Statistics and Application 统计学与应用, 2020, 9(2), 296-303
Published Online April 2020 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2020.92032

Hans汉斯

Statistical Diagnosis Analysis of Hangzhou's GDP Data

Danni Lu, Hongfeng Song, Dengke Xu

Department of Statistics, Zhejiang Agriculture and Forestry University, Hangzhou Zhejiang
Email: 175384319@qq.com

Received: Mar. 29th, 2020; accepted: Apr. 12th, 2020; published: Apr. 20th, 2020

Abstract

Gross domestic product (GDP) is often recognized as the best indicator to measure the economic situation of a country, which can reflect the economic strength and market size of a country or region. In order to explore the main factors affecting Hangzhou's GDP, this paper selects several national economic indicators of Hangzhou from 2000 to 2018, and establishes a regression model based on multiple linear regression. Then the stepwise regression method is used to screen variables, and statistical diagnosis is carried out for the regression model. The strong influence points in the model are screened under certain standards. Based on the data deletion model after deleting the strong influence points, the stepwise regression analysis is carried out again, and the conclusion is that the GDP of Hangzhou is closely related to the year-end resident population, fixed asset investment, resident consumption index and total fiscal revenue.

Keywords

GDP, Multivariate Linear Regression, Multicollinearity, Statistical Diagnosis, Cook Distance, W-K Statistics

杭州市GDP数据的统计诊断分析

陆丹妮，宋红凤，徐登可

浙江农林大学统计系，浙江 杭州
Email: 175384319@qq.com

收稿日期：2020年3月29日；录用日期：2020年4月12日；发布日期：2020年4月20日

摘　要

国内生产总值(GDP)常被公认为是衡量国家经济状况的最佳指标，可以反映一个国家或地区的经济实力和市场规模。为探索影响杭州市GDP的主要因素，本文选择杭州市2000~2018年若干国民经济指标，基

文章引用: 陆丹妮, 宋红凤, 徐登可. 杭州市 GDP 数据的统计诊断分析[J]. 统计学与应用, 2020, 9(2): 296-303.
DOI: 10.12677/sa.2020.92032

于多元线性回归建立回归模型。然后利用逐步回归法进行变量筛选，并对回归模型进行统计诊断，在一定标准下筛选模型中的强影响点，基于删除强影响点后的数据删除模型再次进行逐步回归分析，得到杭州市GDP与年末常住人口、固定资产投资额、居民消费指数以及财政总收入密切相关。

关键词

GDP，多元线性回归，多重共线性，统计诊断，Cook距离，W-K统计量

1. 引言

GDP 是国民经济核算的核心指标，是反映各地区经济实力和国民生活水平的指标，受到人们的广泛关注，GDP 数据质量的提高，对调控宏观经济有很大的促进作用。研究影响 GDP 变化的主要因素，建立一个可靠合理的模型，有助于找到提高国民经济水平和地区发展的关键因素，对影响 GDP 的因素进行定量分析和预测具有重要意义，促进 GDP 不断增长。

郭芳、冷洛[1]通过计量分析，利用回归分析发现影响中国国内生产总值的主要因素是最终消费支出和资本形成总额，即内因是主要因素。文静[2]通过对国内生产总值的变动进行多因素分析，建立多元线性回归模型，提出国家在注重财政支出带来的国内生产总值增加的同时应该关注外资的利用情况。张卜元、刘冰冰、张东旭[3]采取多元回归方法进行模型设定，并进行自相关、异方差及多元共线性检验及修正，得出我国最终消费和资本形成对 GDP 有重大影响。程静[4]通过各种因素对经济增长的作用来进行定量分析，指出能源消耗总量对国内生产总值有影响，并提出科技进步对农村建设的重要性。单翔翔、严浩坤[5]基于多元回归模型，并进行了异方差的检验修正，得出国内生产总值主要受到税收、城乡储蓄存款、财政支出和固定资产投资总额的影响。多元线性回归法显然是研究区域 GDP 及其影响因素的一种可行方案，对于探索影响 GDP 的关键因素有良好的应用价值[5] [6] [7] [8] [9]。

