ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202402_7(2).0003

Research on Marital Status and Its Influencing Factors Using Logistic Regression Model

Xingzhou Wang

School of science, Minzu University of China, Beijing 100081, China

Abstract

Marriage has a crucial impact on families and individuals. The research on marital status is beneficial in understanding the marriage phenomenon and the change in social structure in China. Based on the data from the Chinese General Social Survey (CGSS) in 2018 and 2021, this paper established logistic regression models, conducted an empirical analysis of marital status and its influencing factors, and explored the impacts of personal characteristics, social factors, and economic factors on the marital status. Besides, this paper also conducted prediction research to a certain extent. The conclusion shows that gender has no significant influence on marital status, the floor area of the respondents' housing has a significant positive impact on whether there is a spouse, and the age of the respondents has a significant negative impact on whether there is a spouse. Respondents who reported high levels of self-reported health were more likely to have a spouse than those who reported very poor health. In addition, in terms of model prediction, the data modeling results of 2018 are more conducive to predicting the marital status of the current year than that of 2021.

Keywords

Chinese General Social Survey (CGSS); Logistic Regression Model; Marital Status; Empirical Analysis.

1. Introduction

In social investigation and research, marital status is a crucial variable, which is closely related to gender, individual income, educational experience, individual happiness, and many other factors. Individuals with spouses often have access to emotional support, financial support, and social acceptance, which has a positive impact on the individual's well-being and quality of life. At the same time, marital status is also closely related to family structure and family function, which affects family stability and children's growth. Through the investigation and research of spouses, we can deeply understand the status and role of marriage and family in society, and provide a scientific basis for formulating relevant policies and providing social support.

Chinese General Social Survey (CGSS), started in 2003, is the earliest national, comprehensive, and continuous academic survey project in China. Through systematic and comprehensive collection of social, community, family, and individual data at multiple levels, it aims to summarize the trend of social change and discuss issues of practical significance. As a multi-disciplinary economic and social data collection platform, CGSS data provides valuable data for international comparative research. The application fields of the database cover scientific research, teaching, and government decision-making. Through data analysis, researchers can gain an in-depth understanding of the development and changing trends in Chinese society, thus providing support for research in the field of social science.

The CGSS database for 2018 and 2021 mainly uses individuals and households as analysis units to conduct survey research in mainland China. The respondents were 18 years old and above, and the survey was organized in a continuous cross-sectional survey with annual survey

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202402_7(2).0003

frequency and interview mode. In addition, for the sampling methods and procedures, multi-stage stratified sampling is adopted.

The logistic regression model is a type of statistical model used to deal with binary classification problems. It converts the output of the linear regression model into probability values based on logical functions and is used to predict the probability of an event occurring. In a logistic regression model, the response variable is binary, such as yes or no, success or failure, etc. Use one or more explanatory variables to predict the probability of a response variable. Logistic regression models are widely used in many fields, such as medicine and social science, to predict and explain the probability of binary events. The model has a simple and effective model structure and can deal with linearly separable and linearly indivisible problems.

This paper conducts empirical analysis based on the CGSS database, studies the marital status of people aged 25 and above and its influencing factors, mainly uses logistic regression models for modeling and analysis, excavates significant influencing factors, and finally makes a certain degree of prediction on whether there is a spouse in certain given circumstances. This paper innovatively combines the research results of 2018 and 2021 to make a comparative analysis, specifically compares the differences in marital status between the two years, and finally gives conclusions and policy recommendations.

2. Literature References

Many scholars have done a lot of analysis and research using the CGSS database, among which some representative research contents and conclusions can be summarized as follows: First, based on CGSS data, Wu Dan [1] discussed the impact and mechanism of Internet use on residents' fertility intention. The study found that Internet use had a significant impact on rural areas' fertility intention, and further analysis found that Internet use could improve residents' fertility intention by increasing household annual income. Based on the 2017 CGSS data, Han Hepeng [2] analyzed the factors affecting the re-employment of the elderly in China through the Logistic model. The results showed that gender, age, household registration, health status, marital status, number of children, education level, and whether to participate in the basic endowment insurance were significant factors affecting the re-employment of the elderly. Wang Chonggrun and Zhao Changyan [3] discussed the impact of housing price on family fertility intention based on CGSS micro-data and found that the rising housing price significantly inhibited the fertility intention of women of childbearing age, and when male samples were added, the inhibiting effect of housing price on the fertility intention of women of childbearing age was weakened. Zhou Wei and Ma Hongru [4] used the 2017 CGSS data to empirically analyze the impact of intergenerational mobility of education on the subjective well-being of children. The results show that the higher the educational level of the parents, the more likely the children are to achieve upward educational mobility, and the subjective well-being of the children of upward educational mobility is significantly higher than that of downward educational mobility. Based on the 2017 CGSS data, Wang Zhuo and Su Beibei [5] used logistic regression to analyze the impact of three dimensions of individual characteristics, family endowment, and macro environment on the employment of Chinese youth. The study found that the youth employment situation is not optimistic on the whole, and the youth unemployment problem is prominent in the young age, low education, and low health level. Based on the 2017 CGSS data, Gao Shuang [6] analyzed the impact of popularity promotion on workers' income in ethnic minority areas from two dimensions: listening and speaking. It is found that in ethnic minority areas, the listening premium of Mandarin is higher than the speaking premium of Mandarin. Pei Xudong and Song Juan [7] analyzed the impact of family care on the labor participation rate of urban married women by constructing the Probit model and pointed out that elderly care and child care would have a negative impact on the labor

