What associated factors behind induced abortion in China: A Study on CHNS from 1991 to 2015

Jiyu Yan
Student ID:1851015
CS910 DATA ANALYTICS

***Abstract*—This study is based on data from the China Health and Nutrition Survey(CHNS), with the purpose of exploring the overall induced abortion ratio change and the factors associated with post-abortion women among 18 to 52 years old women with pregnancy history from 1997 to 2015 in 9 provinces in China. Then the education level, health level and other attributes are considered to compare the difference between post-abortion women and non-abortion women. The paper then will attempt to identify what behind induced abortion: the strength of policy implementation, variance of education level, rural/urban or other factors, and finally form models to perdict whether a woman has induced abortion history.**

I. Introduction

The induced abortion of women in past many years is an essential but sensitive topic in China and it related to a lot important issues such as population size and very unbalanced gender ratio. The situation of induced abortion in China is often concerned by the international community, and there is also much debate around this issue.

This study will focus on the data based on latest CHNS(data including 2015 wave published in May 2018), which is a longitudinal survey conducted by the cooperation between institution in USA and China, to examine the impact of health, nutrition and family planning policies and to study how China's social-economic transformation can affect the health and nutritional status of the entire population[1].

The two aspects of this study are shown below.

1) Focus on how induced abortion ratio are changed during these 14 years and between 9 provinces, also consider whether rural or urban make it different.
2) Using this micro-data to find the effect of associated factors such as education level between post-abortion women and non-abortion women.

II. Background

Since the founding of the People's Republic of China, China's population policy has undergone several major changes. At the beginning of the founding of new China, encouraging population births, in the late 1970s and early 1980s, a strict family planning policy was implemented, and gradually loosed after 1995[2], then two children policy is started after 2011. The total birth rate and natural growth rate drop from 33.43% and 25.83% in 1970 to 11.93% and 4.79% in 2011 respectively, then get a little increase to 12.43% and 5.32% in 2017, but still gives China one of the lowest fertility levels in the world nowadays[3]. From figure 1 we could see overall the total birth rate and natural growth rate are drop down before 2006 and the curve tends to be gentle in recent years.

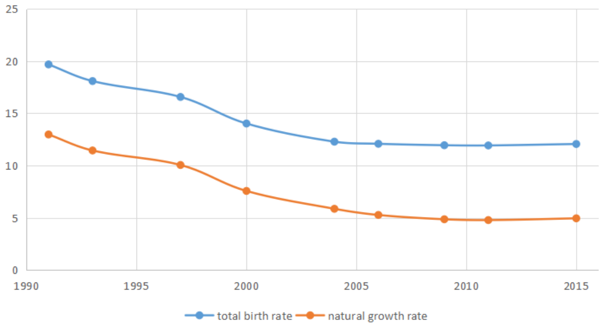


Fig. 1. total birth rate and natural growth rate in China

Induced abortion is performed due to provocation from the outside by intentionally terminating an unwanted pregnancy. Although abortion has been used as a family planning method for many years, it is a problem for women's health, especially in rural areas of developing countries. These procedures are often unsafe due to poor medical conditions, and the impact on well-being, health situation of every mother and her family, even the whole society and country may stand for a long time.

Although there are no clear overall abortion data each year in China for public on Internet, and we can't simply draw the link between population birth rate to 'one-child' policy directly, the implementation of the family planning policy has significantly changed China's demographic structure, prompting China to enter the ranks of aging countries in a very short period of time, forming the whole country a characteristic of "not getting rich but getting older at first"[4]. In response to the socialization problems that may be brought about by the aging of the population, China's population policy has undergone major adjustments in recent years. On December 28, 2013, a single child policy in which a couple who are only child in their family can give birth to two children is legally activated. Just two years later, in October 2015, the 18th CPC Central Committee decided, to fully against population aging, if one of a couple is the only child in his or her family, they could have two children legally. January 1, 2016, comprehensive two children policy is officially implemented.

Abortion data is often difficult to obtain, and the data accuracy is also difficult to guarantee. The overall birth ratio data online is impossible linked to the data of the post-abortion person, and therefore cannot be further analyzed and explained. Therefore there don't have too many research on post-abortion women considering their personal attributes in China[2], [4], [5], [6]. One related research recently shows that the more stringent the abortion policy, the more likely married women are going to have an induced abortion. Women become less vulnerable to enforced induced abortion during a loosened policy period[2].

