第十二章 多元线性回归

第一节 多元线性回归

一、多元线性回归模型

> **一句话抓住本质**
> **多元线性回归不是“变量越多越好”，**
> **而是：**
> **👉 在其他因素都不变时，某一个因素对 Y 的“独立贡献”有多大。**

1.多元线性回归在解决什么问题？（Purpose）

当现实问题变成：

- 一个结果 Y
- **同时受多个因素影响**
- 各因素之间可能相互关联

你关心的是：

> **在控制其他变量后，**
> **每一个 X_j 对 Y 还能产生多大影响？**

👉 这正是一元回归**做不到**的事。

2.模型长什么样？（Model）

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m + e$$

这是整章的**母公式**。

每一部分的真实含义（一定要会解释）

- Y：因变量（响应变量）
- $X_1, \ldots, X_m$：多个自变量
- β_0：截距（所有 $X = 0$ 时的理论均值）
- β_j：**偏回归系数（partial regression coefficient）**
- e：随机误差（模型解释不了的部分）

3.“偏回归系数”到底偏在哪？（Very important）

> **β_j 的定义关键词是：**
> **“在其他自变量保持不变时”。**

直觉解释：

- 改变 X_j
- **冻结** 其他所有 X
- 观察 Y 的平均变化

📌 这就是多元回归存在的**根本理由**：
👉 **控制混杂因素（confounding）**

4.和一元线性回归的关系（一句话打通）

> **一元回归：**
> X → Y（忽略其他世界）
>
> **多元回归：**
> 在“其他因素不变”的条件下
> X → Y

📌 一元回归的 b是 **“总体效应”**
多元回归的 β_j是 **“条件效应”**。

5.多元线性回归的三个核心假设（必须背“逻辑”）

1️⃣ 线性关系假设

- Y 与各 X_j 的关系 **在模型中是线性的**
- 不要求现实世界严格线性
- 但**模型必须近似合理**

2️⃣ 独立性假设

- 各观测值 Y_i 相互独立
- 违反常见于：
  - 重复测量
  - 聚类数据

📌 独立性是统计推断的**地基。**

3️⃣ 残差正态 & 同方差假设（合并理解）

在任意一组 $X_1, \dots, X_m$ 给定时：

- e 的均值为 0
- 方差相同
- 近似正态分布

📌 本质是：
Y 在给定 X 条件下的分布性质一致

6.回归方程在“样本世界”里的样子

总体模型 → 样本估计：

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + \cdots + b_mX_m$$

- b_j：β_j 的估计值
- $\hat{Y}$：**条件均值的估计**

📌 再强调一次：

> **回归预测的是“平均水平”，**
> **不是个体命运。**

7.最小二乘法在多元回归中做了什么？（直觉）

目标函数：

$$Q = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

👉 在所有可能的“m 维平面”中：

> 找到那一个，
> 使 **残差平方和最小。**

区别只是：

- 一元回归：找“直线”
- 多元回归：找“超平面”

8.为什么 b_0 有固定形式？（理解点）

$$b_0 = \bar{Y} - (b_1\bar{X}_1 + \cdots + b_m\bar{X}_m)$$

直觉解释：

> **多元最小二乘解，**
> **同样必须通过样本均值点。**

只是现在这个点是：

$$(\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_m, \bar{Y})$$

9.为什么一定要用统计软件？（现实提醒）

- 多元回归需要解：
  - 多个正态方程
  - 矩阵运算
- 手算几乎不可行

📌 所以：

- **理解模型与结果**
- 而不是手推公式
  才是学习重点。

10.考试级一句话总结（可直接写）

> 多元线性回归模型用于分析因变量与多个自变量之间的线性数量关系，通过引入偏回归系数，在控制其他自变量影响的条件下，评估各自变量对因变量的独立作用，其参数通常采用最小二乘法并借助统计软件进行估计。

