Statistics and Application 统计学与应用, 2019, 8(4), 581-588
Published Online August 2019 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2019.84066

Hans汉斯

Influencing Factors and Regression Analysis of GDP in Yunnan Province

Xinxin Hu

School of Statistics and Mathematics, Yunnan University of Finance and Economics University, Kunming Yunnan
Email: wodeyueding2012@163.com

Received: Jun. 27th, 2019; accepted: Jul. 13th, 2019; published: Aug. 2nd, 2019

Abstract

Based on the statistical yearbook of GDP from 2007 to 2016 in Yunnan province and the related data, using linear regression method, this paper sets up the fitting model to describe the relationship between GDP and related variables in Yunnan province. The heteroscedasticity test, sequence autocorrelation test and abnormal point test for the model are also carried out. The results show that this model can be used to predict the gross domestic product of Yunnan province.

Keywords

GDP, Influencing Factors, Linear Regression Model

云南地区生产总值影响因素和回归分析

胡欣欣

云南财经大学统计与数学学院，云南 昆明
Email: wodeyueding2012@163.com

收稿日期：2019年6月27日；录用日期：2019年7月13日；发布日期：2019年8月2日

摘 要

本文基于统计年鉴中云南省2007~2016年生产总值和与之相关的数据，运用线性回归方法，建立了用于描述云南省地区生产总值与相关变量之间定量关系的拟合模型，并对模型进行了异方差检验、序列自相关检验和异常点的检验。该模型对于云南省地区生产总值的预测有一定的研究作用。

文章引用: 胡欣欣. 云南地区生产总值影响因素和回归分析[J]. 统计学与应用, 2019, 8(4): 581-588.
DOI: 10.12677/sa.2019.84066

关键词

生产总值，影响因素，线性回归模型

1. 引言

地区生产总值是指地区生产总值(地区 GDP)是指本地区所有常住单位在一定时期内生产活动的最终成果。地区生产总值等于各产业增加值之和。地区生产总值是衡量一个地区发展情况的良好尺度，本文使用线性回归模型对云南省地区生产总值的影响因素进行实证研究，通过变量选择方法，筛选得到了对地区生产总值具有显著影响的因素，并建立了拟合模型，该模型通过了异方差性检验。

2. 数据来源与变量选择

2.1. 数据来源

本数据来源于国家统计局网(http://data.stats.gov.cn/easYquerY.htm?cn=E0103)上公布的 2007~2016 的相关数据。

2.2. 变量选择

本文的地区生产总值的影响因素的研究主要考察在众多因素中哪些因素对生产总值有显著的影响。此处首先给出自变量的待选变量集。经查阅资料，此处将城镇单位就业人员工资，全社会固定资产投资总额，地方财政一般预算收入和工业增加值、农林牧业增加值以及建筑业增加值引入待选变量集中，此外由于昆明作为春城花都，常年吸引世界各地的游客前来游玩，故将国际旅游外汇收入也引入待选变量集中。综上，此处选取地方财政一般预算收入(亿元)、全社会固定资产投资总额[1] (亿元)、城镇单位就业人员工资总额(亿元)、工业增加值(亿元)、农林牧业增加值(亿元)、建筑业增加值(亿元)、国际旅游外汇收入(亿元) (为了统一数量级，此处将统计年鉴中的“百万美元”单位换算为“亿元”)为自变量，以地区生产总值(亿元)为响应变量。

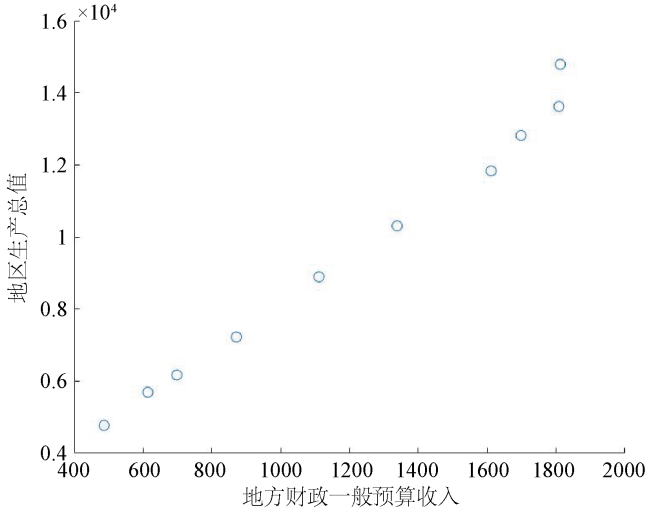
3. 建立模型

3.1. 模型估计[2]

