Stockholms universitet

An analysis of the relationship between income inequality and indicators of social and infra-structural development

Elin Magnusson

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2013:11
Matematisk statistik
Oktober 2013

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Mathematical Statistics
Stockholm University
Bachelor Thesis **2013:11**
http://www.math.su.se

An analysis of the relationship between income inequality and indicators of social and infrastructural development

Elin Magnusson*

Oktober 2013

Abstract

In this paper a multiple regression analysis will be undertaken to examine which social and infrastructural indicators best explain and predict income inequality. To quantify income inequality, the Gini coeffcient is used. The analysis is made on two separate data sets with different ways of measuring income. In the first data set the income from countries worldwide is calculated on the basis of household per capita, while in the second set it is calculated for OECD countries on a modified scale. The two data sets reach between the years 1990 and 2006. The explanatory variables are basic indicators for development such as age and population demographics, school enrolment and infrastructural factors. In the regression analysis it is shown that models for explaining income inequality can be found but that exact predictions cannot be made. Variables used in both final models include *Life expectancy, Urban population* and *Infant mortality rate*. Additionally, age demographic variables are used in both models but the demographic variable used differs between the two.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail:elin.maria.magnusson@gmail.com . Supervisor: Gudrun Brattström.

Preface and acknowledgements

This is a thesis of 15 ECTS credits in Mathematical Statistics at the Department of Mathematics at Stockholm University. I would like to thank my supervisor Gudrun Brattström for help and support throughout the work on this paper.

Disclaimer

This paper is not in any way a collaboration with or acknowledged by the World Bank or the the UNU-WIDER.

Contents

1 Introduction

This analysis will look at income inequality and what social and infrastructural factors there are to explain and predict such inequality. There are several ways of defining and measuring inequality, in this paper the inequality in income is the object of analysis and for that the Gini coeffient (a.k.a Gini index) is a good measurement. Data has been collected internationally and the index is not dependent on specifics such as the form of government in a country, it's national legal system or adherence to international law. The basis for the Gini coefficient is the Lorenz curve (see Appendix B for further calc.), which is used for several measurements of dispersion in data − economic as well as social and biological. The Gini coefficient may in some cases refer to the dispersion of wealth though in this paper only income inequality is used to calculate the Gini coefficient. We want the explanatory variables used in the analysis to have the same characteristics as the Gini coefficient, so that the variables are measured in the same way in each country. E.g. if the meaning of a variable is dependent on the laws in each country, the variable is not comparable between countries, only within each country.

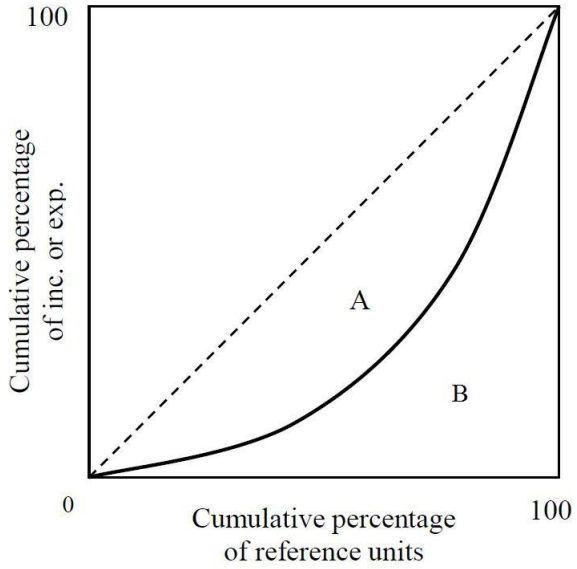
1.1 Background

The Gini coefficient created by Corrado Gini published in 1912, is a measure of the diversity in data and lies between 0 (perfect equality) and 1 (perfect inequality) [1]. It is based on the Lorenz curve, as seen in Figure 1 if the Lorenz curve coincides with the diagonal the data is perfectly equal. In this paper it is the income equality that is measured which theoretically means that if the Gini coefficient is 0 everyone has the same income and if it is 1 one person has all the income in the measured data. There are of course several difficulties surrounding the measuring of equality, for instance the same Gini can come from numerous distributions. The Gini coefficient in this analysis is measured as an index of $0 - 100$ instead of $0 - 1$. The basic calculations of the Gini coefficient is:

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

A is the area between the Lorenz curve and the diagonal and B is the area below the Lorenz curve. For calculation of Lorenz curve see Appendix B.

Figure 1: The Lorenz curve [1]



1.2 Description of data

The data for the Gini coefficient in the following analysis comes from WIID2 [2], the UNU-WIDER World Income Inequality Database, which includes historic data which uses a number of different ways of measuring income/consumption distribution data. The data set in total includes 5314 observations between the years 1867-2006 with 36 variables. The difficulties in comparing income distributions between countries are numerous, and include the fact that there are several methods used to collect information about income that gives different coefficients. In developing countries with a large agricultural sector it is often hard to get accurate data, and in these countries the inequality distribution is often based on consumption instead of income. The 36 variables include the geographic coverage of the surveys underlying the observations, the unit of analysis and the equivalence scale, the income concept, the income share unit and the quality of data.

In Table 1 we see an extract from WIID2, we see 15 of the 36 variables in the data set and 21 of the 5314 observations. So in between *Rep. Gini* and *PopCovr* lies 21 variables that are not used in this paper except for the variable *AreaCovr*. There are two different Gini coefficients reported in WIID2 as seen in Table 1, the first is calculated by WIDER and the second "Rep. GINI" is either the one reported by the source or the Gini coefficient given in the old data base WIID1. In this report the Gini coefficient used is

the one calculated by WIDER, not the "Reported GINI".

Table 1: Extract from WIID2

Country	Year	Gini	Rep. Gini	.....	Pop Covr	Age Covr	IncSharU	Uof Anala	Equivsc	IncDefn	Source1	Survey/ Source2	Qual.
Canada	1993	33,6	33,57	.....	All	All	Census Fam.	Person	Census fam. eq, sqrt	Income, Disp.	Frenette...	Tax data	2
Canada	1994	33,9	33,85	.....	All	All	Census Fam.	Person	Census fam. eq, sqrt	Income, Disp.	Frenette...	Tax data	2
Canada	1995	34,3	34,28	.....	All	All	Census Fam.	Person	Census fam. eq, sqrt	Income, Disp.	Frenette...	Tax data	2
Canada	1996	34,9	34,90	.....	All	All	Census Fam.	Person	Census fam. eq, sqrt	Income, Disp.	Frenette...	Tax data	2
Canada	1997	35,2	35,18	.....	All	All	Census Fam.	Person	Census fam. eq, sqrt	Income, Disp.	Frenette...	Tax data	2
Canada	1998	35,5	35,48	.....	All	All	Census Fam.	Person	Census fam. eq, sqrt	Income, Disp.	Frenette...	Tax data	2
Canada	1999	35,9	35,85	.....	All	All	Census Fam.	Person	Census fam. eq, sqrt	Income, Disp.	Frenette...	Tax data	2
Canada	2000	36,5	36,53	.....	All	All	Census Fam.	Person	Census fam. eq, sqrt	Income, Disp.	Frenette...	Tax data	2
Canada	1990	28,1	28,06	.....	All	All	Eco. Fam.	Person	Eco. fam. eq, sqrt	Income, Disp.	Frenette...	Survey...	1
Canada	1991	28,7	28,73	.....	All	All	Eco. Fam.	Person	Eco. fam. eq, sqrt	Income, Disp.	Frenette...	Survey...	1
Canada	1992	28,3	28,32	.....	All	All	Eco. Fam.	Person	Eco. fam. eq, sqrt	Income, Disp.	Frenette...	Survey...	1
Canada	1993	28,6	28,58	.....	All	All	Eco. Fam.	Person	Eco. fam. eq, sqrt	Income, Disp.	Frenette...	Survey...	1
Canada	1994	28,3	28,34	.....	All	All	Eco. Fam.	Person	Eco. fam. eq, sqrt	Income, Disp.	Frenette...	Survey...	1
Canada	1995	28,8	28,78	.....	All	All	Eco. Fam.	Person	Eco. fam. eq, sqrt	Income, Disp.	Frenette...	Survey...	1
Canada	1996	29,1	29,14	.....	All	All	Eco. Fam.	Person	Eco. fam. eq, sqrt	Income, Disp.	Frenette...	Survey...	1
Canada	1996	29,6	29,62	.....	All	All	Eco. Fam.	Person	Eco. fam. eq, sqrt	Income, Disp.	Frenette...	Survey...	1
Slovenia	2005	24,0	24,00	.....	All	All	Household	Person	Household eq, OECDmod	Income, Disp.	European...	The Eur...	1
Slovenia	2006	24,0	24,00	.....	All	All	Household	Person	Household eq, OECDmod	Income, Disp.	European...	The Eur..	1
Sweden	2004	23,0	23,00	.....	All	All	Household	Person	Household eq, OECDmod	Income, Disp.	Eur...	.	2
Sweden	2005	23,0	23,00	.....	All	All	Household	Person	Household eq, OECDmod	Income, Disp.	European...	The Eur...	1
Sweden	2006	23,0	24,00	.....	All	All	Household	Person	Household eq, OECDmod	Income, Disp.	European...	The Eur...	1
.	.	.	.	.....	.	.	.	.	.	.	.	.	.
.	.	.	.	.....	.	.	.	.	.	.	.	.	.
.	.	.	.	.....	.	.	.	.	.	.	.	.	.

