生物统计基础知识(III)

刘来福
(北京师范大学数学系)

四、回归分析

(一) 函数关系和相关关系

每一自然现象的变化都和它周围其它的现象互相联系着和互相影响着。因此反映这些现象变化规律的各种变量也是互相联系的，互相影响的。从而它们之间就存在着一定的关系。人们通过各种实践发现变量之间的关系大致可分成两种类型。一种类型是我们所熟知的函数关系。其特点是变量之间的关系是完全确定的。例如，对一定数量的饲料M来说，每头牛平均所消耗的饲料的量m与牛的头数n之间的关系就是$m = M/n$，它是一个函数关系。也就是说，如果已知牛的头数和总的饲料的量，那么平均每头牛所消耗的饲料的量就是完全确定的。

但是在生物学当中，在大多数情况下，变量之间的关系并没有这么简单。例如，动物在一定时间内饲料的消耗量与它的增重量之间，根据我们的经验可知它们是有关系的。进食量大，增重也就多。但是要找出这两者之间的确切关系则往往是困难的。因为影响体重增加的因素是复杂的，其中有些我们可以考虑进去，还有相当多是属于我们一时还没有认识和掌握的，或虽已认识但还暂时无法控制和测量的。在所有这些偶然因素影响下我们考虑的变量中有些以随机变量的形式表现出来，从而使得这些变量之间的关系不能确切地表示出来。这种关系在实践中是大量存在的。如在动物饲养实验中的原始体重与一定时间内的增重量的关系，动物的胸围、体长和体重的关系等。这些变量之间都存在着密切的关系，但又不能由一个(或几个)变量的数值精确地求出另一个变量的值。我们称这类变量之间的关系为相关关系。

变量之间所蕴含的相关关系是可以经过多次的试验和分析把它们找出来的。回归分析就是一种处理变量间相关关系的统计方法。它可以从大量观测数据中扬弃随机因素的干扰找出反映事物内在的规律性。

回归分析的内容很多，这里只介绍最简单的一元线性回归。

(二) 回归线的求法

一元线性回归就是通常遇到的配经验直线的问题，也就是如何通过样本资料来研究随机变量之间的线性相关的关系。

例：对白鼠从出生后第6天起每隔3天称一次体重，一直称到第18天，数据如表8。试计算日龄x与体重增长y的回归关系。

表8 白鼠6—18日龄的体重

序号 (i)	1	2	3	4	5
日龄 x_i (天)	6	9	12	15	18
平均体重 y_i (克)	12	17	22	25	29

表8的每对数据(x_i, y_i)对应于横轴为x，纵轴为y的坐标系中的一个点。如果把所有的点都在坐标系中标出，就得到图5。这种图将称之为散点图。从图5可以看出这些点都很接近于一条直线。于是自然就想到用一条直线来表示它们之间的关系。

我们知道，任何一条直线的方程都可写成

$$y = a + bx \qquad (4\text{-}1)$$

的形式，其中b表示直线的斜率，a表示直线与y轴的截距。如果a和b给定了，那么这条直

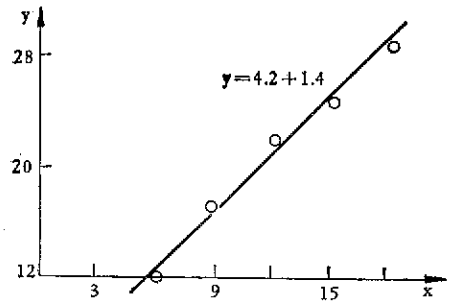

图 5

线就给定了，并且可以在坐标系上画出来。如果直线(4-1)是表 8 的数据所最接近的那条直线，则我们就称之为 y 对 x 的回归直线，b 称为回归系数。

在这个例子中，我们用直尺凭视觉也可以画一条直线，使得直线两边的点差不多相等，而且到这条直线的距离也差不多。这样一条直线就接近于回归直线了。但这样做不同的人就会画出不同的直线，有时还会相差较大。因此就需要给出一个判断回归直线好坏的标准。

