Advances in Applied Mathematics 应用数学进展, 2022, 11(8), 5880-5891
Published Online August 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.118620

Hans汉斯

基于线性模型方法的中国财政收入分析

董凤芝

云南财经大学，云南 昆明

收稿日期：2022年7月22日；录用日期：2022年8月15日；发布日期：2022年8月24日

摘　要

基于《中国统计年鉴》及海关总署内所收集的数据集，基于线性模型分析理论，分析研究了我国国内生产总值、税收收入、进出口贸易总金额、经济活动人口数量四个影响我国财政收入的影响，使用线性回归和逐步回归的方法建立统计模型，最后得出总的回归方程。并对得到的模型方程和参数进行相关假设检验，进行参数估计，通过结果分析，得出了影响我国财政税收的2个主要指标：税收、进出口贸易总额。最后结合所建立的模型进行模型诊断与预测。

关键词

线性模型，假设检验，参数估计，模型诊断

Analysis of China's Fiscal Revenue Based on Linear Model Method

Fengzhi Dong

Yunnan University of Finance and Economics, Kunming Yunnan

Received: Jul. 22nd, 2022; accepted: Aug. 15th, 2022; published: Aug. 24th, 2022

Abstract

Based on the data set collected in the China Statistical Yearbook and the General Administration of Customs, based on the linear model analysis theory, the four influences of China's GROSS DOMESTIC PRODUCT, tax revenue, total amount of import and export trade, and the number of economically active people in China are analyzed and studied, and the statistical model is established by linear regression and stepwise regression methods, and finally the total regression equation is obtained. The relevant hypothesis tests of the obtained model equations and parameters are carried out, the parameters are estimated, and through the analysis of the results, two main indicators affecting

文章引用: 董凤芝. 基于线性模型方法的中国财政收入分析[J]. 应用数学进展, 2022, 11(8): 5880-5891.
DOI: 10.12677/aam.2022.118620

China's fiscal taxation are obtained: taxation and total import and export trade. Finally, the model diagnosis and prediction are carried out in combination with the established model.

Keywords

Linear Models, Hypothesis Testing, Parameter Estimation, Model Diagnostics

1. 简介

一个国家的财政的收入是是指该国国家政府凭借政府的特殊权利，按照有关的法律和法规在一定时期内(一般为一年)取得的各种形式收入的总和，其中包括税收、企事业收入、国家能源交通重点建设基金收入、债务收入、规费收入及罚没收入等。财政的收入水平能反应一个国家的经济实力。在本篇文章中我们使用相关的统计分析方法分析了中国从改革开放到 2019 年年底以来的响应变量财政收入与自变量国家税收、国内的生产总值、进出口贸易总额及经济活动人口之间的相关关系，通过建立模型，找出影响这四个指标中影响中国财政收入的主要因素并进行预测，根据模型的最终形式确立，提出自己的一点建议。

在文献[1]中，罗博炜等人以多元线性回归统计模型为基础，对美国部分地区房价数据进行建模预测，进而探究提高多元回归线性模型精度的方法。

在文献[2]中，王松桂等人通过引进各种线性模型实例，使读者对模型背景有了了解之后，讨论了多元正态及其有关的分布，文章随之又继续讨论线性模型的统计推断的基本理论方法。

在文献[3]中，孙毅等人的基于多元线性回归模型的考试成绩评价并进行相关预测。

在文献[4]中，刘凯通过对广义潜变量模型、EM 算法、因子分析模型做相关理论知识讲解，最后通过深圳某校高一、高二学生的六门成绩进行实际演练。其中将六门成绩分别做了潜变量处理，分别被是文科因子和理科因子的影响进行广义潜变量模型建模分析。并用广义潜变量的方法和以往的因子分析方法做了相应的比较。

本文首先阐述了传统的线性模型的基本理论，用这种模型对数据集建立模型，分析影响我国财政收入的主要因素，对模型进行诊断并进行预测。

2. 研究目的和意义

当前，我国现在还处于并将长期处于社会主义发展阶段，要实现中华民族的伟大复兴史，经济发展这一块必然是国家的重中之重。中国现在在世界的大熔炉中发展的非常好，要想继续观察中国经济发展的走向，研究国家的经济走向是必不可少的。国家的财政收入也是国家经济实力体现的一个重点对象，然而，财政收入又受到很多影响因素的影响，本文着重研究国内生产总值、税收、进出口贸易总额及经济活动人口四个指标对国家财政收入的影响。

