The Report Committee for Haoran Zhang
Certifies that this is the approved version of the following report:

Does Trade Cause Inequality

APPROVED BY

SUPERVISING COMMITTEE:

Tom Shively, Supervisor

Kishore Gawande

Does Trade Cause Inequality

by

Haoran Zhang, B.S.

REPORT

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE IN STATISTICS

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2015

Dedicated to my Parents and Supervisors.

Acknowledgments

I express my sincere thanks to **Professor Kishore Gawande** for giving me the opportunity to study this topic. He conveys a spirit of persistent in regard to research and an excitement in regard to teaching. This master report would not have been possible without his guidance.

I would also like to thank my supervisor, **Professor Tom Shively** for his kind guidance for this report. The report is greatly improved by lots of his wonderful advices.

And lastly, I would like to thank multitudes of people who helped me when I work with this master report, including my classmates, my advisers and all staffs and faculties in Department of Statistics & Data Science in University of Texas at Austin.

Does Trade Cause Inequality

Haoran Zhang, M.S.Stat
The University of Texas at Austin, 2015

Supervisor: Tom Shively

The relationship between international trade and income distribution of countries becomes a hot topic in economics research. This paper use random forest method and stepwise regression method to complete variables selection work from a big panel data set with many economic variables. Analysis of an unbalanced panel of country level data reveals that the trade will reduce income inequality in most situations. The coefficients for trade variables are significant in both two types of models, i.e., with and without considering about country effects. But when we split data set into two groups, the coefficients are significant for developed countries but not significant for developing countries.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In economics, the relationship between international trades and unequal distribution of income becomes a very hot topic recently. There are usually two problems in data analysis for this topic.

First one is that there are too many economic variables to represent and influence international trade and it is very difficult to choose variables without a strong economic background. Many researchers tend to choose variables by making several assumptions and then test different subsets of variables under specific assumptions. The second problem is that most of the economic variables are interacted with each other and the correlation between dependent variable and independent variables will violate assumptions of OLS regression analysis. For example, if we discuss trade and income distribution and we want to find how trade will affect income distribution, the income distribution will affect trading at the same time.

This paper mainly focuses on proposing some solutions for these two problems. To solve the problem of choosing appropriate variables from a large data set, we use Random Forest method and stepwise regression method to select important variables for trade and other economic data by ranking the importance of these variables. To avoid direct causality between dependent variables and independent variables in regression analysis, we use a method named Instrumental Variables to replace correlated independent variable with predicted value from another model. We also draw some conclusions about relationship between income inequality and international trade based on empirical results from our regression results, including consideration about country effects for the models.

Chapter 2

Related Work

There are lots of papers examining the impact of international trade on income inequality. Also, there are papers discussing variables selection and classification by random forest method and stepwise regression. Here we summarize some important papers which are directly relevant to this paper.

Frankel and Romer(1999) took an empirical investigation of the impact of international trade on standards of living. They considered a cross-country regression of income per capital on the ratio of exports or imports to GDP with data of 150 countries. The highlight is to use $Instrumental\ Variable$ (IV) techniques to correct the endogeneity of trade. In general, they show the impact of the trade on income across countries is statistically significant.

Aradhyula, Rahman and Seenivasan (2007) extended Farnkel and Romers conclusion by using a panel data to investigate the impact of trade on country levels. They also divided their data set into two groups - developed countries and developing countries and found the interpretation for coefficients would vary greatly for two groups. In the paper they use trade openness as the key independent variable and improve the result by using instrumental variable method.

Sandri, Valenzuela and Anderson (2006) led a development research group in World Bank and created a new big database containing many new indexes to describe trade and other economic activities of country. Trade reduction index (TRI) and Nominal rate of assistant (NRA) are emphasized in this database, which are the key source for a supplementary database of partial equilibrium indicators of trade and welfare reductions due to government interventions in agricultural markets.

U. Haque, Mark and J.Mathieson have firstly proposed to use stepwise regression method to choose variables from economic variables. They divided the economic variables into several groups by their attributes, then use stepwise regression to choose the variables and measure their importance.

Sandri and Zuccolottoe mentioned about random forest method in their paper *Variable Selection Using Random Forest*. They have compared model selection with stepwise selection to random forest in a specific case and find the later one has more advantages when dealing with large latter of data. Adriansson and Mattsson (2015) did more work to show how random forest method could improve accuracy of prediction in economics. They also compare random forest model with some time series predicted model forecasting GDP growth.

Chapter 3

Methodology

In order to explore which variables have a bigger impact on trade and income inequality and the reasons behind, we use a big panel data set about polity index, trade, income and other important variables for both developed and developing countries. Next, we will describe important variables used in our work, followed by the introduction of variable selection models and regression analysis models.

3.1 Data

We use panel data with variables of different classes and sources for 79 countries and areas over years of 1965-2006. Our original data set consists of 2306 observations and 164 variables.

Based on existing research, several factors impact or directly measure trade and inequality. Our data set contains different kinds of data from following sources. Our baseline data (OL) is the data set used by Liu and Ornelas (2014). This data set contains some basic economic data of countries such as GDP (per capital), war dummy variables, WTO dummy variables, geography and resource abundance variables. In addition, the writers pay much attention to describe the relationship between participation in free trade agreement ($FTAs$) and sustainability of democracy. Thus their data set has many variables related with this relationship.

Another data set SFI is from *the Global Report of Conflict, Governance and State Fragility (2014)* by Monty G.Marshall and Benjamin R.Cole. This data set has many measurements about countries polity performances and democracy index, such as State Fragility Index, Effectiveness Score, Legitimacy Score and many other indexes. In order to add more variables about the polity performance of the countries, we choose a new data set *p4v* from *POLITYTM IV PROJECT, Political Regime Characteristics and Transitions, 1800-2013.* And it has many variables to measure countries democracy and politic activity.

The most important source for data set is $World\ Bank$ (WB). In our data set, we collect data from $World\ Bank$ to describe countries' population growth, international import and export activities, energy use and production and so on. The quality of data is very good for complete time coverage and few missing values.

In order to measure income inequality, we choose $GINI$ index as dependent variables in our research. There are many sources to get $GINI$ index of countries and we collect it from $World\ Bank$ and the database of $United\ Nations\ University$, $UNU - Wider$. Since the later one uses the newest principle to recalculate $GINI$ previous index in history and it tends to have more effective data with a better data clean work, we use *Gini_Wider* from

$UNU - Wider$ as our dependent variable and $Gini\_WB$ from $World\ Bank$ as a reference variable in database.

At last, we will give explanations about $GINI$ index for income distribution with the reasons why we use this index to measure income inequality.

The Gini coefficient is a measured statistical index which indicates the income distribution of a country's residents, which has become the most commonly used measure of inequality in economic research.

The Gini coefficient measures the inequality described by a frequency distribution. A Gini coefficient with value zero represents extremely perfect equality, which assumes all people have totally same income. A Gini coefficient with value one (or 100%) represents maximal or worst inequality of income distribution (for example, only one "richest" person has occupied all the income, and all others have none).

