DHIIVAT: Disease History & Income Inequality Visual Analytics Tool for Exploring Relationship between Disease and Income Inequality in China— Using panel analysis

Lin Lin Jiang ,Ya Jung LEE, Yue Chen WANG

Abstract—With China economic rapid development,people have considered more about health.In order to verify the relationship between wealth and health,we use panel analysis,taking consideration of individual income, income inequality,individual control variables to optimize the linear probability model. Beyond statistical model,we build an interactive visual tool to help adminstrations acquire valueable information from province level data and select variables to customize the model.We design visualization based on dataset specialities and practical need in analysis process,with the use of interactive plots like lorenz curve,error bar, so that the user could have better understanding of implicit problems under the data. Visualization flexibility,statistic modelling function and other potential of the tool is demostrated by our use case.

Index Terms—Visualization, Panel analysis,Income inequality,Gini, logistic probability data, health

1 Introduction

Income inequality is a hot issue these years, many sub-topics have been derived from this theme. Scholars have studied about the root cause that result in this situation, the post-impact of income inequality and so on. Our paper is to explore the relationship between individual income inequality and health in China using panel survey data with 8 provinces across 5 years.

Data from CHNS (China Health and Nutrition Survey) offers good resource for us to explore the income inequality problem in China. It will be an extension study of an existing work,Income inequality and health in China: A panel data analysis.(Bakkeli,2016)

Beyond original work scope, we use updated survey data to verify the explanatory model illustrated in this paper.Also we created interactive visualization toolkit with multiple functionalities to support policy decision process. We use R and R Shiny to build this web-based toolkit.

2 Motivation of the application

China administration has aroused more attention on Chinese health and nutrition, based on this background, a project called The China Health and Nutrition Survey (CHNS) has been proceed for tens of years conducted by Chinese Center for Disease Control and Prevention (CCDC) and National Institute for Nutrition and Health (NINH). CHNS offers abundant data for this disease history and income inequality relationship analysis.

We have found the situation that with economic rapid development, Chinese health situation hasn't improved simultaneously these years. From China cardiovascular diseases report (2015), fertility rate of cardiovascular diseases was increased by year from 2006-2014 in both urban and rural area. So, we want to establish a visual analytic tool to explore the correlation between income and disease history, at mean time, the tool can work as an assistant to help the authority make their decisions.

Lin Lin JIANG is a postgraduate student (MITB FTA Progrmme) at the School of Information Systems, Singapore Management University, Singapore (e-mail: lljiang.2019@mitb.smu.edu.sg)

Ya Jung LEE is a postgraduate student (MITB FTA Progrmme) at the School of Information Systems, Singapore Management University, Singapore (e-mail: yjlee.2019@mitb.smu.edu.sg)

Yue Chen WANG is a postgraduate student (MITB AT Progrmme) at the School of Information Systems, Singapore Management University, Singapore (e-mail: yuechenwang.2019@mitb.smu.edu.sg)

Following are purposes of DHIIVAT:

(1) Interactive reporting interface to view survey data in different provinces along different years.

(2) Explore variables correlation to disease history.

(3) Collaborate linear probability model.

3 Review and critic on past works

For this CHNS (China Health and Nutrition Survey) dataset, many researchers have published paper to study health problem in China from different perspective. For example, gender, rural-urban, society issuance and income. Furthermore, they specific the health problem into many aspects: overall health, individual health, household health, mental health, physical health. This paper is to do further analysis on the relationship between income and physical health based on Bakkeli's research (2016). In Bakkeli's research, the scholar built the model on county level but in this paper, we built province level model.

