Fama-Macbeth回归讨论

邵一淼 191098180

一、引言——从CAPM到Fama-Macbeth

Fama-Macbeth回归是1973年Fama和Macbeth为验证CAPM模型而提出的一种因子统计方法，该模型现如今被广泛用于计量经济学的面板数据分析，而在金融领域在用于多因子模型的回归检验，用于估计各类模型中的因子暴露和因子收益（风险溢价）。

Fama-Macbeth回归是实证资产定价中最常用的方法之一。它的主要用途是验证因子对资产收益率是否产生系统性影响。与投资组合分析不同的是，Fama-Macbeth回归可以在同时控制多个因子对资产收益率的影响下，考察特定因子对资产收益率产生系统性影响，具体体现在因子是否存在显著的风险溢价。

CAPM：$r_i = r_f + (r_m - r_f) * \beta$

这个公式有三个含义：

1、风险与收益的关系是线性的

2、β是对系统性风险的完全度量

3、$r_m - r_f > 0$，在一个风险规避的世界，更高的风险要有更高的收益

要验证CAPM只需要看满不满足上面的三个条件，因此，设定要拟合的模型为：

$r_i = \gamma_0 + \gamma_1 * \beta + \gamma_2 * \beta^2 + \gamma_3 * s + \epsilon$

s是系统性风险，ϵ为残差项

条件1成立有：$\gamma_2 = 0$

条件2成立有：$\gamma_3 = 0$，非beta风险不具有系统性影响

条件3成立有：$\gamma_1 = r_m - r_f > 0$

sharpe-lintner capm假定：$\gamma_0 = r_f$

详细的步骤为（以论文为例）：

1、用四年1926-1929的月收益率，对个股进行时序回归，计算出beta，排序分组为20组

2、用之后五年共60个月1930-1934年的月收益率，重新时间序列回归计算出个股beta和个股残差标准差$s(\epsilon)$，在计算出beta（个股beta的直接简单平均）和组合残差标准差$s_p(\epsilon)$（个股残差标准查直接简单平均）

3、之后四年1935-1938，每一个月都进行一次截面回归，那么四年回归48次。每一次截面回归的因子都是用上期获得的因子（不是上个月的因子）。具体来说，他是个这样的结构化数据：

	ri	**beta**	**beta**2**	**s(e)**
1935-01	group1			
	group2			
	...			
	group20			
1935-02	group1			
	group2			
	...			
	group20			

...使用过去60个月的数据进行时间序列回归，得到因子（第二步）后进行截面回归。

然后进行滚动回归，每次都是用过去的60个月跟新beta与s(e)因子。

所以每个时点的截面回归也就20个样本。

到1935-02，重新滚动分组，从第一步开始，用1930-02~1935-01这60个月的数据进行时间序列回归后得到每组的beta与s(e)

（这里，关于什么是滚动分组：beta是滚动计算的，如果分组里的股票一直不变显然是不太合理哈。因为分组是按照beta分组，是重要保证高beta组的股票的beta始终高。所以分组不能固定，一个股票在上期可能是第一组，下一期可能是第二组。也就是1935-02的分组依据是用四年的月收益率，对个股进行时序回归，计算出beta然后分组）

我们将我们设定的模型加上下标t表示时间，加上p表示组合，因为最后的截面回归就是组合之间的：

$$r_{pt} = \gamma_{0t} + \gamma_{1t}\beta_{p,t-1} + \gamma_{2t}\beta_{p,t-1}^2 + \gamma_{3t}s(\epsilon)_{p,t-1} + u$$

p=1,2,...20

u表示残差，ϵ是时间序列回归得到的残差。

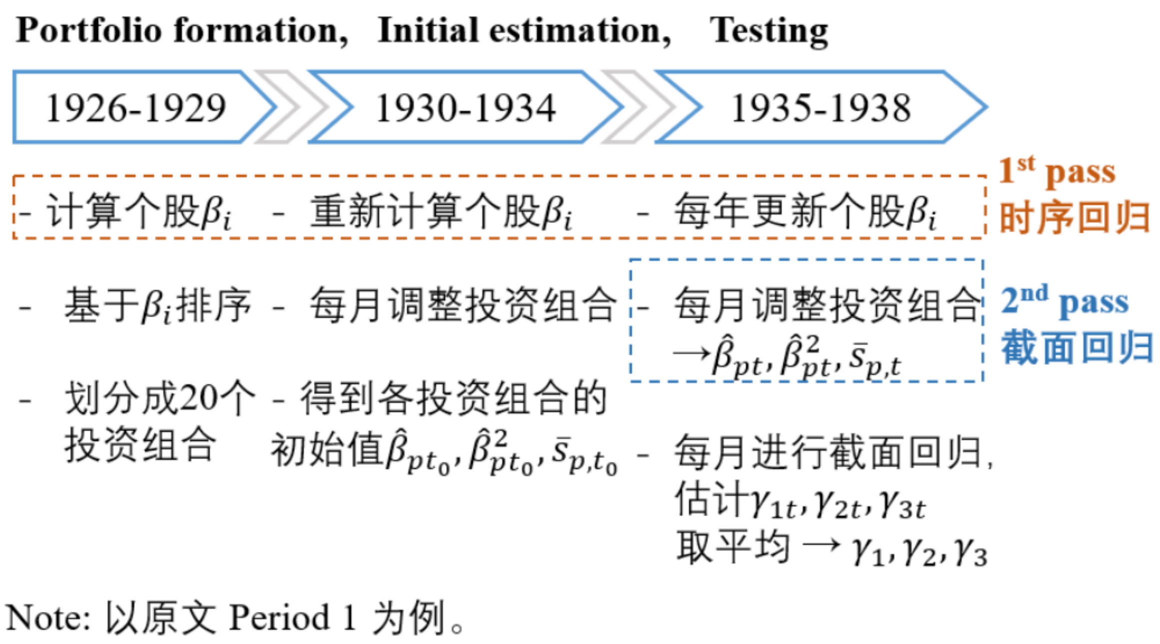
4、对所有截面回归得到的参数求均值，得到我们对参数的最终估计。第三步界面回归完成后，我们得到了这样的结构化数据：

	gamma0	**gamma1**	**...**
1935-01			
1935-02			
1935-03			
...			

