第七章 广义线性模型

复旦大学附属肿瘤医院 周支瑞

7.1 广义线性模型和glm()函数

线性模型的扩展

- 结果变量是类别型的。二值变量（比如：是/否、有效/无效、活着/死亡）和多分类变量（比如差/良好/优秀）都显然不是正态分布。
- 结果变量是计数型的。（比如，一年内哮喘发生的次数，一生中流产的次数，一周交通事故的数目，每日酒水消耗的数量）。这类变量都是非负有限值，而且它们的均值和方差通常都是相关的（正态分布变量间不是如此，而是相互独立）。
- 广义线性模型扩展了线性模型的框架，它包含了非正态因变量的分析。

标准线性模型公式

- 现要对响应变量Y和p个预测变量$X_1 \ldots X_p$间的关系进行建模。在标准线性模型中，可假设Y呈正态分布，关系的形式为：

$$\mu_y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

- 该等式表明响应变量的条件均值是预测变量的线性组合。参数β_j指一单位X_j的变化造成的Y预期的变化，β_0指当所有预测变量都为0时Y的预期值。对于该等式，你可通俗地理解为：给定一系列X变量的值，赋予X变量合适的权重，然后将它们加起来，便可预测Y观测值分布的均值。

广义线性模型公式

- 上使中并没有对预测变量Xj做任何分布的假设。与Y不同，它们不需要呈正态分布。实际上，它们常为类别型变量（比如方差分析设计）。另外，对预测变量使用非线性函数也是允许的，比如你常会使用预测变量X_2或者$X_1 \times X_2$，只要等式的参数（$\beta_0, \beta_1, \ldots, \beta_p$）为线性即可。广义线性模型公式为：

$$g(\mu_y) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

- 其中$g(\mu_Y)$是条件均值的函数（称为连接函数）。另外，你可放松Y为正态分布的假设，改为Y服从指数分布族中的一种分布即可。设定好连接函数和概率分布后，便可以通过最大似然估计的多次迭代推导出各参数值。

glm()函数

- R中可通过glm()函数（还可用其他专门的函数）拟合广义线性模型。它的形式与lm()类似，只是多了一些参数。函数的基本形式为：

  glm(formula, **family=family(link=function)**, data=)

glm()函数的连接函数

- 表13-1列出了概率分布（family）和相应默认的**连接函数（function）**。

表13-1 glm()的参数

分 布 族	默认的连接函数
binomial	(link = "logit")
gaussian	(link = "identity")
gamma	(link = "inverse")
inverse.gaussian	(link = "1/mu^2")
poisson	(link = "log")
quasi	(link = "identity", variance = "constant")
quasibinomial	(link = "logit")
quasipoisson	(link = "log")

lm()函数与glm()函数

- 标准线性模型也是广义线性模型的一个特例。如果令连接函数$g(\mu_Y)=\mu_Y$或恒等函数，并设定概率分布为正态（高斯）分布，那么：

  glm(Y~X1+X2+X3, family=gaussian(link="identity"), data=mydata)

- 生成的结果与下列代码的结果相同：

  lm(Y~X1+X2+X3, data=mydata)

Logistic回归模型

- Logistic回归适用于二值响应变量（0和1）。模型假设Y服从二项分布，线性模型的拟合形式为：

$$\log_e\left(\frac{\pi}{1-\pi}\right)=\beta_0+\sum_{j=1}^{p}\beta_j X_j$$

- 其中π=μ_Y是Y的条件均值（即给定一系列X的值时Y=1的概率），(π/1–π)为Y=1时的优势比，log(π/1–π)为对数优势比，或logit。本例中，log(π/1–π)为连接函数，概率分布为二项分布，可用如下代码拟合Logistic回归模型：

glm(Y~X1+X2+X3, family=binomial(link="logit"), data=mydata) #假设有一个响应变量（Y）、三个预测变量（X1、X2、X3）和一个包含数据的数据框（mydata）

泊松回归模型

- 泊松回归适用于在给定时间内响应变量为事件发生数目的情形。它假设Y服从泊松分布，线性模型的拟合形式为：

$$\log_e(\lambda) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

- 其中λ是Y的均值（也等于方差）。此时，连接函数为log(λ)，概率分布为泊松分布，可用如下代码拟合泊松回归模型：

```
glm(Y~X1+X2+X3, family=poisson(link="log"), data=mydata)
```

可连用的函数

- 与分析标准线性模型时lm()连用的许多函数在glm()中都有对应的形式，其中一些常见的函数如下：

  summary() 展示拟合模型的细节

  coefficients()、coef() 列出拟合模型的参数（截距项和斜率）

  confint() 给出模型参数的置信区间（默认为95%）

  residuals() 列出拟合模型的残差值

  anova() 生成两个拟合模型的方差分析表

  plot() 生成评价拟合模型的诊断图

  predict() 用拟合模型对新数据集进行预测

  deviance() 拟合模型的偏差

  df.residual() 拟合模型的残差自由度

模型拟合和回归诊断

- 当评价模型的适用性时，你可以绘制初始响应变量的预测值与残差的图形

  plot(predict(model, type="response"),residuals(model, type= "deviance"))
- R可列出帽子值（hat value）、学生化残差值和Cook距离统计量的近似值。

  plot(hatvalues(model))

  plot(rstudent(model))

  plot(cooks.distance(model))
- 还可以用其他方法，创建一个综合性的诊断图。在后面的图形中，横轴代表杠杆值，纵轴代表学生化残差值，而绘制的符号大小与Cook距离大小成正比。

  library(car)

  influencePlot(model)

7.2 Logistic回归模型

Logistic回归模型概述

- Logistic回归模型是一种概率模型，它是以某一事件发生与否的概率P为因变量，以影响P的因素为自变量建立的回归模型，分析某事件发生的概率与自变量之间的关系，是一种非线性回归模型。
- Logistic回归模型适用的资料类型：适用于因变量为二项或多项分类（有序、无序）的资料。

Logistic回归模型分类

- 1. 条件Logistic 回归模型：适合于配对或配伍设计资料；
- 2. 非条件Logistic回归模型：适合于成组设计的统计资料

- 因变量可以是：两项分类、无序多项分类、有序多项分类等

Logistic回归模型结构

- Logistic分布函数表达：

$$F(y) = \frac{e^y}{1 + e^y}$$

- Y 的取值在 - ∞ ~ + ∞之间，函数值F(y) 在0 ~ 1之间取值，且呈单调上升的S型曲线。可以将这一特征运用到临床医学和流行病学中描述事件发生的概率与影响因素的关系

