An irreversible trend: what is suppressing fertility?

Yihang Wang, Shuang Du

Introduction

For decades, the birth rate has been a long-standing problem among major countries and societies. While being harsh among merely all the developed countries, the situation of birth rate looks far more optimistic among various less-developed or developing countries, showing a significant regional disparity. Given these facts, an interesting common sense is, the relatively low birth rates of developed countries and the high birth rates of developing countries are both generally considered as an impedance against economic growth. The analysis of birth rate has been attached to great importance for years. Up to now, various studies of birth rates have been sponsored by governments and NGOs, which were expected to provide information for policy making.

Across the world, the birth rate of China has been somehow a special case of study. First and foremost, China has gone through rapid economic growth for over 4 decades to become the second-largest economic entity. On the other hand, the GDP per capita of China has just increased to 10k USD by 2020, merely a half of the threshold to be considered as a developed country. Also, during these 40 years, the family planning policy had a significant change from the well-known “one-child policy” to virtually no limitation. That said, the case of China may be taken as neither a typical developed nor a typical developing country. According to the latest report, there's been a significant drop in the birth rate of China, in the past year, which has aroused widespread concern together with the reality of an aging population. It is widely held as the mainstream opinion that the most important reason for the dramatic decrease in birth rates is urbanization and the modern lifestyle. In other words, the sharp decrease in fertility is mainly because of the rapid growth of the economy and industrialization, thus irreversible. Also, the popularization of high-level education, especially women's education, accounts for this decrease. However, there are also other points of view that the rising living expenses, workload, and the cost of child-raising are to blame for the fall of the fertility rate.

In this project, to statistically illustrate the change of birth rate and the reasons behind, multivariate regressions are to be conducted among a variety of possible factors (as mentioned above: education, industrialization, disposable income, housing price, medical level, working hours, women's rights, urbanization, and industrialization level, etc.). Based on the acquired data from public sources, herein, we chose life expectancy, university recruitment, urban population, mortality, Consumer Price Index as related variables to be considered. Statistical methods like ridge regression, LASSO regression, or PCA will be conducted to eliminate the redundant variables. The difference between different models will be compared. Also, to test the validity of our model, similar linear regression will be conducted based on the data of different countries or societies. We hope this project will bring some insights into the influencing factors of fertility and provide possible forecasts or advice for future policies.

Data Collection

All data was collected from public official resources (World Bank, National Bureau of Statistics, etc.). Due to the limitation of accessible data, we would use the following indicators to build the model: fertility (total birth per woman), life expectancy at birth (in years), university recruitment per 10k, urban population ratio, infant mortality (death per 10k before 3 weeks old), disposable income (in CNY), Consumer Price Index (CPI, 2010 standardized) were taken into account.

Variable	Definition	Explanation
lifexpect	life expectancy	the average period that a person may expect to live
inf.mortality	infant mortality	the number of infant deaths for every 1,000 live births

univ.recruit	university recruitment ratio	the percentage of female population (>18y) who received college educations
urb.popul	urban population ratio	urban population (% of total population)
income.CPI	Household Disposable Income / CPI(base=2008)	the consumption parenting capability

Table 1. Variables and explanations.

The life expectancy (denoted as *lifexpect*) and infant mortality (denoted as *inf.mortality*) were considered as indicators of medical resource per capita, while infant mortality also stands for reproductive risk to some extent. We used university recruitment ratio (denoted as *univ.recruit.*) to measure the education level of young people, who played an essential role in birth rate. The reason why we chose tertiary education rather than others (such as secondary education) was we thought university education had a more significant effect on shaping women's independence awareness. Urban population ratio (denoted as *urb.popul.*) measured the level of modernization, i.e., how many people have been involved in a modern lifestyle. Standardized CPI was used to measure the comprehensive living expenses, coupled with the personal income to describe the consumption parenting capability. Specifically, we used income divided by CPI as the variable in this model (denoted as *income.CPI*).

