Journal of Statistical Modelling: Theory and Applications
Vol. 1, No. 2, 2020, pp. 91-98
Yazd University Press 2020

JSMTA

Research Paper

Analysis of cross countries income inequality panel data: Using random effect regression trees

Hasan Kiaee* [1], Samaneh Eftekhari Mahabadi[2]
[1]Faculty of Islamic Studies and Economics, Imam Sadiq University, Tehran, Iran.
[2]School of Mathematics, Statistics and Computer Science, College of Science, University of Tehran, Tehran 14155-6455, Iran.

Received: September 25, 2020/ Revised: December 31, 2020 / Accepted: February 11, 2020

Abstract: Reducing income inequality is one of the major steps toward economic development. When the level of inequality in the distribution of income and wealth is high in the society, many economic, social and even political problems might happen. So, many studies in the economic literature tried to find the determinants of income inequality and propose some policies to decline it. In this paper, we will address the analysis of income inequality panel data across different countries through 2011 to 2015. One of the commonly used methodologies to analyze panel data is the linear mixed effects model. Since the linearity assumption might be violated, recently, the idea of mixed effect models are combined with the flexibility of tree-based estimation methods which allows for potential higher order interactions as well. In this paper, we apply the resulting estimation method, called the RE-EM tree, to the income inequality panel data. The results show that the RE-EM tree is less sensitive to parametric assumptions and provides improved predictive power compared to simple regression trees without random effects. This is due to the fact that each country applies its own specific poverty reduction measures handled via country-specific random coefficients of RE-EM tree.

Keywords: Income inequality; Gini coefficient; Mixed effects model; Mixed effect Regression trees; Panel data.
Mathematics Subject Classification (2010): 62P20, 62J05.

1 Introduction

Income inequality refers to the fact that different people earn different amounts of money. There are some different measures for evaluating income inequality in the economic literature. The GINI coefficient (or index), developed by the Italian statistician

*Corresponding author: kiaee@isu.ac.ir

Corrado Gini (1912), is an important measure of statistical dispersion that is often used to reflect the extent of income inequality, where a Gini coefficient of 0 expresses perfect equality; and a coefficient of 1 expresses maximal inequality.

There are vast number of researches about the relationship between income inequality and macroeconomic variables over time. Some of these researches address one specific country and try to find a significant regression relationship between income inequality and macroeconomic variable (Azzoni, 2001; Law and Tan, 2009). Others choose a panel regression approach for selected countries or provinces to describe income inequality variation in which some of them are reviewed here. Thalassinos et al. (2012) analyzes the relationship between income inequality and inflation in 13 European countries for the period 2000 to 2009 using panel data methodology. Their results support the hypothesis that inflation has a positive significant effect on income inequality. Halmos (2011) explores the relationship between FDI, exports, GDP and income inequality in Eastern European countries during 1991 to 2006. According to the result, export has significant effect on decreasing income inequality. Ha et al. (2019) examines the impact of urbanization on income inequality in Vietnam, using the panel data regression estimation for 63 provinces in Vietnam from 2006 to 2016. The results show that in the long term, urbanization has an impact on reducing income inequality.

Panel or longitudinal data consists of repeatedly observed measurements for each subject through different occasions. The primary goal of repeated measures study is to characterize the changes in the mean response over time and the factors that influence these changes. Modelling longitudinal data require somewhat more sophisticated statistical techniques because the repeated observations have a sequential nature which implies certain types of correlation structures. Heterogeneous variability must also be accounted for in order to obtain valid inferences.

There are three broad class of models introduced to handle longitudinal data; (i) marginal mean and covariance pattern specification (ii) transitional models (iii) mixed effect models. In the Mixed effect modelling approach, individuals are assumed to have their own subject-specific mean response trajectories over time. More specifically, the mean response is modeled as a combination of population characteristics (fixed effects) assumed to be shared by all individuals, and subject-specific effects (random effects) that are unique to a particular individual. In this framework, introducing random effects in the mean response model automatically induces some covariance pattern on the vector of repeatedly measured responses.

The above mentioned methods for modelling panel data are built upon the linearity assumption for the mean response model which could be mostly violated. So it is natural to suppose, a more flexible relationship than a linear one, which suggests consideration of methods such as nonparametric regression, regression trees, multivariate adaptive regression splines (MARS), neural networks, and so on. Sela and Simonoff (2012) generalized the linear mixed effects model to tree-based models. They focus on the EM algorithm for two-stage mixed effects models given by Laird and Ware (1982). For more information on mixed effects models, including modified estimation procedures and extensions, see Verbeke and Molenberghs (2000).

