Robust Estimation and Outlier Detection on Panel Data

International Conference on Robust Statistics 2017

By:
Ahmed R.M. Al Sayed
Zaidi Isa
S.K. Sek

OUTLINE

1. Introduction
2. Problem Statement
3. Objectives
4. Scope of the Study
5. Literature Review
6. Methodology
7. Results
8. Conclusions

Introduction

- The presence of outliers issue has been clearly noted in different fields of science.
- Outliers are the few observations which have significant different characteristics and stay far away from the other observations in the dataset.
- The presence of outliers in the dataset leads to miss-estimation and biased results.
- We apply robust estimation on panel data using the energy consumption data

Problem Statement

Application on environment science:

- The consumption of non-renewable energy can have crucial effects on environment and economy. Yet there is no consensus on the relationship as results may vary across countries.
- Previous studies did not consider the outliers problem in analysing the relationship.
- Hence, we explore the relationship between energy consumption (EC), GDP and CO_2 in statistical aspect for different groups of panel countries by taking care about outliers in estimation.
- We report partial results on the impacts of GDP and EC on CO_2

Objective

1. To model the impacts of EC and GDP on environmental quality (emission of CO_2) for two different levels of economies

2. To reveal the best robust estimator (M-, Median, S- and MM-estimator) against OLS estimator in the presence of outliers in the panel data.

Scope of the Study

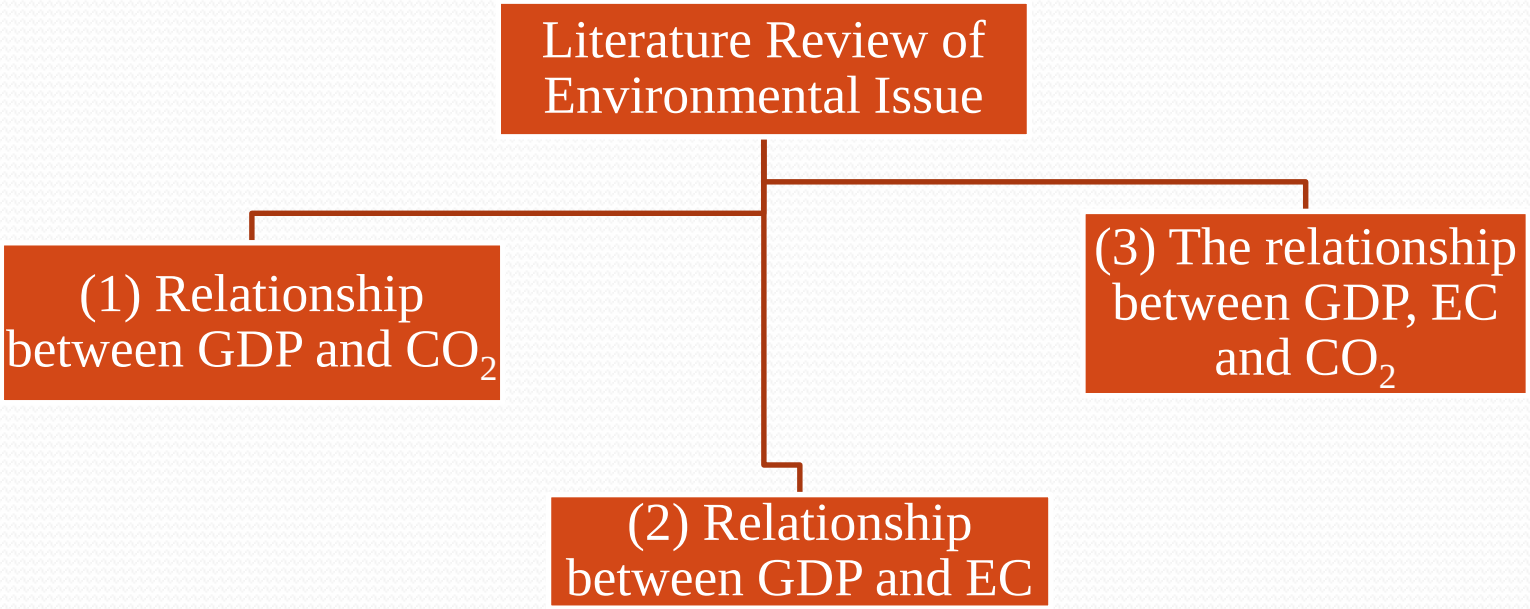
- The panel data includes 29 countries;
  - 17 developed countries
  - 12 developing countries.

- The annual data take the range from 1960 to 2008, and are obtained from the World Bank website.

