ISSN: 2688-9323

Research on the Related Factors of Residents' Consumption Expenditure based on Principal Component Analysis

Xiaoyun Xu

Qufu Normal University, Jining, Shandong, 273165, China

Abstract

China's residents' consumption expenditure is affected by many factors. In order to establish a more reasonable model between consumption expenditure and relevant factors, this paper used the data of various provinces in 2020 and took consumption expenditure as the dependent variable to select residents' food expenditure, residents' clothing expenditure, residents' living expenditure, residents' daily necessities expenditure, residents' transportation and communication expenditure, residents' education, culture and entertainment expenditure, residents' health care expenditure, per capita disposable income and consumer price index are used as explanatory variables. Firstly, some variables are selected by the stepwise regression method, and then the basic hypothesis test and multiple collinearity test are carried out on the linear regression model established by the dependent variable and the selected variable. It is found that the model has multiple collinearity, and then the principal component analysis method is used to solve the problem of multiple collinearity of the model, Finally, a more reasonable linear regression model between consumer expenditure and relevant factors is obtained.

Keywords

Consumption Expenditure; Stepwise Regression; Principal Component Analysis and Linear Regression Model.

1. Introduction

Since the reform and opening, with the rapid development of China's economy, earth shaking changes have taken place in residents' living standards and lifestyles. In the process of economic development, consumption has become an increasingly important factor. At the same time, there are many factors restricting the level of consumption. Therefore, this paper focuses on the relevant factors affecting residents' consumption expenditure, to analyze the economic situation of our country. There are many factors affecting residents' consumption expenditure. The factors selected in this paper are: Residents' food expenditure, residents' clothing expenditure, residents' living expenditure, residents' daily necessities expenditure, residents' transportation and communication expenditure, residents' education, culture and entertainment expenditure, residents' medical care expenditure, per capita disposable income and consumer price index. Through these factors and the data of various provinces in China in 2020, explore the specific relationship between residents' consumption expenditure and its related factors [1].

2. Materials and Methods

2.1. Data Sources

The data of consumption expenditure, residents' food expenditure, residents' clothing expenditure, residents' living expenses, residents' daily necessities expenditure, residents' transportation and communication expenditure, residents' education, culture and

ISSN: 2688-9323

entertainment expenditure, residents' medical and health care expenditure, per capita disposable income and consumer price index in 2020 are all from China's statistical yearbook.

2.2. Description of Data and Variables

Per capita disposable income: the disposable income of residents refers to the sum of final consumption expenditure and savings available to residents, that is, the income available to residents for free disposal. It includes both cash income and in-kind income.

Consumer price index: it is a relative number reflecting the price change trend and degree of consumer goods and services purchased by urban and rural residents in a certain period.

In order to study the dependent variable y the influencing factors of consumer expenditure, the explanatory variables and corresponding variable symbols selected by us are shown in Table 1:

Table 1. Explanatory variables and variable symbols

Food costs for residents	x1
Clothing cost of residents	x2
Residents' living expenses	x3
Consumption of daily necessities for residents	x4
Residents' traffic and communication costs	x5
Residents' educational culture and entertainment expenses	x6
Residents' health care costs	x7
Disposable income per capita	x8
Consumer price index	x9

2.3. Analysis Method and Main Process

2.3.1. Stepwise Regression:

The main idea of stepwise regression is to introduce variables into the model one by one. After each variable is introduced, the selected variables are tested one by one. When the original introduced variables become no longer significant due to the introduction of later variables, they should be eliminated. Introducing a variable or removing a variable from the regression equation is a step of stepwise regression. F test shall be conducted at each step to ensure that there are only significant variables in the regression equation before introducing new variables. This process is repeated until neither significant independent variables are selected into the regression equation nor insignificant independent variables are removed from the regression equation [3].

2.3.2. Principal Component Analysis:

Principal component analysis (PCA), also known as principal component analysis, was first proposed by Hotelling in 1933. Principal component analysis is a multivariate statistical analysis method that transforms multiple indexes into several comprehensive indexes by orthogonal rotation transformation on the premise of losing little information. Generally, the comprehensive index generated by transformation is called principal component, in which each principal component is a linear combination of original variables, and each principal component is not related to each other. In this way, when studying complex problems, we can only consider a few principal components without losing too much information, so it is easier to grasp the main contradiction, reveal the regularity between internal variables, simplify the problem and improve the analysis efficiency [4].

