# 基于自然语言处理技术的股价计量研究

摘要：本文先通过数据爬取从东方股吧爬取了四家来自不同行业的公司舆情数据，然后通过Bi-LSTM（双向长短期记忆网络）对文本数据进行分析得到每日情绪指标。得到情绪指标后我们结合了从Tushare接口获取的其各自的日线数据，并且通过日线数据构造了移动平均、KDJ、MACD等指标，然后针对二者进行融合做出计量探究，经过选择和节选后得到了七个指标，并且由决策树模型给出了建议。

关键词：金融计量，数据挖掘，深度学习，决策树，数据爬取，共线性研究。

## 1.引言

### 1.1研究背景与意义

根据第46次《中国互联网络发展状况统计报告》显示，中国网民规模已达9.40亿，而中国股民数量也超过1.7亿，互联网由原始的信息发布平台逐渐演变成网络媒体及用户传播、获取信息的主要载体，各种网络舆论再此发生、发酵、爆发。各式各样的信息舆论充斥着人们的生活，深刻影响着人们的情绪与态度。

我国股市自改革开放以来发展至今，已经建立起了规模庞大的证券市场。而较西方国家而言，我国的证券市场尚处于起步阶段，还不够成熟，且投资者多为中小投资者，没有形成合理的价值投资取向，在股票操作中的盲目性、投机性问题依然广泛存在。这部分股民在进行投资决策时往往出于主观情绪判断——在乐观情况下会加仓跟进，在悲观情绪下可能会大规模抛售特定股票甚至退出市场。因此，股民情绪对股票市场可以产生重大影响。如何将股民情绪这一非量化因素加入对股票市场的量化研究中，使得股票研究更加全面化一直以来是股票研究的热门课题。

随着深度学习技术的演进，使得将大规模文本数据量化已经成为可能。我国股票市场形成时间较短，容易受到诸多因素干扰而产生剧烈波动。本文在通过金融数据预测股票趋势的同时，处理文本数据、构建情绪指标，将情绪指数与金融数据相结合来预判股票走势和市场整体情况，并基于此对投资者决策提出建议。1.2研究现状

一直以来，对于股价的分析与预测是学术界的热门话题，经济特征、技术指标、财务审计公告、相关新闻、网络舆情等信息源被不断尝试用于研究股票价格及趋势预测。而随着行为金融学理论的发展、深度学习技术的进步，同时由于中国证券市场的不成熟，多数投资者存在较为严重的投机性等理论、技术、现实方面的因素，越来越多的研究通过处理新闻、网络舆情的文本数据来量化投资者情绪、市场行业情绪，结合股票交易数据、企业财务数据等指标，构建股价模型来证明网络舆情、投资者情绪等对股价波动的影响，同时进行股价预测。

1.2.1 文本数据的获取与处理

对于情绪数据的获取与处理，较为常见的方法是通过数据挖掘或网络爬虫获取新闻、微博、股吧等平台的相关数据，再通过自然语言处理获取的数据来进行进一步的研究。孟雪兰等通过人为设定关键词，以百度对相关词条的搜索量作为情绪变动指标，进而通过因子分析法构建投资者情绪指数来预测沪市变动。[[1]](#footnote-1)刘少伟通过数据发掘，获取贴吧发言信息从而构建投资者舆情指数，并结合DHS模型分析其与股价、收益率的关系。[[2]](#footnote-2)孟勇等爬取了2010年1月到2017年8月新浪财经股票板块下的所有文章评论，对共11万条文本数据进行处理，构建文本情绪成分来进行情绪指标的构建。牛雪琪通过网络爬虫技术获取东方财富网股吧评论文本及证券机构对股票行业的研究报告，并对获取的数据进行分词处理与筛选。[[3]](#footnote-3)

1.2.2 情绪指数的构建

除了分析传统的金融指标外，大多中国的学者在文本数据挖掘的基础上采取构建情感词典法，来度量网络舆情及情绪指标的构建。首先确认关键词，其次再对舆情进行层次分析与综合评价测度。孟勇等筛选出带有金融情感的词语，再进行分类，建立积极、消极的金融情绪词典，将BW 方法与文本挖掘技术相结合，构建新的情绪指数。[[4]](#footnote-4)王夫乐等将爬取的新浪微博内容进行情绪分类定性，分为喜悦、悲伤、惊奇、恐惧、愤怒五个维度，以情绪性质的加权平均值作为当日微博情绪。[[5]](#footnote-5)

除此之外，借鉴BW指数的构建方法，选择直接、间接情绪指标中具有代表性的变量来构建情绪综合指数的学者也有很多。李响等选择了月度振幅、交易量、市盈率、消费者信心指数、宏观经济景气指数等九个变量综合构建情绪指数。[[6]](#footnote-6)闵峰等选取换手率、封闭式基金折价率、新增开户数和消费者信心指数四个投资者情绪代理指标来构建投资者情绪指数，与宏观经济变量一起探究其对股指收益率的影响。[[7]](#footnote-7)

## 2.数据准备

### 2.1数据来源

#### 2.1.1非结构化数据

为了能够直接获得与股票直接相关的文本数据，且在格式上得到一定保证，本文主要考虑发布在各大股票网站的股评数据。对比各大股票网站，东方财富股吧网是一个股民流量较大的评论贴吧，单只股票的日评论量顶峰可以达到700条以上。因此，本文情绪分析的主要数据为东方股吧的股民评论数据。本文从东方财富股吧爬取了ST科迪、比亚迪、复兴医药、老白干等四个主要吧及其所属行业贴吧的数十万条评论数据，来作为原始文本数据。

我们调用Python的Request和Beautifulsoup4模块，来作为主要爬取和解析工具，共爬取数据四十余万条，在爬取中插入自定义分隔符作为分列工具，并通过去重、删除空行后还原的股评数据。

在经过处理后，数据构成的主要数据字段如下：

将其保存到文件中以待进一步的处理。

|  |  |
| --- | --- |
| 字段名 | 含义 |
| Click\_Num | 该条新闻/帖子的总被点击数 |
| Talk\_Num | 该条新闻/帖子的总评论数 |
| Text | 该条新闻/帖子的主要内容 |
| Time | 该条新闻/帖子的发布时间 |
| XuHao | 记录本条新闻在该次爬取中的顺序 |

#### 2.1.2结构化数据

在金融研究中，行情数据等结构化数据也起着重要的作用。行情数据包含着单只股票的历史交易额、最高额、交易量等众多信息。通过行情信息，我们可以对单只股票的历史信息得到充分了解，进而评估其收益率与风险情况。

Tushare是一个免费、开源的python财经数据接口包。主要实现对股票等金融数据从数据采集、清洗加工 到 数据存储的过程，能够为金融分析人员提供快速、整洁、和多样的便于分析的数据，为他们在数据获取方面极大地减轻工作量，使他们更加专注于策略和模型的研究与实现上。考虑到Python pandas包在金融量化分析中体现出的优势，Tushare返回的绝大部分的数据格式都是pandas.DataFrame类型，非常便于用pandas/NumPy/Matplotlib进行数据分析和可视化。

在这里，我们通过对Tushare接口对四只股票的在指定期间内的行情数据进行提取。得到的数据字段如下：

|  |  |
| --- | --- |
| 字段名 | 含义 |
| ts\_code | 该股票的股票代码 |
| trade\_date | 交易日期 |
| open | 该股票的该交易日期的开盘价格 |
| high | 该股票的该交易日期的最高价格 |
| low | 该股票的该交易日期的最低价格 |
| close | 该股票的该交易日期的收盘价格 |
| pre\_close | 该股票的该交易日期的前一期的收盘价格 |
| change | 该股票相邻两交易日期的收盘价格变化 |
| pct\_chg | 该股票相邻两交易日期的收盘价格变化率 |
| vol | 该股票该交易日期的交易量 |
| amount | 该股票该交易日期的成交额 |

