1. 摘要:

在本篇报告中我将表明我的六因子模型的复现过程，以及模型各个因子在不同行业中的有效性。在模型中我一共加入六个因子，其中包括来自 Fama French 三因子模型的市场因子（MKT）、规模因子（SMB）、价值因子（HML）；以及动向因子（UMD）、反转因子（fx_HML）、净资产收益率因子（ROE）。本次模型的应用场景为 A 股市场，使用数据为月度数据。模型复现过程共分为以下几步：数据处理、因子计算、建立模型、行业比较。并且，为检验因子数据的准确性，我将自己算出的因子值（复制版本）与 BetaPlus 所提供的数据（经典版本）进行对比。结果发现，复制版本的因子值与经典算法的因子值高度相关，数值范围没有明显差异，且以两种数据所建立的模型效果相似，说明了复制版本数据具有一定准确性。在行业比较方面，我们发现来自 Fama French 的三因子对于不同行业均有非常显著的解释作用，而在此基础上加入的动向因子、反转因子、净资产收益率因子则给不同行业的解释度带来小幅提升。

2. 数据处理:

数据合并

首先，按照模型的因子以及因变量（股票超额收益率）来确定我们所需提取的数据。计算每个因子的所需数据如下表 2.1 所示：

因子计算所需数据

市场因子（MKT)	市场收益率、无风险收益率
规模因子（SMB)	股票市值、股票复权收盘价
价值因子（HML)	账面市值比（Book-to-Market ratio）、股票复权收盘价
动量因子（UMD)	股票复权收盘价
反转因子（FX_HML)	股票复权收盘价
净资产收益率因子（ROE)	净资产收益率、股票复权收盘价
股票超额收益率（因变量)	股票复权收盘价、无风险收益率

表 2.1 因子计算所需数据

由此所提取到的原始数据为日度数据，但我们打算将其转化为月度数据进行处理。在转化的过程中存在数据对齐以及避免使用未来数据等问题，具体解决方式归为一下几点：

1. 日度转化月度数据：选取每月最后一日的数据作为月度数据。例如，股票每月价格均为月末交易日当天所对应的复权收盘价。
2. 数据对齐：股市数据中的月末日期不等同每月最后一天的日期，而是每月最后一个交易日的日期。而在其他数据（如市场收益率、无风险收益率等）中，月末日期就是每月最后一天的日

期。所以在数据合并时，先将股市数据的月末日期转化为每月最后一天的日期，再按日期对齐与其他数据合并。

3. 避免使用未来数据：为确保不会使用未来数据，所有当月更新的数据均在下个月进行使用。例如，公司 4 月 1 号更新的财报数据只能在 5 月开始使用。

将以上数据进行转化之后，我们就可以按照经典算法尽可能还原因子数据。

去除极值

模型的因子数据很大程度上收到了极值的影响。这是因为因子收益率为不同股票组合的平均收益率之差，而每个月中都会有一些股票的涨跌幅较为极端，从而影响了整个组合的平均收益率。所以，在实践中我尽可能排除这些股票带来的影响。

在经典算法中，数据剔除了新股(上市不满 12 个 月)、风险警示股、待退市股和净资产为负股、停牌股票和一字涨跌停股票。我的去除极值的方法相比之下更加粗糙，剔除了上市不满 365 天的新股，距离退市不到 180 天的股票，以及涨跌幅远于平均值 3 个标准差的股票。这种方法会一定程度上影响因子值大小，因为我们将所有涨跌幅极高的股票一刀切掉，而不是像经典算法中通过股票本身的特征来进行剔除。

3. 因子计算

市场收益率&无风险收益率——求市场因子

我们使用能代表 A 股市场整体表现的中证全指来计算市场收益率。收益率的计算方式如下：

$$\text{本月市场收益率} = \frac{\text{本月月末收盘价} - \text{上月月末收盘价}}{\text{本月月末收盘价}}$$

无风险收益率使用的是 BetaPlus 提供的数据，具体来源如下图所示。市场收益率减去无风险收益率即为市场超额收益率，也就是市场因子的数据。

表 1：无风险收益率数据

时间区间	数据来源
2002/08/06 之前	三个月期定期银行存款利率
2002/08/07 至 2006/10/07	三个月期中央银行票据的票面利率
2006/10/08 至今	上海银行间三个月同业拆放利率

股票收益率——求因变量

个股月度收益率是通过月末复权收盘价求得的。模型的因变量为股票超额收益率，即股票收益率减去无风险收益率。

$$\text{股票月度收益率} = \frac{\text{本月月末复权收盘价} - \text{上月月末复权收盘价}}{\text{本月月末复权收盘价}}$$

