# Statistical analysis and prediction of Qingdao urban consumer price index

**Linlin Dai***

*School of Qingdao Huanghai College, No.3111, Ling Hai Road, Qingdao Shandong Province, China*

*Received 6 July 2014, www.cmnt.lv*

**Abstract**

In recent years, Qingdao faces inflationary pressure along with the fast development of economy. The urban consumer price index (CPI) is an important measure of regional inflation level. Thus, analysing the influencing factors of the CIP and determining the key factors is necessary to supply scientific evidence for the economic management departments to adjust the industrial structure. This paper filtrates and eliminates six indexes likely to affect urban CPI of Qingdao based on Studentized residual, correlation analysis, and multiple collinearity diagnosis. This paper also performs a statistical analysis of the urban CPI of Qingdao from 2005 to 2012 and draws the conclusion that food price is the largest index that affects CPI. Basing on this conclusion, corresponding prediction models are established to supply statistics for departments that adjust the economy for the future.

*Keywords:* Studentized residual, correlation analysis, collinearity diagnosis, multiple linear regression model

## 1 Introduction

Along with the rapid development of the Chinese economy, the consumption level is rising and has become an important measure of a regional economy. Consumer price index (CPI) has also become an important measure of the level of inflation. Much research has been conducted on CPI. For example, Zhu Wei and Zhong Weijian have researched the application of the ARMA model in CPI [1]. He Weiwei and Tianhao have analysed the influencing factors of CPI according to the VAR model [2]. Yu Hongyan and Chu Delin have analysed the dynamic correlation between the Chinese customer price index and food price index according to the H-P filtering method [3]. Currently, although the economy of Qingdao is developing rapidly, it also suffers from inflationary pressure. The major reasons for this pressure are that industrial structure is irrational to some extent, and some industries fail to meet the needs of the society. This paper, based on an analysis of Qingdao urban CPI statistics, attempts to determine the positive and negative factors that affect economic development and offers some proposals about optimizing the industrial structure to reduce the pressure of inflation.

The possible factors affecting urban CPI (X1) include the following: the commodity retail price index (X2), the food price index (X3), industrial producers purchase price index (purchasing price index of raw material, fuel, and power) (X4), the producer price index (X5), the price index of investment in fixed assets (X6), and service price index (X7). The index data are collected from the online Statistical Yearbook of Qingdao from 2005 to 2012 and other related information sources. After preparing and analysing the data in advance, the Studentized residuals and related analysis are used to determine the degree of correlation among these indexes and to further filter out the most important indicator of the urban CPI. Multi-collinearity diagnosis is employed for the selected variables to estimate the capability of the multiple regression model. Finally, the selected indicators are used to establish a regression model for the principle component of the urban CPI and to further analyse the effectiveness and accuracy of the model. Given that the real effect factor is fully considered, this model reflects the status of multiple layers in CPI effect well, and both its fitness and significance are superior.

An applied technical route is used as follows: Raw data extraction - gross error analysis (Studentized residual analysis) - related analysis, - multicollinearity diagnosis - multiple regression model is established.

## 2 Establish mathematical model and solve model

### 2.1 DATA PREPROCESSING

#### *2.1.1 Raw data extraction*

Given that a human extracted the raw data used, some measurement errors are inevitable. Thus, the data should be pre-processed first. The index data of some years are incomplete. Thus, we finally chose Qingdao data from 2005 to 2012 to analyse the CPI of this city. Although the sample size decreased, the actual test data can be reflected maximally. The primary index data [4] are shown in Table 1:

* ***Corresponding author*** e-mail: dailinlin79@163.com

TABLE 1 Annual target data of the primary variables (Last year = 100)

| Year | X1 | X2 | X3 | X4 | X5 | X6 | X7 |
|---|---|---|---|---|---|---|---|
| 2005 | 102.3 | 99.3 | 102.1 | 106.5 | 101.4 | 103.1 | 101.4 |
| 2006 | 100.9 | 99.7 | 101.7 | 105.8 | 101.1 | 102.6 | 100.8 |
| 2007 | 104.5 | 102.7 | 111.6 | 106.7 | 101.5 | 104.1 | 102.4 |
| 2008 | 104.7 | 103.9 | 111.7 | 115.9 | 105.3 | 111.0 | 99.8 |
| 2009 | 100.5 | 98.6 | 101.6 | 90.4 | 95.8 | 94.6 | 100.6 |
| 2010 | 102.2 | 101.4 | 106.4 | 112.2 | 103.8 | 104.7 | 100.4 |
| 2011 | 105.0 | 104.5 | 111.1 | 109.4 | 104.9 | 107.0 | 103.1 |
| 2012 | 102.7 | 101.7 | 104.3 | 97.0 | 98.6 | 100.0 | 102.1 |

