Fama-Macbeth Regression

第一次汇报

李畅、姜骁睿、邵一淼、施宇

南京大学 工程管理学院
金融计量学课程

目录

介绍

从 CAPM 到 Fama-Macbeth 回归

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应用于计量经济学的面板数据分析，也是实证资产定价中最常用的方法之一。

Fama-Macbeth 回归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。

考虑最简单的 CAPM 模型：

$$E(r_i) - r_f = \beta_i[E(r_m] - r_f], i = 1, 2, ...N \tag{1}$$

其中，$E[r_i]$ $E[r_m]$ 分别指资产预期收益率和市场组合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，$E(r_m) - r_f$ 是市场组合风险溢价。

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

从 CAPM 到 Fama-Macbeth 回归

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：

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

其中，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：$\gamma_0 = r_f$)

Fama-Macbeth 回归模型设定

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

$$E[R_i] = \beta_i' \lambda, i = 1, 2, ...N \tag{3}$$

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。
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。

Fama-Macbeth 回归模型设定

对于上述模型，Fama-Macbeth 方法分两步进行估计：

1. 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投资组合回报，那么可以通过时间序列回归来分析 $E[R_i]$ 和 β_i 在截面上的关系。
   令 f_t 为资产 i 在 t 期的收益，对各个资产 i 进行时间序列回归，估计 β_i：

$$R_i = \alpha_i + \beta_i' f_t + \epsilon_{it}, t = 1, 2, ..., N \tag{4}$$

   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长度为 $period$：

$$R_i = \alpha_i + \beta_{it}' f_k + \epsilon_{ik}, k = t - period, 2, ..., t \tag{5}$$

Fama-Macbeth 回归模型设定

2. 用第一步得到的 $\hat{\beta}_i$ 作为自变量，在各个时期 t 分别做截面回归：

$$R_{it} = \hat{\beta}_i{}' \lambda_t + \alpha_{it}, i = 1, 2, ...N \tag{6}$$

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 β_{it}.

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：

$$\hat{\lambda} = \frac{1}{T}\sum_{t=1}^{T}\hat{\lambda}_t \tag{7}$$

$$\hat{\alpha}_i = \frac{1}{T}\sum_{t=1}^{T}\hat{\alpha}_{it} \tag{8}$$

Fama-Macbeth 回归模型设定

$\hat{\lambda}$ 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：

$$\sigma^2(\hat{\lambda}) = \frac{1}{T}\sum_{t=1}^{T}(\hat{\lambda}_t - \hat{\lambda})^2 \tag{9}$$

$$\sigma^2(\hat{\alpha}_i) = \frac{1}{T}\sum_{t=1}^{T}(\hat{\alpha_{it}} - \hat{\alpha}_i)^2 \tag{10}$$

上述方法的巧妙之处在于，他把 T 期的回归结果当作 T 个独立的样本。参数的标准误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。
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。

优点

优点

Fama-Macbeth 的优点在于排除了随机扰动在截面上的相关性对标准误的影响。在面板数据中，假设有 T 期，每期为一个截面，每个截面中含有 N 个个体，数据展示如下:

$$\begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1N} \\ X_{21} & X_{22} & \cdots & X_{2N} \\ \vdots & & & \vdots \\ X_{T1} & X_{T2} & \cdots & X_{TN} \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_T \end{pmatrix} \tag{11}$$

其中 X_{it} 代表自变量，Y_t 一般代表当期的个股超额收益率。则面板数据的标准回归为:

$$Y_{it} = X_{it}\beta + \varepsilon_{it} \tag{12}$$

优点

如果 ε 是独立同分布，不包含时间序列相关和截面相关。则估计 $\hat{\beta}$ 如下：

$$\begin{aligned}\hat{\beta} &= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}Y_{it}}{\sum_{i=1}^{N}\sum_{i=1}^{T} X_{it}^2} \\ &= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}(X_{it}\beta+\varepsilon_{it})}{\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}^2} \\ &= \beta + \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}\varepsilon_{it}}{\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}^2}\end{aligned} \tag{13}$$

