

第十七章 多元线性回归

在第十六章一元线性回归中，我们讨论了回归分析中只有一个自变量的情况，也就是一元回归分析，它只能处理两个变量之间的关系。然而在实际研究中，变量之间的关系十分错综复杂，一个因变量往往会同时受到多个自变量的影响。也就是说，因变量需要通过多个自变量来综合解释或预测。例如：学生的学业成绩除了受智商的影响，还受到学习兴趣、学习需要用到多元回归分析。

★在教育与心理学研究中，多元回归分析常用来回答以下四类研究问题：

- (1) 验证自变量对因变量的影响；
- (2) 检验单个或一组自变量的重要性；
- (3) 建立多个自变量对因变量的预测模型；
- (4) 分析多个自变量之间的交互作用。

这一章我们将在一元线性回归分析的基础上，简要介绍多元线性回归分析。

第一节 概述

多元线性回归 (multiple linear regression)：通过建立自变量与因变量之间的回归模型，考察一个因变量与多个自变量之间线性关系的统计分析过程，称为多元线性回归。

目前，在教育心理、社会心理等方面的研究大部分是多因素设计的研究。在这些研究中，多元回归分析的应用更加广泛和深入。例如，当研究人员希望了解哪些因素会影响大学生的学业成绩时，就可以使用多元回归方法来进行分析。首先，研究人员需要根据已有研究资料或理论假设来确定自变量和因变量。在这个例子中，因变量确定为学生学业成绩，自变量包括同伴关系、教师期望、学习能力以及自尊水平。研究人员需要建立回归方程，并通过实测数据来求解模型中的各个参数（回归系数），评价回归模型是否能很好地拟合数据，还要检验各个自变量的作用是否符合预先的假设。如果模型能够很好地拟合实测数据，则可以运用此模型进行预测。

对于初学者，有时会分不清回归与相关的区别，认为多元回归只是将多个变量的相关换了一种呈现形式。而事实并非如此。回归分析重在探讨自变量对因变量变异的解释和预测，对变量的关系有预先指定的明确的方向性，自变量和因变量不能轻易交换位置。多元回归还可以检验多个自变量和因变量之间更为复杂的关系。而相关主要衡量变量之间是否存在关系，以及关系的强弱。即使是考察多个变量之间的相关，如果使用零阶相关（又称简单相关），那么也仅能展现不考虑其他变量的影响时变量两两之间的简单关系。而且变量的地位是平等的，没有使用一个去预测另一个的假设。但回归与相关也非完全割裂，零阶相关是选择自变

量进入回归方程的一个重要的参考指标。本质上，多元回归的基础是变量之间的相关。如果所有变量之间都毫无关系，回归分析也就失去了意义。本章会在多元分析的视角下，介绍一些新的相关系数（如，偏相关和半偏相关），它们可在考虑其他变量影响的同时估计自变量和因变量之间的关系。

一、多元线性回归方程的建立

多元线性回归是分析一个变量与多个变量之间线性关系的最基本的方法，它的模型可以表示为：

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_k X_{ki} + \varepsilon_i \quad (17-1)$$

式中， $X_{1i}, X_{2i}, \cdots, X_{ki}$ 分别是个体 i 在自变量 $X_1, X_2, \cdots, X_k$ 的观测值， Y_i 为是个体 i 在因变量 Y 上的观测值， ε_i 为误差，假设服从均值为 0，方差为 σ^2 的正态分布。 $b_1, b_2, \cdots, b_k$ 为偏回归系数。

在多元线性回归中可以用多个自变量的最优组合来共同预测因变量，因变量的预测值可以表示为：

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_k X_{ki} \quad (17-2)$$

偏回归系数 (partial regression coefficient)： 其它自变量保持不变时，相应自变量的变化所引起的因变量的变化量。

以消费者对商品的需求 (Y) 受商品价格 (X_1) 和消费者收入 (X_2) 影响的二元（两个自变量）线性回归方程为例，回归方程可以表示为：

$$\hat{Y} = a + b_1 X_1 + b_2 X_2 \quad (17-3)$$

式中， $\hat{Y}$ 为自变量 X_1 和 X_2 组合的共同预测值， a 为截距， b_1 和 b_2 分别为自变量 X_1 商品价格和 X_2 消费者收入对应的偏回归系数。按照上文中偏回归系数的定义，在这个二元线性回归方程中，当消费者收入固定不变时，商品价格每变化一个单位，消费者对商品的需求平均变化 b_1 个单位；同理，当商品价格固定不变时，消费者收入每变化一个单位，则消费者对商品的需求平均变化 b_2 个单位。

此外，对于多元回归方程中的参数，估计的方法与第十六章中的一元回归方程相同，都是利用最小二乘法使得残差平方和取得最小值，在此不再赘述回归系数的求解过程，直接以二元线性回归为例，给出回归系数的求解公式。

在二元线性回归中，回归系数 b_1, b_2 和回归截距 a 可以通过下面的公式得到：

$$b_1 = \frac{SP_{X_1 Y} SS_{X_2} - SP_{X_2 Y} SP_{X_1 X_2}}{SS_{X_1 X_1} SS_{X_2 X_2} - SP_{X_1 X_2}^2} = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2} \quad (17-4)$$

$$b_2 = \frac{SP_{X_2 Y} SS_{X_1} - SP_{X_1 Y} SP_{X_1 X_2}}{SS_{X_1} SS_{X_2} - SP_{X_1 X_2}^2} = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2} \quad (17-5)$$

(分母同上)

$$a = \bar{Y} - b_1 \bar{X}_1 - b_2 \bar{X}_2 \quad (17-6)$$

式中， SP_{pq} 表示 $\sum (p - \bar{p})(q - \bar{q})$ ， SS_r 表示 $\sum (r - \bar{r})^2$ 。

