Advances in Applied Mathematics 应用数学进展, 2022, 11(10), 7185-7192
Published Online October 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.1110762

Hans汉斯

基于多元回归模型的人均可支配收入预测分析

代铁林*，牟唯嫣

北京建筑大学理学院，北京

收稿日期：2022年9月18日；录用日期：2022年10月8日；发布日期：2022年10月18日

摘　要

人均可支配收入是反应居民生活水平的重要指标。为探究影响人均可支配收入的因素与人均可支配收入之间的关系，以安徽省各市的数据为例，建立多元线性回归模型，通过逐步回归方法确定重要影响因素，并得到预测值。结果显示，地区生产总值、社会消费品零售总额等对地区人均可支配收入有显著影响。利用逐步回归对变量选择后的模型预测值具有较高精度，这也证明了所提模型的有效性。

关键词

人均可支配收入，多元线性回归模型，逐步回归

Per Capita Disposable Income Forecast Analysis Based on Multiple Regression Models

Tielin Dai*, Weiyan Mu

School of Science, Beijing University of Civil Engineering and Architecture, Beijing

Received: Sep. 18th, 2022; accepted: Oct. 8th, 2022; published: Oct. 18th, 2022

Abstract

Per capita disposable income is an important indicator of residents' living standards. To explore the relationship between the factors affecting per capita disposable income and per capita disposable income, taking the data of cities in Anhui Province as an example, establish a multiple linear regression model, important influencing factors are identified by stepwise regression methods and get the predicted value. The results show that, Regional GDP and total retail sales of consumer

*通讯作者。

文章引用：代铁林，牟唯嫣．基于多元回归模型的人均可支配收入预测分析[J]．应用数学进展，2022，11(10): 7185-7192. DOI: 0.12677/aam.2022.1110762

goods have a significant impact on regional per capita disposable income. The model predicted values after variable selection are highly accurate using stepwise regression, this also proves the effectiveness of the proposed model.

Keywords

Per Capita Disposable Income, Multiple Linear Regression Models, Gradual Regression

Copyright © 2022 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. 引言

21 世纪以来，随着经济的发展，我国国际地位的提升，中华文化的国际影响力也会水涨船高。国民经济发展迅速，人们的生活水平不断提升，随之而来的人均可支配收入也逐渐增多。如在安徽省下辖的市中，人均可支配收入均逐年增加。2020 年安徽省人均可支配收入达到 20,183 元，较 2010 年增加了 27.8% [1]。影响人均可支配收入的因素众多，因此人们对人均可支配收入的研究角度也各不相同。如 Yiman Dong，Tao Zhao [2] (2017)基于计量经济学技术和面板数据集，对家庭人均收入、人均支出和人均 CO 的因果关系进行了研究，实证结果在不同地区表现出不同的因果关系，为节能减排提供了新思路。滕秀花，戴林送[3] (2020)结合灰色预测模型和 Markov 预测模型相的优点，建立灰色 Markov 预测模型，对安徽省城镇居民人均可支配收入进行预测，结果显示预测效果较好。黄志煌[4] (2020)建立包含中国内地 1992~2018 年人均 GDP、常住人口城镇化率、政府财政的支出和城镇居民人均可支配收入四个时间序列变量的 VAR 模型进行实证检验，结果显示中国城镇化水平与城镇居民人均可支配收入之间具有显著的正向相关关系。吴旭[5] (2021)以我国 31 个省、自治区、直辖市为例，对相关数据建立计量模型，结果得到人均可支配收入对居民消费支出的影响最大，人均 GDP 对居民消费支出的影响次之，我国 31 个省市区的居民消费水平差距较大。肖枝洪，马泽巍等[6] (2021)针对 1999~2018 年中国 31 个省级行政区的城镇居民人均可支配收入，提出一种函数型数据聚类方法，聚类结果表明：中国城镇居民人均可支配收入呈不断增长的趋势。左思静，杨宜平[7] (2021)通过对重庆市 1995~2018 年城乡居民人均可支配收入与消费水平数据分析，发现尽管均值回归拟合该数据效果较好。吕学静，杨雪[8] (2022)利用最低生活保障标准测算方法验证贫困家庭的基本生活需求消费与人均可支配收入没有呈现稳定关系，最低生活保障标准与人均可支配收入之间存在长期均衡关系。Fang-Li Ruan，Liang Yan [9] (2022)分析了我国电力消耗、废水排放对人均可支配收入和经济增长的影响。结果显示城市的用电效率下降，将降低平均可支配收入，增加废水排放强度。