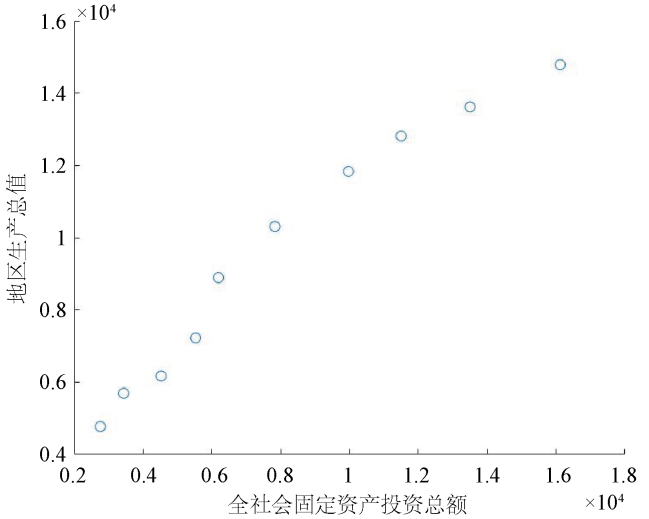
首先，绘制出变量 $X_i\,(i=0,\cdots,7)$和 Y 之间的散点图(见图 1)，观察解释变量与响应变量之间的关系。通过散点图可以初步发现，解释变量 X_i 与生产总值 Y 大致成线性正向影响关系。Y 与 X 之间的 pearson 相关系数分别为 0.9936，0.9798，0.9938，0.9726，0.9968，0.9911，0.9834。

3.2. 用普通最小二乘法(OLS)估计模型[3]

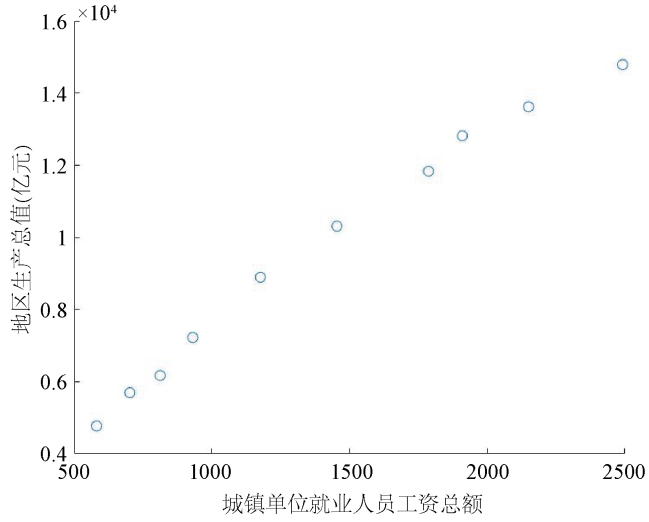
为了进一步分析解释变量 X_i 对生产总值 Y 的影响，本文采用多元线性回归模型对变量之间的关系进行验证。此处建立云南省地区生产总值影响因素分析的七元回归预测模型：


×10⁴
1.6
1.4
1.2
1
0.8
0.6
0.4
地区生产总值
400
600
800
1000
1200
1400
1600
1800
2000
地方财政一般预算收入

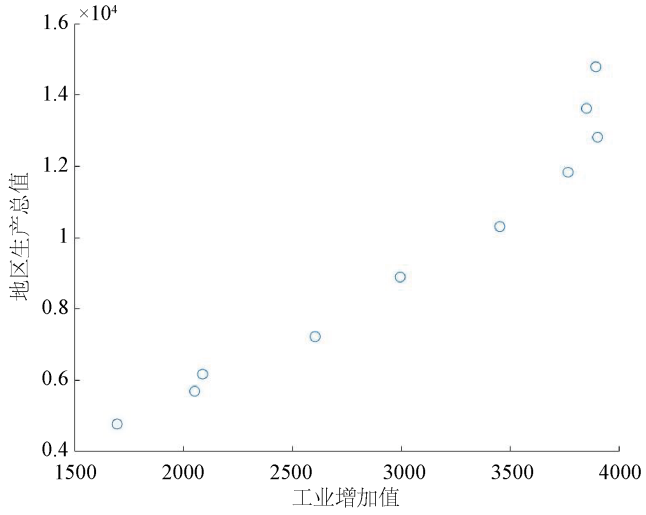
(1)


