

因子模型及相关计量经济学概述

- 因子模型及相关计量经济学概述
 - 1. 从CAPM出发
 - 套利定价理论 Arbitrage Pricing Theory
 - 2. 如何检验?
 - 时间序列回归
 - 两步法横截面回归(two-pass)
 - Fama-MacBeth 回归
 - 3. 方法差异（用什么作为因子收益率？）
 - 排序法
 - 单排
 - 双排
 - 4. 截面多因子模型 vs. 时序多因子模型（用什么作为因子暴露？）
 - 5. 检验异象
 - 排序法检验
 - 时序检验
 - FM回归检验
 - 6. 总结
 - 参考文献

1. 从CAPM出发

$$\mathbb{E}[R_i] - R_f = \beta_i (\mathbb{E}[R_M] - R_f)$$

$$\mathbb{E}[R_i - R_f] = \beta_i (\mathbb{E}[R_M - R_f])$$

$$\mathbb{E}[R_i^e] = \beta_i (\mathbb{E}[R_M^e])$$

- 没有截距
- expectation是什么意思？从实证的角度看，是时间还是横截面？
 - β_i 的下标*i*是什么意思？
 - *i*是研究的兴趣点所在。也就是，不同的 β_i 是否对应了不一样的收益？（横截面选股）
- 直觉上，CAPM说了件什么事情？
 - 对于证券*i*（或证券组合），超额收益率的期望值，等于， β_i 和 另一个东西的超额收益率的期望值的乘积
 - 这“另一个东西”，是因子(factor)。在CAPM里，“因子超额收益”是市场超额收益
 - β_i ，是证券*i*在这个因子上的“因子暴露”(factor exposure, factor loading)

套利定价理论 Arbitrage Pricing Theory

$$\mathbb{E}[R_i^e] = \beta_i' \lambda$$

- K 个因子, λ 是 K 维向量
- APT的出发点是统计上的factor decomposition + 一价定律

- 因此，除了“一价定律”以外，APT实际上没有讨论“经济含义”
- λ 是因子的预期超额收益率向量。也常被称作 factor risk premium。这个收益率应当是正的，因为对于风险，需要给出报酬。
 - 以下 λ 都作为“因子的预期超额收益率向量”，如果不是收益率向量，写作 f 。
 - 有时候因子的收益率是不能简单看到的，因此，因子本身的 factor risk premium 的估计，是一个重要的问题
- 实际数据是否符合这个公式？未必。此时会有定价误差 α_i 出现：

$$\mathbb{E}[R_i^e] = \alpha_i + \beta_i' \lambda$$

- α_i 如果不显著偏离0，可能是统计误差。 α_i 如果显著不为0，则可能是：
 - 某些因子还没有被完全找到，当加入新的因子以后， α_i 不显著了
 - 即使因子都被完全找到了，模型本身也许有问题，例如非线性？
 - 从统计模型的角度说，这种分解总是能做的。但对于实际数据来讲，找到的因子数据是否能符合这条式子，不一定
 - 如果纯粹只做统计的分解(比如分解出latent factors)，对于实际的投资来讲，没有啥意义，因为我们总是希望找到数据之间的联系，而非仅仅是数据本身的某种分解
 - 如何检验模型本身有没有问题？样本内的拟合情况？样本外的预测情况？
 - 模型没错，市场也许对该资产出现了错误定价，形成了套利机会。此时这个资产就被称为一个“异象” (anomaly)
 - 若 α_i 为正，则买入 i 资产，卖出因子组合
 - 这也就是常见的所谓 α ，某种资产的收益中不能被已有的共性的因子描述的部分。如果等式左边是基金经理的资产组合收益，则 α 可能显示了这个基金经理超出常人的能力

总结：因为非系统性风险可以通过分散投资消除，因子描述了驱动个体资产预期收益的共性原因，也即某种系统性风险。因子收益率是这种系统性风险的风险溢价(risk premium)，是个体资产的共性收益。因为因子是系统性风险的衡量，其预期收益应当为正（统计意义上），作为对这种风险的补偿。

