学号 2017302010265

密级

武汉大学本科毕业论文

基于统计的基本面研究：一种机器学习的视角

院（系）名　称：经济与管理学院

专　业　名　称：金融工程

学　生　姓　名：鲍余薇

指　导　教　师：李斌　教授

二〇二一年五月

郑 重 声 明

本人呈交的学位论文，是在导师的指导下，独立进行研究工作所取得的成果，所有数据、图片资料真实可靠。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确的方式标明。本学位论文的知识产权归属于培养单位。

本人签名：_____ 日期：_____

摘　　要

传统的基本面分析方法较为受限。一方面，需要主观预测股票内在价格，其常见计算方法包括现金流贴现法、相对价值法、经济附加值法、实物期权法等。这些方法虽然起到了一定的预测作用，但由于选择的指标过少，或太过于依赖对未来的主观预测结果，亦或计算方法过于程式化等原因，很难排除数据窥视的影响。另一方面，目前研究中可供选择的因子种类较少，难以做到投资策略的推陈出新。

相较而言，本文创新采用大数据统计方法，基于中国 A 股市场 3300 支股票自 2002 年初到 2019 年底的季度财务报表数据，选取 51 个财务指标，采用机器学习中的提升回归树 BRTs 算法对公司内在价值进行回归预测。对比于传统的线性回归方法，机器学习方法能够有效处理高维度的因子数据，挖掘出变量之间的非线性关系，在现有应用中表现良好。

本文将公司市场价值与内在价值的偏差构建为错误定价因子，根据错误定价因子将公司进行分类，据此构建相应的投资组合。实证结果表明，该投资策略构建的多头和多空投资组合能够获得显著的正月度平均收益和风险调节收益，证明了运用错误定价因子进行资产定价与配置的有效性。后续稳健型检验反映了股票市场价值向内在价值的趋近，证明超额收益并非来源于风险因子的遗漏。

本文结论有力支持了基本面分析之于中国股市证券分析的地位，并促进了机器学习与经济学研究的交叉融合。

关键词：基本面分析；错误定价因子；机器学习；提升回归树

ABSTRACT

In traditional fundamental analysis, we have to subjectively predict the intrinsic price of stocks, the calculation methods include Free Cash Flow Valuation (FCFF), relative value assessment, EVA, real option method, etc. Although these methods have played a certain role in forecasting, it is difficult to rule out the influence of data snooping due to the selection of too few indicators, or too much dependence on the subjective prediction results of the future, or the stylized calculation method. Also, there are too few factors to choose in the current research that it is difficult to innovate investment strategies.

Compared to traditional fundamental analysis, we innovatively use statistical model based on big data. Focus on the quarterly financial statements data of 3,300 stocks in China's A-share market from the beginning of year 2002 to the end of year 2019, 51 financial indicators are selected to predict on the company's intrinsic value using boosted regression trees (BRTs) in machine learning. Compared to linear regression, machine learning methods can effectively process high-dimensional factor data and can dig out the nonlinear relationship between variables, which performs well in existing applications.

We define the deviation between the company's market value and the intrinsic value as a mispricing factor, and classify companies according to it, finally construct the corresponding portfolio accordingly. The empirical results show that the long position and the long-short portfolio constructed by this above strategy can obtain significant positive monthly average returns and risk-adjusted returns, which proves the effectiveness of using the mispricing factor for asset pricing and allocation. The follow-up robust test reflects the gradual approach of the stock market value to the intrinsic value, which proves that the excess return does not come from the omission of risk factors.

The conclusions of this article strongly support the status of fundamental analysis in China's stock market and securities analysis, and promote the intersection of machine learning and economic research.

Key words: Quantamental Analysis; Mispricing Factor; Machine Learning; Boosted Regression Trees

目　　录

图片索引

表格索引

1 引言

有效市场假说（Efficient Markets Hypothesis，EMH）是现代金融学的理论基石之一。Fama（1965）首次提出了“有效市场”的概念，并定义为：如果在一个证券市场中，价格已经完全反映了所有可以获得的信息，那么就称这样的市场为有效市场。他认为市场上存在着大量理性的投资者，在任意时点都有无数人正在搜寻线索去精准预测股票未来的价格，为了获取利益而去对股票进行低买高卖，或许对于个人而言仅仅是超额利润的攫取，但在宏观维度上，正是这种高量级的同质活动快速推进着股票市场价格向正确价格的趋近，所有相关的信息已经完全展现在了股票价格中，任何为了预测股价而付出的时间、金钱和努力都是徒劳无功。进而 Fama（1970）在《Journal of Finance》杂志上发表了最具划时代意义的论文《Efficient Capital Markets: A Review of Theory and Empirical Work》，自此正式提出了有效市场假说，并对证券市场的信息分为以下三类：一是与交易相关的历史信息，例如历史成交量、价格等；二是当前的公开信息，如公司财报、分红报告等；三是内部信息。Fama 按照以上三种信息把有效市场分为弱式有效、半强式有效和强式有效三类，其中弱式有效证明了技术分析的无效性；半强式有效排除了基本面分析的作用；而强式有效意味着股票价格已经反映了所有信息，投资者不可能战胜市场。至此确立了“有效市场假说”在现代金融领域的基础性地位。

但是现在越来越多的实证研究表明了中国股票市场的非有效性。自 70 年代开始，随着科技的发展，已经持续识别出了能够提供超额收益的异象（因子）。曾有研究发现了股票的一些异常情况，例如价格的过度波动和可预测性（Shiller，1980）、价格走势逆转（Shleifer et al.，1994）等；在投资组合领域，也发现了规模效应（Bamber et al.，1997）、市盈率效应（Banz，1981）等，前美国金融学会主席 Cochrane（2011）称其为“因子动物园”（Factor Zoo），这些异象说明投资者能够根据基本面等公开信息去预测股票价格，向有效市场假说发起了挑战。

另外，从行为金融角度看，有很多学者指出有效市场假说理论的逻辑存在问题：如果股票价格已经完全反映市场中的信息，那理性投资者在后续的投资活动中就无需花费精力去搜集信息、预测股价了，但这却与市场需要信息才能有效运行相悖。同样，其设置的完美市场假设的条件与实际相悖，在真实的投资活动中，投资者较多地表现为“非理性”，从而使得其决策行为偏离了金融理论所要求和预

测的标准结果，例如羊群效应、月末效应、规模效应和股权溢价之谜等等这些金融异象都被认为是投资者非理性决策行为的主要表现（张元鹏，2015）。虽然行为金融文献并没有完全否定有效市场的界定，但行为金融派的支持者更愿将有效市场看作一种理想状态、不受个人意志（效用和偏好）影响的市场状态（丁志国 等，2017）。

从以上几个方面来看，我国股票市场未能做到完全有效，虽然 A 股市场现已发展成为全球第二大股票市场，总市值近 80 万亿元。但由于成立时间晚、相关制度不完善、散户比例大、交易成本高等问题，导致了定价效率的低下。这时基本面分析更能捕捉市场的非有效性，有效预测未来盈余，带来显著的超额回报（汪荣飞 等，2018）。相对于技术分析依靠图表去预测价格趋势，基本面分析更为关注公司的基本信息，主要包括财务报表或非财务上的信息，并从中评估公司的内在价值。本文将股票价格与内在价格的偏离定义为错误定价，而这种错误定价便是投资者所追捧的超额收益，所以对于错误定价的研究显得极为重要。

由于股票价格是已知的，那么对错误定价的研究关键在于对股票内在价格、公司内在价值的预测。现有研究中常见的计算方法包括现金流贴现法（Free Cash Flow Valuation，FCFF）；相对价值法，即与同行业的相似公司对比，以它们的平均市盈率、市净率及市销率等指标来计算；经济附加值法（Economic Value Added，EVA）；实物期权法等。这些方法虽然起到了一定的预测作用，但由于其选择的指标过少，或太过于依赖对未来的主观预测结果，亦或计算方法过于程式化等原因，很难排除数据窥视（Data snooping）的影响。

近年来，“金融科技”概念异军突起，银行需要和互联网金融巨头进行竞争，只能借助科技进行转型，金融服务逐渐“线上化”；市场要突破传统的金融方式，实现信息安全，需要区块链技术等进行加密等等，金融与 IT 技术的结合已然成为未来的大趋势。而机器学习是计算机科学领域重要的基础学科之一，主要通过人工智能方法模拟人类的行为，去对数据进行学习，并在一次又一次的重复试验中累积经验，逐渐优化算法，让结果更具有效性。随着科学技术的高速发展，计算机的计算效率显著提升，其所能处理的数据以及算法越来越多。机器学习因其能够突破人类计算能力的特点，重新向大家定义了科学研究的深度和广度，不仅局限于计算机领域，各行各业均需要借助其力量进行更为复杂的分析活动。在金融领域，随着中国证券市场的发展、投资者行为的多样化等因素，已很难去进行大规模的金融数据分析，而机器学习则可以做到这一点。

在学术研究中，近年来越来越多学者运用机器学习方法，去关注金融市场预测、资产定价等问题，并取得了丰厚的成果；在现实交易中，量化投资者以数据模型为核心，以程序化交易为手段，以追求绝对收益为目标进行投资，大多采用机器学习算法去实现交易思想。

相比于以往金融研究中运用的传统计量方法，机器学习方法有着以下优势：

- 对比于传统算法，机器学习方法是一种通过模仿人类的学习行为，逐渐累积经验的过程，每次学习过程中，数据预测能力的提升均来自于客观标准，而非人工干预，避免了数据窥视的影响，对于实证结果更具有可解释性。
- 机器学习方法对于数据的接受度更高，在各路专家的集中努力下，机器学习模型有了多个实现算法，且愈发优化，能够高效处理维度更大、结构更复杂的数据，甚至对于以往难以处理的定性文本数据都能简单处理。
- 金融数据适合用机器学习方法处理，金融市场也多表现为动态系统。例如在时间序列数据中，大多数金融属性并不符合线性关系，反而更多呈现为非线性、高维度和噪声性质，很难通过传统的计量方法去解决[11]。但机器学习算法高度灵活，不会预先在主观上对函数形式、变量分布等条件进行假设。而且，很多有关价格预测、资产定价的金融问题，都是涉及到对历史数据的分析，并在准确预测的基础上构建投资组合，数据量较大。

同样，在现有的基本面分析研究中，因子的选择较少，交易策略较为局限。为了突破传统因子研究的局限性，本文创新采用财务指标去构建错误定价因子 $\boldsymbol{M}$，并以此进行资产定价与配置。该因子的构建不参杂任何的主观因素，均通过客观标准去进行优化选择。

基于以上分析，本文采用机器学习方法——提升回归树 BRTs（Boosted Regression Trees）统计预测公司的内在价值。利用 2002 年初到 2019 年底的季度财务报表数据，对 A 股所有公司（剔除金融类股票、ST 股票）进行时间维度的截面回归：公司市场价值对于一系列财务指标进行非参数回归，这些财务指标的组合预测结果即为公司内在价值，回归残差为公司市场价值与内在价值的偏差，将偏差按照公司市场价值标准化后，成功构建出错误定价因子 $\boldsymbol{M}$（Mispricing Factor）。其中，回归系数将分期进行计算，避免了该期其他投资因素的影响；纳入 A 股所有公司，使回归结果更具有普遍性；且后续模型修订根据客观统计标准，排除了数据窥视的影响。

在成功引入错误定价因子 $\boldsymbol{M}$ 后，接下来便是检验其有效性。问题在于：一、

能否通过错误定价因子 M 构建低买高卖的投资组合获取超额收益，进而证明基本面分析的可行性？二、超额利润来自于错误定价还是因子遗漏，股票市场价格是否会向内在价格趋近？本文将针对以上两个问题进行研究。首先从 CSMAR 上搜集 A 股上市公司季度财务数据，通过 BRTs 算法计算每期公司的内在价值和错误定价因子 M；后采用 Fama-MacBeth 回归计算错误定价因子 M 系数 β，检验其是否显著；并且按照错误定价因子 M 将公司分为五组，从价值最被高估到最被低估，运用因子模型得超额收益 α，观察是否呈现单增趋势，以及构建多空投资组合，检验是否有显著为正的超额收益；最后观察长期时间内不同分组超额收益的变动趋势，证明超额收益并非来源于风险因子的遗漏。最后结论证明了错误定价因子 M 的有效性，有力支持了基本面分析之于中国股市证券分析的地位。

本文的研究有一定的现实意义与理论贡献：一、丰富了经济学和管理学领域的研究内容；二、丰富了基本面投资的理论和实践研究；三、丰富了中国股票市场有效性的研究。目前对于中国 A 股市场基本面分析的研究过少且不完善，统计之于数据分析的重要性不言而喻，其能够有效避免数据窥视的影响，是在此研究方向上的创新。同样，建模方法也有所创新，在线性回归的基础上增加了非线性维度的考量，采用机器学习算法去进行数据拟合，更具有信服度。

后文结构如下：第二部分回顾了中国股市有效性、基本面分析等相关文献，第三部分对本文采用的机器学习模型 BRTs 进行详细介绍，第四部分阐述本文的研究设计，介绍了数据和变量的构建方法；第五部分详细分析了本文的实证结果，第六部分为本文的研究结论。

