What contributes to marriage and how to predict it?

Deng Kaisheng

Abstract

In this paper, I analyze the relations between marriage/divorce rate and many other variables grouped by different provinces in China. These variables include population, GDP(Gross Domestic Product), average home price, CPI, annual education funding, consumption per capita, number of health technicians per ten thousand people, savings, unemployment rate and so on. The data from various aspects of society is taken into acount, in this way, I tried to build models to quantitatively reflect what factors affect the marriage/divorce rate.

In the analysis process, I built two models, one is linear regression model, the other is neural network. I use linear regression model to analyze the relationship between marriage rate and other variables, and use neural network to predict the number of marriage couples, once the marriage number is predicted, the marriage rate can be calculated. The techniques used in this project include model selction based on *olsrr* package and *leaks* package in *R*, data visualization using *ggplot* and *Tableau* and the *nerual network toolkit* in MATLAB.

Key words: marriage divorce linear model neural network model selection ggplot Tableau R MATLAB

Contents

1 Topic Analysis

Traditionally, marriage life was based on the principles of the Confucian ideology in China. This ideology formed a culture of marriage that strove for the "Chinese family idea, which was to have many generations under one roof". Confucianism grants order and hierarchy as well as the collective needs over those of the individual. It was the maintenance of filial piety that dictated a traditional behavior code betwen men and women in marriage and in the lifetime preparation for marriage. The segregation of females and the education of males were cultural practices which separated the two sexes, as men and women would occupy different spheres after marriage.

However, things started to change in the last several decades. More Chinese are getting divorced and fewer are getting married, and the government is growing concerned.

According to figures released recently by the Ministry of Civil Affairs, the "curde divorce rate", which measures the number of separations for every 1,000 people in the population — raised by nearly 60% from 2010(2.00‰) to 2017(3.15‰).

By contrast, the data from the Ministry of Civil Affairs shows that the "crude marriage rate", which measures the number of marriages of every 1,000 people in the population — decreased dramatically in 2017(7.7‰), compared with 2016(7.3‰), and this trend looks set to continue.

