第三届“泰迪杯”

全国大学生数据挖掘竞赛

优

秀

作

品

作品名称：基于数据挖掘技术的市财政收入分析预测模型

荣获奖项：二等奖

作品单位：长沙理工大学

作品成员：龙承运　杨芳　蔡碧碧

指导教师：戴志锋

基于数据挖掘技术的市财政收入分析预测模型

摘　要：本文从经济理论及实践考察出发，确定影响财政收入的关联指标为税收、生产总值、全社会固定资产投资、就业人数和其他收入。经广州市年历年统计年鉴，整理出各项指标数据，并利用多元回归分析数学模型识别出影响财政收入的关键因素为：税收、其他收入、全社会固定资产投资。通过函数拟合计算出 2015 年各关键因素的预测值，从而得出 2015 年财政收入预测值为 23741728 万元。通过各关键因素对财政收入的影响程度并结合广州市近年来财政收支的实际预算情况，从经济和非经济视角对广州市未来几年的财政收支预算提出相关建议。

关键词：经济理论；关键因素；多元回归分析；函数拟合；预测值

Analysis and forecast model of financial revenue based on Data Mining Technology

Abstract: This paper from the economic theory and practice study of determine the impact on fiscal revenue related indicators for tax, GDP, whole society fixed assets investment, employment and other income. After years of Guangzhou City Statistical Yearbook, sorting out the index data, and by using multiple regression analysis mathematical model recognition is the key factor in affecting fiscal revenue: taxes, other income, whole society fixed assets investment. By fitting a function calculated 2015 the key factors of predictive value, so as to obtain the fiscal revenue in 2015 predictive value for 23741728_wan million yuan. Through the key factors on the extent of the impact of fiscal revenue and combined with the Guangzhou City in recent years the actual financial revenue and expenditure budget, from the perspective of economic and non economic of Guangzhou City in the next few years, the fiscal revenue and expenditure budget put forward related suggestions.

Key words: Economic theory; key factors; regression analysis; fitting; prediction

目 录

1. 挖掘目标

本文建模的目标是通过挖掘广州历年统计年鉴中财政收入的关联指标，经过财政收入经济理论研究及实践考察，确定财政收入的影响因素。并通过传统的多元回归数学模型识别影响财政收入的关键因素。为做出下一年有效的财政收入预算，为下一年的政策提供指导依据，本文通过函数拟合预测出各关键因素的预测值，并利用财政收入与各关键因素的相关函数，计算出未来几年财政收入预测值。结合社会经济发展和广州市近几年的财政收入及支出等情况，从财政收入和支出预算的角度，对广州市提供相关合理建议。

