基于LDA 模型的中国央行货币政策分析

李冠佑　肖星辰
重庆中学，中国·重庆 404100

【摘 要】中国央行货币政策执行报告包含大量对投资、政策等相关方向有用的信息。本文针对如何提取执行报告中的有效信息，提出基于LDA 主题模型的方法，将文本进行清洗、去除停用词之后，获得每篇文章对应的主题数据，并将主题总结成贷款、金融、政策、融资、增长、改革等6 类，并结合实际情况，得出结论：增长类主题是央行老生常谈的问题，其余5 类只是阶段性地出现在执行报告中。在对LDA 模型稳定性分析的过程中，通过改变分类主题数目的方法，发现主题的增多，只是从部分原始主题中分裂出新的主题，因此验证LDA 模型的稳定性以及正确性。在得到央行货币政策执行报告的LDA 主题数据之后，我们将该数据与常见的宏观经济指标进行Spearman 相关性分析，得到贷款、金融主题与GDP、M2、贷款总额存在着很强的单调性关系。基于LDA主题数据的阶段性特点，再次将主题数据与宏观经济指标进行分时段线性回归分析，进一步得到了贷款主题与GDP、M2、贷款总额，以及金融主题与GDP、M2、贷款总额以及人民币汇率之间的线性回归关系，并且通过显著性检验之后，证实了模型的有效性。

【关键词】LDA；主题模型；相关性分析；央行货币政策

1 绪论

随着人工智能技术和大数据分析技术的不断成熟，技术发展带来的高效率使得对海量数据进行捕捉、管理和处理成为了可能且正被广泛适用于各个领域。人工智能的不断发展使得并行计算变得更快、更便宜、更有效，从而推动了机器学习的长足发展实现了弱人工智能。为了克服机器学习对算法的僵化依赖和对强人工智能的追求，算法的进一步发展推动了深度学习的快速兴起，发展了机器学习的技术。目前深度学习的两大主流之一便是自然语言处理。本研究旨在众多自然语言处理的算法中寻求一个合适的模型，对央行货币政策执行报告中的文本语料进行深入分析。目前文本语料的分析处理大多集中在两个问题上，第一个问题就是文本的表示，也就是对词的分解。第二个问题是文本的学习问题。目前许多文本分类的方法已经应用在了文本语料的挖掘上。

2 LDA 模型介绍及数据处理

2.1 概率模型基础

2.1.1 Beta 分布

在概率统计中，Beta 分布是指一组定义在（0，1）区间上的连续概率分布，存在参数α 和β，（α，$\beta>0$）

$$f(x;\alpha\beta)=\frac{x^{\beta-1}(1-x)^{\beta-1}}{\int_0^1 u^{\beta-1}(1-u)^{\beta-1}du}=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{d-1}(1-x)^{\beta-1}=\frac{1}{B(\alpha,\beta)}x^{a-1}(1-x)^{\beta-1}$$

故Beta分布的概率密度函数为：

上式中Γ（x）=$\int_0^{\infty}t^{x-1}e^{-t}dz$，且随机变量X 满足x ~B（$\alpha$，$\beta$）。图2.1为Beta 分布的概率密度函数图像

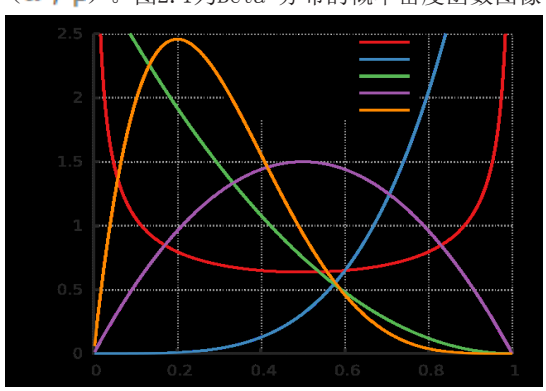

图2.1：Beta分布的概率密度函数

2.1.2 狄利克雷分布

狄利克雷（Dirichlet）分布可以看做是Beta 分布在多维情况下的应用，概率密度函数可定义为：

$$Dir(\vec{p}|\vec{\alpha})=\frac{1}{B(\vec{\alpha})}\prod_{k=1}^{K}P_k^{\alpha_x-1}$$