估计 $\hat{\beta}$ 的标准误如下：

$$\begin{aligned}Var[\hat{\beta}_{ols}] = E(\hat{\beta}_{ols}-\beta)^2 &= E([\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}\varepsilon_{it}]^2[\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}^2]^{-2}) \\ &= E([\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}^2\varepsilon_{it}^2][\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}^2]^{-2}) \\ &= NT\sigma_x^2\sigma_\varepsilon^2(NT\sigma_x^2)^{-2} = \frac{\sigma_\varepsilon^2}{NT\sigma_x^2}\end{aligned} \tag{14}$$

优点

存在截面相关时，参考 Peterson(2009)[1]，放松对 ϵ 的假设，假设 X 中包含一个不可观测的固定时间效应。残差项也包含一个个体的特定效应 γ_i 和一个异质性效应 η_{it}，所以数据的生成过程可以描述如下：

$$\begin{aligned} \varepsilon_{it} &= \gamma_t + \eta_{it} \\ X_{it} &= \mu_t + v_{it} \end{aligned} \tag{15}$$

其中 X 和 ε、γ 和 η、μ 和 v 都相互独立。
则，对截面相关的刻画可以表示如下：

$$\begin{cases} \mathrm{Corr}(X_{it}, X_{js}) & = 1 \quad \text{for } i = j, t = s \\ & = \rho_x = \frac{\sigma_\mu^2}{\sigma_x^2} \quad \text{for } i \neq j, t = s \\ & = 0 \quad \text{for } t \neq s \end{cases} \tag{16}$$

$$\begin{cases} \mathrm{Corr}(\varepsilon_{it}, \varepsilon_{js}) & = 1 \quad \text{for } i = j, t = s \\ & = \rho_\varepsilon = \frac{\sigma_\gamma^2}{\sigma_\varepsilon^2} \quad \text{for } i \neq j, t = s \\ & = 0 \quad \text{for } t \neq s \end{cases} \tag{17}$$

[1] Petersen, M. A. (2009). Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches.The Review of Financial Studies, Vol. 22(1), 435 –480.

优点

对于截面回归，分两阶段进行。第一阶段对每个个体分别进行时间序列回归：

$$Y_{it} = a_i + X_i f_t + \epsilon_{it} \qquad \text{for} t = 1, 2, ..., T \tag{18}$$

其中 X_i 为时序回归中的回归系数。
第二阶段，在 T 期进行截面回归：

$$E_T(Y_i) = X_i \beta + \epsilon_i \qquad \text{for} i = 1, 2, ..., N \tag{19}$$

由15式可得：

$$\begin{aligned} \varepsilon_{iT} &= \gamma_i + \eta_i T \\ X_{iT} &= \mu_i + v_{iT} \end{aligned} \tag{20}$$

所以，在截面回归中，$\hat{\beta}$ 估计如下：

$$\begin{aligned} \hat{\beta} = \frac{\sum_{i=1}^{N} X_i E_T(Y_i)}{\sum_{i=1}^{N} X_i^2} &= \frac{\sum_{i=1}^{N} X_i (X_i \beta + \varepsilon_i)}{\sum_{i=1}^{N} X_i^2} \\ &= \beta + \frac{\sum_{i=1}^{N} X_i \varepsilon_i}{\sum_{i=1}^{N} X_i^2} \end{aligned} \tag{21}$$

优点

$\hat{\beta}$ 的标准误如下：

$$
\begin{aligned}
Var[\beta_{ols}] &= E(\hat{\beta}_{ols} - \beta)^2 \\
&= E([\sum_{i=1}^{N} X_i\varepsilon_i]^2[\sum_{i=1}^{N} X_i^2]^{-2}) \\
&= E([\sum_{i=1}^{N} X_i^2\varepsilon_i^2 + \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} X_iX_j\epsilon_i\epsilon_j][\sum_{i=1}^{N} X_i^2]^{-2}) \\
&= \frac{N\sigma_x^2\sigma_\varepsilon^2 + N(N-1)\rho_x\sigma_x^2\rho_\epsilon\sigma_\epsilon^2}{(N\sigma_x^2)^2} \\
&= \frac{\sigma_\varepsilon^2}{N\sigma_x^2}(1+(N-1)\rho_x\rho_\epsilon)
\end{aligned} \tag{22}
$$