III. The Data

The China Health and Nutrition Survey(CHNS) is an ongoing longitudinal large-scale survey[1], held for the first time in 1989 and there are overall ten times up to now. The latest release data including 2015 survey is just published on-line in May 2018. The survey uses a hierarchical multi-stage cluster random sampling principle. In terms of cities, choose provincial capitals and randomly choose medium-sized cities with poor economic development. In rural areas, the provinces are divided into three categories: high, medium and low. In the economic development level area, one county is selected for each advanced and low level, and two are selected from the intermediate level. In each city or county, four survey points were identified based on the principle of multi-stage cluster random sampling. The city attracted two urban neighborhood committees and two suburban villages, and the county established three villages where the neighborhood committee and county government are located[7]. The survey is divided into three dimensions: community, family and individual. The scope of the survey covers 12 provinces including Heilongjiang, Liaoning, Shandong, Jiangsu, Henan, Hubei, Hunan, Guizhou, Shaanxi, Yunnan,Zhejiang and Guangxi, as well as three direct-controlled municipality including Beijing, Shanghai, and Chongqing. This is a detailed micro-data which led us to the factors behind abortion, as well as personal, family and community characteristics. The survey for pregnancy history are started from 1991 and there are 9 total waves up to now.

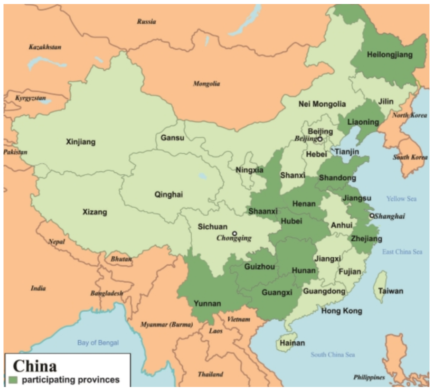


Fig. 2. Map of survey regions[8]

This paper applies to nine survey datasets includes nine waves(1991,1993,1997,2000,2004,2006, 2009,2011,2015) except the first wave because some essential data we need are not included in early 1989 survey.

There are 41 data files including the topic of agriculture, asset, business, childcare, education, urbanindex and so on. For this study 7 files including the most important 'ever married women', 'pregnancy history','and the factors related this study including 'education','health insurance','jobs','wages','urban' are chosen. Based on this detailed data, analysis of the relationship between post-abortion women and non-abortion women is possible.

Each file has many attributes in common, such as 'wave' means the year for survey,'IDind',a 12 digits number which represents a specific person, 'hhid' represents family ID, 'COMMID' means community ID. Based on same 'IDind' and 'wave', all files could be combined to study different factors behind the same person. Some 'IDind' shows more than one time in a file because it's a longitudinal survey and same person could be visited in different waves.

For every specific topic(data file), there are many variables in it and a related document to explain the meaning of all variables and range of the variable value. The original questionnaires also could be found on CHNS website.

IV. Hypothesis

The induced abortion ratio may change a lot in past 14 years due to policy and other factors. The difference between provinces, urban and rural may also have some impact on it.

For mother with induced abortion history and without, there may exist difference on education level, well-being level and financial level etc. For example, non-induced abortion women may be healthier than post-abortion women. On the other hand, how those factors may or may not have impact on whether women have a induce abortion history could also be inferred. For example, women with higher education level may have less possible to do abortion.

V. Data cleaning

There are 1.61GB original data including 41 different files in total which are all in sas7bdat format. All data related to this study are chosen after carefully study based on the original questionnaire documents includes 'ever married women', 'pregnancy history','education','health insurance','jobs','wages','urban'.

All 7 data files are changed to csv format which is easier to handle in many tools and this conversion are made by a python script using sas7bdat package. Among 7 files, the most important one for this study is 'ever married women' and 'pregnancy history'.

```
python sas7bdat_to_csv jobs.sas7bdat
```

All data cleaning need to based on 'pregnancy history'. There are 4126 observations(means 4126 pregnancies, not 4126 different mothers) and 30 variables in this file. Value of 'induced abortion' are changed to '1' and all others to '0' because the basic idea is to see what behind induced abortion. Among 30 variables, IDind_M(mother's ID), wave(survey year),Commid(ID of mother's community) 'S114'(whether abortion for this pregnancy),'S114A'(child gender),T1(province for mother), T2(urban or rural) are chosen for further data combination. Abortion ratio in different years and provinces are shown from this data.