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202402_7(2).0003

participation rate of urban married women, while intergenerational care could alleviate the conflict between women's work and family to a certain extent. He Kelli [8] used the 2015 CGSS data and the Minser income function model and its extended model to analyze the factors affecting the return on the education of graduate students. The results showed that individual characteristics, education level, work experience, and gender significantly affect the return on education of graduate students. Liu Jinshan and Du Lin [9] built a theoretical model of marriage decision-making, aiming to study the impact of housing prices on the marriage rate and finally found that the effect of rising housing prices on the delay of women's first marriage is greater than that of men, and that effect on the delay of marriage of people of marriage age in urban areas is greater than that of rural areas. Yin Donghao, Song Jiayu, and Chen Mingyan et al. [10] used the CGSS database to explore the impact of leisure styles on the happiness of rural elderly people in the context of rural revitalization. The results show that pure time-consuming and achievement leisure activities have a significant positive impact on the well-being of the rural elderly, while social leisure activities have a weaker impact.

Based on the above literature, it can be seen that the research based on CGSS data covers many fields, such as the effect of income promotion in ethnic minority areas, the current situation and influencing factors of youth employment, the impact of family care on the labor participation rate of urban married women, the influencing factors of graduate students' personal education returns, and the relationship between happiness and leisure styles of rural elderly people. The research of these scholars provides inspiring ideas for an in-depth understanding of the CGSS database and offers valuable references for decision-making in related fields.

The following structure is arranged as follows: The third section introduces the variables, data sets, and data sources selected in this study; The fourth section mainly describes the details of data preprocessing and descriptive statistical analysis; The fifth section establishes logistic regression models for modeling and analysis based on CGSS data in 2018, specifically exploring the marital status and its influencing factors; In section 6, similar analysis is carried out on CGSS data of 2021, and the differences between the results and those of 2018 are found. Section 7 analyzes and compares the differences in detail, and makes relevant predictions; The eighth section gives the research conclusions of this paper and puts forward policy suggestions.

3. Data and Variable Description

All the data used in this paper are from the Chinese General Social Survey (CGSS) database. In this paper, the CGSS data of 2018 and 2021 were selected respectively for empirical research, among which the original sample size of 2018 data was 12,787, and the original sample size of 2021 data was 8148. After the data preprocessing process, the sample size in 2018 was 11,017, and the sample size in 2021 was 6169. Specific data preprocessing guidelines and process descriptions are described in detail in the next section.

Table 1 shows the definitions and detailed descriptions of the dependent and explanatory variables selected in this paper. Where the dependent variable “married” represents the marital status, where there is a spouse is 1, and other circumstances (including unmarried, divorced, widowed, etc.) are 0. Among the explanatory variables, “gender” is a categorical variable, and “edu” and “health” represent ordered multiple categorical variables. For specific explanations, see Table 1. It is worth noting that “edu” is divided into 13 grades, 1 means no education, 13 means graduate students and above, and the value shows the increase in years of education from small to large, such as 6 means ordinary high school, and 12 represents regular university undergraduates. The total income of the respondents in the last year, the floor space in the apartment, and age are all continuous numerical variables.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202402_7(2).0003

Table 1. Variable definition table

Variable Type	Variable Name	Representative Significance	Variable Declaration
Dependent	married	Respondent's spouse	1 means you have a spouse, 0 means other
Explanatory	gender	Respondent gender	1 means male, 0 means female
	edu	Highest education level of respondents	It's represented by 1 to 13. Educational attainment increases with the number. One means no education at all, and 13 means postgraduate or above
	health	Respondents rated their current physical health	It's represented by 1 to 5. Health conditions tend to improve as the numbers increase. One is very unhealthy, five is very healthy
	income	Total annual income of respondents last year (Yuan)	Continuous Variable
	area	Floor area (m^2) of respondents' housing units	Continuous Variable
	age	Respondent age (years)	Continuous Variable

4. Descriptive Analysis

4.1. Data Preprocessing

In order to carry out more efficient empirical research, this paper conducted data preprocessing on CGSS datasets in 2018 and 2021. The specific pretreatment process is as follows: First, collect all the original data, then delete all null values, outliers, "don't know", and "refuse to answer" equivalents; Next, in order to study people aged 25 and above, "age" should first be controlled to be greater than or equal to 25. Then, according to the actual research needs, extremely large values of the housing area and the total income of last year needed to be eliminated, so "area" is less than or equal to 1000 and "income" is less than 1 million. Next, the unprocessed indicators are coded, for example, the data obtained in 2021 has given 1 for males and 0 for females, but the data of 2018 needs to be processed manually, so similar processing is done according to the situation of 2021, making the standards of the two study years the same. In addition, for the dependent variables, unmarried, cohabitation, first marriage with a spouse, remarriage with a spouse, separation without divorce, divorce, and widowhood, it is necessary to extract all the cases of "have spouses" and code them as 1, and all other cases are 0. Finally, the obtained new data set applied in the analysis is sorted out and loaded into R software for further analysis. To further understand the numerical characteristics of each variable, a descriptive statistical analysis is required.

