Statistics and Application 统计学与应用, 2020, 9(3), 412-421
Published Online June 2020 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2020.93044

Hans汉斯

Analysis of Per Capita Disposable Income of Rural Residents

Ran Zhao

Qufu Normal University, Qufu Shandong
Email: rt371425@163.com

Received: May 15th, 2020; accepted: May 29th, 2020; published: Jun. 8th, 2020

Abstract

Through the analysis, the factors affecting the per capita disposable income of rural residents are found, and a linear model is established by using the stepwise regression method. Through the significance test of the regression coefficient, the significant factors affecting the per capita disposable income of rural residents can be found. For multicollinearity among independent variables, principal component regression or ridge regression is used to eliminate the collinearity among independent variables, and the regression equation of ridge regression is established $y = 49.504 + 0.432x_1 + 0.142x_2 - 0.549x_3$. Draw a sequence diagram of the per capita disposable income and time of rural residents, and find that the per capita disposable income of rural residents is on the rise. Build a model $\mathrm{ARMA}(p,q)$ based on the data after the difference and extract the trend.

Keywords

Linear Regression, Principal Component Regression, Ridge Regression, $\mathrm{ARMA}(p,q)$ Model

农村居民人均可支配收入的分析

赵　冉

曲阜师范大学，山东　曲阜
Email: rt371425@163.com

收稿日期：2020年5月15日；录用日期：2020年5月29日；发布日期：2020年6月8日

摘　要

通过分析，找到影响农村居民人均可支配收入的因素，利用逐步回归法建立线性模型，通过回归系数的显著性检验，可以找到影响农村居民人均可支配收入的显著因素。对于自变量之间的多重共线性，利用主成分回归或者岭回归消除自变量之间的共线性，建立岭回归的回归方程 $y = 49.504 + 0.432x_1 + 0.142x_2 - 0.549x_3$。

文章引用: 赵冉. 农村居民人均可支配收入的分析[J]. 统计学与应用, 2020, 9(3): 412-421.
DOI: 10.12677/sa.2020.93044

绘制农村居民人均可支配收入与时间的序列图，发现农村居民人均可支配收入呈上升趋势，根据差分后的数据建立 ARMA(p,q) 模型，将趋势提取出来。

关键词

线性回归，主成分回归，岭回归， ARMA(p,q) 模型

1. 引言

通过分析找到影响农村居民人均可支配收入的显著性因素并建立线性回归模型，检查自变量之间的多重共线性并消除，在使用主成分分析消除多重共线性建立模型时，农业各税回归系数的符号不符合实际，尝试使用岭回归建立模型后，回归系数的正负号得到解决，多重共线性也已消除；农村居民人均可支配收入随时间的变化呈上升趋势，可建立模型，将趋势项提取出来。

2. 材料与方法

2.1. 线性回归模型的介绍[1]

根据搜集的数据，建立线性回归模型，进行回归方程以及回归系数的检验，分析影响农村居民人均可支配收入的因素有哪些；若模型存在共线性，则使用主成分分析方法消除共线性进一步建立模型。根据时间发生的顺序将农村居民人均可支配收入在多个时刻的数值记录下来，以得到一时间序列，建立时间序列的模型，分析农民收入的变化趋势。

2.1.1. 线性回归模型的确立

设随机变量 y 与一般变量 $x_1,x_2,\cdots,x_p$ 的线性回归模型为

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p+\varepsilon$$

式中， $\beta_0,\beta_1,\cdots,\beta_p$ 是 $p+1$ 个未知参数， β_0 称为回归常数， $\beta_1,\cdots,\beta_p$ 称为回归系数。y 称为解释变量(因变量)， $x_1,x_2,\cdots,x_p$ 是 p 个可以精确测量并控制的一般变量，称为解释变量(自变量)。

ε 是随机误差，并且假定

$$\begin{cases}E(\varepsilon)=0\\ \mathrm{var}(\varepsilon)=\sigma^2\end{cases}$$