给你一个不会走偏的心智锚点

> **多元回归不是“多塞变量”，**
> **而是“更公平地分配解释权”。**

二、多元线性回归模型的假设检验

> **一句话总览**
> 多元线性回归的检验分三层：
> **① 整体模型有没有用？**
> **② 哪些自变量真正有用？**
> **③ 谁的作用更大？**

1.为什么要做假设检验？（Why）

样本回归系数 b_j：

- 是总体参数 β_j 的估计
- 即使 $\beta_j = 0$，也可能因为抽样误差使 $b_j \neq 0$

👉 **“看起来有关系” ≠ “真的有关系”**
必须通过统计检验确认。

2.模型检验（整体 F 检验）——“这套模型值不值得用？”

1️⃣ 检验的核心问题（Purpose）

> **m 个自变量合在一起，**
> **是否对 Y 有解释力?**

等价于问：

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_m = 0$$

2️⃣ 变异分解（结构必须会）

与一元回归完全同构，只是自由度不同：

$$SS_{总} = SS_{回归} + SS_{残差}$$

含义不变：

- $SS_{总}$：Y 的全部变异
- $SS_{回归}$：**被所有 X 一起解释的部分**
- $SS_{残差}$：模型解释不了的部分

3️⃣ 自由度结构（考试高频）

- 总自由度：

$$v_{总} = n - 1$$

- 回归自由度：

$$v_{回归} = m$$

- 残差自由度：

$$v_{残差} = n - m - 1$$

📌 **记忆技巧**：

回归自由度 = 自变量个数

4️⃣ F 统计量的本质

$$F = \frac{MS_{回归}}{MS_{残差}}$$

直觉解释：

- 如果模型没用：
  回归变异 ≈ 随机误差
  ⇒ F ≈ 1
- 如果模型有用：
  回归变异 ≫ 残差
  ⇒ F 很大

5️⃣ 结论怎么下？

- 若 $F \geq F_{\alpha(m,n-m-1)}$
  👉 拒绝 H_0
  👉 **至少有一个自变量与 Y 存在线性关系**

- 否则：
  👉 模型整体无统计学意义

📌 注意关键词：

> **“至少有一个”**
> 并不是“每一个”。

对于例 12-1，模型检验的假设为

$H_0: \beta_1 = \beta_2 = 0$

$H_1: \beta_1 \neq 0$ 和 $\beta_2 \neq 0$ 至少有一个成立

$\alpha = 0.05$

由 SPSS 或 SAS 统计软件可得到表 12-4 的模型检验结果。结果显示，$F = 21.54$，$P < 0.000\ 1$，拒绝 H_0，即所求回归方程具有统计学意义。

表 12-4　回归方程的方差分析表

变异来源	*SS*	ν	*MS*	*F*	*P*
回归	116.626	2	58.313	21.539	< 0.000 1
残差	46.025	17	2.707		
总变异	162.651	19			

(李康,贺佳主编, 2024, p. 123) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

3.偏回归系数检验——“到底是哪个 X 在起作用？”

整体模型显著 ≠ 每个变量都重要
所以要**逐个检验 β_j**。

4.偏回归系数的 F 检验（原理版）

1️⃣ 核心思想（非常重要）

比较：
“有 Xj” 和 “没 Xj” 的模型，
回归解释力差了多少？”

