财政收入预测分析

张敏

目录

分析财政收入预测背景

1. 财政收入简介和需求

- 财政收入，是指政府为履行其职能、实施公共政策和提供公共物品与服务需要而筹集的一切资金的总和。
- 在我国现行的分税制财政管理体制下，地方财政收入不但是国家财政收入的重要组成部分，而且具有其相对独立的构成内容。

分析财政收入预测背景

2. 财政收入预测数据基础情况

考虑到数据的可得性，本项目所用的财政收入分为地方一般预算收入和政府性基金收入。地方一般预算收入包括以下2个部分。

- 税收收入。
- 非税收收入。

2. 财政收入预测数据基础情况

对1994年至2013年的数据进行分析（本项目所用数据均来自《统计年鉴》）。

各项特征名称及特征说明如下（共13项）：

- **社会从业人数(x1)**：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。
- **在岗职工工资总额(x2)**：反映的是社会分配情况，主要影响财政收入中的个人所得税、房产税以及潜在消费能力。

分析财政收入预测背景

2. 财政收入预测数据基础情况

- **社会消费品零售总额(x3)**：代表社会整体消费情况，是可支配收入在经济生活中的实现。当社会消费品零售总额增长时，表明社会消费意愿强烈，部分程度上会导致财政收入中增值税的增长；同时当消费增长时，也会引起经济系统中其他方面发生变动，最终导致财政收入的增长。
- **城镇居民人均可支配收入(x4)**：居民收入越高消费能力越强，同时意味着其工作积极性越高，创造出的财富越多，从而能带来财政收入的更快和持续增长。
- **城镇居民人均消费性支出(x5)**：居民在消费商品的过程中会产生各种税费，税费又是调节生产规模的手段之一。在商品经济发达的如今，居民消费的越多，对财政收入的贡献就越大。
- **年末总人口(x6)**：在地方经济发展水平既定的条件下，人均地方财政收入与地方人口数呈反比例变化。

分析财政收入预测背景

2. 财政收入预测数据基础情况

- **全社会固定资产投资额(x7)**：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。
- **地区生产总值(x8)**：表示地方经济发展水平。一般来讲，政府财政收入来源于即期的地区生产总值。在国家经济政策不变、社会秩序稳定的情况下，地方经济发展水平与地方财政收入之间存在着密切的相关性，越是经济发达的地区，其财政收入的规模就越大。
- **第一产业产值(x9)**：取消农业税、实施三农政策，第一产业对财政收入的影响更小。
- **税收(x10)**：由于其具有征收的强制性、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，各国都将其作为政府财政收入的最重要的收入形式和来源。

分析财政收入预测背景

2. 财政收入预测数据基础情况

- **居民消费价格指数(x11)**：反映居民家庭购买的消费品及服务价格水平的变动情况，影响城乡居民的生活支出和国家的财政收入。
- **第三产业与第二产业产值比(x12)**：表示产业结构。三次产业生产总值代表国民经济水平，是财政收入的主要影响因素，当产业结构逐步优化时，财政收入也会随之增加。
- **居民消费水平(x13)**：在很大程度上受整体经济状况GDP的影响，从而间接影响地方财政收入。

3. 财政收入预测分析目标

结合财政收入预测的需求分析，本次数据分析建模目标主要有以下2个。

- 分析、识别影响地方财政收入的关键特征。
- 预测2014年和2015年的财政收入。

了解财政收入预测的方法

方法选择——最小二乘估计方法

- 众多学者已经对财政收入的影响因素进行了研究，但是他们大多先建立财政收入与各待定的影响因素之间的多元线性回归模型，运用**最小二乘估计方法**来估计回归模型的系数，通过系数来检验它们之间的关系，模型的结果对数据的依赖程度很大，并且普通最小二乘估计求得的解往往是局部最优解，后续步骤的检验可能就会失去应有的意义。