2. 分析方法与过程

2.1. 总体流程

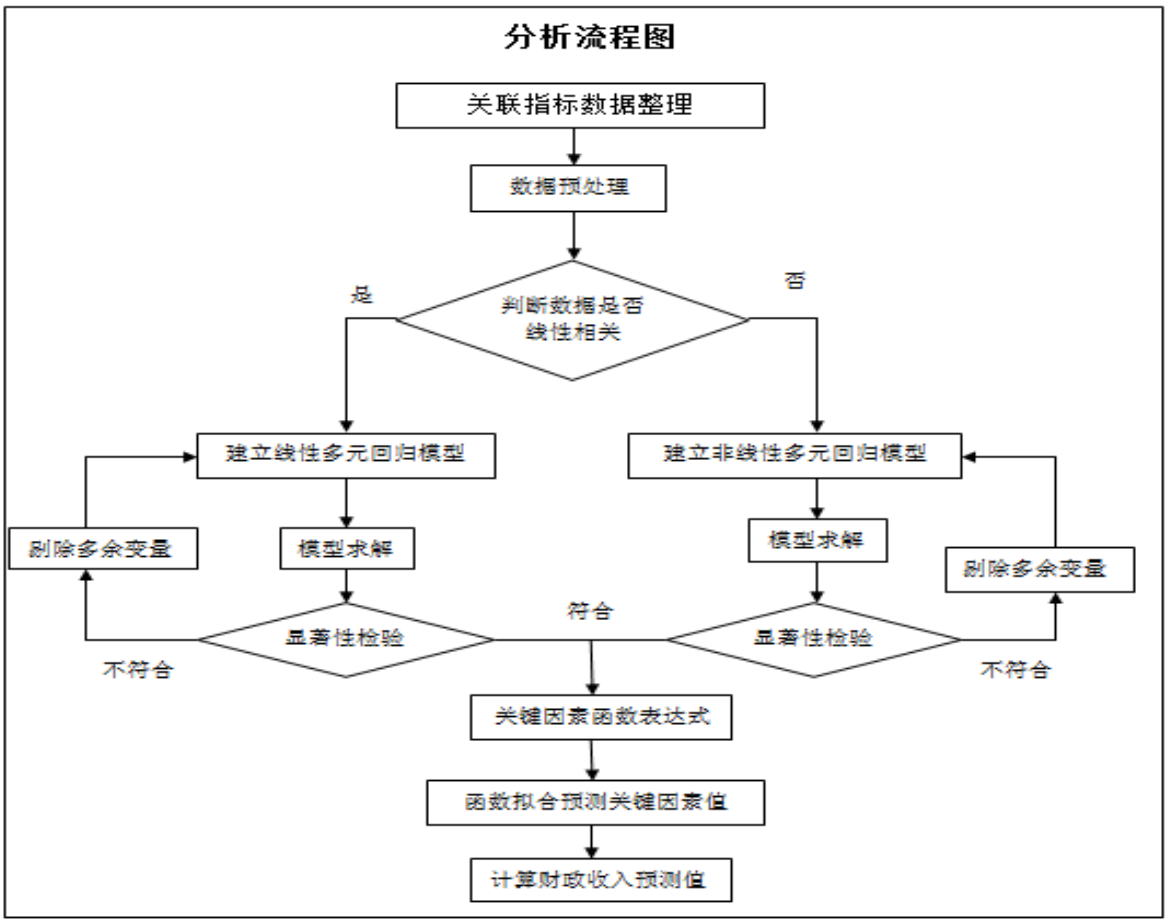

图 1 分析流程图

主要步骤描述：

步骤一：数据预处理，梳理 1999 年—2013 年数据，并对关联指数据标进行分类；

步骤二：判断数据是否线性相关，即判断各指标数据与财政收入数据的相关性，以便建立相应回归模型；

步骤三：建立回归模型，根据多元回归分析原理、步骤，结合财政收入与各指标的关系，建立相应数学模型；

步骤四：模型求解，运用 MATLAB 求出财政收入与各指标的函数表达式；

步骤五：显著性检验，根据回归函数返回的相关系数 R^2 值、$F\_$ 检验值、f 阈值、显著性系数 p 值进行标准检验；

步骤六：剔除多余变量，若假设检验不符合标准，即存在多余变量或异常点，应对其进行剔除；

步骤七：关键因素表达式，根据剔除后的模型求出财政收入与关键因素之间的函数表示；

步骤八：关键因素及财政收入预测，根据多年的数据，可先对各关键因素函数拟合并求预测值，进而运用步骤七得出的函数表达式计算出财政收入。

2.2. 具体步骤

步骤一：确定关联指标

研究财政收入的影响因素离不开一些基本的经济变量。回归变量的选择是建立回归模型的一个极为重要的问题。如果遗漏了某些重要变量，回归方程的效果肯定不会好 而考虑过多的变量，不仅计算量增大许多，而且得到的回归方程稳定性也很差，直接影响到回归方程的应用。影响市财政收入的因素很多，本文通过经济理论对财政收入的解释以及对实践的观察，确定对财政收入影响的因素主要有税收、生产总值、全社会固定资产投资、就业人数和其他收入。

（1）税收。 税收由于具有征收的强制性、无偿性和固定性特点，可以为政府履行其职能提供充足的资金来源。因此，各政府都将其作为政府财政收入的最重要的收入形式和最主要的收入来源。

（2）生产总值。常被公认为衡量区域经济状况的最佳指标。GDP 会促进国民收入，

从而提高居民个人收入水平，直接影响居民储蓄量，并与财政收入的增长保持一定的同向性。

（3）全社会固定资产投资。是建造和购置固定资产的经济活动，即固定资产再生产活动主要通过投资来促进经济增长，扩大税源，进而拉动财政税收收入整体增长。

（4）就业人数。就业人数的上升伴随着居民消费水平的提高，从而间接影响财政收入的增加。

（5）其他收入。结合广州市财政的特点，本文将国有资本经营收入、行政事业性收费、罚没收入、专项收入、政府基金收入等均纳为其他收入。因此，其他收入作为财政收入的组成部分，具有广泛性和不确定性，对财政收入有直接影响。