---

2 偏回归平方和（关键概念）

$$U_j = SS_{回归} - SS_{回归(-j)}$$

含义：

- U_j：**Xj 单独带来的解释增量**
- 自由度 = 1

---

3 F 统计量

$$F_j = \frac{U_j/1}{SS_{残差}/(n-m-1)}$$

判断逻辑：

- 若 F_j 大
  ⇒ 去掉 Xj 会“明显变差”
  ⇒ Xj 有统计学意义

对于例 12-1，胰岛素 X_1 和生长激素 X_2 全部纳入回归方程时，有 $SS_{回归}=116.626\ 5$，$SS_{残差}=46.024\ 6$

把 X_1 从回归方程中取出，建立 X_2 与 Y 的回归方程为

$$\hat{Y}=7.012+0.429X_2$$

此时

$$SS_{回归(-1)}=66.274\ 6$$

$$U_1=SS_{回归}-SS_{回归(-1)}=116.626\ 5-66.274\ 6=50.351\ 9$$

若把 X_2 从回归方程中取出，建立 X_1 与 Y 的回归方程为

$$\hat{Y}=-18.796\ 1-0.458\ 5X_1$$

此时

$$SS_{回归(-2)}=114.703\ 2$$

$$U_2=SS_{回归}-SS_{回归(-2)}=116.626\ 5-114.703\ 2=1.923\ 3$$

由公式（12-9）得

$$F_1=\frac{U_1/1}{SS_{残差}/(n-m-1)}=\frac{50.351\ 9/1}{46.024\ 9/(20-2-1)}=18.598\ 2$$

$$F_2=\frac{U_2/1}{SS_{残差}/(n-m-1)}=\frac{1.923\ 3/1}{46.024\ 9/(20-2-1)}=0.710\ 4$$

查 F 界值表（附表 4），$F_{0.05(1,17)}=3.59$，$F_1>3.59$，$F_2<3.59$。因此，在 $\alpha=0.05$ 水平上，可以认为胰岛素对血糖的作用有统计学意义（$P<0.05$），而生长激素则对血糖变化无影响（$P>0.05$）。

(李康,贺佳主编, 2024, p. 124) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

5.偏回归系数的 t 检验（实际最常用）

1 形式

$$t_{b_j}=\frac{b_j}{S_{b_j}},v=n-m-1$$

- S_{b_j}：偏回归系数的标准误
- 本质：

  估计值 / 抽样误差

2️⃣ 与 F 检验的关系（必考点）

单个自变量时：

$$F_j = t_{b_j}^2$$

📌 所以软件中：

- 看 t 检验即可
- 不必重复做 F

> 2. *t* 检验
>
> 原假设为 $H_0:\beta_j=0$，备择假设为 $H_1:\beta_j\neq 0$。*t* 检验方法比 *F* 检验更为简便，检验统计量为
>
> $$t_{b_j}=\frac{b_j}{S_{b_j}} \qquad (12\text{-}11)$$
>
> 其中 S_{b_j} 为偏回归系数 b_j 的标准误。
>
> 由表 12-3 中的回归系数的估计结果，按公式（12-11），X_1 的检验结果为 $t_{b_1}=-4.313$，$P<0.001$；X_2 的检验结果为 $t_{b_2}=0.843$，$P=0.411$。结论与方差分析检验结果相同，即胰岛素对血糖水平的作用具有统计学意义，生长激素则未显示出其对血糖的作用。

(李康,贺佳主编, 2024, p. 124) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

6.标准化回归系数（解决“谁更重要”）

1️⃣ 为什么原始 b 不能直接比?

因为：

- 单位不同（kg、cm、mmol/L...）
- 变异程度不同

👉 原始 b **没有可比性**

2️⃣ 标准化回归系数是什么?

$$b'_j = b_j(\frac{S_j}{S_Y})$$

等价于:

> **把所有变量都转成 z-score 后再回归**

3️⃣ 如何解释?

- $|b'_j|$ 越大
  ⇒ 在控制其他变量后
  ⇒ **对 Y 的影响越强**
- 常用于:
  - 解释变量重要性
  - 不用于预测

7.整套检验逻辑的一张“脑内流程图”

> **① 先看模型 F 检验**
> → 模型不显著:停
> → 模型显著:继续
>
> **② 看每个自变量的 t 检验**
> → 哪些 $P < \alpha$:真正有用
>
> **③ 用标准化回归系数比较影响大小**

8.考试级一句话总结(可直接写)