3.1 Variable Measurement

There is different measurement for income and health, every choice for measurement has its own pros and cons. Scholars have two ways to measure health, one is self-reported health and the other is physical health. Self-reported health may include bias regarding of age, gender, education (Dowd& Todd,2011), so it will be better choice to choose physical health for the measurement. Fertility rate is a common way to measure physical health, but in this survey data, fertility rate may be 0 through all these years among these families, in other words, it cannot show the difference level of health among households. In Bakkeli's research (2016), the author picked BMI, WHR, MAMC, blood pressure to be the measure of physical health which are objective factors. Due to limitation of CHNS open dataset, our research use cancer history and chronic history to record health situation, because these variables are not like BMI, WHR which are commonly applicable to everyone. With this limitation, the paper will investigate the relationship between disease history and income inequality.

In terms of measure income, previous work uses Individual/ household income, county level income, and income gap (GINI index and Generalized Entropy).In this research, because the model is built on province level and survey data is collected by individual

unit. So，individual income, province level income and income gap are used to measure income part.

3.2 Visualization

Bakkeli's research (2016) only used line graphs to present China Gini-coefficient change with year change. And use facet map to compare Gini-coefficient change in different county. Other individual control variables, as well as the correlation between independent variables and dependent variable, health index, were not presented in visualization way.

According to dataset feature and purpose of DHIIVAT, we apply multiple visualization method to present statistics result, to make the tool more user-friendly for policy makers, the administration not only data scientists.

3.2.1 Data Feature

This data is panel survey data with data collected from 8 different provinces with different sample size across 5 wave(survey year).Considering about this feature, we need to find appropriate way to consider about the uncertainty of survey data(Yu, Z. Y.,2018; Nathan Yau,2018), as well as the way to present data changing with time. Common method to present data uncertainty is error bar, which shows average value, and value range with given confidence interval.

3.2.2 Visualization for compliment Toolkit functionality

The common method to view the raw data from survey is to view distribution. Tracy R. & Robert B. built framework of basic visualization design and their use case. For example, box plot, histogram plot, bar chart.

For Time-series data, Few, S., & Edge, P. (2007) offered multiple available visualization insight to present this kind of temporal data, such as line graph, bubble plot; also illustrated the importance of interpreting slope and direction of change when using line graph.

For Gini-coefficient presentation, the previous work (Bakkeli's research,2016) only uses line graph, however we can use the common interpretation way-Lorenz curve (Kakwani, N. C. ,1977) to present the income inequality.

For model building preparation, the necessary step is to conducted hypothesis test and correlation test between independent variable and dependent variable. (Swinscow& Campbell,1997).

To compile all visual outcome in order (Chris Stolte,2009),we need to respect human viewing behaviour, mapping corresponding graphs at the same page and excluding useless decoration. Above that, we can make full use of interactive function, such as brush, highlight to serve for better data illustration.(Shneiderman, B,1996).

4 Data and Methods

4.1 Variables

The data we use to build this toolkit is taken from the Chinese Health and Nutrition Survey (CHNS) with 5 waves, 2004, 2006, 2009,2011 and 2015. There are total 468 consistent individuals from 8 provinces. Observations are limited to people with a job, aged 16 to 69.

Variables selected from survey result are individual ID, WAVE(survey year), individual income, cancer history, chronic disease history, province population. In addition, individual control variables are involved: age, gender, years of education, majority, marriage status, urban/rural, occupation, occupation type.

This research measures income inequality at province level by every wave. Income inequality indices are represented by the three variables, which are Gini-coefficient, Generalized Entropy (Theil) indices: Theil L, Theil T. Theil L is the mean logarithm deviation, which is sensitive to changes at the bottom income levels. Theil T is also known as the Theil index and is sensitive to changes in upper income levels. We compute average income at province level for model building, as a result of considering about the income gap among provinces.

4.2 Model

Based on panel analysis Principle we need to analyse data horizontally and vertically: Compare same individual's change across years and different individuals in the same year. Fixed effects are WAVE (year) and Individual ID. Because the survey was conducted in different provinces, so we use province average income.

The linear probability model is written below:

$$DiseaseHistory_{ict} = \alpha + \beta_1 Inequality + \beta_2 Income + \beta_3 ProvinceAvgincome + \gamma X_{ict}$$

β stands for the coefficients of independent variables

X is a vector of individual control variables, γ stands for the coefficients of the control variables.

