

基于多元回归分析探究税收的影响因素

杨冬敏

曲阜师范大学统计与数据科学学院, 山东 曲阜

收稿日期: 2023年5月6日; 录用日期: 2023年5月26日; 发布日期: 2023年6月7日

摘 要

税收是国家(政府)公共财政中最重要的一项收入, 也是国家(政府)公共财政的重要收入来源。本文运用SPSS及R软件, 对1996~2022年度(所选资料来源于《中华人民共和国统计年鉴》)的有关统计资料进行分析, 并运用普通最小二乘方法, 通过对影响税收的各因素的共线性、异方差及自相关等特征的研究, 构建出我国税收收入的多元线性回归模型。通过建立的模型, 得出了影响税收增长的重要因子, 从而为税收收入的预测奠定了基础。

关键词

税收, 多元回归, 相关性, 显著性, 预测

Based on Multiple Regression Analysis to Explore the Influence Factors of Tax

Dongmin Yang

School of Statistics and Data Science, Qufu Normal University, Qufu Shandong,

Received: May 6th, 2023; accepted: May 26th, 2023; published: Jun. 7th, 2023

Abstract

Tax is the most important income of the state (government) public finance, and also the important source of income of the state (government) public finance. In this paper, SPSS and R software are used to analyze the relevant statistical data from 1996 to 2022 (the selected data is from the Statistical Yearbook of the People's Republic of China), and the ordinary least squares method is used to study the collinearity, heteroscedasticity and autocorrelation of various factors affecting tax revenue. Construct multiple linear regression model of tax revenue. Through the established model, the important factors affecting the growth of tax revenue are obtained, thus laying the foundation for the forecast of tax revenue.

Keywords

Tax, Multiple Regression, Correlation, Significance, To Predict

Copyright © 2023 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 绪论

从 21 世纪以来,中国在科技与互联网产业飞速发展的基础上,经济的发展迫在眉睫,新形势下的经济发展是经济稳定与和谐发展的结果。《税收结构是否影响经济增长——基于中国四个经济区域实证分析》[1]一文中所述,税收在宏观调控中发挥着重要作用,它不仅是支持公共产品供给的主要资金来源,而且对社会资源配置和缩小收入差距起到了促进作用,实施有效的税收体制已成为社会规划者保持经济可持续增长的重要手段。根据《论税收对产业结构调整的影响》[2]一文所述,税收政策对于我国的产业结构具有较大的影响,它不仅能够调节国家的消费过程,也能调节国家的生产和分配过程。它是一个国家政府赖以生存的经济基础。《税收影响因素分析及建议》[3]一文提出,国内生产总值(GDP)、财政支出、商品零售价格指数和固定资产投资额对中国税收收入的具有显著影响。《我国税收收入影响因素分析》[4]一文中,指出国内生产总值、财政支出和社会消费品零售总额对税收收入的影响较大,对于提高我国税收收入、促进经济平稳增长有着指导意义。

因此,对影响税收的主要因素进行分析,并对国家的税收进行预测,这对国家进行宏观经济调控,调整国民收入的再分配,指导资源的优化配置,推动国家的经济增长和产业发展具有重要的实际意义。随着研究的深入,学者们将更多的影响税收的因素纳入计量模型中量化分析其对税收收入的影响及其程度,然而,经济指标选取的增多,会引起多重共线性问题,因此学者们采取了岭回归、主成分分析等方法解决该问题。左雨婷等以 2001~2018 年安徽省的相关数据,通过构建立岭回归模型研究了三大产业增加值、财政支出、固定资产投资总额和社会消费品零售总额分别对安徽省税收收入的影响,提出优化产业结构、注重税收改革效率及推动居民消费升级的政策建议[5];高天惠等在定性分析影响安徽省税收收入的因素的基础上,运用主成分分析方法降维,得到影响税收的重要因素,并据此给出了相关的政策建议[6];邓光青选择因子分析的方法发现经济发展因子和货币政策因子对税收收入有着正向的相关性,认为政府应放宽政策,增加政府购买[7]。本文采用多元线性回归方法,消除多重共线性、异方差性及序列自相关性等问题,对影响税收增长的主要因素进行了研究。