Then the pregnancy data need to be cleaned to study each mother, means there is no duplicate mother's ID. So all records of same ID in different waves(survey a mother in different years) and same ID, more than one pregnancy obervations(one mother have more than one pregnancies) need to be removed. For the respondents who were interviewed for more than 2 times in the 9 surveys, only the last data(2015) was retained. For duplicate ID further(one mother with more than one pregancies), if there are induced abortion records, one of this records would be retained and if not, choose one record randomly. These methods could make sure all mother's ID are different and now the 'S114' means a mother with or without induced abortion history now. At this step, there are 3094 different mother's records retained.

Other datasets now could be combined now. In each dataset, the same 'IDind' and 'WAVE' could make sure the same mother. No matter how many observations in these datasets(education, health insurance etc.), we could only use 3094 records related to those mothers with pregnancy history. So the column would keep invariant and all rows become longer. All values for 'unknown'(-9 , 9 or other value in specific data files) are carefully deleted.

Among 60 variables in the data 'ever married women', 'S41' and 'S44', number of sons living with woman' and 'number of sons living outside home' are combined to number of sons, same with 'S42','S45' for daughters. 'S69' and 'S72' are combined to binary variable 'want another child sometime'.

Among 17 variables in 'education','A11' for 'completed years of formal education in regular school' and 'A12','Highest Level of education attained' are chosen. Among 52 variables in 'individual medical insurance','M1' for 'Do you have medical insurance' and 'M1A', 'description of person's health' are chosen.

In 'master individual ID file', 'WEST_DOB_Y' which means the western date of birth related to each ID are chosen to calculate the age at each survey wave. 'Primary occupation' and 'C8','average monthly wage last year' are chosen from individual jobs file and individual earnings file.

Finally, 'urbanization index','economic component socre','quality of health score' and so on are chosen from urbanization index file. This combination need to be done by compare 'commid' and 'wave' in original file rather than individual ID.

All data preprocessing are completed by python script for each task and the most common used code are shown as below, it combines the dataset of pregnancy history and ever married women using the same ID and wave.

```
import csv
import pandas as pd
for pitem in...
    for eitem in...
        #same ID and wave
        if pitem[0] == eitem[0] and
            pitem[1] == eitem[1]:
            newline = pitem + eitem
            newlist.append(newline)
name = name0 + name1
test = pd.DataFrame(columns=name, data=
    newlist)
test.to_csv(’final07.csv’)
```

VI. Data Analysis

There are two files after data cleaning, the first one has 4126 different pregnancy records named ’birth records’, the second one with 3094 different mothers divided by ’with induced abortion history group’(called abortion in some figures for convenience) and ’without induced abortion history’(called others in some figures). If there is no specification, all ’abortion’ means ’induced abortion’ only(not included natural abortion), in the contrary ’others’ including all situation without induced abortion: live birth, stillbirth, stillborn fetus, natural abortion. In a word, the two categories here(abortion and others) are mother with at least one induced abortion history, and mother without induced abortion history.

A. 4126 Birth records

The abortion ratio are calculated in this study by the number of induced abortion records divided by all birth records.

1) Different years: Figure 3 clearly shows the induce-abortion ratio declines a lot from year 1991 to 2015. In 1991 the abortion ratio 22.7% is rather high compared to following years, this curve before year 1997 also proves the viewpoint that the population policy is very strict in early years and gradually loosed after around year 1995. Since a clear deep drop down could be found from year 1991 to year 1997. That also means if every family and mother here has the freedom to choose, they may not end up with an induced abortion. At that time people may suffer huge punishment if they don’t obey this policy. From year 2011 to 2015 there is another very deep dropping and it may also related to a policy shown before, 2013 is the first year either husband of wife is the only one child in their family can legally give birth to two children.

