回归分析基础

易侕学院 陈常中

回归（Regression）这个词最早由英国学者高尔顿于 1886 年提出，他在研究父亲的身高和儿子的身高时发现子辈的平均身高是其父辈平均身高以及他们所处族群平均身高的加权平均和。我们把这种现象称为均值回归或者平庸回归（reversion to the mean/reversion to mediocrity）。后来尤勒在高尔顿的基础上提出回归模型中应当加入尽可能多的控制变量。他在 1899 年研究英国济贫法对贫困率的影响的论文中，把地区的人口以及年龄的分布加入到了控制变量当中去，研究结果影响深远。

第一节 回归方程

一、联系函数

一个单元广义线性回归方程表达式为 ：

$$f(Y) = \beta_0 + \beta_1 * X$$

英文中广义（generalize）也可以翻译成“通用”。之所以“通用”是因为方程的左边用的是 f(Y)，即 Y 的联系函数。相当于我们日常生活中所说的用什么方式来联系，如会面、微信、电子邮件和电话等等。选择什么方式取决于 Y 的分布类型。不同的联系方式传递的是不同形式的信息，如语音、文字、表情符号等，不能同日而语。根据 Y 的分布，f(Y)不同，常见的有：

表 7-1 Y 的分布与对应函数

Y 的分布	f (Y)
正态分布（normal）	Y
二项分布（binomial）	Logit（Y）
γ 分布（gamma）	$1/(Y^{-1})$
负二项分布（negative binomial）	Log（Y）
Poisson 分布	Log（Y）

注：表中对数函数的底数默认为自然数

（1）如果 Y 是连续性变量，近似正态分布，联系函数取其本身，英文为 identity，即 f(Y) = Y。β_1反映的是 X 每增加 1，Y 增加多少。

（2）如果 Y 是 0 或 1 的两种状态，即二项分布，联系函数取 logit，logit(Y) = log(P/(1-P))，P 即 Y=1 的率，P/（1-P）即发生 Y 的比值。β_1反映的是 X 每增加 1，发生 Y 的比值的对数增加多少。

因为两个对数的差等于两个数相除后的对数，因此β_1的反对数即为发生 Y（或者说 Y=1）的比值比 ，也是文献上常见的 OR （odds ratio）。

比值（Odds）区别于率（Rate），率为 Y=1 的概率。如用 P 表示率，则 Odds = P/(1-P)。这两者虽有区别，但都是衡量 Y 的一种状态（Y=1）出现的概率大小，特别是当 P 比较小的时候，1-P 几乎等于 1，比值近似于率，比值比（OR）近似于率比（RR）。

当 Y 是连续性变量时，流行病学研究所关心的效应是 X 每增加 1，Y 改变多少。当 Y 为是否患病（0 或 1）两种状态时，流行病学所关心的效应是 X 每增加 1，患病的比值（或概率）改变多少。

我们实际观察到的数据不是每个观测点都在 $f(Y) = \beta_0 + \beta_1*X$ 这条线上，但是围绕这条线上下分散的，就像孙悟空跳不出如来佛的手心一样，这也就是回归的本意。观测值与由方程右边计算出来的预计值之差称为残差。数据分析的目的就是透过现象看本质，现象是观察到的 f(Y)上下波动，看似杂乱无章，本质则是它们都是围绕这条回归线$\beta_0 + \beta_1*X$ 上下波动。

之所以称线性回归方程，是因为 X 对 f(Y) 的影响作用是线性的，X 每增加一个单位，f(Y)增加β_1。X 每增加一个单位意味着什么？如果 X 是连续性变量，如年龄，单位是岁，一个单位表示一岁。如果 X 是 0/1 两分类变量，如 X=0 表示男性，X=1 表示女性，增加一个单位就是从男变为女，也就是男女之差（女性 － 男性）。如果 X 是多分类变量，取值为 0、1、2、3，每增加一个单位就是从 0 到 1，或 1 到 2，或 2 到 3。

二、非线性关系的拟合

如果 X 每增加 1，f(Y)的改变幅度不是固定的β_1，怎么构建方程呢？这种情况只有当 X 是连续性变量或多分类变量时才会出现。如果 X 是多分类变量，首先为每个分类产生一个哑变量，所谓哑变量，其取值是 0 或 1，取值 1 表示原变量 X 落在该分类上，否则取值为 0。取一个分类为对照，将代表其它分类的哑变量放入方程中。以 X 为三分类（0、1、2）为例，首先按下列条件生成 X_0、X_1、X_2三个哑变量：

如 X=0，则：X_0=1，X_1=0，X_2=0

如 X=1，则：X_0=0，X_1=1，X_2=0

如 X=2，则：X_0=0，X_1=0，X_2=1

以 X=0 组为参照组，用 X_1、X_2构建方程为：

$$f(Y) = \beta_0 + \beta_1*X_1 + \beta_2*X_2$$

从这个方程中不难看出，当 X=0 时，$f(Y) = \beta_0$；当 X=1 时，$f(Y) = \beta_0+\beta_1$；当 X=2 时，$f(Y) = \beta_0+\beta_2$。β_1即 X=1 组与 X=0 组 f(Y)的差，β_2即 X=2 组与 X=0 组 f(Y)的差。

如果 X 是连续性变量，X 每增加 1，f(Y)的改变幅度不是固定的β_1，有三种拟合方法：

（1）可以试着用 $f(Y) = \beta_0 + \beta_1 * X + \beta_2 * X^2$ 或再加上 X^3 来拟合，要根据 X 与 Y 的实际关系分析这样拟合是否充分。

（2）用 $f(Y) = \beta_0 + s(X)$，$s(X)$表示平滑曲线函数，即用一条平滑曲线来表达 X 与 f(Y)的关系。曲线形态多样灵活，能确切反映 X 与 Y 的各种形态的关系。缺点是曲线很难用文字描述清楚，而且不便于应用。通常我们首先用平滑曲线拟合来探索，然后尽可能地用简化的回归模型来表达。

（3）把 X 的取值从最小到最大分成几段，然后用 0、1、2、……依次代表这几段，将原来连续性的 X 转换为 0、1、2、……多分类变量，再按上述的多分类变量拟合方法拟合。回归方程拟合段与段之间的差异，忽略段内的差异。因此，分段的原则是使段之间 f(Y)的差异越大越好而段内 f(Y)的差异越小越好。假如 X 与 Y 的关系成 U 型曲线，如下图所示 7-1，如以中间拐点（A）将 X 分成两段，左边 X 小于 A1 这一段 Y 的均值约为 A1，Y 的范围从最小到最大；右边 X 大于 A 这一段 Y 的均值约为 A2，Y 的范围也是从最小到最大。A1 与 A2 没有差别，这样分组的结果是组间没有差异而组内差异很大，这就违背了上述原则。如果将 X 按切点 B 和 C 分成三段（图 7-2），X 小于 B 这一段 Y 的均值约为 M1，X 在 B 和 C 之间这一段 Y 的均值约为 M2，X 大于 C 这一段 Y 的均值约为 M3，这样分组的结果是组间差异大，组内差异小，符合上述原则。