*2.1.2 Gross Error Analysis*

Bulky errors or gross errors, also known as outliers, are individual values in the samples that obviously deviate from the other samples [5]. When processing the data, if the outliners mixed in the measurement data are not eliminated, the statistical analysis would be less accurate. Conversely, if normal data have been eliminated as outliers, the statistical analysis would have less information. Hence, correct result must first determine and process the outliners rationally.

This paper uses the Studentized residual to analyse the gross error of the original data. The basic concepts of Studentized residuals [6] are depicted as follows:
Assuming $n$ times independent measurements are obtained, the column measurement $X$ is obtained as follows:

$$X = \{x_1, x_2, ..., x_i, ..., x_n\} \ (i = 1,2,...,n). \tag{1}$$

The mean $\bar{x}$ and residual $v_i$ respectively are:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \tag{2}$$

$$v_i = x_i - \bar{x}. \tag{3}$$

Definition 1: Sample Studentized residual is the ratio of residuals to the sample standard deviation, that is:

$$y_i = \frac{v_i}{s}. \tag{4}$$

In the following formula, $s$ stands for a sample standard deviation, and the value of $s$ is given by:

$$s = \sqrt{\frac{\sum_{i=1}^{n} v_i^2}{n}}. \tag{5}$$

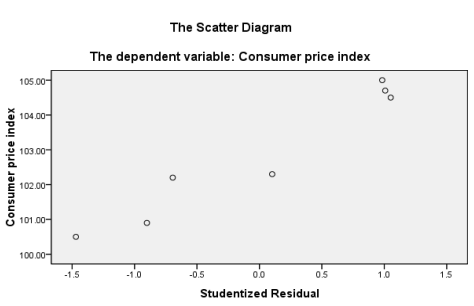
Definition 2: The absolute value of the sample Studentized residual is the ratio of the absolute value of the residuals to the absolute value of the sample standard deviation, that is

$$|y_i| = \frac{|v_i|}{s}. \tag{6}$$

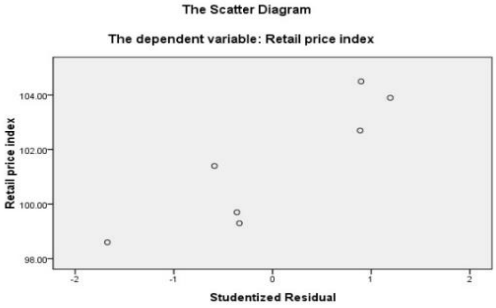
"Studentized" here refers mainly to the common residual without homogeneity of variance. We suppose that expectation is 0 and variance is 1 [7] for easier comparison.

According to the above definition of Studentized residual, the annual data of each indicator are processed. The result is shown in Figure 1. When the absolute value of the Studentized residual is more than 3, the data should be eliminated as variable data, that is, the gross error. However, Figure 1 shows that each absolute value of the indexes' Studentized residuals is less than 3. Thus, the above index data have no gross error and can be directly used for further analysis.

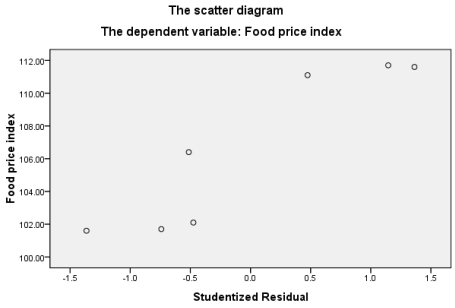
The above value indexes are all prices that share the same unit. Thus, they do not need to be further standardized.