中国的财政收入同其他国家相比，国家财政收入占国内生产总值比重偏低，国家掌握的财政收入偏少，这是这一时期我国面临的备受关注的方面。因此找出影响财政收入的主要因素，建立有效、可靠的预测模型，将会对稳定发展中国的财政收入，维护中国在世界的经济地位有重要意义。

本文通过建立模型，找出影响财政收入的主要因素，线性模型是一种应用最广泛的处理多元数据的统计模型，这种模型可以推广到非常大的数据集，对稀疏数据也很有效，而且该模型训练速度非常快，预测速度也很快。基于其具有操作简便，解释性好等特点，本文基于此对影响我国财政收入的五个变量进行线性拟合，对数据进行合理解释。文章希望可以在国家经济长久健康稳定发展的基础上对相应的主要因素进行调整提出自己的建议。

3. 数据介绍

本文是对来自中国统计出版社出版的《中国统计年鉴》及海关总署(以 2019 年的经济活动人口为预算值)的数据分析，与中国 1978~2019 年共 42 年的财政收入有关，一共有 42 条记录，共有 6 个指标，六个指标都是连续型变量。其中相应变量的具体取值如下表 1 所示。

Table 1. The specific meaning and range of values of the variables
表 1. 变量的具体含义及取值范围

变量类型	变量名	详细说明	取值范围	备注
因变量	y (Fiscal revenue) 财政收入	政府部门在一定时期内所取得的货币收入	11.3226~1903.9008	定量变量 (单位：百亿元)
自变量	X_1 (GDP) 国内生产总值	国家所有常住单位在一定时期内的生产活动的最终成果	36.787~9908.661	定量变量 (单位：百亿元)
	X_2 (Revenue) 税收	国家依法向企业和个人等课税对象征收的款项或实物	5.1928~1580.0046	定量变量 (单位：百亿元)
	X_3 (Total import trade) 进口贸易总额	跨越过境的货品或服务交易	3.5500~3156.2730	定量变量 (单位：百亿元)
	X_4 (Economically active population) 经济活动人口	16 周岁及以上有劳动能力，参加或要求参加社会经济活动的人	406.82~811.04	定量变量 (单位：百万人)
	T (Year) 年份	研究年份	1978~2019	定量变量 (单位：年)

4. 研究方法概述

4.1. 线性回归

4.1.1. 最小二乘估计

含有 p 个自变量的理论线性回归模型具有以下的一般形式：

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \ell \tag{4.1}$$

如果对因变量 Y 和自变量 $X_1, X_2, \cdots, X_p$ 进行了 n 次观察，得到的 n 组对数据集 $y_i, x_{i1}, x_{i2}, \cdots, x_{ip}$，$i = 1, \cdots, n$

若因变量(或被解释变量) y_i 与自变量(或解释变量) $x_{i1}, x_{i2}, \cdots, x_{ip}$ 之间存在如下关系式：

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \tag{4.2}$$

4.1.2. 线性回归模型的基本假设

1) 解释变量一般是非随机变量。

2) 假设随机误差项 $\varepsilon_i\left(i=1,\cdots,n\right)$ 满足(G-M 条件)

$$E\left(\varepsilon_i\right)=0, COV\left(\varepsilon_i,\varepsilon_j\right)=\begin{cases}\sigma^2, i=j\\ 0, \quad i\neq j\end{cases} \tag{4.3}$$

3) 要求样本容量个数要远远大于解释变量的个数。

记

$$y=\begin{pmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{pmatrix}, X=\begin{pmatrix}1 & X_{11} & \cdots & X_{1,p}\\ 1 & X_{21} & \cdots & X_{2,p}\\ \vdots & \vdots & \ddots & \vdots\\ 1 & X_{n1} & \cdots & X_{n,p}\end{pmatrix}, \beta=\begin{pmatrix}\beta_0\\ \beta_1\\ \vdots\\ \beta_p\end{pmatrix}, \ell=\begin{pmatrix}\ell_1\\ \ell_2\\ \vdots\\ \ell_p\end{pmatrix} \tag{4.4}$$

且假设 $rank\left(X\right)=p+1$， $\ell_i\left(i=1,2,\cdots,n\right)$ 互不相关，均值都是 0，并且有公共的方差 σ^2，则的线性回归模型

$$y=X\beta+\ell, E\left(\ell\right)=0, Cov\left(\ell\right)=\sigma^2 \tag{4.5}$$

则称式(4.3)为响应变量 y 关于自变量 $x_1,\cdots,x_p$ 的多元线性回归模型，其中 β_0 称为回归的常数项，$\beta_1,\cdots,\beta_p$ 称为回归系数。当 $p=1$ 时，式(4.5)即为一元线性回归模型。