2. 影响因素及数据来源

影响税收的因素有社会经济发展水平、国家的宏观经济政策、税收征收管理能力等。通过查阅《中华人民共和国统计年鉴》收集到 1996~2022 年度的相关统计资料,包括税收收入、国内生产总值、国民生产总值、财务开支、消费品零售额、全部固定资产投资、货币供应量、进口和出口总值等相关数据。利用相关数据分析影响税收增长的因素,构建税收收入回归模型,对我国税收收入进行预测和控制。

3. 构建理论模型并进行相关性分析

为了对影响税收收入的主要影响因素进行分析,以税收收入 y (亿元)为被解释变量,将国内生产总值 x_1 (亿元),国民生产总值 x_2 (亿元),财务开支 x_3 (亿元),消费品零售额 x_4 (亿元),全部固定资产投资 x_5

(亿元), 货币供应量 x_6 (亿元), 进口和出口总值 x_7 (亿元)等可能与之有关的经济因素当作解释变量, 建立了多元线性回归理论模型。模型为:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \varepsilon$$

其中: β_0 为常数; $\beta_i (i=1,2,3,4,5,6,7)$ 为影响因素的相关系数。做相关性分析, 结果如图 1 所示。

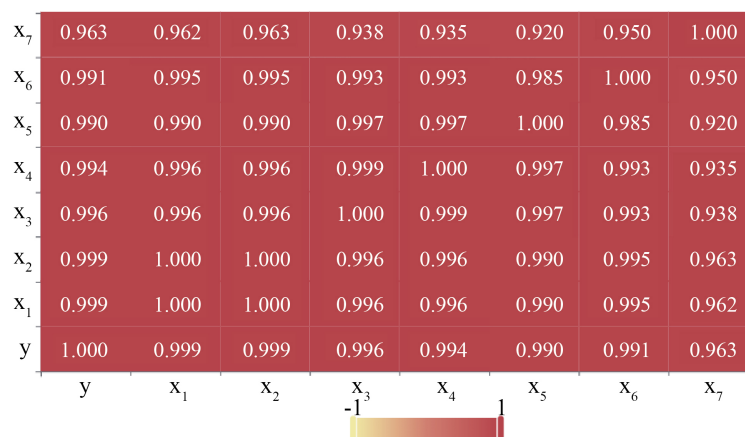


Figure 1. Heat map of correlation results of each index
图 1. 各指标相关性结果热力图

从相关阵可以看出 y 与七个自变量的相关系数都在 0.9 以上, 说明所选自变量是与因变量高度相关的, 用 y 与自变量做多元线性回归是合适的。

4. 回归分析及拟合优度诊断

利用 R 软件对原始数据做回归分析, 输出结果表 1 所示。

Table 1. Statistical table of fitting effect of each variable

表 1. 各变量拟合效果统计表

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-30.77739	794.23522	-0.039	0.969493
x_1	0.38146	0.19726	1.934	0.068169
x_2	-0.17567	0.19588	-0.897	0.381034
x_3	0.55509	0.12377	4.485	0.000254***
x_4	-0.42483	0.09619	-4.416	0.000296***
x_5	0.06265	0.03070	2.041	0.055440
x_6	-0.05819	0.02085	-2.792	0.011634*
x_7	0.01627	0.02593	0.627	0.537916

通过结果可知, 拟合优度 $R^2 = 0.9995$, 修正的可决定系数为 0.9993, 二者皆接近于 1; F 检验的 $F = 5444 > F_{0.05}(7, 19) = 2.54$, $P < 2.2e-16 < 0.05$, 说明模型整体高度显著。

从而通过 R 软件利用最小二乘法得到模型参数的估计值, 得到初步经验回归方程为:

$$y = -30.77739 + 0.38146x_1 - 0.17567x_2 + 0.55509x_3 - 0.42483x_4 + 0.06265x_5 - 0.05819x_6 + 0.01627x_7$$

虽然自变量 $x_1, x_2, x_3, x_4, x_5, x_6, x_7$ 整体上对 y 有显著影响, 但有的自变量对 y 的显著性较差。其中 x_7 消费品零售额的 P 值 = 0.537916 最大, 不显著。 x_2 的 P 值为 0.381034, 也不显著。 x_1, x_5 的 P 值分别为 0.068169, 0.055440, 只是弱显著。

