I. J. Mathematical Sciences and Computing, **2023, 4, 10-19**
Published Online on December 8, 2023 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijmsc.2023.04.02

Birth Rate Study of Henan Province Based on Ridge Regression Model

Mengke Ye*
School of Mathematics and Information Science, Henan Polytechnic University, Jiaozuo, 454000, China
E-mail: 212110010005@home.hpu.edu.cn
*Corresponding Author

Received: 07 June, 2023; Revised: 01 July, 2023; Accepted: 23 July, 2023; Published: 08 December, 2023

Abstract: In order to explore the underlying reasons for the decline in birth rate, this article selects 12 explanatory variables and uses ridge regression method to study the birth rate in Henan Province from 2015 to 2021. Research has shown that four factors, namely the average salary of urban unit employees, the urbanization degree, the ratio of female employees with a university degree or above, and the population mortality rate, can not explain the birth rate. However, the proportion of gross domestic product of the second and third industries, as well as the proportion of female population over 15 years of age who are illiterate, has a positive impact on the car success rate. The gross domestic product per capita, the number of beds per 10000 people in medical institutions, the per capita disposable income of urban residents, the per capita disposable income of rural residents, the adolescent dependency ratio, and the elderly dependency ratio have a negative impact on the birth rate. Through the research in this article, the main factors affecting the birth rate in Henan Province have been identified, and policy recommendations for improving the birth rate have been proposed. The positive impact represents increasing investment in these factors, which can effectively improve the birth rate in Henan Province and solve the serious problems we are currently facing. The negative factor is the opposite.

Index Terms: Birth rate, Influencing factors, Multicollinearity, Ridge regression, Ridge trace map.

1. Introduction

With the rapid development of the economy, people’s living standards are constantly improving. The aging population has become a serious problem faced by more and more countries. At the same time, we have found that human lifespan is constantly increasing, but the birth rate is decreasing. The implementation of China’s “comprehensive two child” policy at the end of October 2015 has brought about changes in the birth rate due to changes in the birth policy. On May 31, 2021, China implemented the three child policy.

The birth rate has become a serious problem we face. There are many methods for studying birth rates. Some scholars set up Poisson regression model to study the influencing factors of birth rate. Some scholars use principal component regression, while others improve and innovate new methods based on existing ones. When selecting influencing factors, some overlook the economic impact. Scholars have conducted a lot of research on the birth rate.

Yiqun Zhou and Guojun Wang [1] found in the empirical study of the impact of birth rate on residents’ savings that the birth rate has a positive relationship with the savings rate. And They further found that the impact of the birth rate on the savings rate will vary according to different regions through regional classification.

Zhigang Guo and Xiwei Wu [2] used poisson regression to study the fertility rate. First, they used poisson regression to estimate the fertility rate through actual data analysis. The second is to show the flexibility of Poisson regression in dealing with fertility issues by introducing more explanatory variables into poisson regression, which provides a new method for social researchers.

Xizhe Peng [3] analyzed the main reasons for differential fertility rates based on a generalized linear model. And they proposed that population policies should be included in comprehensive social development strategies, rather than just as an independent policy related solely to the family planning sector.

Mian Tan [4] divided the factors affecting fertility into economic and social factors and demography factors. Taking Hunan Province as the research object, he applied structural equation model to analyze the factors affecting fertility. The results show that economic and social factors and demography factors have the same impact on fertility.

Zhenwu Zhai and Shujing Li [5] studied the influencing factors of low fertility rate in China in the new era. They propose that reducing childcare costs, formulating and implementing fertility support policies, and creating a favorable social environment can help them achieve true fertility intentions Furthermore, it effectively solves the difficulties encountered in childbirth and increases the fertility rate.

Shuangyue Lan [6] studied the impact of retirement on the fertility rate of offspring families based on CFPS data in her graduation thesis. The research results show that parental retirement has a significant promoting effect on the fertility of offspring families.

Wei Chen [7] published an article Methods of Fertility Rate Studies in China: An Overview of the Past Thirty Years. This article summarizes the methods used by scholars to study fertility rates. A widely used model in analyzing the influencing factors of fertility is the Bongaarts fertility intermediate variable model, which expresses the changes in

total fertility rate as a function of these influencing factors and the sum and natural fertility rate (generally 13.5-17.0).

Liangzhen Guo[8] used ridge regression and BP neural network to study the influencing factors and forecast data of the birth rate in his graduation thesis“Research on Influencing Factors and Trend Forecast of birth rate”. He found that six factors such as GDP in social economy and agriculture, disposable income of urban residents, and disposable income of rural residents played a certain role in promoting the birth rate in Hubei Province, and gave reasonable suggestions.

There are relatively few existing articles that use ridge regression methods to study the issue of birth rate. Ridge regression analysis is a typical analysis method when there is Multicollinearity in the data. Therefore, this article takes the birth rate in Henan Province as an example and uses ridge regression analysis to study its influencing factors. Through the research in this article, a ridge regression model is established to identify the factors that have a significant impact on the birth rate in Henan Province.And some conclusions are drawn. We hope that the policies proposed based on these conclusions can have a certain promoting effect on the birth rate in Henan Province.