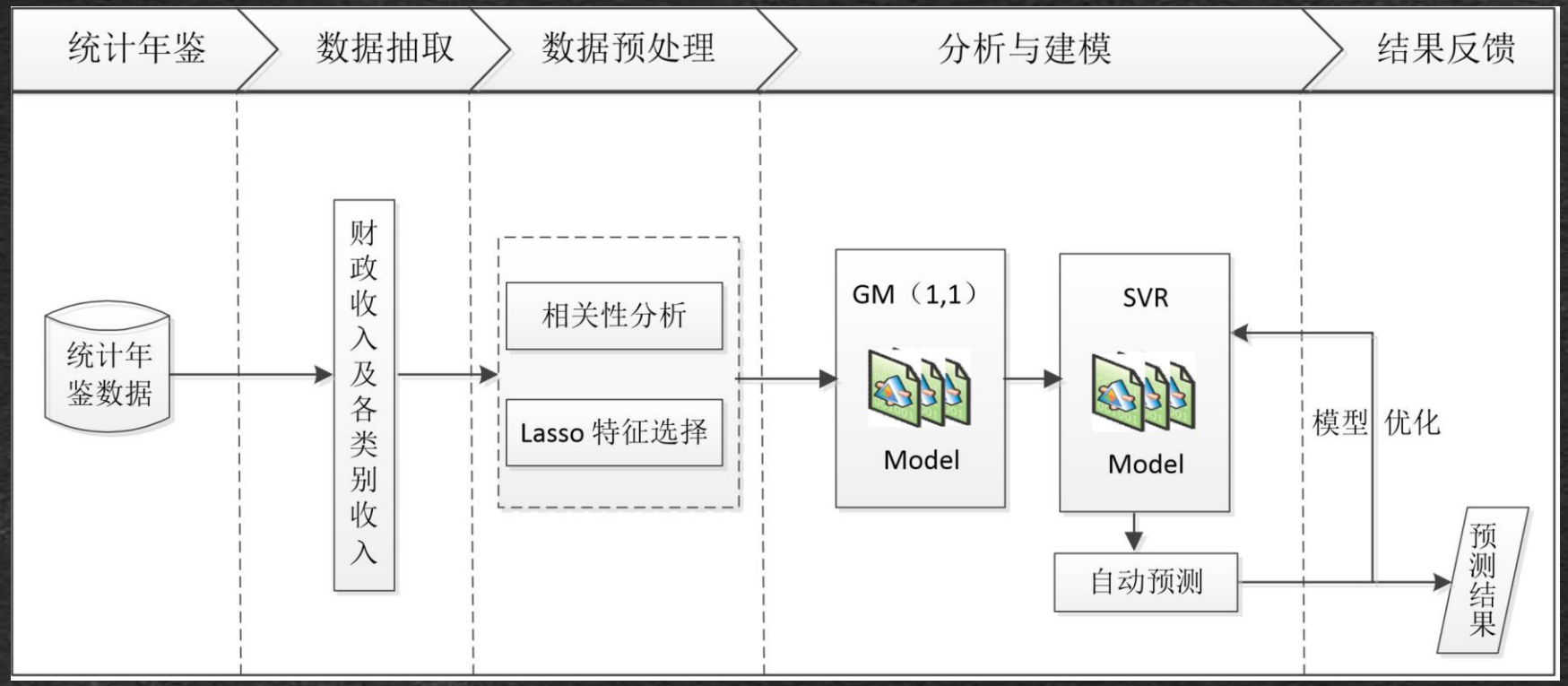
了解财政收入预测的方法

方法选择——Lasso特征选择方法

- 本项目在已有研究的基础上运用**Lasso特征选择方法**来研究影响地方财政收入的因素。在Lasso特征选择的基础上，鉴于灰色预测对少量数据预测的优良性能，对单个选定的影响因素建立灰色预测模型，得到它们在2014年及2015年的预测值。由于支持向量回归较强的适用性和容错能力，对历史数据建立训练模型，把灰色预测的数据结果代入训练完成的模型中，充分考虑历史数据信息，可以得到较为准确的预测结果，即2014年和2015年财政收入。

熟悉财政收入预测的步骤与流程

项目流程



熟悉财政收入预测的步骤与流程

项目流程

本项目的总体流程主要包括以下步骤。

1. 对原始数据进行探索性分析，了解原始特征之间的相关性。
2. 利用Lasso特征选择模型进行特征提取。
3. 建立单个特征的灰色预测模型以及支持向量回归预测模型。
4. 使用支持向量回归预测模型得出2014-2015年财政收入的预测值。
5. 对上述建立的财政收入预测模型进行评价。