×10⁴
1.6
1.4
1.2
1
0.8
0.6
0.4
地区生产总值
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
全社会固定资产投资总额
×10⁴

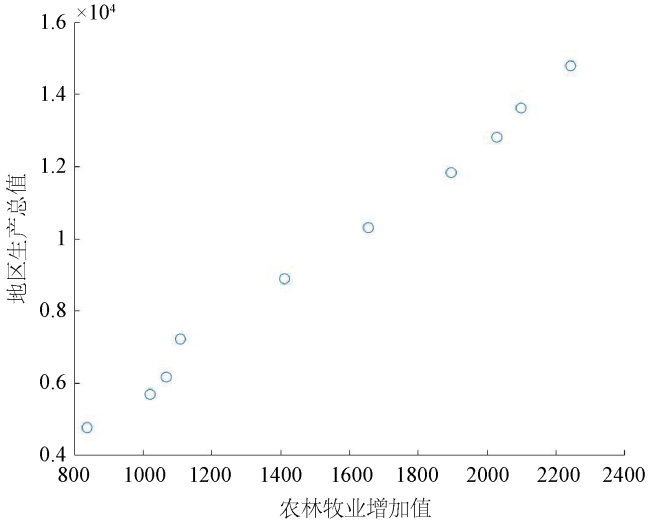
(2)


×10⁴
1.6
1.4
1.2
1
0.8
0.6
0.4
地区生产总值(亿元)
500
1000
1500
2000
2500
城镇单位就业人员工资总额

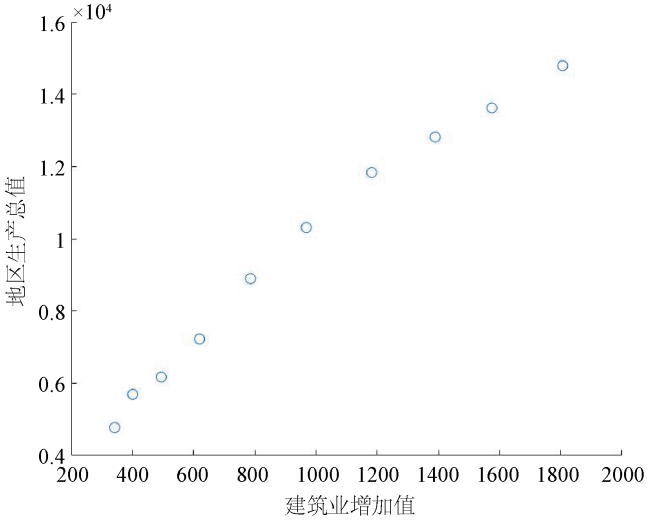
(3)


×10⁴
1.6
1.4
1.2
1
0.8
0.6
0.4
地区生产总值
1500
2000
2500
3000
3500
4000
工业增加值

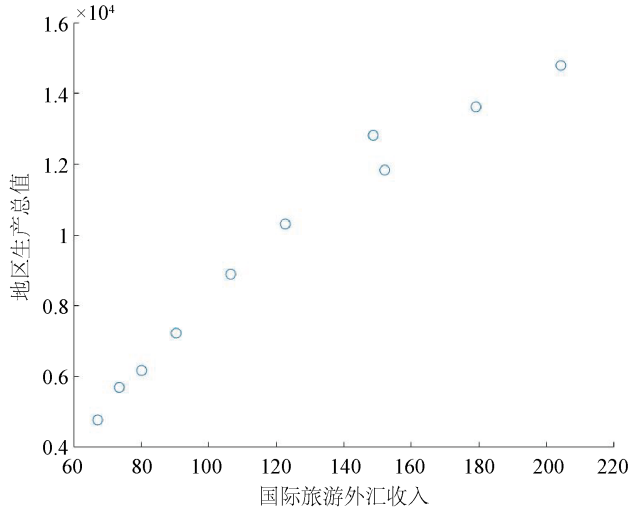
(4)