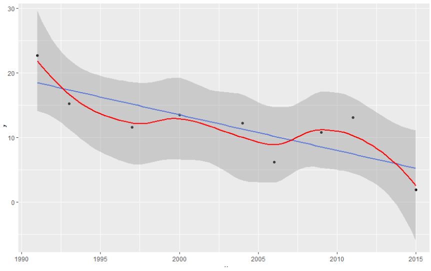


Fig. 3. Abortion ratio in different years

This fitted curve is drawn by R using ggplot command with a linear fitting and a default setting smooth fitting.

$$y = -1.128 * x + 2016.325 \tag{1}$$

(1) is the fitted line and the correlation coefficient here is -0.789. If there’s more data before 1991 and after 2015, more reliable fitted line would be get and it may not be a simple linear relation.

2) Different provinces: Figure 4 shows there are differences in different provinces on induced abortion ratio. Liaoning(20.1%) is the province with highest ratio, located at the near east of Beijing. Guangxi(3.9%) and Guizhou(2.9%) are two provinces with very low abortion ratio, and both locates at the most west south corner in China. Since they are very far from the political center-Beijing, people their may face more gentle one child policy implementations and political environment.

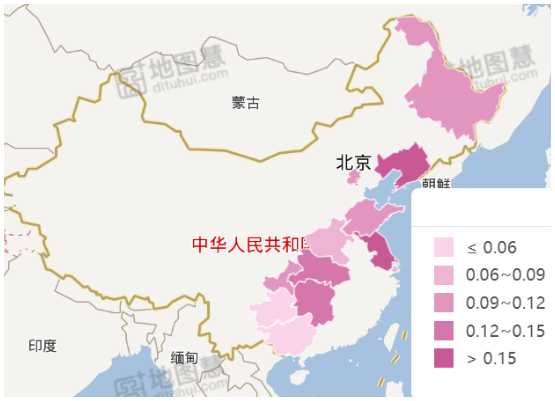


Fig. 4. Abortion ratio in different provinces

3) Different sites: Between rural and urban, the difference of abortion ratio is quite large. The result is same with the early study by Qiao[6], the abortion result in urban(15.7%) is much higher than rural(7.6%) site. On the one hand, urban abortion caused by unintended pregnancy due to contraceptive failure is relatively high[6];

on the other hand, urban people are not willing to have more children, and urban fertility policies are stricter compared to rural site relatively. So once pregnant, it is often possible to do abortion.

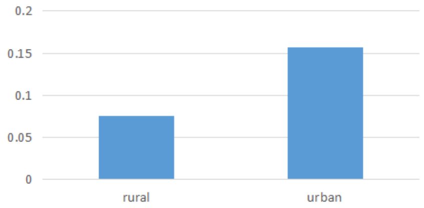


Fig. 5. Abortion ratio in rural/urban

B. 3094 mother records

After study pregnancy records above, all records related to each mother(all with pregnancy history)are analyzed in this second part. There are 3094 records revealing each mother's education, job, wage and other factors. The basic attribute in this study is based on whether a mother has induced abortion history(at least one induced abortion or not). There are 2752 women without abortion history(89%) and 342 post-abortion women(11%). The average age is 31 years old in the data and the range of age is 18-52.

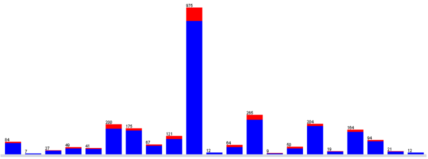


Fig. 6. Education years

1) Education: In figure 6 from Weka, value 23 in the middle is the most common one in education years records which means most women are only graduated from lower middle school. From left to right, means no school, 1 year primary,lower and upper middle school, to 4,5 years college, 6 years college or more at the most right. The red color means women with abortion history.

Figure 7 is a very interesting picture which shows the relation between education level and women abortion ratio. Here the women abortion ratio means the percentage of women with abortion history among all women with pregnancy history which is different from above. Women with 0(None education) and 3(Upper middle school degree) are more likely to do induced abortion compared with 2(lower middle school degree) and 5(university or college degree). From none(0),to primary(1),to lower middle school(2), and from upper middle(3), to technical or vocational degree(4) and university(5) we could conclude that the more education, the less abortion. However, women with upper middle school degree are much more likely to do abortion than lower middle school degree, and that make the overall situation more complex. A linear fitted line could be get by R.