可以看出，截面的相关性会引起截面回归的标准误计算的误差。与14式相比，偏移量与 ρ_x、ρ_ϵ 有关。

优点

对于 Fama-Macbeth 回归，$\hat{\beta}$ 估计如下：

$$\begin{aligned}\hat{\beta}_{FM} &= \sum_{t=1}^{T}\frac{\hat{\beta}_t}{T} \\ &= \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\sum_{i=1}^{N}X_{it}Y_{it}}{\sum_{i=1}^{N}X_{it}^2}\right) \\ &= \beta + \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\sum_{i=1}^{N}X_{it}\varepsilon_{it}}{\sum_{i=1}^{N}X_{it}^2}\right)\end{aligned} \tag{23}$$

$\hat{\beta}$ 估计的标准误如下：

$$\begin{aligned}Var(\hat{\beta}_{FM}) &= \frac{1}{T^2}Var\left(\sum_{t=1}^{T}\hat{\beta}_t\right) \\ &= \frac{Var(\hat{\beta}_t)}{T} + \frac{2\sum_{t=1}^{T-1}\sum_{s=t}^{T}\mathrm{Cov}(\hat{\beta}_t, \hat{\beta}_s)}{T^2} \\ &= \frac{Var(\hat{\beta}_t)}{T} + \frac{T(T-1)}{T^2}\mathrm{Cov}(\hat{\beta}_t, \hat{\beta}_s)\end{aligned} \tag{24}$$

优点

其中，

$$\begin{aligned}\mathrm{Cov}(\hat{\beta}_t,\hat{\beta}_s) &= \mathrm{E}[(\sum_{i=1}^{N} X_{it}^2)^{-1}(\sum_{i=1}^{N} X_{it}\epsilon_{it})(\sum_{i=1}^{N} X_{is}\epsilon_{is})(\sum_{i=1}^{N} X_{is}^2)^{-1}] \\ &= (N\sigma_x^2)^{-2}\mathrm{E}[(\sum_{i=1}^{N} X_{it}\epsilon_{it})(\sum_{i=1}^{N} X_{is}\epsilon_{is})] \\ &= (N\sigma_x^2)^{-2} * 0 \\ &= 0\end{aligned} \tag{25}$$

则标准误如下，截面的相关性并不会引起截面回归的标准误计算的误差。

$$\begin{aligned}Var(\hat{\beta}_{FM}) = \frac{1}{T^2}Var(\sum_{t=1}^{T}\hat{\beta}_t) &= \frac{Var(\hat{\beta}_t)}{T} + \frac{2\sum_{t=1}^{T-1}\sum_{s=t}^{T}\mathrm{Cov}(\hat{\beta}_t,\hat{\beta}_s)}{T^2} \\ &= \frac{Var(\hat{\beta}_t)}{T} + \frac{T(T-1)}{T^2}\mathrm{Cov}(\hat{\beta}_t,\hat{\beta}_s) \\ &= \frac{1}{T}(\frac{\sigma_\epsilon^2}{N\sigma_x^2}) + \frac{T(T-1)}{T^2} * 0 = \frac{\sigma_\epsilon^2}{NT\sigma_x^2}\end{aligned} \tag{26}$$

缺点及改进

EIV 问题

EIV 问题 (errors-in-variables) 是指当统计模型中一个或多个变量并不是真实值，而是存在一个误差时，所产生的模型估计的偏差。
而在 Fama-macbeth 回归中，由于第一步通过时序回归得到的因子载荷是真实 β 的估计量，存在误差，因此在第二步中直接将估计值 $\hat{\beta}$ 作为解释变量就产生了 EIV 问题。
假设存在线性模型：

$$Y_i = \beta_0 + \beta_1 X_i + e_i \quad for\ i = 1 \cdots N \tag{27}$$

真实的 X 无法被观测，而只能观测到存在一定误差的 W，例如：

$$W_i = X_i + u_i \tag{28}$$

其中 $E(u_i|X_i) = 0, Var(u_i|x_i) = \sigma_u^2, \mathrm{Cov}(u_i, X_i) = 0$。

EIV 问题

因此，线性模型 (27) 可以改写为：

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 (W_i - u_i) + e_i \\ &= \beta_0 + \beta_1 W_i + (e_i - \beta_1 u_i) \end{aligned} \tag{29}$$