本文通过多元线性回归模型，对人均可支配收入进行建模，相比简单的线性回归模型，多元线性回归分析可以综合得到多种因素对因变量的影响。数据来源为安徽省统计年鉴。收集了 2020 年安徽省下辖的 16 个市的人均可支配收入等数据。建立了人均可支配收入与社会消费品零售总额、全社会用电量等之间的多元线性回归模型。通过逐步回归方法得到的预测结果与真实数据吻合较好。

2. 数据分析

2.1. 数据来源及标准化

使用 Python 爬虫收集了 2020 年安徽省下辖的 16 个市的各项数据。其中包括人均可支配/元、GDP/

亿、就业人数/万人、每十万人拥有的大专以上人数、社会消费品零售总额/亿元、失业率、全社会用电量/亿千瓦时、进出口总额/亿美元、全社会供水用水量总量/万立方米，分别对应表 1 中的 Y 和 X_1 至 X_8。

首先将数据标准化。即每项数据减去其均值，然后除于标准差。得到数据如表 1 所示。这里只显示前五行标准化后的数据。

Table 1. Normalized portions of the data
表 1. 标准化后的部分数据

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	Y
3.479279	2.280564	2.937248	3.397974	1.365125	2.907045	3.592340	3.442450	1.987280
0.592245	0.849855	0.055667	0.691155	0.197708	0.819944	0.448550	0.494005	0.107075
0.278921	0.426118	1.332766	0.155509	1.233320	0.696302	0.418941	0.527125	1.015484
0.169920	0.741143	0.793807	0.063897	0.875563	0.487357	0.435676	0.518340	1.041717
0.152707	0.236468	0.273507	0.057175	0.367172	0.561758	0.349426	0.412101	0.066867

得到标准化数据后，下面将对数据进行探索性分析。

2.2. 探索性分析及正态性检验

首先通过数据之间的相关性系数进行分析。使用以下对数据的处理均使用 Python 编程实现。计算各项数据间的相关系数，得到相关系数矩阵，通过热力图进行可视化。结果显示在图 1 中。