保存到文件中以便进行下一步研究。

#### 2.1.3数据时间窗口说明

本文中的数据，因为文本数据取自东方股吧的股评数据，故而不同行业之间的长度各存在细微区别。为了与不同行业的文本数据相匹配，故而在提取行情数据时按照爬取到的新闻数据区间来进行提取，以便二者进行匹配。

总体而言，本文所研究的来自四个不同行业的四只股票的数据，大致时间区间取自2015年至2020年十二月。

### 2.2数据处理

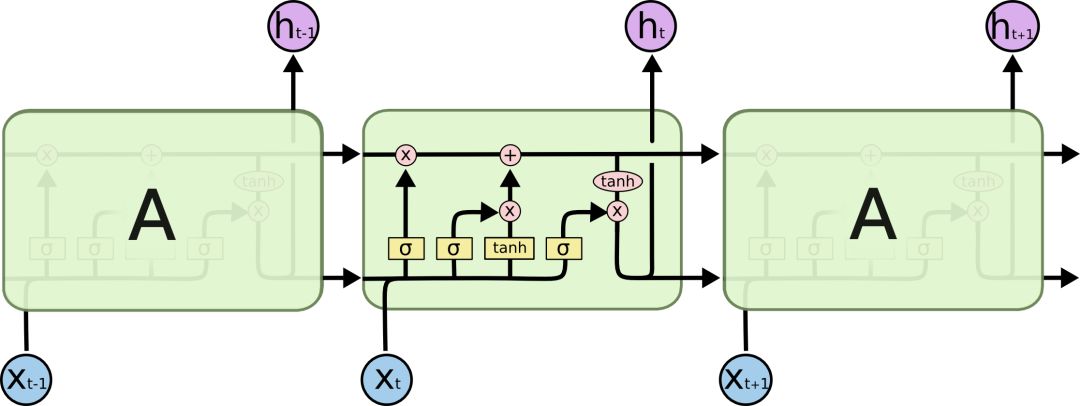
#### 2.2.1非结构化数据

计算机无法直接处理文本数据，在使用文本数据展开计量研究以前，我们必须对文本数据开展量化。

量化文本分析（Quantitative Text Analysis, QTA）是指在社会科学研究中利用计算机技术来自动并且系统地处理大量文本数据的方法。不同的研究往往会使用不同的文本分析模型，但是并不存在一种能够普遍适用于分析各类文本数据的最好模型。在一种文本数据分析中表现良好的模型，并不一定会在分析其他数据中表现良好。

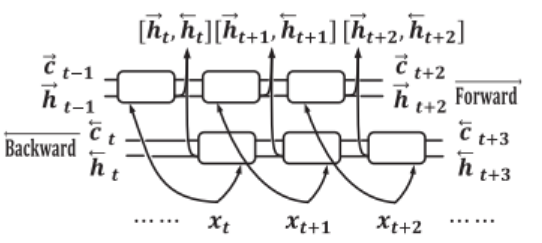
考虑到模型的有效性，以及模型使用的便捷性，我们在此采用双向长短期记忆神经网络模型作为文本的主要量化手段。该模型的介绍如下：

循环神经网络是将序列数据作为输入，在序列的演进方向上进行递归并且所有节点按照链式连接的神经网络。循环神经网络最大的特点是，前面的信息会保存下来并且应用于当前的计算中，也就是说隐藏层的节点变成了有连接的，而且隐藏层的输入不仅包括了输入层的输出，还包括了前一个隐藏层的输出。[[8]](#footnote-8)



循环神经网络虽然可以保存历史信息，但是对于相对复杂的文本处理来说，当前时刻有时候需要更久以前的历史信息，有时又仅仅需要临近的历史信息。长短期记忆（Long Short Term Memory, LSTM）网络则解决了这一问题，使神经网络依然可以从输入数据中学习到长期依赖关系。

前向的LSTM与后向的LSTM结合成BiLSTM。通过增加从后往前传递信息的结构，让网络可以获取到当前时刻之后的内容。



我们将当前时间步记为t，上一时间步记为t-1，时间步t的输入记为，从前往后计算得到的隐藏层的状态记为，从后往前计算得到的隐藏层的状态记为。故而可以得到以下公式：







其中，f代表激活函数，常见的有Relu，tanh等。W是四个权重矩阵，每个权重矩阵都会通过网络的反向传播进行更新。g(x)表示将两部分拼接起来，是网络最终的输出。

在此我们方便起见，使用百度提供的Senta预训练模型。该模型构建了一个情感倾向分析（Sentiment Classification，简称Senta）模型，针对带有主观描述的中文文本，可自动判断该文本的情感极性类别并给出相应的置信度。该模型基于一个双向LSTM结构，情感类型分为积极、消极。输入为可信的文本数组，输出为置信度的数值矩阵，在此我们使用它作为文本处理的核心模型。在得到每日的新闻置信度数组后，我们通过如下公式来构建每日的情趣指标：



其中，W为每日每条新闻的阅读量，X为每日每条新闻的乐观趋向置信度，U为随机误差项。

以此得到每日的情绪指标Score。

#### 2.2.2结构化数据

在获取到结构化数据后，除其本身具有的行情数据以外，我们还借由基本行情数据，生成了一批衍生指标。

（1）移动平均线（Moving Average, MA）

移动平均线是技术分析中的应用的十分普遍的一项指标，主要用于判断行情的趋势。“平均”是指单位周期内的平均收盘价格。“移动是指”将新的交易日收盘价纳入计算周期的同时，删除最早的交易日收盘价。

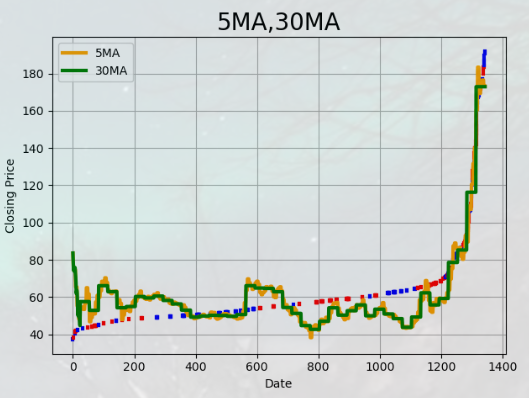
可以由如下的公式概括：







移动平均线体现了在该段周期内持股人的平均持股成本。在股价上涨时移动平均线会随着股价保持上行，反映了该周期内持股人的平均持股成本上升了，反之亦然。一般来说，在上涨过程中，由于主力筹码较多，移动平均线可以视为主力的成本线，在上涨末期时由于主力把筹码派发给了散户，移动平均线可以视为散户的成本线。



例：比亚迪的5日均线与30日均线

(2) 震荡类指标KDJ

KDJ指标最早的雏形是由芝加哥期货交易商George Lane的KD指标，该指标又被称为随机震荡指标（Stochastic Oscillator Indicator）。先使用特定时间跨度中的最后收盘价与该时间跨度内的最高价和最低价的相对位置来推测市场的超买和超卖情况，同时融合了移动平均思想，用更多的信息来捕捉市场中的超买、超卖现象。在KDJ中，又在随机震荡指标的K线和D线的基础上又增加了一条J线，进一步提高了随机震荡指标对市场买卖信号捕捉的准确度。[[9]](#footnote-9)

