专题报告

市值类因子有效性剖析

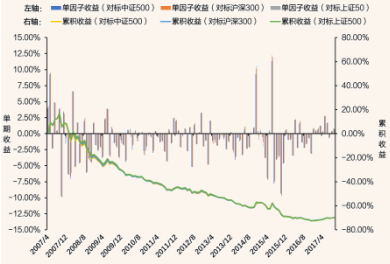
2017 年 12 月 31 日

因子模型系列之二

本报告是招商金工因子模型系列的第二篇报告。本报告在第一篇因子模型报告的基础上，介绍了根据日截面可计算条件来剔除异动数据，搭建可计算股票池的过程和模型被解释变量（即经系统性风险调整后的超额收益）的构建流程，并以市值类因子为例做了算法演示。本报告在整篇报告在整个系列报告中，起到了承上启下的作用。

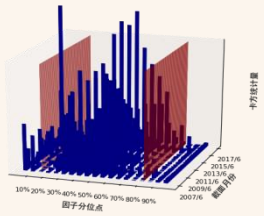
- 在数据库搭建介绍中，报告着重说明了日截面可计算条件，逐日剔除股价异动数据，构建每日可计算股票池。并用该可计算股票池建立等权市场组合和等权基准组合。用以后续构建被解释变量时候使用。
- 阐释了系统性风险估计参数$\Delta\tilde{\beta}^{M}_{i,t_0}$的估计过程，在整段观测窗口下，展示了$\Delta\tilde{\beta}^{M}_{i,t_0}$的分布，并按申万 1 级行业分类分别讨论了$\Delta\tilde{\beta}^{M}_{i,t_0}$在各行业的分布情况。详细叙述了模型被解释变量（即经系统性风险调整后的超额收益）的构建流程。
- 以市值类因子为例进行了算法展示，展示包括因子暴露数据清洗、删失的步骤，独立性检验的结果，标准化赋值的特点；并以不同的估计方法估计了市值类因子的收益。详尽汇报了市值类因子收益估计时候的各类指标，包括单期收益、t统计量概况、Sharpe 值、因子收益同向波动持续月份统计、因子波动量能等，并做了比较和必要总结。这些因子估计中得到的指标，将是我们后期选择因子入因子模型时的重要参考依据。
- 本报告为后续要做的其他各类因子收益估计做了算法上的演示，在整篇报告在整个系列报告中，起到了承上启下的作用。

合计市值单因子收益



资料来源：招商证券、Wind 资讯

合计市值因子χ^2统计量分布



资料来源：招商证券、Wind 资讯

相关报告

《基于增量信息逐层解释的因子模型框架搭建》2017-11

叶涛
021-68407343
yetao@cmschina.com.cn
S1090514040002

研究助理
崔浩瀚
cuihaohan@cmschina.com.cn

正文目录

多因子模型框架回顾 4
基础数据库搭建 5
等权市场组合 5
等权投资基准组合 5
因子模型被解释变量的构建 7
估计截面股票池 7
计算超额系统性风险暴露$\Delta\beta$ 7
因子模型被解释变量 10
单因子收益估计-以市值类因子为例 10
定义市值因子 10
独立性检验 10
因子暴露标准化赋值 16
市值类因子收益估计 16
结论 24

图表目录

图 1 单因子测试流程 4
图 2 等权中证 500 指数与中证 500 指数 5
图 3 等权沪深 300 指数与沪深 300 指数 5
图 4 等权上证 50 指数与上证 50 指数 6
图 5 $\Delta\boldsymbol{\beta}$的历史分布−中证 500 7
图 6 $\Delta\boldsymbol{\beta}$的历史分布−沪深 300 7
图 7 $\Delta\boldsymbol{\beta}$的历史分布−上证 50 8
图 8 合计市值因子$\boldsymbol{\chi 2}$统计量分布（对标中证 500） 11
图 9 合计市值因子$\boldsymbol{\chi 2}$统计量分布（对标沪深 300） 12
图 10 合计市值因子$\boldsymbol{\chi 2}$统计量分布（对标上证 50） 12
图 11 流通市值因子$\boldsymbol{\chi 2}$统计量分布（对标中证 500） 13
图 12 流通市值因子$\boldsymbol{\chi 2}$统计量分布（对标沪深 300） 13
图 13 流通市值因子$\boldsymbol{\chi 2}$统计量分布（对标上证 50） 14
图 14 自由流通市值因子$\boldsymbol{\chi 2}$统计量分布（对标中证 500） 14

图 15 自由流通市值因子$\chi 2$统计量分布（对标沪深 300）..... 15
图 16 自由流通市值因子$\chi 2$统计量分布（对标上证 50）..... 15
图 17 合计市值单因子收益..... 17
图 18 流通市值单因子收益..... 18
图 19 自由流通市值单因子收益..... 19
图 20 因子波动量能..... 20
图 21 合计市值因子收益 t 统计量走势..... 21
图 22 流通市值因子收益 t 统计量走势..... 21
图 23 合计市值因子收益 t 统计量走势..... 22
图 24 合计市值因子(中证 500) V.S. 大小市值指数之差..... 22
图 25 流通市值因子(中证 300) V.S. 大小市值指数之差..... 23
图 26 自由流通市值因子(上证 50) V.S. 大小市值指数之差..... 23

多因子模型框架回顾

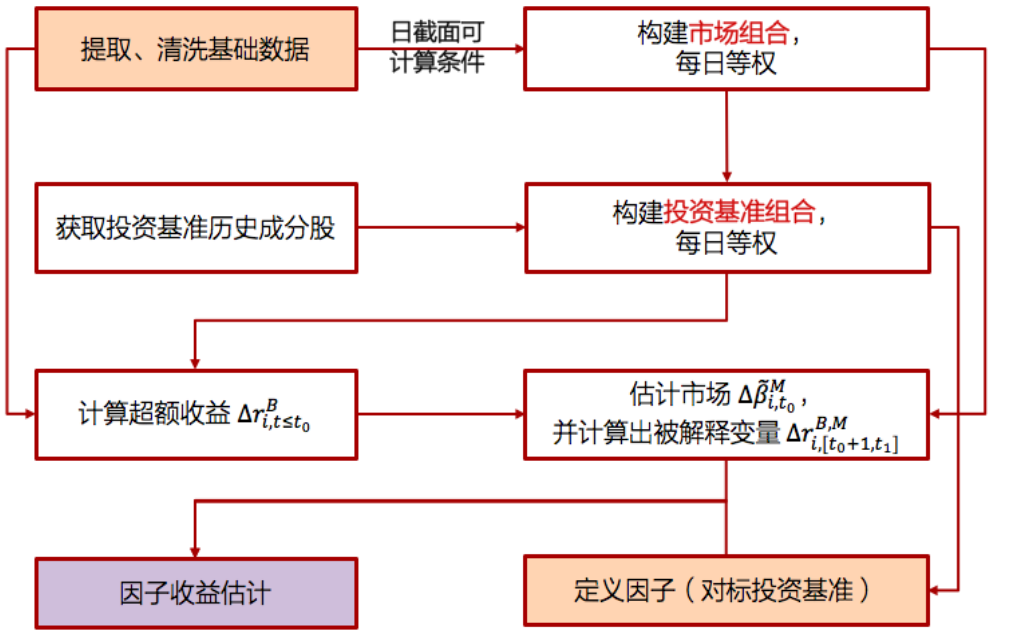
在上一篇报告中，我们详尽介绍了我们的因子模型框架的搭建过程。在因子模型应是选因子而非直接选股的构建理念指导下，基于横截面模型固有的特点，我们选择了横截面模型作为我们因子模型的基本形式，并始终以超额的因子暴露来解释股票的超额收益。其基本形式为：

$$\left(\Delta r^{B,M}_{i,[t_0,t_1]}\right)_{n\times 1} = \left(c_{[t_0,t_1]}\right)_{n\times 1} + \left(\Delta\beta^{(j)}_{i,t_0}\right)_{n\times k}\cdot\left(r^{(j)}_{F,[t_0,t_1]}\right)_{k\times 1} + \left(\varepsilon_{i,[t_0,t_1]}\right)_{n\times 1} \quad \text{式（1）}$$

在因子模型放入多因子的环节，采用增量信息逐层解释方法来增添新因子；用逐层增量解释的方法可以消除因子之间可能存在的多重共线性，使得线性回归得到的估计值最大限度接近最优线性无偏估计量。

从本篇报告开始，我们将逐步实践前述的模型构想，同时也将在实际测算中，不断修正和完善我们的因子框架。本报告将系统阐释我们对数据的处理方式、模型被解释变量的构建方法。并以市值类因子为例，来对单因子估计的方法进行说明。

图 1 单因子测试流程



资料来源：招商证券

本报告的大致流程可以如下概述：

从 Wind 数据库中提取基础数据，根据日截面可计算条件来构建等权市场组合，并与三类常用基准（中证 500、沪深 300 和上证 50）的历史成分股取交集之后，构建等权投资基准组合。

个股绝对收益与计算得到的等权基准对标，可得个股超额收益$\Delta r^{B}_{i,t\le t_0}$。再由日频超额收益序列$\Delta r^{B}_{i,t\le t_0}$与市场组合序列$r_{M,t\le t_0}$直接以普通最小二乘法（OLS）估计个股在$t_0$截面的超额系统性风险暴露$\Delta\tilde{\beta}^{M}_{i,t_0}$,并计算因子模型最后要用到的被解释变量$\Delta r^{B,M}_{i,[t_0+1,t_1]}$。

定义因子，将因子与等权基准对标，可得超额因子暴露，作为模型的解释变量，由被解释变量和解释变量来估计单因子收益。

基础数据库搭建

报告中所用到的行情数据、股本数据、交易日数据等基础数据均取自 Wind 数据库服务的产品 Filesync。为了使得因子收益的估计结果不受异动值的不良影响，我们对原始数据进行了必要处理，根据日截面可计算条件，剔除极有可能出现股价异动的交易日数据，并计算出等权市场组合和等权基准组合。