In this paper, we aim to analyze inequality panel data during 2011-2015. Since there is a considerable non-linear relation between the GINI response variable and the financial indicators through time, traditional linear mixed effect models could not de-

tect significant predictors correctly. Also, simple regression trees as non-parametric modelling tools for independent observations do not lead to powerful predictions, since there are significant positive correlation among repeated GINI measures across countries. Finally, we have fitted Mixed effect regression tree which leads to an interpretable partitions in the feature space considering potential interactions along with powerful predictions.

The remainder of this paper is organized as follows. In Section 2, the motivating data set is described. In Section 3, the linear mixed effect model and its generalization to Random Effects/EM Tree is discussed. The results of data analysis is given in Section 4. Finally, some concluding remarks will be given in Section 5.

2 Data description

The data to be analyzed in this paper is excerpted from World Bank Economic Development Database [†]. To study income inequality, the GINI coefficient is assumed to be the response variable. By examining the literature, it appears that the most important macroeconomic variables which could have significant effect on the GINI coefficient are GDP per capita, Inflation rate, Rural population and Government expenditure. There are 50 countries in the panel with a slightly varying number of observations per country due to missing values. In Figure 1, the box plots of Gini coefficient during 2011 to 2015 are given in the left panel and the scatter plots of selected macroeconomic variables against Gini coefficient for 2015 are presented in the right panel. Small changes in the value of GINI coefficient during time reveals the necessity of longitudinal analysis since there is a highly positive correlation between GINI coefficients during time. Also, the non-linear relationship between macroeconomic variables and Gini coefficient could be concluded from the scatter plots.

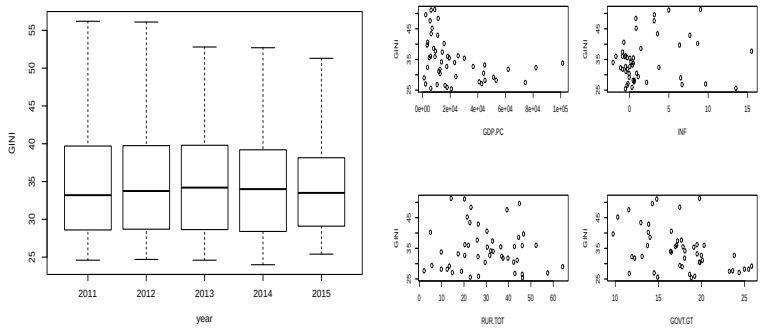

Figure 1: Left panel: Box plots of GINI through time, Right panel: scatter plots of GINI against predictors in 2015.

The response profile plots of the mean GINI coefficient conditional on categorized Inflation rate, GDP per capita and Rural population Ratio are given in Figure 2. Regarded to the Inflation and GDP per capita, countries with the high and low categories have lower GINI coefficients than medium ones. This again shows the nature of non-linear relation between these predictors and GINI. Although, the countries with low

[†]https://databank.worldbank.org/home.aspx

Rural population ratio have little GINI variation during time, the GINI coefficient of high level ones has declined over time. Also, the GINI trajectories of a smaple of countries with high inequality properties ($GINI > 40$) is plotted in Figure 3. This plot shows various trends of GINI index which emphasizes the need for country-specific effects to model the data.

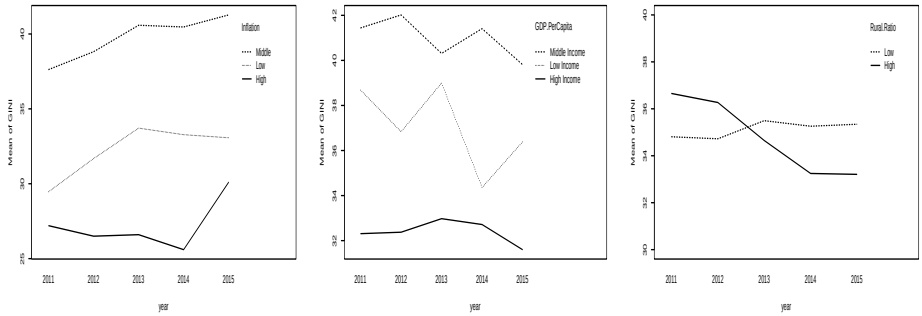

Figure 2: Mean response profile of GINI given inflation, GDP and RUR categories.