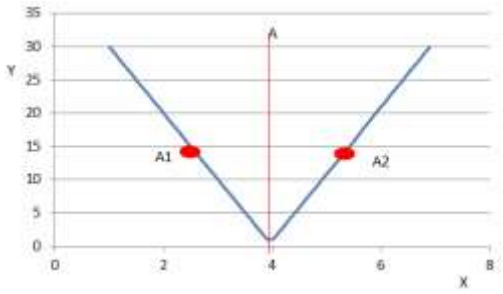

图 7-1 示意图

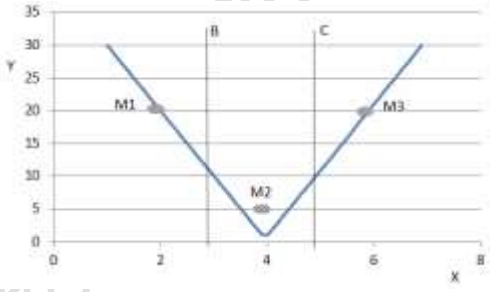

图 7-2 示意图

三、趋势检验

如果 X 是连续性变量，首先要检查 X 与 f(Y)是不是直线关系，即 X 每增加 1，f(Y)的改变幅度是固定的，是直线关系才可以直接把 X 放入方程中拟合。

如何判断 X 与 f(Y)的关系是直线关系呢？最简单直接的方法是平滑曲线拟合（spline smoothing），看曲线形态判断 X 与 Y 是不是直线关系。

如果不用平滑曲线拟合，则怎么分析呢？首先将 X 四等分，如样本量比较大可五等分，为每个等分组生成一哑变量，第一个等分组做参照，将其它几个哑变量放入模型中，观察哑变量的回归系数是否成线性增加或减少，然后做趋势检验。例，下表是 AGE 四分组对 SBP（收缩压）的回归分析及趋势检验结果：

表 7-2 AGE 四分组对 SBP（收缩压）的回归分析及趋势检验

AGE group	回归系数（95%可信区间）P 值
Q1：15.6 - 27.5	Ref.
Q2：27.6 - 33.9	0.4 (-3.5, 4.3) 0.844
Q3：34.0 - 47.8	4.1 (0.3, 8.0) 0.036
Q4: 47.9 - 77.0	26.1 (22.2, 30.0) <0.001
Trend test	0.8 (0.7, 0.9) <0.001

AGE 四等分后，第一四分组 Q1 为参照组，回归系数为 0，第二四分组 Q2 的回归系数为 0.4，第三四分组 Q3 为 4.1，第四四分组 Q4 为 26.1，逐步上升，Q1 与 Q2 比较接近，可以想象如果平滑曲线拟合，前面一段基本上是平的或缓慢上升，到了 Q4 上升比较快。下图 7-3 是曲线拟合图，对照图形与上表结果，可以帮助我们进一步理解回归方程。

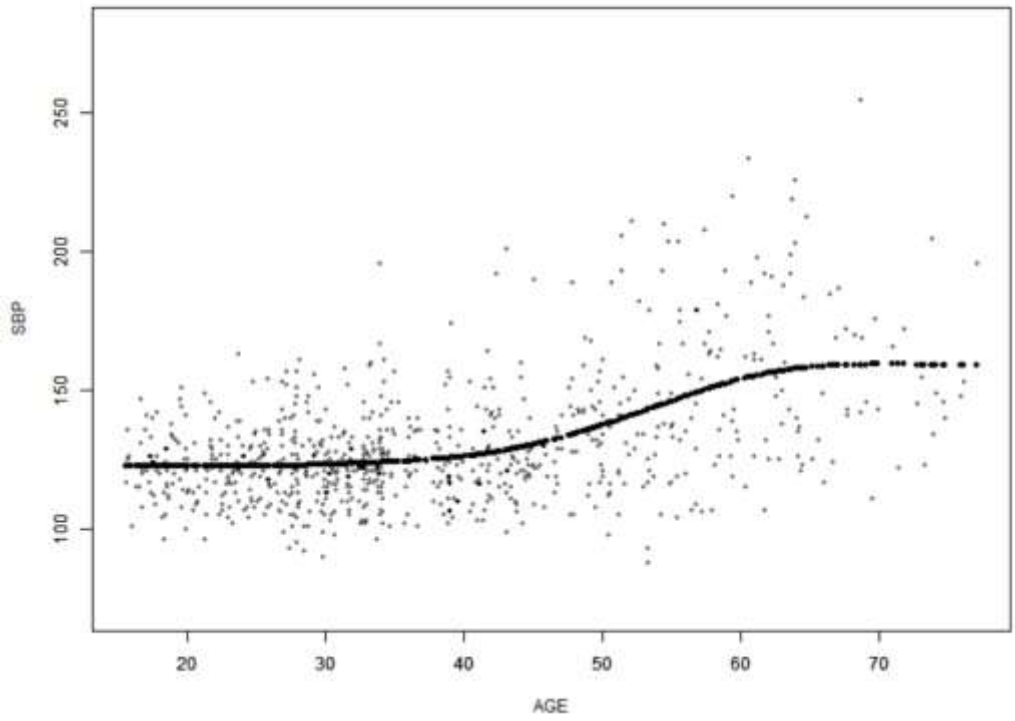

图 7-3 AGE 和 SBP 关系的曲线拟合图

上表中的趋势检验（trend test）是怎么做的呢？首先计算每等分组 X（年龄）的中位数 Xm，用组内中位数替代原始的年龄值，然后运行模型 SBP = $\beta_0 + \beta_1 * Xm$。 此模型的 Xm 不是每个观察对象的年龄，而是其所在年龄等分组的年龄中位数。

为什么不直接用 0、1、2、3 代替 X，而要用各等分组年龄的中位数代替 X 呢？因为用 0、1、2、3 代替 X 则表示组间 X 的差距均是 1，得出来的回归系数表示年龄每增加 1 个四分等级，SBP 增加多少。实际上 Q1 的年龄中位数为 22.6，Q2 为 31.2，Q3 为 40.6，Q4 为 56.3，组间中位数的差距是 8.6、9.4、15.7，是不均等的。用实际年龄中位数代替 X，得出来的回归系数表示年龄每增加 1 岁，SBP 增加多少。

可以想象，如果 Y 与 X 的曲线拟合是一个 U 型曲线，把 X 等分成 5 组，中间的一组（第 3 组）与两段（第 1 组和第 5 组）相差最大，组间比较可能有显著性，如果把第 3 组做为参照组，第 1 组和第 5 组的回归系数可能显著（不等于 0），但趋势检验和直线拟合结果，X 对 Y 的回归系数都得不显著的。

四、多元回归方程

多元回归方程的一般表达式为：

$$f(Y) = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_3 + \cdots\cdots$$

如前面所说的单元回归方程一样，方程中的 β_1、β_2、β_3 分别表示 X_1、X_2、X_3 每增加一个单位 f(Y) 增加多少。虽说如此，两者却有根本区别。如果 X_1、X_2、X_3 对 f(Y) 都有影响，这里所说的 X_1 每增加一

个单位 f(Y)增加β_1，有一个条件，那就是在其它因素（X_2、X_3……）不变的情况下，如果其它因素也在变，那么各自对 f(Y)的贡献加到一起就是最终 f(Y)的改变。换一句话说，β_1就是调整了其它因素（X_2、X_3……）的作用后，X_1对 f(Y)的作用。对 X_2、X_3来说亦是如此。

β_0是什么呢？是 X_1、X_2、X_3……均为零时，f(Y)的期望值（均值）。如果 X 是诸如性别、是否吸烟、职业这类的分类变量，代表的就是 X=0 这一类的人，也就是我们常说的参照组；如果 X 是诸如身高、体重这类的连续性变量，等于零的人是不存在的，β_0是其回归线延申后得出的截距。然而，我们可以把这类连续性变量中心化，即都减去某一有代表性的值或其在人群中的均值，如身高减去 170 厘米、年龄减去 50 岁，这样一来β_0就代表身高为 170 厘米，年龄 50 岁，再加上其它自变量为零所代表的一类人，f(Y)的均值。理解β_0有助于我们进一步理解回归方程。

同理，如果 X_1、X_2、X_3每增加一个单位 f(Y)的改变幅度不是固定的，那么对它的拟合就应该参照上述一元回归方程拟合的方法。

多元回归方程表达式右边所包含的自变量越多，方程的自由度越大，同时意味着残差的自由度越小，残差的自由度等于：样本量-1-方程的自由度。关于“自由度”，可以从字面上理解为可以自由的个数。试想，如果 a、b、c、d 四个变量的总和为 S，当总和 S 固定的时候，a、b、c、d 四个变量中只有三个可以自由取值；如果再把变量 a 取值固定，则 b、c、d 三个变量中只有两个可自由取值。在计算 t 检验时，自由度之所以为 N-1，就是如果所有人观测值的总和固定，只有 N-1 可以自由取值。

