尝试用主成分方法辅助 OLS 回归分析*

李鸿丞 1600011374

2018 年 5 月 21 日

摘要

常用 OLS 回归讨论某个因变量和自变量之间的关系，但当自变量本身对因变量解释力不足时，系数在统计检验上会不显著。为了给出自变量对因变量影响的可靠结论，常用做法是引入新的自变量，以剔除因变量中与原来自变量之间不相关的冗杂部分。本文的目的即是用主成分方法来从数学上对这些“剔除变量”的选择引入一个系统的方法。

关键词：OLS 回归 主成分方法

1 方法介绍

1.1 问题引入

若给定感兴趣的自变量 Y 和两个自变量 X_1、X_2，可以采用 OLS 回归的方法来探究其之间的关系，最常见的问题是：Y 关于 X_i 的回归系数是否显著？X_1 与 X_2 回归系数之间的差别是否显著？OLS 回归在数学上对 Y 做关于 $\{X_i\}$ 的线性回归，实质上是将因变量样本$\boldsymbol{Y}$ 投影到自变量样本张成的线性空间 $U_X = span(\boldsymbol{X_1}, \boldsymbol{X_2})$ 上。然而当$\boldsymbol{Y}$ 在 U_X 上的投影占$\boldsymbol{Y}$ 的比重较小时[1]，X_i 的回归系数估计量 $\hat{\beta}_i$ 的方差会较大，统计检验给出的结论自然也不好。

解决这一问题的常用做法之一是寻找一系列新自变量 $\{Z_j\}$，并将其加入到回归之中去。为了达到较好的效果，我们应该选择与 Y 相关度较高但与每个 X_i 相关度较低的变量组，这一点可以简单地从 $\hat{\beta}_i$ 方差的表达式中看出：

$$Var(\hat{\beta}_i) = \frac{\hat{\sigma^2}}{TSS_X(1 - R_i^2)}$$

其中 $\hat{\sigma^2}$ 是回归对误差项方差的估计，当 $\{Z_j\}$ 与 Y 的相关度高时，这一项被降低且回归系数的检验更显著。然而另一方面，R_i^2 表示 X_i 对其他变量做回归得到的 R^2 值，故如果 $\{Z_j\}$ 与 $\{X_i\}$ 也有很大相关性，R_i^2 增大使得回归系数方差

*总字数：2970

[1]也即 $\{X_i\}$ 对 Y 的解释力不足

增大且检验变得不显著。因此在选择合适的剔除变量时总会面临权衡：新变量组样本$\boldsymbol{Z_j}$ 既要与因变量样本$\boldsymbol{Y}$ 高度相关，又要与 $\{\boldsymbol{X_i}\}$ 保持一定的非相关性。本文只做数学上的努力，故假设已经通过经济学分析找到了许多潜在的剔除变量，记为 $\{V_k\}_N$。但如何从这些变量中提取出最有效的部分是经济学直觉无法解决的。

具体方法还是从空间投影出发，也即 Y、$\{X_i\}_2$ 与 $\{V_k\}_N$ 的相关度等价于将前两者投影到后者张成的线性空间的结果。于是，首先应当通过主成分方法寻找 $\{V_k\}_N$ 对应的主成分元，然后完成投影，并根据投影出来的系数大小选择这一组元中适当的几项构成最终的剔除变量组 $\{Z_j\}$。最后会证明用主成分元选择出的变量是相对最优的。

1.2 方法推导

现有因变量样本$\boldsymbol{Y}$、主自变量样本组 $\{\boldsymbol{X_i}\}_p$ 和潜在剔除变量组 $\{\boldsymbol{V_k}\}_N$，假设后两者分别构成的矩阵列满秩[2]。此时计算出 $\{V_k\}_N$ 各变量间的样本协方差矩阵[3]：

$$\Sigma_V = Cov(\boldsymbol{V}, \boldsymbol{V})$$

由满秩假设与矩阵对称可知此矩阵一定可对角化，也即 $\exists B \in O(N)$ 使得：

$$\Sigma_V = BAB^T = B \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_N \end{bmatrix} B^T \tag{1}$$

据此可作变换：$\boldsymbol{U} = B^T\boldsymbol{V}$ 以及 $Z_k = \frac{U_k}{\sqrt{\lambda_k}}$，从而便得到了归一化后的主成分变量向量$\boldsymbol{Z}$。其具有类似规范正交的形式，这是因为：

$$Cov(\boldsymbol{U}, \boldsymbol{U}) = Cov(B^T\boldsymbol{V}, B^T\boldsymbol{V}) = B^T Cov(\boldsymbol{V}, \boldsymbol{V})B = B^T(BAB^T)B = A$$

$$\Rightarrow Cov(U_i, U_j) = \delta_{ij}\lambda_i \Rightarrow Cov(Z_i, Z_j) = \delta_{ij}$$