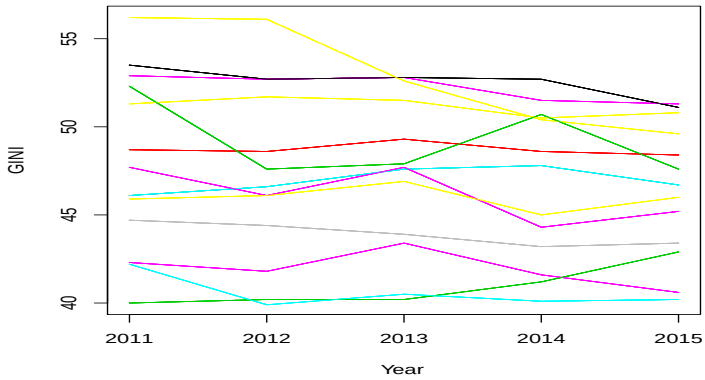

Figure 3: GINI trajectories over time for a sample of countries.

3 The RE-EM tree

Let $Y_i' = (Y_i1, ..., Y_{iT})$ be the vector of panel responses for the $i$the sample through T times $(i = 1, ..., n)$. Also, assume that there is a vector of p covariates recorded for the ith sample at time t $(t = 1, ..., T)$, denoted by $X_{it} = (X_{it1}, ..., X_{itp})$. The linear mixed effect model has the following form:

$$Y_{it} = X_{it}\beta + Z_{it}b_i + \epsilon_{it}, \quad i = 1, ..., n, \quad t = 1, ..., T$$
$$b_i \sim MVN(0, G) \perp \epsilon_{it} \sim N(0, \sigma^2),$$

where β is a $p \times 1$ vector of fixed effects attached to the vector of covariates X_{it}. Also, Z_{it} is a sub vector of X_{it} attached to the $q \times 1$ vector of random subject-specific effects b_i. It is also assumed that the random error terms ϵ_{it} are independent of b_i.

In the above traditional linear mixed effects model, the relation between the marginal mean response, $E(Y_{it}|X_{it})$ and the linear predictor, $X_{it}\beta$, is assumed to have a known linear form, which might be too restrictive an assumption. Since the functional form of this relation is frequently unknown, assuming a linear model may not be the best option. Also, when the number of potential predictors, p, becomes very large, including all of them directly may lead to overfitting and therefore poor predictions.

Assume the more general mixed effect model with the following additive form:

$$\begin{aligned} &Y_{it} = f(X_{it}) + Z_{it}b_i + \epsilon_{it}, \quad i = 1, ..., n, \quad t = 1, ..., T \\ &b_i \sim MVN(0, G) \perp \epsilon_i = (\epsilon_{i1}, ..., \epsilon_{iT}) \sim MVN(0, R), \end{aligned} \tag{1}$$

where, R is a $T \times T$ covariance matrix for the vector of random error terms. Assuming non-diagonal R matrix allows potentially correlated random error terms, as well. If $f(.)$ is a known function that is linear in the parameters, the above model reduces to the linear mixed effects model. However, assuming that $f(.)$ is an unknown potentially non-linear and non-parametric function, Hajjem et al. (2008, 2011) and Sela and Simonoff (2009) independently proposed an estimation method that uses a tree structure to estimate $f(.)$, but also incorporates subject-specific random effects, b_i. In this method, the nodes may split based on any attribute, so that different observations for the same object may be placed in different nodes. However, the method ensures that the longitudinal structure in the errors is preserved.

If the random effects, b_i, were known, (1) implies that we could fit a regression tree to $y_{it}Z_{it}b_i$ to estimate $f(.)$. If the population-level effects, $f(.)$, were known, then we could estimate the random effects using a traditional mixed effects linear model with population-level effects corresponding to the values $f(X_{it})$. Since neither the random effects nor the fixed effects are known, Sela and Simonoff (2012) proposed to alternate between estimating the regression tree, assuming that the estimates of the random effects are correct, and estimating the random effects, assuming that the regression tree is correct. They called the resulting estimator a Random Effects/EM Tree, or RE-EM Tree since alternation between the estimation of different parameters is reminiscent of the EM algorithm, as used by Laird and Ware (1982). Moreover, it should be noticed that the estimation method does not involve a true EM algorithm, so that the usual properties of the EM algorithm do not necessarily apply. Given a RE-EM tree, the associated random effects, and the estimated covariance matrices, the out-of-sample predictions are straightforward.