股票市值& Book-to-Market ratio——求 SMB、HML 因子

具体步骤如下：

1. 计算 SMB、HML 的方法是将取每年四月底的股票市值，根据中位数分为“大市值（B）”和“小市值（S）”两组；
2. 在每年四月底取得 Book-to-Market ratio（BM）数值，再根据第 30 和第 70 百分位分为“低（L）”、”中（M）“、”高（H）“三组。
3. 将两种分类维度取交集来看，我们每年四月底都将所有股票重新排序，并分为“大市值低 BM（BL）”、“大市值中 BM（BM）”、“大市值高 BM（BH）”……共六组。
4. 根据这六组股票的平均收益率(R_{BL}、R_{BM}、$R_{BH……}$) 我们可以求得 SMB 和 HML 因子：

$$SMB = \frac{1}{3}(R_{SL} + R_{SM} + R_{SH}) - \frac{1}{3}(R_{BL} + R_{BM} + R_{BH})$$
$$HML = \frac{1}{2}(R_{HS} + R_{HB}) - \frac{1}{2}(R_{LS} + R_{LB})$$

需要注意的是 1）在计算 BM 时采用的市值数据（分母）为月末数据；而净值数据（分子）从财务报表获得，为每年第一季度公开的数据。2）每年四月份的数据（4.1 ~ 4.30）依然对应去年的数值，这样避免了使用未来数据。3）在计算每组股票月度平均收益率时，我首先求出每支股票的月度收益率，再按市值进行加权平均。

动量和反转因子

传统方法对于某一个月动量的计算需要看本月前 12 个月至前 1 个月的累计涨跌幅变化（例如求今年 2 月份的动量需计算去年 2 月至今年 1 月的累计涨跌幅），根据此期间累计涨跌幅的大小分为高中低三组（30%高 - 40%中 - 30%低）。再将高动量组的当月平均收益率与低动量组的当月平均收益率相减，其差值即为动量因子(momentum High-Minus-Low)。需注意的是计算当月平均收益率时采取的是等权平均，这与计算 SMB、HML 时的方式（按市值加权取平均）不同。

由于我观察到传统方法所计算出的动量不具有很明显的动量效应（动量效应是指前段时间收益率较高的股票，在接下来的表现仍会超过早期收益率低的股票，即，前段时间强势的股票，在未来一段时间

继续保持强势），所以我在传统方法基础上有所改动。我将过去 12 个月至 4 个月期间的涨幅算作动量效应更为明显的区间；将过去 3 个月至 1 个月期间的涨幅算作反转效应更为明显的区间（反转效应，它是指前段时间收益率低的股票，在其后的一段时间内有强烈的趋势经理相当大的逆转，即，前段时间弱的股票，未来一段时间会变强）。再根据两个区间的累计涨跌幅按照上一段所述的计算过程进行分类和计算，分别获得动量因子和反转因子。

ROE 因子

ROE 因子的计算过程按照 BetaPlus 提供的简易算法，将每个季度更新的 ROE 数据进行整理。挑取 ROE 为前 10%和后 10%的股票组合，并求在每月求这两组按市值加权的平均涨跌幅之差。

ROE 因子的计算存在一些不够严谨以及存在个股之间的差异。首先，我们并未考虑在计算 ROE 因子时排除市值和 Book-to-Market ratio 的影响，缺少了控制变量的步骤。其次，由于公司财务报表公布的时间各不相同，导致一些股票的 ROE 数据的更新频率和节点会有差异。并且，ROE 虽然为季度更新的数据，但是一些公司在四月底将去年年底和今年第一季度的数据同时更新。因此在排序时永远按照最新一期的 ROE 进行排序，忽略去年年底的数据。

数据整合结果

	TRADE_DT	S_INFO_WINDCODE	W_SMB	W_HML	ROE_HML	mm_HML	fx_HML	涨跌幅	MKT_minus_free	grow_minus_free
33113	2005-01-31	000001.SZ	-0.0104718	0.00854091	0.00883802	0.001827	-0.005978	-0.080437	-0.060068	-0.081804
33114	2005-01-31	600228.SH	-0.0104718	0.00854091	0.00883802	0.001827	-0.005978	0.012365	-0.060068	0.010998
33115	2005-01-31	000691.SZ	-0.0104718	0.00854091	0.00883802	0.001827	-0.005978	-0.072626	-0.060068	-0.073993
33116	2005-01-31	600756.SH	-0.0104718	0.00854091	0.00883802	0.001827	-0.005978	-0.150981	-0.060068	-0.152348
33117	2005-01-31	600227.SH	-0.0104718	0.00854091	0.00883802	0.001827	-0.005978	-0.077792	-0.060068	-0.079159
...	...	...	...	...	...	...	...	...	...	...
441584	2021-06-30	603267.SH	0.0325423	-0.0348286	-0.00997474	-0.001201	0.000144	-0.006146	0.002862	-0.007556
441585	2021-06-30	603068.SH	0.0325423	-0.0348286	-0.00997474	-0.001201	0.000144	0.146375	0.002862	0.144965
441586	2021-06-30	601698.SH	0.0325423	-0.0348286	-0.00997474	-0.001201	0.000144	-0.014572	0.002862	-0.015982
441587	2021-06-30	603327.SH	0.0325423	-0.0348286	-0.00997474	-0.001201	0.000144	0.015635	0.002862	0.014225
441588	2021-06-30	603317.SH	0.0325423	-0.0348286	-0.00997474	-0.001201	0.000144	-0.185406	0.002862	-0.186816