多元回归方程参与计算的样本是所有自变量（X）和应变量（Y）都完整的观测记录，任何一个变量缺失都将导致该记录被排除在外。

多元回归区别于多元分析。多元回归是多个 X 对一个 Y 的回归分析，多元分析是对多个变量的分析，易侕软件基本统计菜单里有专门的多元分析系列模块。关于多因素分析、多元分析名词的使用有些混乱。英文里 multivariable 与 multivariate 两词也常互换使用，翻译过来就更混乱了。在看到这些名词的时候，不要急于按自己的认识去理解，要对照上下文去理解作者所说的内容。一般来说，英文里的 multivariable 翻译成多因素分析，指的是多个 X 对一个 Y；multivariate 翻译成多元分析，是对多个 Y 的分析。

广义线性模型，Y 的分布类型可以是正态分布、二项分布、泊松分布等等，X 可以是连续性变量、两分类变量或多分类变量，都不要求是正态分布。但要求残差近似正态分布，残差即 f(Y)的预测值与观察值的差，而且希望残差的方差与预测值无关。易侕统计软件的广义线性模型模块，自动给出多元回归方程的残差 QQ 图、残差与预测值的散点图，如下图 7-4 所示。残差 QQ 图可以帮助判断残差是否接近正态分布。残差与预测值的散点图（图 7-5）可以帮助判断残差的方差是否与预测值有关。理想的情况是残差围绕“0”上下对称分布，而且其离散程度恒定，分布范围与预测值无关。相当于方差齐性。如果随预测值增加，残差分布范围偏向正值或偏向负值，或者分布范围越来越大或越来越小，都是不理想的，相当于方差不齐。如果 Y 本身非正态分布，如血铅水平在人群中是偏态分布的，回归方程的残差也是偏态或方差不齐（残差与预测值的散点图成喇叭形），可以尝试用对数转换或平方根转换等方法，将其转换为正态后再做分析，或者尝试用其它形式的联系函数如 log(Y)。

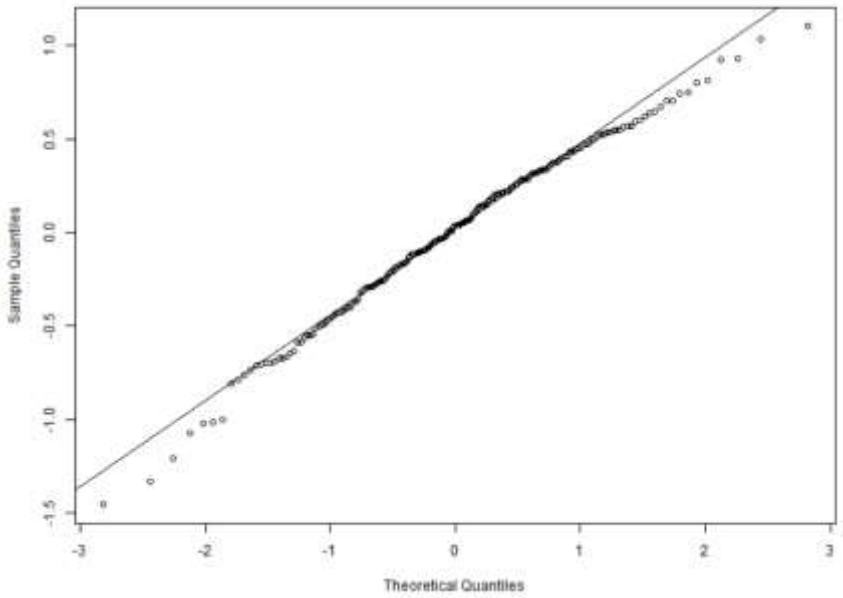

图 7-4 残差 QQ 图

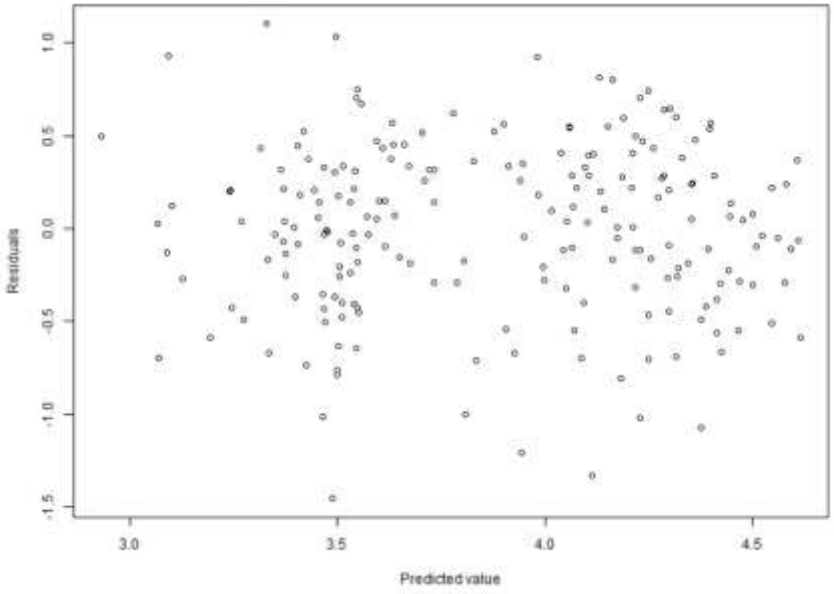

图 7-5 残差与预测值的散点图

第二节 回归方程与 t 检验、方差分析、卡方检验的关系

一、回归分析替代 t 检验：

如分析是否吸烟（SMOKE）对收缩压的影响，SMOKE 是 0/1 两分类的变量，0=不吸烟、1=吸烟。从两组中各随机抽取若干个体，测量收缩压（Y），数据如下图 7-6 所示，每个点到纵坐标的距离，表示所测收缩压的大小。

回归方程为：$Y=\beta_0+\beta_1 SMOKE$，当 SMOKE=0 时，$Y=\beta_0$；SMOKE=1 时，$Y=\beta_0+\beta_1$。β_0 是不吸烟者收缩压的均值，$\beta_0+\beta_1$ 是吸烟者收缩压的均值。β_1 反映了吸烟与不吸烟收缩压均值差，统计检验的目的就是检验 β_1 是否等于零。

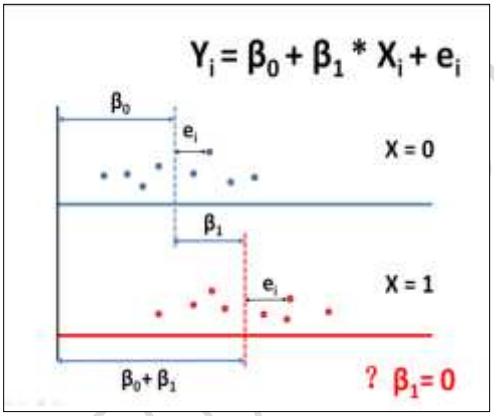

图 7-6 示意图

对回归系数 β_1 是否等于 0 的检验，等同于 t 检验比较两组均值差别是否显著。回归方程不仅给出两组均值差是否显著，而且给出差值大小及其 95%可信区间。

二、回归分析替代方差分析

当 X 分三组或多组，现在以三组为例。X 取值是 0、1、2。如果 X 是等级变量，假定等级间 Y 的差异相同，建立方程 I：

$Y=\beta_0+\beta_1 X$

从这个方程中可以看出：当 X=0 时，$Y=\beta_0$；当 X=1 时，$Y=\beta_0+\beta_1$；当 X=2 时，$Y=\beta_0+2\beta_1$。β_1 表示 X 每增加一个等级，Y 增加多少。

