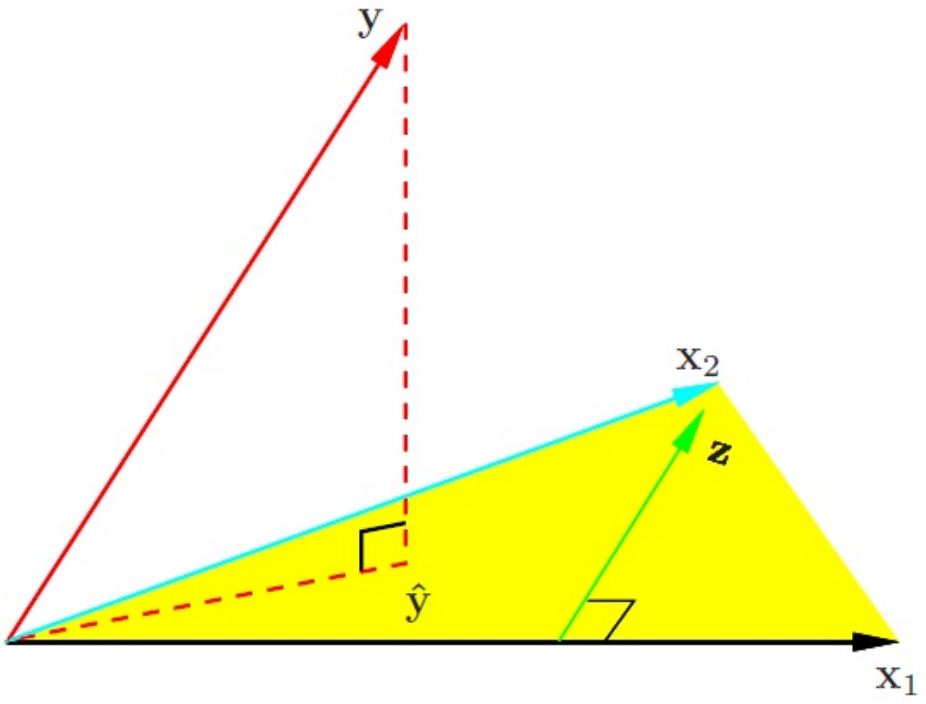
知乎 首发于 川流不息



为什么要进行因子正交化处理?

石川
量化交易 话题的优秀回答者

已关注

黑猫Q形态等 119 人赞同了该文章

摘要

选股多因子模型中常进行因子正交化处理。如果因子之间不满足正交性，则它们会相互影响各自的回归系数，这可能造成回归系数过大的估计误差，对因子的评价产生负面影响。

1 多因子模型求解

在选股多因子模型中，人们常提到的一个概念是**因子正交化处理**。本文就从多因子截面回归求解的角度来简单说说为什么我们喜欢相互正交的因子，以及如果因子之间不正交对回归系数会有什么影响。

一个多因子模型可以写成如下的形式：

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \boldsymbol{\varepsilon}$$

其中 $\mathbf{y}$ 是 $N \times 1$ 阶股票下一期的收益率向量， $\mathbf{X}$ 为 $N \times K$ 阶当期的因子暴露矩阵， $\mathbf{b}$ 为 $K \times 1$ 阶待通过回归求解得到的因子收益率向量， $\boldsymbol{\varepsilon}$ 为 $N \times 1$ 阶残差向量。假设 $\mathbf{X}$ 满足列满秩，则上述模型的 OLS（ordinary least squares）回归求解为：

需要注意的是，在上面这个模型以及 $\mathbf{b}$ 的表达式中，**因子向量 $\mathbf{X}$ 已经包括了所有的 regressors，因此回归模型右侧没有额外的截距项。**这意味着，如果我们假设截距项也是一个因子，则它对应的 $N \times 1$ 阶向量 $[1,1,\ldots,1]^T$ 已经作为 $\mathbf{X}$ 的某一列（通常是第一列）存在于 $\mathbf{X}$ 之中了；如果我们假设截距项不是一个因子，则 $\mathbf{X}$ 中没有 $[1,1,\ldots,1]^T$ 这一列。

在 Barra 的多因子模型 CNE5 中考虑了国家因子，所有个股在该因子上的暴露都是 1，因此它的作用就相当于一个截距因子； $[1,1,\ldots,1]^T$ 这个向量在 Barra 模型中正是 $\mathbf{X}$ 的第一列。另外，对于我们最熟悉的 simple regression model，它的右侧只有一个截距和一个解释变量：

$$y_i = a + bx_i + \varepsilon_i$$

按照上述说明，该模型对应的矩阵 $\mathbf{X}$ 包括两列：一列对应截距，一列对应真正的解释变量 $\mathbf{x}$ ：

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix}$$

从 $\mathbf{b}$ 的表达式来看，它和 $\mathbf{X}^T\mathbf{X}$ 有关。**当 $\mathbf{X}$ 的各列（即回归模型中的不同解释变量，或我们研究问题中的不同因子暴露向量）之间不正交时，则在计算 $\mathbf{X}^T\mathbf{X}$ 乃至最终的 $\mathbf{b}$ 时， $\mathbf{X}$ 不同列之间是相互影响的，而这种影响不是什么好事儿。**

2 简单一元回归

让我们从最简单的一元回归（simple univariate regression）说起。

假设有一元回归模型 $\mathbf{y} = b\mathbf{x} + \boldsymbol{\varepsilon}$ （模型右侧只有一个解释变量，**没有截距项**）。对于两个同阶向量 $\mathbf{m}$ 和 $\mathbf{n}$ ，令 $\langle \mathbf{m}, \mathbf{n} \rangle$ 表示它们的内积，即 $\langle \mathbf{m}, \mathbf{n} \rangle = \sum m_i n_i$ ，则该一元回归模型的 OLS 解为（求解对象就是标量 b ）：

$$b = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\langle \mathbf{x}, \mathbf{x} \rangle}$$