Lorenz curve is usually used to describe Gini coefficient in a mathematical way. The curve plots the proportion of the total income of the population (y axis) that is cumulatively earned by the bottom x% of the population (in Figure 3.1). There is a very important standard line which is at 45 degrees standing for idealized equality. The Gini coefficient could be interpreted as the ratio of the area that lies between the line of equality and the Lorenz curve (marked A area in the figure) over the total area under the line of equality (the sum of marked A and B in the figure), which is given by,

$$Gini = \frac{A}{A+B}$$

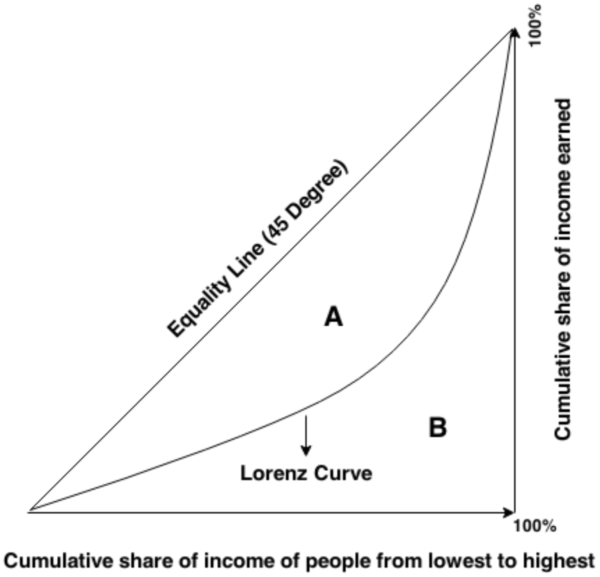

100%
Cumulative share of income earned
Equality Line (45 Degree)
A
B
Lorenz Curve
100%
Cumulative share of income of people from lowest to highest

Figure 3.1: GINI Coefficient Explanation.

If all people have non-negative income, the Gini coefficient can mathematically range from 0 (idealized equality) to 1 (complete inequality); it is also expressed as a percentage ranging between 0 and 100 sometimes (Such as value range for $Gini_Wider$ in our data set).

The Gini coefficient is widely proposed as a measure of inequality of income or wealth for countries. Here we choose 32 typical countries to show their Gini coefficients in year 2005. (See Figure 3.2 below) We find that the Gini coefficients for these 32 countries range from 23 (Sweden) to 56.4 (Brazil).

Most developed countries with good welfare have a Gini coefficient smaller than 40 except United States of America, who also suffers from a problem of huge gap between rich and poor.

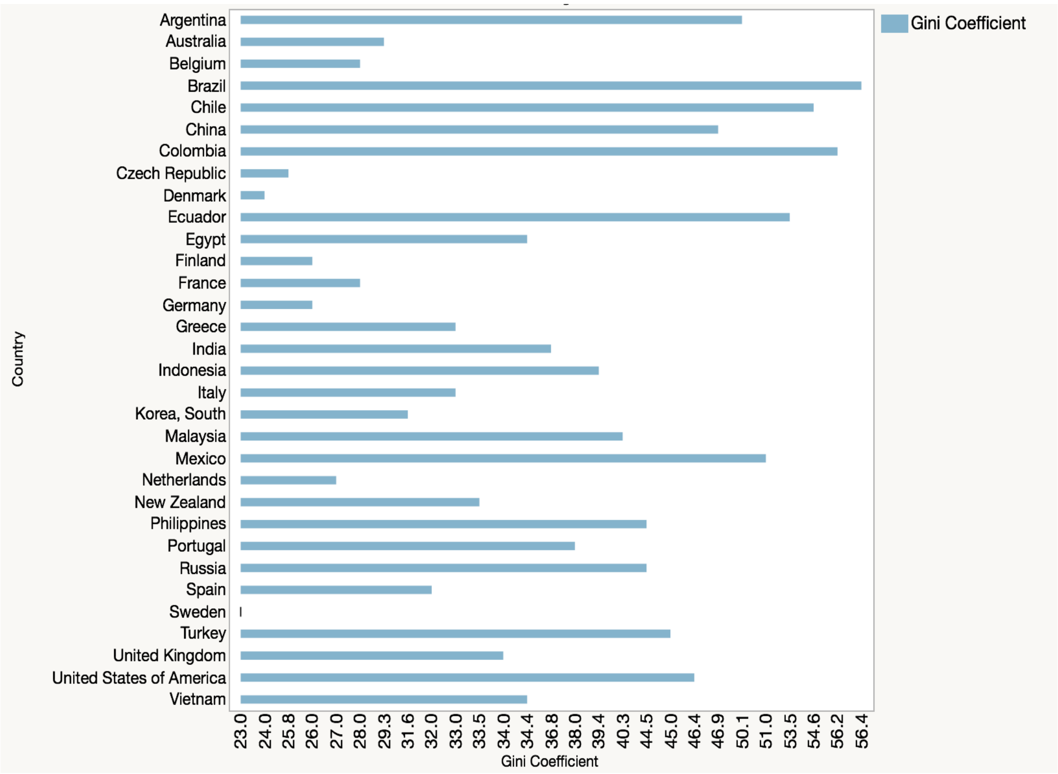


Figure 3.2: Gini Coefficient of 32 Countries in 2005.

3.2 Models and Methods

Our method for this paper follows three main steps: the first step is to use *Random Forest* method to select appropriate variables for regression analysis, then we rank the variables by the importance of variables offered by *Random Forest* measurement. This is relatively a "rough outline" process. The second step is to further perform variable selections by *Stepwise Regression* method with variables kept from last selection. This is a "delicate" process. The last step is to use regression analysis to see how the independent variables we choose from first two steps will affect our dependent variables by interpreting the coefficients in the result. Here we will use $Instrumental$ $Variable$ method to reduce the correlation effects between independent variable and the error term in the model.

3.2.1 The Random Forest Method

The Random Forest (RF) is a very effective algorithm for classifications, regression by using a multitude of decision trees. Leo Breiman and Adele Cutler develop this algorithm, and "Random Forests" is their trademark. Here we simplify some complex concepts about random forest and give a brief introduction about it.

RF adopts the core concept of decision trees which is a popular method for various machine-learning tasks. For example, if we have a relationship data set (or a learning set),

$$D = (X_i, Y_i), \qquad i = 1, 2, \ldots, n$$

All observations are *i.i.d.* from the random vector (X, Y). And we assume that the independent variables X_i may contain t predictors or explanatory variables, $X_i \in R^t$ and $Y_i \in R$ is the corresponding response. The regression process we usually used is to find a relationship of how to use X to predict Y,and we normally use a function like,

$$Y = f(X) + \epsilon$$

Here f is the function of regression containing the information about coefficients. And the main idea of random forest is to use many decision trees (here are binary regression trees), choosing randomly at each node the subset of explanatory variables X with several samples on D.

The RF uses lots of regression trees ($ntrees$, often hundreds) from different subsets of our independent variables. RF selects randomly for each trees and each node, and "each decision tree is built from a bootstrapped sample of full dataset " (Efron and Tibshirani, 1993). The difference for RF with regression is that while the traditional regressions choose all possible variables to evaluate the model, we only use a subset of all variables (fixed numbers of randomly selected variables) to calculate in node. Here we set the numbers of variables to choose as a constant with value $mtry$. Our final outcome for the prediction is an average value over $ntrees$. If we set individual predict for each tree is $r_1, r_2, \dots, r_n$, our final outcome is:

$$r(X) = \frac{1}{N} \sum_{n=1}^{N} h_n(X)$$

Randomization and averaging over trees enables random forest could approximate large classes of function while the errors are controlled relatively small. And it would also consider about interactions and non-liner effects in this process. Thus for our about 160 economical variables, it will consider more relationships among the variables than traditional regression method.

In recent years, there are several packages such as $randomForest$ and $Party$ in statistics software R free to implement random forest algorithm.

3.2.2 The Variable Importance Measure

As we know, Random Forest (RF) is often used for classification and regression. And during RF process, another very useful feature is that it could be used to reduce the data dimensionality and then select the variables.

The first step to measure the variable importance in a data set $D = (X_i, Y_i), i = 1, 2, \ldots, n$ is to fit a random forest to the data. During the fitting process the out-of-bag error here is averaged again and again.

The importance of the j^{th} variables after training is selected by their values, which are permuted among the training data and the out-of-bag error is again computed on this perturbed data set. The importance score for the j^{th} feature is an average score of the difference in out-of-bag error before and after the permutation, which is normalized by the standard deviation of these differences. Variables with large values of score are ranked as more important here.

The $randomForest$ package and $Party$ package in R could give out the importance of the variables. Party package could even give a conditional importance measure that consider about the correlations of variables.