2. Indication Selection and Data Sources

2.1. Indicator selection

Birth Rate(BR): The BR refers to the average number of babies born per thousand people in a certain region or country over a certain period of time. Usually calculated as the ratio of the annual birth population to the total population. The BR is one of the important indicators to measure the population growth rate of a region or country. The BR is influenced by various factors, including economic, cultural, and social policies.

Gross Domestic Product Per Capita(GDPPC): The GDPPC refers to the total value produced by a country or region within a certain period of time divided by the total population of the country or region, which is the average economic value created by each person. The GDPPC is one of the important indicators to measure the economic development level of a country or region. And it is usually used to compare the economic strength and living standards of different countries or regions. The higher the GDPPC, the more developed the economy of the country or region, and the higher the living standards of the people.

Average salary of urban unit employees(ASUE): The ASUE refers to the average obtained by dividing the total wages of all employees in urban units by the total number of employees. This indicator can reflect the income level of urban workers and the supply and demand situation of the job market. Generally speaking, the higher the PDIUR, the higher the income level of urban employees, and the tense supply-demand relationship in the employment market. This indicator is also one of the important indicators for measuring the economic development level of a country or region.

The Proportion of Gross Domestic Product of the Second and Third Industries(PST): The secondary industry refers to industries, including mining, manufacturing, electricity, gas, and water production and supply. The tertiary industry refers to the service industry, including transportation, postal service, communication, wholesale and retail trade, accommodation and catering industry, finance, real estate, leasing, and business services. Gross domestic product (GDP) refers to the total value of all final products and services produced by a country or region within a certain period of time. The PST refers to the proportion of the second and third industries in the gross domestic product of a country or region. This proportion reflects the economic structure and development direction of a country or region. The primary industry in low-income countries accounts for a large proportion of GDP, while the secondary and tertiary industries account for a small proportion of GDP; Secondly, the proportion of the primary industry in the gross domestic product of lower middle and upper middle income countries is very small, while the proportion of the secondary and tertiary industries in the gross domestic product is large; Thirdly, the proportion of the primary industry in the gross domestic product of high-income countries is smaller, the secondary industry is smaller compared to middle-income countries, and the proportion of the tertiary industry in the gross domestic product reaches a higher level. Regions with relatively developed economies will also have a relatively large proportion of the secondary and tertiary industries.

The Number of beds per 10000 people in medical institutions(NBMI): The NBMI refers to the number of medical institution beds per 10000 people in a specific region or country. This indicator can be used to measure whether a region or country has sufficient medical resources, as well as the coverage and quality of its medical services.

Urbanization degree(UD): The UD in a region refers to the degree of urbanization, usually expressed as the percentage of urban population and urban population to the total population. The higher the percentage, the higher UD. The scale effect formed by population aggregation will promote the growth of local employment situation and economic development, and promote the improvement of residents’ income level. Furthermore, it can promote the growth of regional total output value, which may affect the birth rate of the region.

Per Capita Disposable Income of Urban Residents(IUR): The IUR refers to the income earned by urban residents from various economic activities over a certain period of time (usually one year). The total amount of income used for personal consumption and savings after deducting taxes and fees. Then, based on the total population of urban residents, the average income level of each urban resident is calculated by averaging the distribution. This indicator can reflect the economic living standards and consumption capacity of urban residents. And it is also one of the important indicators to measure the level of urban economic development and social-economic status.

Per Capita Disposable Income of Rural Residents(IRR): The IRR refers to the total amount of income that rural residents receive from various economic activities within a certain period of time (usually one year), after deducting necessary living expenses, can be used for disposal, which is the average disposable income of each rural resident. The IRR reflects their living standards, consumption capacity, and economic development level.

Adolescent Dependency Ratio(ADR): The ADR refers to the ratio of the underage population (usually 0-14 years old) to the adult population (usually 15 years old and above) within a specific age group. This ratio can reflect the burden of child rearing in a country or region. If the upbringing of young children is relatively high, it means that more minors need to rely on adults for care and upbringing, which may have social and economic impacts, indirectly affecting the birth rate.

Elderly dependency ratio(EDR): The EDR refers to the ratio of the population aged 65 and above to the population aged 15-64, used to measure the degree of aging and economic burden of a country or region. With the intensification of population aging, the dependency ratio of the elderly population will also increase, which means that more social resources are needed to support the lives and healthcare of the elderly, while also increasing the pressure on the labor market.

Ratio of Female Employees with a University Degree or above(RFUD): The RFUD refers to the proportion of female employees with a university or higher education background. This ratio can reflect women’s participation and quality level in higher education, as well as their competitiveness and job opportunities in the workplace.

Population mortality rate(PMR): Population mortality rate refers to the ratio of the number of deaths to the total population in a certain region or country over a certain period of time. Usually expressed in units of every thousand people. Population mortality rate is one of the important indicators for measuring the health status of a region or country’s population, reflecting the impact of factors such as medical level, hygiene conditions, nutritional status, and lifestyle on the population’s health. The lower the population mortality rate, the better the health status and higher living standards of the population in the region or country.