步骤二：数据预处理

（1） 数据来源

本文数据均以广州统计局提供的《广州统计年鉴》为源，使用 1999—2013 年财政收入、税收、广州市生产总值、全社会固定投资、就业人数和其他收入的数据。确保数据可靠真实。

（2） 数据分类

根据本文对各项指标的定义，通过广州市历年的统计年鉴整理各指标数据见表 1：

表 1 广州市财政收入及关联指标统计数据（单位：万）

指标 / 年份	财政收入 (Y)	税收 (X_1)	生产总值 (X_2)	全社会固定投资 (X_3)	就业人数 (X_4)	其他收入 (X_5)
2013	20881374	8525558	154201434	44545508	7599295	11824379
2012	15796804	7604913	135512072	37583868	7512997	7232049
2011	15351387	7106547	124234390	34122005	7431755	7516336
2010	13991612	6640641	107482828	32635731	7110695	6715166
2009	11076649	5577386	91382135	26598516	6791495	5005569
2007	8389925	4293584	71403223	18633437	6236312	3935683
2006	4767231	3569938	60818614	16963824	5994973	1034604
2005	4088545	3034138	51542283	15191582	5744550	851001
2004	3384477	2490050	44505503	13489283	5407087	767162
2003	3005475	2217300	37586166	11751668	5210706	627067
2002	2719058	2013358	32039616	10092421	5070216	583417
2001	2690984	2137578	28416511	9782093	5029338	546940
2000	2199077	1698857	24927434	9236676	4962579	422507
1999	1881388	1517793	21391758	8782586	4548852	271722

步骤三：多元回归分析

（1） 回归分析基本原理

回归分析是一种处理变量的统计相关关系的一种数理统计方法。其基本思想是：虽然自变量和因变量之间没有严格的、确定的函数关系，但可以设法找出最能代表他们之间关系的数学表达形式。按因变量和自变量的数量对应关系可划分为一个因变量对多个自变量的回归分析及多个因变量对多个自变量的回归分析；按回归模型可划分为线性回归分析和非线性回归分析。

（2） 回归分析基本步骤

1、数据分析

为了确定因变量财政收入 Y 与自变量税收 X_1、生产总值 X_2、全社会固定投资 X_3、就业人数 X_4 和其他收入 X_5 的线性关系，首先利用原始数据作出 Y 对 X_1、X_2、X_3、X_4、X_5 的散点图，如图 2 所示：

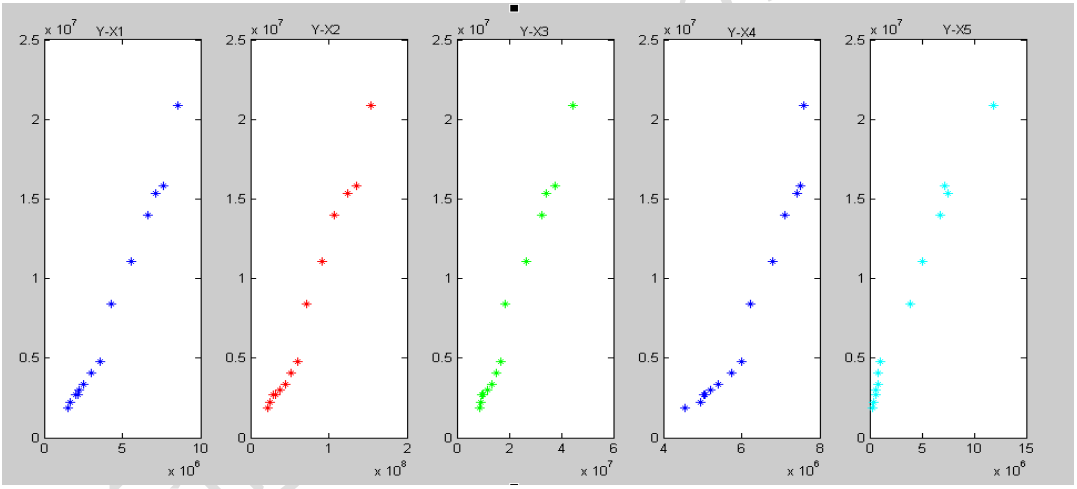

图 2 财政收入与关联指标散点图

从图 2 可以看出，因变量财政收入 Y 与各自变量 X_i 是线性相关的，故应建立一个应变量对多个自变量的线性回归分析模型。

2、建立回归方程

根据传统的回归模型，本文以财政收入 Y 为因变量，税收 X_1、生产总值 X_2、全社会固定投资 X_3、就业人数 X_4 和其他收入 X_5 5 个经济指标为自变量，建立多元函数：