408476 rows × 10 columns

图 3.1：数据整合结果

W_SMB、W_HML 为经过市值加权平均求出的 SMB、HML 因子；mm_HML、fx_HML 为动量因子和反转因子；ROE_HML 为 ROE 因子；MKT_minus_free 为市场因子；grow_minus_free 为因变量股票超额收益率。由上图 3.1 可见，目前的数据为 2005 年 1 月至 2021 年 6 月的月度数据。

和经典算法相比，我的 SMB 因子与经典 SMB 因子的相关系数为 0.98，我的 HML 因子与经典 HML 因子的相关系数为 0.89，我的 ROE 因子和经典 ROE 因子相关系数为 0.88，市场因子数据相同。动量因子和反转因子由于计算过程不同所以在此不做比较。

因子累计收益率对比如下图（3.2、3.3）所示。可以观察到，一些因子在经典算法中累计的数值更高。例如经典算法的 ROE 因子累计最高到 1.5 左右，而我的复制版本则在 0.75 左右。这或许是源自于我们处理极值方法上的差异。经典算法系统性地去除了某一类股票，但剩余的股票可能在某几个月也存在大幅上涨和下跌的情况，从而使得一些因子的绝对值更大。我的方法则是将每一期上涨或下跌较为极端的股票都剔除，这导致剩余股票中每个月都不存在大幅上涨或下跌的情况。这可能会让我所计算出的因子的绝对值更小，因为每组股票平均收益率之间不会存在大的差异。

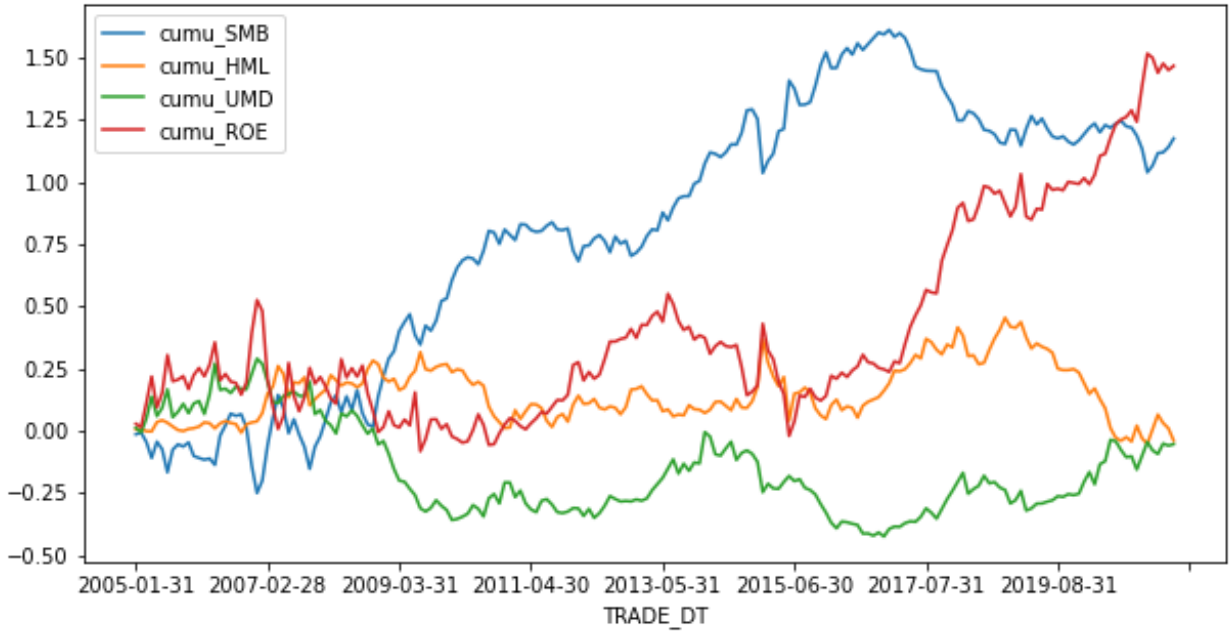

图 3.2：经典算法 SMB（黄）、HML（蓝）、动量因子（绿）、ROE（红）累计收益率

图 3.3：我的算法 SMB（黄）、HML（蓝）、动量因子（绿）、ROE（红）累计收益率

4. 建立模型以及分析

根据已有的数据，我们可以给每一支个股做线性回归，模型（a1）如下图所示。为检验因子是否有效，我们将带有因子的模型（a1, b1, c1）结果与其他带有正态分布的噪音模型（a2, b2, c2）比较，目的在于观察模型解释度是否有显著变化。可见，通过 c1 与 c2 的对比能看出 Fama French 三因子的有效性；b1 和 b2 的对比能看出动量和反转因子的有效性；而 a1 和 a2 可以检验出 ROE 因子的有效性。

$$grow\_minus\_free = \beta_{a1}\, MKT\_minus\_free + \beta_{a2}\, W\_SMB + \beta_{a3}\, W\_HML + \beta_{a4}\, mm\_HML + \beta_{a5}\, fx\_HML + \beta_{a6}\, ROE\_HML \quad (a1)$$

