



**FINDING FACTORS AFFECTING MARRIAGE RATE AND
MARRIAGE PREDICTION IN CHINA USING PANEL
DATA ANALYSIS AND MACHINE LEARNING**

DEYU ZHANG

**MASTER OF SCIENCE
IN
INFORMATION TECHNOLOGY**

**SCHOOL OF APPLIED DIGITAL TECHNOLOGY
MAE FAH LUANG UNIVERSITY**

2024

©COPYRIGHT BY MAE FAH LUANG UNIVERSITY

**FINDING FACTORS AFFECTING MARRIAGE RATE AND
MARRIAGE PREDICTION IN CHINA USING PANEL
DATA ANALYSIS AND MACHINE LEARNING**

DEYU ZHANG

**THIS THESIS IS A PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
INFORMATION TECHNOLOGY**

**SCHOOL OF APPLIED DIGITAL TECHNOLOGY
MAE FAH LUANG UNIVERSITY**

2024

©COPYRIGHT BY MAE FAH LUANG UNIVERSITY

**FINDING FACTORS AFFECTING MARRIAGE RATE AND
MARRIAGE PREDICTION IN CHINA USING PANEL
DATA ANALYSIS AND MACHINE LEARNING**


DEYU ZHANG


THIS THESIS HAS BEEN APPROVED
TO BE A PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF SCIENCE


IN
INFORMATION TECHNOLOGY


2024

EXAMINATION COMMITTEE


.....CHAIRPERSON
(Asst. Prof. Sujitra Arwatchananukul, Ph. D.)


.....ADVISOR
(Asst. Prof. Worasak Rueangsirarak, Ph. D.)


.....CO-ADVISOR
(Surapong Utama, Ph. D.)


.....EXTERNAL EXAMINER
(Asst. Prof. Teerawat Kamnardsiri, Ph. D.)

ACKNOWLEDGEMENTS

First of all, this thesis is dedicated to my parents, and the relatives for being with me all the times. Without my parents' support encourage and endless love, I can't reach goals of my research and studying without any concerns.

Secondly, I would like to thank my supervisors Assistant Professor Dr. Worasak Rueangsirarak and Arjan Dr. Surapong Uttama, Assistant Professor Dr. Santichai Wicha program coordinator. Their constant support and encouragement throughout the process was invaluable. support, guidance, and encouragement were invaluable. Their unwavering presence and wealth of wisdom have been instrumental in my academic growth from the initial stages of starting my research program to the final submission of my dissertation. I greatly value the weekly meetings we hold, which not only serve as important checkpoints to keep me on track academically but also give me great encouragement. I am very grateful for their immeasurable contribution to my development. Also, this research was supported by a writing a thesis grant from Mae Fah Luang University.

In addition to my mentors and university, I would like to thank my illustrious lab partners, whose support has been a constant source of motivation. Our collaborative writing sessions and informal chats, whether conducted via screen during lockdowns or in person when circumstances allow, have provided a lifeline during the most challenging times.

Finally, I would like to express my deepest gratitude to my family for their trust and support in my abilities. Your encouragement has played an integral role in my accomplishments. To my mom, dad, and sister: thank you for everything. I would like to dedicate this master's thesis to you.

Deyu Zhang

Thesis Title Finding Factors Affecting Marriage Rate and Marriage Prediction in China Using Panel Data Analysis and Machine Learning

Author Deyu Zhang

Degree Master of Science (Information Technology)

Advisor Asst. Prof. Worasak Rueangsirarak, Ph. D.

Co-Advisor Surapong Uttama, Ph. D.

ABSTRACT

After China's accession to the WTO and 20 years of rapid development, the marriage rate has shown a downward trend. The main factors leading to the decline in marriage rate are the rapid growth of housing prices and the high price of betrothal gifts. Then, in this study, the adoption of big data analytic is proposed to highlight the significant factors effects to a decision making of new generation Chinese people. The first phase of research aims at fitting machine learning models with the marriage-related data, understanding which attributes affect the marriage rate and predicting the marriage rate. The data collection scope includes seven independent variables related to marriage rate such as GDP, house prices, birth rate, education level etc. over 31 regions in China during 2003-2022. Then the study applied three regression models - Pooled OLS, Random Effects, and Fixed Effects - in predicting China's crude marriage rate. The Random Effects model outperformed both the Pooled OLS and Fixed Effects models, as evidenced by its highest R^2 value (0.2910). However, based on Hausman Test, p-value of $6.458e-16$.the indicate Fixed Effects model was preferable. All models suggested that the average year of education had the most positive effect to the marriage rate while the house price greatly negated the marriage rate. Results showed the

Random Effects model, with an R^2 of 0.2910, as the best fit. Key predictors included GDP, house prices, and gross dependency ratio (negative effects), and sex ratio and education (positive effects). The Effects model excels in prediction, with the lowest MSE (1.6610), RMSE (1.2888) and Random Effects model excels in prediction, with the lowest MSE (1.6610), RMSE (1.2888) and MAE (1.0888).

The second phase of research study aims to analyze the impact of socio-economic factors on the crude marriage rate (CMR) panel data in China from 2003 to 2022 using Dual Machine Learning (DML) for Causal Inference and machine learning models. Four models—XGBoost, LightGBM, CatBoost, and GBDT—were employed for predictions, using 10-fold cross-validation for model evaluation. The results indicated that education and birth rate had the most significant positive impacts on CMR, while GDP showed positive but varying effects, and the female proportion had a notable negative impact. CatBoost performed best in MSE (0.942) and RMSE (0.958), while LightGBM excelled in MAE (0.777). Education, GDP, and birth rate are key factors influencing CMR. CatBoost and LightGBM proved to be effective prediction models, though improvements are needed for regions with significant variability.

After comparing different models, it can be concluded that the Random Effects model performed the best across all evaluation metrics (MSE, RMSE, MAE), demonstrating the advantage of traditional statistical models on this dataset. Although CatBoost performed relatively well among the machine learning models, its overall error was still higher than that of the Random Effects model, with XGBoost and GBDT showing larger errors. This indicates that, in this specific dataset, traditional statistical models outperform more complex machine learning models, highlighting the importance of optimizing model selection based on the characteristics of the data.

Keywords: Marriage Rate, Panel Data, Panel Regression, Hausman Test, Dual Machine Learning, Causal Inference, CatBoost, Marriage Rate Prediction

TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENTS	(3)
ABSTRACT	(4)
LIST OF TABLES	(9)
LIST OF FIGURES	(10)
ABBREVIATIONS AND SYMBOLS	(12)
CHAPTER	
1 INTRODUCTION	1
1.1 Background and Importance of the Research Problem	1
1.2 Research Objectives	6
1.3 Importance of Research	8
1.4 Research Hypothesis	10
1.5 Scope of Research	12
1.6 Research Limitations	13
2 LITERATURE REVIEW	16
2.1 Theoretical Reviews	16
2.2 Related Studies	18
2.3 Related Theories of Term Definition	32
2.4 The Panel Data Regression Model of the Hausman-Test	33
2.5 The Machine Learning Models	39
2.6 Reflection on Literature Review	43

TABLE OF CONTENTS (continued)

	Page
CHAPTER	
3 RESEARCH METHODOLOGY	45
3.1 Overall Methodology	45
3.2 Data Collection	47
3.3 Data Merging	47
3.4 Data Pre-Processing	47
3.5 Data Selection	48
3.6 Development Environment	49
3.7 Exploratory Data Analysis (EDA)	49
3.8 Feature Scaling	50
3.9 Dataset Splitting	53
3.10 Data Visualization Exploration	55
3.11 The Panel Data Regression Model	59
3.12 Evaluation Metrics in Machine Learning	60
4 RESULTS	63
4.1 Distribution of Marriage Rate	63
4.2 Numerical Features Boxplot	68
4.3 Panel Data Regression Models Results	69
4.4 Pooled OLS, Random Effects and Fixed Effects Evaluation Index Results and Prediction	71
4.5 Causal Inference Using Dual Machine Learning (DML) with XGBoost, LightGBM, CatBoost, and GBDT	74

TABLE OF CONTENTS (continued)

	Page
CHAPTER	
4.6 Prediction of Marriage Rate of XGBoost, LightGBM, CatBoost, and GBDT	76
4.7 Discussion	78
4.8 Summary	79
5 CONCLUSIONS	81
5.1 Conclusions	81
5.2 Suggestions	82
REFERENCES	85
CURRICULUM VITAE	94

LIST OF TABLES

Table	Page
3.1 Features Data Selections	48
3.2 Panel Data Structure	60
4.1 The Model Comparison Results of Pooled OLS, Random Effects, Fixed Effects	71
4.2 Pooled OLS, Random Effects and Fixed Effects Evaluation Index Results	72
4.3 The Results of Features, ATE, CATE and HTE	75
4.4 10 K-Fold Cross Validation Results for 4 Models	76
4.5 The Results of XGBoost, LightGBM, CatBoost, and GBDT Evaluation Metric	77
4.6 The Summary of Model Results	80

LIST OF FIGURES

Figure	Page
1.1 Photos Wedding Dinner Party	2
1.2 Photos of Traditional Chinese Wedding Ceremony	4
1.3 Number of Registered Pairs of Marriages and Number of First Marriages, in 1985-2020	6
2.1 The Betrothal Gift Required for Engagement (Excluding House and Car)	19
2.2 Tencent's Guyu Data Released a Survey on the Country's Bride Price Situation in 2020 Based on the Responses of 1,846 Chinese Residents	20
2.3 A Bride Price Map Circulating on Chinese Social Media Platforms, Ranking Regions Based on Average Bride Price in 2022	21
2.4 Amount of Bridge Prices in 7 Economic Zones	21
2.5 Trend of Per Capita Disposable Income of National Residents (Unit: yuan)	22
2.6 Chinese Wedding Room	23
2.7 Comparison of Individual Characteristics of Education by Monthly Income Range (From the Most Recent Survey by the Chinese Household Income Project (CHIP) was Conducted in 2018)	25
2.8 Comparison of Characteristics of Individuals with Bachelor's Degree or Above in Monthly Income Range (From the Most Recent Survey by the Chinese Household Income Project (CHIP) was Conducted in 2018)	26
2.9 National Population by Sex, Educational Level and Age at First Marriage (The Seventh National Population Census in 2020)	29
2.10 Proportion of Population with Educational at Age of First Marriage in Mainland China	30

LIST OF FIGURES (continued)

Figure	Page
2.11 Chinese Education System	31
2.12 China's Gender Ratio in 2020	32
2.13 The Concept of Panel Data Structure	34
2.14 The Processing History of XGBoost, LightGBM and Catboost	40
3.1 Overall Methodology	46
3.2 Correlation Heatmap of Features	50
3.3 Raw Data	51
3.4 Features Scaling Data	52
3.5 Features Describe Data	52
3.6 Two-Way Split (7:3 or 8:2)	53
3.7 Three-Way Split (6:2:2)	54
3.8 The Series of Distribution Plots	56
3.9 Crude Marriage Rate vs. Socioeconomic Factors	58
3.10 Pair Plot of Crude Marriage Rate and Socioeconomic Variables	59
4.1 Spatial and Temporal Distribution Map of Marriage Rate in 2003	64
4.2 Spatial and Temporal Distribution Map of Marriage Rate in 2012	65
4.3 Spatial and Temporal Distribution Map of Marriage Rate in 2022	66
4.4 Numerical Features Boxplot	69
4.5 Actual vs Predicted Marriage Rate in 2022 (Pooled OLS, Random Effects, Fixed Effects)	73
4.6 Actual vs Predicted Marriage Rate in 2022(XGBoost, LightGBM, CatBoost, GBDT)	78

ABBREVIATIONS AND SYMBOLS

CMR	Crude Marriage Rate
GDP	Gross Regional Product (100 million yuan)
RE	Random- Effects
FE	Fixed Effects
ATE	Average Treatment Effect
CATE	Conditional Average Treatment Effect
HTE	Heterogeneous Treatment Effect
DML	Dual Machine Learning
CI	Causal Inference
GBDT	Gradient Boosting Decision Tree
XGBoost	Extreme Gradient Boosting
LightGBM	Light Gradient Boosting Machine
CatBoost	Categorical Boosting
CV-MSE	Cross-Validation Mean Squared Error
CV-RMSE	Cross-Validation Root Mean Squared Error
CV-MAE	Cross-Validation Mean Absolute Error
CV-R ²	Cross-Validation Coefficient of Determination
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error

CHAPTER 1

INTRODUCTION

1.1 Background and Importance of the Research Problem

Marriage is one of the most significant social institutions for both men and women, playing an exceptionally crucial role in various social activities, including happiness, reproduction, child development, gender inequality, crime, and laying the groundwork for resolving labor supply relationships in employment (Chiappori et al., 2002; Zimmermann & Easterlin, 2006; Edlund et al., 2013; Greenwood et al., 2014). Marriage is one of the most significant social institutions for both men and women. As shown in Figure 1.1, shows that the photos wedding dinner party. It plays a crucial role in various social activities, including happiness, reproduction, child development, gender inequality, crime, and providing the foundation for addressing the labor supply in the field of employment. However, in recent decades, marriage rates in many countries have sharply declined, beginning in developed countries such as Western European and American countries, followed by East Asian countries like Japan and South Korea, with China closely following. Since the late 1980s, the first marriage rate in China has been decreasing, while the age of first marriage has been continuously increasing. This trend will directly impact China's overall fertility rate, education, and the participation of people as the main factors in social activities related to labor, causing adverse effects specifically reflected in low fertility rates (Wrenn et al., 2019). Currently, there are three competing explanations in the literature regarding the changes in family formation and marriage - a decrease in fertility rates, the rise of professional women, and the increasing involvement of women in the labor force. Furthermore, there has been a rise in the frequency of women's participation in the labor market, and an enhancement in the level and duration of women's education (Oppenheimer, 1988, 1994; Blossfeld & Jaenichen, 1992; Malhotra, 1997). The female-to-male ratio in Chinese

universities has reached a relatively high level. While these factors partly explain the reasons behind the declining first marriage rate in China, they cannot fully account for the trends in all regions, referring to the 31 provincial-level units in mainland China.



Source Little Red Book Application

Figure 1.1 Photos Wedding Dinner Party

We propose alternative hypotheses regarding the marriage rates in different provinces of mainland China, related to regional housing prices, GDP, personal income, consumption, gender ratios, and betrothal gifts. Specifically, we hypothesize that in various provinces in mainland China, it is customary for men to purchase houses before marriage and to provide betrothal gifts to the women. Furthermore, with housing prices

and betrothal gifts rapidly rising along with economic development, this is leading to a decrease in the first marriage rate under significant livelihood pressures (Wrenn et al., 2019). As shown in Figure 1.2, shows that the Photos of Traditional Chinese Wedding Ceremony. Looking back at the housing prices and rapid growth during this period, which is only 20 years, what factors have led to such a huge change in the Chinese people's mindset? China officially became a member of the World Trade Organization on December 11, 2001, which means that since 2000, China's economy has started to grow rapidly. Prior to this, the form of Chinese bride price ranged from the "three rounds and one ring" in the 1970s (bicycle, sewing machine, radio), the refrigerators, washing machines, and televisions in the 1980s, to computers, air conditioners, and motorcycles in the 1990s. Crossing over to the 21st century, the form of Chinese bride price has shifted primarily to cash, real estate, and cars. The amount of cash has risen from 30,000 RMB to over 300,000 RMB (the amount of bride price varies by region and perception); the first payment for a housing range from tens of thousands to several hundred thousand RMB (housing prices vary in different cities), and buying a car costs around 100,000 RMB. Therefore, paying the bride price has become a significant economic burden for many rural farmers (Chen & Pan, 2023) there has been a significant increase in urban residential property prices in mainland China, surpassing the growth rate of urban household incomes. The price surge in urban areas is particularly higher compared to rural areas, posing increasing challenges for many low- and middle-income families to afford housing (Zhang, 2015; Li et al., 2022). A significant body of research on the influence of increasing property prices on the overall economy suggests that fluctuations in property prices and household wealth can impact fertility rates (Lovenheim & Mumford, 2013; Dettling & Kearney, 2014), employment (Mian & Sufi, 2014; Johnson, 2014), entrepreneurial activities (Corradin & Popov, 2015; Harding & Rosenthal, 2017), education (Lovenheim, 2011; Lovenheim & Reynolds, 2013), wealth disparity (Piketty & Zucman, 2014), investment and financial decisions (Chetty et al., 2017), and consumption patterns (Campbell & Cocco, 2007). We have elaborated on this body of literature and carried out a thorough examination of the influence of different factors such as average housing prices, annual GDP, per capita GDP, disposable income, per capita consumption, dependency ratio,

gender ratio, etc. on the marriage choices of young individuals in different provinces of mainland China from 2003 to 2022.



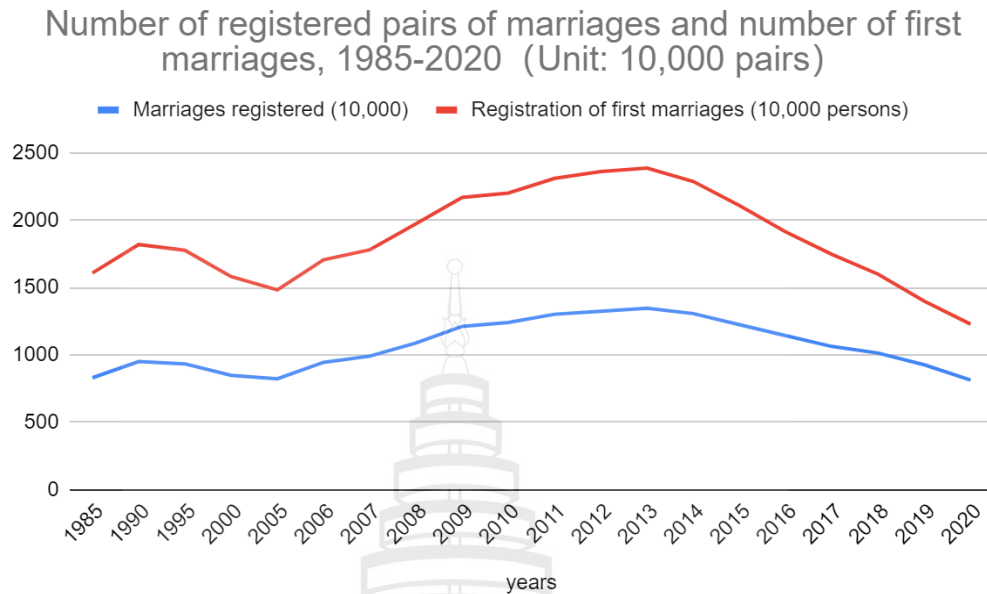
Source Little Red Book Application

Figure 1.2 Photos of Traditional Chinese Wedding Ceremony

Accurately predicting marriage rates is crucial for governments to formulate policies and strategic decisions. Simple linear regression is insufficient for addressing current research questions. Time series analysis can study the temporal dynamics within individual entities, while panel data analysis is more suitable for studying changes

across multiple entities over time. Controlling for unobserved heterogeneity and conducting comparative analysis are essential. However, due to the limited frequency of our study data, which spans only 20 years, using time series analysis may not yield the expected results. This makes panel data a better choice for research involving annual data from different regions, countries, or other entities. This study delves into the intricate dynamics of marriage rates, revealing seven key factors through the application of machine learning techniques. Bresson and Chaturvedi (2023), Jin et al. (2020) proposed an efficient, distribution-free least squares estimation method that utilizes the eigen decomposition of a weight matrix in a dynamic space–time pooled panel data model. The effectiveness of four regression models—Pooled OLS, Random Effects, Fixed Effects, and Panel OLS—is assessed, highlighting the exceptional performance of the Fixed Effects model and the robust predictive capabilities of the Random Effects model. Model training and evaluation are conducted using data from 31 regions across China, focusing on seven critical influencing factors such as GDP, housing prices, and education.

Accurately predicting marriage rates is crucial for governments to formulate policies and strategic decisions. Simple linear regression is insufficient for addressing current research questions. Time series analysis can study the temporal dynamics within individual entities, while panel data analysis is more suitable for studying changes across multiple entities over time. Controlling for unobserved heterogeneity and conducting comparative analysis are essential. As shown in Figure 1.3, shows that the number of registered pairs of marriages and number of first marriages, in 1985-2020. However, due to the limited frequency of our study data, which spans only 20 years, using time series analysis may not yield the expected results. this makes panel data a better choice for research involving annual data from different regions, countries, or other entities. This study delves into the intricate dynamics of marriage rates, revealing seven key factors through the application of machine learning techniques. S. R, Sai Sanjay Shyam et al. (2020) used the effectiveness of four regression models—XGBoost, LightGBM, CatBoost, GBDT models —is assessed, highlighting the exceptional performance of the CatBoost model and the robust predictive capabilities.



Source China Marriage and Family Report (2022)

Figure 1.3 Number of Registered Pairs of Marriages and Number of First Marriages, in 1985-2020

The research is structured as follows. Section 2 begins with an overview of the previous researches on marriage rate and factors affecting it. Section 3 describes research methodology. Section 4 presents results and discussion and section 5 draws a conclusion.

1.2 Research Objectives

The objectives of this research are:

1.2.1 Complete Intake Survey for Research using Panel Data Regression

To find the relationship between marriage rate (dependent variable) and provincial GDP, house prices, gross dependency ratio, birth Rate, female, average years of education per capita, and sex ratio (independent variables). using panel data analysis for all years from 2003 to 2022.

1.2.2 Use Hausman Test to Verify the Hypothesis

Based on the basic definition, we test whether the P-value is less than 0.05, and use the correlation between individual effects and other explanatory variables as the basis for screening fixed effect and random effect models.

1.2.3 Initial Evaluation of Machine Learning on Panel Data

Use machine learning to split the data and use panel data for machine learning, test Pooled OLS, Random Effects, Fixed Effects models, and evaluate MSE, RMSE, MAE.

1.2.4 Predicting Marriage Rates in 2022 Using Panel Data

Use machine learning to split the data and use panel data for machine learning, test Pooled OLS, Random Effects, Fixed Effects models, and evaluate MSE, RMSE, MAE

1.2.5 ATE, CATE, HTE and Dual Machine Learning to Calculate the Impact Factor of Each Independent Variable

The mean treatment effect, which measures the average effect of the treatment (ATE) on the outcome variable in the population, reflects the population-wide causal effect. The conditional mean treatment effect (CATE), which measures the average effect of the treatment on the outcome variable under a particular condition, reflects differences in causal effects across subpopulations. Heterogeneous treatment effects (HTE) refer to the variation in the impact of a treatment across different individuals or groups. In simpler terms, it means that the same treatment can have different effects on different people. General Form.