To be able to compare the Gini coefficient between countries the definitions of measuring income have to be the same in each country. The Canberra Group on Household Income Statistics was a group active between 1996-2000, working for the United Nations Statistic Division and was assembled to amplify the national household income statistics and created guidelines to enhance comparability on income distribution[3]. The group met four times and had representatives from a large group of countries and multiple groups like the Luxembourg Income Study Group at the Centre for Population, United Nations Statistics Division, the World Bank and the Economic Commission for Europe. A final report written by the Canberra Group gives recommendations on which factors surveys should take into account when data is collected and which data is most comparable [7].

The Canberra Group states that the basic statistical unit should be the household i.e. calculating the total income of households instead of for example only looking at personal income. There are several problems with calculating the personal income e.g. children in most countries have no income so an age limit probably should be set and in extension it is not possible to calculate how many persons actually share the income. The income or

consumption should be adjusted to take account of household size using per capita income or consumption. This means that we adjust the income with respect to how many people that is provided by it, or equivalently, how many people that consumes in the household unit. The Canberra group also states that personal weights is preferred for analysis, the weight is given to how many the income represent. For example, if a household has the probability 1 in 500 of being selected in a survey the household has a weight of 500, to calculate the person weights each household unit is multiplied by the number of persons in the specific unit. So the personal weights create a over all distribution of income for individuals, assuming that the household incomes are pooled. Further down we will see that household size is not an absolute measurement. The label "disposable income" is given to the observation if it corresponds to the one described by the Canberra Group, simply the income that after loans, taxes etc. have been paid that can be used for expenditures and savings. In this paper the data that is used follows the recommendations of the Canberra Group when possible.

All data for the explanatory variables come from the World Bank [4], a database of world development indicators. The data catalogue provides over 1200 indicators in a large range of areas and the chosen variables have been collected to be used as basic indicators of development and for mapping simple social factors, such as age demographics. Data for the variables need to be collected and counted in the same way in each country, for example, data for domestic violence is law based and since the laws are different in each country the variable is not comparable between countries. From the data catalogue 41 variables have been chosen. Of these 39 are explanatory factors and the other two are *country* and *year*, which is used to merge the data with the variables from WIID2. In Table 2 we see an extract of the data set downloaded from the World Bank, we see 21 observations of 1426 and the first 6 explanatory variables.

Table 2: Extract from data set of indicators from the World Bank

Country Name	Year	Agri. land (%)	Alt. and nuc. energy (%)	Adj. net enroll. rate, primary, fem. (%)	Adj. net enroll. rate, primary (%)	Pupil/teacher , primary	Fem./ male sec. enroll. (%)	...
Armenia	1990		1,74					...
Austria	1990	42,45	10,98			10,86	91,49	...
Belarus	1990		0				104,27	...
Belgium	1990		23,11				101,07	...
Bolivia	1990	32,73	3,89					...
Botswana	1990	45,91	0,04	89,13	85,64	31,66	111,1	...
Bulgaria	1990	55,67	13,95				99,72	...
Canada	1990	7,45	21,54	95,59	95,23	15,69	100,86	...
Chile	1990	21,38	5,48				105,49	...
China	1990	54,23	1,25		97,04	22,32	73,28	...
Croatia	1990		3,64					...
Cyprus	1990	17,53	0	78,79	78,83	21,38	102,95	...
Czech Republic	1990		6,82			24,49	91,45	...
Denmark	1990	65,77	0,34	97,58	97,5	11,28	102,12	...
Ecuador	1990	28,34	7,12			30,41		...
El Salvador	1990	68,05	20,33					...
Estonia	1990		0					...
Finland	1990	7,87	20,94				118,03	...
France	1990	55,82	38,71		99,91		106,87	...
Gabon	1990	20,01	5,13					...
Georgia	1990		5,25					...
.	.	.	.	.	.	.	.	...
.	.	.	.	.	.	.	.	...
.	.	.	.	.	.	.	.	...
.	.	.	.	.	.	.	.	...
.	.	.	.	.	.	.	.	...

First the WIID2 is sorted and reduced by these criteria which follows the recommendation from the Canberra Group. We keep only the observations where the definitions and coverage of the Gini coefficient is:

- Unit of Analysis= Person
- Income share unit= Household
- Income definition= Disposable income
- Area, Population & Age coverage= All
- Year= 1990-2006

We will now divide data into two separate data sets depending on the equivalence scale. Since the income surveys have been collected on a household basis it is important to scale the income depending on the size of the household, in WIID2 there are mainly four different ways of scaling seen in Table 3. So the equivalence scales are used to calculate the economic "number of persons" in the household unit. As seen in Table 1 other scales are used like the size of a family, the problem is that family, or equivalent, is a vague expression and differs between countries. This is why the Canberra group recommends one of the equivalence scales in the table below. We can see in Table 1 that the observations for "Canada" does not correspond to the definitions above so these are among the variable deleted from the data set.

Table 3: Equivalence scales

Equivalence scale	**Calc. for unit size**
Household per capita	Number of persons in household
Square root	$\sqrt{\text{Number of persons in Household}}$
OECD scale	$1 + 0.7\cdot$ n (add. adults)$+ 0.5\cdot$ n(add. children)
OECD scale modified	$1 + 0.5\cdot$ n (add. adults)$+ 0.3\cdot$ n(add. children)

The data is now divided into two different data sets. First, one with *equivalence scale* "Household eq. OECDmod" from the survey made by the European Commission 2005-2008. Containing only the European Community Household Panel Survey and The European Union Statistics on Income and Living Conditions (EU-SILC) survey, this new data set has 160 observations. We see that in Table 1 that the observation "Sweden, 2004" has the right definitions and coverage to be included in the data set with *equivalence scale* "Household eq. OECDmod", but it lacks the underlying *Survey/Source2* and is therefore excluded from the data set. The second data set contains the observations derived from the reduced WIID2 with the *equivalence scale* "Household per capita", and this new data set has 297 observations. These two data sets now have comparable observations within each set since the corresponding observations in each data set are collected from comparable surveys and the way of calculating the Gini coefficient is the same.

For these two data sets all variables except for *Gini*, *Year* and *Country* is deleted. The data sets are now merged with the data set with the explanatory variables by *Year* and *Country*. All observations lacking a responding Gini coefficient is then deleted. So we now have two data sets one with 160 observations and the other one with 297 and both with *Year* and *Country* as reference variables, *Gini* as responding variable and 39 explanatory variables.

From now on the two data sets will be referred to as GINI-HPC (Household per capita) and GINI-OECD (Household eq. OECDmod) after their corresponding equivalence scale. A table of which countries and which years represented in each data set can be found in Table 20. GINI-HPC have observations with quality $1 - 3$ and GINI-OECD contains only observations with quality 1. The quality of the observations in WIID2 are somewhat based on the Canberra Group's recommendations, and are divided into three parts:

1) *Whether the concepts underlying the observations are known or not*

This might be seen as a given, although this is not always the case. Concepts like disposable, monetary and gross income are not absolute concepts, so the concepts differ and it is not always known what is meant. Especially in older surveys it is not always clear what concepts underlying the observations are.

2) *The coverage of the income/consumption concept*

This adheres closely to the Canberra Group's recommendation, though for example monetary income it has been accepted for most developing countries since home production and in-kind income have little effect on income distribution.

3) *The survey quality*

This point has been divided into three basic points:

- *Coverage issues* — Most important is if the survey coverage is known.
- *Questionnaires* — Need to have enough information on income and expenditures.
- *Data collection methodology* — For example on consumption data either diaries must be used or frequent visits must occur.

These requirements have been checked for every observation/survey and the quality is then set on the basis of the extent to which the data reaches the requirements. The quality can be seen more as a guideline, if an observation has low quality it can still be a usable indicator of inequality in that country. A good example is the observation "Sweden, 2004" in Table 1 that reaches the requirements except that the *Survey/Source2* is missing which lowers the quality.

2 Method

Our main goal is to find a linear model describing the relationship between the Gini coefficient and the explanatory variables, which we do using regression analysis. Another part of the analysis concerns how well the Gini coefficient can be predicted through the chosen models, especially how well it can be predicted with the data from GINI-OECD.

2.1 Regression analysis

Linear regression models are models used to show if a variable, the response, depends or at least is explained linearly by one or several other, explanatory, variables [5].

