第十三章 logistic 回归分析

第一节 Logistic 回归：从建模到检验的一整套体系

一句话总览

Logistic 回归四步走：

① 建模（logit 线性） → ② 估计（MLE） → ③ 检验（LR/Wald + GOF） → ④ 筛选（前进/后退/逐步）

输出核心只看三样：**OR、CI、P**（再加模型拟合）。

一、模型回顾：你到底在拟合什么？（Model）

1 二分类结局与概率

$$Y \in \{0,1\}, P = P(Y = 1 \mid X)$$

2 Logistic 形式（概率尺度）

$$P = \frac{1}{1 + \exp\left[-(\beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m)\right]}$$

3 Logit 线性形式（回归真正做线性的地方）

$$\text{logit}\,(P) = \ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m$$

📌 记住一句话：

线性回归线性在 Y，Logistic 回归线性在 log-odds。

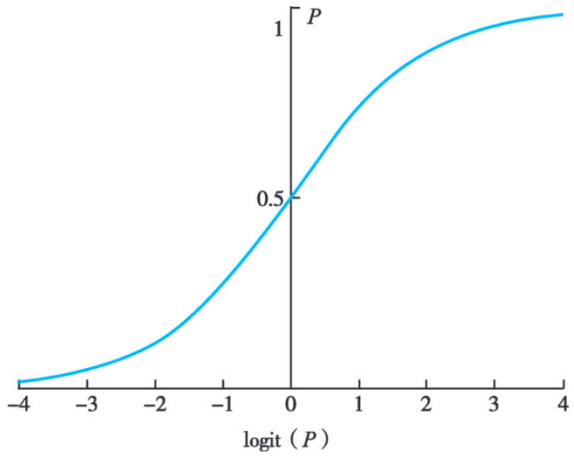

图 13-1　logistic 曲线示意图

(李康,贺佳主编, 2024, p. 132) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

二、参数估计：为什么必须用最大似然（MLE）？

1 似然函数（0/1 数据的“乘积结构”）

对第 i 个观测：

- 若 $Y_i = 1$：贡献 P_i
- 若 $Y_i = 0$：贡献 $1 - P_i$

统一写成：

$$L = \prod_{i=1}^{n} P_i^{Y_i}(1 - P_i)^{1-Y_i}$$

2 为什么没有解析解？（必须迭代）

因为 P_i 是 β 的非线性函数，导致似然对 β 非线性，解不出来，只能用迭代数值方法（Newton–Raphson / Fisher scoring 等）求 MLE：

$$(b_0, b_1, \ldots, b_m)$$

三、从系数到 OR：参数的“可解释版本”

1 OR 与回归系数的对应

当 X_j 从 c_0 变到 c_1（其他变量不变）：

$$\widehat{OR}_j = \exp\left[b_j(c_1 - c_0)\right]$$

最常见的二分类暴露：$c_1 = 1, c_0 = 0$

$$\widehat{OR}_j = \exp\left(b_j\right)$$

2 OR 的置信区间（大样本正态近似）

大样本下 b_j 近似正态，因此：

$$\exp\left(b_j - z_{\alpha/2} S_{b_j}\right)$$

📌 最快判读法：

> **CI 不跨 1 ⇒ 显著；CI 跨 1 ⇒ 不显著。**

四、单因素 vs 多因素 Logistic：它们在“意义”上的差异

1 单因素 Logistic（univariable）

- 结论上常与 χ^2 检验一致：**OR 估计值相同**（本质都是 2×2 的 odds 比）
- 但统计量可能不同：
  - Pearson χ^2（列联表）
  - Wald χ^2（Logistic 输出）
    公式不同，所以数值不同，但**结论通常一致**