等权市场组合

选取 A 股市场所有个股数据（含已退市股票），根据日截面可计算条件，按每日截面进行滚动观测，构建可选股票池。满足以下情况的个股数据，需剔除，不纳入后续的计算，以避免给估计造成不良影响：

1. 新股上市 10 日内（含第 10 日）数据；
2. 证券更名后 10 日内（含第 10 日）的数据；
3. 证券更名后 10 日内（含第 10 日）的数据；
4. 处于特别处理期个股数据（ST 股）和摘帽后 10 日内（含第 10 日）的数据；
5. 个股出现一字板的当日数据；
6. 全天停牌个股当日数据。

由日截面股票池个股可计算条件确定当日入选股票名单，构建每日等权市场组合，即将t日 A 股市场上所有个股对数收益率（以下若无特别强调，所提到的收益率均指对数收益率）取等权平均。

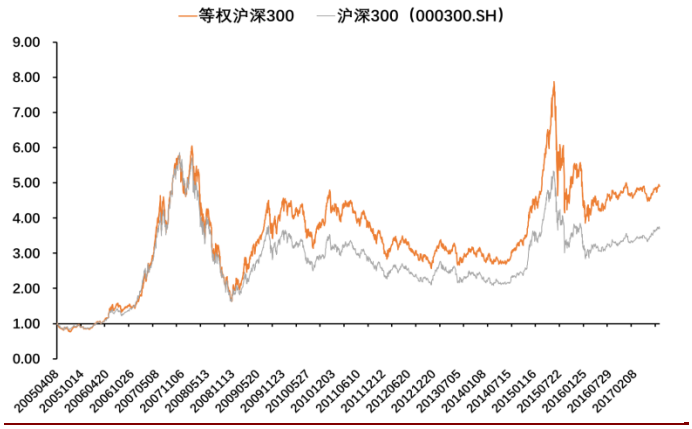
等权投资基准组合

从 Wind 数据库(FileSync)中提取中证 500(000905.SH)、沪深 300(000300.SH)、上证 50（000016.SH）的历史成分股名单。由日截面股票池每日可计算条件确定当日入选投资基准的股票名单（取交集），构建每日等权投资基准组合。

图 2 等权中证 500 指数与中证 500 指数

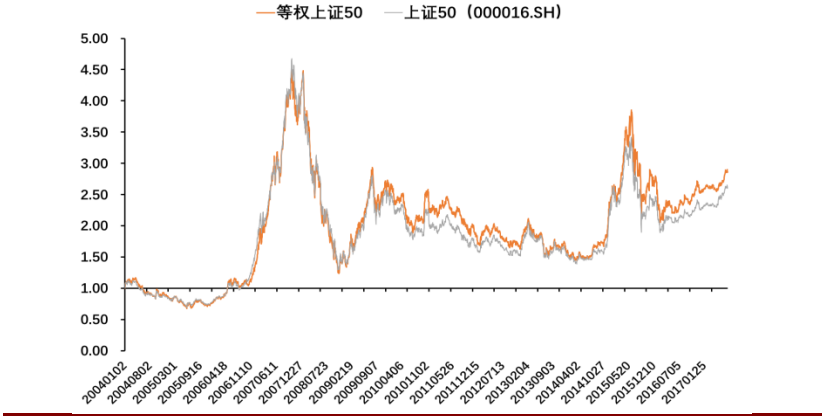
资料来源：招商证券、Wind 资讯

图 3 等权沪深 300 指数与沪深 300 指数



资料来源：招商证券、Wind 资讯

图 4 等权上证 50 指数与上证 50 指数



资料来源：招商证券、Wind 资讯

图 2 至图 4 分别展示了标准基准指数（中证 500、沪深 300、上证 50）与各自对应的等权投资基准之间净值走势的差异。其中橙色线代表等权投资基准；而灰色线代表标准基准指数。从图 2 至图 4 中可以看出，由于等权基准能捕捉到市场反转效应的收益，因此在大部分时间能跑赢标准基准指数。等权上证 50 对标准基准指数的组合差异要小于另两个组合，可能是因为上证 50 成分股的波动较为平缓。

之所以使用等权投资基准作为我们所参考的投资基准，首先是因为根据每日可计算条件构建的等权基准，更能反映市场上真实可交易的成分股收益。譬如，在 2015 年下半年，市场上有大量个股停牌，而标准指数基准仍然将这些停牌个股收益情况纳入指数计算，这势必不能反映当时市场上真实可交易的成分股收益；而根据每日可计算条件来构建的等权基准并不存在这一问题。第二，使用等权方式来构建投资基准是与等权市场组合算法进行统一，以便于后续的各类因子数据处理。

等权投资基准是比标准基准指数更为严格的参考标准，若策略能跑赢等权投资基准，那么能大概率跑赢标准基准指数。

表 1：等权基准指数对标准基准指数的胜率和盈亏比

基准指数	等权中证 500 指数		等权沪深 300 指数		等权上证 50 指数	
窗口长度	胜率	盈亏比	胜率	盈亏比	胜率	盈亏比
5	60.10%	1.035	53.89%	0.997	50.96%	1.021
10	61.20%	1.165	53.71%	1.072	50.61%	1.064
20	64.77%	1.208	52.45%	1.222	52.53%	1.005
40	67.10%	1.458	55.71%	1.172	51.75%	1.030
60	66.64%	1.798	56.33%	1.239	50.77%	1.107
120	75.75%	2.120	60.67%	1.400	51.85%	1.223
240	81.67%	4.198	65.30%	1.606	50.88%	1.454
360	84.29%	5.530	68.01%	2.026	53.50%	1.292
480	85.59%	8.102	74.23%	3.054	62.93%	1.034

资料来源：招商证券、Wind 资讯

按不同的窗口长度进行滚动观测，并计算买入并持有条件下的等权投资基准对对应标准基准指数的胜率和盈亏比（例如等权中 500 指数对标准中证 500 指数的胜率和盈亏比）。纵向看，我们分别以 5 日、10 日、20 日、40 日、60 日、120 日、240 日、360 日和 240 日窗口长度进行滚动观测，随着观测窗口长度的增加，等权投资基准对标准基准指数的胜率和盈亏比基本是在不断上升；横向看，胜率和盈亏比最高的是等权中证 500 指数，而胜率和盈亏比最低的则是等权上证 50 指数。这和图 2 至图 4 展示的情形

一致。

因子模型被解释变量的构建

估计截面股票池

如前文所述，我们的因子模型是横截面模型，在单因子收益估计之前，需先确定估计横截面。考虑到中证 500 指数在 2007 年 1 月 15 日开始才能取得历史成分股名单，为便于对三类基准对应的情况加以比较，将因子模型数据取样窗口定为 2007 年 1 月 15 日至 2017 年 10 月 31 日。

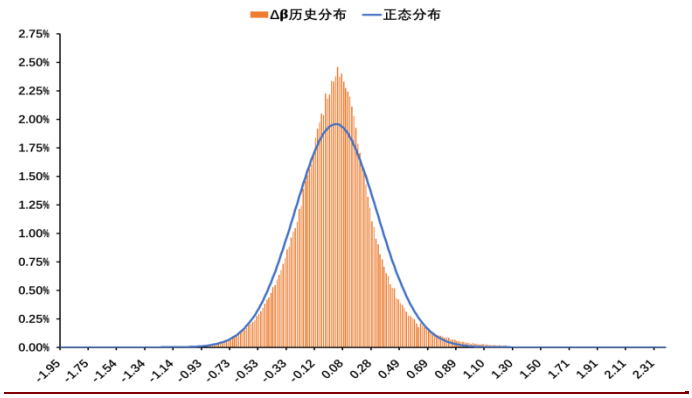
将模型的估计截面暂定为每个自然月的最后一个交易日；而个股截面点定为该自然月满足可计算条件的最后一个交易日。即若个股在某个自然月的最后一个交易日满足可计算条件，则将最后一个交易日作为该个股参与该截面计算的数据取用点；若个股在某个自然月的最后一个交易日不满足可计算条件，则向前寻找个股在该月中满足可计算条件的交易日，取用离最后交易日最近的可计算交易日作为该截面计算的数据取用点。根据上述规则来确认当期估计截面估计股票池，出现以下情况的个股，将从当月截面计算中被剔除，不参与后续计算。

1. 当前自然月最后交易日前向 5 个交易日窗口内个股均不满足可计算条件的个股；
2. 下一自然月最后交易日不满足可计算条件的个股；
3. 下一自然月，满足可计算条件的交易日天数占比低于 1/2 的个股。

计算超额系统性风险暴露$\Delta\tilde{\beta}_{i,t_0}^{M}$

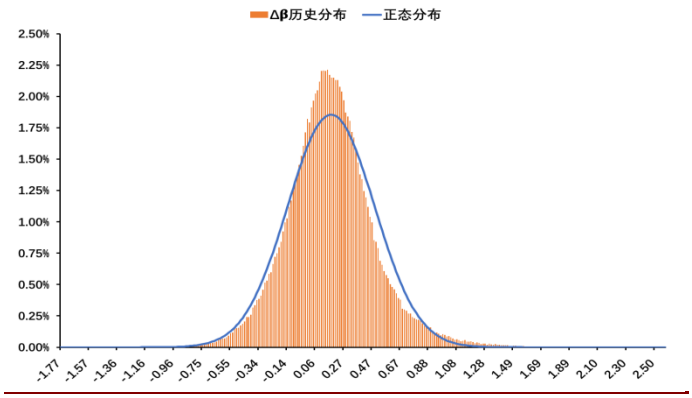
将月估计截面日期表示为t_0，由日频超额收益序列$\Delta r_{i,t\le t_0}^{B}$与市场组合序列$r_{M,t\le t_0}$可以直接以普通最小二乘法（OLS）估计个股$t_0$截面的超额系统性风险暴露$\Delta\tilde{\beta}_{i,t_0}^{M}$以及随机项估计误差$\tilde{\sigma}[\varepsilon_{i,t\le t_0}]$。其中，$\Delta r_{i,t\le t_0}^{B}=r_{i,t\le t_0}-r_{B,t\le t_0}$，即个股的超额收益等于个股绝对收益减去等权基准收益。考虑到投资者的换手频率的适当性以及可观测样本数据量的合理性，这里t取估计截面前 60 个交易日作为观测窗口来估计超额系统性风险暴露$\Delta\tilde{\beta}_{i,t_0}^{M}$。