参数$\gamma_i = \frac{1}{T}\sum\gamma_{it}, i = 0,1,2,3$，标准差就是直接求标准差，那么t统计量也有了：

$t = \frac{\bar{\gamma}}{\sqrt{\frac{s\gamma}{n}}}$

然后就可以进行假设检验了



[重读Fama——从CAPM到Fama-Macbeth回归再到三因子 - 知乎 (zhihu.com)]

二、方法概述

Fama-Macbeth与传统的截面回归类似，本质上也与是一个两阶段回归，不同的是它用巧妙的方法解决了截面相关性的问题，从而得出更加无偏，相合的估计。

时间序列回归

Fama-Macbeth模型与传统截面回归相同，第一步都是做时间序列回归。在因子分析框架中，时间序列回归是为了获得个股在因子上的暴露。如果模型中的因子是 portfolio returns（即使用投资组合收益率作为因子，例如Fama-French三因子模型中的SMB，HML和市场因子），那么可以通过时间序列回归（time-series regression）来分析$E[R_i]$和β_i在截面上的关系。

令f_t为因子组合在t期的收益率，R_{it}为个股i在t期的收益率，用f_t对每只股票的R_{it}回归，即可得到每只股票的全样本因子暴露β_i。

$R_{it} = \alpha_i + \beta_i f_t + \epsilon_{it}, t = 1,2,\ldots,T, \forall i$ （1）

也可滚动计算某个时间段的因子暴露β_{it}，体现个股随市场的变化设置时间段长度为period

$R_{ik} = \alpha_i + \beta_{it} f_k + \epsilon_{ik}, k = t - period, 2, \ldots, t, \forall i$ （2）

截面回归

传统截面回归的第一步是通过时间序列回归得到个股暴露，这一步与Fama-Macbeth回归相同，而第二步回归体现了传统截面回归和Fama-Macbeth的最大不同。

传统截面回归：

在时序回归中回归式在时间序列上取均值，在$E[\epsilon] = 0$的假设下可以得出：

$E[r_i] = \alpha_i + \beta_i E[f]$ （3）

上式正是个股的期望收益与因子暴露在截面上的关系，截距α_i为个股的错误定价。

那么便可通过截面回归找到因子的期望收益率$E[f]$，方法是最小化个股定价错误α_i的平方和。对个股的的收益在时序上取均值得到个股期望收益$E[R_i]$，用全样本的个股因子暴露对个股期望收益做无截距回归。

$E[r_i] = \beta_i \lambda + \alpha_i$ （4）

回归残差α_i为个股的错误定价，λ为因子的期望收益率。

截面回归最大的缺陷在于忽略了截面上的残差相关性，使得OLS给出的标准误存在巨大的低估。

Fama-Macbeth回归

与截面回归相同，Fama-Macbeth回归第一步是通过时间序列回归得到因子暴露值，不同的是，第二步中，Fama-Macbeth在每个t上都做了一次无截距截面回归：

$R_{it} = \beta_i \lambda_t + \alpha_{it}, i = 1, 2, \ldots, N, \forall t$ (5)

上式中的β_i为全样本β，当然若使用滚动回归数据，也可以在不同截面的回归上使用对应时期的$\beta_{i,t}$。

Fama-Macbeth回归相当于在每个t上做一次独立的截面回归，这T次回归的参数取均值作为回归的估计值：

$\hat{\lambda} = \frac{1}{T} \sum_{t=1}^{T} \hat{\lambda}_t$

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

上述方法的巧妙之处在于它把 T 期的回归结果当作 T 个独立的样本。参数的 standard errors 刻画的是样本统计量在不同样本间是如何变化的。在传统的截面回归中，我们只进行一次回归，得到λ和α_i的一个样本估计。而在Fama-Macbeth截面回归中，把T期样本点独立处理，得到T个λ和α_i的样本估计。

若使用全样本因子暴露β_i进行估计,截面回归和Fama-Macbeth的估计结果相同，当使用滚动窗口进行估计时（Fama and MacBeth (1973)中作者使用了滚动窗口），截面回归和Fama-Macbeth回归会得到完全不同的估计结果。

Fama-Macbeth回归很好的解决了截面相关性的问题，但对于时间序列上的相关性仍然无力。

（用一些更通俗的话来说，Fama-Macbeth回归的两步是：

1、估计资产承担风险大小（beta值）。通过对资产收益率的时间序列分析，得到资产承担的风险水平。

2、估计风险溢价时间序列以及统计检验。通过在每个时点的资产收益率对得到的beta值进行截面回归，得到因子在每个时刻的风险溢价。对每个时刻的风险溢价进行平均，并检验均值是否显著异于0.)

三、模型设定

还是从最简单的CAPM模型开始：

$E(r_i) - r_f = \beta_i [E(r_{Mkt}) - r_f], i = 1, 2, \ldots N$

其中，$E[r_i]$、$E[r_{Mkt}]$分别指资产预期收益率和市场组合预期收益率，r_f是无风险利率，β_i是资产i对市场风险的因子暴露，$E(r_{Mkt}) - r_f$是市场组合风险溢价。上式说明了任何资产 的超额收益都由其对系统风险的暴露决定，即刻画了单个资产预期超额收益率在截面上和市场贝塔的线性关系。

用R_i表示资产i的超额收益，即$R_i = r_i - r_f$，用λ表示市场组合风险溢价，即$\lambda = E(r_{Mkt}) - r_f$，那么CAPM可以写为下式：

$E(R_i) = \beta_i' \lambda, i = 1, 2, \ldots, N$

更一般的，如果有k个因子，λ就是k维因子溢价向量，β_i是资产i在k个因子上的k维因子暴露向量。我们希望通过估计λ来检验模型中资产预期超额收益和因子暴露向量β的线性关系是否稳健，并确定定价误差的大小。

对于上述模型，FM方法分两步进行估计：

1、先对各个资产i进行时间序列回归，估计β_i：

$$R_i = \alpha_i + \beta_i' f_t + \epsilon_t^i, t = 1, 2, \ldots, N$$

2、用第一部得到的$\hat{\beta}_i$作为自变量，在各个时期t分别做截面回归：

$$R_i = \hat{\beta}_i{}' \lambda_t + \alpha_{it}, i = 1, 2, \ldots N$$

最后，对每一期的估计结果$\hat{\lambda}_t$、$\hat{\alpha_{it}}$在时序上取平均值得到λ和α_i的估计：

$$\hat{\lambda} = \frac{1}{T} \sum_{t=1}^{T} \hat{\lambda}_t$$

$$\hat{\alpha}_i = \frac{1}{T} \sum_{t=1}^{T} \hat{\alpha_{it}}$$

$\hat{\lambda}$和$\hat{\alpha}_i$的标准误可以由各个时期截面回归$\hat{\lambda}_t$、$\hat{\alpha_{it}}$的标准差得到：

$$\sigma^2(\hat{\lambda}) = \frac{1}{T} \sum_{t=1}^{T} (\hat{\lambda}_t - \hat{\lambda})^2$$

$$\sigma^2(\hat{\alpha}_i) = \frac{1}{T} \sum_{t=1}^{T} (\hat{\alpha_{it}} - \hat{\alpha}_i)^2$$

FM 方法的优势是得到了异方差稳健的标准误，且修正了面板数据在截面上的相关性。此外，由于第二步是每期做一次截面回归，所以允许使用时变的β 做自变量。然而，这种以第一阶段估计量作为第二阶段自变量的方法引入了变量误差问题 (EIV 问题)，且 FM 标准误对残差序列时序相关是不一致的，这就需要对标准误进行修正。

