Advances in Social Science, Education and Humanities Research (ASSEHR), volume 183
2018 4th Annual International Conference on Modern Education and Social Science (MESS 2018)

The Application of logistic model on Study the Influence Factors of Migrants' Happiness Based on Group Lasso

Hai-Xia ZHAO[1,a*] and Hong-Bo SHI[2,b]

[1] School of Statistics, Shan xi University of Finance and Economics, Taiyuan 030006,China

[2] School of Information Management, Shan xi University of Finance and Economics, Taiyuan 030006,China

[a]zhaohaixia_ok@126.com

*Hai-Xia ZHAO

Keywords: Group Lasso-logistic model, variable selection, migrants.

Abstract. Migrants has made a great contribution in the process of urban development, carrying out in-depth research about the influence factors of migrants happiness is of great significance to promote the harmonious development of society. In this paper, based on CHIP2013 data about Migrants, Group Lasso-logistic model was used for variable selection and analyzing, and compared with the methods of full logistic model, forward and backward selection logistic model. The result shows that the Group Lasso-logistic model not only provides the effective sparse results, but the generalization ability is optimal.

Introduction

The contribution of the migrants in the process of economic and social development is obvious to all. However, due to some problems such as household registration system, labor insurance benefits, the migrants has encountered many obstacles that are difficult to overcome in the process of integrating into the urban life, and they are always on the edge of the city. Therefore, it is one of the main contents of building a harmonious society to study the subjective feelings of the migrants on life and try to improve subjective well-being. With the popularization of micro data in recent years, research on this issue has also achieved certain results, mainly in the following aspects: First of all, some scholars have studied the influence of the migrants of happiness from a specific research point. For instance, Xu Shicun [1] examined the impact of urban adaptability on the subjective well-being of the migrants in Heilongjiang Province; Cai Jinghui [2] et al. have studied the happiness of migrants from the perspective of urban scale; Xia lun [3] studied the effect of income on happiness sense of migrants in Beijing; Li Fangzhi and Xiang Shujian [4] studied the influence of the income gap between migrants on the sense of well-being, and the results showed that there was a significant U-shaped relationship between the two; Zhu Hailin et al. [5], Ji Yongbao et al. [6], studied their influence on the happiness of migrants from the perspective of urban sense of belonging and social integration respectively; Second, the study of differences in happiness among different migrants mainly includes Xia Lun [7] and Huang Jiawen [8] , who respectively studied the urban-rural differences and generation differences in the happiness of migrants; Third, taking the migrants in a certain region as an example to study the influencing factors of their happiness, Xia Lun[9], Zhang Huachu[10], Yang Dongliang, Chen Sisi[11], and Xie Yuhui[12] took the migrants of Hubei Province ,Guangzhou city, Beijing Municipality, Shanghai city as examples to examine the influencing factors of their well-being.

In summary, the above research has achieved fruitful results, but also have the following deficiencies: In the research content, the vast majority of studies are started from a certain angle to study its impact on the well-being of the migrants, the systemic research of the influence factors of happiness is mainly based on the migrants of a certain area; On the research methods, the correlation analysis, logistic regression, and multiple logistic regression are generally used. For the case of discrete dependent variables, logistic model has been widely used due to its advantages such as simple calculation and easy interpretation. However, there are many influencing factors of the

happiness of migrants, and more dummy variables need to be introduced into the model. If logistic model is used directly, it will increase the complexity of the model and reduce the accuracy of the model prediction due to the collinearity between the variables and so on.

In view of the above shortcomings, this paper intends to use Logistic model of Group Lasso to study the problem, because there are many qualitative variables involved in the study of the influence factors about the happiness of migrants, and there will be some related virtual variables after quantifying these qualitative variables. As a Group Lasso method for selecting variables at the group level, not only can we avoid the increase of model complexity because of too many variables selected in the Logistic model and the problem of collinearity between variables, but also can make variable selections for related dummy variables as a whole, which is more helpful for interpreting the result of the model.