关于KDJ指标的计算，分为以下几步。

1. 计算未成熟的随机值RSV，计算公式如下所示：

RSV=\*100%

其中，n为时间跨度，RSV取值范围在0-100之间，取值越大越说明收盘价在价格区间的相对位置越高，市场越可能出现超买的现象，反之亦然。

1. 计算得到RSV值后，K值由RSV值的指数移动平均计算得到，即前一日的K值和当期

RSV值经过一定权重相加所得，计算公式如下：

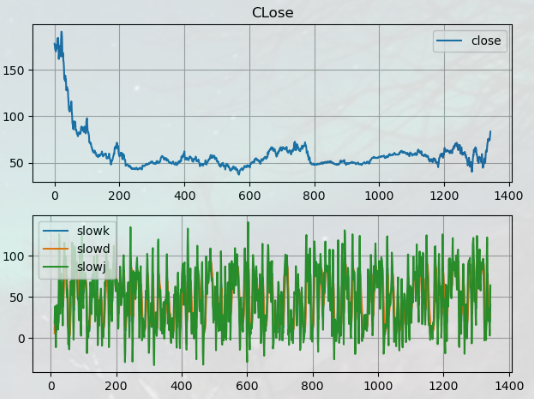


1. D值由K值的指数移动平均计算得到，即前一日的D值和K值经过一定权重调整后相加得到，计算公式如下：



1. J值是KD的辅助指标，进一步反应了K指标和D指标的乘离程度，计算公式如下：





例：比亚迪的收盘价与KDJ值曲线

（3）行情数据

Tushare接口提供了广泛的行情数据，具体字段及其含义已经在上文展示。

二、计量经济实证研究

（一）变量选取

1.被解释变量

比亚迪股份每日收盘价（Y）

2.解释变量

综上，解释变量影响力方面及其与其变动方向整理如下：

|  |  |  |
| --- | --- | --- |
| 代表方面 | 解释变量 | 与Y的预期关系 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

（二）模型假设

（三）样本来源和分析

本文研究对象为，考察样本期为，一共笔数据。其中涉及变量包括：

（四）初步模型估计

1.模型构建

2.样本选择

3.平稳性检验

为了检验所有时间序列变量是否平稳并确定其单整阶数，利用Eviews对所有变量依次进行单位根检验。

1. 对序列Y的平稳性检验

|  |  |  |  |
| --- | --- | --- | --- |
|  | 含趋势项 | 仅含截距项 | None |
| AIC | 3.432065 | **3.430325** | 3.434072 |
| SC | 3.448859 | 3.441521 | **3.439670** |
| HQ | 3.438499 | 3.434614 | **3.434072** |
| P值 | 0.4684 | 0.1988 | 0.7644 |

可以得知，Y在没有截距项、没有趋势项的模式上，不拒绝原假设，是非平稳序列。

|  |  |  |  |
| --- | --- | --- | --- |
|  | 含趋势项 | 仅含截距项 | None |
| AIC | 3.437244 | 3.435529 | **3.433529** |
| SC | 3.454055 | 3.446736 | **3.439133** |
| HQ | 3.443685 | 3.439823 | **3.435676** |
| P值 | 0.0000 | 0.0000 | 0.0000 |

可以得知，Y在含截距项、趋势项的模式上，拒绝原假设，是平稳序列。

（2）平稳性检验的最终结果

按照以上步骤依次对所有变量及其差分进行单位根检验。整理如下：

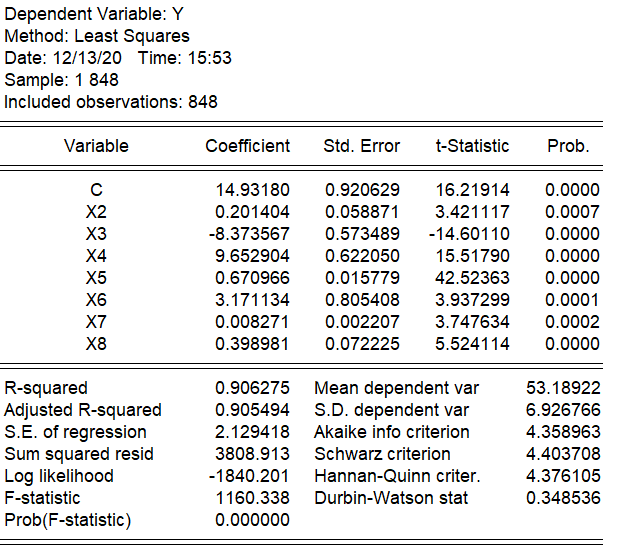
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 变量 | Y | X2 | X3 | X4 | X5 | X6 | X7 | X8 |
| 单整阶数 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |

