2021 年第二届“大湾区杯”粤港澳

全融数学建模竞赛

题 目　券商研报对公司股票走势的影响和投资策略

摘　　要：

针对问题一，针对问题一，因研报提供的影响因素数据较多，为选出有效且代表性较强的特征因子，构建了**单因子选股模型**。首先进行因子的有效性检验，利用 IC 值等进行初步的因子筛选；又考虑特征因子之间存在多重共线性，选用方差膨胀因子 VIF 来进行因子筛选，确定得到一系列有效性较高且具有一定代表性的子因子。最终利用等权法处理不同因子大类下的子因子，得到四个大类因子，分别是估值因子、盈利能力因子、成长因子、规模因子。

针对问题二，通过上述的指标特征提取，确定了影响股票走势的因子。股票的价格是序列相关的，且受到多个指标特征的影响，由此选取三支股票（粤水电、顺络电子、丽珠集团），并采用多元线性回归模型对这三支股票进行建模分析，并预测这三支股票的未来股价。通过计算软件进行**多元回归建模**和预测，得出这三支股票受指标影响，使得走势有所波动，但整体呈现上涨趋势，最后用**多因子模型**来预测股票的未来趋势。因此可下结论，提取的特征指标整体与该三支股票走势是正相关，并且未来短期内可以进行投资。

针对问题三，研报文本可以反应出突发事件的闪现、 舆情、自然灾害等对股票行情的影响。通过对研报的**文本挖掘**，在这里建立基本面类、情绪类、概念主题类三大热词库，其中将基本面类热词、情绪类热词分为正负面两面词库。采用向量空间 **VSM 模型**建立词库。通过对大湾区指数股的研报建立文本向量，考虑使用 TF-IDF 计算词频权重。再分别计算在对应单词下，进行 VSM 分类时正确的概率，以及股票的超额。通过计算研报中单词的对应词频的收益指标的均值来判断事件的闪现、 舆情、自然灾害等等事件对股票行情的影响。

针对问题四，证券公司股票走势需综合考虑公司的内部财务特征以及外部的环境影响。在问题二策略模型中的收益函数上加入外界因素的影响因子，由于外部环境对股票行情的影响也是很重要的一部分，这里考虑采用等权的方式计算综合影响下的组合收益。以收益尽可能大，风险尽可能小的投资目标进行决策。当组合收益函数取得最大值时的投资组合策略即在以收益最大化的目标下的组合策略。

关键词：券商研报　投资策略　多因子分析　多元线性回归模型　VSM 模型

一、 问题的重述

1.1 问题背景

券商研报是指证券公司的研究人员对证券及相关产品的价值，或者影响其市场价格的因素进行分析，所作出的研究报告。而券商研报提供的信息，包括证券的综合分析、上市公司的总结、行业或宏观政策的观点以及相关股票评级等等。想要高回报率地进行投资决策，券商研报是重要的参考资料，可行的方法是从中提取有效的特征指标作为有效因子，根据历史数据完成深度学习，利用学习结果确定投资策略。

1.2 问题要求

我们需要综合研究券商研报的特征指标和外部环境对不同股票走势的影响，建立相关数学模型来完成以下问题：

（1）在湾区指数的 30 支股票之中选取 10 支股票的券商研报，并提取其中的特征指标。

（2）针对选择的 10 支湾区指数股票，建立模型并分析提取的特征指标对股票走势的影响，并提出明确投资策略。

（3）建立相关模型研究突发事件、舆情和自然灾害等因素对选取的 10 支股票行情的影响。

（4）结合以上问题，综合建模分析券商研报和外界环境因素对证券公司走势的影响，修改问题二的投资策略并在此基础上提出新的投资策略。

二、问题的分析

2.1 问题一的分析

针对问题一，要研究对研报的特征指标对股票走势的影响，首先研究选定的

10 支股票的研报，看到研报提供的影响因素数据较多，需要进行指标筛选，于是利用因子有效性检验，也就是 IC 值等进行初步的因子筛选；又考虑特征因子之间存在多重共线性，选用方差膨胀因子 VIF 来进行因子筛选，剔除部分对股票信息贡献不大、且与其他因子存在较强共线性的因子，确定得到一些有效性较高且具有一定代表性的子因子。最终以赋权法处理不同因子大类下的子因子，得到四个大类因子，分别是估值因子、盈利能力因子、成长因子、规模因子。