Logistic分布函数

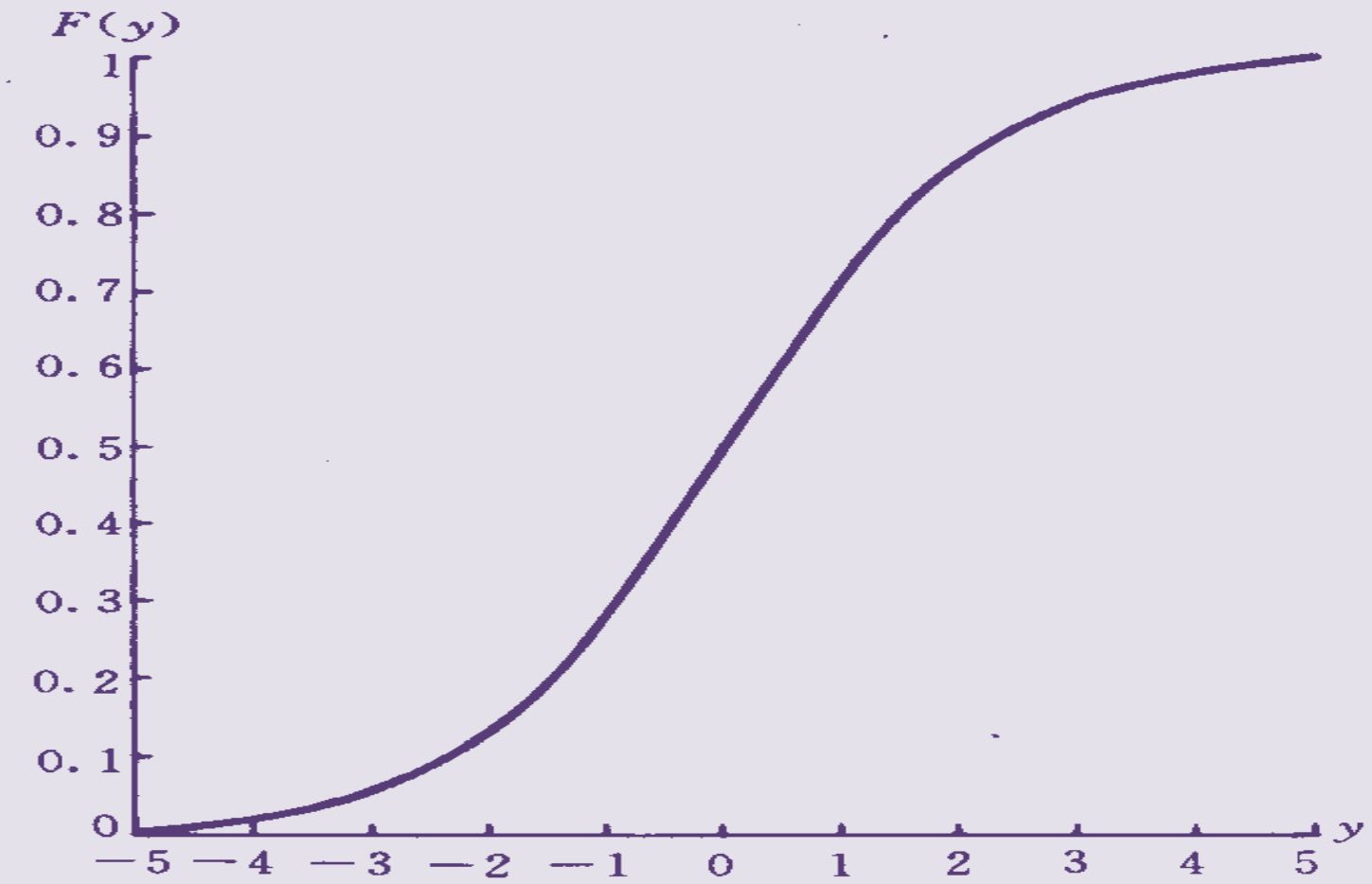


图 28.1　Logistic 分布函数的图型

Logistic回归模型方程

$$P(y = 1 \mid x) = \frac{\exp(\beta_0 + \beta X)}{1 + \exp(\beta_0 + \beta X)}$$

$$Q(y = 0 \mid x) = \frac{1}{1 + \exp(\beta_0 + \beta X)}$$

Logistic回归模型

- 利用Logistic分布函数的特征来表示在自变量X的作用下出现阳性结果或阴性结果的概率。
- 阳性结果的概率记为：P(y=1|x)，（在X 作用下，出现Y=1的概率)
- 出现阴性结果的概率为: Q(y=0|x)，（在X 作用下，出现Y=0的概率)

- 注意：P+Q=1。

Logistic回归模型

- 当只有一个自变量时，Logistic回归模型：

$$P(y = 1 \mid x) = \frac{\exp(\beta_0 + \beta X)}{1 + \exp(\beta_0 + \beta X)} \quad (1)$$

$$Q(y = 0 \mid x) = \frac{1}{1 + \exp(\beta_0 + \beta X)} \quad (2)$$

- 式中，β_0 为回归线的截距，β 是与X有关的参数，称回归系数。

Logistic回归模型

$$\frac{P(y=1 \mid x)}{Q(y=0 \mid x)} = \exp(\beta_0 + \beta X) \quad (3)$$

- 注意：P/Q称为事件的优势，流行病学中称为比值(odds)

Logistic回归模型

- 当有多个X时， Logistic回归模型：

$$P(y=1 \mid x)=\frac{\exp(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots.+\beta_p x_p)}{1+\exp(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots.+\beta_p x_p)} \quad (4)$$

$$Q(y=0 \mid x)=\frac{1}{1+\exp(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots.+\beta_p x_p)} \quad (5)$$

- 式中，β_0 为截距，β_j (j=1,2,...,p), 称偏回归系数

Logistic回归模型

$$\frac{P(y=1 \mid x)}{Q(y=0 \mid x)} = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p) \qquad (6)$$

- 式(1) 或 式(4)称为Logistic回归模型。

$$P(y=1 \mid x) = \frac{\exp(\beta_0 + \beta X)}{1 + \exp(\beta_0 + \beta X)} \qquad (1)$$

$$P(y=1 \mid x) = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p)} \qquad (4)$$

Logistic回归模型

- Logit变换: 将S型曲线转化为直线

$$\frac{P(y=1 \mid x)}{Q(y=0 \mid x)} = \exp(\beta_0 + \beta X) \quad (3)$$

$$\frac{P(y=1 \mid x)}{Q(y=0 \mid x)} = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p) \quad (6)$$

- 对式(3)和式(6) 两边取自然对数得:

$$\ln(P/Q) = \beta_0 + \beta X \quad (7)$$

$$\ln(P/Q) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p \quad (8)$$

Logistic回归模型

$$记\mathrm{logit}(P) = \ln(P/Q)$$

$$\log it(P) = \beta_0 + \beta x$$

$$\log it(P) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_p x_p$$

- 其实变换之后这就是所谓的线性回归方程:
  - ln(P/Q)称为 logit(P)变换;
  - P/Q称为事件的优势，流行病学中称为比值(odds)
- 因此，优势的对数值与影响因素之间呈线性关系。

Logistic回归模型

- 比值比 或 优势比(odds ratio), 简记为 OR
- 暴露组的优势（比值）与非暴露组的优势（比值）之比，称优势比（比值比）（OR）。OR用于说明暴露某因素引起疾病或死亡的危险度大小。

$$OR = \frac{P(1)/[1-P(1)]}{P(0)/[1-P(0)]} \quad (9)$$

Logistic回归模型

- 对式（9）两边取自然对数得：

$$\ln(OR) = \ln\left(\frac{P(1)/[1-P(1)]}{P(0)/[1-P(0)]}\right) = \log it[P(1)] - \log it[P(0)] \quad \textbf{(10)}$$

- P(1)—X取1时，暴露组；P(0)—X取0时，非暴露组

$$\ln(P/Q) = \beta_0 + \beta x$$

$$\ln(OR) = \log it[P(1)] - \log it[P(0)] = (\beta_0 - \beta \times 1) - (\beta_0 - \beta \times 0) = \beta$$

即： $\beta = \ln(OR)$ **或者：** $OR = e^{\beta}$

Logistic回归模型

- β 的统计学意义:

$$OR = \exp(\beta) = e^{\beta}$$

$$\beta = \ln(OR)$$

- 由上式可见，β 的意义是：在其他自变量固定不变的情况下，自变量的暴露水平每改变一个测量单位所引起的优势比（OR）自然对数的改变量，或引起优势比为增加前的e^{β}倍

Logistic回归案例 -- 结局是二分类变量最常用

- 【案例】Hosmer 和 Lemeshow于1989年研究了低出生体重婴儿的独立影响因素。结果变量为：**是否娩出低出生体重儿**（变量名为LOW，1=低出生体重，即婴儿出生体重<2500g; 0=非低出生体重），考虑的影响因素（自变量）有：产妇妊娠前体重（lwt，磅）；产妇年龄（age，岁）；产妇在妊娠期间是否吸烟（smoke，0＝未吸、1＝吸烟）；本次妊娠前早产次数（ptl，次）；是否患有高血压（ht，0＝未患、1＝患病）；子宫对按摩、催产素等刺激引起收缩的应激性（ui，0＝无、1＝有）；妊娠前三个月社区医生随访次数（ftv，次）；种族（race，1＝白人、2＝黑人、3＝其他民族）。

Logistic回归案例 -- 变量筛选方式

- SPSS可进行以下变量筛选方式：

1. 用Enter法把所有自变量全纳入（不做筛选）
2. 用逐步回归筛选自变量（SPSS软件包括6种筛选方式）
3. 先做单因素Logistic回归，$p<0.1$纳入最后的回归方程

- 最终筛选出影响新生儿体重的独立影响因素

Logistic回归案例 1

- 以AER包中的数据框Affairs为例，我们将通过探究婚外情的数据来阐述Logistic回归分析的过程。首次使用该数据前，请确保已下载和安装了此软件包（使用install.packages("AER")）。
- 婚外情数据取自1969年（Psychology Today）所做一个非常有代表性的调查。该数据从601个参与者收集了9个变量，包括一年来婚外私通的频率以及参与者性别、年龄、婚龄、是否有小孩、宗教信仰程度（5分制，1分表示反对，5分表示非常信仰）、学历、职业（逆向编号的戈登7种分类），还有对婚姻自我评分（5分制，1表示非常不幸福，5表示非常幸福）。

Logistic回归案例 1 代码


```
> data(Affairs, package="AER")
> summary(Affairs)
> table(Affairs$affairs)
> Affairs$ynaffair[Affairs$affairs > 0] <- 1
> Affairs$ynaffair[Affairs$affairs == 0] <- 0
> Affairs$ynaffair <- factor(Affairs$ynaffair, levels=c(0,1), labels=c("No","Yes"))
> table(Affairs$ynaffair)
> fit.full <- glm(ynaffair ~ gender + age + yearsmarried + children +religiousness +
education + occupation +rating, data=Affairs, family=binomial())
> summary(fit.full)
> fit.reduced <- glm(ynaffair ~ age + yearsmarried + religiousness + rating,
data=Affairs, family=binomial())
> summary(fit.reduced)
> anova(fit.reduced, fit.full, test="Chisq")
> coef(fit.reduced)
> exp(coef(fit.reduced))
```

评价预测变量对结果概率的影响


```
> testdata <- data.frame(rating=c(1, 2, 3, 4, 5), age=mean(Affairs$age),
yearsmarried=mean(Affairs$yearsmarried),
religiousness=mean(Affairs$religiousness))
> testdata
> testdata$prob <- predict(fit.reduced, newdata=testdata, type="response")
> testdata

> testdata <- data.frame(rating=mean(Affairs$rating), age=seq(17, 57, 10),
yearsmarried=mean(Affairs$yearsmarried),
religiousness=mean(Affairs$religiousness))
> testdata
> testdata$prob <- predict(fit.reduced, newdata=testdata, type="response")
> testdata
```

案例 1 是否还有其他解决办法?

