目录

姓名：朱晗、冯高阳　学号：2022110901007、2022110902012
专业：数据科学与大数据技术　得分：

多元统计分析作业3：上证指数的回归分析

一、问题

把目光投向中国金融市场时，**上证指数**常被视为衡量中国股市整体走势的风向标，它所记录的涨跌背后，不仅蕴含了国内市场自身的结构变迁与情绪流动，也隐含着国际资本联动与宏观经济政策指向的暗流涌动。倘若不仅仅只是简单地观察其上下起伏，而想进一步探寻背后的动力机制，就会发现像恒生指数、人民币对美元汇率、标普500指数、国际原油价格以及各类宏观经济指标等，都可能在悄然影响着上证指数的未来走向。于是，本次研究尝试利用回归分析方法，以2024年的数据为背景，将这些潜在驱动因素一并纳入考虑范围，从而构建一个能更好解释上证指数走势的模型。

正式展开分析之前，有必要对上证指数以及相关的分析框架予以简要回顾。上证指数由上海证券交易所编制，其变动蕴含了企业盈利水平、市场风险偏好、政策预期与国际资本流动的综合效应。在当今环环相扣的全球化经济背景下，当外部市场风起云涌、人民币汇率波动、国际金融环境持续调整时，上证指数也不免被各类外因牵动。回归分析作为常用的统计与计量工具，有助于将包括股市指数、汇率、国际大宗商品价格及宏观经济数据在内的多元要素整合进同一分析框架，以探究其与上证指数的相关程度与影响方向。

二、分析思路

本次研究以上证指数为核心目标，希望通过多元线性回归与*Lasso*回归等统计手段，探索各类外部因素对这一指数的影响。整个过程的逻辑脉络主要分为数据处理、模型构建与性能验证，以及辅助信息呈现三部分。

在数据处理环节，我们从多份数据集出发，将上证指数与恒生指数、沪深300、科创50、美元指数、中证*A*100等多维市场指标同处一室，并确保它们在同一时间尺度上对齐。这就如同在一张有序的数据“棋盘”上，为每个变量找到各自对应的棋位，以便后续的建模能够在一个整齐划一的框架内运行。

其次，辅助信息的呈现（如人民币汇率中间价的时间序列图）为理解上述回归关系提供了宏观层面的参考背景。即便这些汇率数据不直接被输入模型，它们的变化趋势却可能在现实中影响市场资金流向与风险偏好，从而间接左右各指数间的联动。这些图表的目的是协助在解释模型结果时不至于“盲人摸象”，而是能够稍有余地回头看看大环境，为后续进一步的经济学解释与策略制定提供一些更宽泛的坐标系。

接下来，模型构建与性能验证的核心工作是将这些处理过的数据投入回归模型中考验。多因素线性回归模型为我们搭建了一个初步的关联框架：上证指数作为因变量，自变量则是其他市场指标的当期数据。我们能够以*MSE*等指标客观评价模型对未来数据的预测能力。此步骤的意义在于，以严谨可量化的方式检验模型的有效性，而不仅仅停留在主观直觉。之后的*Lasso*回归尝试则通过正则化手段精简模型，将无助于预测的冗余因素“剔除”出局，为后续分析者提供一种从复杂多元中提炼关键因素的操作思路。相较于多因素线性回归，*Lasso*回归为我们展示了另一种更为“精炼”的建模路线。