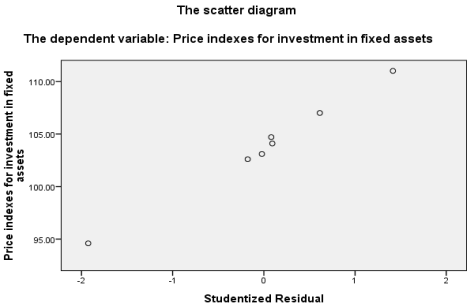

a) Consumer price index



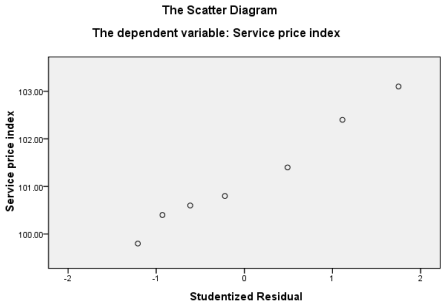

b) Retail price index



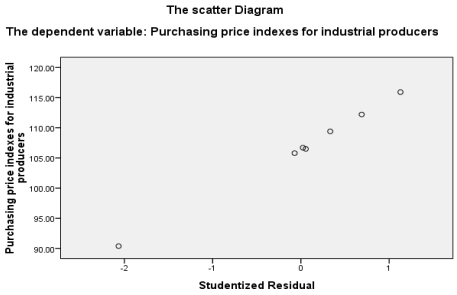

c) Food price index



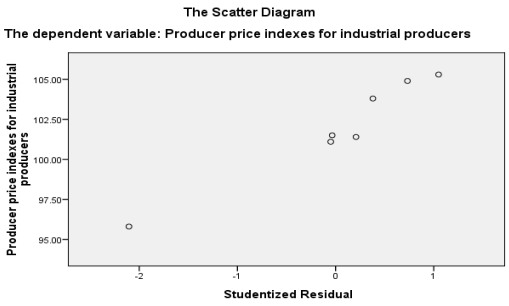

d) Price indexes for investment in fixed assets




e) Service price index




f) Purchasing price indexes for industrial producers




g) Producer price indexes for industrial producers

FIGURE 1 Studentized residual figure of each index variable data

## 2.2 INDEX VARIABLES ANALYSIS BASED ON CORRELATION ANALYSIS

Correlation analysis is a statistical method that studies the quality of relationships between variables. Through correlation analysis, this paper presents a preliminary understanding between urban CPI and other indexes. Thus, factors that are not significant are eliminated to reduce model dimension.

Pearson correlation coefficient between any two variables can be calculated according to Equation (7):

$$\rho = \frac{\sum_{k=1}^{n}(X_{ik}-\overline{X}_i)(X_{jk}-\overline{X}_j)}{\sqrt{\sum_{k=1}^{n}(X_{ik}-\overline{X}_i)^2\sum_{k=1}^{n}(X_{jk}-\overline{X}_j)^2}}, \quad (7)$$

$\overline{X}$ is the average of single variables, and $X_{ik}$ represents the *k-th* variable in the *i-th* data.

SPSS software was used for the correlation analysis of the data in Table 1. The Pearson correlations among each index value can be obtained, as shown in Table 2:

TABLE 2 Pearson correlation coefficient matrix among each index

| | | X1 | X2 | X3 | X4 | X5 | X6 | X7 |
|---|---|---|---|---|---|---|---|---|
| X1 | Pearson correlation | 1 | .936** | .938** | .773* | .436 | .614 | .709* |
| | Significant (both sides) | | .001 | .001 | .024 | .280 | .105 | .049 |
| X2 | Pearson correlation | .936** | 1 | .935** | .777* | .357 | .618 | .740* |
| | Significant (both sides) | .001 | | .001 | .023 | .386 | .102 | .036 |
| X3 | Pearson correlation | .938** | .935** | 1 | .750* | .274 | .628 | .702 |
| | Significant (both sides) | .001 | .001 | | .032 | .511 | .095 | .052 |
| X4 | Pearson correlation | .773* | .777* | .750* | 1 | −.040 | .953** | .963** |
| | Significant (both sides) | .024 | .023 | .032 | | .926 | .000 | .000 |
| X5 | Pearson correlation | .436 | .357 | .274 | −.040 | 1 | −.148 | .008 |
| | Significant (both sides) | .280 | .386 | .511 | .926 | | .727 | .985 |
| X6 | Pearson correlation | .614 | .618 | .628 | .953** | −.148 | 1 | .968** |
| | Significant (both sides) | .105 | .102 | .095 | .000 | .727 | | .000 |
| X7 | Pearson correlation | .709* | .740* | .702 | .963** | .008 | .968** | 1 |
| | Significant (both sides) | .049 | .036 | .052 | .000 | .985 | .000 | |