More specifically, the RE-EM tree is estimated as follows:

1. The estimated random effects, $\hat{b}_i$, is initialized to zero.

2. A regression tree is fitted as an approximation to $f(X_{it})$ where for $i = 1, ..., N$ and $t = 1, ..., T$, $\{y_{it} - Z_{it}\hat{b}_i\}$'s are the values of target variable and the vector of covariates attached to fixed effects, $X_{it} = (x_{it1}, ..., x_{itp})$, construct the feature space.

3. A linear mixed effect model is fitted assuming y_{it} as the response variable and the approximated $f(X_{it})$ in step 2 as the known fixed effect part, and $Z_{it}b_i$ as the random effect part. Since f is estimated based on a regression tree for example with K leaves which partition the feature space into a set of K regions, namely, $R_1, ..., R_K$, then $f(X_{it}) = \sum_{k=1}^{K} I(X_{it} \in R_k)\mu_k$ where μ_K is the mean response parameter for the

individuals in R_k. Actually, the following linear mixed effect model is estimated,

$$y_{it} = \sum_{k=1}^{K} I(X_{it} \in R_k)\mu_k + Z_{it}b_i + \epsilon_{it}$$

and $\hat{b}_i$ is extracted.

4. iterate steps 2 to 4 until the estimated random effects $\hat{b}_i$ converge.

5. Replace the mean responses in the leaves of the fitted tree in step 2, with $\hat{\mu}_k$'s as the output of RE-EM tree.

4 Results of data analysis

In this section, the RE-EM tree is applied for the analysis of Income Inequality data extracted from the World Bank Data Bases webpage [‡]. Figure 4 illustrates the fitted RE-EM tree which have 4 leaves, and partitions the countries during time according to the values of three selected predictors; Rural ratio, Inflation and GDP per capita. Rural ratio has been selected as the most influential predictor in the root of tree. As was expected from the economic theories, countries with higher Rural ratio would have lower levels of GINI. Among countries with lower Rural ratio and lower inflation, those with $4000 < GDP.PC \leq 12000$ have the highest GINI which indicates the worst inequality management. Also, the mean observed GINI and the number of records are given in each leaf. The left panel of Figure 5 shows the boxplots of GINI in 4 leaves of tree during the time. Also, the predicted against actual GINI values are plotted in the right panel. According to the correlation values, the presented RE-EM tree has a high predictability power.

The two main achievements of the fitted RE-EM algorithm are (1) the allowance of non-linear relation between response variable and the predictor variables as well as their potential interactions while preserving simple interpretability of decision trees. From economic point of view, as is illustrated in Figure 4, for countries with lower Rural Population Ratio there is an interaction between Inflation and GDP per capita, while for higher Rural Population Ratios, it does not exist. This could just be presented by a tree approach not the parametric linear models which lead policy makers to choose different poverty reduction measures according to the suggested partitions of the feature space, (2) including random subject-specific effects in the regression tree estimation process allows for the detection of between individual variabilities not accounted for in the simple structured regression tree which leads to more predictability. Actually, this random effects perfectly considers the fact that each country applies its own specific poverty reduction measures which leads to various GINI coefficient paths of different countries through time.

Also, the simple tree without random effects is fitted on these data for which the predicted against actual GINI values are given in Figure 6. Comparing this figure with the right panel of Figure 5, apparently shows that including random effects has improved the tree's prediction power.

[‡]https://databank.worldbank.org/home.aspx

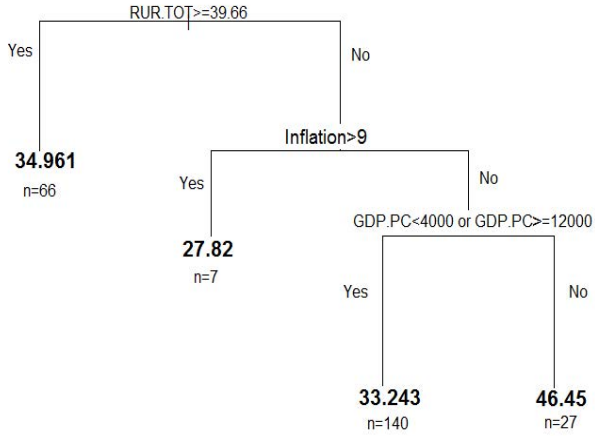

Figure 4: RE-EM tree for Inequality panel data through 2011-2015.