3.2.3 Stepwise Regression

Stepwise regression includes regression models in which the choice of predictive variables is carried out by an automatic procedure. In fact, it could be regarded as a variable selection method where various combinations of variables are tested together. And *Stepwise regression* usually has two common methods to approach the object, *Forward Selection* Method and the *Backward Elimination* Method.

Forward Selection Method, which only includes constant at the beginning, add variables one by one and test them by using a model choosing comparison criterion (for example, see C_P criterion below). The model will keep the variables which improve the model the most, and repeating this process until none improves the model.

Backward Elimination Method, on the opposite, which involves all candidate variables at the beginning, testing by deleting each variable using a chosen model comparison criterion. The model will delete the variable that improves the model the most by being deleted, and repeating this process until no further improvement is possible.

A very important statistics in this process in determining the best model is the C_P criterion. The C_P values will decrease as the number of independent variables in the model increases. And C_P will calculate by the following way.

$$C_P = (N - P - 1)(\tfrac{RMS}{\hat{\sigma}^2}) + (P + 1)$$

Where N is the number of observations. P is the number of independent variables in the models. RMS is the residual mean square of model with P independent variables. $\hat{\sigma}^2$ is the residual mean square of the model with all possible independent variables included in.Thus the best model is to choose model by maximizing R^2, C_P or both.

Since we have choose several variables (about 20 to 30 variables) from *Random Forest* method. Here we choose to use *Backward Elimination* Method in *Stepwise regression* to check which variables are “best” for the object dependent variable.

Here is a simple decision plot (See Figure 3.3) about how we use *Stepwise regression* to drop variables step by step and then decide the model. (If the model has linear model, interactions and higher order terms.)

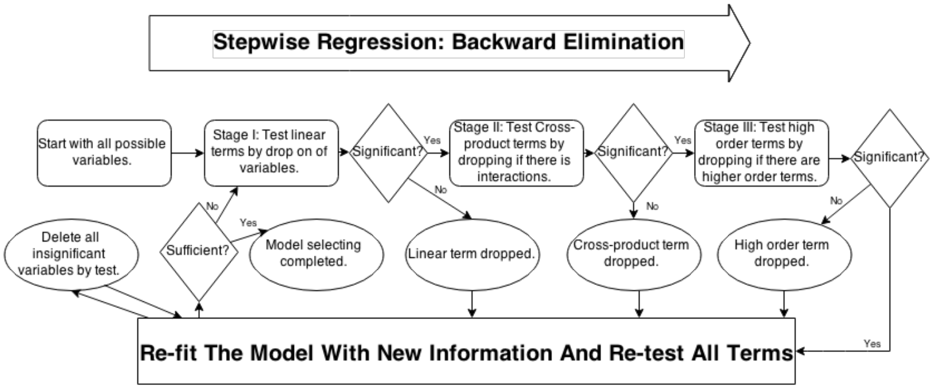

Stepwise Regression: Backward Elimination
Start with all possible variables.
Stage I: Test linear terms by drop on of variables.
Significant?
Yes
No
Stage II: Test Cross-product terms by dropping if there is interactions.
Significant?
Yes
No
Stage III: Test high order terms by dropping if there are higher order terms.
Significant?
No
Yes
Delete all insignificant variables by test.
Sufficient?
No
Yes
Model selecting completed.
Linear term dropped.
Cross-product term dropped.
High order term dropped.
Re-fit The Model With New Information And Re-test All Terms

Figure 3.3: Stepwise Regression Explanations

3.2.4 Instrumental Variables

In economy field, a common thing which will affect the accuracy of regression analysis is that the variables tend to have different kinds of correlations with each other and it causes problem especially when there are strong correlation between dependent variables and independent variables. For example, we assume our OLS regression model is:

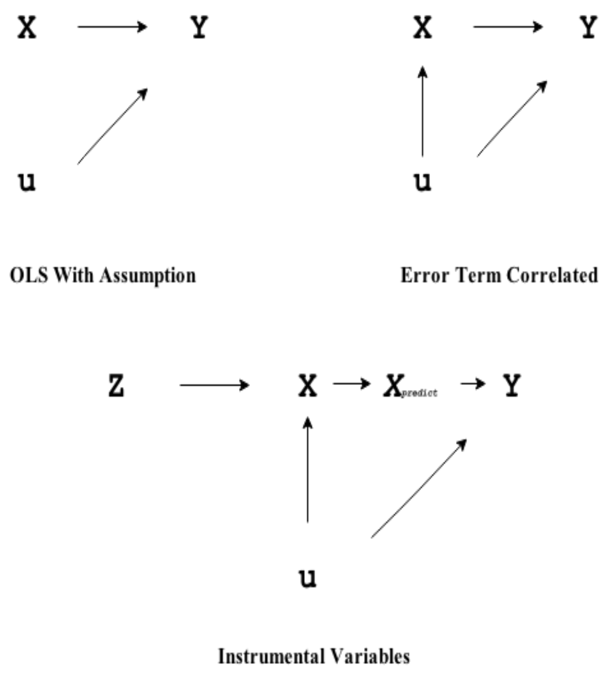
$$y = \alpha x + \mu$$

where μ is error term. We know that regression of y on x is by OLS estimate $\hat{\alpha}$ of α. In common situation, we have an assumption that our independent variables are uncorrelated with the error term, which means there is no association between x and α. But sometimes this assumptions doesn't hold, which leads the results by simple OLS are biased.

A widely used solution is to use $Instrumental\ Variables$ method to avoid the direct correlation between independent variable and error term. The method is to use a new variable which called instrument z which is strongly associated with the change of x but rarely associated with the error term between x and y.

Thus our regression divide to two stages as usual. The first stage is to use dependent variable x and independent variable z, this step need a relatively better fitting requirement because we need to use the predict value of x by z in the next stage. The next step is to use variable y as dependent variable and

x as independent variable, here we will use the estimate of x by the first stage instead of its original value. The following figure (in Figure 3.4) could explain this process easily.



OLS With Assumption

Error Term Correlated

Instrumental Variables

Figure 3.4: Instrumental Variables Explanations

Therefore, one important things in our two stages regressions is that we should keep other independent variables (except instruments) all the same. This method is widely used in econometrics but rarely used in other fields because it is conceptually difficult to interpret.

3.3 Regression Models

After the model selecting process, we will begin to analyze the relationship between inequality and trade by different regression tests. Firstly we will use linear regression test (OLS) to draw some beginning conclusions, then use $Instrumental\ Variable$ method (IV) to take a further research. At last, we will use a longitudinal model to check how this relationship varies across the countries.

3.3.1 Linear Regression Models

Our basic linear regression model is:

$$Gini_i = \alpha_0 + \alpha_{1j}(Trade_j)_i + \alpha_{2k}(Others_k)_i + \epsilon_i \tag{1}$$

In *Equation* (1), we use "*Gini_Wider*" as dependent variable and use some variables which have a direct relationship with international trade as our primary dependent variables. Other variables such as variables related to countries' government policy, economy and resource are regarded as "$Others$" in the equation.

Since "*Gini_WIder*" has the value from 0 to 100 and often varies little for a country in many years, we may also use a log transformation of dependent variables. Thus the model will be:

$$Log(Gini_i) = \alpha_0 + \alpha_{1j}(Trade_j)_i + \alpha_{2k}(Others_k)_i + \epsilon_i \tag{2}$$

In simple linear regression models, we will just pay attention to all of our observations in a cross sectional data format (regardless of countries and years, which may also affect much for the model). It's fine to draw some beginning conclusion by checking the signs of the coefficients. But a better conclusion should be made by the models with more considerations.

3.3.2 Model with Instrumental Variables

One issue for our primary independent variables here is that the international trade indeed affects the inequality, while the inequality will also may affect the international trade. This "endogeneity" make our estimate for regression model biased, especially when we draw some conclusions from coefficient α_{1j} and α_{2k} in *Equation* (1).