这个结论非常简单，但是它十分重要。在上一节中，我们给出了多元回归 OLS 求解的表达式：

$$\mathbf{b} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

比较一元回归模型的标量 b 和多元回归模型的向量 $\mathbf{b}$ 不难发现如下现象：在多元回归模型中如果所有的解释变量两两正交，即 $\langle \mathbf{x}_i, \mathbf{x}_j \rangle = 0, i \neq j$ ，则向量 $\mathbf{b}$ 中的每一个系数 b_i 恰恰等于：

这是因为 $\langle \mathbf{x}_i, \mathbf{x}_j \rangle = 0$ 保证了 $\mathbf{X}^T\mathbf{X}$ 的所有非对角元素都是 0，因此它是一个对角阵。对角阵的逆矩阵就是把该对角阵对角线上的元素都取倒数，所以逆矩阵仍然是对角阵。因此，$\mathbf{X}^T\mathbf{X}$ 的第 i 个对角元素为 $1/\langle \mathbf{x}_i, \mathbf{x}_i \rangle$ 。另一方面，$\mathbf{X}^T\mathbf{y}$ 是一个 $K \times 1$ 向量，它的第 i 个元素是 $\mathbf{x}_i$ $\mathbf{y}$ 和 y 的内积，即 $\langle \mathbf{x}_i, \mathbf{y} \rangle$ 。最终，多元回归的 b_i 正是 $\langle \mathbf{x}_i, \mathbf{y} \rangle / \langle \mathbf{x}_i, \mathbf{x}_i \rangle$ 。

怎么样？ b_i 和一元回归中的 b 的表达式一模一样，说明**当所有解释变量相互正交时，不同的因子（即 $\mathbf{x}_i$ ）对彼此的参数估计（即 b_i ，因子收益率）没有任何影响。**这便是正交的好处。

那么，当因子（解释变量）之间不正交时又会怎样呢？为了回答这个问题，我们首先来看看回归的几何意义。

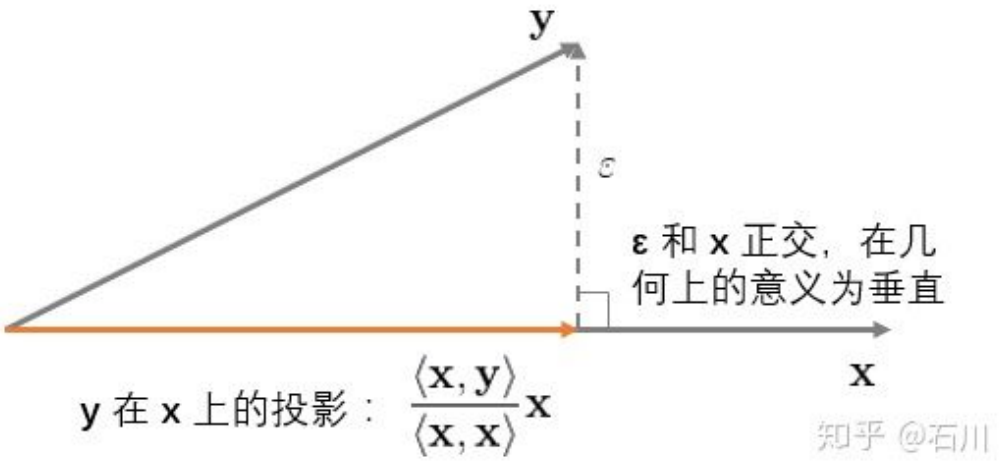
3 回归的几何意义

将 $\mathbf{b}$ 的表达式代入回归模型得到 $\boldsymbol{\varepsilon}$ 的表达式，并计算 $\mathbf{X}$ 和 $\boldsymbol{\varepsilon}$ 的内积有；

$$
\begin{aligned}
\mathbf{X}^T\varepsilon &= \mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{b}) \\
&= \mathbf{X}^T\left(\mathbf{y} - \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}\right) \\
&= \mathbf{X}^T\mathbf{y} - \left(\mathbf{X}^T\mathbf{X}\right)\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \\
&= \mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{y} \\
&= \mathbf{0}
\end{aligned}
$$

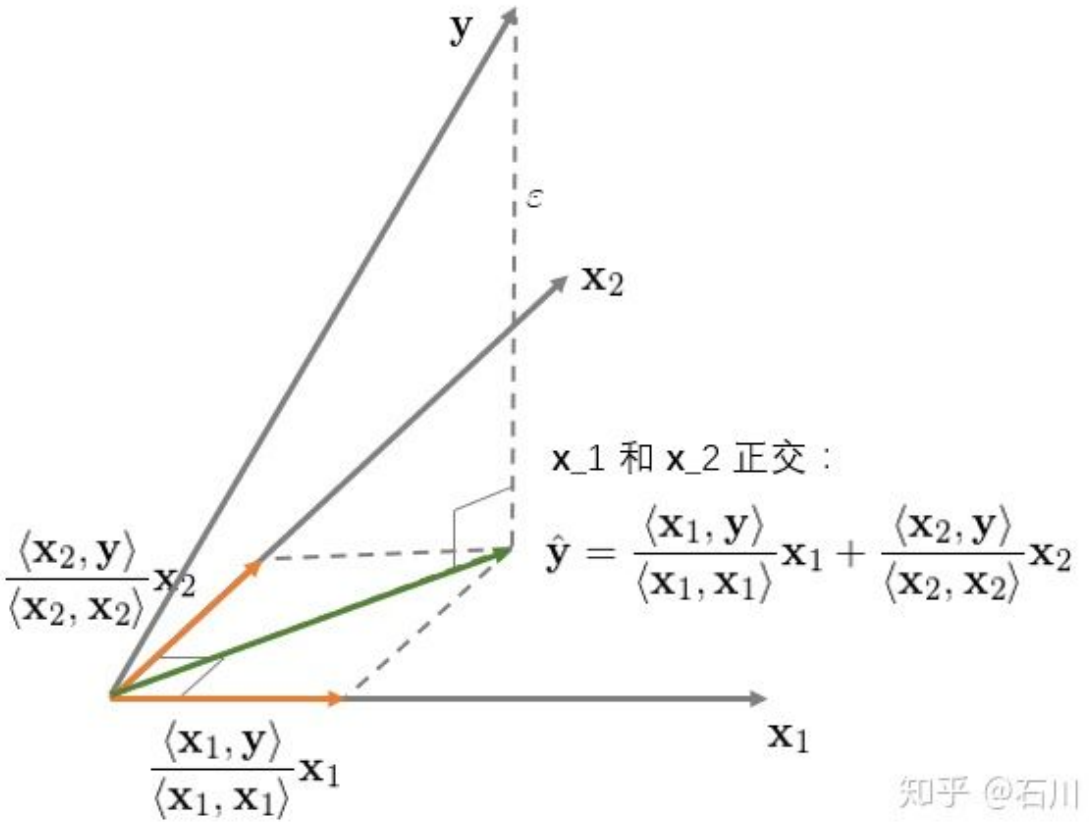
上式说明，OLS 的残差 $\boldsymbol{\varepsilon}$ 和解释变量 $\mathbf{X}$ 正交。来看看这在几何上意味着什么。