2. 如何检验？

时间序列回归

如果：

- λ 是收益率序列，且已知
- 已知 $R_{i,t}^e$ 的时间序列

则 β_i 需要估计。检验目标：

- 是否只用 $\beta_i' \lambda$ 就能解释 $\mathbb{E}[R_i^e]$ ，也即截距是否等于0
- λ 是否预期收益大于0

步骤：

1. 时间序列回归(对每个 i):

$$R_{i(t)}^e = \alpha_i + \beta_i' \lambda_{(t)} + \epsilon_{i(t)}$$

注：回归的样本用括号()表示，比如(t)表示在时间序列上回归

得到估计量 $\hat{\alpha}_i, \hat{\beta}_i$

2. 因子模型的估计为:

$$\frac{1}{T} \sum_{t=1}^T R_{it}^e = \hat{\alpha}_i + \hat{\beta}_i' \left(\frac{1}{T} \sum_{t=1}^T \lambda_t \right)$$

记: $\frac{1}{T} \sum_{t=1}^T \lambda_t \equiv \hat{\mathbb{E}}[\lambda]$, $\frac{1}{T} \sum_{t=1}^T R_{it}^e \equiv \hat{\mathbb{E}}[R_i^e]$ 。此二者即为因子预期收益、证券*i*预期收益的估计

3. 检验 $\hat{\alpha}_i$ 是否为0。单个 $\hat{\alpha}_i$ 只需做 *t* 检验即可。

4. 检验 $\hat{\alpha}_i$ 是否联合为0。Gibbons-Ross-Shanken (GRS) 检验:

$$\frac{T - N - K}{N} \left(1 + \hat{\mathbb{E}}[\lambda]' \hat{\Sigma}_\lambda^{-1} \hat{\mathbb{E}}[\lambda] \right)^{-1} \hat{\alpha}' \hat{\Sigma}^{-1} \hat{\alpha} \sim F_{N, T-N-K}$$

- *T*: 时间序列的样本数; *N*: 横截面样本数; *K*: 因子数
- $\hat{\Sigma}_\lambda = \frac{1}{T} \sum_{t=1}^T \left[\lambda_t - \hat{\mathbb{E}}[\lambda_t] \right] \left[\lambda_t - \hat{\mathbb{E}}[\lambda_t] \right]'$, $K \times K$ 矩阵
- $\hat{\Sigma} = \frac{1}{T} \sum_{t=1}^T \hat{\epsilon}_t \hat{\epsilon}_t'$, $N \times N$ 矩阵

GRS检验的假设:

- 时间序列上 ϵ_{it} 同方差, 无自相关, ϵ_{it} 和 λ_{tk} 独立, 也即 ϵ_{it} i.i.d. 有限样本下, F分布的结果显然也是假设了 ϵ_{it} 时间序列上正态分布
- 大样本下, 无需正态分布假设, 但对有限样本的 χ^2 统计量仍然假设了同方差和无自相关
- 若直接从矩条件出发:

$$\mathbb{E}[\epsilon_t] = 0, \quad N \text{ 个矩条件}$$

$$\mathbb{E}[\lambda_t \otimes \epsilon_t] = 0, \quad K \times N \text{ 个矩条件}$$

给出GMM估计量, 则可以给出Robust的error, 无需假设同方差和自相关。这里矩条件个数和参数个数 (α 有*N*个, β 有*N* × *K*个)一样, 恰好识别

5. 另外, 对于因子 λ_k , 也感兴趣其期望收益是否不为0, 单独做 *t* 检验即可。

两步法横截面回归(two-pass)

如果:

- λ_t 未知, 也即因子收益率不知道, 此时要想办法先估计 λ_t
- 然后检验 $\hat{\alpha}_i$ 是否联合为0

1. 时间序列回归(对每个*i*):

$$R_{i(t)}^e = a_i + \beta_i' f_{(t)} + u_{i(t)}$$

注意:

- $f_{(t)}$ 不是因子的收益率, 而是因子的某个代理变量, 或者就是因子值
- 此时截距项 a_i 不是定价误差
- β_i 是因子暴露的估计值。换句话说, 因为不能直接观察到和因子直接联系的因子收益