The solution here to use $Instrumental\ Variable$ (IV) method to avoid the "endogeneity" of trade. This method is firstly used by Frankel and Romer. Therefore, we need another regression model including some variables which could describe the trade very well but has less relationship with the error term (or correlation) between trade and income inequality. Then we would use the predicted value for trade of this model instead of value of trade in previous model. The model should be like,

$$Trade_i = \beta_0 + \beta_{1j}(Instruments_j)_i + \beta_{2k}(Others_k)_i + \mu_i \tag{3}$$

In *Equation* (3), $(Instruments_j)_i$ are the instrumental variables we prepare to use. And $(Others_k)_i$ in this equation should be exactly the same as $(Others_k)_i$ in *Equation* (1). Our assumption is that the error term ϵ_i in *Equation* (1) will also affect $(Trade_j)_i$ in the same equation, but it will not affect $(Instruments_j)_i$ in *Equation* (3). Thus this regression is our "Stage I" regression and we will use the estimate of the regression to continue "Stage II" regression by *Equation* (1).

3.3.3 Longitudinal Model with Country Effects

The models we mentioned before can discover some relationships between trade and income inequality, but we also ignore some very important effects, which is the effects of different countries to the model. The data set we use is actually a longitudinal (panel) data set and it is unbalanced (the observations of each country are not the same). But in our previous regression, we just use cross sectional data without utilizing the attributes of panel data. If we don't consider about countries effect, our regression will tend to only explain the situations of countries with more observations, thus making our result biased.

To solve this problem, we will use longitudinal model with instrumental variables to analyze by regressions. For we may have some missing values for some years of countries, we will reset and clean our data set by changing the time variable *year* to *period*. Here one period is determined by every 5 years, and period 1 is from year around 1965, period 2 is year around 1970, until period 9 is year around 2005. Some countries may not have the exact years (like 1965,1970 and 1975) we need, but we could find the year around these exact years instead of them. This method will also make our regression more reasonable, since many variables (such as Gini coefficient itself) do not vary much in a short period. One period for about 5 years could better describe the change of the variables. In this way, our model will change to:

$$Gini_{i,t} = \alpha_0 + \alpha_{1j}(Trade_j)_{i,t} + \alpha_{2k}(Others_k)_{i,t} + \epsilon_{i,t} \tag{4}$$

$$Trade_{i,t} = \beta_0 + \beta_{1j}(Instruments_j)_{i,t} + \beta_{2k}(Others_k)_{i,t} + \mu_{i,t} \tag{5}$$

Where i represents different country entities and t represents time variable *period*. If α_0 in *Equation* (4) is estimated directly as a fixed value, this model is a fixed effect model.

Chapter 4

Results and Conclusions

Our results will contain variables selection results by *Random Forest* method, the variables importance rank and key variables explanations, the variables selection results by *Stepwise Regression* method, regression analysis by using *Instrumental Variables* method and a further regression analysis by using longitudinal model. At last we will draw some conclusions from our regression results and make comparisons among different situations.

4.1 Variables Selection by Random Forest

The *Random Forest* process is to use income inequality as dependent variable ($Gini\_Wider$) and all possible features in our database as independent variables. We use both *R* package *randomForest* and package *party* to get the variable importance (*party* has a modified algorithm based on *radomForest*). Our beginning data set has 164 variables and 2306 observations.

Here we have 4 different tests, two for each package. The difference for tests in the same package is the parameter setting of *ntree* and *mtry*, where *ntree* is the depth of calculation in the algorithm and *mtry* is the numbers of variables which are randomly selected at every decision steps. As we have about 2000 observations, we set *ntree* as 50 for this amount of data and set *mtry* as 13 firstly (it is the square root of the numbers of variables, which is a widely used default setting). Then we will try another test by a bigger *ntree* and *mtry*.

After processing all data in *R*, we will output the importance of the variables rank the importance coefficients. The tables below shows top 29 important variables of 4 different test:

Table 4.1: Variable Importance Rank by Random Forest

Rank	**Package "randomForest", mtry=13, ntree=50**	**Package "randomForest", mtry=23, ntree=500**	**Package "party", mtry=13, ntree=50**	**Package "party", mtry=23, ntree=500**
1	Populationgrowth	remote	Populationgrowth	Populationgrowth
2	remote	Populationgrowth	lcontagion_gdpdist	ResourceAbundant
3	ResourceAbundant	Energyuse	ResourceAbundant	lcontagion_gdpdist
4	lcontagion_gdpdist	ResourceAbundant	Energyuse	Energyuse
5	GDP_per_capita	lcontagion_gdpdist	remote	remote
6	Energyuse	GDP_per_capita	llgdppc	gdppcp00
7	mynum_demospell	mynum_demospell	Export_volume_index	lgdppc
8	lgdppc	lgdppc	lFTA_impsh	mynum_demospell
9	llgdppc	GDP	Imports_GDPratio	ldc95_pt_cur
10	dc95_pt_cur	lcontagion_ldist	pop_rural	gdpdeflator
11	lcontagion_ldist	lcontagion_lgdpldist	gdpdeflator	GDP_per_capita
12	lcontagion_lgdpldist	Exports_GDPratio	nra_cov_o	dc95_pt_cur
13	lforeigncap	ldc95_pt_cur	ldc95_pt_cur	llgdppc
14	foreigncap	contagion_lgdpldist	dc95_pt_cur	pop_rural
15	impGDPwdi	llgdppc	gdppcp00	Imports_GDPratio
16	nra_cov_o	var10	lpolcomp	mynum_demospell_C
17	Exports_GDPratio	dc95_pt_cur	lcontagion_ldist	PTA_impsh
18	GDP	lforeigncap	FTA_impsh	lpolcomp
19	ldc95_pt_cur	pop_rural	mynum_demospell	IOnum_l1
20	Population	pop_agric	lgdppc	lforeigncap
21	mynum_demospell_C	mynum_demospell_C	lforeigncap	pop_agric
22	IOnum_l1	Population	GDP_per_capita	FTA_impsh
23	Imports_GDPratio	pop_agreconact	PTA_impsh	polity2
24	pop_agreconact	Imports_GDPratio	legit	agedem_new2
25	Landarea	Export_volume_index	limpGDPwdi	Export_volume_index
26	gdppcp00	foreigncap	lcontagion_lgdpldist	Exports_GDPratio
27	omexpsh	durable	lPTA_impsh	omexpsh
28	gdpdeflator	Landarea	agedem_new	foreigncap
29	contagion_lgdpldist	omexpsh	agedem_new2	nra_cov_o

This selection method, however, will be relatively "restricted" because many variables belong to the same class of data and they are correlated with each other. Therefore, we will combine all high ranked variables in our variable lists together in the first step, then try to divide these variables into several different parts (classes or groups). Finally, we will rank the variables for each class by weighted ranks of 4 different tests. The following table is the result after reorganizing previous ranks.

Table 4.2: Important Variables Classification

Trade	**Economy**	**Polity**	**Population**	**Resource**
Imports_GDPratio	impGDPwdi	polity2	Populationgrowth	Landarea
Import_volume_index	GDP_per_capita	legit	pop rural	ResourceAbundant
Exports_GDPratio	gdppcp00	alpha		Energyuse
Export_volume_index	lgdppc			Energyproduction
nra_cov_o	alpha			remote
nra_covt				
PTA_impsh				
FTA_impsh				

In variable selection above, we omit some variables because they are essentially high correlated with the variables in our table. For a further inspection of how this variables will influence the income inequality, we will use stepwise regression and linear regression to test them.

4.2 Variables Selection by Stepwise Regression

Before we proceed to *Stepwise Regression* process, we will show a data description of what we choose from *Random Forest* method in the following table ($Valid$ in table means the ratio of valid observations to all).

