Statistic regression and **open data** approach for **identifying economic indicators** that influence **e-commerce**

Apollinaire Barme, Simon Tamayo & Arthur Gaudron

To cite this work:
Apollinaire Barme, Simon Tamayo, Gaudron Arthur. Statistic regression and open data approach for identifying economic indicators that influence e-commerce. 20th International Conference on Urban Transportation and City Logistics, May 2018, London, United Kingdom. ⟨hal-01790991⟩

Simon Tamayo
Associate professor
Centre for Robotics – Dept. Mathematics & Systems
MINES ParisTech – PSL Paris, France

BACKGROUND AND POBLEM STATEMENT

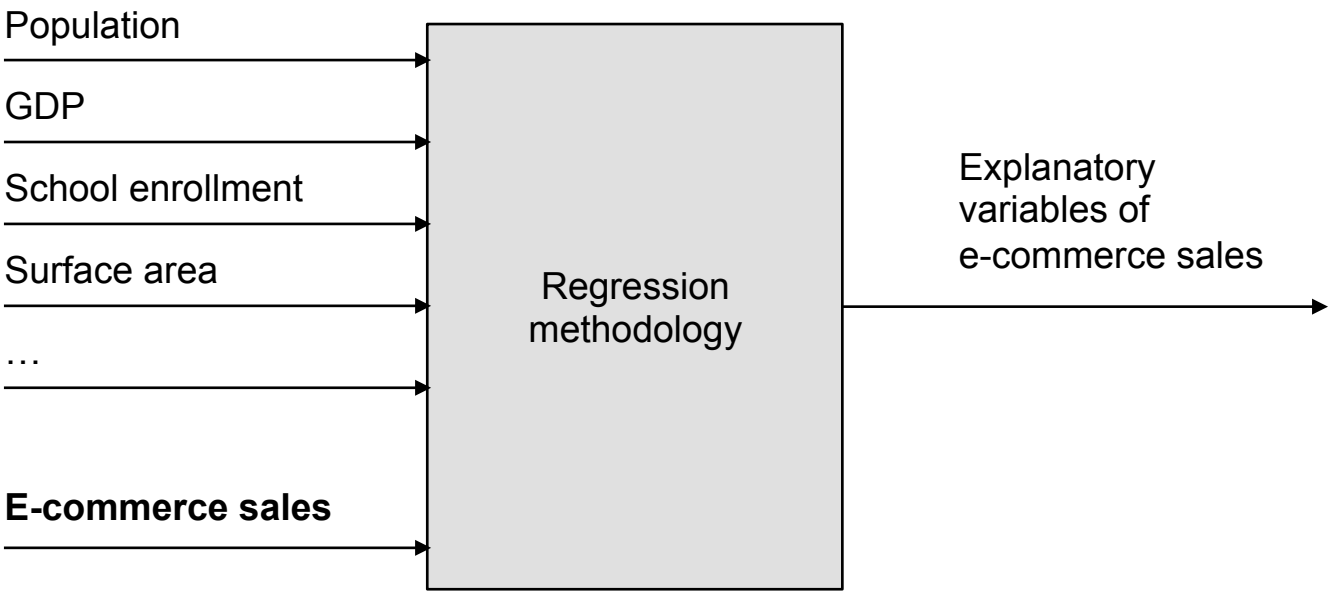
E-commerce embodies a particularly complex system that challenges logistics systems as it is related to a **high fragmentation of receivers** and deliveries of **smaller quantities**.

This paper aims at **identifying linear relations between national e-commerce sales and major economic and demographic indicators**. It deals with the following research questions:

- Which major variables influence e-commerce sales?
- How to identify these variables?
- Would it be possible to predict the sales of e-commerce in a country from these variables?

simon.tamayo@mines-paristech.fr

PROPOSED APPROACH



simon.tamayo@mines-paristech.fr

3 MAIN STEPS

Data Collection

The input data is divided in two:

1 - The dependent vairable, i.e. e-commerce sales of each country (“Global B2C E-commerce Report”).