2　文献综述

对于中国股票市场，已有多位学者研究证明其不具有弱势有效性。贾权 等（2003）对基于市场有效假设的 CAPM 模型以及其他因素与收益率之间的关系进行了实证检验，发现目前我国股市不满足市场有效性的假设，投资者的行为并不是完全理性的；吴振翔 等（2007）通过设计多种投资组合方式，发现在短期（3 个月以内）不能否定市场中无套利的假定；而对于中期和长期（6 个月及 12 个月）统计套利存在，说明我国 A 股股票市场的弱有效性不成立；Lim et al.（2009）运用非线性依赖检验的方法对 1999 年至 2005 年上证 A 股和 B 股的日数据进行检验，发现 A 股和 B 股均不具备弱势有效性；屈博 等（2014）利用 Q 统计量法、方差比检验法、广义谱检验、游程检验等多种方法对 2010 年至 2013 年沪深 300 股指期货的当月合约的 5 分钟高频数据进行检验，实证结果表明，随着考察期长度的增加，市场趋于拒绝弱式有效性。

上述研究说明，中国股市能够通过技术分析和基本面分析获取超额收益，实际上很多实证研究也证明了这个结论。对于技术分析，赵国顺（2009）基于时间序列方法，使用 GARCH 模型和 ARIMA 模型对股价波动趋势有着较好的短期预测效果；王劲松（2010）运用技术指标 MA、KDJ 和 MACD 证明在一定时间跨度上技术分析是有效的；石赛男（2011）的实证研究表明 MACD 指标对于 2004 年至 2009 年内的大、中、小盘股均有一定的预测能力；包懿（2015）基于 2010 到 2014 年三只指数标的和一只个股标的，通过持续持有策略和均线穿越法则策略获取了超额收益。

对于基本面分析，张然 等（2017）采用日历时间组合的方法，证实中国 A 股市场的分析师修正信息具有投资价值，并且这个投资价值主要来源于其能够预测公司基本面信息。汪荣飞 等（2018）基于季度财务指标构建了六组基本面指标，发现均能够有效预测未来盈余，进而发现分析师和投资者均没有意识到基本面指标的价值。常丹婷 等（2020）构建了价值因子、基本面两个维度进行横截面分析，发现只有当估值和基本面预期背离、存在错误定价时，高低估值的对冲组合才能产生显著的超额收益，高达 16.8%。

目前国内有关基本面分析的研究较少，但在成熟的国外资本市场，基本面分析的作用已被大量文献证实。Ou et al.（1988）构建了 68 组基本面指标，发现其

能够成功预测未来股票收益涨跌的概率，多空组合在两年内达到 12.5% 的超额收益。Abarbanell et al.（1997）构建了 12 组基本面指标，研究发现大部分指标能够预测股票未来价值，Abarbanell et al.（1998）进一步利用这些基本面指标构建投资组合，获得了 13.2% 的年化超额收益，并发现超额收益大部分与未来的业绩公告相关。Piotroski（2000）对高账面市值比的公司进行了研究，并用 9 组基本面指标构建出综合指标 FSCORE，用 FSCORE 挑选出真正的价值股并提升了 7.5% 的收益率。同样，Mohanram（2005）着眼于低账面市值比的成长股，利用 8 组基本面指标构建了综合指标 GSCORE。Asness et al.（2019）根据盈利性、成长性、安全性，构建了衡量公司质量的 QMJ（quality minus junk）指标，发现高质量公司股价较高，但低质量公司不定，多头高质量公司空头低质量公司的投资组合信息比率大于 1。Bartram et al.（2018）基于统计方法对在 CRSP 有记录的美国公司进行连续 310 月的检验，发现错误定价信号能够预测未来收益，带来超额利润，基本面分析有效。

上述研究均根据基本面指标构建了综合指标，并且构建对应的多空组合获取了超额收益，证明了基本面分析的有效性。但国内有关基本面分析的研究较少，运用大数据统计方法、机器学习算法的更少。李斌 等（2017）设计了一套基于机器学习和技术指标的量化投资算法 ML-TEA（Machine Learning and Technical Analysis），分别采用支持向量机、神经网络、Adaboost 等机器学习算法，利用 19 项技术指标预测股价涨跌方向，发现这些算法具有更高的预测准确率，而根据预测所构建的投资组合也取得了更好的投资绩效。进而李斌 等（2019）基于 A 股市场的 96 项异象因子，运用 12 种机器学习算法去构建股票收益预测模型及投资组合，并且其投资策略能够获得比传统线性算法和所有单因子更好的投资绩效。Fischer et al.（2018）采用长短期记忆模型（Long Short-Term Memory，LSTM），利用日频收益率数据预测标普 500 中的每只成分股相对于其截面中值收益率的涨跌方向，而相应构建的投资组合绩效显著优于其他线性或机器学习模型。

综上所述，本文将通过 BRTs 模型，基于 A 股公司财务指标去进行其内在价值的预测，并以内在价值与市场价值的差异去构建错误定价因子 M，以此进行我国 A 股市场基本面分析可行性的研究。

3 提升回归树（BRTs）

本文采用提升回归树（BRTs）算法进行数据模拟[32]。

3.1 确定函数形式 $f(.)$

对数据进行回归的方法有很多，可以按照是否提前预设回归形式分为参数回归方法和非参数回归方法。

- 参数回归会对函数形式 $f(.)$ 进行假设，一般为线性模型。正是因为提前预设好了函数形式，所以只需对回归系数 β 进行拟合，比较容易实现；但缺点便是这种提前设定好的函数形式可能并非完全符合现实情况，如果数据背后真实的函数形式为非线性，则按照线性函数去回归，结果可信度极低，也不能有效进行数据预测。
- 而非参数回归不会提前设定好函数形式 $f(.)$，相反，这种方法将寻找一种最贴合数据的函数形式，正是这种灵活性让函数形式更为多样化。由于其自由度较高，所以更为拟合模型的真实函数关系，但这也带来了缺陷，便是需要大量的数据才可以对 $f(.)$ 进行最为准确的预测，若是数据量太小、或是有离群点的影响，函数会对数据过拟合，数据预测效果大幅减弱。

在具体实证过程中，我们可以选用的方法有很多。传统的线性回归方法有最小二乘法线性回归（OLS）、Lasso 回归（Lasso）、岭回归（Ridge）等，传统机器学习算法包括支持向量机（Support Vector Machines，SVM）、梯度提升树（Gradient Boosting Decision Tree，GBDT）、随机森林（Random Forest，RF）等等，以及其他的深度学习算法。

可是，这些方法在灵活性和可解释性两个维度上不能兼得，一般而言参数模型比较受限，自由度不高，但是理解起来比较容易，因为有着具体的函数形式，例如 Lasso 回归、最小二乘法线性回归等方法；而非参数模型更为灵活，不会提供具体的函数形式，难以做到可视化，多为机器学习算法。

3.2 回归树

目前，最流行的两类机器学习算法莫过于神经网络算法（卷积神经网络、循环神经网络、生成式对抗网络和图神经网络）与树形算法（随机森林、GBDT、XG-

Boost 和 LightGBM)[33]。树形算法的基础组成部分便是决策树，由于其易理解、易构建、速度快等特点，被广泛的应用在数据挖掘、机器学习等领域。

根据处理数据类型的不同，决策树又分为两类：分类决策树与回归决策树。分类决策树主要用于处理定性指标、离散型数据，回归决策树用于处理定量指标、连续型数据。

决策树的原理便是模仿树的生长过程，在向下生长的过程中产生各种内部结点（Internal Node），每个内部结点又分支结出叶结点（Leaf Node）。内部结点表示一个特征或属性，即根据该特征把观测值归类到不同分支；而叶结点表示一个类别或者某个值，主要指落入该组所有数据的预测值。对于回归树而言，便是将所有数据 $X_1, X_2, \ldots, X_p$ 分到 J 个空间上不重合的区域 $R_1, R_2, \ldots, R_J$ 里面，对位于每个区域 R_j 的数据，其预测值即为 R_j 内所有观测值的均值。在进行决策过程时，根据输入样本每个特征维度值的大小，从上往下在每一结点进行选择和分支，最终落入 J 个区域中的一个。

理论上来说，分割的区域有着任意的形状，但在实际情况中，为了方便操作和理解，多采用高维矩形的分割形状，最终目标是找到能够使残差平方和 RSS 最小化的分割区域 $R_1, R_2, \ldots, R_J$，RSS 计算方法如下所示，其中 $\hat{y}_{R_j}$ 代表第 j 个区域内所有观测点的平均值。

$$\sum_{j=1}^{J} \sum_{i \in R_j} \left(y_i - \hat{y}_{R_j} \right)^2$$

由于回归树向下分化的可能性太多，无法通过穷举的方法进行选择，所以其采用的方法是递归二元分裂（Recursive Binary Splitting），即从上往下逐步分割，每个结点分到两个方向，对于任意 j 和 s，进行如下分割：

$$R_1(j, s) = \{X \mid X_j < s\} \quad R_2(j, s) = \{X \mid X_j \geq s\}$$

在任意结点上，通过选择最优的系数 j 和 s，使得两组 RSS 最小化。其中 $\hat{y}_{R_1}$ 是组 $R_1(j, s)$ 所有观测点的平均值，$\hat{y}_{R_2}$ 是组 $R_2(j, s)$ 的平均值。

$$RSS = \sum_{i: x_i \in R_1(j,s)} \left(y_i - \hat{y}_{R_1} \right)^2 + \sum_{i: x_i \in R_2(j,s)} \left(y_i - \hat{y}_{R_2} \right)^2$$

按照以上方法确定好了数据空间的切割方式，即生成完整回归树后，对于任意输入数据样本 x，其输出值如下式所示。其中，当 $x \in R_j$ 时，$I(x \in R_j)$ 取 1；

$x \notin R_j$ 时，$I(x \in R_j)$ 取 0。

$$f(x) = \sum_{j=1}^{J} \hat{c}_j I(x \in R_j)$$

3.3 Boosting 提升方法

回归树的优点有很多：便于理解、类似于人的思维方式、数据可视化等等，但缺点便是相比于其他的机器学习算法，其预测准确度仍然较低。所以在处理实际问题时，只用单一的回归树是不够的，可以用套袋法（Bagging）、随机森林（Random Forest）以及提升方法（Boosting）进行模型的改进。本文将利用集成学习中的 Boosting 框架，得到的新模型便是提升回归树（Boosting Regression Tree）。

在概率近似正确（Probably Approximately Correct，PAC）学习的框架下，如果一个概念或一个类，存在一个多项式的算法能够学习它，并且正确率很高，那么就称其为强可学习（Strongly Learnable）的；相反，如果一个概念存在一个多项式的学习算法能够学习它，但是学习的正确率仅比随机猜测略好，那么就称这个概念是弱可学习（Weakly Learnable）的[34]。在 PAC 的学习框架下，一定能够通过组合弱学习器来得到一个强学习器。

Boosting 以一种高度自适应的方法顺序地学习这些弱学习器，其中每个基础模型都依赖于前面的模型，最后按照某种确定性的策略将它们组合起来，以提高分类的性能。即先从初始训练集训练出一个基学习器，再根据基学习器的表现对训练样本分布进行调整，然后基于调整后的样本集训练下一个基学习器，如此反复，直到学习得到的基学习器的达到设定的个数，最终根据这些基学习器的预测表现去施加不同的权重，结合成强学习器[35]。其背后思想便是综合多个方法的预测结果，让结果更有效。

在回归树模型中，实际采用加法模型（Additive Model）与前向分步算法（Forward Stagewise Algorithm）进行提升。

首先采用前向分步算法确定最优参数，确定初始提升树为 $f_0(x) = 0$，第 m 步的模型为：

$$f_m(x) = f_{m-1}(x) + T\left(x; \Theta_m\right)$$

其中，$f_{m-1}(x)$ 为当前模型，第 m 棵决策树参数 Θ_m 将通过最小化损失函数 L 来确定：

$$\hat{\Theta}_m = \operatorname{argmin}_{(\Theta_m)} \sum_{i=1}^{N} L\left(y_i, f_{m-1}\left(x_i\right) + T\left(x_i; \Theta_m\right)\right)$$

当损失函数 L 采用平方误差时，即 $L(y, f(x)) = (y - f(x))^2$，则其损失将变为下式，其中 $r = y - f_{m-1}(x)$ 是当前模型拟合数据时的残差。

$$L\left(y, f_{m-1}(x) + T\left(x; \Theta_m\right)\right) = \left[y - f_{m-1}(x) - T\left(x; \Theta_m\right)\right]^2 = \left[r - T\left(x; \Theta_m\right)\right]^2$$

最终的提升树可以表示为决策树的加法模型。其中, $T\left(x; \Theta_m\right)$ 表示决策树、Θ_m 为决策树的参数、M 为树的个数。

$$f_M(x) = \sum_{m=1}^{M} T\left(x; \Theta_m\right)$$

3.4 梯度提升决策树（GBDT）

在上节中介绍了提升树的方法，其利用加法模型和前向分步算法实现学习的优化过程。当损失函数为平方损失和指数损失函数时，每一步的优化很容易实现。但对于一般的损失函数而言，优化过程却没有那么简单。为了解决这一问题，Friedman（2001）提出了梯度提升算法（Gradient Boost），其利用最速下降法（Steepest Descent），每一次建立模型是在之前模型损失函数的梯度下降方向，使得残差在梯度方向上减少，即利用损失函数的负梯度作为提升树算法残差的近似值，去拟合回归树，这便是提升回归树 BRTs 的一种改进算法——梯度提升决策树（Gradient Boosting Decision Tree，GBDT）[35, 37]。