率，只能观察到因子值或因子的性质的某个代理变量，我们用这个时间序列回归来估计证券*i*的因子暴露。

2. 先求 R_{it}^e 时间上的均值，作为资产*i*的预期收益率的估计值， $\frac{1}{T} \sum_{t=1}^T R_{it}^e \equiv \hat{\mathbb{E}}[R_i^e]$ 。接着做如下横截面回归：

$$\hat{\mathbb{E}}[R_{(i)}^e] = \boldsymbol{\lambda}' \hat{\boldsymbol{\beta}}_{(i)} + \alpha_{(i)}$$

此时 $\boldsymbol{\lambda}$ 是因子预期收益率的估计值。 α_i 是回归的残差。因为是OLS，这个估计本身就是从“最小化 α_i 的平方和”出发来计算的。

- 这个回归也可以加截距。不加截距的回归，是理论；加了截距，是考虑理论也有可能是错的。因此，计量里常见的efficiency和robustness的讨论在这里也成立。也即：
 - 如果理论真的是对的，不加截距的回归能增强参数估计的efficiency，也即更有可能拒绝错误的原假设（因此，若不能拒绝原假设，则可能因子找错了）。加了截距项回归，则更robust，但牺牲了efficiency。如果此时也能拒绝原假设，那么更放心了。
 - 如果理论是错的，因子没找全从而应当加上截距项。此时若回归没加截距项，则遗漏变量，可能估计不一致。

3. 检验 $\hat{\alpha}_i$ 是否联合为0。这里的矩阵和向量都没有加粗，请留意。

$$\text{记 } \hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_{11} & \cdots & \hat{\beta}_{1K} \\ \vdots & \ddots & \vdots \\ \hat{\beta}_{N1} & \cdots & \hat{\beta}_{NK} \end{bmatrix}, \hat{\boldsymbol{\alpha}} = \begin{bmatrix} \hat{\alpha}_1 \\ \vdots \\ \hat{\alpha}_N \end{bmatrix}, \mathbf{u}_t = \begin{bmatrix} u_{1t} \\ \vdots \\ u_{Nt} \end{bmatrix}, \hat{\mathbb{E}}[R^e] = \begin{bmatrix} \hat{\mathbb{E}}[R_1^e] \\ \vdots \\ \hat{\mathbb{E}}[R_N^e] \end{bmatrix}, \hat{\boldsymbol{\lambda}} = \begin{bmatrix} \hat{\lambda}_1 \\ \vdots \\ \hat{\lambda}_K \end{bmatrix}$$

则参数估计为：

$$\hat{\boldsymbol{\lambda}} = \left(\hat{\boldsymbol{\beta}}' \hat{\boldsymbol{\beta}} \right)^{-1} \hat{\boldsymbol{\beta}}' \hat{\mathbb{E}}[R^e]$$

$$\hat{\boldsymbol{\alpha}} = \hat{\mathbb{E}}[R^e] - \hat{\boldsymbol{\beta}} \hat{\boldsymbol{\lambda}}$$

记 $\Sigma_f \equiv \text{cov}(f_t)$, ($K \times K$ 矩阵), $\Sigma \equiv \text{cov}(u_t)$, ($N \times N$ 矩阵)

则

$$\text{cov}(\hat{\boldsymbol{\lambda}}) = \frac{1}{T} \left[\left(\hat{\boldsymbol{\beta}}' \hat{\boldsymbol{\beta}} \right)^{-1} \hat{\boldsymbol{\beta}}' \Sigma \hat{\boldsymbol{\beta}} \left(\hat{\boldsymbol{\beta}}' \hat{\boldsymbol{\beta}} \right)^{-1} \left(1 + \boldsymbol{\lambda}' \Sigma_f^{-1} \boldsymbol{\lambda} \right) + \Sigma_f \right]$$

$$\text{cov}(\hat{\boldsymbol{\alpha}}) = \frac{1}{T} \left[I - \hat{\boldsymbol{\beta}} \left(\hat{\boldsymbol{\beta}}' \hat{\boldsymbol{\beta}} \right)^{-1} \hat{\boldsymbol{\beta}}' \right] \Sigma \left[I - \hat{\boldsymbol{\beta}} \left(\hat{\boldsymbol{\beta}}' \hat{\boldsymbol{\beta}} \right)^{-1} \hat{\boldsymbol{\beta}}' \right]' \cdot \left(1 + \boldsymbol{\lambda}' \Sigma_f^{-1} \boldsymbol{\lambda} \right)$$