Group Lasso-Logistic Model

Since Tibshirani (1996) [13] proposed the Lasso variable selection method, Lasso method has been widely applied to many fields since it has many advantages such as variable selection and parameter estimation simultaneously [14,15]. Subsequently, in order to overcome the disadvantages of Lasso's estimation of bias, it has also been proposed to adapt Lasso, relax Lasso, etc. [16,17]. The above methods are mainly used to make a separate selection of variables, and it does not have group effects, but there are certain correlations between many variables in practice. When the group structure is formed, the results of Lasso method are often difficult to explain[18].

Yuan and Lin (2006) [19] proposed the Group Lasso method for the correlation of variables in linear regression model, and selected variable groups. The form of the Group Lasso method is as follows:

$$\hat{\beta} = \arg\min \frac{1}{2}\left\| Y - \sum_{j=1}^{J} X_j \beta_j \right\|_2^2 + \lambda \sum_{j=1}^{J} \left\| \beta_j \right\|_2 \quad . \tag{1}$$

Where J is the total number of groups that are included in the independent variable, and the he penalty term in the upper formula can be regarded as being between the L1 penalty used in Lasso and the L2 penalty used in the ridge regression. It can not only select variables at the group level, but also have the same property as ridge regression under orthogonal transformation.

Later, Lukas Meier et al. (2008) [20] applied the Group Lasso method to Logistic regression, set the sample is $(x_i, y_i), i = 1,2,\mathrm{L}\ n$, y_i is a binary dependent variable, that is, the independent variable can be a continuous variable or a classification variable, and it can be divided into J groups, and the degree of freedom of the independent variable of group J is set as: $df_j, j = 1,2,\mathrm{L}\ J$.

Set the probability of the occurrence of a dependent variable is:

$$p_\beta(x_i) = \mathrm{P}_\beta[Y = 1 \mid x_i]. \tag{2}$$

Then,

$$\log\left(\frac{p_\beta(x_i)}{1 - p_\beta(x_i)}\right) = \beta_0 + \sum_{j=1}^{J} x_{i,j}^T \beta_j = \eta_\beta(x_i). \tag{3}$$

Where β_0 is the intercept item, β_j is the coefficient vector corresponding to the independent variables of group J, and $\beta = (\beta_0, \beta_1^T, \mathrm{L}\ \beta_J^T)^T$. Then, minimize the following convex function, you

can get parameter estimation $\hat{\beta}_\lambda$,

$$S_{\lambda}(\beta) = -l(\beta) + \lambda \sum_{j=1}^{J} s(df_j) \left\| \beta_j \right\|_2 \quad . \tag{4}$$

$l(\beta)$ is the log likelihood function:

$$l(\beta) = \sum_{i=1}^{n} \{ y_i \eta_{\beta}(x_i) - \log(1 + \exp(\eta_{\beta}(x_i))) \}. \tag{5}$$

The function s(.) still follows the form used in Lukas Meier (2008),

$$s(df_j) = (df_j)^{\frac{1}{2}} \quad . \tag{6}$$

The difference in the harmonic parameter λ will determine the variable selection of the model, and the selection of λ can be carried out according to the AIC, BIC, GCV and other criteria.

Data Sources and Processing

The research data in this paper comes from survey data on migrant workers in CHIP 2013. The explanatory variables selected age, sex and marital status, years of schooling, registered residence, health, labor insurance benefits, lending, home and family terrain, status of housing and contracted land, proportion of migrant workers, trust in friends, family income and living conditions in 14 groups of 36 variables to analyze (see table1). The measure of happiness in the questionnaire is reflected in the question, “Do you feel happy considering all aspects of life?”, the answers to the questions are divided into five categories: “very happy”, “relatively happy”, “general”, “not very happy”, and “very unhappy”. In order to make it easier to compare, we have merged these five categories into two categories: “general above” and “general below”, in the original questionnaire, “very happy”, “relatively happy” and “general” were merged into “general and above” categories, and the “general below” category included “not very happy” and “very unhappy” in the original questionnaire.

The data includes a total of 726 migrant workers, and the remaining sample size is 558 for removing the data that does not meet the filing requirements and the missing data. And this paper first randomly selects 10% as the test set from the “General” and “General below” data, and the remaining 90% is used as the training set. The number of “general above” and “general below” classes in the training set is approximately 9:1. In view of this obvious imbalance, the random oversampling method is used to balance the data to reduce the impact of data imbalance on the model.