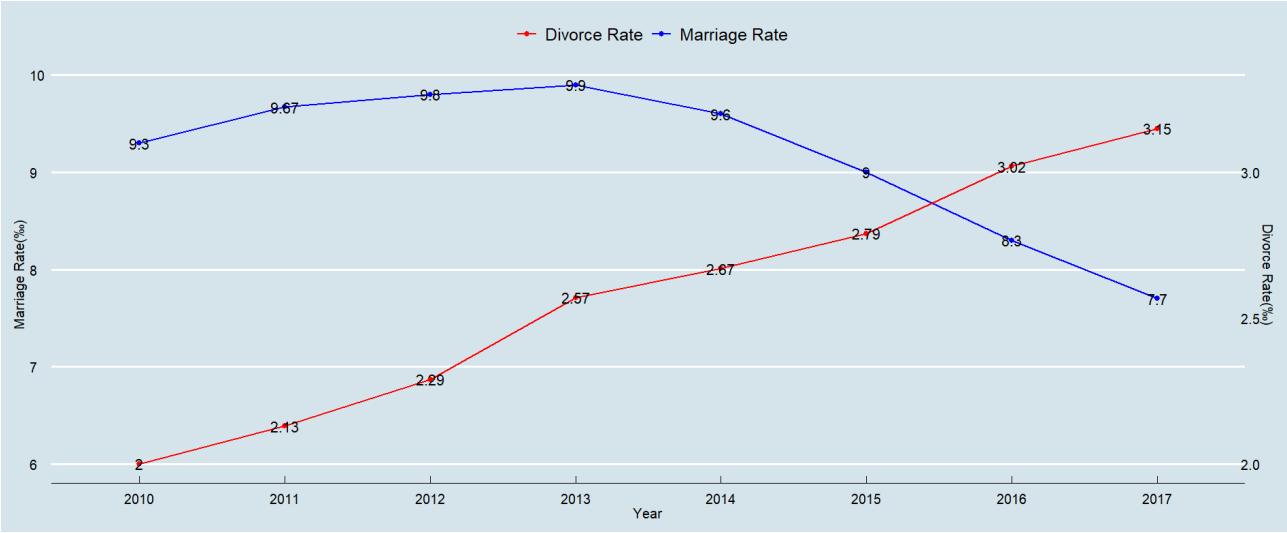

Figure 1: Marriage/Divorce rate in China, 2010 - 2017

Also, marriage rate differs from province to province. In general, eastern China has a lower marriage rate than western China, coastal areas has a lower marriage rate than inland areas. Developed areas like Beijing, Guangdong, Shanghai all have low marriage rate.

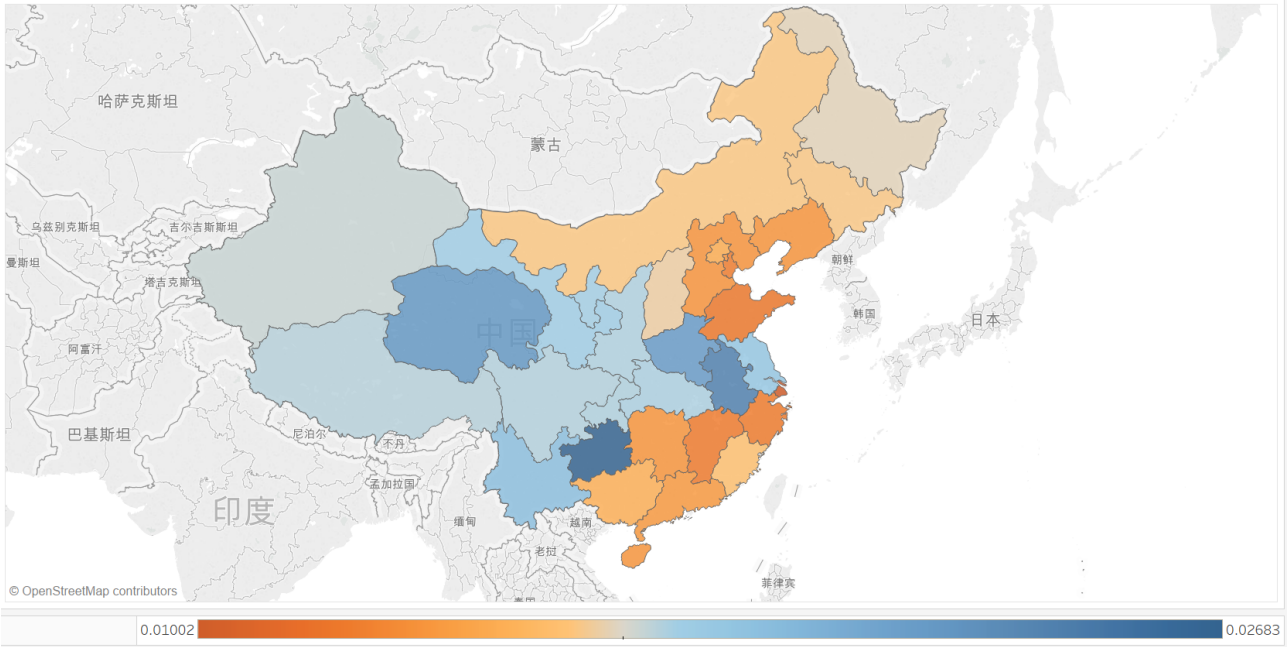

Figure 2: Marriage rate distribution in China, 2018

2 Data Preparation

I collect data from different area of society in the last 10 years, including population, GDP, income, consumption, education, health care, homeprice, unemployment rate, and so on. All the data comes from the National Bereau of Statistics. The detailed description is the following:

Variable Meaning	Variable Name	Unit
Marriage rate	MR	%
Number of marriage couples	MN	10,000
Average sales price of residential commercial housing	HP	yuan / m^2
Consumer Price Index	CPI	
Consumption per capita	CON	yuan
Education funding in total	EDUF	10,000 yuan
Gross domestic product	GDP	100 million yuan
Number of males per 100 females (gender gap)	GG	
Number of health technicians per 10,000 people	HTPT	
Higher education enrollment per 10,000 people	HEET	
Household disposable income per capita	INC	yuan
Population	POP	10,000 people
Savings per GDP	SAVPG	
Unemployment rate	UR	%