各个变量单整阶数

从结果可知，最高阶变量不止一个，该模型满足协整的前提条件，可以进行进一步检验。

4.模型初步估计

使用EG两步法对模型进行协整检验，第一步要对模型进行回归估计，结果如下：

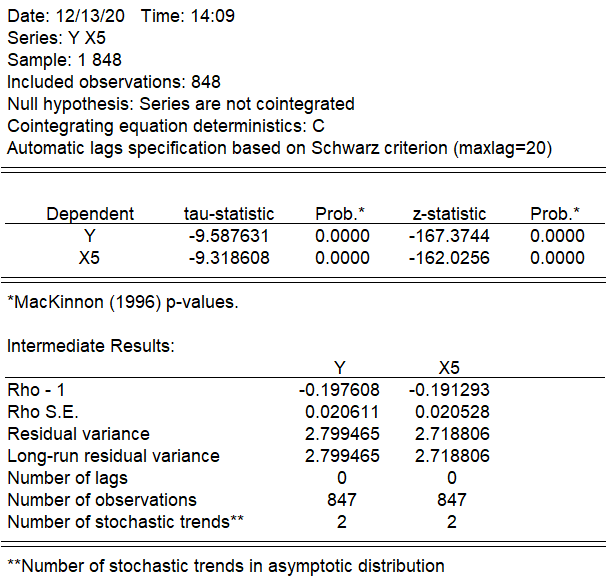


初步回归结果

估计的回归模型为：

5.协整检验

对回归到残差进行ADF检验，结果如下：



Y、X5的协整检验

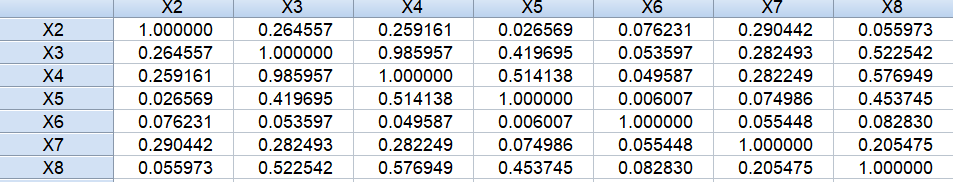
可知，p值小于0.05，说明模型变量间存在协整关系。

（五）多重共线性的检验及补救

1.多重共线性检验

（1）简单相关系数法

通过Eviews计算得到变量间相关系数如下：



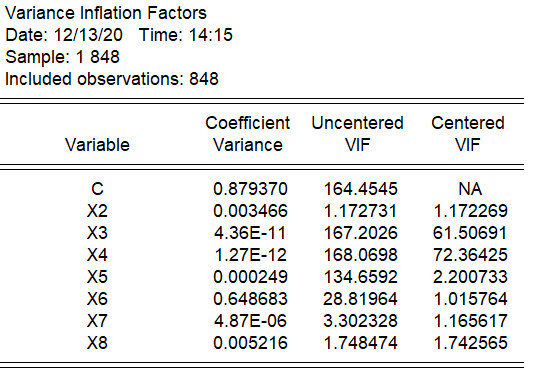
变量相关系数矩阵

由相关系数可知，X3与X4之间相关系数大于0.9，表明变量间高度相关，该模型存在多重共线性。

（2）方差膨胀因子法

为进一步分析变量间多重共线性，采用辅助回归，计算每个变量的方差膨胀因子VIF。

软件计算结果如下：



各个变量VIF

根据经验，VIF大于10表明该解释变量与其他变量存在严重的多重共线性，从表中看出，模X3、X4存在多重共线性。

2.多重共线性补救

为了避免经济含义的改变，本文采用逐步回归法对多重共线性进行补救。

（1）分别对X2、X3、X4、X5、X6、X7、X8进行一元回归，结果如下

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 变量 | X2 | X3 | X4 | X5 | X6 | X7 | X8 |
| 参数估计 | 0.608010 | 43.529236   |  | | --- | |  | | 4.244351 | 0.921386 | 4.68790 | 0.03369 | 2.93305 |
| t统计量 | 3.045978 | 17.23644 | 22.57333 | 65.98643 | 1.80579 | 5.14202 | 19.9826 |
| P值 | 0.0006 | 0.0000 | 0.0000 | 0.0000 | 0.0713 | 0.0000 | 0.0000 |
|  | 0.013952 | 0.259904 | 0.375901 | 0.837314 | 0.00384 | 0.03031 | 0.32066 |

逐步回归的一元回归结果

加入X5的方程最大，因此将X5加入模型。

（2）以X5为基础，顺次加入其他变量，进行回归，结果如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | X2 | X3 | X4 | X5 |  |
| X2  X5 | 0.483206  t=6.953743  p=0.0000 |  |  | 0.918875  t=67.59938  p=0.0000 | 0.846120 |
| X3  X5 |  | 1.056789  t=10.63316  p=0.0000 |  | 0.856874  t=59.27445  p=0.0000 | 0.856513 |
| X4  X5 |  |  | 1.342327  t=13.15658  p=0.0000 | 0.821003  t=55.32305  p=0.0000 | 0.864654 |
|  | X5 | X6 | X7 | X8 |  |
| X5  X6 | 0.921045  t=66.57750  p=0.0000 | 4.272181  t=4.110199  p=0.0000 |  |  | 0.840503 |
| X5  X7 | 0.913378  t=67.55418  p=0.0000 |  | 0.020527  t=7.899143  p=0.0000 |  | 0.848501 |
| X5  X8 | 0,834476  t=58.65614  p=0.0000 |  |  | 0.985296  t=13.46349  p=0.0000 | 0.866049 |

逐步回归的二元回归结果

加入X8时，模型在原来基础上上升最高，两变量p值均小于0.05，表明两变对被解释变量影响都显著。

（3）以X5、X8为基础，顺次加入其他变量，进行回归，结果如下：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | X2 | X3 | X4 | X5 | X8 |  |
| X2  X5  X8 | 0.441245  t=6.987307  p=0.0000 |  |  | 0.834349  t=60.28386  p=0.0000 | 0.960740  t=13.47792  p=0.0000 | 0.873374 |
| X3  X5  X8 |  | 0.632598  t-=0.1357  p=0.0000 |  | 0.813319  t=56.68413  p=0.0000 | 0.787345  t=10.02038  p=0.0000 | 0.871313 |
| X4  X5  X8 |  |  | 0.911423  t=8.329001  p=0.0000 | 0.792349  t=54.31291  p=0.0000 | 0.690175  t=8.757967  p=0.0000 | 0.875783 |
|  | X5 | X6 | X7 | X8 |  |  |
| X5  X6  X8 | 0.836141  t=59.08037  p=0.0000 | 3.144774  t=3.307373  p=0.0010 |  | 0.963577  t=13.19031  p=0.0000 |  | 0.867293 |
| X5  X7  X8 | 0.836237  t=59.98687  p=0.0000 |  | 0.014728  t=6.036857  p=0.0000 | 0.900190  t=12.31998  p=0.0000 |  | 0.871137 |

逐步回归的三元回归结果

加入X2时，模型上升最高，三个变量p值都小于0.05，表明三个变量对被解释变量影响显著。

（4）以X2、X5、X8为基础，顺次加入其他变量，进行回归，结果如下：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | X2 | X3 | X4 | X5 | X8 |  |
| X2  X3  X5  X8 | 0.359418  t=5.51895  p=0.0000 | 0.466735  t=4.414957  p=0.0000 |  | 0.818763  t=57.90815  p=0.0000 | 0.819245  t=10.57695  p=0.0000 | 0.876235 |
| X2  X4  X5  X8 | 0.314872  t=4.881751  p=0.0000 |  | 0.748014  t=4.881751  p=0.0000 | 0.799811  t=55.25039  p=0.0000 | 0.725565  t=9.290457  p=0.0000 | 0.879626 |
|  | X2 | X5 | X6 | X7 | X8 |  |
| X2  X5  X6  X8 | 0.428089  t=6.790907  p=0.0000 | 0.835777  t=60.61204  p=0.0000 | 2.690340  t=2.896536  p=0.0039 |  | 0.942892  t=13.23553  p=0.0000 | 0.874622 |
| X2  X5  X7  X8 | 0.361062  t=5.536136  p=0.0000 | 0.835660  t=60.98754  p=0.0000 |  | 0.010772  t=4.305324  p=0.0000 | 0.902954  t=12.57269  p=0.0000 | 0.876098 |

逐步回归的四元结果

加入X4时模型上升最高，四个变量p值均小于0.05，表明四个变量对被解释变量影响显著。

（5）以X2、X4、X5、X8为基础，顺次加入其他变量，进行回归，结果如下：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | X2 | X3 | X4 | X5 | X8 |  |
| X2  X3  X4  X5  X8 | 0.269507  t=4.642762  p=0.0000 | -8.276932  t=-14.211  p=0.0000 | 9.605712  t=15.20966  p=0.0000 | 0.667228  t=41.67961  p=0.0000 | 0.449929  t=6.179349  p=0.0000 | 0.902910 |
|  | X2 | X4 | X5 | X6 | X8 |  |
| X2  X4  X5  X6  X8 | 0.301186  t=4.680136  p=0.0000 | 0.750086  t=6.6667  p=0.0000 | 0.801163  t=55.58091  p=0.0000 | 2.727265  t=3.010948  p=0.0000 | 0.706820  t=9.063611  p=0.0000 | 0.880201 |
|  | X2 | X4 | X5 | X7 | X8 |  |
| X2  X4  X5  X7  X8 | 0.262090  t=3.974437  p=0.0000 | 0.688109  t=6.051445  p=0.0000 | 0.803606  t=55.69391  p=0.0000 | 0.008451  t=3.406633  p=0.0000 | 0.699066  t=8.962385  p=0.0000 | 0.889557 |

