多因子选股模型

摘要

本文旨在构建并验证一个基于线性回归的多因子选股模型。文章首先介绍了实验的背景、理论基础和目标，随后详细阐述了模型的构建过程，包括线性回归模型的假设、因子的计算和选取方法，以及模型构建的具体步骤。在数值实验部分，文章展示了线性回归的各项假设测试结果、因子数据的计算与处理、因子分析结果，并提供了投资组合的模拟回测结果。通过对 16 个因子的单因子有效性分析和回测实验，文章筛选出 6 个表现较好的因子（mv 因子、bp 因子等）构建多因子资产定价模型，并进行了回测。最终，多因子模型显示出良好的年化收益率和夏普比率，验证了模型的有效性。文章最后对实验结果进行了总结，并对未来的研究方向提出了展望。

一、 引言

1.1 实验背景

随着金融市场的快速发展和投资者对风险管理需求的增加，量化投资策略逐渐受到重视。多因子选股模型作为量化投资中的一种重要策略，通过综合多个因子来预测股票的未来表现，旨在提高投资组合的收益并控制风险。本文旨在通过构建一个基于线性回归的多因子选股模型，来探索其在实际投资中的应用潜力。

1.2 理论基础

多因子模型的理论基础主要来源于资本资产定价模型（CAPM）和套利定价理论（APT）。CAPM 认为，股票的预期收益率可以通过市场风险来解释，而 APT 则进一步扩展了这一理论，认为股票的预期收益率可以通过多个宏观经济因素来解释。本文采用的线性回归模型，正是基于这些理论，通过选取多个因子来构建预测模型。

1.3 实验目标

本文的主要目标是构建一个有效的多因子选股模型，并通过对历史数据的回测来验证模型的有效性。具体目标包括：

1、确定线性回归模型的适用性，并基于此构建多因子选股模型。

2、通过计算和分析多个因子，筛选出对股票未来表现有显著预测能力的因子。

3、构建基于选定因子的投资组合，并进行模拟回测，以评估模型的收益和风险。

4、基于回测结果，提出模型的优化方向和未来研究的可能路径。

通过实现上述目标，本文希望能够为投资者提供一个可靠的量化投资工具，以提高投资决策的科学性和有效性。

二、 模型构建

2.1 构建因子池

在模型构建的第一步，我们首先构建了一个包含多个潜在预测因子的因子池。这些因子涵盖了市场价值（mv）、账面市值比（bp）以及其他 14 个基于不同财务指标和市场行为的 alpha 因子。因子池的构建基于对市场数据的广泛研究和历史表现分析，旨在捕捉影响股票回报的各种潜在因素。

2.2 数据处理

在因子池构建完成后，我们对所选因子进行了数据处理。这包括数据清洗（去除缺失值和异常值）、标准化（使因子具有统一的量纲）和中性化（消除行业和市场的影响），数据处理的目的是为了确保因子数据的质量，提高模型的稳定性和预测能力。本文尝试对不同数据采用不同处理方式，选择使得单因子回测实验中最有效的处理方式组合作为因子的处理结果。

2.3 因子选择

因子选择是基于信息系数（IC）值和单因子回测结果进行的。IC 值衡量了因子与未来回报之间的相关性，而单因子回测则评估了因子在实际投资中的有效性。通过计算每个因子的 IC 值和回测结果，我们筛选出了 6 个有效的因子，这些因子在预测股票回报方面显示出了较高的稳定性和预测能力。

2.4 OLS 回归假设检验

在因子选择完成后，我们对这 6 个因子进行了普通最小二乘（OLS）回归分析。在进行回归分析之前，我们对模型的假设进行了检验，包括线性关系、误差项的独立性、同方差性和正态分布。这些假设检验确保了 OLS 回归模型的有效性和可靠性。

2.5 建立多元线性回归模型

基于通过假设检验的因子，我们建立了一个多元线性回归模型。模型的形式如下：

$$R = \beta_0 + \beta_1 \cdot F_1 + \beta_2 \cdot F_2 + \cdots + \beta_6 \cdot F_6 + \varepsilon$$