$$grow\_minus\_free = \beta_{b1}\, MKT\_minus\_free + \beta_{b2}\, W\_SMB + \beta_{b3}\, W\_HML + \beta_{b4}\, mm\_HML + \beta_{b5}\, fx\_HML + \beta_{b6}\, noise\_f \quad (a2)$$

$$grow\_minus\_free = \beta_{c1}\, MKT\_minus\_free + \beta_{c2}\, W\_SMB + \beta_{c3}\, W\_HML + \beta_{c4}\, mm\_HML + \beta_{c5}\, fx\_HML \quad (b1)$$

$$grow\_minus\_free = \beta_{d1}\, MKT\_minus\_free + \beta_{d2}\, W\_SMB + \beta_{d3}\, W\_HML + \beta_{d4}\, noise\_d + \beta_{d5}\, noise\_e \quad (b2)$$

$$grow\_minus\_free = \beta_{e1}\, MKT\_minus\_free + \beta_{e2}\, W\_SMB + \beta_{e3}\, W\_HML \quad (c1)$$

$$grow\_minus\_free = \beta_{f1}\, noise\_a + \beta_{f2}\, noise\_b + \beta_{f3}\, noise\_c \quad (c2)$$

行业间模型比较

我将模型在不同的行业中运行，判断上述六种模型在不同行业中的解释度差异。根据中信一级行业分类数据，所有股票分为 29 个行业。在个股的时间序列数据中运行上述四个模型，得到每个行业的平均 R 方（Adjusted R Square）。为解决某些股票数据量过小的情况，我排除了数据量小于 20 的股票。

结果发现，噪音和因子的作用存在明显的差异。只有噪音的（c2）模型对于任何行业的 R 方解释度都不超过 0.01，说明股票收益率是无法由随机波动解释的。而（c1）模型在加入 Fama French 三因子后，各行业平均 R 方的增长在 0.35~0.64 之间，说明 Fama French 三因子对于解释超额收益率有很大的提升作用（图 4.1）。对比（b1）（b2），我们发现加入动量、反向因子对于各行业的解释度也有小幅的提升作用（图 4.2）。对比（a1）（a2）可以发现，ROE 因子的作用最不明显，且对于有些行业的解释度不如噪音（图 4.3）。