Table 1 Variable Description

Group	Variable
Sex	X1=1，male；X1=0，female
Age	X2
Years of schooling	X3
Marital status	X4-1=1：First marriage，X4-1=0：other ；X4-2=1：remarriage，X4-2=0：other；X4-3=1：divorce， X4-3=0：other；X4-4=1：Widowed，X4-4=0：other；X4-1~ X4-4 all 0：unmarried
Nature of residence	X5-1=1: agricultural registered permanent residence, X5-1=0:other；X5-2=1：non-agricultural registered permanent residence，X5-2=0:other； X5-1~X5-2 all 0：resident residence
State of health	X6-1=1：great，X6-1=0：other；X6-2=1：good，X6-2=0：other；X6-3=1：general，X6-3=0：other；X6-4=1：not good，X6-4=0：other；X6-1~ X6-4 all 0：very bad
Labor Insurance Benefits	X7-1=1：injury suffered on the job，X7-1=0：other；X7-2=1：unemployment，X7-2=0：other；X7-3=1：reproduction，X7-3=0：other；X7-4=1：more than two items，X7-4=0：other；X7-1~ X7-4 all 0：nothing
Lending situation	X8-1=1：There has been a loan application in the last three years，X8-1=0：No loan application in the past three years, X8-2=1：borrow money in the past three years，X8-2=0：No need to borrow money in the past three years
Home and family terrain	X9-1=1：Plain，X9-1=0：other；X9-2=1：hills，X9-2=0：other；X9-3=1：mountain area，X9-3=0：other；
Status of housing and contracted land	X10-1=1：Old home has its own housing，X10-1=0：Old home does not have its own housing, X10-2=1：The old home has contracted land，X10-2=0：The old home does not have contracted land
Proportion of migrant workers	X11-1=1：Less than 25%，X11-1=0：other；X11-2=1：Greater than 25% less than 50%，X11-2=0：other；X11-3=1：More than 50% less than 75%，X11-3=0：other；X11-1~ X11-3 all 0：Greater than 75%
Trust in friends	X12-1=1：Very unreliable，X12-1=0，other；X12-2=1：Not too credible，X12-2=0，other； X12-3=1：general，X12-3=0，other；X12-4=1：More reliable，X12-4=0，other； X12-1~ X12-4 all 0：Very credible
Family living conditions	X13-1=1：Life is very comfortable and has the ability to pay for all kinds of extra spending，X13-1=0：other； X13-2=1：Life is basically comfortable and incapable of paying all kinds of extra spending，X13-2=0：other； X13-1~ X13-2 all 0：Life is not comfortable and incapable of paying the basic consumption expenditure
Family income status	X14-1=1：Have the ability to deal with the occurrence of all kinds of accidents， X14-1=0：Other； X14-2=1：Have the ability to respond to a number of accidents，X14-2=0：other； X14-3=1：There is no ability to cope with accidents and be able to cope with the occurrence of basic life events，X14-3=0: other；X14-1~ X14-3 all 0：There is no ability to deal with the occurrence of basic events

The Establishment and Estimation of The Model

In this paper, the parameter estimation of Group Lasso-Logistic model is completed by using the grpreg package in R software, according to the BIC criterion, the harmonic parameter is selected, and the estimation result of the model is shown in Table 2.

Table 2 Parameter Estimation

Variable	Group Lasso-logistic model	Backward selection model	Full variable model
X1	0	-0.561	-0.485
X2	0	0	-0.0143
X3	0.070	0.203	0.182
X4-1	-0.098	0	-0.006
X4-2	-0.032	0	0.633
X4-3	-0.263	0	-0.023
X4-4	0.616	21.162	21.190
X5-1	-0.747	-19.137	-18.994
X5-2	0.178	0	0.164
X6-1	0	0	14.744
X6-2	0	0.738	15.676
X6-3	0	0	15.563
X6-4	0	0	31.847
X7-1	0	0	-0.474
X7-2	0	0	14.242
X7-3	0	0	0.351
X7-4	0	0	0.105
X8-1	0.186	0.778	0.5728
X8-2	-0.558	-1.258	-1.274
X9-1	-0.963	-18.699	-18.438
X9-2	-0.473	-17.335	-17.178
X9-3	-1.265	-19.326	-19.089
X10-1	0	0	0.321
X10-2	0	-0.563	-0.501
X11-1	0.107	0	0.202
X11-2	0.516	1.489	1.567
X11-3	0.466	1.298	1.267
X12-1	0.326	0	15.963
X12-2	-1.542	-5.019	-4.727
X12-3	-0.704	-1.597	-1.598
X12-4	-0.922	-2.627	-2.430
X13-1	2.759	18.125	18.457
X13-2	1.681	2.751	2.805
X14-1	1.930	21.162	21.712
X14-2	2.067	19.800	19.894
X14-3	0.704	1.217	1.332