2.1.1 Simple linear regression

The model for simple linear regression is defined as

$$y_i = \alpha + \beta x_i + \epsilon_i$$

where x_i is the explanatory variable for $i = 1...n$, α and β is parameters and ϵ_i is the errors. β and α is estimated with least square.

2.1.2 Multiple regression

The model used for the multiple regression is

$$\mathbf{y} = \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\epsilon}$$

Where $\mathbf{y}$, $\boldsymbol{\beta}$, $\mathbf{X}$ and $\boldsymbol{\epsilon}$ are matrices and vectors for the observed values, the parameters and the errors.

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{pmatrix} \mathbf{X} = \begin{pmatrix} 1 & x_{11} & ... & x_{1p} \\ 1 & x_{21} & ... & x_{2p} \\ . & . & ... & . \\ . & . & ... & . \\ . & . & ... & . \\ 1 & x_{n1} & ... & x_{np} \end{pmatrix} \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_p \end{pmatrix} \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ . \\ . \\ . \\ \epsilon_n \end{pmatrix}$$

for *p-1* variables and n observations, the parameters $\boldsymbol{\beta}$ are estimated with ordinary least squares.

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

2.2 Definitions

The following are some important definitions used in the analysis.

2.2.1 $\mathbf{R}^2$

The coefficient of determination, R^2 [5], is the most commonly used measurement of adjustment in connection with linear models and in particular with multiple regression models. The coefficient of determination is defined as the share of the total variation that the model explains and goes from 0 to 1.

$$R^2 = \frac{SS_{model}}{SS_{total}} = 1 - \frac{SS_{error}}{SS_{total}}$$

Where the sum of squares are defined as

$$SS_{model} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$SS_{error} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$SS_{total} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

For
$\bar{y} = Mean\ of\ response\ variable$
$\hat{y}_i = ith\ predicted\ response$
$y_i = ith\ observed\ response$

2.2.2 PRESS statistic

PRESS stands for Predicted Residual Sum of Squares and is a statistic used to measure the fit of a model and can be used to compare models fitted on the same data set. For the fitted model a form of cross-validation is used were in turn each of the observations used is eliminated where as the model is re-fitted using the remaining observations, so the model is re-fitted n times where n is the number of observations. A statistic is then computed using the residuals and the leverage of the observation used as below [6].

$$\textbf{PRESS} = \sum_{i=1}^{n} (\frac{r_i}{1 - h_i})^2$$

for r_i=residual for ith observation h_i=leverage of ith observation

Where the lowest PRESS indicates the best fit of the models, an over parametrised model tends to give a higher PRESS in comparison with R-square which always gets higher with more variables. This PRESS is the one that SAS calculates and therefore the one used in the analysis.

2.2.3 Stepwise regression

Stepwise regression is a procedure where a regression model is built by starting with zero explanatory variables and adding variables in each step that have a p-value below a chosen α and in each step removing the variables with a p-value above a chosen α , continuing until a model with only significant explanatory variables are left in the model. Although p-value is the most common determining value other statistics can be used for the procedure.

3 Statistical Analysis

We begin by looking at the data sets and each variable, since some of the variables have missing values. The first step is to see how many missing values each variable has in the two data sets. When we perform simple linear regression in the program SAS, which is used for the analysis, as a bonus we get the number of missing values for each variable. We can then look at the Gini coefficient plotted against each variable and the corresponding residuals. When we look at the plots we can also examine if any of the variables need to be transformed. We begin by excluding every variable with more than 20% (237 for GINI-HPC and 128 for GINI-OECD) missing values, this as a first crude reduction of variables.

When the variables have been reduced there are 24 explanatory variables left for GINI-HPC and 28 for GINI-OECD. Later in the analysis there might be more variables that are excluded depending on where the missing values lie since when multiple regression is performed all observations with one missing value or more are eliminated from the regression. This means if all missing values are under the same countries and years then at least 80% of the observations are used but if the missing values lie within different observations there might be very few observations used in the analysis.

We now look at the plots of the Gini coefficient on each of the explanatory variables and the corresponding residuals. We see that in GINI-HPC the variables *Population size*, *Net national income*, *GNI (Gross National Income) per capita* and *GDP (Gross Domestic Product) per Capita* all have both observations and residuals very concentrated in a small area. We take the logarithm of the variables and as can be seen in Figure 5 the residuals now look randomized. In GINI-OECD we take the logarithm of the same variables as in GINI-HPC and as can be seen in Figure 6 also for this data set the residuals are not as concentrated. In Table 21 we see the complete list of explanatory variables, with the transformed variables in the bottom, and which are used for GINI-HPC and for GINI-OECD. In Table 4 we see all the variables used with abbreviations that will be used during the analysis.

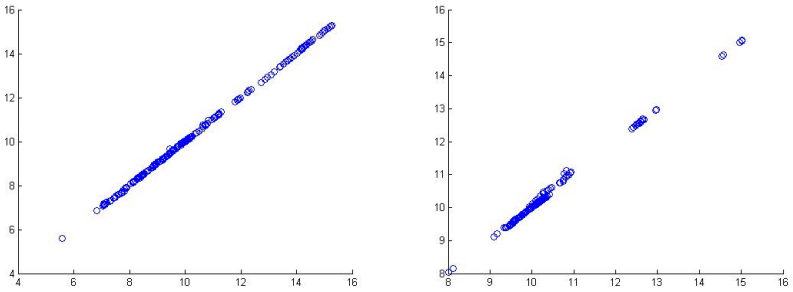
Table 4: Explanatory variable abbreviations

Variable	Abbreviation
Agricultural land (% of land area)	Agri.land
Alternative and nuclear energy (% of total energy use)	Alt& nuc engi
Ratio of female to male secondary enrolment (%)	Fem/male sec.enroll
Death rate, crude (per 1,000 people)	Death rate
Fertility rate, total (births per woman)	Fert.rate
Life expectancy at birth, total (years)	Life exp.
Mortality rate, infant (per 1,000 live births)	Mort. rate
Population ages 0-14 (% of total)	Pop. 0-14
Population ages 15-64 (% of total)	Pop 15-64
Population ages 65 and above (% of total)	Pop 65+
Internet users (per 100 people)	Int. users
Employment to population ratio, 15+, female (%)	Emp/pop fem
Employment to population ratio, 15+, male (%)	Emp/pop male
GDP per person employed (constant 1990 PPP $)	GDP.p.p.emp.
Labour force with secondary education, female (% of female labour force)	Lab.w.sec edu fem
Labour force with secondary education, male (% of male labour force)	Lab.w.sec edu male
Labour force with primary education (% of total)	Lab.w.prim edu
Labour force, female (% of total labour force)	Lab.fem % tot.
Unemployment, total (% of total labour force)	Unemp. tot.
Female legislators, senior officials and managers (% of total)	Fem. leg.
Electric power consumption (kWh per capita)	Elec.pow.con.
Population density (people per sq. km of land area)	Pop.dens
Urban population (% of total)	Urb.pop
Armed forces personnel (% of total labour force)	Armd.personnel
LOG(Population, total)	Log.pop
LOG(Adjusted net national income (current US$))	Log.adj.net.inc
LOG(GNI per capita (constant LCU))	Log.GNI
LOG(GDP per capita (constant LCU))	Log.GDP

As can be hinted by the residuals *GNI per capita* and *GDP per capita* look very correlated, in Figure 2 we see that they are extremely correlated. Since *GNI per capita* have less observations in both data sets it is eliminated from the analysis, only *GDP per capita* is used in the analysis from now on.

Figure 2: Scatterplot for *GNI per capita* on *GDP per capita*

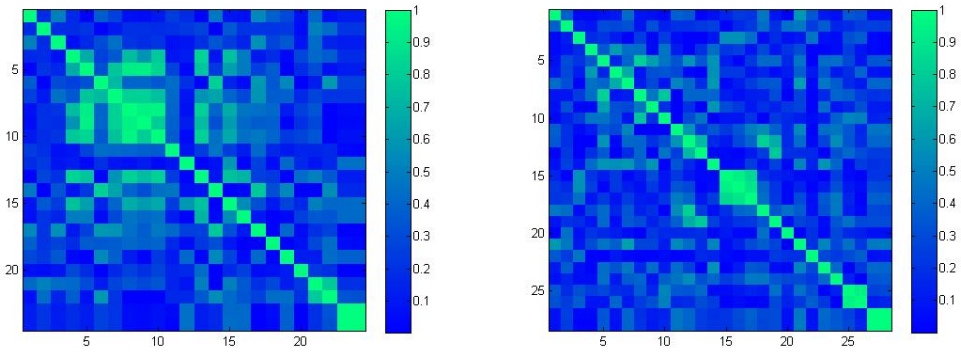
(a) GINI-HPC (b) GINI-OECD



If we continue to look at the relationship between the variables, in Figure 3 we see the absolute value of the correlation coefficients between the variables, in the same order, as listed in Table 21.