其中，R 代表股票的预期回报，$F_1, F_2, \cdots, F_6$代表选定的 6 个因子，$\beta_0, \beta_2, \cdots, \beta_6$ 是模型参数，ε是误差项。

2.6 回测实验

我们使用历史数据对建立的多元线性回归模型进行了回测。回测的频率设置为每周，每次回测都会根据模型预测的超额收益率选出预期表现最好的前 30 支股票，并进行等权重配置。回测过程中，我们记录了包括年化收益率、年化波动率、最大回撤率和夏普比率在内的多个性能指标。

2.7 结果分析

通过对回测结果的分析，我们评估了模型的预测能力和风险控制能力。年化收益率和夏普比率的计算结果表明了模型的收益能力，而年化波动率和最大回撤率则反映了模型的风险水平。通过对这些指标的综合分析，我们得出了模型的有效性，并提出了可能的改进方向。

通过上述步骤，我们成功构建并验证了一个基于多元线性回归的多因子选股模型，为投资者提供了一个科学、系统的投资决策工具。

三、 数值实验

3.1 构建因子池

本文选取了 mv 市场价值因子和 bp 账面市值比因子，并根据参考文献[1]，主观选取如下表 alpha_001 至 alpha_014 共计 14 个基于不同财务指标和市场行为的 alpha 因子构建了因子池，具体因子计算公式展示于附录。

本文针对各种因子，提供了四种处理方式选择：0-1 标准化、正态标准化、Box-Cox 变换以及行业中性化。每种处理方式都有其特定的适用场景和优势，我们对每个因子独立选择最合适的处理方式，并剔除了 1 年内的股票数据以避免近期数据对模型的影响。

在构建因子池的过程中，我们特别剔除了 1 年内的股票数据。这一步骤是为了减少模型过拟合的风险，并确保模型能够捕捉到长期有效的因子效应。通过剔除近期数据，我们可以让模型更加关注那些在长期内稳定有效的因子。