1.2.6 Using Dual Machine Learning (DML) for Causal Inference

Causal inference was proposed to create interpretable, robust, and powerful machine learning models. Zhao et al. (2023) using A double machine learning analysis of green finance influence Exploring the dynamics of urban energy efficiency in China. Hybrid machine learning model using CatBoost and XGBoost methods for enhanced short-term load forecasting (Fuhr et al., 2024) proposed using dual machine learning to estimate causal relationships for method evaluations. Its core approach is to measure

cause-effect relationships. It is ubiquitous in decision-making problems in various fields such as healthcare and economics. A machine learning approach for causal inference that combines machine learning models with dual estimation techniques from economics to reduce bias and improve the accuracy of estimates

1.2.7 Use GBDT's Gradient Boosting Tree XGBoost, LightGBM, CatBoost in Boost Algorithm for Model Evaluation

Four gradient boosting tree models, including GBDT, XGBoost, LightGBM, and CatBoost, are combined with causal analysis machine learning to evaluate the MSE, RMSE, and MAE of each model.

1.3 Importance of Research

This research is vital for exploring the complex relationship between socioeconomic factors and marriage trends. By combining traditional statistical methods and advanced machine learning models, policymakers can obtain a clearer picture of how these factors influence marriage rates, leading to better-informed policies that support societal well-being and economic stability. The insights gained from this research are crucial not just for predicting future trends, but for actively shaping policies that improve family formation and social cohesion in the face of changing demographics.

1.3.1 Understanding the Impact of Socioeconomic Changes on Marriage Trends

Understanding the Impact of Socioeconomic Changes on Marriage Trends This research is crucial for understanding how fluctuations in key economic and demographic variables such as GDP, House Prices, and Gross Dependency Ratio influence marriage rates. By using traditional statistical models (Pooled OLS, Random Effects, Fixed Effects) alongside modern machine learning techniques, it becomes possible to quantify how economic stability, cost of living, and societal dependency affect people's decisions about marriage. Policymakers and social planners can use this

insight to develop informed policies that can address the declining marriage trends observed in many countries.

1.3.2 Evaluating the Predictive Power of Traditional and Machine Learning Models

Comparing traditional statistical approaches with machine learning algorithms allows for a deeper understanding of their predictive capabilities. By examining how well models like XGBoost, LightGBM, CatBoost, and GBDT predict marriage rates, researchers can determine which techniques best capture complex relationships between variables. Given that these machine learning models can handle non-linear relationships and interactions more effectively, they may provide more accurate forecasts, particularly when predicting future trends such as the marriage rate in 2022.

1.3.3 Implications for Long-term Social and Economic Policy

The outcomes of this research have direct implications for the development of social policies. Marriage rates can be a significant indicator of broader societal trends, including fertility rates and family stability. By studying how factors like education, gender distribution, and economic conditions affect marriage rates, governments can create long-term strategies for managing population growth and addressing potential social imbalances (e.g., aging populations, declining birth rates). Accurate predictions of marriage rates can inform policies that encourage family formation and societal stability.

1.3.4 Assessing the Role of Education and Gender Dynamics

This study also helps to uncover the deeper connections between Average Years of Education, Sex Ratio, and Female Population with marriage rates. In modern societies, higher education levels often delay marriage, while gender imbalances (more men or women in the population) can also affect marriage trends. Understanding these effects through a combination of traditional and machine learning models allows for a more comprehensive analysis, guiding educational and gender equality policies that could help balance societal norms.

1.3.5 Improving Policy Accuracy Through Causal Inference

The application of causal inference methods like ATE, CATE, and HTE enables a more precise understanding of the cause-and-effect relationships between socioeconomic variables and marriage rates. This focus on causal relationships allows researchers to evaluate the direct and conditional effects of variables such as GDP, birth rate, and education on marriage. It enhances the policy-making process by ensuring that interventions targeting marriage rates are based on solid evidence of their potential effectiveness, reducing the risk of unintended consequences from policy changes.

1.4 Research Hypothesis

The research hypothesizes several relationships between socioeconomic factors and marriage rates in China. Rising housing prices are expected to negatively impact marriage rates due to financial burdens, while higher GDP may lead to career-focused individuals, lowering marriage rates. Gender imbalances, particularly a higher male-to-female ratio, could also reduce marriage rates. Education, dependency ratio, and consumption levels are additional factors considered. The study also posits that socioeconomic factors affect marriage rates differently across regions, with urban-rural divides highlighting distinct trends in marriage patterns due to varying economic and social conditions.

1.4.1 Research Hypotheses Related to Economic Factors

Hypothesis: There is a significant negative relationship between housing prices and marriage rates across Chinese provinces. **Rationale:** Rising property prices, especially in urban areas, may discourage or delay marriage due to the financial burden of homeownership.

1.4.1.1 Housing Prices and Marriage Rates

Hypothesis: There is a significant negative relationship between housing prices and marriage rates across Chinese provinces. **Rationale:** Rising property prices, especially in urban areas, may discourage or delay marriage due to the financial burden of homeownership.

1.4.1.2 GDP Factors

Hypothesis: Higher GDP and per capita income levels are associated with lower marriage rates in Chinese provinces. Rationale: Economic development may lead to changing priorities, with individuals focusing more on career development or having higher expectations for marriage.

1.4.1.3 Gender Ratio

Hypothesis: The gender ratio (particularly the ratio of males to females) has a significant impact on marriage rates, with a higher ratio of males to females associated with lower marriage rates. Rationale: Gender imbalances may result in difficulties for some individuals to find partners, affecting overall marriage rates.

1.4.1.4 Education Levels

Hypothesis: Higher GDP and per capita income levels are associated with lower marriage rates in Chinese provinces. Rationale: Economic development may lead to changing priorities, with individuals focusing more on career development or having higher expectations for marriage.

1.4.1.5 Dependency Ratio

Hypothesis: Higher GDP and per capita income levels are associated with lower marriage rates in Chinese provinces. Rationale: Economic development may lead to changing priorities, with individuals focusing more on career development or having higher expectations for marriage.

1.4.1.6 Consumption Levels

Hypothesis: There is a significant negative relationship between housing prices and marriage rates across Chinese provinces. Rationale: Rising property prices, especially in urban areas, may discourage or delay marriage due to the financial burden of homeownership.

1.4.2 Hypothesis on the Influence of Region and Urban-Rural Area on Marriage Rate

1.4.2.1 Regional Differences

Hypothesis: Socioeconomic factors affect marriage rates differently across various regions of China, with more developed urban areas showing different trends compared to rural areas. Rationale: Economic development, cultural traditions, and

social environments vary across regions, potentially leading to different marriage patterns.

1.4.2.2 Urban-Rural Divide

Hypothesis: There is a significant negative relationship between housing prices and marriage rates across Chinese provinces. Rationale: Rising property prices, especially in urban areas, may discourage or delay marriage due to the financial burden of homeownership.

1.5 Scope of Research

This study investigates marriage rates in 31 provincial-level units in mainland China from 2003 to 2022, analyzing the impact of socioeconomic factors like GDP, housing prices, birth rate, education, sex ratio, and more. The research employs both quantitative and qualitative methods. Quantitative approaches include multiple linear regression, panel data regression (Fixed and Random Effects), machine learning models (XGBoost, LightGBM, CatBoost, GBDT), and spatial econometrics. Qualitative methods involve interviews and content analysis of social media and news. The study also applies causal inference techniques, such as Average Treatment Effect (ATE) and Dual Machine Learning (DML), ensuring robust predictions through 10-fold cross-validation using MSE, RMSE, MAE, and R-squared metrics.

1.5.1 Geographical Scope

The study covers 31 provincial-level units in mainland China and includes provinces, municipalities, and autonomous regions to capture regional diversity.

1.5.2 Temporal Scope

Analysis of data from 2003 to 2022, a 20-year period and captures a significant period of China's rapid economic development and social change

1.5.3 Variables Under Investigation

Dependent Variable: Marriage rate. Independent Variables: Provincial GDP, Housing prices, Gross dependency ratio, Birth rate, Female population, Average years of education per capita, Sex ratio, Per capita consumption, Betrothal gift practices

(where data is available), Urban-rural population distribution, work pressure indicators (e.g., average working hours)

1.5.4 Methodological Scope

This part of the research applies 10-fold cross-validation on machine learning models to assess the robustness of predictions. Metrics such as MSE, RMSE, MAE, and R-squared are used to compare the performance of XGBoost, LightGBM, CatBoost, and GBDT. This validation process ensures the models' reliability for real-world forecasting.

1.5.4.1 Quantitative Methods

Multiple Linear Regression, Panel Data Regression (Fixed Effects, Random Effects), Time Series Analysis, Spatial Econometrics, Machine Learning Techniques (XGBoost, LightGBM, CatBoost, GBDT).

1.5.4.2 Qualitative Methods

In-depth interviews with unmarried and married individuals across different age groups and regions. Content analysis of social media discussions and news reports on marriage.

1.5.4.3 Causal Inference Methods

Average Treatment Effect (ATE), Conditional Average Treatment Effect (CATE), Heterogeneous Treatment Effects (HTE), Dual Machine Learning (DML).

1.6 Research Limitations

When employing machine learning techniques to study the relationship between socioeconomic factors and China's crude marriage rate, several limitations emerge. These limitations arise from the inherent characteristics of data, methodology, and machine learning models. Below are five key areas where research limitations can be identified:

1.6.1 Data Availability and Quality

One of the significant limitations in this research is the availability and quality of the dataset. The research relies on historical data spanning from 2002 to 2022, which

may not be consistent across all regions of China. Issues such as missing data, reporting inconsistencies, and variations in measurement techniques can affect the accuracy of both traditional and machine learning models. Moreover, socioeconomic data may not capture unobservable factors like cultural attitudes or social policies that influence marriage rates.

1.6.2 Model Assumptions in Traditional Statistical Approaches

Traditional methods like Pooled OLS, Fixed Effects, and Random Effects rely on specific assumptions, such as homoscedasticity, no multicollinearity, and independence of errors. Violation of these assumptions can lead to biased estimates of coefficients, which may affect the interpretation of how socioeconomic factors influence marriage rates. For instance, the assumption of constant error variance may not hold in complex, real-world data, introducing inaccuracies in the results.

1.6.3 Interpretability of Machine Learning Models

Machine learning models such as XGBoost, LightGBM, CatBoost, and GBDT offer superior predictive power, but they often lack interpretability compared to traditional models. While these models can predict outcomes with high accuracy, understanding the exact contribution of each variable to the crude marriage rate becomes challenging due to their “black-box” nature. This lack of transparency makes it difficult to extract clear, actionable insights for policy-making based solely on machine learning results.

1.6.4 Generalization to Future Trends

The predictive models are trained on data from 2002-2021, and their performance is tested on 2022 data. While machine learning models are effective for short-term predictions, their ability to generalize to future trends beyond the training period is limited. Socioeconomic factors influencing marriage rates are subject to policy shifts, economic crises, and unforeseen events (e.g., the COVID-19 pandemic), which machine learning models may not be able to anticipate. This limits their long-term forecasting reliability.

1.6.5 Limitations in Causal Inference Techniques

Although causal inference techniques like double machine learning help in estimating the Average Treatment Effect (ATE), Conditional Average Treatment Effect (CATE), and Heterogeneous Treatment Effect (HTE), these methods still face challenges in identifying true causality. Causal inference is highly dependent on the choice of covariates and model specification, and omitting key variables or introducing measurement errors can lead to incorrect estimates. Additionally, the causal interpretation of machine learning results is more complex compared to traditional econometric models, and it is difficult to fully eliminate biases and confounding fact.



CHAPTER 2

LITERATURE REVIEW

2.1 Theoretical Reviews

In this section, Chen and Wen (2017) reported that over the last twenty years, China's real estate sector has seen significant growth. Major cities like Beijing, Shanghai, Shenzhen, and Guangzhou experienced an average annual increase in housing prices of 24%. In the top 35 cities, housing prices surpass the national average, with an annual growth rate of 17%. By the end of 2021, China's elderly population aged 65 and over had reached 212 million, leading to an old-age dependency ratio of over 20% for the first time in recent years. The total dependency ratio of China's population has been rising for four consecutive years, reaching 46.44%. The old-age dependency ratio is 20.82%, showing a continuous upward trend with an increase of 1.08 percentage points compared to 2020. It is projected that the total fertility rate in China will range between 1.6 and 1.8 in the future. In a low scenario where the total fertility rate remains at 1.3, the population of China will decrease to 620 million by the end of this century. The study explores the link between the marriage market and education, showcasing the impact of girls' education on marriage-related assets such as dowry and bridal gifts. Women with higher education levels receive greater marriage-related assets, which enhances their bargaining power in marriage (Khan, 2024). In China, it typically takes at least 15 years for an individual to complete education from elementary school to university. A bachelor's degree usually takes 16-17 years, a master's degree 18-19 years, and a doctoral degree 22-23 years. The male population is 723.34 million, accounting for 51.24% of the total population, while the female population is 688.44 million, accounting for 48.76%. The sex ratio of the total population is 105.07, slightly lower than in 2010, and the sex ratio at birth is 111.3,

a decrease of 6.8 from 2010. These factors collectively contribute to the changing dynamics of marriage in China, influencing the decline in marriage rates.

Zhao et al. (2023) examined the effects of escalating housing costs on marriage delays in China. They employed the Difference-in-Differences (DID) methodology to assess how increasing prices influence marriage timing. Their findings revealed that rising costs substantially elevate the financial burdens of entering into marriage. Moreover, the delay in marriage due to cost escalation is more noticeable in individuals with advanced female education, more brothers among males, and those from urban areas. This delay in marriage due to cost surge also decreases the inclination for childbirth, resulting in lower fertility rates. Chiplunkar and Weaver (2023) studied Marriage markets and the rise of dowry in India, and they found that between 1930 and 1975, the proportion of Indian marriages involving dowry payments doubled, with the average actual value of payments tripling. Guggenberger (2009) explores the scale properties of two-stage tests within panel data models. The first stage utilizes a Hausman specification test to assess the random effects specification. The second stage applies a test statistic based on either random effects or fixed effects estimates, contingent on the outcome of the Hausman pretest. Pilar Alonso et al. (2024) conducts a study on financial exclusion, depopulation, and aging using panel data regression models to analyze the influence of social demographic characteristics on financial exclusion. Sharma et al. (2023) researchers study House Price Prediction with Machine Learning Algorithms, emphasizing precise prediction using Python libraries like matplotlib, pandas, and NumPy. The widely used Python library for machine learning, scikit-learn, is open-source. Gupta et al. (2022) developed an efficient method for least squares estimation in dynamic space-time panel data models. This method, called eigendecomposition-based bias-corrected least squares procedure, uses eigen decomposition of the weight matrix in dynamic space-time pooled panel data models. Ratnasari et al. (2023) proposed a statistical model to analyze factors influencing the middle-income trap in Indonesia through panel data regression, using observations at the provincial level based on inter-regional decomposed variables.

2.2 Related Studies

The Related Studies chapter mainly introduces The Cultural Customs of Traditional Dowry in China, Tencent Gu Yu Data Released a Survey on the National Bride Price, Housing Prices and Land Transaction Policies, Elderly Care, Aging Rate, Housing Price Growth, and Education Complex System Issues Coexist, Gender Ratio and Male Population.

2.2.1 The Cultural Customs of Traditional Dowry in China

In China, while dowry is the primary form of marital payment, the exchange of dowry between both parties is also common. These components involve the bride, her parents, and the groom. The bride receives a dowry, also known as bridal price, from her husband. Married women typically receive similar dowries to first-time brides, but economists seldom conduct a comprehensive analysis of dowry. Despite the extensive literature on dowry and bridal prices, multiple costs are often paid simultaneously in marriage, such as in the case of the Han ethnic group, including dowry, a car, three types of gold jewelry, and a set of residential properties. It is challenging to address this complexity. Data from Senegal shows that about 85% of marriages involve transfers to the bride's family. Additionally, despite being overlooked in the literature, these marriages involve other marital payments flowing in different directions among the stakeholders. As shown in Figure 2.2, shows that the Betrothal Gift Required for Engagement (Excluding House and Car).



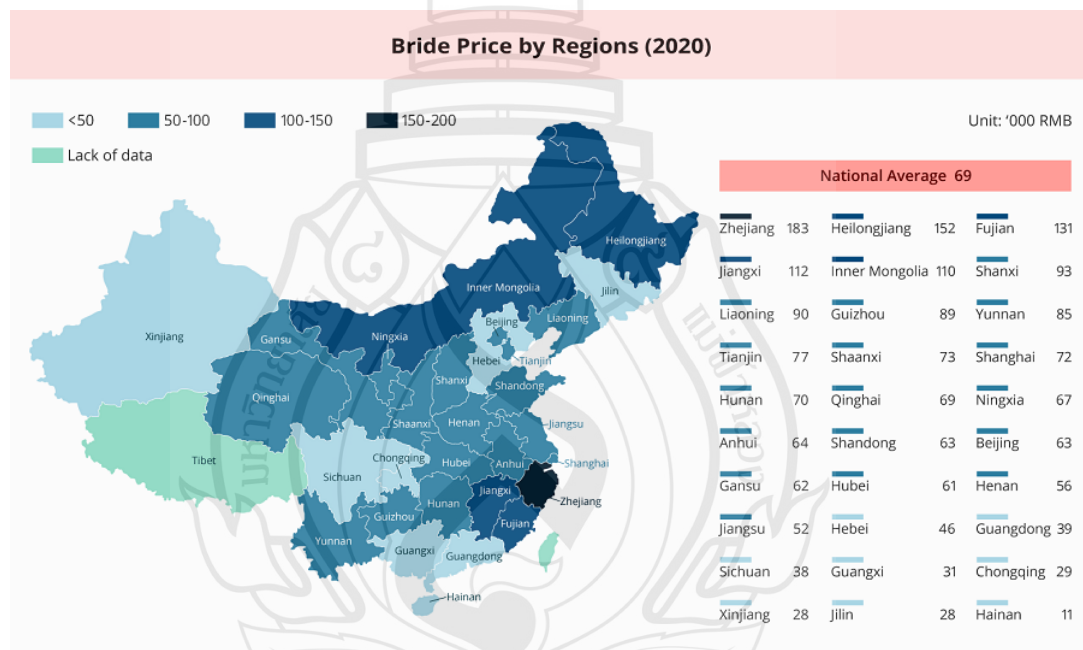
Source Little Red Book Application

Figure 2.1 The Betrothal Gift Required for Engagement (Excluding House and Car)

2.2.2 Tencent Gu Yu Data Published a Survey Regarding the National Bride Price

As shown in Figure 2.2, shows that the Tencent's Guyu Data Released a Survey on the Country's Bride Price Situation in 2020 Based on the Responses of 1,846 in September 2020, Tencent Gu Yu Data conducted a survey examining the national bride price landscape, gathering insights from 1,846 Chinese participants. Zhejiang emerged as the leader, with an average bride price of 183,000 RMB, representing 377% of the national average of 69,095 RMB. Other prominent regions included Heilongjiang at 152,000 RMB, Fujian at 131,000 RMB, Jiangxi at 112,000 RMB, and Inner Mongolia at 110,000 RMB. As shown in Figure 2.2, shows that the bride price map circulating on Chinese social media platforms, ranking regions based on average bride price in 2022. moreover, over 70% of the grooms provided jewelry, such as the traditional "three gold and four silver," comprising a gold necklace, a pair of gold earrings, a gold ring, a silver bowl, a pair of silver chopsticks, a silver hairpin, and a silver bracelet.

Nearly 40% of bride prices also included automobiles and real estate. The bride prices substantially surpass the preliminary data released by Tencent, which primarily focused on this aspect while omitting regional housing prices. Consequently, we will reference Tencent Gu Yu Data's findings based on the input from 1,846 individuals. As shown in Figure 2.2, shows that the amount of bridge prices in 7 economic zones. The provinces divided into the eastern economic zone and the northern economic zone generally have higher bride prices, while the southern, central, southwestern and northwestern economic zones are at lower levels relative to the eastern and northern regions, with the lowest being in the southern region.



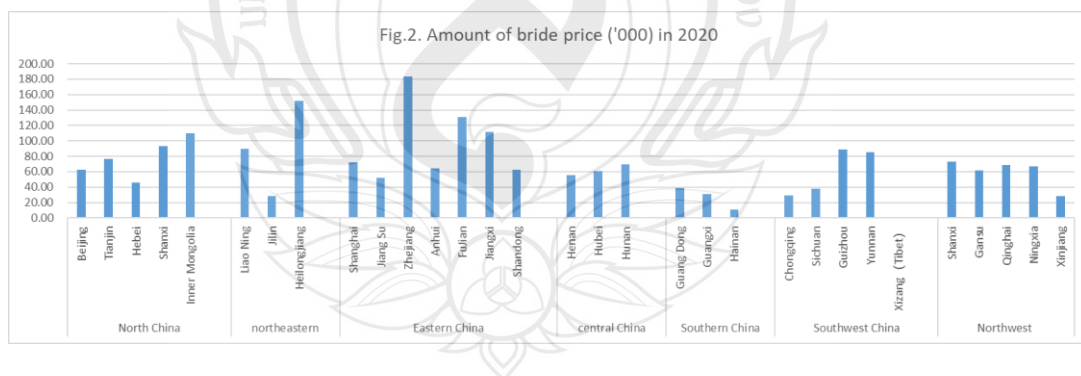
Source Desk and Zaobao (2023)

Figure 2.2 Tencent's Guyu Data Released a Survey on the Country's Bride Price Situation in 2020 Based on the Responses of 1,846 Chinese Residents



Source Desk and Zaobao (2023)

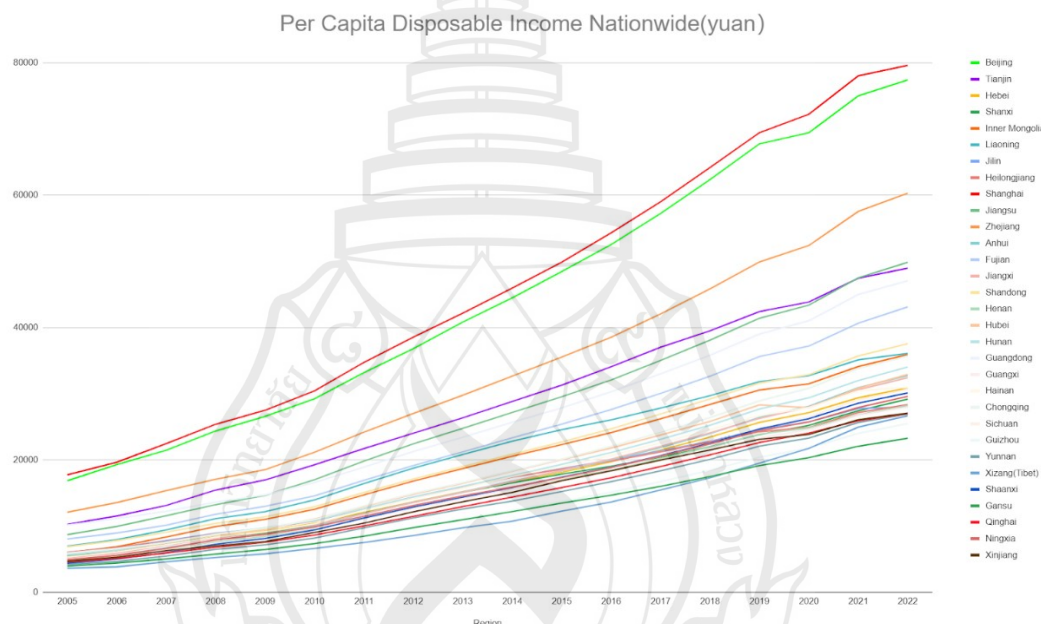
Figure 2.3 A Bride Price Map Circulating on Chinese Social Media Platforms, Ranking Regions Based on Average Bride Price in 2022



Source Shanghai Survey Team of National Bureau of Statistics (2009)

Figure 2.4 Amount of Bridge Prices in 7 Economic Zones

As shown in Figure 2.5, shows that the trend of per capita disposable income of national residents. According to the data analysis, China's per capita GDP in 2022 is 85,698 yuan (12,700 US dollars), and the per capita Gross National Income (GNI) is 84,804 yuan. It is approaching the threshold of high-income countries, but there is still potential to reach the threshold of developed economies with a per capita GDP of over 30,000 US dollars (Statistical Communiqué of the People's Republic of China on the 2022 National Economic and Social Development).