杭州市作为新一线城市，GDP 始终保持稳步增长状态，在全国也处于领先地位。近年来杭州市农业生产形势稳定，优势特色产业增速回升；工业生产总体平稳，新制造业动能增强；服务业态势良好，现代服务业贡献突出；固定资产投资增长快速，商品销售增长稳定，外贸多元发展，居民收入增加，就业社保也不断扩大；另外杭州的优势产业数字经济持续引领 GDP 的增长。本文基于多元线性回归模型，深入研究影响杭州市 GDP 变化的主要因素，并利用统计诊断提高模型的精度，更加准确地剖析杭州市国内生产总值变化的主要原因。

2. 数据的收集和杭州市 GDP 变化的基本情况

本文采用的数据源自浙江省统计信息网及杭州市统计局提供的统计年鉴，从网站中获得杭州市 2000~2018 年的国内生产总值数据及影响 GDP 变化的各项指标数据。以国内生产总值(亿元)为被解释变量，下面五个指标为解释变量：年末常住人口(万人)、全社会就业人员(万人)、固定资产投资(亿元)、居民消费指数(统计年鉴中以 1978 年的居民消费指数为基数 100 进行度量)和财政总收入(亿元)。对此建立多元线性回归模型并进行统计诊断。

图 1 是 2000~2018 年杭州市 GDP 变化情况折线图，折线图显示杭州市 GDP 呈现稳步上升趋势，且上

升幅度较大。从 2000 年的 1382.56 亿元到 2018 年的 13,509.15 亿元，翻了将近十倍。说明杭州市的经济发展水平和国民生活水平较高，位居全国领先位置，根据杭州市各产业的发展状况，GDP 指数仍将继续上升。

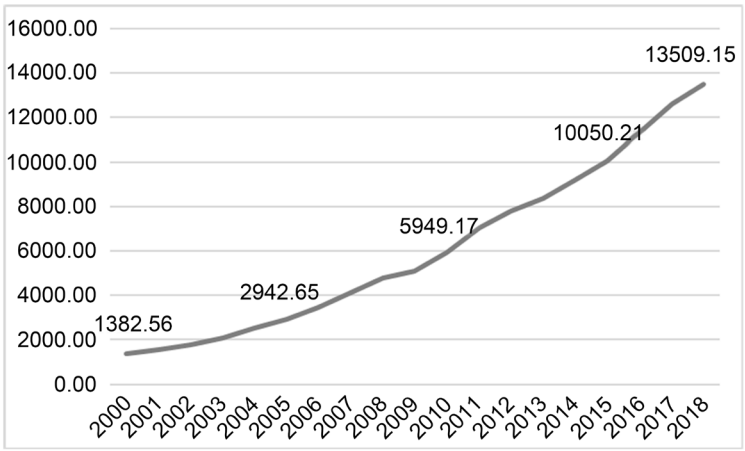

Figure 1. GDP change of Hangzhou in 2000~2018 (100 million yuan)
图 1. 2000~2018 年杭州市 GDP 变化情况(亿元)

3. 模型建立

多元线性回归分析

基于 2000~2018 年杭州市反映国民经济水平(GDP)的相关指标数据，以杭州市生产总值(GDP)为因变量(Y)，以年末常住人口(X1)、全社会就业人数(X2)、固定资产投资(X3)、居民消费指数(X4)、财政总收入(X5)为自变量，拟合多元线性回归模型，得到如下结果。