如果自变量 x 取某个值 x_i 时，观测值为 y_i，而在回归直线上 x_i 所对应的 y 回归值为 $\hat{y}_i$，就有

$$\hat{y}_i = a + bx_i.$$

于是对于每一个 x_i，回归值 $\hat{y}_i$ 与观测值 y_i 的离差 $y_i - \hat{y}_i$ 表示了在点 (x_i, y_i) 回归线与观测值的误差。对于所观测的 n 个点 $(x_i, y_i)(i=1, 2, \cdots, n)$ 来说，如果我们给 a，b 以适当的值，使得离差的平方和

$$Q = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \qquad (4\text{-}2)$$

能够达到最小，则我们就认为由 a，b 决定的这条回归直线就是最好的。

由微积分中求极值的方法可知，使得平方和 Q 达到最小的 a，b 是存在的，它们分别由下式给出

$$b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum xy - (\sum x)(\sum y)/n}{\sum x^2 - (\sum x)^2/n},$$

$$a = \bar{y} - b\bar{x}, \qquad (4\text{-}3)$$

其中 $\bar{x} = \frac{1}{n}\sum x$，$\bar{y} = \frac{1}{n}\sum y$。如果引入记号

$$l_{xx} = \sum x^2 - (\sum x)^2/n,$$
$$l_{yy} = \sum y^2 - (\sum y)^2/n,$$
$$l_{xy} = \sum xy - (\sum x)(\sum y)/n,$$

那么就有

$$b = \frac{l_{xy}}{l_{xx}} \qquad (4\text{-}4)$$

具体计算时可列成下表(表 9)。

表 9 回归分析计算表

序号	x	y	x^2	y^2	xy
1	6	12	36	144	72
2	9	17	81	289	153
3	12	22	144	484	264
4	15	25	225	625	375
5	18	29	324	841	522
Σ	60	105	810	2383	1386

Σ 为各列的总和。由此可以得到

$n = 5$，$\sum x = 60$，$\sum y = 105$，

$\sum x^2 = 810$，$\sum y^2 = 2383$，$\sum xy = 1386$。

于是可以算得

$$\bar{x} = \frac{1}{n}\sum x = 60/5 = 12,$$

$$\bar{y} = \frac{1}{n}\sum y = 105/5 = 21,$$

$$l_{xx} = \sum x^2 - \frac{1}{n}(\sum x)^2 = 810 - 60^2/5 = 90,$$

$$l_{yy} = \sum y^2 - \frac{1}{n}(\sum y)^2 = 2383 - 105^2/5 = 178,$$

$$l_{xy} = \sum xy - \frac{1}{n}(\sum x)(\sum y) = 1386 - 60 \times 105/5 = 126;$$

$$b = l_{xy}/l_{xx} = 126/90 = 1.4,$$

$$a = \bar{y} - b\bar{x} = 21 - 1.4 \times 12 = 4.2.$$

所求的回归方程应该是

$$y = 4.2 + 1.4x。$$

这里的 b 是个正值，它表示随着日龄的增加，白鼠体重增加的趋势(图 5)。

（三）相关系数

前面所介绍的计算回归直线的方法，对于变量 x，y 的任何一组数据 (x_i, y_i)，$(i = 1, 2, \cdots, n)$ 都可以按上述步骤配出一条直线来。甚至对毫无关系的两个变量，由于抽样误差所致，也可以配出一条直线而且有可能是 $b \neq 0$。显然这时所配的直线并没有多大意义。只有当 x 和 y 之间确实存在某种线性关系时，配出的直线才有意义。因此我们还需要进一步找出检验回归直线有无意义的指标和方法。我们把描述两个变量线性关系密切程度的数量指标叫做样本相关系数（或简称相关系数），通常用 r 表示，它由下面的公式来计算

$$r = \frac{l_{xy}}{\sqrt{l_{xx}l_{yy}}} \qquad (4\text{-}5)$$

样本相关系数有如下几个性质：

1. 由 (4-5) 可以看出 r 的正负与 l_{xy} 是一致的。由 (4-4) 又知 l_{xy} 与 b 的正负是一致的。因此有：相关系数 r 与回归系数 b 的正负是一致的。