4.2. Descriptive Statistical Analysis

The specific information of the two data sets is shown in Table 2 and Table 3, which show indicators including mean value, standard deviation, maximum value, and number of observations. It can be seen from the tables that each data set contains 7 variables, of which "married" is the dependent variable, and the remaining variables are the explanatory variables. There are 11017 sample observations after CGSS data processing in 2018 and 6169 sample observations after CGSS data processing in 2021. Additionally, there are no missing values.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202402_7(2).0003

Table 2. Descriptive analysis of variables (2018)

Variable Name	Observed Value	Mean	Standard Deviation	Minimum Value	Maximum Value
married	11,017	0.80	0.40	0	1
gender	11,017	0.47	0.50	0	1
edu	11,017	4.84	3.13	1	13
health	11,017	3.50	1.07	1	5
income	11,017	38464.00	61521.00	0	960000
area	11,017	109.70	83.22	6	1000
age	11,017	53.80	15.43	25	118

Table 3 Descriptive analysis of variables (2021)

Variable Name	Observed Value	Mean	Standard Deviation	Minimum Value	Maximum Value
married	6,169	0.78	0.41	0	1
gender	6,169	0.47	0.50	0	1
edu	6,169	5.12	3.11	1	13
health	6,169	3.44	1.07	1	5
income	6,169	43353.00	70948.00	0	927000
area	6,169	118.30	85.60	7	1000
age	6,169	54.35	15.48	25	95

It can be seen that the maximum income did not exceed one million yuan, and the housing area was controlled at 1000 square meters or less. All samples under the age of 25 were also excluded because it is not of great value to study the marital status of too young group, and it may even interfere with the analysis, which will affect the overall analysis result and prediction effect. Therefore, this paper only researches people aged 25 and above.

Additionally, to explore the relationship between age and the dependent variable "married", a grouping box plot can be drawn to illustrate the situation. Fig. 1 shows the boxplots for the data in 2018 and 2021. The results showed that regardless of the year, the median age of the "other conditions" group was higher than that of the "have a spouse" group.

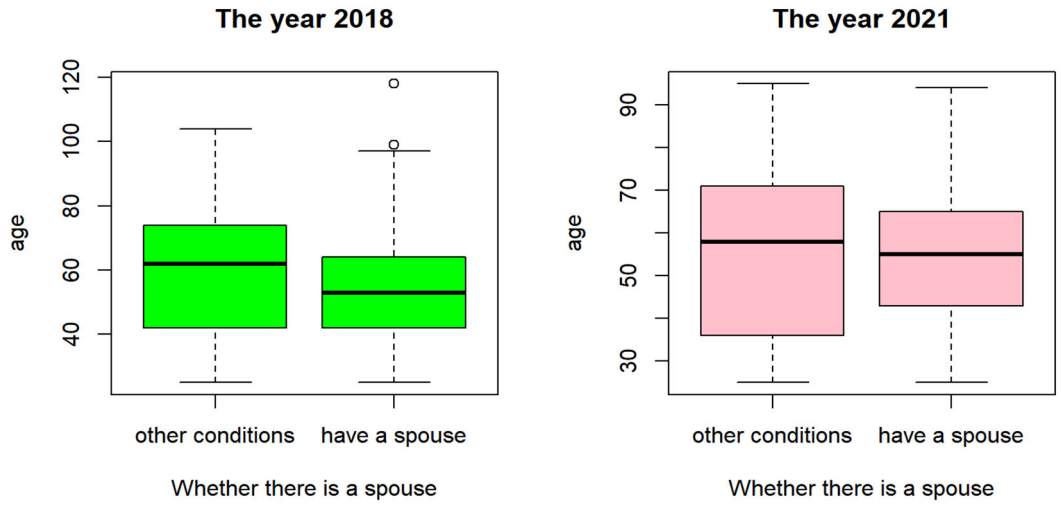

Figure 1. Box plot of "age" and dependent variable "married"

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202402_7(2).0003

Classification calculation and test are required to specifically explore the relationship between continuous explanatory variables and dependent variables, and the results are shown in Table 4. It can be seen from the results that in both 2018 and 2021, the housing floor area of the people with spouses is larger than that of the non-spouse population, and the average age is smaller than that of the non-married population. The total income of last year in 2018, which is the total income of 2017, is higher for people with spouses; Last year's total income in 2021, which is the total income in 2020, is higher for non-spouses.

Table 4. Relationship between the presence of spouses and continuous explanatory variables

Year	married	income_avg	area_avg	age_avg
2018	0	35861	95.3	57.9
	1	39132	113	52.8
2021	0	50313	105	55.1
	1	41386	122	54.2

To explore the relationship between classification explanatory variables and dependent variables, a Chi-square test was carried out for judgment. The null hypothesis of the test is that there is no association between explanatory variables and the presence or absence of spouses. The test results are shown in Table 5. According to the results, except for gender in 2018, all the other variables in the two years have very small P-values. The null hypothesis is rejected at the significance level of 5%, and it is considered that there is a significant correlation between the classification explanatory variable and the dependent variable.

Table 5. The relationship between the presence or absence of a spouse and a categorical explanatory variable

Year	Variable Name	Chi-square Value	Degree of freedom	P value
2018	gender	0.2840	1	0.5941
	edu	153.7900	12	<0.0001
	health	15.5650	4	0.0037
2021	gender	9.2474	1	0.0024
	edu	293.4000	12	<0.0001
	health	75.8170	4	<0.0001

Due to the relationship between the dependent variable and explanatory variables, data modeling will be analyzed below. However, a multicollinearity test should be carried out before modeling to prevent the collinearity between explanatory variables from causing bias in the interpretation of models and coefficients. The results of the multicollinearity test are shown in Table 6. It can be seen that VIF values of all explanatory variables in the two years are less than 10 to measure the degree of multicollinearity, indicating that there is no severe multicollinearity.