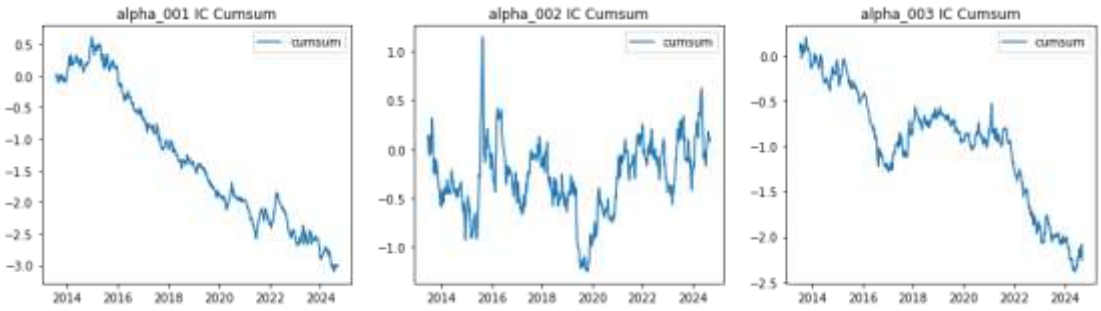
3.2 单因子模型结果分析

回测实验一共选取了包括 mv 因子、bp 因子在内的 16 个因子，计算 IC、IR 值（Pearson 相关系数下），结果如下表所示：

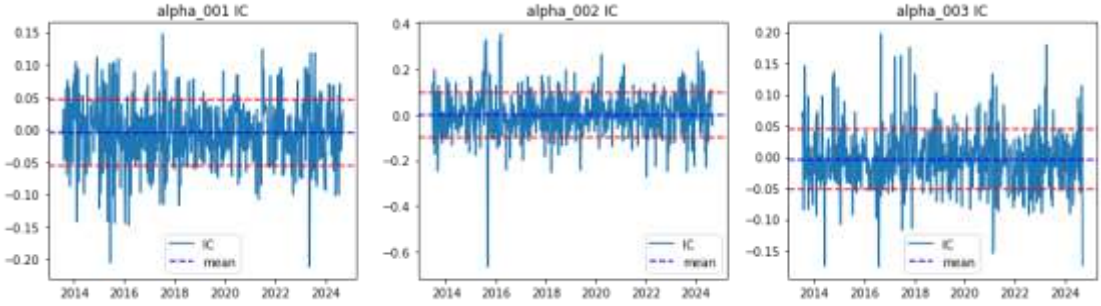
Factor	mean	std	|IC|>0.03	IR
mv	-0.003866	0.108542	77.14%	-0.035616
bp	0.017699	0.078324	68.77%	0.225972
alpha_001	-0.006162	0.050237	52.79%	-0.122648
alpha_002	0.000910	0.097803	67.74%	0.009309
alpha_003	-0.004191	0.047340	44.42%	-0.088535
alpha_004	-0.064599	0.110264	80.67%	-0.585856
alpha_005	-0.008620	0.063379	61.52%	-0.136007
alpha_006	-0.018659	0.106975	76.95%	-0.174425
alpha_007	-0.011054	0.074123	63.94%	-0.149125
alpha_008	0.018469	0.076879	68.77%	0.240237
alpha_009	-0.011853	0.069947	62.45%	-0.169464
alpha_010	-0.029458	0.081090	72.30%	-0.363276

alpha_011	-0.028602	0.090092	74.91%	-0.317480
alpha_012	-0.015526	0.096258	77.14%	-0.161294
alpha_013	-0.020996	0.070389	65.61%	-0.298294
alpha_014	-0.021820	0.105735	79.55%	-0.206367

经过计算因子 Rank IC 值以及相关指标，得到 16 个因子的单因子有效性分析结果如上表所示。其中因子 alpha_001、alpha_003 的 IC 的绝对值大于 0.03 的比例较少，因子 alpha_002 的信息比率 IR 的绝对值过小，绘制出这三个因子的累计曲线以及三个因子的分布情况如下图所示：



因子 alpha_002 的 IC 累积曲线趋势上呈现水平波动，说明因子具有一定预测能力，但稳定性较差；因子 alpha_001、alpha_003 的 IC 累积曲线开始为正，后长期为负值，说明因子可能已经失效，或者与未来回报之间的相关性变差，需要重新评估。



对十个因子分别进行单因子回测，使用最小二乘法拟合一元线性回归模型，预测下一期超额收益率，以周为频率，选取预期超额收益率最大的前 30 支股票作为下期的持仓，采样等权配比，使用 OLS 进行单一因子回测实验，回测结果如下表所示：

Factor	年化收益率	年化波动率	最大回测	夏普比率
mv	-18.23%	46.52%	297.47%	-0.3918
bp	-22.37%	18.45%	-2692.83%	-1.2125
alpha_001	-11.71%	22.28%	-696.75%	-0.5255
alpha_002	76.35%	28.91%	-25.87%	2.6406
alpha_003	-29.88%	22.55%	-1727.95%	-1.3248

