SMB HML因子复现步骤

SMB需要 firm’s market value of equity, HML 需要firm’s book value of equity/ firm’s market value of equity。BE需要用到年度财务报表数据，年度财报数据涉及t-1的问题，今年内都用去年年末的数据。

Jdate为monthly数据原始月末日期

SHROUT is the number of publicly held shares, recorded in thousand

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| PSTK | Balance Sheet | Annual | Number | Millions |
| REVT | Income Statement | Annual | Number | Millions |
| COGS | Income Statement | Annual | Number | Millions |
| XSGA | Income Statement | Annual | Number | Millions |
| AT | Balance Sheet | Annual | Number | Millions |

**数据准备阶段**

首先接入Compustat库，获取book equity相关的三个数据。

**Book Equity** is defined as the Compustat book value of **stockholders' equity** plus balance sheet **deferred taxes and investment tax credit** (if available) minus **book value of preferred stock**.

1. 处理三者的缺失值：前两个用0填充，ps 用the redemption or liquidation or par value of preferred stock (in that order)
2. 生成book equity并只保留正值，负的填充nan
3. Count：a variable that counts number of annual records in Compustat files.
4. 注意，在5-factor model的处理中，我们还计算得到了op, agr, total asset,需要一并保存
5. 导出必要数据（编号、日期、年份、be、count） 存在comp中

从 Fama-French 因子数据中提取月度无风险利率

**正式构建阶段**

接下来接入CRSP 常规化的处理股票monthly data，此块儿还并入了退市数据的处理，得到retadj。

1. 以月末时间对齐、合并退市数据，0填充缺失值，得到退市回报；在个股层面计算每一个股票的ME
2. 删除不必要数据（如合并来的，ME计算时用到的） 留下来的数据排序好存在crsp中。
3. 得到累积ME与最大ME，crsp1只保留有MaxME的主stock,crsp1与SumMe合并、排序、去重得到crsp2；此时crsp2只保留了主 stock的permno并且ME列为permco的累积总ME
4. 注意，在5-factor model的处理中，在这一步我们额外在crsp2中储存了市场组合回报率RMkt （缺失数据用插值法填充）

形成期数据处理（crsp2基础上）

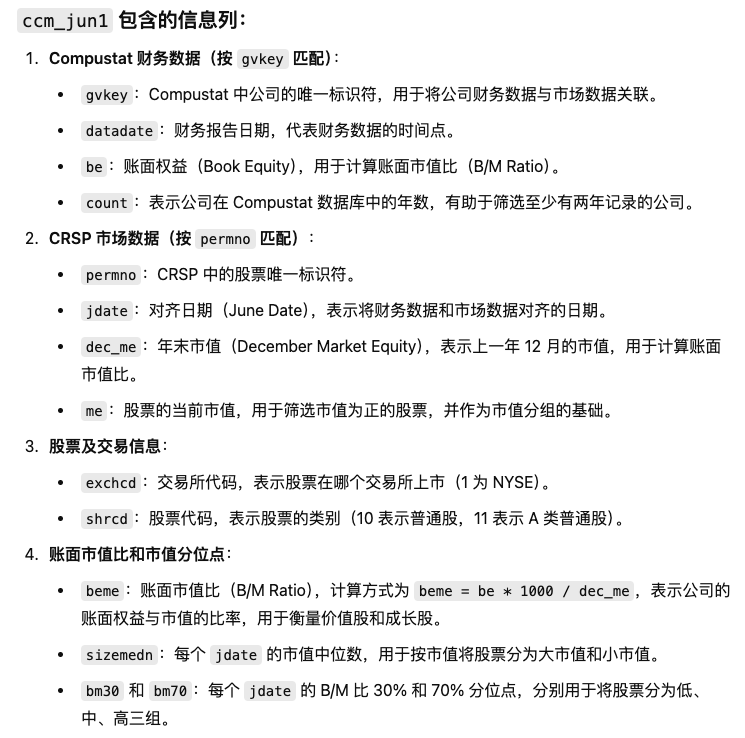
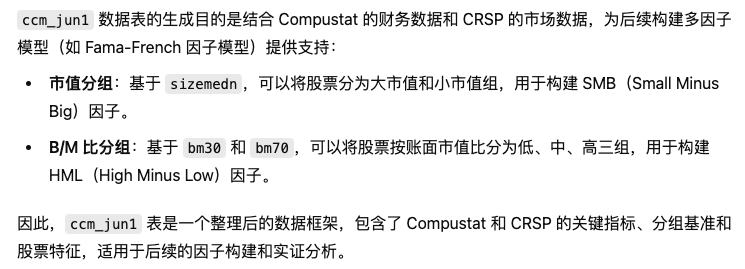
1. 生成年度与月度数据，单独储存12月的数据进decme；
2. 在crsp2中找到对应的形成期开始时间（jdate-6个月）的年月日？
3. 计算cumretx：根据permno与形成期年份分组，计算单只个股每年7月到次年6月的累积收益，得到上一个周期内的累积收益lag cumret （可作为后续分组依据）；同样得到lag market cap，但要注意个股的第一条记录无滞后值，要用特殊办法填充
4. 单独储存ffmonth为1月的数据（也就是真实时间7月的数据）进mebase （作为分组时cap的基准）（permno','ffyear', 'lme），储存的lme代表是上一个月月底的（也就是12月底）me，也叫mebase
5. 合并数据：将crsp2与mebase合并，得到每一个permno在每一个ffyear的基准市值数据；利用1月的lme或是其他月的mebase\*lcumretx（因为其他月的SMB为mebase乘上累积回报）生成加权市值weight；保存进crsp3
6. 将decme年份+1 因为是前一年12月；保留必要数据
7. 获取6月底数据准备构建portfolio：crsp3\_jun 保留crsp3六月数据，与decme合并；根据permno与jdate删除重复值，保证每只个股在每个月只有一个数据。