Table 6. Results of multicollinearity test

2018 VIF Value		2021 VIF Value	
edu	1.53	edu	1.50
age	1.38	income	1.33
income	1.29	age	1.30
health	1.13	health	1.15
gender	1.04	gender	1.04
area	1.02	area	1.02

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202402_7(2).0003

5. Model Analysis – Based on 2018 Data

Next, the logistic regression model was used for data modeling analysis. Firstly, the data of CGSS in 2018 were analyzed after processing. Now all explanatory variables are added to the logistic regression model for fitting, and the results are shown in Table 7.

Table 7. Logistic full model regression results (2018)

	Estimated Value	Standard Deviation	Z value	P value
intercept term	1.5330	0.1764	8.6900	<0.0001***
gender	0.0678	0.0503	1.3470	0.1779
edu2	-0.4888	0.2441	-2.0030	0.0452*
edu3	0.3982	0.0755	5.2770	<0.0001***
edu4	0.7115	0.0795	8.9480	<0.0001***
edu5	0.0759	0.2136	0.3550	0.7224
edu6	0.5920	0.0989	5.9860	<0.0001***
edu7	0.3360	0.1275	2.6350	0.0084**
edu8	-0.1619	0.3638	-0.4450	0.6563
edu9	0.6222	0.1701	3.6580	0.0003***
edu10	-0.0094	0.1370	-0.0690	0.9452
edu11	0.4449	0.1996	2.2290	0.0258*
edu12	-0.5189	0.1226	-4.2310	<0.0001***
edu13	-0.9399	0.2067	-4.5470	<0.0001***
health2	0.0487	0.1246	0.3900	0.6962
health3	0.2718	0.1235	2.2010	0.0278*
health4	0.2978	0.1209	2.4640	0.0137*
health5	0.1728	0.1309	1.3200	0.1867
income	0.0000	0.0000	2.4930	0.0127*
area	0.0035	0.0004	8.8390	<0.0001***
age	-0.0206	0.0018	-11.2630	<0.0001***

Note: "* * *", "* *", "*", respectively ($P < 0.001$, $P < 0.01$, $P < 0.05$, respectively under 0.001, 0.01, 0.05 significance level significantly. The asterisk below has the same meaning.

As can be seen from the results of Table 7, “edu” and “health” are ordered multiple categorical variables, and “edu=1” and “health=1” are the benchmarks for comparison, so they do not appear in the output results of the model. Where “edui” and “healthi” represent the cases where “edu=i” and “health=i,” respectively. It can be seen that all explanatory variables except “gender,” “edu” at some levels, and “health” at some levels were significant at the 0.05 significance level. The influence of “area” and “age” on the dependent variables is positive and negative respectively.

Taking “edu2” as an example, its coefficient value shows that if other conditions remain unchanged, the logarithmic probability of having a spouse decreases by 0.4888 when the highest level of education changes from 1 to 2, and the probability of having a spouse decreases by 11.98% after conversion. The interpretation of the remaining “edu” and “health” variables is the same, and the results are relative to “edu=1” and “health=1”. Keeping all other variables equal, the logarithmic probability of having a spouse (relative to no spouse) increases by 0.0035 for every one square meter increase in the floor “area”; For each additional year of “age”, the logarithmic probability of having a spouse decreases by 0.0206. Also translated into probability, the probability of having a spouse (relative to no spouse) increases by 0.0875% for every square meter increase in “area”; For every additional year of “age,” the probability of having a spouse decreases by 0.515%. In addition, the whole equation of the above model is also significant after the F test.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202402_7(2).0003

In order to avoid over-fitting and choose a more concise model, the stepwise regression method is used for variable selection. First, perform step-based regression, and variables can be eliminated according to information criteria AIC and BIC. The smaller the AIC and BIC are, the higher the fitting accuracy is and the better the prediction effect is. The results of model selection are shown in Table 8. Among them, AIC and BIC models respectively represent the models selected by AIC and BIC criteria and are the models whose AIC and BIC values reach the minimum respectively. The results show that the full model is different from the three models obtained by AIC and BIC criteria, among which the BIC model has the highest AIC value. The BIC value of the full model is the highest.

Table 8. Model selection results (2018)

Model	AIC	BIC
full model	10629.59	10783.04
AIC model	10629.40	10775.55
BIC model	10646.46	10756.07

Table 9 takes the AIC model as an example for analysis and gives the fitting result of the AIC model. It can be seen that the AIC model directly excludes the insignificant variable “gender” in the whole model, and all other variables except “edu5” and “edu10”, “health2” and “health5” are significant at the significance level of 5%. The model as a whole is also significant at the significance level of 5%, and it can be seen that the positive and negative results of the estimation of “area” and “age” are consistent with the whole model. In addition, the probability of having a spouse increased when the respondents' self-perceived health level was healthier than “health=1.”