Figure 3: Correlation between variables

(a) GINI-HPC (b) GINI-OECD



We can see in scatterplots between each variable as well as in the figures above that some variables are very correlated, if we take a look at Table 21 this is to be expected. The cluster of green in the middle of the plot are the variables that are age-related so some correlation between them was to be expected.

In the continuation of the analysis we will look at each of the data sets individually. Since there are a lot of variables it is hard to grasp different dimensions of the data set. To get a better overview and to see if the variables might give us the same information multiple times we perform regression analysis on some of the explanatory variables and then create groups based on the models. Basically, if one or several of the other explanatory variables can explain any of the other variables, we call these intervening variables. If a good fitted model can be created using some of the other explanatory variables in the data set, the program we use might not create the best model possible for the Gini coefficient, since some of the information is already included. This is another reason to create groupings of the explanatory variables so that the groups consist of variables that are not dependent of each other. The last reason to create the intervening models is that when we get a final model for our Gini coefficient we then know more about the explanatory variables in that model.

3.1 GINI-HPC

When we concentrate on the data set GINI-HPC we can see that the correlations in this data set are more prominent than in GINI-OECD. This can be related to several different things, in the analysis we will further explore the relationship between the variables. If we look closer at the scatter plots between the variables we see that they correspond well to the Figure 3A. The variables most correlated to each other in some way are the variables explaining age demographic and the variables infant mortality, fertility rate, crude death rate, employment to population ratio (male) and GDP per person employed.

Even if we know that some of the variables are correlated to each other, this does not necessarily mean that there is a *cause and effect* relationship between the variables, though it hints that some variables can be explained by other variables in the data set and are therefore intervening. This is how we will try to do the partition, that is to say by examining if we can create data sets with fewer explanatory variables by showing that the excluded variables can be explained by other variables within the same data set. When looking closer at the variables we also exclude the variables *Armd.personnel*, *Agri.land*, *Int. users* and *Unemp. tot.* due to lack of observations. In the data set it is now 10 observations having one or more missing values.

After analysing several explanatory variables we do a multiple linear regression with the response variable *crude death rate* and see that we can create a model shown in Table 5 with all P-values for the slopes being < 0.0001. The model also has normally distributed and randomized residuals. As we

can see in Table 5 *the crude death rate* is well explained by the age demographics, fertility rate and mortality rate.

Table 5: Parameter estimates Death rate

Variable	Parameter Estimate	Pr $< \|t\|$
Intercept	113.96957	<.0001
Fert.rate	-0.62514	<.0001
Mort rate	-0.04630	<.0001
Pop 0-14	-0.66786	<.0001
Pop 15-64	-0.65189	<.0001
R-square= 0.9658		

We create a first group (Group 1) without the intervening variable *crude death rate* and try to reduce this group even more. Since all of the explanatory variables in Table 5 are needed to explain the death rate we want to keep these but examine if one or several of these four variables can be explained by the remaining variables in Table 21. So we do models using stepwise regression for each of the four variables as response and the remaining variables as explanatory.

We find that for the variables *Fertility rate* and *Infant Mortality rate* we cannot find any good models, but for the variables *Population ages 0-14 (% of total)* and *Population ages 15-64 (% of total)* there exist models with good fit. As we see in Table 6 and 7 the models both include the same explanatory variables but for the *Population ages 0-14 (% of total)* a model with the variable *Fert. rate* is used. The coefficients for these models have no real value for us in this part of the analysis, but it is worth noticing that the coefficients for the two variables used in both models switches signs between the two intervening models.

Table 6: Parameter estimates Pop 0-14

Variable	Parameter Estimate	Pr $< \|t\|$
Intercept	25.91195	<.0001
Fert rate	4.84593	<.0001
Pop 65+	-0.96276	<.0001
Elec.pow.con.	-0.00016697	<.0001
R-square= 0.9740		

Table 7: Parameter estimates Pop 15-64

Variable	Parameter Estimate	Pr $< \|t\|$
Intercept	73.43597	<.0001
Fert. rate	-4.69670	<.0001
Elec.pow.con.	0.00015520	<.0001
R-square= 0.9039		

A look at the remaining variables in the data set that are not yet explained in any of the models above reveals that the only variable that can be explained well by the other remaining variables are *Labour force, female (% of total labour force)*. In Table 8 we can see the model created with good goodness of fit and with normally distributed residuals.

Table 8: Parameter estimates Lab.fem % tot.

Variable	Parameter Estimate	Pr $< \|t\|$
Intercept	52.26360	<.0001
Emp/pop. fem	0.47868	<.0001
Emp/pop. male	-0.46581	<.0001
Elec.pow.con.	-0.00009448	<0.0001
R-square= 0.9419		

Since we now have a lot more information about the data set and the dependence between the variables we can use this information to "puzzle" groups together so that the groups we create consist of as few variables holding the same information as possible. In each of the groups one or several intervening variables is eliminated depending on the models above, the variables left are either needed to explain a variable that is eliminated or is not used in the intervening model. The groups we create are as follows in Table 9. We begin by creating Group 1 by eliminating the intervening variable *Death rate*, using the model in Table 5. For Group 2 we use the same argument and eliminate the variables *Pop 0-14* and *Pop 15-64* since they are explained by the same variables. In Group 3 we eliminate *Lab.fem % tot.*. In Group 4 instead of eliminating the response variable in the intervening models seen in Table 5, 6, 7 and 8 we keep them and eliminate the explanatory variables used.

Table 9: Grouping of GINI-HPC

Group 1	**Group 2**	**Group 3**	**Group 4**
Alt& nuc engi	Alt& nuc engi	Alt& nuc engi	Alt& nuc engi
Fert.rate	Fert.rate	Fert.rate	Death rate
Life exp.	Life exp.	Life exp.	Life exp.
Mort. rate	Mort. rate	Mort. rate	GDP.p.p.emp
Pop 0-14	Pop 65+	Pop 65+	Lab.fem % tot.
Pop 15-64	GDP.p.p.emp	Emp/pop. fem	Pop.dens.
Emp/pop. fem	Lab.fem % tot	Emp/pop. male	Urb.pop
Emp/pop. male	Elec.pow.con.	GDP.p.p.emp	Log.pop
GDP.p.p.emp	Pop.dens.	Elec.pow.con.	Log.adj.net.inc
Lab.fem % tot.	Urb.pop	Pop.dens.	Log.GDP
Elec.pow.con.	Log.pop.	Urb.pop	
Pop.dens.	Log.adj.net.inc	Log.pop	
Urb.pop	Log.GDP	Log.adj.net.inc	
Log.pop		Log.GDP	
Log.adj.net.inc			
Log.GDP			

We know that even if the intervening models are good models, they are still models, so for the analysis of the Gini coefficient we will also analyse the complete data set. We use stepwise regression on each of the groups and stop at the best PRESS. This means that one by one the variables are included in the model and if the PRESS statistic is lower or the same as before the variable is left in the model, the variables excluded from the finished model will all give a higher PRESS if included. In Table 10 and 11 we see the models statistics.

Table 10: Fit statistics for groups of GINI-HPC

All variables		Group 1		Group 2	
Root MSE	6.20205	Root MSE	5.85298	Root MSE	5.84273
R-Square	0.7053	R-Square	0.7566	R-Square	0.7592
Adj R-Sq	0.7009	Adj R-Sq	0.7522	Adj R-Sq	0.7531
AIC	1271.68834	AIC	1300.09786	AIC	1301.04817
PRESS	10643	PRESS	10061	PRESS	10140

Table 11: Fit statistics for groups of GINI-HPC

Group 3		Group 4	
Root MSE	5.81435	Root MSE	6.51360
R-Square	0.7624	R-Square	0.6996
Adj R-Sq	0.7555	Adj R-Sq	0.6931
AIC	1299.24210	AIC	1362.03106
PRESS	10091	PRESS	12478

We look at the goodness of fit in Table 10 and 11 and see that the two models from Group 1 and Group 3 that has the lowest PRESS also have high R-square so we look closer at these two models. The model for Group 2 also has a relatively small PRESS but has several more explanatory variables so for the simplicity and goodness of fit we do not analyse this model further.

Table 12: Parameter estimates Group 1

Parameter	Estimate	Pr $> \|t\|$
Intercept	-65.577230	>.0001
Life exp.	0.917397	>.0001
Mort. rate	0.403076	>.0001
Pop. 0-14	0.703150	>.0001
Elec.pow.con	-0.001020	>.0001
Urb.pop.	0.252385	>.0001

Table 13: Parameter estimates Group 3

Parameter	Estimate	Pr> $\|t\|$
Intercept	-8.703615	0.5855
Fert. rate	0.493753	0.0092
Mort. rate	0.435718	>.0001
Pop. 65+	-0.815274	>.0001
Emp. fem	-0.237524	0.0010
Emp. male	0.204032	0.0060
Elec.pow.con.	-0.000712	>.0001
Urb.pop.	0.256238	>.0001
Log.GDP	-0.493119	0.0092

As we can see the two models differ by just two extra variables in the latter model. In Figure 7 and 8 we see plot of the diagnostics fit, the models look very much alike when it comes to residuals and prediction plots. Since there is no preferable model in the fit diagnostics or statistics we choose the model from Group 1 based on simplicity. It can be noticed that in both figures

we see that two observations have more leverage then the others, the two observations are the ones for Kenya and Tajikistan.