The Proportion of Female Population over 15 years of age who are illiterate(PFP): The PFP refers to the proportion of female population aged 15 and above who are unable to read and write Chinese. This ratio is usually used to measure the education level and cultural literacy of a country or region. If this proportion is high, it indicates that the education level and cultural literacy in the region are relatively low. This may also affect women’s employment opportunities, social status, and quality of life.

2.2 Data Sources

This paper studied the birth rate of Henan province from 2015 to 2021. The index data comes from China Statistical Yearbook, Henan Statistical Yearbook and National Data.

Table 1. Response variable and explanatory variables.

Indicators classification	**Indicators lable**	**Indicators symbol**	**Indicators unit**
Response variable	y	BR	‰
Explanatory variable	x_1	GDPPC	Ten thousand yuan
	x_2	ASUE	Ten thousand yuan
	x_3	PST	%
	x_4	NBMI	a
	x_5	UD	%
	x_6	IUR	Ten thousand yuan
	x_7	IRR	Ten thousand yuan
	x_8	ADR	%
	x_9	EDR	%
	x_{10}	RFUD	%
	x_{11}	PMR	‰
	x_{12}	PFP	%

3. Ridge Regression Mode

3.1 Introduction To Ridge Regression Estimation

The matrix form of a multiple linear regression model is $Y = X\beta + \varepsilon$. The parameter β is estimated to be by $\hat{\beta} = \left(X^{'}X\right)^{-1} XY$. For the multicollinearity problem, the effect of ordinary least squares becomes worse.A.E. Hoerl first proposed an improved least squares estimation method called ridge estimation in 1962. Later, Hall and Kennard gave a detailed discussion in 1970 [9]. The idea proposed by Ling Hui is very natural. When there is multicollinearity between independent variables,that is When there is multicollinearity between independent variables. That is, $\left|X^{'}X\right| \approx 0$, suppose to add a normal number matrix $kI\left(k > 0\right)$ to $X^{'}X$. Then $X^{'}X + kI$ will approach singularity much less than $X^{'}X$. Considering the dimensionality of variables, first standardize the data so that each column of X after standardization has a mean of 0 and a sum of squares of 1, which is $S_{(ii)} = 1, i = 1, 2, \cdots, p$. The standardized explanatory variable matrix and response variable matrix are represented by X^{*} and Y^{*}, respectively.

Defined as

$$\hat{\beta}(k) = \left(\left(X^{*}\right)^{'} X^{*}\right)^{-1} X^{*}Y^{*} \tag{1}$$

which is called ridge regression estimation, where k is the ridge regression parameter[10].

3.2 Ridge Track Analysis

When the ridge parameter k changes within $(0,\ \infty)$, $\hat{\beta}_j(k)$ is a function of k. The curve drawn by plotting function $\hat{\beta}_j(k)$ in a planar coordinate system is called a ridge trace. In practical applications, appropriate k values are determined and independent variables are selected based on the changing shape of ridge traces.

In ridge regression, ridge trace analysis can be used to understand the role of independent variables and the interrelationships between independent variables.

(1) $\hat{\beta}_{(j)}(0) = \hat{\beta}_{(j)} > 0$, and it's quite large. From the perspective of classical regression analysis, x_j should be regarded as a factor that has a significant impact on y. But $\hat{\beta}_{(j)}(0)$ shows considerable instability. And when k increases slightly from zero, $\hat{\beta}_{(j)}(k)$ significantly decreases and quickly approaches zero, thus losing its predictive ability. From the perspective of ridge regression, it is not important to exclude this variable.

(2) $\hat{\beta}_{(j)}(0) = \hat{\beta}_{(j)} > 0$, but it's very close to 0. From the perspective of classical analysis, x_j has little effect on y. But as k slightly increases, $\hat{\beta}_{(j)}(k)$ suddenly becomes negative. And from the perspective of ridge regression, x_j has a significant impact on y.

(3) $\hat{\beta}_{(j)}(0) = \hat{\beta}_{(j)} > 0$, it means that x_j is relatively significant. But as k increases, $\hat{\beta}_{(j)}(k)$ rapidly decreases and stabilizes at a negative value. From a classical analysis perspective, x_j is a significant factor that has a positive impact on y. From the perspective of ridge regression, x_j is the factor that has a negative impact on y.

(4) If two ridge trace curves, $\hat{\beta}_{(m)}(k)$ and $\hat{\beta}_{(n)}(k)$ are both very unstable. But their sum is generally stable. This situation shows that these two factors have a strong correlation, that is, X_m and X_n have multicollinearity. From the perspective of variable selection, only one of the two needs to be retained.

3.3 Selection of Ridge Parameter

3.3.1 Ridge Trace Method

The intuitive consideration of the ridge trace method is that if the least squares estimation appears unreasonable, such as parameter estimation values and signs that do not conform to economic significance. It is hoped that a certain degree of improvement can be achieved by using an appropriate ridge estimation $\hat{\beta}(k)$. And the selection of the ridge parameter k value is particularly important. The general principle for selecting k value is:

(1) The ridge estimation of each regression coefficient is basically stable.