2 - The explanatory variables, i.e. demographic and economic indicators (collected from “Country Profiles Data” in the World Bank Open Data Portal).

Regression and Selection

We perform linear regression and we compute the Student’s p-values for the different parameters, the Fisher’s p-value for the set and the R^2.

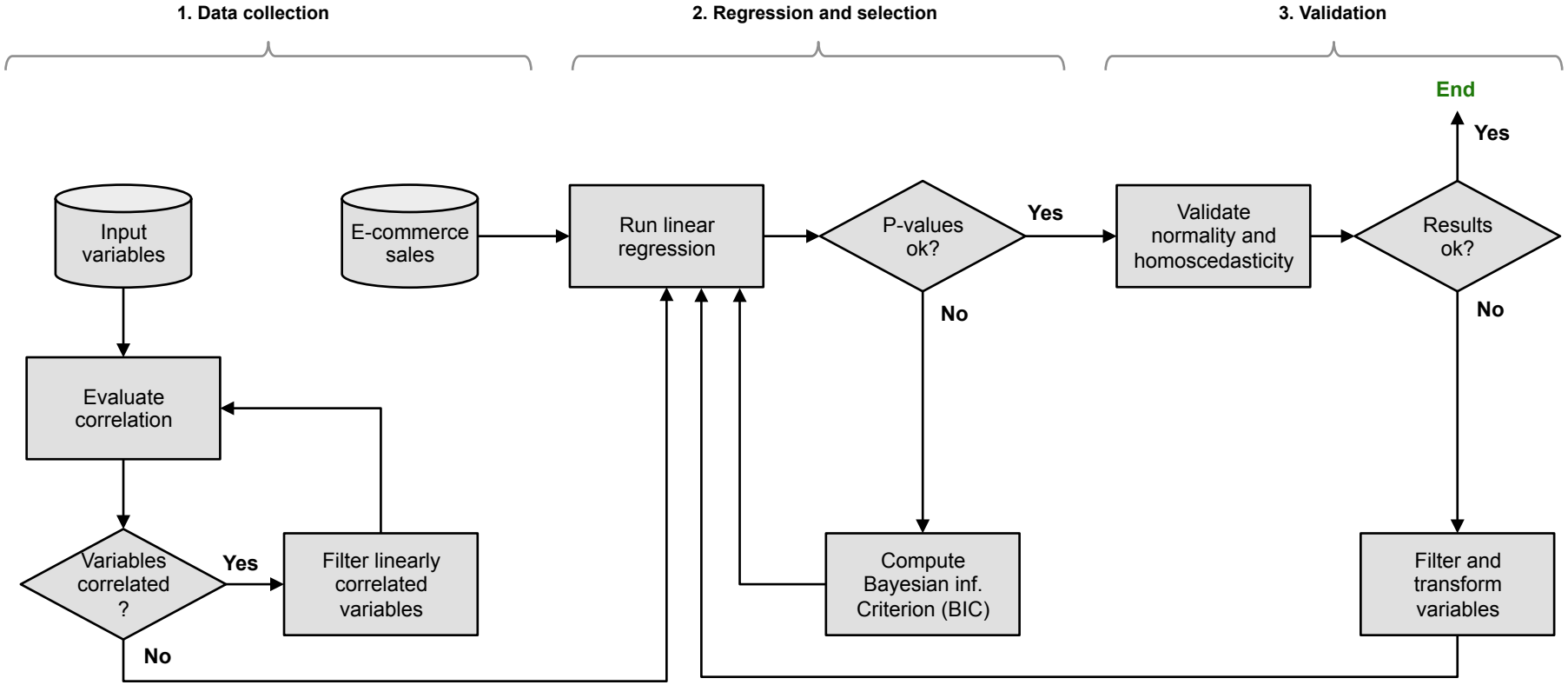
If the Fisher’s p-value is above 0.05 the set is globally irrelevant. In this case, we use the Bayesian Information Criterion, which allows selecting the best-fitted set of variables.

