北京航空航天大學

BEIHANG UNIVERSITY

数理统计课外论文

——基于逐步回归法的国家财政收入回归模型

学院：可靠性与系统工程学院

姓名：曹建钦

学号：20375177

摘要

国家财政收入是国家为了满足社会公共需要而对社会产品所进行的一种社会集中性分配行为，同时它本身也是一种社会宏观的公共管理活动。对财政收入的影响因素进行分析并建立回归模型，有助于对其预测从而更有效地调节资源配置、提高人民生活水平。本文利用逐步回归法及 SPSS 工具对可能影响财政收入的六个因素——工业总产值、农业总产值、建筑业总产值、社会商品零售总额、人口数和受灾面积进行分析并建立回归模型。

关键词：财政收入，最优回归方程，逐步回归，SPSS

Abstract

Financial revenue is a kind of social centralized distribution behavior of social products in order to meet the public needs of the society, and it is also a kind of macro public management activity. The analysis of the influencing factors of the financial revenue and the establishment of the regression model are helpful to predict the financial revenue and adjust the allocation of resources and improve the people's living standard. Using the method of stepwise regression and SPSS, this paper analyzes the six factors which may affect the financial revenue —— the total industrial output value, the total agricultural output value, the total construction output value, the total retail sales of social commodities, the number of population and the disaster area, then establishes the regression model

Key words: financial revenue, optimal regression equation, stepwise regression, SPSS

目录

图目录

表目录

引言

财政收入的影响因素有很多，例如工业总产值、建筑业总产值、社会商品零售总额等。有些因素影响很大而有些因素影响很小，为了找到“最优回归方程”，本文使用逐步回归法及 SPSS 数据分析工具建立起国家财政收入模型，通过此模型能较好地预测国家财政收入，从而对经济进行宏观调控，提供人民生活质量。

一、逐步回归法

（一）、原理

逐步回归的基本思想是通过剔除变量中不太重要又和其他变量高度相关的变量，降低多重共线性程度。将变量逐个引入模型，每引入一个解释变量后都要进行 F 检验，并对已经选入的解释变量逐个进行 t 检验，当原来引入的解释变量由于后面解释变量的引入变得不再显著时，则将其删除，以确保每次引入新的变量之前回归方程中只包含显著性变量。这是一个反复的过程，直到既没有显著的解释变量选入回归方程，也没有不显著的解释变量从回归方程中剔除为止，以保证最后所得到的解释变量集是最优的。

逐步回归法的好处是将统计上不显著的解释变量剔除，最后保留在模型中的解释变量之间多重共线性不明显，而且对被解释变量有较好的解释贡献。但是应特别注意，逐步回归法可能因为删除了重要的相关变量而导致设定偏误。

（二）、操作步骤

（预先给定 F_{out} 和 F_{in}，为了避免死循环，要求 $F_{in}\geqslant F_{out}$）

（1）对 p 个自变量分别与 y 建立回归模型$\hat{y} = \hat{\beta}_{i0}^{(0)} + \hat{\beta}_{i}^{(0)} x_i$，对它们分别计算 F_i，得 F_i 中最大的那个值，比如说 F_{L1}。

①如果 $F_{L1}<F_{in}$，则计算结束，即 y 与所有自变量均线性无关；

②如果 $F_{L1}\geq F_{in}$，则引入 x_{L1}，，建立回归方程

$$\hat{y} = \hat{\beta}_{0}^{(1)} + \hat{\beta}_{1}^{(1)} x_{L1}\ （1.1）$$

（2）建立 y 与自变量子集{ x_{L1} ，x_i }（$i \neq L_1$）的二元回归模型

$$\hat{y} = \hat{\beta}_{i0}^{(0)} + \hat{\beta}_{i1}^{(0)} x_{L1} + \hat{\beta}_i^{(0)} x_i (1.2)$$

将式（1.2）看作全模型，式（1.1）看作减模型求 F_i 值，并取 F_i 中最大的那个值，比如说 F_{L2}。

① 如果 $F_{L2} < F_{in}$，则计算结束，这时建立的回归模型为式（1.1）。

② 如果 $F_{L2} \geqslant F_{in}$，则引入 x_{L2}，，建立回归方程

$$\hat{y} = \hat{\beta}_0^{(2)} + \hat{\beta}_1^{(2)} x_{L1} + \hat{\beta}_2^{(2)} x_{L2} (1.3)$$