4.1.3. 回归方程的显著性检验

多元回归模型(4.3)初步建立后，是否真正解释了预测变量和因变量的关系，还要对所建立的模型进行显著性的检验。其中回归方程的检验，其实就是检验假设：原假设为 p 个回归系数都等于零，也就是检验 $H_0:\beta_1=\beta_2=\beta_3=\cdots=\beta_p=0$，这就意味着因变量 y 与所有的自变量 x_j 都不存在回归关系，多元回归方程没有意义。相应的备择假设 $H_1:\beta_1,\beta_2,\beta_3,\cdots,\beta_p$ 不全为 0， H_0 成立时有：

$$F=\frac{SS_R/p}{SS_E/n-p-1}\sim F\left(p,n-p-1\right)。$$

即 F 服从 F 分布。这样就可以用 F 统计量来检验回归方程是否有意义。

在多元线性回归中模型的线性检验是涉及三个离差平方和，也就是总平方和 SS_T，回归平方和 SS_R 以及残差平方和 SS_E，它们之间具有等式关系： $SS_T=SS_R+SS_E$。我们定义拟合优度为

$$R^2=\frac{SS_R}{SS_T}=1-\frac{SS_E}{SS_T}=1-\frac{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2} \tag{4.6}$$

拟合优度 R^2 能衡量各个自变量对响应变量变动的解释程度，其取值范围在 $\left[0,1\right]$ 之间，R^2 的值越与 1 接近，则自变量对响应变量的解释程度就相对越高，其值越接近于 0，则自变量对响应变量的解释程度就相对越弱。一般来讲，我们所处理的数据集会非常的庞大，其自变量的个数也会随之不断增多，当我们所研究问题涉及的自变量的个数增加，回归的平方和就增加，残差平方和就减少，所以相应的 R^2 也会增大；对应的，当我们所研究问题涉及的自变量的个数减少，回归的平方和增加，残差平方和减少。于是这样就容易把不显著的自变量留在线性回归模型中，为此，需要对该指标加以调整。于是，定义调整的可决系数为

$$R_A^2=1-\frac{SS_E/\left(n-p-1\right)}{SS_T/\left(n-1\right)} \tag{4.7}$$

4.1.4. 回归系数的假设检验

模型有统计学意义所包含的意义并不是说所涉及的每一个回归系数都有意义，因此研究者在建立模

型后有必要对每个回归系数做检验。即做检验：$H_i:\beta_i=0$。此处我们用 t 统计量对回归系数做检验。对于模型(4.3)，β 的 LS 估计为 $\hat{\beta}=(X'X)^{-1}X'y$，记 $C_{p\times p}=(c_{ij})=(X'X)^{-1}$，则有 $\hat{\beta}_i\sim N(\beta_i,\sigma^2c_{ii})$，于是当 H_i 成立时，$\dfrac{\hat{\beta}_i}{\sigma\sqrt{c_{ii}}}\sim N(0,1)$，根据 t 分布的定义，

$$t_i=\frac{\hat{\beta}_i}{\hat{\sigma}\sqrt{c_{ii}}}\sim t_{n-p} \tag{4.8}$$

其中 $\hat{\sigma}^2=\left\|y-X\hat{\beta}\right\|^2/(n-p)$，这里 t_{n-p} 表示在 n-p 的 t 分布，对给定的水平 α，当 $|t_i|>t_{n-p(\alpha/2)}$ 时，拒绝原假设 H_i，否则就接受 H_i。