(2) If the sign of the regression coefficient obtained by least squares estimation is unreasonable, the sign of the ridge estimation becomes reasonable.

(3) The regression coefficient has no absolute value that is not economically meaningful.

(4) The sum of squared residuals does not increase much.

3.3.2 Variance Expansion Factor Method

The variance inflation factor can measure the severity of multicollinearity. Generally, when $c_{(jj)} > 10$, the model has serious multicollinearity. Calculate the covariance matrix of ridge estimation $\hat{\beta}(k)$, and obtain

$$\begin{aligned} D\left(\hat{\beta}(k)\right) &= \operatorname{cov}\left(\hat{\beta}(k), \hat{\beta}(k)\right) \\ &= \operatorname{cov}\left(\left(X^{'}X + kI\right)^{-1} X^{'}y, \left(X^{'}X + kI\right)^{-1} X^{'}y\right) \\ &= \left(X^{'}X + kI\right)^{-1} X^{'} \operatorname{cov}(y, y) X \left(X^{'}X + kI\right)^{-1} \\ &= \sigma^2 \left(X^{'}X + kI\right)^{-1} X^{'}X \left(X^{'}X + kI\right)^{-1} \\ &= \sigma^2 c(k) \end{aligned} \tag{2}$$

Where $c(k) = \left(X^{'}X + kI\right)^{-1} X^{'}X \left(X^{'}X + kI\right)^{-1}$, its diagonal element $c_{jj}(k)$ is the variance inflation factor of ridge estimation. It is not difficult to see that $c_{jj}(k)$ decreases as k increases. The empirical way to select k with variance inflation factor is to select k to make all variance inflation factor $c_{jj}(k) \le 10$. When $c_{jj}(k) \le 10$, the ridge estimate $\hat{\beta}(k)$ of the corresponding k value will be relatively stable[11].

3.3.3 Determining ridge parameter by sum of squares of residuals

Ridge estimation $\hat{\beta}(k)$ reduces mean square error while increasing the sum of squared residuals[12]. We hope to control the increase in the sum of squared residuals $SSE(k)$ of ridge regression within a certain limit, so that a c value greater than 1 can be given, requiring

$$SSE(k) < cSSE \tag{3}$$

find the maximum value that holds the formula[13].

3.4 Selecting variables with ridge regression

The general principle for selecting variables is:

(1) In ridge regression calculation, it is assumed that the design matrix has been centralized and standardized, so that the size of the standardized ridge regression coefficients can be directly compared. We can exclude independent variables with stable and small absolute ridge regression coefficients[14].

(2) When the values are relatively small, the absolute values of the standardized ridge regression coefficients are not very small, but they are unstable and quickly reach zero with increasing values. Independent variables with unstable ridge regression coefficients and zero vibration like this can be considered for elimination.

(3) Exclude independent variables with very unstable regression coefficients for the standardized ridges. If several ridge regression coefficients are unstable, there is no general principle to exclude several variables or which ones. It is determined by the effect of ridge regression analysis after excluding an independent variable[15].

4. Result

4.1 Model results

Due to the incomplete birth rate data in Henan Province and considering the authenticity of the results, the author studied the birth rate situation in Henan Province from 2015 to 2021. By reviewing literature and studying the actual situation of birth rate, we selected 12 indicators that may have an impact on the birth rate. Perform ridge regression analysis on the birth rate using the R(ver4.2.2), where the calculation results of the ridge parameter k and its corresponding regression coefficients are shown in the Table 2. The red, orange, yellow, green, blue, purple, black, grey, cyan, magenta, pink and brown curves represent the ridge traces of the explanatory variables x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11} and x_{12} in Fig.1, respectively. And the red, yellow, green, purple, black, grey, cyan and brown curves represent the ridge traces of the explanatory variables x_1, x_3, x_4, x_6, x_7, x_8, x_9 and x_{12} in Fig. 2, respectively.

Table 2. The variation of the coefficients of each explanatory variable with k from 0 to 2.