**. At the .01 level (both sides) significant correlation
*. At the 0.05 level (both sides) significant correlation

Generally speaking, two vectors will have strong correlation when the absolute value of the Pearson correlation coefficient is greater than 0.8. According to this criterion, the correlation is not strong between urban CPI and investment in fixed assets, service price, industrial producers purchase price, or producer price. Therefore, these four indexes can be eliminated first. Then, the rest of the indicators are analysed.

## 2.3 MULTIPLE COLLINEARITY DIAGNOSIS

Two variables remain after eliminating unimportant variables, and multiple collinearity diagnosis must be used to determine whether a linear correlation is present between these two variables.

Collinearity can be measured through tolerance [8]. If the value is less than 0.1, then a collinearity problem may be present between the two indexes. The tolerance between retail price and food price in SPSS software is shown in Table 3:

TABLE 3 Tolerance calculation table

| Dependent | Independent | Collinearity statistics | |
|---|---|---|---|
| | | *Tolerance* | *VIF* |
| X2 | X3 | 1.000 | 1.000 |
| X3 | X2 | 1.000 | 1.000 |

Table 3 shows that the tolerance is close to 1, and the variance inflation factor is also close to 1. Thus, the independent variables X2 and X3 have a weak collinearity problem.

## 2.4 URBAN CPI BASED ON MULTIPLE LINEAR REGRESSION MODEL

A multiple linear regression model is a linear regression model with several explanatory variables. Such models are used to illustrate the linear relationship among explained variables and several explanatory variables, and its regression equation is given by:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n . \quad (8)$$

In this paper, the multiple linear regression model is analysed using SPSS software, and its process is as follows:

On the basis of previous correlation analysis, two indicators are significant: the commodity retail price index (X2) and the food price index (X3). The original data of these two indexes in Table 1 are extracted and are used to establish multiple linear regression models by using the stepwise regression method. The results are summarized in Tables 4-7:

TABLE 4 Input/remove variables

| Model | Input variable | Remove variable | Methods |
|---|---|---|---|
| 1 | X3 | — | Step by step (guidelines: The probability of F-to-enter <= .050. The probability of F-to-remove >= .100). |

a. dependent: X1

TABLE 5 Model summary

| Model | R | R2 | Adjusted R2 | Standard estimate error |
|---|---|---|---|---|
| 1 | .938a | .880 | .861 | .64430 |

a. Independent: (constant), X3

TABLE 6 Coefficient table

| Model | | Standardized coefficients | | Standard coefficient | t | Sig. |
|---|---|---|---|---|---|---|
| | | *B* | *Standard error* | *Trial version* | | |
| 1 | (constant) | 65.056 | 5.689 | — | 11.435 | .000 |
| | X3 | .356 | .053 | .938 | 6.648 | .001 |

a. Dependent: X1

TABLE 7 Variables have been ruled out

| Model | | Beta In | T | Sig. | Partial correlation | Collinearity statistics |
|---|---|---|---|---|---|---|
| | | | | | | *tolerance* |
| 1 | X2 | .464a | 1.210 | .280 | .476 | .126 |

a. Prediction of the variables in the model: (constant), X3
b. Dependent: X1

In the process of stepwise regression, the variable X2 has been removed. Therefore, only X3, the food price index, has been left as an independent variable. Table 5 shows that the model has extremely good fitting. Table 6

shows that the variable coefficient is significant ($\rho < 0.05$). According to the above regression results, the prediction model of urban CPI can be concluded as:

$$X_1 = 65.056 + 0.356X_3 + \varepsilon . \quad (9)$$

Figure 2 illustrates that the observed and predicted values tend to agree with each other, which further indicates the good quality of the model.