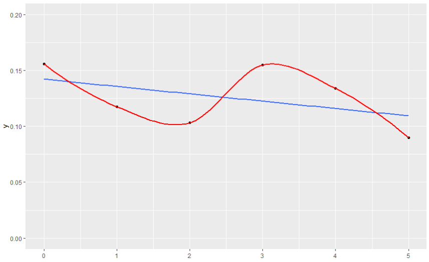


Fig. 7. Post-abortion women ratio with education level

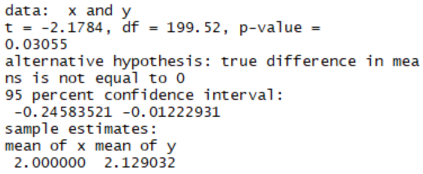
$$y = -31.185 * x + 6.427 \qquad (2)$$

Table1 shows the difference of education year value between women with or without abortion history. Here the data are calculated by R. The mean and 95% confidence interval shows overall the women without abortion history has a little more education years than those post-abortion women.

TABLE I
EDUCATION YEAR VALUE

education value	mean	sd	95%Confidence Interval
induced abortion	22.443	6.764	[21.678,23.210]
birth and others	22.836	6.838	[22.532,23.052]

2) Health: Health value is a description of person's health situation by themselves. 1 means excellent and 4 for poor, the less the better. Women without abortion history is healthier than post-abortion women according to their own feelings since the mean 2 is less than 2.12. The results shown from R in figure 8 reveals there has certain difference between the two groups on health value, since p-value in this t test is $0.03 < 0.05$. Normally there is a statistical difference when this value is less than 0.05. The difference of whether a woman has health insurance is remarkable since 59.2% is larger than 56.8%.



```
data:  x and y
t = -2.1784, df = 199.52, p-value =
0.03055
alternative hypothesis: true difference in mea
ns is not equal to 0
95 percent confidence interval:
 -0.24583521 -0.01222931
sample estimates:
mean of x mean of y
 2.000000  2.129032
```

Fig. 8. R results on health value between women with or without abortion history using t.test()

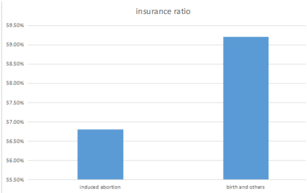


Fig. 9. Have health insurance ratio

3) Jobs and wages: Figure 10 shows the ratio of very job among women. 40% women without abortion history are farmers. Farmers are still the majority in China although the voice of them can't be heard usually. For farmer job, the women without induced abortion history(red) are more than post-abortion women. Since farmers are widely distributed in rural areas, this results also verified the discovery above that people in urban are more likely to do abortion than rural.

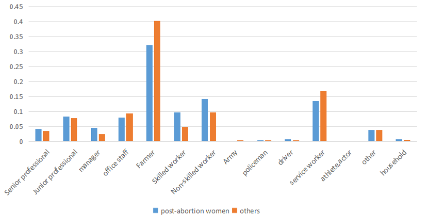


Fig. 10. Jobs difference

Figure 11 and 12 shows there is no significant difference between women with or without abortion history on wages. Here the wage is refer to every month payment and the unit is RMB. It should be mentioned that although the wages are increased over these years but purchasing power is doubtful if other attributes are considered at the same time.

4) Community: All communites mothers live in are compared from the factor of economy, house situation, sociality, population density, overall health and sanity. There are no obvious difference between the environment post-abortion women live in and others live in as shown in figure 13.