如果 X 等级间差异不同，或者当 X 不是等级指标，如职业：0 表示工人，1 表示农民，2 表示干部。这时 0、1、2 只是代码，没有等级关系，这个方程就不合适。这时就要产生 2 个哑变量，一个表示 X=1，另一个表示 X=2。建立方程 II:

$$Y = \beta_0 + \beta_1 (X=1) + \beta_2 (X=2)$$

从这个方程中可以看出：当 X=0 时，$Y=\beta_0$；当 X=1 时，$Y=\beta_0+\beta_1$；当 X=2 时，$Y=\beta_0+\beta_2$。β_1是 X=1 与 X=0 两组 Y 的差，β_2是 X=2 与 X=0 两组 Y 的差。对整个方程是否有统计学意义的检验，等同于方差分析，对β_1、β_2的检验相当于组间比较。

这里我们看到，用回归方程替代方差分析，比较多组均数，直接给出了组间比较的均数差，及其 95%可信区间。对于等级分组，按连续性变量分析，等同于趋势检验。

三、回归分析替代卡方检验

率（p）是阳性数除总数，取值范围在 0~1 之间。比值（odd）是阳性数除阴性数，取值范围是 0 到无穷大。比值与率的换算关系是，odd=p/(1-p)。逻辑回归，Y 的联系函数是比值的对数，其取值范围是负无穷大到正无穷大,方程表达式为:

$$logit(Y)= \beta_0 + \beta_1 X$$

其中 logit(Y) = log[p/(1-p)]，即比值的对数。假设 X=0、1 分别表示两组，p0、p1 分别表示两组的率

X=0 时，$logit[p0/(1-p0)] =\beta_0$

X=1 时，$logit[p1/(1-p1)] =\beta_0+\beta_1$

两式两边相减得：$Log[p1/(1-p1)] - Log[p0/(1-p0)] =\beta_1$

即 $\log([p_1/(1-p_1)]/[p_0/(1-p_0)]) = \beta_1$

其中，$[p_1/(1-p_1)] / [p_0/(1-p_0)]$ 即两组比值的比，简写为 OR。

因此，从回归方程中计算 e^{β_1}即为 OR。当率（p）较小时，两组率的比 p1/p0 （简写为 RR）与比值比（OR）非常接近。我们常说的危险比，通常指的就是比值比（OR），有时候也用率比（RR）。同理，三组（X=0、1、2）率的比较，回归方程为：

$$Logit(Y)= \beta_0 + \beta_1(X=1) + \beta_2(X=2)$$

e^{β_1}即为 X=1 组与 X=0 组相比的 OR，e^{β_2}即为 X=2 组与 X=0 组相比的 OR。如果把 X 按连续性变量分析，回归方程为：$Logit(Y)= \beta_0 + \beta_1 X$。这样得出来的 e^{β_1}即为 X 每增加一个等级导致的危险比。

第三节 回归系数：效应大小是核心结果

流行病学研究的是 X 与 Y 的联系，有没有联系只是一方面，最值得关注的是联系的强度，即 X 对 Y 的影响作用有多大。联系密切的程度不代表影响作用的大小，整天在一起的两个人相互影响作用不一定大于偶尔有联系的两个人。这也是为什么相关不能替代回归，相关性强不等于作用就大。流行病学上把 X 对 Y 的作用大小称为效应。这个效应的大小体现在回归分析上就是高尔顿最初提出的加权平均里的权重，也即回归系数。

一、回归系数的临床意义

初学者在做数据分析时总追求 P<0.05，因为统计上一般把 P<0.05 视为有意义，然后又错误地理解为有差别，最后变成 P 是否小于 0.05 就是是否有差别的标志了。正确的思维模式首先要把关注点从 P 值转移到β值，即回归系数及由其计算出来的其它形式的效应大小（如 OR 值）上来。下面我们来比较一项研究可能的分析结果，以帮助大家理解回归系数的重要意义。

如研究肥胖与高血压关系，下面列出四种可能的分析结果：

1) 胖子与瘦子相比，收缩压有显著差别，P<0.005。

2) 体重指数与收缩压，非常显著相关，P<0.0001。

3) 体重指数每增加 1kg，收缩压增加 0.01 mmHg，P<0.00001。

4) 在控制了其它因素的作用下，体重指数每增加 1kg，收缩压增加 1mmHg，95%可信区间 0.7-1.3mmHg。

上面四种可能的分析结果，哪个才有临床意义呢？答案是第 4 种。前 3 种结果统计上都非常显著，P 值很小，但都没有科研价值和临床应用价值。第 1 种结果只是说两组有差别，差多少？哪组高？都不知道。第 2 种结果只是说两者相关，是正相关还是负相关？相关性强又意味着什么？也不知道。第 3 种结果告诉我们，体重每增加 1 公斤，血压升高 0.01 mmHg。这提供了很重要的作用方向及作用大小信息，比前面两种结果内涵丰富多了，但尚有两个问题：（1）虽然 P 值很小，统计上非常显著，但体重能导致血压的变化幅度太小，没有临床应用价值；（2）所观察到这个作用大小 0.01mmHg 有没有其它因素的作用在里面混杂？不知道。第 4 种结果控制了其它因素的作用，得出的回归系数 1mmHg 是体重指数对收缩压的独立作用，有临床应用价值。

第 4 条结果的临床意义在于：门诊遇到一个高血压并肥胖的患者，医生根据这个研究结果，可以告诉患者仅降低体重这一项就能降低血压多少；如果病人又吸烟又饮酒，可根据相应的文献告诉他，如果戒烟又能降低血压多少，戒酒又能降多少。

比较这些结论，理解统计意义与临床意义的关系，从而理解如何提高一篇论文的科学价值。临床研究不只是要看 P 值是否小于 0.05，更要关注作用大小，特别是独立作用的大小。

从统计方法上看，第 1 种结果是两组比较的 t 检验得出的，第 2 种结果是相关分析得出的，第 3 种结果是回归分析得出的，第 4 种结果也是用回归分析得出的，只不过用的是多元回归分析。回归分

析，给出有临床意义的回归系数，而且可以控制其它因素，分析独立作用大小，掌握回归分析非常必要。

二、哪个因素更重要

临床科研入门的一道门槛是理解回归系数的意义，关注回归系数（即效应大小）的变化。初学者往往被一些统计术语迷惑，不能把关注点集中在回归系数或效应大小上。如 X 的重要性有多大？X 有没有意义，意义有多大？X 与 Y 的相关性强不强？如果有几个因素，最好能按相关性强度，或按 P 值意义大小排序，或按重要性将它们排序。回到生活常识中去分析这些想法，就很容易理解，所谓重要性也罢，意义也罢，是主观的东西。以新冠病毒做比喻，它百年不遇，却对我们的生活改变之大超乎想象，如果要排序，该把它排到第几位呢？

是否重要？哪个更重要？一定要有所指才能有所比，不能笼统地谈重要性大小，谈意义大小，否则就变成了诗词歌赋或抒情散文，而不是科研论文。如果我们把重要性指在效应大小上，把 X 的意义也指在效应大小上，把关注点集中在效应大小上，那就不会被这些名词误导。

前面已经说过，相关性强不代表作用大。统计上相关系数的计算是建立在方差分析的基础上的，方差反映的是变异程度。以 Y 为连续性变量为例，X 在变，Y 也在变。好比你读高中时与你的一个要好的同学一样，他做什么你做什么，整天形影不离，那就是相关性很强。当然不可能他每做一件事你都会跟着做，也就是说总会有 X 变 Y 不变或反向变的时候，所以相关系数不等于 1。尽管你们整天形影不离，但最终你可能选择了学医，他却选择了学文，而影响你学医的可能是另一个与你接触不密切的人，如果用相关与回归表达，那就是你与这个人的相关系数不大，但就决定学什么而言，他对你的回归系数超大。

统计上提到的变量的重要性，多数情况下也是根据方差分析结果而说的，在 Y 的总变异（方差）中，由某因素导致的变异占的比例最大就被认为是最重要的，这无可非议，但也只是从方差的角度来分析的，与回归系数是两码事。而统计上提到的意义的大小，通常是根据 P 值来说，P 值越小意义越大，然而这只是统计意义，反映的是一类错误的概率，更不是效应大小。