Linear Model Analysis

Based on the data as prepared, a variety of linear models were built up under an environment of R studio. We first attempted to build up a linear model as follows:

$$\textit{fertility} \sim \textit{lifexpect} + \textit{univ.recruit.} + \textit{urb.popul.} + \textit{inf.mortality} + \textit{income.CPI}$$

By looking at the model lmfull summary, the R-squared value of 0.8962 is not bad for cross-sectional data of 34 observations. The F-value is highly significant implying that all the explanatory variables together significantly explain fertility. However, coming to the individual regression coefficients, it is seen that the variable lifexpect is not statistically significant. Further, we can plot the model diagnostic checking for other problems such as normality of error term, heteroscedasticity, etc.

The linear regression model also shows that the *lifexpect* variable showed a p-value of 0.69496. Except for it, other variables are all considered significantly correlated with the response. A q-q plot and Shapiro-Wilk test are conducted to test the normality of the residuals, which turn out to be good (p-value = 0.5802).

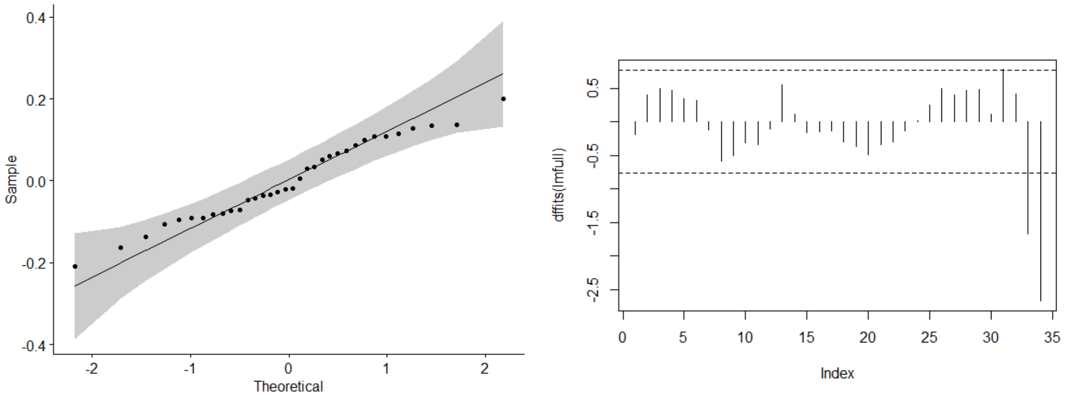


Figure 1. a. (left) The q-q plot of *lmfull*; b. (right) The DFFITS plot of *lmfull* .

Tests for outliers are conducted subsequently, with the help of the *dffits* embedded in R and *outlierTest* from the *car* package. Data No. 35 is pointed out as a high-influential point and NO.34 is pointed out by both methods. In fact, No. 31 surpasses the threshold of *dffits* a little but we decide to ignore this. As a result, No.34 and 35 are excluded from the linear models by creating a new dataset *Fertilitynew*. All the following linear models are based on this new dataset or its subsets.

Additionally, as a multivariable model, the problem of multicollinearity needs to be considered. Based on the variance inflation factors (VIFs), an obvious collinearity is expectedly found in *lmfull*. All the variables have a VIF larger than 10, while three of the VIFs are larger than 100. We also conduct a pairwise correlation test, the result shows that most of the variables are strongly correlated. To fix this problem, a variety of methods would be applied for the variable selection.