图 5 $\Delta\tilde{\beta}_{i,t_0}^{M}$的历史分布-中证 500



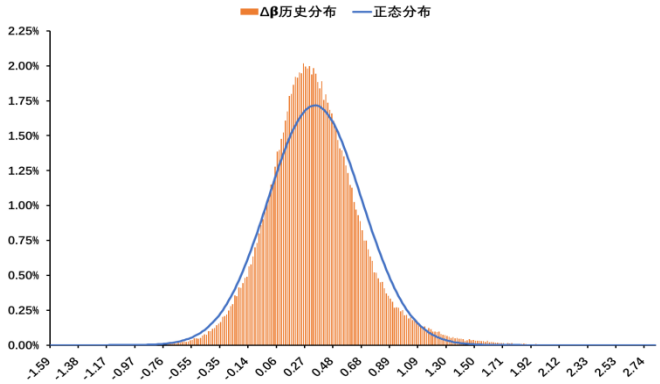
资料来源：招商证券、Wind 资讯

图 6 $\Delta\tilde{\beta}_{i,t_0}^{M}$的历史分布-沪深 300



资料来源：招商证券、Wind 资讯

图 7 $\Delta\tilde{\beta}_{i,t_0}^{M}$的历史分布-上证 50



资料来源：招商证券、Wind 资讯

图 5 至图 7 展示了对应不同基准情况下，$\Delta\tilde{\beta}_{i,t_0}^{M}$的历史分布。橙色条形图展示不同取值范围下$\Delta\tilde{\beta}_{i,t_0}^{M}$的占比，蓝色线表示对应的正态分布。可见$\Delta\tilde{\beta}_{i,t_0}^{M}$的历史分布和正态分布是具有一定相似性的，略带有偏度。

表 2：分行业Δβ统计-中证 500

申万 1 级行业	R^2中位数	R^2平均值	Δβ中位数	Δβ平均值	申万 1 级行业	R^2中位数	R^2平均值	Δβ中位数	Δβ平均值
银行	43.64%	41.75%	-0.672	-0.638	医药生物	2.34%	5.70%	-0.047	-0.053
计算机	5.55%	8.34%	0.298	0.330	休闲服务	2.28%	5.13%	-0.056	-0.056
金融服务	4.93%	13.19%	-0.162	-0.171	采掘	2.24%	5.60%	0.033	0.028
通信	4.37%	7.25%	0.229	0.251	建筑材料	2.13%	4.34%	0.043	0.071
非银金融	4.29%	9.94%	-0.131	-0.130	信息服务	2.12%	4.57%	0.047	0.061
交通运输	3.71%	9.93%	-0.103	-0.095	机械设备	2.10%	4.46%	0.069	0.081
传媒	2.99%	6.19%	0.166	0.175	轻工制造	2.03%	4.72%	-0.004	0.009
钢铁	2.85%	8.04%	-0.099	-0.107	化工	1.98%	4.39%	0.051	0.053
食品饮料	2.79%	6.89%	-0.098	-0.112	信息设备	1.96%	3.91%	0.078	0.084
电气设备	2.75%	5.40%	0.127	0.145	纺织服装	1.93%	4.36%	0.012	0.005
公用事业	2.75%	7.22%	-0.055	-0.065	国防军工	1.87%	3.91%	0.109	0.154
电子	2.72%	5.05%	0.135	0.153	建筑建材	1.85%	4.70%	0.020	0.009
汽车	2.60%	5.75%	0.081	0.084	综合	1.80%	3.60%	0.058	0.048
建筑装饰	2.40%	5.79%	0.052	0.068	房地产	1.73%	4.16%	-0.027	-0.037
家用电器	2.38%	5.23%	0.005	0.007	交运设备	1.71%	3.93%	0.007	0.003
商业贸易	2.38%	5.17%	-0.027	-0.026	农林牧渔	1.68%	3.79%	0.006	0.000
有色金属	2.36%	4.62%	0.082	0.072	**合计**	**2.32%**	**5.50%**	**0.023**	**0.024**

资料来源：招商证券、Wind 资讯

表 3：分行业Δβ统计-沪深 300

申万 1 级行业	R^2中位数	R^2平均值	Δβ中位数	Δβ平均值	申万 1 级行业	R^2中位数	R^2平均值	Δβ中位数	Δβ平均值
银行	33.30%	35.22%	-0.461	-0.421	信息设备	3.53%	5.70%	0.198	0.208
计算机	12.05%	15.06%	0.517	0.547	综合	3.41%	6.10%	0.186	0.189
通信	9.75%	12.82%	0.435	0.465	家用电器	3.39%	6.61%	0.146	0.157
传媒	8.39%	12.01%	0.377	0.396	公用事业	3.23%	6.85%	0.085	0.087
电气设备	7.43%	10.59%	0.338	0.359	纺织服装	3.10%	5.93%	0.146	0.153
国防军工	6.81%	9.40%	0.314	0.372	信息服务	3.05%	5.44%	0.161	0.179
汽车	6.10%	9.47%	0.282	0.299	商业贸易	3.02%	6.08%	0.104	0.114
建筑装饰	5.90%	8.79%	0.262	0.287	轻工制造	2.88%	5.82%	0.135	0.162

申万 1 级行业	R^2中位数	R^2平均值	Δβ中位数	Δβ平均值	申万 1 级行业	R^2中位数	R^2平均值	Δβ中位数	Δβ平均值
金融服务	5.73%	12.46%	-0.05	-0.052	食品饮料	2.84%	6.04%	0.050	0.044
电子	5.31%	8.42%	0.29	0.31	医药生物	2.84%	6.13%	0.102	0.103
非银金融	5.06%	9.77%	0.076	0.086	建筑建材	2.83%	5.66%	0.129	0.125
建筑材料	4.95%	8.29%	0.249	0.283	钢铁	2.79%	6.68%	0.023	0.025
有色金属	4.54%	7.36%	0.223	0.221	农林牧渔	2.54%	5.25%	0.149	0.155
采掘	3.90%	7.45%	0.174	0.176	休闲服务	2.42%	4.79%	0.086	0.096
机械设备	3.84%	7.00%	0.209	0.227	交运设备	2.40%	4.65%	0.120	0.122
化工	3.71%	6.75%	0.189	0.2	房地产	2.37%	4.88%	0.112	0.114
交通运输	3.54%	7.93%	0.032	0.049	**合计**	**3.64%**	**7.05%**	**0.163**	**0.175**

资料来源：招商证券、Wind 资讯

表 4：分行业Δβ统计-上证 50

申万 1 级行业	R^2中位数	R^2平均值	Δβ中位数	Δβ平均值	申万 1 级行业	R^2中位数	R^2平均值	Δβ中位数	Δβ平均值
计算机	19.71%	22.35%	0.749	0.778	金融服务	6.74%	11.89%	0.088	0.088
通信	16.74%	19.79%	0.660	0.694	纺织服装	6.30%	9.98%	0.299	0.313
传媒	16.60%	19.76%	0.606	0.629	家用电器	6.02%	10.12%	0.303	0.323
电气设备	15.56%	18.18%	0.562	0.587	轻工制造	5.78%	9.61%	0.288	0.325
国防军工	15.04%	17.77%	0.545	0.604	农林牧渔	5.70%	9.06%	0.305	0.319
银行	13.51%	21.16%	-0.237	-0.187	信息服务	5.66%	8.01%	0.309	0.322
建筑装饰	12.62%	15.65%	0.477	0.519	公用事业	5.54%	9.58%	0.233	0.248
汽车	12.46%	16.11%	0.503	0.529	建筑建材	5.42%	8.17%	0.255	0.255
非银金融	12.08%	16.44%	0.301	0.317	商业贸易	5.30%	9.45%	0.245	0.266
建筑材料	11.96%	15.38%	0.468	0.510	医药生物	5.16%	9.36%	0.260	0.268
电子	9.91%	13.26%	0.459	0.481	交通运输	5.16%	9.23%	0.173	0.199
有色金属	9.12%	12.31%	0.379	0.384	房地产	4.96%	8.40%	0.263	0.279
采掘	8.94%	12.28%	0.331	0.333	交运设备	4.80%	7.33%	0.246	0.252
机械设备	7.58%	11.29%	0.357	0.384	食品饮料	4.45%	8.06%	0.207	0.207
化工	7.48%	10.99%	0.337	0.355	休闲服务	4.45%	7.85%	0.237	0.258
综合	7.40%	10.51%	0.335	0.349	钢铁	4.24%	7.94%	0.160	0.174
信息设备	7.02%	9.35%	0.338	0.353	**合计**	**6.91%**	**10.82%**	**0.316**	**0.338**

资料来源：招商证券、Wind 资讯

为了更细致地分析$\Delta\tilde{\beta}_{i,t_0}^{M}$的历史分布情况，我们还分行业统计了$\Delta\tilde{\beta}_{i,t_0}^{M}$估计过程中的判定系数$R^2$以及$\Delta\tilde{\beta}_{i,t_0}^{M}$的中位数与平均数（三表分别对应三个不同的基准）。行业分类方法选择窗口期内曾出现过的申万 1 级行业分类，大部分个股在窗口期内出现过一次甚至多次行业变更，我们都对个股对应时期做了精确归类。