5. 回归方法实证分析

5.1. 数据的描述性分析

5.1.1. 变量的描述性分析

从折线图图 1 中我们可以看到我国的各项影响财政收入的经济指标总体上呈现增长趋势。

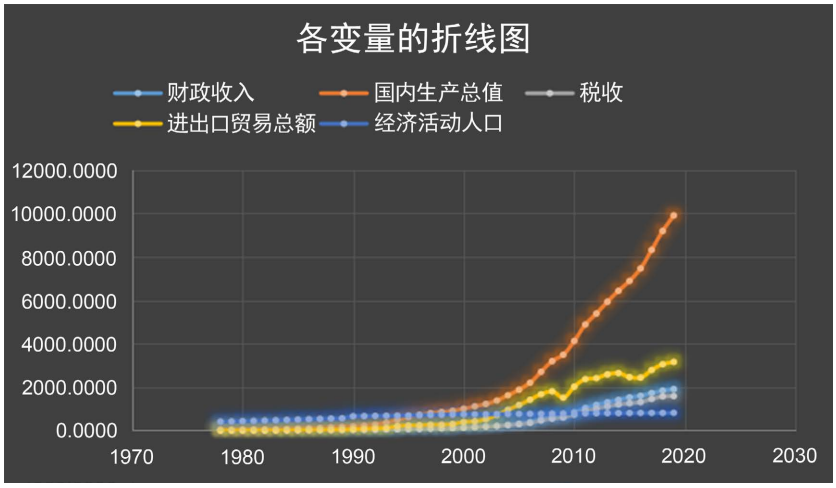


Figure 1. Line chart of variables
图 1. 各变量折线图

5.1.2. 变量之间的散点图矩阵

从图 2 可以看出，我国财政收入与国内生产总值、税收、进出口贸易总额以及经济活动人口均是呈正相关关系，这我们的常理认知是符合的。但是我们从图中也不难看出进出口贸易总额这一因素与其他几个变量之间的相关图中有一点小的跳跃，再结合我们的数据进行观察这个明显的跳跃点的时间是 2009 年，其实这也能说明原因，由于 2008 年金融危机的影响，由于中国出口下降，实体经济严重受损，我国对美国出口的产品中服装和纺织品、鞋袜、玩具、低端生活用品等劳动密集型家用产品所占比重有相当大。又因为这些产品的主要消费群体是美国的中低收入阶层，而他们在这次金融危机中遭受到了很大的经济损失，这必然在很大程度上影响来年这些产品的对美国的出口。

5.1.3. 自变量的箱线图

从图 3 中可以看出，国内生产总值、税收、进出口贸易总额以及经济活动人口这几个变量的分布均较离散，国内生产总值这个变量出现了两个异常点，税收这个变量出现了一个异常点。