In following demonstration of model building use case, we can build three different models with different representation of income inequality.

5 Design Framework

5.1 Toolkit Architecture

DHIIVAT is established on R framework, using free and open source statistical software R (https://www.r-project.org) and the R package shiny (https://cran.r-project.org/web/packages/shiny). This opensource software can be applied to multiple operation systems, operating in Windows, Mac OS X, and Linux environment. Key R packages used in DHIIVAT for visualization purpose, displayed in below table:

Table 1 R Packages introduction

Package	Version	Descriptions
shiny	1.4.0	Interactive web applications for data visualization
ggplot	3.3	High-quality graphs
ggstatsplot	0.4.0	ggplot2' Based Plots with Statistical Details
gglorenz	0.0.1	Plotting Lorenz Curve with the Blessing of 'ggplot2'
ExPanDaR	0.5.1	Use for panel data modelling
corrplot	0.84	Create correlation matrix
plotly	4.9.1	Create interactive Web Graphics

The DHIIVAT toolkit is an off-line interactive visualization tool in html format.

5.2 USER INTERFACE and Functionality

DHIIVAT toolkit are constructed by 4 tabs, which are introduction, EDA analysis, Pairwise correlation, Panel Data Modelling. These four tabs are placing sequentially according to analysis procedure.

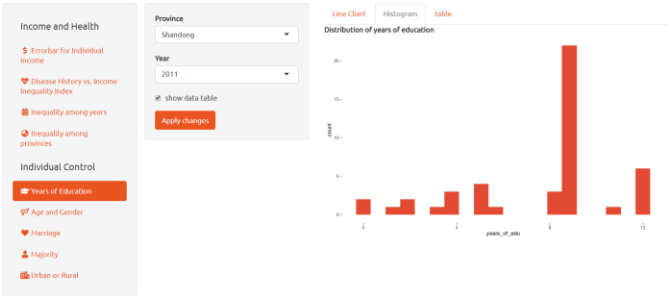


Figure 1 User Interface

In EDA tab, the user can select the variables they want to explore by clicking the variable name under 'Individual Control' menu. The user just tick the box, select from drop-down box to set up the input information, after clicking the submit button, the graph will generate automatically. Present in both graphs and data table format as figure 1 shows, facilitate users to get multi-dimensional understanding of data.

Besides basic information extracted from survey data, further statistics on variables are available for user to explore. At 'Income and Health' menu, it is more about key variables like income ,income inequality and visual exploring relationship between income inequality and disease history.

The following two tabs are served for regression model building. The user can select variables by 'Pairwise correlation' and Hypothesis test in 'Panel Data Modelling' tab. Then the user can select variables by tick the box, and model result appears on the right side after clicking 'Apply Changes'.

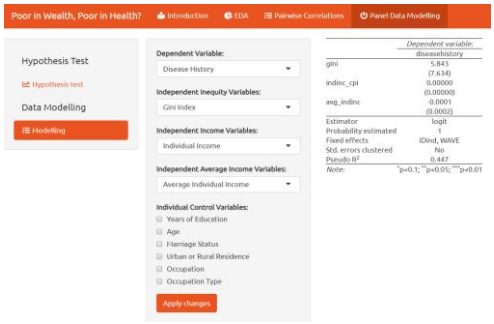


Figure 2 Model Building Panel

6 DEMONSTRATION BY USE CASE

6.1 Data exploration

The first step for real use case may check data of individual control variables. Below is sample for 'Marriage' variable. Marriage age increases gradually with time, which means people's married status are more and more stable, while a sudden drop in 2006. Another obvious trend is that female marriage rate drops and male marriage rate increases very small, although the total marriage situation increases. This trend reveals that more people consider to stay single.

Figure 3 Marriage Status and Rate

Population pyramid can offer important information to the administration, since the health infrastructure planning needs to take aging situation into consideration.