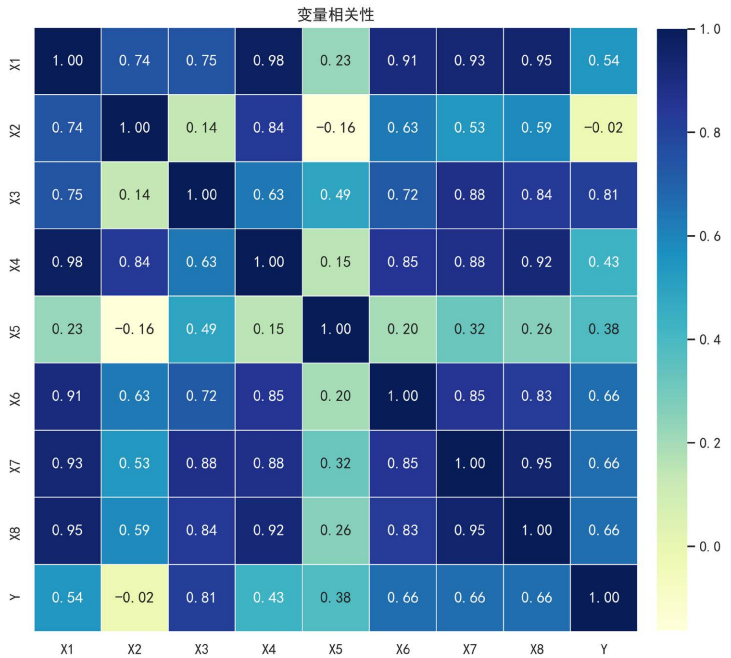

Figure 1. Data correlation coefficient heat plot
图 1. 数据相关性系数热力图

通过图 1 可知，所选变量间的相关性较强。X_2 与因变量 Y 之间的相关性较小，即所选的因素对人均可支配收入均具有一定影响。这也为以下建立模型提供了支撑。

针对数据的分布进行了正态性检验。以标准正态分布的分位数为横坐标，样本值为纵坐标制成散点图，称为 Q-Q 图。数据直方图和 Q-Q 图显示在图 2 中。

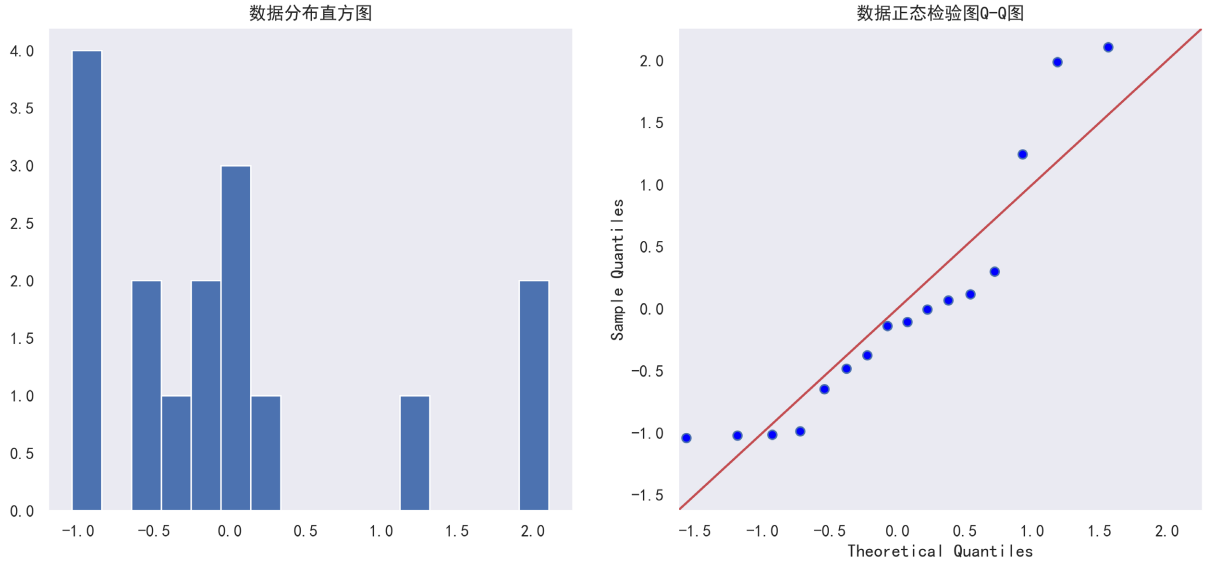

Figure 2. Data distribution histogram and Q-Q plot
图 2. 数据分布直方图和 Q-Q 图

通过图 2 可知，因变量 Y 的数据不具有正态分布特征，并且具有异常值。针对此情况，对数据进行正态性转换，结果在图 3 中显示。

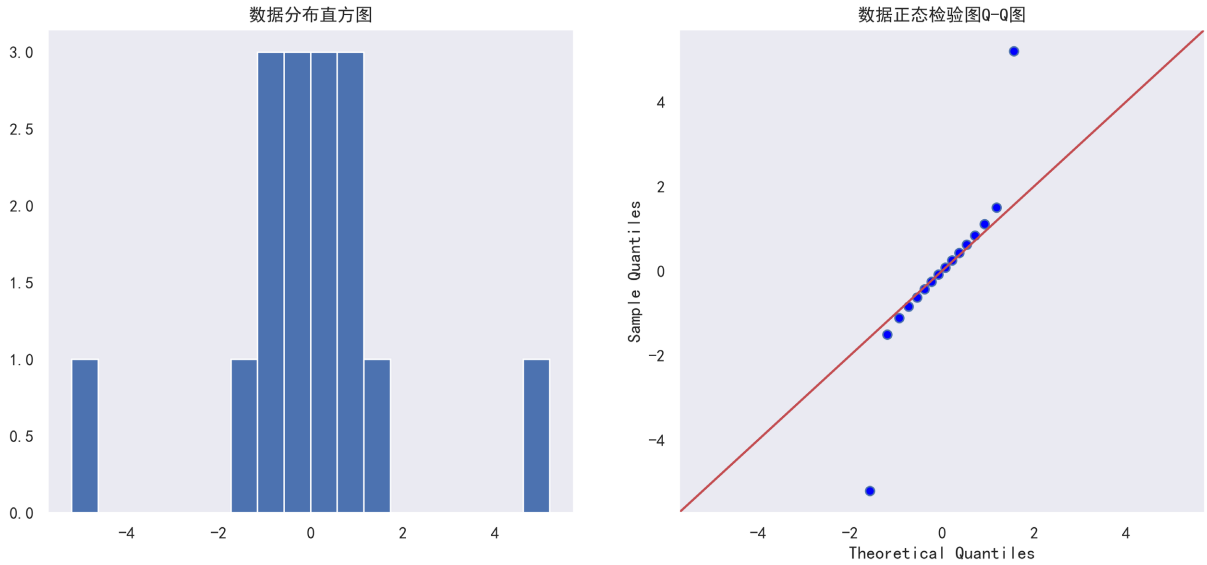

Figure 3. Distribution histogram and Q-Q plot after data transformation
图 3. 数据转化后的分布直方图和 Q-Q 图

通过图 3 可知，转换后的数据为正态分布，且除了两个异常点外，数据为标准正态分布。
在对数据进行探索探索性分析后。下面将使用转化后的数据建立多元线性回归模型。

3. 多元线性回归模型

3.1. 模型简介

一般多元线性回归模型可表示为：

$$Y=\beta_0+\beta_1 x_1+\cdots+\beta_k x_k+\varepsilon \tag{1}$$

其中，因变量 Y 为随机变量，$(\beta_0\beta_1\cdots\beta_k)$ 为回归系数，$x_1\cdots x_k$ 为自变量，ε 为随机误差项。