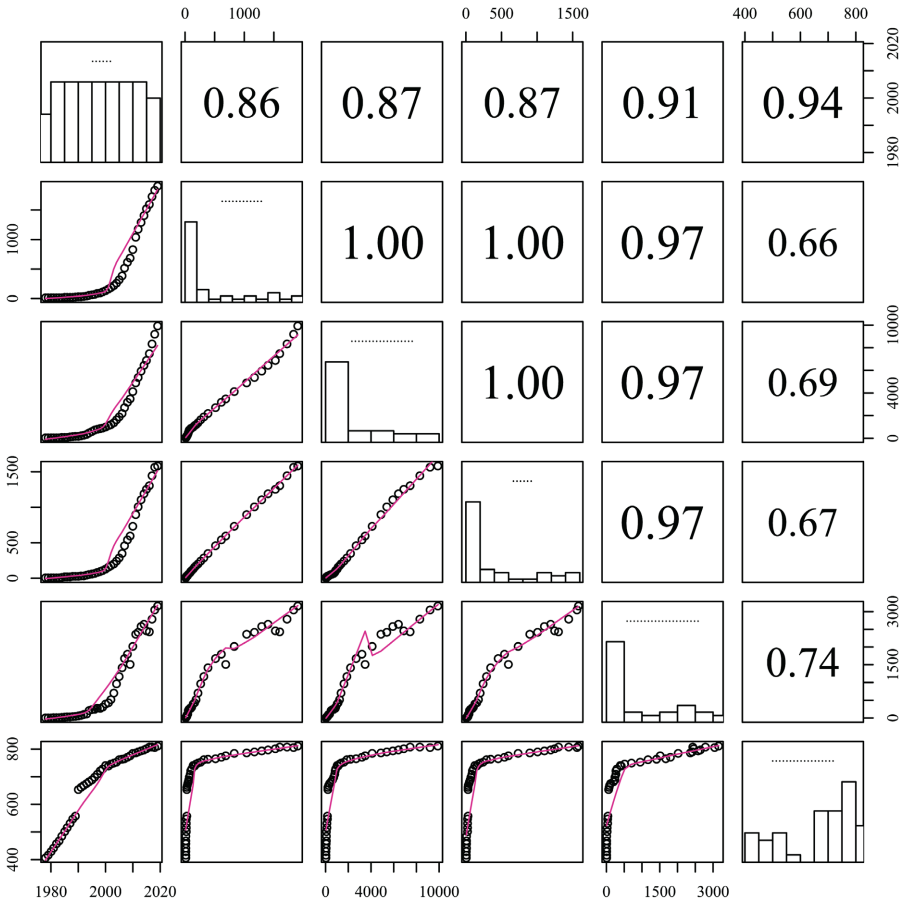


Figure 2. Scatter plot matrix between variables
图 2. 变量之间的散点图矩阵

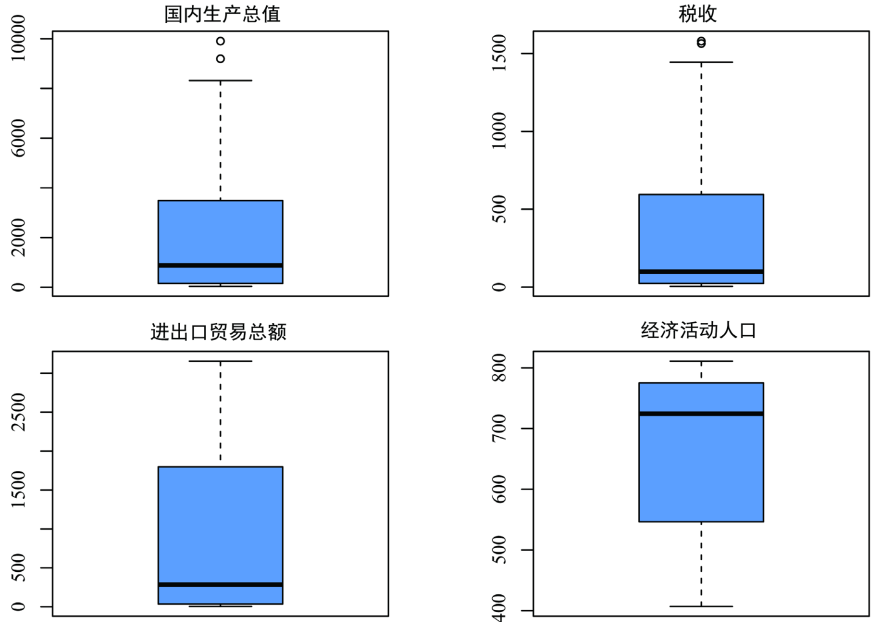


Figure 3. Boxplots of arguments
图 3. 自变量的箱线图

5.2. 线性回归分析

通过编程实现我们得到如下结果：

5.2.1. 变量回归系数

如表 2 所示，

Table 2. Linear regression results
表 2. 线性回归结果

变量	回归系数
截距项	18.331603
国内生产总值	0.005023
税收	1.266439
进出口贸易总额	−0.046212
经济活动人口	−0.036

于是得到线性回归模型：

$$\hat{y} = 18.331603 + 0.005023x_1 + 1.266439x_2 - 0.046212x_3 - 0.036957x_4 \tag{5.1}$$

其中 x_1 为国内生产总值，x_2 为税收，x_3 为进出口贸易总额，x_4 为经济活动人口，y 为财政收入。

由于在很多回归分析中我们所使用的的原始数据自变量 $x_i\ (i = 1,2,3,4)$ 与响应变量都是有单位的，这就涉及到量纲问题。从数值上看，它们样本取值的极差会有很大的差异，均值与标准差也各有不同，因此不能由偏回归系数的大小直接说明对因变量线性影响的大小。为了分析的更加具有说服力，我们应该对自变量采取变量标准化与计算标准化偏回归系数的方法来进行处理以确保分析更加准确。

5.2.2. 变量标准化回归系数

如表 3 所示，

Table 3. Standard linear regression results
表 3. 标准线性回归结果

变量	回归系数
国内生产总值	2.387e−02
税收	1.061e+00
进出口贸易总额	−8.181e−02
经济活动人口	−8.076e−03

此时的线性模型为：

$$\hat{y} = 2.387\text{e}{-023}x_1 + 1.061\text{e}{+00}x_2 - 8.181\text{e}{-02}x_3 - 8.076\text{e}{-03}x_4 \tag{5.2}$$

5.2.3. 模型的检验

如表 4 所示，

Table 4. Parameter estimation and testing
表 4. 参数估计及检验

变量	回归系数 $\hat{\beta}$	标准误 $S_{\hat{\beta}}$	t 值	p 值	标准回归系数
x_0	18.331603	13.091660	1.400	0.170	
x_1	0.005023	0.008680	0.579	0.566	2.387e−02
x_2	1.266439	0.055332	22.888	<2e−16***	1.061e+00
x_3	−0.046212	0.008811	−5.245	6.6e−06***	−8.181e−02
x_4	−0.036957	0.023020	−1.605	0.117	−8.076e−03

Residual standard error: 11.18 on 37 degrees of freedom; Multiple R-squared: 0.9997, Adjusted R-squared: 0.9997; F-statistic: 2.99e+04 on 4 and 37 DF, p-value: <2.2e−16.