本次分析的思路在于将多元数据结构化处理，通过多因素与正则化两类回归模型比较与验证，并借助背景信息的补充来提升对结果的理解。

三、程序

操作步骤与实现过程可分为数据准备、模型构建与评估、以及结果展示三个环节。

在数据准备阶段，程序通过一系列标准化处理与目录检查，确保输出图表与结果文件有序存放，并利用 *pandas* 从事先下载并保存于本地的 *CSV* 文件中读取各类相关金融数据集。

```
sh_data = ef.stock.get_quote_history('000001', beg='2023-01-01', end='2023-12-31')
sh_data.to_csv('../dataset/上证指数_2024.csv', index=False)

# 使用 akshare 获取恒生指数数据（示例：恒生指数代码为'HSI'）
hsi_data = ak.stock_hk_index_daily(symbol='HSI')
hsi_data = hsi_data.rename(columns={'日期': '日期', '收盘': '收盘', '最高': '最高', '最低': '最低', '开盘': '开盘'})
hsi_data.to_csv('../dataset/恒生指数_2024.csv', index=False)
csi300_data = ef.stock.get_quote_history('000300', beg='2023-01-01', end='2023-12-31')
csi300_data.to_csv('../dataset/沪深 300_2024.csv', index=False)
star50_data = ef.stock.get_quote_history('000688', beg='2023-01-01', end='2023-12-31')
star50_data.to_csv('../dataset/科创 50_2024.csv', index=False)
dxy_data = ak.fx_quote_usdx("2023-01-01", "2023-12-31")
dxy_data.to_csv('../dataset/美元指数_2024.csv', index=False)

usdcnh_data = ak.fx_spot_quote('USDCNH', start_date='2023-01-01', end_date='2023-12-31')
usdcnh_data.to_csv('../dataset/USDCNH_2024.csv', index=False)
usdcny_data = ak.fx_spot_quote('USDCNY', start_date='2023-01-01', end_date='2023-12-31')
usdcny_data.to_csv('../dataset/USDCNY_2024.csv', index=False)
rmb_mid_data = ak.fx_rmb_rate(symbol="美元/人民币", start_date="2023-01-01", end_date="2023-12-31")
rmb_mid_data.to_csv('../dataset/人民币汇率中间价_2024.csv', index=False)
```

这些数据包括上证指数、恒生指数、沪深 300、科创 50、美元指数、中证 *A*100 以及人民币汇率中间价等。程序首先定义了 *load_datasets*()函数，将所有数据文件分别读入内存，并存入字典结构中便于后续调用。

```
def load_datasets():
    datasets = {
        "上证指数": pd.read_csv('../dataset/上证指数_2024.csv'),
        "恒生指数": pd.read_csv('../dataset/恒生指数_2024.csv'),
        "沪深 300": pd.read_csv('../dataset/沪深 300_2024.csv'),
        "科创 50": pd.read_csv('../dataset/科创 50_2024.csv'),
        "美元指数": pd.read_csv('../dataset/美元指数_2024.csv'),
        "中证 A100": pd.read_csv('../dataset/中证 A100_2024.csv'),
        "USDCNH": pd.read_csv('../dataset/USDCNH_2024.csv'),
        "USDCNY": pd.read_csv('../dataset/USDCNY_2024.csv'),
        "人民币汇率中间价": pd.read_csv('../dataset/人民币汇率中间价_2024.csv'),
    }
    return datasets
```

在模型构建与评估环节中，程序实现了两个主要模型：多因素线性回归（*Multiple Linear Regression*，*MLR*）与 *Lasso* 回归模型。对于多因素线性回归部分，*multiple_linear_regression*(*datasets*)函数通过将上证指数数据与包括恒生指数、沪深 300、科创 50、美元指数、中证 *A*100 在内的多个自变量数据进行合并匹配，确保各自变量与因变量在同一日期对齐。随后，将合并后的数据集划分为训练集与测试集，通过 *LinearRegression* 模型进行拟合，最终利用测试集数据进行预测，并计算均方误差（*MSE*）作为模型评估指标。程序还将实际值与预测值以图形的方式绘制出来，以直观表现模型对市场走势的拟合情况，并将生成的图表输出至指定目录。

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
model = LinearRegression()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
mse = mean_squared_error(y_test, y_pred)
print(f'MSE: {mse}')