3.2 GINI-OECD

We make the same analysis here as in the last section. If we look at the two data sets and consider what we know so far, a difference in the results for the two data sets is to be expected. A guess could have been that the GINI-OECD can be divided analogous to how the GINI-HPC was divided. If a model is made with response variable *Death rate* and explanatory variables as in the GINI-HPC we can examine the significance of the variables and the R-square and see that the model is not in any way relevant for this data set. So for the GINI-OECD we take the same approach as to the previous data set and start from scratch. When looking at where the missing values lie we decide to exclude *Agri. land* due to lack of observations, it is now 5 observations that has one or more missing values.

Although the same models for reducing the data set cannot be used we can still try to see if there is a sufficient model for the variable *Death rate* as in the previous data set. It turns out that analysing the variable is more complex then in the previous data set, in which only four variables were used for a sufficient model. In this data set the model best suited for data is as shown in Table 14 below. Although the model fit is more than acceptable the model is not as good, in terms of R-square and PRESS, as for the model with same response variable created for GINI-HPC.

Table 14: Parameter estimates for Death rate

Variable	Parameter Estimate	Pr $< \|t\|$
Intercept	59.55075	<.0001
Fem/male sec.enroll	-0.01828	0.0008
Life exp.	-0.63446	<.0001
Mort. rate	0.09777	0.0006
Pop. 65+	0.54270	<.0001
Emp/pop. fem	0.14241	<.0001
Emp/pop. male	-0.08626	<.0001
Lab.fem % tot.	-0.21897	<.0001
Urb.pop	0.01344	<.0001
R-square= 0.9559		

As in the previous data set we look at the variables not used to explain *Death rate* and find that the model shown below in Table 15 for the variable *Fert. rate* has a good fit with plots for residuals looking normally distributed and prediction on observed going along the diagonal.

Table 15: Parameter estimates for Fertility Rate

Variable	Parameter Estimate	Pr $< \|t\|$
Intercept	-2.15946	<.0001
Fem/male sec.enroll	0.00691	<.0001
Pop. 0-14	0.09580	<.0001
Pop. 65+	0.03158	<.0001
Int. users	0.00104	0.0086
GDP.p.p.emp	0.00000652	<.0001
Lab.fem % tot.	0.01389	<.0001
Fem. leg.	-0.00647	0.0021
R-square= 0.8903		

If we continue analysing different intervening models for the remaining variables, we see that some of the variables can be explained but with poor goodness of fit so for the simplicity of creating new groups and for the groups not to be too entangled the two models above are the ones we base our groups on. We create the groups in the same way we did for GINI-HPC, eliminating the intervening variables that are explained by other variables in the group and keep the variables either explaining or not used in the models for the eliminated ones. The reduced groups are created as seen in Table 16. So in the first group we have eliminated both *Death rate* and *Fert. rate*, in Group 2 *Fert. rate* in Group 3 *Death rate* and in Group 4 the explanatory variables for both are deleted.

Table 16: Grouping of GINI-OECD

Group 1	**Group 2**	**Group 3**	**Group 4**
Alt& nuc engi	Alt& nuc engi	Alt& nuc engi	Alt& nuc engi
Fem/male sec.enroll	Fem/male sec.enroll	Fem/male sec.enroll	Death rate
Life exp.	Death rate	Fert.rate	Fert. rate
Mort. rate	Pop. 0-14	Life exp.	Pop. 15-65
Pop. 0-14	Pop. 65+	Mort. rate	Lab.w.sec edu fem
Pop. 15-64	Int. users	Pop. 15-65	Lab.w.sec edu male
Int. users	GDP.p.p.emp	Pop. 65+	Lab.w.prim edu
Emp/pop. fem	Lab.w.sec edu fem	Emp/pop. fem	Elec.pow.con.
Emp/pop. male	Lab.w.sec edu male	Emp/pop. male	Armd.personnel
GDP.p.p.emp	Lab.w.prim edu	Lab.w.sec edu fem	Log.pop
Lab.w.sec edu fem	Lab.fem % tot	Lab.w.sec edu male	Log.adj.net.inc
Lab.w.sec edu male	Elec.pow.con.	Lab.w.prim edu	Log.GDP
Lab.w.prim edu	Pop.dens.	Lab.fem % tot	
Lab.fem % tot.	Urb.pop	Elec.pow.con.	
Fem. leg.	Fem. leg.	Urb.pop	
Elec.pow.con.	Elec.pow.con.	Armd.personnel	
Pop.dens.	Armd.personnel	Log.pop	
Urb.pop	Log.pop.	Log.adj.net.inc	
Armd.personnel	Log.adj.net.inc	Log.GDP	
Log.pop	Log.GDP		
Log.adj.net.inc			
Log.GDP			

So we now do exactly what we did for GINI-HPC, which is to use stepwise regression and stop when our PRESS is as small as possible. In Table 17 and 18 we see the result of the goodness of fit statistics.

Table 17: Grouping of GINI-OECD

All variables		Group 1		Group 2	
Root MSE	2.65641	Root MSE	2.87938	Root MSE	2.86179
R-Square	0.6416	R-Square	0.5703	R-Square	0.5840
Adj R-Sq	0.6221	Adj R-Sq	0.5560	Adj R-Sq	0.5614
AIC	471.54612	AIC	493.84488	AIC	494.78200
PRESS	1194.45243	PRESS	1362.97079	PRESS	1384.75015

Table 18: Grouping of GINI-OECD

Group 3		Group 4	
Root MSE	2.87938	Root MSE	3.19550
R-Square	0.5703	R-Square	0.4938
Adj R-Sq	0.5560	Adj R-Sq	0.4737
AIC	493.84488	AIC	533.95126
PRESS	1362.97079	PRESS	1693.48174

As we can see here, contrary to GINI-HPC, the best fitted model for GINI-OECD is the one we create where all variables were used from the beginning. We also saw in the model created for *Death rate* that the model fit was better in GINI-HPC than for this data set. We should also remember that the model chosen for GINI-HPC was the one created from Group 1, the group where only *Death rate* was eliminated. Let us take a look at the estimated parameters in Table 19

Table 19: Parameter estimates

Parameter	Estimate	Pr $> \|t\|$
Intercept	111.555483	0.0009
Fem/male sec.enroll	0.153972	0.0040
Death rate	-2.967997	$>$.0001
Fert.rate	6.305182	$>$.0001
Life exp.	-1.393607	0.0005
Mort. rate	1.571114	$>$.0001
Pop. 65+	2.177106	$>$.0001
GDP.p.p.emp.	-0.000123	0.0034
Urb.pop.	-0.099475	0.0007

In Figure 9 we see plots over residuals and scatter plot over the predicted on observed. Also in this model the residuals look normally distributed and randomized.

3.3 Conclusion

The models created correspond well to what could be expected from the two data sets, the model for GINI-OECD can be seen as a better more precise model then the model for GINI-HPC since it both have lower PRESS and root-mean-square error. Both models take the age demographic into consideration, something to be expected since the part of the population that actually has an income is very much determined by age. Let us now look at the difference in the coefficients in the two models. Two of the three explanatory variables that are used in both models switch signs between them. In the model for GINI-HPC the *Life expectancy at birth* is a positive coefficient while in the model for GINI-OECD it is negative. So if all other variables stay the same in the OECD-countries it is positive to have a long life expectancy in order to have as little inequality as possible and the other way around for the countries represented in GINI-HPC. The other variable that switches sign is *Urban population* also here the coefficient is negative for the OECD-countries while positive for the GINI-HPC data set. So high urban populations and high life expectancies have completely different effects in the two models. One explanation at least for the difference in urban population is that in an international model a high urban population gives more diversion in social class, i.e. a large urban population with high income and a very poor rural population. In the OECD-countries the difference in economy between the urban and rural population is no longer a major significant factor.

We can also see in the model for GINI-OECD that even though the inequality is lower with a high life expectancy, the Gini coefficient grows with a large *Population 65+*. As mentioned earlier the age demographic is an important factor, as having a large elderly community often means having a large group with a low income. This leaves us in the paradoxical position that if we want equality in income we want to have populations with high life expectancy but where no one ever gets old. This could also be the reason that for GINI-OECD the parameter estimate for *Crude death rate* is negative. Equally we can see that in the model for GINI-HPC the Gini coefficient grows with a large *Population 0-14*, this is also a low income part of a population so this is to be expected.