式中，$\vec{\alpha}$=（$\alpha1$，$\alpha2$，…，αk）为狄利克雷分布的参数，且$\alpha1$，$\alpha2$，…，$\alpha k>0$，B（$\vec{\alpha}$）表示狄利克雷分布的归一化常数，$\int\prod_{k=1}^{K}P_k^{\alpha_x-1}d\vec{p}$。

狄利克雷分布的期望为：

$$E(\vec{p})=\left(\frac{\alpha1}{\sum_{k=1}^{k}\alpha_k}\right),\ \left(\frac{\alpha2}{\sum_{k=1}^{k}\alpha_k}\right),\ \cdots,\ \left(\frac{\alpha k}{\sum_{k=1}^{k}\alpha_k}\right)$$

2.2 语言模型

2.2.1 一元语言模型

语言模型就是确定一串词序列的概率分布，即是判断一句话是否通顺，是否符合该语言的语法。具体来说，语言模型就是为一个长度为k 的文本确定一个概率分布p，从而确定这段文本存在的可能性。假设以S=$W_1,W_2,\cdots,W_k$表示一个句子中的词语序列，其概率可以用下式来表示：

$$P(S)=(W_1,W_2,\cdots,W_{kl})=P(W_1)P(W_2|W_1)\cdots P(W_k|W_1,W_1,\ldots,W_{k-1})$$

如果文本的长度较长，$P(W_k|W_1,W_1,\ldots,W_{k-1})$的估算会非常困难。该方法在应用过程中存在两个不足，一是参数空间会过于庞大，二是现实的数据往往都十分稀疏，导致语言模型应用过程中出现严重的数据稀疏性问题。为了解决参数空间过大的问题，我们引入了马尔可夫假设：随机抽取一个单词，他能够出现的可能性只与它前面出现的一个或者有限的几个词有关。如果一个单词的出现的可能性仅仅依赖着它前面出现的一个词，我们就将其定义为bigram：

$$P(S)=(W_1,W_2,\cdots,W_{kl})=P(W_1)P(W_2|W_1)P(W_3|W_2)\ldots P(W_k|W_{k-1})$$

一般来说，N 元模型就是假设当前词的出现概率只与在他之前的N-1 个词有关，而上述所说的这些概率参数模型，都是可以通过大规模的语料库来计算。根据以上的理论基础，我们可以将一元语言模型的概率重新定义为：

P（S）=（W_1）P（W_2…P（W_n））

上一小节文章描述了Dirichlet 分布的相关知识。其中关键点就是：Dirichlet 先验概率+ 多项分布的数据=Dirichlet 后验分布，其表达式如下所示：

$$Dirichlet(\vec{p}|\vec{\alpha})+MultCount(\vec{n})=Dirichlet(\vec{p}|\vec{\alpha}+\vec{n})$$

这样，在给定参数$\vec{p}$的先验分布$Dir(\vec{p}|\vec{\alpha})$，每个词出现的频次的数据为：

$$\vec{n}\ Mult(\vec{n}|\vec{p},\ N)$$

由于上式为多项分布，所以我们无需通过计算就可以推出后验分布：

$$p(\vec{p}|\omega,\ \vec{\alpha})=Dir(\vec{p}|\vec{\alpha}+\vec{n})=\frac{1}{\Delta(\vec{\alpha}+\vec{n})}\prod_{k+1}^{V}p_k^{n_k+\alpha_k-1}$$

本研究中，我们将均值作为参数估计值并结合狄利克雷分布可得到该文本语料产生的概率为

$$\begin{aligned}p(\vec{p}|\omega,\vec{\alpha}) &= \int p(\omega|\vec{p})p(\vec{p}|\vec{\alpha})d\vec{p}\\ &= \int\prod_{k=1}^{V}p_k^{n_k}Dir(\vec{p}|\vec{\alpha})d\vec{p}\\ &= \int\prod_{k=1}^{V}p_k^{n_k}\frac{1}{\Delta(\vec{\alpha})}\prod_{k=1}^{V}p_k^{\alpha_k-1}d\vec{p}\\ &= \frac{1}{\Delta(\vec{\alpha})}\int\prod_{k=1}^{V}p_k^{n_k+\alpha_k-1}d\vec{p}\\ &= \frac{\Delta(\vec{n}+\vec{\alpha})}{\Delta(\vec{\alpha})}\end{aligned}$$