Table 9. Logistic regression AIC model results (2018)

	Estimated Value	Standard Deviation	Z value	P value
intercept term	1.5340	0.1763	8.7000	<0.0001***
edu2	-0.4799	0.2439	-1.9680	0.0491*
edu3	0.4101	0.0749	5.4730	<0.0001***
edu4	0.7270	0.0787	9.2410	<0.0001***
edu5	0.0875	0.2134	0.4100	0.6819
edu6	0.6076	0.0982	6.1870	<0.0001***
edu7	0.3461	0.1273	2.7200	0.0065**
edu8	-0.1408	0.3629	-0.3880	0.6980
edu9	0.6374	0.1697	3.7550	0.0002***
edu10	0.0037	0.1367	0.0270	0.9782
edu11	0.4545	0.1995	2.2780	0.0227*
edu12	-0.5053	0.1222	-4.1340	<0.0001***
edu13	-0.9294	0.2066	-4.4990	<0.0001***
health2	0.0477	0.1245	0.3830	0.7019
health3	0.2718	0.1235	2.2010	0.0277*
health4	0.2999	0.1208	2.4830	0.0130*
health5	0.1769	0.1308	1.3520	0.1764
income	0.0000	0.0000	2.6850	0.0073**
area	0.0035	0.0004	8.8940	<0.0001***
age	-0.0204	0.0018	-11.1910	<0.0001***

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202402_7(2).0003

The prediction and ROC curve are made below. The ROC curve can be used to measure the advantages and disadvantages of the model visually. AUC is the area between the ROC curve and the y=0 line. In practical applications, the comparison of multiple models can choose the better model by the size of the area, and the selection criterion is that the larger the AUC, the better. Fig. 2 shows the inner sample ROC curves of the above three models, and the AUC values are calculated. The AUC of the full model is 0.6576, the AUC of the AIC model is 0.6577, and the AUC of the BIC model is 0.6526. It can be seen that the AUC values of the AIC model and the full model are close, while the AIC model is slightly better than the full model, but both of them are better than the BIC criterion model. As shown in Figure 6, the area between the ROC curve of the AIC model and the y=0 line is the largest. Therefore, the AIC model is superior, but to further illustrate the prediction accuracy, it is also necessary to conduct an external sample AUC.

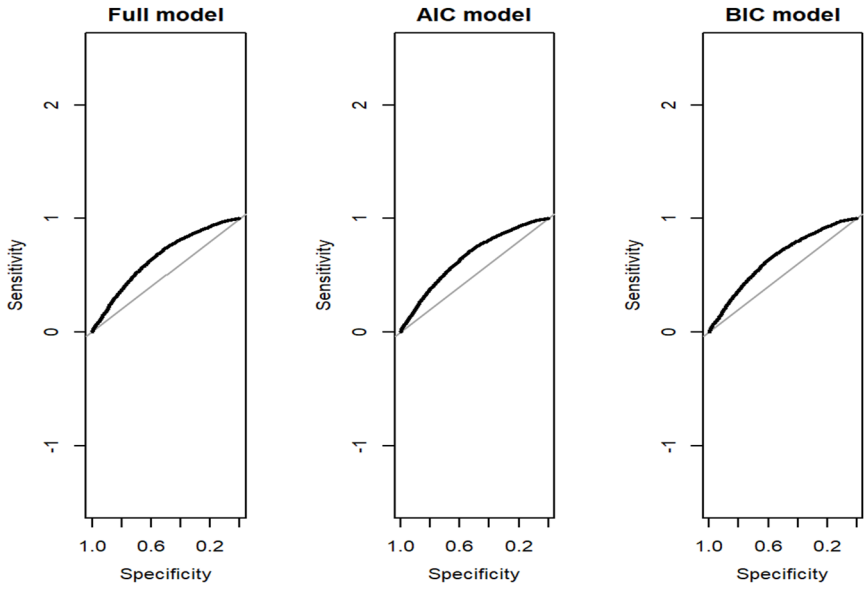

Figure 2. ROC curves of samples within three models (2018)

Next, the model accuracy evaluation test based on external samples is carried out, and the external sample AUC of the model is compared. In each experiment, the original data were randomly arranged, with the first 80% as training samples and the last 20% as verification samples, and the experiment was repeated 200 times. Fig. 3 shows the box plot of the predicted AUC value of the external sample. It can be seen from the box plot that the AIC model still has the highest prediction accuracy. Therefore, the AIC model was chosen as the factor model for a spouse or not in CGSS data modeling in 2018.

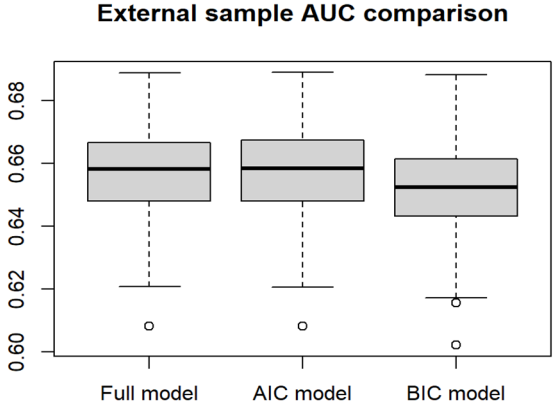

Figure 3. Box plots of external sample forecast (2018)

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202402_7(2).0003

6. Model Analysis – Based on 2021 Data

Then, the logistic regression model was used for data modeling to analyze the CGSS processed data in 2021. Now all explanatory variables are added to the logistic regression model for fitting, and the results are shown in Table 10.

