社会统计分析第十六讲

六月廿四日下午

昨天下午我们讲了净相关系数：$\gamma_{xy\cdot z}$；$\gamma_{xy\cdot zw}$；$\gamma_{xy\cdot zwT}$。这些净相关系数假定变量都是定距变量。它的作用很大，可以控制一个变量，也可以控制两个变量、三个变量等。

假如昨天讲的样本都是随机样本，进行总体性论时就可以用F检定法，即：

$$F=\frac{\text{估计总体之增减误差}}{\text{估计总体之剩余误差}}=\frac{\text{净}\gamma^2/df_1}{(1-\text{净}\gamma^2)/df_2}$$

由于 $\begin{cases} df_1=1 \\ df_2=n-k-2 \\ k=\text{控制变量数} \end{cases}$

代入公式就是

$$F=\frac{\text{净}\gamma^2}{1-\text{净}\gamma^2}(n-k-2)$$

如果我们只控制一个变量Z，即$\gamma_{xy\cdot z}$

$$F=\frac{\gamma^2_{xy\cdot z}(n-3)}{1-\gamma^2_{xy\cdot z}}\quad \text{其中 } df_1=1,\ df_2=n-1-2=n-3$$

这是一般相关系数之检定。

假如我们现在控制两个变量Z、W，即$\gamma_{xy\cdot zw}$，因$df_1=1$；$df_2=n-2-2$；$k=2$

所以 $F=\frac{\gamma^2_{xy\cdot zw}(n-4)}{1-\gamma^2_{xy\cdot zw}}$

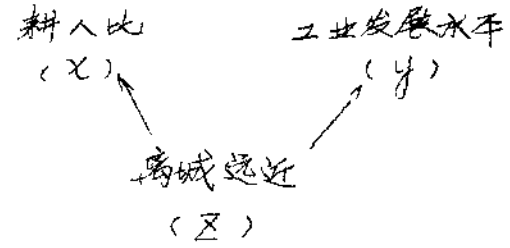
如果我们控制三个变量 Z、W、T。

同理：$F=\dfrac{\gamma^2_{xy\cdot ZWT}(n-5)}{1-\gamma^2_{xy\cdot ZWT}}$

以下控制多少变量可以同理类推。

现在给大家讲一个例子。

昨天讲过的表十一的资料是关于29个生产大队的耕人比，离城远近与工业发展水平的关系。假定这29个生产大队是从一个县里抽出的随机样本，三变量之关系如图所示：



其中 $\gamma_{xy\cdot Z}=-0.47$

$n=29$

这说明控制Z后，X与Y成反比关系。现在我们要问：如果按样本推论总体，在总体里控制Z后，X与Y是否还有关系。

这种推论可表示为 H_1：总体 $\gamma_{xy\cdot Z}\neq 0$

虚无假设则表示为 H_0：总体 $\gamma_{xy\cdot Z}=0$

怎样检定 H_0 呢？用F检定法。

因为这是控制一个变量，所以

$$F=\frac{\gamma^2_{xy\cdot Z}(n-3)}{1-\gamma^2_{xy\cdot Z}}=\frac{-0.47^2(29-3)}{1-(-0.47)^2}=7.372$$

根据 $df_1=1$，$df_2=29-3=26$ 查F表，$F_{0.05}\geq 4.225$。

而因为 $F=7.372>4.225$，所以可以否定 H_0，接受 H_1，即控制Z后，X与Y还有显著关系。

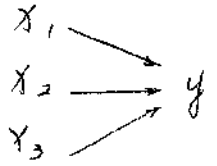
为什么要说显著关系呢？因为它的根据是显著水平。所表示的是X与Y在总体里的关系。今后我说显著关系就是指X与Y在总体内有关系（当然根据某显著水平而言）。

第十章 多因回归与複相关

第一节：什么是多因。

以前讲的是一个变量怎样受另一个变量影响，实际上任何一个变量是受很多变量影响的。

如：



这就是多因的问题。

例如研究犯罪问题，失业可能是犯罪的原因，都市化、人口家庭等都可能也是犯罪的原因。

又如研究青年人找对象为什么困难的原因，可能性别、城乡背景、教育水平、父母职业等各方面都有影响。

现在人们常关心理想子女数目，造成理想子女数目不一的原因可能是职业、年令、性别、教育水平、城乡背景、民族背景、经济水平、夫妻关系、亲友的儿女数目以及住户拥挤等。总之，从实际出发，引起事物变化往往是多因性的。