2.2.2 PLSA 主题模型

2.2.1 节的Unigram Model 是一个简单的语言处理模型；在该模型中，由于假设过于简单，与人类写作过程中产生单词的方式有很大的差异，因此我们需要寻找一个更为贴合人类实际的模型。

从人类构思主题并形成文章内容的过程出发：如果我们想撰写一篇人工智能的文章，我们必须首先想好文章会出现的出题，比如，机器学习占一部分，计算机占一部分，另外还会涉及到其他主题。对应于每个主题，我们需要寻找与之相关的词语，比如计算机主题下，我们会涉及到CPU、计算速度；机器学习主题下，会出现诸如分类的的词语，这是因为不同的词语在不同主题下出现的概率是不一样的。Hoffmn 对以上过程进行了数学化并建立起PLSA 模型：多个主题（Topic）共同构成了一篇文章，多个词语构成一个主题，且每个主题都是词汇上的概率分布。在第m 篇文档中，每个词的生成概率为：

$$p(w|d_m)=\sum_{z=1}^{k}p(w|z)p(z|d_m)=\sum_{z1}^{k}\varphi_{zw}\theta_{mz}$$

所以整篇文档的生成概率为：

$$p\ (\vec{w}|d_m)\ =\ \prod_{i=1}^{n}\sum_{z=1}^{K}p(w|z)\ \ p\ (z|d_m)\ =\ \sum_{z=1}^{K}\varphi_{zwi}\theta_{mz}$$

由于文档之间相互独立，我们也容易写出整个语料的生成概率。求解PLSA 这个Topic Model 的过程汇总，模型参数求解可以使用著名的EM 算法求得局部最优解。

2.2.3 LDA 主题模型

类似于Hoffmn 对Unigram Model 的贝叶斯改造，当我们在如下两个概率参数前加上先验分布，就可以把PLSA 的游戏过程变成贝叶斯过程。当把PLSA 模型的先验分布改为狄利克雷分布，我们就可以得到LDA(Latent Dirichlet Allocation) 模型。

我们发现LDA 的模型生成过程：先按照概率$p(d_i)$ 选中一篇文档d_i，然后从狄利克雷分布α中抽样生成文档d_i的主题分布θ_m，紧接着从主题分布θ_m中抽取文档d_i的第j 个词的主题$z_{m,\ n}$，再接着抽样生成主题$z_{m,\ n}$对应的词分布φ_k，最后从词分布φ_k中生成抽样词$\omega_{m,\ n}$。为了方便求解。前两个过程可以用到Unigram Model 的结论，即对整个语料库下所有主题词进行编号并生成以下概率：

$$p(\vec{z}|\vec{\alpha})=\prod_{m}^{M}\frac{\Delta(\vec{n_w}+\vec{\alpha})}{\Delta(\vec{\alpha})}$$

对于词的生成过程（主题编号的选择不会改变K 个主题下的词分布），可表示为：

$$p(\vec{\omega}|\vec{z},\ \vec{\beta})=p(\vec{\omega}|\vec{\beta})=\prod_{k}^{V}\frac{\Delta(\vec{v_K}+\vec{\beta})}{\Delta(\vec{\beta})}$$

因此，LDA 模型语料库的生成概率可以表示为：

$$p(\vec{\omega},\ \vec{z}|\vec{\alpha},\ \vec{\beta})=\prod_{k}^{K}\frac{\Delta(\vec{v_K}+\vec{\beta})}{\Delta(\vec{\beta})}\cdot\prod_{m}^{M}\frac{\Delta(\vec{n_m}+\vec{\alpha})}{\Delta(\vec{\alpha})}$$

2.2.4 文本预处理与实验设计

《中国货币政策执行报告》是中国人民银行自2001年开始按季度对外公开发布的有关货币政策执行情况的报告。该报告是从官方的角度为大家深入分析宏观经济金融形势，介绍上一季度的货币政策操作，并展望下一季度的政策及其走向。一般报告内容主要包括以下部分：一、货币信贷概况；二、货币政策操作；三、金融市场分析；四、宏观经济分析；五、预测与展望。该报告中包含了官方对过去经济市场发展的总结，以及政策展望，因此对于中小投资者来说是非常重要的，可用于指导投资工作的有力武器。