GBDT 是机器学习中一个长盛不衰的算法模型，具有训练效果好、不易过拟合等优点，其不仅在工业界应用广泛，也被多用于分类、点击率预测、搜索排序等任务，在各类数据挖掘竞赛中也有着不俗表现，据统计，Kaggle 上的比赛有一半以上的冠军方案都是基于 GBDT 去实现的[38]。

由于 GBDT 是在 BRTs 基础上进行改进的，仅在损失函数的优化上有所变化，所以在算法上大同小异。首先确定初始决策树：

$$f_0(x) = \arg\min_{\gamma} \sum_{i=1}^{N} L\left(y_i, \gamma\right)$$

对于从 $m = 1, 2, \ldots, M$ 的所有决策树，以及每个样本 $i = 1, 2, \ldots N$，其损失函数的负梯度为：

$$r_{im} = -\left[\frac{\partial L\left(y_i, f\left(x_i\right)\right)}{\partial f\left(x_i\right)}\right]_{f=f_{m-1}}$$

进而根据算出来的残差 r_{im} 去拟合一个对应叶子结点为 R_{jm} 的决策树，其中

$j = 1, 2, \ldots, J_m$。对于叶结点 $j = 1, 2, \ldots, J_m$，计算出最佳的拟合值：

$$\gamma_{jm} = \arg\min_{\gamma} \sum_{x_i \in R_{jm}} L\left(y_i, f_{m-1}\left(x_i\right) + \gamma\right)$$

以上步骤相当于回归树遍历所有切分变量 j 和切分点 s 以找到最优的 j 和 s，然后在每个结点区域求最优的 γ。

更新第 m 个决策树为：

$$f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I\left(x \in R_{jm}\right)$$

最后得到回归树：

$$\hat{f}(x) = f_M(x) = \sum_{m=1}^{M} f_m(x) = \sum_{m=1}^{M} \sum_{j=1}^{J} c_{mj} I\left(x \in R_{mj}\right)$$

3.5 LightGBM 算法

目前实现 GBDT 模型的算法主要有四种：scikit-learn、XGBoost、LightGBM 和 CatBoost。但在算法原理方面，四种方法各有千秋，后三种方法都在最基础的 GradientBoostingRegressor 函数上进行优化。

- XGBoost 对所有特征进行预排序，并在遍历分割点的时候用 $O(\#data)$ 的代价找到特征上的最好分割点，最后将数据自分割点分裂成左右子结点。相比于 GradientBoostingRegressor 函数采用牛顿法、泰勒展开进行优化[39]。
- LightGBM 是基于 Histogram 的决策树算法，采用 leaf-wise 策略分裂叶子结点，训练速度快、内存占用低、能够有效处理高维大数据。
- CatBoost 能够有效处理类别型特征，以对称树作为基学习器，减少了预测时间，也能够避免数据的过拟合。

由于本文数据维度高、体量大、且全为数值型数据，在保证结果准确的同时也要保持运行速度，所以将采用 LightGBM 算法进行文章后续的实证检验，为避免方法选择的不同造成结果差异，我们同样采用上述其他算法进行验证，发现不同方法结果类似，说明本文的实证结果与算法的选择无关。

对于 LightGBM 算法参数的选择，均采用网格调参（Grid Search）的方式。本文选取了对 BRTs 算法影响程度最大的参数：树的数量 *n_estimators* 和学习率 *learning_rate* 去进行调整，由于 LightGBM 算法采用 leaf-wise 策略进行结点的分裂，为了预防数据的过拟合问题，所以也调整了树的深度 *max_depth*。

由于公司每期的内在价值只能由本期的财务指标进行预测，而非由过去的财务数据，故本文需要逐期进行模型参数的选择，于每期进行一次网格调参，得到预测准确率最高的模型最优参数后，进而预测该期的公司内在价值。

3.6 机器学习方法的局限性

虽然机器学习能够对数据进行很好的非线性拟合，但从另一个方面来说，线性关系是能够被可视化的，也是容易被理解的，而机器学习却是一个“黑匣子”（Black Box），不会展现出具体的回归函数形式，只能做到输入值与输出值的一一对应，很难从逻辑上给出变量之间的联系[11]。特别是在深度学习中，这个问题变得更加明显，例如在深度神经网络算法中，数据的抽象特征由神经网络经过多次计算而得，没有办法将原始数据和最终结果结合到一起。而这一缺陷也会影响很多金融领域的继续深入，例如在资产定价问题上，我们能通过机器学习方法得到某些因子对于投资组合收益的显著预测作用，但很难再去检验为什么是这些因子，或者因子的具体影响途径等等。

另外，机器学习方法十分依赖于训练集数据[40]。虽然机器学习的优势在于能够对数据能够深度学习，并给出有效结果，但结果都是基于输入的数据，数据的质量好坏决定了机器学习预测的上限。而且在金融研究中常用市场的历史数据，而这样的数据一般很难有非常完美、对称的数据结构，或是缺失值太多等等，或多或少都有些局限性。

3.7 机器学习方法的适用性

在本文的研究中，当我们尝试根据财务指标构建错误定价因子时，会遇到高维度的横截面数据预测问题，而且对于回归变量之间的关系是线性还是非线性的，是否存在显着的相互影响作用，这些都不得而知，但我们可以通过使用 BRTs 克服这些限制。

首先，BRTs 在包括金融在内的各个领域内均表现出强大的预测性能。其次，BRTs 可以处理大型的高维数据集，它们可以自动进行变量选择和缩尾工作，对异常值具有鲁棒性。第三，BRTs 不像其他机器学习方法那样是“黑匣子”，相反，该方法因为其可解释性高而闻名。

在实证过程中，可以看到采用 BRTs 方法在预测股票价格方面的出色表现，在后文的结果中，我们构建了相应的投资组合并取得了显著的正收益。此外，我们

同样采用线性方法进行数据模拟，结果显示根据 BRTs 算法提供的投资组合收益高于标准线性回归模型。

另外，在现实问题中鲜有能够完全呈现线性的变量关系，而 BRTs 方法通过非参数方法估计股票特征与收益之间的关系，正是它们之间复杂的关系促使诸如 BRTs 之类的机器学习方法成为了优于传统统计方法的选择。

4 研究设计

相比于传统的线性回归算法，机器学习方法更适合于本文的实证研究。在对于公司内在价值的预测过程中，线性回归模型将变量间的关系进行框定，会造成很大的偏误（Bias），一般而言，现实问题中的变量关系并不完全符合线性，或多或少都会呈现非线性趋势。再加上本文数据维度较高，数据量较大，并且伴有缺失值，机器学习方法均能做到完美解决。

本文的统计方法基于经济学中的一价定律（Law of One Price），类似于任一资产定价模型，我们用合成股票（Synthetic Stock）或复制投资（Replicating Portfolio）的市场价值去代表股票的内在价值，因为每个投资组合的基本特征都与被评估公司的特征相同。例如最简单的 FCFF 法便是进行证券市场上无风险（Risk-free）债券与产生无风险现金流资产之间的比较；在 CAPM 中，股票的复制投资组合是市场投资组合（Market Portfolio）与相同 β 值无风险资产的组合等[41]。

同样，在现有的基本面分析研究中，因子的选择较少，对于超额收益的攫取策略较为局限。近 35 年来已有多项研究证明了例如动量因子（Momentum）、市值因子（Size）等已有异象因子的有效性，针对这些因子的原因剖析也已经有了结果，例如投资者的过度自信（Overconfidence）与处置效应（Disposition Effect）从行为角度解释了动量因子[42, 43]。为了突破传统因子研究的局限性，本文创新采用财务指标去构建错误定价因子 M，并以此进行资产定价与配置。该因子的构建不参杂任何的主观因素，均通过客观标准去进行优化选择。

4.1 模型总体设计

图 4.1 展现了本文的总体设计流程，首先选取股票数据组成证券池，然后挑选公司相应的财务指标，运用 BRTs 模型进行内在价值的预测，以预测出的内在价值与市场价值的差额去构建错误定价因子 M，后以该错误定价因子分组进行投资组合的配置，并以真实股票交易数据进行仿真。

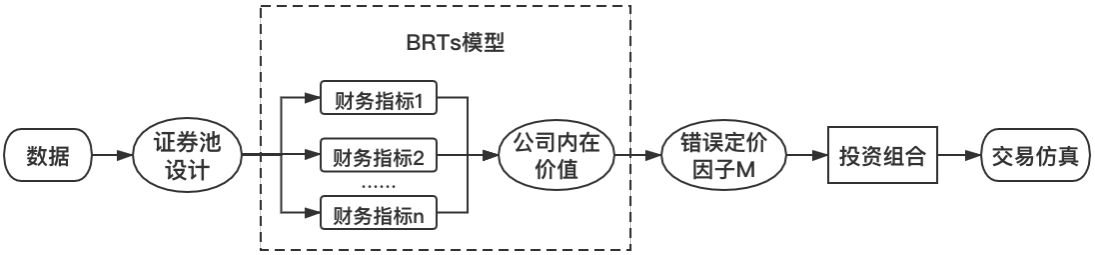

图 4.1　模型总体设计

4.2　数据样本

本文使用的 A 股公司财务报表数据和股票交易数据来自国泰安（CSMAR）数据库，因子模型数据分别来自中央财经大学金融学院[①]、BetaPlus 小组[②]以及 WHUFT 异象因子数据集[③]。由于我国 A 股上市公司的季报数据自 2002 年起才开始公布，为了保障数据的完整性，本文财务指标数据以 2002 年第一季度为数据的起点，由于 2020 年 12 月财务数据较少，原因或在于公司财报未公布、数据库尚未更新完毕等，故最终选择以 2019 年 12 月为止。综上，本文数据样本时间区间为 2002 年 3 月至 2019 年 12 月，频次为季度。

对于股票样本的选择，本文以我国 A 股市场 2018 年底前上市的公司作为数据样本，规避未来收益数据的缺失。由于 ST 股票存在退市风险、金融行业公司财务报表结构有别于上市公司等原因，本文剔除掉 ST 股票、金融类股票，最后剩余 3300 支。

有关数据缺失部分，若某支股票在第 t 期市场价值 $mktcap$ 数据存在缺失，则剔除该股票在 t 期的所有数据；但若有其他公司特征数据缺失，则由后文中所运用的机器学习算法 BRTs 进行自动填充。

在完成以上筛选步骤后，2002 年 3 月至 2019 年 12 月的有效样本共 132468 条。图 4.2 展示了 2002 年 3 月至 2019 年 12 月季频月度有效样本量。总体来看，每个月的样本量随年份呈现上升趋势，从 2002 年 3 月的 938 条有效样本增至 2019 年 12 月的 3225 条。

①详见官网：http://sf.cufe.edu.cn/info/1198/9562.htm
②详见官网：https://www.factorwar.com/data/factor-models/
③详见 Github 主页：https://github.com/WHUFT

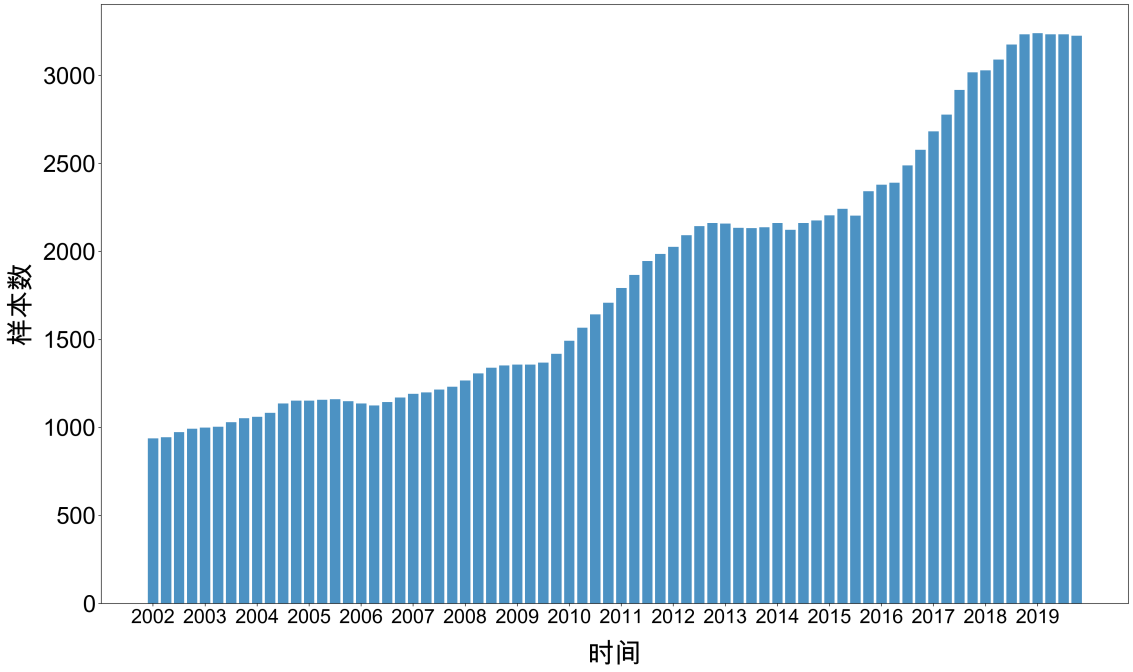

图 4.2　样本数据的月度有效样本量

4.3　变量定义与说明

为了计算公司的错误定价因子 M，首先需要对公司的内在价值进行估计。本文以公司市场价值为因变量，财务指标为自变量进行回归计算，得到的残差即为错误定价；之后通过 Fama-MacBeth 截面回归、因子模型回归等方法验证错误定价因子 M 的有效性。文章中采用的主要变量如表 4.1 所示。