Source Shanghai Survey Team of National Bureau of Statistics (2009)

Figure 2.5 Trend of Per Capita Disposable Income of National Residents (Unit: yuan)

2.2.3 Housing Prices and Land Transaction Policies

As shown in Figure 2.6, shows that the Chinese Wedding Room. To explore the causal relationship between housing prices and marriage rates, we conducted a quasi-natural experiment based on the Urban Land Trading Policy (ULTP) enacted in 2002. Our aim was to uncover the causal link between increasing housing prices and delayed marriages among the youth. In China, housing prices strongly correlate with land

policies (Deng et al., 2012; Ding, 2020; Tian et al., 2020; Wang et al., 2021). Since the introduction of market reforms in 2002, the Chinese government allowed flexible mechanisms for urban land acquisition, enabling businesses to procure land through competitive bidding and auctions. This policy led to a notable rise in housing and land prices, particularly in urban centers, while rural areas saw minimal effects (Hu et al., 2021; Wang et al., 2017). This context allowed us to utilize the Difference-in-Differences (DID) methodology to analyze how housing prices influence the age of first marriage across China. Our findings indicated that the age discrepancy at first marriage between urban and rural youth began to grow around 2002. Our dataset spans 20 years, from 2003 to 2022, across all Chinese provinces. Overall, we found a positive correlation between housing prices, with the most pronounced increases occurring in eastern regions and major cities like Beijing, Shanghai, and Guangdong. Other regions exhibited similar trends subsequently (Zhao et al., 2023).



Source Little Red Book Application

Figure 2.6 Chinese Wedding Room

2.2.4 Income Survey of Families, Individuals and Recent Graduates

As shown in Figure 2.7, shows that the Comparison of Individual Characteristics of Education by Monthly Income Range. the most recent survey by the Chinese Household Income Project (CHIP) was conducted in 2018. The research institute's official website did not disclose the relevant data, but researchers at the institute revealed some data results in an article published on Caixin. Stratified sampling of 70,000 samples showed that the proportion of households with a per capita disposable monthly income (income available for actual use after deducting personal income tax, etc.) of over 10,000 yuan was only 0.61%, while the proportion of households with a per capita disposable monthly income in the 5000-10,000 (yuan) range was 4.52%. The majority of Chinese households had a per capita disposable monthly income in the range of 500-1500 (yuan), accounting for approximately 40.71%. Another set of data showed that only 4.3% of undergraduate fresh graduates had a monthly income (pre-tax, including wages, bonuses, allowances, etc.) of over 10,000 yuan, while 68.1% of undergraduate graduates had a monthly income of 6,000 yuan or less. In 2022, the per capita disposable income for residents reached 32,189 yuan, indicating a 4.7% increase from the prior year. After adjusting for inflation, the real growth amounted to 2.1%. The median disposable income per capita was 27,540 yuan, reflecting a rise of 3.8%. According to census data, urban residents reported an average disposable income of 43,834 yuan, reflecting a year-over-year increase of 3.5%. When adjusted for inflation, the real growth rate was 1.2%. The median disposable income for urban dwellers stood at 40,378 yuan, showing a 2.9% increase. In contrast, rural residents reported a per capita disposable income of 17,131 yuan, marking a notable increase of 6.9% from the previous year. After accounting for inflation, the real increase was 3.8%. The median disposable income for rural residents reached 15,204 yuan, indicating a 5.7% rise. the urban-rural disposable income ratio stood at 2.56, showing a decrease of 0.08 compared to the previous year. According to data on national income distribution across five demographic cohorts, per capita disposable incomes were recorded as follows: 7,869 yuan for the low-income cohort, 16,443 yuan for the lower-middle-income cohort, 26,249 yuan for the middle-income cohort, 41,172 yuan for the upper-middle-income cohort, and 80,294 yuan for the high-income cohort. Additionally, the average monthly earnings for rural migrant workers

nationwide stood at 4,072 yuan, reflecting a 2.8% rise from the previous year. As shown in Figure 2.8, shows that shows the distribution of individuals with a bachelor's degree or higher across different monthly income ranges. The data reveals that 51.8% of individuals have a monthly income over 5000 yuan, indicating that most people with higher education have higher incomes. 28.2% fall within the 2000 - 5000 Yuan range, representing a moderate incomes level. Only 13.5% have a monthly income between 1090 and 2000 yuan, and 6.5% earn less than 1090 yuan. This suggests a positive correlation between education and income, although a small portion of highly educated individuals have lower incomes, reflecting diversity in income distribution.

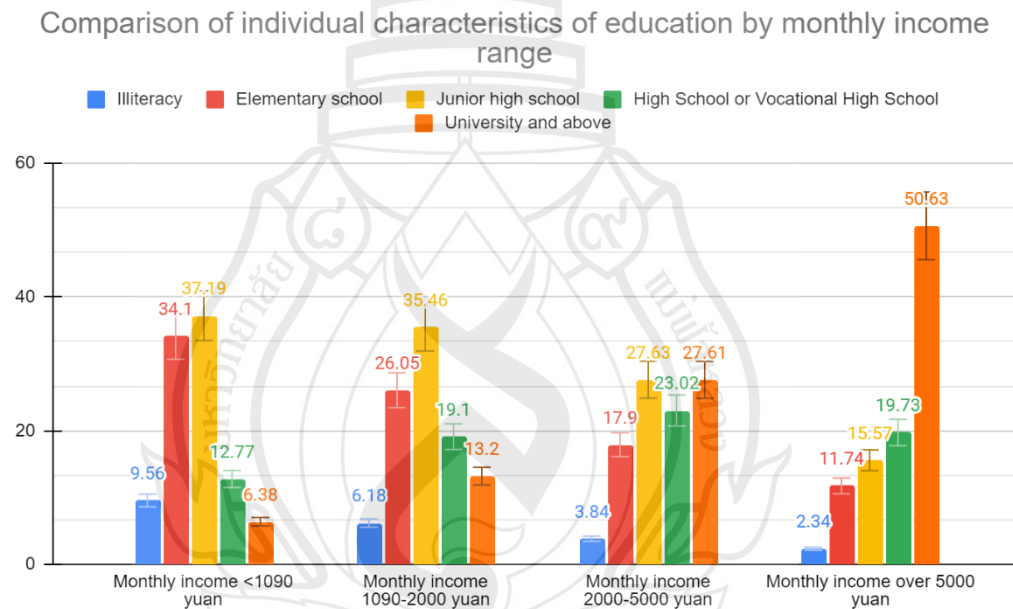


Figure 2.7 Comparison of Individual Characteristics of Education by Monthly Income Range (From the Most Recent Survey by the Chinese Household Income Project (CHIP) was Conducted in 2018)

Comparison of Characteristics of Individuals with Bachelor's Degree or Above in Monthly Income Range

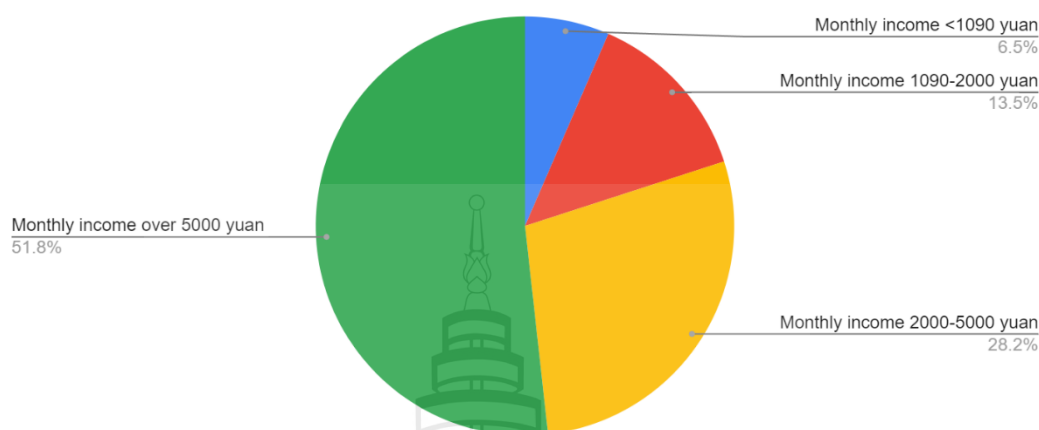


Figure 2.8 Comparison of Characteristics of Individuals with Bachelor's Degree or Above in Monthly Income Range (From the Most Recent Survey by the Chinese Household Income Project (CHIP) was Conducted in 2018)

2.2.5 Elderly Care, Aging Rate, Housing Price Growth, and Education Complex System Issues Coexist

As shown in Figure 2.9, shows that the national population by sex, educational level and age at first marriage. The old-age dependency ratio compares the population aged 65 and above to those in the working-age category of 15 to 64. by the end of 2021, China's elderly population aged 65 and over had reached 212 million, leading to an old-age dependency ratio of over 20% for the first time in recent years, compared to the consistent level below 10% prior to 2000. The total dependency ratio of China's population has been rising for four consecutive years, reaching 46.44%. Specifically, the old-age dependency ratio is 20.82%, showing a continuous upward trend with an increase of 1.08 percentage points compared to 2020 (19.74%). This indicates that for every 100 working-age population, China needs to support nearly 21 elderly people. Additionally, it is projected that the total fertility rate in China will range between 1.6 and 1.8 in the future, leading to a decrease in the total population of China to 1.02 billion by the end of this century. In a low scenario where the total fertility rate remains at 1.3, the population of China will decrease to 620 million by the end of this century.

Over the last twenty years, China has experienced significant growth in its real estate sector. This trend is particularly evident in major cities like Beijing, Shanghai, Shenzhen, and Guangzhou, which have seen an average annual increase in housing prices of 24%. In the top 35 cities, housing prices surpass the national average, with an annual growth rate of 17% (Chen & Wen, 2017). The shift in family planning policies from the 1990s, coupled with the rapid expansion of the housing market, provides an opportunity to explore the causal relationship between marriage rates and individual intentions regarding child-rearing. In traditional Chinese culture, the issue of childbirth is only considered after marriage. Once married, young individuals are not only responsible for supporting their parents but also for raising children. The Chinese concept of “having elders above and young ones below” signifies that when married and with children, the advantage lies in the fact that parents can assist in child care, and their good health means they do not have to bear significant medical expenses if they become ill. The economic reforms since the 1970s have improved the living standards for a large portion of China’s population, although the elderly in rural areas have not experienced as much benefit and are under increasing pressure. In the absence of a rural pension insurance system, the primary economic support for elderly rural residents comes from their own labor income and family assistance, including financial support from adult children. Pension contributions play a relatively substantial role in supporting economically disadvantaged regions at the regional level, and these contributions exhibit urban-rural disparities, with the central region > western region > eastern region and rural areas > urban areas (Li et al., 2023).

As shown in Figure 2.10 proportion of population with educational at age of first marriage in Mainland China. As shown in Figure 2.11 shows that the Chinese education system. The decrease in the marriage rate in China is also connected to the educational level, duration, and age of men and women's education. In normal circumstances, it takes at least 15 years for a Chinese individual to complete education from elementary school to university. With a bachelor’s degree usually taking 4 years for regular majors and 5 years for special majors like clinical medicine and urban and rural planning, the completion of a bachelor’s education requires 16-17 years. A master’s degree (2 years for some taught master’s programs, 3 years for academic ones) takes 18-19 years, and a doctoral degree takes 22-23 years, while pursuing a direct

doctorate (bachelor's degree) takes at least 21 years. Furthermore, during the course of study, parents have already paid most of the students' expenses. Most young people face significant pressure after graduation due to debt and meager salaries. Another factor to consider in the context of the gender ratio imbalance is the adjustment when men delay marriage. The competition for marriage still has a first-order effect on savings. A theory of a frictionless marriage market and a quantitative life cycle model are constructed in the process, wherein factors such as wealth and age play a significant role in the ranking of unmarried men in making marriage decisions (Nie, 2020). From the perspective of students graduating from high school, the proportion of those who choose to pursue a research-oriented university (the 211, 985 Project) and embark on a research-oriented scientific research path, or choose to enter a vocational college from a secondary technical school, or work directly after graduating from junior high or high school is far higher in the marriage rate than those who choose to pursue a master's or doctoral degree after completing an undergraduate degree. From an age perspective, the marriage rate is much higher from 18 to 28 years old than any other period within these 10 years. Based on the data analysis, it has been demonstrated that there is a higher prevalence of marriage among individuals with primary and middle school education over the past decade, suggesting that some individuals were not of legal marriage age at the time of their marriage. The study also explores the link between the marriage market and education, showcasing the impact of girls' education on marriage-related assets such as dowry and bridal gifts. The study estimates the educational impact on the marriage market through indicators like dowry and bridal gifts. Women with higher education levels receive greater marriage-related assets, which enhances their bargaining power in marriage (Khan, 2024). If girls are able to pursue further education, will increase educational achievement result in greater benefits in the marriage market rather than in the labor market – in essence, will brides with higher education levels receive more marriage-related assets? In Pakistani marriage traditions, women receive two types of marriage-related assets (dowry and bridal gifts), and the relationship between education and these assets is uncertain. As the analysis of education and marriage-related assets does not address endogeneity, the findings of this section can only be considered correlational (Khan et al., 2020).

Population demographers and economists (Ryder, 1964; Bergstrom & Lam, 1991) have generally found that the age disparity between spouses serves as a strong equilibrium mechanism in the marriage market. Thus, making moderate adjustments to the marital age gap can offset significant imbalances in the marriage market without causing necessary changes in marriage rates for men and women. Prior investigations have explored the relationship between partners' educational attainment and reported levels of marital satisfaction. Yet, conclusive evidence has not emerged regarding the outcomes of these associations. Research involving a substantial cohort of 10,000 participants in the United States revealed that as educational qualifications rise, the percentage of those reporting marital satisfaction declines (Call & Heaton, 1997). Conversely, a study assessing a smaller group of Belgian individuals (N = 787) found no remarkable link between education and relationship satisfaction. Given these intricate research outcomes, the dynamic between education and marital satisfaction persists as a prominent debate, particularly from a cross-cultural standpoint.

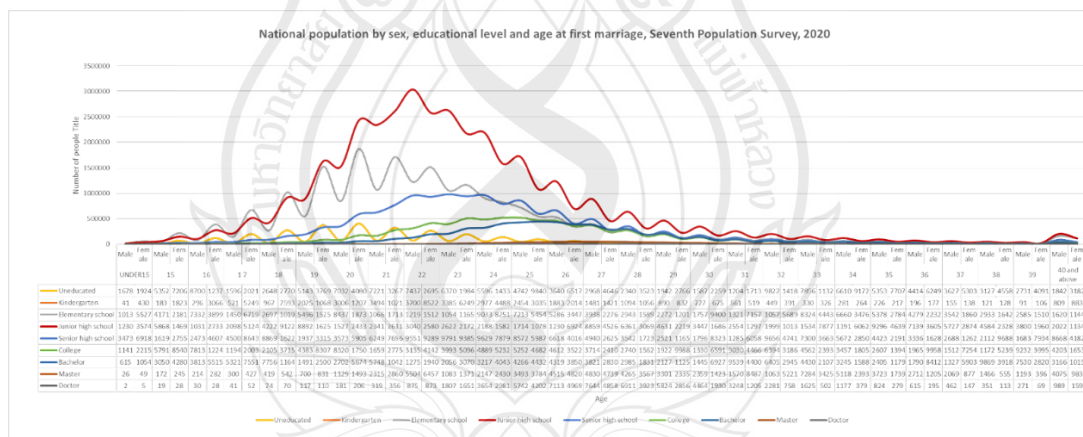
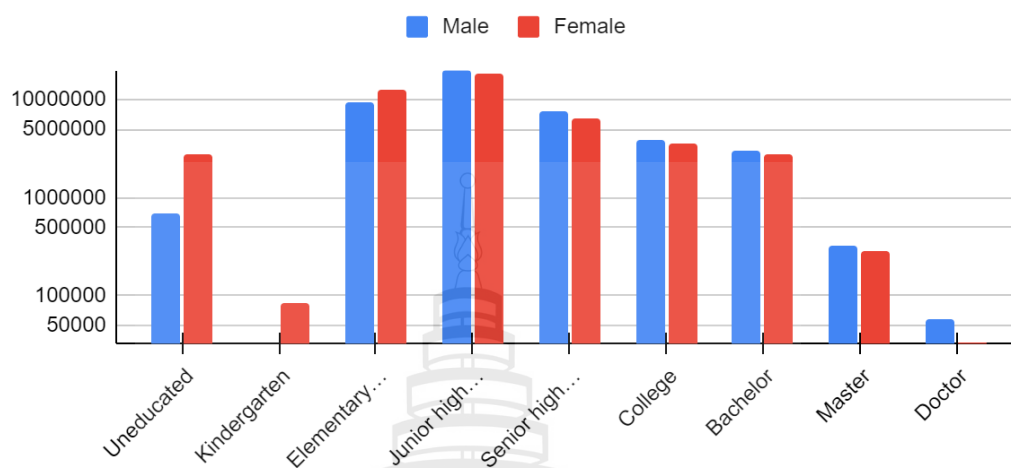


Figure 2.9 National Population by Sex, Educational Level and Age at First Marriage (The Seventh National Population Census in 2020)

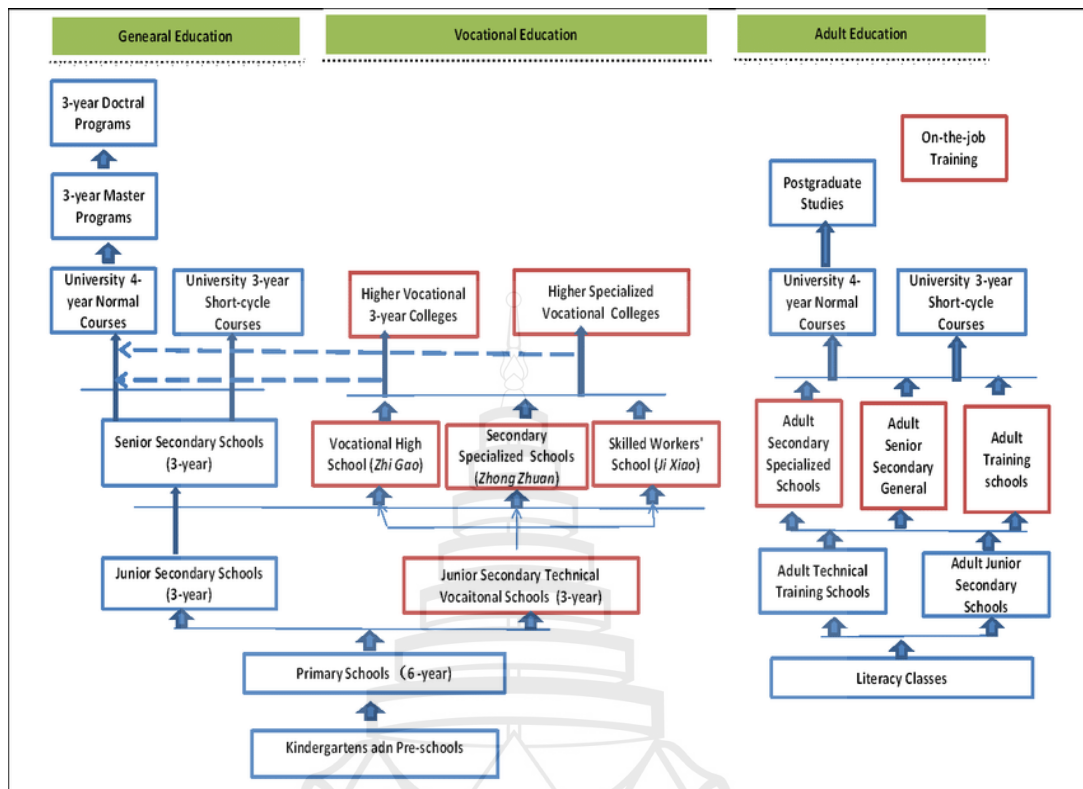
Proportion of population with education at age of first marriage in mainland China



Proportion of population with education at age of first marriage in mainland China

Source Shanghai Survey Team of National Bureau of Statistics (2009)

Figure 2.10 Proportion of Population with Educational at Age of First Marriage in Mainland China



Source Yang (2014)

Figure 2.11 Chinese Education System

2.2.6 Gender Ratio and Male Population

In recent years, there has been significant attention given to China's gender imbalance and the issue of male surplus. This imbalance has resulted in some individuals being unable to adhere to traditional Chinese cultural values when choosing their spouses or even being unable to get married at all. Since 2010, a considerable number of young men in China have struggled to find Chinese spouses, leading to a significant shortage of potential male partners for many years to come. Due to strict fertility policies, over 10% of males born after 1980 are unable to find spouses. The excess of males aged 20 to 49 will persistently rise, projected to hit 20 million by 2015, 30 million by 2025, and an alarming 40 million by 2040. As detailed in China's National Population Development Strategy Report, the population aged 20 to 45 will have 30 million more males than females. It is estimated that from 1983 to 2020, China

will produce at least 51 million surplus males. The male population is 723.34 million, accounting for 51.24% of the total population, while the female population is 688.44 million, accounting for 48.76%. The sex ratio of the total population is 105.07, slightly lower than in 2010, and the sex ratio at birth is 111.3, a decrease of 6.8 from 2010. The gender structure of China's population continues to improve. As shown in Figure 2.12 shows that the male-female ratio in each province.