以中证 500 或者沪深 300 为基准时，银行行业的回归判定系数R^2相对较大，原因是银行股在窗口期内波动较为平缓，与基准偏离较大，所以能被系统性风险解释的超额收益部分也较多，$\Delta\tilde{\beta}_{i,t_0}^{M}$也呈负值。而在上证 50 为基准的情况下，由于上证 50 大部分的成分股来自银行行业，银行行业对上证 50 指数的偏离较小，能被系统风险解释的超额收益部分也相应减少，因而以上证 50 为基准时，银行行业的R^2就有所下降。

另一个R^2较大的行业是计算机行业。由于计算机行业在观测窗口期内走势较强，因此和基准的偏离也较大，能被系统风险解释的超额收益部分也较多。

因子模型被解释变量

计算模型被解释变量，$\Delta r^{B,M}_{i,[t_0+1,t_1]} = \Delta r^{B}_{i,[t_0+1,t_1]} - \Delta\tilde{\beta}^{M}_{i,t_0} \cdot r_{M,[t_0+1,t_1]}$，其中，$\Delta r^{B}_{i,[t_0+1,t_1]}$指$t_0$截面到$t_1$截面一个月间个股的超额收益，$r_{M,[t_0+1,t_1]}$是指指$t_0$截面到$t_1$截面一个月间市场组合收益。$\Delta\tilde{\beta}^{M}_{i,t_0}$是$t_0$截面前 60 个交易日窗口所计算得到的超额系统性风险暴露。模型被解释变量，指的是：即经系统性风险暴露调整后的个股超额收益，通俗地讲，因子模型的被解释变量是指下个月超额收益减掉能被市场系统性风险解释部分后，剩下的那部分超额收益。此构建方式体现因子模型始终以超额的因子暴露来解释超额收益的构建理念。

单因子收益估计-以市值类因子为例

报告用市值类因子为例，来介绍因子模型对于单因子收益估计的计算方法和考量因子优先劣后的指标。

定义市值因子

市值因子选取 A 股合计市值、流通 A 股市值和自由流通 A 股市值（以下分别简称：合计市值、流通市值和自由流通市值）。具体计算方法是：从 Wind 数据库（FileSync）提取可计算交易日的股本数据，而后乘以当日收盘价计算市值。由于 A 股市场市值大小的分布存在明显左偏，为缓解左偏性质对线性估计造成的不良影响，对市值做了对数处理。

表 5：因子计算表

因子名称	Wind 数据库（FileSync）股本字段名	计算公式
合计市值	S_SHARE_TOTALA	ln(S_SHARE_TOTALA × 收盘价)
流通市值	FLOAT_A_SHR	ln(FLOAT_A_SHR × 收盘价)
自由流通	S_SHARE_FREESHARES	ln(S_SHARE_FREESHARES × 收盘价)

资料来源：招商证券、Wind 资讯

用同样的方法计算各基准成分股的市值对数，等权平均后得到基准的市值对数（因子均值）。在每个日截面上与等权基准成分股的相应因子均值进行对标，最后得到作为解释变量的市值因子。

独立性检验

对被解释变量（经系统性风险暴露调整后的个股超额收益$\Delta r^{B,M}_{i,[t_0+1,t_1]}$）和解释变量（和基准对标后的因子暴露）进行独立性检验。报告选用方法为卡方独立性检验。

卡方独立性检验是皮尔森卡方检定中的一种应用，用来检验两个变量之间是否独立。卡方独立性检验的原假设H_0:两个变量相互独立。根据观测数据构造卡方统计量：

$$\chi^2 = \sum \frac{(A-E)^2}{E} = \sum_{i=1}^{k} \frac{(A_i-E_i)^2}{E_i} \qquad \text{式（2）}$$

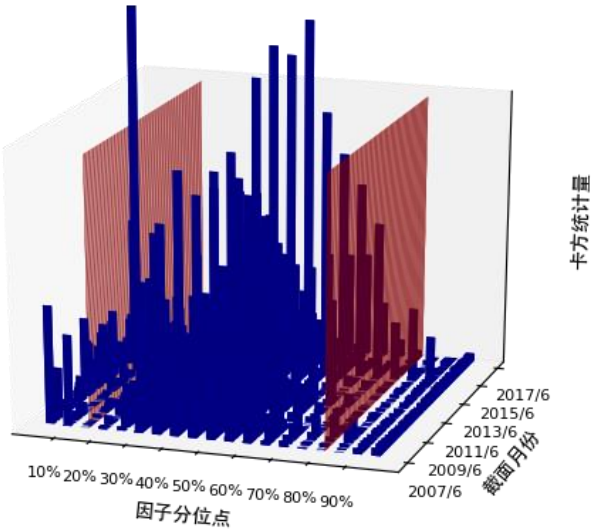
其中，A_i为i水平的观察频数，E_i为i水平的期望频数，当n比较大时，χ^2统计量近

似服从$k-1$个自由度的卡方分布。当观察频数与期望频数完全一致时，χ^2值为 0；观察频数与期望频数越接近，两者之间的差异越小，χ^2值越小；反之，观察频数与期望频数差别越大，两者之间的差异越大，χ^2值越大。

χ^2越大，越有理由拒绝两个变量相互独立的原假设。相较于皮尔森相关系数检验变量间的线性相关性，卡方检验的优势在于其既能检验可能存在的线性相关，又能检验可能存在的非线性相关。

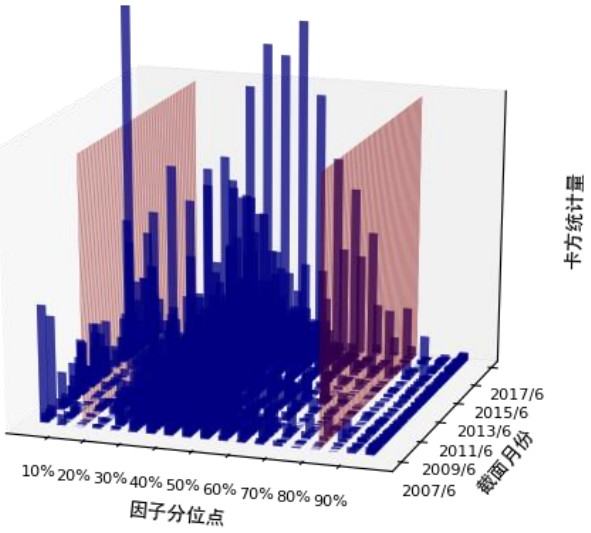
将与基准对标后的市值因子按不同的分位点进行切分，选取切分的分位点对应概率$p \in \{5\%, 10\%, \ldots, 90\%, 95\%\}$，在每个切割的分位点上建立频数列联表，分别进行卡方独立性检验。我们对每个因子，分别对应三个基准做了卡方独立性检验，并以三维条形图的方式来展示每个截面上，各个分位点的卡方值分布。