而后对$\boldsymbol{Y}$ 和 $\{\boldsymbol{X_i}\}_p$ 分别向 $\{\boldsymbol{Z_k}\}_N$ 上作投影，其投影出来的各个分量即是做 OLS 回归的系数估计量：由于 Z_i 之间互不相关，故对任何变量 Y' 的线性模型 $Y' = \sum_{k=1}^{N} \beta_k^{Y'} Z_k + u'$ 的最小二乘估计量 $\hat{\beta}_k^{Y'} = \frac{Cov(Y', Z_k)}{Cov(Z_k, Z_k)} = Cov(Y', Z_k)$ 即为一协方差值，这个表达式的意义是直观而正确的。

然后用 Y 与 X_i 分别替换 Y' 计算这些投影系数。为了得到与 Y 相关性尽量大的变量，则按照对应的 $|\beta_k^Y|$ 从大到小排列 Z_k，越靠前的变量具有越高的

[2]也即样本量 $n > N, p$，且不存在完全共线性

[3]本文所写向量，在没有角标时代表这一类变量，而有角标时则代表某一变量的 n 个观测值构成的向量。例如$\boldsymbol{V}$ 代表所有潜在剔除变量，而$\boldsymbol{V_k}$ 则代表第 k 个变量的样本向量

权重；另外，为了得到与 X_i 相关性尽量小的变量，则按照 $|\beta_k^{X_i}|$ 从小到大排列 Z_k，越靠前亦更重要。为了直观地进行挑选，不妨对每个主自变量设置一个相对权重 w_i，并记 Y' 对应的 Z_k 的排名为 $r_{k,Y'}$，由此计算每个 Z_k 的“加权排名”：

$$R_k = r_{k,Y} + w_i \sum_{i=1}^{p} r_{k,X_i} \tag{2}$$

显然，越小的 R_k 所对应的变量越满足条件，故而适当取靠前的几项 Z_k 即可组成较为有效的剔除变量组。

1.3 相对最优性的证明

相对最优性其实是指，在给定样本以及潜在剔除变量组的情况下，当用主成分方法选定了 J 个上述构造的 Z_j 变量，并记此变量选择[4]为 C_0，则其他任何变量选择 C' 都不可能做到这一点：在对 $\{X_i\}_p$ 的解释力[5]不超过 C_0 的情况下，又对 Y 的解释力比 C_0 强。

首先可令约束条件“取等号”，也即取 C' 使与 C_0 对 $\{X_i\}_p$ 的解释力按照 w_i 加权达到相等。这时候再考察两者对 Y 的解释力。命 Ω_1 为 C_0 所包含的所有 Z_j 组成的集合，Ω_2 为除开 C_0 后的其他 Z_k 组成的集合，Ω_3 为可用以充分且完全地、线性地表示 C' 中所有变量的 Z_k 组成的集合[6]。则假若 $\Omega_3 \not\subset \Omega_1$，那么为了弥补对 $\{X_i\}_p$ 的解释力，必须要在 Ω_2 中选择变量；然而由于在对 $\{X_i\}_p$ 的解释力按照 w_i 加权后相同的条件下，任何 Ω_1 中形成的变量选择均比 Ω_2 中形成的变量选择对 Y 的解释力更强——这是加权排名的结果。因此为了得到更大的对于 Y 的解释力，Ω_3 必须尽量排除 Ω_2 中的元素，也即当达到最大解释力时，应当有 $\Omega_3 \subset \Omega_1$。而且已经控制了 C' 与 C_0 对 $\{X_i\}_p$ 有同等解释力这一条件，故而最终推得：使得拥有对 Y 最大解释力的 C'，必定满足 $\Omega_3 = \Omega_1$，也即解释力不可能超过 C_0。证明结束。

2 应用实例

现以 1950 到 1995 年的年度数据为例 [1]，令 $Y =$“GDP 年增长比”，$X =$“房地产市场产出年增长比”，欲研究后者对前者在所考察的年份中的影响趋势。若简单地对 Y 做关于 X 的 OLS 回归，则发现 X 的回归系数对应的 t 检验 p 值高达 0.811，其 95% 置信区间为 [-0.22,0.28] 无法判断正负，可见此时方差极大回归系数的值提供不了太多信息。

首先要得到一系列潜在剔除变量。很容易想到用 GDP 的其他成分来分解 Y，于是引入 9 个新变量：$V_k, k = 1, 2, ..., 9 =$“第一产业年增长比”、“工业年增

[4]将一组变量称作“变量选择”，如果其中的每个变量都可以写为 $\{V_k\}_N$ 的某个线性组合

[5]本文所谓“解释力”，是指因变量样本向自变量样本张成空间的投影占原来样本的比例

[6]由“变量选择”的定义可知这是可以做到的

长比”、“工业建设年增长比”、“交通与通信业年增长比”、“商业年增长比”、“社会服务支出年增长比”、“银行与保险业年增长比”、“科教文体等福利支出年增长比”和“政府代理与社会组织支出年增长比”[7]。首先调用 MATLAB 软件[8]中有关矩阵运算的语句，直接求出 $V_1 \to V_9$ 组的主成分元 $\boldsymbol{U} = \{U_k\} = B^T\boldsymbol{V}$ 以及对应的本征值向量 $\{\lambda_k\}$。接着再求得 $Z_k = \frac{U_k}{\sqrt{\lambda_k}}$，即可得到规范正交的主成分元了，其关于原变量组$\boldsymbol{V}$ 的线性变换关系被表示如下：