逐步回归的五元结果

加入X3时模型上升最高，四个变量p值均小于0.05，表明四个变量对被解释变量影响显著。

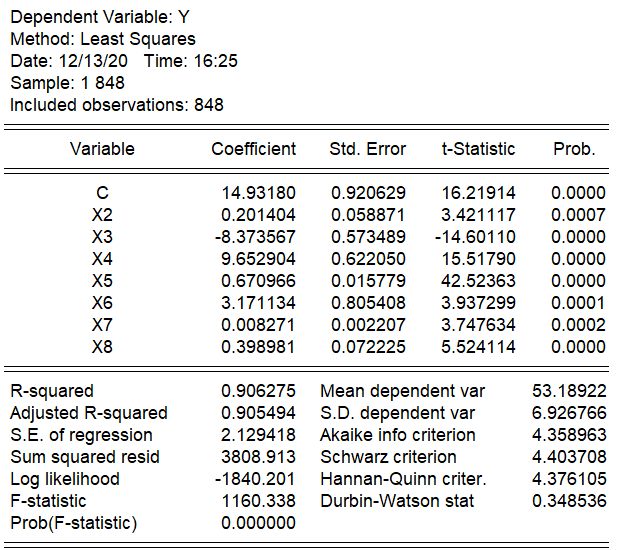
（6）以X2、X3、X4、X5、X8为基础，顺次加入其他变量，进行回归，结果如下：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | X2 | X3 | X4 | X5 | X6 | X8 |  |
| X2  X3  X4  X5  X6  X8 | 0.252747  t=4.38060  p=0.0000 | -8.37571  t=-14.49  p=0.0000 | 9.713888  t=15.5015  p=0.0000 | 0.667247  T=42.0470  P=0.0000 | 3.232004  t=3.98292  p=0.0001 | 0.424426  t=5.85731  p=0.0000 | 0.904708 |
|  | X2 | X3 | X4 | X5 | X7 | X8 |  |
| X2  X3  X4  X5  X7  X8 | 0.21675  t=3.65855  p=0.0000 | -8.27663  t=-14.32  p=0.0003 | 9.54552  t=15.2293  p=0.0000 | 0.67103  t=42.1655  p=0.0000 | 0.00845  t=3.79522  p=0.0000 | 0.423454  t=5.83471  p=0.0000 | 0.904545 |

逐步回归的六元结果

加入X2时模型上升最高，四个变量p值均小于0.05，表明四个变量对被解释变量影响显著。

（7）将全部变量加入模型的结果为：



加入X7的结果

所有变量p值均大于0.05，且修正后的可决系数有所提升。因此不剔除变量，最终的模型估计结果为

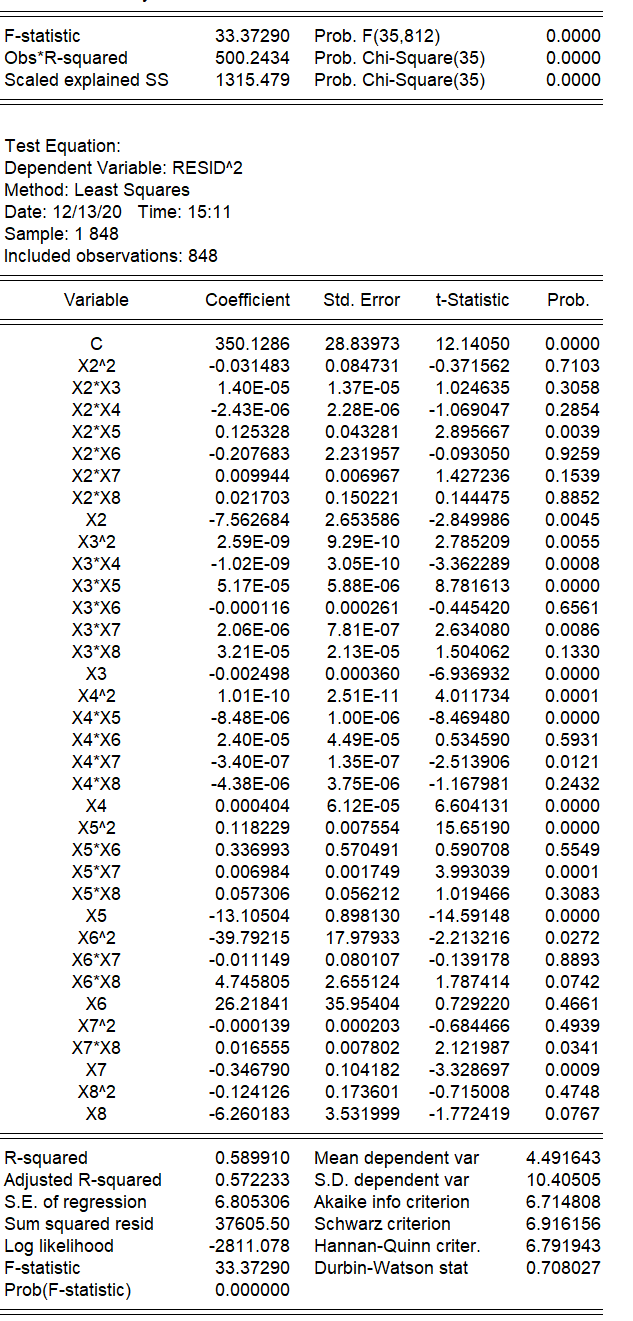
（六）异方差检验与补救

为了避免异方差对模型的影响，还需进行异方差检验。

1.异方差检验

（）White检验

White检验的基本思想是，如果模型存在异方差，其方差与解释变量有关系，分析方差是否与解释变量有某些形式的联系以判断是否存在异方差，一般以OLS估计残差平方与有效样本量的乘积作为其估计量。[[10]](#footnote-10)White检验结果如图：



White检验结果

从表中可知，，在自由度下，自由度p=35，

因此，拒绝原假设，认为模型存在异方差。

（3）ARCH检验

ARCH检验是针对时间序列提出的异方差检验方法，基本思想是做随机误差项方差对其p阶滞后值的自回归，根据滞后变量系数的显著性判断随机误差方差是否受前p期方差的影响。

取p=1,2,3,4,5，分别作ARCH检验，结果整理如下：

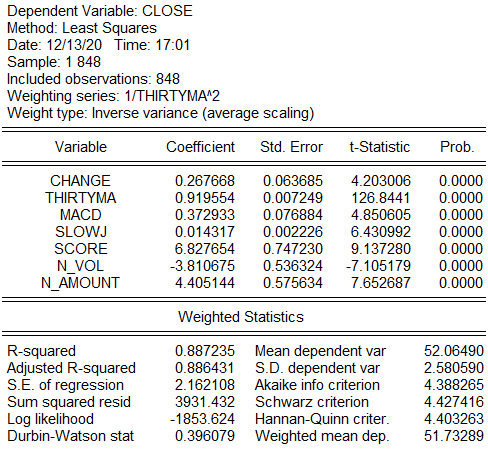
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 阶数 |  | 卡方检验P值 | AIC | SIC |
| 1 | 546.1570 | 0.0000 | 6.491882 | 6.503079 |
| 2 | 547.1891 | 0.0000 | 6.489746 | 6.506556 |
| 3 | 546.5140 | 0.0000 | 6.493280 | 6.515715 |
| 4 | 547.4966 | 0.0000 | 6.491187 | 6.519257 |
| 5 | 553.7701 | 0.0000 | 6.470925 | 6.504641 |

各阶ARCH检验结果

AIC、SIC取值最小的阶数均为5，因此选取阶数为5的结果进行检验，此时P值=0.0000<0.05，因此拒绝原假设，认为该模型存在异方差。

2.异方差修正

采用加权最小二乘法对异方差进行修正，取权重为，回归结果如下：



权重为的WLS结果

对回归结果再次进行White检验，得出p值为0.0000，拒绝原假设，取权重后模型仍然存在异方差，无法将参数估计该会员制。换其他权重再次进行加权最小二乘估计。

选取各变量各形式的加权最小二乘估计，其White检验结果如下表所示：

|  |  |
| --- | --- |
| 权重 | P值 |
|  | 0.0000 |
|  | 0.0001 |
|  | 0.0000 |
|  | 0.0006 |
|  | 0.0000 |
|  | 0.0001 |
|  | 0.0001 |
|  | 0.0000 |
|  | 0.0000 |
|  | 0.0002 |
|  | 0.0001 |
|  | 0.0000 |