图 8 合计市值因子χ^2统计量分布（对标中证 500）



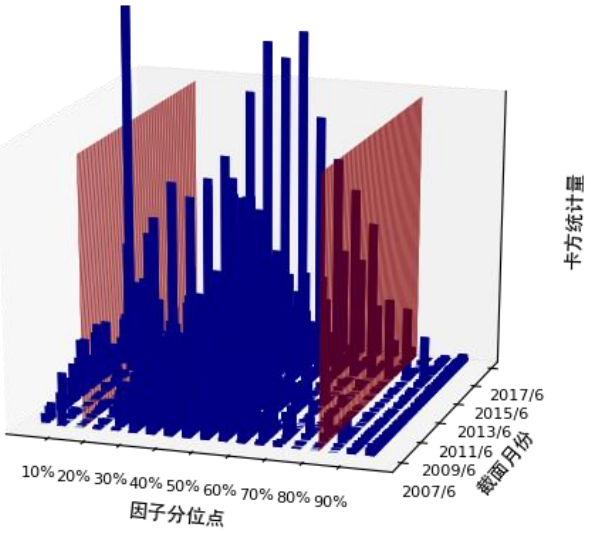
资料来源：招商证券、Wind 资讯

图 9 合计市值因子χ^2统计量分布（对标沪深 300）



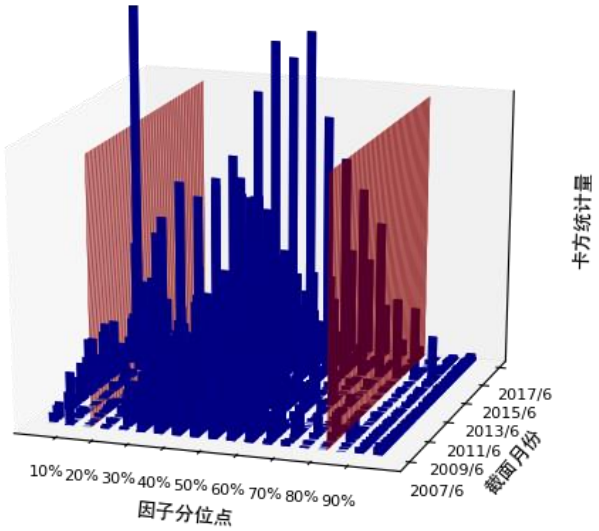
资料来源：招商证券、Wind 资讯

图 10 合计市值因子χ^2统计量分布（对标上证 50）



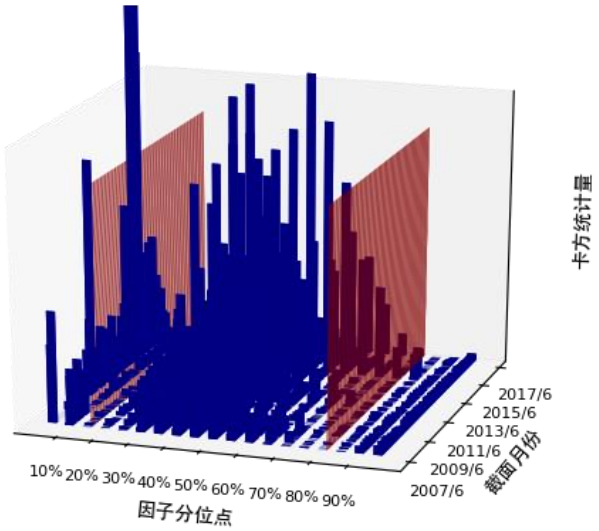
资料来源：招商证券、Wind 资讯

图 11 流通市值因子χ^2统计量分布（对标中证 500）



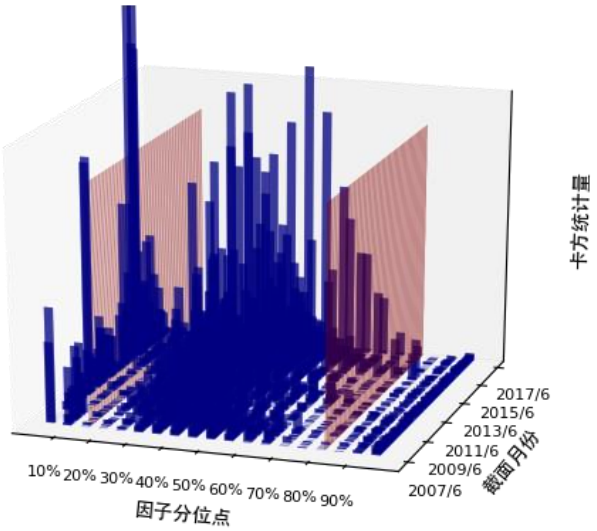
资料来源：招商证券、Wind 资讯

图 12 流通市值因子χ^2统计量分布（对标沪深 300）



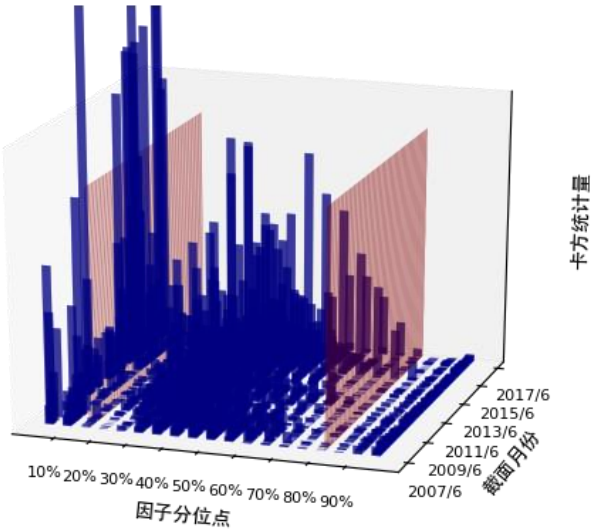
资料来源：招商证券、Wind 资讯

图 13 流通市值因子χ^2统计量分布（对标上证 50）



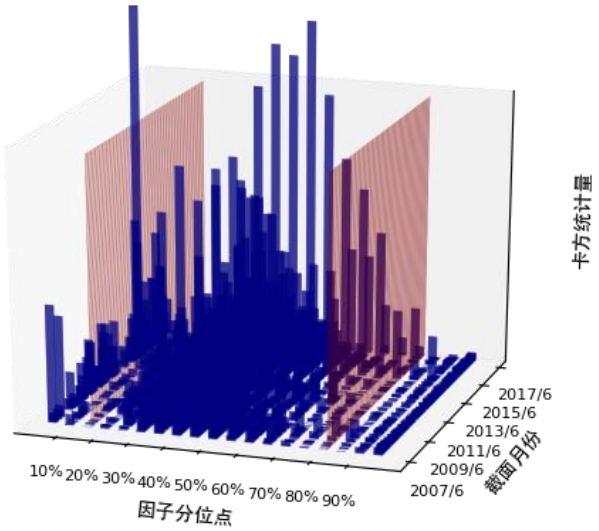
资料来源：招商证券、Wind 资讯

图 14 自由流通市值因子χ^2统计量分布（对标中证 500）



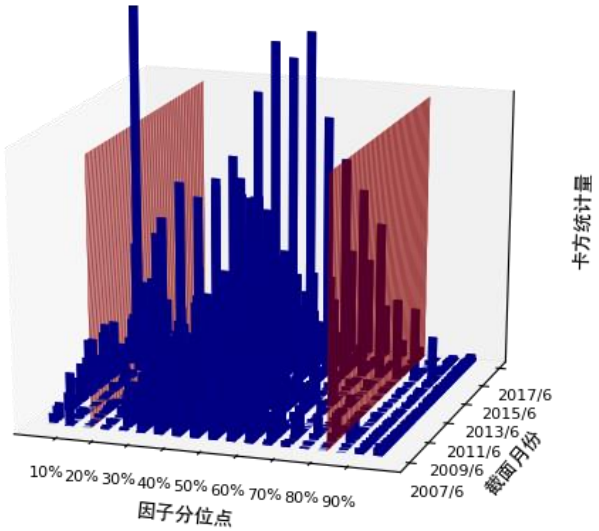
资料来源：招商证券、Wind 资讯

图 15 自由流通市值因子χ^2统计量分布（对标沪深 300）



资料来源：招商证券、Wind 资讯

图 16 自由流通市值因子χ^2统计量分布（对标上证 50）



资料来源：招商证券、Wind 资讯

图 8 至图 16 中，x 轴表示各分位点取值，y 轴表示不同估计截面，z 轴表示χ^2统计量取值。从上述三维条形图可见，三个市值的除了在5%和10%分位点处由于可观测样本数量过少而存在异常值以外，中间分位点部分（$[20\%, 80\%]$）对应卡方统计量明显高于两端部分（$[5\%, 20\%) \cup (80\%, 95\%]$)的卡方统计量。卡方独立性检验的结果展示出，市值因子相关度较高分位点区间是$[20\%, 80\%]$，故而，截取市值因子取值落在$[20\%, 80\%]$区间的个股进行单因子收益估计。

因子暴露标准化赋值

标准化赋值应在不失单调性的条件下使得各因子超额暴露的“量纲”相同,即同时满足单位化与对称性的要求,确保因子收益及波动具有可比性。设定标准化赋值的取值范围为$[-0.5, 0.5]$,将前序步骤处理后的因子超额暴露的取值区间切分为n个等长的子区间,对取值落在第l个子区间的数值。进行标准化赋值$\Delta\beta_{i,t_0}^{(j)} = -0.5 + (l-1)/(n-1)$。这种类似于分级靠档的标准化方法是对传统的中心化方法和打分法的折衷,保留了因子取值之间较大的差异,而将较小的差异进行抹平。本报告因子子区间数n分别取值 5、10、50、100、200 进行标准化赋值。

市值类因子收益估计

对因子超额暴露值完成前述步骤后,分别采用普通最小二乘法(OLS)和加权最小二乘法(WLS)对每个月度截面上的市值因子的单因子收益进行估计。加权最小二乘法的权重为$\omega_{i,t_0}^{WLS} = \left(\tilde{\sigma}\left[\varepsilon_{i,t\le t_0}\right]\sqrt{T_i(t_0,t_1)}\right)^{-1}$,其中,$\tilde{\sigma}\left[\varepsilon_{i,t\le t_0}\right]$为估计市场$\Delta\tilde{\beta}_{i,t_0}^{M}$时估计模型的残差。

我们从以下五个指标来评估单因子收益:

1. 因子收益t统计量的绝对值平均数。由于t统计量值越大,表明因子对于个股超额收益的影响越显著。
2. 因子收益t统计量大于 2 的截面占比。在样本量较多的情况下,当t大于 2 时,因子在5%显著性水平下对因子收益影响显著。
3. 因子收益均值。该值表明因子对收益的贡献程度。
4. 因子收益的波动率(年化)。该值表明因子对收益贡献的波动程度。
5. Sharpe 值:(因子收益/因子收益波动率)$\times\sqrt{12}$