📌 一句话定位：

> χ^2 只告诉你“差异是否存在”，Logistic 还能告诉你“影响有多大（OR）”。

2️⃣ 多因素 Logistic（multivariable）

- 关键价值：**控制混杂（adjustment）**
- 每个 b_j / OR 的含义变成：

> **在控制其他自变量后，X_j 对结局 odds 的独立影响。**

五、影响大小能不能比？——标准化回归系数（提醒点）

教材的提醒很关键：因为自变量单位不同，不能直接用 b_j 比“谁影响最大”。
于是引入“标准化系数”（standardized coefficient）的思路：

> 把 b_j 按 X_j 的尺度进行标准化后比较
>
> 其绝对值，$|b_j'|$ 越大作用越大。

⚠️ 但务实提醒（你做题/写报告要心里有数）：

- Logistic 的核心解释量通常是 **OR（每单位变化）** 或 **按临床有意义尺度变换后的 OR**（例如每 10 岁、每 5 kg/m²）
- 标准化系数更像“比较强弱的辅助指标”，不要替代 OR 的解释主线

六、假设检验体系：模型检验 + 系数检验 + 拟合优度（最关键一节）

> **一句话总览**
> Logistic 的检验分三件事：
> **① 新变量加进来有没有贡献（LR）**
> **② 单个系数是否为 0（Wald）**
> **③ 拟合得像不像数据（GOF）**

A. 回归模型的假设检验：似然比检验（Likelihood Ratio Test, LR）

1️⃣ 核心思想

比较两个模型的对数似然（log-likelihood）：

- 旧模型（不含待检变量）：$\ln L_0$
- 新模型（加入变量后）：$\ln L_1$

2️⃣ 统计量（教材给的 G）

$$G = 2(\ln\ L_1 - \ln\ L_0)$$

当样本量大、在 H_0 下：

$$G \sim \chi^2_\nu, \nu = p - l$$

其中 p,l 是新旧模型自变量个数（差几个变量，自由度就是几）。

📌 结论语言：

- 若 $G \geq \chi^2_{\alpha,\nu}$：新增变量（或变量组）对模型有显著贡献

✅ 优点（你记住就够了）：

> **LR 能比较“模型与模型”，既能检一个变量，也能检一组变量。**

B. 回归系数的假设检验：Wald 检验（Wald test）

检验：

$$H_0: \beta_j = 0$$

统计量：

$$z = \frac{b_j}{S_{b_j}} 或 \chi^2 = (\frac{b_j}{S_{b_j}})^2 \sim \chi^2_1$$

📌 用法：

- 最常见的软件输出就是 Wald χ^2 / P 值

- 适合“单个系数”的显著性判断

⚠️ 教材也提醒了一个考试点：

> Wald 检验对单变量方便，但可能略保守；大样本时 Wald 与 LR 结果通常一致。

C. 拟合优度检验（Goodness-of-fit, GOF）

目的不是检“有没有效应”，而是检：

> **模型预测的频数/概率，和真实观察是否一致?**

常见：Hosmer–Lemeshow（H–L）、Pearson、Deviance（偏差）等，教材以 H–L 为例：

- 本质：把样本按预测概率分组，比较“观察频数 vs 理论频数”的差异（基于 χ^2）

📌 判读（很反直觉但必须牢记）：

> **P 值大（不显著） ⇒ 拟合较好**
> **P 值小（显著） ⇒ 拟合较差**

（因为这里的 H_0 是“拟合得不错”。）

七、变量筛选：Logistic 的逐步回归和线性回归有什么不同?

1️⃣ 目标不变

- 保留有贡献变量
- 剔除无贡献变量
- 得到更简洁可解释模型

2️⃣ 不同点在“检验统计量”

线性回归常用 **F**;
Logistic 回归筛选时常用:

- **LR 统计量（推荐，能检变量组）**
- **Wald 统计量（软件常用）**

3️⃣ 三种方法仍然是那三种

- forward selection
- backward selection
- stepwise selection

📌 但你要记住一条“总原则”：

> **不要迷信逐步筛选得到的“最优模型”。**
> 阈值、顺序、共线性都会让结果飘；最终必须回到专业合理性与可解释性。

八、整套“考试级流程图”（你脑内应该是这样跑的）

① 拟合模型（MLE）
输出 b_j, SE, OR, CI

② 模型整体：LR（模型比较/变量组贡献）
新变量加进来值不值?

③ 单个变量：Wald（系数是否 0）
哪些变量真正显著?

④ 拟合优度：H–L / Pearson / Deviance
模型像不像数据?