复合误差项为 $e_i - \beta_1 u_i$，如果测量误差 u_i 与模型误差 e_i 无关，即 $E(e_i u_i) = 0$，我们可以得出：

$$\mathrm{Cov}(W_i, (e_i - \beta_1 u_i)) = -\beta_1 \sigma_u^2 \tag{30}$$

模型扰动项与解释变量相关，这违反了线性回归模型的基本假设，因此导致估计量有偏且不一致。实际上，因为模型的测量误差是不可观测的，所以在回归估计中忽略误差项 $-\beta_1 u_i$ 会导致估计误差。
当样本量 N 趋于无穷时，我们可以得出：

$$plim\ \hat{\beta}_1 = \beta_1 + \frac{\mathrm{Cov}(X, u)}{Var(X)} = \beta_1 \lambda \tag{31}$$

其中 $\lambda = \sigma_X^2 / \sigma_X^2 + \sigma_u^2$。因此，当存在 EIV 问题时，估计量是有偏且不一致的。

EIV 问题

解决 EIV 问题主要有以下几种方法：

1. Black,Jensen,and Scholes(1972) 以及 Fama and Macbeth(1973) 指出：当使用投资组合时，个股 β 的估计误差会相互抵消，因此对投资组合的 β 估计会更准确，从而在一定程度上降低 EIV 的影响。
   因此 Fama and Macbeth(1972) 在论文中没有使用个股的收益率，而是将个股按照历史 β 的大小构成了不同的投资组合，然后使用投资组合的收益率检验 CAPM。

EIV 问题

② 为了减少 EIV 问题，Jegadeesh et al.(2019)[2]在 Fama-macbeth 回归的第二步中引入了工具变量 (Instrumental Variables)，因此得到的 λ 的估计量为：

$$\lambda_{IV,t} = \frac{B_{IV} r_t}{B_{IV} B_{EV}} \tag{32}$$

上式中，B_{IV} 和 B_{EV} 分别为工具变量和解释变量，B_{EV} 是对 β 的估计量，B_{IV} 是 B_{EV} 的工具变量。

具体来说，在第一步时序回归时按照以下方式进行：

1. 如果当前月是偶数月，则使用过去时间窗口内所有的偶数月个股的收益率进行回归，得到 B_{EV}；使用过去时间窗口内所有的奇数月个股的收益率进行回归，得到 B_{IV}。
2. 如果当前月是奇数月，则使用过去时间窗口内所有的奇数月个股的收益率进行回归，得到 B_{EV}；使用过去时间窗口内所有的偶数月个股的收益率进行回归，得到 B_{IV}。

Jegadeesh et al.(2019) 指出，上述估计可以得到风险溢价的无偏估计。

[2]Jegadeesh N, Noh J, Pukthuanthong K, et al. Empirical tests of asset pricing models with individual assets: Resolving the errors-in-variables bias in risk premium estimation[J]. Journal of Financial Economics, 2019, 133(2): 273-298.

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。

先简单说明 Newey West 的原理：

考虑一个线性模型

$$y = X\beta + \epsilon \tag{33}$$

$$E[\epsilon|X] = 0 \tag{34}$$

$$E[\epsilon\epsilon'] = \sigma^2\Omega \tag{35}$$

$$\hat{\beta} = (X'X)^{-1}X'y = \beta + (X'X)^{-1}X'\epsilon \tag{36}$$

$$\begin{aligned} Var[\beta] &= E[(\beta - \hat{\beta})(\beta - \hat{\beta})'|X] \\ &= \frac{1}{T}(\frac{1}{T}X'X)^{-1}(\frac{1}{T}X'\sigma^2\Omega X)(\frac{1}{T}(X'X)^{-1}) \end{aligned} \tag{37}$$

非球型扰动项及 Newey-West 调整

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:

$$Q = \frac{1}{T} X' \sigma^2 \tag{38}$$

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:

$$S = \frac{1}{T}(\sum_{i=1}^{T} e_t^2 x_T x_t' + \sum_{l=1}^{L} \sum_{t=l+1}^{T} w_l e_t e_{t-1}(x_t x_{t-l}' + x_{t-l} x_t')) \tag{39}$$