lasso = LassoCV(cv=5, random_state=42)
lasso.fit(X_scaled, y)
plt.scatter(X, y, color='blue')
plt.plot(X, lasso.predict(X_scaled), color='red')
plt.title('Lasso 回归')
plt.xlabel('恒生指数')
plt.ylabel('上证指数')
plt.savefig(f'../output/lasso_1/lasso_恒生_上证.png', dpi=300)
```

Lasso 回归环节中，*lasso_regression*(*datasets*)函数则以上证指数与恒生指数这两个变量为例进行说明。程序先对数据进行日期对齐与合并，然后将恒生指数作为自变量，标准化处

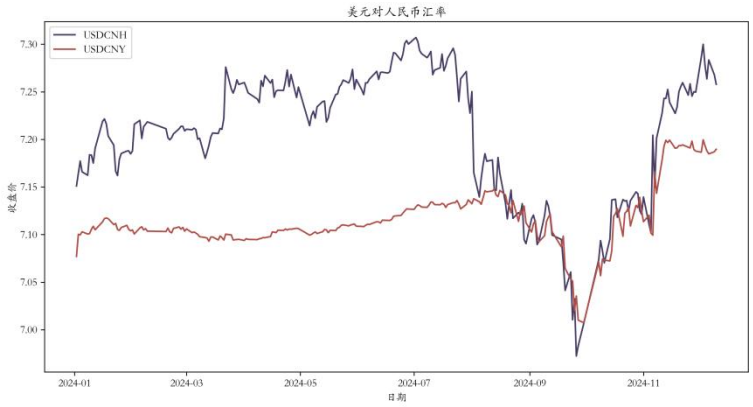
理后输入 *Lasso* 回归模型进行训练。通过 *LassoCV* 方法自动选择正则化参数，程序获得了较为稳健的模型参数。将预测直线与实际数据点在同一坐标系中绘制，使读者可直接观察 *Lasso* 回归模型对两者间关系的拟合表现。