如果建立了模型(1.3)：

（3）当引入 x_{L2}，后，对 x_{L1} 做偏 F 检验，看 x_{L1} 是否需要剔除：

①如果 $F_{L1} > F_{out}$，则不剔除 xL1，，并继续引入下一个变量；

②如果 $F_{L1} \leqslant F_{out}$，则从式（1.3）中剔除 x_{L1}，再继续引入下一个变量。

重复上述步骤，直到所有模型外的变量都不能引入，模型内的变量都不能被剔除为止。

二、财政收入模型建立

（一）、原始数据

1、数据来源

本文从《中国统计年鉴 2021》中查找并整理了 1999 年到 2020 年国家财政收入（以一般公共预算收入总额替代）及其影响因素的相关数据信息，包括工业总产值、农业总产值（第一产业）、建筑业总产值、社会商品零售总额（社会消费品零售总额）、人口数和受灾面积。数据如图 1 所示，据此建立回归模型。

年份	国家财政收入（亿元）	工业总产值（亿元）	农业总产值（亿元）	建筑业总产值（亿元）	社会商品零售总额（亿元）	人口数（万人）	受灾面积（万公顷）
1999	11444.08	36014.4	14549	5180.9	35647.9	125786	49979.5
2000	13395.23	40258.5	14717.4	5534	39105.7	126743	54688
2001	16386.04	43854.3	15502.5	5945.5	43055.4	127627	52214.6
2002	18903.64	47774.9	16190.2	6482.1	48135.9	128453	46946.1
2003	21715.25	55362.2	16970.2	7510.8	52516.3	129227	54505.8
2004	26396.47	65774.9	20904.3	8720.5	59501	129988	37106.26
2005	31649.29	77958.3	21806.7	10400.5	67176.6	130756	38818.23
2006	38760.2	92235.8	23317	12450.1	76410	131448	41091.41
2007	51321.78	111690.8	27674.1	15348	89210	132129	48992.35
2008	61330.35	131724	32464.1	18807.6	114830.1	132802	39990.03
2009	68518.3	138092.6	33583.8	22681.5	132678.4	133450	47213.69
2010	83101.51	165123.1	38430.8	27259.3	156998.4	134091	37425.9
2011	103874.43	195139.1	44781.5	32926.5	183918.6	134916	32470.5
2012	117253.52	208901.4	49084.6	36896.1	210307	135922	24962
2013	129209.64	222333.2	53028.1	40896.8	242842.8	136726	31349.8
2014	140370.03	233197.4	55626.3	45401.7	271896.1	137646	24890.7
2015	152269.23	234968.9	57774.6	47761.3	300930.8	138326	21769.8
2016	159604.97	245406.4	60139.2	51498.9	332316.3	139232	26220.7
2017	172592.77	275119.3	62099.5	57906.6	366261.6	140011	18478.1
2018	183359.84	301089.3	64745.2	65493	380986.9	140541	20814.3
2019	190390.08	311858.7	70473.6	70648.1	408017.2	141008	19256.9
2020	182913.88	313071.1	77754.1	72995.7	391980.6	141212	19957.6

图 1 原始数据

2、变量命名

为了方便数学表达，将国家财政收入命名为因变量 Y，工业总产值等自变量用对应符号 X1，X2，X3，X4，X5，X6 命名，对应关系如表 2 所示。

表 1 符号对应关系

国家财政收入（亿元）	工业总产值（亿元）	农业总产值（亿元）	建筑业总产值（亿元）	社会商品零售总额（亿元）	人口数（万人）	受灾面积（万公顷）
Y	X1	X2	X3	X4	X5	X6

（二）、SPSS 辅助计算

为了方便统计量的计算及模型建立，这里使用 IBM SPSS Statistics26 进行逐步回归分析，具体方法如下：

① 导入原始数据 excel 表。

② 切换变量视图给自变量和因变量打上对应标签。（如图 2）

③ 依次点击“分析”-“回归”-“线性”从而打开线性回归功能。（如图 3）

④ 将 Y 选为因变量，X1~X6 选为自变量，方法选为“步进”即逐步回归。（如图 4）

⑤ 选择“统计”，确定置信水平，勾选个案诊断。（如图 5）

⑥ 选择“选项”，确定 F_{in} 和 F_{out}。（如图 6）

⑦ 完成设置，点击“确定”即可得到分析结果，本文第三部分即对结构的分析和讨论。

	名称	类型	宽度	小数位数	标签	值	缺失	列	对齐	测量	角色
1	年份	数字	4	0		无	无	12	右	标度	输入
2	国家财政收...	数字	9	2	Y	无	无	12	右	标度	输入
3	工业总产值...	数字	8	1	X1	无	无	12	右	标度	输入
4	农业总产值...	数字	7	1	X2	无	无	12	右	标度	输入
5	建筑业总产...	数字	7	1	X3	无	无	12	右	标度	输入
6	社会商品零...	数字	8	1	X4	无	无	12	右	标度	输入
7	人口数（万...	数字	6	0	X5	无	无	12	右	标度	输入
8	受灾面积（...	数字	8	2	X6	无	无	12	右	标度	输入