> 多元线性回归模型的假设检验包括整体模型的 F 检验和偏回归系数的检验。整体检验用于判断自变量集合对因变量是否具有解释作用,而偏回归系数检验用于判断在控制

其他自变量后单个自变量的独立效应；必要时可通过标准化回归系数比较各自变量对因变量影响的相对大小。

三、应用实例

【例 12-2】为了研究影响糖尿病患者糖化血红蛋白（HbA1c）的主要危险因素，某研究者收集了糖尿病患者的糖化血红蛋白 Y（%）、年龄 X_1（岁）、体重指数 X_2（kg/m^2）、总胆固醇 X_3（mmol/L）、收缩压 X_4（mmHg）和舒张压 X_5（mmHg）等数据资料。现从中随机抽取了 20 例，数据见表 12-5，试作多元线性回归分析。

表 12-5　20 例糖尿病患者的数据资料

编号	X_1	X_2	X_3	X_4	X_5	Y	编号	X_1	X_2	X_3	X_4	X_5	Y
1	49	32.19	6.0	148	86	7.6	11	53	23.43	7.1	161	86	7.5
2	67	24.77	2.7	151	98	7.4	12	46	30.56	2.9	146	79	7.3
3	64	25.24	7.0	151	80	7.4	13	59	25.19	6.0	158	80	7.3
4	66	24.26	4.8	157	87	7.2	14	76	27.26	5.4	124	85	6.9
5	68	30.28	3.5	136	83	7.3	15	63	23.93	6.7	133	89	7.5
6	48	26.18	7.6	137	87	7.6	16	74	24.94	7.9	166	82	7.9
7	66	26.36	5.9	157	91	7.5	17	52	22.82	5.3	149	71	7.3
8	47	32.07	5.7	157	89	7.7	18	64	24.34	2.5	126	93	6.8
9	64	28.44	6.1	154	82	7.3	19	54	25.44	2.6	151	83	6.9
10	75	30.65	6.9	137	86	7.7	20	78	28.98	7.2	147	74	7.5

对于表 12-5 的数据用 SPSS 或 SAS 统计软件计算，主要结果见表 12-6 和表 12-7。

表 12-6　回归方程的方差分析表

变异来源	SS	ν	MS	F	P
回归	1.079 06	5	0.215 81	7.32	0.001 5
残差	0.412 94	14	0.029 50		
总变异	1.492 00	19			

表 12-7　偏回归系数估计结果

自变量	偏回归系数	标准误	t	P
常数项	3.875 98	1.011 15	3.83	0.001 8
X_1	−0.001 53	0.004 09	−0.37	0.714 6
X_2	0.031 92	0.013 48	2.37	0.032 8
X_3	0.108 34	0.024 51	4.42	0.000 6
X_4	0.008 50	0.003 68	2.31	0.036 6
X_5	0.010 58	0.006 63	1.60	0.132 8

(李康,贺佳主编, 2024, p. 125) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

由表 12-6 可见，$F=7.32$，$P<0.01$，此回归方程有统计学意义。由表 12-7 可见，自变量 X_2，X_3，X_4 按 $\alpha=0.05$ 检验水准具有统计学意义，但 X_1 和 X_5 不显著。

(李康,贺佳主编, 2024, p. 126) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

四、复相关系数与决定系数（

一句话总览
F 检验回答“有没有用”，
R / R² 回答“有多好用”。

1.为什么要引入 R 和 R²？ （Why）

在多元线性回归中：

- F 检验只能告诉你：
  👉 **模型是否显著**
- 但它**不告诉你**：
  👉 模型解释了 Y 多少变异

而我们真正关心的是：

这套回归方程，对 Y 的解释力有多强?

这就是 **复相关系数 R** 和 **决定系数 R²** 的角色。

2.复相关系数 R ——“整体相关性有多强？”

1️⃣ 定义（Purpose）

复相关系数（multiple correlation coefficient)
表示：**所有自变量 X 的线性组合**与**因变量 Y** 之间的相关密切程度。