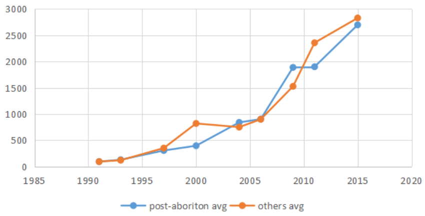


Fig. 11. Avg. wages

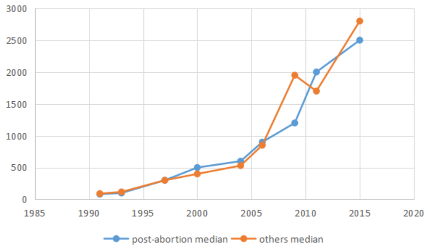


Fig. 12. Wages median

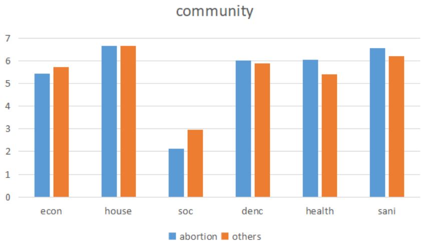


Fig. 13. Community

5) Others: The son/daughter ratio is quite different between post-abortion family and non-abortion history family. Here it is calculated by number of sons divided by number of daughters in all post-abortion women family and others. Ideally, this ratio should be 1 means the number of sons and daughters are same. The 0.82 may be just the data distribution in this survey because it is strange when this ratio is less than 1. However, the 1.56 in post-abortion family is so impressive which means nearly 2 of 3 children in these family are son. This result is related to the traditional concept in China that boys are better than girls, and the widely used B-Ultrasound test which could check the gender of child in mother's womb in the past.

Figure 14 also shows that post-abortion women have more tendency on wanting another baby sometime which makes sense. Figure 15 is the result of T test from R which compares the number of hospital checks for last pregnancy between women with abortion history or not.

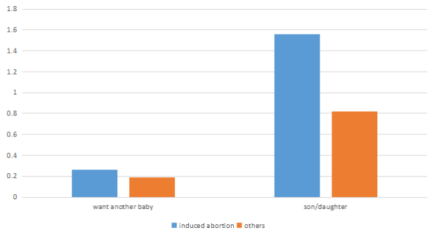
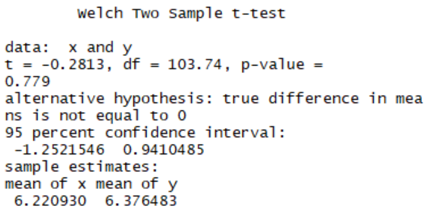


Fig. 14. want another baby, gender ratio



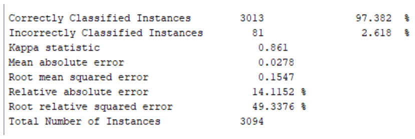
```
        Welch Two Sample t-test

data:  x and y
t = -0.2813, df = 103.74, p-value =
0.779
alternative hypothesis: true difference in mea
ns is not equal to 0
95 percent confidence interval:
 -1.2521546  0.9410485
sample estimates:
mean of x mean of y
 6.220930  6.376483
```

Fig. 15. Num. of hospital checks for last pregnancy

The mean of post-abortion women is slightly less than others but there is no big difference which means all mothers care about their babies.

VII. Classification and Forming a Model

In this section some models are formed to perdict whether a woman has induced abortion history based on combination of factors including survey year(wave), number of pregnancy, from rural or urban, education level and years, contraception or not, whether have health insurance, number of hospital checks when last pregnancy, number of children, want a baby or not, health value and job type. The model is formed by weka and all nominal attrubtes are correctly changed from default numeric type to nominal by the NumericToNominal filter in Weka.



```
Correctly Classified Instances        3013               97.382  %
Incorrectly Classified Instances        81                2.618  %
Kappa statistic                          0.861
Mean absolute error                      0.0278
Root mean squared error                  0.1547
Relative absolute error                 14.1152 %
Root relative squared error             49.3376 %
Total Number of Instances             3094
```

Fig. 16. KNN Statistics from Weka

The baseline here is 89% because 89% women don't have induced abortion history in the data. In a Random forest model with default setting, this held a 96.83% correctly classified instances and kappa statistic is 0.82 here which indicates this classifier is acceptable at average level. The Logistic model held a 89.98% correctly classified instances which is just slightly better than baseline. The third model here is KNN with default parameters, it held 97.38% correctly classified instances with 0.86 kappa statistic. The precision for 0(not abortion) is 98% and for 1(have abortion history) is 92.2%, which is the best model to predict.

VIII. Conclusion

This paper mainly studied two questions.

(a) The change of overall abortion ratio according to 9 different survey waves and 12 provinces in China, and rural/urban site.
(b) To access factors such as education, health and job etc. associated with induced abortion women among women with pregnancy history in China.

For the first question, based on 4126 pregnancy records, it shows that the induced abortion ratio has been declined deeply from 22.7% in 1991 to 1.9% in 2015. It could be concluded that the strength of policy implementation do have a huge impact on women's induced abortion ratio since for every big change,(from 1991 to 1997, and from 2011 to 2015) there are some policies changed at that time.(1995 loosed one child policy, 2013 open two children policy). For provinces, the difference is also very large, Liaoning has 20.1% abortion ratio compared with 2.9% in Guizhou province. Women in urban(15.7%) are more likely to do abortion than rural(7.6%) area.