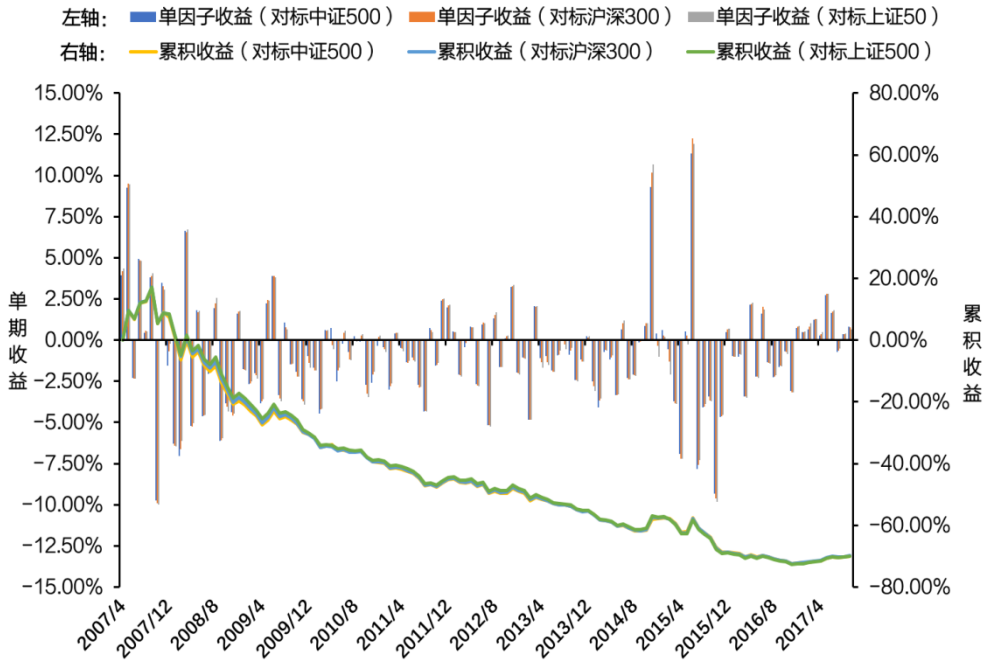
表 6：市值因子收益利率 5 指标汇总（子区间数 200）

估计方法	因子名称	投资基准	t 统计量均值	t 值大于 2 占比	因子收益均值	因子收益波动率（σ）	Sharpe 值	常数项均值
普通最小二乘回归（OLS）	合计市值	中证 500	2.2737	44.72%	-0.84%	11.85%	-0.8501	-0.63%
		沪深 300	2.2783	44.72%	-0.83%	11.95%	-0.8341	-0.40%
		上证 50	2.2735	43.90%	-0.83%	11.92%	-0.8301	-0.40%
	流通市值	中证 500	2.4946	56.91%	-1.02%	11.71%	-1.0416	-0.52%
		沪深 300	2.4985	57.72%	-1.01%	11.74%	-1.0345	-0.30%
		上证 50	2.4966	58.54%	-1.01%	11.71%	-1.0374	-0.29%
	自由流通市值	中证 500	2.3218	56.10%	-1.14%	10.67%	-1.2827	-0.56%
		沪深 300	2.3276	56.10%	-1.14%	10.70%	-1.2724	-0.33%
		上证 50	2.3242	56.10%	-1.13%	10.70%	-1.2671	-0.32%
加权最小二乘回归（WLS）	合计市值	中证 500	2.3776	52.85%	-0.90%	11.22%	-0.9586	-0.17%
		沪深 300	2.4209	52.85%	-0.89%	11.50%	-0.9254	0.03%
		上证 50	2.4331	52.03%	-0.89%	11.54%	-0.9272	0.00%
	流通市值	中证 500	2.6722	60.16%	-1.06%	11.19%	-1.1312	-0.06%
		沪深 300	2.7146	60.16%	-1.02%	11.43%	-1.0745	0.14%
		上证 50	2.7244	60.98%	-1.02%	11.54%	-1.0628	0.11%
	自由流通市值	中证 500	2.5379	57.72%	-1.26%	10.01%	-1.5124	-0.11%
		沪深 300	2.5959	60.16%	-1.25%	10.29%	-1.4509	0.09%
		上证 50	2.5773	59.35%	-1.23%	10.36%	-1.4203	0.06%

资料来源：招商证券、Wind 资讯

表 6 汇报了两种不同的估计方法下三个市值类因子对应不同基准的 5 指标数据(标准化赋值子区间数量 n 为 200)。从t统计量的均值和绝对值大于 2 的截面占比来看，流通市值因子略优于另两个因子。而从 Sharpe 值来看，自由流通市值要优于另两个因子。

图 17 合计市值单因子收益



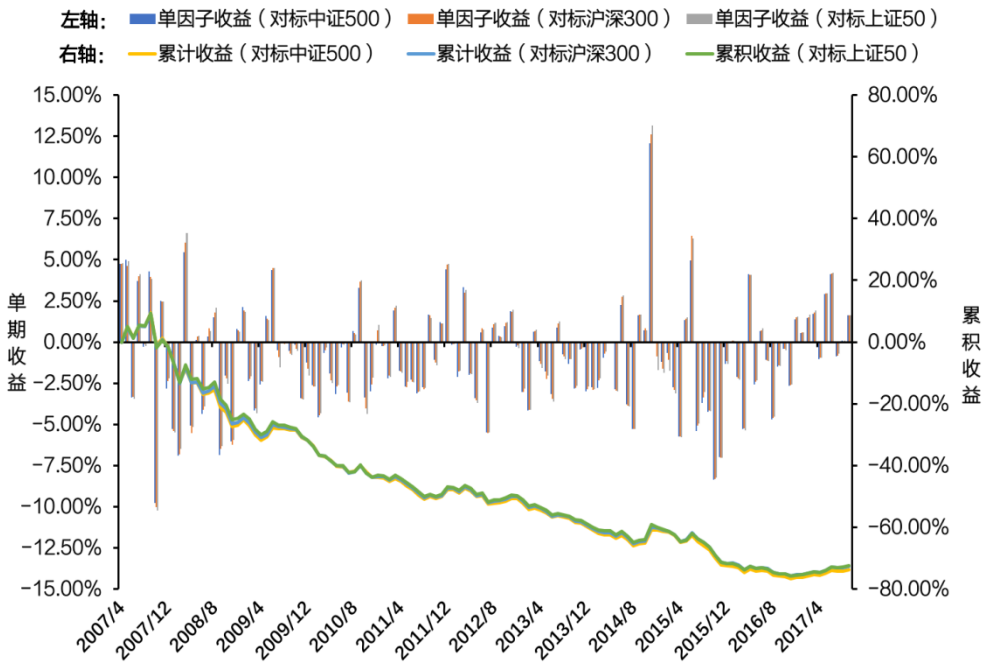
资料来源：招商证券、Wind 资讯

图 17 展示了合计市值因子的单期收益（条形图）和累积收益（折线图）走势。在 2007 至 2009 年间，合计市值因子收益正负向切换较为频繁；2009 年之后直到 2016 年 12 月份，在这很长的一段时间内，合计市值因子的收益基本为负，也就是在这些时段，小市值股票的走势要优于大市值股票走势；而 2016 年 12 月情况又发生了一次反转，市场偏向大市值股票，并一直维持到观测窗口期结束，除了在 2017 年 7 月份出现

了一次短暂的偏小盘行情。

图 17 中最近两次较大的正向收益分别发生在 2014 年 11 月和 2015 年 6 月。2014 年 11 月出现偏大市值的股票的原因是因为牛市开始时，券商股较快地拉升；而 2015 年 6 月偏大市值的原因是，当时股市大跌，而该期大市值股票的跌势要缓于小市值股票。

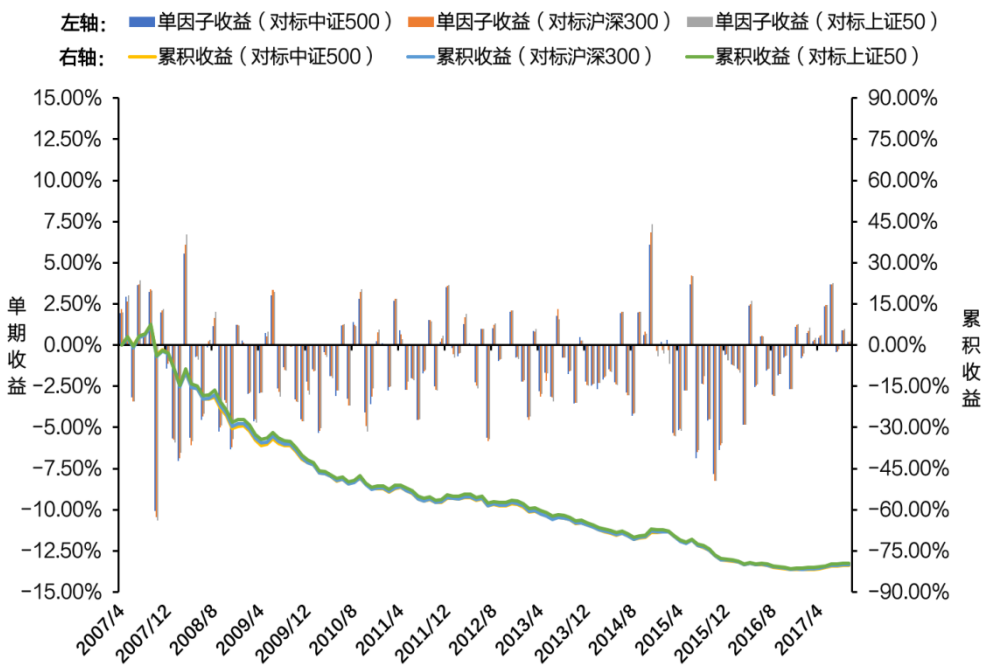
图 18 流通市值单因子收益



资料来源：招商证券、Wind 资讯

图 18 展示了流通市值因子的单期收益（条形图）和累积收益（折线图）走势。流通市值因子收益的走势和合计市值相似，在 2007 至 2009 年间，流通市值因子收益正负向切换较为频繁；2009 年之后直到 2016 年 12 月份，在这很长的一段时间内，流通市值因子的收益基本为负，也就是在这些时段，小市值股票的走势要优于大市值股票走势；而 2016 年 12 月情况又发生了一次反转，市场又偏向大市值股票。最近一次出现较大正收益的截面是 2014 年 11 月。

图 19 自由流通市值单因子收益



资料来源：招商证券、Wind 资讯

图 19 展示了自由流通市值因子的单期收益（条形图）和累积收益（折线图）走势。自由流通市值因子收益的走势和另外两个市值因子的收益走势相似，在 2007 至 2009 年间，自由流通市值因子收益正负向切换较为频繁；2009 年之后直到 2016 年 12 月份，在这很长的一段时间内，自由流通市值因子的收益基本为负；而 2016 年 12 月情况又发生了一次反转，市场又偏向大市值股票。总体来讲，自由流通市值的收益绝对值小于另两个因子。自由流通市值也在 2014 年的 11 月份和 2015 年的 6 月份出现了两次偏正的较大收益。