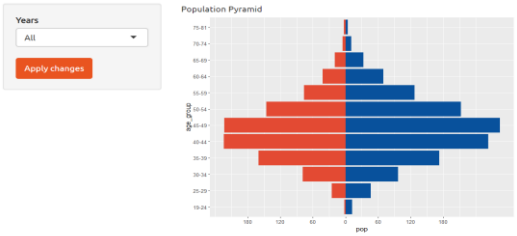


Figure 4 Age pyramid

Select 'Liaoning' from drop-down list to view distribution of ethnic and urban& rural residence in this province. Distribution is shown by histogram graphs below. We can find that most of the people in Liaoning are Han and keep rural household registration.

Figure 5 Ethnic & rural residence

Select provinces 'Guangxi' and 'Heilongjiang', and remain CI as 95%. Although average individual income of both provinces are both increased, the gap between them are expand gradually. The centre point of blue line is always higher than red one, but blue line's length are overlapped and longer, wider range means higher uncertainty, which means we have 95% confidence to say that income in Heilongjiang is definitely higher than income in Guangxi.

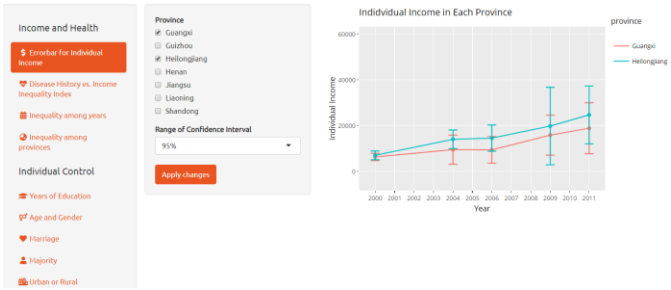


Figure 6 Individual Income

To explore variables correlations, this violin Plot below shows how Inequality Indexes affect Disease History, which is our dependent variable. We can see the distributions of Theil L are different in binary Disease History situation, thus Inequality Index has some influence on our target variable ,besides, this violin plot also displays the number of people in each inequality level. For example, in the level 0.2, there is no people with disease history.

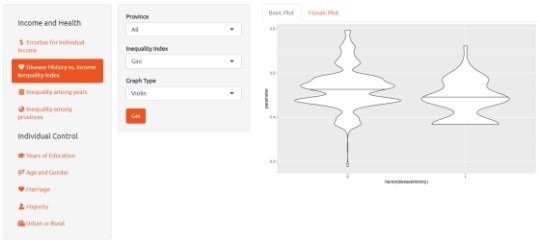


Figure 7 Relationship visualization explore

The other way to explore the relationship is by below Mosaic Plot which groups Inequality Index to four different levels, Equal, Reasonable, Large Gap, and Unequal. From this plot, we can find Equality level has more unhealthy people compared with other

levels, but p-value of Theil T influence on Disease History is larger than 0.05, which means statistically this hypothesis is not correct.

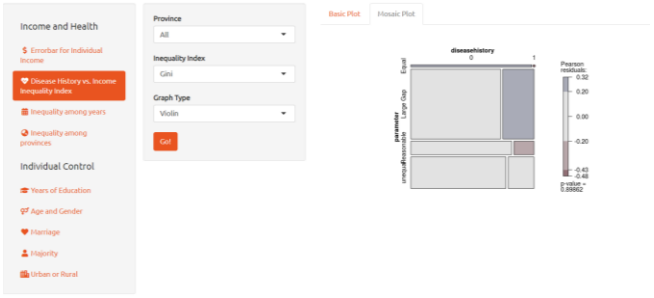


Figure 8 Mosaic Plot

The policy maker may want to compare Gini coefficient comparing among provinces or observe change by years, so that they can better design policy according to province specialities.