- 在婚外情的例子中，婚外偷腥的次数被二值化为一个“是/否”的响应变量，这是因为我们最感兴趣的是在过去一年中调查对象是否有过一次婚外情。如果兴趣转移到量上（过去一年中婚外情的次数），便可直接对计数型数据进行分析。分析计数型数据的一种流行方法是**泊松回归**。

Logistic回归案例 2

- 表11-5 是一个研究吸烟、饮酒与食管癌关系的病例-对照研究资料，试作Logistic 回归分析。设y=1 表示患有食管癌， y=0 表示未患食管癌。令x1=1 表示吸烟， x1=0 表示不吸烟: x2=1表示饮酒， x2=0 表示不饮酒.

表 11-5　　吸烟、饮酒与食管癌关系的病例对照研究资料

分　层	吸　烟	饮　酒	阳性数	阴性数
1	0	0	63	136
2	0	1	63	107
3	1	0	44	57
4	1	1	265	151

Logistic回归案例 2 代码


```
> Example11_4  <- read.table ("example11_4.csv", header=TRUE, sep=",")
> attach(Example11_4)

> fit1 <- glm(y~ x1+ x2, family= binomial(), data=Example11_4)
> summary(fit1)
> coefficients(fit1)
> exp(coefficients(fit1))
> exp (confint(fit1))

> fit2 <- glm(y~ x1 + x2 +x1:x2 ,  family= binomial(), data=Example11_4)
> summary(fit2)
> coefficients(fit2)
> exp(coefficients(fit2))
> exp (confint(fit2))
```

Logistic回归案例 2 代码 续

- **吸烟与饮酒的交互作用**，本例分析吸烟、饮酒危险因素对患食管癌的影响程度以及它们的交互影响程度。设y=1 表示患有食管癌， y=0 表示未患食管癌。令x1=1 表示吸烟， x1=0 表示不吸烟: x2=1 表示饮酒， X2=0 表示不饮酒。这样， x1和x2的交叉水平有4 个，建立4 个哑变量分别代表这4 个水平，记为x11、x10 、x01、x00，它们表示4 种不同的生活方式，即x11表示既吸烟又饮酒， X10 表示吸烟但不饮酒， x01 表示不吸烟但饮酒， x00表示既不吸烟又不饮酒。将前3 个哑变量放进模型， 则可得到前3 种生活方式相对于最后一种生活方式患食管癌的相对危险度。

```
> Example11_4$x11  <- ifelse (x1==1 & x2==1, 1, 0)
> Example11_4$x10  <- ifelse (x1==1 & x2==0, 1, 0)
> Example11_4$x01  <- ifelse (x1==0 & x2==1, 1, 0)
> Example11_4$x00  <- ifelse (x1==0 & x2==0, 1, 0)

> fit3 <- glm(y~ x11 + x10 + x01, family= binomial(), data=Example11_4)
> summary(fit3)
> coefficients(fit3)
> exp(coefficients(fit3))
> exp(confint(fit3))
> detach (Example11_4)
```

Logistic回归案例 3

- 例11-5 为了探讨冠心病发生的有关危险因素， 对26例冠心病病人和28例对照者进行病例-对照研究，各因素的说明及资料见表11-6 和表11-7 。试用Logistic **逐步**回归分析方法筛选危险因素。

表 11-6　冠心病 8 个可能的危险因素与赋值

因　素	变 量 名	赋值说明
年龄（岁）	X_1	<45=1，45~54=2，55~64=3，65~=4
高血压史	X_2	无=0，有=1
高血压家族史	X_3	无=0，有=1
吸烟	X_4	不吸=0，吸=1
高血脂史	X_5	无=0，有=1
动物脂肪摄入	X_6	低=0，高=1
体重指数	X_7	<24=1，24~=2，26~=3
A 型性格	X_8	否=0，是=1
冠心病	Y	对照=0，病例=1

Logistic回归案例 3

表 11-7 冠心病危险因素的病例对照研究资料

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	Y
3	1	0	1	0	0	1	1	0
2	0	1	1	0	0	1	0	0
2	1	0	1	0	0	1	0	0
2	0	0	1	0	0	1	0	0
3	0	0	1	0	1	1	1	0
3	0	1	1	0	0	2	1	0
2	0	1	0	0	0	1	0	0
3	0	1	1	1	0	1	0	0
2	0	0	0	0	0	1	1	0
1	0	0	1	0	0	1	0	0
1	0	1	0	0	0	1	1	0
1	0	0	0	0	0	2	1	0
2	0	0	0	0	0	1	0	0
4	1	0	1	0	0	1	0	0
3	0	1	1	0	0	1	1	0
1	0	0	1	0	0	3	1	0

……其余观测未列出

Logistic回归案例 3 代码


```
> Example11_5 <- read.table ("example11_5.csv", header=TRUE, sep=",")
> attach(Example11_5)
> fullfit <- glm(y~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 , family= binomial(),
data=Example11_5)
> summary(fullfit)
> nothing <- glm(y~ 1 , family= binomial(), data=Example11_5)
> summary(nothing)
> bothways <- step(nothing, list(lower=formula(nothing), upper=formula(fullfit)),
direction="both")
> fit1 <- glm(y~ x6 + x5 + x8 + x1 + x2 , family= binomial(), data=Example11_5)
> summary(fit1)
> fit2 <- glm(y~ x6 + x5 + x8 + x1 , family= binomial(), data=Example11_5)
> summary(fit2)
> coefficients(fit2)
> exp(coefficients(fit2))
> exp (confint(fit2))
> detach (Example11_5)
```

条件Logistic回归案例 4

- 例11-6 某研究机构为了研究胃癌与饮酒的相关关系，收集了病例对照资料如表11-9 所示，其中D 和D' 分别表示患有胃癌和未患有胃癌， E 和E' 分别表示饮酒和不饮酒。试用条件Logistic回归模型分析饮酒对胃癌的影响。

表 11-9　　饮酒与胃癌的病例对照资料

D（y=1）	D'（y=0）	
	E（x=1）	E'（x=0）
E（x=1）	3	14
E'（x=0）	5	d

条件Logistic回归案例 4 代码


```
> install.packages("survival")
> library(survival)
> Example11_6  <- read.table ("example11_6.csv", header=TRUE, sep=",")
> attach(Example11_6)
> model  <-  clogit(outcome~ exposure+ strata(id))
> detach(Example11_6)
```