三、效应值的解读

X 和 Y 的数据类型不同，效应测量值的解释不同，最常见的几种类型如下表：

表 7-3 效应估计和结果变量与危险因素的类型

结果变量（Y）	危险因素（X）	效应测量	统计检验
连续性，如收缩压	分类型，如是否吸烟	吸烟者与不吸烟者收缩压的差（β）及其标准误	H_0: β=0
连续性，	连续性，	体重指数每增加一个单位收缩压增加多	H_0: β=0

如收缩压	如体重指数	少（β）及其 95% 可信区间	
分类型， 如是否高血压	分类型， 如是否吸烟	吸烟者与不吸烟者发生高血压风险比（OR）及其 95% 可信区间	H_0:OR=1
分类型， 如是否高血压	连续性， 如体重指数	体重指数每增加一个单位发生高血压风险比（OR）及其 95% 可信区间	H_0:OR=1

例如下表显示母亲年龄和苯暴露对新生儿出生体重的影响的回归分析结果：

表 7-4　母亲年龄和苯暴露对新生儿出生体重的影响的回归分析

X 变量	N	回归系数（95%可信区间）	P 值
母亲年龄			
20-25	164	Referent	
26-29	526	21 (-50 to 91)	0.559
39-40	102	167 (68 to 266)	0.001
笨暴露	354	-58 （-115 to -2）	0.044

研究者把母亲年龄分成 3 组，第 1 组 20-25 岁，有 164 人，这一组为参照组（Referent）；第 2 组 26-29 岁，有 526 人，回归系数是 21，95%可信区间-50 到 91，p 值 0.559，这个回归系数表示第 2 组比第 1 组出生体重高 21 克，但统计上差别不显著；第 3 组 30-40 岁，有 102 人，与第 1 组相比出生体重高 167 克，95%可信区间 68 到 266，p 值 0.001，有显著性差别。

苯暴露分两分类变量，0 表示无暴露，1 表示有暴露。有 354 人有苯暴露，回归系数-58，表示有暴露组比无暴露组出生体重低 58 克，95%可信区间-115 到-2，p 值 0.044。

第四节 多元回归方程如何调整混杂

一、调整的意义

首先看一个例子，比较吸烟与不吸烟的人的身高，SMOKE 表示是否吸烟 （0=否，1=是），Y 表示身高，单因素回归分析得出：

$$Y = 156.3 + 9.2 * SMOKE$$

即吸烟者比不吸烟者高 9.2cm，P 值小于 0.001。是否能下结论认为吸烟可增加身高呢？从常识上看，不能。那为什么会出现这种统计结果呢？进一步分析发现吸烟的人多是男性，不吸烟者多是女性，

这个方程得出的吸烟者与不吸烟者身高的差异，有很大一部分是因为男女身高的差异。要调整性别的影响，在方程中引进性别（Female）这个变量，Female=0 表示男性，Female=1 表示女性，方程变成：

$$Y = 174.7 + 1.5 * SMOKE - 10.3 * Female$$

吸烟的回归系数从 9.2 变为 1.5，P 值也不再显著了。性别 Female 的回归系数为-10.3，表示女性比男性低 10.3cm。这样一个二元回归方程，就把性别对身高的影响剥离出来，吸烟对身高的回归系数 1.5 就是调整了性别后的吸烟对身高的作用。回归方程就是这样调整混杂，去伪存真，评价所研究的 X 对 Y 的独立作用的。

如果要调整的变量是有序多分类变量，如文化程度，分低、中、高三组，如何放入方程中进行调整？如果是无序多分类变量，如职业，分工人、农民、教师、公务员等，如何调整？如果是连续性变量，如年龄、体重指数等，又该如何调整？

从哪里找答案？一个孩子在路灯下找钥匙，老人问她你的钥匙在哪里丢的，孩子说是在路灯后面的小路上，老人说，那你为什么不去小路上找？孩子回答说，小路上没有灯，漆黑一片，根本找不到，路灯这里亮堂看得清。

这个故事听起来很荒唐，其实仔细想想，我们身边很多人就像这个孩子一样，在一个看似明亮的地方找着跟本不存在的东西。上面提到的问题，还会延申出更多的问题，如多分类变量是否可以合并频数比较少的分组后再调整？连续性变量要不要检查是否符合正态分布，如不符合，要不要先转换成正态，再进入模型进行调整？为什么换一种方法将调整变量放入模型，X 就不显著了？可想而知，如果不明原理，问题是解决不尽的，一个问题解决了，另外三个问题又来。抓住要点了，这一切问题就都不存在了。要点在哪里？只有一句话，那就是调整的目的是什么？是把要调整的因素的作用剥离出去。怎么才算是剥离出去了呢？一句话就是要正确拟合每个调整因素对 Y 的作用。

我们所关注的危险因素 X 与众多混杂因素一起导致了结局的发生或结局指标的高低，调整的目的是把 X 的作用归给 X，把其它因素的作用归给其它因素，一一论功行赏，不能把其它因素的功劳或罪行记到 X 头上。就像一场大战后，要论功行赏一样，我们都希望公平，怎么才算公平呢？

二、如何调整多分类变量

如果混杂因素 Z 是多分类变量，假设分 A、B、C、D 四组，以 A 组为参照，B 的贡献与 A 相当，C 的贡献比 A 高 10，D 的贡献比 A 低 10，清清楚楚，大家都没有争议。如果你把 A、B 两组合并，称为 AB 组，说 C 组比 AB 组高 10，D 组比 AB 组低 10，也没有问题。但如果你把 C、D 两组合并呢？一高一低就抵消了，结果你说 CD 组与 A 组没有差别，与 B 组也没有差别，这时候 C 就肯定不干了。同理，如果你把 A、C 组合并，C 也会不乐意，毕竟人家的功劳比你大，你非要把它划到跟你一样，没有人愿意。因此，无序多分类变量是否可以合并频数较少的分组呢？是可以合并，但怎么合并就要先看它们的哑变量的回归系数，把回归系数接近的亚组合并到一起，它们不会有意见，否则就有不公平，不公平的结果是什么呢？该给某亚组的功劳没有给到位，就有可能被 X 贪功了。用专业术语讲就是有残余的混杂，导致对 X 的作用估计不正确。

如果 Z 是有序多分类变量，就像大哥、二哥、三哥、四弟一样，每人都相差 3 岁（等距），如果他们对 Y 的贡献也像年龄相差一样是等距的或接近等距的，你按排行行赏，如排行每低一次序少（或多）奖励 1 万元，那没有问题，这就是将 Z 按连续性变量放入模型中。但如果他们的贡献不是像年龄那样等距的呢，甚至是无序的呢，可能二哥贡献最大，四弟次之，三哥最差，那排行就没有意义了，应该当作无序多分类变量来处理。

三、如何调整连续性变量

如果 Z 是连续性变量，如年龄，可以理解为有序多分类的延申，只是分类很多很多而已，因为分类过多，每一组的人数只有 1 人，所以如果他们对 Y 的贡献，像他们的年龄那样是有次序的等距的就直接放在模型里，说年龄每增加 1 岁贡献增加或降低多少，这是无可非议的。但如果不是这样，就面临两种选择：

一是合并成几个年龄组，然后按多分类变量分析。怎么合并？因为是有序的，所以必须把相近的合并在一起，至于分几个组，切点切在哪儿，就依照组间差距越大越好，组内差异越小越好的原则。根据对 Y 的贡献大小，希望一个年龄组内的每一位对 Y 的贡献差不多，否则就成了吃大锅饭，干得好与干得不好一个样，就是不公平。