$$Y = C + C_1X_1 + C_2X_2 + C_3X_3 + C_4X_4 + C_5X_5 \qquad \text{式（1）}$$

其中 C 为常数，$C_i(i=1,2,3,4,5)$ 为各指标系数；

利用 MATLAB 软件对上述基本模型进行参数估计得到如下结果：

$$Y = 738237.8445 + 1.3628X_1 - 0.0021X_2 + 0.0206X_3 - 0.2930X_4 + 0.8704X_5 \quad \text{式（2）}$$

其中：$R^2 = 0.9997$，F_检验值=6988.7，阈值 $f = 2.476\times10^{-14}$，显著性 p 值$=1.5027\times10^{10}$

3、回归分析结果检验

第一：异方差检验

可利用 MATLAB 作出各组数据的残差图，见图 3：

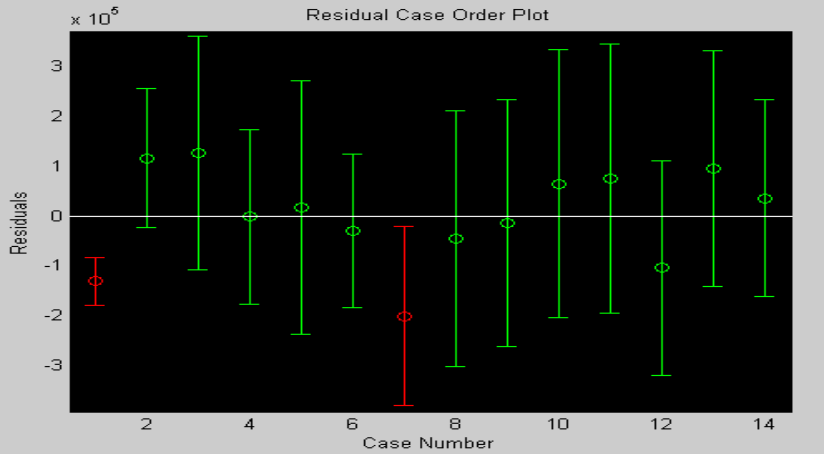

图 3 各组数据残差图

从残差图 3 可以看出，除第一个数据和第七个数据外，其余数据的残差离零点均较近，且从残差的置信区间均包含零点，这说明回归模型结果能较好的符合原始数据，将第二个及第七个数据视为异常点。

第二：指标显著性检验

从 stat 输出的与显著性概率相关的 p 值$=1.5027\times10^{10}>0.05$，这说明回归方程中有些变量可以剔除。

第三：其他检验

在 stat 返回的 4 个值中，$R^2 = 0.9997$，说明模型拟合的很好；，F_检验值=6988.7>阈值 $f = 2.476\times10^{-14}$，符合检验要求。

综上可看出，应对模型进行修正即：去除异常点和剔除不必要指标。

4、模型的修正

首先将数据中的第一个和第七个异常点数据去除，去除后从残差图 4 可见模型分析数据很正常。

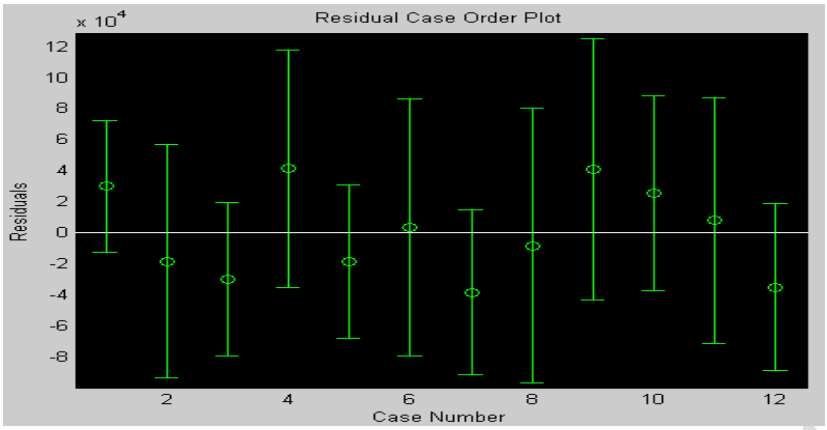

图 4 各组数据残差图

然后对多余经济指标进行剔除，可在 MATLAB 软件包中建一个 M 文件，运用 stepwise 函数进行指标逐个插入，指标选入结果见 5:(蓝色表示选入，红色表示剔除)