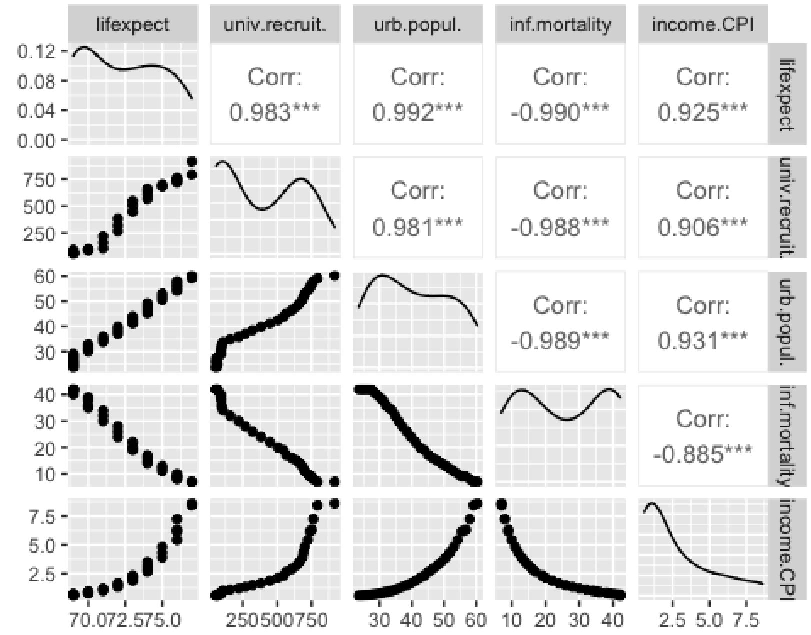


Figure 2. The correlation analysis of variables.

Model Selection

To evaluate the performance of different models, we created subsets *fertilitytrain* as the training set and *fertilitytest* as the testing set, with a splitting ratio of 5:1 based on 32 data points.

Full Model: The full model without variable selection based on *fertilitytrain* was almost the same as *lmfull*. Based on its summary, *lifexpect* and *inf.mortality* both had a p-value larger than 0.05, indicating a poor significance.

Backward Selection: Backward selection generated a new model *lmback2* as follows:

fertility~ univ.recruit.+urb.popul.+income.CPI

The normality of residuals was proved with the same methods. Two variables had a VIF larger than 10 (22.5 for *univ.recruit.* and 30.8 for *urb.popul.*), which indicated there was still a correlation between the three existing variables.

AIC: Apart from backward selection, standard based selection was also conducted. Based on the Akaike Information Criteria (AIC), a new model *lmAIC* was given as below:

fertility~univ.recruit.+urb.popul.+inf.mortality+income.CPI

This model also showed acceptable normality (p-value = 0.1521) but did not fix the problem of collinearity as we had expected.

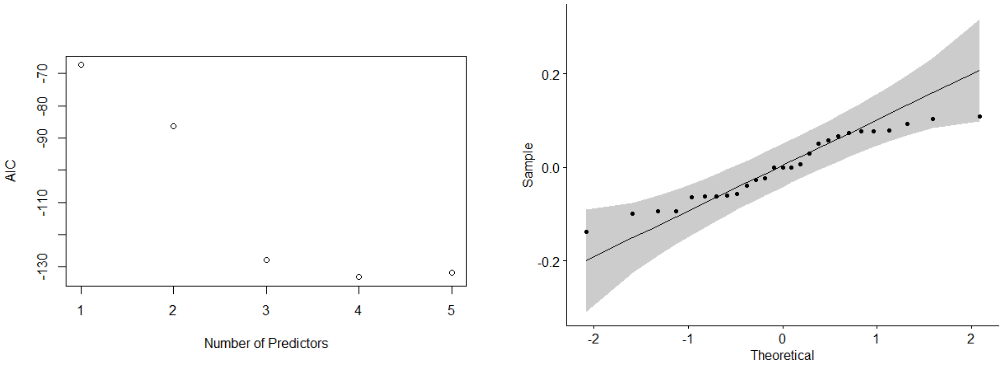
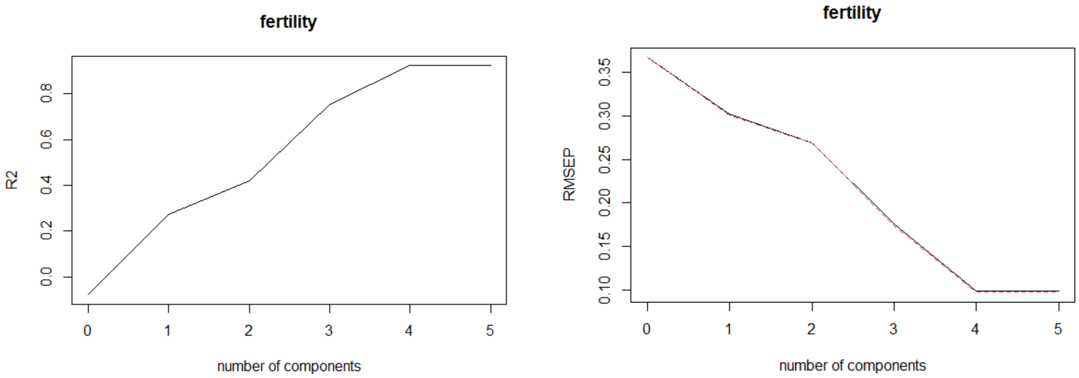