From the estimation results of the Group Lasso-logistic model, the model selected a total of 24 variables in 9 groups, indicating that the years of education, marital status, the nature of the accounts, the status of loans, the topography of the home, the proportion of migrant workers in the village, the proportion of credibility of friends, family life, and income levels all have an important impact on the well-being of the migrants. Among them, the higher the family income level and the living standards, the higher the sense of their happiness. And, the number of years of education contributes to the happiness of the migrants. When the number of years of education increases by 1 year, the incidence of the two categories will increase by 7.3%; Registered permanent also have a significant impact on the well-being of the migrants. The performance of a migrant with agricultural registered permanent residence is significantly lower than that of non-agricultural households; in the past three years, the migrants who had borrowed money from relatives and friends would have a significant decrease in their sense of happiness. The higher the proportion of migrant workers in their home villages, the happier their sense of happiness will be; from the point of view of marital status and trust in relatives and friends, compared with unmarried people, married people have an

inhibitory effect on happiness. Among them, the difference between the first marriage and remarriage is not significant, while the difference in the happiness of the divorced person is reduced by 23.1%. The higher the degree of trust in relatives and friends, the greater the sense of well-being, and the two variables of widowhood (x4-4) and distrust of friends (x12-1) all contribute to the happiness of the migrants, this is because there are very few samples of widowed and very untrustworthy friends in the sample, resulting in such hard-to-interpret results.

Comparison of Models

After parameter estimation of the Group Lasso-logistic model, the paper also estimates the parameters of the Logistic model of the full-variable, forward and backward regression logistic model, and the results are listed in Table 2. In the Logistic model of forward stepwise regression, all variables are retained and not removed. Therefore, the Logistic model of forward stepwise regression is exactly the same as the full-variable Logistic model. The backward step-by-step regression model removes 15 explanatory variables, slightly more than the Group Lasso-logistic model, and the model is more concise; however, the deficiency of this model is that some variables are removed and some variables are retained, for example, the variable groups where x4, x11, etc. are located will cause the lack of interpretation of the model results.

Considering the model prediction accuracy, table 3 lists the correctness rates of the three models on the training set and the test set, it can be seen that the full-variables model in the prediction of the training set is 87.3%, the highest rate of correct, but in the test set, the accuracy of prediction fell to the lowest 74.5%; Similar to the all-variable model, the prediction accuracy of the backward selection model is higher in the training set, but it is reduced by nearly 10 percentage points in the test set. This shows that the generalization ability of the full-variable model and backward selection model is poor; The accuracy of the Group Lasso-logistic model in the training set is slightly lower than that of the other two models, but the accuracy of the prediction in the test set is the highest among the three models. Therefore, from the generalization ability of the model, the Group Lasso-logistic model is indeed the best among the three models.

Table 3 Comparison of model prediction accuracy

Model	Training set	Test set
Group Lasso-logistic model	80.7%	81.8%
Backward selection model	86.6%	76.4%
Full-variable model	87.3%	74.5%

Conclusion

This paper uses the survey data of migrant workers in CHIP2013 to establish the Group Lasso-logistic model, analyzes the influencing factors of migrants happiness, and compared with the logistic model of full-variable logistic regression and backward regression, it concludes that: The number of years of education, marital status, the nature of the household registration, the situation of the loan, the topography of the home, the proportion of migrant workers in the village, the credibility of friends, family life, and income have a significant impact on the happiness of the migrants; Compared with the full-variable logistic model, the Group Lasso-logistic model and the logistic regression model with backward regression can all effectively reduce the variables and reduce the complexity of the model. However, in the interpretation of the results, the Group Lasso-logistic model has certain advantages; Although the accuracy of the Group Lasso-logistic model in the training set is slightly lower than that of the other two models, the accuracy of the

prediction in the test set is the highest. Therefore, the Group Lasso-logistic model is optimal in the extrapolation of the model.