首先考虑最简单的情况，即一元回归 $\mathbf{y} = b\mathbf{x} + \boldsymbol{\varepsilon}$ （再次提醒，没有截距项）。它的几何意义如下图所示：



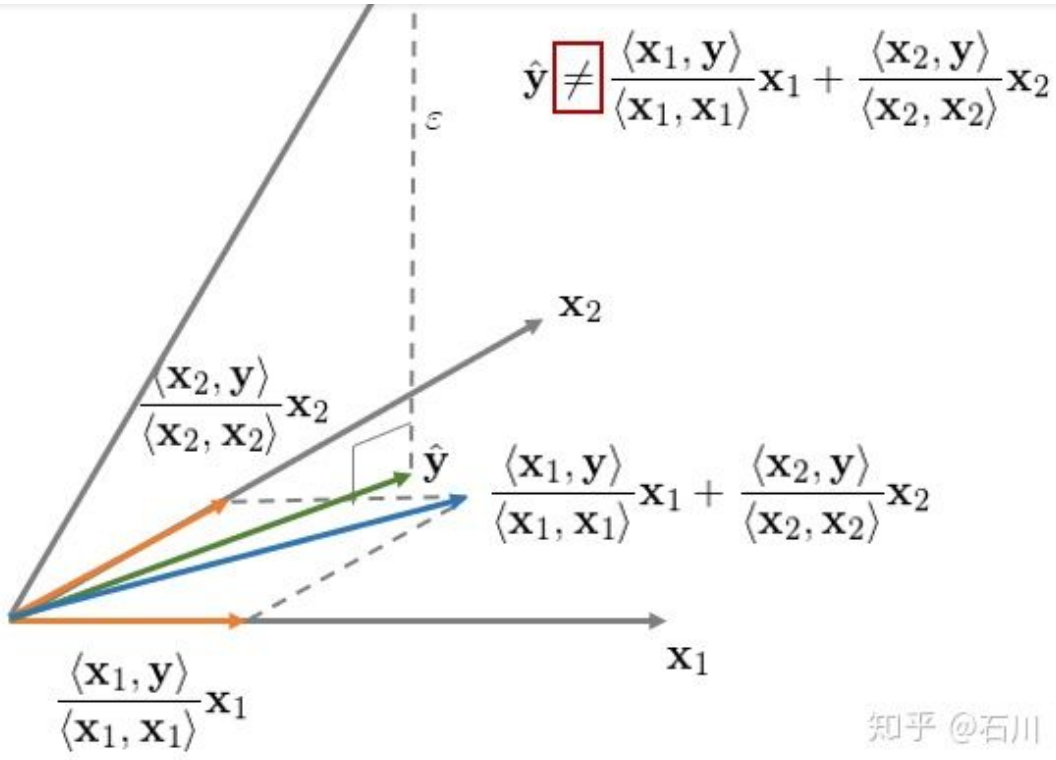
何意义。

再来看看二元回归 $\mathbf{y} = b_1\mathbf{x}_1 + b_2\mathbf{x}_2 + \boldsymbol{\varepsilon}$ ，并首先**假设 $\mathbf{x}_1$ 和 $\mathbf{x}_2$ 之间是正交的。**该回归的几何意义如下：



对于二元回归，它的几何意义是将 $\mathbf{y}$ 垂直投影到由 $\mathbf{x}_1$ 和 $\mathbf{x}_2$ 生成的超平面内，其投影正如上图中绿色向量所示。此外，我们可以分别、独立的将 $\mathbf{y}$ 投影到 $\mathbf{x}_1$ 和 $\mathbf{x}_2$ 上（图中两个橘黄色向量）。在本例中，**由于 $\mathbf{x}_1$ 和 $\mathbf{x}_2$ 相互正交（垂直），因此绿色向量恰好等于两个橘黄色向量之和。**这说明当 $\mathbf{x}_1$ 和 $\mathbf{x}_2$ 正交时，回归系数 b_i 仅由 $\mathbf{x}_1$ 和 $\mathbf{y}$ 决定、其他任何解释变量 $\mathbf{x}_j (j \neq i)$ 对 b_i 均没有影响。

下面来看看 $\mathbf{x}_1$ 和 $\mathbf{x}_2$ **非正交**的情况。该二元回归的几何意义如下：



它和前一种情况最大的区别是，**当 $\mathbf{x}_1$ 和 $\mathbf{x}_2$ 非正交时， $\mathbf{y}$ 在由 $\mathbf{x}_1$ 和 $\mathbf{x}_2$ 生成的超平面内的投影不等于 $\mathbf{y}$ 分别在 $\mathbf{x}_1$ 和 $\mathbf{x}_2$ 上的投影之和。**在这种情况下，解释变量之间对各自的回归系数有不同的作用，因此 OLS 的回归系数 b_i **不再等于** $\langle \mathbf{x}_i, \mathbf{y} \rangle / \langle \mathbf{x}_i, \mathbf{x}_i \rangle$ 。