Table 1: Explanation of the meaning of variables

3 Data Cleaning

The task of data cleaning is to organize it to the format that enables futher processing. Since there is much repeating work, I defined several functions to finish it.

```
# Function: transpose data frame together with years.
Transpose = function(vectorData, name) {
        vectorData = vectorData[, -1]
        colNames = colnames(vectorData)
        vectorData = cbind(colNames, data.frame(t(vectorData)))
        rownames(vectorData) = seq(1, nrow(vectorData))
        colnames(vectorData) = c("year", name)
        return(vectorData)
}
# Function: merge a list of dataframes
mergeFunc = function(dataFrameList, mergeBy) {
    res = dataFrameList[[1]] %>% select(mergeBy)
    for (df in dataFrameList) {
        res = merge(res, df, by = mergeBy, all.x = TRUE)
    }
    return(res)
}
# Function: deal with the NAs using linear interpolation.
# Use "na.interp" directly won't work.
linearInterp = function(x, y, fitData) {
    model = lm(y ~ x, data = fitData)
    interpData = x[is.na(y)]
    y[is.na(y)] = predict(model, newdata = data.frame(x = interpData))
```

```
    return(y)
}
```

After data cleaning, I split the whole dataset into training dataset and testing dataset randomly by ratio 4:1

```
# Prepare training data and testing data for futher modeling.
training_rows = sample(1:nrow(cleanedDataSet), 0.8 * nrow(cleanedDataSet))
training_data = cleanedDataSet[training_rows,]
testing_data = cleanedDataSet[-training_rows,]
write.csv(training_data, "training_set.csv", row.names = F)
write.csv(testing_data, "testing_set.csv", row.names = F)
```

Now let's take a glimpse of the training dataset (first 10 rows and first 9 columns).

year	province	MR	DR	POP	GDP	INC	CON	CPI
2,015	广东省	0.015	0.004	10,849	72,812.550	27,858.860	20,975.700	101.500
2,016	安徽省	0.023	0.007	6,196	24,407.620	19,998.100	14,711.530	101.800
2,016	江西省	0.013	0.004	4,592	18,499	20,109.560	13,258.620	102
2,010	安徽省	0.022	0.004	5,957	12,359.330	9,776.172	6,470.813	103.100
2,014	浙江省	0.016	0.005	5,508	40,173.030	32,657.570	22,551.970	102.100
2,014	江苏省	0.021	0.005	7,960	65,088.320	27,172.770	19,163.560	102.200
2,011	吉林省	0.017	0.007	2,749	10,568.830	13,368.510	9,965.174	105.200
2,018	四川省	0.018	0.008	8,341	40,678.130	22,460.550	17,663.550	101.700
2,015	吉林省	0.017	0.009	2,753	14,063.130	18,683.650	13,763.910	101.700
2,009	辽宁省	0.018	0.006	4,341	15,212.490	13,983.150	9,392.949	100

Table 2: A glimpse of the cleaned data

4 Models

4.1 Analyze — linear model 1

First of all, I plot the correlations between all the independent variables to see if they are highly correlated.