4.4　模型设定

现有研究中，常见的公司内在价值计算方法包括现金流贴现法、相对价值法、经济附加值法、实物期权法等。这些方法虽然能起到一定的预测作用，但由于其选择的指标过少、过于死板，或太过于依赖使用者对未来的主观预测结果，亦或计算方法过于程式化等原因，很难排除掉数据窥视带来的影响。

根据本文的基本面分析原理，公司内在价值反映在一系列财务指标中，如式4.1所示，其中 i 代表不同公司，t 代表不同期数。$mktcap_{i,t}$ 为公司 i 于季度 t 时的市场价值，$f(.)$ 定义了一个参数为 θ 的非线性函数，在本文中为采用BRTs方法背后具体的函数形式，参数 θ 主要包括树的数量 $n\_estimators$、学习率 $learning\_rate$ 以及树的深度 max_depth，均采用网格调参的方法进行最优化，$I_{i,t} = (I_{i,t,1}, I_{i,t,2}, ..., I_{i,t,N})$

表 4.1　变量定义与说明

变量名	定义及说明
$mktcap_{it}$	公司 i 于第 t 期的市场价值
$M_{i,t}$	公司 i 于第 t 期的错误定价因子
$R_{i,t}$	公司 i 于第 t 期月个股收益率
BM	市净率
$beta$	市场投资组合 β 值
$grltnoa$	净经营性资产增长率
PA	毛利率
EP	市盈率
$lagretn$	上一个月的收益率
$mom12$	t-12 月至 t-2 月的累计收益率（共计 11 个月）
$mom36$	t-36 月至 t-13 月累计收益率（共计 24 个月）

代表公司 i 在第 t 期时 N 个财务指标向量，$\epsilon_{i,t}$ 为残差项。

$$mktcap_{i,t} = f(I_{i,t};\theta) + \epsilon_{i,t} \tag{4.1}$$

对于财务指标的选择，为避免数据窥视的影响，并非由笔者主观挑选。相反，本文于 CSMAR 数据库下载了所有财务报表数据，包括资产负债表、利润表与直接法和间接法计算的现金流量表，共计 279 个指标。但由于不同公司个体财务情况差异大、过于细分的指标数据未被使用、或是数据库数据录入缺失等原因，数据库中每个指标或多或少都有缺失值的存在，故为了计算的严谨性，将对 279 个财务指标进行统一筛选。

基于统计，在所有财务指标中，缺失值比例最大的达到 93.6%，最小为 0.01%，均值高达 41.6%，说明有很多指标数据缺失情况严重，具体情况详见附录 A。为了做到对公司内在价值的准确评估，本文选择缺失值比例较小的指标，以避免缺失值造成的数据偏误[44]；同时，为了能完整对公司内在价值进行评估，所选择的财务指标必须多样化。综上，本文以 5% 的缺失值比例为界限，挑选出缺失值比例小于等于 5% 的所有指标，共有 51 个，具体可见附录 B 。

在成功确定函数形式 $f(.)$ 后，即完成对公司内在价值的预测后，将进行式 4.1 的回归，以确定该期的错误定价情况。回归后残差即为公司市值与内在价值的偏差，将该残差基于市值进行标准化后的负值，定义为错误定价因子 M，如式 4.2 所示。

$$M_{i,t} = -1 \times \frac{\epsilon_{i,t}}{mktcap_{i,t}} \tag{4.2}$$

当公司内在价值低于市场价值时，即残差大于 0，对应 M 为负值，说明股票被高估；当公司内在价值高于市场价值时，M 为正值，说明股票被低估。

其中，考虑到市场以及公司的变化情况，公司的内在价值以及错误定价因子 M 每期都会重新计算一次，并对公司进行重新分组调整，避免数据偏误。

总而言之，本文的统计方法不会偏好特定的股票市场、不考虑整个市场在特定期数是否被高估或低估、也不依赖于内在价值的理论模型。相反，本文将公司进行相互比较，使用拟合优度的统计标准来识别公司内在价值如何体现于会计属性。

后续研究中，每期都将根据错误定价因子 M 大小进行公司分组：Q1 到 Q5 五组。其中 Q1 为 M 最小的公司，即股票最被高估；Q5 为 M 最大的公司，即股票最被低估。如果基本面分析有效、财务指标蕴含公司的内在价值，当市场价值最终趋近于内在价值时，最被高估的股票价格会回落，最被低估的股票价格会爬升，以此构建低买高卖的投资组合来获利。

为验证上述猜想，后文将进行因子模型回归、Fama-MacBeth 横截面回归以及稳健性分析，如式 4.3、4.4 所示。

$$R_{i,t+1} = a_t + b_t M_{i,t} + \sum_{k=1}^{K} c_{k,t} X_{i,k,t} + \epsilon_{i,t+1} \tag{4.3}$$

$$R_{i,t} = \alpha_i + \sum_{k=1}^{K} \beta_{i,k} F_{k,t} + \epsilon_{i,t} \tag{4.4}$$

式 4.3 中 $X_{i,k,t}$ 代表公司 i 于季度 t 时第 k 个公司特征，作为控制变量，具体特征见表 4.1 所示；若 b_t 显著，说明错误定价因子 M 能够有效预测股票价格未来走势，证明了其有效性。

式 4.4 中 $F_{k,t}$ 为季度 t 时第 k 个因子值，本文主要采用 Fama-French 三因子模型、Carhart 四因子模型、Fama-French 五因子模型、以及 Hou-Xue-Zhang 四因子模型（后称为 q-因子模型）进行分析；其中 α_i 代表该投资组合的超额收益。如果 Q1 到 Q5 五组超额收益 α 有显著的递增趋势，以及 Q1、Q5 的多空组合超额收益 α 显著大于 0，也能证明错误定价因子 M 的有效性。

4.5 描述性统计

表 4.2 列出了使用 LightGBM 算法对数据进行模拟后，所有相关变量的描述性统计值，这里将样本根据错误定价因子 M 分成了 Q1 到 Q5 五组进行考察。其中第一列为所有数据的平均值，第二列为相应变量与错误定价因子 M 的相关系数，

第三列为 Q1 组（价值最被高估）的变量平均值，一直到最后一列为 Q5 组（价值最被低估）的变量平均值。

表 4.2　描述性统计

			错误定价因子 M 分组				
	所有数据 (1)	相关系数 (2)	Q1 (被高估) (3)	Q2 (4)	Q3 (5)	Q4 (6)	Q5 (被低估) (7)
M	0.4296	1.000	-0.3173	-0.0157	0.2372	0.5847	1.6615
mktcap	7.1297	-0.112	15.2640	8.5644	5.8529	3.8186	2.1325
R_t	0.8750	-0.032	2.6395	1.1614	0.5074	0.2173	-0.1539
R_{t+1}	0.7814	0.040	0.2792	0.4824	0.7857	0.9555	1.4064
R_{t+2}	0.7266	0.038	0.4140	0.4454	0.5725	1.0322	1.1706
lagretn	0.0295	0.012	0.0387	0.0268	0.0252	0.0239	0.0328
BM	0.4304	0.069	0.3321	0.4351	0.4755	0.4871	0.4230
EP	0.0190	-0.051	0.0190	0.0223	0.0207	0.0179	0.0148
beta	1.1420	-0.006	1.1063	1.1414	1.1485	1.1555	1.1609
grltnoa	1.4602	0.020	0.3031	0.3421	0.2945	1.9855	5.0868
PA	0.0133	-0.071	0.0179	0.0153	0.0114	0.0093	0.0123

在描述性统计表中，从第三列到最后一列，第 $t-1$ 期到第 t 期的收益 R_t 逐渐减少，说明购买被高估的股票收益更高，与现实情况相符，同时第 t 期前一个月的收益 *lagretn* 也基本呈现逐渐减少的趋势。但是第 t 期到第 $t+1$ 期的收益 R_{t+1}、$t+1$ 期到第 $t+2$ 期的收益 R_{t+2} 逐渐增加，说明被高估的股票价格下降，逐渐趋于正常，所以收益变低，而被低估的股票价值上涨到正常价值，收益变高，与本文猜想保持一致。

相比于被高估的 Q1 组股票，被低估的 Q5 组股票平均有着更高的市净率 BM、更低的市场价值 *mktcap*；再加上错误定价因子 M 与市净率 BM 有着正相关关系，说明被高估的股票更多是高市值的价值股，而被低估的股票多是低市值的成长股。

错误定价因子 M 与 *beta* 相关系数为负，说明系统性风险并不能解释 M 对于股票未来收益的预测能力。由于 M 与其他异象的相关系数基本都低于 0.05，在后续回归中，将 *beta* 值、过去一个月的收益 *lagretn*、市净率 BM、市盈率 EP、净经营性资产增长率 *grltnoa* 以及毛利率 PA 作为公司层面的控制变量。由于公司市值 *mktcap* 与错误定价因子 M 高度相关，在后续回归中为避免多重共线性问题，控制变量将不包含公司市值 *mktcap*。

5 实证分析

5.1 因子模型

本文依据错误定价因子 ***M*** 的预测结果，构建了 Q1 和 Q5 两组的多空组合，以月度平均收益 *mean* 和风险调节收益 α 作为绩效衡量指标。为了避免 A 股市场做空机制的限制，本文还构建了 Q1 到 Q5 五组的多头组合，如果从 Q1 组到 Q5 组、从最被高估到最被低估的股票能够呈现超额收益的递增趋势，同样能证明错误定价因子 ***M*** 的有效性。

对于因子的选择，使用了常见的 CAPM 模型、Fama−French 三因子模型、Carhart 四因子模型、Fama−French 五因子模型、以及 Hou-Xue-Zhang 四因子模型（后称为 q-因子模型），采用了其中的 6 组因子（CAPM、FF3、FFC、FF5、FF5+MOM、Q）进行研究，具体结果见表 5.1所示，记录了 Q1 到 Q5 五组多头组合和多空组合的月度平均收益 *mean*、风险调节收益 α 及 *t* 值。

表 5.1 因子模型回归

模型	系数	Q1	Q2	Q3	Q4	Q5	Q5-Q1
CAPM	*mean*	-0.180	0.037	0.255	0.421	0.955	1.135
	α	0.0452	0.235	0.540	0.708	1.153	1.123
	t 值	(0.52)	(2.71)	(6.24)	(7.80)	(8.26)	(2.35)
FF3	α	0.0793	0.271	0.573	0.747	1.196	1.096
	t 值	(0.93)	(3.18)	(6.72)	(8.39)	(8.65)	(2.43)
FFC	α	0.146	0.320	0.643	0.823	1.268	1.099
	t 值	(1.70)	(3.74)	(7.52)	(9.22)	(9.13)	(2.40)
FF5	α	-1.186	-1.172	-0.941	-0.762	-0.143	0.981
	t 值	(-9.02)	(-12.90)	(-9.76)	(-9.01)	(-2.82)	(1.76)
FF5+mom	α	-1.122	-1.116	-0.869	-0.687	-0.0727	0.984
	t 值	(-11.14)	(-11.12)	(-8.69)	(-6.56)	(-0.44)	(1.75)
Q	α	0.177	0.380	0.677	0.853	1.298	1.106
	t 值	(2.08)	(4.46)	(7.95)	(9.58)	(9.37)	(2.43)

可以看出在运用 BRTs 模型进行数据模拟后，Q1 到 Q5 的多头组合月度平均收益 *mean*、风险调节收益 α 均呈现递增趋势，并且非常显著；Q1、Q5 的多空组合

也有显著的正收益，证明了错误定价因子 M 的作用，以及基本面分析的有效性。

附录 C 展现了采用线性回归方法的因子模型结果，可以看出结果也非常显著，佐证了本文的结论。但相比而言，采用 BRTs 算法得到的 Q1、Q5 的多空组合收益率更高；并且 BRTs 算法能够有效处理缺失值的问题，而线性回归对于缺失值非常敏感，数据量大大缩水，虽然两种方法结果相似，但对于结果或多或少仍有影响，如果扩展到其他研究中，这种数据的直接丢弃或许是致命的。

5.2 Fama-MacBeth 截面回归

表 5.2 展现了对于所有公司 Fama-MacBeth 的时间序列截面回归结果，两列分别选用了不同的控制变量来表现错误定价因子 M 的预测能力，所有列均控制了行业固定效应，括号内的数字代表 t 值。