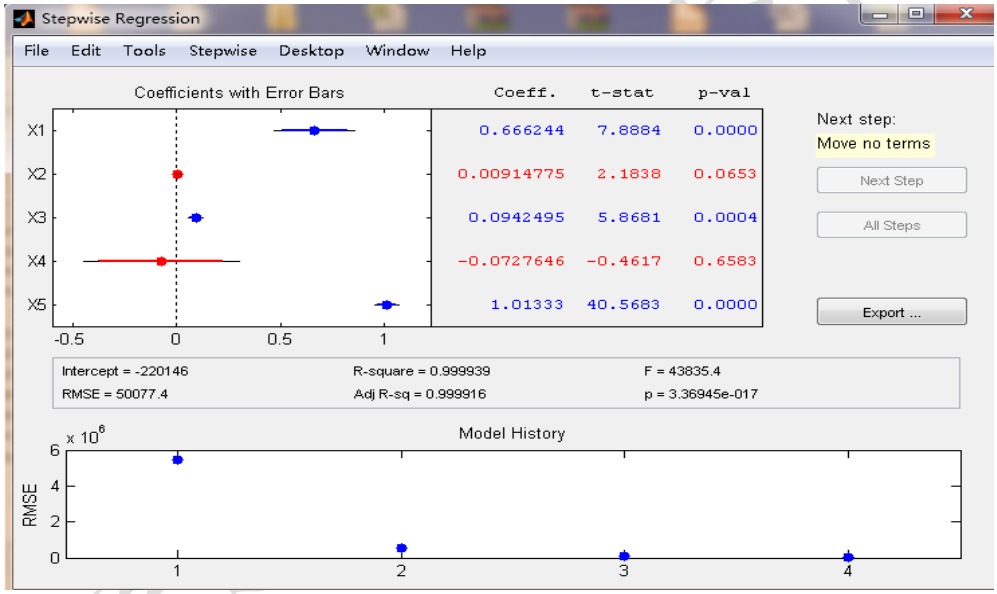

图 5 关键指标选入图

从修正后的模型结果得出：指标显著性参数 $p=3.36945\times10^{-17}<0.05$，符合检验要求。

最后根据修正后的回归模型，建立多元函数关系式：

$$Y=C+C_1X_1+C_2X_3+C_3X_5 \qquad \text{式（3）}$$

用 MATLAB 对该回归模型进行参数估计得到：

$$Y=-40201.49+1.0943X_1+0.0314X_3+0.9049X_5 \qquad \text{式（4）}$$

2.3. 结果分析

从修正后的回归分析模型得出的结果可以看出，自变量财政收入 Y 与因变量税收 X_1、全社会固定投资 X_3、其他收入 X_5 三个经济指标间具有正相关关系，根据各变量系数可知，对财政收入影响最大的税收，其次是其他收入，全社会固定资产投资对财政收入影响相对较弱。

税收作为财政收入的基本因素,其内容非常丰富，它包含广州市 10 个税种的收入（由于农业税在 2006 年取消，故不算之），根据广州市历年统计年鉴各税种数据见附件表（1）。从表中可以看出，税收的主要来源取决于增值税 x_1、营业税 x_2、企业所得税 x_3、城市维护税 x_5，且它们每年收入与税收保持高度正相关，故可利用上面建立的多元回归模型，以税收 y 为因变量，以这四个税种收入 $x_i(i=1\cdots10)$ 为自变量，求得其回归函数为：

$$y=-153152.28+1.5582x_1+1.1173x_2+1.8746x_3+1.1802x_5 \qquad \text{式（5）}$$

故将式（5）代入式（4）可得:

$$Y=-207795.724+1.7051x_1+1.2227x_2+2.0514x_3+1.2915x_5+0.0314X_3+0.9049X_5 \qquad \text{式（6）}$$

故从税收细化来看，增值税 x_1、营业税 x_2、企业所得税 x_3、城市维护税 x_5 是影响财政收入的关键因素。由于其他收入具有广泛性，本文不再对其他收入和全社会固定资产投资进行细化分析。