Table 4.3: Variables Description

Variable Name	**Variable Definition**	**Valid**	**Mean**	**Std. D.**	**MIN**	**MAX**
Import_volume_index	Import volume index (2000 = 100)	0.39	95.27	54.84	6.09	482.72
Export_volume_index	Export volume index (2000 = 100)	0.39	91.74	52.09	7.19	739.26
Imports_GDPratio	Imports of goods and services (% of GDP)	0.81	36.25	26.17	0.00	208.98
Exports_GDPratio	Exports of goods and services (% of GDP)	0.81	29.07	17.38	3.22	121.31
alpha	Estimated weight on social welfare	0.42	8.30	7.86	0.16	37.81
remote	Remoteness of a country to other countries	0.95	2.17	0.04	2.10	2.25
Landarea	Land area (sq. km)	0.88	1325761	2842059	320	1.64e+07
Energyuse	Energy use (kg of oil equivalent per capital)	0.75	2555.00	1985.74	91.08	13690.22
nra_covt	Nominal Rates of Assistance for all_covered products	0.58	0.31	0.56	-0.69	4.05
nra_cov_o	Nominal Rates of Assistance to output of countries	0.58	0.23	0.56	-0.69	4.03
GDP_per_capita	GDP per capital (current US$)	0.83	7527.88	11062.18	56.63	88395.68
Energyproduction	Energy production (kt of oil equivalent)	0.75	107155.40	274729.90	0.00	1688631.00
polity2	Combined Polity Score	0.90	4.30	6.92	-10	10
ResourceAbundant	Resource abundance dummy	0.70	0.41	0.49	0	1
PTA_impsh	Import share form PTA partners	0.90	0.05	0.10	0	0.61
FTA_impsh	Import share form FTA partners	0.90	0.25	0.31	0	0.91
Populationgrowth	Population growth (annual %)	0.90	1.22	1.18	-5.81	6.68
pop_rural	Rural population by FAOSTAT	0.59	4.59e+07	1.45e+08	21000	8.40e+08

4.2.1 Data Clean Work For Further Regression

From data description, we find that some variables have too many missing values (the valid data are less than 50% of all observations). This missing situation is OK for previous variables selection by *Random Forest* process because it could deal with missing value by not considering about them in decision steps. But if we use regressions method of many independent variables at the same time, only the observations with no missing value will be considered, which wastes a large proportion of data.

Furthermore, many variables may have duplicate information choosing by *Random Forest*. For example, many variables about GDP are chosen out from *RF* method, but most of them are just the transformations of GDP. Thus, it needs us to select only few of these kind of variables which could stand for all others.

In addition, our data set is not balanced for countries and years. If some of countries have much more observations than other countries, these countries tend to dominate the results of regression. To remedy this drawback, we need to narrow down the numbers of observations for one country.

The situations all above need us to use a more complete and clean data set to apply further regression analysis. Fortunately, we have already dropped most useless variables by RF method so that we only need to clean data with about 20 remaining variables. Our data clean process contains following steps in the order:

1. First, we drop missing values from the perspective of observations. An observation will be dropped if more than 50% values missing within it.

2. Then we look through the updated dataset from step 1 in perspective of variables and drop variables with more than 50% missing values.

3. We find other sources to replace some missing values. For example, we use the definition of resource abundance in $Wikipedia$ to supplement some dummy variables of $ResourceAbundant$ for some countries.

4. For some missing values, if we could find another value which belongs to and with close year attribute (usually less than 2 years) with it, we will replace the missing with this value.

5. For fixed countries, we regress observations on time if the time variable $year$ is continuous for the country. Then we use estimate value by regression to replace the missing values.

6. We use 0 to replace missing value when it is very hard to find any other sources or use other methods above to supplement missing value if it is easy to interpret this variable by simply replacing with 0.

7. At last, we only keep the observations with value for variable *period* we create in the data set. We have mentioned the reason why we create *period* in previous chapter and this method helps us to control observation numbers for the countries with too many observations.

After the cleaning process, we have a new data set without any missing data and our observations reduce to 433 with 73 countries. For further study of variable selections, we will run a stepwise regression based on all variables we select from random forest.

4.2.2 Stepwise Regression for Gini Coefficient

The stepwise regression begins by using Gini coefficient as dependent variables first. We transfer Gini coefficient to log form and make a comparison between the results. We use the *Backward Elimination* Method to drop the variables and set our criterion at 5% significance level. Table 4.4 shows the results of the stepwise regression.

Table 4.4: Stepwise Regression for Gini Coefficient

	(1)	(2)
VARIABLES	Gini_Wider	Log_Gini
ResourceAbundant	4.744***	0.110***
	(1.000)	(0.0259)
Imports_GDPratio	-0.196***	-0.00474***
	(0.0614)	(0.00159)
Exports_GDPratio	0.185***	0.00444***
	(0.0578)	(0.00149)
Populationgrowth	3.257***	0.0779***
	(0.448)	(0.0116)
polity2	0.246***	0.00642***
	(0.0816)	(0.00211)
nra_covt	-20.17**	-0.593***
	(7.967)	(0.206)
nra_cov_o	20.36**	0.606***
	(8.107)	(0.210)
pop_rural	-2.89e-08***	-8.06e-10***
	(3.66e-09)	(9.47e-11)
Energyproduction	1.11e-05***	3.29e-07***
	(1.97e-06)	(5.10e-08)
Energyuse	-0.00189***	-5.63e-05***
	(0.000370)	(9.55e-06)
GDP_per_capita	-0.000120**	-2.91e-06**
	(4.91e-05)	(1.27e-06)
Constant	37.86***	3.622***
	(1.552)	(0.0401)
Observations	433	433
R-squared	0.473	0.471

Standard errors in parentheses

PTA_impsh, Landarea, FTA_impsh, remote are dropped

*** p<0.01, ** p<0.05, * p<0.1

From the output, we find that four variables are dropped by stepwise regression but most of variables are significant in the test. This relatively indicates the advantage of *RF* method to select variables at a good significant level. The log transformation for dependent variable contributes little to the result so we will not use it in later analysis.

R-squared here is 0.47 though the significant test is good in results. It indicates that it is very hard for us to apply model fit to Gini coefficient by these chosen variables but it is still very good for us to investigate the relationships between them.

The most interesting thing in the output is that, we find the signs for coefficients of *Imports_GDPratio* and *Exports_GDPratio* are totally opposite. This is because there are some relationships between this two variables, and coefficients are affected by the correlation when two variables are included in the model at the same time. Similar thing happens to *nra_covt* and *nra_cov_o*. We will focus on this difference and try to explain more in our following analyses.

4.2.3 Stepwise Regression of Trade Variables

From the results above, we find that there are 4 variables that can represent primary trade variable in Equation (3) we talked about in Chapter 3. Once we have chosen one as dependent variables in Equation (3), others could be instrumental variables as independent variables in this equation.

In this way, we need to inspect how these trade variables will affect each others and select right variables for Equation (3). We take turns to use these 4 variable as candidate dependent variable and then test them by stepwise regression separately. Table 4.5 gives out the result of these tests.

Table 4.5: Stepwise Regression for Trade Variables

	(1)	(2)	(3)	(4)
VARIABLES	Imports_GDPratio	Exports_GDPratio	nra_covt	nra_cov_o
Exports_goods andservices	0.855***			
	(0.0186)			
Populationgrowth	-0.863***		-0.0710***	-0.0716***
	(0.330)		(0.0264)	(0.0259)
nra_covt	-1.589***			
	(0.597)			
pop_rural	-5.87e-09**			
	(2.84e-09)			
Energyproduction	-3.05e-06**		-4.18e-07***	-4.22e-07***
	(1.50e-06)		(1.00e-07)	(9.85e-08)
Energyuse	-0.000756***	0.000746***	5.00e-05**	5.04e-05**
	(0.000231)	(0.000219)	(2.00e-05)	(1.97e-05)
Imports_goods andservices		0.958***	-0.00313**	-0.00322**
		(0.0194)	(0.00154)	(0.00151)
GDP_per_capita		8.80e-05**	1.35e-05***	1.35e-05***
		(3.91e-05)	(2.92e-06)	(2.87e-06)
ResourceAbundant			-0.208***	-0.194***
			(0.0587)	(0.0576)
Constant	9.375***	-2.529***	0.300***	0.294***
	(1.034)	(0.736)	(0.0834)	(0.0819)
Observations	433	433	433	433
R-squared	0.862	0.861	0.308	0.311

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

One important thing should be mentioned here is that we only use variable *nra_covt* if variables *nra_covt* and *nra_cov_o* both exist in models since these two variables are highly correlated with each other. (If one is significant in the model, we are highly confident another one will be significant.) From four different stepwise regression tests, we find that two variables about import and export of countries tend to be more suitable as dependent variable in Equation (3). Two variables about *nra* are not very good here because R-square for their stepwise regression is not good (only about 0.3). The model with Equation (3) needs a big R square regression to ensure a convincing goodness of fit so that we could use the estimate for the second stage model.