R中扩展的Logistic回归及其变种

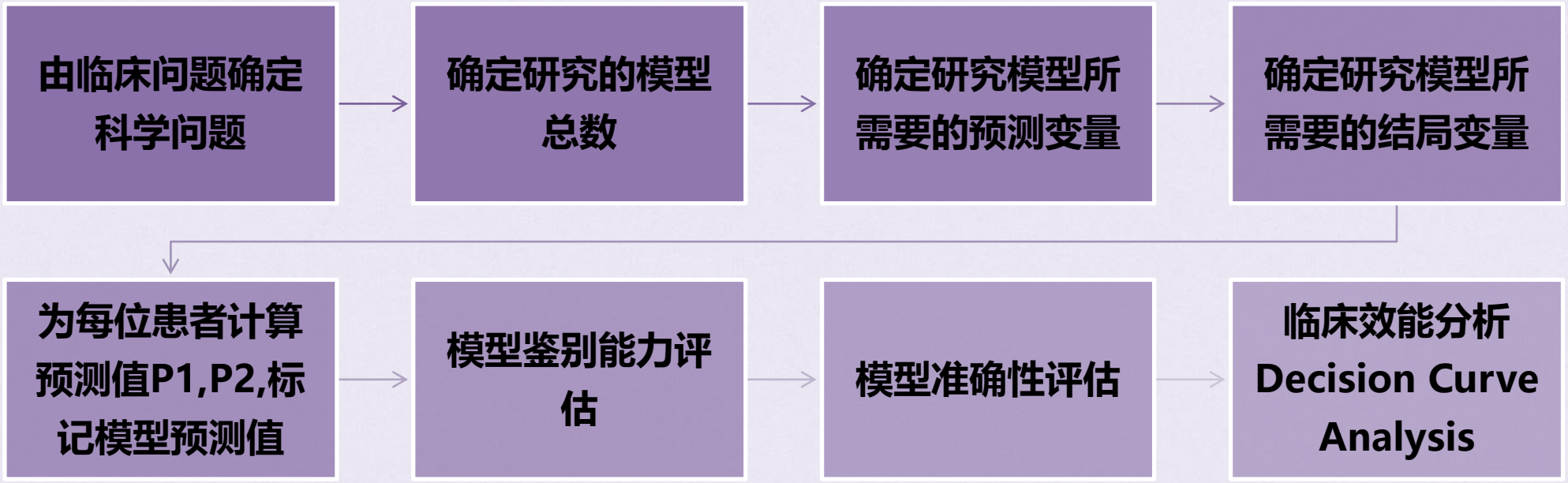
- 稳健Logistic回归 robust包中的glmRob()函数可用来拟合稳健的广义线性模型，包括稳健Logistic回归。当拟合Logistic回归模型数据出现离群点和强影响点时，稳健Logistic回归便可派上用场。
- 多项分布回归 若响应变量包含两个以上的无序类别（比如，已婚/寡居/离婚），便可使用mlogit包中的mlogit()函数拟合多项Logistic回归。
- 序数Logistic回归 若响应变量是一组有序的类别（比如，信用风险为差/良/好），便可使用rms包中的lrm()函数拟合序数Logistic回归。

7.3 Logistic回归模型 -- 列线图绘制

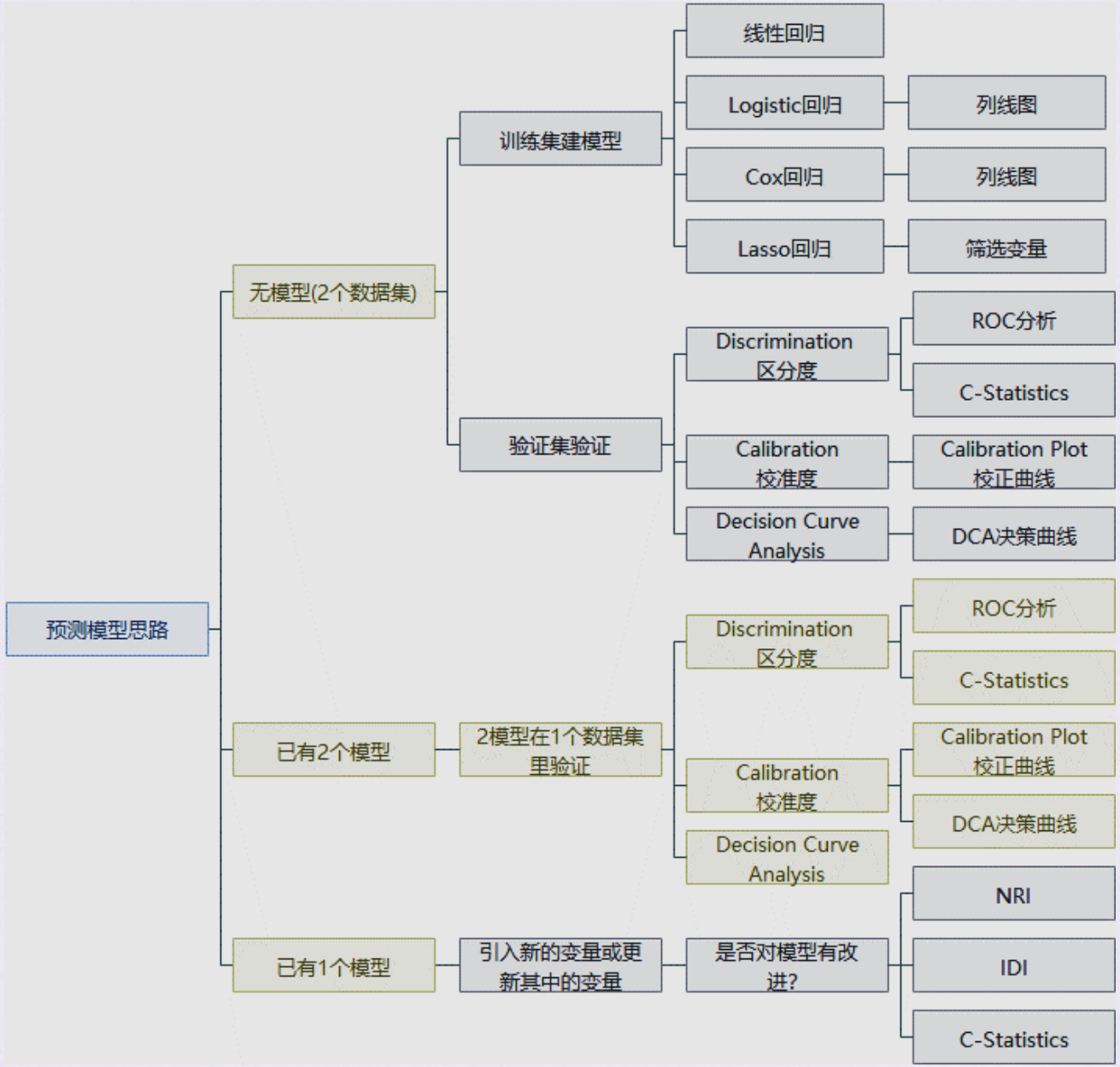
临床预测模型的本质

- 科研预测模型是通过已知来预测未知，而模型就是一个复杂的公式。也就是把已知的东西通过这个模型的计算来预测未知的东西。
- 临床预测模型的本质就是通过回归建模分析，回归的本质就是发现规律。回归是量化刻画，X多大程度上影响Y。尤其是多元线性、Logistic、Cox回归分析等。
- 模型的验证也体现着较高技术难度。模型效能评价是统计分析、数据建模、课题设计的关键所在。

临床预测模型建立 -- 验证步骤



临床预测模型类研究思路



临床预测模型类研究思路【举例】

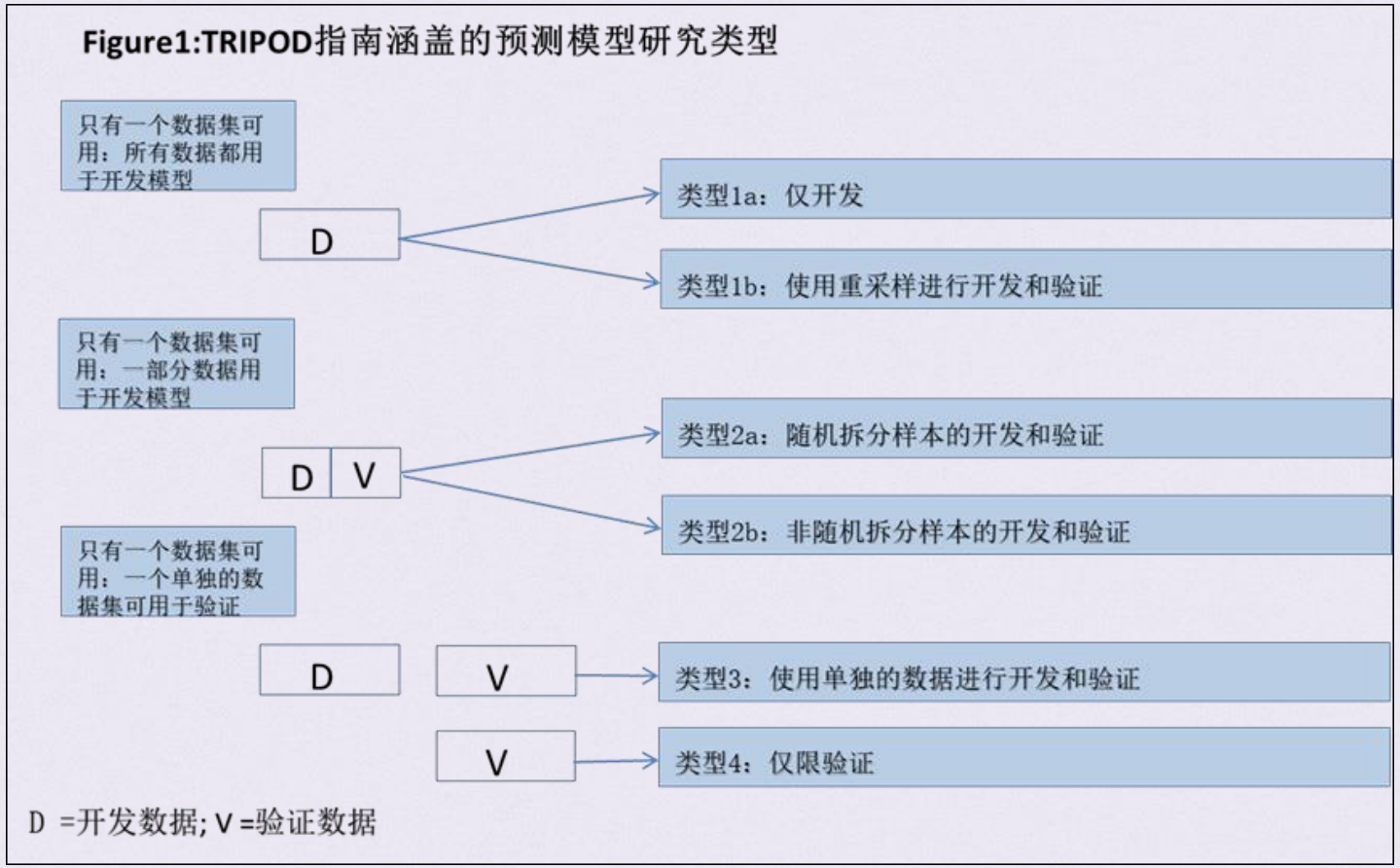
- 临床上有多个心血管疾病风险预测工具：Framingham、QRISK、PROCAM、ASSIGN评分。Heart发表综述《Graphics and Statistics for Cardiology: Clinical Prediction Rules》以心血管风险评分(CVD risk factor)为例探讨如何借助图形优势构建疾病的预测模型，并提出了6个重要步骤。

