为什么回归系数的估计值会有“错误”的符号？

王玉祥　　方　莉

当我们应用多元回归分析时，有时会遇到这样的问题：从实际数据算出的某些回归系数的估计值的符号与从问题的专业知识或直观经验所得的结论相反。例如，在一具体问题中，根据问题的专业知识和经验，有足够的理由认为回归自变量 X_1 的回归系数 β_1 是正的，可是实际算出的 β_1 的最小二乘估计 β_1 却是负的，这时称 β_1 具有“错误”的符号。其实应用回归分析处理实际问题的不少人遇到过这种情况。

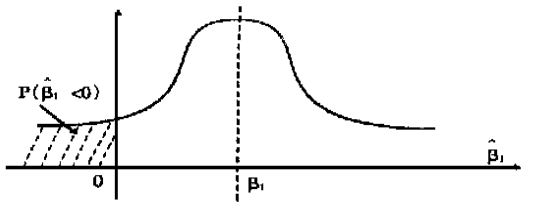
一般说来，下列情况都可能导致某些回归系数的估计具有“错误”符号：(1) 某些自变量取值范围太窄；(2) 模型中丢弃了若干重要自变量；(3) 设计阵含复共线性(Multicollinearity)。下面逐一讨论这三种情况。

一、某些自变量取值范围太窄

用一元回归 $T=\beta_0+\beta_1 X+e$ 来说明这个问题，设由组观测数据 (x_i, y_i)，$i=1,2,\cdots,n$ 算出的回归系数 β_1 的最小二乘估计为 $\beta 1$，则它的方差为

$$V_{ar}(\beta_1)=\frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

其中 $\bar{x}=\frac{1}{n}\sum_{i=1}^{N}x_i$ 从上式可知，β_1 的方差不仅与度量模型误差大小的方差 σ^2 成正比，而且与自变量 x 的取值波动大小 $\sum_{i=1}^{n}(x_i-\bar{x})^2$ 成反比。当 x 从一个较小范围取值时，$\sum_{i=1}^{n}(x_i-\bar{x})^2$ 就很小，从而 β_1 的方差很大。当 $V_{ar}(\beta_1)$ 过大时，即使 β_1 是正值，也存在 β_1 的估计 β_1 以较大概率取负值的可能



性。这一点可以从图 1 看出来，图 1 是 β_1 的概率密度曲线，在误差 e 服从正态分布 $N(0,\sigma^2)$ 条件下，这条曲线是

$N(\beta_1,\sigma^2/\sum_{i=1}^{n}(x_i-\bar{x})^2)$ 图中阴影部分的面积，即为 β_1 取负值的概率，虽然 $\beta_1>0$，但当 $V_{ar}(\beta_1)=\sigma^2/\sum_{i=1}^{n}(x_i-\bar{x})^2$ 很大时，阴影部分的面积可以很大，这时 β_1 以较大概率取负值。

如果通过分析后确认，回归系数估计值取“错误”符号确实是由于某些自变量取值范围太窄，那么修正的办法也就不言自明了。

二、模型中错误地丢弃了若干重要自变量

在回归模型中错误地丢弃若干重要自变量的现象是常见的。为了说明这种情况，先看一个两个自变量的回归例子。

x_{i1}	2	4	5	6	8	10	11	13
x_{i2}	1	2	2	4	4	4	6	6
y_i	1	5	3	8	5	3	10	7

应用最小二乘法，y 对 x_1，x_2 的经验回归方程为 $y=1.063-1.222x_1+3.649x_2$ 如果剔除自变量 x_2，那么 y 对 x_1 的经验回归方程为 $y=1.835+0.463x_1$。可见丢弃了自变量 x_2 之后，使得 x_1 的符号由负变正。

如果某些回归系数估计值的“错误”符号确实来自这种情况，那就要回过头来仔细考虑回归模型的自变量选择问题。目前国内应用较多的选择自变量的方法是逐步回归。这个方法简单易行，且有方便的计算机程序。

三、设计阵含复共线性

为了说明什么是复共线性，需要引进一般多元线性回归模型。假设经过自变量选择，选择了与因变量 Y 有密切关系的 P-1 个自变量 $x_1, x_2 \cdots\cdots, x_{p-1}$。对 $Y, X_1, \cdots, X_{p-1}$ 作了 n 次观测，得到 n 组数据 $y_1, x_{i1} \cdots, x_{ip-1}$，它们满足

$$y_i=\beta_0+\beta_1 x_{i1}+\cdots\beta_{p-1}x_{ip-1}+e_i \quad (i=1,\cdots n)$$

若记

$$Y=\begin{pmatrix} y_1 \\ \cdots \\ y_n \end{pmatrix}, X=\begin{pmatrix} 1 & x_1 & \cdots & x_1 p-1 \\ \vdots & \vdots & & \vdots \\ 1 & x_n 1 & \cdots & x_n p-1 \end{pmatrix} \triangleq (x_0, x_1, \cdots x_{p-1})$$

$$\beta=\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{P-1} \end{pmatrix}, e=\begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$$