EIV问题的修正

由于以第一阶段估计量作为第二阶段自变量引入了 EIV 问题，FM 标准误是不一致的。Shanken (1992) 给出了对 EIV 的修正项。如果残差项ϵ在时间上独立同分布且与因子收益率独立，均方误估计可以由下式给出 (Cochrane，2005；Shanken，1992)：

$$\sigma^2(\hat{\lambda_{OLS}}) = \frac{1}{T} [(\beta'\beta)^{-1} \sum \beta (\beta'\beta)^{-1} (1 + \lambda' \sum_f \lambda) + \sum_f]$$

其中，$\sum_f$是第一阶段回归因子收益率协方差矩阵$cov(f_t, f_t')$，$\sum$是第一阶段回归残差协方差矩阵$cov(\epsilon_t, \epsilon_t')$

时序相关问题和修正

累了，有空再更

四、非球型扰动项及Newey-West 调整

Fama-macbeth每一期使用当期因子暴露和个股下一期的收益率进行截面回归，得到因子的收益率；在全部期进行截面回归后，便可得到每个因子收益率的时间序列。将因子收益率在时序上取均值就得到每个因子的预期收益率，而我们关心的是该因子预期收益率是否显著不为零。

对于任何因子，其收益率序列在时序上很可能存在异方差和自相关性，因此在计算其均值标准误的时候需要进行 Newey-West 调整。然而，这和上面的多因子时序回归很不相同。如何进行 Newey-West 调整呢？

简单说明Newey West的原理：

考虑一个线性模型:

$$
\begin{aligned}
y = X\beta + \epsilon \\
E[\epsilon|X] = 0 \\
E[\epsilon\epsilon'] = \sigma^2\Omega \\
\hat{\beta} = (X'X)^{-1}X'y = \beta + (X'X)^{-1}X'\epsilon
\end{aligned}
$$

$$
\begin{aligned}
Var[\beta] &= E[(\beta-\hat{\beta})(\beta-\hat{\beta})'|X] \\
&= \frac{1}{T}(\frac{1}{T}X'X)^{-1}(\frac{1}{T}X'\sigma^2\Omega X)(\frac{1}{T}(X'X)^{-1})
\end{aligned}
$$

当残差不存在异方差和自相关性时，残差协方差阵为单位阵的倍数，回归系数的协方差估计是一致估计量，当残差存在异方差或自相关性时，协方差阵估计有问题，可以通过Newey West调整解决，具体来说是估计上式中的

$Q = \frac{1}{T}X'\sigma^2\Omega X$

Newey West调整即对Q进行估计，最终给出的估计量具有一致性，表达式如下，用S表示

$S = \frac{1}{T}(\sum_{i=1}^{T} e_t^2 x_T x_t' + \sum_{l=1}^{L}\sum_{t=l+1}^{T} w_l e_t e_{t-1}(x_t x_{t-l}' + x_{t-l}x_t'))$

$where \quad w_l = 1 - \frac{l}{1+L}$

上式中，括号中第一项为仅有异方差时的调整，后面一项为针对自相关的调整，其中，e为样本残差，L为计算自相关性影响的最大滞后阶数，w_l是滞后期l的系数，从公式来看，随着滞后期数的增加，影响减小。将S带入系数协方差阵的估计可以得到协方差的Newey West估计量

$Var[\hat{\beta_{NW}}] = T(X'X)^{-1}S(X'X)^{-1}$

以上是对于OLS的Newey West调整，对于Fama Macbeth回归，是对已经回归出来的一堆beta系数序列的方差进行调整，跟回归有一定差别，可以做一个转换：用回归出来的所有beta做因变量，1做自变量，做一个回归，这样回归出来的系数是所有beta的均值，残差也捕捉了beta中的异方差性和自相关性，对这个回归方程做newey west即可

Turan Bali、Robert Engle、Scott Murray 三位所著的经典教材 Empirical Asset Pricing, the cross section of stock returns（Bali et al. 2016）提出对于单个因子的收益率序列，将其用 1 作为 regressor 回归得到残差 —— 这相当于用因子收益率减去它在时序上的均值。然后把这个残差和 **X = 1** 代入到 Newey-West 调整中即可。

在这个简化版的Newey-West 调整中，Q的估计S简化为：

$S = \frac{1}{T}\sum_{t=1}^{T} e_t^2 + 2\sum_{t=1}^{L}\sum_{t=l+1}^{T} w_l e_t e_{t-1}$

$where \quad e_t = f_t - E_t[f_t] \quad w_l = 1 - \frac{l}{1+L}$

其中 f_t代表被检验因子的收益率时间序列，$E_t[f_t]$ 是它在时序上的均值。由于我们仅仅有一个 regressor，因此上述S其实是一个标量。将它代入到 **V**_OLS 的表达式中，在对其开方，就得到 $E_t[f_t]$ 的标准误:

$s.e.(E_t[f_t]) = \sqrt{S/T}$

对每个因子依次使用上述修正，获得其各自收益率均值的 standard error，然后就可以计算 t统计量以及 p-value 并检验它们的显著性

推导FM中的标准误：

当残差不独立时，估计回归系数和标准误差的另一种方法是法马-麦克贝斯方法(Fama和MacBeth，1973)。在这种方法中，研究人员进行了T次横断面回归。T估计数的平均值为系数估计数：

$$
\begin{aligned}
\hat{\beta}_{FM} &= \sum_{t=1}^{T} \frac{\hat{\beta}_t}{T} \\
&= \frac{1}{T}\sum_{t=1}^{T}(\frac{\sum_{i=1}^{N} X_{it}Y_{it}}{\sum_{i=1}^{N} X_{it}^2}) = \beta + \frac{1}{T}\sum_{t=1}^{T}(\frac{\sum_{i=1}^{N} X_{it}\epsilon_{it}}{\sum_{i=1}^{N} X_{it}^2})
\end{aligned}
$$

法马-麦克贝估计的估计方差计算为:

$$
S^2(\hat{\beta}_{FM}) = \frac{1}{T}\sum_{t=1}^{T}\frac{(\hat{\beta}_t - \hat{\beta}_{FM})^2}{T-1}
$$

上面这个方差公式假设了系数(β_t)的估计量是相互独立的，但这仅在$X_{it}\epsilon_{it}$与$X_{is}\epsilon_{is}$(t!=s)独立的时候才是对的。如果存在firm effect($i.e., \rho_X\rho_\epsilon ! = 0$)。因此，在存在firm effect的情况下，FM方差估计太小了。在这种情况下，FM估计的渐近方差为

$$
\begin{aligned}
Avar(\hat{\beta}_{FM}) &= \frac{1}{T^2}Avar(\sum_{t=1}^{T}\hat{\beta}_t) \\
&= \frac{Avar(\hat{\beta}_t)}{T} + \frac{2\sum_{t=1}^{T-1}\sum_{s=t+1}^{T} Acov(\hat{\beta}_t, \hat{\beta}_s)}{T^2} \\
&= \frac{Avar(\hat{\beta}_t)}{T} + \frac{T(T-1)}{T^2}Acov(\hat{\beta}_t, \hat{\beta}_s)
\end{aligned}
$$