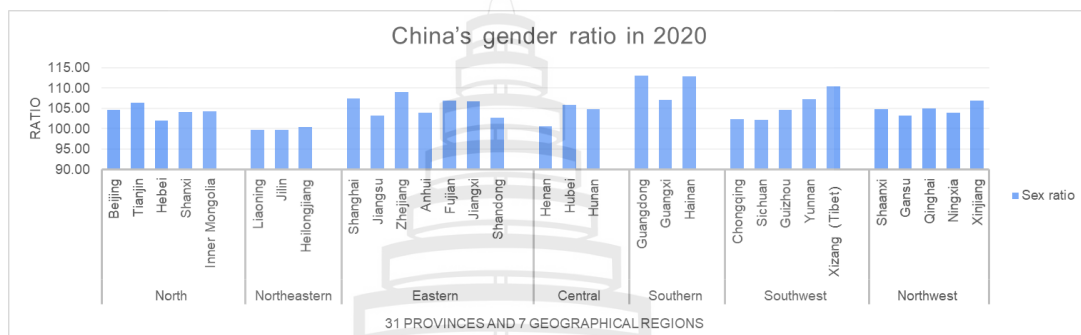


Figure 2.12 China's Gender Ratio in 2020

2.3 Related Theories of Term Definition

In this part of the study, the relevant theories of term definition include two aspects. First, the formula for the crude marriage rate (CMR). Second, the formula for the average years of education.

2.3.1 The Formula for The Crude Marriage Rate (CMR)

According to the United Nations, the crude marriage rate (CMR) is a vital statistics summary rate based on the number of marriages occurring in a population during a given period, usually a calendar year. It is calculated as the number of marriages occurring among the population of a given geographical area during a given year per 1,000 mid-term population of the same area during the same year. The formula for the crude marriage rate (CMR) is;

$$CMR = \frac{\text{Number of Marriage Registration Pairs}}{\text{Mid - Term Population}} * 1000$$

Where:

CMR = Crude marriage rate

Mid – Term Population = Arithmetic mean of the population on 1 January and the population on 31 December of a year. It is used to calculate annual rates

2.3.2 The Formula for The Average Years of Education

This study annually calculated the educational achievements of individuals aged 6 and above from 2003 to 2022, outlining the equation for average years of education per capita in this context. The formula for the average years of Education is;

$$\text{Average years of education} = \frac{(a*6+b*9+c*12+e*15+f*16+g*19)}{\text{Total Population Age 6 and Above}}$$

Where: a, b, c, d, e, f and g refer to the number of populations get Certificate

a = Primary School Certificate

b = Junior High School Certificate

c = Senior High School

d = Vocational Secondary School

e = College

f = Bachelor

g = Postgraduate (Master, PhD)

2.4 The Panel Data Regression Model of the Hausman-Test

Panel data is a combination of cross-sectional data and time series data. Cross-sectional data involves observing multiple entities' variables at a specific point in time, while time series data involves observing a single entity repeatedly over time. Panel data integrates both features by gathering data from identical subjects over time, similar to observing the same individuals at consistent intervals on a timeline.

2.4.1 Panel Data Regression

Panel data integrates the features of cross-sectional and time-series data. Cross-sectional data captures a single snapshot of multiple subjects and their variables at a specific

moment. In contrast, time-series data focuses on repeated measurements of a single subject over time. Panel data combines these approaches by collecting data from multiple identical subjects across different time points. As shown in Figure 2.14, shows that the concept of panel data structure. The panel data model contains both cross-section and time dimensions. Let i ($i = 1 \dots N$) represent the cross-section (individual), t ($t = 1 \dots N$) represent the time, and set the following model, the formula for the Panel Data Regression is;

$$y_{it} = a_i + \lambda_t + x_{it}\beta + \epsilon_{it}$$

Where:

a_i = Represents individual effects, representing factors that do not change over time

y_{it} = $N \times 1$ dependent variable

x_{it} = $N \times k$ independent variables

ϵ_{it} = Model error term

β = Parameters to be estimated, represents the marginal impact of x_{it} on y_{it}

λ_t = Represents the time effect, which is used to control the impact of factors that change over time (the time dummy variable includes the time trend term, which is mainly used to control technological progress)

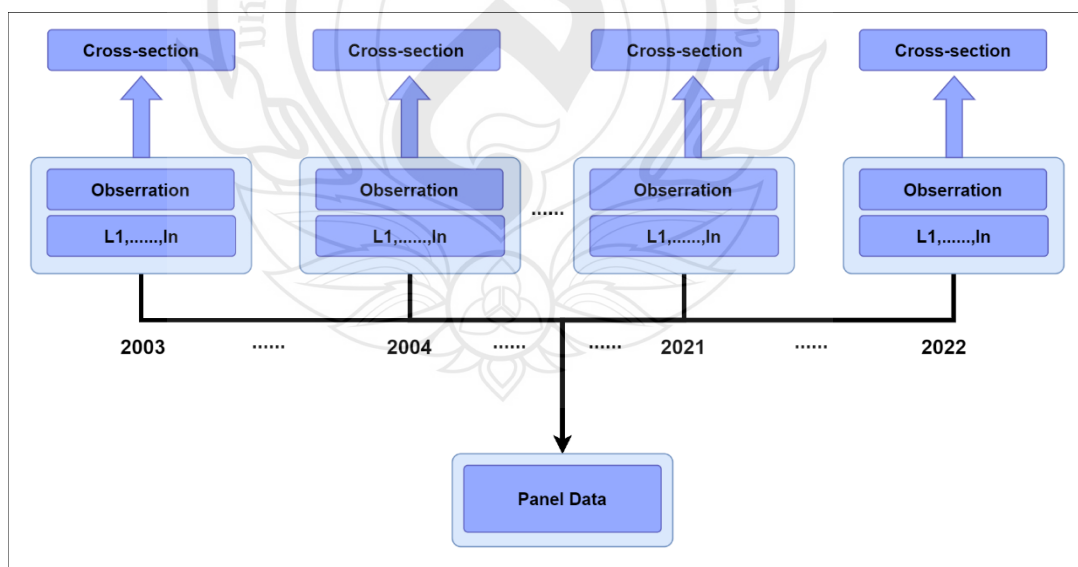


Figure 2.13 The Concept of Panel Data Structure

Obviously, a_i and λ_t cannot be directly observed or quantified in most cases, so they cannot be included in the model. This often leads to the problem of omitted variables in cross-sectional analysis. One of the main uses of panel data models is to deal with these unobservable individual effects or time effects. When a_i is equal for all i , the model degenerates into a mixed data model (Pooled OLS).

According to the size of the number of individuals N and the number of periods T , panel data can usually be divided into macro panels and micro panels: macro panels are generally “large T and small N ”, and micro panels are generally “small T and large N ”. Depending on the size of N and T , the parameter estimation method used and the focus of analysis are also different. The formula for the Panel Data Regression is;

$$y_{it} = x_{it}\beta + x_i + \alpha + u_{it}$$

Where:

$t = 1 \dots T$

$i = 1 \dots N$

y = Independent variable

$X = IV(s)$

β = Coefficients

α = Individual effects

u = Idiosyncratic error

2.4.2 Pooled OLS Model

The multiple pooled ordinary least squares (OLS) method presents itself as a basic OLS model applied to panel data. It overlooks both temporal and individual traits, concentrating solely on inter-individual dependencies. Nonetheless, basic OLS necessitates the absence of correlation—also termed exogeneity—between unobserved independent variables and instrumental variables (IVs).

2.4.3 Random-Effects (RE) Model

Random-effects (RE) models analyze the individual impacts of unobserved, independent variables treated as random variables over time. These models can toggle between ordinary least squares (OLS) and fixed effects (FE), allowing for an

examination of relationships both among individuals and within them. Breusch and Pagan (1980) introduced the Lagrange Multiplier (LM) statistic derived from the residuals of the panel random effect model. The formula for the null hypothesis for assessing the random effects is;

$$H_0 = \sigma_\alpha^2 = 0; H_1: \sigma_\alpha^2 \neq 0;$$

The corresponding test statistic LM is;

$$LM = \frac{NT}{2(T-1)} \left[\frac{\sum_{i=1}^N [\sum_{t=1}^T e_{it}]^2}{\sum_{i=1}^N [\sum_{t=1}^T e_{it}^2]} \right]$$

Under the null hypothesis, the statistic follows a chi-square distribution with 1 degree of freedom. Rejection of the null hypothesis indicates the presence of a random effect.

2.4.4 The Fixed Effects (Panel OLS) Model

The Fixed Effects Model, or Panel Data Fixed Effects Model, is a statistical tool for analyzing panel data, which combines cross-sectional and time series dimensions from observations of the same group over time.

The fundamental objective of the fixed effects test is to assess the significance of the differences in the intercept terms across individuals, thereby evaluating if $\alpha_1 = \alpha_2 = \dots = \alpha_N = 0$. This aligns with the principles of hypothesis testing. the following null hypothesis is set;

$$H_0 = \alpha_1 = \alpha_2 = \dots = \alpha_{N-1} = 0$$

If the outcome refutes the null hypothesis, it suggests substantial variations in the intercept terms among individuals. Hence, one must incorporate the fixed effect in the model. Conversely, if not, the mixed OLS model remains the more fitting choice. The F statistic can usually be used to test whether the above hypothesis is true;

$$F = \frac{(R_u^2 - R_r^2) / (N - 1)}{(1 - R_u^2) / (NT - N - K)} \sim F(N - 1, NT - N - K)$$

Where:

R_u^2 = Goodness-of-fit coefficient for the fixed effects model (unconstrained model)

Goodness-of-fit coefficient for mixed data models (constrained models)

R_r^2 = Goodness-of-fit coefficient for mixed data models (constrained models)

N = Section

T = Number of periods

K = Number of explanatory variables

If the null hypothesis is discarded, it indicates a significant individual effect, thereby suggesting that the fixed effect model is superior to the mixed data model. Likewise, a corresponding F statistic can be developed to evaluate the significance of the period effect.

2.4.5 The Fixed Effects or Random Effects

After testing to show that the individual effect (α_i) needs to be included in the model, should α_i be considered as part of the random interference term (random effects model) or as a parameter to be estimated.

The Hausmann test

The Hausmann test serves to differentiate between fixed effects models and random effects models in panel analysis. In this context, Guggenberger (2009) argues that the random effects model (RE) is favored under the null hypothesis owing to its higher efficiency, whereas the fixed effects model (FE) maintains at least comparable consistency under the alternative hypothesis.

According to the fundamental definition, we can utilize the correlation between individual effect α_i and other explanatory variables to inform the screening of fixed effect and random effect models. At this point, the Hausman test becomes applicable. The core idea is this: if we assume that α_i and other explanatory variables are uncorrelated, then the parameter estimates derived from the fixed effect model using within-group transformation and the random effect model using the GLS method will both be unbiased and consistent. Nevertheless, the former remains invalid if the null hypothesis holds. In such cases, the parameter estimates from the fixed effect model maintain consistency, while those from the random effect model do not. Thus, under

the null hypothesis, there should be no significant difference in the parameter estimates between the two models, allowing for a statistical test based on their differences.

Assume that β_{within} is the within-group estimator of the fixed effect model, and β_{GLS} is the GLS estimator of the random effect model. Under the null hypothesis, we have:

$$cov = (\beta_{GLS}, \beta_{GLS} - \beta_{within}) = 0$$

According to the variance formula;

$$var(\beta_{within} - \beta_{GLS}) = var(\beta_{within}) + var(\beta_{GLS}) - 2 cov(\beta_{within}, \beta_{GLS})$$

also because;

$$var(\beta_{GLS}) = cov(\beta_{within}, \beta_{GLS})$$

Therefore, there is;

$$var(\beta_{within} - \beta_{GLS}) = var(\beta_{within}) - var(\beta_{GLS}) = \psi$$

The Hausman test is based on the following Wald statistic;

$$W = (\beta_{within} - \beta_{GLS})' \psi^{-1} (\beta_{within} - \beta_{GLS}) - x^2(K - 1)$$

If the null hypothesis is rejected, it indicates that the individual effect α_i is related to the explanatory variable. At this time, the results of the random effect model are inconsistent, and the fixed effect model should be selected.

2.5 The Machine Learning Models

Panel data is a combination of cross-sectional data and time series data. Cross-sectional data involves observing multiple entities' variables at a specific point in time, while time series data involves observing a single entity repeatedly over time. Panel data integrates both features by gathering data from identical subjects over time, similar to observing the same individuals at consistent intervals on a timeline.

2.5.1 The XGBoost, LightGBM, CatBoost, GBDT

In essence, 'Boosting' refers to a methodology that integrates multiple base models into a singular composite model. By constructing additional simple base models, often termed weak models or weak learners, the resulting composite model enhances its predictive strength. Boosting trains these weak learners in a sequential manner, with each one refining the performance of its predecessor. The robust learner achieved through this process is known as an 'ensembled model' within Boosting Algorithms.

When integrating weak learners, one employs either the average or weighted average of the prior weak learner's error functions to refine subsequent learning iterations. Boosting emphasizes misclassified rules or rules yielding high errors by adjusting weaker rules. An increase in weights highlights data points that were misclassified by earlier weak models. During the evaluation phase, each model's performance is assessed according to the test error from each weak model, and predictions are combined using a weighted voting mechanism. Boosting techniques effectively reduce prediction bias. As shown in Figure 2.9, shows that the processing history of XGBoost, LightGBM, and CatBoost.

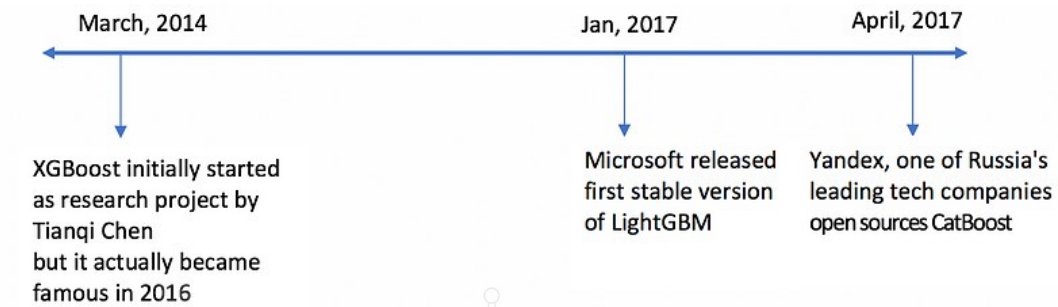


Figure 2.14 The Processing History of XGBoost, LightGBM and Catboost

2.5.1.1 GBDT (Gradient Boosting Decision Trees)

GBDT is an ensemble model that constructs a series of decision trees by iteratively fitting the residuals (errors) from the previous iteration. Each tree tries to reduce the errors from the previous model, thus improving the final prediction. The formula for the GBDT is;

$$\operatorname{argmin}_h \sum_{i=1}^n L(y_i, F_m(x_i) + h(x_i))$$

Where:

L = The loss function

y_i = the actual value

$F_m(x_i)$ is the predicted value of the mth tree

2.5.1.2 XGBoost (Extreme Gradient Boosting)

XGBoost is an extension of GBDT that optimizes GBDT using faster computation, parallel processing, and regularization techniques, (Chen & Guestrin, 2016) designed to enhance model performance and scalability. The formula for the XGBoost is;

$$L(\theta) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k)$$

Where:

$L(\theta)$ is objective function to be minimized during model training. It combines the prediction loss (accuracy of predictions) and a regularization term (to avoid overfitting).

2.5.1.3 LightGBM (Light Gradient Boosting Machine)

LightGBM is an improved version of GBDT designed to handle large-scale datasets efficiently. Ke et al. (2017) uses a histogram-based decision tree algorithm and introduces a leaf-wise growth strategy, improving both training speed and model performance. The formula for the LightGBM is;

$$L(\theta) = \sum_i l(y_i, \hat{y}_i) + \lambda \cdot \Omega(T)$$

2.5.1.4 CatBoost

CatBoost is a GBDT algorithm particularly suited for handling categorical features. Ye et al. (2020) enhances GBDT with techniques to avoid data leakage and overfitting, making it highly effective for datasets with categorical variables. Prokhorenkova et al. (2018) designed to enhance model performance and scalability. The formula for the CatBoost is;

$$\hat{y}(x) = \sum_{t=1}^T \eta f_t(x)$$

Where:

$f_t(x)$ is the output of the t -th tree

η is the learning rate, controlling the contribution of each tree

2.5.2 Dual Machine Learning (DML) for Causal Inference

Causal inference was proposed to create interpretable, robust, and powerful machine learning models. Kumar et al. (2024) core approach is to measure cause-effect relationships. Cohrs et al. (2024) is ubiquitous in decision-making problems in various fields such as healthcare and economics. A machine learning approach for causal inference that combines machine learning models with dual estimation techniques from economics to reduce bias and improve the accuracy of estimates.

2.5.2.1 ATE (Average Treatment Effect)

The mean treatment effect, which measures the average effect of the treatment on the outcome variable in the population, reflects the population-wide causal effect. The formula for the ATE is;

$$ATE = E[Y(T = 1) - Y(T = 0)]$$

Where:

$Y(T = 1)|T = 1$ and $Y(T = 0)|T = 1$ are the potential treated and control outcomes of the treated group, respectively

ATT can also be called local average treatment effect (LATE)

2.5.2.2 CATE (Conditional Average Treatment Effect)

The conditional mean treatment effect, which measures the average effect of the treatment on the outcome variable under a particular condition, reflects differences in causal effects across subpopulations. The formula for the CATE is;

$$CATE = E[Y(T = 1)|X = x] - E[Y(T = 0)|X = x]$$

Where:

$Y(T = 1)$

$X = x$ and $Y(T = 0)$

X are the potential treated and control outcomes of the subgroup with $X = x$ respectively

CATE is also known as the heterogeneous treatment effect

2.5.2.3 HTE (Heterogeneous Treatment Effect)

HTE refers to the variation in the impact of a treatment across different individuals or groups. Huang et al. (2022) present a robust framework for causal learning aimed at estimating average treatment effects (ATE). Numerous decision-making challenges in economics and healthcare focus on accurately assessing ATE using observational data. In simpler terms, it means that the same treatment can have different effects on different people. General Form. The formula for the HTE is;

$$Y_i = \alpha + \beta T_i + \gamma X_i + \epsilon_i$$

Where;

Y_i is the outcome variable (e.g., treatment effect)

T_i is the treatment indicator variable (1 indicates treatment received, 0 indicates no treatment)

X_i represents individual characteristics (covariates), α , β , and γ are parameters to be estimated, ϵ_i is the error term

$$y_i = \alpha + \beta T_i + \delta(T_i * X_i) + \epsilon_i$$

Where;

δ represents the effect of the interaction between treatment effect and covariates, reflecting the heterogeneity of treatment effects.

2.5.3 The Panel Data Model

Panel data integrates the features of cross-sectional and time-series data. Cross-sectional data captures a single snapshot of multiple subjects and their variables at a specific moment. In contrast, time-series data focuses on repeated measurements of a single subject over time. Panel data combines these approaches by collecting data from multiple identical subjects across different time points. As shown in Figure 2.14, shows that the concept of panel data structure.

2.6 Reflection on Literature Review

The complexity of marriage economics in mainland China is influenced by various factors. One of the key contributing factors is the difference in economic scale across regions, which impacts individual circumstances. These differences lead to varying inputs and outputs, influencing the creation of wealth for both parties within their work areas. They are essential components of a region's wealth-producing capacity and significantly impact marriage rates and regional differences across the country. Additionally, these diverse factors play a crucial role in determining overall marriage rates and influencing aspects such as marriage, childbirth, and education. Therefore, it is essential to objectively research and address this issue, employing advanced machine

learning methods to provide a more comprehensive analysis. The literature review provides a comprehensive overview of the complex factors influencing marriage trends, particularly in China and India. It highlights the interplay between economic factors, cultural practices, demographic shifts, and individual choices. The use of diverse methodological approaches, from traditional statistical methods to advanced machine learning techniques, demonstrates the multifaceted nature of this research area. Moving forward, there's a need for more integrated approaches that consider the dynamic interplay of various factors affecting marriage decisions. Additionally, expanding the geographical scope of research and incorporating more diverse cultural perspectives could enrich our understanding of global marriage trends and their societal implications.



CHAPTER 3

RESEARCH METHODOLOGY

This section defines the marriage rate problem in China and outlines the methodology, including data collection, merging, and preprocessing. It employs exploratory data analysis (EDA), feature scaling, and normal distribution testing. Panel data regression models and machine learning methods are used, with evaluation indicators like MSE, RMSE, and MAE, ensuring data readiness for analysis.

3.1 Overall Methodology

In this section, we define the marriage rate problem in China, collect data, focus on independent variables, and perform quantitative analysis. Missing data is handled in the preprocessing step. We propose a hybrid model to predict marriage rate using panel data and causal inference Using Dual Machine Learning (DML) with XGBoost, LightGBM, CatBoost, and GBDT (Guggenberger, 2010; Ratnasari et al., 2023). By combining OLS, fixed effects, random effects, fixed effects models, and P. Hausman test, we finally evaluate model performance and predictions with RMSE and MAE. As shown in Figure 3.1 shows the Overall Methodology of our research paper, focusing on crude marriage rates using panel data analysis and Machine learning model evaluation and prediction. Zhao et al. (2024) using A double machine learning analysis of green finance influence Exploring the dynamics of urban energy efficiency in China. Hybrid machine learning model using CatBoost and XGBoost methods for enhanced short-term load forecasting. Fuhr and Berens (2024) proposed using dual machine learning to estimate causal relationships for method evaluations. We finally evaluate model performance and predictions with MSE, RMSE and MAE. We focus on the most critical characteristics that have a significant impact on the crude marriage rate, such as distance from GDP, house price, gross dependency ratio, Birth Rate,

Female, Average years of education, and Sex Ratio. As shown in Figure 3.1, shows the Overall Methodology of our research paper, focusing on crude marriage rates using panel data analysis and Machine learning model evaluation and prediction.