统计量：

$$\hat{\boldsymbol{\alpha}}' \text{cov}(\hat{\boldsymbol{\alpha}})^{-1} \hat{\boldsymbol{\alpha}} \sim \chi_{N-K}^2$$

因子预期收益率 $\boldsymbol{\lambda}$ 的检验：从 $\text{cov}(\hat{\boldsymbol{\lambda}})$ 取出对角线上的元素，开方即是标准误 σ_{λ_k} 。用 $\hat{\lambda}_k / \sigma_{\lambda_k}$ 做自由度为T-1的t检验即可。

- $(1 + \lambda' \Sigma_f^{-1} \lambda)$ 是Shanken(1992)的调整，因为第一步的因子暴露 β 是估计出来的，要把估计误差考虑进来。
 - 在月度数据上检验模型，这个误差实际影响很小
- 也有对应的GLS版本和GMM版本，具体可查阅 Cochrane (2005)。

Fama-MacBeth 回归

1. 时间序列回归(对每个 i):

$$R_{i(t)}^e = a_i + \beta_i' f_{(t)} + u_{i(t)}$$

2. 在每个时间 t 上，做横截面回归

$$R_{(i)t}^e = \lambda_t' \hat{\beta}_{(i)} + \alpha_{(i)t}$$

- 也可以有截距项
- 和两步法横截面回归的区别：
 - 两步法的第二步先取时间平均，然后做一次横截面回归
 - FM 在每个时间 t 上都做横截面回归

好处：相当于增加了样本。

3. 假设误差在时间上独立，则有以下统计量：

$$\hat{\lambda} = \frac{1}{T} \sum_{t=1}^T \hat{\lambda}_t, \quad \hat{\alpha}_i = \frac{1}{T} \sum_{t=1}^T \hat{\alpha}_{it}$$

$$\sigma^2(\hat{\lambda}) = \frac{1}{T^2} \sum_{t=1}^T (\hat{\lambda}_t - \hat{\lambda})^2, \quad \sigma^2(\hat{\alpha}_i) = \frac{1}{T^2} \sum_{t=1}^T (\hat{\alpha}_{it} - \hat{\alpha}_i)^2$$

- 检验因子预期收益率是否为0，用 $\hat{\lambda}_t$ 序列做 t 检验即可
- 联合检验 $\hat{\alpha}_i$ 是否为0:

记 $\hat{\alpha} = \frac{1}{T} \sum_{t=1}^T \hat{\alpha}_t$ ，则可用

$$\text{cov}(\hat{\alpha}) = \frac{1}{T^2} \sum_{t=1}^T (\hat{\alpha}_t - \hat{\alpha})(\hat{\alpha}_t - \hat{\alpha})'$$

作为方差协方差矩阵的估计（横截面上的相关性，和两步法类似），检验：

$$\hat{\alpha}' \text{cov}(\hat{\alpha})^{-1} \hat{\alpha} \sim \chi_{N-K}^2$$

- 时间独立假设在资产定价问题里可以接受，因为收益率在时间上相关性很弱。公司金融里则未必。
- 即使时间上不独立，也可以很容易把估计量更改为调整了时间相关性的估计量，例如 $\sigma^2(\hat{\lambda}) = \frac{1}{T} \sum_{j=-\infty}^{\infty} \text{cov}_T(\hat{\lambda}_t, \hat{\lambda}_{t-j})$ ，当然可以用Newey-West调整
- 另外，第一步的时间序列回归，可以是每个 t 上得到不一样的 β_{it} 。比如，对应于 t 期，用 rolling window 做回归，用的数据是 $t-h$ 到 $t-1$ 期的(FM原文用的是5年的月收益率)，得到 $\hat{\beta}_{it-1}$ 。接下来在第二步时就用这些每一期不一样的 $\hat{\beta}_{it-1}$ 对 R_{it}^e 做横截面回归

- 第二步的回归就可以这样理解：在每个时间点上， $\hat{\beta}_{it-1}$ 是这个N个资产在横截面上的差异，这些差异能否带来第t期的 R_{it}^e 的不一样

3. 方法差异（用什么作为因子收益率？）

问题：如果因子收益已知（也即这类因子的特征可以直接从可交易的证券中观察到，例如size, momentum, 而非GDP），那么时间序列上直接求平均就是因子预期收益的估计。此时如果仍然做两步法横截面回归或者FM回归，又得到一个从OLS而来的所谓的因子预期收益的估计，这两者有什么不同？该用哪个？