非正交 $\mathbf{x}_i$ 之间的相互作用如何影响回归系数 b_i 呢？通过连续正交化来求解多元线性回归可以回答这个问题。

4 用正交化过程求解多元回归

还是拿我们最熟悉的 simple regression model 为例；该模型有两个解释变量 —— 截距项和 $\mathbf{x}$ 。

$$y_i = a + bx_i + \varepsilon_i$$

令 $\mathbf{x}_0$ 表示截距项对应的解释变量，即 $\mathbf{x}_0 = [1, 1, \ldots, 1]^T$ ； $\mathbf{x}_1$ 表示上式中的解释变量 $\mathbf{x}$ 。假设 $\mathbf{x}_0$ 和 $\mathbf{x}_1$ 非正交（正交的话我们就不用费劲了）。对于简单回归模型，回归系数 a （对应 $\mathbf{x}_0$ ）和 b （对应 $\mathbf{x}_1$ ）的解为：

$$\begin{aligned} b &= \frac{\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2} \\ a &= \frac{1}{n}\sum y_i - b\left(\frac{1}{n}\sum x_i\right) \end{aligned}$$

是和 $\mathbf{z}_0$ 互相垂直（正交）的向量，记为 $\mathbf{z}_1$ 。由一元回归的性质可知：

$$\mathbf{z}_1 = \mathbf{x}_1 - \frac{\langle \mathbf{z}_0, \mathbf{x}_1 \rangle}{\langle \mathbf{z}_0, \mathbf{z}_0 \rangle} \mathbf{z}_0 = \mathbf{x}_1 - \bar{x}\mathbf{1}$$

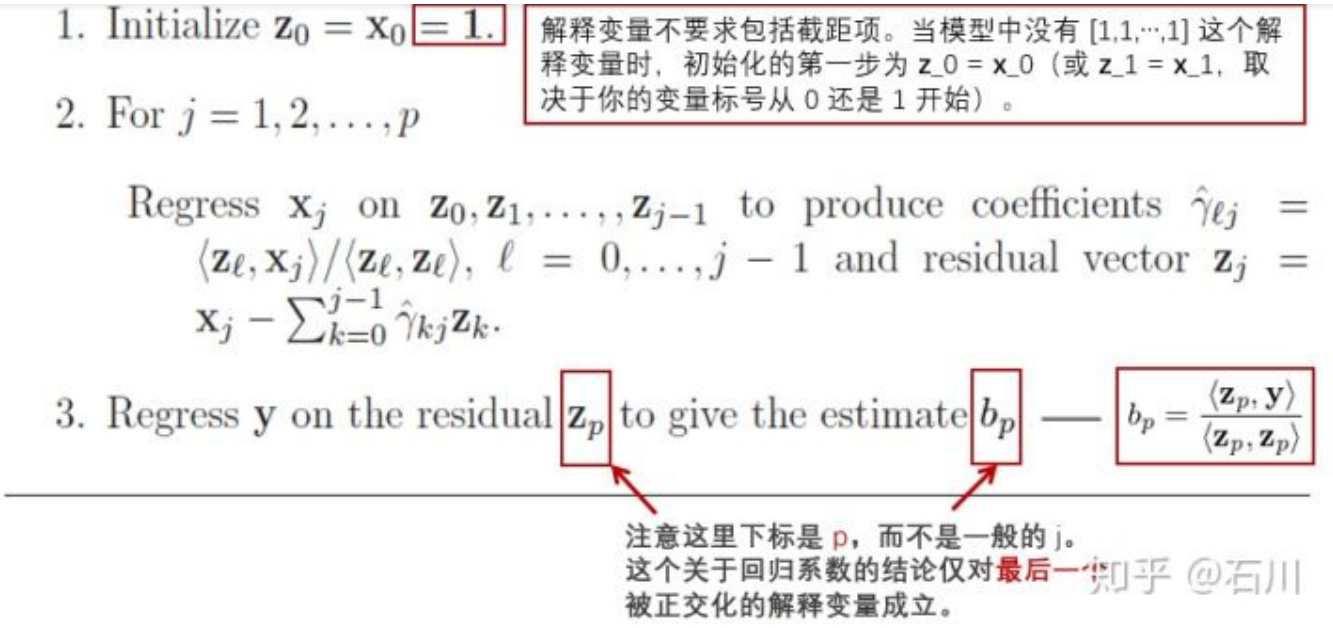
其中 $\bar{x}$ 表示 $\mathbf{x}$ 的均值， $\mathbf{1}$ 表示列向量 $[1, 1, \ldots, 1]^T$ ，即 $\mathbf{z}_0$ 。

So far so good？接下来，注意了：

将 $\mathbf{y}$ 用上面得到的 $\mathbf{z}_1$ 进行一元回归（不带截距），得到的回归系数就是上述 simple regression model 中解释变量 $\mathbf{x}$ 的回归系数 b ！

$$\begin{aligned} b &= \frac{\langle \mathbf{z}_1, \mathbf{y} \rangle}{\langle \mathbf{z}_1, \mathbf{z}_1 \rangle} = \frac{\langle \mathbf{x}_1 - \bar{x}\mathbf{1}, \mathbf{y} \rangle}{\langle \mathbf{x}_1 - \bar{x}\mathbf{1}, \mathbf{x}_1 - \bar{x}\mathbf{1} \rangle} \\ &= \frac{\sum x_i y_i - \bar{x} \sum y_i}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i}{\sum (x_i^2 - 2x_i \bar{x} + \bar{x}^2)} \\ &= \frac{\sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i}{\sum x_i^2 - 2(\sum x_i)\bar{x} + n\bar{x}^2} = \frac{\sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i}{\sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2} \\ &= \frac{\sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i}{\sum x_i^2 - n\bar{x}^2} = \frac{\sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2} \end{aligned}$$