1) 回归方程的假设检验

检验假设：$H_0:\beta_1=\beta_2=\beta_3=\beta_4=0$，$H_1:\beta_1,\beta_2,\beta_3,\beta_4$ 不全为 0。

从表 4 我们可以看出，模型整体显著，对财政收入有显著性影响的变量有税收、进出口贸易总额，拟合优度 0.9997，调整的可决系数为 0.9997。

由方差分析结果可见(见表 4)，模型的 F 值为 2.99e+04，$P<0.0001$，所以我们的模型回归是有意义的。

2) 回归系数的假设检验

检验假设：$H_0:\beta_i=0$，$H_1:\beta_i\neq 0$ $(i=1,2,3,4)$。

通过 t 检验的最终结果我们可以看到，回归系数 b_2,b_3 的 P 值都小于 0.010，我们可认为解释变量税收 x_2 和进出口贸易总额 x_3 表现显著；而 b_1 的 P 值大于 0.050，因此，不能否定对 $\hat{\beta}_1=0$ 的原假设，我们可以认为国内生产总值 x_1 对财政收入 y 没有显著的影响。b_4 的 P 值小于 0.5，拒绝原假设，这说明经济活动人口对财政收入 y 时是有影响的，但我们也能发现，经济活动人口所对应回归系数为负值，这与经济现实是不相符合的，这可能是因为解释变量之间存在高度的共线性引起的，这对于分析是一个很重要的关注点。

线性回归模型还原：

$$\hat{y}=1.266439x_2-0.046212x_3 \tag{5.3}$$

其中 x_2 为税收，x_3 为进出口贸易总额，y 为财政收入。

模型解读：

控制税收不变时，进出口总贸易额每增加一个单位，财政收入减少−0.046212 个单位；控制进出口总贸易额保持不变时，每增加一个单税收位，财政收入增加 1.266439 个单位。

5.2.4. 模型诊断

从以下图 4、图 5、图 6 中可以看出，线性回归模型大致保持水平状态，有较好的线性关系，不存在异方差现象，残差服从正态分布，存在三个强影响点。剔除这三个样本点重新建立线性模型的拟合优度 0.998，调整的可决系数为 0.998，模型检验对模型的影响不是很大，固采用原模型。