差异之处：

横截面回归时，

$$\hat{\lambda} = (\hat{\beta}'\hat{\beta})^{-1} \hat{\beta}'\mathbb{E}[R^e]$$

注意 $\hat{\beta} = \begin{bmatrix} \hat{\beta}_{11} & \cdots & \hat{\beta}_{1K} \\ \vdots & \ddots & \vdots \\ \hat{\beta}_{N1} & \cdots & \hat{\beta}_{NK} \end{bmatrix}$ 是因子暴露矩阵， $\hat{\beta}_{ij}$ 是第*i*个资产在第*j*个因子上的暴露

$(\hat{\beta}'\hat{\beta})^{-1} \hat{\beta}'$ 是一个 $K \times N$ 矩阵，第*k*行可以看成是第*k*个因子的投资组合，所以 $(\hat{\beta}'\hat{\beta})^{-1} \hat{\beta}'$ 的第*j*行、第*i*列的值是第*j*个因子在第*i*个资产上的权重

此时，

$$\left[(\hat{\beta}'\hat{\beta})^{-1} \hat{\beta}' \right] \hat{\beta} = (\hat{\beta}'\hat{\beta})^{-1} (\hat{\beta}'\hat{\beta}) = I_K$$

因此： $(\hat{\beta}'\hat{\beta})^{-1} \hat{\beta}'$ 这个因子的投资组合权重矩阵有如下性质：

- 第*k*个因子在所有资产上的投资组合 $(\hat{\beta}'\hat{\beta})^{-1} \hat{\beta}'$ 的第*k*行给出的权重，在第*k*个因子上的暴露是1
- 第*k*个因子在所有资产上的投资组合 $(\hat{\beta}'\hat{\beta})^{-1} \hat{\beta}'$ 的第*k*行给出的权重，在其他因子上的暴露是0
- 因此，横截面回归得到的因子预期收益率的估计，更“纯净”

排序法

因子的概念是抽象的：

- 因子收益率是什么？如何估计？
- 因子暴露是什么？如何估计？

实践中，只能考虑用可以交易的证券的组合，来模拟某个因子收益率。理想的模拟组合，应当满足：

1. 只在这个被模拟的因子上暴露，在其他因子上不暴露
 2. 如果有多种满足条件1的组合，应当选择特质性风险较小的那个组合
- 条件1很好理解：如果同时在其他因子上有暴露，则这个投资组合的收益率不能仅仅归因于这个因子，同其他因子也有关系。

- 条件2说的是： $R_{it}^e = \alpha_i + \beta_{ki}\lambda_{kt} + 0 * \lambda_{[k]t} + \epsilon_{it}$ 中的 ϵ_{it} 的变化要比较小，否则也会干扰这个投资组合的模拟程度。中括号表示除 k 之外的其他因子，其 β 为 0

鸡生蛋还是蛋生鸡？

- 按照理想中的模型，估计因子暴露需要知道因子收益率
- 要知道因子收益率，需要知道因子模拟组合，而构建因子模拟组合，需要知道因子暴露

如果因子是可以交易的证券的某个特征，比如 Book/Market ratio，则利用排序法可以模拟构建因子收益率。

排序法：

别估计因子暴露了，就用这个想法：如果证券对应的这个特征的指标高，则它在这个因子上的因子暴露就高

- 这个特征并不一定是因子暴露。也没有什么模型（也可以说是非参模型）
- 因此，如果因子不是可以交易的证券的特征，比如 GDP，那么得回到两步回归的方法来估计因子暴露和因子收益率

抛开所有理论的直觉：拥有某个特征的某种证券(因子暴露更高)，在未来的收益会更好(因子预期收益为正)

单排

1. 确定特征，选择股票池
2. 按照特征的取值大小分组，例如 10 组。构建多空组合：做多最高组的股票，做空最低组的股票。通常要求多空两组的金额相同，无需自有资金。由于多空两组通常包含多只证券，还需考虑单只证券的权重，例如按市值加权
3. 定时更新(rebalance)

双排

双排可以控制另一个特征(因子)的影响

一般常见的 double sorting 有两种，一种是 independent sort, 一种是 dependent sort.