记 $\beta=(\beta_0\beta_1\cdots\beta_k)'$，$Y=(y_1\cdots y_k)'$，$\varepsilon=(\varepsilon_1\cdots\varepsilon_n)'$，$X=(X_1\cdots X_k)'$。则(1)式可用矩阵表示为：

$Y=X\beta+\varepsilon$

假设误差服从正态分布$(0,\sigma^2)$，使用最小二乘估计得到的参数估计值为：

$$\hat{\beta}=(X'X)^{-1}X'Y$$

$$\hat{Y}=X\hat{\beta}$$

设 e 为残差，$e=Y-\hat{Y}$，则 σ^2 的最小二乘估计为：

$$\hat{\sigma}^2=\frac{e'e}{n-k-1}$$

3.2. 模型应用

假设人均可支配收入 Y 与 $X_1\cdots X_8$ 之间服从多元线性回归模型，即 Y 与 $X_1\cdots X_8$ 之间满足下式：

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4+\beta_5X_5+\beta_6X_6+\beta_7X_7+\beta_8X_8+\varepsilon \tag{2}$$

使用最小二乘估计进行多元线性回归。结果显示在表 2 中。

Table 2. Multiple linear regression results
表 2. 多元线性回归结果

	coef	std err	t	P > |t|
Intercept	6.1e−12	0.268	2.29e−11	1.000
X_1	−6.1121	3.296	−1.855	0.106
X_2	−4.9305	1.728	−2.853	0.025
X_3	−1.9657	1.472	−1.335	0.224
X_4	6.9156	4.148	−1.667	0.139
X_5	−0.5168	0.370	−1.398	0.205
X_6	3.5219	0.823	4.278	0.004
X_7	0.8711	1.245	0.700	0.507
X_8	0.9998	1.773	0.564	0.590