Tests and Validation

Once the subset of parameters is validated in terms of p-values, three graphs resulting from the regression are analyzed: (1) “Residuals versus fitted” that shows the positions of the residuals; (2) “Scale location” that shows the variances of the residuals; and (3) “Residuals vs. leverage” that allows detecting aberrant points.

simon.tamayo@mines-paristech.fr

PROPOSED METHODOLOGY



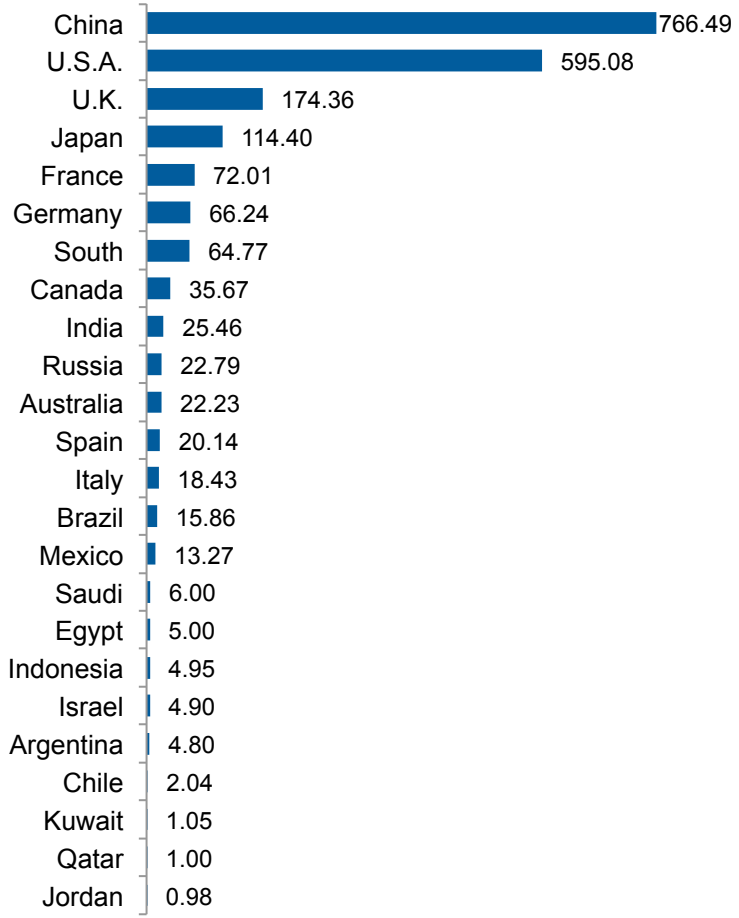
simon.tamayo@mines-paristech.fr

APPLICATION

Input variables

1. Population	19. Urban population growth
2. Population growth	20. Energy use
3. Surface area	21. CO2 emissions
4. Population density	22. Electric power consumption
5. GNI, Atlas method	23. GDP
6. GNI per capita, Atlas method	24. GDP growth
7. GNI, PPP	25. Inflation, GDP deflator
8. GNI per capita, PPP	26. Exports of goods and services
9. Life expectancy at birth	27. Imports of goods and services
10. Fertility rate	28. Gross capital formation
11. Adolescent fertility rate	29. Time required to start a business
12. Mortality rate, under-5	30. Mobile cellular subscriptions
13. Immunization, measles	31. Individuals using the Internet
14. School enrollment, primary	32. High-technology exports
15. School enrollment, secondary	33. Merchandise trade
16. Forest area	34. Net barter terms of trade index
17. Terrestrial and marine protected areas	35. Personal remittances, received
18. Improved sanitation facilities	36. Foreign direct investment net inflows

E-commerce sales (10^6 us$)



simon.tamayo@mines-paristech.fr

REGRESSION RESULTS (1/2)