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

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。

非球型扰动项及 Newey-West 调整

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:

$$Var[\hat{\beta}_{NW}] = T(X'X)^{-1}S(X'X)^{-1} \tag{40}$$

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。
在这个简化版的 Newey-West 调整中，Q 的估计 S 简化为:

$$S = \frac{1}{T}\sum_{t=1}^{T} e_t^2 + 2\sum_{t=1}^{L}\sum_{t=l+1}^{T} w_l e_t e_{t-1} \tag{41}$$

$where \quad e_t = f_t - E_t[f_t] \quad w_l = 1 - \frac{l}{1+L}$
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回归量，因此上述 S 其实是一个标量。将它代入到 V_OLS $E_t[f_t]$ 的标准误:

$$s.e.(E_t[f_t]) = \sqrt{S/T} \tag{42}$$

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。

遗漏变量与 mimicking-protfolio

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。
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：

$$r_t = \beta\gamma + \beta_{vt} + u_t \tag{43}$$

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_i（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。

遗漏变量偏差

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 $v1t$ 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。

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超额收益写成：

$$\gamma_g^{MP} = (w^g)^T E(\tilde{r}_t) \tag{44}$$

遗漏变量偏差

由于测试资产 r_t 遵循与 r_t 相同的定价模型。我们可以写为:

$$\tilde{r}_t = \tilde{\beta}\gamma + \tilde{\beta}v_t + \tilde{u}_t \tag{45}$$

同样的，我们可以写出 mimicking-portfolio 的第一个因子的风险溢价:

$$\gamma_g^{MP} = \tilde{\beta}\sum^{v}\tilde{\beta}^T + \sum^{\tilde{u}})^{(} - 1)(\tilde{\beta}\sum^{v} e_1)^{T} \tilde{\beta}\gamma \tag{46}$$

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 $\sum^{\tilde{u}} \approx 0$
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。

Stata 实证

Stata 实证

下面我们用投资组合数据估计一个简单的 CAPM 模型，来说明 FM 两阶段回归的基本步骤。其中，数据来源于「Ken French 主页」的研究数据，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，和「Fama/French 3 Factors」中的无风险收益、市场超额收益数据。主要变量如下：

port$_n$$um$	投资组合序号
t	时期（月度）
mktrf	市场组合超额收益率
rpe	25 个投资组合超额收益率
beta	第一阶段回归估计量
Lbeta:	beta 的一阶滞后项（前一月 beta ）

Stata 实证

经过初步处理后数据形式如下:

	port_num	t	mktrf	rpe
1.	1	2020m9	-0.036	-0.002
2.	1	2020m10	-0.021	-0.032
3.	2	1926m7	0.030	-0.033
4.	2	1926m8	0.026	-0.056
5.	2	1926m9	0.004	-0.018

Stata 实证

主要描述性统计量如下:

```
. tabstat mktrf rpe if (t>=ym(1930,1) & t<=ym(1938,12)), ///
          stats(mean sd min max p1 p25 p50 p75 p99)      ///
          column(s) long  format(%4.3f)

 variable |  mean     sd     min     max      p1     p25    p50    p75    p99
----------+------------------------------------------------------------------
    mktrf | 0.005  0.107  -0.291  0.389  -0.238  -0.062  0.004  0.057  0.371
      rpe | 0.013  0.179  -0.532  1.546  -0.319  -0.088  0.002  0.078  0.702
-----------------------------------------------------------------------------
```