本研究涉及到的数据资源包括2001-1 季度到2019-1季度的《中国货币政策执行报告》，其中2016-1 季度报告缺失，共计72 个PDF 文档。所有数据均从中国人民银行官网货币政策司下载得到。对文本的预处理包括原始文档的格式转化，分词、以及修正为符合LDA模型的输入格式. 处理过程如下：

(1) 使用Python对原始72个PDF文档的文字部分进行批量读取并写入到TXT格式，保存文本文件；

(2) 使用jieba 进行分词，去除stopwords，并保存；

(3) 将分词之后的文本文件进一步修改格式，输入到LDA 模型。

LDA 模型参数设置：主题数为6 类；迭代次数3000次；α= 0.1，β= 0.1，每个类的高频词显示个数为20 个。

2.3 结论与分析

2.3.1 LDA 模型结果及其可视化

(1)增长主题所占面积最大，该主题占比在60% 到70%范围内波动，这完全符合人民银行保证经济增长的目标。

(2)改革主题在2003-1 季度到2012-2 季度出现面积较大，最大值出现在2007 末，达到34%，说明央行在该时间段内除增长以外对改革这一话题的关注度最高。经查阅发现，2013 年十五届三中全会国务院国有资产监督管理委员会挂名成立，到2012 年十七届三中全会国资委出台规定，民间投资主体可通过出资入股等形式参与国企改制重组。

这两个事件为国企改革的标志时间，与我们的改革主题模型吻合。

(3)融资主题出现在2008-1 到2013-2 季度，在时间上正好对应了2008 年以及2012年的经济金融危机时期。2010-4 季度占比最大值为32%，这一时间段除增长以外央行对融资主题关注度最高。

(4)贷款主题出现在2001-1 到2006-4 季度，最高值为60%， 出现在2001-3 季度，高于增长主题的35%。金融主题从2014-2 季度进入央行的视野，并且一直增加，在2018年-4 季度达到峰值39%。

(5)在该报告中，并非6大主题每次都会出现；例如2001-1季度到2006-4季度，基本上只有增加、贷款跟改革主题出现；在2014-2季度至今，只存在增长、金融跟政策三大主题。这说明了货币政策的发布紧跟时下热点经济问题，同时也表明中国的央行在调控宏观经济中注重时效性、目的性明确。

3 LDA 主题与经济数据相关性分析

3.1 主要经济指标的选择与处理

根据上述章节所得到的主题，我们选择与主题密切相关的宏观经济指标，包括：GDP（国内生产总值）、CPI（居民消费价格指数）环比、企业景气指数、企业家信心指数、上证所A 股成交金额、M2 （货币供应量）、人民币汇率（兑美元）、贷款总额（数据来源：东方财富网数据中心）。由于上述部分指标只有月度数据，根据指标意义的不同，我们采用不同的方法得到该指标的季度数据。处理过程如下所示：

（1）人民币汇率（兑美元）：以每季度对应的月度数据求平均所得：$X_i=\frac{a_1+a_2+a_3}{3}$，式中$X_i$为人民币汇率季度数据，$a_i$为该季度所对应三个月的汇率月度数据。

（2）M2（货币供应量）、贷款总额：以每季度对应的月度数据求和所得：$X_i=a_1+a_2+a_3$，式中X_i为经济指标季度数据，a_i为该季度所对应三个月的经济指标月度数据。

（3）CPI（居民消费价格指数）环比：假设现在有月份0;1; : : : ; 6 对应月环比数值为a_0， a_1，a_2，…a_6，第0 月代表要计算的第一个季度前的那个月，因为我们只有环比数据，所以考虑到该季度的第1 个月的环比数值包含第0 月的信息；那么月份1; 2; 3 对应的则为第0 季度的CPI 环比数据b0，同理4; 5; 6 则可以求出第1 季度的环比数据b_1。计算方法如下：