怎么样？我们并没有直接对该模型求解，而是通过正交化的方式就求出了解释变量 $\mathbf{x}_1$ 的回归系数 b 。反应快的小伙伴也许马上会问 a 呢？ a 是否等于 $\langle \mathbf{z}_0, \mathbf{y} \rangle / \langle \mathbf{z}_0, \mathbf{z}_0 \rangle$ 呢？别急，我们一会儿就聊 a ，但是在那之前先来看一个**通过连续正交化求解多元回归的算法**（Hastie et al. 2016）：



该算法的核心是通过连续的正交化计算把一组非两两正交的向量 $\mathbf{x}_i$ 转换成一组两两正交的向量 $\mathbf{z}_i$ ，并以此方便的求出最后一个被正交化的解释变量的多元回归系数。虽然它只有三步，但是每一步都值得解读一下：

1. 第一步是初始化，在所有解释变量中（如果回归中有截距项，就把 $[1, 1, \ldots, 1]^T$ 看做一个解释变量）任意挑选一个当作 $\mathbf{x}_0$ 进行初始化 $\mathbf{z}_0 = \mathbf{x}_0$ 。

2. 第二步是**根据我们自己选定的递归顺序（任意顺序都可以）**，对 $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p$ 依次进行正交化。例如，对 $\mathbf{x}_j$ 的正交化处理就是用它和之前已经被处理过后的正交向量 $\mathbf{z}_0, \mathbf{z}_1, \cdots, \mathbf{z}_{j-1}$ **逐一独立**一元回归得到系数 $\langle \mathbf{z}_k, \mathbf{x}_j \rangle / \langle \mathbf{z}_k, \mathbf{z}_k \rangle, k = 0, 1, \cdots, j-1$ ，进而用 $\mathbf{x}_j$ 减去 $(\langle \mathbf{z}_k, \mathbf{x}_j \rangle / \langle \mathbf{z}_k, \mathbf{z}_k \rangle)\mathbf{z}_k, k = 0, 1, \cdots, j-1$ 之和，得到的残差就是最新的正交化后的向量 $\mathbf{z}_j$ 。

3. 使用 $\mathbf{y}$ 和 $\mathbf{z}_p$ 进行一元回归，得到的系数 $\langle \mathbf{z}_p, \mathbf{y} \rangle / \langle \mathbf{z}_p, \mathbf{z}_p \rangle$ 正是这个多元回归 OLS 求解中原始解释变量 $\mathbf{x}_p$ 的回归系数 b_p 。注意，这一结论仅对**最后一个**（第 p 个）被正交化后的解释变量成立。换句话说，对于别的解释变量 $j < p$ ， $\langle \mathbf{z}_j, \mathbf{y} \rangle / \langle \mathbf{z}_j, \mathbf{z}_j \rangle$ 并不是多元回归中原解释变量 $\mathbf{x}_j$ 的回归系数。

看到这里，有的小伙伴可能会问，这个算法确实不错，但是费了半天劲算出了一大堆相互正交的向量 $\mathbf{z}_j$ ，但是求解回归系数的结论仅对最后一个被正交化的解释变量成立，这不是坑爹吗？

答案是并不坑爹！这是因为上述算法中的关键一点是，**正交化这些解释变量的顺序是任意的。**我们可以选任何一个来初始化，也可以选任何一个作为最后一个被正交化的解释变量。无论我们怎么选，上述过程都保证了最后一个被正交化的解释变量的回归系数满足 $b_p = \langle \mathbf{z}_p, \mathbf{y} \rangle / \langle \mathbf{z}_p, \mathbf{z}_p$ 。因此，**我们只需要依次挑选这些解释变量作为最后一个被正交化的，就可以通过上述步骤方便的求出它们的回归系数。**而它所反映出来的本质是：

仍能够对 $\mathbf{y}$ 产生的增量贡献。

这个算法叫作多元回归的 Gram-Schmidt（格拉姆-施密特）正交化过程。

本小节开始的 simple regression model 已经验证了上述结论。我们使用 $\mathbf{x}_0$ 将 $\mathbf{x}_1$ 正交化处理得到 $\mathbf{z}_1$ ，然后用 $\mathbf{y}$ 和 $\mathbf{z}_1$ 回归得到的正是 $\mathbf{x}_1$ 的回归系数 b ；如果将 $\mathbf{x}_1$ 选为 $\mathbf{z}_0$ ，然后用它正交化 $\mathbf{x}_0 = [1, 1, \dots, 1]^T$ ，就可以方便的求出回归系数 a 。

让我们来好好审视一下这个结论，即：

$$b_p = \frac{\langle \mathbf{z}_p, \mathbf{y} \rangle}{\langle \mathbf{z}_p, \mathbf{z}_p \rangle}$$

上式说明，解释变量 $\mathbf{x}_p$ 的回归系数 b_p 和正交化后的 $\mathbf{z}_p$ 的大小（ $\mathbf{z}_p$ 自己的内积为分母）有关。**如果 $\mathbf{x}_p$ 和其他解释变量高度相关（即非常不正交），那么 $\mathbf{z}_p$ 就会很小，则会导致 b_p 非常不稳定（一点点样本数据的变化都会导致 b_p 的大幅变化）。**当 y_i 满足独立同分布时，假设它的方差为 σ^2 ，可以证明回归系数 b_p 的方差和 $\mathbf{z}_p$ 的大小成反比，即 $\|\mathbf{z}_p\|^2$ 越小， b_p 的误差越大：

$$\mathrm{var}(b_p) = \frac{\sigma^2}{\langle \mathbf{z}_p, \mathbf{z}_p \rangle} = \frac{\sigma^2}{\|\mathbf{z}_p\|^2}$$

在多因子模型中， b_p 代表的是因子 p 的收益率。**为避免因子收益率的估计非常不稳定，要求不同的因子之间尽量满足正交化。**举例来说，在 Barra 的 CNE5 模型中，非线性规模因子和规模因子之间进行了正交化处理；残差波动率因子和规模以及 BETA 因子也进行了正交化处理。