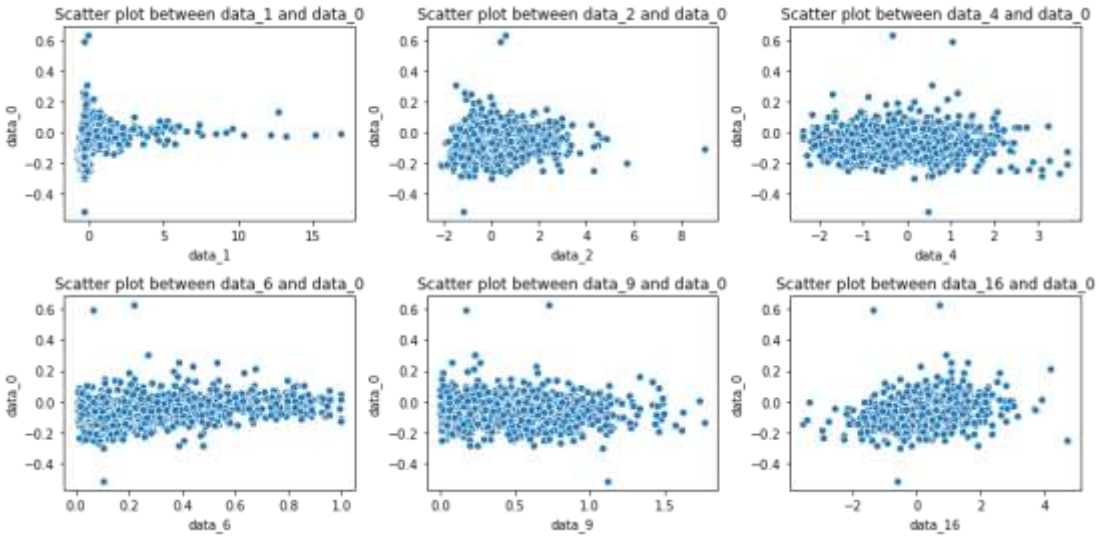
alpha_004	88.50%	27.74%	-229.79%	3.1900
alpha_005	-27.10	19.95%	-563.88%	-1.5121
alpha_006	-30.16%	22.66%	-1898.16%	-1.5809
alpha_007	112.31%	31.60%	-195.69%	3.5541
alpha_008	-50.92%	25.48%	-4468.90%	-1.9980
alpha_009	-20.17%	21.52%	-2594.18%	-0.9373
alpha_010	-57.03%	43.98%	-1955.57%	-1.2967
alpha_011	-41.25%	20.59%	-2798.25%	-2.0037
alpha_012	-0.65%	32.78%	-3733.77%	-0.0198
alpha_013	-37.04%	26.77%	-2019.08%	-1.3833
alpha_014	12.83%	42.57%	-2622.79%	0.3013

可以得知，因子 alpha_001、alpha_003、alpha_005、alpha_006 等因子的年化收益率为负值，且 alpha_010、alpha_014 因子的年化波动率较高，说明因子的稳定性较差，这些因子的期望收益率较低；alpha_002 、alpha_004 年化收益率和波动率表现较好，但是最大回撤率非常高，说明该因子具有收益能力，但是抗风险能力较弱，容易受到异常情况和突发情况的影响，可以作为多因子选股模型中的一个因子；alpha_007 因子收益率和回测表现良好，但是年化波动率较高，说明因子稳定性较差。

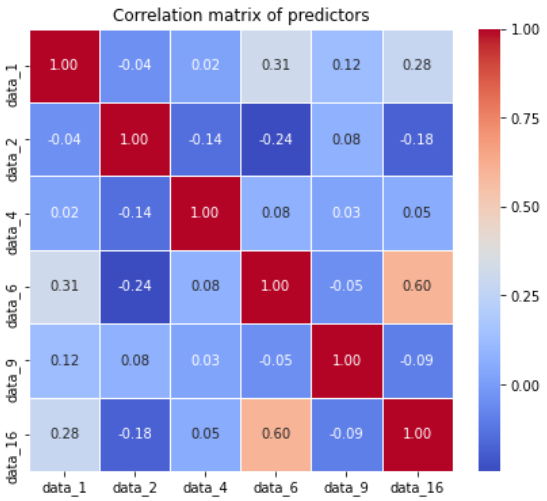
综合以上分析，选取了 mv 因子、bp 因子以及 4 个有效的 alpha 因子：alpha_002、alpha_004、alpha_007、alpha_014，进行下一步操作。