k	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
0.0	-0.1974	-0.6275	0.2319	-0.2947	0.4460	-0.0754	-0.0018	-0.3737	0.5393	-0.0346	-0.1861	0.6029
0.1	-0.2019	-0.2367	0.2124	-0.2071	0.1325	-0.0960	-0.0539	-0.1910	0.1477	-0.0345	-0.1441	0.4249
0.2	-0.1565	-0.1795	0.2011	-0.1696	0.0517	-0.0953	-0.0797	-0.1669	0.0351	-0.0406	-0.1394	0.3491
0.3	-0.1346	-0.1530	0.1942	-0.1520	0.0123	-0.0944	-0.0918	-0.1567	-0.0175	-0.0431	-0.1355	0.3104
0.4	-0.1218	-0.1374	0.1888	-0.1417	-0.0109	-0.0937	-0.0986	-0.1509	-0.0477	-0.0442	-0.1317	0.2866
0.5	-0.1133	-0.1273	0.1843	-0.1349	-0.0261	-0.0932	-0.1029	-0.1472	-0.0671	-0.0448	-0.1281	0.2703
0.6	-0.1074	-0.1202	0.1802	-0.1299	-0.0369	-0.0927	-0.1058	-0.1445	-0.0806	-0.0451	-0.1246	0.2582
0.7	-0.1030	-0.1149	0.1764	-0.1262	-0.0449	-0.0923	-0.1077	-0.1425	-0.0904	-0.0452	-0.1212	0.2488
0.8	-0.0997	-0.1109	0.1728	-0.1232	-0.0511	-0.0920	-0.1091	-0.1408	-0.0977	-0.0453	-0.1180	0.2412
0.9	-0.0970	-0.1076	0.1695	-0.1208	-0.0560	-0.0917	-0.1101	-0.1395	-0.1034	-0.0453	-0.1149	0.2349
1.01.	-0.0948	-0.1050	0.1663	-0.1188	-0.0599	-0.0914	-0.1108	-0.1383	-0.1078	-0.0453	-0.1119	0.2295
1	-0.0930	-0.1029	0.1633	-0.1170	-0.0632	-0.0912	-0.1113	-0.1373	-0.1114	-0.0453	-0.1090	0.2248
1.2	-0.0915	-0.1011	0.1604	-0.1155	-0.0659	-0.0909	-0.1116	-0.1364	-0.1142	-0.0452	-0.1062	0.2206
1.3	-0.0902	-0.0995	0.1577	-0.1142	-0.0683	-0.0907	-0.1118	-0.1355	-0.1166	-0.0452	-0.1035	0.2168
1.4	-0.0891	-0.0982	0.1550	-0.1130	-0.0703	-0.0905	-0.1119	-0.1347	-0.1185	-0.0452	-0.1009	0.2134
1.5	-0.0882	-0.0971	0.1524	-0.1119	-0.0720	-0.0903	-0.1120	-0.1340	-0.1201	-0.0452	-0.0983	0.2103
1.6	-0.0874	-0.0960	0.1499	-0.1109	-0.0735	-0.0901	-0.1120	-0.1334	-0.1214	-0.0453	-0.0959	0.2075
1.7	-0.0866	-0.0951	0.1475	-0.1100	-0.0747	-0.0899	-0.1119	-0.1327	-0.1225	-0.0453	-0.0936	0.2048
1.8	-0.0860	-0.0943	0.1452	-0.1092	-0.0759	-0.0897	-0.1118	-0.1321	-0.1234	-0.0453	-0.0913	0.2024
1.9	-0.0854	-0.0936	0.1430	-0.1084	-0.0769	-0.0895	-0.1117	-0.1315	-0.1241	-0.0454	-0.0891	0.2000
2.0	-0.0848	-0.0929	0.1408	-0.1077	-0.0778	-0.0894	-0.1115	-0.1310	-0.1247	-0.0454	-0.0869	0.1979

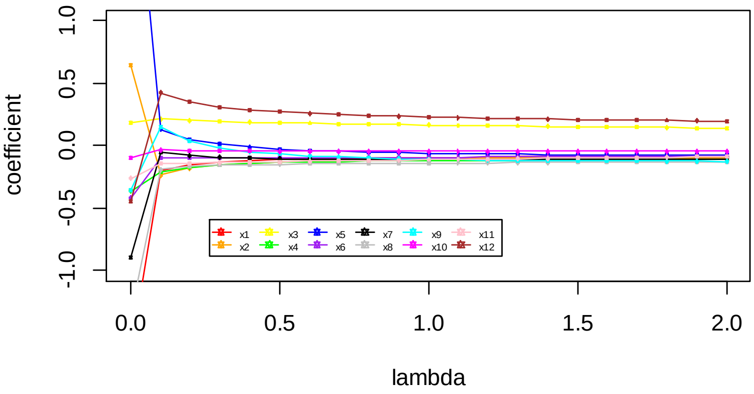

Fig.1. Ridge coefficient trace for the standardized birth rate data.

Using ridge trace method analysis, the first column in Table 2 shows the ridge parameter k, which ranges from 0 to 2 and has a step size of 0.1. There are a total of 21 k values. The 2nd to 13th columns are the normalized ridge regression coefficients of the data, with the first row $k = 0$, which is the normalized regression coefficient of the ordinary least squares estimation. From Fig. 1, it can be seen that the curve shapes of x_1 and x_5 are quite similar. According to ridge track analysis , the variable x_1 that may have a greater impact on the birth rate is left behind, and the variable x_5 is removed. x_2 and x_{12} are similar to the above situation, with similar curve shapes and a relatively stable sum. Excluding x_2, variables x_{10} and x_{11} have relatively stable and small absolute ridge regression coefficients. According to the principle of selecting variables in the ridge plot, these independent variables can be deleted. At the same time, it can be observed that when the ridge parameter is between 0.3 and 0.5, stability has been achieved.