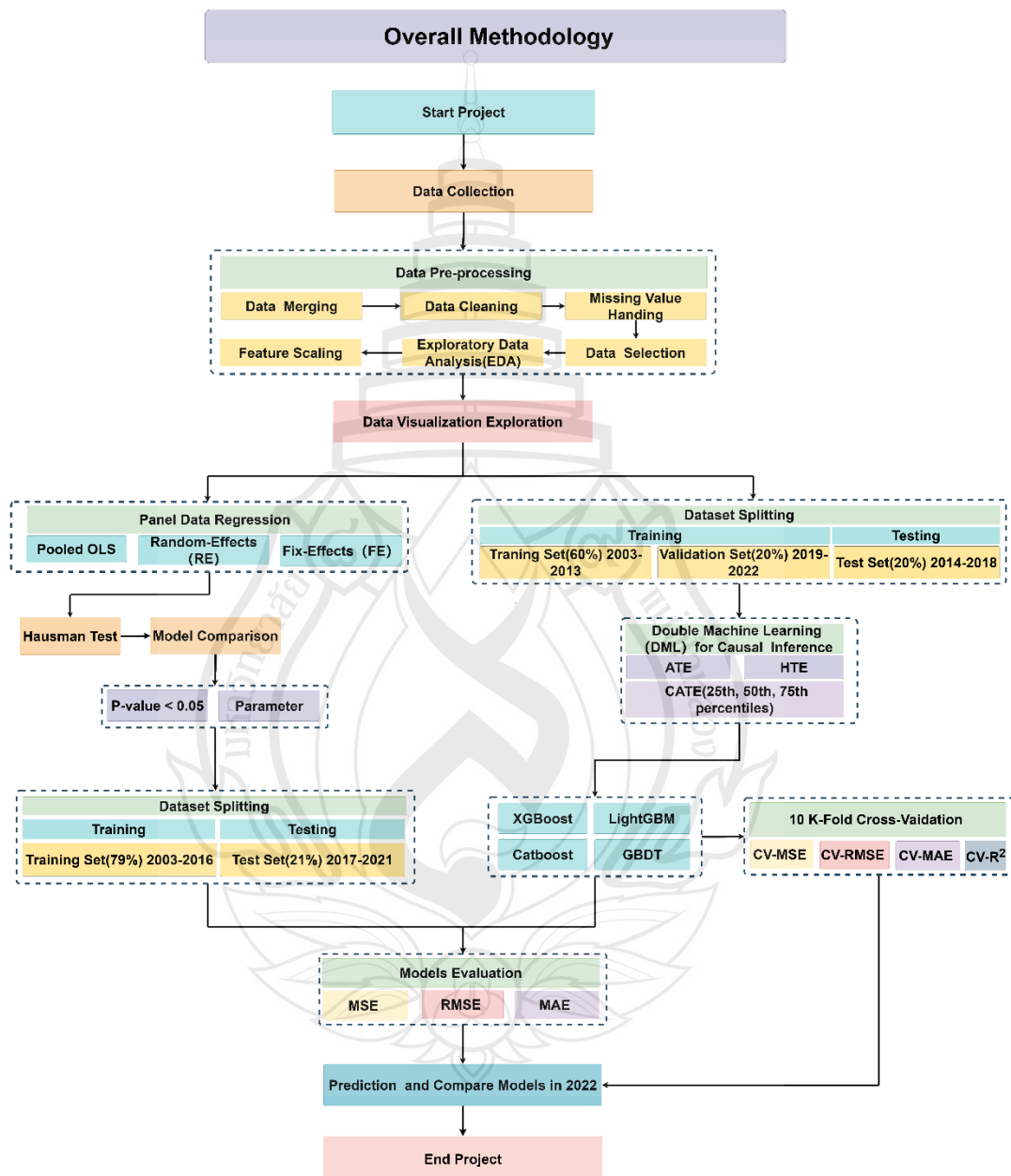


Figure 3.1 Overall Methodology

3.2 Data Collection

In this section, thorough explanations of the key variables impacting the crude marriage rate are presented, encompassing data collection and processing. The crude marriage rate and economic factors data are obtained from the English website of National Bureau of Statistics of China (<https://data.stats.gov.cn/english/>) and the China Statistical Yearbook (<https://www.stats.gov.cn/sj/ndsj/>), the China website of the National Bureau of Statistics of China (<https://www.stats.gov.cn/sj/ndsj/>) and the Tencent Gu Yu Data 2020 (<https://www.stats.gov.cn/sj/ndsj/>). Each relevant dataset undergoes annual updates on the National Bureau of Statistics of China's website and encompasses the years 2003 to 2022, culminating in a 20-year dataset.

3.3 Data Merging

In the study, due to the independence of the datasets provided by the National Bureau of Statistics of China and the China Statistical Yearbook, there is no established correlation between each dataset. Additionally, the relevant independent variables and dependent variables available for download on the public platform consist of several separate Excel spreadsheets. The maximum span of publicly available datasets is 20 years, specifically from 2003 to 2022. To facilitate future research due to the multitude of tables, we consolidated the useful data for the study, creating distinct Excel and CSV format files. The columns in the CSV file data include Region, Year, Dependent Variables, and Independent Variables.

3.4 Data Pre-Processing

Scikit-learn, a Python library for machine learning, was utilized to combine and preprocess data from various independent CSV files in the research. The missing values in certain columns were handled through the Simple Imputer tool available in the SciPy library, ensuring the integrity of the data.

3.5 Data Selection

After merging and preprocessing the data, we decided to select the data column features to include the 31 provinces in mainland China, the years are 2003-2022, the dependent variable y is Crude marriage rate, and the independent variables are (X_1-X_7) . Table 2.1 offers an explanation of the arrangement of the newly screened features.

Table 3.1 Features Date Selection

No.	Features	Features Explanation
1	Region	31 provinces in mainland China (No data from Hong Kong, Macao and Taiwan)
2	Year	2003-2022
3	Crude_arriage_rate (Y)	The Number of Marriages Occurring in a Population During a Given Period
4	GDP (X_1)	Gross Regional Product (100 million yuan)
5	House_Prices (X_2)	Average Selling Price of Commercialized Residential Buildings (yuan / square meters)
6	Gross_Dependency_Ratio (X_3)	Gross Dependency Ratio (Sample Survey) (%)
7	Birth_Rate (X_4)	Birth Rate (%)
8	Female (X_5)	Female Population Aged 15 and Over (Sample Survey) (person)
9	Average_Years_of_Education (X_6)	Average years of education per capita
10	Sex_Ratio (X_7)	Sex Ratio (Female=100) (Sample Survey) (female=100)

3.6 Development Environment

In this research, we utilized Python versions 3.11 and 3.12 as our primary programming language. For data processing and analysis, we employed several libraries: NumPy for foundational scientific computing, pandas for efficient data manipulation and analysis, and SciPy for advanced mathematical functions and scientific calculations. In the domain of machine learning, we utilized scikit-learn, a widely-used library for implementing various machine learning algorithms. For data visualization, we relied on Matplotlib, which serves as the basic plotting library, and Seaborn, which is built on top of Matplotlib to provide enhanced statistical data visualization capabilities. Additionally, for statistical analysis, we incorporated statsmodels for statistical modeling and econometrics, along with scipy.stats, a module within SciPy dedicated to statistical functions. This environment facilitated comprehensive analysis and modeling throughout the research.

3.7 Exploratory Data Analysis (EDA)

As shown in Figure 3.2, there is a heat map of a crude marriage rate dataset. The values on the diagonal are all 1, indicating that each variable is perfectly positively correlated with itself. In the matrix, the correlation between House Prices and Average years of education per capita is the highest, with a correlation coefficient of 0.66, suggesting a strong positive relationship between them and possibly a similar trend in their changes. The correlation between Birth Rate and average years of education per capita is the lowest, with a correlation coefficient of -0.61, indicating a significant negative relationship, meaning that when one variable increases, the other tends to decrease. Additionally, the correlation between GDP and House Prices is 0.43, showing a moderate positive correlation, while the correlation between variables 3 and 4 is 0.40, indicating a certain degree of positive correlation as well. The correlation coefficients between other variables are relatively small, mostly ranging from -0.4 to 0.4, which suggests that the linear relationships among these variables are weak or almost non-existent. Overall, most correlations among these variables are not significant, with only

a few variable pairs showing notable relationships. Therefore, in practical analysis, we can focus on these variable pairs with significant correlations (such as House Prices and Average years of education per capita, and Birth Rate and rage years of education per capita) to further explore their mutual influences.



Figure 3.2 Correlation Heatmap of Features

3.8 Feature Scaling

Data normalization, known as feature scaling, is a key preprocessing step in many regression-oriented machine learning models. It involves standardizing numerical attributes to a common scale. In this study, Min-Max Scaling was used to normalize attributes to a range of 0 to 1, Chaurasia and Haq (2023) aiming to reduce

the impact of dimension variations and improve model training efficiency and reliability. Figure 3.3 shows the Raw Data, Figure 3.4 shows the Features Scaling Data, and Figure 3.5 shows the Features Describe Data. The number 1 represents GDP, 2 represents House Prices, 3 represents Gross Dependency Ratio, 4 represents Birth Rate, 5 represents Female, 6 represents Average years of education, and 7 represents Sex Ratio. It provides statistical descriptions for seven features, covering 620 data points. The mean values for features 3 and 4 are close to 0.5, indicating they are nearly binary distributed, while other features have smaller mean values. The maximum and minimum values for features 1 and 2 are 1 and 0, respectively, suggesting they may be binary variables. Feature 5 has a small mean and standard deviation, indicating a narrow distribution. The quartiles show the distribution for each feature, with features 3 and 4 having higher upper quartiles, indicating higher values for these features.

Region	Year	Crude_Ma	GDP	House_Pri	Gross_De	Birth_Rate	Female	Average_y	Sex_Ratio
Beijing	2003	6.38736	5267.2	4456	27.8	5.1	6111	10.3457	106.08
Beijing	2004	8.40589	6252.5	4747.14	26.7	6.1	6256	10.5586	105.01
Beijing	2005	6.24187	7149.8	6162.13	26.7	6.29	90424	10.6858	102.65
Beijing	2006	10.6184	8387	7375.41	26.9	6.26	6554	10.9501	97.21
Beijing	2007	6.97494	10425.5	10661.2	24.7	8.32	6645	11.0853	99.06
Beijing	2008	8.26652	11813.1	11648	25	8.17	6608	10.9696	103.38
Beijing	2009	9.70968	12900.9	13224	25	8.06	6692	11.1726	104.27
Beijing	2010	6.98267	14964	17151	38.9	7.48	1938	11.0092	102.03
Beijing	2011	8.49802	17188.8	15517.9	21.3	8.29	7725	11.555	104.09
Beijing	2012	8.3205	19024.7	16553.5	21.9	9.05	7649	11.8363	105.18
Beijing	2013	7.65177	21134.6	17854	22.7	8.93	7574	12.0284	107.62
Beijing	2014	7.77983	22926	18499	23	9.75	7946	11.8542	102.24
Beijing	2015	7.54113	24779.1	22300	26.2	7.96	144262	12.1464	109.45
Beijing	2016	7.5262	27041.2	28489	29.2	9.32	7885	12.3891	105.85
Beijing	2017	6.85962	29883	34117	30.6	9.06	7860	12.6651	102.65
Beijing	2018	6.25	33106	37420.2	27.8	8.24	8017	12.6754	98.75
Beijing	2019	5.85388	35445.1	38433	28	8.12	7429	12.782	101.56
Beijing	2020	5.18045	35943.3	42684	38.9	13.2	1938	12.2878	102.03
Beijing	2021	4.70078	41045.6	46941	35.8	6.35	10004	12.6609	104.26
Beijing	2022	4.16209	41610.9	47784	37.3	5.67	9660	12.7146	104.18
Tianjin	2003	6.23145	2257.8	2393	33.2	7.14	4365	9.24681	97.44
Tianjin	2004	7.48047	2621.1	2950.34	31.4	7.31	4377	9.64492	98.37
Tianjin	2005	6.04027	3158.6	3987.22	28.8	7.44	60578	9.5129	100.81
Tianjin	2006	7.72093	3538.2	4649.25	29.3	7.67	4362	9.72924	96.81

Figure 3.3 Raw Data

Region	Year	Crude_Ma	Crude_Ma	GDP	House_Pri	Gross_Dej	Birth_Rate	Female	Average_y	Sex_Ratio
Beijing	2003	0.00639	0.00639	0.03941	0.07458	0.22193	0.12055	0.00773	0.73061	0.44728
Beijing	2004	0.00841	0.00841	0.04705	0.0808	0.19321	0.18904	0.00794	0.75414	0.41268
Beijing	2005	0.00624	0.00624	0.05401	0.11102	0.19321	0.20205	0.13447	0.76821	0.33635
Beijing	2006	0.01062	0.01062	0.06361	0.13694	0.19843	0.2	0.00839	0.79744	0.16041
Beijing	2007	0.00697	0.00697	0.07942	0.20712	0.14099	0.3411	0.00853	0.81239	0.22025
Beijing	2008	0.00827	0.00827	0.09018	0.22819	0.14883	0.33082	0.00847	0.79959	0.35996
Beijing	2009	0.00971	0.00971	0.09862	0.26185	0.14883	0.32329	0.0086	0.82204	0.38875
Beijing	2010	0.00698	0.00698	0.11462	0.34573	0.51175	0.28356	0.00145	0.80397	0.3163
Beijing	2011	0.0085	0.0085	0.13187	0.31085	0.05222	0.33904	0.01015	0.86432	0.38292
Beijing	2012	0.00832	0.00832	0.14611	0.33297	0.06789	0.3911	0.01004	0.89543	0.41818
Beijing	2013	0.00765	0.00765	0.16248	0.36074	0.08877	0.38288	0.00992	0.91667	0.49709
Beijing	2014	0.00778	0.00778	0.17637	0.37452	0.09661	0.43904	0.01048	0.8974	0.32309
Beijing	2015	0.00754	0.00754	0.19074	0.4557	0.18016	0.31644	0.2154	0.92972	0.55627
Beijing	2016	0.00753	0.00753	0.20829	0.58789	0.25849	0.40959	0.01039	0.95655	0.43984
Beijing	2017	0.00686	0.00686	0.23033	0.70809	0.29504	0.39178	0.01035	0.98708	0.33635
Beijing	2018	0.00625	0.00625	0.25533	0.77865	0.22193	0.33562	0.01059	0.98821	0.21022
Beijing	2019	0.00585	0.00585	0.27347	0.80028	0.22715	0.3274	0.00971	1	0.3011
Beijing	2020	0.00518	0.00518	0.27733	0.89107	0.51175	0.67534	0.00145	0.94535	0.3163
Beijing	2021	0.0047	0.0047	0.31691	0.98199	0.43081	0.20616	0.01358	0.98661	0.38842
Beijing	2022	0.00416	0.00416	0.32129	1	0.46997	0.15959	0.01306	0.99255	0.38583
Tianjin	2003	0.00623	0.00623	0.01607	0.03052	0.36292	0.26027	0.0051	0.60909	0.16785
Tianjin	2004	0.00748	0.00748	0.01889	0.04243	0.31593	0.27192	0.00512	0.65312	0.19793
Tianjin	2005	0.00604	0.00604	0.02306	0.06457	0.24804	0.28082	0.0896	0.63852	0.27684
Tianjin	2006	0.00772	0.00772	0.026	0.07871	0.2611	0.29658	0.0051	0.66244	0.14748

Figure 3.4 Features Scaling Data

	1	2	3	4	5	6	7
count	620.000000	620.000000	620.000000	620.000000	620.000000	620.000000	620.000000
mean	0.147114	0.112171	0.497402	0.530187	0.055889	0.553571	0.379543
std	0.159548	0.128166	0.175902	0.204381	0.133562	0.137955	0.120939
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.040104	0.040372	0.381201	0.389212	0.009564	0.488183	0.316300
50%	0.095086	0.081771	0.511749	0.554110	0.019711	0.556310	0.361902
75%	0.194760	0.136058	0.613577	0.675342	0.036913	0.624929	0.442594
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

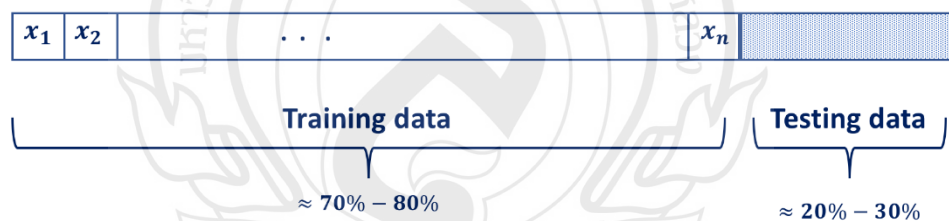
Figure 3.5 Features Describing Data

3.9 Dataset Splitting

The training set is the portion of the dataset that is reserved for fitting the model. In other words, the model looks at the data in the training set and learns from it to directly improve its parameters. The validation set is the dataset used to evaluate and fine-tune the machine learning model during training, helping to assess the performance of the model and make adjustments. The test set is the dataset used to evaluate the final performance of the trained model.

3.9.1 Two-Way Split (7:3 or 8:2)

As shown in Figure 1.3, shows that the Two-Way Split (7:3 or 8:2). Splitting data into a training set and a test set, often used in larger datasets or scenarios where extensive parameter tuning isn't necessary. This approach is straightforward but may result in limited test data, affecting the stability of performance evaluation. In this study, 2003 to 2016 was used as the training set, and 2017 to 2021 was used as the test set. The data from 2022 was reserved for prediction.



Source Kecojevic (2020)

Figure 3.6 Two-Way Split (7:3 or 8:2)

3.9.2 Three-Way Split (6:2:2)

As shown in Figure 1.3, shows that the Three-Way Split (6:2:2). Often used when data is smaller in scale. Here, data is divided into a training set (60%), validation set (20%), and test set (20%), which allows for performance evaluation during training

and tuning via the validation set, while keeping an independent test set to assess the model's final generalization capability. The training set is used to fit the model, the validation set is used for hyperparameter tuning and selecting the best model configuration, and the test set is kept entirely separate from training and tuning to evaluate the model's generalization performance. The data we study belongs to a small-scale sample set (tens of thousands of samples). The commonly used allocation ratio in machine learning is 60% training set, 20% validation set, and 20% test set. We use 2003-2013 as the training set, 2014 to 2018 as the training set, and 2019 to 2022 as the validation set.



Source Ding (2020)

Figure 3.7 Three-Way Split (6:2:2)

3.9.3 Benefits of Using a Three-Way Split

Using a three-way split for small-scale data offers key advantages by reducing data leakage, accurately evaluating generalization, and preventing overfitting during tuning. With this approach, the model is tuned using only the validation set, keeping the test set fully isolated and preserving its integrity for final evaluation. This separation ensures that test results better reflect real-world performance and allows for unbiased generalization assessment. Overall, a three-way split is particularly valuable when both parameter tuning and independent testing are needed, enabling a more accurate measure of a model's practical effectiveness on limited data.

3.10 Data Visualization Exploration

The distributions of various socioeconomic factors show distinct patterns. GDP, house prices, and the female population exhibit strong right skewness, indicating that most values are concentrated at lower levels with a few higher outliers. The gross dependency ratio and sex ratio distributions are more symmetric, resembling normal curves, though the sex ratio has a slight skew near a balanced 1.0. The birth rate shows moderate right skewness, while the average years of education per capita follows an almost normal distribution, centered around 8-10 years. Overall, these factors reveal diverse trends in their respective distributions.

3.10.1 Normal Distribution Test

In general, several variables such as GDP, house prices, and the female population show significant skewness, which could affect model performance and may require transformation or normalization for further analysis. Other variables like the gross dependency ratio, birth rate, and average years of education exhibit more balanced or normal-like distributions.

3.10.1.1 GDP

The distribution of GDP shows a strong right skew, indicating that most of the data points are clustered at lower GDP values, with a few outliers representing significantly higher GDP levels.

3.10.1.2 House Prices

House prices are heavily skewed to the right, with the majority of values concentrated at the lower end. The distribution suggests that higher house prices are relatively rare.

3.10.1.3 Gross Dependency Ratio

The distribution of the gross dependency ratio is more symmetric, resembling a bell curve, with most values falling around the center. There is a noticeable spread across the range.

3.10.1.4 Birth Rate

The birth rate distribution shows moderate skewness, with a concentration around lower values and a wider spread as the birth rate increases.

3.10.1.5 Female Population

Similar to GDP and house prices, the distribution of the female population is highly right-skewed, indicating a large number of regions or periods with relatively low female population numbers and few with very high populations.

3.10.1.6 Average Years of Education per Capita

The distribution of average years of education is almost normal, centered around 8-10 years, with few extreme values on either side.

3.10.1.7 Sex Ratio

The sex ratio follows a nearly normal distribution with a slight skew, concentrated around a value close to 1.0, indicating a balanced ratio between male and female populations.

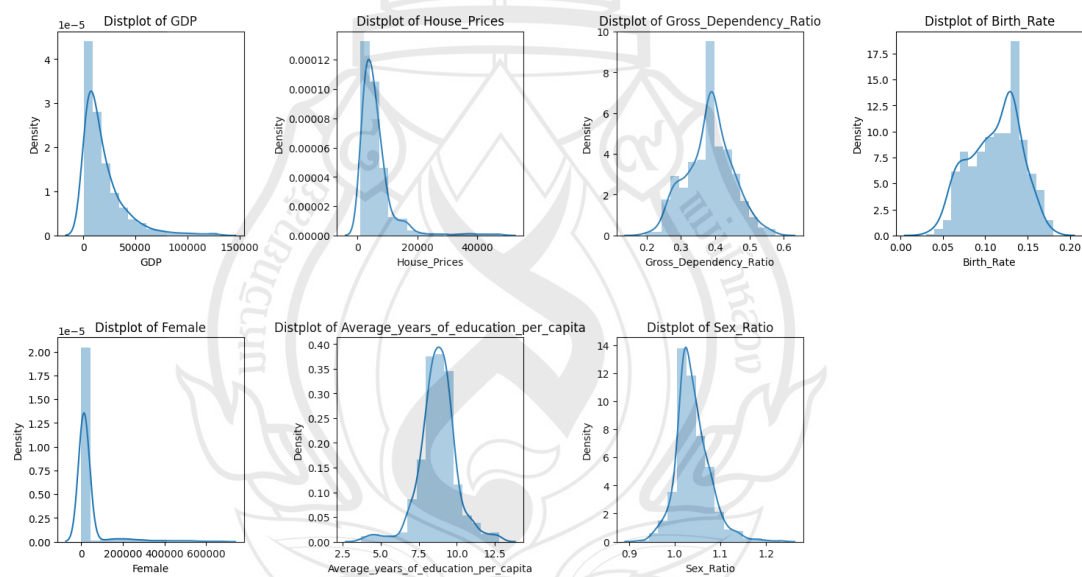


Figure 3.8 The Series of Distribution Plots

3.10.2 Crude Marriage Rate vs. Socioeconomic Factors

The figure 3.7 shows that the scatter plots collectively indicate that house prices and birth rates may have some significant influence on the crude marriage rate, while other factors like GDP, gross dependency ratio, and sex ratio seem to exhibit weaker or more complex relationships. The influence of education appears concentrated in a particular range but requires further investigation.

3.10.2.1 Crude Marriage Rate vs. GDP

There seems to be a weak negative relationship between GDP and the crude marriage rate. As GDP increases, the crude marriage rate slightly decreases, but the data points are dispersed, indicating high variability.

3.10.2.2 Crude Marriage Rate vs. House Prices

This plot shows a clear negative correlation. As house prices increase, the crude marriage rate tends to decrease significantly, suggesting that higher housing costs might be a deterrent to marriage.

3.10.2.3 Crude Marriage Rate vs. Gross Dependency Ratio

There appears to be a more scattered and weak relationship. No clear trend is visible between the gross dependency ratio and the crude marriage rate, suggesting minimal or no direct correlation.

3.10.2.4 Crude Marriage Rate vs. Birth Rate

The relationship here seems somewhat positive, with higher birth rates being associated with higher crude marriage rates, although the relationship is not strictly linear.

3.10.2.5 Crude Marriage Rate vs. Female Population

This plot shows little to no visible relationship between the female population and the crude marriage rate. The data points are quite scattered, suggesting minimal influence.

3.10.2.6 Crude Marriage Rate vs. Average Years of Education per Capita

There seems to be a concentration of points around a particular range of years of education (around 7–10 years). Beyond that range, there is less data, and no clear pattern emerges.

3.10.2.7 Crude Marriage Rate vs. Sex Ratio

The data points are quite scattered, showing no obvious relationship between the sex ratio and the crude marriage rate. The plot suggests a weak or no correlation between these variables.