从 Compustat 和 CRSP 数据库中建立公司和股票的对应关系，

1. 获取linking table CCM，用今日日期填充缺失日期
2. 根据gvkey合并 Compustat 和 CCM 表得到ccm1，包含了 Compustat 中公司和 CRSP 中股票的关联信息
3. 处理日期，创建年末和跳转日期jdate。因为ccm1包含公司数据，应在下一财务年内使用。
4. 在ccm2中过滤有效的公司-股票链接，即确保公司财务数据仅在链接有效期间与 CRSP 的股票数据匹配，保留有效数据。
5. 根据jdate和permno link comp and crsp\_jun获得ccm\_jun
6. 在ccm\_jun中计算B/M ratio,具体数据为上一财务年Comp中的be/六月时记录的去年12月底的dec\_me

利用NYSE 股票构建市值（Size）和账面市值比（B/M）分组的bp

1. 从ccm\_jun筛选符合条件的 NYSE 股票存入nyse
2. 计算市值中位数分点sizemedn，方便后续分为大公司和小公司,存入nyse\_sz
3. 计算 B/M 比的 30% 和 70% 分位点,存入nyse\_bm
4. 合并市值和 B/M 分位点进nyse\_breaks，将其与ccm\_jun按日期jdate合并得到ccm\_jun1



为 Fama-French 因子模型构建分组

1. 用两个函数定义Size和B/M分组，sz\_bucket和bm\_bucket
2. 创建posbm为数据有效标签：beme > 0 且 me > 0 且至少两年记录的股票为1
3. 在ccm\_jun1中，对posbm=1的应用分组函数并创建标识变量，szport, bmport
4. 创建nonmissport: 确保 B/M 组合信息完整, 只包含那些成功分配了 B/M 比组合的股票,减少噪音影响；写的时候可以改进一下，变成两个组合都要完整，把szport为空的也踢出去

按每年6月份分组的组合信息（szport 和 bmport）与股票的月度记录 合并，构建出一个筛选后的数据集 ccm4用于后续多因子分析

1. 提取 6 月份的组合分组信息june并新增ffyear标记形成年份
2. 将 crsp3 中的月度市场数据与 6 月的分组数据 june 合并，按 permno 和 ffyear 匹配；筛选符合条件的记录得到ccm4

ccm4 是一个包含所有符合条件的股票的 月度数据，并附加了每年 6 月的 市值组合（szport） 和 账面市值比组合（bmport）

基于 Fama-French 因子模型构建 SMB（小市值减大市值）和 HML（高账面市值比减低账面市值比）因子

1. wavg 函数用于按指定权重（wt 列）计算某个分组（如 szport、bmport）的加权平均收益率（retadj），加权方式为市值权重。
2. 利用ccm4根据jdate 和两个标签分组，生成加权收益率存入vwret，计算公司数量存入vwret\_n，为两表都生成sbport 列将市值和账面市值比组合标签合并，（如 SH 表示小市值高 B/M 比）。
3. 转置 vwret 和 vwret\_n，生成的 ff\_factors 和 ff\_nfirms 表分别以 jdate 为索引，将各组合的收益率和公司数量按列排列，便于计算 SMB 和 HML 因子。
4. 计算 HML 和 SMB 因子：（最后要重命名日期列）

* HML 因子：

WH：高 B/M 比组合的平均收益率，WH = (BH + SH) / 2。

WL：低 B/M 比组合的平均收益率，WL = (BL + SL) / 2。

WHML：HML 因子，表示高 B/M 组合与低 B/M 组合的收益差异。

* SMB 因子：

WB：大市值组合的平均收益率，WB = (BL + BM + BH) / 3。

WS：小市值组合的平均收益率，WS = (SL + SM + SH) / 3。

WSMB：SMB 因子，表示小市值组合与大市值组合的收益差异。

1. **HML 因子公司数量**：计算高 B/M 和低 B/M 组合的公司数量；**SMB 因子公司数量**：计算小市值和大市值组合的公司数量。

ff\_factors：包含 SMB 和 HML 因子的收益率时间序列。

ff\_nfirms：记录每个月份在 SMB 和 HML 因子分组中的公司数量统计。

Compare with ff

1. 从 Fama-French 数据库提取 SMB 和 HML 因子，对齐月末时间
2. 将 Fama-French 因子数据 \_ff 与自定义因子 ff\_factors 按日期合并，生成 \_ffcomp 表。smb 和 hml表示官方数据；WSMB 和 WHML表示自定义计算的 SMB 和 HML 因子；筛选出 1970 年以后的数据 \_ffcomp70，用于对比。
3. 使用 pearsonr 计算 1970 年后的 smb 和 WSMB、hml 和 WHML 因子间的皮尔森相关系数
4. 检查数据类型和结构，将 date 转换为日期格式并设置为索引，便于后续可视化。