第六章　回归、相关与方差分析

6.1　回归分析基本概念

本章所要讨论的题目都是在数理统计学中应用很广泛的分支.它们有一个共同点,即都是研究变量之间的关系.这些变量可以是随机的,也可以是非随机(可以理解为能由人所控制)的,但不能全部为非随机的.它们的不同之处在于:回归分析着重在寻求变量之间近似的函数关系,相关分析则不着重这种关系,而致力于寻求一些数量性的指标,以刻画有关变量之间关系深浅的程度.第三章中讨论过的相关系数,就是这样的一个指标.方差分析着重考虑一个或一些变量对一特定变量的影响有无及大小,由于其方法是基于样本方差的分解,故得名.以上只是一个很一般的描述,在以后的叙述中将加以充实和确切化.

我们先来谈回归分析."回归"一词的来由将在后面加以解释.在现实世界中存在着大量这样的情况:两个或多个变量之间有一些联系,但没有确切到可以严格决定的程度.例如,人的身高 X 和体重 Y 有联系,一般表现为 X 大时,Y 也倾向于大,但由 X 并不能严格地决定 Y.一种农作物的亩产量 Y 与其播种量 X_1,施肥量 X_2 有联系,但 X_1,X_2 不能严格决定 Y.工业产品的质量指标 Y 与工艺参数和配方等有联系,但后者也不能严格决定 Y.

在以上诸例及类似的例子中,Y 通常称为因变量或预报量,X,X_1,X_2 等则称为自变量或预报因子.因变量自变量的称呼借用自函数关系,它不十分妥贴,因为,有时变量间并无明显的因果关系存在.例如,不好说一个人的身高是因体重是果,因为你也可以反过来说,该人身高是因其体重大.预报量与预报因子的名称来源于实际.因为在应用中,多是借助于一些变量之值去预测另一些

变量之值.比如说,用播种量和施肥量去预测产量.这名称也非十分完善,因为在回归分析的某些应用中,并无预报的含义.迄今为止,对 X(或($X_1,X_2\cdots$))和 Y 并无一种一致采用或公认为妥贴的称呼,为简单计,今后我们将固定使用自变量和因变量这一对名词.

为什么由 X_1,X_2 等不能严格决定 Y? 理由很清楚.拿农作物那个例子来说,影响产量 Y 的因素(变量)很多,远不止播种量 X_1 和施肥量 X_2 二者,其他如灌溉情况,气温变化情况,灾害(病虫害、风灾之类),都影响到 Y.这些因素中,有可以人为控制的(如已考虑的 X_1,X_2),有原则上可控但因技术、经济力量不及,或研究工作目标有限未予控制的,还有一大批难于控制的随机因素.因此,已考虑的因素 X_1,X_2 只能在一定程度上决定产量 Y,其余则委之于随机误差.因此,在回归分析中,因变量总是看作为随机变量.至于自变量则情况较复杂:有随机的,如人的身高体重那个例子,不是给定身高去测体重,而是随机地抽出一个人,同时测其身高体重,故二者都是随机变量.也有非随机的,农作物例中的播种量和施肥量即是,它们的取值可以由人控制.从数理统计学的理论上说这二者有差别.但从实用上说,人们往往把随机自变量当作非随机去处理,但对结果的解释要小心,以后再谈.在本章 6.2 和 6.3 这两节中,除有特别声明,我们将一律把自变量视为非随机的.

现设在一个问题中有因变量 Y,及自变量 $X_1,\cdots,X_p$.可以设想 Y 的值由两部分构成:一部分由 $X_1,\cdots,X_p$ 的影响所致,这一部分表为 $X_1,\cdots,X_p$ 的函数形式 $f(X_1,\cdots,X_p)$.另一部分则由其他众多未加考虑的因素,包括随机因素的影响所致,它可视为一种随机误差,记为 e.于是得到模型:

$$Y = f(X_1,\cdots,X_p) + e$$

e 作为随机误差,我们要求其均值为 0:

$$E(e) = 0$$

于是得到:$f(X_1,\cdots,X_p)$就是在给定了自变量 $X_1,\cdots,X_p$ 之值的

条件下,因变量 Y 的条件期望值.可写为*

$$f(X_1,\cdots,X_p) = E(Y \mid X_1,\cdots,X_p)$$

函数 $f(x_1,\cdots,x_p)$称为 Y 对 $X_1,\cdots,X_p$ 的"回归函数",而方程

$$y = f(x_1,\cdots,x_p)$$

则称为 Y 对 $X_1,\cdots,X_p$ 的"回归方程".有时在回归函数和回归方程之前加上"理论"二字,以表明它是直接来自模型,也可以说是模型的一个组成部分,而非由数据估计所得.后者称为"经验回归函数"和"经验回归方程".