📌 本质一句话：

回归预测值 $\hat{Y}$ 与真实 Y 的相关系数

2️⃣ 取值范围（与简单相关的关键区别）

- 简单相关系数：
  $-1 \le r \le 1$
- **复相关系数：**

$$0 \le R \le 1$$

👉 原因很重要（常被问）：

R 描述的是“解释强度”，**不带方向性**
⇒ 只能为非负值。

3️⃣ 计算公式（结构记忆）

$$R = \sqrt{\frac{SS_{回归}}{SS_{总}}} = \sqrt{1 - \frac{SS_{残差}}{SS_{总}}}$$

直观理解：

- 分子：模型解释的变异
- 分母：Y 的全部变异
- 开平方：把“解释比例”还原成“相关强度”

4️⃣ R 的直觉解释

- R 越大⇒ 回归方程整体预测 Y 越准
- R = 0⇒ 模型几乎不解释 Y
- R = 1⇒ 完全线性解释（现实中极少）

📌 但注意：

R 本身不直接用于解释“解释了多少”
👉 这一步交给 R^2。

3.决定系数 R^2 ——“解释了多少变异？”

1️⃣ 定义（核心概念）

决定系数（coefficient of determination）
是复相关系数的平方：

$$R^2 = R^2$$

表示：

回归方程能够解释 Y 总变异的比例

2️⃣ 计算公式（考试必会）

$$R^2 = \frac{SS_{回归}}{SS_{总}} = 1 - \frac{SS_{残差}}{SS_{总}}, 0 \le R^2 \le 1$$

结构意义非常清晰：

- 分子：模型解释的那一部分
- 分母：全部变异
- 比值：解释比例

3️⃣ 如何解释 R^2（一定要会“翻译成人话”）

- R^2 越接近 1⇒ 模型拟合越好
- R^2 越接近 0⇒ 模型解释能力越弱

📌 **标准考试表述模板**（你给的例子非常典型）：

> 例：$R^2 = 0.7232$
> 表明由年龄、体重指数、总胆固醇、收缩压和舒张压
> **可解释该样本中糖化血红蛋白变异的 72.32%。**

4.R / R^2 与 F 检验的关系（结构性理解）

这是很多人**概念混乱的地方**，一句话给你理顺：

> **F 检验：** → 判断“解释是否显著存在”
>
> **R^2：** → 衡量“解释的强弱程度”

关键逻辑：

- F 不显著 👉 R^2 再大也**不可靠**
- F 显著 👉 R^2 才有**解释价值**

📌 所以顺序永远是：

> **先看 F → 再解释 R^2**

5.常见误区（这一节的“雷区”）

⚠️ 误区 1：R^2 大 ≠ 模型一定好

- 自变量越多 ⇒ R^2 **一定不会减小**
- 但可能只是：👉 **“堆变量”造成的假提升**（这也是为什么后面会引出 **调整 R^2**）

⚠️ 误区 2：R^2 不能说明因果

R^2 只说明：**统计解释程度**

不说明：

- 因果方向
- 机制是否合理
- 模型是否可推广

⚠️ 误区 3：不能用 R^2 比较“不同因变量”的模型

因为：

- Y 的变异尺度不同
- SS_总 本身不可比

6.这一节的“脑内结构图”

SS 总 = SS 回归 + SS 残差

↓

- **R**：
  回归预测值 $\hat{Y}$ 与 Y 的相关强度
- **R^2**：
  回归模型解释 Y 变异的比例

↓

F 检验判断“有没有”
R^2 判断“有多少”

7.考试级一句话总结（可直接抄）

复相关系数 R 表示回归方程中全部自变量的线性组合与因变量之间的相关密切程度，其平方 R^2 称为决定系数，反映回归模型对因变量变异的解释比例。R^2 越大，说明回归模型对数据的拟合程度越好，但其解释应以整体模型的显著性检验为前提。

第二节｜多元逐步回归

一句话总览
多元逐步回归的本质不是“找最大 R^2”，
而是：
在控制统计显著性的前提下，
用最少的自变量，构建一个解释清晰、预测稳定的模型。

一、为什么需要逐步回归？（Why）

在多元线性回归中，当自变量较多时，通常会同时面临三类问题：

1. **并非所有 X 都真的有用**
   - 有些回归系数不显著
   - 只是“被带进模型”的
2. **自变量之间可能相关**
   - 信息重叠
   - 多重共线性（multicollinearity）
3. **模型过复杂，反而更差**
   - 参数多 → 残差均方变大
   - 解释困难、预测不稳