ISSN: 2688-9323

2.3.3. Variance Expansion Factor Method:

After the central standardization of the independent variable, note that the standardized design matrix is X^*, and $X^{*'}X^*$ is the correlation matrix of the independent variable. $C=(c_{ij})=(X^{*'}X^*)^{-1}$. The main diagonal element $VIF_j=c_{jj}$ is called the variance expansion factor of the independent variable VIF_j. The size of $VIF_j \geq 10$ reflects whether there is multicollinearity between independent variables. Therefore, it can be used to measure the severity of multicollinearity. Experience shows that when $VIF_j \geq 10$, there is serious multicollinearity between independent variable x_j and other independent variables [5].

2.4. Main Process

(1) Firstly, based on the original data, a theoretical linear regression model of consumer expenditure y and all explanatory variables is established:

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\beta_6x_6+\beta_7x_7+\beta_8x_8+\beta_9x_9+\varepsilon. \quad (1)$$

Statistical software was used to estimate the values of the unknown parameters in the model, and the model saliency test, equation significance test and goodness of fit test were carried out.

(2) Stepwise regression method was used to select variables, and linear regression equation between dependent variable y and selected variables was established to model saliency test, equation significance test and goodness of fit test.

(3) The new regression equation was tested by heteroscedasticity test, autocorrelation test and collinearity test;

(4) Principal component analysis is used to eliminate the problem of collinearity of the model, and a reasonable regression equation is obtained to explain the relationship between dependent variable y and explanatory variables.

3. Results and Analysis

3.1. Establish a Linear Regression Model of Dependent Variables to All Explanatory Variables

The results obtained by using statistical software are shown in:

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9,
    data = a)

Residuals:
     Min       1Q   Median       3Q      Max
-119.889  -29.908   -6.642   25.251  148.834

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.951e+03  3.130e+03   0.624    0.540
x1           1.040e+00  1.721e-02  60.436   < 2e-16 ***
x2           1.168e+00  1.023e-01  11.416  1.82e-10 ***
x3           9.818e-01  2.614e-02  37.560   < 2e-16 ***
x4           8.962e-01  1.142e-01   7.849  1.12e-07 ***
x5           1.055e+00  4.085e-02  25.818   < 2e-16 ***
x6           9.435e-01  5.450e-02  17.311  6.59e-14 ***
x7           1.093e+00  4.484e-02  24.376   < 2e-16 ***
x8           8.727e-03  8.368e-03   1.043    0.309
x9          -2.159e+01  3.036e+01  -0.711    0.485
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 66.28 on 21 degrees of freedom
Multiple R-squared:  0.9999,    Adjusted R-squared:  0.9999
F-statistic: 3.473e+04 on 9 and 21 DF,  p-value: < 2.2e-16
```

Figure 1. Linear regression model of dependent variable to all explanatory variables

ISSN: 2688-9323

It can be seen from the results that the coefficients and intercept terms of X_8 and X_9 in the explanatory variables ***cannot*** pass the significance test, and the other coefficients and equations can pass the significance test. At this time, the linear regression equation obtained is:

$$y = 1.04x_1 + 1.168x_2 + 0.9818x_3 + 0.8962x_4 + 1.055x_5 + 0.9435x_6 + 1.093x_7 + 0.0087x_8 - 21.59x_9 + 1951. \quad (2)$$

3.2. Stepwise Regression is Used to Select Variables

Since some of the coefficients in the linear regression equations of y and all explanatory variables cannot pass the significance test, the stepwise regression method is used to select the independent variables.

First, calculate the correlation coefficients between all explanatory variables and dependent variable y, and the results are as follows:

Table 2. Correlation coefficient between explanatory variable and dependent variable

x1	*x2*	*x3*	*x4*	*x5*	*x6*	*x7*	*x8*	*x9*
0.8995145	*0.6544698*	*0.9674977*	*0.9366022*	*0.8913604*	*0.8747567*	*0.7651212*	*0.9854305*	*-0.355435*

According to the results, the explanatory variable with the strongest correlation with y is x8, and the correlation coefficient reaches 0.985. Therefore, stepwise regression is carried out based on the linear relationship between y and x8.