其中，

$$
\begin{aligned}
ACov(\hat{\beta}_t, \hat{\beta}_s) =& plim_{N\to\infty}[(\frac{\sum_{i=1}^{N} X_{it}^2}{N})^{-1})(\frac{\sum_{i=1}^{N} X_{it}\epsilon_{it}}{N})] \\
&* (\frac{\sum_{i=1}^{N} X_{is}\epsilon_{is}}{N})(\frac{\sum_{i=1}^{N} X_{is}^2}{N})^{-1}] \\
&= (\sigma_X^2)^{-2} plim_{N\to\infty}[(\frac{\sum_{i=1}^{N} X_{it}\epsilon_{it}}{N})(\frac{\sum_{i=1}^{N} X_{is}\epsilon_{is}}{N})] \\
&= (\sigma_X^2)^{-2} plim_{N\to\infty}[\frac{\sum_{i=1}^{N} X_{it}X_{is}\epsilon_{it}\epsilon_{is}}{N}] \\
&= (\sigma_X^2)^{-2}\frac{N\rho_X\sigma_X^2\rho_\epsilon\sigma_\epsilon^2}{N^2} \\
&= \frac{\rho_X\rho_\epsilon\sigma_\epsilon}{N\sigma_X^2}
\end{aligned}
$$

代入得FM系数估计的渐近方差的表达式：

$$
\begin{aligned}
Avar(\hat{\beta}_{FM}) &= \frac{Avar(\hat{\beta}_t)}{T} + \frac{T(T-1)}{T^2}ACov(\hat{\beta}_t, \hat{\beta}_s) \\
&= \frac{1}{T}(\frac{\sigma_\epsilon}{N\sigma_X^2}) + \frac{T(T-1)}{T^2}(\frac{\rho_X\rho_\epsilon\sigma_\epsilon^2}{N\sigma_X^2}) \\
&= \frac{\sigma_\epsilon^2}{NT\sigma_X^2}(1+(T-1)\rho_X\rho_\epsilon)
\end{aligned}
$$

五、遗漏变量偏差与mimicking portfolio

如果遗漏了一些定价因素，则线性资产定价模型中风险溢价的标准估计量就会是有偏差的，即遗漏变量偏差。mimicking-portfolio approach通常选择一小部分投资组合（例如，按规模和按市场数量分类的投资组合）来预测收益率。但是理论上不能保证控制或跨越组合足以纠正遗漏的变量偏差。

假设$v_t=(v_{1t},v_{2t})^T$是一个包含两个潜在相关因子的向量。
考虑模型：
$r_t=\beta\gamma+\beta v_t+u_t$
其中，u_t是特异性风险，$\beta=(\beta_1:\beta_2)$是风险暴露矩阵，$\gamma=(\gamma_1:\gamma_2)$是两个因子的风险溢价。我们关心的是估计第一个因子$v_{1t}$（记为$g_t$)的代理风险溢价，在上面的简单假设中他的风险溢价就是γ_1。

Fama-Macbeth的做法是：首先，将每个测试资产的超额回报进行时间序列回归，估计出资产的风险暴露，β_1和β_2。然后，对估计的β_1和β_2进行截面回归，得到γ_1和γ_2的风险溢价估计。

而mimicking-portfolio方法通过将因子投射到一组可交易的资产回报上来估计g_t的风险溢价，因此构建一个与g_t最大相关的可交易投资组合（这也是为什么他也被称作是最大相关的模拟投资组合）。g_t的风险溢价随后被估计为其模拟投资组合的平均超额收益。

首先考虑估计截面回归中$g_t=v_{1t}$的风险溢价，此时遗漏了v_{2t}。这个遗漏会在两个阶段中都产生偏差：
1、 只要v_{2t}与v_{1t}相关，时序回归中就会产生β_1的有偏估计。偏差的大小取决于这些因子的时序相关性。
2、第二次偏差发生在第二步截面回归中，我们本希望将平均收益回归到整个风险暴露矩阵β上，但是因为v_{2t}被忽略了，所以实际只使用了$\tilde{\beta}_1$。偏差的大小取决于β_1和β_2的截面相关性。
\end{enumerate}
这两种偏差（第一步遗漏的v_{2t}和第二步遗漏的β_2)都会影响g_t的风险溢价的估计。

在mimicking-portfolio方法中，遗漏变量偏差可能会来源于g_t被投射到的资产的遗漏。考虑$g_t=v_{1t}$对一组测试资产的超额收益的映射，$\tilde{r}_t$。这个映射产生的系数是$w^g=Var(\tilde{r_t})^{-1}Cov(\tilde{r}_t,g_t)$，这些是$g_t$的模拟投资组合的权重，其超额回报是$r_t^g=(w^g)^T\tilde{r}_t$。

因此，我们可以将模拟组合的期望超额收益写成：
$\gamma_g^{MP}=(w^g)^TE(\tilde{r_t})$
由于测试资产$\tilde{r}_t$遵循与r_t相同的定价模型。我们可以写为：
$\tilde{r_t}=\tilde{\beta}\gamma+\tilde{\beta}v_t+\tilde{u_t}$

同样的，我们可以写出mimicking-portfolio的第一个因子的风险溢价：
$\gamma_g^{MP}=((\tilde{\beta}\sum^v\tilde{\beta}^T+\sum^{\tilde{u}})^{-1}(\tilde{\beta}\sum^v e_1))^T\tilde{\beta}\gamma$
其中e_1是列向量$(1,0)^T$，$\sum^v$是因子协方差矩阵，$\sum^{\tilde{u}}$是映射中使用的资产的特异性风险的协方差矩阵。

上述公式表明，并非所有的g_t投影到的资产的选择都会带来γ_1的一致估计，也就是说不能保证
$\gamma_g^{MP}=\gamma_1$

但是只要满足以下两个条件，那么就有$\gamma_g^{MP}=\gamma_1$，如果被选作为投资组合的资产：
1、很多样化$\tilde{\sum}^u\approx 0$
2、完全涵盖真实的因子v_t，这样$\tilde{\beta}$是可逆的并且$v_t=\tilde{\beta}^{-1}\tilde{r}_t$
然而，当这些条件不满足时，模拟投资组合估计通常会有偏差。特别是，如果在预测中使用的资产集忽略了一些有助于跨越v_t中所有风险因素的投资组合，那么模拟投资组合估计将会有偏差。

六、Stata实现

为简单说明Fama-Macbeth两阶段回归的主要步骤，以下用投资组合数据估计一个简单的 CAPM 模型。数据主要使用了[25 Portfolios Formed on Size and Book-to-Market] 中的 25 个投资组合 1926.7-2020.10 期间的月度收益率(RP.csv)，和[Fama/French 3 Factors] 中的无风险收益、市场超额收益数据(Mkt-RF.csv)。