- Dependent variable:
  - Carbon dioxide emissions, metric tons per capita (CO_2).
- Independent variables:
  - Gross domestic product in US$ (GDP)
  - Aggregate energy consumption, kilo tons of oil equivalent per capita (EC).

Literature Review

Three strands of studies



Differences results may due to data, estimation approaches and omitted variables bias (Stern & Common (2001), Dinda (2004), Yang & Zhao (2014))

… continue

- GDP and CO_2

- Testing the validity of EKC hypothesis
- Leads to inconclusive results

- GDP and EC

- Pioneered by Kraft and Kraft (1978)
- 4 types of hypothesis:

(1) neutral or no causal relationship;
(2) conservation hypothesis or uni-directional effect (GDP to EC)
(3) growth hypothesis or uni-directional effect (EC to GDP)
(4) feedback hypothesis or bi-directional causality effect

- Most studies detected relationship

- GDP, EC and CO_2

- Most studies detected cointegrating relationship

Methodology

Panel data analysis

It is a combination of longitudinal data observed over a period of time.

- **Advantage of using panel data regression**
  - It controls the variation of time series and cross sections simultaneously.
  - It allows covering more observations by pooling the time series data and cross sections.
  - It controls the individual heterogeneity which gives more informative data, less collinearity among the variables.

Methodology

The constructed model is

$$CO2_{it} = \alpha_i + \beta_1 EC_{it} + \beta_2 GDP_{it} + \beta_3 D_{it} + \varepsilon_{it}$$

$$D_{it} = \begin{cases} 1 & developing \\ 0 & developed \end{cases}$$

Methodology

OLS and Robust Estimators in Panel Data Regression.

- Robust estimators tries to reduce the effect of the outliers by giving less weight to those observations, not by excluding or ignoring them.
- We compare 4 estimators: Median-estimator, M-estimator, S-estimator and MM-estimator
- Benchmark: OLS

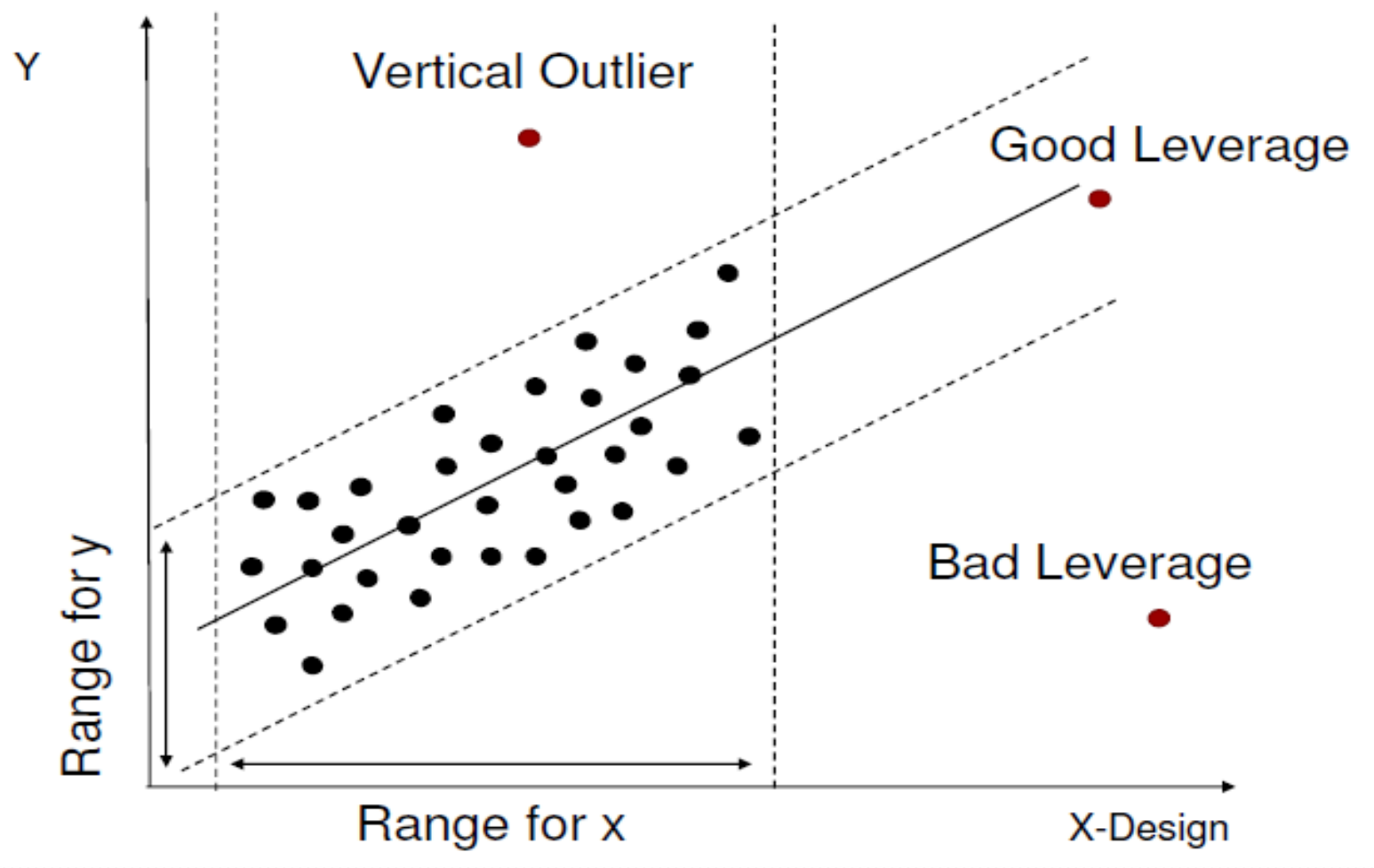
- The characteristics of robust estimators:
  - to perform as accurate as OLS estimator when the assumptions of OLS estimator have met.
  - to perform much accurate than the OLS estimator when the assumptions of the latter have not met.