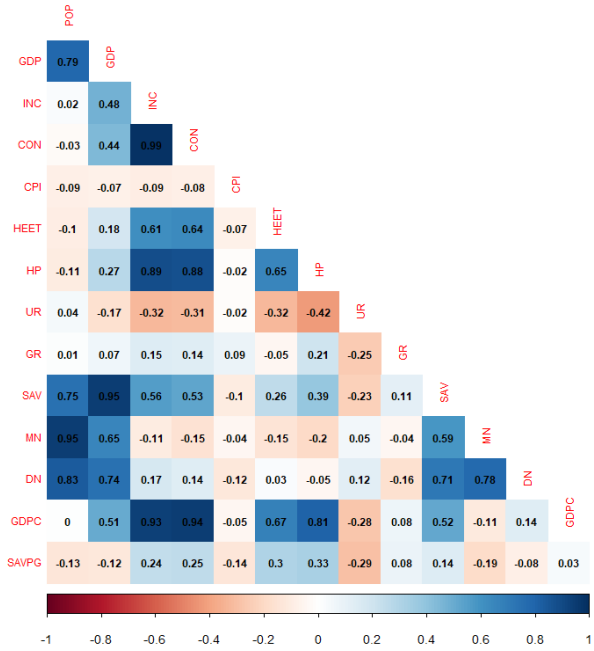

Figure 3: The correlation between independent variables

We can see the correlations between some of the variables are large and we should pay attention to that and try to avoid a very large variance inflation factor in our model.

Then, I trained linear model with all the independant variables, and used variable selection techniques according to the following principles:

- GDP per capita and population can reflect income already.
- GDP per capita is better predictor than GDP.
- The income, consumption and savings has the relationship like $incom = consumption + savings$, so I just keep SAV, and SAV per GDP is better preditor than SAV.
- Homprice and CPI seem to have no influence on marriage rate.

After that, I found the p-value of GG insignificant, so I removed GG in my model and get the final linear model.

	Dependent variable:		
	MR		
	linear_model.whole	linear_model.modify	linear_model.final
POP	0.00000*** (0.00000)	0.00000*** (0.00000)	0.00000*** (0.00000)
GDP	0.00000*** (0.00000)		
INC	−0.00000 (0.00000)	−0.00000 (0.00000)	
CON	0.00000** (0.00000)		
CPI	0.0001 (0.0001)		
HEET	0.00000 (0.00000)		
HP	0.000 (0.00000)		
UR	−0.001*** (0.0003)	−0.001** (0.0003)	−0.001** (0.0003)
GG	0.0001 (0.0001)	0.00003 (0.0001)	
SAV	−0.00000*** (0.00000)		
GDPC	−0.001 (0.0004)	−0.0004 (0.0003)	−0.001*** (0.0001)
SAVPG	0.004* (0.002)	−0.002 (0.001)	−0.003*** (0.001)
Constant	0.011 (0.013)	0.024*** (0.002)	0.025*** (0.002)
Observations	310	310	310
R^2	0.298	0.213	0.209
Adjusted R^2	0.270	0.198	0.199
Residual Std. Error	0.003 (df = 297)	0.003 (df = 303)	0.003 (df = 305)
F Statistic	10.530*** (df = 12; 297)	13.701*** (df = 6; 303)	20.204*** (df = 4; 305)
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 3: Summaries of 3 linear models

The coeffients of the final linear model are the shown in the following table:

Variable	(Intercept)	POP	UR	GDPC	SAVPG
Coefficient	0.02492	2.072×10^{-7}	-7.973×10^{-4}	-6.554×10^{-4}	-3.16×10^{-3}

Table 4: Coeffients of final linear model

Therefore, we can reach a equation indicating that:

$$MR = 0.02492 + 2.072 \times 10^{-7} \times POP - 7.973 \times 10^{-4} \times UR - 6.554 \times 10^{-4} \times GDPC - 3.16 \times 10^{-3} \times SAVPG \quad (1)$$

4.2 Prediction

4.2.1 Linear Model 2

This time, I use the similar techniques to select variables, the summary of the three linear models are as follows:

	Dependent variable:		
		MN	
	linear_model.whole	linear_model.modify	linear_model.best
POP	0.012*** (0.0004)	0.012*** (0.0004)	0.012*** (0.0003)
GDP	−0.00002 (0.0002)	−0.00003 (0.0001)	
INC	0.00005 (0.0002)		
CPI	0.542* (0.280)	0.538* (0.279)	0.516* (0.275)
HEET	0.0002 (0.001)		
UR	−2.979*** (0.745)	−3.013*** (0.723)	−3.090*** (0.692)
GG	−0.206* (0.118)	−0.207* (0.116)	−0.208* (0.114)
SAV	−0.001*** (0.0002)	−0.001*** (0.0002)	−0.001*** (0.0001)
GDPC	0.659 (0.832)	0.898*** (0.304)	0.809*** (0.274)
SAVPG	0.602 (5.276)	1.236 (4.825)	
Constant	−49.664* (29.644)	−49.359* (29.508)	−45.429 (28.319)
Observations	248	248	248
R^2	0.942	0.942	0.942
Adjusted R^2	0.939	0.940	0.940
Residual Std. Error	6.643 (df = 237)	6.617 (df = 239)	6.596 (df = 241)
F Statistic	382.588*** (df = 10; 237)	482.061*** (df = 8; 239)	646.765*** (df = 6; 241)
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 5: Model selection of linear model prediction

Then test the performance on testing set.

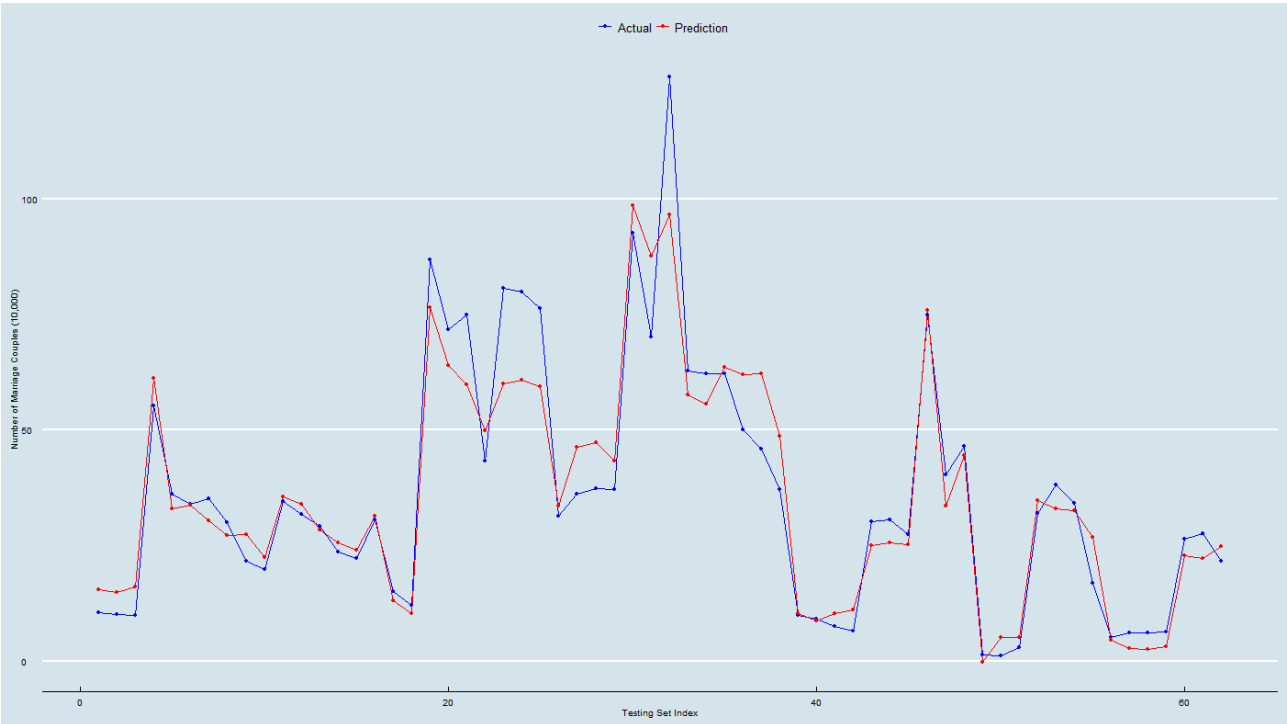

Figure 4: Linear model performance on testing set

It is obvious that the model can predict the marriage number in a large degree. The mean square error is 67.8

4.2.2 Neural Network

According to the experience in linear model variable selection, I choose **POP, UR, GG, SAV, GDPC** as predictors.