👉 因此，我们希望：

保留“真正有统计学意义”的自变量，
剔除“贡献小、冗余高”的变量。

二、逐步回归在“回归体系”中的位置

先明确一件事（非常重要）：

> **逐步回归 ≠ 一种新的回归模型**
> 它是：
> **多元线性回归中的“变量筛选策略”**

本质仍然是：

- 线性模型
- 偏回归系数
- F / t 检验
- 残差分析

三、三种经典自变量筛选方法

1️⃣ 向前选择法（Forward Selection）

基本思路：

> 从“无变量”开始，
> **一次引入一个最有用的变量。**

操作逻辑：

1. 从所有 X 中
   👉 选择 **偏回归平方和最大且显著** 的变量
2. 引入模型
3. 在剩余变量中重复
4. 直到 **没有变量再满足引入标准**

优点：

- 模型由简到繁
- 计算直观

致命问题（必考理解点）：

> **先进入的变量，**
> **可能在后续引入新变量后变得不显著，**
> **但却不会被自动剔除。**

👉 容易留下“历史遗留变量”。

2️⃣ 向后选择法（Backward Selection）

基本思路：

> 从“全模型”开始，**一次剔除一个最没用的变量。**

操作逻辑：

1. 建立包含所有自变量的模型
2. 找到：
    - 偏回归平方和最小
    - 且 **无统计学意义** 的变量
3. 将其剔除
4. 重复，直到不能再剔除

适用条件（非常具体，容易考）：

- 样本量较大：

$$n > 100$$

- 或自变量较少：

$$m < 10$$

局限：

- 若样本量小、变量多
  👉 初始模型不稳定
- 共线性强时
  👉 剔除顺序可能“失真”

3️⃣ 逐步选择法（Stepwise Selection）

> **向前 + 向后 的结合版**

也是**实际最常用**的一种。

核心思想（一定要背逻辑）：

> **每引入一个新变量，都要重新检查：**
> **已进入模型的变量是否还“值得留下”。**

操作特征：

- 引入（forward）
- 剔除（backward）
- **交替进行**
- 直到：
  - 无新变量可引入
  - 无旧变量可剔除

👉 形成一个“动态平衡”的模型。

四、逐步回归的统计判据（怎么“进 / 出”？）

1️⃣ 基本检验工具

逐步回归的判断基础是：

偏回归平方和的 F 检验

即比较：

- “有该变量”
- “无该变量”

模型解释力是否显著变化。

2️⃣ F 检验水准的设定（极易忽略，但很重要）

在逐步回归前，必须**先设定**：

- **引入标准**：α_引入
- **剔除标准**：α_剔除

一般要求：

α_引入 ≤ α_剔除

常用水平：

- 0.05
- 0.10
- 0.20

📌 现实权衡：

- α 太严格 → 模型过简

- α 太宽松 → 筛选失效

五、校正决定系数（Adjusted R^2）在逐步回归中的角色

1️⃣ 为什么不能只看 R^2?

因为：

> **R^2 随自变量个数增加“单调不减”**

👉 会奖励“堆变量”的模型。

2️⃣ 校正决定系数的定义

$$R_c^2 = 1 - (1 - R^2)\frac{n - 1}{n - p - 1} = 1 - \frac{MS_{误差}}{MS_{总}}$$

其中：

- p：进入模型的自变量个数

3️⃣ 如何使用?

