文章编号：1006-4354 (2012) 01-0015-04

通径分析在 Excel 和 SPSS 中的实现

杜　鹃

(宝鸡市气象局，陕西宝鸡　721006)

摘　要：选取一组气象要素为数据源，分别利用 Excel2007 和 SPSS17.0 进行通径分析计算和显著性检验。分析结果表明：Excel2007 通径分析实现步骤较为复杂，SPSS17.0 实现通径分析简单快捷，且可直接得到最优回归方程和直接通径系数。两者在作统计分析前都要对数据进行正态分布检验，分析结果要进行显著性检验，直至回归方程中各变量检验均为显著水平为止，最优的路径选择由决策系数决定。

关键词：通径分析；回归分析；Excel2007；SPSS17.0

中图分类号：P413　　**文献标识码**：A

在研究多个相关变量间的线性关系时，除了可采用多元线性回归分析和偏相关分析外，还可采用通径分析（path analysis）。通径分析是数量遗传学家 Sewall Wright 于 1921 年提出来，经遗传育种学者不断改进和完善形成的一种多元统计技术。它通过对自变量和因变量间的相关分解来研究因变量（性状）的相对重要性，已在众多领域广泛应用。但由于分解相关系数的通径分析技术计算量比较大，很多计算手工难以实现，计算结果不够精确。通径分析过程可通过 SAS、DPS 等专业统计学软件实现，但一些软件处理需进行复杂编程，目前以 Excel 和 SPSS 运用最为广泛[1]。为介绍如何利用这两软件实现通径分析自动化，选取靖边县 1980—2008 年有关气象资料，说明通径分析在 Excel2007 和 SPSS17.0 中的具体实现过程，为有关资料的通径分析提供参考。

1　原理与方法

通径分析是在多元回归的基础上将相关系数分解为直接通径系数（某一自变量对因变量的直接作用）和间接通径系数（该自变量通过其他自变量对因变量的间接作用）。通径分析的理论证明，任一自变量 X_i 与因变量 Y 之间的简单相关系数（r_{iy}）等于 X_i 与 Y 的直接通径系数（P_{iy}）与所有 X_i 与 Y 的间接通径系数（P_{ij}）之和，即 X_i 对 Y 的总作用。当许多自变量共同影响一个因变量时，每个自变量对因变量的重要性是不同的，其中一个自变量可能通过其他自变量对因变量起作用，此时可用间接通径系数表示。如 X_i 通过 X_j 对 Y 的间接通径系数 $P_{ij}=r_{ij}\,P_{jy}$。为检验模型

收稿日期：2011-08-31

作者简介：杜鹃（1984—），女，陕西宝鸡人，学士，助理工程师，主要从事办公政务工作。

参考文献：

[1] 李建芳，李建军，郭清厉，等. 宝鸡市气候服务系统 [J]. 陕西气象，2005 (2)：15-16.

[2] 魏凤英. 现代气候统计诊断与预测技术[M]. 2 版. 北京：气象出版社，2007：5.

[3] 谭浩强. Visual Basic 程序设计教程 [M]. 北京：清华大学出版社，2002：2-10.

[4] Visual Basic 的数据库编程 [EB/OL]. [2005-10-08]. http://tech.163.com/05/1008/10/1VHM7G8700091589_2.html.

[5] 刘志铭. Visual Basic 数据库开发实例解析 [M]. 北京：机械工业出版社，2003：8.

[6] 雷向杰，杨凌，刘文泉，等. 陕西省短期气候预测业务质量评估 [J]. 陕西气象，2009 (5)：29-32.

[7] 雷向杰. 短期气候预测质量评估方法与业务考核办法 [J]. 陕西气象，2008 (6)：25-28.

的优劣，需要进一步计算剩余通径系数（b_{ey}），如果剩余通径系数较小，说明已找出主要变量，否则表明误差较大或者还有更重要的因素需要考虑在内。剩余通径系数（b_{ey}）表示为

$$b_{ey}=(1-D^2)^{1/2},$$

其中 $D^2=\sum_{k=1}^{i}P_{ky}\,r_{ky}$，是决定系数。

在通径分析中，一般认为最难计算的就是通径系数。事实上，通过软件进行线性回归计算，计算结果给出的线性回归方程的标准系数即通径系数，再乘以相关系数即可获得间接通径系数[2]。

2 通径分析在 Excel2007 中的实现

2.1 描述性统计分析

资料源于陕西省气象局档案馆，选取靖边县1980—2008年8月的月蒸发量Y、月降雨量X_1、月平均温度X_2、月日照时数X_3、月平均相对湿度X_4和月平均风速X_5为实验数据，用以分析各气候要素对蒸发量的影响。Excel2007的通径分析功能在“工具”菜单栏中的“数据分析”中。描述性统计和正态检验结果显示各气象因子的偏斜度都较小，均接近于0，说明各组数据近似满足正态分布要求，可进行通径分析。