Besides, we find that using *Imports_GDPratio* as dependent variable could contain more meaningful independent variables than using with *Exports_GDPratio*. Thus we tend to pay more attentions to the model with *Imports_GDPratio* in later analyses.

4.3 Regressions with Instrumental Variables

In this section, we will give several comparisons between regressions using instrumental variables method and not using this method. We will use *Imports_GDPratio* or *Exports_GDPratio* as endogenous variables separately.

We will test many combinations of variable groups by using the variables we selected before. What we expect is that the variables in both tests of Equation (3) (the first stage) and Equation (1) (the second stage) are significant. The following results are just part of our many tests, with as many independent variables significant as possible. We use generalized method of moments (GMM) for coefficient estimates of regression with *IV* method.

4.3.1 One Instrumental Variable Situations

Table 4.6 shows the regression with *Imports_GDPratio* as endogenous variable. Here the instrumental variable is *Exports_GDPratio.* Table 4.7 gives the results of reverse situation.

Table 4.6: Imports_GDPratio with One Instrumental Variable

	Dependent variable: Gini_Wider	
	(1)	(2)
VARIABLES	OLS Estimate	IV GMM Estimate
Imports_GDPratio	-0.114***	-0.0828***
	(0.0274)	(0.0287)
pop_rural	-2.20e-08***	-2.11e-08***
	(3.58e-09)	(3.89e-09)
Energyuse	-0.00203***	-0.00204***
	(0.000245)	(0.000265)
nra_covt	-3.297***	-3.314***
	(0.834)	(0.734)
Constant	48.54***	47.59***
	(1.095)	(1.147)
Observations	433	433
R-squared	0.278	0.275

First Stage Statistics: F(4, 428) = 570.86, R-square = 0.8582

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

From two tables, we find that the coefficients of trade variables are all negative in 4 different tests. And the coefficients of trade change differently for *import* and *export* after we use *IV* method to improve our model. In *import* part, the effects of trade to Gini coefficient become smaller by using

Table 4.7: Exports_GDPratio with One Instrumental Variable

Dependent variable: Gini_Wider		
	(1)	(2)
VARIABLES	OLS Estimate	IV GMM Estimate
Exports_GDPratio	-0.0725***	-0.118***
	(0.0263)	(0.0290)
pop_rural	-2.04e-08***	-2.15e-08***
	(3.58e-09)	(3.96e-09)
Energyuse	-0.00198***	-0.00191***
	(0.000251)	(0.000262)
nra_covt	-3.231***	-3.151***
	(0.845)	(0.729)
Constant	46.97***	48.16***
	(0.998)	(1.080)
Observations	433	433
R-squared	0.261	0.256

First Stage Statistics: F(4, 428) = 435.81, R-square = 0.8613

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

IV method, we could find that income inequality decreases by 8.28% for every 1% increase in countries *import* activities about goods and services, which means an increase in import of *import* will protect income inequality from increasing. In *export* part, the effects of trade to Gini coefficient become bigger by using *IV* method, with income inequality decreasing by 11.8% for every 1% increase in countries *export* activities about goods and services.

In addition, we could find that *nra* index is negative here and statistically significant at 1% level. One unit increase in *nra* index will lead to about three units decrease in Gini coefficient, which means a bigger nra index will more or less decrease income inequality. As we know, a bigger nra index means the country will protect more about their domestic industry of commodities. Countries usually tend to protect common commodities of domestic industry than luxury ones, so we could regard it as an egalitarian policy by the country.

4.3.2 Two Instrumental Variable Situations

As we know, if we choose either one of trade variable with import or trade variable with export as dependent variable in Equation (3), another one will be the instrumental variable. Table 4.5 indicates that nra index will also be significant as instrumental variable. Here we will inspect how the situation will change if we use two instrumental variables for $(Instruments_j)_i$ in Equation (3) of the model.

Table 4.8 and table 4.9 show us the regression output with two instrumental varables for both *import* and *export* variables.

From the output we find that there is no big change to the results of coefficients with just one instrumental variables. For nra index has been one part of instrumental variables in the model, we could not interpret the coefficient of it. The variables of rural population and total energy consumption are still significant in the model. The coefficient for rural population is very small so we may ignore it. And the coefficient of energy consumption of countries indicates that more energy use will reduce income distribution inequality. The explanation for this is maybe the energy needs of countries will lead to the

Table 4.8: Imports_GDPratio with Two Instrumental Variables

	Dependent variable: Gini_Wider	
	(1)	(2)
VARIABLES	OLS Estimate	IV GMM Estimate
Imports_GDPratio	-0.116***	-0.0651**
	(0.0278)	(0.0294)
pop_rural	-2.15e-08***	-2.04e-08***
	(3.64e-09)	(3.82e-09)
Energyuse	-0.00241***	-0.00260***
	(0.000229)	(0.000273)
Constant	48.70***	47.03***
	(1.113)	(1.161)
Observations	433	433
R-squared	0.251	0.242

First Stage Statistics: F(4, 428) = 570.86, R-square = 0.8582
Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

desire for more trade activities, or big energy use itself means large production of commodities for trade.

4.4 Regressions with Fixed Effects of Countries

We have drawn some conclusions from the results above, but we lose some power when we use these regressions. All regressions above only considered about the value of dependent variables and independent variables without considering about country effects in this model. As our data set is a panel data set essentially, if some countries have more observations with certain linear

Table 4.9: Exports_GDPratio with Two Instrumental Variables

	Dependent variable: Gini_Wider	
	(1)	(2)
VARIABLES	OLS Estimate	IV GMM Estimate
Exports_GDPratio	-0.0780***	-0.119***
	(0.0267)	(0.0301)
pop_rural	-2.00e-08***	-2.12e-08***
	(3.64e-09)	(3.95e-09)
Energyuse	-0.00235***	-0.00239***
	(0.000235)	(0.000269)
Constant	47.22***	48.09***
	(1.012)	(1.109)
Observations	433	433
R-squared	0.236	0.229

First Stage Statistics: F(4, 428) = 435.81, R-square = 0.8613
Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

patterns, these observation will dominate our regression and lead to biased results.

Therefore we will use a longitudinal regression by Equation (4) and Equation (5) with consideration of country effects for the whole model. When we use a new model, we need to change variables setting for many different times to find a "best" model, except for fixing our trade variables setting as previous models. Here we use fixed effect model for our longitudinal analysis to assume that the effects of one country for the regression remain the same.

Table 4.10 and table 4.11 below gives the results of instrumental variables regression with country fixed effects.