2.2 问题二的分析

在问题一的基础上，首先得到了具有一定有效性的特征指标因子，需要建立模型来分析特征指标对股票市价走势的影响，并给出投资策略。

结合题意，该题重点在于建立合适的模型来分析特征指标是如何影响股票价格的走势。

由于股票市场的股票价格关系为序列相关的，这意味着股票的历史信息可以用来预测未来的股价，且股票的价格之间存在一定的线性关系，因此，可以采用多元线性回归模型的方法，分别对所给的不同股票种类的数据进行建模、求解、分析。对于提出的投资策略，则需要对股票进行预测。由于多因子模型的基础理论认为股票的收益是由一些共同的因子来驱动的，因此选用多因子模型来进行预测，并给出投资策略。

2.3 问题三的分析

分析研报的内容，考虑突发事件的闪现、 舆情、自然灾害等的影响，研报文本可以反应出事件对股票行情的影响。通过对研报的文本挖掘，可以将单词分进行分类，在这里建立基本面类、情绪类、概念主题类三大热词库，其中将基本面类热词、情绪类热词分为正负面两面词库。采用向量空间 VSM 模型进行词频分析并建立词库。通过对大湾区指数股的研报为数据库，建立对应单词的文本向量，考虑使用 TF-IDF 计算词频权重。再分别计算在对应单词下，进行 VSM 分类时正确的概率，以及股票的超额。通过计算研报中单词的对应词频的收益指标

的均值来判断事件的闪现、 舆情、自然灾害等等事件对股票行情的影响。

2.4 问题四的分析

针对问题四，证券公司股票走势需综合考虑公司的内部财务特征以及外部的环境影响。在问题二策略模型中的收益函数上加入外界因素的影响因子，由于外部环境对股票行情的影响也是很重要的一部分，这里考虑采用等权的方式计算综合影响下的组合收益。以收益尽可能大，风险尽可能小的投资目标进行决策。当组合收益函数取得最大值时的投资组合策略即在以收益最大化的目标下的组合策略。

三、基本假设

- **假设 1** 多元线性回归模型中的随机误差具有零均值和等方差，既随机误差在不同的样本点之间是不相关的；
- **假设 2** 多元线性回归模型中的解释变量 x_1，x_2，…，x_p 是确定性变量，不是随机变量。

四、符号说明

符号	意义
IC	信息系数；
IR	信息比率；
VIF	方差膨胀因子；
x_p	第 p 个特征因子。
R_q	股票投资组合的收益

五、模型建立

5.1 基于有效性分析与去共线性的特征提取

考虑到研报提供的特征因子数量众多，因此对众多特征因子经过有效性检验筛选和共线性检验，筛选出有效的特征因子。

5.1.1 数据清洗

在对券商研报的数据进行特征提取的过程中，不同股票的券商研报拥有的特征因子数据不同，故提取相同的特征因子标准化后进行分析。

5.1.2 基于因子有效性检验的单因子选股模型

①回归法

股票下一期的收益率作为因变量，待测试的因子作为自变量，根据自己的需要也可加入行业虚拟变量进行行业中性。回归得到的系数也就是因子的收益率。构造回归方程之后，一般会参考系数 t 值的绝对值均值，系数 t 值绝对值序列大于 2 的占比，年化因子收益率，年化因子收益波动率，因子收益的夏普比等参数来看因子的有效性。注意这里如果是想做风格中性的多因子模型，这里需要把风格因子也放进方程中，也就是检验这个 alpha 因子在风格因子存在的情况下对股票的收益是否还具备解释力度。