由于回归系数 b 刚好是回归直线的斜率，它的正负表示了当 x 增加时，变量 y 是增加，还是减小。如果 x 增加 y 也随着增加，这时 b 是正数，从而 r 也是正的，我们称之为正相关。否则，如果 x 增加时 y 减小，这时 r 为负值，就称为负相关。

2. 如果利用 a 的表达式 (4-3)，把总误差 Q 的表达式 (4-2) 改写为

$$\begin{aligned} Q &= \sum_{i=1}^{n}(y_i - bx_i - a)^2 \\ &= \sum_{i=1}^{n}(y_i - bx_i - \bar{y} + b\bar{x})^2 \\ &= \sum_{i=1}^{n}[(y_i - \bar{y}) - b(x_i - \bar{x})]^2 \\ &= \sum_{i=1}^{n}(y_i - \bar{y})^2 \\ &\quad - 2b\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) \\ &\quad + b^2\sum(x_i - \bar{x})^2 \\ &= l_{yy} - 2bl_{xy} + b^2 l_{xx} \end{aligned}$$

再利用 b 的表达式 (4-4) 得

$$\begin{aligned} Q &= l_{yy} - 2\frac{l_{xy}}{l_{xx}}l_{xy} + \frac{l_{xy}^2}{l_{xx}^2}l_{xx} \\ &= l_{yy} - \frac{l_{xy}^2}{l_{xx}} = l_{yy}\left(1 - \frac{l_{xy}^2}{l_{xy}l_{yy}}\right) \\ &= l_{yy}(1 - r^2) \qquad (4\text{-}6) \end{aligned}$$

因为 $Q \geqslant 0$ 且 $l_{yy} \geqslant 0$，所以有 $1 - r^2 \geqslant 0$。于是就得到相关系数的另一个性质：$|r| \leqslant 1$。

3. 据此我们就可以在散点图上说明当 r 取不同值时，散点分布的情形。

(1) $r = 0$，这时有 $l_{xy} = 0$，因此 $b = 0$。这也就是说回归方程应该是 $y = a$。它说明 y 与 x 没有线性关系。在这种情形下散点的分布丝毫也没有呈线性的趋势(图 6a)。

(2) $|r| = 1$，这时由 (4-6) 式可以看出应该有 $Q = 0$，由 (4-2) 式可以得到 $y_i = \hat{y}_i$，

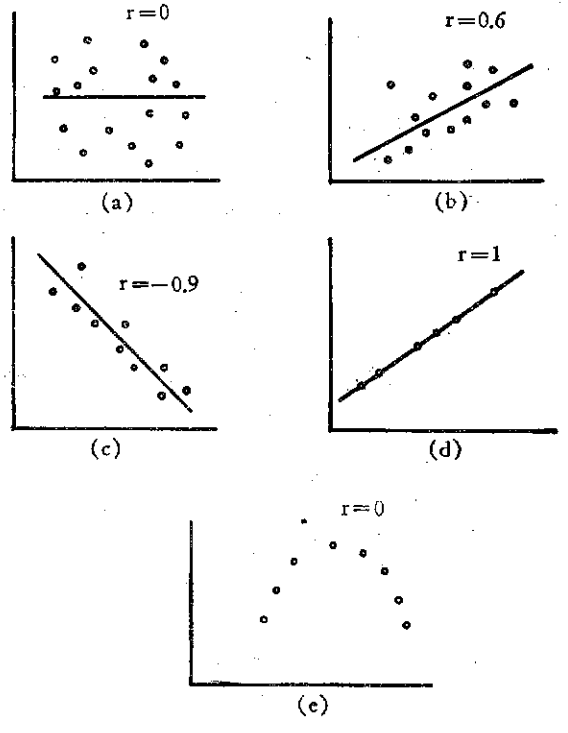

图 6

($i = 1, 2, \cdots, n$)。也就是说所有的观察值都在这条回归直线上，它们呈线性函数关系。我们称这样的关系为 y 与 x 完全线性相关（图 6d)。