Figure 3. a. (left) The AIC plot of *lmAIC*; b. (right) The q-q plot of *lmAIC*.

PCR & PLSR: Based on the R^2 and root mean square error percentage (RMSEP), the result indicates that predictions should be derived from the model with 4 principal components. The PLSR method gave out a very similar outcome (not presented here). Based on the mechanism (diagonalization of X^TX) of PCR and PLSR, the problem of collinearity was fixed. However, both models (*lmpcr* and *lmplsr*) didn't pass the tests of normality, with p-values of 0.003526 and 0.002228 respectively.



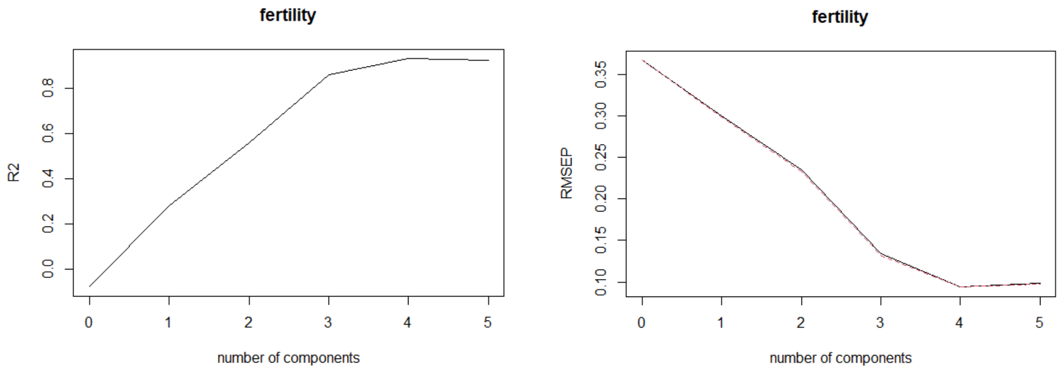


Figure 4. a. (upper left) The R^2 vs. ncomp of *lmpcr*; b. (upper right) The RMSEP vs. ncomp of *lmpcr*; c. (lower left) The R^2 vs. ncomp of *lmplsr*; d. (lower right) The RMSEP vs. ncomp of *lmplsr*.