数据说明：仓库中RP.csv中存储的是25 个投资组合 1926.7-2020.10 期间的月度收益率，每行代表一个月份，每列代表一个投资组合；Mkt-RF.csv存储的是1926.7-2020.10 期间的无风险收益、市场超额收益数据，每行代表一个月份，Mkt-RF和RF列代表市场超额收益率和无风险收益。

数据预处理：

变量	含义
port_num	投资组合编号，1~25
t	时期，如1936m7格式
rpe	超额收益，投资组合收益-无风险收益

第一阶段：

pass1 1930.1-1938.11：25*48次时序回归（1930.1-1934.12->1933.12-1938.11）

估计$beta_{it}, i=1,2..25$，窗口为五年，每次向后移动一个月

```
bys port_num: asreg rp mktrf if (t>=ym(1930,1) & t<=ym(1938,12)) , wind(t 60)
rmse se newey(4)
```

```
. list in 46/50
```

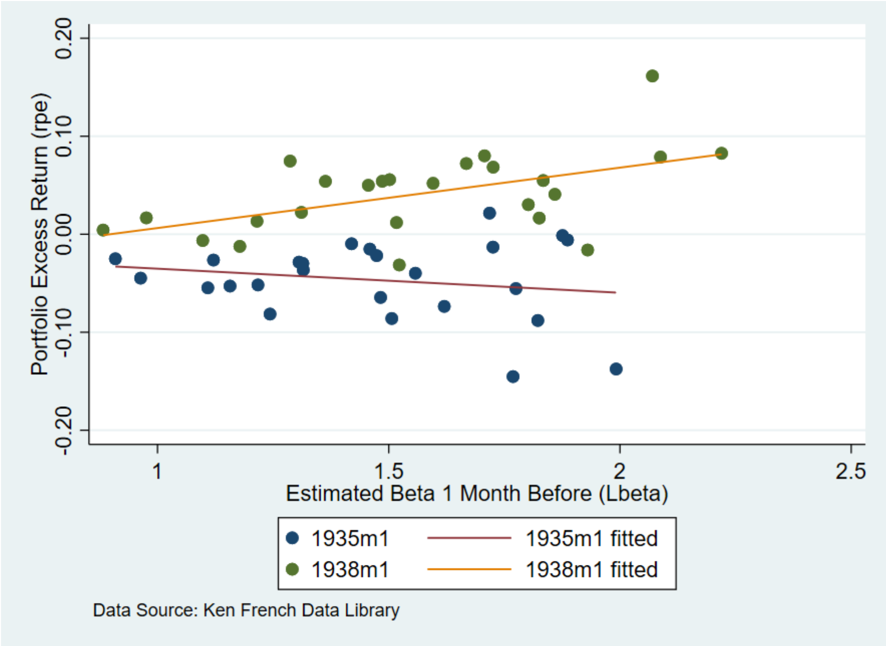
	port_num	t	mktrf	rpe	_rmse	_Nobs	_R2	_adjR2	_b_mktrf	_b_cons	_se_mk~f	_se_cons
46.	1	1938m9	0.01	-0.14	0.16	60	0.42	0.41	1.90	-0.00	0.33	0.01
47.	1	1938m10	0.08	0.12	0.16	60	0.42	0.41	1.91	-0.00	0.33	0.01
48.	1	1938m11	-0.02	-0.04	0.16	60	0.44	0.43	1.97	-0.00	0.38	0.01
49.	1	1938m12	0.04	0.04	0.15	60	0.47	0.46	1.97	0.00	0.38	0.02
50.	2	1934m12	0.00	-0.01	0.22	60	0.48	0.47	1.62	0.03	0.30	0.03

port_num	_b_mktrf	_se_mk~f	_R2	_rmse
1	1.89	0.32	0.48	0.21
2	1.82	0.30	0.52	0.19
3	1.87	0.20	0.72	0.12
4	1.75	0.17	0.70	0.12
5	1.95	0.22	0.65	0.15
6	1.39	0.15	0.66	0.11
7	1.51	0.13	0.77	0.09
8	1.55	0.16	0.79	0.09
9	1.62	0.15	0.80	0.09
10	1.74	0.15	0.75	0.10
11	1.25	0.08	0.81	0.06
12	1.20	0.06	0.90	0.04
13	1.36	0.09	0.92	0.04
14	1.47	0.09	0.86	0.06
15	1.79	0.10	0.89	0.07
16	0.96	0.05	0.92	0.03
17	1.13	0.07	0.94	0.03
18	1.32	0.06	0.93	0.04
19	1.50	0.09	0.91	0.05
20	1.98	0.15	0.86	0.08
21	0.90	0.03	0.97	0.02
22	1.09	0.03	0.98	0.02
23	1.26	0.06	0.95	0.03
24	1.50	0.06	0.92	0.05
25	1.69	0.18	0.78	0.10
Total	1.50	0.13	0.81	0.08

(_b_mkrtf就是beta)

为了截面回归更方便，直接将自变量取滞后项(beta滞后一个月)

在做截面回归之前，先看一下rpe和beta估计值的关系



该图画出了 1935m1 和 1938m1 两个时间节点上投资组合超额收益率 rpe 和上一月 估计值 **Lbeta** 的关系，横轴是 Lbeta，纵轴是 rpe。

接下来使用xtfmb进行第二阶段估计，也可以用asreg fmb，还可以用statsby

```
.   global regvar "rpe Lbeta"

.         *xtfmb
.         xtfmb $regvar

Fama-MacBeth (1973) Two-Step procedure              Number of obs      =      1200
                                                    Num. time periods =        48
                                                    F(   1,     47)     =      0.91
                                                    Prob > F           =    0.3437
                                                    avg. R-squared     =    0.2567

------------------------------------------------------------------------------
             |            Fama-MacBeth
         rpe |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       Lbeta |   .0122536   .0128116     0.96   0.344     -.01352     .0380272
       _cons |  -.0029518   .0131503    -0.22   0.823   -.0294067     .0235032
------------------------------------------------------------------------------
```

[基于机器学习方法的宏观因子模拟投资组合构建 - 知乎 (zhihu.com)]

[多因子回归检验中的 Newey-West 调整 - 知乎 (zhihu.com)]

[Fama-Macbeth 回归和Newey-West调整 - 云+社区 - 腾讯云 (tencent.com)]