(3) 如果 $0 < |r| < 1$，这时 x 与 y 间可能存在着一定的线性关系。由 (4-6) 式可以看出，当 r 的绝对值比较小时，Q 值就比较大。这说明观测点与回归线的总误差比较大，也就是说散点离回归直线较为分散。当 r 的绝对值逐渐增大以致于接近于 1 时，Q 就逐渐地减小接近于零，也就是说散点就逐渐靠近回归直线以致于呈函数关系(图 6b, c)。

由此可以看出，相关系数 r 确实可以表示两个变量 x 与 y 之间线性相关的密切程度。$|r|$ 愈小，x，y 之间的线性相关程度愈小；反之，$|r|$ 愈大，愈接近 1，x 与 y 间的线性相关关系就愈密切。

还要指出，相关系数只表示 x 与 y 间线性关系的密切程度。当 $|r|$ 很小，甚至等于零时，也不一定表示 x 与 y 之间不存在着相互依赖的关系。图 6e，虽然 $r = 0$，但从图上可看出 x 与 y 间的依赖关系是明显的。只不过这种关系不是线性关系而已。

(四) 线性相关关系的显著性

对于随机变量，由于抽样误差的存在只有当样本相关系数的绝对值大到一定程度时才表示它们之间存在着线性关系。这时我们所配的回归直线才有意义。在这种情况，我们说相关系数或相关关系是显著的。

对于两个正态随机变量 x，y，使其样本相关系数 r 在一定的显著水平 α 下的临界值还与样本对的个数 n 有关。r 值表（见表 10）给出了对不同的 n 值，在两种显著水平 α(0.05, 0.01) 下相关系数的临界值 r_α。

在我们的例子中，相关系数为

$$r = \frac{126}{\sqrt{90 \times 178}} = \frac{126}{\sqrt{16020}}$$

$$= \frac{126}{126.57} = 0.9954$$

由表 10 查得当 $n = 5$ 即 $n - 2 = 3$ 时 $r_{0.05} = 0.878$，$r_{0.01} = 0.959$。由于

$$|r| = 0.9954 > 0.959 = r_{0.01},$$

因此，白鼠的日龄与体重的线性关系是极显著的。故做的回归直线有意义。

(五) 能化为直线回归的曲线回归

在生物学中，有时两个变量之间的关系并

表 10 r 值表

$n-2$	0.05	0.01	$n-2$	0.05	0.01	$n-2$	0.05	0.01
1	0.997	1.000	10	0.576	0.708	19	0.433	0.549
2	0.950	0.990	11	0.553	0.684	20	0.423	0.537
3	0.878	0.959	12	0.532	0.661	25	0.381	0.487
4	0.811	0.917	13	0.514	0.641	30	0.349	0.449
5	0.754	0.874	14	0.497	0.623	35	0.325	0.418
6	0.707	0.834	15	0.482	0.606	40	0.304	0.393
7	0.666	0.798	16	0.468	0.590	45	0.288	0.372
8	0.632	0.765	17	0.456	0.525	50	0.273	0.354
9	0.602	0.735	18	0.444	0.561	60	0.250	0.325

表 11 6—16 日龄鸡胚胎的干重

日 龄 x	6	7	8	9	10	11	12	13	14	15	16
干重(克) y	0.029	0.052	0.079	0.125	0.181	0.261	0.425	0.738	1.130	1.882	2.812
$w = \lg y$	−1.538	−1.284	−1.102	−0.903	−0.742	−0.583	−0.372	−0.132	0.053	0.275	0.449

不一定是线性关系，而是某种曲线相关关系。例如一些比较简单的生长现象就是按照指数增长的规律变化，其它还有如二次关系、对数关系、双曲线关系等等都是曲线的关系。做曲线回归的方法也很多。这里只通过一个例子简单地介绍一下通过适当的变换把曲线回归的问题转化为直线回归问题来解决的方法。