2.2 回归分析与显著性检验

按照蒸发量为因变量Y，其他要素为自变量录入数据，选定回归分析控制项，建立以月蒸发量为因变量，其他因子为自变量的多元回归线性方程：$Y=280.733+0.026X_1+1.081X_2+0.217X_3-3.123X_4+23.365X_5$，模型检验达极显著水平（$F=30.137$，显著性$F$检验$F_{Sig}=2.25E-09$），偏回归系数显著性检验结果为$X_4$、$X_5$极显著，$X_3$（$t=2.372$，$p=0.026$）接近显著水平，$X_1$（$t=0.519$，$p=0.609$）、$X_2$（$t=0.653$，$p=0.520$）不显著。因此先剔除$X_1$、$X_2$以建立最优回归方程。剔除$X_1$、$X_2$后建立的回归方程为：$Y=345.263\,7+0.192\,9X_3-3.402\,6X_4+23.228\,9X_5$，方差分析结果表明该方程达到显著水平（$F=52.774$，$F_{Sig}=5.866\,1E-11$）。各偏回归系数显著性检验结果（表1）表明，$X_4$、$X_5$达极显著水平，$X_3$达显著水平，说明对$Y$做关于$X_3$、$X_4$、$X_5$的通径分析是有意义的。

表1 各偏回归系数（剔除X_1、X_2后）显著性检验结果

	系 数	标准误差	t	p
a	345.263 7	52.012 8	6.638 1	5.90E−07
X_3	0.192 9	0.084 0	2.294 8	0.030 4
X_4	−3.402 6	0.483 0	−7.045 4	2.20E−07
X_5	23.228 9	7.180 8	3.234 9	0.003 4

2.3 通径分析及显著性检验

2.3.1 直接通径系数计算与显著性检验 首先，在“数据分析”中计算得到各要素间的相关系数。X_3、X_4、X_5与Y的相关系数分别为0.674、−0.878、0.537，查相关系数显著性临界值表得$r_{0.01}=0.505$，X_3、X_4、X_5与Y间都满足$|r|>r_{0.01}$，故均达到极显著相关水平。

将求得的自变量间和自变量与因变量Y间的简单相关系数$r_{x_ix_j}$和r_{x_iy}建立正则方程组

$$\begin{cases}P_{x_1y}+r_{x_1x_2}P_{x_2y}+r_{x_1x_3}P_{x_3y}=r_{x_1y}\\ r_{x_2x_1}P_{x_1y}+P_{x_2y}+r_{x_2x_3}P_{x_3y}=r_{x_2y}\\ r_{x_3x_1}P_{x_1y}+r_{x_3x_2}P_{x_2y}+P_{x_3y}=r_{x_3y}\end{cases}\quad(1)$$

用矩阵表示为

$$\begin{pmatrix}1 & r_{x_1x_2} & r_{x_1x_3}\\ r_{x_2x_1} & 1 & r_{x_2x_3}\\ r_{x_3x_1} & r_{x_3x_2} & 1\end{pmatrix}\begin{pmatrix}P_{x_1y}\\ P_{x_2y}\\ P_{x_3y}\end{pmatrix}=\begin{pmatrix}r_{x_1y}\\ r_{x_2y}\\ r_{x_3y}\end{pmatrix},\quad(2)$$

令 $\boldsymbol{R}=\begin{pmatrix}1 & r_{x_1x_2} & r_{x_1x_3}\\ r_{x_2x_1} & 1 & r_{x_2x_3}\\ r_{x_3x_1} & r_{x_3x_2} & 1\end{pmatrix}$，$\boldsymbol{P}=\begin{pmatrix}P_{x_1y}\\ P_{x_2y}\\ P_{x_3y}\end{pmatrix}$，

$\boldsymbol{S}=\begin{pmatrix}r_{x_1y}\\ r_{x_2y}\\ r_{x_3y}\end{pmatrix}$，则 $\boldsymbol{RP}=\boldsymbol{S}$[3]，