This Lorenz curve shows 4 years GINI coefficient in Liaoning. After comparing these curves, we found that the income inequality situation increases from 2000 to 2011. However, the change between 2004 and 2009 doesn't have huge difference.

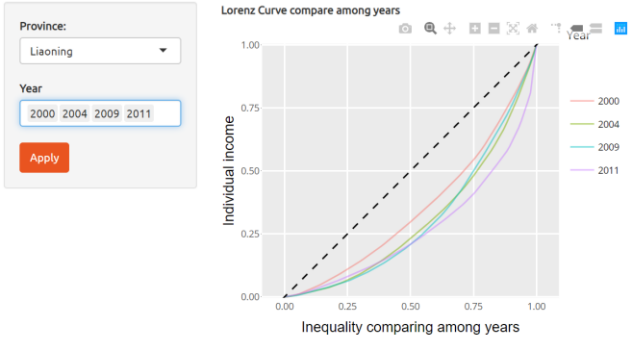


Figure 9 Lorenz Curve

6.2 Variable Selection

Before clarify all variables to collaborate a model, it is import to do variables selection. We can use correlation matrix as below to find highly correlated independent variables and make final decisions.

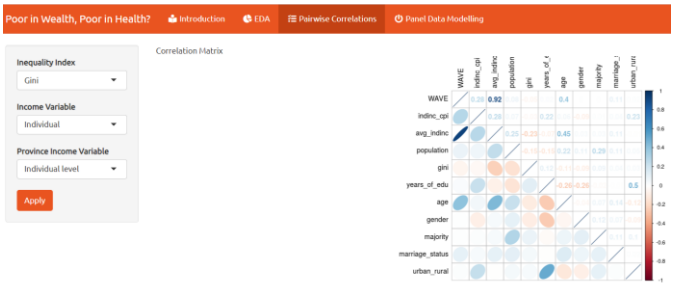


Figure 10 Pairwise correlation

From this Correlation Matrix we can directly understand the relationship among all variables. If choosing Gini Index as our inequality variable, we can find year (WAVE) and individual income (indinc_cpi) have highly positive correlation, 0.92. Other variables such as age vs. average individual income and years of education vs. urban residence have middle positive correlation. There is no obvious negative correlation among these variables. Users can consider drop some variables with low correlation like gender, population, and marriage status.

Also we can check for whether continuous variables are statistically significant to dependent variable—disease history by hypothesis test. (figure 11)

For example, when choosing individual income as individual feature, test method as parametric. We can see that the p value <0.001,so we can reject null hypothesis, which means individual income has some influence on people's health.

Figure 11 Hypothesis test

There are four available testing methods in this functionality, which are parametric(ANOVA/t-test),non-parametric(Wilcoxon),Robust(ANOVA/t-test with median) and Naïve Bayes factor illustrated by ggbetweenstats package documentation (CRAN,2020).The robust test and naïve bayes factor test are not for common usage as comparing models. Naïve byes only valid for between-subjects tests as the website clarified.

6.3 Model building

We decide to test the effects of individual income vs. people's health situation, and income equality vs. people's health situation using Linear Probability Model on our panel data. Considering we have three different indexes to measure income equality, three models are built which use Gini Index, Theil T, and Theil L respectively. Other independent variables and dependent variable are keeping the same.

After fixing Individual ID and Year to certain values, we consider three variables, income equality, individual income, and average individual income in different provinces as our major variables, and see do they have significant influences on dependent variable, Disease History.

From our dashboard, we give users many flexibility that they can choose which inequality variable and individual control variables they want to include in model. However, to show complete and formal results in this report, we put all variables into our models and use following three pictures to illustrate.