同时, x_2, x_4 和 x_6 系数的符号与预期相反, 且每个自变量的显著性和这些自变量与因变量 y 两两之间的相关系数并不一致, 产生这个问题的原因可能是自变量之间存在共线性。由相关系数阵中各解释变量之间相关性值都接近于 1 可以进一步看出, x_2, x_4 和 x_6 之间的确存在严重的相关性。研究进一步利用逐步回归的方法来解决变量间的多重共线性问题。

5. 多重共线性修正

分别对 y 和 $x_1, x_2, x_3, x_4, x_5, x_6, x_7$ 间的关系进行一元线性回归拟合, 得到 7 个一元线性回归模型。

$$y = \beta_0 + \beta_1 x + \varepsilon$$

通过对数据进行拟合, 每个模型的参数估计值拟合结果如表 2 所示。可以看出变量 x_1 的 R^2 值最高, 具有最佳的拟合效果。表明国内生产总值(x_1)在税收收入中起主要作用。各解释变量的一元回归模型拟合优度的大小顺序为: $x_1, x_2, x_3, x_4, x_6, x_5, x_7$ 。

Table 2. Unitary linear regression fitting effect statistical table
表 2. 一元线性回归拟合效果统计表

自变量	x_1	x_2	x_3	x_4	x_5	x_6	x_7
β_1	0.19174	0.19283	0.73581	0.442022	0.223538	0.310153	0.991238
t	109.7469	107.3433	47.10349	38.14153	31.18633	32.55841	13.78774
R-squared	0.998259	0.998181	0.990624	0.98577	0.978864	0.980574	0.900522
Adjusted R-squared	0.998177	0.998094	0.990177	0.985093	0.977858	0.979649	0.895785
F 统计量	12044.38	11522.58	2218.738	1454.776	972.5874	1060.05	190.1019

然后, 利用 SPSS 软件以 x_1 为基础依次加入其他解释变量进行逐步回归, 回归结果如表 3 和表 4 所示:

Table 3. Statistical table of stepwise regression effect
表 3. 逐步回归效果统计表

模型	R	R 方	调整后 R 方	标准估算的 错误	R 方 变化量	F 变化量	自由度 1	自由度 2	显著性 F 变化量	德宾 - 沃森
1	0.999 ^a	0.997	0.997	2326.45796	0.997	8980.963	1	25	0.000	
2	0.999 ^b	0.998	0.997	2165.71244	0.000	4.849	1	24	0.038	
3	0.999 ^b	0.998	0.998	2030.36484	0.000	4.306	1	23	0.049	
4	1.000 ^a	0.999	0.999	1169.79858	0.001	47.287	1	22	0.000	2.043

Table 4. Stepwise regression coefficient effect statistics table**表 4.** 逐步回归系数效果统计表

模型	未标准化系数		标准化系数		显著性
	B	标准错误	Beta	t	
1	(常量)	-4935.794		-7.293	0.000
	x_1	0.189	0.999	94.768	0.000
2	(常量)	-5351.765		-8.106	0.000
	x_1	0.229	1.213	12.410	0.000
	x_6	-0.067	-0.215	-2.202	0.038
3	(常量)	-3922.894		-4.237	0.000
	x_1	0.197	1.043	8.502	0.000
	x_6	-0.083	-0.267	-2.810	0.010
	x_3	0.165	0.222	2.075	0.049
4	(常量)	-832.678		-1.194	0.245
	x_1	0.213	1.126	15.698	0.000
	x_6	-0.084	-0.270	-4.930	0.000
	x_3	0.725	0.973	7.759	0.000
	x_4	-0.369	-0.832	-6.877	0.000

1) 在 x_1 基础上, 引入拟合优度位于第五位的 x_6 , 得到最小二乘回归结果, 回归方程为:

$$y = -5351.765 + 0.229x_1 - 0.67x_6$$

$R^2 = 0.998$, 修正的可决系数为 0.997, $F=5184.24$, $DW=0.297$ 。

2) 第二步再引入拟合优度位于第三位的 x_3 , 得到最小二乘回归结果, 回归方程为:

$$y = -3922.894 + 0.197x_1 - 0.083x_6 + 0.165x_3$$

$R^2 = 0.998$, 修正的可决系数为 0.998, $F=3399.74$, $DW=0.356$ 。

3) 第三步再引入拟合优度位于第四位的 x_4 , 得到最小二乘回归结果, 回归方程为:

$$y = -832.678 + 0.213x_1 - 0.084x_6 + 0.725x_3 - 0.369x_4$$

$R^2 = 0.999$, 修正的可决系数为 0.999, $F=8899.592$, $DW = 2.043$ 。

从此模型中可以看出, 国内生产总值(x_1)、货币供应量(x_6)、财务开支(x_3)和消费品零售额对税收收入(x_4)有显著影响。

6. 异方差性

6.1. 残差图分析

通过构建国内生产总值(x_1)与税收收入 y 之间对应关系的散点图(见图 2)、财务开支(x_3)与税收收入 y 之间对应关系的散点图(见图 3)、消费品零售额(x_4)与税收收入 y 之间对应关系的散点图(见图 4)、货币供应量(x_6)与税收收入 y 之间对应关系的散点图(见图 5)来研究税收与相关主变量的异方差性。从中可以看出, 随着 x_1 、 x_6 、 x_3 和 x_4 的增加, 各项税收也不断增加, 但是离散程度也逐步扩大, 表明变量之间可能存在增量异方差性。

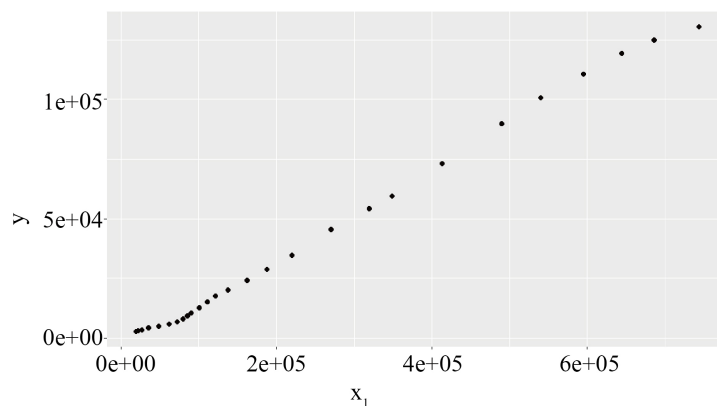


Figure 2. Scatter plot of the relationship between gross product and tax revenue

图 2. 生产总值(x_1)与税收收入 y 之间对应关系的散点图

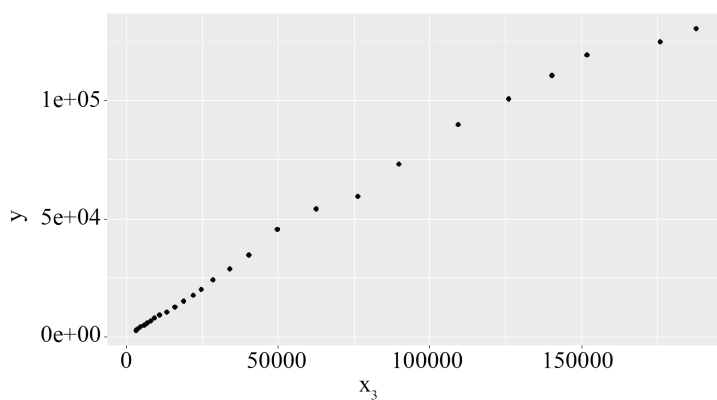


Figure 3. Scatter plot of the relationship between financial expenditure and tax revenue

图 3. 财务开支(x_3)与税收收入 y 之间对应关系的散点图

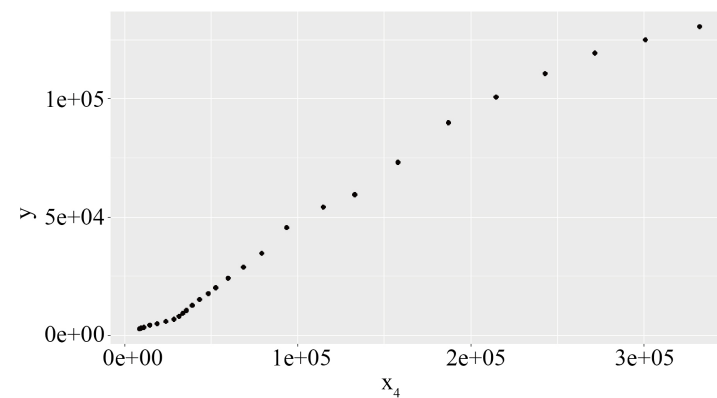


Figure 4. Scatter plot of the relationship between retail sales of consumer goods and tax revenue