②分层回归法

用待测试的因子对股票进行打分，根据分数将股票分为 N 组来进行分组回测，根据因子测试的周期来进行调仓的操作。指标的话，就是看做多第一组股票组合并做空最后一组股票组合的（or 基准指数）的收益率，夏普比率等。还有就是根据每一组的收益率可以很好的看出因子是否具有单调性。

③IC、IR 指标

IC 即信息系数，表示所选股票的因子值与股票下期收益率的截面相关系数，通过 IC 值可以判断因子值对下期收益率的预测能力。信息系数的绝对值越大，该因子越有效。IC 为负表示因子值越小越好，IC 为正表示因子值越大越好。IC 的计算方法是：计算全部股票在调仓周期期初排名和调仓周期期末收益排名的线性相关度。$IC\epsilon[-1,1]$，IC 的绝对值越大，预测能力越好。IC 最大值为 1，表示该因子选股 100%准确，对应的是排名分最高的股票，选出来的股票在下个调仓周期中，涨幅最大；相反，如果 IC 值为-1，则代表排名分最高的股票，在下个调仓周期中，跌幅最大，是一个完全反向的指标。IC 值的计算方法包括两种，Normal IC 和 Rank IC，分别对应 Pearson 相关系数和 Spearman 相关系数。

（1）Normal IC

Normal IC 是由 t 期因子载荷预测得到的 t+1 期收益预测值与收益实际值的相关系数

$$IC_A = Pearson(f_A, r)$$

其中，IC_A 为因子 A 在该期的 IC 值，f_A 为用 t+1 期收益率的预测值，r 为 t+1 期实际收益率。使用 Normal IC 的前提条件是收益服从正太分布，也可以直接使用当期因子值与下期实际收益率的 Pearson 相关系数。

（2）Rank IC：

Rank IC 采用了秩相关系数，适用于各种收益分布，公式如下

$$IC_{rank} = Pearson(f_{index}, r_{index})$$

IR 即信息比率（Information Ratio），是超额收益的均值与标准差之比，表示因子在在多个调仓周期中获得稳定 Alpha 的能力，可以根据 IC 近似计算，公式如下

$$IC\_IR = \frac{mean(IC)}{STD(IC)}$$

该公式是从超额收益出发，逐步推导得出的。IR= IC 的多周期均值/IC 的标准方差，代表因子获取稳定 Alpha 的能力。整个回测时段由多个调仓周期组成，每一个周期都会计算出一个不同的 IC 值，IR 等于多个调仓周期的 IC 均值除以这些 IC 的标准方差。所以 IR 兼顾了因子的选股能力（IC 代表）和因子选股能力的稳定性（IC 的标准方差的倒数代表）。
策略 IR 表示策略稳定战胜指数的能力

$$IR_{策略} = \frac{超额收益}{超额收益波动率}$$

现对特征因子进行单调性检验。有时因子在统计上并不表现出对于未来收益率很好的预测能力，但是可能由于该因子的复杂逻辑，其在策略中仍能获得超额收益。于是可以用更直接的方法检验该因子的选股能力。

Step1 每个调仓期，按照因子指标大小对股票池中所有股票分组，一般根据券池大小为 5 组或 10 组。

Step2 组内按等权重或者市值加权进行历史数据回测

Step3 多次调仓期后观察回测结果，包括累计收益，最大回撤，IR 值，胜率等指标和净值曲线的层次划分。如果优势组各指标越好，净值曲线层次划分越明显，说明单调性越强，该因子越有效。

5.1.3 基于去多重共线性的多因子选股模型

多因子选股模型目前已经成为 A 股量化权益投资的核心策略，因子共线性问题对策略的效果与稳健性具有基础性的重要影响，因子之间共线性问题的存在会使得投资组合的实际风格暴露偏离预期，容易在某些因子上产生过多的风险暴露。如果能合理解决因子共线性问题，将有利于提升多因子选股模型的稳健性，更好地控制投资组合的风险暴露。