图 2 给变量打上标签

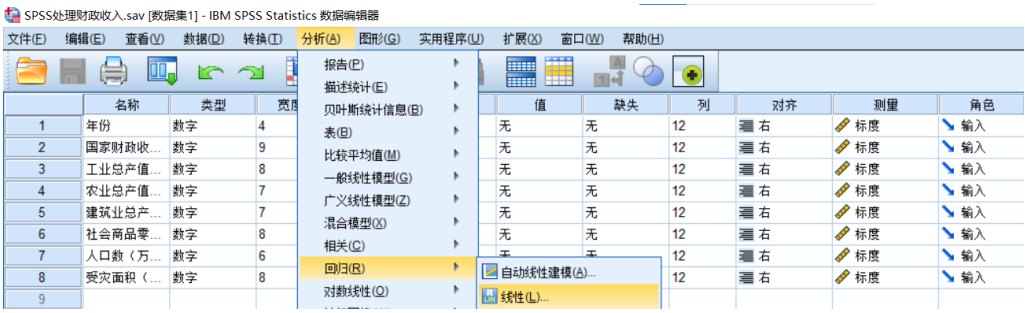

图 3 启用线性回归功能

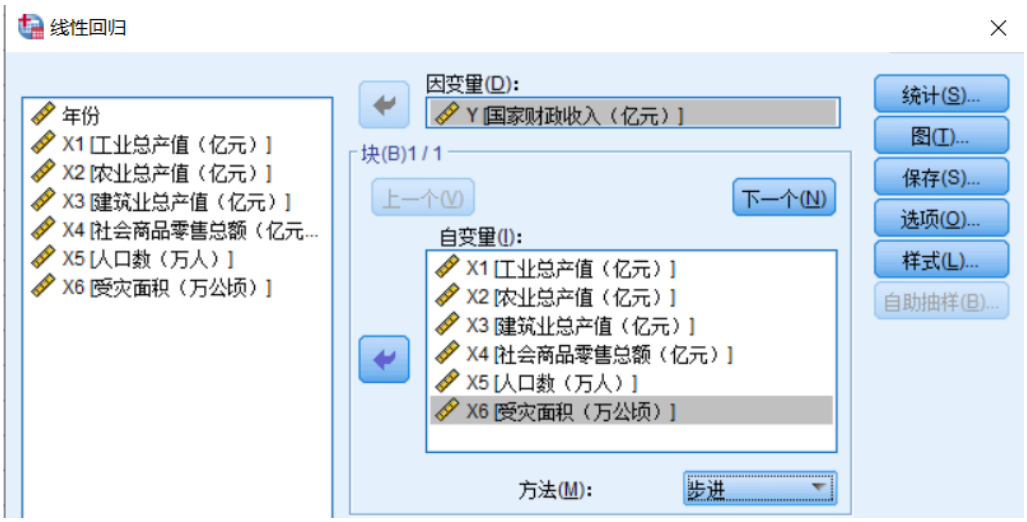

图 4 选定变量类型及逐步回归方法

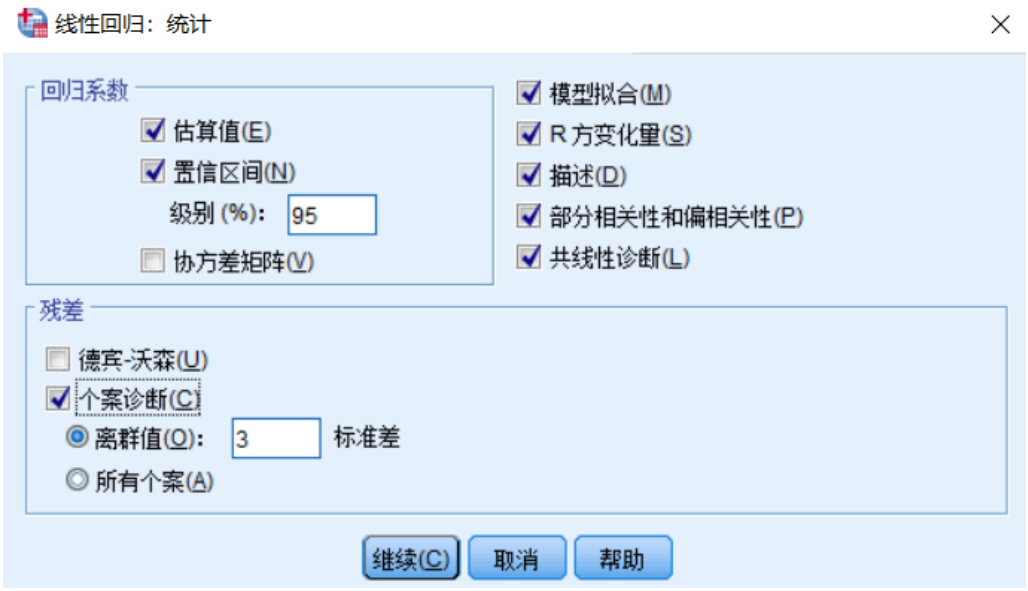

图 5“统计”中选择置信区间及个案诊断

图 6 确定 F_{in} 和 F_{out}

（三）、计算结果及模型

1、各 X（自变量）与 Y（因变量）间的 Person 相关性

表 2 各影响因素与因变量 Person 相关性

相关性							
		X1	X2	X3	X4	X5	X6
Pearson相关性	Y	.995	.993	.990	.994	.984	-.933

X1、X2、X3、X4、X5 和 X6 与 Y 之间的相关性如表 2 所示。其相关系数依次为 0.995，0.993，0.990，0.994，0.984，-0.993，同时显著性检验单尾 P 值（相关系数为 0 的概率）均为 0，初步无法排除变量。

2、逐步回归法筛选变量

表 3 输入/除去的变量

输入/除去的变量[a]			
模型	**输入的变量**	**除去的变量**	**方法**
1	X1	.	步进（条件：要输入的 F 的概率 <= .050，要除去的 F 的概率 >= .100）。
2	X4	.	步进（条件：要输入的 F 的概率 <= .050，要除去的 F 的概率 >= .100）。
3	X3	.	步进（条件：要输入的 F 的概率 <= .050，要除去的 F 的概率 >= .100）。

a. 因变量：Y

选定 F 概率小于等于 0.05 时输入，F 概率大于等于 0.1 时除去。结果如表 3 所示，可以得知系统在逐步分析时产生了 3 个模型，模型 1 为按照 F 检验将与 Y 关系最密切的变量 X1 引入而建立，之后引入 X4 同时 X1 未被剔除，生成模型 2，最后引入 X3 同时 X1 和 X4 未被剔除，生成模型 3。按照逐步回归法最后得到模型三，即最优回归方程含有的变量为 X1（工业总产值）、X4（社会商品零售总额）、X3（建筑业总产值）。

3、各模型拟合情况

表 4 模型摘要

模型摘要[d]									
					更改统计				
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
1	.995[a]	.989	.989	6834.29276	.989	1865.310	1	20	.000