目录

Pearson相关系数

相关性分析是指对两个或多个具备相关性的特征元素进行分析，从而衡量两个特征因素的相关密切程度。

在统计学中，常用Pearson相关系数来进行相关性分析。

Pearson相关系数是用来度量两个特征X和Y之间的相互关系（线性相关的强弱），是最简单的一种相关系数，通常用r或ρ表示，取值范围在[-1,1]之间。

Pearson相关系数的一个关键的特性就是它不会随着特征的位置或是大小的变化而变化。例如，把X变为$a+bX$，把Y变为$c+dY$，其中a，b，c，d都是常数，不会改变相互之间的相关系数。

Pearson相关系数

若两个向量$\boldsymbol{X}=[x_1,x_2,\ldots,x_n]$，$\boldsymbol{Y}=[y_1,y_2,\ldots,y_n]$，则它们之间的Pearson相关系数如下。

$$r=\frac{n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{\sqrt{n\sum_{i=1}^{n}x_i{}^2-\left(\sum_{i=1}^{n}x_i\right)^2}\sqrt{n\sum_{i=1}^{n}y_i{}^2-\left(\sum_{i=1}^{n}y_i\right)^2}}$$

当$0<r<1$时，表示$\boldsymbol{X}$和$\boldsymbol{Y}$呈现正相关关系；当$-1<r<0$时，表示$\boldsymbol{X}$和$\boldsymbol{Y}$呈现负相关关系。若$r=0$，表示$\boldsymbol{X}$和$\boldsymbol{Y}$不相关；若$r=1$，表示$\boldsymbol{X}$和$\boldsymbol{Y}$完全正相关；若$r=-1$，表示$\boldsymbol{X}$和$\boldsymbol{Y}$完全负正相关。$|r|$越接近1，说明$\boldsymbol{X}$和$\boldsymbol{Y}$差距越小，相关性越大。

分析计算结果

Pearson相关系数矩阵

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	y
x1	1.00	0.95	0.95	0.97	0.97	0.99	0.95	0.97	0.98	0.98	-0.29	0.94	0.96	0.94
x2	0.95	1.00	1.00	0.99	0.99	0.92	0.99	0.99	0.98	0.98	-0.13	0.89	1.00	0.98
x3	0.95	1.00	1.00	0.99	0.99	0.92	1.00	0.99	0.98	0.99	-0.15	0.89	1.00	0.99
x4	0.97	0.99	0.99	1.00	1.00	0.95	0.99	1.00	0.99	1.00	-0.19	0.91	1.00	0.99
x5	0.97	0.99	0.99	1.00	1.00	0.95	0.99	1.00	0.99	1.00	-0.18	0.90	0.99	0.99
x6	0.99	0.92	0.92	0.95	0.95	1.00	0.93	0.95	0.97	0.96	-0.34	0.95	0.94	0.91
x7	0.95	0.99	1.00	0.99	0.99	0.93	1.00	0.99	0.98	0.99	-0.15	0.89	1.00	0.99
x8	0.97	0.99	0.99	1.00	1.00	0.95	0.99	1.00	0.99	1.00	-0.15	0.90	1.00	0.99
x9	0.98	0.98	0.98	0.99	0.99	0.97	0.98	0.99	1.00	0.99	-0.23	0.91	0.99	0.98
x10	0.98	0.98	0.99	1.00	1.00	0.96	0.99	1.00	0.99	1.00	-0.17	0.90	0.99	0.99
x11	-0.29	-0.13	-0.15	-0.19	-0.18	-0.34	-0.15	-0.15	-0.23	-0.17	1.00	-0.43	-0.16	-0.12
x12	0.94	0.89	0.89	0.91	0.90	0.95	0.89	0.90	0.91	0.90	-0.43	1.00	0.90	0.87
x13	0.96	1.00	1.00	1.00	0.99	0.94	1.00	1.00	0.99	0.99	-0.16	0.90	1.00	0.99
y	0.94	0.98	0.99	0.99	0.99	0.91	0.99	0.99	0.98	0.99	-0.12	0.87	0.99	1.00