在多个候选模型中：

> **选择校正 R^2 最大的模型，**
> **作为“统计意义下的最优方程”。**

📌 注意措辞：

> **“最优”是相对的、准则依赖的。**

六、逐步回归的根本局限

这是老师**最爱考“理解题”的地方**：

> 逐步回归得到的“最优模型”，
> **高度依赖于：**
>
> - F 检验水准的选择
> - 变量进入 / 剔除顺序
> - 样本本身

👉 因此：

> **无法保证在所有准则下都是最优模型。**

七、最重要的一句话（方法论底线）

> **逐步回归只能提供统计上的候选模型，**
> **回归方程是否合理，必须结合专业背景与实际意义判断。**

在生物医学 / 流行病学里尤为重要。

八、考试级一句话总结

> 多元逐步回归是在多元线性回归中通过向前选择、向后剔除或双向筛选的方法，引入具有统计学意义的自变量并剔除不显著变量，以获得较为简洁且解释性较强的回归模型。其筛选结果依赖于检验水准和变量进入顺序，所得“最优”模型需结合专业知识进行判断。

对例 12-2 进行多元逐步回归，采用逐步法筛选自变量，选入水准为 0.10，剔除水准为 0.15。逐步回归的过程及分析结果归纳如表 12-8、表 12-9 和表 12-10，具体步骤如下。

表 12-8　逐步回归过程

步骤（l）	引入变量	剔除变量	变量个数 p	R^2	$SS_{回}^{(l)}(X_j)$	$SS_{残}^{(l)}$	F 值	P 值
1	X_3		1	0.480	0.717	0.775	16.640	0.001
2	X_2		2	0.564	0.124	0.651	3.259	0.089
3	X_4		3	0.671	0.160	0.491	5.220	0.036

表 12-9　方差分析表

变异来源	SS	ν	MS	F	P
总变异	1.492	19			
回归	1.001	3	0.334	10.889	＜0.001
残差	0.491	16	0.031		

表 12-10　回归系数估计及检验结果

变量	回归系数	标准误	标准回归系数	t	P
常数项	4.799	0.667	0.000	7.194	＜0.001
X_2	0.031	0.014	0.341	2.287	0.036
X_3	0.097	0.024	0.611	4.125	0.001
X_4	0.008	0.004	0.330	2.285	0.036

最后得到的回归方程为

$$\hat{Y} = 4.799 + 0.031X_2 + 0.097X_3 + 0.008X_4$$

根据上述结果，可以认为体重指数 X_2、总胆固醇 X_3 和收缩压 X_4 是影响糖化血红蛋白的主要因素。

(李康,贺佳主编, 2024, p. 127) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

第三节 多元线性回归的注意事项

一句话总览

多元回归最容易翻车的不是算错，

而是：**不满足条件、样本不够、编码不当、共线性、盲信筛选。**

一、应用条件（Assumptions）——“能不能用多元线性回归？”

多元线性回归原则上要求：

1️⃣ 因变量类型

- **Y 应为连续型变量**（continuous outcome）

2️⃣ 误差项分布与方差结构

模型误差 e 要满足：

- **正态性（normality of errors）**
  e 近似服从正态分布
  （注意：强调的是**残差**而不是 Y 本身）
- **方差齐性（homoscedasticity）**
  不同 X 取值下，误差方差大致相同
  （否则标准误、t/F 检验会失真）

3️⃣ 观测值独立性（independence）

- **Y 的观测值相互独立**
  对“传染性疾病”等可能出现**聚集/传播相关**的数据要谨慎
  （常见问题：同家庭/同班级/同病房导致相关）

📌 一句话记忆：

> 条件不满足，回归系数可能还能算出来，
> 但 t / F / P 值可能不可信。

二、样本含量（n 与 m 的比例）——“稳不稳定？”

1️⃣ 核心风险

当自变量个数 m 较多，而样本量 n 相对不大时：

- 回归方程**不稳定**
- 可能出现 **R^2 很大**的“假象”
- 预测在新样本上容易崩

2️⃣ 经验法则（考试常用）

> n 至少应为自变量个数 m 的 **5–10 倍**
> 即：$n \geq 5m \sim 10m$

📌 直觉解释：
变量多但样本少，相当于“用太多参数去贴合噪声”。

三、定性变量数量化（Categorical encoding）——“怎么把分类 X 放进模型？”