$$a = b_0 = \bar{Y} - b_1 \bar{X}_1 - b_2 \bar{X}_2$$

二、多元线性回归的基本假设

使用多元回归分析的前提假设与一元回归分析的基本相似，包括：

(1) 各个自变量均与因变量 Y 呈线性关系，且自变量不包含测量误差， Y 为连续等距的变量。

(2) 误差 ε_i 服从均值为 0 的正态分布。

★(3) 误差 ε_i 满足方差齐性的假设，即无论自变量 X_i 的取值大小，误差 ε_i 的方差相等，均为 σ^2 。

(4) 误差之间相互独立的假设，即不同个体对应的误差相互独立。

(5) 误差与各个自变量之间相互独立，即误差的大小与自变量取值大小没有关系。

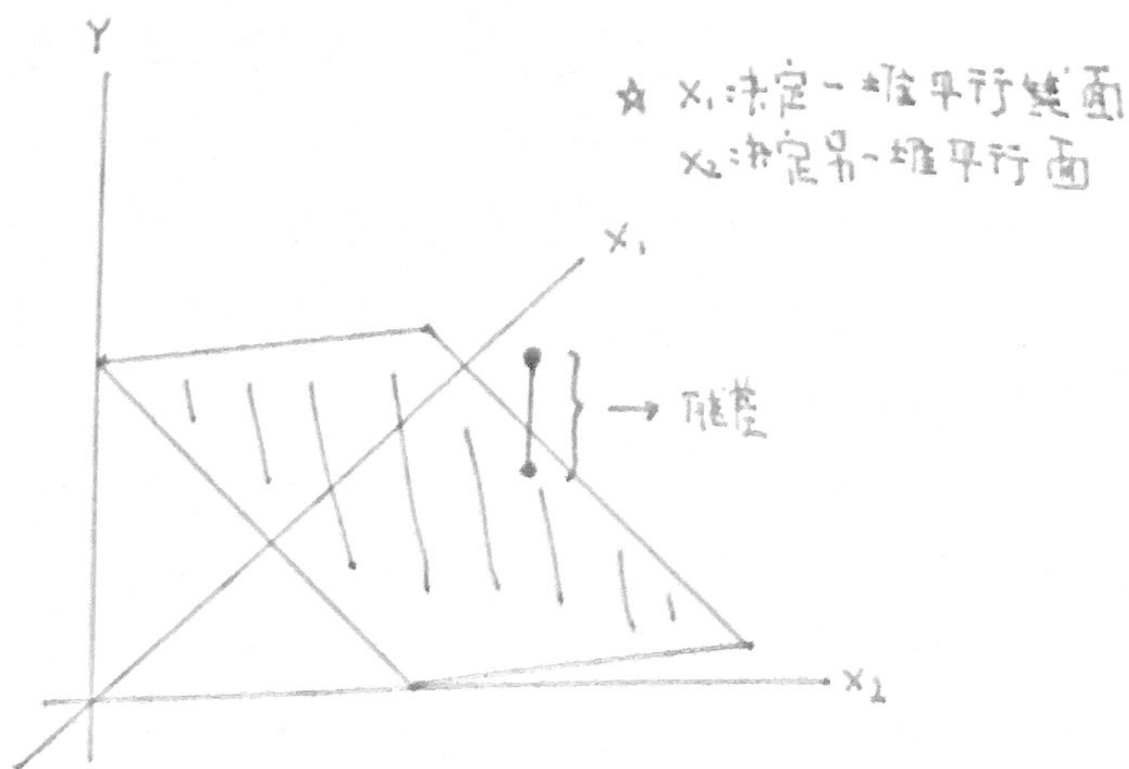
(6) 多个自变量之间不存在多重共线性，即多个自变量之间不存在过高的相关。我们将在第五节讨论多重共线性的相关议题。

当满足这六个假设时，可建立多元回归方程，探究因变量与多个自变量之间的相关关系，并可根据已知的自变量，对因变量进行解释和预测。

第二节 多元回归的类型

在多元线性回归分析中，根据自变量进入回归模型的方式，常见的有三种主要类型：标准多元回归、序列（分层）回归和统计（逐步）回归。三种回归的区别主要体现在如何解释自变量之间因相关而产生的重叠，以及按照怎样的规则决定自变量进入的顺序。

本节将使用韦恩图表示变量之间的关系。以图 17-1 为例，共有三个自变量（ X_1 、 X_2 和 X_3 ）和一个因变量（ Y ）。 X_1 、 X_2 和 X_3 共同解释 Y 的部分为 $a + b + c + d + e$ ，其中三个自变量单独解释的部分分别是 a 、 c 和 e ， X_1 和 X_2 共同解释的部分为 b ， X_2 和 X_3 共同解释的部分为 d ， X_1 和 X_3 没有共同解释的部分，因此没有重叠的部分。



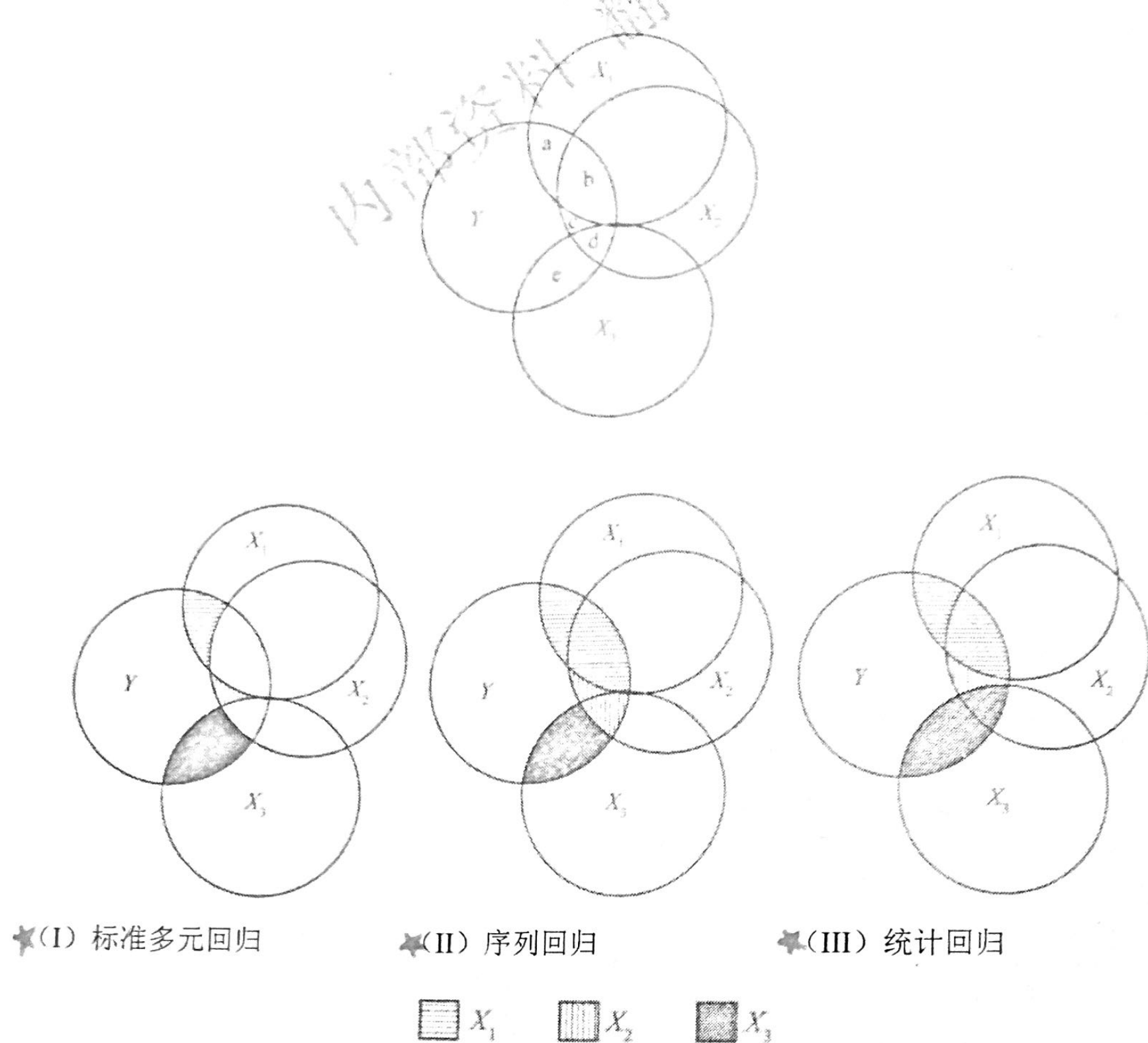


图 17-1 三种回归类型解释因变量方差的韦恩图

一、标准多元回归

在标准多元回归 (standard multiple regression) 中, 所有自变量同时进入回归方程, 故又称为同时回归 (simultaneous regression)。标准多元回归度量了每个自变量进入方程后所增加的对因变量变异解释的贡献。换言之, 标准多元回归在计算单个自变量的贡献时, 该自变量与其它所有自变量共同解释的部分都被排除, 仅计算剩余的可解释的部分。正是由于这一点, 在标准多元回归中, 所有重叠的部分将不计入任何自变量的贡献。