图 4. 消费品零售额(x_4)与税收收入 y 之间对应关系的散点图

进一步以标准化残差 ZRE_i 为纵坐标, 以因变量 y 为横坐标画散点图(见图 6), 标准化残差值随 y 值的增大而增大, 具有明显的规律, 因而可认为模型的随机误差项 ε_i 的方差是非齐性的, 存在异方差性。

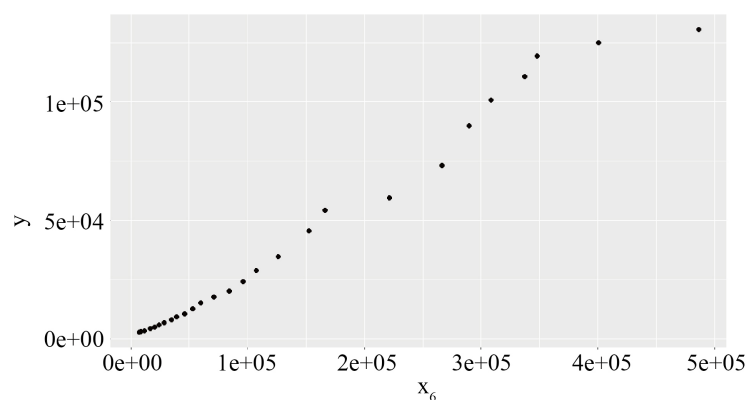


Figure 5. Scatter plot of the relationship between money supply and tax revenue

图 5. 货币供应量(x_6)与税收收入 y 之间对应关系的散点图

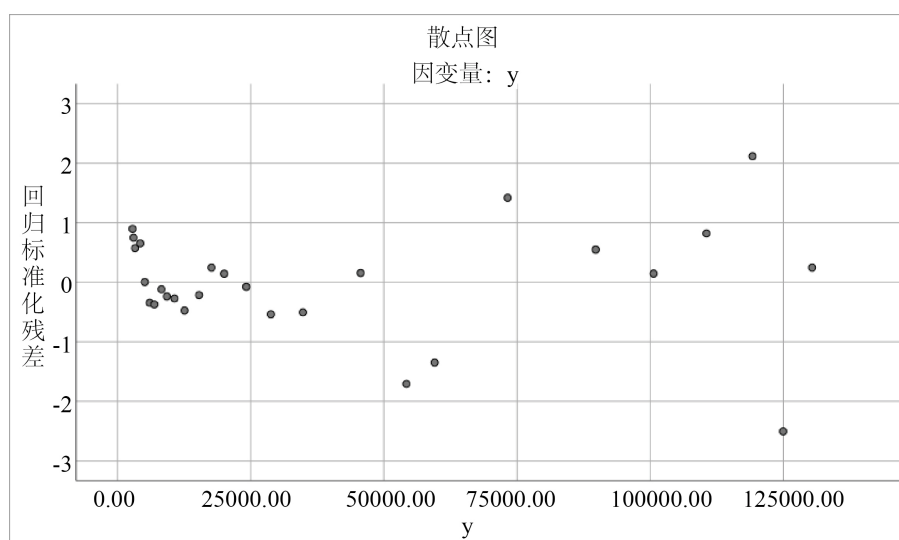


Figure 6. Interpreted variables and standardized residual scatter plots

图 6. 被解释变量和标准化残差散点图

6.2. 异方差修正——加权最小二乘估计

利用 SPSS 软件分别计算解释变量 x_1 、 x_6 、 x_3 和 x_4 与普通残差的等级相关系数如表 5 所示, 由结果可知 x_6 与普通残差的等级相关系数为 0.723 最大, 所以选取 x_6 构造权函数。

Table 5. Coefficient of correlation between explanatory variables and ordinary residuals

表 5. 解释变量与普通残差的等级相关系数

Spearman's rho		x_1	x_3	x_4	x_6
ABSE	相关系数	-0.094	0.476	0.603**	0.723**
	Sig. (双尾)	0.641	0.012	0.001	0.000
	N	27	27	27	27