⑤ 若要筛选：forward/backward/stepwise
统计筛选 + 专业判断双保险

九、综合示例｜例 13-1：慢性心力衰竭患者不良预后的 Logistic 回归分析

一、研究目的与资料概况

以某医院心内科收治的慢性心力衰竭患者为研究对象，探讨影响患者不良预后（心源性再住院或死亡）的相关因素。
结局变量为二分类变量：

$$Y=\begin{cases}1, & \text{发生不良结局}\\ 0, & \text{未发生不良结局}\end{cases}$$

候选自变量包括人口学特征（性别、年龄）、临床分级（纽约分级）、既往病史（高血压、家族史）、体格指标（BMI 分级）及多项实验室检查指标（血糖、白细胞、白蛋白等）。

二、单因素分析：χ² 检验与单因素 Logistic 回归

（一）χ² 检验（以高血压为例）

将结局变量与高血压构成四格表，计算 Pearson χ² 值及 OR 值，用于判断两者是否存在统计学关联。

表 13-3　例 13-1 单变量数据四格表（高血压）

	$Y=1$	$Y=0$	合计
有高血压	197	307	504
无高血压	80	205	285
合计	277	512	789

$\text{Pearson}\,\chi^2=9.699,\ OR=ad/bc=1.644$

(李康,贺佳主编, 2024, p. 135) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

结果显示，高血压组与无高血压组在不良结局发生率上存在显著差异，Pearson χ² 检验有统计学意义，同时可计算得到 OR 值用于描述效应大小。

（二）单因素 Logistic 回归分析

以不良预后为因变量，以高血压为自变量建立单因素 Logistic 回归模型，估计回归系数、Wald χ²、P 值及 OR 值及其 95% 置信区间。

表 13-4　例 13-1 单变量 logistic 回归结果（高血压）

因素 X	回归系数 b	标准误 S_b	Wald χ^2	P 值	OR 值	OR 值 95% 置信区间	
						下限	上限
常数项	−0.941	0.132	50.952	＜0.001			
高血压	0.497	0.160	9.620	0.002	1.644	1.201	2.252

(李康,贺佳主编, 2024, p. 135) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

分析结果表明，高血压的 OR 值与四格表法计算结果一致，提示在二分类自变量情况下，**单因素 Logistic 回归与 χ² 检验在 OR 估计上等价。**
但两者所用统计量不同：χ² 检验基于频数差异，而 Logistic 回归基于 Wald χ²，对回归系数进行显著性检验。

三、多因素 Logistic 回归模型分析

（一）模型建立

将所有具有临床意义或在单因素分析中具有潜在关联的变量同时纳入模型，采用 SPSS 软件进行多因素 Logistic 回归分析，估计各变量在控制其他因素后的独立效应。

表 13-5　例 13-1 logistic 回归参数估计结果

因素 X	回归系数 b	标准误 S_b	Wald χ^2	P 值	OR 值	OR 值 95% 置信区间		标准化回归系数
						下限	上限	b_j'
常数项	−3.633	0.586	38.421	＜0.001				
性别	−0.022	0.175	0.016	0.899	0.978	0.694	1.379	−0.006
年龄	0.029	0.008	12.604	＜0.001	1.030	1.013	1.047	0.174
BMI			1.721	0.632				
BMI（1）	0.149	0.355	0.176	0.675	1.161	0.579	2.328	0.095
BMI（2）	−0.102	0.187	0.295	0.587	0.903	0.626	1.303	−0.065
BMI（3）	−0.257	0.231	1.239	0.266	0.773	0.492	1.216	−0.164
家族史	0.017	0.221	0.006	0.940	1.017	0.659	1.568	0.004

(李康,贺佳主编, 2024, p. 135) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

续表

因素 X	回归系数 b	标准误 S_b	Wald χ^2	P 值	OR 值	OR 值 95% 置信区间		标准化回归系数
						下限	上限	b_j'
纽约分级	0.355	0.112	10.076	0.002	1.426	1.145	1.774	0.145
高血压	0.490	0.176	7.741	0.005	1.632	1.156	2.305	0.130
血脂	0.146	0.202	0.528	0.468	1.158	0.780	1.719	0.034
血糖	0.343	0.170	4.089	0.043	1.410	1.011	1.966	0.092
白细胞	0.479	0.196	5.958	0.015	1.614	1.099	2.372	0.128
红细胞	−0.095	0.207	0.210	0.647	0.909	0.606	1.365	−0.026
红细胞分布宽度	−0.283	0.195	2.096	0.148	0.754	0.514	1.105	−0.076
血红蛋白	0.035	0.207	0.029	0.864	1.036	0.691	1.553	0.009
白蛋白	0.838	0.185	20.459	＜0.001	2.311	1.608	3.323	0.222