如图 17-1 (I) 所示, 排除和其他自变量共同解释 Y 的部分 b 和 d 后, X_1 、 X_2 和 X_3 单独解释 Y 的部分依次为 a 、 c 和 e , 这也是 X_1 、 X_2 和 X_3 在标准多元回归中各自对因变量 Y 解释的最终贡献, 而 b 和 d 没有被纳入该回归方程可解释的范围。

在标准多元回归中, 类似 X_2 这样与其他自变量和因变量相关较高的自变量, 与其他自变量一同进入方程后, 有可能会得到与理论预期相反的结果。因此去掉其它变量解释过的部分后, 该自变量独特的贡献会变得很低。正因如此, 当自变量之间相关较高时, 对标准多元回归的结果解释应当更为谨慎, 除了每个自变量单独的贡献, 还需考虑变量间的相关。

二、序列回归

序列回归 (sequential regression), 又称分层回归 (hierachical regression)。在序列回归中,

自变量将根据研究者指定的顺序依次进入回归方程，回归分析中根据定义的变量顺序，作多次回归分析，每次都在前面进入变量的基础上增加一个（或一组）新的变量重新进行回归分析。由于存在前后顺序，衡量一个（或一组）自变量的贡献时，与其它变量共同解释的部分会归为先进入的变量。（主动选择）（一般把最重要的预测变量放在最后）

如图 17-1 (II) 所示，假设研究者指定 X_1 首先进入， X_2 其次进入， X_3 最后进入。在进行序列回归分析时，会按照指定的顺序，作三次回归分析。第一个分析的模型中只包含变量 X_1 ；第二个分析的模型中包含变量 X_1 和 X_2 ；第三个分析的模型中包含变量 X_1 、 X_2 和 X_3 。衡量自变量贡献时，由于 X_1 进入时方程中没有其它自变量，它与其它变量重叠的部分都被“占为己有”，故 X_1 解释的部分为 a 和 b ； X_2 进入时，与 X_1 重叠的部分 b 已经被归为 X_1 ，仅计算 X_2 剩余可解释的部分，即为 c 和 d ； X_3 最后进入，它与之前所有进入的自变量相互重叠的部分都已被“瓜分”，可增加解释的部分仅为 e 。可以看出，与标准回归不同，序列回归将重叠部分按一定顺序全部分配给了各自变量，相比于标准多元回归，在该例中 X_2 解释的部分大大提高。

研究者一般会出于逻辑和理论的考量指定自变量进入方程的顺序。一种可能是让最具理论重要性的自变量优先进入方程，以保证最重要的变量能最大地解释因变量的变异。例如，预测篮球运动员成绩时让训练时间第一个进入回归方程；相反，另一种可能是让最重要的变量最后进入方程，而让控制变量或研究者不关心的变量最先进入。这样就相当于在控制其他变量的情况下考察最感兴趣的变量对因变量解释和预测的贡献大小。这也是大多数研究者应用序列回归的意义：控制了之前加入变量的影响后，考察新加入变量是否有显著影响。在实际应用序列回归时，一次可以进入一组自变量，即每层中可同时估计多个自变量，此时对于该层来说相当于一个标准多元回归，不同层之间按照序列回归的方式解释结果。

三、统计回归

统计回归 (statistical regression) 完全以统计标准决定自变量进入回归方程的顺序，没有考虑变量的意义和理论解释的问题，是一种数据驱动的回归类型。一个变量进入或被排除出方程完全根据该样本下计算出的统计指标，一些细微的差异将会对衡量变量重要性造成较大的影响。

如图 17-1 (III) 所示，由于 X_1 与 Y 有最高的相关，故最先进入。由此，首先解释了 a 和 b 两个部分。然后，在此基础上比较 X_2 和 X_3 ：如果 X_2 进入，则可解释的部分为 c 和 d ；如果 X_3 进入，则为 d 和 e 。由于 e 的面积大于 c ，此时让 X_3 进入对预测 Y 的准确性提升将更为明显，故 X_3 第二个进入方程。 X_2 最后进入时可解释的部分仅剩 c 。可以发现，统计回归可能会使与因变量有较高相关的自变量最后进入回归方程，并得到其贡献最小的结论。因此研究者使用统计回归时也应关注变量间的相关，不然解释结果时很容易出错。

统计回归有三种常用的选择变量是否包含在回归模型中的方式：向前选择、向后选择和逐步选择。向前选择 (forward selection) 的作法是：从不含自变量的回归方程开始，依次加

入满足显著性标准的自变量，直到没有自变量再达到这一标准。向后选择 (backward selection) 的作法是：首先将所有的自变量都加入回归方程，逐渐剔除那些对因变量贡献不显著的自变量。逐步回归 (stepwise regression) 是向前选择和向后选择的结合。它的做法是：从不含自变量的回归方程开始，每次加入一个对预测因变量贡献最大的自变量，加入新变量后检验方程已有自变量中是否有预测不显著的，如果有便将其剔除。如此反复进行，直到没有变量可以进入或剔除。

由于统计回归可选出对预测因变量最有用的一组自变量，如果研究者重在关注回归方程的预测性，可使用此类回归。另外，统计回归还常用于自变量 (预测变量) 较多的探索性分析中。

四、三种回归的比较

从使用目的来看，统计回归更侧重于数据驱动的探索和建立模型，因此，当对因变量与多个自变量的关系了解不多时，可使用统计回归。标准多元回归常用于在对变量影响顺序没有事先假设的情况下，同时分析一组自变量对因变量的影响，其重点在于解释所有自变量对因变量整体的影响，或者在控制其他自变量的条件下单个自变量对因变量的影响。序列回归一般是在研究者对模型有理论支持或验证假设时使用，常用来分析新增加的一个 (或一组) 自变量对因变量的单独贡献。

从变量的进入顺序和过程的可控性来看，标准回归由于所有变量一次性进入而变得完全不可控，统计回归由于完全由样本数据计算的统计指标决定各个自变量的取舍而同样不可控，只有序列回归是由研究者决定变量的进入顺序，是最为可控的回归类型。

正是由于上述差别，三种回归有各自的独特之处：标准回归会剔除所有自变量的重叠部分，可以体现出每个自变量的单独贡献。序列回归可以在某些变量进入方程的前提下探讨另一些变量的贡献。统计回顾可以辨别具有多重共线性的变量，可为未来的研究剔除某些明显冗余的变量。实际应用中，研究者应根据拟解决的研究问题选用合适的方法。

第三节 自变量的重要性

一、多元测定系数

正如一元线性回归，在多元回归分析中，因变量 Y 变异的总平方和可以分解为两部分：一部分是回归平方和，它是回归方程所能解释的部分，反映了由于 $X_1, X_2, \dots, X_k$ 和 Y 的线性关系而导致的 Y 的变异。另一部分是残差平方和，它是由于观察值没有落在回归线上而引起的，是包含 $X_1, X_2, \dots, X_k$ 的回归方程没能解释的 Y 的变异部分。在多元线性回归中，将回归平方和与总平方和的比值称为多元测定系数 (coefficient of multiple determination)，用 R^2 表示，计算公式为：

$$R^2 = \frac{\sum(\hat{Y} - \bar{Y})^2}{\sum(Y - \bar{Y})^2} = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T} = \frac{b_1 SS_{YX_1} + b_2 SS_{YX_2}}{SS_{YY}} \quad (17-7)$$