二是用曲线拟合调整连续性的 Z，因为是曲线，无法用一个或几个回归系数表达，需要的话贴个图，一图抵万言。

一个战斗英雄自 20 岁参军，屡立战功，为国家做出了很大的贡献，但不幸的是在他 40 岁的时候，因为一次战斗负伤，终身残废，离不开人照顾与药物维持，消耗了很多人力与资源，直到 60 岁去世。最后评价他对国家的贡献时，他 20 岁到 40 岁是正贡献，40 到 60 岁是负贡献，能说正负抵消，没有贡献吗？不能！一个连续性的混杂因素 Z 也可能如这个战斗英雄一样，它前面一段对 Y 的贡献，与后面一段对 Y 的贡献可能正好是反的，曲线拟合的结果就是一条 U 形曲线。如果粗糙地用一条直线拟合，就变成一条水平线，说这个 Z 对 Y 没有贡献，甚至不需要调整，这是不负责任的，就需要用曲线拟合，或分段拟合，或分成几个组拟合。易侕统计(EmpowerStats)软件曲线拟合和连续性变量分组模块，正是为此设计的。

调整的原则是尽可能做到把他人的功劳归给他人，尽量做到公平，该细致的时候要细致，该简单的时候要简单，细致的目的是为了公平，简单的目的也是为了公平。连续性变量如曲线拟合显示其与 f(Y)的关系不是直线，就不能粗糙地直接把原变量放入模型中，就需要考虑用曲线拟合或分组后按分类变量拟合，这就是该细致的时候要细致。无序多分类变量，可以合并的分组就应该合并；有序多分类变量如果哑变量的回归系数呈等级趋势，就应该用原变量直接放入模型（直线拟合）。因为这样可以提高模型的检验效率，这就是该简单的时候要简单。就像论功行赏一样，可以按集体行赏就按集体行赏，不必要纠缠于这个集体内谁贡献多那么一点点，谁少那么一点点。否则，不仅会产生矛盾，更重要的是无法实行。所谓检验效率即能检测出有显著差异的能力。回归模型每增加一个参数，残差的方差就会减少一点，但同时残差的自由度也减少了一个，检验效率与残差的方差和自由度有关。残差的方差越小效率越高，残差的自由度越高效率越高。

第五节 需要调整的混杂因素

要正确估计 X 对 Y 的独立作用，需要调整其它因素的混杂，哪些因素是混杂因素需要调整呢？要回答这个问题，不妨让我们先回顾一下流行病学病因通路模型，假设 Y 的发生有如下可能的通路：

1：X + A + B ==》Y

2：D + C + E ==》Y

3：X + F + B ==》Y

4：A + G + H ==》Y

X 的效应体现在通路 1 和 3 上，通过这两个通路发生 Y 的人数越多，X 的效应越大。调整的目的是使得有 X 组与无 X 组相比，它们通过通路 2 和 4 发生 Y 的人数相同。如果把 A、B、C、D、E、F、G、H 都调整了，换句话理解就是使这些因素在两组的水平是一样的，是不是可以了呢？当然是可以的！然而是不是必须呢？不一定。假设 D、C 两因素本来在有 X 组与无 X 组是均衡的，也就是说 D、C 与 X 无关，就没有必要调整。虽然理论上调整所有可能的混杂因素没有错，但可以想象的是这样一个大而全的方程里很可能有些因素不需要调整，把它们放在方程里，模型的参数多了，检验效率下降了。如果这个方程里 X 的作用只是接近显著，这时候把那些不需要调整的变量拿掉，X 的作用就可能显著了。

一、不能根据 P 值确定是否要调整

要调整的混杂因素首先是 Y 的危险因素，同时与 X 有关。这句话说起来简单，执行起来就不能那么简单了。譬如，什么叫与 X 有关？是不是 P 值小于 0.05 就是有关，否则就是无关呢？同理，如何判定是 Y 的危险因素，是根据 P 值来确定呢？有人说，P 小于 0.05 要求太严了点，放大一点改为 P 小于 0.10。不管改为多少，这种简单的一刀切做法总让人觉得不踏实。就好像给犯人量刑一样，贪污多少就是死罪，少一分就可以免了，这必然会激发人为了活命，想方设法甚至不惜弄虚作假把贪污额控制在这个切点之下。在数据分析时，我们会不会想方设法操控 P 值把不想调整的排除掉，把想调整的纳进来呢？会，这是人性决定的，所以我们感觉到不踏实。因此，不能简单地按 P 值来判断某因素是否与 X 有关，是否与 Y 有关，以此来决定要不要调整它。

正确的做法是首先根据现有的知识，阅读当前文献并结合临床实践经验，将所有与 Y 可能有关的因素列出来，然后分析这些因素与 X 的关系，再确定调整策略。这里用“策略”这个词，目的一是表达“不简单”，二是强调调整的目的。策略是为了达到一个战略目标而制定的，我们的目标是呈现 X 对 Y 的独立作用。

判断某因素是否要调整，不能简单地看 P 值，要看 X 的回归系数（效应值）。下面用一个实例，帮助理解如何确定调整混杂因素。下表 7-5 显示 X_1、X_2、X_3、X_4、X_5对结局变量 Y 的作用。

表 7-5 结局变量 Y 和各因素的回归分析

	单因素分析	多因素分析（方程一）
X_1	0.30 (-0.28, 0.87) 0.311	0.11 (-0.42, 0.65) 0.679
X_2	0.47 (0.36, 0.59) <0.001	0.47 (0.36, 0.58) <0.001
X_3	0.41 (0.13, 0.68) 0.004	0.28 (0.01, 0.55) 0.044
X_4	3.32 (0.37, 6.27) 0.028	2.30 (-0.59, 5.19) 0.119
X_5	5.22 (2.91, 7.53) <0.001	4.81 (2.60, 7.02) <0.001

注：表中数据为β（95% CI ）P 值

根据 P 值，X_1的 P 值最大，从方程一中剔除，得方程二，再根据 P 值把 X_4从方程中剔除，得方程三，如下表 7-6 所示：

表 7-6 结局变量 Y_2和各因素的多因素分析

	方程二	方程三
X_1	-	-
X_2	0.47 (0.36, 0.57) <0.001	0.47 (0.36, 0.58) <0.001
X_3	0.28 (0.01, 0.55) 0.046	0.36 (0.11, 0.61) 0.005
X_4	2.28 (-0.60, 5.17) 0.122	-
X_5	4.93 (2.80, 7.06) <0.001	4.77 (2.65, 6.90) <0.001

如果分析的目的，是确定 X_3对 Y 的作用，方程二回归系数是 0.28，方程三是 0.36，这两个值相差 20%以上。方程二调整了 X_4，更确切地反映了 X_3对 Y 的作用。P 值受样本量的影响，不能因为 X_4的 P 值不显著就不调整。如果分析的目的，是确定 X_2或 X_5对 Y 的作用，可用方程三，因为方程二与方程三相比 X_2、X_5的回归系数没有差别或差别不大，不影响对其效应的估计。

调整的目的是为了正确估计 X 的独立作用。因此，调整后 X 的回归系数的变化不大的因素可以不调整；反之，即使调整因素的 P 值比较大，也需要调整。

二、中介变量的调整

现在让我们假定 X 导致 Y 的发生过程中，有两种途径，一是直接导致 Y 的发生，二是通过一个中间过程 S，再导致 Y 的发生。如：

$$X \rightarrow Y$$

$$X \rightarrow S \rightarrow Y$$

这时，S 肯定是 Y 的危险因素，X 导致 S 导致 Y 的发生。如某因素可以导致宫内生长发育迟缓从而导致低出生体重，也可通过导致早产从而导致低出生体重。调整 S 的结果是什么呢？如果有无 X 两组，S 相同，那么就阻断了通过 S 导致 Y 的这条途径，估计的是第一条通路直接导致 Y 的作用。这就是为什么必要时要列出不同的调整方案，如调整 I、调整 II、调整 III。调整因素不同，结果的解释不同。至于哪些因素是这里所说的 S，那就要结合现有的知识分析其与 X 的关系。