Table 10. Logistic full model regression results (2021)

	Estimated Value	Standard Deviation	Z value	P value
intercept term	0.8086	0.2302	3.5130	0.0004***
gender	0.0117	0.0647	0.1810	0.8562
edu2	-0.7924	0.3325	-2.3830	0.0172*
edu3	0.3505	0.1136	3.0850	0.0020*
edu4	0.5436	0.1147	4.7400	<0.0001***
edu5	0.0396	0.2599	0.1520	0.8790
edu6	0.5222	0.1382	3.7790	0.0002***
edu7	0.2473	0.1684	1.4690	0.1420
edu8	0.4571	0.5681	0.8050	0.4210
edu9	0.5210	0.2062	2.5260	0.0115*
edu10	-0.1127	0.1887	-0.5970	0.5502
edu11	0.3469	0.2265	1.5320	0.1256
edu12	-0.4514	0.1602	-2.8180	0.0048**
edu13	-1.0440	0.2699	-3.8670	0.0001***
health2	0.2016	0.1516	1.3300	0.1835
health3	0.3977	0.1423	2.7950	0.0052**
health4	0.3923	0.1424	2.7540	0.0059**
health5	0.2870	0.1541	1.8620	0.0626
income	0.0000	0.0000	-1.2060	0.2278
area	0.0026	0.0005	5.6940	<0.0001***
age	-0.0076	0.0024	-3.1560	0.0016**

According to the results of Table 10, similar to the analysis results of 2018, the influences of “area” and “age” on the dependent variables are positive and negative respectively, and “gender” is not significant. However, “income” is not significant in the 2021 dataset. “edu=1” and “health=1” are the benchmarks for comparison and therefore do not appear in the model's output, “edui” and “healthi” represent cases where “edu=i” and “health=i,” respectively. It can be seen that all explanatory variables except “gender,” “income,” “edu” at some levels, and “health” at some levels are significant at the 0.05 significance level. Taking “edu2” as an example again, its coefficient value shows that if other conditions remain unchanged, the logarithmic probability of having a spouse decreases by 0.7924 when the highest level of education changes from 1 to 2, and the probability of having a spouse decreases by 18.83% after conversion. The interpretation of the remaining “edu” and “health” variables is the same, and the results are relative to “edu=1” and “health=1.”

Keeping all other variables equal, the logarithmic probability of having a spouse (relative to no spouse) increases by 0.0026 for every one square meter increase in the floor “area”; For each additional year of “age,” the logarithmic probability of having a spouse decreases by 0.0076. Also translated into probability, the probability of having a spouse (relative to no spouse) increases by 0.0650% for every square meter increase in “area”; For every additional year of

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202402_7(2).0003

"age," the probability of having a spouse decreases by 0.190%. In addition, the whole equation of the above model is also significant after the F test.

The stepwise regression method is still used for variable selection below. The results of model selection are shown in Table 11. Among them, the AIC and BIC models respectively represent the models selected by the AIC and BIC criteria and are the models whose AIC and BIC values reach the minimum respectively. According to the results of Table 11, the whole model is different from the three models obtained by the AIC and BIC criteria, in which the BIC model still has the highest AIC value and the whole model has the highest BIC value.

Table 11. Model selection results (2021)

Model	AIC	BIC
full model	6343.413	6484.686
AIC model	6340.842	6468.661
BIC model	6343.679	6444.588

Table 12 takes the AIC model as an example to analyze and gives the fitting result of the AIC model. It can be seen that the AIC model removes gender and income variables that are not significant in the whole model, and some of the remaining variables are not significant at the significance level of 5%, and the significance performance of the coefficient is not as good as the modeling results of 2018. However, the model as a whole is significant at the significance level of 5%, and it can be seen that the positive and negative results of the estimation of "area" and "age" are still consistent with the whole model. In addition, all "health" levels remain positive.

Table 12. Logistic regression AIC model results (2021)

	Estimated Value	Standard Deviation	Z value	P value
intercept term	0.7938	0.2297	3.4550	0.0006***
edu2	-0.8086	0.3316	-2.4380	0.0148*
edu3	0.3519	0.1125	3.1290	0.0018**
edu4	0.5386	0.1128	4.7740	<0.0001***
edu5	0.0229	0.2585	0.0890	0.9293
edu6	0.5111	0.1359	3.7620	0.0002***
edu7	0.2304	0.1666	1.3830	0.1665
edu8	0.4500	0.5673	0.7930	0.4277
edu9	0.4958	0.2042	2.4280	0.0152*
edu10	-0.1445	0.1852	-0.7810	0.4350
edu11	0.3014	0.2221	1.3570	0.1749
edu12	-0.5000	0.1527	-3.2740	0.0011**
edu13	-1.1385	0.2564	-4.4400	<0.0001***
health2	0.2021	0.1516	1.3330	0.1825
health3	0.3947	0.1423	2.7740	0.0055**
health4	0.3863	0.1423	2.7140	0.0066**
health5	0.2795	0.1539	1.8160	0.0694
area	0.0026	0.0005	5.7260	<0.0001***
age	-0.0074	0.0024	-3.1070	0.0019**

The prediction and ROC curve are made below. Fig. 4 shows the inner sample ROC curves of the above three models and calculates the AUC values respectively. The AUC of the full model is

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202402_7(2).0003

0.6248, that of the AIC model 0.6239, and that of the BIC model 0.6205. It can be seen that compared with the data modeling results of 2018, the AUC value of the 2021 model is generally low, and the prediction ability is weak. In the 2021 data modeling results, the AUC value of the full model was the highest.