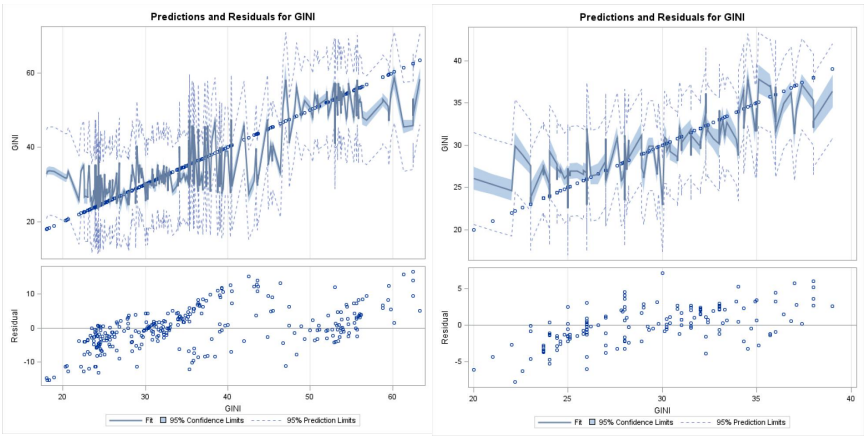
In both models the coefficient for *Infant mortality rate* is positive, this variable is a very good indicator of health care in a country and the assumption that a good health care system implies a more equal income seems correct. In GINI-HPC the *Electric power consumtion* is the last variable and the coefficient is positive, a large power consumption can be an example of several

things including both demographics and the level of the country's industrialisation. In the model for the OECD countries when the indicator of an equal school system *Ratio of female to male secondary enrolment (%)* grows the Gini coefficient grows as well. It should be mentioned that the observations used in the GINI-OECD data set varies between 93.85-117.3 with only 66 out of 160 observations being below 100 . A high fertility rate also seems to increase income inequality, while the indicator on economy *GDP per person employed* results in less inequality.

Now let us turn our attention to the plot between the observed value and the predicted values, in Figure 4 we see the predicted values plotted by the observed.

Figure 4: Predictions for Gini coefficient by observed

(a) GINI-HPC (b) GINI-OECD



As we see the model for GINI-HPC has a far greater range than GINI-OECD both in the predicted values and between the observed, something we could already see in the Figures 7 and 9 where the former looks more clustered. The idea all along has been to use the model for GINI-HPC as an indicator, not a prediction, of what increases and decreases the Gini coefficient, and as mentioned in the data background the quality of the data used is not always perfect. On the other hand the quality of data for the Gini coefficient for the OECD-countries is very good, but again as we can see the model we have created can be used as an excellent indication, but not an exact prediction of the Gini coefficient.

4 Discussion

The analysis could have been improved in many ways, mostly in terms of the data. With this sort of data an exact model for prediction is often very hard to find. It can be more interesting to have a model showing several indicators, as in this paper, than to have a more exact model including e.g. quantiles of income, which should make the prediction better since quantiles are a part of calculating the Gini coefficient. For both data sets more observations would have been desirable both in terms of observations for the Gini coefficient using the same measurements for calculation and more observations for the explanatory variables. Some of the variables had to be excluded from the analysis because the low number of observations, and the analysis would have been better if these variables could be included in the analysis. It would have been preferable if the quality of data was the same for all variables. For the data from the World Bank it is sometimes hard to know the quality of data because it is often collected from statistical institutions in each country, not from a independent organisation. The quality overall is good but variance in data can of course increase the variance in the models.

Another desirable improvement would have been to have had more current data, as the latest observations are from 2006, and since then a lot has happened in terms of both infrastructural and social factors. Several OECD-countries have had government changes and it would be interesting to know, for example, if the financial crisis in 2008 had an effect on what factors could be good indicators of low inequality in income. One way of analysing this would have been to create two data sets, modelling data between 1990-2000 and 2001-2011, which was not possible in this instance given the date range of the data used.

5 References

[1] http://website1.wider.unu.edu/wiid/WIID2c.pdf, 26 febuary 2013

[2] http://www.wider.unu.edu/research/Database/en_ GB/
wiid/_ files/79789834673192984/default/WIID2C.xls, 26 febuary 2013

[3] http://unstats.un.org/unsd/methods/citygroup/canberra.htm , 11 October 2013

[4] http://databank.worldbank.org/data/views/variableselection/
selectvariables.aspx?source=world-development-indicators, 17 April 2013

[5] Ajit C. Tamhane och Dorothy D. Dunlop (2000) "Statistics and Data Analysis. From Elementary to Intermediate", Prentice Hall

[6] http://support.sas.com/documentation/cdl/en/statug/63033/HTML/
default/viewer.htm#statug_ glmselect _ sect025.htm, 26 September 2013

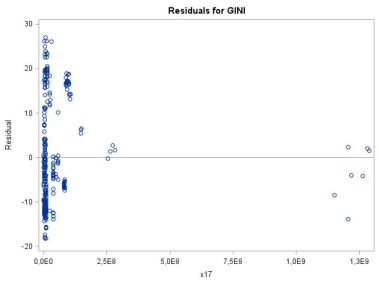
[7] http://www.unece.org/fileadmin/DAM/stats/groups/cgh/
Canbera_ Handbook _ 2011 _ WEB.pdf, 11 October 2013

6 Appendix

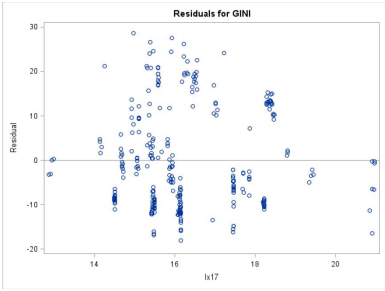
A Figures

Figure 5: Transformations for GINI-HPC

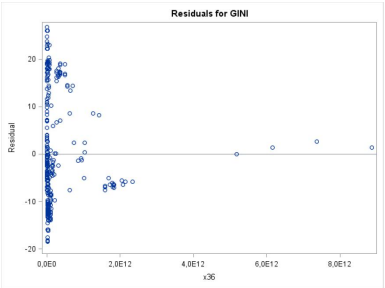
(a) Population size



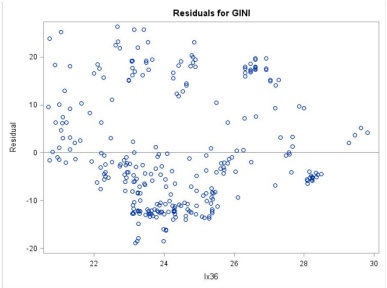
(b) log(Population size)



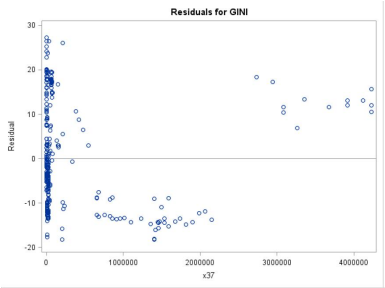
(c) Net national income



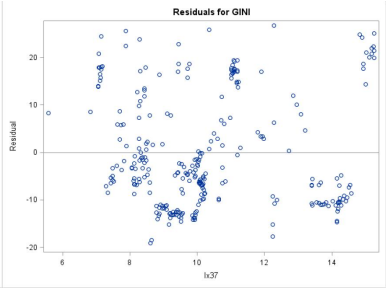
(d) log(Net national income)



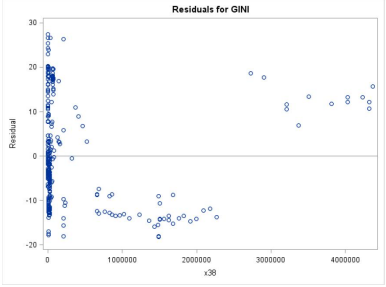
(e) GNI per capita



(f) log(GNI per capita)



(g) GDP per capita



(h) log(GDP per capita)

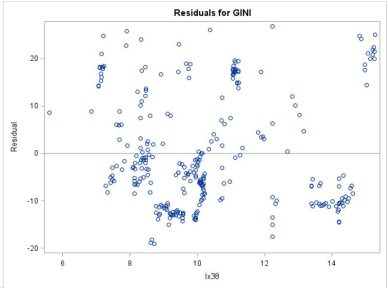
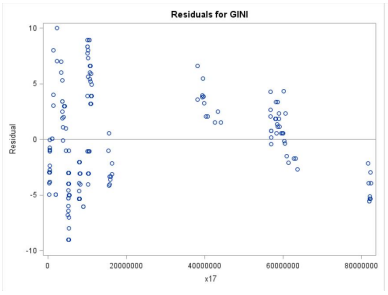
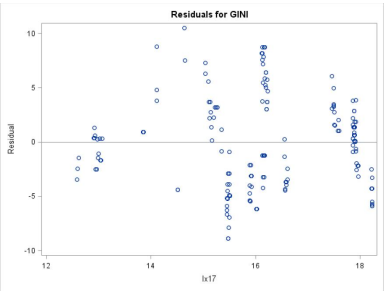