结果显示，收益率方差较大，且存在异常值。由于缩尾处理后对整体结果没有太大影响，因此后文使用了未经缩尾的原始数据。

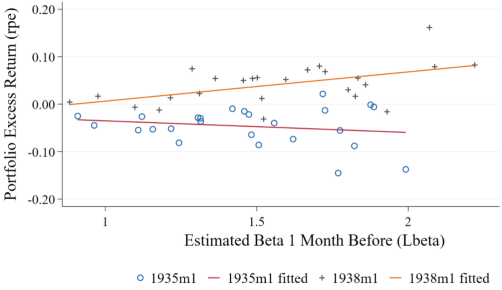
Stata 实证

第一阶段：时序回归使用 bysort 和 asreg 实现对每个投资组合的滚动窗口回归：

```
     +------------------------------------------------------------------------------------------
-------------+
     | port_num        t   mktrf     rpe   _rmse   _Nobs    _R2   _adjR2   _b_mktrf   _b_cons   _se_mk~
f   _se_cons |
     |------------------------------------------------------------------------------------------
-------------|
  1. |        1   1938m9    0.01   -0.14    0.16      60   0.42     0.41       1.90     -0.00       0.2
9       0.02 |
  2. |        1  1938m10    0.08    0.12    0.16      60   0.42     0.41       1.91     -0.00       0.2
9       0.02 |
  3. |        1  1938m11   -0.02   -0.04    0.16      60   0.44     0.43       1.97     -0.00       0.2
9       0.02 |
  4. |        1  1938m12    0.04    0.04    0.15      60   0.47     0.46       1.97      0.00       0.2
8       0.02 |
  5. |        2  1934m12    0.00   -0.01    0.22      60   0.48     0.47       1.62      0.03       0.2
2       0.03 |
     +------------------------------------------------------------------------------------------
-------------+
```

Stata 实证

第二阶段：截面回归
在正式做回归之前，大致查看下超额收益和市场 β 估计值的关系：



Data Source: Ken French Data Library

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 *beta* 估计值 Lbeta 的关系，横轴是 Lbeta，纵轴是 rpe。通过 FM 第二阶段回归，市场组合风险溢价 λ 就是在 1935m1 到 1938m12 间 48 个月每月进行截面回归得到的斜率的均值。

Stata 实证

下面使用 xtfmb 进行第二阶段估计：

```
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 |     0.012       0.013     0.96   0.344        -0.014       0.038
 _cons |    -0.003       0.013    -0.22   0.823        -0.029       0.024
------------------------------------------------------------------------------
```

Stata 实证

verbose 选项给出每一时期截面回归的结果，lag(4) 选项汇报考虑四阶自相关的 Newey-West 修正标准误。

```
Fama-MacBeth (1973) Two-Step procedure    Number of obs     =     1200
                                          Num. time periods =       48
                                          F(  1,     47)    =     0.75
                                          Prob > F          =   0.3899
Newey-West corrected SE (lag length: 4)   avg. R-squared    =   0.2567
------------------------------------------------------------------------------
       |          Fama-MacBeth
   rpe |   Coef.   Std. Err.       t    P>|t|      [95% Conf. Interval]
-------+----------------------------------------------------------------------
 Lbeta |   0.012       0.014     0.87   0.390       -0.016       0.041
 _cons |  -0.003       0.012    -0.25   0.806       -0.027       0.021
------------------------------------------------------------------------------

Coefficient estimates and R-squared of the cross-sectional regressions in step 1

  +-----------------------------------------+
  |    t       Lbeta    constant         R2 |
  |-----------------------------------------|
  | -300    -.024564    -.010588   .0349527 |
  | -299    -.163997    .1709705   .7057621 |
  | -298   -.0481288   -.0088595   .1374627 |
                 (截取部分结果)
  | -255    .1308773   -.0390214   .2675762 |
  | -254   -.0661581     .058338   .3761048 |
  | -253   -.0134165    .0538785   .0088549 |
  |-----------------------------------------|
  | Mean    .0122536   -.0029518   .2567224 |
  |    N          48          48         48 |
  +-----------------------------------------+
```

总结

总结

本次汇报我们主要讨论了以下问题:

1. Fama-Macbeth 回归是什么？Fama-Macbeth 回归是最初是为了验证 capm 而提出的两步截面回归检验方法，在计量金融学中被广泛用于面板数据的分析。
2. Fama-Macbeth 有什么优点？Fama-Macbeth 排除了残差在截面上的相关性对标准误的影响.
3. Fama-Macbeth 有哪些不足和改进？Fama-Macbeth 回归对于残差在时序上的相关性无能为力，同时存在 EIV 和遗漏变量问题。
4. 如何使用 Stata 完成一次 Fama-Macbeth？asreg 和 xtfmb 命令。