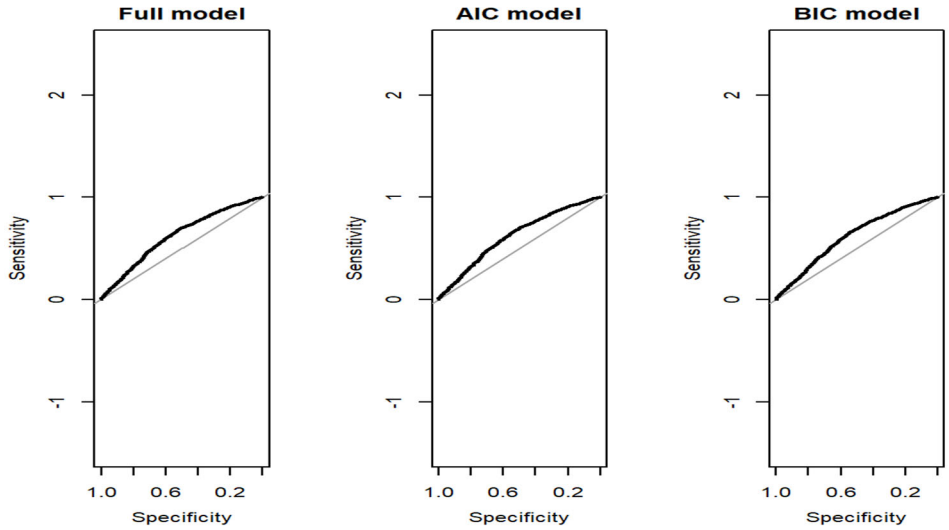

Figure 4. ROC curves of samples within three models (2021)

Next, the model accuracy evaluation test based on external samples is conducted to compare the external sample AUC of the model. The experiment was repeated 200 times, and the original data were randomly arranged in each experiment, with the first 80% as the training sample and the last 20% as the verification sample. Fig. 5 shows the box plot of the predicted AUC value of the external sample. As can be seen from the box plot, the prediction accuracy of the full model is still the highest. Therefore, the full model is chosen as the factor model of whether there is a spouse in CGSS data modeling in 2021.

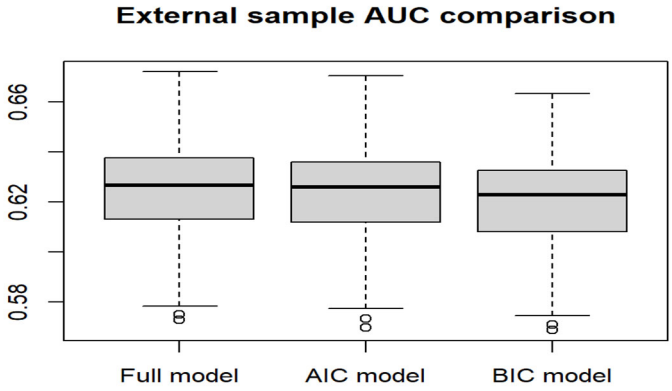

Figure 5. Box plots of external sample forecast (2021)

7. Comparison and Analysis of Results for 2018 and 2021

This section makes a comparative analysis of the optimal models obtained in the previous two sections. Table 13 shows the prediction of the probability of having spouses of respondents in their respective samples by the optimal logistic regression model obtained in the two years. As can be seen from the results in the table, the average predicted probability of respondents

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202402_7(2).0003

having a spouse in 2018 and 2022 is 0.7957 and 0.7797, respectively, and the median is 0.8135 and 0.7981 respectively. In addition, other indicators of the two years are close. Overall, the predictions made by the model were optimistic about the respondents having a spouse.

Table 13. Descriptive statistical analysis with spousal probability prediction

	Minimum Value	First Quartile	Median	Mean	Third Quartile	Maximum Value
2018 AIC model	0.3433	0.7436	0.8135	0.7957	0.8639	0.9960
2021 full model	0.3347	0.7463	0.7981	0.7797	0.8296	0.9828

Since the 2018 data modeling AUC curve exceeded 0.65, while the 2022 AUC was only around 0.62, the accuracy and effectiveness of the 2018 data modeling for prediction will be better. Two sets of prediction Settings are given below, and their explanatory variable Settings are shown in Table 14. It can be seen that the first respondent is a 40-year-old male, who graduated from high school and is in a relatively healthy state, with a total income of 800,000 yuan last year and a housing construction area of 100 square meters. The second respondent is a 30-year-old female with a college degree and is in a very healthy state, with a total income of 200,000 yuan last year and a housing construction area of 60 square meters.

Table 14. Predictive Setting

Explanatory Variable	Setting 1	Setting 2
gender	1	0
edu	6	12
health	4	5
income	800000	200000
area	100	60
age	40	30

The predicted results of the two sets are shown in Table 15. From the results, the probability of the respondent having a spouse in different years can be obtained. It can be seen that the predicted probabilities of 2018 and 2021 are quite different, and the probability of male respondents in setting 1 in 2018 having a spouse is more than 95%, which is much higher than the average given by the AIC model in the same year. However, the full model for 2021 predicts that the probability of this male respondent having a spouse for that year is less than 78%, and lower than the mean and median forecast for 2018. In addition, considering the female respondents in setting 2, the predicted probability of having a spouse in any given year was lower than the mean or median for the same year.