Figure 6: Transformations for GINI-OECD

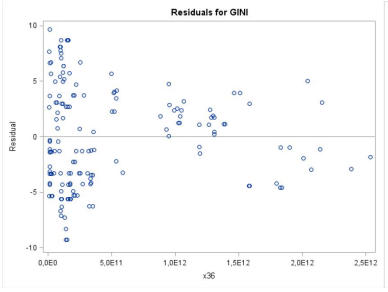
(a) Population size



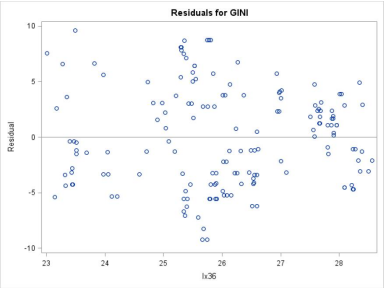
(b) log(Population size)



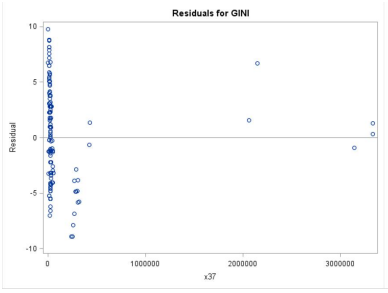
(c) Net national income



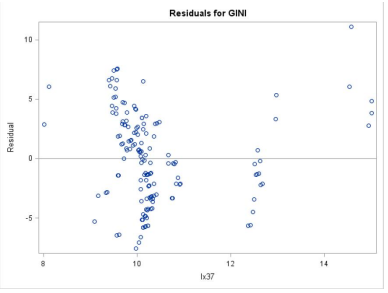
(d) log(Net national income)



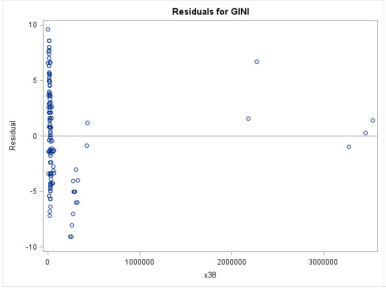
(e) GNI per capita



(f) log(GNI per capita)



(g) GDP per capita



(h) log(GDP per capita)