Through the above analysis, we have decided to exclude x_2, x_5, x_{10}, x_{11}. The remaining variables are x_1, x_3, x_4, x_6, x_7, x_8, x_9, x_{12}, and use y to establish a regression equation with the remaining 8 independent variables. Reduce the range of ridge parameter values to 0 to 0.5, with a step size of 0.05, and use the R(ver.4.2.2) to calculate and output the results as shown in Table 3 and Fig. 2.

Table 3. The variation of the coefficients of each explanatory variable with k from 0 to 2.

k	x_1	x_3	x_4	x_6	x_7	x_8	x_9	x_{12}
0.00	-1.2496	0.3660	-1.2336	1.3624	0.1554	0.1184	0.4172	0.7412
0.05	-0.2233	0.2811	-0.5552	-0.0505	-0.0661	0.1070	0.2708	0.5959
0.10	-0.2078	0.2788	-0.4183	-0.1092	-0.1264	0.0968	0.1341	0.4730
0.15	-0.1933	0.2701	-0.3573	-0.1293	-0.1515	0.0822	0.0623	0.4098
0.20	-0.1822	0.2605	-0.3211	-0.1380	-0.1638	0.0677	0.0178	0.3712
0.25	-0.1737	0.2511	-0.2964	-0.1421	-0.1703	0.0542	-0.0126	0.3447
0.30	-0.1670	0.2423	-0.2781	-0.1441	-0.1737	0.0417	-0.0346	0.3254
0.35	-0.1616	0.2341	-0.2638	-0.1448	-0.1755	0.0304	-0.0513	0.3106
0.40	-0.1571	0.2266	-0.2521	-0.1450	-0.1762	0.0201	-0.0643	0.2989
0.45	-0.1533	0.2197	-0.2424	-0.1447	-0.1763	0.0107	-0.0748	0.2893
0.50	-0.1500	0.2133	-0.2341	-0.1443	-0.1760	0.0021	-0.0834	0.2813

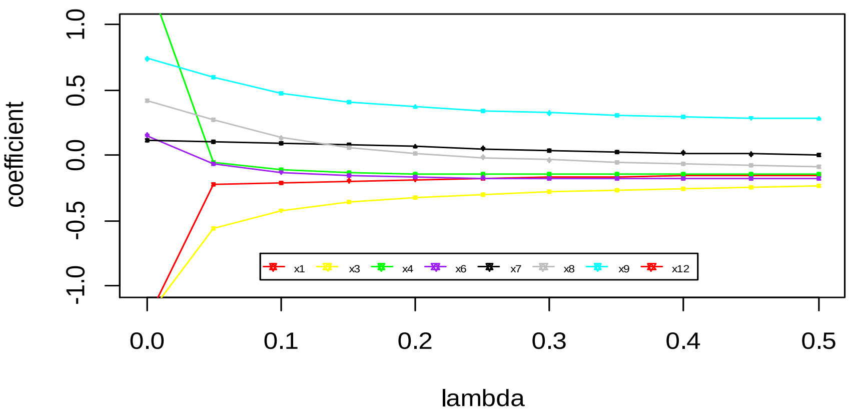

Fig.2. Ridge coefficient trace for the standardized birth rate data (after removing some explanatory variables).

From Fig. 2, it can be seen that after excluding x_6, x_7, x_{10}, and x_{11}, the range of change in the ridge regression coefficient decreases and the overall performance is stable. From Fig. 2 of the ridge plot, it can be seen that when the ridge parameter $k > 0.3$, the value of the ridge parameter is basically stable. Using the ridge package in R(ver.4.2.2), call the function LinearRidge to perform ridge regression analysis when k is within the range of 0 to 0.5. By observing the results obtained by R(ver.4.2.2), we can find that at $k = 0.35$, all explanatory variables are at a significance level of 10. Through hypothesis testing, as shown in Table 4.

Table 4. Parameter estimation of Ridge regression model

Variable	Parameter estimation	Estimation reeor	Error	T-values	$\Pr > \|t\|$
Intercept	-6.225e-16	NA	NA	NA	NA
x_1	-1.177e-01	-2.883e-01	4.803e-02	6.003	1.94e-09 ***
x_3	1.134e-01	2.779e-01	1.294e-01	2.148	0.0317 *
x_4	-1.487e-01	-3.643e-01	6.038e-02	6.033	1.61e-09 ***
x_6	-1.260e-01	-3.087e-01	4.709e-02	6.555	5.57e-11 ***
x_7	-1.499e-01	-3.671e-01	5.182e-02	7.083	1.41e-12 ***
x_8	-1.069e-01	-2.618e-01	1.219e-01	2.148	0.0317 *
x_9	-1.440e-01	-3.528e-01	6.346e-02	5.560	2.70e-08 ***
x_{12}	2.019e-01	4.947e-01	6.759e-02	7.318	2.51e-13 ***
Ridge parameter $k = 0.35$					

Note: ‘‘*’’ represents a significance level of 10%, ‘‘* *’’ represents a significance level of 5%, and ‘‘* * *’’ represents a significance level of 1%.