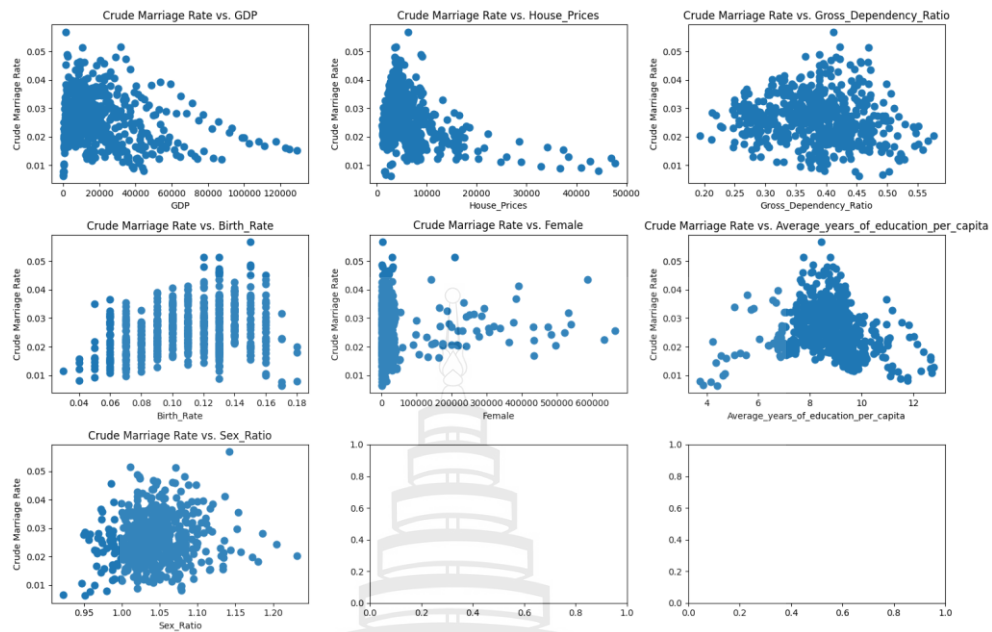


Figure 3.9 Crude Marriage Rate vs. Socioeconomic Factors

3.10.3 Pair Plot of Crude Marriage Rate and Socioeconomic Variables

The pair plot helps visualize multiple variables in relation to the crude marriage rate and among each other. Key observations include a possible negative relationship between house prices and the crude marriage rate, as well as some potential positive trends between birth rate, average years of education, and the marriage rate. However, many of the relationships appear weak or non-linear, and further statistical analysis would be required to confirm these observations.

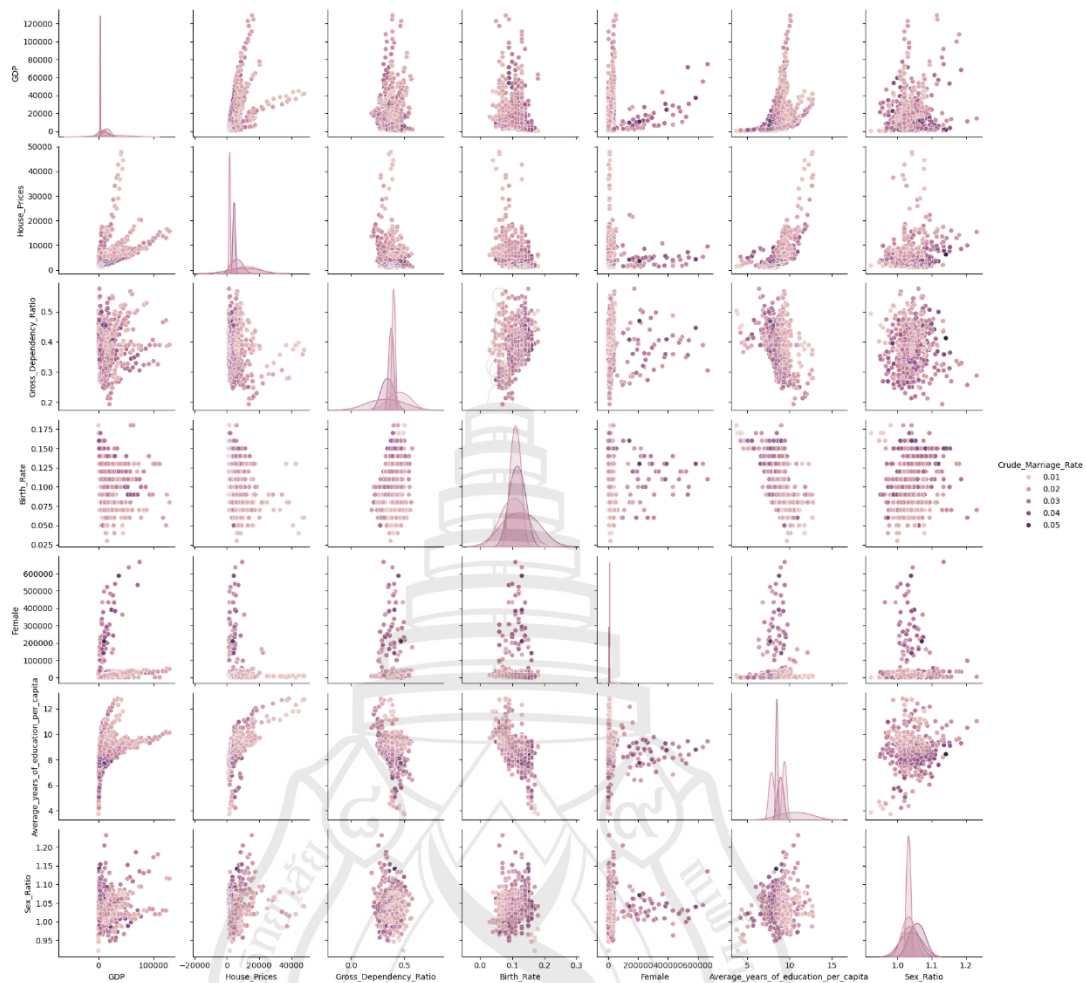


Figure 3.10 Pair Plot of Crude Marriage Rate and Socioeconomic Variables

3.11 The Panel Data Regression Model

Panel data combines cross-sectional and time series data. Cross-sectional data shows entities and variables at one time, while time series data tracks entities over time. (Ratnasari et al., 2023) analysis aims to determine the characteristics of MIT in Indonesia and the factors that influence it and obtain a panel data regression model formed from MIT modeling. Analysis using panel data regression method. The panel data regression model was obtained based on panel data, namely data consisting of a combination of cross-section and time series data. Panel data merges features of both data types into one model (Guggenberger, 2010). The impact of using Hausman pretests

on the size of hypothesis tests. It can be likened to a timeline where the same individuals are observed at regular intervals. Table 3.2 provides a detailed explanation of the panel data structure and its relevant characteristics.

Table 3.2 Panel Data Structure

Research Object	Time	Y	X ₁	X ₂	...	X _K
1	1	Y ₁₁	Y ₁₁₁	Y ₂₁₁	⋮	X _{K11}
	2	Y ₁₂	Y ₁₁₂	Y ₂₁₂	⋮	X _{K12}

	T	Y _{1T}	Y _{11T}	Y _{21T}	⋮	X _{K1T}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
N	1	Y _{N1}	Y _{1N1}	Y _{2N1}	⋮	Y _{KN1}
	2	Y _{N2}	Y _{1N2}	Y _{2N2}	⋮	Y _{KN2}

	T	Y _{NT}	Y _{1NT}	Y _{2NT}	⋮	Y _{KNT}

The advantage of panel data lies in our ability to control heterogeneity in the regression model by treating heterogeneity as fixed or random. There are various types of panel data regressions. The interpretation of Formula (3) is based on this symbol representation.

3.12 Evaluation Metrics in Machine Learning

(Gupta et al., 2022) Use the regression model evaluation indicators RMSE, MAE, indicators are mainly used to evaluate the prediction error rate and model performance in regression analysis.

3.12.1 Mean Squared Error (MSE)

Mean squared error (MSE) is a common measure of the quality of an estimator, such as a machine learning model. It calculates the average squared difference between

the predicted values and the actual values. A lower MSE value indicates a better fit of the model to the data. The formula for the MSE is;

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y})^2$$

3.12.2 Root Mean Squared Error (RMSE):

Root mean squared error (RMSE) is the square root of the mean squared error (MSE). It is another common measure of the quality of an estimator, and it represents the average error in the predictions. The formula for the RMSE is;

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y})^2}$$

Where:

N is the number of samples, Y_i is the true value, \hat{y} is the predicted value

3.12.3 Mean Absolute Error (MAE)

Mean absolute error (MAE) measures estimator quality by averaging absolute differences between predicted and actual values. a lower MAE value indicates a better fit of the model to the data. The formula for the MAE is;

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}|$$

3.12.4 K-Fold Cross Validation

10-K-Fold Cross Validation is a commonly used model evaluation method to estimate the performance of machine learning models on unseen data. The first step is data partitioning, which randomly divides the data set into 10 equal parts (called “folds”). The second step is model training and validation, which is performed 10 times. one fold is used as the validation set and nine folds are used as the training set. The third step is to calculate the average performance: the evaluation results of the 10 validations are averaged as the final performance indicator of the model. The formula for the 10-K-Fold Cross Validation is;

$$L_{cv} = \frac{1}{10} \sum_{k=1}^{10} L^k$$

Where:

L_{cv} is represents the average loss (or error) across all 10 folds in the cross-validation process

L^k is denotes the loss (or error) calculated for the k-th fold during the cross-validation



CHAPTER 4

RESULTS

4.1 Distribution of Marriage Rate

As shown in Figure 4.1, Figure 4.2, and Figure 4.3 shows the significant changes in China's marriage rate in the past 20 years. The social, economic and cultural factors behind it are very complex and deserve in-depth discussion. From 2003 to 2022, China's marriage rate has shown an overall trend of “rising first and then falling”. It reached a relative peak in 2012 and then dropped significantly in 2022. The western region has always maintained a low marriage rate, and the eastern coastal region had a high marriage rate before 2012, but it has dropped significantly after 2022. The marriage rate in the central region is relatively stable, but it also began to decline in 2022. Economic development, cost of living, cultural changes and population mobility are key factors affecting changes in marriage rates. Especially in 2022, the phenomena of late marriage, non-marriage and high marriage costs will become more obvious.

4.1.1 Overall Trends and Characteristics of Marriage Rates in 2003, 2012 and 2022

These three figures show the spatial and temporal evolution of marriage rates in 31 provinces in China in the past 20 years, providing important clues for studying the social, economic and cultural factors behind the marriage rates.

4.1.1.1 Spatial and Temporal Distribution Map of Marriage Rate in 2003

The Figure 4.1 shows that areas with low marriage rates are mainly concentrated in remote western provinces such as Tibet and Xinjiang (dark blue), while the marriage rates in the eastern coastal and central provinces are relatively high, mainly distributed in Shandong, Jiangsu and other places (light green and yellow). The

economically developed southeastern coastal areas (such as Guangdong and Fujian) also show relatively high marriage rates.

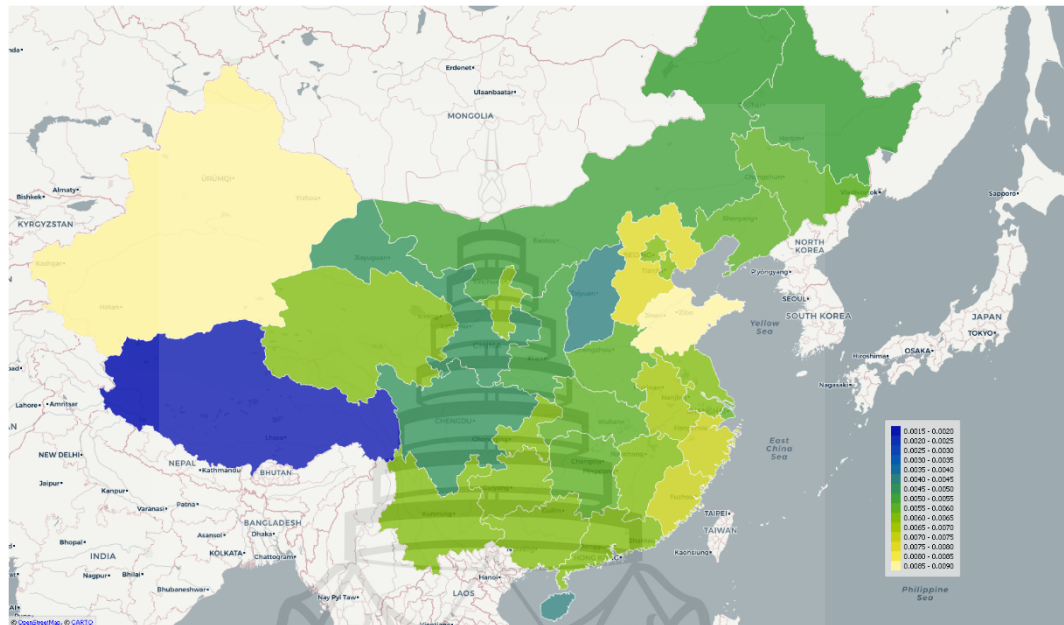


Figure 4.1 Spatial and Temporal Distribution Map of Marriage Rate in 2003

4.1.1.2 Spatial and Temporal Distribution Map of Marriage Rate in 2012

The Figure 4.2 shows an overall upward trend: in 2012, the overall marriage rate increased compared to 2003, and the gap between the marriage rates in the central and western provinces and the coastal provinces narrowed. The marriage rates in western provinces such as Tibet, Xinjiang and Ningxia are still low (blue), but slightly higher than in 2003. The marriage rates in the eastern coastal areas (such as Shanghai, Zhejiang and Guangdong) were at a high level in 2012 (light yellow).

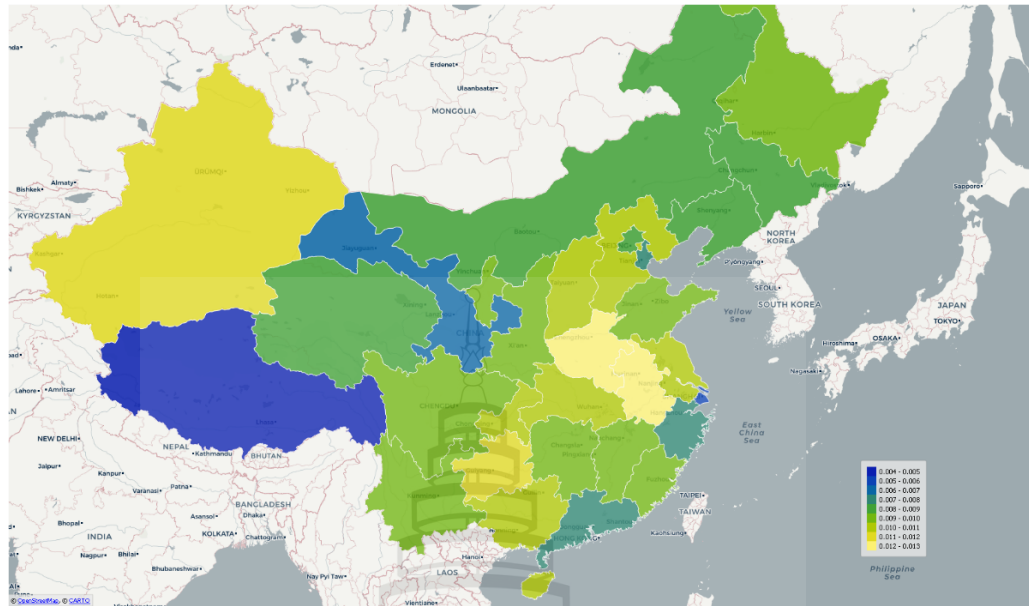


Figure 4.2 Spatial and Temporal Distribution Map of Marriage Rate in 2012

4.1.1.3 Spatial and Temporal Distribution Map of Marriage Rate in 2022

The Figure 4.3 shows a general decline in the marriage rate: by 2022, the marriage rate in many provinces has declined compared to 2012, especially in the developed eastern coastal and central provinces, such as Jiangsu, Zhejiang, and Shandong (blue and green). The marriage rate in western provinces (such as Tibet and Qinghai) has not increased significantly compared with previous years and remains at a low level. The marriage rate in Northeast China (such as Liaoning and Jilin) has dropped significantly compared with 2012, and is even lower than the national average.

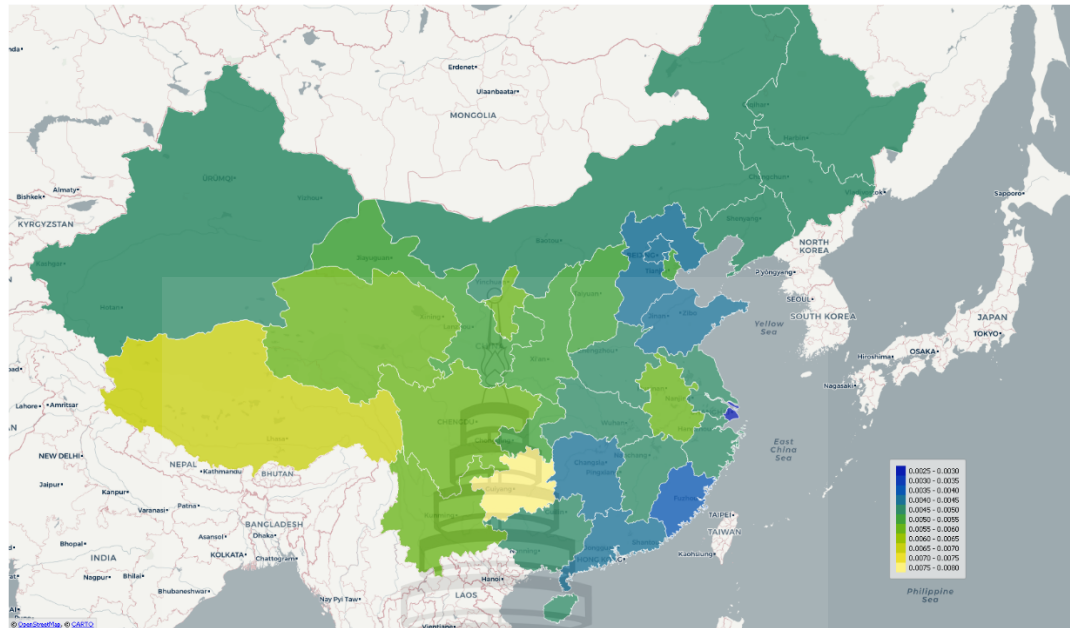


Figure 4.3 Spatial and Temporal Distribution Map of Marriage Rate in 2022

4.1.2 Regional Difference Analysis

To deeply analyze the changes in marriage rates in China's 31 provinces in 2003, 2012, and 2022, we can re-summarize these charts from multiple dimensions, including time trends, regional differences, potential driving factors, and combine them with the context of China's social and economic development to better understand these data.

4.1.2.1 Western Region

The marriage rates in western provinces such as Tibet and Xinjiang have always been low, especially in 2003 and 2022. This may be related to the region's economic development level, cultural customs and low population mobility.

4.1.2.2 Eastern Region

Eastern coastal provinces such as Jiangsu, Zhejiang, and Shanghai showed high marriage rates in 2003 and 2012, but saw a sharp decline in 2022. Rapid economic development and high living costs in these provinces, coupled with high population mobility and a trend toward late marriage, may have led to the decline in marriage rates.

4.1.2.3 Central Region

In central regions, such as Henan and Hubei, the marriage rate has changed relatively steadily. Although there are certain fluctuations, the overall trend is consistent with the national trend, that is, it has gradually declined since 2012.

4.1.3 The Impact of Social and Economic Context

The impact of social and economic background is mainly discussed from four aspects: economic factors, population mobility, changes in cultural and social concepts, and policies and social security.

4.1.3.1 Economic Factors

China's economic development and urbanization process have advanced rapidly in the past 20 years, especially in the eastern coastal provinces. The economic boom promoted high marriage rates between 2003 and 2012, especially in developed provinces, but it was accompanied by rising housing prices and increased living pressure, which led to rising marriage costs, especially in 2022.

4.1.3.2 Population Mobility

A large number of rural populations has migrated to cities. In areas with faster economic development, population mobility has intensified. The pressure of urban life and the accelerated pace of life have led to an increase in late marriage and non-marriage. This trend is most evident in the eastern coastal areas.

4.1.3.3 Changes in Culture and Social Attitudes

With the rapid economic development, the concept of marriage in Chinese society has also undergone profound changes. The younger generation has a more liberal and personalized attitude towards marriage, and the traditional concept of early marriage has gradually been replaced by the concept of late marriage or even no marriage, especially in developed regions and large cities, which has directly affected the overall decline in the marriage rate.

4.1.3.4 Policy and Social Security

China's family planning policy and the two-child policy and three-child policy in recent years have also indirectly affected the marriage rate. Although the government has tried to stimulate the birth rate through policies in recent years, the effect is limited, and the decline in the marriage rate is still a trend that cannot be ignored.

4.2 Numerical Features Boxplot

The figure 4.4 shows that the boxplot analysis of marriage rate in China reveals various key features including GDP, housing prices, fertility, gender ratio, female population, years of education, and male-female gender ratio. Specifically, the results reveal the following: GDP values span from 0 to 0.6, primarily falling between 0.2 and 0.4, with a few notable exceptions indicating significant GDP disparities between provinces. Housing prices vary from 0 to 0.8, showing a broad distribution and numerous anomalies, highlighting substantial differences in housing prices across provinces, including some with exceptionally high prices. Fertility rates are concentrated mostly between 0.2 and 0.8, with significant outliers suggesting notable differences among provinces. Gender ratios range from 0.2 to 1.0, showing a relatively tight concentration without clear outliers, indicating minor variations among provinces. Female population values range from 0 to 1.0, with the majority falling between 0.4 and 0.8, showcasing discrepancies in female population levels across provinces. Years of education range from 0.6 to 1.0, with a concentrated distribution and no significant outliers, suggesting minimal differences in education levels among provinces, with most having higher education rates. Male-female gender ratio values span from 0.4 to 1.0 with a symmetrical distribution, implying uniformity in male-female gender ratios among provinces. Analyzing these metrics offers crucial insights for forecasting marriage rates in China using Ridge and polynomial regression models, assessing model performance through various cross-validation techniques such as Holdout, LOOCV, and K-Fold CV to ensure accurate predictions and enhance comprehension of marriage rate trends.