三、共线性筛查

调整因素之间的关系也需要关注，如果某两个或多个调整因素之间相关性很强，或某一个因素是其它几个因素合成的，如 BMI 是由体重和身高计算得来的，不能把这些因素同时放入方程中，这就是共线性问题。在构建多元回归方程时，如果一组自变量存在共线性，即变量之间的相关性过强，或一个变量可由其它几个变量生成，可以使得模型估计失真。好比在一个人群中（所有的自变量）有几个人（共线性变量）拉帮结派搞小团体，左右集体决策，其他人（其余变量）的话语权可能被完全剥夺。

一个简单的方法来确定自变量之间的共线性是根据方差膨胀因子（VIF）来判断。要计算某自变量的 VIF，首先构建一个线性回归方程，用所有其它自变量解释该变量。如一组自变量：X1、X2、X3...，构建 X1 的线性回归方程：

$$X1 = X2 + X3 + ...$$

取该方程的 R^2（即方程可以解释 X1 变异的部分），$VIF=1/(1-R^2)$。同理，要计算 X2 的 VIF，首先构建方程：

$$X2 = X1 + X3 + ...$$

VIF 值越高，共线性越高。通常 VIF<10 是可以接受的，超过则应该从方程中剔除。

四、协变量检查与筛选

首先对所有自变量进行共线性筛查，剔除 VIF 大于等于 10 的协变量。然后逐个按以下步骤检查和筛选。

在分析 X 对 Y 的作用时，是否要调整变量“Z” 呢？

步骤一：先看“Z”与 Y 有没有联系，用单因素分析，看“Z”的 P 值。

$$Y = \beta_0 + \beta_1 Z$$

步骤二：再看调整“Z”与不调整“Z”，X 对 Y 的作用是否有变化。先运行基本模型，记录β_1，再在该模型中加入“Z”，看β_1变化多大？

基本模型：$Y = \beta_0 + \beta_1 X$

基本模型中引进 Z：$Y = \beta_0 + \beta_1 X + \beta_2 Z$

基本模型中可加入一些必须要调整的变量。

步骤三：再运行一个完整的模型，即调整所有可能的因素，然后从模型中剔除“Z”，看 X 的回归系数β_1的变化。

完整模型：$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 A_2 + \cdots\cdots$

完整模型中剔除 Z：$Y = \beta_0 + \beta_1 X + \beta_3 A_2 + \cdots\cdots$

按照上述思路，比较不同的模型，观察 X 的回归系数的变化，确定哪些因素要调整，工作量很大。而且特别要注意的是：因为可能调整因素 Z 有缺失值，导致调整 Z 的模型比不调整 Z 的模型所用的样本量小。而比较两个模型所用的样本要求是一样的，否则不可比。需要先要限定分析样本，然后才能合理比较。易侕协变量检查与筛选模块即为实现上述过程而设计的。

简单地采用逐步回归法筛选变量确立多元回归方程，分析 X 的独立作用是不合理的。首先，逐步回归是根据 P 值判断是否纳入方程的，而 P 值受样本量影响，应该调整的因素可能因为 P 值超过设定的界限而被排除在外（如前所述），这样一来，同一个变量在小样本里可能不需要调整，而在大样本里需要调整，这就造成小样本的研究结果不可信，这显然是不合理的。

五、协变量的临床意义最重要

构建多元回归模型评价 X 对 Y 的独立作用大小时，要调整哪些协变量不只是统计问题，更是一个临床问题。这里，我们用“协变量”代表与 Y 有关的变量，区别于“混杂变量”。混杂是既与 Y 有关又与 X 有关的变量，调整混杂才能正确估计 X 对 Y 的作用大小。调整只与 Y 有关的协变量呢？一般来说不会改变 X 的回归系数，但可以减小模型的残差。如果某因素对 Y 的作用比较强，虽然它与 X 无关，不影响 X 的回归系数，但把它放在方程中，可大大降低模型的残差的方差。残差的方差越小，模型越敏感，也就是模型能检测到 X 有作用的能力（检验效率）越高，具体体现是 X 的回归系数的标准误变小，X 对 Y 的效应的 95%可信区间变窄，P 值更显著。

混杂变量是一定要调整的，某因素构成混杂的条件是其与 X 和 Y 均有关。这里的关键是什么叫有关？什么样的关系才需要调整？什么样的关系又不能调整？这需要结合专业知识和常识来思考。

简单地说 Y 的下游变量不能调整，X 的下游变量调整与否的结果解释不同。所谓“下游”指的是有其引起的。如 X→S→Y→D，S 是 X 的下游变量，D 是 Y 的下游变量。在这条通路中，调整 S 就阻断了 X→S→Y 这条通路，得出来的是通过其它 X→Y 的通路发生的作用。D 是 Y 的下游，如果 D 固定的话，Y 也就固定了，把末端堵死了，这条通路也就不通了。所以 Y 的下游变量不能调整。

Y 的下游变量不能调整，还有一个原因是，如果调整了 X、Y 共同的下游变量，有可能导致本来 X 和 Y 没有联系变得有联系。例如，在成年人中，身高（X）与慢性咳嗽（Y）常识上分析应该是没有联系。然而，身高与肺功能 FEV1（一秒肺活量）有正相关，身高高的人 FEV1 高；慢性咳嗽与 FEV1 也有关，慢性咳嗽导致 FEV1 降低。FEV1 可以说是身高的下游变量，也是慢性咳嗽的下游变量。如果调整了 FEV1，就会发现身高与慢性咳嗽有联系了，身高高的人发生慢性咳嗽的风险高。用常识不难理解这个问题，如果 X→Z，Y→Z，可用一简单表达式 Z=X+Y 表示，本来 X 与 Y 没有关系，但如果固定 Z 了，则 X 与 Y 就有关了。

进一步不难理解如果调整了 X 的上游变量会出现什么结果，如 Z→X→Y，X 和 Z 同时在方程里，X 的回归系数代表 Z 固定的时候 X 对 Y 的作用，如果 X 完全是有 Z 决定的，在共线性筛查的时候，Z 就会被除掉；如果还有另外的通路，如 U→X→Y，U 没有调整，则得出的是 U 通过 X 导致 Y 的作用，而不是 X 对 Y 的作用。

说到这里，希望大家进一步理解明确 X、Y、Z（调整变量）之间的时相关系的重要性，队列研究和 RCT 的优势就在于此。在横断面研究中，能出现诸如上例在研究身高与慢性咳嗽的联系中，错误调整肺功能的问题，因为我们无法确定慢性咳嗽发生时间和 FEV1 发生改变的时间。如果是在 Y 发生之前测量的 Z，不可能是 Y 的下游变量。试想，如果是一个队列研究，刚开始时所有人都没有慢性咳嗽（Y=0），测量基线水平的 FEV1 和身高，最后分析基线的身高与发生慢性咳嗽的关系时，调整基线的 FEV1 是没有问题的，因为基线的 FEV1 不可能受后面发生的慢性咳嗽影响。

确定协变量与 X、Y 的先后关系和上下游关系，不是统计问题，是临床问题，需要结合医学专业知识来分析，对调整后的结果解读更离不开临床意义，这不是统计知识所能替代的。

通常一篇文章既要呈现调整又要呈现未调整的模型，目的是可以比较调整与未调整的回归系数，判断是否有混杂及混杂导致的 X 回归系数的变化。如果调整的模型 X 的回归系数变化较大，通常我们需要进一步分析清楚主要是调整哪（几）个协变量所起的作用。混杂因素本身也是 Y 的危险因素，通过这个过程揪出重要的混杂因素，进一步通过分层分析是否既有混杂又有交互作用。

分析目的是确定 X 与 Y 有没有联系？X 对 Y 的独立作用估计是多少？注意这里用到估计两字，为什么？因为总体真值是不可知的。如果我们用高度精密的仪器测量一把尺子多次，就会发现几次测出来的长度不同。连一把尺子的长度都测不准，更何况是人的身高、体重、血压、X 的效应，等等，这