通过表 2 可知，回归结果较显著，P 值较大的变量有 3 个。计算决定系数 $R^2=0.874$

矫正决定系数 $Adj\_R^2=0.730$。$\text{Prob}=0.0139$。

对模型进行优化，得到预测结果，在图 4 中显示。

通过图 4 可知，预测结果的均方误差 MSE 较大，可能是受到异常值的影响。我们将剔除异常值后再次进行预测的结果显示在图 5 中。

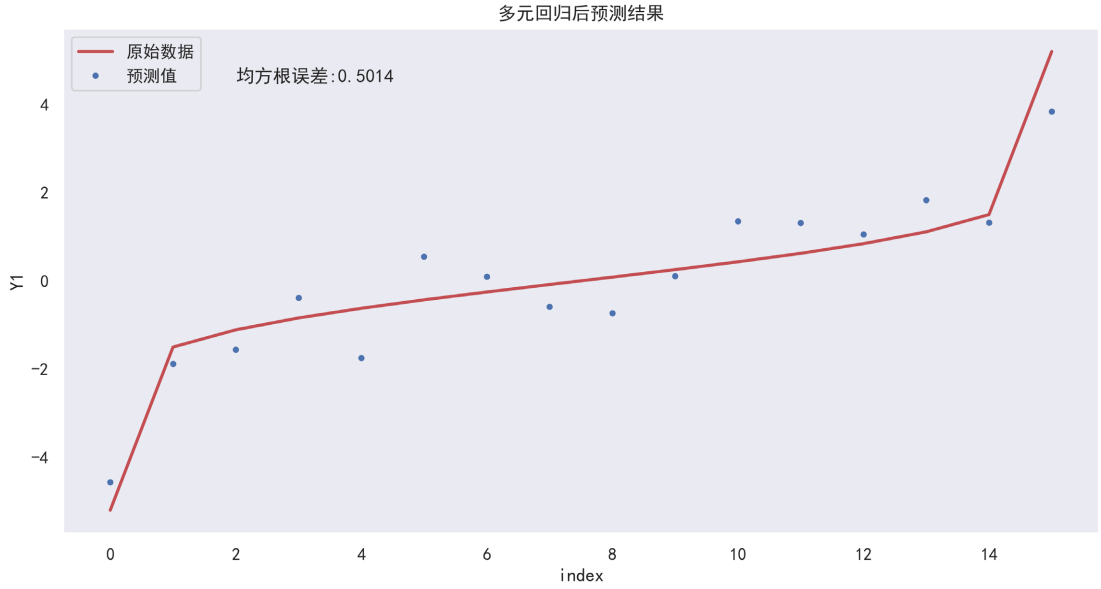

Figure 4. Optimized multiple regression prediction results
图 4. 优化后的多元回归预测结果

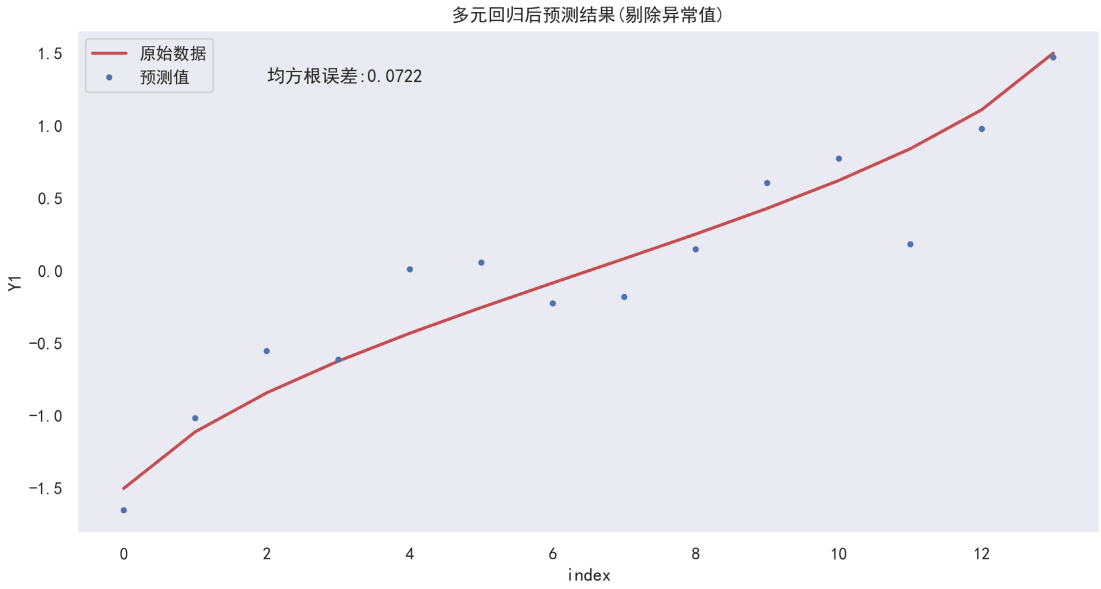

Figure 5. Multiple regression prediction results after outliers are excluded
图 5. 剔除异常值后多元回归预测结果

通过图 5 可知，在剔除异常值后，预测的结果更靠近于真实数据曲线。并且预测结果的 MSE 减小较多，说明异常值的影响得到较好消除。

3.3. 逐步回归

通过逐步回归对变量进行筛选，结果显示在表 3 中，这里只展示前 10 行是结果。

通过表 3 可知，在 4 个变量的情况下，已经得到较高的 R_squared。于是只使用 4 个变量进行多元线性回归，选取的 4 个变量为(X_1, X_2, X_4, X_6)，此时的 $R^2 = 0.819790$ 。回归结果显示在表 4 中。

Table 3. The result of gradual regression
表 3. 逐步回归的结果

variable	bic	aic	Cond	R_squared
(X_2, X_4)	22.588209	20.671037	3.609060	0.758523
(X_2, X_4, X_6)	22.620659	20.064429	6.270598	0.799545
(X_1, X_2)	22.905865	20.988693	2.816752	0.752981
$(X_1, X_2, X_4, X_5, X_6)$	23.446344	19.612000	35.910581	0.854153
(X_1, X_2, X_4, X_6)	23.769172	20.573886	33.563317	0.819790
$(X_2, X_4, X5, X_6)$	23.830747	20.635461	7.184406	0.818996
(X_2, X_3)	24.140960	22.223788	1.243170	0.730199
(X_2, X_4, X_5)	24.292953	21.736724	5.267011	0.774112
(X_1, X_2, X_6)	24.860683	22.304453	6.868695	0.764764
(X_2, X_4, X_7)	24.869210	22.312981	10.925808	0.764620

Table 4. Select the regression results for 4 variables
表 4. 选取 4 个变量的回归结果

	Coef	std err	t	P > |t|
Intercept	0.194	0.125	−0.155	0.881
X_1	−1.62	1.569	−1.034	0.031
X_2	−1.50	0.467	−3.232	0.012
X_4	2.2646	1.474	1.537	0.043
X_6	0.7928	0.496	1.600	0.048