	Z_1	Z_2	Z_3	Z_4
V_1	-0.0506	0.1357	-0.1311	0.0051
V_2	0.1973	0.0237	-0.0319	-0.0085
V_3	-0.0452	-0.0177	0.0067	0.0340
V_4	-0.2215	0.0206	0.0161	-0.0425
V_5	-0.0268	-0.0587	-0.0010	0.0680
V_6	0.0332	0.0777	0.0518	0.0246
V_7	0.0122	-0.0004	-0.0075	0.0099
V_8	0.0105	-0.1444	-0.0723	-0.0656
V_9	0.0479	0.0698	0.0548	-0.0917

Z_5	Z_6	Z_7	Z_8	Z_9
0.0202	-0.0008	0.0099	0.0033	0.0001
-0.0538	-0.0252	-0.0150	0.0074	0.0133
-0.0130	0.0597	0.0291	0.0171	0.0158
-0.0363	-0.0285	-0.0212	0.0010	0.0115
0.0275	-0.0639	0.0300	0.0103	0.0057
0.0624	0.0040	-0.0453	0.0008	0.0125
0.0009	0.0018	0.0268	-0.0367	0.0133
0.0414	0.0058	-0.0143	0.0024	0.0077
0.0257	-0.0126	0.0458	0.0075	0.0028

得到线性变换矩阵后用 STATA 程序[9]完成生成 $\{Z_k\}$ 的步骤，并对 Y 和 X 分别关于 $\{Z_k\}$ 做 OLS 回归，得到的系数估计量绝对值以及相应的排序如下：

表 1: Y 变量系数排序

排名	1	2	3	4	5	6	7	8	9
变量	Z_9	Z_8	Z_3	Z_6	Z_2	Z_5	Z_7	Z_1	Z_4
系数绝对值	7.4	2.2	1.7	1.6	1.1	0.7	0.4	0.3	0.27

[7]所有原始数据都被保存在 Data.xls 文件中

[8]见 Data.m 文件，结果被保存在 Z.xls 文件中

[9]见 Data.do 文件，其中三组排序的结果被保存在 rank.xls 文件中

表 2: X 变量系数排序

排名	1	2	3	4	5	6	7	8	9
变量	Z_3	Z_7	Z_9	Z_4	Z_6	Z_8	Z_5	Z_1	Z_2
系数绝对值	0.3	0.4	0.6	0.9	1.0	1.2	1.7	2.0	2.8

若取 X 的权重为 $w = 1$，则可以列出加权排名（当两者加权相等时，应按照 X 系数排序小者在前的原则）：

表 3: 加权排名结果

变量	Z_3	Z_9	Z_8	Z_7	Z_6	Z_4	Z_5	Z_2	Z_1

至此可以开始按照上述排序进行尝试了，最终发现当取剔除变量为 Z_3、Z_9 和 Z_8 时已经达到了相当好的效果！其在 STATA 中运行 OLS 回归的结果为：

	R^2	联合显著性 F	RMSE
	0.9041	0.0000	2.7351
变量	回归系数	系数显著性 t	置信区间
X	0.0904	0.030	[0.0093,0.1715]
Z_3	-1.7163	0.000	
Z_9	7.4646	0.000	
Z_8	2.3319	0.000	

可见回归结果中各变量系数显著性均很高，X 系数的 95% 置信区间也显示其值为一致大于零的。由此可以得出比之前更确切的结论：房地产市场产出年增长比对 GDP 年增长比于 1952-1950 年间有正向的影响，其影响系数为大约 9% 量级。

3 方法反思和结论

这一方法有其有效性，已反映在了上述例子中。然而也有局限性：首先，这一方法没有提供寻找新变量的经济学思路，当 $\{V_k\}$ 本身解释力不够好时无论怎么选择变量都不容易得到好结果；其次，这一方法的最优性是建立在先验地设定 w_i 权重的基础上的，如果对权重设置不正确，也很可能导致回归失调的问题；再者，这一方法并没有给出加入剔除变量后 $\{X_i\}$ 的解释力受到“侵蚀”程度的度量，当选择不当时很可能使得某个 X_i 的影响被扭曲。

究其实质，这是一套建立在 OLS 回归框架下的显著性优化方法，为一些有重要意义但是在统计上暂时不显著的相关关系的探究提供了一条解决途径。虽然本文阐述的方法本身有缺点，但我相信这一方法在实际工作中应当能发挥作用。

参考文献

[1] Hsueh,T.& Q.Li. China's National Income, 1952-1995[M]. Boulder:Westview Press, 1999