```
Step:  AIC=266.13
y ~ x1 + x7 + x5 + x3 + x4 + x6 + x2

       Df Sum of Sq      RSS    AIC
<none>                  98996 266.13
+ x8    1      4521    94475 266.69
+ x9    1      1964    97032 267.51
- x4    1    266780   365776 304.65
- x2    1    951273  1050269 337.35
- x6    1   1556555  1655552 351.46
- x7    1   2613704  2712700 366.76
- x5    1   3154910  3253906 372.40
- x1    1  19097818 19196814 427.42
- x3    1  50524610 50623606 457.48
```

Figure 2. Stepwise regression results

The statistical software is used for stepwise regression to select the final explanatory variables: x1, x2, x3, x4, x5, x6, x7, and the eliminated variables are the independent variables corresponding to the coefficients that have not passed the significance test in the linear regression equation.

Establish a linear regression model for y and the remaining variables, and the results are as follows:

The new linear regression equation is:

$$y = 1.05x_1 + 1.24x_2 + 1.01x_3 + 0.87x_4 + 1.06x_5 + 0.96x_6 + 1.09x_7 - 285.49. \quad (3)$$

ISSN: 2688-9323

```
Call:
lm(formula = y ~ x1 + x7 + x5 + x3 + x4 + x6 + x2, data = a

Residuals:
     Min       1Q   Median       3Q      Max
-108.023  -28.584   -4.816   27.890  163.594

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -285.48822   83.63328  -3.414  0.00238 **
x1             1.04733    0.01572  66.611  < 2e-16 ***
x7             1.09361    0.04438  24.642  < 2e-16 ***
x5             1.06407    0.03930  27.074  < 2e-16 ***
x3             1.00867    0.00931 108.344  < 2e-16 ***
x4             0.87027    0.11054   7.873 5.64e-08 ***
x6             0.96104    0.05054  19.017 1.45e-15 ***
x2             1.23925    0.08336  14.866 2.76e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Residual standard error: 65.61 on 23 degrees of freedom
Multiple R-squared:  0.9999,    Adjusted R-squared:  0.9999
F-statistic: 4.557e+04 on 7 and 23 DF,  p-value: < 2.2e-16
```

Figure 3. Linear regression model of dependent variable and selected variable

It can be seen from the figure that all coefficients and regression equations have passed the significance test, and the goodness of fit has reached 99.99%.

3.3. Test of Linear Regression Equation

3.3.1. Test of Heteroscedasticity for Linear Regression Equation

(1) Heteroscedasticity

In the basic assumptions of the regression model, it is assumed that the random term error $\varepsilon_1, \varepsilon_2, \cdots \varepsilon_n$ has the same variance, and heteroscedasticity is a violation of the basic assumptions, that is: $\text{var}(\varepsilon_i) \neq \text{var}(\varepsilon_j), \text{i} \neq \text{j}$.

(2) Heteroscedasticity test

The results of heteroscedasticity test for x1-x7 these seven independent variables are as follows:

```
        Spearman's rank correlation rho

data:  a[, 1] and abs(e)
S = 4244, p-value = 0.4369
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1443548

        Spearman's rank correlation rho

data:  a[, 2] and abs(e)
S = 3712, p-value = 0.1716
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.2516129

        Spearman's rank correlation rho

data:  a[, 3] and abs(e)
S = 3990, p-value = 0.2904
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1955645
```

(a)

ISSN: 2688-9323

```
        Spearman's rank correlation rho

data:  a[, 4] and abs(e)
S = 4950, p-value = 0.9922
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.002016129

        Spearman's rank correlation rho

data:  a[, 5] and abs(e)
S = 5086, p-value = 0.8923
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.02540323

        Spearman's rank correlation rho

data:  a[, 6] and abs(e)
S = 4208, p-value = 0.4139
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.1516129

        Spearman's rank correlation rho

data:  a[, 7] and abs(e)
S = 4506, p-value = 0.6231
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.09153226
```

(b)

Figure 4. Heteroscedasticity test results

It can be seen from the results that the P values of the test results are greater than the significance level $\alpha = 0.05$, so there is no sufficient reason to believe that their variables have heteroscedasticity.

3.3.2. Autocorrelation Test of Linear Regression Equation

(1) Autocorrelation

In the basic assumptions of the regression model, it is assumed that the random term error is uncorrelated, and autocorrelation is a violation of the basic assumptions, that is: $\mathrm{cov}(\varepsilon_i, \varepsilon_j) \neq 0, \mathrm{i} \neq \mathrm{j}.$

(2) Autocorrelation test

The autocorrelation test results by statistical software are shown in the figure below:

```
        Durbin-Watson test

data:  both
DW = 2.5156, p-value = 0.2587
alternative hypothesis: true autocorrelation is not 0
```

Figure 5. Autocorrelation test

The P values obtained are greater than the given significance level $\alpha = 0.05$, so there is no sufficient reason to believe that the respective variables have autocorrelation. Therefore, it is considered that there is no heteroscedasticity and autocorrelation in this linear model.

ISSN: 2688-9323

3.3.3. Multiple Collinearity Test of Linear Regression Equation

(1) Multicollinearity

A basic assumption of the multiple linear regression model is that the rank of the design matrix X is required:

$$\text{rank}(X) = p+1 \tag{4}$$

That is, the column vectors in X are required to be linearly independent. If there are p + 1 numbers $c_0, c_1, c_2, \cdots, c_p$ that are not all 0, so that:

$$c_0 + c_1 x_{i1} + c_2 x_{i2} + \cdots + c_p x_{ip} = 0, i = 1, 2, \cdots, n \tag{5}$$

Then there is complete multicollinearity between independent variables $x_1, x_2, \cdots, x_p$. In practical problems, it is common to make the above formula approximately true, that is, there is a p + 1 number $c_0, c_1, c_2, \cdots, c_p$ that is not all 0, so that:

$$c_0 + c_1 x_{i1} + \cdots c_2 x_{i2} + \cdots + c_p x_{ip} \approx 0, i = 1, 2, \cdots, n \tag{6}$$

(2) multicollinearity test

The variance expansion factor method is used to test multicollinearity. The variance expansion factors of the seven explanatory variables are calculated as follows:

Table 3. Variance expansion factors of independent variables

x1	x2	x3	x4	x5	x6	x7
5.090573	3.803057	5.943221	10.43284	5.577058	5.168216	3.741155

Because of $VIF_4 \geq 10$, it is considered that there is serious multicollinearity between independent variable x_4 and other variables.

3.4. Eliminating Collinearity

The variance percentage reflects the proportion of data variation that can be explained by the principal components, that is, the information proportion containing the original data. The variance percentage of each principal component calculated by the statistical software is:

```
Importance of components:
                            Comp.1       Comp.2       Comp.3       Comp.4
Standard deviation     3518.3374123 917.10428282 458.07663799 3.101012e+02
Proportion of Variance    0.9088471   0.06175232   0.01540608 7.060293e-03
Cumulative Proportion     0.9088471   0.97059944   0.98600552 9.930658e-01
                            Comp.5       Comp.6       Comp.7
Standard deviation     2.468973e+02 1.562942e+02 95.178778682
Proportion of Variance 4.475572e-03 1.793502e-03  0.000665114
Cumulative Proportion  9.975414e-01 9.993349e-01  1.000000000
```

Figure 6. Cariance percentage of each principal component

The variance percentage of the first principal component = 90.88%, which contains more than 90% of the information of seven original variables. For principal component analysis, the cumulative variance percentage of the selected principal component is required to be more than 80%, so it is enough to take one principal component here. Using the eigenvector corresponding to the largest eigenvalue as the coefficient, calculate the score of the first principal component y1 as follows:

ISSN: 2688-9323

```
> y1
 [1] -18832.598 -11150.783  -6899.389  -5722.513  -7141.337  -7575.651
 [7]  -6101.806  -6152.659 -19819.542 -10908.969 -13143.413  -7381.619
[13] -10975.460  -7202.537  -7509.931  -6040.275  -7538.755  -7679.971
[19] -12113.945  -6330.041  -7657.516  -7751.859  -7317.820  -5396.059
[25]  -6150.567  -5203.884  -6363.957  -5979.505  -6441.823  -6018.708
[31]  -5979.712
```

Figure 7. Score of the first principal component

Do the ordinary least squares regression of y to y1, and the results are as follows:

```
Call:
lm(formula = y ~ y1)