×10⁴
1.6
1.4
1.2
1
0.8
0.6
0.4
地区生产总值
800
1000
1200
1400
1600
1800
2000
2200
2400
农林牧业增加值

(5)


×10⁴
1.6
1.4
1.2
1
0.8
0.6
0.4
地区生产总值
200
400
600
800
1000
1200
1400
1600
1800
2000
建筑业增加值

(6)



(7)

Figure 1. The scatter plot between the variable X_i ($i = 0,\cdots,7$) and Y((1)~(7))
图 1. 散点图((1)~(7))

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \mu$$

其中，X_1 系地方财政一般预算收入(亿元)、X_2 系全社会固定资产投资总额(亿元)、X_3 系城镇单位就业人员工资总额(亿元)、X_4 系工业增加值(亿元)、X_5 系农林牧业增加值(亿元)、X_6 系建筑业增加值(亿元)、X_7 系国际旅游外汇收入(亿元），Y 系地区生产总值(亿元)。β_i ($i = 0,\cdots,7$)为各解释变量对应的参数，μ 为随机误差项。回归方程的参数估计值及检验结果如表 1 所示：

Table 1. The result
表 1. 检验结果

B	−592.0935638		**R**	−28.82337198		**STATS**	0.999865087
	−1.003571793			−48.71111925			2117.481046
	−0.03798688			64.14180971			4.72E−04
	−1.586163353			−18.363136			7783.073529
	1.439753515			81.93407372			
	2.271980028			−24.36149496			
	4.253160936			−10.50518321			
	16.51859195			−11.55886781			
BINT	−5777.876278	4593.68915		−15.1716182			
	−6.531785918	4.524642332		11.41890798			
	−0.782036473	0.706062713	**RINT**	−159.9517622	102.3050182		
	−15.80920849	12.63688178		−316.6328979	219.2106594		
	−0.33337036	3.212877391		−400.803646	529.0872654		
	−5.188910392	9.732870447		−188.9341391	152.2078671		
	−2.762196045	11.26851792		−231.2976245	395.1657719		
	−60.47366393	93.51084783		−1136.258304	1087.535315		
				−577.3128279	556.3024615		
				−522.2354876	499.117752		
				−514.2038755	483.8606391		
				−104.1184798	126.9562958		

所得到的模型为

$$hY = -592.094 - 1.004X_1 - 0.038X_2 - 1.586X_3 + 1.440X_4 + 2.272X_5 + 4.253X_6 + 16.519X_7 + \mu$$

线性方程的回归检验的 P 值为 $0.00047 \ll 0.5$，R^2 为 0.999，这意味着在 5%的显著性水平下，因变量与自变量之间的线性关系是显著的。而在系数的 t 检验中，p 值最小的是 0.155，故在 5%显著性水平下所有系数均不显著，即每个解释变量对被解释变量的线性影响均不是显著的[4]。这说明模型自变量之间很可能存在多重共线性。T 检验中的解释变量都不显著，可能是由于某些自变量对因变量的影响被其他自变量所掩盖。为了检验多重共线性[5]的存在，进一步对各变量之间的相关关系进行分析研究，结果如表 2 所示：

Table 2. System resulting data of standard experiment
表 2. 标准试验系统结果数据

	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7
Y	1.0000	0.9936	0.9798	0.9938	0.9726	0.9968	0.9911	0.9834
X_1	0.9936	1.0000	0.9556	0.9777	0.9878	0.9935	0.9729	0.9637
X_2	0.9798	0.9556	1.0000	0.9940	0.9104	0.9706	0.9967	0.9933
X_3	0.9938	0.9777	0.9940	1.0000	0.9428	0.9894	0.9978	0.9961
X_4	0.9726	0.9878	0.9104	0.9428	1.0000	0.9725	0.9346	0.9221
X_5	0.9968	0.9935	0.9706	0.9894	0.9725	1.0000	0.9824	0.9763
X_6	0.9911	0.9729	0.9967	0.9978	0.9346	0.9824	1.0000	0.9929
X_7	0.9834	0.9637	0.9933	0.9961	0.9221	0.9763	0.9929	1.0000