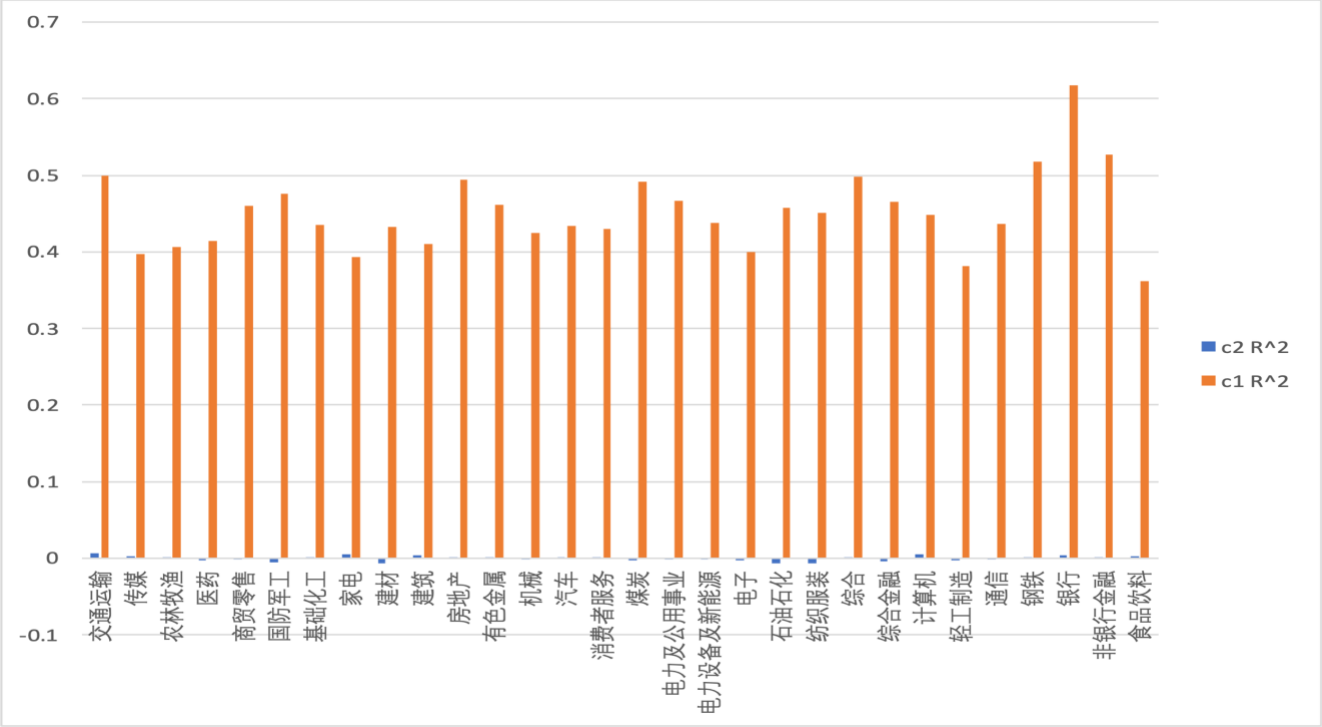

图 4.1：模型 c1 和 c2 的 R 方结果比较

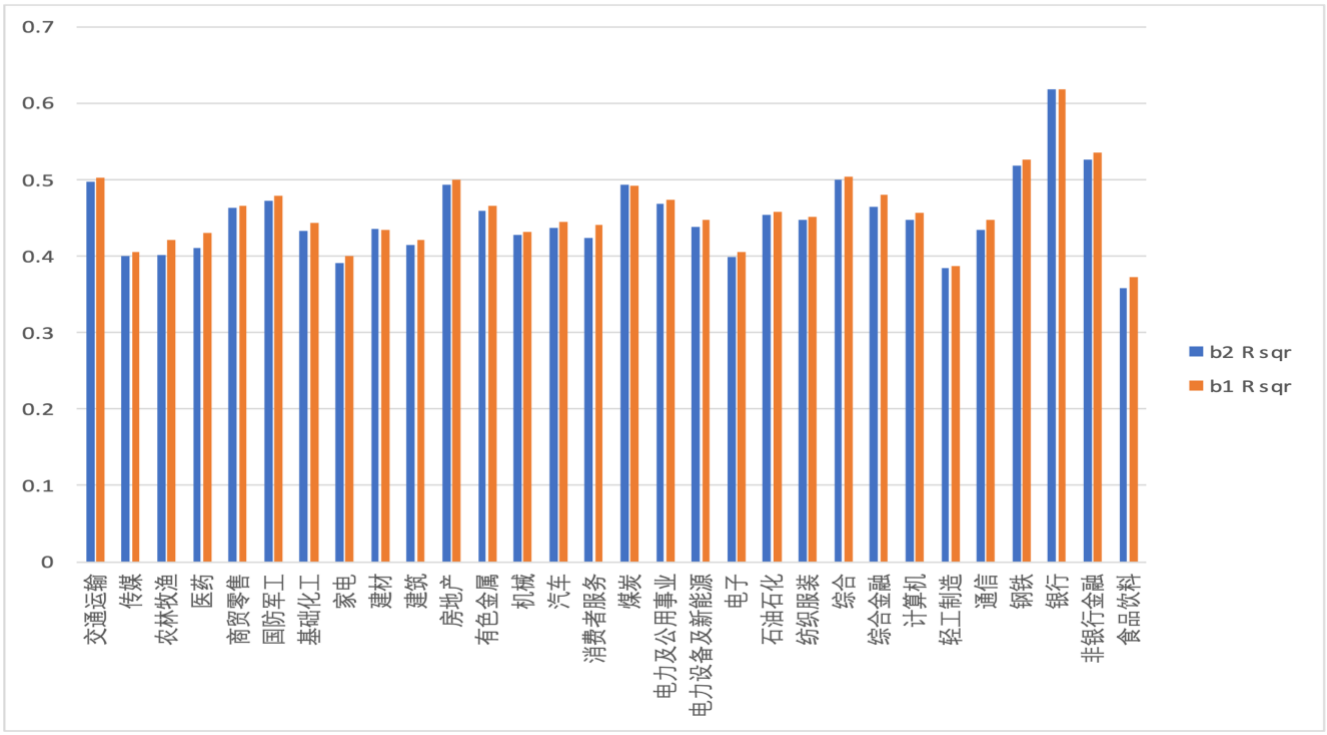

图 4.2：模型 b1 和 b2 的 R 方结果比较

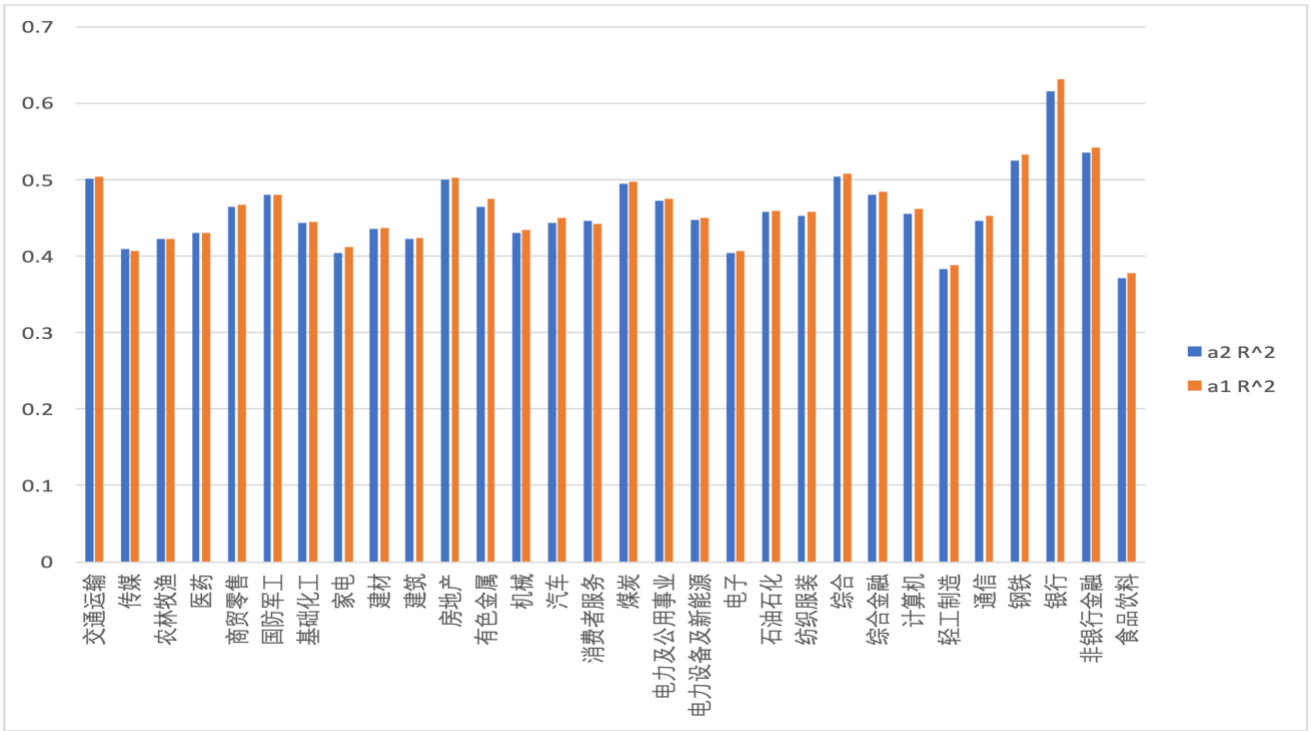

图 4.3：模型 a1 和 a2 的 R 方结果比较

从上图看，模型对于银行股票的解释度尤为突出。模型（c1）对于银行股票的解释度已经超过 0.6，并且通过观察模型结果可以看出，市场因子 MKT 和规模因子 HML 的显著性在其中起到很大作用。从图表中对比（a1）（a2）来看，加入 ROE 因子也一定程度地增加了模型在银行股票中的解释度。并且在 25 个银行股票数据中分别运行模型（a1）时，有 14 个模型的 ROE 因子都显著（p-value < 0.05），且每个模型平均有 2.7 个因子为显著。

模型对于食品行业股票的解释度是所有行业中最低的。通过因子的显著性来看，模型在食品行业股票中的显著因子的确是最少的，每个模型平均有 1.78 个显著因子。这说明了食品股票的收益率变化有更多市场整体变化无法解释的部分。

动量、反转、ROE 因子相对 Fama French 三因子对因变量解释度较低的原因并非是这三个因子无法用来解释股票超额收益率，而是因为在 Fama French 三因子的加入模型的基础之上，剩下的三个因子没有给解释度带来明显提升。也就是说，动量、反转、ROE 因子也难以解释 Fama French 三因子所解释不到的变化，这说明更多的其他因子需要加入进来并从不同维度去解释股票超额收益率的变化。为验证这个猜测，我们仅使用动量、反转、ROE 因子建立模型（d），结果如下。可见，这三个因子本身具有一定的解释作用（图 4.4）。

$$grow\_minus\_free = \beta_d\ mm\_HML + \beta_d\ fx\_HML + \beta_d\ ROE\_HML \quad (d)$$