用 Weight Estimate 估计幂指数 m , 对数似然值如表 6 所示, 取对数似然值最大的 m 为 2。

Table 6. Logarithmic likelihood values of different powers
表 6. 不同幂指数的对数似然值

幂	对数似然值
-2	-243.182
-1.5	-238.97
-1	-235.176
-0.5	-231.631
0	-226.555
0.5	-219.334
1	-212.049
1.5	-207.103
2	-206.665
2.5	-210.847

加权最小二乘法修正结果 1 如表 7 和加权最小二乘法修正结果 2 见表 8 所示，可以看出在给定显著水平 $\alpha = 0.05$ 时， x_1 、 x_6 、 x_3 和 x_4 的系数检验统计量 t 的 p 值都远小于 0.05，则拒绝原假设，认为 x_1 、 x_6 、 x_3 和 x_4 对 y 的影响是显著的；拟合优度也发生了变化，从原来的 $R^2 = 0.9995$ ，修正的可决系数 0.9993 变为拟合优度 $R^2 = 0.999973$ ，修正的可决系数为 0.999968，说明模型更加显著。从而消除共线性和异方差性之后，可得经验回归方程：

$$y = -466.6158 + 0.207991x_1 + 0.781413x_3 - 0.233524x_4 - 0.402295x_6$$

Table 7. Modified result 1 by weighted least square method
表 7. 加权最小二乘法修正结果 1

	系数	标准误差	t	概率
(常量)	-466.6158	139.4642	-3.345775	0.0029
x_1	0.207991	0.003425	60.72185	0
x_3	0.781413	0.020043	38.98637	0
x_4	-0.233524	0.069708	-3.350037	0
x_6	-0.402295	0.014392	-27.95195	0

Table 8. Modified result 2 by weighted least square method
表 8. 加权最小二乘法修正结果 2

变量名	值
可决系数	0.999973
调整可决系数	0.999968
回归的标准误差	133.3699
残差平方和	391325.6
对数似然估计值	-206.665
F 统计量	200061.5

Continued

被解释变量的均值	11474.22
被解释变量的标准差	36182.94
施瓦茨准则	13.02968
Durbin-Watson (DW)统计量	0.469998

7. 序列自相关性

7.1. 自相关性分析——DW 检验

根据表 8 估计的结果得 $DW = 0.469998$ ，查 DW 表。因为样本量 n 为 27，解释变量为 x_1 、 x_6 、 x_3 和 x_4 ，所以 k 为 4，得下限临界值 $d_L = 1.16$ ，上限临界值 d_U 为 1.65。又因为统计量 $0 < 0.469998 = DW < d_L$ ，说明残差项存在正自相关。

7.2. 自相关性修正——一阶差分法

首先计算差分 $\Delta y_t = y_t - y_{t-1}$ ， $\Delta x_{it} = x_{it} - x_{i(t-1)}$ ，($i=1,3,4,6$)，用 Δy_t 对 Δx_{it} ($i=1,3,4,6$) 做过原点的最小二乘回归，来消除自相关，利用 SPSS 软件得到修正结果如表 9 和表 10 所示，DW 值从 0.469998 上升到 1.44634，即 $d_L < 1.44634 = DW < d_U$ ，消除了序列的自相关性。从而，最终得到消除了多重共线性、异方差性和自相关性的经验回归方程：

$$y = -6426.857 + 0.261690x_1 + 0.23927x_3 - 0.233524x_4 - 0.051258x_6$$

$R^2 = 0.999682$ ，修正的可决系数为 0.999603， $F = 12584.13$ ， $DW = 1.44634$ 。

Table 9. Autocorrelation correction result 1

表 9. 自相关性修正结果 1

	系数	标准误差	t	概率
(常量)	-6426.857	2408.353	-2.668569	0.0148
x_1	0.261690	0.022989	11.38344	0
x_3	0.23927	0.076105	3.152591	0.005
x_4	-0.233524	0.069708	-3.350037	0.0032
x_6	-0.051258	0.022989	11.38344	0

Table 10. Autocorrelation correction result 2

表 10. 自相关性修正结果 2

变量名	值
可决系数	0.999682
调整可决系数	0.999603
回归的标准误差	865.8781
残差平方和	14994899
对数似然估计值	-209.339
F 统计量	12584.13