分析计算结果

分析

- 由上表可知，居民消费价格指数（x11）与财政收入（y）的线性关系不显著，呈现负相关。其余特征均与财政收入呈现高度的正相关关系。
  - 按相关性大小，依次是x3，x4，x5，x7，x8，x10，x13，x2，x9，x1，x6和x12。
- 各特征之间存在着严重的多重共线性：
  - 特征x1，x4，x5，x6，x8，x9，x10与除了x11之外的特征均存在严重的共线性。
  - 特征x2，x3，x7与除了x11和x12外的其他特征存在着严重的多重共线性。
  - x11与各特征的共线性不明显。
  - x12与除了x2，x3，x7，x11之外的其他特征有严重的共线性。
  - x13与除了x11之外的各特征有严重的共线性。
  - x2和x3，x2和x13，x3和x13等多对特征之间存在完全的共线性。
- 由上述分析可知，选取的各特征除了x11外，其他特征与y的相关性很强，可以用作财政收入预测分析的关键特征，但这些特征之间存在着信息的重复，需要对特征进行进一步筛选。

目录

了解Lasso回归方法

1. 概念

- Lasso回归方法属于正则化方法的一种，是压缩估计。
- 它通过构造一个惩罚函数得到一个较为精炼的模型，使得它压缩一些系数，同时设定一些系数为零，保留了子集收缩的优点，是一种处理具有复共线性数据的有偏估计。

2. 基本原理

- Lasso以缩小特征集（降阶）为思想，是一种收缩估计方法。
- Lasso方法可以将特征的系数进行压缩并使某些回归系数变为0，进而达到特征选择的目的，可以广泛地应用于模型改进与选择。
- 通过选择惩罚函数，借用Lasso思想和方法实现特征选择的目的。这种过程可以通过优化一个“损失” + “惩罚”的函数问题来完成。

2. 基本原理

Lasso参数估计被定义如下。

$$\hat{\beta}(lasso) = \underset{\beta}{\operatorname{argmin}}^2 \left\| y - \sum_{j=1}^{p} x_i \beta_i \right\|^2 + \lambda \sum_{j=1}^{p} |\beta_i|$$

其中，λ为非负正则参数，控制着模型的复杂程度，λ越大对特征较多的线性模型的惩罚力度就越大，从而最终获得一个特征较少的模型，$\lambda \sum_{j=1}^{p} |\beta_j|$称为惩罚项。调整参数λ的确定可以采用交叉验证法，选取交叉验证误差最小的λ值。最后，按照得到的λ值，用全部数据重新拟合模型即可。

3. 适用场景

- 当原始特征中存在多重共线性时，Lasso回归不失为一种很好的处理共线性的方法，它可以有效地对存在多重共线性的特征进行筛选。
- 在机器学习中，面对海量的数据，首先想到的就是降维，争取用尽可能少的数据解决问题，从这层意义上说，用Lasso模型进行特征选择也是一种有效的降维方法。
- Lasso从理论上说，对数据类型没有太多限制，可以接受任何类型的数据，而且一般不需要对特征进行标准化处理。

4. Lasso回归方法优缺点

- **优点**：可以弥补最小二乘法和逐步回归局部最优估计的不足，可以很好地进行特征的选择，可以有效地解决各特征之间存在多重共线性的问题。
- **缺点**：如果存在一组高度相关的特征时，Lasso回归方法倾向于选择其中的一个特征，而忽视其他所有的特征，这种情况会导致结果的不稳定性。

虽然Lasso回归方法存在弊端，但是在合适的场景中还是可以发挥不错的效果。在财政收入预测中，各原始特征存在着严重的多重共线性，多重共线性问题已成为主要问题，这里采用Lasso回归方法进行特征选取是恰当的。

分析Lasso回归结果

分析系数表

用Python编制相应的程序后运行得到如下表所示的结果。

x1	x2	x3	x4	x5	x6	x7
-0.0001	0.000	0.124	-0.010	0.065	0.000	0.317
x8	x9	x10	x11	x12	x13	
0.035	-0.001	0.000	0.000	0.000	-0.040	

由上表可看出，利用Lasso回归方法识别影响财政收入的关键影响因素是社会从业人数（x1）、社会消费品零售总额（x3）、城镇居民人均可支配收入（x4）、城镇居民人均消费性支出（x5）、全社会固定资产投资额（x7）、地区生产总值（x8）、第一产业产值（x9）和居民消费水平（x13）。