(a) 先将每个月的环比数据变成实际比例：$Pi=\frac{a_1+100}{100}$，式中Pi为实际比例；

(b) 然后得到每个月相对于第0 月的增长比例： $Ci=\prod_{k=1}^{i} p_i$，式中Ci为每个月相较于第0 月的比例；

(c) 最后即可得到第一季度的环比数据b_1：

$$b_1=\frac{\sum_{i=1}^{3} c_i}{\sum_{i=1}^{6} c_i}$$

同理，可以得到每个季度CPI 环比数据。

3.2数据分析

3.2.1 数学模型

对主题数据进行正态性检验，我们可以发现其并不满足正态分布，因此不能使用Pearson 相关系数来描述变量之间关系.2.。在这里，我们采用Spearman 秩相关系数。Spearman 秩相关系数是Spearman 在1904 年提出的一种秩统计参数，用于度量变量之间的单调性特征，其与数据分布无关，即：非参数性质。

定义**3.1** (Spearman 秩相关系数)**.** Spearman 秩相关系数，是在Pearson 相关系数的基础上，利用两个变量中元素在各自集合中的排序来计算其相关性。

假设原始变量为x; y，长度为n，按照升序或者降序进行排序，得到x′ ; y′ ，则x′ ; y′ 为每对变量在排序之后的秩次。原始位置相同的变量的秩次差为di = x′ —y′ 。Spearman 秩相关次数计算公式如下：

$$p_s=1-\frac{6\sum d_i^2}{n(n^2-1)}$$

Pearson 相关系数研究的是变量之间的线性相关性，当变量之间线性关系越强（正相关或逆相关），Pearson 系数的绝对值越趋近于±1。而Spearman 则研究单调相关性，即：两个变量之间满足单调递增，则Spearman 系数趋近于1；反之，两变量之间跃接近于单调递减变化，则趋近于-1。

3.2.2 结论与分析

在本文中，我们将研究基于LDA 主题模型得到的6 个主题数据与7 个宏观经济数据之间的相关性关系。

(1) 贷款主题与GDP、上证所A 股成交金额、M2 和贷款总额存在的明显的负相关关系，最高值为 0:9，说明这些经济指标随着贷款主题数据的增大而减小，也就是当央行货币执行报告中统计与贷款相似词频越高，即可得GDP、上证所A 股成交金额、M2 和贷款总额的预测值越低，该数值只能说明变量之间的单调性关系，因为不能根据贷款主题数据，直接预测经济指标的数值大小。

(2) 金融主题同样与GDP、上证所A 股成交金额、M2 和贷款总额存在正相关，相关系数为0.59。

(3) 政策主题与GDP 的正相关性与人民币汇率的负相关性。融资主题与宏观数据之间不存在明显的相关性，因此，这一主题对于我们分析LDA 模型与宏观数据之间相关性影响不产生作用。

(4) CPI 环比与6 个主题数据的Spearman 系数绝对值很小，说明CPI 环比说明与LDA 主题不存在相关性关系。因此，我们不能用LDA 模型得到政策执行报告与CPI 这一指标之间的关系。

(5) 在上图中，我们可以看到GDP 列与M2 列、贷款总额列数值基本一致。由Spearman 秩相关系数的意义，可以得到这三个宏观数据之间应该满足强相关性。

3.3 分时段线性回归分析

3.3.1 回归分析简介

回归分析是一种经典的数理统计分析方法，根据研究变量的数量可以分为一元回归分析：常常用于研究两个变量之间的关系；或多元回归分析，即变量与多变量的关系。除了研究变量之间线性关系的线性回归之外，也存在非线性回归分析方法。回归分析一般也可用于预测及控制模型。在线性模型中，我们通常使用一元线性方程描述两个变量之间的关系，即：y=a+bx+ε

式中，a 和b 为回归系数，“为误差项，代表随机因素。对于该线性回归模型，一般满足误差项”的正态性、无偏性、共方差性等假设。

对回归分析的求解，最后落到对回归系数a; b 以及误差项的估计。最小二乘法是最常用的参数估计方法，其思想为使数据集中的每个点到该样本的回归方程的点最近。即：

$$\begin{cases} \text{贷款主题与 GDP：} y=-82295.811x+70146.48 \\ \text{贷款主题与 M2：} y=-1482401.131x+1232162.109 \\ \text{贷款主题与贷款总额：} y=-901375.34x+812992.283 \end{cases}$$