些指标还不像尺子那样，这些指标本身就在不断变化。然而，这种测不准不等于我们就可以不去追求真理。我们需要做的是尽我们现有的认知水平和工具尽可能地接近真理。

第六节 缺失变量的处理

在使用多元回归方程拟合数据前，需要检查数据缺失情况。通常在危险因素研究中，X 与 Y 缺失的情况我们比较容易掌握，而调整变量的缺失情况常被忽略。当调整变量比较多时，或有几套不同的调整方案时，因调整变量缺失，方程用到的样本量差异可能很大，因此我们在比较不同的模型时，需要特别关注模型所用到的样本量。易侕统计软件也会自动报告每个方程所用到的样本量。

在分析 X 对 Y 的独立作用时，如果发现某变量（Zi）缺失比较多，而该变量又必须在方程中调整，怎么办呢？

一、多重插补

插补相当于用现有数据预测缺失数据（Zi），插入预测值。如何预测？相当于根据当前的样本数据构建一个预测方程，用未缺失的变量预测缺失变量。具体操作时，我们给出一串变量（Vs），要包含无缺失的变量和待插补的变量。插补程序自动根据给定的变量构建预测模型，对缺失数据进行插补，输出完整数据。在做多重插补时，尽可能多的放入一些无缺失或缺失比较少的变量，以提高插补效果。需要注意的是：

1. X 和 Y 的缺失不宜插补，因为我们不能用插补的数据分析 X 与 Y 的关联关系。

2. 给定的变量（Vs）里通常不应包含 Y。因为插补的基本原理是通过预测方程由未缺失数据预测缺失变量，我们不能用 Y 预测一个缺失的 Zi，然后又用这个 Zi 来对 Y 进行回归。这样做给人以“监守自盗”的感觉。

3. 通常我们需要通过插补生成 5 套（或更多）数据，然后分别用每套数据构建模型，最后综合来自各套数据的模型计算 X 的回归系数。不能只报告一套插补数据的回归结果，否则会低估 X 的回归系数的标准误。易侕软件有专门的“缺失数据多重插补”模块和“合并多套数据回归系数”模块。

二、引进缺失哑变量

如果缺失变量（Zi）是分类型变量，添加一类“缺失”，如性别，原来编码为 0=男、1=女，现在添加一类：2=缺失，然后按无序分类变量处理；

如果缺失变量（Zi）是连续性变量，生成是否缺失哑变量 Zim，Zim=0 表示 Zi 不缺失，Zim=1 表示缺失。然后将 Zi 缺失的数据赋值为 0，用新变量 Zia 表示，把 Zia 和 Zim 同时放入方程中。其原理如方程：

方程 1： $f(Y)= \beta_0 + \beta_1 Zia + \beta_2 Zim + \cdots\cdots$

这时：

当 Zi 缺失时，Zia=0，Zim=1，f(Y)= β_0 + β_2

当 Zi 不缺失时，Zia=Zi，Zim=0，f(Y)= β_0 + β_1 Zia 。

再看看处理前，直接用 Zi 拟合，方程为：

方程 2： f(Y)= β_0 + β_{10} Zi + ……

方程 1 中的 β_1 与方程 2 中的 β_{10} 完全相同，都表示 Zi 每增加一个单位 f(Y)改变多少。然而方程 2 只用到 Zi 不缺失的记录。而方程 1 因 Zi 缺失都赋值 0，Zi 与 Zim 都不缺失，用到所有的记录。方程 1 中的 β_0 + β_2 代表了 Zi 缺失者 f(Y)的平均值， β_2就是该平均值与回归方程截距的差值。

同理，如果 Zi 是分类变量，缺失组哑变量的回归系数代表缺失组 f(Y)的均值与参照组的差。

从上述的方程 1 中，我们应该注意到对缺失变量（Zi）缺失的记录，我们无法知道其 Zi 值是多少，只是用一个统一回归系数，来代表每个缺失的观测对象 Zi 的作用。实际上对 Zi 缺失者来说，他们的 Zi 一定是有差异的，其对 Y 的作用一定不一样，对这种作用我们无法有区别地进行调整。引进缺失哑变量进入模型的结果只是让缺失记录都参与了分析，增加了残差的自由度，提高了模型的对其它变量的统计检验效率。

第七节 回归分析应用实例

曾有这样一个求助：诊断工具是两个连续变量，想知道（1）用 Logistc 回归将两个连续变量生成联合变量；（2）分别为每个变量找一个切点生成两分类变量，然后采用并联或串联的方法。比较三种方法的 AUC，如何提高联合的诊断效能？

一、关于切点的常识性思考

首先，连续性变量蕴含的信息多于分类型变量，用连续性变量去预测应该更好。

以高血压为例，SBP（收缩压）达 140mmHg 或 DBP 达 90mmHg 就诊断为高血压。是否高血压是两个状态，数字表达为 1=是，0=否。如果我们抛开统计，用常识去思考这几个问题：

1. 是否高血压这两种状态真的如 0 与 1 一样是两种绝然不同的状态，中间没有过度带吗？
2. SBP 的切点可以是 141mmHg 或 139mmHg 吗？ DBP 的切点可以是 91mmHg 或 89mmHg 吗？
3. SBP 140mmHg 的人真的就有病？139mmHg 的人真的就正常吗？

万物都在变化中，而变化都有一个从量变到质变的过程。就某一个人来说，其血压导致健康状况的变化也是这样。就一个人群来说，虽然有的人 SBP 是 140 仍然健康，有的人 139 就表现出疾病状态，

但总体来说，SBP 是 140 的人一定比 139 的健康状态差的比例高一点，141 的比 140 的又差一点。理解了这一点，就能理解连续性的 SBP 与 DBP 的实际测量值，蕴含的信息一定比 SBP 是否大于等于 140，DBP 是否大于等于 90 的两分类变量多。

这样一来，为两个连续性变量各找一个最佳切点，生成两个两分类变量，然后用串联或并联的方法联合，一定不如用 logistic 回归方程直接将两个连续性变量联合好。

二、回归方程解读串联和并联

再看串联与并联问题，都可以用多元回归方程来替代。看两个自变量的简单方程：

方程 1 logit（Y）= a + b1*X1 + b2*X2。

如果把 X1 与 X2 按切点转换成两分类变量放在方程里，方程 1 就完全替代了两个两分类变量的串联或并联。因为：

- 当 X1=1，X2=1 时，方程右边等于 a + b1+b2，以这个值为切点，就是串联，也就是 X1、X2 都等于 1 时判断 Y=1，否则 Y 等于 0。
- 当 X1=0，X2=0 时，方程右边等于 a，以这个值为切点，就是并联，也就是 X1、X2 都等于 0 时判断 Y=0，否则 Y 等于 1。

比较是 1 个变量 X1 好，还是 X2 好，还是 X1+X2 好，无非是三个回归方程（只用 X1，只用 X2,用 X1+X2）的比较，也就是对 X1 与 X2 的回归系数的显著性的检验。因此，如果完全理解了回归方程，串联或并联也都尽收眼底。

三、如何优化预测方程

如果保持连续性变量在方程中，毕竟连续性变量信息量多，应该说用连续性变量一定不会差。然而在构建方程时需要考虑到：如果 X1、X2 与 logit（Y）不是直线性关系时，直接放在方程里，有可能还不如以两分类变量的方式放入方程中。两分类的 X1 与 X2 的方程的拟合优度反而更好。这就回到了如何构建回归方程，以达到最大拟合优度这个问题上来了。方程的拟合优度的另外一种表现形式就是 AUC。

首先我们要做 X1、X2 对 Y 的平滑曲线拟合，根据曲线关系图，拓展上述方程。如果有分段式的效应可以拓展到分段线性模型（piecewise regression），如果呈抛物线式的可以考虑添加 X1、X2 的平方项。再进一步，看 X1 与 X2 有没有交互作用，如果发现有交互作用，在方程中引进交互作用项，进一步提高拟合优度。如果不管怎么拟合，X1 与 X2 中都有一个没有必要放入方程中，那也就说明联合两指标没有必要。

总而言之，只要掌握了回归方程的原理，掌握了如何构建回归方程，用 logistic 回归方程就可以最好地回答是否有必要联合两指标？如何联合两指标？以达到最优诊断效能。