则得到线性回归模型 $Y=X\beta+e$，通常称 X 为该回归模型的设计阵。如果 X 的列向量 $x_0, x_1, \cdots, x_{p-1}$ 之间有近似线性关系

$$a_0x_0+a_1x_1+\cdots+a_{p-1}x_{p-1}\approx 0$$

则称设计阵 X 含复共线关系或复共线性。

记$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$为 X'/X 的特征根。当 X 含复共线性时,X'/X 的特征根至少有一个很接近于零。即至少$\lambda_P \approx 0$。另一方面 β 的最小二乘估计 β 的协方差阵为 $Cov(\beta) = \sigma^2 (X'X)^{-1}$(注意,这里假定了 $E(e) = 0$,$Cov(e) = \sigma^2 I$)所以,常数项β_0和回归系数β_i的最小二乘估计的方差之和 $\sum_{i=0}^{p-1} V_{ar}(\beta_i) = \sigma^2 tr(X'X)^{-1} = \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i}$ 这里 tr(A) 表示方阵 A 的对角线元素之和。如果设计阵 X 含复共线性,有一些$\lambda_i \approx 0$,于是$\sum_{i=0}^{p-1} V_{ar}(\beta_i)$就很大,与(一)款同样的理由,可导致一些$\beta_i$与对应的$\beta_i$的符号相反。在应用上有多种方法判定设计阵 X 是否含有复共线性,可采用方差扩大因子法、条件数法、特征分析法等。

如果设计阵 X 确实含有复共线性,此时不仅回归系数估计值可能有"错误"符号,而且回归系数最小二乘估计的性质总体上说显著不好。为了克服这个缺陷,本世纪 60 年代以来,统计学家提出了许多种有偏估计,以期改进最小二乘估计。目前付诸实用且最有影响的是岭估计 $\beta(R) = (X'X + kI_p)^{-1} X'y$ 这里 k 是待定常数,称为岭参数。对不同的 k,上式确定了不同的岭估计。在岭估计中,确定 k 的原则之一是,使所有回归系数估计值$\beta_i(R)$都具有较合理的符号。

总之,在应用上如果发现某些回归系数的估计值取"错误"符号,此时应该回过头来,仔细分析原始数据,检查初始模型,寻找导致"错误"符号的真正原因,然后根据不同情况,采取适当纠正方法,或扩大某些自变量取值范围,或重新考虑自变量选择,或采用某种有偏估计。

参考文献

陈希孺,王松桂,回寻分析 —原理、方法及应用,安徽教育出版社,1987

作者单位:晋中教育学院
阳泉市卫生防疫站

(上接第 41 页)严格考核与兑现奖惩;四是完善计量,检验工作,健全收发领退计量、检验手续,防止浪费、丢失、短少;五是加强会计档案管理,使记帐、算帐、报帐等工作符合国家统一会计制度的要求,逐步实现会计基础工作制度化、规范化、科学化,为会计工作的正常进行提供基础性的保证。

3. 加强企业内部管理制度建设,建立完善的会计内部控制制度。会计内控制度应对那些对会计记录和会计报表的可靠性有直接影响的会计处理程序作出明确的规定,使各项经济业务的会计处理工作有章可循,职责分明,为会计信息的真实性提供制度保障。

4. 提高会计人员的业务素质和职业道德,充分发挥会计职能,向信息使用者提供真实的会计信息。笔者认为小型企业的财会负责人必须取得会计师资格,大、中型企业的财会负责人必须取得高级会计师的资格。只有素质提高了,才能防止会计信息的人为失真。要建立健全会计人员职业道德规范,鼓励会计人员模范地执行法规制度,遵守职业道德,学习专业知识,敢于、善于动真碰硬,揭发会计信息失真的现象。敢于抵制各种违法乱纪行为,把好"关口"。

5. 建立会计信息质量监督体系。要继续深化会计改革,健全宏观经济监督机制,充分发挥好政府监督、社会监督的作用,建立健全对会计报表的管理和会计信息质量的检查制度,要继续帮助和指导企业建立健全内部财务管理制度,监督其有效运行。

6. 实行会计负责人委派制,各地应建立一支高素质的会计队伍统一管理。对辖区内单位的财会负责人实行委派,使会计负责人的管理、工资等与该单位脱勾,并定期实行轮换。这样做的最大好处就是免除了会计负责人的后顾之忧,可以大胆坚持原则,不听命于单位领导,从而有利于从源头上,从根本上防止会计信息失真。

此外,我们也应努力减少会计核算中的人为因素,增加会计核算的刚性,使会计信息更加准确、可靠。

(作者单位:山西省煤炭规划设计院)

(上接第 43 页)1. 它们都是总指数。

2. 当平均数指数作为综合指数的变形形式时,这两种指数都能进行因素分析。

3. 它们都是由综合指数发展而来的,都是综合指数的派生指数。

笔者认为,该文作者之所以得出文中那些结论,除了对平均数影响因素的理解有失偏颇外,重要的是忽略了统计学与数学的本质区别。社会经济统计学是从数量方面描述社会经济现象总体的规模、水平、状态,分析事物之间的相互关系,揭示事物发展变化的原因,探索事物发展变化的规律性的一门学科。因此在设计指标及指标体系,对指标进行因素分析分解,建立指数和指数体系时,就不能只从数学公式的推演是否成立考虑,而应当从事物的内在联系出发,从分析问题、解决问题的需要出发来进行。只有这样,才能正确理解和运用统计学中的各种指标和分析方法,对社会经济现象进行分析,得出正确的结论。

作者单位:吕梁地区会计学校
山西杏花村汾酒厂