5.2.5. 逐步回归结果

分别用 spss 和 R 操作结果如下表所示，

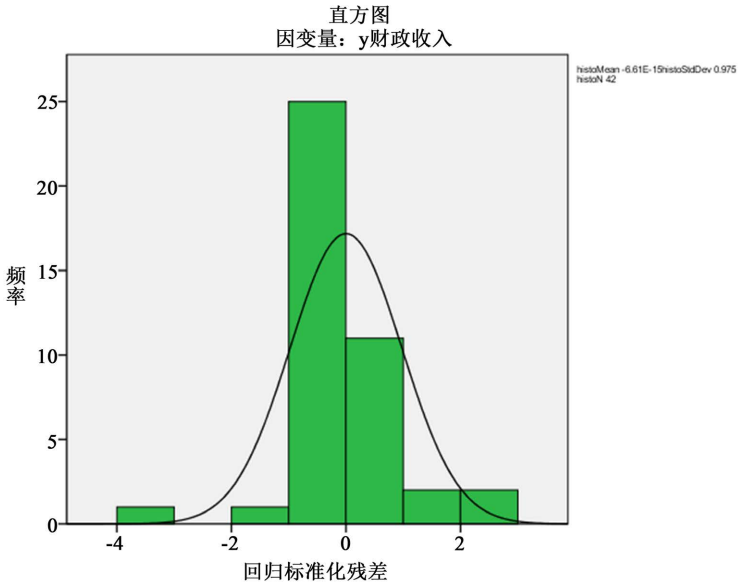


Figure 4. Regression normalized residuals
图 4. 回归标准化残差

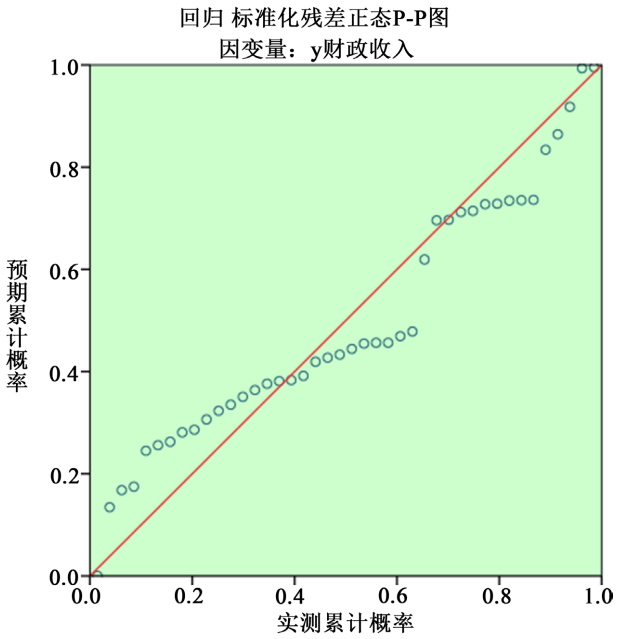
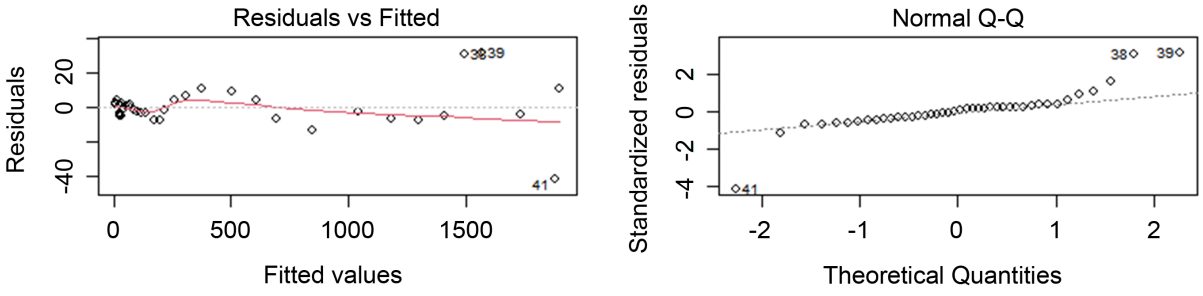


Figure 5. Measured cumulative probability
图 5. 实测累计概率



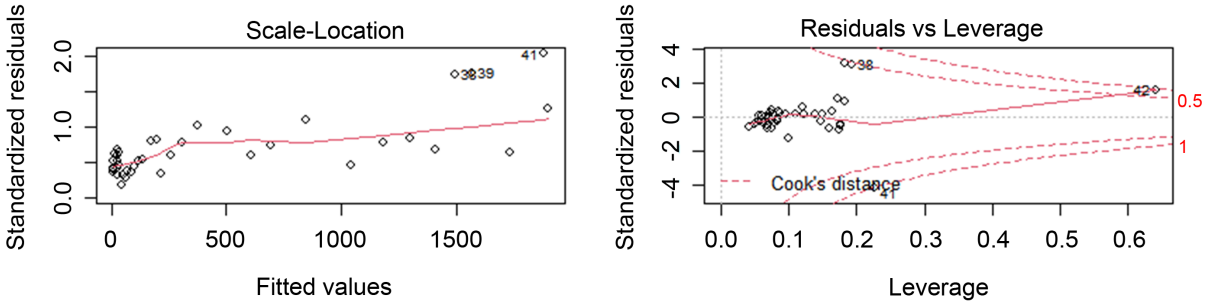


Figure 6. Linear regression model diagnostics
图 6. 线性回归模型诊断

Table 5. Model summary
表 5. 模型摘要

模型	R	R 平方	调整后 R 平方	估计值的标准误差
2	1.000[b]	0.9997	0.9997	11.08

a.预测值：(常数)，x_2, x_3；b.因变量：y。

Table 6. Variance part analysis
表 6. 方差份析

模型	平方和	df	平均值平方	F	显著性
1 回归	14952177.022	2	7476088.511	58923.506	0.000*c*
残差	4948.237	39	126.878		
总计	14957125.258	41			

a.预测值：(常数)，x_2, x_3；b.因变量：y。

Table 7. Coefficient
表 7. 系数

模型	未标准化系数		标准化系数	T	显著性	B 的信赖区间	
	B	标准错误	Beta			下限	上限
2 (常数)	−2.367	2.388		−0.991	0.328	−7.197	2.463
x_2	1.305	0.015	1.093	86.040	0.000	1.275	1.336
x_3	−0.054	0.007	−0.096	−7.581	0.000	−0.069	−0.040

