Trends in Social Sciences and Humanities Research
ISSN: 2959-9938
DOI: 10.61784/tsshr3012

ANALYSIS OF FACTORS INFLUENCING THE ENGEL INDEX BASED ON REGRESSION MODELS

GaoBo Peng
School of Mathematics and Statistics, Guangxi Normal University, Guilin 541006, Guangxi, China.
Corresponding Email: 1062041360@qq.com

Abstract: This paper explores the macroeconomic indicators affecting the Engel Index. Initially, a multicollinearity test reveals the presence of multicollinearity among the independent variables. To eliminate multicollinearity, ridge regression modeling is employed. The analysis identifies six independent variables, with only one having an insignificant regression coefficient. The results show that labor force, urbanization rate, and trade balance are negatively correlated with the Engel Index, while the added value of the primary industry and the gross national income index are positively correlated with the Engel Index.

Keywords: Engel Index; Multicollinearity; Ridge regression; Principal component regression

1 INTRODUCTION

The Engel coefficient is an economic indicator introduced by German economist and statistician Ernst Engel in the 19th century. It is used to measure the standard of living and level of wealth of a household or country. The Engel coefficient is calculated as the ratio of food expenditure to total consumption expenditure. It reflects the proportion of total expenditure that a household or country spends on food. According to Engel's findings, when a household or country has lower income, the proportion of food expenditure is higher; conversely, as income increases, the proportion of food expenditure decreases. Therefore, a higher Engel coefficient indicates a lower standard of living and greater poverty, whereas a lower Engel coefficient indicates a higher standard of living and greater wealth.

The Engel coefficient is widely used not only at the household level but also in national-level analyses. Research by Lancaster et al. shows that the Engel coefficient is widely used due to its simplicity and ability to intuitively reflect living standards [1]. Kaus studied the consumption tendencies of residents in over 50 countries across 12 categories, demonstrating that Engel's law can reveal patterns of consumption [2]. Lewbel's research indicates that the basic meaning of Engel's law is that the rate of increase in food expenditure is slower than the rate of increase in income, showing that the proportion of food expenditure decreases as income increases [3]. Chai et al. found that Engel's law can reveal the relationship between household income and consumption structure [4]. Wang et al. discovered that the Engel coefficient varies among different income groups, indicating a relationship between income levels and the Engel coefficient [5]. Dorothy Brady's research showed that families without children have lower Engel coefficients, and the coefficient increases with the number of children in the family [6]. Angus Deaton and Anne Case analyzed the impact of various cultural customs on residents' dietary activities and found that different cultural habits significantly affect the Engel coefficient [7]. Crawford et al. showed that the Engel coefficient is primarily influenced by the number of women in the household; the fewer the women, the higher the Engel coefficient [8]. John found that the household head's gender, age, and family size also affect the Engel coefficient [9].

2 MAIN THEORIES

2.1 Collinearity

Collinearity is an important concept in statistics and regression analysis, especially in multiple regression models. Collinearity refers to the phenomenon where there is a high correlation among the independent variables. This can lead to instability in the regression coefficients, thereby affecting the explanatory power and predictive ability of the model. Collinearity is classified into perfect collinearity and near-perfect collinearity. Perfect collinearity occurs when an independent variable can be exactly represented by other independent variables, making regression analysis infeasible. Near-perfect collinearity occurs when there is an approximate linear relationship among independent variables. Although regression analysis can still be performed, the results may be unstable.

Collinearity primarily affects regression analysis in the following ways: firstly, it causes instability in regression coefficients, leading to large fluctuations in estimates, and even sign reversal. Secondly, it increases the standard errors of the regression coefficients, affecting the significance tests. Lastly, collinearity makes it difficult to distinguish the independent effects of the variables, reducing the explanatory power of the model. Methods to detect collinearity include the variance inflation factor (VIF), eigenvalue decomposition, and the condition number. A VIF value exceeding 10 typically indicates strong collinearity. Eigenvalues close to zero or a condition number greater than 30 also suggest severe collinearity.