在结束本小节的讨论之前，我还想介绍一个有意思也有用的特性。本节的论述说明我们可以任选一个解释变量作为最后一个，然后根据连续正交化方便的求出它的回归系数。这意味着如果我们有20个解释变量，需要进行 20 次上述操作。那么，**是否存在什么办法仅通过进行一次连续正交化就求出所有的回归系数 $b_j, j = 0, 1, \cdots, p$ 呢？答案是肯定的。**

假设我们按照某给定顺序 $\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_p$ 进行了连续正交化过程，得到了 $\mathbf{z}_0, \mathbf{z}_1, \cdots, \mathbf{z}_p$ ，且我们现在知道 $b_p = \langle \mathbf{z}_p, \mathbf{y} \rangle / \langle \mathbf{z}_p, \mathbf{z}_p \rangle$ 。由于 b_p 正是解释变量 $\mathbf{x}_p$ 的回归系数，因此 $b_p \mathbf{x}_p$ 正是 $\mathbf{x}_p$ 所解释的 $\mathbf{y}$ 的部分。如果从 $\mathbf{y}$ 中剔除 $b_p \mathbf{x}_p$ ，并把得到的 $\mathbf{y} - b_p \mathbf{x}_p$ 用 $\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_{p-1}$ 回归，则结果就和 $\mathbf{x}_p$ 无关了。在这个新的回归中， $\mathbf{x}_{p-1}$ 就变成了最后一个被正交化的解释变量，其对应的正交向量为 $\mathbf{z}_{p-1}$ 。因此， $\mathbf{x}_{p-1}$ 的回归系数就是用新的 $\mathbf{y} - b_p \mathbf{x}_p$ 和 $\mathbf{z}_{p-1}$ 回归的结果：

以此类推，我们可以按照 $b_p, b_{p-1}, \cdots, b_0$ 的**倒序**求解出多元回归中所有解释变量的回归系数 b_j （Drygas 2011）：

$$\begin{aligned} b_p &= \frac{\langle \mathbf{z}_p, \mathbf{y} \rangle}{\langle \mathbf{z}_p, \mathbf{z}_p \rangle} \\ b_j &= \frac{\langle \mathbf{z}_j, \mathbf{y} - \sum_{i=j+1}^{p} b_i \mathbf{x}_i \rangle}{\langle \mathbf{z}_j, \mathbf{z}_j \rangle}, \ j = p-1, p-2, \cdots, 0 \end{aligned}$$

最后用本小节开始的 simple regression model 检验一下。我们用上述方法求解截距项的回归系数 a 看看。根据定义有 $\mathbf{z}_0 = \mathbf{1}$ 并假设已知 b 。则根据上面的表达式可得：

$$\begin{aligned} a &= \frac{\langle \mathbf{1}, \mathbf{y} - b\mathbf{x} \rangle}{\langle \mathbf{1}, \mathbf{1} \rangle} \\ &= \frac{\sum y_i - b \sum x_i}{n} \\ &= \frac{1}{n} \sum y_i - b \left(\frac{1}{n} \sum x_i \right) \end{aligned}$$

这正是直接求解 simple regression model 得到的回归系数 a （请往前滚屏比较看看）。

5 一个例子

本节用一个例子来验证一下上一节的各种公式。假设有四个解释变量 $\mathbf{x}_0$ 到 $\mathbf{x}_3$ ，以及 $\mathbf{y}$ ：

$$\mathbf{x}_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \mathbf{x}_1 = \begin{bmatrix} 1 \\ 2 \\ 4 \\ 3 \\ 7 \\ 6 \\ 3 \\ 5 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} 0 \\ 1 \\ 2 \\ 8 \\ 9 \\ 4 \\ 6 \\ 2 \end{bmatrix}, \mathbf{x}_3 = \begin{bmatrix} 1 \\ 3 \\ 3 \\ 2 \\ 5 \\ 8 \\ 0 \\ 4 \end{bmatrix}, \mathbf{y} = \begin{bmatrix} 3 \\ 1 \\ 4 \\ 1 \\ 5 \\ 9 \\ 2 \\ 6 \end{bmatrix}$$

直接使用回归系数 $\mathbf{b}$ 的表达式求解，则它们的回归系数分别为： $b_0 = 0.38548073$，$b_1 = 0.96332683$，$b_2 = -0.36300685$，$b_3 = 0.37189391$。按照 $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$

$$
\begin{aligned}
\mathbf{z}_0 &= \mathbf{x}_0 \\
\mathbf{z}_1 &= \mathbf{x}_1 - \frac{\langle \mathbf{x}_1, \mathbf{z}_0 \rangle}{\langle \mathbf{z}_0, \mathbf{z}_0 \rangle} \mathbf{z}_0 \\
\mathbf{z}_2 &= \mathbf{x}_2 - \frac{\langle \mathbf{x}_2, \mathbf{z}_0 \rangle}{\langle \mathbf{z}_0, \mathbf{z}_0 \rangle} \mathbf{z}_0 - \frac{\langle \mathbf{x}_2, \mathbf{z}_1 \rangle}{\langle \mathbf{z}_1, \mathbf{z}_1 \rangle} \mathbf{z}_1 \\
\mathbf{z}_3 &= \mathbf{x}_3 - \frac{\langle \mathbf{x}_3, \mathbf{z}_0 \rangle}{\langle \mathbf{z}_0, \mathbf{z}_0 \rangle} \mathbf{z}_0 - \frac{\langle \mathbf{x}_3, \mathbf{z}_1 \rangle}{\langle \mathbf{z}_1, \mathbf{z}_1 \rangle} \mathbf{z}_1 - \frac{\langle \mathbf{x}_3, \mathbf{z}_2 \rangle}{\langle \mathbf{z}_2, \mathbf{z}_2 \rangle} \mathbf{z}_2
\end{aligned}
$$

$$
\mathbf{z}_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \mathbf{z}_1 = \begin{bmatrix} -2.875 \\ -1.875 \\ 0.125 \\ -0.875 \\ 3.125 \\ 2.125 \\ -0.875 \\ 1.125 \end{bmatrix}, \mathbf{z}_2 = \begin{bmatrix} -1.51082251 \\ -1.37662338 \\ -2.10822511 \\ 4.75757576 \\ 2.29437229 \\ -1.83982684 \\ 2.75757576 \\ -2.97402597 \end{bmatrix}, \mathbf{z}_3 = \begin{bmatrix} -0.00844984 \\ 1.08972192 \\ -1.02611768 \\ 1.06099247 \\ -0.48286987 \\ 2.17022584 \\ -1.56337379 \\ -1.24012905 \end{bmatrix}
$$