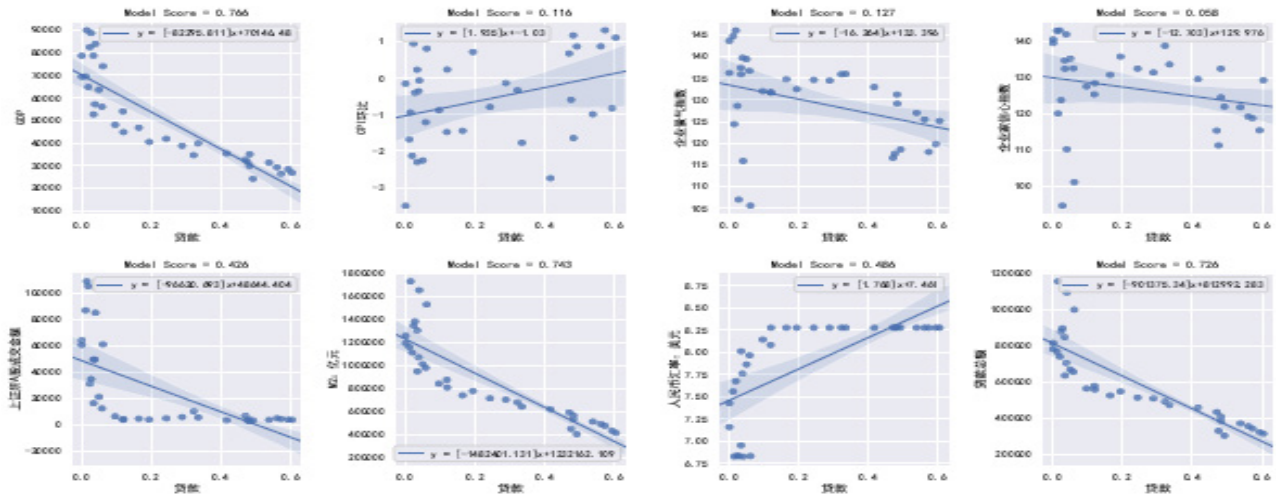

图 3.3: 贷款主题与宏观经济数据回归分析结果

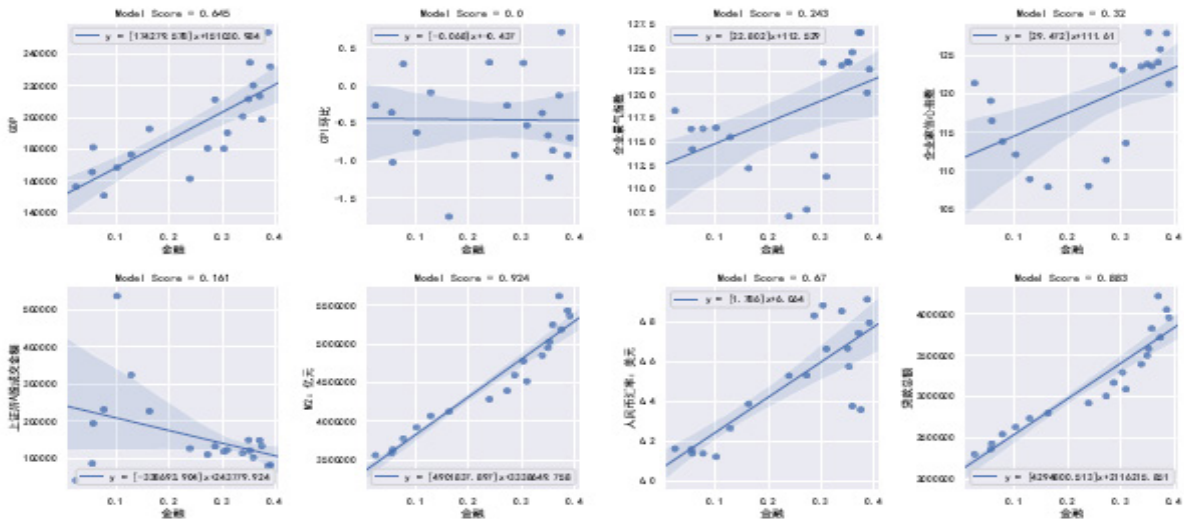

图 3.4: 金融主题与宏观经济数据回归分析结果图

$$\min_{a,\ b}\sum_{i=1}^{n}\varepsilon^2=\min\sum_{i=1}^{n}(y_1-a-bx_i)^2$$

式中，a；b 为回归系数，为误差项。

3.3.2 结论与分析

根据截取的分段主题数据与对应时间的宏观经济数学进行线性回归分析并进行t 检验；得到如图3.3表所示：

根据图3.3，可以得到，贷款主题与GDP、M2 和贷款总额的决定系数r2 分别为0.766、0.743、0.726，可以视为线性回归的拟合效果较好；在对这三个模型进行t 检验时，P值远远小于0.05，因此，可以得到贷款主题的权重数值与这三个宏观指标的函数关系为：

该结果与Spearman 模型的相关性相吻合，但是又明显地增加了相关性分析的分辨率。在Spearman秩相关性分析中，我们得到贷款数据与GDP、M2 以及贷款总额之间的相关系数基本相同，也就是只能得到贷款数据与这三个宏观指标满足明显的单调关系，但是在线性回归分析中，我们发现贷款与这三个宏观经济指标之间的决定系数r2 并不相同。

另外，LDA 模型结果与宏观数据之间的强线性关系为我们提供了另外一种预测宏观指标的方法，即：可以通过过去三个月的经济状况来预测接下来的季度报告中主题权重大小，也根据该数据预测宏观经济数据。

下图为金融主题数据与宏观经济数据所做回归分析的结果：

$$\begin{cases}\text{金融主题与 GDP：}\ y=174279.578x+151030.984\\ \text{金融主题与 M2：}\ y=4901837.897x+3338649.758\\ \text{金融主题与人民币汇率：}\ y=1.786x+6.064\\ \text{金融主题与贷款总额：}\ y=4294800.513x+2116215.851\end{cases}$$