Table 1. Multiple linear regression results
表 1. 多元线性回归结果

Model	Unstandardized Coefficients		Standardized	*t*	Sig.
	B	Std. Error	Beta		
(常量)	−8740.869	1997.825		−4.375	**0.001**
年末常住人口(*X*1)	2.909	1.008	0.080	2.675	**0.019**
全社会就业人数(*X*2)	3.702	1.008	0.096	3.675	**0.003**
固定资产投资(*X*3)	0.476	0.120	0.253	3.966	**0.002**
居民消费指数(*X*4)	10.702	2.977	0.249	3.595	**0.003**
财政总收入(*X*5)	1.352	0.343	0.342	3.939	**0.002**

a. 因变量：GDP， $R^2 = 0.999$， $R_a^2 = 0.999$，Prob(F) = 0.000，**AIC = 189.35**

表 1 是根据原始数据建立的多元回归模型，从回归结果可知多元回归模型通过了 F 检验，并且 5 个自变量的回归系数都是显著的，因此可以得到多元回归模型为：

$$Y = 2.909X_1 + 3.702X_2 + 0.476X_3 + 10.702X_4 + 1.352X_5 - 8740.869 \quad (1)$$

由回归方程可知，5 个自变量与因变量之间均呈现正线性关系，年末常住人口增加，全社会就业人数增加，固定资产投资提高，居民消费指数提高，财政总收入增加，都会带动 GDP 的增长，这些因素都是影

响 GDP 的关键因素。为了进一步提高模型的精度，需要对模型进行统计诊断，找到原始数据中的强影响点，并对模型进行改进。

4. 统计诊断

4.1. 影响分析

4.1.1. 基于 Cook 距离的统计诊断[10]

对给定的多元线性回归模型：$Y=X\beta+\varepsilon, E(\varepsilon)=0, Var(\varepsilon)=\sigma^2 I_n$ 和删除第 i 个数据点以后的模型：$Y_{(i)}=X_{(i)}\beta+\varepsilon_{(i)}, \varepsilon_{(i)}\sim N(0,\sigma^2 I_{n-1})$，度量第 i 个数据点 (x_i^T, y_i) 对参数 β 的估计量影响大小的 Cook 距离定义为：

$$CD_i=\frac{\left(\hat{\beta}-\hat{\beta}_{(i)}\right)^{\mathrm{T}} X^{\mathrm{T}} X\left(\hat{\beta}-\hat{\beta}_{(i)}\right)}{(p+1)\hat{\sigma}^2}$$

其中 $X=(\tilde{x}_1,\cdots,\tilde{x}_n)$ 为已知的列满秩矩阵，$\tilde{x}_i=(1,x_i^T)^T, x_i=(x_{i1},\cdots,x_{ip})^T$，$X_{(i)}$ 是删掉矩阵 X 的第 i 行向量 $\tilde{x}_i$ 后得到的矩阵；$\hat{\beta}$ 是原多元线性回归模型中参数 β 的估计，$\hat{\beta}_{(i)}$ 是数据删除模型中参数 β 的估计；p 是自变量个数。在给定模型下，Cook 距离也可以简化为：$CD_i=\frac{h_{ii}}{1-h_{ii}}\frac{r_i^2}{p+1}=\frac{h_{ii}}{1-h_{ii}}\frac{n-p-1}{p+1}b_i$，其中 $b_i=\frac{r_i^2}{n-p-1}\sim \mathrm{Beta}\left(\frac{1}{2},\frac{n-p-2}{2}\right)$，所以 Cook 距离不仅与残差 r_i 的大小有关，还与杠杆值 h_{ii} 的大小有关，它度量了 β 和 σ^2 的估计量之间的距离，因此残差和杠杆值相对大的点很有可能是强影响点。

4.1.2. 基于 W-K 统计量的统计诊断[10]