Estimator	Based on	Formula	Drawback
OLS	minimizing the sum of squared residuals.	$\hat{\beta}_{OLS} = \arg\min_{\beta} \sum_{i=1}^{n} r_i^2(\beta)$	it may provide **bias** estimation in the presence of outliers.
Median estimator	minimizing the sum of absolute values of residuals.	$\hat{\beta}_{L_1} = \arg\min_{\beta} \sum_{i=1}^{n} \lvert r_i(\beta) \rvert$	it does not tackle the existence of **bad leverage points**.
M-estimator	minimizing the weighted sum of residuals by including loss function ρ.	$\hat{\beta}_M = \arg\min_{\beta} \sum_{i=1}^{n} \rho\left\{\frac{r_i(\beta)}{\sigma}\right\}$	it is not robust regards the **bad leverage points.**
S-estimator	minimizing the measure of dispersion σ of the residuals which is less sensitive to outliers than the variance.	$\hat{\beta}_S = \arg\min_{\beta} \hat{\sigma}^s \{r_1(\beta), r_2(\beta), \ldots . r_n(\beta)\}$ It has low Gaussian efficiency.	
MM-estimator	is based on the combination of M-estimator's efficiency and S-estimator's robustness.	$\hat{\beta}_{MM} = \arg\min_{\beta} \sum_{i=1}^{n} \rho\left\{\frac{r_i(\beta)}{\hat{\sigma}^s}\right\}$	

Methodology

3. Outliers in Panel Data

Properly there are three types of outliers;