Methods to mitigate collinearity include: removing highly correlated independent variables, using principal component analysis (PCA) to transform the variables and adding penalty terms through ridge regression to reduce the regression coefficients.

2.2 Ridge Regression

Arthur E. Hoerl and Robert W. Kennard proposed ridge regression in 1970. It is a biased estimation method that improves upon ordinary least squares (OLS). The basic idea is: for a linear regression model:

$$Y = X\beta + \varepsilon \tag{1}$$

the least squares estimate of the parameters is: $\widehat{\beta} = (X^TX)^{-1}X^TY$ If there is strong multicollinearity among the independent variables, meaning the determinant of $X^TX$is close to zero (nearly singular), the estimation results may be highly biased. When the number of effective equations is less than the number of unknowns, there is no unique solution, leading to infinitely many solutions. Using the least squares method in this scenario results in instability and unreliability. By adding a normal matrixλI, where is the ridge parameter and is the identity matrix, we obtain the ridge regression estimate:

$$\widehat{\beta} = (X^TX + \lambda I)^{-1}X^TY \tag{2}$$

The cost function modifies the residual sum of squares (RSS) by adding a penalty term $\sum_{j=1}^{p} \beta_j^2$ for the coefficients, ensuring that while minimizing the RSS, the coefficients do not become excessively large:

$$J_\beta(\beta) = RSS + \lambda \sum_{j=1}^{p} \beta_j^2 = RSS + \lambda ||\beta||^2 \tag{3}$$

Typically, the results of ridge regression models are slightly lower than those of ordinary regression models, but their significance is much higher. Ridge regression is particularly useful in studies with collinearity problems and large amounts of ill-conditioned data.The fonts, their sizes, and styles can be seen in the table & figure below.

3 DATA SOURCES AND VARIABLE DESCRIPTIONS

3.1 Data Sources

The data used in this paper comes from the National Bureau of Statistics, covering a period of seventeen years from 2004 to 2020. The data spans seven dimensions: Gross National Income Index (last year=100), labor force (in millions), urbanization rate (%), trade balance (in billions of RMB), added value of the primary industry (%), Consumer Price Index (CPI, last year=100), and the Engel Index (%). Among these, the urbanization rate is not direct data but is calculated according to the National Bureau of Statistics' definition: urbanization rate = urban population / total population (both based on the permanent population, not the registered population). The urban population and total population data are sourced from the National Bureau of Statistics.

3.2 Variable Descriptions

This paper selects the Engel Index (%) as the dependent variable, denoted as. The independent variables are labor force (in millions), urbanization rate (%), trade balance (in billions of RMB), added value of the primary industry (%), Consumer Price Index (CPI, last year=100), and Gross National Income Index (last year=100), denoted as,x_2, x_3, x_4, x_5, x_6 respectively.

4 EMPIRICAL ANALYSIS

4.1 Collinearity Analysis

First, the pairs function in R was used to create scatter plots, providing an overall view of the multidimensional data. See Figure 1.

From the scatter plot matrix, it is evident that some variables are correlated. For instance, the dependent variable shows a strong negative correlation with the independent variables and, and a strong positive correlation with. Additionally, and are also strongly positively correlated.

The strong correlations among the independent variables can be confirmed by examining their correlation coefficients, as shown in Table 1.

Table 1 Correlation Coefficients

	x_1	x_2	x_3	x_4	x_5	x_6
x_1	1.0000000	0.8071398	0.6484744	0.8420276	0.1139538	0.6551198
x_2	0.8071398	1.0000000	0.8002923	0.9474073	0.5091849	0.8482647
x_3	0.6484744	0.8002923	1.0000000	0.7670221	0.4033238	0.6397549
x_4	0.8420276	0.9474073	0.7670221	1.0000000	0.3173843	0.7060632
x_5	0.1139538	0.5091849	0.4033238	0.3173843	1.0000000	0.6955236
x_6	0.6551198	0.8482647	0.6397549	0.7060632	0.6955236	1.0000000