1st regression results
22 variables
R^2 = 0.4654 (not yet satisfactory)
Fisher's *p-value* = 0.5231

	Estimate	Std. Error	Pr(>|t|)
(Intercept)	6.172e+03	2.159e+06	0.998
Population	1.672e+02	3.637e+02	0.726
Population growth	1.683e+05	2.131e+05	0.574
Surface area	-1.758e+01	1.827e+01	0.512
Population density	-1.205e+03	1.178e+03	0.493
GNI per capita	7.935e+00	8.269e+03	0.999
Fertility rate	-2.645e+05	3.002e+05	0.540
Adolescent fertility	-4.510e+03	7.103e+03	0.640
Immunization measles	7.421e+03	1.074e+04	0.615
School enrollment prim.	7.691e+03	2.295e+04	0.794
School enrollment sec.	-3.092e+03	1.011e+04	0.811
Terr. and marine pro.	-6.950e+03	7.973e+03	0.544
GDP	1.565e-02	3.430e-02	0.727
GDP growth	2.524e+04	8.458e+04	0.815
Inflation	8.252e+02	8.159e+03	0.936
Imports of goods	-9.231e+03	2.014e+04	0.726
Gross capital formation	-9.278e+03	1.437e+04	0.635
Time to start a business	-2.460e+03	6.188e+03	0.759
Mobile cellphone subscriptions	2.637e+03	4.145e+03	0.639
Individuals using the internet	-4.983e+03	1.159e+04	0.742
High technology exports	2.006e+04	1.610e+04	0.431
Merchandise trade	-5.175e+02	9.343e+03	0.965
Net barter term	-1.376e+03	3.787e+03	0.778

Bayesian Information Criterion
14 variables

	Estimate	Std. Error	Pr(>|t|)
None		2.7841e+10	547.29
Gross capital formation	5.1831e+09	3.3024e+10	548.29
Population	7.6372e+09	3.5478e+10	550.01
Mobile cellphone subscriptions	1.2806e+10	4.0647e+10	553.28
Fertility rate	1.4667e+10	4.2508e+10	554.35
Population growth	1.6154e+10	4.3995e+10	555.18
Import of goods	1.6181e+10	4.4022e+10	555.19
Individuals using the internet	1.7300e+10	4.5141e+10	555.79
Immunization measles	1.9147e+10	4.6988e+10	556.76
Adolescent fertility	2.3058e+10	5.0899e+10	558.68
Terrestrial and marine protected	2.4698e+10	5.2539e+10	559.44
Surface area	3.3475e+10	6.1316e+10	563.14
High technology exports	4.7518e+10	7.5359e+10	568.09
Population density	4.7832e+10	7.5673e+10	568.19
GDP	8.1840e+10	1.0968e+11	577.10

2nd regression results
22 variables
R^2 = 0.9139
Fisher's *p-value* = 6.306e-05

	Estimate	Std. Error	Pr(>|t|)
(Intercept)	2.748e+05	3.516e+05	0.454586
Population	1.238e+02	7.877e+01	0.150571
Population growth	1.206e+05	5.276e+04	0.048152
Surface area	-1.743e+01	5.298e+00	0.009384
Population density	-8.596e+02	2.186e+02	0.003446
Fertility rate	-1.518e+05	6.972e+04	0.057413
Adolescent fertility	-4.945e+03	1.811e+03	0.023218
Immunization measles	6.920e+03	2.782e+03	0.034541
Terrestrial and marine protected	-5.989e+03	2.119e+03	0.019865
GDP	2.435e-02	4.734e-03	0.000608
Imports of goods	-6.371e+03	2.786e+03	0.048002
Gross capital formation	-4.446e+03	3.435e+03	0.227746
Mobile cellphone subscriptions	9.998e+02	4.914e+02	0.072391
Individuals using the internet	-6.251e+03	2.643e+03	0.042264
High technology exports	1.716e+04	4.378e+03	0.003515

simon.tamayo@mines-paristech.fr

REGRESSION RESULTS (2/2)