WLS后White检验结果

由上表可知，各种权重下p值均小于0.05，无法有效消除异方差。

因此，本文转换补救思路，从消除非球形扰动项改为在非球形扰动项的情况下，对模型估计量方差进行修正。此时对模型进行自相关检验以确定选择可行最小二乘法或New-West一致标准误差法。

(七)自相关检验与补救

1.自相关检验

（1）DW检验

首先判单该模型是否满足前提：

①模型七个解释变量均为非随机

②假设该模型随机误差项自回归形式为一阶线性回归形式

③模型中解释变量不包含滞后的被解释变量

④该回归模型含截距项

⑤数据序列无缺失项

综上，该模型适用于DW模型。

样本量为848，七个解释变量，5%显著性水平时，查DW统计表可知，由于目前难以得知样本量为848的情况，故合理推测，样本量为800时，应更大，又因为模型的DW=0. 348536<,也必然小于n=848时的，表明该模型存在正自相关。

（2）BG(LM)检验

选取阶数为1-5，比较其AIC、SIC取值，得到最优阶数。

结果如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 阶数 | LM | P值 | AIC | SIC |
| 1 | 625.6319 | 0.0000 | 3.022775 | 3.073113 |
| 2 | 629.7425 | 0.0000 | 3.006473 | 3.062405 |
| 3 | 631.8847 | 0.0000 | 2.998970 | 3.060494 |
| 4 | 631.8956 | 0.0000 | 3.068395 | 3.001278 |
| 5 | 634.3237 | 0.0000 | 2.992337 | 3.065047 |

BG检验各界检验结果

AIC、SIC取值最小的阶数为5，因此选取阶数为5的结果进行检验，此时P值为0.0000，拒绝原假设，认为该模型存在自相关。

综上，该模型存在自相关。

2.自相关的补救

采用广义差分法对自相关进行补救

假定为一阶自回归形式：

对模型进行广义差分：

)))

))))

令

该模型随机误差项满足古典假定，可以利用最小二乘法进行参数估计，并反代得到原参数。但是进行差分的系数未知，要对进行估计。

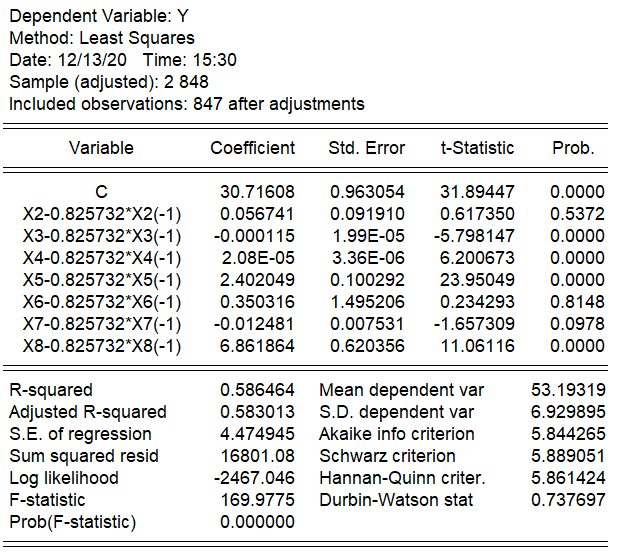
（1）利用DW估计

由DW=0. 348536计算得：

生成新序列，

（i=2,3,4,5,6,7,8）

对广义方差后的模型进行最小二乘估计，结果如下：



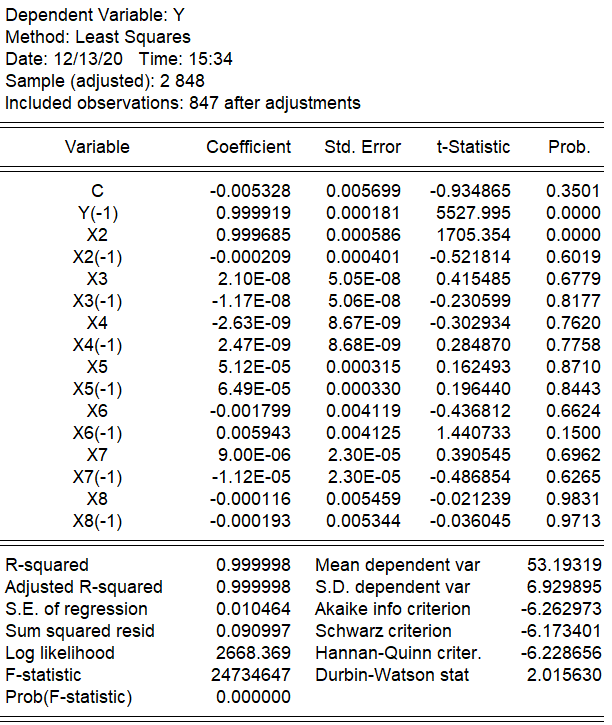
按DW估计广义差分结果

由回归结果可知，DM=0.737697，由于使用了广义差分，样本容量变为846。DW<临界值，表明广义差分后该模型依然存在自相关，该方法估计的不可用。

（2）德宾两步法估计

①第一步，做回归

用OLS法估计结果如下:

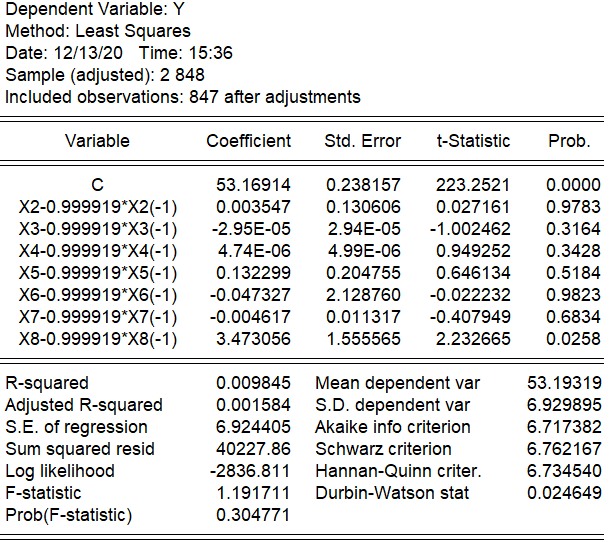


其中系数的估计即，可知

②第二步：生成新序列，

,（i=2,3,4,5,6,7,8）

对广义差分后的新模型进行最小二乘估计，结果如下：



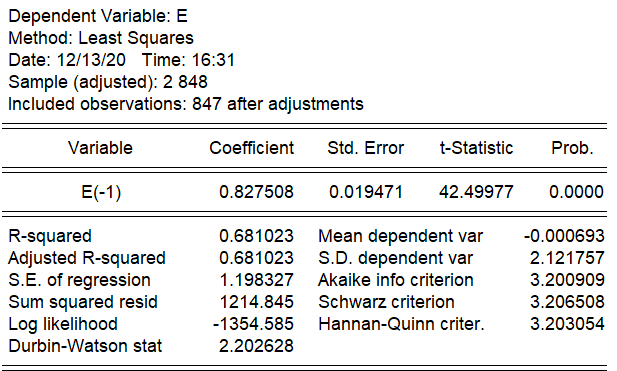
杜宾两步法广义查分估计结果

由回归结果可知，DW=0.024649

由于受用了广义差分，样本容量变为846。依然沿用对原模型进行DW检验的方法。DW<临界值，表明广义差分后该模型依然存在自相关，该方法估计的不可用。

（3）用残差序列估计

使用原方程OLS估计后的残差进行过原点的自回归，结果如下：

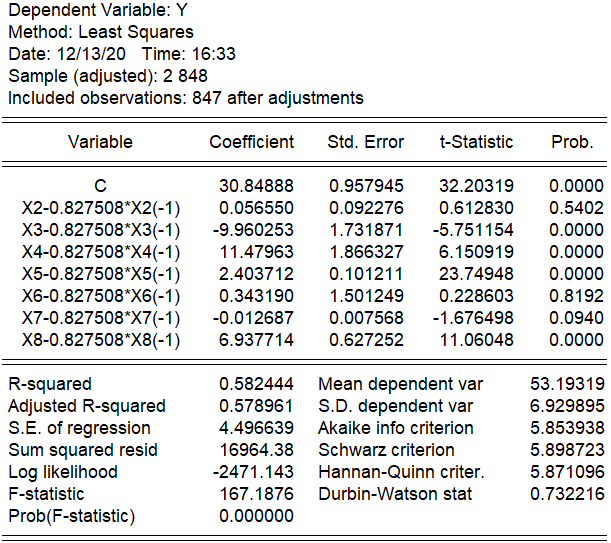


用残差序列估计的结果

可知，生成新序列

,（i=2,3,4,5,6,7,8）

对广义差分后的模型进行最小二乘回归，结果如下：

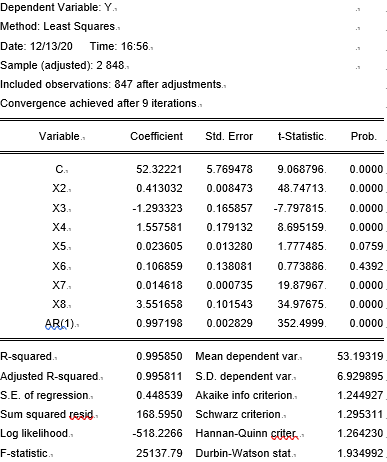


用残差序列估计广义查分结果

由回归结果得DW=0.732216。依然沿用对原模型进行DW检验的方法。DW<临界值，表明广义差分后该模型依然存在自相关，该方法估计的不可用。

（4）科克伦-奥克特迭代法估计

用Eviews进行迭代，最终结果为：



科克伦-奥克特迭代法结果

经过9次迭代，得到，DW=1.934992，较为接近2，模型自相关很可能已经消除。

选取阶数1-5对广义差分进行BG检验，结果如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 阶数 | LM | P值 | AIC | SIC |
| 1 | 9.32687 | 0.0477 | 1.224206 | 1.280188 |
| 2 | 6.75330 | 0.0955 | 1.217554 | 1.279135 |
| 3 | 19.54100 | 0.1235 | 1.166341 | 1.233520 |
| 4 | 9.79713 | 0.0505 | 1.168373 | 1.241150 |
| 5 | 12.77920 | 0.0134 | 1.166890 | 1.245265 |

各阶BG检验结果

AIC、SIC取值最小的阶数均为3，因此选取阶数为1时的结果进行检验，此时P值为0.6235>0.05，因此不能拒绝原假设，认为经过广义差分后模型自相关已消除。

3.对广义差分后的模型进行异方差检验

（1）ARCH检验

取p=1,2,3,4,5，分别作ARCH检验，结果整理如下

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 阶数 |  | 卡方检验P值 | AIC | SIC |
| 1 | 2.79258 | 0.0352 | 0.828563 | 0.839769 |
| 2 | 1.22062 | 0.0990 | 0.771232 | 0.788058 |
| 3 | 1.27738 | 0.1363 | 0.756993 | 0.748389 |
| 4 | 1.8640 | 0.1400 | 0.709948 | 0.738044 |
| 5 | 1.0762 | 0.0843 | 0.685934 | 0.719680 |

广义差分后各阶ARCH检验

AIC、SIC取值最小的阶数均为5，因此选取阶数为5时的结果进行检验，此时p值=0.0843>0.05，因此不拒绝原假设，认为该模型已不存在异方差。

综上，广义方差后的模型通过计量经济学检验，还原为原模型结果如下：

（5.769478） (0.008478) (0.165857) (0.179132) (0.13280)

(0.138081) (0.000735) (0.101543)

t=(9.068796) (48.74713) (-7.797815) (8.695159) (1.777485)

(0.773886) (19.87967) (34.97675)

p=(0.0000) (0.0000) (0.0000) (0.0000) (0.0759)

(0.4392) (0.0000) (0.0000)

4.对广义差分后的模型进行统计检验

**t检验**

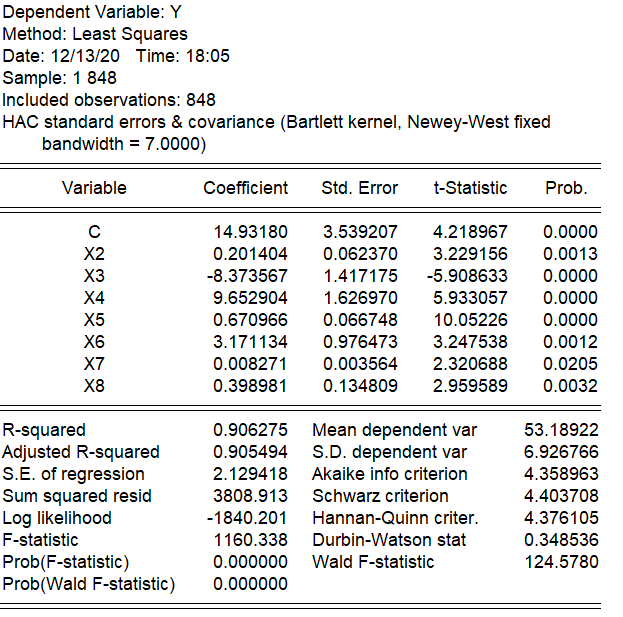
分别针对，在给定显著性水平下，除外，p值均小于0.05，拒绝原假设，对被解释变量影响显著。

因此，本文决定尝试Newey—West一致标准误差法，看是否显著。

（八）Newey—West 异方差和自相关一致标准误差法

由以上讨论可知，通过最小二乘估计无法有效消除异方差，通过广义差分法得到的模型，解释变量对被解释变量没有显著影响，且该模型存在自相关与异方差问题，因此应当采用Newey—West一致标准误差法，在存在非球形扰动项的情况下，对模型估计方差的估计式进行检验。

用Eviews对模型进行回归，结果如下：



Newey—West一致标准误差法回归结果

和原回归结果相比，参数估计值没有改变，参数标准差的估计值半大，解决了由异方差和自相关导致的对参数估计不准的问题。

三、结论

1.模型最终回归结果

综上，模型最终估计为：

（3.539207） (0.062370) (1.417175) (1.626970) (0.066748)

(0.976473) (0.003564) (0.134809)

t=(4.218967) (3.229156) (-5.908633) (5.933057) (10.05226)

(3.247538) (2.320688) (2.959589)

p=(0.0000) (0.0012) (0.0000) (0.0000) (0.0000)

(0.0012) (0.0205) (0.0032)