目录

了解灰色预测算法

1. 概念

- 灰色预测法是一种对含有不确定因素的系统进行预测的方法。
- 在建立灰色预测模型之前，需先对原始时间序列进行数据处理，经过数据处理后的时间序列即称为生成列。
- 灰色系统常用的数据处理方式有累加和累减两种。

2. 基本原理

灰色预测是以灰色模型为基础的，在众多的灰色模型中，GM(1，1)模型最为常用。

设特征$\boldsymbol{X}^{(0)} = \{\boldsymbol{X}^{(0)}(i), i = 1,2,\ldots,n\}$为一非负单调原始数据序列，建立灰色预测模型：

- 首先对$X^{(0)}$进行一次累加得到一次累加序列$\boldsymbol{X}^{(1)} = \{\boldsymbol{X}^{(1)}(k), k = 1,2,\ldots,n\}$。
- 对$X^{(1)}$可建立下述一阶线性微分方程，如下，即GM(1,1)模型。

$$\frac{d\boldsymbol{X}^{(1)}}{\mathrm{dt}} + a\boldsymbol{X}^{(1)} = \mu$$

- 求解微分方程，得到预测模型如下。

$$\hat{\boldsymbol{X}}^{(1)}(k+1) = \left[\hat{\boldsymbol{X}}^{(1)}(0) - \frac{\hat{\mathrm{u}}}{\hat{\mathrm{a}}}\right] \mathrm{e}^{-\hat{\mathrm{a}}\mathrm{k}} + \frac{\hat{\mathrm{u}}}{\hat{\mathrm{a}}}$$

- 由于GM(1,1)模型得到的是一次累加量，将GM(1,1)模型所得数据$\hat{\boldsymbol{X}}^{(1)}(\mathrm{k}+1)$经过累减还原为$\hat{\boldsymbol{X}}^{(0)}(k+1)$，即$\boldsymbol{X}^{(0)}$的灰色预测模型如下。

$$\hat{\boldsymbol{X}}^{(0)}(k+1) = \left(e^{-\hat{\mathrm{a}}} - 1\right)\left[\boldsymbol{X}^{(0)}(n) - \frac{\hat{\mathrm{u}}}{\hat{\mathrm{a}}}\right] e^{-\hat{\mathrm{a}}k}$$

2. 基本原理

- 后验差检验模型精度如下表所示。

P	C	模型精度
>0.95	<0.35	好
>0.80	<0.5	合格
>0.70	<0.65	勉强合格
<0.70	>0.65	不合格

3. 适用场景

- 灰色预测法的通用性比较强些，一般的时间序列场合都可以用，尤其适合那些规律性差且不清楚数据产生机理的情况。

4. 灰色预测优缺点

- **优点**：具有预测精度高、模型可检验、参数估计方法简单、对小数据集有很好的预测效果。
- **缺点**：对原始数据序列的光滑度要求很高，在原始数据列光滑性较差的情况下灰色预测模型的预测精度不高甚至通不过检验，结果只能放弃使用灰色模型进行预测。

了解SVR算法

1.基本原理

SVR（Support Vector Regression，支持向量回归）是在做拟合时，采用了支持向量的思想，来对数据进行回归分析。给定训练数据集$\mathbf{T} = \{(\vec{x}_1, y_1), (\vec{x}_2, y_2), \dots, (\vec{x}_n, y_n)\}$，其中$\vec{x}_1 = \left(x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(n)}\right)^{\mathrm{T}} \in \mathbb{R}^n$，$y_i \in \mathbb{R}$，$i = 1,2,\dots,n$。对于样本$(\vec{x}_i, y_i)$通常根据模型输出$f(\vec{x}_i)$与真实值$y_i$之间的差别来计算损失，当且仅当$f(\vec{x}_i) = y_i$时损失才为零。

SVR的基本思路是：允许$f(\vec{x}_i)$与y_i之间最多有ϵ的偏差。仅当$|f(\vec{x}_i) = y_i| > \epsilon$时，才计算损失。当$|f(\vec{x}_i) = y_i| \leq \epsilon$时，认为预测准确。

