Regression Analysis on Chinese Agricultural Output and Its Influencing Factors Based on R

Meichen Dong [1] Yingying Zhang [1]

[1] Department of Statistics and Actuarial Science, College of Mathematics and Statistics, Chongqing University, China

Abstract

Seven factors which influence China's agricultural output are selected to analyze the relationship between agricultural output and the factors. Parameter estimation, hypothesis testing, and regression analysis are applied to build up the model. Multiple linear regression and principal component regression are used to model the data. Based on the data collected from National Bureau of Statistics of China, three principal components, namely basic element component, balance component, and utility component are derived. Reasonable explanations that are consistent with China's reality are made from the principal components and factors to the contribution of the agricultural output. Finally, several reasonable suggestions are given according to the analysis.

Keywords: parameter estimation, hypothesis testing, multiple linear regression, principal component regression, agricultural output

1. Introduction

Agriculture is one of China's basic industries. Therefore the analysis of the relationship between agricultural output and its influencing factors helps to promote the production. The main influencing factors[1] include: The total power of agricultural machinery, employment in primary industry, effective irrigation area, chemical fertilizers, rural electricity consumption, crop acreage, agricultural disaster affected area.

This paper will set up multiple linear regression and principal component regression models to analyze the significant factors that affect agricultural output. The results of the model will be described and explained, and then put forward the proposal on China's agriculture-related aspects.

1.1. Symbols and assumptions

Symbols:

Y (10^8 yuan): Agricultural output;

X_1 (10^4 kw): Total power of agricultural machinery;

X_2 (10^4): Employment in primary industry;

X_3 (10^3 hectares): Effective irrigation area;

X_4 (10^4 tons): Chemical Fertilizers;

X_5 (10^8 kwh): Rural electricity consumption;

X_6 (10^3 hectares): Agricultural disaster affected area;

X_7 (10^4 tons): Crop acreage.

Note: The "disaster affected area" refers to areas suffered from flood, drought, pests, frost, cold, hail, and other natural disasters, leaving the crop yields less than those of normal years. Therefore, it is possible that the reduction of some disaster affected areas is little, so its unit is 10^3 hectares.

Assumptions:

The selected variables have linear relationship with agricultural output.

The collected data is real and effective; some abnormal data are properly processed.

Complex data that can not be obtained and the factors whose effects are small are neglected.

2. Regression models and solutions

2.1. Multiple linear regression

The matrix form[2,3] of the multiple linear regression is expressed as

$$\begin{cases} \boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}, \\ \boldsymbol{\varepsilon} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n), \end{cases}$$

where $\boldsymbol{Y}$ is a dependent variable; $\boldsymbol{\beta}$ is a coefficient vector of the regression equation; $\boldsymbol{\varepsilon}$ is a random error vector which has a standard multivariate normal distribution; $\boldsymbol{X}$ is a matrix carrying the sample data.

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \ \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \ \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}.$$

As proved, the least squares estimation (LSE) of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y}$.

We have the results of LSE as follows:

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}},$$

$$\hat{\sigma}^2 = \hat{\boldsymbol{\varepsilon}}^T \hat{\boldsymbol{\varepsilon}} / (n - p - 1),$$

where $\hat{\sigma}$ represents the standard deviation of the residuals, the smaller the better.

The squared correlation coefficient is denoted as

$$R^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \frac{SS_R}{SS_T},$$

where

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}),$$

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2,$$

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2,$$

$$SS_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2,$$

$$SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2.$$

R^2 is used to measure how closely Y is related to $X_1, X_2, \cdots, X_p$. The bigger the R^2 the better the model fits.

2.2. Principal component regression[2]

The sample data matrix is

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_{(1)}^T \\ \boldsymbol{X}_{(2)}^T \\ \vdots \\ \boldsymbol{X}_{(n)}^T \end{bmatrix} = \left[\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_p\right].$$

The sample covariance matrix is expressed as

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{k=1}^{n}(\boldsymbol{X}_{(k)} - \overline{\boldsymbol{X}})(\boldsymbol{X}_{(k)} - \overline{\boldsymbol{X}})^T = (s_{ij})_{p\times p},$$

where

$$\overline{\boldsymbol{X}} = \frac{1}{n}\sum_{k=1}^{n}\boldsymbol{X}_{(k)} = (\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_p)^T,$$

$$s_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j),$$
$$i, j = 1, 2, \cdots, p.$$

The sample correlation matrix is:

$$\boldsymbol{R} = \frac{1}{n-1}\sum_{k=1}^{n}\boldsymbol{X}_{(k)}^*\boldsymbol{X}_{(k)}^{*\,T} = (r_{ij})_{p\times p},$$

where

$$\boldsymbol{X}_{(k)}^* = \left[\frac{x_{k1} - \bar{x}_1}{\sqrt{s_{11}}}, \frac{x_{k2} - \bar{x}_2}{\sqrt{s_{22}}}, \cdots, \frac{x_{kp} - \bar{x}_p}{\sqrt{s_{pp}}}\right]^T,$$