2.4 函数拟合及预测

结合上述步骤对影响财政收入关键因素的分析，对于预测未来几年财政收入的问题也得到解决。本文先通过对关键因素的预测从而实现对年度财政收入的预测。具体过程如下：

（1） 对各关键影响因素发展趋势进行函数拟合

1、对税收（增值税）函数拟合得到：

$$y=71.584x^5-717868x^4+3\times10^9x^3-6\times10^{12}x^2+6\times10^{15}x-2\times10^{18} \qquad \text{式（7）}$$

其中拟合程度 $R^2=0.9926$，2015 年预测值为：2556425

对税收（营业税）函数拟合得到：

$$y=-160.41x^4+1\times10^6x^3-4\times10^9x^2+5\times10^{12}x-3\times10^{15} \qquad \text{式（8）}$$

其中 $R^2=0.9915$，2015 年预测值为：1794048

对税收（企业所得税）函数拟合得到：

$$y=-172.66x^4+1\times10^6x^3-4\times10^9x^2+6\times10^{12}x-3\times10^{15} \quad \text{式（9）}$$

其中 $R^2=0.97$，2015 年预测值为：1207906

对税收（城市维护税）函数拟合得到：

$$y=28.447\ x^5-285428x^4+1\times10^9x^3-2\times10^{12}x^2+2\times10^{15}x-9\times10^{17} \quad \text{式（10）}$$

其中 $R^2=0.9777$，2015 年预测值为：1232896

2、对其他收入函数拟合得到：

$$y=136.38\times x^6-2\times10^6\times x^5+8\times10^9\times x^4-2\times10^{13}\times x^3+3\times10^{16}\times x^2-3\times10^{19}x+9\times10^{21}$$

式（11）

其中 $R^2=0.9843$，2015 年预测值为：15195370

3、对全社会固定资产投资函数拟合得到：

$$y=187117x^2-7\times10^8x+7\times10^{11} \quad \text{式（12）}$$

其中 $R^2=0.9945$，2015 年度预测值为：52953324

4、将各预测值代入到式（4）可得 2015 年广州市财政收入为 23741728 万元

3. 结论

本文通过多元回归分析方法来识别影响广州市财政收入的关键因素，得出税收、其他收入、全社会固定投资为影响广州市财政收入的关键因素，并通过函数拟合先对各关键因素进行预测，得出广州市 2015 年税收（增值税）预测值为 2556425 万元、税收（营业税）预测值为 1794048 万元、税收（企业所得税）预测值为 1207906 万元、税收（城市维护税）预测值为 1232896 万元、其他收入预测值为 15195370 万元、全社会固定资产投资预测值为 52953324 万元、财政收入为 23741728 万元。

根据本文得出的各指标预测值，结合社会经济发展和广州市近几年的财政收入及支出情况，考虑经济因素和非经济因素的影响，对广州市的财政收支提出以下几点建议：

（1）保持税收收入占 GDP 的合理比例，提高税收征管执行力度。税收是广州市财政收入的主要来源，且税收水平对财政收入的影响很大，因此税务机关要加强税务源监控，强化税务知识培训，完善税务法律体系，转变工作作风，依法治税、依法征税，

通过加强各部门配合，不断提高税收征收率，保持税收随经济的发展平稳增长。

（2）鼓励非公有制经济的发展，增加广州市固定资产的投资，带动财政收入增长。个体、私营等非公有制经济是生产力发展的重要力量，尽管国有经济目前在整体上仍然是政府最大的税收来源，但非公有制经济直接和间接创造的税收将占越来越大的比重，并且积极鼓励发展的行业实现“消费型”增值税，调整消费税、营业税，增加居民消费能力，提高税收对财政的贡献率。

（3）转变招商引资模式，提升招商引资质量，重点引进高附值、税收贡献大的产业，围绕现代服务业、高新技术、新能源等领域加大招商引资力度。

4. 参考文献

[1] 纪跃芝.影响财政收入增长的相关因素分析[J].统计与决策,2009.19.

[2] 白萍.影响我国财政收入的多元线性回归模型[J].统计与决策,2005.10.

[3] 樊孝菊.地方财政一般预算收入主要影响因素的实证分析[J].科技创业月刊,2008.10.

[4] 周忠辉，丁建勋，王丽丽.我国财政收入影响因素的实证研究[J].当代经济，2011（4）：84-85.

[5] 高铁梅.计量经济分析方法与建模 EVIEWS 应用及实例[M].北京：清华大学出版社，2006.