Continued

被解释变量的均值	43000.08
被解释变量的标准差	43446.26
施瓦茨准则	16.85487
Durbin-Watson (DW)统计量	1.44634

8. 结论

修正后的经验回归模型为：

$$y = -6426.857 + 0.261690x_1 + 0.23927x_3 - 0.233524x_4 - 0.051258x_6$$

该模型最终修正了多重共线性、异方差性及序列自相关性等问题，提升了模型的精确度，并且使得模型作为一个整体，显着性提高，拟合度增强。

由此可得，假设在其他变量不变的情况下，国内生产总值每增长 1 亿元，税收收入会增长 0.261690 亿元；假设在其他变量不变的情况下，当财务开支增长 1 亿元，税收收入增加 0.23927 亿元；假设在其他变量不变的情况下，当消费品零售额增加 1 亿元，税收收入减少 0.233524 亿元；假设在其他变量不变的情况下，货币供应量增长 1 亿元，税收收入减少 0.051258 亿元。

通过上述的模型可以发现：一方面，国内生产总值、货币供应量、财务开支和消费品零售总额对税收有明显的影 响。另一方面，国民生产总值、全部固定资产投资和进出口总值对税收收入的影响不够明显，但是这并不能表明这 3 个因素对税收总量没有影响，因为国民生产总值、全部固定资产投资和进口和出口总值可以通过对国内生产总值、货币供应量、财务开支和消费品零售额的影响来对税收收入产生间接影响。

9. 建议

针对上述影响因素的分析，提出以下建议。

9.1. 优化产业结构

通过上述分析可知，税收与国内生产总值有着密切的关系，所以需要持续稳定地推进生产力发展，对产业结构进行调整优化，持续提升我国的经济发展水平和质量，为税收收入创造一个良好的发展环境。

9.2. 调节财政开支结构

由于税收的规模很大程度上受财政支出的制约，因此，政府要保持税收和财政支出之间的平衡和协调，积极调整和优化财政收支结构，使资源得到合理分配，如增加科技、民生等重点领域的支出、压缩一般支出，从严控制机关和事业单位的支出，强化绩效管理，使财政收支结构得到优化，实现税收的稳定增长。

9.3. 推动居民消费升级

消费对经济增长起到了支撑作用，推动居民消费升级，可以使消费对经济的作用更加明显。税收收入的增加是与经济发展密不可分的，通过提高居民消费水平来实现税收收入的稳定增长。伴随着时代的发展，人们对生活质量的要求也在不断提高，消费的方式以及产品的需求种类都发生了很大的变化。为此，企业应该积极开发出高质量、低成本、高效率、高效率的产品，以适应新的消费需要。

9.4. 调整、改革、完善税制

我国应在税制调整、税制改革和税制完善三个方面进行整合,并将三者有机地结合在一起,充分地发挥各自的功能。税收的增减都有可能与税收的调整和改进有关,因此,如果我们能更多地从完善税收制度的角度来考虑问题,就能在保证税收收入稳定的前提下,实现税收制度的科学化和合理化。

参考文献

- [1] 张霜. 税收结构是否影响经济增长——基于中国四个经济区域实证分析[C]//新兴经济体研究会,中国国际文化交流中心,广东工业大学. 开放合作命运与共论文集(上). 2020: 23.
- [2] 石镇国. 论税收对产业结构调整的影响[J]. 现代经济信息, 2019(8): 309.
- [3] 张松. 税收影响因素分析及建议[J]. 河北农机, 2019, 249(3): 110-111.
- [4] 方红丽. 我国税收收入影响因素分析[J]. 经济研究导刊, 2022, 501(7): 120-122+147.
- [5] 左雨婷, 武丽. 安徽省税收收入影响因素研究——基于 2001-2018 年数据的实证分析[J]. 赤峰学院学报: 自然科学版, 2020, 36(5): 66-71.
- [6] 高天惠, 张玉静, 朱家明, 崔文凯. 基于主成分分析对安徽省税收收入影响因素的实证研究[J]. 哈尔滨师范大学自然科学学报, 2020(5): 12-17.
- [7] 邓光青. 税收收入的影响因素研究——基于 SPSS 软件[J]. 纳税, 2020(24): 3-5+62.