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}, \; i, j = 1, 2, \cdots, p.$$

We can figure out the principal components based on the correlation matrix $\boldsymbol{R}$. Let $\lambda_1^* \geq \lambda_2^* \geq \cdots \geq \lambda_p^* \geq 0$ denote the eigenvalues of $\boldsymbol{R}$, $\boldsymbol{a}_1^*, \boldsymbol{a}_2^*, \cdots, \boldsymbol{a}_p^*$ denote the unit eigenvectors that are orthogonal to each other. Let

$$\boldsymbol{Z}_{(i)}^* = \boldsymbol{Q}^{*T}\boldsymbol{X}_{(i)}^*, \; i = 1, 2, \cdots, n,$$

where $\boldsymbol{Q}^* = (\boldsymbol{a}_1^*, \boldsymbol{a}_2^*, \cdots, \boldsymbol{a}_p^*)$. Thus

$$\boldsymbol{Z}^* = \begin{bmatrix} z_{11}^* & z_{12}^* & \cdots & z_{1p}^* \\ z_{21}^* & z_{22}^* & \cdots & z_{2p}^* \\ \vdots & \vdots & & \vdots \\ z_{n1}^* & z_{n2}^* & \cdots & z_{np}^* \end{bmatrix} = \begin{bmatrix} \boldsymbol{Z}_{(1)}^{*T} \\ \boldsymbol{Z}_{(2)}^{*T} \\ \vdots \\ \boldsymbol{Z}_{(n)}^{*T} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_{(1)}^{*T}\boldsymbol{Q}^* \\ \boldsymbol{X}_{(2)}^{*T}\boldsymbol{Q}^* \\ \vdots \\ \boldsymbol{X}_{(n)}^{*T}\boldsymbol{Q}^* \end{bmatrix}$$
$$= \boldsymbol{X}^*\boldsymbol{Q}^* = (\boldsymbol{Z}_1^*, \boldsymbol{Z}_2^*, \cdots, \boldsymbol{Z}_p^*).$$

So the relationship between Y and $X_1, X_2, \cdots, X_p$ is expressed as

$$Y = \beta_0^* + \beta_1^* Z_1^* + \cdots + \beta_m^* Z_m^*,$$

$$Z_i^* = a_{1i}^* X_1^* + a_{2i}^* X_2^* + \cdots + a_{pi}^* X_p^* = \sum_{k=1}^{p} \frac{a_{ki}^*(X_k - \bar{X}_k)}{\sqrt{s_{kk}}},$$

$$i = 1, 2, \cdots, m,$$

so we can figure out $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$,

where

$$\beta_0 = \beta_0^* - \sum_{k=1}^{p} \sum_{i=1}^{m} \frac{\beta_i^* a_{ki}^* \bar{X}_k}{\sqrt{s_{kk}}},$$

$$\beta_k = \sum_{i=1}^{m} \frac{\beta_i^* a_{ki}^*}{\sqrt{s_{kk}}}, \; k = 1, 2, \cdots, p.$$

2.3. Solutions

We collected 34 sets of data from National Bureau of Statistics of China. The R software program used in the analysis is as follows (the data file is *lsn.csv*):

```
rm(list=ls())
DA=read.csv("lsn.csv",header=T)
lm.ag=lm(Y~X1+X2+X3+X4+X5+
X6+X7,data=DA)
summary(lm.ag)
```

In analysis, we find that not all of the seven independent variables $X_1, X_2, \cdots, X_7$ are statistically significant because of the multivariable multicollinearity. To see how bad the multicollinearity is, we consider the condition number κ of the matrix $\boldsymbol{X}^T\boldsymbol{X}$. If $\kappa > 1000$, then serious multicollinearity exists. The R program is as follows:

```
XX=cor(DA[1:7])
kappa(XX,exact=TRUE)
```

In this model, $\kappa = 1233.023 > 1000$, so the principal component regression (PCR) is needed to analyze this problem as the multicollinearity exists. The R program is as follows:

```
ag.pr=princomp(~X1+X2+X3+X4
+X5+X6+X7,data=DA,cor=T)
summary(ag.pr,loadings=TRUE)
```

The results are reported in Tables 1, 2, and 3.

Table 1. The summary result of the principal component analysis.

	Comp.1	Comp.2	Comp.3	Comp.4
Standard deviation	2.17067	1.21864	0.793493	0.367783
Proportion of Variance	0.67312	0.2122	0.089947	0.019323
Cumulative Proportion	0.67312	0.88527	0.975222	0.994545
	Comp.5	Comp.6	Comp.7	
Standard deviation	0.1504372	0.1083134	0.0618171	
Proportion of Variance	0.0032330	0.0016760	0.0005459	
Cumulative Proportion	0.9977781	0.9994541	1	

Table 2. The loading matrix of the principal component analysis.