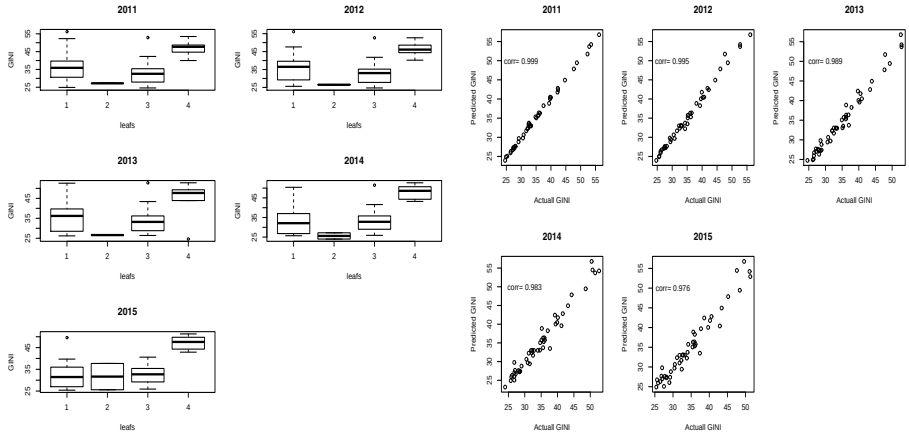

Figure 5: Left panel: Boxplots of GINI split by the RE-EM tree's leaves during time, Right panel: scatter plots of predicted against actual GINI in different years based on RE-EM tree.

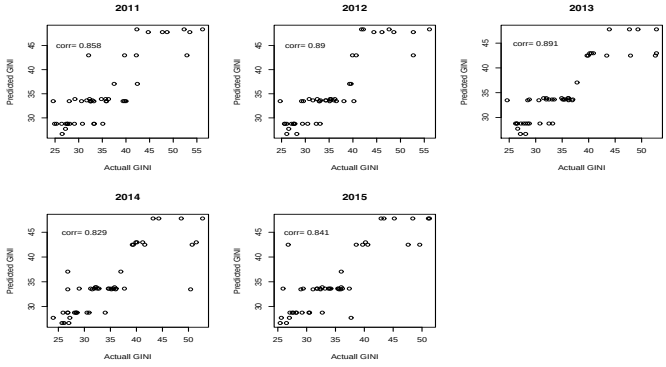

Figure 6: Scatter plots of predicted against actual GINI based on simple tree without random effect.

5 Discussion

In this paper, we have analyzed cross-country inequality panel data based on 50 countries during 2011 to 2015. All previous literature in this field has applied different models based on linearity assumption which could be violated as were shown in the descriptive plots given in the text. Hence, we proposed to fit a random effect tree to analyze these longitudinal data which leads to better interpretation and predictability power. Based on the fitted RE-EM tree rural ratio as the root, Inflation and GDP per capita as the following branches are the important macro economic variables influencing GINI index.

References

Azzoni, C.R. (2001). Economic growth and regional income inequality in Brazil. *Annals on Regional Sciences*, **35**, 133–152.

Gini, C. (1912) Variabilita‘ e Mutabilita‘: Contributo Allo Studio Delle Distribuzioni e Relazioni Statistiche, vol. III (part II). Bologna: Cuppini.

Ha, N.M., Le, N.D. and Trung, P. (2019). The impact of urbanization on income inequality: A study in Vietnam. *Journal of Risk and Financial Managment,* **12**(3), 146–160.

Hajjem, A., Bellavance, F., and Larocque, D. (2008). Mixed-effects regression trees for clustered data. LesCahiers du GERAD G-2008-57.

Hajjem, A., Bellavance, F. and Larocque, D. (2011). Mixed effects regression trees for clustered data. *Statistics and Probability Letters,* **81**, 451–459.

Halmos, K. (2011). The effect of FDI, export and GDP on income inequality in 15 eastern european countries. *Acta Polytechnica Hungarica,* **8**(1), 123–136.

Laird, N.M. and Ware, J.H. (1982). Random effects models for longitudinal data. *Biometrics*, **38**, 963–974.

Law S.H. and Tan. H.B. (2009). The role of financial development on income inequality in Malaysia. *Journal of Economic Development,* **34**(2), 153–168.

Sela, R.J. and Simonoff, J.S. (2009). RE-EM trees: a new data mining approach for longitudinal data. NYU Stern Working Paper SOR-2009-03.

Sela R.J. and Simonoff J.S. (2012). RE-EM trees: A data mining approach for longitudinal and clustered data. *Machine Learning*, **86**, 169–207

Thalassinos, E., Uğurlu, E. and Muratoğlu, Y. (2012). Income inequality and inflation in the EU. *European Research Studies*, **XV**(1), 128–140.

Verbeke, G. and Molenberghs, G. (2000). Linear Mixed Models for Longitudinal Data. New York: Springer.