§Model 1

In this model, we use Gini Index as our inequality variable but neglect all individual control variables and run the model. The Pseudo R^2 is 0.447,so the model cannot work well for explanatory purpose.(Shows in figure 12)

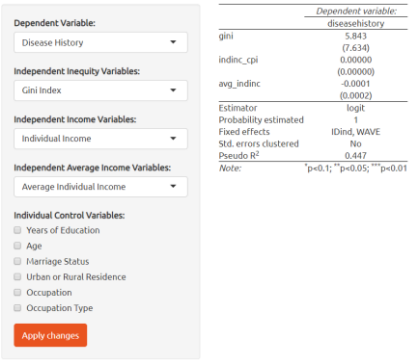
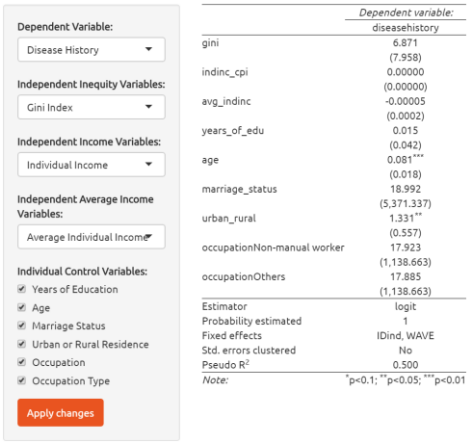


Figure 12 Model 1

§Model 2

On the basis on model 1, we take all individual control variables into consideration to the new model. Result shows in figure 13, Pseudo R^2 is 0.5, which means this model has some interpretive ability among dependent and independent variables. Besides, we can see only two individual control variables, Age and Urban_Rural, have significant influence on people's health situation. When people's age increase, they are more likely to have disease history. On the other hand, people who have urban household registration are more easily to have diseases compared with those who from rural area.



	Dependent variable:
	diseasehistory
gini	6.871
	(7.958)
indinc_cpi	0.00000
	(0.00000)
avg_indinc	-0.00005
	(0.0002)
years_of_edu	0.015
	(0.042)
age	0.081***
	(0.018)
marriage_status	18.992
	(5,371.337)
urban_rural	1.331**
	(0.557)
occupationNon-manual worker	17.923
	(1,138.663)
occupationOthers	17.885
	(1,138.663)
Estimator	logit
Probability estimated	1
Fixed effects	IDind, WAVE
Std. errors clustered	No
Pseudo R^2	0.500
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 13 Model 2

§Model 3 & Model 4

In model 3 and model 4, inequality variable is replaced by Theil T and Theil L Index respectively. Changing inequality index doesn't affect Pseudo R^2 value. Only Age and Urban_Rural variables have significant effects on Disease History.

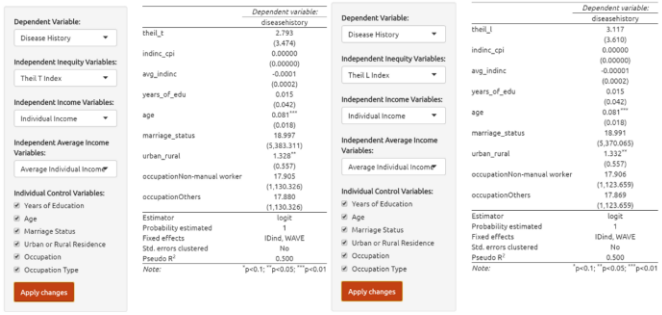
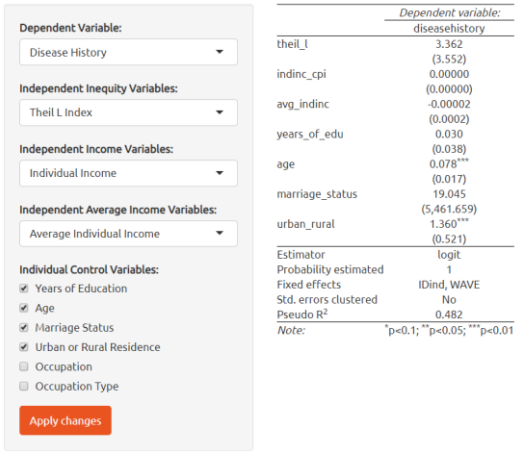