多因问题要从三方面理解：

①：了解所有这些因素，同时对Y的影响如何。

②：比较之下，哪个因素对Y的影响要重要。

③：知道各因素（自变量）之值后，如何予测依变量Y之值。

第二节：复相关系数（multiple Correlation）

这里要讲的内容，假定全部变量都是定距变量。复相关系数是研究一组自变量同时影响 y 时，它们的共同效力怎么样。

复相关系数用 R 表示。

R 的统计值是 0——1，没有正负之分，因为研究一组变量只看效果大小，不能有正负抵消的问题，R^2 有 PRE 的解释，叫多元决定系数。

现在看复相关系数的公式。

如果研究 $\left.\begin{matrix} X_1 \\ X_2 \end{matrix}\right\} \longrightarrow y$

那么复相关决定系数 $R^2_{y\cdot 12}$ 的公式如下：

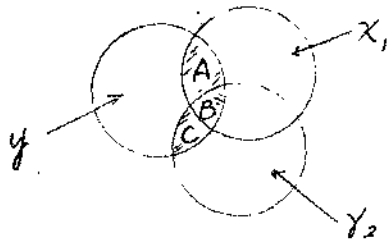
$$R^2_{y\cdot 12} = r^2_{y1} + (r^2_{y2\cdot 1})(1-r^2_{y1})$$

复相关系数 $R_{y\cdot 12}$ 的公式如下：

$$R_{y\cdot 12} = \sqrt{R^2_{y\cdot 12}}$$

简单讲一下公式的来历。

如图。



y 表示不知 X_1 和 X_2 时预测 y 的全部误差，（A+B）是已知 X_1 预测 y 时消减的误差，（B+C）是已知 X_2 预测 y 时消减的误差，已知 X_1、X_2 预测 y 时，同时消减的误差 = A+B+C，亦即 $R^2_{y\cdot 12}$，公式中的 r^2_{y1} 就是 A+B 那一部分，$(r^2_{y2\cdot 1})(1-r^2_{y1})$ 就是剩余的部分 C，加起来就表示 X_1 和 X_2 两个变量同时影响 y 时所消减之误差，所以它也有 PRE 的解释。

但是这个公式计算起来麻烦，根据统计学家推算，可简化为如下：

$$R^2_{y\cdot 12}=\frac{r^2_{y1}+r^2_{y2}-2r_{y1}r_{y2}r_{12}}{1-r^2_{12}}$$

假如研究 y 同时受 x_1, x_2, x_3 的影响，即 $\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}\longrightarrow y$ 那么，

$$R^2_{y\cdot 123}=R^2_{y\cdot 12}+(r^2_{y3\cdot 12})(1-R^2_{y\cdot 12})$$

复相关系数的公式便是：

$$R_{y\cdot 123}=\sqrt{R^2_{y\cdot 123}}$$

如果是：

$\begin{pmatrix}x_1\\x_2\\x_3\\x_4\end{pmatrix}\longrightarrow y$ 则其公式是：

$$R^2_{y\cdot 1234}=R^2_{y\cdot 123}+(r^2_{y4\cdot 123})(1-R^2_{y\cdot 123})$$

（注意：关于 R 之修正，参看第十九讲）

现举例如下：

假定仍用表十一的资料。

$\begin{cases}x_1=\text{离城市远近}\\x_2=\text{公社大队耕人比}\\y=\text{大队工业化程度}\end{cases}$ $\begin{pmatrix}x_1\\x_2\end{pmatrix}\longrightarrow y$

$\begin{cases}r_{y1}=-0.51\\r_{y2}=-0.64\\r_{y3}=+0.67\end{cases}$ 则 $R^2_{y\cdot 12}=\dfrac{r^2_{y1}+r^2_{y2}-2r_{y1}r_{y2}r_{12}}{1-r^2_{12}}$