3.3 OLS 回归假设检验

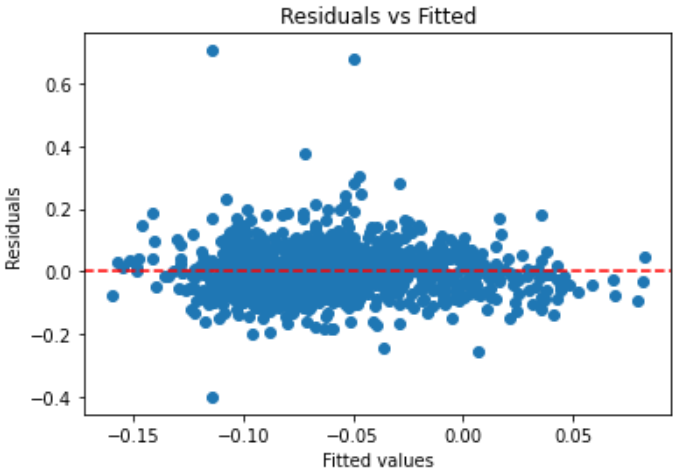
1、线性性，绘制散点图如下，表现不够明显。



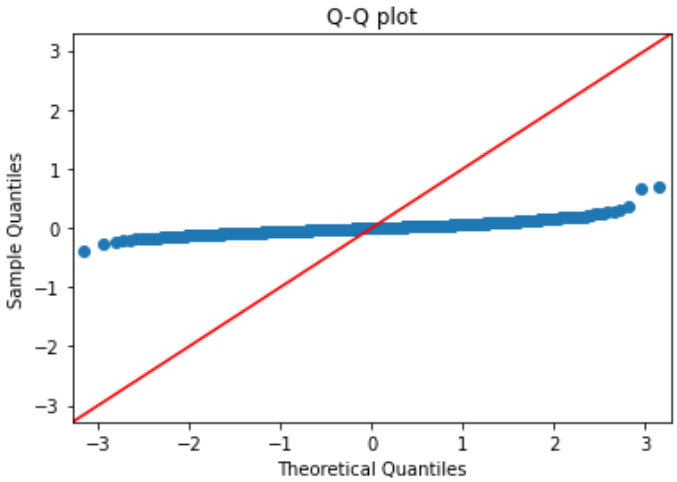
2、自变量独立性，绘制相关系数热力图如下，没有明显的相关性。



3、同方差性，绘制多元回归残差图如下，除去异常值，没有明显的异方差性。



4、正态性，绘制残差 Q-Q 图如下，残差可能不满足正态性假设。



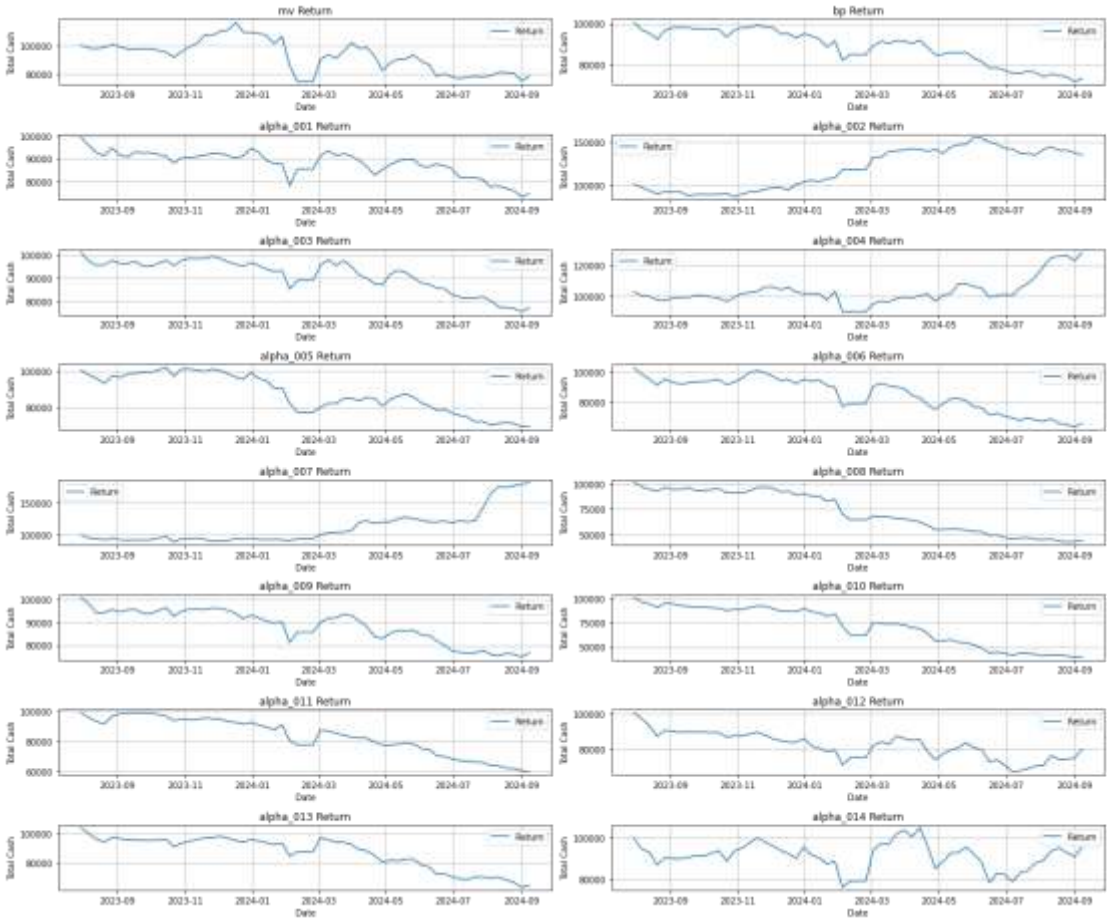
继续进行 Shapiro-Wilk 检验，计算得检验 p 值为$6.34 \times 10^{-27} \ll 0.05$，认为数据不服从正态分布。