式中, $\Sigma(Y - \bar{Y})^2$ 为总平方和, $\Sigma(\hat{Y} - \bar{Y})^2$ 为解释 Y 变异的回归平方和。

多元测定系数 (coefficient of multiple determination): 在多元回归分析中, 将回归平方和与总平方和的比值称为多元测定系数, 用 R^2 表示。

复相关系数: 多元测定系数的算术平方根, 用 R 表示, 又称为多元相关系数 (coefficient of multiple correlation)。

多元测定系数可以表示得到的回归方程在多大程度上解释了因变量 Y 的变异, 取值范围为 $[0, 1]$ 。 R^2 越接近 1, 表示方程中的自变量对因变量 Y 的解释力越强。例如, $R^2 = 0.64$ 说明 Y 的变异中有 64% 可由回归方程解释。

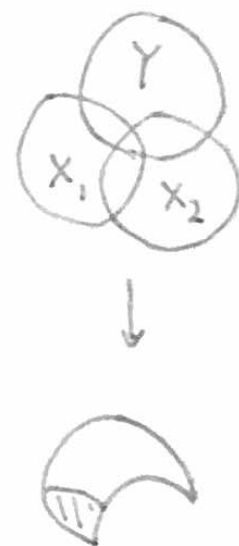
多元测定系数的算术平方根称为复相关 (multiple correlation)。复相关表示 Y 与 $X_1, X_2, \dots, X_k$ 之间的线性相关程度, 取值范围为 $[0, 1]$ 。复相关系数越接近 1, 表示 Y 与这组自变量 $X_1, X_2, \dots, X_k$ 之间的线性关系越密切。当复相关系数等于 1 时, 所有的观测值都落在回归线上; 当复相关系数等于 0 时, 表示 Y 的变异与 $X_1, X_2, \dots, X_k$ 的线性变化无关。因此, 复相关系数还有一种计算方法, 即求因变量的观测值和由回归方程得到的预测值的简单相关系数的绝对值。

二、偏相关系数

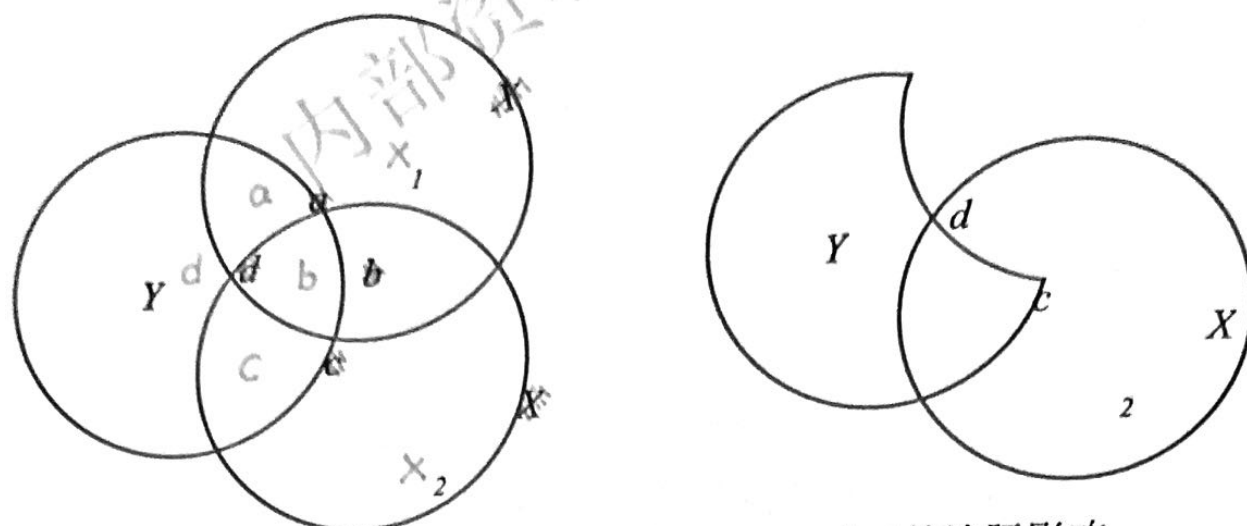
偏相关 (partial correlation) 指的是控制其它自变量后 Y 和 X 的相关。它等于从 Y 和 X 中除去其它预测变量的影响之后, Y 和 X 中剩余部分的简单相关。通常会将控制变量的个数称为阶。因此控制一个、两个、三个变量的偏相关就分别称为一阶 (first-order)、二阶 (second-order)、三阶 (third-order) 偏相关。因为简单相关没有控制变量, 所以也称为零阶 (zero-order) 相关。

将偏相关系数求平方后便可以得到偏测定系数 (coefficient of partial determination)。相较于多元测定系数 R^2 是方程中所有自变量解释 Y 的变异占 Y 总变异的比, 偏测定系数是在控制其它自变量的条件下, 单一自变量对因变量的边际解释力, 记为 R_p^2 。

以图 17-2 (I) 为例, X_2 和 X_1 对因变量 Y 既有共同解释的部分 b , 又有各自单独解释的部分 a 和 c 。假设研究者进行序列回归, 先做 Y 对 X_1 的一元回归, 再做 Y 对 X_1 和 X_2 的二元回归, 那么偏测定系数反映的就是新加入回归的变量 X_2 所解释的变异 $\frac{c}{c+d}$ 。但需要注意的是, 这一解释比例并不是以 Y 的总变异为整体, 而是以前一次回归所未能解释的部分为整体, 故称反映的是 X_2 独立于 X_1 的对 Y 的边际贡献。而 X_2 与 X_1 相关部分对 Y 的贡献则已经包含在