2.1.2. 回归参数的普通最小二乘估计

即寻找参数 $\beta_0,\beta_1,\cdots,\beta_p$ 的估计值 $\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_p$，使离差平方和

$Q(\beta_0,\beta_1,\cdots,\beta_p)=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_{i1}-\cdots-\beta_px_{ip})^2$ 达到极小。

当 $(X'X)^{-1}$ 存在时，即得回归参数的最小二乘估计为：

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

2.1.3. 回归方程、回归系数的检验

对多元线性回归方程的显著性检验就是要看自变量 $x_1,x_2,\cdots,x_p$ 从整体上对随机变量 y 是否有明显的影响。

原假设 $H_0:\beta_1=\beta_2=\cdots=\beta_p=0$

构造 F 检验统计量如下：

$$F=\frac{SSR/p}{SSE/(n-p-1)}$$

当原假设成立时，F 服从自由度为 $(p,n-p-1)$ 的 F 分布。

当 $F>F_\alpha(p,n-p-1)$ 时，拒绝原假设 H_0，否则认为在显著性水平 α 下，y 与 $x_1,x_2,\cdots,x_p$ 有显著的线性关系，即回归方程是显著的。

检验 x_j 是否显著等价于检验

$$H_{0j}:\beta_j=0,\quad j=1,2,\cdots,p$$

如果接受原假设，则 x_j 不显著；如果拒绝原假设，则 x_j 是显著的。

据此可以构造 t 统计量

$$t_j=\frac{\hat{\beta}}{\sqrt{c_{jj}}\hat{\sigma}}$$

式中

$$\hat{\sigma}=\sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n}e_i^2}$$

2.1.4. 共线性诊断

① 方差扩大因子法

$c_{jj}=\frac{1}{1-R_j^2}$ 作为方差扩大因子的定义，证明见参考文献[2]，当 $VIF_j\geq 10$ 时，说明自变量 x_j 与其余自变量之间有严重的多重共线性。(注意：有些教材认为 $VIF_j\geq 4$ 存在多重共线性。详见参考文献[3])。

② 条件数

记 $X'X$ 的最大特征根为 λ_m，称

$$k_i=\sqrt{\frac{\lambda_m}{\lambda_i}},\quad i=0,1,\cdots,p$$

为特征根 λ_i 的条件数。

通常认为 $0<k<10$ 时，设计矩阵 X 没有多重共线性；$10\leq k<100$ 时，存在较强的多重共线性；$k\geq 100$ 时，存在严重的多重共线性。

2.1.5. 主成分的定义与导出[4]

设 X 是 p 维随机变量，并假设 $\mu=E(X)$，$\Sigma=\mathrm{var}(X)$。考虑如下线性变换

$$\begin{cases}Z_1=a_1^{\mathrm{T}}X\\Z_2=a_2^{\mathrm{T}}X\\\quad\vdots\\Z_p=a_p^{\mathrm{T}}X\end{cases}$$

易见

$$\mathrm{var}\left(Z_i\right)=a_i^{\mathrm{T}}\sum a_i,\ \ i=1,2,\cdots,p$$

$$\mathrm{cov}\left(Z_i,Z_j\right)=a_i^{\mathrm{T}}\sum a_j,\ \ i,j=1,2,\cdots,p,i\neq j$$

我们希望 Z_1 的方差达到最大，即 a_1 是约束优化问题

$$\max\ \ a^{\mathrm{T}}\sum a$$

$$\text{s.t.}\ \ a^{\mathrm{T}}a=1$$

的解。因此， a_1 是 $\sum$ 最大特征值(不妨设为 λ_1)的特征向量。此时，称 $Z_1=a_1^{\mathrm{T}}X$ 为第一主成分。类似地，希望 Z_2 的方差达到最大，并且要求 $\mathrm{cov}\left(Z_1,Z_2\right)=a_1^{\mathrm{T}}\sum a_2=0$。由于 a_1 是 λ_1 的特征向量，所以，选择的 a_2 应与 a_1 正交。类似于前面的推导， a_2 是 $\sum$ 第二大特征值(不妨设为 λ_2)的特征向量。称 $Z_2=a_2^{\mathrm{T}}X$ 为第二主成分。