使用 Drygas (2011) 提出的解法按照 b_3, b_2, b_1, b_0 的顺序求解各个回归系数 b_j ：

$$
\begin{aligned}
b_3 &= \frac{\langle \mathbf{z}_3, \mathbf{y} \rangle}{\langle \mathbf{z}_3, \mathbf{z}_3 \rangle} \\
b_2 &= \frac{\langle \mathbf{z}_2, \mathbf{y} - b_3 \mathbf{x}_3 \rangle}{\langle \mathbf{z}_2, \mathbf{z}_2 \rangle} \\
b_1 &= \frac{\langle \mathbf{z}_1, \mathbf{y} - b_3 \mathbf{x}_3 - b_2 \mathbf{x}_2 \rangle}{\langle \mathbf{z}_1, \mathbf{z}_1 \rangle} \\
b_0 &= \frac{\langle \mathbf{z}_0, \mathbf{y} - b_3 \mathbf{x}_3 - b_2 \mathbf{x}_2 - b_1 \mathbf{x}_1 \rangle}{\langle \mathbf{z}_0, \mathbf{z}_0 \rangle}
\end{aligned}
$$

上述公式求出 $b_0 = 0.38548073$，$b_1 = 0.96332683$，$b_2 = -0.36300685$，$b_3 = 0.37189391$，和使用回归系数 $\mathbf{b}$ 的表达式求解的结果完全一致。另外，我们也可以分别选择 $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2$ 替换 $\mathbf{x}_3$ 作为最后一个被正交化的解释变量（前三个变量的顺序也不重要），并利用 $b_p = \langle \mathbf{z}_p, \mathbf{y} \rangle / \langle \mathbf{z}_p, \mathbf{z}_p \rangle$ 求解，得出的 b_j 也和上面的完全相同。

6 正交等于不相关?

从定义出发，两个因子向量 $\mathbf{x}_1$ 和 $\mathbf{x}_2$ 正交意味着它们的内积，即 $\langle \mathbf{x}_1, \mathbf{x}_2 \rangle$ 为零。而 $\mathbf{x}_1$ 和 $\mathbf{x}_2$ 的相关系数为零则意味着 $\langle \mathbf{x}_1 - \mathbf{E}[x_1] \cdot \mathbf{1}, \mathbf{x}_2 - \mathbf{E}[x_2] \cdot \mathbf{1} \rangle$ 为零，因为在计算相关系数时，必须先分别减去其均值，这就是个 centering 的过程。由于 $\langle \mathbf{x}_1, \mathbf{x}_2 \rangle$ 为零不一定意味着 $\langle \mathbf{x}_1 - \mathbf{E}[x_1] \cdot \mathbf{1}, \mathbf{x}_2 - \mathbf{E}[x_2] \cdot \mathbf{1} \rangle$ 也为零，因此正交不一定等于不相关。

举个例子，$[4, 2]^T$ 和 $[3, -6]^T$ 的内积为零，这两个向量正交。而各自减去均值后，$[4, 2]^T$ 和 $[3, -6]^T$ 分别变为 $[1, -1]^T$ 和 $[4.5, -4.5]^T$ 。这两个新向量在一条直线上、内积不为零，因此 $[4, 2]^T$ 和 $[3, -6]^T$ 的相关系数不为零（事实上，它们的相关系数等于 1）。从多元回归求解的角度来说，我们在乎的是他们是否正交，而非 centering 之后的内积是否为零（即是否不相关）。

不过对于因子暴露向量来说，因为个股在每个因子上的暴露都经过 demean 处理了，所以每个因子向量的均值已经是零了（这里考虑的就是简单等权均值的情况，而不是像 Barra 那种用市值作为权重进行去均值的情况）。从这个意义上说，因子向量之间正交和它们之间不相关等价。

7 结语

一不留神又写了这么长。从阅读体验来说，实在抱歉。

本文掰扯了一大堆公式其实就是想说明下面这句话：

在多元线性回归中，解释变量 $\mathbf{x}_j$ 的回归系数 b_j 等于 $\mathbf{x}_j$ 在被其他 $\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, \cdots, \mathbf{x}_p$ 调整之后（即正交化，从而排除其他 $\mathbf{x}_i$ 对 $\mathbf{x}_j$ 的影响）仍能够对 $\mathbf{y}$ 产生的增量贡献。如果 $\mathbf{x}_j$ 和其他解释变量高度相关，则它的回归系数 b_j 会有很大的估计误差。这对于多因子模型中评价因子收益非常不利。

如果看完之后你对这句话有一定的体会，那我的功夫就没白花。

在计算机算法进行多元回归求解的时候，并不是试图按照 $\mathbf{b}$ 的公式计算 $\mathbf{X}^T\mathbf{X}$ 的逆矩阵，而采用的正是正交化的思路。在正交化的过程中可以非常容易的得到 $\mathbf{X}$ 的 QR 分解，其中 $\mathbf{Q}$ 是正交阵、 $\mathbf{R}$ 是上三角阵。这也极大的化简了回归系数 $\mathbf{b}$ 以及 $\mathbf{y}$ 预测值的求解。由于篇幅原因（我也好意思说篇幅……），本文就不给出 QR 分解的具体表达式了，感兴趣的读者请参考 Hastie et al. (2016)。

参考文献

- Drygas, H. (2011). On the Relationship between the Method of Least Squares and

- Hastie, T., R. Tibshirani, and J. Friedman (2016). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd Ed. Springer.