1. 选择一组预测变量作为潜在CVD影响因素纳入到风险评分中
2. 选择一个合适的统计模型来分析预测变量和CVD之间的关系
3. 从已有的预测变量中，选择足够重要的变量纳入到风险评分中
4. 构造风险评分模型
5. 评价风险评分模型
6. 在临床实践中解释风险评分的使用。

$$\ln\left(\frac{P}{1-P}\right)=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p=\beta_0+\sum_{j=1}^{p}\beta_jx_j$$

TRIPOD声明: 临床预测模型分类

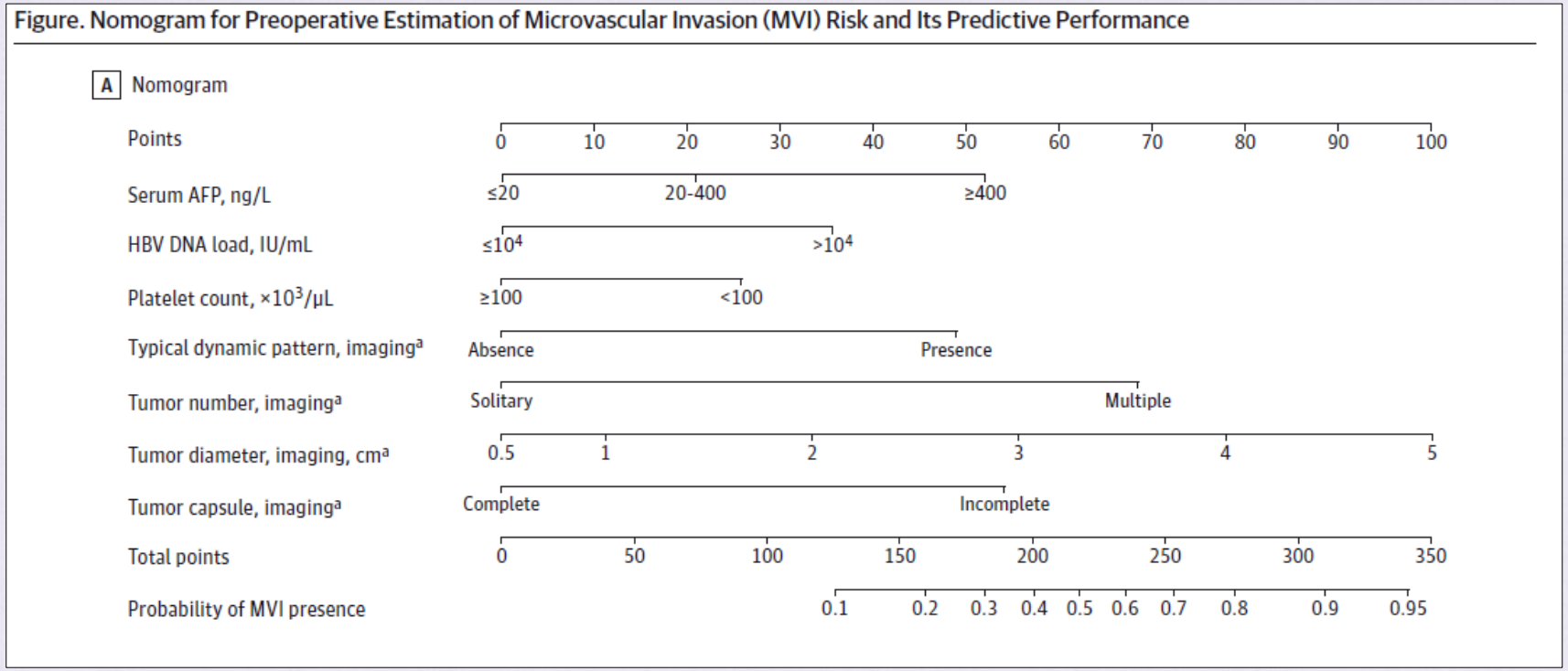
- 1、既往没有模型：为了解决一个全新的问题，展示自己好也就够了
- 2、既往已有模型：评价这两个模型的优劣
- 3、改进现有的模型：增加其他预测指标或更新现有指标



www.tripod statement.org

列线图案例解读

- Nomogram for Preoperative Estimation of Microvascular Invasion Risk in Hepatitis B Virus–Related Hepatocellular Carcinoma Within the Milan Criteria. JAMA Surgery, 2015. SCI IF=7.9



列线图案例解读

Table 1. Participant Characteristics

- SI conversion factors: To convert albumin to grams per liter, multiply by 10; α-fetoprotein tomicrograms per milliliter, multiply by 1; ALT and GGT to microkatals per liter, multiply by 0.0167; creatinine tomicromoles per liter, multiply by 88.4; glucose to millimoles per liter, multiply by 0.0555; platelets to X 10^9/L, multiply by 1; red blood cells to X 10^{12}/L, multiply by 1; total bilirubin to micromoles per liter, multiply by 17.104; white blood cells to X 10^9/L , multiply by 0.001.
- Tumor boundary on imaging was categorized as (1) smooth, presenting as a nodular-shaped tumor on all axial, coronary, and sagittal imaging or (2) not smooth, presenting as single nodule with no clear boundary.