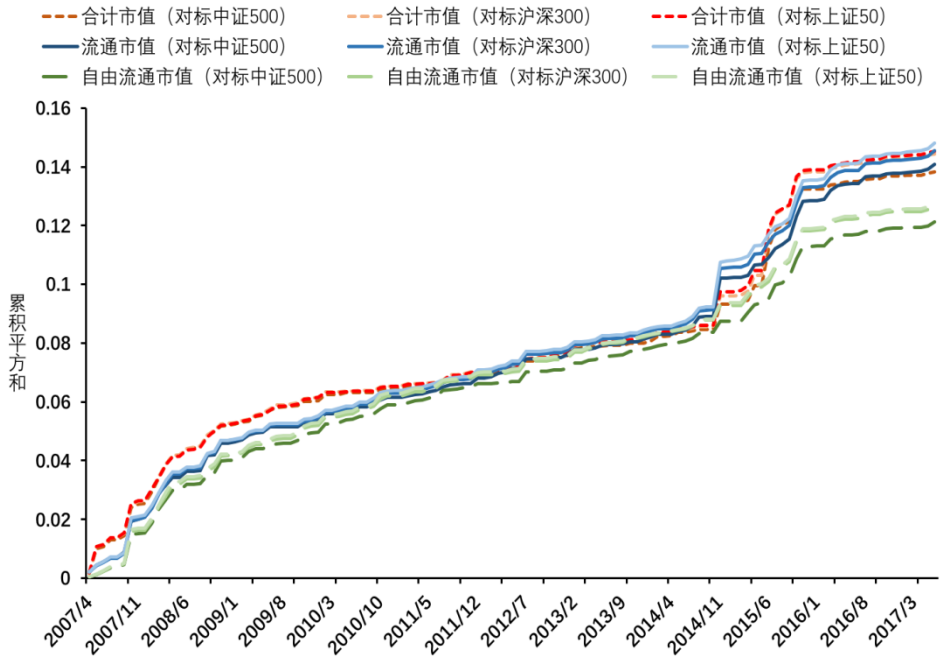
表 7：因子同向波动持续月份数统计

因子名称	合计市值			流通市值			自由流通市值		
对标基准	中证 500	沪深 300	上证 50	中证 500	沪深 300	上证 50	中证 500	沪深 300	上证 50
1 个月	28	32	33	25	25	25	25	24	24
2 个月	7	11	10	11	14	14	12	13	13
3 个月	10	8	7	5	6	6	8	9	9
4 个月	3	1	2	4	2	2	4	1	1
5 个月	4	4	4	3	3	3	2	2	2
6 个月	2	1	1	-	1	1	2	2	2
7 个月	1	1	1	-	-	-	-	-	-
8 个月	-	1	1	-	-	-	-	-	-
9 个月	-	-	-	1	-	-	1	1	1
10 个月	-	-	-	1	1	1	-	-	-
11 个月	-	-	-	1	-	-	1	1	1
12 个月	-	-	-	-	-	-	-	-	-
13 个月	-	-	-	-	1	1	-	-	-

资料来源：招商证券，Wind 资讯

分别统计了在不同基准下，三个因子的收益同向持续月份月数和对应出现的次数。流通市值收益持续性表现优于另两个因子。

图 20 因子波动量能

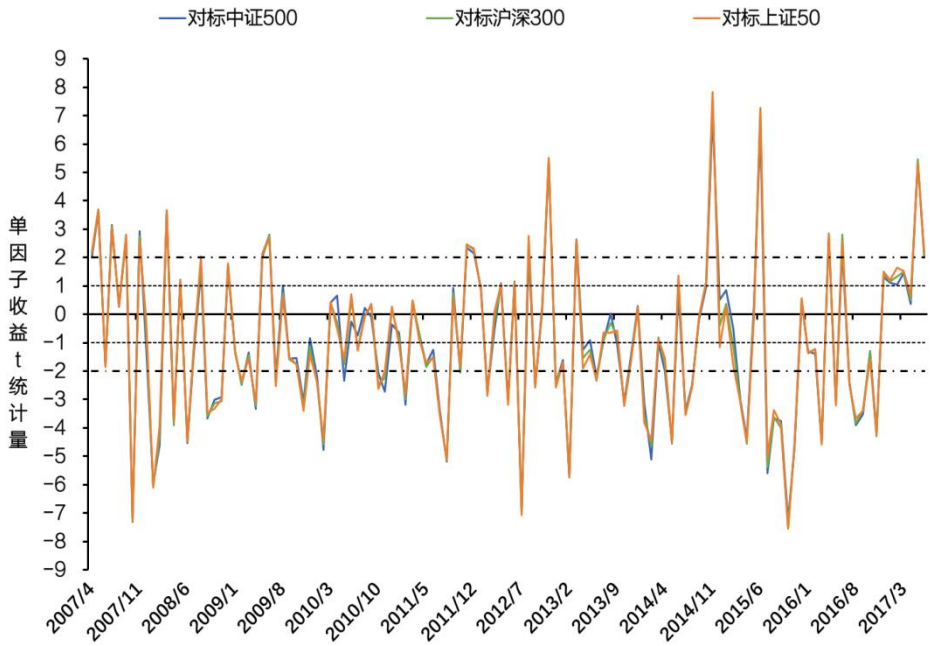


资料来源：招商证券、Wind 资讯

图 20 汇报了三个因子的波动量能。在挑选因子放入因子模型时，除了要考虑因子对于个股的区分度、因子持续性外，还要斟酌因子对于超额收益的贡献度。波动量能就是用以描述因子对超额收益贡献度的指标。我们用单因子单期收益的累积平方和来刻画因子的波动量能。

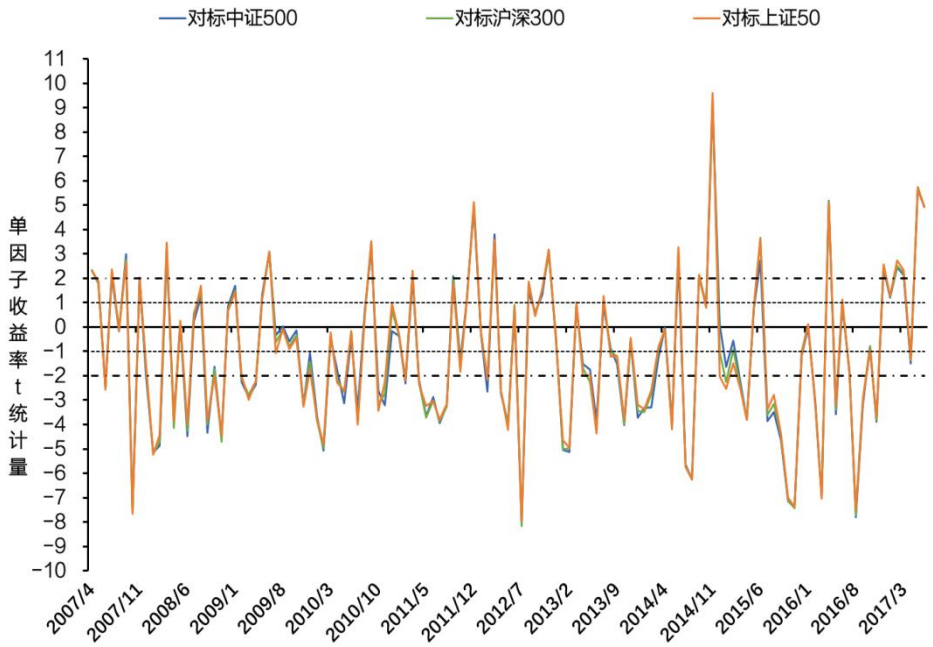
市值类因子的波动量都比较强。相比之下，合计市值和流通市值的波动量能略好于自由流通市值。在时间维度上，两段牛市时候的因子波动量能较强，而其他时段波动量能较为平缓。