其中，第一列仅采用了 M 作为解释变量，结果非常显著；第二列新加入了公司特征作为控制变量：$beta$ 值、市净率 BM、三个动量因子 $lagretn$、$mom12$、$mom36$、市盈率 EP、净经营性资产增长率 $grltnoa$ 与毛利率 PA，让回归更为完整，错误定价因子 M 同样显著，说明错误定价因子 M 可以用来预测股票未来收益。

5.3 股票特征重要性

如图 5.1 所示，在对公司内在价值的回归计算过程中，对 51 个财务指标进行重要性排序。其中，y 轴显示了所有财务指标，x 轴代表了不同财务指标的重要性比例，可见最为重要的指标有股本、净利润，分别占到了约 10%、6%，对应说明评估公司内在价值的重要因素一般在于股东行为以及公司业务的利润大小等。

5.4 稳健性分析

上述小节 5.2、5.1 的结果证明了错误定价因子 M 以及基本面分析的有效性，公司内在价值能够完全反映于财务指标。为检验超额利润是否真正来自于错误定价而非因子遗漏，即市场价值是否会逐渐趋向于内在价值，本文进行以下稳健性分析。

在第 t 期计算各公司的错误定价因子 M，并分为 Q1 到 Q5 五组，构建 Q1 与 Q5 两组的多空组合，基于该投资组合计算未来 36 期的收益情况，同上文采用六组因子（CAPM、FF3、FFC、FF5、FF5+MOM、Q）进行研究，结果如图 5.2、5.3、5.4、5.5、5.6、5.7 所示。

表 5.2　Fama-MacBeth 截面回归

	(1) R_{t+1}	(2) R_{t+1}
M	1.623***	6.601**
	(10.82)	(1.98)
$beta$		20.08***
		(8.23)
BM		27.61
		(1.48)
EP		-252.2
		(-0.52)
$lagretn$		-13.03***
		(-3.39)
$mom12$		8.121***
		(5.43)
$mom36$		3.545**
		(2.42)
$grltnoa$		-0.757
		(-0.17)
PA		-176.7*
		(-1.80)
行业固定效应	控制	控制
N	129169	50153
R^2	0.0557	0.572

t statistics in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

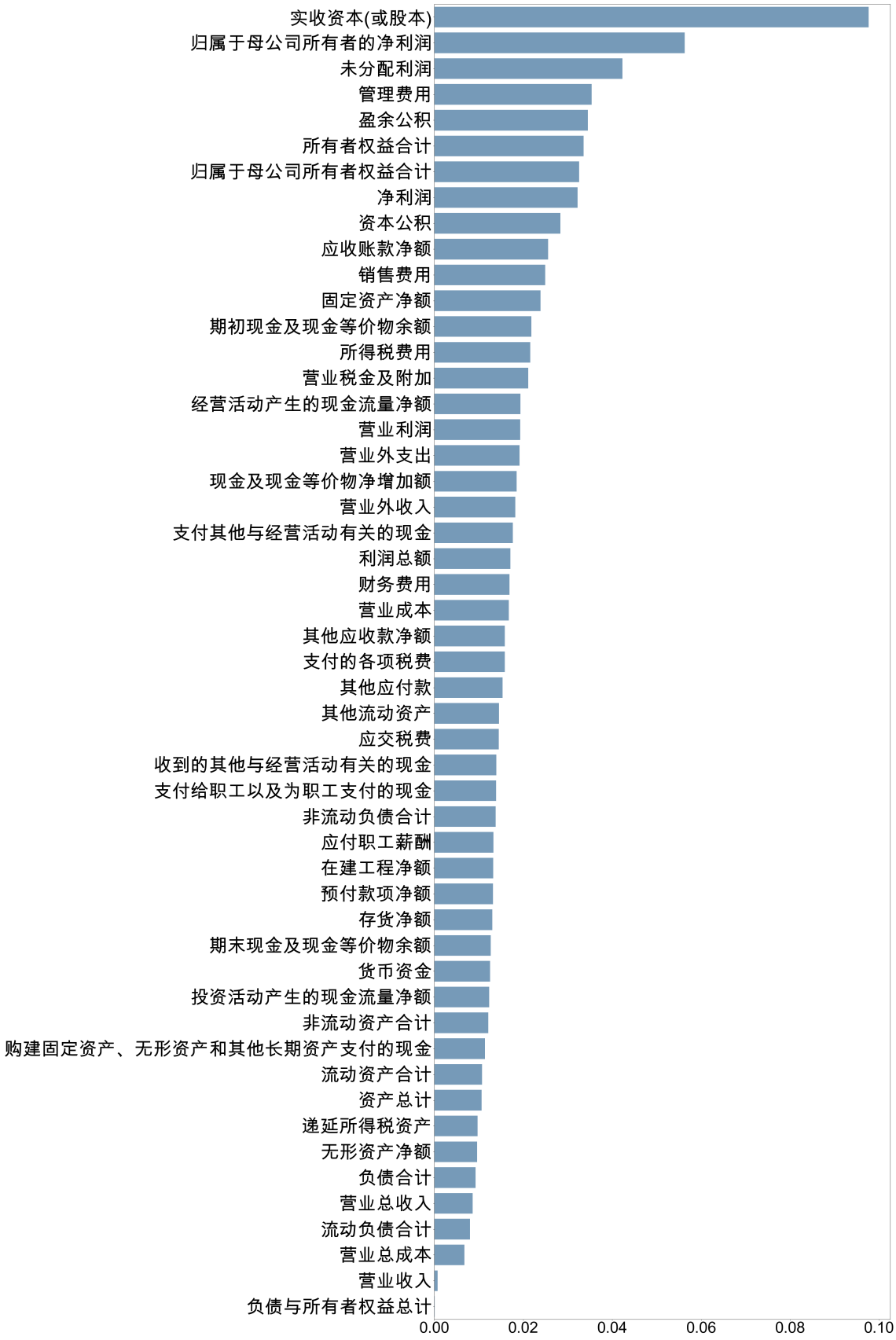
实收资本(或股本)
归属于母公司所有者的净利润
未分配利润
管理费用
盈余公积
所有者权益合计
归属于母公司所有者权益合计
净利润
资本公积
应收账款净额
销售费用
固定资产净额
期初现金及现金等价物余额
所得税费用
营业税金及附加
经营活动产生的现金流量净额
营业利润
营业外支出
现金及现金等价物净增加额
营业外收入
支付其他与经营活动有关的现金
利润总额
财务费用
营业成本
其他应收款净额
支付的各项税费
其他应付款
其他流动资产
应交税费
收到的其他与经营活动有关的现金
支付给职工以及为职工支付的现金
非流动负债合计
应付职工薪酬
在建工程净额
预付款项净额
存货净额
期末现金及现金等价物余额
货币资金
投资活动产生的现金流量净额
非流动资产合计
购建固定资产、无形资产和其他长期资产支付的现金
流动资产合计
资产总计
递延所得税资产
无形资产净额
负债合计
营业总收入
流动负债合计
营业总成本
营业收入
负债与所有者权益总计
0.00
0.02
0.04
0.06
0.08
0.10

