“case-control”。

同样，在 R 窗口中输入语句计算 model2 的决策曲线：

```
full_model<-decision_curve(formula=model2,          （model2 作为完全模型）
data=dcaData,
family=binomial(link='logit'),                      （构建 Logistic 回归模型）
thresholds=seq(0,1,by=0.01),                        （以 0.01 为步长取阈值概率）
confidence.intervals=0.95,                          （计算 95%置信区间）
study.design = "cohort")                            （研究设计为队列研究）
```

下面用 plot_decision_curve()函数绘制决策曲线，在 R 窗口中输入语句：

```
plot_decision_curve(list(baseline_model,full_model),　（绘制两个模型的决策曲线）
curve.names=c("Baseline model","Full model"),　（设置曲线名称）
col=c("blue","red"),lty=c(1,2),lwd=c(3,2,2,1),　（设置不同模型曲线的颜色，样式和宽度）
legend.position="topright",　（设置图注的位置）
confidence.intervals=FALSE,　（确定是否绘制曲线的置信区间）
cost.benefit.axis=FALSE)　（确定是否绘制成本效益坐标轴）
```

绘制出的决策曲线图如图 3.23 所示。


扫码看彩图

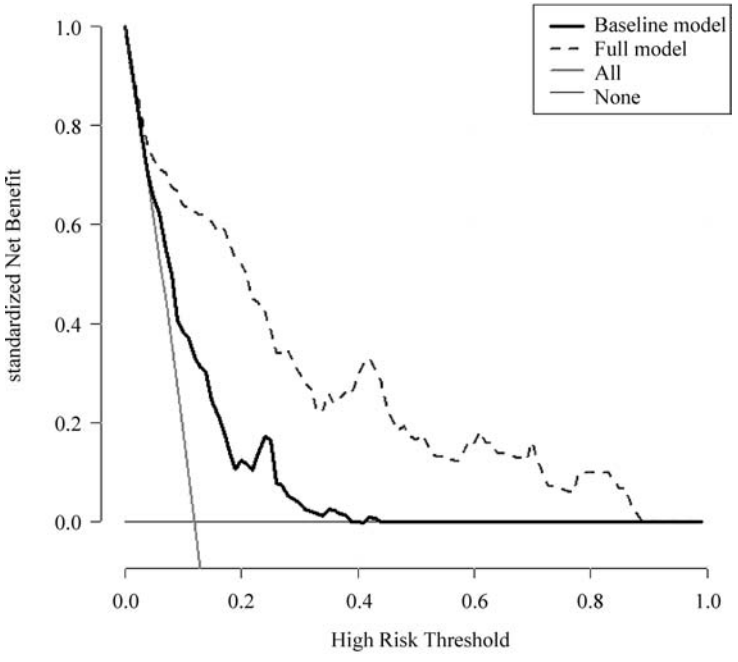


图 3.23　两个 Logistic 回归模型的决策曲线分析

图 3.23 中横坐标为阈值概率，纵坐标为标准净增益。All 和 None 表示两种极端情况：None 表示所有样本都净获益为 0；All 表示所有样本都有净获益，该净获益是斜率为负值的反斜线。从图 3.23 可以看到，full_model 模型的净收益比 baseline_model 模型高。这说明两个肿瘤标志物对癌症预测起了一定的作用。