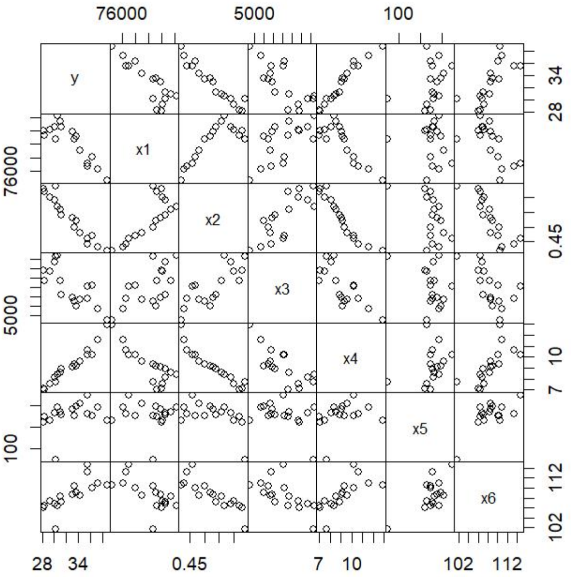

Figure 1 Scatter Plot Matrix

From the table, we see that the correlation coefficients between and,,, are 0.8071398, 0.6484744, 0.8420276, and 0.6551198, respectively, indicating strong positive correlations. The correlation coefficients between and,, are 0.8002923, 0.9474073, and 0.8482647, respectively, indicating very strong positive correlations. The correlation coefficients between $x_3$andx_4, are 0.7670221 and 0.6397549, respectively, indicating strong positive correlations. The correlation coefficient between and is 0.7060632, indicating a strong positive correlation. The correlation coefficient between and x_6 is 0.6955236, indicating a strong positive correlation.

Using the kappa function in R, the condition number of the independent variable matrix was found to be 2610187. Using the vif function in R, the variance inflation factors (VIF) are obtained as shown in Table 2.

Table 2 Variance Inflation Factors

x_1	x_2	x_3	x_4	x_5	x_6
5.407842	31.326773	2.852933	18.563677	3.668128	7.660773

A condition number greater than 100 or a VIF value greater than 10 indicates severe collinearity among the independent variables. In summary, the six factors selected as potential influencers of the Engel Index exhibit severe collinearity.

4.2 Ridge Regression

As required by ridge regression, the data needs to be standardized. To determine the ridge parameter, ridge trace plots need to be created, which can be done using the lm.ridge function and the matplot function in R. See Figure 2 for the ridge trace plots.

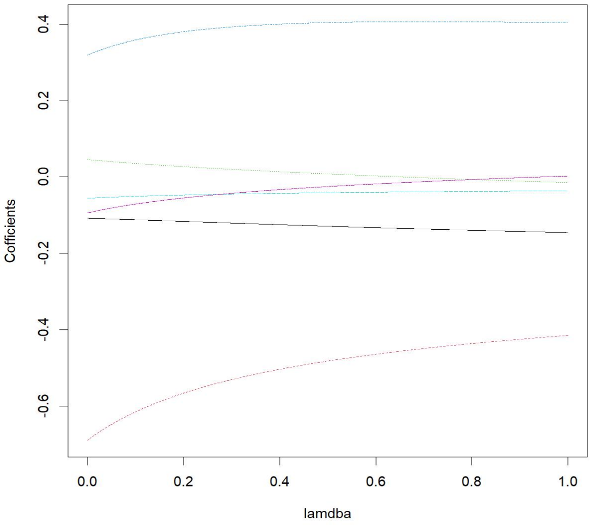

Figure 2 Ridge Trace Plots

From the plots, it can be seen that finding the value of the ridge parameter where all the curves stabilize is difficult, as most curves stabilize very slowly. Thus, it is hard to directly determine an appropriate ridge parameter. However, the select function in R can be used to calculate the value of based on several statistical criteria, or generalized cross-validation can be used to determine the ridge parameter kkk directly. In this section, generalized cross-validation is employed to determine the value of.

The lm.ridge function can then be used to provide the parameter estimates, but the summary function cannot be used to view the results. Therefore, the linearRidge function is used in this section to obtain the parameter estimates. The results are shown in Table 3.