之后求自变量相关矩阵$\boldsymbol{R}$的行列式$|\boldsymbol{R}|$（行列式值不为0，系数矩阵有逆矩阵，方程组有唯一解）来判断系数矩阵是否有逆矩阵。在公式栏中输入MDETERM（）函数，求得$|\boldsymbol{R}|=0.569$，故系数方程有逆矩阵$\boldsymbol{R}^{-1}$。在公式栏输入MINVERSE（）函数求得$\boldsymbol{R}^{-1}$，则$\boldsymbol{R}^{-1}\boldsymbol{S}$为直接通径系数矩阵，令$b_i=P_{iy}$，输入MMULT（）函数即可求得$b_i$。$X_3$对$Y$的直接通径系数$b_3=0.211$，$X_4$、

X_5 对 Y 的直接通径系数分别为 $b_4=-0.666$、$b_5=0.255$。

各通径系数显著性检验结果与回归系数检验结果一致，回归方程多元决定系数 $D^2=0.864$，表明因变量变异中 86.4% 可由回归部分解释，误差为 13.6%。据此可求出误差 e 对 Y 的直接通径系数，即剩余通径系数 $b_{ey}=0.369$。

2.3.2　间接通径系数计算及结果分析　X_i 通过 X_j 对 Y 的间接作用可由各自变量间的相关系数 r_{ij} 与 X_i 所对应的直接通径系数 b_i 乘积得到。如 X_3 通过 X_4 对 Y 的间接作用为 $P_{34}=r_{34}b_4=0.395$，X_3 对 Y 的总作用 r_{3y} 为直接作用与间接作用之和。决策系数 R^2（3）$=b_3^2+2b_3$（$P_{34}+P_{35}$）$=0.240$。从通径分析结果（表 2）可看出，各变量对 Y 的直接作用大小排序为 $b_5>b_3>b_4$，与 Y 的相关来看 $r_{3y}>r_{5y}>r_{4y}$，X_4 对 Y 的直接作用和通过其他变量对 Y 的间接作用皆为负；X_3 通过 X_4 对 Y 的间接作用大于 X_5 通过 X_3、X_4 对 Y 的间接作用之和，即使 $b_3<b_5$，但总作用 $r_{3y}>r_{5y}$。在复杂的路径信息中可用决策系数来选择最优路径。决策系数 R^2（4）$>R^2$（3）$>R^2$（5），故 X_4 为主要决策变量。

表 2　各自变量对因变量 *Y* 的通径系数

自变量	直接作用 b_i	间接作用 P_{ij}	总作用 r_{iy}	决策系数 R^2（i）
X_3	0.211	0.395（通过 X_4）；0.067（通过 X_5）	0.674	0.240
X_4	−0.666	−0.125（通过 X_3）；−0.087（通过 X_5）	−0.878	0.726
X_5	0.255	0.056（通过 X_3）；0.227（通过 X_4）	0.537	0.209
e	0.369		0.369	0.136

3　通径分析在 SPSS17.0 中的实现

3.1　对因变量正态性检验

SPSS17.0 中的通径分析在“分析”菜单下实现。首先对因变量 Y 进行正态分布检验，因变量 Y 的偏度为 −0.171，接近于 0，接近正态分布，为负偏态。峰值也为负值，说明 Y 的正态分布图为平峰图，Y 服从正态分布，可进行回归分析。

3.2　相关性分析及检验

选择“分析—相关—双变量”菜单输入变量值，各变量与 Y 相关性大小排序为 $r_{2y}>r_{3y}>r_{5y}>r_{1y}>r_{4y}$，且 X_1、X_4 与 Y 呈负相关（表 3）。但这样的相关关系不能完整反映两变量间线性相关程度，因为每两个变量间相关都有其他变量作用，要排除其他变量作用，需得出各变量间的偏相关系数。选择“分析－相关－偏相关”，当其他四变量被控制时，计算 X_1 与 Y 的偏相关系数 $R_{1y}=0.108$。自由度＝23，$r_{0.05}=0.396$，$r_{0.01}=0.505$，故 R_{1y} 不显著，同理计算求得 R_{2y} 也不显著，R_{3y}、R_{4y} 和 R_{5y} 显著。各变量与 Y 的偏相关系数与相关系数相比，除绝对值大小顺序不同，符号也不同。

表 3　自变量与因变量 *Y* 的相关系数

自变量	相关系数	偏相关系数
X_1	−0.411	0.108
X_2	0.684	0.135
X_3	0.674	0.443
X_4	−0.878	−0.685
X_5	0.538	0.537

3.3　回归分析及检验

选择“分析—回归—线性”，在方法选项中保持默认“强制进入法”，该方法要求系统在建立回归方程时把所选中的全部自变量都保留在方程中。表 4 给出回归方程中各变量的截距、标准误差、标准系数（即通径系数）及对应的显著性检验结果。由通径系数可看出各变量对 Y 的直接影响大小排序为 $X_5>X_3>X_2>X_1>X_4$，且 X_4 的直接影响为负。