图 5.1　股票特征重要性

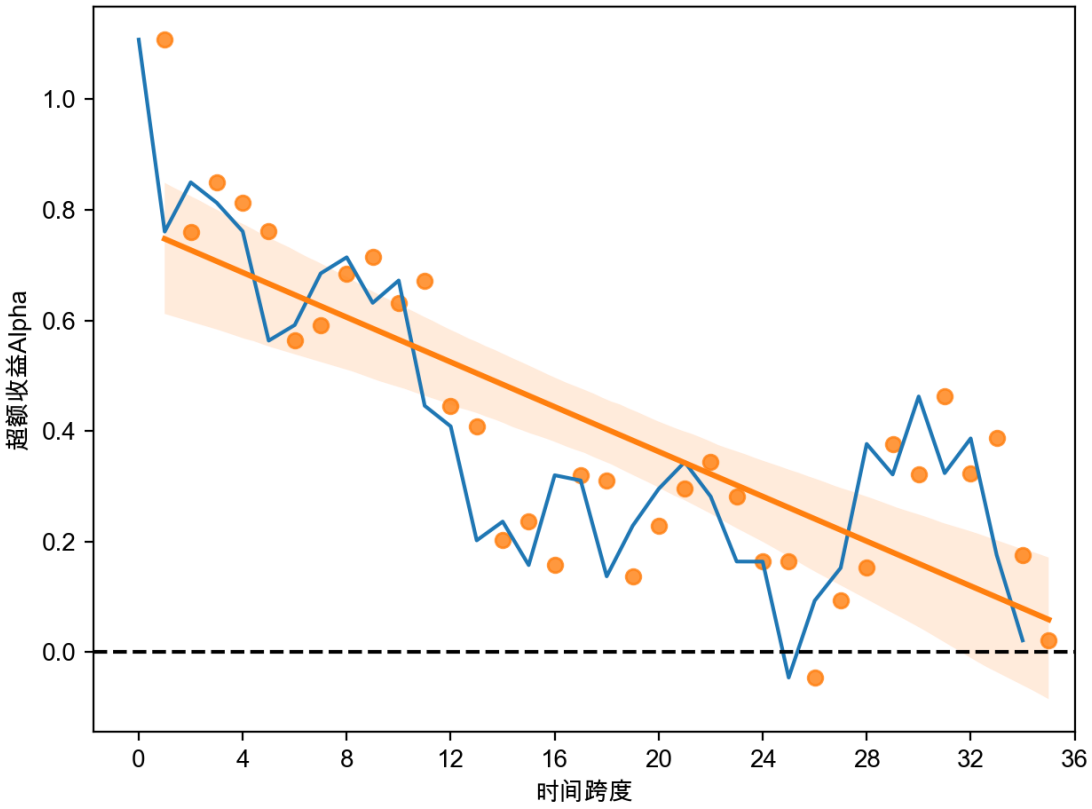

图 5.2　信号延迟（CAPM 模型）

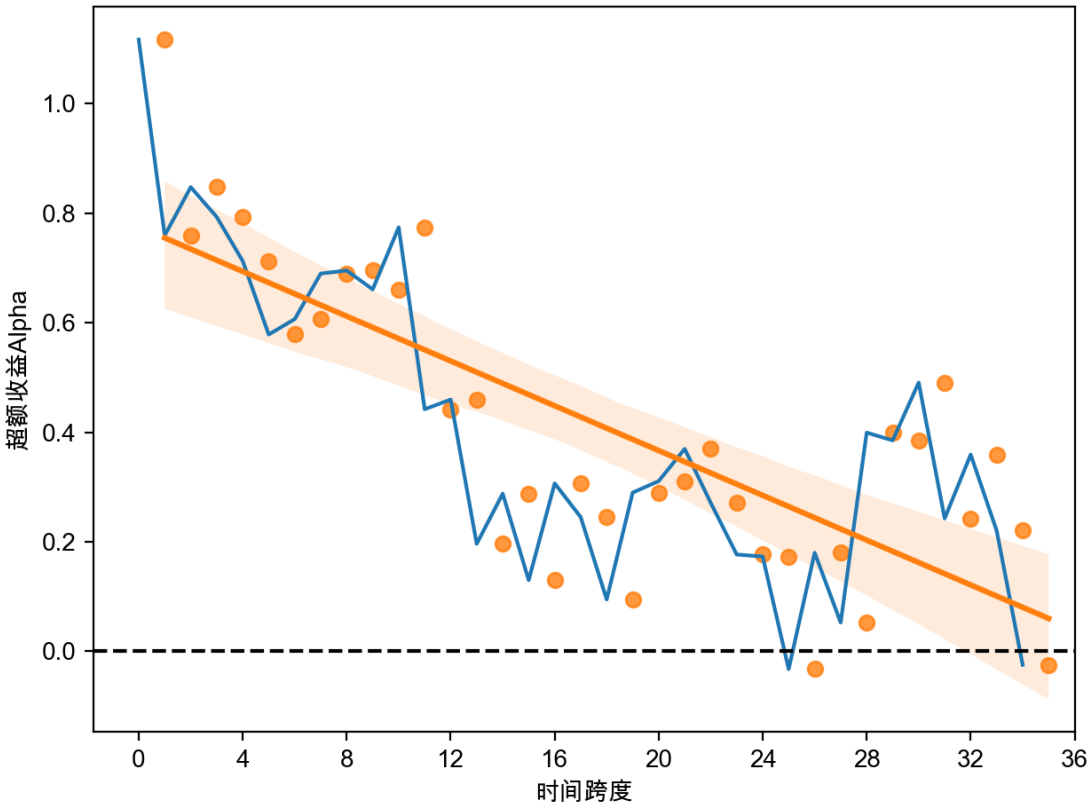

图 5.3　信号延迟（FF3 模型）

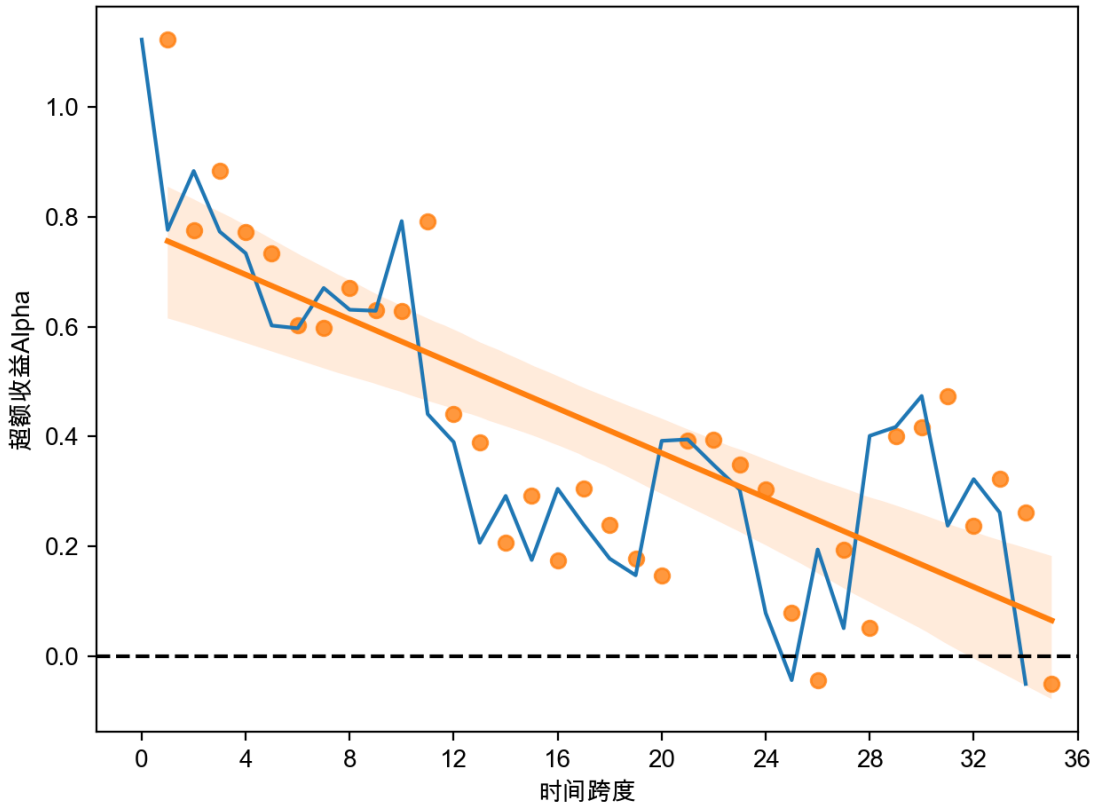

图 5.4　信号延迟（FFC 模型）

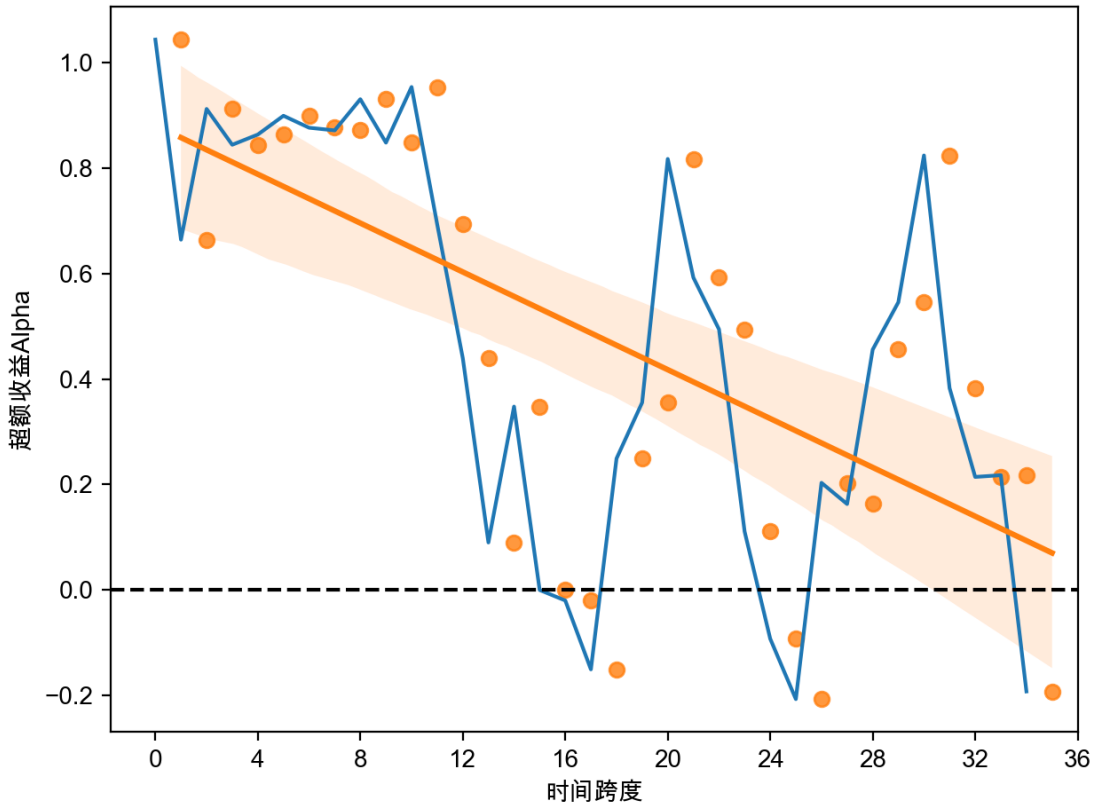

图 5.5　信号延迟（FF5 模型）

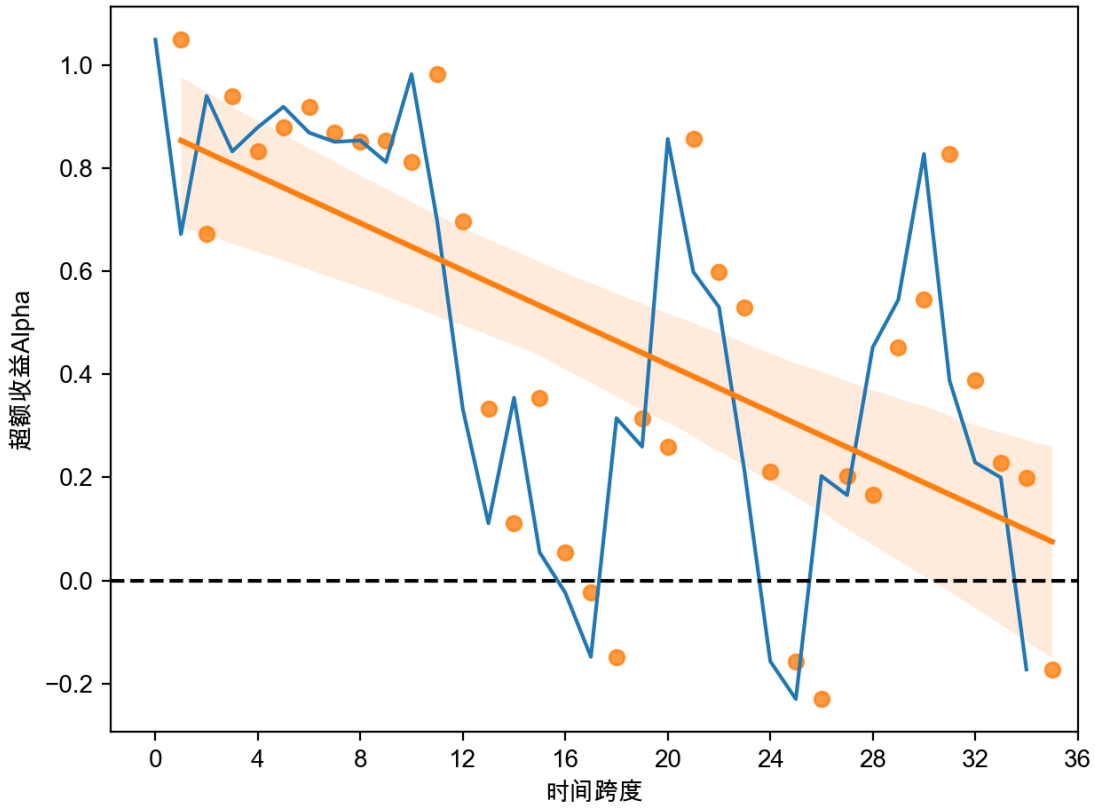

图 5.6　信号延迟（FF5+MOM 模型）

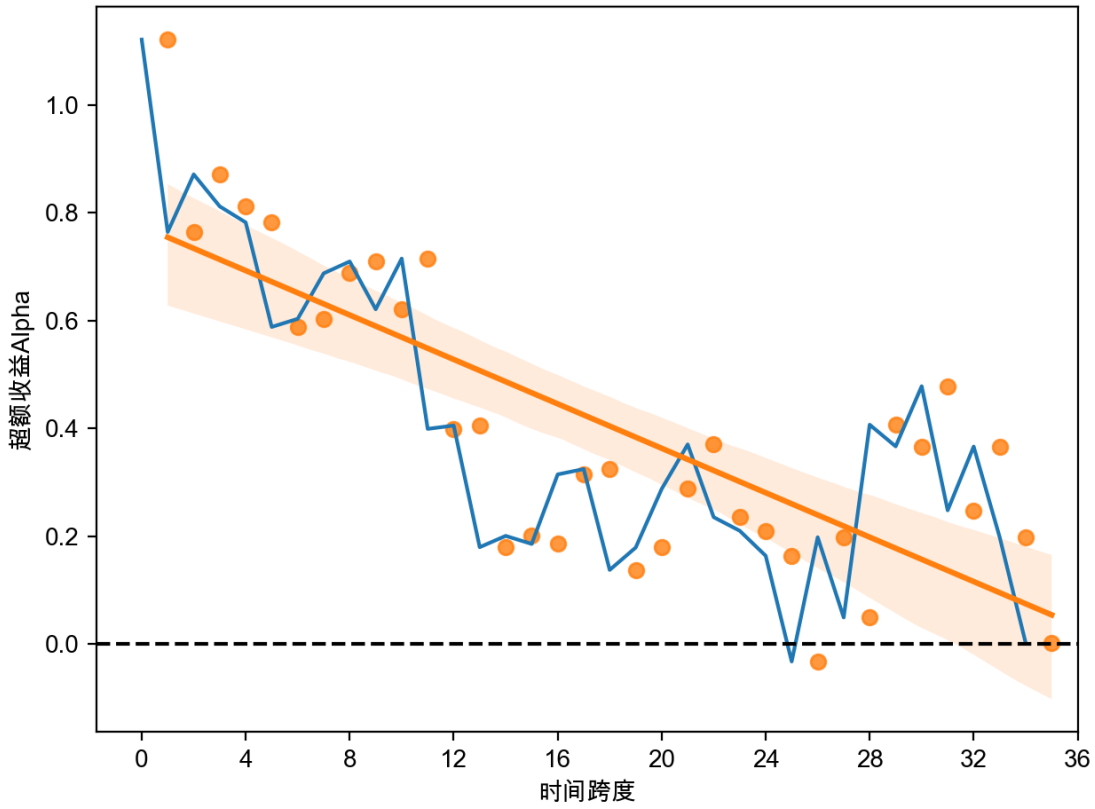

图 5.7　信号延迟（Q 因子模型）

在六组模型中，多空组合收益均在当前第 t 期拥有最高的超额收益 α，但后面随着时间增长，收益呈现明显的下降趋势，说明当前时点的错误定价已经随着时间而抵消，股票的市场价格能够灵活随着信息进行调整，进而逐渐趋近于内在价值，收益也就越来越小，证明了本文在每一期重新进行内在价值计算、根据错误定价因子 M 对公司进行分组的正确性。若采用过往时点的结果进行未来股票收益的分析，会产生一定的偏误，所以对于运用错误定价因子 M 进行投资组合的构建，要做到及时更新。

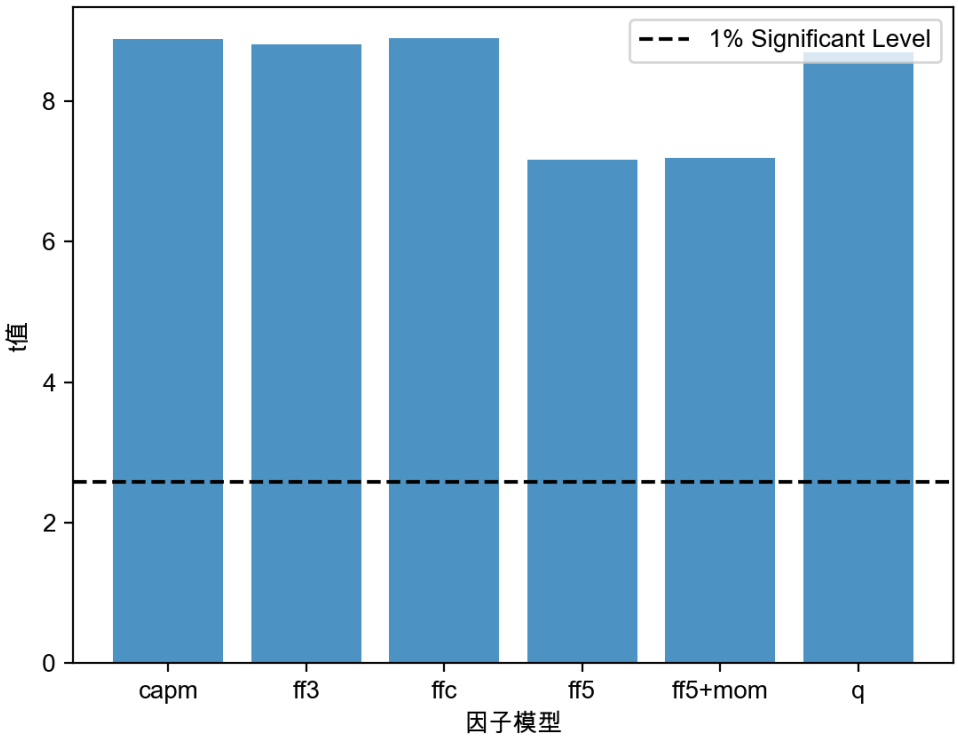

图 5.8　五组模型多空组合显著性检验

图 5.8 是对于六组因子模型中多空组合收益是否显著大于 0 的检验，所有模型结果均在 1% 的显著性水平下成立，证明了运用财务指标衡量公司内在价值的可行性、根据错误定价因子 M 进行公司分组的有效性。