Table 3 Regression Coefficients

(Intercept)	x_1	x_2	x_3	x_4	x_5	x_6
6.5e-16	-6.4e-04	-7.2e-04	-5.7e-04	7.2e-04	2.5e-04	5.5e-04

From the table, it can be seen that labor force (in millions), urbanization rate (%), and trade balance (in billions of RMB) are negatively correlated with the Engel Index (%). In contrast, the added value of the primary industry (%), Consumer Price Index (CPI, last year=100), and Gross National Income Index (last year=100) are positively correlated with the Engel Index (%). The magnitudes of the regression coefficients are also similar.

Next, the summary function in R can be used to output the results of the linear ridge regression model test. The specific results are shown in Table 4.

Table 4 Model Test Results

Coefficient	Estimate	Scaled Estimate	Std. Error (scaled)	t value (scaled)	Pr(>|t|)
(Intercept)	6.54E-16	NA	NA	NA	NA
x1	-0.00065	-0.00258	0.000726	3.559	0.000372
x2	-0.00072	-0.00289	0.000725	3.981	6.85e-05
x3	-0.00057	-0.00228	0.000726	3.143	0.001671
x4	0.00073	0.002919	0.000725	4.023	5.74e-05
x5	0.000251	0.001002	0.000726	1.38	0.167575
x6	0.000555	0.00222	0.000726	3.06	0.002213

From the figure, it can be seen that only the regression coefficient of did not pass the significance t-test, indicating that five variables are significant. Therefore, using ridge regression to analyze the relationship between the independent variables and the dependent variable is appropriate.

4.3 Principal Component Regression

In this section, PCR is used for modeling. Six independent variables must undergo principal component analysis to reduce the data's dimensionality and simplify the model. The princomp function in R is employed for principal component regression modeling, and the summary function is used to obtain the specific results. See table 5 for the detailed results.

Table 5 Model Test

Component	Standard Deviation	Proportion of Variance	Cumulative Proportion
Comp.1	2.0807435	0.7215823	0.7215823
Comp.2	1.0133088	0.1711324	0.8927147
Comp.3	0.60044424	0.06008888	0.95280358
Comp.4	0.41187548	0.02827357	0.98107715
Comp.5	0.30507618	0.01551191	0.99658906
Comp.6	0.143058208	0.003410942	1.00000000

The results show that 72.2% of the data variance can be explained by the first principal component, 17.1% of the variance by the second principal component, and 6.0% by the third principal component. The variance explained by the first principal component is about four times that of the second principal component and approximately twelve times that of the third principal component. The contributions of the second and third principal components decline sharply relative to the first principal component, and the last few principal components explain very little variance. This indicates that most of the variations in the independent variables can be explained by a few dimensions or principal components.

Using the summary function, the linear combinations of the vectors constructing each principal component can also be obtained. See table 6 for the specific results.

Table 6 Composition of Principal Components

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
x1	0.399	0.443	0.375	0.371	0.605	0.0
x2	0.471	0.0	-0.341	-0.203	0.0	0.785
x3	0.408	0.0	-0.835	0.356	0.0	0.0
x4	-0.445	-0.254	0.0	0.637	0.0	0.573
x5	-0.267	0.804	0.0	0.141	-0.510	0.0
x6	-0.427	0.292	-0.399	-0.441	0.574	0.222

The results show that all six independent variables collectively determine the first principal component. However, the first three independent variables are positively correlated with the first principal component, while the last three are negatively correlated. The second principal component is constructed only by the labor force (in millions), added value of the primary industry (%), Consumer Price Index (CPI, last year=100), and Gross National Income Index (last year=100).

Based on the contributions of each principal component, this section uses only the first two principal components for the linear regression model. The lm function in R is used to output the least squares estimates, and the summary function is used to display the results of the multiple linear model test. See table 7 for the specific results.