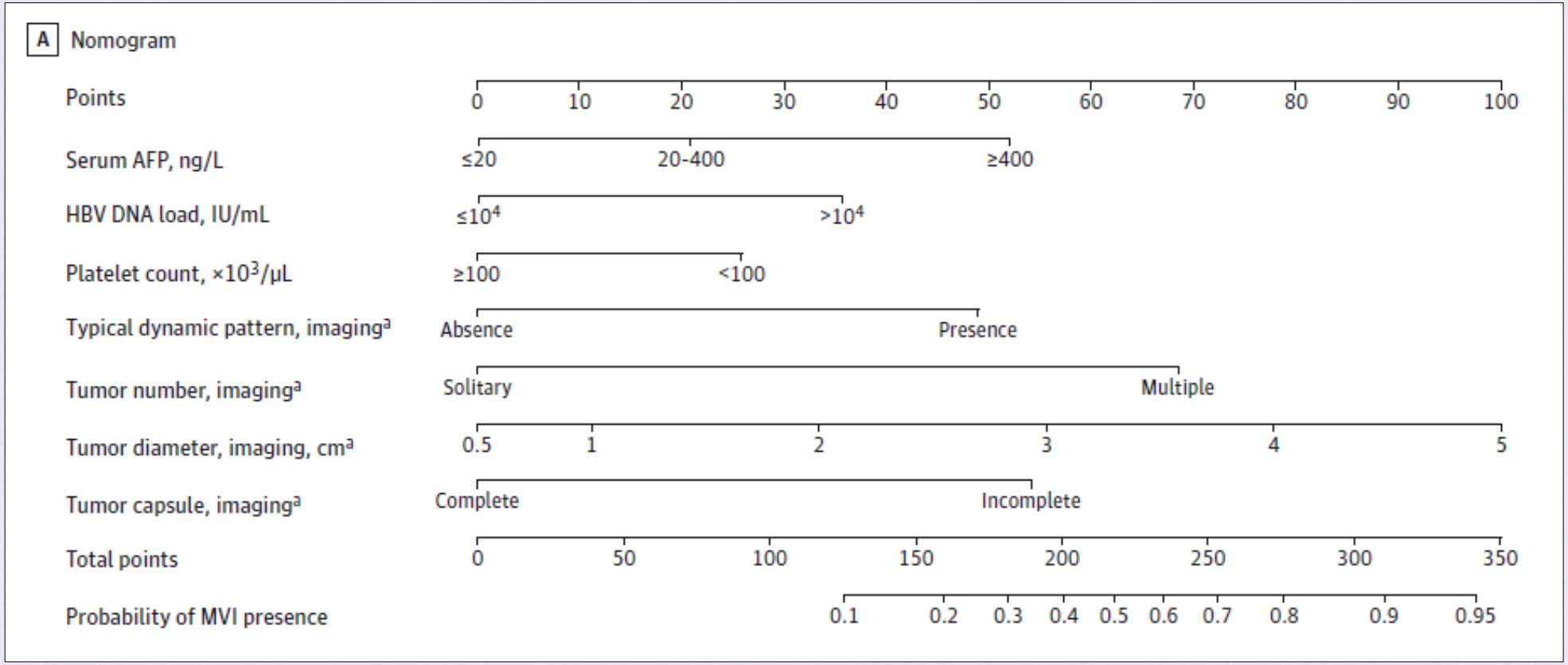
列线图案例解读

Table 2. Univariate Logistic Regression Analysis of MVI Presence Based on Preoperative Data in the Training Cohort

列线图案例解读

Table 3. Multivariate Logistic Regression Analysis of MVI Presence Based on Preoperative Data in the Training Cohort

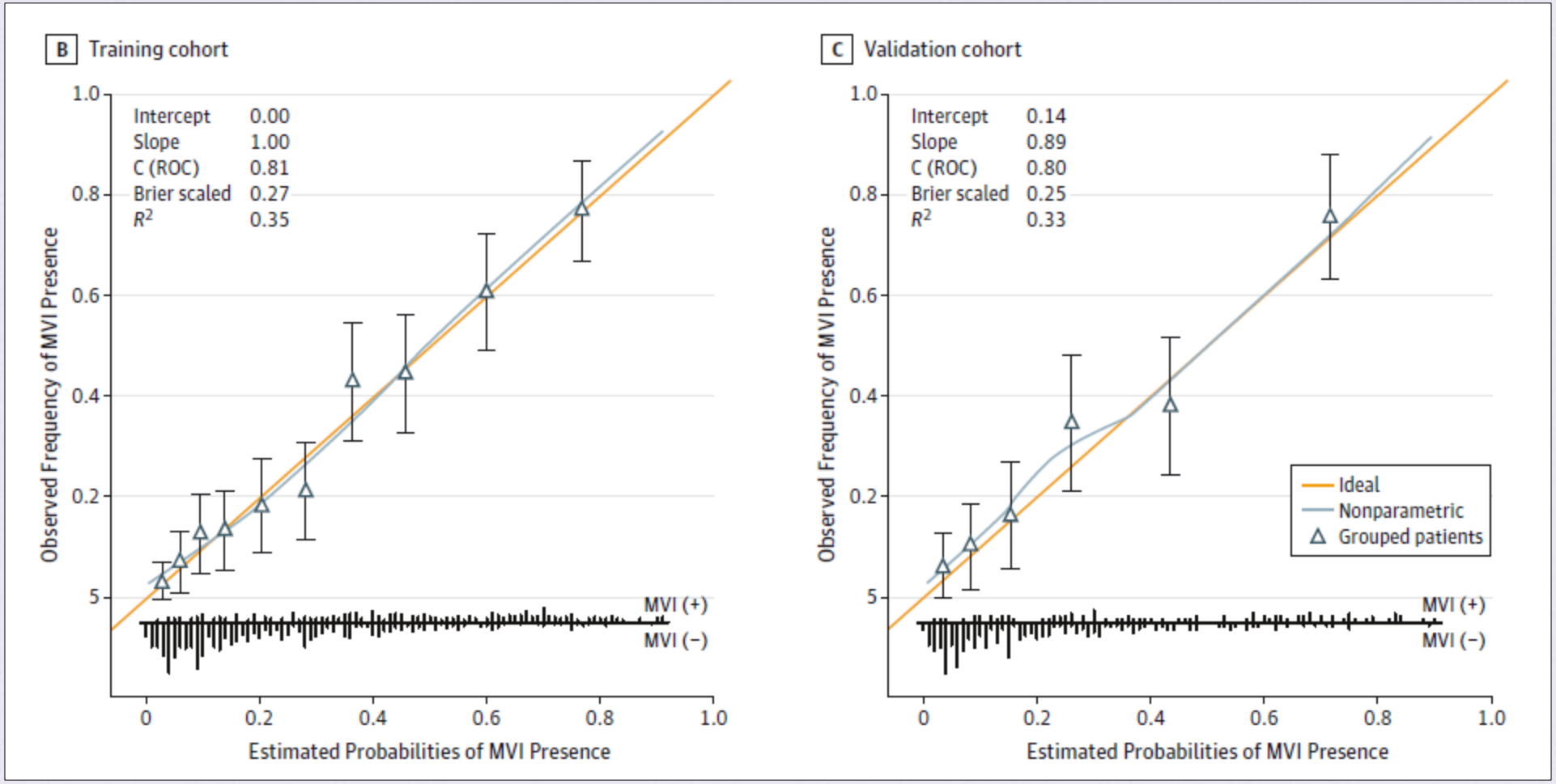
列线图案例解读


A Nomogram
Points
0 10 20 30 40 50 60 70 80 90 100
Serum AFP, ng/L
≤20 20-400 ≥400
HBV DNA load, IU/mL
≤10^4 >10^4
Platelet count, ×10^3/μL
≥100 <100
Typical dynamic pattern, imaging[a]
Absence Presence
Tumor number, imaging[a]
Solitary Multiple
Tumor diameter, imaging, cm[a]
0.5 1 2 3 4 5
Tumor capsule, imaging[a]
Complete Incomplete
Total points
0 50 100 150 200 250 300 350
Probability of MVI presence
0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.95

列线图案例解读

- **A**, Nomogram to estimate the risk of MVI presence preoperatively in hepatitis B virus (HBV)–related hepatocellular carcinoma within theMilan criteria. To use the nomogram, find the position of each variable on the corresponding axis, draw a line to the points axis for the number of points, add the points from all of the variables, and draw a line from the total points axis to determine the MVI probabilities at the lower line of the nomogram.

列线图案例解读


B Training cohort
Intercept 0.00
Slope 1.00
C (ROC) 0.81
Brier scaled 0.27
R^2 0.35
Observed Frequency of MVI Presence
Estimated Probabilities of MVI Presence
MVI (+)
MVI (−)
C Validation cohort
Intercept 0.14
Slope 0.89
C (ROC) 0.80
Brier scaled 0.25
R^2 0.33
Ideal
Nonparametric
Grouped patients

列线图案例解读

- **B**, Validity of the predictive performance of the nomogram in estimating the risk of MVI presence in the training cohort (n = 707).
- **C**, Validity of the predictive performance of the nomogram in estimating the risk of MVI presence in the validation cohort (n = 297).
- The distribution of the predicted probabilities of MVI presence is shown at the bottom of the graphs, separating those with (+) and without (−)MVI. The triangles indicate the observed frequencies of MVI presence by the deciles of the predicted probability. AFP indicates α-fetoprotein; C index, concordance index; and ROC, receiver operating characteristic.

Table 4. Accuracy of the Prediction Score of the Nomogram for Estimating the Risk of MVI Presence

Variable	Value (95% CI)	
	Training Cohort	Validation Cohort
Area under ROC curve, concordance index	0.81 (0.78-0.85)	0.80 (0.75-0.86)
Cutoff score	200	200
Sensitivity, %	73.5 (67.0-79.3)	61.8 (50.9-71.9)
Specificity, %	76.6 (72.6-80.3)	80.8 (74.7-85.9)
Positive predictive value, %	57.2 (52.0-64.9)	57.9 (49.2-68.5)
Negative predictive value, %	87.2 (83.2-89.4)	83.2 (76.0-87.7)
Positive likelihood ratio	3.1 (2.6-3.8)	3.2 (2.3-4.4)
Negative likelihood ratio	0.35 (0.28-0.44)	0.47 (0.36-0.62)

Abbreviations: MVI, microvascular invasion; ROC, receiver operating characteristic.