Table 4.10: Analysis of Import with Country Effects

	Dependent variable: Gini_Wider			
	(1)	(2)	(3)	(4)
VARIABLES	OLS Estimate1	IV Estimate1	OLS Estimate2	IV Estimate2
Imports_GDPratio	-0.0455	-0.0385	-0.0863**	-0.115***
	(0.0292)	(0.0256)	(0.0375)	(0.0432)
remote	110.3***	111.3***	246.3***	245.6***
	(12.86)	(14.22)	(78.00)	(78.06)
Energyproduction	7.03e-07	7.74e-07	1.83e-05***	1.93e-05***
	(1.78e-06)	(1.24e-06)	(5.53e-06)	(5.58e-06)
nra_covt	-2.719***	-2.697***	1.529	1.528
	(0.854)	(0.765)	(1.107)	(1.108)
pop_rural	-1.91e-08***	-1.90e-08***	-2.54e-08	-2.50e-08
	(3.79e-09)	(3.39e-09)	(1.64e-08)	(1.64e-08)
Constant	-198.4***	-200.9***	-495.2***	-493.0***
	(28.35)	(30.97)	(169.7)	(169.9)
Observations	433	433	433	433
R-squared	0.286	0.285	0.073	
country FE	NO	NO	YES	YES
IV method	NO	YES	NO	YES
Number of countries			73	73

First Stage Statistics without FE: F(5, 427) = 549.31, R-square = 0.8603
First Stage Statistics with FE: F(5, 355) = 235.39, R-square = 0.7870
Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

From the output of two tables, we could find that *import* and *export* variables are significant in fixed effects test combined with variable *remote* and *Energyproduction*, and *nra* index will not be significant in the model, which is different from the situations without country fixed effects. The sign for the coefficients of trade variables are still negative in tests, which means more international trade will more or less reduce income inequality in some degree. And the effects level is that income inequality decreases by about 10% for every 1% increase in countries trade activities. The conclusion is highly consistent with previous results.

Table 4.11: Analysis of Export with Country Effects

	Dependent variable: Gini_Wider			
	(1)	(2)	(3)	(4)
VARIABLES	OLS Estimate1	IV Estimate1	OLS Estimate2	IV Estimate2
Exports_GDPratio	-0.0324	-0.0462*	-0.0907***	-0.0905**
	(0.0269)	(0.0257)	(0.0340)	(0.0392)
remote	112.2***	110.3***	256.8***	256.7***
	(12.77)	(14.14)	(77.86)	(77.88)
Energyproduction	9.88e-07	9.14e-07	1.90e-05***	1.90e-05***
	(1.77e-06)	(1.22e-06)	(5.54e-06)	(5.60e-06)
nra_covt	-2.592***	-2.602***	1.419	1.419
	(0.850)	(0.752)	(1.105)	(1.105)
pop_rural	-1.89e-08***	-1.92e-08***	-2.50e-08	-2.50e-08
	(3.79e-09)	(3.41e-09)	(1.64e-08)	(1.64e-08)
Constant	-203.1***	-198.4***	-518.1***	-518.1***
	(28.08)	(30.75)	(169.4)	(169.4)
Observations	433	433	433	433
R-squared	0.284	0.283	0.077	
country FE	NO	NO	YES	YES
IV method	NO	YES	NO	YES
Number of countries			73	73

First Stage Statistics without FE: F(5, 427) = 549.31, R-square = 0.8603
First Stage Statistics with FE: F(5, 355) = 239.27, R-square = 0.7132
Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

In addition, the variable *remote* becomes significant at 1% level when considering about country effects since the value for *remote* usually remains the same for one country and this variable is easily dropped in regression before if we don't consider about country effects. The coefficient for *remote* indicates that the geographic disadvantage of a country will lead to a bad income distribution. The geographic position of a country sometimes reveal its opportunity of trade and often affect countries' trade openness.

At last, we will try to investigate the relationship between trade and inequality for developed countries and developing countries separately. Table 4.12 and table 4.13 give out the results. For simplicity, we only show the results of *import* variable and *export* variable has very similar results.

Table 4.12: Analysis of Import with Country Effects for Developed Countries

	Dependent variable: Gini_Wider			
	(1)	(2)	(3)	(4)
VARIABLES	OLS Estimate1	IV Estimate1	OLS Estimate2	IV Estimate2
Imports_GDPratio	-0.0626	-0.0708**	-0.170***	-0.220***
	(0.0389)	(0.0312)	(0.0649)	(0.0730)
remote	56.62***	55.22***	238.3**	261.1***
	(14.12)	(18.27)	(94.92)	(96.27)
Energyproduction	1.88e-06	1.93e-06	4.04e-06	5.26e-06
	(2.52e-06)	(2.05e-06)	(8.21e-06)	(8.26e-06)
nra_covt	-0.166	-0.180	1.829	1.761
	(0.811)	(0.693)	(1.202)	(1.205)
pop_rural	-1.01e-08	-1.47e-08	1.18e-07	7.97e-08
	(5.88e-08)	(5.74e-08)	(2.69e-07)	(2.70e-07)
Constant	-86.14***	-82.80**	-476.7**	-524.0**
	(31.18)	(39.39)	(203.6)	(206.3)
Observations	225	225	225	225
R-squared	0.165	0.165	0.064	
country FE	NO	NO	YES	YES
IV method	NO	YES	NO	YES
Number of countries			35	35

First Stage Statistics without FE: F(5, 219) = 559.74, R-square = 0.9328
First Stage Statistics with FE: F(5, 185) = 176.90, R-square = 0.7431
Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Two tables show us the results that our previous conclusions are almost unchanged for developed countries. The effects of trade to income inequality of 35 developed countries even seem to be bigger than the average effects of all 73 countries. *remote* is always significant at a good level whether we consider about country effects or not. But *Energyproduction* will not be significant in tests. This is maybe because most developed countries tend to produce energy with higher technology, less pollution and fewer labors, which means a less effect on market, working employment structure and trade.

At the same time, the results for developing countries are very interest-

Table 4.13: Analysis of Import with Country Effects for Developing Countries

Dependent variable: Gini_Wider

VARIABLES	(1) OLS Estimate1	(2) IV Estimate1	(3) OLS Estimate2	(4) IV Estimate2
Imports_GDPratio	-0.107***	-0.0864**	-0.00943	-0.0138
	(0.0407)	(0.0343)	(0.0515)	(0.0619)
remote	141.9***	142.7***	483.5***	479.2***
	(26.58)	(24.35)	(152.7)	(156.5)
Energyproduction	8.97e-06**	8.87e-06**	2.94e-05***	2.95e-05***
	(3.64e-06)	(3.97e-06)	(7.53e-06)	(7.60e-06)
nra_covt	-3.453	-3.472	-2.569	-2.531
	(2.793)	(2.617)	(2.643)	(2.660)
pop_rural	-3.36e-08***	-3.30e-08***	-2.54e-08	-2.55e-08
	(4.87e-09)	(4.08e-09)	(1.83e-08)	(1.84e-08)
Constant	-262.8***	-265.2***	-1,018***	-1,008***
	(58.39)	(53.36)	(336.0)	(344.6)
Observations	208	208	208	208
R-squared	0.291	0.290	0.146	
country FE	NO	NO	YES	YES
IV method	NO	YES	NO	YES
Number of countries			38	38

First Stage Statistics without FE: F(5, 202) = 154.47, R-square = 0.7989
First Stage Statistics with FE: F(5, 165) = 101.01, R-square = 0.6471
Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

ing. The coefficients of trade of income inequality become not significant when considering about fixed country effects even though they are still significant with no consideration of fixed country effects. It indicates that many studies just use cross sectional data to draw a conclusion between trade and inequality is suspicious here.

4.5 Possible Further Improvements

There exists some possibilities for further improvement of our analysis.

The first thing is to use more countries for analysis. For we use a fixed effects model for country effects, we may not include some random effects of the countries we don't use.

The second possible improvement is to clean the data set with some different method to get a balanced panel data set. It should be better interpreted with balanced panel data.

Another thing could be considered is to add time effects to the model. As we know, for most countries, the Gini coefficients don't change much in recent 10 years while it varies a lot around year 1980. Maybe we could introduce some dummy variables about year to the model or consider year effects within countries, which may lead to more interesting conclusions.

The last thing we may improve is to consider how to predict Gini coefficient by trade and other variables. For most of the studies, R-square for second stage regression is far from good for constructing a predict model for Gini coefficient and we may achieve the predicting work by some alternative method.