Table 7 Model Test

|Coefficient	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.157e-16	6.781e-02	0.000	1.0000
z1	-4.364e-01	3.259e-02	-13.391	2.26e-09
z2	-2.253e-01	6.692e-02	-3.367	0.0046
Residual Standard Error	0.2796 on 14 degrees of freedom			
Multiple R-squared	0.9316			
Adjusted R-squared	0.9218			
F-statistic	95.33 on 2 and 14 DF			
p-value	7.009e-09			

It can be seen that the regression coefficients of the first and second principal components pass the significance t-test, indicating that all variables are significant. Additionally, the model's significance test is also significant, as the F-statistic is 95.33, and the corresponding p-value is 7.009e-09, which is less than 0.01. The R-squared and adjusted R-squared values are 0.9316 and 0.9218, respectively, showing that the model can explain about 92% of the data variance. This collectively indicates that using PCR modeling is appropriate. It can also be seen that both the first and second principal components negatively impact the Engel Index. Combining the results of the linear combinations constructing each principal component, it can be concluded that the labor force, urbanization rate, trade balance, and Consumer Price Index are negatively correlated with the Engel Index, while the added value of the primary industry and Gross National Income Index are positively correlated with the Engel Index.

5 RESULTS AND DISCUSSION

Factors such as the labor force, urbanization rate, trade balance, added value of the primary industry, and Gross National Income Index all have significant impacts on the Engel Index. Specifically, the labor force, urbanization rate, and trade balance are negatively correlated with the Engel Index, while the added value of the primary industry and Gross National Income Index are positively correlated with the Engel Index.

From the data in Table 1, it can be observed that the Engel Index in China has been decreasing year by year, indicating an improvement in the living standards of Chinese residents. However, in 2022, the Engel Index increased compared to the previous year, likely due to the impact of the COVID-19 pandemic.

Based on the conclusions drawn from the regression model, to continue improving the living standards of Chinese residents and further reduce the Engel Index, it is necessary to continue increasing the values of the labor force, urbanization rate, and trade balance while decreasing the values of the added value of the primary industry and Gross National Income Index.

In terms of the labor force, the number of workers in China has been decreasing in recent years due to the aging population, which has led the country to gradually relax its family planning policies.

Regarding the urbanization rate, this value has been on an upward trend and should continue to be maintained.

Concerning the trade balance, it has generally been steadily increasing. However, there have been some years of decline since 2016, which may be related to the sanctions imposed by the United States on Chinese enterprises. Therefore, to maintain an increasing trade balance, China needs to develop its core competitiveness.

Regarding the added value of the primary industry, the added value of agriculture is relatively low. An excessive proportion of agriculture can lead to an increase in the Engel Index. However, agriculture is fundamental and its total amount should not be reduced. Therefore, China should vigorously develop the secondary and tertiary industries while ensuring a stable agricultural foundation.

As for the Gross National Income Index, it is influenced by inflation, which is inevitable. Hence, China should aim to keep the inflation rate under control.

COMPETING INTERESTS

The author have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Lancaster G, Ray R, Valenzuela M R. A Cross-Country Study of Equivalence Scales and Expenditure Inequality on Unit Record Household Budget Data. Review of Income & Wealth, 1999, 45(4): 455-482.

[2] Kaus W. Beyond Engel's law - A cross-country analysis. The Journal of Socio-Economics, 2013.

[3] Benjamin, S., Loeb. The Use of Engel's Laws as a Basis for Predicting Consumer Expenditures. Journal of Marketing, 1955, 20(1): 20-27.

[4] Chai A, Moneta A. Retrospectives: Engel Curves.The journal of economic perspectives, 2010, 24(1): 225-240.

[5] Wang X, Woo WT. The size and distribution of hidden household income in China. Asian Economic Papers. 2011, 10(1): 1-26.

[6] Brady DS, Barber HA. The pattern of food expenditures. The Review of Economics and Statistics. 1948, 30(3): 198-206.

[7] Angus Deaton, Anne Case. Analysis of Household Expenditure. LSMS working paper, 1992(4).

[8] Crawford I, Laisney F, Preston I. Estimation of household demand systems with theoretically compatible Engel curves and unit value specifications. Journal of econometrics. 2003, 114(2): 221-241.

[9] Gibson J. Why Does the Engel Method Work? Food Demand, Economies of Size and Household Survey Methods.Oxford Bulletin of Economics & Statistics, 2002(4): 341-359.