Figure 14 Model 3 & Model 4

§Model 5

The user may wonder whether individual control really can help to increase the explanatory effect of this model. So, in this model, we keep Theil L as inequality index, and select all individual control variables except Occupation and Occupation Type. The pseudo R^2 decreases a little comparing to model 4 to 0.482. We can get the conclusion that the more individual control variables, the better explanatory effect.(Figure 15)



	Dependent variable:
	diseasehistory
theil_l	3.362
	(3.552)
indinc_cpi	0.00000
	(0.00000)
avg_indinc	-0.00002
	(0.0002)
years_of_edu	0.030
	(0.038)
age	0.078***
	(0.017)
marriage_status	19.045
	(5,461.659)
urban_rural	1.360***
	(0.521)
Estimator	logit
Probability estimated	1
Fixed effects	IDind, WAVE
Std. errors clustered	No
Pseudo R^2	0.482
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 15 Model 5

7 Conclusion

From model building process and outcome ,we can draw the conclusion that income inequality has no significant effect on our independent variable Disease history. Although the outcome is different from our original purpose, we can still prove that people's age and living in urban or rural area can affect their health situation. Besides these two variables, individual income is statistically significant to our dependent variable Disease history from our hypothesis result(shown as figure 11).

In terms of DHIIVAT use in real case, they are two potential users:

On one hand, our model can provide direct visualization and analysis results for scholars who have interests to research the influences of economy and demographic factors on people's health situation. On the other hand, when China government need to set up some policies related to people's health, they may want to collect basic demographic data in different provinces such as gender, age, marriage status, education level, and their household registration are urban or rural.

From our EDA part, related organization can collect all insights rapidly and find out which provinces need more help on these parts. They can use our model to check which variables have significant influence on people's health. For example, from our conclusion, we found that elderly people and people who are born in urban areas may have higher possibility to have diseases. Hence, related organization should pay more attention on medical care for the elders or give more benefits to people from rural areas but work in cities and have healthy issues.

8 Future Work

According to our raw data, there is serious problem on our dependent variable. Compared with the article- Income inequality and health in China: A panel data analysis published by Nan Zou (Bakkeli,2016).We only consider about two disease history variables as our Disease History variable to measure health rather than four commonly applicable health variables in original paper.

Cancer history and chronic diseases are limited to measure people health. Only considering these two factors, we don't have enough unhealthy cases to build model, which affects our model results. We suggest that people who want to use our model can use charged channels to get more complete data for research, which can help to get more accurate and interesting insights.

In terms of statistical method, unlike original paper, we built province level model rather than county level model. For fixed effects, Bakkeli fixed different disease to build model but we fixed ID and WAVE. Also we use logit estimator rather than OLS. These

change may be reason that we draw different result comparing to Bakkeli's research. We need to collect abundant data and conduct further analysis to optimize our model and re-verify the result.

In terms of use of DHIIVAT, we can add more functionality like insert data, change data, delete date so that we can offer more flexibility to satisfy user's need.

REFERENCES

[1] Bakkeli, N. (2016). Income inequality and health in China: A panel data analysis. Social Science & Medicine, 157, 39–47.

[2] China cardiovascular diseases report 2015: a summary from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5329726/ Accessed 25 Apr 2020

[3] Dowd, J., & Todd, M. (2011). Does Self-reported Health Bias the Measurement of Health Inequalities in U.S. Adults? Evidence Using Anchoring Vignettes From the Health and Retirement Study. Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 66B(4), 478–489.

[4] Yu, Z. Y.(2018). Visualizing Uncertainty What We Talk about When We Talk about Uncertainty (Doctoral dissertation, Northeastern University).

[5] Nathan Yau, Visualizing the Uncertainty in Data. FLOWINGDATA Retrieve from: https://flowingdata.com/2018/01/08/visualizing-the-uncertainty-in-data/;2018. Accessed 25 Apr 2020.