Appendices

Appendix 1

Countries Included in Data Set

No.	Developed Countries	Developing Countries
1	Argentina	Bangladesh
2	Australia	Benin
3	Austria	Brazil
4	Belgium	Bulgaria
5	Canada	Cameroon
6	Chile	China
7	Cyprus	Colombia
8	Czech Republic	Cote D Ivoire
9	Denmark	Dominican Republic
10	Estonia	Ecuador
11	Finland	Egypt
12	France	Ethiopia
13	Germany	Ghana
14	Greece	India
15	Hungary	Indonesia
16	Iceland	Kazakhstan
17	Ireland	Kenya
18	Israel	Lithuania
19	Italy	Malaysia
20	Japan	Mexico
21	Korea, South	Morocco
22	Latvia	Mozambique
23	Malta	Nicaragua
24	Netherlands	Nigeria
25	New Zealand	Pakistan
26	Norway	Philippines
27	Poland	Romania
28	Portugal	Russia

29	Slovakia	Senegal
30	Slovenia	South Africa
31	Spain	Sri Lanka
32	Sweden	Tanzania
33	Switzerland	Thailand
34	United Kingdom	Turkey
35	United States of America	Ukraine
36		Vietnam
37		Zambia
38		Zimbabwe
Total	**35**	**38**

Appendix 2

Sample of Longitudinal Data Set

country	**year**	**period**	**Gini_Wider**	**...**	**nra_covt**	**nra_cov_o**	**...**	**pop_rural**	**Populationgrowth**	**Resource**	**...**
Argentina	1965	1	36	...	-0.2637031	-0.2662039	...	5266000	1.490681143	1	...
Argentina	1970	2	40.9	...	-0.2178071	-0.2092732	...	5061000	1.537528592	1	...
Argentina	1975	3	34.70000076	...	-0.4148859	-0.4207015	...	4957000	1.639382372	1	...
Argentina	1980	4	40.5	...	-0.0586246	-0.0535293	...	4808000	1.500248116	1	...
Argentina	1985	5	39.79999924	...	-0.21641	-0.2347707	...	4534000	1.507176912	1	...
Argentina	1990	6	44.4	...	-0.267808	-0.2664403	...	4234000	1.402801086	1	...
...	...	...	...	...	...	...	...	...	...	...	...
China	1990	6	29.39999962	...	-0.335186	-0.3373787	...	838785000	1.467303211	1	...
China	1995	7	28.971	...	0.0334617	0.0313356	...	837098000	1.086509151	1	...
China	2000	8	39.028	...	0.0176127	0.0117018	...	818974000	0.787956593	1	...
China	2004	9	46.9	...	0.0136095	0.0089088	...	794634000	0.593932815	1	...
Colombia	1964	1	57.20000076	...	-0.0167407	0.0046368	...	9139000	2.954724116	1	...
Colombia	1970	2	55.2	...	-0.1426976	-0.1143815	...	9794000	2.608170314	1	...
...	...	...	...	...	...	...	...	...	...	...	...
France	1995	7	30.2	...	0.4151784	0.4151785	...	14596000	0.361612779	0	...
France	2000	8	28.2	...	0.3284382	0.3284382	...	14399000	0.684623197	0	...
France	2005	9	28	...	0.1591581	0.1591581	...	11251000	0.753310125	0	...
Germany	1964	1	23.89999962	...	0.9944111	0.9944111	...	16878000	0.805140883	0	...
Germany	1970	2	20.4	...	0.9750646	0.9750646	...	15923000	0.332661393	0	...
Germany	1975	3	36.59999847	...	0.5284097	0.5284097	...	14813000	-0.372846359	0	...
...	...	...	...	...	...	...	...	...	...	...	...
Korea, South	1998	8	36.9	...	1.274514	1.274514	...	9557000	0.721865018	0	...
Korea, South	2004	9	31.6	...	2.264316	2.264316	...	9440000	0.375613356	0	...
Latvia	1992	6	33.3	...	-0.4603626	-0.4603782	...	800000	-1.376795066	0	...
Latvia	1995	7	28.5	...	0.0120464	-0.0095242	...	784000	-1.425810812	0	...
Latvia	2000	8	33.7	...	0.3121814	0.2337998	...	787000	-0.963935407	0	...
Latvia	2005	9	36	...	0.1412498	0.1412498	...	738000	-1.080571457	0	...
...	...	...	...	...	...	...	...	...	...	...	...
Philippines	1994	7	42.9	...	0.2435255	0.2435255	...	31433000	2.2795756	0	...
Philippines	2000	8	49.441	...	0.3886384	0.3886384	...	31384000	2.126511513	0	...
Philippines	2003	9	44.53	...	0.1545934	0.1545934	...	31184000	2.018955625	0	...
Poland	1992	6	23.9648	...	-0.0385695	-0.0650757	...	14945000	0.306681391	0	...
Poland	1995	7	33	...	0.0808306	0.0753812	...	14938000	0.135721034	0	...
Poland	2000	8	34.17555	...	0.1079459	0.1021384	...	14826000	-1.044335398	0	...
...	...	...	...	...	...	...	...	...	...	...	...
Zambia	1975	4	59	...	-0.628291	-0.628291	...	3302000	3.375293059	1	...
Zambia	1991	6	48.4	...	-0.6898849	-0.6898849	...	5138000	2.43949672	1	...
Zambia	1996	7	52.40000153	...	-0.4559333	-0.4559333	...	6073000	2.589902446	1	...
Zambia	2003	8	42.08	...	-0.1896852	-0.1896852	...	6985000	2.501029295	1	...
Zimbabwe	1990	6	56.6	...	-0.4031563	-0.4031563	...	7432000	2.854307075	1	...
Zimbabwe	1995	7	73.3	...	-0.2679456	-0.2679456	...	8014000	1.826900354	1	...

Bibliography

[1] Nils Adriansson and Ingrid Mattsson. Forecasting gdp growth, or how can random forests improve predictions in economics? 2015.

[2] Paul D Allison. Measures of inequality. *American sociological review*, pages 865–880, 1978.

[3] Satheesh Aradhyula, Tauhidur Rahman, and Kumaran Seenivasan. Impact of international trade on income and income inequality. In *American Agricultural Economics Association Annual Meeting, Portland July 29-August*, volume 1, page 2007, 2007.

[4] Avik Chakrabarti. Does trade cause inequality? *Journal of Economic Development*, 25(2):1–22, 2000.

[5] Tom S Clark and Drew A Linzer. Should i use fixed or random effects. *Unpublished paper*, 25, 2012.

[6] Cletus C Coughlin. Measuring international trade policy: a primer on trade restrictiveness indices. *Federal Reserve Bank of St. Louis Review*, 92(September/October 2010), 2010.

[7] Klaus Deininger and Lyn Squire. A new data set measuring income inequality. *The World Bank Economic Review*, 10(3):565–591, 1996.

[8] Jeffrey A Frankel and David Romer. Does trade cause growth? *American economic review*, pages 379–399, 1999.

[9] Ajit Kumar Ghose. *Global economic inequality and international trade.* International Labour Office, Employment Sector, 2001.

[10] Nadeem Ul Haque, Donald J Mathieson, and Nelson C Mark. *The relative importance of political and economic variables in creditworthiness ratings.* International Monetary Fund, 1998.

[11] Ronald R Hocking. A biometrics invited paper. the analysis and selection of variables in linear regression. *Biometrics*, pages 1–49, 1976.

[12] Max Roser and Jesus Crespo Cuaresma. Why is income inequality increasing in the developed world? *Review of Income and Wealth*, 2014.

[13] Marco Sandri and Paola Zuccolotto. Variable selection using random forests. In *Data analysis, classification and the forward search*, pages 263–270. Springer, 2006.

Vita

Haoran Zhang was born in Hubei, China on 10 May 1991. He received the Bachelor of Science degree in Statistics from Huazhong University of Science and Technology in Wuhan, China in June 2013.

After graduating from undergraduate school, he applied to the University of Texas at Austin for enrollment in their statistics program. He was accepted and started graduate studies in August, 2013.

Permanent address: 1300 Crossing Place, APT822
Austin, Texas 78741

This report was typeset with LaTeX[†] by the author.

[†] LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.