```
x = pd.to_datetime(df.index.to_numpy())
plt.figure(figsize=(10, 8))
for col in df.columns:
    plt.plot(x, df[col], label=col)
plt.xticks(ticks=x[::15], labels=df.index[::15], rotation=45)
plt.title('人民币汇率中间价')
plt.legend(loc='upper center', bbox_to_anchor=(0.5, -0.15), ncol=7)
plt.tight_layout()
plt.savefig('../output/display/人民币汇率中间价.png', dpi=300)
```

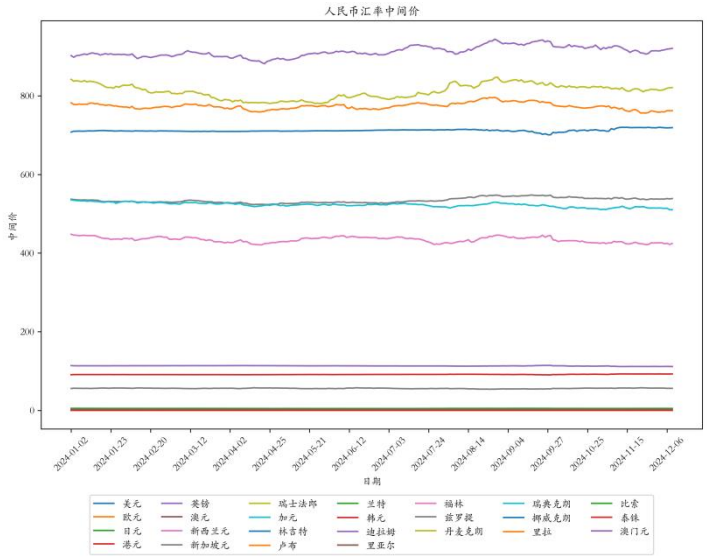
最后，为进一步展示数据特点与多维度指标间的关系，程序绘制人民币汇率中间价随时间变化的趋势图。该函数从数据集中选取相关列后，以折线图的形式直观呈现不同货币对的中间价变化情况，并将结果图表输出到指定目录。

四、结果与分析

选取了具有代表性的三类数据进行初步可视化，展示了从不同维度对上证指标所进行的初步探查，为后续数据分析与模型构建提供参考。

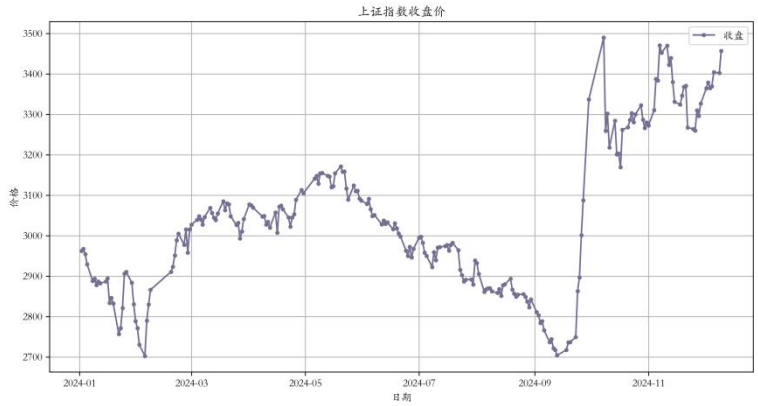


首先，关于美元对人民币的汇率对比图，可以观察到 *USDCNH*（离岸人民币）与 *USDCNY*（在岸人民币）在时间轴上的价格变动轨迹。两条曲线在整体趋势上具有一定的相似性，但仍存在价差和波动差异。这种差异通常源于离岸与在岸市场间的交易机制、市场预期及资金流动性差异。



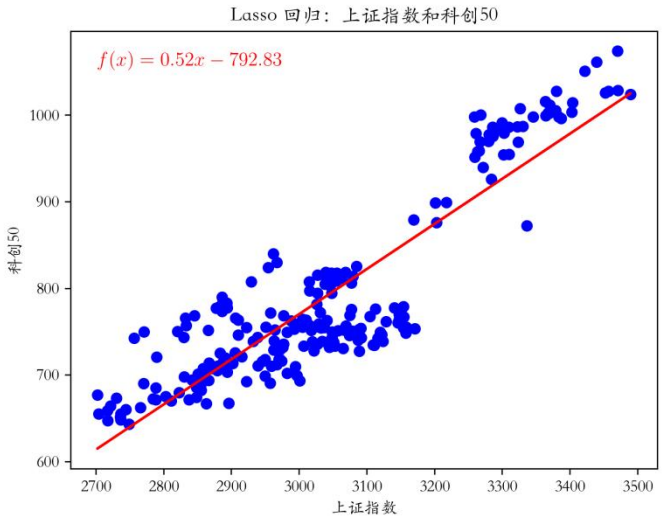
在人民币汇率中间价的多元对比图中，不同颜色的曲线对应美元、欧元、英镑、港元、

日元、新加坡元、瑞士法郎等多种国际货币对人民币的中间价走向。通过这些多币种的汇率中间价对比，可以一览人民币在多个国际货币体系中的相对强弱与价值变动。



上证指数的收盘价走势图刻画出在所选时间范围内中国 *A* 股市场整体表现的变化趋势。线图上价格的上下波动反映了市场情绪、经济数据和政策信息的交织影响。从曲线可见，指数在某些时段可能经历了明显的上行或下行调整，这些拐点往往与宏观经济政策、国际市场联动、企业盈利预期以及投资者情绪的转变相关。

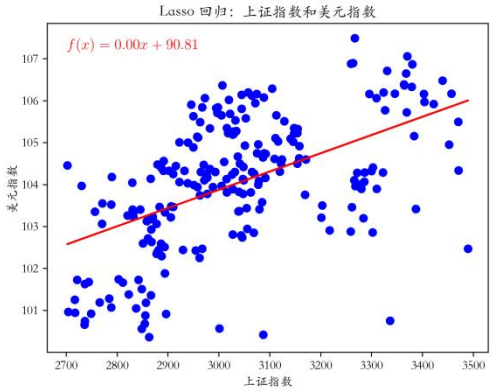
运用 *Lasso* 回归，对上证指数与多个指标（科创 50、美元指数、中证 *A*100、恒生指数）之间的线性关系进行拟合的结果。通过在散点图中叠加一条由 *Lasso* 回归估计的线性拟合直线，我们可以在一定程度上判断各自变量与上证指数间的相关关系强弱与方向性。以下分别对几张图表进行描述与解释：

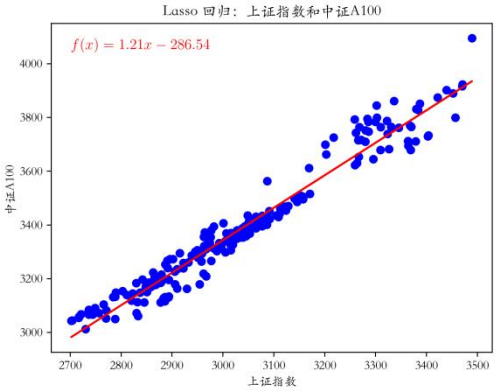