为选用多个经典因子，对选定 10 支股票进行了相关实证分析，并探究能够合理解决因子共线性的实用性方法。

（1）多重共线性的检验[1]

①相关系数检验法

解释变量之间如果有较强的相关性，则表明可能存在较强的多重共线性。

②条件数法

条件数定义为正规方程组系数矩阵的最大与最小特征根之间比值的平根。通过正规方程组系数矩阵 X'X 的条件数来判断解释变量之间多重共线性的严重程度。为了消除解释变量之间量纲不同的问题，计算条件数之前首先要对故股票特征数据进行单位化处理，即需要将矩阵 X 中各列分别除以

$$\sqrt{X_j^T X_j} = \sqrt{\sum_{i=1}^{n} x_{ij}^2}$$

记经过单位化处理后的矩阵$X^{'}X$的最大和最小特征根分别为$\lambda_{\max}$和$\lambda_{\min}$，则其条件数定义为$\lambda = \left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{1/2}$。当条件数等于 1 的时候，表示不存在共线性问题。条件数越大说明共线性越强，通常设定判断阈值为 30。

③方差膨胀因子(VIF)检验法

自变量 X 的方差膨胀因子记为VIF($VIF = \frac{1}{1-R^2}$)，R^2是以 X 为被解释变量对其他解释变量进行回归所得到的模型拟合优度(也称为判定系数)。所以方差膨胀因子(VIF) 检验法与判定系数检验法是一-致的。一般判断是否存在多重共线性的VIF阈值为 10。

（2）剔除不重要变量

在前面的建模中，为了不漏掉重要的影响因素，考虑了过多的自变量。当涉及的自变量较多时，大多数回归方程都受到多重共线性的影响。于是做自变量的选元，剔除一些自变量。现以十支股票作为样本，每股价格为因变量，不同的经过第一题有效性筛选的特征因子作为自变量，建立多元线性回归方程如下

$$y = a_1 x_1 + a_2 x_2 + \ldots + a_n x_n + b$$

利用该多元线性回归方程的变量的方差扩大因子进行判断该因子与其他因子的多重共线性。当回归方程中的全部自变量都通过显著性检验后，若回归方程中仍然存在严重的多重共线性，有几个变量的方差扩大因子大于 10，即把方差扩大因子最大者所对应的自变量首先剔除，再重新建立回归方程，如果仍然存在严重的多重共线性，再继续剔除方差扩大因子最大、且由R^2变化可得的对因变量影响不大的自变量，直到回归方程中不再存在严重的多重共线性为止，在这个过程中，同时考虑特征因子对股票的经济意义，决定保留或剔除某自变量。[2]

5.1.4 基于赋权法构建特征指标模型

经过上文的特征因子筛选，得到了对应特征因子大类下的重要子因子，如下

表 1-1 筛选出的特征因子

因子大类	重要子因子
估值因子	市盈率、企业价值倍数、市现率
成长因子	净利润增长率、每股收益增长率、净资产增长率
盈利能力因子	毛利率、净资产收益率
规模因子	流通市值、流通股本

对每个重要子因子标准化后予等权法处理，也就是得到每个因子大类的Y_i的取值方程如下

$$Y_i = \frac{1}{k}\sum_{i=1}^{k} z_k$$

k 为当前因子大类的所拥有的子因子数量。

5.2 问题二的模型建立与求解

5.2.1 基于分析研报特征指标对股票走势影响的模型构建

（1）多元线性回归模型的构建

定义：设随机变量 y 与一般变量 x_1，x_2，…，x_p 的线性回归模型为

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \tag{2.1}$$

其中，β_0，β_1，…，β_p 是 $p+1$ 个未知参数，称为回归系数，β_0 是回归常数，ε 是随机误差。

（2）对模型的基本假设

为了方便进行模型的参数估计，对回归方程进行一些假定，如下：

①解释变量 x_1，x_2，…，x_p 是确定性变量，不是随机变量

②随机误差项具有零均值和等方差

③正态分布的假定为：

$$\begin{cases} \varepsilon_i \sim N(0,\sigma^2),\ i = 1,2,3..,n \\ \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n \text{相互独立} \end{cases}$$