先进入方程的 X_1 中了。



(I) X_1 、 X_2 对 Y 的单独影响和共同影响

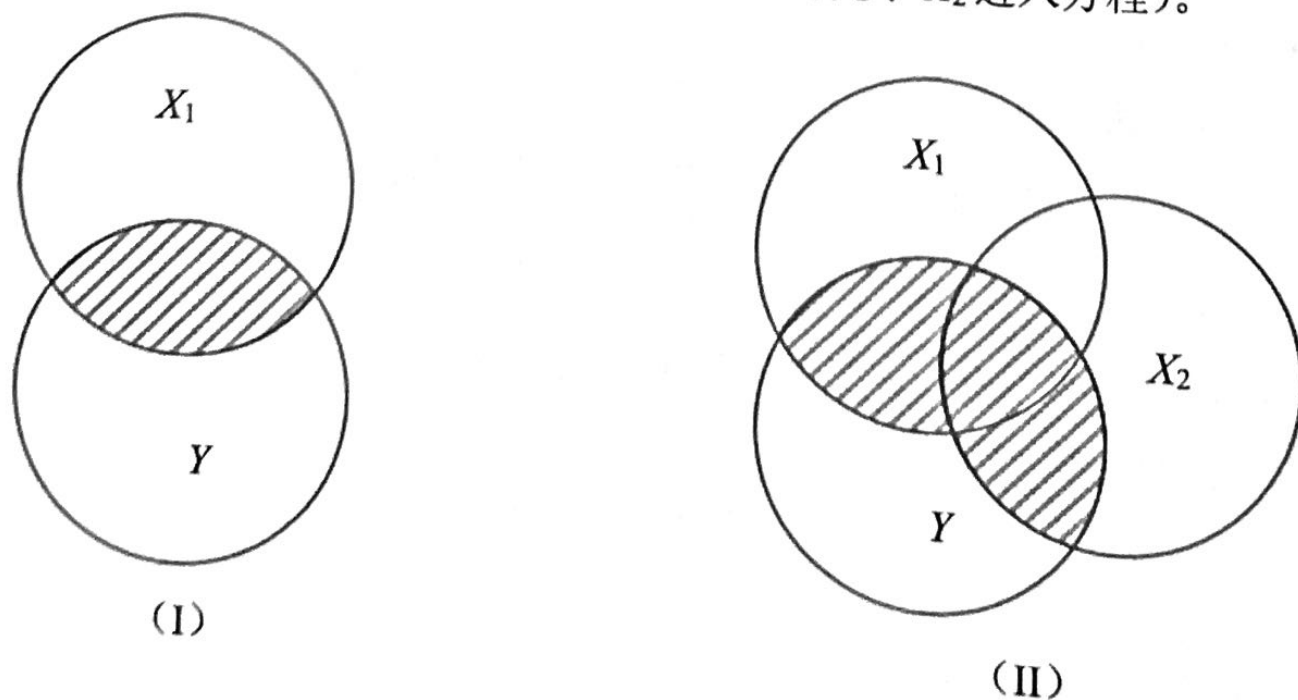
(II) X_2 对 Y 的边际影响

图 17-2 偏测定系数示意图

三、半偏相关系数

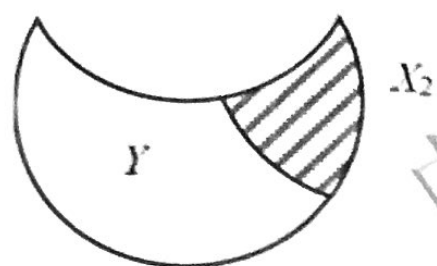
半偏相关 (semi-partial correlation) 又叫作部分相关 (part correlation), 是指从 X_k 中剔除 $X_1, X_2, \dots, X_{k-1}$ 的影响后 X_k 与 Y 之间的相关, 它反映了 X_k 对 Y 的独特贡献。以两个自变量为例, 与偏相关不同的是, 半偏相关仅从 X_2 中剔除与 X_1 相关的部分, 并没有在 Y 中剔除与 X_1 相关的部分。如图 17-3 所示, (II) 的阴影部分表示控制 X_1 后 X_2 与 Y 的偏相关, 而 (III) 的阴影部分则表示控制 X_1 后 X_2 与 Y 的半偏相关。

半偏相关描述了一个自变量对因变量的单独贡献, 是衡量自变量重要性的有效指标。但在不同类型的回归中, 半偏相关的解释方式也有不同。在标准多元回归中, 半偏相关的平方 R_s^2 可解释为剔除该自变量后 R^2 的减少量。通常情况下, 所有自变量的半偏相关的平方和会小于 R^2 。半偏相关平方和与 R^2 之间的差值代表共同变异, 也就是两个或多个自变量共同解释的变异。在序列回归和统计回归中, 半偏相关的平方 (记为 R_s^2) 可解释为加入该自变量的 R^2 的增加量, 且所有自变量的半偏相关的平方和正好等于回归方程的 R^2 。此时研究的问题就变为: 当一些自变量优先进入方程并解释了共同变异后, 加入该自变量还可以使 R^2 提高多少? 于是自变量在何时进入方程便尤为重要。图 17-4 表示在标准多元回归和序列回归中, 不同相关系数的计算结果 (在序列回归中指定 X_1 先于 X_2 进入方程)。

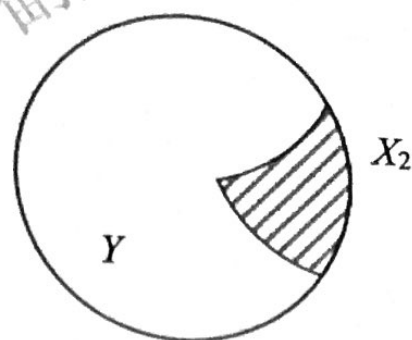


(I)

(II)

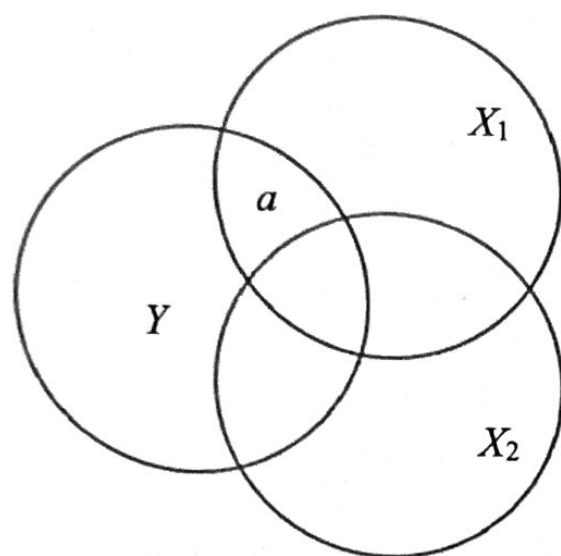


(III) 偏相关



(IV) 半偏相关

图 17-3 偏相关和半偏相关比较示意图



★图 17-4 两类回归中不同测定系数的示意图及计算结

测定系数	自变量	标准多元回归	序列回归
R^2	X_1	$\frac{a+b}{a+b+c+d}$	$\frac{a+b}{a+b+c+d}$
	X_2	$\frac{c+b}{a+b+c+d}$	$\frac{c+b}{a+b+c+d}$
R_s^2 ($\frac{1}{2}T_{\text{偏}}$)	X_1	$\frac{a}{a+b+c+d}$	$\frac{a+b}{a+b+c+d}$
	X_2	$\frac{c}{a+b+c+d}$	$\frac{c}{a+b+c+d}$
R_p^2 ($T_{\text{偏}}$)	X_1	$\frac{a}{a+d}$	★ $\frac{a+b}{a+b+c+d}$
	X_2	$\frac{c}{c+d}$	$\frac{c}{c+d}$

四、标准化回归系数

在多元线性回归分析中,由于方程中有多个自变量,所以经常需要判断在这些自变量中,哪些是影响因变量的主要因素,而哪些是次要因素。这时就可以采取比较各自变量偏回归系数 $b_1, b_2, \dots, b_k$ 的绝对值大小的方法。但是因为每个自变量的测量单位并不一致,所以不能直接对这些回归系数进行比较。此时,需要先将所有自变量 X_j 和因变量 Y 进行标准化,得到标准化的变量 $Z_j = \frac{X_j - \bar{X}_j}{S_j}$ 和 $Z_Y = \frac{Y - \bar{Y}}{S_Y}$ 之后再进行回归分析,最终得到的标准化回归方程如下:

$$\hat{Z}_Y = \beta_1 Z_1 + \beta_2 Z_2 + \dots + \beta_k Z_k \quad (17-8)$$

因为标准化的 Z 变量是无量纲的变量,所以此时的回归系数 β_j 就称为标准化回归系数

(standardized regression coefficient)。由于对自变量和因变量都进行了标准化的处理，解决了不同自变量测量单位不同的问题， β_j 之间就可以相互比较，其绝对值的大小就代表了对 Y 作用的大小。因为标准化后的变量均值都为 0，所以在标准化回归方程中不存在常数项。另外，我们也可以通过公式 $\beta_j = b_j \cdot \frac{S_Y}{S_{X_j}}$ 计算标准化回归系数。其中， S_Y 和 S_{X_j} 分别为因变量 Y 和自变量 X_j 的标准差。