$$=\frac{(-0.51)^2+(-0.64)^2-2(-0.51)(-0.64)(0.67)}{1-(0.67)^2}$$

$$=\frac{0.23}{0.55}=0.42$$

$R_{y\cdot 12}=\sqrt{R^2_{y\cdot 12}}=\sqrt{0.42}=0.65$ 即：0 ——△—— 0.65 —— +1

说明了 X_1, X_2 对 y 同时有影响，消减误差达 42%。

统计推论：

设 $\begin{cases} H_1: R \neq 0 \text{（总体）} \\ H_0: R = 0 \end{cases}$

$df_1 = k$

$df_2 = n - k - 1$

k — 自变量的数目，则

$$F = \frac{R^2}{1-R^2}\left(\frac{n-k-1}{k}\right)$$

代入上述二十九个大队的例子：$n = 29$　$k = 2$　$R^2 = 0.42$ 则，

$$F = \frac{0.42}{1-0.42}\left(\frac{29-2-1}{2}\right) = 9.414$$

$df_1 = k = 2$

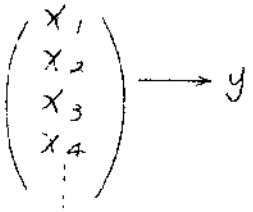
$df_2 = n - k - 1 = 29 - 2 - 1 = 26$

要求 $\alpha < 0.05$ 则 $F_{0.05} \geq 3.369$，否定 H_0，说明了 X_1, X_2 对 y 都同时有影响（在总体内）

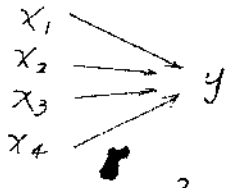
第三节：<u>多因直线回归分析</u>

上述偏相关系数。

其模式是：



现讲多因直线回归分析，其模式则是：



多因直线迴归分析的作用是：

一、估计 y 值；

二、比较各个 X 的效力。

以前讲过的简单直线迴归分析，是研究一个 X 与一个 Y 的关系，而现在讲的多因直线迴归分析，则是讲的多个 X 与一个 y 的关系。

还是让我们再以简单直线迴归分析谈起。

$X \longrightarrow Y$

$y' = bx + a$，这方程式内之 b，其实是 b_{yx}，表示 b 为 y 受 X 影响的 b，亦即说，$b_{yx} \neq b_{xy}$，因 $X' = b_{xy} Y + a$。故

$$b_{yx} = \frac{\triangle y'}{\triangle X}$$

b 的大小受易度单位影响，所谓易度单位，就是指易度以什么单位来计算。比如时间的易度单位，可以是年，可以是月，也可以是日，易度的单位不同，b 值的大小，也是不相同的。比如：

教育年期（X）⟶ 理想子女数目（y）

就是以年为易度单位，因此，如果易度单位改为月，b 的大小亦会随之改变。

因为 b 的大小视易度单位而定，故难于比较，例如：

$$\left.\begin{aligned} y' &= b_1 X_1 + a_1 \\ y' &= b_2 X_2 + a_2 \end{aligned}\right\} b_1 与 b_2 之大小视于 X_1 与 X_2 之易度单位，故 b 可用于予测，而不可用于比较。$$

要为了比较 b，则需使 b 标准化，亦即使每一变易之值变为标准值（Standard Score），就不受易度单位影响。

例如：

$X \Longrightarrow Z_x$

$y \Longrightarrow Z_y$

$$Z_x = \frac{X - \overline{X}}{S_x}$$

$$Z_y = \frac{y - \overline{y}}{S_y}$$

这样，X y就都变为标准值Z

变易X的标准化就是以它自己的均值为0点，然后，每个变易值之距离以它自己变易的标准差（S）为单位。y变易同理。

例如《表八》：

X	y	Z_x	Z_y
2	5	−1.09	+1.59
2	4	−1.09	+1.06
3	4	−0.55	+1.06
3	3	−0.55	+0.53
4	1	0.00	−0.53
4	1	0.00	−0.53
4	0	0.00	−1.06
6	0	+1.09	−1.06
8	0	+2.19	−1.06
$\overline{X}=4$	$\overline{y}=2$	$\overline{Z}_x=0$	$\overline{Z}_y=0$
$S_x=1.83$	$S_y=1.89$	$S'_x=1$	$S'_y=1$

怎样把X变成标准值呢？以表中第一名妇女所给的数值计算如下：

$$\begin{cases} Z_x = \dfrac{2-4}{1.82} = -1.09 \\ Z_y = \dfrac{5-2}{1.89} = 1.59 \end{cases}$$