Manual filter
11 variables
Variables with *p-values* > 0.05 were removed

	Estimate	Std. Error	Pr(>|t|)
(Intercept)	2.748e+05	3.516e+05	0.454586
Population	1.238e+02	7.877e+01	0.150571
Population growth	1.206e+05	5.276e+04	0.048152
Surface area	-1.743e+01	5.298e+00	0.009384
Population density	-8.596e+02	2.186e+02	0.003446
Fertility rate	-1.518e+05	6.972e+04	0.057413
Adolescent fertility	-4.945e+03	1.811e+03	0.023218
Immunization measles	6.920e+03	2.782e+03	0.034541
Terrestrial and marine protected	-5.989e+03	2.119e+03	0.019865
GDP	2.435e-02	4.734e-03	0.000608
Imports of goods	-6.371e+03	2.786e+03	0.048002
Gross capital formation	-4.446e+03	3.435e+03	0.227746
Mobile cellphone subscriptions	9.998e+02	4.914e+02	0.072391
Individuals using the internet	-6.251e+03	2.643e+03	0.042264
High technology exports	1.716e+04	4.378e+03	0.003515

3rd regression results
11 variables
R^2 = 0.8915
Fisher's *p-value* = 8.296e-06

	Estimate	Std. Error	Pr(>|t|)
(Intercept)	1.138e+05	2.329e+05	0.63382
Population growth	8.968e+04	2.904e+04	0.00940
Surface area	-1.229e+01	4.957e+00	0.02897
Population density	-6.228e+02	1.825e+02	0.00516
Fertility rate	-1.087e+05	4.149e+04	0.02234
Adolescent fertility rate	-4.062e+03	1.158e+03	0.00432
Immunization measles	7.285e+03	2.577e+03	0.01524
Terrestrial and marine protected	-4.148e+03	1.744e+03	0.03489
GDP	2.889e-02	4.506e-03	3.35e-05
Imports of goods	-5.082e+03	2.193e+03	0.03895
Individuals using the internet	-5.770e+03	1.308e+03	0.00085
High technology exports	1.266e+04	3.060e+03	0.00138