通过表 4 可知，筛选的 4 个变量进行回归后，P 值均小于 0.05，说明拟合结果较好。使用逐步回归筛选后的变量进行预测，结果显示在图 6 中。

通过图 6 可知，在仅选择 4 个变量进行回归后，得到的 MSE 仍较小。筛选后的变量对因变量 Y 有线性回归关系。

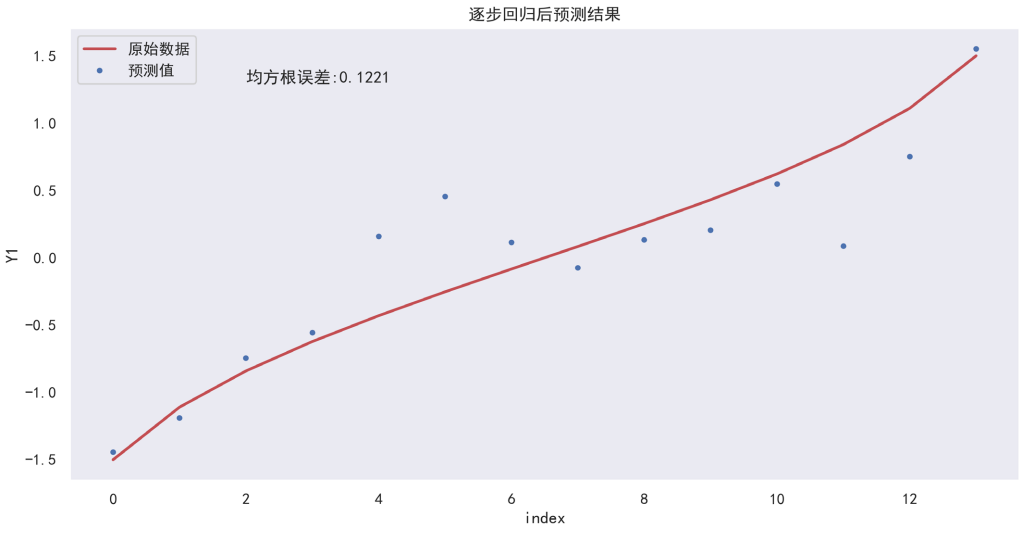

Figure 6. Stepwise regression of the predicted result after filtering variables
图 6. 逐步回归筛选变量后的预测结果

4. 结论

本文通过多元线性回归模型，对人均可支配收入进行建模。对收集的数据首先进行标准化，然后做相关性分析，探索变量间的相关性。通过正态性检验和正态性变换将数据转化为标准正态分布。建立了因变量的多元线性回归。通过逐步回归筛选出重要影响变量。并得到预测值。结果显示，地区生产总值GDP、就业人数、社会消费品零售总额和全社会用电总量与人均可支配收入之间呈显著线性回归关系。通过筛选后的变量进行预测得到的结果具有较好的精度。这也验证了模型的有效性。

基金项目

国家自然科学基金/National Natural Science Foundation of China (NO.12171462)。

参考文献

[1] 安徽省统计局(2022) [EB/OL]. http://tjj.ah.gov.cn/, 2022-10-14.

[2] Dong, Y.M. and Zhao, T. (2017) Difference Analysis of the Relationship between Household per Capita Income, per Capita Expenditure and per Capita CO_2 Emissions in China: 1997-2014. *Atmospheric Pollution Research*, **8**, 310-319. https://doi.org/10.1016/j.apr.2016.09.006

[3] 滕秀花, 戴林送. 应用灰色 Markov 模型预测安徽省城镇居民人均可支配收入[J]. 安庆师范大学学报(自然科学版), 2020, 26(4): 24-26. https://doi.org/10.13757/j.cnki.cn34-1328/n.2020.04.006

[4] 黄志煌. 城镇化对城镇居民人均可支配收入的影响——基于 VAR 模型分析[J]. 广西质量监督导报, 2020(7): 49-51+48.

[5] 吴旭. 我国居民消费水平的影响因素和现状分析[J]. 统计与管理, 2021, 36(10): 4-10. https://doi.org/10.16722/j.issn.1674-537x.2021.10.001

[6] 肖枝洪, 马泽巍, 曾伟. 基于函数型数据的城镇居民人均可支配收入聚类分析[J]. 重庆三峡学院学报, 2021, 37(2): 44-56. https://doi.org/10.13743/j.cnki.issn.1009-8135.2021.02.006

[7] 左思静, 杨宜平. 重庆市城乡居民人均可支配收入和消费的分位数回归估计[J]. 重庆工商大学学报(自然科学版), 2021, 38(1): 120-128. https://doi.org/10.16055/j.issn.1672-058X.2021.0001.018

[8] 吕学静, 杨雪. 城市最低生活保障标准动态调整机制研究——基于消费视角的省级面板数据[J]. 人口与发展, 2022, 28(2): 104-112+47.

[9] Ruan, F.-L. and Yan, L. (2022) Interactions among Electricity Consumption, Disposable Income, Wastewater Discharge, and Economic Growth: Evidence from Megacities in China from 1995 to 2018. *Energy*, **260**, Article ID: 124910. https://doi.org/10.1016/j.energy.2022.124910