由上可见，只要变为标准值，不管任何变量，其均值皆为0，其标准差皆为1。因此：

$$\begin{cases} y = bX + a \\ Z'y = \beta Z_x, \end{cases}$$

β 是标准化回归系数 = 通用权数 (beta weight)

$$\beta = \beta_{yx} = \frac{\Delta Z'y}{\Delta Z x}$$

$$= b_{yx}\left(\frac{S_x}{S_y}\right)$$

为何 $Z'y = \beta Z_x$ 的后面没有 a ？

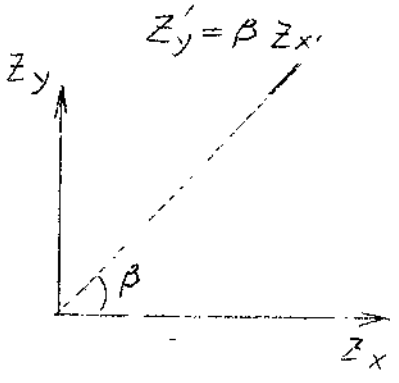
因为 $a = \bar{y} - b\bar{x}$

$A = \bar{Z}_y - \beta \bar{Z}'_x$

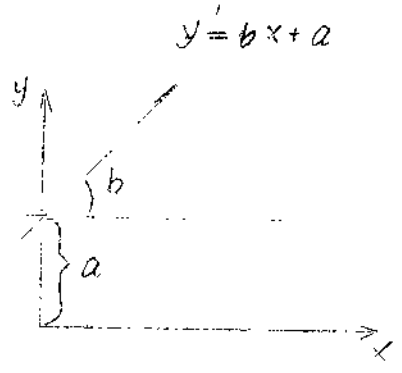
而 $\bar{Z}_y = 0 \quad \bar{Z}_{x'} = 0$

所以 $A = 0$ 如下图所示：

图(一)



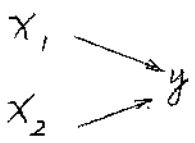
图(二)

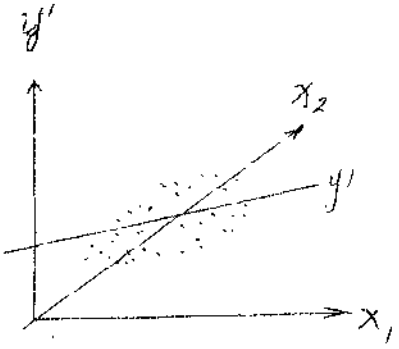


由图(一)可见，当 $Z'y = \beta Z_x$ 时后面没有截距 A，故 $Z'y = \beta Z x$ 这条线通过0点。

由图(二)可见，当 $y' = bX + a$ 时后面有截距 a，故 $y' = bx + a$ 这条线不一定通过0点。

现在讲两个变量同时影响 y。





$y' = b_1 X_1 + b_2 X_2 + a_{1.2}$ 多元直线回归方程式

$z'_y = \beta_1 z_1 + \beta_2 z_2$ 标准化多元直线回归方程式

三条线都为直角关系，表示三度空间，用一条直线穿过所有的点，这条直线根据最小平方法求出。所以根据 X_1, X_2 的数值来估计 y 值，其错误最小。

标准化直线回归方程式有比较效力的作用。

以上两个方程式有不同的作用。首先，让我们解释 b_1 和 b_2 这两个系数。

b_1 实际上应当写作 $b_{y1.2} = \frac{\triangle y'}{\triangle X_1}$ = 净回归系数，即在控制 X_2 以后，X_1 变化引起 y 变化的数值。

b_2 实际上应当写作 $b_{y2.1} = \frac{\triangle y'}{\triangle X_2}$ = 净回归系数，即在控制 X_1 以后，X_2 变化引起 y 变化的数值。