2.统计检验

（1）拟合优度检验

由结果可知，，这说明解释变量联合起来可以解释被解释变量超过90%的变异，拟合效果较好。

（2）F检验

Wald F=124.6780，检验p值=0.000000，小于0.05的显著性水平，因此认为解释变量联合起来对被解释变量影响效果显著。

（3）t检验

七个系数的估计值t检验的P值均小于0.05的显著性水平，因此认为解释变量联合起来对被解释变量影响效果显著。

## 三、政策建议

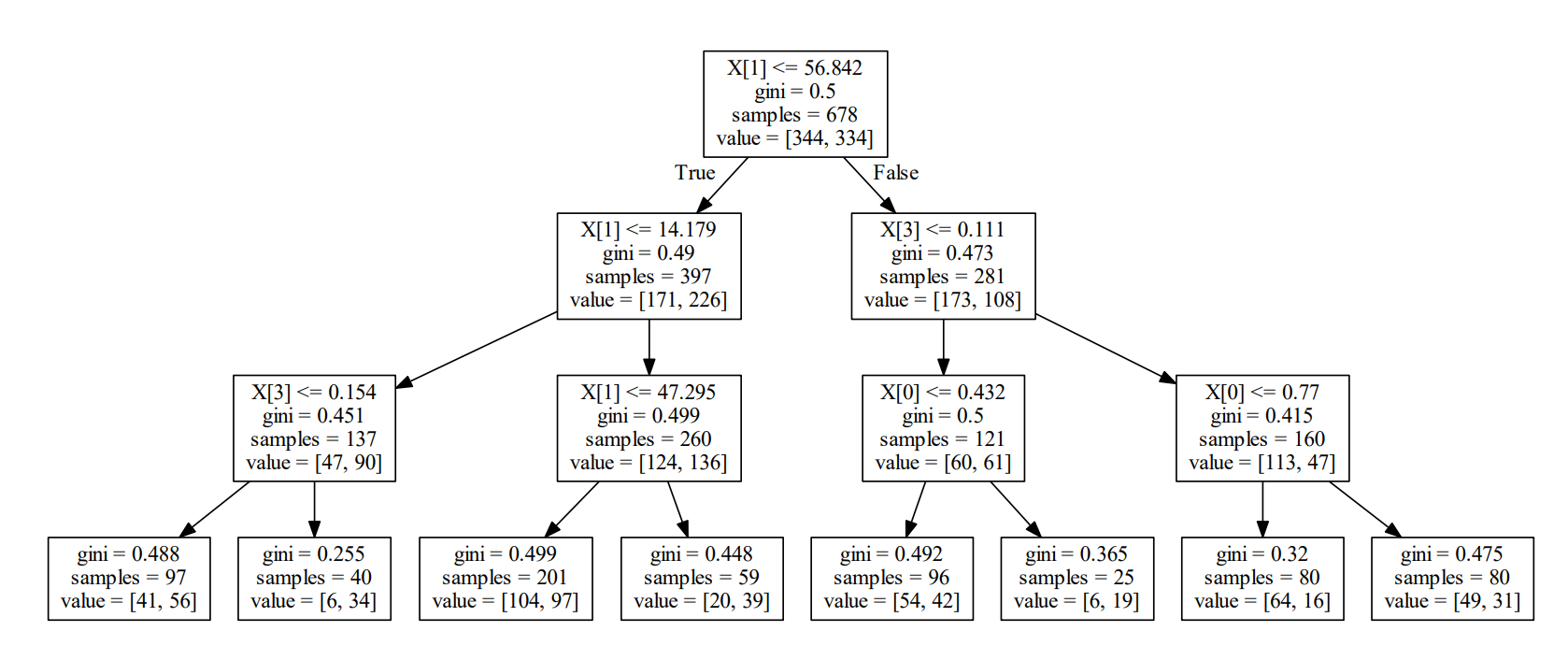
通过上文的解释研究过后，我们可以得出七个解释变量联合后对被解释变量的影响效果显著。

决策树模型是一种简单易用的非参数分类器。它不需要对数据有任何的先验假设，计算速度较快，结果容易解释，而且稳健性强[[11]](#footnote-11)。决策树模型就是由决策点、策略点(事件点)及结果构成的树形图，一般应用于序列决策中，通常以最大收益期望值或最低期望成本作为决策准则，通过图解方式求解在不同条件下各类方案的效益值，然后通过比较，做出决策。[[12]](#footnote-12)在前人的研究基础上[[13]](#footnote-13)，证明决策树是一种良好的决策模型，在此我们决定引入决策树模型来辅助散户进行决策。

首先，为了方便决策树进行训练，我们通过将每日股价波动映射为0-1变量作为训练标签。具体而言，当前日股价大于或者等于今日股价时，我们认为股价没有上涨，记今日上涨趋势为1。反之，则记今日上涨趋势为0 。

然后将标准化后的七个属性作为训练特征，上涨趋势作为决策标签，输入到决策树模型中。决策树有着诸多参数，经过网格搜索（GridSearchCV）来搜索得到最优参数，得到最优参数后训练得到最优模型。

调用决策树可视化包，最优的训练模型如下图所示：



图：决策树模型

由此图我们给出建议：

在预测股票走势，决定投资方向时应该首先注意最近三十日均线与MACD值，其后再关注每日情绪指标以及KDJ中的J值，然后做出投资决策。

1. 孟雪兰,孟祥兰,胡杨洋.基于文本挖掘和百度指数的投资者情绪指数研究[J].宏观经济研究,2016,(1). [↑](#footnote-ref-1)
2. 刘少伟.投资者舆情指数对股价波动风险影响的研究[J].金融管理研究,2018,(2):200-227. [↑](#footnote-ref-2)
3. 牛雪琪.整合网络股评舆情的股价涨跌预测建模与评价[D].东北财经大学,2019. [↑](#footnote-ref-3)
4. 孟勇,常静.投资者情绪对规模效应的影响[J].统计与信息论坛,2019,(4):98-104. [↑](#footnote-ref-4)
5. 王夫乐,王相悦.社会情绪是否会影响股市受益—来自新浪微博的证据[J].山西财经大学学报,2017,(2):39. [↑](#footnote-ref-5)
6. 李响,田路,王谦,李群.投资者情绪对股票投资收益的影响研究[J].数学的实践与认识,2020,(18):258-268. [↑](#footnote-ref-6)
7. 闵峰,黄创霞,文凤华,杨晓光.宏观经济、投资者情绪和股票市场收益[J].系统科学与数学,2017,(2):370-382. [↑](#footnote-ref-7)
8. 《神经网络与深度学习 基于Tensorflow框架和Python技术实现》，包子阳，北京，电子工业出版社，2019.04。 [↑](#footnote-ref-8)
9. 引自《量化投资 以R语言为工具》，蔡立耑，北京：电子工业出版社，2016.01。 [↑](#footnote-ref-9)
10. 庞皓，《计量经济学》（第四版），北京，科学出版社，2019.01。 [↑](#footnote-ref-10)
11. 李慧, 余明. 基于决策树模型的湿地信息挖掘与结果分析[J]. 地球信息科学学报, 2007, 9(002):60-64. [↑](#footnote-ref-11)
12. 徐晓萍, 马文杰. 非上市中小企业贷款违约率的定量分析——基于判别分析法和决策树模型的分析[J]. 金融研究, 2011(3):111-120. [↑](#footnote-ref-12)
13. 陈涛. 基于聚类和决策树算法的股票配置策略研究[D].北京交通大学,2019. [↑](#footnote-ref-13)