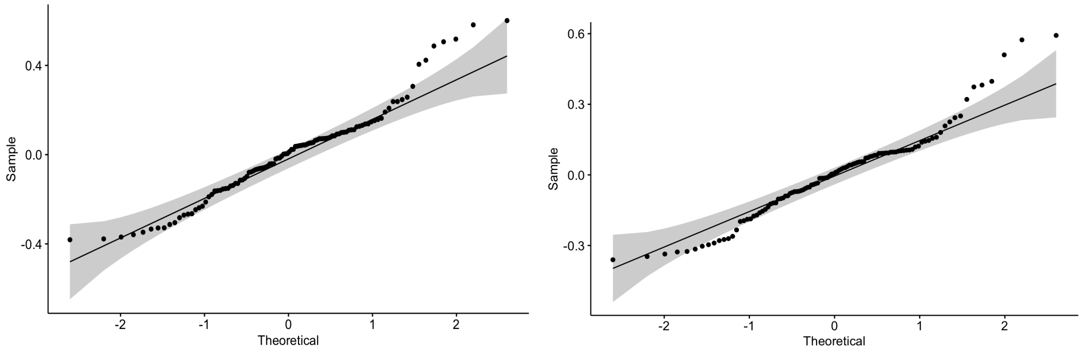


Figure 5. a. (left) The q-q plot of *lmpcr*; b. (right) The q-q plot of *lmplsr*.

Ridge Regression & LASSO: Then, two new models (*lmridge* and *lmlasso*) based on ridge regression and LASSO regression were built up. The lambda values were determined by the cross validation (CV) method, which both turned out to be 0.01. Unlike ridge regression, in *lmlasso*, *lifexpect* and *inf.mortality* were automatically omitted for insignificance, while the coefficient of *univ.recruit*. was much lower than before (around 0.01 now). Both models have fitted the requirement of normality, while *lmlasso* had a larger p-value of 0.2387 than 0.08765 of *lmridge.*

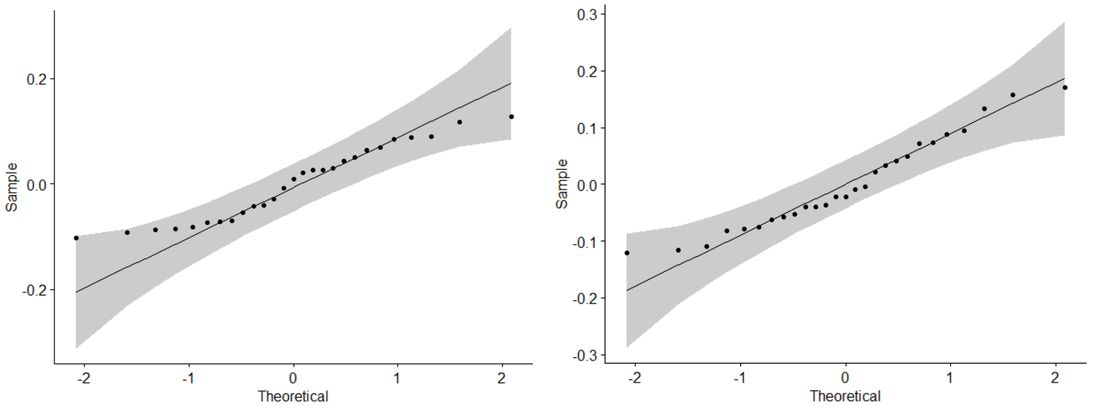


Figure. 5. a. (left) The q-q plot of *lmridge*; b. (right) The q-q plot of *lmlasso*.

Conclusion: For model performance evaluation, we utilized the RMSE, which is the standard deviation of the residuals (prediction errors), to measure the performance for each model:

Method	**RMSE**
Full Model	0.06446588
Backward Selection	0.04200628
AIC	0.06129241
PCR	0.07551262
PLS	0.07124705
Ridge Regression	0.09737348
LASSO	0.07385271

Table 2. RMSE comparison of different models.

The RMSE result shows that Backward Selection had a relatively better performance. But given the situation that we need to fix the collinearity and satisfy the normality of residuals, only LASSO and ridge regression models should be acceptable. We can see that LASSO regression has a better performance than Ridge Regression.

We also tried to conduct PCR or PLSR based on model *lmback2*, to see whether it had fixed the problem of collinearity. However, the resulting models did not show good normality.