例：表 11 记录着 6 到 16 天内鸡胚胎干重量的资料。如果把 x，y 的数据标在直角坐标系中，从散点图可以看出重量 y 随着日龄 x 的增加上升得愈来愈快（图 7）。散点基本上呈指数曲线的形式，又由生物学中的指数增长的模型启发我们很自然就想到了 x，y 之间可能按下述的指数关系变化。

$$y = ab^x, \qquad (4\text{-}7)$$

其中 a，b 是需要估计的参数。对这个式子稍加变换，两边取对数就得到

$$\lg y = \lg a + (\lg b)x。$$

如果用 w 表示变量 $\lg y$，就有

$$w = a' + b'x, \qquad (4\text{-}8)$$

这里 $a' = \lg a$，$b' = \lg b$。

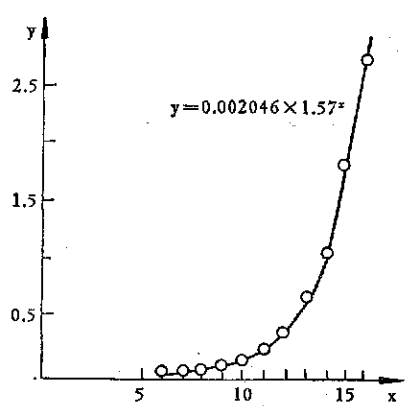

图 7

式（4-8）表明变量 x 和 $w = \lg y$ 之间是线性相关的关系。

如果把干重 y 的观测值分别取对数，所得的值做为 w 的观测值记入表 4-4 的最后一行。那么我们就可以用资料 (x_i, w_i) 按前面介绍的方法做直线回归，得到

$$b' = 0.1959,$$

$$a' = -2.689,$$

$$r = 0.9992,$$

$$w = 0.1959x - 2.689。$$

由于相关系数 r 相当高，故 w 与 x 的直线回归关系极显著。

再把 x 与 w 的关系变回到 x 与 y 的关系上去。由于 $\lg a = -2.689$，$\lg b = 0.1959$，因此得 $a = 0.002046$，$b = 1.57$。于是就得到了所求的回归曲线

$$w = 0.002046 \times 1.57^x,$$

或者写成下面的形式

$$w = 0.002046e^{0.4511x}。$$

在用化为直线回归的方法做曲线回归时，主要在于曲线类型的确定和变换函数的选择。

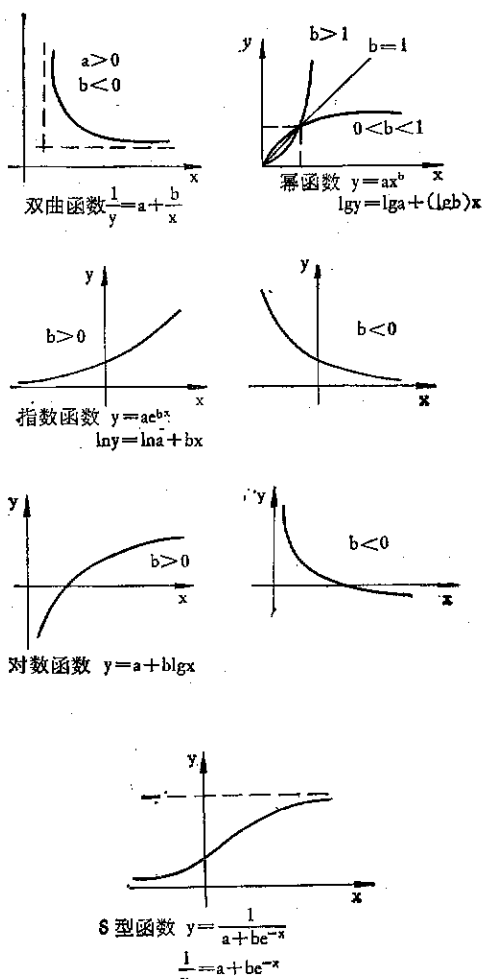

图 8

至于究竟选择那一类曲线来配合观测数据，一般来说是不容易的，主要是靠生物学的知识来定。如果在生物学上的含意尚不清楚，就可以从数学上去选择。这时可以根据散点图中散点所呈的大致形状与已知函数的图形进行比较来选择。也可以从生物学与数学两方面相配合来选择曲线的类型。至于变换函数则要依赖于所要配合的曲线的类型而定。图 8 中我们给出了几个常用的函数图形及所用的变换函数，供读者使用时参考。

另外在做曲线回归时，也有个回归关系的显著性的问题。对于曲线回归的显著性检验限于篇幅就不做介绍了，读者可参考有关的书籍。