编辑于 2019-07-03

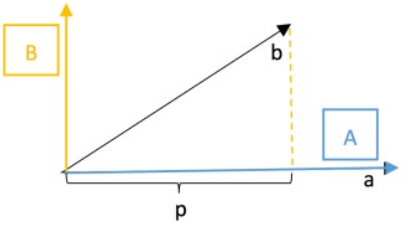
多因子模型 BARRA模型 计量经济学

文章被以下专栏收录

推荐阅读

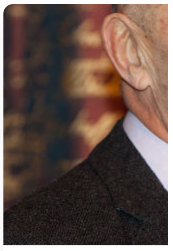
996笑话数则（跟风）

KirishimaHiiragi



正交矩阵和 Gram-Schmidt 正交化[MIT线代第十七课]

忆臻 发表于机器学习算...



论文笔记 I Systemat

张天涯

知乎 首发于 川流不息

Syous 1 年前

其实这段放在计量经济学框架里也是共通的。追求残差和自变量的正交化过程，相当于保证系数b的unbias估计结果；后面的正交化过程，也就是利用一个IV来做类似于2SLS过程。伍德里奇的《Econometric analysis of cross section and panel data》也讲的很详细。

3

石川 (作者) 回复 Syous 1 年前

谢谢！需要学习的真的太多了。

1

Syous 回复 石川 (作者) 1 年前

哪里哪里.....我是跟着大佬你在学习的，我贴上面这段是为了其他不同背景的读者方便参考阅读的，毕竟殊途同归.....可不是给你看的。

1

会飞的瓶盖 1 年前

厉害厉害。大佬更了。

赞

会飞的瓶盖 1 年前

最大的价值就在最后一段，前面算实证。

就是新增因子在排除存量因子共同（重叠）作用后，剩下部分的预测能力。如果有就是有效的。。。无论正负。

赞

大亏货 1 年前

说句实在的为什么需要正交化，是因为现有的已公开的模型太落后，只能处理正交化的自由度

赞

哈德逊河畔的鹅 1 年前

理论上来说需要正交化，但是这样处理之后怎么处理因子失去经济意义的问题？另外您觉得需要做正态变换的处理吗？

赞

石川 (作者) 回复 哈德逊河畔的鹅 1 年前

感谢留言。先简单回答第一个问题吧。可以参考 Barra 的做法：它们把大类内的小因子叫做 descriptor，大类因子叫 factor。每个 factor 类内的 descriptors 采取普通的线性

是高度相关的，只是用不同的代理指标描述某个风格。

我之前写了一篇这个zhuanlan.zhihu.com/p/38...，里面提到了同时使用多个代理指标（descriptor）来组合成大类因子是否有数据挖掘的成分呢？在这方面，AQR 的观点是使用多个指标并不是一种因子激增（因此没有数据挖掘问题），而是提高因子健壮性的一种方法，因为无论哪个单一指标都无法完美的代表我们的目标因子。这种处理类似于机器学习中的集合学习算法，它和随机森林以及 AdaBoost 算法比单一的决策树算法分类效果更好有异曲同工之妙。

最后，如果某个大类因子和其他大类因子非常相关，那即便有失去经济意义的问题，我认为也有必要正交化。

1

哈德逊河畔的鹅 回复 石川 (作者) 1 年前

谢谢您！这种普通线性加权应该就跟Asness构造quality minus junk类似。另外听AQR最近刚出的podcast他们prefer little negative correlation。不过有一点我一直比较疑惑，如果按照BARRA的方法来做，在每个时点都进行截面回归，最后预测收益率和协方差矩阵，那么不论是对因子的变换或者正交化，需要在每个截面上保持一致吗？那如果因子在时序上表现不同怎么办？例如a和b可能现在比较相关需要正交，但是以前可能不那么相关就不需要。

赞

陈泽宇 11 个月前

您好，最近在看有关知识，看了你的文章后，获益良多。 另外我有一个私人问题，如果您有时间的话，请您能给我指点一下。我现在在法国一所工程师学校读硕士，本科在国内某电读计算机，学校明年有个项目是和慕尼黑大学合作培养金融数学mathmatical financial硕士，很是吸引我。索性去年成绩也不错，应该可以被选上。但是我对金融方面两眼黑，虽然网上都说这个金融数学理论较多，但是我之前还是计算机方面学的多一点。 本来打算学个什么机器学习，大数据。 所以，依照您对金融数学的了解，这个专业的难度和前景，以及国内形式怎么样，能分享一二吗？

非常感谢您对一个迷茫的学生的指点。

赞

彭程 8 个月前

barra里的因子好像都没正交过，这是为什么呢？

赞

石川 (作者) 回复 彭程 8 个月前

barra 正交化了。比如在 CNE5 里面用 size 和 beta 对 residual vol 进行了正交化处理。可以再仔细看看文档。:)

王——

8 个月前

石老师，我最近也在研究因子正交化的内容，发现一些问题。我们考查因子相关性是从pearson相关系数角度考察的，正交化后的因子的pearson相关系数确实为0。可是对于排序打分的做法而言，相关性以秩相关系数考察更合理。正交化后秩相关系数可能显著不为0，这个问题怎么解决您有什么建议吗？

赞

追风 8 个月前

正交的时候用Lowdin对称正交时因子旋转的最少，保留的经济意义最多

赞

卢宣 5 个月前

作者每一篇文章的正文或者评论都会引出另一篇写的文章，看的停不下来[捂脸]再次感谢！

赞 回复 踩 举报

马铭岑 1 个月前

石川请问多因子模型时间序列上，可以使用斯密特正交吗？如果不能使用，时间序列如何解决因子之间多重共线性？希望有哪位大神能回答，万分感激[拜托]

赞

搏击长空 1 个月前

关于QR分解：jianshu.com/p/9da4bc2d5...

赞