七、一些可能的问题

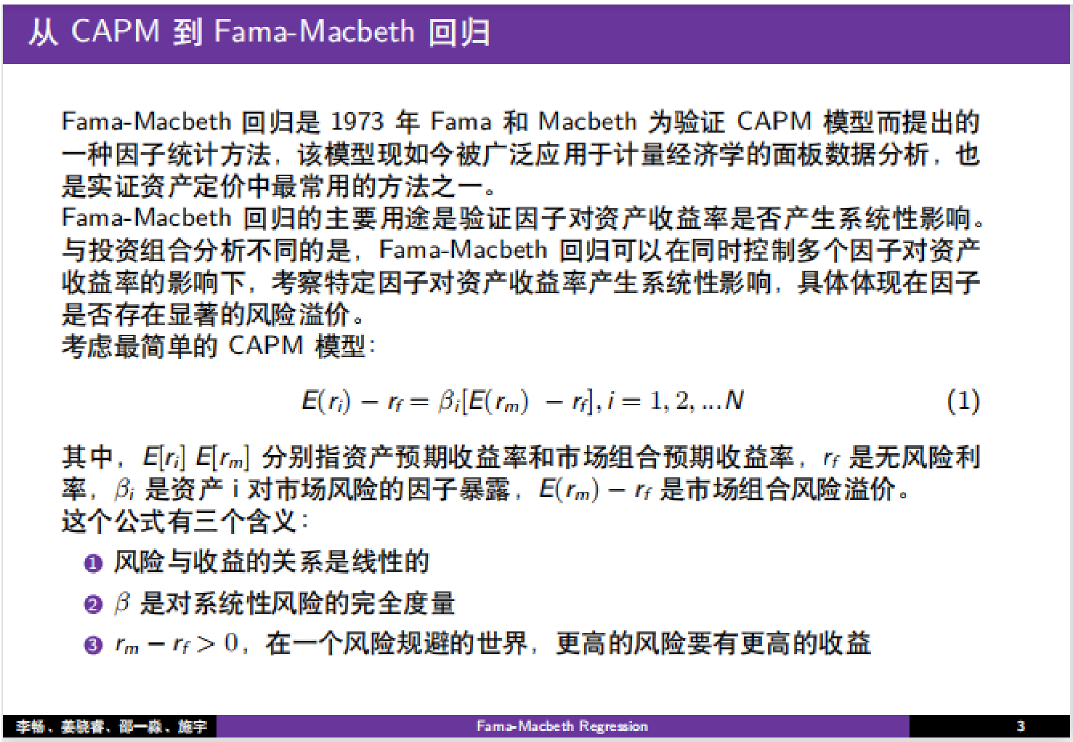
1、面板数据

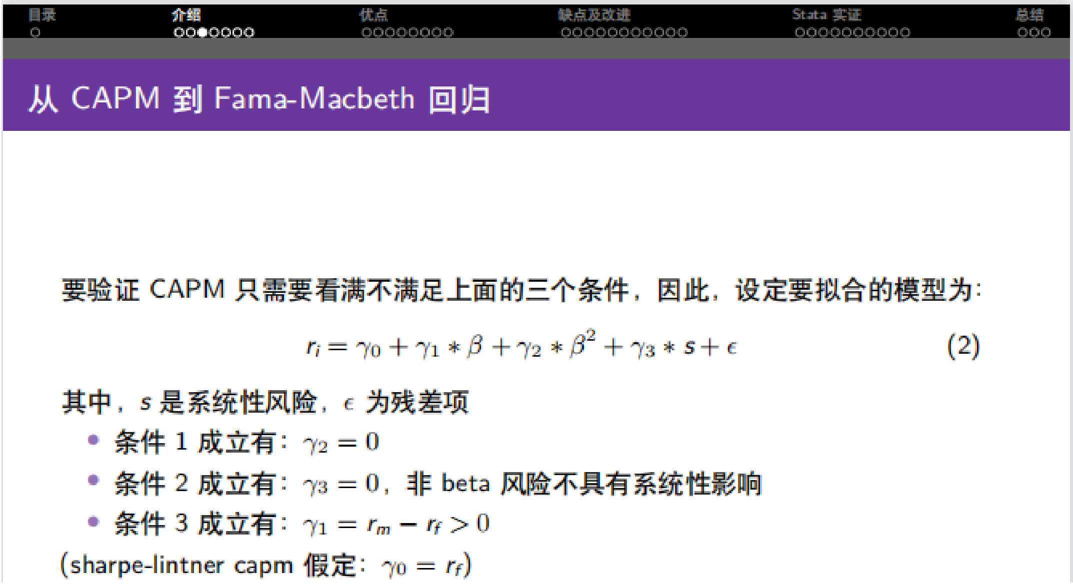
面板数据，即Panel Data，也叫"平行数据"，是指在时间序列上取多个截面，在这些截面上同时选取样本观测值所构成的样本数据。或者说他是一个m*n的数据矩阵，记载的是n个时间节点上，m个对象的某一数据指标。

[Fama-Macbeth小组第一次汇报v2.0 (maifile.cn)]

第一次汇报上半部分的问题的课后思考

1、关于公式2的理解





以下参考原文(Fama, E. F., & MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. *Journal of political economy*, *81*(3), 607-636.)：

$E(\hat{R}_i) = E(\hat{R}_0) + [E(\hat{R}_m) - E(\hat{R}_0)]\beta_i$　(*)

上式有三个可以验证的条件：

C1、风险与收益的关系是线性的

C2、β是对系统性风险的完全度量

C3、$r_m - r_f > 0$，在一个风险规避的世界，更高的风险要有更高的收益

为了验证条件CI-C3，我们必须确定一些有效的投资组合。这反过来又要求在投资者做出投资组合决策时，明确说明市场平衡的特征。

假设资本市场是完美的。此外，假设所有投资者都可以从无成本获得的信息中得出对任何资产或投资组合未来价值分布的相同和正确的评估——这种假设通常被称为“同质预期”。最后，假设允许卖空所有资产。然后Black（1972）表明，在市场均衡的情况下，所谓的市场投资组合，由权重定义

$$x_{im} = \frac{total \quad market \quad value \quad of \quad all \quad units \quad of \quad asset \quad i}{total \quad market \quad value \quad of \quad all \quad assets}$$

总是有效的。

由于市场组合包含所有正金额资产，市场组合是检验预期回报风险条件CI-C3的一个方便的参考点。同质期望假设意味着对回报分布的事前评估和事后回报分布之间的对应关系，这也是对这三个假设进行有意义的检验所必需的。

方程（ * ）是根据预期回报计算的。但其影响必须通过逐期证券和投资组合回报的数据进行检验。我们希望选择一个逐周期返回的模型，允许我们使用观察到的平均回报来测试预期回报条件C1-C3，但这个模型仍然尽可能普遍。我们建议采用以下对（ * ）进行随机推广的方法：

$$\bar{R_{it}} = \hat{\gamma_{0t}} + \hat{\gamma_{1t}}\beta_i + \hat{\gamma_{2t}}\beta_i^2 + \hat{\gamma_{3t}}s_i + \hat{\eta_{it}}$$

下标t代表时期t，所以$\hat{R_{it}}$是t-1期到t期的资产i的回报。上式允许$\hat{\gamma_{0t}}$和γ_{1t}在不同时期随机变化。C3假设是预期风险收益$\hat{\gamma_{1t}}$(在式(*)中是$[E(\hat{R_{mt}}) - E(\hat{R_{0t}})]$)是正的，换句话说就是 $E(\hat{\gamma_{1t}}) = E(\hat{R_{mt}}) - E(\hat{R_{0t}}) > 0$