对给定的多元线性回归模型：$Y=X\beta+\varepsilon, E(\varepsilon)=0, Var(\varepsilon)=\sigma^2 I_n$ 和删除第 i 个数据点以后的模型：$Y_{(i)}=X_{(i)}\beta+\varepsilon_{(i)}, \varepsilon_{(i)}\sim N(0,\sigma^2 I_{n-1})$，第 i 个数据点 (x_i,y_i) 删除前后对 $\tilde{x}_i$ 处拟合值的影响可以定义为：

$$WK_i=\frac{\hat{y}_i-\tilde{y}_{(i)}}{\hat{\sigma}_{(i)}\sqrt{h_{ii}}}=\frac{\tilde{x}_i\left(\hat{\beta}-\hat{\beta}_{(i)}\right)}{\hat{\sigma}_{(i)}\sqrt{h_{ii}}}$$

其中 $X=(\tilde{x}_1,\cdots,\tilde{x}_n)$ 为已知的列满秩矩阵，$\tilde{x}_i=(1,x_i^T)^T, x_i=(x_{i1},\cdots,x_{ip})^T$，$X_{(i)}$ 是删掉矩阵 X 的第 i 行向量 $\tilde{x}_i$ 后得到的矩阵；$\hat{\beta}$ 是原多元线性回归模型中参数 β 的估计，$\hat{\beta}_{(i)}$ 和 $\hat{\sigma}_{(i)}$ 是数据删除模型中参数 β 和 σ 的估计；h_{ii} 是杠杆值的大小。对于本文的线性模型，W-K 统计量可以表示为：

$$WK_i=\sqrt{\frac{h_{ii}}{1-h_{ii}}}t_i,\ (WK_i)^2=CD_i\left(X^{\mathrm{T}}X,\hat{\sigma}_{(i)}^2\right),\ t_i=\frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}}$$

WK_i 度量了 $(\hat{\beta},\hat{\sigma}^2)$ 与 $(\hat{\beta}_{(i)},\hat{\sigma}_{(i)}^2)$ 之间的差异，比 Cook 距离更加通用，也更容易地被应用到线性模型以外的各种更复杂的统计模型。

4.2. 实例研究

利用 R 语言软件计算出所有样本点的 Cook 距离和 W-K 统计量，并制作折线图，图 2 是 Cook 距离散点图，图 3 是 W-K 统计量散点图，在一定标准下筛选出强影响点，并在图中标注。

强影响点对模型的参数估计和统计推断会产生一定影响，通过上述两个统计量可以得到 Cook 距离和 W-K 统计量的折线图，统计量的值越大，该数据点的影响也就越大。本例中 Cook 距离记作 CD_i，W-K 统计量记作 $dffit$，不同情况下有不同的评价标准，此处将 $CD_i>0.5$，$|dffit|>1.5$ 的数据点视作强影响点[11] [12] [13] [14]，由此本例共有 3 个强影响点，如表 2 所示。