2	.998[b]	.997	.996	3954.03076	.007	40.750	1	19	.000
3	.999[c]	.998	.998	2708.94739	.002	22.479	1	18	.000

a. 预测变量：（常量），X1

b. 预测变量：（常量），X1，X4

c. 预测变量：（常量），X1，X4，X3

d. 因变量:Y

为了保证建立的线性回归模型是“最优”的：一方面是该模型中包含所有

对因变量 Y 有显著性影响的自变量，另一方面是该模型中所包含的自变量个数尽可能地少，不含有无意义的变量，而且还应该是模型中 R^2 达到最大者，前面筛选后模型三有所有对 Y 有显著影响的自变量，同时表 4 中的数据也表示模型 3 中 R^2 最大，故可以认为模型 3 为 3 个模型中最好的，也验证了逐步回归法在寻找最优回归方程时的有效性。

4、方差分析

进一步对模型三进行方差分析（分析结果如表 5），可以看出其 F 值为 3993.885，并且显著性概率 Sig<0.001，可以认为回归效果是显著的，可以认为 Y 与 X1、X3、X4 之间有线性关系。

表 5 方差分析结果

ANOVA[a]						
模型		平方和	自由度	均方	F	显著性
1	回归	87124058046.657	1	87124058046.657	1865.310	.000[b]
	残差	934151149.233	20	46707557.462		
	总计	88058209195.891	21			
2	回归	87761156370.675	2	43880578185.338	2806.676	.000[c]
	残差	297052825.215	19	15634359.222		
	总计	88058209195.891	21			
3	回归	87926118068.863	3	29308706022.954	3993.885	.000[d]
	残差	132091127.027	18	7338395.946		
	总计	88058209195.891	21			

a. 因变量：Y

b. 预测变量：(常量)，X1

c. 预测变量：(常量)，X1，X4

d. 预测变量：(常量)，X1，X4，X3

5、回归方程

表 6 回归系数结果

系数a											
模型		未标准化系数		标准化系数	t	显著性	相关性			共线性统计	
		B	标准错误	Beta			零阶	偏	部分	容差	VIF
1	(常	-17733.3	2884.07		-6.149	0					