Write the standardized ridge regression equation for the sample data as follows:

$$\begin{aligned}\hat{y}^{*} &= -0.1177x_1^{*} + 0.1134x_3^{*} - 0.1487x_4^{*} - 0.126x_6^{*} \\ &\quad - 0.1499x_7^{*} - 0.1069x_8^{*} - 0.1444x_9^{*} + 0.2019x_{12}^{*}\end{aligned}$$

The corresponding non standardized ridge regression equation at this time is:

$$\begin{aligned}\hat{y} &= 3.536192 - 0.31249x_1 + 0.284111x_3 - 0.0394x_4 - 0.59734x_6 \\ &\quad - 1.2361x_7 - 0.08547x_8 - 0.09358x_9 + 0.262948x_{12}\end{aligned}$$

According to the results of ridge regression analysis, the explanatory variables x_2, x_5, x_{10}, and x_{11} have no significant impact on the birth rate in Henan Province, while x_3 and x_{12} have a positive correlation with the birth rate. x_1, x_4, x_6, x_7, x_8, x_9 are significantly negatively correlated with birth rate.

Based on the output of the ridge regression model, the following conclusions can also be drawn:

(1) The larger the PST represents the more developed economy, we can think that the improvement of economic level can stimulate the fertility desire of most families in Henan Province, have enough capital to raise children, and thus increase the birth rate.

(2) The higher the PFP, the fewer educational opportunities for women, the lower the possibility of contacting fresh things from outside, and the easier it is to satisfy the current situation. Therefore, it is very important to improve women’s educational level and cultural accomplishment. The higher the education level of women, the higher their expectations for their own occupation, the greater the conflict between reproduction and employment, and the higher the cost of childbearing. Therefore, women with higher education tend to have fewer children. Compared with highly educated women, illiterate women are more likely to be constrained by their current circumstances and have no more opportunities to enter the workplace to realize their own value. The higher the illiteracy rate, the higher the birth rate.

(3) The ADR and the EDR represent, respectively, how many children and the elderly should bear per 100 working-age people. The larger these two indicators are, the greater the pressure that contemporary young people are under. In order to reduce your own stress, there will be the idea of having fewer children.

(4) The NBMI represents the development of the medical industry. As the economy grows, the NBMI increases. Higher the NBMI usually means better medical services and wider medical coverage, while lower indicators may mean shortage of medical resources and inadequate medical services. The abundance of medical resources can not be separated from the efforts of the broad masses of people, which reflects to a certain extent that people are busy with their work and have less time to have children. At the same time, we find that with the increase of the IUR, the IRR and the GDPPC, the birth rate is decreasing. Regarding the growth of these favorable factors, the reasons why the birth rate is still decreasing. We can find that in recent years, the concept of childbearing has changed, the age of marriage and childbearing has been postponed, and the desire for childbearing of young people has been reduced due to other factors that increase the cost of childbearing. The outbreak of new coronary pneumonia has delayed marriage and childbearing arrangements to some extent. Several international studies have found that fertility levels have decreased in many countries and regions since the epidemic. In 2020, Japan’s birth population decreased from the previous year, and Sout Korea’s birth population decreased from the previous year. China’s growth is still increasing, that is, the growth is decreasing, and the total amount has not decreased.[16] The concept of “bringing up children and preventing old age” gradually withdraws from the thoughts of the contemporary people and no longer needs children to protect their old age life.

4.2 Research conclusions and policy recommendations

This chapter takes the birth rate of Henan Province as the response variable, and takes economic and population factors as explanatory variables, builds a ridge regression model. Based on the data of Henan Province, it is concluded that there are four factors that can not explain the birth rate: the ASUE, the UD, the RFUD and the PMR. The PST and the PFP have a positive impact on the birth rate. The GDPPC, the NBMI, the IUR, the IRR,the ADR and the EDR have a negative impact on the birth rate.

As the economy improves, we have seen a decrease in birth rates in recent years. This requires policy support, such as discrimination women face in employment, and some businesses may have gender biases. Instead of guaranteeing women’s welfare benefits during pregnancy, childbirth and breastfeeding, they try to make pregnant women leave their jobs in order to protect the rights and interests of enterprises. This is one reason why women are reluctant to have children, and they are more willing to realize their value in the workplace. In order to truly solve this problem, the government should formulate appropriate policies, perfect the law on protection of women’s rights and interests, establish a social security system, and truly realize the legalization and institutionalization of women’s rights and interests.

Over-squeezing of employees in some enterprises can lead to a decline in the quality of life of young people and not too much time to enjoy life. This requires relevant policies to restrict the enterprise, so long working hours should not be allowed. While vigorously developing the economy, we should formulate some“human-friendly” policies to protect the rights and interests of employees. At the same time, the concept of equality between men and women is advocated to improve women’s own quality.

5. Conclusion

The research object of this article is the birth rate in Henan Province. The ridge regression method was used to analyze the correlation and role between various influencing factors and the research object. Establish a ridge regression model to gain a clearer understanding of the relationship between birth rate and influencing factors.