5.2.2 基于分析研报特征指标对股票走势影响的模型求解

基于问题一得出的结论，可以得知如下需要分析的特征指标，分别为：估值因子、成长因子、盈利能力因子和规模因子这四类因子大类。这些特征指标，既为所建模型的回归系数 β_1，…，β_p。由于不同的股票具有各自的代表性，故我们选取已国泰君安为代表购买的三支股票：粤水电、顺络电子、丽珠集团为代表，进行数据的代入和分析。可以得到以下三个结果：

表 2-1 粤水电的回归系数表

	Coefficients		Coefficients
Intercept	8.56E-02	X Variable 5	1.88E-05

X Variable 1	1.39E-04	X Variable 6	3.60E-06
X Variable 2	-2.26E-03	X Variable 7	7.83E-02
X Variable 3	5.65E-02	X Variable 8	-8.02E-06
X Variable 4	-2.87E-02	X Variable 9	-4.92E-06

表 2-2 顺络电子的回归系数表

	Coefficients		**Coefficients**
Intercept	-1.80E+00	X Variable 5	5.64E+00
X Variable 1	3.15E-03	X Variable 6	9.77E-02
X Variable 2	7.18E-05	X Variable 7	1.17E+00
X Variable 3	-2.09E-04	X Variable 8	1.12E-01
X Variable 4	-1.96E+00	X Variable 9	-2.60E-03

表 2-3 丽珠集团的回归系数表

	Coefficients		**Coefficients**
Intercept	4.88E-01	X Variable 5	3.42E-01
X Variable 1	2.32E-04	X Variable 6	4.64E-02
X Variable 2	-3.26E-04	X Variable 7	6.95E-05
X Variable 3	-2.30E-04	X Variable 8	3.72E+03
X Variable 4	5.08E-04	X Variable 9	1.32E-01

从以上三个表（表中展示出前十个回归系数）可以得到三支股票的多元回归模型。不难看出，每支股票都有一半以上的回归系数 β_p 为正数，既该特征指标对于股票的影响力呈现出正相关的关系，因此可以从整体模型的角度来看，每支股票的利润的走势都呈现出上涨的趋势。

我们得到了多元线性回归方程的公式，就可以利用统计软件对数据进行未来的预测。

5.2.3 投资策略

（1）模型构建

由题目可知，这是由 N（N=10）只股票组成的资产组合，记第i只股票在该组合中的权重为w_i，收益为r_i。那么该组合的收益率R_q可表示为

$$R_q = \sum_{i=1}^{N} w_i r_i, \quad N = 10$$

记市场上驱动股票的因子的数量为 k，u_i表示股票的特质收益，则r_i为

$$r_i = \sum_{k=1}^{K} X_{ik} f_k + u_i$$

整个投资组合在风险组合在第k个风险因子上的暴露程度可以表示为：

$$X_k^q = \sum_{i=1}^{N} w_i X_{ik}$$

那么，

$$R_q = \sum_{k=1}^{K} X_{ik} f_k + \sum_{i=1}^{N} w_i, \quad N = 10$$

多因子模型是风险与收益关系的定量表达，因子则是可能影响或解释股票期望回报率的解释变量[3]。表达式如下:

$$\widetilde{y_j} = \sum_{k=1}^{K} X_{jk} * \widetilde{f_k} + \widetilde{u_j} \quad (2.2)$$

其中，$\widetilde{y_j}$ 为股票的收益率，X_{jk} 为股票 j 在因子 k 上的因子暴露（因子载荷），$\widetilde{f_k}$ 为因子 k 的因子收益，$\widetilde{u_j}$ 为股票 j 的残差收益率。