6 结论与启示

本文基于 2002-2019 年的季度财务报表数据，选取缺失值比例低于 5% 的 51 个财务指标，采用 BRTs 模型、LightGBM 算法完成了对公司内在价值的预测，并以估计的内在价值与市场价值的差异构建了错误定价因子 M，当 M 为负值时，说明公司价值最被高估，当 M 为正值时，说明公司价值最被低估。根据错误定价因子 M 对公司进行分类，并进行相应的高卖低买策略，研究证明从最被高估到最被低估的投资组合，超额收益 α 呈现显著的递增关系；并且其多空组合产生了显著的正收益，但收益保持下降关系，这也强调了信息迭代的重要性，需经常性对数据进行重新调整与分析。以上结论在 Fama-MacBeth 截面回归和季度投资组合的五种因子模型检验中均显著符合预期，能够通过未来 36 期的稳健性测试，具有可观的经济意义。综上所述，本文的实证结果表明，基于季度财务报表的统计分析能够有效运用于中国 A 股市场，基本面分析起到了一定的预测作用。

总体来说，BRTs 模型在本文的股票价格预测、错误定价情况与资产组合配置研究中表现良好，进一步丰富和拓展了该领域下的实证研究结果。与传统的线性研究相比，基于机器学习算法的统计方法有着以下几个特点：

(1) 本文所采用的数据包含了从 02 年初到 19 年底的绝大多数 A 股市场股票，并且选取了 51 个财务指标进行预测，整体数据体量更大，维度更高。与经典的线性回归方法相比，机器学习算法数据分析能力更为优越，并且对于高维数据的处理更为谨慎。而且本文采用的 BRTs 模型比线性模型更易让人理解背后的原理，也符合常人的思维方式；同样，树模型对于缺失值并不敏感，而线性模型会去掉有缺失值的整行数据，对于数据整体结构的影响太大。

(2) 运用机器学习方法对中国 A 股市场资产定价与配置的研究现在仍很少，但该领域对于未来金融研究领域的发展有着重要的铺垫作用，因为其将金融与计算机结合到了一起，成功将计算机科学的逻辑思维运用到了金融领域的数据分析问题上，为后续研究提供了大量的相关实证研究结果[11]。一方面，随着机器学习算法的不断优化，提高了对于资产价格预测的准确性，同样让资产配置有了更多可能的组合。但另一方面，机器学习算法的“黑匣子”特征也会影响具体问题的结果，因为输入变量与输出结果可能很难具有逻辑上的联系，在研究过程中我们更关心实证的结果而非理论，机器学习方法可能会限

制问题的深入理解。所以，要做到将机器学习方法有效融入到金融领域的研究中，仍是未来需要解决的一个关键问题，需要做到两者的平衡。

基本面分析的有效性也反映出我国资本市场的效率低下，虽然我国 A 股市场总市值近 80 万亿元，已发展为全球第二大股票市场。但仍然由于成立时间晚、相关制度不完善、散户比例大、交易成本高等问题，导致了定价效率的低下[10]。而市场效率的提高需要从多方面努力。

(1) 投资者对于市场信息应仔细甄别，避免盲目跟风，根据自身的风险承担水平构建适合自己的投资组合；提高认知水平，学习相关财务知识，能够做到财务报表的简单阅读，市场基本面信息的理解。

(2) 监管部门应加强建立与完善监管披露机制，让信息传递更具透明性与高效率。在现有科技社会，推进市场数据的信息化，利用大数据技术做到数据的及时披露，及时筛选并纠正市场上的错误信息，引导投资者理性投资；相反，该举措也能做到信息的向上反馈，让政策制定者更为清晰的了解到市场的实际运行情况，为国家宏观政策的制定提供依据。但同时也要做到对用户个人隐私的保护。

(3) 上市公司需严格遵守会计规则，做到财务报表的及时、准确、公平披露，不得误导投资者，并做好内幕信息的知情人登记工作，减少投资者层面上的信息不对称。

参考文献

[1] FAMA E F. The Behavior of Stock-Market Prices[J]. The Journal of Business, 1965, 38(1): 34-105.

[2] FAMA E F. Efficient Capital Markets: A Review of Theory and Empirical Work[J]. The Journal of Finance, 1970, 25(2,): 383-417.

[3] SHILLER R J. Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends?: w0456[R]. National Bureau of Economic Research, 1980.

[4] SHLEIFER A, VISHNY R W. Politicians and Firms*[J]. The Quarterly Journal of Economics, 1994, 109(4): 995-1025.

[5] BAMBER L S, BARRON O E, STOBER T L. Trading Volume and Different Aspects of Disagreement Coincident with Earnings Announcements[J]. The Accounting Review, 1997, 72(4): 575-597.

[6] BANZ R W. The relationship between return and market value of common stocks [J]. Journal of Financial Economics, 1981: 3-18.

[7] COCHRANE J H. Presidential Address: Discount Rates[J]. The Journal of Finance, 2011, 66(4): 1047-1108.

[8] 张元鹏. 投资者真的是理性的吗——行为金融学对法玛的 “市场有效假说” 的质疑与挑战[J]. 学术界, 2015(01): 116-125.

[9] 丁志国, 金博, 徐德财. 有效市场的检验——行为金融对 EMH 理论的批判[J]. 当代经济研究, 2017(03): 51-59.

[10] 汪荣飞, 张然. 基本面分析在中国 A 股市场有用吗?——来自季度财务报表的证据[J]. 金融学季刊, 2018, 12(01): 81-105.

[11] 赵琪, 徐维军, 季昱丞, 等. 机器学习在金融资产价格预测和配置中的应用研究述评[J]. 管理学报, 2020, 17(11): 1716-1728.

[12] 贾权, 陈章武. 中国股市有效性的实证分析[J]. 金融研究, 2003(07): 86-92.

[13] 吴振翔, 陈敏. 中国股票市场弱有效性的统计套利检验[J]. 系统工程理论与实践, 2007(02): 92-98.

[14] LIM K P, BROOKS R. Are Chinese stock markets efficient? Further evidence from a battery of nonlinearity tests[J]. Applied Financial Economics, 2009, 19(2): 147-

155.

[15] 屈博, 庞金峰. 基于序列性质检验的中国股票市场弱式有效性研究[J]. 浙江金融, 2014(11): 59-67.

[16] 赵国顺. 基于时间序列分析的股票价格趋势预测研究[D]. 厦门大学, 2009.

[17] 王劲松. 股市常用技术分析方法的有效性实证研究[D]. 西南财经大学, 2010.

[18] 石赛男. 股票技术分析中 MACD 指标的有效性检验[D]. 西南财经大学, 2011.

[19] 包懿. 中国股市技术分析有效性研究[D]. 上海交通大学, 2015.

[20] 张然, 汪荣飞, 王胜华. 分析师修正信息、基本面分析与未来股票收益[J]. 金融研究, 2017(07): 156-174.

[21] 常丹婷, 李峰. 价值因子背后的逻辑——估值-基本面预期差与错误定价[J]. 投资研究, 2020, 39(08): 142-159.

[22] OU J A, PENMAN S H. F Inancial Statement Analys Is and the Predict Ion of Stock Returns*[M]. 1988.

[23] ABARBANELL J S, BUSHEE B J. Fundamental Analysis, Future Earnings, and Stock Prices[J]. Journal of Accounting Research, 1997, 35(1): 1.

[24] ABARBANELL J S, BUSHEE B J. Abnormal Returns to a Fundamental Analysis Strategy[J]. The Accounting Review, 1998, 73(1): 19-45.

[25] PIOTROSKI J D. Value Investing: The Use of Historical Financial Statement Information to Separate Winners from Losers[J]. Journal of Accounting Research, 2000, 38: 1-41.

[26] MOHANRAM P S. Separating Winners from Losers among Low Book-to-Market Stocks using Financial Statement Analysis[J]. Review of Accounting Studies, 2005: 38.

[27] ASNESS C S, FRAZZINI A, PEDERSEN L H. Quality minus junk[J]. Review of Accounting Studies, 2019, 24(1): 34-112.

[28] BARTRAM S M, GRINBLATT M. Agnostic fundamental analysis works[J]. Journal of Financial Economics, 2018, 128(1): 125-147.

[29] 李斌, 林彦, 唐闻轩. ML-TEA: 一套基于机器学习和技术分析的量化投资算法[J]. 系统工程理论与实践, 2017, 37(05): 1089-1100.

[30] 李斌, 邵新月, 李玥阳. 机器学习驱动的基本面量化投资研究[J]. 中国工业经济, 2019(08): 61-79.

[31] FISCHER T, KRAUSS C. Deep learning with long short-term memory networks for financial market predictions[J]. European Journal of Operational Research, 2018, 270(2): 654-669.

[32] ELITH J, LEATHWICK J R, HASTIE T. A working guide to boosted regression trees[J]. Journal of Animal Ecology, 2008, 77(4): 802-813.

[33] MICROSTRONG. Regression Tree 回归树[J]. 人工智能, 2019/9/16 上午 2:05:19.

[34] MICROSTRONG. 深入理解提升树（Boosting tree）算法[J]. 人工智能, 2019/10/1 上午 4:42:28.

[35]『机器学习笔记』GBDT 原理-Gradient Boosting Decision Tree_AaronChou 的博客-CSDN 博客[M]. 2017.

[36] FRIEDMAN J H. Greedy function approximation: A gradient boosting machine.[J]. The Annals of Statistics, 2001, 29(5).

[37] GBDT：梯度提升决策树 - 简书[M]. 2016.

[38] MICROSTRONG. 深入理解 LightGBM[J]. 人工智能, 2020/1/4 下午 9:19:46.

[39] Optimization - XGBoost Loss function Approximation With Taylor Expansion[J]. Cross Validated, 2016.

[40] 罗琦, 游学敏, 吕纤. 基于网络数据挖掘的资产定价研究述评[J]. 管理学报, 2020, 17(01): 148-158.

[41] ROSS S. A simple approach to the valuation of risky streams[J]. Journal of Business, 1978, 51(3): 453-475.

[42] DANIEL K, HIRSHLEIFER D, SUBRAHMANYAM A. Investor Psychology and Security Market Under- and Overreactions[J]. The Journal of Finance, 1998, 53(6): 1839-1885.

[43] GRINBLATT M, HAN B. Prospect theory, mental accounting, and momentum[J]. Journal of Financial Economics, 2005, 78(2): 311-339.

[44] YAN X S, ZHENG L. Fundamental Analysis and the Cross-Section of Stock Returns: A Data-Mining Approach[J]. The Review of Financial Studies, 2017, 30(4): 1382-1423.