Table 15. Prediction of the Outcome

Year and Model	Setting 1	Setting 2
2018 AIC model	0.9557	0.7460
2021 full model	0.7759	0.6131

8. Conclusion

In conclusion, a logistic regression model was established for the data after CGSS pretreatment in 2018 and 2021 for analysis, to explore whether there were spouse-influencing factors, and to make appropriate predictions in this paper. The following conclusions can be drawn from the data modeling and prediction results:

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202402_7(2).0003

First of all, in both 2018 and 2021, the age of the respondents and the floor area in the respondents' housing significantly affected the marital status, while gender was not significant in any year. The optimal model selected by the model in 2018 is the AIC model, while in 2021 it is the full model with all explanatory variables. In the AIC model of 2018, the respondents' total income last year was significant, but in the full model of 2021, the data of 2021 was used to conclude that “income” was not significant. In addition, some levels of “health” and “edu” were significant in different years. From the perspective of prediction, the predictive effect and accuracy of data modeling in 2018 may be better than that in 2021.

In addition, the influence of the floor area in the respondents' housing on the dependent variable is positive, and the influence of the respondents' age is negative. The coefficient of respondents' total income last year is very small, which may be related to the unit of yuan; Respondents who thought they were healthy were more likely to have a spouse than those who thought they were very unhealthy. The positive and negative signs of different changes in the highest level of education will often change, and the probability of having a spouse does not increase completely with the increase in education. Therefore, it can be inferred that the housing area is one of the more critical factors, and the population with a spouse may generally have a stable residence and a comfortable accommodation environment. In addition, with a certain increase in age, the probability of finding a spouse decreases. And healthy people are favored over unhealthy people, so the probability of having a mate increases.

According to the above conclusions, the following policy recommendations can be put forward: First, to encourage housing construction and improvement, especially to provide a larger building area of housing units. The provision of suitable housing conditions can increase people's marital stability and the quality of family life. Secondly, strengthen health education and measures to promote physical health in order to improve people's health level. Individuals in good health are more likely to have stable marital relationships, so by encouraging healthy lifestyles, people's chances of mate choice can be increased. In addition, providing people with equal opportunities for career development and increasing income levels can help improve marital stability. Governments can help promote income equity and provide training opportunities to raise people's income levels, thereby increasing their choice of spouse. Attention should also be paid to the quality of education and skills training, and the provision of a comprehensive education system, including vocational education and skills training, helps to improve people's employment opportunities and income levels, thereby affecting their spouse's situation. Finally, marriage counseling and support can also be provided among young people to help them better cope with marriage problems and challenges. Future research could further explore other potential influencing factors and the relationship between marital status and other social phenomena, thereby deepening the understanding of marital status and social development.

Acknowledgments

Minzu University of China.

References

[1] D. Wu. Internet Usage and Residents' Fertility Intentions: Empirical Evidence Based on CGSS Data (Henan University Journal (Social Science Edition), China 2023), p.32-37+153. (In Chinese).

[2] H. P. Han. Analysis of Factors Affecting the Reemployment of the Elderly in China: An Empirical Study Based on CGSS 2017 (China Collective Economy, China 2023), No.731(03), p.165-168. (In Chinese).

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202402_7(2).0003

[3] Z. R. Wang, C. Y. Zhao. The Impact of Housing Prices on Family Fertility Intentions: Evidence from Micro Data of CGSS (China Real Estate, China 2023), No.776(03), p.6-17. (In Chinese).

[4] W. Zhou, H. R. Ma. The Influence of Intergenerational Educational Mobility on Subjective Well-being of Descendants: An Empirical Study Based on CGSS 2017 (Journal of Chongqing Jiaotong University (Social Science Edition), China 2023), p.77-85. (In Chinese).

[5] Z. Wang, B. B. Su. Research on the Employment Situation of Chinese Youth and Its Influencing Factors: An Empirical Analysis Based on CGSS 2017 Data (Northwest Population Journal, China 2022), p.42-53. (In Chinese).

[6] S. Gao. Analysis of the Income Promotion Effect of Universal Education in Minority Areas: An Empirical Analysis Based on CGSS 2017 Survey Data (Modern Marketing (Academic Edition), China 2021), No.204(12), p.147-149. (In Chinese).

[7] X. D. Pei, J. Song. The Influence of Family Care on the Labor Participation Rate of Married Urban Women: An Empirical Analysis Based on CGSS 2017 Data (Journal of Xi'an Shiyou University (Social Science Edition), China 2023), p.20-26. (In Chinese).

[8] K. L. He. Analysis of Factors Affecting Personal Education Returns for Graduate Students: An Empirical Study Based on CGSS 2015 (Journal of Economic Research Guide, China 2022), No.524(30), p.137-140. (In Chinese).

[9] J. S. Liu, L. Du. Does Rising House Prices Delay First Marriage? An Empirical Analysis Based on CGSS Data (Journal of Beihang University (Social Sciences Edition), China 2022), p.1-10. (In Chinese).

[10] D. H. Yin, J. Y. Song, M. Y. Chen, et al. Impact of Leisure Modes on the Well-being of Rural Elderly from the Perspective of Rural Revitalization: An Empirical Analysis Based on CGSS Data (Anhui Agricultural Science Bulletin, China 2022), p.13-16. (In Chinese).

[11] Information on http://cgss.ruc.edu.cn.