图中红色拟合直线的斜率约为 0.52，截距约为-792.83。换言之，两者之间存在正相关关系。考虑到科创 50 指数是代表中国科技创新型企业板块的重要参考指标，与上证指数中许多传统行业成分股有所不同，其表现某种程度上反映了 *A* 股科技成长性与市场对创新企业前景的预期。市场风险偏好提升、流动性改善或宏观经济预期乐观时，既能带动传统板块的上涨，也倾向于同时推高科技创新领域的估值。

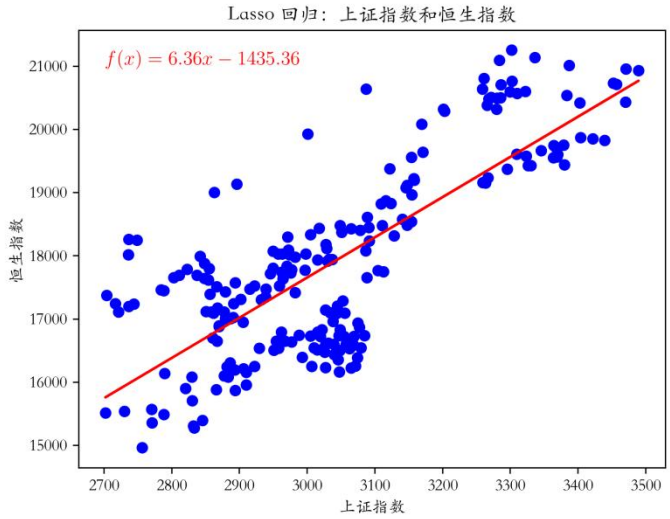
在上证指数与美元指数的拟合中，斜率几乎接近 0，截距约为 90.81。这种结果表明，上证指数与美元指数之间在本数据环境下的线性相关度不高。理论上，美元指数反映国际美元相对于一篮子货币的强弱，一般而言，若美元趋强，国际资金倾向流出新兴市场或风险资产，从而对 *A* 股可能产生间接影响。但本图中接近水平的拟合线与高度分散的点云表明，在所选时间段和数据样本下，这种宏观逻辑并未明显体现出来，或被其他更强烈的市场因素

所中和。这从科学解释的角度来看，可能意味着样本期间内 *A* 股走势更多受到本土政策、产业结构调整和投资者结构变化等因素驱动，而非仅由国际汇率或全球资金流向决定。





此图中拟合直线的斜率约为 1.21，截距约为-286.54。与科创 50 相比，中证 *A*100 是从大样本中精选出的代表性强、流动性好的蓝筹或核心资产指数，因此与上证指数的基础成分更为接近。这种强烈的正向相关关系说明两者走势高度同步。上证指数和中证 *A*100 都对整体市场流动性、政策导向、企业盈利预期保持敏感，当市场整体趋强时，高质量股票群体所组成的 *A*100 表现更为积极，且涨幅可能相对更高，体现出 *A* 股核心资产在多头行情中获得更强劲的支持。



考虑到恒生指数以香港市场为基础，尽管与内地 *A* 股市场有所区隔，但由于地理及资金流动渠道的原因，中港资本市场之间存在一定联动。当上证指数趋于走强时，香港市场往往受到溢出效应与情绪传递的影响，同步出现上行趋势。这种高倍率的斜率可解释为恒生指数基期数值整体较高，因此同样幅度（点数）变化上，恒生指数的绝对涨幅看起来较大。不过无论如何，这一明显的正相关在经济层面上反映出内地与香港股市间日益紧密的资本与信息交流，对外资配置中国资产的渠道扩展与政策导向，以及两地投资者情绪关联度的提升。

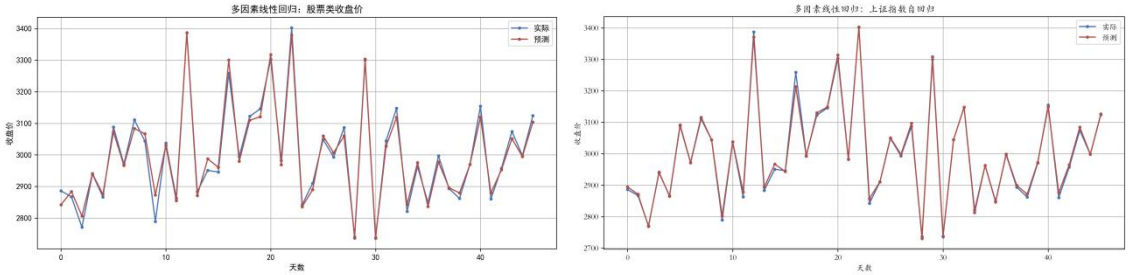
这些基于 *Lasso* 回归所得的拟合直线并不能单纯表明因果关系，而仅是线性相关度的一种统计刻画。*Lasso* 回归的使用在这里主要帮助我们进行特征选择与模型简化，以减少多重共线性和过拟合的问题。当系数较大且显著时，例如上证指数与中证 *A*100 的关系，我们更容易获得稳定的线性关联；而当系数接近 0 或点云分布无明显线性态势时（如与美元指数的关系），则说明该特征在模型中解释力有限或不稳定。从经济金融学角度而言，不同指数之间的正相关往往体现了共同的市场驱动力，如流动性环境、投资者风险偏好和宏观经济预

期等；指数间相关度的差异化则显示出不同市场或板块对国际环境、政策调控与产业结构变化的敏感度不同。

随后在 *LASSO* 回归后进行了**多元线性回归分析**，回归模型的整体表现表明，所选取的多个因素共同对上证指数的走向具有相当高的解释能力，这些自变量的组合足以对上证指数的波动趋势给出相对紧密的拟合。在当前的研究框架下，上证指数的短期或中期变化，并非孤立地依赖某一单一因素，而是受到多重市场和宏观变量的合力影响。