一般情况下对于协方差阵 $\sum$，存在正交阵 Q，将它化为对角阵，即

$$Q^{\mathrm{T}}\sum Q=\Lambda=diag\left(\lambda_1,\lambda_2,\cdots,\lambda_p\right)$$

且 $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p$，则矩阵 Q 的第 i 列就对应于 a_i，相应的 Z_i 为第 i 主成分。

2.2. ARMA 模型

设 $\left\{\varepsilon_t:t=0\pm1,\pm2,\cdots\right\}\sim WN\left(0,\sigma_\varepsilon^2\right)$，则序列 $\left\{X_t:t=0\pm1,\pm2,\cdots\right\}$ 满足的 p 阶常系数线性差分方程

$$X_t=\phi_0+\phi_1X_{t-1}+\cdots+\phi_pX_{t-p}+\varepsilon_t-\theta_1\varepsilon_{t-1}-\cdots-\theta_q\varepsilon_{t-q},\ \ t=0\pm1,\pm2,\cdots$$

为 p 阶自回归 q 阶移动平均模型，记为 $\mathrm{ARMA}\left(p,q\right)$ 模型。其中 $\phi_p,\theta_q\neq0$，关于 λ 的代数方程 $\lambda^p-\phi_1\lambda^{p-1}-\cdots-\phi_{p-1}\lambda-\phi_p=0$ 与 $\lambda^q-\theta_1\lambda^{q-1}-\cdots-\theta_{q-1}\lambda-\theta_p=0$ 无公共根。称 $\phi_0=0$ 的模型为中心化 $\mathrm{ARMA}\left(p,q\right)$ 模型。

利用延迟算子 B 可将模型表示为

$$\Phi\left(B\right)X_t=\phi_0+\Theta\left(B\right)\varepsilon_t,\ \ t=0,\pm1,\pm2,\cdots$$

其中 $\Phi\left(B\right)=1-\phi_1B-\phi_2B^2-\cdots-\phi_pB^p$ 和 $\Theta\left(B\right)=1-\theta_1B-\theta_2B^2-\cdots-\theta_qB^q$ 分别为 B 的 p 和 q 次多项式。

$\mathrm{AR}\left(p\right)$ 序列的偏自相关系数是 p 阶截尾的； $\mathrm{MA}\left(q\right)$ 序列的自相关系数是 q 阶截尾的。

3. 结果与分析

3.1. 回归方程的建立

根据统计年鉴搜集能够影响农村居民人均可支配收入的因素的数据。其中农村居民人均可支配收入作为因变量，农产品生产价格指数、受灾面积、耕地面积、支农支出、农业各税、农业机械总动力、农用化肥施用量、乡村就业人数、农业产值作为自变量。

3.1.1. 建立回归方程

使用 SPSS 软件建立回归方程。由于自变量个数较多，我们采用逐步回归法建立线性回归方程，进行回归方程、回归系数的检验以及共线性诊断等。回归方程的检验见图 1~3。

若记 y 为农村居民人均可支配收入， x_1 为支农支出， x_2 为农业产值， x_3 为农业各税，由图 3 可以建立回归方程

$$y=20.213+0.450x_1+0.146x_2-0.627x_3$$

模型摘要[d]

模型	R	R方	调整后R方	标准估算的误差	更改统计					德宾-沃森
					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	0.992[a]	0.983	0.983	554.0311	0.983	1578.680	1	27	0.000	
2	0.996[b]	0.993	0.992	375.9516	0.009	32.637	1	26	0.000	
3	0.997[c]	0.994	0.993	347.4305	0.001	5.444	1	25	0.028	0.809

a.预测变量：(常量),支农支出(亿元)

b.预测变量：(常量),支农支出(亿元),农业产值(亿元)

c.预测变量：(常量),支农支出(亿元),农业产值(亿元),农业各税(亿元)

d.因变量：农村居民人均可支配收入(元/人)

Figure 1. Summary of each model of stepwise regression method
图 1. 逐步回归法的各模型摘要

ANOVA[a]