For the second question, education, health, jobs, wages, community, son/daughter ratio are evaluated between women with abortion history and without abortion history among all women with pregnancy history between age 18-52. Some factors don't have too much effect including community they live in, wages, number of hospital checks during last pregnancy. The overall average education level of non-abortion group is higher than abortion group. Non-abortion women group also have more people with health insurance, more farmer, a little more healthier, compared with post-abortion group. In addiation, son/daughter ratio in post-abortion group is much higher than non-abortion group and post-abortion women are also more interested to have another baby.

At last, KNN, logistic regression and Random Forest model are used to predict whether a woman has induced abortion history based on some factors above. The KNN model got the best result on training data with 97.38% correctly classified instances.

IX. Extensions

A. *Controlling other factors*

Since the observations are not averagely distributed among 9 waves and 12 provinces in the data, it may affect the results when we sum all observations from different years and provinces to study factors such as education and health. If time is enough, those factors should be evaluated by controlling for other factors such as year, province, age and so on.

B. Provinces and economy

In figure 4 those two provinces with lowest abortion ratio at south west are farest from capital Beijing in survey, and both are also not very rich provinces. The relation between a province economy and abortion ratio may also be a question to study. Further, other factors within a province such as nationality structure of population may also affect abortion ratio.

C. Children in post-abortion family

Since there are Idind_C value indicated every child ID in women pregnancy data which could be found in other files in the CHNS data, the question of whether there's difference between children in post-abortioin family and other family could be answered. The possible evaluations could be obesity rate, whether live with parents and so on.(We only study son/daughter ratio)

Acknowledgment

This research uses data from China Health and Nutrition Survey (CHNS). We thank the National Institute for Nutrition and Health, China Center for Disease Control and Prevention, Carolina Population Center (P2C HD050924, T32 HD007168), the University of North Carolina at Chapel Hill, the NIH (R01-HD30880, DK056350, R24 HD050924, and R01-HD38700) and the NIH Fogarty International Center (D43 TW009077, D43 TW007709) for financial support for the CHNS data collection and analysis files from 1989 to 2015 and future surveys, and the China-Japan Friendship Hospital, Ministry of Health for support for CHNS 2009, Chinese National Human Genome Center at Shanghai since 2009, and Beijing Municipal Center for Disease Prevention and Control since 2011.

References

[1] B. M. Popkin, S. Du, F. Zhai, and B. Zhang, "Cohort profile: The china health and nutrition survey—monitoring and understanding socio-economic and health change in china, 1989–2011," *International journal of epidemiology*, vol. 39, no. 6, pp. 1435–1440, 2009.

[2] C. Wang, "The impact of the state's abortion policy on induced abortion among married women in china: a mixed methods study," *Chinese Sociological Review*, vol. 49, no. 4, pp. 316–339, 2017.

[3] C. S. Yearbook *et al.*, "National bureau of statistics of china," *China Statistical Yearbook*, 2018.

[4] W. Kaiyong, D. Jun, and W. Fuyuan, "Influence of the implementation of the universal two-child policy on demographic structure and population spatial distribution in china," *Progress in Geography*, vol. 35, no. 11, pp. 1305–1316, 2016.

[5] Y. Che, E. Dusabe-Richards, S. Wu, Y. Jiang, X. Dong, J. Li, W.-H. Zhang, M. Temmerman, and R. Tolhurst, "A qualitative exploration of perceptions and experiences of contraceptive use, abortion and post-abortion family planning services (pafp) in three provinces in china," *BMC women's health*, vol. 17, no. 1, p. 113, 2017.

[6] Q. Xiaochun, "Analysis of induced abortion of chinese women," *Institute of Population Research*, vol. 26, no. 3, pp. 16–25, 2002.

[7] L. XIN and P. LI, "Food consumption patterns of chinese urban and rural residents based on chns and comparison with the data of national bureau of statistics," *Journal of Natural Resources*, vol. 33, no. 1, pp. 75–84, 2018.

[8] a UNC Carolina Population Center project, "Map of survey regions," https://www.cpc.unc.edu/projects/china/about/proj_desc/chinamap.