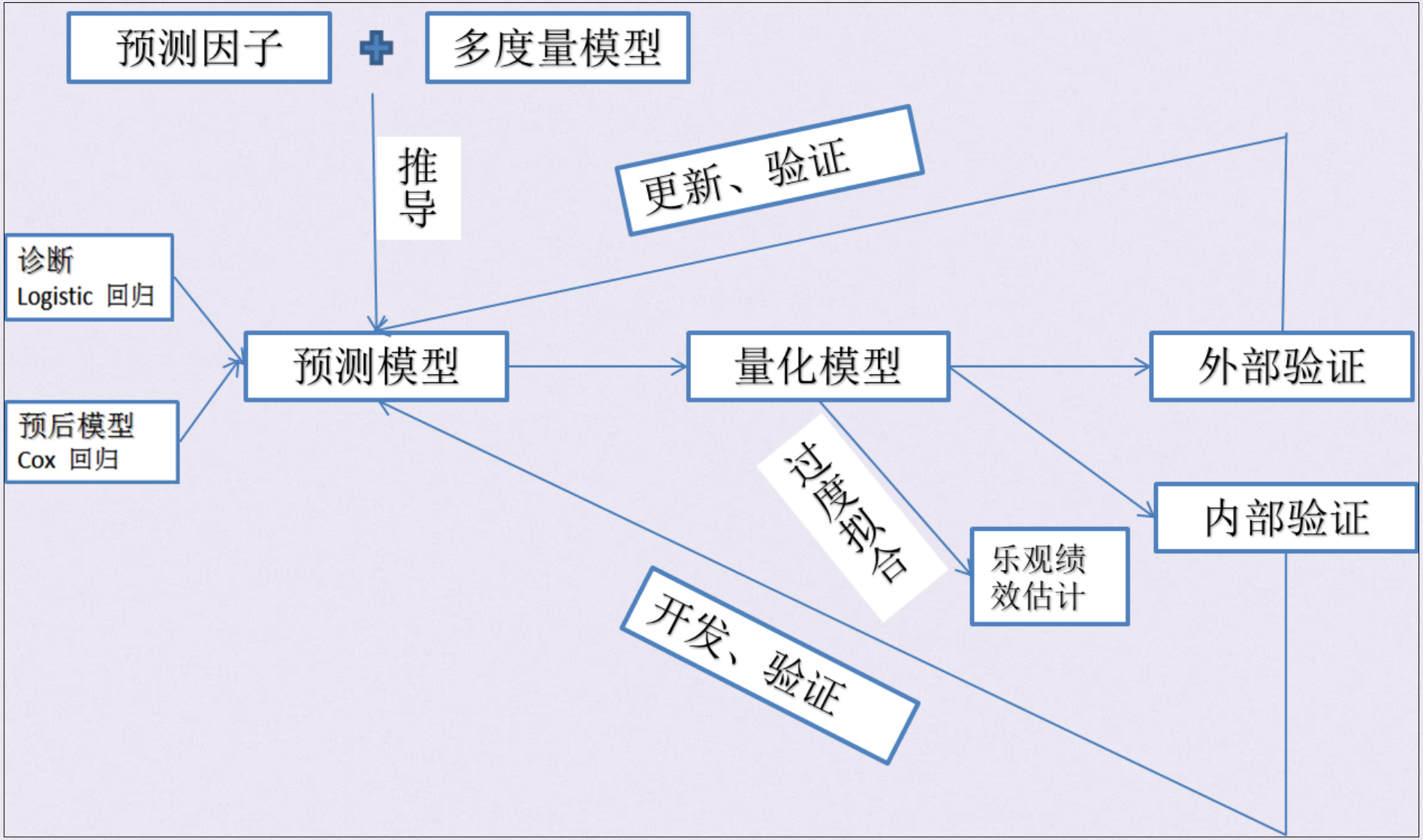
列线图案例解读 统计方法

- The significance of each variable in the training cohort was assessed by univariate logistic regression analysis for investigating the independent risk factors of presence of MVI. All variables associated with MVI at a significant level were candidates for stepwise multivariate analysis.
- A nomogram was formulated based on the results of multivariate logistic regression analysis and by using the rms package of R, version 3.0 (http://www.r-project.org/). The nomogram is based on proportionally converting each regression coefficient in multivariate logistic regression to a 0- to 100-point scale. The effect of the variable with the highest β coefficient (absolute value) is assigned 100 points. The points are added across independent variables to derive total points, which are converted to predicted probabilities.

列线图案例解读 统计方法

- The predictive performance of the nomogram was measured by concordance index (C index) and calibration with 1000 bootstrap samples to decrease the overfit bias.
- For clinical use of the model, the total scores of each patient were calculated based on the nomogram. Receiver operating characteristic curve analysis was used to calculate the optimal cutoff values thatwere determined by maximizing the Youden index (ie, sensitivity + specificity − 1).Accuracy of the optimal cutoff value was assessed by the sensitivity, specificity, predictive values, and likelihood ratios.

预测类案例研究思路总结：



Logistic回归案例 5 列线图&校正曲线绘制

- 【案例】Hosmer 和 Lemeshow于1989年研究了低出生体重婴儿的影响因素。结果变量为是否娩出低出生体重儿（变量名为LOW，1=低出生体重，即婴儿出生体重<2500g; 0=非低出生体重），考虑的影响因素（自变量）有：产妇妊娠前体重（lwt，磅）；产妇年龄（age，岁）；产妇在妊娠期间是否吸烟（smoke，0＝未吸、1＝吸烟）；本次妊娠前早产次数（ptl，次）；是否患有高血压（ht，0＝未患、1＝患病）；子宫对按摩、催产素等刺激引起收缩的应激性（ui，0＝无、1＝有）；妊娠前三个月社区医生随访次数（ftv，次）；种族（race，1＝白人、2＝黑人、3＝其他民族）。

Logistic回归案例 5 列线图&校正曲线绘制 代码


```
> library(foreign)
> library(rms)
> mydata<-read.spss("lweight.sav")
> mydata<-as.data.frame(mydata)
> head(mydata)
> mydata$low <- ifelse(mydata$low =="低出生体重",1,0)
> mydata$race1 <- ifelse(mydata$race =="白种人",1,0)
> mydata$race2 <- ifelse(mydata$race =="黑种人",1,0)
> mydata$race3 <- ifelse(mydata$race =="其他种族",1,0)
> attach(mydata)
> dd<-datadist(mydata)
> options(datadist='dd’ )

> fit1<-lrm(low~age+ftv+ht+lwt+ptl+smoke+ui+race1+race2,data=mydata,x=T,y=T)
> fit1
> summary(fit1)
> nom1 <- nomogram(fit1, fun=plogis,fun.at=c(.001, .01, .05, seq(.1,.9, by=.1), .95, .99, .999),
lp=F, funlabel="Low weight rate")
> plot(nom1)
> cal1 <- calibrate(fit1, method='boot', B=1000)
> plot(cal1,xlim=c(0,1.0),ylim=c(0,1.0))
```

Logistic回归案例 5 列线图&校正曲线绘制 代码 续


```
> mydata$race <- as.factor(ifelse(mydata$race=="白种人", "白种人","黑人及其他种族"))
> dd<-datadist(mydata)
> options(datadist='dd’ )
> fit2<-lrm(low~age+ftv+ht+lwt+ptl+smoke+ui+race,data=mydata,x=T,y=T)
> fit2
> summary(fit2)
> nom2 <- nomogram(fit2, fun=plogis, fun.at=c(.001, .01, .05, seq(.1,.9, by=.1), .95, .99, .999),
lp=F, funlabel="Low weight rate")
> plot(nom2)
> cal2 <- calibrate(fit2, method='boot', B=1000)
> plot(cal2,xlim=c(0,1.0),ylim=c(0,1.0))

> fit3<-lrm(low~ht+lwt+ptl+smoke+race, data=mydata, x=T, y=T)
> fit3
> summary(fit3)
> nom3 <- nomogram(fit3, fun=plogis, fun.at=c(.001, .01, .05, seq(.1,.9, by=.1), .95, .99, .999),
lp=F, funlabel="Low weight rate")
> plot(nom3)
> cal3 <- calibrate(fit3, method='boot', B=1000)
> plot(cal3,xlim=c(0,1.0),ylim=c(0,1.0))
```

7.3 Logistic回归模型 -- C-Statistics 计算

C-Statistics 如何计算?

- SPSS中如何计算C-Statistics

以案例5演示SPSS的操作:

[案例] Hosmer 和 Lemeshow于1989年研究了低出生体重婴儿的影响因素。结果变量为是否娩出低出生体重儿（变量名为LOW，1=低出生体重，即婴儿出生体重<2500g; 0=非低出生体重），考虑的影响因素（自变量）有：产妇妊娠前体重（lwt，磅）；产妇年龄（age，岁）；产妇在妊娠期间是否吸烟（smoke，0＝未吸、1＝吸烟）；本次妊娠前早产次数（ptl，次）；是否患有高血压（ht，0＝未患、1＝患病）；子宫对按摩、催产素等刺激引起收缩的应激性（ui，0＝无、1＝有）；妊娠前三个月社区医生随访次数（ftv，次）；种族（race，1＝白人、2＝黑人、3＝其他民族）。

C-Statistics 如何计算?