模型		平和方	自由度	均方	F	显著性
1	回归	484576600.2	1	484576600.2	1578.680	0.000[b]
	残差	8287662.098	27	306950.448		
	总计	492864262.3	28			
2	回归	489189432.6	2	244594716.3	1730.546	0.000[c]
	残差	3674829.630	26	141339.601		
	总计	492864262.3	28			
3	回归	489846563.5	3	163282187.8	1352.704	0.000[d]
	残差	3017698.810	25	120707.952		
	总计	492864262.3	28			

a.因变量：农村居民人均可支配收入(元/人)

b.预测变量：(常量),支农支出(亿元)

c.预测变量：(常量),支农支出(亿元),农业产值(亿元)

d.预测变量：(常量),支农支出(亿元),农业产值(亿元),农业各税(亿元)

Figure 2. Analysis of variance of each model by stepwise regression method
图 2. 逐步回归法的各模型方差分析

系数[a]

模型		未标准化系数		标准化系数	t	显著性	共线性统计	
		B	标准误差	Beta			容差	VIF
1	(常量)	1553.056	133.420		11.640	0.000		
	支农支出(亿/元)	0.603	0.015	0.992	39.733	0.000	1.000	1.000
2	(常量)	267.923	242.490		1.105	0.279		
	支农支出(亿/元)	0.313	0.052	0.515	6.051	0.000	0.040	25.264
	农业产值(亿/元)	0.112	0.020	0.486	5.713	0.000	0.040	25.264
3	(常量)	20.213	247.971		0.082	0.936		
	支农支出(亿/元)	0.450	0.076	0.739	5.952	0.000	0.016	63.009
	农业产值(亿/元)	0.146	0.023	0.631	6.296	0.000	0.024	41.053
	农业各税(亿/元)	-0.627	0.269	-0.369	-2.333	0.028	0.010	102.380

a.因变量：农村居民人均可支配收入(元/人)

Figure 3. Coefficient value, test and collinearity diagnosis of regression equation
图 3. 回归方程的系数取值、检验及共线性诊断

由图 1 可知：回归方程的 $R^2=0.994$，回归方程的 p 值小于 $\alpha(0.05)$，回归方程是显著的；回归系数也都小于 α，因此回归系数也是显著的，显然使用线性回归方程去拟合模型是较好的。

但上图也可以看到，自变量 x_1,x_2,x_3 之间存在共线性，若能够消除共线性，则模型将会进一步改善。

3.1.2. 主成分分析消除共线性

由图 4 可知，第一个主成分的累计贡献率以达到 99%，另外两个主成分可以舍去，达到降维的目的。

总方差解释

成分	初始特征值			提取载荷平方和		
	总计	方差百分比	累积%	总计	方差百分比	累积%
1	2.973	99.107	99.107	2.973	99.107	99.107
2	0.020	0.681	99.788			
3	0.006	0.212	100.000			