注：BMI 正常＝0，偏低＝1，超重＝2，肥胖＝3。

(李康,贺佳主编, 2024, p. 136) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

（二）多因素分析结果解释

多因素 Logistic 回归结果显示：

- **年龄**、**纽约分级**、**高血压**、**血糖**、**白细胞计数**、**白蛋白**等变量在控制其他因素后仍与不良预后显著相关；
- 性别、BMI 分级、家族史及部分血液学指标在多因素模型中未显示统计学显著性。

各回归系数对应的 OR 值反映了在其他变量保持不变的条件下，各因素对不良结局发生 odds 的独立影响。

（三）标准化回归系数与影响强度比较

由于各自变量量纲不同，不能直接比较回归系数 b_j 的大小。
为比较不同因素对结局变量影响的相对强弱，引入标准化回归系数：

$$b_j' = \frac{b_j S_j}{\pi/\sqrt{3}}$$

其中 S_j 为自变量 X_j 的标准差。

标准化回归系数的绝对值越大，说明该变量对结局变量的相对影响越大。

从结果可见，白蛋白的标准化回归系数绝对值最大，提示其对不良预后的影响最强；家族史的影响最小。

四、自变量筛选：逐步 Logistic 回归分析

在多因素分析基础上，采用向前逐步选择法（forward selection），以似然比检验（Likelihood Ratio Test）作为变量进入与剔除模型的统计依据，设定进入水准 α 入 = 0.05，剔除水准 α 出 = 0.10。

表 13-6　例 13-1 的 logistic 回归模型自变量筛选结果

因素 X	回归系数 b	标准误 S_b	Wald χ^2	P 值	OR 值	OR 值 95% 置信区间	
						下限	上限
常数项	−4.311	0.580	55.329	＜0.001			
年龄	0.029	0.008	14.132	＜0.001	1.029	1.014	1.044
高血压	0.469	0.171	7.510	0.006	1.599	1.143	2.236
纽约分级	0.324	0.107	9.178	0.002	1.383	1.121	1.705
血糖	0.341	0.167	4.169	0.041	1.407	1.014	1.952
白细胞	0.369	0.171	4.629	0.031	1.446	1.033	2.023
白蛋白	0.740	0.173	18.360	＜0.001	2.097	1.494	2.942

(李康,贺佳主编, 2024, p. 137) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

筛选结果显示，最终进入模型的变量包括：

- 年龄
- 高血压
- 纽约分级
- 血糖
- 白细胞计数
- 白蛋白

上述变量构成了预测慢性心力衰竭患者不良预后的最终 Logistic 回归模型。

五、小结（方法论视角）

本例完整展示了 Logistic 回归分析在临床研究中的标准应用流程：

1. **单因素分析**：χ² 检验与单因素 Logistic 回归用于初步筛选和效应方向判断；
2. **多因素建模**：控制混杂因素，评估自变量的独立作用；
3. **Wald 检验**：判断单个回归系数的显著性；

4. **似然比检验**：用于变量筛选与模型比较；
5. **OR 与置信区间**：作为结果解释的核心指标。

该方法在流行病学和临床研究中具有良好的解释性和实用价值。

第二节｜条件 Logistic 回归

> **一句话总览**
> **一旦数据来自“配对 / 分层设计”，**
> **普通 Logistic 回归就不再合适，**
> **条件 Logistic 回归是“设计驱动”的必选模型。**