- R中如何计算C-Statistics

1. rms包中lrm函数拟合logistic回归模型，模型参数可直接读取C，Dxy
2. ROCR包中performance函数计算AUC
3. Hmisc包中的somers2函数直接计算C, Dxy

7.3 Logistic回归模型 -- 亚组分析森林图绘制

Subgroup	No. of Patients	Aspirin n(%)	Clopidogrel-Aspirin n(%)	Hazard Ratio (95%CI)	P value*
Overall	5170	362(14.0)	275(10.6)	0.78(0.65-0.93)	
Age					0.70
<65 yr	3029	202(13.2)	145(9.7)	0.73(0.57-0.95)	
≥65 yr	2141	160(15.2)	130(12.0)	0.75(0.57-0.98)	
Sex					0.22
Male	3420	235(13.9)	172(9.9)	0.67(0.53-0.85)	
Female	1750	127(14.1)	103(12.1)	1.04(0.75-1.44)	
Index event					0.93
Minor stroke	3725	271(14.6)	209 (11.2)	0.73(0.59-0.90)	
TIA	1445	91(12.5)	66(9.2)	0.93(0.61-1.41)	
ABCD^2 score					0.33
4	747	38(10.1)	35(9.4)	0.94(0.49-1.83)	
>4	698	53(15.1)	31(9.0)	0.70(0.35-1.38)	
Previous stroke					0.16
Yes	1033	72(14.0)	67(13.0)	0.81(0.54-1.23)	
No	4137	290(14.0)	208(10.1)	0.74(0.60-0.91)	
Previous TIA					0.43
Yes	174	14(17.5)	9(9.6)	1.27(0.33-4.85)	
No	4996	348(13.9)	266(10.7)	0.77(0.64-0.93)	
History of hypertension					0.58
Yes	3399	259(15.4)	206(12.0)	0.72(0.58-0.90)	
No	1771	103(11.4)	69(7.9)	0.82(0.57-1.18)	
Previous diabetes					0.75
Yes	1093	91(16.8)	72(13.1)	0.74(0.49-1.11)	
No	4077	271(13.3)	203(10.0)	0.81(0.66-1.00)	
Systolic pressure					0.53
<140 mm Hg	1376	76(11.0)	62(9.1)	0.98(0.62-1.54)	
≥140 mm Hg	3790	286(15.1)	212(11.2)	0.71(0.58-0.87)	
Time to randomization					0.26
<12 hr	2573	189(14.8)	157(12.1)	0.81(0.63-1.04)	
≥12 hr	2597	173(13.3)	118(9.1)	0.72(0.54-0.95)	
Aspirin taken within 24 hr					0.86
Yes	597	47(15.2)	35(12.2)	0.65(0.35-1.19)	
No	4573	315(13.8)	240(10.5)	0.78(0.64-0.94)	

Chance, Circulation, 2015, 132(1)

Logistic回归案例 6 亚组分析森林图绘制

- 化疗药物的反应率

Response	No.of patients	OR(95% CI)			
Overall	1472	1.66(1.28,2.15)	1.66	1.28	2.15
Clinical stage					
Advanced	951	1.67(1.23,2.26)	1.67	1.23	2.26
non-advanced	521	1.67(1.05,2.66)	1.67	1.05	2.66
ctDNA assays					
qPCR	1280	1.73(1.30,2.29)	1.73	1.3	2.29
ARMS	192	1.58(0.50,5.04)	1.58	0.5	5.04
Ethnicity					
Asian	326	1.93(0.84,4.43)	1.93	0.84	4.43
Caucasian	1146	1.57(1.20,2.06)	1.57	1.2	2.06

Logistic回归案例 6 亚组分析森林图绘制 代码


```
> library(forestplot)
> rs_forest <- read.csv('rs_forest.csv',header = FALSE)
# 读入数据的时候大家一定要把header设置成FALSE，保证第一行不被当作列名称。
# tiff('Figure 1.tiff',height = 1600,width = 2400,res= 300)
> forestplot(labeltext = as.matrix(rs_forest[,1:3]),
           #设置用于文本展示的列，此处我们用数据的前三列作为文本，在图中展示
           mean = rs_forest$V4, #设置均值
           lower = rs_forest$V5, #设置均值的上限
           upper = rs_forest$V6, #设置均值的下限
           is.summary = c(T,T,T,F,F,T,F,F,T,F,F),
           #该参数接受一个逻辑向量，用于定义数据中的每一行是否是汇总值，若是，则在对应位置设置
为TRUE，若否，则设置为FALSE；设置为TRUE的行则以粗体出现
           zero = 1, #设置参照值，此处我们展示的是HR值，故参照值是1，而不是0
           boxsize = 0.4, #设置点估计的方形大小
           lineheight = unit(10,'mm'),#设置图形中的行距
           colgap = unit(3,'mm'),#设置图形中的列间距
           lwd.zero = 2,#设置参考线的粗细
           lwd.ci = 1.5,#设置区间估计线的粗细
           col=fpColors(box='#458B00', summary= "#8B008B",lines = 'black',zero = '#7AC5CD'),
           #使用fpColors()函数定义图形元素的颜色，从左至右分别对应点估计方形，汇总值，区间估计
线，参考线
           xlab="The estimates",#设置x轴标签
           graph.pos = 3)#设置森林图的位置，此处设置为3，则出现在第三列
```

7.4 泊松回归模型

泊松回归案例

- 为阐述泊松回归模型的拟合过程，并探讨一些可能出现的问题，我们使用robust程辑包中的Breslow癫痫数据集（Breslow, 1993）。特别地，我们将讨论在治疗初期的8周内，抗癫痫药物对癫痫发病数的影响。请提前安装robust包。
- 我们就遭受轻微或严重间歇性癫痫的病人的年龄和癫痫发病数收集了数据，包含病人被随机分配到药物组或者安慰剂组前8周和随机分配后8周两种情况。响应变量为sumY（随机化后8周内癫痫发病数），预测变量为治疗条件（Trt）、年龄（Age）和前8周内的基础癫痫发病数

泊松回归案例 代码


```
# look at dataset
> data(breslow.dat, package="robust")
> names(breslow.dat)
> summary(breslow.dat[c(6, 7, 8, 10)])

# plot distribution of post-treatment seizure counts
> opar <- par(no.readonly=TRUE)
> par(mfrow=c(1, 2))
> attach(breslow.dat)
> hist(sumY, breaks=20, xlab="Seizure Count",
    main="Distribution of Seizures")
> boxplot(sumY ~ Trt, xlab="Treatment", main="Group Comparisons")
> par(opar)

# fit regression
> fit <- glm(sumY ~ Base + Age + Trt, data=breslow.dat, family=poisson())
> summary(fit)
```

泊松回归案例 代码 续


```
# interpret model parameters
> coef(fit)
> exp(coef(fit))

# evaluate overdispersion 过度离势检验
> deviance(fit)/df.residual(fit)
> library(qcc)
> qcc.overdispersion.test(breslow.dat$sumY, type="poisson")

# fit model with quasipoisson
> fit.od <- glm(sumY ~ Base + Age + Trt, data=breslow.dat,
                family=quasipoisson())
> summary(fit.od)
```

泊松回归结果解释

- 在泊松回归中，年龄的回归参数为0.0227，表明保持其他预测变量不变，年龄增加1岁，癫痫发病数的对数均值将相应增加0.03。截距项即当预测变量都为0时，癫痫发病数的对数均值。由于不可能为0岁，且调查对象的基础癫痫发病数均不为0，因此本例中截距项没有意义。通常在因变量的初始尺度（癫痫发病数而非发病数的对数）上解释回归系数比较容易。为此，需要指数化回归系数。
- 指数化后，保持其他变量不变，年龄增加1岁，期望的癫痫发病数将乘以1.023。这意味着年龄的增加与较高的癫痫发病数相关联。更为重要的是，1单位Trt的变化（即从安慰剂到治疗组），期望的癫痫发病数将乘以0.86，也就是说，保持基础癫痫发病数和年龄不变，服药组相对于安慰剂组癫痫发病数降低了20%。

主要参考文献

【1】Robert I. Kabacoff 著, 《R语言实战 》(第2版), 人民邮电出版社, 2016

【2】Peter Dalgaard 著, 《R语言统计入门》》(第2版), 人民邮电出版社, 2014

【3】薛毅 陈立萍 著,《R语言实用教程》, 清华大学出版社, 2014

【4】张铁军 陈兴栋 刘振球 著,《R语言与医学统计图形》, 人民卫生出版社, 2018

【5】汪海波 萝莉 汪海玲 著, 《R语言统计分析与应用》， 人民邮电出版社，2018

Thanks!

感谢您的观看！