提取方法：主成分分析法。

Figure 4. Cumulative contribution rate of variance
图 4. 方差的累计贡献率

为获得因变量 y 与自变量 x_1, x_2, x_3 的回归方程，可以使用 *spss* 分两次进行线性回归得到；也可使用 *R* 软件编程一步得到。

下面为简化过程，我们利用 *R* 软件进行求解。

Table 1. The regression coefficients solved by using *R*
表 1. 利用 *R* 求解的回归系数

Intercept	x_1	x_2	x_3
558.468	0.203	0.077	0.566

因此，由表 1 可知，还原后的主成分回归方程为

$$y = 558.468 + 0.203x_1 + 0.077x_2 + 0.566x_3$$

通过分析可以发现，农业各税与农村居民人均可支配收入是负相关的，因此，回归系数应该为负，但由回归方程可知，各自变量对因变量的影响都是正的，这显然与实际不符。

3.1.3. 岭回归

尝试使用岭回归进行回归方程系数的估计。绘制的岭迹图见图 5。

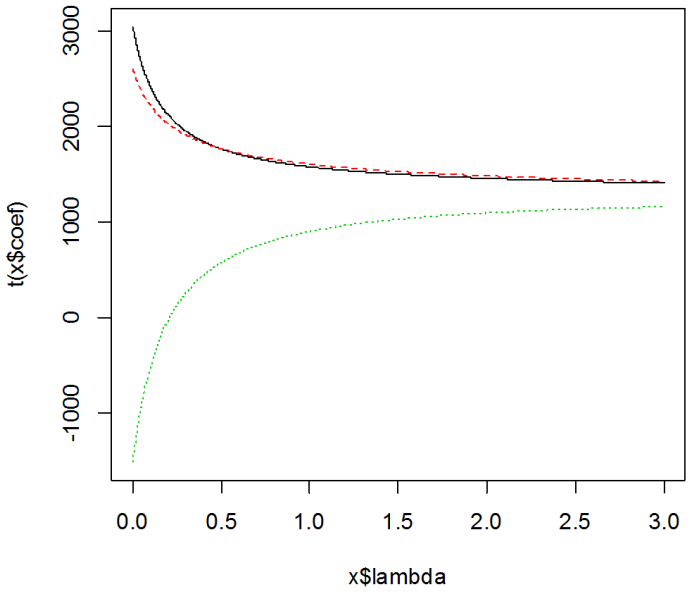

Figure 5. Ridge trace
图 5. 岭迹图

```
> #利用 select 函数找出最优岭参数lambda，会有三个值，任选一个即可。
> lm.ridge(y~x1+x2+x3,data=datal,lambda=0.013)
                          x1          x2          x3
49.5039083   0.4323035   0.1420429  -0.5493528
> #把选取的lmdba 参数写到岭回归函数中去，在这里lambda=0.013。
```

Figure 6. Estimation of ridge regression coefficient
图 6. 岭回归系数估计

由图 6 可知岭回归的回归系数，建立的回归方程为

$$y = 49.504 + 0.432x_1 + 0.142x_2 - 0.549x_3$$

显然，此问题使用岭回归能较好的建立回归方程。

从以上结果可知，影响农村居民人均可支配收入的因素有三个，分别是支农支出、农业产值、农业各税。其中，支农支出、农业产值对农村居民人均可支配收入的影响是正相关的，农业各税对农村居民人均可支配收入的影响是负相关的。