$a_{1.2}$ = 截距 $= \bar{y} - b_1(\bar{x}_1) - b_2(\bar{x}_2)$

计算标准化回归系数同理：

$\beta_1 = \beta_{y1.2}$ = 控制 X_2 以后的标准净回归系数 = 回归权数。

$\beta_2 = \beta_{y2.1}$ = 控制 X_1 以后的标准净回归系数 = 回归权数。

怎样计算 $b_1, b_2, \beta_1, \beta_2$ 呢？根据统计学家推算，用最小

平方法，那么计算公式为：

$$\begin{cases} \beta_1 = \beta_{y1.2} = \dfrac{\gamma_{y1} - \gamma_{y2}\gamma_{12}}{1-\gamma_{12}^2} \\ \beta_2 = \beta_{y2.1} = \dfrac{\gamma_{y2} - \gamma_{y1}\gamma_{12}}{1-\gamma_{12}} \end{cases}$$

$$\begin{cases} b_1 = \beta_1 \left(\dfrac{S_y}{S_1}\right) \\ b_2 = \beta_2 \left(\dfrac{S_y}{S_2}\right) \end{cases}$$

$$a_{12} = \overline{y} - b_1(\overline{x}_1) - b_2(\overline{x}_2)$$

例：

(X_1) 离城远近 ——→ 工业化程度 (y)

(X_2) 耕地与人口比率 ——→ 工业化程度 (y)

假如：$\gamma_{y1} = -0.51$　　$S_y = 3.61$

$\gamma_{y2} = -0.64$　　$S_1 = 3.02$

$\gamma_{12} = +0.67$　　$S_2 = 0.21$

$R = 0.65$　　$\overline{y} = 4.72$

$\overline{x}_1 = 5.74$

$\overline{x}_2 = 1.23$

那么我们问离城远近和耕人比率那一个因素对工业化影响更大一些？计算：

$$\beta_1 = \frac{(-0.51) - (-0.64)(0.67)}{1-(0.67)^2} = 0.15$$

$$\beta_2 = \frac{(-0.64) - (-0.51)(0.67)}{1-(0.67)^2} = -0.54$$

$$b_1 = (-0.15)\left(\frac{3.61}{3.02}\right) = -0.18$$

$$b_2 = (-0.54)\left(\frac{3.61}{0.21}\right) = -9.28$$

$$a_{12} = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 = 4.72 - (-0.18)(5.47) - (-9.28)(1.23)$$
$$= 17.11$$

所以 $\begin{cases} y' = -0.18x_1 - 9.28x_2 + 17.11 \\ z'_y = -0.15z_1 - 0.54z_2 \end{cases}$

可见，耕地与人口比率比离城远近对工业化的影响更大些。

怎么预测呢？以表十一中第十五个大队为例。该大队离城7.8公里，耕人比1.23亩/每人 $\begin{cases} x_1 = 7.8 \\ x_2 = 1.23 \end{cases}$

根据 x_1, x_2 估计 y（工业化程度）。将以上数值代入公式：

$y' = -0.18(7.8) - 9.28(1.32) + 17.11 = 3.5$。

根据计算，我们就可以回答这个大队有3.5%的人口是参加工业的。（实际上是3.9%，相差不多）。

现在讲三个变量：

公式是：

$$\begin{cases} y' = b_1x_1 + b_2x_2 + b_3x_3 + a_{123} \\ z'_y = \beta_1z_1 + \beta_2z_2 + \beta_3z_3 \end{cases}$$

其中：$b_1 = b_{y1.23} = \beta_1\left(\frac{S_y}{S_1}\right)$

$b_2 = b_{y2.13} = \beta_2\left(\frac{S_y}{S_2}\right)$

$b_3 = b_{y3.12} = \beta_3\left(\frac{S_y}{S_3}\right)$

$a_{123} = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 - b_3\bar{x}_3$

从以上几个式子可以看出，只要求出 $\beta_1, \beta_2, \beta_3$，其它就易计算了。

据统计学家推论：

$$\beta_1 + \gamma_{12}\beta_2 + \gamma_{13}\beta_3 = \gamma_{y1} \quad \cdots\cdots ①$$

$$\gamma_{12}\beta_1 + \beta_2 + \gamma_{23}\beta_3 = \gamma_{y2} \quad \cdots\cdots ②$$

$$\gamma_{31}\beta_1 + \gamma_{32}\beta_2 + \beta_3 = \gamma_{y3} \quad \cdots\cdots ③$$

又因为 γ 不计方向（$x \longleftrightarrow y$），故 $\gamma_{32} = \gamma_{23}$。

运用代数运算，可根据上面三个方程式解出三个未知数 β_1、β_2、β_3。

设：$\begin{cases} \gamma_{y1} = 0.6735 \\ \gamma_{y2} = 0.5320 \\ \gamma_{y3} = 0.3475 \end{cases}$ $\begin{cases} \gamma_{12} = 0.1447 \\ \gamma_{13} = 0.3521 \\ \gamma_{23} = 0.0225 \end{cases}$