5、多重共线性，计算 VIF 值如下，VIF 值均小于 10，认为不存在多重共线性。

Variable	VIF
const	5.6133
data_1	1.1268
data_2	1.1199
data_4	1.0361
data_6	1.6850
data_9	1.0334
data_16	1.6117

3.4 多因子模型结果分析

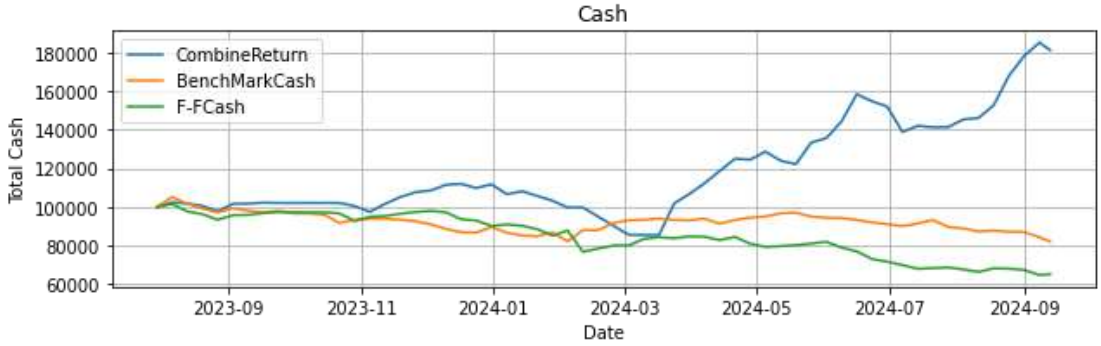
设定各个因子初始总资金为 100000 元，采样等比例配置股票各股数，一周调整一次仓位，采用满仓操作，回测资金变化图如下所示：



综合以上结果，决定选取因子 mv、bp、alpha_001、alpha_002、alpha_007、alpha_014，共记 6 个因子构建多因子资产定价模型，采用与单因子回测实验相同的条件以及资产配置方案建立回测模型。多因子回测结果如下表所示：

Factor	年化收益率	年化波动率	最大回测	夏普比率
	85.66%	35.11%	-2.2221	2.4399

回测资金变化图如下所示：



可以看到组合收益能够超过基准收益与简化 Fama-French 三因子模型，证明组合有一定的盈利能力；但是在 2024 年 3 月出现异常持续损失，说明模型仍有需要改进的地方，需要在未来进一步深化。

四、 总结与展望

4.1 总结

本文通过构建和验证一个基于线性回归的多因子选股模型，对量化投资策略在实际金融市场中的应用潜力进行了深入研究。文章首先介绍了实验的背景、理论基础和目标，然后详细阐述了模型的构建过程，包括线性回归模型的假设、因子的计算和选取方法，以及模型构建的具体步骤。在数值实验部分，文章展示了线性回归的各项假设测试结果、因子数据的计算与处理、因子分析结果，并提供了投资组合的模拟回测结果。通过对 16 个因子的单因子有效性分析和回测实验，文章筛选出 6 个表现较好的因子构建多因子资产定价模型，并进行了回测。最终，多因子模型显示出良好的年化收益率和夏普比率，验证了模型的有效性。