第四节 多元回归分析中的有效性检验

多元回归方程的有效性检验可以分为回归方程整体的显著性检验、新加入变量的显著性检验和回归系数的显著性检验三个部分，以下展开介绍。

一、回归方程的显著性检验

多元回归方程的显著性检验就是检验 Y 与 $X_1, X_2, \dots, X_k$ 的线性关系是否显著，也就是判断在总体回归系数中是否至少有一个不等于 0。虚无假设 H_0 为：在总体中， $X_1, X_2, \dots, X_k$ 的变化都不引起 Y 的线性变化，此时所有总体回归系数均为 0，表达式为：

$$H_0: B_1 = B_2 = \dots = B_k = 0$$

而备择假设 H_1 就为：

$$H_1: \text{至少有一个 } B_j \neq 0 \quad (j = 1, 2, \dots, k)$$

虚无假设 H_0 和备择假设 H_1 都是根据总体回归方程作出的假设，因此在此处都使用 B 表示总体回归系数，而在表达样本的回归系数时一般用 b 表示。

在多元回归中，同样可以使用方差分析的思想对回归方程作检验。前文提到的多元测定系数，将因变量的总变异平方和分解为回归平方和与残差平方和，就是利用了方差分析的思想。总平方和 $\Sigma(Y - \bar{Y})^2$ 对应的自由度为 $n - 1$ ，其中 n 为样本量；回归平方和 $\Sigma(\hat{Y} - \bar{Y})^2$ 中 k 个自变量对应的偏回归系数分别为 $B_1, B_2, \dots, B_k$ ，自由度为 k ；残差平方和 $\Sigma(Y - \hat{Y})^2$ 的自由度为 $n - k - 1$ 。各部分平方和除以其对应的自由度之后可以得到相应的均方，可以用回归均方除以残差均方构造 F 检验统计量，从而对回归方程作方差分析。

★ 表 17-1 多元回归中的方差分析表

变异来源	平方和	自由度	均方	F 值	概率
回归	$\Sigma(\hat{Y} - \bar{Y})^2$	k	$\frac{\Sigma(\hat{Y} - \bar{Y})^2}{k}$	$F = \frac{\Sigma(\hat{Y} - \bar{Y})^2 / k}{\Sigma(Y - \hat{Y})^2 / (n - k - 1)}$	p
残差	$\Sigma(Y - \hat{Y})^2$	$n - k - 1$	$\frac{\Sigma(Y - \hat{Y})^2}{n - k - 1}$		
总变异	$\Sigma(Y - \bar{Y})^2$	$n - 1$			

在虚无假设成立的情况下, F 值服从 $(k, n-k-1)$ 的 F 分布。若 F 值不显著, 则接受 H_0 , 即认为 Y 与 $X_1, X_2, \dots, X_k$ 的线性关系不显著; 若 F 值显著, 则拒绝 H_0 , 即认为此多元线性回归方程是显著的。将公式 (17-3) 代入 F 值的计算公式, 得:

$$F = \frac{\frac{\sum(\hat{Y} - \bar{Y})^2}{k}}{\frac{\sum(Y - \hat{Y})^2}{(n-k-1)}} = \frac{\frac{R^2}{k}}{\frac{(1-R^2)}{(n-k-1)}} \quad (17-9)$$

由公式 (17-9) 可知, 对回归方程整体的显著性检验等价于对多元测定系数 R^2 的显著性检验。换言之, 所有自变量合起来对因变量的解释率与 0 有显著差异, 同样可以说明多元回归方程是有意义的。

二、新加入变量的显著性检验

在序列回归和统计回归中, 如果想检验在已有自变量解释一定的 R^2 后新加入的一个或一组自变量是否显著地增加了 R^2 , 或者说 R^2 的改变量是否显著, 仍然可以使用 F 检验。其中, F 值的计算公式如下:

$$F = \frac{(R_{wi}^2 - R_{wo}^2)/m}{(1-R^2)/(n-k-1)} \quad (17-10)$$

式中, R_{wi}^2 是新加入自变量后回归方程的多元测定系数, R_{wo}^2 是没有加入新自变量时回归方程的多元测定系数, m 是新加入自变量的个数。可以发现, 公式 (17-10) 的分母与公式 (17-9) 方程整体检验中 F 值的分母是相同的。

三、回归系数的显著性检验 (t 检验)

对回归系数进行显著性检验的零假设 H_0 为: $B_j = 0$, 而备择假设 H_1 为: $B_j \neq 0$ 。当检验不显著时, 则不拒绝 H_0 , 即认为自变量 X_j 与 Y 不存在线性关系, 可以从回归方程中删除 X_j ; 而当检验显著时, 则拒绝 H_0 , 即认为自变量 X_j 与 Y 存在线性关系, 进而在回归方程中保留 X_j 。

在标准多元回归中, 对自变量的偏回归系数的显著性采用 t 检验。得到回归系数的估计值 b_j 后, 还需要计算回归系数的标准误 SE_{b_j} , 将回归系数的估计值除以标准误便得到 t 值, 该 t 值的自由度为 $n-k-1$ 。如果 t 检验显著, 可以认为该自变量的回归系数与 0 存在显著差异, 说明控制了方程中其它自变量的影响后, 该自变量对因变量仍有显著的预测作用。

【例 17-1】某医院想研究患者的身体健康症状、心理健康症状和应急生活事件 3 个因素对其看病次数的影响。随机选取了 360 名患者进行调查。

(1) 作标准多元回归分析, 结果如下:

表 17-2 标准多元回归模型拟合的总体情况表

Model	R	R^2	Adjusted R^2	Std. Error of Estimate	Durbin-Watson
1	.459	.211	.204	10.128	.318

模型拟合的总体情况如表 17-2 所示, 由 R^2 可知: 三个自变量总共可以解释因变量变异

的 21.1%。Durbin-Watson 的检验结果为.318，按一般标准认为残差可能存在正相关。

表 17-3 标准多元回归方程检验的方差分析表

	Model	SS	df	MS	F	p
1	Regression	9766.257	3	3255.419	31.736	<.001
	Residual	36518.118	356	102.579		
	Total	46284.375	359			

回归方程检验的方差分析结果如表 17-3 所示。表中列出了回归项和残差项的平方和、自由度和均方，还列出了 F 值以及显著性水平 p 。从表中可以看出，自变量对因变量的线性回归方程是显著的， $F(3, 356) = 31.736$ ， $p < .001$ ，即说明身体健康症状、心理健康症状和应急生活事件能够显著预测患者看病的次数。