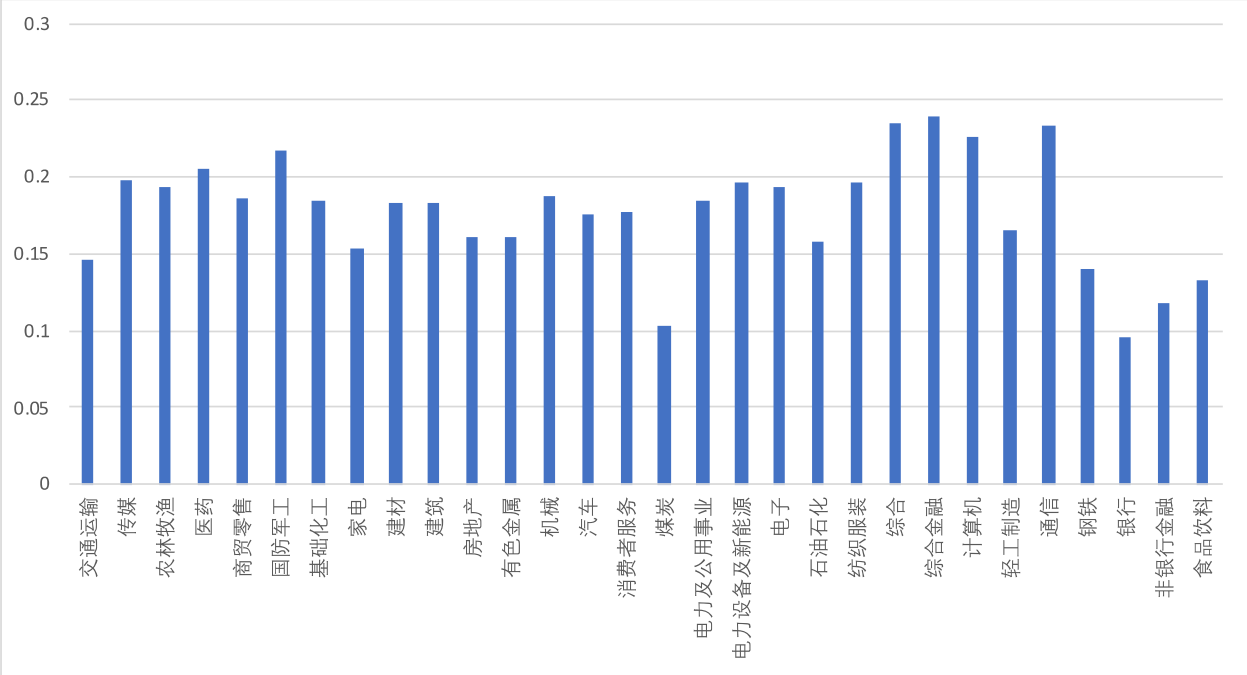

图 4.3：模型 d 的 R 方结果

5. 市值排名前 1000 股票池因子

将研究范围由所有 A 股转为市值排名前 1000 的 A 股后，因子值也相应有所改变。由于我所选取的市值数据为每年四月底市值，因此市值排名也相应地在每年四月底调整一次，并根据这 1000 支股票算出相应的 SMB，HML，动量，反转因子。如下图（5.1、5.2）所示：