3.2. 农民收入的变化趋势

3.2.1. 描述性时序分析

从图 7 上来看，农村居民人均可支配收入呈上升趋势，可根据自相关以及偏自相关图确定模型。

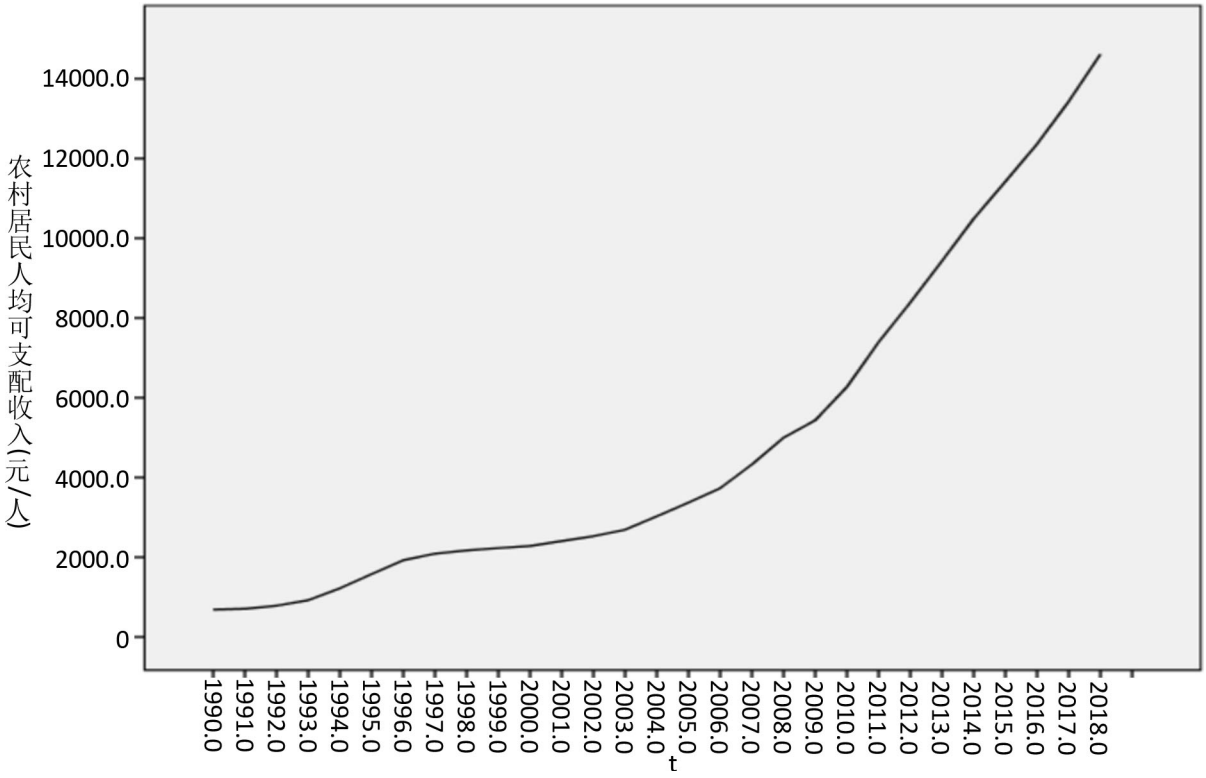

Figure 7. Time series of per capita disposable income of rural residents
图 7. 农村居民人均可支配收入的时序图