	量)										
	X1	0.667	0.015	0.995	43.189	0	0.995	0.995	0.995	1	1
2	(常量)	-9412.92	2117.332		-4.446	0					
	X1	0.352	0.05	0.526	7.042	0	0.995	0.85	0.094	0.032	31.394
	X4	0.233	0.036	0.477	6.384	0	0.994	0.826	0.085	0.032	31.394
3	(常量)	-13956.4	1738.558		-8.028	0					
	X1	0.451	0.04	0.673	11.247	0	0.995	0.936	0.103	0.023	42.937
	X4	0.413	0.046	0.846	9.074	0	0.994	0.906	0.083	0.01	104.385
	X3	-1.454	0.307	-0.517	-4.741	0	0.99	-0.745	-0.043	0.007	142.655

a.因变量：Y

根据表 6 可得到模型 3 回归方程系数，多元线性回归方程如下：

$$Y = -13956.4 + 0.451X_1 - 1.454X_3 + 0.413X_4$$

6、未引入变量的检验

SPSS 提供了 3 个模型各自舍弃变量的统计量计算结果（表 7 所示），从中可以看到模型 3 舍弃的 X2、X5 和 X6 的 P 值分别为 0.066，0.147 和 0.791，均大于 0.05，不能引入，故模型合理，六个影响因素作用时农业总产值、人口数和受灾面积对于财政收入的影响可以忽略。

表 7 舍弃变量的统计量

排除的变量[a]								
模型		输入 Beta	t	显著性	偏相关	共线性统计		
						容差	VIF	最小容差
1	X2	.324[b]	1.504	.149	.326	.011	93.064	.011
	X3	.310[b]	2.259	.036	.460	.023	42.904	.023
	X4	.477[b]	6.384	.000	.826	.032	31.394	.032
	X5	.038[b]	.246	.808	.056	.024	42.088	.024
	X6	-.066[b]	-1.068	.299	-.238	.138	7.239	.138
2	X2	.018[c]	.125	.902	.029	.009	108.421	.009
	X3	-.517[c]	-4.741	.000	-.745	.007	142.655	.007
	X5	.053[c]	.600	.556	.140	.024	42.119	.013
	X6	-.036[c]	-.993	.334	-.228	.136	7.372	.029
3	X2	.184[d]	1.966	.066	.430	.008	121.210	.006
	X5	-.098[d]	-1.521	.147	-.346	.019	53.583	.006
	X6	-.007[d]	-.269	.791	-.065	.127	7.861	.007

a. 因变量：Y

b. 模型中的预测变量：(常量), X1

c. 模型中的预测变量：(常量), X1, X4

d. 模型中的预测变量：(常量), X1, X4, X3

结论

可建立以下财政收入的回归模型：

（Y：财政收入；X1：工业总产值；X3 建筑业总产值；X4 社会商品零售总额）

$$Y = -13956.4 + 0.451X_1 - 1.454X_3 + 0.413X_4$$

从回归方程中可以看出 1999-2020 年之间工业总产值对国家财政收入影响非常大，同时，建筑业总产值和社会商品零售总额也有不可忽视的影响。而农业总产值、受灾面积和人口数对其影响不大，这说明了我国在 1999 年-2020 年间工业、建筑业和零售业飞速发展，在国家财政收入中占了相当大的份额，这也是符合 1999-2020 年国家发展战略的。受灾面积对财政收入影响很小也说明了国家变得越来越富强，抵抗风险的能力很高。这些都说明我国处于经济飞速发展的阶段，正向社会主义现代化国家迈进。

当然，本模型也存在一些问题和可以完善的部分，一方面国家财政收入同国家政策和社会发展阶段有着千丝万缕的关系，因此，如果截取的时间段发生改变，其模型也可能发生变化；另一方面，虽然线多元性回归方程拟合很成功，但不排除非线性拟合比线性拟合更优的可能性。

参考文献

[1] 中华人民共和国统计局.中国统计年鉴[M].北京：中国统计出版社,2021

[2] 孙海燕，周梦，李卫国，冯伟.数理统计[M].北京：北京航空航天大学出版社，2016.10

[3] 钟海燕，殷锋.IBM SPSS 统计分析与应用[M].北京：中国经济出版社，2018.09