一、为什么需要条件 Logistic 回归？（Why）

在医学与流行病学研究中，**配对设计（matched design）**非常常见，用来在设计阶段控制混杂因素，例如：

- 病例–对照按 **年龄、性别** 匹配
- 每个病例配：
  - 1 个对照（1:1）
  - 或 M 个对照（1:M，通常 $M \leq 3$）

这样做的目的不是提高样本量，而是：

> **在设计阶段“硬性消除”强混杂因素的影响。**

⚠️ 关键后果（非常重要）

一旦使用了配对设计，就意味着：

- **组内病例与对照可比**
- **组间病例与对照不可比**

👉 因此：

> **不能把所有个体“混在一起”建模**
> 否则会直接破坏研究设计。

这正是普通（非条件）Logistic 回归**不再适用**的原因。

二、条件 Logistic 回归的适用场景（When）

只要满足以下任一情况，就应考虑条件 Logistic 回归：

1. **1:1 或 1:M 配对病例–对照研究**
2. 明确存在“匹配组 / 层（strata）”结构
3. 匹配因素本身：
   - 不需要估计效应
   - 但必须被严格控制

📌 典型例子：

> 每个病例与同年龄、同性别的对照配对
> → 年龄、性别不再进入模型
> → 它们已经被“条件化”掉了。

三、数据结构与符号体系（Structure）

设：

- 共 n 个匹配组（层）
- 每组：
  - 1 个病例（$Y = 1$）
  - M 个对照（$Y = 0$）
- X_{itj}：
  第 i 个匹配组中，第 t 个个体的第 j 个危险因素

📌 核心思想：

分析完全发生在“组内比较”层面。

四、条件 Logistic 回归模型（Model）

1 模型形式

对第 i 个匹配组：

$$P_i = \frac{1}{1 + \exp\left[-(\beta_{0i} + \beta_1 X_1 + \cdots + \beta_m X_m)\right]}, i = 1, \ldots, n$$

2 模型中最关键的一点

每个匹配组都有自己的截距 β_{0i}

这意味着：

- β_{0i} 吸收了：
  - 匹配因素（年龄、性别等）
  - 组层面的所有固定效应
- 我们 **不估计 β_{0i}**
  它们只是“技术性参数”

📌 真正关心、真正可解释的，仍然只有：

$$\boldsymbol{\beta_1, \ldots, \beta_m}$$

——组内暴露差异对应的效应。

五、条件 Logistic 的“条件”到底在哪里？（核心理解）

名字里的 **conditional** 不是修辞，而是统计含义：

模型是“在给定每个匹配组中病例数已知”的条件下，
对暴露分布进行建模。

直觉翻译：

- 每组“必然有 1 个病例”
- 问题变成：

  为什么在这一组中，
  是这个人而不是那个人成为病例？

六、参数估计与检验（与普通 Logistic 的关系）

1️⃣ 参数估计方法

- **仍然使用最大似然估计（MLE）**
- 但构造的是**条件似然（conditional likelihood）**
- 软件自动完成，无需手动区分

2️⃣ 回归系数的解释

完全一致：

- $\exp(\beta_j)$ = **组内调整后的 OR**
- 解释为：

  在同一匹配组内，
  暴露因素 X_j 对发病 odds 的影响

3️⃣ 假设检验与拟合优度

与**非条件 Logistic 回归相同**：

- 回归系数检验：
  - Wald
  - 似然比检验（LR）
- 模型拟合：
  - Hosmer–Lemeshow（或等价方法）

📌 实际操作中：

> **统计软件会直接给出结果，**
> **使用者只需选对模型类型。**

七、条件 Logistic vs 普通 Logistic（必须会对比）

方面	普通 Logistic	条件 Logistic
数据结构	个体独立	配对 / 分层
截距	一个 β_0	每层一个 β_{0i}
匹配因素	需显式建模	被条件化消除
比较方式	组间 + 组内	**严格组内比较**
适用场景	非配对数据	**配对病例–对照**

八、最容易犯的原则性错误（一定要避开）

> **错误：**
> 对配对病例–对照资料，
> 直接用普通 Logistic 回归，把匹配变量当协变量加进去。

📌 问题在于：

- 会破坏配对设计
- 会低估标准误
- OR 和 P 值都可能偏差

👉 **正确做法只有一个：**

设计是配对的，分析就必须是条件的。

九、考试级一句话总结（可直接写）

条件 Logistic 回归适用于配对或分层的病例–对照资料，通过在匹配组内进行比较来控制混杂因素的影响。模型在每个匹配组中引入特异性截距，并采用最大似然法进行参数估计，其回归系数的解释、假设检验及拟合优度检验方法与非条件 Logistic 回归基本一致，是分析配对病例–对照研究的合适方法。

第三节｜Logistic 回归的应用及注意事项（结构化总结）

一句话总览
Logistic 回归的价值不在“算 OR”，
而在于：
在复杂现实中，控制混杂、量化风险、支持决策。

一、Logistic 回归的主要应用场景（When & Why）

1️⃣ 流行病学危险因素分析（最经典用途）

核心问题

在存在混杂因素的情况下，
某一因素是否仍然是独立危险因素?