2. 适用场景

- 由于支持向量机拥有完善的理论基础和良好的特性，人们对其进行了广泛的研究和应用，涉及分类、回归、聚类、时间序列分析、异常点检测等诸多方面。
- 具体的研究内容包括统计学习理论基础、各种模型的建立、相应优化算法的改进以及实际应用。
- 支持向量回归也在这些研究中得到了发展和逐步完善，已有许多富有成果的研究工作。

3. SVR算法优缺点

- **优点**：支持向量回归不仅适用于线性模型，对于数据和特征之间的非线性关系也能很好抓住；支持向量回归不需要担心多重共线性问题，可以避免局部极小化问题，提高泛化性能，解决高维问题；支持向量回归虽然不会在过程中直接排除异常点，但会使得由异常点引起的偏差更小。
- **缺点**：计算复杂度高，在面临数据量大的时候，计算耗时长。

4. 主要参数介绍

- sklearn库的LinearSVR函数实现了线性支持向量回归，其使用语法如下。

*class sklearn.svm.**LinearSVR**(epsilon=0.0, tol=0.0001, C=1.0, loss='epsilon_insensitive'...)*

- *常用参数及说明如下。*

参数名称	说明
epsilon	接收float。用于loss参数中的 参数。默认为0.1。
tol	接收float。指定终止迭代的阈值。默认为0.0001。
C	接收float。表示罚项系数。默认为1.0。
loss	接收string。表示损失函数，有两个选项。默认为epsilon_insensitive。 1.epsilon_insensitive：此时损失函数为L_ϵ（标准的SVR）。 2.squared_epsilon_insensitive：此时损失函数为L_ϵ^2。

续表

参数名称	说明
fit_intercept	接收boolean。表示是否计算模型的截距。默认为True。
intercept_scaling	接收float。如果提供了，则实现X变成向量[X, intercept_scaling]。此时相当于添加了一个人工特征。该特征对所有实例都是常数值。默认为1。 1.此时截距变成：intercept_scaling *人工特征的权重ω_s。 2.此时人工特征也参与了罚项的计算。
dual	接收boolear。选择解决对偶问题或原始问题。如果为True，则解决对偶问题；如果是False，则解决原始问题。默认为True。
verbose	接收int。表示是否开启verbose输出。默认为0。
random_state	输入int，或者一个RandomState实例，或者None。表示使用的随机数生成器的种子。默认为None。 1.如果为整数，则它指定随机数生成器的种子。 2.如果为RandomState实例，则指定随机数生成器。 3.如果为None，则使用默认的随机数生成器。
max_iter	接收int。指定最大迭代次数。默认为1000。