（2）模型预测

通过模型建立与分析，可得出粤水电、顺络电子以及丽珠集团三支股票的未来走势依旧是良好的，有上涨的趋势，可以进行投资。

5.3 基于文本挖掘分析外部环境对股票行情的影响

股票的走势与行情不只是受到公司内部财务特征的影响，也会受到外部环境的影响，如突发事件的闪现、 舆情、自然灾害等的影响。分析师的分析会结合突发事件的闪现、 舆情、自然灾害对股票的可能影响进行分析。从研报文本可以获取分析师面对该股票的情况的投资情绪倾向，借助研报的热词来分析外部定性影响对股票行情的影响。

5.3.1 热词库构建的可行性分析

分析研报的内容，考虑突发事件的闪现、 舆情、自然灾害等的影响，这里建立基本面类、情绪类、概念主题类三大热词库，其中将基本面类热词、情绪类热词分为正负面两面词库。参考文献[5]，在近年来，各大基金公司和券商纷纷进行量化投资的研究之中，研究研报的分析师数量基数庞大，研报获取和稳定性较为可观。在 2011-2014 年的热词库研究中，“基本面+情绪面的热词库选股策略的表现是稳定的，同样的，“概念主题”热词库选股策略也是比较符合理想状况的。

5.3.2 基于基于 VSM 模型热词库的构建

向量空间 VSM 模型的基本思想是：给定一个文本，用一个向量表示该文本的语义，向量的每一维度分别对应一个单词，其数值是该单词在该文本中出现的

词频或者是 TF-IDF。则每个文本相当于一个向量，特征数为单词的总数，通过计算向量之间的余弦值来比较文本的近似层度。

Step1:向量空间模型的建立。以大湾区指数股相关的研报文本为基础，得到 n 个关键词特征，构建空间向量$(t_1,t_2,...,t_n)$

Step2：文本特征项权重的建立。每个文本的词的贡献度不同，则每个词的权重不同。这里考虑采用 TF-IDF 来确定权重。

在一个文本中，一个词出现的次数越多，那么这个词的贡献度越大，通过下面公式计算词的权重。其中tf表示词频，$length$表示文本长度，那么

$$TF_w = \frac{\text{在某一词条}w\text{出现的次数}}{\text{该类中所有的词条数目}}$$

$$T = \frac{tf}{length}$$

若一个词在多个文档中出现过，那这个词对某一个文档的贡献度就越小，这里定义逆文档频率来度量这个概念。记df为文档频率，N为文档总数，则逆文档频率计算公式如下

$$IDF = \log\left(\frac{df}{N+1}\right)$$

那么，

$$TF - IDF = TF \times IDF$$

通过对湾区指数的文本挖掘与分析，构建如下热词库：

表 3-1 基本面类词库

词库类型	正反类型	词库
基本面词库	正面	收购、并购、转型、超预期、成长、定增、改革、拐点、增长、提升、价值、低估
	反面	业绩下滑、业绩同比下降

表 3-2 情绪面类词库

词库类型	正反类型	词库
情绪面类词库	正面	龙头、巨大、翻倍、强烈、重大
	负面	不利于、不及预期、不达预期

表 3-3 概念主题类热词库

词库类型	年份	词库
概念主题类热词库	2020	新冠疫情、20cm、新能源车、赛道、白酒、碳中和、基金、需求侧改革

Step3：分析词频胜率和超额。计算每个热词对应的时期的机器学习分析结果（即胜率p）的正确概率以及股票价格的涨跌程度（即超额Q_i，其中$i=1,2,...,n$）。那么，每个热词对应反应出来的收益期望为：

$$E(Q^{(j)})=pQ_i^{(j)}$$

Step4：通过计算文本中含有热词库的词的影响均值，来计算事件对股票行情的影响情况。

$$E(Q)=\sum E(Q^{(j)})$$

若$E(Q)$大于 0，则影响为正向的，$E(Q)$越大，影响越大；若$E(Q)$小于 0，则影响为正向的，$E(Q)$越小，影响越小。

5.4 基于问题二改进的规划模型的策略研究

在问题二的基础上，需要加上外部环境对股票行情的影响，进而需修改投资策略模型进行模型优化。由题目可知，这是由 N（N=10）只股票组成的资产组合，记第i只股票在该组合中的权重为w_i，收益为r_i。那么该组合的收益率R_q依