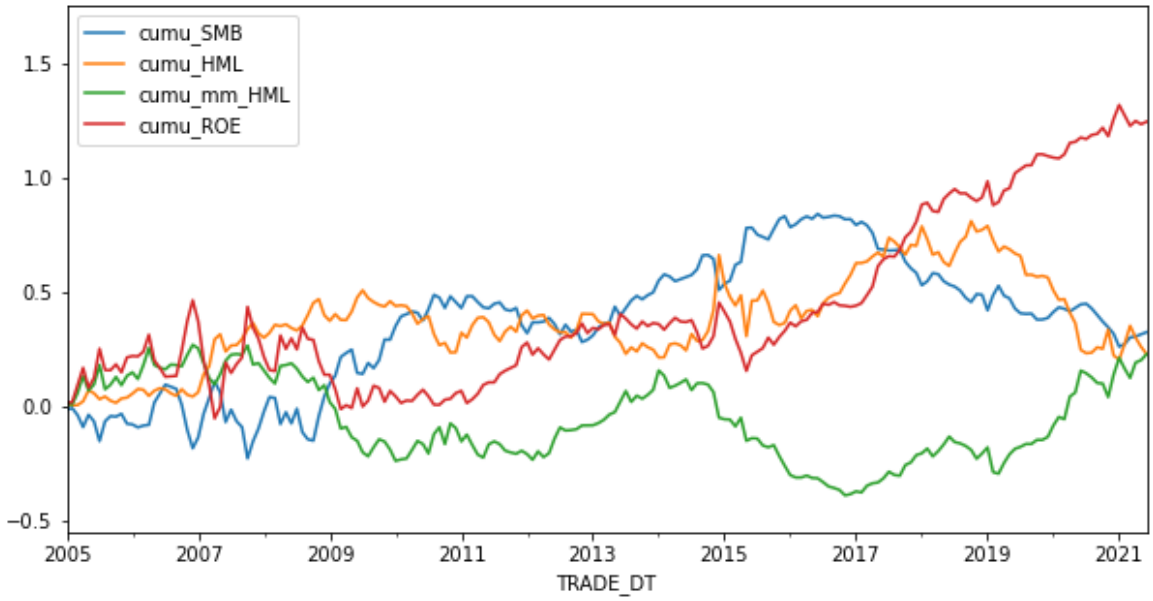

图 5.1：市值前 1000 股票池 SMB（黄）、HML（蓝）、动量因子（绿）累计收益率

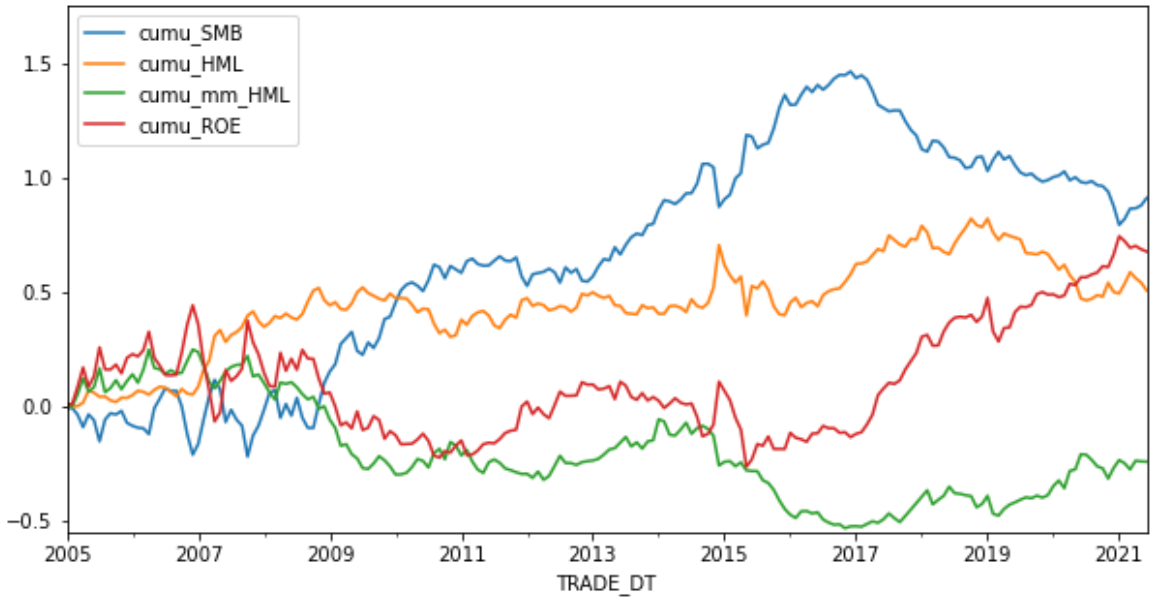

图 5.2：A 股 SMB（黄）、HML（蓝）、动量因子（绿）累计收益率

由此可见，从市值前 1000 股票池所得出的的因子累计收益率从绝对值上看更小。累计收益率绝对值更小是因为不同组合之间的收益率差异不显著，所以因子收益率变化不稳定。这一现象在 SMB 因子的体现上尤为明显，在 2007 年至 2017 年初，A 股市场 SMB 的累计值由-0.25 左右达到将近 1.5，而市值前 1000 股票池的 SMB 因子仅由-0.2 左右达到 0.8 左右。这说明在此期间小市值股票收益率的确高于大市值股票，但这种趋势更多地体现在整个 A 股市场，对于市值前 1000 的股票作用并不明显。

其中比较例外的因子是 ROE 因子，由上图可见，市值前 1000 股票池的 ROE 因子累计收益率自 2016 后稳定增加，并且程度高于 A 股市场的 ROE 因子。这说明 A 股 ROE 因子的累计值的增长可能主要是从市值前 1000 股票池中具有高净资产收益率的股票带来的——如果我们看向市值在 1000 名以后的股票，或许 ROE 因子的趋势不会如此明显。由下图 5.3 可见，市值在 1000 名后的股票池的确在 ROE 收益率上的变化较为平缓，说明了 ROE 因子收益率主要在来源于市值前 1000 的股票中。

对比 A 股市场的股票和市值前 1000 以后的股票池，我们同样发现从市值前 1000 之后股票池所得出的的因子累计收益率从绝对值上看更小。

结合以上内容，可见这些因子在 A 股市场中会带来更加明显的超额收益，而在大盘股和中小盘股中因子超额收益相对没有那么高。这说明如果我们现实中在大盘股或小盘股中投资时，这些因子的作用或许不会像理论中那样明显。

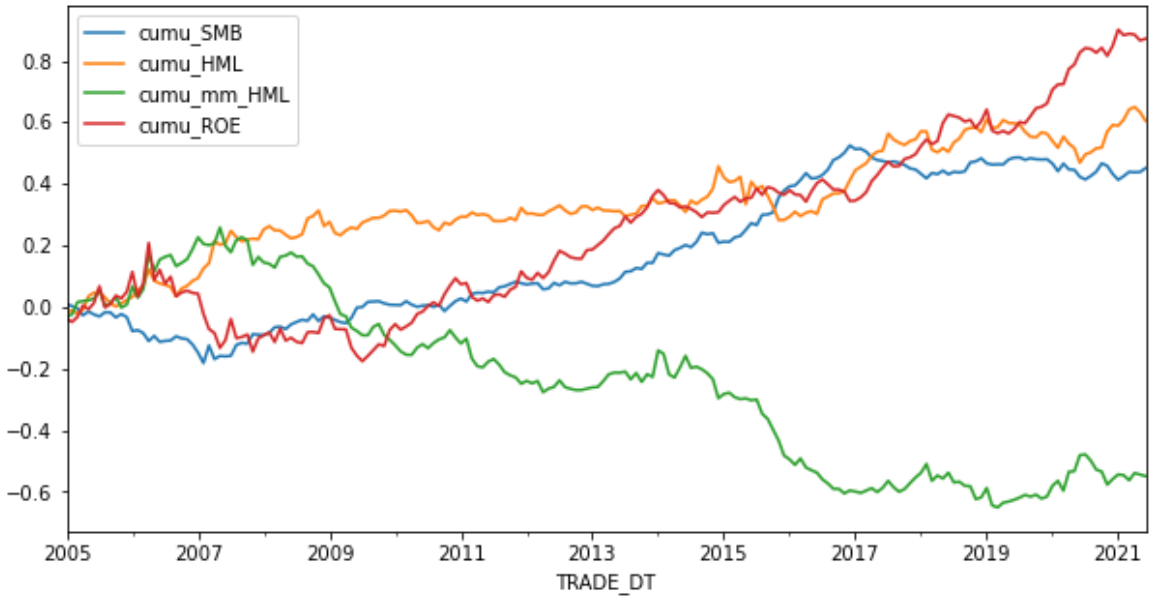

图 5.3：市值前 1000 后 x 股票池 SMB（黄）、HML（蓝）、动量因子（绿）累计收益率