控制混杂的两条路径

1. **设计阶段**
    - 分层设计、配对设计

2. **分析阶段（更常用）**
    - 将混杂因素一并引入 **多因素 logistic 回归模型**

📌 关键优势：

> Logistic 回归可以**同时控制多个混杂因素**，
> 得到**调整后的 OR（adjusted OR）**。

研究设计的适用性

Logistic 回归可用于：

- 横断面研究
- 病例–对照研究
- 队列研究

👉 除病例–对照研究的**截距项解释不同**外，
回归系数（OR）的意义一致。

2️⃣ 临床研究与临床试验数据分析

典型问题

> 试验组与对照组
> **在年龄、病情、性别等方面不完全均衡，**
> **会不会“高估或低估”治疗效果？**

Logistic 回归的作用

- 将：
    - 年龄
    - 病情严重程度
    - 疾病亚型
    - 试验中心
      等非处理因素作为协变量
- 得到：

调整混杂因素后的治疗效应

📌 对分层设计的临床试验：

可用 logistic 回归
对分层因素进行**统一调整与分析。**

3️⃣ 药物 / 毒物的剂量–反应分析

核心用途

- 拟合 **剂量–反应曲线**
- 估计有效剂量（如 ED_{50}）
- 检验剂量–反应趋势

多药物联合作用（非常重要）

可在模型中加入交互项：

$$P = \frac{1}{1 + \exp\left[-(\beta_0 + \beta_1 A + \beta_2 B + \gamma AB)\right]}$$