The overall analysis of China's fertility dataset reveals that

1. To solve the collinearity problem with the original dataset, we introduced several methods such as Backward Selection, AIC, PCR, PLS, Ridge Regression, LASSO regression. Using RMSE as a prediction criterion and considering residual normality, we concluded that LASSO regression had the best performance.
2. Variables *univ.recruit*, *urb.popul*, *income.CPI* all have significant influences on the fertility rate in China during the 1986-2016 period.
   The final prediction model under LASSO regression is :
   fertility = 4.233 + 0.001univ.recruit.-0.085urb.popul.+0.217income.CPI

The model shows that China's fertility rate has positive relationships with university recruitment rate and household financial status, where the financial status has a stronger impact on the fertility rate. We can also see that the fertility rate has a negative relationship with the urban population ratio. It indicates that urbanization has reduced the fertility rate, while the potential reason could be that urban residents would likely have higher work pressure, less leisure time, and higher child-raising costs.

Application

In this section, we will apply the same methods above to Japan's fertility dataset. The purpose is to find out if the contributing factors on China's fertility rate had the same effects on Japan's fertility rate. We chose Japan as the application dataset for the following reasons:

1. Japan and China have a close cultural background which could limit the influence on cultural differences.
2. Japan is known as a typical aging society with an extremely low birth rate. As China is now facing the aging problem as a result of the previous Birth Control

Policy, we expect to gain some insights into some potential factors on improving the fertility rate when China becomes an aging society one day.

As a developed country, Japan has superior advantages in economics, social welfare, and education than China. This background could result in a difference in the final model representations. However, we would like to use the prediction result of Japan as a forecast of China's future fertility rate.

Due to the limitation of data accessibility, in Japan's dataset, life expectancy (*lifexpect*), infant mortality (*inf.mortality*), urban population rate (*urb.popul.*), and fertility rate (*fertility*) were collected under the same criteria as in China's dataset. Meanwhile, we chose university recruitment gross rate (*univ.recruit.*, in %) instead of university recruitment per 10k, and household disposable income (HHDI) instead of personal disposable income to give the standardized *HHDI.CPI*.

Linear Model Analysis

We first tried the linear regression on the new dataset with the *lmfull* model:

fertility~lifexpect+urb.popul.+fem.enroll.+inf.mortality+HHDI.CPI

The model summary shows an adjusted R square at 0.3662, which indicates the variables may not sufficiently interpret the fertility rate. We can slo see from the result that three variables (*lifexpect, urb.popul., and inf.mortality)* are not statistically significant under 95% confidence interval.

A q-q plot and Shapiro-Wilk test are used to test the normality of the residuals.The result shows a p value at 0.4817, which is higher than 0.05, indicating a high significance in normality of residuals. As the same method used above, the *dffits* embedded in R and *outlierTest* from the *car* package. Data No. 20 is pointed out as a high-influential point. As a result, No.20 is excluded from the linear models by creating a new dataset *Fertilitynew*. All the following linear models are based on this new dataset or its subsets.

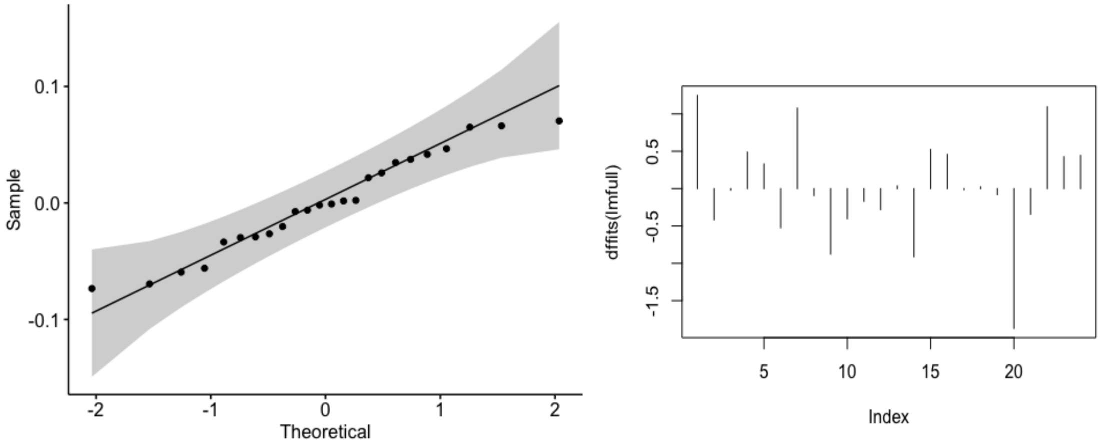


Figure 6. a. (left) The q-q plot of *lmfull*; b. (right) The DFFITS plot of *lmfull*.