上式中的β_i^2是为了验证线性性的。C1的假设是$E(\hat{\gamma_{2t}}) = 0$尽管$\hat{\gamma_{2t}}$被允许随着时期随机变化。s_i也是类似的，这意味着资产i不具有除了与β_i以外的系统性风险。C2假设是$E(\hat{\gamma_{3t}}) = 0$但是$\hat{(}\gamma_{3t})$也可以随着时间随机变化。

扰动项$\hat{\eta_{it}}$被认为是有均值0并且独立于其他任何变量。如果所有投资组合回报的分布是正态的，那么变量$\hat{\eta_{it}}, \hat{\gamma_{0t}}, \hat{\gamma_{1t}}, \hat{\gamma_{2t}}, \hat{\gamma_{3t}}$必须具有多元正态（或对称稳定）分布。

c1-c3是双参数模型所隐含的预期回报和风险的条件。但该模型，尤其是对完美市场的基本假设，意味着一个有效的资本市场，因为每个时间点的价格都完全反映了可用的信息。当然，“效率”这个词的使用不要与投资组合效率混淆。这个术语，如果有点不幸的话，也至少是标准的。

结合条件C1的市场效率要求对随机非线性系数$\hat{\gamma_{2t}}$的时间序列的审查不会导致$\hat{\gamma_{2t}}$的预期未来值的非零估计。在形式上，$\hat{\gamma_{2t}}$必须是一个fair game。在实际操作中，虽然非线性可以在事后观察到，因为$\hat{\gamma_{2t}}$是一个公平的博弈，但在（ * ）总结的双参数模型有效的假设下，投资者预先行动总是合适的。也就是说，在他的投资组合决策中，他总是假定证券的风险与其预期回报之间存在线性关系。同样，双参数模型中的市场效率要求非β风险系数$\hat{\gamma_{3t}}$和返回干扰的时间序列$\hat{\eta_{it}}$，是fair game。公平博弈假设也适用于 $\hat{\gamma_{1t}} - [E(\hat{R_{mt}}) - E(\hat{R_{0t}})]$的时间序列，即t期的风险溢价与期望值之间的差值。

在Fama(1970b)的术语中，这些都是关于一个由双参数模型产生预期回报的市场的资本市场效率的“弱形式”的命题。这些主张很弱，因为它们只关心价格是否完全反映了过去回报时间序列中的任何信息。”“强形式”的测试将涉及到对所有现有信息调整价格的速度。

然后我的理解（如有不对欢迎指正）：

1、β和γ是什么?

以CAPM为例，β是市场风险因子，β^2是非线性的市场风险，γ_2是回报率与非线性市场风险的关系。

这里比较容易产生误解的地方在于，（2）式中的β和普通CAPM中的β含义不同，普通CAPM中的β一般是指市场组合的风险暴露，而（2）式中的β却是指市场风险。

2、s是什么?

s是除了β以外的其他宏观风险。γ_3是他的风险暴露。

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: journal of the Econometric Society*, 817-838.

Newey, W. K., & West, K. D. (1987). Hypothesis testing with efficient method of moments estimation. *International Economic Review*, 777-787.

Newey, W. K., & West, K. D. (1986). A simple, positive semi-definite, heteroskedasticity and autocorrelationconsistent covariance matrix.

Shanken, J. (1992). On the estimation of beta-pricing models. *The review of financial studies*, *5*(1), 1-33.

2、异方差与自相关部分推导

考虑如下的线性模型

$$\begin{aligned} y &= X\beta + \epsilon \\ E[\epsilon|X] &= 0 \\ E[\epsilon\epsilon'|X] &= \sigma^2\Omega = \sum \end{aligned}$$

其中，y是T*1阶向量（T 代表时序的总期数）；**X** 是 T × K 阶矩矩阵（其中 K 是 regressors 的个数）；**ε** 是 T × 1 阶残差向量；**Σ**（T × T 阶）是残差的协方差矩阵。回归的目的是为了得到回归系数 **β**（K × 1 阶矩阵）并检验它们的显著性。

在球形扰动项的假设下，Ω是单位阵I

在广义线性回归中，残差常常表现出异方差和自相关两种特性。此时Ω显然不是单位阵

对于异方差（但仍然可以假设独立），通常有：

$$\sigma^2\Omega = \sigma^2\begin{bmatrix} w_{11} & 0 & \dots & 0 \\ 0 & w_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & w_{TT} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_T^2 \end{bmatrix}$$

对于自相关（但同方差），通常有

$$\sigma^2\Omega = \sigma^2\begin{bmatrix} 1 & \rho_1 & \dots & \rho_{T-1} \\ \rho_1 & 1 & \dots & \rho_{T-2} \\ \dots & \dots & \dots & \dots \\ \rho_{T-1} & \rho_{T-2} & \dots & 1 \end{bmatrix}$$

如果Ω已知，通常使用GLS。但当GLS未知时，OLS往往是首选。对上述线性模型进行OLS求解就可以得到β的OLS估计，记为:$\hat{\beta}$

$\hat{\beta} = (X'X)^{-1}X'Y = \beta + (X'X)^{-1}X'\epsilon$

对上式两边取期望，则当$E[\epsilon|X] = 0$的假设成立时易知$E[\hat{\beta}] = \beta$，进而推导出$\hat{\beta}$的协方差矩阵，记为V_{OLS}：

$$\begin{aligned} V_{OLS} &= E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'|X] \\ &= E[(X'X)^{-1}X'\epsilon\epsilon'X(X'X)^{-1}|X] \\ &= (X'X)^{-1}X'(\sigma^2\Omega)X(X'X)^{-1} \\ &= \frac{1}{T}(\frac{1}{T}X'X)^{-1}(\frac{1}{T}X'[\sigma^2\Omega]X)(\frac{1}{T}X'X)^{-1} \end{aligned}$$

当残差不存在异方差以及自相关性时，**Ω = I** 而上面的协方差矩阵也可以简化成我们最熟悉的经典 OLS 里面的形式，也就是各种 OLS 软件包给出的参数的标准误（协方差矩阵对角线元素的平方根）和 t-statistic 的结果。然而，当残差存在异方差或者自相关时，OLS 得到的 β 的方差的估计是不准确的，从而影响对 β 进行统计检验。

在 **Ω** 未知的情况下，需对 **V**_OLS 进行估计。上面的表达式可以看成是三个矩阵相乘的形式，其中第一个和第三个仅和 **X** 有关，因此**核心目标就是估计中间矩阵（middle matrix）**。为了方便讨论，令**Q** 代表中间的矩阵：

$$\begin{aligned} Q &= \frac{1}{T} X'[\sigma^2 \Omega] X \\ &= \frac{1}{T} \sum_{i=1}^{T} \sum_{j=1}^{T} \sigma_{ij} x_i x_j' \end{aligned}$$

其中$x_i = [x_{i1}, x_{i2}, \ldots, x_{iK}]'$，即X的第i行的转置。只要我们能找到矩阵Q的估计，就能进而求出$\hat{\beta}$的协方差矩阵V_{OLS}。

针对残差的假设不同，最常见的两种估计是 White 估计（仅假设异方差）以及 Newey and West 估计（考虑异方差及自相关）。

当残差仅有异方差但没有自相关时，我们需要的估计量 **Q** 简化为：

$Q = \frac{1}{T} \sum_{i=1}^{T} \sigma^2 x_i x_i'$

（因为i!=j时，协方差矩阵对应的格子是0）

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: journal of the Econometric Society*, 817-838.

