

相关系数与偏相关系数在相关分析中的应用^{*}

严丽坤

(云南财贸学院 统计信息学院, 云南 昆明 650221)

摘 要:在相关分析中,通常利用相关系数来分析或测定这些变量之间的线性相关程度,然而,简单相关系数受其他因素的影响,反映的往往是非本质的联系。要准确地反映两个经济变量之间的内在联系,需要计算偏相关系数。通过偏相关系数与相关系数的比较,来确定这两个变量之间的内在线性联系会更真实,更可靠。

关键词:相关分析;相关系数;偏相关系数

中图分类号:F222 **文献标识码:**A **文章编号:**1007-5585(2003)03-0078-03

Application of Correlation Coefficient and Biased Correlation Coefficient in Related Analysis

YAN Li - kun

(Statistics and Information School, Yunnan University of Finance and Economics, Kunming 650221, China)

Abstract: In the related analysis, the correlation coefficient is generally used to analyze and determine the linear's related degrees of variables. Because it is affected by other factors, what the simple correlation coefficient reflects is not the essence of relations. In order to reflect exactly the internal relations between the two economic variables, it needs calculating the biased correlation coefficient. By comparison with biased correlation coefficient and correlation coefficient, it will be more real and reliable to determine the internal linear relations between the two variables.

Key Words: Related Analysis; Correlation Coefficient; Biased Correlation Coefficient

相关分析是处理变量与变量之间关系的一种统计方法。近年来这种统计方法已广泛应用于经济学、医学、生物学、社会学等诸多领域,并取得了一定实效。从所处理的变量多少来看,如果研究的是两个变量间的关系称为简单相关;如果研究的是两个以上变量间的关系称为多元相关。从变量之间的关系形式上看,有线性相关分析及非线性相关分析。从统计思想和方法来看,线性相关是最基本的方法。

在相关分析中,通常利用两个变量之间的简单相关系数和一个变量与多个变量之间的复相关系数来分析或测定这些变量之间的线性相关程度,并据此进行线性回归分析、预测和控制等。相关系数 r 绝对值愈大(愈接近1),表明变量之间的线性相关程度愈高;相关系数绝对值愈小,表明变量之间的线性相关程度愈低。相关系数为零时,表明变量之间不存在线性相关关系。故此,人们通常利用相关系数的大小来解释变量间相互关系的大小。例如,以云南省人均收入与人均食品支出为资料计算最为简单的恩格尔函数所表明的在商品价格不变的条件下食品支出与收入水平的关系(如表1所示):

^{*} 收稿日期:2003-01-16

作者简介:严丽坤(1963—),女,浙江江山人,云南财贸学院讲师,主要从事统计理论与统计教学研究。

以食品支出为 Y 、人均收入为 X 作相关性检验,得 $r_{XY} = 0.96465$, 相应的概率值 $p = 0.0001 < 0.05 = \alpha$ 。这说明人均收入的增长与食品支出的增长程度呈高度正相关。为定量分析收入增长对食品支出的影响,对两者进行回归分析,得线性回归方程: $Y_T = 362.6418 + 0.29425X_T$

经检验人均收入与食品支出之间的线性关系极其显著。在不考虑其他因素的情况下,收入每提高一个单位,食品支出将增加 0.29425 元;362.6418 表示即使在收入为 0 的情况下,食品支出也需要 362.64 元。食品支出在总收入中平均所占比重为: $\frac{Y_T}{X_T} = 362.6418 / X_T + 0.29425$

上式中的 $\frac{Y_T}{X_T}$ 即所谓的恩格尔系数。显而易见,在此,恩格尔系数会随着 X_T 的增加而递减。这与德国统计学家恩格尔提出的反映食品支出与收入水平之间关系的恩格尔定律的结论是一致的。在这里,相关系数 $r = 0.96465$,既表明了收入增长与食品支出关系,又反映了收入增长与食品支出具有内在的线性联系的程度。

然而,简单相关系数受其他因素的影响,反映的往往是表面的非本质的联系。例如,一种商品的需求既受收入水平的影响又受其价格的影响。按照经济学理论,在一定收入水平下,该商品的价格越高,商品的需求量就越小。也就是说,需求与价格之间应当是负相关。可是,在现实经济生活中,由于收入和价格常常都有不断提高的趋势,如果不考虑收入对需求的影响,仅仅利用需求和价格的时间序列数据去计算简单相关系数,就有可能得出价格越高需求越大的错误结论。

高度相关并不意味着因果性。如果观察到样本相关系数具有某个大的正值或负值,就作结论说 x 的变化引起了 y 的变化,这是错误的。以云南省 GDP 和全省年末人口数资料(表 2)来说明:

众所周知,GDP 是衡量经济的重要指标,其数值的大小受众多因素的影响。如果 GDP 为 A 、年末人口数为 B 进行相关性分析,则 $r_{AB} = 0.9816$, 且 $p = 0.0001 < 0.05 = \alpha$ 。这意味着人口 B 与国内生产总值 A 之间存在高度正相关,亦即 B 越大, A 也越大。如果我们把这个高度正相关理解为这两个变量内在联系或因果关系,就会得出这样的结论:要想提高云南省的 GDP,只需大量增加人口。这无论从经济理论还是从日常生活哲理来说都是极其荒谬的。我们认为人口和 GDP 之间不应该有什么本质的内在联系。但是,是什么原因导致人口与国民收入具有如此高度的正相关关系呢?只要我们认真研究就会发现,人口与 GDP 都与另外一个变量——时间 T 有联系,它们都随着时间