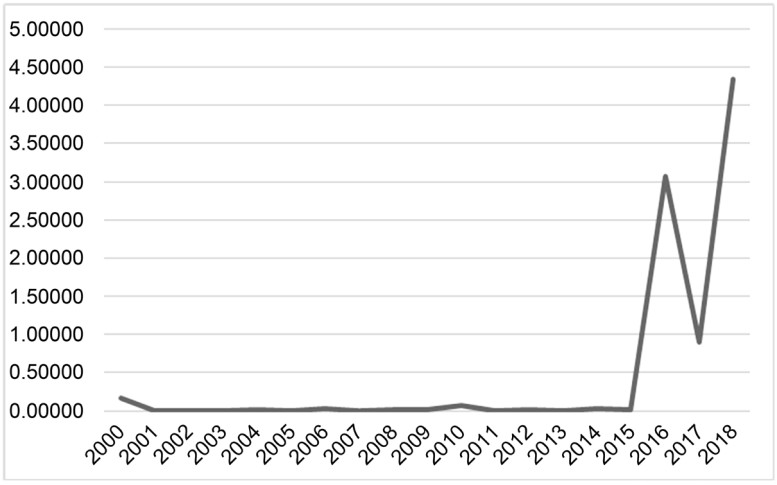

Figure 2. Scatter plot of Cook distance
图 2. Cook 距离散点图

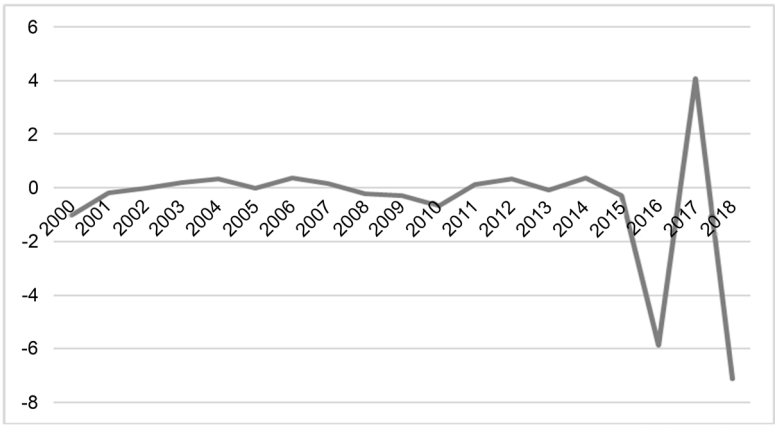

Figure 3. Scatter plot of W-K statistics
图 3. W-K 统计量散点图

Table 2. Summary of strong impact points
表 2. 强影响点汇总表

异常点	CD_i	*dffit*	GDP
2016	3.070	−5.85847	11313.72
2017	0.906	4.08337	12603.36
2018	4.340	−7.10142	13509.15

这三年数据的 Cook 距离和 W-K 统计量相对较大，这是因为杠杆值的大小与数据点距离数据中心的距离相关，离中心值越远，杠杆值就越大，对数据产生的影响就越大，因此被认为是强影响点。杭州市 GDP 的强影响点呈现出聚集成堆的现象，即这 3 个强影响点就是近三年的数据。随机性成分一般是造成 GDP 出现强影响性的主要原因，随机性成分往往是在现实经济运行过程中出现的各种不可预测、非重复性或者基本没有规律的突发性情况，而且与不同情况下的经济发展政策和市场行情相关。对于杭州来说，虽然存在一系列无法掌控的挑战，但是杭州凭借自身的发展资源和优势，区域经济持续稳向走好，质量向优发展，人民的生活水平不断提高，产业结构日益凸显优势，高端技术和人才不断增加，吸引外资的潜力逐渐提升，发展的

韧性和获得感明显增强，经济增长呈现出质量高、速度稳的新特点。另外近阶段由于国民经济已经经历过较大的增长，增长速度略微减缓，但是仍然保持良好的增长势头，与之前的经济状况相比，杭州市的经济发展到了一个独具特点的新阶段，虽然经济发展持续向好，但是仍然需要不断提升，获得更大的突破。

5. 模型修正

5.1. 数据删除模型概述

5.1.1. 数据删除模型

考虑多元线性回归模型：$Y=X\beta+\varepsilon$，$E(\varepsilon)=0$，$Var(\varepsilon)=\sigma^2 I_n$，其中 $X=(\tilde{x}_1,\cdots,\tilde{x}_n)^{\mathrm{T}}$ 为已知的列满秩阵，$\tilde{x}_i=(1,x_i^{\mathrm{T}})^{\mathrm{T}}$，$x_i=(x_{i1},\cdots,x_{ip})^{\mathrm{T}}$；$\beta=(\beta_1,\cdots,\beta_p)^{\mathrm{T}}$，$Y=(y_1,\cdots,y_n)^{\mathrm{T}}$，$\varepsilon=(\varepsilon_1,\cdots,\varepsilon_n)^{\mathrm{T}}$。