致谢

回想起在武汉的七年生活，总会感觉到不可思议，感谢幸运之神的眷顾。

从初三开始，我对自己的定位就非常清晰：不是那种整天只会学习的那种书呆子，是特别喜欢上网冲浪紧跟潮流的那种“不良少女”，娱乐活动是我生活中不可或缺的一部分。

人生最大的转折点便是 2014 年，在没有竞赛基础的情况下误打误撞过了华师一附中的专县生考试。在考试的前两晚，我爸在家里拿着在淘宝上买的华师一专县生真题给我讲到了凌晨两点，还记得考数学的时候，选择题都是靠代入选项猜的，后面一看竟然都蒙对了。妈妈在食堂等我的时候遇到了一群从宜昌城区来考试的家长们，但他们都是学校统一培训过，并且组织带队过来的，所以也瞧不起从县城过来的我们。

高中开始，妈妈和大姨一起到武汉来工作和陪读，误打误撞赶上了好时候，自此在武汉有了真正意义上的家。特别感谢我的父母，成功从小县城跃居到了武汉，高中和大学期间没有让我吃过亏，在花钱方面从不吝啬，也从不限制我的发展方向。

之后便是三年丰富多彩的高中生活，感谢我的母校华师一附中，让我深刻体会到了“把时间还给学生，把方法教给学生”的教育模式，在校期间大多数时间还是非常快乐的，以至于到现在都能受益。另一方面，感谢优秀的高中同学们，都是超高质量的人脉资源。班上一共有 7 个人在武大，可以经常聚会或是喊着跑腿，大学期间从未孤独。

感谢高二的同桌陈同学，整个高三生活都因为我们而变成了彩色，怀念总是给来很晚的你打水的早晨，拿现金去吃洪湖人家铁板牛蛙的中午，或是一起走敏行环路吃外卖的晚上。虽然天各一方，未来过去我只想见你。

妈妈在怀孕的时候梦到在樱花树下打滚，或许冥冥之中注定着我要来武汉大学。很庆幸自己在填志愿时还是选择了武大，有着体面的专业，在学生工作和课外活动方面都很尽兴。不得不感叹在武汉读书简直太方便了，坐上牌坊的 586 就可以回家，说着湖北塑普也没有突兀的感觉，还有好多在武汉读书的高中同学的陪伴。

当然，最感谢的还是本科期间认识的老师和同学们。感谢两个辅导员以及班

导马亮老师，帮助我作为班长做好 1703 班的学生工作；感谢刘岩老师，在大三的时候跟随参与了因果推断研讨班，参加了 CBD 数据库的建设工作，免费蹭了很多饭，领了工资，在指导下完成了保研论文，以及无数的推荐信；感谢李斌老师，同样也是无数的推荐信，大四进组蹭了好多次活动，入门了量化投资领域，加入了金融科技研讨班，并完成了这篇毕业论文；最后感谢我可爱的大学同学们，一起吃了好多好多顿饭，特别是金工姐妹花孔彤阳同学，一起参加了很多比赛，八了很多卦，合作了很多项目，以及硕士阶段能一起去到人大。

我是一个很有仪式感的人，但也是一个多愁善感很爱哭的人，从小到大，只要遇到很令人激动或者感动的事情都会忍不住掉眼泪，甚至每次和别人吵架的时候也会忍不住哭，我妈说是我太过于娇气。由于华师一并不是高考考点，大家都被分到了水高、十四中等学校，在高考前几天便放假了，大家准备回家进行最后的冲刺，没有人说再见。还记得当时我用教室上面的电脑播放着 closer，一边唱一边收拾着东西，到了要走的时候，教室里的人已经寥寥无几，还剩下坐在角落的梁同学，便跑过去抱着她一边哭一边说高考加油呀。现在就能预想到毕业典礼当天我穿着学士服不知道抱着哪个同学失声痛哭的样子。

写完论文便是毕业，便是和一段人生告别。在追梦过程中，我获得的已经比梦想本身更多，下一站——北京见。

附录 A　279 个财务指标及其缺失值比例

CSMAR 指标代码	缺失值比例	CSMAR 指标代码	缺失值比例	CSMAR 指标代码	缺失值比例
B001000000	0.000121	A001124000	0.289429	C0i1017000	0.579615
B001300000	0.000121	A002209000	0.297673	C0i1005000	0.580208
B001100000	0.000976	A001214000	0.309606	C0i1006000	0.58089
A001000000	0.001013	A002205000	0.31065	C0i1007000	0.581362
A004000000	0.001018	A001120000	0.313851	D000111000	0.623499
A003100000	0.001044	B006000000	0.314638	A0b1201000	0.629681
A002000000	0.001091	A001215000	0.316659	A002210000	0.630673
A003101000	0.001186	B006000101	0.318333	D000107000	0.650545
A001212000	0.001212	A003106000	0.337113	A0F3109000	0.663511
A003000000	0.001427	C003004000	0.338729	D000112000	0.682659
A003105000	0.001532	C002005000	0.340173	A003111000	0.708262
A002113000	0.001579	C003008000	0.34511	A001206000	0.716075
B002000101	0.001632	C003001000	0.352877	B002200000	0.717014
B002000000	0.001642	B001302101	0.35963	B002300000	0.717172
A003102000	0.002545	A001220000	0.35995	A0f1108000	0.722619
A002112000	0.003148	C002010000	0.366504	A0f2106000	0.72296
A001101000	0.004177	B006000102	0.374034	B0f1213000	0.724639
A003103000	0.004424	A001211000	0.378999	D000108000	0.727142
A001218000	0.004906	A001202000	0.418612	B0f1105000	0.727772
A001111000	0.009015	A001107000	0.430329	B0f1208000	0.729356
B001200000	0.010579	B001301000	0.436117	A0f1300000	0.729603
A002120000	0.010988	D000100000	0.440793	A0f2300000	0.729719
A001121000	0.011251	D000101000	0.441312	A0b2103201	0.734914
A002100000	0.011833	D000114000	0.441517	A0b1103000	0.735228
A001100000	0.011838	D000115000	0.441827	A0d1126000	0.735286
A001200000	0.012227	D000103000	0.442036	A0b1104000	0.735433
A001112000	0.013476	D000200000	0.445815	A0b2103101	0.735564
B001101000	0.016493	D000204000	0.450894	B0d1104201	0.735926
B001207000	0.016718	D000113000	0.450952	B0d1104301	0.736131
A001123000	0.016844	D000205000	0.451298	A0d1101101	0.736168
A001222000	0.017191	B005000000	0.451797	B0d1104101	0.736215
A002200000	0.021095	D000109000	0.45887	A0d1102101	0.736399
B001211000	0.02131	D000104000	0.464616	A0b1105000	0.73834
B001210000	0.022134	A003102101	0.466605	C007000000	0.739132
B001400000	0.023713	A001219000	0.470562	A0d1218101	0.739138
A001125000	0.023965	A001204000	0.471464	B0i1204101	0.7394
B001500000	0.025938	B001500101	0.472571	B0i1203101	0.739505

CSMAR 指标代码	缺失值比例	CSMAR 指标代码	缺失值比例	CSMAR 指标代码	缺失值比例
B001201000	0.027775	C003001101	0.472608	A0d2101101	0.739537
B002100000	0.029076	C003006101	0.480584	B0i1208103	0.739537
B001209000	0.038438	Bbd1102000	0.480731	A0i2116000	0.739552
C001000000	0.043643	D000102000	0.481938	A0i1209000	0.739558
C005000000	0.044173	B0d1104000	0.484877	B0i1103303	0.739558
C001020000	0.044483	D000110000	0.485727	A0i2124000	0.739563
C001022000	0.044918	C002009000	0.485764	B0i1103203	0.739573
C001021000	0.045097	B0i1103000	0.493871	B0i1203203	0.739573
C002000000	0.046293	B0i1204000	0.493923	B0i1204203	0.739573
A001213000	0.046603	B0i1203000	0.494495	A0i2118000	0.739678
C006000000	0.048739	B0i1103101	0.495088	A0i1225000	0.739678
C005001000	0.04887	C002004000	0.496489	B0i1103111	0.739746
C001013000	0.049368	A002105000	0.510038	A0i2117000	0.739788
C002006000	0.049484	D000106000	0.51388	A0i2119101	0.739809
C001001000	0.055093	A001109000	0.530582	A0i1224000	0.739836
C001014000	0.055891	A001203000	0.531039	A0i2119201	0.739888
A001221000	0.0565	A003107000	0.531726	A0i1116201	0.739888
C003000000	0.058373	D000105000	0.532204	A0i1116101	0.739888
A002109000	0.059386	A001216000	0.535258	A0i2119401	0.739967
A002108000	0.059931	B001304000	0.537399	A0i2119301	0.739967
B001302000	0.062099	C003003000	0.543097	A0i1116401	0.739972
A003200000	0.065835	A001207000	0.544488	A0i1116301	0.74003
A002101000	0.066449	A001217000	0.546545	A0F3108000	0.740088
C003006000	0.068275	A0f3104000	0.549415	A0i1210000	0.740103
A001110000	0.090797	D000116000	0.551588	A0d2202000	0.740114
A002206000	0.093898	B001303000	0.554668	A0i1115000	0.740523
A001223000	0.097393	Bbd1102101	0.556568	B001305000	0.75107
A002107000	0.101103	C0f1009000	0.559239	D000206000	0.776893
B002000201	0.111908	A0f1122000	0.559648	D000207000	0.779165
A002208000	0.122145	A0f2110000	0.562476	D000203000	0.783526
C003005000	0.125981	C0f1018000	0.562796	D000201000	0.786627
A001205000	0.126847	Bbd1102203	0.564045	D000202000	0.789655
C003002000	0.145686	B0d1104401	0.56425	B001211203	0.821875
A002201000	0.151096	A0f2104000	0.564995	B001308000	0.821917
A002114000	0.15776	C0b1015000	0.567262	B001216000	0.839711
A002125000	0.158993	A0f1106000	0.567629	B001211101	0.841007
C002003000	0.165983	B0d1104501	0.568527	B001500201	0.84393
A002204000	0.207558	A0d2122000	0.570647	D000117000	0.853428
B003000000	0.214963	A0b2102000	0.572032	B001400101	0.85921
A002126000	0.215472	A0i2111000	0.573155	B001307000	0.877986
C001012000	0.218615	C0i2008000	0.57348	A003112000	0.915421
C002002000	0.21947	C0d1011000	0.573612	A001228000	0.9243

CSMAR 指标代码	缺失值比例	CSMAR 指标代码	缺失值比例	CSMAR 指标代码	缺失值比例
C002007000	0.221459	C0b1002000	0.573732	A003112201	0.928598
B004000000	0.223038	A0d2123000	0.574252	A003112101	0.933121
A002115000	0.228737	A0i1113000	0.574273	A002127000	0.935703
A002207000	0.23324	A0i1114000	0.574593	A003112301	0.935986
A001119000	0.236656	C0b1016000	0.574834	A002128000	0.940788
C004000000	0.239379	B0i1206000	0.575055	A001127000	0.949598
C003007000	0.261014	B0i1205000	0.575091	A001229000	0.959726
B001212000	0.261382	A0i2121000	0.575112	D000118000	0.982552
A002203000	0.263124	C0d1010000	0.575133	A001128000	0.983875
C002001000	0.266498	B0i1202000	0.575144	A001226000	0.991331
A0b2103000	0.281175	C0d1008000	0.577652	A001227000	0.99449
A0d1102000	0.28233	C0b1003000	0.577752	B001302201	0.995603
A0i2119000	0.285546	C0b1004000	0.578303	B001306000	0.999769
A0i1116000	0.286024	C0i1019000	0.579541	B002000301	0.999895

附录 B　缺失值比例小于 5% 的 51 个财务指标

CSMAR 指标代码	指标名称
A004000000	负债与所有者权益总计
B001101000	营业收入
B001200000	营业总成本
A002100000	流动负债合计
B001100000	营业总收入
A002000000	负债合计
A001218000	无形资产净额
A001222000	递延所得税资产
A001000000	资产总计
A001100000	流动资产合计
C002006000	购建固定资产、无形资产和其他长期资产支付的现金
A001200000	非流动资产合计
C002000000	投资活动产生的现金流量净额
A001101000	货币资金
C006000000	期末现金及现金等价物余额
A001123000	存货净额
A001112000	预付款项净额
A001213000	在建工程净额
A002112000	应付职工薪酬
A002200000	非流动负债合计
C001020000	支付给职工以及为职工支付的现金
C001013000	收到的其他与经营活动有关的现金
A002113000	应交税费
A001125000	其他流动资产
A002120000	其他应付款
C001021000	支付的各项税费
A001121000	其他应收款净额
B001201000	营业成本
B001211000	财务费用
B001000000	利润总额
C001022000	支付其他与经营活动有关的现金
B001400000	营业外收入
C005000000	现金及现金等价物净增加额
B001500000	营业外支出
B001300000	营业利润
C001000000	经营活动产生的现金流量净额
B001207000	营业税金及附加

CSMAR 指标代码	指标名称
B002100000	所得税费用
C005001000	期初现金及现金等价物余额
A001212000	固定资产净额
B001209000	销售费用
A001111000	应收账款净额
A003102000	资本公积
B002000000	净利润
A003100000	归属于母公司所有者权益合计
A003000000	所有者权益合计
A003103000	盈余公积
B001210000	管理费用
A003105000	未分配利润
B002000101	归属于母公司所有者的净利润
A003101000	实收资本 (或股本)

附录 C　因子模型线性回归结果

模型	系数	Q1	Q2	Q3	Q4	Q5	Q5-Q1
CAPM	*mean*	0.073	0.231	0.255	0.654	0.767	0.694
	α	0.225	0.415	0.391	0.628	0.879	0.685
	t 值	(2.18)	(4.26)	(4.16)	(6.81)	(9.15)	(2.22)
FF3	*α*	0.312	0.505	0.481	0.724	0.977	0.688
	t 值	((3.06)	(5.25)	(5.19)	(8.00)	(10.33)	(2.20)
FFC	*α*	0.410	0.602	0.573	0.791	1.018	0.626
	t 值	(4.03)	(6.28)	(6.20)	(8.74)	(10.74)	(2.01)
FF5	*α*	-1.186	-1.172	-0.941	-0.762	-0.143	1.114
	t 值	(-11.68)	(-10.65)	(-9.86)	(-7.02)	(-2.94)	(3.04)
FF5+mom	*α*	-1.285	-1.085	-0.959	-0.661	-0.268	1.073
	t 值	(-10.71)	(-9.61)	(-8.81)	(-6.19)	(-2.39)	(2.96)
Q	*α*	0.351	0.544	0.525	0.763	1.017	0.706
	t 值	(3.44)	(5.66)	(5.67)	(8.43)	(10.75)	(2.40)