4. 主要参数介绍

- 使用sklearn构建的SVR模型属性及其说明如下表所示。

属性名称	说明
coef_	返回array。给出各个特征的权重。
intercept_	返回array。给出截距，即决策函数中的常数项。

分析预测结果

通过灰色预测模型得出的预测值

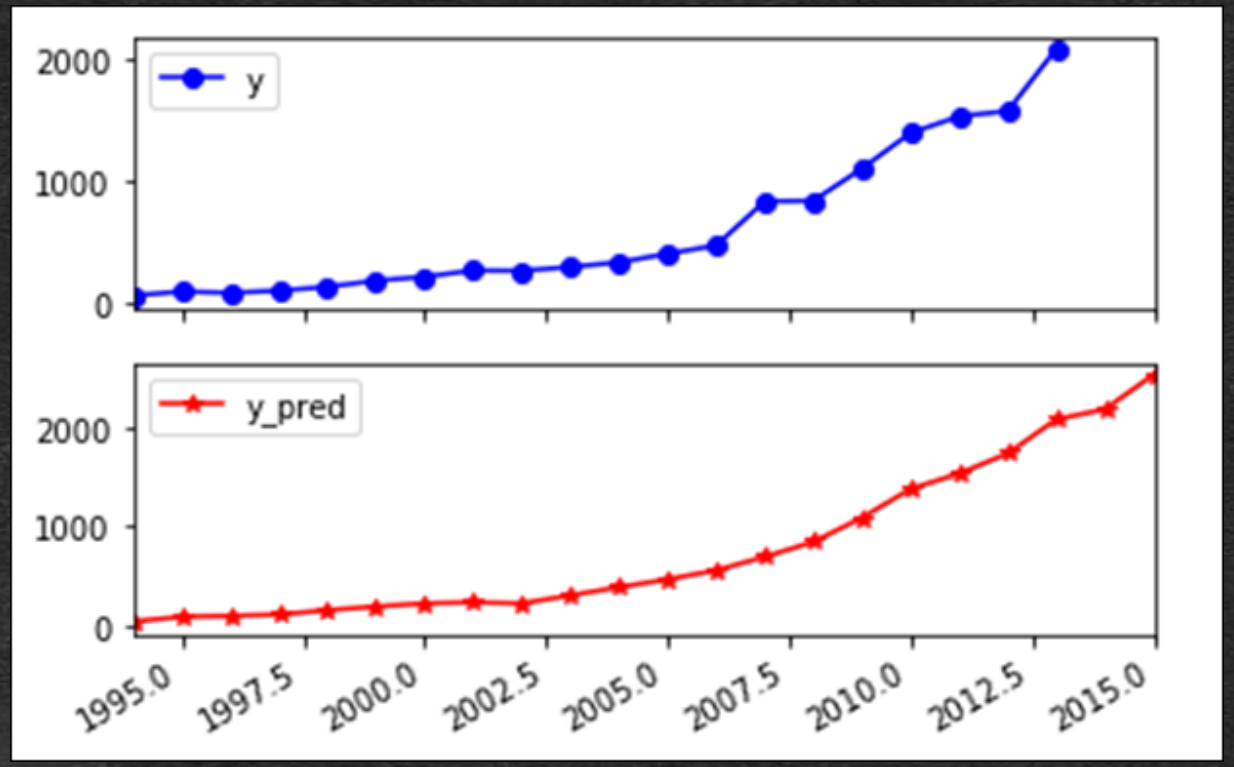
- 社会从业人数（x1）、社会消费品零售总额（x3）、城镇居民人均可支配收入（x4）、城镇居民人均消费性支出（x5）、全社会固定资产投资额（x6）、地区生产总值（x7）、第一产业产值（x8）和居民消费水平（x13）特征的2014年及2015年通过建立的灰色预测模型得出的预测值，如下表所示。

	2014预测值	2015预测值	预测精度等级
x1	8142148.2	8460489.3	好
x3	7042.31	8166.92	好
x4	43611.84	47792.22	好
x5	35046.63	38384.22	好
x6	8505523	8627139	好
x7	4600.4	5214.78	好
x8	18686.28	21474.47	好
x13	44506.47	49945.88	好

分析预测结果

预测结果对比

- 将上表的预测结果代入地方财政收入建立的支持向量回归预测模型，得到1994年至2015年财政收入的预测值，其中Y_pred表示预测值。



分析预测结果

模型评价指标

- 采用回归模型评价指标对地方财政收入的预测值进行评价，得到的结果如下表所示。

平均绝对误差	34.2036806008
中值绝对误差	17.4157390837
可解释方差值	0.990889695375
R方值	0.990878079078

- 可以看出平均绝对误差与中值绝对误差较小，可解释方差值与R方值十分接近1，表明建立的支持向量回归模型拟合效果优良，可以用于预测财政收入。

目录

小结

本项目财政收入预测，主要介绍了原始数据的**相关性分析**、**特征的选取**、**构建灰色预测**和**支持向量回归预测模型**、**模型的评价**四部分内容。

- 在财政收入相关数据的**相关性分析**中，采用了简单相关系数对数据进行了分析。
- 在**特征选取**中，运用了广泛使用的Lasso回归模型。
- 在**模型的构建**阶段，针对历史数据首先构建了**灰色预测模型**，对所选特征的2014年与2015年的值进行预测。然后根据所选特征的原始数据与预测值，建立**支持向量回归模型**，得到财政收入的最终预测值。
- 最后用平均绝对误差、中值绝对误差、可解释方差值和R方值进行**模型的评价**。

本项目建立的财政收入预测模型，通过图 8-2可以看出，很好的拟合了财政收入的变化情况。同时，模型还具有很高的预测精度，可以用来指导实际的工作。

Thank you!