为了研究各个数据点在统计推断中的作用，就是要检测第 i 个数据点是否为异常点或强影响点。数据删除模型即多元线性回归模型中删除第 i 个数据点之后，研究该数据点删除前后对回归模型参数 β 的估计量以及其他统计量是否有举足轻重的影响。数据删除模型的矩阵形式可以表示为：

$$Y_{(i)}=X_{(i)}\beta+\varepsilon_{(i)},\varepsilon_{(i)}\sim N(0,\sigma^2 I_{n-1})$$

其中 $Y_{(i)}$ 和 $\varepsilon_{(i)}$ 表示 Y 和 ε 删除第 i 分量后的向量，而 $X_{(i)}$ 表示删掉矩阵 X 的第 i 行向量 $\tilde{x}_i$ 后得到的矩阵。

5.1.2. 数据删除模型的参数估计

数据删除模型中参数 β 和 σ^2 的最小二乘估计可表示为：

$$\hat{\beta}_{(i)}=\hat{\beta}-\frac{(X^{\mathrm{T}}X)^{-1}\tilde{x}_i e_i}{1-h_{ii}},\hat{\sigma}^2_{(i)}=\frac{n-p-1-r_i^2}{n-p-2}\hat{\sigma}^2$$

上式也表明第 i 个数据点所对应的残差 e_i 越大，则估计量 $\hat{\beta}_{(i)}$ 和 $\hat{\beta}$ 之间的差异越大，那么第 i 个数据点对模型的影响也越大。另外学生化内残差 r_i^2 越大，估计量 $\hat{\beta}_{(i)}$ 和 $\hat{\beta}$ 之间的差异越大。第 i 个数据点对应的杠杆值 h_{ii} 越大，那么估计量 $\hat{\beta}_{(i)}$ 和 $\hat{\beta}$ 之间的差异也会越大，即高杠杆点对模型的参数估计有较大的影响。

5.2. 实例研究

基于删除 3 个强影响点之后的数据，分别对因变量和自变量做一般多元线性回归和逐步回归，得到删除强影响点之后的多元回归模型。

Table 3. Modified multiple regression results
表 3. 修正后的多元回归结果

Model	Unstandardized Coefficients		Standardized	*t*	Sig.
	B	Std. Error	Beta		
(常量)	−13588.167	3345.005		−4.062	**0.002**
年末常住人口(*X*1)	11.994	4.582	0.130	2.618	**0.026**
全社会就业人数(*X*2)	1.212	1.392	0.040	0.871	0.404
固定资产投资(*X*3)	0.370	0.136	0.210	2.725	**0.021**
居民消费指数(*X*4)	10.964	2.046	0.274	5.358	**0.000**
财政总收入(*X*5)	1.513	0.454	0.354	3.331	**0.008**

a. 因变量：GDP，$R^2=1.000$，$R_a^2=0.999$，Prob(F) = 0.000，**AIC = 138.26**

Table 4. Stepwise regression results
表 4. 逐步回归结果

Model	Unstandardized Coefficients		Standardized	*t*	Sig.
	B	Std. Error	Beta		
(常量)	−15326.525	2654.159		5.775	**0.000**
年末常住人口(*X*1)	14.944	3.050	0.162	4.900	**0.000**
固定资产投资(*X*3)	0.280	0.087	0.159	3.214	**0.008**
居民消费指数(*X*4)	11.622	1.880	0.291	6.181	**0.000**
财政总收入(*X*5)	1.691	0.402	0.395	4.211	**0.001**

a. 因变量：GDP， $R^2=1.000$， $R_a^2=0.999$，Prob(F) = 0.000，**AIC = 137.43**

表 3 是删除强影响点之后得到的多元回归结果，表 4 是利用逐步回归法得到的回归结果。基于数据删除模型的回归模型拟合度和解释程度更高，经过变量筛选后，模型的 AIC 有所降低，此时的模型更接近于实际情况。由表 4 可得修正后的回归模型为：