- independent sort 就是对于 variable 1, 计算 quantile, 按照 quantile 分组。对于 variable 2, 计算 quantile, 再分组。取之间的交集。
 - independent sort 的好处是对于 variable 1, 2 的顺序没有要求。坏处是如果 variable 1 和 variable 2 高度相关，则可能导致某些交叉组里的成员很多，某些交叉组里的成员又很少，甚至没有。
- dependent sort 就是对于 variable 1, 计算 quantile, 分组。然后在每组内部，按照 variable 2 计算 quantile, 再分成小组。
 - dependent sort 相当于是一个 "conditional on", 也就是 "基于" variable 1 的关于 variable 2 的分组。看的是在 variable 1 各个组内部 variable 2 的情况。
 - 因此，dependent sort 的顺序很重要

4. 截面多因子模型 vs. 时序多因子模型 (用什么作为因子暴露?)

估计因子暴露似乎有很多种套路：

- 最传统的 Fama-French 方法：先按照特征(特征并不是因子暴露)排序得到因子收益率，

时序回归得到因子暴露。这种传统方法被称作“时序多因子模型”

- 就把特征作为因子暴露，不做时序回归了。直接开始做第二步在每个时间 t 上做横截面回归，然后考察 λ_t 的均值是否不为0。这种被称作“横截面多因子模型”。
- 甚至还可以这么做：
 - 先用公司特征和 R_{it}^e 做每个 t 上的横截面回归，得到 λ_t
 - 再用 λ_t 和 R_{it}^e 做时序回归，又得到 $\hat{\beta}_{it}$
- Fama French (2020)给出了这几种方法的比较。结论：横截面多因子模型的解释力比较好。

5. 检验异象

排序法检验

如果一个因子和另一个因子的相关性很强，则按照一个因子进行排序可能捕捉的是另一个因子的风险。此时可以用双排法：

同时用两个因子分别独立排序，把交集拿出来检验。

多个因子时，多重排序法存在问题：交集会越来越小，甚至没有交集。

时序检验

1. 已知一个异象的收益率序列 $R_{Anomaly,t}^e$ （例如，可以用exposure构建的排序股票组合）。已知其他因子收益率序列 λ_t 。把 $R_{Anomaly,t}^e$ 放等式左边， λ_t 放等式右边，做时间序列回归
2. 看截距项是否显著不为0 (NW 调整)

例：

$$R_{Anomaly,t}^e = \alpha + \beta_1 R_{Mt}^e + \beta_2 SMB_t + \beta_3 HML_t + \epsilon_t$$

FM回归检验

1. 已知异象的指示变量(exposure)，已知其他因子的factor exposure。用下一期个股收益放等式左边，factor exposures 放等式右边
2. 在每个时间 t 上做横截面回归，得到参数估计 $\hat{\lambda}_t$ 的时间序列
3. 看 $\hat{\lambda}_t$ 是否平均不为0 (NW调整)

例：

$$R_{it+1}^e = \gamma_t + \lambda_{0t} Anomaly_{it} + \lambda_{1t} \beta_{it} + \lambda_{2t} Size_{it} + \lambda_{3t} BM_{it} + u_{it}$$

细节：求简单均值时，如何做 NW 调整来计算标准误？

解答：做单变量回归。 $\lambda_{0t} = \text{intercept} + u_t$ ，intercept的估计值就是 λ_{0t} 的均值，同时，软件可以轻松计算NW调整，无需自己手动计算。

6. 总结

个人经验：

- 各种方法差异不大。一种方法给出了显著的结果，另一种方法通常也不会不显著。
- 为了稳健起见，学术论文通常考察多种方法，各种检验。
- 然而，p-hacking、数据挖掘出来的假象难以避免。
- 另一方面，市场本身不是平稳的，风险源一直都在变，样本外的表现不稳定。

- 实践中，应当仔细思索因子的经济含义，避免陷入数据挖掘的狂热。
 - The cost of computing has dropped exponentially, but the cost of thinking is what it always was. That is why we see so many articles with so many regressions and so little thought. --Zvi Griliches

参考文献

Cochrane, John H., 2005, *Asset Pricing*. Revised edition. (Princeton University Press, Princeton, N.J).