代入方程式①、②、③，则：

$$\beta_1 + 0.1447\beta_2 + 0.3521\beta_3 = 0.6735$$

$$0.1447\beta_1 + \beta_2 + 0.0225\beta_3 = 0.5320$$

$$0.3521\beta_1 + 0.0225\beta_2 + \beta_3 = 0.3475$$

解得：$\begin{cases} \beta_1 = 0.5593 \\ \beta_2 = 0.4479 \\ \beta_3 = 0.1405 \end{cases}$

又设：$\begin{cases} S_y = 2.9469 \\ S_1 = 2.6651 \\ S_2 = 2.1151 \\ S_3 = 2.2100 \end{cases}$ $\begin{cases} \bar{y} = 5.50 \\ \bar{x}_1 = 4.95 \\ \bar{x}_2 = 5.50 \\ \bar{x}_3 = 5.40 \end{cases}$

把以上数字代入 b_1、b_2、b_3 公式及 a_{123} 公式：

$$b_1 = 0.5593\left(\frac{2.9469}{2.6651}\right) = 0.6184$$

$$b_2 = 0.4479\left(\frac{2.9469}{2.1151}\right) = 0.6240$$

$$b_3 = 0.1405\left(\frac{2.9469}{2.2100}\right) = 0.1873$$

$$a_{123} = 5.50 - (0.6184)(4.95) - (0.6240)(5.50) - (0.1873)(5.40)$$
$$= -2.0045$$

所以：

$$\begin{cases} y' = 0.6184X_1 + 0.6240X_2 + 0.1873X_3 - 2.0045 \\ Z'_y = 0.5593Z_1 + 0.4479Z_2 + 0.1405Z_3 \end{cases}$$

据 y' 公式，可知道 X_1、X_2、X_3 可预测 y 值，

据 Z'_y 公式，可比较 X_1、X_2、X_3 的效力，$X_1 > X_2 > X_3$

又如：失业率 X_1 ——→ 犯罪率 y

都市化 X_2 ——→ 犯罪率 y

教育水平 X_3 ——→ 犯罪率 y

解得：$Z'_y = 0.75Z_1 + 0.21Z_2 - 0.34Z_3$

比较 X_1、X_2、X_3 的效力，$X_1 > X_3 > X_2$，其中 X_3（教育水平）对 y（犯罪率）有负方向的影响。

但是，我们在使用 β（回归权数）时，还要注意统计累赘（redundancy）问题。

因为，假如有三个以上自变量，我们常用的公式是 $Z'_y = \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3$ 来比较 X_1、X_2、X_3 的效力。其中的 β 都是净回归，即：

$\beta_1 = \beta_{y1.23}$　即控制 X_2、X_3

$\beta_2 = \beta_{y2.13}$　即控制 X_1、X_3

$\beta_3 = \beta_{y3.21}$　即控制 X_1、X_2

如果，X_1 与 X_2 关系很密，而 X_3 与前二者之关系不大时，就会产生如下现象：

控制 X_2、X_3 时，结果 β_1 数值很小；控制 X_1、X_3 时，β_2 数值也会很小了而控制 X_1、X_2 时，由于不影响 X_3 之效力，β_3 二数值就会较大了。

这样，β_3 就显得很重要：我们得出 X_3 的效力大于 X_1、X_2 就不一定准确了。这就是统计累赘问题。

因此，在运用回归权数比较 X_1、X_2、X_3 的效力时，X_1、X_2、X_3 之间的相关系数要差不多，$r_{12} \fallingdotseq r_{23} \fallingdotseq r_{13}$，才能避免统计累赘。

当然，如果我们只根据偏回归系数（b）来推测 y' 的值，就不必考虑这个问题了。

社会统计分析第十七讲

六月二十五日下午讲

$$y' = b_1 X_1 + b_2 X_2 + a_{12}$$

$$Z'_y = \beta_1 Z_1 + \beta_2 Z_2$$

上面两个公式是多因直线回归分析方程式。现在我们讨论推论到总体的情况。一般有两种做法：

① $\begin{cases} H_1: & \text{在总体为 } b_1 \text{ 或 } b_2 \neq 0 \text{ 即 } \beta_1 \text{ 或 } \beta_2 \neq 0 \text{ 或 } R \neq 0 \\ H_0: & \text{〃} \quad b_1 = b_2 = 0 \text{ 即 } \beta_1 = \beta_2 = 0 \text{ 或 } R = 0 \end{cases}$