3.2.2. ARMA 模型

二阶差分后的时序图(图 8)平稳，根据差分后的自相关图以及偏自相关图，可以建立差分后 AR(1) 模型(见图 9)。

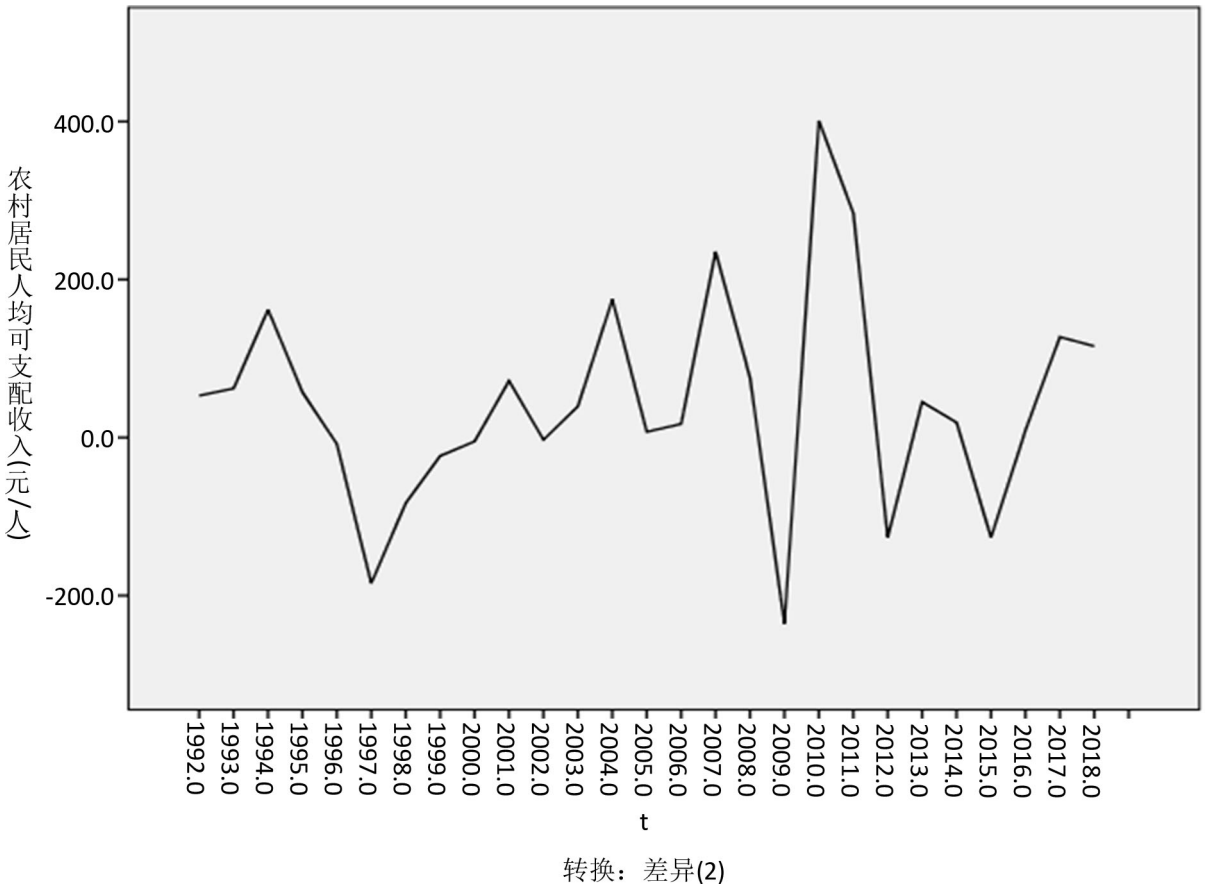

Figure 8. Sequence diagram after second-order difference
图 8. 二阶差分后的时序图

ARIMA模型参数

					估算	标准误差	t	显著性
农村居民人均可支配收入(元/人)-模型_1	农村居民人均可支配收入(元/人)	平方根	常量		23.204	64.556	0.359	0.722
			AR	延迟 1	0.130	0.203	0.641	0.528
			差异		2			
	t	不转换	分子	延迟0	-0.011	0.032	-0.357	0.724

Figure 9. Model parameters in $\mathrm{ARIMA}(1,2,0)$
图 9. $\mathrm{ARIMA}(1,2,0)$ 的模型参数