旧可表示为

$$R_q = \sum_{i=1}^{N} w_i r_i，\ N = 10$$

记市场上驱动股票的因子的数量为 k，则u_i表示股票的特质收益。在这里股票的收益还需考虑外部环境的影响，根据问题三采用$E(Q^{(i)})$来表示，这里考虑采用等权计算收益，那么r_i为

$$r_i = 0.5 \times \left(\sum_{k=1}^{K} X_{ik} f_k + u_i \right) + 0.5 \times E(Q^{(i)})$$

那么，

$$R_q = \frac{1}{2} \sum_{k=1}^{K} X_{ik} f_k + \frac{1}{2} \sum_{i=1}^{N} w_i + \frac{1}{2} E(Q^{(i)})，\ N = 10$$

在综合考虑收益尽可能大，风险尽可能小的投资目标进行决策。那么我们的投资目标是：

$$\max\ R_q$$

当R_q取得最大值是的投资组合策略即在以收益最大化的目标下的组合策略。

六、模型的评价

6.1 模型的优点：

（1）多元线性回归模型可以准确地计算各个因素之间的相关程度与回归拟合程度的高低，从而提高了预测的效果；

（2）多元线性回归模型考虑了多个因素，将多个变量考虑进模型，相较于一元回归模型等普通的回归模型，更加符合适用于实际经济问题。

（3）多因子模型可实现数据的降维，其因子的统计量通常比基于资产的统计量具有更好的稳定性，且相较于均值-方差模型这种传统模型，多因子模型具有更加丰富的信息源和解释变量[4]。

6.2 模型的不足与改进：

（1）多元回归模型中，因素之间大多也有一些联系，当某一个自变量变动时，往往也会导致另一个变量的波动，这时就会在一定程度上影响到结果。

（2）在数据的获取和处理上存在一些不足，可考虑从多个方面获取数据进行组合。报表的数据具有延时性，可考虑用前提数据进行替代。

七、参考文献

[1]马池坤. 针对多因子选股模型因子共线性问题的 A 股市场实证分析[D]. 山东大学, 2018.

[2] 何晓群.应用回归分析[M]. 北京:电子工业出版社, 2017. 156-161.

[3]吴雁南,赵子铱.多因子模型资产定价应用评述[J].企业科技与发展,2021(08):64-66.

[4]国金基金. 一篇文章告诉你市场上热议的多因子模型怎么构建，如何区分优劣. [2018-9-13]. https://guba.eastmoney.com/news,jjdp,782959993.html

[5]张成伟,郑诚.基于改进 VSM 的文本信息检索研究[J].计算机技术与发展,2009,19(1):71-73.

[6]徐陆彤,徐松涛.白酒行情分析——基于酒企研报[J].上海商业,2021(6):81-83.

[7]杨戈,杨麓涛.基于爬虫和 TFIDF-NB 算法的微博情感分析[J].电子技术应用,2021,47(4):59-62+66.

附录

1. 代码

(R 语言)

```
library(readr)

x<- read_csv("/Users/candyning/Desktop/粤水电.csv")

y=read_csv("/Users/candyning/Desktop/顺络电子.csv")

z=read_csv("/Users/candyning/Desktop/丽珠集团.csv")

x1=ts(x)

x2=ts(y)

x3=ts(z)

xx1<-lm(y~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10,data=x1)

xx2<-lm(y~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10,data=x1)

xx3<-lm(y~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10,data=x1)

summary(xx1)

plot(xx1)

dfp1<-predict(xx1,interval="prediction")

dfp2<-predict(xx2,interval="prediction")

dfp3<-predict(xx3,interval="prediction")
```