表 1 云南省人均收入、人均食品支出

年份	人均收入(元)	人均食品支出(元)
1991	1703.16	763.84
1992	2061.74	861.6
1993	2639.07	1066.99
1994	3433.97	1441.93
1995	4064.93	1808.71
1996	4977.95	1971.54
1997	5558.29	2109.53
1998	6042.78	2222.58
1999	6178.68	2194.25
2000	6324.64	2091.7
2001	6797.71	2105.66

资料来源:《云南统计年鉴》(2002),中国统计出版社。

表 2 云南省 GDP、年末人口数

年份	GDP(亿元)	年末人口数(万人)
1991	517.41	3782.1
1992	618.69	3831.6
1993	779.21	3885.2
1994	973.97	3939.2
1995	1206.68	3989.6
1996	1491.62	4041.5
1997	1644.23	4094.0
1998	1793.9	4143.8
1999	1855.74	4192.4
2000	1955.09	4240.8
2001	2074.71	4287.4

资料来源:《云南统计年鉴》(2002),中国统计出版社。

的变化而变化,而且都是同方向变化的。我们计算人口与时间 T 的相关系数 $r_{BT} = 0.9865$, 国民生产总值 A 与时间 T 的相关系数 $r_{AT} = 0.988$ 。可以看出人口与 GDP 都与时间 T 高度正相关。正是由于人口和 GDP 都与时间高度正相关,所以才导致人口与国民收入之间高度正相关。这就告诉我们,两个经济变量之间的高度相关关系,有时并不是这两个经济变量本身的内在联系所决定的,它完全可能由另外一个变量的媒介作用而形成高度相关。所以,我们绝不能只根据相关系数很大,就认为两者经济变量之间有直接内在的线性联系。此时要准确地反映两个经济变量之间的内在联系,就不能简单的计算相关系数,而是需要考虑偏相关系数。

偏相关系数是在对其他变量的影响进行控制的条件下,衡量多个变量中某两个变量之间的线性相关程度的指标。所以,用偏相关系数来描述两个经济变量之间的内在线性联系会更合理、更可靠。偏相关系数不同于简单相关系数。在计算偏相关系数时,需要掌握多个变量的数据,一方面考虑多个变量之间可能产生的影响,另一方面又采用一定的方法控制其他变量,专门考察两个特定变量的净相关关系。在多变量相关的场合,由于变量之间存在错综复杂的关系,因此偏相关系数与简单相关系数在数值上可能相差很大,有时甚至符号都可能相反。

偏相关系数的取值与简单相关系数一样,相关系数绝对值愈大(愈接近 1),表明变量之间的线性相关程度愈高;相关系数绝对值愈小,表明变量之间的线性相关程度愈低。例如,我们以上讨论的人口与 GDP 的关系,如果令时间 T 不变,计算人口与 GDP 的偏相关系数为 $r_{AB.T} = 0.2743$, 且 $P = 0.4431 > 0.05 = \alpha$ 。由此可以看出,如果去掉时间的因素,人口与 GDP 之间的偏相关系数很小,即如果固定时间不变,人口与 GDP 之间的关系是微弱的,且经检验人口与 GDP 之间没有什么内在的必然的线性联系。人口越多未必 GDP 越高,这是比较合乎实际的。

由以上的实例可以看出,在相关分析中,我们切不可只根据相关系数很大,就认为两个经济变量之间有内在的线性联系或因果关系。因为相关系数只表明两个变量的共变联系,尽管这种共变联系有时也体现了两个变量的内在联系(如物价与需求量),但在很多情况下,这种共变联系是由某个或某些变量的变化所引起的。所以,我们在研究经济变量之间的相关关系时,当由样本计算的两个变量的相关系数很大时,我们要认真检查一下这种相关是否与经济理论和经济意义相符合;如果不符,一定是由于其他变量的变化所引起的。这时,我们就需要研究和探索引起这两个变量高度相关的变量;去掉这些变量变化的影响因素,计算偏相关系数,最后确定这两个变量之间的内在线性联系。当我们研究多个经济变量时,有时计算其中两个变量的相关系数与经济理论和经济意义相符。但由于其他变量影响的作用,这个相关系数可能扩大或缩小了这两个变量之间的真实联系,这时,通过偏相关系数与相关系数的比较,来确定这两个变量之间的内在线性联系会更真实,更可靠。所以,在相关分析中,我们除了使用相关系数以外,还应该使用偏相关系数,这是非常重要,也是十分必要的。

参考文献:

- [1] 袁卫,庞皓,曾五一. 统计学[M]. 北京:高等教育出版社,2000.
- [2] 云南省统计局. 云南统计年鉴[Z]. 北京:中国统计出版社,2002.

责任编辑、校对:李品秀