Using ridge regression model to analyze the birth rate of Henan Province, the paper takes the GDPPC, the ASUE, the PST, the NBMI, the UD, the IUR, the IRR, the ADR,the EDR, the RFUD, the PMR and the PFP were used as explanatory variables, and the influencing factors of the birth rate in Henan province from 2015 to 2021 were analyzed. Due to the existence of multicollinearity in the data, we want to leave more influencing factors for analysis and research subjectively, so we choose ridge regression method. Using R(ver4.2.2) to establish a ridge regression model, based on the output results, it was found that the explanatory variables the ASUE, the UD, the RFUD, and the PMUR were still not significant. After removing these variables, the ridge regression model was established again. When the ridge parameter $k = 0.35$, the remaining 8 explanatory variables were all significant, indicating that the ridge regression model was stable at this time. The results of the ridge regression model show that, the PST and the PFP have a positive impact on the birth rate. The GDPPC, the NBMI, the IUR, the IRR, the ADR and the EDR have a negative impact on the birth rate.

The innovation of this article lies in:

(1) A Systematic Study on the Birth Rate Problem within Henan Province.

(2) The use of ridge regression to study the influencing factors of birth rate in Henan Province has rarely been used in previous studies.

Population problem is the basic problem of a country. It is closely related to economic development and people’s social problems. Excessive population may even cause large social problems such as poverty and disease. Therefore, the country must pay enough attention to the population problem.

In this article, we investigated the influencing factors of the birth rate in Henan Province. Through the research in this article, policy recommendations have been proposed to improve the birth rate. I believe that through these suggestions, the birth rate in Henan Province can be effectively improved.

It is worth noting that in this article, we only considered data from Henan Province from 2015 to 2021. And only one method was used to study the birth rate, which makes this article somewhat limited. More years of data can be selected and multiple methods can be used from different perspectives to make the results more realistic in future research.

References

[1] Yiqun Zhou, Guojun Wang. (2016) Empirical analysis on Chinas birth rate and saving rate based on provincial panel data. CHINA POPULATIONESOUCES AND ENVIONMENT, 26 (11), 266-269.

[2] Zhigang Guo,Xiwei Wu. (2006) Applicati on of Poisson Regressi on i n Fertility Study. POPULATION JOURNAL, (4), 2-95.

[3] Xizhe Peng. (1990) Application of generalized linear model in differential fertility analysis. China Academic Journal Electronic Publishing House, (02), 49-52.

[4] Mian Tan. (2011) Analysis of the influencing factors of population fertility rate in Hunan Province. China Academic Journal Electronic Publishing House, 211-212.

[5] Zhenwu Zhai, Shujing Li. (2023) The influencing factors of low fertility rate in China in the new era. JOURNAL OF UNIVERSITY OF JINAN(Social Science Edition), 33 (1),13-24.

[6] Shuangyue Lan. (2022) The effect of retirement on fertility in offspring families: an empirical test based on CFPS data. Southwestern University of Finance and Economics.

[7] Wei Chen. (2009) Methods of Fertility Rate Studies in China: An Overview of the Past Thirty Years. POPULATION JOURNAL, (3), 3-8.

[8] Liangzhen Guo. (2022) Study on the influencing factors and trend forecast of birth rate in Hubei Province. Central China Normal University.

[9] Robert I.Kabacoff. (2016) R in Action. POSTS & TELECOM PRESS.

[10] Bradley Efron, Trevor Hastie. (2017) Computer Age Statistical Inference Algorithms, Evidence, and Data Science. Cambridge University Press.

[11] Xiaoqun He. (2017) Applied Regression Analysis(R Language Edition). PUBLISHING HOUSE OF ELECTRONICS INDUSTRY.

[12] Shisong Mao,Yiming Cheng,Xiaolong Pu. (2011) Probability theory and mathematical statistics (Second Edition). PHigher Education Press.

[13] Shunan Zhang. (2022) The National Bureau of Statistics responded to the "birth rate decline": there are three reasons. https://www.163.com/dy/article/GTTU1HOA0519C6T9.html.

[14] Liying Wan. (2016) Analysis and Application of Ridge Regression. JOURNAL OF XUCHANG UNIVERSITY, 35(2), 19-23.

[15] Xiaogang Dong, Yajing Diao, Hunling Li, Chunjie Wang, Linan Wen. (2018) The analysis of the fiscal revenue factors under the ridge regression,LASSO regression and the Adaptive-LASSO regression. Journal of Jilin Normal University(Natural Science Edition), 39(2), 45-53.

[16] Hailong Zhu, Pingping Li. (2022) Analysis of influencing factors of fiscal revenue in Anhui province based on ridge regression and lasso regression. Journal of Jiangxi University of Science and Technology, 43(1), 59-65.

Authors' Profiles

Mengke Ye is a student for Master degree for applied Mathematics in the School of Mathematics and Information Science at Henan Polytechnic University in China and her research interests are in applied statistics.

How to cite this paper: Mengke Ye, "Birth Rate Study of Henan Province Based on Ridge Regression Model", International Journal of Mathematical Sciences and Computing(IJMSC), Vol.9, No.4, pp. 10-19, 2023. DOI: 10.5815/ijmsc.2023.04.02