由表可以看出，各变量之间的确存在一定的线性关系。对七个自变量采用逐步回归的方法进行变量筛选[6]，得到的结果如图 2：

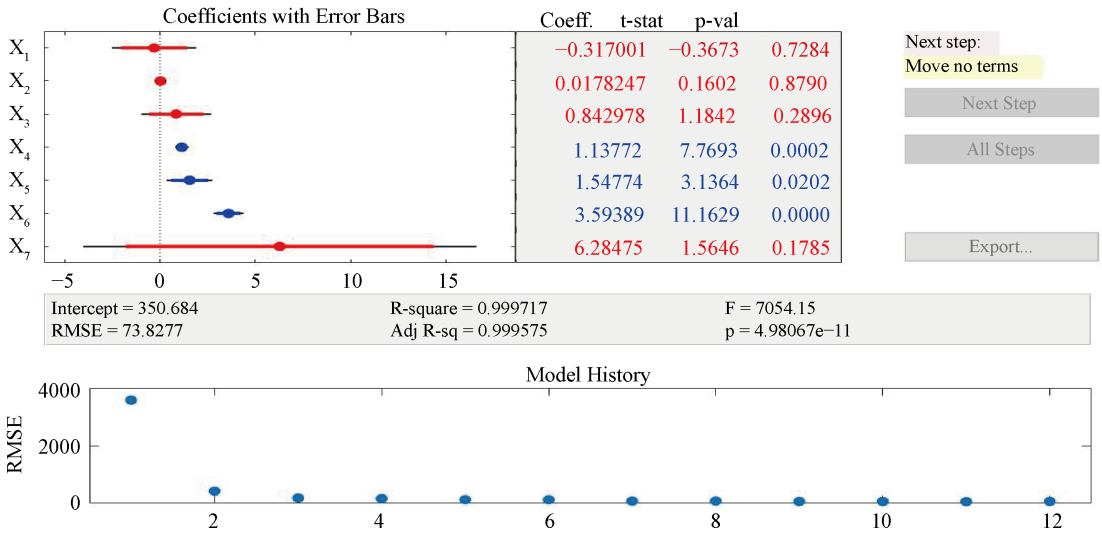


Figure 2. The result of stepwise
图 2. 逐步回归结果

逐步回归[7]的结果显示选择的自变量应当为工业增加值(X_4)，农林牧业增加值(X_5)，建筑业增加值(X_6)。在5%的显著性水平下，他们的P值分别为0.0002、0.0202和0.0000，表明这三个解释变量对模型方程的影响是显著的。

为了进一步确证变量选择结果，此处使用AIC准则[8]对一些重点待选模型进行比较。比较结果见表3。

Table 3. The model selection
表3. 模型选择

方程参数个数	0	1	2	3
AIC	10.1881	10.3881	10.5881	10.7881

由表可见，AIC准则提供的变量选择的结果与逐步回归法一致，均选择X_4，X_5和X_6。

将模型方程进行二次拟合，结果如表4所示

Table 4. The result of refit
表4. 二次拟合结果

B	350.6842228		R	−33.52757785		STAT	0.99971655
	1.137715241			−13.63017709			7054.1519
	1.547737596			14.28206732			4.98E−11
	3.593891533			−30.80403968			5450.522315
				127.0311839			
				−8.572269564			
				18.06150944			
				−105.0181031			
				−16.75386847			
				48.93127516			
BINT	−67.8708	769.2392	RINT	−186.859	119.8043		
	0.779395	1.496036		−176.098	148.8378		
	0.340233	2.755242		−157.319	185.8833		
	2.806107	4.381676		−108.111	46.50304		
				9.489593	244.5728		
				−182.997	165.8529		
				−145.033	181.1563		
				−240.495	30.45866		
				−190.971	157.4633		
				−57.0083	154.8708		

故得出模型方程为：$hY = 350.684 + 1.138X_4 + 1.548X_5 + 3.594X_6$。

3.3. 异方差性检验

a. 残差图分析法[9]