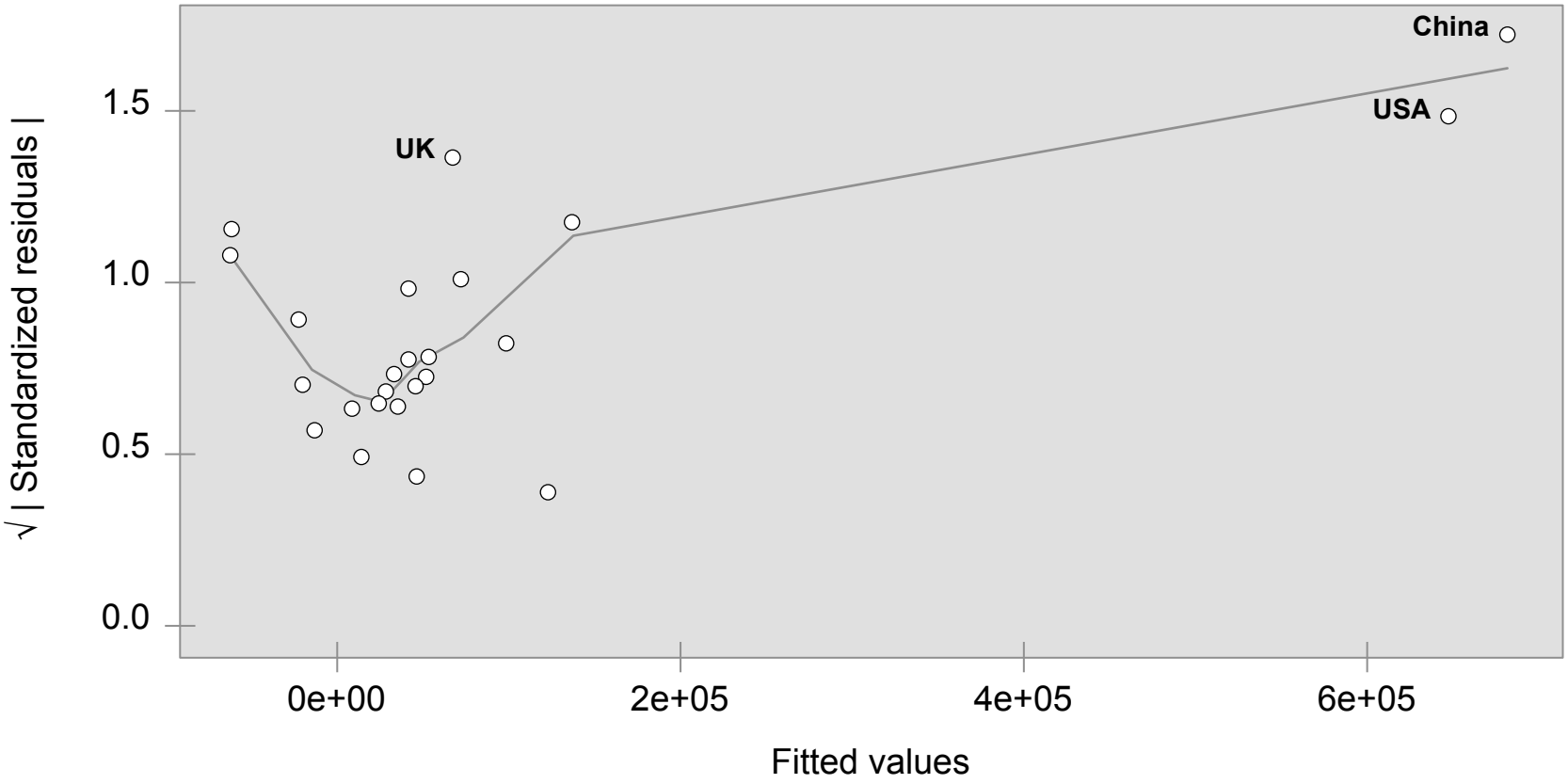
Which major variables influence e-commerce sales?

A set of 11 explanatory variables has been established. It includes the following indicators in order of significance:

1. **GDP**
2. **Individuals using Internet**
3. **High technology exports**
4. **Adolescent fertility rate**
5. **Population density**
6. **Population growth**
7. **Immunization, measles**
8. **Fertility rate**
9. **Surface area**
10. **Terrestrial and marine protected areas**
11. **Imports of goods and services.**