表 17-4 标准多元回归系数估计及其检验表

Model	Unstandardized		Standardized		p
	B	Std. Error	β	t	
(Constant)	-4.208	1.369		-3.705	.002
身体健康症状	1.778	.271	.360	6.551	.000
心理健康症状	.022	.150	.008	.149	.881
应急生活事件	.016	.193	.193	3.826	.000

回归系数估计及其检验结果如表 17-4 所示。从表中可以看出截距项、身体健康症状和应急生活事件的偏回归系数的 t 检验都达到了显著水平，说明截距项和这些偏回归系数都具有统计学意义，这两个自变量与因变量的多元线性关系显著。而心理健康症状的偏回归系数则没有达到显著水平，表明该自变量对因变量的贡献可以忽略不计。这些结果可以说明身体健康症状和应急生活事件在控制其它变量恒定的情况下各自能显著预测患者看病的次数。

(2) 作序列多元回归分析，结果如下：

由于本例中残差分布不满足正态及残差方差齐性假设，因此对因变量作了对数转换。研究者此时关注的问题是：分别控制身体健康和应急事件的影响后，心理健康是否还显著影响患者看病的次数？换言之，对于身体健康和生活压力相似的人，心理问题更多的人看病的次数是否更多？据此在进行序列回归分析时，身体健康症状首先进入回归方程，其次应急生活事件进入，心理健康症状最后进入方程。即：

结果变量：看病次数（对数转换）。三个模型的预测变量分别是：

Model 1: 身体健康症状

Model 2: 身体健康症状，应急生活事件

Model 3: 身体健康症状，应急生活事件，心理健康症状

模型拟合的总体情况表如表 17-5 所示。由于建立了三个回归模型，共有三个模型拟合结果。从模型一到模型三， R^2 依次增大，这是增加了预测变量的缘故。值得注意的是，模型三的调整的 R^2 小于模型二，而标准误大于模型二，这一结果提醒我们模型三加入的预测变

量可能对因变量不具有良好的预测作用。结果中还有增加新自变量后的 R^2 变化的 F 检验结果。结果表明, 仅使用身体健康症状就可以显著预测患者的看病次数, $F(1, 358) = 162.971$, $p < .001$, $\Delta R^2 = .313$; 但在此基础上加入应急生活事件同样能显著提高回归模型的 R^2 , $F(1, 357) = 18.368$, $p < .001$, $\Delta R^2 = .034$; 但最后加入心理健康症状, 并没有带来 R^2 的显著改变, $F(1, 356) = .351$, $p = .554$, $\Delta R^2 = .001$ 。这同样表明模型三加入的心理健康症状对预测因变量没有显著的贡献。

表 17-5 序列多元回归的模型拟合的总体情况表

Model	R	R^2	Adjusted R^2	Std. Error of Estimate	Change Statistics				
					ΔR^2	F Change	df_1	df_2	p
1	.559	.313	.311	.79308	.313	162.971	1	358	.000
2	.589	.346	.343	.77452	.034	18.368	1	357	.000
3	.589	.347	.342	.77522	.001	.351	1	356	.554

回归方程检验的方差分析结果如表 17-6 所示。表中列出了三个回归方程的 F 检验的结果, 可以看出自变量对因变量的多元线性回归都是显著的。

表 17-6 序列多元回归方程检验的方差分析表

	Model	SS	df	MS	F	p
1	Regression	102.505	1	102.505	162.971	< .001
	Residual	225.175	358	.629		
	Total	327.680	359			
2	Regression	113.524	2	56.762	94.623	< .001
	Residual	214.156	357	.600		
	Total	327.680	359			
3	Regression	113.735	3	37.912	63.084	< .001
	Residual	213.945	356	.601		
	Total	327.680	359			

回归系数估计及其检验结果如表 17-7 所示。表中列出了三个回归方程中截距项和含有自变量的偏回归系数和 t 检验结果。在模型一中, 身体健康症状可以显著预测看病次数, $\beta = .559$, $p < .001$ 。由于没有其它自变量影响, 身体健康症状的零阶相关、偏相关和半偏相关三个系数相等。在模型二中同时存在身体健康症状和应急生活事件, 此时仅关注这一层新加入的自变量, 即应急生活事件。结果表明控制了身体健康症状后, 应急生活事件仍可以显著预测看病次数, $\beta = .192$, $p < .001$ 。其偏相关将身体健康症状可解释的部分同时从因变量和应急生活事件中去掉后, 应急生活事件能解释因变量变异的大小, 为.221; 而半偏相关表示仅在应急生活事件中去掉与身体健康症状重叠的部分后, 应急生活事件能解释因变量变异的大小, 为.183。由于存在其它自变量的影响, 身体健康症状的偏相关和半偏相关相较于上一层有所减小, 减少的部分即为新加入的应急生活事件能解释的部分。模型三中的结果可以回

答本例最关注的问题,即控制了身体健康症状和应急生活事件后,心理健康症状不能显著预测看病次数, $\beta = .030$, $p < .554$ 。偏相关和半偏相关分别为.031和.025,与零阶相关.336相差较大,再次验证了这一结果,这也体现了序列回归分析相比于简单相关分析的优势,即能够在控制其它自变量的情况下考察关键变量与因变量的关系。此外,在每层只加入一个新的自变量时,对偏回归系数的 t 检验等价于对回归方程的偏测定系数的 F 检验。

表 17-7 序列多元回归系数估计及其检验表

Model	Unstandardized		Standardized			Correlations		
	<i>B</i>	Std. Error	β	<i>t</i>	<i>p</i>	Zero-order	Partial	Part
(Constant)	.576	.099		5.834	.000			
身体健康症状	.232	.018	.559	12.766	.000	.559	.559	.559
(Constant)	.407	.104		3.910	.000			
身体健康症状	.210	.019	.504	11.282	.000	.559	.513	.483
应急生活事件	.001	.000	.192	4.286	.000	.337	.221	.183
(Constant)	.402	.105		3.833	.000			
身体健康症状	.204	.021	.491	9.822	.000	.559	.462	.421
应急生活事件	.001	.000	.185	4.038	.000	.337	.209	.173
心理健康症状	.007	.011	.030	.593	.554	.336	.031	.025

最后,在文献中汇报序列回归分析结果时,推荐整理成如表 17-8 的形式。

表 17-8 预测患者看病次数的序列回归分析结果

Model	第一步			第二步			第三步		
	<i>b</i>	<i>SE</i>	β	<i>b</i>	<i>SE</i>	β	<i>b</i>	<i>SE</i>	β
身体健康症状	.232	.018	.559***	.210	.019	.504***	.204	.021	.491***
应急生活事件				.001	.000	.192***	.001	.000	.185***
心理健康症状							.007	.011	.030
ΔR^2		.313***			.034***			.001	

注: *** $p < .001$ 。

第五节 多重共线性

在多元线性回归分析中,因为涉及多个自变量,所以还需要考虑多个自变量之间相关过高有可能带来的影响,即多重共线性问题。

★多重共线性 (multicollinearity): 指线性回归模型中自变量之间由于存在较高的相关而使模型参数估计不准确或难以估计等问题。

如果两个自变量之间的相关系数很高,或者一个自变量可以由其它自变量线性表示,即

认为存在多重共线性问题。共线性会导致分析的结果难以解释,例如比较典型的表现是模型整体上的测定系数很高,且因变量与各自变量的相关系数也很高,但自变量的回归系数均不显著。这一结果在解释上存在整体和局部变量重要性不一致的矛盾,很容易导致对自变量重要性作出错误判断。在大多数的实际问题中,自变量间存在相关是非常普遍的现象。但是,如果多个自变量之间相关重叠性太高(高度相关),自变量间会相互削弱对因变量的边际影响,使回归系数在数值上降低,而标准误增大,使得回归系数变得不可靠,而且回归方程虽然显著,但每个自变量不显著。