The variance inflation factors (VIFs) method indicates obvious collinearity between variables in Japan’s fertility dataset, where most of the variables have a VIF larger than 10. We will apply the LASSO regression, which was found to have the best performance in the previous analysis.

Under LASSO regression, the fitted model is :

fertility = 1.476 + 0.003 HHDI.CPI - 0.014 fem.enroll

Different from the linear model we get based on China’s Fertility dataset, the LASSO regression generates a model which only contains two variables. It means that under the LASSO regression model, the fertility rate is only affected by female school enrollment rate and financial status. We will discuss the details of generating these differences in the summary section.

Summary

In this project, we tried to analyze the influencing factors of the fertility rate. Based on the acquired data, linear models of China have been compared. It can be concluded that the most important factors are related to education level, living expenses, urbanization, and industrialization. This is consistent with common sense: a higher education rate (especially female education) and a higher employment rate in a non-agricultural industry could lead to a significant decrease in fertility. It could be explained by the awakening of female independence, while could also be credited to the adaptation to a more industrialized society. Thus, we can hardly decide which one should be the most prominent factor. Referring to the history of the past 4 decades, the boost of economy and education, the rapid expansion of cities and middle-class, and the subsequent drop of fertility happened simultaneously. However, there were several defects in this model that need to be discussed.

First and foremost, our dataset was not comprehensive or large enough. In the ideal case, we planned to use different indicators to illustrate one main factor, for example, the living expenses should include different variables like housing, education, healthcare, and daily necessities. Based on a series of variables for one factor, we might be able to illustrate the importance of different variables in detail. Also, if given more data points (for another decade), our model might be more convincing. Due to the limitation of data resources, we only obtained a set of gross data, which was only enough for a rough model. Resultantly, several hypotheses of fertility were verified, but we were limited on moving further for deeper insights.

Another point to mention is that we did not take the influence of policy into account. From 1980 to 2015, the famous One-child Policy had been strictly implemented. However, it shifted to Two-child and Three-child policies in recent years. In fact, a minor burst in fertility was observed in the next several months after the policy changed. However, it may not greatly affect our model and the linearity. The intention of fertility has already dropped to a very low level, so the recent policy did not bring too many infants.

One more major concern is the validity of the final models. In the LASSO and ridge regression, we had some troubles in the determination of optimal lambda value with the CV method. After adjustments of the lambda sequence, we obtained extremely low lambda (<0.01) in both models. Though the models were not the same as the primitive ones (lmfull), we still doubted that they might be very similar and may still have the problem of collinearity.

For Japan’s fertility data set, we were not able to obtain a full dataset with all variables from the 1970s, when the fertility rate of Japan was not as low as it is nowadays. In this dataset, the fertility rate is more like a horizontal line with several minor fluctuations. It was even worse than the accessible fertility data having only one decimal place. This led to an extremely low correlation between all the independent variables and the response variable. However, we still obtained similar results as in China's model. One possible reason for the less significance of urbanization in Japan’s model is there may be a threshold of this variable. When lower than the threshold, the growth of the urbanization rate can lead to a remarkable decrease in fertility, while it has little effect on fertility after reaching the threshold. Given a larger dataset, we may be able to verify this hypothesis.

Contributions

Shuang and Yihang collected the datasets, established the models, and completed the report together. Specifically, Yihang was more focused on the R programming, and Shuang was more focused on the analysis of the linear models.