simon.tamayo@mines-paristech.fr

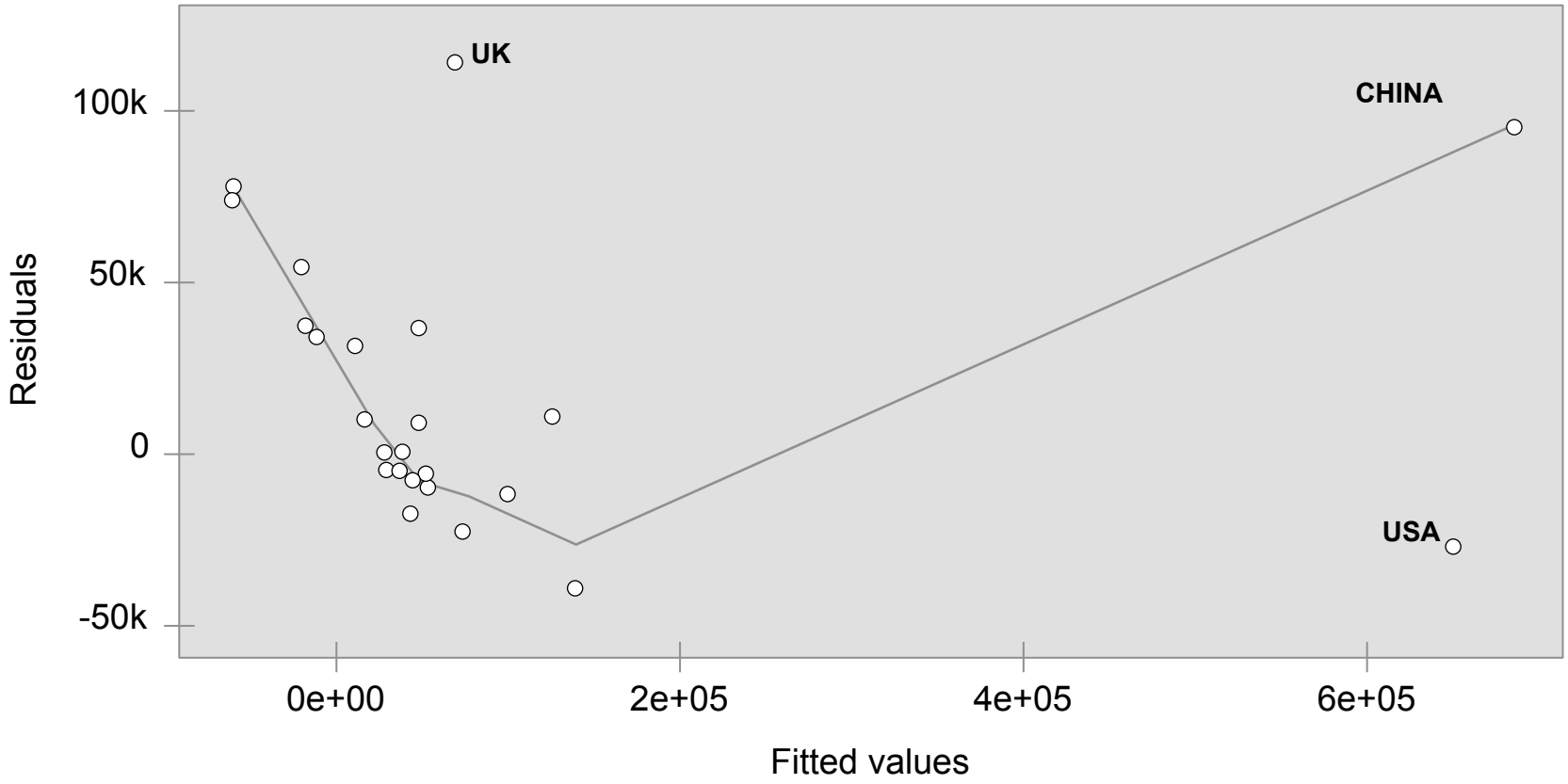
VALIDATION: HOMOSCEDASTICITY



Breush-Pagan test was also performed: the hypothesis of homoscedasticity of the residuals is validated