模型表示为：

$$\hat{y} = 1.305x_2 - 0.054x_3 \tag{5.4}$$

综合观察表 5、表 6、表 7 用逐步回归分析所得的结果与我们一开始做的线性回归分析不谋而合。

5.3. 预测与评估

最后，我们对建立的模型进行预测和评估，用于预测的自变量为国内生产总值、税收、进出口贸易总额和经济活动人口。各变量的均值，在线性模型上进行预测并计算得我国的平均财政收入，并与真实

值做比较，为便于比较将其放在表 8 中。从表中，我们可以直接看出，就这个数据集而言，线性模型预测效果很好。

Table 8. Forecas
表 8. 预测

方法	预测值	真实值
线性回归	443.6537	443.6537

6. 结论

6.1. 结果分析

本研究为了研究我国国家财政收入与自变量国内生产总值、税收、进出口贸易总额和经济活动人口之间的关系，从相关分析结果可以看到，我国的财政收入和自变量国内生产总值、税收、进出口贸易总额和经济活动人口的相关系数分别为 0.9967401、0.9995907、0.9679219、0.6581953，表现出非常密切的关系($r > 0.8$，$p < 0.001$)。财政收入与税收收入的关系是密切的($r = 0.9995907$，$p < 0.001$)。通过建模我们发现，财政收入与税收和进出口贸易总额上关系是很紧密的，当然由于本文只是选取了个人主观想法上的几个因变量，这与实际上的财政收入影响因素是远远不够的，因此，如果文章还想进一步研究就得把影响财政收入的所有因素继续考虑进来分析。还有一个地方值得一提的是，尽管这里对所建立的模型进行了相应的预测，但是我们要知道，国家的经济状况不会随我们建立的模型而保持一成不变，就比如这次新型冠状病毒的影响，国家的经济肯定会受到重大的影响。所以我们在进行预测时根据的是历史数据。对未来的数据我们要根据情况进行分析。换句话讲，我们所做的工作其实是一种对过去的经济状况的一种分析，个人觉得这样描述比较准确一些。

6.2. 合理性建议

由于我们所选取的数据是一部分数据进行分析，所以对最完整的模型(所有影响财政收入的因素)还是会有一定的差异，这也就不难解释我们的模型对这几个影响因素的参数估计会有负值的情况，按常理应该是正相关的。

通过分析可知，我国的财政收入受到税收、进出口贸易总额的显著影响。其中税收和进出口贸易总额的影响又更加明显。总的来说，面对现在这样一个大熔炉的时代，全球化发展成为一种不可逆转的潮流，发展中国家要珍惜这样一个发展机会，要加强对外的交流合作，争取朝着好的方向发展。税收收入要能够与国家的发展相适应，不能盲目调高税率；同时作为一名合格的公民，要本着对自己，对祖国负责的态度，坚决抵制偷税漏税等不法行为。

一个国家的财政收入水平能反应一个国家经济实力，影响财政收入的根本性因素是经济发展水平，因此国家只有大力发展经济，使经济持续稳定地增长，国家的财政收入才能增加，才能实现中华民族伟大的复兴史，使我们国家能够长久屹立于世界舞台上。因此，在人口较多的中国，国家要增加财政收入，当然我们的政府要顺应发展控制好我们国家的财政收入，收入过高过低都对国家发展不利。增加财政收入要处理好国家、社会、企业、个人的关系。在保证国家财政收入稳步增长的基础上，使得社会创造出良好的经济发展环境，进而逐步使人民的生活水平得到不断提高。同时，国家特别是我们发展中国家应该同其他发展较好的国家进行对比，在此基础上不断完善调整自身发展。值得注意的是，国家财政收入占国内生产总值比重偏低，国家掌握的财政收入偏少是我们国家面临的一个问题。

参考文献

[1] 罗博炜, 洪智勇, 王劲屹. 多元线性回归统计模型在房价预测中的应用[J]. 计算机时代, 2020(6): 51-54.

[2] 王松桂. 线性模型引论[M]. 北京: 科学出版社, 2004: 121-123

[3] 孙毅, 刘仁云, 王松, 冷晓冰, 臧雪柏. 基于多元线性回归模型的考试成绩评价与预测[J]. 吉林大学学报(信息科学版), 2013, 31(4): 404-408.

[4] 刘凯. 分组广义线性潜变量模型在教育中的应用[D]: [硕士学位论文]. 吉林: 东北师范大学, 2013.