[6] Few, S., & Edge, P. (2007). Visualizing Change. Visual Business Intelligence Newsletter, September.

[7] http://www.perceptualedge.com/articles/visual_business_intelligence/visualizing_change.pdf Accessed 26 Apr 2020

[8] Kakwani, N. C. (1977). Applications of Lorenz curves in economic analysis. Econometrica: Journal of the Econometric Society, 719-727.

[9] Tracy R., Robert B. Which chart or graph is right for you?

[10] https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51Acceessed 26 Apr 2020

[11] Chris Stolte(2009). Enhancing Visual Analysis by Linking Multiple Views of Data.

[12] https://www.tableau.com/sites/default/files/whitepapers/enhancing-visual-analysis-tsi-_0.pdf Accessed 26 Apr 2020

[13] Swinscow, T. D. V., & Campbell, M. J. (1997). Correlation and regression. Statistics at square one, 11. https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/11-correlation-and-regression Accessed 26 Apr 2020

[14] Shneiderman, B(1996). The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. Proceedings 1996 IEEE Symposium on Visual Languages, 336-343.

[15] CRAN. (2020) Package 'ggstatsplot'.ggbetweenstats.15-20

[16] https://cran.r-project.org/web/packages/ggstatsplot/ggstatsplot.pdf

[17] Accessed 26 Apr 2020

[18] Indrajeet Patil(2020) ggbetweenstats

[19] https://indrajeetpatil.github.io/ggstatsplot/articles/web_only/ggbetweenstats.html Accessed 26 Apr 2020

APPENDIX

Tabel 1 Variables description

Variable	Description
Health	
Disease History	Binary variable. Defined as 0 if individual has none of chronic disease or cancer. 1 means has medical history.
Income inequality	
Gini	The Gini-coefficient in the county level, sensitive to changes at middle income levels.
Theil L	The mean logarithm deviation of the Generalized Entropy (Theil) indices, which is sensitive to changes at the bottom income levels.
Theil T	The Theil index and is sensitive to changes in upper income levels.
Income variables	
Individual income	The sum of each individual's income source, by adding up all individual income and revenue, minus individual expenditures. Household subsidies and other income that cannot be allocated to individuals in the household are not considered as a part of individual income.
Province mean income(ind.)	Captures the degree of economic development in a province-level unit, calculated by averaging individual income in a province for all observations in the CHNS. "Ind" refer to individual.
Household income	The sum of all individual incomes in a household.
Province mean income(hh.)	Calculated by averaging household income in a province for all observation in the CHNS. "hh" refer to household.
Individual controls	
Age	The age of respondent.
Gender	Binary variable. Defined male as 0 and female as 1.
Marriage Status	Binary variable. Married as 1, unmarried as 0.
Majority	If the nationality of object is Han, then defined as "1", else 0.
Years of education	Calculated from the beginning of primary school, 6 years of primary school graduation, 9 years of junior high school graduation, 12 years of high school graduation, and 16 years of university graduation.
urban	Binary variable. If respondent holds urban household registration then defined as 1, else 0.
Occupation	
Services class	Includes "senior professional/technical", "administrator/executive/manager" and "army officer/police officer".
Non-manual worker	Includes "junior professional technical" and "office staff".
Skilled worker/supervisor	Includes "skilled worker" and "ordinary soldier, policeman", "driver" and "athlete, actor, musician".
Semi-/non-skilled worker	Includes "non-skilled worker" and "service worker".
Farmer	As originally defined by CHNS data.
Others	The rest of original occupation covered by CHNS data.

Occupation Type	
State	Includes "government", "state service/institute" and "state-owned enterprise".
Collective	Includes "small collective enterprise" and "large collective enterprise".
Family farming	As original variable "family contract farming" of CHNS data.
Individual enterprise	As variable "private, individual enterprise", which originally defined by CHNS
Private three-cap Enterprise	The same as "three- capital enterprise" in CHNS data.
Others	Includes "unknown" data in CHNS.