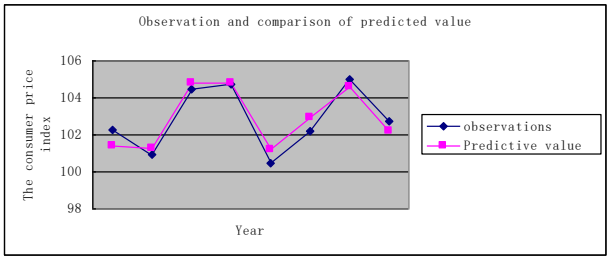



FIGURE 2 Comparison chart of observed and predicted CPI values

In this paper, the statistical analysis and prediction of the Qingdao urban residents' consumption index have been proposed. The close relationship between the urban CPI and food prices will greatly help relevant government departments make correct decisions. Government departments must strictly control food prices to ensure the stability of the urban CPI, which will maintain the balance of the industrial economy in Qingdao City as well as maintain a healthily developed and sustained economy.

## 3 Conclusion

This article performs statistical analyses and predictions of influencing factors on urban CPI in Qingdao and concludes that a close relationship exists between urban CPI and food price. Figure 2 illustrates that a stable long-run equilibrium relationship exists between CPI and food price index. Over the past decade, food price has been the main influencing factor of CPI. Additionally, through Studentized residual and correlation analysis, we can infer that non-food prices also affect CPI to some extent, and the effect will gradually increase with the development of the society. Therefore, the government department concerned must focus on the effect of food prices when considering CPI and price level. Meanwhile, in the long process of revising food prices to stabilize CPI and price level, other relative prices, especially some affecting factors that change widely or greatly, must not be ignored. Doing so will promote the sound, balanced, and sustained development of the industrial economy.

## Acknowledgments

This research is supported by Shandong Province Higher Educational Humanities and Social Sciences Program (No. J12WG75).

## References

[1] Zhu W, Zhong W J 2008 Application of ARMA Model in Predicting the Consumer Price Index *Finance & Economy* **16** 82-3 *(in Chinese)*
[2] He W W, Tian H 2007 Analysis of Influencing Factors of Consumer Price Index *Statistics and Decision* **23** 66-8 *(in Chinese)*
[3] Yu H Y, Chu D Y 2011 Analysis of The dynamic Correlation between China's Consumer Price Index and Food Price *Index Price: Theory & Practice* **2** 52-3
[4] Qingdao Statistical Information Network *statistical yearbook of Qingdao*. http://www.stats-qd.gov.cn/statsqd/Columns/2013nj.asp?typeid=1930&videos=&jms=230/ 29 Jan 2013 *(in Chinese)*
[5] Li Q K, Lv Z P 2008 Fuzzy Adaptive Kaiman Filtering Algorithm Based on the Statistics of *t* Distribution *Acta Geodaetica et Cartographica Sinica* **37**(4) 228-9 *(in Chinese)*
[6] Liu X L, Li Q H 2008 The Discussion about Robust Estimation based on The Equivalent Weight Function constructed by Studentized Residual *Modern Surveying and Mapping* **31**(6) 207-8
[7] Wang Z Y, Zhang H B, Liu Z M 2006 The Novel Method for Outliers' Rejection of the Studentized Residual Error *Chinese Journal of Scientific Instrument* **27**(6) 624-8 *(in Chinese)*
[8] Lin L, Ding W X, Brower D L 2012 Collinearity alignment of probe beams in a laser-based Faraday effect diagnostic *Review of Scientific Instruments* **83**(10) 320-3

**Author**

**Linlin Dai, born in August, 1979, Liaocheng District, Shandong Province, P.R. China**

**Current position, grades:** lecturer at the School of finance and economics Qingdao Huanghai college, China.
**University studies:** Master Degree in education at Guizhou Normal University in China. Bachelor Degree in science at Liaocheng University in China.
**Scientific interests:** Economics statistics.
**Publications:** 6 papers.
**Experience:** teaching experience of 8 years, 2 scientific research projects.