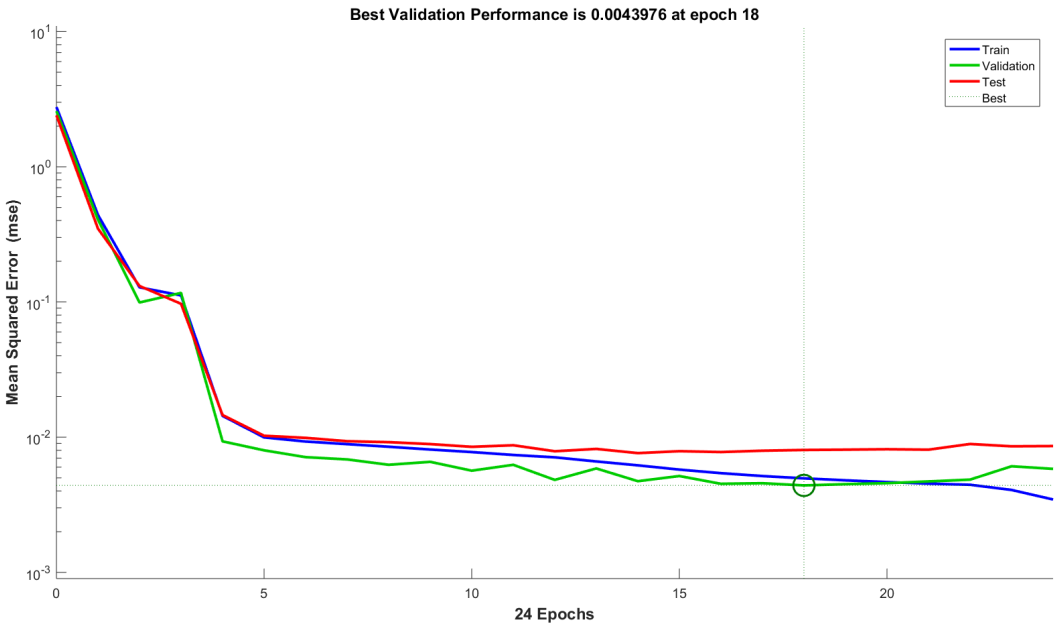

Figure 5: Neural network training process

According to the curve of training prcess, the performance of neural network is very close to the best.

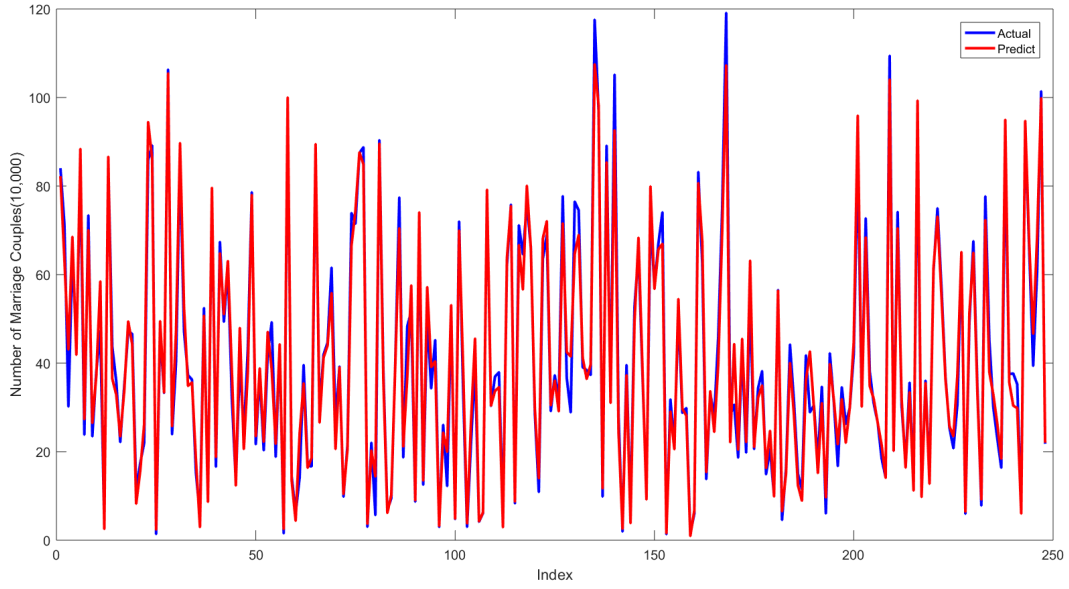

Figure 6: Neural network performance on training set

The network performs perfectly on training set.