Acknowledgement

This research was financially supported by the National Social Science Fund project (16BGL207), Shanxi University of Finance and Economics Youth Fund project (Z06038) and Research project on Philosophy and Social Sciences in Colleges and Universities (2015325).

References

[1] Xu Shicun, Analysis of the Influence of Urban Adaptation on the Subjective Happiness of Migrants: A Case Study of Heilongjiang Province, J. Journal of Population Sciences. 04(2015)36-47.

[2]Cai Jinghui, RenBin, Huang Xiaoning, The Impact of Urban Scale on the Well-being of Migrants: Empirical Evidence from RUMIC (2009), J. Journal of Guizhou University of Finance and Economics.01(2016)89-99.

[3] Xia Lun, Research on the Relationship between Migrants Income and Subjective happiness——Based on Survey Data of Migrants in Beijing, J. Journal of Xihua University (Philosophy and Social Sciences).03(2014) 101-107.

[4] Li Fangzhi, Xiang Shujian, Research on the Impact of Income Gap between the Migrants on Subjective Happiness, J. Statistics and Information Forum. 07(2016)107-112.

[5] Zhu Hailin, Chen Jiancheng, Bai Wei, Zhang Yujing, Study on the Relationship between the City Migrants belonging and Subjective Well-being -- Based on Survey Data of the Migrants in Beijing City, J. Journal of Hebei University of Science and Technology (Social Science). 01(2015) 33-38.

[6] Ji Yongbao, Gao Jingyun, Yang Jun, The influence of social integration of Migrants on the Sense of Happiness: A case study of Shandong Province, J. City issues. 07(2016) 95-103.

[7] Xia Lun, Analysis of Urban-rural Differences in Subjective Happiness of Migrants, J. Statistics and Decision. 09(2015) 110-115.

[8] Huang Jiawen, The Subjective Happiness of Migrants and Its Intergenerational Differences, J. Journal of South China Agricultural University (Social Sciences Edition). 02(2015)122-133.

[9] Xia Lun, Analysis of Influencing Factors of Subjective Happiness of Migrants, J. Statistics and Decision. 09(2014) 93-96.

[10] Zhang Huachu, Influencing Factors of Subjective Happiness of Migrants: A Case Study of Guangzhou City, J. Urban Issues. 10(2014) 90-95.

[11]YANG Dongliang, CHEN Sisi, The Influencing Factors of the Happiness of Migrants in Beijing, J. Journal of Population. 05(2015) 63-72.

[12] Xie Yuhui, Chen Hongsheng, Liu Yuqi, Li Zhigang, Study on the Happiness of Migrants in China's Big Cities: Taking Shanghai as an Example, J. Modern Urban Research. 12(2015) 2-8.

[13] Tibshirani R, Regression shrinkage and selection via the Lasso, J. Journal of the Royal Statistical Society: Series B (Methodological). 58(1996) 267-288.

[14] Fang kuangnan, Zhang Guijun, Zhang Huiying, A method of personal credit risk early warning based on Lasso-logistic model, J. Journal of Quantitative Economics and Technology. 02(2014) 125-136.

[15]Yu Shenghua, Gong Shanghua, Grain price forecast based on Lasso and Support Vector Machine, J. Journal of Hunan University (Social Sciences). 01(2016)71-75.

[16] Zou H, The adaptive Lasso and its oracle properties, J. Journal of the American Statistical

Association.101(2006)1418-1429.

[17] Meinshausen N, Relaxed Lasso, J. Computational Statistics & Data Analysis. 52(2007) 374-393.

[18] Hu Xiaoning, He Xiaoqun, Personal Credit Evaluation Based on Group Lasso, J. Mathematics in Practice and Theory. 03(2015)89-94.

[19]Yuan M,Lin Y, Model selection and estimation in regression with grouped variables, J. Journal of the Royal Statistical Society: Series B(Statistical Methodology). 68(2006)49-67.

[20] Meier L, van de Geer S, Buhlmann P, The group lasso for logistic regression, J. Journal of the Royal Statistical Society: Series B. 70(2008)53-71.