因此，可建立方程

$$\nabla^2 X_t = 23.204 + 0.130\nabla^2 X_{t-1} + \varepsilon_t$$

根据图 10，可以看到模型的拟合效果。进而也可以进行数据的预测。

4. 讨论

4.1. 回归方程的参数解释

支农支出是指国家对农业、农村、农民的财政支持，显然，财政支出越大，农民的收入也越高；农业产值能够反映农民一年的生产规模，农业产值越大，农民收入越高；

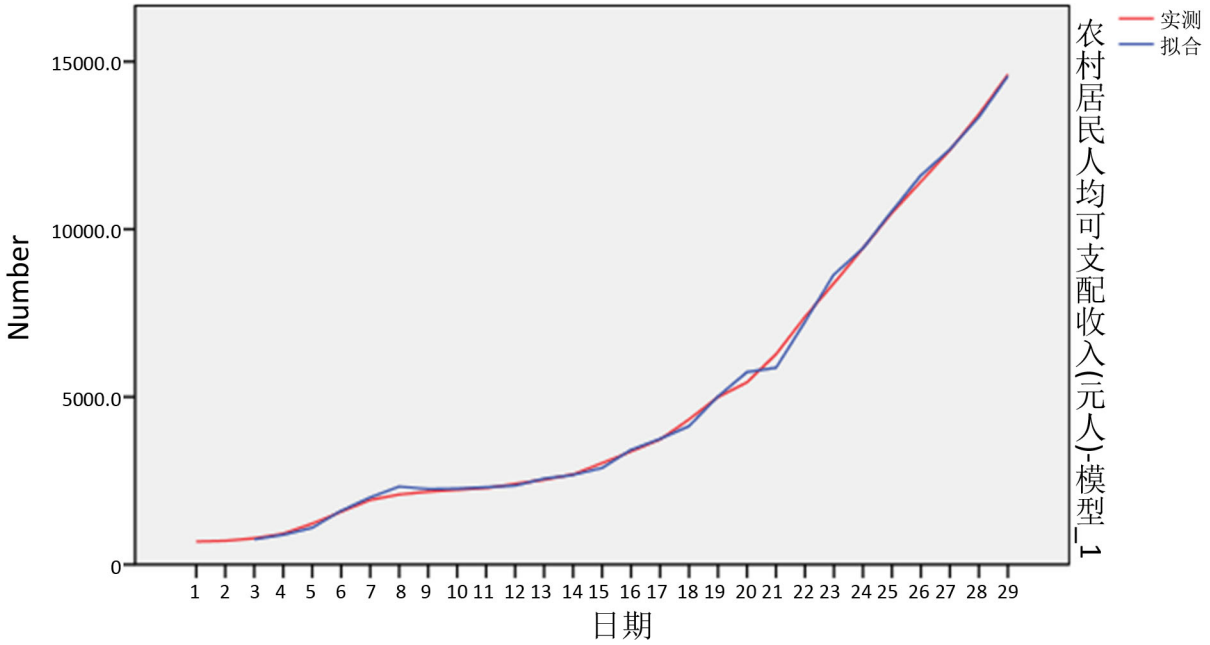

Figure 10. Fitting effect of the model
图 10. 模型的拟合效果

农业各税包括耕地占用税、契税等，是国家对从事农业生产、有农业收入的单位和个人征收的一种税，农业各税越高，农民收入越低。

以上分析可以发现，国家加大对农业的支出、提高农业产值以及降低农业税可以有效地提高农民的收入。

4.2. 预测

通过建立线性回归模型、ARIMA$(1,2,0)$，可对农村居民人均可支配收入进行预测，求解置信区间等等。

4.3. 模型分析

不同时期影响农村居民人均可支配收入的因素是不太相同的，不同国家、不同地区也是不同的，要想对一个地区影响农民收入的因素进行分析，需要重新获取数据，重新分析，但分析的基本想法是不变的。

参考文献

[1] 何晓群, 刘文卿. 应用回归分析[M]. 第 5 版. 北京: 中国人民大学出版社, 2019.

[2] 国家统计局农村社会经济调查司, 编. 中国农村统计年鉴-2019 [Z]. 北京: 中国统计出版社, 2019.

[3] 国家统计局, 编. 中国统计年鉴-2019 [Z]. 北京: 中国统计出版社, 2019.

[4] 薛毅, 陈立萍. 统计建模与 R 软件[M]. 北京: 清华大学出版社, 2007.

附　录：R 程序代码

```
install.packages("car")
install.packages("MASS")
install.packages("lars")
library(foreign)
library(car)
library(MASS)
library(lars)
data <- read.csv("数据.CSV")
data<-na.omit(data)
data1<-data[c(6,7,11,12)]
lm.sol <- lm(y~x1+x2+x3,data=data1)
summary(lm.sol)
student.pr <- princomp(~x1+x2+x3,data=data1,cor=T)
summary(student.pr,loadings=TRUE)
pre <-predict(student.pr)
data1$z<-pre[,1]
lm.sol <- lm(y~z,data=data1)
summary(lm.sol)
data1$z <-pre[,1]
student.pr <- princomp(~x1+x2+x3,data=data1,cor=T)
summary(student.pr,loadings=TRUE)
beta <-coef(lm.sol)
A<-loadings(student.pr)
x.bar <-student.pr$center
x.sd <- student.pr$scale
coef <- (beta[2]*A[,1])/x.sd
beta0 <- beta[1]-sum(x.bar*coef)
c(beta0,coef)
#绘制岭迹图
plot(lm.ridge(y~x1+x2+x3,
data=data1,lambda=seq(0,3,0.001)))
select(lm.ridge(y~x1+x2+x3,
data=data1,lambda=seq(0,0.3,0.001)))
#利用 select  函数找出最优岭参数 lambda,会有三个值，任选一个即可。
lm.ridge(y~x1+x2+x3,data=data1,lambda=0.013)
#把选取的 lmbda 参数写到岭回归函数中去，在这里 lambda=0.013。
```