各变量决策系数 R^2（1）$=-0.039\ 9$，R^2（2）$=0.098\ 2$，R^2（3）$=0.263\ 3$，R^2（4）$=0.669\ 6$，R^2（5）$=0.210\ 5$。大小排序为 R^2

(4) $>R^2$ (3)$>R^2$ (5) $>R^2$ (2) $>R^2$ (1)。各回归系数 t 检验结果表明 X_1、X_2 不显著，X_3、X_4、X_5 对 Y 影响十分显著，因此将 X_1、X_2 从回归方程中剔除，以建立最优模型。

表 4 五元一次方程回归系数及检验结果

	非标准化系数		标准误差(试用版)	t	p
	系数	标准系数			
a	280.825	106.329		2.641	0.015
X_1	0.026	0.050	0.046	0.522	0.607
X_2	1.799	2.759	0.076	0.652	0.521
X_3	0.217	0.092	0.237	2.367	0.027
X_4	−3.124	0.692	−0.611	−4.513	0.000
X_5	23.399	7.660	0.257	3.055	0.006

注：自由度为 23，$t_{0.05}=2.069$，$t_{0.01}=2.819$

在方法选项中选择“逐步进入法”直接得到最优模型，该方法是根据方差分析结果选择符合判据的自变量且对因变量贡献最大的进入回归方程。表 5 给出各模型决定系数的平方根 D，决定系数 D^2，调整后的 D^2 和标准估计误差。D^2 值越大反映自变量与因变量的共变量比率越高，模型与数据拟合度越好，从表中可看出模型 3 的 D^2 最大，定为最优模型。从模型 3 回归系数显著性结果（表 6）可得回归方程：$Y=345.303+0.192X_3-3.402X_4+23.257X_5$，回归方程显著性检验结果表明：回归平方和为 18 379.142，残差平方和为 2 901.789，对应的 F 统计量的值为 52.781，$p<0.05$，表明建立的回归方程有效。表 6 中各回归系数 t 检验也都在极显著水平，说明自变量 X_3、X_4、X_5 对因变量 Y 均有显著影响。各自变量对 Y 的标准系数，即直接通径系数 b_i 和 Excel 中计算结果一致。

表 5 模型拟合概述结果

模型	D	D^2	调整 D^2	标准估计的误差
1	0.878①	0.771	0.762	13.445 19
2	0.914②	0.835	0.822	11.618 75
3	0.929③	0.864	0.847	10.773 65

预测变量：①a（常量），X_4；②a（常量），X_4，X_5；③a（常量），X_4，X_5，X_3。

表 6 模型 3 的回归系数及显著性检验结果

	非标准化系数		标准误差(试用版)	t	p
	系数	标准系数			
a	345.303	51.985		6.642	0.000
X_4	−3.402	0.483	−0.666	−7.047	0.000
X_5	23.257	7.186	0.255	3.237	0.003
X_3	0.192	0.084	0.211	2.289	0.031

4 结论与讨论

4.1 Excel2007 和 SPSS17.0 都未提供专门计算通径系数的模块，但都可通过多元线性方程拟合获得通径系数。Excel2007 计算步骤较为复杂，但对于熟悉编程的用户，可利用 Excel2007 提供的 VBA 功能简化数据的处理步骤。SPSS17.0 计算通径系数步骤简单，结果详细，且可进行逐步回归，一次性得到最优回归方程，操作简单。

4.2 Excel2007 和 SPSS17.0 做通径分析时，都需先对数据进行正态性检验，数据服从正态分布才能继续进行统计分析。若通径分析结果中存在对因变量作用不显著的自变量，须逐一剔除，直至线性回归方程中各变量检验都为显著为止。

4.3 通过通径系数计算可得到各变量和因变量间的直接与间接关系，以及各自变量间的相互制约关系，决策系数的排序结果为最优路径选择提供准确依据。

4.4 在对通径分析结果进行解释时，若某一自变量和因变量相关系数基本等于它的直接作用，说明相关性反应二者真实关系，且通过此因变量进行直接选择效果较好。若相关系数为正，直接作用为负或值较小，说明间接效应是相关的主要原因，此时要对间接变量同时加以考虑。

参考文献：

[1] 任红松，朱家辉，杨斌，等．Excel 在通径分析中的应用［J］．农业网络信息，2006 (3)：90-92.

[2] 杜家菊，陈志伟．使用 SPSS 线性回归实现通径分析的方法［J］．生物学通报，2010，45 (2)：4-5.

[3] 崔党群．通径分析的矩阵算法［J］．生物数学学报，1994，9 (1)：73-74.