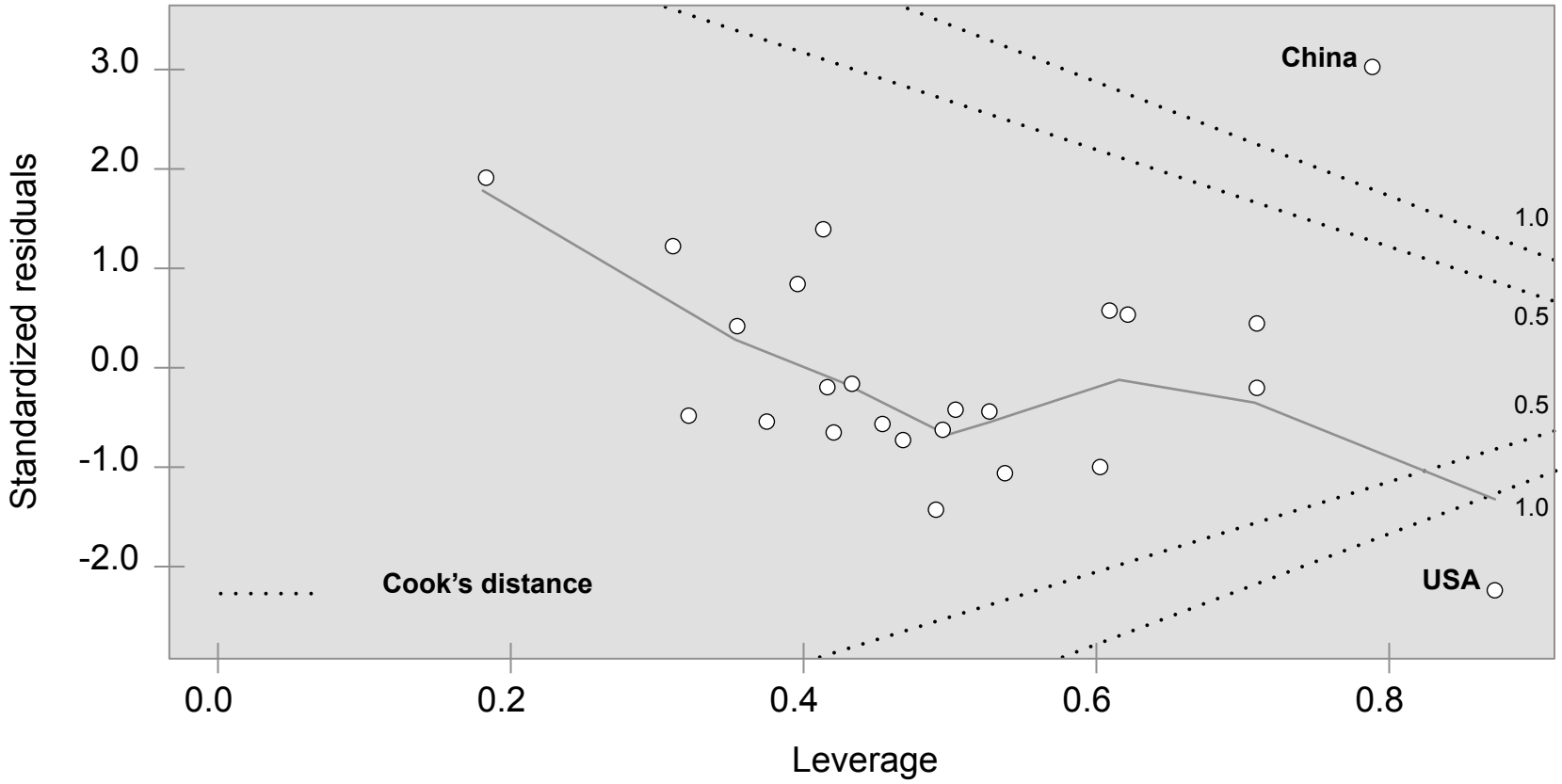
simon.tamayo@mines-paristech.fr

VALIDATION: LINEARITY



simon.tamayo@mines-paristech.fr

VALIDATION: ATYPICAL POINTS



simon.tamayo@mines-paristech.fr

CONCLUSION AND PERSPECTIVES

- This study proposed a methodology to relate e-commerce sales to economic and demographic indicators. A set of 11 explanatory variables were identified.
- A methodology was proposed in order to validate the significance of a set of input indicators.
- Low p-value and high R^2 results indicated that the resulting model is adequate.
- However, the input set of observations is still very limited: only 24 countries form the set of dependent variables.

- The next version of this work will include historical data from several years in order to increase the sample size and identify trends over time.

simon.tamayo@mines-paristech.fr

MERCI POUR VOTRE ATTENTION

QUESTIONS?

Simon TAMAYO
60 bd. Saint-Michel, 75006, Paris
Tel. +33 1 40 51 94 52
simon.tamayo@mines-paristech.fr
www.mines-paristech.fr