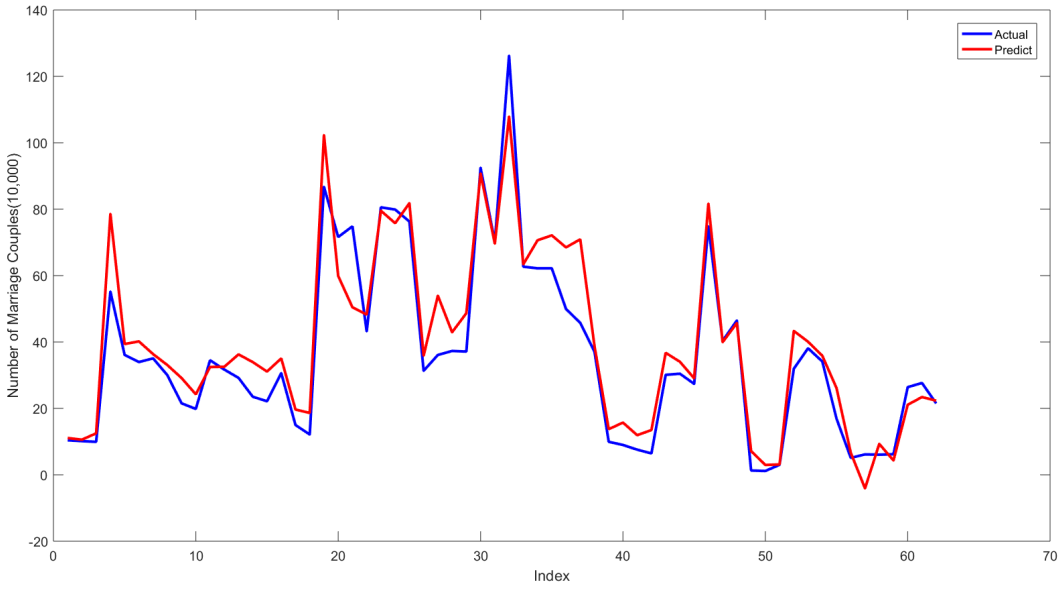

Figure 7: Neural network performance on testing set

The performance on testing set is satisfactory.

4.3 Prediction Results

Let's take a glimpse of the prediction results.

year	province	actual_MN	predicted_MN
2016	内蒙古自治区	19.840	22.447
2014	辽宁省	34.510	35.569
2015	辽宁省	31.700	33.995
2017	辽宁省	29.190	28.375
2014	吉林省	23.520	25.668
2016	吉林省	22.150	23.897
2016	黑龙江省	30.630	31.434
2009	上海市	14.990	12.978
2018	上海市	12.148	10.183

Table 6: Some of the prediction results

5 Model Intepretation

There are some interesting conclusions that we can draw from the linear model.

- Population
  The model indicates that the population has positive influence on the marriage rate, which seems reasonable since more population means people may have more chances to meet each other, there is more interaction between individuals. Henan province(河南省) is a good example.

- Unemplyment rate
  The model also shows that unemployment rate has a negtive coefficient, which means higher unemployment rate will result in lower marriage rate. That is reasonable because higher unemployment rate means more unstable factors of society.

- GDP per capita
  GDP per capita is also a negtive factor, which means the more developed provinces tend to have lower marriage rate. This may be due to the ecnomic pressure and fast pace in more developed areas. This pattern can also be seen in figure2, where coastal areas tend to have lower marriage rate.