设 ξ 为一随机变量,则 $E(\xi-c)^2$ 作为 c 的函数,在 $c=E(\xi)$ 处达到最小.由这个性质,可以对理论回归函数 $f(x_1,\cdots,x_p)$作下面的解释:如果我们只掌握了因素 $X_1,\cdots,X_p$,而希望利用它们的值以尽可能好地逼近 Y 的值,则在均方误差最小的意义下,以使用理论回归函数为最好.

但在实际问题中,理论回归函数一般总是未知的,统计回归分析的任务,就在于根据 $X_1,\cdots,X_p$ 和 Y 的观察值,去估计这个函数,及讨论与此有关的种种统计推断问题,如假设检验问题和区间估计问题.所用的方法,在相当大的程度上取决于模型中的假定,也就是对回归函数 f 及随机误差 e 所作的假定.先说回归函数 f.一种情况是对 f 的数学形式并无特殊的假定,这种情况称为"非参数回归".另一种情况,即目前在应用上最多见的情况,是假定 f 的数学形式已知,只其中若干个参数未知.例如,$p=2$,而已知 $f(x_1,x_2)$形如

$$f(x_1,x_2) = c_1 + c_2 e^{c_3 x_1} + c_4 \log x_2$$

其中 $c_1,\cdots,c_4$ 是未知参数,要通过观察值去估计.这种情况称为"参数回归".其中在应用上最重要且在理论上发展得最完善的特

* 以往我们定义条件期望时,是假定所有的变量都为随机的.如今自变量 $X_1,\cdots,X_p$ 并非随机,故记号 $E(Y \mid X_1,\cdots,X_p)$只是一种借用.可以简单地理解为:Y 的分布依赖于参数 $X_1,\cdots,X_p$,故其期望值也应与 $X_1,\cdots,X_p$ 有关.

例，是 f 为线性函数的情形：

$$f(x_1,\cdots,x_p)=b_0+b_1x_1+\cdots+b_px_p$$

这种情况叫做“线性回归”，是我们今后讨论的主要对象.线性回归的限制看来较强.不过，如果自变量变化的范围不太大，而曲面 $y=f(x_1,\cdots,x_p)$弯曲的程度也不过分，则在较小的范围内，它可以近似地用一个平面(即线性函数)去代替之，而不致引起过大的误差.其次，有些形式上看是非线性的回归函数，可能通过自变量的代换转化为线性的，见 6.3 节.因此，线性回归模型有比较大的适用面，加之它处理上简便，成为一个极其重要的模型.

对随机误差 e，我们已假定其均值 $E(e)=0$. e 的方差σ^2 是回归模型的一重要参数，因为

$$E[Y-f(X_1,\cdots,X_p)]^2=E(e^2)=\mathrm{Var}(e)=\sigma^2$$

σ^2 愈小，用 $f(X_1,\cdots,X_p)$逼近 Y 所导致的均方误差就愈小，回归方程也就愈有用. σ^2 的大小由什么决定呢？这就在于以下两点：

1. 在选择自变量时，是否把对因变量 Y 有重要影响的那些都收进来了.如果是这样，则未被考虑的即作为随机误差去处理的那些因素，总的起作用就较小，因而 σ^2 也就会较小.反之，若遗漏了或因条件关系，使某些对 Y 有重要影响的因素未被考虑，则其影响进入随机误差 e，将导致 σ^2 增大.

2. 回归函数的形状是否选得准.比如，理论回归函数 $f(x_1,\cdots,x_p)$本是一个非线性函数，而你用一个线性函数 $g(x_1,\cdots,x_p)$，则二者的差距 $f-g$ 就作为一种误差进入 e 内，而加大了它的方差.

因此在应用上，通过观察数据对误差方差 σ^2 作估计，也是很重要的.如果估计值很大，超过了该项应用所能承受的范围，则估计所得的回归方程意义就不大.在这个时候，就有必要再考虑一下自变量的选择是否抓着了主要因素，以及所用的回归方程的形式是否太不符合实际.