$$Y=14.944X_1+0.280X_3+11.622X_4+1.691X_5-15326.525 \quad (2)$$

此时回归模型通过 F 检验，并且排除了全社会就业人数这一变量，其余 4 个自变量的回归系数依旧显著。并且 4 个回归系数均为正，说明这 4 个自变量与因变量之间呈现正比例关系，随着年末常住人口的增多、固定资产投资额的增加、居民消费指数的提高及财政总收入的增加，杭州市 GDP 呈现上升趋势。因此，如若想要使杭州市 GDP 持续稳定增加，就要努力吸引人才和资金，提高就业率和工资水平，进而提高居民的生活消费水平。另外还要注意协调财政总收入和总支出，保持财政收入稳定增长，尽量避免不必要的财政支出。

6. 结论与建议

根据原始回归模型与数据删除模型的比较可知，近几年杭州市 GDP 与全社会就业人数的关系更大。而在删除了近 3 年的数据后，全社会就业人数的回归系数不再显著，此时得到的回归模型只有 4 个变量显著，分别是年末常住人口、固定资产投资额、居民消费指数以及财政总收入，且均为正相关关系。

根据以上模型的求解结果可知，杭州市 GDP 的发展与这些因素密不可分，进一步得出杭州市经济发展面临的主要问题。政府应该采取一系列可行措施：继续发展第三产业，扩大开放力度，吸引更多外资，努力提高固定资产投资额；提高就业率，吸引更多行业人才，特别是适应杭州发展的高端技术人员，发挥好杭州的产业优势；改善居民的生活水平，以此提高居民的收入和居民消费指数；掌握好政府与市场之间的关系，保证地区财政收入的稳定增长，并适当降低财政支出。从各个关键方面入手，共同促进杭州市国民经济的稳定、高效、全面可持续发展。

基金项目

浙江省高校重大人文社科攻关计划项目(2018QN037)。

参考文献

[1] 郭芳, 冷洛.国内生产总值影响因素的计量分析[J]. 云南财贸学院学报(社会科学版), 2008, 23(1): 90-92.
[2] 文静. 影响国内生产总值的因素分析[J]. 商业文化(下半月), 2011(5): 109.

[3] 张卜元, 刘冰冰, 张东旭. 我国国内生产总值影响因素实证分析[J]. 合作经济与科技, 2016(2): 10-11.

[4] 程静. 国内生产总值影响因素分析[J]. 经济研究导刊, 2014(7): 7+46.

[5] 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.

[6] 赵深淼, 张英, 刘洋. 北京市国内生产总值影响因素分析[J]. 佳木斯大学学报(自然科学版), 2017, 35(4): 681-683.

[7] 黄潇逸. 基于多元线性回归分析的地区生产总值影响因素研究[C]//浙江省地理学会. 浙江省地理学会 2018 年学术年会暨“城市国际化研究”高峰论坛论文摘要集. 浙江省地理学会: 浙江省地理学会, 2018: 15.

[8] 李彦芙. 基于多元线性回归模型的江苏省 GDP 增长影响因素研究[J]. 特区经济, 2019(4): 84-88.

[9] 王璐. 带线性约束的多元线性回归模型的统计诊断[D]: [硕士学位论文]. 南京: 南京理工大学, 2008.

[10] 唐年胜, 李会琼. 应用回归分析[M]. 北京: 科学出版社, 2014: 97-102, 105-108.

[11] 黄守坤. 回归诊断中例外数据点及大影响点的处理[J]. 统计与决策, 2002(6): 47.

[12] 赵喜仓, 渠田田, 许鲜欣. 数据删除模型在 GDP 诊断中的应用[J]. 统计与决策, 2011(10): 8-10.

[13] 胡章刚. 线性回归诊断若干问题研究[D]: [硕士学位论文]. 武汉: 武汉科技大学, 2010.

[14] 王彤. 线性回归模型的稳健估计及多个异常点诊断方法研究[D]: [博士学位论文]. 西安: 第四军医大学, 2000.