Residuals:
     Min       1Q   Median       3Q      Max
-2040.98  -615.88   -90.03   707.83  2063.01

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5524.10129  454.63705   12.15 6.69e-13 ***
y1            -1.87200    0.05057  -37.02  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 990.6 on 29 degrees of freedom
Multiple R-squared:  0.9793,    Adjusted R-squared:  0.9786
F-statistic:  1370 on 1 and 29 DF,  p-value: < 2.2e-16
```

Figure 8. Ordinary least squares regression of y to y1

The principal component regression equation obtained is:

$$y = -1.872y_1 + 5524.1. \tag{7}$$

The coefficient and regression equation passed the significance test, and the goodness of fit reached 97.86%. Therefore, it is considered that this model can be used to fit the relationship between y and y1.

3.5. Obtain the Final Equation

Establish the linear regression equation of principal component y1 and other variables:

```
Call:
lm(formula = y1 ~ x1 + x2 + x3 + x4 + x5 + x6 + x7)

Residuals:
       Min         1Q     Median         3Q        Max
-3.352e-12 -7.842e-13 -1.161e-13  8.151e-13  3.605e-12

Coefficients:
             Estimate Std. Error    t value Pr(>|t|)
(Intercept)  2.287e-12  2.174e-12  1.052e+00    0.304
x1          -4.234e-01  4.086e-16 -1.036e+15   <2e-16 ***
x2          -4.671e-02  2.166e-15 -2.156e+13   <2e-16 ***
x3          -8.688e-01  2.420e-16 -3.591e+15   <2e-16 ***
x4          -8.864e-02  2.873e-15 -3.085e+13   <2e-16 ***
x5          -1.696e-01  1.021e-15 -1.660e+14   <2e-16 ***
x6          -1.258e-01  1.313e-15 -9.579e+13   <2e-16 ***
x7          -1.059e-01  1.153e-15 -9.179e+13   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.705e-12 on 23 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 1.886e+31 on 7 and 23 DF,  p-value: < 2.2e-16
```

Figure 9. Linear regression equation of principal components and other variables

The equation obtained is:

ISSN: 2688-9323

$$y1 = -0.423x_1 - 0.047x_2 - 0.869x_3 - 0.089x_4 - 0.17x_5 - 0.126x_6 - 0.106x_7. \quad (8)$$

Substitute into equation $y = -1.872y_1 + 5524.1$, and the final equation is:

$$y = 0.792x_1 + 0.088x_2 + 1.627x_3 + 0.167x_4 + 0.318x_5 + 0.236x_6 + 0.198x_7 + 5524.1. \quad (9)$$

4. Conclusion

(1) Residents' food expenses, clothing expenses, living expenses, daily necessities expenses, transportation and communication expenses, education, culture and entertainment expenses and medical care expenses have a positive impact on Residents' consumption expenditure, which is the factor to promote the increase of consumption expenditure.
(2) Residents' living expenses have the greatest impact on consumer spending, and residents' clothing expenses have the least impact on consumer spending.
(3) For every additional unit of residents' food expenditure, the consumption expenditure will increase by 0.792 units; For each additional unit of clothing expenditure, the consumption expenditure will increase by 0.088 units; For every additional unit of housing expenditure, the consumption expenditure will increase by 1.627 units; For every additional unit of daily necessities, the consumption expenditure will increase by 0.167 units; For each additional unit of transportation and communication expenditure, the consumption expenditure will increase by 0.318 units; For every additional unit of education and culture expenditure, the consumption expenditure will increase by 0.236 units; For every additional unit of health care expenditure, consumer expenditure will increase by 0.198 units.

References

[1] Zhang Wenjie Empirical analysis of relevant factors affecting residents' consumption expenditure in various regions of China [J] China business theory, 2016,5:114.

[2] National Bureau of statistics of the people's Republic of China China Statistical Yearbook 2021 [Z], Beijing, 2021.

[3] He Xiaoqun, Liu Wenqing Applied regression analysis [M]. Version 5 Beijing: China Renmin University Press, 2021:143.

[4] He Xiaoqun, Liu Wenqing Applied regression analysis [M]. Version 5 Beijing: China Renmin University Press, 2021:183.

[5] He Xiaoqun, Liu Wenqing Applied regression analysis [M]. Version 5 Beijing: China Renmin University Press, 2021:157.