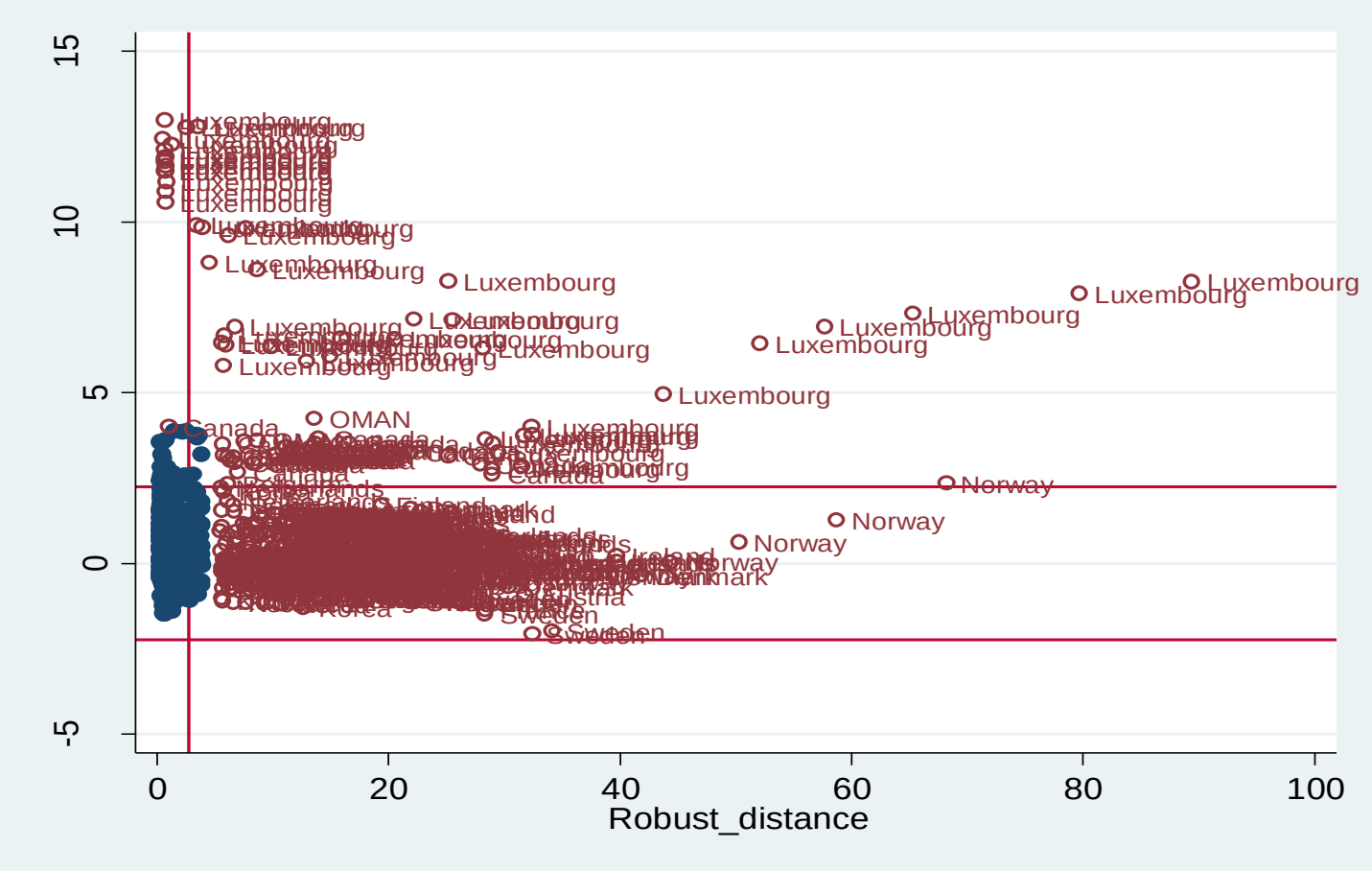


Methodology

Types		Location	The effects
Leverage points	bad leverage	x-dimension but staying **far from the regression line**.	It has a high influence towards OLS estimation model (**intercept and slopes**).
	good leverage	x-dimension but they are **outlying close to the regression line**	It may affect the estimated **standard errors** in statistical inference
Vertical outliers		y-dimension but **not in the x-dimension.**	It affects the OLS estimator in estimating the **intercept**.

RESULTS

- The diagnostic outliers plot using Mahalanobis distance



RESULTS

Coefficient	OLS-estimator	Median estimator	M-estimator	S-estimator	MM-estimator
GDP	-1.3*** (0.20)	-2.21*** (0.15)	-2.24*** (0.12)	-1.8*** (0.24)	-2.27*** (0.14)
EC	2.37*** (0.21)	3.12*** (0.16)	2.74*** (0.11)	2.55*** (0.19)	2.77*** (0.12)
Dummy	-1790.1*** (33.7)	-794.4*** (24.1)	-682.1*** (19.9)	-158.2*** (29.6)	-606.7*** (25.5)
Constant	3.51*** (0.48)	1.85*** (0.35)	1.29*** (0.26)	1.34*** (0.34)	1.32*** (0.27)
RMSE	4.92	5.14	4.68	4.79	4.88
R-sq	0.41	0.39	0.62	0.53	0.47

RESULTS

- All coefficients are significant at 1% significant level for all estimated models.
- The results show that the lowest RMSE value is in the model which used M-estimator with 4.68.
- the results of the standard error support that the lowest standard error values in estimated model by using M-estimator.
- The mean of CO_2 is lower in developing countries than that in developed countries.

CONCLUSION

- The result indicates that the panel data contains different types of the outliers throughout the whole period of 1960 to 2008, but most of those outliers are found in developed countries group.
- Robust Mahalanobis distance is used to detect the outliers.
- In general, robust estimators appeared to have better properties than OLS estimator when the dataset has several types of outliers.

CONCLUSION

- M-estimator is the best robust model fitting the data by considering the influence of different types of outliers, i.e. leverage points and the vertical outliers.

POLICY RECOMMENDATIONS

- Overall evidence implies that energy is an important factor that contributes to economic growth, as well as environment degradation.
  - to bring the importance of renewable energy to the forefront of the wider energy usage.
  - Strategic plans should include utilization of environmentally friendly technology and renewable clean energy.
  - Environmental policy is crucial to balance economic growth and environmental quality for sustainable growth in both developed & developing economies.

Thank You
for your kind attention