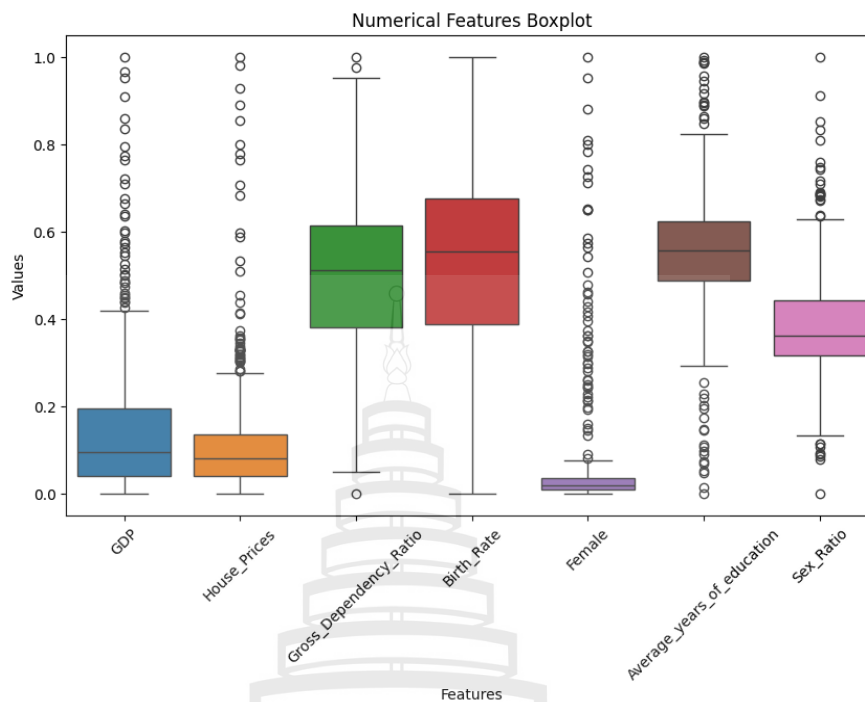


Figure 4.4 Numerical Features Boxplot

4.3 Panel Data Regression Models Results

The table compares three panel data models: Pooled OLS, Random Effects, Fixed Effects, using various evaluation metrics. The Table IV presents the results of three different regression models: Pooled OLS, Random Effects, and Fixed Effects, the models are compared based on their R-squared values, and p-values, as well as the Parameter values for the independent variables (X_1 to X_7).

4.3.1 R^2

The R^2 value indicates the proportion of variance in the dependent variable that is predictable from the independent variables. The Random Effects model has the highest R^2 value (0.2910), suggesting it explains the most variance in the crude marriage rate data among the three models.

4.3.2 P-value (Hausman Test)

Through the Hausman test, The Hausman test further supported the suitability of the Fixed Effects model with a statistically significant p-value of 6.458e-16. This model is best suited to capture the reverse effects on the marriage rate, showing a direct positive correlation with the data, indicating that the selected predictor factors have a significant and consistent impact on the marriage rate.

4.3.3 Parameter Value

The table 4.1 lists the parameter (Const, X_1 , X_2 , X_3 , X_4 , X_5 , X_6 , X_7) for each model. In the Pooled OLS and Fixed Effects models, the constant term (Const) is the same (3.2812), while in the Random Effects model, it is slightly higher (4.2974). The coefficients of the respective variables (X_1 to X_7) differ across the models, but show a similar overall trend. For instance, both X_1 (GDP), X_2 (house prices) and X_3 (gross dependency ratio) exhibit negative effects in all models, with X_3 being especially significant (-7.5618 in Pooled OLS and Fixed Effects, -5.6215 in Random Effects). X_4 (birth rate) also consistently shows negative effects in all models. Conversely, both X_6 and X_7 (sex ratio) demonstrate positive effects across all models, with X_6 (average years of education) showing the strongest effect in the Random Effects model (7.3140). Additionally, while the coefficient of X_5 (Female) is positive in the Pooled OLS and Fixed Effects models (0.2708), it is negative in the Random Effects model (-0.0707). Overall, the Random Effects model excels in explaining the impact of variables, particularly demonstrating higher explanatory power in controlling individual effects.

Table 4.1 The Model Comparison Results of Pooled OLS, Random Effects, Fixed Effects

The Model Comparison Results			
	Pooled OLS	Random Effects	Fixed Effects
R^2	0.2465	0.2910	0.2465
P-value (Hausman Test)		6.458e-16	
Const	3.2812	4.2974	3.2812
X_1	-0.8548	-2.2743	-0.8548
X_2	-7.5618	-5.6215	-7.5618
X_3	-2.2973	-4.5810	-2.2973
X_4	4.2011	3.9717	4.2011
X_5	0.2708	-0.0707	0.2708
X_6	7.0050	7.3140	7.0050
X_7	0.8529	1.0652	0.8529

4.4 Pooled OLS, Random Effects and Fixed Effects Evaluation Index Results and Prediction

The study shows that Table 4.2 lists the results of three different regression models. As shown in Fig.6, the actual marriage rate and predicted marriage rate in 2022. Detailed analysis of the study below.

4.4.1 The Panel Data Results of Pooled OLS, Random Effects and Fixed Effects

The Table 4.2 presents evaluation indicator results of the evaluation metrics (MSE, RMSE, MAE) of three machine learning models: pooled OLS, random effects, and fixed effects. The random effects model performed well, with the lowest MSE (1.661), RMSE (1.2888), and MAE (1.0888), indicating that it is more accurate in predicting crude marriage rates. The pooled OLS performed moderately, with an MSE of 2.436, RMSE of 1.5608, and MAE of 1.4103, while the fixed effects model

performed the worst, with the highest MSE (4.7498), RMSE (2.1794), and MAE (1.8471). This highlights the effectiveness of the random effects model in capturing regional differences in marriage rates.

Table 4.2 Pooled OLS, Random Effects and Fixed Effects Evaluation Index Results

The Summary of Model Results			
	Pooled OLS	Random Effects	Fixed Effects
MSE	2.4363	1.6610	4.7498
RMSE	1.5608	1.2888	2.1794
MAE	1.4103	1.0888	1.8471

When evaluating the three models - Pooled OLS, Random Effects, and Fixed Effects - we find that the Random Effects model performs best in error metrics (MSE: 1.6610, RMSE: 1.2888, MAE: 1.0888), while the Fixed Effects model has the highest error indicators, with Pooled OLS falling between the two. However, despite Random Effects' superior performance in these metrics, the Fixed Effects model is considered to potentially yield the best results, a judgment that may be based on factors not shown in the table. The Fixed Effects model might be preferred because it better controls for time-invariant individual characteristics, reducing omitted variable bias; mitigates endogeneity problems caused by omitted variables; is more suitable for capturing inter-individual differences in panel data; provides consistent estimates; may offer more meaningful interpretations in policy analysis; is more appropriate for samples containing most of the population individuals; and if a Hausman test was conducted, the results might support using the Fixed Effects model. Therefore, although error indicators show the Random Effects model fits better, the Fixed Effects model may be more suitable when considering practical applications and theoretical explanations. This reminds us that when selecting the best model, we should not only consider goodness of fit but also weigh multiple factors such as data structure, research objectives, and theoretical foundations.

4.4.2 Prediction of Marriage Rate

The Figure 4.5 compares the actual marriage rates by region in 2022 with the predicted values from the random effects model. The graph compares the actual and predicted Crude Marriage Rates for 2022 across various regions in China using three machine learning models: Pooled OLS, Random Effects, and Fixed Effects. The actual rates, represented by blue circles, vary significantly by region, with lower rates in metropolitan areas like Beijing, Shanghai, and Tianjin, and higher rates in regions such as Xinjiang, Henan, and Shandong. The Pooled OLS model, shown by orange crosses, displays less variability and tends to underestimate rates in high-rate regions and overestimate in low-rate regions. The Random Effects model, indicated by green squares, aligns more closely with actual rates than Pooled OLS but still shows discrepancies, particularly in extreme-value regions. The Fixed Effects model, represented by red diamonds, most accurately follows actual rates, though it misses some peaks and troughs. Overall, the Fixed Effects model is the most reliable, highlighting significant regional variability in marriage rates and the limitations of the Pooled OLS model in assuming regional homogeneity. Future improvements could involve adding more variables and using advanced machine learning techniques to enhance prediction accuracy.

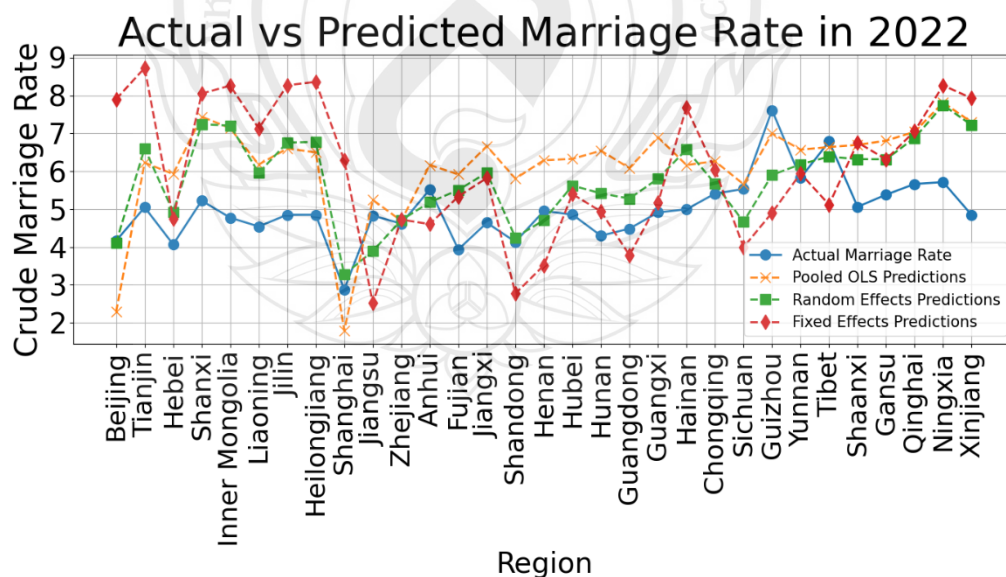


Figure 4.5 Actual vs Predicted Marriage Rate in 2022 (Pooled OLS, Random Effects, Fixed Effects)

4.5 Causal Inference Using Dual Machine Learning (DML) with XGBoost, LightGBM, CatBoost, and GBDT

The Table 4.3 compares three panel data models: Pooled OLS, Random Effects, Fixed Effects, using various evaluation metrics. The Table 4.3 presents the results of three different regression models: Pooled OLS, Random Effects, and Fixed Effects, the models are compared based on their R-squared values, and p-values, as well as the Parameter values for the independent variables (X_1 to X_7).

4.5.1 Results of ATE and CATE for Independent Variable Features

The Table 4.3 illustrates the impact of various independent variables on the Crude Marriage Rate (CMR), analyzed using Double Machine Learning (DML) to estimate the Average Treatment Effect (ATE), Conditional Average Treatment Effect (CATE), and Heterogeneous Treatment Effect (HTE). Notably, the Birth Rate (X_4) and Average Years of Education (X_6) exhibit the highest ATE values of 5.492 and 5.666, respectively, indicating a strong positive influence on CMR. In contrast, the Female (X_5) variable shows a significant negative ATE (-2.353), suggesting a decrease in CMR. The HTE values reveal variability in the effects; for example, GDP (X_1) shows substantial heterogeneity (4.968), implying varied impacts across different subpopulations. House Prices (X_2) has a negative HTE of -10.145, indicating a consistent negative effect on CMR across different groups. The CATE values at the 25th, 50th, and 75th percentiles highlight the conditional effects, with notable variations observed in House Prices (X_2) and Birth Rate (X_4), suggesting context-specific influences on CMR. These findings underscore the nuanced and multifaceted relationships between these variables and the Crude Marriage Rate.

Table 4.3 The Results of Features, ATE, CATE and HTE

Features	The Results of Features, ATE, CATE and HTE				
	ATE	CATE (25th, 50th, 75th percentiles)			HTE
		25th	50th	75th	
GDP (X ₁)	3.196	0.000	0.331	-1.065	4.968
House_Prices (X ₂)	-0.443	0.000	-1.336	1.222	-10.145
Gross_Dependency_Ratio (X ₃)	0.352	0.626	-0.962	0.737	0.675
Birth_Rate (X ₄)	5.492	1.211	0.097	1.165	4.932
Female (X ₅)	-2.353	0.000	0.000	-0.708	-3.314
Average_years_of_education (X ₆)	5.666	0.456	-0.163	-0.883	5.544
Sex_Ratio (X ₇)	-0.706	-1.041	0.029	-0.184	-0.619

4.5.2 K-Fold Cross Validation Results

The table compares four models: XGBoost, LightGBM, CatBoost and GBDT, The Table IV presents the results of four different regression models: 10 K-Fold Cross Validation Results for 4 Models are compared based on their R-squared values.

The Table 4.4 shows the performance comparison of four machine learning models in 10-fold cross validation. In the 10-fold cross validation, the performance of the four models, XGBoost, LightGBM, CatBoost, and GBDT, were compared. The results show that CatBoost performs best in all evaluation indicators, with an MSE of 0.942, an RMSE of 0.958, a MAE of 0.704, and an R² of 0.780. GBDT follows closely with an MSE of 0.956, an RMSE of 0.963, a MAE of 0.763, and an R² of 0.774. XGBoost also shows strong performance with an MSE of 0.979, an RMSE of 0.975, a MAE of 0.746, and an R² of 0.771. LightGBM has the highest MSE and RMSE values, at 1.060 and 1.021 respectively, a MAE of 0.777, and the lowest R² at 0.749. In summary, CatBoost outperforms the other three models in terms of prediction accuracy, especially in the MAE and R² indicators.

Table 4.4 10 K-Fold Cross Validation Results for 4 Models

Evaluation	10 K-Fold Cross Validation Results for 4 Models			
Metric	XGBoost	LightGBM	CatBoost	GBDT
CV-MSE	0.979	1.060	0.942	0.956
CV-RMSE	0.975	1.021	0.958	0.963
CV-MAE	0.746	0.777	0.704	0.763
CV-R ²	0.771	0.749	0.780	0.774

4.6 Prediction of Marriage Rate of XGBoost, LightGBM, CatBoost, and GBDT

The study shows that table 4.5 lists the results of four different regression models. As shown in Figure 4, the performance indicators of the four models are compared. As shown in Fig.3, the actual marriage rate and predicted marriage rate in 2022. Detailed analysis of the study below.

4.6.1 The Model Results of XGBoost, LightGBM, CatBoost, and GBDT

The Table 4.5 shows the evaluation index results of the four machine learning models. In the model evaluation, the performance of XGBoost, LightGBM, CatBoost, and GBDT are compared by MSE, RMSE, and MAE. The results show that CatBoost performs best in MSE and RMSE indicators, which are 3.534 and 1.880 respectively, but is slightly inferior to LightGBM in MAE. LightGBM has an MSE of 3.862 and an RMSE of 1.965, but performs best in the MAE indicator with a value of 1.536. XGBoost's MSE and RMSE are 4.070 and 2.017 respectively, slightly higher than those of CatBoost and LightGBM, with an MAE of 1.634. GBDT has the worst performance, with an MSE of 4.425, an RMSE of 2.103, and an MAE of 1.689. Overall, CatBoost performs best in MSE and RMSE, while LightGBM performs well in MAE.

Table 4.5 The Results of XGBoost, LightGBM, CatBoost, and GBDT Evaluation Metric

Evaluation Metric	The Summary of Model Results			
	XGBoost	LightGBM	CatBoost	GBDT
MSE	4.070	3.862	3.534	4.425
RMSE	2.017	1.965	1.880	2.103
MAE	1.634	1.536	1.539	1.689

4.6.2 Prediction of Marriage Rate

This Figure 4.6 shows the compares the actual and predicted crude marriage rates across different regions in 2022, using four machine learning models: XGBoost, LightGBM, CatBoost, and GBDT. The solid lines depict actual marriage rates, while the dotted lines represent the models' predictions. Overall, the actual rates exhibit significant regional fluctuations, while the models' predictions follow a more consistent trend. For instance, in Beijing, XGBoost closely matches the actual values, while LightGBM is lower, and CatBoost is higher. In Hebei, XGBoost and CatBoost are accurate, whereas LightGBM and GBDT show deviations. In Inner Mongolia, LightGBM performs better, while XGBoost and CatBoost underpredict. In Shanxi, all models overpredict. Liaoning shows higher predictions from all models, contrasting with the lower actual rate. This overprediction pattern is also seen in Hainan and Guangxi, while Heilongjiang and Jiangsu are underpredicted. These results highlight the need for improvement, especially in regions with high variability in marriage rates, and provide insights for selecting more accurate prediction tools.

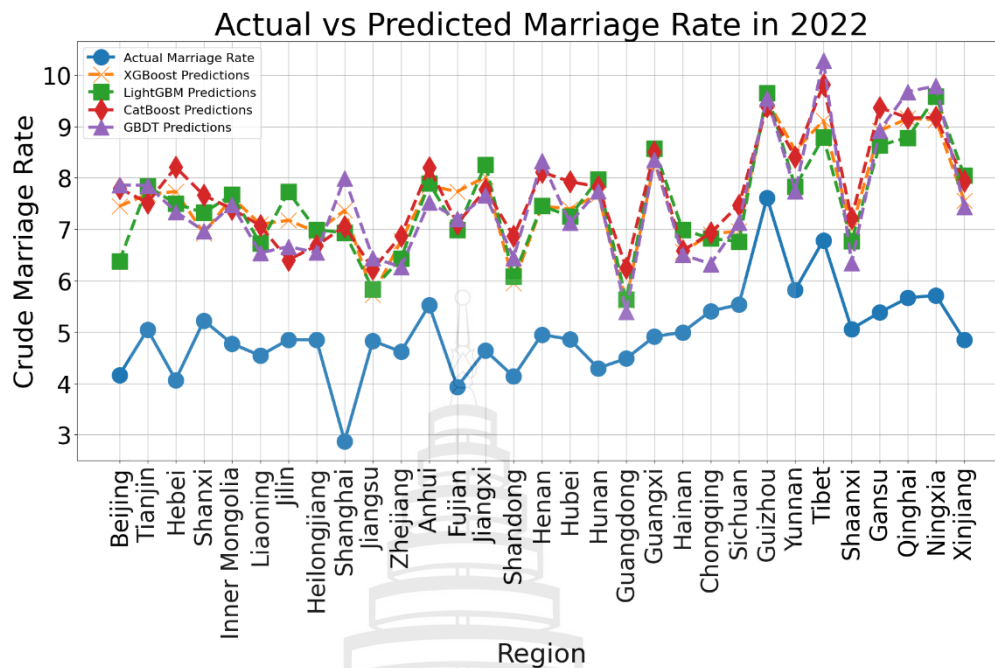


Figure 4.6 Actual vs Predicted Marriage Rate in 2022 (XGBoost, LightGBM, CatBoost, GBDT)

4.7 Discussion

The findings of this study reflect the broader social and economic transformation occurring in China over the past two decades. The observed decline in marriage rates, particularly in 2022, is indicative of deeper societal shifts, including the rising costs of living, increasing housing prices, and changing cultural norms around marriage and family life. Regions with higher economic development, such as the eastern coastal areas, have witnessed the most dramatic declines, driven by factors such as population mobility, urbanization, and the growing prevalence of late marriage and non-marriage.

The study's use of machine learning models offers an innovative approach to understanding the predictors of marriage rates. CatBoost, in particular, outperformed other models, suggesting that it is better suited to capturing non-linear relationships and regional variability in marriage rates. However, the performance of all machine learning models demonstrated some limitations in accurately predicting extreme values, particularly in provinces with either very high or very low marriage rates. The causal

inference results also highlight the importance of considering heterogeneity in socioeconomic factors, as their impact on marriage rates varies across different regions.

Furthermore, the study's integration of panel data models with machine learning techniques demonstrates the benefits of combining traditional econometric approaches with advanced computational methods to analyze complex social phenomena. By comparing the effectiveness of these models, the research offers a comprehensive framework for analyzing marriage rates, which could be adapted to other regions or social trends in future studies.

4.8 Summary

The results of this research provide valuable insights into the shifting marriage trends in China, revealing the intricate relationship between socioeconomic factors and the crude marriage rate. The analysis highlights that the marriage rate in China has undergone a notable shift, rising significantly from 2003 to 2012 and subsequently experiencing a sharp decline by 2022. The study identifies that this fluctuation is largely driven by rapid economic development, rising housing prices, and evolving cultural attitudes towards marriage, particularly in eastern coastal regions like Shanghai, Zhejiang, and Jiangsu.

Among the models applied, the Random Effects model emerged as the most effective traditional regression approach, explaining the highest proportion of variance in marriage rates across provinces. However, in the machine learning analysis, CatBoost outperformed other models, including XGBoost, LightGBM, and GBDT, particularly in terms of MSE, RMSE, and R^2 , demonstrating its robustness in capturing the complex relationships between the independent variables and the marriage rate.

The results of causal inference using DML further underscore the profound impact of birth rate, average years of education, and female population on the crude marriage rate, with significant heterogeneity observed across different regions. The analysis also reveals that housing prices and gross dependency ratio exert a consistently negative effect on the marriage rate, with house prices showing the strongest negative impact across various regions. Overall, the machine learning models

demonstrate superior predictive power compared to traditional statistical approaches, but also highlight areas for improvement in handling regional variability.

Table 4.6 shows Model Evaluation Index Results. After comparing different models, it can be concluded that the Random Effects model performed the best across all evaluation metrics (MSE, RMSE, MAE), demonstrating the advantage of traditional statistical models on this dataset. Although CatBoost performed relatively well among the machine learning models, its overall error was still higher than that of the Random Effects model, with XGBoost and GBDT showing larger errors. This indicates that, in this specific dataset, traditional statistical models outperform more complex machine learning models, highlighting the importance of optimizing model selection based on the characteristics of the data.

Table 4.6 The Summary of Model Results

The Summary of Model Results			
	MSE	RMSE	MAE
Pooled OLS	2.436	1.560	1.410
Random Effects	1.661	1.288	1.088
Fixed Effects	4.749	2.179	1.847
XGBoost	4.070	2.017	1.634
LightGBM	3.862	1.965	1.536
CatBoost	3.534	1.880	1.539
GBDT	4.425	2.103	1.689

CHAPTER 5

CONCLUSIONS

5.1 Conclusions

In recent years, China has experienced significant changes in its marriage rate, a phenomenon influenced by a complex interplay of socioeconomic factors. This study aims to investigate the factors that influence the crude marriage rate (CMR) across different regions of China over a span of 20 years, focusing on the period from 2003 to 2022. By examining the socioeconomic variables such as GDP, housing prices, gross dependency ratio, birth rate, female population, average years of education, and sex ratio, the research endeavors to reveal the underlying drivers behind the fluctuations in marriage rates. The analysis involves both traditional statistical models like Pooled OLS, Random Effects, and Fixed Effects, and machine learning methods including XGBoost, LightGBM, CatBoost, and GBDT. These methods are employed to predict the marriage rate and compare the effectiveness of each model through cross-validation techniques and various evaluation metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2).

Given the dramatic rise and subsequent fall in marriage rates, particularly in the economically developed eastern provinces, this research delves into the regional differences and seeks to explore the causes behind the significant trends in marriage behavior in China. Additionally, causal inference using machine learning is applied to investigate the Average Treatment Effect (ATE), Conditional Average Treatment Effect (CATE), and Heterogeneous Treatment Effect (HTE) of the socioeconomic factors on the crude marriage rate, allowing for a more nuanced understanding of these relationships.