根据图3.4，金融主题数据与M2 和贷款总额的决定系数r2 为0.924 和0.883，线性回归模型的效果很好；同时，该

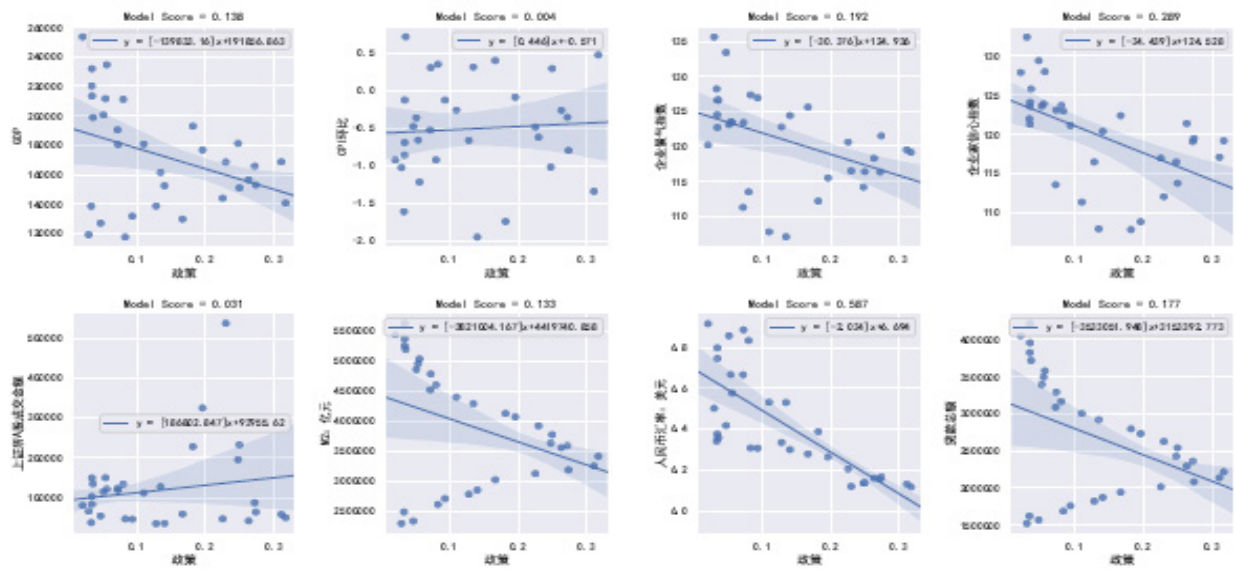

图 3.5: 政策主题与宏观经济数据回归分析结果图

主题数据同样与GDP、人民币汇率的决定系数也大于0.6，说明模型效果较好。

下式为金融主题与GDP、M2、人民币汇率以及贷款总额之间的回归方程：

图3.5为政策主题仅与人民币汇率线性相关性稍微明显，与其他指标的散点图成混乱状态；可以视为没有线性关系。

同样，在对融资、增长和改革主题数据进行相同的回归分析之后发现，这个三个指标对经济数据没有任何线性回归，因此不能用来做预测模型。

4 结论 与展望

4.1 论文主要结论

在量化投资以及利用人工智能方法进行数据挖掘的大背景下，我们摒弃了以往直接采用宏观数据或者股票市进行机器学习等方法，对央行报告这一类文本信息进行挖掘，得到其有效信息，并研究该信息与宏观经济指标之间的关系。根据本文对LDA 方法的介绍以及LDA 模型在金融经济领域的应用，本文获得如下研究成果：

(1) 本研究基于LDA 模型对2001-1 季度到2019-1 季度的央行货币政策执行报告进行主题研究，提取了报告中主要的六大主题，并将其定义为贷款、改革、金融、融资、增长、政策主题。及其次，文章对提取的六大主题的时间序列数据进行可视化展示，结果发现增长为央行的第一大宗旨，且6 个主题并不会出现在央行的每一次执行报告中，即央行执行报告的关注点每次仅关注部分主题。

(2) 对LDA 模型进行灵敏度分析。文章对参数主题个数的选取进行调整，验证LDA 模型的稳定性。结果发现，参数增大时，LDA 模型结果仅表现为原始主题到新主题的分裂；LDA 模型具有非常好的稳定性，其结果是可信赖的。

(3) 本研究分析了研究主题与宏观经济指标的相关性：由于主题数据序列存在大量的0 值情况，因此主题数据不满足正态分布特性；我们选择Spearman 相关系数研究其单调相关性，得到贷款与GDP、M2 和贷款总额满足明显的单调性特征；另外，又根据该单调性间接验证了宏观数据GDP、M2 和贷款总额之间的强线性特征。