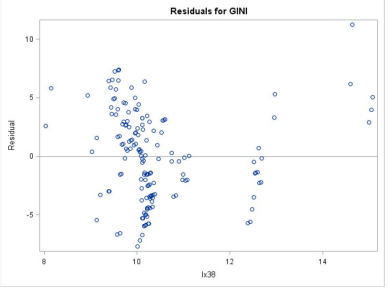


Figure 7: Fit diagnostics Group 1 GINI-HPC

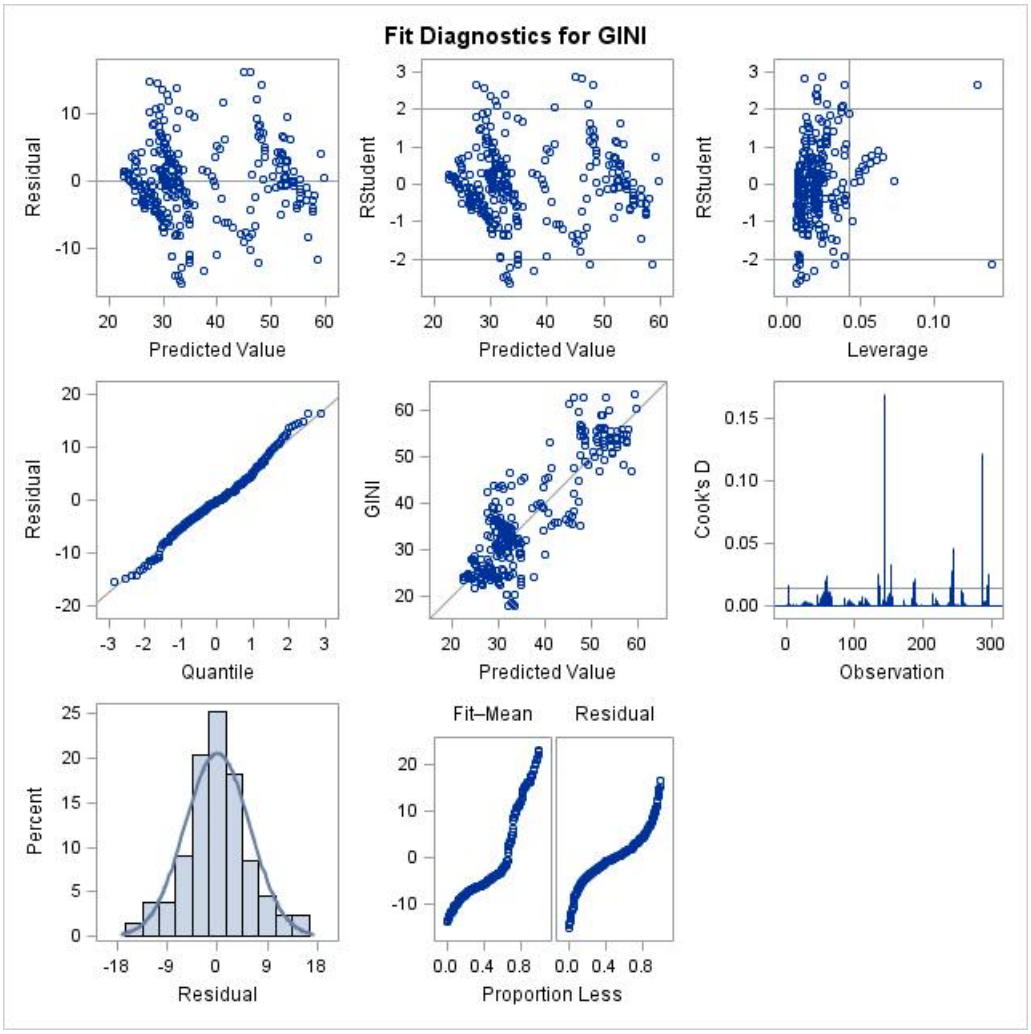


Figure 8: Fit diagnostics Group 3 GINI-HPC

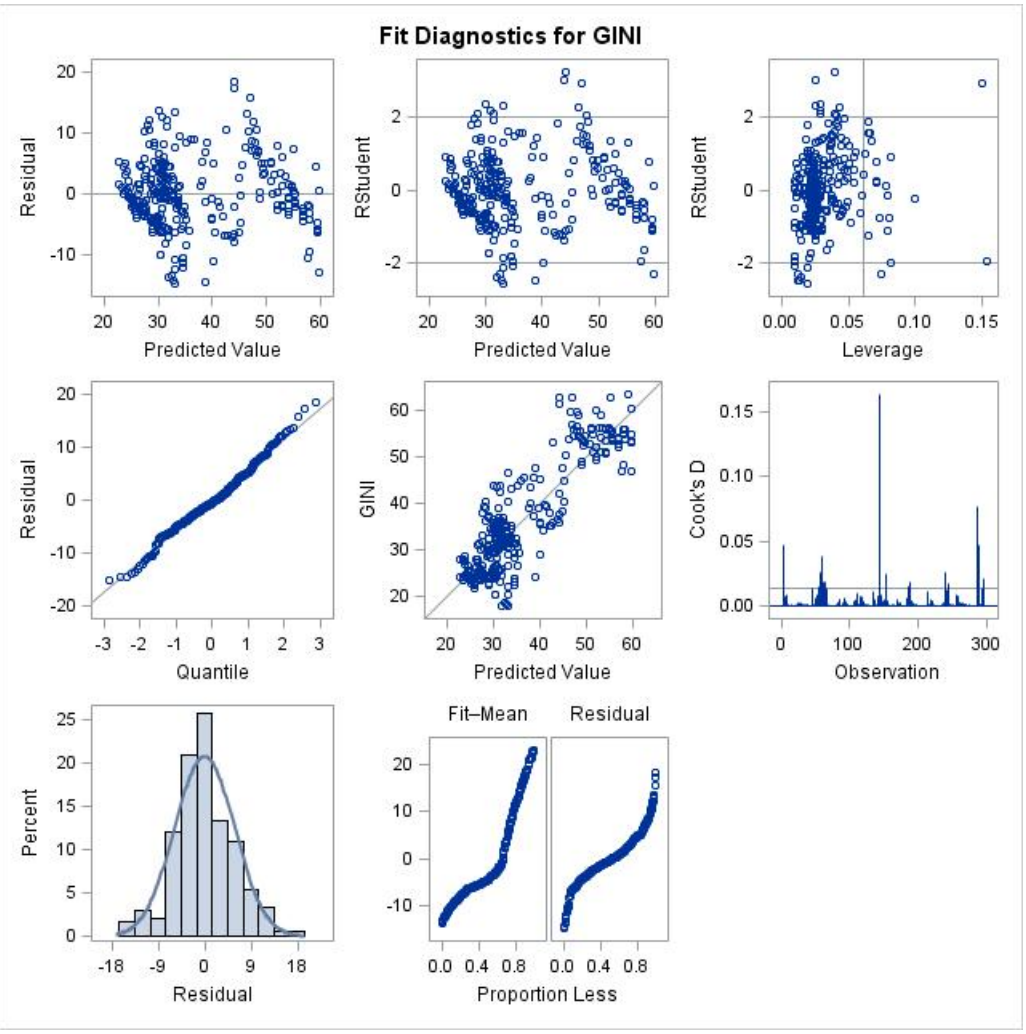


Figure 9: Fit diagnostics All Var. GINI-OECD

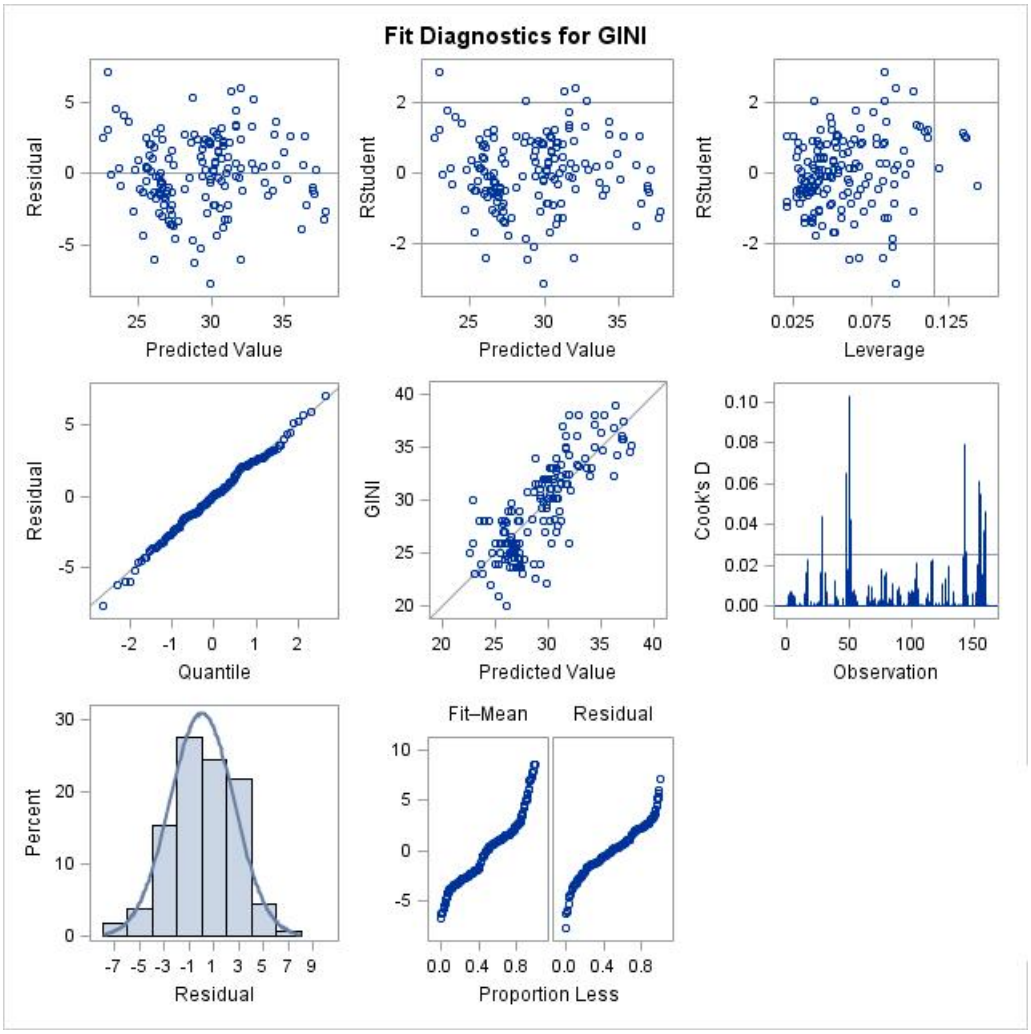


Table 20: Countries and years used in analysis

GINI-HPC		GINI-OECD	
Armenia	1996,2002-06	Austria	1995-2006
Belgium	1997, 2000	Belgium	2003-2006
Bulgaria	1992-2006	Cyprus	2005-2006
Belarus	1995-2006	Czech Republic	2005-2006
Bolivia	1999, 2000	Denmark	1995, -97, -99, -01, 2003-2006
Botswana	1994	Estonia	2004-2006
Chile	1992-1996,1998-2000	Spain	1995-2001, 2004-2006
China	1991, -95, -96, -00, 02,-03	Finland	1996-2001, 2004-2006
Czech Republic	1991-1993,-96	France	1995-2001, 2004-2006
Germany	1992-2004	Greece	1995-2001, 2003-2006
Denmark	1992	Hungary	2005-2006
Ecuador	1994-1995, 1998-2000	Ireland	1995-2001, 2003-2006
Estonia	1995, -97, -98, -00	Iceland	2004-2006
Guatemala	1998, 2000	Italy	1995-2001, 2004-2006
Honduras	1997-1998	Lithuania	2005-2006
Croatia	1998	Luxembourg	1995-2001, 2003-2006
Hungary	1991, 1993-2006	Latvia	2005-2006
Israel	1992, -97, -01	Malta	2005-2006
Italy	1991, -93, -95, -98, -00	Netherlands	1995-2001, 2005-2006
Kazakhstan	1996	Norway	2005-2006
Kenya	1999	Poland	2005-2006
Kyrgyzstan	1993, 1996-2006	Portugal	1995-2001, 2004-2006
Lithuania	1997-2004	Slovakia	2005-2006
Luxembourg	1991, -94, -97, -00	Slovenia	2005-2006
Latvia	1995-2000, 2002-2004	Sweden	2005-2006
Moldova	1997, 2000-2002	United Kingdom	1995-2001, 2005-2006
Mexico	1992, -94,-96, -98, -00, -02, -04, -05	Germany	1995-2001, -05, -06
Nicaragua	1993, -98		
Peru	1994, -97, -00		
Poland	1991-2005		
Paraguay	1995, -99		
Romania	1991		
Russia	1992, -95, -00		
El Salvador	1997-2000		
Somalia	2002		
Serbia	2003-2006		
Slovak republic	1991-1993, 1996-2006		
Slovenia	1991-2003, -05		
Tajikistan	1999		
Turkey	1994		
Ukraine	1999-2002		
USA	1991, -94, -97, -00		
Uzbekistan	-01		

Table 21: Explanatory variables

Variable	Included in data set	
	GINI-OECD	GINI-HPC
Agricultural land (% of land area)	x	x
Alternative and nuclear energy (% of total energy use)	x	x
Ratio of female to male secondary enrolment (%)	x	x
Death rate, crude (per 1,000 people)	x	x
Fertility rate, total (births per woman)	x	x
Life expectancy at birth, total (years)	x	x
Mortality rate, infant (per 1,000 live births)	x	x
Population ages 0-14 (% of total)	x	x
Population ages 15-64 (% of total)	x	x
Population ages 65 and above (% of total)	x	x
Internet users (per 100 people)	x	x
Employment to population ratio, 15+, female (%)	x	x
Employment to population ratio, 15+, male (%)	x	x
GDP per person employed (constant 1990 PPP $)	x	x
Labour force with secondary education, female (% of female labour force)	x	
Labour force with secondary education, male (% of male labour force)	x	
Labour force with primary education (% of total)	x	
Labuor force, female (% of total labour force)	x	x
Unemployment, total (% of total labour force)	x	x
Female legislators, senior officials and managers (% of total)	x	
Electric power consumption (kWh per capita)	x	x
Population density (people per sq. km of land area)	x	x
Urban population (% of total)	x	x
Armed forces personnel (% of total labour force)	x	x
LOG(Population, total)	x	x
LOG(Adjusted net national income (current US$))	x	x
LOG(GNI per capita (constant LCU))	x	x
LOG(GDP per capita (constant LCU))	x	x

B Lorenz curve

The Lorenz curve is a curve that is used to calculate inequality, although often used for income or wealth it can be used in varying ways such as diversity in demographics or populations in ecology. The curve is based on the cumulative distribution. The Lorenz curve shows the cumulative share of income aggregating to each category of the population, from lowest to richest.
For
P= Cumulative share of population
C= Cumulative share of income
I= Share of income

		Perfect equality		Ex. of income dist.	
Income category	P	I	C	I	C
Richest 20%	100	20	100	40	100
2nd richest 20%	80	20	80	30	60
3rd richest 20%	60	20	60	15	30
4th richest 20%	40	20	40	10	15
Poorest 20%	20	20	20	5	5

The examples in Table B is demonstrated in Figure 10 below, where the curve corresponds to the Ex. of income dist. and the diagonal to perfect equality.

Figure 10: The Lorenz curve [1]