Given The findings of this study reflect the broader social and economic transformation occurring in China over the past two decades. The observed decline in marriage rates, particularly in 2022, is indicative of deeper societal shifts, including the rising costs of living, increasing housing prices, and changing cultural norms around marriage and family life. Regions with higher economic development, such as the eastern coastal areas, have witnessed the most dramatic declines, driven by factors such as population mobility, urbanization, and the growing prevalence of late marriage and non-marriage.

The study's use of machine learning models offers an innovative approach to understanding the predictors of marriage rates. CatBoost, in particular, outperformed other models, suggesting that it is better suited to capturing non-linear relationships and regional variability in marriage rates. However, the performance of all machine learning models demonstrated some limitations in accurately predicting extreme values, particularly in provinces with either very high or very low marriage rates. The causal inference results also highlight the importance of considering heterogeneity in socioeconomic factors, as their impact on marriage rates varies across different regions.

Furthermore, the study's integration of panel data models with machine learning techniques demonstrates the benefits of combining traditional econometric approaches with advanced computational methods to analyze complex social phenomena. By comparing the effectiveness of these models, the research offers a comprehensive framework for analyzing marriage rates, which could be adapted to other regions or social trends in future studies.

5.2 Suggestions

In summary, addressing the decline in marriage rates requires a multifaceted approach that considers economic, social, and cultural factors. By leveraging advanced machine learning models and further refining regional analyses, future research can contribute to more effective policy interventions and a deeper understanding of marriage trends in China.

5.2.1 Policy Recommendations

Given the findings of this research, several key suggestions can be made for policymakers and future researchers.

5.2.1.1 Housing Affordability

Rising housing prices are a major deterrent to marriage in economically developed regions. Policymakers should consider measures such as subsidized housing for young couples, rent control policies, and increased support for affordable housing projects, particularly in high-demand urban areas.

5.2.1.2 Education and Gender Equality

The study suggests that education and female population dynamics influence marriage patterns. Policies that promote gender equality in employment and provide career support for women could help alleviate the pressures that come with educational and career pursuits, which are often seen as barriers to marriage.

5.2.1.3 Targeted Regional Policies

Since the socioeconomic determinants of marriage rates vary significantly across regions, localized policies that address specific regional needs would be more effective. For instance, less developed regions may benefit from economic incentives to encourage earlier marriages, while urbanized areas might focus on reducing the financial burden associated with marriage and family formation.

5.2.2 Future Research Directions

Future research directions include Incorporate More Socioeconomic Variables, Exploration of Divorce and Birth Rates, Longitudinal and Dynamic Modeling.

5.2.2.1 Incorporate More Socioeconomic Variables

Future studies could benefit from including additional variables such as migration patterns, unemployment rates, healthcare access, and government family policies. These factors are likely to have significant impacts on marriage rates but were not included in this study.

5.2.2.2 Exploration of Divorce and Birth Rates

Extending the research to include divorce and birth rates would provide a more comprehensive understanding of family formation and dissolution trends in China. Analyzing how these rates interact with marriage trends could offer additional

insights for policymakers concerned with the broader implications of declining marriage rates.

5.2.2.3 Longitudinal and Dynamic Modeling

The Time-series analysis or dynamic panel models could be used in future studies to forecast future marriage trends and examine how past trends influence future behaviors. This would help policymakers better anticipate demographic shifts and plan accordingly.

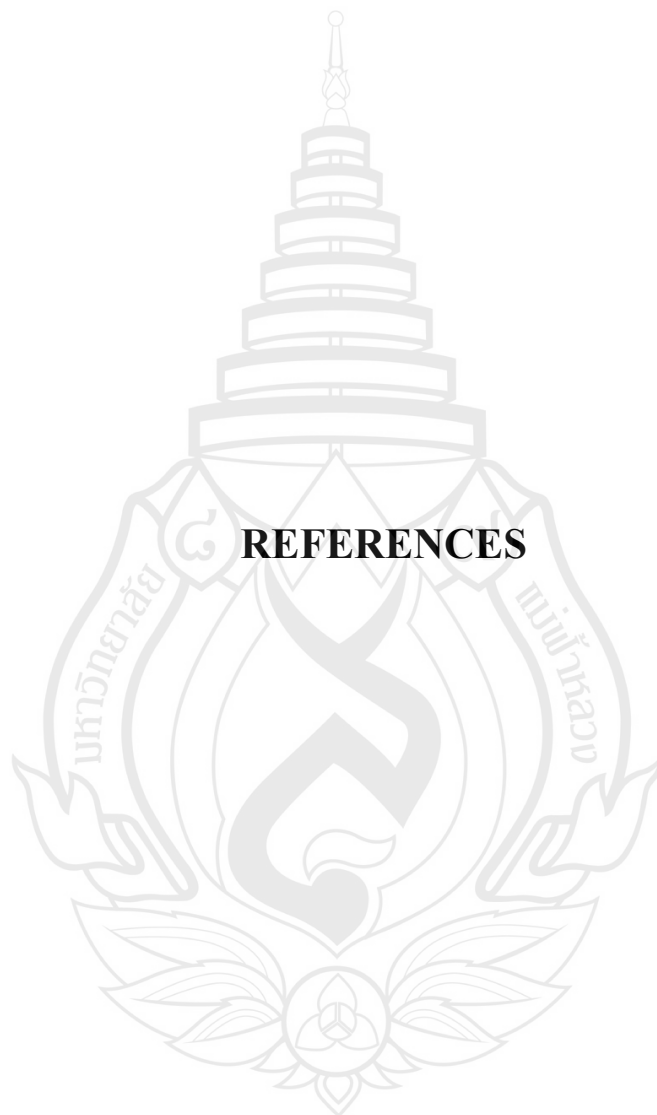
5.2.3 Model Enhancements

5.2.3.1 Neural Networks and Deep Learning

The introduction of neural network-based models could offer improved accuracy in capturing complex, non-linear relationships in the data, especially in regions with extreme variations in marriage rates.

5.2.3.2 Region-Specific Machine Learning Models

Tailoring machine learning models to specific regions might enhance prediction accuracy, given the significant variability in the factors driving marriage rates across China. Additionally, ensemble methods combining traditional econometric models with machine learning could offer a hybrid approach for more reliable prediction.



REFERENCES

REFERENCES

- Bergstrom, T., & Lam, D. (1991). *The two-sex problem and the marriage squeeze in an equilibrium model of marriage market*. Center for Research on Economic & Social Theory.
- Blossfeld, H., & Jaenichen, U. (1992). Educational expansion and changes in women's entry into marriage and motherhood in the Federal Republic of Germany. *Journal of Marriage and Family*, 54(2), 302-315.
<https://doi.org/10.2307/353062>
- Bresson, G., & Chaturvedi, A. (2023). Dynamic space–time panel data models: An eigendecomposition-based bias-corrected least squares procedure. *Spatial Statistics*, 56, 100758. <https://doi.org/10.1016/j.spasta.2023.100758>
- Call, V. R. A., & Heaton, T. B. (1997). Religious influence on marital stability. *Journal for the Scientific Study of Religion*, 36(3), 382-392.
<https://doi.org/10.2307/1387856>
- Campbell, J. Y., & Cocco, J. F. (2007). How do house prices affect consumption? Evidence from micro data. *Journal of Monetary Economics*, 54(3), 591–621.
<https://doi.org/10.1016/j.jmoneco.2005.10.016>
- Chaurasia, A., & Haq, I. U. (2023). Housing price prediction model using machine learning. In *2023 International Conference on Sustainable Emerging Innovations in Engineering and Technology (ICSEIET)*, Ghaziabad, India, 2023 (pp. 497-500). IEEE.
- Chen, J., & Pan, W. (2023). Bride price and gender role in rural China. *Heliyon*, 9(1), e12789. <https://doi.org/10.1016/j.heliyon.2022.e12789>

- Chen, K., & Wen, Y. (2017). The great housing boom of China. *American Economic Journal: Macroeconomics*, 9(2), 73-114.
<https://doi.org/10.1257/mac.20140234>
- Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*.
<https://dl.acm.org/doi/pdf/10.1145/2939672.2939785>
- Chetty, R., Sándor, L., & Szeidl, Á. (2017). The effect of housing on portfolio choice. *The Journal of Finance*, 72(3), 1171–1212. <https://doi.org/10.1111/jofi.12500>
- Chiappori, P., Fortin, B., & Lacroix, G. (2002). Marriage market, divorce legislation, and household labor supply. *Journal of Political Economy*, 110(1), 37–72.
<https://doi.org/10.1086/324385>
- Chiplunkar, G., & Weaver, J. (2023). Marriage markets and the rise of dowry in India. *Journal of Development Economics*, 164, 103115.
<https://doi.org/10.1016/j.jdeveco.2023.103115>
- Cohrs, K., Varando, G., Carvalhais, N., Reichstein, M., & Camps-Valls, G. (2024). Causal hybrid modeling with double machine learning - Applications in carbon flux modeling. *Machine Learning Science and Technology*, 5(3), 035021. <https://doi.org/10.1088/2632-2153/ad5a60>
- Corradin, S., & Popov, A. (2015). House prices, home equity borrowing, and entrepreneurship. *The Review of Financial Studies*, 28(8), 2399–2428.
<https://doi.org/10.1093/rfs/hhv020>
- Deng, Y., Gyourko, J., & Wu, J. (2012). *Land and house price measurement in China*. National Bureau of Economic Research. <https://doi.org/10.3386/w18403>
- Desk, C., & Zaobao, L. (2023). *China is cracking down on exorbitant bride price rates to save marriages*. <https://www.thinkchina.sg/society/china-cracking-down-exorbitant-bride-price-rates-save-marriages>

- Detting, L., & Kearney, M. S. (2014). House prices and birth rates: The impact of the real estate market on the decision to have a baby. *Journal of Public Economics*, 110, 82–100. <https://doi.org/10.1016/j.jpubeco.2013.09.009>
- Ding, Y. (2020). *Some strategies for machine learning projects: Error analysis can help you improve system performance*. <https://towardsdatascience.com/some-strategies-for-machine-learning-projects-5f2f32c34635>
- Edlund, L., Li, H., Yi, J., & Zhang, J. (2013). Sex ratios and crime: Evidence from China. *The Review of Economics and Statistics*, 95(5), 1520–1534. https://doi.org/10.1162/REST_a_00356
- Fuhr, J., Berens, P., & Papies, D. (2024). *Estimating causal effects with double machine learning -- A method evaluation*. <https://ideas.repec.org/p/arx/papers/2403.14385.html>
- Greenwood, J., Guner, N., Kocharkov, G., & Santos, C. (2014). Marry your like: assortative mating and income inequality. *The American Economic Review*, 104(5), 348–353. <https://doi.org/10.1257/aer.104.5.348>
- Guggenberger, P. (2009). The impact of a Hausman pretest on the size of a hypothesis test: The panel data case. *Journal of Econometrics*, 156(2), 337–343. <https://doi.org/10.1016/j.jeconom.2009.11.003>
- Gupta, R., Sharma, A., Anand, V., & Gupta, S. (2022). Automobile Price Prediction using Regression Models. *2022 International Conference on Inventive Computation Technologies (ICICT)*. <https://doi.org/10.1109/iciict54344.2022.9850657>
- Harding, J. P., & Rosenthal, S. S. (2017). Homeownership, housing capital gains and self-employment. *Journal of Urban Economics*, 99, 120–135. <https://doi.org/10.1016/j.jue.2016.12.005>

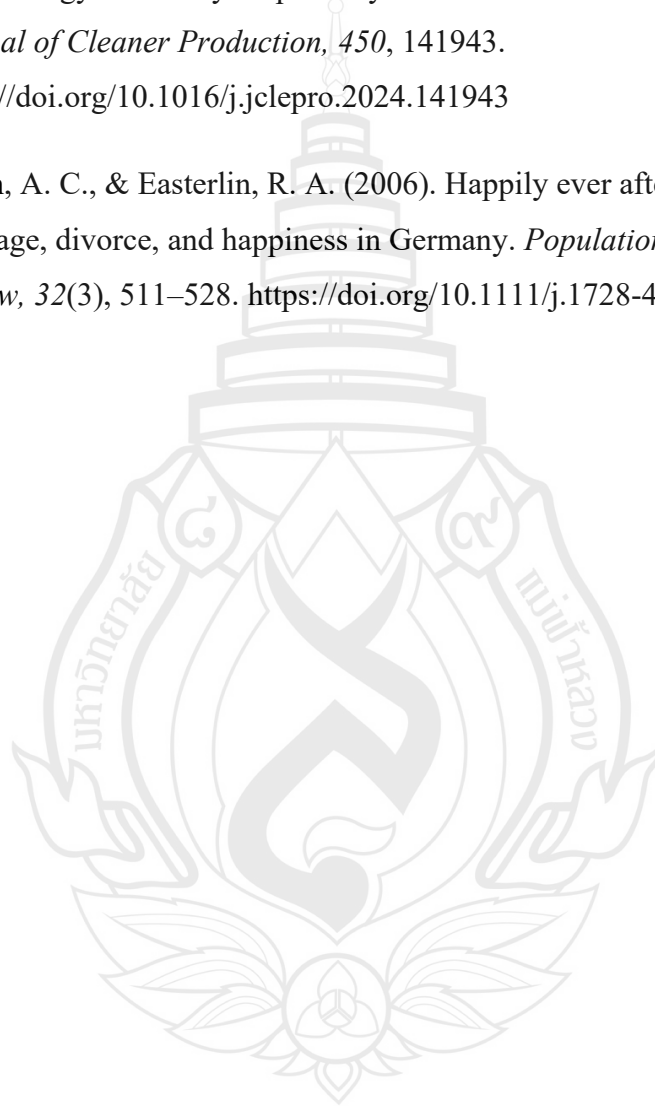
- Hu, M., Wu, L., Xiang, G., & Zhong, S. (2021). Housing prices and the probability of marriage among the young: evidence from land reform in China. *International Journal of Emerging Markets*, 18(2), 420–438.
<https://doi.org/10.1108/IJOEM-09-2020-1116>
- Huang, Y., Leung, C. H., Wu, Q., Yan, X., . . . Huang, Z. (2022). Robust causal learning for the estimation of average treatment effects. In *2022 International Joint Conference on Neural Networks (IJCNN)*.
<https://doi.org/10.1109/ijcnn55064.2022.9892344>
- Johnson, W. R. (2014). House prices and female labor force participation. *Journal of Urban Economics*, 82, 1–11. <https://doi.org/10.1016/j.jue.2014.05.001>
- Ke, G., Meng, Q., Finley, T., Wang, T., . . . Liu, T.-Y. (2017). *LightGBM: A highly efficient gradient boosting decision tree*. https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf
- Kecojevic, T. (2020). *Machine learning*. http://dataliteracy.rbind.io/module4/what_is_ml/
- Khan, S. (2024). Female education and marriage in Pakistan: The role of financial shocks and marital customs. *World Development*, 173, 106413.
<https://doi.org/10.1016/j.worlddev.2023.106413>
- Khan, S., Klasen, S., & Pasha, A. (2020). *Asset ownership and female empowerment: Evidence from a natural experiment in Pakistan*. https://edi.opml.co.uk/wp-content/uploads/2020/10/EDI_Klasen_October2020-1.pdf
- Kumar, A., Dodda, S., Kamuni, N., & Arora, R. K. (2024, March 31). *Unveiling the impact of macroeconomic policies: A double machine learning approach to analyzing interest rate effects on financial markets*.
<https://arxiv.org/abs/2404.07225>

- Li, B., Jansen, S. J., Van Der Heijden, H., Jin, C., & Boelhouwer, P. (2022). Unraveling the determinants for private renting in metropolitan China: An application of the Theory of Planned Behavior. *Habitat International*, 127, 102640. <https://doi.org/10.1016/j.habitatint.2022.102640>
- Li, Q., Liang, D., & Zhang, L. (2023). Have pensions reduced the relative poverty? --- -- empirical analysis from China CHARLS data. *Heliyon*, 9(12), e22711. <https://doi.org/10.1016/j.heliyon.2023.e22711>
- Lovenheim, M. (2011). The effect of liquid housing wealth on college enrollment. *Journal of Labor Economics*, 29(4), 741–771. <https://doi.org/10.1086/660775>
- Lovenheim, M., & Mumford, K. J. (2013). Do family wealth shocks affect fertility choices? evidence from the housing market. *The Review of Economics and Statistics*, 95(2), 464–475. https://doi.org/10.1162/REST_a_00266
- Lovenheim, M., & Reynolds, C. L. (2013). The effect of housing wealth on college choice: Evidence from the housing boom. *Journal of Human Resources*, 48(1), 1–35. <https://doi.org/10.1353/jhr.2013.0001>
- Malhotra, A. (1997). Gender and the timing of marriage: Rural-urban differences in Java. *Journal of Marriage and Family*, 59(2), 434-450. <https://doi.org/10.2307/353481>
- Mian, A., & Sufi, A. (2014). What explains the 2007-2009 drop in employment?. *Econometrica*, 82(6), 2197–2223. <https://doi.org/10.3982/ECTA10451>
- Nie, G. (2020). Marriage squeeze, marriage age and the household savings rate in China. *Journal of Development Economics*, 147, 102558. <https://doi.org/10.1016/j.jdeveco.2020.102558>
- Oppenheimer, V. K. (1988). A theory of marriage timing. *American Journal of Sociology*, 94(3), 563–591. <https://doi.org/10.1086/229030>

- Oppenheimer, V. K. (1994). Women's rising employment and the future of the family in industrial societies. *Population and Development Review*, 20(2), 293-342. <https://doi.org/10.2307/2137521>
- Piketty, T., & Zucman, G. (2014). Capital is back: Wealth-income ratios in rich countries 1700–2010. *The Quarterly Journal of Economics*, 129(3), 1255–1310. <https://doi.org/10.1093/qje/qju018>
- Pilar Alonso, M., Gargallo, P., Lample, L., López-Escolano, C., Miguel, J. A., & Salvador, M. (2024). Measuring the relationship between territorial exclusion and depopulation – A municipal classification proposal to guide territorial balance. *Journal of Rural Studies*, 111, 103421. <https://doi.org/10.1016/j.jrurstud.2024.103421>
- Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Neural Information Processing Systems*, 31, 6639–6649.
- Ratnasari, V., Audha, S. H., & Dani, A. T. R. (2023). Statistical modeling to analyze factors affecting the middle-income trap in Indonesia using panel data regression. *MethodsX*, 11, 102379. <https://doi.org/10.1016/j.mex.2023.102379>
- Ryder, N. B. (1964). The process of demographic translation. *Demography*, 1, 74-82. <https://doi.org/10.1007/BF03208446>
- Shanghai Survey Team of National Bureau of Statistics. (2009). *Average years of education per capita*. <https://tjj.sh.gov.cn/zcjd/20091102/0014-86153.html>
- Sharma, S., Arora, D., Shankar, G., Sharma, P., & Motwani, V. (2023). House price prediction using machine learning algorithm. In *2023 7th International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 982-986). IEEE.

- Shyam, R., Ayachit, S. S., Patil, V., & Singh, A. (2020). Competitive analysis of the top gradient boosting machine learning algorithms. In *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*. IEEE.
<https://doi.org/10.1109/icacccn51052.2020.9362840>
- Tian, L., Yan, Y., Prof, G. C. L., Wu, Y., & Shao, L. (2020). Breaking the land monopoly: Can collective land reform alleviate the housing shortage in China's mega-cities?. *Cities*, *106*, 102878.
<https://doi.org/10.1016/j.cities.2020.102878>
- Wang, J., Yang, P., Zhang, L., & Hou, X. (2021). A low-FODMAP diet improves the global symptoms and bowel habits of adult IBS patients: A systematic review and meta-analysis. *Front Nutr.*, *19*(8), 683191.
<https://doi.org/10.3389/fnut.2021.683191>
- Wang, Y., Yu, Y., & Su, Y. (2017). Does the tender, auction and listing system in land promote higher housing prices in China?. *Housing Studies*, *33*(4), 613–634. <https://doi.org/10.1080/02673037.2017.1373750>
- Wrenn, D. H., Yi, J., & Zhang, B. (2019). House prices and marriage entry in China. *Regional Science and Urban Economics*, *74*, 118–130.
<https://doi.org/10.1016/j.regsciurbeco.2018.12.001>
- Yang, P. (2014). *Understanding vocational education market in China*.
https://www.researchgate.net/figure/Chinese-Education-System_fig2_337589737
- Ye, Z., Geng, Y., Chen, J., Chen, J., & Chen, H. (2020). Zero-shot text classification via reinforced self-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 3014–3024). Association for Computational Linguistics.
- Zhang, C. (2015). Income inequality and access to housing: Evidence from China. *China Economic Review*, *36*, 261–271.
<https://doi.org/10.1016/j.chieco.2015.10.003>

- Zhao, C., Chen, B., & Li, X. (2023). Rising housing prices and marriage delays in China: Evidence from the urban land transaction policy. *Cities*, 135, 104214. <https://doi.org/10.1016/j.cities.2023.104214>
- Zhao, X., Zeng, B., Zhao, X., Zeng, S., & Jiang, S. (2024). Impact of green finance on green energy efficiency: A pathway to sustainable development in China. *Journal of Cleaner Production*, 450, 141943. <https://doi.org/10.1016/j.jclepro.2024.141943>
- Zimmermann, A. C., & Easterlin, R. A. (2006). Happily ever after? Cohabitation, marriage, divorce, and happiness in Germany. *Population and Development Review*, 32(3), 511–528. <https://doi.org/10.1111/j.1728-4457.2006.00135.x>





CURRICULUM VITAE

CURRICULUM VITAE

NAME Deyu Zhang

EDUCATIONAL BACKGROUND

2016 Bachelor of Arts Animation
Southwest Forestry University, China

WORK EXPERIENCE

2018-2021 Yunnan Technology and Business University
Yunnan, China

2018-2019 Sunwoda Electronic Co., Ltd.
Huizhou, China

2016-2021 Yunnan College of Foreign Affairs & Foreign
Language
Yunnan, China

PUBLICATION

- Zhang, D., Rueangsirarak, W., Uttama, S., Zhao, X., Long, W., & Zhao, Y. (2024). Using machine learning models to evaluate and predict STEM postgraduates' graduation rates: A case study of China. In *2024 9th International STEM Education Conference (iSTEM-Ed)*, Cha-am, Hua Hin, Thailand, 2024 (pp. 1-6). IEEE. <https://doi.org/10.1109/iSTEM-Ed62750.2024.10663160>
- Zhang, D., Rueangsirarak, W., & Uttama, S. (2024). Predicting China's marriage rate: Causal inference using Dual Machine Learning (DML) with XGBoost, LightGBM, CatBoost, and GBDT, In *2024 5th International Conference on Big Data Analytics and Practices (IBDAP)*, Bangkok, Thailand, 2024 (pp. 73-78). IEEE. <https://doi.org/10.1109/IBDAP62940.2024.10689698>