多元回归自变量通常是连续型，但分类变量必须先**数量化。**

1️⃣ 二分类变量（Binary）

- 直接用 0/1 编码即可
  例如：男=1，女=0（或反过来都行，但解释要一致）

$$X=\begin{cases}1, & \text{男性}\\ 0, & \text{女性}\end{cases} \quad \text{或} \quad X=\begin{cases}1, & \text{女性}\\ 0, & \text{男性}\end{cases}$$

(李康,贺佳主编, 2024, p. 128) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

2️⃣ 多分类变量（k 类）

用 **k−1 个哑变量（dummy variables）** 表示

关键点只有一个：

- **必须留一个“参照组（reference group）”**
- k 类 → 放进模型的是 k−1 列 0/1

📌 含义：
每个哑变量的系数，都是“该组相对参照组”的差异（在控制其他变量后）。

3️⃣ 有序变量（Ordinal）

例如轻/中/重：

- 可以按 1/2/3 直接赋值进入模型
- 样本量大时，也可拆成哑变量更稳妥（避免强行假设“间距相等”）

$$X_1=\begin{cases}1, & \text{A 型}\\ 0, & \text{其他血型}\end{cases},\quad X_2=\begin{cases}1, & \text{B 型}\\ 0, & \text{其他血型}\end{cases},\quad X_3=\begin{cases}1, & \text{AB 型}\\ 0, & \text{其他血型}\end{cases}$$

对此也可以用表 12-11 表述。

表 12-11　血型变量的哑变量赋值方法

血型 X	分组编码 G	哑变量		
		X_1	X_2	X_3
O 型	1	0	0	0
A 型	2	1	0	0
B 型	3	0	1	0
AB 型	4	0	0	1

(李康,贺佳主编, 2024, p. 128) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

4️⃣ 重要提醒：k>2 时，逐步回归不能“单独挑某一个哑变量”

这点非常关键、也很容易错：

> 当分类数 $k > 2$ 时，
> **k−1 个哑变量应作为一个整体同时进出模型。**

否则会出现“只进来其中一类”的荒谬结果，解释会变得不一致甚至错误。

四、多重共线性（Multicollinearity）——“为什么显著性突然乱套？”

1️⃣ 发生条件

当自变量之间存在较强线性关系（信息高度重叠）：

- 参数估计会变得**不稳定**
- 标准误被放大
- 系数符号可能反常、P 值飘忽

2️⃣ 直观后果

你会看到典型怪象：

- 模型整体 F 检验显著
  但单个变量 t 检验都不显著
- 换一批变量/换一下进入顺序，结果大变

3️⃣ 教材给的直觉例子

年龄、吸烟年限、饮酒年限往往彼此相关：相关很高时，OLS（最小二乘）就可能不合理或解释困难。

五、关于变量筛选（Stepwise 等）——“最快得到结果，但最危险”

1️⃣ 可以用，但不能迷信

逐步回归能让问题“快速简化”，但：

所谓“最优”方程 **不一定最好。**
没选入的变量 **也未必没意义。**

2️⃣ 为什么结果会变?

因为逐步回归高度依赖:

- 引入/剔除的检验水准（α 的设定）
- 变量进入顺序
- 变量组合改变后，某变量的 P 值也会变

📌 一句话总结风险:

逐步回归给你的是“统计学上的候选模型”，
不是“真理模型”。

3️⃣ 正确姿势（考试也能写）

建立回归方程时:

最好结合研究目的与专业知识，
预先确定核心变量，
统计筛选只能作为辅助。

六、考试级一句话总结（可直接写）

多元线性回归要求误差项近似正态、方差齐性及观测值独立，样本量应相对自变量数足够（经验上 n 为 m 的 5–10 倍）。分类自变量需数量化，k 类分类变量用 $k-1$ 个哑变量表示且应整体进出模型。自变量间强相关会导致多重共线性，使参数估计不稳定、显著性检验失真；逐步回归虽便于筛选变量，但结果依赖检验水准与变量组合，应结合专业知识判断模型合理性。

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