- 若 $\gamma \neq 0$：
  👉 存在 **协同或拮抗作用**

📌 这在：

- 联合用药
- 毒物复合暴露
  中非常关键。

4️⃣ 预测与判别（风险评估工具）

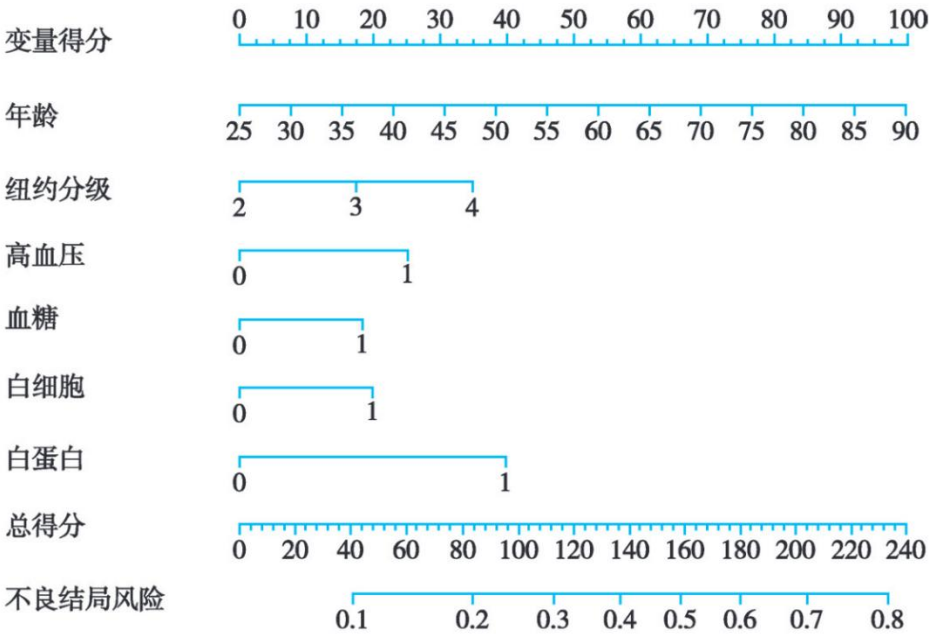

图 13-2　logistic 回归预测列线图

(李康,贺佳主编, 2024, p. 140) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

Logistic 回归的本质优势

> 它是一个 **概率模型**，
> 输出的是 **事件发生概率**，而不是“是/否”。

实际应用

- 个体患病风险预测
- 不良结局风险评估
- 临床决策支持

列线图（Nomogram，应用热点）

- 将 logistic 回归模型 **可视化**
- 变量 → 分值 → 总分 → 概率
- 方便医生和患者使用

📌 **重要限制（必考）**：

条件 logistic 回归
因不估计常数项
👉 **不能用于预测。**

二、Logistic 回归使用的关键注意事项（雷区集中区）

表 13-10　年龄变量离散化处理的赋值方法

年龄/岁 X	等级变量 G	哑变量		
		X_1	X_2	X_3
＜40	1	0	0	0
40~＜50	2	1	0	0
50~＜60	3	0	1	0
60~	4	0	0	1

(李康,贺佳主编, 2024, p. 140) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

1️⃣ 自变量取值方式（非常容易影响结论）

（1）二分类变量

- 常用 0 / 1 编码
- 解释清晰，最稳妥

（2）多分类变量

- 转化为 **哑变量（dummy variables）**
- g 类 → g – 1 个哑变量

⚠️ 重要规则：

当 $g > 2$ 时，
$g-1$ 个哑变量必须作为一个整体进出模型，
不能单独筛选。

（3）连续变量的三种处理方式（必会比较）

方法一：直接作为连续变量

- 优点：信息不丢失，检验效率高
- 缺点：参数解释不直观
  （如“年龄增加 1 岁”的 OR 可能无实际意义）

方法二：分组后按 1,2,…,g 赋值

- 简单，但隐含“组距等效”的假设

方法三：分组后转为哑变量（推荐）

- 解释最直观
- 但需要合理分组

📌 实务建议：

统计合理性 + 专业可解释性
比“计算方便”更重要。

2️⃣ 样本含量要求（Logistic 回归的硬前提）

原则

Logistic 回归的统计推断
建立在大样本近似基础上。

经验标准

- 病例与对照：
  - 各 ≥ 30–50 例
- 变量越多 → 样本需求越大
- **配对资料**：

  匹配组数 ≥ 自变量个数的 **20 倍**

📌 样本太小的风险：

- OR 极端
- CI 极宽
- Wald 检验失真

3️⃣ 变量间的交互作用（不是自动生成的）

什么是交互?

一个变量的效应
取决于另一个变量的水平

如何处理?

- **依赖专业知识先判断是否可能存在**
- 在模型中加入乘积项检验

📌 重要提醒：

不要机械地“全加交互项”，
那是制造不稳定模型。

4️⃣ 多分类 Logistic 回归（结局变量扩展）

当因变量不是二分类，而是：

- **无序多分类**
  （如疾病亚型）
- **有序多分类**
  （如无/轻/中/重，疗效分级）

可分别使用：

- 无序多分类 logistic 回归
- 有序多分类 logistic 回归

📌 核心结论：

> **除参数解释略有差别外，**
> **分析流程与二分类 logistic 基本一致。**

三、这一节的“判断级总逻辑”

> **① 是否存在混杂?**
> → 是：用多因素 logistic
>
> **② 是否是配对设计?**
> → 是：用条件 logistic（不能预测）
>
> **③ 变量怎么编码最有意义?**
> → 统计合理 + 专业可解释
>
> **④ 样本量够不够?**
> → 不够：任何 P 值都不可信

四、考试级一句话总结（可直接写）

> Logistic 回归广泛应用于流行病学危险因素分析、临床研究数据分析、剂量–反应关系研究及风险预测。通过多因素建模可在分析阶段控制混杂因素，得到调整后的效应估

计。实际应用中应注意变量取值方式、样本含量要求、交互作用的合理引入以及多分类结局变量的模型选择；条件 logistic 回归因不估计截距，不能用于预测分析。

[BIBLIOGRAPHY] Please click Zotero - Refresh in Word/LibreOffice to update all fields