图 21 合计市值因子收益 t 统计量走势



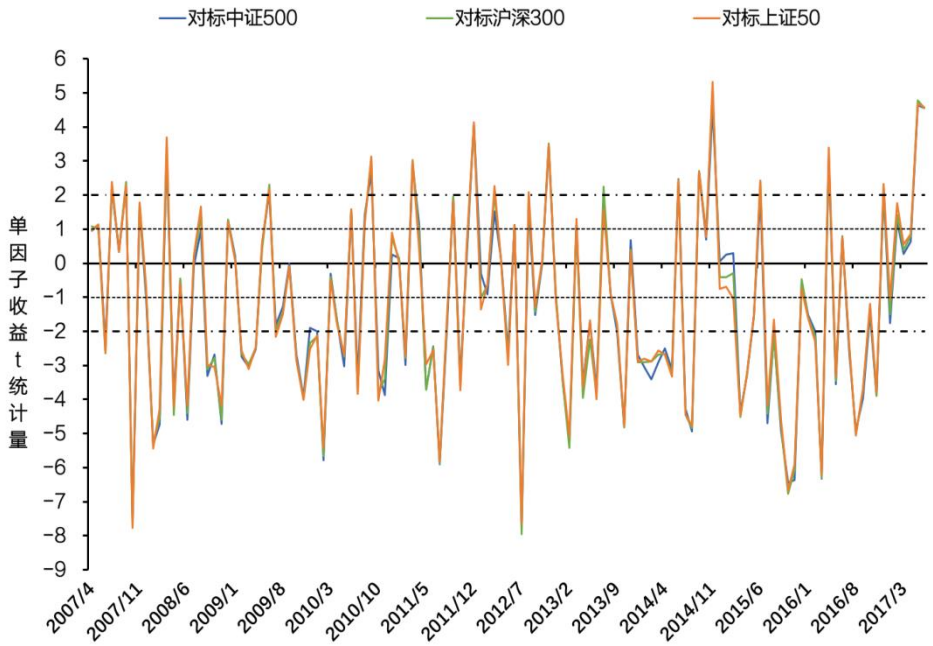
资料来源：招商证券、Wind 资讯

图 22 流通市值因子收益 t 统计量走势



资料来源：招商证券、Wind 资讯

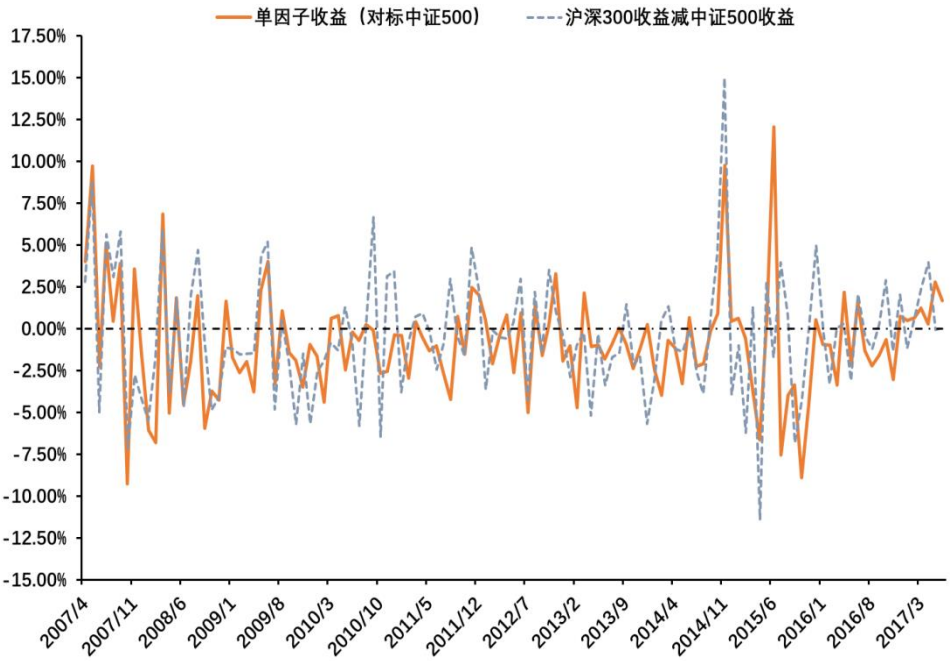
图 23 合计市值因子收益 t 统计量走势



资料来源：招商证券、Wind 资讯

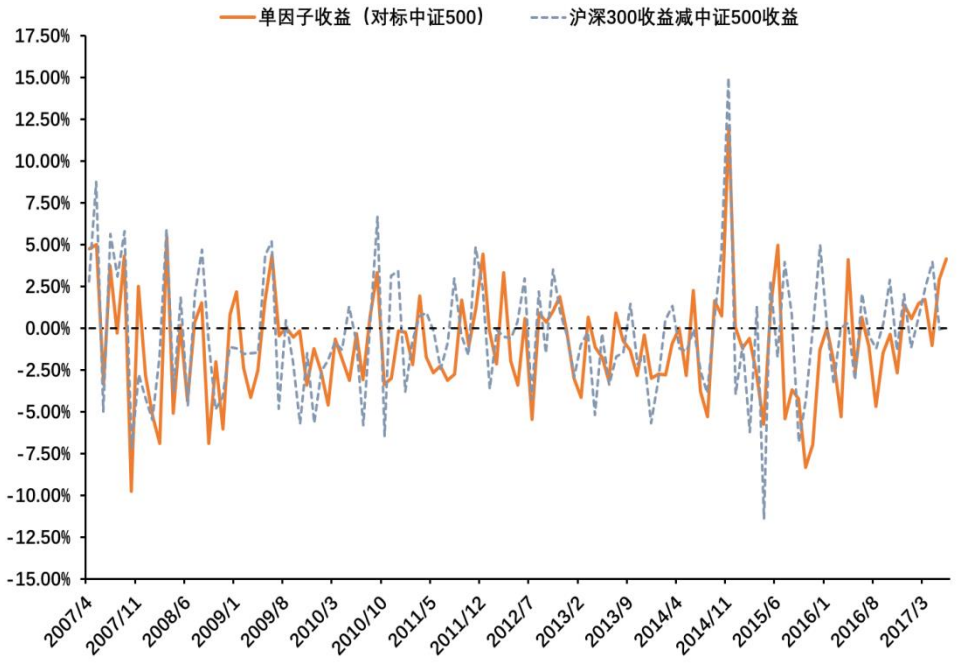
图 21 至图 23 展示了三个因子在进行单因子收益估计时的t统计量走势。80%以上的t统计量绝对值大于 2，只有合计市值和流通市值在 2010 年前后出现过短暂的 t 统计量绝对值小于 2 的时期。t统计量走势说明市值类因子的对于个股经系统性风险调整后的超额收益解释性较强。

图 24 合计市值因子(中证 500) V.S. 大小市值指数之差



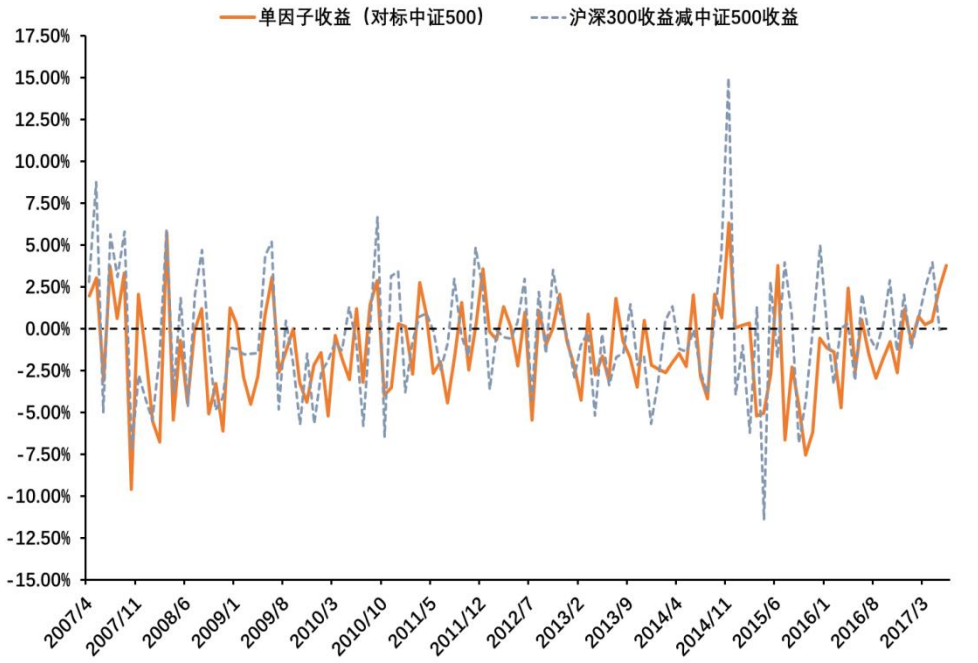
资料来源：招商证券、Wind 资讯

图 25 流通市值因子(中证 300) V.S. 大小市值指数之差



资料来源：招商证券、Wind 资讯

图 26 自由流通市值因子(上证 50) V.S. 大小市值指数之差



资料来源：招商证券、Wind 资讯

图 24 至图 26 展示了市值类因子收益走势和大小市值指数之差的走势拟情况，这里的市值类因子收益走势和大小市值指数之差是用沪深 300 的收益减去中证 500 的收益得到。在 2009 年之前的一段时期，市值类因子收益走势和大小市值指数之差的走势拟合度较好；而 2009 年之后拟合性较差。可能的原因是，在 2009 年之前，A 股市场个股数量较少，中证 500 对于小市值股票的代表性较好，但是随着时间的推移，A 股市场的股票越来越多，导致中证 500 对于小市值股票的代表性不似从前，时至今日，中证

500 更多代表中偏大市值的股票走势，故而后期拟合较差。

结论

本报告是招商金工因子模型系列的第二篇报告。本报告在第一篇因子模型报告的基础上，介绍了数据库搭建过程和模型被解释变量构建流程，并以市值类因子为例做了算法展示。整篇报告在整个系列报告中，起到了承上启下的作用。

在数据库搭建介绍中，报告着重说明了截面每日可计算条件，以及在该可计算条件下构建起来的等权市场组合和等权基准组合。

阐释了系统性风险估计参数$\Delta\tilde{\beta}_{i,t_0}^{M}$计算过程和模型被解释变量的构建流程。以市值类因子为例进行了算法展示，展示包括因子清洗、删失的步骤，独立性检验的结果，标准化赋值的特点；并以不同的估计方法估计了市值类因子的收益。详尽汇报了市值类因子收益估计时候的各类指标，包括单期收益、t统计量概况、Sharpe 值、同向波动持续月份统计、因子波动量能等。这些因子估计中得到的指标，将是我们后期选择因子入因子模型时的重要参考依据。

本报告为后续要做的其他各类因子收益估计做了算法上的演示，再后续的研究中，将用我们的因子模型框架，来对各大类因子进行收益估算、并统计比较各个指标的差异，继续推进我们的因子系列研究。

分析师承诺

负责本研究报告的每一位证券分析师，在此申明，本报告清晰、准确地反映了分析师本人的研究观点。本人薪酬的任何部分过去不曾与、现在不与，未来也将不会与本报告中的具体推荐或观点直接或间接相关。

叶涛：首席分析师。上海交通大学管理学硕士，2005 年起从事金融工程研究，曾先后任职于易方达基金机构投资部、上投摩根基金研究部、申万菱信基金投资管理总部、长江证券研究部、广发证券发展研究中心，2014 年 3 月加盟招商证券研究发展中心。

崔浩瀚：研究助理。浙江大学经济学硕士，2017 年 7 月加盟招商证券研究发展中心金融工程组。

投资评级定义

公司短期评级

以报告日起 6 个月内，公司股价相对同期市场基准（沪深 300 指数）的表现为标准：

强烈推荐：公司股价涨幅超基准指数 20%以上
审慎推荐：公司股价涨幅超基准指数 5-20%之间
中性：公司股价变动幅度相对基准指数介于±5%之间
回避：公司股价表现弱于基准指数 5%以上

公司长期评级

A：公司长期竞争力高于行业平均水平
B：公司长期竞争力与行业平均水平一致
C：公司长期竞争力低于行业平均水平

行业投资评级

以报告日起 6 个月内，行业指数相对于同期市场基准（沪深 300 指数）的表现为标准：

推荐：行业基本面向好，行业指数将跑赢基准指数
中性：行业基本面稳定，行业指数跟随基准指数
回避：行业基本面向淡，行业指数将跑输基准指数

重要声明