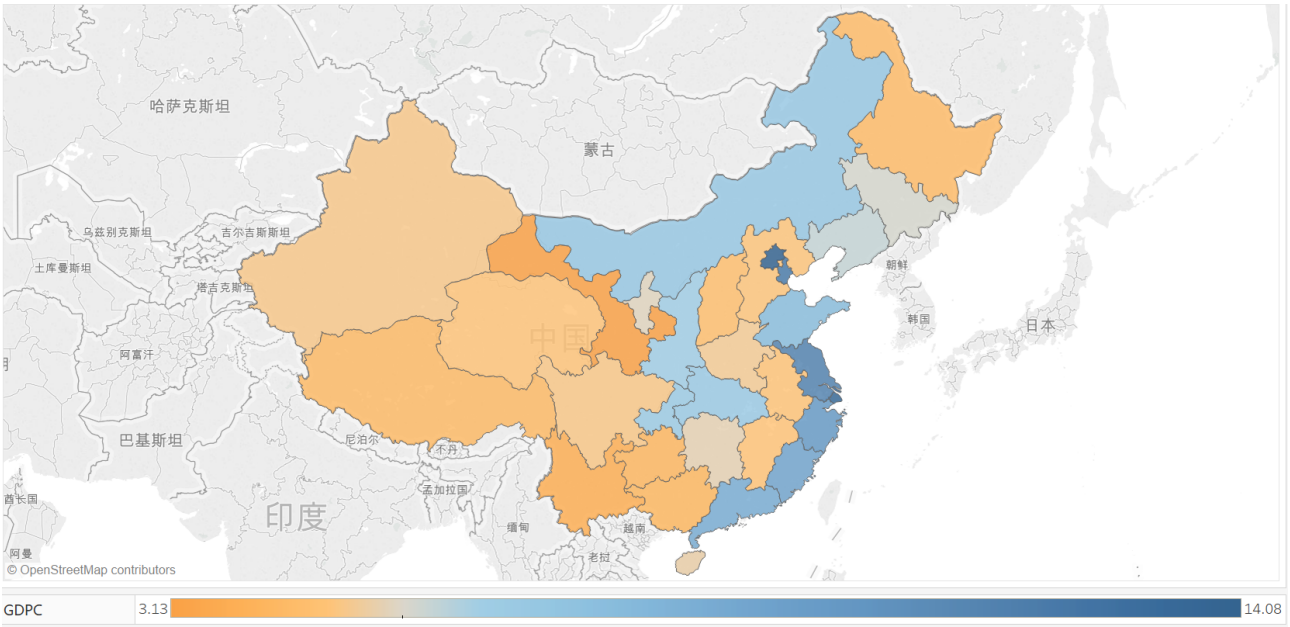

Figure 8: The distribution of GDP per capita in China, 2018

- Savings per GDP
  It may seem weird at first glance that savings per GDP has a negtive coefficient. However, it makes sense once thinking further. Higher level of savings per GDP indicates that in one area, people tend to save their income in the bank rather than concume it, that is, a more conservative consumption habbit. This habbit can undermine marriage rate since getting married is very costly, especially in big cities, for males.

- Gender gap
  It's surprising that gender gap doesn't have a large influence on marriage rate. Upon further analysis, the reasons may due to the gap between urban and suburban areas. In China, the gender gap is focused in suburban areas, where the development level is lower than that of urban areas. Therefore, many females in urban area are unwilling to marry a male in the suburban area due to worse living environment, lower income, poorer health care and so on.
  Another reason may be the population immgration, which is very common in China, especailly in big cities. People from different areas may meet and marry, thus undermine the influence of gender gap in a certain area.

6 References

[1] National Bureau of Statistics
[3] Ministry of Civil Affairs of the People's Republic of China

[2] South China Moring Post: Marriage rate down, divorce rate up as more Chinese couples say ‘I don’t’ or ‘I won’t any more’

[3] ggplot2 version 2.2.0 - Demonstration of dual y-axes using sec.axis

[4] ggthemes — all your figure are belong to us