变量名	系数	标准差	*t* 值	*P* 值	[0.025]	[0.975]
const	97.7243	130.568	0.748	0.455	-159.64	355.088
恒生指数	-0.0021	0.002	-0.918	0.359	-0.007	0.002
沪深 300	1.6217	0.12	13.472	0	1.384	1.859
科创 50	-0.1204	0.043	-2.811	0.005	-0.205	-0.036
美元指数	3.3879	1.45	2.336	0.02	0.53	6.246
中证 *A*100	-0.8986	0.119	-7.579	0	-1.132	-0.665

然而，当深入到具体的回归系数层面，并非所有自变量都在统计上对上证指数表现出明确的影响，其中有些变量即使理论上应与上证指数相关，却在本模型中未能达到显著性标准。出现这种情况的原因可能是多方面的：一方面，所选时间段和数据结构可能使某些传统上重要的指标在样本期内并未突出其应有的影响力；另一方面，不同指数间在市场映射上可能存在重叠和替代效应，当多个对市场有相似解释力的指标同时进入模型时，模型难以清晰区分它们各自的边际贡献。部分指标呈现与直观预期不符的关系倾向。例如，当直观上某一指标应与上证指数同向变动时，实际结果却显示出相反方向的关系。潜在原因包括：多重共线性导致系数不稳定，各因素间存在高度相关性使得模型在系数估计上出现偏差；或是样本期内发生了一系列特殊事件与政策变化，影响了市场资金流向，使传统逻辑暂时失效。



综合而言，多元线性回归在整体拟合上取得了优异表现，某些自变量与上证指数呈现出显著且合理的联系，特别是与国内市场密切相关的指标，对上证指数的正相关性和高显著性不仅反映了国内资本市场中核心资产群体与整体市场的联动趋势，也为后续深入研究资本配置、政策影响和投资者行为的机理提供了实证支撑。从实务角度看，这样的结论帮助研究者和投资机构在资产配置时更有据可循，能在市场环境变化时快速根据相关指数动态调整投资策略。

五、难点与联想

在本次分析中，最大的难点之一是数据的前期处理和多重共线性问题。由于不同市场指标的获取途径和数据格式不统一，尤其是国际金融数据和国内指数的时间序列对齐工作，往往需要耗费大量精力进行数据清洗与整理。此外，多重共线性在多元回归模型中的存在使得部分自变量的边际影响难以准确量化。这种现象尤其在与国内市场密切相关的指数（如沪深 300 与中证 *A*100）中体现得尤为明显，它们之间高度相关的特性可能导致模型中回归系数的不稳定甚至偏误，从而影响最终结果的解读。因此，在建模中引入 *Lasso* 正则化手段，不

仅简化了模型，同时也为后续分析提供了更可靠的特征选择方法。

本次结果启发我们，在实际的投资分析和市场研究中，单一维度的指标往往不足以全面反映市场走势，系统性整合多维数据可能是提升预测能力的有效途径。此外，研究还显示，市场的内外联动关系并非一成不变，具体影响机制会因时间周期、政策环境和国际市场动荡等外生变量而改变。未来若将时间序列分析与非线性回归模型（如机器学习算法）结合，可能更好地捕捉复杂市场动态。