如果要处理有关的检验和区间估计问题，比方说，取定了线性

回归函数 $b_0+b_1x_1+\cdots+b_px_p$，有对未知系数 b_i 等作假设检验和区间估计的问题，则只有在假定随机误差 e 服从正态分布 $N(0,\sigma^2)$时，才有满意的小样本方法. 因此，在实用回归分析中，常假定误差服从正态分布. 经验证明：对多数应用问题来说，这个假定是可以接受的，如果没有这个假定，那就需要使用大样本方法.

回归分析的应用，可以归纳为以下几方面.

第一方面是纯描述性的. 为简单计，以一个自变量 X 的情况为例，因变量总记为 Y. 假定在工作中我们经常要记录 X 和 Y 之值(比如说，X 代表月份，Y 代表该月的产值)，而积累了一批数据(X_1, Y_1)，(X_2, Y_2)，$\cdots$，(X_n, Y_n). 把它们标在直角坐标系上，称为散点图. 这往往是杂乱无章的，但仍可能有某种趋势存在. 如图 6.1 中的点虽系杂乱无章，但大体呈现出一种直线走向的趋势. 用回归分析的方法可找出一条较好地代表这些点的走向的直线 l. 在一定程度上，这条直线 l 描述了所观察到的这批数据所遵从的规律，虽不十分准确，但有时很有用.

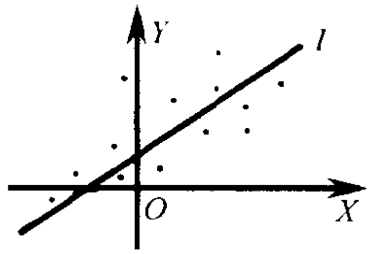

图 6.1

这种应用之所以称为描述性的，是因为它只是对数据的一种“总结”，它只涉及现有数据，不超出其外，用统计的语言说，它并不企图对数据(X_1, Y_1)，$\cdots$，(X_n, Y_n)所来自的总体作任何推断.

第二方面是估计回归函数 f. 仍拿人的身高 X 和体重 Y 这个例子来说，姑且把 X 视为自变量而 Y 为因变量. 若假定(X, Y)服从二维正态分布，则如在第二章中已证明的，Y 对 X 的回归函数 $f(x)$，即条件期望 $E(Y|X=x)$，为 x 的线性函数b_0+b_1x. 如果通过样本对 b_0 和 b_1 作出了估计 $\hat{b}_0$ 和 $\hat{b}_1$，则用 $\hat{b}_0+\hat{b}_1x$ 去估计 b_0+b_1x. 在本例中，后者就是在身高为 x 的人群中的平均体重. 这在应用上很有意义，因为在不少问题中，我们所关心的正是这个平均值. 再拿亩产 Y 与播种量X_1与施肥量 X_2 的关系这个例子

来说,也许我们所关心的正是在一定播种量 x_1 和一定施肥量 x_2 之下,平均亩产能达到多少.这就是 Y 对 X_1,X_2 的回归函数 $f(x_1,x_2)$.

第三方面是预测,即在特定的自变量值$(x_{10},\cdots,x_{p0})$之下,去预测因变量 Y 将取的值 y_0.例如,随意碰到一个人测出其身高为 x_0,而没有秤其体重或秤了没有把结果告诉你,让你去预测这人体重有多少.这与估计身高为 x_0 的人群的平均体重 $f(x_0)=E(Y|X=X_0)$不同.后者并非特定的一个身高为 x_0 的人的体重,而是全体这样的人体重的平均值,而预测的对象则是这个特定的人的体重.从模型上可以这样看:设在 $X=X_0$ 处进行观察,随机误差为 e_0,而 Y 之值为 y_0,则 $y_0=f(x_0)+e_0$.为了预测 y_0,需要对 $f(x_0)$进行估计,同时也对随机误差值 e_0 作估计,把二者相加得出 y_0.随机误差 e_0 之值凭机会而定,没有什么好的估计方法,只能根据其均值为0这一点,将其值估计为0.于是 Y 的预测值就取为回归函数 $f(x)$在这个点 x_0 处的估计 $\hat{f}(x_0)$.