★共线性现象会造成很多不良的影响,主要表现在以下几个方面:

- (1) 回归系数的置信区间变宽,系数变得不稳定,由样本推到总体不可靠;
- (2) 回归系数不能很好地反映单个自变量对因变量的独立影响;
- (3) 使变量的偏测定系数变小;
- (4) 当方程用于预测时,回归结果变得不可靠。

●需要注意的是,多重共线性指的是变量之间的线性相关,当自变量之间存在非线性相关时,不一定会产生严重的多重共线性问题。

★在研究过程中,及时地发现共线性问题是很重要的,这就要求研究人员掌握一些检验共线性现象是否存在的方法,以下列出一些常用的方法,以供参考:

- (1) 方程的测定系数很高,但自变量的回归系数不显著,或者偏测定系数很小时,则变量间有可能是存在共线性问题;
- (2) 只有两个自变量时,如果它们之间的相关系数很高,或者存在多个自变量时,一个自变量可以由其它自变量线性表示出,都有可能是存在共线性问题;
- (3) 分别构造不含某一自变量的回归模型,与包含所有自变量的回归模型比较,如果测定系数很接近,则有可能是存在共线性问题;
- (4) 另外,可以采用一些指标检验共线性。这些指标在一般的统计分析软件中都可以直接输出,常用的指标及其判断标准有:

容忍度(tolerance)越接近1越好。当容忍度小于0.1时,就认为自变量之间存在多重共线性,经验上认为该变量不能被接受。

方差膨胀因子(Variance Inflate Factor, VIF)越大,说明多重共线性越强。当方差膨胀因子值大于10时,表明多重共线性较强。

条件指数(Condition Index)越大,说明多重共线性越强。当条件指数值小于10时,表明存在的多重共线性弱。

当方程中存在严重的多重共线性时,应想办法消除这种影响,以下是一些补救办法:

- (1) 去掉与因变量相关程度低、而与其他预测变量高度相关的变量;
- (2) 去掉可以被其余自变量线性表示的变量;
- (3) 增加样本量;

- (4) 采用新的样本数据;
- (5) 利用先验信息组合变量, 尽量保留原有自变量中有用的信息, 减少自变量个数;
- (6) 换用其它形式的回归, 如岭回归;
- (7) 变量转换, 可以对变量进行必要的转换 (如倒数或对数变换等), 以减弱变量间的多重共线性。

第六节 在文献中的结果报告

SPSS 操作

1. 操作过程

在 SPSS 中打开数据文件, 选择 Analyze→Regression→Linear 命令, 打开主对话框。将因变量选入到 Dependent 列表框中, 将自变量选入 Independent(s)列表框中。选择自变量进入回归方程的方法: Enter/Remove/Forward/Backward/Stepwise。可在 Statistics、Plots、Save、Options 按钮中, 添加其他设置。之后单击 OK 按钮即可运行程序。

2. 结果分析

在正式出版的论文和研究报告中, 多元回归分析一般需要报告: 进入方程的自变量, 模型的总体拟合情况 (测定系数), 标准化和非标准化的回归方程, 回归方程的显著性 (回归方程检验的方差分析表), 各自变量回归系数的显著性等。可以用表格的形式呈现, 再结合文字描述。

例如, 考察生产线和销售形象对顾客继续购买该产品的可能性的影响作用, 使用多元回归分析的方法, 得到的结果如表 17-9 所示:

表 17-9 模型拟合的总体情况表

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.688	.473	.437	.6705	2.022

通过回归分析, 两个自变量 (生产线和销售形象) 均进入回归方程。方程的测定系数 R^2 为 0.473, 说明生产线和销售形象一共可以解释顾客继续购买该产品的可能性变异的 47.3%。Durbin-Watson 检验的值为 2.022, 约等于 2, 可以说明残差与自变量之间相互独立。

表 17-10 回归方程检验的方差分析表

Model	Sum of Squares	df	Mean Square	F	Sig.
1 Regression	49.702	2	24.851	44.884	.000
Residual	109.073	197	.554		
Total	158.775	199			

从表 17-10 中可以看出, 自变量对因变量的线性回归方程是显著的, $F(2, 197) = 44.884$, $p < 0.001$, 说明生产线和销售形象能够影响顾客继续购买该产品的可能性。

表 17-11 回归系数估计及其检验表

Model	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	β	t	Sig.
(Constant)	4.594	.334		13.749	.000
生产线	.303	.040	.446	7.550	.000
销售形象	.250	.047	.315	5.336	.000

从表 17-11 中可以看出生产线和销售形象的偏回归系数 t 检验都达到了显著水平, $p < 0.05$, 说明这两个自变量能够显著预测顾客继续购买该产品的可能性。

非标准化回归方程可以写为: $\hat{Y} = 4.594 + 0.303 \times \text{生产线} + 0.250 \times \text{销售形象}$

标准化回归方程可以写为: $\hat{Y} = 0.446 \times \text{生产线} + 0.315 \times \text{销售形象}$

小结

1. 多元回归分析是同时考察多个自变量对一个因变量影响的一种统计方法, 在教育与心理学研究中有着广泛的运用。

2. 多元线性回归的基本假设包括:

(1) 各个自变量均与因变量 Y 呈线性关系, 且自变量不包含测量误差, Y 为连续等距的变量; (2) 误差 $\varepsilon_i \sim N(0, \sigma^2)$; (3) 误差 ε_i 满足方差齐性的假设; (4) 误差之间相互独立; (5) 误差与各个自变量之间相互独立; (6) 多个自变量之间不存在多重共线性。

3. 多元回归有三种主要类型: 标准多元回归, 序列(分层)回归, 统计回归, 它们的区别在于自变量进入方程的顺序和解释自变量贡献的方式, 应根据研究需要选择合适的回归类型。

4. 多元回归中有多种可衡量自变量重要性的指标, 包括多元测定系数 R^2 、偏相关系数、半偏相关系数、标准化回归系数等。灵活使用这些指标可帮助选择合适的自变量, 对自变量的重要性作出合理的解释。

5. 在多元回归中, 可借鉴方差分析的思想对回归方程的整体显著性进行 F 检验, 还可对新加入自变量对 R^2 的改变 (ΔR^2) 的显著性进行 F 检验, 同时, 也可以采用 t 检验对单个自变量偏回归系数的显著性进行检验。

6. 在大多数的实际问题中, 自变量间存在相关是非常普遍的现象, 但是基于最小二乘法的多元回归假设自变量之间不存在多重共线性。在作多元回归分析之前, 应采用统计指标(如容忍度, 方差膨胀因子和条件指数)判断自变量之间是否存在共线性。

7. 共线性现象会对回归分析的结果造成很多不良影响, 例如可能会带来模型的测定系数很高, 且因变量与各自变量的相关系数也很高, 但自变量的回归系数均不显著, 导致分析结果难以解释等问题, 实际应用中应想办法消除这些影响。