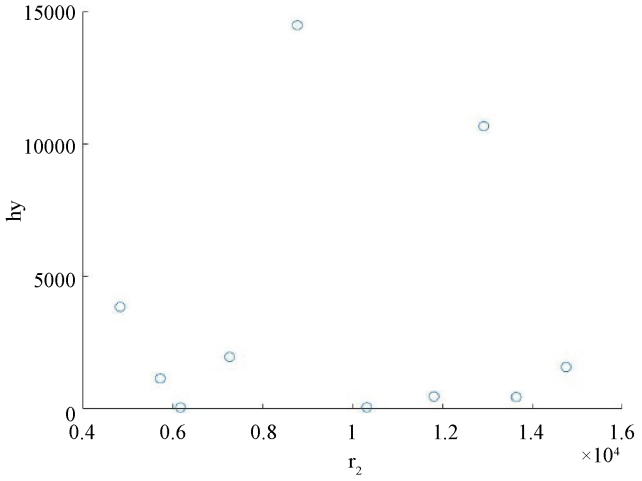


Figure 3. The residual plot
图 3. 残差图

由异方差性检验原理可知，当线性回归模型满足其假设条件时，即模型中不存在明显的异方差性，残差图上的 n 个数据点的散布应该是随机的，无任何规律。观察图 3 可得，数据点的分布较为随机，故此模型不存在明显的异方差性。

b. 斯皮尔曼(Spearman)检验[10]

由 MATLAB 程序运算可得，P 值为 0.8810，大于 0.5，故模型不存在异方差。

4. 结论

经过上述分析，我们建立了 $hY = 350.684 + 1.138X_4 + 1.548X_5 + 3.594X_6$ 这样一个方程模型。从这个模型来看，在地方财政一般预算收入、全社会固定资产投资总额、城镇单位就业人员工资总额、工业增加值、农林牧业增加值、建筑业增加值和国际旅游外汇收入这些自变量中，对地区生产总值影响最为显著的是工业增加值、农林牧业增加值和建筑业增加值。工业增加值、农林牧业增加值、建筑业增加值均与地区生产总值成正相关。这说明工业、农业和建筑业对地区生产总值的提高具有积极作用，这与我们的常识了解也是相一致的。欲提高一个地区的生产总值，应大力促进其工农建三方面产业的发展。

参考文献

[1] 陈静. 我国各地区生产总值的影响因素分析及建议[J]. 商, 2015(7): 150-151.
[2] 杨武. 安徽省国内生产总值影响因素的多元回归分析[J]. 南方农机, 2018, 49(5): 11-13.
[3] 李实. 云南省生产总值影响因素实证分析[J]. 中国市场, 2011(31): 105-107.
[4] 朱琳, 陈飞. 云南失业率影响因素分析和回归诊断[J]. 当代经济, 2013(4): 90-91.
[5] 王雪雪. 我国地区生产总值影响因素的实证分析[J]. 时代金融, 2017(15): 18-22.
[6] 单翔翔, 严浩坤. 基于多元回归模型分析我国国内生产总值的影响因素[J]. 时代金融, 2018(9): 238-239.
[7] 吴喜之. 应用回归及分类[M]. 北京: 中国人民大学出版社, 2016: 49-51.
[8] 王燕. 应用时间序列[M]. 北京: 中国人民大学出版社, 2012: 82-83.
[9] 唐年胜, 李会琼. 应用回归分析[M]. 北京: 科学出版社, 2014: 114-115.
[10] 杨林涛. 非参数统计视角下的异方差检验设计及其应用[J]. 数量经济技术经济研究, 2014, 31(11): 118-131.

知网检索的两种方式：

1. 打开知网首页：http://cnki.net/，点击页面中“外文资源总库 CNKI SCHOLAR”，跳转至：http://scholar.cnki.net/new，搜索框内直接输入文章标题，即可查询；
   或点击“高级检索”，下拉列表框选择：[ISSN]，输入期刊 ISSN：2325-2251，即可查询。
2. 通过知网首页 http://cnki.net/顶部“旧版入口”进入知网旧版：http://www.cnki.net/old/，左侧选择“国际文献总库”进入，搜索框直接输入文章标题，即可查询。

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：sa@hanspub.org