	Z_1^*	Z_2^*	Z_3^*	Z_4^*	Z_5^*	Z_6^*	Z_7^*
X_1^*	0.453	0.107		-0.224	-0.359		0.773
X_2^*		-0.71	-0.602	-0.309		0.162	
X_3^*	0.454			-0.274	0.816	0.217	
X_4^*	0.456					-0.853	-0.226
X_5^*	0.441	0.203	0.11	-0.241	-0.439	0.39	-0.592
X_6^*	0.426	-0.11	-0.219	0.845		0.21	
X_7^*		-0.651	0.75				

Table 3. The $\hat{\sigma}$, R^2, and the significance of the regression coefficients as the number of principal components (PC num) varies.

PC num	$\hat{\sigma}$	R^2	Significance of regression coefficients
2	3184	0.9786	significant
3	2172	0.9904	significant
4	2208	0.9904	Z_4^* is not significant
5	2247	0.9904	Z_4^*, Z_5^* are not significant
6	1950	0.9930	Z_4^*, Z_5^* are not significant
7	1892	0.9937	Z_4^*, Z_5^*, Z_7^* are not significant

Table 1 shows that the cumulative proportion of the first two principal components is 88.5%, so we need at least two principal components. In Table 3, for the standard deviation of the residuals $\hat{\sigma}$, the smaller the better; for the squared correlation coefficient R^2, the bigger the better. Also, by comparison, we find out that three principal components are optimal. The R program for the PCR is as follows:

```
pre=predict(ag.pr)
DA$z1=pre[,1];DA$z2=pre[,2];DA
$z3=pre[,3]
lm.sol=lm(Y~z1+z2+z3,data=DA)
summary(lm.sol)
```

Regression coefficients and regression equations have passed the test as all variables and constants have the significant level of "***". We get a regression equation

$$Y = 21941.3 + 9233.5Z_1^* + 3778.2Z_2^* + 2840.6Z_3^*.$$

This relationship can also be expressed by original variables:

$$\begin{aligned} Y = &-22588.18 + 0.1934973X_1 \\ &-1.222116X_2 + 0.8399706X_3 \\ &+2.570691X_4 + 2.518641X_5 \\ &+0.4320846X_6 + 0.02847791X_7. \end{aligned} \tag{1}$$

3. Conclusions

3.1. Analysis of principal components[4]

From Table 2, we see that

$$Z_1^* = 0.453X_1^* + 0.454X_3^* + 0.456X_4^* + 0.441X_5^* + 0.426X_6^*,$$

For Z_1^*, the coefficients of X_1^*, X_3^*, X_4^*, X_5^*, and X_6^* are close to 0.45, and these variables all indicate basic elements for agricultural production, so we can define this component as a basic element component. It has a positive effect to the output.

$$Z_2^* = 0.107X_1^* - 0.71X_2^* + 0.203X_5^* - 0.11X_6^* - 0.651X_7^*,$$

For Z_2^*, we can see this component emphasizes the power for agriculture and weakens the utility items. It informs us to utilize agricultural resources reasonably, so we can define this component as a balance component. It has a positive effect to the output.

$$Z_3^* = -0.602X_2^* + 0.11X_5^* - 0.219X_6^* + 0.75X_7^*.$$

For Z_3^*, taking into account of the efficiency of labor utilization and the trend of mechanization, we can reasonably define this principal component as a utility component .

3.2. Economic significance analysis[4,5]

For the regression equation (1), the constant represents other factors that are ignored. The coefficients of total power of agricultural machinery (X_1), effective irrigation area (X_3), chemical fertilizers (X_4), rural electricity consumption (X_5), crop acreage (X_7) all represent the positive impacts on the agricultural output (Y). For employment in primary industry (X_2), we consider this negative coefficient as reasonable because of the trend of mechanization and the most efficient ratio of labor and machine. For agricultural disaster affected area (X_6), although the affected area cannot help increase the output, however, we explain it in this way: when the irrigation area is larger, the area's risk of being affected by disasters turns higher.

Finally, we strongly recommend that better machines are designed to improve the energy efficiency and thus reduce the energy consumption; invest more on agricultural infrastructure in order to promote the process of mechanization.

4. References

[1] Chunhua Chang, "Analysis on factors of Chinese agricultural output," *Rural Economy and Science*, 2006 (8): 51-52.

[2] Yi Xue, Liping Chen, *Statistical Modeling with R Software*, Beijing: Tsinghua University Press, 2007.

[3] Hu Yang, Qiongsun Liu, Bo Zhong, *Mathematical Statistics*, Beijing: Higher Education Press, 2004.

[4] Han Liu, Zuwen Cao, "Investment in infrastructure, human capital accumulation and economic growth in agriculture," Economic Issues, 2012 (12): 84-90.

[5] Yue Zhao, Nan Yang, "Empirical analysis on grain production in China," *Knowledge Economy*, 2012 (9): 78.