White(1980)指出使用X及样本残差e可以求出Q的渐进估计（记为S_0):

$S_0 = \frac{1}{T} \sum_{i=1}^{T} e_i^2 x_i x_i'$

将上述Q的估计S_0代入到 V_OLS 的表达式中，可以得到$\hat{\beta}$的协方差矩阵的估计：

$$\begin{aligned} V_{OLS} &= \frac{1}{T} (\frac{1}{T} X'X)^{-1} (\frac{1}{T} \sum_{i=1}^{T} e_i^2 x_i x_i') (\frac{1}{T} X'X)^{-1} \\ &= T(X'X)^{-1} S_0 (X'X)^{-1} \end{aligned}$$

这意味着哪怕我们对异方差的取值或结构一无所知，我们仍然可以根据最小二乘的结果进行适当的推断。

在实际问题中，除了异方差外，仍需考虑残差的自相关性。为此，一个自然的想法是将上述 **Q** 的估计延伸到对角线之外的元素，即：

$S = \frac{1}{T} \sum_{i=1}^{T} \sum_{j=1}^{T} e_{ij} x_i x_j'$

但是这种方法有两个问题，使得他不是正确的：

1、表达式一共有T*T项求和，但是他的系数仅仅是1/T，所以S可能不完全收敛

2、即使S收敛，他也很可能不是正定的，从而使得最后估计的$\hat{\beta}$的协方差矩阵不是正定的，这显然有违常理。

于是，Newey and West (1987）给出了当残差同时存在异方差和自相关时，**Q** 的估计，记为 **S**：

$S = \frac{1}{T} \{ \sum_{t=1}^{T} e_t^2 x_t x_t' + \sum_{l=1}^{L} \sum_{t=l+1}^{T} w_l e_t e_{t-l} (x_t x_{t-l}' + x_{t-l} x_t') \}$

$where \quad w_l = 1 - \frac{l}{1+L}$

上式中，大括号中的第一项对应仅有异方差情况下的 S_0，而后面第二项则是针对自相关性的修正。其中 L 是计算自相关性影响的最大滞后阶数（Newey and West 1994 给出了自动计算 L 取值的自适应算法），w_l 是滞后期 l 的系数，其隐含的意思是自相关性的影响随着滞后期 l 的增大而减小。在实际计算时，考虑到自由度的问题，为了得到无偏估计可以将上式中大括号外面的 1/T 换成 1/(T - K)；大括号内部的求和项仍是 T 项及 L × T 项。

将S带入到V_{OLS}的表达式中，得到Newey West估计量：

$V_{OLS} = T(X'X)^{-1}S(X'X)^{-1}$

在时序OLS回归中，Newey West调整同时作用于多个回归量的回归系数，从而求出$\hat{\beta}$的协方差矩阵，常见于因子分析中的投资组合测试中，具体方法为：

1、使用目标因子投资组合的收益率序列和（多个）已有因子收益率在时序上OLS回归（同时带截距项，代表超额收益部分；假设已有因子 + 截距项一共 K 个 回归变量），得到残差；

2、使用截距项和已有因子收益率序列X和残差e，通过Newey-West 调整求出 V_OLS；

3、将 V_OLS 的对角线元素开平方，其平方根就是参数$\hat{\beta}$的标准误（一共 K 个，对应 K 个 regressors）；

4、使用$\hat{\beta}$的估计和 Newey-West 调整后的标准误计算出这些参数的 t-statistics，从而判断它们的显著性。

在因子分析中，Fama-MacBeth regression 是最常见的截面回归方法（Fama and MacBeth 1973）。在该回归中，每一期使用当期因子暴露和个股下一期的收益率进行截面回归，得到因子的收益率；在全部期进行截面回归后，便可得到每个因子收益率的时间序列。将因子收益率在时序上取均值就得到每个因子的预期收益率，而我们关心的是该因子预期收益率是否显著不为零。

对于任何因子，其收益率序列在时序上很可能存在异方差和自相关性，因此在计算其均值标准误的时候需要进行 Newey-West 调整。然而，这和上面的多因子时序回归很不相同。如何进行 Newey-West 调整呢？

对于单个因子的收益率序列，将其用 1 作为 regressor 回归得到残差 —— 这相当于用因子收益率减去它在时序上的均值。然后把这个残差和 X = 1 代入到 Newey-West 调整中即可。

在这个简化版的Newey West调整中，Q的估计S简化为：

$$S = \frac{1}{T}\sum_{t=1}^{T} e_t^2 + 2\sum_{l=1}^{L}\sum_{t=l+1}^{T} w_l e_t e_{t-l}$$
$$where \quad e_t = f_t - E_t[f_t], w_l = 1 - \frac{l}{1+L}$$

其中 f_t 代表被检验因子的收益率时间序列，$E_t[f_t]$ 是它在时序上的均值。由于我们仅仅有一个 regressor，因此上述S 其实是一个标量。将它代入到 V_OLS 的表达式中，在对其开方，就得到 $E_t[f_t]$ 的标准误：

$s.e.(E_t[f_t]) = \sqrt{S/T}$

对每个因子依次使用上述修正，获得其各自收益率均值的 standard error，然后就可以计算 t-statistic 以及 p-value 并检验它们的显著性。

3、遗漏变量偏差中的定价模型

假设$v_t = (v_{1t}, v_{2t})^T$是一个包含两个潜在相关因子的向量。假设他们都是去均值化的，他们可以被看作是 factor innovations（？）。原论文中主要考虑 nontradable factors，他们的均值和风险溢价没有直接相关性。这就是为什么我们直接根据因素创新来编写模型。当然，如果这些因素是 tradable，那么这些因素本身的平均值——减去无风险率——就是风险溢价，在这种情况下，我们在这里讨论的方法仍然作为风险溢价的替代估计器有效。

假设观察到无风险率，我们用超额回报来表示该模型：
$r_t = \beta\gamma + \beta v_t + u_t$
其中，u_t是特异性风险，$\beta = (\beta_1 : \beta_2)$是风险暴露矩阵，$\gamma = (\gamma_1 : \gamma_2)$是两个因子的风险溢价。我们关心的是估计第一个因子$v_{1t}$（记为$g_t$)的代理风险溢价，在上面的简单假设中他的风险溢价就是γ_1。