以上的虚无假设就是说两个自变量 X_1 和 X_2 同时与倚变量 y 无关。问题是我们如何来鉴定这一虚无假设，此时用以前讲过的 F 检定法。

即：$F = \dfrac{R^2}{1-R^2}\left(\dfrac{n-k-1}{k}\right)$，而 $df_1 = k$，$df_2 = n-k-1$

（其中 k 为自变量的数目）

② $\begin{cases} H_1: b_1 \neq 0 \\ H_0: b_1 = 0 \end{cases}$ 和 $\begin{cases} H_1: b_2 \neq 0 \\ H_0: b_2 = 0 \end{cases}$

也就是说将自变量分开来进行检定，用F检定法先对b_1检定，然后同理对b_2检定。下面是对b_1的检定公式：

$$F = \frac{\gamma_{y(1\cdot2)}^2}{1-R_{y\cdot12}^2}(n-k-1)$$

$$df_1 = 1$$

$$df_2 = n-k-1$$

这里注意：

$$\gamma_{y(1\cdot2)} \neq \gamma_{y1\cdot2}$$

因为$\gamma_{y1\cdot2}$ = 净相关像数，

而$\gamma_{y(1\cdot2)}$ = 局部相关像数（Part Correlation）

根据统计推论，局部相关像数的计算公式为：

$$\gamma_{y(1\cdot2)}^2 = \frac{\gamma_{y1} - \gamma_{y2}\gamma_{12}}{\sqrt{1-\gamma_{y2}^2}}$$

所以对b_1的检定首先应算出$\gamma_{y(1\cdot2)}$的值，然后代入即可计算得F值。同理对b_2的检定公式：

$$F = \frac{\gamma_{y(2\cdot1)}^2 \cdot (n-k-1)}{1-R_{y\cdot12}^2}$$

其中 $\gamma_{y(2\cdot1)}^2 = \frac{\gamma_{y2} - \gamma_{y1}\gamma_{12}}{\sqrt{1-\gamma_{y1}^2}}$ ，$df_1 = 1$，$df_2 = n-k-1$

第四节 逐步回归分析之程序

上一节谈到，在多元回归方程式（如$y' = b_1X_1 + b_2X_2 + b_3X_3 + a_{123}$）中，各个自变量（如$X_1$、$X_2$、$X_3$）同时影响依变量$y'$。

逐步回归分析则不同，它是各个自变量依先后次序进入回归方程式的。如：第一步，$y' = b_2 X_2 + a_2$

第二步，$y' = b_2 X_2 + b_1 X_1 + a_{21}$

第三步，$y' = b_2 X_2 + b_1 X_1 + b_3 X_3 + a_{213}$

这样就能依次序逐步分析一个两个乃至三个变量对y的不同影响，即每放入一个变量就能计算一次消减误差之增大。见下表。

次序	方程式	PRE	增加之PRE	含义(阴影部分)
第一步	$y' = b_2 X_2 + a_2$	r^2_{y2}	r^2_{y2}	
第二步	$y' = b_2 X_2 + b_1 X_1 + a_{21}$	$R^2_{y.12}$	$R^2_{y.12} - r^2_{y2}$	
第三步	$y' = b_2 X_2 + b_1 X_1 + b_3 X_3 + a_{213}$	$R^2_{y.123}$	$R^2_{y.123} - R^2_{y.12}$	

为什么要研究增加之PRE呢？这是为了挑选比较有效力的自变量，譬如当放入第二个变量时，PRE增加不少，说明放入之变量效力很高，而再放入第三个变量时，PRE的增加甚微数，说明此变量对y影响很小，可以不必考虑。这在自变量众多时帮助取舍自变量，以简化资料。

这样产生的一个问题是，哪一个自变量首先放入呢？即先后次序如何确定？有两种方法。第一种是根据研究理论来确定先后次序，社会学家对影响倚变量之诸因素加以分析，认为在理论上较重要的自变量首先放入方程式。例如：

y = 个人现代化

X_1 = 入息

X_2 = 教育

X_3 = 智能