由这里得出两条结论:一是预测问题与回归函数问题虽然在实质上很不一样(如前面所曾解释的),但二者之解则一样.因为这一点,有些著作没有强调这二者的区别所在.二是预测的精度要比估计回归函数的精度差.因为在预测中,除了估计回归函数有一个误差外,还要加上一个随机误差 e_0.这一点在考虑区间估计时能更清楚地看出来.

第四方面是控制.在这类应用中,不妨把自变量解释为输入值,因变量解释为输出值.目标是要把输出值控制在给定的水平 y_0.若通过数据估计出了经验回归方程 $y=\hat{f}(x_1,\cdots,x_p)$,则根据这方程可调整自变量 $X_1,\cdots,X_p$ 的取值,以达到上述目的.例如,自变量 X 是用药量,而 Y 是某种生理指标,例如血压,调整用药量以使血压达到某种认为是正常的水平.

我们提一下"回归设计"这个概念.为了估计理论回归函数 $f(x_1,\cdots,x_p)$,需要对自变量 $X_1,\cdots,X_p$ 和因变量 Y 进行观测.有

两种情况：一是自变量也是随机的，如人的身高体重那个例子，这时除了一般地保证抽样的随机性以外，就没有多少可做的事情了．例如在一大群人中抽取若干以量测其身高体重，则只须尽力保证人群中的每一个有同等的被抽出的机会．

另一种情况是自变量是非随机的，其取值在一定限度内可由人去控制．这时，为保证取得最大的效果，应对自变量在各次试验中所取的值进行适当的规划．例如，若在将来的应用中自变量多取某区域 B 上之值，则在进行试验时就要让自变量多在这个范围内取值．也可以设想，试验点在空间的排列可能需要有某种对称性，以便于统计分析．这些问题的研究构成了回归分析的一个分支，叫做回归设计，它也可以看作是试验设计这个统计学分支的一个组成部分，本章将不讨论这方面的问题．

最后我们来解释一下“回归”这名称的由来．这个术语是英国生物学家兼统计学家 F．高尔顿在 1886 年左右提出来的．人们大概都注意到，子代的身高与其父母之身高有关．高尔顿以父母之平均身高 X 作为自变量，某成年子女身高的平均 Y 为因变量．他观察了 1074 对父母及某成年子女身高的平均，将所得 (X, Y) 值标在直角坐标系上，发现二者的关系近乎一条直线，有如图 6.1．总的趋势是 X 增加时 Y 倾向于增加——这是意料中的结果．有意思的是，高尔顿对所得数据作了深入一层的考察，而发现了某种有趣的现象．

高尔顿算出这 1074 个 X 值的算术平均为 $\overline{X}=68$ 英寸（1 英寸为 2.54 厘米），而 1074 个 Y 值的算术平均为 $\overline{Y}=69$ 英寸，子代身高平均说增加了 1 英寸，这个趋势现今人们也已注意到．以此为据，人们可能会这样推想：如果父母平均身高为 a 英寸，则这些父母的子代平均身高，应为 $a+1$ 英寸，即比父代多 1 英寸．但高尔顿观察的结果与此不符：他发现：当父母平均身高为 72 英寸时，他们的子代身高平均只有 71 英寸，不仅达不到预计的 $72+1=73$ 英寸，反而比父母平均身高小了．反之，若父母平均身高为 64 英寸，则观察数据显示子代平均身高为 67 英寸，比预计的 $64+1=65$ 英

寸要多.

高尔顿对此的解释是:大自然有一种约束机制,使人类身高分布保持某种稳定形态而不作两极分化.这就是一种使身高“回归于中心”的作用.例如,父母身高平均为72英寸,比他们这一代平均身高68英寸高出许多,“回归于中心”的力量把他们子代的身高拉回来一些:其平均只有71英寸,反比父母平均身高小,但仍超过子代全体平均69英寸.反之,当父母平均身高只有64英寸——远低于他们这一代的平均值68,而“回归于中心”的力量将其子代身高拉回去一些,其平均值达到67,增长了3英寸,但仍低于子代全体平均值69.

正是通过这个例子,高尔顿引入了回归这个名词.现在我们觉得,高尔顿的例子只反映了变量关系中的一种情况,在其他涉及变量关系的众多情况中,多不必如此,故拿这个名称作为变量关系统计分析的称呼,实不见得恰当.但这个名词现今已沿用成习,如硬要改变,反觉多此一举了.