4.2 展望

尽管本文的多因子选股模型在回测中显示出了较好的性能，但仍存在一些局限性和改进空间。未来的研究可以在以下几个方向进行拓展和深化：

1、因子优化：进一步研究和开发新的因子，尤其是那些能够捕捉市场非线性特征和动态变化的因子，以提高模型的预测能力和稳健性。

2、模型改进：探索更先进的统计和机器学习方法，如随机森林、支持向量机或神经网络，以提高模型的预测精度和适应不同市场条件的能力。

3、风险管理：加强对投资组合风险的管理和控制，通过引入风险平价、最大回撤限制等技术，提高投资组合在不同市场环境下的稳定性。

通过这些研究方向的深入探索，可以进一步提升多因子选股模型的实用性和有效性，为投资者提供更加科学和系统的投资决策工具。

五、附录

因子计算公式

Factor	Formula
mv	总市值
bp	1/市净率
alpha_001 #1	(-1*CORR(RANK(DELTA(LOG(VOLUME), 1)), RANK(((CLOSE - OPEN) / OPEN)), 6))
alpha_002 #2	(-1 * DELTA((((CLOSE - LOW) - (HIGH - CLOSE)) / (HIGH - LOW)), 1))
alpha_003 #3	SUM((CLOSE=DELAY(CLOSE,1)?0:CLOSE-(CLOSE>DELAY(CLOSE,1)?MIN(LOW,DELAY(CLOSE,1)):MAX(HIGH,DELAY(CLOSE,1)))),6)
alpha_004 #4	((((SUM(CLOSE, 8) / 8) + STD(CLOSE, 8)) < (SUM(CLOSE, 2) / 2)) ? (-1 * 1) :(((SUM(CLOSE, 2) / 2) <((SUM(CLOSE, 8) / 8) - STD(CLOSE, 8))) ? 1 : (((1 <(VOLUME / MEAN(VOLUME,20))) || ((VOLUME /MEAN(VOLUME,20)) ==1)) ? 1 : (-1 * 1))))
alpha_005 #5	(-1 * TSMAX(CORR(TSRANK(VOLUME, 5), TSRANK(HIGH, 5), 5), 3))
alpha_006 #6	(RANK(SIGN(DELTA((((OPEN * 0.85) + (HIGH * 0.15))), 4)))* -1)
alpha_007 #7	((RANK(MAX((VWAP - CLOSE), 3)) + RANK(MIN((VWAP - CLOSE), 3))) *RANK(DELTA(VOLUME, 3)))
alpha_008 #8	RANK(DELTA(((((HIGH + LOW) / 2) * 0.2) + (VWAP * 0.8)), 4) * -1)
alpha_009 #24	SMA(CLOSE-DELAY(CLOSE,5),5,1)
alpha_010 #29	(CLOSE-DELAY(CLOSE,6))/DELAY(CLOSE,6)*VOLUME
alpha_011 #17	RANK((VWAP - MAX(VWAP, 15)))^DELTA(CLOSE, 5)
alpha_012 #135	SMA(DELAY(CLOSE/DELAY(CLOSE,20),1),20,1)

alpha_013 #40	SUM((CLOSE>DELAY(CLOSE,1)?VOLUME:0),26)/SUM((CLOSE<=DELAY(CLOSE,1)?VOLUME:0),26)*100
alpha_014 #122	(SMA(SMA(SMA(LOG(CLOSE),13,2),13,2),13,2)-DELAY(SMA(SMA(SMA(LOG(CLOSE),13,2),13,2),13,2),1))/DELAY(SMA(SMA(SMA(LOG(CLOSE),13,2),13,2),13,2),1)

六、参考文献

[1]. 李辰, & 刘富兵. (2017). 基于短周期价量特征的多因子选股体系. 国泰君安证券研究.