(4) 为了消除主题时间序列中0 值产生的影响，我们采用分时段的方式，截取非0 段的主题时间数据与宏观经济指标进行线性回归分析，并绘制散点图，可以得到贷款与GDP、M2、贷款总额，以及金融与GDP、M2、人民币汇率和贷款总额之间的强线性关系。

4.2 工作展望

本文基于LDA 模型对央行货币政策执行报告进行主题挖掘，得到了上述研究成果的同时，也存在着一定的不足。

(1) LDA 模型是自然语言处理中的一种文本挖掘算法；未来的研究中，可以采用其他NLP 算法对央行货币政策执行报告进行主题挖掘，并比较不同算法之间的联系与区别；

(2) 在本文进行主题分类过程中发现每类主题下的高频词出现了不可避免的重复，这也对主题的确定产生了一定的影响。在后续工作中，我们可以调整每个主题下高频词的显示个数和调整停用词的方式进行模型修正。

(3) 本文在对LDA 模型结果与宏观数据指标进行相关性分析的过程中采用了简单明了的相关性分析得到大致走势，然后利用线性回归得到部分指标之间的强线性关系。在后期的研究中，可以考虑数据滞后性的滞后性对结果的影响。

参考文献：

[1]中国人民银行货币政策司,2001.央行货币政策执行报告[EB/OL].(2001-1:2019-1).http://www.pbc.gov.cn/.

[2]李凯风,2010.基于Spearman 分析的中国开放式基金业绩持续性实证研究[J].改革与战略,206(26):72-76.

[3]BLEI DM,YNA,JORDAN MI, 2003. Latent Dirichlet Allocation[J].Journal of Machine Learning Research (3):993-1022.

[4]LAS CRUCES N, T.DUNNING,1994.Statistical Identification of Language[J].Computing Research Laboratory, New Mexico State University.

[5]SEGOND F, SCHILLER A, GREFENSTETTE G, et al., 1997. An Experiment in Semantic Tagging Using Hidden Markov Model.